Theor Appl Genet (1992) 84:803–811

THEORETICAL AND APPLIED GENETICS

Accuracy of mapping quantitative trait loci in autogamous species

Johan W. van Ooijen

Department of Genetics, Agricultural University, Dreijenlaan 2, 6703 HA Wageningen, The Netherlands

Received November 4, 1991; Accepted March 10, 1992
Communicated by J.W. Snape

Summary. The development of linkage maps with large numbers of molecular markers has stimulated the search for methods to map genes involved in quantitative traits (QTLs). A promising method, proposed by Lander and Botstein (1989), employs pairs of neighbouring markers to obtain maximum linkage information about the presence of a QTL within the enclosed chromosomal segment. In this paper the accuracy of this method was investigated by computer simulation. The results show that there is a reasonable probability of detecting QTLs that explain at least 5% of the total variance. For this purpose a minimum population of 200 backcross or F_2 individuals is necessary. Both the number of individuals and the relative size of the genotypic effect of the QTL are important factors determining the mapping precision. On the average, a QTL with 5% or 10% explained variance is mapped on an interval of 40 or 20 centiMorgans, respectively. Of course, QTLs with a larger genotypic effect will be located more precisely. It must be noted, however, that the interval length is rather variable.

Key words: Mapping – Molecular markers – Quantitative trait loci

Introduction

The inability to identify the genotype (the alleles) of the genes determining a quantitative trait prevents location of the individual genes by normal linkage mapping. Biometrical approaches have been proposed and used to study association between genetic markers and quantitative trait loci (QTLs) (see, for example, Sax 1923; Thoday 1961). However, the lack of a sufficient number of genetic markers, distributed evenly over the entire genome, has hampered further development of these methods.

This problem has been overcome by the development of linkage maps occupied by large numbers of molecular markers, such as restriction fragment length polymorphisms (RFLPs) (e.g., Helentjaris 1987; Zamir and Tanksley 1988; Keim et al. 1990). Several methods have been proposed for mapping QTLs, two of which are based on estimating linkage between a single marker and a QTL (Weller 1986; Luo and Kearsey 1989). However, a disadvantage of a method based on a single marker is that the power decreases with increasing distance between the marker and the QTL. To increase the efficiency, more markers per chromosome need to be employed. The linkage information on a putative QTL at a certain chromosomal segment between two co-dominant markers is maximal when the segregation information of these two markers is used simultaneously. Hence, adding segregation information of markers outside such a segment does not provide additional linkage information on a QTL within this segment. If, however, one (or both) of the two markers is dominant, then extra information on a QTL within the segment can be gained by employing neighbouring markers. The co-dominance requirement holds for most RFLP markers, but does not hold for the recently developed RAPD (random amplified polymorphic DNA; Williams et al. 1990; Welsh and McClelland 1990) markers.

Estimation methods based on two, so-called, flanking markers have been published by Jensen (1989), Lander and Botstein (1989), and Knapp et al. (1990). These authors apply similar methods based on maximum likelihood. Jensen's method is specific for doubled haploids. It

Present address: Centre for Plant Breeding and Reproduction Research (CPRO-DLO), P.O. Box 16, 6700 AA Wageningen, The Netherlands

allows estimation of segregation distortion, which is a common phenomenon in doubled haploids and wide (interspecific) crosses. However, this approach only considers one pair of marker loci. Knapp et al. (1990) present a number of models for various population types, but here again only a single pair of marker loci is involved. The approach of Lander and Botstein integrates information on all employed markers into one, so-called, QTL likelihood map. From this map an approximate position of a QTL is indicated by a support interval. It is the integration of all markers that makes this method very appealing, giving detailed insight into the part of the genome investigated. In order to provide an indication of the attainable resolution of this procedure, the method is demonstrated in their paper with a single simulated backcross progeny.

Applications of the Lander and Botstein approach have been presented by Paterson et al. (1988, 1991). In their 1990 paper Paterson et al. described the fine mapping of the QTLs which were originally found in their 1988 paper. The results were not satisfactory. For instance, a QTL for soluble solids in tomato was found in a segment on chromosome 1 (TG158-TG27) that showed only a non-significant effect in the 1988 study. On the other hand, a very significant effect on fruit mass in the left region of chromosome 1 (TG24-TG59) as described in their 1988 study was not detected here. Instead, a factor on the other side of the chromosome (TG245-TG255) was found. Are these discrepancies occasional, or are they a sign of the inaccuracy of the method? The aim of this paper is to gain insight into the accuracy of the QTL mapping procedure as described by Lander and Botstein (1989). To this end a computer simulation study was performed in which the probability of detecting a QTL and the precision of the mapping were investigated for backcross and F_2 populations.

Of course, the accuracy of any QTL mapping procedure depends on a number of factors: the heritability of the trait, the number of genes involved, the interactions of the genes, the distribution of the genes over the genome, the statistical distribution of the random non-genetic factors, the type of segregating population studied, the size of this population, the genome size, and the number of marker loci employed, as well as their distribution over the genome. Because of the large number of factors involved, we restrict ourselves to a few relatively simple cases, which are described in the Methods section.

The QTL mapping procedure

The QTL mapping procedure, as described by Lander and Botstein (1989) in a somewhat concise form, will be presented here more extensively, applied to a first generation backcross (BC_1) and an F_2 population. Essentially, the method is a maximum likelihood approach to the segregation of a mixture of probability distributions (compare Titterington et al. 1985; McLachlan and Basford 1988). For each position in the genome (e.g., every 1 centiMorgan) we want to calculate how likely the presence of a segregating QTL at that position is. In order to do that, we have to specify the mixture model. Requirements are a known and accurate linkage map, and the genotypes of segregating marker loci at regular positions on this map of all individuals in the cross progeny. The linkage information for a position on a chromosome is maximal when the genotype information of the two markers flanking this position is used simultaneously. For each genotype class of these two markers there is a mixture, which consists of two (BC_1) or three (F_2) normal distributions with means μ_0, μ_1, or μ_2 according to the QTL genotype, qq, Qq, or QQ, respectively, and equal residual variance σ_r^2. The corresponding probability density functions are denoted as $f_q(x)$, with $q \in \{0,1,2\}$ for qq, Qq, and QQ, respectively. Because we are looking at a specific position on the genome, we are able to separate the effects from QTLs on other chromosomes. Thus, the residual variation is caused by QTLs (affecting the same trait) on other chromosomes, and by random non-genetic factors. For each marker class there would be a separate mixture model. But, since these models have the same components $f_q(x)$, and the mixing proportions are related through the linkage map, one general model can be specified. Given the marker genotype m for two flanking markers (A and B), their map distance (translated into recombination frequency r), and the position of the QTL (Q) between the markers (r_a and r_b denote the recombination frequencies between A and Q, and Q and B, respectively; $0 \le r_a \le r$; when there is no interference, then: $r = r_a + r_b - 2r_a r_b$), the phenotypic value x of an individual has the probability density function (pdf):

$$f(x \mid m; r_a) = \sum_{q=0}^{2} \pi_{mq} f_q(x)$$

with:

m = marker genotype,
r_a = recombination frequency between marker A and QTL Q,
π_{mq} = probability for QTL genotype $q \in \{0,1,2\}$ depending on marker genotype m and position determined by r_a,
$f_q(x)$ = normal pdf with mean μ_q and variance σ_r^2.

The mixing proportions π_{mq} are determined by the segregation ratios for the QTL within a marker class, and add up to one. These can easily be computed using the expected genotype frequencies presented in Tables 1 and 2. For instance, the mixing proportions for the BC_1 marker genotype AB/AB are:

$$\pi_{AB/AB0} = 0, \pi_{AB/AB1} = \frac{r_a r_b}{r_a r_b + (1-r_a)(1-r_b)},$$

$$\pi_{AB/AB2} = \frac{(1-r_a)(1-r_b)}{r_a r_b + (1-r_a)(1-r_b)}.$$

The likelihood, or joint pdf of the entire progeny is:

$$L = L(\mu_0, \mu_1, \mu_2, \sigma_r^2; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f(x_i \mid m_i; r_a) \quad (1)$$

Table 1. Expected genotype frequencies in BC_1 progeny from the backcross AQB/aqb × AQB/AQB, multiplied by two

Genotype	Frequency of QTL genotype		
	qq	qQ	QQ
A.B/AQB	0	$r_a r_b$	$(1-r_a)(1-r_b)$
A.b/AQB	0	$r_a(1-r_b)$	$(1-r_a)r_b$
a.B/AQB	0	$(1-r_a)r_b$	$r_a(1-r_b)$
a.b/AQB	0	$(1-r_a)(1-r_b)$	$r_a r_b$

Table 2. Expected genotype frequencies in F_2 progeny from a self fertilised AQB/aqb, multiplied by four

Genotype		Frequency of QTL genotype		
		qq	qQ	QQ
AA	BB	$r_a^2 r_b^2$	$2r_a(1-r_a)r_b(1-r_b)$	$(1-r_a)^2(1-r_b)^2$
	Bb	$2r_a^2 r_b(1-r_b)$	$2r_a(1-r_a)[r_b^2+(1-r_b)^2]$	$2(1-r_a)^2 r_b(1-r_b)$
	bb	$r_a^2(1-r_b)^2$	$2r_a(1-r_a)r_b(1-r_b)$	$(1-r_a)^2 r_b^2$
Aa	BB	$2r_a(1-r_a)r_b^2$	$2[r_a^2+(1-r_a)^2]r_b(1-r_b)$	$2r_a(1-r_a)(1-r_b)^2$
	Bb	$4r_a(1-r_a)r_b(1-r_b)$	$2[r_a^2+(1-r_a)^2][r_b^2+(1-r_b)^2]$	$4r_a(1-r_a)r_b(1-r_b)$
	bb	$2r_a(1-r_a)(1-r_b)^2$	$2[r_a^2+(1-r_a)^2]r_b(1-r_b)$	$2r_a(1-r_a)r_b^2$
aa	BB	$(1-r_a)^2 r_b^2$	$2r_a(1-r_a)r_b(1-r_b)$	$r_a^2(1-r_b)^2$
	Bb	$2(1-r_a)^2 r_b(1-r_b)$	$2r_a(1-r_a)[r_b^2+(1-r_b)^2]$	$2r_a^2 r_b(1-r_b)$
	bb	$(1-r_a)^2(1-r_b)^2$	$2r_a(1-r_a)r_b(1-r_b)$	$r_a^2 r_b^2$

with:

x_i = phenotypic value of individual i,
m_i = marker genotype of individual i,
N = number of individuals.

The values of μ_0, μ_1, μ_2 and σ_r^2 that maximise this likelihood are the maximum likelihood estimates. In order to maximise L with respect to $\theta \in \{\mu_0, \mu_1, \mu_2, \sigma_r^2\}$, we may maximise ln(L) by differentiation:

$$\frac{\partial \ln L}{\partial \theta} = \sum_{i=1}^{N}\left[\frac{1}{f(x_i|m_i;r_a)}\sum_{q=0}^{2}\frac{\partial \pi_{mq} f_q(x_i)}{\partial \theta}\right] = \sum_{i=1}^{N}\sum_{q=0}^{2}\left[\frac{\pi_{mq} f_q(x_i)}{f(x_i|m_i;r_a)}\frac{\partial \ln f_q(x_i)}{\partial \theta}\right]. \quad (2)$$

Define:

$$\alpha_q(x_i|m_i;r_a) = \frac{\pi_{mq} f_q(x_i)}{f(x_i|m_i;r_a)},$$

i.e., the probability of an individual to be of QTL genotype q at the locus determined by r_a, if it has phenotypic value x, and conditional on its marker genotype m. When the derivatives are set equal to zero, the resulting equations can be solved for θ (see Appendix):

$$\hat{\mu}_q = \frac{\sum_{i=1}^{N}[\alpha_q(x_i|m_i;r_a)x_i]}{\sum_{i=1}^{N}\alpha_q(x_i|m_i;r_a)}, \quad (3a)$$

and:

$$\hat{\sigma}_r^2 = \frac{1}{N}\sum_{i=1}^{N}\sum_{q=0}^{2}[\alpha_q(x_i|m_i;r_a)(x_i-\mu_q)^2]. \quad (3b)$$

These solutions are, in fact, the weighted means and weighted mean square, respectively; the weights are the conditional probabilities $\alpha_q(\cdot)$. The solutions are recursive, unless the QTL coincides with one of the two markers, i.e., $r_a = 0$ or $r_a = r$. In the latter case it is easy to show that the solutions are the analysis of variance estimators of the marker class means and residual variance [except for the degrees of freedom, for which N is used instead of (N−2) or (N−3) with BC_1 or F_2, respectively]. In the former case, however, there are no explicit solutions. In that case solutions to the likelihood equation (1) can be obtained with several numerical methods. A convenient and fast iterative method is the EM algorithm (E, expectation; M, maximisation; Dempster et al. 1977). The EM algorithm is a general method for calculating maximum likelihood estimates from incomplete data. Applied to the analysis of mixture distributions, the memberships of the distributions are the missing data. To be specific for QTL mapping: the genotype at the Q-locus is unknown.

The E-step concerns the expectation of the missing data, conditional on the known data and some initial approximate values of $\theta \in \{\mu_0, \mu_1, \mu_2, \sigma_r^2\}$. The M-step determines new values for θ by maximising the log-likelihood using the initial values of θ and the expected values of the missing data calculated in the E-step. This is done with equations (3). The E- and M-steps are executed alternately, each time replacing the former values with the newly estimated ones, until the log-likelihood stops increasing. However, for QTL mapping the expectations of the genotype at the Q-locus are fully determined by the fixed map position, i.e., r_a, which are the mixing proportions π_{mq}. Hence, the E-step in this procedure always results in the same values. Because the mixing proportions are fixed the method is not an exact form of EM. It is possible to incorporate the estimation of mixing proportions in a comparable estimation procedure (see, for example, Knapp et al. 1990). This, however, will lead to discrepancies with the map distance between the markers, which is assumed to be known accurately from previous experiments.

The mixing proportions are probabilities, and the real distribution of QTL genotypes in a finite sample will not be exactly equal to these proportions. Since in the procedure the mixing proportions are fixed, this will lead to errors. These are assumed to be of minor importance when the map distance between markers is small, because: (1) in marker classes, that are non-recombinant for the markers, the probabilities for two out of the three QTL genotypes are very small, since these require at least one double recombination; (2) each recombinant marker class is relatively small in number.

The iterative solution employs equations (3). Appropriate starting values for the QTL genotype means and residual variance are the population mean and variance, respectively. A new set of parameter values is calculated by substituting the current set of values in the right hand side of equations (3). For each iteration the log-likelihood value is calculated; the iterations are stopped when this value does not increase more than a certain predefined fractional tolerance value (e.g., 10^{-6}).

When the null hypothesis, $H_0: \mu_0 = \mu_1 = \mu_2$ (in effect meaning there is no QTL) is true, the likelihood can be calculated:

$$L_0(\mu_{pop}, \sigma_{pop}^2; x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} f(x_i)$$

with:

μ_{pop} = population (BC_1 or F_2) mean
σ_{pop}^2 = population variance
f(x) = normal pdf with mean μ_{pop} and variance σ_{pop}^2.

A likelihood ratio test statistic for the alternative hypothesis, that a QTL is segregating at the current position, is transformed into a so-called LOD score (LOD=log of odds):

$$\text{LOD} = {}^{10}\log\left[\frac{L(\mu_0,\mu_1,\mu_2,\sigma_r^2;x_1,x_2,\ldots,x_N)}{L_0(\mu_{pop},\sigma_{pop}^2;x_1,x_2,\ldots,x_N)}\right].$$

For a single test the LOD score, when H_0 is true, is asymptotically distributed as a chi-square random variable $(\frac{1}{2}\,{}^{10}\log e)\chi^2$ with 1 degree of freedom for a BC_1 and 2 df for an F_2 population. The difference between BC_1 and F_2 is caused by the fact that with a BC_1 actually the existence of just two QTL genotypes is tested, whereas with an F_2 there would be three QTL genotypes. It should be mentioned that there is some debate about the correctness of the asymptotic approximation (see Titterington et al. 1985, section 5.4). Knapp et al. (1990), for instance, would use 4 df for an F_2.

The LOD score is calculated for positions at regular distance [e.g., 1 centiMorgan (cM)] between the outer markers of each chromosome, always employing the two nearest flanking markers. A QTL likelihood map is constructed by plotting the LOD score against the genome map, an example is shown in Fig. 1. Such a map can be regarded as the likelihood profile for the position of a QTL, although in principle it is a connected series of likelihood profiles for chromosome segments between two neighbouring markers. The fact that LOD scores for neighbouring segments are calculated with one shared marker and one marker different, results in a curve that is angled at the marker locations.

The maximum likelihood estimator of the position of the QTL is the point on the map for which the graph has its maximum. To obtain a sort of a confidence interval of the position of the QTL, a so-called one-LOD support interval is constructed by taking the two positions, left and right of the point estimate of the QTL, that have a LOD score of one less than the maximum (Fig. 1). One LOD less corresponds to a probability of a factor ten less than the most likely position. In a similar way a support interval of an arbitrary LOD value can be constructed.

The location of the QTL is of course of primary interest. Secondary, but nonetheless important, are the additive genetic effect and the dominance deviation of the QTL. For these estimates the values of μ_0, μ_1, and μ_2 at the estimated position of the QTL are used. Only the F_2 allows for estimation of additive and dominance effects; in the case of the BC_1 the dominance effect cannot be determined, whereas the estimated additive effect is biased by dominance, if present.

The model assumes normal distributions as the mixing components. However, if the number of genes involved in the studied trait is small, say less than five, then the normality assumption may not be correct, especially if the heritability is very high (say >0.75). In such a situation the genetic part of the residual variation is multinomial (van Ooijen 1989). It is assumed that in such a case QTLs can be mapped initially with the present method, resulting in few QTLs with relatively large estimated effects, and afterwards an appropriate model may be fitted.

In practice, there will always be individuals for which the marker genotype can not be determined, mostly because of technical difficulties in the laboratory. The method, described above, can readily be extended for these missing marker values. In such a case the likelihood contribution of an individual for a given position in the genome is based upon the closest known flanking markers. In the EM iterations the $\alpha_q(\cdot)$'s have to be based on these markers. If only a marker on one side happens to be known, the likelihood contribution can only be based on this one marker, and appropriate mixing proportions (π_{mq}) and $\alpha_q(\cdot)$'s will have to be calculated. If in the extreme case no flanking markers are known, one might either not use this individual in the QTL mapping procedure, since it will not be very informative, or else use the expected segregation ratios of a QTL independent of a marker (0: ½: ½ for a BC_1, or ¼: ½: ¼ for an F_2) as the mixing proportions.

Methods

Because of the large number of factors that influence the accuracy of QTL mapping, we restrict ourselves to a few relatively simple cases. A first generation backcross (BC_1) and an F_2 population are studied. The cross parents are homozygous. The population size is 100, 200, or 400 individuals. Only one chromosome with a single segregating QTL is simulated, with an additive but no dominance effect. The random non-genetic factors follow a normal distribution with zero mean and an appropriate variance σ_e^2. The random non-genetic variance was chosen such that the QTL explained 1%, 5%, or 10% of the total variance, i.e., genetic plus non-genetic variance equals 100%. For a population size of 100, and for an explained variance of 1% the simulation consisted of 1000 replications, and 500 replications otherwise.

Chromosome length is 120 cM, which is the average chromosome length of tomato (*Lycopersicon esculentum*), a crop spe-

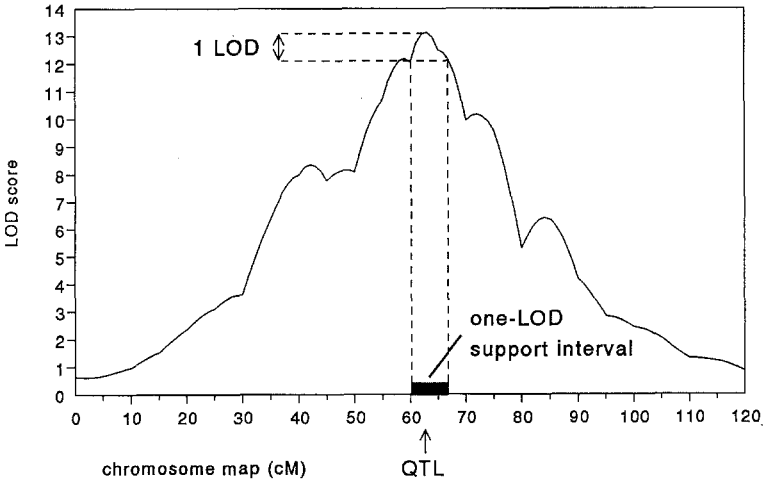

Fig. 1. Example of a QTL likelihood map. Result of a simulated trial with an F_2 population of 400 individuals, a QTL (at map position 62.5 cM) that explains 10% of the total variance, and a segregating marker every 5 cM. The construction of a one-LOD support interval is demonstrated

cies of major mapping interest. The average chromosome length for various studied plant species is roughly 100 cM (O'Brien 1990). Interference is assumed to be absent. There is a segregating marker locus every 5 cM (starting at position 0 cM) and the map positions are assumed to be known precisely. In practice one would not determine all marker loci at once. Initially one would start with markers, say, every 20 cM, and when the LOD score tends to be significant, additional markers would be employed. In this computer simulation, however, all markers are determined. The position of the QTL is in the middle between two markers at 62.5 cM, presumably the worst possible location between two markers, because when the QTL is closer to one marker the power of the mapping is assumed to be higher.

A model for a single QTL is fitted, as described in the previous section. The LOD score is calculated at positions every 1 cM. When a maximum LOD score for a chromosome exceeds the significance threshold, support intervals are constructed of 0.5, 1, 2, and 3 LOD. The significance threshold has been determined beforehand, also by simulation (see below). Concerning the support intervals the following observations were made: (1) whether the interval enclosed the QTL, and (2) the length of the interval.

LOD significance threshold

As has been mentioned above, when the null hypothesis is true, the LOD score for a single test behaves as a chi-square random variable (multiplied by a certain constant). However, in the making of a QTL likelihood map a series of correlated tests are performed, and not just a single test, the correlation being caused by linkage. Such a series is, in fact, a stationary stochastic process. In order to obtain a significance threshold for the maximum of a QTL likelihood map, one needs the distribution of the maximum value of the corresponding stationary process. Lander and Botstein (1989; proposition 2) present an approximating equation to obtain a critical value of the maximum LOD score for a backcross progeny, but not for an F_2. Therefore, the probability distribution function of the maximum LOD score (of one chromosome), when H_0 is true, was obtained by simulation. For both a BC_1 and an F_2 a population of 100 individuals was simulated, analogous to the method described above, except that the QTL was missing. In each replication the QTL likelihood map was calculated, and the maximum LOD score for the chromosome determined. The results, based on 16 000 replications, are presented in Fig. 2. Table 3 and Fig. 3 give the right hand tail of the distributions.

If we are to simulate tomato with its twelve chromosomes, and we want an overall significance level of 0.05, then per (independent) chromosome we need a significance level of $1-\sqrt[12]{(1-0.05)} = 1-0.9957 = 0.0043$. According to Table 3 the corresponding LOD thresholds are 3.0 and 3.7 for BC_1 and F_2, respectively. The value of BC_1 is somewhat larger than indicated by Lander and Botstein (their Fig. 4), presumably due to a larger chromosome length (120 vs 100 cM). The LOD threshold for an F_2 is larger than for a BC_1 due to the fact that the LOD score for an F_2 has two degrees of freedom versus one for a BC_1.

Results

The fraction of the replications that resulted in a significant maximum LOD score, in fact the probability of detecting the QTL, is shown in Table 4. It is clear that this depends on the number of individuals in the test, and on the size of the genetic effect of the QTL. Only when a population of at least 200 individuals is employed, is there a reasonable chance of detecting a QTL that explains 5% or more of the total variance. The results also indicate that these fractions are slightly smaller for an F_2 than for a BC_1 in situations with the same number of individuals and the same explained variance of the QTL. This might be due to an inappropriate LOD threshold, but the number of replications (16 000) used in its determination would seem adequate, especially when viewing the smoothness of the graphs of the distribution functions (Fig. 3). So, apparently, the power to detect a QTL, when there is one, is slightly less for an F_2 than for a BC_1. This must be due to the fact that three mixing components have to be estimated in an F_2 vs two in a BC_1. However, these situations are not directly comparable, since a QTL that explains a certain fraction of the total variance in a BC_1, explains more than that fraction in an F_2. The additive genetic variance of a QTL in an F_2 is twice that

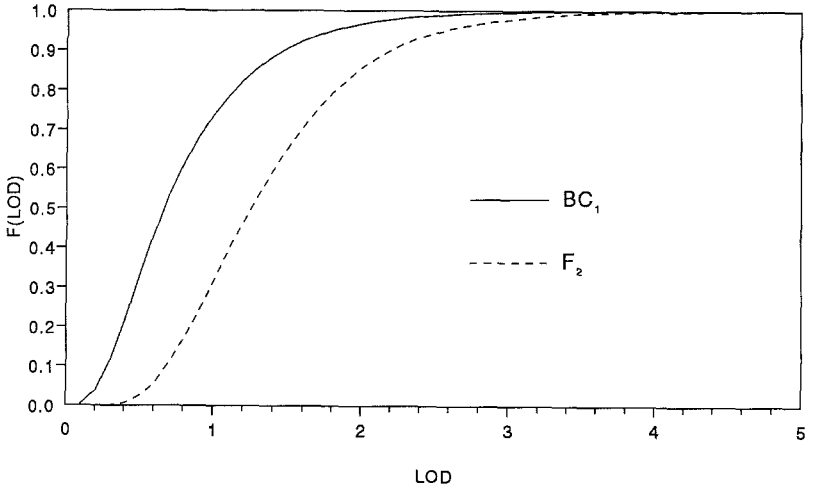

Fig. 2. Cumulative distribution function of the maximum LOD score of a chromosome of 120 cM length with no QTL segregating and a segregating marker locus every 5 cM, for a BC_1 and an F_2 progeny. Determined by 16 000 simulated trials

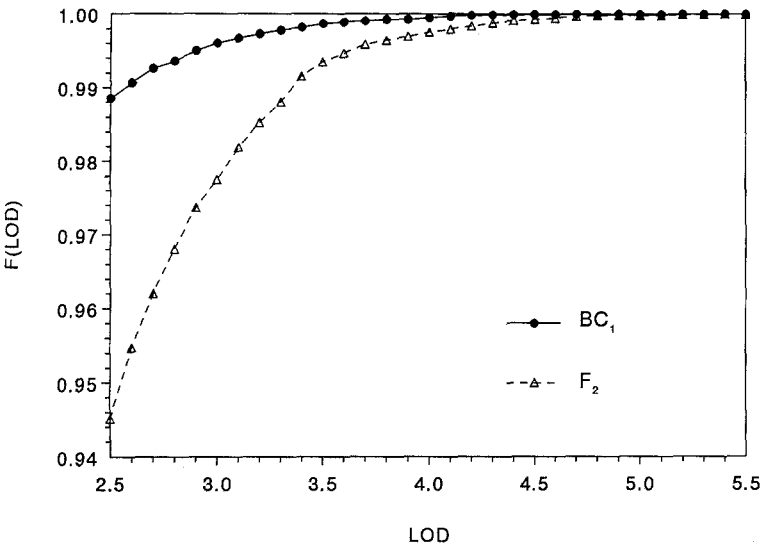

Fig. 3. Upper tail of the cumulative distribution function of the maximum LOD score of a chromosome of 120 cM length with no QTL segregating and a segregating marker locus every 5 cM, for a BC_1 and an F_2 progeny. Determined by 16 000 simulated trials

Table 3. Upper tail of the cumulative distribution function of the maximum LOD score of a chromosome of 120 cM with no QTL segregating and every 5 cM a segregating marker locus, $F_{BC1}(LOD)$ and $F_{F2}(LOD)$, respectively, for a BC_1 and an F_2 progeny. Determined by 16 000 simulated trials

LOD	$F_{BC1}(LOD)$	$F_{F2}(LOD)$	LOD	$F_{BC1}(LOD)$	$F_{F2}(LOD)$
2.0	0.96713	0.85638	3.7	0.99906	0.99588
2.1	0.97444	0.87950	3.8	0.99925	0.99644
2.2	0.97800	0.89981	3.9	0.99925	0.99694
2.3	0.98238	0.91850	4.0	0.99950	0.99756
2.4	0.98525	0.93331	4.1	0.99969	0.99800
2.5	0.98856	0.94513	4.2	0.99981	0.99838
2.6	0.99063	0.95475	4.3	0.99988	0.99881
2.7	0.99263	0.96206	4.4	0.99988	0.99913
2.8	0.99363	0.96806	4.5	0.99994	0.99931
2.9	0.99500	0.97369	4.6	0.99994	0.99944
3.0	0.99606	0.97750	4.7	0.99994	0.99969
3.1	0.99669	0.98188	4.8	0.99994	0.99981
3.2	0.99731	0.98525	4.9	0.99994	0.99981
3.3	0.99781	0.98806	5.0	0.99994	0.99981
3.4	0.99819	0.99156	5.1	0.99994	0.99981
3.5	0.99869	0.99350	5.2	1.00000	0.99994
3.6	0.99888	0.99469	5.3	1.00000	1.00000

Table 4. Fraction of the simulations that obtained a significant LOD score. N, population size; σ^2_{exp}, fraction of the total variance explained by the QTL

N	σ^2_{exp}					
	BC_1			F_2		
	0.01	0.05	0.10	0.01	0.05	0.10
100	0.01	0.11	0.41	0.02	0.06	0.31
200	0.02	0.41	0.87	0.02	0.29	0.79
400	0.07	0.84	1.00	0.05	0.76	1.00

variance in a BC_1, which leads Lander and Botstein (1989) to the conclusion that an F_2 is nearly twice as powerful for detecting a certain QTL as a BC_1. However, this two-fold genetic variance in an F_2 holds for the other QTLs as well (if there is more than one QTL). Therefore, the total variance (due to all QTLs plus non-genetic variance) also increases substantially, while the non-genetic variance remains the same. Hence, the fraction of the variance explained by a single QTL in an F_2 will be less than twice this fraction, down to even the same fraction, in a BC_1, depending on the total number of QTLs and the proportion of their genetic effects. Table 5 demonstrates a few examples of the relation between the fraction explained variance in a BC_1 and an F_2. The effect of dominance is to reduce or enlarge the explained variance in the BC_1 depending on whether the backcross is to the dominant or the recessive homozygote, respectively, whereas in an F_2 both homozygotes are always present. Thus, since the comparison of BC_1 with F_2 depends particularly on the number of QTLs determining the trait, this has to be based upon the explained variance of the QTL, which also stresses the fact that the mapping procedure is a method of segregating a mixture of probability distributions.

When a significant maximum LOD score was obtained, support intervals were constructed for four different LOD levels, and the frequency with which the actual QTL was enclosed was observed (Table 6). This also depends on the number of individuals and on the fraction of the variance explained by the QTL. Since the support intervals serve as a confidence interval, one requires a 95% confidence rate, as is normal in biological research. In that case it appears that a support level of two LOD is necessary for the simulated situations.

The length of the support interval is of interest. For breeding purposes it is important to know down to which

Table 5. Relation between the fraction of the total variance explained by a QTL in a BC_1 and this fraction in an F_2 depending on the total number (nr) of unlinked QTLs with an additive effect of the same size. Dominance is absent

BC_1	nr	F_2	BC_1	nr	F_2	BC_1	nr	F_2
0.01	1	0.0198	0.05	1	0.0952	0.10	1	0.1818
	2	0.0196		2	0.0909		2	0.1667
	3	0.0194		3	0.0870		3	0.1538
	4	0.0192		4	0.0833		4	0.1429
	5	0.0190		5	0.0800		5	0.1333
	10	0.0182		10	0.0667		10	0.1000

Table 6. Fraction of the support intervals that enclosed the QTL, N, population size; LOD, level of the support interval; σ^2_{exp}, fraction of the total variance explained by the QTL

N	LOD	σ^2_{exp}			
		BC_1		F_2	
		0.05	0.10	0.05	0.10
200	0.5	0.59	0.74	0.57	0.66
	1.0	0.78	0.90	0.73	0.85
	2.0	0.94	0.99	0.92	0.98
	3.0	1.00	1.00	0.99	0.99
400	0.5	0.70	0.82	0.66	0.77
	1.0	0.84	0.95	0.84	0.94
	2.0	0.96	0.99	0.96	0.99
	3.0	1.00	1.00	0.99	1.00

Table 7. Mean and standard deviation (in brackets) of the length of the support interval. Units, cM; N, population size; LOD, level of the support interval; σ^2_{exp}, fraction of the total variance explained by the QTL

N	LOD	σ^2_{exp}			
		BC_1		F_2	
		0.05	0.10	0.05	0.10
200	0.5	10 (4)	9 (4)	9 (3)	9 (3)
	1.0	17 (8)	15 (6)	16 (6)	14 (6)
	2.0	37 (15)	29 (13)	33 (13)	28 (12)
	3.0	70 (26)	50 (24)	60 (23)	47 (22)
400	0.5	9 (4)	7 (3)	9 (3)	7 (2)
	1.0	15 (7)	11 (4)	14 (6)	11 (4)
	2.0	31 (15)	18 (8)	27 (13)	19 (8)
	3.0	53 (26)	27 (12)	47 (22)	28 (13)

range on a chromosome a QTL can be located. If the range is large, it may contain more interesting loci, and hence further research would be needed to obtain a better resolution of the area. For the molecular biologist it is important to know if it is practical to consider cloning the gene, bearing in mind, of course, the variability in the relation between linkage and physical distance. Table 7 presents the average and standard deviation of the length of the support intervals. Support intervals of LOD level 0.5 or 1 show reasonably short lengths, but as we have seen above, they often do not enclose the QTL. If we use a safe support level of two LOD, we can expect a length of around 20 to 40 cM. However, if we look at the standard deviations in Table 7, it is clear that these lengths are rather variable.

Discussion

We have seen that the QTL mapping procedure of Lander and Botstein (1989) enables one to determine a map position of QTLs quite well. Of course there are limitations. QTLs with a small additive effect ($\sigma^2_{exp} = 1\%$) are very unlikely to be detected. A population size of at least 200 individuals is necessary, unless one is only interested in genes with a very large effect ($\sigma^2_{exp} > 10\%$). A population size of 400 individuals seems currently the largest practically feasible with respect to the RFLP side of the work. Therefore, it can be expected that with this mapping procedure QTLs with an explained variance of at least 5% stand a good chance of being detected. Paterson et al. (1991) in their mapping experiment with an F_2 of 350 individuals found various QTLs for a number of traits, of which the estimated explained variance was always above 4%. This is in agreement with our results, although this fraction will probably stand in relation to the applied significance threshold of the LOD score and the population size. In their paper, amongst other things, the effect of environment on QTL expression was investigated. Only four genes were detected that were expressed in the three environments studied, while 25 others were expressed in two or one environment. This might indeed reflect genotype × environment interaction. However, the residual variation caused by random non-genetic factors (environment) may differ largely over environments, and also the size of replication (in this case the number of F_3 progeny) influences the residual non-genetic variation. The magnitude of the non-genetic residual variation directly influences the explained variance of a QTL, and thus its LOD score. So, a comparison across environments definitely needs good estimates of the non-genetic residual variance in all environments. These estimates must be employed, in one way or another, in the assessment of the QTL likelihood maps; simply applying the same LOD significance threshold does not seem to be correct.

As mentioned in the introduction, the results of the fine mapping by Paterson et al. (1990) were not satisfactory. A non-significant effect of chromosome 1 on soluble solids in their 1988 study (Paterson et al. 1988) had become significant with their fine mapping technique; the

size of the estimated additive effect being approximately the same. However, in the light of the present simulation results, it would appear that the non-significant peak in the soluble solids LOD score in the 1988 study might just as well have been an artefact, because it is well below the LOD significance threshold. It remains unclear, why the very significant effect of chromosome 1 on fruit mass is not detected with the fine mapping experiment. If we construct a two-LOD support interval, it seems very possible, that the responsible QTL lies just outside the region of the available markers.

The precision of estimating the location of the QTL may be somewhat disappointing, especially for gene cloning purposes. For these purposes fine mapping techniques, such as studying near-isogenic lines, will be necessary to obtain a better resolution. But for breeding purposes the precision seems adequate, although this will always depend on the importance of the other genes in the area of the mapped gene. Of course, genes with a genotypic effect larger than the ones studied in this paper will be mapped with greater precision.

Only a few relatively simple cases have been studied so that many questions still remain. An important one is, what would be found if two segregating QTLs are located on the same chromosome. In contrast to other methods, this mapping procedure with its QTL likelihood map, may, at least theoretically provide some insight into the possibility of two segregating QTLs. An example is given by Lander and Botstein (1989, Fig. 3), in which the QTLs are 80 cM apart. If the QTLs are closer to each other they will behave more like one locus with an added effect of the linked alleles. In the case where the loci are in repulsion in the F_1, the possibility exists that not even one QTL will be discovered. Much attention is needed on the strategy to be followed for fitting models with multiple QTLs.

Another important question relates to the fact that the analysis is based on a marker linkage map, which is accurately known from many previous experiments. For current mapping experiments the linkage map is often estimated from the same individuals on which the QTL mapping is done. Such a map might be less precise, but it will relate better to the recombination events in the cross. Probably there is no such thing as "the map" of a crop species, but rather each cross (especially wide crosses) may have its own map with some areas with higher and other areas with lower recombination rates, although the order of markers will be unchanged. Thus, while a map based on many previous experiments may be precise, it may also be biased for the cross that is being analyzed. The question is whether this has a large impact on the quality of the QTL mapping procedure.

As mentioned before, in practice there will always be missing data on the marker genotypes. Since this leads to lower LOD scores, and since the amount of missing data will vary across markers, this has an effect on the comparability of LOD scores on different parts of a chromosome, unless that amount is not excessive. More or less the same applies to dominant markers. Dominance of markers will lead to lower LOD scores, and hence comparing LOD scores based on dominant, with LOD scores based on co-dominant markers becomes difficult.

Another point of attention is the rather concave behaviour of the QTL likelihood map often observed between two neighbouring markers, which does affect the corresponding estimates of means and residual variance. Examples of this can be seen in Fig. 1, but it can be much more serious in cases with markers further apart (data not shown). Since the positions on the map coinciding with markers give unbiased analysis of variance estimates of means and residual variance, this indicates a possibly large bias at the positions in between two neighbouring markers. This might have consequences for the estimation of the position of the QTL, especially when markers are not equally spaced in a region with a segregating QTL.

There are two final remarks to make. One is that if a QTL is estimated to lie near the end of a chromosome, it will be very convenient to know that the marker at the end of the map is genuinely telomeric, thus ensuring that the QTL support interval does not extend beyond this marker. Thus, telomeric markers are very important. The second remark is that many important quantitative traits are not normally distributed. For instance, many diseases are scored on some ordinal scale, and can be treated as ordered categorical data. A threshold model (e.g., Falconer, 1981) can be incorporated into the described QTL mapping prodecure; threshold models can be fitted to data by maximum likelihood (Jansen 1991). Another solution would be to use a nonparametric method at each marker, such as the Wilcoxon rank-sum test when there are only two marker classes (BC_1), or else use the Kruskal-Wallis test.

Acknowledgements. The author thanks Hans Jansen of CPRO-DLO Wageningen for his help on the EM algorithm and for comment on the manuscript. He also acknowledges drs Piet Stam and Pim Zabel, both of the Agricultural University Wageningen, and dr Mart van Grinsven of Zaadunie Enkhuizen for helpful comments and suggestions on the manuscript. This research was cosponsored by a group of eight Dutch tomato breeding companies.

Appendix

In order to solve the derivatives (2) for θ we need the derivatives of the logarithm of the pdf $f_q(x)$:

$$f_q(x) = \frac{1}{\sqrt{2\pi\sigma_r^2}} \exp\left[\frac{(x-\mu_q)^2}{-2\sigma_r^2}\right],$$

so:

$$\ln f_q(x) = -\tfrac{1}{2}\ln 2\pi - \tfrac{1}{2}\ln \sigma_r^2 + \frac{(x-\mu_q)^2}{-2\sigma_r^2}.$$

The derivatives are:

$$\frac{\partial \ln f_q(x)}{\partial \mu_{q'}} = \frac{x-\mu_q}{\sigma_r^2} \quad \text{if} \quad q = q',$$

and:

$$\frac{\partial \ln f_q(x)}{\partial \mu_{q'}} = 0 \quad \text{if} \quad q \neq q',$$

and:

$$\frac{\partial \ln f_q(x)}{\partial \sigma_r^2} = \frac{-1}{2\sigma_r^2} + \frac{(x-\mu_q)^2}{2(\sigma_r^2)^2}.$$

When we substitute these derivatives into equation (2) we get:

$$\frac{\partial \ln L}{\partial \mu_q} = \sum_{i=1}^{N}\left[\alpha_q(x_i|m_i;r_a)\frac{x_i-\mu_q}{\sigma_r^2}\right],$$

and:

$$\frac{\partial \ln L}{\partial \sigma_r^2} = \sum_{i=1}^{N}\sum_{q=0}^{2}\left\{\alpha_q(x_i|m_i;r_a)\left[\frac{-1}{2\sigma_r^2} + \frac{(x_i-\mu_q)^2}{2(\sigma_r^2)^2}\right]\right\}.$$

When these equations are set equal to zero, it is easy to obtain the solutions (3).

References

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Statist Soc Ser B 39:1–38

Falconer DS (1981) Introduction to quantitative genetics, 2nd edn. Longman, London

Helentjaris T (1987) A genetic map for maize based on RFLPs. Trends Genet 3:217–221

Jansen J (1991) Fitting regression models to ordinal data. Biom J 33:807–815

Jensen J (1989) Estimation of recombination parameters between a quantitative trait locus (QTL) and two marker gene loci. Theor Appl Genet 78:613–618

Keim P, Diers BW, Olson TC, Shoemaker RC (1990) RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics 126:735–742

Knapp SJ, Bridges WC, Birkes D (1990) Mapping quantitative trait loci using molecular marker linkage maps. Theor Appl Genet 79:583–592

Lander ES, Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Luo ZW, Kearsey MJ (1989) Maximum likelihood estimation of linkage between a marker gene and a quantitative trait locus. Heredity 63:401–408

McLachlan GJ, Basford KE (1988) Mixture models: inference and applications to clustering. Marcel Dekker, New York

O'Brien SJ (ed) (1990) Genetic maps, 5th edn. Cold Spring Harbor Laboratory Press, New York

Ooijen JW van (1989) The predictive value of estimates of quantitative genetic parameters in breeding of autogamous crops. PhD-thesis, Agricultural University, Wageningen

Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature 335:721–726

Paterson AH, DeVerna JW, Lanini B, Tanksley SD (1990) Fine mapping of quantitative trait loci using selected overlapping recombinant chromosomes, in an interspecies cross of tomato. Genetics 124:735–742

Paterson AH, Damon S, Hewitt JD, Zamir D, Rabinowitch HD, Lincoln SE, Lander ES, Tanksley SD (1991) Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics 127:181–197

Sax K (1923) The association of size differences with seed-coat pattern and pigmentation in *Phaseolus vulgaris*. Genetics 8:552–560

Thoday JM (1961) Location of polygenes. Nature 191:368–370

Titteringon DM, Smith AFM, Makov UE (1985) Statistical analysis of finite mixture distributions. Wiley, New York

Weller JI (1986) Maximum likelihood techniques for the mapping and analysis of quantitative trait loci with the aid of genetic markers. Biometrics 42:627–640

Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218

Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535

Zamir D, Tanksley SD (1988) Tomato genome is comprised largely of fast-evolving, low copy-number sequences. Mol Gen Genet 213:254–2617