A COMPOSITE INTERVAL MAPPING METHOD FOR LOCATING MULTIPLE QTLS

Zhao-Bang Zeng
Program in Statistical Genetics, Department of Statistics
North Carolina State University, Raleigh, NC 27695-8203 USA

SUMMARY

Recently we developed a new statistical method for mapping quantitative trait loci (QTLs). Compared to current QTL mapping methods, our method has a number of advantages and can more accurately and efficiently locate multiple QTLs. The method is based on a procedure which combines the interval mapping with an interval test. Theoretically, this method is more complex and versatile than the current methods and allows a host of models be fitted in data to explore fully the complex information in a data set. We have now implemented the method for several commonly used experimental designs and for multiple trait analysis.

INTRODUCTION

Correct identification of genes affecting quantitative trait variation by using molecular marker information is a very important first step to any subsequent marker-assisted selection or gene introgression. There have been several statistical methods and computer programs (notably the interval mapping of Lander and Botstein (1989)) developed to utilize a complete marker linkage map to systematically search major quantitative trait loci (QTLs) in experimental organisms. There are however several problems with the current QTL mapping methods. The major deficiencies of the current QTL mapping methods, including Lander and Botstein's interval mapping method, are as follows: (*i*) The test for a QTL is not formulated as an interval test (a test which should distinguish whether there is a QTL on an interval or not and should be independent of the effects of QTLs at other regions of the chromosome). (*ii*) If there is more than one QTL on a chromosome, the test statistic will be compounded and the estimated positions and effects of the identified "QTLs" by current methods are likely to be biased. (*iii*) It is also not efficient to use only two markers at a time to do the test, as the information from other markers is not utilized.

Recently we have developed a new statistical method of QTL mapping (Zeng, 1993; 1994) to improve the precision and efficiency of mapping multiple QTLs. The method is based on a procedure which combines interval mapping with a multiple regression analysis. The basis of the method is an interval test in which the test statistic on a marker interval is made to be independent of the effects of QTLs located on other regions of the chromosome. This is achieved by fitting other genetic markers in the statistical model as a control when performing interval mapping by using partial regression theory.

In this paper, I briefly discuss some properties and utility of the method, and also summarize some recent developments on the method.

RESULTS AND DISCUSSION

Composite interval mapping: Suppose that we have a data set from a backcross population from two inbred lines with measurements on a quantitative trait and information on t molecular markers in n individuals. To test for a QTL in a marker interval between two markers i and $i+1$, we can use the following linear model to perform the test

$$y_j = b_0 + b^* x_j^* + \sum_{k \neq i, i+1} b_k x_{jk} + e_j \qquad \text{for } j = 1, 2, \cdots, n \tag{1}$$

where y_j is the trait value of the jth individual, b_0 is the mean of the model, b^* is the effect of the putative QTL expressed as a difference in effects between the homozygote and heterozygote, x_j^* is an indicator variable, taking a value 1 or 0 with probability depending on the genotypes of the markers i and $i+1$ and the testing position of the putative QTL, b_k is the partial regression coefficient of the phenotype y on the

kth marker, x_{jk} is the type of the kth marker in the jth individual, taking a value 1 or 0 depending on whether the marker type is homozygote or heterozygote, and e_j is a random variable. The summation of other markers in the model depends on the choice of the model. By fitting different markers in the model, a host of models can be created and different models have different advantages and disadvantages.

Statistically this is a mixture model. Assuming that e_j's are identically and independently normally distributed with mean zero and variance σ^2, the likelihood function is given by $L_1 = \prod_{j=1}^{n}[p_j(1)f_j(1) + p_j(0)f_j(0)]$, where $p_j(1)$ gives a prior probability of $x_j^* = 1$, $p_j(0) = 1 - p_j(1)$, $f_j(1)$ and $f_j(0)$ specify a normal density function for the random variable y_j with a mean $b_0 + b^* + \sum_{k\neq i,i+1} b_k x_{jk}$ and $b_0 + \sum_{k\neq i,i+1} b_k x_{jk}$, respectively, and a variance σ^2. By using the standard maximum likelihood procedures, the maximum likelihood (ML) estimates of the parameters b^*, b_k's and σ^2 are found to be the solutions of

$$\hat{b}^* = (\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'\hat{\mathbf{P}}/\hat{c} \tag{2}$$

$$\hat{\mathbf{B}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{P}}\hat{b}^*) \tag{3}$$

$$\hat{\sigma}^2 = [(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}})'(\mathbf{Y} - \mathbf{X}\hat{\mathbf{B}}) - \hat{c}\hat{b}^{*2}]/n \tag{4}$$

where $\mathbf{Y}$ is a $(n \times 1)$ vector of y_j's, $\hat{\mathbf{B}}$ is a $((t-1) \times 1)$ vector of the ML estimates of b_k's (including b_0 but excluding b^*), $\mathbf{X}$ is an $(n \times (t-1))$ matrix of x_{jk}'s, $\hat{\mathbf{P}}$ is a $(n \times 1)$ vector with elements $\hat{P}_j$ specifying the ML estimate of the posterior probability of $x_j^* = 1$:

$$\hat{P}_j = p_j(1)\hat{f}_j(1)/[p_j(1)\hat{f}_j(1) + p_j(0)\hat{f}_j(0)] \tag{5}$$

and $\hat{c} = \sum_{j=1}^{n} \hat{P}_j$. The prime indicates transposition of a vector or matrix.

These estimates can be found by iterations of the above equations via the ECM algorithm (Meng and Rubin 1993) (ECM stands for Expectation/Conditional Maximization) beginning with the initial estimate $\hat{b}^* = 0$ or the least squares estimates of b^* and $\mathbf{B}$ using $x_j^* = p_j(1)$. In each iteration, the algorithm consists one E-step, Eq. (5), and three CM-steps, Eqs. (2), (3) and (4). The convergence of the algorithm to ML estimates has been proven by Meng and Rubin (1993). The advantage of this algorithm over the full EM algorithm (maximizing $\hat{b}^*$ and $\hat{\mathbf{B}}$ simultaneously in the M step) is that the inverse, $(\mathbf{X}'\mathbf{X})^{-1}$, does not need to be updated, and thus the efficiency of the numerical evaluation is improved substantially.

The test for the hypotheses $H_0 : b^* = 0$ and $H_1 : b^* \neq 0$. is based on the likelihood ratio LR $= -2\log(L_0/L_1)$, where L_0 and L_1 represent the likelihood values under H_0 and H_1. As shown by Zeng (1993), this test is an interval test, a test with the test statistic independent of the effects of possible QTLs located outside a defined region on the chromosome being tested. Theory and properties of such a test were established and explained by Zeng (1993).

Like Lander and Botstein' interval mapping, this test can be performed at any position in a genome. Thus it creates a systematical strategy to search for QTLs in a genome. As the test statistic is almost independent for each interval, a test on each interval is more likely to test for a single QTL only.

There are several advantages of this method in comparison with the current QTL mapping methods. First, by confining the test to one region at a time, it reduces a multiple dimensional search problems (for multiple QTLs) to a one dimensional search problem. Second, by conditioning on linked markers in the test, the sensitivity of the test statistic to the position of individual QTLs can be increased and the precision of QTL mapping can be improved significantly. Third, by selectively and simultaneously utilizing all the available information in a data set to make inference, the efficiency of QTL mapping is improved. See Zeng (1994) for more detailed discussion on the method.

Model selection: As explained above, statistical analysis for mapping QTLs can be made based on a number of models. Then what is the best model for analysis for a given data set, *i.e.* how many markers and what markers should be used in (1)? Generally in practice, for the same data set, many models can be applied and compared. It would, however, be tedious to do that exercise on every data set generated.

Numerical methods can be provided to guide the search for the "best" model for mapping QTLs. This is particularly important for automation of data analysis.

Several factors influence the choice of a model, such as the sample size and the number and density of markers in data. When the sample size is small, not many markers (and parameters) can be fitted in the model, as the degree of freedom of the test can be reduced very significantly by fitting too many parameters. Markers can be selected for fitting as background control by stepwise regression. When linked markers are also fitted to create an interval test, the genetic distance between the boundaries and the testing position (called the testing window of the interval test) can be controlled. Empirical study has shown that when this distance is more than 15 cM (which, of course, depends on sample size), the power reduction associated with the interval test can be largely alleviated. One problem to apply the method to data is that, in practice, markers are unevenly distributed in genomes. In some regions, markers are in dense, and in some other regions markers are spare. So if we use markers to fix the testing window, the window size can vary from region to region. That could cause a problem for comparing results between different regions. One solution for that can be to use virtue indicator markers constructed from their respective flanking markers to fix the size of the testing window for interval test. All these procedures can be automated.

Significance values: Appropriate choice of a critical value for a test is a major issue in statistics. For mapping QTLs, the choice for a critical value is complicated by the situation of multiple tests in multiple (correlated) locations. Also for different models, the critical values tend to be different. The general patterns and simple approximations of the critical values for the interval test have been discussed by Zeng (1994). These can serve as a guide for choosing an appropriate critical value for the test in different situations. However, more studies on the issue need to be made as the methods are extended to other experimental designs and data structures.

Analysis on multiple traits and multiple environments–testing pleiotropy, close linkage, and QTL-environment interaction: Many QTL mapping data have observations on multiple traits or on one trait in multiple environments. With such data, we can ask questions like: Does a QTL have pleiotropic effects on multiple traits? Does a QTL show genotype-environment interaction effects? Statistically this involves multiple trait analysis, as expressions of a trait in different environments can be regarded as different traits or different trait states. Two experimental designs are considered. In design I, different traits are measured in the same individual or genotype, and in design II, traits expressed in different environments are measured in different individuals.

For design I, the statistical model for analysis can be (for two traits in F_2 design)

$$\begin{pmatrix} y_{1j} \\ y_{2j} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix} x_j^* + \begin{pmatrix} d_1^* \\ d_2^* \end{pmatrix} z_j^* + \sum_k \left[\begin{pmatrix} b_{1k} \\ b_{2k} \end{pmatrix} x_{jk} + \begin{pmatrix} d_{1k} \\ d_{2k} \end{pmatrix} z_{jk} \right] + \begin{pmatrix} e_{1j} \\ e_{2j} \end{pmatrix} \qquad (6)$$

where d_1^* and d_2^* are the dominance effects of the putative QTL at the position being tested on traits 1 and 2, d_{1k} and d_{2k} are dominance effects of marker k on the traits, z_j^* is an indicator variable, taking a value 1 or 0 depending on the probability of the putative QTL to be heterozygote or homozygote, z_{jk} takes a value 1 if marker k of individual j is heterozygote and 0 otherwise, e_{1j} and e_{2j} are error terms and assumed to be bivariate normal distributed, and other terms have similar interpretations as those in (1). This is an extension of model (1) to two trait analysis for F_2 design. For design II, the model can be $y_{ij} = \mu_i + b_i^* x_{ij}^* + d_i^* z_{ij}^* + \sum_k (b_{ik} x_{ijk} + d_{ik} z_{ijk}) + e_{ij}$ where y_{ij} is the value of trait i (or a trait in environment i) in individual j and $e_{ij} \sim N(0, \sigma_i^2)$.

Several tests can be made in this setting of analysis. First, to test the presence a QTL at a position, we can test the hypothesis $b_1^* = 0$, $b_2^* = 0$, $d_1^* = 0$, and $d_2^* = 0$. Given that this hypothesis is rejected, we can proceed to test the hypotheses $b_1^* = 0$ and $d_1^* = 0$ or $b_2^* = 0$ and $d_2^* = 0$ to see whether the QTL has pleiotropic effects, and also in the case of multiple environments, we can test the hypothesis $b_1^* = b_2^*$ and $d_1^* = d_2^*$ to see whether there is QTL-environment interaction. In certain cases, tests can also be made

to test the hypotheses of pleiotropy vs. close linkage. Under the null hypothesis of pleiotropy, one QTL (one position) is formulated to have effects on both traits, and under the alternative hypothesis of close linkage, two QTLs (two positions) are formulated to have effects on two traits.

There are two advantages of this analysis. First, by multivariate analysis, the accuracy and efficiency of mapping QTLs can be further improved in some cases, compared with mapping on each trait separately. We have observed from our simulation studies (Jiang, Zeng and Weir, unpublished) that the joint analysis can usually give better resolution for mapping QTLs. Second, it provides a formal method to test biologically interesting questions.

Testing epistasis: Mapping for QTLs can proceed position by position for testing one QTL at a time conditional on some of other markers. After that, if sufficient interest arises, interaction effects of pairwise identified QTLs (or chromosome regions) can be tested for significance. A genetic model of two gene interaction has been proposed for statistical analysis (Mather and Jinks, 1977). Consider the simplest case of two gene pairs, A-a and B-b. These can give rise to nine different genotypes in an F_2 design

	AA	Aa	aa
BB	$a_1 + a_2 + i_{aa}$	$d_1 + a_2 + i_{da}$	$-a_1 + a_2 - i_{aa}$
Bb	$a_1 + d_2 + i_{ad}$	$d_1 + d_2 + i_{dd}$	$-a_1 + d_2 - i_{ad}$
bb	$a_1 - a_2 - i_{aa}$	$d_1 - a_2 - i_{da}$	$-a_1 - a_2 + i_{aa}$

where a and d define the additive and dominant effects of loci A and B, and i_{aa}, i_{ad}, i_{da} and i_{dd} are four interaction effects of loci A and B. These eight parameters correspond to the eight degrees of freedom among the nine observations. These specifications are very general for statistical analysis.

The statistical model for testing interaction effects of two QTLs identified to be located between markers i, $i+1$ and markers i', $i'+1$ can be defined as $y_j = b_0 + a_1 x^*_{1j} + d_1 z^*_{1j} + a_2 x^*_{2j} + d_2 z^*_{2j} + i_{aa} w^*_{aaj} + i_{ad} w^*_{adj} + i_{da} w^*_{daj} + i_{dd} w^*_{ddj} + \sum_{k \neq i, i+1, i', i'+1} (b_k x_{kj} + c_k z_{kj}) + e_j$, where the stared variables are corresponding indicator variables with values specified in the above table with probabilities depending on the genotypes of two pairs of flanking markers and the testing positions. In F_2 design, with two pairs of flanking markers, there are 81 ($= 9 \times 9$) different combinations of marker genotypes, and for each marker genotype the probability of an individual to possess one of the nine QTL genotypes is different. However, given the marker genotype, the probability of QTL genotype should be independent for two marker intervals, irrespective of whether the two marker intervals are linked or not (assuming that there is no cross-over interference), that is $Prob(\text{Q1Q2}|\text{MP1/MP2}) = Prob(\text{Q1}|\text{MP1})Prob(\text{Q2}|\text{MP2})$ for QTL genotype Q1Q2 given two marker pair genotypes MP1 and MP2.

The likelihood function is given by $L = \prod_{j=1}^{n} \sum_{k=1}^{9} p_{jk} f_{jk}$ where the summation is for the nine genotypes of QTLs A and B, p_{jk} is the prior probability of individual j having QTL genotype k, and f_{jk} is the corresponding normal density function for individual j with QTL genotype k specified by the model. Testing on parameter i's can be made individually or collectively by likelihood ratio tests with constrained likelihoods calculated with corresponding i's set up to be zero.

This is an appropriate testing procedure which combines interval mapping with multiple regression analysis. Properties and behavior of this method for testing QTL epistasis will be discussed elsewhere.

This study is supported in part by grants NIH GM 45344 and NSF DEB- 9220856.

REFERENCES

LANDER, E. S. and BOTSTEIN, D. (1989) Genetics 121: 185-199.
MATHER, K. and JINKS, J. L. (1977) Introduction to Biometrical Genetics.
MENG, X.-L. and RUBIN, D. B. (1993) Biometrika 80: 267-268.
ZENG, Z.-B. (1993) Proc. Natl. Acad. Sci. USA 90: 10972-10976.
ZENG, Z.-B. (1994) Genetics (in press).