QTL Mapping

Z-B Zeng, North Carolina State University, Raleigh, NC, USA

© 2001 Elsevier Inc. All rights reserved.

This article is reproduced from the previous edition, volume 3, pp 1587–1593, © 2001, Elsevier Inc.

Introduction

Quantitative trait locus (QTL) mapping is a genome-wide inference of the relationship between genotype at various genomic locations and phenotype for a set of quantitative traits in terms of the number, genomic positions, effects, and interaction of QTL. The primary purpose of QTL mapping is to localize chromosomal regions that significantly affect the variation of quantitative traits in a population. This localization is important for the ultimate identification of responsible genes and also for our understanding of genetic mechanisms of the variation.

Mapping QTL can also help us to understand how many QTL significantly contribute to the trait variation in a population. How much variation is due to the additive effects of QTL and how much is due to dominant and epistatic effects of QTL? What is the nature of genetic correlation between different traits in a genomic region, pleiotropy, or close linkage? Do QTL interact with environments? These questions are related to the genetic architecture of quantitative traits in the population, and are intimately related to many applications in quantitative genetics, such as marker-assisted prediction or selection and marker-assisted gene introgression.

Data for mapping QTL consist of types of a number of polymorphic genetic markers and quantitative trait values for a number of individuals. Marker data are categorical and can be classified into different categories and recorded in digital form, such as 1 or 0 for the presence or absence of a particular molecular band at a particular marker, or the two marker genotypes (homozygote and heterozygote) for a backcross population from two inbred lines. Based on segregation analysis, these markers can be ordered in linkage groups or linearly on chromosomes to represent a genetic linkage map. Quantitative trait data are usually continuous, such as body weight, but can also be discrete, such as litter size. While marker data contain information about segregation of a genome in a population, quantitative trait data contain information about the variation of traits in the population. The two data sets are connected by QTL. A part of the trait variation is caused by the segregation of QTL which are linked to some of the markers in the genome. So the statistical task of mapping QTL is to relate quantitative trait variation to genetic marker variation in terms of a quantitative genetic model that includes many genetic architecture parameters such as number, positions, effects, and interactions of genes that affect the quantitative traits of interest.

Traditional experimental designs for locating QTL start with two parental lines differing both in trait values and in the marker variants they carry. Suppose two pure-breeding lines, p_1 and p_2, have marker genotypes MN/MN and mn/mn for two markers. Crossing these lines produces f_1 offspring that is doubly heterozygous. It is denoted as MN/mn, where the slash separates the contributions from the two parents. Each f_1 individual can produce four possible gametes, or marker allele combinations for transmission to the next generation. The proportions of these four gametes MN, Mn, mN, and mn are $(1-r_{MN}/2$, $r_{MN}/2$, $r_{MN}/2)$ and $(1-r_{MN}/2)$, respectively, where r_{MN} is the recombination frequency between the two markers. This segregation of gametes can be observed, for example, from backcross populations b_1 and b_2 ($b_1=f_1\times p_1$ and $b_2=f_1\times p_2$) and also f_2 populations ($f_2=f_1\times f_1$). If a number of genetic markers and quantitative traits are observed in these and other populations, mapping can be performed to locate QTL.

One-Marker Analysis

The simplest method of associating markers with quantitative trait variation is to test for trait value differences between different marker groups of individuals for a particular marker. For example, let $\tilde{\mu}_{M/M}$ and $\tilde{\mu}_{M/m}$ be the observed trait means of the groups of individuals with marker genotypes M/M and M/m for a marker in a backcross population, we can test for significance between means $\tilde{\mu}_{M/M}$ and $\tilde{\mu}_{M/m}$ using the usual t test with the statistic $t=(\tilde{\mu}_{M/m}-\tilde{\mu}_{M/m})/\sqrt{s^2((1/n_{M/M})+(1/n_{M/m}))}$, where s^2 is the pooled sampling variance and $n_{M/M}$ and $n_{M/m}$ are corresponding sample sizes in each marker class. The hypotheses to be tested can be H_0: $\mu_{M/M}=\mu_{M/m}$ and H_1: $\mu_{M/M}\neq\mu_{M/M}$.

To understand the relevance of this test to QTL mapping, we need to know what exactly is tested in genetic terms. Suppose that there are m QTL contributing to the genetic variation in a backcross population from two inbred lines. Ignoring epistasis, the expected difference between $\tilde{\mu}_{M/M}$ and $\tilde{\mu}_{M/M}$ is $\varepsilon(\tilde{\mu}_{M/M}-\tilde{\mu}_{M/m})=\sum_{i=1}^{m}(1-2r_i)a_i$, where ε denotes expectation, a_i is the effect of the ith QTL expressed as a difference in effects between the recurrent parent homozygote and the heterozygote, and r_i is the recombination frequency between the marker and the ith QTL. Essentially, this means that we test a composite parameter that constitutes gene effects and recombination frequencies for (potentially) a number of genes. Of course, many QTL may not be linked to the marker, and thus have 0.5 recombination frequency. The above hypotheses are then equivalent to H_0: all r_i = 0.5 and H_1: at least one $r_i<0.5$, because the a_i's are usually nonzero by experimental design. If $\tilde{\mu}_{M/M}$ and $\tilde{\mu}_{M/M}$ are found to be significantly different, we conclude that the marker is linked to one or possibly more QTL. This analysis, however, cannot determine whether a significant marker effect is due to one or multiple QTL and whether the effect is due to distantly linked QTL with large effects or closely linked QTL with small effects. With a dense linkage map, the second problem can be alleviated.

 doi:10.1016/B978-0-12-374984-0.01248-1

Interval Mapping

Because single-marker analysis cannot separate r and a in test and estimation even when there is only one QTL on a chromosome, Lander and Botstein (1989) proposed a maximum likelihood method that uses a pair of adjacent markers to test the effect of a genomic position within a chromosomal interval bracketed by two adjacent markers. This is an attempt to disentangle r and a in analysis. This method is called interval mapping. Specifically, for a backcross population, they proposed the following linear model to test for a QTL located on an interval between two adjacent markers $y_j = \mu + b * x_j^* + e_j$ for $j = 1, 2, \ldots, n$, where y_j is a quantitative trait value of the jth individual, μ is the mean of the model, x_j^* is an indicator variable, taking a value 1 or 0 for the two possible QTL genotypes with probability depending on the genotypes of markers and the genomic position being tested, $b*$ is the effect of the putative QTL, e_j is a residual variable (usually assumed to be normally distributed with mean zero and variance σ^2), and n is the sample size. Since x_j^* is usually unobserved for a particular genomic position and can take different values, statistically this is a mixture model. The likelihood function of the model is as follows:

$$L(\mu, b*, \sigma^2) = \prod_{j=1}^{n} \left[p_{1j}\varphi\left(y_j|\mu + b*, \sigma^2\right) + p_{0j}\phi\left(y_j|\mu, \sigma^2\right)\right]$$

where $\phi(y_j|\mu, \sigma^2)$ is a normal density function of y_i with mean μ and variance σ^2, and p_{kj} is the probability of $x_j^* = k$, given marker data and the testing position of the putative QTL.

The test statistic can be constructed using a likelihood ratio (LR) $\mathrm{LR} = -2\ln\left(L(\hat{\hat{\mu}}, b^* = 0, \hat{\hat{\sigma}}^2)/L(\hat{\mu}, b^*, \hat{\sigma}^2)\right)$ to compare the null hypothesis H_0: $b* = 0$ with the alternative hypothesis H_1: $b* \neq 0$, assuming that the putative QTL is located at the point of consideration, where $\hat{\mu}, b^*,$ and $\hat{\sigma}^2$ are the maximum likelihood estimates of μ, $b*$, and σ^2 under H_1, and $\hat{\hat{\mu}}, \hat{\hat{\sigma}}^2$ are the estimates of μ, σ^2 under H_0 with $b*$ constrained to zero.

In human linkage analysis, the LR test statistic, however, has traditionally been expressed in terms of LOD (for log odds) score $\mathrm{LOD} = -\log_{10}(L(\hat{\hat{\mu}}, b^* = 0, \hat{\hat{\sigma}}^2))/(L(\hat{\mu}, \hat{b}^*, \hat{\sigma}^2))$. Extending this tradition, many QTL mapping analyses also use LOD score as a test statistic. There is a one-to-one correspondence between LR and LOD, and LR can be translated into LOD as follows:

$$\mathrm{LOD} = \frac{1}{2}\,(\log_{10} e)\mathrm{LR} = 0.217\ \mathrm{LR}$$

This test can be performed at any genomic position covered by markers, and thus, the method involves a systematic strategy of searching for QTL. If the LR test statistic at a genomic region exceeds a predefined critical threshold, a QTL is estimated at the position of the maximum test statistic. The estimates of locations and effects of QTL are asymptotically unbiased statistically with this maximum likelihood approach if there is only one QTL on a chromosome.

It is important to determine an appropriate critical threshold for a test statistic above which a QTL can be claimed with a certain confidence. The determination of the critical threshold is based on the distribution of a test statistic under the null hypothesis. This distribution for LR at a given position is generally asymptotically chi-square with a degree of freedom that is equal to the number of parameters under the test. However, because the test is usually performed in the whole genome, there is a multiple testing problem, and the distribution of the maximum LR or LOD score over the whole genome under the null hypothesis becomes very complicated. Theoretical and numerical analyses have indicated that the threshold at 5% significance level over a whole genome is generally between 2 and 3.5 on LOD score, depending on the size of genome, density of markers, sample size, and genetic model. Alternatively, the relevant threshold for a given data set can be estimated numerically from the data by using a permutation test.

The model of interval mapping is relatively simple in terms of genetics. Because of it, it has a critical problem that, if there are two or more QTL on a chromosome, the test statistic at a genomic position will be affected by all those linked QTL. Therefore, the estimated positions and effects of 'QTL' identified by this method can be biased. Moreover, some genomic regions which do not contain QTL can still show a significant peak on LOD score if there are multiple QTL in the nearby regions. This is the so-called 'ghost' gene phenomenon. This defect is similar to the defect in single-marker analysis that is discussed above.

Composite Interval Mapping

Ideally, when we test a marker interval for a QTL, we would like to have our test statistic be independent of the effects of possible QTL located in other regions of the chromosome. If such a test can be constructed, we can break down the effects of linked QTL by statistical means to avoid the confounding effects of multiple linked QTL in the search for each individual QTL. In other words, we can test independently each interval for the presence of a QTL. Such a test can be constructed by using a combination of interval mapping and multiple regression.

In a multiple regression analysis of a trait on multiple markers or other explanatory variables, each regression coefficient is a partial regression coefficient conditional on other variables fitted in the model. Largely because of the linear structure of genes on chromosomes, a partial regression coefficient of a trait on a marker or a testing position of interest possesses a very important property that the coefficient is expected to depend only on those QTL within an interval that is bracketed by two fitted flanking markers. The flanking markers are fitted in the model as cofactors to block the effects of other possibly linked QTL to the test. This treatment makes the partial regression coefficient independent of QTL effects on other linked or unlinked intervals, and is the basis of composite interval mapping. The linear independence, however, depends on the assumption of no crossing-over interference and no epistasis. Interference and epistasis introduce nonlinearity into the model.

Specifically, to test for a QTL on an interval between two adjacent markers, we can extend the interval mapping model to $y_j = \mu + b * x_j^* + \sum_k b_k x_{jk} + e_j$, where x_{jk} is an indicator variable referring to the genotype of marker k that is selected to control the genetic background, b_k is the partial regression coefficient associated with marker k, and $b*$ now is also a partial regression coefficient associated with the putative QTL. In this case, the likelihood function becomes

$$L(b*, \mathbf{b}, \sigma^2) = \prod_{j=1}^{n} \left[p_{1j}\phi\left(y_j|\mathbf{x}_j\mathbf{b} + b*, \sigma^2\right) + p_{0j}\phi\left(y_j|\mathbf{x}_j\mathbf{b}, \sigma^2\right)\right]$$

where $xjb = \mu + k_{\Sigma} bkxjk$.

An LR test statistic can also be constructed to compare the hypotheses H_0: $b^* = 0$ with H_1: $b^* \neq 0$. However, since b^* is a partial regression coefficient, the null hypothesis is a composite hypothesis, conditional on other partialression coefficients in the model. Thus, the method is called composite interval mapping. Many statistical issues of composite interval mapping were discussed in Zeng (1994).

Like interval mapping, this test can be performed at any position in a genome covered by markers. Thus, it also gives a systematic strategy to search for QTL in a genome. The main advantage of composite interval mapping, as compared with interval mapping, is the ability to separate effects and locations of multiple linked QTL in mapping. This is shown in **Figure 1** as an example. **Figure 1** summarizes the analyses of mapping body weight loci on mouse chromosome *x* from a backcross population (Dragani *et al.*, 1995). The test statistic, LOD score, of the interval mapping and composite interval mapping analyses is plotted against the linkage map location of the chromosome referenced by 14 microsatellite markers. The value of LOD score at each map position indicates the strength of evidence for a QTL at the position. If the LOD score at a genomic region exceeds a predetermined threshold, one or more QTL are indicated in that region.

For the interval mapping, the threshold is 3.3 for the experimental design. By this criterion, the LOD score in the most part of chromosome *x* is above the threshold, and shows significant peaks in several marker intervals. However, not all significant peaks could be interpreted as QTL because of linkage effects, the 'ghost' gene phenomenon, and statistical sampling effects. Although the analysis strongly supports the existence of segregating QTL on chromosome *x*, it is not clear from the interval mapping analysis how many QTL are on the chromosome and where they are located.

The LOD score of the composite interval mapping analysis shows two distinct major peaks. This suggests that there are at least two body weight QTL on chromosome X in the mouse genome, one is mapped near marker *Rp18-rs11* and the other near *DXMIT60*. The two QTL together explain 25% of the phenotypic variance in the mapping population. In this case, the composite interval mapping analysis achieved a much better resolution in mapping QTL.

Multiple Interval Mapping

Composite interval mapping still has some limitations. One limitation is that the analysis can be affected by an uneven distribution of markers in the genome, meaning that the test statistic in a marker-rich region may not be comparable to that in a marker-poor region. It is also difficult to estimate epistasis of multiple QTL and the contribution of multiple QTL to the phenotypic variance. These limitations can be removed if multiple QTL are searched and mapped simultaneously. This is the idea of multiple interval mapping which fits multiple putative QTL, including epistasis, in a model to search, test, and estimate the positions, effects, and interactions of multiple QTL simultaneously.

Multiple interval mapping (Kao *et al.*, 1999; Zeng *et al.*, 1999) consists of four components: (1) an evaluation procedure designed to analyze the likelihood of the data given a genetic model (number, genomic positions, and epistatic pattern of QTL); (2) a search strategy optimized to select the best

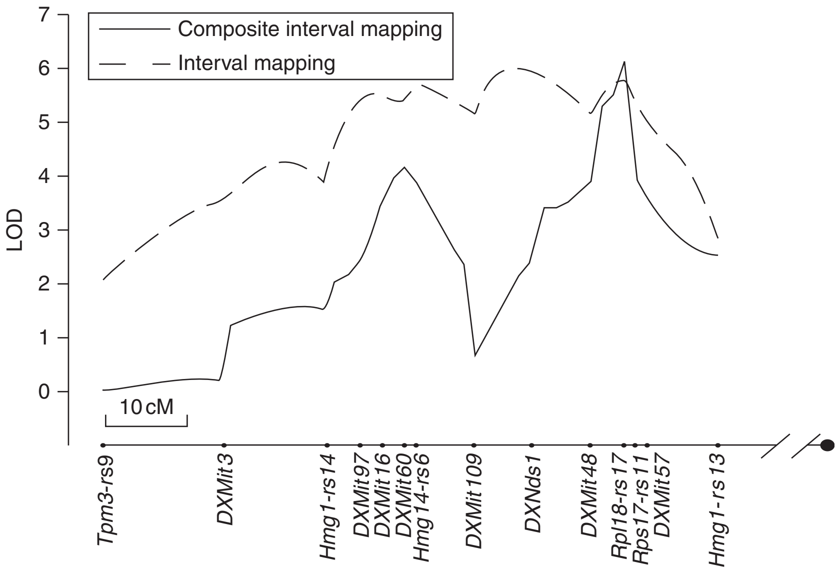

Figure 1 Genetic mapping of body weight loci on chromosome X in a mouse backcross population. LOD score curves of a composite interval mapping analysis (solid curve) and an interval mapping analysis (dashed curve) are shown on a map containing 14 molecular markers. By the interval mapping analysis, it seems that most of the chromosome shows significant effects on body weight. The composite interval mapping analysis strongly indicates that there are two body weight loci on chromosome X segregating in the population.

genetic model (among those sampled) in the parameter space; (3) an estimation procedure for all the genetic parameters of quantitative traits, including the number, positions, effects, and epistasis of QTL, and genetic variances and covariances explained by QTL effects, given the selected genetic models; and (4) a prediction procedure to estimate or predict the genotypic values of individuals and their offspring based on the selected genetic model and estimated genetic parameter values for marker-assisted selection or prediction.

For m putative QTL in a backcross population, the multiple interval mapping model is defined by

$$y_i = \mu + \sum_{r=1}^{m} \alpha_r x_{ir}^* \sum_{r<s\epsilon(1,\ ,m)}^{t} \beta_{rs}(x_{ir}^* x_{is}^*) + e_i.$$

where y_i is the phenotypic value of individual i, while x_{ir}^* is a coded variable denoting the genotype of putative QTL r (defined by 1/2 or − 1/2 for the two genotypes). The variable x_{ir}^* is unobserved, but its conditional probability given observed marker phenotypes can be analyzed. Parameters of the model include the mean (μ), the marginal effects ($\alpha_r{}'s$) and epistatic effects ($\beta'_{rs}\,s$) of the putative QTL, and the variance (σ^2) of the residual effect (e_i, assumed to be normally distributed with mean zero). To avoid overparameterization, a subset of significant pairwise QTL epistatic effects, indicated by $r \neq s\ \varepsilon\ (1,\ldots, m)$, are selected to be included in the model. Since the genotypes of an individual at many genomic locations are not observed (but marker phenotypes are), the model contains missing data, and thus, the likelihood function of the data given the model is a mixture of normal distributions:

$$L(\mathbf{E}, \mu, \sigma^2 | \mathbf{Y}, \mathbf{X}) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi(y_i | \mu + \mathbf{D}_j \mathbf{E}, \sigma^2 \right]$$

The term in square brackets is the weighted sum of a series of normal density functions, one for each of the 2^m possible multiple QTL genotypes. p_{ij} is the probability of each multilocus genotype conditioned on marker data. The QTL parameters (α's and β's) are contained in the column vector $\mathbf{E}$, while the row vector $\mathbf{D}_j$ specifies the configuration of x^*'s associated with each α and β for the jth QTL genotype.

The analysis of the likelihood can be performed through a numerical expectation/maximization (EM) algorithm. The EM algorithm is an iterative procedure involving an E step (expectation) and an M step (maximization) in each iteration. In the $(t+1)$th iteration, the E step is as follows:

$$\pi_{ij}^{[t+1]} = \frac{p_{ij}\phi\left(y_i | \mu^{[t]} + \mathbf{D}_j \mathbf{E}^{[t]}, \sigma^{2\,[t]}\right)}{\sum_{j=1}^{2^m} p_{ij}\phi\left(y_i | \mu^{[t]} + \mathbf{D}_j \mathbf{E}^{[t]}, \sigma^{2\,[t]}\right)}$$

and the M step is as follows:

$$E_r^{[t+1]} = \frac{\sum_i \sum_j \pi_{ij}^{[t+1]} D_{jr} \left[(y_i - \mu^{[t]}) - \sum_{s=1}^{r-1} D_{js} E_s^{[t+1]} - \sum_{s=r+1}^{m+1} D_{js} E_s^{[t]} \right]}{\sum_i \sum_j \pi_{ij}^{[t+1]} D_{jr}^2}$$
$$\text{for } r = 1, \ldots, m+t$$

$$\mu^{[t+1]} = \frac{1}{n} \sum_i \left(y_i - \sum_j \sum_r \pi_{ij}^{[t+1]} D_{jr} E_r^{[t+1]} \right)$$

$$\sigma^{2\,[t+1]} = \frac{1}{n} \left[\sum_i \left(y_i - \mu^{[t+1]} \right)^2 - 2 \sum_i \left(y_i - \mu^{[t+1]} \right) \sum_j \sum_r \pi_{ij}^{[t+1]} D_{jr} E_r^{[t+1]} + \sum_r \sum_s \sum_i \sum_j \pi_{ij}^{[t+1]} D_{jr} D_{js} E_r^{[t+1]} E_s^{[t+1]} \right]$$

where E_r is the rth element of $\mathbf{E}$ and D_{jr} is the rth element of $\mathbf{D}_j$. Given an initial value for parameters $\mathbf{E}$, the algorithm can rotate between E and M steps until the convergency of estimates.

The test for each QTL effect, say E_r, is performed by an LR test conditioned on the other QTL effects:

$$\mathrm{LOD} = \log_{10} \times \frac{L(E_1 \neq 0, \ldots, E_{m+t} \neq 0)}{L(E_1 \neq 0, \ldots, E_{r-1} \neq 0, E_r = 0, E_{r+1} \neq 0, \ldots, E_{m+t} \neq 0)}$$

For given positions of m putative QTL and $m+t$ QTL effects, the likelihood analysis can proceed as outlined above. Then the main task is to search and select the best genetic model (number, genomic positions, and epistatic pattern of QTL) that fits the data well. Search for multiple QTL in a multiple (unknown) dimension space is a very difficult task. Several issues have to be considered and balanced in designing an efficient algorithm for the search process. On the one hand, we need to consider the reliability and robustness of an algorithm and, on the other hand, we also need to consider its efficiency and applicability. Several methods have been used for this process, such as stepwise model selection, genetic algorithms, and Markov chain Monte Carlo.

The stepwise model selection consists of a number of components. There is a search step that searches the genome for the position of new QTL given the current genetic model (a forward step); an epistasis step that searches for significant interaction effects of the newly identified QTL with other QTL in the model (a part of the forward step); an evaluation step that evaluates each QTL effect fitted in the model for significance under the new model and drops any nonsignificant effect (a backward step); an optimization step that optimizes the estimates of genomic position of each QTL fitted in the model under the new model; a stopping rule that determines the termination of the search process; and an estimation step that reports estimates of various genetic architecture parameters, composite genetic parameters, and individual genotypic values. Estimation of individual genotypic values and prediction of offspring genotypic values of two individuals can provide a basis for marker-assisted selection.

As an example, Figure 2 shows the mapping result by multiple interval mapping for a morphological shape difference between two *Drosophila* species (Zeng *et al.*, 2000). Two *Drosophila* species, *D. simulans* and *D. mauritiana*, were crossed to make F_1 hybrids. Because F_1 males are sterile, females of F_1 population were backcrossed to each of the parental species to create two backcrosses. There are about 500 individuals in each backcross. The trait is the morphology of the posterior lobe of the male genital arch analyzed as the first principal component in an elliptical Fourier analysis. After extensive search analysis using multiple interval mapping, the model is stabilized at 19 QTL with six significant epistatic terms in the backcross to *D. mauritiana*. Figure 2 depicts the likelihood profile (LOD score) for each QTL that spans between its neighbors. The peak of each likelihood profile provides an estimate of the position of a QTL on the genetic linkage map.

Mapping QTL is not restricted for backcross and F_2 populations of inbred lines. Mapping methods can be extended and

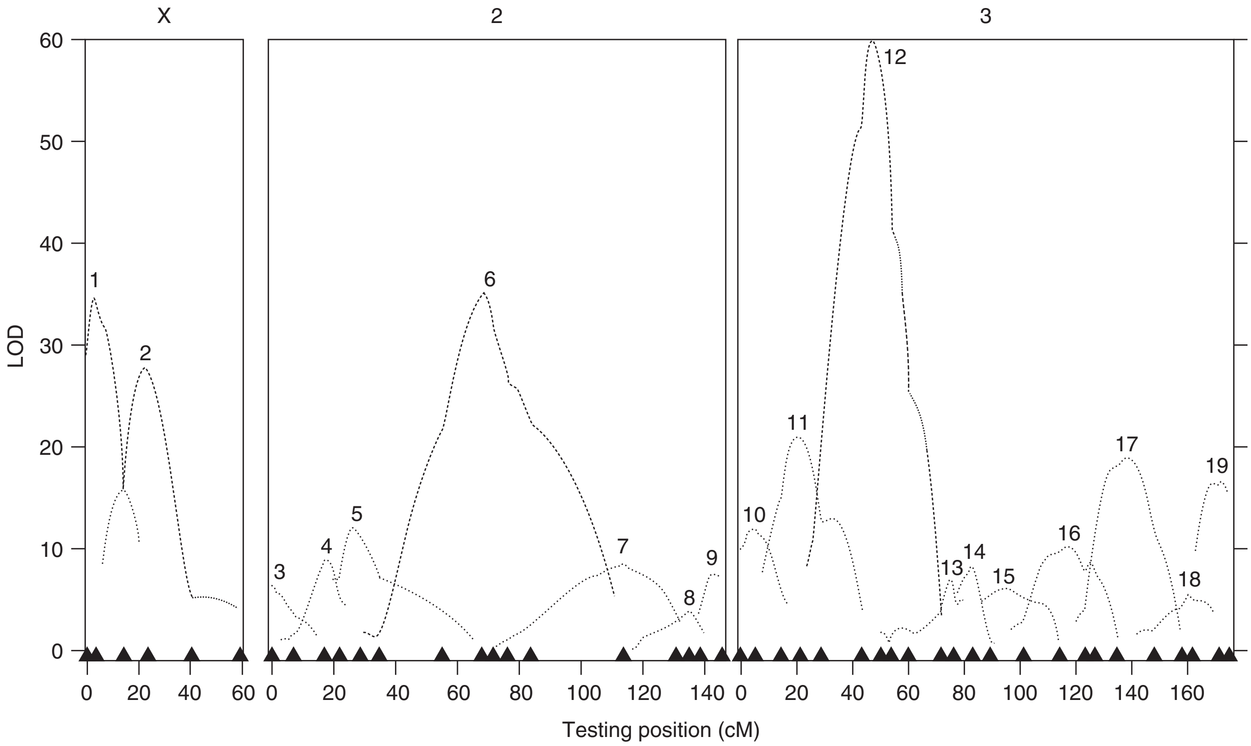

Figure 2 LOD profiles of the 19 putative QTL mapped by multiple interval mapping for chromosomes X, 2 and 3 of *Drosophila*. Marker positions are shown by triangles.

applied to other crosses of populations or species or to segregating populations. For species such as human, the mapping of QTL has to be made with current segregating populations. No matter what population is analyzed, the general idea of QTL mapping analysis is based on the inference of genotypes and an appropriate model that relates a trait to the genotypes or combinations of genotypes at a number of genomic positions. However, for mapping QTL with segregating populations, statistical analyses become much more complicated due to a number of limiting factors in data, such as small family size, unknown linkage phases between markers and QTL, and complicated family structures. Many statistical methods for mapping QTL from segregating populations have been developed. These include, for example, the sib-pair methods, the identity-by-descent mapping, and some Bayesian methods that incorporate Markov chain Monte Carlo algorithms. More studies are needed to generalize these and other methods to make them applicable to a wide variety of populations or experimental designs, data structures, and genetic models.

See Falconer and Mackay (1996) and Lynch and Walsh (1997) for more general discussion on the genetic basis of QTL and on genetic and statistical analyses for mapping QTL.

See also: Linkage Map; Marker; Multifactorial Inheritance; QTL (Quantitative Trait Locus); Quantitative Trait.

References

Dragani TA, Zeng Z-B, Canzian F, *et al.* (1995) Molecular mapping of body weight loci on mouse chromosome X. *Mammalian Genome* 6: 778–781.

Falconer DS and Mackay TFC (1996) *Introduction to Quantitative Genetics*, 4th edn. Harlow: Longman.

Kao C-H, Zeng Z-B, and Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. *Genetics* 152: 1203–1216.

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

Lynch M and Walsh B (1997) *Genetics and Analysis of Quantitative Characters*. Sunderland, MA: Sinauer Associates.

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

Zeng Z-B, Kao C-H, and Basten CJ (1999) Estimating the genetic architecture of quantitative traits. *Genetic Research* 74: 279–289.

Zeng Z-B, Liu J, Stam LF, *et al.* (2000) Genetic architecture of a morphological shape difference between two *Drosophila* species. *Genetics* 154: 299–310.