Proc. Assoc. Advmt. Anim. Breed. Genet. 26: 90-93

EVALUATING THE ACCURACY OF SNP-BLUP AND ADMIXTURE FOR BREED COMPOSITION PREDICTION IN LIMITED AND WELL-REPRESENTED CATTLE BREEDS

C.A. Ryan

Teagasc, Moorepark, Co. Cork, Ireland

SUMMARY

The present study provides the first comprehensive comparison of SNP-BLUP and Admixture for predicting breed composition in datasets comprising both underrepresented and well-represented cattle breeds. When analysing 6 breeds with small reference populations, both SNP-BLUP and Admixture demonstrated high accuracy (≥98% correct assignment) with minimal mean absolute difference (0.01) in predicted proportions. However, when the analysis included all 19 breeds (13 well-represented and 6 breeds with limited representation) the accuracy of SNP-BLUP decreased to 82%, while Admixture maintained 99% accuracy. The performance gap was largely resolved by reducing the training population size of well-represented breeds to match the breeds with limited representation (50 animals per breed). This adjustment increased the accuracy of SNP-BLUP to 93%. These results demonstrate that unequal training population sizes negatively impact SNP-BLUP's performance due to its statistical assumptions, while Admixture remained robust regardless of training population balance.

INTRODUCTION

An animal's breed composition is typically assumed to reflect the average of its parents' breed proportions. However, due to chromosomal recombination during gametogenesis, offspring from crossbred parents may deviate from these expectations. Several tools have been developed to estimate breed composition, including Admixture (Alexander *et al.* 2009), and SNP-BLUP (Single Nucleotide Polymorphism Best Linear Unbiased Prediction), which is commonly used in genomic evaluations. These methods employ fundamentally different statistical approaches. Admixture uses a likelihood-based approach to estimate ancestry proportions by modelling allele frequencies across populations. In contrast, SNP-BLUP is based on the infinitesimal model, where SNP effects are assumed to be small, random, and drawn from a shared variance structure. Given the widespread use of SNP-BLUP in genomic evaluations, its applicability to predicting breed composition offers the advantage of utilising existing pipelines, potentially streamlining the process and enhancing computational efficiency compared to standalone software tools**.** Despite these methodological differences, only two previous studies have compared SNP-BLUP and Admixture for breed composition prediction, and both were limited to specific population structures; Ryan *et al.* (2023) focused on well-represented breeds with large training populations (500 purebred animals per breed) and Struken *et al.* (2017) focused on uniformly small populations (< 60 purebred animals per breed). Therefore, a critical knowledge gap exists regarding the comparative performance of SNP-BLUP and Admixture in datasets containing breeds with limited numbers as well as well-represented breeds. The objective of this study was to evaluate the accuracy of SNP-BLUP for predicting the breed composition of 6 breeds of cattle with limited numbers of purebreds and to compare its performance to that of Admixture. An additional objective was to investigate the impact of including data from 13 well-represented breeds on the accuracy of breed composition predictions for breeds with limited numbers.

MATERIALS AND METHODS

Genotypes for 49,213 SNPs were available post-quality control for a dataset of 19 breeds.

Purebred animals were verified using principal component analysis and unsupervised Admixture analysis (breed composition > 0.9). For the 13 well-represented breeds, the dataset from Ryan *et al.* (2023) was used, consisting of 500 animals per breed in the training population, 3,146 purebred validation animals, and 4,330 crossbred animals. For the additional 6 breeds with limited purebred representation (Piedmontese, Dexters, Montbeliarde, Irish Maol, Jersey, and Romagnola), training population sizes ranged from 22 (Jersey) to 264 (Dexter), with a total of 652 animals, and a purebred validation population of 333 animals. An additional crossbred validation population of 228 animals from these 6 breeds was also included.

Breed composition was estimated using SNP-BLUP in the MIX99 software (MiX99 Development Team, 2022) and Admixture following the methodology outlined in Ryan *et al.* (2023). Analyses were conducted first using only the 6 breeds with limited numbers (K = 6), followed by a combined analysis of all 19 breeds (K= 19). Breed assignment was considered accurate when the predicted proportion for an animal in the purebred validation population was ≥0.90 for a specific breed. The differences in the main breed proportion estimates for crossbred animals predicted using SNP-BLUP and Admixture were compared. To investigate the impact of training population size imbalance, subsets of 200 and 50 animals were randomly sampled from the 13 well-represented breeds to match the smaller training populations of the rare breeds.

RESULTS AND DISCUSSION

When the 6 breeds with limited representation were analysed independently, both SNP-BLUP and Admixture accurately assigned ≥98% of the purebred validation population to their respective breeds (Figure 1). The mean absolute difference between the Admixture and SNP-BLUP breed proportion estimates was minimal (0.011 ± 0.03), consistent with previous studies using balanced training populations consisting of 500 animals per breed (Ryan *et al.* 2023) or uniformly small training populations of < 60 animals per breed (Strucken *et al.* 2017). This demonstrates that both methods perform well when analysing breeds with similar representation levels.

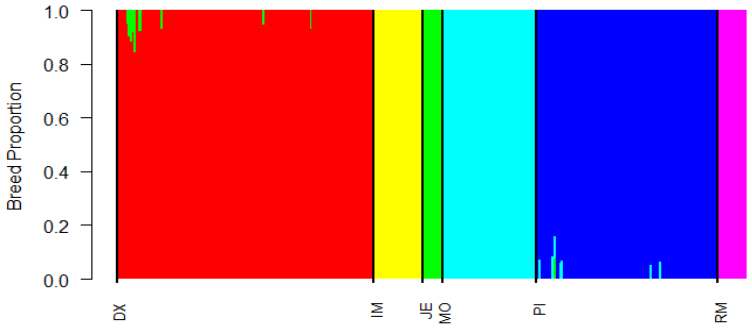

Figure 1. Estimated breed proportions from Admixture for each purebred individual in the validation population for K=6. Each animal is represented by a thin vertical line and each colour represents one inferred population. The breed proportion of each animal is represented by the length of each colour in that animal's vertical bar. Breeds include Dexters (DX), Irish Maol (IM), Jersey (JE), Montbeliarde (MO), Piedmontese (PI), and Romagnola (RM)

However, the mean absolute difference between SNP-BLUP and Admixture breed composition predictions for the purebred validation population increased to 0.077 when 13 additional, larger breeds (from Ryan *et al.* 2023) were included in the analysis (K=19). Admixture maintained a high overall accuracy, correctly assigning 98% of purebred validation animals, while SNP-BLUP correctly assigned only 82% (Table 1). For crossbred animals, the discrepancy between methods was larger, with a mean absolute difference (standard deviation) of 0.12 (±0.15).

The greatest discrepancies occurred for Shorthorns (mean absolute difference = 0.24) and Simmentals (mean absolute difference = 0.50). SNP-BLUP failed to correctly assign any purebred Shorthorn or Simmental animals, instead estimating their average breed proportions as 0.76 and 0.50, respectively. This misclassification was asymmetric: Shorthorns were predominantly assigned as part Shorthorn and part Irish Maol, and Simmentals as part Simmental and part Montbéliarde. This bias likely stems partly from the genetic similarity between the breeds but also from SNP-BLUP's modelling assumptions. As a linear mixed model, SNP-BLUP applies shrinkage to SNP effects, regressing estimates toward a genomic mean that is disproportionately influenced by well-represented breeds due to their larger contribution to allele frequency estimates (Meuwissen *et al.* 2001). In imbalanced datasets, this shrinkage may potentially create a systematic misclassification pattern where animals from well-represented breeds (e.g., Shorthorn, Simmental) are erroneously assigned partial ancestry to breeds with limited representation (e.g., Irish Moiled, Montbéliarde). This may occur through two interacting mechanisms. First, shared haplotype segments between breeds are preferentially attributed to the less-represented breed because their smaller training populations provide weaker constraints on the shrinkage process. Second, breed-distinguishing SNPs for well-represented breeds have their effects disproportionately shrunk toward zero due to the overwhelming influence of numerically dominant breeds on the genomic mean. In contrast, breeds with limited representation likely maintain more stable predictions because their sparse representation minimises their contribution to the global genomic mean, and their unique alleles, being rare in the overall population, are less affected by shrinkage and thus retain stronger predictive value. This potentially explains why Admixture, which models population-specific allele frequencies without shrinkage, maintains high accuracy regardless of training population structure. Admixture successfully assigned 99.6% of Shorthorns and 100% of Simmentals correctly.

Table 1. The percentage of animals within each breed correctly assigned (i.e., estimated breed proportion for a specific breed was predicted to be ≥0.90) to their respective breeds using SNP-BLUP and Admixture when all 19 breeds were included in the analysis

Breed	SNP-BLUP	Admixture
Angus	99.4	100
Aubrac	100	100
Blonde d'Aquitaine	100	99.5
Belgian Blue	99.0	100
Charolais	99.7	100
Dexter	99.2	97.7
Friesian	94.8	98.7
Hereford	99.2	100
Holstein	98.1	74.7
Irish Maol	96.1	100
Jersey	100	100
Limousine	79.9	100
Montbeliarde	92	100
Piedmontese	93.7	97.9
Parthenaise	100	100
Romagnola	86.6	100
Saler	98.6	100
Shorthorn	0	99.5
Simmental	0	100

To test whether training population imbalance was indeed the primary cause of SNP-BLUP's performance decline, the size of the training population for each of the 13 well-represented breeds

was systematically reduced. When the training population of the well-represented breeds was reduced to 250 animals per breed, Shorthorns and Simmentals were still assigned as part Irish Maol and Montbéliarde, respectively. However, when training populations were equalised at 50 animals per breed, SNP-BLUP's overall accuracy improved to 93%, confirming that unequal training population sizes, not genetic similarity, were the primary cause of misclassification. Notably, Admixture's accuracy remained stable regardless of training population structure, consistent with its likelihood-based framework.

These findings have important practical implications for genomic analyses. While SNP-BLUP offers computational efficiency and integration with existing genomic evaluation pipelines, the study herein demonstrate that similar training population sizes are essential, the size of which will be limited by the breed with the smallest representation. However, it should be noted that Admixture has its own limitations, including potential sensitivity to input file order (Crum *et al.* 2019) and greater computational demands.

CONCLUSION

The present study provides the first systematic evaluation of breed composition prediction methods in datasets containing both breeds with limited representation and well-represented cattle breeds. While SNP-BLUP and Admixture performed comparably for breeds with small reference populations in isolation (≥98% accuracy), SNP-BLUP's accuracy declined substantially (to 82%) when well-represented breeds were added due to training population imbalance. This performance gap was largely resolved by equalising training population sizes (93% accuracy with 50 animals per breed), demonstrating that SNP-BLUP's limitations stem from statistical assumptions rather than intrinsic genetic relationships. In contrast, Admixture maintained 99% accuracy regardless of dataset structure. These results provide clear guidance for breed prediction methodologies in different scenarios: (1) for balanced datasets or when computational efficiency is paramount, SNP-BLUP performs adequately; (2) for imbalanced datasets where maintaining large reference populations is valuable, Admixture provides more robust results despite greater computational demands.

ACKNOWLEDGEMENTS

This publication has emanated from research supported by the European Commission in the frame of the Horizon 2020 INTAQT project (INnovative Tools for Assessment and Authentication of chicken meat, beef and dairy products' QualiTies, Grant agreement ID: 101000250).

REFERENCES

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

Crum T.E., McHugo G.P., Dover M.J. and MacHugh D.E. (2019) *PLoS ONE* **14**: e0215912.

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genetics* **157**: 1819.

MiX99 Development Team (2022) MiX99: A software package for solving large mixed model equations. Release I/2022. Natural Resources Institute Finland (Luke).

Ryan C.A., Berry D.P., O'Brien A., Pabiou T. and Purfield D.C. (2023) *Front. Genet.* **14**: 1125678.

Strucken E.M., Al-Mamun H.A., Esquivelzeta-Rabell C., Gondro C., Mwai O.A. and Gibson J.P. (2017) *Genet. Sel. Evol.* **49**: 1.