A Unified Approach to Utilize Phenotypic, Full Pedigree, and Genomic Information for a Genetic Evaluation of Holstein Final Score

I. Misztal[1], *I. Aguilar*[1,2], *D. Johnson*[3], *A. Legarra*[4], *S. Tsuruta*[1] *and T. J. Lawlor*[5]
[*]*Department of Animal and Dairy Science, University of Georgia, Athens 30602, USA*
[†]*Instituto Nacional de Investigación Agropecuaria, Las Brujas, 90200, Uruguay*
[‡]*LIC, Private Bag 3016, Hamilton 3240, New Zealand*
[§]*INRA, UR631 SAGA, BP 52627, 32326 Castanet-Tolosan, France*
[#]*Holstein Association USA Inc., Brattleboro, VT 05302-0808, USA*

Abstract

A relationship matrix in a genetic evaluation system was augmented for incorporation of genomic information to create a single-step procedure. The procedure was applied to a national evaluation for final score in U.S. Holsteins. Computing was 2% longer than the traditional evaluation. Accuracies and biases of prediction of young bulls were affected by scaling of the genomic relationship matrix. With "optimal" scaling, reliabilities were higher and the inflation of EBV was lower compared to a multi-step approach. Accurate genomic evaluations can be obtained by modifying the relationship matrices in current evaluation systems. Advantages of the single-step procedure include a dramatic simplification of computations, ability to use more complicated models (including multi-trait), increased resistance to selection bias introduced by use of genomic evaluations, and improved accuracy for ungenotyped animals.

Introduction

Genomic evaluation is currently a multi-step procedure (VanRaden, 2008; Hayes *et al.*, 2009). It includes the traditional genetic evaluation, creation of pseudo-observations [daughter deviations (DDs) or deregressed evaluations], genomic selection, and possibly combining of genomic and polygenic data in a selection index. Those steps have numerous options and parameters

Current experiences with genomic evaluations from the multi-step procedure seem mixed. Although genomic evaluations are more accurate than parent averages (PAs) and approach the accuracy of evaluations for progeny-tested bulls, they also seem inflated (VanRaden *et al.*, 2009a). Inflation of genetic evaluations by genomic information causes top young bulls to have an unfair advantage over older progeny-tested bulls. Some of the problems with genomic evaluations may be caused by incorrect parameters and strong assumptions used in multi-step procedures. Problems with biases in genomic selection are likely to increase with current methodology because traditional evaluations do not include the genomic information on which selection decisions are based (Patry and Ducrocq, 2009).

For genomic selection, which involves only genotyped animals, estimating single-nucleotide polymorphism (SNP) marker effects or using best linear unbiased prediction with a genomic relationship matrix are equivalent except for some numerical issues (VanRaden, 2008; Goddard, 2009). Misztal *et al.* (2009) proposed a single-step procedure in which the pedigree-based relationship matrix is augmented by contributions from the genomic relationship matrix, which allows all animals (genotyped or not) to be included in analyses. Assuming that the combined matrix cannot be inverted, they also suggested a computing procedure based on a nonsymmetric system of mixed model equations that was suitable for millions of animals; the matrix could be semi-positive definite. Legarra *et al.* (2009) derived a joint relationship matrix based on pedigree and genomic relationships. Even though the matrix seemed complicated, computations were feasible even for large data sets. Johnson (personal communication, 2009) derived an inverse of the last matrix, which allows for simplified computations.

The single-step procedure provides one unified framework, eliminates a number of assumptions and parameters, and provides an opportunity for a more accurate genomic evaluation than with the multi-step procedure. The purpose of this study was to present implementation of a single-step procedure in a national evaluation setting and compare its performance to a multi-step procedure.

Data

Data were U.S. Holstein information for final score as used for May 2009 official evaluations (Holstein Association USA, 2009). A total of 10,466,066 records were available for 6,232,548 cows. Pedigrees included 9,100,106 animals. Genotypes for 6,508 bulls were generated using the Illumina BovineSNP50 BeadChip and DNA from semen contributed by U.S. and Canadian artificial-insemination organizations to the Cooperative Dairy DNA Repository; genotypes were provided by the Animal Improvement Programs Laboratory, Agricultural Research Service, USDA (Beltsville, MD).

Relationship Matrix Including Both Pedigree and Genomic Information

"Raw" genomic relationships ($\mathbf{G}_b$) were created using either equal allele frequencies (G05) or estimated allele frequencies for the base population (GB; Gengler *et al.*, 2007). To facilitate inversion, final analyses used the following $\mathbf{G}$:

$$\mathbf{G} = 0.95\mathbf{G}_b + 0.05\mathbf{A}_{22},$$

where $\mathbf{A}_{22}$ is a numerator relationship matrix for genotyped animals. Instead of $\mathbf{A}^{-1}$, genomic analyses used $\mathbf{H}^{-1}$:

$$\mathbf{H}^{-1} = \mathbf{A}^{-1} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \lambda\left(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1}\right) \end{bmatrix},$$

where λ is a scaling factor. For a young bull's solution (u_i):

$$u_i = \frac{u_{sire} + u_{dam} + \lambda \sum_{j,\, j \neq i} (a_{22}^{ij} - g^{ij}) u_j}{2 + \lambda(g^{ii} - a_{22}^{ii})},$$

λ determines the weight for genomic contributions. At $\lambda = 0$, $\mathbf{H}^{-1}$ becomes $\mathbf{A}^{-1}$.

Models and Analyses

A repeatability animal model was used for analysis as is currently done for U.S. national evaluations of Holstein conformation traits (Holstein Association USA, 2009). The first two analyses used final scores through 2004 only. The first analysis (P2004) used only the pedigree-based relationship matrix; the second analysis (PG2004) used relationships based on both pedigree and genomic information. The third analysis (P2009) used the complete data set and the pedigree-based relationship matrix. The fourth analysis (MG2004) used predictions from P2004 to obtain genomic predictions using a multi-step approach as described by VanRaden *et al.* (2009b). The MG2004 approach assumed equal variances per SNP marker effect.

Comparisons were based on regressions:

$$\mathbf{x2009} = \mu + \delta\mathbf{x2004} + \mathbf{e},$$

where **x2009** are DDs or estimated breeding values (EBVs) for genotyped bulls without daughters in 2004 and with ≥40 daughters in 2009, **x2004** are predictions from various 2004 evaluations, μ is a mean, δ is a regression coefficient, and **e** is the residual error. The most accurate method for prediction of young bulls would have μ close to 0, δ close to 1, and R^2 as high as possible.

Software

Initial software for the construction of **G** and the multi-step evaluation was provided by Paul VanRaden. Additional software for the construction of **G** was contributed by Ben Hayes. The software was modified for efficient matrix multiplication, matrix inversion, and parallelization. Computation of pedigree-based relationship matrix ($\mathbf{A}_{22}$) was based on the formulas of Misztal *et al.* (2009) using Colleau's algorithm (2002). Genetic evaluation was by modified BLUP90IOD (Tsuruta *et al.*, 2001; Misztal *et al.*, 2002).

Results and Discussion

Table 1 shows R^2 and δ for 2009 DDs and EBV on various 2004 predictions for young bulls with a λ assumed to be 1. For DDs, prediction with parent average resulted in R^2 of 24% and δ of 0.79. The δ showed that a high parent average overestimated genetic evaluation with progeny included by 27%. With the multi-step approach, R^2 increased to 40% and δ to 0.88 (inflation decreased to 14%).

Table 1. R^2 and regression coefficients (δ) for 2009 DDs and EBVs on 2004 young bull predictions.

Evaluation	DD		EBV	
method	R^2	δ	R^2	δ
Parent average	24	0.79	36	0.82
Multi-step	40	0.88	50	0.83
Single-step				
G05	41	0.77	49	0.71
GB	38	0.69	45	0.64

With the single-step approach, R^2 was 38 to 41% and δ was 0.69 to 0.77 depending on **G**. Compared with the multi-step approach, the G05 R^2 was 1 percentage unit higher, which indicated a slightly higher accuracy for breeding values. The G05 δ was 0.08 lower, which indicated more inflation of the early prediction. Because G05 had the highest R^2 and the lowest inflation, subsequent single-step comparisons were based on G05.

Results based on 2009 EBVs were generally similar to those for 2009 DDs except for a slight advantage for the multi-step approach. The δ indicated greater inflation of predictions than with DYDs. Inflation on the 2009 EBV scale is meaningful for producers because their comparisons are based on breeding values and not on DDs.

Because parent average was similar for evaluations with and without **G**, the inflation resulted from indirectly placing too much weight on genomic relationships.

Table 2 shows R^2 and δ as for various λ. As λ varied from 1.0 to 0.5, R^2 stayed the same or declined for DDs but had an interim maximum for EBVs. At λ of 0.7, R^2 increased to 51% for EBVs, which is 1 percentage unit better than for the multi-step method. Also, δ was 0.01 higher than for the multi-step method, which showed that δ could be increased to 0.94 with only a slight decrease in R^2. Thus, bias can be controlled through λ.

Table 2. R^2 and regression coefficients (δ) for 2009 DDs and EBVs on 2004 young bull predictions based on single-step G05 evaluation with various λ.

	DD		EBV	
λ	R^2	δ	R^2	δ
1.0	41	0.77	49	0.71
0.9	41	0.83	50	0.77
0.8	41	0.86	51	0.80
0.7	40	0.89	51	0.84
0.6	40	0.91	50	0.87
0.5	39	0.94	50	0.90
0.3	35	0.93	47	0.91

Why a weighting factor is needed is not clear. For example, in another study with chickens (results not reported), the highest accuracy was obtained without a weighting factor (λ = 1.0). One possibility is that genetic parameters used in the evaluation are not optimal for prediction of young bulls. Other issues are preferential treatment of bull dams and the nature of final score, for which the definition changes over time (Tsuruta *et al.*, 2005; Koduru, 2006).

The accuracy of the single-step method depends on the choice of **G** and λ. With the proper choice, it is more accurate than the multi-step procedure. One reason why the choice of **G** is critical is that genomic and relationship matrices should be compatible both in scale and in structure.

For this study, **G** was constructed with equal variances assumed for SNP marker. When variances are not equal [e.g., as in BayesA or BayesB (Meuwissen *et al.*, 2001)], an equivalent **G** can be constructed by scaling contributions for different markers. Although the generalization of the single-step method to multi-trait is obvious when **G** is identical for each trait, separate **G** per trait may require single-trait analyses. For many traits, the

benefits and simplicity of multiple analyses using the same **G** may overcome the loss of accuracy due to using less than the optimal **G** for each trait.

Conclusions

Evaluations by the single-step procedure are simpler to implement, more accurate, and less biased than those from a multi-step procedure. Generalizations to complex models such as random regression and multi-trait are automatic. Additional benefits include resistance to selection bias due to genomic information used in selection and improved accuracy for ungenotyped animals.

Acknowledgments

The authors thank the Cooperative Dairy DNA Repository and the Animal Improvement Programs Laboratory, USDA, for providing the genotypic data and Holstein Association USA for financial support; contributions from and discussions with Curt Van Tassel, Paul VanRaden, George Wiggans, and Jeff O'Connell were greatly appreciated. Editorial help by Suzanne Hubbard is gratefully acknowledged.

References

Colleau, J.J. 2002. An indirect approach to the extensive calculation of relationship coefficients. *Genet. Sel. Evol. 34,* 409–421.

Gengler, N., Mayeres, P. & Szydlowski, M. 2007. A simple method to approximate gene content in large pedigree populations: Application to the myostatin gene in dual-purpose Belgian Blue cattle. *Animal 1,* 21–28.

Goddard, M. 2009. Genomic selection: Prediction of accuracy and maximisation of long term response. *Genetica 136,* 245–257.

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

Holstein Association USA. 2009. *Holstein Total Performance Index Sire Summaries*. Holstein Association USA, Inc., Brattleboro, VT.

Koduru, V.K.R. 2006. Changes of Holstein sire PTAs and trends from 1st to 2nd crop evaluations for final score. *M.S. Dissertation,* University of Georgia, Athens.

Legarra, A., Aguilar, I. & Misztal, I. 2009. A relationship matrix including full pedigree and genomic information. *J. Dairy Sci. 92,* 4656–466.

Meuwissen, T.H.E., Hayes, B.J. & Goddard, M.E. 2001. Prediction of total genetic value using genome-wide dense marker maps. *Genetics 15,* 1819–1829.

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

Misztal, I., Tsuruta, S., Strabel, T., Auvray, B., Druet, T. & Lee, D.H. 2002. Blupf90 and related programs (BGF90). *Proc. 7th World Congress on Genetics Applied to Livestock Production,* Montpellier, France, Commun. No. 28-07.

Patry, C. & Ducrocq, V. Bias due to genomic selection. *Interbull Bulletin 39,* 77–82.

Tsuruta, S., Misztal, I. & Stranden, I. 2001. Use of the preconditioned conjugate gradient algorithm as a generic solver for mixed-model equations in animal breeding applications. *J. Anim. Sci. 79,* 1166–1172.

Tsuruta, S., Misztal, I. & Lawlor, T.J. 2005. Changing definition of productive life in US Holsteins: Effect on genetic correlations. *J. Dairy Sci. 88,* 1156–1165.

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

VanRaden, P.M., Tooker, M.E. & Cole, J.B. 2009a. Can you believe those genomic evaluations for young bulls? *J. Dairy Sci. 92(E-Suppl. 1),* 314(abstr. 279).

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