THE IMPORTANCE OF POPULATION STRUCTURE ON THE ACCURACY OF GENOMIC PREDICTION IN A MULTI-BREED SHEEP POPULATION

H.D. Daetwyler[1,2], K.E. Kemper[1,3], J.H.J. van der Werf[2,3], and B.J. Hayes[1,2]

[1]Department of Primary Industries, 1 Park Drive, Bundoora 3083, Vic, Australia, [2]CRC for Sheep Industry Innovation, Armidale, 2351, NSW, Australia, [3]University of Melbourne, Parkville 3010, Vic, Australia, [3]University of New England, Armidale 2351, NSW, Australia

SUMMARY

Population structure, due to breed, strains and sire family, influences the accuracy of genomic prediction. We investigated principle component analysis as a way to account for population structure in within and across breed genomic prediction of greasy fleece weight and eye muscle depth in multi-breed sheep data. Population structure (including for example half sib family relationships) is responsible for a large proportion of the accuracy of genomic prediction. Correcting for it increased accuracy of greasy fleece weight across breed prediction, but reduced accuracy of across breed prediction for eye muscle depth for breeds not in the reference set. However, the correction reduced within breed accuracy.

INTRODUCTION

Genomic prediction (Meuwissen *et al.* 2001) is a method of estimating an individual's genetic merit using genetic markers and phenotypic records. It has been demonstrated that relatedness of reference to validation sets influences the accuracy of genomic prediction (Habier *et al.* 2007; Habier *et al.* 2010). The more related the reference and validation, the higher the accuracy. In multi-breed populations, population structure, as well as within breed relatedness also includes within and across breed associations. So in multi-breed populations the accuracy of genomic prediction could be expected to have two main components: i) prediction based on genomic relationships arising from population structure, both within and across breeds and ii) prediction based on linkage disequilibrium (LD) between markers and QTL. The two components are correlated, because breed relatedness increases LD across breeds and within breed relationships increase linkage. It is currently unclear the extent to which the two sources contribute to accuracy in multi breed populations. However, the distinction is important as accuracy due to LD is more likely to persist across generations and even across breeds if marker and QTL phase is consistent. In contrast, the accuracy due to relatedness does not persist across breeds or even across generations (Habier *et al.* 2007; De Roos *et al.* 2009). An across-breed strategy for genomic prediction would be suited to species with multiple prominent breeds (Hayes *et al.* 2009; Daetwyler *et al.* 2010). Attempts to account for population structure have included fitting a pedigree, fitting breed effects, and principle components (PCs; e.g. Price *et al.* 2006). Principle component analysis (PCA) is attractive when pedigrees are not available, but it may not adequately correct for population structure in diverse population samples (McVean 2009). Guidelines are lacking on whether and when correcting for population structure is advantageous in genomic prediction.

Here we investigate the influence of population structure on the accuracy of genomic prediction both within and across breed in a large multi-breed sheep dataset. In addition, we explore how PCA performs in accounting for population structure and investigate the behaviour of accuracy as a varied number of PCs are fitted in the model.

METHODS

Two phenotypic traits were investigated in sheep, yearling greasy fleece weight (GFW) and ultrasound scanned eye muscle depth (EMD). GFW and EMD have heritabilities of 0.37 and 0.23, respectively (Safari *et al.* 2005; Mortimer *et al.* 2010). The reference population included 3341 and 7431 animals for GFW and EMD respectively. Whereas the GFW reference was mostly Merino sheep (MER), the EMD data contained greater proportions of Border Leicester (BL), Polled Dorset (PD) and White Suffolk (WS). The datasets have been described in more detail in Daetwyler *et al.* (2010). Breed group size ranged from 3307 animals for purebred MER to 5 for a BL/East Friesian/PD. A total of 196 rams sired the total reference population and the size of the resulting half-sib families ranged from 385 to 1. The size of the ram half-sib families was often larger than the number of animals in the respective breed-cross groups.

The genomic predictions estimated in the reference population were tested in a validation population consisting of purebred rams with high accuracy Australian sheep breeding values (ASBVs). Genomic prediction accuracies were calculated within the following breeds: MER, BL, PD, and WS, as the Pearson correlation of genomic breeding values and validation ram ASBVs. ASBV accuracy for GFW was low in PD and WS and correlations are therefore not presented, the remaining ram ASBVs mean accuracies were all above 0.83. All animals were genotyped using the Illumina 50K ovine SNP chip (Illumina Inc., San Diego, USA), which reacts to 54,977 SNPs. Quality control reduced the number of SNP to 48640.

The following genomic best linear prediction (GBLUP) model was fitted in ASReml (Gilmour *et al.* 2009): $\mathbf{y} = \mathbf{Xb} + \mathbf{Zg} + \mathbf{e}$, where **y** was a vector of phenotypic records, **X** and Z were design matrices, **b**, **g**, and **e** were vectors of fixed, additive genetic and residual effects, respectively. The following distributions were assumed: $\mathbf{g} \sim N\,(0,\ \sigma_g^2\mathbf{G})$ and $\mathbf{e} \sim N\,(0,\ \mathbf{I}\sigma_e^2)$, where **G** was a genomic relationship matrix calculated as in Yang *et al.* (2010). Fixed effects were sex, birth type, rearing type, contemporary group (birth year×site×management group) and age at trait recording. Weight at scanning was fitted for EMD. Sire and dam breed effects were fitted in some analyses.

PCA was performed on **G** using the R function **eigen**. We coded dummy variables to contrast animals of a particular breed or breed cross with all other animals. The dummy variables were correlated with the first 200 PCs, with the expectation that correlations would be high for PCs associated with this breed-cross group. This was repeated for individual ram half-sib families. The impact of PCs on genomic prediction was gauged by fitting a range of 0 to 200 PCs as fixed covariates in GBLUP analysis. Sire and dam breed were not fitted in models with PCs. A chromosome specific **G** was calculated for chromosome 1 and was fitted with and without 200 PCs instead of the genome-wide **G**, to assess what component of genetic variance **G** was associated with. Predictions from a multi-breed reference set including all breeds are denoted Case 1. The accuracy of across breed prediction was also investigated in subsets of the multi-breed reference populations excluding the breed to be predicted (Case 2), which were used to predict BL, PD and WS rams. An increasing number of PCs was fitted to evaluate their impact on across breed prediction accuracy.

RESULTS AND DISCUSSION

Our dummy correlations revealed that the PC at which a group, be it a breed-cross or a half-sib family, is differentiated from the rest is greatly dependent on its size. While MER were differentiated in PC1, the largest ram half-sib family was differentiated long before other smaller breed groups. This raises doubts about whether PCA can be used to only correct for breed effects while leaving structure due to families intact. Considering the results in this study, the general practice of fitting only the first few PCs seems inadequate in diverse data, indeed fitting any number of PCs reduced within breed accuracy (Figure 1).

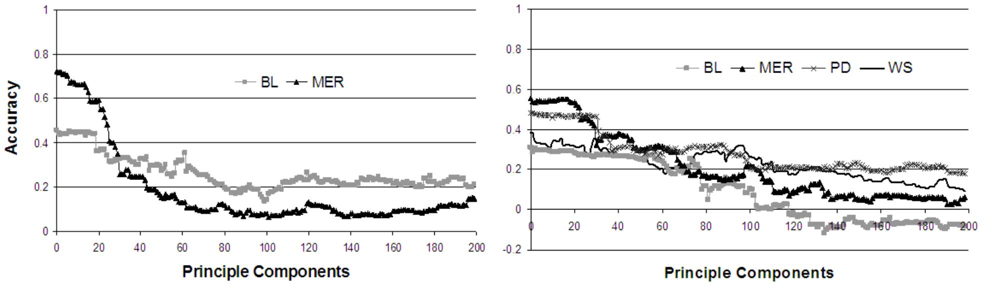

Figure 1. Accuracy of genomic prediction in GFW (left) and EMD (right) when an increasing number of PCs are fitted in addition to the base model in multi-breed reference populations including all breeds (Case 1).

An increasing number of PCs from zero to 200 were fitted in GBLUP to determine their effect on accuracy. Figure 1 shows the decay in accuracy as more PCs are fitted in both GFW and EMD. In GFW, a clearer trend of diminished accuracy as more PCs were fitted was observed in MER and BL. The MER group reached a lower plateau at approximately PC 50 whereas BL reached this plateau at approximately PC 80. In EMD, the various breeds were more equally represented in the reference population and all four validation breeds reached lower plateaus between PC 110 and 130. We speculate that these lower plateaus are a measure of the accuracy due to LD of markers and QTL, as the majority of the effect of population structure has been accounted for. These plateaus will not continue indefinitely, as eventually the PCs account for variation due to LD. While initially MER had the highest accuracy at low PCs, the PD and WS breeds had higher accuracies once the lower plateau was reached (Table 1). This trend of lower MER accuracy at late PCs was consistent in both traits and may be due to the lower effective population sizes (N_e) of BL, PD, and WS when compared to MER (e.g. less LD between SNP and QTL in MER).

Table 1. Accuracy in GFW and EMD in four breeds for a reference set including all breeds (Case1), where Ch1-NoPC and Ch1-200PC are the accuracy of chromosome 1 with and without PCs. Case 2 is the across breed accuracy in multi-breed data excluding the breed to be predicted with and without fitting PCs.

		Across Breed Accuracy Case 1 All Breeds in Reference				Across Breed Accuracy Case 2	
Trait	Breed	Total No PC	Plateau 200 PC	Ch1 No PC	Ch1 200 PC	No PC	With PC
GFW	MER	0.72	0.15	0.62	-0.09	NA	NA
	BL	0.46	0.21	0.43	0.03	0.05	0.20
EMD	MER	0.56	0.06	0.46	-0.01	NA	NA
	BL	0.31	-0.08	0.15	-0.17	0.08	0.01
	PD	0.48	0.18	0.41	0.14	0.33	0.27
	WS	0.39	0.09	0.48	0.40	0.26	0.17

Fitting a chromosome specific relationship matrix revealed that a large proportion of accuracy was due to population structure because the accuracy achieved with a single chromosome was high (Table 1), and it is extremely unlikely that most QTL underlying genetic variation reside only on chromosome 1. In GFW, fitting 200 PCs reduced the percentage of total accuracy in MER and BL. In EMD, the percent of total accuracy of chromosome 1 was reduced in MER and PD when fitting

200 PCs, but increased BL and WS. As can be seen below, it is possible that across breed prediction may have been improved by fitting more PCs and this may have contributed to greater proportional accuracies in some cases. Fitting sire and dam breed in the model only marginally reduced the accuracy from chromosome 1 (results not shown), demonstrating that it only weakly accounted for population structure.

The accuracy achieved from across breed prediction is an ultimate measure of the accuracy due to LD when the reference set excludes the breed to be predicted (Table 1), as across breed prediction accuracy cannot arise from within breed population structure (although it is a lower limit as only QTL segregating in multiple breeds will be exploited). When the highest across breed accuracy was used, fitting PCs resulted in increased accuracy for BL in GFW. In EMD, no advantage of fitting PCs was observed in any breed. The inconsistent results highlight the need for extensive data exploration to maximise the accuracy for a particular breed and trait.

The main reason for the large disparity between accuracy due to population structure and accuracy due to LD is the sheep SNP chip is not dense enough to ensure high LD between SNP and QTL, reducing the accuracy of this component.

CONCLUSIONS

A large proportion of the accuracy of genomic prediction in sheep is due to population structure at the current medium SNP density. This makes across breed prediction difficult and predictions unstable over many generations. There was an inconsistent trend that accounting for population structure with PCs lead to increases in across breed accuracy. However, adjusting for population structure always decreased the within breed accuracy. In the short term, increasing the number of animals of the target breed in the reference population would yield the quickest increase in accuracy. With higher density SNP, a strategy could be pursued where across breed prediction would account for population structure but within breed prediction would not. An across breed strategy is expected to be more effective in BL, PD and WS due to smaller effective population size than in MER.

ACKNOWLEDGEMENTS

The authors would like to thank the CRC for Sheep Industry Innovation and the Sheep Genomics project (FMFS), including MLA and AWI Ltd, for funding and providing data. We also thank Daniel Brown, Andrew Swan, Ken Geenty, Klint Gore, Cedric Gondro, James Kijas and farm staff from the Sheep CRC Information Nucleus Flock and the FMFS for their assistance in data collection and processing.

REFERENCES

Daetwyler H. D., Hickey J. M., Henshall J. M., *et al.* (2010). *Anim. Prod. Sci.* **50:** 1004.
De Roos A. P. W., Hayes B. J. and Goddard, M. E. (2009). *Genetics* **183:** 1545.
Gilmour A. R., Gogel B., Cullis B. R., *et al.* (2009). *2009 ASReml user guide release 3.0*.
Habier D., Fernando R. L. and Dekkers J. C. M. (2007). *Genetics* **177:** 2389.
Habier D., Tetens J., Seefried F.-R., *et al.* (2010). *Genet. Sel. Evol.* **42:** 5.
Hayes B. J., Bowman P. J., Chamberlain *et al.* (2009). *Genet. Sel. Evol.* **41:** 51.
McVean G. (2009). *PLoS Genet.* **5,** e1000686.
Meuwissen T. H. E., Hayes B. J. and Goddard M. E. (2001). *Genetics* **157:** 1819.
Mortimer S. I., van der Werf J. H. J., Jacob R. H., *et al.* (2010). *Anim. Prod. Sci.* **50:** 1135.
Price A. L., Patterson N. J., Plenge R. M., *et al.* (2006). *Nat. Genet.* **38:** 904.
Safari E., Fogarty N. M. and Gilmour A. R. (2005). *Livest. Prod. Sci.* **92:** 271.
Yang J., Benyamin B., McEvoy B. P., *et al.* (2010). *Nat. Genet.* **42:** 565.