Mapping-Linked Quantitative Trait Loci Using Bayesian Analysis and Markov Chain Monte Carlo Algorithms

Pekka Uimari,[1] and Ina Hoeschele

Department of Dairy Science, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0315

Manuscript received September 13, 1996
Accepted for publication March 10, 1997

ABSTRACT

A Bayesian method for mapping linked quantitative trait loci (QTL) using multiple linked genetic markers is presented. Parameter estimation and hypothesis testing was implemented via Markov chain Monte Carlo (MCMC) algorithms. Parameters included were allele frequencies and substitution effects for two biallelic QTL, map positions of the QTL and markers, allele frequencies of the markers, and polygenic and residual variances. Missing data were polygenic effects and multi-locus marker-QTL genotypes. Three different MCMC schemes for testing the presence of a single or two linked QTL on the chromosome were compared. The first approach includes a model indicator variable representing two unlinked QTL affecting the trait, one linked and one unlinked QTL, or both QTL linked with the markers. The second approach incorporates an indicator variable for each QTL into the model for phenotype, allowing or not allowing for a substitution effect of a QTL on phenotype, and the third approach is based on model determination by reversible jump MCMC. Methods were evaluated empirically by analyzing simulated granddaughter designs. All methods identified correctly a second, linked QTL and did not reject the one-QTL model when there was only a single QTL and no additional or an unlinked QTL.

CURRENTLY, several methods are available for mapping genes in outbred livestock populations, including methods based on linear regression (LS), maximum likelihood analysis (ML), residual maximum likelihood (REML) with expected covariance matrix of random quantitative trait locus (QTL) effects, and Bayesian analysis. LS is computationally fast, easy to implement even with standard statistical packages, provides estimates of QTL position but not of any other genetic parameters, and is restricted to specific mating designs. The REML method (GRIGNOLA *et al.* 1994, 1996a,b; GRIGNOLA and HOESCHELE 1996) postulates normally distributed QTL allelic effects and is computationally somewhat more demanding than LS analysis, provides estimates of variance contributions of QTL, and allows use of pedigree information in fitting polygenic and QTL effects. REML with an expected covariance matrix of QTL effects conditional on observed marker information is an approximation to ML and Bayesian analyses. These latter methods account for the distribution of genotypes on the pedigree given observed phenotypic and marker data, and enable the estimation of all genetic parameters assuming specific QTL models (*e.g.*, biallelic or normal-effects QTL), but are computationally very demanding. Approximate implementations of ML linkage analysis for livestock populations are available (*e.g.*, WELLER 1986; GEORGES *et al.* 1995; MACKINNON and WELLER 1995). Bayesian analysis provides alternative parameter estimators to the standard ML estimators conditional on a most likely QTL position and alternative tests for linkage based on posterior probabilities of linkage (HOESCHELE and VANRADEN 1993a,b). Additional fixed and random effects can be included in the analysis, as well as allele frequencies and map positions of markers. Bayesian analysis utilizes pedigree information in fitting QTL and polygenic effects, and is suitable for different breeding designs. A more complete review of different statistical methods for gene mapping in livestock populations is given by HOESCHELE *et al.* (1996).

In earlier contributions, Bayesian linkage analysis for outbred livestock populations proposed by HOESCHELE and VANRADEN (1993a,b) was implemented via Markov chain Monte Carlo (MCMC) algorithms for a single marker and a biallelic QTL (THALLER and HOESCHELE 1996a,b). UIMARI *et al.* (1996a) extended this method to Bayesian analysis mapping a biallelic QTL using multiple linked markers. HOESCHELE *et al.* (1996) modified the latter method to fit a normal-effects QTL model.

The current paper was motivated by evidence of detecting a single "ghost QTL" with LS analysis, when two linked QTL were actually segregating (HALEY and KNOTT 1992; MARTINEZ and CURNOW 1992). The same phenomenon has been reported by GRIGNOLA and HOESCHELE (1996) using REML. To avoid detecting a ghost QTL, a two-dimensional search for two QTL can

Corresponding author: Ina Hoeschele, Department of Dairy Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061-0315. E-mail: inah@vt.edu

[1] *Present address:* Centre for Genetic Improvement of Livestock, Department of Animal and Poultry Science, University of Guelph, Guelph, Ontario N1G 2W1, Canada.

be performed with LS and REML methods (GRIGNOLA and HOESCHELE 1996; SPELMAN *et al.* 1996; UIMARI *et al.* 1996b). An alternative approach has been suggested by ZENG (1993, 1994) and XU and ATCHLEY (1995), referred to as Composite Interval Mapping, where phenotypes are regressed on linked and unlinked markers, or marker variances are fitted, respectively. Furthermore, JANSEN (1993, 1996) and JANSEN and STAM (1994) developed maximum likelihood interval mapping including selected markers as cofactors to account for the effects of other QTL.

In this paper, we extend Bayesian linkage analysis to two linked QTL. Three different MCMC algorithms are compared for testing the hypotheses of zero, one and two QTL linked to the markers. The first approach follows the idea of THALLER and HOESCHELE (1996a) and UIMARI *et al.* (1996a), where a linkage indicator variable is included in the joint posterior distribution. The second approach employs the idea of variable selection by KUO and MALLICK (1994), and the third approach applies Bayesian model determination by reversible jump MCMC (GREEN 1995). The methods are evaluated by analyzing simulated granddaughter designs.

MATERIALS AND METHODS

Bayesian inferences about the number of linked QTL and parameters were based on the joint posterior distribution of missing data and parameters given observed marker (M) and phenotypic data (y) of a quantitative trait. Parameter vector θ included gene frequencies (p_1 and p_2), substitution effects (α_1 and α_2) and map positions (d_{Q1} and d_{Q2}) of the QTL (QTL_1 and QTL_2), vector of allele frequencies ($\mathbf{q}$) and map positions ($\mathbf{d}$) of m marker loci, an overall mean and additional fixed effects (β), and polygenic (σ_u^2) and residual (σ_e^2) variances. The position of the first marker was taken as the origin of the linkage group ($d_1 = 0$). The missing data included polygenic effects (u) and multi-locus marker-QTL genotypes (MG) for the entire pedigree. Following UIMARI *et al.* (1996), the MG genotype was defined such that in each Gibbs cycle the linkage phases of the markers and QTL are known. This approach enables the sampling of QTL allele frequencies from standard Beta distributions and the sampling of marker allele frequencies from Dirichlet distributions. Map positions were converted to recombination rates using Haldane's no interference map function. Below, $P(.)$ will denote the joint probability of discrete variables and $f(.)$ the joint probability density of continuous variables or a combination of discrete and continuous variables.

Also included in the joint posterior distribution and in the Gibbs sampler were indicator variables for the linkage model (no QTL, one QTL, or two QTL linked to the markers). Three MCMC algorithms that differed in the definition of the linkage indicator variables and the one (or single-)-QTL model were compared.

The first approach (MCMC scheme 1) follows the idea of THALLER and HOESCHELE (1996a,b) and UIMARI *et al.* (1996a), where a linkage indicator variable ℓ is included in the joint posterior distribution. The following notation is used to denote the four different states of linkage: $l = 00$ denotes two unlinked QTL, $l = 01$ and $l = 10$ denote one linked and one unlinked QTL, and $l = 11$ denotes two linked QTL on the chromosome under a study. Note that the single-QTL model in scheme 1 postulates one linked and another unlinked QTL, and polygenic effects. The joint posterior density of the parameters, the missing data, and the linkage indicator variable given observed marker and phenotypic data is

$$f(\theta, MG, u, l|y, M) \propto f(l)f(\theta|l)f(u|\theta)P(MG|\theta) \times P(M|MG)f(y|\theta, u, MG), \quad (1)$$

where parameters are assumed to be independent *a priori*, or

$$f(\theta|l) = f(\beta)f(\alpha_1)f(\alpha_2)f(p_1)f(p_2) \times f(d_{Q1}, d_{Q2}|l)f(\mathbf{q})f(\mathbf{d})f(\sigma_u^2)f(\sigma_e^2) \quad (2)$$

and observations are independent conditional on missing data, or

$$P(M|MG)f(y|\theta, u, MG) = \prod_{i=1}^{n} P(M_i|MG_i) \times \prod_{i=1}^{N} f(y_i|\beta, \alpha_1, \alpha_2, u_i, MG_i, \sigma_e^2), \quad (3)$$

where n (N) is the number of individuals with marker (phenotypic) data, and $f(\beta)$ = constant. Prior distributions are as follows: uniform on $[p_l, p_u]$ for p_1 and p_2 where $0 < p_l < p_u < 1$; uniform on $[\alpha_l, \alpha_u]$, with $0 < \alpha_l < \alpha_u$ and α_u being a large constant, for α_1 and α_2; and uniform on $[0, c]$ for σ_u^2 and σ_e^2, where c is a large constant. Marker allele frequencies are Dirichlet (1) at each locus. The marker positions are order statistics from a uniform distribution on the assumed prior length of the linkage group, and d_{Q1} and d_{Q2} are order statistics from a uniform distribution (QTL_1 is to the left of QTL_2). If $l = 00$, d_{Q1} and d_{Q2} are uniform on the remainder of the genome not including the chromosome under study ($[T_u - T_l, T]$). If $l = 10$, d_{Q1} and d_{Q2} are uniform on the chromosome under study ($[T_l, T_u]$) and on $[T_u - T_l, T]$, respectively. Similarly, if $l = 01$, d_{Q1} and d_{Q2} are uniform on $[T_u - T_l, T]$ and $[T_l, T_u]$, respectively; and if $l = 11$, d_{Q1} and d_{Q2} are order statistics from the uniform $[T_l, T_u]$, where T is the total length of the genome, and T_l and T_u are assumed lower and upper limits, relative to the origin of the linkage group, of the chromosome under a study, respectively. Furthermore, $f(u|\theta) = f(u|\sigma_u^2)$ is the density of $N(0, \mathbf{A}\sigma_u^2)$, where $\mathbf{A}$ is a known additive genetic relationship matrix, and $P(MG|\theta)$ is the joint probability of a set of multilocus genotypes on the pedigree.

Univariate conditional sampling distributions were used for all parameters in (1) except that l, d_{Q1}, and d_{Q2}

were sampled jointly using the method of composition (TANNER 1993). Linkage indicator variable l was sampled according to the conditional probability

$$P(l = j|MG, \mathbf{d}) = \frac{P(l = j)P(MG|\mathbf{d}, l = j)}{\sum_{k=00}^{11} P(l = k)P(MG|\mathbf{d}, l = k)}, \quad (4)$$

where $k = 00$, 01, 10, and 11, and

$$P(MG|\mathbf{d}, l = 00) = P(MG|\mathbf{d}, r_{Q1} = 0.5, r_{Q2} = 0.5)$$

$$P(MG|\mathbf{d}, l = 01) = \int_0^{T_u} \times P(MG|\mathbf{d}, d_{Q2}, r_{Q1} = 0.5) f(d_{Q2})\, dd_{Q2}$$

$$P(MG|\mathbf{d}, l = 10) = \int_{T_l}^{d_m} \times P(MG|\mathbf{d}, d_{Q1}, r_{Q2} = 0.5) f(d_{Q1})\, dd_{Q1}$$

$$P(MG|\mathbf{d}, l = 11) = \int_0^{T_u} \int_{T_l}^{\min(d_{LMQ2}, d_{Q2}-c)} \times P(MG|\mathbf{d}, d_{Q1}, d_{Q2}) f(d_{Q1}) f(d_{Q2})\, dd_{Q1} d_{Q2}, \quad (5)$$

where r_{Qk} ($k = 1, 2$) is recombination rate between QTL_k and the first marker, and 0 and d_m are the positions of the first and the last marker, respectively. Note that (4) is obtained by conditioning on all variables except d_{Q1} and d_{Q2}. For the last conditional probability of MG in (5), the inner integration is for QTL_1 with integration limits equal to the left end of the chromosome and the minimum of the position of the left flanking marker of QTL_2 (d_{LMQ2}) and a position c cM to the left of QTL_2, with c cM being the allowed minimum distance between QTL. The outer integration for QTL_2 is from the origin of the linkage group to the chromosome end. These integration spaces were defined to assure that at least one marker is located between the QTLs. Samples of the QTL map positions were obtained by discretizing their conditional (on l, $\mathbf{d}$, and MG) distribution, *i.e.*, by computing the conditional probability densities of d_{Q1} and d_{Q2} on a two-dimensional grid over their joint sample space. Sampling distributions for the other parameters and missing data can be found in UIMARI *et al.* (1996a).

The second approach (MCMC scheme 2) follows the idea of variable selection in linear and generalized linear models of KUO and MALLICK (1994). Note that for this scheme, the single QTL model postulates that the trait is affected by one linked QTL and polygenes, but not by an additional unlinked QTL. The linear model for phenotype y of individual i given its QTL genotypes is

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + \lambda_1 z_{i1}\alpha_1 + \lambda_2 z_{i2}\alpha_2 + u_i + e_i, \quad (6)$$

where $\mathbf{x}_i'$ is row i of design covariate matrix $\mathbf{X}$, $\boldsymbol{\beta}$ is a vector of fixed effects, z_{i1} and z_{i2} are coefficients equal to -1, 0 or $+1$ depending on the genotype at each biallelic QTL, and λ_1 and λ_2 are indicator variables taking values 0 or 1. If $\lambda_k = 1$ ($k = 1, 2$), QTL_k with substitution effect α_k has an effect on mean phenotype. The joint posterior density of the parameters, the indicator variables and missing data given observed marker and phenotypic data is

$$f(\theta, MG, u, \boldsymbol{\lambda}|y, M) \propto P(\boldsymbol{\lambda}) f(\theta) f(u|\theta) P(MG|\theta) \times P(M|MG) f(y|\theta, u, MG, \boldsymbol{\lambda}), \quad (7)$$

where $\boldsymbol{\lambda} = [\lambda_1, \lambda_2]'$. Vector $\boldsymbol{\lambda}$ was sampled according to the conditional probability

$$P(\lambda_1 = j, \lambda_2 = j'|\theta, u, MG, y) = \frac{P(\lambda_1 = j, \lambda_2 = j') f(y|\theta, u, MG, \lambda_1, \lambda_2)}{\sum_{k=0}^{1}\sum_{k'=0}^{1} P(\lambda_1 = k, \lambda_2 = k') f(y|\theta, u, MG, \lambda_1, \lambda_2)}. \quad (8)$$

Conditional on λ, the sampling distributions for the parameters and missing data are as before, except that if $\lambda_k = 0$ ($k = 1, 2$), then α_k is sampled from its prior distribution.

The third approach (MCMC scheme 3) is based on Bayesian model determination by reversible jump Markov chain Monte Carlo (GREEN 1995). In this approach, one linkage indicator variable is defined as in the single QTL analysis of UIMARI *et al.* (1996a) with $l = 0$ (no linkage) and $l = 1$ (at least one QTL linked), and an additional QTL model indicator variable is denoted by h with $h = 1$ (one QTL linked) and $h = 2$ (two QTL linked). This analysis can alternatively be conducted in two steps by first performing the single QTL analysis of UIMARI *et al.* (1996), and if linkage ($l = 1$) is sampled frequently, subsequently running an analysis with $l = 1$ fixed and h being sampled. The two-step approach is chosen here. As for MCMC scheme 2, the one-QTL model ($h = 1$) postulates a single linked QTL and polygenes, but not an additional unlinked QTL.

The dimension of the parameter vector and the sampling spaces of the MG differ for the cases of one *vs.* two linked QTL. The reversible jump MCMC approach allows switching between different sample spaces within a single Gibbs chain. Define $\theta_1 = [\beta, \alpha_{11}, p_{11}, d_{Q11}, \mathbf{d}, \mathbf{q}, \sigma_u^2, \sigma_e^2]$ as the parameter vector under the one-QTL model, $\theta_2 = [\beta, \alpha_{21}, \alpha_{22}, p_{21}, p_{22}, d_{Q21}, d_{Q22}, \mathbf{d}, \mathbf{q}, \sigma_u^2, \sigma_e^2]$ as the parameter vector under two-QTL model, $\mathbf{MG}_1$ as the vector of multi-locus marker-QTL_1 genotypes, and $\mathbf{MG}_2$ as the vector of marker-QTL_1-QTL_2 genotypes. Then, a move from one model or sample space to another is proposed according to the probabilities of "birth" and "death", where birth refers to a move from a lower to a higher dimensional space, here from the one-QTL to the two-QTL model, and death is the opposite move. When the current model is the one-QTL model, probabilities of birth and death are

$$b_1 = c\min\{1, P(h = 2)/P(h = 1)\} \quad \text{and} \quad d_1 = 0,$$

and when the current model is the two-QTL model,

$b_2 = 0$ and $d_2 = c \min\{1, P(h = 1)/P(h = 2)\}$,

where $P(h = 1)$ and $P(h = 2)$ are prior probabilities of the one-QTL and the two-QTL model, respectively, and c is a constant to be specified. [GREEN (1995) recommends setting c such that $b_k + d_k = 0.9$ for $k = 1, 2$].

A move from the one-QTL to the two-QTL model requires obtaining values for α_{21}, α_{22}, p_{21}, p_{22}, d_{Q21}, and d_{Q22}, while the reverse move requires obtaining values for α_{11}, p_{11}, and d_{Q11}. We chose to set $\alpha_{21} = \alpha_{11}$ and generate $\alpha_{22} = w_\alpha$ with $w_\alpha \sim U[\alpha_l, \alpha_u]$. New p_{21} and p_{22} were generated from the current p_{11} as $p_{21} = p_{11} + w_p$ and $p_{22} = p_{11} - w_p$, where $w_p \sim U[p_{\min}, p_{\max}]$ and $p_{\max} = \min\{p_u - p_{11}, p_{11} - p_l, 0.20\}$ and $p_{\min} = -p_{\max}$. New QTL positions were generated from d_{Q11} as $d_{Q21} = d_{Q11}$ and $d_{Q22} = d_{Q11} + w_d$, where $w_d \sim U[d_{\min}, d_{\max}]$ and $d_{\min} = \max\{d_{Q1} - d_{LMQ1}, d_{RMQ1} - d_{Q1}, 20$ cM$\}$ and $d_{\max} = d_{\min} + 40$ cM (d_{RMQ1} is the position of the right flanking marker of QTL$_1$). For the reverse move from the two-QTL to the one-QTL model, $\alpha_{11} = \alpha_{21}$, $p_{11} = 0.5$ ($p_{21} + p_{22}$), and $d_{Q11} = d_{Q21}$. These proposal distributions for the parameters satisfy the dimension-matching requirement of reversible jump MCMC (GREEN 1995), *i.e.*, the dimension of $[\theta_1, w_\alpha, w_p, w_d]$ and $[\theta_2]$ are equal. Other proposal distributions consistent with dimension matching can be used.

Because $\mathbf{MG}_1$ does not contain genotypes for QTL$_2$, $\mathbf{MG}_2$ has to be generated also subject to the dimension-matching requirement. Marker and QTL$_1$ ($\mathbf{G}_1$) genotypes were the same as under the one-QTL model, and QTL$_2$ genotypes ($\mathbf{G}_2$) were sampled using the conditional probabilities $P(\mathbf{G}_2|\boldsymbol{\theta}_2, \mathbf{MG}_1)$. For the reverse move from the two-QTL to the one-QTL model, $\mathbf{G}_1$ was kept and $\mathbf{G}_2$ deleted.

A proposed move is accepted or rejected in a Metropolis-Hasting step. A move from the one-QTL model to the two-QTL model is accepted according to the probability

$$\min\left\{1, \frac{P(h=2)}{P(h=1)} \frac{f(\boldsymbol{\theta}_2)}{f(\boldsymbol{\theta}_1)f(\mathbf{w})} \frac{P(\mathbf{MG}_2|\boldsymbol{\theta}_2)}{P(\mathbf{MG}_1|\boldsymbol{\theta}_1)P(\mathbf{G}_2|\boldsymbol{\theta}_2, \mathbf{MG}_1)} \times \frac{f(\mathbf{y}|\boldsymbol{\theta}_2, \mathbf{MG}_2, \mathbf{u})}{f(\mathbf{y}|\boldsymbol{\theta}_1, \mathbf{MG}_1, \mathbf{u})} \frac{f(\mathbf{M}|\mathbf{MG}_2)}{f(\mathbf{M}|\mathbf{MG}_1)} \frac{d_2}{b_1} \times J\right\}, \quad (9)$$

where

$$J = \left|\frac{\partial \boldsymbol{\theta}_2}{\partial(\boldsymbol{\theta}_1, \mathbf{w})}\right|$$

is a Jacobian and $\mathbf{w} = [w_\alpha, w_p, w_d]$. After simplification the probability is

$$\min\left\{1, \frac{P(h=2)}{P(h=1)} \frac{f(\boldsymbol{\theta}_2)}{f(\boldsymbol{\theta}_1)f(\mathbf{w})} \frac{P(\mathbf{MG}_1|\boldsymbol{\theta}_2)}{P(\mathbf{MG}_1|\boldsymbol{\theta}_1)} \times \frac{f(\mathbf{y}|\boldsymbol{\theta}_2, \mathbf{MG}_2, \mathbf{u})}{f(\mathbf{y}|\boldsymbol{\theta}_1, \mathbf{MG}_1, \mathbf{u})} \frac{d_2}{b_1} \left|\frac{\partial \boldsymbol{\theta}_2}{\partial(\boldsymbol{\theta}_1, \mathbf{w})}\right|\right\}. \quad (10)$$

TABLE 1

Granddaughter designs with different map positions and substitution effects (in genetic standard deviations) of two biallelic QTL

	QTL$_1$		QTL$_2$	
Design	Position (d_{Q1}) (cM)	Effect (α_1)	Position (d_{Q2}) (cM)	Effect (α_2)
I	30	1.0	70	0.75
II	30	1.0	70	0.375
III	30	1.0	—	—
IV	30	1.0	Unlinked	0.75

A move from the two-QTL to the one-QTL model is accepted according to the reciprocal of the above formula.

Simulation: The methods were evaluated empirically by analyzing simulated granddaughter designs (WELLER *et al.* 1990). The simulated pedigree structure was identical to that of THALLER and HOESCHELE (1996b) and UIMARI *et al.* (1996a) with 2000 sons, 20 sires, and nine ancestors of sires. All dams were unrelated. Phenotypic data consisted of daughter yield deviations (DYDs) (VANRADEN and WIGGANS 1991) of 2000 sons. Heritability of the quantitative trait was set to 0.3, and reliability of DYDs was 0.7. Marker information consisted of five markers with five alleles each at equal frequencies and was available for all animals in the pedigree excluding dams. Markers were 20 cM apart. QTL were assumed to be biallelic. Locations and effects of the QTL varied, and are listed in Table 1. Depending on the design, one linked QTL, one unlinked and one linked QTL, or two linked QTL were simulated, and the remaining genetic variation was polygenic. QTL and marker genotypes were simulated for all sons, sires, and ancestors. Each design was replicated four times and analyzed with all three MCMC schemes.

RESULTS

Starting values: For all analyses the same starting values were used. Starting values for the parameters in θ were arbitrary, and true values were used as starting values for d (this can be justified by the fact that marker distances are usually well estimated in advance). Starting values for MG and u were obtained by first sampling sires and sons jointly by ignoring ancestors of the sires, and then sampling the paternal ancestors conditional on offspring genotypes but ignoring parental genotypes. The starting model for the first MCMC scheme was the one-QTL model or $l = 10$; $\lambda_1 = 1$, $\lambda_2 = 0$ for the second scheme, and $h = 1$ (with $l = 1$ fixed) for the third scheme. The starting value for d_{Q1} was 25 cM for all MCMC schemes. The starting position for the second QTL was "unlinked" for the first MCMC scheme, 65 cM for the second MCMC scheme, and no starting position was required for the third MCMC

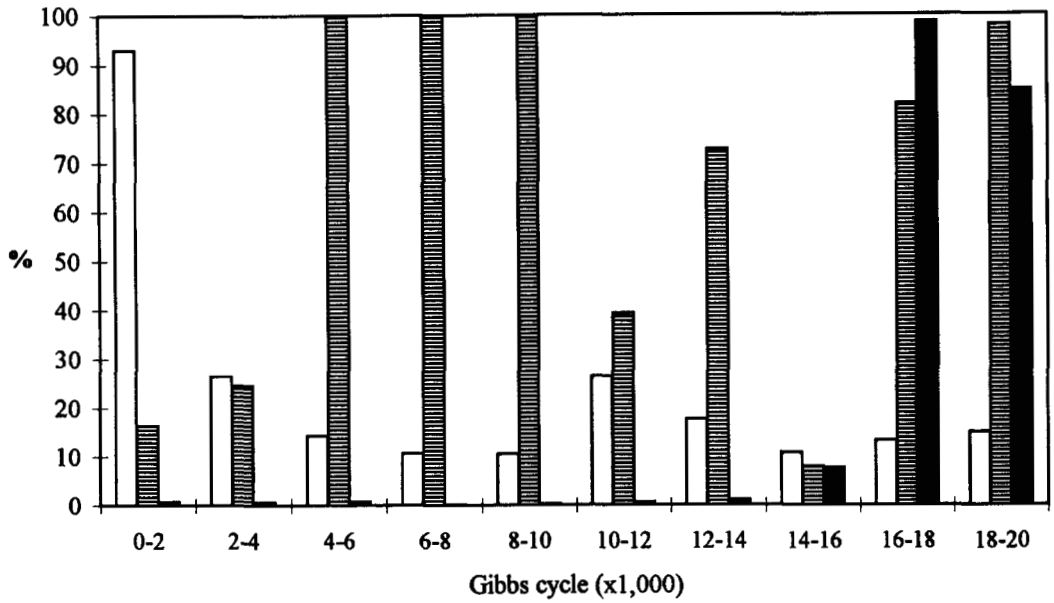

FIGURE 1.—Proportions (%) of samples where the two-QTL model was selected. Proportions were calculated within consecutive blocks of 2000 Gibbs cycles, for design II and MCMC schemes 1 (□), 2 (▤), and 3 (■).

scheme, where the one-QTL model did not include d_{Q2} in the parameter vector.

Prior distributions: For MCMC schemes 1 and 2, prior probabilities for $l = 00$, $l = 10$, $l = 01$, and $l = 11$ (or $\lambda = [0, 0]'$, $\lambda = [1, 0]'$, $\lambda = [0, 1]'$, and $\lambda = [1, 1]'$) were 0.8, 0.08, 0.08, and 0.04, respectively. For reversible jump MCMC, equivalent values were $P(l = 0) = 0.80$ and $P(l = 1) = 0.20$, and $P(h = 1) = 0.80$ and $P(h = 2) = 0.20$. Limits for the prior distributions of the QTL substitution effects varied depending on the MCMC scheme used. The upper limits for α_1 and α_2 were set to $2.0\sigma_a$, where σ_a is a prior guess of the total genetic variance. For the first MCMC scheme, the lower limit was set to $0.25\sigma_a$. This limit was chosen to ensure that the model with one linked QTL is sampled by moving the second QTL away from the linkage group instead of allowing its effect to become very small and still linked, if the one-QTL model is true. For the second MCMC scheme, the lower limit for the QTL effect was set to 0, allowing an unlinked QTL to reach a small value, and thus allowing the sampling of $\lambda_k = 0$ ($k = 1, 2$). For reversible jump MCMC, the lower limit was set to 0.

Diagnostics from Gibbs output: After 2000 cycles of burn-in, the length of a single Gibbs chain was set to 20,000 for model selection. Few chains with 100,000 cycles were run to verify that parameter estimates from the shorter chains were consistent with those from the longer chains. Based on the estimated autocorrelations for lags 1–5000 (GEYER 1992), an effective sample size (ESS), which estimates the number of independent samples with information content equal to that of the dependent sample (SORENSEN *et al.* 1995), was computed for each parameter. Analysis of the simulated data with a Gibbs chain of 20,000 took ~26, 7 and 20 hr for the MCMC schemes 1, 2 and 3, respectively, on an IBM-SP2 with RS/6000 390/590 processors. The most time consuming parts of the programs were the sampling of MG genotypes, the joint sampling of d_{Q1} and d_{Q2}, and the calculation of the ratio in (10) for the reversible jump MCMC.

To monitor the movement of the samplers between QTL models, Figure 1 contains the frequencies of cycles with the one-QTL model sampled, calculated within consecutive blocks of 2000 cycles, for all three MCMC schemes (see below).

Model selection: Frequencies of the sampled models (the one-QTL and the two-QTL model), averaged over four replicates (numbers for individual replicates given in parentheses), are presented in Table 2. All MCMC schemes yielded similar results, were able to select the correct two-QTL model when the second QTL had a relatively large effect on phenotype (Table 2), and were able to move between the two models when the second QTL had a small effect on phenotype (Figure 1). In Figure 1 the proportion of samples, where two-QTL model was selected, within consecutive blocks of 2000 Gibbs cycles is presented for a replicate for granddaughter design II. More similar frequencies across blocks should indicate faster movement of the chain between models. From this figure and similar findings for other replicates, it appears that the movement for scheme 1 is best.

The single-QTL model for the MCMC scheme with linkage indicator variable l includes a second unlinked QTL (design IV), while the single-QTL model for the other MCMC schemes includes only a single QTL in addition to polygenes (design III). Due to the difference in single-QTL models between MCMC schemes, all MCMC schemes were applied to designs III and IV representing the two alternative single-QTL models. All schemes correctly sampled the single-QTL model much more frequently than the two-QTL model for designs III and IV. The lower frequency of the single-QTL

TABLE 2

Frequency (%) of samples where two-QTL model was selected using linkage indicator variable ℓ (MCMC scheme 1), variable selection of KUO and MALLICK (1994) (MCMC scheme 2), and reversible jump MCMC of GREEN (1995) (MCMC scheme 3)

	Granddaughter design[a]			
MCMC scheme	I	II	III	IV
1	100 (100, 100, 99, 100)[b]	43 (55, 82, 24, 11)	18 (10, 10, 18, 33)	38 (73, 21, 27, 32)
2	100 (100, 100, 100, 100)	58 (96, 23, 51, 64)	16 (12, 3, 44, 4)	11 (20, 14, 4, 7)
3	93 (94, 100, 97, 80)	52 (82, 45, 64, 19)	2 (2, 4, 2, 2)	3 (4, 1, 6, 1)

[a] Designs are defined in Table 1.
[b] First number is average frequency across four replicates. Individual frequencies of the four replicates are given in parentheses.

model for MCMC scheme 1 and design IV was due to the first replicate, where the position of the second QTL was sampled frequently in the left flank at a recombination rate of 0.43 with the first marker, hence the two-QTL model was sampled with an almost unlinked second QTL. A model with two unlinked QTL and polygenic effects, or with polygenic effects only, was included in MCMC schemes one and two, respectively. However, these models were never sampled, a finding that is consistent with results from a previous study (UIMARI *et al.* 1996a), where nonlinkage was never sampled (after burn-in) when there was a single linked QTL with a large effect (UIMARI *et al.* 1996a).

To facilitate movement between QTL models for MCMC scheme 3, adjustments to parameters and missing data present under both the one- and two-QTL models were investigated. For example, when attempting a move from the one- to the two-QTL model, the mean was adjusted for the new genotypic mean as

$$\mu = \mu_* + (p_{21} - 0.5)\alpha_{21} + (p_{22} - 0.5)\alpha_{22} - (p_{11} - 0.5)\alpha_{11},$$

where * denotes a parameter in the one-QTL model. Furthermore, the total genetic variances in the one- and two-QTL models were set equal in the transition between models, or

$$0.5p_{11}(1 - p_{11})\alpha_{11}^2 + \sigma_{u^*}^2 = 0.5p_{21}(1 - p_{21})\alpha_{21}^2 + 0.5p_{22}(1 - p_{22})\alpha_{22}^2 + \sigma_u^2.$$

When currently at the one-QTL model and with new QTL parameters for the two-QTL model generated, the above equation was solved for σ_u^2, and subsequently polygenic values (u) were rescaled by multiplication with $\sigma_u/\sigma_{u'}$. The results presented in Table 2 were computed without adjustments, as adjustments were found not to decrease but rather increase the staying rate for the acceptance ratio in (9).

Parameter estimation: Estimates of the parameters for the three MCMC schemes are given in Table 3. Estimates of the parameters for both QTL were close to the true values for all MCMC schemes, when the second QTL explained 28% of the total genetic variance and was linked with the first QTL (design I). When the second QTL explained only 7% of the total genetic variance (design II), its position was not well estimated, and MCMC schemes 2 and 3 gave better estimates than scheme 1. The higher estimate for α_2 from the first compared to the other MCMC schemes was probably due to the lower limit of $0.25\sigma_a$ used in MCMC scheme 1. Estimates of the position of the second QTL in designs III and IV are meaningless because most of the time the one-QTL model was sampled. Also, estimates of p_2 and α_2 are meaningless for MCMC schemes 2 and 3, when the two-QTL model was sampled with low frequency. However, for MCMC scheme 1, these parameters represent the frequency and the effect of an unlinked QTL. It should be noted that these results are based on a small number of replicates.

A few (six) chains were run with 100,000 cycles for MCMC scheme 2 and design I. Means of the parameter estimates from these runs were close to those of the shorter chains (20,000 cycles) given in Table 3. Effective sample size varied from 65 (α_1) to 693 (d_{Q2}). The highest posterior correlations among parameters were found between variance components (~ -0.9), which is in close agreement with previous single-QTL studies (THALLER and HOESCHELE 1996b; UIMARI *et al.* 1996a). The other correlations ranged from -0.3 to 0.3 without any noticeable pattern. Polygenic variance was overestimated while residual variance was underestimated, a finding in agreement with previous studies (*e.g.*, UIMARI *et al.*, 1996a). The marginal posterior distribution of polygenic variance is skewed to the right, causing an upward bias of the posterior mean estimator. This bias is a small sample bias resulting from the small number of halfsib families. Due to the strong, negative posterior correlation between polygenic and residual variances, the residual variance is underestimated. Posterior correlations between the variances and the QTL parameters are small to moderate, hence inaccuracy in the estimation of the variances should have little impact on the estimation of the other parameters.

When data simulated with the two-QTL model were

TABLE 3

Average parameter estimates of three MCMC schemes

	True value	Granddaughter design[a] I	II	III	IV	One-QTL model
p_1	0.50	0.54 (0.04)[b]	0.53 (0.09)	0.59 (0.02)	0.45 (0.03)	0.56 (0.04)
		0.54 (0.04)	0.58 (0.05)	0.56 (0.03)	0.46 (0.03)	
		0.52 (0.04)	0.53 (0.02)	0.54 (0.02)	0.48 (0.03)	
α_1	54.77	49.00 (3.15)	49.63 (2.91)	53.40 (0.86)	47.76 (3.24)	53.3 (1.55)
		51.69 (2.86)	50.56 (2.68)	51.54 (2.96)	49.51 (2.65)	
		53.23 (1.93)	49.87 (4.19)	53.68 (0.70)	49.28 (3.19)	
d_{Q1}	0.30	0.33 (0.019)	0.34 (0.035)	0.32 (0.006)	0.31 (0.015)	0.41 (0.034)[c]
		0.32 (0.004)	0.34 (0.035)	0.31 (0.010)	0.31 (0.014)	
		0.35 (0.031)	0.32 (0.004)	0.31 (0.005)	0.32 (0.010)	
p_2	0.50	0.48 (0.13)	0.36 (0.04)	0.45 (0.14)	0.39 (0.07)	
		0.46 (0.08)	0.50 (0.07)	0.58 (0.05)	0.46 (0.10)	
		0.70 (0.01)	0.73 (0.04)	0.61 (0.03)	0.47 (0.07)	
α_2	41.08/20.54/0.00/41.08	45.9 (1.23)	30.33 (2.11)	28.6 (1.26)	34.51 (4.51)	
		44.9 (0.99)	23.52 (3.41)	12.57 (6.44)	11.77 (3.03)	
		37.55 (2.36)	25.94 (3.46)	5.11 (0.51)	4.62 (1.45)	
d_{Q2}	0.70/0.70/NL[d]/NL	0.73 (0.03)	−0.81 (0.20)	−1.29 (0.17)	−1.26 (0.16)	
		0.69 (0.03)	0.53 (0.10)	0.55 (0.04)	0.59 (0.04)	
		0.66 (0.01)	0.64 (0.04)	0.65 (0.02)	0.69 (0.01)	
μ	0.00	−0.39 (3.47)	−0.98 (6.10)	4.57 (2.58)	−3.95 (2.55)	2.13 (2.46)
		−1.94 (2.87)	2.3 (1.55)	2.75 (1.16)	−1.53 (1.74)	
		(1.76)	4.15 (2.46)	−0.39 (0.18)	−0.73 (1.76)	
σ_u^2	164.1/322.3/375.0/164.1	355.7 (62.1)	463.6 (53.9)	437.8 (64.6)	440.6 (59.9)	505.9 (18.7)
		311.9 (49.1)	448.0 (52.6)	435.8 (4.39)	458.0 (53.2)	
		440.8 (53.8)	465.9 (43.8)	479.4 (80.3)	508.5 (68.3)	
σ_r^2	750.0	517.7 (83.2)	450.0 (60.6)	486.6 (118.6)	442.1 (122.3)	517.7 (83.2)
		525.4 (95.6)	621.7 (98.0)	662.9 (17.5)	726.8 (60.3)	
		414.5 (95.6)	589.6 (88.8)	553.7 (132.3)	637.5 (62.2)	

Empirical standard errors of means across replicates within design and method are shown in parentheses.

[a] Designs are defined in Table 1.

[b] The parameter estimates in rows 1, 2, and 3 are for MCMC schemes 1, 2, and 3, respectively.

[c] QTL positions conditional on marker intervals are as follows: interval 3, 0.33 [56%]; interval 4, 0.48 [38%]; interval, 0.61 [6%]. Marginal posterior probabilities of QTL location in an interval are in brackets.

[d] QTL not linked.

analyzed using the single-QTL model by applying the restriction $l \leq 1$ in MCMC scheme 1, parameter estimates were close to the true values for the QTL with the larger effect (Table 3). The QTL position was sampled near the true position of the larger QTL, as a ghost position in between the QTL, or in the interval of the smaller QTL. In most replicates the position was sampled exclusively within the interval of the larger QTL, with most of the variance contribution of the smaller QTL being incorporated into the polygenic variance.

DISCUSSION

Bayesian linkage analysis allowing for linked QTL and implemented with three different MCMC schemes was investigated. MCMC scheme 1 was an extension of previous work by UIMARI *et al.* (1996), where a linkage indicator variable was incorporated into the analysis. MCMC scheme 2 was based on the work of KUO and MALLICK (1994) on variable selection in (generalized) linear models, and scheme 3 was an application of reversible jump MCMC (GREEN 1995).

All methods performed well for the simulated granddaughter designs. Reversible jump MCMC was very sensitive to the way in which new variables were generated when moving from the one- to the two-QTL model and back. The most successful way to generate new variables (parameters and genotypes at the second QTL) was to set the values for all variables pertaining to the first QTL equal to the current values under the single-QTL model, and to generate new values for the second QTL. The only exception was that gene frequencies were generated using the old value as a basis for new gene frequencies at both QTL and an average of the gene frequencies at both QTL when returning from the two- to the one-QTL model (see METHODS). Furthermore, the generating distribution of $\mathbf{G}_2$ was crucial for MCMC scheme 3 to perform well. The optimum [in the sense of acceptance rate based on (10)] generating distribution for the additional QTL_2 genotypes, when moving

from the one- to the two-QTL, model is $P(\mathbf{G}_2|\boldsymbol{\theta}_2, \mathbf{MG}_1, y)$, because $\mathbf{G}_2$ is sampled conditional on all other genotypes and on phenotypes. However, the resulting scheme would be computationally unfeasible, because the normalizing constants for $P(\mathbf{MG}_2|\boldsymbol{\theta}_2)$ and $P(\mathbf{MG}_1|\boldsymbol{\theta}_1)$ are unknown and require summation over all possible genotypes of all animals in the pedigree. Therefore, the $\mathbf{G}_2$ were generated using the conditional probabilities $P(\mathbf{G}_2|\boldsymbol{\theta}_2, \mathbf{MG}_1)$. With $d_{21} = d_{11}$, the ratio of $P(\mathbf{MG}_1|\boldsymbol{\theta}_2)$ to $P(\mathbf{MG}_1|\boldsymbol{\theta}_1)$ in (10) reduced to a term dependent only on p_{11} and p_{21}.

In MCMC scheme 3, the second QTL was always generated to the right of the first QTL by setting $d_{22} = d_{12} + w_d$ with w_d positive. However, if the true position of the second QTL is to the left of the first QTL, this procedure may fail. Therefore, first the single-QTL analysis should be performed using different starting positions for the QTL on the chromosome, and subsequently, several two-QTL analyses may be run, one with $d_{22} = d_{12} + w_d$ and another with $d_{22} = d_{12} - w_d$.

The three MCMC schemes differed in their requirements for the parameter spaces of d_{Q1} and d_{Q2}. In MCMC scheme 1, including the flanks was necessary to enable moves from the single-QTL model (with the second QTL unlinked) to the two-QTL model (with the second QTL linked). For MCMC schemes 2 and 3, flanks did not need to be considered. If there is evidence that a QTL may be located in either of the flanks, the parameter space can be expanded to include the flanks. Expanding the parameter spaces of d_{Q1} and d_{Q2} to include flanks increases computing time, and the QTL position may occasionally be sampled in the flanks even if the true position is within the linkage group, as can be seen in Table 3 (design II).

THALLER and HOESCHELE (1996a) and SATAGOPAN *et al.* (1996) performed QTL model selection based on Bayes factors, which were estimated using different MCMC algorithms. THALLER and HOESCHELE (1996a) found that MCMC sampling with model indicators gave much more stable results than MCMC estimates of Bayes factor. SATAGOPAN *et al.* (1996) stabilized Bayes factor estimation using normal and multivariate t weighting densities. In general, estimation of Bayes factors via MCMC is not a trivial task.

In conclusion, this paper demonstrates that it is feasible to fit linked QTL simultaneously using Bayesian analysis with MCMC algorithms. The Bayesian analysis provides estimates of all genetic parameters and can fit alternative QTL models (*e.g.*, a normal-effects QTL), rather than a biallelic QTL (HOESCHELE *et al.* 1996). We recommend pursuing an analysis with two linked QTL after results from the single-QTL study indicate QTL presence on the chromosome of interest. In this work, also compared were three MCMC schemes for QTL-model selection, which performed well, yielding similar results for the simulated designs. The schemes differed, however, in their CPU time requirements and ease of implementation, with MCMC scheme 2 outperforming the others in these two criteria.

This project was funded by the National Science Foundation (award no. BIR-9596247). Technical support provided by D. SATTERTHWAITE from the Digital Western Region is gratefully acknowledged. This work was performed using the resources of the Cornell Theory Center, which receives funding from the National Science Foundation, New York State, IBM Corporation, and others.

LITERATURE CITED

GEORGES, M., D. NIELSEN, M. MACKINNON, A. MISHRA, R. OKIMOTO *et al.*, 1995 Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. Genetics **136:** 907–920.

GEYER, C. J., 1992 Practical Markov chain Monte Carlo (with discussion). Stat. Sci. **7:** 467–511.

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

GRIGNOLA, F. E., I. HOESCHELE and K. MEYER, 1994 Empirical best linear unbiased prediction to map QTL. Proceedings of the 5th World Congress on Genetically Applied Livestock Production, Vol. **21:** 245–258.

GRIGNOLA, F. E., I. HOESCHELE and B. TIER, 1996a Residual maximum likelihood to map quantitative trait loci: methodology. Genet. Sel. Evol. **28:** 479–490.

GRIGNOLA, F. E., I. HOESCHELE, Q. ZHANG, and G. THALLER, 1996b Residual maximum likelihood to map quantitative trait loci: a simulation study. Genet. Sel. Evol. **28:** 491–504.

GRIGNOLA, F. E., Q. ZHANG and I. HOESCHELE, 1997 Mapping linked quantitative trait loci using residual maximum likelihood. Genet. Sel. Evol. (in press).

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

HOESCHELE, I., and P. M. VANRADEN, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. **85:** 953–960.

HOESCHELE, I., and P. M. VANRADEN, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85:** 946–952.

HOESCHELE, I., P. UIMARI, F. GRIGNOLA, Q. ZHANG and K. GAGE, 1996 Statistical mapping of polygene loci in livestock. Proc. Int. Biometrics Soc. (in press).

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 trait into multiple loci via interval mapping. Genetics **136:** 1447–1455.

KUO, L., and B. MALLICK, 1994 Variable selection for regression models. Technical Report, Department of Statistics, University of Connecticut, Stors, CT.

MACKINNON, M. J., and J. I. WELLER, 1995 Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. Genetics **141:** 755–770.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the location and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85:** 480–488.

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.

SORENSEN, D. A., S. ANDERSEN, D. GIANOLA and I. KORSGAARD, 1995 Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. **27:** 229–249.

SPELMAN, R. J., W. COPPIETERS, L. KARIM, J. A. M. VAN ARENDONK, and H. BOVENHUIS, 1996 Quantitative trait loci analysis for five milk production traits on chromosome *six* in the Dutch Holstein-Friesian population. Genetics **144:** 1799–1808.

TANNER, M. A., 1993 *Tools for Statistical Inference.* Springer-Verlag, Berlin.

THALLER, G., and I. HOESCHELE, 1996a A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: I. Methodology. Theor. Appl. Genet. **93:** 1161–1166.

THALLER, G., and I. HOESCHELE, 1996b A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci: II. A simulation study. Theor. Appl. Genet. **93:** 1167–1174.

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

UIMARI, P., Q. ZHANG, F. E. GRIGNOLA, I. HOESCHELE and G. THALLER, 1996b Analysis of QTL workshop I granddaughter design data using least-squares, residual maximum likelihood, and Bayesian methods. J. Quant. Trait Loci **2:** 7.

VANRADEN, P. M., and G. R. WIGGANS, 1991 Derivation, calculation and use of national animal model information. J. Dairy Sci. **74:** 2737–2746.

WELLER, J. I., 1986 Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics **42:** 627–640.

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

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

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: P. D. KEIGHTLEY