Proc. Assoc. Advmt. Anim. Breed. Genet. 22:39-43

GENOMIC PREDICTION USING QTL REGIONS IDENTIFIED FROM REGIONAL HERITABILITY MAPPING FOR PARASITE RESISTANCE IN AUSTRALIAN SHEEP

M. Al Kalaldeh, J.H.J. van der Werf and C. Gondro

University of New England, School of Environmental and Rural Science, Armidale 2351, Australia

SUMMARY

Genomic selection uses genomic information to predict the breeding value of animals and can achieve higher prediction accuracy than pedigree based selection. This study aimed to compare the accuracy of genomic prediction using a medium-density (50k) SNP panel, as well as an imputed high-density (600k) SNP panel, with and without including pre-selected SNPs from QTL regions identified by regional heritability mapping (RHM). The proportion of variance explained by the pre-selected SNPs combined in a genomic relationship matrix (GRM) was considerably smaller than that explained by all SNPs from the 600k panel (25% of the genomic heritability). To obtain a better estimate of the variance explained by the pre-selected SNPs, both GRMs from the pre-selected SNPs (GRM_s) and their complementary SNPs from the 600k panel (GRM_c) were fitted in a single model. The total heritability explained by both GRM_s and GRM_c when fitted together was similar to the heritability explained by fitting all SNPs in a single GRM. The GRM_s explained a smaller proportion (18%) of the total heritability, whereas the GRM_c explained 82%. Fitting either the 50k or the 600k SNP panels resulted in similar prediction accuracy for parasite resistance (~0.37). However, when both GRM_s and GRM_c were fitted together in the prediction model, genomic accuracy was increased by 10%. These results indicate that accuracy of genomic prediction can be improved by including QTL information explicitly in the prediction models.

INTRODUCTION

Traditional genetic improvement relies on the use of pedigree information and phenotypic records of farm animals to estimate their breeding values. This has led to substantial genetic gain in most livestock species, especially for the traits that are easy to measure. However, the process is often inefficient for low-heritable, expensive or difficult to measure traits. An example is parasite resistance, measured by indicator traits such as worm egg counts (WEC), which is an important health issue that affects the sheep industry worldwide. Genomic selection offers an alternative to conventional breeding programs and can increase the rate of genetic gain by using genomic information to predict the breeding values of selection animals (Hayes *et al*., 2009).

In genomic selection, the genomic breeding values (GBV) for selection candidates are predicted based on the estimates of marker effects across the whole genome. The accuracy of predicting genomic breeding values depends on the heritability of the trait, the size of the reference population and the level of relatedness between the reference population and selection candidates (Habier *et al*., 2010). Moreover, the accuracy is highly influenced by the level of linkage disequilibrium between the SNP markers and the QTL (quantitative trait loci) affecting the trait (Goddard 2009). Depending on the genetic architecture of the trait, the chosen statistical method used to build the prediction model will have a significant impact on prediction accuracy. Models that incorporate pre-selected SNPs from QTL regions have been shown to improve the accuracy of genomic prediction (Brondum *et al.* 2015).

The objective of this study was to compare the accuracy of genomic prediction based on a medium-density (50k) SNP panel, high-density (600k) SNP panel, and including pre-selected SNPs

from QTL regions identified by regional heritability mapping for parasite resistance in Australian sheep.

MATERIALS AND METHODS

Animals. Parasite resistance, as measured by WEC, was investigated in a multi-breed sheep population from the Sheep Cooperative Research Centre information nucleus flock (INF). A total of 7,539 animals with both genotype data and WEC phenotypes were included in this analysis. Various breeds were represented in the population (Table 1) but with a significant proportion of Merino sheep, and only this breed had a substantial proportion of purebred animals. The remaining breeds were mainly represented by their crosses with Merino (van der Werf *et al*. 2010).

Table 1. Proportions of different breeds in the population

Breed	**BL**	**COR**	**COOP**	**EF**	**WD**	**PD**	**TEX**	**AF**	**PS**	**MER**
Proportion (%)	11.1	0.8	10	0.7	0.4	1.8	2.3	2	1.1	69.8

Border Leicester: **BL**, Corriedale: **COR**, Coopworth: **COOP**, East Friesian: **EF**, White Dorper:**WD**, Poll Dorset: **PD**, Texel: **TEX**, Australian Finnsheep: **AF**, Prime Samm: **PS**, Merino:**MER**

Genotypes. Animals were genotyped using the 50k Ovine marker panel (Illumina Inc., SanDiego, CA, USA). SNPs were removed if they had a minor allele frequency (MAF) < 1%, an Illumina Gentrain score (GC) less than 0.6, a call rate less than 95%, or not in Hardy-Weinberg equilibrium. Furthermore, positions of SNPs were obtained from the latest sheep genome *Ovis_aries_v3.1*, and any SNP with unknown position was removed. After applying these quality measures, 7,539 animals and 48,198 SNPs were retained. The imputation from the medium-density panel to the high-density (HD) SNP panel was performed using the Fimpute algorithm (Sargolzaei *et al*. 2014).

Cross-validation experimental design. Animals were randomly split into ten non-overlapping subsets (i.e. each subset with ~ 753 animals). For each experiment, one of the ten subsets served as a validation population and the remaining of the data served as the training population. The whole process was repeated ten times so that each subset served once as the validation population.

Regional heritability mapping (RHM). RHM was performed ten times, once for each validation set. The input to RHM consists of phenotype and genotype data (600k SNPs) on animals in the combined nine training sets. Data on animals in the validation set was not included in the RHM input. In RHM, each chromosome was divided into regions of pre-defined number of SNPs, and the variance attributable to each region was estimated. Window size of 200 SNPs was used to build genomic relationship matrix (GRM) and the window was shifted every 100 SNPs so that each two adjacent windows overlap midway. The significance was evaluated by the likelihood ratio test (LRT), comparing the RHM model which includes the regional effect with the base model composed of mean, fixed effects and random animal and error terms, but without the regional effect. The base model **(1)** and the RHM model **(2)** fitted to the data were as follows:

$$\boldsymbol{y = Xb + Za + e} \quad \textbf{(1)}$$
$$\boldsymbol{y = Xb + Za + Z_2 g + e} \quad \textbf{(2)}$$

where y is a vector of cube root transformed WEC records; b is a vector of fixed effects; $\boldsymbol{a}$ is a vector of random additive genetic effects, $\boldsymbol{g}$ is a vector of random regional genetic effect estimated from SNPs within each region (window), $\boldsymbol{e}$ is a vector of residuals which was assumed to be distributed as $\sim N(0, I\sigma_e^2)$, where $\boldsymbol{\sigma_e^2}$ is the residual variance. X, Z and Z_2 are incidence matrices

relating fixed, additive genetic and regional genetic effects to phenotypes. $\boldsymbol{a}$ was assumed to be distributed as $\sim N(0, A\sigma_a^2)$, where $\boldsymbol{A}$ is the numerator relationship matrix (NRM) calculated from deep pedigree records and $\boldsymbol{\sigma_a^2}$ is the additive genetic variance explained by pedigree; and $\boldsymbol{g}$ was assumed to be distributed as $N(0, G\sigma_g^2)$, where G is the regional genomic relationship matrix constructed from SNPs within each region, and σ_g^2 is the regional genomic variance. The fixed effects included in the models were breed proportions, age of animals, age of dam, gender, rearing type × birth type and contemporary groups (combination of flock site, birth year and management group effects).

Selection of SNP markers. Genomic regions obtained from each of the ten-fold cross-validation RHM analyses were ranked based on their LRT and significant regions were selected. For each fold, the top five ranked regions across the ten-fold experiments were the same. SNPs located within the top five ranked regions were used to build a GRM (GRM_s) and the proportion of the variance explained by these pre-selected SNPs was estimated by replacing the NRM in model (1) by the GRM obtained from the pre-selected SNPs. Variance was not only estimated using the GRM for the selected SNPs, but also by using a complementary GRM (GRM_c) based on the remaining SNPs from the 600k SNP panel. To obtain a better estimate of the variance explained by the selected SNPs, both the GRM_s and GRM_C were fitted together in the same model.

Accuracy of genomic prediction. To evaluate the impact of the selected SNPs on prediction accuracy, genomic predictions for the validation animals was calculated and correlated with the phenotypes of the same animals. The $GRM_s$was fitted and the genomic best linear unbiased prediction (GBLUP) analysis was performed. The prediction model that includes both GRM_s and GRM_c was also evaluated. Genomic breeding values (GBV) were calculated following the ten-fold cross-validation procedure as described above. Prediction accuracy was calculated as the correlation between the predicted GBVs of the validation set and the adjusted phenotypes, which were corrected for fixed effects, divided by the square root of the trait heritability. Furthermore, the regression coefficient (slope) of the adjusted phenotypes on the GBVs was calculated to assess the bias of genomic predictions.

RESULTS AND DISCUSION

The RHM results for ten-fold experiments are shown in the Manhattan plots in Figure1. The top five ranked regions remained consistent across the ten-fold cross-validation experiments. These five regions include three windows (107 -108 Mb, 110 -112 Mb, 117 -118 Mb) on OAR2, three overlapping windows between 28 to 36 Mb on OAR6, a window between 17 to 18 Mb on OAR18, a window between 7.2 to 6.8 Mb on OAR20 and a window between 40 to 41 Mb on OAR24. 1600 SNPs located within these regions were selected to build a GRM and, the heritability explained by the pre-selected SNPs was 0.05 compared to 0.19 explained by all the SNPs from the 600k panel.

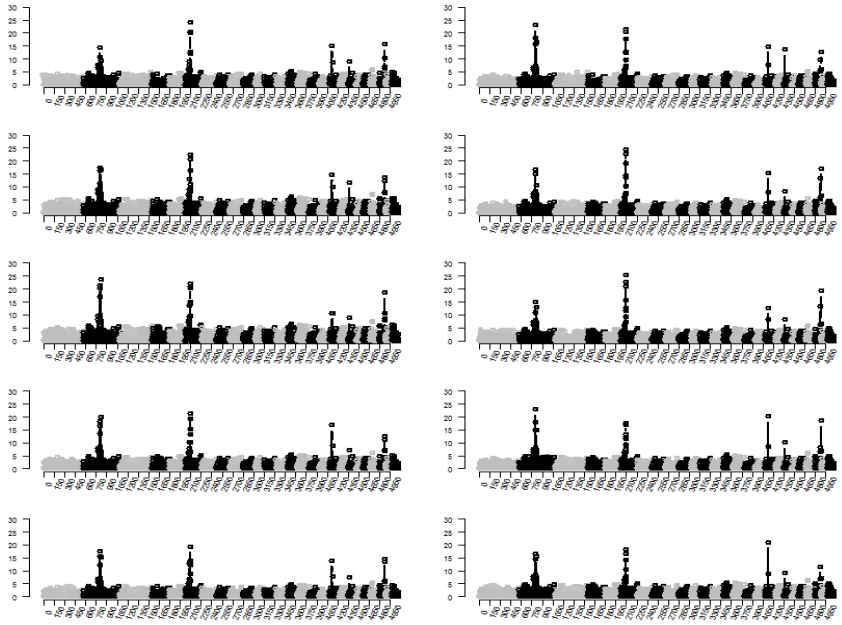

Figure 1. Manhattan plots of regional heritability mapping (RHM) results across the ten-fold cross-validation experiments. The x-axis represents the number of windows and the y-axis represents the corresponding likelihood ratio test (LRT) for each window.

Another way of testing the importance of the pre-selected SNPs was to investigate how much heritability was lost when the pre-selected SNPs were excluded from the GRM. Fitting only GRM_C, containing all SNPs in the 600k panel minus the pre-selected SNPs from the target regions, resulted in a similar heritability estimate as fitting all the SNPs. To assess the relative importance of the GRM from the selected SNPs and the GRM from the remaining SNPs, both GRM_s and GRM_C were fitted simultaneously in the same model. The proportion of variance explained when both GRM_s and GRM_C were fitted simultaneously was similar to the proportion of the genetic variance explained by fitting all the SNPs from the 600k. The GRM from the selected SNPs explained 18% of the total heritability, whereas 82% of the total heritability was explained by all the remaining SNPs (Table 2).

Table 2. The proportion of phenotypic variance (h^2) explained for parasite resistance

Selection criteria	GRM	GRM_s	GRM_C	logL
G (50k)	0.178 ± 0.020			-10673
G(600k)	0.194 ± 0.021			-10670
G(regions)		0.050 ± 0.009		-10682
GRMc			0.188 ± 0.021	-10673
G(Regions)+GRMc		0.034 ± 0.008	0.152 ± 0.021	-10638

G (50k): GRM from the 50k SNP panel, G (600k): GRM from the 600k SNP panel, G (regions): GRM_s from the pre-selected SNPs; GRMc: complementary GRM (GRMc)

Using any of the 50k and the 600k SNP panels resulted in a similar prediction accuracy for parasite resistance (~0.37, Table 3). When the GRM_s from the pre-selected SNPs was fitted alone,

the prediction accuracy dropped by 18% compared to fitting all SNPs from the 600k panel. However, when both GRM_s and GRM_C were fitted together, higher prediction accuracy was observed than fitting all the SNPs in a single GRM. This is likely because a model with two components of genetic effects allows effects of the pre-selected SNPs to have larger variance than all the remaining SNPs in the panel, thus putting more weight on the pre-selected SNPs from the QTL regions. Moreover, the slopes of all models were not significantly different from 1, which indicates no significant bias in the predictions. It should however be noted that the RHM regions are not independent since they were the same across all 10-fold repeats and this can of course favourably influence the prediction accuracy. While suboptimal for a fair comparison of accuracy of prediction this lack of independence is not unexpected nor undesirable in practice since QTLs should have a real biological effect on a trait and are expected to be consistently identifiable in different datasets with similar power. If the RHM regions changed with each subset of the data, there would be greater cause for concern.

Table 3. Cross-validation prediction accuracy for parasite resistance averaged over the ten validation sets, and slope for the regression of adjusted phenotypes on the predicted breeding values

Selection criteria	Accuracy	SE(accuracy)	Slope	SE(slope)
G (50k)	0.368	0.036	0.915	0.197
G(600k)	0.374	0.036	0.916	0.193
G(regions)	0.307	0.035	0.841	0.219
G(Regions)+GRMc	0.411	0.036	0.848	0.164

CONCLUSION

The results in this study show that there is little advantage of using the imputed high density SNP panel over the medium-density panel for genomic prediction with this trait. However, by incorporating information from QTL regions explicitly into the genomic prediction model, prediction accuracy of parasite resistance increased by 10% based on the current SNP panel density. These results suggest that QTL information should be beneficial in genomic prediction, not just for parasite resistance but also for other economically important traits in sheep**.**

ACKNOWLEDGEMENT

This project was supported by Sheep CRC, Next-Generation BioGreen 21 Program PJ01134906 and PJ012611, Rural Development Administration, Republic of Korea and Australian Research Council (DP130100542).

REFERENCES

Hayes B.J., Bowman P.J., Chamberlain A.C., Verbyla K. and Goddard M.E. (2009) *Genet. Sel. Evol.* **41**: 51.

Habier D., Tetens J., Seefried F., Lichtner P. and Thaller G. (2010. *Genet. Sel. Evol.* **42**: 5.

Goddard, M.E., (2009) *Genetica.* **136**: 245.

Brondum R. F., Su G., Janss L., Sahana G., Guldbrandtsen B., Boichard D., *et al.* (2015) *J Dairy Sci.* **98**: 4107.

Van der Werf J.H.J., Kinghorn B.P. and Banks R.G. (2010) *Anim. Prod. Sci.* **50**: 998.

Sargolzaei M., Chesnais J. and Schenkel F. (2014) *BMC Genom.* **15**: 478.