Detection of Quantitative Trait Loci in Outbred Populations With Incomplete Marker Data

Marco C. A. M. Bink and Johan A. M. Van Arendonk

Animal Breeding and Genetics Group, Wageningen Institute of Animal Sciences, Wageningen Agricultural University, 6700 AH Wageningen, The Netherlands

ABSTRACT

Augmentation of marker genotypes for ungenotyped individuals is implemented in a Bayesian approach via the use of Markov chain Monte Carlo techniques. Marker data on relatives and phenotypes are combined to compute conditional posterior probabilities for marker genotypes of ungenotyped individuals. The presented procedure allows the analysis of complex pedigrees with ungenotyped individuals to detect segregating quantitative trait loci (QTL). Allelic effects at the QTL were assumed to follow a normal distribution with a covariance matrix based on known QTL position and identity by descent probabilities derived from flanking markers. The Bayesian approach estimates variance due to the single QTL, together with polygenic and residual variance. The method was empirically tested through analyzing simulated data from a complex granddaughter design. Ungenotyped dams were related to one or more sons or grandsons in the design. Heterozygosity of the marker loci and size of QTL were varied. Simulation results indicated a significant increase in power when ungenotyped dams were included in the analysis.

RECENT advances in molecular genetics technology have led to the availability of moderate resolution genetic marker maps for plant and livestock species (e.g., Barendse et al. 1994). Animal and plant breeders are currently using these genetic markers to identify chromosomal regions containing quantitative trait loci (QTL; e.g., Paterson et al. 1988; Stuber et al. 1992; Andersson et al. 1994; Georges et al. 1995). The power of QTL detection is an important factor in the analysis of experiments, that is, to maximize the chance of detecting QTL and minimize the risk of false-positives.

Weller et al. (1990) outlined the granddaughter design to map QTL in dairy cattle. In this design, marker genotypes are determined for granddams and their sons (paternal half sibs), and quantitative trait phenotypes are measured on daughters of sons. This scheme capitalizes on the existing structure in dairy cattle populations and minimizes the number of marker genotypes needed for a given power of detection (Weller et al. 1990). Traditional methods such as (multiple) linear regression and maximum likelihood interval mapping assume unrelated grandsire families and only two generations of genotyped individuals. However, relationships between families, such as related granddams and maternal grandsires and their sons frequently occur in outbred populations. Furthermore, available data may involve multiple generations of genotyped or phenotyped individuals. Exploiting all relationships between individuals and all information collected over generations seems a very appropriate approach to increase power of QTL detection.

Parameter estimation in complex animal (and plant) breeding pedigrees may be tackled by Bayesian analysis, and a comprehensive overview is given by Wang (1998). In Bayesian analysis, prior assumptions and the likelihood of the data at hand form the joint posterior density of all unknown variables in a model underlying the observed phenotypes. Markov chain Monte Carlo (MCMC) methods provide means for exploration of complex non-standard joint densities, and marginal posterior densities of parameters of interest can be approximated. There are a variety of techniques for their implementation (Gelfand 1994) of which Gibbs sampling (Geman and Geman 1984) is the most commonly used. Bayesian linkage analysis in combination with MCMC methods have been applied in human genetics (e.g., Thomas and Cortessis 1992; Heath 1997a), in plant genetics (e.g., Satagopan et al. 1996; Sillanpää and Arjas 1998), and in animal genetics (e.g., Thaller and Hoeschele 1996a; Uimari et al. 1996; Hoeschele et al. 1997).

A second assumption in methods currently employed for QTL linkage analysis of half-sib or full-sib designs, is that all individuals have observed marker genotypes. The incompleteness of marker data may be due to genotyping expenses or lack of DNA. This has hampered the implementation of a full pedigree evaluation in QTL mapping. Augmentation of missing genotypes via the
Gibbs sampler has been suggested (e.g., Thomas and Cortessis 1992). However, when genotypes are missing on parents and the locus has more than two alleles, the Gibbs sampler may be theoretically reducible, i.e., not be able to reach all permissible genotypes from the starting configuration (e.g., Sheehan and Thomas 1993). This reducibility problem does not occur if at least one parent has observed marker genotypes, which may hold for dairy cattle data, where semen of sires is stored and available for DNA typing. In this study, we concentrate on this latter situation.

In this study a Bayesian approach is presented that estimates variance due to a single QTL, together with polygenic and residual variances, allowing ungenotyped individuals. We adapt the method of Jansen et al. (1998) to describe marker information on an individual in terms of allelic constitution of its homologues and identity by descent (IBD) values. We extend the genotype sampling approach of Bink et al. (1998a) from single marker to multiple linked markers. The approach described will be evaluated by the analysis of simulated data from a granddaughter design with many maternal ties between sons, and between sires and sons. Emphasis is on the accuracy of estimates of dispersion parameters. The position of the QTL relative to multiple linked markers is fixed in this study, and possibilities to estimate this parameter are discussed. We also discuss an extension of our approach to pedigrees with no restrictions on incompleteness of marker data.

**MATERIALS AND METHODS**

**Marker genotypes:** Consider a $q$ member population on which marker scores are observed. Let $g$ denote the $ith$ individual’s genotype at all marker loci (excluding the QTL genotype). The genotype $g$ includes full multilocus information about alleles and their IBD pattern, but this information can be observed only partially. For each possible genotypic configuration $g$ on the population (that is, being consistent with observed marker scores) a scalar probability of occurrence may be calculated. The number of possible genotypic configurations exponentially increases when considering marker data on many individuals for many marker loci and containing many missing marker scores. The Gibbs sampler has been successfully used to explore a large number of genotypic configurations and their probability of occurrence (e.g., Guo and Thompson 1992; Janss et al. 1995). Jansen et al. (1998) introduced different descriptions of the genotype of founders (that is, individuals with both parents unknown) and nonfounders in the population. They specified the genotypic state of any founder by the alleles at each of its homologues, and they expressed the state of any nonfounder by IBD values indicating parental origin of its alleles.

For illustration, consider the small pedigree in Table 1. Two founder individuals (individual 1 and 2) have observed marker scores and the linkage phase is assumed to be known for convenience (it limits the number of genotypic configurations that are consistent with observed marker scores). Marker alleles of these individuals are arbitrarily assigned to their first and second homologues, where first and second correspond
to paternally and maternally inherited gametes, respectively. Note that these individuals do not have IBD values because their parents are unknown. On the basis of observed marker scores, three genotypes are allowed for the ungenotyped nonfounder (individual 3). For completeness, alleles of nonfounders’ homologues are also given. Marker data may provide full information on the IBD pattern. For example, the allele a at marker 1 (and 2) for individual 4 is identical by descent to the first allele in its sire, and the IBD value equals 1. More often the IBD patterns are not constant, due to allelic switches in parent or offspring. For example, the IBD values for the allele c at marker 1 for individual 4 depend on the genotypes of individual 3. Note that for a homozygous parent, the IBD value of alleles transmitted to its offspring can be either 1 or 2.

The major advantage of the approach of Jansen et al. (1998) is that in each state of the Markov chain, each marker is informative for each offspring. This means that one can directly use information on the closest flanking markers and one does not have to search for informative segregations. Uncertainty about transmission of alleles is incorporated into the analysis by updating allelic constitution of genotypes in founders and by updating the IBD pattern for nonfounders, as is described later.

**QTL model:** In animal genetics models, allelic effects at the QTL in an outbred population may be represented by a normally distributed random effect, where covariances between allelic effects depend on identical by descent probabilities that are derived from marker information (Fernando and Grossman 1989; Van Arendt et al. 1994; Wang et al. 1995). Let \( v \) denote the vector of additive effects of QTL alleles, containing 2q elements for q individuals. That is, two unique QTL allelic effects are fitted for each individual. For individual \( i \), let \( v_i^p \) and \( v_i^m \) denote the paternally and maternally inherited QTL allele, respectively. Let \( P(a = b) \) denote the probability that alleles \( a \) and \( b \) are identical by descent. Then we can write

\[
\begin{align*}
v_i^p &= P(v_i^p = v_1^p)v_1^p + P(v_i^p = v_2^p)v_2^p + \epsilon_i^p, \quad (1a) \\
v_i^m &= P(v_i^m = v_i^m)v_i^m + P(v_i^m = v_i^m)v_i^m + \epsilon_i^m, \quad (1b)
\end{align*}
\]

where \( s \) and \( d \) denote the sire and dam of the individual, and \( \epsilon_i^p \) and \( \epsilon_i^m \) are residuals (Fernando and Grossman 1989). Consider the situation in which the QTL is located between marker \( k \) and \( k + 1 \) and IBD values are known for these markers (either 1 or 2, see also example in Table 1). Let \( x = p \) or \( m \), when the parent considered is the sire or dam, respectively, and \( n_{i,al} \) is the recombination fraction between marker \( k \) and the QTL. Then the probability of IBD for the QTL can be represented as

\[
P(v_i^x = v_{i,\text{parent}}^x) = \begin{cases} 
(1 - r_{x,i}) \times (1 - r_{x,i+1}) & 1 \\
(1 - r_{x,i}) \times (1 - r_{x,i+1}) & 2 \\
(r_{x,i}) \times (1 - r_{x,i+1}) & 1 \\
(r_{x,i}) \times (1 - r_{x,i+1}) & 2 
\end{cases}
\]

For simplicity, we assume recombination fractions to be equal in males and females. The probability \( P(v_i^x = v_{i,\text{parent}}^x) \) equals \( 1 - P(v_i^x \neq v_{i,\text{parent}}^x) \). The residuals \( \epsilon_i^p \) and \( \epsilon_i^m \) are normally distributed, that is

\[
(\epsilon_i^p, \epsilon_i^m) \sim N(0, \sigma^2_\epsilon I_2),
\]

where \( \sigma^2_\epsilon \) is half the additive genetic variance explained by the QTL, and

\[
\begin{align*}
\delta^p &= \left(1 - \left(\frac{1}{2} P(v_i^p = v_1^p)\right)^2 + \left[2 \times P(v_i^p = v_2^p) \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m)\right]\right) \\
\delta^m &= \left(1 - \left(\frac{1}{2} P(v_i^m = v_1^m)\right)^2 + \left[2 \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m)\right]\right)
\end{align*}
\]

When both parents of individual \( i \) are not inbred at the QTL, this equation reduces to

\[
\begin{align*}
\delta^p &= \left(1 - \left(\frac{1}{2} P(v_i^p = v_1^p)\right)^2 + \left[\frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m)\right]\right) \\
\delta^m &= \left(1 - \left(\frac{1}{2} P(v_i^m = v_1^m)\right)^2 + \left[\frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times \frac{1}{2} \times P(v_i^m = v_2^m) \times P(v_i^m = v_2^m)\right]\right)
\end{align*}
\]

and when parent \( x \) is unknown, \( \delta^p = 1 \). Our model is an approximation of a mixture model in which the QTL allelic effect is exactly identical to one of the parental QTL allelic effects (see also Hoeschele et al. 1997 and Jansen et al. 1998). Changes in allelic effects between parent and (grand) offspring might be due to mutations; the fact that a QTL represents a cluster of closely linked QTL, or epistatic effects.

Let \( G \) denote the genotypic relationship matrix for the QTL \((2q \times 2q)\), where the \((i,j)\) element represents the probability of QTL allele \( i \) being identical by descent to QTL allele \( j \). Then, the conditional density of \( v \) can be given as

\[
(v|G, \sigma^2) \propto \left( G \sigma^2 G^{-1} \right)^{-1/2} \exp\left(-1/2 \sigma^2 (v - \mu)^T G^{-1} (v - \mu) \right)
\]

Van Arendt et al. (1994) presented a recursive algorithm to efficiently construct matrix \( G \) and its inverse \( G^{-1} \). Matrix \( G^{-1} \) has a nice sparse structure: The nonzero elements in \( G^{-1} \) pertaining to an individual’s QTL allelic effect arise from its own contribution (to its parents) plus those of its offspring, i.e., its neighborhood set (e.g., Shesghan and Thomas 1993). The determinant of \( G^{-1} \) and the term \( vG^{-1}v \) can be efficiently computed using partitioned matrix theory (Seagull 1982). After some algebra, the conditional density of \( v \) is

\[
p(v|G, \sigma^2) \propto \prod_{i=1}^{q} \left( (\delta^p)^{-1} \right) \\
\times \exp\left(-1/2 \sigma^2 (\delta^p)^{-1} \times (v_i - \bar{v}_i)^2 \right) \\
+ \left( (\delta^m)^{-1} \times (v_i - \bar{v}_i)^2 \right),
\]

where \( \bar{v}_i = P(v_i^p = v_{i,\text{parent}}^p) \times v_{i,\text{parent}}^p + P(v_i^m = v_{i,\text{parent}}^m) \times v_{i,\text{parent}}^m \) with parent being a sire or dam for \( x \) being the paternally or maternally derived allele of the individual, respectively. And, for example, the full conditional density of the paternal QTL effect of male \( i \), \( v_i^p \),

\[
p(v_i^p|G, \sigma^2) \propto \prod_{i=1}^{q} \left( (\delta^p)^{-1} \right) \\
\times \exp\left(-1/2 \sigma^2 (\delta^p)^{-1} \times (v_i^p - \bar{v}_i^P)^2 \right) \\
+ \left( (\delta^m)^{-1} \times (v_i^p - \bar{v}_i^P)^2 \right),
\]

where \( \bar{v}_i^P = P(v_i^p = v_{i,\text{parent}}^p) \times v_{i,\text{parent}}^p + P(v_i^m = v_{i,\text{parent}}^m) \times v_{i,\text{parent}}^m \) with parent being a sire or dam for \( x \) being the paternal or maternally derived allele of the individual, respectively.

**Updating marker genotypes:** Three classes of individuals are distinguished when updating genotypic information: (1) genotyped founders (with offspring); (2) genotyped non-founders; and (3) ungenotyped parents (unigenotyped non-parents are not considered). Examples in Table 1 of each category are individuals 1 and 2, individuals 4, 5, and 6, and
individual 3, respectively. The sampling of genotypes is described for each of these categories in the subsequent section. Note that when sampling genotypes for markers flanking the QTL, observed trait phenotypes are taken into account via the individuals’ QTL effects.

Category 1: genotyped founders: To take all possible linkage phases in the genotypes of genotyped founders into account, linkage phases are sampled interval by interval and founder by founder, as suggested by Jansen et al. (1998). For a particular set of two neighboring markers, e.g., j and (j + 1), one can use information on the individual, its mates, and offspring to calculate the conditional probabilities for two options, “phase switch” and “no phase switch,” and subsequently sample one of the options. In a phase switch, the distal part of its homologue 1 (marker j + 1 to end) is attached to the proximal part of homologue 2 (map origin to marker j) and vice versa. Also, the IBD values at the distal part of the chromosome in its offspring are switched (1 becomes 2 and vice versa).

Updating of linkage phase for the marker interval containing the QTL actually involves two interval updates, i.e., the interval “left flanking marker—QTL” and “QTL—right flanking marker.” The conditional probabilities of the two linkage phases now also include information from the random QTL, using Equation 5 (the QTL has no IBD patterns). For the left interval, the phase switch option involves a switch in founder QTL effects. This affects the computation of Equation 5, and in a phase switch the founder QTL effects do switch (nothing changes for the QTL effects in its offspring). For the right interval, order of QTL effects within a founder is unaffected.

Category 2: genotyped nonfounders: To generate complete genotypes of nonfounders, one can sample a new IBD pattern given genotype of parents. This can be done individually by individual and marker locus by marker locus. If we update the IBD at a certain marker locus, then the two flanking marker loci (with “known” IBD) are fully informative and no other marker loci are needed. We consider at most four IBD patterns (two per known parent), discarding the ones inconsistent with the individual’s marker score. The IBD values of the individual’s offspring are used when one of the consistent IBD patterns for the individual involves an allelic switch in the individual. When only one parent is known, population allelic frequencies are used. When the individual’s alleles are switched (heterozygous), its offspring’s IBD values are switched as well (1 becomes 2 and vice versa).

When a marker flanks the QTL, the conditional probabilities include information on the QTL effects of the individual and its parents (and also of its offspring if one of the consistent IBD patterns causes an allelic switch within the individual) by using Equation 5 for each consistent IBD pattern.

Category 3: ungenotyped parents: This is the most complicated category because genotypes cannot be updated individually by individual. To illustrate this, suppose a sire with genotype a/b, an ungenotyped dam, and their two offspring (g1 = a/b, g2 = a/c). Starting with g1 = b/c, the first offspring will have a/b, i.e., the a-allele at its paternal homologue and the b-allele at its maternal homologue. Then, updating individual by individual will not allow a switch to the configuration g1 = a/c that would be consistent with the first offspring having a/b instead of a/b. To avoid this problem, we update an ungenotyped parent and its offspring in a block, allowing an allelic switch in the offspring. This allelic switch needs of course to be consistent with the other parent’s marker genotype. The genotype for the ungenotyped parent is sampled from its marginal (with regard to its offspring) distribution, and the IBD of its offspring is subsequently updated from its full conditional (with regard to the parent) distribution. Updates are done marker locus by marker locus. When one or both parents (of the ungenotyped parent) are unknown, the conditional probabilities also involve population allelic frequencies. Note that for an augmented homozygous genotype, the offspring’s IBD value may equal 1 or 2 and both values are taken into account. This also holds for an augmented heterozygous genotype when parent and offspring have the same alleles. When a marker flanks the QTL, the conditional probabilities include information from the QTL using Equation 5. After updating an ungenotyped parent, its genotyped offspring are updated (as described under category 2).

Allele frequencies: The allele frequencies at a particular marker locus in a population are likely unknown and can be treated as such. Let \( \eta_i \) denote the counts of allele i at marker locus m at “founder” homologues, i.e., homologues of founders plus the nonparental homologue of nonfounders with only one parent identified. Then, allele frequencies at each marker locus can be sampled from a Dirichlet distribution with parameters \( \eta_i + 1 \) (for Dirichlet distribution, see Gelman et al. 1995, p. 482).

Mixed linear model: Let \( b \) be a vector of fixed effects, and let \( u \) be a \( q \times 1 \) vector of residual additive (polygenic) effects (not linked to the marker linkage group under consideration). Then the model underlying the \( n \) phenotypes is given as
\[
y = Xb + Zu + ZTv + e,
\]
with
\[
b \sim U(b_{\text{min}}, b_{\text{max}}), \quad u \sim N(0, A_1 s_1^2), \quad v \sim N(0, G_1 r_1^2), \quad e \sim N(0, R_1 s_2^2),
\]
where \( y \) is an \( N \times 1 \) vector of phenotypes, \( X \) and \( Z \) are known design matrices relating records in \( y \) to fixed effects and to individuals, \( T \) is a known incidence matrix relating each individual to its two QTL alleles, \( e \) is a vector of residuals, \( b_{\text{min}}, b_{\text{max}} \) are vectors with minimum and maximum values for fixed effects, \( A \) is the additive genetic relationship matrix (e.g., Henderson 1988), \( s_1^2 \) is the polygenic variance, \( R \) is a known diagonal matrix, and \( s_2^2 \) is the residual variance.

The model is parameterized in terms of the heritability, \( h^2 = s_1^2 / s_2^2 \), proportion of the additive genetic variance due to the QTL \( (\gamma = 2s_1^2 / s_2^2) \) and residual variance \( (1 - \gamma) \), where \( s_1^2 \) is the additive genetic and \( s_2^2 \) is the phenotypic variance. In the remainder of the article \( \gamma \) is referred to as proportion of QTL. In this study, the QTL position relative to the origin of the marker map is assumed known, but this assumption may be removed as shown by Bink et al. (1998c).

Prior knowledge on dispersion parameters: Different priors may be useful to explore the amount of information coming from the data for a particular parameter in the model. In a previous study, Bink et al. (1998b) showed that the posterior density of \( \gamma \) was clearly affected by using different beta distributions to represent prior knowledge on the proportion of QTL \( (\gamma) \), indicating lack of information on \( \gamma \) from the data. In this study, two beta distributions are considered to represent prior knowledge on \( \gamma \). A beta (1,1) prior is uniform between 0 and 1 and is denoted UNIFORM. A beta (1,9) prior has the mode at zero with mean equal to 0.1 and is denoted PEAKED AT ZERO. On the basis of Bink et al. (1998b), priors on \( h^2 \) and \( s_2^2 \) were taken uniformly over the interval \([0, 1]\) and \([0, \infty)\), respectively.

Implementation of MCMC sampling: Bayesian inferences about the parameters are here computed using the Gibbs sampler and the Metropolis-Hastings (MH) algorithm (Metropolis et al. 1953; Hastings 1970) on the basis of the joint posterior distribution of the missing data and the parameters given the observed data \( y \) and marker data \( m \). The missing data are the fixed effects \( b \), the random QTL \( y \) and polygenic \( u \) effects, and marker genotypes (i.e., linkage phase between
alleles at the markers and marker scores for ungenotyped individuals. Now let \( \theta \) denote \( \{b, u, v, h, \gamma, \alpha_i^2\} \).

To reduce the number of genetic effects (polygenic and QTL) that must be sampled (in a granddaughter design), a reduced animal model (RAM; Quaas and Pollak 1980) is used. That is, the genetic effects of ungenotyped granddaughters are absorbed into the parental genetic effects, as described by Bink et al. (1998b).

The sampling distributions for all elements in \( \theta \) are similar to those in Bink et al. (1998b). For parameters \( b, u, \) and \( v \), the full conditional densities are normals and values are drawn by using the Gibbs sampler. A scalar-wise sampling strategy may lead to slow convergence of the Markov chain (Stam and Roberts 1993), especially when elements in \( \theta \) are highly correlated. A full-block sampling strategy, i.e., sampling all correlated elements in \( \theta \) at once, may improve convergence significantly (Liu et al. 1994), but may also be hard to implement in animal breeding applications (Garcia-Cortes and Sorensen 1996). Within the RAM, block sampling as suggested by Jannss et al. (1995) is applied to polygenic effects of grandsires together with those of their sons. Similarly, block sampling is applied, within the RAM, to the QTL effects of grandsires together with the paternally derived QTL effects in their sons and also to the QTL effects of elite dams together with maternally derived QTL effects of their sons. First a new realization is drawn for the parental effect from the reduced conditional density, after absorption of genetic effects of sons. Second, new realizations are drawn for the sons, conditional on the new value of the parental genetic effect.

The full conditional density for \( \alpha_i^2 \) is an inverse chi-square distribution with degrees of freedom equal to \( \text{dim}(e) - 2 \), and sampling is done via the Gibbs sampler. The sampling distributions for \( b, u, v, h, \gamma \) are nonstandard and samples of these parameters are obtained using MH algorithms (Bink et al. 1998b). In the MH algorithms for updating \( h, \gamma \), we used the random walk approach as a candidate generating density (Chib and Greenberg 1995). Lengths of sampling intervals in the random walk need to be determined empirically to arrive at desired acceptance rates for the MH algorithm (e.g., 20–50%, Chib and Greenberg 1995).

**Data simulation:** In this study, we simulated the segregation of a QTL in a granddaughter design. The pedigree material consisted of 20 unrelated grandsires, 400 elite dams, and 800 sons, equally distributed over the 20 grandsires. Two hundred elite dams were daughters of randomly assigned grandsires and the remaining 200 were unrelated to the grandsires. There were no maternal relationships between dams. Dams may have 1, 2, 3, 4, or 5, or 6 sons with probability 0.50, 0.25, 0.10, 0.70, 0.050, and 0.025, respectively (relaxing fixed probabilities, a truncated Poisson distribution may apply). Mating of dams with grandsires was at random, but father-daughter matings were avoided. As a result of this strategy, ~300 dams are related to at least two males in the pedigree (e.g., multiple sons and/or grandsire). About 400 sons are also maternal grandsons of grandsires. These numbers approximately reflect a Dutch granddaughter experiment design as described by Spelman et al. (1996). Polygenic and QTL effects for grandsires and founder dams were sampled from \( N(0, \alpha_i^2) \) and \( N(0, \sigma_t^2) \), respectively. The polygenic effect of individual \( i \) is simulated as \( u_i = \frac{1}{2}(u_{Si} + u_{Di}) + \Phi \). When individual \( i \) has unknown parents, zeros are substituted for \( u_{Si} \) and \( u_{Di} \). The term \( \Phi \) represents Mendelian sampling that follows a normal distribution with mean zero and variance equal to 0.50 \( \alpha_i^2 \), 0.75 \( \sigma_t^2 \), or 1.0 \( \sigma_t^2 \), when 2, 1, or 0 parents are known. Inheritance of QTL effects (and the linked marker alleles) from parent to offspring occurs at random. When a parent is unknown the QTL effect is drawn from \( N(0, \sigma_t^2) \). Individual phenotypes, observed on 100 daughters for each son, were generated as

\[
y \sim N(\frac{1}{2}u_S + \mu_v + (1 - p)v_S, \frac{1}{2}(\sigma_u^2 + \sigma_v^2 + \sigma_t^2)),
\]

where \( p \) is a QTL variable. No phenotypes were simulated for dams and all males. The phenotypic variance and the heritability of the trait were equal to 100 and 0.40, respectively. The proportion of genetic variance due to the QTL (= \( \gamma \)) was equal to 0.10 or 0.25, representing a small and large QTL, respectively (Table 2).

For each individual a 100-cM chromosome was simulated with six markers at 20-cM intervals. The position of the QTL was 30 cM from the origin of the linkage group. Each marker contained either two (low informative markers) or four (high informative markers) alleles with equal frequencies, assuming Hardy-Weinberg equilibrium within marker alleles and linkage equilibrium between alleles of different markers (Table 2).

**Approaches to analyze data:** Marker data in a granddaughter design typically comprise marker genotypes for grandsires and their sons. Three different approaches for analysis are presented in Table 3. The first approach (denoted PAT_RLT) considers only paternal relationships between males in the pedigree, all with marker genotypes. The second approach (denoted ALL_RLT) considers all relationships between individuals in the pedigree, and allows ungenotyped parents (dams) with the condition that all their mates (grandsires) have marker genotypes observed. The third approach (denoted ALL_GTP) also considers all relationships, as in ALL_RLT, but all dams had observed marker genotypes. This third approach was included as a control for two reasons, first to verify whether the results from approach ALL_RLT made sense and second whether approach ALL_RLT could compete with a situation where dams were genotyped.

**Post-MCMC analysis, Bayesian inferences:** For each parameter an effective sample size (ESS) was computed that estimates the number of independent samples with information content equal to that of the dependent samples (Sorensen et al. 1995). From the Bayesian perspective, inference about parameter vector \( \theta \) can be addressed via the posterior density \( P(\theta | y) \). The

----

**Table 2**

<table>
<thead>
<tr>
<th></th>
<th>Proportion of genetic variance explained by the QTL (( \gamma ))</th>
<th>Alleles per marker locus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small QTL, high informative markers</td>
<td>0.10</td>
<td>4</td>
</tr>
<tr>
<td>Large QTL, high informative markers</td>
<td>0.25</td>
<td>4</td>
</tr>
<tr>
<td>Large QTL, low informative markers</td>
<td>0.25</td>
<td>2</td>
</tr>
</tbody>
</table>

*Proportion of genetic variance explained by the QTL (\( \gamma \)).

*Alleles have equal frequencies.

**Table 3**

| Approaches for analysis of data from complex granddaughter designs |
|---|---|---|
| PAT_RLT | Paternal | Males |
| ALL_RLT | Paternal and maternal | Males |
| ALL_GTP | Paternal and maternal | Males and females |
TABLE 4

Average effective samples, average posterior mean estimates, average posterior standard deviations, and the average ln of the odds ratio test statistic across 10 replicates for proportion QTL ($\gamma$)

<table>
<thead>
<tr>
<th>Prior ($\gamma$) = UNIFORM</th>
<th>Prior ($\gamma$) = PEAKED AT ZERO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Effective samples</td>
<td>Average estimate</td>
</tr>
<tr>
<td>PAT_RLT$^b$</td>
<td>134</td>
</tr>
<tr>
<td>ALL_RLT$^c$</td>
<td>101</td>
</tr>
<tr>
<td>ALL_GTP$^d$</td>
<td>132</td>
</tr>
<tr>
<td>Large QTL, low informative markers</td>
<td></td>
</tr>
<tr>
<td>PAT_RLT</td>
<td>192</td>
</tr>
<tr>
<td>ALL_RLT</td>
<td>280</td>
</tr>
<tr>
<td>ALL_GTP</td>
<td>253</td>
</tr>
<tr>
<td>Large QTL, high informative markers</td>
<td></td>
</tr>
<tr>
<td>PAT_RLT</td>
<td>110</td>
</tr>
<tr>
<td>ALL_RLT</td>
<td>121</td>
</tr>
<tr>
<td>ALL_GTP</td>
<td>158</td>
</tr>
</tbody>
</table>

Simulated proportion QTL was small ($\gamma = 0.10$) or large ($\gamma = 0.25$), and information content per marker was high ($n = 4$) or low ($n = 2$). Prior knowledge on proportion QTL ($\gamma$) was UNIFORM or PEAKED AT ZERO.

$^a$ln(odds), ln(posterior mode ($\gamma$)/ posterior ($\gamma = 0$)). Numbers in parenthesis, number of replicates with ln(odds) below the critical value of 3.0.

Approach: PAT_RLT, analysis includes paternal relationships.

Approach: ALL_RLT, analysis includes all relationships, marker genotypes on males.

Approach: ALL_GTP, analysis includes all relationships, marker genotypes on all individuals.

**RESULTS**

**Running the MCMC sampler**: The MCMC sampler was run for 100,000 cycles preceded by a burn-in period of 500 cycles. Each 250th sample was stored for further analysis. This chain length proved to be sufficient to obtain at least 100 effective samples (Sorensen et al. 1995) in most runs. When the effective sample size was <75, the particular replicate was repeated with a different seed and this procedure was sufficient to obtain enough effective samples. Among all parameters, lowest effective sample sizes were found for parameter $\gamma$, indicating that estimating this parameter is most difficult. Effective sample sizes decreased for smaller QTL and for lower informative markers (Table 4). The prior density of $\gamma$ did not seriously affect the effective sample size (Table 4). The MCMC sampler was run on a HP 9000 K260 server, and computing times of a single chain for approach PAT_RLT, ALL_RLT, and ALL_GTP were 23 min, 2 hr 12 min, and 1 hr 1 min, respectively. This indicates that updating the marker haplotypes and IBD patterns for ungenotyped individuals was the most time consuming part of the MCMC sampler.

**Parameter estimates**: Heritability: In all replicates, estimates for parameters $h^2$ and $\sigma_e^2$ were very accurate, independent of approach or size of $\gamma$. For example, for data with a large QTL and low informative markers, the posterior mean estimates of $h^2$ (simulated 0.40) were, averaged over 10 replicates, 0.393, 0.394, and 0.394 for approach PAT_RLT, ALL_RLT, and ALL_GTP, respectively. The averages of estimates of the posterior standard deviation were 0.023, 0.022, and 0.023 for approach PAT_RLT, ALL_RLT, and ALL_GTP, respectively. Similar levels of accuracy were found for estimates of the residual variance. The use of individual phenotypes allows a clear dissection of the phenotypic variance into genetic and residual components. This result was also found by Bink et al. (1998b) and Van Arendonk et al. (1998), but was not found by others (Thaller and Hoeschele 1996b; Uimari et al. 1996; Uimari and Hoeschele 1997), which used the average phenotype of daughters of a sire instead of all individual phenotypes.

Small QTL, high informative markers: The marginal posterior density was flatter and shifted toward the mean of the UNIFORM prior (0.5), when using only paternal relationships compared to using all relationships (Figure 1). The posterior density for PAT_RLT was more...
similar to those of the other two approaches when using the PEAKED AT ZERO prior. Including all relationships led to posterior densities with a smaller standard deviation, that is, higher accuracy of estimates. Including genotypes for dams (ALL_GTP) did not further improve the accuracy. Including all relationships led to smaller estimated HPD90 regions for \( \gamma \) (Figure 1). The HPD90 regions were smaller when the PEAKED AT ZERO prior was used, especially when only paternal relationships were considered. Averaged over 10 replicates, the posterior mean of \( \gamma \) for approach PAT_RLT and the UNIFORM prior was 0.15, which was clearly larger than the simulated value (0.10). Apparently, the data did not provide sufficient information to reduce the effect of the UNIFORM prior, which has an expected mean of 0.5. When the PEAKED AT ZERO prior on \( \gamma \) was used, the estimated posterior mean was equal to the simulated value, which is also the expected mean of the prior (Table 4).

Large QTL, high informative markers: In approach PAT_RLT, the marginal posterior density for parameter \( \gamma \) was relatively flat when the UNIFORM prior was used (Figure 2). The marginal posterior density for \( \gamma \) was clearly shifted toward zero when applying the PEAKED AT ZERO prior in approach PAT_RLT. The other two approaches (ALL_RLT and ALL_GTP) gave similar and more stable densities with the two priors for \( \gamma \), indicating more information coming from the data compared to PAT_RLT. The HPD90 region was largest for approach PAT_RLT with a UNIFORM prior (Figure 2). The PEAKED AT ZERO prior led to a downward shift of the HPD90 regions, in particular for approach PAT_RLT. The PEAKED AT ZERO prior led also to estimated posterior means that were smaller than the simulated values for all approaches (Table 4). The UNIFORM prior led to an upward bias in the estimated posterior mean for approach PAT_RLT but not for the other approaches.

Large QTL, low informative markers: Low informative markers (two alleles per locus) resulted in relatively flat posterior densities for \( \gamma \) (Figure 3), but differences were observed between the three approaches. The use of all relationships improved the accuracy, but in this case the use of all genotypes gave an additional improvement over ALL_RLT. The PEAKED AT ZERO prior led to posterior densities that were closer to zero in all approaches but especially for PAT_RLT. The estimated HPD90 region was again largest for approach PAT_RLT with the UNIFORM prior. The HPD90 regions for approaches ALL_GTP and ALL_RLT were very similar for the UNIFORM prior. However, the HPD90 region for approach ALL_RLT was shifted more toward zero than the region for approach ALL_GTP with the PEAKED AT ZERO prior (Figure 3). The posterior mean estimates were all higher than the simulated value for the UNIFORM prior and below the simulated value for the PEAKED AT ZERO prior. Differences between estimated and simulated values were largest for approach PAT_RLT.

**Hypothesis testing, detection of QTL:** The hypothesis of the presence of a QTL at a particular position in a linkage map was tested via a posterior odds ratio. For a small QTL the ln(\text{odds}) averaged over 10 replicates for approach PAT_REL was 2.69, which was below the critical threshold of 3.0. For approach PAT_REL only 3 out of 10 replicates yielded significant evidence for the presence of a QTL (Table 4). This was very similar to the power of QTL detection found by Bink et al. (1998b). Approach ALL_RLT resulted in an average ln(\text{odds}) of 5.58, and the QTL was significantly detected in 7 out of 10 replicates. Approach ALL_GTP failed to significantly detect the small QTL in only one of the replicates.

For a large QTL and high informative markers, approach PAT_RLT was detected the QTL in at least 8 out of 10 replicates, i.e., two and one failures for UNIFORM and PEAKED AT ZERO priors, respectively (Table 4). The approaches ALL_RLT and ALL_GTP detected the QTL in all replicates. The average ln(\text{odds}) was clearly higher for the large QTL. Note that the posterior odds of approach ALL_RLT for a small QTL [ln(\text{odds}) = 5.64] was even a little higher than the posterior odds of approach PAT_RLT for a large QTL [ln(\text{odds}) = 5.58], when high informative markers were considered.

Reducing heterozygosity of the markers resulted in lower averaged estimates of the ln(\text{odds}) for all cases. The detection rate for approach PAT_RLT with a low informative marker was 50% or lower depending on the prior (Table 4). In all except one case, the QTL was still significantly detected by approaches ALL_RLT and ALL_GTP.

**DISCUSSION**

A variety of statistical gene mapping methods have been developed and applied to outbred populations (see Bovenhuis et al. 1997; Hoeschele et al. 1997). Computationally inexpensive methods, such as regression interval mapping, allow data permutation to determine genome-wide threshold values for test statistics and can be extended more easily to incorporate multiple QTL; however, these methods can only use certain types of relatives (e.g., half-sibships or full-sibships). Bayesian analysis is computationally more demanding but takes full account of the uncertainty associated with all unknowns in the QTL mapping problem and offers the opportunity to analyze general pedigree data and to fit other random components such as polygenic effects (e.g., Thaller and Hoeschele 1996a). Bayesian linkage analysis has been applied in animals (e.g., Thaller and Hoeschele 1996a; Uimari et al. 1996), plants (e.g., Satagopan et al. 1996), and humans (e.g., Thomas and Cortessis 1992). Application of these methods to large
Figure 1.—Marginal posterior inferences for proportion QTL ($\gamma$) for data with small QTL ($\gamma = 0.10$) and high informative markers. Marginal posterior density is given for replicate 1, with UNIFORM prior (top), and with PEAKED AT ZERO prior (middle). Ninety percent highest posterior density regions, averaged over 10 replicates (bottom). Three approaches: PAT_RLT, analysis includes paternal relationships; ALL_RLT, analysis includes all relationships, marker genotypes on males; ALL_GTP, analysis includes all relationships, marker genotypes on all individuals.
Figure 2.—Marginal posterior inferences for proportion QTL ($\gamma$) for data with large QTL ($\gamma = 0.25$) and high informative markers. Marginal posterior density is given for replicate 1, with UNIFORM prior (top), and with PEAKED AT ZERO prior (middle). Ninety percent highest posterior density regions, averaged over 10 replicates (bottom). Three approaches: PAT_RLT, analysis includes paternal relationships; ALL_RLT, analysis includes all relationships, marker genotypes on males; ALL_GTP, analysis includes all relationships, marker genotypes on all individuals.
Figure 3.—Marginal posterior inferences for proportion QTL ($\gamma$) for data with large QTL ($\gamma = 0.25$) and low informative markers. Marginal posterior density is given for replicate 1, with UNIFORM prior (top), and with PEAKED AT ZERO prior (middle). Ninety percent highest posterior density regions, averaged over 10 replicates (bottom). Three approaches: PAT_RLT, analysis includes paternal relationships; ALL_RLT, analysis includes all relationships, marker genotypes on males; ALL_GTP, analysis includes all relationships, marker genotypes on all individuals.
pedigrees with missing genotypes, as described in this article, has not been explored in depth (Hoeschel e et al. 1997). The procedures of Janss et al. (1995), i.e., block sampling of ungenotyped dams and their offspring, and Janssen et al. (1998), i.e., sampling IBD patterns, were implemented to achieve good mixing of the sampler in the full pedigree analysis with incomplete marker information. To accommodate missing marker data, special precautions need to be taken for the sampling procedure to avoid reducibility, i.e., not all possible genotype configurations can be reached from any valid starting configuration. Reducibility especially occurs in situations in which offspring are genotyped but both parents are not. In livestock, the number of offspring per sire is usually large and genetic material from males is often stored, which facilitates genotyping of the male parent. When genetic material is not available, genotypes of males can often be inferred from their offspring. In the present study, it is assumed that marker genotypes on at least one parent are known. This assumption does not limit the application of the presented approach to livestock, but it might be limiting in situations where family sizes are smaller. Sh D eeh an and Thomas (1993) allowed non-Mendelian segregation of alleles (e.g., genotype AB transmitting allele C) to solve the theoretical reducibility. Inferences were based on samples from only those Gibbs cycles with strict Mendelian segregation, which may be an inefficient procedure in large animal breeding populations. Instead of using a fixed probability on non-Mendelian segregation, one may consider a simulated tempering scheme (Geyer and Thompson 1995) that allows this probability to randomly increase from and decrease to zero. Ulmari et al. (1996), Grignola et al. (1996b), and Hoeschel e et al. (1997) investigated the effect of ignoring relationships among families on estimates of QTL location and genetic parameters. Virtually no difference was found between analyses with and without relationships between families for situations with much and little information about the QTL. In our study a large impact of including additional relationships was found (Table 4). This apparent discrepancy in literature can be explained by the relationships considered. In the earlier studies, relationships between the grandsires were included, which leads to additional information on estimating the paternally inherited QTL alleles. In the present study, the ungenotyped dams of the sons were included, which provides information for estimating the maternally inherited QTL alleles. The impact of including additional relationships is clearly demonstrated in Figures 1–3. Including additional relationships resulted in improved estimates of parameter g1, i.e., lower posterior standard deviations and smaller HPD90 regions (Table 4, Figures 1–3). These results strongly suggest that including all relationships in complex pedigrees does improve power of QTL detection.

The pedigree we analyzed consisted of ∼100,000 individuals. The largest proportion of individuals was offspring of sires that only had phenotypic records. The complexity of the problem was reduced by applying a RAM (Quaas and Pollak 1980) in which genetic effects of ungenotyped nonparents are absorbed into those of their parents as presented by Bink et al. (1998b). The procedure presented in this article, which applies a RAM, offers the opportunity to combine the information from different experimental designs, e.g., a grand-daughter design, a grand-granddaughter design (W. Coppieter et al., A. Kvasz, J.-J. Arranz, B. Grisart, J. Riquet et al., unpublished results), or a daughter design, and also the information collected in a closed breeding population spanning several generations. Despite higher computational requirements, the application of a RAM in a Bayesian context more naturally treats missing genotypes than the restricted maximum-likelihood procedures described by Grignola et al. (1996a).

In this study, we assumed a fixed QTL position relative to known markers. Bink et al. (1998c) showed that the position of the QTL can be included as an additional parameter in the model. Appropriate sampling of QTL position was facilitated through the use of simulated tempering (Geyer and Thompson 1995). Simulated tempering, which has also been applied in radiation hybrid mapping (Heath 1997b), proved especially useful to improve mixing by relaxing the distance between closely linked loci. Alternatively, George et al. (1998) mapped a biallelic QTL relative to multiple markers via Bayesian model choice by implementing the reversible jump sampler (Green 1995).

In conclusion, the work presented shows that detection of QTL in data from complex pedigrees is feasible by the use of MCMC and Bayesian analysis. It is shown that using all existing relationships increases the power of detection and the accuracy of the estimates. This work also lays the foundation to study the number of QTL and their relative positions within marker linkage maps.

The authors thank Ritsert Jansen, Luc Janss, Henk Bovenhuis, and Dick Quaas for stimulating discussion. The authors acknowledge financial support from Holland Genetics.

**LITERATURE CITED**


