Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Inbred Line Cross Data

Mikko J. Sillanpää and Elja Arjas

Rolf Nevanlinna Institute, University of Helsinki, FIN-00014 Finland

Manuscript received April 7, 1997
Accepted for publication November 17, 1997

ABSTRACT

A novel fine structure mapping method for quantitative traits is presented. It is based on Bayesian modeling and inference, treating the number of quantitative trait loci (QTLs) as an unobserved random variable and using ideas similar to composite interval mapping to account for the effects of QTLs in other chromosomes. The method is introduced for inbred lines and it can be applied also in situations involving frequent missing genotypes. We propose that two new probabilistic measures be used to summarize the results from the statistical analysis: (1) the (posterior) QTL-intensity, for estimating the number of QTLs in a chromosome and for localizing them into some particular chromosomal regions, and (2) the locationwise (posterior) distributions of the phenotypic effects of the QTLs. Both these measures will be viewed as functions of the putative QTL locus, over the marker range in the linkage group. The method is tested and compared with standard interval and composite interval mapping techniques by using simulated backcross progeny data. It is implemented as a software package. Its initial version is freely available for research purposes under the name Multimapper at URL http://www.rni.helsinki.fi/~mjs.

WHEN two purely homozygous, inbred, very genetically divergent lines are crossed, all offspring (F_1 generation) are genetically identical, being heterozygous at each locus. In the haplotypes of the F_1 individuals, the locus next to each quantitative trait locus (QTL) has the same allele as it had in the parental haplotype. This is because, in this ideal case, parents are homozygous at each locus and recombination events cannot change haplotypic arrangements. Therefore, linkage disequilibrium (nonrandom allelic association) in this group is maximal.

When an F_2 or backcross generation is produced, linkage disequilibrium is reduced slightly but still remains at a high level. The degree of reduction depends on the distance and on the recombination fraction between the considered QTL and the nearby marker locus. If mating is continued till F_3 and the succeeding generations, disequilibrium area surrounding a QTL is reduced further in each generation. This is why the backcross or F_2 intercross data from inbred lines is particularly suitable for QTL mapping.

The commonly used QTL mapping methods for plants and animals introduced recently use offspring data from divergent inbred lines in backcross or F_2 intercross design. The reason for using such a design is to maximize linkage disequilibrium and the amount of heterozygosity (information content) in meioses. Examples of such methods are interval mapping (Lander and Botstein 1989), least squares method (Haley and Knott 1992), and composite interval mapping (Jansen 1993; Jansen and Stam 1994; Zeng 1993, 1994). For some species, it is very impractical, time consuming and also expensive to produce inbred lines. In such cases, methods have been developed for considering crosses between outbred lines that are genetically divergent and show two very separate groups of phenotypic values, for example, due to different selection histories. One such procedure was presented in Haley *et al.* (1994), where the analysis was done in terms of line origins, with the assumption that crossed lines are "fixed" for different genes (or alleles) and would then show homozygosity in most of the QTL loci. This method (design) requires genotypic data from the parental and grandparental generations in addition to genotypic and phenotypic offspring data. Recently Jansen (1996) introduced a general method for line crosses by applying the EM-algorithm (Dempster *et al.* 1977), where the evaluation of the expectation step was conducted by a Markov chain Monte Carlo (MCMC) technique analogous to that of Guo and Thompson (1992). All these methods are based on the assumption that the distribution of the phenotypes is Gaussian. A robust method in this respect was developed recently by Kruglyak and Lander (1995). Modifications of (composite) interval mapping to binary traits were presented by Visscher *et al.* (1996a) and Xu and Atchley (1996), and a QTL mapping method for ordinal categorical traits by Hackett and Weller (1995).

In interval mapping (Lander and Botstein 1989),

Corresponding author: Mikko J. Sillanpää, Rolf Nevanlinna Institute, Research Institute of Mathematics, Statistics, and Computer Science, P.O. Box 4, FIN-00014 University of Helsinki, Finland.
E-mail: mjs@rolf.helsinki.fi

which is currently used routinely for QTL mapping (Mapmaker/QTL by Lincoln *et al.* 1992), it is possible to calculate likelihood scores for a putative QTL placed in any position between two adjacent flanking markers. By changing the flanking markers one at a time, it is possible to determine the likelihood curve over the whole genome. The procedure is based on regression of phenotypes on QTL genotypes, and because QTL genotypes are unknown, results are obtained by using an iterative EM algorithm in which convergence to a local maximum is guaranteed. For each fixed location separately, the algorithm searches the parameter vector value giving the highest likelihood score. These (profile likelihood) scores are then used to draw the LOD-score curve corresponding to different QTL positions.

The composite interval mapping procedure introduced by Zeng (1993, 1994), and the multiple-QTL mapping of Jansen (1993) and Jansen and Stam (1994), are in principle similar to interval mapping except that also some markers outside the tested interval (also in other chromosomes) are fitted to the model as covariates in order to reduce background noise caused by other QTLs and/or polygenic variation. The most significant markers for such reduction can be chosen in a preliminary analysis by using stepwise regression. These background control markers lie near the QTLs and they are used instead of the true QTLs (whose locations are unknown), yielding a better resolution than would be possible if those QTL effects were not considered at all. Theoretical support is provided by the fact that when a QTL has an effect on the trait one can see the same effect indirectly through the closest marker which is in linkage disequilibrium with the QTL (Tanksley 1993). The degree by which the corresponding phenotypic effect is reduced is determined by the recombination fraction between the marker and the QTL. However, there should be at least some effect, because linkage disequilibrium is almost at its maximum value in the backcross and F_2 intercross population. The improvement in resolution of the composite interval mapping over standard interval mapping is sometimes huge (*e.g.*, Kuittinen *et al.* 1997), thus yielding more putative QTL findings. However, the method does not yet seem to be widely used in practice, even though it is implemented in some software packages (*e.g.*, QTL Cartografer by Basten *et al.* 1996, and MapQTL by van Ooijen and Maliepaard 1996, and PLABQTL by Utz and Melchinger 1996).

The purpose of this paper is to present an approach for high resolution QTL mapping in inbred line crosses from incomplete data when the phenotypic distribution is assumed to be Normal. Our modeling approach is based on regression and the assumption that individual QTL effects are additive. The method belongs to the general framework of variable dimensional Bayesian models (*e.g.*, Green 1995), applying MCMC algorithms (see appendix a) to obtain the posterior distribution of the number of influential QTLs as well as estimating their locations in the (analyzed) chromosome and the corresponding phenotypic effects. Such a possibility was hinted at in a conference presentation by A. F. M. Smith (1996), and it has been explicitly implemented by Satagopan and Yandell (1996), Stephens and Fisch (1996), and Uimari and Hoeschele (1997).

MCMC methods (Metropolis *et al.* 1953; Hastings 1970) are not new techniques, but their widespread application in statistics did not start before the introduction of Gibbs sampling (Geman and Geman 1984). For a review of applications in gene mapping see Thompson (1994), Thomas and Gauderman (1995) and references therein. An excellent introduction to the Gibbs sampling and to the Metropolis-Hastings (M-H) algorithm can be found from Casella and George (1992), and Chib and Greenberg (1995). More advanced papers are Geyer (1992), and Besag *et al.* (1995). Variable dimensional parameterizations are considered in Arjas and Gasbarra (1994), and Green (1995).

Bayesian inference for gene mapping has been considered by Tai (1989), Thomas and Cortessis (1992), Smith and Roberts (1993), and Stephens and Smith (1993). A Bayesian QTL mapping method was proposed recently by Satagopan *et al.* (1996), where a prespecified number of QTLs was assumed to be present in the considered chromosome. The method did not take into account effects of QTLs in other chromosomes. Judgement concerning the actual numbers of QTLs (in the model) was proposed to be made by using Bayes factors from separate MCMC runs with different numbers of QTLs in each. Satagopan and Yandell (1996), and Stephens and Fisch (1996) considered all chromosomes simultaneously, treating the number of QTLs as a random variable. [The example in Satagopan and Yandell (1996) included only one chromosome, however.] Satagopan and Yandell (1996) used an M-H within Gibbs scheme in estimation, in contrast to an M-H scheme applied by Stephens and Fisch (1996), as well as here. Stephens and Fisch (1996) simulated 10 chromosomes in 6 different datasets, with a different heritability in each. They also tested several priors. However, they did not consider missing data.

Bayesian QTL mapping in outbred livestock population, using granddaughter design, has been studied by Uimari *et al.* (1996a,b) and Uimari and Hoeschele (1997). A Gibbs sampler was used in single and bi-QTL models which included both major gene and polygenic effects. Uimari and Hoeschele (1997) also tested the idea of a variable dimensional model, considering the cases of either zero, one or two QTLs and using the convention that the second QTL, if present, was always "to the right" of the first one. They concluded that their method was sensitive to the way in which the parameters of the second QTL were chosen. Effects of major QTLs in the other chromosomes were not taken into account in these studies, because composite interval mapping related techniques in multi-generational pedigrees are difficult to apply. Apart from Uimari *et al.* (1996a), techniques for handling missing genotypic data were not considered.

The contents of this paper are as follows. Next, we describe our statistical model. The results from simulation experiments are described thereafter. The final section contains a discussion of the method and of the experiences we had. In appendix a we outline the MCMC algorithm used in the estimation.

MODEL

QTL search is usually concentrated on a given chromosome (or chromosomal segment). Therefore everything in the following is with respect to such a chosen fixed linkage group (chromosome) unless the contrary is stated. Let $y = (y_1,y_2, ..., y_{N_{ind}})$ denote the vector of known phenotypes (missing phenotypes are not considered here), where N_{ind} is the number of individuals in the experiment and y_i is the phenotype of the ith individual. Suppose that the observations y_i are distributed according to a normal law resulting from some design in inbred linecross data. We shall view the unknown number of QTLs, denoted by N_{qtl}, as an unobserved random variable. Denote the QTL locations by $l = (l_1,l_2, ..., l_{N_{qtl}})$ and let χ be an $N_{qtl} \times N_{ind}$ matrix, where the qth column $x_q = (x_{q1},x_{q2}, ...,x_{qN_{ind}})'$ is the QTL genotype vector in location l_q, with element x_{qi} referring to the ith individual. Let G^* and G be the corresponding complete and incomplete (observed) marker information respectively; G^* and G are taken to be $N_{ind} \times N$ matrices, where N is the number of markers. Let I be the chromosomal interval with the first and the last markers of the chromosome as endpoints. A fixed marker map (*i.e.*, known recombination fractions between markers whose order is known), denoted by m, is assumed known before the analysis.

We denote complete and incomplete (observed) marker information in other chromosomes respectively by G_o^* and G_o. Let X_o^* be a subset of the complete marker information in other chromosomes, consisting of selected columns (marker genotype vectors) of G_o^*. Using an obvious set notation, $X_o^* \subset G_o^*$. Similarly, let X_o be a subset of incomplete marker information, $X_o \subset G_o$. We assume that X_o^* is chosen to correspond to known background control information (a selected set of markers that are hoped to be close to influential QTLs outside the interval I). Again, we arrange X_o^* into the form of an $N_{ind} \times N_{bc}$ matrix and denote its (i,k)th element by X_{ik}^*, corresponding to the genotype of the ith individual's kth background control. Here, N_{bc} is the number of background controls.

In the chosen design (*i.e.*, experimental cross), let N_{gen} be the number of possible genotypes and let $\alpha = (\alpha_1, ..., \alpha_{N_{gen}})$ be the vector including all possible genotypes at any (marker or QTL) locus, ordered so that all homozygote genotypes come before heterozygotes. In this setting, genotypes AB and BA are considered to be the same. The regression parameters are the following: a is the regression intercept (mean value), $b_q = (b_{q1}, ..., b_{qN_{gen}})$ is a vector of regression coefficients (classification variables) where b_{qj} is the regression coefficient for the qth QTL genotype α_j at location l_q, $\sigma^2 = Var(e_i)$ is the residual variance and $C = (c_{kj})$ is an $N_{bc} \times N_{gen}$ matrix of regression coefficients c_{kj} for background controls. We consider the following statistical model for y:

$$y_i = a + \sum_{q=1}^{N_{qtl}} \sum_{j=1}^{N_{gen}} b_{qj} 1_{\{x_{qi}=\alpha_j\}} + \sum_{k=1}^{N_{bc}} \sum_{j=1}^{N_{gen}} c_{kj} 1_{\{X_{ik}^*=\alpha_j\}} + e_i . \qquad (1)$$

Here $1_{\{x_{qi}=\alpha_j\}}$ is the indicator variable (dummy), which takes value one if the ith individual's qth QTL genotype x_{qi} at location l_q is α_j, and otherwise its value is zero. Similarly, $1_{\{X_{ik}^*=\alpha_j\}}$ is the indicator variable taking value one if the ith individual's marker genotype in the kth background control is α_j and otherwise its value is zero. Here we assume that the residuals e_i are independent and normally distributed according to $N(0, \sigma^2)$. In order to maintain the traditional way of considering gene effects one can make the following convention. In the case of three possible genotypes, say α = (AA,BB,AB), the constraint $b_{q2} = -b_{q1}$ will produce an additive effect for homozygotes and make b_{q3} correspond to a dominance effect of a heterozygote for each q. For the background control parameters, the constraint $c_{k2} = -c_{k1}$ for $k = 1, ..., N_{bc}$ will have a similar interpretation. In case of backcross, the corresponding constraints are $b_{q1} = 0$ and $c_{k1} = 0$, respectively.

We use the shorthand notation $\delta = (a,b_1, ..., b_{N_{qtl}}, \sigma^2,C)$ and $\theta = (\delta,\chi,l,G^*,X_o^*,N_{qtl})$. Notation $A^* \sim A$ means that A^* is consistent with A in cases where A is incomplete (observed) and A^* is complete information. From Bayes' theorem, we get $p(\theta|y,G,X_o,m) = 1/p(y,G,X_o|m)$ $p(y,\theta,G,X_o|m)$, where the joint density $p(y,\theta, G,X_o|m)$ can be factored into a likelihood and a (joint) prior density as $p(y,\theta,G,X_o|m) = p(y,G,X_o|\theta,m)p(\theta|m)$. Here the likelihood can be further written into the form of the product $p(y,G,X_o|\theta,m) = p(G|\theta,m)p(X_o|G,\theta,m)p(y|\theta,m,G,X_o) = 1_{\{G^*\sim G\}}1_{\{X_o^*\sim X_o\}}p(y|\theta,m)$. In this, $1_{\{G^*\sim G\}}$ and $1_{\{X_o^*\sim X_o\}}$ are indicators taking values one and zero depending on whether the complete genotypes G^* in the chromosome and in the background control sites X_o^* (in the other chromosomes) are consistent with their observed incomplete counterparts or not.

As for the joint prior distribution $p(\theta|m)$, given the marker map m, we make the following (conditional) independence assumptions:

(i) Given N_{qtl}, the vector δ consisting of the parameters of the linear model (1) is independent of the other coordinates of θ, *i.e.*, of (χ,l,G^*,X_o^*), as well as of m;

(ii) X_o^*, the vector consisting of the true background control genotypes in other chromosomes, is independent of (χ,l,G^*,N_{qtl}), the true marker and QTL-information in the considered chromosome as well as of the marker map m;

(iii) the true genotypes at marker locations, G^*, are conditionally independent of the number of QTLs (N_{qtl}) and their locations (l) given the marker map m.

Then the joint prior density function can be factored further and be presented in the product form

$$p(\theta|m) = p(G^*|m)p(N_{qtl}|m)p(l|m,N_{qtl}) \\ p(\chi|G^*,l,m,N_{qtl})p(\delta|N_{qtl})p(X_o^*). \qquad (2)$$

The density $p(X_o^*)$ is not conditioned on the fixed marker map, because the prior for genotypes can be thought not to be dependent on the marker order or distances. The posterior density of θ is then proportional to the joint density

$$p(\theta|y,G,X_o,m) \propto p(\theta|m)p(y,G,X_o|\theta,m) = \\ p(\theta|m)p(y|\theta,m)1_{\{G^*\sim G,X_o^*\sim X_o\}}\,, \qquad (3)$$

where

$$p(y|\theta,m) = \prod_{i=1}^{N_{ind}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\right. \\ \left.(y_i - (a + \sum_{q=1}^{N_{qtl}} \sum_{j=1}^{N_{gen}} b_{qj}1_{\{x_{qi}=\alpha_j\}} + \sum_{k=1}^{N_{bc}} \sum_{j=1}^{N_{gen}} c_{kj}1_{\{x_{ik}^*=\alpha_j\}}))^2\right] \qquad (4)$$

is the likelihood function (normal density) constructed from residuals e_i. Here complete background control genotypes X_{ik}^* are determined uniquely from X_o^*. The expression for the posterior distribution depends on the prior densities, where the parameter values are restricted to only that part of the parameter space which is consistent with what is already observed.

In specifying prior densities it is important to note that the number of possible alleles and genotypes (N_{gen}) depends on the (experimental) design in question. Therefore, also the prior densities $p(\chi|G^*,l,m,N_{qtl})$, $p(G^*|m)$ and $p(X_o^*)$ should reflect the design. Crosses between inbred lines have two alleles, forming two (three) different genotypes in backcross (F_2 intercross) designs.

Let M_s^* be the complete genotype vector in the sth marker position, *i.e.*, the sth column in G^*, and let $M_{s,i}^*$ be its ith component. In case there are some unobserved genotypic data, *i.e.*, $G^*\backslash G \neq 0$, we consider the following conditional independence structure for the prior $p(G^*|m)$: we assume that $p(M_s^*|G_{-s}^*,m) = p(M_s^*|M_{s-1}^*,M_{s+1}^*,m) = \Pi_{i=1}^{Nind}\, p(M_{s,i}^*|M_{s-1,i}^*, M_{s+1,i}^*,m)$, where G_{-s}^* includes all the other columns in G^* except the sth. In a backcross design, the prior for an individual i with complete genotype information G^* can be computed as the product

$$P(G_i^*|m) = p(M_{1,i}^*)\prod_{s=1}^{N-1} p(M_{s+1,i}^*|M_{s,i}^*,m), \qquad (5)$$

where

$$p(M_{s+1,i}^*|M_{s,i}^*,m) = r_{s,s+1}1_{\{M_{s,i}^*\neq M_{s+1,i}^*\}} + (1 - r_{s,s+1})1_{\{M_{s,i}^*=M_{s+1,i}^*\}} \qquad (6)$$

is the probability of having genotype $M_{s+1,i}^*$ at the marker position $s + 1$ in case there is genotype $M_{s,i}^*$ at position s, and $p(M_{1,i}^*)$ is the probability of genotype $M_{1,i}^*$ at a marker locus 1 in individual i. Here $r_{s,s+1}$ is the recombination fraction between the markers s and $s + 1$. (Note that $r_{s,s+1} = r_{s+1,s}$, and therefore also $p(M_{s+1,i}^* \mid M_{s,i}^*,m) = p(M_{s,i}^* \mid M_{s+1,i}^*,m)$.)

In the case of F_2 intercross, for each individual the transition probabilities $p(M_{s+1,i}^*|M_{s,i}^*,m)$ from position s to $s + 1$ can be arranged into the 3×3 matrix containing all possible transitions between states AA, BB, and AB

$$\begin{pmatrix} (1 - r_{s,s+1})^2 & r_{s,s+1}^2 & 2r_{s,s+1}(1 - r_{s,s+1}) \\ r_{s,s+1}^2 & (1 - r_{s,s+1})^2 & 2r_{s,s+1}(1 - r_{s,s+1}) \\ r_{s,s+1}(1 - r_{s,s+1}) & r_{s,s+1}(1 - r_{s,s+1}) & 1 - 2r_{s,s+1}(1 - r_{s,s+1}) \end{pmatrix}. \qquad (7)$$

The prior distribution of N_{qtl}, the number of QTLs, is here assumed to be truncated Poisson, where the Poisson mean λ and the maximum number N_{qtlmax} of QTLs are fixed control parameters such that $0 \leq \lambda \leq N_{qtlmax}$. The upper bound N_{qtlmax} is introduced for computational reasons.

In the following, we shall use the generic term "object" for any marker or QTL in the chromosome. The prior distribution of QTL genotypes at locations l is assumed to have the following product form:

$$p(\chi|G^*,l,m,N_{qtl}) = \prod_{q=1}^{N_{qtl}} p(x_q|G^*,x_1,.,x_{q-1},l,m) \\ = \prod_{q=1}^{N_{qtl}} \prod_{i=1}^{N_{ind}} p(x_{qi}|G_{iL}^{*q},G_{iR}^{*q},r_q)\,. \qquad (8)$$

Here, G_{iL}^{*q} and G_{iR}^{*q} are the genotypes of the left and the right flanking object (marker or QTL) for the qth QTL in individual i, chosen among the complete set of the markers in the chromosome and the QTLs at positions $l_1,.,l_{q-1}$. When the location of a QTL and the corresponding flanking object genotypes are known, the QTL genotype is independent of the genotypes of other objects (markers or QTLs) in this list. We denote by $r_q = (r_{q1}, r_{q2})$ the resulting recombination fractions between the QTL at l_q and the corresponding flanking objects. As is often done in QTL mapping applications, the same recombination rates for male and female meioses are assumed also here. The algorithms for constructing the probabilities of different QTL genotypes (last term in Equation 8) under backcross and F_2 designs can be found in appendices b1 and b2, respectively. In this construction, Haldane's map function is used for converting the distance between l_q and the left flanking object to a corresponding recombination fraction for each q. Haldane's formula $r_{q2} = (r_{fm}^q - r_{q1})/(1 - 2r_{q1})$ then gives the recombination fraction between l_q and the right flanking object, with r^{q}_{fm} being the recombination fraction between the two flanking objects of the qth QTL. Note that Equation 8 allows more than a single QTL within the same marker interval. Apart from being more general than models in which at most one QTL is allowed, this feature is thought to improve the mixing properties of the MCMC sampling algorithm.

An obvious choice for the priors of all the QTL locations is the uniform distribution on the considered chromosome, corresponding to the assumption that no prior knowledge concerning the QTL loci is available. However, if some 'non-data dependent' knowledge has been obtained, for example, using cytogenetical methods (*e.g.*, physical exclusion mapping), one can specify a prior which has most of its mass on some narrower chromosomal area. (Note that the uniform prior densities here do not cause any integrability problems, because all chromosomes are of finite length.) We shall assume equal prior probabilities for background control genotypes and use uniform prior distributions for all regression parameters. The natural range for the residual variance, for example, is between zero and the phenotypic variance.

The main features of the proposed model are summarized graphically in Figure 1. Our main interest is in the number of QTLs and in their positions in the considered chromosome. In order to arrive at a meaningful description of the results from the estimation we consider the QTLs as forming a nonhomogenous Poisson process over the chromosome. The results of the statistical analysis can then be expressed in an intuitive and coincise manner in terms of the corresponding estimated intensity. In practice, we divide the chromosome into intervals (bins) $\Delta_1, \Delta_2, ..., \Delta_{N_{bins}}$ of equal length (according to the Haldane distance). The interval length $\|\Delta_j\|$ chosen by the analyst reflects the resulting mapping resolution. Denote the number of MCMC cycles (sampling iterations) by N_{cycs}, and let

$$\hat{\lambda}_j = \left[\frac{1}{N_{cycs}} \sum_{k=1}^{N_{cycs}} \sum_{q=1}^{N_{qtl}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j\}}\right] / \|\Delta_j\| \tag{9}$$

be the approximate posterior QTL intensity on interval Δ_j obtained from the Monte Carlo simulation. Here $\Sigma_{q=1}^{N_{qtl}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j\}}$ is the number of QTLs in Δ_j in round k of the simulation. The product $\hat{\lambda}_j \|\Delta_j\|$ gives then an obvious approximation of the posterior expected number of QTLs in interval Δ_j. (Note that some bins might occasionally contain more than just one QTL during the same iteration cycle.) We combine the estimates $\hat{\lambda}_j$ into a single QTL-intensity function by writing $\hat{\lambda}(s) = \Sigma_j \hat{\lambda}_j 1_{\{s \in \Delta j\}}$, that is, $\hat{\lambda}(s) = \hat{\lambda}_j$ for $s \in \Delta_j$.

For assessing phenotypic effects in the backcross design, let $D_j(d)$ be the cumulative distribution function (c.d.f.) associated with the phenotypic effect of a putative QTL in bin Δ_j. There will then be one such c.d.f. for each bin. In the backcross design, let

$$\hat{D}_j(d) = \frac{\Sigma_{k=1}^{N_{cycs}} \Sigma_{q=1}^{N_{qtl}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j,\, b_{q2}^{(k)} - b_{q1}^{(k)} \leq d\}}}{\Sigma_{k=1}^{N_{cycs}} \Sigma_{q=1}^{N_{qtl}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j\}}} \tag{10}$$

be the empirical estimator of $D_j(d)$. To obtain corresponding formulas for the additive, $\hat{D}_j^a(d)$, and dominance, $\hat{D}_j^d(d)$, genetic effects (in unconstrained model) in an F_2 design, we must replace the indicator function in the numerator of Equation 10 by $1_{\{l_q^{(k)} \in \Delta j,\, b_{q1}^{(k)} - \mu_q^{(k)} \leq d\}}$, and $1_{\{l_q^{(k)} \in \Delta j,\, b_{q3}^{(k)} - \mu_q^{(k)} \leq d\}}$, respectively. Here $\mu_q^{(k)} = (b_{q1}^{(k)} + b_{q2}^{(k)})/2$. In the constrained model (where the constraint $b_{q2} = -b_{q1}$ is assumed for each q) the corresponding indicators are $1_{\{l_q^{(k)} \in \Delta j,\, b_{q1}^{(k)} \leq d\}}$ and $1_{\{l_q^{(k)} \in \Delta j,\, b_{q3}^{(k)} \leq d\}}$. For each empirical c.d.f. we determine its median and the 2.5- and 97.5-percent quantiles. The statistics are then drawn as functions of j, to get curves as shown in Figure 2. The estimates are stable when the denominator $N_{cycs}\hat{\lambda}_j\|\Delta_j\|$ in Equation 10 is large. In practice one should therefore concentrate on bins j for which $\hat{\lambda}_j$ is not too small. But these are precisely those bins which are most likely to contain a QTL anyway.

SIMULATION ANALYSIS

In order to test the performance of this method, a data set was simulated by the QTL Cartografer software (Basten *et al.* 1996). In particular, we wanted to compare its performance in the case of a simple backcross design to interval mapping (IM), and to composite interval mapping with five background controls (CIM/05). A backcross population ($N_{ind} = 250$) was generated of individuals having three 100-cM length chromosomes and in each 11 equally spaced markers 10 cM apart from each other. The trait was assumed to have heritability (the proportion of phenotypic variance explained by the simulated QTLs) 0.7 and phenotypic variance 1.0. The effects and the locations of the six simulated QTLs can be found in Table 1. A second data set was generated from this complete simulated set by randomly deleting 30 percent of the marker genotypes. The second set was used to test how well our method is capable of handling situations where a large proportion of genotypes are unknown.

First, the two data sets were analyzed by using the IM and CIM/05 methods. The background controls for the analyses were chosen by standard stepwise regression (QTL Cartografer software). The background control markers for the complete data were marker two in chromosome *1*, markers zero, one, and six in chromosome *2*, and marker two in chromosome *3*. In the analyses where 30 percent of genotypes were missing, the background controls were marker two and three in chromosome *1*, markers one and six in chromosome *2* and marker three in chromosome *3*. The same background controls were also used in the corresponding Bayesian analyses. In the CIM, "window width" was 10 cM (*i.e.*, the background controls less than 10 cM from the analyzed interval were not included in the model).

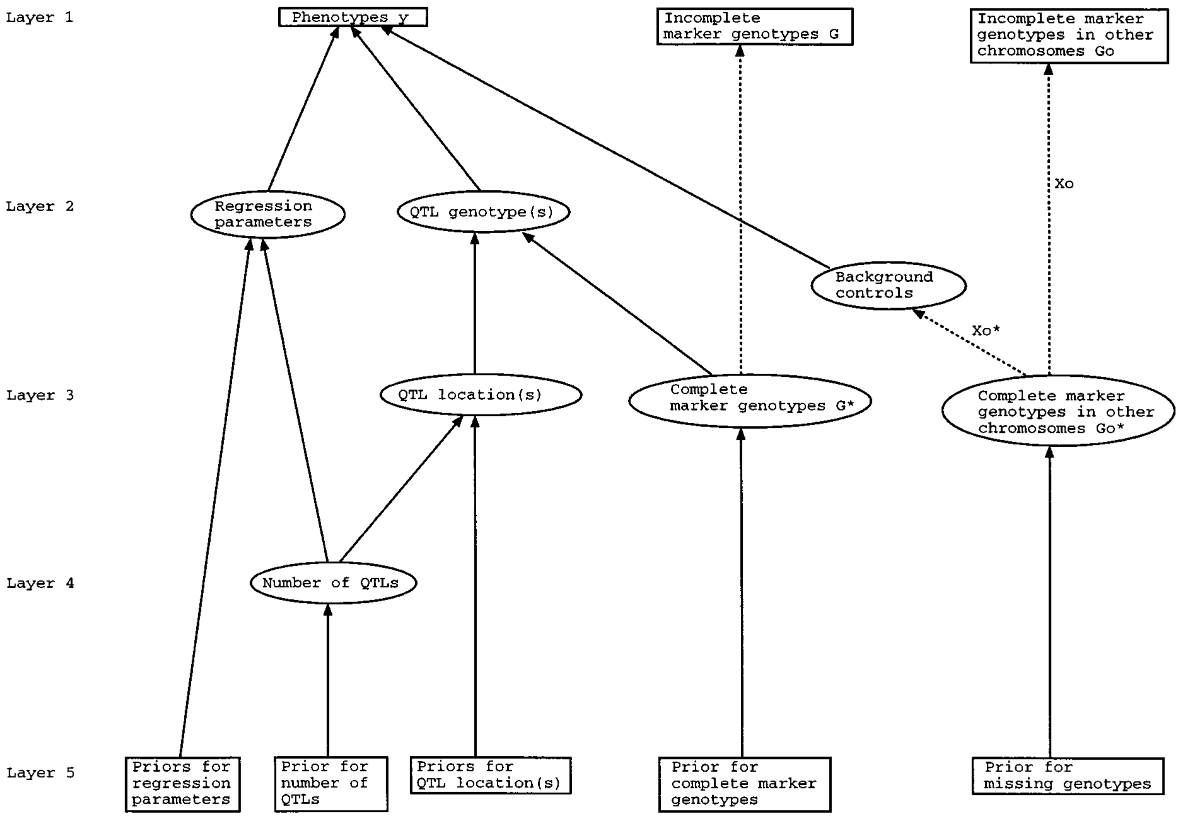
Layer 1
Layer 2
Layer 3
Layer 4
Layer 5
Phenotypes y
Incomplete marker genotypes G
Incomplete marker genotypes in other chromosomes Go
Regression parameters
QTL genotype(s)
Xo
Background controls
Xo*
QTL location(s)
Complete marker genotypes G*
Complete marker genotypes in other chromosomes Go*
Number of QTLs
Priors for regression parameters
Prior for number of QTLs
Priors for QTL location(s)
Prior for complete marker genotypes
Prior for missing genotypes

Figure 1.—Hierarchical structure of the model. Boxes refer to fixed values and ellipses to random variables. Layer one is observed, layer five given, and the others are unknown (sampled). Solid arrows indicate the direction of hierarchical dependency. Dotted arrows describe direct functional relationship.

Several test runs were made in order to carefully adjust the range of the proposal distributions (*i.e.*, parameters that control the maximal step size) of the Metropolis-Hastings algorithm. Such a range is specified for each of the following: (1) QTL locations, (2) regression mean, (3) residual variance, and regression coefficients of both (4) the QTLs and (5) background control genotypes. These control parameters influence directly the rejection rates; if they are chosen carelessly the chain may not convergence to the correct limiting distribution within a reasonable time. The values used in the final analyses are given in Table 2.

The (C-program implementing a Metropolis-Hastings-Green) chain was run 500,000 rounds for all analyses in chromosomes *1* and *2* and 1,000,000 and 1,500,000 rounds respectively for the complete and incomplete data in chromosome *3*. Computations were made on an UltraSparc Model 170 workstation, with running times varying between 1 hr and 30 min, and 6 hr and 20 min, depending on the chromosome and on other work load on the computer. The initial value for the number of QTLs was three, and the corresponding locations were 20.0 cM, 50.0 cM, and 80.0 cM in all MCMC runs. The Poisson mean (hyperparameter) was set to $\lambda = 2$ and the maximum number of QTLs (in the analyzed chromosome) to $N_{qtlmax} = 3$. The prior of the residual variance was chosen to be uniform over the range [0.0, 0.89], the right endpoint being equal to the phenotypic variance estimate from the data. The prior of the intercept was taken to be uniform over [−100, 100], and that for QTL and background control genotypic regression coefficients uniform over [−2, 2]. In all cases the chosen ranges were certain to cover all realistic parameter values. Finally, the prior for QTL locations was uniform over [0, 100].

Results: The likelihood ratio statistic (LRS) curves (in base 10 logarithmic scale) from IM and CIM/05 runs, and the Bayesian posterior QTL intensities in the considered three chromosomes, are shown on the left side of Figures 2 (complete data) and 3 (incomplete data). The phenotypic effect estimates given by the IM and CIM/05 methods, as well as the curves consisting of the pointwise (*i.e.*, in different locations of the putative QTLs) medians and the 2.5- and 97.5-percent quantiles of the posterior distributions of phenotypic effects are shown in the same figures on the right. For an obvious reason, the phenotypic effect estimates deserve serious consideration only in those chromosomal regions in which the statistical analysis suggests that there actually might be a QTL. Judging by the level of the QTL-intensity, we have shaded such areas in the figures.

Approximate posterior probability distributions of

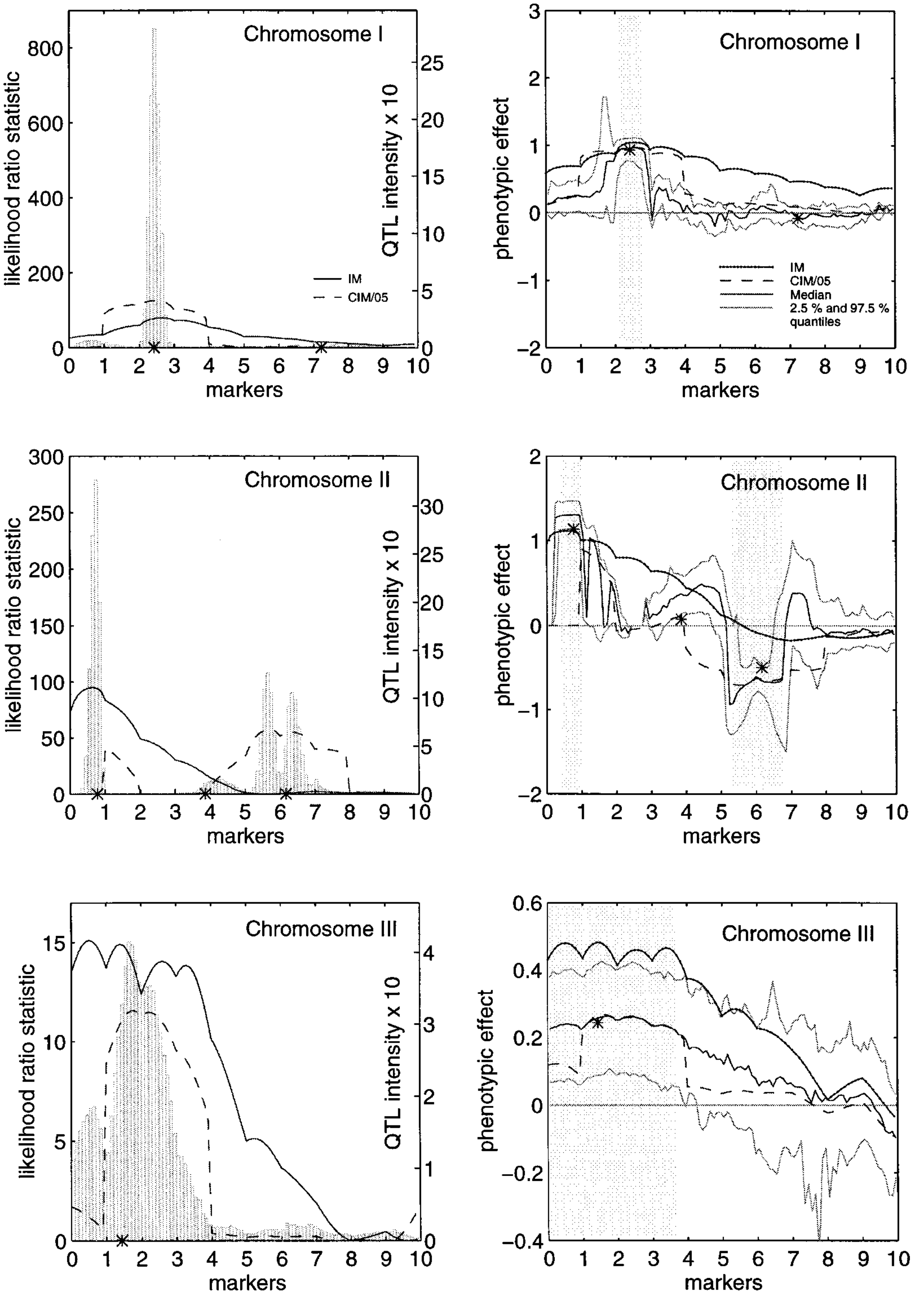

Figure 2.—The complete data analysis. On the left, the results from interval mapping (IM, solid line) and composite interval mapping with five background controls (CIM/05, broken line) are shown. Simulated true QTL locations are indicated with an asterix (*). The histogram corresponds to the (approximate) posterior QTL intensity over the chromosome, with binlength 1 cM. The left (right) y-axis corresponds to the likelihood ratio statistic (posterior QTL intensity). On the right, corresponding phenotypic effect estimates are shown. The solid line is the posterior median and the grey lines are the 2.5- and 97.5 percent quantiles of the posterior distribution of the phenotypic effect of a putative QTL. Shaded regions are suggested credible intervals for QTL localization (see Table 5). The phenotypic effect estimates (posterior median and quantiles) are reliable only in these regions.

the number of QTLs in different chromosomes, for both the complete and the incomplete data, are given in Tables 3 and 4. The posterior expectation of the number of QTLs in each chromosome (shown in Tables 3 and 4) coincides with the area under the corresponding QTL-intensity curve (in Figures 2 and 3).

Perhaps the most natural question to ask in this context is the following: "What is the (posterior) probability that some particular chromosomal area, say I, contains at least one QTL, given the evidence in the data?" Denoting the number of QTLs in I by N^I_{qtl}, one can calculate various MCMC approximations of that probability as shown in Equation 11 below, where $\lambda(s)$ is the QTL-intensity at point s, and $\hat{\lambda}(s)$ is its estimator:

$$\begin{aligned} P(N^I_{qtl} \geq 1 \mid \text{data}) &\approx \frac{1}{N_{cycs}} \sum_{k=1}^{N_{cycs}} 1_{\{l^{(k)}_q \in I \text{ for some } q\}} \\ &\approx 1 - exp\{-\int_I \hat{\lambda}(s)ds\} \\ &\approx \int_I \hat{\lambda}(s)ds. \qquad (11) \end{aligned}$$

TABLE 1

The locations and the phenotypic (additive genetic) effects of the simulated QTLs

Chromosome	Left marker	θ_L	$distance_L$	θ_R	Additive effect
1	2	0.0398	0.0415	0.0553	0.9337
1	7	0.0216	0.0221	0.0721	−0.0796
2	0	0.0720	0.0777	0.0218	1.1358
2	3	0.0791	0.0861	0.0137	0.0727
2	6	0.0164	0.0167	0.0767	−0.4984
3	1	0.0413	0.0431	0.0538	0.2444

The left column refers to the chromosome in which the considered QTL is. The next column refers to the nearest left flanking marker of the QTL in the chromosome. θ_L (θ_R) is the recombination fraction between the QTL position and the left (right) flanking marker. $distance_L$ gives the distance (in Morgans) between the left flanking marker and the QTL position, converted from θ_L by using Haldane's map function.

The last approximation, where the integral on the right is actually an expression for the (posterior) expected number of QTLs in I, is reasonable only if the integral is small. Table 5 gives a few such approximations for some chromosomal areas, based on either the complete or the incomplete data. These regions represent a moderate to high posterior QTL-intensity region surrounding the mode. Making the interval I longer will obviously always increase both the probability $P(N^I_{qtl} \geq 1 \mid \text{data})$ and the expectation $E(N^I_{qtl} \mid \text{data})$, but this will be at the cost of less accurate localization of the genes. In other words, given the evidence in the data, there is always a trade-off between accuracy and probability, just as in confidence intervals in classical statistical inference. In this sense, Table 5 represents only one possible summary of our findings. We have not made an attempt to establish a standard for forming such intervals or corresponding cut-off points here. Indeed, we think that the intensity curve in itself is the best summary of the information concerning the number of QTLs and their locations in the chromosome as obtained from the statistical analysis.

We do not display empirical threshold values (corresponding to permutation tests) for IM and CIM in Figures 2 and 3 because they were not the main issues here. However, the LOD score 3.0 would correspond to the likelihood ratio statistic $3.0 \times 2/log_{10}(e) \approx 13.82$, which can be used as a rule-of-thumb threshold when examining the figures (even though it was originally derived for monogenic traits in human). In practice the CIM method needs a somewhat higher threshold than IM (Zeng 1994). Note that the LRS curves and posterior QTL-intensity decrease sharply close to some marker positions. This is because genotype information concerning putative QTLs tends to be more accurate close to markers, and unless a QTL actually coincides with a marker locus, there will be strong evidence against placing a QTL in exactly, or very close to, these positions (see Kuittinen *et al.* 1997).

In Figure 2, the IM and CIM curves, apart from CIM in chromosome *3*, contain enough evidence (in the sense that the LRS is greater than the threshold value 13.82) of QTL activity. From Figure 2 it can be seen that our method performs well in all chromosomes, by

TABLE 2

Ranges of the proposal distributions *R*(.), proposal probabilities, numbers of iterations, and background control markers (BGCs) from other chromosomes, used in the simulation analyses

	$R(l_q)$	$R(a)$	$R(\sigma^2)$	$R(b_{qj})$	$R(c_{kj})$	$p_a = p_d$	Iterations	BGCs
Complete data								
Chromosome *1*	1.0 cM	0.01	0.01	0.1	0.1	1/3	500,000	2
Chromosome *2*	1.0 cM	0.01	0.01	0.1	0.1	1/3	500,000	0,1,6
Chromosome *3*	2.0 cM	0.01	0.01	0.1	0.1	1/3	1,000,000	2
Incomplete data								
Chromosome *1*	1.0 cM	0.01	0.01	0.1	0.1	1/3	500,000	2,3
Chromosome *2*	1.0 cM	0.01	0.01	0.1	0.1	1/3	500,000	1,6
Chromosome *3*	2.0 cM	0.01	0.01	0.1	0.1	1/3	1,500,000	3

Notation is as follows: $R(l_q)$ is the range of proposals for the QTL location parameters, $R(a)$ for the regression mean, $R(\sigma^2)$ for the residual variance, $R(b_{qj})$ for the regression coefficients of the QTL genotypes, and $R(c_{kj})$ for the regression coefficients of the background control genotypes.

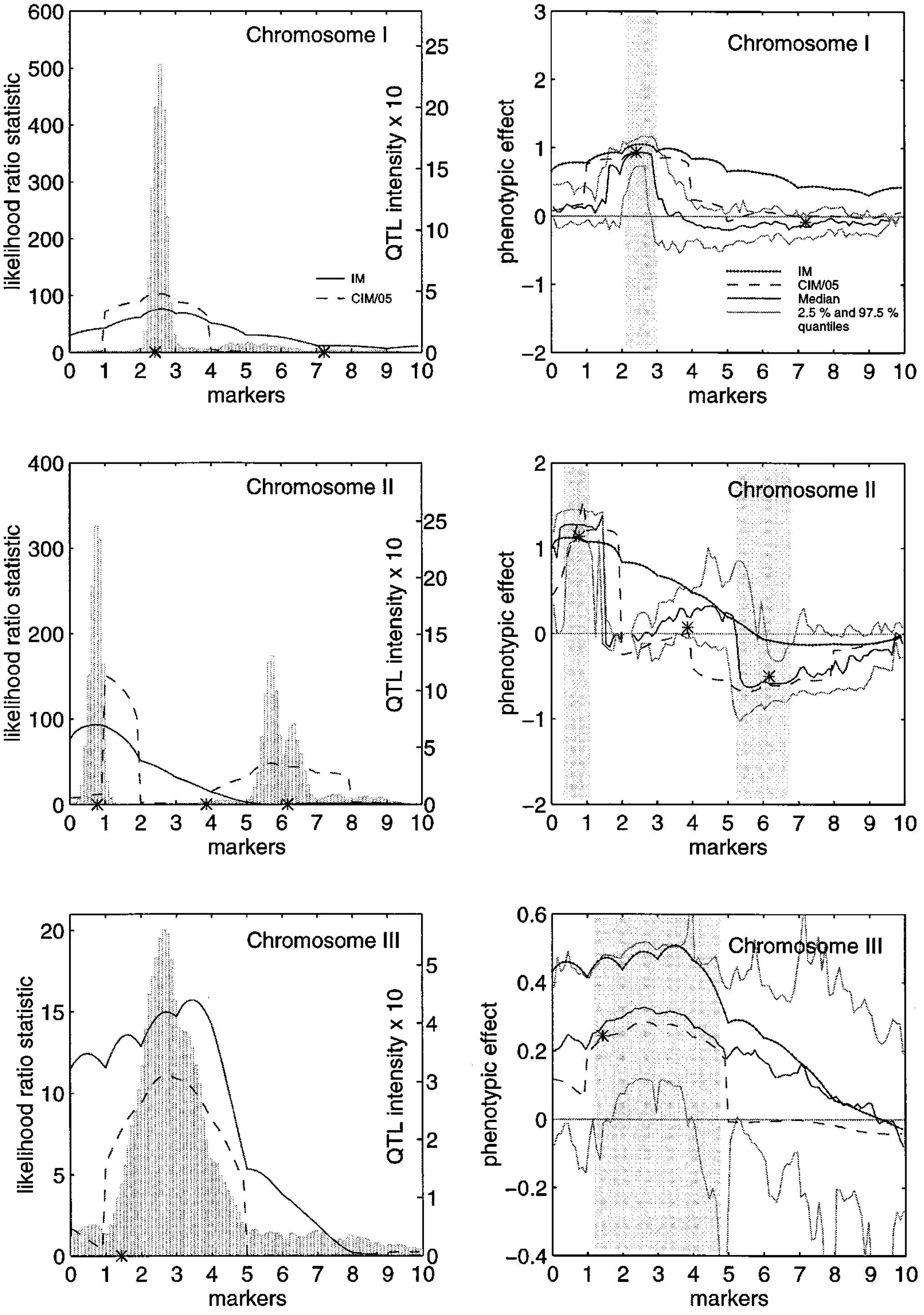

Figure 3.—The incomplete (30% of genotypes missing) data analysis. On the left, the results from interval mapping (IM, solid line) and composite interval mapping with five background controls (CIM/05, broken line) are shown. Simulated true QTL locations are indicated with an asterix (*). The histogram corresponds to the (approximate) posterior QTL intensity over the chromosome, with binlength 1 cM. The left (right) y-axis corresponds to the likelihood ratio statistic (posterior QTL intensity). On the right, corresponding phenotypic effect estimates are shown: the solid line is the posterior median and the grey lines are the 2.5- and 97.5-percent quantiles of the posterior distribution of the phenotypic effect of a putative QTL. Shaded regions are suggested credible intervals for QTL localization (see Table 5). The phenotypic effect estimates (posterior median and quantiles) are reliable only in these regions.

giving well localized high intensities for QTLs close to their true locations. Note that the highest QTL-intensities are often found near the modes of CIM or IM curves. The weakest QTLs in chromosomes 1 and 2 remained undetected by any of these methods, apparently because of their small phenotypic effects. All three methods estimated the phenotypic effect of the most influential QTLs very accurately, but IM and CIM/05 do not give confidence intervals for the estimates [unless they are estimated separately with a permutation test (Churchill and Doerge 1994) or by bootstrapping]. Our method gave posterior credible intervals for the phenotypic effects in all cases.

In the case of incomplete data, the mode of the intensity given by our method differs by 12 cM from the simulated true location in chromosome *3* (Figure 3). This is apparently a consequence of reduced genotype information in the incomplete data set, as well as of the

TABLE 3

The posterior distribution of the number of QTLs and its posterior expectation in the three chromosomes (complete genotype data)

	$P(N_{qtl} = x \mid y,G,X_o,m)$				
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$E(N_{qtl} \mid y,G,X_o,m)$
Chromosome *1*	0.0000	0.8916	0.1070	0.0014	1.1098
Chromosome *2*	0.0000	0.0000	0.7414	0.2586	2.2586
Chromosome *3*	0.1516	0.8018	0.0443	0.0023	0.8973

weak phenotypic effect attributable to this particular QTL. Note in this case the low maximal level of the posterior QTL-intensity, compared to the levels reached in the other two chromosomes.

For a more explicit comparison of the three methods, one-lod-support intervals (see Ott 1991, pp. 66–67) were determined around the modes of the IM and CIM/05 curves estimated from the complete data. The threshold values defining the one-lod-support intervals are the maximal value (the mode) minus 1.0 LOD (*i.e.*, $2/log_{10}(e)$ units in the LRS scale). An alternative would have been to estimate confidence intervals for the QTL locations, by applying a permutation test (see Churchill and Doerge 1994) or by bootstrapping, as in Visscher *et al.* 1996b. The results from the comparison are summarized in Table 6. (The numerical accuracy of the estimated support interval depends on the chosen step size which specifies how frequently the LRS is evaluated along the chromosome.)

Considering first the complete data set and the first chromosome with the LRS evaluated at every 0.1 cM, we obtained the one-lod-support interval [23.0 cM, 29.2 cM] using the IM, and [21.6 cM, 27.3 cM] using the CIM method. The latter almost coincides with the interval $\mathcal{I}$ = [21 cM, 28 cM] (see Tables 5 and 6), obtained from Figure 2 so that it contains practically all moderate to high posterior QTL-intensity values. All these intervals covered the true QTL at 24.15 cM. The posterior probability that the region $\mathcal{I}$ contains at least one QTL is 0.63. The true phenotypic effect of the second QTL (at 72.21 cM) in chromosome *1* was so small that this QTL was not detected by any of the three methods discussed here.

At the "left end" of chromosome *2*, we obtained the one-lod-support interval [2.9 cM, 9.0 cM] when using the IM, and [10.0 cM, 11.9 cM] when using the CIM method. The latter interval is actually so narrow that the true QTL at 7.77 cM falls just outside it. The posterior QTL-intensity in this region is concentrated almost completely on the interval $\mathcal{I}$ = [4 cM, 10 cM] which again contains the true QTL. At the "right end" of chromosome *2*, the LRS curve arising from CIM is bimodal, resulting in two overlapping one-lod-support intervals, [53.6 cM, 59.6 cM] and [52.9 cM, 67.3 cM]. The shorter CIM interval is so narrow that it does not contain the true QTL at 61.67 cM, but the wider one does. In this case, a natural choice for a Bayesian credible region would be the interval $\mathcal{I}$ = [53 cM, 68 cM], which is very close to that obtained by CIM and for which the posterior probability of there being at least one QTL in the region is 0.64. The IM method failed to reveal any statistically significant QTL activity in this part of chromosome *2*.

In chromosome *3*, we obtained the one-lod-support interval [0.0 cM, 39.8 cM] with the IM method. With the CIM method, we could not determine a corresponding support interval because the maximum LRS value was less than the critical value LOD 3.0. Because of the small phenotypic effect of the simulated QTL, the Bayesian method did not localize this QTL as well as in chromosomes *1* and *2*. Perhaps a natural interval to suggest, if at all, would be [0 cM, 37 cM], which would be nearly identical to the interval suggested by the IM-analysis and which would cover 90.2 percent of the total area under the QTL-intensity curve in that chromosome.

Whenever possible, all three methods estimated the

TABLE 4

The posterior distribution of the number of QTLs and its posterior expectation in the three chromosomes (incomplete genotype data)

	$P(N_{qtl} = x \mid y,G,X_o,m)$				
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$E(N_{qtl} \mid y,G,X_o,m)$
Chromosome *1*	0.0000	0.7803	0.2058	0.0139	1.2336
Chromosome *2*	0.0000	0.0000	0.7966	0.2034	2.2034
Chromosome *3*	0.0820	0.8292	0.0785	0.0103	1.0171

TABLE 5

Approximate (posterior) probability ($1 - exp\{-\int_I \hat{\lambda}(s)ds\}$) that a given chromosomal area I contains at least one QTL calculated for different areas I

	I	*Length* (I)	$P(N^I_{qtl} \geq 1 \mid data)$	$E(N^I_{qtl} \mid data)$
Complete data				
Chromosome *1*	[21 cM, 28 cM]	7 cM	0.63	0.9901
Chromosome *2*	[4 cM, 10 cM]	6 cM	0.63	0.9940
Chromosome *2*	[53 cM, 68 cM]	15 cM	0.64	1.0148
Chromosome *3*	[0 cM, 37 cM]	37 cM	0.55	0.8093
Incomplete data				
Chromosome *1*	[21 cM, 30 cM]	9 cM	0.64	1.0106
Chromosome *2*	[3 cM, 11 cM]	8 cM	0.63	0.9960
Chromosome *2*	[52 cM, 68 cM]	16 cM	0.63	1.0019
Chromosome *3*	[12 cM, 48 cM]	36 cM	0.57	0.8434

The (posterior) expected number of QTLs in I, calculated as the integral of the QTL intensity over I is also determined.

locations of the QTLs fairly well (see Table 6). In all these analyses the Bayesian estimates (posterior medians) of individual phenotypic effects were close to the true values, and they were practically the same as what was obtained when applying the CIM method. In this respect, the IM method turned out to be much inferior.

The results from analysing the incomplete data were largely similar, see Tables 5 and 6, and Figure 3 for details. The main difference was that the one-lod-support intervals become somewhat wider, as did the corresponding Bayesian credible regions. The differences were not very large, however.

DISCUSSION

We have presented here a novel method for high resolution mapping of multiple QTLs and for the estimation of their phenotypic effects, using the general framework of Bayesian variable dimensional models. For the estimation of the parameters (see appendix a) we use the Metropolis-Hastings algorithm with "reversible jumps" (Green 1995) between models with different numbers of QTLs. In the case of more than one QTL, this construction improved the mixing properties of the sampler when compared to a single-QTL model (results not shown). The effects on the QTLs in other chromosomes are taken into account indirectly, through nearby markers. When the genotype of a marker is missing, information from surrounding markers is utilized by applying the conditional probabilities as specified in Equation 5.

The main advantage in using Bayesian, instead of the more traditional frequentist inferential methods, is that they enable the analyst to quantify probabilistically the uncertainty involved in each claim made about QTLs, without needing to use problematic mental constructs such as "the relative frequency of incorrect decisions made in a long sequence of trials repeated under similar

TABLE 6

True locations (QTL), estimated locations and one-lod-support intervals for IM and CIM, and Bayesian point estimates (modes of the QTL intensity) together with the support intervals (I) from Table 5

	QTL	IM	CIM	I
Complete data				
Chromosome *1*	24.15	26.3 [23.0, 29.2]	24.4 [21.6, 27.3]	24.3 [21, 28]
Chromosome *2*	7.77	6.3 [2.9, 9.0]	10.0 [10.0, 11.9]	7.3 [4, 10]
Chromosome *2*	61.67	no peak	56.9 [53.6, 59.6], 63.3 [52.9, 67.3]	56.7 [53, 68]
Chromosome *3*	14.31	5.2 [0.0, 39.8]	LRS too low	16.5 [0, 37]
Incomplete data				
Chromosome *1*	24.15	26.1 [22.5, 29.2]	25.0 [21.8, 28.1]	25.6 [21, 30]
Chromosome *2*	7.77	7.3 [2.8, 11.3]	10.0 [10.0, 11.8]	7.9 [3, 11]
Chromosome *2*	61.67	no peak	56.9 [53.3, 59.8], 63.2 [52.0, 68.8]	57.1 [52, 68]
Chromosome *3*	14.31	34.4 [0.0, 43.7]	LRS too low	26.6 [12, 48]

LRS was evaluated at every 0.1 cM in IM and CIM estimation. Bayesian modes (intervals) were obtained with binlength 0.1 cM (1.0 cM).

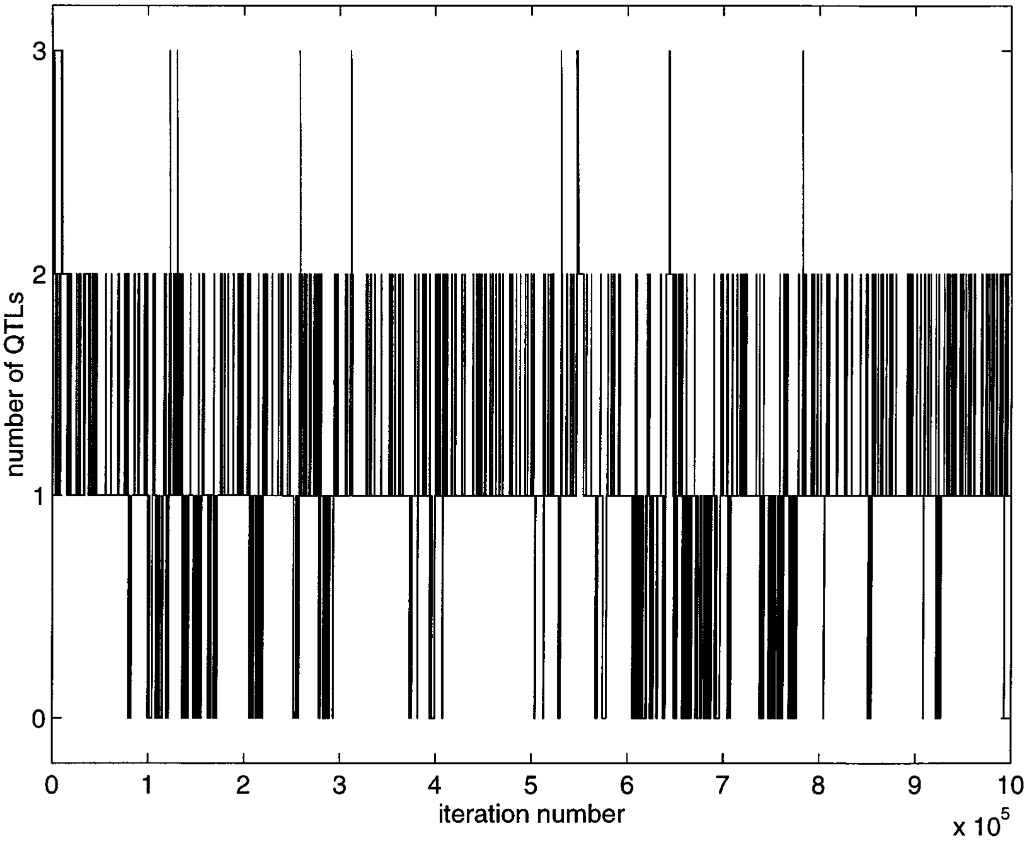

Figure 4.—The sample path $k \rightarrow N_{qtl}^{(k)}$ from a simulation trial of 1,000,000 iterations with complete data set (Chromosome *3*).

conditions.” The transformation of the prior distribution into the posterior through Bayes’ formula, corresponds directly to the natural intuition of “learning from the data.” Depending on the goals of the study, the specification of the prior can be “neutral” if the goal is to present a statistical summary of the information in the data, or, if available, it can also reflect an expert’s prior knowledge about unobserved quantities, in which case the posterior will be a synthesis of such expert knowledge and empirical evidence coming from the data. Another major advantage of the Bayesian approach is the relative ease by which missing data (such as missing genotypes) problems can be handled, together with the estimation of all other unobservables. Finally, the application of MCMC methods gives considerable freedom in building large hierarchical statistical models corresponding to the analyst’s perception of the underlying genetic structures and dependencies.

The results of the statistical analysis are summarized by two new measures: the posterior QTL-intensity, considered as a function of its location in the chromosome, and the posterior distribution of the phenotypic effect of the corresponding putative QTL. Such probabilistic summary measures seem to correspond directly to the immediate objectives of QTL mapping, that is, localizing the important QTLs in different chromosomes and estimating their effects on the phenotype(s). In particular, in situations where one can expect that there are several QTLs in the considered chromosomal area, the posterior QTL-intensity captures the essential information about their number and positions in an easily interpretable probabilistic form. In this way we can avoid completely the difficult inferential problems concerning the “correct” threshold values of a LOD score (or a LRS) which arise in testing multiple QTL hypotheses. Moreover, our method does not appear to produce false positives easily, as one can expect to get a low QTL-intensity in regions where there is no, or is only little, QTL activity.

The performance of our method was compared to interval mapping (IM) and composite interval mapping (CIM) by using a simulated backcross population of 250 offspring. A second data set was obtained by randomly deleting 30 percent of the marker genotypes in the complete set.

In the execution of the MCMC sampling we used an overparametrized regression model which has one extra coefficient for each QTL and for each background control locus. Therefore the model intercept and the genotypic coefficients are not identifiable as such, but their contrasts (phenotypic effects) are. We also tested the alternative updating scheme used by Satagopan and Yandell (1996) where, when the number of QTLs was proposed to be changed, its effect was first balanced against a corresponding change in the overall mean. Our overparametrized model seemed to have better mixing properties, however.

The MCMC methods require some amount of computational effort and will, in practice, need the capacity of a workstation. To ensure sufficient mixing, we performed some relatively long test runs before a final QTL analysis. Another possibility is to apply some diagnostic tools, such as CODA (Best *et al.* 1995). Because we ran long simulation trials, no sampled values were rejected because of burn-in. The mixing properties of the sampling algorithm do not seem to be very sensitive to the prespecified proposal probabilities. In the case of adding or deleting a QTLs, the only restriction appeared to be that the proposal probabilities of changing the dimension should not be too small. As an illustration of the degree of mixing which was typically encountered, we show (Figure 4) how N_{qtl} was varying in a simulation run of length 1,000,000. In the final runs, the observed rejection rates for both adding and deleting steps varied between 0.998 and 0.999. By comparison, the rejection rates for updating QTL locations were rather low in general because the proposals were usually limited to a fairly narrow interval around the current value. (Larger jumps were allowed when a QTL was either deleted or added.)

An important issue which has so far come up only implicitly in our analysis is that the number of QTLs and their phenotypic effects cannot be considered in isolation from each other. The idea that any gene with a non-zero effect on phenotypes is a QTL which in principle could be detected from data seems like an idealization far from reality, and as a consequence, “the correct number of QTLs” will exist as an objectively defined quantity only in simulated data. Here we have controlled this problem by bounding the number of QTLs in each chromosome by a given constant. An alternative would be to modify our method by paying attention only to influential QTLs, in the sense that their phenotypic effect exceeds some given threshold T. The only change this would make into our formulas is that, in Equations 9 and 10, we would have to add into indicators the restriction $|b_{q2}^{(k)} - b_{q1}^{(k)}| \geq T$.

Another question we have not discussed is how to choose the markers which are to be used as covariates in the analysis. Instead of using the results from a preliminary stepwise regression analysis as we have done here, one could “learn” from Bayesian QTL analyses of other chromosomes and choose certain markers from high posterior QTL-intensity areas as QTL representatives (background controls). Alternatively, one could choose covariates from amongst a set of candidate genes in case such genes are available. A final possibility would be to consider the entire genome in a single variable dimensional QTL analysis, always using the “current” QTLs as controls (Satagopan and Yandell 1996; Stephens and Fisch 1996). For this our method would need some adjustments, however.

An initial version of the program source code (written in C language) is freely available for research purposes from Rolf Nevanlinna Institute’s web page (http://www.rni.helsinki.fi/~mjs). A more user-friendly documented version is currently being developed. The present framework will be extended later to cover outbred linecross and (human) pedigree data. Epistatic effects (interactions between QTLs) and multiple trait analysis would also be worth considering in the future.

M.S. thanks Kari Auranen, Jukka Ranta and Dario Gasbarra for many useful discussions about Metropolis-Hastings algorithms, and Matti Taskinen for helpful hints in the programming and computer work. We are also grateful to Outi Savolainen, Leena Peltonen, Claus Vogl, Janne Pitkäniemi, Pekka Uimari and four anonymous referees for their constructive comments on an earlier version of the paper. This work was supported by a research grant from the Academy of Finland.

LITERATURE CITED

Arjas, E., and D. Gasbarra, 1994 Nonparametric Bayesian inference from right censored survival data using the Gibbs sampler. Statist. Sinica **4:** 505–524.

Basten, C. J., B. S. Weir and Z.-B. Zeng, 1996 QTL Cartografer, the reference manual and tutorial for QTL mapping. (available at http://statgen.ncsu.edu.)

Besag, J., P. Green, D. Higdon and K. Mengersen, 1995 Bayesian computation and stochastic systems. Statist. Sci. **10:** 3–66.

Best, N. G., M. K. Cowles and S. K. Vines, 1995 CODA: Convergence Diagnosis and Output Analysis software for Gibbs Sampler output: Version 0.3. Cambridge: Medical Research Council Biostatistic Unit.

Casella, G., and E. I. George, 1992 Explaining the Gibbs sampler. American Statistician **46:** 167–174.

Chib, S., and E. Greenberg, 1995 Understanding the Metropolis-Hastings algorithm. American Statistician **49:** 327–335.

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

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc., Ser. B. **39:** 1–38.

Geman, S., and D. Geman, 1984 Stochastic relaxation, gibbs distribution, and the Bayesian restoration of images. IEEE Trans. Pattn. Anal. Mach. Intell. **6:** 721–741.

Geyer, C. J., 1992 Practical Markov Chain Monte Carlo. Statist. Sci. **7:** 473–511.

Geyer, C. J., 1996 Likelihood inference for spatial point processes. (available at http://www.stats.bris.ac.uk/MCMC/)

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

Guo, S. W., and E. A. Thompson, 1992 Monte Carlo method for combined segregation and linkage analysis. Am. J. Hum. Genet. **51:** 1111–1126.

Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.

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

Haley, C. S., S. A. Knott, and J.-M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

Hastings, W. K., 1979 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

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

Jansen, R. C., 1996 A General Monte Carlo method for mapping multiple quantitative trait loci. Genetics **142:** 305–311.

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

Kruglyak, L., and E. S. Lander, 1995 A Nonparametric approach for mapping quantitative trait loci. Genetics **139:** 1421–1428.

Kuittinen, H., M. J. Sillanpää, and O. Savolainen, 1997 Genetic basis of adaptation: flowering time in Arabidopsis Thaliana. Theor. Appl. Genet. **95:** 573–583.

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

Lincoln, S., M. Daly, and E. S. Lander, 1992 Mapping genes controlling quantitative traits with MAPMAKER/QTL 1.1 Whitehead Institute Technical Report. 2nd edition.

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:** 1087–1092.

Ott, J., 1991 *Analysis of Human Genetic Linkage.* Revised edition. The John Hopkins University Press, Baltimore.

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

Satagopan, J. M., 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, Joint Statistical Meetings, Chicago, IL. (available at ftp://ftp.stat.wisc.edu/pub/yandell/revjump.html)

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.

Smith, A. F. M., 1996 Bayesian curves and CARTs. First European Conference on Highly Structured Stochastic Systems, Rebild, May 1996.

Smith, A. F. M., and G. O. Roberts, 1993 Bayesian computation via the Gibbs sampler and related Markov Chain Monte Carlo methods. J. Roy. Statist. Soc., Ser. B, **55:** 3–23.

Stephens, D. A., and R. D. Fisch, 1996 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Technical report. (available at http://www.ma.ic.ac.uk/statistics/techrep.html)

Stephens, D. A., and A. F. M. Smith, 1993 Bayesian inference in multipoint gene mapping. Ann. Human Genet. **57:** 65–82.

Tai, J. J., 1989 Application of Bayesian decision procedure to the inference of genetic linkage. J. Am. Statist. Assoc. **84:** 669–673.

Tanksley, S. D., 1993 Mapping polygenes. Annu. Rev. Genet. **27:** 205–233.

Thomas, D. C., and V. Cortessis, 1992 A Gibbs sampling approach in linkage analysis. Hum. Hered. **42:** 63–76.

Thomas, D. C., and W. J. Gauderman, 1995 Gibbs sampling methods in genetics, pp. 419–440 in *Markov Chain Monte Carlo in Practice*, edited by W. R. Gilks, S. Richardson and D. J. Spiegelhalter. Chapman & Hall, London.

Thompson, E. A., 1994 Monte Carlo likelihood in genetic mapping. Statist. Sci. **9:** 355–366.

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

Uimari, P., G. Thaller, and I. Hoeschele, 1996a The use of multiple markers in a Bayesian method for mapping quantitative trait loci. Genetics **143:** 1831–1842.

Uimari, P., Q. Zhang, F. Grignola, I. Hoeschele, and G. Thaller, 1996b Analysis of QTL workshop I Granddaughter design data using least-squares, residual maximum likelihood and Bayesian methods. J. QTL 1996: 2, art. 7.

Utz, H. F., and A. E. Melchinger 1996 PLABQTL: A program for composite interval mapping of QTL. J. Quant. Trait Loci **2:** 7.

van Ooijen J. W., and C. Maliepaard, 1996 Plant Genome IV. Abstract at: http://probe.nalusda.gov:8000/otherdocs/pg/pg4/abstracts/p316.html.

Visscher, P. M., C. S. Haley, and S. A. Knott, 1996a Mapping QTLs for binary traits in backcross and F2 populations. Genet. Res., Camb. **68:** 55–63.

Visscher, P. M., R. Thompson, and C. S. Haley, 1996b Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.

Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetic **143:** 1417–1424.

Zeng, Z.-B., 1993 Theoretical basis for separation of multiple linked gene effects in 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: Z-B. Zeng

APPENDIX A: ESTIMATION OF MODEL PARAMETERS VIA MARKOV CHAIN MONTE CARLO

Markov chain Monte Carlo (MCMC) methods enable one to calculate expectations with respect to the posterior in an approximate manner in situations where computation by traditional methods, because of the high dimension of the parameter space, would be complicated or impossible. In MCMC, each expectation is approximated by a sample average where the sample is drawn by simulation from an ergodic Markov chain constructed in such a way that its limiting distribution coincides with the posterior. The fact that the number of QTLs has not been fixed in advance, and is actually estimated in conjunction with the other model parameters, leads us to consider this problem within the general framework of variable dimensional parameter estimation.

We have preferred to use the Metropolis-Hastings algorithm (M-H) instead of its special case, the Gibbs sampler. The main motivation behind this choice has been that the M-H algorithm is so easy to implement, without needing to work with analytically involved full conditional distributions. An additional bonus of M-H is its greater flexibility and the relative ease by which it can be extended to more complex designs and pedigree structures.

For algorithmic reasons (mixing properties), we do not use any constraints for genotypic coefficients, even if this kind of overparametrization represents an uncommon modeling practice. Genotypic coefficients are here not identifiable as such, but their contrasts can be estimated.

Following Green's (1995) construction of a Markov chain with reversible jumps, we apply acceptance probabilities of the form min{1, (posterior ratio) × (proposal ratio) × (Jacobian)}. Throughout our consideration here, the Jacobian is one, being the determinant of an identity matrix. This is because the location(s) of existing QTL(s) do not determine the location of a proposed new QTL, nor does the deletion of one QTL influence the position(s) of the remaining QTL(s). In the following, we describe the steps of our Metropolis-Hastings-Green cycle. Initial values of the number of QTLs, of their locations, as well as the Poisson mean λ, the maximal number of QTLs allowed and the ranges of the uniform priors, are all given by the analyst (see simulation analysis). Initial values of the regression parameters are generated to be close to the center of their prior range. Initial values for the QTL genotypes and missing marker genotypes are generated from their priors. Figure 1, showing the hierarchical structure of the model, may help to understand the relationships between the parameters in the following updating scheme.

Step 1: We consider here three different move types: (1.1) modify the location(s) and configuration(s) of existing QTL(s), (1.2) add one QTL to the model, and (1.3) delete one QTL from the model. These move types have proposal probabilities p_m, p_a, and p_d, respectively, such that $p_a = 1_{\{Nqtl<N_{qtlmax}\}}c$, $p_d = 1_{\{Nqtl>0\}}c$, and $p_m = 1 - p_a - p_d$. Here c is a given positive constant in [0, 1/2]. In step 1.1 we do not fix the order of QTLs unlike in Richardson and Green (1997).

Step 1.1: $N_{qtl}^{(t)}$: $= N_{qtl}^{(t-1)}$. The following cycle is repeated for each QTL, $q = 1, \ldots, N_{qtl}^{(t)}$: A new proposal for a QTL location, l_q^{new}, is sampled from a symmetric uniform density around the previous value. The proposal is accepted with probability

$$\min\left\{1, \frac{p(x_q{=}x_q^{(t-1)}|G^{*(t-1)},x_1^{(t)},.,x_{q-1}^{(t)},l_1^{(t)},.,l_{q-1}^{(t)},l_q^{new},l_{q+1}^{(t-1)},.,l_{N_{qtl}}^{(t-1)},m)}{p(x_q{=}x_q^{(t-1)}|G^{*(t-1)},x_1^{(t)}.,.x_{q-1}^{(t)},l_1^{(t)},.,l_{q-1}^{(t)},l_q^{t-1)},.,l_{N_{qtl}}^{(t-1)},m)}\right\}.$$

When nothing else is changed, the likelihood remains unchanged even if the proposal is accepted and therefore these two likelihoods cancel from the acceptance probability expression. If a proposal is accepted then $l_q^{(t)} = l_q^{new}$, and otherwise $l_q^{(t)} = l_q^{(t-1)}$. New QTL genotypes are proposed separately for each individual so that x_{qi}^{new} is sampled from $p(x_{qi} \mid G_{qi,l}^{*(t-1)}, G_{qi,r}^{*(t-1)}, r_q)$ (constructed as show in appendix b). Individual proposals are accepted with probabilities $\{1, L_{i1}/L_{i2}\}$, where the likelihoods $L_{i1} = p(y_i \mid \delta^{(t-1)}, x_{qi}^{new}, \chi_{-qi}^{(t-1)}, l^{(t)}, X_{i,o}^{*(t-1)}, N_{qtl}^{(t)}, m)$ and $L_{i2} = p(y_i \mid \delta^{(t-1)}, x_{qi}^{(t-1)}, \chi_{-qi}^{(t-1)}, l^{(t)}, X_{i,o}^{*(t-1)}, N_{qtl}^{(t)}, m)$ for individual i are evaluated at the new and the old QTL genotypes, respectively. If the proposal for individual i is accepted then $x_{qi}^{(t)} = x_{qi}^{(new)}$, and otherwise $x_{qi}^{(t)} = x_{qi}^{(t-1)}$. Here $\chi_{-qi}^{(t-1)}$ includes all except the qth QTL genotypes in individual i in round t for QTLs having indices lower than q and in round $t - 1$ for higher indices.

Step 1.2: Add one new QTL. A proposal $(N_{qtl}^{(t)} = N_{qtl}^{(t-1)} + 1)$ for the number of QTLs is made. The new QTL location, $l_{(N_{qtl}^{(t-1)}+1)}$, is proposed from the uniform density on I. The QTL genotypes at location $l_{(N_{qtl}^{(t-1)}+1)}$ are proposed from $p(x_{(N_{qtl}^{(t-1)}+1)}|G^{*(t-1)},x_1^{(t-1)},.., x_{N_{qtl}}^{(t-1)}, l^{(t-1)}, l_{(N_{qtl}^{(t-1)}+1)},m)$. The regression coefficients of the new QTL genotypes are drawn from their priors. The proposal is accepted with probability

$$\min\left\{1, L_1/L_2 \times \frac{\lambda}{(N_{qtl}^{(t-1)} + 1)^2} \times \frac{p_d}{p_a}\right\},$$

where the likelihoods $L_1 = p(y \mid \theta^{(N_{qtl}^{(t-1)}+1)},m)$ and $L_2 = p(y \mid \theta^{(N_{qtl}^{(t-1)})}, m)$ are evaluated at the new and the old parameter value, respectively. If the proposal is accepted, then $N_{qtl}^{(t)} = N_{qtl}^{(t-1)} + 1$, and the new QTL location, the corresponding genotypes and the regression coefficients are all accepted simultaneously; otherwise $N_{qtl}^{(t)} = N_{qtl}^{(t-1)}$. The term $\lambda/(N_{qtl}^{(t-1)} + 1)^2$ in the Hastings ratio comes from the product

$$\begin{array}{lll} \dfrac{\lambda}{(N_{qtl}^{(t-1)} + 1)^2} = & \dfrac{\lambda}{(N_{qtl}^{(t-1)} + 1)} & (= \text{the Poisson prior ratio}) \\ & \times \dfrac{1}{(N_{qtl}^{(t-1)} + 1)} & \begin{array}{l}(= \text{the proposal probability} \\ \quad \text{ratio of selecting a} \\ \quad \text{particular QTL for a delete} \\ \quad \text{step and selecting a QTL} \\ \quad \text{for an add step)}\end{array} \\ & \times\ 1 & \begin{array}{l}(= \text{the ratio of two uniform} \\ \quad \text{proposal densities)}\end{array} \end{array}$$

The term in the denominator is squared only because we are not fixing the order of the QTLs (p. 23 in Geyer 1996, where the order of quantities is fixed).

Step 1.3: Delete one QTL. The proposal is made that the number of QTLs is decreased from $N_{qtl}^{(t-1)}$ by one, each of the deletions being equally likely. The deletion is accepted with probability

$$\min\left\{1, \frac{L_1}{L_2} \times \frac{(N_{qtl}^{(t-1)})^2}{\lambda} \times \frac{p_a}{p_d}\right\},$$

where the likelihoods $L_1 = p(y|\theta^{(N_{qtl}^{(t-1)}+1)},m)$ and $L_2 = p(y|\theta^{(N_{qtl}^{(t-1)})},m)$ are evaluated at the new and the old parameter value, respectively. If the proposal is accepted, then $N_{qtl}^{(t)} = N_{qtl}^{(t-1)} - 1$, and otherwise $N_{qtl}^{(t)} = N_{qtl}^{(t-1)}$.

Step 2: The marker genotype proposals G_{new}^*, denoted by $G_{i,new}^*$ for individual i, in the chromosome are sampled for all individuals from the distribution, where all the consistent genotypes are considered as equally likely. $p(G_{new}^* \mid m)$ is evaluated for all the individual according to Equation 5. The marker genotype proposals are accepted for each individual i separately with probability

$$\begin{aligned} \min\{1, \Pi_{q=1}^{N_{qtl}^{(t)}}[&p(x_{qi} = x_{qi}^{(t)} \mid G_{i,new}^*,x_{1i}^{(t)}, ., x_{(q-1)i}^{(t)}, \\ &l^{(t)},m)p(G_{i,new}^* \mid m)/[p(x_{qi} = x_{qi}^{(t)} \mid G_i^{*(t-1)}, x_{1i}^{(t)},.., \\ &x_{(q-1)i}^{(t)},l^{(t)},m)p(G_i^{*(t-1)} \mid m)]]\}. \end{aligned}$$

[Note that $p(M_{1,i}^*)$ for each individual i in Equation 5 cancels from the Hastings ratio]. If the proposals for individual i are accepted, then $G_i^{*(t)} = G_{i,new}^*$, and otherwise $G_i^{*(t)} = G_i^{*(t-1)}$.

Step 3: New proposals for the regression parameters are sampled from the symmetric uniform densities around their previous values (random walk). Denoting the likelihoods $L_1 = p(y|\delta^{new},\chi^{(t)},l^{(t)},X_o^{*(t-1)},m,N_{qtl}^{(t)})$ and $L_2 = p(y|\delta^{(t-1)},\chi^{(t)},l^{(t)}, X_o^{*(t-1)}, m,N_{qtl}^{(t)})$, the proposals (in the prior range) are accepted simultaneously with probability $\min\{1,L_1/L_2\}$. If new regression parameter values are accepted, then $\delta^{(t)} = \delta^{new}$, and otherwise $\delta^{(t)} = \delta^{(t-1)}$.

Step 4: The genotype proposals for the background control markers in other chromosomes are sampled directly from the uniform prior density $p(X_o^*)$ for all the individuals. All proposals which are consistent with the incomplete observations are considered as equally likely. The proposals are accepted, separately for each individual i, with probability $\min\{1,L_1/L_2\}$, where the corresponding likelihoods $L_1 = p(y_i \mid \delta^{(t)},\chi_i^{(t)},l^{(t)},X_{i,o}^{*new}$,

$m,N_{qtl}^{(t)})$ and $L_2 = p(y_i|\delta^{(t)},\chi_i^{(t)},l^{(t)},X_{i,o}^{*(t-1)}, m,N_{qtl}^{(t)})$ are evaluated at the new and the old parameter values, respectively. If the proposals for individual i are accepted, then $X_{i,o}^{*(t)} = X_{i,o}^{*new}$, and otherwise $X_{i,o}^{*(t)} = X_{i,o}^{(t-1)}$.

Step 5: Go back to the start, Step 1, until a prespecified number of cycles has been reached.

APPENDIX B

This appendix contains pseudo code algorithms for calculating the conditional probabilities of different QTL genotypes given the flanking objects (Equation 8) for each individual i in backcross (appendix b1) and F_2 intercross (appendix b2) designs. The main idea of the algorithms is described in detail in the backcross case.

B_1:

```
for $i = 1, . . ., N_{ind}$
  $total = 0$
  for $j = 1, . . ., N_{gen}$
    $alleleshare_L = \max[[(\alpha_{j1}\ equal\ G^*_{iL}(1)) + (\alpha_{j2}\ equal\ G^*_{iL}(2))],$
                    $[(\alpha_{j1}\ equal\ G^*_{iL}(2)) + (\alpha_{j2}\ equal\ G^*_{iL}(1))]]$
    $alleleshare_R = \max[[(\alpha_{j1}\ equal\ G^*_{iR}(1)) + (\alpha_{j2}\ equal\ G^*_{iR}(2))],$
                    $[(\alpha_{j1}\ equal\ G^*_{iR}(2)) + (\alpha_{j2}\ equal\ G^*_{iR}(1))]]$
    if $(alleleshare_L\ equal\ 2)\ p_1 = (1 - r_2)$
    else $p_1 = r_1$
    if $(alleleshare_R\ equal\ 2)\ p_2 = (1 - r_2)$
    else $p_2 = r_2$
    $p_{ij} = (p_1 \times p_2)$
    $total = total + p_{ij}$
  for $j = 1, . . ., N_{gen}$
    $p_{ij} = p_{ij} / total$
```

B_1 Inbred lines; backcross ($P \times F_1$): A pseudo code algorithm for constructing conditional QTL genotype ($N_{gen} = 2$) probabilities given the flanking objects for each individual i. Backcross offspring always get the same phase from the parental strain and the genotypic offspring ratio is 1 : 1 for genotypes AA and AB. The alleles in the jth genotype are α_{j1} and α_{j2}. The alleles in the left (right) flanking object (marker or QTL) genotype in individual i are denoted by $G^*_{iL}(1)$ and $G^*_{iL}(2)$ ($G^*_{iR}(1)$ and $G^*_{iR}(2)$). The algorithm first calculates the numerator $p(x_i = \alpha_j|G^*_{iL},r_q) \times p(G^*_{iR}|x_i = \alpha_j,r_q)$ for $j = 1, . . ., N_{gen}$. Inside the same loop, the algorithm accumulates the denominator according to the Chapman-Kolmogorov equation $p(G^*_{iR}|G^*_{iL},r_q,m) = p(G^*_{iR}|G^*_{iL},m) = \Sigma_{j=1}^{N_{gen}}\ [p(x_i = \alpha_j \mid G^*_{iL},r_q) \times p(G^*_{iR}|x_i = \alpha_j,r_q)]$. The conditional probabilities are then calculated from the formula $p(x_i = \alpha_j|G^*_{iL},G^*_{iR}, r_q) = p(x_i = \alpha_j|G^*_{iL},r_q) \times p(G^*_{iR}|x_i = \alpha_j,r_q) / p(G^*_{iR}|G^*_{iL},m)$ for $j = 1, . . ., N_{gen}$. $p(G^*_{iR}|G^*_{iL},m)$ is a normalizing constant (*total*) which makes probabilities add to one. [It is also equivalent to the probability of getting a certain flanking object haplotype from an F1 parent, which is $r_{fm} = (r_1 + r_2 - 2r_1r_2)$ or $(1 - r_{fm})$.]

B2:

```
for $i = 1, . . ., N_{ind}$
  $total = 0$
  for $j = 1, . . ., N_{gen}$
    $heterozygote_{QTL} = (\alpha_{j1}$ not equal $\alpha_{j2})$
    $heterozygotes_{(R)} = (G^*_{iR}(1)\ not\ equal\ G^*_{iR}(2))\ AND\ (G^*_{iR}(1)$
        $not\ equal\ G^*_{iR}(2))$
    if $(heterozygote_{QTL})\ N_{comb} = 2$
    else $N_{comb} = 1$
    if $(heterozygotes_{(R)})\ N_{comb} = N_{comb} \times 2$
    $p_{ij} = 0$
    $x_1 = \alpha_{j1}$
    $x_2 = \alpha_{j2}$
    $M_{r1} = G^*_{iR}(1)$
    $M_{r2} = G^*_{iR}(2)$
    for $k = 1, . . ., N_{comb}$
      if $((heterozygote_{QTL})\ AND\ ((k\ equal\ 2)\ OR\ (k\ equal\ 4)))$
        swap $(x_1, x_2)$
      if $([(not\ heterozygotes_{QTL})\ AND\ (k\ equal\ 2)]\ OR$
        $[(heterozygotes_{(R)})\ AND\ (k\ equal\ 3)])$ swap $(M_{r1}, M_{r2})$
      if $(x_1\ equal\ G^*_{iL}(1))\ p_1 = (1 - r_1)$
      else $p_1 = r_1$
      if $(x_1\ equal\ M_{r1})\ p_2 = (1 - r_2)$
      else $p_2 = r_2$
      if $(x_2\ equal\ G^*_{iL}(2))\ p_3 = (1 - r_1)$
      else $p_3 = r_1$
      if $(x_2\ equal\ M_{r2})\ p_4 = (1 - r_2)$
      else $p_4 = r_2$
      $p_{ij} = p_{ij} + ((p_1 \times p_2) \times (p_3 \times p_4))$
    $total = total + p_{ij}$
  for $j = 1, . . ., N_{gen}$
    $p_{ij} = p_{ij} / total$
```

B_2 Inbred lines; F_2 intercross ($F_1 \times F_1$): A pseudo code algorithm for constructing the conditional probabilities of QTL genotypes ($N_{gen} = 3$) given the flanking objects for each individual i. Different combinations for haplotypic assignments must be taken into account as well as the genotypic offspring ratio, which is 1 : 2 : 1 for the genotypes AA, AB, and BB, respectively. The alleles in the jth genotype are α_{j1} and α_{j2}. The alleles in the left (right) flanking object genotype in individual i are denoted by $G^*_{iL}(1)$ and $G^*_{iL}(2)$, [$G^*_{iR}(1)$ and G^*_{iR})].