Proc. Assoc. Advmt. Anim. Breed. Genet. 22:341-344

USE OF GENOMIC DATA TO DETERMINE BREED COMPOSITION OF AUSTRALIAN SHEEP

P.M. Gurman, A.A. Swan and V. Boerner

Animal Genetics and Breeding Unit*, University of New England, Armidale, NSW 2351 Australia

SUMMARY

The Australian sheep industry is characterised by the use of many sheep breeds and regular crossing among them. For the purposes of genetic evaluation, sheep are currently assigned breed proportions based on pedigree. SNP genotypes have been used in many applications to reveal population structure including livestock breeds. In this paper, we investigate the use of SNP genotypes to determine breed structure in Australian sheep breeds using the genotype database assembled for genetic evaluation. Algorithms implemented in two software programs, ADMIXTURE and BreedComp were able to identify sheep breeds and genetic groups within the Merino breed. These results can potentially lead to more accurate identification of breed content, and more accurate predictions of breeding value through improved allocation to genetic groups.

INTRODUCTION

A number of sheep breeds play an important role in the Australian sheep meat and wool industries, with the Merino dominant for wool production, Border Leicester, Coopworth and composite breeds used for maternal performance, and Poll Dorset, Texel, and Suffolk and White Suffolk used as terminal sires for meat production. Crossing among these breeds is common not only at the commercial level, but also in seed-stock flocks where some breeders seek to exploit breed differences. Considerable genetic diversity is present within the Merino breed, such that many flocks are considered to be different genetic groups for the purposes of evaluation. Currently, the evaluation system accounts for breed and within-breed genetic group differences using the Westell-Quaas approach (Westell *et al.* 1988), in which the breed composition of each animal is modelled through the pedigree. With increasing availability of genomic data, the utility of this data in estimating sheep breeds and genetic groups within breeds has been examined. Various authors have investigated the use of genomic data to identify population structures in beef breeds (Sölkner *et al.* 2010; Kuehn *et al.* 2011a; VanRaden *et al.* 2011; Frkonja *et al.* 2012) and sheep breeds (Dodds *et al.* 2013). In this paper, we investigate the use of genomic data to identify breed and within-breed population structures in Australian sheep.

MATERIALS AND METHODS

50K SNP genotypes (as described by Moghaddar *et al.* (2015)) and pedigree-based breed proportions were collated for 31,125 sheep from the reference and industry populations used for genomic evaluation in Australia. These data contained records for straight-bred sheep with 623 Border Leicester, 1,966 Poll Dorsets, 28 Texels, 37 Suffolks, 39 White Suffolks, and 14,440 Merinos, where "straight-bred" is defined here as containing at least 0.95 of that breed proportion from the pedigree. Of the recorded straight-bred Merino sheep, some were recorded as straight-bred of a particular Merino group with 456 as 'ultra-fine', 2,907 as 'fine-medium' and 967 as 'strong'. The genomic relationship matrix (G) was calculated using the method by Yang *et al.* (2010) and a singular value decomposition performed on the G matrix, such that $G = U\Sigma V$. Vectors of the U

*AGBU is a joint venture of NSW Department of Primary Industries and the University of New England

matrix were then analysed visually in 'R' (R Core Team 2016) to determine if breeds and genetic groups could be identified.

Analytical tools were used to predict breed and genetic group proportions including supervised ADMIXTURE (Alexander *et al.* 2009) and constrained genomic regression, hereby referred to as BreedComp (Boerner, 2017), which is a constrained (i.e. estimated proportions are less than one) version of the approach by Chiang *et al.* (2010) and Kuehn *et al.* (2011). For both algorithms, a training set of animals was created, comprising animals from straight-bred animal clusters for each breed or genetic group of Merinos based on previous pedigree-based breed proportions. This training set contained seven breeds, 497 Border Leicesters, 1902 Poll Dorsets, 21 Texels, 36 Suffolks, 339 'ultra fine' Merinos, 488 'fine-medium' Merinos and 216 'strong' Merinos, resulting in 3,499 sheep used for training and 27,626 sheep for validation. While a group of straight-bred White Suffolks were recorded, these animals did not appear to be genetically different to sheep recorded as partly White Suffolk. This breed was not included in the training animals, with animals previously assigned to this breed attributed to proportions of the other breeds by the algorithms.

Prediction accuracy was measured using Root Mean Squared Error (RMSE), based on the differences between current pedigree based breed assignments, and those estimated by the predictive algorithms. RMSE values for each breed or genetic group were calculated by $RMSE = \sqrt{(\sum_{i=1}^{n}(\hat{q}_i - q_i)^2)/n}$ where $\hat{q}_i$ is the breed proportion predicted using genomic data for the i^{th} animal, q_i is the breed proportion from the pedigree for the i^{th} animal, and n is the number of animals modelled. RMSE was calculated for each breed individually, as well as an overall value across all breeds. Algorithms producing lower RMSE values were deemed to produce more accurate estimates.

RESULTS AND DISCUSSION

Breeds could be differentiated in plots of the first two vectors of the U matrix (see Figure 1). Distinct clusters of straight-bred Border Leicester, Poll Dorset and Merino animals could be identified in the extremes of these plots from their pedigree based breed assignments and a small cluster of Suffolk animals could be identified in the Suffolk plot. Further, groups of crossbred animals could also be differentiated in between the clusters of straight-bred animals, e.g. a group of ¼ Merino and ¾ Border Leicester can be seen in the lower middle of the their respective plots. This is also true for other clusters of animals in these plots, which can be attributed visually to varying combinations of breeds.

It was also possible to identify genetic groups of Merino animals in the 4th and 5th vectors of the U matrix (see Figure 2), with differentiation of ultra-fine, fine-medium and strong sheep possible. It is evident from these plots that some sheep previously assigned to the fine-medium group may instead belong to the ultra-fine group. This would suggest that historically the fine-medium category has become a default category for sheep that have been hard to group. It can also be seen in Figures 1 and 2 that many animals appear to have been assigned to the wrong breed or genetic group.

The ADMIXTURE and BreedComp algorithms were able to predict breed and genetic group proportions based on the training animals provided to them, with BreedComp appearing to provide slightly more accurate estimates. RMSE values are presented, where possible (see Table 1), with BreedComp producing lower RMSE values for a larger number of breeds than ADMIXTURE. Caution is warranted in interpretation of these values presented here because of errors in the pedigree-based assignments. In addition, some Merino sheep of unknown type have been allocated to a default category.

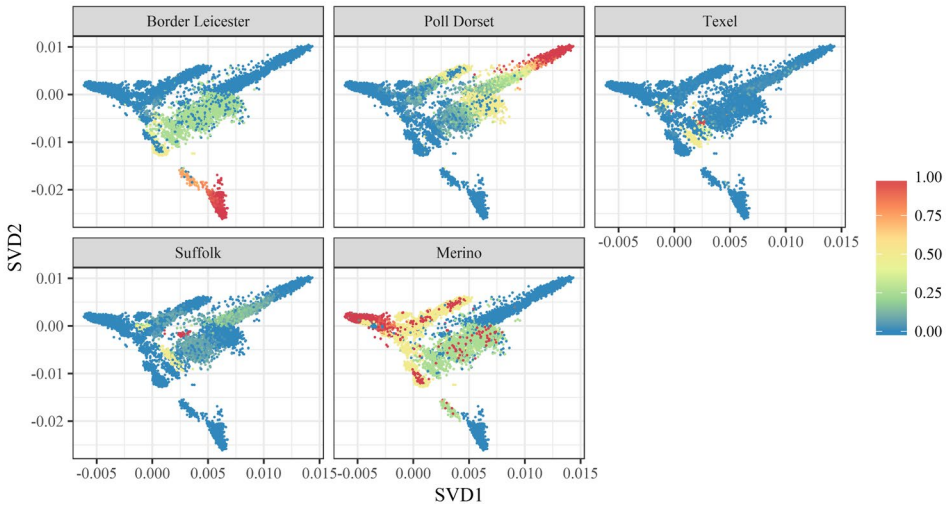

Figure 1. Pedigree based breed proportions of main sheep breeds in each genotyped animal. Red points indicate animals with 100% content of the given breed, and blue points 0%

BreedComp was able to identify all breeds and genetic groups of sheep included in the training animals. ADMIXTURE did not identify the Texel breed from the training animals, instead identifying a fourth genetic group of merinos. Some sheep were reassigned by both algorithms, for instance, many sheep were reclassified from the fine-medium Merino group to the ultra-fine group. This can be seen in the RMSE values for these categories (see Table 1) which were larger than for the other categories. Importantly for routine application, BreedComp was approximately 45 times faster than ADMIXTURE, with BreedComp running for 4.1 minutes on a single CPU core, while ADMIXTURE ran for 3.1 hours on 28 CPU cores. BreedComp appears to be slightly more accurate and faster than ADMIXTURE for these data.

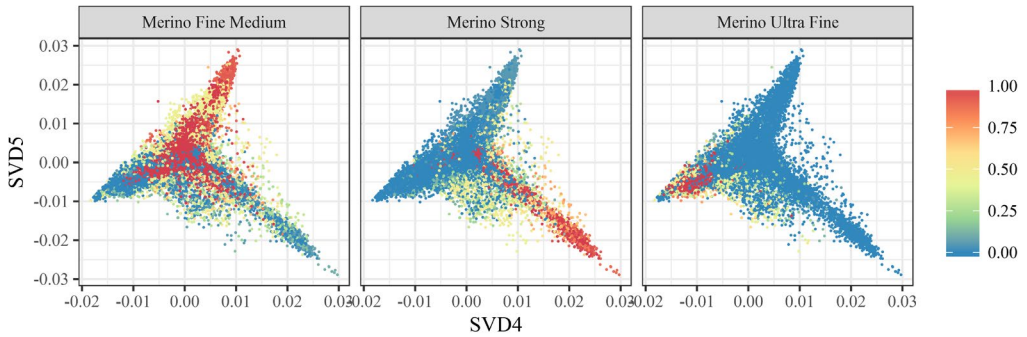

Figure 2. Pedigree based proportions of each genetic group for Merinos. Red points indicate animals with 100% content of the given breed, and blue points 0%

Table 1.RMSE for breed composition predictions for ADMIXTURE and BreedComp

Breed	ADMIXTURE[1]	BreedComp
Overall	0.181	0.170
Border Leicester	0.059	0.055
Poll Dorset	0.059	0.095
Texel	-	0.027
Suffolk	0.172	0.072
Merino (Ultra-Fine)	0.242	0.298
Merino (Fine-Medium)	0.254	0.247
Merino (Strong)	0.190	0.190

[1]ADMIXTURE was unable to identify the Texels thus a RMSE value was not calculated

Another issue with using these algorithms to estimate breed proportions is their inability to estimate breed proportions for breeds lacking genotyped straight-bred animals. For instance, a small cluster of sheep can be identified just to the upper right of the Merinos that are ½ Merino and ½ Dorper. These sheep are currently being assigned by BreedComp as ¾ Merino and ¼ Poll Dorset. Without the inclusion of straight-bred Dorper sheep in the data, these sheep cannot be correctly classified. A small cluster can also be identified that contains a portion of Coopworth.

Application of the BreedComp algorithm would allow for more accurate estimation of breeding values for Australian sheep, especially for animals without pedigree information and only genomic data. Animals previously assigned to the wrong breed or genetic group can also be reassigned, further improving genetic evaluation systems.

REFERENCES

Alexander D.H., Novembre J., Lange K. (2009) *Genome Res.* **19**:1655.

Boerner V. (2017) Proc. Assoc. Adv. Anim. Breed. Genet. **22**: these proceedings.

Chiang C.W.K., Gajdos Z.K.Z., Korn J.M., Kuruvilla F.G., Butler J.L., *et al.* (2010) *PLOS Genet.* **6**:e1000866.

Dodds K.G., Newman S.-A.N., Auvray B., McEwan J.C. (2013) Proc. Assoc. Adv. Anim. Breed. Genet. **20**:274.

Frkonja A., Gredler B., Schnyder U., Curik I., Sölkner J. (2012) *Anim. Genet.* **43**:696.

Kuehn L.A., Keele J.W., Bennett G.L., McDaneld T.G., Smith T.P.L., *et al*. (2011a) *J. Anim. Sci.* **89**:1742.

Moghaddar N., Gore K.P., Daetwyler H.D., Hayes B.J., van der Werf J.H.J. (2015) Genet. Sel. Evol. **47**:97.

Sölkner J., Frkonja A., Raadsma H.W., Jonas E., Thaller G., *et al*. (2010) *Interbull Bull.* 62.

VanRaden P.M., Olson K.M., Wiggans G.R., Cole J.B., Tooker M.E. (2011) *J. Dairy Sci.* **94**:5673.

Westell R.A., Quaas R.L., Van Vleck L.D. (1988) *J. Dairy Sci.* **71**:1310.

Yang J., Benyamin B., McEvoy B.P., Gordon S., Henders A.K., *et al*. (2010) *Nat. Genet.* **42**:565.

R Core Team (2016) R: A Language and Environment for Statistical Computing R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/.