Effect of Genetic Architecture on Accuracy of Multi Breed Genomic Prediction

***Y.C.J. Wientjes*[*,†], M.P.L. Calus[*], M.E. Goddard[‡,§] and B.J. Hayes[‡,#,‖].**

[*]Animal Breeding and Genomics Centre, Wageningen UR Livestock Research, Wageningen, the Netherlands, [†]Animal Breeding and Genomics Centre, Wageningen University, Wageningen, the Netherlands, [‡]The Department of Environment and Primary Industries, Bundoora, Australia, [§]University of Melbourne, Parkville, Australia, [#]Dairy Futures Cooperative Research Centre, Bundoora, Australia, [‖]La Trobe University, Bundoora, Australia

ABSTRACT: Objectives were to investigate effect of genetic architecture and including random across and within breed effects in GBLUP on accuracy of multi breed genomic prediction. High-density genotypes and imputed synonymous, missense and premature stop codon mutations using sequence data were available for 3000 Holstein Friesians and 3000 Jerseys. Phenotypes of traits with different genetic architectures, regarding allele frequency spectra and number of breed specific QTL, were simulated by sampling 100 QTL from a mutation class. Accuracies of genomic breeding values were estimated using GBLUP including random across and within breed effects. Increase in accuracy by adding individuals of another breed to the reference population and accuracy of across breed genomic prediction was low. Genetic architecture influenced accuracies; accuracies reduced when QTL allele frequencies were lower and QTL were more breed specific. Including a random within breed effect did not affect accuracies.

Keywords: Multi breed; Genomic prediction; Genetic architecture

Introduction

Accuracy of genomic prediction depends on the size of the reference population; the larger the size of the reference population, the more accurate breeding values can be predicted for other individuals of which genomic information is available (e.g. Meuwissen et al. (2001); Daetwyler et al. (2008)). Therefore, it is appealing to enlarge initially small reference populations by using information of different breeds. However, the added benefit of adding another breed to the reference population may be affected by differences in linkage disequilibrium (LD) between breeds. If the phase of LD between a SNP and a QTL differs between breeds, the apparent effect of the SNP will vary between breeds. In some cases, the QTL might only segregate in some breeds although the SNPs segregate more widely. Estimating SNP effects across as well as within breed might be a way to benefit from increasing the reference population by adding another breed, even when SNP effects differ between breeds. This can for example be done with a GBLUP model including random across and within breed effects. The first objective was to investigate the effect of genetic architecture on accuracy of multi breed genomic prediction. The second objective was to investigate the effect of a GBLUP model with a random across and within breed effect on accuracy of multi breed genomic prediction. Different genetic architectures were simulated by sampling QTL from three different classes of mutations (synonymous, missense, and premature stop codon, imputed from real whole genome sequence data), and allele substitution effects from two different models.

Materials and Methods

Genotypes. High-density genotypes were available for 3000 Australian Jersey cows and 3000 Australian Holstein Friesian cows. SNPs with low quality were deleted using the same criteria as described in Erbe et al. (2012). After the quality check, 606,384 SNPs remained. Genotypes for 80,515 synonymous, 97,296 missense, and 4,064 premature stop codon mutations were imputed using Beagle (Browning and Browning (2007)), based on sequence information of the 1,000 bull genome consortium (Daetwyler et al. (2014)). All imputed mutations were used, independent of the reliability of imputation, to prevent a positive selection for mutations in high LD with a SNP on the used SNP chip. Allele frequency spectra and number of breed specific mutations of the different mutations in imputed data were comparable to real sequence data; minor allele frequencies were highest and number of breed specific mutation lowest for synonymous mutations, followed by missense and finally premature stop codon mutations, which segregated predominately within breed.

Phenotypes. Phenotypes of all individuals were simulated by randomly sampling 100 QTL from imputed and segregating 1) synonymous, 2) missense, or 3) premature stop codon mutations. Two scenarios were used to sample allele substitution effects using an equal chance on a positive or negative effect: 1) RANDOM, in which effects were randomly sampled from a gamma distribution (shape 0.4, scale 1.66), and 2) VAR; in which each QTL explained the same variance, i.e. effects were depending on allele frequency. A purely additive model was assumed to calculate true breeding values (TBVs). The simulated heritability was 0.8 and to ensure an equal heritability per breed, environmental effects were simulated based on TBVs corrected for breed effects. For each scenario, simulations were replicated ten times.

Accuracy of genomic prediction. Each replicate, three different reference populations were used and the accuracy of estimated breeding values was calculated for a validation population of 1,000 randomly selected Holstein Friesian cows and 1,000 randomly selected Jersey cows (Table 1). The data was analyzed with the following GBLUP type of model in ASReml (Gilmour et al. (2009)):

$$\mathbf{y} = \mathbf{1}_n \mu + \mathbf{Z}\mathbf{g}_a + \mathbf{Z}\mathbf{g}_w + \mathbf{e}$$

in which $\mathbf{y}$ is a vector containing simulated phenotypes, $\mathbf{1}_n$ is a vector consisting of ones, μ is the mean, $\mathbf{g}_a$ and $\mathbf{g}_w$ are

genomic breeding values predicted either across or within breed ($\mathbf{g}_a \sim N(0, \mathbf{G}_a\sigma^2_{g_a})$ and $\mathbf{g}_w \sim N(0, \mathbf{G}_w\sigma^2_{g_w})$), $\mathbf{Z}$ is a matrix allocating genomic breeding values to individuals and $\mathbf{e}$ is a vector containing the residuals $\sim N(0, \sigma^2_e)$. The used $\mathbf{G}_a$ matrix rescales pedigree and genomic inbreeding levels of both breeds to a common base immediately before the divergence of the two breeds (Erbe et al. (2012)). The $\mathbf{G}_w$ matrix is formed from the $\mathbf{G}_a$ matrix by setting the across breed elements to zero. In this model, two genomic breeding values were predicted for each animal: one using information from all breeds and one using only information from its own breed. For each validation animal, a genomic estimated breeding value (GEBV) was calculated as the sum of the estimated genomic breeding value across and within breed. The correlation between GEBVs and TBVs represented the accuracy of genomic prediction. The model was also run without a random within breed effect to check the advantage of including this effect in the model.

Table 1. Overview of the number of individuals from each breed used in the different reference populations and as validation animals

	Reference population[1]		Validation animals	
Nr.	Nr. of HF[1]	Nr. of J[2]	Nr. of HF[1]	Nr. of J[2]
1	2,000	2,000	1,000	1,000
2	2,000	500	1,000	1,000
3	2,000	0	1,000	1,000

[1] HF = Holstein Friesian
[2] J = Jersey

Results and Discussion

Accuracies of genomic prediction. Accuracies are shown in Figure 1 for the RANDOM scenario (A) as well as the VAR scenario (B). When the number of Jerseys in the reference population decreased from 2000 to 0, accuracies of Jerseys decreased from 0.56-0.69 to 0.08-0.17 in the RANDOM scenario and from 0.19-0.50 to 0.02-0.07 in the VAR scenario. Accuracies of Holstein Friesian animals decreased by only ~0.01 when the number of Jerseys in the reference population decreased. So, the potential benefit of using information from another breed in predicting genomic breeding values is low, especially when rare alleles had a large effect.

For all reference populations and genetic architectures, accuracy was higher in the RANDOM scenario than in the VAR scenario. This is because in both scenarios the effect of QTL with high minor allele frequency (MAF) is estimated more accurately, but in the random scenario these QTL explain more of the variance than in the VAR scenario, where QTL explained the same proportion of the genetic variance regardless of MAF.

In general, accuracies were higher when synonymous mutations were used as QTL and lowest when premature stop codon mutations were used as QTL, with a more pronounced difference in the VAR scenario. This indicates that lower minor allele frequencies and more breed specific QTL resulted in a lower accuracy of genomic prediction, which was expected. Due to ascertainment bias of the SNPs on the chip (Matukumalli et al. (2009)), LD between SNPs and QTL reduces when the allele frequency of QTL becomes more extreme. This means that the SNPs do not explain all the QTL variance which results in a lower accuracy, which is also shown in other studies (Daetwyler et al. (2013); De los Campos et al. (2013)). This effect is particularly strong in the VAR scenario because QTL at low MAF explain more variance than they do in the RANDOM scenario.

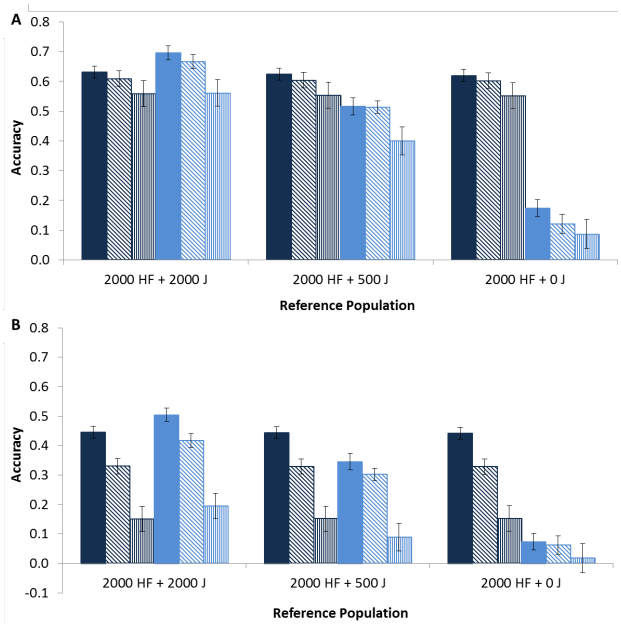

Figure 1 - Accuracies of genomic prediction (± standard errors) for Holstein Friesian (HF; dark) and Jersey (J; light) animals using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance with 100 QTL underlying the trait sampled from synonymous (solid fill), missense (diagonal fill) or premature stop codon mutations (vertical fill) and using 3 reference populations.

Estimated variance components. The estimated across and within breed variance together form the estimated total genetic variance. Estimates for those variances are shown in Figure 2 for the RANDOM scenario (A) and VAR scenario (B). Differences between the replicates were large, resulting in reasonably large standard errors of across breed variances (RANDOM: ~0.11; VAR: ~0.06) and within breed variances (RANDOM: ~0.08; VAR: ~0.05). This indicates that the power to disentangle the across and within breed effect was low.

In the RANDOM scenario (VAR scenario), estimated heritabilities were on average 0.76 (0.56) when QTL were sampled from synonymous mutations; 0.76 (0.41) when QTL were sampled from missense mutations; and 0.70 (0.17) when QTL were sampled from premature stop codons. Those results indicate that estimated heritabilities only slightly underestimated the simulated heritability of 0.8 in the RANDOM scenario, while estimated heritabilities were much lower in the VAR

scenario. This indicates that it was more difficult to pick up all the genetic variance when rare alleles had a large effect.

The proportion of the total genetic variance explained by the within breed component was always very low for the reference populations consisting of only one breed, which was expected. For the other, multi-breed, reference populations, the proportion explained by the within breed component was in the RANDOM scenario (VAR scenario) ~25% (~47%) when QTL were sampled from synonymous mutations, ~38% (~53%) when QTL were sampled from missense mutations, and ~49% (~66%) when QTL were sampled from premature stop codon mutations. So, the proportion of the total genetic variance explained by the within breed variance was larger when the number of breed specific QTL was higher and the minor allele frequency lower. This effect was more pronounced when rare alleles had a large effect, i.e. in the VAR scenario. This was expected, since breed specific QTL do not contribute to the across breed variance. Besides that, LD phase between QTL and SNP is less consistent across breeds for QTL with a lower allele frequency, due to ascertainment bias of the SNPs (Matukumalli et al. (2009)).

For all scenarios, accuracies and genetic variances were equal for the models with or without a random within breed effect. This is probably related to the fact that the power to disentangle the across and within breed effect was low.

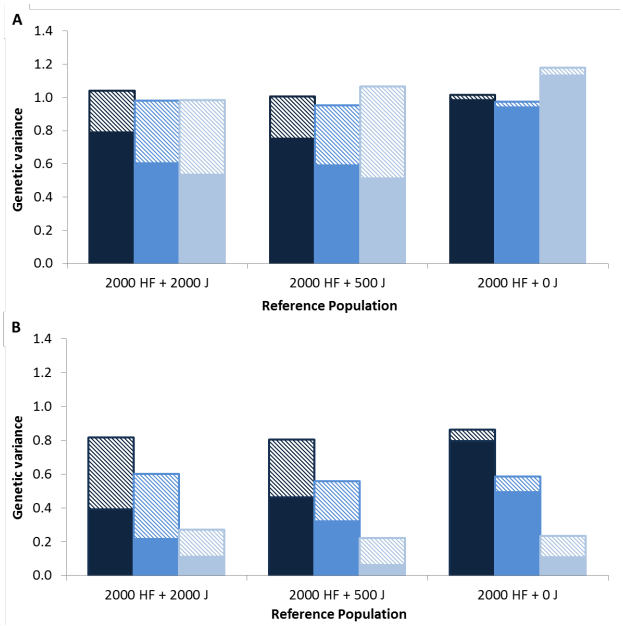

Figure 2 - Across (solid fill) and within (diagonal fill) breed genetic variances using simulated allele substitution effects (A) randomly sampled from a gamma distribution or (B) with each QTL explaining an equal proportion of the genetic variance with 100 QTL underlying the trait sampled from synonymous (dark), missense (medium) or premature stop codon mutations (light) and using 3 reference populations (HF = Holstein Friesian, J = Jersey).

Conclusion

The results show that the potential benefit of using information from another breed in predicting genomic breeding values is low if GEBVs are calculated using BLUP, especially when rare alleles had a large effect. Accuracy of both single breed and multi breed genomic prediction is influenced by the genetic architecture of the QTL underlying the trait, with lower accuracies by decreasing minor allele frequencies of the QTL. Therefore, the genetic architecture (allele frequency spectra of QTL, proportion of QTL segregating across breeds) is demonstrated to be a key parameter determining the accuracy of multi breed genomic predictions. Finally, adding a random within breed effect to a GBLUP model did not influence the accuracy of genomic prediction, most likely because the power to disentangle a random across and within breed effect was low.

Acknowledgements

Financial support from CRV BV (Arnhem, The Netherlands) and of EU FP7 IRSES SEQSEL (Grant no. 317697) is acknowledged. The 1,000 bulls genomes consortium is acknowledged for providing the sequence data, and in particular Paul Stothard for providing the annotations.

Literature Cited

Browning, S. R., and Browning, B. L. (2007). *Am. J. Hum. Genet.* 81: 1084-1097.

Daetwyler, H. D., Calus, M. P. L., Pong-Wong, R. et al. (2013). *Genetics* 193: 347-365.

Daetwyler, H. D., Capitan, A., Pausch, H. et al. (2014). *Nat. Genet.* In press.

Daetwyler, H. D., Villanueva, B., and Woolliams, J. A. (2008). *PLoS ONE* 3: e3395.

De los Campos, G., Vazquez, A. I., Fernando, R. et al. (2013). *PLoS Genet.* 9: e1003608.

Erbe, M., Hayes, B. J., Matukumalli, L. K. et al. (2012). *J. Dairy Sci.* 95: 4114-4129.

Gilmour, A. R., Gogel, B., Cullis, B. et al. 2009. VSN International Ltd, Hemel Hempstead, UK.

Matukumalli, L. K., Lawley, C. T., Schnabel, R. D. et al. (2009). *PLoS ONE* 4: e5350.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). *Genetics* 157: 1819-1829.