FINDING THE OPTIMAL REFERENCE POPULATION FOR GENOMIC PREDICTION OF AUSTRALIAN RED DAIRY CATTLE

I. van den Berg[1], I.M. MacLeod[1] and J.E. Pryce[1,2]

[1]Agriculture Victoria Research, AgriBio, 5 Ring Road, Bundoora, Victoria, 3083 Australia
[2]School of Applied Systems Biology, La Trobe University, Bundoora, Victoria, 3083 Australia

SUMMARY

Genomic prediction for breeds with a small population size, such as the Australian Red, is challenging, because reliability depends on the size of the reference population and its relatedness to the animals evaluated. Our objective was to find the optimal reference population for Australian Red, comparing within breed and multi breed prediction for milk yield, fat yield, protein yield and somatic cell count.

Our results show that while multi breed prediction can result in higher accuracies than within breed prediction, adding fewer animals that are more closely related to the validation population can result in a higher reliability than adding a much larger number of individuals that are more distantly related.

INTRODUCTION

Genomic prediction for breeds with a relatively small population size, such as Australian Red cattle, is challenging, because the reliability of prediction is dependent on the size of the reference population (Goddard 2009). Sharing reference populations across breeds or countries may increase the size of the reference population, though this has only been advantageous for closely related breeds, such as the Nordic Red cattle breeds (Brøndum *et al*. 2011). Australian Red cattle are influenced by several Red dairy breeds, including Scandinavian Red cattle breeds, Ayrshire, Shorthorn, Illawarra and Red and White Holstein (http://www.aussiereds.com.au).

Multi breed prediction often analyses the same trait in different breeds as a single trait with a breed effect to account for differences across breeds. Not all QTL impact the expression of quantitative traits in the same way across breeds (Raven *et al*. 2014) and there may be QTL by breed interactions resulting in different effects of QTL for different breeds. Therefore, it may be appropriate to fit the same trait in different breeds as multiple correlated traits (Olson *et al*. 2012).

Because linkage disequilibrium is maintained over much shorter distances across breeds than within breed (de Roos *et al.* 2008), prediction reliability is expected to decrease faster across breeds than within a breed when the distance between causal mutations and prediction markers increases (van den Berg *et al.* 2016). Consequently, the standard 50K SNP chip may not be dense enough for accurate prediction from Holstein to Australian Red, and variants close to causal mutations could potentially result in a higher reliability.

The objective of this study was to find the optimal reference population for Australian Red dairy cattle. Within and multi breed reference populations were compared, with multi breed populations containing either a low number of Holstein animals that are relatively closely related to Australian Red cattle based on a genomic relationship matrix between Holstein and Australian Red cattle, or larger numbers of more distant Holstein and Jersey individuals, used a single trait model or a multi trait model that fitted the same trait in different breeds as multiple correlated traits.

MATERIALS AND METHODS

We calculated the reliability of genomic prediction in Australian Red bulls for different reference populations. The reference population contained up to 3,248 Holstein bulls, 48,386 Holstein cows,

807 Jersey bulls, 8,734 Jersey cows and 3,041 Australian Red cows. Genome-wide complex trait analysis (GCTA) (Yang *et al.* 2011) was used to first construct a genomic relationship matrix of the full reference population and perform a principal component analysis (PCA). In total, 10 reference populations were used. The largest reference population contained 3,041 Australian Red cows, 51,634 Holstein and 9,541 Jersey individuals, and the smallest only the Australian Red cows. Additional reference populations contained the Australian Red cows and either all Holstein individuals or only Holsteins with a value for the first principal component (PC1) above a certain threshold. Figure 1 shows the first two principal components of the PCA, and indicates the groups used to construct different reference populations. The number of individuals in each of these seven subsets is shown in Table 1. The validation population contained 280 Australian Red bulls. Deregressed proofs (DRP) for milk (MY), fat (FY) and protein yield (PY) and somatic cell count (SCC) were calculated following Garrick *et al.* (2009) and used as phenotypes.

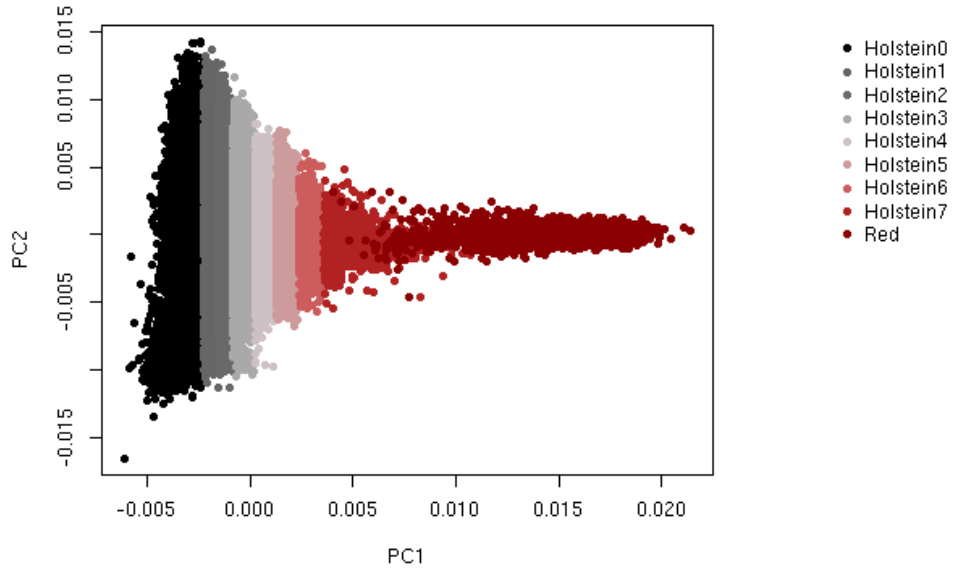

Figure 1. First two principal components (PC1 and PC2) of the genomic relationship matrix of the multibreed reference population containing Holstein and Australian Red individuals. Different colours show different subsets of animals that are used to construct different reference populations

Table 1. Number of Holstein and Australian Red (Red) individuals in different reference populations based on the first principal component

Breed	H1-7+R	H2-7+R	H3-7+R	H4-7+R	H5-7+R	H6-7+R	H7+R
Holstein	39,788	29,809	19,835	9,880	4,915	2,436	1,197
Red	3,041	3,041	3,041	3,041	3,041	3,041	3,041

Genotypes were available for the Illumina BovineSNP50 chip (50K, real or imputed). Because the LD between QTL and prediction markers on the 50K chip may not be conserved across breeds, we also analysed genotypes on a custom chip with 46,516 imputed sequence variants selected by Xiang *et al.* (2019) that are expected to be enriched for dairy trait QTL (XT).

For each of the reference populations, we used the GBLUP model as implemented in MTG2 (Lee and van der Werf 2016) to predict GEBV of the validation population. The reliability of genomic prediction was calculated as the squared correlation between DRP and GEBV divided by the average reliability of individuals in the validation population. The model either considered the same trait in different breeds as a single trait, fitting a breed effect to correct for breed differences (ST-GBLUP), or fitted the same trait in different breeds as different, correlated traits, using a multi trait model (MT-GBLUP).

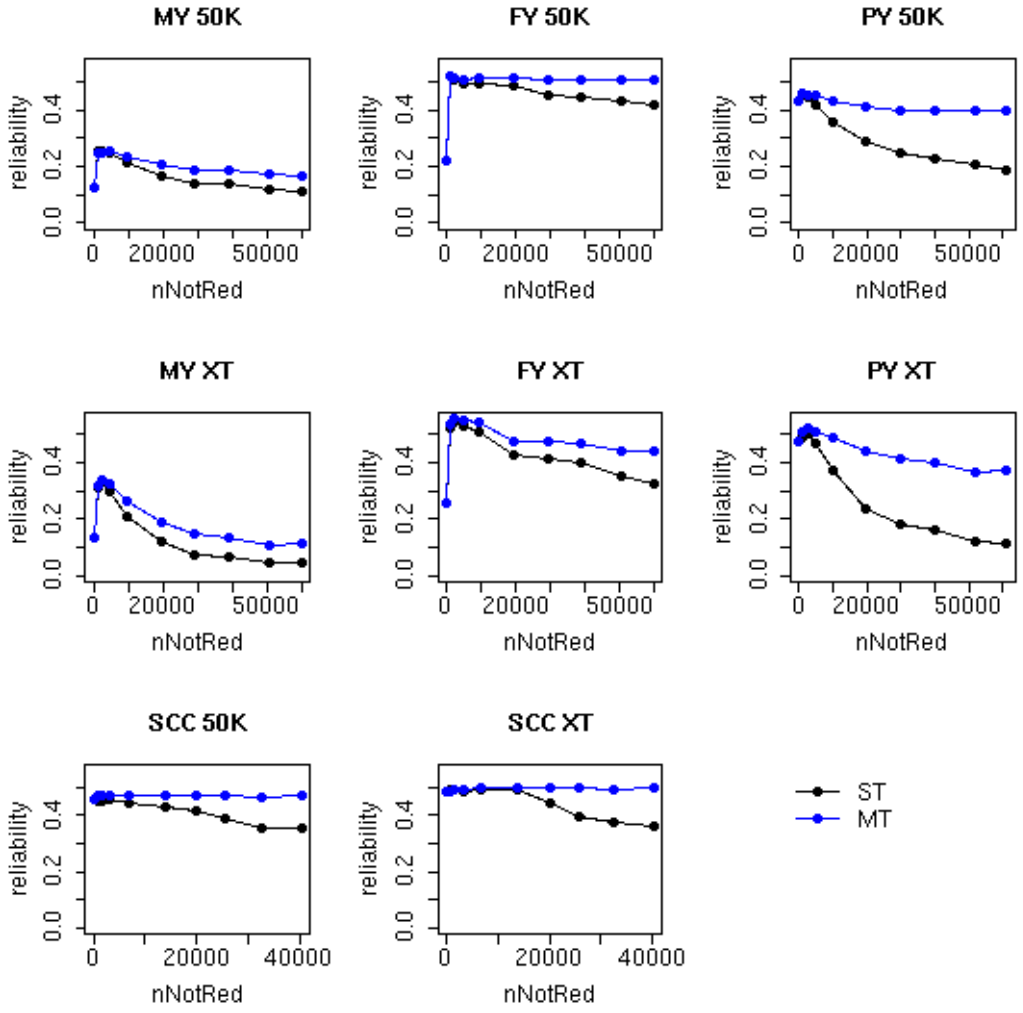

Figure 2. Reliability of genomic prediction as a function of the number of Holstein and Jersey individuals in the reference population (nNotRed) for milk yield (MY), fat yield (FY), protein yield (PY) and somatic cell count (SCC), using variants on the 50K SNP chip (50K) or selected sequence variants (XT). For the multi breed reference populations, the same trait in different breeds was analysed using a single trait model fitting a breed effect (ST) or a multi trait model considering the trait as multiple correlated traits in different breeds (MT)

RESULTS AND DISCUSSION

Figure 2 shows the reliability as a function of the composition of the reference population. The overall pattern was similar for all traits: the highest accuracies were obtained using a multi breed reference population with a limited number of Holstein individuals that are relatively closely related to the Australian Red. population.

For all traits tested, the highest reliability was obtained with the MT model and a multi breed reference population. The XT variants only led to a small difference in reliability compared to the 50K variants. For MY, FY and PY, the reference population resulting in the highest reliability contained around 2,400 Holsteins (with reliabilities of 0.34, 0.56, 0.52 for MY, FY and PY, respectively), while for SCC, the highest reliability (0.50) was obtained with 13,822 Holsteins. Adding Jerseys to full reference population containing all Holstein individuals resulted in a similar reliability as obtained without the Jerseys.

Except for FY, the reliability obtained with the full multi breed reference population was lower than the reliability obtained with the within breed reference population. The decrease in reliability when adding larger numbers of Holstein individuals to the reference population was larger with the ST model than with the MT model.

The GBLUP prediction models assume all variants are equally important to predict the trait. Models that can allocate higher importance to variants linked to causal mutations, such as Bayesian variable selection models or a weighted GBLUP, may result in higher and be less prone to the decrease in reliability we observed when adding larger numbers of Holstein individuals to the reference populations.

CONCLUSIONS

Our results show that while multi breed prediction can result in higher accuracies than within breed prediction, adding fewer animals that are more closely related to the validation population can result in a higher reliability than adding a much larger number of individuals that are more distantly related. To implement genomic prediction in Australian Red cattle, an international reference population containing other Red breeds is likely to lead to a higher reliability than a multi breed Australian reference population.

ACKNOWLEDGEMENTS

We acknowledge Bolormaa Sunduimijid for providing the imputed sequence variants, DataGene for providing access to data used in this study, and our partners in the 1000 Bulls Genomes Project for access to the reference genomes.

REFERENCES

Brøndum R.F., Rius-Vilarrasa E., Strandén I., Su G., Guldbrandtsen B., Fikse W.F. and Lund M.S. (2011) *J. Dairy Sci.* **94**: 4700.

de Roos A.P.W., Hayes B.J., Spelman R.J. and Goddard M.E. (2008) *Genetics* **179**: 1503.

Garrick D.J., Taylor J. F. and Fernando R.L. (2009) *Genet Sel Evol* **41**:44.

Goddard M.E. (2009) *Genetica* **136**: 245.

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

Olsen K.M., VanRaden P.M., and Tooker M.E. (2012) *J. Dairy Sci.* **95**: 5378.

Raven, L-A., Cocks B.G. and Hayes B.J. (2014) *BMC Genomics* **15**: 62.

van den Berg I., Boichard D., Guldbrandtsen B. and Lund M.S. (2016) *G3* **6**: 2553.

Xiang R., van den Berg I., MacLeod I.M., Hayes B.J., Prowse-Wilkins C.P., Wang M., Bolormaa S., Liu Z., Rochfort S.J., Reich C.M., Mason B.A., Vander Jagt C.J., Daetwyler H.D., Chamberlain A.J. and Goddard M.E. (2019) *Submitted*

Yang J., Lee S.H., Goddard M.E. and Visscher P.M. (2011) *Am. J. Hum. Genet.* **88**: 76.