Large-Scale Single-Step Genomic Evaluation for Milk Production Traits

B. L. Harris, A. M. Winkelman and D. L. Johnson
LIC, Private Bag 3016, Hamilton, New Zealand

Abstract

The single-step method of genomic evaluation for milk volume, fat yield and protein yield was applied to the national dairy herd of New Zealand. Genomic information from a 50K SNP chip was available on 5402 Holstein Friesian (HF), Jersey (J) and HFxJ sires. The genomic relationship matrix (GRM) or the Euclidean distance matrix (EDM) in a Gaussian kernel was used to augment the pedigree-based relationship matrix in the mixed model equations. Scale parameters of 0.3, 0.5 and 0.7 were used for the GRM and 0.5, 0.7 and 0.9 for the EDM. Traditional breeding values (BVs) were compared to genomic breeding values (GBVs) in two youngest cohorts of the progeny-tested sires (N=525). An increasing scale parameter was associated with an increased inflation of the GBVs. The EDM resulted in lower inflation of fat GBVs than the GRM. The effect was smaller and more variable for the other traits. Augmenting the relationship matrix with the GRM versus the EDM and changing the magnitude of the scale parameters had little impact on the accuracy of the evaluation.

Introduction

A single-step method of genomic evaluation, that simultaneously uses phenotypic, genomic and relationship information, was first proposed by Misztal *et al.* (2009). The method entails augmenting the pedigree-based relationship matrix by a genomic relationship matrix (GRM) that is then incorporated into the mixed model equations (MME). Misztal *et al.* (2010) have enhanced the single-step method by modifying the augmented relationship matrix to adjust for the scale of the genomic predictions, thereby providing a way to adjust for inflation of the genomic breeding values (GBVs). The calculation of the GRM requires estimates of the base population SNP frequencies. This is straightforward in single-breed evaluations. However, because SNP frequencies differ by breed, an across-breed evaluation has multiple base populations, with crossbred animals descending from more than one population. Harris and Johnson (2010) describe a method to adjust the GRM for a multi-breed, pure-and crossbred population. The method requires the individual animal breed proportions to be known and is computationally intensive. The Euclidean distance matrix (EDM) in a Gaussian kernel, proposed by Gianola and van Kaam (2008) is an alternative method of incorporating genomic information into the MME that does not require information on individual breed proportions.

The aim of this study was to assess the single-step method of genomic evaluation for milk volume, fat yield and protein yield in the New Zealand (NZ) dairy population. A full description of current across-breed, multiple-trait (MT) random regression (RR) test-day model (TDM) used for national genetic evaluation is provided by Harris *et al.* (2006). Genomic information of sires was included in the MME via either the GRM or the EDM. The accuracy and inflation of the genomically enhance evaluations was assessed by regressing the traditional breeding values (BVs) on the GBVs.

Methods and Materials

Data

Herd test data for milk volume, fat yield and protein yield extracted from NZ’s national database were used in this study. The data included records from seasons 1986 to 2011 (In NZ, a season starts in June and ends in May the next year. Hence, an animal calving in season 2011 finishes her lactation in 2012). Pedigree data and breed information from the 1940s onwards were used to calculate the additive genetic relationship matrix and genetic groups. The data set included a total of 171,994,721 testday records from 22,483,802 animals (cows and ancestors). The cow population consisted primarily of Holstein Friesians (HF) (52%), Jerseys (J) (18%) and Friesian-Jersey Crosses (HFxJ) (29%), with the remaining cows being other crosses and breeds. A total of 5,402 sires born in seasons 1965 through 2007 were genotyped on the 50K SNP chip (which contained 38,108 SNPs after quality control).

Statistical Model for National Genetic Evaluation

The model for the national genetic evaluation of each production trait was a MT RR TDM where lactations 1, 2, 3 and lactations 4 through 6 are modelled as different genetic traits. Separate random permanent environmental effects were fitted for each of the 6 lactations. The order of the MME for each trait was approximately 550 millon. Full details of the statistical model are given by Harris *et al.* (2006). BVs were calculated for days 3 to 270 within each lactation. Results for the BVs averaged over the 4 lactations are used in this study.

Genomic information was included in the national evaluation using the single-step method (Misztal *et al.*, 2009). For this method, the pedigree-based relationship matrix was augmented by either the GRM or EDM. The inverse of the augmented relationship matrix (H) was calculated as follows:

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

with,

$$\mathbf{B} = \lambda(\mathbf{G}^{-1} - \mathbf{A}_{22}^{-1})$$

where **G** is the GRM or EDM, A^{-1}_{22} is the submatrix of the pedigree relationship matrix pertaining to genotyped animals and λ is the scale parameter.

The across-breed GRM was calculated by adapting the method of vanRaden (2008) to multiple breeds. Essentially, this required the SNP marker matrix to be adjusted so that the SNP markers had a mean of zero within breed and within- and across-breed variances equivalent to those of the pedigree relationship matrix. The GRM was corrected for differences between genotyped and non-genotyped base populations by extending the method of Vitezica *et al.* (2011) to multiple breeds. The EDM was calculated using the method of Gianola and van Kaam (2008). A bandwidth of 4/3 was used for all traits and scale parameters. The ranges of the scale parameters for the GRM and EDM were chosen based on Harris et al. (2011). Scale parameters of 0.3, 0.5 and 0.7 were used for the GRM and 0.5, 0.7 and 0.9 for the EDM.

The effect of incorporating genomic information in the MME was assessed by comparing traditional BVs to the GBVs. The traditional analyses included all herd test data up and including season 2011. The genomic analysis included all herd test data up and including season 2009. In these analyses, sires born in 2006 and 2007 (N=525; N=251 HF, N=104 HFxJ, N=170 J), whose first-crop daughters completed their first lactations in seasons 2010 and 2011, respectively, were the test population. Genotyped sire born prior to 2006 will be referred to as the training sires. The accuracy of prediction was calculated as the correlation between the traditional BVs (obtained using all data) and GBVs of test animals. The inflation was assessed using the regression slope of traditional BVs on GBVs, a slope of unity indicating no inflation.

Computational Strategy

The MME were solved using a preconditioned conjugate gradient (PCG) solver (Strandén and Lidauer, 1999) and iteration on data with code reordering (Tsurata *et al.*, 2001). The prior solutions for the single-step method were the solutions from the traditional genetic evaluation. The matrix **B** was precalculated for genotyped animals prior to solving the mixed model equations. Matrix inversion and multiplication were done using the Intel MKL libraries. The PCG solver for the single-step model used the same procedure as the traditional model with the addition of a new routine that updated the PCG multiplication step, the product of projection vector and the

MME, for the product of B and projection vector. The new routine used direct multiplication of B and the projection vector elements pertaining to the genotyped animals. This multiplication was repeated for each trait.

Results and Discussion

The single-step models all converged within 500 iterations after the inclusion of the genomic information. The time increase, per iteration, from adding the genomic information to the model was approximately 5 seconds resulting in a total iteration taking 2 minutes and 36 seconds.

Tables 1, 2 and 3 show the inflation and accuracy of the GBVs of the test sires for milk, fat and protein, respectively. The GBVs of the HF and J sires were inflated (i.e. regression coefficient less than 1.0) regardless of the whether the relationship matrix was augmented with the GRM or EDM. The inflation increased (regression coefficient decreased) with increasing scale parameter. The same trend was found for the FJX sires, but in the cases where the coefficients were greater than 1.0, an increased scale parameter was associated with decreased deflation of the GBVs. The biggest difference between the GRM and EDM was found for the fat GBVs, where the EDM resulted in lower inflation. The EDM tended to result in lower inflation for the milk and protein, but the effect was smaller and less consistent than it was with fat. A single value of the scaling parameter across breeds is a compromise – if the weighted, across-breed mean of the inflation factor was close to unity, one or more breed(s) would have inflated GBVs and the remaining one(s) would have deflated BVs. Augmenting the relationship matrix with the GRM versus the EDM and changing the magnitude of the scale parameters had little impact on the accuracy of the evaluation.

The within-breed correlations among the GBVs of the training sires, calculated using the different relationship matrices (GRM and EDM) and scaling factors, were greater than 0.997 for all traits. The correlations among the GBVs of the test sires ranged between 0.90 and 0.99. The lowest correlations were observed between the two most extreme scenarios, GRM + □=0.3 and EDM + □=0.9.

The GBV means and standard deviations for the different augmented relationship matrices and scaling factors were nearly identical for training sires across all traits. In contrast, the means for the test sires were regressed more towards the breed means and the standard deviation of the GBVs increased as the scaling factor increased for all traits and both augmented relationship matrices.

The single-step procedure outlined in this paper was computationally feasible for a complex genetic evaluation model with a large amount of data. Augmentation of the relationship matrix with an EDM tended to result in lower levels of inflation of the GBVs than did the GRM. The choice of the optimal scale parameters will be more challenging in across-breed genomic evaluations compared to single-breed evaluations.

Table 1. The inflation and accuracy[1] of the GBVs for milk of the test sires.

Across-breed Genomic Relationship Matrix						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.7	0.80	0.81	1.04	0.68	0.69	0.87
0.5	0.84	0.87	1.08	0.69	0.70	0.88
0.3	0.87	0.92	1.14	0.68	0.70	0.88
Euclidean distance matrix in a Gaussian Kernel						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.9	0.84	0.81	1.04	0.65	0.66	0.85
0.7	0.91	0.90	1.09	0.67	0.68	0.87
0.5	0.93	0.95	1.14	0.68	0.70	0.88

[1] Correlation, HF = Holstein Friesian, J = Jersey and X = Holstein Friesian x Jersey Crossbred Sires

Table 2. The inflation and accuracy[1] of the GBVs for fat of the test sires.

Across-breed Genomic Relationship Matrix						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.7	0.83	0.87	0.89	0.70	0.75	0.71
0.5	0.88	0.94	0.95	0.71	0.76	0.71
0.3	0.92	0.98	0.97	0.70	0.75	0.69
Euclidean distance matrix in a Gaussian Kernel						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.9	0.91	0.92	0.98	0.69	0.74	0.71
0.7	0.94	0.97	1.02	0.70	0.75	0.71
0.5	1.00	1.00	1.06	0.70	0.75	0.71

[1] Correlation, HF = Holstein Friesian, J = Jersey and X = Holstein Friesian x Jersey Crossbred Sires

Table 3. The inflation and accuracy[1] of the GBVs for protein of the test sires.

Across-breed Genomic Relationship Matrix						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.7	0.72	0.73	1.04	0.58	0.65	0.83
0.5	0.80	0.78	1.11	0.60	0.65	0.84
0.3	0.86	0.83	1.18	0.60	0.64	0.85
Euclidean distance matrix in a Gaussian Kernel						
Value for □	Inflation			Accuracy		
	HF	J	HFxJ	HF	J	HFxJ
0.9	0.73	0.70	1.06	0.54	0.63	0.83
0.7	0.83	0.77	1.13	0.57	0.64	0.85
0.5	0.90	0.83	1.17	0.59	0.64	0.86

[1] Correlation, HF = Holstein Friesian, J = Jersey and X = Holstein Friesian x Jersey Crossbred Sires

References

Gianola, D.G & van Kaam, J.B.C.H.M. 2008. Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics 178*, 2289–2303.

Harris B.L., Winkelman, A.M., Johnson, D.L. & Montgomerie, W.A. 2006. Development of a national production testday model for New Zealand. *Interbull Bulletin 35,* 23-27.

Harris, B.L, Johnson, D.L. & Spelman, R.J. 2011. Choice of parameters for removal of inflation genomic breeding values for dairy cattle. Proc. Assoc. Advmt. *Anim. Breed. Genet. 19,* 359-362.

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.

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.

Strandén, I. & Lidauer, M. 1999. Solving large mixed linear models using preconditioned conjugate gradient iteration. *J. Dairy Sci. 82,* 2779–2787.

Tsurata, 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.

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

Vitezica, Z.G., Aguilar, I., Misztal, I. & Legarra, A. 2011. Bias in genomic predictions for populations under selection. *Genet. Res. (Camb). 93(5),* 357-366.