*Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.***Genomic prediction for parasite resistance in sheep using whole-genome sequence data**

M. Al Kalaldeh[1, 2], N. Duijvesteijn[1, 2], N. Moghaddar[1, 2], J.P. Gibson[1, 2]& J.H.J. van der Werf[1, 2]

[1]Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351, Australia.
[2] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia malkalal@myune.edu.au (Corresponding Author).

Summary

This study aimed to compare QTL mapping precision and the accuracy of genomic prediction for WEC in sheep using variants selected from the high-density SNP panel (600k) and the imputed whole-genome sequence (WGS) data and to evaluate the prediction accuracy when the selected variants were used. A total of 11,431 animals with both phenotype and genotype data were included in this analysis. Variants selection was performed on an independent QTL discovery set that comprised 6,431 animals, whereas effects of the selected SNPs were trained and validated using the remaining 5,000 animals. Mapping precision of three regions on chromosomes 2 (107-117 Mb), 6 (35-39 Mb) and 18 (17-19 Mb) was similar to those obtained from the WGS data. However, mapping precision based on the WGS data was more precise in the regions on chromosomes 3 (147-150 Mb) and 24 (38-42 Mb) compared to those obtained from the 600k SNPs. SNPs located within the target regions were used to evaluate their effects on the accuracy of genomic prediction. Genomic prediction models including a separate genomic relationship matrix (GRM) based on selected SNP genotypes fitted alongside a GRM based on the standard 50k array improved the prediction accuracy for WEC from 13% to 26% compared to fitting only the 50k. The highest accuracies were obtained when variants were selected from the WGS data (), which implies that WGS data may include variants that have high LD with rare causative mutations in those regions that are not fully captured by the 600k SNP panel. The results of this paper show that the use of WGS variants located within or close to QTL regions can improve the prediction accuracy of WEC compared to using variants selected from the high-density SNP panel.

Keywords: QTL, GWAS, RHM, prediction accuracy

Introduction

Gastrointestinal parasite infections present a major health issue affecting sheep in Australia and worldwide. Selection for parasite resistance, measured by worm egg counts (WEC), is difficult as the trait is not easily measured. Genomic selection offers an alternative to traditional selection methods and can improve the rate of genetic gain by using single nucleotide polymorphism (SNP) markers to predict the breeding values of animals (Meuwissen *et al*., 2001). Increasing the SNP density may increase the level of linkage disequilibrium (LD) between SNPs and QTL affecting the traits and is therefore beneficial for efforts to map quantitative trait loci for WEC. SNPs in QTL regions might then be found to possibly increase the accuracy of genomic selection. A recent study on production traits in sheep (Moghaddar *et al.,* 2017) has shown that genomic predictions using the high-density

(600k) SNP panel resulted in a very small improvement in the accuracy compared to using the medium-density (50k) SNP panel.

Using all genotyped variants from WGS may not necessarily improve the prediction accuracy except when only variants in high LD with causative mutations are used. Genomic prediction models that incorporate pre-selected variants from QTL regions have shown to improve the prediction accuracies (for example: van den Berg *et al.,* 2016). Genomic prediction models using genomic best linear unbiased prediction (GBLUP) assume equal variance for the effects of SNP markers. However, selected SNPs from QTL regions might require more emphasis in the analysis. The objective of this study was to compare the mapping precision and the accuracy of genomic prediction for parasite resistance in sheep using pre-selected variants identified from the high-density 600k SNP chip and the imputed whole-genome sequence (WGS) data and to evaluate the prediction accuracy when using selected SNPs.

Material and methods

Animals

Parasite resistance, as measured by worm egg count (WEC), was investigated in a multi-breed sheep population from the Sheep Cooperative Research Centre information nucleus flock (INF). A total of 11,431 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 purebred Merino sheep. The remaining breeds were mainly represented by their crosses with the Merino (van der Werf *et al*. 2010).

Table 1. Proportions of different breeds in the population

Breed	BL	COR	COOP	EF	WD	PD	TEX	DR	AF	SF	WS	PS	MER
Proportion (%)	10.9	0.8	6.7	0.5	0.8	6.7	1.7	0.5	1.3	1.6	2.9	0.9	64.6

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

To avoid obvious bias, the QTL discovery set should be independent from the training and validation sets. Therefore, a QTL discovery set was generated from 6,431 animals representing a random mix of all breeds and crosses in the whole population, whereas the remaining 5,000 animals were assigned to the training/validation set. Validation groups were chosen from the training/validation set using two different methods. The first method was through ten-fold cross-validation, where animals were split into ten non-overlapping subsets and one of these subsets served as a validation group and the remaining data served as a training group. The whole process was repeated ten times so that each subset served once as a validation set. The second method was by creating a validation group that has a distant relationship with the remaining animals in the training/validation set. Genomic relationships were calculated based on the whole-sequence data. Animals were then ranked based on their median genomic relationships and the 10 % of animals having the lowest relationships were then selected as a validation group.

Genotypes

Animals were genotyped using the 50k Ovine marker panel (Illumina Inc., SanDiego, CA, USA). The imputation from the medium-density panel to the high-density (HD) SNP panel was performed using the FImpute algorithm, whereas the imputation from the high-density (HD) SNP panel to whole-genome sequence (WGS) genotypes was performed by Minimac3. All variants with R^2 imputation accuracy (reported by Minimac) ≤ 0.4 (Bolormaa, et al., 2018), variants with a MAF<0.01 and those on the X chromosome were excluded leaving a total of 28,525,455 variants in the RHM and GWAS.

Selection of prediction markers

Both genome-wide association studies (GWAS) and regional heritability mapping (RHM) approaches were performed to select subsets of prediction markers to be included in the prediction models. For GWAS, WEC were first pre-adjusted using the base model, which include fixed effects and random additive genetic effects based on a pedigree-based relationship matrix, using ASReml (Gilmour *et al.*, 2009). Variants from the WGS genotypes were then regressed, one at the time, on the adjusted phenotypes using R statistical language (http://www.r-project.org). RHM was performed on WGS data using MTG2 software. In RHM, each chromosome was divided into regions with a window size of 12000 SNPs (~ 1 Mb), and the variance explained by each region was estimated. The significance of each region on the genome was assessed by the likelihood ratio test (LRT), comparing the RHM model, which included the regional SNP effect, with the base model composed of mean, fixed effects and random animal and error terms. Fixed effects included the breed proportions, age of animals, age of dam, sex, contemporary groups, rearing type and birth type. Significant regions were selected at false discovery rate (FDR) adjusted p-value of 0.05. Significant regions as well as those with p-values just below the genome-wide significance level were selected for further analysis on 50k, 600k SNPs and WGS data using smaller window sizes (0.5 Mb and 0.25 Mb) to compare the mapping resolution of the three SNP panels.

To evaluate the impact of the selected SNPs on prediction accuracy, genomic predictions for the validation animals were calculated and correlated with the phenotypes of the same animals. The GRM from the 50k SNPs was fitted and a genomic best linear unbiased prediction (GBLUP) analysis was performed. The prediction models that include both a GRM from the selected variants and a GRM from the 50k SNPs were also evaluated and compared to the model where only one GRM from the 50k SNPs was fitted. The accuracy of prediction was calculated as the correlation between the genomically estimated breeding value from the GBLUP model with the animals' phenotype divided by the square root of the heritability.

Results and Discussion

The GWAS and RHM results using the WGS data are shown in the Manhattan plots in Figures1 and 2, respectively. None of the SNPs from GWAS reached the FDR significance threshold of 0.05. However, the RHM mapped four regions to chromosomes 2, 6, 18, and 24 which passed the genome-wide significance level. These include: one region of four windows (107 -117 Mb) on OAR2, three overlapping windows between 35 to 39 Mb on OAR6, a window between 17 to 18 Mb on OAR18, and a window between 40 to 42 Mb on OAR24.. Fine-mapping RHM using 0.25 Mb windows on GWS data had a more precise mapping resolution than larger window sizes. RHM results using 0.25 Mb window size from the 600k

SNPs were similar to those from WGS for regions on chromosomes 2, 6, and 18. RHM results from the GWS were more precise for chromosomes 3 and 24. RHM results based on the 50k genotypes, on the other hand, performed not as well as the WGS or 600k SNP panels, apart from the significant region on chromosome 2 (Figures 2 to 5). The RHM results from the 600k and WGS data were used to select subsets of SNPs for genomic prediction. Selection was based on p-values using arbitrary cut-offs of 3 and 4 for the. The total number of variants from the WGS data with a higher than 3 and 4 were 80,285 and 44,581 before LD pruning and 10,543 and 6,368 after LD pruning, respectively. The total number of variants from the 600k SNP panel that had a higher than 3 and 4 were only 1,435 and 827, respectively. Accuracies of genomic prediction obtained by using a GRM from the 50k SNPs and scenarios where a GRM from the 50k SNPs was fitted alongside a GRM from pre-selected SNPs are shown in Table 2. The prediction accuracies were evaluated in a 10-fold cross-validation design with random allocation of animals to subsets. No significant differences were observed between the different scenarios based on the cross-validation design. This is largely due to the close relationships between animals across the CV groups, where SNPs can capture co-segregation of alleles as well as the LD between QTL and SNPs. Prediction accuracies were also evaluated by creating a validation group that had a distant relationship with all the remaining animals in the training/validation set. Distant relationships between validation and training groups helps the genomic predictions to benefit more from the LD between the SNPs and QTL affecting the trait. Fitting a GRM from the pre-selected SNPs with a GRM from the 50k SNPs improved the prediction accuracy substantially compared to fitting the 50k SNPs alone. Prediction accuracies based on the selected variants from the WGS were generally higher than those obtained from the 600k SNPs. The highest accuracy of genomic prediction was obtained in scenarios where 50k SNPs were fitted with SNPs selected from the WGS data with a higher than 3. Pruning SNPs based on LD () resulted in a smaller number of SNPs included in the prediction model, however, the prediction accuracies were similar to those obtained using all the SNPs in the target regions before pruning.

Conclusion

Our results show that RHM outperformed GWAS in detecting regions for parasite resistance when the whole-genome sequence data were used. RHM captures the genetic variation in a region by integrating the effects of rare and common variants, and therefore is able to capture more of the genetic variance undetected by GWAS. The 600k SNPs provided mapping precision similar to those obtained from the WGS data for the regions on chromosomes 2, 6, and 18 when analysed with a 0.25 Mb window size. However, the WGS provided better resolution for the regions on chromosomes 3 and 24. Using the pre-selected SNPs from the WGS data improved the prediction accuracies of parasite resistance more so than using the pre-selected SNPs from the 600k SNPs when all regions on chromosomes 2, 3,6,18 and 24 were used. This is probably because WGS data is more likely to include SNPs with strong LD with other rare causal variants in those regions not tagged by the 600k SNPs.

Acknowledgments

This project was funded and supported by the Sheep CRC Information Nucleus flocks.

The authors acknowledge and thank Sunduimijid Bolormaa and Iona Macleod for providing the imputed sequence data. Klint Gore is acknowledged for his help on retrieving data.

List of References

Bolormaa, S., A. Chamberlain, J.H.J. v. d. Werf, H.D. Daetwyler & I.M. MacLeod, 2018. Evaluating the accuracy of imputed whole genome sequence in sheep. Proc. 11th World Congr. Genet. Appl. Livest. Prod.

Moghaddar, N., A.A. Swan & J.H.J. Van Der Werf, 2017. Genomic prediction from observed and imputed high-density ovine genotypes. Genet. Selec. Evol. 49: 40-last page.

Meuwissen, T.H., B.J. Hayes & M.E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genet. 157: 1819-1829.

Van der Werf, J., B. Kinghorn & R. Banks, 2010. Design and role of an information nucleus in sheep breeding programs. Anim. Prod. Sci. 50(12): 998-1003.

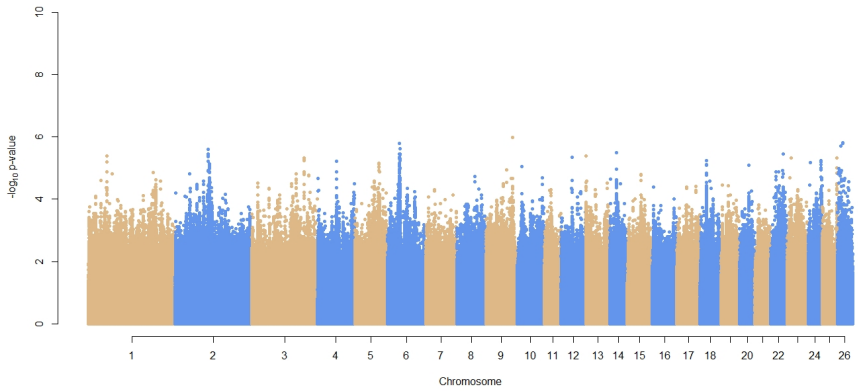


Figure 1. Manhattan plot of genome-wide assoication studies (GWAS) based on whole-genome sequence data (WGS). The x-axis represents the number of windows and the y-axis represents the for each SNP.

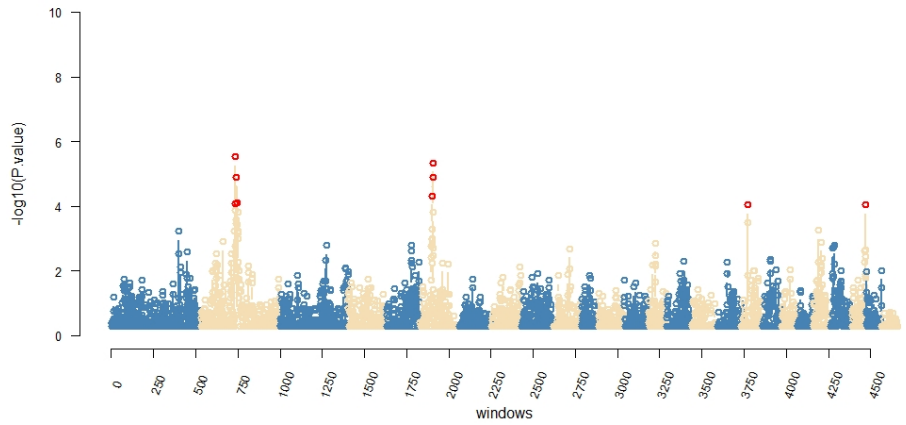


Figure 2. Manhattan plot of regional heritability mapping (RHM) based on whole-genome sequence data (WGS). The x-axis represents the number of windows and the y-axis represents

the for each window.

Table 2. Prediction accuracy for parasite resistance and slopes for the regression of adjusted phenotypes on the predicted breeding values.

Selection criteria	distantly-related	Random cv	distantly-related	Random cv
	Prediction accuracy		slope	
50k only			**0.66**	**0.99**
50k+			**0.67**	**0.94**
50k+			**0.94**	**0.96**
50k +			**0.75**	**0.96**
50k+			**1.07**	**0.97**
50k+			**0.83**	**0.97**
50k+			**1.06**	**0.98**

50k: fitting only the GRM from the 50k SNPs**, 50k +** : fitting simultaneously a GRM from the 50k SNPs and a GRM from the sequence variants in windows with , **50k +** : fitting simultaneously a GRM from the 50k SNPs and a GRM from the sequence variants in windows with , **50k+**: fitting simultaneously a GRM from the 50k SNPs and a GRM from the sequence variants in windows with after LD pruning, **50k+**: fitting simultaneously a GRM from the 50k SNPs and a GRM from the sequence variants in windows with after LD pruning, **50k+**: fitting simultaneously a GRM from the 50k SNPs and a GRM from the 600K variants in windows with , and **50k+**: fitting simultaneously a GRM from the 50k SNPs and a GRM from the 600K variants in windows with .

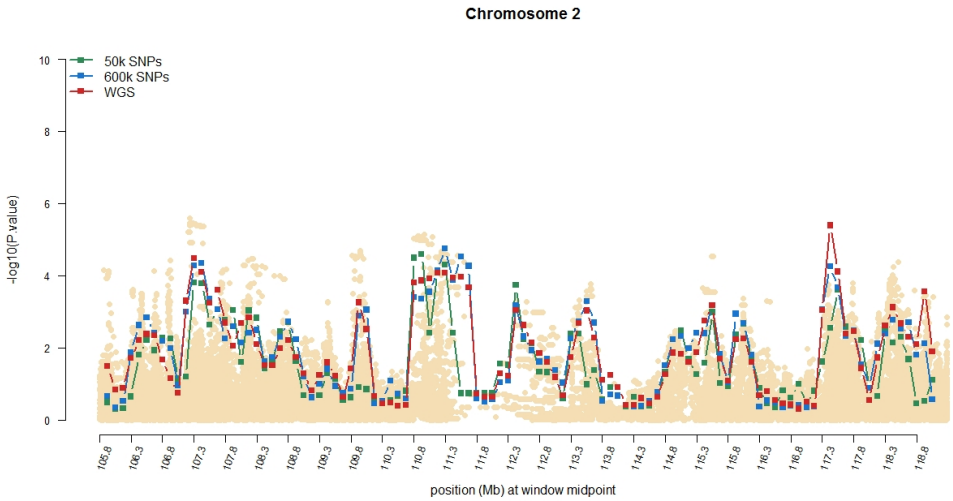
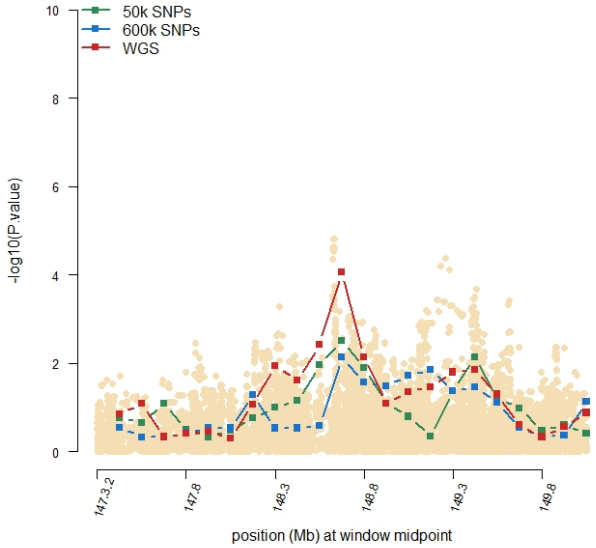
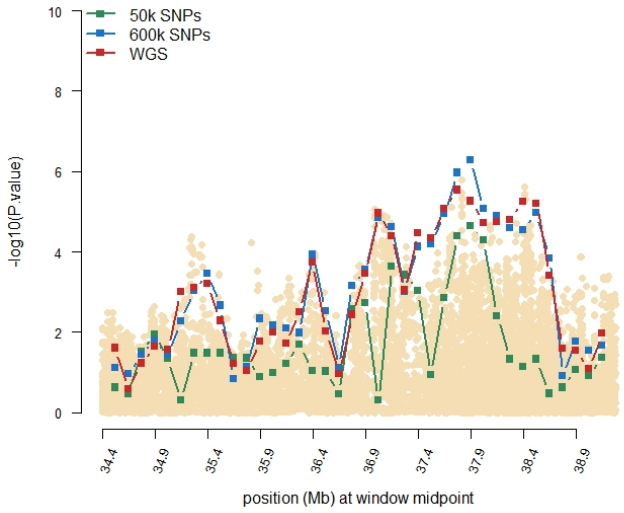


Figure 2. GWAS and RHM results of the identified region on chromosome 2. The solid lines show the RHM results using 0.25 Mb window size from the three SNP panels (50k = green; 600k = blue; sequence = red), where each window was positioned at window midpoint. Wheat-coloured dots show the GWAS results within the region.

Chromosome 3



Chromosome 6



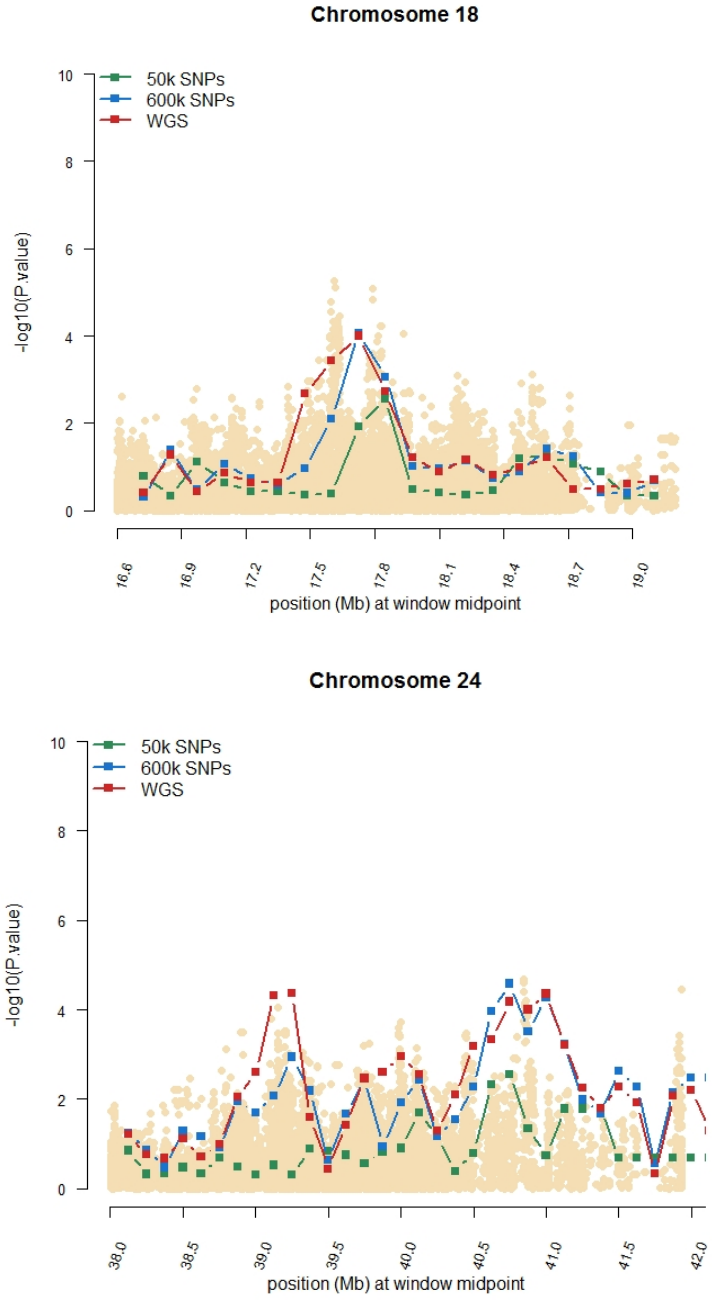


Figure 3. GWAS and RHM results of the identified region on chromosome 3(a), chromosome 6 (b), chromosome 18 (c), and chromosome 24 (d). The solid lines show the RHM results using 0.25 Mb window size from the three SNP panels (50k = green; 600k = blue; sequence = red), where each window was positioned at window midpoint. Wheat-coloured dots show the GWAS results within the region.