Exploiting sequence variants for genomic prediction in Australian sheep using Bayesian models

M. Khansefid[1,2], S. Bolormaa[1,2], A. A. Swan[2,3], J. H. J. van der Werf[2,4], N. Moghaddar[2,4], N. Duijvesteijn[2,4], H. D. Daetwyler[1,5,2] & I. M. MacLeod[1,2]

[1] Agriculture Victoria, AgriBio Centre for AgriBioscience, Bundoora, VIC 3083, Australia
majid.khansefid@ecodev.vic.gov.au (Corresponding Author)
[2] Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW 2351, Australia
[3] Animal Genetics and Breeding Unit, University of New England, Armidale, NSW 2351, Australia
[4] School of Environmental and Rural Science, University of New England, Armidale, NSW 2351, Australia
[5] School of Applied Systems Biology, La Trobe University, Bundoora, VIC 3086, Australia

Summary

The accuracy of genomic predictions could be potentially improved by creating competitively priced low to medium density custom SNP chips, that include sequence SNPs strongly associated with a range of economically important traits. The SheepGenomesDB and Australia Sheep CRC have recently completed whole-genome sequencing of 726 sheep, enabling the imputation of approximately 46,000 Australian sheep of multiple breeds and crosses that were previously genotyped with lower density SNP chips. Subsets of these sheep are recorded for a range of growth and meat quality traits. We used this dataset to discover putative causal SNPs associated with these traits and then combined these SNPs with the 50k SNP chip genotypes for Bayesian genomic prediction. The genomic predictions were validated in purebred Merino and Border Leicester × Merino crossbreds. On average there was a 5% increase in the accuracy of genomic breeding values by adding the top sequence SNPs to the 50k SNP genotypes compared to using only the 50k genotypes.

Keywords: sequence variants, genomic prediction, BayesR, meat quality, growth, sheep

Introduction

In genomic prediction, dense marker information and phenotypes in the reference population are used to estimate the markers effects. These estimated marker effects, can then be used to calculate the genomic estimated breeding values (GEBVs) for the target or validation animals which are genotyped, but do not have any phenotypic records (Meuwissen *et al.*, 2001).

Genomic prediction provides an attractive alternative to traditional selection for hard to measure traits or traits that cannot be measured in the selection candidates, such as milk production in dairy bulls and meat quality in sheep. However, in comparison with dairy cattle, the adoption of genomic selection in sheep genetic evaluations needs extra considerations due to the diversity of breeds and composites resulting in small reference population sizes within breed and high genotyping costs relative to economic returns (van der Werf *et al.*, 2014). Therefore, it is important to keep genotyping costs as low as possible and to increase the reference population size through the use of multi-breed populations for genomic evaluations. So far, information from other breeds has shown to be of limited value in genomic prediction.

This could be because the genotyped markers may not always tag the same QTL across breeds (Goddard *et al.*, these proceedings). Although it is possible to impute genotyped animals to whole-genome sequence (WGS), for routine genetic evaluations the use of millions of genotypes is not computationally feasible and the accuracy of imputation is less than real genotyping (Bolormaa *et al.*, these proceedings). Ideally then, the industry requires a relatively inexpensive customized low to medium density SNP chip, that includes putative causal mutations from sequence affecting traits of interest.

In this study we used a multi-breed reference population in sheep to assess the potential of exploiting imputed WGS to increase the accuracy of genomic predictions. Specifically, we investigated 6 growth and meat quality traits. We applied Bayesian methods to compare genomic prediction using only a 50k SNP marker panel versus a 50k panel with added potential QTL variants discovered in imputed sequence.

Material and methods

Animals and phenotypes

The animals were a mixture of breeds and crosses that had been phenotyped and genotyped as part of the Sheep CRC dataset and industry evaluations. They were divided to three non-overlapping groups. The aim was to generate a QTL discovery population, a reference population for genomic prediction and a validation population, that were independent of each other to avoid bias. The animals for the validation group were selected to have the lowest possible genetic relationships with the reference sets, to ensure that the genomic prediction is least affected by these genetic relationships. The number of animals in the trait specific reference and validation sets are shown in Table 1. The validation consisted of two breed groups with large numbers of genotyped animals in our study: 1) purebred Merino (MER) and 2) Border Leicester × Merino crossbreds (BL × MER). There was a more diverse mixture of breeds and crosses in both the reference and QTL discovery populations. The QTL discovery population was used only to identify potential causal mutations using a genome-wide association study (GWAS) with imputed WGS. The genomic prediction equations were derived only from the reference population, and these were validated separately in each of the two validation groups.

We studied 6 traits, consisting of carcass fat depth at C site (CCFAT), carcass and post-weaning eye muscle depth (CEMD and PEMD), intermuscular fat percentage (IMF), shear force measured at day 5 after slaughter (SF5) and post-weaning weight (PWT). The descriptions of the phenotypes are available in Bolormaa *et al.* (2016). Phenotypes pre-adjusted for various fixed effects including birth-rearing type, sex, and contemporary groups.

Genotypes and QTL discovery

All animals had real or imputed HD genotypes and were imputed to WGS (details given in Bolormaa *et al.*, these proceedings). When using the 50k panel for genomic prediction, we used 36,955 SNPs with known location and minor allele frequency (MAF) > 0.005. The imputed WGS included 31,154,082 variants. Genotypes on the X chromosome were excluded.

We used only the discovery population for each trait to run a GWAS using the 31 million imputed genotypes. We set a lenient GWAS p-value threshold of $<10^{-3}$ and from the start of the chromosome selected the most significant SNPs below this threshold within a 100

Kb window, and sliding 50 Kb, repeating this along each chromosome. After selection of these top sequence variants, PLINK software (Purcell *et al.*, 2007) was used to prune one of any pair of SNPs with an r^2 linkage disequilibrium score > 0.95. The GWAS model tested each of the sequence variants, one at a time, for each trait using the Wombat software (Meyer, 2007):

$$\mathbf{y= Xb + s_i\alpha_i + Qq + Zg + e} \quad (1)$$

in which **y** and **b** are vectors of phenotypes and fixed effects, respectively, $\mathbf{s_i}$ is the vector of genotypes (coded 0,1,2) for each animal at the i^{th} SNP fitted as a covariate, $\boldsymbol{\alpha_i}$ is the i^{th} SNP effect, $\mathbf{q} \sim N(0, \mathbf{I}\sigma^2_q)$ contains random breed effects and **Q** is a matrix with breed proportion of each animal according to pedigree information and σ^2_q is the variance among breed groups, $\mathbf{g} \sim N(0, \mathbf{G}\sigma^2_{SNP})$ contains random additive effects where **G** is the genomic relationship matrix (GRM) and σ^2_{SNP} is the variance explained by all SNPs. The GRM was constructed with HD SNPs, **e** is the vector of random residual effects, **X**, **Q** and **Z** are the design matrices connecting phenotypes to their corresponding fixed effect, random breed effect and random additive effect, respectively. Variance components for random effects were first computed without $\mathbf{s_i\alpha_i}$ in the model and were then fixed for these values in the GWAS analysis.

Genomic prediction

We applied the BayesR method (Erbe *et al.,* 2012) and the BayesRC method (MacLeod *et al.,* 2016) for genomic prediction. Prior to running the Bayesian analysis, the reference and validation phenotypes were pre-adjusted for data source, and breed proportions using a model similar to equation 1, but excluding the single SNP effect ($\mathbf{s_i\alpha_i}$). This pre-adjustment of the phenotypes mitigates possible effects due to population structure (such as Merino strain) that could lead to biased results in the validation sets. Prior to analysis genotypes were centred and standardised to a variance of 1. The Bayesian models fitted only the SNP effects, modelled as a mixture of four normal distributions with a mean of zero and variance: $\sigma^2_1=0\sigma^2_g$, $\sigma^2_2=0.0001\sigma^2_g$, $\sigma^2_3=0.001\sigma^2_g$ and $\sigma^2_4=0.01\sigma^2_g$, where σ^2_g is the additive genetic variance. The Bayesian GEBVs were calculated using one of three possible genotype sets: top sequence variants (top), 50k panel (50k), or 50k plus top sequence variants (50k+top).

The key difference between BayesR and BayesRC is that the later allowed for the selected top sequence SNPs to be allocated to a separate category or class than the remaining 50k SNPs. Each category is then independently modelled as a mixture of the four distributions but with the same priors. If the separate category of SNP is enriched for causal variants this can improve the fit of the model. Each Bayesian model was replicated with 5 MCMC chains, each with 40,000 iterations (20,000 burn-in). The accuracy of genomic prediction was calculated as Pearson's correlation between adjusted phenotypes and GEBVs divided by the square root of trait heritability (50k result, Table 2). The bias of predictions was defined as the regression coefficient of adjusted phenotypes on GEBVs.

Results and discussion

The number of top sequence variants and animals in the reference and validation sets are shown in Table 1. The genetic variances explained by SNPs and the estimated heritabilities in different models are shown in Table 2. The amount of σ^2_g explained by the top SNPs was much lower than using the 50k SNPs in the model which indicates that only a

proportion of QTLs were captured in the top sequence variants. Typically GWAS QTL account for only a proportion of the expected genetic variance because they lack power to detect many of the small or rare QTL affecting quantitative traits. Adding the top SNPs to the 50k, increased the amount of σ^2_g explained by the SNPs in BayesR but not in BayesRC.

The accuracy and bias of genomic predictions for the two different validation sets are shown in Figure 1 and 2. To more generally compare the results of different models, the accuracy and bias for each of the two validation groups was averaged for each trait and shown in Table 3. On average, the accuracy of genomic prediction increased by about 5% by adding the top sequence variants to 50k genotypes. The accuracy of predictions increased in BayesRC in comparison with BayesR when the top SNPs were highly predictive (such as PWT). However, this improvement was marginal probably because the added sequence SNPs were those with relatively large effects, and BayesR may already have captured them appropriately by allocation to the distribution with the highest variance. Although using only top SNPs to calculate GEBVs gave an average accuracy similar to the 50k, the bias of predictions for each of the breeds in the validation set was much more variable with top SNPs alone than for denser genotypes (Figure 1 and 2). For example, the regression of genomic predictions on phenotypes for CCFAT in pure Merino and Border Leicester × Merino crossbreds were 0.68 and 1.65 when using only top SNPs, and this reduced to 1.10 and 0.92 using 50k and top SNPs in the BayesR model. This suggests that these top sequence variants are not always segregating in all validation breeds or their effects are not equal across different breeds.

In conclusion, adding sequence SNPs associated with economic traits and adding them to low density SNP panels can increase the accuracy of genomic prediction while minimising genotyping costs for industry applications.

Acknowledgements

The authors would like to acknowledge the Cooperative Research Centre for Sheep Industry Innovation (Sheep CRC) for funding of this project, Sheep Genetics and MLA for providing access to phenotypic data from industry animals and we would also like to acknowledge Klint Gore (University of New England, Armidale, NSW, 2351, Australia) for preparing the SNP chip genotype data and Sheep CRC and Sheep Genetics staff across Australia for recording the phenotypic data.

List of References

Bolormaa, S., B. J. Hayes, J. H. J. van der Werf, D. Pethick, M. E. Goddard, & H. Daetwyler, 2016. Detailed phenotyping identifies genes with pleiotropic effects on body composition. BMC genomics 17(1), 224.

Erbe, M., B. J. Hayes, L. K. Matukumalli, S. Goswami, P. J. Bowman, C. M. Reich, B. A. Mason, & M. E. Goddard, 2012. Improving accuracy of genomic predictions within and between dairy cattle breeds with imputed high-density single nucleotide polymorphism panels. J. Anim Sci. 95(7): 4114-4129.

MacLeod, I. M., P. J. Bowman, C. J. Vander Jagt, M. Haile-Mariam, K. E. Kemper, A. J. Chamberlain, C. Schrooten, B. J. Hayes & M. E. Goddard, 2016. Exploiting biological priors and sequence variants enhances QTL discovery and genomic prediction of complex traits. BMC genomics 17(1): 144.

Meuwissen, T. H., B. J. Hayes & M. E. Goddard, 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157(4): 1819-1829.

Meyer, K., 2007. WOMBAT—A tool for mixed model analyses in quantitative genetics by

restricted maximum likelihood (REML). Journal of Zhejiang University-Science B 8(11): p 815-821.

Purcell, S., B. Neale, K. Todd-Brown, L. Thomas, M. A. Ferreira, D. Bender, J. Maller, P. Sklar, P. I. De Bakker, M. J. Daly, and P. C. Sham, 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. 81(3): 559-575.

van der Werf, J. H. J., R. G. Banks, S. A. Clark, S. J. Lee, H. D. Daetwyler, B. J. Hayes & A. A. Swan, 2014. Genomic selection in sheep breeding programs. In Proceedings of the 10th World Congress of Genetics Applied to Livestock Production 17-22.

Table 1. Number of top-SNPs and animals in discovery, reference and validation sets in different traits.

		Number of animals		Number of animals in validation set	
Trait *(unit)*	top-SNPs	Discovery	Reference	MER	BL × MER
CCFAT *(mm)*	4,426	4,452	7,635	912	536
CEMD *(mm)*	4,377	4,473	7,714	904	519
PEMD *(mm)*	4,283	7,114	9,715	1,766	586
IMF *(%)*	4,354	3,905	6,353	843	474
SF5 *(N)*	4,901	4,344	7,392	868	531
PWT *(Kg)*	4,652	8,937	11,067	3,118	543

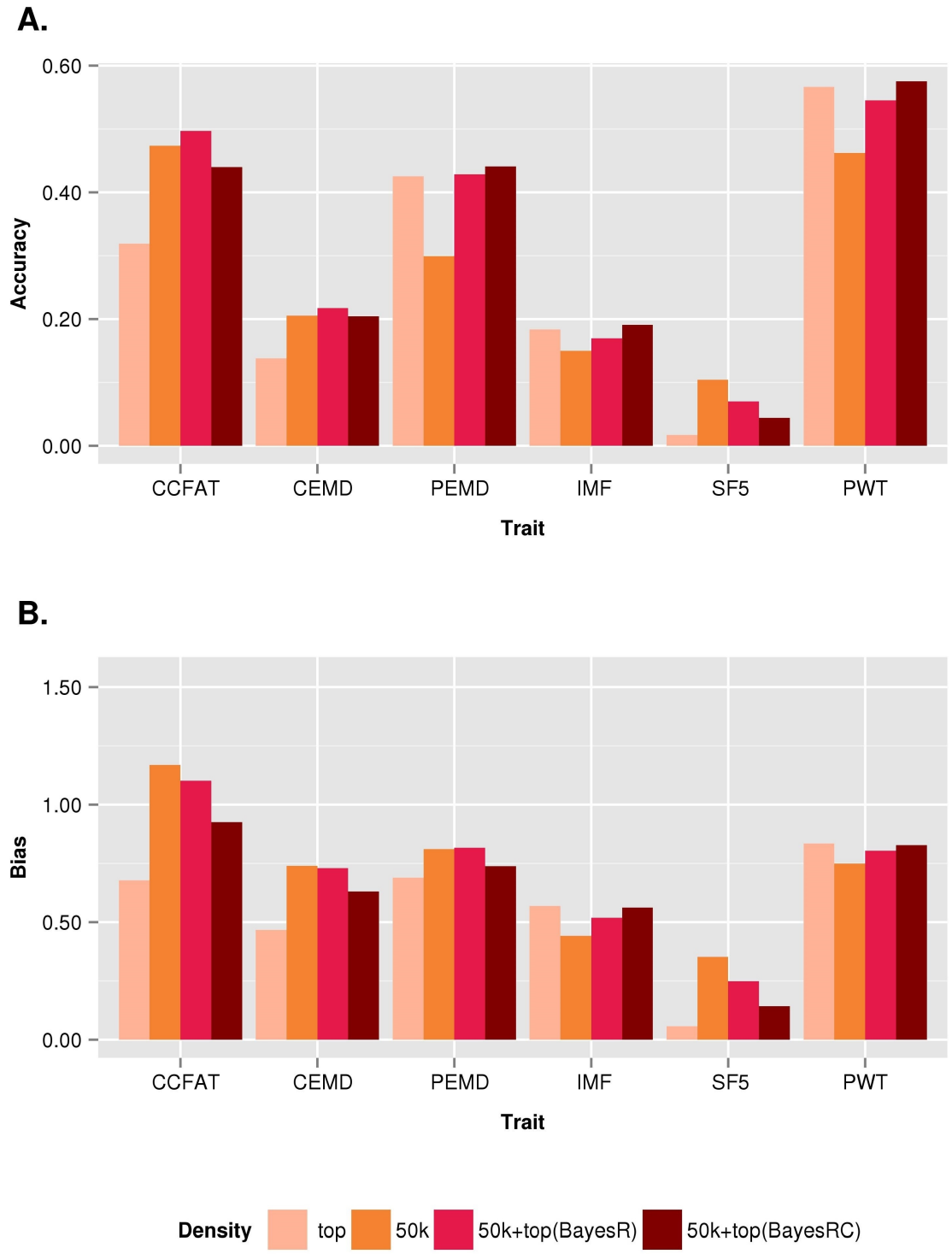
Table 2. The estimated genetic variance (and heritability) in different models.

	BayesR			BayesRC
Trait *(unit)*	top	50k	50k+top	50k+top
CCFAT *(mm)*	0.24 (0.09)	0.49 (0.19)	0.51 (0.20)	0.49 (0.19)
CEMD *(mm)*	0.55 (0.07)	1.19 (0.14)	1.28 (0.15)	1.27 (0.15)
PEMD *(mm)*	0.63 (0.13)	1.13 (0.24)	1.22 (0.26)	1.15 (0.24)
IMF *(%)*	0.11 (0.16)	0.26 (0.37)	0.26 (0.37)	0.25 (0.36)
SF5 *(N)*	6.20 (0.14)	10.71(0.23)	11.30(0.25)	10.27(0.23)
PWT *(Kg)*	3.82 (0.12)	7.19 (0.21)	7.41 (0.22)	7.18 (0.22)

Table 3. The accuracy (and bias) of genomic prediction averaged across two validation groups (pure Merino and Border Leicester × Merino) in different models.

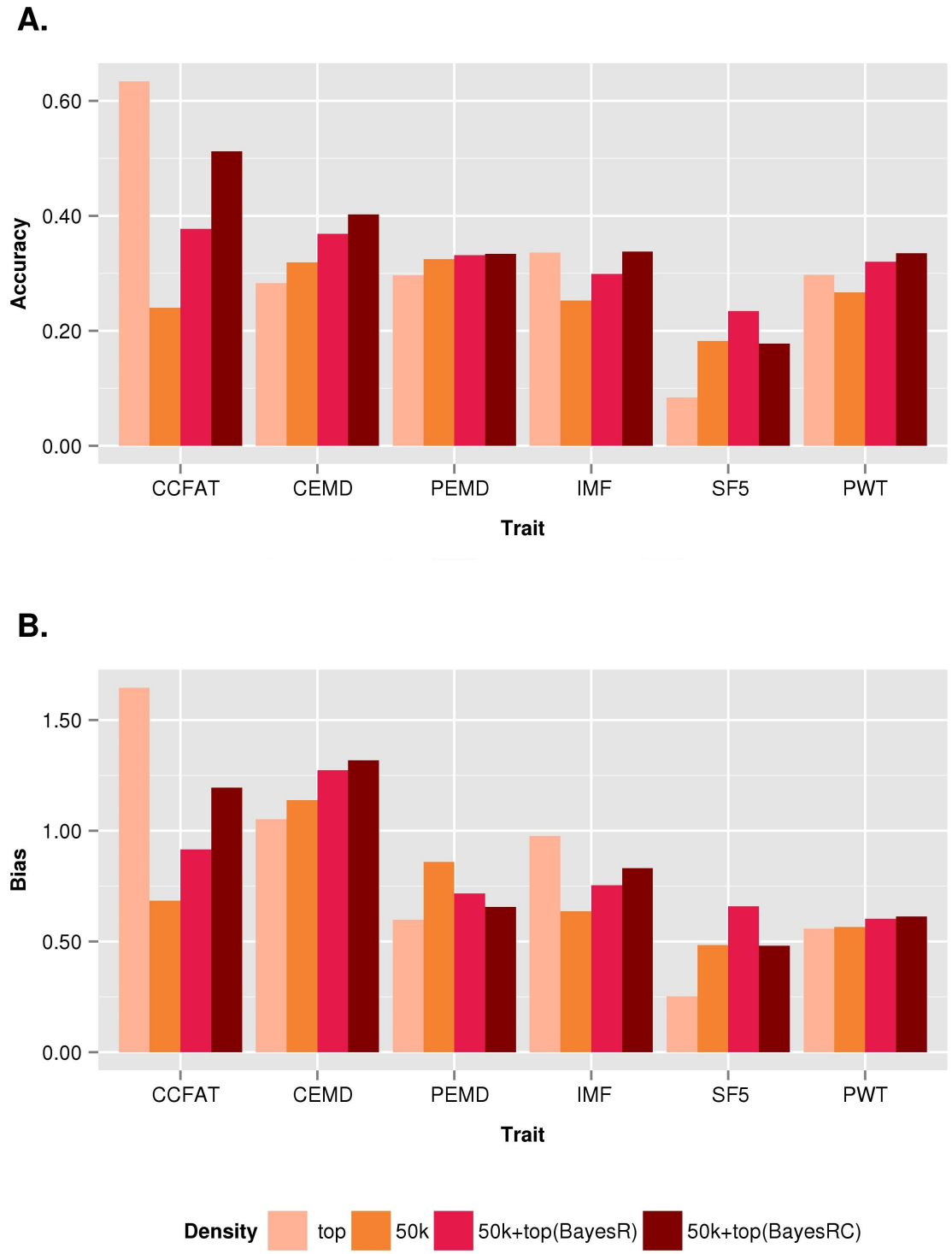
	BayesR			BayesRC
Trait *(unit)*	top	50k	50k+top	50k+top
CCFAT *(mm)*	0.48 (1.16)	0.36 (0.93)	0.44 (1.01)	0.48 (1.06)
CEMD *(mm)*	0.21 (0.76)	0.26 (0.94)	0.29 (1.00)	0.30 (0.97)
PEMD *(mm)*	0.35 (0.63)	0.32 (0.82)	0.38 (0.75)	0.38 (0.68)
IMF *(%)*	0.26 (0.77)	0.20 (0.54)	0.23 (0.64)	0.26 (0.70)
SF5 *(N)*	0.05 (0.15)	0.14 (0.42)	0.15 (0.45)	0.11 (0.31)
PWT *(Kg)*	0.43 (0.70)	0.36 (0.66)	0.43 (0.70)	0.45 (0.72)

Figure 1. Accuracy (A) and bias (B) of genomic predictions in purebred Merino for different traits in different models.



** CCFAT = carcass fat depth at C site, CEMD = carcass eye muscle depth, PEMD = post-weaning eye muscle depth, IMF = intermuscular fat percentage, SF5 = shear force measured at day 5 after slaughter and PWT = post-weaning weight.*

Figure 2. Accuracy (A) and bias (B) of genomic predictions in Border Leicester × Merino crossbreds for different traits in different models.



** CCFAT = carcass fat depth at C site, CEMD = carcass eye muscle depth, PEMD = post-weaning eye muscle depth, IMF = intermuscular fat percentage, SF5 = shear force measured at day 5 after slaughter and PWT = post-weaning weight.*