A Fast EM Algorithm For Genomic Selection

R.K. Shepherd[*], *T.H.E. Meuwissen*[†] and J.A. Woolliams[§]

Introduction

Genomic selection is being adopted in many livestock breeding programs. Some industry applications use BLUP methods (called GS-BLUP) which are computationally fast but assume each marker effect is normally distributed with the same variance. The accuracy of prediction can often be increased by using models which not only allow marker variance to vary but also allow a large proportion of markers to have no effect. Meuwissen et al. (2001) called these methods BayesA and BayesB respectively. However implementing these Bayesian methods is computationally slow, particularly for large SNP panels. This paper gives details of an Expectation Maximisation (EM) algorithm called emBayesB for implementing a BayesB-like model which is both fast and accurate.

Material and methods

Theory. If we knew precisely which SNP were in linkage disequilibrium (LD) with QTL, then the problem would be much easier. This missing information is crucial in formulating an EM algorithm. An unknown variable z_j is defined which indicates if the *j*th SNP is in LD with QTL ($z_j = 1$) or not ($z_j = 0$). If $z_j = 1$, the SNP effect g_j is assumed to be from a double exponential (DE) distribution with parameter λ; while if $z_j = 0$, the SNP effect is assumed to be distributed as a Dirac Delta (DD) function which has all its probability mass at 0. We assume *a priori* that a fraction γ of the SNPs are in LD with QTL. Hence the prior distribution for g_j can be written as $p(g_j) = \gamma \left[0.5\lambda \exp(-\lambda \mid g_j \mid) \right] + (1-\gamma)\, \delta(g_j)$ ie. a mixture of a DE and the Dirac Delta function $\delta(g_j)$. A linear data model $\mathbf{y} = \mathbf{Bg} + \mathbf{e}$ is assumed to relate record y_i of individual *i* to the *j*th SNP effect g_j where element b_{ij} of 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 we are able to develop an iterative sequence of E and M-steps which converge to maximum *a posteriori* (MAP) parameter estimates. At iteration *k* the E-step involves calculating $\gamma_j^k \left(= E\left[z_j \mid \mathbf{y}\ \&\ \textit{all current estimates}\right]\right)$, the posterior probability that SNP *j* is in LD with QTL given the data and all current parameter estimates. This is done

[*] Computing Sciences, CQUniversity, Rockhampton 4702, Australia
[†] IHA, Norwegian University of Life Sciences, Box 5003, N1432 As, Norway
[§] The Roslin Institute & R(D)SVS, University of Edinburgh, Roslin, Midlothian EH25 9PS, UK

analytically and fast. Given the data and the current values of γ_j^k, the M-step calculates $\hat{g}_j = \gamma_j^k DE_{\text{mode}}$, $\hat{\gamma} = \frac{1}{m}\mathbf{1}'\boldsymbol{\gamma}^{\mathbf{k}}$, $\hat{\lambda} = \mathbf{1}'\boldsymbol{\gamma}^{\mathbf{k}} \Big/ |\hat{\mathbf{g}}|' \boldsymbol{\gamma}^{\mathbf{k}}$ and $\hat{\sigma}_e^2 = \frac{1}{n}(\mathbf{y} - \mathbf{B}\hat{\mathbf{g}})'(\mathbf{y} - \mathbf{B}\hat{\mathbf{g}})$ where $\boldsymbol{\gamma}^{\mathbf{k}}$ is the vector of posterior probabilities at iteration k and DE_{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 converges quickly to produce MAP estimates of the g_j, ML estimates of $\gamma, \lambda, \sigma_e^2$ and posterior probabilities γ_j^k. More details are in Shepherd et al. (submitted).

Data Simulation. The QTLMAS XII common dataset (Lund et al. (2009)) was analysed. An initial population of 100 founders (50 of each sex) was simulated. For each of the next 50 generations, 100 progeny (50 male and 50 female) were produced by randomly sampling parents. Then for the next and last 6 generations, 15 males and 150 females were randomly selected for a hierarchical mating to produce 100 progeny per male and 10 progeny per female, giving a total of 1500 pedigreed progeny per generation. The validation data consisted of 1200 individuals with only genotype records and was a random selection of 400 progeny per generation from each of the last 3 generations. The training data was the genotype and phenotype records of the 4665 individuals from the preceding 4 generations. There were 6000 biallelic markers at 0.1 cM spacing on the six 100cM chromosomes, giving 1000 markers per chromosome. The genomic location and allele substitution effects of the 48 simulated biallelic and additive QTL are shown in Figure 1. The number of QTL, which explain more than 0.1, 1, 5 and 10% of the total genetic variation in the training data, was 28, 15, 6 and 4 respectively. An individual's true breeding value (TBV) was the sum of the effects of all of the individual's QTL. A trait with heritability of 0.3 was produced by adding a normally distributed error term to the TBV of each individual.

Statistical Analysis. The equation $\mathbf{GEBV} = \mathbf{B}\hat{\mathbf{g}}$ was determined using the phenotypes and SNP genotypes of the 4665 individuals in the training data set. The number of SNP analysed was 5726 as only SNP with a minor allele frequency greater than 0.05 were used. The initial parameter estimates for emBayesB were $g_j = 0, \gamma = 0.01$ plus λ and σ_e^2 for a total phenotypic variance of 4.4 and heritabilities of 0.3 or 0.5. The prediction equation was used to calculate the GEBV of the 1200 individuals in the validation data set using only the genotype of their 5726 SNP. The correlation between TBV and GEBV was calculated for each of the 3 generations (400 individuals) of validation data as well as the linear regression of TBV on GEBV, which has a slope of 1 if the GEBV are unbiased. GEBV were also calculated using least squares (LS), GS-BLUP and the ICE algorithm of Meuwissen et al. (2009) for the same 5726 SNP. The SNP effects for LS and GS-BLUP were calculated by solving $(\mathbf{B}'\mathbf{B} + \alpha\mathbf{I})\hat{\mathbf{g}} = \mathbf{B}'\mathbf{y}$ where $\alpha = \left(1 - h^2\right)\Big/h^2$ for GS-BLUP and $\alpha = 0$ for LS.

Results and discussion

emBayesB was the most accurate method of prediction using all 1200 validation records (Table 1). The ICE algorithm produced a similar correlation in the generation following

Table 1: Correlation and regression coefficient (in brackets) of TBV on GEBV for each generation, and for all 3 generations, of the validation data for 4 methods of prediction.

	Least	Initial h^2 estimate = 0.3			Initial h^2 estimate = 0.5		
Gen	**Squares**	**GS-BLUP**	**ICE**	**emBayesB**	**GS-BLUP**	**ICE**	**emBayesB**
1	0.41 (0.20)	0.77 (0.85)	0.88 (0.88)	0.88 (1.00)	0.74 (0.71)	0.84 (0.81)	0.86 (0.98)
2	0.36 (0.15)	0.77 (0.89)	0.86 (0.88)	0.89 (1.03)	0.73 (0.71)	0.81 (0.81)	0.86 (1.01)
3	0.30 (0.11)	0.72 (0.78)	0.81 (0.78)	0.86 (0.92)	0.68 (0.62)	0.77 (0.71)	0.84 (0.90)
All	0.34 (0.15)	0.75 (0.85)	0.85 (0.86)	0.87 (1.00)	0.71 (0.69)	0.81 (0.79)	0.85 (0.98)

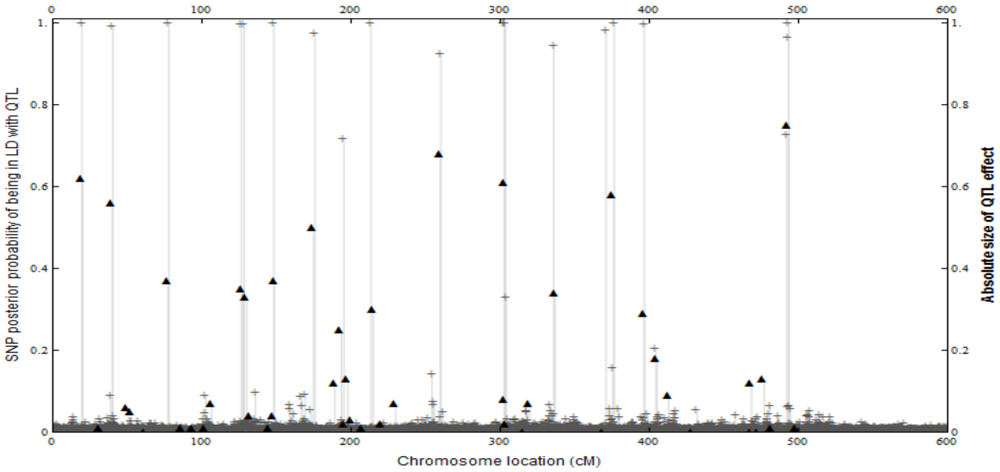

Figure 1. Absolute value of each QTL effect (▲) for the 48 simulated QTL plus the emBayesB posterior probability (+) of each SNP being in LD with at least one QTL

training but, like GS-BLUP, the accuracy decreased considerably by the 3rd generation after training. The accuracy of emBayesB decreased the least over generations. Using emBayesB produced unbiased estimates of TBV using all 1200 validation records but mainly in the first two generations after training (Table 1). All other methods produced GEBV which overestimated TBV in all generations. Lund et al. (2009) reported correlations of 0.84 to 0.87 and unbiased estimates of TBV for Bayesian MCMC methods applied to this data set.

Most SNP had small posterior probabilities of being in LD with QTL (Figure 1). Only 24 of the 5726 SNP had posterior probabilities greater than 0.1. emBayesB detected all 15 QTL with allele substitution effects greater than 0.2 by calculating posterior probabilities of 0.72 or more for nearby SNP (Figure 1). There were 15 QTL which each explained more than 1% of the total additive genetic variation (V_A) and, in total, over 95% of the total V_A. emBayesB detected each of these 15 QTL (Figure 2). The distance from each of the detected QTL to the nearest high probability SNP averaged 0.7cM, with the largest distance being 1.7cM. It appears that the reason why the accuracy of emBayesB decreases least with generations post training is that emBayesB models QTL location accurately, and thus the segregation (in later generations) is also modeled better than the segregation with GS-BLUP or ICE.

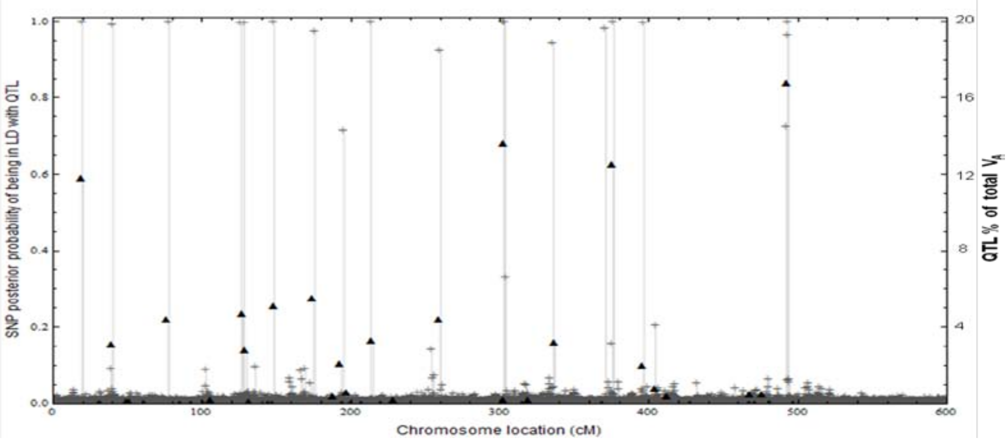

Figure 2. Percentage of total V_A (▲) explained by each of the 48 QTL plus the emBayesB posterior probability (+) of each SNP being in LD with at least one QTL

Each emBayesB analysis took only a few minutes to compute as did all the methods in Table 1 as each method uses Gauss-Seidel iteration, often in combination with fast computation of analytical formulas. With emBayesB most of the computation time is spent constructing the $\mathbf{B}'\mathbf{B}$ matrix which will be even larger as the size of SNP panels increase. However Gauss-Seidel iteration with residual update (Leggarra and Misztal (2008)) can be used to avoid the calculation of this matrix as shown by Shepherd et al. (submitted). Analysing this data using a Bayesian MCMC method took approximately 2 days (R. Pong-Wong, pers. comm.). The speed advantage of emBayesB is because no sampling of posterior distributions is required. However emBayesB will require further computation to obtain standard errors.

Conclusion

emBayesB is a fast, accurate and unbiased EM algorithm for implementing genomic selection by mapping QTL in genome-wide dense SNP marker data. Its accuracy is similar to Bayesian MCMC methods but it takes only a fraction of the time. The decline in accuracy over the generations following training is less for emBayesB than for GS-BLUP as emBayesB estimates QTL location accurately, and so is better able to model QTL segregation in later generations using marker LD. The current dataset contained QTL of large effect and it will be interesting to see if similar conclusions hold for datasets without QTL of large effect.

References

Legarra, A. And Misztal, I. (2008). *J.Dairy Sci.*, 91:360-366.

Lund, M., Sahana, G., de Koning, D-J., *et al.* (2009). *BMC Proc.*, 3(Suppl 1):S1.

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

Meuwissen, T., Solberg, T., Shepherd, R., and Woolliams, J. (2009). *Gen. Sel. Evol.*, 41**:**2.

Shepherd, R., Meuwissen, T. and Woolliams, J. (submitted). *BMC Bioinformatics*