Proceedings of the World Congress on Genetics Applied to Livestock Production, 11.456

Genomic evaluation based on selected variants from imputed whole-genome sequence data in Australian sheep populations

N. Moghaddar[1,2]*, I.M. MacLeod*[1,3]*, N. Duijvesteijn*[1,2]*, S. Bolormaa*[1,3]*, M. Khansefid*[13]*, H. Al-Mamun* [1,2]*, S. Clark*[1,2]*, A.A. Swan*[1,4]*, H.D. Daetwyler*[1,3]*, and J.H.J van der Werf*[1,2]

[1]*Sheep-CRC, Armidale, NSW 2351, Australia*
[2]*School of Environmental & Rural Science, University of New England, Armidale, NSW 2351, Australia*
n.moghaddar@une.edu.au (Corresponding Author)
[3]*Bioscience Research, Agriculture Victoria, Bundoora, VIC 3083, Australia*
[4]*Animal Genetics and Breeding Unit (AGBU), Armidale, NSW 2351, Australia*

Summary

This study investigates improvement in accuracy of genomic prediction for growth and eating quality traits in Australian sheep populations based on selected variants from imputed whole genome sequence (WGS) data combined with a 50k-SNP array. Selection of SNP variants was based on single trait multi-breed genome wide association studies (GWAS) on WGS data in an independent data subset. Genomic prediction was based on genomic best linear unbiased prediction (GBLUP) using training sets of between 6,353 and 11,067 multi-breed purebred and crossbred animals. Four different genotype sets were compared: 50k SNP genotypes, WGS variants, selected sequence variants from GWAS and selected sequence variants combined with 50k genotypes. The latter set was modeled as either one or as two subsets with different variance components. Results showed a substantial improvement in prediction accuracy when selected sequence variants from GWAS were added to the standard 50k-SNP array. Absolute value of increase in accuracy across different traits was on average 6.2% and 4.1% for purebred and crossbred Merino sheep, respectively, when selected sequence variants and 50k genotypes were fitted as two variance components simultaneously. The improvement in prediction accuracy across different traits was on average 4.4% and 3.8% for purebred and crossbred Merino sheep, respectively, when selected sequence variants combined with 50k SNP arrays were fitted as one variance component.

Keywords:
Whole Genome Sequence data, genomic prediction

Introduction

Accuracy of genomic prediction is highly dependent on the size of the training population and the effective number of chromosome segments as a function of the effective population size (N_e). In multi-breed/crossbred populations such as in Australian sheep population, it is difficult to achieve a sufficiently high genomic prediction accuracy for genetically diverse breeds with high N_e or for minor breeds and it would be beneficial if across breed information could be used. So far, there is limited evidence that information from other breeds can notably increases the prediction accuracy, when prediction is based on common genome wide single nucleotide polymorphism (SNP) arrays like 50k genotypes. This could be due to low linkage disequilibrium (LD) between genetic markers and QTL and also because of possible differences in QTL effects across different breeds and strains within a breed. Low LD suggests denser marker genotypes could improve genomic prediction accuracy across breeds. However, using high density (600k) genotypes has shown only a small improvement in prediction accuracy in sheep compared to using a 50k SNP array (Moghaddar *et al.*, 2017). Whole genome sequence data contains millions of variants, including causative mutations responsible for genetic variation of a trait. However, causal mutations and variants in strong linkage with them, are likely to be only a small subset of the whole sequence variants. Studies in dairy cattle show no to limited improvements in prediction accuracy when using the complete set of sequence variants or when a subset of significant GWAS sequence variants were combined with a common medium density SNP array (e.g. van Binsbergen et *al.*, 2015; van den Berg *et al.*, 2016; Brøndum *et al.*, 2015; VanRaden *et al.*, 2017). This study investigates the accuracy of genomic prediction for growth and eating quality traits in a multi-breed Australian sheep population using selected sequence variants combined with 50k genotypes compared to using whole genome sequence data or only 50k genotypes.

Materials and Methods

Phenotypes. Phenotypes of six growth and eating quality traits collected between 2008 and 2015 from a combined multi-breed "Research" (Sheep-Cooperative Research Center Information Nucleus Flocks) and "Industry" database (Sheep Genetics, AGBU) were used in this study. The investigated traits were post weaning weight (PWT), post weaning eye muscle depth (PEMD), carcass eye muscle depth (CEMD), carcass fat (CFAT), intra muscular fat (IMF) and shear force at day 5 aging (SF5). The phenotypes were first adjusted for significant non-genetic effects, including contemporary group and data source and for maternal effects and genetic groups derived from pedigree. All adjusted phenotypes were then divided into three non-overlapping data subsets including a GWAS discovery subset, a genomic prediction training subset and two validation subsets. The GWAS discovery subset was selected randomly and included all the main breeds and crosses in the research and industry data (across all birth years) and ranged from 4,282 to 4,900 animals for different traits. The prediction training subset was similar in composition to the GWAS discovery and ranged from 6,353 to 11,067 animals across different traits. The validation subset was comprised of two groups; a purebred Merino subset ranging from 848 to 3,118 animals and a crossbred of Border Leicester x Merino sheep (BLxMer) subset ranging from 315 to 868 animals for different traits. Validation subsets were selected to have low genomic relationship to the training subset by assigning complete half-sib families to subsets.

SNP Marker Genotypes

50k genotypes: A total of 35,980 animals from research and industry flocks were used across different traits in this study in which animals had already been genotyped with 50k-ovine SNP genotypes (67%) or imputed from low density 12k-ovine SNP genotypes to 50k genotypes (33%). Imputation from 12k to 50k genotypes was based on Beagle 3.3.2 software (Browning, 2009). After performing quality control on genotypes the number of SNPs used in the final 50k array was equal to 48,599.

High Density (HD) genotypes: 2,266 key animals including all sires and dams/progeny with a high relationship to the rest of the population of purebreds and crossbreds of the four main sheep breeds (Merino, Border Leicester, Poll Dorset and White Suffolk) were genotyped with the ovine HD SNP chip (600k). After performing quality control the sporadic missing genotypes in HD genotypes were imputed with Beagle 3.3.2. The total number of SNPs in the final HD genotype array was 510,065. All 2,266 animals with observed HD genotypes were used as a reference set to impute 50k genotypes to HD using Minimac 3.0 (Das *et al.*, 2016). Prior to imputation pre-phasing was performed using Eagle 2.0 software (Loh *et al.*, 2016) separately for HD genotyped animals and for 50k genotyped animals.

Whole Genome Sequence (WGS) genotypes:

A complete description of the imputation to whole genome sequence in these Sheep-CRC animals is provided by Bolormaa *et al.,* (these proceedings). Briefly 376 animals from the main Australian sheep breeds from research flocks were sequenced (with ~10x coverage) by the Sheep-CRC. This data was then combined with WGS data on an additional 350 European sheep breeds sequenced within the "Sheep Genomes DB" project (Daetwyler *et al.*, 2017). The combined WGS data on 726 animals were used as the sequence imputation reference set to impute 35,518 animals with real or imputed HD genotypes to sequence. As described for HD imputation, genotypes were pre-phased using Eagle2 (Loh *et al.,* 2016) and then imputed with Minimac3 (Das *et al.*, 2016). All variants with Minimac 3.0 imputation quality statistic (R^2) lower than 0.4 were discarded, resulting in a final set of 31,154,249 SNP and InDel genotypes for every animal.

Statistical methods.

Pre-adjusted phenotypes described above were used in GWAS analysis and in genomic prediction. The GWAS was implemented based on single variant regression method in the independent multi-breed GWAS discovery data subset. A genomic relationship matrix **G** (calculated based on HD genotypes) was also fitted in GWAS to account for the population structure not captured for by genetic groups. Sequence variants to be used in genomic prediction were selected based on the most significant SNP (-*Log Pvalue ≥ 3*) and then pruned for high LD (≥ 0.95) in each 100 kb window with a sliding window size of 50kb. GBLUP was performed for the training and validation subset with pre-adjusted phenotypes and fitting **G** constructed from: a) 50k genotypes, b) all sequence variants, c) selected sequence variants, d) selected sequence variants combined with 50k genotype and fitted as one variance component and e) 50k genotypes and selected sequence variants fitted as two variance components simultaneously. MTG2 software (Lee and van der Werf, 2016) was used for genomic prediction. Accuracy of genomic prediction was evaluated in validation data subsets (purebred Merinos and crossbred Merinos) based on the correlation between genomic breeding values (GBV) and the adjusted phenotypes divided by the square root of trait heritability (calculated from the 50K genotype GBLUP: Table 1). Bias of genomic prediction was assessed as deviation from unity of the coefficient of regression of adjusted-phenotypes on GBV in validation subsets.

Results and Discussion

Variance components.

Table 1 shows the summary statistics of the data including size of the genomic prediction training subset, phenotypic mean and standard deviation plus estimated additive genetic variance and heritability for different models. Results showed a small but consistently higher additive genetic variance and hence heritability based on WGS data or 50k genotypes plus selected sequence variants for different traits compared to only 50k genotypes. The heritability based on WGS variants was always the highest and the sum of heritability based on fitting 50k genotypes and selected sequence variants (fitted as two component) was in-between the heritability from 50k genotypes and WGS data. Heritability of selected sequence variants *(-LogPvalue ≥ 3*) varied between 3% and 11% for different traits when selected sequence variants fitted with the 50k genotypes simultaneously.

Accuracy and bias of genomic prediction. Figures 1 and 2 show the accuracy of genomic prediction based on different models for purebred and crossbred Merino validation subsets respectively. Compared to using 50k genotypes, using all sequence variants resulted in a 1.4% and 2.6% (absolute value) improvement in prediction accuracy for purebred and crossbred Merinos for different traits. A notably higher and almost consistent improvement in prediction accuracy was observed when selected sequence variants from GWAS were combined with 50k genotypes. Improvement in absolute value of prediction accuracy was 6.2% and 4.1% for the purebred and crossbred validation populations when selected sequence variants were fitted as two separate variance components, and 4.4% and 3.8% when selected sequence variants and 50k genotypes were fitted as one variance component. An exception to this was SF5 in both purebred and crossbred validation subset in which 50k-SNP genotypes or all sequence variants outperformed the 50k-SNP genotypes combined with selected sequence variants. Coefficients of regression of adjusted phenotypes on GBV are shown in Table 2 for purebred and crossbred Merino validation subsets, which indicate lower bias of GBV based on using all sequence variants or selected sequence variants combined with 50k genotypes. A notably higher bias was observed when using only selected sequence variants.

In this study the GWAS data subset was selected to be multi-breed and higher accuracy was observed for both purebred and crossbred Merino sheep when selected sequence variants were included in genomic prediction. This result could be partly due to Merino being the dominant breed in the data (purebred and crossbred). In the case of SF5, genomic prediction accuracy decreased with the addition of selected sequence variants to 50k genotypes which could be related to the lower power of GWAS in SF5 due to smaller sample size.

This study showed a small improvement in genomic prediction accuracy using all sequence data compared to 50k genotypes. Marginal improvement in accuracy from using all WGS variants is in line with several other studies in cattle (e.g. van Binsbergen *et al.,* 2015) and could be because of the difference in genomic relationships among animals based on 50k or all sequence variants. In this study, using the CFAT data set as an example, the correlation between the elements of the genomic relationship matrix based on using all sequence SNP variants versus only 50k genotypes was high and equal to 0.976.

Regression coefficients of adjusted phenotypes were lower than one in most cases, which shows GBVs are over-dispersed. The results also show more bias for traits with low prediction accuracy including IMF and SF5 which could be related to smaller size of both training and validation data subsets for those traits.

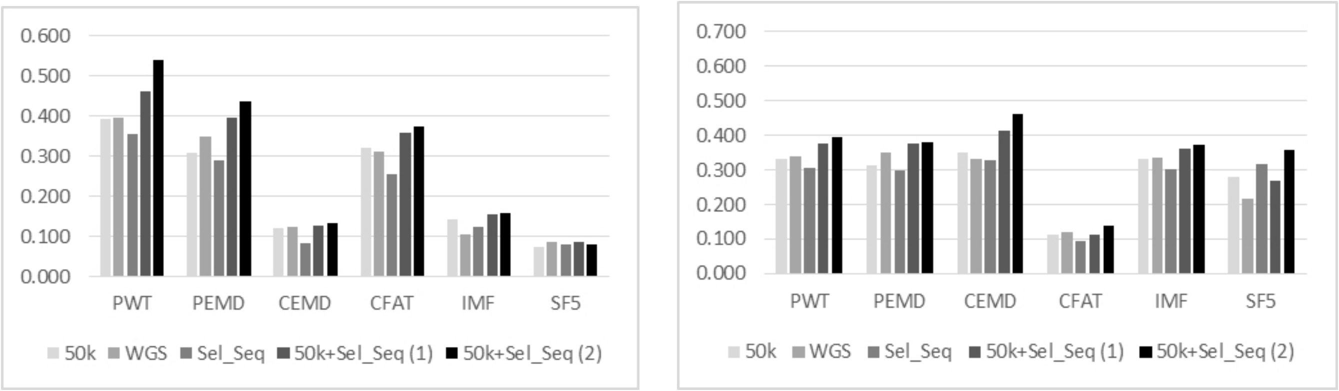
Table 1. Genomic prediction training sub set size, phenotypic mean (standard deviation), additive genetic variance and heritability based on GBLUP using 50k, WGS data and 50k plus selected variants.

Trait	size[1]	Phenotypic Mean (sd)[2]	$V_{a,50k}$	$V_{a,WGS}$	$V_{a,(50k,sel\text{-}variants)}$	h^2_{50k}	h^2_{WGS}	$h^2_{(50k,sel\text{-}variants)}$
PWT *(kg)*	11,067	46.70 (13.10)	6.97	8.62	5.37 , 2.04	0.21	0.25	0.16 , 0.06
PEMD *(mm)*	9,715	26.34 (4.65)	1.06	1.36	0.75 , 0.42	0.23	0.28	0.19 , 0.09
CEMD *(mm)*	7,714	29.56 (4.85)	1.38	1.70	1.22 , 0.30	0.16	0.19	0.14 , 0.03
CFAT *(mm)*	7,635	4.04 (2.27)	0.51	0.58	0.32 , 0.19	0.19	0.21	0.13 , 0.07
IMF *(%)*	6,353	4.35 (1.14)	0.26	0.30	0.23 , 0.05	0.38	0.42	0.33 , 0.07
SF5 *(Newtons)*	7,392	25.30 (14.05)	10.90	13.31	5.68 , 4.99	0.24	0.29	0.14 , 0.11

[1]: population size, [2]: phenotypic mean and standard deviation, V_a: additive genetic variance, PWT: post weaning weight, PEMD: post weaning eye muscle depth, CEMD: carcass EMD, CFAT: carcass fat, IMF: intra muscular fat, SF5: shear force at day five, 50k: 50k-SNP genotypes, WGS: whole genome sequence SNPs, sel-variants: selected sequence variants.

Figure 1. Accuracy of genomic prediction in purebred Merinos

Figure 2. Accuracy of genomic prediction in crossbred Merinos



1: Combined 50k genotypes and selected sequence varaiants fitted as one component. 2: 50k genotypes and selected sequence variannts fitted as two separate components simoultanously.

Conclusion

This study demonstrated that imputed whole genome sequence data can be used to improve the accuracy of genomic prediction in multi-breed sheep populations. Accuracy of prediction was increased by about 5% on average when selected sequence variants were added to the common 50k genotype array. GBLUP methods can best accommodate such selected variants by fitting them as a separate variance component from the common 50k SNPs.

Table 2. Regression coefficient of adjusted phenotypes from GBV in purebred and crossbred Merino validation sets.

	Purebred Merinos				Crossbred Merinos			
Trait	50k	WGS	Sel-Variants	50k+Sel-Variants	50k	WGS	Sel-Variants	50k+Sel-Variants
PWT	0.92	0.91	1.14	1.06	0.89	0.89	0.80	0.88
PEMD	0.87	0.90	0.74	0.88	0.92	0.95	0.68	0.84
CEMD	0.89	0.88	0.70	0.77	1.14	1.00	0.69	1.62
CFAT	1.06	1.10	0.61	0.91	0.36	0.72	1.44	1.07
IMF	0.51	0.49	0.44	0.50	0.85	0.88	0.84	0.88
SF5	0.34	0.46	0.61	0.35	0.64	0.56	1.44	0.65

Acknowledgements

The authors thank Sheep Genetics and MLA for providing access to phenotypic data from industry animals.

List of References

Brøndum, R.F., G. Su, L. Janss, G. Sahana, B. Guldbrandtsen, D. Boichard, M.S. Lund., 2015. Quantitative trait loci markers derived from whole genome sequence data increases the reliability of genomic prediction. J Dairy Sci. 98(6):4107

Browning, S. R. and B. L. Browning., 2009. A unified approach to genotype imputation and haplotype phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84: 210

Daetwyler, H. D., R. Brauning, A. J. Chamberlain., *et al*., 2017. 1000 bull genomes and sheep genome db projects: enabling cost-effective sequence level analysis globally. AAABG 2017.

Das, S, L. Forer, S. Schönherr, *et al*., 2016. Next-generation genotype imputation service and methods. Nature Genetics. 48: 1284

Gilmour, A. R., B. G. Gogel, B. R. Cullis, R.Thompson., 2009. ASReml User Guide Rrelease 3.0.

HemelHempstead:VSN International Ltd

Lee and van der Werf., 2016. MTG2: an efficient algorithm for multivariate linear mixed model analysis based on genomic information. Bioinformatics 32: 1420

Loh P.R,. P. Danecek. P.F. Palamara. *et al*., 2016. Reference-based phasing using the Haplotype Reference Consortium panel. Nature Genetics 48:1443

Moghaddar, N., A. Swan and J.H.J Van der Werf, J.H.J., 2017. Genomic prediction from observed and imputed high-density ovine genotypes Genet. Sel. Evol. 49: 40

Van Binsbergen. R., M. Calus, M. Bink, *et al*., 2015. Genomic prediction using imputed whole-genome sequence data in Holstein Friesian cattle. Genet Sel Evol 47:71

Van den Berg, I., D. Boichard and M.S. Lun., 2016. Sequence variants selected from a multi-breed GWAS can improve the reliability of genomic predictions in dairy cattle. Genet Sel Evol. 48: 83

VanRaden P.M., M.E. Tooker, J.R O'Conell *et al*., 2017. Selecting sequence variants to improve genomic predictions for dairy cattle. Genet. Sel. Evol 49:32

Yang J., B. Benyamin., B.P., McEvoy., *et al*., 2010. Common SNPs explain a large proportion of heritability for human height. Nature Genetics. 42(7): 565