GENOMIC SELECTION USING A FAST EM ALGORITHM 2. ANALYSIS OF SIMULATED DATA

R.K. Shepherd[1], T.H.E. Meuwissen[2] and J.A. Woolliams[3]

[1]FABIE, CQUniversity, Rockhampton, QLD 4702
[2]IAAS, Norwegian University of Life Sciences, Box 5003, N1432 As. NORWAY
[3]The Roslin Institute & R(D)SVS, University of Edinburgh, Roslin, Midlothian, EH25 9PS UK

SUMMARY

The paper reports on a fast EM algorithm for genomic selection by mapping QTL in genome-wide dense SNP marker data. The algorithm called emBayesB was used to analyse a 6000 SNP dataset simulated for the QTLMAS XII workshop. True breeding value was accurately predicted by GEBV with a correlation of 0.85 in the validation data, while the regression coefficient of 0.97 indicated unbiased predictions of breeding value. The results were similar to Bayesian MCMC estimates but were calculated in a fraction of the time. emBayesB was also able to accurately map the location of individual QTL which explain more than 1% of the total genetic variation.

INTRODUCTION

Genomic selection is a recent tool for genetic improvement in animal breeding. It involves the use of dense DNA markers covering the whole genome so that all QTL are in linkage disequilibrium (LD) with at least one marker. It has recently become economically feasible due to the commercial availability of dense genotyping chips, containing thousands to millions of single nucleotide polymorphisms (SNP), and the development of high throughput systems, all at cost effective prices. Genomic selection involves two main steps: first the estimation of SNP effects using phenotype and genotype data in a reference population (training data), and then the prediction of genomic breeding values (GEBV) using only marker genotypes (and the previously estimated SNP effects) in the population under selection (validation data).

Mixed model (or GS-BLUP) methods and Bayesian MCMC (Markov Chain Monte Carlo) methods are the main contenders for calculating GEBV. Bayesian MCMC methods generally have a higher accuracy of predicting GEBV than GS-BLUP methods but are slow computationally (Lund *et al.* 2009). The prior information used in a Bayesian approach can be incorporated in an Expectation Maximisation (EM) algorithm through the calculation of a posterior mode. Also EM algorithms can be significantly faster than Bayesian MCMC methods if well formulated. This paper reports on a fast EM algorithm for genomic selection by mapping QTL in genome-wide dense SNP marker data. The algorithm is called emBayesB because it is an EM implementation of the important features of the BayesB method of Meuwissen *et al.* (2001).

MATERIALS AND METHODS

Data simulation. The data analysed was the QTLMAS XII common dataset. Full details are available in Lund *et al.* (2009). Initially a population of 100 founders (50 of each sex) was simulated. For 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 of 1200 individuals with only genotype records was produced using the last 3 generations by randomly selecting 400 progeny per generation. The training data of 4665 individuals with genotype and phenotype records consisted of the progeny 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 two alleles at each marker were sampled with equal probability in the founders. QTL effects were sampled from a gamma distribution. The genomic location and allele substitution effects of the 48 simulated biallelic and additive QTL are shown in Figure 2. 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.

EM algorithm. The methodology used in the EM algorithm emBayesB is presented in the companion paper (Shepherd and Woolliams 2009) and so is not reported here.

Validating emBayesB. The algorithm emBayesB used phenotypes and SNP genotypes of the 4665 individuals in the training data set to calculate the prediction equation $\mathbf{GEBV = B\hat{g}}$ by Gauss-Seidel iteration. 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 assumed for the EM algorithm were $\hat{g}_j = 0$, $\hat{\gamma} = 0.01$, $\hat{\lambda} = 7.8$ and $\hat{\sigma}_e^2 = 1.4$, some of which result from an assumed heritability of 0.5 and a total phenotypic variance of 2.8. 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 the validation data as well as the linear regression of TBV on GEBV, which has a slope of 1 if the GEBV are unbiased.

RESULTS AND DISCUSSION

The main advantage of emBayesB over Bayesian MCMC methods is the speed of computation. The Bayesian methods are computationally slow as they use intense MCMC techniques (eg. Gibbs sampling) to obtain, after considerable computation, random samples from each SNP's posterior distribution for the thousands of unknown SNP effects. The emBayesB algorithm is fast as it uses Gauss-Seidel iteration to compute a weighted analytical mean of the modes in each SNP's posterior distribution as the best estimate of a SNP effect (Shepherd and Woolliams 2009). Using a Bayesian MCMC approach for this data took 2 days on a Unix box (R. Pong-Wong, pers. comm.) whereas emBayesB converged in a few minutes on a laptop PC.

Using emBayesB produced a correlation of 0.85 between TBV and GEBV for the validation data. This correlation is considerably larger than the range of correlations of 0.5 to 0.77 for the GS-BLUP models used on the same data, but similar to correlations of 0.84 to 0.87 for Bayesian MCMC methods, as reported by Lund *et al.* (2009). One Bayesian method used a 5 SNP haplotype to produce a correlation of 0.92. But the correlation for this method would usually be smaller as haplotypes are estimated with error in real life, not known exactly as in this data. Using emBayesB produced a slope of 0.97 for the regression of TBV on GEBV in the validation data. This slope indicates unbiased predictions of breeding value which was also found for the Bayesian MCMC methods but in general not for the GS-BLUP methods. The Bayesian MCMC methods assume a SNP mixture distribution in which relatively few markers explain a large variance, while a large number of markers explain a very small variance. Due to the similar information in the priors of emBayesB and the Bayesian MCMC methods, it is no surprise that they produce similar predictions as measured by the accuracy and the regression coefficient of TBV on GEBV.

As expected most SNP (5702 in total) have small posterior probabilities of being in LD with QTL (Figure 1). A surprising result was that only 24 of the 5726 SNP had posterior probabilities greater than 0.1. On chromosome 6 all SNP had posterior probabilities less than 0.06 which is due to the absence of QTL on this chromosome. emBayesB detects all 15 QTL with allele substitution

effects greater than 0.2 by calculating posterior probabilities of 0.72 or more for nearby SNP (Figure 2).

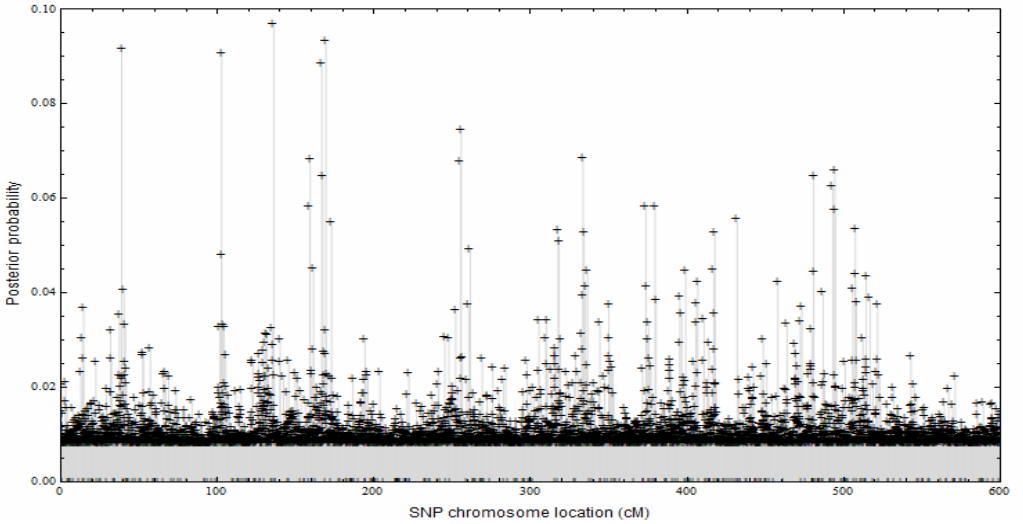

Figure 1. Posterior probability (+) for 5702 SNP of each being in LD with at least one QTL.

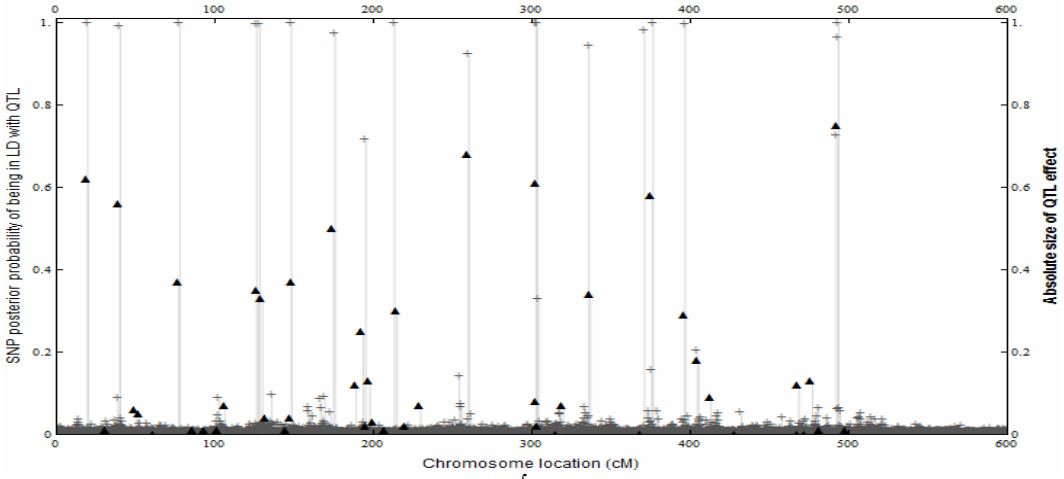

Figure 2. Absolute value of the QTL effect (∫) for the 48 simulated QTL plus the posterior probability (+) for 5726 SNP of each being in LD with at least one QTL.

There are 15 QTL which each explain more than 1% of the total additive genetic variation (V_A) and in total, explain over 95% of the total V_A. emBayesB detected each of these 15 QTL (Figure 3). The distance from each of the detected QTL to the nearest high probability SNP averaged 0.7cM, with the largest distance being 1.7cM. As the genetic variation explained by a QTL dropped below 1% so did the posterior probability of a SNP being in LD with it (Figure 3). It was found that the EM estimates of the high probability SNP effects were not very accurate for estimating the effects of individual QTL. However using only SNP with a large posterior probability as fixed effects in a multiple linear regression resulted in QTL estimates which were

quite accurate. Basically it seems that the posterior probabilities can be used to screen out most of the SNP, leaving only those in LD with QTL, for a multiple regression approach if estimates of large (> 1% V_A) QTL effects are required.

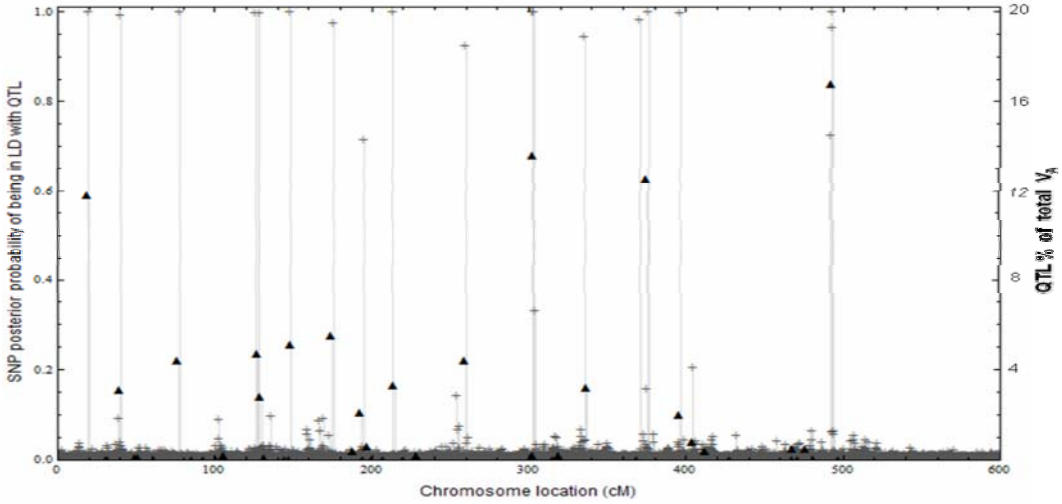

Figure 3. Percentage of total V_A (ĵ) explained by each of the 48 QTL plus the posterior probability (+) for 5726 SNP of each being in LD with at least one QTL.

In general emBayesB found only one nearby SNP with a high posterior probability for each of the 15 QTL which individually explain more than 1% of the total V_A (Figure 3). However multiple high probability SNP were reported nearby for QTL at 303cM, 376cM and 493cM. The reason for these multiple nearby high probability SNP seems to be related to the fact that these QTL explain more V_A than any other QTL and that this larger QTL variation induces higher correlation among neighbouring SNP through LD. Further investigation is needed to determine if this result is a feature of emBayesB.

As the size of SNP panels increase, the Bayesian MCMC methods will become even slower. Further research is required to see if emBayesB will be a suitable algorithm for analysing data from large SNP panels as convergence is not guaranteed with the Gauss-Seidel iteration.

CONCLUSIONS

emBayesB is a fast and accurate EM algorithm for implementing genomic selection by mapping QTL in genome-wide dense SNP marker data. Its accuracy is comparable to Bayesian MCMC methods but it takes only a fraction of the time. The current study analysed only one replicate of a common dataset so there is a need to analyse more data to understand its capabilities.

ACKNOWLEDGMENTS

This paper reports collaborative research instigated while RKS was on sabbatical at the Roslin Institute with support from CQUniversity and the Roslin Institute.

REFERENCES

Lund, M.S., Sahana, G., de Koning, D-J., Su G. and Carlborg, O. (2009) *BMC Proc.* **3**(Suppl 1):S1
Meuwissen, T.H.E., Hayes, B.J. and Goddard, M.E. (2001) *Genetics* **157:**1819.
Meuwissen, T.H.E., Solberg, T.R., Shepherd, R., and Woolliams, J.A. (2009) *Gen. Sel. Evol.* **41:**2.
Shepherd R.K. and Woolliams J.A. (2009) *Proc. Assoc. Advmt. Anim. Breed. Genet.* **18**: 80.