Accurate and Efficient Mapping of Quantitative Trait Loci

E.S. Lander[1] and D. Botstein[2]

[1]Whitehead Institute for Biomedical Research
Cambridge, Massachusetts 02142 and
Harvard University, Cambridge, Massachusetts 02138

[2]Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 and
Genentech, Inc., South San Francisco, California 94080

The advent of complete genetic linkage maps consisting of codominant DNA markers such as restriction fragment length polymorphisms (RFLPs) (Botstein et al. 1980) has made feasible the resolution of multiple Mendelian factors underlying quantitative genetic differences between strains. In principle, such dissection of polygenic traits is straightforward: (1) A backcross or intercross is performed between two strains differing in a trait of interest; (2) progeny are scored both for the trait and for codominant markers spaced throughout the genome; and (3) a correlation is sought between the trait and the inheritance pattern of one or more markers. When a significant correlation is found, the presence of a *q*uantitative *t*rait *l*ocus (QTL) is declared.

In practice, there are a number of methodological problems. To address these problems, we have recently developed a comprehensive approach to QTL mapping (Lander and Botstein 1989).

Detecting QTLs by Interval Mapping Using LOD Scores

The traditional approach to detecting the presence of a QTL near a marker locus and inferring the phenotypic effects (α and $\alpha+\beta$) of the QTL alleles is to perform linear regression of the quantitative phenotype on the genotype at the marker locus (e.g., Soller and Brody 1976). A shortcoming of this approach is that it tests for the presence of QTLs only exactly at a marker locus—not in the intervals between markers. As a consequence, (1) whenever the QTL does not lie exactly at the marker locus, recombination decreases the apparent phenotypic effect and thus causes linear regression systematically to underestimate the phenotypic effect β; (2) since it diminishes the apparent phenotypic effect, recombination also increases the number of progeny needed to detect linkage to a QTL; (3) individuals with missing genotypic data at a marker locus cannot be used in the analysis; and (4) although the approach detects the presence of a QTL in the neighborhood of a marker, it supplies no estimate of its position.

In order to remedy these problems, we exploit the full power of a genetic linkage map by generalizing linear regression along lines commonly used in human genetics (Ott 1985). Linear regression is a special case of the method of maximum likelihood, the statistical principle that advocates estimating parameters by the value that maximizes the probability of the observed data having occurred. Even when a QTL is at some distance from a marker locus, one can compute the probability P_ϕ that an individual will show phenotype ϕ as a function of the allelic effects α and $\alpha+\beta$ at the QTL; the phenotypic variance σ^2 not attributable to the QTL; the position θ of the QTL relative to the nearest informative markers; and the genotypes at these marker loci. Specifically,

$$P_\phi(\alpha,\beta,\sigma^2,\theta) = \sum_g \ p_\theta(g) \quad f_{\alpha,\beta,\sigma^2}(\phi)$$

where the summation is taken over all possible genotypes at the QTL; $p_\theta(g)$ is the probability that the genotype is g at the QTL based on the position of the QTL relative to the markers and the observed genotype at the markers; and

$$f_{\alpha,\beta,\sigma^2}(\phi)$$

is the probability that an individual with QTL genotype g will exhibit phenotype ϕ based on the values of the parameters. For the entire data set, the *likelihood function* $L(\alpha,\beta,\sigma^2,\theta)$ is simply the product of $P_\phi(\alpha,\beta,\sigma^2,\theta)$ taken over all individuals. At any given position θ, one can then apply numerical analysis to find the maximum likelihood estimates (MLEs) of the QTL parameters $\alpha^*,\beta^*,\sigma^{*2}$. The strength of the evidence for the presence of a QTL at a given position θ is provided by the odds ratio.

$$\text{Odds ratio} = L(\alpha^*,\ \beta^*,\ \sigma^{*2},\theta)\ /\ L(\alpha^{**},\ 0,\ \sigma^{**2},\ \theta)$$

where α^{**} and σ^{**2} are the MLEs under the assumption that there is no linked QTL (i.e., $\beta = 0$). Essentially, the odds ratio denotes how much more probable it is for the data to have

arisen if there is a QTL at the given position than if there is no linked QTL. Following the convention in human genetics (Ott 1985), evidence for linkage is reported in terms of the LOD score = $\log_{10}$ (odds ratio). The evidence for the presence of a QTL can be conveniently displayed by a QTL *likelihood map*, indicating the LOD score at all points along the length of a chromosome (Fig. 1). When the LOD score crosses a predetermined threshold *T*, the presence of a QTL is declared. The approximate position of the QTL can be represented by, for example, a 1.0-LOD support interval, defined as the region within which the LOD score remains within 1.0 log unit of its maximum. The support interval is similar to a confidence interval for the location of the QTL.

Among the advantages of interval mapping over simple regression are the following: (1) Because the phenotypic effect β is the MLE for a correctly specified model, it follows from general properties of MLEs that it is asymptotically unbiased. (2) Because flanking markers allow the genotype at QTLs to be inferred more accurately, somewhat fewer progeny are needed. (3) The use of QTL likelihood maps allows estimation of position of a QTL within an appropriate support interval. (4) Individuals with missing data at a marker are not discarded from the analysis, since information about QTL genotype can be extracted from the nearest flanking, informative markers. Although interval mapping is more powerful, it should be noted that the method reduces to simple linear regression when the QTL lies exactly at a marker locus ($\theta = 0$) and there are no missing data: The estimated allelic effects agree with those obtained by regression, and the LOD score is closely related to the F-statistic used for judging the statistical significance of regressions.

One disadvantage of the new method is that standard computer packages for linear regression cannot be used for QTL mapping. Instead, special purpose computer programs must be written. Accordingly, we have recently written MAPMAKER-QTL, a computer program that implements interval mapping for backcrosses (E.S. Lander and S.E. Lincoln, unpubl.). We are currently extending this program to the analysis of intercrosses.

Appropriate Threshold for Detecting QTLs

Although the traditional approach to QTL mapping allows a 0.05 chance of false positives at any given point in the genome

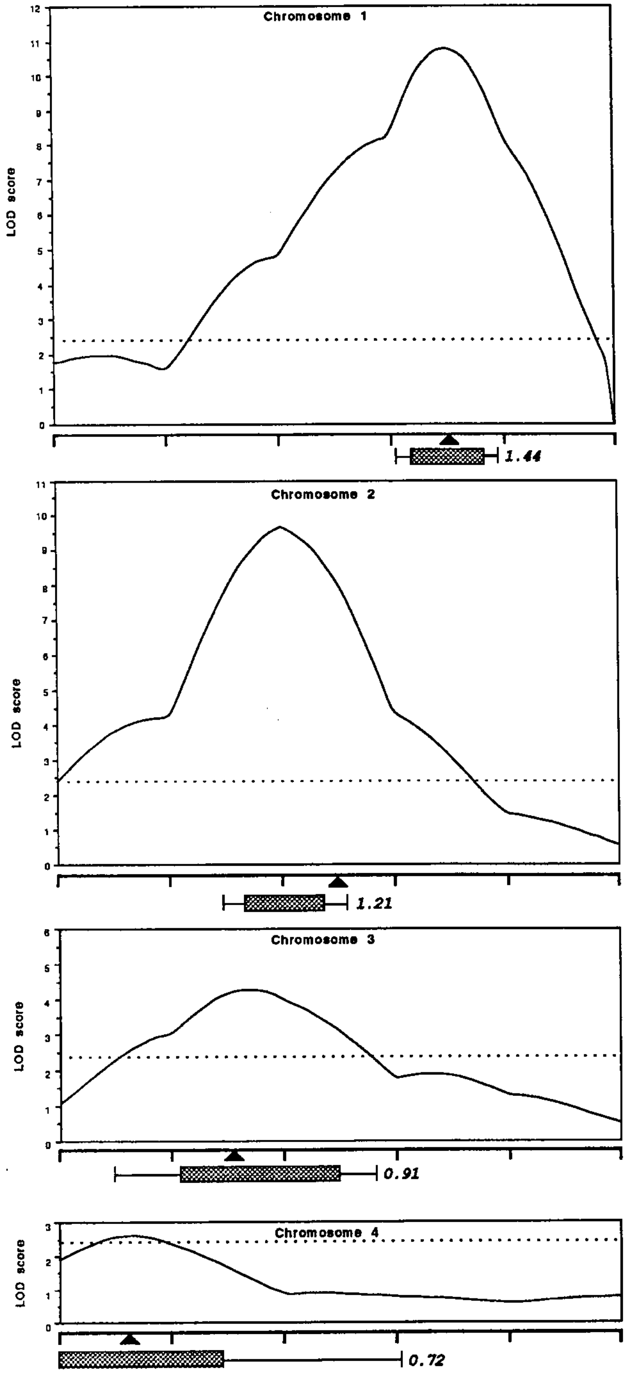

Figure 1 *(Continued on following page.)*

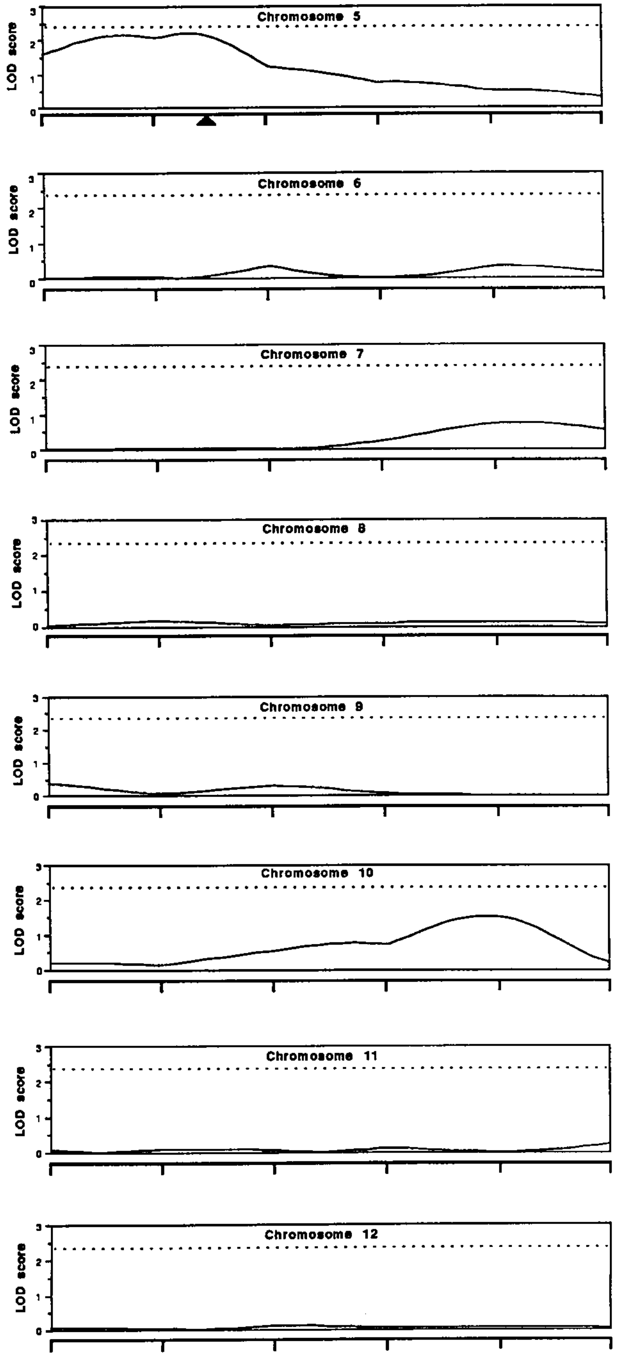

Figure 1 *(See following page for legend.)*

(Soller and Brody 1976), this choice neglects the fact that many markers are being tested. Indeed, one can show (Lander and Botstein 1989) that this standard will yield a probability >90% of at least one false-positive QTL occurring *somewhere* in the genome.

To guard against false positives then, how high a LOD score threshold T should be employed? Assuming that there are no QTLs in a genome, one can show (Lander and Botstein 1989) that the QTL likelihood map will follow the square of a stochastic process known as the Orenstein-Uhlenbeck diffusion (corresponding to a particle diffusing by Brownian motion while coupled to the origin by a weak Hookean spring). By applying the large-deviation theory of the Orenstein-Uhlenbeck diffusion (Leadbetter et al. 1983), one can derive limits on how high the LOD score might become by chance. For a typical plant genome, one can show that a LOD score threshold of about 2.5 is required to keep the chance <5% that a false positive will occur *somewhere* in the genome (Lander and Botstein 1989). This threshold corresponds approximately to the 0.001 confidence level for each marker tested.

Application to the Tomato

In collaboration with colleagues, we have recently applied these methods and computer programs to a genetic dissection of fruit weight, soluble solids concentration, and pH in the tomato (Paterson et al. 1988). Analysis of 237 progeny of an interspecific backcross in tomato revealed six QTLs affecting fruit weight, four QTLs affecting soluble solids concentration, and five QTLs affecting pH. In each case, the QTLs detected accounted for more than 50% of the observed genetic variance in the backcross.

Figure 1 LOD scores for a hypothetical quantitative trait. The LOD scores are based on simulated data for 250 backcross progeny in an organism with 12 chromosomes of 100 cM each. For each individual, crossovers were generated assuming no interference, and genotypes were recorded at RFLP markers spaced every 20 cM throughout the genome (indicated by tick marks on the chromosomes below each graph). The quantitative phenotype for each individual was generated by summing individual alleles at five QTLs and adding random environmental normal noise. Alleles at the QTLs had effects 1.5, 1.25, 1.0, 0.75, and 0.50 and were located, respectively, on chromosomes 1, 2, 3, 4, and 5 at (arbitrarily chosen) genetic positions 70, 49, 27, 8, and 30 cM from the left end (indicated by black triangles on the chromosomes), and the random noise had S.D. 1. No QTLs were located on chromosomes 6–12. The dotted line at LOD = 2.4 indicates the required significance level for a genome of this size. The four largest QTLs attained this LOD threshold. Gray bars indicate one-log confidence intervals for the position of the QTLs: Outside this region, the odds ratio has fallen by a factor of 10. Thin lines extending from the gray bars indicate two-log confidence intervals. MLEs of the phenotypic effect are indicated to the right of the confidence intervals. Data were analyzed with MAPMAKER-QTL computer package (S.E. Lincoln and E.S. Lander, unpubl.).

Number of Progeny Required

Using the traditional approach of linear regression at marker loci, the number of progeny required to map a QTL is inversely proportional to the square of the allelic effect of the QTL (measured in units of environmental standard deviations), except when the allelic effect becomes so large that the trait is qualitative (Soller and Brody 1976). Thus, detection of small phenotypic effects may require quite large progeny sizes. Fortunately, various methods can be used to increase the efficiency of QTL mapping by decreasing the constant of proportionality. In practical cases, (1) *interval mapping* can decrease the number of progeny by about 25% by allowing inheritance at QTL loci to be followed more accurately; (2) *selective genotyping* of the progeny with the most extreme phenotypes can decrease the number of progeny that must be genotyped by about fivefold by exploiting the fact that the largest expected LOD score is contributed by the most extreme progeny; (3) *progeny testing* can decrease the number of progeny required by reducing "environmental" noise; and (4) *simultaneous search* of multiple intervals can decrease the number of progeny required by decreasing the unexplained genetic variance. Elsewhere (Lander and Botstein 1989), we describe these methods in more detail and provide calculations of the number of progeny required to map QTLs having a given phenotypic effect.

Selection of Strains

If the goal is to find QTLs having a substantial effect on a trait of interest, it is possible to choose parental strains to maximize the chance of success (Lander and Botstein 1989). Ideally, one should select strains in which a large difference in the trait (1) has resulted from natural or artificial selection in opposite directions and (2) is principally caused by a small number of QTLs. Concerning the latter point, one can apply the classic formula of S. Wright for the number k of effective factors in a cross (see Wright 1968) to estimate the number of QTLs. Although this estimate is only accurate when all the QTLs have equal phenotypic effects, it can be shown that k provides a good estimate of the number of QTLs that have a *large* phenotypic effect (Lander and Botstein 1989).

By combining the methods described above, it is possible to design and analyze crosses to achieve accurate and efficient mapping of QTLs.

REFERENCES

Botstein, D., R.L. White, M. Skolnick, and R.W. Davis. 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32:** 314.

Lander, E.S. and D. Botstein. 1989. Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* **121:** 185.

Leadbetter, M.R., G. Lindgren, and H. Rootzen. 1983. *Extremes and related properties of random sequences and processes.* Springer Verlag, New York.

Ott, J. 1985. *Analysis of human genetic linkage.* Johns Hopkins University Press, Baltimore.

Paterson, A.H., E.S. Lander, J.D. Hewitt, S. Peterson, S.E. Lincoln, and S.D. Tanksley. 1988. Resolution of quantitative traits into Mendelian factors by using a complete RFLP linkage map. *Nature* **335:** 721.

Soller, M. and T. Brody. 1976. On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crosses between inbred lines. *Theor. Appl. Genet.* **47:** 35.

Wright, S. 1968. Genetics of quantitative variability. In *Evolution and the genetics of populations*, vol. 1: *Genetic and biometric foundations*, p. 383. University of Chicago Press, Chicago.

Use of RFLPs for Analysis of Quantitative Trait Loci in Maize

J. Romero-Severson, J. Lotzer, C. Brown, and M. Murray

Agrigenetics Corporation, Madison, Wisconsin 53716

Pinpointing the location and effect of specific quantitative trait loci (QTLs) can result in dramatic gains from selection in the early generations of a quantitative trait improvement program. Mapped restriction fragment length polymorphisms (RFLPs) provide a set of codominant, densely distributed genetic markers, some of which, by virtue of linkage to a QTL, may display associations with the genotypic variance of a quantitative trait. However, there are no standard experimental designs and statistical analyses for detecting QTLs with RFLPs. Some factors to consider and a comparison of statistical approaches follow.

Population Structure

The most important components of a QTL mapping strategy are the differential between the parents for the trait and the power of the field design to distinguish the difference between genetic and environmental effects. Other considerations include population structure, the numbers of individuals to sample, and the number of RFLP markers to use. For quantitative traits with relatively few genes (<10), either a simple F_3 progeny test with RFLP analysis of the F_2 or an inbred-backcross approach (Wehrhahn and Allard 1965) should suffice. For the more challenging traits, development of recombinant inbred lines may be necessary (Burr et al. 1988).

The numbers of individuals to sample and the extent of genomic coverage needed are largely determined by economic and genetic limitations. The population sizes in the thousands which theory suggests for QTL mapping (Lebowitz et al. 1986) are presently impractical and may be unnecessary if the genes involved have disproportional effects. Moreover, theoretically good genomic coverage (an RFLP marker every 10 map units) may not be available in the population in question.