Maximum Likelihood Mapping of Quantitative Trait Loci Using Full-Sib Families

Sara A. Knott* and Chris S. Haley†

**Institute of Cell, Animal and Population Biology, University of Edinburgh, Scotland, and †Agricultural and Food Research Council, Institute of Animal Physiology and Genetics Research, Roslin, Midlothian, Scotland*

Manuscript received April 1, 1992
Accepted for publication August 21, 1992

ABSTRACT

A maximum likelihood method is presented for the detection of quantitative trait loci (QTL) using flanking markers in full-sib families. This method incorporates a random component for common family effects due to additional QTL or the environment. Simulated data have been used to investigate this method. With a fixed total number of full sibs power of detection decreased substantially with decreasing family size. Increasing the number of alleles at the marker loci (*i.e.*, polymorphism information content) and decreasing the interval size about the QTL increased power. Flanking markers were more powerful than single markers. In testing for a linked QTL the test must be made against a model which allows for between family variation (*i.e.*, including an unlinked QTL or a between family variance component) or the test statistic may be grossly inflated. Mean parameter estimates were close to the simulated values in all situations when fitting the full model (including a linked QTL and common family effect). If the common family component was omitted the QTL effect was overestimated in data in which additional genetic variance was simulated and when compared with an unlinked QTL model there was reduced power. The test statistic curves, reflecting the likelihood of the QTL at each position along the chromosome, have discontinuities at the markers caused by adjacent pairs of markers providing different amounts of information. This must be accounted for when using flanking markers to search for a QTL in an outbred population.

INTEREST in the detection of loci affecting a quantitative trait of importance (quantitative trait loci or QTL) has mainly concentrated on the use of populations derived from the cross between two divergent lines (usually F_2 or backcross populations), particularly two inbred lines. This situation is optimum for the detection of QTL as populations derived from crosses of divergent lines are likely to have QTL of large effect segregating, the QTL and markers have high heterozygosities and, in some situations, the linkage phase between the markers and QTL need not be inferred from the data. However, for many species, including man, such populations are not available and would be difficult or impossible to set up.

Outbreeding populations have several disadvantages for the detection of QTL. In selected populations QTL of large effect on traits under selection may be at extreme frequencies or fixed. If a QTL has only two alleles then, assuming random mating, the heterozygosity is at most 0.5. That is, even in the best situation, on average, 50% of parents do not produce segregating gametes at this locus. Marker heterozygosities are also less than one, although if highly polymorphic markers are used, such as microsatellites, heterozygosity may be high. In the absence of linkage disequilibrium the phase of linkage between the QTL and the markers varies across families, and hence information has to be accumulated on a within family basis (SOLLER 1990).

Where family sizes are large enough (*e.g.*, in dairy cattle and poultry) it may be possible to use least squares based methods to find marker-QTL linkages within single pedigrees without the need to accumulate evidence on individual markers or intervals across pedigrees. In many cases, however, this will not be possible and several methods for accumulation of evidence across pedigrees have been developed for single markers. When based upon analysis of variance techniques, the evidence for linkage is contained in the between marker classes within families sum of squares (*e.g.*, NEIMANN-SORENSEN and ROBERTSON 1961; HILL 1975; SOLLER and GENIZI 1978). GELDERMANN (1975) introduced a χ^2 based test in which the squared differences between animals in different marker classes within families are accumulated. Other methods for the detection of linkage between QTL and markers have been developed for human studies and deserve exploration in the context of animal populations. These include the method of HASEMAN and ELSTON (1972), which involves the regression of the squared

phenotypic difference between two full sibs on the number of alleles at a marker locus that they share which are identical by descent, and a similar method by GOLDGAR (1990) that can take account of multiple markers and more than two full-sibs.

The development of maximum likelihood (ML) methods for the detection of segregating QTL in outbred populations is much more difficult than for inbred lines. In the F_2 analyses only the QTL parameters and the residual variance are unknown, whereas in segregating populations the allele frequency at the QTL must be estimated in addition to these parameters and the linkage phase between markers and QTL (and sometimes between markers) is unknown and all possible haplotype combinations must be considered. There is also variation between families due to genetic effects unlinked to the markers and to environmental effects. The necessity to include this additional between family variation in the model needs to be investigated. In segregation analysis not accounting for this source of variation parameter estimates are biased and spurious genes can be detected (KNOTT, HALEY and THOMPSON 1992). Maximum likelihood methods have been considered in human genetics but with the assumption that the QTL accounts for all the genetic variance and ignoring any common family environmental effect (for example DEMENAIS, LATHROP and LALOUEL 1988). In order to account for some of the additional correlations between relatives BONNEY, LATHROP and LALOUEL (1988) suggested the use of a regressive model in conjunction with linkage analysis.

In this paper the likelihood necessary to apply interval mapping (using flanking markers) in full-sib families is described, including an approximation which makes feasible the incorporation into the model of a polygenic or environmental between family variance component. In order to explore the properties of the method it is applied to a limited number of data sets simulated containing a QTL of relatively large effect. The influence of family structure and marker information content on the test statistic and the parameter estimates are considered. Marker data from grandparents are relatively easy to incorporate and will provide prior information on the phase of linkage between markers in the parents and do not substantially increase the computation involved. Simulated data are used to assess the benefit of the inclusion of this additional information. A full-sib population structure would be relevant to a number of species, including pigs, poultry and man where the natural structure is one of full-sib families and cattle, where the possibilities of multiple ovulation and embryo transfer enable full-sib families to be produced and maintained within a herd.

TABLE 1

Parameters required in the models for analysis

Model[a]	μ	r_A	a	d	p	σ_w	σ_b
1. Linked QTL	x	x	x	x	x	x	
2. Linked QTL + common family effect	x	x	x	x	x	x	x
3. Unlinked QTL + common family effect	x		x	x	x	x	x
4. Unlinked QTL	x		x	x	x	x	
5. Common family effect	x					x	x
6. Random environmental effect	x					x	

x indicates that the parameter is optimized in this model.
[a] Models 1 to 5 also included a residual environmental effect.

DERIVATION OF THE FLANKING MARKER LIKELIHOOD

Model: The population is assumed to be composed of a number of unrelated full-sib families whose parents mated at random. Marker information is available for both the parents and their offspring. In some situations it is assumed marker information is also available from grandparents. Phenotypic information is recorded only on the offspring.

A QTL is assumed to be in linkage equilibrium with all markers in the parental population. Linkage between a QTL and a marker will, however, generate linkage disequilibrium within families. Hence, information on linked QTL comes from segregation within families. For simplicity we consider only two alternative alleles at the QTL but any number of alleles at the marker loci.

The model for the phenotype of the jth full sib in the ith family with QTL genotype q can be written as follows:

$$y_{ij} = \mu + g_q + u_i + e_{ij}$$

where: y_{ij} is the phenotype of the jth offspring of the ith family, μ is the QTL mid-homozygote value, g_q is the effect of QTL genotype q as a deviation from the mid-homozygote value (μ), u_i is the random effect for family i and e_{ij} is the individual residual random component. The family and residual random effects are assumed to be normally and independently distributed with variances σ_b^2 and σ_w^2, respectively. g_q is equal to a for Q_1Q_1, d for Q_1Q_2 and Q_2Q_1 and $-a$ for Q_2Q_2.

Initially we will assume that there are no additional loci or common family environmental effects contributing to the trait and thus u_i can be omitted (model 1; see Table 1).

In subsequent notation the subscripts s, d and o refer to the sire, dam and offspring, respectively.

Flanking marker likelihood: We require the likelihood of the data given that a QTL is located between two adjacent markers. This likelihood is composed of the probabilities of the parents' genotypes and the

likelihood of the offspring phenotypes and genotypes given the parents. To begin with we will consider the likelihood of a full-sib family given the marker and QTL genotypes and phase of linkage in the two parents. In general the phase of linkage between markers in the parents is unknown and information on phase comes from the offspring, hence the offspring likelihood has to include both the likelihood of the phenotypes and the marker genotypes.

Given the parental information the likelihood of each offspring is independent of its full sibs. For each offspring both the phase of linkage between the markers and the origin (from the sire or dam) of the two marker haplotypes have to be considered, so there are a maximum of four possibilities (for an offspring heterozygous at both marker loci whose parents were also both heterozygous for the same alleles at both marker loci). The number of possible QTL genotypes that an individual might have depends on the QTL genotypes under consideration for the parents, with a maximum of three when both parents are assumed to be heterozygous at the QTL. For an offspring that is heterozygous at the QTL, the probability of it inheriting either allele Q_1 from its sire and Q_2 from the dam or vice versa has to be considered, so, effectively we might consider four QTL genotypes (accounting for the parental origin of the alleles). Given the parental information and an assumption about interference between recombination events the probability of each of the possible QTL and marker combinations for each full sib offspring can be calculated from the recombination fractions between the postulated QTL and the flanking markers. Thus transmission probabilities can be obtained for all offspring and all QTL genotypes in a similar way as for a backcross or F_2 population (LANDER and BOTSTEIN 1989) (although in the full-sib case the probability of the QTL genotype joint with, rather than conditional on, the observed marker genotypes is calculated). These probabilities will be denoted trans(m_o, $q_o|m_s$, q_s, m_d, q_d) (where m refers to the marker phase and q to the QTL genotype). We have assumed that there is no interference so that a recombination event in one interval does not influence the frequency of recombination in an adjacent interval. With a prior estimate for the recombination fraction between the markers, r, the transmission probabilities can be written in terms of a single unknown parameter, the recombination fraction between one of the markers (say A) and the QTL (r_A).

The likelihood of the full-sib progeny (of family i) can be written as follows:

$$L = \prod_{j=1}^{n_i} \sum_{m_j=1}^{M_j} \sum_{q_o=1}^{Q} \text{trans}(m_j, q_o|m_s, q_s, m_d, q_d) \frac{1}{(2\pi\sigma_w^2)^{1/2}} \exp\left[\frac{-(y_{ij} - \mu - g_{q_o})^2}{2\sigma_w^2}\right]$$

where: n_i is the number of full sibs in family i, M_j is the number of possible marker phases for offspring j (including whether marker haplotypes from sire or dam), Q is the number of genotypes at the QTL (*i.e.*, 4) and trans(m_j, $q_o|m_s$, q_s, m_d, q_d) is the probability of the offspring marker and QTL genotypes given the parental genotypes and phase of linkage.

The QTL genotype and phase of linkage are not known in any parent and hence all possible genotypes and phases need to be considered. For each combination the likelihood of the full-sibship can be calculated and this weighted by the prior probability of the phase of marker linkage and QTL genotype assumed for each parent. Hence the likelihood for one family (i) can be written as follows:

$$L_{1i} = \sum_{m_s=1}^{M_s} p(m_s) \sum_{q_s=1}^{Q} \text{freq}(q_s) \sum_{m_d=1}^{M_d} p(m_d) \sum_{q_d=1}^{Q} \text{freq}(q_d) \prod_{j=1}^{n_i} \sum_{m_j=1}^{M_j} \sum_{q_o=1}^{Q} \text{trans}(m_j, q_o|m_s, q_s, m_d, q_d) \frac{1}{(2\pi\sigma_w^2)^{1/2}} \exp\left[\frac{-(y_{ij} - \mu - g_{q_o})^2}{2\sigma_w^2}\right]$$

where: $p(m)$ is the probability of marker phase m, freq(q) is the frequency of QTL genotype q in the parental generation and M_s and M_d are the number of possible marker phases for the sire and dam of family i. The remaining notation has been described previously.

There are two possible phases of linkage between the two marker loci for each parent, if the parent is heterozygous at each marker the two alternatives need to be considered. With no prior information the probability ($p(m)$) of each phase is one half. However, marker information may be available on ancestors of the parents which may alter this probability. Taking account of the QTL phase relative to the markers there are four QTL genotypes to be considered for each parent. The probability of each genotype is obtained from the genotype frequencies of the QTL in the parental population. We have assumed that the sires and dams come from the same population and that this population is in Hardy-Weinberg equilibrium.

For N independent full-sib families the total likelihood can simply be obtained by taking the product of the likelihood L_{1i} for each full-sib family.

$$L_1 = \prod_{i=1}^{N} L_{1i} \quad (1)$$

Incorporating a between family effect: Mean differences between full-sib families could be caused by a QTL, but also by additional loci affecting the trait of interest or by environmental factors. Model 2 incorporates an effect common to each family to account

for these between family differences. This random common family effect is assumed to be normally distributed (with a mean of zero and variance σ_b^2) and independent of the QTL and the within family environmental component. Equation 1 has to be extended to include the likelihood of this family effect (u_i) and the likelihood of the offspring phenotype given the QTL genotype has to take account of the family effect. The value of this parameter, however, is unknown and for a given family mean is expected to differ according to the different QTL genotypes considered for the parents. Hence, the likelihood has to include an integration over all possible values of the family effect.

The following likelihood is obtained:

$$L_2 = \prod_{i=1}^{N} \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma_b^2)^{1/2}} \exp\left[\frac{-u_i^2}{2\sigma_b^2}\right] \cdot \sum_{m_s=1}^{M_s} p(m_s) \sum_{q_s=1}^{Q} \text{freq}(q_s) \sum_{m_d=1}^{M_d} p(m_d) \sum_{q_d=1}^{Q} \text{freq}(q_d) \prod_{j=1}^{n_i} \sum_{m_j=1}^{M_j} \sum_{q_o=1}^{Q} \text{trans}(m_j, q_o | m_s, q_s, m_d, q_d) \frac{1}{(2\pi\sigma_w^2)^{1/2}} \exp\left[\frac{-(y_{ij} - \mu - q_{qo} - u_i)^2}{2\sigma_w^2}\right] . du_i \quad (2)$$

where: N is the number of full-sib families.

Equation 2 can be integrated to give a likelihood that can be calculated exactly, however, this exact form involves a summation over all the possible combinations of QTL genotype and marker phase in the parents and offspring and quickly becomes infeasible to calculate as the number of full sibs increases. An alternative is to approximate the integration. A numerical approximation can be used (for example, HILDEBRAND 1974), replacing the integration with a weighted summation and we have shown previously that this gives a very close approximation to the exact likelihood (KNOTT, HALEY and THOMPSON 1992). A description of the method is given in the APPENDIX.

TESTING FOR A QTL

To detect the presence of a QTL linked to a marker (or within an interval) the maximized likelihood given in Equation 2 or 1 has to be compared to the maximum likelihood under a model without a linked QTL. The likelihood equation is written below for a model that assumes the trait is controlled by a QTL that is unlinked to the markers and incorporates a random family effect (model 3). Marker data are omitted from this likelihood which is now a segregation analysis likelihood.

$$L_3 = \prod_{i=1}^{N} \int_{-\infty}^{+\infty} \frac{1}{(2\pi\sigma_b^2)^{1/2}} \exp\left[\frac{-u_i^2}{2\sigma_b^2}\right] \sum_{q_s=1}^{Q} \text{freq}(q_s) \sum_{q_d=1}^{Q} \text{freq}(q_d) \prod_{j=1}^{n_i} \sum_{q_o=1}^{Q} \text{trans}(q_o | q_s, q_d) \frac{1}{(2\pi\sigma_w^2)^{1/2}} \exp\left[\frac{-(y_{ij} - \mu - g_{qo} - u_i)^2}{2\sigma_w^2}\right] . du_i \quad (3)$$

where: trans ($q_o | q_s, q_d$) is the Mendelian transmission probability of QTL genotype q_o given the parents' genotypes q_s and q_d. The remaining parameters have been described previously.

Numerical approximation to the integration can be used to calculate the likelihood under this model. Other models that can be fitted include an unlinked QTL with a residual random component but no common family effect (model 4), a model incorporating a random effect for common family effect but omitting the QTL (model 5) and a model that assumes that the data are controlled by a random environmental effect only (model 6). The likelihoods for these (likelihoods 4–6) can be obtained by appropriate simplication of Equation 3. Table 1 summarizes the models in terms of the parameters each requires.

With nested hypotheses, WILKS (1938) has shown that twice the difference in the natural log likelihoods is asymptotically distributed as a χ^2 with degrees of freedom equal to the number of parameters estimated in the full model and fixed in the reduced model. In our case some of the required conditions for this distribution are violated (*i.e.*, parameters are fixed on bounds in the reduced model causing some parameter redundancy, *e.g.*, if $p = 1$ then parameters a and d are meaningless). From work with segregation analysis, however, where similar problems are encountered (KNOTT, HALEY and THOMPSON 1992) we would suggest that a χ^2 distribution with these degrees of freedom would be suitable. For the models including linkage of a QTL (models 1 and 2 without and with a common family effect, respectively), the likelihood of the offspring phenotypes and their marker genotypes are included, whereas for the remaining models (3–6) the likelihood of the phenotypes only is used. To be comparable, therefore, the likelihood of the observed marker genotypes given the recombination fraction between the marker loci has to be included. Under models 3–6, the genetic models for the phenotypes do not depend on the markers, therefore the natural logarithm of the likelihood of the phenotypes can simply be added to the log likelihood of the marker genotypes. This is equivalent to subtracting the log marker likelihood from log likelihood under model 1 or 2 to give the likelihood of the phenotype

data for the offspring given the model, conditional on the observed marker genotypes.

SIMULATION AND ANALYSES

Parents were generated by random allocation of genotypes at each locus depending on the frequency of alleles and assuming Hardy-Weinberg equilibrium. Offspring were generated assuming no interference, so that a recombination event in one interval does not effect the occurrence of a recombination event in an adjacent interval.

We explored the mapping of a QTL in a segment of "chromosome" 50 cM in length. A QTL with two alleles at equal frequency and additive effect of two residual (*i.e.*, environmental and non-QTL genetic effects) standard deviations between homozygotes was simulated to be in the middle of this segment. Markers were placed symmetrically 10 and 25 cM from the QTL. Markers were simulated to have eight alleles, which could be reduced to two, at equal frequency. A polygenic component was included by simulating 25 unlinked loci each with small, additive effect on the trait and alleles at equal frequency, so that they accounted for one third of the total variance in the population. The QTL and residual effects also each accounted for one third of the total variance.

We considered a fixed number of progeny (1000) and altered the number of full-sib families (and thus parents and grandparents) that the population comprised. Marker data were simulated for all three generations but phenotypes only for the progeny. Fifty, 100 and 250 families were considered. The first situation could be a pig population, the second either pigs or cattle and the third cattle or humans.

For the population comprising 100 families of size 10, data were also simulated with the same marked chromosome segment and unlinked polygenic component but without the QTL. The polygenic component accounted for one half of the total variance.

Ten replicates of each data set were simulated and analyzed. The same phenotypic data were analyzed, using either the markers 20 cM apart or those 50 cM apart and with either eight or two alleles at each marker. The data were also analyzed both including and ignoring marker information from the grandparental generation. The maximum value of the likelihood of the data under all of the models presented previously (1–6) was obtained. Additionally, the likelihood using linkage to a single marker was also maximized. This likelihood is the same as that given for flanking markers (Equation 2), except the problem of phase of linkage of the marker loci no longer has to be considered only phase between the marker locus and QTL. The transmission probabilities now involve only a single recombination fraction and can be calculated given the observed marker genotypes in the offspring and the postulated QTL genotype.

The likelihoods were maximized using the quasi-Newton algorithm E04JAF (Numerical Algorithms Group 1990). The simulated value was used as a prior estimate for the recombination fraction between the markers and was fixed in the analyses, hence, for example, for model 2 there were seven unknown parameters (r_A, p, a, d, μ, σ_b, σ_w). Initial estimates of the parameters are required for the maximization process and these were approximately the values used to simulate the data. For models incorporating either of the genetic components but not both, initial estimates were set such that the remaining genetic part explained the total simulated genetic variance.

RESULTS

QTL detection: From the analyses of each set of data the following test statistics (*i.e.*, twice the log difference between the likelihoods) were calculated from likelihoods (1) to (6) (with expected degrees of freedom (d.f.)):

i: Linked QTL + common family effect (2) *vs.* unlinked QTL + common family effect (3) (1 d.f.).
ii: Linked QTL (1) *vs.* unlinked QTL (4) (1 d.f.).
iii: Unlinked QTL + common family effect (3) *vs.* unlinked QTL (4) (1 d.f.).
iv: Unlinked QTL + common family effect (3) *vs.* common family effect (5) (3 d.f.).
v: Unlinked QTL (4) *vs.* random environmental effect (6) (3 d.f.).
vi: Common family effect (5) *vs.* random environmental effect (6) (1 d.f.).

The models for likelihoods (1) to (5) also included a residual (random) environmental effect.

Table 2 gives the mean test statistic for a linked QTL (*i.e.*, tests i and ii) over the replicate analyses. The single marker results (test i_s) are given as the mean of the higher of the test statistics obtained for the two markers. For tests that ignored marker information (tests iii to vi) the mean values over the ten replicate simulations are given in Table 3.

When the data were simulated without a linked QTL the test of a linked versus an unlinked QTL is expected to have a mean of one and a variance of two. When flanking markers were used this test statistic was frequently negative. The test of a linked QTL versus no QTL is expected to have a mean of four and a variance of eight. Adding the test statistics from test i and test iv gives the mean test statistic for this test when a common family component is included, and adding the test statistics from tests ii and v gives the mean test statistic for this test when a common family component is omitted. The last test is grossly inflated because the null hypothesis cannot account

TABLE 2

Mean test statistics for three tests of linkage

Stimulated QTL effect	No. of families	No. of FS	Marker alleles	Interval size	Mean test statistic[a]		
					Test i_f	Test i_s	Test ii_f
0.00	100	10	8	20	0.6	1.6	−18.9
					(4.4)	(2.1)	(12.4)
	100	10	2	20	−1.1	0.6	−14.2
					(1.9)	(1.1)	(5.3)
	100	10	8	50	−3.3	0.9	−12.8
					(5.8)	(2.2)	(6.5)
	100	10	2	50	−1.5	0.6	−6.2
					(3.4)	(1.2)	(3.8)
0.33	50	20	8	20	84.8	63.2	73.2
					(23.7)	(19.0)	(20.3)
	50	20	2	20	52.2	37.5	44.6
					(21.6)	(14.3)	(22.4)
	50	20	8	50	42.9	26.1	38.1
					(18.8)	(11.7)	(16.3)
	50	20	2	50	19.9	16.0	17.0
					(14.4)	(10.6)	(13.5)
0.33	100	10	8	20	57.5	44.5	52.4
					(11.3)	(11.9)	(12.5)
	100	10	2	20	26.7	18.1	25.2
					(10.6)	(7.8)	(12.6)
	100	10	8	50	21.9	16.2	19.0
					(11.0)	(9.9)	(12.8)
	100	10	2	50	9.5	7.4	9.0
					(8.2)	(6.0)	(11.2)
0.33	250	4	8	20	19.9	17.4	18.1
					(10.5)	(8.3)	(10.9)
	250	4	2	20	12.5	10.3	11.3
					(10.2)	(7.7)	(12.6)
	250	4	8	50	5.5	4.8	4.2
					(3.6)	(2.7)	(5.3)
	250	4	2	50	2.2	2.7	0.0
					(4.2)	(3.3)	(6.4)

Each mean is based on 10 replicate simulations and the empirical standard deviation is given in parentheses. In all cases the polygenic variance explained one third of the total variance, and the QTL either contributed one third (0.33) or was omitted (0.00). The QTL had two alleles and was additive in effect. All marker alleles were simulated to be at equal frequency. FS = full sibs.

[a] Test i, linked QTL + common family *vs.* unlinked QTL + common family (1 d.f.); test ii, linked QTL *vs.* unlinked QTL (1 d.f.); *f* and *s* refer to the use of flanking and single markers, respectively.

for between family differences. Any model that can take account of differences between families fits much better than one that cannot, hence the high test statistics for tests v and vi.

The mean test statistics obtained when a QTL is simulated provide some indication of the relative powers of the different marker situations and population structures. Flanking markers are more powerful than single markers (test i_f compared with test i_s), and the incorporation of a between family effect increases the power (test i_f compared with ii_f). Decreasing the number of full sibs per family decreases the test statistic, as does increasing the interval between the markers (or distance to a single marker) and decreasing the number of alleles at the marker loci. Segregation analysis (test iv) is much less powerful for detection of a QTL than when marker information is included (*i.e.*, tests i and ii) and has virtually no power when the family size is reduced to ten full sibs.

Table 4 gives the increase in mean test statistics, compared with the results in Table 2, obtained by the incorporation of information from the grandparental marker genotypes for the same data and using flanking markers. Including this information has little impact but has greatest effect in the analyses with a large number of small families and when eight alleles are segregating at markers 50 cM apart.

Parameter estimation: Mean estimates for the position (given in recombination units from one marker, say *A*), additive effect and allele frequency of the QTL estimated using the likelihood in Equation 2 are given in Table 5 along with the mean variance estimates. The mid-homozygote value and dominance effect were also estimated but have not been pre-

TABLE 3

Mean test statistics ignoring marker information

Simulated QTL effect	No. of families	No. of FS	Mean test statistic[a]			
			Test iii	Test iv	Test v	Test vi
0.00	100	10	8.2	2.6	122.9	128.6
			(7.0)	(2.3)	(27.9)	(29.5)
0.33	50	20	10.6	7.4	262.4	265.6
			(7.0)	(4.5)	(41.9)	(44.3)
0.33	100	10	11.0	4.6	201.7	208.1
			(6.5)	(3.5)	(38.9)	(41.1)
0.33	250	4	6.9	2.9	102.5	106.5
			(5.8)	(2.2)	(22.9)	(25.0)

The means are from the same data as used for Table 2. The empirical standard deviations are given in parentheses. FS = full sibs.

[a] Test iii, unlinked QTL + common family *vs.* unlinked QTL (1 d.f.); test iv, unlinked QTL + common family *vs.* common family (3 d.f.); test v, unlinked QTL *vs.* environment (3 d.f.); test vi, common family *vs.* environment (1 d.f.).

TABLE 4

Increase in the mean test statistics for tests of linkage when grandparental marker information is incorporated

No. of families	No. of FS	Marker alleles	Interval size	Mean test statistic[a]	
				Test i*f*	Test ii*f*
50	20	8	20	0.0	0.0
50	20	2	20	0.1	−0.1
50	20	8	50	1.5	1.3
50	20	2	50	0.2	0.2
100	10	8	20	0.3	0.3
100	10	2	20	0.1	0.3
100	10	8	50	1.6	1.6
100	10	2	50	0.0	0.0
250	4	8	20	0.6	0.4
250	4	2	20	0.2	0.2
250	4	8	50	2.1	2.5
250	4	2	50	0.3	0.6

The same data as used for Table 2 with a QTL has been reanalyzed incorporating grandparental marker information into the likelihood of a linked QTL. The results are presented as the increase in the mean test statistic compared with ignoring this information (*i.e.*, as given in Table 2). FS = full sibs.

[a] Test i, linked QTL + common family *vs.* unlinked QTL + common family (1 d.f.); test ii, linked QTL *vs.* unlinked QTL (1 d.f.).

sented, however, in all cases the mean estimates were close to the expected value of zero. Results are shown for all the analyses of data with 100 families (without grandparental information). The other population structures showed the same trends and so, for these, only the means from the analyses of data with eight marker alleles and a 20 cM interval have been given. In general, the estimates are close to the simulated values for all situations. Estimates are also given in Table 5 for the most and least powerful situations for the analyses using single markers (giving the mean over the estimates from the marker associated with the highest likelihood) and for the analyses using flanking markers but omitting the common family component. Using single markers the distance from the marker was, on average, underestimated. Omitting the common family effect from the likelihood caused the QTL effect to be overestimated, on average, but did not appear to affect the other parameters.

Identifying the correct interval: The results given so far assume that the correct interval containing the QTL can be identified. For a given position of the QTL (*i.e.*, fixing the recombination fraction between the marker and postulated QTL) the remaining parameters can be estimated and the test statistic calculated. With the assumption of no interference, Haldane's mapping function (HALDANE 1919) can be used to convert recombination fractions to distances in centimorgans. Figure 1 gives the profile of the test statistic along the chromosome for one set of data with a single QTL at 10 cM with additive effect that explains one third of the phenotypic variance. Polygenic and environmental variances also each explained one third of the total variance. The data consisted of 100 full-sib families of size ten. The same phenotypic data has been analyzed and the two lines are the curves obtained with either two alleles at each marker or eight alleles. Both curves show a maximum at approximately the location of the QTL. The shape of the curve around this maximum indicates the accuracy with which this location is obtained. Using eight alleles at the markers the test statistic is greater and the maximum is much more clearly defined. The surfaces are not smooth but have discontinuities at the markers.

Figure 2 gives the profile of the test statistic for the same set of phenotypic data as used for Figure 1. However, in this case the markers flanking the QTL each have only two alleles and the additional marker has eight alleles. In this case the overall maximum has shifted into the interval bounded by the more informative markers rather than being in the correct interval. If the marker at 20 cM is omitted, so that a single interval 35 cM long is considered, the suggested position of the QTL moves back towards the simulated position. The small kink in this latter curve occurs where the best model changes from one with little dominance (on the left hand side of the kink) to one with overdominance (on the right hand side).

DISCUSSION

We have shown how the principles of interval mapping can be applied to the analysis of data from full-sib families. This requires the inclusion of a between family polygenic or environmental component in the model, which is made computationally feasible by the

TABLE 5

Mean parameter estimates

No. of families	No. of FS	Marker alleles	Interval size	Mean parameter estimates				
				r_A	a	p	σ_w	σ_b
Flanking markers								
50	20	8	20	0.100	10.17	0.468	8.74	4.45
				(0.004)	(0.28)	(0.015)	(0.05)	(0.27)
100	10	8	20	0.095	10.24	0.478	8.60	4.34
				(0.005)	(0.20)	(0.015)	(0.05)	(0.39)
100	10	2	20	0.106	10.05	0.448	8.69	4.25
				(0.007)	(0.32)	(0.024)	(0.08)	(0.49)
100	10	8	50	0.218	8.49	0.489	8.69	4.61
				(0.013)	(1.05)	(0.030)	(0.09)	(0.51)
100	10	2	50	0.199	8.85	0.526	8.80	4.66
				(0.023)	(0.88)	(0.037)	(0.15)	(0.50)
250	4	8	20	0.090	9.61	0.459	8.77	4.35
				(0.014)	(0.43)	(0.035)	(0.12)	(0.20)
Single markers								
100	10	8	20	0.054	10.18	0.508	8.65	4.33
				(0.016)	(0.37)	(0.025)	(0.14)	(0.44)
100	10	2	50	0.116	9.31	0.561	8.85	4.69
				(0.035)	(0.73)	(0.050)	(0.17)	(0.50)
Flanking markers omitting between family variance								
100	10	8	20	0.097	11.78	0.472	8.75	
				(0.005)	(0.19)	(0.015)	(0.07)	
100	10	2	50	0.207	12.08	0.493	8.57	
				(0.008)	(0.20)	(0.024)	(0.10)	
Simulated values for all analyses								
			20	0.091	10.00	0.500	8.66	5.00
			50	0.197				

Mean estimates, over the 10 replicate simulations, of the recombination fraction between the QTL and marker (r_A), the additive effect at the QTL (a) and the allele frequency (p) and the between and within family variances (σ_b and σ_w). The standard errors of the means are given in parentheses.

use of a numerical approximation to the integration over the random family component.

The use of full-sib family data is much less powerful than in the F_2 for the detection of a QTL of equivalent size. LANDER and BOTSTEIN (1989) suggested an approximation to predict the non-central portion of the test statistic (the additional central portion being equal to the degrees of freedom) for F_2 populations (the accuracy of which has been confirmed by simulation studies of F_2 populations, KNOTT and HALEY 1992). This approximation would predict a test statistic of about 325 for a QTL which explains 33% of the variance in an F_2 population of 1000 individuals and was flanked by two markers each 10 cM away (230 for a QTL explaining 25% of the variance, which is the average proportion of variance explained by the QTL within a full-sib family). The highest mean test statistic we observed in this study is 85 and there are several factors contributing to this difference in power. First, loss of information compared with an F_2 population arises because linkage phase between markers and QTL varies between parents and thus information on linkage is obtained from within families segregating for the QTL (SOLLER 1990). If the QTL has only two alternative alleles, then assuming Hardy-Weinberg equilibrium, at most 50% of parents are expected to be heterozygous at the QTL and, hence, in 25% of families the QTL will not be segregating. These families will not provide information about the effect and linkage of the QTL. In the F_2 population all markers are informative, whereas with an outbreeding population in linkage equilibrium some marker loci will be homozygous and hence provide no linkage information.

Single marker analyses were less powerful than analyses using flanking markers, with the single marker test statistics being, on average, only 70% of the test statistics using flanking markers. This is unlike the situation found with an F_2 population (for example, KNOTT and HALEY 1992) where there is little difference in power. In the F_1 parents of an F_2 population all markers are heterozygous and thus informative whereas, in an outbreeding population some parental markers are not heterozygous. Thus using single markers the frequency of parents with completely uninformative markers is higher than it is with flanking markers in an outbreeding population. Presenting the results from the single marker giving the highest

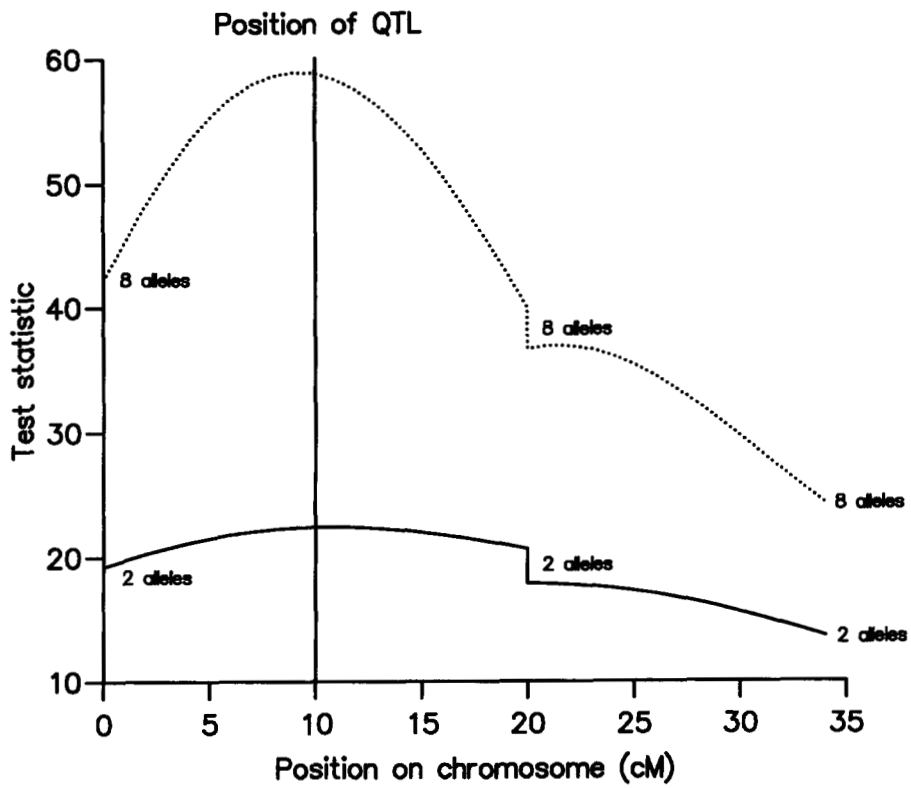

FIGURE 1.—Test statistic (test i_f) curves produced from the analysis of a single set of phenotypic data. The data were generated with a single QTL with two alleles at equal frequency with additive effect and explaining one third of the phenotypic variance. It was simulated to be 10 cM from the end of the chromosome. An additional third of the phenotypic variance was explained by many polygenes not located on this chromosome. Marker loci were positioned at 0, 20 and 35 cM and could have either two or eight alleles with alleles at equal frequency in both cases. The two curves result from analysis using either eight alleles at all three loci or only two at each.

test statistic has minimized the difference between the use of single and flanking markers. On the other hand, the QTL was placed at the mid-point of the interval which gives a comparison least favorable to the single marker analyses.

With a fixed total number of offspring, increasing the number of full-sib families (and hence, decreasing the number of full sibs per family) decreases the test statistic. With four full sibs, even with the large QTL simulated here, the method was not very powerful and if the markers are widely spaced and have few alleles a gene is unlikely to be detected. Information on the phase of linkage between the markers and QTL in the parents comes from the full sibs. With increasing numbers of full sibs linkage phase can be more accurately determined.

The inclusion of grandparental marker data provides additional information only on the phase of marker linkage in the parents. The largest improvement in test statistic from inclusion of grandparental marker information was observed when a 50-cM interval was being considered, there were eight alleles at each marker locus and the family size was small. This is consistent with expectations as information on the phase of marker linkage in the parents is only useful when the parent is heterozygous at both markers. With only two alleles at the marker loci only about 25% of parents are expected to have both markers heterozygous and so grandparental information will only help in these cases. On the other hand with eight alleles at the markers over 76% are expected to be heterozygous at both markers. With closely linked markers each with a high number of alleles and a reasonable family size the offspring will provide good information about the phase, but with fewer offspring and the markers further apart the grandparental information is valuable.

We have presented results from tests of linkage, *i.e.*, a comparison of a linked with an unlinked QTL. With this test when using flanking markers the hypotheses compared are not nested because only a restricted set of linked hypotheses are considered in which the QTL falls at or between the markers (*i.e.*, we compare the likelihood of a QTL in the interval with that for a completely unlinked QTL). This test, therefore, can give negative test statistics (as observed in Table 2), *i.e.*, close linkage may be less likely than no linkage and the test statistic distribution under the null hypothesis will not be χ^2. Note that in contrast, when only a single marker is used the hypotheses are nested as the recombination fraction for a linked QTL can vary from 0.0 to 0.5 and negative test statistics are

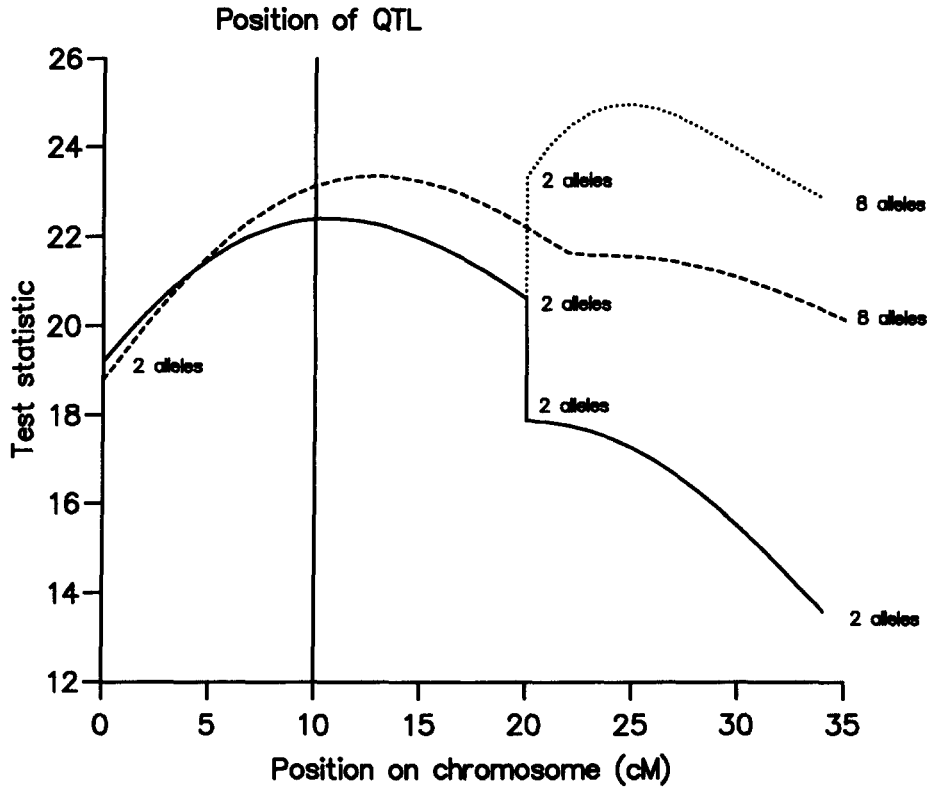

FIGURE 2.—Test statistic curves (test i_f) produced from the analysis of a single set of phenotypic data. The same set of phenotypic data as used for figure 1 has been analysed. The solid line gives the test statistic curve shown previously (Figure 1) for the analysis using two alleles at the three marker loci. The dotted line is obtained analysing the second interval with two alleles at the first marker and eight at the second. The dashed line results from analyses omitting the middle marker and with the first marker having two alleles and the last eight.

not observed. Also with single markers, the χ^2 distribution is not expected to hold because of the selection of the marker giving the highest likelihood. In the context of F_2 data, where large simulation studies are more feasible, we have suggested that for flanking markers linked *vs.* no QTL provides a robust and unbiased test (KNOTT and HALEY 1992). For example, any non-normality inflates the likelihood under a model with an unlinked QTL leading to a negative test statistic when the likelihood under the linked hypothesis is compared with this likelihood, whereas the test of a linked versus no QTL in F_2 is less affected. By analogy, this suggests, the use in full-sib family data of the test of the likelihood under a model with a linked QTL and common family effect compared with the likelihood under a model with just a family effect. From the results presented here, the power of this test is similar to that of the test with an unlinked QTL and common family component as the null hypothesis and the test would not suffer from negative values. In any event it is obvious from Table 3 that the null hypothesis has to be able to take account of differences between families, otherwise any model that can explain some of these differences is much more likely even if it is not the correct explanation of the data. Omitting the common family component, and comparing a linked QTL with an unlinked QTL (test ii) has an appropriate null hypothesis (it can explain between family differences) but the power is reduced compared with test i_f (or i_f + iv), presumably because a linked QTL does not adequately explain unlinked genetic variance.

Mean test statistics, rather than power, have been presented to indicate the relative benefit of the situations considered. Although we are ultimately interested in power the test statistics provide additional information, for example an indication of differences between situations that would in fact both provide 100% power of detection. Also it is the test statistic, not power, that scales with the sample size and the amount of variance explained by the QTL, at least in F_2 populations (KNOTT and HALEY 1992). Given a significance threshold test statistics can be converted to power making use of non-central χ^2 distributions. Setting the significance threshold should take account of the number of tests performed and their non-independence, but the level at which to set it is problematic (LANDER and BOTSTEIN 1989; RISCH 1991).

Using flanking markers with the complete model (Equation 2) the parameter estimates obtained for the effect, frequency and position of the QTL were, on average, close to the simulated values. The variance

components also correctly reflected the simulated values. The underestimate of the recombination fraction using single markers partly results from taking the marker associated with the higher test statistic which, in general, gives a lower estimate of recombination fraction. If the common family component from the likelihood was omitted, the estimate for the effect of the QTL was inflated. This presumably occurs because the linked QTL has to go some way towards explaining the between family variance due to the simulated polygenic effect.

The analysis is carried out on a within family basis and hence the presence of related full-sib families (*e.g.*, having the same sire) should not lead to spurious detection of QTL or bias the estimates of the effect of a QTL substantially. The estimates of the variance components and the population allele frequency of the QTL may be biased. However, it should be relatively easy to extend this model to include the additional information correctly if the same sire has been used for several full-sib families (in a hierarchical design).

The magnitude of the test statistic increases with the information content of the markers, thus intervals bounded by markers with a high information content tend to increase the test statistic for this interval even if the QTL is in an adjacent interval (see Figure 2). Hence, care will have to be taken when comparing intervals with different information not to incorrectly position a QTL in a more informative interval. The problem may be partially resolved by analyzing two adjacent intervals as one, omitting the intervening marker. A better solution would be to analyze three or more markers jointly (optimally a whole chromosome at a time). Analyzing more markers simultaneously will not only remove discontinuities between intervals but will also be more powerful, however, whilst theoretically feasible this is computationally impractical at present.

In the analyses presented here a prior estimate for the recombination fraction between the markers has been used. This recombination fraction could be estimated jointly with the QTL parameters, however, this would increase the computation and with a correct assumption about interference the joint estimate is likely to be similar to a prior estimate based only on the marker data in the sample, as observed for F_2 data (KNOTT and HALEY 1992). The recombination fraction has been assumed to be constant across sexes. Where prior information or where the marker data suggests that recombination rates are different in the two sexes this could be easily incorporated. If enough data were available two fractions could be estimated, one for each sex, alternatively, map distances could be assumed to be in a constant ratio with each other, the ratio dictated by the marker data.

In conclusion, the ML method suggested using flanking markers and incorporating a random common family effect can be used to detect a QTL and to estimate its effect and frequency. To obtain adequate power we need to aim to create a fairly dense map with highly informative markers. Ensuring that they were all highly informative would also reduce problems associated with intervals varying in information content. For the same total number of offspring power increases substantially with family size and with larger families use of additional marker information from grandparents is of little benefit. To minimize bias in the estimates the null hypothesis should allow for between family variance. The analyses are computationally intensive (each simulation and set of analyses took on average 24 hr for a single run on a Sequent Symmetry computer at about 20 MIPS) and approximations are required in order to allow further investigation of the use of full sibs and other outbreeding population structures.

We acknowledge the support of the Agricultural and Food Research Council (AFRC) and the Ministry of Agriculture, Fisheries and Food (MAFF) in the United Kingdom and by the BRIDGE programme of the Commission of the European Communities.

LITERATURE CITED

ABRAMOWITZ, M., and I. STEGUN, 1972 *Handbook of Mathematical Functions*. Dover, New York.

BONNEY, G. E., G. M. LATHROP and J. M. LALOUEL, 1988 Combined linkage and segregation analysis using regressive models. Am. J. Hum. Genet. **43:** 29–37.

DEMENAIS, F., G. M. LATHROP and J. M. LALOUEL, 1988 Detection of linkage between a quantitative trait and a marker locus by the lod score method: sample size and sampling considerations. Ann. Hum. Genet. **52:** 237–246.

GELDERMANN, H., 1975 Investigations on inheritance of quantitative characters in animals by gene markers. I. Methods. Theor. Appl. Genet. **46:** 319–330.

GOLDGAR, D. E., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.

HALDANE, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

HASEMAN, J. K., and R. C. ELSTON, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:** 3–19.

HILDEBRAND, F. B., 1974 *Introduction to Numerical Analysis* (International Series in Pure and Applied Mathematics). McGraw-Hill, New York.

HILL, A. P., 1975 Quantitative linkage: a statistical procedure for its detection and estimation. Ann. Hum. Genet. **38:** 439–449.

KNOTT, S. A., and C. S. HALEY, 1992 Aspects of maximum likelihood methods for the mapping of quantitative trait loci in line crosses. Genet. Res. (in press).

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.

LANDER, E. S., and D. BOTSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

NEIMANN-SORENSEN, A., and A. ROBERTSON, 1961 The association between blood groups and several production character-

istics in three Danish cattle breeds. Acta Agric. Scand. **11:** 163–196.

Numerical Algorithms Group, 1990 *The NAG Fortran Library Manual–Mark 14*. NAG Ltd., Oxford.

RISCH, N., 1991 A note on multiple testing procedures in linkage analysis. Am. J. Hum. Genet. **48:** 1058–1064.

SOLLER, M., 1990 Genetic mapping of the bovine genome using deoxyribonucleic acid-level markers to identify loci affecting quantitative traits of economic importance. J. Dairy Sci. **73:** 2628–2646.

SOLLER, M., and A. GENIZI, 1978 The efficiency of experimental designs for the detection of linkage between a marker locus and a locus affecting a quantitative trait in segregating populations. Biometrics **34:** 47–55.

WILKS, S. S., 1938 The large sample distribution of the likelihood ratio for testing composite hypotheses. Ann. Math. Stat. **9:** 60–62.

Communicating editor: B. S. WEIR

APPENDIX

Numerical approximation to the integration: Hermite integration can be used to approximate Equation 2 by replacing the integration with a weighted summation. Suitable values of the parameter to be integrated (u_i) and appropriate weights are obtained from the Hermite polynomial, and have been tabulated (*e.g.*, found in ABRAMOWITZ and STEGUN 1972). These standard values for the abscissae and weights should be multiplied by $\sqrt{2}\sigma_b$ and $1/\sqrt{\pi}$, respectively. Preliminary analyses showed that over 30 points in the summation were required to obtain an accurate value for the likelihood in this situation. With fewer points in the summation problems were encountered with maximization. A reduction in the number of points required for an accurate approximation can be obtained by the incorporation of a location parameter (l_i) so that the summation for each family (i) is not taken about zero. Also a scaling parameter (v_i) can be incorporated to adjust the range of the abscissae. Effectively an integration over u_i is replaced with an integration over z_i where $z_i = (u_i - l_i)/\sqrt{2}\ v_i$. A suitable value for the location parameter is the family mean due to polygenic and common environmental effects. In our model the predicted value of this mean family effect is expected to differ depending on the QTL genotypes being considered for the parents. Hence, a different value for the location parameter for each QTL genotype mating combination in the parents (now denoted l_{ic}) was considered. For each mating combination the full-sib family mean was adjusted according to the expected contribution from the QTL, and then scaled to give the location parameter. For family i with QTL genotypes q_s and q_d for the sire and dam the following formula is obtained:

$$l_{ic} = \frac{n_i}{n_i + \lambda}\left(\bar{y}_{i.} - \mu - \sum_{q=1}^{Q} \text{trans}(q \mid q_s, q_d) g_q\right)$$

where λ is equal to σ_w^2/σ_b^2, $\bar{y}_{i.}$ is the full-sib mean and trans($q \mid q_s, q_d$) is the Mendelian transmission probability of QTL genotype q given parents are q_s and q_d. Other parameters have been defined previously. For example, when considering the QTL genotypes Q_1Q_1 and Q_1Q_2 for the two parents the transmission probabilities would be 0.5, 0.5, 0 and 0 for the genotypes Q_1Q_1, Q_1Q_2, Q_2Q_1 and Q_2Q_2 respectively. These values, estimated for each family and QTL genotype combination in the parents, were used as the value around which the summation was taken. A suitable scaling parameter (v_i) is $\sqrt{\sigma_w^2/(n_i + \lambda)}$. For a model omitting the QTL the use of these location and scaling parameters would give an exact value for the likelihood with a single point in the summation. The likelihood can now be written as follows:

$$L = \prod_{i=1}^{N} \frac{v_i}{\sigma_b (2\pi\sigma_w^2)^{n_i/2}} \sum_{t=1}^{T} \frac{1}{\sqrt{\pi}} W_t \sum_{m_s=1}^{M_s} p(m_s) \sum_{q_s=1}^{Q} \text{freq}(q_s)$$

$$\sum_{m_d=1}^{M_d} p(m_d) \sum_{q_d=1}^{Q} \text{freq}(q_d) \exp\left[\frac{-(z_t\sqrt{2}v_i + L_{ic})^2}{2\sigma_b^2} + z_t^2\right]$$

$$\prod_{j=1}^{n_i} \sum_{m_o=1}^{M_j} \sum_{q_o=1}^{Q} \text{trans}(m_j, q_o \mid m_s, q_s, m_d, q_d)$$

$$\exp\left[\frac{-(y_{ij} - \mu - q_{qo} - (z_t\sqrt{2}v_i + L_{ic}))^2}{2\sigma_w^2}\right]$$

where W_t and z_t are obtained from standard tables and T is the number of points in the summation. The use of different location parameters and a scaling parameter substantially reduced the number of points required so that six gave a good approximation.