Proceedings, 10th World Congress of Genetics Applied to Livestock Production

Across-breed genomic prediction in dairy cattle

B. L. Harris, A. M. Winkelman and D. L. Johnson
Livestock Improvement Corporation, Hamilton, New Zealand

ABSTRACT:
In most countries, genomic evaluations are obtained from within-breed analyses. New Zealand (NZ) is an exception that has been doing genomic evaluations using a multi-breed reference population, including purebred and crossbred individuals, since 2007. This paper summarizes across-breed genomic prediction experiences in NZ. A number of areas are discussed including, breed stratification found in the SNP data, the prediction of breed proportions from SNPs, methods for building a multi-breed genomic relationship matrix (GRM) and the accuracy of across-breed genomic prediction from different across breed GRMs. In the NZ population, accounting for the population structure in the GRMs had little effect on the validation accuracies and inflation measures from genomic selection analyses. It is suggested that prediction models that utilize breed-specific haplotype blocks based on high density SNP chips, along with a large training population, may improve genomic prediction in multi-breed populations.

Keywords: dairy cattle; genomics; crossbreeding

Introduction

Genomically enhanced breeding values are widely used for the selection of young dairy sires. In most countries, the genomic predictions are from within-breed analyses using a single-breed reference population. Unlike other countries, New Zealand (NZ) has purebred populations, as well as a large crossbred population, that have been evaluated using a multi-breed animal model since 1996. The incorporation of genomic data provides additional challenges in a multi-breed analysis. Genomic relationships are a function of allele frequencies, which may differ among breeds because of different origins and selection pressures. We will refer to differences in allele frequencies across breed as breed stratification. In pedigree-based genetic evaluations, the numerator relationship matrix (NRM) accounts for relatedness, but not breed effects. The breed effects are modeled using genetic groups. When combining pedigree and genomic relationships, we need to maintain the separation of breed and relatedness and therefore take account of the stratification when incorporating genomics.

The first genomically enhanced breeding values (GBVs) in NZ were calculated using a two-step method (vanRaden, 2008) in which direct genomic values were calculated using the breed-adjusted genomic relationship matrix of Harris and Johnson (2010).

In beef cattle, within-breed SNP estimation has shown to be of limited value for the genomic prediction of other breeds, (Saatchi and Garrick, 2013). In dairy cattle De Roos et al. (2009), Ibanez-Esciche et al. (2009), Pryce et al. (2011) and Schrooten et al. (2013) have shown that across-breed genomic evaluations are more accurate than within-breed evaluations. The closer the genetic distance between the breeds and the greater marker density, the greater the improvement in accuracy, assuming consistency of QTL and SNP marker phase and marker effect size.

Crossbreeding in NZ dairy industry has been steadily increasing since the early 1980s. The crosses are mainly between the Holstein Friesian and Jersey breeds. In 2013, the proportion of crossbreed heifer calves reared was 47%, compared to 37% and 9.3% for Holstein Friesians and Jerseys, respectively. In 2001 progeny tested crossbred sires became available to the NZ industry. This paper will focus on across-breed genomic prediction in NZ.

The objectives of this paper are to investigate:

1. Breed stratification in genomic data,
2. The prediction of breed proportions using genomic data,
3. Methods for building a multi-breed genomic relationship matrix (GRM), and
4. The accuracies of multi-breed genomic prediction from different GRMs.

Breed Stratification

Genotypes were obtained from the Illumina BovineSNP50 Beadchip panel. There were 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. Data on 16,437 sires containing three black and white strains, Jerseys (J) and their crosses were used to explore breed stratification. The black and white strains are categorized as Holsteins (HOL), Holstein Friesians (HF) and Friesians (FR) based on a decreasing fraction of recorded overseas ancestry. The black and white strains will collectively be referred to as the Holstein/Friesian Breeds (HFB). Breed categories were defined based on being at least 15/16ths of the given breed. The percentages for each breed category were 11% HOL, 2% FR, 34% HF 29% J and 24% FJ, the latter representing crosses between HFB and Jerseys. Only data from sires were used because of high certainty of recorded breed. The posterior mean of allele frequency was calculated within breed category, assuming a uniform prior on [0,1], as $p_j = (1 + \sum_i g_{ij})/(2 + 2N)$, where g_{ij} is j^{th} the genotype (coded 0, 1 or 2 indicating allele dosage) of the i^{th} animal and N is the number of animals.

The correlations between estimated allele frequencies are given in Table 1. The correlations among the HFB were higher than those between the HFB and J, reflecting the known breed differentiation. The lowest correlation was between J and HOL. Between-breed differences of the allele frequencies were calculated. Values of +1.0 or -1.0 indicates that alternate alleles are fixed in the two breeds. Figure 1 shows box plots for the distribution of these differences between J and HOL, FR and HOL, and FR and J. The largest differences were between J and HOL and smallest between FR and HOL, reflecting the genetic distances among the pure breed categories.

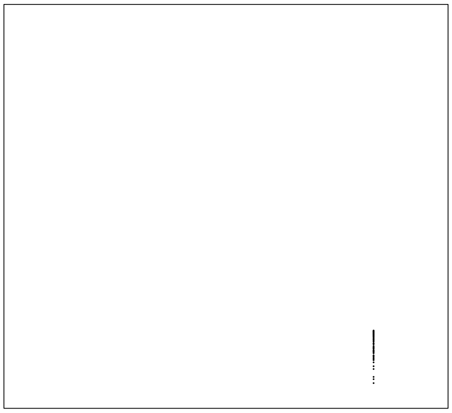

Figure 1: Distribution of estimated allele frequency differences between three purebred categories.

Table 1. The correlations between estimated allele frequencies for five breed categories.

Breed§	HOL	J	FJ	FR
J	0.65			
FJ	0.85	0.91		
FR	0.87	0.71	0.88	
HF	0.95	0.70	0.91	0.94

§HOL = Holstein, J = Jersey, FJ = HF x Jersey, HF = Holstein x FR, and FR = Friesian

Principal components analysis can be used to explore population stratification by identifying major axes of variation (Price et al., 2006). Let M be the normalized genotype matrix with elements $g_{ij}/\sqrt{p_j(1-\ p_j)}$, where g_{ij} is the j^{th} genotype of the i^{th} individual and p_j is the frequency of the j^{th} allele. Major axes of variation are described by the eigenvectors with the largest eigenvalues of the matrix MM^T. These axes of variation can describe the nonrandom breed and pedigree structure within the population (Daetwyler et al., 2012). Of the 16,437 eigenvalues, the largest 4659 explained significant ($P < 0.05$) amounts of variation based on the TW statistic (Patterson et al., 2006). The first two of these were significant at a level of $P < 10^{-12}$. These two principal components are shown in Figure 2. The first component differentiates the HFB strains from J, with the FJ being intermediate between the two breeds. The second component differentiates among the HFB strains, clearly distinguishing between the HOL and FR strains, with the HF positioned in between these two. However, each of the first six major axes of variation differed significantly ($P < 0.05$) between breed categories. The last four of these axes also explained variation caused by individual sires having large half-sib families within the population structure, similar to Daetwyler et al. (2012).

Figure 2: Two largest axes of variation of the sire SNP data.

Prediction of breed from SNPs

Multi-breed genomic evaluation requires accurate identification of breed composition of individual animals. In NZ, the breed composition of progeny test sires is known with the greatest degree of accuracy. Increasing numbers of female genotypes are being included in the reference population (Harris et al., 2013). For a large number of these genotyped females only the sire is genotyped and the possibilities of pedigree and breed composition errors exist.

Accurate prediction of breed composition based on SNP data has been demonstrated by Kuehn et al. (2011) and Frkonja et al., (2012) for admixed cattle populations. Frkonja et al., (2012) employed various prediction methods including Bayes B, LASSO and PLSR, and found correlations of 0.93-0.97 when predicting the proportions of Simmental and Red Holstein Friesian in Swiss Fleckvieh cattle. Kuehn et al., (2011) found accuracies ranging from 89 to 83% in beef breed crosses from genomic BLUP predictions of breed using the Bovine 50k and 3k SNP panels, respectively.

Genomic BLUP was used to predict the proportion of HF and HOL identified by pedigree records in the 16,437 sires using the selected Bovine 50k SNPs. A cross-validation study was undertaken where the training population was defined as a random half of the data and the remaining data was the test population. This process was repeated 100 times. The accuracy and the bias of prediction were calculated for each sample. The average accuracy of prediction was 0.99 for both HF and HOL proportion with an average bias of 0.009 and 0.012 (deviation from an expected regression of 1.0) for HF and HOL, respectively. Finally, genomic BLUP was used to calculate SNP solutions from the entire sire dataset and these solutions were applied to 64,946 cow genotypes. The correlation between the SNP-based breed prediction and the pedigree-record breed proportion was 0.94 and 0.88 for HF and HOL, respectively. Figure 3 illustrates the Mahalanobis distance plot between the SNP-predicted and pedigree-recorded HF proportion for cows. Observations above the red line in Figure 3 are viewed as outliers in terms of difference between the predictions based on SNP and pedigree HF proportion. The red line in Figure 3 is positioned at a Mahalanobis distance of 2.5, values above 2.5 are considered outliers. Approximately 3.9% of the cows would be considered as outlier observations. These cows may need to be excluded from the genomic evaluation because they would have incorrect genetic breed group solutions obtained from the traditional genetic evaluation. Another option would be retain the genotypes but replace the pedigree breed proportions with SNP based predictions.

Figure 3: Mahalanobis distance between SNP-predicted and pedigree-recorded HF breed proportion for 64,946 cows.

Estimation of multi-breed genomic relationship matrices

In dairy cattle, genomic relationships based on genome wide SNP data are used to estimate genomic breeding values (GBV), where the genomic relationships replace, or are combined with, the pedigree-based relationships. In some approaches, the calculation of the GRM requires estimates of the base-population SNP frequencies. This is straightforward in single breed populations but more complicated in multi-breed populations because SNP frequencies differ by breed and crossbred animals descend from more than one population. We discuss four methods of creating genomic relationships in a multi-breed population.

The first method (Harris and Johnson, 2010) is based on regressing the MM^T matrix on the expected value for the NRM, after taking into account the covariance between relatives in a multi-breed population. The covariances between relatives are calculated based on the methods outlined by Luo et al. (1993) using breed origins of known ancestors of the genotyped individuals. This method is computationally intensive, but feasible, for a relatively small population of genotyped animals (for example, all genotyped bulls within an evaluation). However, the required Cholesky decomposition is too computationally demanding for large genotyped populations, as would be the case when tens of thousands of cows are genotyped.

The second method (Makgahlela et al., 2013) is based on adjusting the genotypes using current or base population estimates of allele frequencies specific to breed type. This method was applied in the current study. The GRM was calculated as M*M^{*T}/m, where m is the total number of SNP and

$$M_{ij}^{*} = (g_{ij} - 2p_{ij})/\sqrt{p_{ij}(1-p_{ij})}$$

where p_{ij} is the j$^{\mathrm{th}}$ allele frequency corresponding to the breed composition of animal i. The p_{ij} values were estimated using linear regression. HF and J allele frequencies were estimated within the HF, J and FJ populations separately and also across all data. The concordance between the estimates from all data and the purebred data sets was high with correlations of 0.999 and standard deviations of the difference in estimates of 0.011 and 0.013 for HF and J estimates, respectively. Lower values were obtained if only FJ animals were used, with correlations of 0.991 and 0.986 and standard deviations of the difference in estimates of 0.04 and 0.051 for HF and J estimates, respectively. Figure 4 shows the distributions of Jersey allele frequency differences estimated from all the data (N=16,437) or 5351 FJ compared to 4509 Jersey sires. Although the correlations from the FJ data compared to the Jersey are high, there are substantial differences in estimated allele frequency for a number of SNP, with the extremes in the ± 0.2 range.

Figure 4: Distributions of Jersey allele frequency differences estimated from the all the sire data (n=16,437) or 5351 Jersey-Holstein Friesian compared to 4509 Jersey sires.

The third method used the Euclidean distance matrix (EDM) in a Gaussian kernel, as proposed by Gianola and van Kaam (2008). Harris et al. (2011) used the EDM in a multi-breed genomic evaluation. The advantage of using the EDM, instead of the GRM, is that the EDM does not require information on individual breed proportions or within-breed base allele frequencies. The Gaussian kernel transforms a distance measure into a correlation as

$\mathrm{Exp}(-d_{ij}/h)$, where d_{ij} is the Euclidean distance between individuals i and j, and h is bandwidth parameter. Thus, the larger the difference between individuals, the smaller the correlation, with zero distance corresponding to a correlation of unity. The EDM has the advantage of being positive definite even when the number of individuals is greater than the number of SNP, provided there are no identical twins in the SNP data. One disadvantage of the EDM is that the diagonal elements are unity regardless of the degree of inbreeding. Another disadvantage is that the EDM is not directly comparable to the NRM.

A fourth method to build a multi-breed GRM is based on the procedure outlined by Price et al. (2006). Genotypes are adjusted for major axes of variation that explain breed stratification in SNP data. The method adjusts each SNP by using the axes of variation as covariates in a multiple regression analysis. This can be calculated efficiently as $\bar{M} = M - EE^T M$, where the $\bar{M}$ is the adjusted SNP marker matrix, E is the matrix of eigenvectors and M is SNP marker matrix. This method is valid only when the eigenvectors are calculated on the same SNP data including all individuals otherwise a multiple regression SNP by SNP is required. The adjusted SNP marker matrix is used to calculate the GRM as $\bar{M}\bar{M}^T/m$. This method has the advantage that the breed information is removed without the use of the pedigree. The disadvantages of this method are that the calculation of the eigenvectors can be computational demanding large data sets and the assumption of two major axes may not remove all of the breed information but the use of more than two may remove pedigree structure as well as breed stratification.

The accuracy of across-breed genomic prediction from the use of different across breed GRMs

Evaluations were obtained for the different GRMs. The GRMs described in the previous section, except the GRM outlined Harris and Johnson (2010) that was computationally infeasible, were calculated using NZ male and female genomic data. Additionally a GRM that ignored breed stratification was calculated. The analyses included all of LIC's genotyped bulls born in 2007 or earlier, a selection of CRV Ambreed (another AI company in NZ) genotyped bulls and genotyped females born in 2006 or earlier. The phenotype for all analyses was the deregressed BV (DRBV) as described in Harris and Johnson (2010). The genomic analyses were run using the DRBV for protein yield that would have been available at the end of season 2008. Genomic BVs were from calculated from a hybrid single-step method (Harris et al. 2013). This method has the advantage of providing a computationally efficient genomic evaluation using traditional national DRBV, rather than phenotypic records, as the starting point. A preconditioned conjugate-gradient method is used to solve these equations after first using matrix inversion to calculate the inverse of the GRM and associated partitioned **A** matrix.

In NZ, a season starts in June and ends in May the next year. Hence, an animal calving in season 2008 finishes her lactation in 2009. 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. Validation followed the procedure of (Mantysaari, 2010). The accuracy of prediction was calculated as the correlation between the deregressed progeny test BVs (obtained using data available at the end of season 2013) 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. Table 2 shows the breed composition of the training (sires born prior to 2005) and test sire populations. Table 3 shows the year of birth and breed composition of the cows with genotypes. Prior to 2005, the female genotypes were available on bull dams. In 2005 and 2006, the genotypes were mainly from cows in specialized progeny test herds.

Table 2. The numbers of genotyped sire in the training and test data sets by breed.

Type	HF§	J§	FJ§
Training	3611	1816	367
Testing	685	391	334

§ J = Jersey, FJ = HF x Jersey, HF = Holstein x Friesian

Table 3. The numbers of genotyped cows in the training data set by year of birth and breed.

Birth Year	HF§	J§	FJ§
1990-2004	1773	347	222
2005	3237	3146	1533
2006	4096	4239	1887

§ J = Jersey, FJ = HF x Jersey, HF = Holstein x Friesian

Discussion

In this study accounting for the population structure in the GRMs has little effect on the validation accuracies and inflation measures. Similar results were found by Makgahlela et al., (2013) and Makgahlela et al., (2014). Thomasen et al., (2013) found that including breed information in a genomics study containing US and Danish Jersey cattle improved the genomic predictions. However, including population structure, derived from SNP data, in the prediction model did not improve reliabilities of the genomic predictions. Daetwyler et al., (2012) attempted to decompose the accuracy of genomic selection from population structure or linkage disequilibrium in a multi-breed sheep population. They found accounting for population structure via the use of eigenvectors from the genomic relationships decreased the accuracy of genomic prediction for most breeds. We also found lower genomic prediction accuracies when the GRM was adjusted by the first two eigenvectors to remove breed stratification. It is conceivable that adjusting by the GRM for major eigenvectors removes population structure other than breed stratification such as structure relating large to half sib families and within family selection. Previous results from genomic evaluation in the NZ, comparing models for evaluating a multi-breed population, suggested a genetic architecture of many QTL with small effects (Harris et al., 2011). The results suggest that a considerable proportion of the genomic prediction accuracy is due to population structure rather than LD between SNP markers and QTL. Removing population structure via eigenvector adjustment of SNP markers is likely to lead to decreased genomic prediction accuracy.

Genomic prediction has been applied successfully in single-breed dairy cattle populations such as the large Holstein populations in North America and Europe. Less success has been achieved in smaller populations and in admixed or multi-breed populations. This could be due to a combination of factors: insufficient training populations, larger effective population sizes and low levels of LD between SNP markers and QTL. For the NZ population, the genomic prediction accuracy obtained from the multi-breed analysis is higher that obtained by within-breed analysis (Harris et al., 2011) and higher than that obtained by parent average from traditional genetic evaluation. Similar results have been reported by Schrooten et al., (2013) for the Holland Genetics genomic program in NZ. The improvement in accuracy of multi-breed analysis compared to the within-breed analysis is likely to be function of the increased training population size in the multi-breed analysis. Harris et al., (2103) reported increases in the levels of genomic prediction accuracy as large numbers of females were added to the NZ training population, suggesting that the training population size in NZ is still suboptimal for multi-breed genomic prediction.

There may be genetic mechanisms that account for lower genomic prediction accuracy in multi-breed analyses than in single breed analyses. The current multi-breed genomic models assume homogeneity of QTL and SNP marker phase and marker effect size across breeds and between the training and test populations. The greater the genetic distance between breeds the less likely this assumption is maintained due to genotype-by-genetic background interactions (de Roos et al., 2009). Also, Deng (2001) has suggested that population admixture can conceal the underlying QTL effects. It was thought that increased SNP marker density was required to adequately model LD across breeds. Multi-breed analyses undertaken by Su et al. (2011) and Harris et al. (2011) comparing 50K SNP panels with 800K SNP panels showed little improvement in accuracy. However, it is questionable whether the training population sizes used in these studies had sufficient statistical power to exploit the increase in marker density.

The majority of multi-breed genomic prediction studies have used GRMs or EDMs based on arrays of individual SNPs. The use of high density SNP data to form haplotype blocks for use in a multi-breed genomic analysis could improve the accuracy and remove the need for homogeneity of QTL and SNP marker phase and marker effect size. If the SNP marker density is sufficient to produce breed-specific haplotype blocks that are associated with the QTL alleles segregating within a given breed, then a greater proportion of the genetic variance will be explained by the haplotype blocks within a multi-breed model.

Conclusion

Multi-breed genomic prediction is a challenging area of research. The results to date have lower accuracy compared to then those achieved from large within-breed genomic analyses, particularly in the Holstein breed. The breed stratification in the admixed populations can be categorized by examination of the eigenvectors on the relationship matrices. The breed content for individual animals can be predicted with a high degree of accuracy from SNP data using simple GBLUP analyses. Adjusting GRMs to take account of breed information or the use of EDMs in multi-breed genomic prediction models provides only small improvement in accuracy or reduction of inflation in the resulting GBVs compared to ignoring breed.

A number of factors could potentially result in the lower accuracies: small training population size, large effective population sizes for admixed populations, low levels of LD between SNP markers and QTL making the implicit assumption of homogeneity of QTL and SNP marker phase and marker effect sizes. Prediction models that utilize breed-specific haplotype blocks based on high density SNP chips and the inclusion of large numbers of females in the training populations may improve the prediction accuracies by mitigating the above factors in the analyses.

Table 4. The accuracy of the genomic breeding values for different genomic relationship matrices (GRM).

Breed‡	IGN§	MAK§	EDM§	EVA§
HF	0.60	0.63	0.62	0.55
FJ	0.62	0.64	0.64	0.50
Jersey	0.70	0.73	0.73	0.55

§ [1]ING = GRM ignoring breed, MAK = multibreed GRM (Makgahlela et al. (2013), EDM = Euclidean distance matrix, EVA = eigenvector adjusted GRM

‡ J = Jersey, FJ = HF x Jersey, HF = Holstein x Friesian

Table 5. The bias of the genomic breeding values for different genomic relationship matrices (GRM).

Breed‡	IGN§	MAK§	EDM§	EVA§
HF	0.96	1.01	0.99	0.77
FJ	1.03	1.04	0.99	0.63
Jersey	1.09	1.08	1.06	0.72

§ [1]ING = GRM ignoring breed, MAK = multibreed GRM (Makgahlela et al. (2013), EDM = Euclidean distance matrix, EVA = eigenvector adjusted GRM

‡ J = Jersey, FJ = HF x Jersey, HF = Holstein x Friesian

Literature Cited

Daetwyler, H. D., Kemper, K. E., van der Werf, J. H. J. et al. (2012). J. Anim. Sci. 90:3375-3384.

Deng, H. W. (2001) Genetics 159:1319-1323

De Roos, H. D., Kemper, K. E., van der Werf, J. H. J. et al. (2009). Genetics. 183:1545-1553.

Frkonja, A. B., Gredler U., Schnyder, I. et al. (2012). J. Anim. Genet. 43:696-703.

Gianola, D. .G and van Kaam, J. B. C. H. M. (2008). Genetics 178: 2289–2303.

Gengler N., Mayeres P. and Szydlowski M. (2007). Animal 1:21-28

Harris, B. L., Johnson, D. L., and Spelman R. J. (2008). Proc. ICAR 36th Session. 325-330.

Harris, B. L. and Johnson, D. L. (2010). J. Dairy Sci. 93:1243-1252

Harris, B. L., Creagh, F., Winkelman A. M., et al., (2011). Interbull Bull 44:3-7.

Harris, B. L., Winkelman A. M. and Johnson, D. L. (2013). Interbull Bull 47:23-27.

Ibanez-Esriche N., Fernando, R. L., Toosi A., et al., (2009). Genet. Evol. Sel. 41:12-23

Kuehn L. A., Keele J. W., Bennett G. L., et al., (2011) J. Anim. Sci. 89:1742-1750.

Lo, L. L., R. L. Fernando, and M. Grossman. (1993). Theor. Appl. Genet. 87:423–430.

Mantysaari, E., Liu, Z. and VanRaden, P. (2010) Interbull Bull. 41:17-21.

Makgahlela M. L., Strandén I., Nielsen U. S., et al., (2013). J. Dairy Sci. 96:5364-5375.

Makgahlela M. L., Strandén I., Nielsen U. S., et al., (2014). J. Dairy Sci. 97:1117-1127.

Su G., Brømdum R. F., Ma P., et al., (2011). J. Dairy Sci. 95:4657-4665.

Price A. L., Patterson N, Plenge R. M., et al., (2006). Nat Genet. 38:904-909.

Patterson N., Price A. L. and Reich D. (2006). PLoS Genetics 2:2074-2093

Pryce J. E., Gredler B., Bolorman S., et al., (2011). J. Dairy Sci. 94:2625-2630.

Saatchi M. and Garric D. J. (2013). Proc. Assoc. Advmt. Anim. Breed. Genet. 20:207-210.

Schrooten C., Schopen G. C. B., Parker A., et al., (2013). Proc. Assoc. Advmt. Anim. Breed. Genet. 20:138-141.

Thomasen J. R., Sorensen A. C., Su G., et al., (2013). J. Anim. Sci. 91:3105-3112

vanRaden P. R. (2008) J. Dairy Sci. 91:4414-4423.