Proc. Assoc. Advmt. Anim. Breed. Genet. 23:71-74

GENOMIC PREDICTION IN A NUMERICALLY SMALL SHEEP BREED POPULATION USING IMPUTED SEQUENCE VARIANTS

N. Moghaddar[1,2], D.J. Brown[2,3], A.A. Swan[2,3], I.M. MacLeod[4] and J.H.J. van der Werf[1,2]

[1]Animal Science, ERS, University of New England, Armidale, NSW 2351, Australia
[2]CRC for Sheep Industry Innovation, Armidale, NSW 2351, Australia
[3]Animal Genetics & Breeding Unit*, University of New England, Armidale, NSW, 2351 Australia
[4]School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3083, Australia

SUMMARY

The accuracy of genomic prediction for a numerically small sheep breed was investigated based on a large multi-breed admixed reference set using moderate or high density SNP genotypes, imputed whole genome sequence genotypes or selected sequence variants based on a genome wide association study (GWAS). Reference set with weight and eating quality phenotypes was divided into a GWAS sub set (n=4,000), a training set (n=13,466 to 38,098) and a validation set with data of 143 to 169 purebred Dorper sheep. Genomic BLUP was used to estimate genomic breeding values and prediction accuracy was evaluated in the validation set based on the correlation between GBV and corrected phenotypes. Results showed a prediction accuracy between 20% and 30% based on 50k genotypes across different trait, which increased on average by 2.5% to 7.0% by using HD genotypes or selected sequence variants derived from an independent GWAS.

INTRODUCTION

Genomic prediction has been successfully implemented in breeding programs of the main livestock species. In numerically small breeds, it is difficult to establish a reasonably large reference population and prediction based on other main breeds was shown to be of limited value, (Kachman *et al*. 2013; Moghaddar *et al*. 2014). Low GBV predictability from other breeds would be partly because of low linkage disequilibrium (LD) across breeds between genetic markers and the causative mutation, a different distribution of QTL effect and QTL frequency between breeds, or due to genotype by background genotypes interaction. The problem of low LD maybe overcome when using denser marker sets or whole genome sequence (WGS) variants in genomic prediction. This study evaluated the accuracy of genomic prediction for growth and eating quality traits in purebred Dorper sheep based on a large multi-breed admixed sheep reference population, and to compare predictions based on common 50k or HD SNP genotypes, imputed WGS genotypes or using selected sequence variants based on an association study.

MATERIALS AND METHODS

Phenotypes and Animals. Data on post weaning weight (PWT), carcass scanned fat (CCFAT)and eye muscle depth (CEMD), intramuscular fat (IMF) and shear force at 5 days aging (SF5) recorded in research and industry flocks between 1999 and 2017 were used in this study. Figure 1 shows the genetic diversity of the sheep breeds used in this study as a plot of the first versus the second principal component derived from a genomic relationship matrix (GRM). Phenotypes were corrected for fixed environmental effects separately for research and industry animals. The fixed effects of the model were flock, year, sex, management groups, birth and rearing type, age of dam, age at and weight at measurement (for scanned traits). Random maternal effects were fitted for post weaning weight. Corrected phenotypes from research and industry data were combined and then corrected for source

* A joint venture of NSW Department of Primary Industries and the University of New England

of data (research/industry) and random effect of breed proportion derived from a multi generation pedigree using ASReml 3.0 (Gilmour *et al.* 2009). Between 143 and 169 purebred Dorper sheep with phenotypes and genotypes were used as validation set to represent a numerically small breed. Two data subsets were formed for a genome wide association study (GWAS); n=4000, either randomly assigned or selected based on possible higher relationship to the validation set. The rest of population (between 17,466 and 42,098 across different traits) was used as genomic prediction training set.

Genotypes. Animals were genotyped with the Illumina 50k-ovine (~70%) or 12k-ovine SNP panel (~30%), which yielded a final 44,101 and 11,377 SNP per animal respectively. Genotypes were imputed to HD genotypes based on 2,266 animals as reference set and then to WGS based on 726 animals as reference set. The final set was comprised of 31,154,249 SNP and InDels. Selection of sequence variants was based on significant SNP (–Log Pvalue ≥ 3.5) in GWAS performed on sequence data and then pruned locally for high LD (≥0.95). Association analysis was based on regression of corrected phenotypes on single sequence variant in linear mixed model (LMM) using Gemma V0.96 (Zhou and Stephens 2012).

Genomic prediction. GBV were calculated based on GBLUP with MTG2 2.02 (Lee *et al*. 2016) using the following SNP arrays: 1) 50k (44,101) genotypes, 2) HD (452,998) genotypes, 3) WGS (30,724,780) and 4) 50k and selected sequence variants (2,583-2,865). The following model was used to estimate variance components and genomic breeding values in scenarios 1, 2 and 3: y=Xb + Za + e, where y is a vector of corrected phenotypes, b is a vector of fixed effect (only mean), a is a vector of random additive genetic effects and e is a vector of random residual effects. X and Z are incidence matrices that relate fixed and additive genetic effects to phenotypes respectively. The additive genetic effects were assumed to be normally distributed with a covariance structure based on the GRM derived from the respective SNP panels. The genomic prediction model in scenario 4 was based on fitting two genetic component simultaneously, with covariance structure based on a GRM from 50k genotypes and selected variants, respectively. Accuracy of genomic prediction in purebred Dorper sheep was evaluated based on Pearson correlation coefficient between GBV and corrected phenotypes in the validation set divided by the square root of the trait's heritability.

RESULTS AND DISCUSSION

Slightly higher heritability, but consistent across different traits, was observed based on imputed HD genotypes and imputed sequence data compared to 50k genotypes (Table 1). Higher heritability is related to stronger LD between markers and QTLs and better estimation of realized genetic relationship.

The sum of the heritability based on fitting two random components simultaneously was on average similar to heritability estimates based on 50k or HD genotypes. Figures 2 and 3 compare the accuracy of genomic prediction for Dorper sheep according to using 50k or imputed HD genotypes, imputed WGS variants and 50k SNPs plus selected imputed WGS variants, respectively. Results show a higher accuracy of genomic evaluation by including the effect of selected sequence variants in the prediction model as an additional random effect. The extra accuracy was on average 0.065 and 0.077 higher when fitting selected sequence variants from a random or selected GWAS population, respectively. SF5 and IMF showed the highest increase in prediction accuracy; 0.11 and 0.09 when using selected variants derived from random or selected GWAS populations, respectively. Accuracy of genomic evaluation from using all called sequence variants (~$31x10^6$ variants) was not consistently higher than 50k genotypes. SF5 showed an increase of 0.05 and the prediction accuracy was equal or even lower than 50k genotypes. Prediction from imputed HD genotypes was more accurate (2.4%) compared to prediction using 50k genotypes in most cases except for PWT and IMF. Results show a base of between 20% and 32% genomic prediction accuracy on growth and eating quality traits using 50k genotype data for Dorper sheep based on the use of a large multi-breed reference population (13,466

to 38,098). This base prediction accuracy was expected and would be related to the use of the large multi-breed reference set which includes breeds that are genetically close to Dorper sheep (Figure.1).

Table 1. Heritability (h^2) estimates based on 50k, HD, WGS and 50k and Selected Sequence variants for different traits

Trait	No of Records	h^2,50k	h^2,HD	h^2,WGS	h^2(50k,Sel_SNPs)
Post Weaning Weight (PWT)	38,098	0.182	0.182	0.184	0.174, 0.04
Carcass Scanned Fat (CCFAT)	14,369	0.185	0.214	0.229	0.163,0.06
Carcass Eye Muscle Depth (EMD)	14,507	0.148	0.151	0.149	0.135,0.02
Intra Muscular Fat (IMF)	13,466	0.404	0.434	0.455	0.412,0.03
Shear Force day5 Aging (SF5)	14,394	0.172	0.178	0.196	0.146,0.03

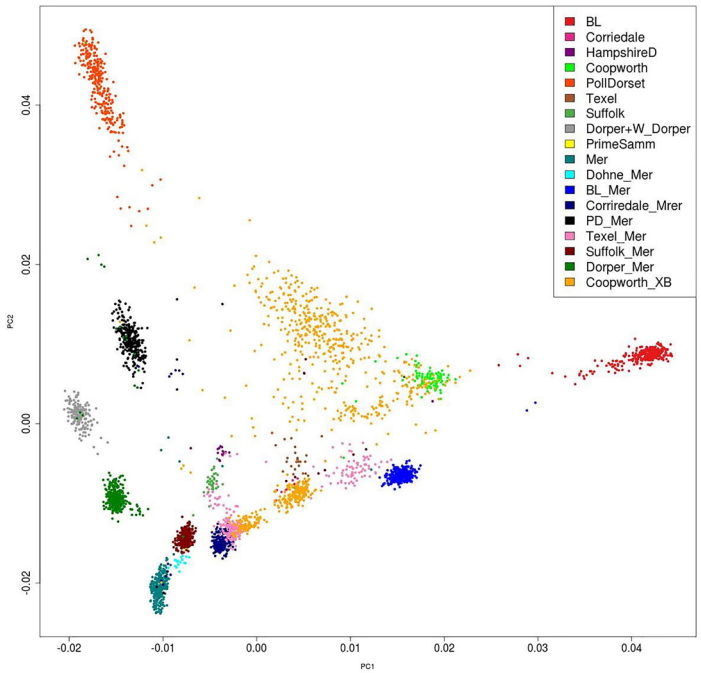

Figure 1. Genetic diversity of the sheep breeds as a plot of the first vs second principal components

Improvement in prediction accuracy by using selected sequence variants in the current study is in similar range to previous study in main sheep breeds (Moghaddar *et al.* 2018) and is in line with the results of studies on multi-breed dairy cattle. In dairy cattle, Van den Berg *et al.* (2016) showed on average up to 7% higher genomic prediction reliabilities (R^2) across milk yield, protein and fat from a multi-breed reference population. Brøndum *et al.* (2015) reported up to 5% improvement in genomic prediction reliability on a range of production traits in multi-breed dairy cattle based on including selected sequence data from GWAS in GBLUP. Using a complete set of imputed WGS a marginal, zero or even some drop in GBV accuracy observed. This is because WGS provide a very large amount of genetic markers of which a small subset would be at or in high LD with causative

mutations. Majority of these imputed sequence variants would not be able to capture genetic variance and their contribution would be limited to capturing the family relationships between animals, which would be similar or slightly higher to the relationship captured by 50k genotypes. Similar results of no improvement in prediction accuracy from using all the sequence variants data have been reported in Holstein-Friesian dairy cattle (VanRaden *et al.* 2015).

The extra prediction accuracy based on selected variants derived from a GWAS subset that used data from animals closely related to the target breed appears to be slightly higher (2% on average) than using a random GWAS subset. The differences may be not statistically significant and requires more verification in further studies, particularly based on larger GWAS populations. However, higher accuracy would be related to probably larger proportion of SNPs derived from a more related GWAS subset in association with gene that segregate in target breed. This indicates that while multi-breed GWAS population is more powerful to find larger numbers of causal genomic regions (Duijvesteijn *et al.* 2018; van der Berg *et al.* 2016), our study showed more genetically related GWAS population to target population is preferable to obtain more accurate genomic breeding values. The GWAS results, which showed there are some significant genomic regions limited to a random or a selected GWAS subsets, support these results.

CONCLUSIONS

Genomic prediction accuracy for a numerically small breed population increased by 2.5% and 7% based on using imputed high-density marker genotypes and imputed sequence variants derived in an independent population respectively. Selection of sequence variants from a genetically more related population was in favour of higher genomic prediction accuracy in small breed populations.

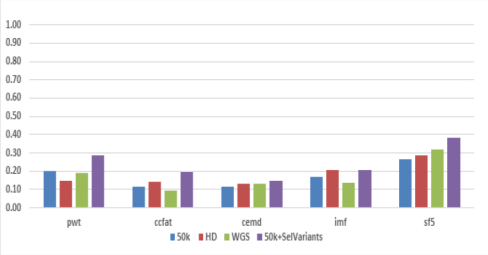

Figure 2. Accuracy of genomic prediction from 50k, HD and using selected SNPs from random GWAS set

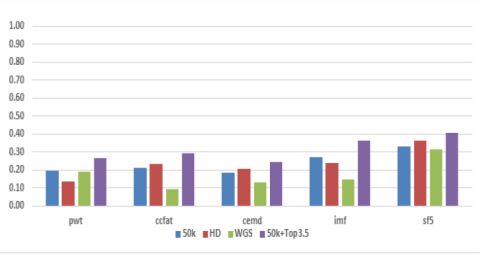

Figure 3. Accuracy of genomic prediction using 50k, HD, WGS and selected SNPs from selected GWAS set

REFERENCES

Brøndum R.F., Su G., Janss L. and Sahana G. (2015) *J. Dairy Sci*. **98**: 4107.

Kachman S., Spangler M., Bennett G., Hanford K., Kuehn L., Snelling W., Thallman R., Saatchi M., Garrick D., Schnabel R., Taylor J. and Pollak E. (2013) *Genet. Sel. Evol*. **45**: 30.

Gilmour A.R., Gogel B.J., Cullis B.R. and Thompson R (2009). VSN International Ltd; 2009.

Lee, S.H. and van der Werf, J.H.J (2016). Bioinformatics 32, 1420.

Moghaddar N., Swan A.A. and van der Werf J.H.J. (2014) *Genet. Sel. Evol*. **46**: 58.

Van den Berg I., Boichard D. and Lund M.S. (2016) *Genet. Sel. Evol*. **48**: 83.

VanRaden P.M., Tooker M.E., O'Conell J.R., Cole J.B. and Bickhart D.M. (2017) *Genet. Sel. Evol.* **49**:32.

Zhou X. and Stephens M. (2012) *Nature Genetics*. **44**: 82.