Contributed paper

EVALUATING THE BENEFITS OF INCLUDING PREDICTIVE SNP MARKERS IN SINGLE STEP EVALUATION IN SHEEP USING CROSS-VALIDATION

L. Li[1], P.M. Gurman[1], A.A. Swan[1], N. Moghaddar[2] and J.H.J. van der Werf [2]

[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

A SNP array of 50k SNP markers was used in single-step GBLUP (SS-GBLUP) models to estimate breeding values in the Australian sheep genetic evaluation system. In 2019, Neogen launched a new GeneSeek Genomic Profiler Ovine 50k chip, which included ~5000 SNPs that were identified based on Sheep CRC research as highly predictive for growth, carcass and eating quality traits. The objective of this work was to apply a five-fold cross-validation approach to compare different models for the use of predictive SNPs for post-weaning weight (PWT), carcass eye muscle depth (CEMD), carcass fat at C site (CCFAT), intramuscular fat (IMF) and shear force (SF5) based on the LAMBPLAN terminal sire genetic evaluation. Correlation and regression coefficients between adjusted phenotypes and SS-GBLUP EBVs for validation animals from the different models were calculated. The results indicated that adding predictive SNPs slightly improved the correlation and regression coefficient of EBVs, but there was no advantage in giving them more weight via a separate term in the model, confirming that the current industry evaluation model using a single genomic relationship matrix is the best of the tested models for these traits.

INTRODUCTION

Single-step genomic BLUP (SS-GBLUP) procedures have been implemented in the Australian sheep genetic evaluation system since 2017 (Brown *et al.* 2018). Prior to 2020, the genomic relationship matrix (GRM) used in SS-GBLUP analyses was built using an ovine 50k panel of common SNPs. Recent genome-wide association studies have identified ~5000 predictive SNP markers for carcass and eating quality traits in sheep (Moghaddar *et al.* 2019). In 2019, Neogen launched a GeneSeek Genomic Profiler Ovine (GGP) 50k panel, which included these predictive SNPs. To accommodate these markers, the set of SNPs used in routine genetic evaluations was modified to be the union of all SNPs included on all panels used for sheep genotyping, resulting in a set of 60,410 SNPs. This set of SNPs was then implemented in the sheep SS-GBLUP analyses in a single genomic relationship matrix (GRM) from 2020. However, this method assumes equal weighting for all SNPs. An alternative approach is to use an additional term in the model, using a separate GRM based on predictive SNPs, effectively giving them more weight to those SNPs. In this study, models with one or two GRMs fitted in the SS-GBLUP model for the calculation of breeding values were investigated using a five-fold cross-validation approach. The correlation and regression of SS-GBLUP EBVs with adjusted phenotypes from the different models were compared.

MATERIALS AND METHODS

Phenotype data. This study was conducted using data from the LAMBPLAN terminal sire industry evaluation, due to the new predictive SNPs targeting growth, carcass and eating quality traits. The data consisted of records from animals measured for the main slaughter traits in the Sheep

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

CRC Information Nucleus Flock (van der Werf *et al*. 2010) and the MLA Resource Flock databases which are used in the industry evaluation. Phenotypes were pre-adjusted for a combination of birth type, rearing type, age, and age of dam, depending on the trait. Five traits from two data sets were used in SS-GBLUP analyses to estimate breeding values for cross-validation (Table 1). The first data set included 9688 animals that had all five traits observed as well as SNP genotype information (the "small data set"). To investigate whether the extra phenotypes from ungenotyped animals affected the cross-validation results for those genotyped animals, the second data set extended the small data set by including all ungenotyped animals with at least one trait observed for any of the five traits in the analysis (the "large data set"). A summary of the two data sets is presented in Table 1. Pedigree information was extracted from the LAMBPLAN database and included 44,874 and 1,985,749 animals for the small and large data sets, respectively.

Table 1. Traits (units), number of animals (N), mean and standard deviation (sd) for the small (animals with all phenotypes and genotypes) and large (all animals including ungenotyped animals with at least one phenotype) data sets in this study

Trait	Unit	Small data set			Large data set		
		N	mean	sd	N	mean	sd
Post-weaning weight (PWT)	kg	9688	58.58	9.47	1,674,789	58.00	9.71
Carcass eye muscle depth (CEMD)	mm	9688	31.31	3.87	16,753	31.43	3.77
Carcass fat at C site (CCFAT)	mm	9688	4.13	1.96	16,560	4.63	2.48
Intramuscular fat (IMF)	%	9688	4.24	0.99	14,832	4.35	1.04
Shear force (SF5)	Newtons	9688	34.88	15.22	14,840	34.24	15.16

The five-folds subsets derived from the 9688 genotyped animals were used as the cross-validation data set for SS-GBLUP analyses. Animals were crosses between terminal sire breed rams and Merino ewes or Border Leicester x Merino ewes. The main ram breeds represented were White Suffolk (323 sires,3801 progeny), Poll Dorset (319 sires, 4080 progeny), Suffolk (40 sires, 499 progeny), White Dorper (35 sires, 309 progeny), Texel (31 sires, 413 progeny) and Dorper (29 sires, 235 progeny). Five-fold subsets were randomly allocated stratified by ram breeds and sire families with five replicates with the average number of sires and progeny ranging from 161 to 167 and from 1679 to 2043 for each subset, respectively.

Genomic data. Three sets of SNPs were used in this study: unselected (random) SNPs (55,709), the predictive SNPs (4,701) and the combined set (60,410). The first set was a combination of the original ISAG 50k sheep panel and the additional random SNPs from the Neogen GGP 50k, where the actual number of SNPs used is the set remaining after applying quality control measures. The predictive 4,701 SNPs (Moghaddar *et al*. 2019) were those originating from the CRC research that were then commercialised on the GGP 50k. Genomic relationship matrices (GRMs) were constructed based on these SNP sets, using the implementation of the breed-adjusted GRM as described by Gurman *et al*. (2019) and as implemented in the LAMBPLAN terminal sire SS-GBLUP analysis. Three genomic relationship matrices were calculated: $\mathbf{G}_r$, based on the random SNPs; $\mathbf{G}_p$, based on the predictive SNPs and $\mathbf{G}_{rp}$, based on the combined set.

Models. The multivariate linear mixed model used for estimating breeding values was $\mathbf{Y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Zt} + \mathbf{e}$, where **Y** is data in the multivariate form; **Xb** is the fixed contemporary group effects (defined as combinations of the management group, flock, year, sex, breed type and date of measurement); **ZQg** is the random genetic group effects; **Zt** represents combined effects of breeding values based on pedigree and genomic effects from different SNP sets, and **e** is residuals. Maternal effects were included as permanent environment effects for PWT. Four combinations of polygenic and genomic effects were compared to identify appropriate models: 1) A model: $\boldsymbol{Zt} = \boldsymbol{Za}$; 2) A+G$_r$ model: $\boldsymbol{Zt} = \boldsymbol{Za} + \boldsymbol{Zu_r}$; 3) A+G$_{rp}$ model: $\boldsymbol{Zt} = \boldsymbol{Za} + \boldsymbol{Zu_{rp}}$; and 4) A+G$_r$+G$_p$ model: $\boldsymbol{Zt} = \boldsymbol{Za} + \boldsymbol{Zu_r} + \boldsymbol{Zu_p}$, where $\mathbf{a}$, $\mathbf{u}_r$, $\mathbf{u}_p$ and $\mathbf{u}_{rp}$ are N($\mathbf{0}$, $\boldsymbol{A}\otimes\boldsymbol{\Sigma_a}$), N($\mathbf{0}$, $\mathbf{H}_{\boldsymbol{r}}\otimes\boldsymbol{\Sigma}_{g_r}$), N($\mathbf{0}$, $\mathbf{H}_p\otimes\boldsymbol{\Sigma}_{g_p}$) and N($\mathbf{0}$, $\mathbf{H}_{rp}\otimes\boldsymbol{\Sigma}_{g_{rp}}$) respectively, with $\mathbf{H}_r$, $\mathbf{H}_p$ and $\mathbf{H}_{rp}$ matrices derived from combining the genomic relationship matrixes $\mathbf{G}_r$, $\mathbf{G}_p$ and $\mathbf{G}_{rp}$ with pedigree relationship matrix **A**, respectively. $\boldsymbol{\Sigma_a}$, $\boldsymbol{\Sigma}_{g_r}$, $\boldsymbol{\Sigma}_{g_p}$, and $\boldsymbol{\Sigma}_{g_{rp}}$ are the multivariate genetic variance-covariance matrices due to those corresponding relationship matrices as estimated by Gurman *et al.* (2021).

The average accuracy of the different models was assessed by the correlation coefficient between EBVs and phenotypes adjusted for contemporary group effects (solutions from the same models with the full data set) for the animals in the test set which were removed from the analysis. Note that correlations were presented without scaling by heritability. The bias was evaluated based on the regression coefficient of adjusted phenotype on EBVs. This process was repeated for all five cross-validation sets.

RESULTS AND DISCUSSION

The average correlation and regression coefficient for validation animals across the five cross replicates from cross-validation are shown in Table 2 for the small data set and in Table 3 for the large data set. Results from both data sets show that the average correlation increased by the largest amount when adding genomic information, from model A to model A+G$_r$, with much greater improvement for carcass and eating quality traits (17.6 ~ 43.5% increase) than growth traits (5.3 ~ 7.9 % increase for PWT). The correlation was also generally higher in the large data set compared to the small data set. There were small improvements in correlation when adding predictive SNPs in the combined GRM, from A+G$_r$ to A+G$_{rp}$, but no apparent benefit was observed in fitting predictive SNPs in a separate GRM in model A+G$_r$+G$_p$. The results confirm that the current LAMBPLAN model (A+G$_{rp}$), including predictive SNPs in a combined GRM is an appropriate solution to exploit the additional benefits of these SNPs.

Table 2. Average correlation and regression coefficients for validation animals for post-weaning weight (PWT), carcass eye muscle depth (CEMD), carcass fat at C site (CCFAT), intramuscular fat (IMF), and shear force (SF5) for models A, A+G$_r$, A+G$_{rp}$ and A+G$_r$+G$_p$ across 5 replicates for the small data set

Models	PWT	CEMD	CCFAT	IMF	SF5
			Correlation		
A	0.38	0.17	0.17	0.23	0.18
A+G$_r$	0.40	0.20	0.23	0.33	0.23
A+G$_{rp}$	0.41	0.21	0.24	0.36	0.25
A+G$_r$+G$_p$	0.40	0.19	0.22	0.34	0.23
			Regression coefficient		
A	0.97	1.01	0.90	0.92	0.92
A+G$_r$	0.93	0.97	0.98	1.10	1.01
A+G$_{rp}$	0.94	0.99	1.00	1.15	1.03
A+G$_r$+G$_p$	0.93	0.83	0.88	1.08	0.91

[1] Standard deviation for correlation and regression coefficients ranged from 0.002 to 0.008

Table 3. Average correlation and regression coefficients for validation animals for post-weaning weight (PWT), carcass eye muscle depth (CEMD), carcass fat at C site (CCFAT), intramuscular fat (IMF), and shear force (SF5) and for models A, A+G$_r$, A+G$_{rp}$ and A+G$_r$+G$_p$ across 5 replicates for the large data set

Models	PWT	CEMD	CCFAT	IMF	SF5
			Correlation		
A	0.38	0.19	0.22	0.31	0.20
A+G$_r$	0.41	0.23	0.27	0.39	0.25
A+G$_{rp}$	0.41	0.24	0.28	0.41	0.26
A+G$_r$+G$_p$	0.41	0.22	0.26	0.39	0.24
			Regression coefficient		
A	0.87	0.91	0.95	0.93	0.96
A+G$_r$	0.81	0.82	0.89	1.08	0.94
A+G$_{rp}$	0.81	0.83	0.90	1.12	0.95
A+G$_r$+G$_p$	0.82	0.74	0.83	1.08	0.84

[1] Standard deviation for correlation and regression coefficients ranged from 0.002 to 0.008

Regression coefficient estimates were generally within an acceptable range around the expected value of 1 in both data sets, although there was a greater degree of over-prediction (regression coefficient < 1) in the large data set relative to the small data set. This could be due to the variance components used in both data sets were estimated using the small data set. Over-prediction regression coefficient was also more remarkable for the weight trait, PWT. It is interesting to note that IMF is the only trait with under-prediction regression coefficient (regression coefficient >1), especially for the A+G$_{rp}$ model.

CONCLUSIONS

Cross-validation analyses comparing the predictive ability of breeding values demonstrated the benefits of including genomic information, and that predictive SNPs do increase correlation by a small amount, and they can be included in a single genomic relationship matrix with all SNPs rather than used for an additional random term. This method is equivalent to the current industry evaluation model for these traits, highlighting that the current method is the more accurate of those investigated.

ACKNOWLEDGEMENTS

This work was funded by MLA Project L.GEN.1815.

REFERENCES

Brown D.J., Swan A.A., Boerner V., Li L., Gurman P.M., McMillan A., van der Werf J.H.J. (2018) *Proc. World Congr. Genet. Appl. Livest. Prod.* Species-Ovine:460.

Gurman P.M., Bunter K.L., Boerner V., Swan A.A., Brown D.J. (2019) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **23**:254.

Gurman P.M., Li L., Swan A.A., Moghaddar N., and van der Werf J.H.J. (2021) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **24**:.

Moghaddar N., Khansefid M., van der Werf J.H.J., Bolormaa S., Duijvesteijn N., Clark S.A., Swan A.A., Daetwyler H.D. and MacLeod I.M., 2019. *Genet. Sel. Evol.* **51**(1), 72.