Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score

I. Aguilar, I. Misztal, D.L. Johnson, Andres Legarra, S. Tsuruta, T.J. Lawlor

J. Dairy Sci. 93:743–752
doi:10.3168/jds.2009-2730

Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score[1]

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

ABSTRACT

The first national single-step, full-information (phenotype, pedigree, and marker genotype) genetic evaluation was developed for final score of US Holsteins. Data included final scores recorded from 1955 to 2009 for 6,232,548 Holsteins cows. BovineSNP50 (Illumina, San Diego, CA) genotypes from the Cooperative Dairy DNA Repository (Beltsville, MD) were available for 6,508 bulls. Three analyses used a repeatability animal model as currently used for the national US evaluation. The first 2 analyses used final scores recorded up to 2004. The first analysis used only a pedigree-based relationship matrix. The second analysis used a relationship matrix based on both pedigree and genomic information (single-step approach). The third analysis used the complete data set and only the pedigree-based relationship matrix. The fourth analysis used predictions from the first analysis (final scores up to 2004 and only a pedigree-based relationship matrix) and prediction using a genomic based matrix to obtain genetic evaluation (multiple-step approach). Different allele frequencies were tested in construction of the genomic relationship matrix. Coefficients of determination between predictions of young bulls from parent average, single-step, and multiple-step approaches and their 2009 daughter deviations were 0.24, 0.37 to 0.41, and 0.40, respectively. The highest coefficient of determination for a single-step approach was observed when using a genomic relationship matrix with assumed allele frequencies of 0.5. Coefficients for regression of 2009 daughter deviations on parent-average, single-step, and multiple-step predictions were 0.76, 0.68 to 0.79, and 0.86, respectively, which indicated some inflation of predictions. The single-step regression coefficient could be increased up to 0.92 by scaling differences between the genomic and pedigree-based relationship matrices with little loss in accuracy of prediction. One complete evaluation took about 2 h of computing time and 2.7 gigabytes of memory. Computing times for single-step analyses were slightly longer (2%) than for pedigree-based analysis. A national single-step genetic evaluation with the pedigree relationship matrix augmented with genomic information provided genomic predictions with accuracy and bias comparable to multiple-step procedures and could account for any population or data structure. Advantages of single-step evaluations should increase in the future when animals are preselected on genotypes.
Key words: best linear unbiased predictor, genomic prediction, single nucleotide polymorphism, genetic evaluation

INTRODUCTION

Genomic evaluations are currently calculated with a multiple-step procedure (VanRaden, 2008; Hayes et al., 2009). A typical evaluation requires 1) traditional evaluation with an animal model, 2) extraction of pseudo-observations such as deregressed evaluations or daughter deviations (**DD**), 3) estimation of genomic effects for genotyped animals usually using simple sire models, and possibly 4) combining the genomic index with traditional parent averages (**PA**) and EBV (Hayes et al., 2009; VanRaden et al., 2009b). Those steps are dependent on many parameters and assumptions. For example, estimation of genomic effects has several options (Meuwissen et al., 2001; Gianola et al., 2006; VanRaden, 2008; de los Campos et al., 2009). The SNP marker effects can be estimated with different assumptions regarding the prior distribution of such effects. Genomic effects can also be estimated with a simple model that includes a genomic relationship matrix derived from genotypes and variances of the SNP marker effects (Nejati-Javaremi et al., 1997). Both methods are

Received September 14, 2009.
Accepted November 10, 2009.
[1]Appendix A of this paper was developed by D. L. Johnson (Livestock Improvement Corp., Hamilton, New Zealand).
[2]Corresponding author: iaguilar@inia.org.uy

equivalent except for numerical properties (VanRaden, 2007).

Initially, genomic evaluation was tested with simulated data and a variety of assumptions (VanRaden, 2008). Experiences with actual data from dairy cattle (Hayes et al., 2009; VanRaden et al., 2009b) indicated that using a large number of markers with equal variance for all markers is appropriate for most traits. Limiting the number of SNP markers to only those with large effects resulted in reduced accuracy (Cole et al., 2009). However, little (if any) loss of accuracy occurred for most dairy cattle traits by assuming equal rather than different variance for each SNP marker (Cole et al., 2009; VanRaden et al., 2009b). Further, assuming equal variance allows the use of the same genomic relationship matrix for all traits.

Current experiences with genomic evaluations from the multiple-step procedure seem mixed. Genomic evaluations are more accurate than PA and approach the accuracy of evaluations for progeny-tested bulls, but they also seem inflated (VanRaden et al., 2009a). Although their inflation is lower than that of current PA, the potentially great utilization of top genomically evaluated young sires increases the importance of high accuracy and minimum bias. 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 multiple-step procedures. However, effects of those parameters and assumptions are extremely difficult to verify, particularly in the presence of selection. An alternative explanation for the mixed results is that observed regressions and estimated reliabilities are biased downward by selective genotyping. A more serious problem is when pseudo-observations are poorly defined or of poor quality (e.g., for animals with small progeny numbers), which is often the case for monogastric species and for beef cattle.

Misztal et al. (2009) proposed a single-step evaluation in which the pedigree-based relationship matrix is augmented by contributions from the genomic relationship matrix. They also suggested a computing procedure based on a nonsymmetric system of mixed model equations that was suitable for millions of animals. Legarra et al. (2009) derived a joint relationship matrix based on pedigree and genomic relationships. Even though the matrix was expensive and complex to create, computations were feasible even for large data sets.

The single-step procedure provides a unified framework, eliminates several assumptions and parameters, and provides the opportunity to calculate more accurate genomic evaluations than the multiple-step procedures. The objective of this study was to use a single-step procedure for genomic evaluation in a national evaluation setting and compare its performance to a multiple-step procedure.

MATERIALS AND METHODS

Data

Data were US Holstein information for final score 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 were available for 9,100,106 animals. Genotypes for 6,508 bulls were generated using the Illumina BovineSNP50 BeadChip (Illumina, San Diego, CA) and DNA from semen contributed by US and Canadian AI organizations to the Cooperative Dairy DNA Repository (Beltsville, MD); genotypes were provided by the Animal Improvement Programs Laboratory, Agricultural Research Service, USDA (Beltsville, MD).

Relationship Matrix with Pedigree and Genomic Information

Misztal et al. (2009) suggested that a numerator relationship matrix ($\mathbf{A}$) can be modified to a matrix ($\mathbf{H}$) that includes both pedigree-based relationships and differences between pedigree-based and genomic-based relationships ($\mathbf{A}_\Delta$): $\mathbf{H} = \mathbf{A} + \mathbf{A}_\Delta$. In their examples, they used

$$\mathbf{H} = \begin{bmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{G} \end{bmatrix} = \mathbf{A} + \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix},$$

where subscripts 1 and 2 represent ungenotyped and genotyped animals, respectively, and $\mathbf{G}$ is a genomic relationship matrix. In tests, such $\mathbf{H}$ did not work because off-diagonals of $\mathbf{H}$ were not functions of $\mathbf{G}$. Assume, for example, that no animal in $\mathbf{G}$ has records; then, according to $\mathbf{H}$, the predicted breeding value for genotyped animals ($\mathbf{u}_2$) would be $\mathbf{u}_2 \mid \mathbf{u}_1 = \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{u}_1$, where $\mathbf{u}_1$ is the predicted breeding value for ungenotyped animals, and $\mathbf{G}$ would have no role whatsoever.

Legarra et al. (2009) suggested deriving the joint density of $\mathbf{u}_1$ and $\mathbf{u}_2$ as $p(\mathbf{u}_1, \mathbf{u}_2) = p(\mathbf{u}_1 \mid \mathbf{u}_2)p(\mathbf{u}_2)$. The conditional distribution $p(\mathbf{u}_1 \mid \mathbf{u}_2)$ is based on pedigree through the selection index or multivariate normal properties; $p(\mathbf{u}_2)$ is based only on genomic information, possibly from genomic relationships. The covariance of the joint distribution of $\mathbf{u}_1$ and $\mathbf{u}_2$ is thus $\mathbf{H}$:

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

which could be implemented in tests by using computing algorithms such as in Misztal et al. (2009) with only a few more computations per round of iteration than for traditional evaluations. Convergence was readily obtained for medium-sized data sets (up to 1 million); however, for larger data sets, convergence was strongly dependent on the type of **G** used.

An inverse of **H** that allows for drastically simpler computations (see Appendix A) is

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

where $\mathbf{A}_{22}^{-1}$ is the inverse of a pedigree-based relationship matrix for genotyped animals only. This expression has also been independently derived by Christensen and Lund (2009). However, the new formula introduces a small problem: **G** is usually singular and, therefore, cannot be inverted without additional steps.

Models and Analyses

A repeatability animal model was used for analysis as is currently done for US national evaluation of Holstein conformation traits (Holstein Association USA, 2009). The first 2 analyses used final scores through 2004 only. The first analysis (**Ped$_{04}$**) used only the pedigree-based relationship matrix; the second analysis (**PedGen1$_{04}$**) used relationships based on both pedigree and genomic information in a single-step approach. The third analysis (**Ped$_{09}$**) used the complete data set and only the pedigree-based relationship matrix. The fourth analysis (**PedGenM$_{04}$**) used predictions from Ped$_{04}$ and a multiple-step approach to obtain genomic predictions (**GP**) as described by VanRaden et al. (2009b). Options in the last analysis were genomic relationship matrix and base allele frequencies. Both PedGen1$_{04}$ and PedGenM$_{04}$ assumed equal variances per SNP marker effect.

"Raw" genomic relationships ($\mathbf{G}_b$) were created as

$$\mathbf{G}_b = \frac{\mathbf{ZZ}'}{k},$$

where **Z** is an incidence matrix for SNP effects with elements

$$z_{ij} = \begin{cases} 0 - 2p_j & \text{if homozygous 11} \\ 1 - 2p_j & \text{if heterozygous 12 or 21,} \\ 2 - 2p_j & \text{if homozygous 22} \end{cases}$$

for animal i and SNP j with allele frequency p_j. Several allele frequencies were used to center the matrix: 0.5, base population estimated by linear regression of gene content (Gengler et al., 2007), and current population. The scaling parameter k was defined as

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

(VanRaden, 2008), which assumes a priori independence of SNP effects (Gianola et al., 2009).

Another scaling parameter has been proposed by Gianola et al. (2009) with

$$k = \left[(p_0 - q_0)^2 + 2\left(\frac{\sum p_j(1 - p_j)}{n} \right) \left(\frac{\alpha + \beta + 2}{\alpha + \beta} \right) \right] n,$$

where $p_0 = \alpha/(\alpha + \beta)$ is the expected allele frequency, $q_0 = (1 - p_0)$; α and β are parameters of the beta distribution fitting the base allelic frequency, and n is the number of SNP. That modification accounts for random ascertainment of SNP and their frequencies.

Matrices $\mathbf{G}_b$ were sometimes singular or close to singularity. To facilitate inversion, final analyses used a weighted **G** as proposed by VanRaden (2008): $\mathbf{G} = 0.95\mathbf{G}_b + 0.05\mathbf{A}_{22}$. The weights were not critical, and replacing them with 0.98 and 0.02 caused negligible differences.

Because GP could be scaled incorrectly, a series of analyses used $\mathbf{H}^{-1}$:

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

where λ scales differences between genomic and pedigree-based information. More precisely (see Appendix B), λ sets the value of **G** in **H** to a new value ($\mathbf{G}^*$):

$$\mathbf{G}^* = \left[\lambda\mathbf{G}^{-1} + (1 - \lambda)\mathbf{A}_{22}^{-1} \right]^{-1},$$

thus blending genomic and pedigree information. For $\lambda = 1$, $\mathbf{G}^* = \mathbf{G}$; for $\lambda = 0$, $\mathbf{G}^* = \mathbf{A}_{22}$ and $\mathbf{H} = \mathbf{A}$. In fact, this corresponds to the following prior for genotyped animals:

$$p\left(u_2\right) = p\left(u_2 \mid \mathbf{G}, \lambda\right) p\left(u_2 \mid \mathbf{A}_{22}, \lambda\right)$$
$$= N\left(0, \mathbf{G} / \lambda\right) N\left(0, \mathbf{A}_{22} / \left(1-\lambda\right)\right).$$

Comparisons were based on the regressions

$$\mathrm{DD} = \mu + \delta \mathrm{EBV}_{04} + \mathrm{e}$$

and

$$\mathrm{EBV}_{09} = \mu + \delta \mathrm{EBV}_{04} + \mathrm{e},$$

where DD were deregressed evaluations (VanRaden et al., 2009b) from genotyped bulls without daughter records in 2004 but with daughter records in 2009 that were computed with complete final score data but without genomic information; EBV_{09} are breeding values based on final scores up to 2009 but without genomic information; μ is a mean; δ is a regression coefficient; EBV_{04} are breeding values based on final scores up to 2004; and e is residual error. Breeding values were calculated for 2 sets of genotyped bulls: 1) 2,575 young bulls with no daughter records in 2004 but with daughter records in 2009 and 2) 3,933 evaluated bulls with daughter records in 2004. The most accurate method for prediction for young bulls would have μ close to 0, δ close to 1, and R^2 as high as possible.

Both DD and EBV_{09} regressions were examined to allow more detailed comparison. Although DD computed through deregressed evaluations allow partial removal of the effect of PA, the removal is contingent on the accuracy of approximate reliabilities. Also, the goal of GP is not to predict DD but to predict future breeding values.

Software

Initial software for the construction of **G** and the multiple-step evaluation was provided by P. M. VanRaden (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD). Additional software for creating **G** was contributed by B. J. Hayes (Biosciences Research Division, Department of Primary Industries Victoria, Bundoora, Australia). Software refinement included rearrangements of code in Fortran 95 for efficient matrix multiplication, matrix inversion, and parallelization. Computation of $\mathbf{A}_{22}$ followed the formulas of Misztal et al. (2009), which used the algorithm of Colleau (2002). Genetic evaluation was performed by modified BLUP90IOD (Tsuruta et al., 2001; Misztal et al., 2002), which uses iteration on data with the preconditioned conjugate gradient algorithm.

Table 1. Coefficients of determination (R^2) and coefficients (δ) for regression of 2009 daughter deviations (DD) or corresponding estimated breeding values (EBV_{09}) for bulls progeny tested from 2005 through 2009 on 2004 predictions obtained by different algorithms

	DD		EBV_{09}	
Prediction method	R^2 (%)	δ	R^2 (%)	δ
Parent average	24	0.76	36	0.79
Multiple-step	40	0.86	50	0.82
Single-step[1]				
G5	41	0.76	49	0.70
GB	38	0.68	45	0.63
GC	37	0.71	45	0.66
GG – G5	41	0.79	50	0.73
GG – GB	38	0.77	46	0.71
GG – GC	39	0.79	46	0.73

[1]Assumed allele frequency of 0.5 (G5), base population (GB), current population (GC), or calculated as in [30] of Gianola et al. (2009) (GG).

RESULTS AND DISCUSSION

Precomputation of **G** and $\mathbf{A}_{22}$ took 650 s and 45 s, respectively, on an Opteron 64-bit processor with a clock speed of 3.02 GHz and a cache size of 1 MB, using one processor; their inversion took approximately 150 s. Time per 1 preconditioned conjugate gradient round for $\mathrm{PedGen1}_{04}$ was 13 s, which was 2% greater than 1 round for Ped_{04}. Convergence rates (not shown) for $\mathrm{PedGen1}_{04}$ and Ped_{04} were almost identical. A complete analysis with $\mathrm{PedGen1}_{04}$ took approximately 2 h. Memory requirement for precomputation of **G** was 2.7 GB.

Table 1 shows R^2 and δ for regression of 2009 DD and corresponding EBV_{09} on various 2004 predictions for young bulls. For PA, R^2 was 24% with δ of 0.76. The δ showed that PA overestimated the genetic evaluation with progeny included by 27%. For the multiple-step approach, R^2 increased to 40% and δ to 0.86. The increase in R^2 of 16% compared with PA R^2 was slightly higher than the increase of 13% reported by VanRaden et al. (2009b). VanRaden et al. (2009a) reported a regression coefficient of 0.74. Differences from the results of VanRaden et al. (2009a,b) were caused in part by slightly different data (theirs included Canadian evaluations but fewer genotypes and US records) and methodology details (e.g., different computation of approximate reliabilities).

For the single-step approaches (Table 1), R^2 for DD varied between 37 and 41%, and δ varied between 0.68 and 0.79 depending on **G**. The highest single-step increase in R^2 over prediction from PA was 1% higher than the multiple-step increase, which indicated that single-step breeding values were slightly more accurate than those by the multiple-step as implemented here.

Table 2. Coefficients of determination (R^2) and coefficients (δ) for regression of 2009 daughter deviations (DD) or corresponding breeding values (EBV_{09}) for bulls progeny tested from 2005 through 2009 on 2004 predictions from a single-step approach using an allele frequency of 0.5 and different relative variances for the genomic matrix (λ)

	DD		EBV_{09}	
λ	R^2 (%)	δ	R^2 (%)	δ
1.0	41	0.76	49	0.70
0.9	41	0.81	50	0.76
0.8	41	0.84	51	0.79
0.7	40	0.88	51	0.83
0.6	40	0.90	50	0.85
0.5	39	0.92	50	0.88
0.3	35	0.91	47	0.89

The best δ was 0.07 lower than the multiple-step δ, which indicated greater inflation of prediction for young bulls. The highest single-step R^2 and δ (least inflation) were for **G** based on equal allele frequencies with extra benefits from modifications by Gianola et al. (2009). For simplification, subsequent comparisons used the equal allele frequency **G** but without the modifications.

The R^2 values obtained using **G** matrix with equal allele frequency was greater compared with a **G** matrix created using base allele frequency. This was in the opposite direction to a similar study (VanRaden et al., 2008). In addition, the latter study reported correlations of 0.6 between genomic- and pedigree-based inbreeding coefficients, whereas a correlation of 0.2 using base allele frequencies was found in the current study. Further analyses need to be done to address such differences.

Results for EBV_{09} (Table 1) generally were similar to those for DD but with a slight advantage for the multiple-step approach. The δ indicated much greater inflation than for DD. Inflation on the EBV_{09} scale is important for producers because their comparisons are based on EBV and not on DD. It is debatable whether the results with EBV_{09} are valid in this case, because they contain information from PA. On the other hand, DD computed using approximated reliabilities may contain an extra noise.

Parent average was, in general, similar for runs with and without G. Thus, inflation higher than that in PA could be caused by too much indirect weight on genomic relationships. Inflation could be lowered by weighting $(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})$ by λ (see Appendix B). Table 2 shows R^2 and δ for DD and EBV_{09} with such a weighting. As λ decreased from 1.0 to 0.5, R^2 gradually decreased for DD but had an interim maximum for EBV_{09}. At $\lambda = 0.7$, EBV_{09} R^2 increased to 51%, which was 1% better than for the multiple-step approach (Table 1); δ was also higher than for the multiple-step approach by 0.01. The δ can be increased to 0.92 with only a slight decrease in R^2. Because the primary interest of breeders is to identify animals with the highest genetic merit, a moderate reduction in bias (i.e., higher δ) would be preferred to a small increase in overall accuracy (R^2).

Accuracy of the single-step approach was dependent on the choice of **G** and the weighting placed on the difference between **G** and **A**. With the proper choice, accuracy of the single-step approach was superior to that of the multiple-step approach. One reason why the choice of **G** is critical is that genomic and pedigree relationship matrices should be compatible in both scale and structure. The importance of structure can be seen from the decomposition of the genomic breeding value in Appendix B. The weight of PA relative to genomic information depends on λ and even more on diagonals of $\mathbf{G}^{-1}$ and $\mathbf{A}_{22}^{-1}$. In general, the diagonal of $\mathbf{G}^{-1}$ depends on the genomic relationships and measures the amount of information provided to individual i by other animals.

The primary influence of the weighting factor (λ) appears to be related to the proportion of the additive variance explained by the genomic information (Appendix B). Snelling et al. (2009) found that different numbers of SNP genotypes used for the construction of **G** resulted in different decomposition of the additive variance between the genomic and polygenic effects. Genomic information from the best genotyped bulls would add relationship information for several animals and most likely result in higher additive variance. The Canadian official genomic evaluation system for Holsteins (Van Doormaal et al., 2009) assumes that only 80% of the additive variance is explained by the SNP information. Other factors behind the weighting factor may be related to final score as a trait in US Holsteins. For example, heritability based on records of grade animals is lower than with records on registered animals (Koduru, 2006). 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). Future studies with more traits and species will clarify the influence of the weighting factor as well as alternative weighting factors. Although our decomposition between the genomic and polygenic effects involved inverses of the respective matrices, it can also be done on the direct scale, by assuming that only part of the genetic variance is explained by the genomic information (Christensen and Lund, 2009).

What **G** should be is still undetermined. As implemented for this study, **G** was constructed so that linear effects were assumed for SNP genotypes while also collecting information about realized relationships (Van Raden, 2008). Other alternatives exist. For example, matrix **K** in González-Recio et al. (2008) included a similarity index across genotypes. Probabilities for

identity by descent can also be used and averaged across loci (Villanueva et al., 2005).

Use of regression coefficients to measure bias was described by Reverter et al. (1994) and forms the basis of Method R estimation of variance components. However, the use of δ to calibrate GP might be problematic. First, it relies on the same set of equations being used for old and recent evaluations, which was not true for this study; the "old" evaluation ($PedGen1_{04}$) used **H**, whereas the "recent" evaluation (Ped_{09}) used **A**. Second, as seen by experience with Method R, the estimated regression coefficient has large error and might be biased, especially with selection (Cantet et al., 2000; Schenkel and Schaeffer, 2000). On the other hand, little bias and very efficient computations were reported by Druet et al. (2001), who traced the bias to the use of fixed effects estimated from subsets of the data.

For this study, **G** was constructed with equal variances assumed for SNP marker effects. When variances are not equal; for example, as in Bayes-A or Bayes-B (Meuwissen et al., 2001), an equivalent **G** can be constructed by scaling contributions from different markers. Such construction requires precomputing those variances based on genotyped individuals and pseudo-data.

The generalization of the single-step approach to multiple traits is obvious when **G** is identical for each trait. However, separate **G** matrices for each trait may require single-trait analyses. For several traits, the benefits and simplicity of multiple-trait analysis using the same **G** may overcome the loss of accuracy from using less than the optimal **G** for each trait.

The single-step approach to evaluation as described in this study is easy to implement just by modifying the relationship matrix for current evaluations. Aside from simplification of genomic evaluation, the procedure is expected to improve evaluations for all ungenotyped animals. Updated PA and PTA for descendants of genotyped animals are possible using multiple-step methods with additional calculations (see http://aipl.arsusda.gov/reference/changes/eval0901.html). Advantages of single-step evaluations should increase in the future when animals have been pre-selected on genotypes. Traditional evaluations expect that Mendelian sampling averages zero, but in the future only animals with positive Mendelian sampling may receive phenotypes.

A substantial part of any current genomic selection is validation for young animals. In contrast, in BLUP based on pedigree information, such a validation is rarely performed and is implicitly replaced by variance component estimation, although some validation is performed indirectly for analyses used by Interbull multiple-trait across-country (MACE) evaluations (Interbull, 2001). With some assumptions, it is possible that the parameters of a single-step procedure are regular variance components plus weighting factors, either as proposed in this study or different. In such a case, the validation steps can be replaced by parameter estimation, greatly simplifying the use of the genomic information. Ways to estimate values of weighting factors by REML, Markov chain Monte Carlo, or other methods remain to be investigated.

CONCLUSIONS

Full genomic and pedigree evaluations by the single-step approach were as good as those obtained with the multiple-step approach in terms of accuracy and bias. Generalization for complex data structures or more complicated models are straightforward. Additional computational cost was small relative to pedigree evaluation. The highest accuracy was obtained with a scaled genomic relationship matrix created under the assumption of equal allele frequencies. The main advantages of the single-step approach are its simplicity and automatic weights for the various sources of information for the overall breeding value. Moreover, advantages of single-step evaluations should increase in the future when animals are preselected on genotypes.

ACKNOWLEDGMENTS

The authors thank P. M. VanRaden and G. R. Wiggans (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD), J. R. O'Connell (University of Maryland School of Medicine, Baltimore), C. P. Van Tassell (Bovine Functional Genomics Laboratory, ARS, USDA, Beltsville, MD), and L. Varona (Universidad de Zaragoza, Spain) as well as Holstein Association USA Inc. (Brattleboro, VT) and the Cooperative Dairy DNA Repository (Beltsville, MD) for providing genotypic data. Financing from Agence National de la Recherche project AMASGEN (Jouy en Josas, France) is acknowledged. Editing assistance was provided by Suzanne Hubbard (Animal Improvement Programs Laboratory, ARS, USDA, Beltsville, MD). Helpful comments and suggestions from the two reviewers are acknowledged.

REFERENCES

Cantet, R. J. C., A. N. Birchmeier, M. G. Santos-Cristal, and V. S. de Avila. 2000. Comparison of restricted maximum likelihood and method R for estimating heritability and predicting breeding value under selection. J. Anim. Sci. 78:2554–2560.

Christensen, O. F., and M. S. Lund. 2009. Genomic relationship matrix when some animals are not genotyped. Page 299 in Proc. 60th Annual Meeting EAAP, Barcelona, Spain. Wageningen Press, Wageningen, the Netherlands.

Cole, J. B., P. M. VanRaden, J. R. O'Connell, C. P. Van Tassell, T. S. Sonstegard, R. D. Schnabel, J. F. Taylor, and G. R. Wiggans. 2009. Distribution and location of genetic effects for dairy traits. J. Dairy Sci. 92:2931–2946.

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

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel, and J. M. Cotes. 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182:375–385.

Druet, T., I. Misztal, M. Duangjinda, A. Reverter, and N. Gengler. 2001. Estimation of genetic covariances with Method R. J. Anim. Sci. 79:605–615.

Gengler, N., P. Mayeres, and M. Szydlowski. 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.

Gianola, D., G. A. de los Campos, W. G. Hill, E. Manfredi, and R. L. Fernando. 2009. Additive genetic variability and the Bayesian alphabet. Genetics 183:347–363.

Gianola, D., R. L. Fernando, and A. Stella. 2006. Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173:1761–1776.

González-Recio, O., D. Gianola, N. Long, K. A. Weigel, G. J. M. Rosa, and S. Avendaño. 2008. Nonparametric methods for incorporating genomic information into genetic evaluations: An application to mortality in broilers. Genetics 178:2305–2313.

Hayes, B. J., P. J. Bowman, A. J. Chamberlain, and M. E. Goddard. 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.

Interbull. 2001. Interbull guidelines for national & international genetic evaluation systems in dairy cattle with focus on production traits. Interbull Bull. 28. http://www-interbull.slu.se/bulletins/bulletin28/Interbull%20Guidelines-2001.pdf Accessed Nov.9, 2009.

Koduru, V. K. R. 2006. Changes in genetic evaluations from 1st to 2nd crop for final score in Holsteins. MS Thesis. Univ. Georgia, Athens.

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. E., 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., 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, T. Strabel, B. Auvray, T. Druet, and D. H. Lee. 2002. Blupf90 and related programs (BGF90). Commun. No. 28–07 in Proc. 7th World Congress on Genetics Applied to Livestock Production, Montpellier, France.

Nejati-Javaremi, A., C. Smith, and J. P. Gibson. 1997. Effect of total allelic relationship on accuracy of evaluation and response to selection. J. Anim. Sci. 75:1738–1745.

Reverter, A., B. L. Golden, R. M. Bourdon, and J. S. Brinks. 1994. Technical note: Detection of bias in genetic predictions. J. Anim. Sci. 72:34–37.

Schenkel, F. S., and L. R. Schaeffer. 2000. Effects of nonrandom parental selection on estimation of variance components. J. Anim. Breed. Genet. 117:225–239.

Snelling, W. M., L. A. Kuehn, R. M. Thallman, J. W. Keele, and G. L. Bennett. 2009. Genomic heritability of beef cattle growth. J. Anim. Sci. 87(E-Suppl. 2):396. (Abstr.)

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

Tsuruta, S., I. Misztal, and I. Stranden. 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.

Van Doormaal, J., G. Kistemaker, P. G. Sullivan, M. Sargolzaei, and F. S. Schenkel. 2009. Canadian implementation of genomic evaluations. Interbull Bull. 40. http://www-interbull.slu.se/bulletins/bulletin40/Pre/ITB_Van_Doormaal.pdf Accessed Nov. 9, 2009.

VanRaden, P., M. Tooker, and N. Gengler. 2008. Effects of allele frequency estimation on genomic predictions and inbreeding coefficients. J. Dairy Sci. 91(E-Suppl. 1):506. (Abstr.)

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. Interbull Bull. 37:33–36.

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

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

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

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

Villanueva, B., R. Pong-Wong, J. Fernández, and M. A. Toro. 2005. Benefits from marker-assisted selection under an additive polygenic genetic model. J. Anim. Sci. 83:1747–1752.

APPENDIX A

Let the inverse of the numerator relationship matrix ($\mathbf{A}$) be:

$$\mathbf{A}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} \end{bmatrix},$$

where animals are partitioned into 2 groups with group 2 denoting genotyped animals.

To derive an inverse for the combined relationship matrix of Legarra et al. (2009), using the properties of the inverse of partitioned matrix, useful identities from $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ are

$$\mathbf{A}^{11}\mathbf{A}_{11} + \mathbf{A}^{12}\mathbf{A}_{21} = \mathbf{I}, \quad [A1]$$

$$\mathbf{A}^{21}\mathbf{A}_{12} + \mathbf{A}^{22}\mathbf{A}_{22} = \mathbf{I}, \quad [A2]$$

$$\mathbf{A}^{11}\mathbf{A}_{12} + \mathbf{A}^{12}\mathbf{A}_{22} = \mathbf{0}, \quad [A3]$$

$$\mathbf{A}^{21}\mathbf{A}_{11} + \mathbf{A}^{22}\mathbf{A}_{21} = \mathbf{0}, \text{ and} \quad [A4]$$

$$\left(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)^{-1} = \mathbf{A}^{11}. \quad [A5]$$

Using [A1] through [A4] and multiplying the whole-population matrix

$$\mathbf{H} = \mathbf{A} + \begin{bmatrix} \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{A}_{12}\mathbf{A}_{22}^{-1}(\mathbf{G} - \mathbf{A}_{22}) \\ (\mathbf{G} - \mathbf{A}_{22})\mathbf{A}_{22}^{-1}\mathbf{A}_{21} & \mathbf{G} - \mathbf{A}_{22} \end{bmatrix}$$

by

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

gives $\mathbf{H}^{-1}\mathbf{H} = \mathbf{I}$.

A direct approach to getting $\mathbf{H}^{-1}$ comes from the distribution function. Based on the conditional distribution

$$\mathbf{u}_1 \mid \mathbf{u}_2 \sim N\left(\mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2,\ \mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\right)$$

and [A1] through [A5], the full distribution can be written as

$$\begin{aligned} p(\mathbf{u}_1, \mathbf{u}_2) &= p(\mathbf{u}_1, \mathbf{u}_2 \mid \mathbf{u}_2)p(\mathbf{u}_2) \\ &= p(\mathbf{u}_1 \mid \mathbf{u}_2)p(\mathbf{u}_2) \\ &\propto \exp[-0.5(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)'\mathbf{A}^{11}(\mathbf{u}_1 - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{u}_2)]\exp[-0.5\mathbf{u}_2'\mathbf{G}^{-1}\mathbf{u}_2] \quad [A6] \\ &= \exp\left(-0.5\begin{bmatrix}\mathbf{u}_1' & \mathbf{u}_2'\end{bmatrix}\begin{bmatrix} \mathbf{A}^{11} & -\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -\mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11} & \mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}\mathbf{A}_{21}\mathbf{A}^{11}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right) \\ &= \exp\left(-0.5\begin{bmatrix}\mathbf{u}_1' & \mathbf{u}_2'\end{bmatrix}\begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{G}^{-1} + \mathbf{A}^{22} - \mathbf{A}_{22}^{-1} \end{bmatrix}\begin{bmatrix}\mathbf{u}_1 \\ \mathbf{u}_2\end{bmatrix}\right). \end{aligned}$$

The matrix in [A6] is the inverse of the variance matrix of the full distribution. Therefore,

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

APPENDIX B

To illustrate the role of λ and decomposition of joint predictions in PA, genomic prediction (GP), and pedigree prediction from the subset of genotyped relatives (PP_{22}), consider $\mathbf{H}^{-1}$ after including λ:

$$\mathbf{H}^{-1} = \begin{bmatrix} \mathbf{A}^{11} & \mathbf{A}^{12} \\ \mathbf{A}^{21} & \mathbf{A}^{22} + \lambda(\mathbf{G}^{-1} + \mathbf{A}_{22}^{-1}) \end{bmatrix}. \quad [B1]$$

Denote $\mathbf{H}^{-1}$ as $\{h^{ij}\}$, $\mathbf{A}^{-1}$ as $\{a^{ij}\}$, $\mathbf{G}^{-1}$ as $\{g^{ij}\}$, and $\mathbf{A}_{22}^{-1}$ as $\{a_{22}^{ij}\}$. Consider the equation for breeding value u_i of individual i without records or progeny, in the spirit of VanRaden and Wiggans (1991); k indicates genotyped individuals (in $\mathbf{A}_{22}$), and j indicates all individuals (in $\mathbf{A}$):

$$\sum_j h^{ij} u_j = 0 \text{ and}$$

$$\lambda \sum_k g^{ik} u_k + (1-\lambda) \sum_k a_{22}^{ik} u_k + \sum_j a^{ij} u_j - \sum_k a_{22}^{ik} u_k = \lambda \sum_k g^{ik} u_k - \lambda \sum_k a_{22}^{ik} u_k + \sum_j a^{ij} u_j = 0.$$

Thus, for $\lambda = 0$, only contributions from pedigree relationships remain.

Consider more specifically young animal i without records or progeny. The equation with inbreeding ignored is

$$-u_s - u_d + 2u_i + \lambda \sum_j (g^{ij} - a_{22}^{ij}) u_j = 0,$$

where s and d correspond to sire and dam, respectively. Then,

$$\begin{aligned} u_i &= \frac{u_s + u_d + \lambda \sum_{j,j \neq i} (a_{22}^{ij} - g^{ij}) u_j}{2 + \lambda(g^{ii} - a_{22}^{ii})} \\ &= \left(\frac{u_s + u_d}{2} \right) \left(\frac{2}{2 + \lambda(g^{ii} - a_{22}^{ii})} \right) + \left(\frac{\lambda \sum_{j,j \neq i} a_{22}^{ij} u_j}{2 + \lambda(g^{ii} - a_{22}^{ii})} \right) - \left(\frac{\lambda \sum_{j,j \neq i} g^{ij} u_j}{2 + \lambda(g^{ii} - a_{22}^{ii})} \right) \\ &= 2(w) PA + \lambda(w) g^{ii} GP - \lambda(w) a_{22}^{ii} PP_{22}, \end{aligned} \quad [B2]$$

where

$$PA = \left(\frac{u_s + u_d}{2} \right),$$

$$GP = \frac{-\sum_{j,j \neq i} g^{ij} u^j}{g^{ii}}, \text{ and}$$

$$PP_{22} = \frac{-\sum_{j,j\neq i} a_{22}^{ij} u^j}{a_{22}^{ii}};$$

that is, parent average, genomic prediction, and subset pedigree prediction, with weights summing to 1. These are the same sources of information as in VanRaden et al. (2009b) except that they are estimated jointly. Note that PP_{22} might be different from PA because 1) both parents might not be genotyped and 2) only the subset of genotyped animals is considered if PP_{22} is computed independently (as in PedGenM$_{04}$). If $\lambda = 0$, only PA remains; if $\lambda = 1$, then weighting of the 3 sources of information depends on the elements $a^{ii} = 2$, g^{ii}, and a_{22}^{ii}, which measure precision of the 3 information sources relative to other breeding values. That approach is similar to the reliabilities used to combine the 3 information sources in VanRaden et al. (2009b).