Proc. Assoc. Advmt. Anim. Breed. Genet. 22:593-596

ACCOUNTING FOR POPULATION STRUCTURE IN GENOMIC PREDICTION OF AUSTRALIAN MERINO SHEEP

N. Moghaddar[1, 3], D.J. Brown[2, 3], A.A. Swan [2, 3] and J.H.J. van der Werf[1, 3]

[1] School of Environmental and Rural Science, University of New England, Armidale, 2351
[2] Animal Genetics and Breeding Unit (AGBU), University of New England, Armidale, 2351
[3] Cooperative Research Centre for Sheep Industry Innovation, Armidale, Australia.

SUMMARY

The aim of this study was to compare different ways of accounting for population structure for genomic prediction of three economic traits in an Australian Merino sheep population. Population structure was accounted for either by fitting genetic groups (GG) derived from pedigree, or fitting principal components (PCs) calculated from the genomic relationship matrix based on 50k density SNP marker genotypes. Genomic breeding values (GBV) were calculated using genomic best linear unbiased prediction (GBLUP) and the GBV accuracy was evaluated based on 5 fold cross-validation across half-sib families. Best linear unbiased estimation (BLUE) of GG or PC effects were added to the GBV. Results showed that accounting for population structure either by fitting GG or PCs improved the accuracy of genomic prediction. Furthermore, fitting the first two PCs gave a similar accuracy to fitting GG derived from pedigree. The improvement in GBV accuracy after accounting for population structure in studied traits was not high (3.8% when averaged across traits) which may be because the genomic relationship matrix will implicitly account for some of the population structure effect when the GG or PCs are not fitted in analysis. In the case of missing or incomplete pedigrees, PCs can be used to account for population structure and to improve the prediction accuracies.

INTRODUCTION

Differences in average genetic effects of breeds or strains within breeds (population structure) may affect the accuracy of genetic merit evaluation of selection candidates. Population structure could bias the genomic estimated breeding values (GBV) and hence affect the realized selection response. Australian Merino sheep is a highly diverse population due to different breeding objectives within the various types of Merino, and due to different production environments. The Merino breed consists of many sub-populations according to wool quality, e.g. strong wool, fine wool and ultra-fine wool Merinos. Accounting for population structure is a very importance feature of MERINOSELECT which is the national genetic evaluation of Australian Merino sheep (Brown *et al*. 2015; Swan *et al*. 2014)

The effect of population structure can be accounted for in the estimation of breeding values (based on phenotype and pedigree), according to genetic groups derived from pedigrees. However, in the case of incomplete pedigree information, population structure can be derived from genotypes by using Principal Components (PCs) from the genomic relationships matrix (GRM) (Price *et al.* 2006). Fitting PCs explicitly in the model is likely more accurate than accounting for the structure implicitly through the GRM (Van der Werf *et al*. 2013). The aim of this study was to compare fitting genetic groups based on pedigree with fitting PCs based on the genomic relationship matrix when accounting for population structure in genomic prediction of Australian Merino sheep.

MATERIALS AND METHODS

Reference population, phenotypes and validation population. The traits studied were post weaning weight (PWW, 6,388 records), ultrasound scanned eye muscle depth (PEMD, 4,012

records) measured between 150 and 290 days from birth and yearling greasy fleece weight (YGFW, 5,200 records) on Merino sheep. Animals originated from the "Sheep Cooperative Research Centre Information Nucleus Flock" (INF) and the Resource Flock (RF) which consisted of eight sites located across different regions of Australia and these were linked to each other by using common sires through artificial insemination between 2007 and 2015. More information on the scope and design of the INF is provided by Van der Werf *et al.* (2010). The accuracy of genomic prediction was evaluated based on the average of 5-fold cross-validation, where whole half sib families were sampled such that half sibs could not appear in training as well as validation set. The accuracy was calculated as the correlation between the GBV and the corrected phenotype, divided by the square root of the trait heritability.

Genotypes. Genotypes were available based on real 50K Ovine marker panel (Illumina Inc., San Diego, CA, USA) or 12K which was imputed to 50K. The 50K and 12K marker panel provided respectively 48,559 and 12,646 SNP genotypes after applying quality control. The sporadic missing genotypes were imputed first using Beagle 3.0 (Browning 2009). Animals genotyped with 12K marker density then were imputed to 50K density using Beagle 3.0 and using all Merino animals genotyped with 50K marker density as reference set. Accuracy of imputation was shown to be high (on average 0.96).

Statistical methods. Genomic best linear unbiased prediction (GBLUP) was used to calculate the Genomic Breeding Values (GBV) using the ASReml (Gilmour *et al.* 2009) program. The model fitted for each trait was: $y = Xb + Z_1g + Z_2m + e$ where y is a vector of phenotypes, b is a vector with fixed effects, g is the random additive genetic effect of the animal, m is a vector with maternal effects and e is vector of random residual effects, X, Z_1 and Z_2 are incidence matrices relating effects to animals. The parameters g, m and e are considered normally distributed as: $g \sim N(0, G\sigma_g^2)$, $m \sim N(0, I\sigma_m^2)$ and $e \sim N(0, I\sigma_e^2)$, respectively and G was the genomic relationship matrix calculated based on 50k markers genotypes using the VanRaden (2008) method. The common fixed effects in all models were birth type, rearing type, gender, age at measurement and contemporary group which was flock × birth year × management group. In the GG models 5 genetic groups were fitted as a regression (fixed continuous variable) on proportion of Merino sub-population (strains) where the proportions for individual animals were derived from a deep pedigree. In the PC models principal components were fitted by regression on up to ten eigenvectors associated with the largest 10 principal components.

RESULTS AND DISCUSSION

Tables 1, 2 and 3 compare the accuracy of genomic prediction between different models of fitting GG or PCs to account for population structure for PWW, PEMD and YGFW, respectively. Results show higher prediction accuracy for three different traits studied when population structure was accounted for in the model and then solutions for GG or PCs' effects were added to the GBV. This result was in line with a previous study by Daetwyler *et al.* (2013) who showed higher genomic prediction accuracy within Australian sheep breeds by accounting for population structure using PCs. However, the improvement in accuracy compared to only fitting the GRM in this study was not very high and on average 3.4% in absolute value.

Results showed fitting the first two largest PCs resulted in similar prediction accuracy to fitting GG from pedigree. Brown *et al.* (2015) and Swan *et al.* (2014) also showed strong correlation between using GG derived from pedigree and PCs calculated from genomic relationship matrix to correct the impact of population structure on estimation of genetic merits of animals. In this study the accuracy of GBV (GG/PC effect inclusive) was not increased by fitting more PCs. Results also showed a continuous decrease in GBV accuracy if the GG or PC effect solution was not added to GBV (Tables 1-3).

Table 1. Variance components, (SE) and average accuracy of genomic predictions from 5 fold cross-validation for PWW based on fitting genetic groups (GG) or Principal Components (PCs)

Model	Ve[1]	Va[2]	Vdam[3]	r(GBV1,Res)[4]	r(GBV2,Res+GG) [5]
No GG	13.83 (0.73)	12.05 (0.91)	2.08 (0.61)	NA	0.348
GG	14.61 (0.74)	10.22 (0.89)	2.28 (0.61)	0.243	0.368
1PC	14.24 (0.73)	10.98 (0.90)	2.23 (0.61)	0.218	0.342
2PC	14.41(0.73)	10.55 (0.89)	2.30 (0.61)	0.215	0.355
3PC	14.96 (0.74)	9.33 (0.88)	2.44 (0.61)	0.194	0.322
4PC	14.94 (0.74)	9.36 (0.88)	2.43 (0.61)	0.194	0.322
5PC	14.93 (0.74)	9.40 (0.88)	2.43 (0.61)	0.191	0.322
10PC	14.99 (0.74)	9.24 (0.88)	2.45 (0.61)	0.178	0.316

[1]Residual variance, [2]Additive genetic variance, [3]Dam permanent environmental effect, [4]Average of correlation between GBV (corrected for GG or PC effects) and corrected phenotypes (adjusted for GG effects). [5]Average of correlation between GBV (plus solution for GG or PCs) and corrected phenotypes (not adjusted for GG effect).

Table 2. Variance components, (SE) and accuracy of genomic prediction for PEMD based on fitting genetic groups (GG) or Principal Components (PCs)

Model	Ve[1]	Va[2]	r(GBV1,Res)[3]	r(GBV2,Res+GG) [4]
GG not fitted	5.066 (0.22)	2.251 (0.25)	NA	0.384
GG fitted	5.398 (0.23)	1.728 (0.25)	0.348	0.420
1PC	5.146 (0.22)	2.121 (0.25)	0.341	0.412
2PCs	5.237 (0.22)	1.976 (0.25)	0.320	0.422
3PCs	5.504 (0.22)	1.565 (0.25)	0.317	0.394
4PCs	5.496 (0.23)	1.552 (0.25)	0.316	0.393
5PCs	5.510 (0.23)	1.550 (0.25)	0.316	0.393
10PCs	5.524 (0.23)	1.550 (0.25)	0.311	0.387

[1]Residual variance, [2]Additive genetic variance, [3]Average of correlation between GBV (corrected for GG or PC effects) and corrected phenotypes (adjusted for GG effects). [4]Average of correlation between GBV (plus solution for GG or PCs) and corrected phenotypes (not adjusted for GG effect).

Tables 1, 2 and 3 also show the additive genetic, residual and dam variance (for PWW and YGFW only) for different models. Results show a continuous decrease in additive genetic variance and an increase in residual variance by fitting GG or fitting 1 to 10 PCs. The change in dam effect was very small in PWW and YGFW.

Results of this study showed that accounting for population structure according to pedigree or genomic information improves the total genetic merit prediction accuracy. However, the increase in prediction accuracy in traits studied was not very high compared to fitting only the GRM. This indicate that it is likely that the GRM could account for only part of the effect of population structure implicitly as was indicated before (Van der Werf *et al.* 2013).

The reason for lower accuracy of GBVs (corrected for PCs) by fitting more PCs would be because fitting more PCs can capture part of the total additive genetic variance between different flocks and between half-sib families within flocks.

In term of estimating the total genetic merits for animals with pedigree information the results show the GG model seems to work slightly better than PCs model. However, fitting the first two largest PCs derived from the GRM can also sufficiently account for population structure. This shows that in the case of missing, incomplete or not reliable pedigree information and if the animals were genotyped, PCs could be used to account for population structure to obtain higher prediction accuracies within a breed. This could be more important in prediction of unbiased breeding values on the national scale such as Australian Sheep Breeding values (ASBV) with probable larger impact of genetic groups.

Table 3. Variance component, (SE) and accuracy of genomic prediction for YGFW based on fitting genetic groups (GG) or Principal Components (PCs)

Model	Ve[1]	Va[2]	V(dam)[3]	r(GBV1,Res)[4]	r(GBV2,Res+GG) [5]
GG not fitted	0.160 (0.01)	0.128 (0.01)	0.016 (0.01)	NA	0.564
GG fitted	0.163 (0.01)	0.121 (0.01)	0.017 (0.01)	0.532	0.611
1PC	0.153 (0.01)	0.131 (0.01)	0.020 (0.01)	0.524	0.562
2PCs	0.156 (0.01)	0.127 (0.01)	0.021 (0.01)	0.519	0.604
3PCs	0.157 (0.01)	0.122 (0.01)	0.021 (0.01)	0.509	0.569
4PCs	0.161 (0.01)	0.122 (0.01)	0.021 (0.01)	0.509	0.566
5PCs	0.163 (0.01)	0.121 (0.01)	0.022 (0.01)	0.508	0.566
10PCs	0.167 (0.01)	0.116 (0.01)	0.021 (0.01)	0.487	0.560

[1]Residual variance, [2]Additive genetic variance, [3]Dam permanent environmental effect, [4]Average of correlation between GBV (corrected for GG or PC effects) and corrected phenotypes (adjusted for GG effects), [5]Average of correlation between GBV (plus solution for GG or PCs) and corrected phenotypes (not adjusted for GG effect).

ACKNOWLEDGEMENTS

The authors would like to extend their gratitude to Klint Gore (University of New England, Armidale, NSW, 2351, Australia) for preparing and cleaning genotype data and managing the CRC information nucleus database, and all staff involved at the Sheep CRC Information Nucleus sites across Australia.

REFERENCES

Brown D.J., Swan A.A., Gill J.S., *et al*. (2015). *Proc. Assoc. Advmt. Anim. Breed. Genet.* **20**: 66.

Browning B.L. and Browning S.R. (2009) *Am. J. Hum. Genet.* **84**: 210.

Gilmour A.R., Gogel B.J., Cullis B.R., *et al.* (2009) ASReml User Guide Release 3.0.

Price A.L., Patterson N.J., Plenge R.M., *et al* (2006) *Nature Genet.* **38**: 904.

Swan A.A., Brown D.J., Daetwyler H.D *et al.* (2014) *Proc. 10th World Congress Gen. Appl. Livest. Prod.*, Vancouver, Canada.

Van der Werf J.H.J., Kinghorn B.P. and Banks R.G. (2010) *Anim. Prod. Sci.* **50**: 998.

Van der Werf J.H.J., Brown D.J. and Swan A.A. (2013) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **20**: 340.

VanRaden P.M. (2008) *J. Dairy Sci.* **91**: 4414.