Fine Mapping of Complex Trait Genes Combining Pedigree and Linkage Disequilibrium Information: A Bayesian Unified Framework

Miguel Pérez-Enciso[1]

Institut National de la Recherche Agronomique, Station d'Amélioration Génétique des Animaux, 31326 Castanet-Tolosan, France

Manuscript received September 4, 2002
Accepted for publication December 19, 2002

ABSTRACT

We present a Bayesian method that combines linkage and linkage disequilibrium (LDL) information for quantitative trait locus (QTL) mapping. This method uses jointly all marker information (haplotypes) and all available pedigree information; *i.e.*, it is not restricted to any specific experimental design and it is not required that phases are known. Infinitesimal genetic effects or environmental noise ("fixed") effects can equally be fitted. A diallelic QTL is assumed and both additive and dominant effects can be estimated. We have implemented a combined Gibbs/Metropolis-Hastings sampling to obtain the marginal posterior distributions of the parameters of interest. We have also implemented a Bayesian variant of usual disequilibrium measures like D' and r^2 between QTL and markers. We illustrate the method with simulated data in "simple" (two-generation full-sib families) and "complex" (four-generation) pedigrees. We compared the estimates with and without using linkage disequilibrium information. In general, using LDL resulted in estimates of QTL position that were much better than linkage-only estimates when there was complete disequilibrium between the mutant QTL allele and the marker. This advantage, however, decreased when the association was only partial. In all cases, additive and dominant effects were estimated accurately either with or without disequilibrium information.

AN ultimate goal of quantitative trait loci (QTL) studies is to clone the gene(s) responsible for the genetic differences between individuals and, eventually, identify the causal mutation(s). Certainly, this is a daunting task that will be accomplished only gradually. One of the most severe limitations, at the moment, is that the QTL position is estimated with too large an error to allow positional cloning when a classical linkage analysis is employed. The 95% confidence interval for the QTL position usually spans over 5–20 cM, at a minimum. The wide confidence interval occurs because the number of meioses in the genotyped pedigree is usually very small; only between two and three generations are generally employed. Linkage disequilibrium (LD)-based methods, in contrast, capitalize on the number of generations that occurred since the appearance of mutation and can produce extremely accurate estimates of the gene position, within kilobases in some instances (Hastbacka *et al.* 1994). Nevertheless, the chance of success of the LD strategy depends on a number of population parameters, such as the degree of admixture in the sampled population, the actual level of association between the causal mutation and the polymorphisms, or the correct ascertainment of phases and of genotypes at the QTL. Of course these parameters are usually unknown but do dramatically affect the results (Terwilliger and Weiss 1998). In fact, a pure LD analysis is likely to result in a large number of false positives as illustrated recently, *e.g.*, in Alzheimer's disease (Emahazion *et al.* 2001).

A promising approach is thus to combine both linkage and linkage disequilibrium (LDL) methods to add their advantages in a single unified theoretical framework. More specifically, there is an urgent need for robust methods that provide accurate estimation of the QTL position. Consider for the sake of illustration a simple design where a number of nuclear families are typed, *i.e.*, parents and offspring. The theoretical advantages of combining linkage disequilibrium and pedigree (linkage) information in QTL analysis are manifold: (i) A marker for which a parent is homozygous does not contribute information in a linkage analysis, yet it does in LD analysis; (ii) conversely, two parents may share the same haplotype but not necessarily the same QTL genotypes, and a pure LD analysis would be misleading but the phenotype of offspring together with the ascertainment of alleles transmitted can be used to determine which are the most likely QTL genotypes of the parents; (iii) an individual without relatives but with phenotype records can be included in the LD analysis, in contrast to a pure linkage study; and (iv) a comparison of the analyses including or not the LD information can assess the validity of the LD model assumptions (*i.e.*, one mutation t generations ago).

Several authors have addressed the problem of combining LD and linkage mapping for quantitative trait loci (Zhao *et al.* 1998; Allison *et al.* 1999; Almasy *et*

[1]*Address for correspondence:* INRA, Station d'Amélioration Génétique des Animaux, BP 27, Cedex 31326 Castanet-Tolosan, France. E-mail: mperez@toulouse.inra.fr

al. 1999; Fulker *et al.* 1999; Wu and Zeng 2001; Farnir *et al.* 2002; Meuwissen *et al.* 2002), whereas Xiong and Jin (2000) proposed a method suited to disease susceptibility genes. Zhao *et al.* (1998) developed a semiparametric procedure based on the score-estimating equation approach and that addressed the particular case of single-nucleotide polymorphisms. This is one of the first articles to provide a theoretical framework for LDL mapping but the estimating equation approaches are difficult to implement in practice; they require complex computations adapted to each family structure. For instance, the method sums over all possible phases and computes their probabilities, which is extremely complex to do in practice beyond a few markers. The statistical properties of these estimators are also unknown.

Fulker *et al.* (1999) developed a sib-pair analysis in a likelihood framework. The approach followed by Allison *et al.* (1999) is a generalization of the transmission disequilibrium test (TDT) for quantitative traits (Allison 1997), where a between- and within-family association parameter is modeled via a mixed model. Neither the Fulker *et al.* (1999) nor Allison *et al.* (1999) methods are very suited to analyzing complex pedigrees as they consider sib pairs (Fulker *et al.* 1999) or parent-offspring trios (Allison *et al.* 1999) and their theoretical framework is difficult to generalize to more complex settings. TDT in particular is not an optimum choice to deal with very polymorphic markers like microsatellites and makes use of only a limited amount of the total information contained in a typical pedigree. Meuwissen *et al.* (2002), in turn, proposed to model the QTL alleles as a random variable, where the covariance between base population haplotypes allows the inclusion of the LD information (Meuwissen and Goddard 2000), and the covariance between non-base population haplotypes was computed as in Fernando and Grossman (1989) and Goddard (1992). They estimated the position via maximum likelihood. The model followed by these authors is different from the usual LD, where a diallelic QTL is assumed. The key issue in their method is to compute the identity-by-descent probabilities between the base population haplotypes, and this was done by considering the number of identity-by-state alleles shared by any two haplotypes, along the lines also suggested by McPeek and Strahs (1999). They assumed that phases are known, which is a reasonable assumption only if families are very large, *e.g.*, as in dairy cattle. Otherwise, QTL positioning can be dramatically affected if a phase is incorrectly specified. Farnir *et al.* (2002) developed an analytical approach for combining linkage and LD in half-sib families, where the disequilibrium information is incorporated via Terwilliger's (1995) approach. Their method would be very cumbersome to generalize to more complex populations; in addition, phases are assumed to be known and it is not a true multipoint method. The method of Wu and Zeng (2001), intended for natural populations, is also difficult to apply to complex pedigrees.

Here we present a Bayesian method that combines linkage and LD information for QTL mapping within a unified theoretical framework. Our LDL method uses jointly all marker information, as well as all available pedigree information; *i.e.*, it is not restricted to any specific experimental design and it is not required that phases be known. If desired, infinitesimal genetic effects or environmental noise (fixed) effects can also be fitted. A diallelic QTL is assumed and both additive and dominant effects can be estimated. We have implemented a combined Gibbs/Metropolis-Hastings sampling to obtain the marginal posterior distributions of the parameters of interest. We illustrate the method with simulated data.

THEORY

We assume that the goal of the analysis is to fine map a QTL that has been previously located within a given genome region. The genetic model presupposes that a single mutation occurred t generations ago on a gene affecting the trait studied. Thus, initially, a single ancestral (founder) haplotype harbored the mutation. The number of haplotypes carrying the mutation increases in successive generations provided that the mutation is not lost and, due to recombination, the initial allele combination is eroded. The amount of disequilibrium between markers and QTL decreases proportionally to genetic distance and to the number of generations elapsed since mutation. Here we use the population model for linkage disequilibrium decay described in Morris *et al.* (2000), with modifications described below. Briefly, a binary variable S_{ki} is defined such that, at any kth marker locus and ith individual, the locus will be either identical by descent (IBD) with the original haplotype carrying the mutation ($S_{ki} = -$) or not ($S_{ki} = +$), with minus and plus signs standing for the mutant and wild haplotype alleles, respectively. By convention we denote the QTL by locus 0. A Markov chain Monte Carlo (MCMC) method was provided by Morris *et al.* (2000) to obtain the transition probabilities of a locus being IBD or not at locus $k + 1$ conditional on being IBD or not at locus k.

Now suppose that the QTL additive and dominance effects are a and d, respectively; *i.e.*, the mean phenotype of the individuals homozygous for the wild allele (+/+) minus that of individuals homozygous for the mutant allele (−/−) is $2a$, whereas the mean phenotype of heterozygous individuals, (+/−) or (−/+), is d. Suppose further that a number m of individuals have been typed for DNA markers, contained in matrix **M**, and that phenotypic measurements (**y**) are available on a subset of n individuals. The linear explicative model is

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{w}_a a + \mathbf{w}_d d + \mathbf{Zu} + \mathbf{e} = \mathbf{X^*}\,\boldsymbol{\beta}^* + \mathbf{Zu} + \mathbf{e}, \quad (1)$$

TABLE 1

Main symbols used

n	Number of phenotypic records
m	Number of individuals in the pedigree
$\mathbf{y}$	Phenotypic records, dimension n
$\mathbf{M}$	Marker information, contains the alleles for each individual and marker; dimension $m \times$ no. of markers $\times$ 2
$\mathbf{S_0}$	Identity-by-descent status of the QTL allele of the base generation individuals with the causative mutation; it can take values *wild* (+) or *mutant* (−) allele, dimension 2 × no. of base generation individuals
a	Additive QTL effect; the average value of individuals with genotype (+/+) − (−/−) is $2a$
d	Dominance effect; phenotypic value of individuals with genotype (+/−) or (−/+)
$\mathbf{u}$	Infinitesimal genetic value; it contains all genetic effects except the QTL under study, dimension m
$\boldsymbol{\beta}$	Fixed (noise environmental) effects, dimension the sum of levels for each fixed effect
σ^2_u	Infinitesimal genetic variance
σ^2_e	Residual variance
δ	QTL position, in morgans
t	Time (no. of generations) since mutation
$\mathbf{T}$	$2 \times m$ matrix with QTL segregation indicators. The genotype of all individuals is unambiguously determined by $\mathbf{T}$ and $\mathbf{S}_0$
$\mathbf{H}$	Marker phases; contains indicator variable to identify whether the allele in vector $\mathbf{M}$ is of paternal or maternal origin; dimension $m \times$ no. of markers

where $\boldsymbol{\beta}$ is a fixed-effects (environmental/nongenetic effects) vector; $\mathbf{w_a}$ is a vector with indicator variables taking values 1 or −1 if the QTL genotype of each individual is +/+ or −/−, respectively, and zero for heterozygous individuals; $\mathbf{w_d}$ contains values 1 if individual QTL genotype is +/− or −/+, zero otherwise; and $\mathbf{u}$ and $\mathbf{e}$ contain the infinitesimal genetic values (polygenic effects) and residuals, respectively, whereas $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices. The matrix $\mathbf{X^*}$ contains $\mathbf{X}$ plus two additional columns for $\mathbf{w_a}$ and $\mathbf{w_d}$; similarly vector $\boldsymbol{\beta}^*$ is $\boldsymbol{\beta}$ plus elements a and d.

The goal of the analysis is to obtain estimates of the set of parameters, $\boldsymbol{\theta} = \{\mathbf{S_0}, a, d, \mathbf{u}, \boldsymbol{\beta}, \sigma^2_u, \sigma^2_e, \delta, t, \mathbf{T}, \mathbf{H}\}$, where $\mathbf{S_0}$ is a matrix containing the IBD status of the two individual QTL alleles with the causal mutation, taking values + or −; σ^2_u is the infinitesimal genetic variance; σ^2_e, the residual variance; and δ is the QTL position. $\mathbf{T}$ is a QTL segregation indicator vector containing, for each individual and haplotype, a binary variable specifying whether the QTL allele is IBD with the paternal or maternal parental allele (Thompson 1994). Note that $\mathbf{S}_0$ needs to be specified only for the base population individuals (those without known parents) and that the QTL genotypes for the whole population are unambiguously determined once $\mathbf{S}_0$ and $\mathbf{T}$ are specified. Finally, $\mathbf{H}$ is a vector containing the phases (paternal or maternal) for each of the markers. It can be seen that $\mathbf{w_a}$ and $\mathbf{w_d}$ in (1) are completely determined by $\mathbf{S_0}$ and $\mathbf{T}$ and are not additional random variables; a redundant notation was used solely for the sake of clarity in (1). The main symbols that are used throughout the article are detailed in Table 1 for the reader's convenience.

The Bayesian inference is based upon the posterior distribution of the parameters,

$$p(\boldsymbol{\theta}|\mathbf{y}, \mathbf{M}) \propto p(\mathbf{y}, \mathbf{M}|\boldsymbol{\theta})p(\boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{M}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \quad (2)$$

where $p(\mathbf{y}, \mathbf{M}|\boldsymbol{\theta})$ is the likelihood (in the Bayesian sense), and $p(\boldsymbol{\theta})$ is the *a priori* distribution for the parameters. Note that phenotypes and markers are conditionally independent. Ideally, inferences about each of the parameters in $\boldsymbol{\theta}$, say θ_l, should be based on the marginal posterior distribution, *i.e.*,

$$p(\theta_l|\mathbf{y}, \mathbf{M}) = \int_{\theta_{-l}} p(\theta_l, \boldsymbol{\theta}_{-l}|\mathbf{y}, \mathbf{M})\partial\boldsymbol{\theta}_{-l}, \quad (3)$$

where $\boldsymbol{\theta}_{-l}$ indicates the vector of parameters except the lth unknown. Typically this multidimensional integral is unfeasible and we need to resort to stochastic procedures like Gibbs or Metropolis-Hastings sampling schemes (Sorensen and Gianola 2002). In the following, we describe all conditional distributions that we need to sample from. Unless otherwise stated, we make the usual assumptions of flat priors for all parameters, except for $p(\mathbf{u}) = \text{Normal}(\mathbf{0}, \mathbf{A}\sigma^2_u)$, where $\mathbf{A}$ is the additive relationship matrix between individuals (Lynch and Walsh 1998).

The rest of this section is devoted to presenting the main conditional distributions to sample from to obtain the posterior distribution of the parameters of interest. For the reader less interested in the mathematical details, this part can be summarized as follows. For the base population individuals (those without ancestors genotyped) we use their marker haplotypes and the phenotypic information of their descendants, in addition to the prior allele frequencies, to ascertain the more likely QTL genotypes. The LD signal is incorporated into the model via the distribution $p(\mathbf{M}|\boldsymbol{\theta})$, which quantifies the probability of an individual carrying a certain marker haplotype conditional on its QTL genotype and other

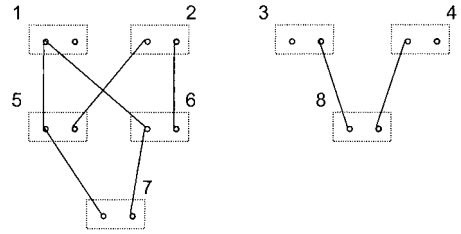

FIGURE 1.—Representation of a pedigree via the transmission coefficients **T**. Each small circle represents an allele of the QTL, identical-by-descent alleles are connected with a solid line, and individual genotypes, 1–8, are boxed with dashed lines.

population parameters, like the age of the mutation. We assume a star-shaped genealogy. We suppose that base population individuals are genotyped for most of the markers but not that phases are known; they are inferred from the offspring genotypes. LD or allele frequency priors do not contribute any information to obtain the genotypes of the descendant individuals (conditionally on the genotypes of the base population) and are sampled following the most likely recombinants as inferred from marker information. Once the QTL alleles are sampled, most of the remaining parameters are obtained via a classical Gibbs sampling within the mixed-model context (SORENSEN and GIANOLA 2002). In contrast, Metropolis-Hastings is required for the QTL position; here we identify where recombinants have occurred at two alternative positions and the resulting likelihoods using available phenotypic information are compared (UIMARI and SILLANPÄÄ 2001).

Base population QTL genotypes (S_0): In the absence of LD information, only the phenotypes of the individuals that have received a given base population allele provide information about the likely value of that allele. This is illustrated in the simple pedigree of Figure 1; the solid lines represent the transmitted alleles, stored in **T**. Suppose that we are sampling the IBD status of first individual and first allele (S_{011}), conditional on all other parameters including the $\mathbf{S}_0$ of the remaining individuals (denoted by $\boldsymbol{\theta}\_$). The phenotypes of individuals 1, 5, 6, and 7 influence the probability $p(S_{011}|\boldsymbol{\theta}\_, \mathbf{y}, \mathbf{M})$. In contrast, $p(S_{012}|\boldsymbol{\theta}\_, \mathbf{y}, \mathbf{M})$, corresponding to the second QTL allele, involves only the phenotype of individual 1, as this allele was not transmitted. If that individual does not have phenotype recorded, $p(S_{012}|\boldsymbol{\theta}\_, \mathbf{y}, \mathbf{M})$ is strictly proportional to the prior frequencies for each QTL allele, when LD information is not being used. We denote by ψ_i the set of individuals that have received at least one allele for individual i and have phenotypes, *i.e.*, $\psi_1 = \{1, 5, 6, 7\}$, $\psi_2 = \{5, 6\}$, and $\psi_3 = \psi_4 = \{3, 4, 8\}$. Note that the set ψ may vary from iteration to iteration as a new **T** is sampled. If LD information is being used, $p(S_0|\boldsymbol{\theta}\_, \mathbf{y}, \mathbf{M})$ also depends on the marker alleles of the base population individuals. Using all sources of information, the QTL IBD status of the ith base population individual can be sampled from the fully conditional distribution,

$$\begin{aligned} p(S_{0i1}, S_{0i2}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}\_) &\propto p(\mathbf{y}|\boldsymbol{\theta})p(\mathbf{M}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &= \left[\prod_{j\in\Psi_i} p(y_j|S_{0i1}, S_{0i2}, \mathbf{S}_{0-}, a, d, u_i, \boldsymbol{\beta}, \sigma_e^2, \mathbf{T})\right] \\ &\quad \times p(\mathbf{M}_i|S_{0i1}, S_{0i2}, t, \mathbf{H}_i, \delta) \times p(S_{0i1}, S_{0i2}) \\ &= p(\mathbf{y}_{j\in\Psi_i}|\boldsymbol{\theta}) \times p(\mathbf{M}_i|\boldsymbol{\theta}) \times p(S_{0i1}, S_{0i2}), \end{aligned} \quad (4)$$

where y_j is the phenotype of the jth individual having received at least one allele from individual i, and $\mathbf{S}_{0-}$ denotes the rest of IBD status not sampled. We now show which are the distributions involved in (4). The first term is a product of Normal densities $N(e_j, \sigma_e^2)$, with

$$e_j = y_j - \mathbf{x}_j'\boldsymbol{\beta} - u_j - w_{aj}a - w_{dj}d,$$

where, $\mathbf{x}_j'$ is the column vector of $\mathbf{X}$ corresponding to the jth individual's observation.

The distribution $p(\mathbf{M}_i|\boldsymbol{\theta})$ in (4) is the probability of having marker alleles linked in haplotype 1 or 2 (say $\mathbf{M}_{i1}$ or $\mathbf{M}_{i2}$) conditional on a given QTL genotype, its position relative to DNA markers, and the parameter governing the LD decay (t). Both haplotypes are conditionally independent; thus $p(\mathbf{M}_i|\boldsymbol{\theta}) = p(\mathbf{M}_{i1}, \mathrm{M}_{i2}|S_{0i1}, S_{0i2}, t, \mathbf{H}_i, \delta) = p(\mathbf{M}_{i1}|S_{0i1}, t, \mathbf{H}_i, \delta)p(\mathbf{M}_{i2}|S_{0i2}, t, \mathbf{H}_i, \delta)$, where $\mathbf{M}_{i1}$ contains the marker alleles received from the father and $\mathbf{M}_{i2}$, those of mother's origin. Consider the marker alleles of a given individual i at haplotype h ($\mathbf{M}_{ih}$); in our notation L markers are to the left and R markers to the right of the current QTL position. Then,

$$\begin{aligned} &p(\mathbf{M}_{ih}|S_{0ih}, t, \mathbf{H}_i, \delta) \\ &= p(M_{ih,-L}, \ldots, M_{ih,-2}, M_{ih,-1}, M_{ih1}, M_{ih2}, \ldots, M_{ihR}|S_{0ih}, t, \mathbf{H}_i, \delta) \\ &= p(M_{ih,-L}, \ldots, M_{ih,-2}, M_{ih,-1}|S_{0ih}, t, \mathbf{H}_i, \delta) \\ &\quad \times p(M_{ih1}, M_{ih2}, \ldots, M_{ihR}|S_{0ih}, t, \mathbf{H}_i, \delta) \;=\; Q_{ihL}Q_{ihR}, \end{aligned}$$

where M_{ihk} denotes the allele at marker k (starting from the QTL) of haplotype h, ith individual. Note that k takes negative values for markers to the left of the QTL. Dropping subscripts i and h and the conditioning on t, **H,** and on δ for clarity, we find

$$\begin{aligned} Q_R &= p(M_1, M_2, \ldots, M_R|S_0) \\ &= \sum_{S_1} p(M_2, \ldots, M_R|S_1)p(M_1|S_1)p(S_1|S_0). \end{aligned}$$

This process is repeated sequentially from the QTL position toward the extremes of the interval,

$$\begin{aligned} Q_R &= \sum_{S_1}\sum_{S_2} p(M_3, \ldots, M_R|S_2)p(M_2|S_2)p(S_2|S_1)p(M_1|S_1)p(S_1|S_0) \\ &= \sum_{S_1}\sum_{S_2} \cdots \sum_{S_R} \prod_{k=1}^{R} p(M_k|S_k)p(S_k|S_{k-1}), \end{aligned} \quad (5)$$

where S_k is the IBD state of marker allele k of individual i with the original mutant haplotype.

At any marker locus, k, the locus will be either IBD with

the original haplotype carrying the mutation ($S_k = -$) or not ($S_k = +$). The term $p(M_k|S_k)$ contains the marker allele probabilities conditional on S_k; $p(M_k|S_k = +)$ is simply given by the population allele frequencies. In contrast, $p(M_k|S_k = -)$ will be 1 for the allele that carried the mutant haplotype and 0 for the remaining alleles. The vector $p(\mathbf{M}|S_L = \ldots S_{-1} = S_1 = S_R = -)$ is the original haplotype that carried the mutation. Of course this haplotype is unknown but can be inferred as shown by Morris *et al.* (2000). Here we have preferred to consider both S_k and $p(M_k|S_k)$ as nuisance parameters; *i.e.*, we are not usually interested directly in them, and thus we integrate them out in (5). As a result, $p(M_k|S_k = -)$ is no longer 0's and 1's but can take any value between the two extremes. The appendix shows how $p(M_k|S_k)$ is updated.

The transition probabilities $p(S_k|S_{k-1})$ can be obtained as detailed in Morris *et al.* (2000) and depend on the effective size and time since mutation. Four transition probabilities need to be specified, which are

$$p(S_k = -|S_{k-1} = -) = \exp(-\phi t\delta_{k,k+1}) + [1 - \exp(-\phi t\delta_{k,k+1})]\alpha,$$

$$p(S_k = +|S_{k-1} = -) = [1 - \exp(-\phi t\delta_{k,k+1})](1 - \alpha),$$

$$p(S_k = -|S_{k-1} = +) = [1 - \exp(-\phi t\delta_{k,k+1})]\alpha,$$

and

$$p(S_k = +|S_{k-1} = +) = \exp(-\phi t\delta_{k,k+1}) + [1 - \exp(-\phi t\delta_{k,k+1})](1 - \alpha)$$

(Morris *et al.* 2000), where ϕ is the ratio of 1 M/1 Mb DNA (typically 1/100), $\delta_{k,k+1}$ is the distance (morgans) between loci k and $k + 1$, and α is the probability of recombining with a haplotype carrying the mutation. This parameter is in fact highly confounded with t (Kaplan *et al.* 1995) and we did not try to estimate it; rather, we set $\alpha = 0.001$. This had a negligible impact on the results.

Expression (5) is extremely difficult to compute. However, we can rearrange as

$$Q_R = \sum_{S_1} p(M_1|S_1)\,p(S_1|S_0) \ldots \sum_{S_{R-1}} p(M_{R-1}|S_{R-1})\,p(S_{R-1}|S_{R-2}) \times \sum_{S_R} p(M_R|S_R)\,p(S_R|S_{R-1}).$$

Thus, starting from the outermost marker, R, it is feasible to compute Q_R using the recursive formula

$$q_k = \sum_{S_k} p(M_k|S_k)\,p(S_k|S_{k-1})\,q_{k+1}$$

with initial values $q_{-L} = q_R = 1$; $Q_R = \Sigma_{k=R}^{1}\, q_k$, and similarly $Q_L = \Sigma_{k=-L}^{1}\, q_k$. Note that each coefficient q_k is a vector with two elements corresponding to states $S_k = +$ and $S_k = -$. At the end of the computations we obtain the probabilities of individual haplotypes given $S_0 = +$ and $S_0 = -$. There can be numerical problems in obtaining Q_R or Q_L for a large number of markers as the number of possible haplotypes increases exponentially with the number of markers, especially for highly polymorphic markers like microsatellites. However, since the relevant statistic is the ratio $Q(S_0 = +)/Q(S_0 = -)$, q_R and q_L can be initialized to a very large number.

Finally, $p(S_{0i1}, S_{0i2})$ in Equation 4 is the *a priori* probability of the IBD state of the two QTL alleles with the original mutant haplotype. When the individual is not inbred, $p(S_{0i1}, S_{0i2}) = p(S_{0i1})p(S_{0i2})$, where the prior probabilities are the same for any base population allele. If the ith individual is known to be inbred from the available pedigree with inbreeding coefficient f_i, $p(S_{0i1}, S_{0i2}) = (1 - f_i)p(S_{0i1})p(S_{0i2}) + f_i p(S_{0i2})\eta(S_{0i1}|S_{0i2})$, with η being an indicator 1/0 function that makes S_{0i1} take the same value as S_{0i2}. The prior probability of an allele being identical by descent with the original mutant haplotype is α if the base population individuals have been sampled at random from the population, *i.e.*, $p(S_{0i} = -) = \alpha$ and $p(S_{0i} = +) = 1 - \alpha$ for every individual. Otherwise, *e.g.*, case/control study or selective genotyping, the probabilities have to be modified accordingly (Morris *et al.* 2000).

In summary, to sample the IBD states at the QTL position we evaluate Equation 4 at all four possible QTL genotypes, *i.e.*, (+/+), (+/−), (−/+), (−/−), for each base population individual in turn, and we take a random number according to the genotype probabilities. Both alleles are thus sampled simultaneously. Nevertheless, this strategy can be ameliorated by sampling larger blocks of base population IBD states. Suppose IBD states of base population individuals 1 through c are sampled; then

$$p(S_{0i1}, S_{0i2}, \ldots, S_{0c2}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}_-) \propto \prod_{j\in\Psi} p(y_j|S_0, a, d, u_i, \boldsymbol{\beta}, \sigma_e^2, \mathbf{T}) \times \prod_{i=1}^{c} p(\mathbf{M}_i|S_{0i1}, S_{0i2}, t, \mathbf{H}_i, \delta) \times \prod_{i=1}^{c} p(S_{0i1}, S_{0i2}), \quad (6a)$$

where $j \in \psi$ means any individual having received at least one allele from any of individuals 1 through c. An issue of interest is to determine which $\mathbf{S}_0$ elements are to be sampled together to minimize the risk of reducibility. Here we sampled jointly those origins that coincided in the maximum number of individuals. For instance, if only four origins were to be sampled together in the pedigree of Figure 1, two blocks with the IBD status of individuals (1, 2) and (3, 4) rather than (1, 3) and (2, 4) would be chosen. Note that a pure linkage approach can be easily implemented sampling from

$$p(S_{0i1}, S_{0i2}, \ldots, S_{0c2}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}_-) \propto \prod_{j\in\Psi} p(y_k|\mathbf{S}_0, a, d, u_i, \boldsymbol{\beta}, \sigma_e^2, \mathbf{T}) \times \prod_{i=1}^{c} p(S_{0i1}, s_{0i2}) \quad (6b)$$

instead of from (6a).

The rest of the sampling distributions required are

detailed in the APPENDIX. Once all variables are initialized, the Markov chain Monte Carlo (MCMC) chain consists of iterating successively via Equations 4 or 6 and A2–A7a plus updating the phases (**H**), $p(\mathbf{M}|\mathbf{S})$, and the transmission indicators (**T**). Obviously, in a linkage-only approach,(6b), the sampling is simplified by not sampling $p(\mathbf{M}|\mathbf{S})$ and time since mutation (t). The procedure is otherwise identical.

Two-marker disequilibrium measures: LD measurements like D' (HEDRICK 1987; LEWONTIN 1988) rely on the possibility of ascertaining the linkage phases and the alleles themselves, which is not possible with quantitative traits because the QTL genotypes are not known. Nevertheless, phases and QTL alleles are generated each iteration so we can define a Bayesian estimate of D' between any marker and the QTL, computing D' at the current configuration using the formula $D' = \sum_{i=1}^{n_1} \sum_{j=1}^{2} p_i q_j |D'_{ij}|$, where i is the ith allele of the marker, with frequency p_i, the marker has n_1 alleles, index j refers to the jth QTL allele, with frequency q_j, and $D'_{ij} = D_{ij}/D_{\mathrm{MAX}}$ is the usual measure for diallelic markers. Here we provided the mean of the posterior distribution, obtained as D' averaged over iterations. We also computed the recommended measure by PRITCHARD and PRZEWORSKI (2001) denoted by r^2 (or Δ^2 in DEVLIN and RISCH 1995), which is defined as $r^2 = \sum_{i=1}^{n_1} \sum_{j=1}^{2} D_{ij}^2 / p_i q_j$. One of the interesting properties of r^2 is that r^2 times the number of haplotypes is distributed as a chi square with $n_1 - 1$ d.f. (WEIR 1996), although this is an approximation and does not hold for large r (HUDSON 1985). Nevertheless that property is not needed here as we are able to derive the full posterior distribution of r^2 between any marker and the QTL and assess the relevant highest density region that covers the point 0 (no disequilibrium). Here we report that $r = \sqrt{r^2}$ to make it comparable with D'. Both D' and r were calculated using only the base population individuals.

SIMULATION

Two population types that can typically be found in livestock, with "simple" and "complex" pedigrees, were simulated. The simple population consisted of 40 unrelated full-sib families, 10 offspring per family. The complex population was a four-generation pedigree, with a base population of 80 unrelated parents that produced 40 full-sib families of size 5 (generation 2), whereas generations 3 and 4 consisted of 20 full-sib families (5 offspring per family). Parents were chosen at random except in generation 1, where all parents had an equal number of offspring. Both simple and complex pedigrees had a total of 480 individuals. The explored region spanned 25 cM and contained six microsatellites at positions 0, 5, 10, 15, 20, and 25 cM, together with 10 single-nucleotide polymorphisms (SNPs) located at positions 11, 12, 13, 14, 16, 17, 18, 19, 21, and 22 cM. SNP allele frequencies were 0.3 and 0.7, whereas there were six alleles at equal frequencies for each microsatellite. The QTL was located in position 18 cM, its additive effect was $a = 1$, there was no dominance ($d = 0$), and the residual variance was $\sigma_e^2 = 1$. Phenotypic records were simulated for generation 2 in the simple population and for all individuals in the complex pedigree. All individuals were genotyped. The mutant QTL allele frequency in the population studied was 0.3. Two situations were considered: The mutant QTL allele was either completely associated with SNP allele "2" (frequency = 0.3) in position 18 cM or partially associated with the SNP allele "1" (frequency = 0.7). In the former case, all haplotypes with the SNP allele 2 in position 18 cM carried the mutant QTL allele; in the latter case, initially ~42% (0.3/0.7) of haplotypes with SNP allele 1 harbored the QTL mutant allele. The original haplotype carrying the mutation was 1111111111211111 with complete association and 1111111111111111 in the second case. It was assumed that the mutant allele appeared 100 generations ago, and the decay in disequilibrium was simulated following the model in MORRIS *et al.* (2000). We compared the results using the LDL method (Equation 6a) with those when only linkage information was used (Equation 6b).

Three replicates of each case were run, resulting in 12 analyses in total. The only fixed effect included in the analyses was the general mean. The maximum change in QTL position was set to 0.5 cM in each direction. We ran 50,000 iterations of the MCMC chain, discarding the first 4000 iterations. Eight origins were sampled jointly; thus $p(S_{0i1}, S_{0i2}, \ldots, S_{0c2}|\mathbf{y}, \mathbf{M}, \boldsymbol{\theta}_-)$ can take $2^8 = 256$ values because the QTL is assumed to be diallelic. Phases were updated in blocks of six. Each complete iteration took ~3.5 sec on an alpha workstation with processor 21164A. The computing time per iteration is highly dependent on the number of paths and phases updated simultaneously.

RESULTS

Table 2 presents the mean and SD of the marginal posterior distributions for the main parameters in the case of complete association. The posterior distributions for the additive and dominant effects in the first replicate are plotted in Figure 2a and provide a whole picture about the uncertainty regarding these parameters. Results were very similar for all replicates so only one is presented. The estimates of the genetic effects and the residual variance were quite accurate, and the SDs of their posterior distributions were small, indicating that there is enough information in the data to estimate these parameters. The 95% highest density region contained the true values of a, d, and σ_e^2 in all cases. In particular, it was correctly detected that gene action was additive. A rigorous test of dominance, nevertheless, would imply computing the Bayes factors between the

TABLE 2

Posterior distribution statistics: complete association

Pedigree[a]	Replicate	Analysis[b]	Parameters[c] a/σ_e	d/σ_e	σ_e^2	Position (M)	t
Simple	1	LDL	1.06 (0.08)	0.02 (0.10)	0.96 (0.07)	0.169 (0.031)	73 (14)
		L	1.00 (0.09)	−0.05 (0.13)	1.00 (0.07)	0.148 (0.044)	—
	2	LDL	1.07 (0.08)	0.08 (0.13)	0.99 (0.07)	0.183 (0.027)	79 (20)
		L	1.07 (0.10)	0.14 (0.15)	0.99 (0.08)	0.144 (0.062)	—
	3	LDL	1.08 (0.10)	0.08 (0.10)	0.88 (0.07)	0.180 (0.014)	90 (18)
		L	1.02 (0.11)	−0.01 (0.12)	0.92 (0.08)	0.192 (0.028)	—
Complex	1	LDL	0.94 (0.08)	−0.05 (0.09)	1.13 (0.08)	0.182 (0.024)	71 (16)
		L	0.88 (0.09)	0.01 (0.10)	1.18 (0.09)	0.197 (0.033)	—
	2	LDL	0.99 (0.08)	−0.01 (0.10)	1.03 (0.07)	0.187 (0.020)	101 (21)
		L	0.95 (0.09)	0.01 (0.12)	1.07 (0.08)	0.168 (0.042)	—
	3	LDL	0.96 (0.08)	0.05 (0.09)	1.09 (0.08)	0.169 (0.017)	141 (15)
		L	0.91 (0.09)	0.05 (0.10)	1.11 (0.08)	0.160 (0.031)	—

All haplotypes with SNP allele 2 carried the QTL mutant allele.

[a] Simple pedigree populations consist of independent full-sib families; complex population is a four-generation pedigree with random mating.

[b] LDL analysis combines both linkage disequilibrium and pedigree information; L analysis uses only linkage.

[c] Mean of the marginal posterior distribution (SD of the marginal posterior distribution).

two competing models. Interestingly, there was little difference between using or not using the linkage disequilibrium information. This means that most, if not all, information to estimate the QTL genetic effects comes from classical linkage analysis. The effect of population structure was also negligible. However, including LD does affect the estimate of the QTL position (Table 2, Figure 3) with complete association between the SNP and the QTL alleles: (1) The mode of the posterior distribution always coincided with the true position and this was not necessarily the case in the linkage-only approach; (2) LDL estimates were always less biased; and (3) the SDs of the posterior distributions were always smaller in the LDL than in the linkage-only method. In general, the relative advantage of LDL over linkage-only was larger in the two-generation than in the complex pedigrees. This can occur because more meioses are available for mapping in the four- than in the two-generation pedigree but also because in the complex pedigree there were fewer offspring per family, making it less accurate for estimating the QTL genotype and the marker phases of the base population individuals, and this has a much larger effect on LDL than in linkage-only analysis.

Results concerning the incomplete association scenario are presented in Table 3 and Figure 4. As expected, the estimates of the QTL effects were similar to those in Table 2, albeit the SDs were somewhat larger in particular for the dominance effect. Replicate 2 of the complex pedigree had unusually large SD of the posterior distributions of a and d. But more importantly, the accuracy of the QTL position was generally much smaller with incomplete than with complete association (note that the scales of the y-axes are different in Figures 3 and 4). It is also apparent that the mode of the posterior distribution coincided with the true position only

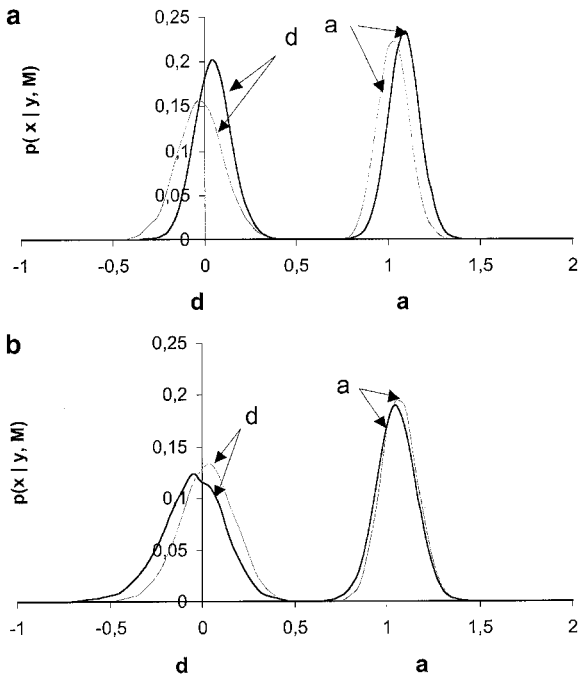

FIGURE 2.—Marginal posterior probabilities of additive (a) and dominant effects (d), expressed in residual standard deviation units. The thick line corresponds to the LDL estimate and the thin shaded line, to the linkage-only estimate. (a) First replicate of the simple pedigree, complete association; (b) first replicate of the simple pedigree, incomplete association. The true values were $a = 1$ and $d = 0$.

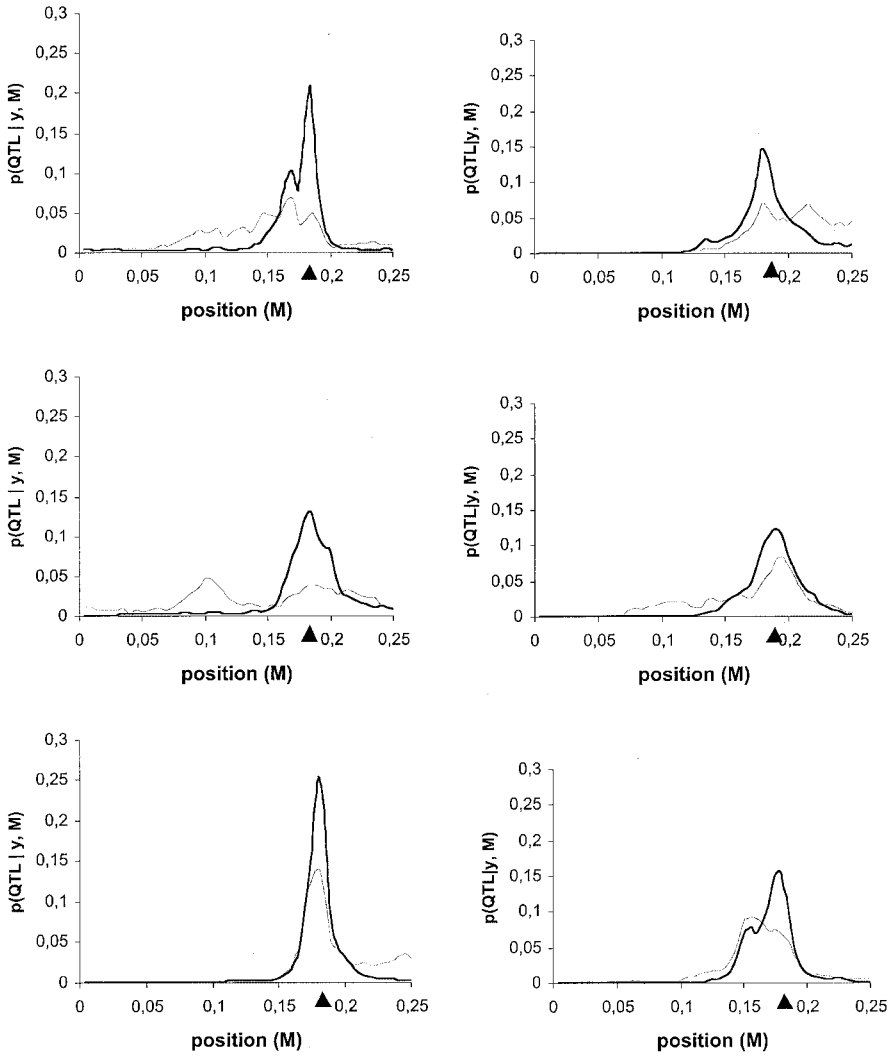

FIGURE 3.—Marginal posterior probabilities of QTL location with complete association between QTL and SNP genotype. (Left) Simple population graphs; (right) complex population graphs. The three replicates are shown below each other. The solid thick lines refer to estimates obtained using linkage and linkage disequilibrium, and the thin shaded lines refer to estimates obtained using linkage information only. The QTL was located in position 18 cM (indicated by the arrowhead).

once (replicate 1, complex pedigree) although it was close, positions 0.16–0.17 M, in the remaining replicates with the LDL approach. In some instances (replicate 1, simple pedigree) the posterior density was very flat and covered almost the whole region under study. In principle, linkage-only estimates should not be greatly affected by either complete or incomplete association, because the accuracy depends mainly on the informativity of markers to identify recombinant haplotypes. This seems to be the case if we exclude the rather outlying replicate 1 (simple pedigree, Figure 4). The average SD of the QTL position posterior density was 4 cM in the linkage-only approach for both complete and incomplete association scenarios. In contrast, it was 2.2 and 3 cM using LDL in the complete and incomplete scenarios, respectively.

Contrary to the estimates of QTL genetic effects or position, the LD decay parameter t was loosely estimated (Tables 2 and 3). This means that there is little information in the data to estimate them. In fact, we observed that $p(\mathbf{M}|\boldsymbol{\theta})$ was quite flat for different values of t. A positive reading is that the exact figures for t did not affect the final results to a large extent, as we found similar output when we fitted these parameters to a variety of values, in agreement with previous results (MEUWISSEN and GODDARD 2000).

Finally, Figure 5 draws a plot of the simple disequilibrium measures between each marker and the QTL, D' and r, for the three simple pedigrees. D' and r measures obtained under both statistical methods LDL and linkage-only are plotted. The two top and bottom plots correspond to the complete and incomplete LD scenarios, respectively. The most striking feature is, perhaps, the extreme differences in behavior between D' and r. Under complete LD, the pattern of r was much more stable behavior than that of D', as there was very little variation between replicates and r peaked clearly at the QTL position (18 cM). In contrast, D' had a much larger variability between replicates and was clearly multimodal in several instances. Nevertheless, these two measures showed clear maxima at or close to the true QTL position under complete disequilibrium. The picture changes dramatically in the incomplete LD scenario.

TABLE 3

Posterior distribution statistics: incomplete association

Pedigree[a]	Replicate	Analysis[b]	Parameters[c] a/σ_e	Parameters[c] d/σ_e	Parameters[c] σ_e^2	Position (M)	t
Simple	1	LDL	1.03 (0.10)	−0.07 (0.17)	0.87 (0.07)	0.161 (0.053)	81 (12)
		L	1.04 (0.10)	0.00 (0.15)	0.86 (0.07)	0.118 (0.073)	—
	2	LDL	1.03 (0.11)	−0.08 (0.18)	0.87 (0.07)	0.172 (0.039)	82 (15)
		L	1.04 (0.10)	−0.01 (0.16)	0.86 (0.07)	0.143 (0.059)	—
	3	LDL	1.03 (0.11)	−0.07 (0.18)	0.87 (0.07)	0.177 (0.030)	75 (17)
		L	1.04 (0.10)	−0.01 (0.11)	0.86 (0.07)	0.171 (0.048)	—
Complex	1	LDL	0.78 (0.08)	0.09 (0.11)	1.10 (0.08)	0.182 (0.015)	93 (20)
		L	0.78 (0.09)	0.10 (0.13)	1.10 (0.08)	0.188 (0.021)	—
	2	LDL	0.87 (0.15)	0.10 (0.23)	1.15 (0.10)	0.183 (0.034)	80 (13)
		L	0.89 (0.16)	0.16 (0.22)	1.13 (0.10)	0.195 (0.035)	—
	3	LDL	0.99 (0.08)	−0.01 (0.10)	1.02 (0.07)	0.192 (0.030)	130 (20)
		L	1.01 (0.08)	0.03 (0.11)	1.01 (0.07)	0.156 (0.040)	—

Initially, 43% of haplotypes with SNP allele 1 carried the QTL mutant allele.

[a] Simple pedigree population consists of independent full-sib families; complex population is a four-generation pedigree with random mating.

[b] LDL analysis combines both linkage disequilibrium and pedigree information; L analysis uses only linkage.

[c] Mean of the posterior distribution (SD of the posterior distribution).

Here r had maxima only at the nearest microsatellites (15 and 20 cM) but a very flat curve was apparent in clear contrast with the complete LD case. The pattern for D' was not as affected by incomplete LD (Figure 5, bottom left) although the profile was somewhat flatter than that with complete LD. Again, we observed a large variability between replicates. It is apparent that the LD statistics D' and r were higher when using LDL than when using linkage-only methods, although the general pattern was comparable (compare thick solid lines *vs.* thin shaded lines in Figure 5).

DISCUSSION

We have provided a coherent and unified theoretical framework to combine linkage and LD information, as exemplified in Equations 4, 6a, and 6b. The method worked well with simulated data. Here we have used the exponential growth model as described by Morris *et al.* (2000) but the Bayesian framework is flexible and other population models can be incorporated by modifying $p(\mathbf{M}|\boldsymbol{\theta})$ appropriately in Equation 4 or 6. An important feature of the method presented here is that it provides the joint haplotype probability conditional on the QTL genotype, *i.e.*, $p(M_{-L}, \ldots M_R, | \mathbf{S}_0, \boldsymbol{\theta}_-)$, whereas Morris *et al.* (2000) wrote the likelihood as $p(\mathbf{M}|\mathbf{S}_0, \boldsymbol{\theta}_-) = \prod_k p(M_k|\theta)$, which differs from that used here, Equation 5. Take, without loss of generality, two markers. Morris *et al.* (2000, p. 162, bottom) used

$$P(M_1, M_2|S_0) = P(M_1|S_0)P(M_2|S_0),$$

where

$$P(M_1|S_0) = \sum_{S_1} p(M_1|S_1)p(S_1|S_0)$$

and

$$P(M_2|S_0) = \sum_{S_2}\sum_{S_1} p(M_2|S_2)p(S_2|S_1)p(S_1|S_0) = \sum_{S_2} p(M_2|S_2) \sum_{S_1} p(S_2|S_1)p(S_1|S_0).$$

In contrast, we used the actual joint distribution, which is

$$P(M_1, M_2|S_0) = \sum_{S_2}\sum_{S_1} p(M_2|S_2)p(S_2|S_1)p(M_1|S_1)p(S_1|S_0) = \sum_{S_2} p(M_2|S_2) \sum_{S_1} p(S_2|S_1)p(M_1|S_1)p(S_1|S_0).$$

(Equation 5). Unless complete independence exists (which does not make sense in a haplotype analysis), a joint distribution is not equal to the product of the marginals, and our approach should provide more power, even in a LD-only analysis, than that of Morris *et al.* (2000).

Our results show that it is indeed possible to go beyond the 20-cM confidence interval to locate QTL in populations of reasonable size with moderate family sizes and without an extremely dense genotyping. But they also point out that the advantages of combining LD information into the usual linkage framework should not be overemphasized and that its impact may vary dramatically depending on a number of factors. First, the usefulness of LDL over linkage-only methods is heavily dependent on the nature of the association,

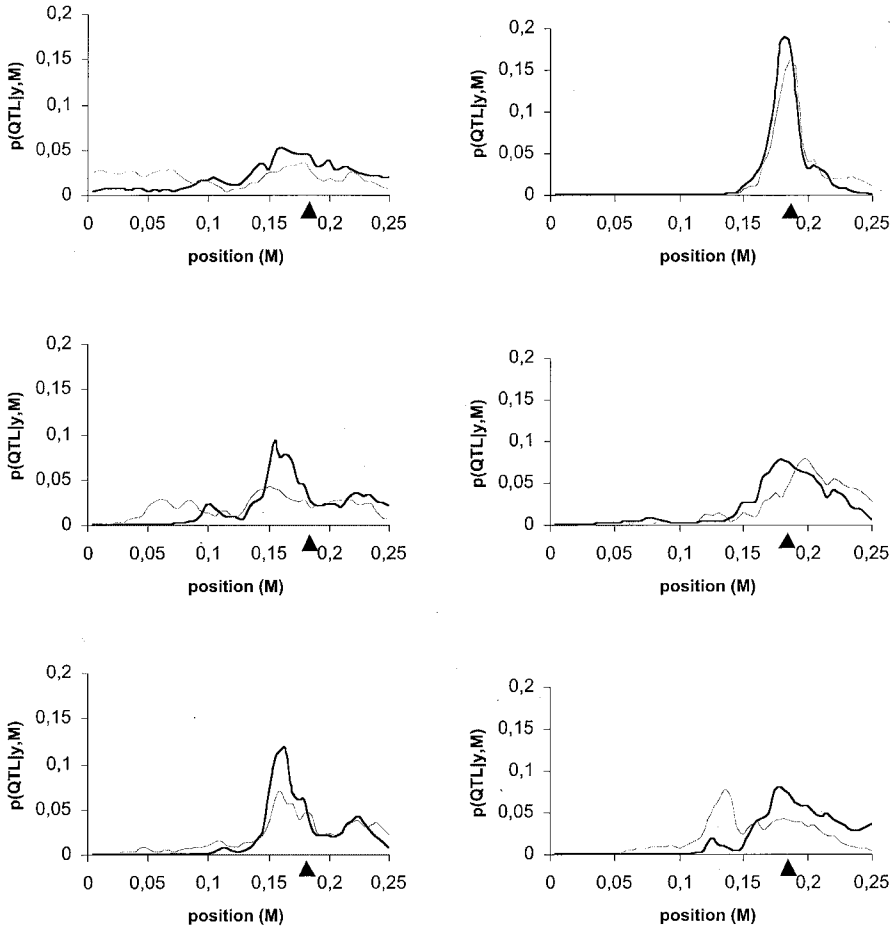

FIGURE 4.—Marginal posterior probabilities of QTL location with incomplete association between QTL and SNP genotype. (Left) Simple population graphs; (right) complex population graphs. The three replicates are shown below each other. The solid thick lines refer to estimates obtained using linkage and linkage disequilibrium, and the thin shaded lines refer to estimates obtained using linkage information only. The QTL was located in position 18 cM (indicated by the arrowhead).

e.g., on whether there is complete LD between the marker and the QTL allele. Second, in the population structure, for accurate LD mapping it is extremely important to determine correctly the phases and the QTL genotypes. Having a small number of base population individuals with large families seems a better option than having a complex pedigree spanning several generations, although the optimum structure will depend on the strength of LD; *e.g.*, if LD is extreme, a large number of base populations animals will be better because we will have more "independent" haplotypes. Finally, chance will affect the results: Mendelian transmission, recombination, and environmental noise are stochastic processes that may result in very different data sets starting from identical initial conditions. A sample of this variability is in Figures 3 and 4, and very interesting experimental results are presented, *e.g.*, in EMAHAZION *et al.* (2001).

Our relatively pessimistic conclusions contrast with much more optimistic views of the advantages of LDL mapping in livestock, more specifically in dairy cattle (FARNIR *et al.* 2002; MEUWISSEN *et al.* 2002). Of course parts of the discrepancies are due to the different methodological approaches. It should also be mentioned that the accuracy of QTL estimates may also be affected by the method of computing the posterior distribution from the MCMC samples (HOTI *et al.* 2002). However, the dairy cattle population structure is ideally suited for LD mapping; very large families and small effective population sizes make it possible to accurately estimate phases and QTL genotypes and reduce genetic heterogeneity. This is not the case for most livestock species and certainly not the case in humans. Results from the group of M. Georges are very illustrative (RIQUET *et al.* 1999; FARNIR *et al.* 2002). Initially, RIQUET *et al.* (1999) located a QTL using only LD information, but that position was shifted to a significantly different position in a later analysis that combined LD and linkage. The primary reason was that sires had different genotypes assigned in each analysis. The population sizes that we used here prevented us from an accurate estimation of both the QTL genotypes of base populations and of some of the phases; these two facts together make it that no one-to-one correspondence between haplotype and QTL genotype can be established unequivocally. As a result, linkage-only methods do not compare too badly with the LDL strategy. MCMC methods take care of the uncertainty but at the price of increasing the variance of the posterior density and thus the accuracy.

In this work, we have also proposed Bayesian equiva-

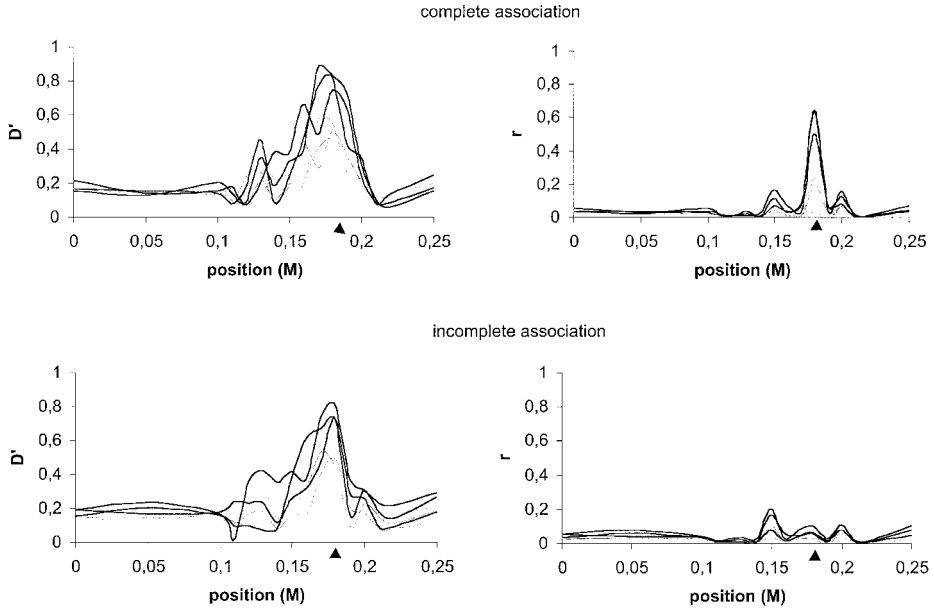

FIGURE 5.—Plots of disequilibrium measures D' and r between each marker and the QTL. The top (bottom) row corresponds to the three replicates with complete (incomplete) association in the simple pedigree. Estimates obtained with the LDL method are shown as thick solid lines and those with linkage only, as thin shaded lines. The QTL was located in position 18 cM (indicated by the arrowhead).

lents for the classical LD measures D' and $r = \sqrt{r^2}$. Interestingly, r and D' exhibited distinct behaviors depending on whether there was a complete association between the QTL and the SNP (Figure 5); r decreased more markedly than D' as we moved away from the QTL with complete association, but the reverse was true with incomplete association. NORDBORG and TAVARÉ (2002) have shown that the D' measure fluctuates more widely than r, which is in agreement with our results. It is important to note that there may be a large variability in disequilibrium decay, as has been evidenced by simulation (*e.g.*, NORDBORG and TAVARÉ 2002; PRITCHARD and PRZEWORSKI 2001) or with experimental data (REICH *et al.* 2001). In particular, it is difficult to compare LD measures of SNPs with those of microsatellites. Disequilibrium measures depend necessarily on allele frequencies and, as argued (NORDBORG and TAVARÉ 2002), they should because gene history and frequency are inextricably linked. Here disequilibrium measures decreased much more rapidly with SNPs than with multiallelic markers. It is also important to bear in mind that the pattern in disequilibrium decay between QTL and marker does not necessarily parallel the posterior distribution of the QTL position, as is evident from comparing the graphs in Figures 3 and 4 (simple pedigree) with those in Figure 5.

Certainly, further extensions and testing of this approach are warranted, particularly to overcome some of the potential risks of using LD. First of all, stratification may cause spurious disequilibrium. In principle, a LDL methodology should be more robust than a pure LD strategy but this remains to be tested and it is uncertain whether stratification has such a large impact on quantitative traits mapping as it does with binary traits. Genetic heterogeneity is also a major problem in quantitative trait loci mapping. In this case there will be a number n_f of original haplotypes carrying a distinct or the same mutation affecting the trait. In our model, this amounts to considering more than either + or − IBD states; an IBD indicator variable should be included and probabilities $p(\mathbf{M}|\mathbf{S}_0 = k, k = 1, n_f)$ should be estimated. In the likely case that n_f is not known, a reversible-jump MCMC strategy could be used. LIU *et al.* (2001) and MORRIS *et al.* (2002) have recently presented an alternative approach to allow for multiple mutations in a pure LD-mapping strategy. Missing markers are dealt with by using only available information for computing phases and segregation indicators. This is a reasonable approximation if the percentage of missing genotypes and the pedigree's complexity are not large; otherwise the transmission coefficients **T** are not properly calculated. This should not be too much of a concern in the special case of fine mapping, where one is usually analyzing a few generations and very dense genotyping. However, this is a much more important limitation in marker-assisted selection or in linkage analysis of complex populations. Here we have implicitly assumed a star-shaped genealogy, which is not realistic in many instances. The dependence among sampled base population haplotypes, *i.e.*, the fact that recombination histories are correlated, can be included in the model via, *e.g.*, coalescent techniques assuming a given effective size (MEUWISSEN and GODDARD 2001). A simple strategy is to consider that prior allele states in any two haplotypes are not independent, *i.e.*, $p(S_{0i}, S_{0i'}) \neq p(S_{0i})p(S_{0i'})$, but rather use the additive relationship coefficient ($\rho_{i,i'}$), computed using all available pedigrees as a measure of association; then $p(S_{0i}, S_{0i'}) = (1 - \rho_{ii'})p(S_{0i})p(S_{0i'}) + \rho_{ii'}\, p(S_{0i})\eta(S_{0i}|S_{0i'})$, as explained in the THEORY section. Much more complicated is the issue of conditioning on the actual known pedi-

gree previous to the first genotyped individuals and their observed marker alleles.

To conclude, fine mapping complex trait genes is a topic of very active research and a major challenge in both human and animal genetics. Given the diversity of genetic architectures and population histories, it is unlikely that a single statistical approach will be valid for all cases. One of the advantages of the Bayesian approach presented here is that the different sources of knowledge are conditionally independent (Equations 4 and 6) so that we can consider, *e.g.*, different population genetic models to model LD simply by changing equation $p(\mathbf{M}|\boldsymbol{\theta})$ appropriately. Additionally, the degree of uncertainty about the parameters can be fully described via the marginal posterior distribution.

We are thankful for helpful discussions with M. Sillanpää, D. Milan, J. M. Elsen, B. Goffinet, and L. L. G. Janss. A referee is thanked for the suggestions. This work was funded by the Bureau des Ressources Génétiques, project no. 20, and Action en Bioinformatique (France).

LITERATURE CITED

Allison, D. B., 1997 Transmission-disequilibrium tests for quantitative traits. Am. J. Hum. Genet. **60:** 676–690.

Allison, D. B., M. Heo, N. Kaplan and E. R. Martin, 1999 Sibling-based tests of linkage and association for quantitative traits. Am. J. Hum. Genet. **64:** 1754–1763.

Almasy, L., J. T. Williams, T. D. Dyer and J. Blangero, 1999 Quantitative trait locus detection using combined linkage, disequilibrium analysis. Genet. Epidemiol. **17** (Suppl. 1): S31–S36.

Devlin, B., and N. Risch, 1995 A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics **29:** 311–322.

Emahazion, T., L. Feuk, M. Jobs, S. L. Sawyer, D. Fredman *et al.*, 2001 SNP association studies in Alzheimer's disease highlight problems for complex disease analysis. Trends Genet. **17:** 407–413.

Farnir, F., B. Grisart, W. Coppieters, J. Riquet, P. Berzi *et al.*, 2002 Simultaneous mining of linkage and linkage disequilibrium to fine map quantitative trait loci in outbred half-sib pedigrees: revisiting the location of a quantitative trait locus with major effect on milk production on bovine chromosome 14. Genetics **161:** 275–287.

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

Fulker, D. W., S. S. Cherny, P. C. Sham and J. K. Hewitt, 1999 Combined linkage and association sib pair analysis for quantitative traits. Am. J. Hum. Genet. **64:** 259–267.

Goddard, M. E., 1992 A mixed model analysis of data on multiple genetic markers. Theor. Appl. Genet. **83:** 878–886.

Hastbacka, J., A. de la Chapelle, M. M. Mahtani, G. Clines, M. P. Reeve-Daly *et al.*, 1994 The diastrophic dysplasia gene encodes a novel sulfate transporter: positional cloning by fine-structure linkage disequilibrium mapping. Cell **78:** 1073–1087.

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

Hedrick, P. W., 1987 Gametic disequilibrium measures: proceed with caution. Genetics **117:** 331–341.

Henderson, C. R., 1984 *Applications of Linear Models in Animal Breeding.* University of Guelph, Guelph, ON, Canada.

Hoti, F. J., M. J. Sillanpaa and L. Holmstrom, 2002 A note on estimating the posterior density of a quantitative trait locus from a Markov chain Monte Carlo sample. Genet. Epidemiol. **22:** 369–376.

Hudson, R. R., 1985 The sampling distribution of linkage disequilibrium under an infinite alleles model without selection. Genetics **109:** 611–631.

Janss, L. L. G., R. Thompson and J. A. 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.

Kaplan, N., W. G. Hill and B. S. Weir, 1995 Likelihood methods for locating disease genes in nonequilibrium populations. Am. J. Hum. Genet. **56:** 18–32.

Lewontin, R. C., 1988 On measures of gametic disequilibrium. Genetics **120:** 849–852.

Liu, J. S., C. Sabatti, J. Teng, B. J. Keats and N. Risch, 2001 Bayesian analysis of haplotypes for linkage disequilibrium mapping. Genome Res. **11:** 1716–1724.

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

McPeek, M. S., and A. Strahs, 1999 Assessement of linkage disequilibrium by the decay of haplotype sharing, with application to fine scale genetic mapping. Am. J. Hum. Genet. **65:** 858–875.

Meuwissen, T. H. E., and M. E. Goddard, 2000 Fine mapping of quantitative trait loci using linkage disequilibria with closely linked marker loci. Genetics **155:** 421–430.

Meuwissen, T. H., and M. E. Goddard, 2001 Prediction of identity by descent probabilities from marker-haplotypes. Genet. Sel. Evol. **33:** 605–634.

Meuwissen, T. H., A. Karlsen, S. Lien, I. Olsaker and M. E. Goddard, 2002 Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. Genetics **161:** 373–379.

Morris, A. P., J. C. Whitaker and D. J. Balding, 2000 Bayesian fine-scale mapping of disease loci, by hidden Markov models. Am. J. Hum. Genet. **67:** 155–169.

Morris, A. P., J. C. Whittaker and D. J. Balding, 2002 Fine-scale mapping of disease loci via shattered coalescent modeling of genealogies. Am. J. Hum. Genet. **70:** 686–707.

Nordborg, M., and S. Tavaré, 2002 Linkage disequilibrium: what history has to tell us. Trends Genet. **18:** 83–90.

Pritchard, J. K., and M. Przeworski, 2001 Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. **69:** 1–14.

Reich, D. E., M. Cargill, S. Bolk, J. Ireland, P. C. Sabeti *et al.*, 2001 Linkage disequilibrium in the human genome. Nature **411:** 199–204.

Riquet, J., W. Coppieters, N. Cambisano, J. J. Arranz, P. Berzi *et al.*, 1999 Fine-mapping of quantitative trait loci by identity by descent in outbred populations: application to milk production in dairy cattle. Proc. Natl. Acad. Sci. USA **96:** 9252–9257.

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.

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

Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and McMc Methods in Quantitative Genetics.* Springer Verlag, New York.

Terwilliger, J. D., 1995 A powerful likelihood method for the analysis of linkage disequilibrium between trait loci and one or more polymorphic marker loci. Am. J. Hum. Genet. **56:** 777–787.

Terwilliger, J. D., and K. M. Weiss, 1998 Linkage disequilibrium mapping of complex disease: Fantasy or reality? Curr. Opin. Biotechnol. **9:** 578–594.

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.

Uimari, P., and M. J. Sillanpää, 2001 Bayesian oligogenic analysis of quantitative and qualitative traits in general pedigrees. Genet. Epidemiol. **21:** 224–242.

Wang, C. S., J. J. Rutledge and D. Gianola, 1993 Marginal inferences about variance components in a mixed linear model using Gibbs sampling. Genet. Sel. Evol. **25:** 41–62.

Weir, B. S., 1996 *Genetic Data Analysis II.* Sinauer, Sunderland, MA.

Wu, R., and Z-B. Zeng, 2001 Joint linkage and linkage disequilibrium mapping in natural populations. Genetics **157:** 899–909.

Xiong, M., and L. Jin, 2000 Combined linkage and linkage disequilibrium mapping for genome screens. Genet. Epidemiol. **19:** 211–234.

Zhao, L. P., C. Aragaki, L. Hsu and F. Quiaoit, 1998 Mapping of complex traits by single nucleotide polymorphisms. Am. J. Hum. Genet. **63:** 225–240.

Communicating editor: C. Haley

APPENDIX: SAMPLING DISTRIBUTIONS

Mixed-model effects (a, d, u, and β): The mixed-model equations (Henderson 1984) are, conditional on $\mathbf{w_a}$, $\mathbf{w_d}$, σ^2_u, and σ^2_e,

$$\begin{bmatrix} \mathbf{X^{*\prime}X^*} & \mathbf{X^{*\prime}Z} \\ \mathbf{Z'X^*} & \mathbf{Z'Z} + \mathbf{A}^{-1}\lambda \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^* \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X^{*\prime}y} \\ \mathbf{Zy} \end{bmatrix}, \quad \text{(A1)}$$

or $\mathbf{Cb} = \mathbf{d}$, where $\mathbf{C}$ is the left-hand-side matrix in (A1) above, $\mathbf{d}$ is the right-hand-side vector, and $\mathbf{b}$ contains $\boldsymbol{\beta}^*$ and $\mathbf{u}$, with $\lambda = \sigma^2_e/\sigma^2_u$. Wang *et al.* (1993) showed that the fully conditional distribution of any element b_i of $\mathbf{b} = [\boldsymbol{\beta}^*, \mathbf{u}]$ is

$$b_i \sim \text{Normal}(d_i - \sum_{j=1, j\neq i}^{N} c_{ij} d_j, \sigma^2_e/c_{ii}), \quad \text{(A2)}$$

where d_i is the ith element of the right-hand-side vector, and c_{ij} is element (i, j) of $\mathbf{C}$, which has dimension N.

Variance components (σ^2_u and σ^2_e): The fully conditional distributions are

$$p(\sigma^2_u|\mathbf{S}_0, a, d, \mathbf{u}, \boldsymbol{\beta}, \sigma^2_e, \mathbf{y}) = (\mathbf{u'}\,\mathbf{A}^{-1}\mathbf{u})\chi^{-2}_m \quad \text{(A3)}$$

and

$$p(\sigma^2_e|\mathbf{S}_0, a, d, \mathbf{u}, \boldsymbol{\beta}, \sigma^2_u, \mathbf{y}) = (\mathbf{y} - \mathbf{X^*}\,\boldsymbol{\beta}^* - \mathbf{Zu})' \times (\mathbf{y} - \mathbf{X^*}\boldsymbol{\beta}^* - \mathbf{Zu})\chi^{-2}_n \quad \text{(A4)}$$

(Wang *et al.* 1993), where χ^{-2}_q stands for an inverted chi-square distribution with q d.f. Equations A3 and A4 assume a naïve ignorance prior. Conjugate informative priors with prior variance O^2 and ν d.f., respectively, result in *posteriori* conditional distributions of the type $(\text{QF} + O^2\nu)\chi^{-2}_{q+\nu}$, where QF is the quadratic form in (A3) or (A4) (Wang *et al.* 1993; Sorensen and Gianola 2002).

Linkage disequilibrium parameters [t, $p(M_{kj}|S_k)$]: The fully conditional distribution of t is not a known distribution and, thus, we resort to Metropolis-Hastings sampling. A new proposed age of mutation t^{new} is accepted with probability

$$\min\left\{1, \frac{p(\mathbf{M}|\mathbf{S_0}, t^{\text{new}}, \mathbf{H}, \delta)}{p(\mathbf{M}|\mathbf{S_0}, t, \mathbf{H}, \delta)}\right\}. \quad \text{(A5)}$$

The probabilities $p(M_{kj}|S_k)$ contain the allele probabilities for each allele j of marker k conditional on the IBD state of the marker with the original mutant haplotype. This variable is updated each iteration as follows. For each base population individual, the probabilities that at marker k the alleles are IBD with the mutant haplotype are calculated given the IBD state at the QTL position, S_0 equaling either + or −. The original frequencies of allele j at marker k in the nonmutant population are obtained from

$$p(M_{kj}|S_k = +, \boldsymbol{\theta}\_) = \sum_{i=1}^{F}\sum_{h=1}^{2} p(S_{kih} = +|S_{0ih})\eta_{ihjk}/(2F),$$

where F is the number of base population haplotypes, η_{ihjk} is an indicator variable taking value = 1 if the individual i has allele j at marker k and haplotype h, and zero otherwise. Similarly, we compute

$$p(M_{kj}|S_k = -, \boldsymbol{\theta}\_) = \sum_{i=1}^{F}\sum_{h=1}^{2} p(S_{kih} = -|S_{0ih})\eta_{ihjk}/(2F),$$

which is the probability that the original mutant haplotype contains allele j at marker k. An alternative option is to sample the original mutant haplotype as in Morris *et al.* (2000). However, and unless we are interested in reconstructing the original haplotype, we prefer the approach here, whereby the founder haplotype, that where QTL mutation occurred, is treated as a nuisance parameter and integrated out.

Phase sampling (H): Phases that could not be determined unambiguously were sampled using a block Gibbs sampling algorithm. A parameterizable number of marker phases were sampled jointly for each individual in turn. The algorithm works as follows. First, unknown phases for a given individual are identified, say n_h unknown phases. Second, an indicator variable is constructed taking all possible values (2^{n_h}). For instance, suppose that there are four markers and that the phases of first and last markers are known or not sampled (*i.e.*, missing marker), then the indicator variable may take values −00−, −01−, −10−, and −11−, where "−" stands for not sampled, "0" for paternal, and "1" for maternal origin. Finally, the probability associated with each value is calculated using all available marker information and current phases in parents and offspring and a new phase block is sampled. Here a maximum of six phases were sampled jointly.

Segregation indicators (T): **T** was usually updated together with the QTL position, as explained below. A new proposal for **T** was sampled conditioning on marker and phase information using Mendelian rules.

QTL position (δ): This is one of the most critical steps of the Bayesian procedure. A variety of strategies have been proposed in the literature (Satagopan *et al.* 1996; Heath 1997; Uimari and Hoeschele 1997; Sillanpaa and Arjas 1998). In a typical sampling scheme, individual S_0 would be updated conditional on the other genotypes, but this is a risky option as the chain will get stuck easily (Janss *et al.* 1995). Uimari and Sillanpää (2001) proposed a dual sampling scheme. In some iterations, δ is updated using the acceptance ratio

$$\min\left\{1, \frac{p(\mathbf{T}|\delta^{new}, \mathbf{H})}{p(\mathbf{T}|\delta, \mathbf{H})}\right\}. \quad \text{(A6)}$$

However, using (A6) may prevent δ from "jumping" between adjacent marker intervals because the above acceptance ratio is very sensitive to the percentage of QTL recombinant haplotypes, which in turn depends on the marker interval. In other iterations **T** and δ were updated simultaneously. A new **T** was generated as described using a new position, δ^{new}, and both $\mathbf{T}^{new}$ and δ^{new} were accepted with probability

$$\min\left\{1, \frac{p(\mathbf{y}|\mathbf{T}^{new}, \mathbf{S}_0, a, d, \mathbf{u}, \beta, \sigma_e^2)}{p(\mathbf{y}|\mathbf{T}, \mathbf{S}_0, a, d, \mathbf{u}, \beta, \sigma_e^2)}\right\} \quad \text{(A6′)}$$

(Uimari and Sillanpää 2001). Otherwise **T** and δ remained unchanged. Here, sampling was normally performed via (A6′), except every five iterations when (A6) was used.