J. Dairy Sci. 96:5364–5375
http://dx.doi.org/10.3168/jds.2012-6523

The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations

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, 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, Oulu, Finland

ABSTRACT

Different approaches of calculating genomic measures of relationship were explored and compared with pedigree relationships (**A**) within and across base breeds in a crossbreed population, using genotypes for 38,194 loci of 4,106 Nordic Red dairy cattle. Four genomic relationship matrices (**G**) were calculated using either observed allele frequencies (AF) across breeds or within-breed AF. The **G** matrices were compared separately when the AF were estimated in the observed and in the base population. Breedwise AF in the current and base population were estimated using linear regression models of individual genotypes on breed composition. Different **G** matrices were further used to predict direct estimated genomic values using a genomic BLUP model. Higher variability existed in the diagonal elements of **G** across breeds (standard deviation = 0.06, on average) compared with **A** (0.01). The use of simple observed AF across base breeds to compute **G** increased coefficients for individuals in distantly related populations. Estimated breedwise AF reduced differences in coefficients similarly within and across populations. The variability of the current adjusted **G** matrix decreased from 0.055 to 0.035 when breedwise AF were estimated from the base breed population. The direct estimated genomic values and their validation reliabilities were, however, unaffected by AF used to compute **G** when estimated with a genomic BLUP model, due to inclusion of breed means in the model. In multibreed populations, **G** adjusted with breedwise AF from the founder population may provide more consistency among relationship coefficients between genotyped and ungenotyped individuals in an across-breed single-step evaluation.
Key words: allele frequency, genomic relationship matrix, admixed population, genomic prediction

INTRODUCTION

The use of marker genotypes to estimate relationships among individuals in a population has become increasingly important in many fields of genetics. In livestock breeding, knowledge of relationships is used routinely to estimate genetic variation and animal breeding values (EBV; VanRaden, 2008; Hayes et al., 2009; Su et al., 2012), monitor inbreeding (Fernández et al., 2011; Toro et al., 2011), and for conservation of animal genetic resources (Eding and Meuwissen, 2001). Traditionally, relationship coefficients are calculated from the pedigree data. Pedigree relationships are obtained as 2 times the expected average identity by descent (**IBD**) sharing between 2 relatives (Malécot, 1948) and have been applied successfully within the framework of mixed-model equations for best linear unbiased prediction of EBV. Presently, with the increasing availability of genetic markers covering the whole genome, pedigree-based relationships can be replaced or combined with realized relationships calculated from marker data in the prediction of genomic breeding values (genomic EBV; Habier et al., 2007; Hayes et al., 2009)

Realized relationships derived from molecular markers are based on the actual IBD sharing or identity by state for genomic regions and, therefore, have more variation between closely related animals than pedigree relationships (VanRaden, 2008; Hayes et al., 2009). Moreover, realized relationships capture unrecorded pedigrees. Several different methods of calculating genomic relationship matrices (**G**) have been developed, for genotyped animals only (VanRaden, 2008; Yang et al., 2010) and when genotyped and ungenotyped individuals are combined (Legarra et al., 2009; Misztal et al., 2009; Christensen and Lund, 2010). In the latter approach, an arbitrary weight on pedigree relationships (**A**) is often used to measure the amount of variation not explained by markers. Although variability exists in the accuracy of predictions among the above methods (Forni et al., 2011), generally these accuracies are at

---

Received December 24, 2012.
Accepted April 24, 2013.
[1] Corresponding author: mahlako.makgahlela@helsinki.fi

least twice of those estimated from pedigree data only (see, for example, Su et al., 2012).

Genomic relationship matrix **G** can be constructed by using a matrix having genotype information for each individual and marker (VanRaden, 2008). Each genotype is a deviation from the marker-specific population mean, which is calculated using population allele frequencies. The estimation of **G** has been shown to be closer to **A** when inferences use allele frequencies (**AF**) in the distant ancestral population (VanRaden, 2008; VanRaden et al., 2011) instead of AF in the currently genotyped population. This is because the expected **G** coefficients would be expressed relative to the same base population as **A**. The limitation is that base population AF are generally not available with field data and their estimation can be challenging. Gengler et al. (2007) showed an efficient way of calculating gene content and base population AF within a breed. However, in practice, currently genotyped populations are assumed to be the base population. This results from centering the genotype matrix used to build **G** with current-data AF so that the average genomic relationship between animals within the current population become 0 and scaling **G** such that the additive genetic variance would be comparable to that obtained through conventional methods (Powell et al., 2010; Forni et al., 2011).

The use of observed AF within a breed may not have major practical implications in genomic BLUP (**GBLUP**) models. In the context of structured populations, the effect of using across-breed AF to make **G** may have consequences for the estimation of relationships, mainly attributable to varying source of AF between breeds. Eding and Meuwissen (2001) demonstrated that average relatedness between 2 populations could be expressed in terms of population-specific AF. VanRaden et al. (2011) used the average of 3 breedwise AF for estimation in the combined 3-breed population. Although these approaches would be beneficial for multiple populations with distinctive subpopulations, a need still exists for approaches in populations that constitute mainly crossbred animals. The Nordic Red dairy cattle (**RDC**) comprise 3 subpopulations by country of birth [i.e., Denmark (**DNK**), Sweden (**SWE**), and Finland (**FIN**)]. Over the years of crossbreeding, the majority of animals (~98%) in the Nordic RDC are composites of base breeds. The absence of pure base-breed animals remains a major challenge for the estimation of breedwise AF. The objective of this study was to investigate whether the use of estimated breedwise AF in the calculation of genomic relationships would provide a more accurate estimate of **G** than using AF across breeds, and to determine the effect on **G** when AF are estimated in the base population versus the currently genotyped population.

MATERIALS AND METHODS

Data

This study was carried out in a structured population with 60 pure base breed bulls and 4,046 bulls of combinations of base breeds in the Nordic RDC. Genotypes for all 4,106 bulls were attained using the Illumina Bovine SNP50 BeadChip (Illumina Inc., San Diego, CA). For quality purposes, markers from the X chromosome, without map position in the UMD3.0 genome assembly (Zimin et al., 2009) and with minor allele frequency (**MAF**) <5% were discarded. In addition, animal genotypes with a GenCall score (Illumina Inc., 2005) <60% and marker loci with call rates <5% in a large reference sample from the same genotyping laboratory, consisting of Danish Holstein bulls, were discarded. Finally, missing genotypes were imputed using fastPHASE software chromosome by chromosome (Scheet and Stephens, 2006). Due to unavailability of pure base breeds, informative SNP above were selected based on across-breed AF. After quality control, a total of 38,194 SNP markers were available for analyses. The entire RDC pedigree, containing over 4 million records, was used to calculate breed proportions (**BP**) for individual bulls (Lidauer et al., 2006). A breed was defined only if the average BP in the data was greater than 10%. Breeds used in this study were the Swedish Red (**SRB**), Finnish Ayrshire (**FAY**), and Norwegian Red (**NRF**), and the remaining breeds with BP less than 10% were combined into the breed "other." A more detailed description about the population structure, breeds contained, and definitions of the final 4 breeds and their trends is provided by Makgahlela et al. (2013). The pedigree for genotyped bulls contained 22,300 animals.

Phenotypes were individual daughter deviations (**IDD**) of 1,995,606 RDC cows for milk, protein, and fat yields, obtained from March 2010 official evaluations of the Nordic cattle genetic evaluations. By definition, the IDD are cow performances adjusted for fixed effects, nongenetic random effects, and genetic effects of the cow's dam (Mrode and Swanson, 2004). Here, however, IDD were computed by using animal model deregression from 305-d combined EBV (Mäntysaari et al., 2011). For validation of predictions using different **G** matrices, the data were split into sets of 3,300 training bulls born between 1980 and 1999, and 806 validation bulls born between 1998 and 2005. The training data had older bulls, which were evaluated for the first time during the 2005 Nordic cattle routine evaluation.

Estimation of Pedigree and Genomic Relationships

Pedigree relationships of genotyped bulls were estimated using the RelaX2 computer program (Strandén