How Would Different Models of Genetic Variation Affect Genomic Selection?

S.A. Clark [1,2], J.M. Hickey [1,2], J. H. J. van der Werf [1,2]

Introduction

Genomic selection (GS) is a method to predict breeding values in livestock; however the underlying mechanism by which it predicts is not fully clear. Initially it was thought that GS predicted effects of quantitative trait loci (QTL), in linkage disequilibrium (LD) with markers (Meuwissen et al. 2001). Increasingly, there has been a realization that GS predicts 'relationships' between animals (Habier et al., 2010). Literature suggests that a very small part of the additive genetic variance is explained by variation at known QTL (Maher, 2008). Fearnhead et al, (2004) noted that this is often inconsistent with high estimates of heritability and they proposed that a rare variant model might explain this "missing heritability". Given this debate, it is relevant to obtain greater understanding into what GS is actually predicting for several reasons. Firstly, the LD paradigm predicts that GS can estimate breeding values with higher accuracies as long as marker densities are increasing, possibly even allowing prediction of breeding value across breeds (Goddard et al., 2006). Moreover, accurate prediction would persist for several generations into the future. In contrast, if the relationship paradigm is true, then our predictive ability based on genomic data would remain for only one or two generations ahead. Consequently continuous measurement of phenotypes of individuals that are at least somewhat related to selection candidates would be needed.

Models underlying genetic variation could range from an infinitesimal model based on many genes, each with very small effects to a model based on the effects of a limited number of genes (QTL model). Although experimental data is needed to provide more evidence about the true model underlying genetic variance, simulation can be used to explore the behavior of various prediction methods used in genomic selection. Broadly, these methods vary in how much they allow individual loci to contribute to variation, from equal variation across all loci (gBLUP) to only specific variation at certain loci (Bayes B), and therefore align with the infinitesimal model and the QTL model, respectively. However, it has been shown before that using traditional BLUP when assuming the infinitesimal model is quite robust against drastic deviations from that model (Maki-Tanila and Kennedy, 1986). It is unknown how well the Bayesian methods would perform when the true model of variation is more 'infinitesimal'.

The objectives of this research were to study the accuracy and robustness of various methods used for genomic selection under a range of underlying genetic models, and to compare the accuracy of genomic selection when the validation animals were one generation, several generations and across different selection lines from the prediction animals.

[1] School of Environmental and Rural Science, University of New England, Armidale, NSW, 2351, Australia.

[2] Cooperative Research Centre for Sheep Industry Innovation, Armidale, NSW, 2351, Australia.

Material and methods

Base Genotype Simulations. Genotype simulations were conducted using the Markovian Coalesence Simulator (MaCS) (Chen et al., 2008) to simulate base haplotypes. Thirty chromosomes each with base haplotypes of a 100 cM region (100,000,000 base pairs) of sequence data were simulated with a per site mutation rate of 2.5*10e-8. The present and historical effective population size was based primarily on the results of Villa-Angulo et al. (2009) for Holstein. Consequently, the current Ne was set to 100, the Ne 1,000 years ago being 1,256, the Ne 10,000 years ago being 4,350, and the Ne 100,000 years ago being 43,500. The simulated haplotypes were then dropped through a simulated popluation structure of three base generations containing 2000 animals. The pedigree was then split into two divergent lines each of 10 generations, each generation containing 1000 animals, half male and half female. Five percent of males, were selected randomly and all females had two offspring per generation.

Genetic variation was simulated using 5 different models:, the Infinitesimal model (IM), the QTL model (QM) and a rare variant model (RM), both with 100 and 1000 QTL. The IM was simulated using a traditional polygenic simulation model and breeding values and residual effects were drawn from a random normal distribution. In contrast breeding value for the QM was assigned using the QTL effects and the genotype of each individual animal. QTL were assigned from segregating SNPs in generation 1 of line 1. Effects of each QTL were drawn from an inverted chi-squared distribution with a shape and scale parameter of 0.4 and 1.66 respectively (Meuwissen et al., 2001). Similarly, the RM was simulated as the QM however all QTL were assigned to SNPs with allele frequency <0.01. In both QM and RM all of the genetic variance was explained by QTL. Heritability for all models was 0.3 of total phenotypic variance. To ensure that the proportion of genetic variation remained constant the residual variance of breeding values was scaled using **u`u**/ (n-1) where u is a vector of breeding value of animals in generation 1.

Statistical analyses and breeding value estimation. Three methods were used to estimate breeding values. 1) Bayes B (Meuwissen et al, 2001), a Monte Carlo Markov Chain method which assumes that markers have different variances and allows only a proportion (5% in this study) of SNPs to have an actual effect, implemented using AlphaBayes (Hickey & Tier, 2009). 2) gBLUP, which assumes an equal variance for each marker and uses a genomic relationships matrix among individuals in a reference set and test set allowing it to compute variance components and best linear unbiased prediction (BLUP) from a mixed model and 3) Traditional BLUP which ignores genomic data and relies on information from ancestors using a numerator relationship matrix Both mixed model methods were implemented using ASReml (Gilmour et al., 2006).

Three training populations (2000 animals) were assigned: 1) Generations 1 and 2 of line 1; 2) Generations 8 and 9 of line 1; and 3) Generations 8 and 9 of line 2. Estimations of 60,000 SNP effects in each training set were then used to predict the breeding value of animals in the 10th generation of line one (1000 animals). Eight replicates were performed and the estimated breeding values for each method where compared to the simulated true breeding value. BLUP acted as a control using the entire pedigree of line 1, with the phenotypes of animals from generations 8 and 9 of line 1 to predict the breeding value of generation 10.

Results and discussion

Estimation methods: Bayes B was the most robust method of estimating breeding value using genomic data. It was the most accurate method of predicting breeding value when QTL were present especially when there was a small number of QTL (Table 1). Bayes B is based on a model that allows prediction of large QTL effects and this clearly helps the accuracy of GS if large QTL effects really exist. Similarly under the RV model Bayes B had the ability to estimate rare variants, again it more accurately predicted breeding value when there were few QTL. As the model of variation became more polygenic the superiority of Bayes B decreased. However it still retained a similar predictive ability to gBLUP under the IM. A plausible explanation for this result is that Bayesian GS methods can be similar to mixed models in that they predict breeding values using numerator relationships based on the information in genomic data, rather than using traditional methods based on pedigree relationships. In contrast gBLUP was more constant when the number of QTL varied, although when the validation population was less related the accuracy of prediction using gBLUP was decreased. This is expected given that gBLUP relies on the relatedness of animals to predict breeding value. Furthermore, this was illustrated when both GS methods were applied to the infinitesimal model in training set 2 or 3 as all ability to predict breeding value was lost.

Table 1. Average Correlation Between Estimated and True Breeding Values and average genetic variance

Model		Training Set	Bayes B	gBLUP	BLUP	Genetic Variance
QTL	**100QTL**	**1**	0.83	0.56	0.46	0.32 (0.29-0.34)
		2	0.77	0.37		
		3	0.76	0.33		
	1000QTL	**1**	0.65	0.59	0.47	0.31 (0.27-0.32)
		2	0.49	0.38		
		3	0.47	0.34		
Rare	**100QTL**	**1**	0.73	0.46	0.42	0.2 (0.15-0.37)
		2	0.67	0.26		
		3	0.63	0.21		
	1000QTL	**1**	0.40	0.37	0.36	0.12 (0.06-0.22)
		2	0.31	0.25		
		3	0.25	0.19		
Infinitesimal		**1**	0.39	0.40	0.45	0.29 (0.27-0.32)
		2	-0.01	0.00		
		3	0.00	-0.01		

Model of Genetic Variation: As expected high accuracies were observed when breeding values were predicted under the QM. However, these high accuracies are not observed when GS is used to predict breeding value in real populations and accuracies are commonly closer to 0.5 (Moser et al., 2010). This suggests that for most traits the IM or a similar polygenic model is closer to reality than the QM. In contrast, the results of the RV appeared to be somewhat erratic, and the low correlation was found especially for the BLUP method. Large variability in genetic variance was observed in the later generations and this explains the low correlation. Due to the low allele frequencies of QTL in generation 1, it was easy to 'lose' variation due to drift giving large fluctuations in results. As a consequence of all QTL being rare, changes in the frequency of these alleles had a large effect on the overall genetic variance in the population, both increasing and decreasing the variance. The final model of variation, the infinitesimal model, completely changed the way in which GS predicted breeding value, even the Bayes B method estimated relationships between animals, however accuracies were low. Also as mentioned before without QTL to persist through generations, GS was unable to predict breeding value in training sets 2 and 3.

Conclusion

This research suggests that Bayes B is a superior method to gBLUP as it performs well in all scenarios, even in the IM case. It is the most robust method of analysis because it is able to predict breeding value from QTL effects and also has the ability to estimate relationships if these QTL where no longer evident or if the size of the effects are small. Moreover, the underlying model of variation greatly affects the predictive ability of genomic selection. Given the alternative models, a complete model of any of those tested is unlikely and the true genetic model is more likely to be a combination of all of these models.

References

Chen G. K., Marjoram P. and Wall J. D., (2009) *Genome Res.* 19: 136-142

Fearnhead N. S., Wilding J. L., Winney B., et al. (2004) *Proc Natl Acad Sci. USA* 101:15992–15997.

Gilmour A.R., Gogel B.J., Cullis B.R., et al., (2006) VSN International Ltd, Hemel Hempstead, HP1, 1ES, UK

Goddard M. E., Hayes B., McPartlan H., et al. (2006), *Proc 8th WCGALP*, 22-708

Habier D., Tetens J., Seefried F., et al., (2010) *Genet. Sel. Evol.*, 42:5

Hayes B.J., Bowman P.J., Chamberlain A.J., et al. (2009), *J Dairy Sci* , 92(2):433–443

Hickey, J.M. and Tier B., (2009) AlphaBayes: User Manual. UNE, Australia.

Maher, B., (2008), *Nature*. 456(7218):18-21.

Moser G., Tier B., Crump R.E., etal., (2009) *Genet. Sel. Evol.*, 41:56

Maki-Tanila A., and Kennedy B. W., (1986) *Proc 3rd WCGALP*, XII: 443–448.

Meuwissen T. H. E., Hayes B. J., and Goddard M. E. (2001) *Genetics*, 157:1819-29

Villa-Angulo R., Matukumalli L. K., Gill C. A., et al.,(2009), *BMC Genetics*, 10:19