Estimating the number of quantitative trait loci via Bayesian model determination

Jaya M. Satagopan
Memorial Sloan-Kettering Cancer Center

Brian S. Yandell
University of Wisconsin

Jaya M. Satagopan, Department of Epidemiology and Biostatistics, MSKCC, 1275 York Ave., New York, NY 10021

Key Words: Posterior distribution, Reversible jump Markov chain Monte Carlo

1 Introduction

Identifying the number of Quantitative Trait Loci (QTL) affecting a trait of interest and their autosomal locations are important research topics among plant breeders and molecular biologists since this will facilitate future research. Various methods have been developed to identify a single QTL. For an extensive bibliography, we refer to Satagopan (1995). Recently, Satagopan et al. (1996) used a Bayesian approach via Markov chain Monte Carlo to simultaneously detect multiple loci affecting the trait. The number of QTL was estimated by fitting various models with different number of QTL and then by comparing these models using Bayes factors (Kass and Raftery 1993). The Bayes factor must be estimated carefully in order to ensure stability of the estimates (Newton and Raftery 1994). In this paper we use a Bayesian model determination approach via MCMC to estimate the number of loci by including the number of QTL as an unknown parameter following Green (1995).

2 QTL Model

Consider a simple linear additive model for a phenotypic trait affected by s unknown QTL. 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. For notation, let $\boldsymbol{Q}_i = \{Q_{ij}\}_{j=1}^{s}$ denote the QTL genotypes for the ith individual, and let $\boldsymbol{\alpha} = \{\alpha_j\}_{j=1}^{s}$ and $\boldsymbol{\delta} = \{\delta_j\}_{j=1}^{s}$ denote the additive and dominance effects of the s loci, respectively. The observed phenotype y_i for the ith individual in a sample of size n may be given by the following linear model :

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

where ϵ_i is a random, mean 0, deviation with variance σ^2, and $(\mu, \boldsymbol{\alpha}, \boldsymbol{\delta})$ determine the expected response given the QTL genotype $\boldsymbol{Q}_i$. The genetic parameters are the QTL loci $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^{s}$ and the model unknown $\boldsymbol{\theta} = (\mu, \boldsymbol{\alpha}, \boldsymbol{\delta}, \sigma^2)$, where λ_j is the distance of the jth QTL from one end of the linkage group.

Assume that a linkage map has been developed for the genome based on m markers with genotypes $\boldsymbol{M}_i = \{M_{ik}\}_{k=1}^{m}$ for the ith individual, with ordered markers $\{1, 2, \cdots, m\}$ and known distances $\boldsymbol{D} = \{D_k\}_{k=1}^{m}$, where D_k is the genetic map distance between markers 1 and k.

In practice, we observe the phenotypic trait $\boldsymbol{y} = \{y_i\}_{i=1}^{n}$ and the marker genotypes $\boldsymbol{M}_i$ but not the QTL genotypes $\boldsymbol{Q}_i$. However, the probability distribution of the QTL genotypes, given the number of QTL, their locations, the marker genotypes and

the distance between the markers, can be modeled in terms of recombination between the loci and the markers.

The likelihood of the parameters s, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ from the ith individual may be expressed as

$$L(s, \boldsymbol{\lambda}, \boldsymbol{\theta} | y_i, \boldsymbol{M}_i, \boldsymbol{D}) = \sum_{\boldsymbol{q}_i} \pi(y_i | s, \boldsymbol{Q}_i = \boldsymbol{q}_i, \boldsymbol{\theta}) \; \pi(\boldsymbol{Q}_i = \boldsymbol{q}_i | s, \boldsymbol{\lambda}, \boldsymbol{M}_i, \boldsymbol{D}) \qquad (2)$$

with the sum over the set of all possible QTL genotypes for the ith individual, $\boldsymbol{q}_i = \{q_{ij}\} \in \{-1, 0, 1\}^s$. $\pi(\boldsymbol{Q}_i | s, \boldsymbol{\lambda}, \boldsymbol{M}_i, \boldsymbol{D})$ is the probability of QTL genotypes given the number of loci, their locations, flanking marker genotypes and distance between the markers. When the data $\boldsymbol{y} = \{y_i\}_{i=1}^n$ are n independent observations, the likelihood can be expressed, after suppressing the notation for conditioning on $\{\boldsymbol{M}_i\}_{i=1}^n$ and $\boldsymbol{D}$, as

$$L(s, \boldsymbol{\lambda}, \boldsymbol{\theta} | \boldsymbol{y}) = \prod_{i=1}^{n} \sum_{\boldsymbol{q}_i} \pi(y_i | s, \boldsymbol{Q}_i = \boldsymbol{q}_i, \boldsymbol{\theta}) \; \pi(\boldsymbol{Q}_i = \boldsymbol{q}_i | s, \boldsymbol{\lambda}) \qquad (3)$$

which is a familiar mixture model likelihood.

Our aim is to make inference about s, $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ using this likelihood which 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 posterior inference summaries for all the components in the model. This is based on the joint posterior of all the unknowns $(s, \boldsymbol{\lambda}, \boldsymbol{Q}, \boldsymbol{\theta})$ given by

$$\pi(s, \boldsymbol{\lambda}, \boldsymbol{\theta}, \boldsymbol{Q} | \boldsymbol{y}) \propto \pi(\boldsymbol{y} | s, \boldsymbol{Q}, \boldsymbol{\theta}) \; \pi(\boldsymbol{Q} | s, \boldsymbol{\lambda}) \; \pi(s, \boldsymbol{\lambda}, \boldsymbol{\theta}) \qquad (4)$$

with $\pi(\boldsymbol{y} | s, \boldsymbol{Q}, \boldsymbol{\theta}) = \prod \pi(y_i | s, \boldsymbol{Q}_i, \boldsymbol{\theta})$ the data probability mass given the QTL genotypes, $\pi(\boldsymbol{Q} | s, \boldsymbol{\lambda}) = \prod \pi(\boldsymbol{Q}_i | s, \boldsymbol{\lambda})$ the probability mass of the QTL genotypes of all the observations given their locations (and $\boldsymbol{M}$ and $\boldsymbol{D}$), and $\pi(s, \boldsymbol{\lambda}, \boldsymbol{\theta})$ the prior density of the genetic parameters. We assume prior independence of $\boldsymbol{\lambda}$ and $\boldsymbol{\theta}$ given s. A natural choice for prior of $\boldsymbol{\lambda}$ (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, $\{\alpha_j\}_{j=1}^s$, $\{\delta_j\}_{j=1}^s$ and σ^2 makes its form simple while increasing diffuseness makes the prior objective. We use a Poisson distribution (with parameter γ) to specify the prior for s, truncated to $s < s_{\max}$ for a suitable choice of $s_{\max}$.

In the Bayesian approach we infer the parameters based on their marginal posterior distribution, which can be obtained from the joint posterior (equation 4) 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.

3 Parameter Estimation

We use MCMC to study the joint posterior density given by equation (4), by constructing a Markov chain with this target distribution. The Markov chain is a random sequence of states

$$(s^0, \boldsymbol{\lambda}^0, \boldsymbol{Q}^0, \boldsymbol{\theta}^0), (s^1 \boldsymbol{\lambda}^1, \boldsymbol{Q}^1, \boldsymbol{\theta}^1), \cdots (s^N \boldsymbol{\lambda}^N, \boldsymbol{Q}^N, \boldsymbol{\theta}^N)$$

started at an arbitrary point $(s^0, \boldsymbol{\lambda}^0, \boldsymbol{Q}^0, \boldsymbol{\theta}^0)$ having positive posterior density, and proceeding by simple rules that modify the unknowns s, $\boldsymbol{\lambda}$, $\boldsymbol{Q}$, and $\boldsymbol{\theta}$.

The dimension of the target distribution (4) changes when s is changed. In this case, standard MCMC theory does not hold. We use the reversible jump algorithm (Green 1995) to move between different models by updating s as in the following steps: (i) a birth step which can increase the number of QTL from s to $s+1$ with probability $b_s = c \; \min(1, p(s+1)/p(s))$; (ii) a death step which can decrease the number of QTL from $s+1$ to s with probability $d_s = c \; \min(1, p(s)/p(s+1))$. Step (iii) updates the genetic parameters and QTL

genotypes (λ, Q, θ) for a specific s with probability $1 - b_s - d_s$. The probabilities b_s and d_s are constrained such that $d_0 = 0$ and $b_{s_{max}} = 0$. c is a constant such that $b_s + d_s \leq 0.9$, $0 \leq s \leq s_{max}$. Steps (i) and (ii) can change the dimension of the model. We use a hybrid sampler (Tierney 1994) to randomly choose one of the above 3 steps at each transition of the chain.

For a given s, the steps updating λ, Q and θ 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.

3.1 Updating the number of QTL

3.1.1 Birth step

For convenience let $x = (s, \lambda, Q, \theta)$. The birth process involves proposing a new QTL, its genotype and the corresponding effect. Denote the proposed parameters of the birth process as

$$x^* = \left(s+1, (\lambda, \lambda_{s+1}), (Q, Q_{s+1}), (\theta, \alpha_{s+1}, \delta_{s+1})\right) .$$

1. Choose an interval for birth not containing any other QTL, with probability $\frac{1}{(m-1-s)}$.

2. 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 $\frac{1}{(D_{k+1} - D_k)}$.

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

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

for $i = 1, \cdots, n$.. M_{ik} and $M_{i,k+1}$ are the flanking marker genotypes for the new QTL. These probabilites can be obtained in terms of recombinations between the QTL and the flanking markers.

4. Sample the effects of the new QTL (α_{s+1} and δ_{s+1}) independently from a normal distribution, say $N(0, \tau^2)$.

5. Set the model mean as $\mu \leftarrow \mu - \alpha_{s+1} \sum_i Q_{i,s+1}/n - \delta_{s+1}(1 - \sum_i |Q_{i,s+1}|/n)$.

Hence, the proposal distribution for the birth step is given by

$$\begin{aligned} q_b &= q(\lambda_{s+1}, Q_{s+1}, \alpha_{s+1}, \delta_{s+1}) \\ &= \left(\frac{1}{m-s}\frac{1}{D_{k+1} - D_k}\right) \times \pi(Q_{s+1} | \lambda_{s+1}) \times \\ &\quad \phi\left(\frac{\alpha_{s+1}}{\tau}\right) \times \phi\left(\frac{\delta_{s+1}}{\tau}\right) , \end{aligned} \tag{5}$$

where $\phi(.)$ is the density of a standard normal distribution, and $\pi(Q_{s+1} | \lambda_{s+1})$ is the probability of the new QTL genotypes given the flanking marker genotypes and intermarker distances as in step 3 above.

3.1.2 Death step

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

1. Choose one of the $s+1$ loci. Drop a locus with uniform probability $\frac{1}{s+1}$. This reduces the number of loci to s.

2. Drop the corresponding effects α_{s+1} and δ_{s+1}, and deterministically update the model mean as $\mu \leftarrow \mu + \alpha_{s+1}(\sum_i Q_{i,s+1}/n) + \delta_{s+1}(1 - \sum_i |Q_{i,s+1}|/n)$.

3. Drop the corresponding genotypes Q_{s+1}.

The proposal for the death step is written as

$$q_d = \frac{1}{s+1} .$$

3.1.3 Acceptance Probability

The acceptance probability of a birth step is given by $\min(1, A)$ where

$$A = \frac{\pi(x^* | y)}{\pi(x | y)} \frac{d_{s+1}}{b_s} \frac{q_d}{q_b} \tag{6}$$

For the death step, the acceptance probability is $\min(1, A^{-1})$ with appropriate relabelling of $\boldsymbol{x}$ and $\boldsymbol{x}^*$. The samples $(s^t, \boldsymbol{\lambda}^t, \boldsymbol{Q}^t, \boldsymbol{\theta}^t)_{t=1}^N$ can be used for posterior inference. In particular, $\{s^t\}_{t=1}^N$ can be used for model selection. In addition, the samples $\boldsymbol{x}_s^t = \{(\boldsymbol{\lambda}^t, \boldsymbol{Q}^t, \boldsymbol{\theta}^t) : s^t = s\}$ can be used to examine the properties of a particular model with s QTL (Green 1995). We again refer the reader to Satagopan (1995) for a bibliography on MCMC methods.

4 Example

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 double haploid line from the *Brassica napus* cv. Stellar was crossed to a single plant of cv. Major. Three groups of 105 DH lines were given 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). 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. EM algorithm for a single QTL model (Lander and Botstein 1989) showed two LOD peaks on linkage group 9 (LOD = 8.37 and 6.91). 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 the two QTL model to best fit the data (Satagopan et al. 1996). We further investigate this using the MCMC Bayesian model determination idea of the previous section.

We use the following model for the number days to flowering for the ith 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.

The overall mean μ, and genetic effects $\boldsymbol{\alpha}$ are given independent Gaussian prior centered at 0 and variance $\tau^2 = 10$, allowing for the possibility of extreme QTL effects. The phenotypic variance σ^2 is assumed to have an inverse gamma prior with parameter 2. The λ_j's are assumed to have uniform prior as described earlier. The number of QTL s has a Poisson prior with parameter $\gamma = 2$. We set $s_{\max} = 5$.

The starting values are chosen as follows. $s^0 = 2$, the 2 loci λ_1^0 and λ_2^0 are at the two ends of the linkage group. Model mean μ^0, and QTL effects α_1^0 and α_2^0 are 0. The variance σ_0^2 is 1. After an initial burn–in of 5,000 states, 200,000 states were sampled at every 50th cycle giving a working set of 4,000 states. Table 1 shows the posterior frequencies of each QTL model visited. It can be observed

s	EPP
0	0.0
1	0.25
2	0.67
3	0.078
4	0.0025
5	0.000125

Table 1: The number of QTL (s) and their estimated posterior probability (EPP).

that the mode of the posterior distribution of s is 2. Table 2 gives the parameter estimates and es-

timated Monte Carlo error when when $s = 2$ (two QTL model). The estimated Monte Carlo error for

Parameter	Estimate	EMC
μ	3.062	0.010
σ^2	0.077	0.011
Effect 1	-0.075	0.030
Locus 1	39.53 cM	0.025
Effect 2	-0.128	0.022
Locus 2	78.02 cM	0.044

Table 2: Parameter estimates and estimated Monte Carlo error (EMC) when $s = 2$.

s is 0.059. Figure 1 shows the marginal posterior densities of the two loci (obtained when the chain sampled a 2 QTL model, i.e. $s^t = 2$). The results are similar to those in Satagopan et al. (1996).

5 Discussion

In this paper we have fit a model which considers the unknown number of QTL as an additional parameter instead of fitting different models by varying s and then comparing these to determine the model that best fits the data. Application to the *Brassica* flowering data shows the results to be similar to those obtained using Bayes factors (Satagopan et al. 1996). Inference for the number of QTL (s) and other parameters is based on the marginal posterior densities.

This MCMC approach has been shown to be efficient in several applications (Green 1995; Mallick 1995). We have presented a preliminary work on application to the QTL problem. Another run of the Markov chain was obtained by starting from a single QTL model ($s^0 = 1$). Starting values of the other parameters were the same as before. 200,000 runs of the chain were subsampled at every 50th interval after an initial burn–in of 5,000 runs. Table 3 shows the marginal posterior probability of s for this run. The mode of the distribution is once again $s = 2$. However, the posterior probability of

s	EPP
0	0.0
1	0.415
2	0.522
3	0.058
4	0.005
5	0.0

Table 3: The number of QTL (s) and their estimated posterior probability (EPP) with $s^0 = 1$

a single QTL ($s = 1$) is higher than before. Estimated posterior probability of the number of QTL with these two different starting values may agree better with much longer runs. Simulations in this vein are currently in progress.

The prior distribution for s, and the birth and death updating schemes are not unique. Currently simulation studies are under way to examine the performance of this approach by considering (i) alternative proposal distributions for the birth and death steps, (ii) different starting values s^0, and (iii) other prior distributions for s.

Acknowledgement

We thank Tom Osborn and Marcio Ferreira for providing the *Brassica* flowering time data. We also acknowledge partial support for research and computing from a USDA hatch grant through the College of Agriculture and Life Sciences, University of Wisconsin– Madison. JMS acknowledges research support from NCI grant CA 69396. Preliminary work for this research was done when JMS was at the University of Wisconsin–Madison.

References

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.

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

Kass, R. E. and A. E. Raftery (1993, March). Bayes factors and model uncertainty. Technical Report 571, Carnegie Mellon University.

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.

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., 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.

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

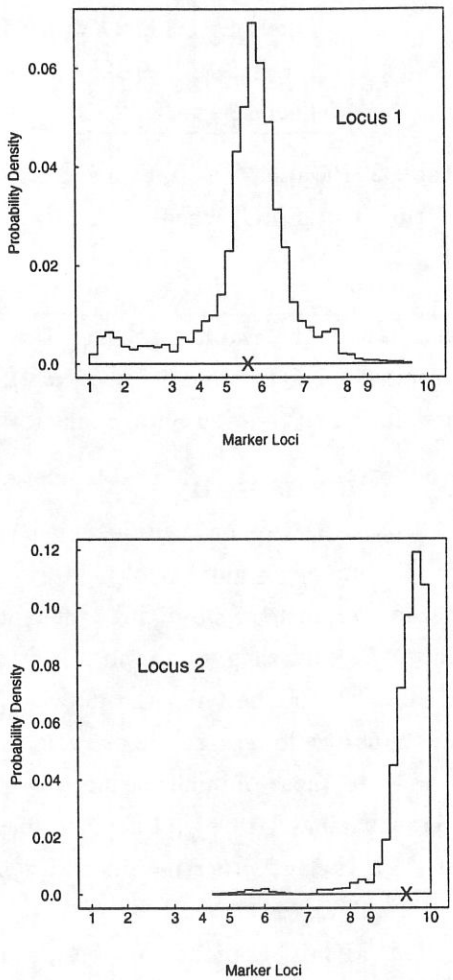

Figure 1: Marginal posterior densities of the two loci λ_1 and λ_2. The densities are obtained based on the samples when the chain selected a 2 QTL model.