Bayesian Mapping of Quantitative Trait Loci Under Complicated Mating Designs

Nengjun Yi and Shizhong Xu

Department of Botany and Plant Sciences, University of California, Riverside, California 92521

Manuscript received April 10, 2000
Accepted for publication December 19, 2000

ABSTRACT

Quantitative trait loci (QTL) are easily studied in a biallelic system. Such a system requires the cross of two inbred lines presumably fixed for alternative alleles of the QTL. However, development of inbred lines can be time consuming and cost ineffective for species with long generation intervals and severe inbreeding depression. In addition, restriction of the investigation to a biallelic system can sometimes be misleading because many potentially important allelic interactions do not have a chance to express and thus fail to be detected. A complicated mating design involving multiple alleles mimics the actual breeding system. However, it is difficult to develop the statistical model and algorithm using the classical maximum-likelihood method. In this study, we investigate the application of a Bayesian method implemented via the Markov chain Monte Carlo (MCMC) algorithm to QTL mapping under arbitrarily complicated mating designs. We develop the method under a mixed-model framework where the genetic values of founder alleles are treated as random and the nongenetic effects are treated as fixed. With the MCMC algorithm, we first draw the gene flows from the founders to the descendants for each QTL and then draw samples of the genetic parameters. Finally, we are able to simultaneously infer the posterior distribution of the number, the additive and dominance variances, and the chromosomal locations of all identified QTL.

THE availability of dense molecular marker maps provides a large opportunity to locate genes responsible for variation of quantitative traits in plants, animals, and humans. Experimental design and methodology are two important issues in quantitative trait loci (QTL) mapping. Most QTL mapping techniques require designed line crosses, *e.g.*, F_2 or backcross (BC). These crosses do not exist in natural populations and are not commonly used in some plant species in the breeding industry. It is not economical to design such line cross experiments solely for the purpose of QTL mapping if these crosses are not regularly used in a breeding program. Almost all natural populations and most domesticated plant populations consist of complicated pedigree structures. Even if inbred lines are used, different crosses may be connected by some common ancestors. A mating design combining information from multiple crosses is more powerful than one involving a single cross (Muranty 1996; Xu 1998). Multiple crosses increase the polymorphic levels of QTL alleles and may permit the detection of QTL that are undetectable in a single line cross. Animal and human geneticists have paid considerable attention to the relative power of simple pedigree analysis and more complicated family structures in linkage analysis of QTL and found that large complex pedigree designs are usually more powerful (*e.g.*, Weller *et al.* 1990; Wijsman and Amos 1997; Slate *et al.* 1999). The main reasons behind this increased power are: (1) a complex pedigree increases the chance that founder alleles are equally represented in the mapping population; and (2) there are more informative meioses in a complex pedigree than in a simple pedigree for the same number of genotypes.

A variety of methods have been developed for QTL mapping (Hoeschele *at al.* 1997; Lynch and Walsh 1998). These methods can be classified into three categories: least-squares analysis (LS), maximum-likelihood analysis (ML), and Bayesian analysis. These methods differ in computational requirement, efficiency in terms of extracting information, flexibility with regard to handling different data structures, and ability in mapping multiple QTL. The simple LS method is efficient in terms of computational speed, but cannot extract all information from the data and is restricted to specific mating designs. ML interval mapping (Lander and Botstein 1989) is one of the most widely used methods for QTL analysis in a single cross. The interval mapping method has been extended to composite interval mapping and multiple interval mapping (Jansen 1993; Zeng 1993; Kao *et al.* 1999). These extensions are designed particularly for mapping multiple QTL in a single line cross. However, it is not straightforward to apply these methods to QTL mapping in general pedigrees. The identical-by-descent-based variance component method can be applied to general pedigrees (Almasy and Blangero 1998). This method not only incorporates full pedigree information but also is robust to the number of QTL alleles. In addition, it does not require the knowledge of marker linkage phases (Schork 1993;

Corresponding author: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: xu@genetics.ucr.edu

Amos 1994; Xu and Atchley 1995). The identical-by-descent (IBD)-based variance component approach has become a very useful strategy for QTL mapping in humans. If the linkage phase information is indeed known, by ignoring such information, the IBD method may be suboptimal. It is also questionable to apply this method to plants where the pedigree sizes are usually large due to the need to invert large IBD matrices repeatedly for each QTL position considered.

Bayesian analysis is preferable because of its convenience and flexibility in the use of full pedigrees and mapping multiple QTL, although it is computationally very demanding. Bayesian mapping fully takes into account the uncertainties associated with all unknowns in the QTL mapping problem, including the number and locations of QTL, effects of QTL, and the genotypes of markers and QTL. In plant line-crossing experiments, Bayesian mapping has been developed by using the Markov chain Monte Carlo algorithm, in particular, for detection of multiple QTL (Satagopan and Yandell 1996; Satagopan *et al.* 1996; Sillanpää and Arjas 1998, 1999; Stephens and Fisch 1998; Yi and Xu 2000). In animals and humans, Bayesian mapping has been designed not only to map multiple QTL, but also to extract full pedigree information (Heath 1997; Uimari and Hoeschele 1997). However, most existing Bayesian mapping methods assume a biallelic QTL model. Although this assumption may be reasonable for a single line cross, it is less so for complex pedigrees. Therefore, Bayesian mapping needs to be extended to multiallelic systems.

There are several differences between plants and animals or humans in the context of general pedigrees: (i) many plant species are self-compatible and one must deal with a system that involves a mixture of selfing and outcrossing; (ii) inbreeding and line crossing are common mating designs in most plant breeding populations; (iii) a breeding population of plants usually contains fewer founders than an animal or human population and the founders can be pure inbred lines; and (iv) family sizes of plants are usually large compared with animals and humans. These differences provide additional opportunities for detecting QTL. Unfortunately, the Bayesian mapping methods developed for human and animal pedigrees cannot handle the unique properties for plant pedigrees. This poses unique challenges for plant geneticists to develop new QTL mapping statistics.

In this article, we develop a Bayesian method of QTL mapping under arbitrarily complicated mating designs, including a group of independent or related F_2 or backcross populations and complicated multiple-generation cross populations derived from inbred or outbred founders. The method is so flexible that it can handle a variety of genetic models, such as arbitrary number of QTL alleles, dominance effects, and fixed and random models. The Bayesian method is implemented via a reversible jump Markov chain Monte Carlo (MCMC) algorithm, which allows simultaneous estimation of the number, the locations, and effects of identified QTL.

STATISTICAL METHODS

Mixed model: Assume that the mapping population consists of *n* individuals with arbitrary pedigree relationships, and among the *n* individuals there are *m* founders and $(n - m)$ nonfounders. The whole population may consist of a single large pedigree or multiple independent pedigrees. A founder in a pedigree is defined as an individual with no parents included in the pedigree. In contrast, a nonfounder is defined as an individual with both parents included in the pedigree. Founders are assumed to be unrelated but can be inbred. The descendants may be related in an arbitrary way.

Let $\mathbf{y}$ represent an $n \times 1$ vector for the observed values of a quantitative trait. When the trait is controlled by multiple genes acting independently, $\mathbf{y}$ can be described by the linear model

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{j=1}^{l}(\mathbf{u}_j^{\mathrm{p}} + \mathbf{u}_j^{\mathrm{m}} + \mathbf{v}_j) + \mathbf{e}, \tag{1}$$

where $\mathbf{X}$ is a known design matrix for a vector of nongenetic effects $\mathbf{b}$ (including the overall mean), l is the number of QTL on all chromosomes, $\mathbf{u}_j^{\mathrm{p}}$ and $\mathbf{u}_j^{\mathrm{m}}$ are $n \times 1$ vectors for the paternal and maternal allelic effects for the jth QTL, $\mathbf{v}_j$ is an $n \times 1$ vector for the dominance effects for the jth QTL, and $\mathbf{e}$ is the vector of residual (environmental) effects. This model is written in the original form of the animal model (Fernando and Grossman 1989) except that we have included the dominance effects.

Denote $\mathbf{a}_j$ as a $2m \times 1$ vector for the effects of the founder alleles (with *m* ancestors, each with two alleles) and $\mathbf{d}_j$ as a vector of interaction effects (dominance effects) between all possible pairs of the $2m$ founder alleles at the jth QTL. The dimension of $\mathbf{d}_j$ is $m(2m + 1)$. The dimension of $\mathbf{d}_j$ can be reduced greatly in some mating designs where it is impossible for some founder alleles to be combined in any descendant. The QTL effects of all individuals can be expressed as linear functions of the allelic effects and their interactions in the founders, *i.e.*, $\mathbf{u}_j^{\mathrm{p}} = \mathbf{Z}_j^{\mathrm{p}}\mathbf{a}_j$, $\mathbf{u}_j^{\mathrm{m}} = \mathbf{Z}_j^{\mathrm{m}}\mathbf{a}_j$, and $\mathbf{v}_j = \mathbf{W}_j\mathbf{d}_j$, leading to

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \sum_{j=1}^{l}(\mathbf{Z}_j^{\mathrm{p}} + \mathbf{Z}_j^{\mathrm{m}})\mathbf{a}_j + \sum_{j=1}^{l}\mathbf{W}_j\mathbf{d}_j + \mathbf{e}. \tag{2}$$

This model is written in the form of a reduced animal model (Cantet and Smith 1991) in which each allele is traced back to one of the founder alleles through $n \times 2m$ matrices $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$, also called the allelic inheritance matrices. Note that the $n \times m(2m + 1)$ dominance design matrix $\mathbf{W}_j$ is a function of $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$. The allelic inheritance matrices are not observable, but their distributions are deduced from molecular markers linked to the chromosome on which the jth QTL resides. Therefore, the distributions of $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$ are functions of marker information and the chromosomal location of the jth QTL. The environmental effects are assumed to follow a $N(\mathbf{0}, \mathbf{I}\sigma_{\mathrm{e}}^2)$ distribution.

The observables in model (2) include the phenotypic values, $\mathbf{y} = \{y_i\}_{i=1}^n$, the covariate $\mathbf{X}$, and the marker data $\mathbf{M}^*$. The marker data include the locations of markers on chromosomes and the observed (possibly incomplete) marker genotypes. The observed marker genotypes in some individuals may not be fully informative and the patterns of allelic inheritance of such markers may also be unknown. The list of unobservables includes the number of QTL l, the QTL locations $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^l$, the complete marker genotype matrix $\mathbf{M}$, the QTL allelic inheritance matrices $\mathbf{Z}^{\mathrm{p}} = \{\mathbf{Z}_j^{\mathrm{p}}\}_{j=1}^l$ and $\mathbf{Z}^{\mathrm{m}} = \{\mathbf{Z}_j^{\mathrm{m}}\}_{j=1}^l$, the QTL allelic effects $\mathbf{a} = \{\mathbf{a}_j\}_{j=1}^l$, the QTL dominance effects $\mathbf{d} = \{\mathbf{d}_j\}_{j=1}^l$, and the residual variance σ_{e}^2. The location parameter, λ_j, is expressed as the distance of the jth QTL from one end of the chromosome. The complete marker genotype means marker genotype with known linkage phase. A complete genotype for a nonfounder means a known allelic inheritance pattern. The QTL dominance design matrices are suppressed in the list of unknowns because they are completely determined by the QTL allelic inheritance matrices.

In a Bayesian framework, the unknowns in the model are considered to be drawn from appropriate prior distributions. The joint posterior distribution of all unobservables $\boldsymbol{\theta} = \{l, \boldsymbol{\lambda}, \mathbf{a}, \mathbf{d}, \mathbf{b}, \mathbf{M}, \mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{m}}, \sigma_{\mathrm{e}}^2\}$ given the observables $\{\mathbf{y}, \mathbf{X}, \mathbf{M}^*\}$ and prior information can be expressed as

$$
\begin{aligned}
p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{X}, \mathbf{M}^*) \propto\ & p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{m}}, \mathbf{M}|l, \boldsymbol{\lambda}, \mathbf{M}^*) \\
& \times p(l)p(\mathbf{b})p(\sigma_{\mathrm{e}}^2) \\
& \times \prod_{j=1}^{l}\{p(\mathbf{a}_j)p(\mathbf{d}_j)p(\lambda_j)\}.
\end{aligned} \tag{3}
$$

The likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$ depends on the distribution of $\mathbf{y}$. For normally distributed traits, it has the form

$$
p(\mathbf{y}|\boldsymbol{\theta}) \propto (\sigma_{\mathrm{e}}^2)^{-n/2} \times \exp\left\{-\frac{1}{2\sigma_{\mathrm{e}}^2}\mathbf{R}^{\mathrm{T}}\mathbf{R}\right\}, \tag{4}
$$

where $\mathbf{R} = \mathbf{y} - \mathbf{Xb} - \Sigma_{j=1}^l(\mathbf{Z}_j^{\mathrm{p}} + \mathbf{Z}_j^{\mathrm{m}})\mathbf{a}_j - \Sigma_{j=1}^l\mathbf{W}_j\mathbf{d}_j$.

The priors of the allelic and dominance effects of QTL, $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$, depend on the inbreeding coefficients of the founders (see APPENDIX). The inclusion of inbreeding coefficients of founders enables the proposed method to treat inbred founders. The prior distribution of the number of QTL, $p(l)$, is assumed to be a truncated Poisson distribution with mean μ and a predefined maximum number $l_{\max}$. When no information regarding the locations is available, the prior probability that a QTL is on a chromosome is proportional to the length of the chromosome. Within a chromosome, each QTL has a uniform distribution of residing at any location on that chromosome. The prior distributions of $\mathbf{b}$ and σ_{e}^2 are assumed to be uniform on predefined intervals, although other priors can be used. Finally, $p(\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{m}}, \mathbf{M}|l, \boldsymbol{\lambda}, \mathbf{M}^*)$ is the joint conditional distribution of QTL allelic inheritance matrices and complete marker genotypes.

In some situations, we need to add an extra layer to the hierarchical model. The distributions $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$ depend on other unknown quantities $\sigma_{\mathrm{a}_j}^2$ and $\sigma_{\mathrm{d}_j}^2$, the allelic and dominance variance of the jth QTL. In other words, we replace $p(\mathbf{a}_j)$ by $p(\mathbf{a}_j, \sigma_{\mathrm{a}_j}^2) = p(\mathbf{a}_j|\sigma_{\mathrm{a}_j}^2)p(\sigma_{\mathrm{a}_j}^2)$ and $p(\mathbf{d}_j)$ by $p(\mathbf{d}_j, \sigma_{\mathrm{d}_j}^2) = p(\mathbf{d}_j|\sigma_{\mathrm{d}_j}^2)p(\sigma_{\mathrm{d}_j}^2)$. The parameters of interest now are $\sigma_{\mathrm{a}_j}^2$ and $\sigma_{\mathrm{d}_j}^2$, with $\mathbf{a}_j$ and $\mathbf{d}_j$ being treated as missing values. The joint posterior distribution of all variables is then factorized as

$$
\begin{aligned}
p(\boldsymbol{\theta}, \mathbf{V}_{\mathrm{a}}, \mathbf{V}_{\mathrm{d}}|\mathbf{y}, \mathbf{X}, \mathbf{M}^*) \propto\ & p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{Z}^{\mathrm{p}}, \mathbf{Z}^{\mathrm{m}}, \mathbf{M}|l, \boldsymbol{\lambda}, \mathbf{M}^*) \\
& \times p(l)p(\mathbf{b})p(\sigma_{\mathrm{e}}^2) \\
& \times \prod_{j=1}^{l}\{p(\mathbf{a}_j|\sigma_{\mathrm{a}_j}^2)p(\sigma_{\mathrm{a}_j}^2)p(\mathbf{d}_j|\sigma_{\mathrm{d}_j}^2) \\
& \qquad \times p(\sigma_{\mathrm{d}_j}^2)p(\lambda_j)\},
\end{aligned} \tag{5}
$$

where $\mathbf{V}_{\mathrm{a}} = \{\sigma_{\mathrm{a}_j}^2\}_{j=1}^l$, $\mathbf{V}_{\mathrm{d}} = \{\sigma_{\mathrm{d}_j}^2\}_{j=1}^l$, and the distributions $p(\mathbf{a}_j|\sigma_{\mathrm{a}_j}^2)$ and $p(\mathbf{d}_j|\sigma_{\mathrm{d}_j}^2)$ are multivariate normal as given in the APPENDIX. We use uniform prior distributions for $p(\sigma_{\mathrm{a}_j}^2)$ and $p(\sigma_{\mathrm{d}_j}^2)$ within some predetermined intervals. Other terms in Equation 5 are the same as in Equation 3.

Equations 3 and 5 correspond to two different approaches in QTL mapping, *i.e.*, the fixed-model and the random-model approaches, respectively. If there are only a few founders who are not randomly sampled from a large reference population, our interest may be only in the values of the actual allelic effects and the dominance effects for the founders at hand. Under the fixed-model approach the priors for the allelic and dominance effects are treated as variable with known distributions. The fixed-model approach is very common in designed line-crossing experiments, *e.g.*, F_2 and BC designs, where the average effect of allelic substitution is the parameter of interest. When the founders are randomly sampled from a reference population, we are usually interested in the variances of the genetic effects in the population from which the founders are sampled. In this case, the distributions for the allelic and dominance effects of founders depend on some unknown parameters. When the number of founders is so small that a meaningful estimate of the allelic or dominance variance cannot be inferred from the limited number of alleles sampled, we may still use the fixed-model approach, even if the founders are a random sample. In this study, we concentrate on the random model approach.

Reversible jump MCMC: In Bayesian analysis, inferences about the parameters of interest are based on the joint posterior distribution of all unknowns. Since the joint posterior distribution does not have a standard form, MCMC samplers are used to generate samples from the joint posterior distribution (Metropolis *et al.* 1953; Hastings 1970; Geman and Geman 1984; Green 1995). The MCMC algorithm consists of the following steps:

a. Updating QTL allelic effects $\mathbf{a} = \{\mathbf{a}_j\}_{j=1}^{l}$;
b. Updating QTL dominance effects $\mathbf{d} = \{\mathbf{d}_j\}_{j=1}^{l}$;
c. Updating QTL allelic variances $\mathbf{V}_{\mathrm{a}} = \{\sigma^2_{\mathrm{a}_j}\}_{j=1}^{l}$;
d. Updating QTL dominance variances $\mathbf{V}_{\mathrm{d}} = \{\sigma^2_{\mathrm{d}_j}\}_{j=1}^{l}$;
e. Updating the fixed effects $\mathbf{b}$ and residual variance σ^2_{e};
f. Updating complete marker genotypes $\mathbf{M}$ and QTL allelic inheritance matrices $\mathbf{Z}^{\mathrm{p}}$ and $\mathbf{Z}^{\mathrm{m}}$;
g. Updating QTL locations $\boldsymbol{\lambda} = \{\lambda_j\}_{j=1}^{l}$;
h. Birth of a QTL (adding one new QTL to the model) or death of a QTL (removing one existing QTL from the model).

The proposed algorithm starts from an initial point and proceeds to update each of the unknowns in turn. One complete pass over these eight update steps defines a cycle of iteration. Updating steps (a)–(g) are conventional and do not alter the dimension of the variable vector. We use Metropolis-Hastings algorithms to implement steps (a)–(e) and (g), and the Gibbs sampler to update step (f). Step (h) involves changing QTL number by one and making necessary corresponding changes to ($\mathbf{a}$, $\mathbf{d}$, $\mathbf{V}_{\mathrm{a}}$, $\mathbf{V}_{\mathrm{d}}$, $\mathbf{Z}^{\mathrm{p}}$, $\mathbf{Z}^{\mathrm{m}}$, $\boldsymbol{\lambda}$). A reversible jump step is needed to change the number of QTL.

Several methods have been available for updating marker genotypes in general pedigrees (*e.g.*, Sobel and Lange 1996; Heath 1997; Uimari and Hoeschele 1997; Bink and Van Arendonk 1999). In the simulation study (see the next section), we adopt the method of Bink and Van Arendonk (1999) to update the marker genotypes. For more complicated situations, a descent graph sampler of Sobel and Lange (1996) is needed (see discussion). Updating the fixed effects $\mathbf{b}$ and residual variance σ^2_{e} is also straightforward.

Updating QTL effects: The allelic effects of QTL are updated founder by founder and locus by locus. Denote $\mathbf{a}_k^j$ as a 2 × 1 vector for the allelic effects of the *k*th founder at the *j*th QTL (appendix). To update $\mathbf{a}_k^j$, two random variables, δ_1 and δ_2, are simulated independently from the symmetric uniform distribution around zero (random walk). The length of this uniform distribution is determined empirically and should result in a reasonable rate of average acceptance rate. A new proposal value of $\mathbf{a}_k^j$ takes $\mathbf{a}_k^{j*} = \mathbf{a}_k^j + (\delta_1, f_k\delta_1 + (1 - f_k)\delta_2)^{\mathrm{T}}$, where f_k is the inbreeding coefficient of the *k*th founder. The new proposal is accepted with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\mathbf{a}_k^{j*}|\sigma^2_{\mathrm{a}_j})}{p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{a}_k^j|\sigma^2_{\mathrm{a}_j})}\right\}, \tag{6}$$

where $\boldsymbol{\theta}^*$ means all elements of $\boldsymbol{\theta}$ except that $\mathbf{a}_k^j$ is replaced by the proposal $\mathbf{a}_k^{j*}$.

The dominance effects of QTL are updated for two founders at a time, again in a locus-by-locus basis. Denote $\mathbf{d}_{kk'}^j$ as a vector of dominance effects between alleles of founder k and alleles of founder k' at the *j*th QTL. The dimension of $\mathbf{d}_{kk'}^j$ is three or four, depending on whether k equals k' (see the appendix). To update $\mathbf{d}_{kk'}^j$, a new proposal $\mathbf{d}_{kk'}^{j*}$ is simulated by random walk, denoted by

$$\mathbf{d}_{kk'}^{j*} = \mathbf{d}_{kk'}^j + (\delta_1, f_k\delta_1 + (1 - f_k)\delta_2, f_k\delta_1 + (1 - f_k)\delta_4)^{\mathrm{T}}$$

if $k' = k$; otherwise,

$$\begin{aligned}\mathbf{d}_{kk'}^{j*} = \mathbf{d}_{kk'}^j &+ (\delta_1, f_k\delta_1 + (1 - f_k)\delta_2, f_{k'}\delta_1 \\ &+ (1 - f_{k'})\delta_3, f_kf_{k'}\delta_1 + (1 - f_k)f_{k'}\delta_2 \\ &+ f_k(1 - f_{k'})\delta_3 + (1 - f_k)(1 - f_{k'})\delta_4)^{\mathrm{T}},\end{aligned}$$

where δ_1, δ_2, δ_3, and δ_4 are sampled independently from the symmetric uniform distribution around zero. The new proposal is accepted with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)p(\mathbf{d}_{kk'}^{j*}|\sigma^2_{\mathrm{d}_j})}{p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{d}_{kk'}^j|\sigma^2_{\mathrm{d}_j})}\right\}, \tag{7}$$

where $\boldsymbol{\theta}^*$ means all elements of $\boldsymbol{\theta}$ except that $\mathbf{d}_{kk'}^j$ is replaced by the proposal $\mathbf{d}_{kk'}^{j*}$.

Updating QTL variances: QTL allelic and dominance variances are updated locus by locus. To update $\sigma^2_{\mathrm{a}_j}$ and $\sigma^2_{\mathrm{d}_j}$, new proposals $\sigma^{2^*}_{\mathrm{a}_j}$ and $\sigma^{2^*}_{\mathrm{d}_j}$ are sampled from the symmetric uniform densities around their previous values. The proposals are accepted with probabilities

$$\min\left\{1, \frac{\prod_{k=1}^{\mathrm{m}} p(\mathbf{a}_k^j|\sigma^{2^*}_{\mathrm{a}_j})}{\prod_{k=1}^{\mathrm{m}} p(\mathbf{a}_k^j|\sigma^2_{\mathrm{a}_j})}\right\} \quad \text{and} \quad \min\left\{1, \frac{\prod_{k=1}^{\mathrm{m}}\prod_{k'=k}^{\mathrm{m}} p(\mathbf{d}_{kk'}^j|\sigma^{2^*}_{\mathrm{d}_j})}{\prod_{k=1}^{\mathrm{m}}\prod_{k'=k}^{\mathrm{m}} p(\mathbf{d}_{kk'}^j|\sigma^2_{\mathrm{d}_j})}\right\}, \tag{8}$$

respectively.

Updating QTL allelic inheritance $\mathbf{Z}^p$ *and* $\mathbf{Z}^m$*:* Each allele in the descendants can be traced back to one of the founder alleles. This is reflected by $\mathbf{a}_j^{\mathrm{p}} = \mathbf{Z}_j^{\mathrm{p}}\mathbf{a}_j$ and $\mathbf{a}_j^{\mathrm{m}} = \mathbf{Z}_j^{\mathrm{m}}\mathbf{a}_j$, where each row of matrices $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$ has one element taking 1 and all other elements being 0. It is not convenient to generate realizations of $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$ directly, but we can easily generate a sample of $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$ indirectly through the following recursive approach.

Consider a pedigree with m founders. Let the $2m$ founder alleles be numbered consecutively from 1 to $2m$. Then allele $2k - 1$ and $2k$ are the two alleles of the *k*th founder. Assume that individuals are entered into the pedigree in a chronological order so that the parents are evaluated before their progeny. Denote $\mathbf{Z}_j^{\mathrm{p}}(i)$ and $\mathbf{Z}_j^{\mathrm{m}}(i)$ as the *i*th rows of $\mathbf{Z}_j^{\mathrm{p}}$ and $\mathbf{Z}_j^{\mathrm{m}}$, respectively; *i.e.*, $\mathbf{Z}_j^{\mathrm{p}}$ (i) and $\mathbf{Z}_j^{\mathrm{m}}(i)$ store the allele identifications of the paternal and maternal alleles of individual i, respectively. For example, if the paternal allele of individual i is traced back to allele 3 of the founders and the maternal allele is traced back to allele 10 of the founders, then the third element of $\mathbf{Z}^{p}_{j}(i)$ is 1 and all other elements are 0, and the tenth element of $\mathbf{Z}^{m}_{j}(i)$ is 1 and all other elements are 0. We now describe the recurrent process of building matrices $\mathbf{Z}^{p}_{j}$ and $\mathbf{Z}^{m}_{j}$. Assume that we have already built matrices $\mathbf{Z}^{p}_{j}$ and $\mathbf{Z}^{m}_{j}$ up to the first $(i - 1)$th rows and are ready to build the ith rows. If the ith individual is a founder, say the kth founder, then the $(2k - 1)$th element of $\mathbf{Z}^{p}_{j}(i)$ and the $(2k + 1)$th element of $\mathbf{Z}^{m}_{j}(i)$ are 1. If the ith individual is not a founder but the progeny of individuals i_1 (father) and i_2 (mother), then

$$\mathbf{Z}^{p}_{j}(i) = u^{p}_{ij}\mathbf{Z}^{p}_{j}(i_1) + (1 - u^{p}_{ij})\mathbf{Z}^{m}_{j}(i_1)$$

and

$$\mathbf{Z}^{m}_{j}(i) = u^{m}_{ij}\mathbf{Z}^{p}_{j}(i_2) + (1 - u^{m}_{ij})\mathbf{Z}^{m}_{j}(i_2),$$

where u^{p}_{ij} and u^{m}_{ij} are the paternal and maternal segregation (meiosis) indicators, respectively, for individual i at the jth QTL. If the paternal allele of the father is passed to individual i, then $u^{p}_{ij} = 1$; otherwise, $u^{p}_{ij} = 0$. The value of u^{m}_{ij} is similarly defined but for the allelic inheritance of the mother. Note that $\mathbf{Z}^{p}_{j}(i_1)$, $\mathbf{Z}^{m}_{j}(i_1)$, $\mathbf{Z}^{p}_{j}(i_2)$, and $\mathbf{Z}^{m}_{j}(i_2)$ have been previously built because $i_1 \leq i - 1$ and $i_2 \leq i - 1$. Therefore, to trace the allelic origin, one only needs to simulate the segregation indicators for each descendant.

Simulating the segregation indicators (u^{p}_{ij}, u^{m}_{ij}) is straightforward. The segregation indicators (u^{p}_{ij}, u^{m}_{ij}) can take four possible values, *i.e.*, (1, 1), (1, 0), (0, 1), and (0, 0). The conditional posterior distribution is thus a discrete distribution over the four possible allelic inheritance patterns and depends on the position of the QTL, the segregation indicators of flanking loci (markers or QTL), the phenotypic value of the progeny, and other parameter values in the model. We then sample a value from the posterior distribution and convert the segregation indicators into the design matrices using the recursive equation. The recursive algorithm for QTL allelic inheritance can be applied to complicated designs with mixture of outcrossing and selfing.

Updating QTL locations: Similar to the method of Sillanpää and Arjas (1998, 1999), we do not fix the order of QTL when updating the QTL locations. Elements of $\boldsymbol{\lambda}$ are modified one at a time using the Metropolis algorithm. For the jth QTL, a proposal λ^{*}_{j} is sampled from a symmetric uniform distribution in the neighborhood of the previous value λ_j. In the meantime, new proposals for the segregation indicator matrices, denoted by $\mathbf{U}^{p*}_{j}$ and $\mathbf{U}^{m*}_{j}$, are generated according to the method of updating QTL allelic inheritance matrices. The new allelic inheritance matrices, $\mathbf{Z}^{p*}_{j}$ and $\mathbf{Z}^{m*}_{j}$, are calculated using the recursive equations. Denote the generating distribution of $(\mathbf{U}^{p*}_{j}, \mathbf{U}^{m*}_{j})$ by $q(\mathbf{U}^{p*}_{j}, \mathbf{U}^{m*}_{j})$. The proposals are accepted with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^{*}, \lambda^{*}_{j}, \mathbf{Z}^{p*}_{j}, \mathbf{Z}^{m*}_{j})}{p(\mathbf{y}|\boldsymbol{\theta})} \times \frac{p(\mathbf{U}^{p*}_{j}, \mathbf{U}^{m*}_{j}|\lambda^{*}_{j}, \mathbf{G}^{L}_{j}, \mathbf{G}^{R}_{j})}{p(\mathbf{U}^{p}_{j}, \mathbf{U}^{m}_{j}|\lambda_{j}, \mathbf{G}^{L}_{j}, \mathbf{G}^{R}_{j})} \times \frac{q(\mathbf{U}^{p}_{j}, \mathbf{U}^{m}_{j})}{q(\mathbf{U}^{p*}_{j}, \mathbf{U}^{m*}_{j})}\right\}, \tag{9}$$

where $\boldsymbol{\theta}^{*}$ contains all elements of $\boldsymbol{\theta}$ except $\boldsymbol{\lambda}_j$, $\mathbf{Z}^{p}_{j}$, $\mathbf{Z}^{m}_{j}$. Let $\mathbf{G}^{L}_{j}(\mathbf{G}^{R}_{j})$ denote the complete genotypes of all pedigree members at the left (right) flanking locus of the corresponding location. If the proposals are accepted, we update the location of the jth QTL and also modify the segregation indicators, the allelic inheritance, and dominance design matrices at the jth QTL at the same time.

Updating QTL number: The reversible jump mechanism is needed to change the QTL number in the model. In this study, a reversible pair is used: birth/death of a QTL. In every cycle of the simulation, we make a random choice between attempting to add one new QTL into the model or delete one existing QTL from the model, with probabilities p_{a} and $p_{d} = 1 - p_{a}$, respectively. Of course, $p_{a} = 0$ if $l = l_{\max}$ and $p_{d} = 0$ if $l = 0$, and otherwise we choose $p_{a} = 0.5$, for $0 < l < l_{\max}$.

For a birth step, we need to generate a new location λ_{l+1}, a new allelic variance $\sigma^{2}_{a_{l+1}}$, a new dominance variance $\sigma^{2}_{d_{l+1}}$, a new vector of allelic effects of the founder alleles $\mathbf{a}_{l+1}$, a new vector of dominance effects between all possible pairs of the founder alleles $\mathbf{d}_{l+1}$, and new inheritance and dominance design matrices of all pedigree members $\mathbf{Z}^{p}_{l+1}$, $\mathbf{Z}^{m}_{l+1}$, and $\mathbf{W}_{l+1}$ for the new QTL. The new location λ_{l+1} and the variances $\sigma^{2}_{a_{l+1}}$ and $\sigma^{2}_{d_{l+1}}$ are sampled from the corresponding prior densities. The allelic effects $\mathbf{a}_{l+1}$ and the dominance effects $\mathbf{d}_{l+1}$ are then simulated from the distributions $p(\mathbf{a}_{l+1}|\sigma^{2}_{a_{l+1}})$ and $p(\mathbf{d}_{l+1}|\sigma^{2}_{d_{l+1}})$ described in the appendix. The segregation indicator matrices, denoted as $\mathbf{U}^{p}_{l+1}$ and $\mathbf{U}^{m}_{l+1}$, are generated from $q(\mathbf{U}^{p}_{l+1}, \mathbf{U}^{m}_{l+1})$ using the method of updating QTL allelic inheritance matrices, and new inheritance and dominance design matrices $\mathbf{Z}^{p}_{l+1}$, $\mathbf{Z}^{m}_{l+1}$, and $\mathbf{W}_{l+1}$ are then calculated. The proposal is accepted with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^{*})}{p(\mathbf{y}|\boldsymbol{\theta})} \times \frac{\mu \times p(\mathbf{U}^{p}_{l+1}, \mathbf{U}^{m}_{l+1},|\lambda_{l+1}, \mathbf{G}^{L}_{l+1}, \mathbf{G}^{R}_{l+1})}{l + 1} \times \frac{p_{d}/(l + 1)}{p_{a} \times q(\mathbf{U}^{p}_{l+1}, \mathbf{U}^{m}_{l+1})}\right\}, \tag{10}$$

where $\boldsymbol{\theta}^{*} = (\boldsymbol{\theta}, \lambda_{l+1}, \mathbf{a}_{l+1}, \mathbf{d}_{l+1}, \mathbf{Z}^{p}_{l+1}, \mathbf{Z}^{m}_{l+1})$ with l in $\boldsymbol{\theta}$ replaced by $(l + 1)$; $\mathbf{G}^{L}_{l+1}(\mathbf{G}^{R}_{l+1})$ denotes the complete genotypes of all pedigree members at the left (right) flanking locus of the location λ_{l+1}.

The death step is somewhat simpler. A random choice is made among the existing QTL, and the chosen QTL is then proposed to delete from the model. If the jth existing QTL is proposed to delete, the acceptance probability for the deletion is

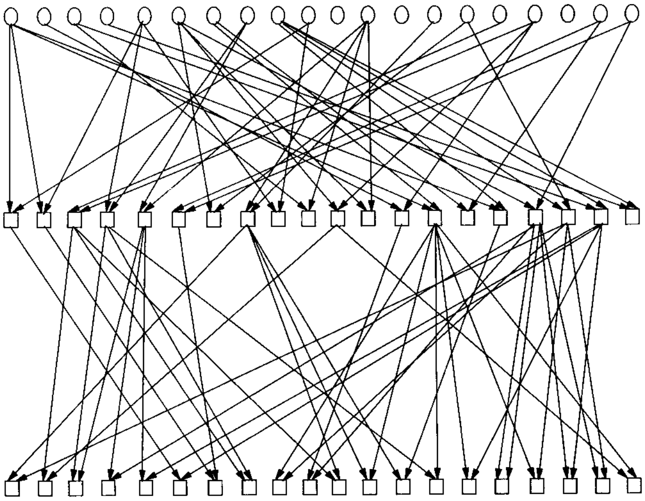

Figure 1.—The simulated pedigree consisting of 2020 individuals over one base generation (founders) and two descendant generations. The 20 open circles in the first row represent 20 founders sampled from a large random base population. By random mating (including selfing), the 20 founders form 20 full-sib families (open squares in the second row), each with 50 sibs, making a total of 1000 sibs in the F_1 generation. Among the 1000 F_1 individuals, 20 were randomly selected to form 20 full-sib families of the F_2 generation (open squares in the third row), each family consisting of 50 sibs, leading to 1000 F_2 individuals.

$$\min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}^*)}{p(\mathbf{y}|\boldsymbol{\theta})} \times \frac{l}{\mu \times p(\mathbf{U}_j^{\mathrm{p}}, \mathbf{U}_j^{\mathrm{m}}|\lambda_j, \mathbf{G}_j^{\mathrm{L}}, \mathbf{G}_j^{\mathrm{R}})} \times \frac{p_{\mathrm{a}} \times q(\mathbf{U}_j^{\mathrm{p}}, \mathbf{U}_j^{\mathrm{m}})}{p_{\mathrm{d}}/l}\right\}, \tag{11}$$

where $\boldsymbol{\theta}^*$ means all elements of $\boldsymbol{\theta}$ except the items corresponding to the *j*th QTL. Other terms of this equation are defined similarly as in Equation 10.

A SIMULATION STUDY

Design of the simulation experiment: The proposed method was evaluated empirically by analyzing a simulated large complex pedigree. The pedigree consists of 2020 individuals covering three discrete generations. The pedigree is depicted in Figure 1. Twenty founders were randomly sampled from an outbred base population; *i.e.*, founders were noninbred and genetically unrelated to each other. These founders are numbered from 1 to 20 and represented by open circles in the first row of Figure 1. With completely random mating among the founders, 20 full-sib families were formed, each represented by an open square in the second row of Figure 1. Because of the complete randomness, some founders (11, 13, and 18) were not represented in the next generation while others (*e.g.*, 1, 6, 9, etc.) were overrepresented. Each full-sib family contains 50 members so that a total of 1000 individuals are available in the F_1 generation (the second row of Figure 1). From the 1000 individuals, we randomly selected 20 as parents of the next generation. These 20 parents were randomly mated to form 20 full-sib families in the F_2 generation. Each full-sib family is represented by an open square in the third row of Figure 1. Again, each family consists of 50 members, leading to 1000 individuals in the F_2. The mating was completely arbitrary, including selfing, full-sib, and half-sib mating. Although we did not simulate parent-offspring mating (overlapping generation), nothing prevents us from doing that.

A quantitative trait was modeled as being controlled by three QTL residing on two chromosomes of length 100 and 70 cM, respectively. The $40 = 2 \times 20$ allelic effects and $820 = 40(40 + 1)/2$ dominance effects of each QTL in the founders were simulated from their corresponding normal distributions. The residual variance was set at $\sigma_e^2 = 1.0$. The fixed effect contains only the overall mean, which was set at $b = 0.0$. The true locations, and allelic and dominance variances of the three simulated QTL are given in Table 1. Marker data were generated for all individuals. Eleven and 8 codominant markers were respectively placed on the two chromosomes with a marker distance of 10 cM between two neighboring markers. Six equally frequent alleles were simulated at each marker locus. With this assignment of marker allele frequencies, many loci were partially informative and some were even uninformative at all. No phenotypic records were available for the 20 founders. The linkage phases of markers in the founders were reshuffled and eventually reconstructed via the MCMC process. Two sets of data were analyzed: data I include all 2020 individuals and data II include only founders and the F_1 generation, a total of 1020 individuals.

The initial value for the QTL number was set at two and the corresponding locations were at 50 cM of chromosome 1 and 40 cM of chromosome 2, respectively. The prior Poisson mean of the QTL number was $\mu = 2$ and the maximum number of QTL was $l_{\max} = 6$. The starting values were 0.05 for all QTL allelic and dominance variances and 0.0 and 2.0 for the overall mean and the residual variance, respectively. The initial marker linkage phases were assigned randomly for each founder. The initial QTL allelic inheritance for each individual was determined by the initial QTL locations and the initial complete genotypes of the flanking markers.

A flat prior was assigned to the overall mean. The priors for all variance components were chosen to be uniform on (0.0, 2.0], the right endpoint being equal to the true phenotypic variance. The prior for the QTL locations was uniform over the whole genome. The tuning parameters of the proposal distributions were chosen to be 2.0 cM for QTL locations and 0.05 for all other parameters.

The proposed MCMC sampler was run for 5×10^5 cycles in each of the MCMC analyses. The first 400 samples (burn-in) were discarded. To reduce serial correlation in the samples, we saved only one in every 50

TABLE 1

The true locations and allelic and dominance variances of the three simulated QTL

Chromosome	Location (cM)	Allelic variance ($\sigma^2_{\mathrm{a}_j}$)	Dominance variance ($\sigma^2_{\mathrm{a}_j}$)	Heritability
1	25	0.15	0.10	0.2
1	75	0.10	0.00	0.1
2	25	0.15	0.10	0.2

The heritability is defined as proportion of the phenotypic variance explained by the locus of interest.

cycles of simulations so that the total number of samples kept in the analysis was 10^4.

Results: The estimated posterior distributions of the QTL number in the analyses of the two data sets are given in Table 2. In each of the data sets, it is immediately apparent that there are three QTL controlling the trait. The posterior expectations are essentially the same as the true number of QTL for both data sets. The posterior modes of QTL numbers are consistent with the true number of QTL as well. The posterior for data II is more widely spread than that for data I, indicating that QTL can be more accurately detected using the extended families.

QTL locations were estimated using the posterior QTL intensity function (Sillanpää and Arjas 1998, 1999). In practice, we divided each chromosome into many small intervals of equal length, say 1 cM, and then calculated the proportion of QTL in each interval from the MCMC samples. The posterior QTL intensities for data I and II are presented in Figures 2 and 3, respectively. The QTL intensity graphs are concentrated around the true locations of the simulated QTL. Three peaks of the graph for data I appear in [24, 25] and [75, 76] on chromosome 1 and in [24, 25] on chromosome 2. The corresponding peaks for data II are in [25, 26] and [75, 76] on chromosome 1 and in [25, 26] on chromosome 2. These results not only support quite strongly a model having three QTL but also indicate that the QTL locations are estimated accurately for both data sets. Finally, we noted that it took only a few thousand iterations for QTL locations to converge to their stationary states, regardless of which initial position was chosen, indicating that the algorithm for updating QTL locations is very efficient.

The overall mean and the residual variance were estimated from all MCMC samples. The posterior distributions of these two parameters for data I are depicted in Figure 4. The posterior mean and standard error are 0.2175 and 0.2790 for the overall mean, and 1.2037 and 0.0552 for the residual variance, respectively. It can be seen that the overall mean and the residual variance were slightly overestimated.

Following the idea of Sillanpää and Arjas (1999), two methods were used to assess the QTL effects (variances in our case). In the first method, we constructed the location-wise posterior densities for the variances. In the second method, we used only the posterior samples in which QTL locations fall into the regions with sufficiently high estimated QTL intensities to estimate the allelic and dominance variances. Let $f_{\mathrm{a}}(\Delta_k)$ and $f_{\mathrm{d}}(\Delta_k)$ be the cumulative distribution functions associated with the allelic and dominance variances of a putative QTL in small interval Δ_k. We used the means of samples in Δ_k to assess $f_{\mathrm{a}}(\Delta_k)$ and $f_{\mathrm{d}}(\Delta_k)$. Therefore, $f_{\mathrm{a}}(\Delta_k)$ and $f_{\mathrm{d}}(\Delta_k)$ can be expressed as

$$f_{\mathrm{a}}(\Delta_k) = \frac{\sum_{m=1}^{10^4}\sum_{q=1}^{l^{(m)}}\sigma^2_{\mathrm{a}_q}1(\lambda_q^{(m)}\in\Delta_k)}{\sum_{m=1}^{10^4}\sum_{q=1}^{l^{(m)}}1(\lambda_q^{(m)}\in\Delta_k)}$$

and

$$f_{\mathrm{d}}(\Delta_k) = \frac{\sum_{m=1}^{10^4}\sum_{q=1}^{l^{(m)}}\sigma^2_{\mathrm{d}_q}1(\lambda_q^{(m)}\in\Delta_k)}{\sum_{m=1}^{10^4}\sum_{q=1}^{l^{(m)}}1(\lambda_q^{(m)}\in\Delta_k)},$$

respectively, where $l^{(m)}$ is the number of QTL in the mth posterior sample, $\lambda_q^{(m)}$ is Note that $f_{\mathrm{a}}(\Delta_k)$ and $f_{\mathrm{d}}(\Delta_k)$ are meaningful only when a sufficient number of samples are contained in Δ_k. The plots of $f_{\mathrm{a}}(\Delta_k)$ and $f_{\mathrm{d}}(\Delta_k)$ for data I are presented in Figure 2. The chromosome regions with sufficiently high posterior QTL intensity are given in Table 3. The posterior samples in which QTL locations fell into these regions were used to estimate the QTL variances. Figure 5 depicts the posterior distri-

TABLE 2

Estimate of the posterior distribution of the QTL number and its expectation

	Estimated distribution, for l =							Estimated expectation
	0	1	2	3	4	5	6	
Data I	0.0002	0.0001	0.0024	0.9459	0.0511	0.0003	0.0000	3.048
Data II	0.0000	0.0029	0.0230	0.7791	0.1921	0.0029	0.0000	3.1691

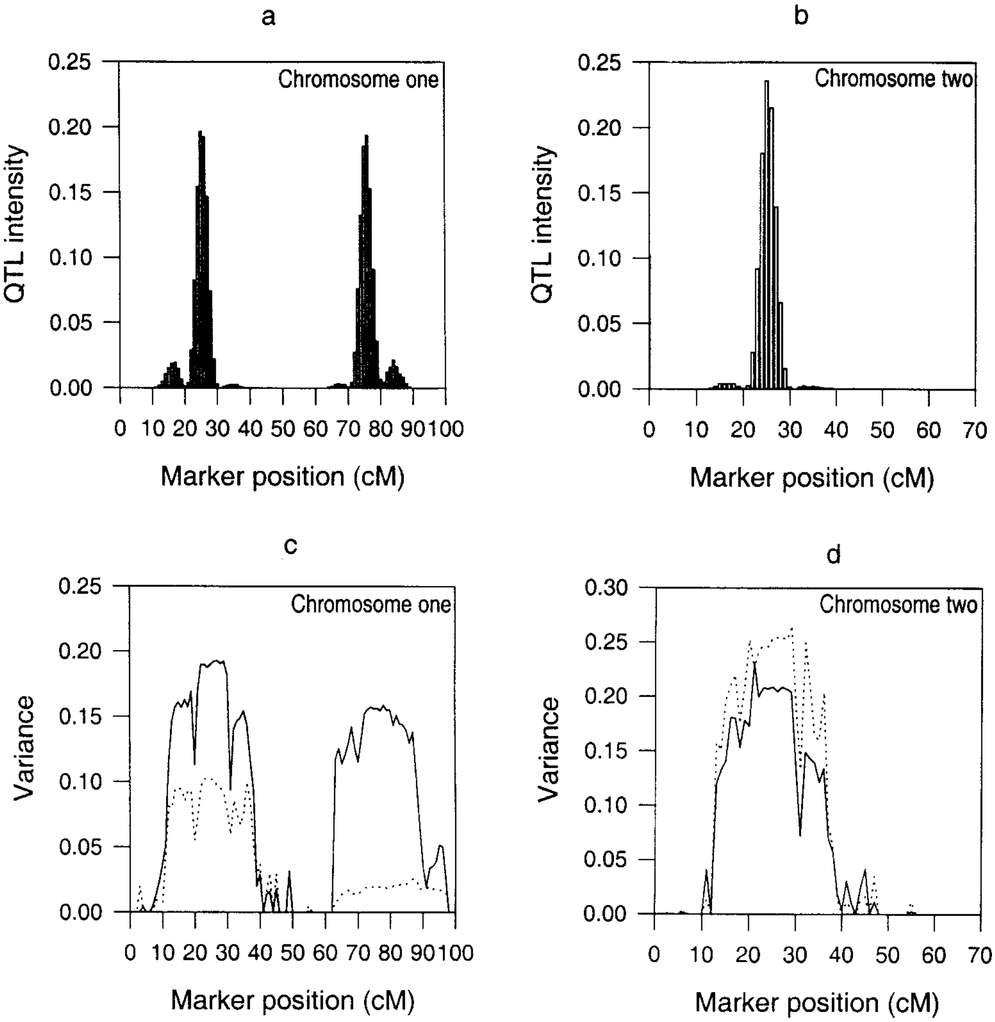

Figure 2.—Histograms of the posterior QTL intensity (a and b), and QTL allelic and dominance variance estimates (c and d) over two chromosomes with bin length of 1 cM for the analysis of data I. In c and d, the solid and dotted curves represent allelic and dominance variances, respectively.

butions for the QTL variances from the analysis of data I. We also calculated the means and the standard errors of the posterior samples for the QTL variances (see Table 3). In most situations, it appears that the estimates of QTL variances are close to the corresponding true values with small standard errors. The posterior means and the estimation errors of the QTL locations are also given in Table 3. It can be seen that the estimated QTL locations are very close to the corresponding true values.

DISCUSSION

There are many statistical methods and computer programs available for QTL mapping. Most of them are specialized in one or two particular types of designs, *e.g.*, BC, F_2, or multiple nuclear families. Here, we introduce a unified methodology of QTL mapping for arbitrarily complicated mating designs, ranging from a simple line cross to multiple independent sib-pairs. Although we developed the method based on a random-model approach, it works equally well for a fixed model. The difference between the random and the fixed models under the Bayesian framework is judged only by whether the distributions of allelic and dominance effects of QTL, *i.e.*, $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$, are treated as priors or not. If they are treated as prior distributions, the parameters involved in the prior distributions are assessed before the experiment, and there is no attempt to estimate them. The model is then called the fixed model. On the other hand, if $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$ are not the ultimate priors but the distribution of missing values $\mathbf{a}_j$ and $\mathbf{d}_j$, we are then interested in the parameters in $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$, *e.g.*, $\sigma^2_{a_j}$, which in turn need to be assigned a prior. We are essentially interested in making an inference for $\sigma^2_{a_j}$. In this case, the model is called a random model. For the fixed model, the update steps for QTL variances in the proposed algorithm are no longer required. Therefore, programming-wise, the difference between a random and a fixed model depends on the turning on/off of a single statement.

The ability to handle arbitrarily complicated pedigrees and the flexibility of switching between fixed and random models possessed by our Bayesian mapping arise from the use of an "allelic approach" as opposed to the traditional "genotypic approach." In this study, we dealt exclusively with the allelic (haplotype) values

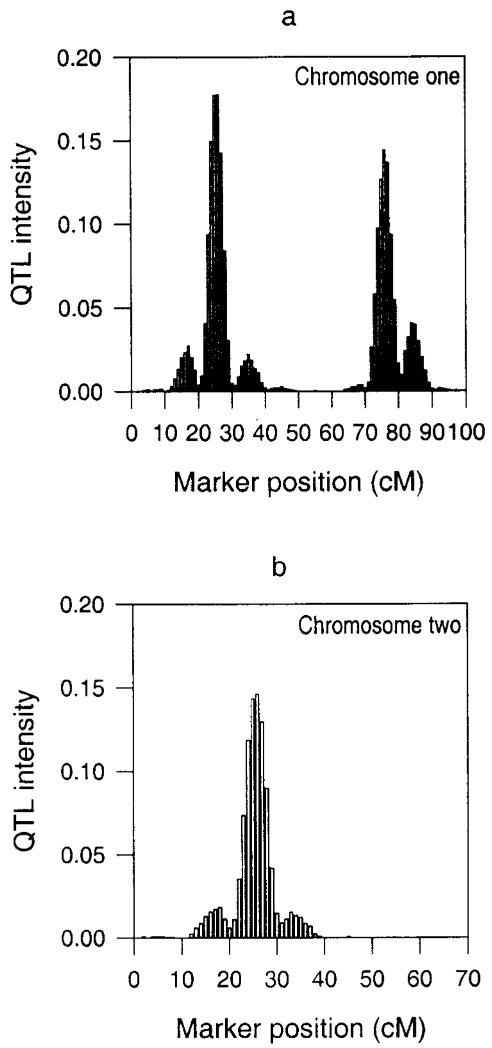

Figure 3.—Histograms of the posterior QTL intensity over chromosomes 1 (a) and 2 (b) with bin length of 1 cM for the analysis of data II.

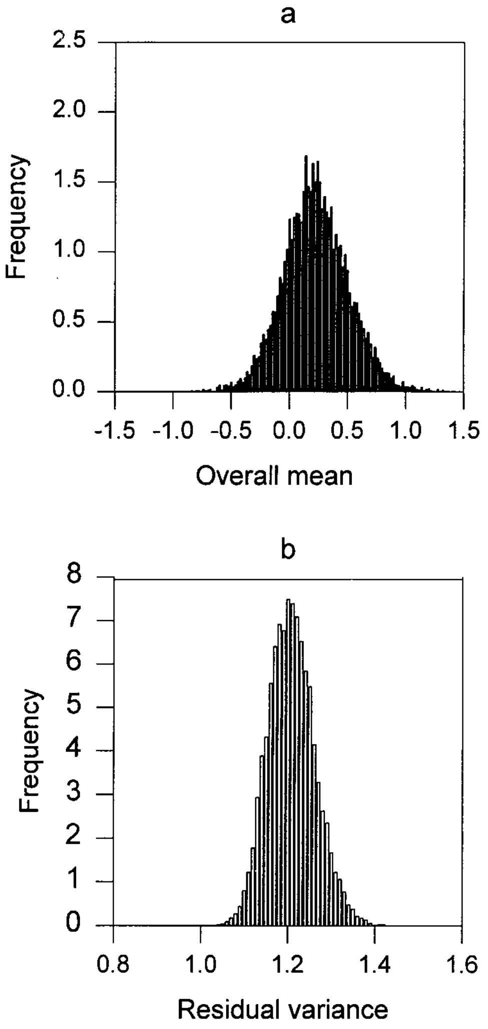

Figure 4.—Approximate posterior distributions of the overall mean (a) and residual variance (b) for data I.

rather than the genotypic values. We also sampled the "allelic inheritance" from parents to offspring rather than sampling the "genotypic transition." As a result, there is no need to consider the number of alleles and the total number of genotypes per locus in the mapping population. Instead, consideration is needed only when we assess the prior distribution of the founder alleles. This treatment has greatly simplified the algorithm and increased the robustness of the method.

In QTL mapping experiments of plants, founders are not usually a random sample from a reference population. They are often selected to be complementary for some traits of interest. As a consequence, the fixed-model approach can be used. Under a random model, however, the update steps for the additional parameters included in the priors $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$ depend on the form of their prior distributions. We have only explained the reversible jump MCMC algorithm under normal-effects QTL. Under the model of multiple QTL with discrete effects, we can modify the update steps (c) and (d) in the proposed algorithm. Assume that there are k alleles at a certain QTL with allelic effects $\{a_i\}_{i=1}^{k}$, dominance effects $\{d_{ij}\}_{i \geq j}$, and frequencies $\{p_i\}_{i=1}^{k}$ in the base population, where p_i, a_i, and d_{ij} are the frequency and effect of the ith allele and the dominance effect between alleles i and j, respectively. The priors for a_i and d_{ij} can be assigned as independent normal with known mean and variance. The prior for $\{p_i\}_{i=1}^{k}$ can take a symmetric Dirichlet. In this case, the parameters of interest are $\{a_i\}_{i=1}^{k}$, $\{d_{ij}\}_{i \geq j}$, and $\{p_i\}_{i=1}^{k}$. The frequencies $\{p_i\}_{i=1}^{k}$ can be updated using a Gibbs sampler because its full conditional distribution also remains Dirichlet (Gelman *et al.* 1995; Richardson and Green 1997). To update allelic effects $\{a_i\}_{i=1}^{k}$ and dominance effects $\{d_{ij}\}_{i \geq j}$, we first simulate a proposal for each parameter by using a random walk, then update the alleles of each founder by sampling from a multinomial distribution, and finally use the Metropolis-Hastings algorithm to update these parameters. Similar algorithms for the bial-

TABLE 3

Highest posterior QTL intensity interval, Bayesian estimates of QTL locations, and allelic and dominance variances

Chromosome	Interval (cM)	Sum of the QTL intensity	QTL location (cM)	Allelic variance $\sigma^2_{a_j}$	Dominance variance $\sigma^2_{a_j}$
1	20–∼30	0.9074	24.9302	0.1904	0.0995
			(1.6444)	(0.0738)	(0.0583)
	70–∼80	0.8717	75.5748	0.1547	0.0194
			(1.9312)	(0.0576)	(0.0137)
2	20–∼30	0.9765	24.8091	0.2067	0.2115
			(1.5393)	(0.0817)	(0.0870)

Standard errors of the estimates are given in parentheses.

lelic QTL model have been proposed in Uimari and Hoeschele (1997) and Heath (1997).

A problem may arise in real data analysis in which the number of alleles of a putative QTL in the base population is unknown, nor are the distributions of the effects of the QTL. In this situation, two strategies should be considered:

1. We can use the fixed-model approach to solve the random-model problem; that is, we first estimate the allelic and dominance effects of founders and then convert them into the QTL variances (Xu 1998). This method is expected to be efficient in the case where the number of founders is small.
2. In the case of many founders, one may use normal distributions to approximate the prior distributions $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$. In fact, drawing inferences about the multiallelic QTL variance via the normal distribution is a natural way to characterize genetic variation in the base population. In addition, normal distribution of the allelic effects is usually a very robust assumption. This has been verified in the context of ML mapping by Xu and Atchley (1995) who found that, for data simulated under a biallelic model, the analysis based on normal distribution provided very accurate estimates of QTL variances.

Further investigation is needed to investigate the robustness of normal distribution or other distributions in the framework of Bayesian analysis.

The proposed reversible jump MCMC algorithm performed well for the simulated data. The following points are noteworthy. First, the update step of QTL locations is different from existing algorithms in Bayesian mapping (*e.g.*, Satagopan *et al.* 1996; Heath 1997; Stephens and Fisch 1998; Sillanpää and Arjas 1998, 1999; Bink *et al.* 2000). We updated QTL location, allelic inheritance, and dominance design matrices simultaneously. The acceptance probability depended on the proposed location, the genotypes of flanking loci of the proposed location (markers or QTL), and the phenotypic value of the progeny as well as other parameter values in the model. The proposed method greatly improved the mixing of QTL locations in the simulation study. For the normal QTL effects model, we found that the method used in Stephens and Fisch (1998) and Sillanpää and Arjas (1998) resulted in QTL position stuck within the starting marker interval; *i.e.*, the chain was essentially reducible. Furthermore, our algorithm was much simpler than those of Heath (1997) and Bink *et al.* (2000). Second, the mixing of QTL number was sensitive to the way in which the proposals for the new QTL were generated when one QTL was added to the model and one QTL was removed from the model. The proposal distribution of allelic inheritance matrix $\mathbf{Z}^{\mathrm{p}}_{l+1}$ and $\mathbf{Z}^{\mathrm{m}}_{l+1}$ was crucial for the reversible jump step to perform well. It was found that QTL number mixed poorly when $p(\mathbf{Z}^{\mathrm{p}}_{l+1}, \mathbf{Z}^{\mathrm{m}}_{l+1}|\lambda_{l+1}, \mathbf{G}^{\mathrm{L}}_{l+1}, \mathbf{G}^{\mathrm{R}}_{l+1})$ was used to generate $\mathbf{Z}^{\mathrm{p}}_{l+1}$ and $\mathbf{Z}^{\mathrm{m}}_{l+1}$ in the normal-effects model approach, although such proposal distributions of the genotypes of new QTL worked well in line crosses and the biallelic-effects model (Uimari and Hoeschele 1997; Sillanpää and Arjas 1998, 1999). Third, we found very little influence of the starting values of unknowns on the mixing of the number of QTL. For example, starting with $l_0 = 6$, l quickly dropped to 3 after several hundred iterations and subsequently behaved the same as that started with $l_0 = 3$. We also found that the convergence speed of different parameters was quite different. As expected, the dominance variance converged most slowly due to too many dominance effects in the simulated data. Finally, the implement of the algorithm was computationally demanding due to the intricacy of our MCMC sampler. The analysis of data I with a chain of 5×10^5 cycles took ∼3 days on a SUN SPARC 5 workstation. The most time-consuming parts of our program were the updates of complete marker genotypes, QTL allelic inheritance, and dominance design matrices. The computational speed can be improved substantially with more efficient programming skills.

The assessment of the convergence and autocorrelation of the MCMC with the use of the reversible jump sampler remains a significant problem because the dimension keeps changing from one cycle to another. When the dimension changes, the identities of the QTL

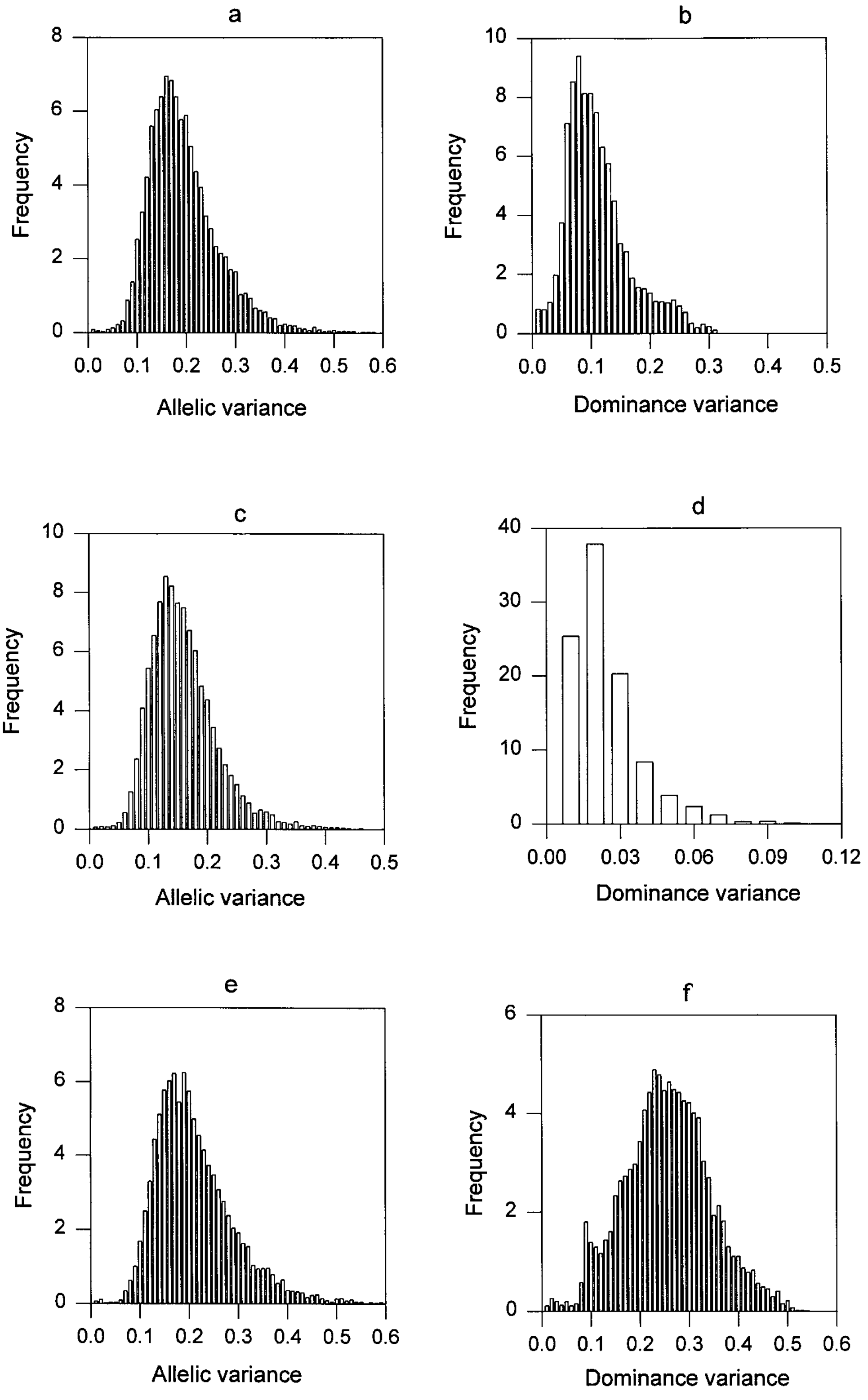

FIGURE 5.—Approximate posterior distributions of the QTL allelic and dominance variances for data I. (a and b) Allelic and dominance variances determined from interval 20–~30 cM of chromosome one, respectively; (c and d) allelic and dominance variances determined from interval 70–~80 cM of chromosome 1, respectively; (e and f) allelic and dominance variances determined from interval 20–~30 cM of chromosome 2, respectively.

also change. The parameters in one cycle of the iteration may be different from those in the next cycle of iteration. Therefore, the convergence criteria developed for the MCMC with fixed dimension are hardly applicable to the reversible jump MCMC (*e.g.*, BROOKS 1997). In our simulation study, therefore, we empirically determined the burn-in period, the length of the MCMC chain, and the interval length of subsampling to reduce the serial correlation. We used the plots of the changes in the number of QTL against the number of iterations to determine an approximate burn-in period (plots not shown). The length of subsampling intervals was chosen to eliminate obvious changing trends for all parameters. Compared with the reversible jump algorithm for line-cross designs (*e.g.*, Sillanpää and Arjas 1998), the acceptance rates for adding a new QTL to the model and deleting a QTL from the model in our simulation were relatively low. This result is expected because too many new values need to be generated for a birth step in the complicated mating design. The acceptance proportion for updating QTL locations was rather high (~75%).

We used the method of Bink and Van Arendonk (1999) to update marker genotypes in the simulation study. This method is established on a marker-by-marker and individual-by-individual basis. In general, this kind of single-site update does not always lead to an irreducible sampler because of the strong dependency of close relatives and strong dependency of adjacent loci. In our simulation study, we did not find insufficient mixing in marker genotype sampling, because the marker loci were not tightly linked. However, a block sampling, *e.g.*, the genotypes at a given locus being updated simultaneously for all individuals or sampling several loci jointly, is expected to be preferred over a single-site sampling in complex pedigrees and tightly lined loci. Such a sampling strategy will be incorporated into the proposed algorithm. The marker genotype sampler used in this study is suitable only in the case where there is no missing marker nonfinal offspring. When there are missing markers in nonfinal offspring, more sophisticated samplers, *e.g.*, the descent graph sampler (Sobel and Lange 1996), are required to update the marker complete genotypes. The descent graph sampler can be used to sample the gene flow patterns in arbitrarily complicated pedigrees. This powerful computational algorithm can be incorporated into our model.

Following the convention in human pedigree analysis, we have assumed that all founders are included in the model. In open-pollinated trees, however, seeds collected from one tree (mother) are usually pollinated from multiple unknown trees (fathers). Because the fathers (founders) are not identified, their contribution to the progeny is difficult to evaluate. Our model, in theory, can include these founders in the pedigree but treat their marker genotypes as missing. A method that excludes the missing founders while still analyzing the data properly is under development.

We thank Claus Vogl and Lori Weingartner for helpful comments on the manuscript. This research was supported by the National Institutes of Health grant GM55321 and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program 97-35205-5075 to S.X.

LITERATURE CITED

Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigree. Am. J. Hum. Genet. **62:** 1198–1211.

Amos, C. I., 1994 Robust variance-components approach for assessing genetic linkage in pedigrees. Am. J. Hum. Genet. **54:** 535–543.

Bink, M. C. A. M., and A. M. Van Arendonk, 1999 Detection of quantitative trait loci in outbred populations with incomplete marker data. Genetics **151:** 409–420.

Bink, M. C. A. M., L. L. G. Janss and R. L. Quaas, 2000 Markov chain Monte Carlo for mapping a quantitative trait locus in outbred populations. Genet. Res. **75:** 231–241.

Brooks, S. P., 1997 Discussion to Richardson and Green (1997). J. R. Stat. Soc. Ser. B **59:** 774–775.

Cantet, R. J. G., and C. Smith, 1991 Reduced animal model for marker assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **23:** 221–233.

Fernando, R. L., and M. Grossman, 1989 Marker-assisted selection using best linear unbiased prediction. Genet. Sel. Evol. **21:** 467–477.

Gelman, A. J. B., H. S. Carlin, H. S. Stern and D. B. Rubin, 1995 *Bayesian Data Analysis.* Chapman & Hall, London.

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.

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

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

Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. Am. J. Hum. Genet. **61:** 748–760.

Hoeschele, I., P. Uimari, F. E. Grignola, Q. Zhang and K. M. Gage, 1997 Advances in statistical methods to map quantitative trait loci in outbred populations. Genetics **147:** 1445–1457.

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

Kao, C. H., Z. B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

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

Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinauer Associates, Sunderland, MA.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, 1953 Equation of state calculations by fast computing machines. J. Chem. Phys. **21:** 1087–1091.

Muranty, H., 1996 Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity **76:** 156–165.

Richardson, S., and P. J. Green, 1997 On Bayesian analysis of mixtures with an unknown number of components. J. R. Stat. Soc. Ser. B **59:** 731–792.

Satagopan, R. J., and B. S. Yandell, 1996 Estimating the number of quantitative trait loci via Bayesian model determination. Special Contributed Paper Session on Genetic Analysis of Quantitative Traits and Complex Diseases. Biometric Section, Statistical Meeting, Chicago, IL.

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

Schork, N. J., 1993 Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am. J. Hum. Genet. **53:** 1306–1319.

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

Sillanpää, M. J., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics **151:** 1605–1619.

Slate, J., J. M. Pemberton and P. M. Visscher, 1999 Power to detect QTL in a free-living polygynous population. Heredity **83:** 327–336.

Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: applications to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58:** 1323–1337.

Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov Chain Monte Carlo. Biometrics **54:** 1334–1347.

Uimari, P., and I. Hoeschele, 1997 Mapping linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics **146:** 735–743.

Weller, J. I., Y. Kashi and M. Soller, 1990 Power of daughter and granddaughter designs for determining linkage between marker loci and quantitative trait loci in dairy cattle. J. Dairy Sci. **73:** 2525–2537.

Wijsman, E. M., and C. I. Amos, 1997 Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet. Epidemiol. **14:** 719–735.

Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148:** 517–524.

Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141:** 1189–1197.

Yi, N., and S. Xu, 2000 Bayesian mapping of quantitative trait loci for complex binary traits. Genetics **155:** 1391–1403.

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.

Communicating editor: Y.-X. Fu

APPENDIX

The prior distributions for allelic and dominance effects of the founders: Denote $\mathbf{a}_k^j$ as a 2 × 1 vector for the allelic effects of the kth founder at the jth QTL, and $\mathbf{d}_{kk'}^j$ as a vector of interaction effects (dominance effects) between the kth and the k'th founder alleles at the jth QTL. The dimension of $\mathbf{d}_{kk'}^j$ is 3 or 4, depending on whether k equals k'. Explicitly, $\mathbf{a}_k^j = (a_{k(1)}^j, a_{k(2)}^j)^{\mathrm{T}}$, $\mathbf{d}_{kk}^j = (d_{kk(1)}^j, d_{kk(2)}^j, d_{kk(3)}^j)^{\mathrm{T}}$, and $\mathbf{d}_{kk'}^j = (d_{kk'(1)}^j, d_{kk'(2)}^j, d_{kk'(3)}^j, d_{kk'(4)}^j)^{\mathrm{T}}$ $(k \neq k')$. The dimension of $\mathbf{d}_{kk}^j$ is 3 because each founder carries two alleles and potentially contributes 3 possible interactions in the descendants. The dimension of $\mathbf{d}_{kk'}^j$ is 4 because 2 founders carry a total of 4 alleles and potentially contribute 2 × 2 = 4 possible interactions.

The priors for $\mathbf{a}_k^j$ and $\mathbf{d}_{kk'}^j$ are assumed to be independent normals:

$$\mathbf{a}_k^j \sim N(\mathbf{0}, \mathrm{Var}(\mathbf{a}_k^j)), \quad \mathbf{d}_{kk'}^j \sim N(\mathbf{0}, \mathrm{Var}(\mathbf{d}_{kk'}^j)).$$

If the inbreeding coefficient, f_k, for the kth founder equals 1, the two elements of $\mathbf{a}_k^j$ are identical, and so are the three elements of $\mathbf{d}_{kk}^j$. Therefore, $\mathbf{a}_k^j$ and $\mathbf{d}_{kk}^j$ each turns into a scalar, $a_{k(1)}^j$ and $d_{kk(1)}^j$. Similarly, if both f_k and $f_{k'}$ are unity, $d_{kk'(1)}^j = d_{kk'(2)}^j = d_{kk'(3)}^j = d_{kk'(4)}^j$. Note that $\sigma_{\mathrm{a}_j}^2$ and $\sigma_{\mathrm{d}_j}^2$ are set at prefixed constants under a fixed model. Otherwise, they are treated as unknown parameters to be estimated. If $f_k < 1$, we have

$$\mathrm{Var}(\mathbf{a}_k^j) = \sigma_{\mathrm{a}_j}^2 \begin{pmatrix} 1 & f_k \\ f_k & 1 \end{pmatrix}, \quad (\mathrm{Var}(\mathbf{a}_k^j))^{-1} = \frac{1}{\sigma_{\mathrm{a}_j}^2 (1 - f_k^2)} \begin{pmatrix} 1 & -f_k \\ -f_k & 1 \end{pmatrix}$$

and

$$\mathrm{Var}(\mathbf{d}_{kk}^j) = \sigma_{\mathrm{d}_j}^2 \begin{pmatrix} 1 & f_k & f_k \\ f_k & 1 & f_k \\ f_k & f_k & 1 \end{pmatrix},$$

$$(\mathrm{Var}(\mathbf{d}_{kk}^j))^{-1} = \frac{1}{\sigma_{\mathrm{d}_j}^2 \, (1 + f_k)(1 + 2f_k)} \begin{pmatrix} 1 + f_k & -f_k & -f_k \\ -f_k & 1 + f_k & -f_k \\ -f_k & -f_k & 1 + f_k \end{pmatrix}.$$

If $f_k = 1$ and $f_{k'} \neq 1$, then $d_{kk'(1)}^j = d_{kk'(2)}^j$ and $d_{kk'(3)}^j = d_{kk'(4)}^j$. Therefore, $\mathbf{d}_{kk'}^j$ is equivalent to the 2 × 1 normal random vector $(d_{kk'(1)}^j, d_{kk'(3)}^j)^{\mathrm{T}}$, with covariance

$$\sigma_{\mathrm{d}_j}^2 \begin{pmatrix} 1 & f_{k'} \\ f_{k'} & 1 \end{pmatrix}.$$

Similarly, if $f_{k'} = 1$ and $f_k \neq 1$, $\mathbf{d}_{kk'}^j$ is equivalent to 2 × 1 normal random vector $(d_{kk'(1)}^j, d_{kk'(2)}^j)^{\mathrm{T}}$, with covariance

$$\sigma_{\mathrm{d}_j}^2 \begin{pmatrix} 1 & f_k \\ f_k & 1 \end{pmatrix}.$$

Finally, for the case where $f_k \neq 1$ and $f_{k'} \neq 1$, we can get

$$\mathrm{Var}(\mathbf{d}_{kk'}^j) = \sigma_{\mathrm{d}_j}^2 \begin{pmatrix} 1 & f_k & f_{k'} & f_k f_{k'} \\ f_k & 1 & f_k f_{k'} & f_{k'} \\ f_{k'} & f_k f_{k'} & 1 & f_k \\ f_k f_{k'} & f_{k'} & f_k & 1 \end{pmatrix},$$

and

$$(\mathrm{Var}(\mathbf{d}_{kk'}^j)) = \frac{1}{\sigma_{\mathrm{d}_j}^2 \, (1 - f_k^2)(1 - f_{k'}^2)} \begin{pmatrix} 1 & -f_k & -f_{k'} & f_k f_{k'} \\ -f_k & 1 & f_k f_{k'} & -f_{k'} \\ -f_{k'} & f_k f_{k'} & 1 & -f_k \\ f_k f_{k'} & -f_{k'} & -f_k & 1 \end{pmatrix}.$$

Under the assumption that the founders are independent from each other, the priors $p(\mathbf{a}_j)$ and $p(\mathbf{d}_j)$ can be expressed as the following forms, respectively:

$$p(\mathbf{a}_j) = \prod_{k=1}^{m} p(\mathbf{a}_k^j) \quad \text{and} \quad p(\mathbf{d}_j) = \prod_{k=1}^{m} \prod_{k'=k}^{m} p(\mathbf{d}_{kk'}^j).$$