Experiences with a single-step genome evaluation[1]

Ignacy Misztal,*[2] Samuel E. Aggrey,† and William M. Muir‡

*Department of Animal and Dairy Science, and †Department of Poultry Science, University of Georgia, Athens 30602; and ‡Department of Animal Sciences, Purdue University, West Lafayette, IN 47907

ABSTRACT Genomic selection can be implemented based on the genomic relationship matrix (GBLUP) and can be combined with phenotypes from nongenotyped animals through the use of best linear unbiased prediction (BLUP). A common method to combine both sources of information involves multiple steps, but is difficult to use with complicated models and is nonoptimal. A simpler method, termed single-step GBLUP, or ssGBLUP, integrates the genomically derived relationships (G) with population-based pedigree relationships (A) into a combined relationship matrix (H) and allows for genomic selection in a single step. The ssGBLUP method is easy to implement and uses standard BLUP-based programs. Experiences with field data in chickens, pigs, and dairy indicate that ssGBLUP is more accurate yet much simpler than multi-step methods. The current limits of ssGBLUP are approximately 100,000 genotypes and 18 traits. Models involving 10 million animals have been run successfully. The inverse of H can also be used in existing programs for parameter estimationm, but a properly scaled G is needed for unbiased estimation. Also, as genomic predictions can be converted to SNP effects, ssGBLUP is useful for genomic-wide association studies. The single-step method for genomic selection translates the use of genomic information into standard BLUP, and variance-component estimation programs become a routine.

Key words: single-step genomic relationship matrix, single nucleotide polymorphism, best linear unbiased prediction, genome-wide association study

2013 Poultry Science 92:2530–2534
http://dx.doi.org/10.3382/ps.2012-02739

INTRODUCTION

Currently, the most common scenario for obtaining genomic predictions via SNP arrays and information from phenotypes of nongenotyped relatives involves a multistep approach including 1) running a regular BLUP evaluation based on pedigrees, 2) extracting pseudo-observations for genotyped individuals (e.g., daughter deviations or de-regressed EBV), 3) estimating SNP effects using pseudo-observations as records, and 4) possibly combining the genomic predictions with parent averages (VanRaden, 2008). One can start with step 3 using just phenotypic records associated with genotyped individuals; however, accuracy is reduced because phenotypic information from relatives cannot be used. Step 3 usually requires estimating weights for SNP effects, mostly via Bayesian procedures such as BayesX (where X could be A, B, C, and so on; Meuwissen et al., 2001; Hayes et al., 2009b). A procedure in which all weights are assumed equal (the infinitesimal model), leads to a genomic relationship matrix **G** and is called GBLUP (Habier et al., 2007). However, methods based on **G** vs. BLUP estimation of SNP effects (Meuwissen et al., 2001) are equivalent (VanRaden, 2008).

Current experiences for quantitative traits with SNP panels of around 50 to 60K indicate that (a) at least 1,000 genotypes of high accuracy animals are required for a noticeable increase in accuracy and (b) GBLUP is generally as accurate as BayesX procedures (VanRaden et al., 2009; Hayes et al., 2009c). These results indicate that the number of QTL are high (Daetwyler et al., 2010) and the infinitesimal model is approximately correct for most traits. Thus, the increased accuracy of genomic selection mainly results from better estimation of relationships with G rather than by estimating effects of major genes as with BayesX. Further, use of larger SNP panels (>500K) has provided only small improvements in accuracy, indicating that the effect of increasing SNP can be viewed as reducing the sampling error of **G**. Thus the primary mode by which genomic information improves genetic evaluation is through better estimation of relationships among animals, including Mendelian sampling (Goddard et al., 2010).

Received August 31, 2012.
Accepted November 22, 2012.
[1]Presented as part of the Experimental Design for Poultry Production and Genomics Symposium at the Poultry Science Association's annual meeting in Athens, Georgia, July 12, 2012.
[2]Corresponding author: ignacy@uga.edu

Using G for genomic selection through multi-step methodology is complicated and includes several approximations. Pseudo-observations are dependent on other estimated effects and approximated accuracy of EBV. All the approximations reduce accuracy and can inflate GEBV. Also, because of its complexity, the multistep approach is prone to errors, which have been observed in many commercial releases in dairy cows.

Because almost all the genomic information is included in a genomic relationship matrix, Misztal et al. (2009) proposed a single-step methodology where pedigree and genomic relationships are combined into matrix **H**, which is subsequently used in BLUP. Then, compared with a multistep evaluation, step 1 is modified to use matrix **H**, and steps 2 to 4 are eliminated. Legarra et al. (2009) and Christensen and Lund (2010) developed such a matrix, and Aguilar et al. (2010) demonstrated that a single-step methodology can be simple, fast, and accurate. The purpose of this paper is to present the theory of, and experiences with, the single-step methodology (**ssGBLUP**).

ssGBLUP

Legarra et al. (2009) developed matrix **H** that combines pedigree and genomic relationships:

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1} & 0 \\ 0 & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ \mathbf{I} \end{bmatrix} [\mathbf{G} - \mathbf{A}_{22}] [\mathbf{I} \quad \mathbf{I}] \begin{bmatrix} \mathbf{A}_{22}^{-1}\mathbf{A}_{21} & 0 \\ 0 & \mathbf{I} \end{bmatrix},$$

where **I** is an identity matrix, subscripts 1 and 2 denote ungenotyped and genotyped animals, respectively, and **G** is a genomic relationship matrix as in VanRaden (2008). Aguilar et al. (2010) and Christensen and Lund (2010) found that the inverse of matrix **H** as above has a simple form:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{G^{-1}} - \mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Replacing $\mathbf{A}^{-1}$ with $\mathbf{H}^{-1}$ in existing software for genetic evaluation or for estimation of variance components makes those programs applicable for genomic studies. Efficient computation of $\mathbf{H}^{-1}$ requires efficient computation of $\mathbf{G}^{-1}$ and ${\mathbf{A}_{22}}^{-1}$, where the last matrix is an inverse of a pedigree-based relationship matrix for genotyped animals only. Aguilar et al. (2011a) demonstrated that with efficient computing, such matrices can be created in a few minutes of computer time for about 10K genotypes and in about 1 h for 30K genotypes.

Matrix **G** is calculated by matrix multiplication (VanRaden, 2008). Let p_j be allele frequency for genotype 2 in marker j, and let m_{ij} be genotypes for ith animal and jth marker such that

$$m_{ij} = \begin{cases} 0 - 2p_j \text{ - homozygous 11} \\ 1 - 2p_j \text{ - heterozygous 12 to 21} \\ 2 - 2p_j \text{ - homozygous 22} \end{cases}$$

so that average $m_{i.}$ is 0. Then $\mathbf{G} = \mathbf{MM}'/k$, where the scale parameter k is usually computed as

$$k = 2\sum p_j(1 - p_j).$$

Thus, allele frequencies affect both the mean and scale of **G**. When equal allele frequencies are assumed, averages of the diagonal or off-diagonal elements may be much larger than in $\mathbf{A}_{22}$. Scaling **G** by regression on **A** as in VanRaden (2008) may result in **G** not being positive definite. Matrix **G** can be made compatible with $\mathbf{A}_{22}$ by using observed allele frequencies scaled such that the same averages for the diagonal and off-diagonal elements of $\mathbf{A}_{22}$ result (Chen et al., 2011a). When genotyped animals include clones, **G** as constructed above is singular and cannot be inverted. Therefore, a common strategy is to replace **G** with $\alpha\mathbf{G} + (1 - \alpha)\mathbf{A}_{22}$, where α is close to 1.0 (e.g., 0.95). In Christensen et al. (2012), the optimal α was 0.8, although variations in accuracy with different α were minimal.

Applications of ssGBLUP

The ssGBLUP has been used for several large-scale analyses including dairy (Tsuruta et al., 2011; Aguilar et al., 2011b; VanRaden, 2012), pigs (Forni et al., 2011; Christensen et al., 2012), and chickens (Chen et al., 2011b). Experiences indicate that 1) ssGBLUP is generally as accurate as or more accurate than multistep methods, 2) the inflation of GEBV is usually smaller, and 3) the operation is much simpler. Another reason for using ssGBLUP is the ability to account for selection bias when selection is based on genotypes only (VanRaden, 2012). Some experiences are species related. For example, in dairy there are small differences in GEBV with different **G** in dairy, where the "training" animals are high-accuracy bulls. In swine and chickens, different **G** affects estimates of variances components, calculated accuracy and, to a smaller degree, real accuracy (Forni et al., 2011; Simeone et al., 2012). Chen et al. (2011a) found that the scale of **G** influences ranking of genotyped versus ungenotyped animals. The optimal **G** would have the same averages of diagonals and off-diagonals as $\mathbf{A}_{22}$. Vitezica et al. (2011) derived a formal proof and showed that, under well-formed **G**, ssGBLUP is more accurate and less biased than a multistep approach. The optimal scaling is equivalent to

$$\mathbf{G}c = \text{Fst}/2 + (1 - \text{Fst})\mathbf{G}c,$$

where Fst is a fixation index of genotyped animals relative to the base population and **G**c is a genomic rela-

tionship matrix computed using realized allele frequencies. In practice, Fst is equal to twice the average of off-diagonals of $\mathbf{A}_{22}$.

ssGBLUP in Chicken

The ssGBLUP was applied in chickens for a large experiment (Chen et al., 2011b). However, results of selection over a few generations were short of expectations, although accuracies were generally in line with expectations (Muir et al., 2012). These were caused by many procedural issues and bias in the initial generation due to incorrect scaling (Muir et al., 2012). Therefore, the practical results of ssGBLUP or any other methodology for genomic selection in a commercial situation is dependent on attention to detail, using the mature methodology, and knowledge of issues of genomic selection specific to a given population.

Convergence and Biases

Poor convergence rate, large reranking, or both have been observed in several analyses. These problems were traced to incompatability between $\mathbf{G}$ and $\mathbf{A}_{22}$. For well-formed G and with many generations of complete pedigrees, $\mathbf{G}$ and $\mathbf{A}_{22}$ are very similar, with SD <0.04 (Wang and Misztal, 2011). However, differences exceeding 1.0 are observed in practice. Such differences are due to short or incomplete pedigrees, pedigree mistakes, incorrect assignment of genotypes, poor quality of genotypes, and the unaccounted presence of multiple/lines breeds. In general, $\mathbf{A}_{22}$ is affected by the number of generations and completeness of the pedigree. Matrix $\mathbf{G}$ is affected by allele frequencies, number of SNP, quality of genotypes, and scaling. These factors, when unaccounted for, can cause large differences between $\mathbf{G}$ and $\mathbf{A}_{22}$, and subsequently poor convergence rate with iterative methods.

One particular source of differences between $\mathbf{G}$ and $\mathbf{A}_{22}$ and subsequently poor convergence rate is heterogeneous base populations or missing parents at many generations. One remedy is to reduce the number of pedigrees to 4 to 5 so that the effect of missing parents in generations before the cutoff is eliminated. Experience indicates that accuracy will remain the same or even increase slightly after cutting. Another important source of difference is the presence of multiple lines or breeds. In such a case, the distribution of the diagonal of $\mathbf{G}$ may be multi-modal (Simeone et al., 2011). The solution is to construct $\mathbf{G}$ reflecting differences among line/breeds, such as in Harris and Johnson (2010). The issue of multibreed ssGBLUP is currently a hot research topic.

Fine Tuning

The theory for constructing $\mathbf{H}$ also makes many assumptions that may not hold in practice. Those assumptions include that the same genetic parameters in the genotyped sample as in the complete population, and existence of complete data on all traits for which selection occurred to account for selection bias. Several studies found that better accuracies and lower biases of GEBV can be achieved by fine-tuning α, β, τ, and ω in $\mathbf{H}$ defined as below:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} 0 & 0 \\ 0 & \tau(\alpha\mathbf{G} + \beta\mathbf{A}_{22})^{-1} - \omega\mathbf{A}_{22}^{-1} \end{bmatrix}.$$

Generally, $\alpha \approx 0.8$ and $\beta = 1 - \alpha$. Also, $\tau < 1$ and $\omega < 1$. Note that large ω causes H to be nonpositive definite, causing slower convergence or even divergence when mixed model equations are solved by iteration, whereas smaller ω will generally cause better convergence. Another mechanism to improve the convergence rate is to remove old pedigrees or even remove old data altogether, especially when the base population is heterogeneous (i.e., parents are missing across generations).

Large Number of Genotypes

When $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}{}^{-1}$ are computed explicitly, the cost is cubic with the number of genotypes. Subsequently, costs with >50 to 100K genotypes are excessive. Research in several labs (Harris and Johnson, VanRaden, Legarra and Ducrocq) focuses on the implementation of ssGBLUP where the inverse is not needed. These methods rely on the fact that the product of $\mathbf{Gq}$ can be obtained with linear time with respect to the number of genotypes as

$$\mathbf{Gq} = \mathbf{MDM'q} = \{\mathbf{M}[\mathbf{D}(\mathbf{M'q})]\}$$

when the products are computed sequentially as products of a matrix and a vector.

Approximate Accuracies

The additional accuracy of GEBV due to the genomic information can be expressed in terms of record equivalents (**RE**). A contribution due to the genomic information from animal j to animal i in terms of RE is approximately (Misztal et al., 2013):

$$\mathrm{RE}_{ij} \sim (g_{ij} - a_{22,ij})^2 \mathrm{acc}^2,$$

where acc is the accuracy. As the mean difference between the genomic and pedigree relationship is very small, with SD <0.04 in poultry as found by Wang and Misztal (2011), individual contributions are very small. However, large differences between the relationships due to mistakes in pedigree or genotypes greatly inflate RE. Also, the contribution is small from animals with small individual accuracy. This means that a genomic prediction benefits much more from genotyping

animals with high accuracy (males with many progeny) than from animals with their own records only. The formula above has mainly an educational value as contributions from many animals overlap, creating double counting. For formulas that seem to work well but are more cryptic, see Misztal et al. (2013).

Genome-Wide Associations with ssGBLUP

Much of genomic work especially in academia aims at finding genome-wide associations (**GWAS**) and therefore solutions of SNP effects are desirable. The ssGBLUP can be adapted for GWAS (Wang et al., 2012). This involves the following steps: 1) convert GEBV to SNP effects; 2) estimate individual SNP variances; 3) incorporate variances in G; and 4) possibly re-compute GEBV and iterate. Experiences indicate that such methodology is similar in accuracy to BayesB, however, at much lower cost. It also allows GWAS with effects that are hard to account for with multistep methods, such as maternal or random regressions, and complicated models including multiple-trait analysis. The extra benefit of GWAS is availability of SNP weights. Such weights can be incorporated into **G**, possibly increasing the accuracy for traits with major genes. However, in multiple-trait analysis the genomic relationship matrix needs to be identical for all the traits. In such analysis, major markers can possibly be modeled as separate effects, different for each trait.

CONCLUSIONS

The single-step methodology provides for easy incorporation of genomic data into a genetic evaluation. The methodology is ready for single-breed evaluations with up to 50K genotypes; extensions to multiple breeds and a larger number of genotypes are under way. Also, tools for approximating accuracies and for exploring genome-wide associations are available. The realized accuracy of a genomic evaluation is dependent on many factors, including the quality of genomic data and the structure of the population. In commercial settings, the use of any genomic methodology needs to be accompanied by ongoing validation and subsequent troubleshooting if the results fall short of expectations. See frequently asked questions (FAQ) on genomic selection (Misztal, 2011). Such work needs the expertise of well-qualified scientists with comprehensive training in animal breeding and genetics.

ACKNOWLEDGMENTS

This study was partially funded by the Holstein Association, Smithfield Premium Genetics, Pig Improvement Company, and Agriculture and Food Research Initiative (Washington, DC) grants 2009-65205-05665 and 2010-65205-20366 from the USDA National Institute of Food and Agriculture Animal Genome Program (Washington, DC).

REFERENCES

Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta, and T. J. Lawlor. 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., I. Misztal, A. Legarra, and S. Tsuruta. 2011a. Efficient computation of genomic relationship matrix and other matrices used in single-step evaluation. J. Anim. Breed. Genet. 128:422–428.

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

Chen, C. Y., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2011a. Effect of different genomic relationship matrices on accuracy and scale. J. Anim. Sci. 89:2673–2679.

Chen, C. Y., I. Misztal, I. Aguilar, S. Tsuruta, T. H. E. Meuwissen, S. E. Aggrey, T. Wing, and W. M. Muir. 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.

Christensen, O., and M. Lund. 2010. Genomic prediction when some animals are not genotyped. Genet. Sel. Evol. 42:2.

Christensen, O. F., P. Madsen, B. Nielsen, T. Ostersen, and G. Su. 2012. Single-step methods for genomic evaluation in pigs. Animal 6:1565–1571.

Daetwyler, H. D., R. Pong-Wong, B. Villanueva, and J. A. Woolliams. 2010. The impact of genetic architecture on genome-wide evaluation methods. Genetics 185:1021–1031.

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

Goddard, M. E., T. H. E. Meuwissen, and B. J. Hayes. 2010. Genomic selection in farm animal species—Lessons learnt and future perspectives. Proc. 9th World Congr. Genet. Appl. Livest. Prod. Paper 0701.

Habier, D., R. L. Fernando, and J. C. M. Dekkers. 2007. The impact of genetic relationship information on genome-assisted breeding values. Genetics 177:2389–2397.

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

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 2009a. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92:433–443.

Hayes, B. J., H. D. Daetwyler, P. J. Bowman, G. Moser, B. Tier, R. Crump, M. Khatkar, H. W. Raadsma, and M. E. Goddard. 2009b. Accuracy of genomic selection: Comparing theory and results. Proc. Assoc. Advmt. Anim. Breed. 18:34–37.

Hayes, B. J., P. M. Visscher, and M. E. Goddard. 2009c. Increased accuracy of artificial selection by using the realized relationship matrix. Genet. Res. 91:47–60.

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

Meuwissen, T. H., B. J. Hayes, and M. E. Goddard. 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829.

Misztal, I. 2011. FAQ for genomic selection—Editorial. J. Anim. Breed. Genet. 128:245–246.

Misztal, I., A. Legarra, and I. Aguilar. 2009. Computing procedures for genetic evaluation including phenotypic, full pedigree, and genomic information. J. Dairy Sci. 92:4648–4655.

Misztal, I., S. Tsuruta, I. Aguilar, A. Legarra, P. M. VanRaden, and T. J. Lawlor. 2013. Methods to approximate reliabilities in single-step genomic evaluation. J. Dairy Sci. 96:647–654. http://dx.doi.org/10.3168/jds.2012-5656.

Muir, W. M., Ragavendran, A., Tosa, G. J. M., T. H. E. Meuwissen, I. Misztal, M. Groenen, T. Wing, R. Okimoto, A. Vereijken, and H. Cheng. 2012. Genomic selection in poultry, results with broilers and comparison with traditional BLUP. Proc. Plant Anim. Genome Meet. W591.

Simeone, R., I. Misztal, I. Aguilar, and A. Legarra. 2011. Evaluation of the utility of genomic relationship matrix as a diagnostic

tool to detect mislabeled genotyped animals in a broiler chicken population. J. Anim. Breed. Genet. 128:386–393.

Simeone, R., I. Misztal, I. Aguilar, and Z. Vitezica. 2012. Evaluation of a multi-line broiler chicken population using a single-step genomic evaluation procedure. J. Anim. Breed. Genet. 129:3–10.

Tsuruta, S., I. Aguilar, I. Misztal, and T. J. Lawlor. 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. 2012. Avoiding bias from genomic pre-selection in converting daughter information across countries. Interbull Bull. 45:1–5.

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

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

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

Wang, H., and I. Misztal. 2011. Comparisons of numerator and genomic and relationship matrices. J. Anim. Sci. 89(E-Suppl. 1):163.

Wang, H., I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir. 2012. Genome-wide association mapping including phenotypes from relatives without genotypes. Genet. Res. (Camb.) 94:73–83.