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 $\mathbf{H}$: