Bayesian Mapping of Multiple Quantitative Trait Loci From Incomplete Outbred Offspring Data

Mikko J. Sillanpää* and Elja Arjas*,†

**Rolf Nevanlinna Institute, FIN-00014 University of Helsinki, Finland and †National Public Health Institute, FIN-00300 Helsinki, Finland*

Manuscript received July 6, 1998
Accepted for publication December 28, 1998

ABSTRACT

A general fine-scale Bayesian quantitative trait locus (QTL) mapping method for outcrossing species is presented. It is suitable for an analysis of complete and incomplete data from experimental designs of F_2 families or backcrosses. The amount of genotyping of parents and grandparents is optional, as well as the assumption that the QTL alleles in the crossed lines are fixed. Grandparental origin indicators are used, but without forgetting the original genotype or allelic origin information. The method treats the number of QTL in the analyzed chromosome as a random variable and allows some QTL effects from other chromosomes to be taken into account in a composite interval mapping manner. A block-update of ordered genotypes (haplotypes) of the whole family is sampled once in each marker locus during every round of the Markov Chain Monte Carlo algorithm used in the numerical estimation. As a byproduct, the method gives the posterior distributions for linkage phases in the family and therefore it can also be used as a haplotyping algorithm. The Bayesian method is tested and compared with two frequentist methods using simulated data sets, considering two different parental crosses and three different levels of available parental information. The method is implemented as a software package and is freely available under the name Multimapper/outbred at URL http://www.rni.helsinki.fi/~mjs/.

INBRED line cross designs are routinely used for quantitative trait locus (QTL) mapping in experimental organisms, because then full heterozygosity and perfect coupling between alleles in the QTL and in nearby marker loci are found in all F_1 individuals. Furthermore, the biallelic nature of the design suits well the tradition in genetics, where QTL are treated as biallelic and all different heterozygous QTL effects are considered jointly as a dominance effect. Depending on the organism, an attempt to produce inbred lines is not always practical or even possible (Haley *et al.* 1994); then methods developed for outbred designs are to be used.

Presently, there are QTL mapping methods suitable for the analysis of outbred populations (for a review see Hoeschele *et al.* 1997) as well as general pedigrees (*e.g.*, Heath 1997). However, an application of general pedigree analysis methods tends to be statistically inefficient and might actually not be possible for data arising from controlled outcrossing experiments. Therefore "design-specific" methods are needed. Their main advantages over general purpose pedigree methods are (1) incorporation of design-specific properties (such as a control of maximum number of possible QTL genotypes) into the analysis; and (2) background variation can be controlled by marker covariates, instead of using polygenic components or unlinked QTL.

When interval mapping, where a putative QTL is placed somewhere between markers, is applied to outbred offspring data, linkage phases (haplotypes) of parents must be considered. They are needed to determine whether paternally (or maternally) derived alleles at two neighboring loci are of the same grandparental origin or not. For a comparison, note that the grandparental line origin of alleles found in inbred line-cross offspring is automatically known in all marker positions.

Haley *et al.* (1994) presented a QTL mapping method for outbred line-cross data (F_2) concerning two divergent breeding populations, where a fixation of different influential QTL alleles in different grandparental lines was assumed. Their method requires genotyped grandparents to establish haplotypes for parents. They reduce the allelic space by using grandparental origin indicators instead of original marker alleles (see also Lander and Green 1987; Thompson 1994; Kruglyak *et al.* 1995; where a grandparental origin indicator is a binary digit in the inheritance vector). This work was extended to four segregating alleles by Knott *et al.* (1997). Somewhat earlier Maliepaard and Van Ooijen (1994) and Jansen (1996) presented more general algorithms for outcrossing experiments. Their methods assume neither a fixation of QTL in the crossed grandparental lines nor the availability of grandparental genotypes,

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

but instead require known haplotypes for parents. The method of Jansen (1996) was recently generalized by Jansen *et al.* (1998) to more complex populations, where parental haplotypes were not required to be known in advance. In this method, the full genotypic and allelic origin information is considered in all founders, but only segregation indicators (*i.e.*, grandparental origins) are used in nonfounders. No information is lost because the actual allelic forms for nonfounders can be traced from the pedigree by following each gene flow backward. Moreover, this treatment was shown to lead to more efficient mixing of the sampler than did methods in which the genotypes for nonfounders also are stored. The same idea was mentioned by Thompson (1994), and it was used by Sobel and Lange (1996) for descent graphs in pedigree analysis. Jansen *et al.* (1998) tested many different models.

Recently we presented a Bayesian QTL mapping method from incomplete inbred line-cross data (Sillanpää and Arjas 1998). This article also contains numerous references to other Bayesian works on QTL mapping. In this framework, the number of influential QTL in the analyzed chromosome is treated as an unobserved random variable, and then the algorithmic ideas of Green (1995) are applied to deal with the varying dimension of the parameter space. We used an idea similar to composite interval mapping (Jansen 1993; Zeng 1993, 1994; Jansen and Stam 1994; Kao and Zeng 1997) to account for the influence of some QTL in other chromosomes. We also advocated the use of the posterior QTL intensity as a new probabilistic summary measure for the inference. Now we generalize this approach to cover also backcross and F_2 (full-sib) offspring data, or multiple F_2 families from outcrossing experiments. In the method, the assumption concerning the fixation of QTL alleles in the crossed lines, as well as the degree in which the haplotypes or genotypes in parents or in grandparents are known, are optional. The assumption concerning fixation of QTL, together with the design (BC or F_2), determines the maximal number of QTL genotypes that can segregate in a family structure. We assume that the offspring are at least partly genotyped and that corresponding quantitative phenotypic measurements from the trait are available. If the parents and/or grandparents are not genotyped, we use information from progeny to impute consistent multiple random haplotypes for the parents, following a Markov Chain Monte Carlo (MCMC) scheme. We also use grandparental origin indicators as in Haley *et al.* (1994), but the coding is redone for each haplotype arrangement (imputation) in parents. As a byproduct, this approach produces the linkage-phase distributions for each offspring and their parents. Therefore it can also be used for haplotyping, in data with at least partially genotyped parents (see discussion).

If the F_2 family sizes in the studied plant or animal organism are relatively small, one has to combine information from several families. A complication arising from family pooling is that there will then typically be a large number of founders and therefore possible QTL alleles in the data. (Note also that the applicability of marker covariates needs to be considered.) To keep the maximum number of QTL genotypes low (≤ 4) in the combined data, one can assume one of the following alternatives: (1) Grandparents in each family have been drawn from the same two gene pools (lines), in which case they all represent two different QTL alleles in each trait locus. (Fixation of different QTL alleles in these two lines has been assumed.) (2) All families to be combined are related and share the same two grandparents, *i.e.*, all parents belong to the same F_1 generation. (Fixation of different QTL alleles in the two lines is again assumed.) (3) All families in the (combined) data are related and share the same four grandparents (numbered from 1 to 4) in such a way that one parent in each family is always progeny of grandparents 1 and 2 and the other parent is always progeny of grandparents 3 and 4; parents descending from grandparents 1 and 4, or 2 and 3, are excluded. Fixation of different QTL alleles in all four grandparental lines and that these lines show somewhat different phenotypic values has been assumed. If these assumptions are met, the resulting offspring population will have four different QTL alleles segregating in each trait locus.

In the following, we focus mainly on data from a one-family experiment. Our model is described next, followed by the results from simulation experiments and a discussion. In two appendixes, parameter estimation and summary measures for statistical inference are considered.

MODEL

We use the notation of Sillanpää and Arjas (1998) for the following entities: phenotype vector (y), the number of offspring individuals (N_{ind}), the number of QTL (N_{qtl}), QTL location vector (l), QTL genotype matrix (χ), the number of background controls (N_{bc}), incomplete and complete background control genotype information including parents (X_o and X_o^*), the number of QTL genotypes (N_{gen}), QTL genotypic effect (regression coefficient) vectors ($b_1, b_2, \ldots, b_{N_{\text{qtl}}}$), genotypic effects for background controls (C), residual variance (σ^2), fixed marker map m, and consistency between complete and incomplete information ($A^* \sim A$).

Let $I = (I_i)$ be the indicator vector, where element $I_i = 1_{\{y_i \text{ observed}\}}$ takes the value one or zero depending on whether y_i is observed or not. Let H^* and H be the corresponding complete and incomplete (observed) haplotype information (genotype + allelic origin information:paternal/maternal) in the marker positions. In each case, we indicate the split between maternally and paternally inherited haplotypes by writing $H^* = (H^{*\text{F}}, H^{*\text{M}})$ and $H = (H^{\text{F}}, H^{\text{M}})$. Here H^* and H are taken to

be $(N_{\text{ind}} + 2) \times N$ matrices, where N is the number of markers in the considered chromosome. Note that incomplete haplotype information often covers complete genotypic information but not the allelic origin.

In the chosen experimental design, let $\alpha = (\alpha_1, \ldots, \alpha_{N_{\text{gen}}})$ be the vector containing all possible QTL genotypes at any locus, so that their actual allelic forms are unknown. These QTL genotypes correspond to combinations of QTL alleles that were present in the crossed grandparents (founders) and that were transmitted to the F_1 parents. Let $\gamma_k = (\gamma_{k1}, \ldots, \gamma_{kN_{\text{gen}}^{\text{bc}(k)}})$ be the vector containing all possible background control genotypes and let $N_{\text{gen}}^{\text{bc}(k)}$ be their number (maximally four) at the kth background control. Let $B = (B_i)$, where B_i is a vector of covariates (*e.g.*, age, sex, or treatment) for offspring i. Let ρ be a vector of regression coefficients of these covariates (including also class means if some covariate is a classification variable). In case there is no individual control, we let all B_i reduce to $B_i = 1$ and ρ to a common regression intercept $\rho = a$. Here we consider only the case where no covariate values are missing.

We consider the following composite interval mapping (Kao and Zeng 1997) model for y:

$$y_i = \rho' B_i + \sum_{q=1}^{N_{\text{qtl}}} \sum_{j=1}^{N_{\text{gen}}} b_{qj} 1_{\{x_{qi}=\alpha_j\}} + \sum_{k=1}^{N_{\text{bc}}} \sum_{j=1}^{N_{\text{gen}}^{\text{bc}(k)}} c_{kj} 1_{\{x^*_{ik}=\gamma_{kj}\}} + e_i. \quad (1)$$

Here $1_{\{x_{qi}=\alpha_j\}}$ and $1_{\{X^*_{ik}=\gamma_{kj}\}}$ are indicator variables (*cf.* Sillanpää and Arjas 1998). For contrast parameterization, we can impose constraints $b_{q1} = 0$ and $c_{k1} = 0$ here for all values of q and k (see appendix b).

We use the shorthand notation $\delta = (b_1, \ldots, b_{N_{\text{qtl}}}, \sigma^2, \rho, C)$ and $\theta = (\delta, \chi, l, H^*, X_o^*, N_{\text{qtl}})$. Under natural conditional independence assumptions (*cf.* Sillanpää and Arjas 1998) the joint prior density function for θ can be presented in the product form

$$p(\theta|m) = p(H^*|m)p(N_{\text{qtl}}|m)p(l|m, N_{\text{qtl}}) \times p(\chi|H^*, l, m, N_{\text{qtl}})p(\delta|N_{\text{qtl}})p(X_o^*). \quad (2)$$

The posterior density of θ is then proportional to the right-hand side of

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

where $p(y \mid \theta, I, m)$ is the likelihood function (normal density) constructed from those independent residuals e_i in (1) in which the observation indicator $I_i = 1$. Here the complete background control genotypes are determined uniquely from X_o^*.

The ingredients of the prior density (2) are specified as follows. Denote complete haplotype information at the marker positions of the ith offspring by H_i^*, and similarly that of male and female parents by H_M^* and H_F^*. We consider the simple product form prior for the complete haplotypes in the family: $p(H^*|m) = p(H_F^*|m)\, p(H_M^*|m) \prod_{i=1}^{N_{\text{ind}}} p(H_i^*|H_F^*, H_M^*, m)$. Furthermore, for each offspring i we can further factorize the prior and compute it as the product

$$\begin{aligned} p(H_i^*|H_F^*, H_M^*, m) &= p(H_i^{*F}|H_F^*, m)p(H_i^{*M}|H_M^*, m)\ 1_{\{H_i^{*F}\sim H_F^*,\ H_i^{*M}\sim H_M^*\}} \\ &= p(\mathcal{G}_{1,i}^F(H^*)) \prod_{s=1}^{N-1} [p(\mathcal{G}_{s+1,i}^F(H^*)|\mathcal{G}_{s,i}^F(H^*))] \\ &\quad \times p(\mathcal{G}_{1,i}^M(H^*)) \prod_{s=1}^{N-1} [p(\mathcal{G}_{s+1,i}^M(H^*)|\mathcal{G}_{s,i}^M(H^*))] \\ &\quad \times 1_{\{H_i^{*F}\sim H_F^*,\ H_i^{*M}\sim H_M^*\}}. \end{aligned} \quad (4)$$

Here, $\mathcal{G}_{s,i}^F(x)$ ($\mathcal{G}_{s,i}^M(x)$) is a function of haplotype information x, and it determines the grandparental origin of the maternal (paternal) allele of individual i at marker locus s. The probabilities $p(\mathcal{G}_{1,i}^M(H^*) = \text{F}) = p(\mathcal{G}_{1,i}^M(H^*) = \text{M}) = p(\mathcal{G}_{1,i}^F(H^*) = F) = p(\mathcal{G}_{1,i}^F(H^*) = \text{M}) = \frac{1}{2}$ are the prior probabilities of different grandparental origins under Mendelian segregation for paternally and maternally inherited alleles at marker locus 1 in offspring i. When only maternally inherited alleles of offspring i are considered, then

$$p(\mathcal{G}_{s+1,i}^F(H^*)|\mathcal{G}_{s,i}^F(H^*)) = r_{s,s+1}1_{\{\mathcal{G}_{s+1,i}^F(H^*)\neq\mathcal{G}_{s,i}^F(H^*)\}} + (1 - r_{s,s+1})1_{\{\mathcal{G}_{s+1,i}^F(H^*)=\mathcal{G}_{s,i}^F(H^*)\}} \quad (5)$$

is the conditional probability that in individual i the marker at position $s + 1$ is of grandparental origin $\mathcal{G}_{s+1,i}^F(H^*)$ provided that the marker at position s has grandparental origin $\mathcal{G}_{s,i}^F(H^*)$. Here $r_{s,s+1}$ is the recombination fraction between the markers s and $s + 1$. The structure of $p(\mathcal{G}_{s+1,i}^M(H^*)|\mathcal{G}_{s,i}^M(H^*))$ derived for paternally inherited alleles is similar.

Let the complete background control marker information in parents F and M be $X_{o,F}^*$ and $X_{o,M}^*$, respectively. We assume the following prior form for background control genotypes in the other chromosomes: $p(X_o^*) = p(X_{o,F}^*)p(X_{o,M}^*) \prod_{i=1}^{N_{\text{ind}}} p(X_{o,i}^*|X_{o,F}^*, X_{o,M}^*)$, where $p(X_{o,i}^*|X_{o,F}^*, X_{o,M}^*) \propto p(X_{o,i}^*)1_{\{X_{o,i}^*\sim X_{o,F}^*,\ X_{o,i}^*\sim X_{o,M}^*\}}$. We also assume marker independence and that all (consistent) genotypes are *a priori* equally likely.

The prior distribution of the number of QTL is assumed to be truncated Poisson (see Sillanpää and Arjas 1998). For all QTL locations, we assume the uniform prior distribution on the considered chromosome. The prior for QTL genotype coefficients is assumed to be normal with zero mean and zero correlation, the variance being a hyperparameter specified by the analyst.

As in Sillanpää and Arjas (1998), we use the term *object* to represent any marker or QTL in the considered chromosome and the term *flanking object* (of the QTL q) to represent any combination of two entities [markers and/or QTL: $1, \ldots, (q - 1)$] having their loci closest to the QTL q. Now, denote by $H_{i,LR}^{*q}$ the complete (grand-

TABLE 1

The number of possible alleles and genotypes at a QTL in outbred linecross designs (backcross and F_2 intercross with and without assumed fixation of QTL alleles in different lines)

Outbred (line) cross	Backcross	F_2
Fixed grandparental lines have been assumed	2 alleles 2 genotypes	2 alleles 3 genotypes
General (no assumed fixation)	3 alleles 4 genotypes	4 alleles 4 genotypes

Here genotype AB is considered to be the same as BA. F_2 without assumed fixation corresponds to a full-sib family of outcrossing (cross-pollinating) species.

parental origin-coded) haplotype of the left and the right flanking object of the qth QTL in offspring i. We denote by $r_q = (r_{q1}, r_{q2})'$ the resulting recombination fractions between the QTL at l_q and the corresponding flanking object (after an application of Haldane's map function). Here we assume that the recombination rates in male and female meioses are the same (even though an extension to different recombination fractions would be straightforward; see Haley *et al.* 1994). The prior distribution of ordered genotypes of QTL is now assumed to have the product form $p(\chi \mid H^*, l, m, N_{\text{qtl}}) = \prod_{q=1}^{N_{\text{qtl}}} p(x_q|x_1, \ldots, x_{q-1}, H^*, l, m) = \prod_{q=1}^{N_{\text{qtl}}} \prod_{i=1}^{N_{\text{ind}}} p(x_{qi}|H^{*q}_{i,\text{LR}}, r_q)$. Note that QTL are not automatically (conditionally) independent from each other (see Sillanpää and Arjas 1998).

The QTL analysis of the offspring is done in terms of parental haplotypes. The numbers of possible QTL alleles and QTL genotypes in BC and F_2 designs are found in Table 1. Given the QTL genotype vector $\alpha = (\alpha_1, \ldots, \alpha_{N_{\text{gen}}})$, the prior probabilities for $s = 1, \ldots, N_{\text{gen}}$ are calculated from the equation

$$p(x_{qi} = \alpha_s|H^{*q}_{i,\text{LR}}, r_q) = \frac{p(x_{qi} = \alpha_s|H^{*q}_{i\text{L}}, r_q) \times p(H^{*q}_{i,\text{R}}|x_{qi} = \alpha_s, r_q)}{p(H^{*q}_{i\text{R}}|H^{*q}_{i\text{L}}, r_q)} \quad (6)$$

Here $H^{*q}_{i\text{L}} = (\mathcal{G}^{\text{F}}_{\text{L}(q),i}(H^*), \mathcal{G}^{\text{M}}_{\text{L}(q),i}(H^*))$, and $H^{*q}_{i\text{R}} = (\mathcal{G}^{\text{F}}_{\text{R}(q),i}(H^*), \mathcal{G}^{\text{M}}_{\text{R}(q),i}(H^*))$ are the left- and right-ordered flanking object (QTL or marker) genotypes in the grandparental origin form. Haplotype coding and the evaluation of the probability in (6) in F_2 and backcross designs are illustrated in Figures 1 and 2.

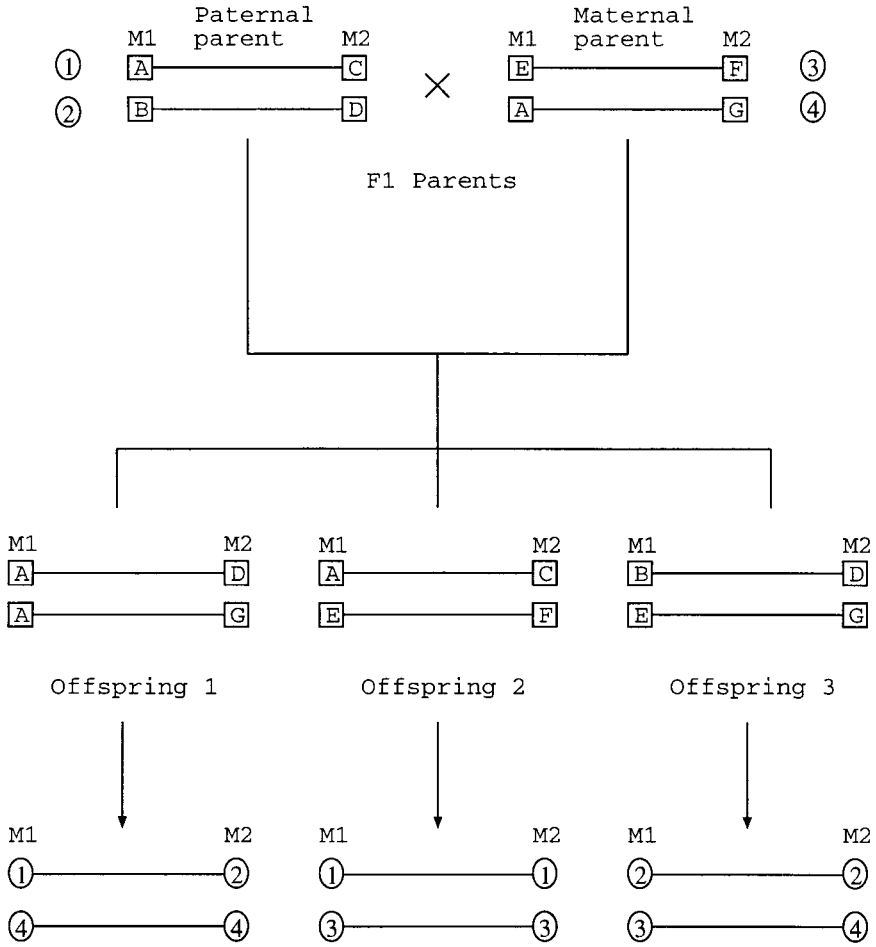


Figure 1.—Genotypes of two marker positions in an F_2 full-sib family with three offspring (an example). Haplotypes of all individuals are known. Arrows indicate how alleles are coded with respect to their grandparental "line" origins. The paternal (maternal) haplotype of offspring becomes a sequence of lines 1 and 2 (3 and 4). For an illustration, suppose that there is a QTL (q) between the markers and that each haplotype of the parents contains a different QTL allele (also numbered from 1 to 4) in that locus. The four QTL genotypes are then combinations of parental chromosomes and they show the following correspondence here: 13 = (A–C, E–F), 14 = (A–C, A–G), 23 = (B–D, E–F), and 24 = (B–D, A–G). Denote the recombination fraction between the flanking markers by r_{LR}, and that between the QTL and the left (right) flanking marker by r_{q1} (r_{q2}). By applying Equation 6, the probability of QTL genotype 13 occurring in offspring 1 is given by $((1 - r_{q1})r_{q2} \times r_{q1}r_{q2}) / (r_{\text{LR}}(1 - r_{\text{LR}}))$, that of 14 by $((1 - r_{q1})r_{q2} \times (1 - r_{q1})(1 - r_{q2})) / (r_{\text{LR}}(1 - r_{\text{LR}}))$, that of 23 by $(r_{q1}(1 - r_{q2}) \times r_{q1}r_{q2}) / (r_{\text{LR}}(1 - r_{\text{LR}}))$, and that of 24 by $(r_{q1}(1 - r_{q2}) \times (1 - r_{q1})(1 - r_{q2})) / (r_{\text{LR}}(1 - r_{\text{LR}}))$.

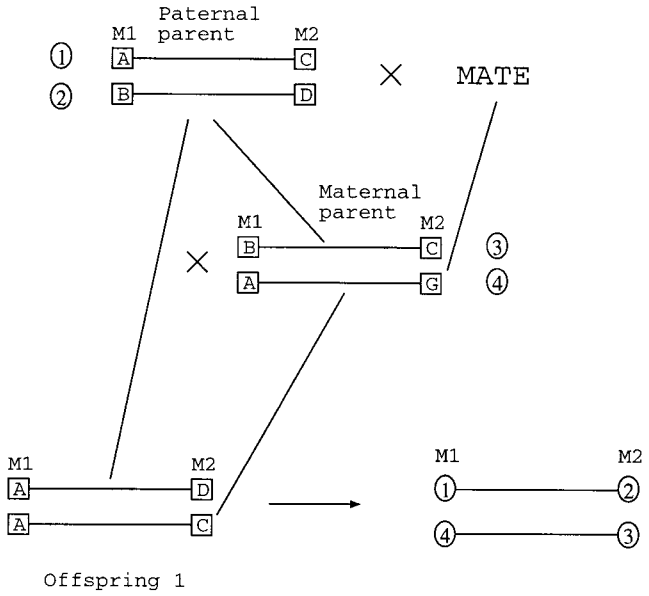


Figure 2.—Genotypes at two marker positions in a backcross family with one offspring (an example). Haplotypes of all individuals are known. Arrows indicate how alleles are coded with respect to their grandparental origins. Paternal parent of backcross progenies is also one of the grandparents, *i.e.*, code 3 means {1 or 2}. In case fixed (grandparental) lines are assumed, codes 2 and 3 can be replaced by 1, and thus paternal meioses are not considered in the QTL genotype probability calculations at all, *i.e.*, if an offspring chromosome inherited from the nonfounder parent, here the maternal parent, is of type 1–2, then for a QTL between the markers the genotype probabilities of types 1*X* and 2*X* are given by $(1 - r_{q1})r_{q2}/r_{LR}$ and $r_{q1}(1 - r_{q2})/r_{LR}$, respectively; here *X* indicates the other (not considered) allele.

SIMULATION ANALYSIS

To test the performance of this method, an outcrossing F_2 population consisting of $N_{ind} = 200$ offspring was generated by a simulation program provided by J. W. Van Ooijen (Centre for Biometry Wageningen, CPRO-DLO, The Netherlands). We considered two 100-cM long chromosomes, both having 11 evenly spaced markers, at every 10 cM. The simulated trait had a genetic (QTL) variance 4.47 and a phenotypic variance 6.35, resulting in heritability 0.7. Two sets of parental crosses were generated: In the first set the parental mating type was fully informative (AB × CD) at all marker loci, and in the second set the degree of informativeness, as well as the corresponding linkage phases, varied from locus to locus. The simulated true underlying parental cross in the second set is shown in Figure 3; it is underlying in the sense that after the simulation this information was "forgotten" and not used in the Bayesian analyses (as explained below). The genotype-specific phenotype effects and the locations of the three simulated QTL can be found from Table 2. All haplotypic assignments in the offspring were assumed unknown. In the statistical analyses, three specifications regarding the amount of parental information were considered: (1) All genotypes and haplotypic assignments in parents were assumed known; (2) all genotypes were assumed known but their phases unknown in parents; and (3) all parental and grandparental marker information was assumed unknown (missing). The performance of our method was compared to that of "all-markers" interval mapping (IM; Maliepaard and Van Ooijen 1994) and to multiple QTL mapping with two background controls [MQM/02; both implemented in the MAPQTL program of Van Ooijen and Maliepaard 1996; MAPQTL (tm) version 3.0; CPRO-DLO, Wageningen, The Netherlands]. Note that in the IM and MQM methods the genotypes and the linkage phases in parents must be known.

In addition, the simulated data in which each QTL had four alleles were analyzed (in cases 1 and 3), having incorrectly assumed fixed grandparental lines (where

CHROMOSOME 1

Markers	0	1	2	3	4	5	6	7	8	9	10
	a	a	a	a	a	a	a	a	a	a	a
	a	b	b	b	a	a	b	b	b	b	b
×	a	c	a	a	a	a	c	a	b	a	a
	b	d	b	a	b	b	d	b	a	a	b

Figure 3.—Parental cross (maternal parent × paternal parent) in two simulated chromosomes.

CHROMOSOME 2

Markers	0	1	2	3	4	5	6	7	8	9	10
	a	a	a	a	a	a	a	a	a	a	a
	b	b	b	b	b	b	b	b	b	b	b
×	a	b	c	a	a	a	c	c	c	a	b
	a	a	d	b	a	b	d	d	d	a	a

TABLE 2

The locations and the individual phenotypic effects (eff{·}) of different heterozygous genotypes of the three simulated QTL

Chromosome	Left marker	Location(cM)	Eff{AC}	Eff{AD}	Eff{BC}	Eff{BD}
1	3	32.7	−2.3	−0.7	+0.7	+2.3
1	5	58.0	+0.3	+1.1	−0.3	−1.1
2	4	41.2	+1.5	+1.5	−1.5	−1.5

The left column refers to the chromosome in which the considered QTL is located. The next column refers to the nearest left flanking marker of the QTL in the chromosome. Location is the distance (in centimorgans) between the QTL position and the leftmost marker in the linkage group. Parental mating type (maternal parent × paternal parent) in all three QTL positions is AB × CD.

TABLE 3

The ranges $R(\cdot)$ of the proposal distributions for different parameters, the corresponding proposal probabilities, the numbers of iterations, and the indices of the background control markers from other chromosomes, which were used in the simulation analyses

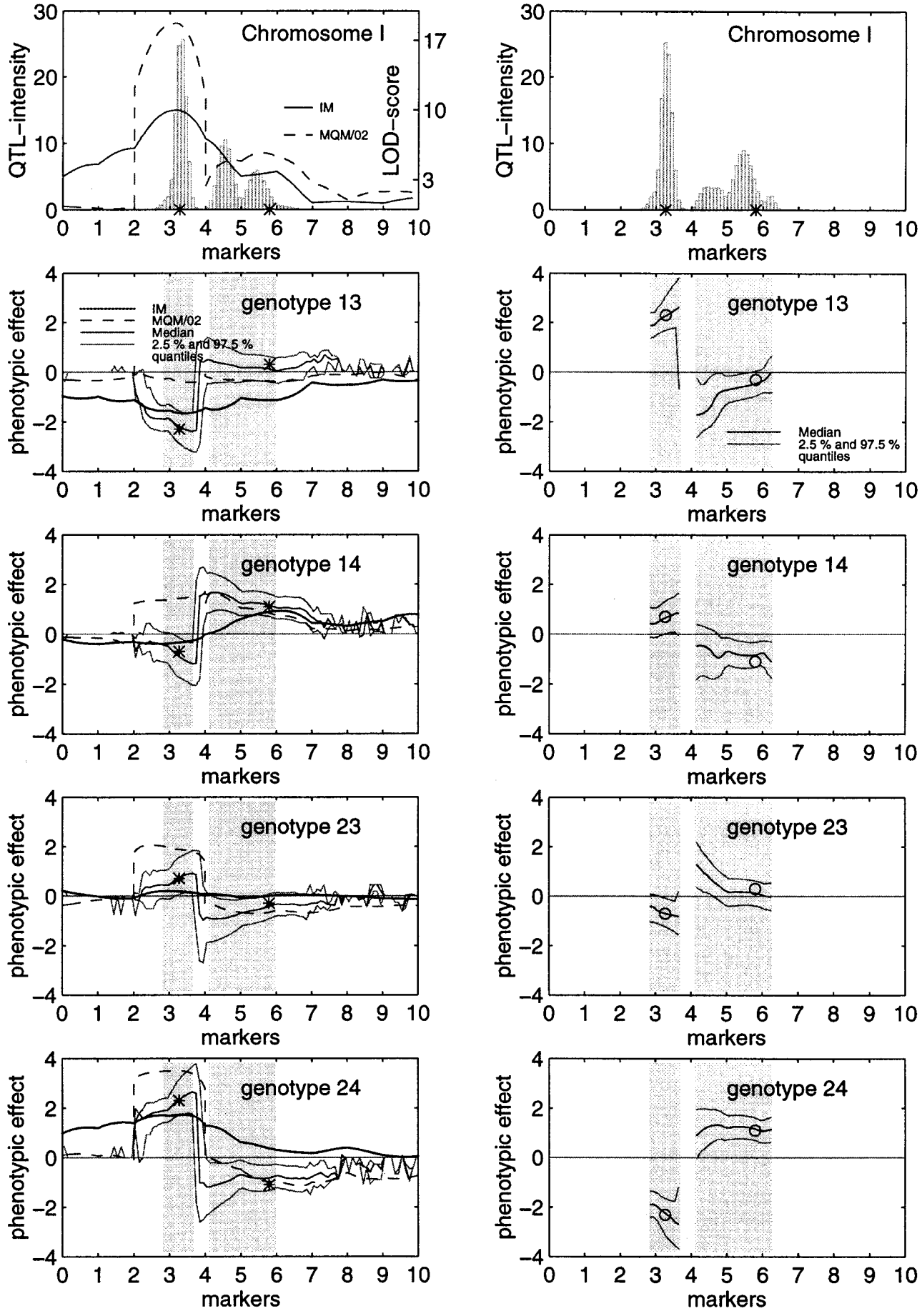
	$R(l_q)$	$R(a)$	$R(\sigma)$	$R(b_{qj})$	$R(c_{kj})$	$p_a = p_d$	No. of iterations	BGCs
Chromosome 1	2.0	1.0	0.2	1.5	2.0	1/3	5,000,000	3
Chromosome 2	2.0	1.0	0.2	1.5	2.0	1/3	5,000,000	4

BGCs, background control markers.

grandfathers were assumed to originate from the same line). This was done to see how this erroneous assumption influences the results.

In all Bayesian analyses described here, our C-program implementing a Metropolis-Hastings chain was run 5,000,000 cycles in a Pentium II/266MHz computer. No values were deleted because of burn-in, but the chain was thinned so that only every fifth iteration was saved, resulting in 1,000,000 sampled values for each parameter. After a preprocessing stage (see appendix a), background controls were chosen. When analyzing a real data set, they can be determined by a single marker regression or by performing several analyses. Here, however, we simply chose marker 3 in chromosome 1 and marker 4 in chromosome 2 as background controls. Very likely, a few reanalyses would have led to the same conclusion. As no covariates (age, sex, etc.) were used, there was a common intercept ($\rho = a$ and $B_i = 1$ for all i). The running times, in circumstances where there was practically no other load in the computer, varied around 9 hr. The initial value for the number of QTL was three, and the corresponding locations were 20.0 cM, 50.0 cM, and 80.0 cM. The Poisson mean (hyperparameter) was set to $\lambda = 2$ and the maximum number of QTL (in the analyzed chromosome) to three. The residual standard deviation was chosen to be uniform over the range [0.0, 2.55], the right endpoint being equal to the phenotypic standard deviation estimate from the data. The prior of the intercept was taken to be uniform on [−13, 13], those of the QTL genotypic regression coefficients were independent normal distributions with mean zero and variance 100, and the prior of the background control genotypic regression coefficients was uniform on [−13, 13]. Finally, the prior of the QTL locations was uniform over [0, 100]. The control parameter values used in the final analyses are given in Table 3. The proposal distribution for the genotypic effects (coefficients) was chosen to be $N(0, 0.5)$ in cases where the addition of a new QTL to the model was proposed.

Figure 4.—Results from the estimation when all markers are fully informative. (Top) Graphs of the posterior QTL intensity in chromosome 1 when all parental genotypic information is known (left), and when only parental genotypes (but not linkage phases) are known (right). Here the histogram corresponds to the (approximate) posterior QTL intensity over the chromosome, with binlength 1 cM. On the top left, the results from interval mapping (IM, solid line) and multiple QTL mapping with two background controls (MQM/02, broken line) are shown. The left (right) *y*-axis corresponds to the posterior QTL intensity (LOD score). Note the logarithmic scale of the LOD score. From the remaining eight panels, four (on the left) describe phenotypic effect estimates of different genotypes in chromosome 1 with all parental genotypic information known, and four (on the right) those in chromosome 1 when only parental genotypes are available. The solid line is the pointwise posterior median, and the gray lines the 2.5 and 97.5% quantiles of the posterior distribution, of the phenotypic effect of a putative QTL. (*) The simulated true QTL. (o) The labeling corresponding to best fit of the phenotypic effects of the QTL to their estimates. Shaded regions are suggested credible intervals for QTL localization. The estimated phenotypic effects are reliable only in these regions.


Chromosome I
QTL-intensity
LOD-score
IM
MQM/02
markers
genotype 13
phenotypic effect
IM
MQM/02
Median
2.5 % and 97.5 % quantiles
genotype 14
genotype 23
genotype 24

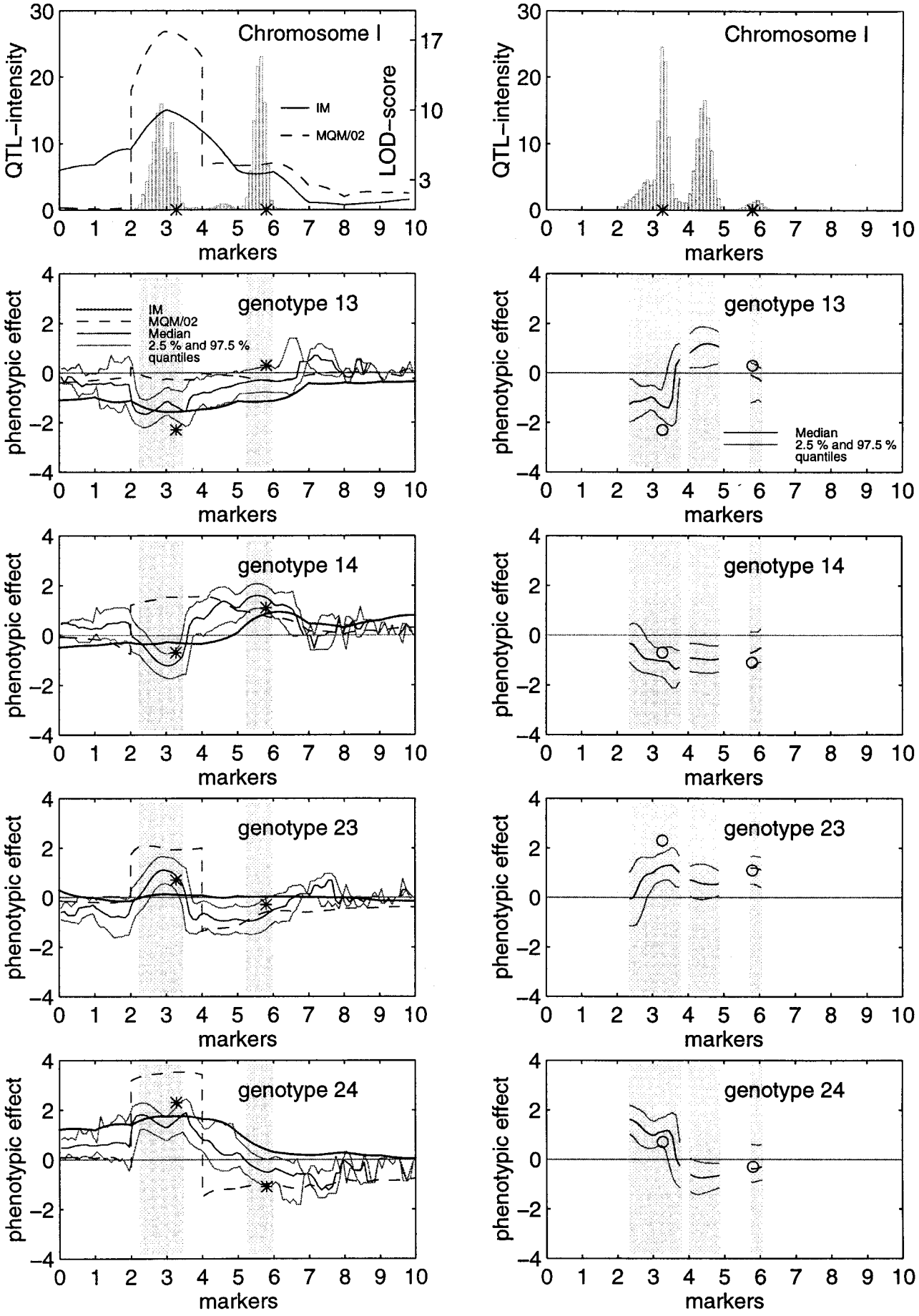

Chromosome I
QTL–intensity
LOD–score
IM
MQM/02
markers
genotype 13
genotype 14
genotype 23
genotype 24
phenotypic effect
IM
MQM/02
Median
2.5 % and 97.5 % quantiles
Median
2.5 % and 97.5 % quantiles

TABLE 4

The posterior distribution of the number of QTL and its expectation in four analyses of chromosome 1

	$P(N_{qtl} = x \mid y, I, H, X_o, m)$				
	$x = 0$	$x = 1$	$x = 2$	$x = 3$	$E(N_{qtl} \mid y, I, H, Xo, m)$
Nonconstant information					
Complete parental info.	0.0000	0.0479	0.9312	0.0209	1.9730
Partial parental info.	0.0000	0.0460	0.9379	0.0161	1.9701
Fully informative data					
Complete parental info.	0.0000	0.0775	0.9212	0.0013	1.9237
Partial parental info.	0.0000	0.1097	0.8881	0.0021	1.8924

"Complete parental info." refers to the analysis where all parental genotypes and linkage phases were known and "partial parental info." refers to the case where only parental genotypes were available.

In the IM and MQM/02 analyses, walking speed was set to 0.5 cM, which is the smallest admissible value in the MAPQTL software. We used the same background controls in MQM/02 as in the Bayesian analyses.

RESULTS

The Bayesian posterior QTL intensities (see appendix b) in chromosome 1, when all parental information was present (case 1) or when parental linkage phases were absent (case 2), are shown in Figure 4 (top) when all markers are fully informative, and Figure 5 (top) when marker information varies from marker to marker. The curves consisting of the pointwise medians and the 2.5 and 97.5% quantiles of the posterior distribution of the phenotypic effects of the four genotypes, as functions of the putative QTL location, are shown in the same figures when all parental information is present (left), or when parental linkage phases are unknown (right). Approximate posterior distributions of the number of QTL in chromosome 1, obtained from these four different analyses, are shown in Table 4. The analyses where all parental information was absent (case 3) are not summarized in figures or in tables. This is because in theory case 3 is not fully identifiable, resulting in probabilistic summary measures (the posterior QTL intensity and the posterior distribution of the number of QTL) that are not unique. These problems are described and considered more in the discussion.

Table 5 gives a brief summary of our findings concerning the localization of QTL as suggested by the QTL intensities in Figures 4 and 5. The table makes direct reference to (approximate) posterior probabilities that a particular chromosomal region I of high QTL-intensity concentration contains a given number of QTL. Also the corresponding posterior expectations are calculated. The analyses support quite strongly the hypothesis of two QTL in chromosome 1.

In the analyses where all markers were fully informative (Figure 4, top), the two posterior QTL-intensity graphs (from cases 1 and 2) became nearly identical, regardless of whether parental linkage phase information was available or not. Both posterior QTL-intensity graphs were nicely concentrated around the left QTL at 32.7 cM. The graphs surrounding the right (weaker) QTL at 58 cM were much wider, and there was also some bias to the left. However, the true simulated QTL is still inside the regions [41 cM, 60 cM] and [41 cM, 63 cM] of elevated posterior QTL intensities. In this case (Figure 5, top left), the MQM analysis performed well in both QTL localizations in chromosome 1, but the IM analysis managed to localize only the left QTL. (Note that the posterior QTL-intensity graphs covering the regions [41 cM, 60 cM] and [41 cM, 63 cM] are

Figure 5.—Results from the estimation when marker information varies from marker to marker. (Top) Graphs of the posterior QTL intensity in chromosome 1 when all parental genotypic information is known (left), and when only parental genotypes (but not linkage phases) are known (right). In these panels, the histogram corresponds to the (approximate) posterior QTL intensity over the chromosome, with binlength 1 cM. (top left) The results from interval mapping (IM, solid line) and multiple QTL mapping with two background controls (MQM/02, broken line) are shown. The left (right) *y*-axis corresponds to the posterior QTL intensity (LOD score). Note the logarithmic scale of the LOD score. From the remaining eight panels, four (on the left) describe phenotypic effect estimates of different genotypes in chromosome 1 with all parental genotypic information known, and four (on the right) those in chromosome 1 when only parental genotypes are available. The solid line is the pointwise posterior median, and the gray lines the 2.5 and 97.5% quantiles of the posterior distribution, of the phenotypic effect of a putative QTL. (*) The simulated true QTL. (o) The labeling corresponding to best fit of the phenotypic effects of the QTL to their estimates. Shaded regions are suggested credible intervals for QTL localization. The estimated phenotypic effects are reliable only in these regions.

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 in four analyses

Chromosome 1	I	Length (I)	$P(N^I_{qtl} \geq 1 \mid \text{data})$	$E(N^I_{qtl} \mid \text{data})$
Nonconstant information				
Complete parental info.	[22 cM, 35 cM]	13 cM	0.63	0.9955
Complete parental info.	[52 cM, 60 cM]	8 cM	0.59	0.8975
Partial parental info.	[23 cM, 38 cM]	15 cM	0.65	1.0411
Partial parental info.	[40–49] and [57–61]	13 cM	0.57	0.8502
Fully informative data				
Complete parental info.	[28 cM, 37 cM]	9 cM	0.63	0.9881
Complete parental info.	[41 cM, 60 cM]	19 cM	0.59	0.8933
Partial parental info.	[28 cM, 37 cM]	9 cM	0.63	0.9858
Partial parental info.	[41 cM, 63 cM]	22 cM	0.58	0.8681

The (posterior) expected number of QTL in I, calculated as the integral of the QTL intensity over I, is also determined. "Complete parental info." refers to the analysis where all parental genotypes and linkage phases were known and "partial parental info." refers to the case where only parental genotypes were available.

multimodal. This is apparently the same phenomenon that is typical to the LOD-score curve at marker points: often there is more evidence, because of marker genotyping, against placing a putative QTL exactly at a marker locus than against placing it somewhere nearby.) The graph leaves somewhat uncertain why, of the two modes, the one that is farther away from the true simulated QTL at 58 cM ended up being higher in the first case.

It can be seen from Figure 5 that the nonconstant marker information analysis (case 1) results in high posterior QTL intensities surrounding both simulated QTL in chromosome 1. The IM and MQM analyses localized quite well the "left" QTL at 32.7 cM, but localization of the "right" QTL at 58 cM was poor with both methods. Somewhat surprisingly, in the Bayesian method, the left, more influential, QTL was not localized as accurately as the right QTL when linkage phases were available in parents. This may be a consequence of the fact that there is a highly informative marker very close to the right QTL, whereas this is not the case with the left QTL (see Table 6). As could be expected, the localization was somewhat less accurate when the parental genotypes or their linkage phases were not available.

Consider next the estimation of the phenotypic effects, indicated by asterisks in Figures 4 and 5. As could be expected, the estimation was most successful in the case (displayed in Figure 4, left) where marker information was complete and where complete parental information was available. In the case of nonconstant marker information, but still assuming complete knowledge of the parental genotypes and linkage phases, the estimates were somewhat less accurate, with some of the true values being just outside the 95% credible boundaries (Figure 5, left). When analyzing real data, the true labeling [*i.e.*, assigning of the QTL genotypes (13, 14, 23, 24) to the true grandparental alleles] of the phenotypic effects is almost always unknown (except for the QTL

TABLE 6

Estimated informativeness of different marker loci of the simulated data set, with two degrees of parental genotype information

	Markers										
	0	1	2	3	4	5	6	7	8	9	10
					Chromosome 1						
Parents known	0.5	1.0	0.485	0.5	0.5	0.5	1.0	0.5	0.575	0.5	0.495
Parents unknown	0.667	1.0	0.485	0.667	0.667	0.667	1.0	0.5	0.575	0.667	0.495
					Chromosome 2						
Parents known	0.5	0.5	1.0	0.505	0.5	0.465	1.0	1.0	1.0	0.5	0.48
Parents unknown	0.667	0.5	1.0	0.505	0.667	0.465	1.0	1.0	1.0	0.667	0.48

In the first case, parental mating type (with or without knowing their haplotypic arrangements) is known in each marker locus (parents known) and in the second, all parental information is unknown (parents unknown). In the latter case, marker informativeness, *i.e.*, the proportion of offspring alleles whose grandparental origin can be uniquely determined at a locus, is calculated as an expectation (weighted sum) over consistent parental mating types.

TABLE 7

Point estimates and their support regions from different analyses

QTL	IM	MQM	I_1	I_2
		Nonconstant information		
32.7	30.0 [26.5, 35.5]	29.5 [26.5, 36.0]	28.25 [22, 35]	32.85 [23, 38]
58.0	60.0 [0.0, 63.5]	60.0 [20.0, 66.0]	56.45 [56, 60]	44.45 [40, 49] and [57, 61]
		Fully informative data		
32.7	31.5 [26.0, 36.5]	31.5 [27.0, 34.5]	32.95 [28, 37]	32.85 [28, 37]
58.0	60 [0.0, 63.5]	55.5–56.0 [45.0, 64.0]	45.55 [41, 60]	54.55 [41, 63]

True QTL locations, their estimated locations and one-lod-support intervals from the IM and MQM analyses, and Bayesian point estimates (modes of the QTL intensity) together with suggested support intervals (I_1) when parental genotypes and haplotypes are available, and (I_2) when no parental haplotypes are available (from Table 5). LOD score was evaluated every 0.5 cM in the IM and MQM estimation. The posterior modes (intervals) were obtained with binlength 0.1 cM (1.0 cM), using all 5,000,000 sampled values in I_1 and 1,000,000 values in I_2. Note that in the IM and MQM analyses, there is no counterpart to estimates in column I_2.

genes that have been positionally cloned). If parental genotype and/or linkage phase information are missing, the labeling of the genotypic effects according to the grandparental origin of the alleles also becomes nonunique in the simulated case. For this reason, when comparing the phenotypic effect estimates with the true values used in the simulation, we have to make sure that each estimate is matched correctly with a combination of two grandparental QTL alleles. Such reassignment of the QTL genotypes is indicated on the right-hand side of Figures 4 and 5 by circles. In chromosome 1, note that the genotype labels are not consistent with each other in case 2.

The performance of the IM and MQM methods in the estimation of the phenotypic coefficients of the putative QTL was not particularly good. Moreover, they do not provide confidence intervals for such point estimates. Confidence intervals would have to be determined separately, for example, by employing bootstrap techniques.

The point estimates of QTL locations and their support regions are summarized in Table 7 for four different analyses of chromosome 1.

When considering chromosome 2 (which was analyzed only in cases 1 and 3), the posterior QTL-intensity graphs (see Figure 6) were all nicely concentrated around the simulated true QTL at 41.2 cM, regardless of whether the markers were fully informative or not. Also, the IM and MQM methods were able to localize the QTL at 41.2 cM quite well.

The performance of the analyses (cases 1 and 3), when it was incorrectly assumed that the grandparental lines are fixed (pictures not shown), was quite poor in chromosome 1. The only exception was the case where all parental information was available and all markers were fully informative. Then the simulated QTL at 32.7 cM was localized rather well, and there was also some indication of QTL activity around the QTL at 58 cM. Assuming fixation in the situation where all markers were fully informative but where all parental information was absent, only the latter QTL resulted in a high (but broad) QTL-intensity concentration.

DISCUSSION

We have presented here a Bayesian procedure for mapping multiple QTL from incomplete outbred offspring data, thus extending our earlier method (Sillanpää and Arjas 1998) to a more general experimental design. A test version of the software (written in C language) is available at http://www.rni.helsinki.fi/~mjs/. The method is capable of handling situations where marker information from parents and/or grandparents is missing in varying degrees, as well as cases where some of the marker information from the offspring is unavailable. In contrast to Sillanpää and Arjas (1998), the present model was not overparameterized, because this did not seem to improve the mixing properties of the sampler.

Following Sillanpää and Arjas (1998), we use the posterior QTL intensity as a probabilistic summary measure for the localization of QTL. During the MCMC sampling, we do not restrict the order of the QTL in any way to label them. If order-based labeling is preferred, it can be established afterward from the MCMC realizations. This is an alternative to imposing constraints on the MCMC simulation as was done, *e.g.*, in Satagopan *et al.* (1996), Satagopan and Yandell (1996), Richardson and Green (1997), and in Uimari and Hoeschele (1997).

We tested the performance of our method by using simulated F_2 data sets (two informativeness levels), with varying degrees of parental marker information (three levels). It seems intuitively plausible, and it also became clear from our simulations, that the availability of parental linkage phase information is more important in the case where the markers are not fully informative. The situation where also a part of the offspring marker genotypes is missing was not considered in the test analyses.

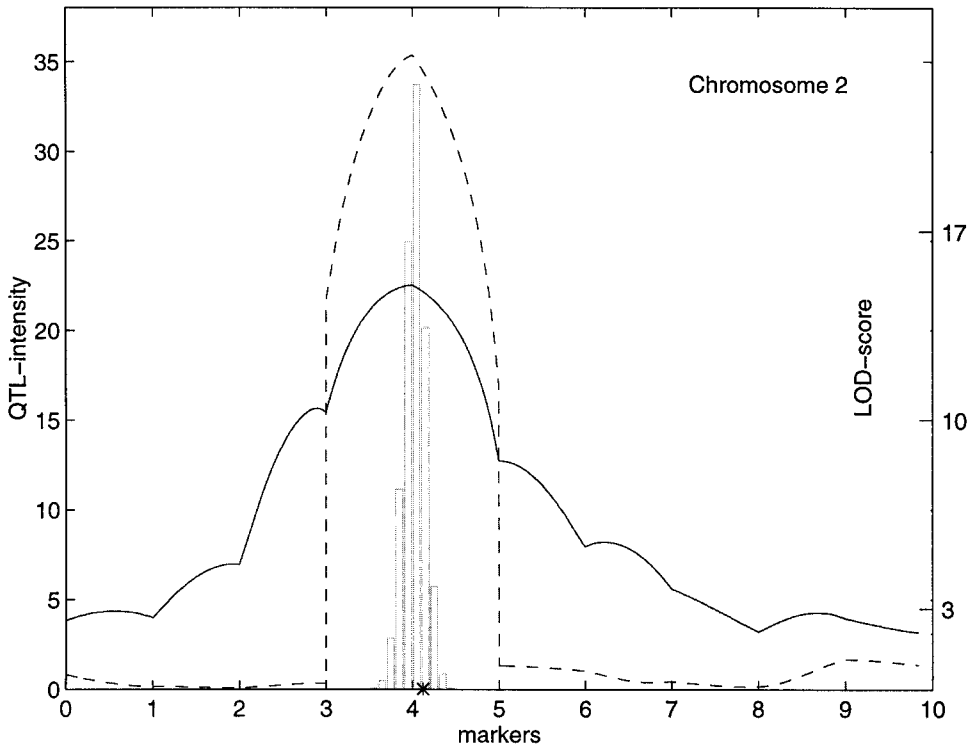


Figure 6.—The posterior QTL intensity in chromosome 2 when all parental haplotype information is known and the marker information varies from marker to marker. The results from interval mapping (IM, solid line) and multiple QTL mapping with two background controls (MQM/02, broken line) are shown. The histogram corresponds to the (approximate) posterior QTL intensity over the chromosome, with binlength 1 cM. The left (right) *y*-axis corresponds to the posterior QTL intensity (LOD score). (*) The simulated true QTL.

Standardization of the phenotypic data is recommended before applying Bayesian QTL mapping in practice. Then the same proposal windows and other control parameters can be applied to different data sets, instead of performing separate test trials for each. Another advantage is that the numerical accuracy may be improved because computers' ability to store floating point numbers is maximal when dealing with numbers between zero and one.

The marker covariates can be chosen by an application of simple linear regression at each marker (putative QTL) position, omitting individuals whose genotype at that locus was unknown (because data augmentation would need linkage phase information). In doing so, one should pay attention to how much information a potential covariate marker carries and how many missing values there are. If an interesting region does not contain any fully informative markers, one can often find two closely linked markers such that each marker alone is informative only with respect to one (and a different) parent.

Parental mating type is usually not constant in outcrossing experiments. Thus a systematic application of some index describing the proportion of informative meioses locally present in the data will help the analyst to quantify the possibility of localizing a QTL in different areas of the considered chromosome. One such measure is displayed in Table 6. The influence of marker informativeness (*cf.* marker polymorphism in Kruglyak 1997) can be seen clearly from our simulation analysis (Table 6 and Figures 4 and 5) where, in the uninformative areas, intensity graphs are much more spread out, or even biased in some direction.

The phenotypic effects can be estimated reliably only in chromosomal regions in which the posterior QTL intensity is sufficiently high. As an alternative to the locationwise posterior densities for phenotypic effects shown in Figures 4 and 5, the posterior density can be constructed as an expectation over several pointwise values (of phenotypic effects), each being associated with a putative QTL location within a particular region of high posterior QTL intensity. One such posterior density is shown in Figure 7.

There appear to be two possible philosophies about how the indexing of QTL genotypes should be interpreted. Considering QTL genotype 13, for example, the first interpretation says that lines 1 and 3 are names for the parental haplotypes. In this case the remaining uncertainty concerning linkage phase is in how the grandparental alleles are assigned to these haplotypes. According to the second interpretation, lines 1 and 3 are names for the grandparental lines (alleles), and uncertainty is in the assignment of the parental haplotypes to these lines. Obviously, these two ways of thinking lead to different results only when there is some uncertainty in the parental linkage phases. We have adopted here the first interpretation, even though the second one is in some sense more fundamental in the context of QTL mapping.

We stress that in situations where all parental information is missing (case 3) it will be problematic to assign unique grandparental origins to the estimated pheno-

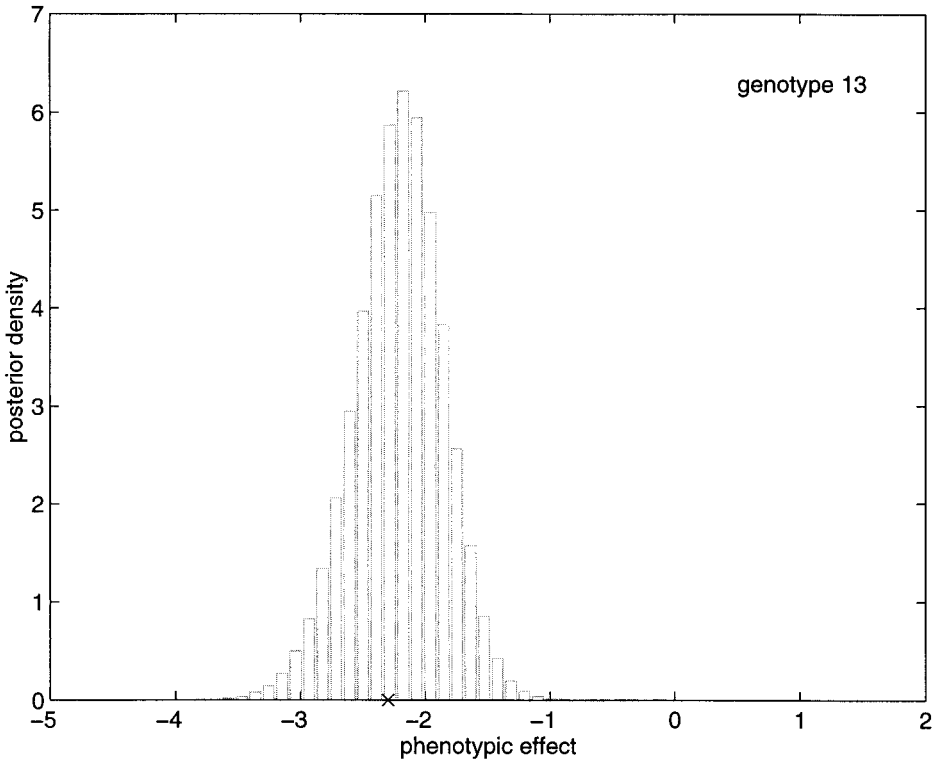


Figure 7.—Approximate posterior distribution of the phenotypic effect of genotype 13 of the left QTL (chromosome 1), determined from the interval from 27 cM to 39 cM. All parental information was available and all markers were fully informative.

type effects. In this situation, both parents have symmetric pairs of haplotype configurations that are *a posteriori* equally likely to be the correct underlying mating structure. As a consequence, under these circumstances the correspondence between QTL genotypes (13, 14, 23, and 24; *cf.* Figure 1) and their grandparental alleles is not unique. In our program, the assignment can actually change from one iteration cycle to another within one MCMC run, let alone in different runs. (In practice such changes are rare because of the strong local dependence between offspring and their parents and between adjacent loci.) In case 3, the parental phase reconstruction can actually change suddenly in some region of the chromosome to a symmetrical mating type. (This can only be checked from the simulated data.) Also the resulting posterior QTL-intensity curves can differ in such regions in different MCMC runs.

In cases 1 and 2, the very strong local dependency structure between parents and offspring and between adjacent loci will in practice prevent such phase transitions during the same MCMC run. Therefore, to avoid problems of this kind, we strongly recommend that at least one of the parents should be genotyped in several marker loci along the chromosome, as equidistant as is possible.

Locally, of course, if there is a fully informative (reference) marker, in case 3 we can also avoid such identifiability problems and averaging in estimation by fixing the assignments (segregation indicators) arbitrarily at the reference marker and then using the fact that, as long as the genetic distance from the marker is short, haplotype assignment can be made in a way that is with high probability consistent with that chosen at the reference marker locus. If this informative marker is near a contemplated QTL, this technique will also facilitate the estimation of the corresponding phenotypic effects, by keeping the four haplotypic assignments (and thus the corresponding QTL allele combinations) apart. A more negative aspect of this technique is that it works only locally, as simultaneous haplotype assignments at two or more marker positions might not agree with the true haplotype configuration. As a consequence, the estimation would need a new MCMC run for each such local assignment.

M.S. thanks Matti Taskinen for his advice in the programming work, and Päivi Hurme and Outi Savolainen for many useful discussions about the designs. We are grateful to Johan Van Ooijen for providing his simulation program, which was used to generate test data sets, and to Pekka Uimari and three anonymous referees for their constructive comments on the manuscript. This work was supported by a research grant (no. 38352) from the Academy of Finland, and by the ComBi Graduate School.

LITERATURE CITED

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

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.

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.
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.
Jansen, R. C., D. L. Johnson and J. A. M. Van Arendonk, 1998 A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. Genetics **148:** 391–399.
Janss, L. L., G. R. Thompson and J. A. M. Van Arendonk, 1995 Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor. Appl. Genet. **91:** 1137–1147.
Jensen, C. S., and A. Kong, 1997 Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. Manuscript available at MCMC preprint service (http://www.stats.bris.ac.uk/MCMC/).
Jensen, C. S., and N. Sheehan, 1998 Problems with determination of noncommunicating classes for Monte Carlo Markov Chain applications in pedigree analysis. Biometrics **54:** 416–425.
Kao, C.-H., and Z.-B. Zeng, 1997 General formulas for obtaining the MLEs and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53:** 653–665.
Knott, S. A., D. B. Neale, M. M. Sewell and C. S. Haley, 1997 Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. Theor. Appl. Genet. **94:** 810–820.
Kong, A., 1991 Analysis of pedigree data using methods combining peeling and Gibbs sampling, pp. 379–385 in *Computer Science and Statistics Proceedings of the 23rd Symposium on the Interface*, edited by E. M. Keramidas and S. M. Kaufman. Interface Foundation, Fairfax Station, VA.
Kruglyak, L., 1997 The use of a genetic map of biallelic markers in linkage studies. Nat. Genet. **17:** 21–24.
Kruglyak, L., M. J. Daly and E. S. Lander, 1995 Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am. J. Hum. Genet. **56:** 519–527.
Lander, E. S., and P. Green, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.
Lin, S., 1995 A scheme for constructing an irreducible Markov Chain for pedigree data. Biometrics **51:** 318–322.
Lin, S., E. Thompson and E. Wijsman, 1994 Finding noncommunicating sets for Markov Chain Monte Carlo estimation on pedigrees. Am. J. Hum. Genet. **54:** 695–704.
Maliepaard, C., and J. W. Van Ooijen, 1994 QTL mapping in a full-sib family of an outcrossing species, pp. 140–146 in *Biometrics in Plant Breeding: Applications of Molecular Markers*, edited by J. W. Van Ooijen and J. Jansen. CPRO-DLO, Wageningen, The Netherlands.
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, 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.
Sheehan, N., and A. Thomas, 1993 On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics **49:** 163–175.
Sillanpää, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373–1388.
Sobel, E., and K. Lange, 1996 Descent graphs in pedigree analysis: application to haplotyping, location scores, and marker-sharing statistics. Am. J. Hum. Genet. **58:** 1323–1337.
Thompson, E. A., 1994 Monte Carlo likelihood in genetic mapping. Stat. 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.
Van Ooijen J. W., and C. Maliepaard, 1996 Plant Genome IV. Abstract at: http://probe.nalusda.gov:8000/otherdocs/pg/pg4/abstracts/p316.html/.
Wijsman, E. M., 1987 A deductive method of haplotype analysis in pedigrees. Am. J. Hum. Genet. **41:** 356–373.
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: PREPROCESSING AND PARAMETER ESTIMATION

Before the actual statistical analysis, the data go through a preprocessing stage. In this process, we infer as much of the marker genotype and linkage phase information as is possible by direct logical deduction from known parts of the family structure. The deduction rules applied here (sequentially until there are no new assignments) are similar to the genotyping rules of Wijsman (1987). If grandparental genotypes are present, these deduction rules are first applied to the grandparents and parents, and then to the parents and offspring. In this process, sets of consistent parental mating types are determined for each marker (see below) and they are later repeatedly applied for the estimation.

Let us consider a multi-allelic marker in the chromosome to be analyzed, where, after the logical deductions, the genotypes of the parents are still unknown. Further, consider the genotype or complete haplotype imputations for both parents by updating them one at a time. In such situations, when the genotype of one parent has been imputed, some offspring genotypes may in fact uniquely determine the genotype of the other parent. To avoid this and to make the sampler work more efficiently, genotypes of parents are considered jointly, and they have to form a pair that is consistent with the offspring genotypes. Therefore, we go through all possible allele combinations in parents, one at a time at each marker locus, and check whether any of them is inconsistent with the offspring genotypes. All inconsistent pairs are eliminated. In a backcross, one needs to check an additional consistency in genotypes of related parents.

Sometimes a block-update is preferred over a single-site-update in MCMC applications to pedigrees (Kong 1991; Janss *et al.* 1995; Heath 1997; Jensen and Kong 1997). This is because a local dependence resulting from inheritance constraints can be so strong that the sampler in practice will be reducible during the available time if single-site updating dynamics are used (see Sheehan and Thomas 1993; Lin *et al.* 1994; Lin 1995; Jensen and Sheehan 1998). Even biallelic loci can be practically

reducible in some designs; see Janss *et al.* (1995). Single outbred family (F_2 design) with many offspring is an extreme example of this kind of strong dependence structure. Therefore, haplotypes for the entire family are updated as one block (Step 2 below) at each marker. (In some cases, due to the dependency between adjacent loci, good mixing properties of the sampler may be difficult to achieve, even when block-updating is applied within one locus.)

In the following, we describe only those parts of the estimation algorithm that are different from those in Sillanpää and Arjas (1998; see also the graphical representation of the model therein):

Step 2. The following is repeated for each marker, $j = 1, \ldots, N$: A new ordered genotype proposal (family-block) at the jth position is constructed as follows:

1. If one or both genotypes in parents are unknown, a consistent pair of genotypes is proposed. Each consistent genotype-pair is considered as equally likely.
2. If unknown, their allelic origins are also proposed considering each configuration as equally likely.
3. Incomplete offspring genotypes are completed by taking one allele (with equal transmission probabilities) from each parent. These transmissions simultaneously specify the allelic origins and the grandparental origins, which are then updated accordingly.
4. Unknown allelic origins of known offspring genotypes are determined by using deduction. Origins of a homozygote can be assigned randomly, and an offspring allele not found in one parent must originate from the other parent. If some origins are left uncertain, they are proposed with equal probabilities.
5. Grandparental origins are determined for offspring alleles having a heterozygous parent, but are randomly assigned for alleles inherited from homozygotes.

The family-block proposal $H_{(j)}^{*\text{new}}$ is accepted, separately for each marker j, with probability

$$\min\{1,\ p(\chi = \chi^{(t)} | H^{*(t,\text{new}(j))},\ l^{(t)},\ m,\ N_{\text{qtl}}^{(t)}) \times \prod_{i=1}^{N_{\text{ind}}} f_{j,i}(H^{*(t,j-1)},\ H^{*(t,\text{new}(j))}) \,/\, [p(\chi = \chi^{(t)} | H^{*(t,j-1)},\ l^{(t)},\ m,\ N_{\text{qtl}}^{(t)}) \times \prod_{i=1}^{N_{\text{ind}}} f_{j,i}(H^{*(t,j-1)},\ H^{*(t,j-1)})]\}.$$

If the proposals for marker j are accepted, then $H_{(j)}^{*(t)} = H_{(j)}^{*\text{new}}$, and otherwise $H_{(j)}^{*(t)} = H_{(j)}^{*(t-1)}$. Here the notation $H_{(j)}^{*(t)}$ refers to the family-block haplotype in the jth marker in the tth round, while vector $H^{*(t,\text{new}(j))} = (H_{(1)}^{*(t)}, \ldots, H_{(j-1)}^{*(t)}\ H_{(j)}^{*\text{new}}, H_{(j+1)}^{*(t-1)}, \ldots, H_{(N)}^{*(t-1)})$, vector $H^{*(t,j)} = (H_{(1)}^{*(t)}, \ldots, H_{(j)}^{*(t)}\ H_{(j+1)}^{*(t-1)}, \ldots, H_{(N)}^{*(t-1)})$, and function $f_{j,i}(H_1^*, H_2^*) = \{p(\mathcal{G}_{j+1,i}^{\text{F}}(H_1^*) | \mathcal{G}_{j,i}^{\text{F}}(H_2^*)) \times p(\mathcal{G}_{j,i}^{\text{F}}(H_2^*) | \mathcal{G}_{j-1,i}^{\text{F}}(H_1^*)) \times p(\mathcal{G}_{j+1,i}^{\text{M}}(H_1^*) | \mathcal{G}_{j,i}^{\text{M}}(H_2^*)) \times p(\mathcal{G}_{j,i}^{\text{M}}(H_2^*) | \mathcal{G}_{j-1,i}^{\text{M}}(H_1^*))\}$.

Step 3. Random walk proposals for regression parameters are generated in three different blocks: (1) mean, environmental covariates, and residual standard deviation; (2) all QTL genotypic coefficients; and (3) all background control coefficients. Denote by L_1 (L_2) the likelihood and by p_1 (p_2) the normal density prior for the QTL genotypic coefficients evaluated at the new (old) values. The proposals are accepted separately for each block with probability $\min\{1, L_1 \times p_1 / (L_2 \times p_2)\}$. If accepted, then $\delta^{(t)} = \delta^{\text{new}}$, and otherwise $\delta^{(t)} = \delta^{(t-1)}$. (In block 3, the acceptance ratio is evaluated separately for each background control.)

Step 4. Imputation for the missing background control markers is done as in Sillanpää and Arjas (1998) except for the following: A consistent genotype pair is first proposed for the parents. Then all offspring with a missing genotype in the corresponding background control position are completed by sampling alleles according to Mendelian transmission probabilities.

APPENDIX B

As in Sillanpää and Arjas (1998), we divide the chromosome into bins $\Delta_1, \Delta_2, \ldots, \Delta_{N\text{bins}}$, where $\hat{\lambda}_j$ is the approximate posterior QTL intensity on interval Δ_j, obtained from the Monte Carlo simulation of N_{cycs} iteration cycles. In a backcross or an F_2 intercross, let

$$\widehat{D_j^x}(d) = \frac{\sum_{k=1}^{N\text{cycs}} \sum_{q=1}^{N_{\text{qtl}}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j,\ b_{qx}^{(k)} - \mu_q^{(k)} \le d\}}}{\sum_{k=1}^{N\text{cycs}} \sum_{q=1}^{N_{\text{qtl}}^{(k)}} 1_{\{l_q^{(k)} \in \Delta j\}}} \tag{7}$$

be the empirical estimator of c.d.f. $D_j^x(d)$ associated with the phenotypic effect of heterozygous QTL genotype x at a putative QTL in bin Δ_j and $\mu_q^{(k)} = \sum_{x=1}^{N\text{gen}} b_{qx}^{(k)} / N_{\text{gen}}$. If fixation of QTL alleles in different grandparental lines is assumed, we can use distribution functions similar to those presented for F_2 in Sillanpää and Arjas (1998).