INCREASING THE ACCURACY OF GENOMIC PREDICTION IN CROSSBRED DAIRY CATTLE

M. Khansefid[1], M.E. Goddard[1,2], M. Haile-Mariam[1], C. Schrooten[3], G. de Jong[3], E. O'Connor[4], J.E. Pryce[1,5], H.D. Daetwyler[1,5] and I.M. MacLeod[1]

[1]AgriBio Centre for AgriBioscience, Agriculture Victoria, Bundoora, VIC, 3083 Australia
[2]Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Parkville, VIC, 3010 Australia
[3]CRV, 6800 AL Arnhem, the Netherlands
[4]CRV Ambreed, Hamilton, 3216 New Zealand
[5]School of Applied Systems Biology, La Trobe University, Bundoora, VIC, 3083 Australia

SUMMARY

This study assessed the accuracy of genomic prediction for crossbred dairy cows (mixed crosses between Holstein and Jersey) when purebreds and crossbreds were combined in a single mixed-breed reference set. The reference population consisted of 36,695 bulls and cows. There were six validation breed groups including crossbred cows (from New Zealand and Australia) as well as purebred cows (Holstein or Jersey cows from New Zealand). The effect of using genotypes of different marker densities (50K or HD) and different analytical models (GBLUP or emBayesR) on the accuracy and bias of genomic predictions was studied. The results showed that on average for milk traits (milk, fat and protein yields), the accuracies increased using HD genotypes compared to 50K genotypes, regardless of the prediction model. However, emBayesR outperformed GBLUP in all validation populations with the highest increase observed for Australian crossbreds when HD genotypes were used. Additionally, the bias of genomic prediction was reduced when using HD compared to 50K genotypes in both GBLUP and emBayesR models.

INTRODUCTION

Genomic prediction (GP) within breeds is generally very accurate using the standard 50K SNP panel when the linkage disequilibrium (LD) between markers and the causal mutations is preserved over long distances (e.g. using 50K in a purebred Holstein reference to predict into young Holstein bulls). However, using a single breed reference population for GP in crossbreds generally has low reliability, similar to the low accuracy for across breed prediction (e.g. Kemper *et al.* 2015). This is likely due in part to the fact that LD decays faster in crossbreds compared to purebreds so that markers that accurately predict QTL effects in Holsteins may not always be in LD with the same causal mutation allele in the crossbred. This is particularly the case when crossbreeding occurs over several generations as is common in the New Zealand dairy industry (New Zealand Dairy Statistics, www.dairynz.co.nz/dairystatistics).

The use of a mixed breed reference population to increase the reference population size, can potentially increase the accuracy of GP if the markers segregating across breeds have the same LD phase with the causal mutation alleles (Kemper *et al.* 2015). Moreover, inclusion of crossbreds in the reference population should also help to find the most predictive markers closest to causal mutations because LD would be preserved over shorter distances. This also helps to limit the number of multiple SNPs in high LD with QTLs.

Therefore, the aim of this study was to increase the accuracy of GP in crossbred dairy cattle using a mixed breed and crossbred reference population, increasing the density of markers (HD versus 50K) and using models which tend to calculate individual SNP effects (Bayesian) rather than haplotype effects (GBLUP). The dairy crossbreds were cows that had varying proportions of Holstein and Jersey (approximately 50%:50% cross = "HJ", approximately 75% Holstein = "HHJ" and approximately 75% Jersey = "HJJ"). The accuracy of GP in the crossbreds was also compared with purebred Holstein "H" and Jersey "J".

MATERIALS AND METHODS

Animals. The reference set consisted of 7,463 purebred bulls mainly from New Zealand and the Netherlands (953 Red H, 5,409 H and 1,101 J) as well as 29,232 purebred and crossbred cows from New Zealand (NZ) (7,623 H, 9,262 HHJ, 7,807 HJ, 1,157 HJJ and 3,383 J). There were five NZ cow validation populations: 1,002 H, 863 HHJ, 868 HJ, 324 HJJ and 532 J. An Australian (AU) cow validation set of 344 HJ was included to demonstrate GP in a less related group.

Relatedness between validation and reference. The validation sets were selected to reduce high relationships with the reference: no sires or half-sib brothers of validation cows were included in the reference. It has been demonstrated that the strength of the top 10 genomic relationships between validation and reference animals (Rel.$_{\text{Top10}}$) gives a good indicator of the relative accuracy of GP (Clark *et al.* 2012). Therefore, this is reported for each validation population.

Phenotypes. Milk, fat, and protein yields were analysed separately but the results are reported as the average across three traits. The phenotypes for CRV bulls were de-regressed proofs (DRP) on the Australian scale, derived from international MACE (2018) breeding values (Liu 2009). The NZ and AU cow phenotypes were also DRP which were processed together by DataGene (2018) using test day records and correcting for known fixed effects as for the official Australian dairy cattle evaluations (https://datagene.com.au/).

Genotypes. Two sets of imputed genotypes were used in GP: the standard Illumina 50K SNP panel (40,850 SNP) and Illumina HD 800k SNP panel (633,375 SNP), where the latter included an additional custom set of ~ 1200 variants. In the HD genotype set, one of each pair of SNP in LD $r^2 > 0.95$ was pruned out leaving 316,396 SNP. The majority of genotypes were first imputed from low density chips (~ 10k SNPs) up to 50K and then imputed from 50K to HD using FImpute (Sargolzaei *et al.* 2014). The SNPs with minor allele frequency (MAF) < 0.002 were removed.

Models. The GBLUP (Meuwissen *et al.* 2001) analysis used the following model (with MTG2 software: Lee and Van der Werf 2016):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \qquad (1)$$

where, **y** is the vector of phenotypes (MY, PY or FY DRP) for the animals in the reference, **X** is a design matrix allocating phenotypes to fixed effects (sex and breed group), **b** is the vector of fixed effect solutions, **Z** is a design matrix allocating records to individual additive genetic values in **u**, $\mathbf{u} \sim N(0, \mathbf{G}\sigma^2_g)$ is a vector of genomic breeding values (GEBVs) in which σ^2_g is the additive genetic variance and **G** is the GRM constructed from animal genotypes (50K or pruned HD), and $\mathbf{e} \sim (0, \mathbf{E}\sigma^2_e\mathbf{)}$ is a vector of random residual effects in which σ^2_e is the error variance and **E** is a diagonal matrix constructed as diag($1/w_i$) where w_i is the weighting coefficient for each animal. Weighting coefficients were calculated differently for cows and bulls following Equation 5 and 6 of Garrick *et al.* (2009), with heritability h^2=0.33, repeatability t=0.56 and proportion of variance not explained by markers is c=0.2.

We also analysed the data with "emBayesR" (Wang *et al.* 2016: in-house software):

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wv} + \mathbf{e} \qquad (2)$$

where, **y**, **X**, **b** and **e** are as for equation 1, **v** is the vector of SNP effects (50K or pruned HD), **W** is a design matrix of SNP marker genotypes (50K or pruned HD). In emBayesR model, the initial EM (Expectation-Maximisation) phase was set for a maximum of 1,500 iteration with the convergence parameter set as 1×10^{-7} and the BayesR phase was set to complete 5,000 iterations. For each trait, the emBayesR model was run in 5 independent replicated analyses (MCMC chains) to check for convergence and the results were averaged across the 5 chains. The accuracy of GP for each validation breed group was defined as the Pearson's correlation coefficient between GEBVs and DRPs ($r_{GEBV,DRP}$). The bias of GP was assessed by calculating the regression coefficient of DRP on GEBVs ($b_{DRP,GEBV}$) (no bias $b_{DRP,GEBV} = 1$).

RESULTS AND DISCUSSION

The accuracy and bias of GEBVs in each of the six validation breed groups are shown in Figure 1, where the values are averaged across MY, FY, and PY.

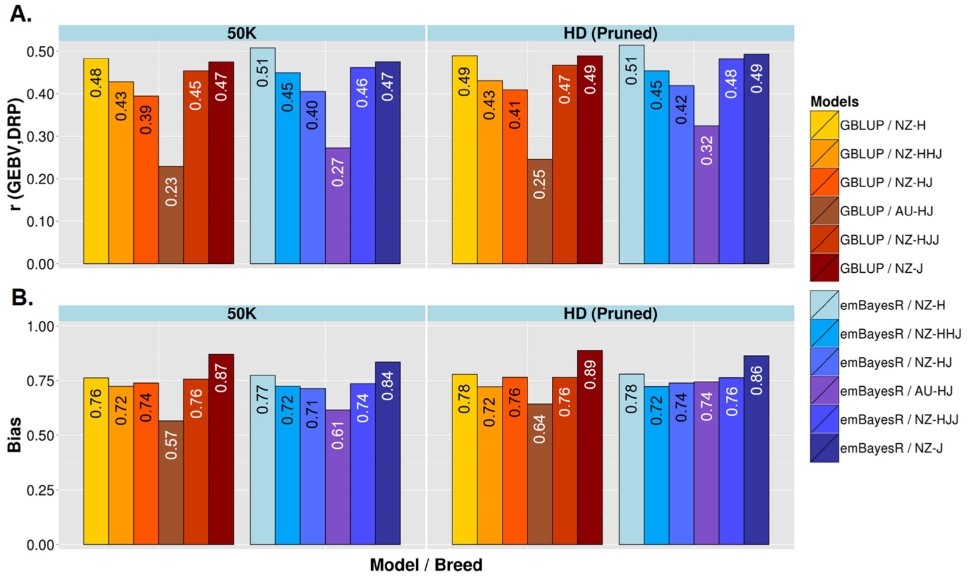

Figure 1. The accuracy (A) and bias (B) of GP in validation breed groups using different marker density genotypes and analytical models

Crossbred vs. purebred. Regardless of SNP density and method used, the accuracy of GP in purebred H was highest followed by pure J cows. It is not surprising that H was the most accurate because the H breed dominated the reference population. Furthermore, the relationships between the different validation sets and the reference may also partly explain the results. The $Rel._{Top10}$ in purebred H and J cows was 0.250 and 0.345, respectively, but was generally lower in crossbred cows: HHJ=0.209, NZ-HJ=0.207, AU-HJ=0.195 and HJJ=0.282. Therefore, this may be partly contributing to the observed lower in accuracy of GP for: crossbreds compared to purebreds, as well as HJ compared to other crosses. The lower accuracy observed in crossbreds could also be partly due to the lower reliability of some crossbred phenotypes compared to purebreds and potentially, genotype imputation in crossbreds may be less accurate than for purebreds. Although pure J validation had the strongest relationships with the reference, the accuracy of GP in pure J was slightly lower than H. This may occur because the proportion of J in the reference is very low compared to H, therefore if some QTL segregate only in J they may not be accurately predicted. The GP in pure J and H was generally less biased than crossbreds. However, the bias across the different crossbred validations was almost the same. Although all validations showed some bias, $b_{DRP,GEBV}$ lower than 1 is a common observation in dairy cattle (Khansefid *et al.* 2014).

50K vs. pruned HD genotypes. There was a consistent increase in accuracy when using pruned HD instead of 50K genotypes, regardless of validation breed group and prediction method (on average ~ 2%; from 0.41 to 0.43). Additionally, the bias of GP was reduced when using denser genotypes (on average ~ 3%; from 0.73 to 0.76). This suggests that increasing the marker density enables more precise estimates of QTL effects because markers tend to be closer and in stronger LD with the causal variants. Moreover, in HD genotypes the markers tend to have the same LD phase with the causal mutation alleles across different breeds. Therefore, for the validation breed group AU-HJ, in which the cows were least related to the reference, the amount of gain from using denser markers was expected to be greatest. However, the amount of gain in AU-HJ accuracy compared to other validation sets was greater only when emBayesR was used in GP. This suggests that to obtain the most benefit from increased marker density, the Bayesian model works better than GBLUP because it provides a more precise estimate of QTL effects. In AU-HJ using HD genotypes instead of 50K genotypes, also reduced the bias of predictions more than other validation breed groups.

GBLUP vs. emBayesR. The accuracy of GP was increased using emBayesR instead of GBLUP in all validation breed groups regardless of marker density (on average ~ 2%; from 0.42 to 0.44). This is likely because the genetic architecture of the milk traits is better modelled by the Bayesian sparse mixture model compared to the quasi-infinitesimal GBLUP model (Goddard *et al.* 2016). In GBLUP the effect of causal QTLs tends to be spread across many markers that are in LD with the causal mutations and all effects come from the same normal distribution. However, in emBayesR the SNP effects are estimated more precisely because we allow a mixture distribution of SNP effects where some may be small medium or large, and a proportion of SNP may have no effect on the trait. This Bayesian model would therefore be expected to show the most benefit when the validation animals are less related to the reference group. Using emBayesR instead of GBLUP did not have a large effect on the bias of GP, except in AU-HJ where the bias of prediction reduced.

CONCLUSIONS

The accuracy of GP in crossbreds was lower than purebreds. Using HD instead of 50K genotypes and emBayesR instead of GBLUP increased the accuracy and reduced the bias of genomic predictions.

ACKNOWLEDGEMENTS

We thank DataGene (particularly Drs K Konstantinov and G Nieuwhof), CRV and DairyNZ staff for processing and provision of cow data. The authors acknowledge financial support from CRV and DairyBio (a joint venture between Agriculture Victoria and Dairy Australia).

REFERENCES

Clark S.A., Hickey J.M., Daetwyler H.D. and van der Werf J.H. (2012) *Genet. Sel. Evol.* **44**: 4.

Kemper K.E., Reich C.M., Bowman P.J., Vander Jagt C.J., Chamberlain A.J., Mason B.A., Hayes B.J. and Goddard, M.E. (2015) *Genet. Sel. Evol.* **47**: 29.

Khansefid M., Pryce J.E., Bolormaa S., Miller S.P., Wang Z., Li C. and Goddard M.E. (2014) *J. Animal Sci.* **92**: 3270.

Garrick D.J., Taylor J.F. and Fernando R.L. (2009) *Genet. Sel. Evol.* **41**: 55.

Goddard M.E., Kemper K.E., MacLeod I.M., Chamberlain A.J. and Hayes B.J. (2016) *Proc R Soc Lond [Biol]* **283**: 20160569.

Meuwissen T.H., Hayes B J. and Goddard M.E. (2001) *Genetics* **157**: 1819.

Lee S.H. and van der Werf J.H.J. (2016) *Bioinformatics* **32**: 1420.

Liu Z. (2009) Deregressing MACE proofs for genomic evaluations, Proteje Meeting, Brussels, Belgium.

Sargolzaei M., Chesnais J.P. and Schenkel F.S. (2014) *BMC Genomics* **15**: 478.

Wang T., Chen Y.P.P., Bowman P.J., Goddard M.E. and Hayes B.J. (2016) *BMC Genomics* **17**:744.