Genomic evaluations for crossbred dairy cattle*

B. L. Harrist†

Graphical Abstract

Summary

Genomic evaluation in crossbred dairy cattle populations combines pedigree, phenotypic, SNP marker, and breed information into a single analysis. The paper covers a recent large-scale single step marker model implementation in New Zealand. Brief details outlining the solving methods, account for multiple breed and breed crosses, and the data sizes are discussed.

Highlights

- Genomic evaluations for crossbred dairy cattle from a New Zealand perspective are discussed.
- This review provides a short history of the across-breed genetic or genomic evaluations used in New Zealand since 1996.
- Implementation details for the current across-breed single-step marker model are outlined.

*Presented as part of the Breeding and Genetics Symposium: Crossbreeding at the ADSA Annual Meeting, July 2021. Livestock Improvement Corporation Ltd., Private Bag 3016, Hamilton 3140, New Zealand. †Corresponding author: bevin.harris@lic.co.nz. © 2022, The Authors. Published by Elsevier Inc. and FASS Inc. on behalf of the American Dairy Science Association®. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>). Received October 07, 2021. Accepted December 01, 2021.

Genomic evaluations for crossbred dairy cattle*

B. L. Harrist†

Abstract: This short review discusses genomic evaluations for crossbreed dairy cattle from a New Zealand perspective. I briefly introduce the cow population and farm systems that have resulted in high crossbreeding rates and provide a short history of across-breed genetic and genomic evaluations used in New Zealand since 1996. I provide implementation details for the current across-breed single-step marker model (SSMM), including data sizes, solving methods, and how to account for multiple breed and breed crosses, along with a summary of the validation results. Finally, I outline future research areas that are being undertaken to improve our across-breed genomic evaluation.

Dairy farmers in New Zealand (NZ) have been exploiting the benefits of milking crossbred cattle for more than 30 yr. Almost all crossbred cows are derived from the Jersey (J) and Holstein-Friesian (HF) breeds. From the 1999 season to the 2019 season, the percentage of crossbred cows in the population increased from 21 to 49%. Correspondingly, the percentage of purebred J and HF cows has dropped from 15 and 56% to 8 and 33%, respectively (Livestock Improvement Corp./Dairy NZ, 2020). In response to the increasing numbers of crossbred cows in the population, crossbred sires have been available for AI since 2006.

Most NZ dairy farms are managed using a predominantly pasture-based system, with all cows calving in spring. The herd is required to calve within 365 d to maintain the seasonal calving system. This system favors cows of moderate size and good fertility. The combination of better fertility and productivity has been the primary driving force for the increased adoption of crossbred cows, which is evident in the numbers given in Table 1. Table 1 provides a simple measure of productivity (kg of milk solids per kg of live weight) for crossbred and purebred cows. The hybrid vigor estimates from the July 2020 national genetic evaluation for first-lactation milk fat and protein yields and mature live weight were 16.5, 10.3, and 11.7 kg for the Jersey-Holstein and 9.4, 6.3, and 9.1 kg for the Jersey-Friesian components, respectively. The crossbred cows have the highest productivity in part due to hybrid vigor. Dairy farmers in NZ have been provided with a Production Worth index for each cow in their herd. This index provides an average lifetime net profit per unit of feed that includes hybrid vigor for crossbred cows.

Genomic SNP data allow insights into the genetic background of both the pure and crossbred populations. There are 2 ways of illustrating the genetic differences between the breeds. The first is to explore population stratification using the Admixture software (Alexander et al., 2009). Data from 120K HF, J, and J-HF crossbred animals genotyped on the 50K Illumina SNP chip were analyzed using this software, with the number of categories set to 3. There is a striking difference between the HF and J purebred animals. The J animals show a high degree of genomic uniformity, whereas the HF purebred animals are an admixture of Holstein and Friesian strains with a small amount of Jersey. The Jersey background in the

HF animals is due to conversion of J herds to HF herds in the 1960s and early 1970s. Figure 1 shows a scatterplot of the breed classes using the first 2 principal components from the SNP chip data on 23K sires. This scatterplot illustrates a large amount of breed structure in the genomic data. The figure also shows the genetic distances among the breeds and the large degree of genomic variability among the HF-J crossbred sires. The magnitude admixture, the breed structure in the NZ data, and the shared pedigree links across the breeds and crosses justify the need for an across-breed genomic evaluation where one set of SNP effects is estimated for all breeds and breed crosses.

Before undertaking the first across-breed genomic evaluations in 2008, NZ had undertaken across-breed genetic evaluations since 1996. In 1996, the across-breed evaluation was based on an animal model with genetic grouping that included breed proportions. The models also had effects for hybrid vigor (Harris et al., 1996). The traits evaluated included milk production, live weight, survival, and 15 linear type traits. In 2006, milk production traits and SCS were evaluated using a test-day random regression model. Fertility and BCS traits were included in the mid-2000s.

The first across-breed genomic evaluations in 2008 were based on a 2-step genomic BLUP (GBLUP; Harris and Johnson, 2010), which had a low prediction accuracy and upward bias from overfitting the SNP effects and a small reference population. The reference population contained only approximately 5,000 J, HF, and J-HF crossbred sires. Daughter yield deviations (DYD) from the traditional genetic evaluation were inputs to the models.

By 2015, there were close to 100K genotyped individuals in the reference population. The reference population included approximately 75K cows. Complete age cohorts of cows were selected from herds for genotyping that extensively used young nonproven sires for breeding. The inclusion of cow genotypes in the reference improved the prediction accuracy, but the magnitude of the increase was considerably lower than when adding 75K progeny-tested sire genotypes. A new genomic evaluation system was implemented in 2015. The method used a simplified single-step approach (Winkelman et al., 2015). The simplified approach refers to only the inclusion of genotype individuals and their ancestors in a single-step GBLUP framework (Miszta et al., 2009) rather

*Presented as part of the Breeding and Genetics Symposium: Crossbreeding at the ADSA Annual Meeting, July 2021. Livestock Improvement Corporation Ltd., Private Bag 3016, Hamilton 3140, New Zealand. †Corresponding author: bevin.harris@lic.co.nz. © 2022, The Authors. Published by Elsevier Inc. and FASS Inc. on behalf of the American Dairy Science Association®. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>). Received October 07, 2021. Accepted December 01, 2021.

Table 1. A simple measure of productivity (kg of milk solids per kg of live weight) for crossbred and purebred cows

Breed	Average milk fat (kg/cow)	Average protein (kg/cow)	Average milk solids (kg/cow)	Mature live weight (kg)	kg of milk solids/kg of live weight
HF-Jersey cross	202	164	366	500	0.73
Holstein-Friesian (HF)	199	169	368	550	0.67
Jersey	178	134	311	450	0.69

than the entire population. The genomic evaluations were dropped down through the pedigree from the genotyped individuals to nongenotyped progeny in descending order of birth year so that the nongenotyped progeny received genomic evaluations. As with the previous genomic evaluation, the traditional genetic evaluation DYDs were used as inputs to this model. It became apparent with the increasing numbers of genotyped animals entering the genomic evaluation each year that the simplified single-step approach would become computationally intractable due to the method requiring the inverse of the genomic relationship matrix. The genomic relationship matrix used in this model was based on Euclidean distance in a Gaussian kernel, which was not easily extendable to the more computationally efficient APY (algorithm for proven and young) algorithms (Misztal, 2016) developed for single-step GBLUP, which reduces the size of inverse of the genomic relationship matrix. There was also a concern that any genomic preselection bias in the traditional genetic evaluation could result in biased DYDs. To overcome the computational limitations and the potential bias from using DYDs, research into using a single-step-marker model (SSMM; Fernando et al., 2016) began in late 2017. The SSMM was chosen over the single-step GBLUP approaches because the dimensions of the SSMM equations are based on the number of SNP markers rather than the number of genotyped animals. The

number of genotyped animals continues to increase by approximately 25K/yr. The SSMM was implemented in February 2020.

Our current SSMM genomic evaluation includes 32 million animals. Most animals are HF, J, or HF-J crosses, with Ayrshire being the next largest breed group. Close to 200K genotypes are included in the analysis. Sire genotypes make up 5% of the genotypes, and 38 phenotypes are evaluated.

The SSMM generic model statement for all traits is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}_g\mathbf{M}_g\mathbf{m} + \mathbf{Z}_n\mathbf{u}_n + \mathbf{Z}_a + \mathbf{Z}_p + \mathbf{e},$$

where $\mathbf{y}$ is the phenotype vector; $\mathbf{X}$, $\mathbf{Z}_g$, $\mathbf{Z}_n$, and $\mathbf{Z}$ are incidence matrices; $\mathbf{M}_g$ is the SNP marker matrix, where g refers to genotyped animals; $\mathbf{m}$ is the vector of SNP marker effects with each column centered to have mean of zero; $\mathbf{b}$ is the vector of fixed effects including hybrid vigor, genetic groups, and breed covariates; $\mathbf{u}$ is the vector of genomic breeding values for nongenotyped animals, and n refers to nongenotyped animals; $\mathbf{a}$ is the vector of polygenic effects; $\mathbf{p}$ is the vector of permanent effects; and $\mathbf{e}$ is the random residual effect. Only traits with repeated records fitted the permanent effects. The model includes the polygenic effects because we assume that the SNP do not capture all the genetic variance. Three significant features of SSMM mixed model equations

Figure 1. A scatterplot of the breed classes using the first 2 principal components from the 50K SNP data on 23K sires. HF = Holstein-Friesian.

(MME) deserve mentioning. First, the size of genomic components in the MME is defined by the number of SNP markers, not the number of genotyped animals. Therefore, large increases in the numbers of genotyped animals do not increase the computational burden. Second, MME are much denser than traditional genetic evaluation MME, particularly in areas containing the SNP marker matrix, which can affect the convergence of MME. Last, the MME requires the inverse numerator relationship matrix for nongenotyped animals, containing tens of millions of rows. This inverse can be done efficiently (Fernando et al., 2016) using sparse Cholesky decomposition and is suitable for large-scale parallel computing implementations. This component is $\mathbf{M}_g' \mathbf{A}^{gm} (\mathbf{A}^{nm})^{-1} \mathbf{A}^{ng} \mathbf{M}_g$, precalculated before solving the MME equations, where $\mathbf{A}^{ij}$ is the i th and j th sub-block of inverse numerator relationship matrix. The precalculation takes approximately 3 h for 31 million animals using 148 CPU cores.

Fernando et al. (2014) describe the inclusion of a J equation to account for arbitrary centering of the marker effects. The arbitrary centering of the marker effects results from the situation where a subset of individuals in the pedigree are genotyped and the SNP marker equations can be centered at different locations relative to the pedigree equations. In multi-breed settings, a J equation is fitted for each breed. Genotyped crossbred animals would have multiple J values corresponding to the breed group proportions present in that animal. The product of J covariates and the estimated J effects are included in the estimated genomic breeding value. A limitation of this approach is that it requires all animals to be fully connected to the pedigree. For animals that are not connected or are only partly connected to the pedigree, the J coefficients will not sum to 1 and the estimated genomic breeding values will be biased toward zero. The bias can occur for animals with missing parents or minor breeds with few or no genotyped individuals. The bias was overcome by replacing the J equations with breed covariates. The breed covariates were calculated by dropping the breed fractions down through the pedigree from the oldest to youngest animals. The progeny always had the sum of one half of the sire and dam breeds to ensure that the covariates were consistent across generations. This approach is identical to that of Fernando et al. (2014) if all nongenotyped animals are fully connected to the pedigree. Covariates for hybrid vigor for 7 breed group combinations were fitted as fixed effects. The inclusion of covariates for recombination losses in advance breed crosses was also explored. However, the estimates for the recombination loss covariates were small in magnitude and often not biologically meaningful for most of the traits evaluated. Consequently, the recombination loss covariates were removed from the models. Genetic group covariates were fitted for animals with missing parents. The genetic groups were assigned by sex of missing parent and birth year (in 5-yr blocks) using the approach of Westell et al. (1988). The genetic groups were fitted directly as fixed effects and not via the numerator relationship matrix.

A standard preconditioned conjugate gradient iterative solver with a diagonal preconditioner was used to solve the SSMM MME. This approach produced an unstable and extremely slow convergence pattern, partly due to the size and complexity of the equations, especially those associated with the marker effects that form dense blocks. Using a block diagonal preconditioner was found to solve these convergence issues. The blocks in the preconditioner

Table 2. Bias and accuracy of prediction for the 3-yr cohorts of test sires from the validation regression model for breeding worth (BRW) index

Breed	Genomic BRW		Pedigree BRW	
	Bias	Accuracy	Bias	Accuracy
Holstein-Friesian (HF)	0.90	0.73	1.08	0.42
HF × Jersey cross	0.96	0.61	0.79	0.27
Jersey	1.0	0.74	0.91	0.33

were associated with the major factors fitted in the model: fixed effects, SNP marker solutions, nongenotyped animal genomic breeding values, polygenic breeding values, and permanent environmental effects. The blocks contained the matrix inverses for each block. The strategies to obtain the matrix inverses involved Cholesky decomposition for the breeding value blocks, and direct inverses for the other blocks with equation absorption were necessary. Similar approaches are used to solve large-scale engineering and operations research problems (Kardani et al., 2013).

Before the implementation of the SSMM system, a validation study was undertaken. The data for model validation consisted of genotyped bulls born before 2010 and genotyped females born before 2012. The genomic analyses were run using the phenotypic data that would have been available at the end of season 2013. Four cohorts of young sires born from seasons 2010 to 2013 were the test population. The method used has been described in detail by Mäntysaari et al. (2011). We distinguished between the 3 breed classes (HF, J, and HF-J) to avoid having biased results from an across-breed validation analysis, particularly when the breed means for the traits are different. The difference in breed means will positively contribute to the accuracy and inflation measures. The validation accuracies of the genomic breeding values averaged 22, 29, and 29% greater than those from pedigree-based breeding values for the HF, HF-J and J sires, respectively. The inflation measures were also closer to unity for the genomic breeding values compared with the pedigree-based breeding values. The accuracies were also higher and the inflation measures closer to 1 than those calculated for the previous genomic selection system (Winkelman et al., 2015).

A second validation was done in 2020 using the national breeding objective Breeding Worth (BRW), which is an economic efficiency measure, net profit per unit of feed, used across all breeds. Farmers are more concerned about the BRW performance of genomics than the performance of individual traits. Pedigree-based and genomic BRW predictions were run using data up to the end of the 2016 season, and validation used 3-yr cohorts of young sires. The predictions were compared with the BRW results based on progeny data at the end of the 2020 season. The genomic prediction of BRW outperformed the pedigree-based prediction in terms of prediction accuracy and bias for all 3 breed classes (Table 2).

We are currently researching 3 areas for future enhancements. The first is the development and implementation of across-breed multiple-trait SSMMs. Prototype models for milk production traits, fertility, and SCS are being tested. For these traits, each lactation is treated as a separate trait. The fertility model is scheduled for implementation in December 2021. The primary research effort has been building efficient equation solvers to ensure that models converge quickly and optimizing the computational workload to allow the models to run in a reasonable time. Recently, a small

number of dominance effects with large phenotypic impacts have been discovered in the New Zealand dairy cattle population (Reynolds et al., 2021). The second area is the inclusion of a limited number of dominance effects in the SSMM. The central focus of the work is how to model dominance effects in genotyped and non-genotyped individuals within an SSMM. Finally, a collaborative project to identify breed-specific haplotypes surrounding QTL is underway. This research utilizes the 170K HF, J, and HF-J animals imputed to sequence and the NZ Jersey and NZ Holstein Friesian bovine maps. The aim is to include breed-specific haplotypes in future SSMM analyses.

References

- Alexander, D. H., J. Novembre, and K. Lange. 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* 19:1655–1664. <https://doi.org/10.1101/gr.094052.109>.
- Fernando, R. L., H. Cheng, B. L. Golden, and D. J. Garrick. 2016. Computational strategies for alternative single-step Bayesian regression models with large numbers of genotyped and non-genotyped animals. *Genet. Sel. Evol.* 48:96.
- Fernando, R. L., J. C. Dekkers, and D. J. Garrick. 2014. A class of Bayesian methods to combine large numbers of genotyped and non-genotyped animals for whole-genome analyses. *Genet. Sel. Evol.* 46:50–62.
- Harris, B. L., J. M. Clark, and R. G. Jackson. 1996. Across-breed evaluation of dairy cattle. *Proc. N.Z. Soc. Anim. Prod.* 56:12–15.
- Harris, B. L., and D. L. Johnson. 2010. Genomic predictions for New Zealand dairy bulls and integration with national genetic evaluation. *J. Dairy Sci.* 93:1243–1252. <https://doi.org/10.3168/jds.2009-2619>.
- Kardani, O., A. V. Lyamin, and K. Krabbenhøft. 2013. A comparative study of pre-conditioning techniques for large sparse systems arising in finite element limit analysis. *Int. J. Appl. Math. (Sofia)* 43:195–203.
- Livestock Improvement Corp. Ltd/DairyNZ Ltd. 2020. New Zealand Dairy Statistics 2019–2020. Accessed June 1, 2021. <https://www.lic.co.nz/about/dairy-statistics/>.
- Mäntysaari, E., Z. Liu, and P. M. VanRaden. 2011. Interbull validation test for genomic evaluations. *Interbull Bull.* 41:17–21.
- Misztal, I. 2016. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. *Genetics* 202:401–409.
- 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. <https://doi.org/10.3168/jds.2009-2064>.
- Reynolds, E. G. M., C. Neeley, T. J. Lopdell, M. Keehan, K. Dittmer, C. S. Harland, C. Couldrey, T. J. J. Johnson, K. Tiplady, G. Worth, M. Walker, S. R. Davis, R. G. Sherlock, K. Carnie, B. L. Harris, C. Charlier, M. Georges, R. J. Spelman, D. J. Garrick, and M. D. Littlejohn. 2021. Non-additive association analysis using proxy phenotypes identifies novel cattle syndromes. *Nat. Genet.* 53:949–954. <https://doi.org/10.1038/s41588-021-00872-5>.
- Westell, R. A., R. L. Quaas, and L. D. VanVleck. 1988. Genetic groups in an animal model. *J. Dairy Sci.* 71:1310–1318. [https://doi.org/10.3168/jds.S0022-0302\(88\)79688-5](https://doi.org/10.3168/jds.S0022-0302(88)79688-5).
- Winkelman, A. M., D. L. Johnson, and B. L. Harris. 2015. Application of genomic evaluation to dairy cattle in New Zealand. *J. Dairy Sci.* 98:659–675. <https://doi.org/10.3168/jds.2014-8560>.

Notes

This study received financial support from the NZ Ministry of Primary Industries, SFF Futures Programme: Resilient Dairy–Innovative breeding for a sustainable dairy future (Grant number PGP06-17006).

The author has not stated any conflicts of interest.