Heredity (2010) 105, 483–494
© 2010 Macmillan Publishers Limited All rights reserved 0018-067X/10 $32.00
www.nature.com/hdy

ORIGINAL ARTICLE

An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects

S Xu
Department of Botany and Plant Sciences, University of California, Riverside, CA, USA

The least absolute shrinkage and selection operator (Lasso) estimation of regression coefficients can be expressed as Bayesian posterior mode estimation of the regression coefficients under various hierarchical modeling schemes. A Bayesian hierarchical model requires hyper prior distributions. The regression coefficients are parameters of interest. The normal distribution assigned to each regression coefficient is a prior distribution. The variance parameter in the normal prior distribution is further assigned a hyper prior distribution so that the variance parameter can be estimated from the data. We developed an expectation–maximization (EM) algorithm to estimate the variance parameter of the prior distribution for each regression coefficient. Performance of the EM algorithm was evaluated through simulation study and real data analysis. We found that the Jeffreys' hyper prior for the variance component usually performs well with regard to generating the desired sparseness of the regression model. The EM algorithm can handle not only the usual regression models but it also conveniently deals with linear models in which predictors are defined as classification variables. In the context of quantitative trait loci (QTL) mapping, this new EM algorithm can estimate both genotypic values and QTL effects expressed as linear contrasts of the genotypic values.

Heredity (2010) **105,** 483–494; doi:10.1038/hdy.2009.180; published online 6 January 2010

Keywords: empirical Bayes; Lasso; mixed model; QTL mapping; shrinkage estimation

Introduction

Mapping quantitative trait loci (QTLs) has long been treated as a variable selection problem (Broman and Speed, 2002; Manichaikul *et al*., 2009) because the number of markers (predictors) can be larger than the sample size, making ordinary least square method infeasible. Ridge regression (Hoerl and Kennard, 1970) is one of the solutions to handle relatively large regression models and has been applied to QTL mapping (Whittaker *et al*., 2000). However, the results of the usual ridge regression are not satisfactory because all regression coefficients are shrunken by the same shrinkage factor. Xu (2003) developed a Bayesian shrinkage method to estimate QTL effects, in which different regression coefficients are shrunken using different shrinkage factors. This kind of selective shrinkage analysis discriminates against small regression coefficients and favors for large regression coefficients. As a result, it performs far better than the classical ridge regression. The original Bayesian shrinkage analysis of Xu (2003) was implemented through the Markov chain Monte Carlo sampling algorithm, which is time consuming for large models coupled with large sample sizes. Xu (2007) recently proposed an empirical Bayesian method to improve the computational efficiency, while still preserving the desired sparseness of the final model.

In the empirical Bayesian method of Xu (2007), estimation of variance components is achieved by repeated callings of the Nelder and Mead (1965) simplex algorithm. This method only applies to numerically coded predictors. In many situations, in which the predictors are discrete classification variables, the special algorithm of Xu (2007) that only applies to numerically coded predictors cannot be used. For example, in QTL mapping of F_2 populations that are derived from the cross of two inbred lines, there are three possible genotypes at each locus. We have to code the three genotypes numerically as 1, 0 and -1, to capture the additive effect (a), and as 0, 1 and 0, to capture the dominance (d) effect. With other mapping populations, for example, four-way cross (Xu, 1998), the numerical coding is more complicated. In association mapping, in which the number of genotypes may vary from one locus to another, an optimal numerical coding system may not even exist. Therefore, a method that can handle classification predictor variables is more general than the simplex algorithm adopted by Xu (2007). With the general method, we can directly estimate the genotypic values and their variances, and then convert the genotypic values into additive effect, dominance effect and whatever effect of interest. Whereas the simplex algorithm in the empirical Bayesian method of Xu (2007) cannot handle classification predictor variables, the expectation–maximization (EM) algorithm (Dempster *et al*., 1977) can do it in a straightforward manner. Therefore, we propose an EM algorithm to estimate the variance components under this general setting of the predictors.

It is well known that the least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996) estimation of regression coefficients has a Bayesian interpretation. When the variance parameter in the normal prior of each regression coefficient is assigned an exponential prior, the Bayesian posterior mode estimate of the regression coefficient is the Lasso estimate (Tibshirani,

Correspondence: Dr S Xu, Department of Botany and Plant Sciences, University of California, 900 University Avenue, Riverside, CA 92521, USA. E-mail: xu@genetics.ucr.edu

Received 12 August 2009; revised 26 October 2009; accepted 12 November 2009; published online 6 January 2010

1996; Park and Casella, 2008; Yi and Xu, 2008). With the simplex algorithm adopted by Xu (2007), extension to the Lasso estimate is not obvious. But such an extension is straightforward when an EM algorithm is applied. Similar EM algorithm has been proposed by Figueiredo (2003) and Yi and Banerjee (2009), who treated the variance components as missing values. In the proposed EM algorithm, we will treat the regression coefficients as missing values when we estimate the variance parameters. This makes the estimates of regression coefficients empirical Bayesian estimates. As a result, theory and method of classical mixed-effect model apply to the empirical Bayesian estimation of QTL effects.

Theory and methods

Model

Let y be an $n \times 1$ vector for the phenotypic values of a quantitative trait, where n is the number of individuals in the mapping population. The linear model for y is

$$y = \sum_{j=1}^{q} X_j \beta_j + \sum_{k=1}^{p} Z_k \gamma_k + \varepsilon. \tag{1}$$

where β_j is the jth non-QTL effect (for example, the year effect), X_j is the corresponding design matrix, γ_k is a vector of genotypic values for locus k and Z_k is the corresponding incidence matrix determined by the genotypes of locus k. The dimensions of γ_k and Z_k depend on the number of genotypes for locus k. The residual error vector ε is assumed to be distributed as $\varepsilon \sim N(0, \sigma^2 I_n)$, where I_n is an $n \times n$ identity matrix and σ^2 is an unknown residual error variance. We are interested in estimating all the nuisance parameters (β), the genotypic values for all QTLs (γ) and the prior variances of all QTL effects simultaneously from the same model. If we evaluate markers of the entire genome, p can be very large and sometimes may be even larger than the sample size, although q can be relatively small. In this case, we need to adopt a shrinkage method to estimate γ, which are the most important parameters in QTL analysis.

Prior distribution

Let m_k be the number of genotypes at locus k. For example, in a F_2 population, each locus has three possible genotypes, and thus, $m_k = 3$ for all $k = 1,\ldots, p$. The dimension of Z_k is $n \times m_k$ and the dimension of γ_k is $m_k \times 1$. We adopt the normal prior for γ_k, for example,

$$p(\gamma_k|\sigma_k^2) = N(\gamma_k|0, \sigma_k^2 I_{m_k}) \tag{2}$$

Under this prior, model (1) becomes a typical mixed model so that y has a multivariate normal distribution with mean μ and variance–covariance matrix V, where

$$\mu = \sum_{j=1}^{q} X_j \beta_j \tag{3}$$

and

$$V = \sum_{k=1}^{p} Z_k Z_k^T \sigma_k^2 + I\sigma^2 \tag{4}$$

Following Yi and Xu (2008), we consider two classes of prior for σ_k^2. The first class is the scaled inverse χ^2 prior, whose density is

$$p(\sigma_k^2|\tau,\omega) = \mathrm{Inv} - \chi^2(\sigma_k^2|\tau,\omega)$$
$$\propto (\sigma_k^2)^{-\frac{1}{2}(\tau+2)} \exp\left(-\frac{\omega}{2\sigma_k^2}\right) \tag{5}$$

In the scaled inverse χ^2 distribution, τ and ω are hyperparameters representing the degree of prior belief and the scale. Two special cases of the scaled inverse χ^2 distribution are particularly interesting, because they represent priors commonly used in data analysis. One special case is $\xi = (\tau, \omega) = (-2, 0)$, which is equivalent to the uniform prior $P(\sigma_k^2) \propto 1$. This uniform prior leads to the usual maximum likelihood estimate of the variance component. The other special case is $\xi = (\tau, \omega) = (0, 0)$, which represents the Jeffreys' prior (Figueiredo, 2003), that is, $P(\sigma_k^2) = 1/\sigma_k^2$. This prior does not have hyperparameters at all, and thus, is extremely convenient to use in real data analysis (Figueiredo, 2003).

The second class of prior is the exponential prior,

$$p(\sigma_k^2|\lambda) = \mathrm{Expon}\left(\sigma_k^2 \middle| \frac{\lambda^2}{2}\right) = \frac{\lambda^2}{2} \exp\left(-\frac{\lambda^2}{2}\sigma_k^2\right) \tag{6}$$

where λ^2 is the shrinkage factor (hyperparameter). This exponential prior will generate the Lasso estimation (Tibshirani, 1996; Park and Casella, 2008; Yi and Xu, 2008) of the QTL effects.

Posterior mode

Our EM algorithm treats γ as the missing value. This is different from the EM algorithm of Figueiredo (2003) and Yi and Banerjee (2009), who treated σ_k^2 as the missing value. The EM steps will be given after we describe the formulas for the maximization steps and the expectation steps. The target function for maximization in our EM algorithm is the expected complete-data log likelihood function, in which the regression coefficients are treated as missing values. For the scaled inverse χ^2 prior, the part of the expected complete-data log likelihood function relevant to σ_k^2 is

$$L(\sigma_k^2|\tau,\omega) = -\frac{\tau + 2 + m_k}{2} \ln(\sigma_k^2) - \frac{1}{2\sigma_k^2}[E(\gamma_k^T \gamma_k) + \omega] \tag{7}$$

where $\mathrm{E}(\gamma_k^T\gamma_k) = \mathrm{E}(\gamma_k^T\gamma_k \mid \theta, y)$ is a short notation for the conditional expectation of the quadratic term of γ_k, given the current values of parameters (θ) and the data (y). Setting $\frac{\partial}{\partial \sigma_k^2} L(\sigma_k^2|\tau,\omega) = 0$ and solving for σ_k^2, we obtain

$$\sigma_k^2 = \frac{E(\gamma_k^T \gamma_k) + \omega}{\tau + 2 + m_k} \tag{8}$$

When $\xi = (\tau, \omega) = (-2, 0)$, we have $\sigma_k^2 = E(\gamma_k^T\gamma_k)/m_k$, equivalent to the solution when a uniform prior is used (typical mixed model solution for a variance component). When $\xi = (\tau, \omega) = (0, 0)$, we get $\sigma_k^2 = E(\gamma_k^T\gamma_k)/(2 + m_k)$, a stronger shrinkage than the uniform prior.

For the exponential (Lasso) prior, the part of the expected complete-data log likelihood function relevant to σ_k^2 is

$$L(\sigma_k^2|\lambda) = -\frac{m_k}{2} \ln(\sigma_k^2) - \frac{E(\gamma_k^T \gamma_k)}{2\sigma_k^2} - \frac{1}{2}\lambda^2 \sigma_k^2 \tag{9}$$

Setting $\frac{\partial}{\partial\sigma_k^2} L(\sigma_k^2 | \lambda^2) = 0$ and solving for σ_k^2 leads to two solutions, with the positive one being

$$\sigma_k^2 = \frac{\sqrt{m_k^2 + 4\lambda^2 E(\gamma_k^{\mathrm{T}}\gamma_k)} - m_k}{2\lambda^2} \tag{10}$$

Formulas for the fixed effects and residual variances follow the standard procedure of mixed model methodology (Lindstrom and Bates, 1988). For the fixed effects, we have

$$\beta = (X^{\mathrm{T}}V^{-1}X)^{-1}(X^{\mathrm{T}}V^{-1}y) \tag{11}$$

For the residual error variance, we use

$$\sigma^2 = \frac{1}{n}(y - X\beta)^{\mathrm{T}}(y - X\beta - \sum_{k=1}^{p} Z_k E(\gamma_k)) \tag{12}$$

where $E(\gamma_k) = E(\gamma_k | \theta, y)$ is a short notation for the conditional expectation of γ_k. Finding the posterior modes of the parameters belongs to the maximization steps. We have noticed that these maximization steps depend on $E(\gamma_k)$ and $E(\gamma_k^{\mathrm{T}}\gamma_k)$, which are the conditional expectations of the linear and quadratic terms of the missing value.

Best linear unbiased prediction

The expectation of the quadratic term required in the maximization steps is expressed as

$$E(\gamma_k^{\mathrm{T}}\gamma_k) = E(\gamma_k^{\mathrm{T}})E(\gamma_k) + \mathrm{tr}[\mathrm{var}(\gamma_k)] \tag{13}$$

where

$$E(\gamma_k) = \sigma_k^2 Z_k^{\mathrm{T}} V^{-1}(y - X\beta) \tag{14}$$

is the conditional expectation and

$$\mathrm{var}(\gamma_k) = I\sigma_k^2 - \sigma_k^2 Z_k^{\mathrm{T}} V^{-1} Z_k \sigma_k^2 \tag{15}$$

is the conditional variance of the missing vector γ_k. Derivation of equations (14) and (15) are given in Appendix A. Both the expectation and the variance depend on the parameters and thus iterations are needed. Once the iterations converge, the conditional expectation $\mathrm{E}(\gamma_k)$ is called the best linear unbiased prediction (BLUP) and the square root of the variance $\mathrm{var}(\gamma_k)$ is called the prediction error of γ_k. However, BLUP is defined on the basis of true parameters. The conditional expectation of γ_k after the iterations converge is conditional on estimated parameters. Technically, the conditional expectation given in equation (14) is not called as BLUP, but is called as empirical Bayesian estimate. Therefore, we will call the BLUP of QTL effects as the estimated QTL effects subsequently, although they are predicted QTL effects under the mixed model framework.

EM steps

Now let us define $\theta = \{\beta, \sigma^2, \sigma_1^2, \ldots, \sigma_p^2\}$ as the parameter vector and $\xi = \{\tau, \omega\}$ or λ^2 as the hyperparameters. The genotypic values γ are treated as missing values. The EM steps are described below.

Step (0) Choose ξ or λ^2, set $t = 0$ and initialize parameters with $\theta = \theta^{(t)}$.

Step (1) Calculate $E(\gamma_k^T\gamma_k)$ using equations (13–15), which is the E-step.

Step (2) Update θ using equations (8, 10–12), which is the M-step.

Step (3) Let $t = t + 1$, and repeat Steps (1) and (2) until convergence is reached.

Linear contrasts

The EM algorithm is described with γ being defined as the genotypic values that are not equivalent to QTL effects. The QTL effects can be defined as linear contrasts or linear combinations of the genotypic values. There are two ways to obtain the QTL effects, one of which is to recode matrix Z so that γ directly represent the QTL effects. For example, if Z_k for the jth individual of locus k is coded as 1, 0 and −1, for the three genotypes, the corresponding γ_k would be the additive effect of QTL k. The second way of obtaining QTL effects is through linear contrasts of the genotypic values. The Z_k retains its original definition as a matrix of dummy variables so that γ_k represents a vector of genotypic values. In this case, an extra step is required to obtain the QTL effects after the EM algorithm converges. First, we need to obtain the BLUP and prediction error of γ_k. Second, we define coefficients of a linear contrast and use them to convert the estimated genotypic values into a QTL effect. For example, in the F_2 line crossing example, the three components of γ_k represent the three genotypic values denoted by $\gamma_{k=}[\mathrm{G}_{11}\ \mathrm{G}_{12}\ \mathrm{G}_{22}]^{\mathrm{T}}$ for the three genotypes (A_1A_1, A_1A_2 and A_2A_2). The coefficients of the linear contrast for the additive effect may be defined as $H_{\mathrm{a}} = [1/2\ 0\ -1/2]^{\mathrm{T}}$. The additive effect for QTL k is then defined as $a_k = H_{\mathrm{a}}^{\mathrm{T}}\gamma_k$. Similarly, the dominance effect may be defined as $\mathrm{d}_k = H_{\mathrm{d}}^{\mathrm{T}}\gamma_k$, where $H_{\mathrm{d}} = [-1/4\ 1/2\ -1/4]^{\mathrm{T}}$. Define $H = H_{\mathrm{a}} || H_{\mathrm{d}}$ as the horizontal concatenation of matrices H_{a} and H_{d} (notation used in SAS language), the QTL effects (including both the additive and the dominance effects) are then obtained by using the formula $\eta_k = [a_k\ d_k]^{\mathrm{T}} = H^{\mathrm{T}}\gamma_k$. The estimated QTL effects for locus k are then

$$E(\eta_k) = \sigma_k^2 H^{\mathrm{T}} Z_k^{\mathrm{T}} V^{-1}(y - X\beta) \tag{16}$$

with a variance–covariance matrix

$$\mathrm{var}(\eta_k) = H^{\mathrm{T}}(I\sigma_k^2 - \sigma_k^2 Z_k^{\mathrm{T}} V^{-1} Z_k \sigma_k^2)H \tag{17}$$

The coefficients of linear contrasts, denoted by matrix H, can be defined in many different ways. It is up to the investigator to choose his/her own favorite scale. Therefore, the genotypic effect model is more flexible than the QTL effect model. Finally, it is possible to test the hypothesis H_0:$\eta_k = 0$ using the Wald test statistic

$$\mathrm{Wald} = \hat{\eta}_k^{\mathrm{T}} \mathrm{var}^{-1}(\hat{\eta}_k)\hat{\eta}_k \tag{18}$$

for each locus. Under the null hypothesis, Wald test statistic follows approximately a χ^2 distribution with two degrees of freedom. This allows us to calculate the P-value for each locus. Therefore, the Wald test statistics is often called the χ^2 statistics.

The variance of the prior distribution of the genotypic value is σ_k^2 for the kth QTL. After the linear contrasts (combinations), the additive effect a_k has a prior $N(0, H_{\mathrm{a}}^{\mathrm{T}} H_{\mathrm{a}} \sigma_k^2) = N(0, 1/2\ \sigma_k^2)$ and the dominance effect d_k has a prior $N(0, H_{\mathrm{d}}^{\mathrm{T}} H_{\mathrm{d}} \sigma_k^2) = N(0, \frac{3}{8}\sigma_k^2)$. The two effects are no longer independent because the prior covariance between a_k and d_k is $H_{\mathrm{a}}^{\mathrm{T}} H_{\mathrm{d}} \sigma_k^2 = -1/4\sigma_k^2$. The additive effects estimated using the allelic effect model and the genotypic effect model with linear contrast will not be affected by the coding (see results of simulations described later).

Simulation study

Experimental setup

We simulated a single large chromosome of 2400 cM (centiMorgan) long evenly covered by 481 co-dominance markers (5 cM per marker interval). The simulated population was an F_2 family derived from the cross of two inbred lines with sample size $n = 500$. The genotype indicator variable for individual j at locus k is defined as $Z_{jk} = \{1, 0, -1\}$ for the three genotypes (A_1A_1, A_1A_2, A_2A_2), respectively. Dominance effects were not simulated and also not included in the model for this simulation experiment, but will be considered in a separate experiment presented later. A total of 20 QTLs were simulated, with the sizes and locations of the QTLs listed in Table 1. These parameter values were used to generate a quantitative trait with a population mean $\beta = 10.0$ and a residual error variance $\sigma^2 = 10.0$. The total genetic variance for the trait is

$$V_G = \sum_{k=1}^{20}\sum_{k'=1}^{20} \gamma_k \gamma_{k'} \mathrm{cov}(z_k, z_{k'}) = \frac{1}{2}\sum_{k=1}^{20}\sum_{k'=1}^{20} \gamma_k \gamma_{k'} (1 - 2r_{kk'}) \quad (19)$$

where $r_{kk'}$ is the recombination coefficient between QTLs k and k', $\mathrm{cov}(z_k, z_{k'}) = \mathrm{var}(z)(1-2r_{kk'})$ is the covariance between Z_k and $Z_{k'}$ and $\mathrm{var}(Z) = 1/2$ is the variance of Z (assuming no segregation distortion). The total genetic variance for the quantitative trait is $V_G = V_Q + V_L = 66.384$, which is the sum of the genetic variances due to QTL (V_Q) and covariance between linked QTLs (V_L), where

$$V_Q = \frac{1}{2}\sum_{k=1}^{20} \gamma_k^2 = 46.7804 \quad (20)$$

and

$$V_L = \sum_{k'>k}^{20} \gamma_k \gamma_{k'} (1 - 2r_{kk'}) = 19.6034 \quad (21)$$

The residual error variance for the trait is $\sigma^2 = V_E = 10.0$. Therefore, the total phenotypic variance is $V_P = V_G + V_E = 76.384$. The proportion of the genetic variance contributed by each QTL is $0.5\gamma_k^2/V_G$ for the kth QTL (given in the column headed with Prop-G in Table 1). The corresponding proportion of the phenotypic variance contributed by the kth QTL is $0.5\gamma_k^2/V_P$ and given in the column headed with Prop-P in Table 1. The true QTL effects are depicted in Figure 1, which will be used as the standard for comparison with estimated QTL effects using various model and prior setups.

Table 1 QTL parameters used in the simulation studies

QTL	*Position (cM)*	*Marker*	*Effect*	*Prop-G*	*Prop-P*
1	50	11	4.47	0.1505	0.1308
2	125	26	3.16	0.0752	0.0654
3	205	42	−2.24	0.0378	0.0328
4	235	48	−1.58	0.0188	0.0163
5	355	72	2.24	0.0378	0.0328
6	360	73	3.16	0.0752	0.0654
7	610	123	1.10	0.0091	0.0079
8	630	127	−1.10	0.0091	0.0079
9	800	161	0.77	0.0045	0.0039
10	900	181	1.73	0.0225	0.0196
11	905	182	3.81	0.1093	0.0950
12	920	185	2.25	0.0381	0.0331
13	1100	221	−1.30	0.0127	0.0111
14	1210	243	−1.00	0.0075	0.0065
15	1305	262	−2.24	0.0378	0.0328
16	1335	268	1.58	0.0188	0.0163
17	1345	270	1.00	0.0075	0.0065
18	1365	274	−1.73	0.0225	0.0196
19	1800	361	0.71	0.0038	0.0033
20	2300	461	0.89	0.0060	0.0052

Abbreviations: cM, centiMorgan; Prop-G, the proportion of genetic variance contributed by the QTL; Prop-P, the proportion of phenotypic variance contributed by the QTL; QTL, quantitative trait loci.
The QTL effects are referred to the additive effects only.

Allelic effect model

Under the allelic effect model, we numerically coded the three genotypes with $Z_k = \{1, 0, -1\}$ for the three genotypes $\{A_1A_1, A_1A_2, A_2A_2\}$. The QTL effects were directly estimated without taking linear contrasts of the genotypic values. For 481 markers, the Z matrix has a dimensionality of 500×481. Three different priors were chosen for this data analysis: (1) $\xi = (\tau, \omega) = (-2, 0)$ representing uniform prior for σ_k^2; (2) $\xi = (\tau, \omega) = (0, 0)$ representing the Jeffreys' prior for σ_k^2; (3) the Lasso prior $\lambda^2 = 5.1758$. This particular Lasso prior value was chosen using the following empirical method,

$$\lambda^2 = \left[\frac{1}{p}\sum_{k=1}^{p} \sigma_k^2\right]^{-1/2} \quad (22)$$

More information about this empirical Lasso parameter will be discussed later. The results for the three different priors are presented in graphical form for the reason that a tabular form of presentation is hard to show all the small estimated QTL effects. The results are depicted in Figure 2, showing that the Jeffreys' prior appears to be better than the Lasso prior, but both are better than the uniform prior. The QTL effect profile of the Jeffreys' prior mimics the true QTL effect profile (see Figure 1) more closely than the other two priors. Compared with the Jeffreys' prior, the Lasso prior tends to split major QTL effects into a few small effects in the neighborhood of the true QTL. Therefore, the Lasso-estimated QTL effect profile tends to have many small 'bumps' along the genome.

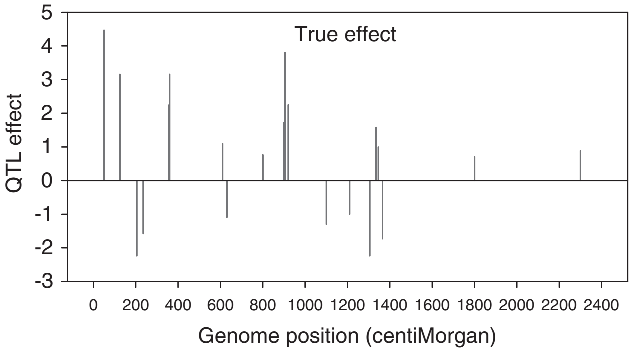

Figure 1 True quantitative trait loci (QTL) effects (additive model) and locations of QTLs in a simulated genome of 2400 cM (centiMorgan) in length.

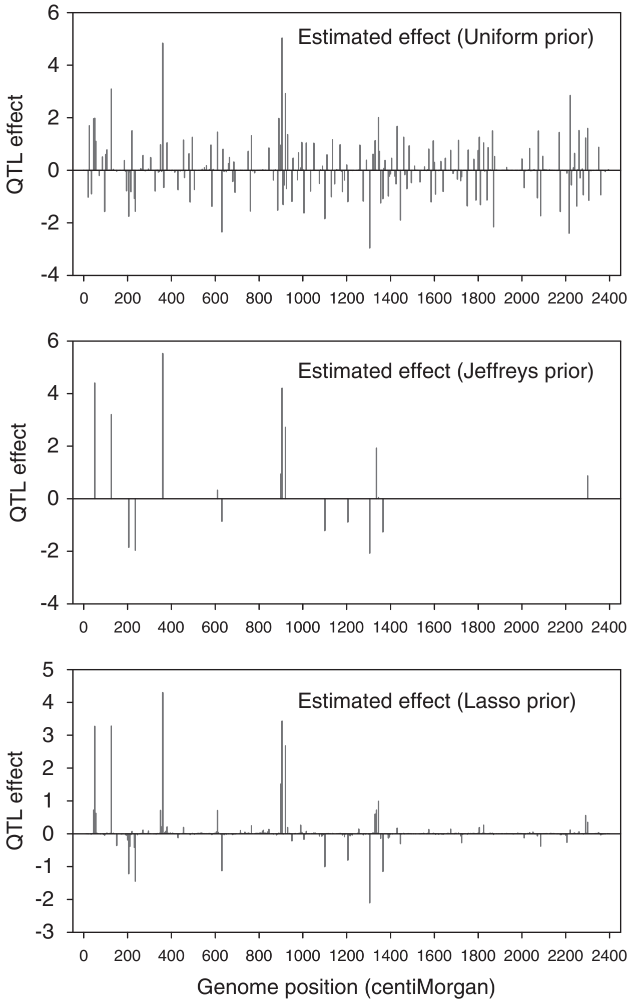

Figure 2 Estimated additive effects and locations of quantitative trait loci (QTLs) for the simulated data under the allelic effect model. The uniform prior is equivalent to $\xi=(\tau,\,\omega)=(-2,\,0)$. The Jeffreys' prior is equivalent to $\xi=(\tau,\,\omega)=(0,\,0)$. The Lasso prior is $\lambda^2=5.1758$.

We used the mean squared error (MSE) of the estimated QTL effects to further evaluate the performance of the three priors. The MSE is defined as

$$\mathrm{MSE}(\tau,\omega)=\frac{1}{481}\sum_{k=1}^{481}(\gamma_k^{\mathrm{Inv}-\chi^2}-\gamma_k)^{\mathrm{T}}(\gamma_k^{\mathrm{Inv}-\chi^2}-\gamma_k) \tag{23}$$

for the scaled inverse χ^2 prior and

$$\mathrm{MSE}(\lambda^2)=\frac{1}{481}\sum_{k=1}^{481}(\gamma_k^{\mathrm{Lasso}}-\gamma_k)^{\mathrm{T}}(\gamma_k^{\mathrm{Lasso}}-\gamma_k) \tag{24}$$

for the Lasso prior, where $\gamma_k^{\mathrm{Inv}-\chi2}$ is the BLUP value obtained under the scaled inverse χ^2 distribution, $\gamma_k^{\mathrm{Lasso}}$ is the BLUP value obtained under the Lasso prior distribution and γ_k is the true value. The MSE comparison shows that $\mathrm{MSE}(-2,\,0)=0.351129659$, $\mathrm{MSE}(0,\,0)=0.034842259$ and $\mathrm{MSE}(5.1758)=0.033882049$. Therefore, the Jeffreys' prior and the Lasso prior perform equally well, and both are better than the uniform prior. The noisy signals of the Lasso prior have not increased the MSE compared with the Jeffreys' prior. In fact, they have improved (decreased) the MSE slightly.

Genotypic effect model

The same data set was also analyzed using the genotypic effect model, in which the Z matrix was coded as dummy variables. For 481 markers, the Z matrix has $481\times3=1443$ columns, and thus, 1443 genotypic values were estimated. To compare this analysis with the allelic effect model, we used linear contrast H_{a} (described earlier) to convert the three genotypic values of each locus into an additive effect. The dominance effects, however, were not simulated (zero effects for all loci). Again, the three priors chosen in the allelic effect model analysis were used here, that is, $\xi=(\tau,\,\omega)=(-2,\,0)$, $\xi=(\tau,\,\omega)=(0,\,0)$ and $\lambda^2=4.786525$. The results are almost duplicates of the allelic effect model. The additive effect profiles for the three priors are almost the same as that obtained in the allelic effect model (data not shown). The estimation errors are also very close for the two models (data not shown). The MSEs of the three priors are $\mathrm{MSE}(-2,\,0)=0.417594$, $\mathrm{MSE}(0,\,0)=0.0682055$ and $\mathrm{MSE}(4.786525)=0.031560243$, respectively. The Lasso prior appeared to perform slightly better than the Jeffreys' prior. The genotypic effect model and the allelic effect model can be used interchangeably for QTL mapping in line crosses. For line crossing experiments such as BC and F_2, there is no advantage of using the genotypic effect model except that this model provides estimated genotypic values so that investigators can directly interpret the results regarding which parent is carrying the 'high' or 'low' allele at each locus.

Simulation with dominance effects

To examine the efficiency of the EM algorithm for estimating the dominance effects, we simulated another data set with all other settings being the same as the simulated data set described before except that we added six dominance effects to the genome. The sizes and the locations of the dominance effects are depicted in Figure 3a (the upper panel). For simplicity, we only report the result for the Jeffreys' prior $\xi=(\tau,\,\omega)=(0,\,0)$ under the genotypic effect model. The estimated additive effects and the dominance effects are depicted in Figure 3b (the lower panel). The estimated genotypic values and other relevant information for the data analysis are presented in Table 3. We used $\hat{a}_k=H_{\mathrm{a}}^{\mathrm{T}}\hat{\gamma}_k$ and $\hat{d}_k=H_{\mathrm{d}}^{\mathrm{T}}\hat{\gamma}_k$ to convert the genotypic values γ_k into additive (a) and dominance (d) effects. The variance–covariance matrix of the estimated QTL effects $\hat{\eta}_k=[\hat{a}\ \hat{d}_k]^{\mathrm{T}}$ are then calculated and used to generate the Wald test statistic and the P-value using

$$P\text{-value}=1-P_{\chi^2}^{-1}(\mathrm{Wald},\,2) \tag{25}$$

where $P_{\chi2}^{-1}$ denote the inverse of the χ^2 distribution function with two degrees of freedom. We used an arbitrary cutoff point to determine the 'significance' of each locus using P-value <0.01 as the criterion of significance. The Wald test statistics and the P-values are listed in Table 2 for all the 24 simulated loci. All but four of the 24 loci were detected. The four loci that failed to reach the cutoff P-value are markers 123, 127, 243 and 270. Markers 123 and 127 are 20 cM apart from each other and each had an additive effect of 1.1 but with opposite signs. Marker 243 had an additive effect of $a=-1.0$, explaining only 0.65% of the phenotypic variance. Marker 270 had an additive effect of $a=1.0$, also explaining only 0.65% of the phenotypic variance. In fact, this marker is only 10 cM apart from marker 268, which had an additive effect of $a = 1.58$. The effect of marker 270 was absorbed by marker 268, because the estimated effect of marker 268 is $a = 2.147$, slightly less than $2.58 = 1.58 + 1.0$ (sum of the additive effects of the two loci).

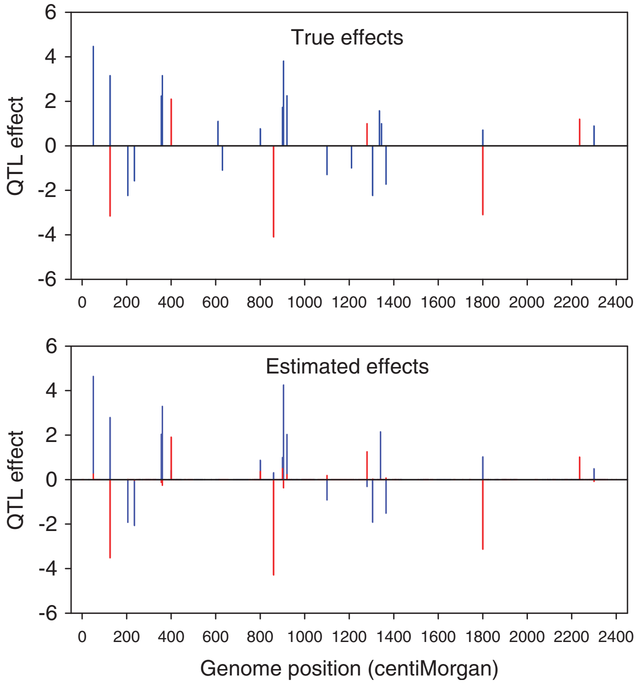

Figure 3 Additive and dominance quantitative trait loci (QTL) effects in the second simulation experiment. The upper panel shows the true QTL effects and the lower panel shows the estimated QTL effects under the genotypic effect model with the Jeffreys' shrinkage prior. The blue needles with diamonds represent the additive effects and the red needles with triangles represent the dominance effects.

Alternative values of hyperparameters

For the same simulated data set without dominance effects (described in the experimental setup section), we chose a few alternative hyperparameters for the scaled inverse χ^2 distribution and a few alternative Lasso parameters to evaluate the performance of the new method. We only evaluated the allelic effect model for its simplicity and quickness. For the scaled inverse χ^2 prior, we first let $\xi = (\tau, \omega) = (\tau, 0)$ and only varied τ from 0 to -1, decremented by 0.1. This type of priors was proper and suggested by ter Braak *et al.* (2005). In addition, we let $\xi = (\tau, \omega) = (-0.5, \omega)$ and varied ω from 0 to 1, incremented by 0.1. For the Lasso prior, we chose λ^2 in the neighborhood of $\lambda^2 = 5.1758$ (empirical value obtained earlier for this data set) ranging from 1 to 10, incremented by 1. We used the MSE to evaluate the performance of the method under various hyperparameter values. The MSE of these priors are presented in Table 3. For the set of priors in the $\xi = (\tau, \omega) = (\tau, 0)$ series (Prior I), the minimum MSE occurs at $\tau \approx -0.6$. For the set of priors in the $\xi = (\tau, \omega) = (-0.5, \omega)$ series (Prior II), the minimum MSE occurs at $\omega \approx 0.05$. A slight increase of ω will dramatically increase the MSE. Therefore, $0 \leqslant \omega \leqslant 0.1$ seems to be optimal. For the Lasso priors (Prior III), the minimum MSE occurs when $6.0 \leqslant \lambda^2 \leqslant 10.0$. The empirical value of $\lambda^2 = 5.1758$ is not far away from the optimal values. Note that these optimal hyperparameters are sample specific and may not be generalized to other samples. More discussion on the optimal hyperparameters will be presented later.

Table 2 Estimated genotypic values of the three genotypes (A_1A_1, A_1A_2 and A_2A_2), and the corresponding additive and dominance effects of QTL obtained under the genotypic effect model using the Jeffreys' shrinkage prior $\xi = (\tau, \omega) = (0, 0)$

QTL	*cM*	*Marker*	A_1A_1	A_1A_2	A_2A_2	*a*	*d*	*std(a)*	*std(d)*	*Wald*	P-*value*
1	50	11	−4.808	0.334	4.474	4.641	0.251	0.224	0.156	432.579	0.000
2	125	26	−0.444	−4.692	5.136	2.790	−3.519	0.237	0.157	650.268	0.000
3	205	42	1.919	0.008	−1.927	−1.923	0.006	0.259	0.166	55.461	0.000
4	235	48	2.091	−0.049	−2.043	−2.067	−0.037	0.262	0.166	62.400	0.000
5	355	72	−1.950	−0.175	2.126	2.038	−0.131	0.472	0.258	18.866	0.000
6	360	73	−3.120	−0.345	3.465	3.292	−0.259	0.472	0.261	49.376	0.000
7	400	81	−1.694	2.547	−0.854	0.420	1.910	0.238	0.156	151.211	0.000
8	610	123	0.000	0.000	0.000	0.000	0.000	0.003	0.003	0.002	0.999
9	630	127	0.001	0.001	−0.001	−0.001	0.000	0.008	0.007	0.022	0.989
10	800	161	−1.117	0.492	0.625	0.871	0.369	0.214	0.151	23.673	0.000
11	860	173	2.546	−5.720	3.173	0.313	−4.290	0.242	0.161	708.659	0.000
12	900	181	−1.323	0.664	0.659	0.991	0.498	0.417	0.251	11.473	0.003
13	905	182	−4.004	−0.498	4.502	4.253	−0.374	0.484	0.275	77.208	0.000
14	920	185	−2.179	0.299	1.880	2.029	0.224	0.324	0.185	41.878	0.000
15	1100	221	0.790	0.250	−1.040	−0.915	0.187	0.207	0.150	19.889	0.000
16	1210	243	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000
17	1280	257	−0.521	1.668	−1.147	−0.313	1.251	0.218	0.154	66.361	0.000
18	1305	262	1.902	0.029	−1.931	−1.916	0.022	0.239	0.161	64.450	0.000
19	1335	268	−2.161	0.028	2.133	2.147	0.021	0.288	0.171	55.829	0.000
20	1345	270	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
21	1365	274	1.465	0.097	−1.562	−1.514	0.073	0.258	0.164	34.922	0.000
22	1800	361	1.065	−4.173	3.108	1.021	−3.130	0.216	0.158	406.707	0.000
23	2235	448	−0.707	1.352	−0.645	0.031	1.014	0.213	0.152	44.571	0.000
24	2300	461	−0.437	−0.111	0.548	0.492	−0.083	0.182	0.133	7.665	0.022

Abbreviations: *a*, additive effect; cM, centiMorgan; *d*, dominance effect; QTL, quantitative trait loci.
Note that std(*a*) and std(*d*) are standard errors of the estimated additive and dominance effects, respectively.

Table 3 The mean squared error (MSE) of alternative prior choice for the simulated data set reported in the 'experimental setup' section under the allelic effect model

Hyperparameter (ϕ)	*Prior I* $\xi = (-0.1\phi, 0)$	*Prior II* $\xi = (-0.5, 0.1\phi)$	*Prior III* $\lambda^2 = \phi$
−5	0.044865	—	—
0	0.034842	0.034136	0.336384
0.5	0.034338	0.029194	0.092581
1	0.033919	0.038428	0.063728
2	0.033553	0.066319	0.047295
3	0.033433	0.103673	0.040110
4	0.033323	0.143936	0.035931
5	0.034136	0.190627	0.033142
6	0.030042	0.240278	0.031770
7	0.031679	0.291576	0.030694
8	0.030630	0.341711	0.030058
9	0.031118	0.393937	0.030101
10	0.031358	0.444304	0.029975
15	0.041820	0.696547	0.031143
50	—	—	0.051102

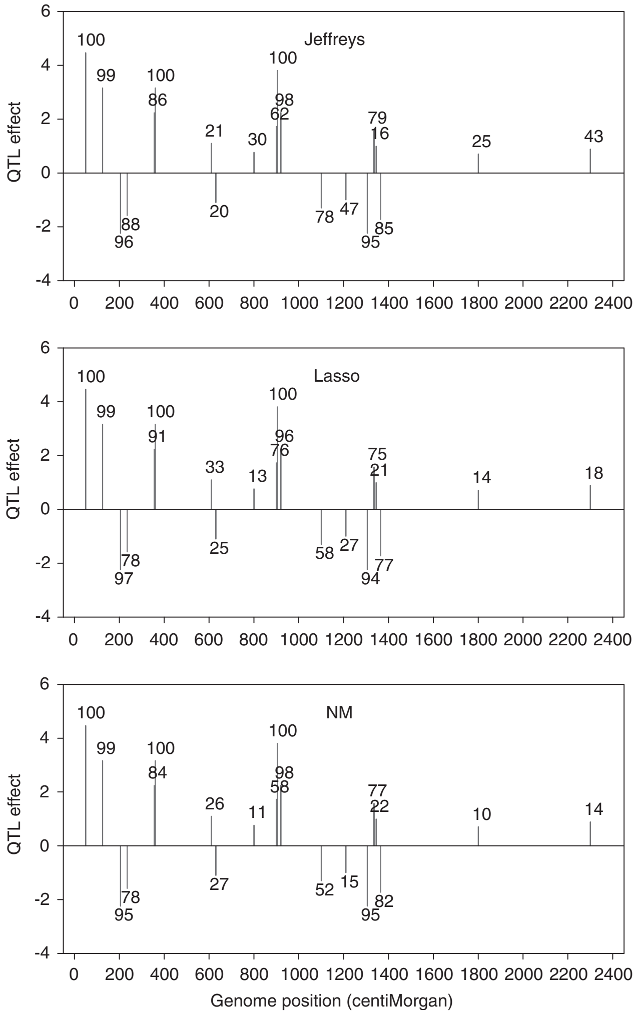

Figure 4 Average estimated quantitative trait loci (QTL) effects over 100 replicated simulations and the empirical statistical powers for the simulated QTL. The upper panel shows the results of the Jeffreys' prior. The panel in the middle shows the results of the Lasso prior. The lower panel shows the results of the Nelder–Mead (NM) algorithm with hyperparameter $\xi = (\tau, \omega) = (-0.5, 0.05)$. The heights of the needles represent the average estimated QTL effects. The numbers at the tip (end) of the needles represent the statistical powers measured in percentage, for example, the integer 99 represents 99%.

Power and false-positive rate

The Bayesian methods presented here can be reinterpreted for classical power analysis using replicated simulation experiments. In this section, we used the same QTL parameters given in Table 1 and the same experimental setup to simulate 100 additional samples for power analysis. We used the allelic effect model to estimate parameters and calculate the test statistics. As we only considered the additive effects, the test statistic for each locus is defined as the squared QTL effect divided by the squared prediction error of the estimated QTL effect. Under the null hypothesis, this test statistic approximately follows a χ^2 distribution with one degree of freedom. This allows us to calculate the *P*-value for each locus. We chose 0.01 as the threshold for the *P*-value to determine the significance of a locus with a QTL effect and the false-positive status of a locus with no QTL effect. In other words, if a QTL has a *P*-value <0.01 in a particular replication, the QTL is claimed to be detected in that replication and the proportion of the replicates in which the QTL is detected out of the 100 replications is the empirical statistical power for that QTL. As the power was evaluated for each QTL, the false-positive rate (FPR) should also be defined in a locus-specific manner. A locus with no QTL effect is labeled false positive if the *P*-value is smaller than the 0.01 threshold. The FPR of the non-QTL locus is then defined as the proportion of the replicates labeled as false positive out of the 100 replications. The FPR is also called the Type I error. We simulated 20 QTLs out of 481 loci. The distance between any consecutive loci is 5 cM. We observed that the effect of a QTL failing to be detected was very often picked up by a marker in the neighborhood. If a neighboring marker reaches the significance level, this QTL is also claimed to be detected. Therefore, for every true QTL, three consecutive loci (with the true QTL in the center) are claimed as QTLs. A non-QTL is defined as a locus that is separated by at least one neutral marker from a true QTL.

The 100 replicated samples were analyzed using three different priors (methods): the Lasso method (the Lasso parameter was empirically estimated), the Jeffreys' method (Jeffreys's prior was used) and the method of Xu (2007) implemented with the Nelder and Mead (1965) simplex algorithm. The three methods are denoted as Lasso, Jeffreys and NM, respectively. For some reasons, the NM method cannot handle the Jeffreys' prior. Therefore, $\xi = (\tau, \omega) = (-0.5, 00.5)$ was used as the prior for the NM method. The average estimated QTL effects for all the 481 loci over the 100 replications are depicted in Figure 4, for all the three methods. The heights of the needles represent the average estimated QTL effects. The empirical statistical powers (numerical values) for the loci are placed at the tips of the needles in Figure 5. The three methods have similar powers, with the Jeffreys method slightly better than the Lasso method, which is slightly better than the NM method. Figure 4 shows the corresponding biases of the estimated QTL effects for the three methods. The biases are typically between −0.6 and 0.6. Two loci show large biases for the Jeffreys' prior, from −0.8 to 0.8. The Bayesian shrinkage method is expected to be biased. The biases observed from the

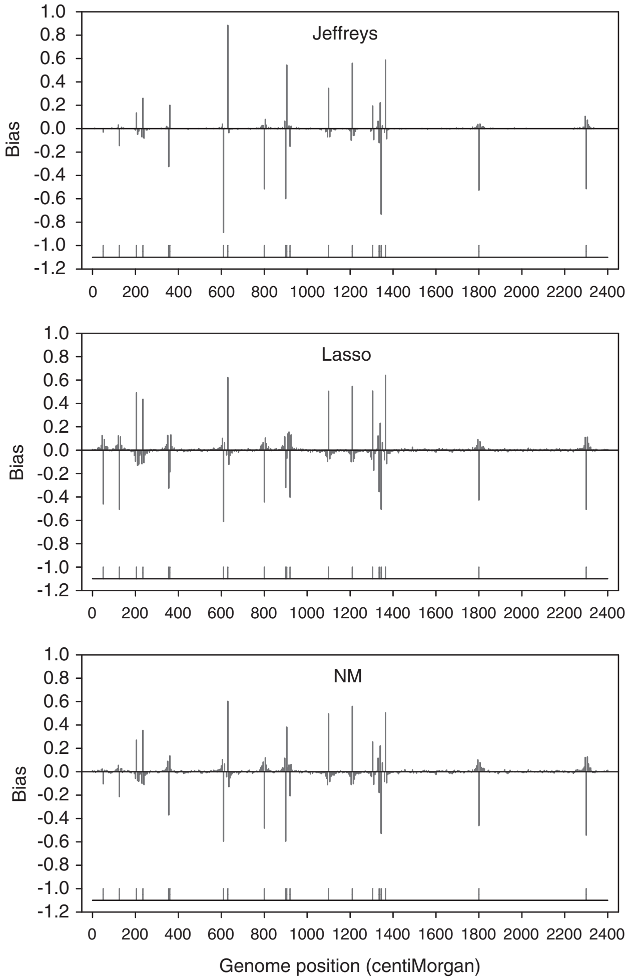

Figure 5 The biases (average estimated quantitative trait loci (QTL) effect–true QTL effect) of the estimated QTL effects for three methods. The upper, middle and lower panels represent the Jeffreys' method, the Lasso method and the Nelder–Mead (NM) method, respectively. The true QTL positions are marked with the ticks of the horizontal axis.

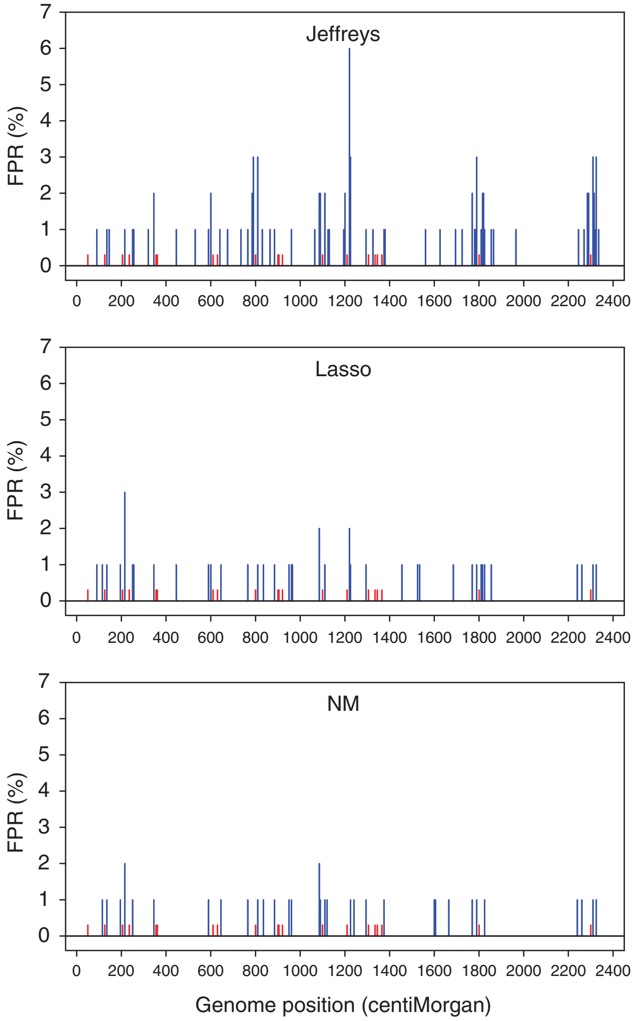

Figure 6 The false-positive rate (FPR) profiles for the three methods. The upper, middle and lower panels represent the Jeffreys' method, the Lasso method and the Nelder–Mead (NM) method, respectively. The quantitative trait loci (QTL) positions are marked with the red ticks.

repeated simulation experiments are not too serious compared with the actual values of the QTL effects.

Figure 6 presents the FPR profiles for the three methods. Most of the non-QTLs have zero FPR. A small percentage of the loci have one false positive out of the 100 replications. For the Jeffreys' method, one locus has 6% FPR, six loci have 3% FPR and 14 loci have 2% FPR. The largest FPR occurs near a true QTL position with a small effect. The Lasso method has one locus with 3% FPR and two loci with 2% FPR. The NM method has the lowest FPR. Overall, all the three methods have quite low FPR.

The average numbers of iterations required to converge were 23.51, 15.96 and 11.81, respectively, for the three methods (Lasso, Jeffreys and NM). The corresponding total computing times for completing the analysis of 100 replications were 128 min (Lasso), 89 min (Jeffreys) and 100 min (NM) for the three methods. The longer computing time for the Lasso method was due to the large number of iterations required for the program to converge. The average estimated QTL parameters along with the estimated population mean and residual variance obtained from 100 replicated simulations are provided in the supplemental material for interested readers. The original simulated data sets are also given in the supplemental material.

Real data analysis

We used a real data set from recombinant inbred lines of *Arabidopsis* (Loudet *et al.*, 2002) as an example to show the application of the method. The two parents initiating the line cross were Bay-0 and Shahdara with Bay-0 as the female parent. The recombinant inbred lines were actually F_7 progeny of single seed descendants (selfing) of the F_2 plants. The residual heterozygosity was low (Loudet *et al.*, 2002). Flowing time was recorded for each line in two environments: long day (16 h photoperiod) and short day (8 h photoperiod). We used the short-day flowering time as the quantitative trait for QTL mapping. The two parents had very little difference in short-day flowering time. The sample size (number of recombinant lines) was 420. A couple of lines did not have the phenotypic records, and their phenotypic values were replaced by the population mean for convenience of data analysis. A total of 38 microsatellite markers were used for QTL mapping. These markers are more or less evenly distributed along five chromosomes with an average 10.8 cM per marker interval. The marker names and positions can be found in the original article (Loudet *et al.*, 2002).

We inserted a pseudo marker in every 2 cM of the genome. With the inserted pseudo markers, the total number of loci subject to analysis is 200 (38 true markers plus 162 pseudo markers). All the 200 putative loci were evaluated simultaneously in a single model. Therefore, the model for the short-day flowering time trait is

$$y = X\beta + \sum_{k=1}^{200} Z_k \gamma_k + \varepsilon \qquad (26)$$

where X is a 420×1 vector of unity, β is the population mean (intercept), Z_k is a 420×1 vector coded as 1 for one genotype and 0 for the other genotype for locus k. If locus k is a pseudo marker, $Z_k = \Pr(\text{genotype} = 1)$, which is the conditional probability of marker k being of genotype 1. Finally, γ_k is the QTL effect of locus k. We only used the allelic effect model for the real data analysis.

The data were analyzed using three different priors, (1) $\xi = (\tau, \omega) = (-2, 0)$ corresponding to the uniform prior, (2) $\xi = (\tau, \omega) = (0, 0)$ representing the Jeffreys' prior and (3) the Lasso prior with $\lambda^2 = 3.2739$. The estimated QTL effects are depicted in Figure 7. The Jeffreys' prior (the panel in the middle of Figure 7) produced the cleanest signals of QTL effects. Four QTLs were detected in three chromosomes. The uniform prior (the panel at the top of Figure 7) and Lasso prior (the panel at the bottom of Figure 7) also produced four peaks corresponding to the same positions as those detected by the Jeffreys' prior. However, additional signals also occur for these two priors. The estimated QTL effects and QTL positions along with the *t*-test statistics and other information under the Jeffreys' prior are given in Table 4.

We also performed an interval mapping on the short-day flowering time trait. The results are depicted in Figure 8. Results of chromosome 1, 2, 3 and 4 agree well with our Bayesian analysis. However, interval mapping cannot separate the two QTLs in chromosome 5. Detailed result of interval mapping can be found in the original study (Loudet *et al.*, 2002).

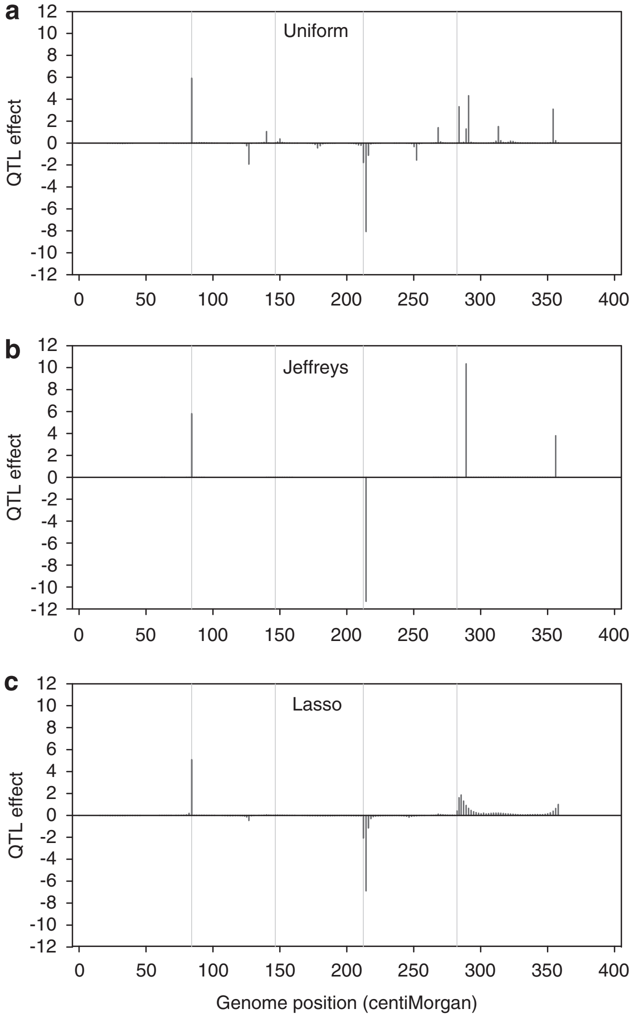

Figure 7 Estimated effects and locations of quantitative trait loci (QTLs) for the trait of short-day flowering time of *Arabidopsis* (Loudet *et al.*, 2002). The five chromosomes are merged into a single genome and separated by the dotted green reference lines. The upper panel represents the results using the uniform prior. The panel in the middle represents the results using the Jeffreys' prior. The lower panel gives the results of the Lasso prior with $\lambda^2 = 3.2739$.

Table 4 The estimated QTL parameters for the *Arabidopsis* data using the Jeffreys' prior under the allelic effect model

QTL	*Chr.*	*cM*	*Effect*	*StdErr*	t-*Test*	*Variance*	*Prop-P*
1	1	84.1	5.8068	0.4909	11.8271	8.1768	0.0967
2	4	2.0	−11.3116	0.5043	−22.4302	29.6081	0.3503
3	5	6.825	10.3596	0.5985	17.3090	20.3803	0.2411
4	5	73.733	4.8789	0.4881	7.7745	4.9297	0.0583

Abbreviations: Chr, chromosome; cM, centiMorgan; Prop-P, the proportion of phenotypic variance contributed by the QTL; QTL, quantitative trait loci; StdErr, standard error.

The estimated population mean and the residual error variance are $\beta = 59.7088$ and $\sigma^2 = 29.8626$, respectively. The total genetic variance contributed by all the four QTLs is $V_G = V_Q + V_L = 63.0590 + (-8.4471) = 54.6478$. The overall phenotypic variance is $V_P = V_G + \sigma^2 = 54.6478 + 29.8629 = 84.5108$. The total proportion of the phenotypic variance contributed by the four QTL is $H^2 = V_Q / V_P = 0.7466$.

Discussion

The EM algorithm developed in this study is not a new method of QTL mapping. It is an alternative algorithm used to find the empirical Bayesian estimates of QTL effects. All properties of the empirical Bayesian method of Xu (2007) implemented through the simplex algorithm apply to the EM algorithm. These properties (for example, dealing with epistatic effects) have been investigated by Xu (2007), and thus, were not further explored in the current study. The advantages of the EM algorithm over the simplex algorithm are the flexibility to handle both the allelic effect model and the genotypic effect model, and the ability to deal with the Lasso prior. Although the simplex method in general can handle genotypic effect models, the fast algorithm to invert the variance matrix described by Xu (2007) cannot be applied, because that algorithm only holds for the allelic effect model in which each regression coefficient has its own variance. Another advantage of the EM algorithm is its transparency of the formulation, as apposed to the simplex algorithm, so that programming of the EM

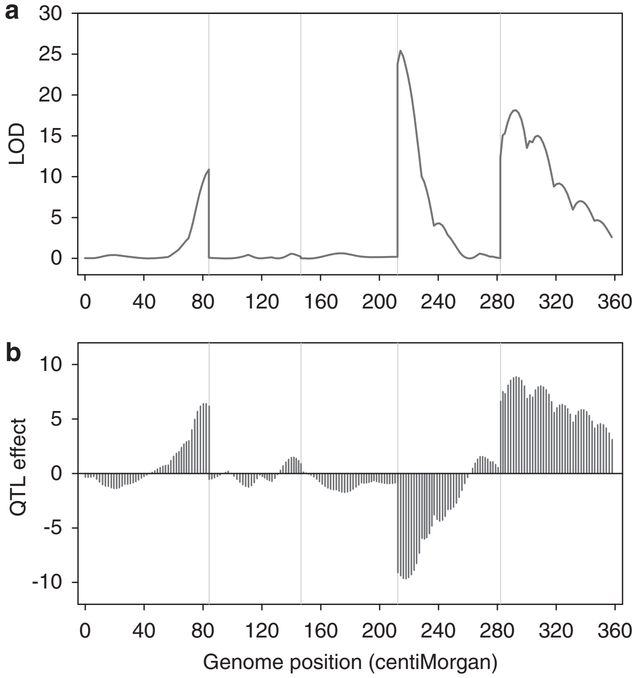

Figure 8 LOD (logarithm (base 10) of odds) score profile of the *Arabidopsis* short-day flowering time quantitative trait loci (QTL) mapping resulted from interval mapping. The upper panel shows the LOD score profile. The lower panel shows the estimated QTL effect profile.

algorithm becomes much easier. Similar to any other EM algorithms, our EM algorithm also has its own limit in terms of slow convergence when the parameters are near the local optimum. Therefore, the simplex algorithm adopted in the original empirical Bayes (Xu, 2007) still has its value in terms of fast convergence and robustness to the initial values.

The empirical Bayesian estimation of QTL effects is a kind of posterior mode estimation, and thus, is different from the fully Bayesian estimation implemented through the MCMC algorithm (Xu, 2003; Wang *et al*., 2005). If the Markov chain is sufficiently long, results of the MCMC sampling would be better than the posterior mode estimation. However, the posterior mode estimation is a quick method to achieve the results that are almost as good as the fully Bayesian estimation. For the same simulated data, the EM algorithm took about 1 min to complete the estimation, whereas the MCMC-implemented sampling algorithm took about one-half hour (data not shown). In addition, our experience showed that the Jeffreys' prior usually performs well compared with other hyperparameter values. However, the Jeffreys' prior is improper in the sense that a marginal posterior distribution of σ_k^2 does not exist (ter Braak *et al*., 2005). Although we are not interested in σ_k^2 *per se*, but use σ_k^2 as a shrinkage factor to control the estimate of γ_k, an improper posterior σ_k^2 always presents a warning signal regarding the convergence of the chain. Theoretically, all parameters should converge to the stationary distribution to validate the MCMC algorithm. The posterior mode estimation does not have such a concern.

An obvious question with the posterior mode estimation is how to choose the hyperparameter $\xi = (\tau, \omega)$ or λ^2. We have noticed that the hyperparameter has a large role in the final estimates of QTL effects. A common way of choosing the hyperparameter is to use a cross-validation test. Tibshirani (1996) in the original Lasso method took a fivefold cross-validation approach. We can adopt the same cross-validation method to help determine the optimal hyperparameter. If desired, cross-validation can be conducted by the users, because standard x-fold cross-validation is straightforward and easy to program. However, using cross-validation to determine the optimal parameter may also have its own problems. For example, the optimal Lasso parameter λ^2 may depend on both the sample size and the dimensionality of the model. Assume that we decide to use the recommended fivefold cross-validation to determine the optimal λ^2. The optimal value found in the fivefold validation may not be optimal at all if a threefold cross-validation is performed. What is the optimal x in the x-fold cross-validation? Suppose that the fivefold cross-validation is the choice and we do not want to use any other folds, the optimal λ^2 in fact is only optimal for sample size *4n/5*, but our sample size is actually n. The question may keep coming one after another.

If one decides not to use a cross-validation to determine the hyperparameters, we offer the following suggestions based on our own experience of data analyses. The scale parameter ω in $\xi = (\tau, \omega)$ can be set to zero or close to zero, say 0.001, and thus, we only have one hyperparameter τ to worry about. We should start with the Jeffreys' prior $\xi = (\tau, \omega) = (0, 0)$ and then choose an improved value from there. A cross-validation can be used to evaluate a few alternative values around $\tau = 0$. Given that the algorithm is computationally efficient, a wide range of values of τ can be evaluated within a short period of time.

The Lasso prior should be found using the cross-validation method suggested by Tibshirani (1996). By trial and error, we found that equation (22) usually is a good choice for the Lasso parameter. Let $\bar{v} = p^{-1}\sum_{k=1}^{p}\sigma_k^2$ be the average of the QTL variance components. The empirical Lasso prior is simply $\lambda^2 = \sqrt{1/\bar{v}}$. Intuitively, when all QTLs have very large variance components, the average should also be large, and thus, the Lassos prior should be small (little shrinkage). If all QTL effects have small variance components, the average should also be small, leading to strong shrinkage. If we treat λ^2 as an unknown parameter and estimate it through maximization of the expected complete-data log likelihood function, the solution would be $\lambda^2 = 1/\bar{v}$. However, this value did not work, because the shrinkage was too strong so that all regression coefficients would be shrunken to zero. It's square root worked just fine, but provided no theoretical proof. We used this empirical shrinkage parameter for the simulated data (500 individuals and 481 markers) and found that the optimal value λ^2 was in the range between 6 and 10. It turned out that the empirical value of $\lambda^2 = 5.1758$ is not far away from that optimal range.

Programming the EM algorithm developed in the study is made straightforward by following the EM steps described earlier. However, users can download the SAS/IML code that we used to analyze the simulated data. The SAS/IML code (EM-Lasso) along with the data is posted on our website (www.statgen.ucr.edu). Skilled SAS users may use PROC MIXED and PROC IML interactively with the SAS MACRO to call the iterative process. We can use PROC MIXED to calculate β and γ, with variance parameters held at the values provided in a SAS data set. PROC MIXED is extremely efficient in estimating β and predicting γ. PROC IML can be used to calculate the variance components using the predicted γ and their standard errors generated by PROC MIXED. The calculated variance components are stored in a SAS data set, which in turn is called by PROC MIXED as the input parameter values. Finally, we can use a SAS MACRO to connect the two procedures iteratively and call the macro to achieve the EM estimates of QTL effects. There is a newly released mixed model procedure in SAS called PROC HPMIXED. This new procedure is a simplified version of PROC MIXED, designed with the purpose of fast speed. We can replace PROC MIXED by PROC HPMIXED to improve the computational efficiency.

Finally, association study for quantitative traits involves no new statistical methods beyond the methods presented for linkage studies. The two only differ by the populations used for marker analysis. Association study uses randomly selected individuals from a target population for mapping. As a result, the inference space is the entire population from which the individuals are sampled. Linkage study, however, uses all individuals from the same family of line cross, and thus, the inference space is only the two lines initiating the cross. Association study can narrow down the actual genes because of cumulative historical recombinants, whereas the linkage study cannot unless the sample size is extremely large. The EM algorithm developed here can be used for both linkage study and association study, except that the fixed effects in the association study should be designed so that they can capture population admixture and other complicated factors unique to association study. The genotypic effect model is more useful than the allelic effect model in association study, because the number of genotypes per locus may vary from one locus to another. When the number of genotypes per locus is very large, linear contrasts for QTL effect conversion are not easy to define. In this case, association of marker k with a trait is actually indicated by the estimated value of σ_k^2.

Conflict of interest

The author declares no conflict of interest.

Acknowledgements

I wish to thank three anonymous reviewers and the associate editor for their critical and constructive comments on the manuscript. This project was supported by the National Research Initiative (NRI) Plant Genome of the USDA Cooperative State Research, Education and Extension Service (CSREES) 2007-02784.

References

Broman KW, Speed TP (2002). A model selection approach for the identification of quantitative trait loci in experimental crosses. *J R Stat Soc Series B* **64**: 641–656.

Dempster AP, Laird NM, Rubin DB (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc B* **39**: 1–38.

Figueiredo MAT (2003). Adaptive sparseness for supervised learning. *IEEE Trans Pattern Anal Mach Intell* **25**: 1151–1159.

Giri NC (1996). *Multivariate Statistical Analysis*. Marcel Dekker Inc: New York. pp 53–63.

Han L, Xu S (2008). A Fisher scoring algorithm for the weighted regression method of QTL mapping. *Heredity* **101**: 453–464.

Hoerl AE, Kennard RW (1970). Ridge regression: application to nonorthogonal problems. *Technometrics* **12**: 68–82.

Lan H, Chen M, Flowers JB, Yandell BS, Stapleton DS, Mata CM *et al*. (2006). Combined expression trait correlations and expression quantitative trait locus mapping. *PLoS Genet* **2**: e6.

Lan H, Stoehr JP, Nadler ST, Schueler KL, Yandell BS, Attie AD (2003). Dimension reduction for mapping mRNA abundance as quantitative traits. *Genetics* **164**: 1607–1614.

Lindstrom MJ, Bates DM (1988). Newton-Raphson and EM algorithms for linear mixed-effects models for repeated-measures data. *J Am Stat Assoc* **83**: 1014–1022.

Loudet O, Chaillou S, Camilleri C, Bouchez D, Daniel-Vedele F (2002). Bay-0 x Shahdara recombinant inbred line population: a powerful tool for the genetic dissection of complex traits in Arabidopsis. *Theor Appl Genet* **104**: 1173–1184.

Manichaikul A, Moon JY, Sen S, Yandell BS, Broman KW (2009). A model selection approach for the identification of quantitative trait loci in experimental crosses, allowing epistasis. *Genetics* **181**: 1077–1086.

Nelder JA, Mead R (1965). A simplex method for function minimization. *Comput J* **7**: 308–313.

Park T, Casella G (2008). The Bayesian Lasso. *J Am Stat Assoc* **103**: 681–686.

ter Braak CJ, Boer MP, Bink MC (2005). Extending Xu's Bayesian model for estimating polygenic effects using markers of the entire genome. *Genetics* **170**: 1435–1438.

Tibshirani R (1996). Regression shrinkage and selection via the Lasso. *J R Stat Soc Series B* **58**: 267–288.

Wang H, Zhang Y, Li X, Masinde GL, Mohan S, Baylink DJ *et al.* (2005). Bayesian shrinkage estimation of quantitative trait loci parameters. *Genetics* **170**: 465–480.

Whittaker JC, Thompson R, Denham MC (2000). Marker-assisted selection using ridge regression. *Genet Res* **75**: 249–252.

Xu S (1998). Iteratively reweighted least squares mapping of quantitative trait loci. *Behav Genet* **28**: 341–355.

Xu S (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* **163**: 789–801.

Xu S (2007). An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics* **63**: 513–521.

Yi N, Xu S (2008). Bayesian LASSO for quantitative trait loci mapping. *Genetics* **179**: 1045–1055.

Yi N, Banerjee S (2009). Hierarchical generalized linear models for multiple quantitative trait locus mapping. *Genetics* **181**: 1101–1113.

Appendix A

Derivation of BLUP

Let us rewrite model (1) of the main text as

$$y = X\beta + \sum_{k' \neq k}^{p} Z_{k'}\gamma_{k'} + Z_k\gamma_k + \varepsilon \tag{A1}$$

This allows us to obtain

$$\begin{aligned}\operatorname{cov}(y, \gamma_k^{\mathrm{T}}) &= \operatorname{cov}\left(X\beta + \sum_{k' \neq k}^{p} Z_{k'}\gamma_{k'} + Z_k\gamma_k + \varepsilon, \gamma_k^{\mathrm{T}}\right) \\ &= \operatorname{cov}(Z_k\gamma_k, \gamma_k^{\mathrm{T}}) = Z_k\operatorname{var}(\gamma_k) = Z_k\sigma_k^2\end{aligned} \tag{A2}$$

The joint distribution of y and γ_k is multivariate normal with expectation and covariance matrix given below,

$$E\begin{bmatrix} y \\ \gamma_k \end{bmatrix} = \begin{bmatrix} E(y) \\ E(\gamma_k) \end{bmatrix} = \begin{bmatrix} X\beta \\ 0 \end{bmatrix} \tag{A3}$$

and

$$\begin{aligned} \operatorname{var}\begin{bmatrix} y \\ \gamma_k \end{bmatrix} &= \begin{bmatrix} \operatorname{var}(y) & \operatorname{cov}(y, \gamma_k^{\mathrm{T}}) \\ \operatorname{cov}(\gamma_k, y^{\mathrm{T}}) & \operatorname{var}(\gamma_k) \end{bmatrix} \\ &= \begin{bmatrix} V & Z_k\sigma_k^2 \\ Z_k^{\mathrm{T}}\sigma_k^2 & I_{m_k}\sigma_k^2 \end{bmatrix} \end{aligned} \tag{A4}$$

According to the theorem of multivariate normal distribution (Giri, 1996), the conditional distribution of γ_k, given y is multivariate normal with expectation and variance given in the following equations,

$$\begin{aligned} E(\gamma_k|y) &= \operatorname{cov}(\gamma_k, y^{\mathrm{T}})\operatorname{var}^{-1}(y)[y - E(y)] \\ &= Z_k^{\mathrm{T}}\sigma_k^2 V^{-1}(y - X\beta) \end{aligned} \tag{A5}$$

and

$$\begin{aligned} \operatorname{var}(\gamma_k|y) &= \operatorname{var}(\gamma_k) - \operatorname{cov}(\gamma_k, y^{\mathrm{T}})\operatorname{var}^{-1}(y)\operatorname{cov}(y, \gamma_k^{\mathrm{T}}) \\ &= I_{m_k}\sigma_k^2 - \sigma_k^2 Z_k^{\mathrm{T}} V^{-1} Z_k \sigma_k^2 \end{aligned} \tag{A6}$$

These two equations correspond to equations (14) and (15) of the main text, respectively.

Supplementary Information accompanies the paper on Heredity website (http://www.nature.com/hdy)