J. Dairy Sci. 96:647–654
http://dx.doi.org/10.3168/jds.2012-5656

Methods to approximate reliabilities in single-step genomic evaluation

I. Misztal,*[1] S. Tsuruta,* I. Aguilar,† A. Legarra,‡ P. M. VanRaden,§ and T. J. Lawlor#
*Department of Animal and Dairy Science, University of Georgia, Athens 30602-2771
†Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay
‡INRA, UR631-SAGA, BP 52627, 31326 Castanet-Tolosan Cedex, France
§Animal Improvement Programs Laboratory, Agricultural Research Service, US Department of Agriculture, Beltsville, MD 20705-2350
#Holstein Association USA Inc., Brattleboro, VT 05302-0808

ABSTRACT

Reliability of predictions from single-step genomic BLUP (ssGBLUP) can be calculated by matrix inversion, but that is not feasible for large data sets. Two methods of approximating reliability were developed based on the decomposition of a function of reliability into contributions from records, pedigrees, and genotypes. Those contributions can be expressed in record or daughter equivalents. The first approximation method involved inversion of a matrix that contains inverses of the genomic relationship matrix and the pedigree relationship matrix for genotyped animals. The second approximation method involved only the diagonal elements of those inverses. The 2 approximation methods were tested with a simulated data set. The correlations between ssGBLUP and approximated contributions from genomic information were 0.92 for the first approximation method and 0.56 for the second approximation method; contributions were inflated by 62 and 258%, respectively. The respective correlations for reliabilities were 0.98 and 0.72. After empirical correction for inflation, those correlations increased to 0.99 and 0.89. Approximations of reliabilities of predictions by ssGBLUP are accurate and computationally feasible for populations with up to 100,000 genotyped animals. A critical part of the approximations is quality control of information from single nucleotide polymorphisms and proper scaling of the genomic relationship matrix.
Key words: genomic prediction, reliability, single-step evaluation, best linear unbiased predictor

INTRODUCTION

A single-step genomic BLUP (**ssGBLUP**) is a modification of BLUP to use genomic information. In ssGBLUP, the pedigree-based numerator relationship matrix (**A**) and a relationship matrix based on genomic information (**G**) are combined into a single matrix **H** (Legarra et al., 2009). The inverse of **H** has a simple form and can substitute for the inverse of **A** in existing software (Aguilar et al., 2010). Compared with multistep methods (VanRaden, 2008), ssGBLUP is simpler and applicable to complicated models. The ssGBLUP has been successfully used for chickens (Chen et al., 2011b), pigs (Forni et al., 2011), and dairy cattle (Aguilar et al., 2010, 2011b; Tsuruta et al., 2011). The computing limit of ssGBLUP is currently up to about 100,000 genotypes of progeny-tested animals (Aguilar et al., 2011a), with no limit on the number of animals or traits. However, recent developments (Ducrocq and Legarra, 2011; Legarra et al., 2011) may allow ssGBLUP to be used with an unlimited number of genotypes.

In a genetic evaluation, computing reliability of EBV is of interest. When the system of equations is small, reliability can be computed by direct matrix inversion of the BLUP mixed-model equations (Henderson, 1984). When the system of equations is large, inversion is impossible and reliability needs to be approximated. Several approximation methods for animal models exist for nongenomic evaluations. The approximation method of Misztal and Wiggans (1988), which is easy to compute, involves the effective number of records and a sum of contributions to an animal from its parents and progeny. That approximation is iterative, although a noniterative modification exists (VanRaden and Wiggans, 1991). The approximation method of Misztal and Wiggans (1988) was extended to repeatability models (Wiggans et al., 1988; Misztal et al., 1993), multiple-trait models that include maternal effect (Strabel et al., 2001), and random regression models (Sánchez et al., 2008). The advantage of approximation is simplicity and computing ease.

An approximation of reliability when genomic information is available needs to fulfill a few obvious conditions. First, more genotypes should result in equal or higher reliability. Second, a young genotyped animal should create no additional information for other animals. Third, the extra information contributed to the reference population should be small or none for a

Received April 24, 2012.
Accepted September 18, 2012.
[1]Corresponding author: ignacy@uga.edu

young animal with ancestors that are not genotyped. However, a young animal should contribute information to its nongenotyped parents. For example, genotypes can be imputed for nongenotyped parents that have several genotyped progeny. Similarly, the single-step equations adjust parent EBV through linear rather than nonlinear imputation methods. Fourth, no extra reliability should be gained for an animal from different lines or breeds. The purpose of this study was to extend the approximation algorithm of Misztal and Wiggans (1988) to ssGBLUP.

MATERIALS AND METHODS

Data

Data were simulated using QMSim software (Sargolzaei and Schenkel, 2009) for an additive trait with heritability of 0.5, 2 chromosomes, and 60 QTL. Performance was simulated for 15,800 individuals in 5 generations, and 1,500 individuals of the last 3 generations were genotyped. Each animal in the simulation had a single phenotypic record. Details of the simulation were reported by Wang et al. (2012).

Derivation of Approximation Methods

Reliability of animal i (rel_i) can be approximated as $1 - [\alpha/(\alpha + d_i)]$, where α is the ratio of error variance to animal genetic variance and d_i is the amount of information for animal i in units of effective number of records (Misztal and Wiggans, 1988). The information can be calculated by inversion of the left-hand side (**LHS**) of the mixed-model equations as $\text{LHS}^{ii}_{uu} = 1/(\alpha + d_i)$, where uu denotes the block of the LHS for the animal effect for the animal effect and ii denotes the diagonal element corresponding to animal i. Then, d_i can be partitioned as $d^r_i + d^p_i + d^g_i$, where d^r_i is the contribution from records (phenotypes), d^p_i is the contribution from pedigrees, and d^g_i is the contribution from genomic information. With pedigree information, contributions to an animal are from progeny and parents only. With genomic information, contributions are from all animals with genomic information.

For simplicity, assume a single-trait mixed model:

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

where $\mathbf{y}$ is a vector of observations, $\mathbf{b}$ is a fixed effect, $\mathbf{u}$ is the random additive animal effect, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices relating $\mathbf{b}$ and $\mathbf{u}$ to $\mathbf{y}$, and $\mathbf{e}$ is the random residual effect. When relationships are known, LHS is

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\boldsymbol{\alpha} \end{bmatrix},$$

where $\mathbf{A}$ is the numerator relationship matrix, and the diagonal elements of the inverse of the LHS for animal i can be presented as

$$\text{LHS}^{ii}_{uu} = 1/\left(\alpha + d^r_i + d^p_i\right). \quad [1]$$

If $\mathbf{D}^r = \left\{d^r_i\right\}$ and $\mathbf{D}^p = \left\{d^p_i\right\}$ are known, Equation 1 can be simplified to $\text{LHS}^{ii}_{uu} = \left[\left(\mathbf{D}^r_i + \mathbf{D}^p_i + \mathbf{I}\boldsymbol{\alpha}\right)^{-1}\right]_{ii}$, where $\mathbf{I}$ is an identity matrix, or approximated as

$$\text{LHS}^{ii}_{uu} \approx \left[\left(\mathbf{D}^r_i + \mathbf{A}^{-1}\boldsymbol{\alpha}\right)^{-1}\right]_{ii}.$$

Misztal and Wiggans (1988) estimated the contributions from relationships separately for each relationship in an iterative formula:

$$\begin{bmatrix} 1.5\alpha + d_s - d^r_{s_i} & 0.5\alpha & -\alpha \\ 0.5\alpha & 1.5\alpha + d_d - d^r_{d_i} & -\alpha \\ -\alpha & -\alpha & 2\alpha + d_i - d^r_{i_s} - d^r_{i_d} \end{bmatrix}^{-1} = \begin{bmatrix} 1/(\alpha + d_s) & \ldots & \ldots \\ \ldots & 1/(\alpha + d_d) & \ldots \\ \ldots & \ldots & 1/(\alpha + d_i) \end{bmatrix},$$

where d_i, d_s, and d_d are total amounts of information from animal i and its sire (s) and dam (d), respectively; $d^r_{s_i}$ and $d^r_{d_i}$ are contributions to sire and dam information from records of animal i, respectively; and $d^r_{i_s}$ and $d^r_{i_d}$ are contributions to information for animal i from records of its sire and dam, respectively. Nonmatrix formulas for the same contributions, but expressed in daughter equivalents, were derived by VanRaden and Wiggans (1991).

When genomic information is available, the LHS of ssGBLUP is

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{H}^{-1}\boldsymbol{\alpha} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\boldsymbol{\alpha} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G}^{-1} - \mathbf{A}^{-1}_{22} \end{bmatrix}\boldsymbol{\alpha} \end{bmatrix},$$