Bayesian model determination for quantitative trait loci

Jaya M. Satagopan

Department of Epidemiology and Biostatistics

Memorial Sloan–Kettering Cancer Center

1275 York Avenue, New York, NY 10021

e-mail: satago@biosta.mskcc.org

Brian S. Yandell

Department of Statistics

University of Wisconsin

1210 W. Dayton Street, Madison, WI 53706

e-mail:yandell@stat.wisc.edu

September 29, 1998

ABSTRACT

A reversible jump Markov chain Monte Carlo (MCMC) algorithm is illustrated to infer the number of quantitative trait loci (QTL) affecting a phenotypic trait, their chromosomal locations, and their effects. A multi-loci model is fit to quantitative trait and molecular marker data, with the trait response modeled as a linear function of the additive and dominance effects of the unknown QTL genotypes. The number of QTL is unknown and must be estimated as well. Inference summaries for the number of loci, their locations and effects are obtained from the corresponding marginal posterior densities obtained by integrating the likelihood using reversible jump MCMC, rather than by optimizing the joint likelihood surface. Using simulated data and flowering time data from *Brassica napus* we observe that the choice of prior distribution plays an important role in inference. This prior distribution greatly influences how well the chain mixes as well as the posterior distribution of the number of loci. However, the posterior mode of the number of loci is not affected by the choice of its prior. Further, neither the choice of prior nor the starting value for the number of loci affect the estimated chromosomal locations and their effects.

Keywords: Reversible jump Markov chain Monte Carlo; Metropolis-Hastings algorithm; *Brassica napus*.

1. INTRODUCTION

Green (1995) described a reversible jump Markov chain Monte Carlo (MCMC) approach to sample from a distribution of interest when the dimension of the parameter vector is not fixed. Here we demonstrate the utility of this approach to a specific problem of estimating the number of quantitative trait loci (QTL) affecting a trait, a problem of increased research among plant breeders and molecular biologists.

Consider a trait $Y = \{y_i\}_{i=1}^n$ from n individuals, affected by an unknown number of QTL, say s. The effect of the genotypes at these loci on the ith trait value can be described by the following simple linear model:

$$y_i \quad = \quad \mu + \sum_{j=1}^{s} \alpha_j Q_{ij} + \epsilon_i, \tag{1}$$

where μ is the model mean, $Q_i = \{Q_{ij}\}_j^s$ are the genotypes at the s loci for the ith individual, $\alpha = \{\alpha_j\}_{j=1}^s$ are the effects of these s loci, and ϵ_i is a zero mean random deviation with variance σ^2. Specifying a distribution, such as Normal, for the error terms induces a probability distribution on the trait values. In practice, the number of loci (s) and their chromosomal locations are unknown, and the goal of any QTL linkage analysis is to estimate these unknowns. Further, the genotypes at the loci are unobserved. Typically, molecular markers are genotyped and a linkage map is created, providing estimated chromosomal location of these markers and inter-maker genetic distances. Given such molecular marker information, the probability distribution of the unobserved QTL genotypes can be determined easily.

Various statistical methods have been used to estimate s and the chromosomal locations of the QTL. These include EM algorithm to estimate a single locus also known as interval mapping (Lander and Botstein 1989),and multiple regression methods combined with interval mapping to identify multiple loci (Haley and Knott 1992; Zeng 1993, 1994; Jansen 1993; Jansen and Stam 1994) by scanning the region between contiguous pairs of molecular markers along the genome of interest. In all these

methods, the logarithm of odds (LOD) score is examined for any significant evidence for the presence of a QTL in the scanned region. The LOD score is the logarithm (base 10) of likelihood ratios when comparing a null hypothesis of no QTL affecting the trait, versus an alternative hypothesis of one or many loci affecting the trait. The appropriate significance threshold is approximated by the asymptotic distribution of the LOD statistic under the null hypothesis, which is not very easy to obtain. Permutation tests can be used to obtain the appropriate threshold values (Churchill and Doerge 1994; Doerge and Churchill 1996).

Alternatively, Bayesian methods have been used to obtain posterior inference about the location of the putative loci (Thomas and Cortessis 1992; Hoeschele and VanRanden 1993a, 1993b). Recently, Markov chain Monte Carlo (MCMC) methods were used to determine the chromosomal locations of the QTL for a fixed s (Satagopan et al. 1996). After fitting various models with different values of s, Bayes factors (Kass and Raftery 1995) were used to compare these models in order to estimate the number of QTL affecting the trait. However, Bayes factors must be estimated carefully in order to ensure its stability (Newton and Raftery 1994).

Rather than fitting different models by varying s, an alternative approach would be to consider s as a further unknown paramter to be estimated. Satagopan and Yandell (1996) used reversible jump MCMC to fit multiple loci on a single linkage group and to estimate the probability distribution of the number of loci affecting the trait of interest. Sillanpää and Arjas (1998) and Stephens and Fisch (1998) used a similar approach and fit a multi–locus model to F_2 and backcross breeding schemes. The later two approaches considered fitting multiple loci on several chromosomes of the genome of interest.

In this paper, we illustrate a reversible jump MCMC algorithm which uses some of the regression methods of Seber (1977) such as including additional covariates in a regression model. We first propose a stochastic model describing the distribution of the data conditional on the unknown number of QTL and the unknown multiple

QTL genotypes. Standard genetic theory is used to describe the distribution of such unobserved genotypes given genetic parameters and the existence of multiple QTL. As part of our Bayesian analysis, a third level in the hierarchy is a probability distribution over the unknown number of QTL and the genetic parameters. Reversible jump MCMC is used to estimate the marginal posterior distribution of the unknown number of loci, and the marginal posterior distributions of the model parameters conditional upon the number of loci affecting the trait. The number of QTL, their locations and other model parameters are estimated from the respective marginal posterior distributions. Simulated data and flowering time data for *Brassica napus* are used to illustrate the proposed method.

2. QTL MODEL

Consider the simple linear model given by equation (1). At each marker locus and the putative QTL, associate 1 with one homozygous parent type, –1 with the other homozygous parent type and 0 with the heterozygote. In general, the relation between the trait and putative QTL genotypes would be determined by a conditional distribution $\pi(y_i|s, Q_i, \theta)$ with $\theta = (\mu, \alpha, \sigma^2)$ the unknowns. The model could include dominance and epistasis (interaction), and need not be linear nor rely on normality.

In practice, we observe the phenotypic trait Y and a set of marker genotypes M_i but not the QTL genotypes Q_i. Assume that a linkage map has been developed based on m markers with genotypes $M_i = \{M_{ik}\}_{k=1}^m$ for the ith individual, with ordered markers $\{1, 2, \cdots, m\}$. For convenience, suppose this map consists of exactly one linkage group. The markers are assumed to be at known distances $D = \{D_k\}_{k=1}^m$ along the map, with D_k the genetic map distance between markers 1 and k and $D_1 = 0$. Figure 1 illustrates 10 markers on linkage group 9 of *Brassica napus* and the inter–marker distances. Let λ_j be the (unknown) distance of the jth QTL from one end of the linkage group, and $\Lambda = \{\lambda_j\}_{j=1}^s$.

The conditional probability distribution of the QTL genotypes, $\pi(Q_i|s,\Lambda) = \pi(Q_i|s,\Lambda,M_i,D)$, given the number of QTL, their location, the marker genotypes and the distance between the markers, can be modeled in terms of recombination between the loci and the markers. From now on we suppress the notation for conditioning on markers M_i and intermarker distances D. Under the Haldane assumption of independence of recombination events (Ott 1991, pp. 14–19), each QTL genotype Q_{ij} is conditionally independent of nonflanking marker and other QTL genotypes given the flanking marker genotypes. For example, suppose the jth QTL is between markers k and $k+1$. The conditional distribution $\pi(Q_i|s,\Lambda)$ can be written as:

$$\begin{aligned}
\pi(Q_i|s,\Lambda) &= \pi(Q_i|s,\Lambda,M_i,D) \\
&= \prod_{j=1}^{s} \pi(Q_{ij}|\Lambda,M_i,D) \\
&\qquad \text{(assuming the loci segregate independently)} \\
&= \prod_{j=1}^{s} \pi(Q_{ij}|\lambda_j,M_{i,k},M_{i,k+1},D_k,D_{k+1}). \\
&\qquad \text{(under Haldane assumption of independent recombinations)} \qquad (2)
\end{aligned}$$

The marginal likelihood of the parameters s, Λ and θ for the ith individual may be obtained from the joint distribution of traits and QTL genotypes

$$\pi(y_i,Q_i|s,\Lambda,\theta) = \pi(y_i|s,Q_i,\theta)\ \pi(Q_i|s,\Lambda) \qquad (3)$$

by summing over the set of all possible QTL genotypes for the ith individual, $q_i = \{q_{ij}\}_{j=1}^{s} \in \{-1,0,1\}^s$. Therefore,

$$L(s,\Lambda,\theta|y_i) = \sum_{q_i} \pi(y_i,Q_i = q_i|s,\Lambda,\theta)\ . \qquad (4)$$

When the data Y are n independent observations, the marginal likelihood for the trait data is the product over individuals, a familiar mixture model likelihood,

$$L(s,\Lambda,\theta|Y) = \prod_i L(s,\Lambda,\theta|y_i)\ . \qquad (5)$$

Our aim is to make joint inference about the number of QTL, their positions (loci) and the sizes of their effects. The joint likelihood is a mixture of densities, and hence, is difficult to evaluate when there are multiple QTL. Rather than attempt optimization of the likelihood surface, we apply Bayesian analysis and integrate this likelihood, modified by a prior, to produce inference summaries for all the components in the model.

In a Bayesian approach as discussed here we infer the parameters based on their marginal posterior distribution, which can be obtained from the joint posterior given below by integrating over the other unknowns. Exact solution to such high–dimensional integrals are difficult, but Markov chain Monte Carlo (MCMC) approximation is quite feasible. The joint posterior distribution of all the unknowns (s, Λ, Q, θ) is proportional to

$$\pi(s, \Lambda, Q, \theta|Y) \propto \pi(s, \Lambda, \theta) \prod_i \pi(y_i, Q_i|s, \Lambda, \theta) \tag{6}$$

with $Q = \{Q_i\}_{i=1}^n$ the s QTL genotypes for all the n individuals and $\pi(s, \Lambda, \theta)$ a prior density for the model and genetic parameters. We construct a Markov chain with this target distribution resulting in a random sequence of states

$$(s^0, \Lambda^0, Q^0, \theta^0), (s^1, \Lambda^1, Q^1, \theta^1), \cdots (s^N, \Lambda^N, Q^N, \theta^N)$$

starting at an arbitrary point $(s^0, \Lambda^0, Q^0, \theta^0)$ having positive posterior density, and proceeding by simple rules that modify the unknowns s, Λ, Q, and θ. This has to be done carefully since the dimension of (6) changes when s is changed, and standard MCMC theory does not hold in this case. In Section 3, we illustrate the application of reversible jump MCMC to move between models with different number of loci.

2.1 Reversible jump MCMC

This is a random–sweep Metropolis–Hastings algorithm for general state spaces (Richardson and Green 1997) and proceeds as follows. Suppose $x_s = (s, \Lambda, Q, \theta)$ is the current state of the chain indexed by s, the current number of loci. The chain can either move to a "birth" step (where the number of loci increases to $s+1$ from s), or to a "death" step (where the number of loci decreases to $s-1$ from s), or continue with the "current" number (s) of loci. Green (1995) describes these moves and the acceptance probabilities of the "birth" and "death" steps. Suppose the chain moves from a parameter space indexed by s_1 to a space indexed by s_2 (for example, $s_1 = s$, and $s_2 = s+1$). The acceptance probability for the Metropolis–Hastings move is given by the following:

$$\min\left\{1, \frac{\pi(x_{s_1}|y)}{p(x_{s_2}|y)} \frac{\pi_2(u_2)}{\pi_1(u_1)} \left|\left|\frac{\partial(x_{s_2}, u_2)}{\partial(x_{s_1}, u_1)}\right|\right|\right\} \quad . \tag{7}$$

where u_1 and u_2 are random vectors of dimensions d_1 and d_2 with densities π_1 and π_2, respectively, such that $s_1 + d_1 = s_2 + d_2$. The last term in the above acceptance probability is the Jacobian of transformation from the space of dimension s_1 to a space of dimension s_2.

3. A REVERSIBLE JUMP ALGORITHM FOR MULTIPLE QTL

Let $s_{\max}$ be the maximum allowed value for s, the number of loci. The above reversible jump algorithm can be readily adapted to MCMC inference for multiple QTL by considering moves between different models, updating s as in the following steps:

1. a birth step which can increase the number of QTL from s to $s+1$, with probability $b_s = c \times \min\{1, \pi(s+1)/\pi(s)\}$;

2. a death step which can decrease the number of QTL from $s+1$ to s, with probability $d_s = c \times \min\{1, \pi(s)/\pi(s+1)\}$; and

3. updating the genetic and model parameters, and the QTL genotypes without changing s, with probability $1 - b_s - d_s$;

where c is a uniform random number from $(0, 0.5)$. Hence, $b_s + d_s < 1$ so that step 3 does not have a zero probability. The probabilities b_s and d_s are constrained such that $d_0 = 0$ and $b_{s_{\max}} = 0$. Steps 1 and 2 can change the dimension of the parameter space. We use a hybrid sampler (Tierney 1994) to randomly choose one of the above three steps at each transition of the chain.

Assume prior independence of Λ and θ given s. Therefore,

$$\pi(s, \Lambda, \theta) = \pi(\Lambda|s)\pi(\theta|s)\pi(s).$$

A natural choice for prior of Λ (given s), when no information regarding the locations is available, is the uniform distribution for s ordered variables on $[0, D_m]$. Specifying a conjugate prior for μ, α, and σ^2 makes its form simple while increasing diffuseness makes the prior objective. The prior on s, $\pi(s)$, could be Poisson or Uniform$(0, s_{\max})$, for some suitable Poisson mean or $s_{\max}$.

The steps updating Λ, Q and θ for a given s are described in detail in Satagopan et al. (1996). In this section we focus on birth and death type moves for updating the number of QTL (s). More specifically, given the current state (s, Λ, Q, θ), we proceed to sample s for the next state as follows.

The birth and death steps involve adding a locus ($s \to s+1$) or dropping a locus ($s \to s-1$), with subsequent rescaling of the QTL effects. The model given by equation (1) for a fixed s can be rewritten in matrix form as

$$\begin{aligned} Y &= 1\ \mu + \sum_{j=1}^{s} Q_j\ \alpha_j + \epsilon \\ &= X\beta + \epsilon\ , \end{aligned} \tag{8}$$

where $\beta^T = (\mu\ \alpha_1\ \alpha_2\ \cdots\ \alpha_s)$ is a column vector of model mean and QTL effects, and hence the model parameters are $\theta = (\beta, \sigma^2)$. $X = (1\ Q_1\ \cdots\ Q_s)$ is the $n \times (s+1)$ design matrix with the first column all ones, corresponding to the model mean μ, and the other columns correspond to the s QTL genotypes. It is computationally convenient to set up the Cholesky decomposition (Anderson et al. 1995) for the design matrix $X = FG$ in which $F = (F_1 : F_2)$ is orthogonal and G is upper triangular.

3.1 Birth Step

The birth step involves proposing a new QTL, its genotype and the corresponding effect. Denote the proposed parameters of the birth step as $(s+1, \lambda_{s+1}, Q_{s+1}, \alpha_{s+1})$. Hence, the birth step considers the following model:

$$Y \;=\; X\beta^* + \alpha_{s+1}Q_{s+1} + \epsilon\ , \tag{9}$$

when the current model is given by equation (8).

B1. Choose an interval for birth not containing any other QTL, with probability $1/(m-1-s)$. Here, $m-1$ is the total number of intermarker intervals, and s is the number of QTL in the model before the birth step.

B2. Suppose we choose the interval between markers k and $k+1$. Choose a locus λ_{s+1} in this interval uniformly between (D_k, D_{k+1}) with probability $1/(D_{k+1} - D_k)$.

B3. Sample the QTL genotypes for this new locus according to

$$\pi(Q_{i,s+1}|\lambda_{s+1}, M_{ik}, M_{i,k+1}, D_k, D_{k+1}), \qquad i = 1, \cdots, n\ .$$

M_{ik} and $M_{i,k+1}$ are the flanking marker genotypes for the new QTL. These probabilities can be obtained in terms of recombinations between the QTL and the flanking markers (Knapp, Bridges, and Birkes 1990).

B4. In order to sample the new QTL effect α_{s+1}, first obtain U, a random number from a standard normal distribution. The scalar $V = Q_{s+1}^T F_1 F_1^T Q_{s+1} = ||F_1^T Q_{s+1}||^2$ is used to reweight the new QTL effect as

$$\alpha_{s+1} = V^{-1} Q_{s+1}^T F_1 F_1^T Y + \sigma V^{-1/2} U \ .$$

B5. Modify the regression parameters β to get new parameters β^* as

$$\beta^* \ = \ \beta - G^{-1} F_1^T Q_{s+1} \alpha_{s+1} \ .$$

If we allow for multiple QTL to be present between a pair of flanking markers, the QTL probability distribution in step B3 can be calculated conditional upon the genotypes of other QTL in that marker interval (details are not presented here). The proposal probability for the birth step based on the above updating scheme is given by

$$q_b = \frac{1}{m-s-1} \times \frac{1}{D_{k+1} - D_k} \times \pi(Q_{s+1}|\lambda_{s+1}) \times \pi(U) \ . \tag{10}$$

3.2 Death Step

For notational convenience, assume that the death step attempts to move from a model with $s+1$ loci to s loci. (If necessary, renumber the loci so this is the case). The death step considers the model given by equation (8), when the current model is as in equation (9). The proposal for the death step is uniform over all $s+1$ loci,

$$q_d \ = \ 1/(s+1) \ . \tag{11}$$

The death step to move from $s+1$ to s loci proceeds as follows:

D1. Choose one of the $s+1$ loci with probability q_d. This reduces the number of loci from $s+1$ to s. Let α_{s+1} and Q_{s+1} correspond to the effect and QTL genotypes of the locus to be dropped.

D2. Drop the effect α_{s+1}. Modify the regression parameters β of the smaller model to

$$\beta + G^{-1}F_1^T Q_{s+1}\alpha_{s+1} \ ,$$

with $X = F_1 G$ for the design matrix of the remaining s QTL genotypes.

D3. Drop the corresponding QTL genotypes Q_{s+1}.

The idea for updating the model mean and QTL effects in the birth and death steps (**B4, B5**, and **D2**) is similar to that of introducing additional regression parameters (Seber 1977, section 3.7; Mallick 1995). This adjustment to the model parameters is done to obtain the best fit in the new model subspace.

3.3 Acceptance Probability

The acceptance probability for the birth and death steps are $\min(1, A)$ and $\min(1, A^{-1})$, respectively where

$$A = \frac{\pi(s+1, \{\Lambda, \lambda_{s+1}\}, \{Q, Q_{s+1}\}, (\theta, \alpha_{s+1})|Y)}{\pi(s, \Lambda, Q, \theta|Y)} \times \frac{d_{s+1}}{b_s} \ \frac{q_d}{q_b} \ \frac{V^{1/2}}{\sigma} \ . \tag{12}$$

The first term on the right hand side corresponds to the ratio of the posteriors in the larger model (9) with $s+1$ QTL and the smaller model (8) with s QTL. The other terms are the ratio of the probability of death and birth moves, the ratio of death and birth proposals, and the Jacobian of transformation from smaller to larger models. This Jacobian is derived in the Appendix. For a move which does not involve the change of parameter dimension, the acceptance probability is a Metropolis–Hastings acceptance probability based simply on the ratio of densities.

3.4 Inference

The sampled states of the reversible jump Markov chain can be used to obtain inference summaries about the parameters of interest. The frequencies of the sampled values of s gives an estimate of its marginal posterior density. An estimate of s, such as mode, can be obtained from this marginal posterior density. Inference for other paramerers can be obtained conditional upon a given value of s. For example, the empirical averages of the sampled values of Λ for a given s is an estimate of the QTL locations. Confidence intervals can be given by high posterior density (HPD) regions obtained from the corresponding marginal posterior densities (Box and Tiao 1973).

4. RESULTS

The proposed algorithm for reversible jump MCMC is illustrated using simulated data and using flowering time data for *Brassica napus*. Sensitivity to the choice of prior must be examined in a Bayesian analysis. Here we examine sensitivity to the choice of QTL prior mean. We first present the analysis of simulated data, followed by the analysis of flowering time data. Sensitivity to the choice of QTL prior mean is discussed for both these analyses.

4.1 Simulated Data

Trait data for two QTL were generated under models inspired by the *Brassica napus* study detailed in section , with 105 individuals and 10 molecular markers on linkage group 9 of this genus (Figure 1). Two QTL were placed, one between markers 4 and 5 at 30.8cM, and the second between markers 8 and 9 at 66.7cM. Data on 105 independent individuals (number similar to flowering time data) were generated using the linear model of equation (1), with the error term having a $N(0, 1)$ distribution.

The effects of the two loci were $(\alpha_1, \alpha_2) = (3, 3)$ and the model mean $\mu = 0$, in the same units as the trait data.

The prior distributions for the reversible jump analysis were chosen as follows. The prior on s was taken as Poisson with mean 4, restricted to the range $0, 1, \cdots, 10$. The prior on model mean μ, and the effects α were normal centered at 0 and variance 10, allowing the possibility of extreme QTL effects. The QTL positions Λ had an ordered uniform prior on the entire linkage group 9. The model variance σ^2 had an inverse gamma(2,2) prior. The starting values for the reversible jump MCMC analysis were as follows. $s^0 = 1$, and the starting value for a single putative locus (λ_1) was at the center of the linkage group between markers 5 and 6. The starting values μ^0 and α_1^0 were 0, and $\sigma_0^2 = 0.5$.

After a burn–in period of 100,000 sweeps, 500,000 runs of reversible jump MCMC were obtained. These values were sampled at every 200th cycle, giving a working set of 2,500 samples. The mode of the marginal posterior distribution of s was at 2 with probability 0.42. Other plausible values for the number of loci were $s = 3$ with marginal posterior probability 0.39, and $s = 4$ with probability 0.15. In all the runs, the number of QTL (s) never exceeded 6. Corresponding to the sampled values of $s = 2$, the two loci were estimated at 31cM and 67cm with corresponding effects 3.06 and 2.92. The estimated model mean μ and model variance σ^2 were -0.03 and 1.17, respectively. The overall features of the true model were recovered in this analysis. The effects of the two loci are "large" in this simulation. Therefore, it may not be surprising that all the features of the simulated data were recovered in the reversible jump MCMC sampling. We were interested in investigating whether similar results can be obtained when the underlying QTL have very small effects.

A second set of data were generated based on the results of the flowering time data analysis in Satagopan et al. (1996). Two QTL were simulated between markers 5 and 6 at 40cM, and between markers 9 and 10 at 76cM. The effects of the two loci were -0.06 and -0.12, respectively, in the same units as the trait data. The model

mean mean μ was 3.0. The model variance σ^2 was 0.1. We first considered a Poisson distribution with mean 5 for the number of loci s, restricted to the range $0, \cdots, 10$. Prior for the other unknown parameters were the same as for the above simulation. Starting values for the reversible jump MCMC run were the same as for the above simulation.

Table 1 gives the estimated posterior distribuition of s for this simulation. Under a QTL prior mean of 5, the marginal posterior of $s = 2$ has the highest probability, although other values for s are plausible. Figure 2 gives the estimated marginal posterior distributions of the two loci conditional upon $s = 2$. The estimated QTL positions and 95% HPD confidence intervals are also indicated. The estimated positions are very close to the true loci and the HPD confidence intervals include the true positions as well. Sampled values of $s = 1$ correspond to sampling a single QTL position. In this case, if the true number of loci is 2, one would expect the marginal posterior distribution of the single locus to be multi–modal. Figure 3 gives the estimated marginal posterior distribution of a single QTL conditional upon $s = 1$. This distribution has 3 modes, one between markers 5 and 6, a second mode between markers 9 and 10, and a third mode occuring in the interval between markers 8 and 9. The two intervals between markers 7 and 9 are sampled frequently when $s = 1$. This is a common feature known as "ghosting" which occurs while fitting a single locus model when the true underlying model involves multiple loci affecting the trait, and was previously observed by Haley and Knott (1992). Figure 4 gives the estimated marginal posterior distribution of the three loci conditional upon $s = 3$. The estimated positions of locus 2 and locus 3 are very close to the simulated data, and are close to the estimated values of the loci conditional upon $s = 2$. Locus 1 is estimated to lie between markers 3 and 4. One may then want to determine whether there are indeed 3 loci affecting the trait. We can examine this by looking at the HPD confidence intervals. The 95% HPD region of locus 1 spans two regions, one between markers 1 and 2, and a second region between markers 3 and 7. This second region is

contained within the 95% HPD interval for locus 2. This, coupled with the fact that the posterior probability of $s = 3$ is lower than the probability that $s = 2$, indicates that a two loci model is more likely than a three loci model.

4.2 Sensitivity Analysis

In a Bayesian analysis, as discussed in this paper, it is important to check for sensitivity to the choice of prior distributions. Satagopan (1995) studied sensitivity to the choice of prior for an MCMC approach to fit multiple QTL model when s is fixed. The results of the analysis were not affected greatly by the choice of prior distributions and hyperparameters. Here we discuss sensitivity to the choice of prior distribution of s. The choice of QTL prior mean and how well the sampled Markov chain "mixes" appear to interact closely.

A Poisson prior was chosen for s throughout our analysis. The estimated marginal posterior distribution of s was affected by the choice of mean (and, hence, variance) for this Poisson prior. The prior distribution of s enters the acceptance ratio (equation 12) calculation through the ratio d_{s+1}/b_s and thus influences the "mixing" of the chain. Increasing the prior mean (or variance) of s tends to increase its posterior variance. Marginal posterior of s for the second set of simulations for various choices of QTL prior means are given in Table 2. In this case, the chain did not mix well and continued to stay at smaller values of s (0 and 1) for a long time when the prior mean was chosen to be small (< 3). For example, when the QTL prior mean was 1 the following marginal posterior probabilities were observed for various starting values of s for the second set of simulations: a no QTL model had a marginal posterior probability of 0.38, a one QTL model had a probability of 0.53, and a two QTL model had a probability of 0.08. The chain did not sample loci larger than 3. The chain moved between different models more frequently when the prior mean was larger than 3. Increasing the prior mean seemed to favor a higher number of loci. As discussed

in the previous section, the loci estimated in the model of a specific dimension (or, a given value of s) were also the ones estimated in models of other dimensions (Figures 2, 3, and 4).

Bayes factor comparing a model with s_1 loci against a model with s_2 loci is defined as

$$\begin{aligned} B(s_1, s_2) &= \frac{\pi(y|s_1)}{\pi(y|s_2)} \\ &= \frac{\pi(s_1|y)}{\pi(s_2|y)} \times \frac{\pi(s_2)}{\pi(s_1)} \\ &\quad \text{(by Bayes theorem)} . \end{aligned} \tag{13}$$

We calculated Bayes factors for the second set of simulations. The value of B(2,3), Bayes factor comparing models with 2 and 3 loci, for QTL prior mean 3, 4, 5, and 6 are 3.17, 2.73, 3.18, and 2.67, respectively. These values are fairly similar and do not seem to be affected by the choice of QTL prior mean. Similarly, B(1,2), Bayes factor comparing models with 1 and 2 loci, for QTL prior means 3, 4, 5, and 6 are 1.74, 1.51, 1.48, and 1.35, respectively. The results were not sensitive to choices of starting values for s and for other parameters.

4.3 Analysis of flowering time data

The *Brassica* genus has been widely studied for disease resistance, freezing tolerance, flowering time and seed oil content, among various other traits of economic importance. Here we analyze double haploid (DH) progeny from *Brassica napus* to detect QTLs for flowering time. A DH line from the *Brassica napus* cv. Stellar (an annual canola cultivar) was crossed to a single plant of cv. Major (a biennial rapeseed cultivar) which was used as a female. One hundred and five DH lines, the F_1 hybrid and progeny from self-pollination of the parents Major and Stellar were evaluated in the field for flower initiation. The plants were divided into 3 groups and each group

was exposed to one of the 3 treatments – no vernalization, 4 weeks vernalization and 8 weeks vernalization. Materials and methods and preliminary analysis of the experiment are given in Ferreira et al. (1995). DNA extraction and linkage map construction are described in Ferreira, Williams, and Osborn (1995). To illustrate reversible jump MCMC, we consider only flowering data for 105 progeny from 8 weeks vernalization treatment and genotypes of 10 markers from linkage group 9. One out of 105 phenotypic data was missing and 9% of the marker genotypes were missing. Figure 5 shows the two LOD peaks (LOD values = 8.37 and 6.91) on linkage group 9 obtained using the EM algorithm for a single QTL model (Lander and Botstein 1989). Fixing a QTL at the higher peak showed an increase in the LOD score of 1.72 for a second putative QTL. Fitting single, two and 3 QTL models using the Bayesian approach and comparing them using Bayes factors showed that a two QTL model best fit the data (Satagopan et al. 1996). We further investigate this using the above reversible jump MCMC algorithm.

We use the simple linear model given by equation (1) for the number days to flower for the ith DH line, where y_i is logarithm of the number of days to flower, and μ, α_j and Q_{ij} are defined as earlier. Note that since the DH lines are homozygous at every locus, $Q_{ij} \in \{-1, 1\}$. The random errors ϵ_i are assumed to have independent Gaussian distributions with mean 0 and common variance σ^2.

The Bayesian formulation of the problem requires specification of prior distribution on the set of model parameters $\theta = (\mu, \alpha, \sigma^2)$ and the loci Λ. For simplicity we assume prior independence of the model parameters. The overall mean μ, and genetic effects α are given independent Gaussian prior centered at 0 and variance 10 allowing for the possibility of extreme QTL effects. The phenotypic variance σ^2 is assumed to have an inverse gamma(2,2) prior. The λ_j's are assumed to have uniform prior as described earlier. The number of QTL s has a Poisson prior with mean 5. We set $s_{\max} = 10$.

The starting values are chosen as follows. $s^0 = 1$, the single starting locus λ_1^0

is at the center of linkage group 9. Model mean μ^0, and the single QTL effect α_1^0 are 0. The variance σ_0^2 is 0.5. After an initial burn–in of 100,000 states, 1,000,000 states were sampled at every 200th cycle giving a working set of 5,000 states. Table 1 gives the estimated marginal posterior distribution of s. The estimated mode of this distribution is at $s = 2$. Figures 6A and 6B give the estimated posterior distributions of the two loci conditional upon $s = 2$. Locus 1 is estimated at 34.3cM near marker 5. Locus 2 is estimated at 71.1cM between markers 9 and 10. The estimated effects of the two loci are -0.05 and -0.13 respectively. The estimated model mean is 3.06 and the estimated model variance is 0.08. Figure 6C shows the estimated posterior distribution of a single locus conditional on $s = 1$. This distribution has 3 modes, two around the same regions as the loci estimated when $s = 2$, and a third one in between these two modes. The third mode corresponds to a ghost QTL sampled under a one QTL model when the trait is affected by multiple loci. The 95% HPD confidence region for a single locus spans two regions, one between markers 5 and 7, and a second region between markers 7 and 10. These regions are also included in the confidence intervals of the two loci sampled under a two QTL model (Figures 6A and 6B). Based on these plots and the posterior distirbution of s, a two QTL model is more likely although other models (such as three loci) are plausible. Using Bayes factors as a model selection criterion, Satagopan et al. (1996) observed that a two loci model was more likely for the flowering time data. Butruille (1998) conduted breeding experiments focusing on linkage group 9 (now called N2) and found evidence to support two linked QTL. This work is being extended to fine mapping of the region, exploiting synteny with *Arabidopsis thalianus* to locate the candidate gene.

Sensitivity to the choice of QTL prior mean was examined for the flowering time data. Results were similar to those observed in the second set of simulations. The posterior probability of $s = 2$ was the highest for various choices of prior mean. A very small prior mean resulted in poor mixing of the chain. Increasing the prior mean also increased the variance of the estimated marginal posterior of s. The results were

not affected by the choice of starting values, nor were they affected by the choice of prior for the other unknown parameters.

5. DISCUSSION

One of the goals of any linkage study is to determine the number of loci affecting the trait of interest. One can estimate this quantity by either fitting a single locus model and including additional loci in a step–wise manner (Lander and Botstein 1989), or by considering this as a model selection problem by comparing single versus multi–locus models (Satagopan et al. 1996; Hoeschele and VanRanden 1993b), or by considering the unknown number of loci as an additional parameter to be estimated as presented here. An attractive feature of the reversible jump MCMC approach is that we can obtain an estimate of the probability distribution of the number of loci in the form of marginal posterior distribution of s. For the simulated data and flowering time, we have considered a simple linear model without dominance or epistasis. These terms can be easily incorporated into the model. Epistasis, which denotes interaction between two genetic loci, can be included as an interaction term in the model. Although we have considered analysis of only one linkage group, this method can be easily extended to include multiple chromosomes. This can be done by including an additional step in the algorithm which will choose a chromosome before choosing an interval for a birth or a death step within that chromosome. Here we have demonstrated application to a double haploid progeny which is homozygous at every genetic locus. Other breeding schemes can also be considered. For the flowering time data analysis, 1,000,000 iterations of the chain were sampled at every 200th cycle. The stored samples were examined for serial correlation and convergence.

Satagopan and Yandell (1996) gave a reversible jump MCMC approach for multiple loci on a single linkage group. Later Sillanpää and Arjas (1998) and Stephens and Fisch (1998) used a similar approach for multiple loci on different linkage groups.

This paper modifies the method of Satagopan and Yandell (1996) by updating the model mean and effects within the birth and death steps (steps B4, B5 and D2) to provide a good fit for the data conditional upon the number of loci and the genotypes at the estimated locations. This idea is similar to modifying the regression parameters when additional covariates are included in the model (Seber 1977). Further, by updating the model parameters as in steps B4, B5 and D2, we achieved better mixing of the chain than in our earlier work. The advantage of updating the parameters as in the earlier three works on reversible jump approach is that the Jacobian is very easy to evaluate (in fact, the Jacobian is 1). However, the Jacobian must be evaluated carefully to implement the method presented here.

Reversible jump MCMC method is computationally intensive. But given the speed of modern computers it is feasible to implement this approach. However, the choice of priors, particularly prior of s, must be carefully assessed in every application. The prior mean of s plays an important role in how well the chain mixes. Further, the posterior variance of s also depends on its prior mean (which is the same as the prior variance). It remains to be investigated whether very long runs of the chain are required for small prior means of s. This prior mean appears in the acceptance probability of the birth (and death) step as the ratio of birth and death proposals. In our analyses of flowering time data, between 14% and 33% of the proposed birth/death type moves were accepted for various choices of prior mean of s. Another choice of prior for s is Uniform over $\{0, 1, \cdots, s_{\max}\}$. The ratio of birth and death proposals under this prior would be 1. The role of $s_{\max}$ in mixing of the sampled chain must be investigated.

ACKNOWLEDGEMENT

The authors would like to thank Professor Thomas C. Osborn, Department of Agronomy, University of Wisconsin, for providing the *Brassica napus* flowering time data. Brian S. Yandell was supported in part by a USDA/CSREES Hatch grant.

APPENDIX: JACOBIAN FOR THE BIRTH STEP

The birth step involves changing the parameters from (s, Λ, Q, β) to

$$\left(s+1, \{\Lambda, \lambda_{s+1}\}, \{Q, Q_{s+1}\}, \{\beta^*, \alpha_{s+1}\}\right) \ .$$

The contribution to the Jacobian comes only from transforming β in the smaller model to (β^*, α_{s+1}) in the larger model as described in section in steps **B4** and **B5**. Hence, the Jacobian is given by

$$\begin{aligned} J &= \det \begin{pmatrix} \partial\beta/\partial\beta^* & \partial\beta/\partial\alpha_{s+1} \\ \partial U/\partial\beta^* & \partial U/\partial\alpha_{s+1} \end{pmatrix} \\ &= \frac{\partial(\beta, U)}{\partial(\beta^*, \alpha_{s+1})} \\ &= \det \begin{pmatrix} I_{s+1} & \partial\beta/\partial\alpha_{s+1} \\ 0_{s+1}^T & (1/\sigma)V^{1/2} \end{pmatrix} \\ &= (1/\sigma)V^{1/2} \quad \square \end{aligned}$$

Here I_{s+1} is an identity matrix of dimension $s+1$, $\partial\beta/\partial\alpha_{s+1}$ is a column vector of length $s+1$, and 0_{s+1} is a column vector of 0s of length $s+1$. The Jacobian is the product of the diagonal elements since the matrix is upper triangular.

References

Anderson, E., Z. Bai, C. Bischof, J. Demmel, J. Dongarra, et al. (1995). *LAPACK Users' Guide* (Second ed.). Society for Industrial and Applied Mathematics, Philadelphia.

Box, G. E. and G. C. Tiao (1973). *Bayesian Inference in Statistical Analysis*. John Wiley and Sons, New York.

Butruille, D. V. (1998). *Introgression of winter germplasm and use of molecular markers to document its effects on agronomic traits of spring inbreds and hybrids of* Brassica napus. Ph. D. thesis, University of Wisconsin. Department of Agronomy.

Churchill, G. A. and R. W. Doerge (1994). Empirical threshold values for quantitative trait mapping. *Genetics 138*, 963–971.

Doerge, R. W. and G. A. Churchill (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics 142*, 285–294.

Ferreira, M. E., J. M. Satagopan, B. S. Yandell, P. H. Williams, and T. C. Osborn (1995). Mapping loci controlling vernalization requirement and flowering time in *brassica napus*. *Theoretical and Applied Genetics 90*, 727–732.

Ferreira, M. E., P. H. Williams, and T. C. Osborn (1995). RFLP mapping of Brassica napus using F1-derived doubled haploid lines. *Theoretical and Applied Genetics 89*, 615–621.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika 82*, 711–732.

Haley, C. S. and S. A. Knott (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity 69*, 315–324.

Hoeschele, I. and P. VanRanden (1993a). Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. *Theoretical and Applied Genetics 85*, 953–960.

Hoeschele, I. and P. VanRanden (1993b). Bayesian analysis of linkage between genetics markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. *Theoretical and Applied Genetics 85*, 946–952.

Jansen, R. C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics 135*, 205–211.

Jansen, R. C. and P. Stam (1994). High resolution of quantitative traits into multiple loci via interval mapping. *Genetics 136*, 1447–1455.

Kass, R. E. and A. E. Raftery (1995, March). Bayes factors and model uncertainty. Technical Report 571, Carnegie Mellon University.

Knapp, S. J., W. C. Bridges, and D. Birkes (1990). Mapping quantitative trait loci using molecular marker linkage maps. *Theoretical and Applied Genetics 79*, 583–592.

Lander, E. S. and D. Botstein (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics 121*, 185–199.

Mallick, B. K. (1995). Bayesian curve estimation by polynomials of random order. Technical Report 95–19, Department of Mathematics, Imperial College.

Newton, M. A. and A. E. Raftery (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society B 56*, 3–48.

Ott, J. (1991). *Analysis of Human Genetic Linkage* (Revised ed.). The Johns Hopkins University Press, Baltimore.

Richardson, S. and P. J. Green (1997). On Bayesian analysis of mixture with an unknown number of components. *Journal of the Royal Statistical Society, Series B 59*, 731–792.

Satagopan, J. M. (1995). *A Markov chain Monte Carlo approach to detect polygene loci for complex traits*. Ph. D. thesis, University of Wisconsin.

Satagopan, J. M. and B. S. Yandell (1996). Estimating the number of quantitative trait loci via Bayesian model determination. In *Proceedings of the Section on Biometrics*, Joint Statistical Meetings, Chicago.

Satagopan, J. M., B. S. Yandell, M. A. Newton, and T. C. Osborn (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics 144*, 805–816.

Seber, G. A. (1977). *Linear Regression Analysis*. John Wiley and Sons, New York.

Sillanpää, M. J. and E. Arjas (1998). Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. *Genetics 148*, 1373–1388.

Stephens, D. A. and R. D. Fisch (1998). Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. *Biometrics 00*, 000–000. (To appear).

Thomas, D. C. and V. Cortessis (1992). A Gibbs sampling approach to linkage analysis. *Human Heredity 42*, 63–76.

Tierney, L. (1994). Exploring posterior distributions using Markov chains (with discussion). *Annals of Statistics 22*, 1701–1762.

Zeng, Z. B. (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *PNAS USA 90*, 10972–10976.

Zeng, Z. B. (1994). Precision mapping of quantitative trait loci. *Genetics 136*, 1457–1468.

Table 1: Estimated posterior probability of the number of loci for 2 data sets. Data 1: Simulated data with "small" QTL effects where $(\alpha_1, \alpha_2) = (-0.06, -0.12)$, and Data 2: flowering time data for *Brassica napus*. Probabilities are given for various choices of QTL prior means.

Data		QTL prior mean			
Set	s	3	4	5	6
1	0	0.050	0.020	0.000	0.000
	1	0.440	0.310	0.260	0.180
	2	0.380	0.410	0.440	0.400
	3	0.120	0.200	0.230	0.300
	4	0.020	0.050	0.060	0.100
	5	0.000	0.010	0.010	0.020
	6	0.000	0.000	0.000	0.000
2	0	0.000	0.000	0.000	0.000
	1	0.430	0.280	0.210	0.120
	2	0.430	0.440	0.410	0.340
	3	0.120	0.220	0.280	0.330
	4	0.020	0.050	0.090	0.160
	5	0.000	0.000	0.010	0.040
	6	0.000	0.000	0.000	0.005

Figure Captions

Figure 1: Linkage group 9 of *Brassica napus* with 10 markers positioned according to their genetic distances. The centiMorgan distance between consecutive pairs of markers are shown on the right.

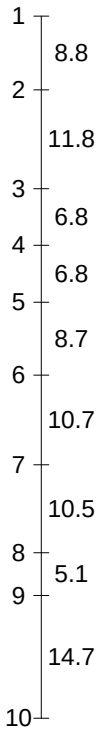
Figure 2: Histogram of estimated marginal posterior distributions of two loci conditional upon $s = 2$ for the simulated data with "small" QTL effects. The X denotes the estimated locations. 95% HPD intervals are indicated by the parantheses.

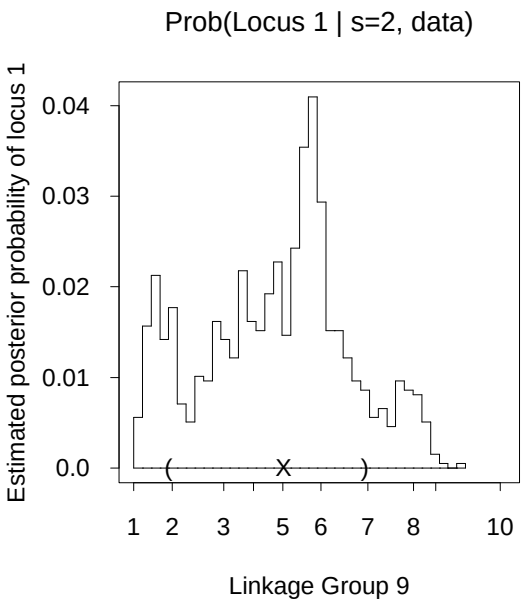
Figure 3: Histogram of estimated marginal posterior distribution of a single locus conditional upon $s = 1$ for the simulated data with "small" QTL effects.

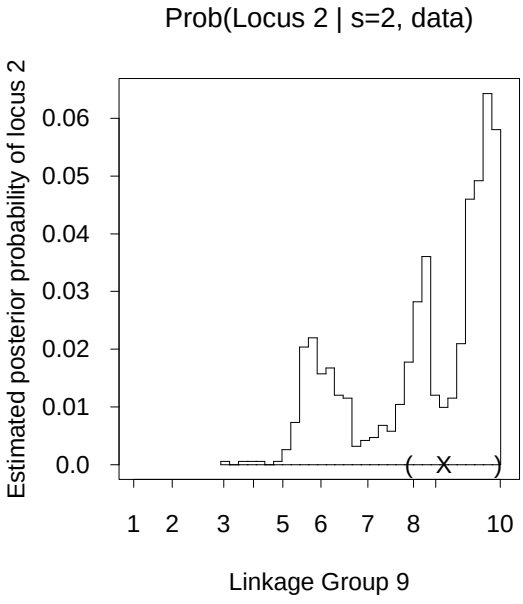
Figure 4: Histogram of estimated marginal posterior distribution of 3 loci conditional upon $s = 3$ for the simulated data with "small" QTL effects. X denotes the estimated loci. 95% HPD intervals are indicated by parantheses.

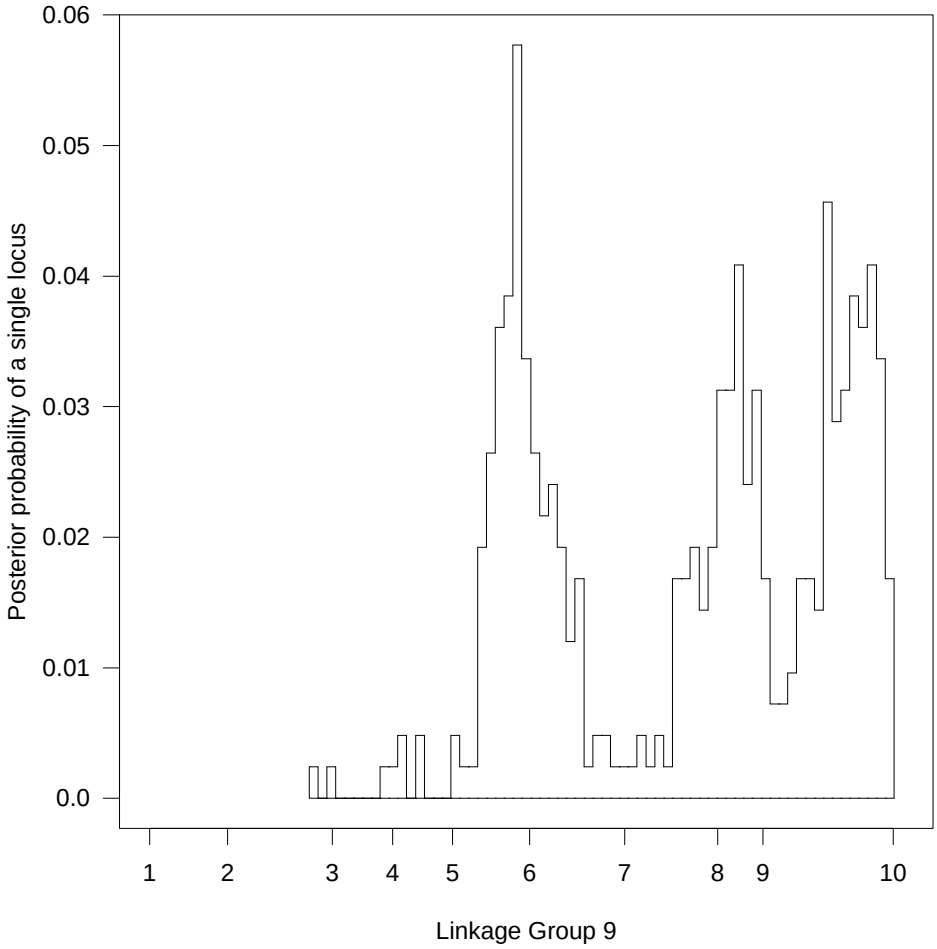
Figure 5: LOD score using EM algorithm for a single QTL model. The horizontal axis is linkage group 9 with all 10 markers positioned according to their genetic distances. The vertical axis is the LOD score. The horizontal line at LOD = 3 corresponds to the conventional LOD cut-off.

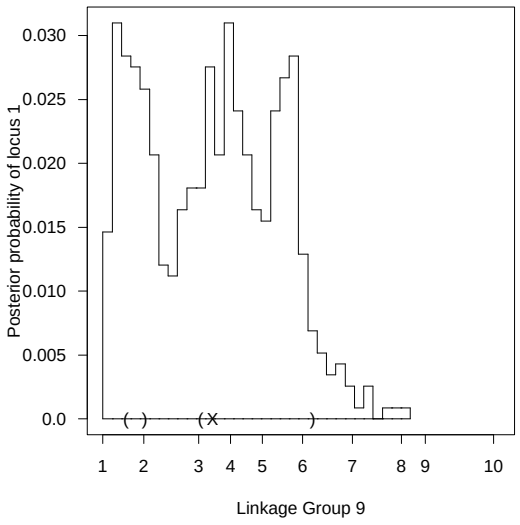
Figure 6: Estimated posterior distributions of the locations for *Brassica napus* flowering time data. Figures A and B give the posterior distribution for locus 1 and locus 2 conditional upon a two loci model. Figure C gives the posterior distribution of a single locus conditional upon a one locus model. X denotes the estimated loci. 95% HPD intervals are indicated by parantheses.

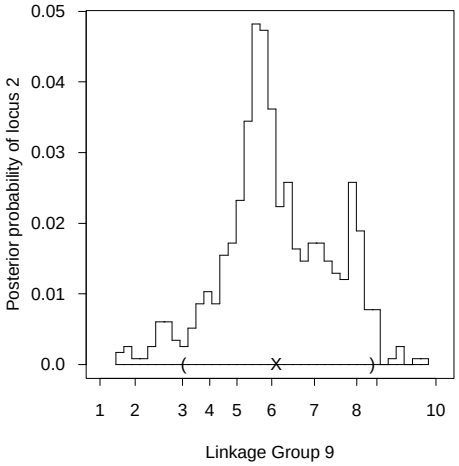

1
8.8
2
11.8
3
6.8
4
6.8
5
8.7
6
10.7
7
10.5
8
5.1
9
14.7
10

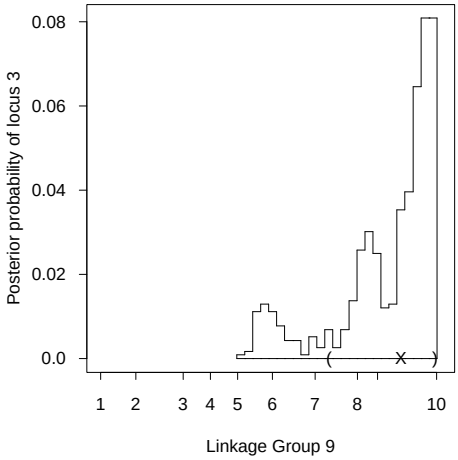

Prob(Locus 1 | s=2, data)
Estimated posterior probability of locus 1
0.0
0.01
0.02
0.03
0.04
1
2
3
5
6
7
8
10
Linkage Group 9

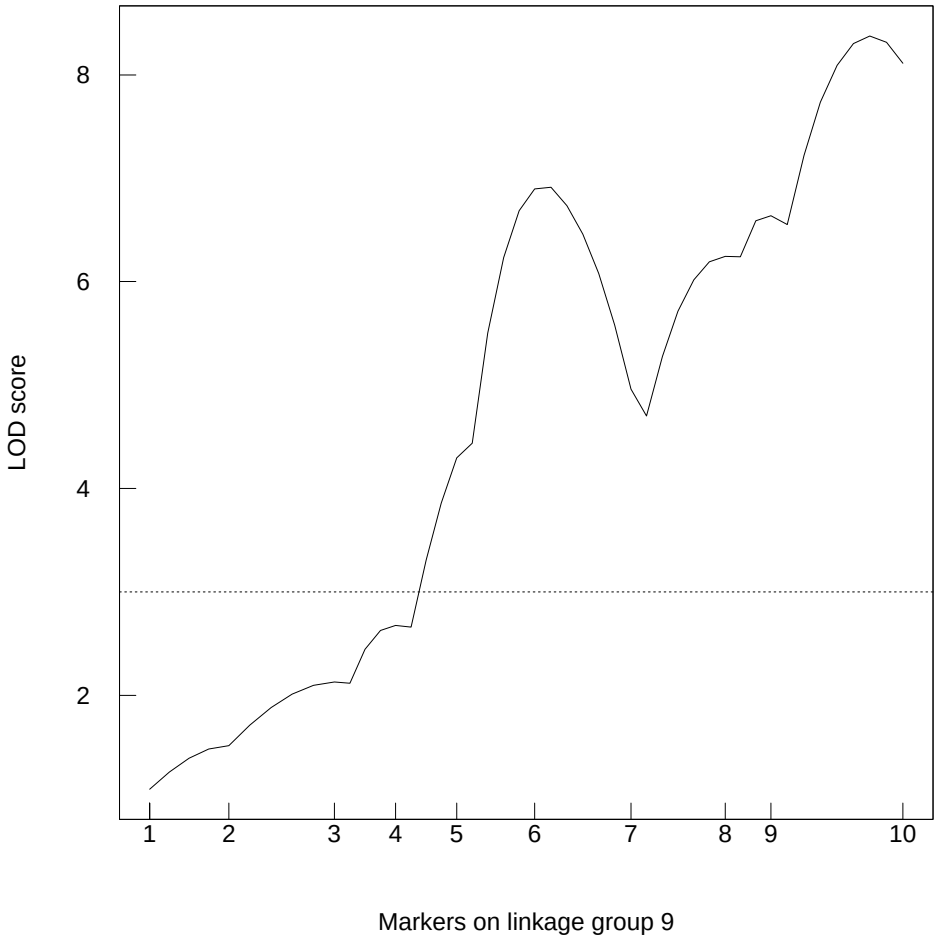

Prob(Locus 2 | s=2, data)
Estimated posterior probability of locus 2
0.0
0.01
0.02
0.03
0.04
0.05
0.06
1
2
3
5
6
7
8
10
Linkage Group 9

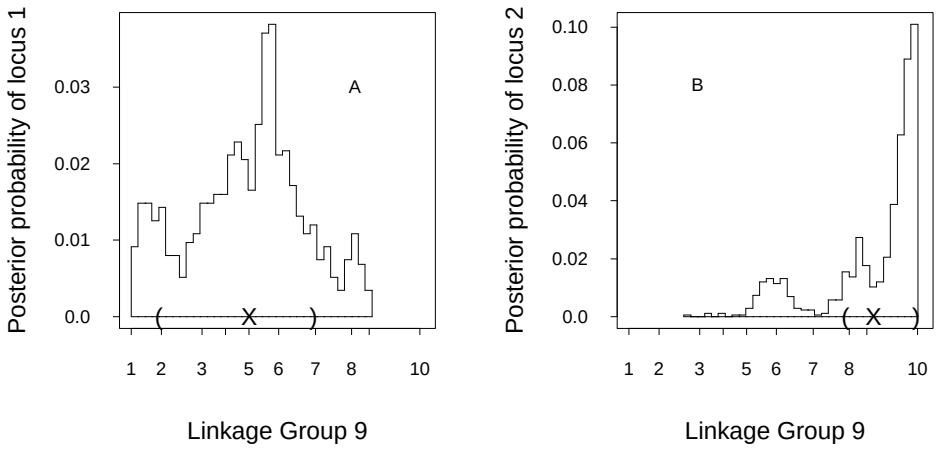

0.06
0.05
0.04
0.03
0.02
0.01
0.0
Posterior probability of a single locus
1
2
3
4
5
6
7
8
9
10
Linkage Group 9

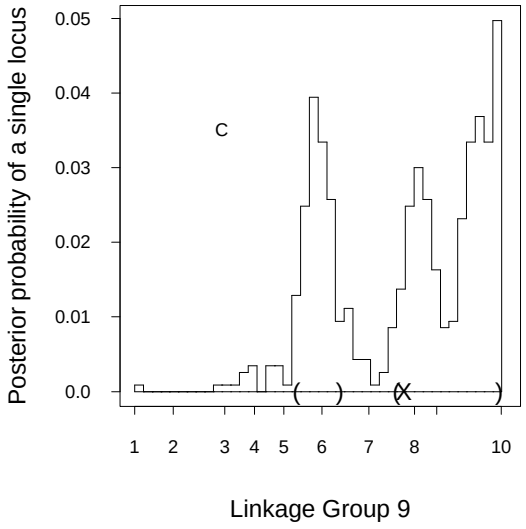

Posterior probability of locus 1
0.030
0.025
0.020
0.015
0.010
0.005
0.0
1 2 3 4 5 6 7 8 9 10
Linkage Group 9


Posterior probability of locus 2
0.05
0.04
0.03
0.02
0.01
0.0
1 2 3 4 5 6 7 8 10
Linkage Group 9


Posterior probability of locus 3
0.08
0.06
0.04
0.02
0.0
1 2 3 4 5 6 7 8 10
Linkage Group 9


LOD score
8
6
4
2
1
2
3
4
5
6
7
8
9
10
Markers on linkage group 9


Posterior probability of locus 1
0.03
0.02
0.01
0.0
A
1 2 3 5 6 7 8 10
Linkage Group 9
Posterior probability of locus 2
0.10
0.08
0.06
0.04
0.02
0.0
B
1 2 3 5 6 7 8 10
Linkage Group 9


Posterior probability of a single locus
0.05
0.04
0.03
0.02
0.01
0.0
C
1 2 3 4 5 6 7 8 10
Linkage Group 9