J. Dairy Sci. 97:1117–1127
http://dx.doi.org/10.3168/jds.2013-7167

Using the unified relationship matrix adjusted by breed-wise allele frequencies in genomic evaluation of a multibreed population

M. L. Makgahlela,*†[1] I. Strandén,*† U. S. Nielsen,‡ M. J. Sillanpää,§# and E. A. Mäntysaari†
*Department of Agricultural Sciences PO Box 27 FIN-00014 University of Helsinki, Finland
†MTT Agrifood Research Finland, Biotechnology and Food Research, Biometrical Genetics, FIN-31600 Jokioinen, Finland
‡Danish Agricultural Advisory Service, Udkaersvej 15, 8200 Aarhus, Denmark
§Department of Mathematical Sciences, and
#Department of Biology and Biocenter Oulu, PO Box 3000 FIN-90014 University of Oulu, Finland

ABSTRACT

The observed low accuracy of genomic selection in multibreed and admixed populations results from insufficient linkage disequilibrium between markers and trait loci. Failure to remove variation due to the population structure may also hamper the prediction accuracy. We verified if accounting for breed origin of alleles in the calculation of genomic relationships would improve the prediction accuracy in an admixed population. Individual breed proportions derived from the pedigree were used to estimate breed-wise allele frequencies (AF). Breed-wise and across-breed AF were estimated from the currently genotyped population and also in the base population. Genomic relationship matrices (**G**) were subsequently calculated using across-breed ($\mathbf{G_{AB}}$) and breed-wise ($\mathbf{G_{BW}}$) AF estimated in the currently genotyped and also in the base population. Unified relationship matrices were derived by combining different **G** with pedigree relationships in the evaluation of genomic estimated breeding values (GEBV) for genotyped and ungenotyped animals. The validation reliabilities and inflation of GEBV were assessed by a linear regression of deregressed breeding value (deregressed proofs) on GEBV, weighted by the reliability of deregressed proofs. The regression coefficients (b_1) from $\mathbf{G_{AB}}$ ranged from 0.76 for milk to 0.90 for protein. Corresponding b_1 terms from $\mathbf{G_{BW}}$ ranged from 0.72 to 0.88. The validation reliabilities across 4 evaluations with different **G** were generally 36, 40, and 46% for milk, protein, and fat, respectively. Unexpectedly, validation reliabilities were generally similar across different evaluations, irrespective of AF used to compute **G**. Thus, although accounting for the population structure in $\mathbf{G_{BW}}$ tends to simplify the blending of genomic- and pedigree-based relationships, it appeared to have little effect on the validation reliabilities.

Key words: allele frequency, adjusted unified relationship matrix, genomic estimated breeding value, admixed cattle population

INTRODUCTION

Genomic evaluations use genome-wide dense SNP data to predict individual breeding values to be used for selection (Meuwissen et al., 2001). Several reports have shown encouraging results in the application of genomic evaluations within breeds (Hayes et al., 2009b; Su et al., 2010). In dairy cattle, evaluations allow breeders to identify genetically superior bulls at a much earlier age and have been widely applied for breeding purposes (Hayes et al., 2009b; Kearney et al., 2009; Reinhardt et al., 2009; Su et al., 2010). Recent studies have tackled prospects of genomic evaluations in combined purebred (Hayes et al., 2009a; Pryce et al., 2011; Olson et al., 2012) and admixed (Brøndum et al., 2011; Makgahlela et al., 2013a) populations, and have emphasized the potential of this method for multibreed evaluations. The studies concluded that the prediction accuracy across multiple populations was more than that of the parental average but not as effective as the prediction within breeds. The observed low accuracy has been associated with structured reference populations and insufficient linkage disequilibrium (**LD**) between SNP markers and QTL (de Roos et al., 2009; Ibáñez-Escriche et al., 2009).

Genomic evaluations in dairy cattle are generally implemented in multiple steps (Van Doormaal et al., 2009; VanRaden et al., 2009; Harris and Johnson, 2010; Su et al., 2012b). Each step is characterized by different parameters and assumptions, which may overall compromise the prediction performance. In a single-step approach, the pedigree and genomic information are incorporated into a single relationships matrix, which then enters the mixed-model equations for simultaneous prediction of genomic EBV (**GEBV**) for genotyped and ungenotyped individuals (Misztal et al., 2009; Christensen and Lund, 2010). In comparison to multi-

Received June 21, 2013.
Accepted October 16, 2013.
[1]Corresponding author: mahlako.makgahlela@helsinki.fi

step approaches, this provides a unified framework, minimizes possible errors, and provides the opportunity for more accurate genomic predictions (Aguilar et al., 2010). Although computationally expensive, the prediction accuracies from single-step analyses are generally higher than observed for multi-step procedures (Aguilar et al., 2010; Forni et al., 2011).

A crucial component in single-step analyses is the optimal construction of **G**, and optimal weighting of the pedigree and genomic information (Forni et al., 2011; Meuwissen et al., 2011; Christensen et al., 2012). Genomic relationships capture unrecorded pedigrees, and are expressed relative to the mean relatedness of the genotyped individuals, because genomic relationship matrices (**G**) are often calculated using allele frequencies (**AF**) of the genotyped individuals (VanRaden, 2008; Yang et al., 2010). To be compatible with the pedigree-based relationship matrix (**A**), **G** is scaled with complex scaling parameters (Chen et al., 2011; Meuwissen et al., 2011; Christensen et al., 2012). In multi-breeds, the construction of **G** using AF across breeds, however, tended to increase **G** coefficients for animals that were distant from the mean relatedness of the genotyped population (Simeone et al., 2011; Makgahlela et al., 2013b), and closely followed **A** when AF within breeds were used to derive **G** (Makgahlela et al., 2013b). It was earlier found that **G** computed with either AF within or across breeds generated similar validation reliabilities using the genomic BLUP (**GBLUP**) of genotyped animals only, due to inclusion of the breed mean into the model (Makgahlela et al., 2013b).

We hypothesize that if the accuracy of genomic evaluations depends on the structure of the reference population, then accounting for breed origin of alleles in the calculation of **G** in crossbreeds could improve the reliability of GEBV estimated using the unified relationship matrix (single-step GBLUP, **ssGBLUP**). The Nordic Red dairy cattle (**RDC**) population has been shown to have a cross-breeding structure (Brøndum et al., 2011; Makgahlela et al., 2013a), low marker-QTL LD, and large effective population size (Rius-Vilarrasa et al., 2011). Therefore, we investigated if accounting for breed composition in **G** calculated with currently genotyped AF could improve the reliability of GEBV. Further, we studied if there would be gain in reliability if **G** is calculated using estimated base population AF.

MATERIALS AND METHODS

Phenotype and Genotype Data

Data were deregressed breeding values (deregressed proofs, **DRP**) for 2,816,745 cows derived from the EBV using an iterative procedure of Jairath et al. (1998) and Schaeffer (2001). The cow EBV for milk, protein, and fat and their corresponding effective daughter contributions (**EDC**) were obtained from March 2010 official Nordic cattle genetic evaluations (http://www.nordicebv.info/Routine+evaluation/). The cow EDC was calculated using ApaX99 software, with an approach described by Interbull (2004) but excluding information provided by the dam. Cows with records had an EDC, indicating the amount of information in the individual animal. Deregression was carried out using the DeRegress option (Strandén and Mäntysaari, 2010) in the MiX99 software package (Lidauer and Strandén, 1999) with full animal model pedigree file. The heritabilities used in deregression were those reported to Interbull (Table 1). In deregression, individual EBV were weighted by their EDC, and the reliability of DRP was as follows: $r^2_{DRP_i} = EDC_i/(EDC_i + \lambda)$, with EDC_i being EDC for individual i and λ being the variance ratio for the trait analyzed. Hereafter, cow DRP and their corresponding EDC contained the original information that would allow the animal model to solve back the original EBV for bulls and cows (Mäntysaari et al., 2011)

Genotype data for bulls were obtained using the Illumina Bovine SNP50 BeadChip (Illumina Inc., 2005). Markers from the X chromosome, without map position in the UMD3.0 genome assembly (Zimin et al., 2009) and with call rate less than 5% in a large reference sample of Danish Holstein bulls analyzed in the same laboratory were discarded. Further edits removed marker loci with minor allele frequency less than 5% and animal genotypes with a GenCall score (Illumina Inc., 2005) less than 60%. Finally, we imputed genotypes for missing markers using fastPHASE software (Scheet and Stephens, 2006). After pruning, 4,106 genotyped bulls with 38,194 SNP markers were available for our analyses. These data were divided into a training set of 3,300 bulls and a test set of 806 young bulls. Routine evaluations of the 2005 Nordic cattle genetic evaluations (genetic evaluation is available at

Table 1. The parameters used in the deregression of breeding values (deregressed proofs, DRP): the heritability (h^2), variance ratio $\left(\lambda = \frac{1-h^2}{h^2}\right)$, and mean reliability of DRP $\left(\bar{r}^2_{DRP}\right)$ for bulls in the training and testing data sets by trait

			$\bar{r}^2_{DRP}$	
Trait	h^2	λ	Training bulls	Testing bulls
Milk	0.40	1.50	0.96	0.95
Protein	0.28	2.57	0.95	0.93
Fat	0.32	2.13	0.95	0.94