A Bayesian Approach to Detect Quantitative Trait Loci Using Markov Chain Monte Carlo

Jaya M. Satagopan,* Brian S. Yandell,[†] Michael A. Newton[†] and Thomas C. Osborn[‡]

**Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10021-6094, [†]Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706-1685 and [‡]Department of Agronomy, University of Wisconsin, Madison, Wisconsin 53706-1514*

Manuscript received August 10, 1995
Accepted for publication June 24, 1996

ABSTRACT

Markov chain Monte Carlo (MCMC) techniques are applied to simultaneously identify multiple quantitative trait loci (QTL) and the magnitude of their effects. Using a Bayesian approach a multi-locus model is fit to quantitative trait and molecular marker data, instead of fitting one locus at a time. The phenotypic trait is modeled as a linear function of the additive and dominance effects of the unknown QTL genotypes. Inference summaries for the locations of the QTL and their effects are derived from the corresponding marginal posterior densities obtained by integrating the likelihood, rather than by optimizing the joint likelihood surface. This is done using MCMC by treating the unknown QTL genotypes, and any missing marker genotypes, as augmented data and then by including these unknowns in the Markov chain cycle along with the unknown parameters. Parameter estimates are obtained as means of the corresponding marginal posterior densities. High posterior density regions of the marginal densities are obtained as confidence regions. We examine flowering time data from double haploid progeny of *Brassica napus* to illustrate the proposed method.

PLANT breeders and molecular biologists are interested in using molecular markers to identify genetic loci associated with quantitative traits (quantitative trait loci or QTL). Over the past few years statistical issues related to the identification of QTL have become topics of increased research, leading to the development of various linear models, test statistics, and confidence regions for the quantitative trait loci.

Likelihood based methods have been developed to identify a single QTL for a phenotypic trait. Two key approaches to this problem are the single marker *t*-test, also known as point analysis (SOLLER *et al.* 1976), and interval mapping (LANDER and BOTSTEIN 1989). Under the single marker *t*-test approach, the hypothesis of no difference between the two parental genotypes is tested at every marker locus using a *t*-test, identifying markers closely linked to the putative QTL. The main disadvantages of this procedure are that the exact location of the QTL cannot be estimated, and the gene effects are under estimated when the locus is far from marker loci.

The interval mapping method models the unknown genotypes of the putative QTL conditional on flanking markers. Estimates of parameters and unknown genotypes are obtained by an EM algorithm (LANDER and BOTSTEIN 1989). The profile likelihood of the QTL is used to obtain a support interval for the likely location of the gene. The LOD score test statistic, defined as the logarithm of the likelihood ratio to the base ten, tests for the presence of a putative QTL at every locus. Under the hypothesis of no QTL, the asymptotic distribution of the LOD score at any locus is proportional to a chi-squared distribution when the required regularity conditions hold. The locus corresponding to the maximum LOD score is the estimated location of a QTL if the maximum LOD score exceeds the chi-squared threshold at any predetermined level of significance. CHURCHILL and DOERGE (1994a) suggested randomly permuting trait values to determine the exact null-distribution of the LOD statistic. DUPUIS (1994) extended the work of LANDER and BOTSTEIN (1989), giving joint large sample confidence intervals for a single locus and its effect as LOD support intervals and Bayesian Credible Sets.

Many quantitative traits may be modified by multiple genes having effects of different magnitudes. A stepwise approach has been used to locate multiple QTL for both single marker *t*-test and interval mapping methods. After a single locus model is fitted, the residuals are examined for the presence of a second QTL, and so on. However, it is well known that the step-wise fitting of models result in biased estimates of gene effects. The step-wise approach may identify ghost (false) QTL when there is actually no QTL or when there are two or more QTL (KNOTT and HALEY 1992; MARTINEZ and CURNOW 1992). Hence, we need a method that can look for multiple QTL simultaneously.

Several recent works examined approximate methods to infer multiple QTL effects. HALEY and KNOTT (1992) used a multiple regression approach to fit a

Corresponding author: Jaya M. Satagopan, Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, 1275 York Ave., New York, NY 10021. E-mail: satago@biost.mskcc.org model with two QTL by searching the chromosome of interest in two dimensions. They found results similar to the maximum likelihood method using simulated data. Interval mapping combined with multiple regression have been used to detect multiple loci (ZENG 1993, 1994; JANSEN 1993; JANSEN and STAM 1994), with some selected markers outside the interval as cofactors in regression to reduce genetic variation. In all these approaches, the appropriate significance threshold to compare test statistics is typically approximated by the asymptotic distribution of the LOD statistic (but see KNOTT and HALEY 1992; CHURCHILL and DOERGE 1994a).

Bayesian approach to linkage analysis has been used previously (THOMAS and CORTESSIS 1992; HOESCHELE and VANRANDEN 1993a,b). While the maximum likelihood analysis implicitly assumes prior linkage between a marker and QTL, a Bayesian approach formalizes this by incorporating the prior linkage information into the analysis. Computing the likelihood involves summing over all possible combinations of QTL genotypes. The EM algorithm, as suggested by LANDER and BOTSTEIN (1989), overcomes this computational difficulty by calculating the expected QTL genotypes that would maximize the likelihood. In this paper we illustrate a Bayesian approach using MCMC and simply sample from the joint posterior of the unknown parameters and missing data. Inference summaries for the loci and their effects are based on their marginal posteriors. In computing the confidence intervals, non-Bayesian approaches do not properly account for uncertainties in the other parameters. Hence, the confidence intervals for the loci do not have their nominal coverage probabilities. The Bayesian approach at least addresses this concern by averaging over such uncertainties since it considers the marginal posterior of the loci given the data, although it may not completely overcome it.

HOESCHELE and VANRANDEN (1993a,b) computed the posterior probability of linkage between a QTL and a single (or a pair of) marker(s) for daughter and granddaughter designs in animal models. THOMAS and CORTESSIS (1992) estimated the LOD scores for linkage in a large pedigree using a Bayesian approach via the Gibbs sampler. However, they do not discuss confidence intervals for the locations of multiple loci and their effects, or estimate the number of loci affecting the trait of interest.

This paper is organized as follows. We first propose a stochastic model describing the distribution of data conditional on the unknown multiple QTL genotypes. Next, 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 genetic parameters. MCMC is used to study the resulting joint posterior of the parameters given the data, and estimated marginal posteriors are the basis for QTL estimates. Bayes factor is used to estimate the number of QTL. This proposal is illustrated using phenotypic trait data for *Brassica napus,* on days to flowering.

QTL MODEL

Consider a simple linear additive model for a phenotypic trait. 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. Consider a quantitative trait expressed by a single gene, subject to environmental variation. The observed phenotype y_i for the ith individual in a sample of size n may be given by the following linear model:

$$y_i = \mu + \alpha Q_i + \delta(1 - |Q_i|) + \epsilon_i, \tag{1}$$

where ϵ_i is a random, mean 0, deviation with variance σ^2, and (μ, α, δ) determine the expected response given the QTL genotype Q_i. The location of the putative locus, its genotypes and effects can be estimated from this model by assuming an appropriate distribution for the traits (LANDER and BOTSTEIN 1989). When the quantitative trait is expressed by multiple genes acting independently, Model (1) may be extended to

$$y_i = \mu + \sum_{j=1}^{s} \alpha_j Q_{ij} + \sum_{j=1}^{s} \delta_j(1 - |Q_{ij}|) + \epsilon_i. \tag{2}$$

For notation, let $Q_i = \{Q_{ij}\}_{j=1}^{s}$ denote the QTL genotypes for the ith individual, and let $\alpha = \{\alpha_j\}_{j=1}^{s}$ and $\delta = \{\delta_j\}_{j=1}^{s}$ denote the additive and dominance effects of the s loci, respectively. The genetic parameters are the QTL loci $\lambda = \{\lambda_j\}_{j=1}^{s}$ and the model unknowns $\theta = (\mu, \alpha, \delta, \sigma^2)$, where λ_j denotes the distance of the jth QTL from one end of the linkage group. Interaction terms can be incorporated in the above model (Equation 2) when there is evidence for epistasis between the putative loci (KNOTT and HALEY 1992). Specifying the error density, such as normal, leads to a corresponding probability density $\pi(y_i | Q_i, \theta)$.

Assume that a linkage map has been developed for the genome. For convenience we consider only one linkage group with ordered markers $\{1, 2, \ldots, m\}$ and known distances $D = \{D_k\}_{k=1}^{m}$, where D_k is the genetic map distance between markers 1 and k. Define $M_i = \{M_{ik}\}_{k=1}^{m}$ as the marker genotypes of the ith individual.

In practice, we observe the phenotypic trait y_i and the marker genotypes M_i but not the QTL genotypes Q_i. However, the probability distribution of the QTL genotypes, given the location of the putative loci, the marker genotypes and the distance between the markers, can be modeled in terms of recombination between the loci and the markers. Suppose the jth QTL is between markers k_j and $1 + k_j$ ($D_{k_j} \leq \lambda_j < D_{1+k_j}$), and that no other QTL lies in this marker interval, which is a reasonable assumption given a dense map. Under the Haldane assumption of independence of recombination events (OTT 1991), this QTL genotype is conditionally independent of nonflanking marker genotypes and other QTL genotypes given the genotype of flanking markers k_j and $1 + k_j$. With π denoting a probability density or mass function, the QTL genotype distribution for the ith individual becomes

$$\pi(Q_i | \lambda, M_i, D) = \prod_{j=1}^{s} \pi(Q_{ij} | \lambda_j, M_i, D)$$

(assuming the loci segregate independently)

$$= \prod_{j=1}^{s} \pi(Q_{ij} | \lambda_j, M_{ik_j}, M_{ik_j+1}, D_{k_j}, D_{k_j+1}) \quad (3)$$

(by Haldane independence assumption).

Each component in the above product can be obtained in terms of recombination between the markers k_j and $k_j + 1$, k_j and the jth QTL, and the jth QTL and $k_j + 1$ (KNAPP *et al.* 1990).

The likelihood of the parameters λ and θ from the ith individual may be expressed as

$$L(\lambda, \theta | y_i, M_i, D) = \sum_{q_i} \pi(y_i, Q_i = q_i | \lambda, \theta, M_i, D)$$

$$= \sum_{q_i} \pi(y_i | Q_i = q_i, \theta) \pi(Q_i = q_i | \lambda, M_i, D), \quad (4)$$

with the sum over the set of all possible QTL genotypes for the ith individual, $q_i = \{q_{ij}\} \in \{-1, 0, 1\}^s$. When the data $y = \{y_i\}_{i=1}^n$ are n independent observations, the likelihood becomes a product of factors of the form given by Equation 4. This can be expressed, after suppressing the notation for conditioning on $\{M_i\}_{i=1}^n$ and D, as

$$L(\lambda, \theta | y) = \prod_{i=1}^{n} \sum_{q_i} \pi(y_i | Q_i = q_i, \theta) \pi(Q_i = q_i | \lambda), \quad (5)$$

a familiar mixture model likelihood.

Our aim is to make inference about λ and θ using this likelihood, but evaluating this likelihood is not trivial. The likelihood is a finite mixture of densities and becomes very difficult to evaluate when there are multiple QTL. JANSEN (1993) demonstrates the EM algorithm for maximizing the likelihood in θ with λ fixed.

Rather than attempt optimization of the likelihood surface, we apply Bayesian analysis and therefore integrate this likelihood, modified by a prior, to produce inference summaries for all the components in the model. We can infer the position of QTL and their corresponding effects. In addition, we can estimate the number of QTL using Bayes factor for model selection. The definition of Bayes factors and justification for its use are given in the section HOW MANY QTL.

The extent to which the choice of the prior distribution over the parameter space affects the final inference is a measure of robustness and requires checking in each application. The prior could be chosen based on related studies or information from the literature. Here we use diffuse normal priors (specific details of prior for the *B. napus* data are provided in the subsection Prior distribution). Bayes theorem combines the data and the prior to produce a posterior distribution over all unknown quantities. With $Q = \{Q_i\}_{i=1}^n$ denoting the QTL genotypes for all the n observations, the posterior density of λ, θ and Q is given by

$$\pi(\lambda, \theta, Q | y) \propto \pi(y | Q, \theta) \pi(Q | \lambda) \pi(\lambda, \theta), \quad (6)$$

with $\pi(y | Q, \theta) = \prod \pi(y_i | Q_i, \theta)$, the data probability mass given the QTL genotypes; $\pi(Q | \lambda) = \prod \pi(Q_i | \lambda)$, the probability mass of the QTL genotypes of all the observations given their locations (and M and D); and $\pi(\lambda, \theta)$, the prior density of the genetic parameters. We assume prior independence of the parameters

$$\pi(\lambda, \theta) = \pi(\lambda) \pi(\mu) \pi(\sigma^2) \prod_{j=1}^{s} \{\pi(\alpha_j) \pi(\delta_j)\}.$$

A natural choice for prior of λ 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, $\{\alpha_j\}_{j=1}^s$, $\{\delta_j\}_{j=1}^s$ and σ^2 makes its form simple while increasing diffuseness makes the prior objective.

In the Bayesian approach we infer the genetic parameters based on their marginal posterior distribution, which can be obtained from the joint posterior (Equation 6) by integrating over the other unknowns. Exact solution to such high-dimensional integrals are difficult, but Monte Carlo approximation, as described in the following section, is quite feasible.

PARAMETER ESTIMATION

Markov chain Monte Carlo methods are used commonly to evaluate complex integrals involving likelihoods and to summarize posterior distributions in Bayesian problems. Here we use MCMC to study the joint posterior density given by Equation 6, by constructing a Markov chain with this target distribution.

The target distribution (6) lives on a high-dimensional product space. The QTL distances $\lambda = \{\lambda_j\}_{j=1}^s$ range from 0 to the maximum length D_m, such that $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_s \leq D_m$. The QTL genotypes are elements of a discrete space, $Q \subset \{-1, 0, 1\}^{s \times n}$. Each additive and dominance effect, and the overall mean μ vary in the real line, and the environmental variance $\sigma^2 > 0$. We construct a Markov chain over the resulting product space. Various approaches have been suggested to construct such Markov chains, for example, the Gibbs sampler (GEMAN and GEMAN 1984), Metropolis-Hastings algorithm (HASTINGS 1970) and hybrid schemes (TIERNEY 1994). Our method is a Gibbs sampler, with Metropolis steps to update QTL positions.

Algorithm: The Markov chain is a random sequence of states

$$(\lambda^0, Q^0, \theta^0), (\lambda^1, Q^1, \theta^1), \ldots, (\lambda^N, Q^N, \theta^N)$$

started at an arbitrary point (λ^0, Q^0, θ^0) having positive posterior density, and proceeding by simple rules that modify the three unknowns λ, Q, and θ. Each step in this chain is a cycle of three smaller steps first updating λ then Q, followed by θ. The step updating λ is a Metropolis-Hastings step. The steps updating Q and θ are Gibbs sampler steps. More specifically, given the current state (λ, Q, θ), we proceed as follows.

Updating $\boldsymbol{\lambda}$: Elements of λ are modified one at a time using the Metropolis algorithm. For the jth locus, a proposal λ_j^* is generated from a uniform distribution on the interval ($\max(\lambda_{j-1}, \lambda_j - d)$, $\min(\lambda_{j+1}, \lambda_j + d)$). This distribution is denoted by $u(\lambda_j, \lambda_j^*)$ and maintains ordering of the loci (note $\lambda_0 = 0$, and $\lambda_{s+1} = D_m$). The tuning parameter $d > 0$ affects the conditional variance of the proposal. The proposal is accepted with probability $\min(\alpha, 1)$, as given below in Equation 7.

$$\alpha(\lambda_j, \lambda_j^*) = \frac{\pi(\lambda_j^* | \lambda_{-j}, Q, \theta, y)\, u(\lambda_j^*, \lambda_j)}{\pi(\lambda_j | \lambda_{-j}, Q, \theta, y)\, u(\lambda_j, \lambda_j^*)}, \quad (7)$$

where $\lambda_{-j} = \{\lambda_{j'}: 1 \leq j' \leq s, j \neq j'\}$. If the proposal is not accepted, the state remains unchanged, and the algorithm proceeds to update the next QTL position. The unknown loci can be updated more than once between updates of the other parameters. One update per cycle is sufficient. However, if there is evidence that the chain is mixing slowly (*i.e.*, slow convergence to the equilibrium distribution), then more than one update of λ between other updates will improve mixing.

Updating $\boldsymbol{Q}$: The conditional independence of the QTL genotypes Q (see Equation 3) imply that their updates can proceed separately for each individual. Hence for each individual i and QTL j, draw the QTL genotype Q_{ij} from its full conditional density $\pi(Q_{ij} | \lambda, Q_{i(-j)}, \theta, y)$, where $Q_{i(-j)} = \{Q_{ij'}: 1 \leq j' \leq s, j \neq j'\}$. For instance, the full conditional that an unknown genotype is 1 is Bernoulli with probability p_{ij} given by

$$\begin{aligned} p_{ij} &= \pi(Q_{ij} = 1 | \lambda, Q_{i(-j)}, \theta, y) \\ &= \frac{\pi(Q_{ij} = 1 | \lambda_j)\pi(y_i | \theta, Q_i, Q_{ij} = 1)}{\Sigma_q \pi(Q_{ij} = q | \lambda_j)\pi(y_i | \theta, Q_i, Q_{ij} = q)}. \quad (8) \end{aligned}$$

Updating $\boldsymbol{\theta}$: We update each component of θ by considering their corresponding full conditionals. The parameter μ is updated from its full conditional $\pi(\mu | \lambda, Q, \alpha, \delta, \sigma^2, y)$. If this full conditional is completely known and is easy to sample from, then we sample μ directly from this full conditional as in a Gibbs sampler. For example, for the Brassica data to be discussed later, this full conditional is Gaussian given by Equation 20 in APPENDIX A. Hence updating μ is equivalent to drawing a random sample from this Gaussian distribution. If the full conditional is not easy to sample from, then a Metropolis-Hastings algorithm can be used.

To update $\alpha = \{\alpha_j\}_{j=1}^s$, each component α_j can be updated using its full conditional $\pi(\alpha_j | \lambda, Q, \mu, \alpha_{-j}, \delta, \sigma^2, y)$, where $\alpha_{-j} = \{\alpha_k: 1 \leq k \leq s, k \neq j\}$. For the Brassica example this full conditional is again Gaussian (Equation 21, APPENDIX A) and can be easily updated. The dominance parameter $\delta = \{\delta_j\}_{j=1}^s$ can be updated as α. We have considered double haploid Brassica progeny in the example and hence $\delta = 0$. The variance σ^2 can be updated from its full conditional $\pi(\sigma^2 | \lambda, Q, \mu, \alpha, \delta, y)$. This full conditional is an inverse gamma distribution for our example (Equation 22, APPENDIX A).

Missing marker data: In practice, some marker data are missing. The missing marker genotypes affect the model of QTL genotypes given the markers. For a single individual, the probabilities of QTL genotypes can be computed on whatever marker information is available for that individual. However, all contributions to the likelihood are based on the (flanking) markers at common recombination distances from the putative QTL. By incorporating the missing marker genotypes as further unknowns in the MCMC approach, these recombination distances need not be recomputed for every individual.

The MCMC approach handles the missing data very naturally. Partition the marker genotypes as known and missing, $M = (M^+, M^-)$. Further, assume that the estimated distances between markers are not affected by missing marker genotypes. We now consider the unknowns to consist of (λ, Q, θ, M^-). The equilibrium distribution of interest is

$$\pi(\lambda, \theta, Q, M^- | y, M^+, D).$$

Updating the loci (λ), genotypes (Q), and the unknown parameters (θ) proceeds as before. In addition, the missing marker genotypes are updated individually based on their full conditional densities. For instance, suppose the kth marker corresponding to the ith individual is missing. $\pi(M_{ik}^- = 1 | \lambda, Q_i, \theta, y, M_i^+, D)$ is the full conditional that this missing marker genotype is 1. This full conditional is Bernoulli with probability q_{ik} given by

$$q_{ik} = \frac{\pi(M_{ik}^- = 1 | \lambda, Q_i, M_i^+)}{\Sigma_m \pi(M_{ik}^- = m | \lambda, Q_i, M_i^+)}. \quad (9)$$

Further discussions on this probability are provided in APPENDIX B.

Justification: The MCMC algorithm is justified by the fact that if f is any function of the unknowns that is square integrable with respect to the equilibrium distribution π, then

$$\bar{f}_N = \frac{1}{N}\sum_{t=1}^{N} f(\lambda^t, Q^t, \theta^t) \rightarrow E_\pi[f(\lambda, Q, \theta) | y],$$

$$\text{almost surely as } N \rightarrow \infty, \quad (10)$$

where (λ^t, Q^t, θ^t) are samples from the Markov chain (*e.g.*, TIERNEY 1994).

Estimates: Suppose we are interested in estimating the location λ_1. We can use the empirical means of $f(\lambda^t$,

Q^t, θ^t) $= \lambda_1^t$. By (10), $\bar{\lambda}_1 = \Sigma_{t=1}^N \lambda_1^t / N$ is a simulation consistent estimator of $E_\pi(\lambda_1 | y)$, itself a statistic and a Bayes estimator of the unknown QTL location. Estimates of other loci and parameters arise similarly as empirical averages of the corresponding MCMC samples.

The marginal posterior densities of the parameters of interest (μ, α, δ, σ^2) can be obtained from the sample values either by Rao-Blackwellization (GELFAND and SMITH 1990; LIU *et al.* 1994), or by kernel density estimation. For example, the Rao-Blackwell density estimator of $\pi(\alpha_1 | y)$ is

$$\bar{\pi}(\alpha_1 | y) = \frac{1}{N} \sum_{t=1}^{N} \pi(\alpha_1 | \lambda^t, Q^t, \mu, \alpha_{-1}^t, \delta^t, (\sigma^t)^2, y). \quad (11)$$

On the other hand, we choose to estimate the marginal posterior density of λ_1 using a histogram kernel

$$\bar{\pi}(\lambda_1 | y) = \frac{1}{Nh} \sum_{j=0}^{\lfloor D_m / h \rfloor} I_{(jh,(j+1)h]}(\lambda_1) \sum_{t=1}^{N} I_{(0,1]} \left(\frac{\lambda_1^t}{h} - j \right), \quad (12)$$

where h is the bin-width of the histogram, D_m is the length of the linkage group, and $I_{(a,b]}(x) = 1$ if $a < x \leq b$ and 0, otherwise. Marginal posterior density estimates of other unknowns can be obtained similarly.

Confidence intervals: Confidence intervals for the parameters of interest can be obtained as high posterior density (HPD) regions (BOX and TIAO 1973). The HPD region with coverage rate $1 - \alpha$ for λ is defined as a region $\mathcal{R}$ in the parameter space of λ such that $\pi(\lambda \in \mathcal{R} | y) = 1 - \alpha$ and, for $\lambda^1 \in \mathcal{R}$ and $\lambda^2 \notin \mathcal{R}$, $\pi(\lambda^1 | y) \geq \pi(\lambda^2 | y)$. RITTER (1992) describes a method for calculating approximate HPD regions from MCMC samples. Simply, for each simulated λ^t, calculate the posterior density $\pi_t := \pi(\lambda^t | y)$ and rank order these values as $\pi_{(1)} > \cdots > \pi_{(N)}$. Next, form cumulative sums $c_{(t)} = \Sigma_{j=1}^t \pi_{(j)}$. Define t^* as the smallest index for which $c_{(t_*)} > 1 - \alpha$. The approximate HPD region is defined by the point cloud formed by the sample vectors with marginal posterior values larger than or equal to π_{t_*}. Marginal HPD regions for any component λ_j, or any other parameter, can be computed similarly, using smooth density estimates of the corresponding marginal.

Implementation issues: The asymptotic variance of $\bar{f}_N$ is an estimate of the Monte Carlo variance (GEYER 1992) and is given by (GREEN and HAN 1992)

$$N \operatorname{Var}(\bar{f}_N) \approx \sigma^2 \sum_{t=-\infty}^{\infty} \rho_t(f), \quad (13)$$

where ρ_t is the lag t autocorrelation. This is very difficult to compute in practice. GEYER (1992) suggests various time series approaches to estimate the Monte Carlo error.

Another major implementation issue is to determine the length of the chain (N). TIERNEY (1994) and SMITH and ROBERTS (1993) suggest examining the series $f(\lambda^t, Q^t, \theta^t)$ using time series plots or autocorrelation plots. This could provide evidence that the chain is not sufficiently long. Further, to ensure that we are using only those samples after the chain has attained equilibrium distribution, a common approach is to discard the initial few runs (burn-in period) and consider only the remaining samples for estimation purposes. Subsampling the chain at regular intervals is also a common practice to reduce serial correlation between the samples. GEYER (1992) and TIERNEY (1994) provide detailed discussions on implementation issues.

HOW MANY QTL?

In the frequentist approach, the LOD score is used as a test statistic to detect QTLs. CHURCHILL and DOERGE (1994a) state that the finite sample size and distribution of the quantitative trait could cause one to doubt the reliability of the asymptotic distribution of the LOD score. Here we present an alternative approach to detect QTLs based on Bayesian model selection criteria. Instead of calculating the likelihood of the parameters, which is hard to compute, we use the samples from the Markov chain to calculate Bayes factor (JEFFREYS 1961; KASS and RAFTERY 1995). In particular, we suggest selecting the number of QTL affecting the trait by running MCMC under different models (example, models with $s = 1, 2, \cdots$) and comparing them using Bayes factors.

Let model$_1$ and model$_2$ be two models that are to be compared. The posterior odds in favor of model$_1$ against model$_2$ for data y can be expressed, using Bayes theorem, as

$$\frac{\pi(\text{model}_1 | y)}{\pi(\text{model}_2 | y)} = \frac{\pi(\text{model}_1)}{\pi(\text{model}_2)} \frac{\pi(y | \text{model}_1)}{\pi(y | \text{model}_2)}, \quad (14)$$

where the first factor is the prior odds, and

$$B_{12} = \frac{\pi(y | \text{model}_1)}{\pi(y | \text{model}_2)} \quad (15)$$

is the ratio of marginal probabilities of y given the two models and is called the Bayes factor. When the QTL genotypes are known, the marginal probability of the data under model$_1$ and model$_2$ can be written as

$$\pi(y | \text{model}_j) = \int \pi(y | \lambda_j, \theta_j) \pi(\lambda_j, \theta_j) d(\lambda_j, \theta_j), \quad (16)$$

with prior $\pi(\lambda_j, \theta_j)$ and data probability density $\pi(y | \lambda_j, \theta_j)$ under model$_j$, $j = 1, 2$. Note that the posterior probability of model$_j$ is

$$\pi(\text{model}_j | y) \propto \pi(y | \text{model}_j) \pi(\text{model}_j)$$

with some arbitrarily chosen prior $\pi(\text{model}_j)$. For instance, a uniform prior can be used for a finite collection of models $j = 0, 1, \ldots, J$.

KASS and RAFTERY (1995) suggest using the harmonic

TABLE 1

Bayes factor for double haploid progeny

	n = sample size		
LOD	50	100	200
0.0	50	100	200
0.5	15.81	31.62	63.25
1.0	5	10	20
1.5	1.58	3.16	6.32
2.0	0.5	1.0	2.0
2.5	0.16	0.31	0.632
3.0	0.05	0.10	0.2
5.0	0.0005	0.001	0.002
10.0	5e-9	1e-8	2e-8

mean estimator of $\pi(y|\text{model}_j)$ that is given by (Newton and Raftery 1994)

$$\hat{\pi}(y|\text{model}_j) = N\Big/\left\{\sum_{t=1}^{N} 1/\pi(y|Q^t, \theta^t)\right\}, \qquad j = 1, 2. \quad (17)$$

This consistent estimator may be unstable, with infinite variance. Taking h to be an arbitrary density on the target space, the more general

$$\hat{\pi}(y|\text{model}_j) = \left\{\frac{1}{N}\sum_{t=1}^{N}\frac{h(\lambda^t, Q^t, \theta^t)}{\pi(y|Q^t, \theta^t)\pi(\lambda^t, Q^t, \theta^t)}\right\}^{-1} \quad (18)$$

may be more stable, being asymptotically normal provided $\int h^2(x)/(\pi(y|x)\pi(x))\,dx < \infty$ (Kass and Raftery 1995). We use $h(\lambda, Q, \theta) = h(\lambda)\pi(Q|\lambda)\pi(\theta)$ with $h(\lambda)$ a normal density restricted to $0 \leq \lambda_1 \leq \cdots \leq \lambda_s \leq D_m$ with center $\bar{\lambda}$ and variance acting as a tuning parameter.

The harmonic mean estimator (17) is unstable because the "complete likelihoods" $\pi(y|Q^t, \theta^t)$ are normal ordinates and hence drop quickly for extreme Q^t or θ^t. Noting that any harmonic mean of likelihoods is a consistent estimator for the marginal probability of the data, we obtain another estimator by integrating σ^2. This results in the harmonic mean of the heavier tailed t densities (when the prior for σ^2 is inverse Gamma, Bernardo and Smith 1994, page 139). For an alternative approach if $\pi(\lambda|Q, \theta, y)$ were known completely, see Frühwirth-Schnatter (1995).

Frequentist significance tests can be used to reject a hypothesis. For example, when testing H_0, one QTL affects the trait *vs.* H_1, more than one QTL affects the trait, one may reject H_0. Bayes factors offer a way to evaluate evidence in favor of a null hypothesis, particularly in situations comparing more than two models. Further, Bayes factors provide a way of incorporating external (prior) information to evaluate the hypotheses of interest (Kass and Raftery 1995).

Interpretation of LOD score using Bayes factor: The LOD score can be interpreted in terms of the Bayes factor (Kass and Raftery 1995, page 778),

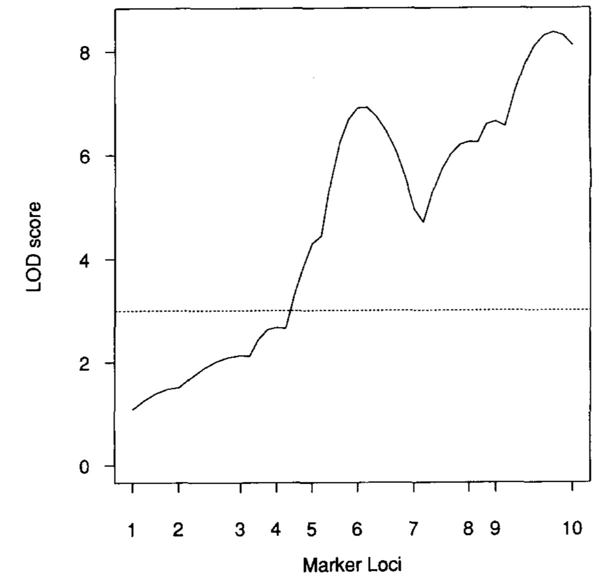

FIGURE 1.—Order of markers (horizontal axis) and LOD score (vertical axis) obtained from MAPMAKER/QTL for days to flowering with 8 weeks vernalization. Dotted horizontal line corresponds to LOD score of 3.0.

$$\text{LOD} \approx -\log_{10}(B_{12}) - \tfrac{1}{2}(d_1 - d_2)\log_{10}(n), \quad (19)$$

where d_j is the number of parameters in model$_j$, and n is the total number of observations. Recall that LOD = $\log_{10}(L_1/L_2)$ with $L_j = \pi(y|\hat{\lambda}_j, \hat{\theta}_j)$ and $\hat{\theta}_j$ and $\hat{\lambda}_j$ the maximum likelihood estimates under model$_j$.

Table 1 shows Bayes factors for commonly used LOD scores and sample sizes n = 50, 100, 200. A Bayes factor near 1 (say, between 0.3 and 3) does not favor either model. In practice, B_{12} larger than 100 (smaller than 0.01) decisively supports model$_1$ (model$_2$), following Jeffreys (1961). For example, when n = 100, a LOD score of three decisively favors model$_2$ (Bayes factor of 0.1) but LOD 1.5 would not favor either model. However, when the sample size is 200, a LOD score of 1.5 would substantially favor model$_1$.

FLOWERING TIME IN BRASSICA

The data and model structure: 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 *B. napus* to detect QTLs for flowering time. A double haploid line from the *B. napus* cv. Stellar (an annual canola cultivar) was crossed to a single plant of cv. Major (a biennial rapeseed cultivar) that was used as a female. One hundred 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 three groups and each group

TABLE 2

LOD scores (from stepwise regression using EM algorithm) and Bayes factors (using normal weighting density in Equation 19) for Brassica model selection

Models	LOD	LOD(BF)	BF
0–1	8.375	8.57	2.8e-7
1–2	1.715	2.48	0.12
2–3	0.706	1.50	1.1

LOD(BF) is the approximation of LOD using the Bayes factor.

was exposed to one of the three 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.* (1995a). DNA extraction and linkage map construction are given in FERREIRA *et al.* (1995b). To illustrate 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 1 shows the profile along linkage group 9 of the LOD score for flowering time for 8 weeks vernalization obtained using the EM algorithm for a single QTL model (LANDER and BOTSTEIN 1989). We can observe that the LOD score has two peaks, the larger between markers 9 and 10 (LOD = 8.37) and the second around marker 6 (LOD = 6.91), suggesting the possibility of two QTL in that linkage group. We fixed a QTL at the higher peak and found an increase in the LOD score of 1.715 for a second putative QTL around marker 6. Fixing both these QTL and looking for a third led to only an increase of only 0.706 in the LOD score (see Table 2). We further examine this using MCMC.

A number of models could be compared. We examine the following: (1) there is no QTL *vs.* there is at least one QTL, (2) there is only one QTL *vs.* there are at least two QTLs, and (3) there are only two QTLs *vs.* there are at least two QTLs. Earlier investigations showed some evidence of increasing variance with increasing days to flower. Further, it appeared that putative genes at QTL might have a multiplicative effect, that is, additive on a logarithmic scale. Therefore, we use the following model for the number days to flowering for the *i*th DH line:

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

where y_i is log 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, there is no dominance term δ_j in the above model. The random errors ϵ_i are assumed to have independent Gaussian distributions with mean 0 and common variance σ^2.

Prior distribution: 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. When some information about the unknowns is available, the priors may be chosen by putting more weight in a desired range. For example, it is believed that alleles from Stellar parent result in shorter time to flowering than alleles from Major parent. The overall mean μ is given a Gaussian prior, centered at zero with variance 10 to make the distribution diffuse. The QTL effects α_j, $j = 1, \ldots, s$, are given independent normal priors, also centered at 0 with large variance (10) allowing for the possibility of extreme QTL effects. The phenotypic variance σ^2 is assumed to have an inverse gamma prior. In the absence of prior information about the QTL locations, any position along the linkage group could be a possible position for the QTLs. Hence λ_j's are assumed to have uniform prior along the entire linkage group 9 such that $0 < \lambda_1 < \cdots < \lambda_s < D_m$.

The full conditional densities of the parameters under the above priors are given in APPENDIX A. For each analysis, the Markov chain ran 400,000 cycles and was sampled every 200 cycles, without any initial burn-in, to give a working set of 2000 states. Time series methods to estimate the Monte Carlo error using Tukey-Hanning window suggested that these are large enough samples to precisely estimate posterior quantities. Table 3 gives the parameter estimates and estimated Monte Carlo standard errors for the single, two and three QTL models.

TABLE 3

Parameter estimates and Bayes factor for the three models relative to the two QTL model

s	μ	σ^2	Effect	Locus	Effect	Locus	Effect	Locus	BF
1	3.060	0.081					−0.165	71.9	0.12
	(0.010)	(0.027)					(0.034)	(0.045)	
2	3.061	0.078	−0.066	42.2			−0.128	76.4	
	(0.005)	(0.010)	(0.047)	(0.026)			(0.044)	(0.070)	
3	3.060	0.080	−0.041	38.6	−0.044	57.8	−0.111	77.9	1.1
	(0.021)	(0.004)	(0.060)	(0.019)	(0.067)	(0.035)	(0.050)	(0.016)	

Monte Carlo standard errors in parentheses. *s*, the number of QTL in each model; BF, Bayes factor.

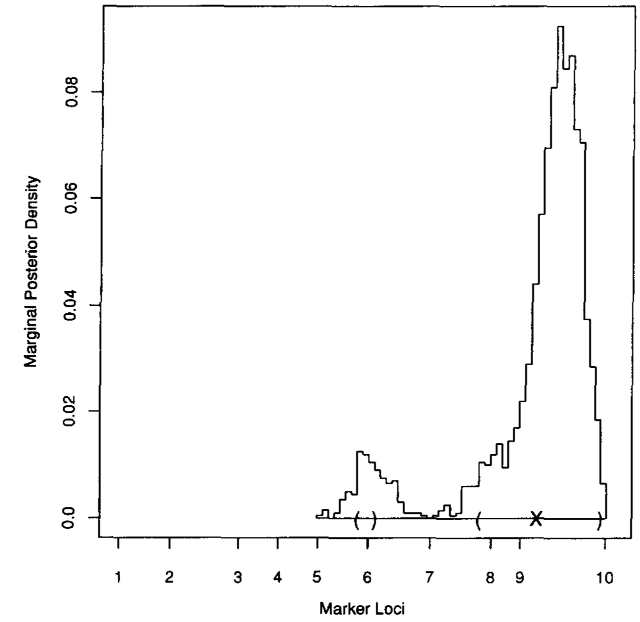

FIGURE 2.—Single QTL model. Marginal posterior density of the location of a single putative QTL, as obtained from MCMC, is shown. The estimate of the location is shown by ×. 90% HPD confidence interval is shown by parentheses.

Single QTL model: The starting value for the single putative locus was $\lambda_1^0 = 24$ cM. A single QTL was estimated to be at $\bar{\lambda}_1 = 71.9$ cM between markers 9 and 10. The estimated effect was $\bar{\alpha}_1 = -0.165$ cM, implying that the logarithm of days to flowering is decreased if the QTL has alleles from the Stellar parent. This is consistent with the initial knowledge about flowering time of the parents. The histogram estimate of $\pi(\lambda_1 | y)$ is presented in Figure 2. We observe that the marginal posterior density of λ_1 appears to be bimodal with a high mode between markers 9 and 10, and a smaller one between markers 5 and 6, which is consistent with LOD scores obtained from the EM algorithm (Figure 1). The autocorrelation function of the single locus (λ_1) subsampled at every 100th and every 200th cycle of the chain is shown in Figure 3. Autocorrelation between the every 100th subsample was significant even at lag 30. However, autocorrelation between every 200th subsample was very small. The marginal posterior of the effect α_1 using Equation 11 is shown in Figure 4. From these marginal posterior densities, the corresponding confidence intervals were obtained as HPD regions. The 90% HPD confidence region for λ_1 was the union of two intervals: (41, 44) cM between markers 5 and 6, and (62, 83) cM between markers 7 and 10 (represented using parentheses in Figure 2). This gave further support to the possibility of two loci on chromosome 9 controlling days to flowering. To explore this, we fit a two QTL model.

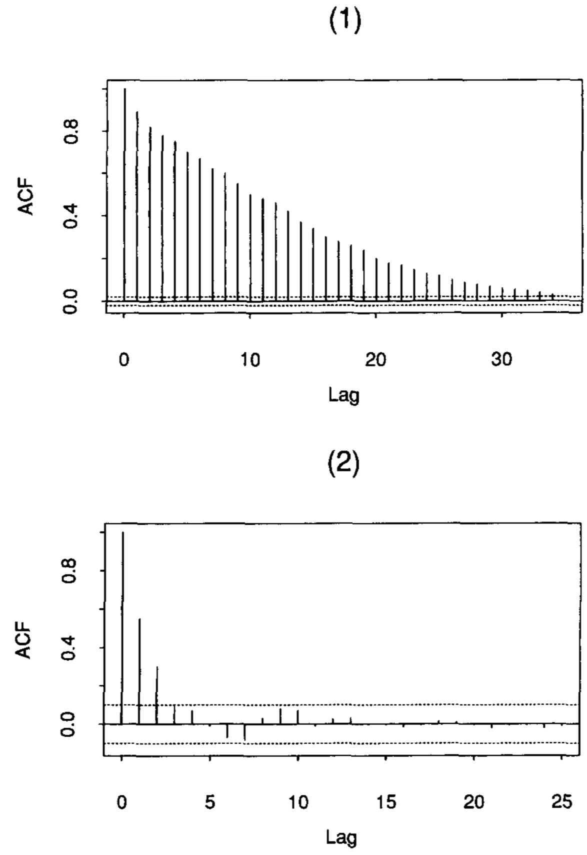

FIGURE 3.—Single QTL model. Autocorrelation function of the single locus (λ_1) from the single QTL model subsampled at (1) every 100th run and (2) every 200th run of the chain.

Two QTL model: The starting values for the two loci were $\lambda_1^0 = 24$ cM and $\lambda_2^0 = 59$ cM. The first locus λ_1 was estimated at 42.2 cM between markers 5 and 6, and locus λ_2 at 76.4 cM, between markers 9 and 10. The effect α_2 of locus λ_2 was nearly twice that of α_1. The locus λ_2 having the larger effect corresponded to the high mode in the single QTL model. The posterior

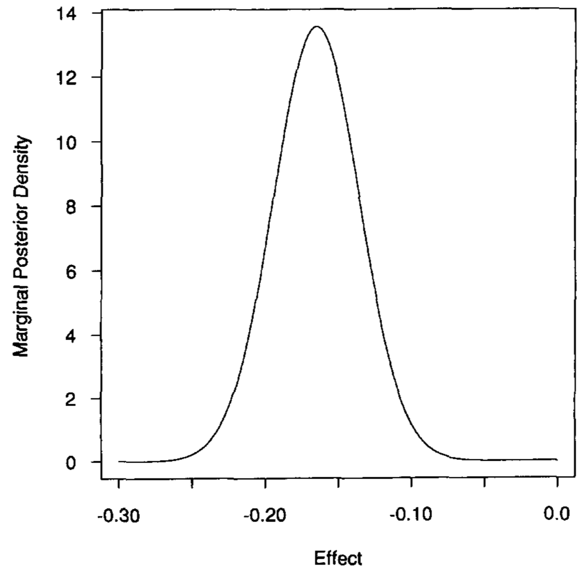

FIGURE 4.—Single QTL model. Marginal posterior density of the QTL effect α_1 estimated by Rao-Blackwellization.

TABLE 4

Posterior correlation of the parameters from a two QTL model

	Locus 2	μ	Effect 1	Effect 2	σ^2
Locus 1	0.163	−0.017	−0.152	0.157	0.010
Locus 2		−0.154	0.080	−0.141	−0.163
μ			−0.181	0.181	0.047
Effect 1				−0.987	−0.050
Effect 2					0.071

correlation between the parameters are shown in Table 4. Posterior correlation between the two loci was small. However, their effects were highly correlated. The joint HPD regions for the loci were obtained by first estimating the joint posterior density of $\lambda = (\lambda_1, \lambda_2)$ from their histogram. The joint HPD regions for the loci and effects are presented in Figures 5 and 6 superimposed on scatterplots of the Markov chain realizations. The marginal posterior distributions of the loci are shown in Figure 7. The locus with the larger effect is estimated at 76.4 cM, between markers 9 and 10, while the one with the smaller effect is estimated at 42.2 cM, between markers 5 and 6. The locus with the larger effect corresponds to the higher peak in the single QTL model (Figures 1 and 2) and is highly concentrated, likely due to the large effect of this locus. The distribution of the other QTL is somewhat concentrated but has a larger tail, possibly due to its small effect. We fit a three QTL model to investigate whether there is evidence for any more loci.

Three QTL model: The starting values for the loci were $\lambda_1^0 = 24$ cM, $\lambda_2^0 = 42$ cM and $\lambda_3^0 = 59$ cM. Marginal posterior distributions of the three QTL (Figure 8) indicate that loci 1 and 3 correspond to those in the previ-

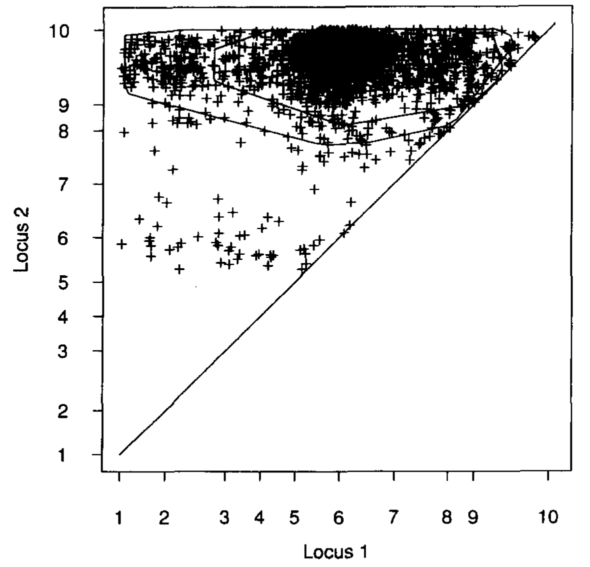

FIGURE 5.—Two-QTL model. Joint 50, 90 and 95% HPD confidence regions for $\lambda = (\lambda_1, \lambda_2)$.

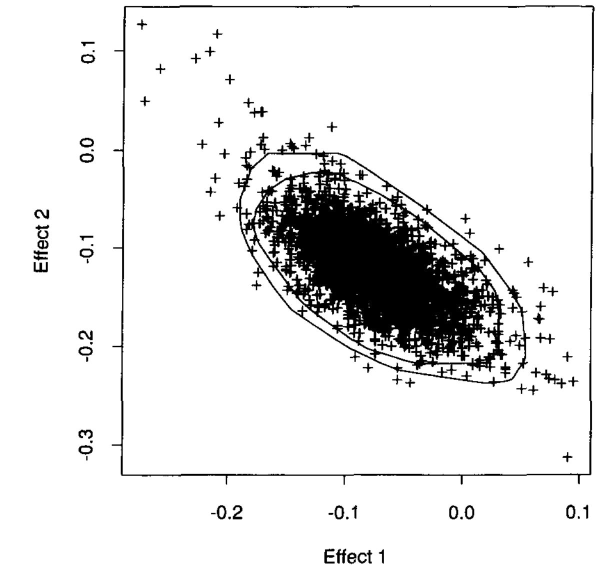

FIGURE 6.—Two-QTL model. Joint 50, 90 and 95% HPD confidence regions for the effects (α_1, α_2) obtained from their joint posterior density.

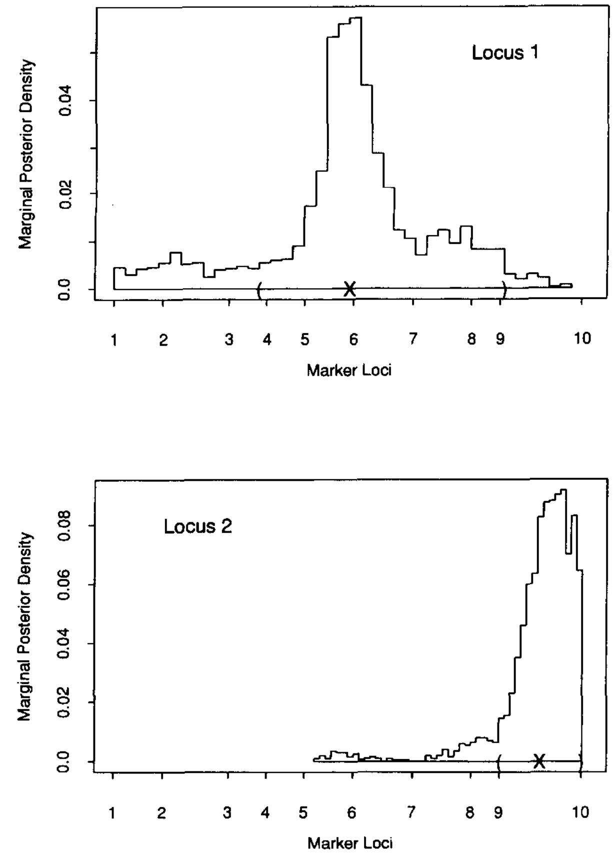

FIGURE 7.—Two QTL model. (A and B) Marginal posterior densities of the location of two putative loci. Estimates of the loci are shown by ×. 90% HPD confidence intervals are shown by parentheses. (C) The joint posterior density of the effects of the loci. Each point corresponds to a sample obtained from MCMC. 90% HPD confidence region and marginal posterior densities are also shown.

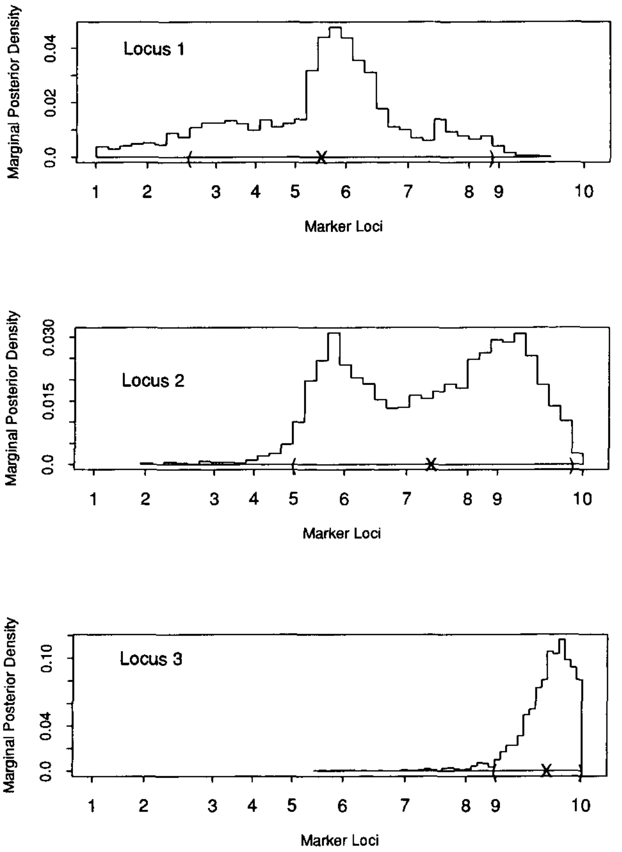

FIGURE 8.—Three-QTL model. (A–C) Marginal posterior densities of the three putative loci are shown. Estimated locations are shown by ×. 90% HPD confidence intervals are shown by parentheses.

ous model. The distribution of putative locus 2 is bimodal with modes in the same regions as the other two loci. This seems to indicate that there may be only two QTL supported by the data in this linkage group. Note that the effects of loci 1 and 3 are similar to the corresponding effects in the two QTL model (Table 3).

Model selection: Examining the plots of loci and effects (Figures 2–8) suggested a two QTL model. LOD scores obtained from the stepwise regression approach using EM algorithm (LANDER and BOTSTEIN 1989) did not support a two-QTL model at the threshold of 2 (Table 2). The harmonic mean estimate of Bayes factor (Equation 17) comparing one and two QTL models was 0.011. Comparison between two and three QTL models resulted in a Bayes factor of 1.8. Preliminary investigations with normal weighting densities h (in Equation 18) resulted in a Bayes factor of 0.12 to compare one and two QTL models. Comparing two and three QTL models gave a Bayes factor of 1.1. Using the harmonic of multivariate t densities (in Equation 18) gave Bayes factors of 0.12 and 1.23 to compare one and two, and two and three QTL models respectively.

We were concerned that the estimates of Bayes factors were not very stable when there were several QTL in the model. Ten repeated runs of the Markov chain found the harmonic mean estimate of Bayes factors comparing one and two QTL in the range 0.00028–0.55, and for comparing two and three QTL in the range 1.8–440. Hence, five repeated runs of the chain were obtained to study the stability of the estimates using normal and multivariate t weighting densities (Equation 18). Both these approaches gave stable estimates of Bayes factors. Estimates from multivariate t weights were found to be more stable than normal weights.

All the three approaches to estimate Bayes factors favored a two QTL model over a single QTL model. These estimates did not distinguish between a two or a three QTL model but supported both the models equally. Hence, we infer that the two QTL model is appropriate.

A small simulation experiment: To investigate the performance of the MCMC method for the flowering data, 100 parametric bootstrap data sets, with two QTL and true parameters as those estimated from the flowering data, were simulated. Both one and two QTL models were fit for each of the 100 simulated data using the same chain length, burn-in and subsampling as the flowering data. Bayes factor comparing the two models favored the two QTL model 72% of the times. Note that the effect of one of the QTLs, in the two QTL model for the flowering data, is very small (−0.066). Hence it is not surprising that the two QTL model is not supported 28% of the time. We also observed the number of times the estimates of the two loci from the 100 data sets were within the corresponding 90% HPD confidence region obtained for the flowering data and observed that 95% of the estimates were contained within the 90% HPD confidence region from the flowering data. This is shown in Figure 9. For the effects of the two loci, we observed that 92% of the estimates were within the 90% HPD confidence region.

DISCUSSION

In this paper we have fit a model that allows for multiple loci, instead of fitting one locus at a time or accounting for other putative loci by using markers as cofactors. Simultaneous fitting of multiple loci leads to multi-dimensionality problems (HALEY and KNOTT 1992; JANSEN and STAM 1994; ZENG 1994). We have handled this by using the MCMC approach to simultaneously look for multiple loci and their effects. The marginal posterior distribution of the parameters of interest can be obtained, which is not available with other existing methodologies. Inference for parameters is based on the marginal posterior densities. Mean and high posterior density region from the marginal posteriors are obtained as parameter estimates and inference regions, respectively. The parameter estimates, standard errors and Monte Carlo errors were not affected by considering burn-in period of 5000, 10,000 and 50,000 initial

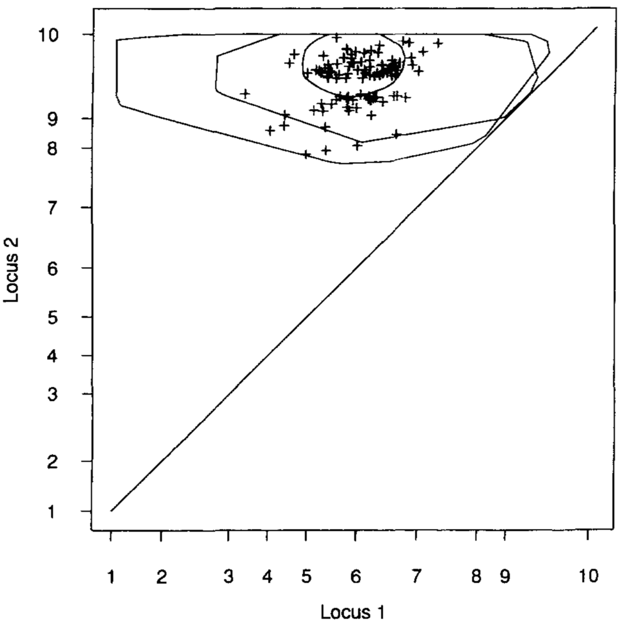

FIGURE 9.—Simulated data. Estimates of the two QTL from 100 parametric bootstrap data sets for 8-wk vernalization data. Also shown are the 50, 90 and 95% HPD confidence region for the loci from the flowering data.

samples. Hence we did not use any burn-in period. Rejection rates for λ from the Metropolis-Hastings step were ~50%. Bayes factors were used to estimate the number of QTL affecting the trait. Using a time-shared DEC-ALPHA machine, we found that the time to simulate a chain of length 400,000 was ~10 min for a single QTL model, 25 min for a two-QTL model and ~75 min for a three-QTL model.

We first considered a no QTL model and estimated μ and σ^2 for this model using the MCMC approach. Bayes factor to compare this model with a single QTL model was 2.8e-07, strongly favoring a single QTL model over none. Hence we proceeded with single-, two- and three-QTL models as described in the previous section. Preliminary studies of the harmonic mean estimators of Bayes factors, by repeated runs of the Markov chain, suggested some instability for comparing two- and three-QTL models. The harmonic mean of multivariate t densities and the weighted harmonic mean reported here were more stable across simulations. However, further investigation of these and other approaches is under way.

Although we have analyzed data from a double haploid population in this paper, other population designs can be readily incorporated by considering the appropriate distribution $\pi(Q|\lambda)$. The model considered in this paper assumes no epistatic effect of the markers or QTL. However, epistatic effects can be included as interaction terms in the model. The approach described here looks for QTL in one linkage group only. It is possible to search a genome consisting of distinct linkage groups by a simple extension of these methods by at each cycle selecting a linkage group with probability proportional to its length and then selecting a proposal locus uniformly along that linkage group. In addition, the approach used here can be generalized to consider other distributions of traits, such as t (BESAG *et al.* 1995) and generalized linear models by suitably altering $\pi(y|\theta, Q)$.

Model checking is very critical in any statistical approach, especially in complex hierarchical models as this one. This includes checking sensitivity to the prior and goodness of various model assumptions. Currently, work is under way to use a Bayesian model determination criterion to estimate the number of QTL (s) by incorporating them as further unknowns in the analysis. This would enable one to estimate the posterior distribution of the number of QTL and hence estimate the number of QTL affecting the trait without having to consider Bayes factors to estimate this quantity. We are also working on extending the search to more than one linkage group by considering the entire genome as a whole.

We thank contributions from members of T.O.'s laboratory, especially MARCIO FERREIRA, CARLOS THORMANN and DAVID BUTRUILLE. Discussions with BOB MAU helped clarify various points about the Metropolis-Hastings algorithm for updating the loci. We would also like to acknowledge useful comments from three anonymous refrees that improved the presentation considerably. This work was done when J.S. was at the University of Wisconsin-Madison. Work by B.Y. and J.S. was supported in part by USDA-CSRS 93-3574. Research by T.O. supported in part by USDA-NRICGP 94-37300-0326.

LITERATURE CITED

BERNARDO, J., and A. SMITH, 1994 *Bayesian Theory*. John Wiley and Sons, New York.

BESAG, J., P. GREEN, D. HIGDON and K. MENGERSON, 1995 Bayesian computation and stochastic systems (with discussion). Stat. Sci. **10:** 1–66.

BOX, G., and G. TIAO, 1973 *Bayesian Inference in Statistical Analysis*. Wiley Interscience, New York.

CHURCHILL, G., and R. DOERGE, 1994a Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

DUPUIS, J., 1994 *Statistical Problems Associated with Mapping Complex and Quantitative Traits from Genomic Mismatch Scanning Data*. Ph.D. Dissertation. Stanford University.

FERREIRA, M., J. SATAGOPAN, B. YANDELL, P. WILLIAMS and T. OSBORN, 1995a Mapping loci controlling vernalization requirement and flowering time in *Brassica napus*. Theoret. Appl. Genet. **90:** 727–732.

FERREIRA, M., P. WILLIAMS and T. OSBORN, 1995b RFLP mapping of *Brassica napus* using F1-derived doubled haploid lines. Theoret. Appl. Genet. **89:** 615–621.

FRÜHWIRTH-SCHNATTER, S., 1995 Bayesian model discrimination and Bayes factors for linear Gaussian state space models. J. R. Stat. Soc. B **57:** 237–246.

GELFAND, A. E., and A. F. M. SMITH, 1990 Sampling-based approaches to calculating marginal densities. J. Am. Stat. Assoc. **85:** 398–409.

GEMAN, S., and D. GEMAN, 1984 Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans. Pattern Anal. Machine Intell. **6:** 721–741.

GEYER, C., 1992 Practical Markov chain Monte Carlo. Stat. Sci. **7:** 437–511.

GREEN, P. J., and X. L. HAN, 1992 Metropolis methods, Gaussian proposals, and antithetic variables. Lecture Notes Stat. **74:** 142–164.

HALEY, C., and S. KNOTT, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

HASTINGS, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

HOESCHELE, I., and P. VANRANDEN, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theoret. Appl. Genet. **85:** 953–960.

HOESCHELE, I., and P. VANRANDEN, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theoret. Appl. Genet. **85:** 946–952.

JANSEN, R., 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.

JEFFREYS, H., 1961 *Theory of Probability.* Oxford University Press, London.

KASS, R. E., and A. E. RAFTERY, 1995 Bayes factors. J. Am. Stat. Assoc. **90:** 773–795.

KNAPP, S., W. BRIDGES and D. BIRKES, 1990 Mapping quantitative trait loci using molecular marker linkage maps. Theoret. Appl. Genet. **79:** 583–592.

KNOTT, S., and C. HALEY, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. **60:** 139–151.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

LIU, J. S., W. H. WONG and A. KONG, 1994 Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. Biometrika **81:** 27–40.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theoret. Appl. Genet. **85:** 480–488.

NEWTON, M. A., and A. E. RAFTERY, 1994 Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). J. R. Stat. Soc. B **56:** 3–48.

OTT, J., 1991 *Analysis of Human Genetic Linkage.* (revised ed.). John Hopkins University, Baltimore.

RITTER, C., 1992 *Modern Inference in Nonlinear Least Squares Regression.* Ph.D. Dissertation. University of Wisconsin.

SMITH, A., and G. ROBERTS, 1993 Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. J. R. Stat. Soc. B **55:** 3–23.

SOLLER, M., T. BRODY and A. GENIZI, 1976 On the power of experimental designs for the detection of linkage between markers and quantitative loci in cross between inbred lines. Theoret. Appl. Genet. **47:** 35–39.

THOMAS, D., and V. CORTESSIS, 1992 A Gibbs sampling approach to linkage analysis. Hum. Hered. **42:** 63–76.

TIERNEY, L., 1994 Exploring posterior distributions using Markov chains (with discussion). Ann. Stat. **22:** 1701–1762.

ZENG, Z.-B., 1993 Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

ZENG, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: B. S. WEIR

APPENDIX A

Full conditional densities: The parameters are updated in a cycle using the following full conditional densities.

$$\mu \mid \lambda,\ Q,\ \alpha,\ \sigma^2,\ y$$

$$\sim \mathrm{N}\left(\frac{\dfrac{\eta}{\tau^2} + \dfrac{\Sigma_{i=1}^{n}(y_i - \Sigma_{j=1}^{s}\alpha_j Q_{ij})}{\sigma^2}}{\dfrac{1}{\tau^2} + \dfrac{n}{\sigma^2}},\ \frac{1}{\dfrac{1}{\tau^2} + \dfrac{n}{\sigma^2}}\right), \quad (20)$$

$$\alpha_{j^*} \mid \lambda,\ Q,\ \mu,\ \{\alpha_j\}_{j\neq j^*},\ \sigma^2,\ y$$

$$\sim \mathrm{N}\left(\frac{\Sigma_{i=1}^{n} Q_{ij^*}(y_i - \mu - \Sigma_{j\neq j^*}\alpha_j Q_{ij})}{\sigma^2\left(\dfrac{1}{\tau^2} + \dfrac{\Sigma_{i=1}^{n} Q_{ij^*}^2}{\sigma^2}\right)},\ \frac{1}{\dfrac{1}{\tau^2} + \dfrac{\Sigma_{i=1}^{n} Q_{ij^*}^2}{\sigma^2}}\right),$$

$$j^* = 1, \ldots, s \quad (21)$$

$$\sigma^2 \mid \lambda,\ Q,\ \mu,\ \alpha,\ y$$

$$\sim \mathrm{IG}\left(3 + \frac{n}{2},\ \frac{1}{3 + \dfrac{\Sigma_{i=1}^{n}(y_i - \mu + \Sigma_{j=1}^{s}\alpha_j Q_{ij})^2}{2}}\right), \quad (22)$$

where $\eta = 0$ is the prior mean of μ, and $\tau^2 = 10$ is the prior variance of μ and α_j^*.

APPENDIX B

Missing marker data: Each missing marker genotype is updated individually based on its full conditional. The Bernoulli probability q_{ik} that the kth missing marker for the ith individual is 1 is given by

$$q_{ik} = \pi(M_{ik}^- = 1 \mid \lambda,\ Q,\ \theta,\ M^+,\ y)$$

$$= \pi(M_{ik}^- = 1 \mid \lambda,\ Q_j,\ \theta,\ M_i^+,\ y_i)$$

(by independence of the individuals)

$$= \frac{\pi(M_{ik}^- = 1 \mid \lambda,\ Q_j,\ M_i^+)}{\Sigma_m \pi(M_{ik}^- = m \mid \lambda,\ Q_j,\ M_i^+)}. \quad (23)$$

If k is not a flanking marker for any putative QTL, this Bernoulli probability is given by

$$q_{ik} = \frac{\pi(M_{ik}^- = 1 \mid M_i^+)}{\Sigma_m \pi(M_{ik}^- = m \mid M_i^+)}. \quad (24)$$

Each component in the above probability can be obtained in terms of recombination between marker k and its flanking marker(s) (KNAPP *et al.* 1990).

If marker $k - 1$ (or $k + 1$) and QTL j flank marker k, then q_{ik} can be written in terms of recombinations between $k - 1$ (or $k + 1$), k and j.