Approximation of Genomic Accuracies in Single-Step Genomic Evaluation

***I. Misztal*[1], *S. Tsuruta*[1], *I. Aguilar*[2], *A. Legarra*[3] *and T.J. Lawlor*[4]**
[1]*Department of Animal and Dairy Science, University of Georgia, Athens, GA 30602, USA*
[2]*Instituto Nacional de Investigación Agropecuaria, Las Brujas 90200, Uruguay*
[3]*INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France*
[4]*Holstein Association USA Inc., Brattleboro, VT 05302-0808, USA*

Abstract

Reliability of predictions from single-step genomic BLUP (ssGBLUP) can be calculated by inversion, but that is not feasible for large data sets. Two proposed approximations of reliability are based on decomposition of a function of reliability into contributions from records, pedigree, and genotypes. The first approximation involves inversion of a matrix that contains inverses of the genomic relationship matrix (**G**) and the pedigree relationship matrix for genotyped animals ($\mathbf{A}_{22}$). The second approximation involves only the diagonal elements of those inverses. The approximations were tested with a simulated data set. The correlations between exact and approximated contributions due to genomic information were 0.92 for the first approximation and 0.56 for the second approximation; contributions were inflated 60 and 260%, respectively. The respective correlations for reliabilities were 0.98 and 0.72. After correction for inflation, those correlations increased to 0.99 and 0.89. Approximations of reliabilities of predictions by ssGBLUP are accurate and computationally feasible. A critical part of the approximations is quality control of SNP information and proper scaling of **G**.

Key words: genomic prediction, accuracy, reliability, single-step evaluation, BLUP

Introduction

A single-step genomic BLUP (ssGBLUP) is a modification of BLUP to use genomic information. In ssGBLUP, the pedigree-based 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 to 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.*, 2010), and dairy cattle (Aguilar *et al.*, 2010; Tsuruta *et al.*, 2011; Aguilar *et al.*, 2011b). 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. Recent developments (Legarra *et al.*, 2011; Ducrocq and Legarra, 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 inversion. When the system of equations is large, inversion is impossible and reliability needs to be approximated. Several approximations for animal models exist for non-genomic evaluations. An approximation by Misztal and Wiggans (1998) that is easy to compute involves the effective number of records and a sum of contributions to an animal from its parents and progeny. This approximation is iterative although a non-iterative modification exists (VanRaden and Wiggans, 1991). The approximation of Misztal and Wiggans (1998) was extended to repeatability (Wiggans *et al.*, 1988; Misztal *et al.*, 1993), multiple-trait including maternal effect (Strabel *et al.*, 2001), and random regression (Sánchez *et al.*, 2008) models. 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 result in equal or higher reliability. Second, a young genotyped animal creates no additional information for other animals. Third, the extra information from genomics is small or none for a young animal with ancestors that are not genotyped. Fourth, no extra reliability is gained for an animal from different lines or breeds.

The purpose of this study was to extend the algorithm of Misztal and Wiggans (1988) to ssGBLUP.

Data

Data were simulated using QMSim (Sargolzaei and Schenkel, 2009) for an additive trait with heritability of 0.5, two chromosomes, and 60 QTL. Performance was simulated for 15,800 individuals in five generations, and 1,500 individuals of the last three generations were genotyped.

Derivations

Reliability of animal i (rel_i) can be approximated as $1 - [\alpha/(\alpha + d_i)]$, where α is the variance ratio 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 $\mathrm{LHS}_{uu}^{ii} = 1/(\alpha + d_i)$, where u is ?. Then d_i can be partitioned as $d_i^r + d_i^p + d_i^g$, where d_i^r is contribution from records (phenotypes), d_i^p is contribution from pedigrees, and d_i^g is 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 with the additive animal effect as the only random effect. When relationships are known, LHS is

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

and the diagonal elements of the inverse of LHS for animal i can be presented as $\mathrm{LHS}_{uu}^{ii} = 1/(\alpha + d_i^r + d_i^p)$. If $\mathbf{D}^r = \{\mathbf{d}_i^r\}$ and $\mathbf{D}^p = \{\mathbf{d}_i^p\}$ are known, the formula can be simplified to $\mathrm{LHS}_{uu}^{ii} = [(\mathbf{D}_i^r + \mathbf{D}_i^p + \alpha\mathbf{I})^{-1}]_{ii}$

or approximated as

$$\mathrm{LHS}_{uu}^{ii} \approx [(\mathbf{D}_i^r + \alpha\mathbf{A}^{-1})^{-1}]_{ii}.$$

Misztal and Wiggans (1988) estimated the contributions from relationships separately for each relationship in an iterative formula. Non-matrix formulas for the contributions were derived by VanRaden and Wiggans (1991).

When genomic information is available,

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

and the diagonal elements of the inverse of LHS for animal i are $\mathrm{LHS}_{uu}^{ii} = 1/(\alpha + d_i^r + d_i^p + d_i^g)$. If $\mathbf{D}^r$ and $\mathbf{D}^p$ are known, the formula can be approximated as

$$\mathrm{LHS}_{guu}^{ii} \approx \{[\mathbf{D}_i^r + \mathbf{D}_i^p + \alpha(\mathbf{I}^{-1} + \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})]^{-1}\}_{ii}.$$

In this equation, $\mathbf{G}$ accounts for genomic information, and $\mathbf{A}_{22}$ accounts for an adjustment to prevent double counting.

The last equation can be the basis for the following algorithm (called Approx1) to approximate reliabilities with genomic information:

1. Approximate reliabilities with an algorithm that ignores genomic information.

2. Convert those reliabilities to effective number of records for genotyped animals only:
   $d_i = \alpha[1/rel_i\text{-}1)]$.

3. Calculate the inverse:
   $\mathbf{Q} = [\mathbf{D} + \alpha(\mathbf{I}^{-1} + \mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})]^{-1}$.

4. Calculate genomic reliabilities: $r_i = 1 - \alpha q^{ii}$.

5. Possibly adjust reliabilities of non-genotyped animals if those are functions of reliabilities of genotyped animals.

Algorithm based on diagonal elements

When off-diagonals of some matrices are ignored, the formula for **Q** can be simplified to

$$\mathbf{Q} = \{\mathbf{D} + \alpha[\mathbf{I}^{-1} + \text{diag}(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})]\ \}^{-1}.$$

This algorithm (called Approx2) is based on findings that diagonal information in $\mathbf{G}^{-1}$ contains the information in $\mathbf{A}^{-1}$ plus genomic information.

Analyses

Total information per animal was calculated using an animal model with pedigree relationship only and using ssGBLUP. Contributions due to genomics were calculated as differences in information from the two analyses. Approximations used the non-genomic information from the pedigree-only analysis. Matrix **G** was constructed using current allele frequencies and subsequently rescaled so that means of diagonal and off-diagonal elements were identical to those of $\mathbf{A}_{22}$ (Chen *et al.*, 2011a; Vitezica *et al.*, 2011). Initially reliabilities were calculated from the sum of all contributions. For approximations only, reliabilities were calculated with genomic contributions regressed to have a mean equal to that for exact contributions.

Results

Table 1 shows statistics for exact and approximated genomic contributions as well as correlations between exact and approximated contributions. Correlation with the exact method was 0.92 for Approx1 and 0.56 for Approx2. Both contributions were inflated: by 60% for Approx1 and by over 3 times for Approx2. Inflation resulted from ignoring off-diagonal elements in $\mathbf{X}'\mathbf{X}$, $\mathbf{Z}'\mathbf{Z}$, and $\mathbf{A}^{-1}$.

Table 2 shows statistics for exact and approximated reliabilities as well as correlations between exact and approximated reliabilities. Correlation with the exact method was 0.98 for Approx1 and 0.72 for Approx2. Both contributions were inflated. After regressing genomic contributions (Table 3), reliabilities were no longer inflated, and correlation with the exact method increased to 0.99 for Approx1 and 0.89 for Approx2. In practice, the coefficient of regression is unknown and has to be derived experimentally, e.g., to match realized reliabilities.

Approx1 is computationally feasible if ssGBLUP is feasible because ssGBLUP requires the inverses of **G** and $\mathbf{A}_{22}$ to be computable. Approx2, a simplification of Approx1, generally offers little benefit over Approx1 except when diagonal elements of the inverses of **G** and $\mathbf{A}_{22}$ can be computed at a low cost.

Table 1. Statistics for genomic contributions from three methods to estimate reliability.

Method	Mean	Range	Correlation with exact contribution
Exact[1]	2.4 ± 0.4	1.7–4.7	—
Approx1	3.9 ± 0.6	2.9–8.3	0.92
Approx2	8.6 ± 4.2	4.5–62	0.56

[1]ssGBLUP.

Table 2. Statistics for reliabilities from three methods to estimate reliability.

Method	Mean, %	Range, %	Correlation with exact contribution
Exact[1]	81 ± 2	77–90	—
Approx1	85 ± 2	83–93	0.98
Approx2	91 ± 2	86–98	0.72

[1]ssGBLUP.

Table 3. Statistics for reliabilities from three methods to estimate reliability after rescaling genomic contributions.

Method	Mean, %	Range, %	Correlation with exact contribution
Exact[1]	81 ± 2	77–90	—
Approx1	81 ± 2	78–92	0.99
Approx2	81 ± 4	75–96	0.89

[1]ssGBLUP.

For Approx1 and Approx2, reliability calculated by inversion is assumed to reflect real reliability. This was confirmed by Hayes *et al.* (2009) in a simulation study. However, predicted reliabilities were inflated compared with realized reliabilities in a study by VanRaden *et al.* (2009). Several explanations exist for the inflation. First, inflation could result from several approximations and assumptions inherent in multiple-step procedures. Second, genetic relationships fade over generations under selection (Muir, 2007), and thus contributions from older generations may be inflated. Third, effects of major genes (if they exist) may not be fully accounted for by the method. Fourth, the analysis model may be deficient, e.g., by ignoring selection, censoring, or preferential treatment. For example, the genetic parameters for several chicken traits in two lines were different between complete data sets or genotyped subsets in chicken (Chen *et al.*, 2011b), and origins of those differences were difficult to explain. Differences among predicted and realized reliabilities were not obvious before the era of genomic selection as interest in realized reliabilities was limited. Probably the best way to tackle the issue of inflated predicted reliabilities is by research on causes of such inflation, both with and without genomic information.

Approx1 and Approx2 are based on differences between $\mathbf{G}$ and $\mathbf{A}_{22}$. Chen *et al.* (2011a) found that number of SNP and assumed allele frequencies affected statistics of $\mathbf{G}$ and $\mathbf{G}^{-1}$. They recommended that $\mathbf{G}$ be constructed with current allele frequencies and then rescaled to match statistics of $\mathbf{A}_{22}$. They also found that decreasing the number of SNP when constructing $\mathbf{G}$ inflated $\mathbf{G}$ (although inflation was small when number of SNP was >20,000). In populations with multiple lines with different allele frequencies (e.g., Simeone et al., 2011), $\mathbf{G}$ needs to be rescaled for different lines to avoid less accurate approximations of accuracy (e.g., Harris and Johnson, 2010). Wang and Misztal (2011) found that for properly scaled $\mathbf{G}$, the SD of a difference between elements of $\mathbf{G}$ and $\mathbf{A}_{22}$ is < 0.04. Similar quantity found by Hill and Weir (2011). Larger differences of up to 1.0 are due genotyping mistakes, pedigree mistakes, incomplete pedigree and mixing of lines. Such differences can lead to inflated approximations of reliability for selected animals.

Figures 1 and 2 show exact and approximated genomic contributions after scaling. Most of the Approx1 contributions are similar, but some are inflated. For Approx2, the fit for most animals is not as good, and inflation for selected animals is larger. Reasons for inflation for some animals will be studied subsequently.

Figures 3 and 4 show exact and approximated reliabilities after scaling. The fit for Approx1 is very good, whereas that for Approx2 is not as good. The fit for reliabilities is better than for genomic contributions because of an upper bound of 1 and the stabilizing effect of contributions from records and pedigrees.

Conclusions

Two algorithms to approximate reliabilities in ssGBLUP were presented. The first algorithm was relatively accurate and inexpensive for <30,000 genotypes. It required some heuristics to regress inflated genomic contributions.

Acknowledgements

This project was supported by AFRI grants 2009-65205-05665 and 2010-65205-20366 from USDA NIFA, Holstein Association USA, and PIC. Discussions with Paul VanRaden and George Wiggans were greatly appreciated. Editorial help by Suzanne Hubbard is gratefully acknowledged.

Figure 1. Exact (y axis) and approximated (x axis) genomic contributions after scaling for Approx1.

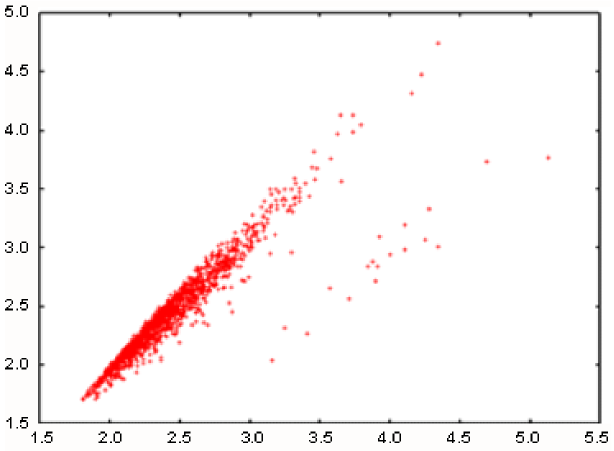


Figure 2. Exact (y axis) and approximated (x axis) genomic contributions after scaling for Approx2.

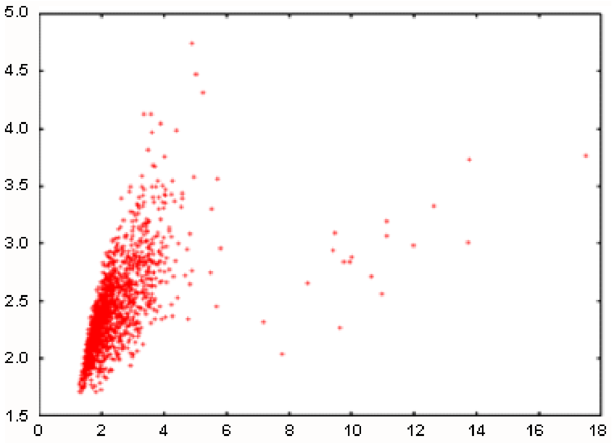


Figure 3. Exact (y axis) and approximated (x axis) genomic reliabilities after scaling for Approx1.

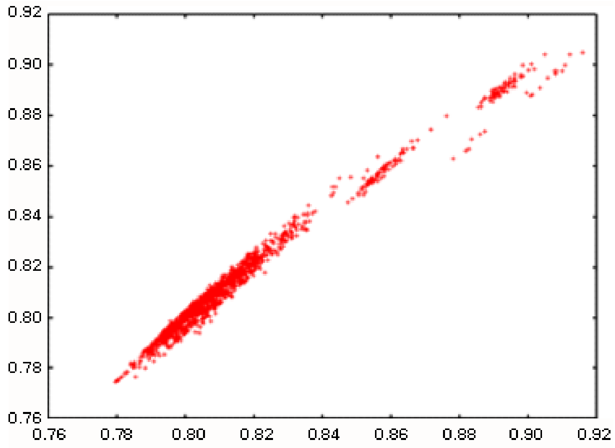
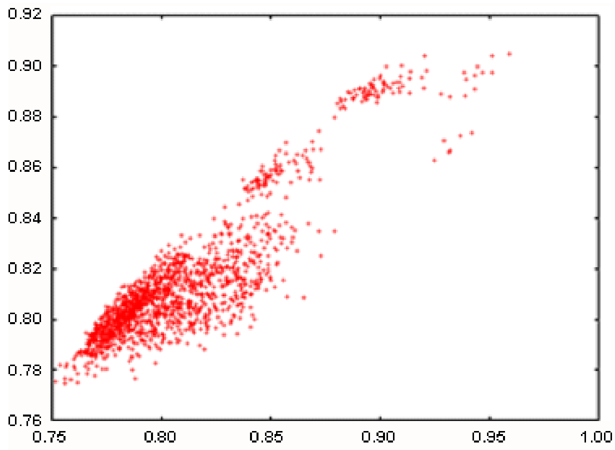


Figure 4. Exact (y axis) and approximated (x axis) genomic reliabilities after scaling for Approx2.



References

Aguilar, I., Misztal, I. Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2010. A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. *J. Dairy Sci. 93,* 743–752.

Aguilar, I., Misztal, I., Legarra, A. & Tsuruta, S. 2011a. Efficient computation of the genomic relationship matrix and other matrices used in single-step evaluation. *J. Anim. Breed. Genet.* doi: 10.1111/j.1439-0388.2010.00912.x.

Aguilar, I., Misztal, I. Tsuruta, S., Wiggans, G.R. & Lawlor, T.J. 2011b. Multiple trait genomic evaluation of conception rate in Holsteins. *J. Dairy Sci. 94,* 2621–2624.

Chen, C.Y., Misztal, I., Aguilar, I., Legarra, A. & Muir, B. 2011a. Effect of different genomic relationship matrix on reliability and scale. *J. Anim. Sci.* (accepted)

Chen, C.Y., Misztal, I., Aguilar, I., Tsuruta, S., Meuwissen, T.H.E., Aggrey, S.E., Wing, T. & Muir, W.M. 2011b. Genome-wide marker-assisted selection combining all pedigree phenotypic information with genotypic data in one step: An example using broiler chickens. *J. Anim. Sci. 89,* 23–28.

Ducrocq, V. & Legarra, A. 2011, An iterative implementation of the single step approach for genomic evaluation which preserves existing genetic evaluation models and software. *Interbull Bulletin 44,* 138-142.

Forni, S., Aguilar, I. & Misztal, I. 2011. Different genomic relationship matrices for single-step analysis using phenotypic, pedigree and genomic information. *Genet. Sel. Evol. 43,* 1.

Harris, B.L. & Johnson, D.L. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci. 93,* 1243–1252.

Hayes, B.J., Visscher, P.M. & Goddard, M.E. 2009. Increased reliability of artificial selection by using the realized relationship matrix. *Genet. Res. 91,* 47–60.

Hill, W.G. & Weir, B.S. 2011. Variation in actual relationship as a consequence of Mendelian sampling and linkage. *Genet. Res. 93,* 47–64.

Legarra, A., Aguilar, I. & Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci. 92,* 4656–4663.

Legarra, A., Misztal, I. & Aguilar, I. 2011. The single step: genomic evaluations for all. *Proc. European Assoc. Anim. Prod.*, Stavanger, Norway, 17:1.

Misztal, I., Lawlor, T.J. & Short, T.H. 1993. Implementation of single- and multiple-trait animal models for genetic evaluation of Holstein type traits. *J. Dairy Sci. 76,* 1421–1432.

Misztal, I. & Wiggans, G.R. 1988. Approximation of prediction error variance in large-scale animal models. *J. Dairy Sci. 71 (Suppl. 2),* 27–32.

Muir, W.M. 2007. Comparison of genomic and traditional BLUP-estimated breeding value reliability and selection response under alternative trait and genomic parameters. *J. Anim. Breed. Genet. 124,* 342–355.

Sánchez, J.P., Misztal, I. & Bertrand, J.K. 2008. Evaluation of methods for computing approximate reliabilities in maternal random regression models for growth trait in beef. *J. Anim. Sci. 86,* 1057–1066.

Sargolzaei, M. & Schenkel, F.S. 2009. QMSim: A large-scale genome simulator for livestock. *Bioinformatics 25,* 680–681.

Simeone, R., Misztal, I., Aguilar, I. & Vitezica, Z. 2010. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. *J. Anim. Breed. Genet.* DOI: 10.1111/j.1439-0388.2011.00939.x.

Strabel, T., Misztal, I. & Bertrand, J.K. 2001. Approximation of reliabilities for multiple-trait models with maternal effects. *J. Anim. Sci. 79,* 833–839.

Tsuruta, S., Aguilar, I., Misztal, I. & Lawlor, T.J. 2011. Multiple-trait genomic evaluation of linear type traits using genomic and phenotypic data in US Holsteins. *J. Dairy Sci. 94*, 4198-4204.

VanRaden, P.M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci. 91,* 4414–4423.

VanRaden, P.M., Van Tassell, C.P., Wiggans, G.R., Sonstegard, T.S., Schnabel, R.D., Taylor, J.F. & Schenkel, F.S. 2009. Invited review: Reliability of genomic predictions for North American Holstein bulls. *J. Dairy Sci. 92*, 16–24.

VanRaden, P.M. & Wiggans, G.R. 1991. Derivation, calculation, and use of national animal model information. *J. Dairy Sci. 74,* 2737–2746.

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genet. Res. Camb. 93*, 357–366.

Wang, H. & Misztal, I. 2010. Comparisons of numerator and genomic and relationship matrices. *J. Anim. Sci. 89 (Suppl. 1),* 163.

Wiggans, G.R., Misztal, I. & Van Vleck, L.D. 1988. Animal model evaluation of Ayrshire milk yield with all lactations, herd-sire interaction, and groups based on unknown parents. *J. Dairy Sci. 71,* 1319–1329.