Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 323-326

ASCERTAINING THE BREED COMPOSITION OF AN ADMIXED SHEEP POPULATION USING GENOMIC INFORMATION: A SIMULATION STUDY

S. Sahoo[1], M.H. Ferdosi[1], J.H.J. van der Werf[2] and S. de las Heras-Saldana[1]

[1] Animal Genetics Breeding Unit,[*] University of New England, Armidale, NSW 2351 Australia
[2] School of Environmental & Rural Science, University of New England, Armidale, NSW 2351 Australia

SUMMARY

The aim of this study was to evaluate the reliability of estimating the breed composition of an admixed population using simulated data. Comprehending breed composition will help to discern the animal's ancestry and to define breed groups in genetic evaluation. The simulation was performed in two steps: initially, a historical population was simulated and then a recent population comprising two subpopulations was simulated. Haplotypes of these subpopulations were coded differently and selected randomly and independently for 5 generations, followed by a cross between them in generation 6, with further inter-crossing for the last 4 generations. True breed composition (TBC) was calculated from coded haplotypes for the crossbred population and later compared with results obtained from Admixture and BreedComp algorithms. BreedComp provided more accurate breed composition estimates than Admixture in this study. These findings suggest that the approach can be extrapolated to real data to assign animals into genetic groups to improve the prediction accuracy of breeding values.

INTRODUCTION

Animal breeders strive to select the best animals to genetically improve breeding populations (Niehoff *et al.* 2024). Due to selection and admixture, many populations are heterogeneous, which needs to be accounted for in genetic evaluations. It is thereby important to identify genetic groups to account for the genetic variation in admixed populations. Genetic groups or unknown parent groups (UPGs) can be assigned to animals of unknown parentage to account for differences in genetic merit (Westell *et al.* 1988). Different methods have been implemented to define genetic groups, based on birth year, region or flocks of origin, or generation from pedigree records. However, this approach is problematic when the pedigree is incomplete (Masuda *et al.* 2022). With the advent of genomic data, a potentially powerful alternative method for breed of origin is available (Eiriksson *et al.* 2022), but little is known about the accuracy of using genomic data to define genetic groups for admixed populations.

Since genomic data became available, modified versions of best linear unbiased prediction (BLUP) such as genomic BLUP (GBLUP) and single-step GBLUP have been used to improve the estimation of breeding value (Legarra *et al.* 2009). Previous studies have reported that the complexity of the relationship matrix increased with the integration of pedigree and genomic data. Due to different base populations being used for the numerator relationship matrix (**A**) and genomic relationship matrix (**G**), incompatibility between pedigree and genomic data produced biased estimates of breeding value (Masuda *et al.* 2022). A generalised UPG, also known as the metafounders (MF) model, was developed to ensure the compatibility between pedigree and genomic relationships by adjusting **A** to match **G,** balancing and accounting for relationships between and within base populations of **A** and **G** (Legarra *et al.* 2015; Himmelbauer *et al.* 2024).

[*] A joint venture of NSW Department of Primary Industries and Regional Development and the University of New England

However, incomplete pedigree in real data makes it difficult to properly define MF, as an accurate definition of relationships is needed. In this context, accurate estimation of breed composition based on genotype data could enhance the definition of genetic groups in admixed populations and improve compatibility with pedigree data.

Gurman *et al.* (2017) compared the estimates of clustering approaches such as Admixture (Alexander *et al.* 2009) and BreedComp (Boerner and Wittenburg 2018) with the breed composition estimated using pedigree data of Australian sheep breeds in which BreedComp provided slightly better estimates than Admixture. Similarly, findings by Boerner and Wittenburg (2018) highlighted that BreedComp performed better in crossbred animals artificially generated from real genotyped data of 11 different cattle breeds. However, pedigree-based breed assignment can be inaccurate due to missing pedigree records, affecting its comparison with genomic-based breed assignment methods.

Simulation studies can be used to determine the accuracy of these methods to estimate breed composition using genomic data, as tracing the alleles passed to subsequent generations using haplotype information can aid in determining the accurate breed contribution from either parent. Although simulation studies have been used to determine breed composition with clustering algorithms (Himmelbauer *et al.* 2024), the accuracy of the clustering methods using genomic data to evaluate its efficiency is largely unexplored.

Hence, this study used simulated data to assess the reliability of estimating breed composition based on genomic information to infer the parental contribution to each animal in an admixed population.

MATERIALS AND METHODS

Simulation overview. A historical sheep population was simulated using QMSim (Sargolzaei and Schenkel 2009). The genome consisted of 26 chromosomes with 49,400 SNPs, a mutation rate of 2.5×10^{-5}, an effective population size (Ne) of 100, and a crossover interference of 25cM. A total of 2000 animals were simulated descending from 20 sires and 1000 dams.

The historical population was divided into populations A and B using custom code in R. To track back the ancestral populations, alleles in haplotypes were recoded as -1 and 1 for population A, and -2 and 2 for population B, respectively. All 1,000 dams in both populations were selected along with 54 males in population A and 26 males in population B to create a difference in effective population size between these two populations. A recombination function was added to the population using the hsphase R package (Ferdosi *et al.* 2014). Mating within populations A and B was conducted randomly and independently for 5 generations.

In generation 6, 26 sires from population B were mated with 1000 dams from population A. The resulting cross population was used to simulate four more generations of inter-crosses, using 54 sires and 1,000 dams from generations 7 to 10.

From animals in generation 5 to 10 (n = 1000 per generation), the haplotypes were recoded (0, 1) format and converted to genotypes to estimate the breed composition using supervised Admixture (K = 2) (Alexander *et al.* 2009) and BreedComp (Boerner 2017) analysis. The true breed composition (TBC) was measured based on counts of alleles at each locus coded according to the breed of origin. The accuracy of the estimates was calculated as the difference between TBC with estimates obtained from Admixture and BreedComp for the crossbred generations 6 to 10, and these differences were plotted using Python v3.11 (Python Software Foundation 2023). Population parameters, such as average inbreeding coefficient and effective population size for crossbred population, were calculated using *--het* in Plink (Purcell *et al.* 2007) and SNeP software (Barbato *et al.* 2015), respectively.

RESULTS AND DISCUSSION

Effective population size and inbreeding coefficients. Inbreeding coefficients (F) and effective population size (Ne) were estimated for generations 5 to 10 (Table 1). Ne for population A at generation 5 (264) is higher than that of population B (167) because of fewer sires being used in population B, which also explains the higher F-value in population B. In generation 6, there was an increase in Ne as a result of the cross between populations A and B, which were mated randomly for 5 generations. As F-value was calculated using observed and expected homozygotes, a negative estimate of F explained the increase in the level of heterozygosity (Purcell *et al.* 2007) as observed in generation 6. From generation 7 to 10, Ne declined in the crossbred population as it followed inter se mating within the population leading to an increase in inbreeding coefficients.

Table 1. Estimation of inbreeding coefficient (F) and effective population size (Ne) from generation 5 to 10

Generation	5 (PopA)	5 (PopB)	6	7	8	9	10
F	0.0084	0.0234	-0.024	0.0007	0.0015	0.0085	0.0116
Ne	264	167	338	329	310	308	292

Breed composition estimates. Density plot depicts the distribution for the difference between TBC and estimates from Admixture and BreedComp (Figure 1). BreedComp exhibited a lower sigma value (σ = 0.08) than Admixture (σ = 0.13) revealing that the difference between TBC and estimates of BreedComp were closer to zero as compared to Admixture results, indicating slightly more accurate estimates. Similar findings were reported by (Gurman *et al.* 2017; Boerner and Wittenburg 2018) for Australian sheep and cattle breeds respectively, where BreedComp was able to perform better than Admixture.

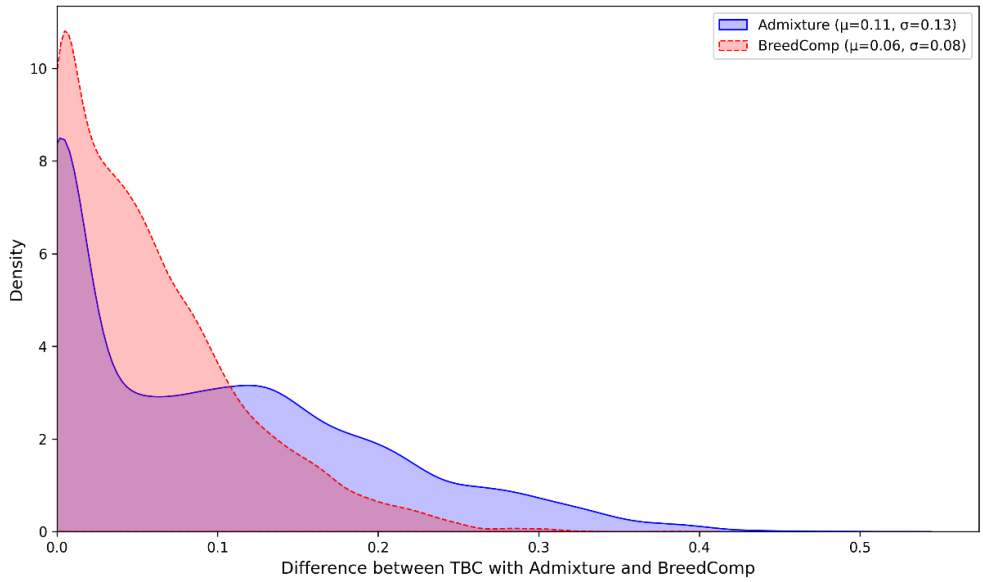

Figure 1. Density Plot illustrating the difference of TBC with estimates of Admixture and BreedComp for crossbred population

As Ne estimates were calculated on the basis of linkage disequilibrium (LD) (Barbato *et al.* 2015), a decrease in Ne estimates from generation 7 to 10 (Table 1) reflects a high level of LD in the crossbred population. Neither Admixture nor BreedComp algorithms account for LD. However, BreedComp relaxes marker independence by accounting for linear dependencies between markers, possibly explaining why BreedComp could provide better estimates of breed composition compared to Admixture results (Boerner and Wittenburg 2018).

The BreedComp algorithm can determine breed composition, enabling proper grouping of animals based on genomic data. Reliable breed composition estimates will help in clustering similar animals into MFs, which can be achieved by defining them based on genotype data. This approach helps overcome the limitation of missing and inaccurate pedigree records. Further research will be needed to investigate applying this method to estimate the breed composition of real animals based on genomic data to assign animals properly into MFs to test if this improves the accuracy of the prediction of breeding value which will enhance the overall productivity of the farm.

CONCLUSIONS

This study provided insights into the use of Admixture and BreedComp algorithms to estimate reliability for evaluating breed composition in admixed sheep populations using genomic information. BreedComp provided more accurate estimates of breed composition than Admixture in this study. The application of algorithms like BreedComp will help in precise assignment of animals to define metafounders based on genomic data, thereby improving the accuracy of breeding value in genetic evaluation.

ACKNOWLEDGMENTS

Shweta Sahoo acknowledges the financial support provided by the UNE-IPRA scholarship.

REFERENCES

Alexander D.H., Novembre J. and Lange K. (2009) *Genome Res* **19**: 1655.
Barbato M., Orozco-terWengel P., Tapio M. and Bruford M.W. (2015) *Front. Genet.* **6**: 109.
Boerner V. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22**: 97.
Boerner V. and Wittenburg D. (2018) *Front. Genet.* **9**: 185.
Eiriksson J.H., Byskov K., Su G., Thomasen J.R. and Christensen O.F. (2022) *J. Dairy Sci* **105**: 5178.
Ferdosi M.H., Kinghorn B.P., van der Werf J.H.J., Lee S.H. and Gondro C. (2014) *BMC Bioinformatics* **15**: 172.
Gurman P.M., Swan A.A. and Boerner V. (2017) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **22**: 341.
Himmelbauer J., Schwarzenbacher H., Fuerst C. and Fuerst-Waltl B. (2024) *J. Dairy Sci.* **107**: 8170.
Legarra A., Aguilar I. and Misztal I. (2009) *J. Dairy Sci.* **92**: 4656.
Legarra A., Christensen O.F., Vitezica Z.G., Aguilar I. and Misztal I. (2015) *Genetics* **200**: 455.
Masuda Y., VanRaden P.M., Tsuruta S., Lourenco D.A.L. and Misztal I. (2022) *J. Dairy Sci.* **105**: 923.
Niehoff T.A.M., Ten Napel J., Bijma P., Pook T., Wientjes Y.C.J., Hegedus B. and Calus M.P.L. (2024) *Genet. Sel. Evol.* **56**: 41.
Purcell S., Neale B., Todd-Brown K., Thomas L., Ferreira M.A., Bender D., Maller J., Sklar P., de Bakker P.I., Daly M.J. and Sham P.C. (2007) *Am. J. Hum. Genet.* **81**: 559.
Python Software Foundation. 2023. Python Language Reference, version 3.11. *Python Software Foundation*, Wilmington, DE. Available at https://www.python.org
Sargolzaei M. and Schenkel F.S. (2009) *Bioinformatics* **25**: 680.
Westell R.A., Quaas R.L. and Van Vleck L.D. (1988) *J. Dairy Sci.* **71**: 1310.