Impact of Including a Large Number of Female Genotypes on Genomic Selection

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

Abstract

A method to calculate genomically enhanced breeding values (BVs) for LIC's in-house genomic evaluation system has been assessed. The method is a hybrid single-step (SS) method of incorporating information from the pedigree-based and the genomic relationship matrices into the mixed model equations. The predictions of the test bulls obtained using the SS method were approximately 5% more accurate than those obtained from the current two-step genomic evaluation system. The hybrid SS method produced BVs that were 20 to 30% less inflated than those of the current system. To compare the bias and accuracy of including just the female genotypes independent of method changes, the hybrid method was used where only sire genotypes were included. The inclusion of the female genotypes gave small improvements in both the bias and accuracy in the validation results.

Introduction

Genomically enhanced breeding values (GEBVs) have been published in New Zealand (NZ) since 2008. The calculation and publication of the evaluations is overseen by New Zealand Animal Evaluation Ltd. (NZAEL), an industry-good body that is tasked with ensuring optimum genetic improvement of the NZ dairy herd. The current method of calculating GEBVs is a multi-step method in which the traditional breeding values (BVs) are blended with direct genomic values (DGVs) using a selection index approach, as originally described by VanRaden (2009). The DGVs are obtained using best linear unbiased prediction (BLUP) with a genomic relationship matrix (GRM), and the phenotype is the deregressed breeding value (DRBV). Genomic information on females is not included in this analyses, other than that of a limited number of bull dams. The blended method does not have an in-built system of controlling inflation and adjustments are made to the GEBVs post-calculation.

The current national system of genomic evaluation will be discontinued by NZAEL in September 2013. LIC has made made the decision to develop an in-house genomic evaluation system. A project was undertaken to assess methodologies to calculate GEBVs that would overcome some of the deficiencies of the current system. These include the need for the (possibly ad hoc) blending of the traditional and genomics BVs, the post-calculation inflation adjustments and not utilising the increasing number of female genotypes that are becoming available. The desired outcome of overcoming the deficiencies is reduced bias and increased accuracy of the GEBVs over the current system. An alternative to the multi-step procedure is the single-step (SS) method, 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 GRM that is then incorporated into the mixed model equations (MME). Misztal *et al.* (2010) have enhanced the SS method by modifying the augmented relationship matrix to adjust for the scale of the genomic predictions, thereby providing a way to control inflation of the GEBVs. This method was used with genomic information on 5402 bulls, within the New Zealand national evaluation (Harris *et al.,* 2012). However, attempts to include the genomic information on 50,000+ cows resulted in convergence problems when solving the MME and was deemed to be infeasible for routine evaluation, at present. An alternative to the SS is the hybrid SS method (Harris *et al.,* 2011). The method includes all genotyped bulls and cows as well as all their ancestors. Hence the size of the equations is

considerably smaller than that of the full national SS method. The smaller size makes the system computationally feasible to solve and the use of the SS methodology retains the advantages of providing a means of controlling bias within the evaluation and obviates the need for blending of the DGVs with the traditional BVs.

The objective of this paper is to outline the new LIC genomic evaluation and compare theis to the current genomic evaluation for milk, milkfat, and protein traits.

Materials and Methods

Phenotypes and Genotypes

The three traits in the national breeding objective were considered in this study. They are milk, fat, and protein yield. The phenotype for all analyses was the deregressed BV (DRBV) for each trait as described in Harris and Johnson (2010). The genotypes were from the Illumina BovineSNP50 Beadchip panel, 34,963 SNPs after removing SNP for low call rates, minor allele frequencies 2%, non-Mendelian inheritance, failed Hardy-Weinberg tests and low imputation accuracy were used.

Animals for model validation

The data for model validation consisted of all of LIC's genotyped bulls born in 2007 or earlier, a selection of CRV Ambreed genotyped bulls and genotyped females born in 2006 or earlier. All bulls were progeny tested. The genomic analyses were run using the phenotypic data that would have been available at the end of season 2008. Sires born in seasons 2005, 2006 and 2007, whose first-crop daughters completed their first lactations in seasons 2009, 2010 and 2011, respectively, were the test population. Their genotypes, but not their phenotypes, were included in the analyses. Genotyped sires born prior to 2005 will be referred to as the training sires. The accuracy of prediction was calculated as the correlation between the DRBVs (obtained using data available at the end of season 2013) and GEBVs of test animals. The bias was assessed using the regression slope from this anaylsis.

Animals for national evaluation

Following model validation, GEBVs were calculated for the national population for the three traits. The national evaluation consisted of all genotyped animals, and their ancestors, born in 2012 or earlier. Table 1 contains a summary of the numbers of animals in both the model validation and national datasets.

Current Genomic Evalution

The current system is an across-breed evaluation where the DGVs are estimated using BLUP that includes the GRM that is adjusted for breed frequencies as outlined by Harris and Johnson (2010). Genotypes are included for all bulls and a limited number of bull dams. The phenotype was the DRBV. The GEBVs were obtained by blending the DGVs and the DRBVs using a selection index method as outlined by VanRaden (2009). The blended BV is adjusted for inflation by scaling the Mendelian sampling (MS) component of the GEBVs.

Hybrid single-step evaluation

Problems with equation convergence when large numbers of female genotypes were included in a full SS genomic evaluation have led to the development of a hybrid SS method (Harris *et al.*, 2010) which has the advantage of providing a computationally efficient genomic evaluation using traditional national BV, based on pedigree, as the starting point. Moreover, the same model can be used for all traits. The method sssentially is the deregression procedure is then reversed but with the genomic relationship matrix replacing the pedigree relationship matrix. A preconditioned conjugate-gradient method is used to solve these equations after first using matrix inversion techniques to calculate the inverse of the GRM and associated partitioned A matrix.

Calculation of reliability uses matrix inversion of MME pertaining to the genotyped individuals after absorption of the equations for non-genotyped individuals. The reliabilities of ungenotyped ancestors can be updated for the genomic information provided by their progeny using methods outlined in Harris and Johnson (1998). For genotyped animals, we determine the contribution to their reliabilities from genomics using the reliabilities determined above and the reliabilities from the national evaluation based on pedigree. Then, working from youngest to oldest, this genomic information is blended with the pedigree based reliabilities of the ancestors. Genomic BV for ungenotyped descendants are then obtained using the technique outlined in Harris and Johnson (2010).

Table 1. Numbers of animals in both the model validation and national datasets for the current and hybrid single step models.

	Validation Dataset		National Dataset	
Method of analysis	Current GBLUP	Hybrid SS	Current GBLUP	Hybrid SS
Genotyped males	7102	7102	14343	14343
Genotyped females	454	17559	454	47574
Non-genotyped males	0	8541	0	12888
Non-genotyped females	0	76545	0	165521

Table 2. Bias and accuracy of GEBVs based on the current GBLUP system.

Trait	Breed	Bias	Accuracy
milk	Friesian	0.64	0.62
milk	Jersey	0.80	0.74
milk	FJ Cross	0.77	0.68
fat	Friesian	0.62	0.59
fat	Jersey	0.72	0.62
fat	FJ Cross	0.67	0.59
prot	Friesian	0.59	0.60
prot	Jersey	0.70	0.62
prot	FJ Cross	0.70	0.61

The genomic BVs are then recombined with the genetic group solutions. This single-step method avoids the blending step involved in the current method of genomic evaluation.

Genomic Relationships

The genomic information was incorporated into the MME via the GRM or the Euclidean distance matrix (EDM) in a Gaussian kernel, as proposed by Gianola and van Kaam (2008). Harris and Johnson (2010) describe a method to adjust the GRM for a multi-breed population. This method is computationally intensive, but feasible for a relatively small population of genotyped animals. The method is infeasible for large genotyped populations, as would be the case when tens of thousands of cows are genotyped. Makgahlela et al. (2013) outline a method of calculating the breed-adjusted GRM that is computationally feasible for large populations. This method was used to calculate an across-breed GRM. Evaluations are done using the hybrid SS method where the pedigree-based relationship matrix is augmented with either the GRM or EDM. A number of scaling factors are used to control the inflation of the resulting GEBVs.

Table 3. Bias and accuracy of GEBVs based on the hybrid SS method.

Trait	Breed	Bias	Accuracy
milk	Friesian	0.96	0.72
milk	Jersey	1.01	0.70
milk	FJ Cross	1.02	0.83
fat	Friesian	1.03	0.75
fat	Jersey	0.99	0.70
fat	FJ Cross	1.06	0.70
prot	Friesian	0.99	0.62
prot	Jersey	0.99	0.64
prot	FJ Cross	1.06	0.73

Scale parameter: milk=0.7, fat=0.8 and protein=0.6

Results

Validation Data

Based on validation testing, the across breed GRM method has proved inferior to the definition based on EDM. In addition, convergence monitoring has at times identified

some instability perhaps due to near singularity conditions for the GRM. Only the results from hybrid SS with EDM will discussed. Table 2 contains the bias and accuracy of the GEBVs of test animals evaluated using the current system. The bias estimates are those obtained prior to any adhoc post processing inflation adjustments. For most of the traits, the estimates are substantially less than one, indicating considerable inflation of the GEBVs. Table 3 shows the inflation and accuracy of the test sires GEBVs obtained using the hybrid SS EDM model with different scale parameters ranging from 0.3 to 1.0, with the optimal scale parameter for each trait being different (see Table 3). The GEBVs are considerably less inflated than those of the current model. The variation in the scale parameter had a small affect on the inflation and accuracy of the GEBVs.

Table 4. Bias and accuracy of GEBVs based on the hybrid SS method excluding female genotypes.

Trait	Breed	Bias	Accuracy
milk	Friesian	0.93	0.69
milk	Jersey	0.97	0.66
milk	FJ Cross	0.98	0.81
fat	Friesian	1.09	0.72
fat	Jersey	1.07	0.64
fat	FJ Cross	1.10	0.67
prot	Friesian	0.95	0.61
prot	Jersey	0.94	0.62
prot	FJ Cross	0.98	0.71

Scale parameter: milk=0.7, fat=0.8 and protein=0.6

To compare the bias and accuracy of including just the female genotypes independent of method changes, the sire genotypes from the validation dataset that are used in the current GBLUP system was also used in the hybrid SS method. Table 4 contains the bias and accuracy of the GEBVs of test animals evaluated using the hybrid SS method without the female genotypes. Comparing table 3 and 4 the inclusion of the female genotypes increased accuracy by 1% to 4% and generally reduced bias by a factors ranging from 0.02 to 0.06.

National Data

Analysis from of the BVs/GEBVs from the national dataset showed that for proven sires, the GEBVs from the hybrid SS evaluation had a very close relationship to the traditional BVs, whereas the association was lower with the current GEBVs, regardless of blending. Comparing unproven sires BVs/GEBVs, the genomic information is expected to add information to the PA BV so the association would be lower than is the case for proven sires. Differences exist between the blended GEBVs of the current system and the GEBVs from the hybrid SS evaluation. Overall, the hybrid SS evaluations were closer to the PA BVs than were the GEBVs from the current system.

Reliability

Figure 1 shows the whisker plots for protein GEBVs obtained from the current GBLUP genomic evaluation and the hybrid SS evaluation. Data from bulls born in 2005 to 2012 are included. Bulls born in 2008 or earlier have daughter information while bulls born in 2009 or later have no daughter information. The reliabilities from the current system are higher than expected. Using a validation accuracy of 0.60 (current system) the predicted GEBV reliability should be approximately 0.61 (Mantysaari, 2010). The reliabilities obtained using the hybrid SS method are more in line with this expectation.

Conclusion

A hybrid single-step method of genomic evaluation has been developed and assessed. The system can accommodate genotypes from males and a large number of females. The system uses phenotypic records from genotyped animals as well as all their ancestors. The bias is controlled by the choice of scaling factor. The resulting GEBVs are more credible than those of the current system. The associated reliabilities are lower and in line with expectation.

References

Gianola, D. & 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. & Johnson, D.L. 1998. Approximate reliability of genetic evaluations under an animal model. *J. Dairy Sci. 81*, 2723-2728.

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.

Harris, B.L., Johnson, D.L. & Spelman, R.J. 2011. Choice of parameters for removal of inflation in genomic breeding values for dairy cattle. *AAABG 2011*, 359.

Harris, B.L., Winkelman, A.M. & Johnson, D.L. 2012. Large-scale single-step genomic evaluation for milk production traits. *Interbull Bulletin 46*, 20-24.

Makgahlela, M.L., Stranden, I., Nielsen, U.S., Sillanpaa, M.J. & Mantysaari, E.A. 2013. The estimation of genomic relationships using breedwise allele frequencies among animals in multibreed populations. *J. Dairy Sci. 96*, 5364-5375.

Mantysaari, E., Liu, Z. & VanRaden, P. 2010. Interbull validation test for genomic evaluations. *Interbull Bulletin 41*, 17-21.

Misztal, I., Aguilar, I., Johnson, D.L., Legarra, A., Tsuruta, S. & Lawlor, T.J. 2009. A unified approach to utilize phenotypic, full pedigree, and genomic information for a genetic evaluation of Holstein final score. *Interbull Bulletin 40*, 240-243.

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

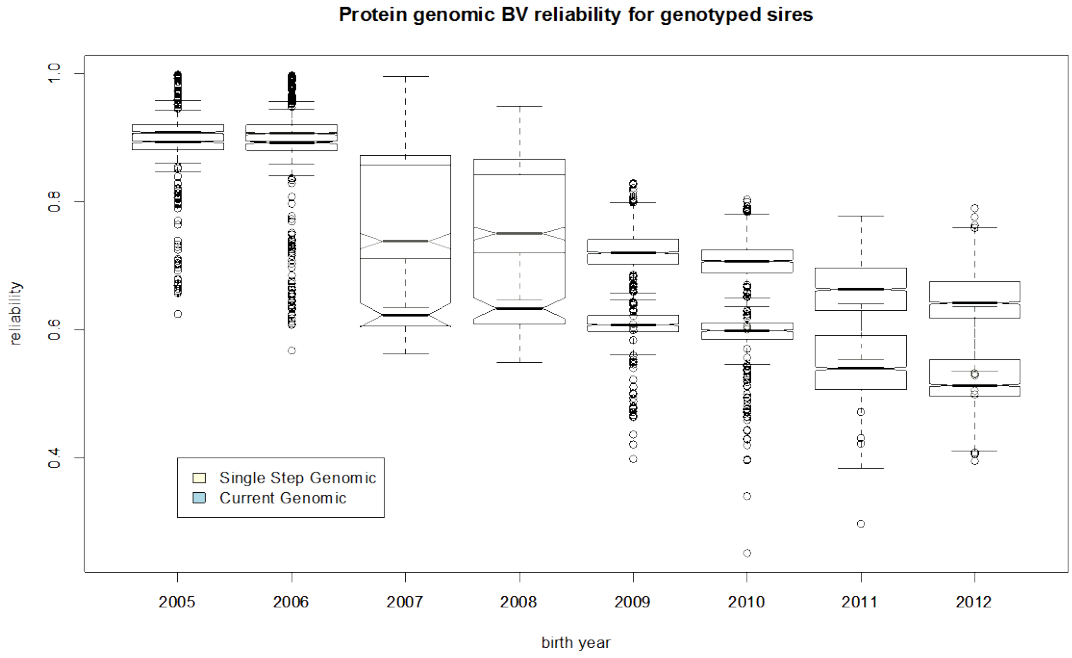

Figure 1. Whisker plots for protein GEBVs obtained from the current genomic GBLUP evaluation and the hybrid SS genomic evaluation.