A Mixture Model Approach to the Mapping of Quantitative Trait Loci in Complex Populations With an Application to Multiple Cattle Families

Ritsert C. Jansen,* David L. Johnson† and Johan A. M. Van Arendonk‡

**Centre for Biometry Wageningen, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), Wageningen NL-6700 AA, The Netherlands, †Livestock Improvement Corporation, Hamilton, New Zealand and ‡Wageningen Institute of Animal Sciences, Wageningen NL-6700 AH, The Netherlands*

Manuscript received March 3, 1997
Accepted for publication September 17, 1997

ABSTRACT

A mixture model approach is presented for the mapping of one or more quantitative trait loci (QTLs) in complex populations. In order to exploit the full power of complete linkage maps the simultaneous likelihood of phenotype and a multilocus (all markers and putative QTLs) genotype is computed. Maximum likelihood estimation in our mixture models is implemented via an Expectation-Maximization algorithm: exact, stochastic or Monte Carlo EM by using a simple and flexible Gibbs sampler. Parameters include allele frequencies of markers and QTLs, discrete or normal effects of biallelic or multiallelic QTLs, and homogeneous or heterogeneous residual variances. As an illustration a dairy cattle data set consisting of twenty half-sib families has been reanalyzed. We discuss the potential which our and other approaches have for realistic multiple-QTL analyses in complex populations.

THE number of genes identified in humans, plants and animals has increased notably during the last decade. The largest increase in number of identified genes has occurred for qualitative single gene traits. In contrast, progress in mapping quantitative trait loci (QTLs) has been slow, except for species for which inbred lines are available. Human pedigrees are often complex and small, and analysis of pedigree data requires sophisticated statistical techniques, the development of which has become a bottleneck in QTL mapping (Guo and Thompson 1992). In outbred populations of animals or plants, this bottleneck is also real but less severe, because of the high reproduction rate and the option to design experiments. The main problems to be dealt with in the analysis of complex populations can be summarized as follows:

1. The number of alleles and the allele frequencies in the (base) population are unknown for QTLs as well as for marker loci;
2. If a parent is homozygous at a marker locus, it is impossible to trace which allele from a pair of parental homologous chromosomes has been transmitted to a descendant;
3. When two parents are heterozygous and carry the same alleles at a marker locus, the parental origin of alleles of a heterozygous descendant cannot be determined;
4. The genotype at a QTL cannot be observed, and we therefore do not know which parents are heterozygous for the QTL;
5. Markers may have been selectively genotyped for only a subset of the population;
6. Linkage phases between markers and between markers and QTLs may be unknown.

Markers are used to follow the inheritance of genome segments from parent to offspring (the pattern of identity-by-descent or IBD of alleles). If the IBD pattern at a certain marker (or QTL) locus is unknown, then neighboring markers may be informative, *i.e.*, linked markers may indicate the likely IBD pattern at the locus under study (*cf.* Haley *et al.* 1994; Jansen 1996a; Knott *et al.* 1996). Phenotype also contains information on genotype, but this information is often ignored to simplify computations (see Modeling QTLs below). In general, all markers and phenotype should be used simultaneously so as to recover at each map position as much information as possible.

One can assume fixed or random effects models or mixed models for the relation between phenotype and "known" genotype. As stated above the information about genotype can be incomplete for various reasons. We therefore enter the area of so-called *mixture* models, where the possible genotypic configurations are the components of the mixture. An important monograph on mixture models is written by Titterington *et al.* (1985). A popular statistical algorithm for handling mixture problems is the expectation-maximization (EM) algorithm. It is an iterative approach, it is relatively easy to program, and it produces maximum likelihood estimates and also empirical Bayesian *a posteriori*

Corresponding author: Ritsert C. Jansen, Centre for Biometry Wageningen, Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, NL-6700 AA Wageningen, The Netherlands. E-mail: r.c.jansen@cpro.dlo.nl

estimates (Dempster *et al.* 1977). Jansen (1992) and Jansen and Stam (1994) described a general and flexible EM algorithm for recovering information about a multilocus genotype in populations obtained from crosses between inbred lines. In a next article, Jansen (1996a) developed a Monte Carlo method for multilocus analysis in a simple outbred cross between two plant cultivars. Here, we will make our EM approach applicable for complex population structures in which additional dependencies between individuals may exist. We propose a stochastic EM algorithm and a Monte Carlo EM algorithm in which a Markov chain of possible genotypic configurations is generated via the Gibbs sampler. With large progeny groups, however, the chain may show slow changing of genotype states or can even remain stuck in a certain "subspace" (Janss *et al.* 1995). To avoid these problems, we introduce a simple and flexible scheme, based on different descriptions of the genotype of founders and non-founders of the population. Furthermore we demonstrate how our EM approach can be used to fit models for single or multiple QTLs with fixed or random effects. Recently, data for paternal half-sib families of dairy cattle have been adopted for comparison of analytical approaches developed in the animal breeding community (Bovenhuis *et al.* 1998). As an illustration, we will analyze these data. We postulate several mixture models and fit them by using our Monte Carlo EM algorithm.

There is a growing need for sophisticated analytical tools to genetically dissect multigenic traits in complex populations. The theory on QTL mapping is developing very fast, and we therefore start with a section in which we briefly review and classify the various recent developments (Jansen 1996a; Knott *et al.* 1996; Satagopan *et al.* 1996; Satagopan and Yandell 1997; Thaller and Hoeschele 1996; Uimari *et al.* 1996a; Xu 1996; Grignola *et al.* 1997). In the discussion, we focus on the potential that our and other approaches have for realistic multiple-QTL analyses in complex populations.

MODELING QTLS

Recent analytical approaches can be classified according to three criteria: modeling the full mixture of possible phenotype-genotype combinations or not, assuming fixed or random QTL effects, and adopting the (restricted) maximum likelihood or Bayesian approach.

Consider an N-member population on which trait values and marker scores are observed. Let y_i denote the ith individual's trait value and let g_i denote its combined genotype at all marker loci and one or more putative QTLs. We write $\mathbf{y} = (y_1 \ldots y_N)'$ and $\mathbf{g} = (g_1 \ldots g_N)'$. The population may consist of multiple generations and marker and trait data need not be observed for all individuals. For a given genotypic constitution $\mathbf{g}$ on the population, one can model the relation between $\mathbf{y}$ and the "known" genotype $\mathbf{g}$ by assuming a model with discrete and fixed QTL-effects and normally distributed error. The distribution $f(\mathbf{y}|\mathbf{g})$ is then a multivariate normal distribution and the mean, $\mu(\mathbf{y}|\mathbf{g})$, is modeled in terms of genetic parameters (θ) such as additivity and dominance of (multiple) QTL effects. On the other hand, one may prefer a random model in case of multiallelic QTLs. It is then often assumed that QTL effects are independent realizations from a normal distribution, which represents the distribution over many alleles in the base population. Now $f(\mathbf{y}|\mathbf{g})$ is a multivariate normal distribution with variance-covariance matrix $v(\mathbf{y}|\mathbf{g})$ expressed in terms of genetic parameters. Multiple QTLs, random or fixed family (polygenic) effects and additional (experimental) effects such as QTL-QTL interaction or QTL-environment interaction can be included resulting in so-called mixed models. Also other types of distribution can be assumed in addition to the commonly assumed normal distribution.

The genotype $\mathbf{g}$ includes full multilocus information about alleles and their IBD pattern, but unfortunately this information can be observed only partially. For each possible genotypic configuration $\mathbf{g}$ on the population (that is, a configuration which is consistent with the observed marker data), we can calculate a scalar probability $P(\mathbf{g})$ of occurrence. $P(\mathbf{g})$ is a function of (known or unknown) recombination and allele frequencies. The exact methods use *mixture* distributions to model the full relation between phenotype and possible genotypes: $f(\mathbf{y}) = \sum P(\mathbf{g})f(\mathbf{y}|\mathbf{g})$, where summation is over all possible genotypes $\mathbf{g}$. Jansen (1992, 1994, 1996a), Thaller and Hoeschele (1996), Uimari *et al.* (1996a), Xu (1996) and Satagopan *et al.* (1996, 1997) consider mixture models for QTL mapping. Jansen uses maximum likelihood for QTLs with discrete effects, Xu uses maximum likelihood for QTLs with normal effects, Thaller and Hoeschele, Uimari *et al.* and Satagopan *et al.* use Bayesian methods for QTLs with discrete effects.

An exact mixture analysis can be computationally demanding, especially if the number of possible genotypes $\mathbf{g}$ is huge. Approximate *expectation* methods first calculate an expected trait mean $\mu(\mathbf{y}) = \sum P(\mathbf{g})\mu(\mathbf{y}|\mathbf{g})$ if a discrete QTL-effects model is used, where summation is again over possible genotypes. In the normal QTL-effects models, an expected variance-covariance matrix $v(\mathbf{y}) = \sum P(\mathbf{g})v(\mathbf{y}|\mathbf{g})$ is calculated. Next it is assumed that $f(\mathbf{y})$ is normally distributed with mean $\mu(\mathbf{y})$ or variance-covariance matrix $v(\mathbf{y})$. Knott *et al.* (1996) use the expectation method for discrete QTL-effects models and Grignola *et al.* (1997) use it for normal-QTL effects models. Zeng (1994) assumes discrete QTL-effects and uses a combination of the mixture and the expectation method: the expectation method to deal with missing marker (cofactor) data and the mixture method for the putative QTL.

Phenotypes contain information on QTL genotypes. Moreover, if markers are linked to QTLs, phenotypes also contain information about incomplete marker genotypes. The exact methods take into account marker plus phenotype information to retrieve information. In contrast, in the approximate expectation methods, genotype probabilities are calculated on the basis of marker data only and this calculation is done only once, namely before QTL analysis.

MAXIMUM LIKELIHOOD IN MIXTURE MODELS VIA EM ALGORITHMS

Let θ denote the vector of all parameters for fixed and random model terms and for recombination and allele frequencies. In QTL mapping, the genetic map is usually assumed to be known (*i.e.*, the recombination frequencies are known). Jansen (1992, 1996a) developed formulae for simple two-generation designs with $f_\theta(\mathbf{y},\mathbf{h}) = \Pi\, f_\theta(y_i,h_i)$, where the product is over the members of the population and $\mathbf{h} = (h_1 \ldots h_N)'$ denotes the observed marker data. Here we consider a general population structure in which case additional dependency of individuals can exist so that $f_\theta(\mathbf{y},\mathbf{h})$ can not be expressed as a simple product of member likelihoods. To simplify notation, we will write f_θ again as f (and P_θ as P).

The simultaneous likelihood $\mathcal{L}(\theta)$ of all observed trait and marker data is a mixture likelihood with the possible genotypes as components

$$\mathcal{L}(\theta) = f(\mathbf{y},\mathbf{h}) = \sum_g P(\mathbf{g}) f(\mathbf{y},\mathbf{h}|\mathbf{g})$$

where $f(\mathbf{y},\mathbf{h}|\mathbf{g}) = f(\mathbf{y}|\mathbf{g})$ if $\mathbf{h}$ is consistent with $\mathbf{g}$ and $f(\mathbf{y},\mathbf{h}|\mathbf{g}) = 0$ otherwise; $f(\mathbf{y}|\mathbf{g})$ may refer to a fixed, random or mixed model. Parameter estimation can be carried out by maximum likelihood. The likelihood equations are

$$\begin{aligned} 0 &= \frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta) = \frac{1}{f(\mathbf{y},\mathbf{h})}\frac{\partial}{\partial\theta}\sum_g P(\mathbf{g}) f(\mathbf{y}|\mathbf{g}) \\ &= \sum_g \frac{f(\mathbf{y}|\mathbf{g})\cdot P(\mathbf{g})}{f(\mathbf{y},\mathbf{h})}\frac{\partial}{\partial\theta}\ln\left(P(\mathbf{g}) f(\mathbf{y}|\mathbf{g})\right) \\ &= \sum_g P(\mathbf{g}|\mathbf{y},\mathbf{h})\frac{\partial}{\partial\theta}\ln P(\mathbf{g}) + \sum_g P(\mathbf{g}|\mathbf{y},\mathbf{h})\frac{\partial}{\partial\theta}\ln f(\mathbf{y}|\mathbf{g}), \end{aligned}$$

where summation is over possible genotypes $\mathbf{g}$ consistent with $\mathbf{h}$.

Exact EM: The likelihood equations can be solved by applying an EM algorithm (Jansen 1992; Jansen and Stam 1994). Each iteration consists of two steps. First, in the so-called E-step, the conditional probability $P(\mathbf{g}|\mathbf{y},\mathbf{h})$ is evaluated for all possible genotypes $\mathbf{g}$, given the current parameter estimates and given the observed incomplete information $\mathbf{h}$ on the genotype (using Baye's theorem). Next, in the so-called M-step, the likelihood equations are solved by fixing the weights $P(\mathbf{g}|\mathbf{y},\mathbf{h})$, which gives updated parameter estimates. The likelihood equation can be split into two terms: the first term refers to the genetic linkage between loci, the second term to the phenotype-complete genotype relation. Each term can be recognized as a likelihood equation for non-mixture problems that can be solved with standard statistical routines or packages for (weighted) regression or variance component models (see also Jansen 1992).

Stochastic EM: In each cycle of the EM algorithm, the likelihood equation can be estimated by using a single Monte Carlo realization

$$\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta) \triangleq \frac{\partial}{\partial\theta}\ln P(\mathbf{g}^{(j)}) + \frac{\partial}{\partial\theta}\ln f(\mathbf{y}|\mathbf{g}^{(j)}),$$

where in the jth EM cycle a single complete genotype $\mathbf{g}^{(j)}$ is generated given $\mathbf{y}$, $\mathbf{h}$ and the current parameter estimates [*i.e.*, by using the distribution $P(\mathbf{g}^{(j)}|\mathbf{y},\mathbf{h})$]. This expression can be treated as a standard likelihood equation, and it can be solved with standard statistical software. The "posterior" distribution of parameter estimates obtained over many EM cycles and after a suitable burn-in period is approximately centered at the maximum likelihood estimate and the mean of the distribution can be used as an ML estimate (Celeux and Diebolt 1985). This "posterior" distribution can also be plotted for parameters of interest. Also a preliminary short stage of stochastic EM can be run that yields good starting values for Monte Carlo EM.

Monte Carlo EM: In each cycle of the EM algorithm, the likelihood equation can be estimated using a number (M) of Monte Carlo realizations

$$\frac{\partial}{\partial\theta}\ln\mathcal{L}(\theta) \triangleq \frac{1}{M}\sum_j\frac{\partial}{\partial\theta}\ln P(\mathbf{g}^{(j)}) + \frac{1}{M}\sum_j\frac{\partial}{\partial\theta}\ln f(\mathbf{y}|\mathbf{g}^{(j)}),$$

where in the jth Monte Carlo sample, complete genotypes $\mathbf{g}^{(j)}$ are generated given $\mathbf{y}$, $\mathbf{h}$ and the current parameter estimates [*i.e.*, by using the distribution $P(\mathbf{g}^{(j)}|\mathbf{y},\mathbf{h})$]. This expression can be treated as a likelihood equation for standard non-mixture problems of $N \times M$ observations. For sufficiently large genotype samples, Monte Carlo EM will inherit the properties of exact EM.

The Monte Carlo samples can also be used for likelihood-ratio estimation in the final EM step. The likelihood ratio is estimated as

$$\frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \triangleq \frac{1}{M}\sum_j\frac{f_{\theta_1}(\mathbf{y}|\mathbf{g}^{(j)})\cdot P_{\theta_1}(\mathbf{g}^{(j)})}{f_{\theta_0}(\mathbf{y}|\mathbf{g}^{(j)})\cdot P_{\theta_0}(\mathbf{g}^{(j)})}$$

where complete genotypes $\mathbf{g}^{(j)}$ are sampled given $\mathbf{y}$, $\mathbf{h}$ and θ_0 (Jansen 1996a). Estimation of the likelihood ratio is most effective when θ_1 and θ_0 are close. To increase efficiency of estimation, θ_1 can be related to θ_0 via one (or more) intermediate models spanning the range between θ_1 and θ_0. For instance, calculate $\mathcal{L}(\theta_1)/\mathcal{L}(\theta_0) = \mathcal{L}(\theta_1)/\mathcal{L}(\theta_2) \times \mathcal{L}(\theta_2)/\mathcal{L}(\theta_0)$. We used this approach to compare models with and without a putative QTL at a certain map position (see our dairy cattle ap-

plication in the next section). Geyer (1993) described an alternative approach to increase efficiency of estimation. He noted that the sampling distribution under θ_0 need not be in the model and suggested to sample from finite mixtures of distributions in the model of interest

$$P_{\text{mix}}(\mathbf{g}|\mathbf{y},\mathbf{h}) = \sum_{k=1}^{K} \pi_k(\mathbf{y},\mathbf{h}) P_{\theta_k}(\mathbf{g}|\mathbf{y},\mathbf{h})$$

where π_k are the mixing proportions. Again, the θ_ks can be chosen to span the range between the two parameters of interest. If we have a computer code to sample from $P_\theta(\mathbf{y};|\mathbf{g})$ for any θ, then we can also sample from the mixture: run the program for each of the θ_ks and combine the samples to form a sample of the mixture (Geyer 1993). This method can be used to efficiently calculate the likelihood ratios for pairs $(\theta_k,\theta_{k'})$ that are components of the sampling mixture.

Gibbs sampler: Unfortunately there may be no direct feasible way to generate the Monte Carlo samples because of the huge amount of possible genotypic states **g** in complex populations with many loci. A solution to this problem is to utilize the Gibbs sampler (*cf.* Guo and Thompson 1992; Janss *et al.* 1995). Jansen (1996a) considered a situation with multiple loci in a simple outbred cross between two plant cultivars and described a simple Gibbs sampler in which the offspring genotype is updated in a stepwise manner for only a single locus and a single individual at a time, while taking "for granted" the remaining part of the genotype. In this way, the number of possible genotypic states is small and sampling can easily be done (of course states have to be consistent with observed marker data).

In this article, we deal with general population structures in which case there are serious problems if we specify genotype in terms of allelic constitution only. For instance, in a half-sib analysis, a parent (sire) with current QTL-state *a*/*b* produces offspring carrying allele *a* or *b*, and a change of the parent's genotype to *a*/*a* or *b*/*b* is practically prohibited if the male has many offspring (Janss *et al.* 1995). To avoid this kind of problem, we now introduce different descriptions of the genotype of founders and non-founders in the population. We specify the genotypic state of any founder by the alleles at each of its two homologues. We express the state of any non-founder by IBD values indicating parental origin of its alleles. In the above half-sib example, an offspring of the sire *a*/*b* inherits the allele of either the first homologue of its parent (IBD = *1*) or the second (IBD = *2*). In a marker-QTL-marker situation, a sire may have current genotype *aaa*/*bbb*, *i.e.*, *aaa* on homologue *1* and *bbb* on homologue *2*. An offspring of this sire may have current genotype *aab*/*ccc*, with *c* alleles originating from the dam, and we will write this as *112*/*ccc* by using an IBD indicator rather than the actual allele type (*a* or *b*). The same can be done in other types of populations (*e.g.*, with more generations).

We will now briefly consider the three steps used in our Gibbs sampler for updating the genotype of founders (their QTL states and linkages phases between loci) and the genotype of non-founders (via IBD pattern).

Step 1: To take all possible QTL states of founders into account, one can sample allelic configurations QTL by QTL. For instance, consider a change of marker-QTL-marker genotype *aaa*/*bbb* of a sire to *aaa*/*bab*, *aba*/*bab* or *aba*/*bbb* without changing the "known" IBD pattern (*e.g.*, offspring *112*/*ccc*). One can calculate the corresponding conditional probabilities given "known" IBD pattern and given phenotypes, and next sample one of the four possible states.

Step 2: To take all possible linkages phases in the genotypes of founders into account, one can sample linkage phases interval by interval and founder by founder. For instance, consider a change of the linkage phase between the proximal and distal part of the chromosome at a certain interval for a certain founder. In the case of a phase switch, the distal part of its homologue *1* is attached to the proximal part of homologue *2* (*i.e.*, becomes part of the new homologue *2*) and vice versa. IBD values are used for the description of genotypes of non-founders, and, in case of a phase switch, one should change the IBD values at the distal part of the chromosome accordingly (*1* becomes *2* and vice versa). One can calculate the conditional probabilities for the two options "phase switch" and "no phase switch" and sample one of them.

Step 3: To generate genotypes of non-founders, one can sample a new IBD pattern given "known" genotype of founders. This can be done individual by individual and locus by locus. If we update the IBD at a certain marker locus, then the two flanking loci (with "known" IBD!) are fully informative and no other loci are needed (no matter whether the two flanking loci are markers and/or QTLs). If we update the IBD at a putative QTL, the update step also depends on the expected phenotype (*i.e.*, 'fitted values') given "known" genotype at other putative QTLs. As stated above, only genotypic states consistent with observed marker data are allowed.

Our notation has two important advantages. First, one can now change the genotype of founders independently of the IBD pattern and vice versa, therewith avoiding the problems discussed by Janss *et al.* (1995). Second, an IBD pattern is generated at all loci and therefore IBD is "known" in the Gibbs sampling process even if a parent is homozygous for a marker or if both parents and offspring are heterozygous carrying the same alleles at a certain marker. Computer implementation is then rather straightforward (using three-locus information instead of multilocus information), not only for the single-QTL models, but also for the more powerful multiple-QTL models. However if two (or more) loci are closely linked, it may be more efficient to

update the genotype at these loci together ("in blocks") to reduce auto-correlation in the Gibbs sampler, at the cost of more computer programming (Janss *et al.* 1995). In our application, we blocked marker and QTL when fitting a QTL close to or on top of the marker.

APPLICATION

In this section, our focus will be on the dairy cattle experiment (Bovenhuis *et al.* 1998). Data were available for 20 sires and their families of half-sib sons. Nine molecular markers on chromosome *6* were scored in sires and sons, and marker alleles were encoded within families (code *a* and *b* for alleles of paternal type and code *c* for maternal alleles that differ from the sire alleles). Protein percentage data were available for sons only (obtained as averages for milk production data of daughters of the sons). We refer to Spelman *et al.* (1996) for more information on the experiment. It should be noted that we analyzed the data as they were released to the animal breeding community. Spelman *et al.* (1996) analyzed slightly different data (corrected for some suspicious data).

Many different models can be formulated by making (combinations of) assumptions about the number of genes involved (monogenic, oligogenic or polygenic inheritance), about the number of alleles per QTL (biallelic or multiallelic), about the allelic effects (fixed or random model terms), about interaction effects (QTL-family, QTL-QTL), about residual variance (homogeneity of residual variance over families or heterogeneity), about the effect of the dam (ignored or included in the model), or about linkage phases between loci (unknown, or fixed at a likely configuration). As an illustration, some of the proposed models will be applied to the dairy cattle data (see Table 1 for a description), and our results are compared with those published by Spelman *et al.* (1996).

We have to do with a "serious" mixture problem, since there are several sources of missing information: each marker isuninformative (homozygous) for several families; in a number of cases, it cannot be assessed which of the two marker alleles of an offspring originates from the sire; all QTL scores and some marker scores of offspring are missing; all marker scores of dams are missing; marker and QTL allele frequencies are unknown; and linkage phases of all loci are unknown. Clearly the total number of possible configurations **g** consistent with observed marker data is huge, making an exact analysis demanding. We assume that the recombination frequencies are known (fixed genetic map). Let y_{ij} be the trait value of the jth son of the ith sire.

Model I: Spelman *et al.* (1996) used the expectation method developed by Knott *et al.* (1994). To simplify the computational work, the most likely linkage phase was determined and taken for each sire, and when different phases were equally likely, one was chosen at random. Effects of dams were ignored. Information on marker allele frequencies was not used. In the approximate method, $P(\mathbf{g})$ is calculated on the basis of marker data only and this calculation is done only once, namely before QTL analysis. At the map position under study their *expectation* QTL model reads

$$y_{ij} = \mu + s_i + a_{i1} \times \text{Prob}_{ij1} + a_{i2} \times \text{Prob}_{ij2} + e_{ij}$$

where s_i is the (polygenic) fixed effect of the ith sire,I a_{i1} is the fixed effect of the QTL allele at the first homologue of the ith sire at the map position under study, Prob_{ij1} is the (previously calculated) probability that the jth son of the ith sire has received this allele given the observed marker data, a_{ij2} and Prob_{ij2} are defined analogously, and e_{ij} is a random normally distributed residual with homogeneous residual variance over families. Note that this model can also be reparameterized in terms of the allele-substitution effects $(a_{i2} - a_{i1}) \times \text{Prob}_{ij2}$ of the sires, and in this sense, the model is a multiallelic-QTL model.

Model II: A mixture model for a multiallelic QTL is considered (similar to Spelman's model I). At the map position under study the model for phenotype given "known" genotype reads

$$y_{ij} = \mu + s_i + a_{i1} \times q_{ij1} + a_{i2} \times q_{ij2} + e_{ij}$$

where $q_{ij1} = 1 - q_{ij2}$ is an IBD indicator: $q_{ij1} = 1$ if the son has inherited the QTL allele of its sire's first homo-

TABLE 1

Outline of the models used in the cattle application

Model	Approach used to deal with missing marker and QTL information	QTL	Residual variance over families	QTL contribution of dam
I	Expectation	Multiallelic	Homogeneous	Ignored
II	Mixture	Multiallelic	Homogeneous	Ignored
III	Mixture	Multiallelic	Homogeneous	Ignored
IV	Mixture	Biallelic	Homogeneous	Included

logue at the map position under study, otherwise $q_{ij1} = 0$. It was assumed that residual variance was homogeneous over the families. Parameters now include (known) recombination frequencies and (unknown) marker allele frequencies within families (markers were encoded within families). The mixture distribution is obtained by summing over all possible genotypes. In contrast to Spelman *et al.* (1996), we here take all possible linkage phases into account and we also use information on marker allele frequencies and information on phenotype plus marker observations in the calculation of genotype probabilities Prob_{ij1} and Prob_{ij2} (at each EM iteration).

Model III: As model II, but now we assume heterogeneous residual variance over families (that is, a separate variance parameter per family was used). As an example of fitting multiple QTLs, we also extended model III and fitted two QTLs simultaneously.

Model IV: Models I–III are multiallelic-QTL models, and dam contributions were ignored. Now we consider a biallelic QTL and we also include the (unobserved) dam contributions. The estimate of the polygenic effect of a sire is affected by the QTL genotype being considered for that sire (Knott *et al.* 1992). To deal with that we use μ_{QQ}, μ_{Qq} and μ_{qq} instead of μ. If the ith sire has genotype QQ, the model for phenotype given "known" genotype reads

$$y_{ij} = \mu_{QQ} + s_i + a_1 \times q_{ij1} + a_2 \times q_{ij2} + a_1 \times d_{ij1} + a_2 \times d_{ij2} + e_{ij}$$

where a_1 and a_2 are the effects of the two QTL alleles at the map position under study, and the additional variable $d_{ij1} = 1 - d_{ij2}$ is the indicator for the dam contribution: $d_{ij1} = 1$ if the ith son of the jth sire has inherited allele a_1 from its dam, otherwise $d_{ij1} = 0$. The term μ_{QQ} is replaced by μ_{Qq} if the sire's genotype is Qq and by μ_{qq} if it is qq. We assumed homogeneity of residual variance over families. The mixture distribution is again obtained by summing over all possible genotypes and the parameters include (known) recombination frequencies, (unknown) marker allele frequencies within families and (unknown) QTL allele frequencies in the base population. Marker scores of dams are missing. In our analysis, phenotypes of the sons and marker scores of the sires and sons are used to recover as much information on dam contribution as possible.

QTL likelihoods: Figure 1 shows the four QTL likelihood plots for models I–IV. At each map position, the value of the test statistic is plotted for the comparison of the two models with and without a QTL at the given map position. The solid curve of model I is obtained by converting the F values reported by Spelman *et al.* (1996) into likelihood ratio values (likelihood ratio test $\approx pF$ where $p = 20$ is the d.f. for the test, see Haley and Knott 1992). Like in model I, the tests for a multiallelic QTL in models II and III have 20 d.f. In contrast, the test for the biallelic QTL in model IV has $\sim$2 d.f.: one for the QTL effect (assuming additivity) and one for the frequency of the QTL allele.

To obtain empirical critical values, Spelman *et al.* (1996) analyzed original marker data with randomly permuted trait values over many permutations. A QTL for protein percentage was detected near marker two with a single-test significance value of 0.01% and an experiment-wise significance value of 1%. The evidence was mainly coming from two families (families 1 and 16).

Parameter estimation for models II–IV was implemented via Monte Carlo EM, using 1000 Gibbs cycles per EM iteration and using the genotypic state in each tenth cycle as a Monte Carlo realization. In the final EM iteration, QTL likelihood was evaluated at marker positions by running 25,000 Gibbs cycles, using every 20th cycle for Monte Carlo evaluation of the likelihood ratio, and using $\leq$20 intermediate models spanning the range between the model with the QTL and that without a QTL. Running 25,000 Gibbs cycles for model III took $\sim$15 min CPU time on a DEC AlphaServer 2100 at 275 MHz.

Figure 1 clearly shows that all curves peak in the first marker bracket, that is, there is similarity of the four QTL likelihood curves in the region between markers one and four. In contrast, large differences between the curves appear in the region between markers six and nine. We will propose several explanations for these dissimilarities below.

Comparison of models I and II: Note that these models assume homogeneity of residual variance, although variances differ significantly between families (not

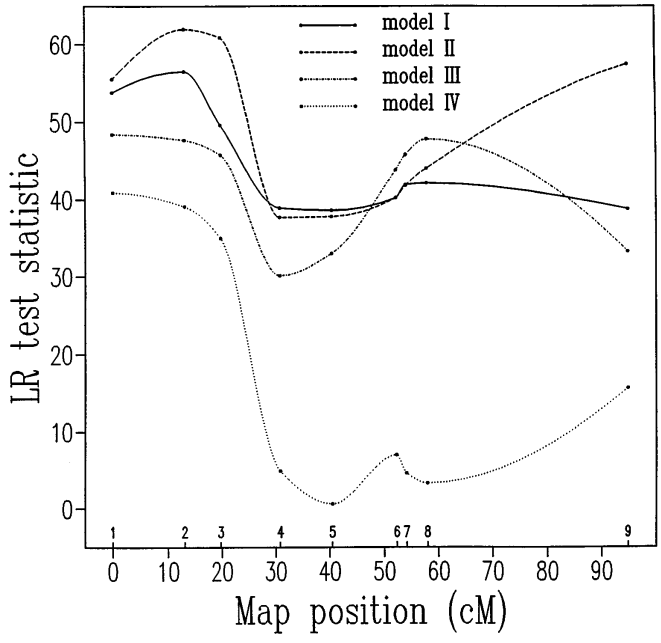

Figure 1.—QTL likelihood under various models for protein percentage in the dairy cattle experiment (see Table 1 for a description of the models).

shown). The significant heterogeneity of residual variance is (partially) due to (non-)segregation of the putative QTL near marker two. The QTL likelihood curves differ in particular at marker nine. This marker is loosely linked to the other markers and also uninformative for families with the largest values of total variance (families 1 and 16 are among them; data not shown). Under model II—with the faulty assumption of homogeneity of residual variance—we can fit a mixture to those families to reduce their within-family residual variance to the average residual variance; this significantly improves the fit to the data and explains the high value of the test statistic at marker nine under model II.

Comparison of models II and III: In model II, we assume homogeneity of residual variance, whereas in model III we allow for heterogeneity of residual variance. Clearly, under model III, the high QTL likelihood at marker nine has disappeared. Analysis by models II and III demonstrates that the assumption of homogeneous variance is not appropriate when fitting a QTL at marker nine. Between markers one and four the QTL likelihood is much higher under model II than under model III (single-test significance levels of 0.0003 and 0.04%, respectively). Under model III, a larger total variance for a certain family can be met by a larger estimate of residual variance, and therefore the evidence for QTL activity will now originate only from differences between the means of genotype classes at a marker. Under model II, reduction of residual variance (ultimately to the average residual variance) will increase the test statistic value. Therefore QTL likelihood is expected to be higher under model II than under model III if a QTL is segregating near marker two. However, differences between total variances can partially originate from segregation of different sets of QTLs at other parts of the genome and, if that is the case, a certain degree of artificial inflation of QTL likelihood is expected under models I and II (although model I is probably more robust).

QTL activity near marker eight is suggested by model III. We extended model III and fitted a *two-QTL* model with QTLs at markers two and eight (we have chosen marker two, because markers one and three are uninformative for families 1 and 16). We compared this model with the single-QTL model with a QTL at marker two only. This (conditional) likelihood ratio test for QTL activity near marker eight was significant at a 0.2% single-test significance level. It is interesting to note that this region is known to contain multiple casein loci that affect protein percentage (Bovenhuis *et al.* 1992).

Comparison of models II–IV: Models II and III assume a multiallelic QTL but ignore the dam contribution. Model IV assumes a biallelic QTL and takes the dam contribution into account. It is somewhat difficult to compare the QTL likelihood curves because of the difference of degrees of freedom involved in the two tests: 20 in models II and III and only two in model IV. The peak in the first marker bracket is more significant under model IV than under model II or III (single-test significance levels of >0.00001, 0.0003 and 0.04%, respectively). Similar power for models I–IV is expected if the true situation is indeed a biallelic-QTL configuration with small QTL effect (Knott *et al.* 1996). The models indicate a large QTL effect (more than one genetic standard deviation) (Spelman *et al.* 1996), so that power will be improved significantly by taking the dam contribution into account (model IV). Combining these results, evidence is provided for the presence of a biallelic QTL near marker two. In contrast, models II, III and IV clearly differ for QTL likelihood near marker eight, and we conclude that putative QTL activity in this region cannot be explained by the presence of a single biallelic QTL. The true situation may consist of a multiallelic QTL or cluster of biallelic QTLs. It is known that several casein loci are clustered in this region of chromosome *6* (Bovenhuis *et al.* 1992).

Comparison of models IV and I: Spelman *et al.* (1996) used model I and reported that the QTL near marker two affects protein percentage in families 1 and 16. The results from our analysis, using model IV, are in agreement with the previous results: the sire is heterozygous in all Gibbs cycles for family 16 and in $\sim$95% of the Gibbs cycles for family 1 (the conditional probability of being heterozygous is high). Under model IV, the other 18 families are homozygous for the QTL in most Gibbs cycles, but they still follow a mixture distribution: each son inherits one of the two QTL alleles from his dam (equivalent to standard segregation analysis within families).

DISCUSSION

Currently, single-QTL methods are still widely used in plant, animal and human genetics, but they are intrinsically inappropriate for complex traits affected by multiple QTLs. In experimental plant applications, multiple-QTL models (MQM) are now used more and more frequently; background QTLs are taken into account by including them (via linked markers) as cofactors in the model (proposed by Jansen 1992; see Jansen 1996b for a review). This can be done in plants because complete marker maps are available for many plant species and also because experimental plant populations, *e.g.*, F_2 or BC, are easier to deal with from the analytical point of view. In animal and human applications, the effects of background QTLs are often modeled by a single variance component term because complete markers maps are not (yet) available for livestock or human populations. In general populations, a marker can be segregating in some families whereas the QTL is not and vice versa. Then we cannot use a marker linked to a putative QTL as the cofactor in the

(expectation or mixture) model, as also indicated by *e.g.*, Spelman *et al.* (1996). In such cases, we should really include the QTL instead of the marker as cofactor in the model, although we can put the putative QTL close to or even on top of a marker. Eventually dense marker maps may become available in human and animal applications, and, with cofactors for background QTLs, it may not be necessary to include other parameters for genetic background control. Modeling via cofactors will also make it possible to explain differences in variances between families originating from segregating of different sets of genes and therefore residual variance may be assumed to be homogeneous; this can not be achieved by a model term for polygenic background effect. We believe that our mixture model approach via stochastic or Monte Carlo EM brings MQM mapping in complex populations within reach. Moreover, our approach uses data imputation via the Gibbs sampler (to generate one possible genotype in stochastic EM and multiple genotypes in Monte Carlo EM) and with these "known" genotypes standard software routines for linear regression, variance component or mixed models can be applied. Our Gibbs sampler is implemented in an easy locus-by-locus and individual-by-individual manner. In particular, the stochastic EM algorithm is relatively easy to program. We have mainly used stochastic EM to provide starting values for Monte Carlo EM. More research has to be done to compare the efficiency of stochastic and Monte Carlo EM in various situations.

A Bayesian approach developed by Satagopan *et al.* (1997) offers an alternative to the MQM mapping approach. These authors assume a Poisson prior distribution for the unknown number of QTLs with discrete effects. By using recent MCMC techniques, the "birth" or "death" of a QTL can be sampled to have great flexibility with respect to the number of QTLs in the model. Other groups now work on similar approaches (I. Hoeschele, personal communication; M. Sillanpää, personal communication).

We expect that ML or Bayesian approaches for multiple-QTL with discrete effects are computationally manageable in complex populations. In contrast, in the case of multiple QTLs of normal effects the computation of multiple variance components may already be much more intensive for three or more QTLs. Therefore, practical computational considerations may prevent the use of variance component models, although multiallelic QTLs may exist, and drawing inferences about multiallelic QTL variance via normal QTL-effects models would be the natural way to characterize genetic variation in the (base) population.

Although the structure of a population may be very complex, a simplified analysis may often be possible. This can be done by either focusing on a well-designed and simple subset of the entire population or by relaxing assumptions and ignoring possible sources of (genetic) variation. For instance, with multiple families, one can estimate allele contrasts for the parents of the families without considering their relationships; one can ignore full-sib relationships within families and perform half-sib analyses for males and females separately; one can select the most likely linkage phases in parents and ignore other configurations; etc. One can then first use an approximate (expectation) method that is computationally inexpensive (Knott *et al.* 1996; Grignola *et al.* 1997) and apply the data simulation method ("parametric bootstrapping") (Jansen 1994) or permutation method (Churchill and Doerge 1994) to obtain genome-wide significance thresholds for QTL detection. In this way, the entire genome would be screened relatively fast to pinpoint regions for further investigation by exact methods that need more computer time. Knott's approximate method uses one step of regression (least squares) analysis at each map location, whereas Jansen's exact method uses multiple cycles of regression analysis (iteratively reweighted least squares). The exact approach is computationally more demanding than the approximate approach. But this may be just a matter of seconds only if markers are highly informative. Moreover, the power and efficiency of the methods will then be similar. In more complex situations, however, we expect the exact (mixture) approaches to be more powerful and efficient than the approximate (expectation) methods at the cost of more computation. Particularly when markers are not fully informative, when individuals are selectively genotyped, when QTLs with large(r) effects are present, or when population structure is complex and much information is lost by simplification, the power and precision can increase considerably by the exact mixture approaches.

As indicated by our analysis of the cattle data set, it can be useful to compare various models with rather different assumptions such as for instance biallelic *vs.* multiallelic QTLs or homogeneous *vs.* heterogeneous residual variance over families. Our analysis suggested the presence of a biallelic QTL near marker two of chromosome 6, and the presence of a cluster of biallelic QTLs or a multiallelic QTL in the region of known casein genes near marker eight. We also demonstrated the pitfall of detecting ghost QTLs when erroneously assuming homogeneity of residual variance. Spelman's, Uimari's and our analyses produce slightly different but still consistent results (Spelman *et al.* 1996; Uimari *et al.* 1996b). This may not be too surprising because multilocus information for paternal inheritance was relatively high for markers 2–8 (see Figure 5 in Spelman *et al.* 1996). The data set was adopted by the animal breeding community to stimulate the development and comparison of (recent) analytical approaches to QTL mapping in complex situations. These cattle data have generated our study, but our methods can handle more complex situations. To investigate proper-

ties of the new methodology in a more thorough way, simulation studies are currently being carried out. For instance, we now study the performance of our mixture approach in the presence of selective genotyping. When marker scores are missing, we sample possible allelic configurations by using the Gibbs sampler, as for the case of the unknown QTL. Preliminary results indicate that the estimates of QTL effects are not biased by selection.

We are grateful to Livestock Improvement Corporation, Holland Genetics and the Department of Genetics of the University of Liege for data access.

LITERATURE CITED

Bovenhuis, H., J. A. M. Van Arendonk and S. Korver, 1992 Associations between milk protein polymorphisms and milk production traits. J. Dairy Sci. **75:** 2549–2559.

Bovenhuis, H., J. A. M. Van Arendonk, G. Davis, J. M. Elsen, C. S. Haley *et al.*, 1998 Detection and mapping of Quantitative Trait Loci in farm animals. Livestock Production Science (in press).

Celeux, G., and J. Diebolt, 1985 The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. Comp. Statist. Q. **2:** 73–82.

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

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

Geyer, C. J., 1993 Estimating normalizing constants and reweighting mixtures in Markov chain Monte Carlo. Technical Report 568, School of Statistics, University of Minnesota.

Grignola, F. E., I. Hoeschele and B. Tier, 1997 Mapping quantitative trait loci via residual maximum likelihood. I. Methodology. Genet. Sel. Evol. **28:** 479–490.

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

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

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

Jansen, R. C., 1992 A general mixture model for mapping quantitative trait loci by using molecular markers. Theor. Appl. Genet. **85:** 252–260.

Jansen, R. C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics **138:** 871–881.

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

Jansen, R. C., 1996b Complex plant traits: time for polygenic analysis. Trends Plant Sci. **1:** 89–94.

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

Janss, L. L. G., R. Thompson and J. A. M. Van Arendonk, 1995 Application of Gibbs sampling for inference in a mixed major gene-polygenic inheritance model in animal populations. Theor. Appl. Genet. **91:** 1137–1147.

Knott, S. A., C. S. Haley and R. Thompson, 1992 Methods of segregation analysis for animal breeding data: a comparison of power. Heredity **68:** 299–311.

Knott, S. A., J. M. Elsen and C. S. Haley, 1994 Multiple marker mapping of quantitative trait loci in half sib populations. Proceedings of the 5th World Congress on Genetics applied to Livestock Production, Guelph, Canada **21:** 33–56.

Knott, S. A., J. M. Elsen and C. S. Haley, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. **93:** 71–80.

Maliepaard, C., and J. W. Van Ooijen, 1994 QTL mapping in a full-sib family of an outcrossing species, pp. 140–146 in *Biometrics in Plant Breeding: Applications of Molecular Markers*, edited by J. W. Van Ooijen and J. Jansen. CPRO-DLO, Wageningen, The Netherlands.

Satagopan, J. M., and B. S. Yandell, 1997 Estimating the number of quantitative trait loci via Bayesian model determination, in *Proceedings of the Section on Biometrics*, American Statistical Association, Alexandria, VA (in press).

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.

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.

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

Titterington, D. M., A. F. M. Smith and U. E. Makov, 1985 *Statistical Analysis of Finite Mixture Distributions.* Wiley, New York.

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

Uimari, P., Q. Zhang, F. Grignola, I. Hoeschele and G. Thaller, 1996b Analysis of QTL workshop I granddaughter design data using least-squares, residual maximum likelihood and Bayesian methods. JQTL: http://probe.nalusda.gov:8000/otherdocs/jqtl

Van Arendonk, J. A. M., B. Tier and B. P. Kinghorn, 1993 Simultaneous estimation of effects of marker and polygenes on a trait showing quantitative genetic variation, pp. 192 in *Proc. XVII Int. Congress of Genetics.* Birmingham, U.K.

Xu, S., 1996 Computation of the full likelihood function for estimating variance at a quantitative trait locus. Genetics **144:** 1951–1960.

Zeng, Z.-B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: P. D. Keightley