Proc. Assoc. Advmt. Anim. Breed. Genet. 20:463-466

USING CROSS-VALIDATION IN A FAST EM ALGORITHM FOR GENOMIC SELECTION AND COMPLEX TRAIT PREDICTION

R.K. Shepherd[1], M.J. Drumm[1] and J. Yang[2]

[1]School of Engineering & Technology, Central Queensland University, Rockhampton QLD 4702
[2]The University of Queensland Diamantina Institute, Woolloongabba, QLD 4102

SUMMARY

This paper reports on changes to the EM algorithm emBayesB which estimates QTL effects using dense genome-wide SNP marker data. To overcome convergence issues, modifications were made to the original algorithm which included cross-validation for the estimation of model parameters. The modified algorithm called emBayesB_CV was used to analyse a trait simulated on real human genotypes consisting of 294,831 SNP measured on 3925 individuals. Three datasets were simulated for a trait determined by 10, 100 or 1000 additive QTL. The results showed that the modified algorithm emBayesB_CV was not only computationally fast, but also more accurate than GBLUP in predicting breeding value. However prediction accuracy declined as the size of QTL effects decreased due to the result that although emBayesB_CV could accurately locate the chromosomal location of large QTL effects, this was not the case for small QTL effects.

INTRODUCTION

Genomic prediction of breeding values is a new tool for selection in livestock breeding programs and for risk prediction with complex human diseases. In animal breeding genomic selection uses information from high-density genome-wide SNP markers to predict the breeding value of candidates for selection. Firstly SNP effects have to be estimated in the population of interest by analysing the relationship between phenotype and the SNP genotypes (called training). Then genomic estimated breeding values (GEBV) are calculated by summing the estimated SNP effects across the genome of each candidate. Usually the accuracy of GEBV is assessed in an independent dataset by calculating the correlation between GEBV and either True Breeding Value (TBV) or phenotype (called validation). Bayesian models can be used to include important prior beliefs about the QTL effects and are usually more accurate than BLUP methods using the realised relationship matrix (called GBLUP). But Bayesian prediction is computationally slow for large SNP panels, whereas GBLUP is much faster. emBayesB is an Expectation Maximisation (EM) algorithm which not only incorporates important prior information about QTL effects, but is also computationally fast like GBLUP. However convergence issues are known to occur with emBayesB unless arbitrary bounds are placed on the estimated parameter of the SNP effect distribution like in Shepherd *et al*. (2010). This paper investigates cross-validation for parameter estimation in addition to other modifications to the emBayesB algorithm.

MATERIALS AND METHODS

EM theory. Full details are in Shepherd *et al*. (2010). If we knew which SNP were in linkage disequilibrium (LD) with QTL, then the problem would be much easier. So we assume *a priori* that a fraction γ of the SNP are in LD with QTL and that SNP in LD with QTL have effects from a double exponential (DE) distribution with parameter λ. A linear data model $\mathbf{y} = \mathbf{Bg} + \mathbf{e}$ is assumed to relate phenotype y_i of individual *i* to the j^{th} SNP effect g_j where element b_{ij} of the $n \times m$ matrix $\mathbf{B}$ is the number (0, 1 or 2) of reference alleles (usually standardised) of SNP *j* for individual *i*. The errors are assumed normal and independent such that $\mathbf{y} \mid \mathbf{g} \sim N\left(\mathbf{Bg}, \mathbf{I}\sigma_e^2\right)$.

Using EM theory an iterative sequence of E and M-steps is developed which should converge to maximum *a posteriori* (MAP) parameter estimates. At iteration *k*, the E-step involves the calculation of p_j^k, the posterior probability that SNP *j* is in LD with QTL given the data and all current parameter estimates. This is done analytically and fast. Then given the data and the current values of p_j^k, the M-step calculates $\hat{g}_j = p_j^k DE_{j(\text{mode})}$, $\hat{\gamma} = \frac{1}{m}\mathbf{1}'\mathbf{p^k}$, $\hat{\lambda} = \mathbf{1}'\mathbf{p^k} \Big/ \left|\hat{\mathbf{g}}\right|' \mathbf{p^k}$ and $\hat{\sigma}_e^2 = \frac{1}{n}\left(\mathbf{y} - \mathbf{B}\hat{\mathbf{g}}\right)'\left(\mathbf{y} - \mathbf{B}\hat{\mathbf{g}}\right)$ where $\mathbf{p^k}$ is the vector of posterior probabilities at iteration k and $DE_{j(\text{mode})}$ is the posterior mode of g_j conditional on all current estimates using a DE prior only. Iterating between the E and M-steps the algorithm should converge quickly to produce MAP estimates of $\mathbf{g}$, posterior probabilities $\mathbf{p}$ and ML estimates of $\gamma, \lambda, \sigma_e^2$.

emBayesB_CV. Modifications were made to the original algorithm due to convergence issues. Firstly a complete E-step was performed before updating estimates of each SNP effect g_j. This was not done in the original algorithm. Also the total genetic variance in a dataset was estimated using GBLUP and the estimate of h^2 was fixed in emBayesB_CV. Then the parameter γ was estimated by *k*-fold cross-validation in the training data, with λ being determined by the equation $\lambda^2 = 2m\gamma \Big/ h^2\sigma_y^2$ using the fixed values of h^2 and γ. To speed up convergence, γ was updated each iteration using the M-step equation $\hat{\gamma} = \frac{1}{m}\mathbf{1}'\mathbf{p^k}$ with corresponding updates calculated for λ.

Data simulation. Genotypes of 3925 unrelated humans from GWA Studies were used as described in Yang *et al*. (2010). SNP were randomly selected (from the 294,831 SNP available) to be biallelic QTL and then removed as SNP in the analysis. Three datasets were simulated consisting of 10, 100 or 1000 additive QTL which meant the number of SNP used in each analysis was 294821, 294731 and 293831 respectively. QTL effects were independently simulated from a normal distribution and summed to produce the TBV of each individual. A trait with heritability 0.8 was produced by adding a normally distributed error term to the TBV of each individual. The number of QTL, which explain more than 0.1, 1, 5 and 10% of the total phenotypic variation, was 9, 8, 7 and 5 respectively in the 10 QTL dataset, whereas the number of QTL was 83, 29, 4 and 0 respectively in the 100 QTL dataset. For the 1000 QTL dataset, the number of QTL, which explain more than 0.1, 1 and 5% of the total phenotypic variation, was 297, 4 and 0 respectively.

Statistical analysis. The dataset for each of the three QTL scenarios, was initially split into a training set and a validation set consisting of 3500 and 425 records respectively. The value of γ was estimated using 5-fold cross-validation in the training set. For each γ value, the prediction equation $\mathbf{GEBV = B\hat{g}}$ was estimated using 4 folds (2800 individuals) and then used to calculate GEBV in the left out fold (700 individuals). This procedure was repeated 5 times, so that each fold was left out once, in order to produce GEBV for all 3500 individuals. The value of γ which maximised the correlation between GEBV and phenotype in the training data was chosen. Then this value of γ was used to estimate the SNP effects $\hat{\mathbf{g}}$ using all 3500 training records. To validate the estimated SNP effects $\hat{\mathbf{g}}$, the correlation between TBV and GEBV was calculated for the 425 validation records, as well as the linear regression of TBV on GEBV, which has a slope of 1 if the GEBV are unbiased. The SNP effects were also estimated by GBLUP in the training set

and then validated in the validation dataset. For GBLUP the estimated SNP effects were solutions to the training set equations $(\mathbf{B}'\mathbf{B} + \alpha\mathbf{I})\hat{\mathbf{g}} = \mathbf{B}'\mathbf{y}$ where $\alpha = \sigma_e^2 / \sigma_g^2 = m(1 - h^2) / h^2$.

RESULTS AND DISCUSSION

Using cross-validation to estimate γ, fixing h^2 and using separate E & M-steps solved the problem of convergence which can occur with emBayesB_CV. In Table 1 each emBayesB_CV run uses 5-fold cross-validation with an initial fixed γ and takes approximately 30 minutes on a single compute node of the High Performance Computing (HPC) facility at CQUniversity. Searching for the optimum γ usually took between 10 and 20 such runs which can be run simultaneously on a multiple node HPC facility. Hence for large SNP panels, emBayesB_CV will be significantly faster than a full Bayesian analysis as found by Shepherd *et al*. (2010).

Table 1. Correlation between GEBV and TBV ($\mathbf{r}_{TBV}$) for the validation data of GBLUP and emBayesB_CV using the 10, 100 or 1000 QTL datasets. b is the linear regression coefficient of TBV on GEBV while h^2 is the fixed heritability. The estimated number of SNP in LD with QTL ($m\hat{\gamma}$) is given as well as the estimated parameter $\hat{\lambda}$ of the SNP effect distribution.

No.	No.	Fixed	GBLUP	emBayesB_CV		
QTL	SNP (m)	h^2	$\mathbf{r}_{TBV}$ (b)	$\mathbf{r}_{TBV}$ (b)	$m\hat{\gamma}$	$\hat{\lambda}$
10	294821	0.8	0.15 (0.7)	0.88 (1.0)	93	10.5
100	294731	0.7	0.13 (0.6)	0.74 (1.0)	163	2.9
1000	293831	0.6	0.21 (0.9)	0.33 (0.8)	203	1.3

Table 1 shows the correlation between GEBV and TBV for the validation data. emBayesB_CV was significantly more accurate than GBLUP. The poor performance of GBLUP was due to the fact that unrelated individuals were chosen in the original dataset (Yang *et al*. 2010). Livestock populations have high levels of relatedness and so GBLUP would be expected to do much better in livestock populations. emBayesB_CV was most accurate when there were 10 simulated QTL ($r_{TBV} = 0.88$) with the accuracy decreasing to 0.74 and 0.33 for 100 and 1000 QTL respectively (Table 1). This decline in accuracy is not unexpected. As the number of QTL increase, the size of individual QTL effects decrease and thus it becomes more difficult to detect the location of the QTL as shown in Figure 1. Figure 1A shows chromosome 6 which has the two largest QTL effects in the 10 QTL dataset. It can be seen that the location of the 2 QTL is accurately determined by SNP with large posterior probabilities. However Figure 1B shows that the 69 small QTL effects on chromosome 4 in the 1000 QTL dataset, are not accurately located by SNP with large posterior probabilities. In fact there are only 4 SNP with posterior probabilities greater than 0.9 on chromosome 4 and none are located close to QTL.

Although emBayesB_CV was not able to accurately locate QTL with small effects, it was able to predict aggregate breeding value more accurately than GBLUP by detecting SNP in regions of trait variation. For example, in the 1000 QTL dataset, the chromosome explaining the largest percentage of the phenotypic variance (6.7%) was chromosome 4 which contained 69 QTL, and 13 detected SNP with posterior probabilities greater than 0.05 of being in LD with QTL (Figure 1B). On the other hand the chromosome explaining the smallest percentage of the phenotypic variance (0.9%) in the 1000 QTL dataset, was chromosome 21 which contained 13 QTL, and only 4 detected SNP with posterior probabilities greater than 0.05 of being in LD with QTL.

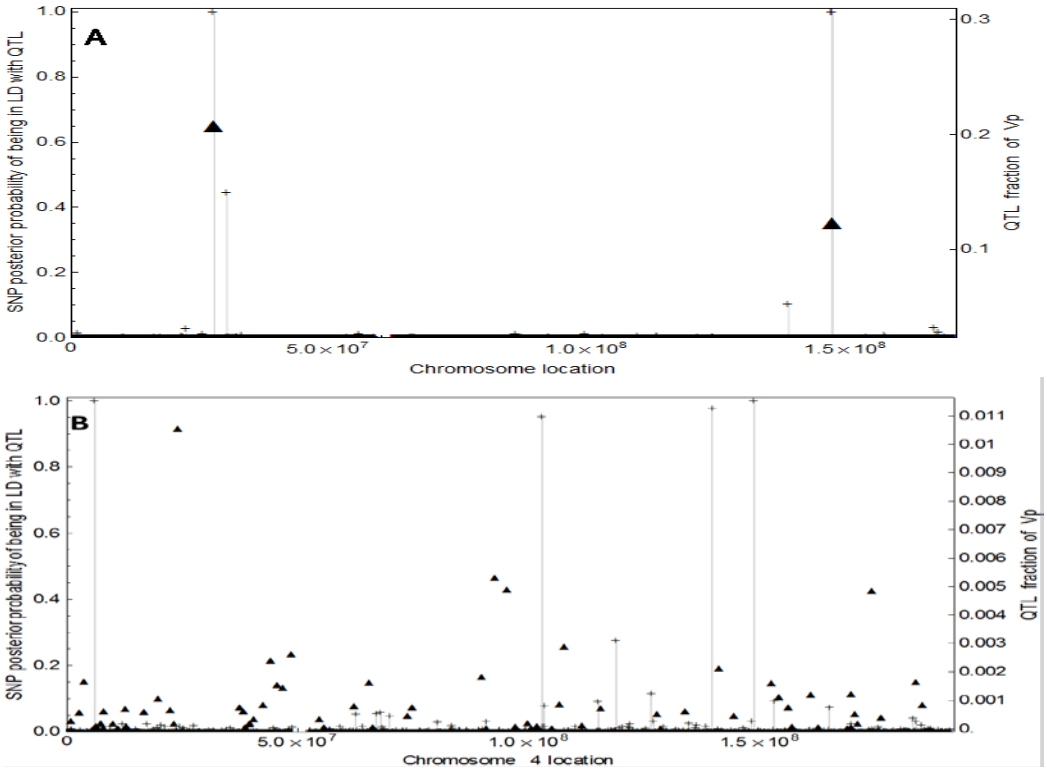

Figure 1. Posterior probability (+) of a SNP being in LD with QTL and the fraction of the total phenotypic variance (▲) explained by a QTL on chromosome 6 for the 10 QTL dataset (A) and on chromosome 4 for the 1000 QTL dataset (B).

The number of SNP with posterior probabilities greater than 0.5 of being in LD with QTL, was 51, 76 and 49 for the 10, 100 and 1000 QTL datasets. Also using the formula:

$No.\ SNP\ in\ LD = m\hat{\gamma}$ where $\hat{\gamma}$ estimates the proportion of SNP in LD with QTL,

Table 1 shows that emBayesB_CV predicts 93, 163 and 203 SNP are in LD with QTL for the 10, 100 and 1000 QTL datasets. This shows the difficulty emBayesB_CV has in detecting QTL as the effects get smaller. Increasing the number of genotyped individuals will help detect smaller QTL effects.

CONCLUSIONS

emBayesB_CV is a computationally fast method of predicting breeding value using dense genome-wide SNP marker data which was significantly more accurate than GBLUP for the scenarios investigated in this paper. emBayesB_CV overcame the convergence issue which often occurs with emBayesB. The chromosomal location of large QTL effects can be accurately located with emBayesB_CV. But this is not the case for QTL of small effect.

REFERENCES

Shepherd R. K., Meuwissen T.H.E. and Woolliams J.A. (2010) *BMC Bioinformatics* **11**: 529
Yang J., Benyamin B., McEvoy B. P., Gordon S., *et al.* (2010) *Nature Genetics* **42**: 565