Proc. Assoc. Advmt. Breed. Genet. 21: 449-452

OPTBR: COMPUTATIONALLY EFFICIENT GENOMIC PREDICTIONS AND QTL MAPPING IN MULTI-BREED POPULATIONS

Tingting Wang[1,2,3], Yi-Ping Phoebe Chen[1], Kathryn E. Kemper[4], Michael E. Goddard[2,3,4] and Ben J. Hayes[1,2,3]

[1] Faculty of Science, Technology and Engineering, La Trobe University, Victoria, Australia
[2] AgriBio, Centre for AgriBioscience, Biosciences Research, DEDJTR, VIC 3083, Australia
[3] Dairy Futures Cooperative Research Centre, VIC 3083, Australia
[4] Faculty of Veterinary and Agricultural Sciences, University of Melbourne, Victoria, Australia

SUMMARY

As genomic data used for prediction of complex traits rapidly expand in size, the importance of computational efficiency of genomic prediction algorithms becomes paramount. In this paper we describe an expectation-maximisation (EM) algorithm for genomic prediction (OptBR) with the speed-up scheme that is up to 30 times faster than MCMC implementations. The algorithm is flexible for joint analysis of data from different sources, as it includes weightings for the accuracy of phenotype, and can accommodate effects of factors such as breed, age, sex and additional covariates. A further advantage of the method is that QTL mapping is performed simultaneously with genomic prediction.

INTRODUCTION

Genomic predictions are increasingly used to identify breeding individuals in livestock and crop improvement programs. The prediction equation to calculate genomic predictions is derived from a reference population genotyped for thousands of single nucleotide polymorphisms (SNPs), and with phenotypes for the target trait (Meuwissen *et al.* 2001), or through an alternative implementation where genomic relationships derived from the SNP are used to predict breeding values for selection candidates (e.g. VanRaden 2008). Across many species, a key finding is that reference populations must be very large to achieve high accuracies of genomic prediction. One way to increase the size of the reference population is to combine information across populations from the same species. For example in dairy and beef cattle small to moderate increases in prediction accuracy have been reported by using a multi-breed reference population (Lund *et al.* 2014; Kemper *et al.* 2015; Bolormaa *et al.* 2013). Another finding from these studies is that the increase in accuracy of prediction from combining information across populations can depend on the method of prediction.

For multi-breed predictions, methods which assume *a priori* that SNP effects are all non-zero and small, and all from the same normal distribution (SNP-BLUP and GBLUP) do not perform as well as methods that assume *a priori* that some SNP may have zero, small or moderate to large effects (BayesB, or BayesR) (Lund *et al.* 2014; Kemper *et al.* 2015). Compared to BLUP methods, these models use priors which assume a large proportion of SNP have effects close to zero, or actually zero, while a small proportion of SNP have moderate to large effects. This is important not only to improve genomic predictions across breeds, but also to improve the precision of QTL mapping using such methods. While the Bayesian methods are very attractive, the major difficulty with these methods is long computation time, which becomes intractable with very large data sets. The long computational time arises because Bayesian methods are typically implemented using MCMC. To speed up Bayesian methods, several heuristic convergence methods have been proposed e.g. fastBayesB (Meuwissen *et al.* 2009) or fastBayesA (Sun *et al.* 2012). All of these methods reported reduced computation time but in some cases the prediction accuracy was reduced compared to their MCMC counterparts.

Our aim was to develop a computationally efficient algorithm (OptBR for Optimized BayesR) for simultaneous multi-breed prediction and QTL mapping. OptBR implements an EM algorithm on the hierarchical prior assumption for SNP effects and other parameters from BayesR (Erbe *et al.* 2012). Also, OptBR retains the advantage of Predicted Error Variance (PEV) correction of emBayesR (Wang *et al.* 2015) to improve the accuracy. OptBR has four improvements compared with emBayesR which allow it to be applied to very large data sets, which may encompass multiple populations. These advantages include 1) weighting of phenotypes to allow for different errors in measurement across populations; 2) multi-breeds are accounted for by introducing fixed effects into the prediction models; 3) a polygenic term to capture variation not explained by the SNP, and 4) a speed-up scheme to make it 30 times faster than BayesR implemented with MCMC.

MATERIALS AND METHODS

Genotypes and phenotypes. OptBR was implemented on 630K SNPs panels (with total 632,003 SNPs), that was imputed from 777K and 54K Illumina Bovine SNP arrays. Phenotypes for milk yield, protein yield, fat% and fertility were daughter trait deviations (DTD) for bulls, and trait deviations (TD) for cows. For genomic prediction, the data was separated into references set and validation sets. The reference data included 16,214 Holstein and Jersey bulls and cows, while the validation set included 251 Holstein bulls (bulls born after 2007), or a third breed, 114 Australian Red bulls (Australian reds bulls were never included in the reference set).

Data Model. The statistical model is $\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}\mathrm{g} + \mathbf{Wv} + \mathbf{e}$ where $\boldsymbol{\beta}$ is a vector of fixed effects including breed, $\mathbf{g}$ is a vector of the SNP effects, $\mathbf{v}$ is a vector of polygenic effects ~ $N(0, \mathbf{A}\sigma_{\mathrm{v}}^2)$, $\mathbf{e}$ is a vector of residuals ~ $N(0, \mathbf{E}\sigma_{\mathrm{e}}^2)$ where $\mathbf{E}$ is diagonal and accounts for error in TD and DTD, with σ_{e}^2 the error variance. Three design matrices $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ allocate phenotype ($\mathbf{y}$) to the vectors $\boldsymbol{\beta}$, $\mathbf{g}$, and $\mathbf{v}$ separately. The SNP effects are assumed to be drawn from a mixture of normal distributions with zero mean and variance either $0\ or\ 0.0001 * \sigma_{\mathrm{g}}^2\ or\ 0.001 * \sigma_{\mathrm{g}}^2\ or\ 0.01 * \sigma_{\mathrm{g}}^2$ with probability $\mathbf{Pr}_{\boldsymbol{k}}$ $(k = 1 \cdots 4)$ drawn from a Dirichlet distribution with parameters (1,1,1,1).

Expectation maximisation algorithm. To implement the EM algorithm we rewrite the statistical model for the i^{th} SNP as $\mathbf{y} = \mathbf{X\beta} + \mathbf{Z}_{\mathbf{i}}\mathrm{g}_{\mathrm{i}} + \mathbf{u_1} + \mathbf{Wv} + \mathbf{e}$ where $\mathbf{u_1} = \boldsymbol{Zg} - \mathbf{Z}_{\mathbf{i}}\mathrm{g}_{\mathrm{i}}$, that is $\mathbf{u_1}$ is the sum of all SNP effects other than SNP *i*. This form of the model allows us to treat $\mathbf{u_1}$ as missing data and take expectations of the likelihood over $\mathbf{u_1}$ and hence estimate g_{i} allowing for the errors in the estimates of all the other SNP effects. We take the expectation of the log Likelihood of $\mathbf{y}$ using $Var(\mathbf{u_1}|\mathbf{y}) = \mathbf{PEV}(\mathbf{u_1})$ where the prediction error variance (PEV) is derived from a BLUP approximation to the mixture model. We then maximize the expected likelihood with respect to each of the parameters including g_{i}, the mixing proportions ($\mathbf{Pr}$), $\boldsymbol{\beta}$ and $\mathbf{v}$ as well as σ_{e}^2. We also trialled a speed-up scheme: when the SNP effect g_{i} is very small ($|\mathrm{g}_{\mathrm{i}}| \geq 0.00000001$) after 50 iterations, it was not updated in future iterations but left at its current value.

RESULTS AND DISCUSSION

To compare computing times for OPTBR and BayesR, three reference data sets related to milk yield were used, which have 632,003 SNPs with different numbers of animals ranging from 3,049 in RefI (Holstein bulls Only), 11,527 in RefII (Holstein bulls and cows), to 16,214 in RefIII (Holstein and Jersey bulls and cows) seen in Figure 1. The results demonstrate the advantage of OptBR over BayesR, and the advantage of the speed-up scheme. For instance, in the largest dataset time to convergence was 720 hours for BayesR but 28 hours for OptBR_Sp.

The accuracies of prediction using the EM were similar to BayesR with the exception of fat% (Table 1). A detailed investigation of the speed-up scheme was assessed using milk yield. Table 1 shows that the speed up procedure did not sacrifice any accuracy (Table 1).

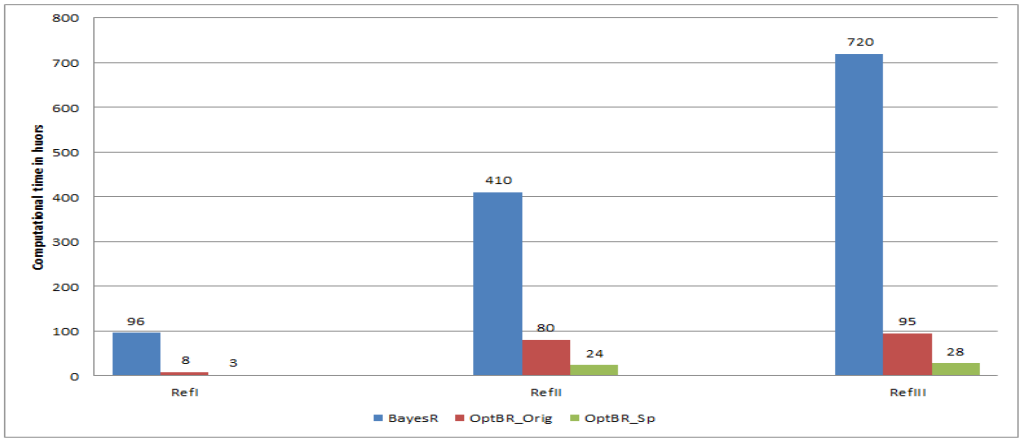

Figure 1. The computational time in hours compared between BayesR, OptBR_Orig, and OptBR_Sp on three reference data sets (RefI with 3,049 animals**,** RefII with 11,527 animals**,** and RefIII with 16,214 animals**).**

Table 1. The impact of the speed-up scheme C1 on accuracy (Acc.), the proportion of variants in each distribution (Pr) and error variance (σ_e^2) using milk yield as an example.

	Acc.	Pr	σ_e^2
OptBR_Orig	0.66	[0.998371, 0.001583, 0.000007, 0.000039]	239409
OptBR_Sp	0.68	[0.997545, 0.002394, 0.000009, 0.000052]	247965

The results in Table 2 demonstrate the robust prediction ability of our algorithm OptBR for multi-breeds and across breed prediction. On milk production traits, both BayesR and OptBR have 3%~7% advantage over GBLUP. On the fertility, three methods had the similar performance. The prediction accuracy for Australian red bulls was not as high as for Holstein, which is not surprising given there were no Australian Reds in the data set. The bias is the coefficient of regressing the phenotype of validation set on Genomic Estimated Breeding Value (GEBV), which shows the underestimation of three methods for SNP effects on most of the traits except Fertility.

Table 2. The accuracy (Acc.) and bias of predictions for BayesR, GBLUP and OptBR from the Holstein and Jersey multi-breed reference population using either the Holstein or Australian Red validation populations.

	Milk Yield		Protein Yield		Fat%		Fertility	
	Acc.	Bias	Acc.	Bias	Acc.	Bias	Acc.	Bias
Holstein validation								
BayesR	0.68	0.84	0.68	0.88	0.81	0.90	0.44	1.53
GBLUP	0.63	0.83	0.65	0.85	0.74	0.85	0.44	1.66
OptBR	0.68	0.90	0.68	0.79	0.77	0.83	0.44	1.27
Australian Reds validation								
BayesR	0.22	0.60	0.12	0.49	0.45	0.92	0.27	1.03
GBLUP	0.16	0.54	0.11	0.51	0.32	0.90	0.29	0.97
OptBR	0.24	0.70	0.12	0.42	0.41	0.89	0.29	1.10

We compared the ability of BayesR and OptBR to map QTL by investigating the number of SNPs with high posterior probabilities of having a non-zero effect (Figure 2). The number and position of QTL was similar between BayesR and OptBR. For milk yield, similar to BayesR, OptBR finds SNPs near to the genes *CSF2RB* located on chromosome 5, SNPs near the casein complex on chromosome 6 (~87Mb), and SNPs related to *CCL28/GHR* on chromosome 20. The well-known gene *DGAT1* (on chromosome 14) is mapped by both BayesR and OptBR.

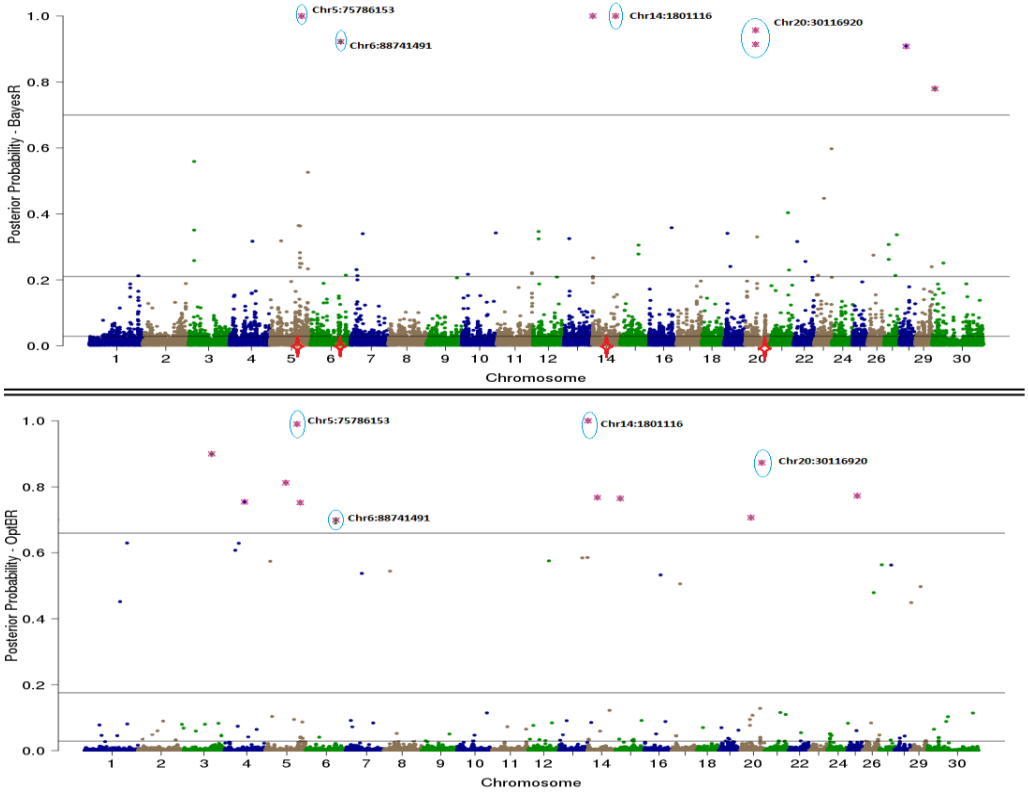

Figure 2. Posterior probability of non-zero SNP effect for milk yield from BayesR (top) and OptBR (bottom) across all chromosomes.

The results suggest that OptBR will be useful for simultaneous genomic prediction and QTL mapping, particularly for very large data sets where computational efficiency is very important.

REFERENCES

Meuwissen T.H.E., Hayes B.J. and Goddard M.E. (2001) *Genetics.* **157**:1819.
VanRaden P.M. (2008) *J Dairy Sci*. 91(11):4414-23.
Lund M.S., Su G., Janss L., Guldbrandtsen B. *et al.* (2014) *Livest Sci*. **166:**101.
Yang J., Benyamin B., McEvoy B.P, Gordon S., *et al.* (2010). *Nat Genet.* **42:**565.
Kemper K.E., Reich C.M., Bowman P.J., Vander Jagt C.J., *et al.* (2015) *Genet Select Evol*. **47**:29.
Bolormaa S., Pryce J.E., Kemper K., Savin K., *et al.* (2013) *J Anim Sci*. **91**:3088.
Meuwissen T.H., Solberg T.R., Shepherd R. and Woolliams J.A. (2009) *Gen Sel Evol*. **41**:2.
Sun X., Qu L., Garrick D.J., Dekkers J.C.M. *et al.* (2012) *PLOS ONE*. **7(11)**: e49157.
Wang T., Chen Y.P.P., Goddard M.E., Meuwissen T.H. *et al. Gen Sel Evol*. **47**:34
Erbe M., Hayes B.J., Matukumalli L.K., Goswami S., *et al.* (2012) *J. Dairy Sci*. **95**:4114.