Mapping Mendelian Factors Underlying Quantitative Traits Using RFLP Linkage Maps

Eric S. Lander,[*,†,‡] **and David Botstein**[‡,§]

[*]*Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142,* [†]*Harvard University, Cambridge, Massachusetts 02138,* [‡]*Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, and* [§]*Genentech, South San Francisco, California 94080*

Manuscript received August 2, 1988
Accepted for publication October 6, 1988

ABSTRACT

The advent of complete genetic linkage maps consisting of codominant DNA markers [typically restriction fragment length polymorphisms (RFLPs)] has stimulated interest in the systematic genetic dissection of discrete Mendelian factors underlying quantitative traits in experimental organisms. We describe here a set of analytical methods that modify and extend the classical theory for mapping such quantitative trait loci (QTLs). These include: (i) a method of identifying promising crosses for QTL mapping by exploiting a classical formula of SEWALL WRIGHT; (ii) a method (interval mapping) for exploiting the full power of RFLP linkage maps by adapting the approach of LOD score analysis used in human genetics, to obtain accurate estimates of the genetic location and phenotypic effect of QTLs; and (iii) a method (selective genotyping) that allows a substantial reduction in the number of progeny that need to be scored with the DNA markers. In addition to the exposition of the methods, explicit graphs are provided that allow experimental geneticists to estimate, in any particular case, the number of progeny required to map QTLs underlying a quantitative trait.

THE conflict between the Mendelian theory of particulate inheritance and the observation that most traits in nature exhibit continuous variation was eventually resolved by the concept that quantitative inheritance can result from the segregation of multiple genetic factors, modified by environmental effects (JOHANNSEN 1909; NILSSON-EHLE, 1909; EAST 1916). Breeding studies confirmed numerous predictions of this theory (EAST 1916) and pioneering genetic mapping studies (SAX 1923; RASMUSSON 1933; THODAY 1961; TANKSLEY, MEDINA-FILHO and RICK 1982; EDWARDS, STUBER and WENDEL 1987) showed that it was even possible occasionally to detect genetic linkage to the putative *quantitative trait loci* (QTLs). Unfortunately, systematic and accurate mapping of QTLs has not been possible because of the difficulty in arranging crosses with genetic markers densely spaced throughout an entire genome. Recently, such studies have become possible in principle with the advent of restriction fragment length polymorphisms (RFLPs) as genetic markers (BOTSTEIN *et al.* 1980) and the increasing availability of complete RFLP maps in many organisms.

Systematic genetic dissection of quantitative traits using complete RFLP linkage maps would be valuable in a broad range of biological endeavours. Agricultural traits such as resistance to diseases and pests, tolerance to drought, heat, cold, and other adverse conditions, and nutritional value could be mapped and introgressed into domestic strains from exotic relatives (RICK 1973; HARLAN 1976). Aspects of mammalian physiology such as hypertension, atherosclerosis, diabetes, predispositions to cancer and teratomas, alcohol sensitivity, drug sensitivities and some behaviours could be investigated in animal strains differing widely for these traits (TANASE *et al.* 1970; DE JONG 1984; PAIGEN *et al.* 1985; PROCHAZKA *et al.* 1987; HESTON 1942; KALTER 1954; MALKINSON and BEER 1983; SHIRE 1968; STEWART and ELSTON 1973; ELSTON and STEWART 1973; FESTING 1979). Evolutionary questions about speciation could be elucidated by determining the number and nature of the genes involved in reproductive barriers (COYNE and CHARLESWORTH 1986). An example of such genetic dissection is reported in a companion paper (PATERSON *et al.* 1988): In an interspecific cross in tomato, QTLs affecting fruit weight, concentration of soluble solids and fruit pH are mapped to within about 20–30 cM by means of a complete RFLP linkage map.

The purpose of this paper is to discuss the general methodology for mapping QTLs in experimental organisms. Although the basic idea has been clear since SAX (1923), the systematic approach made possible by complete RFLP linkage maps raises a number of questions. With complete coverage of the genome assured by the map, is it possible to design a cross so as to make it highly likely that QTLs will be found? Can the estimation of QTL effects and positions be made more accurate through the use of flanking markers? When searching an entire genome for QTLs, what precautions are needed to avoid false positives? In view of the time and expense of complete RFLP genotyping, how can the number of progeny that must be genotyped be minimized? To address these issues, we explore below ways to:

(*i*) *Identify promising crosses for QTL mapping.* Genetic dissection of a quantitative trait will succeed only when some of the QTLs segregating in the cross have relatively large phenotypic effects. By exploiting a classical formula of SEWALL WRIGHT, we show that it is often possible to recognize such crosses in advance and thereby to ensure that QTLs will in fact be identified.

(*ii*) *Exploit the full power of complete linkage maps.* The traditional approach to mapping QTLs (SAX 1923; SOLLER and BRODY 1976) involves studying single genetic markers one-at-a-time. In general, the drawbacks of the method include that (a) the phenotypic effects of QTLs are systematically underestimated, (b) the genetic locations of QTLs are not well resolved because distant linkage cannot be distinguished from small phenotypic effect, and (c) the number of progeny required for detecting QTLs is larger than necessary. By adapting the method of LOD scores used in human genetic linkage analysis, we show how to remedy these problems by the approach of *interval mapping* of QTLs. In addition, the traditional approach neglects the problem that testing many genetic markers increases the risk that false positives will occur. We determine the appropriate degree of statistical stringency to prevent such errors in mapping QTLs.

(*iii*) *Decrease the number of progeny to be genotyped.* In typical cases, a reduction of up to sevenfold can be achieved by combining two approaches: interval mapping and selective genotyping. *Selective genotyping* involves growing a larger population, but genotyping only those individuals whose phenotypes deviate substantially from the mean. Additional methods for increasing the power of QTL mapping include reducing environmental noise by progeny testing and reducing genetic noise by studing several genetic regions simultaneously.

Although the RESULTS section is mathematical in parts, the DISCUSSION presents the methodology in terms of explicit graphs that allow an experimental geneticist to design crosses to dissect a quantitative trait by using a complete RFLP linkage map.

RESULTS

The basic methodology for mapping QTLs involves arranging a cross between two inbred strains differing substantially in a quantitative trait: segregating progeny are scored both for the trait and for a number of genetic markers. Typically, the segregating progeny are produced by a B_1 backcross (F_1 × Parent) or an F_2 intercross (F_1 × F_1). For simplicity, only the backcross will be discussed in detail. As we note below, the F_2 intercross is analogous and requires only about half as many progeny.

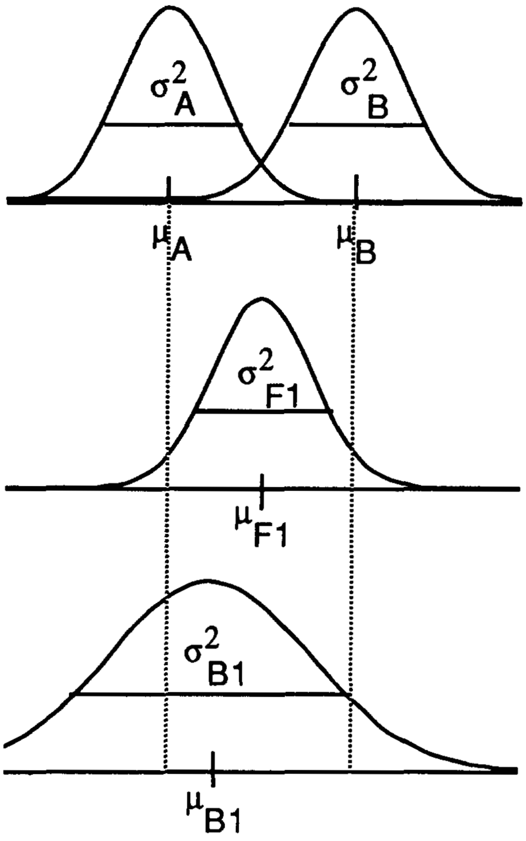

FIGURE 1.—Phenotype distributions. Schematic drawing of phenotypic distributions in the *A* and *B* parental, F_1 hybrid and B_1 backcross populations.

Definitions and assumptions: Let *A* and *B* be inbred strains differing for a quantitative trait of interest, and suppose that a B_1 backcross is performed with *A* as the recurrent parent. Let

$$(\mu_A, \sigma_A^2),\ (\mu_B, \sigma_B^2),\ (\mu_{F_1}, \sigma_{F_1}^2) \text{ and } (\mu_{B_1}, \sigma_{B_1}^2)$$

denote the mean and variance of the phenotype in the *A*, *B*, F_1 and B_1 populations, respectively (Figure 1). Let $D = \mu_B - \mu_A > 0$ denote the phenotypic difference between the strains. The cross will be analyzed under the classical assumption (MATHER and JINKS 1971; FALCONER 1981) that the phenotype results from summing the effects of individual QTL alleles, and then adding normally distributed environmental (*i.e.*, nongenetic) noise. In particular, we assume complete codominance and no epistasis. These assumptions imply that

$$\mu_{F_1} = \tfrac{1}{2}(\mu_A + \mu_B), \tag{1a}$$

$$\mu_{B_1} = \tfrac{1}{2}(\mu_A + \mu_{F_1}), \quad \text{and} \tag{1b}$$

$$\sigma_A^2 = \sigma_B^2 = \sigma_{F_1}^2 < \sigma_{B_1}^2. \tag{1c}$$

The variances within the *A*, *B* and F_1 populations equal the *environmental variance*, σ_E^2, among genetically identical individuals, while the variance within the B_1 progeny also includes *genetic variance*, $\sigma_G^2 = \sigma_{B_1}^2 - \sigma_E^2$. Frequently, phenotypic measurements must be mathematically transformed so that parental phenotypes are approximately normally distributed and the relations (1abc) are approximately satisfied. For example, WRIGHT (1968) obtained an excellent fit to the theory by applying a log-transformation (appropriate when the standard deviations scale with the mean) to tomato fruit weight.

By the phenotypic effect δ of a QTL, we will mean the additive effect of substituting *both A* alleles by *B* alleles. A single allele has effect $\frac{1}{2}\delta$, since additivity is assumed. In a backcross, the segregation of a QTL with effect δ contributes an amount $\delta^2/16$ to the genetic variance σ_G^2. The *variance explained* by the QTL is written $\sigma_{\text{exp}}^2 = \delta^2/16$, while the *residual variance* is $\sigma_{\text{res}}^2 = \sigma_{B_1}^2 - \sigma_{\text{exp}}^2$.

Choosing strains

The ability to map QTLs underlying a quantitative trait depends on the magnitude of their phenotypic effect: the smaller the effect that one wishes to detect, the more progeny will be required. Before attempting genetic dissection of a quantitative trait, it would thus be desirable to identify crosses segregating for QTLs with relatively large phenotypic effects and to estimate the magnitude of the effects. In fact, this can often be accomplished by exploiting a classical formula of WRIGHT.

WRIGHT (quoted by CASTLE 1921; WRIGHT 1968) proved that the number k of QTLs segregating in a backcross between two strains with phenotypic difference D can be estimated by the formula:

$$k = D^2/16\sigma_G^2, \qquad (2)$$

provided that the following assumptions hold: (i) the QTLs have effects of equal magnitude, (ii) the QTLs are unlinked, and (iii) the alleles in the high strain all increase the phenotype, while those in the low strain decrease the phenotype. (To see this, recall that the variance explained by a single such QTL would be $\sigma_{\text{exp}}^2 = (D/k)^2/16$ and thus the total genetic variance explained by the k QTLs would be $\sigma_G^2 = (1/k)(D^2/16)$.)

The quantity k is called the *number of effective factors* in the cross. If the assumptions are satisfied, then each QTL affects the phenotype by (D/k) and explains $(1/k)$ of the genetic variance in the backcross. Unfortunately, if these assumptions are not satisfied (as will be likely in practice; *cf.* PATERSON *et al.* 1988), the number of effective factors k may seriously underestimate the number of QTLs. In principle, the number of QTLs is unlimited. In this case, must there exist *any* QTLs affecting the phenotype by (D/k)? More generally, for any $0 \leq \varepsilon \leq 1$, must there exist QTLs affecting the phenotype by $\varepsilon(D/k)$? And, how much of the total phenotypic difference D and the genetic variance σ_G^2 can be attributed to such QTLs? Proposition 1 (proven in APPENDIX [A1]) supplies an answer:

Proposition 1. *Consider a cross in which the phenotypic difference between the strains is D and the number of effective factors is k. Assume that the QTLs are unlinked and that the alleles in the "high" strain all increase the phenotype. Let S_ε denote the set consisting of those QTLs that alter the phenotype by at least $\varepsilon(D/k)$. No matter how many QTLs are segregating and no matter what their individual phenotypic effects, the QTLs in S_ε must together account for a fraction $\geq D_\varepsilon$ of the total phenotypic difference D between the strains and must together explain a fraction $\geq V_\varepsilon$ of the genetic variance in the second generation, where*

$$D_\varepsilon = [\tfrac{1}{2}\varepsilon + \sqrt{(1-\varepsilon)k + \tfrac{1}{4}\varepsilon^2}]/k \quad \text{and}$$

$$V_\varepsilon = 1 - \varepsilon(1 - D_\varepsilon).$$

Considering the case $\varepsilon = 1$, the proposition states that the QTLs with phenotypic effect (D/k) must account for a phenotypic difference of at least (D/k). In other words, there *must* exist at least *one* QTL having phenotypic effect $\geq (D/k)$.

Suppose that we are willing to search for QTLs with somewhat smaller effects. How much of the phenotypic difference can be attributed to QTLs with effect $\geq \frac{1}{2}(D/k)$? Taking $\varepsilon = \frac{1}{2}$ and considering various values of k, we have:

k	Minimum proportion (%) of phenotypic difference D accounted for by QTLs with effect $\geq \frac{1}{2}(D/k)$	Minimum proportion (%) of genetic variance σ_G^2 explained by QTLs with effect $\geq \frac{1}{2}(D/k)$
2	64	82
3	50	75
4	42	71
5	37	69

A small value of k thus implies that the cross must be segregating for QTLs with relatively large effects ($\geq\frac{1}{2}(D/k)$), which together account for a substantial proportion of the phenotypic difference and explain a substantial proportion of the genetic variance in the backcross.

In other words, WRIGHT's formula can be used to indicate the presence of *some* QTLs with large effects—even though the number k of effective factors may not be a reliable estimate of the total *number* of QTLs. Note that Proposition 1 provides only a lower bound on the total effect attributable to the QTLs in S_ε: in general, these QTLs will have an even greater effect.

How serious a limitation is posed by the two assumptions remaining in Proposition 1?

(i) The first assumption is not essential: admitting the possibility of linked QTLs simply allows that some large QTL effects may eventually prove to be due to several nearby genes. Such questions may be safely neglected at first.

(ii) The second assumption is more important. Fortunately, it is possible to choose crosses in which it is likely to be satisfied. The ideal situation would be two strains arising from brief, intense artificial selection for and against the trait in a large outbred population, followed by inbreeding: in such a case, classical selection theory (*e.g.*, FALCONER 1981) shows that a "high" strain is unlikely to fix a "low" allele at QTLs with relatively large effect; moreover, the force of selection will be greatest on the QTLs with the largest effects. Many such strains have been developed by artificial selection to study various physiological traits. As a reasonable alternative, one could use strains that appear to have resulted from natural selection for the trait.

Judicious choice of strains can essentially ensure that *some* QTLs will be detected in a reasonable progeny size calculable in advance. When studying strains resulting from selection, a sensible approach might be to use enough progeny to map QTLs having effect δ between $\frac{1}{2}(D/k)$ and (D/k). Of course, one could choose to study more progeny and might well be rewarded with the detection of QTLs with smaller effects.

Unselected strains exhibiting extreme phenotypic differences may also merit attention. Despite the lack of a mathematical guarantee, QTLs with large effects may nonetheless be segregating. When there is no prior evidence of both high and low alleles in the same strain, one may wish to proceed as in the previous paragraph. When there is evidence (as when many segregating progeny exhibit phenotypes more extreme than either parent; (*cf.* PATERSON *et al.* 1988), the analysis above does not apply and the detection level must be chosen somewhat arbitrarily.

Assuming that the desired detection level δ has been chosen (by Proposition or arbitrarily), we next consider the method for mapping QTLs and the number of progeny required.

Mapping QTLs: traditional approach

The traditional approach (SAX 1923; SOLLER and BRODY 1976; TANKSLEY, MEDINA-FILHO and RICK 1982; EDWARDS, STUBER and WENDELL 1987) for detecting a QTL near a genetic marker involves comparing the phenotypic means for two classes of progeny: those with marker genotype *AB*, and those with marker genotype *AA*. The difference between the means provides an estimate of the phenotypic effect of substituting a *B* allele for an *A* allele at the QTL. To test whether the inferred phenotypic effect is significantly different from 0, one applies a simple statistical test—amounting to linear regression (*i.e.*, one-way analysis of variance) under the assumption of normally-distributed residual environmental variance.

Consider a QTL that contributes σ^2_{exp} to the genetic variance. Supposing that such a QTL were located exactly at a marker locus, the number of progeny required for detection would be approximately (SOLLER and BRODY 1976)

$$(Z_\alpha)^2(\sigma^2_{\text{res}}/\sigma^2_{\text{exp}}), \tag{3}$$

where this progeny size affords a 50% probability of detection if such a QTL is actually present and a probability α of a false positive if no QTL is linked. Here, Z_α is defined by the equation $\textit{Probability}(z > Z_\alpha) = \alpha$ where z is a standard normal variable (*i.e.*, Z_α is the number of standard deviations beyond which the normal curve contains probability α). SOLLER and BRODY (1976) suggest allowing a false positive rate of $\alpha = 0.05$. For a given false positive rate, the required progeny size thus scales essentially inversely with the square of the phenotypic effect of the QTL or, equivalently, inversely with the variance explained.

Although it captures the key features of QTL mapping, the traditional approach has a number of shortcomings:

(i) If the QTL does not lie at the marker locus, its phenotypic effect may be seriously underestimated. If the recombination fraction is θ, the inferred phenotypic effect of the QTL is biased downward by a factor of $(1 - 2\theta)$. [Proof: If the two QTL genotype classes have phenotypic means 0 and 1, then the two marker genotype classes will have means θ and $(1 - \theta)$.]

(ii) If the QTL does not lie at the marker locus, substantially more progeny may be required. In particular, the variance explained by the marker decreases by a factor of $(1 - 2\theta)^2$ and the number of progeny consequently increases by a factor of $1/(1 - 2\theta)^2$. For an RFLP map with markers every 10, 20, 30 or 40 cM throughout the genome, the progeny size would need to be increased by 22%, 49%, 82% or 123%, respectively, to account for the possibility that the QTL might lie in the middle of an interval—*i.e.*, at the maximum distance from the nearest RFLP. (These calculations use the Haldane mapping function, corresponding to no interference.)

(iii) The approach does not define the likely position of the QTL. In particular, it cannot distinguish between tight linkage to a QTL with small effect and loose linkage to a QTL with large effect.

(iv) The suggested false positive rate of $\alpha = 0.05$ neglects the fact that many markers are being tested. While the chance of a false positive at any given marker is only 5%, the chance that at least one false positive will occur *somewhere* in the genome is much higher.

These difficulties stem from the fact that single markers are analyzed one-at-a-time. To remedy these problems, we generalize the approach so that we may exploit the full power of an RFLP linkage map to scan the intervals between markers as well.

QTL mapping: interval mapping using LOD scores

Method of maximum likelihood: The traditional approach, involving linear regression of phenotype on genotype, is a special case of the *method of maximum likelihood*. Formally, the phenotype ϕ_i and genotype g_i for the *i*th individual are assumed to be related by the equation

$$\phi_i = a + bg_i + \varepsilon,$$

where g_i is encoded as a (0, 1)-indicator variable equal to the number of B alleles, ε is a random normal variable with mean 0 and variance σ^2, and a, b, and σ^2 are unknown parameters. Here, b denotes the estimated phenotypic effect of a single allele substitution at a putative QTL.

The linear regression solutions $(\hat{a}, \hat{b}, \hat{\sigma}^2)$ are in fact *maximum likelihood estimates* (MLEs) for the parameters—that is, they are the values which maximize the probability $L(a, b, \sigma^2)$ that the observed data would have occurred. Here,

$$L(a, b, \sigma^2) = \Pi_i\, z((\phi_i - (a + bg_i)), \sigma^2), \qquad (4)$$

where $z(x, \sigma^2) = (2\pi\sigma^2)^{-1/2}\exp(-x^2/2\sigma^2)$ is the probability density for the normal distribution with mean 0 and variance σ^2. Under the method of maximum likelihood, the MLEs are compared to the constrained MLEs obtained under the assumption that $b = 0$, corresponding to the assumption that no QTL is linked. These constrained MLEs are easily seen to be $(\hat{\mu}_A, 0, \hat{\sigma}^2_{B_1})$. The evidence for a QTL is then summarized by the *LOD score*:

$$\text{LOD} = \log_{10}(L(\hat{a}, \hat{b}, \hat{\sigma}^2)/L(\hat{\mu}_A, 0, \hat{\sigma}^2_{B_1})),$$

essentially indicating how much more probable the data are to have arisen assuming the presence of a QTL than assuming its absence. (The choice of $\log_{10}$ accords with longstanding practice in human genetics (MORTON 1955), although $\log_e$ would be slightly more convenient below.) If the LOD score exceeds a predetermined threshold T, a QTL is declared to be present. The important issues are: (i) What LOD threshold T should be used in order to maintain an acceptably low rate of false positives? (ii) What is the expected contribution to the LOD score (called the ELOD) from each additional progeny? The number of progeny required is then T/ELOD to provide even odds of detecting the QTL with the desired false positive rate.

When only a single genetic marker is being tested, these questions are easily answered. (i) By a general result about maximum likelihood estimation in large samples (KENDALL and STUART 1979), LOD is asymptotically distributed as $\frac{1}{2}(\log_{10} e)\chi^2$, where χ^2 denotes the χ^2 distribution with 1 d.f. A false positive rate of α will thus result if the LOD threshold is chosen so that $T = \frac{1}{2}(\log_{10} e)(Z_\alpha)^2$. For the 5% error rate suggested by SOLLER and BRODY (1976), the threshold is $T = 0.83$. We postpone temporarily the question of the appropriate threshold when many markers are being tested. (ii) For a QTL contributing σ^2_{exp} to the backcross variance, the expected LOD score per progeny (ELOD) is

$$\text{ELOD} = \tfrac{1}{2}\log_{10}(1 + \sigma^2_{\text{exp}}/\sigma^2_{\text{res}}) \qquad (5a)$$

$$\approx \tfrac{1}{2}(\log_{10} e)(\sigma^2_{\text{exp}}/\sigma^2_{\text{res}}) \qquad (5b)$$

$$\approx 0.22(\sigma^2_{\text{exp}}/\sigma^2_{\text{res}}) \qquad (5c)$$

where (5a) follows from well-known results about linear regression and (5b) follows from Taylor expansion for small values of $(\sigma^2_{\text{exp}}/\sigma^2_{\text{res}})$. Combining these two results, the number of progeny required so that the LOD score is expected to exceed T is

$$T/\text{ELOD} \approx (Z_\alpha)^2(\sigma^2_{\text{res}}/\sigma^2_{\text{exp}}). \qquad (6)$$

This confirms that the maximum likelihood approach agrees with the result (3) from the traditional approach above, when examining effects at a single marker locus. The more general framework of maximum likelihood, however, allows the method to be extended to more complex situations described below.

Interval mapping: If genetic markers have been scored throughout the genome, the method of maximum likelihood can be used as above to estimate the phenotypic effect and the LOD score for a putative QTL at *any* given genetic location (*cf.* LANDER and BOTSTEIN 1986a, b). The main difference is that the QTL genotype g_i for individual i is unknown; the appropriate likelihood function is therefore

$$L(a, b, \sigma^2) = \Pi_i[G_i(0)L_i(0) + G_i(1)L_i(1)], \qquad (7)$$

where $L_i(x) = z((\phi_i - (a + bx)), \sigma^2)$ denotes the likelihood function for the individual i assuming that $g_i = x$ and $G_i(x)$ denotes the probability that $g_i = x$ conditional on the genotypes and positions of the flanking markers. (Given a map function, G is easily computed. For example, if the flanking markers both have genotype AA in an individual and they lie at recombination fraction θ and θ' from the putative QTL, then the probability of the QTL genotype being AB is $\theta\theta'$, assuming no interference.) Note that (7) reduces to (4) in the special case that the QTL lies at a marker locus and the genotype g_i is thus known with certainty.

Finding the maximum likelihood solution (a^*, b^*, σ^{2*}) to (7) can be regarded as a linear regression problem with missing data: none of the independent variables (genotypes) are known; only probability distributions for each are available. Standard computer programs for linear regressions cannot be used:

instead, one must write a computer program to maximize the likelihood function explicitly. While any maximization method (*e.g.*, Newton's method) can be used, we have found it convenient to use recent techniques for maximum likelihood estimation with missing data (LITTLE and RUBIN 1987)—specifically, the EM algorithm (DEMPSTER, LAIRD and RUBIN 1977; LANDER and GREEN 1987). We have written a computer program MAPMAKER-QTL (S. LINCOLN and E. S. LANDER, unpublished) to compute LOD scores for putative QTLs in a backcross population. (A more complete program, also capable of handling F_2 intercrosses, is under development and will be made available.)

To illustrate the method, we have analyzed simulated data from many backcrosses. Figure 2 presents a *QTL likelihood map*, showing how the LOD score varies throughout a genome, for a simulated data set involving 250 backcross progeny segregating for five QTLs with various allelic effects. Based on the assumed genome size and density of markers, a LOD score of 2.4 is required (see below) for declaring the presence of a QTL. In the example, the four largest QTLs are detected while the fifth does not attain statistical significance. The approximate position of the QTLs is indicated by one-lod support intervals, defined by the points on the genetic map at which the likelihood ratio has fallen by a factor of 10 from the maximum.

QTL likelihood maps are closely analogous to location score maps used in human genetics, which display the classical LOD score for a qualitative trait and which often indicate gene positions by means of one-lod support intervals (OTT 1985).

Among the advantages of the approach are:

(i) The QTL likelihood map represents clearly the strength of the evidence for QTLs at various points along the entire genome.

(ii) In contrast to the traditional approach, the inferred phenotypic effects are asymptotically unbiased. This is an immediate consequence of the fact that they are MLEs for a correctly specified model (KENDALL and STUART 1979).

(iii) The probable position of the QTL is given by support intervals, indicating the range of points for which the likelihood ratio is within a factor of 10 (or 100, if desired) of the maximum.

(iv) Interval mapping requires fewer progeny than the traditional approach for the detection of QTLs. In meioses in which the flanking markers do not recombine, the genotype of the QTL is known almost certainly—up to the chance of a double crossover (*e.g.*, at most 1% in the case of a 20 cM RFLP map). In essence, the flanking markers can be thought of as a single tightly linked *virtual* marker in such meioses. Supposing that genetic markers are available every *d* cM and considering the (worst) case of a QTL in the middle of an interval, one can show (APPENDIX [A2]) that

$$\mathrm{ELOD}_{\text{interval mapping}} \approx (1 - 2\theta)^2\ \mathrm{ELOD}_0/(1 - \psi), \quad (8a)$$

where ψ is the recombination fraction corresponding to d cM, θ is the recombination fraction corresponding to $\frac{1}{2}d$ cM, and ELOD_0 is the expected LOD score for a marker located exactly at the QTL. By contrast, recall that

$$\mathrm{ELOD}_{\text{single markers}} \approx (1 - 2\theta)^2\ \mathrm{ELOD}_0. \quad (8b)$$

Interval mapping thus decreases the required number of progeny by a factor of $(1 - \psi)$. For maps with $d =$ 10, 20, 30 and 40 cM, the savings are 9%, 16%, 23% and 28%, respectively (where, as earlier, we assume the Haldane mapping function).

(v) QTL likelihood maps can also be used to distinguish a pair of linked QTLs from a single QTL, provided that they are not so close that recombination between them is very rare. Holding fixed the position of one QTL, the increase in LOD score caused by a second putative QTL can be computed for each position along the chromosome. An example is shown in Figure 3.

In addition to being tested on numerous simulated data sets, interval mapping has recently been applied in a companion paper (PATERSON *et al.* 1988) to an interspecific backcross in tomato: six QTLs affecting tomato fruit weight, four QTLs affecting the concentration of soluble solids, and five QTLs affecting fruit pH were mapped to about 20–30 cM.

In general, interval mapping should prove valuable for analyzing and presenting evidence for QTLs and for decreasing the number of progeny required to detect QTLs of a given magnitude.

Appropriate threshold for LOD scores: When an entire genome is tested for the presence of QTLs, the usual nominal significance level of 5% is clearly inadequate. Indeed, applying this standard which corresponds to a LOD score of 0.83 would have resulted in a spurious QTL being declared on chromosome *10* in Figure 2.

The appropriate threshold depends on the size of the genome and on the density of markers genotyped. To determine the correct LOD threshold, the issue is: If *no* QTLs are segregating, what is the chance that the LOD score will exceed the threshold *T somewhere* in the genome? It is useful to consider two limiting situations: (i) the *sparse-map* case, in which consecutive markers are well-separated, and (ii) the *dense-map* case, in which the spacing between consecutive markers approaches zero.

In the *sparse-map* case, occurrences of spuriously high LOD scores are essentially independent. To achieve an overall significance level of α when M

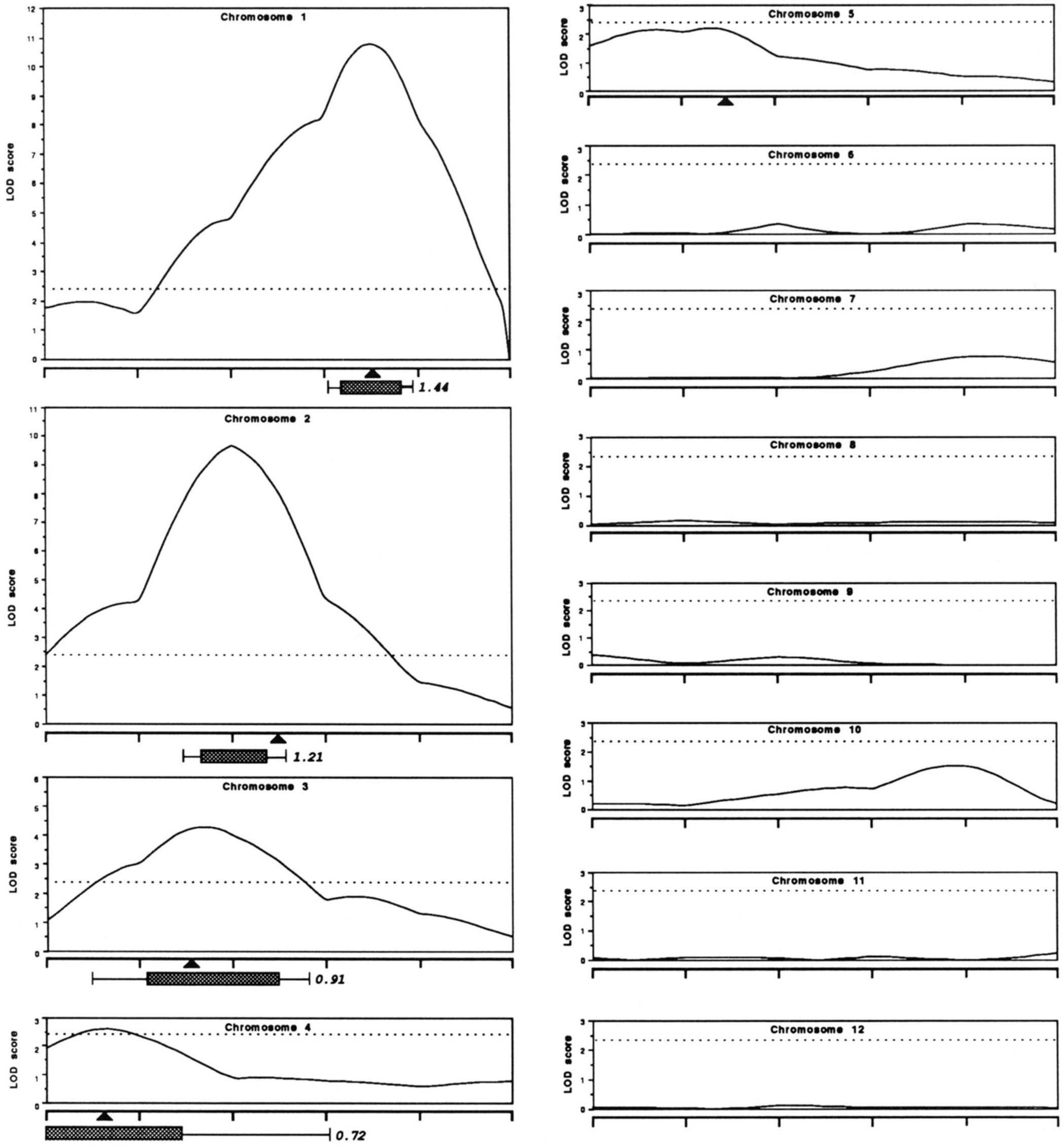

FIGURE 2.—LOD scores for a hypothetical quantitative trait. The LOD scores are based on simulated data for 250 backcross progeny in an organism with 12 chromosomes of 100 cM each. For each individual, crossovers were generated assuming no interference and genotypes recorded at RFLP markers spaced every 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 allelic effects at five QTLs and adding random environmental normal noise. Alleles at the QTLs had effects $\frac{1}{2}\delta$ = 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) Random environmental noise had standard deviation 1. No QTLs were located on chromosomes *6–12*. The dotted line at LOD = 2.4 indicates the required significance level. The four largest QTLs attained this LOD threshold. The grey bars indicate one-log support intervals for the position of the QTLs: outside this region, the odds ratio has fallen by a factor of 10. The thin lines extending from the gray bars indicate two-log confidence intervals. Maximum likelihood estimates 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, unpublished).

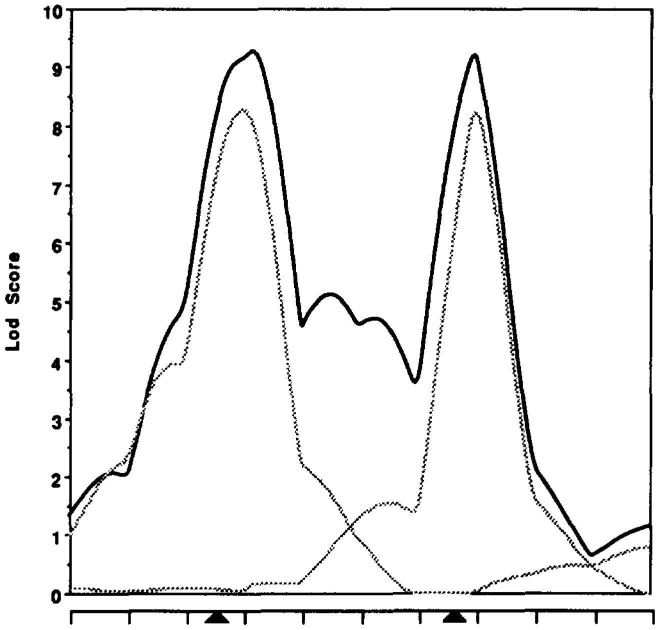

FIGURE 3.—LOD scores for a chromosome containing two QTLs. Data for 250 backcross progeny were simulated with a chromosome of 200 cM containing two QTLs with phenotypic effects $\frac{1}{2}\delta = 0.9$ at 50 cM and 130 cM from the left. The black curve shows the LOD scores, which suggests the presence of two QTLs. To test this, the gray curves were generated by computing the difference of (i) the LOD score with a QTL fixed at one position and a second QTL varying along the chromosome (computed by bivariate missing data regression) minus (ii) the LOD score with simply a QTL fixed at the position. After controlling for each peak, there remains strong evidence for the presence of a second peak. If the two QTLs are brought closer together, the number of progeny required to resolve them increases.

intervals are tested, a nominal significance level of α/M should be required for each individual test, corresponding to a LOD threshold of $\frac{1}{2}(\log_{10} e)(Z_{\alpha/M})^2$.

In the *dense-map* case, occurrences of spuriously high LOD scores at nearby markers are no longer independent events. As the number M of intervals tested tends to infinity (with each interval growing smaller), the required nominal significance level for each individual test approaches a nonzero limit independent of M. In fact, we prove in the APPENDIX [A3] that, in the limit of an infinitely dense-map and a large progeny size, the LOD score varies according to the square of an ORENSTEIN-UHLENBECK diffusion process. Well-known in physics and engineering, the ORENSTEIN-UHLENBECK diffusion describes a particle executing Brownian motion while being coupled to the origin by a weak spring. The extreme value properties of this diffusion have been extensively studied (LEADBETTER, LINDGREN and ROOTZEN 1983) and the results immediately translate into statements about how high a LOD score will be expected to occur by chance, given the size of the genome. Specifically, for a high threshold T, we have (see APPENDIX [A3]) the following result:

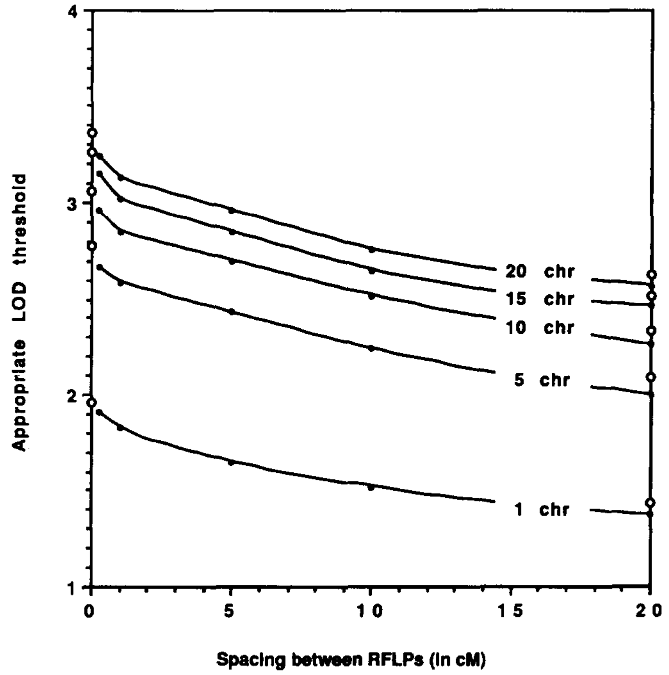

FIGURE 4.—LOD thresholds. Appropriate LOD threshold so that the chance of a false positive occurring *anywhere* in the genome is at most 5%, as a function of genome size and density of RFLPs scored. Chromosomes are assumed to be 100 cM in length—although approximately the same LOD threshold applies to any genome of the same total genetic length. The open circles at 0 cM correspond to the dense-map approximation and those at 20 cM correspond to the sparse-map approximation (see text), while each filled circle represents empirical results from 10,000 simulated trials. For example, a LOD threshold of about 2.4 would be required when using a 15 cM RFLP map of the tomato genome (≈1000 cM).

Proposition 2: *Consider an organism with C chromosomes and genetic length G, measured in Morgans. When no QTLs are present, the probability that the LOD score exceeds a high level T is $\approx (C + 2Gt)\ \chi^2(t)$, where $t = (2 \log 10)T$ and $\chi^2(t)$ denotes the cumulative distribution function of the χ^2 distribution with 1 d.f. In order to make the probability less than α that a false positive occurs somewhere in the genome, the appropriate LOD threshold is thus $\approx T_\alpha = (2 \log 10)t_\alpha$, where t_α solves the equation $\alpha = (C + 2Gt_\alpha)\chi^2(t_\alpha)$.*

For both the sparse-map and dense-map cases, a standard χ^2 table may thus be used to calculate the LOD score threshold corresponding to a 5% chance that even a single false positive will occur. For intermediate situations, we used extensive numerical simulation to determine the appropriate LOD thresholds as a function of genome size and marker spacing (Figure 4). Typically, a LOD threshold of between 2 and 3 is required to ensure an overall false positive rate of 5%. For instance, analyzing the domestic tomato (C = 12, $G \approx 11$) with a 20 cM RFLP map requires a LOD threshold of 2.4—equivalent to applying a nominal significance level of about $\alpha' = 0.001$ for each individual test performed. If the nominal 5%

significance level (LOD > 0.83) were used instead, one can show that the probability would exceed 90% that a false positive would arise *somewhere* in the genome. (Although a formal proof relies on the properties of ORENSTEIN-UHLENBECK diffusions, this essentially follows because $1 - (1 - 0.05)^{11/0.20} \approx 0.94$.) Indeed, a LOD score of 1.5 occurred by chance on chromosome 10 in the simulated data shown in Figure 2.

Number of progeny required: Given the ELOD for a QTL as a function of its phenotypic effect (Equation 8) and the LOD threshold T (Figure 4), a progeny size of T/ELOD will ensure a 50% chance of detecting linkage to such a QTL no matter where it lies in the genome. If it is desired to increase the chance of success to $100\beta\%$, standard arguments (KENDALL and STUART 1979) show that the progeny size should be further increased by a factor of $[1 + (Z_{1-\beta}/Z_{\alpha'})]^2$, where α' is the nominal significance level corresponding to a LOD score of T.

A technical note: The approximate progeny sizes given above (Equations 3, 5a, 5b, 6, 8a and 8b) are exact in the case of QTLs with small effects. Slight modifications are required for QTLs with large effects; see APPENDIX [A4].

Increasing the power of QTL mapping

Although interval mapping increases the efficiency of QTL mapping, large numbers of progeny may still be required. We therefore discuss additional methods to increase the power of QTL mapping, the most important of which is *selective genotyping.*

Selective genotyping of the extreme progeny: Some progeny contribute more linkage information than others. As a general principle, the individuals that provide the most linkage information are those whose genotype can be most clearly inferred from their phenotype. For example, LANDER and BOTSTEIN (1986b) have pointed out that the vast majority of linkage information about human diseases with incomplete penetrance comes from the affected individuals: since the genotype of unaffected individuals is uncertain, they provide relatively little information.

Applying this principle to quantitative genetics, the highest ELODs are provided by the progeny that deviate most from the phenotypic mean. When the cost of growing progeny is less than the cost of complete RFLP genotyping (as is frequently the case), it will thus be more efficient to increase the number of progeny grown but to genotype only those with the most extreme phenotypes. The increase in efficiency can be estimated as follows, with a more precise argument given in the APPENDIX [A5]. Since regression minimizes squared deviations from the mean, the ELOD conditional on an individual's phenotype ϕ is proportional to $(\phi - \mu_{B1})^2$. Thus, the proportion of individuals with extreme phenotype ϕ such that $|\phi - \mu_{B1}| \geq L$ is

$$Q(L) = 2 \int_L^\infty z(x)dx,$$

while the proportion of the linkage information contributed by such individuals is

$$S(L) = 2 \int_L^\infty x^2 z(x)\,dx$$
$$= Q(L)[1 + 2Lz(L)/Q(L)] \approx Q(L)[1 + L^2] \quad (9)$$

using integration by parts and the asymptotic approximation $z(L)/Q(L) \approx \frac{1}{2}L$ for large L (accurate to within only about 10–15% for small L). Accordingly, the same total linkage information would be obtained by growing a population that was larger by a factor of $h(L) = 1/S(L)$, but only genotyping individuals with extreme phenotypes. The number of progeny to genotype would fall by a factor of $g(L) = S(L)/Q(L) \approx [1 + L^2]$. Graphs of $Q(L)$, $S(L)$, $h(L)$ and $g(L)$ are shown in Figure 5. We observe that:

(i) Progeny with phenotypes more than 1 SD from the mean comprise about 33% of the total population but contribute about 81% of the total linkage information. By growing a population that was only about 25% larger and genotyping only these extreme progeny, the same total linkage information would be obtained from genotyping only about 40% as many individuals.

(ii) Progeny with phenotypes more than 2 SD from the mean comprise about 5% of the total population but contribute about 28% of the total linkage information. By growing a population that was about 3.6-fold larger and genotyping only these extreme progeny, the same total linkage information would be obtained from genotyping about 5.5-fold fewer individuals (since $h(2) = 3.6$ and $g(2) = 5.5$).

(iii) It is probably unwise to go beyond the 5% tails of the distribution. From a practical point of view, true phenotypic outliers may represent artifacts. Moreover, the increase in population size required for $L > 2$ outweighs the decreased number of individuals to genotype.

The strategy of *selective genotyping* will substantially increase efficiency whenever growing and phenotyping additional progeny requires less effort than completely genotyping individuals at all RFLP markers—which is typically the case in many organisms.

It sould be noted that standard computer programs for linear regression *cannot* be used (even for single marker analysis) when only the extreme progeny have been genotyped: phenotypic effects would be grossly overestimated because of the biased selection of progeny. As in the case of interval mapping, missing-data methods are required (LITTLE and RUBIN 1987). Con-

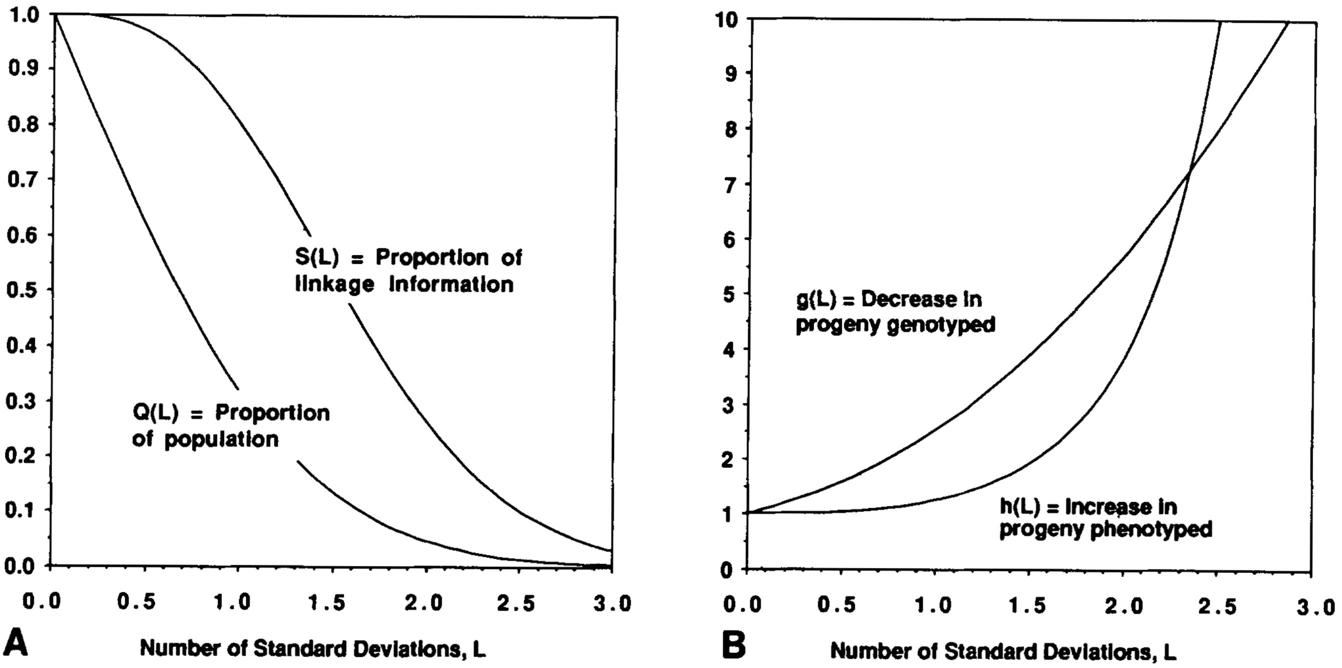

FIGURE 5.—Selective genotyping. A, Progeny having phenotypes exceeding mean by $\geq L$ standard deviations make up a proportion $Q(L)$ of population but account for a proportion $S(L)$ of the total LOD score for the progeny. B, If only individuals having phenotypes exceeding mean by $\geq L$ standard deviations are typed, the number of progeny genotyped may be decreased by a factor of $g(L)$ if the number of progeny grown and phenotyped is increased by a factor of $h(L)$.

veniently, the maximum likelihood methods discussed above will produce the correct results provided that the *phenotypes* are recorded for all progeny: *genotypes* for the nonextreme progeny may simply be entered as missing. Using the MAPMAKER-QTL program, we have thus been able to apply the method to both simulated and experimental data sets.

Decreasing environmental variance via progeny testing: As shown above, the number of progeny needed to map a QTL is proportional to

$$(\sigma^2_{\rm res}/\sigma^2_{\rm exp}) = [(\sigma^2_G + \sigma^2_E)/\sigma^2_{\rm exp}] - 1.$$

Typically, the environmental variance exceeds the genetic variance. If σ^2_E could be reduced, QTL mapping would become considerably more efficient. If the environmental noise results from measurement error, one might either average replicate measurements or try to develop a better assay. More often, environmental noise results from true physiological differences between genetically identical individuals. In this case, it may be possible to reduce σ^2_E through progeny testing: an individual's phenotype could be inferred indirectly from the average phenotype of n of its self or backcross offspring, since the variance of the average will be smaller. The effectiveness of this strategy may be limited, however, by unknown effects of dominance and epistasis. The approach will work best with recombinant inbred lines (see below), where isogenic individuals can be tested and averaged.

Simultaneous search: Just as environmental noise can be decreased via progeny testing, genetic noise can be reduced by simultaneously studying several intervals containing QTLs. If the genetic variance is large, such an approach may further decrease the number of progeny required. In the APPENDIX [A6], we discuss the extension of interval mapping to such simultaneous search (*cf.* LANDER and BOTSTEIN 1986a, b), the question of the appropriate LOD score when considering sets of intervals, and the approximate increase in the power of QTL mapping.

F_2 intercrosses and recombinant inbred strains: Although the discussion above concerns the backcross, it applies directly to F_2 intercrosses and recombinant inbred strains, with the following modifications:

(i) In an F_2 intercross, a QTL with phenotypic effect δ contributes variance $\delta^2/8$ and thus WRIGHT's formula (2) becomes $k = D^2/8\sigma^2_G$. Since F_2 intercrosses provide information about twice as many meioses as backcrosses of the same size, fewer progeny are required for detecting QTLs having purely additive effects: only 50–60% as many progeny are needed, depending on the density of the markers used (calculations not shown). If a QTL is partly dominant, one of the backcrosses will be more efficient and one less efficient for mapping it. The magnitude of dominance effects can be estimated by explicitly incorporating them into the maximum likelihood analysis via an additional parameter (see APPENDIX [A3]).

(ii) Recombinant inbred strains are analyzed in the same manner as backcrosses, except that the multigenerational breeding scheme that is used to construct recombinant inbred strains increases the effective genetic length of the genome. Compared to a backcross, the density of crossovers is doubled in a recombinant inbred strain produced through selfing and is quadrupled in a recombinant inbred strain produced by sib mating (HALDANE and WADDINGTON 1931). A genetic length of $2G$ or $4G$ must be used in place of G when computing the appropriate LOD threshold--leading to an increase of 0.3 or 0.6, respectively, in the threshold required. Although the higher threshold will increase the number of progeny required, the effect is typically offset by the ability to decrease the number of progeny by reducing the environmental variance through replicate phenotypic measurements within each recombinant inbred strain (*cf.* progeny testing above). Recombinant inbred strains will thus typically be more efficient for QTL mapping than equal number of backcross progeny. However, this advantage may often be negated by the considerable time and effort required to construct large numbers of such strains.

DISCUSSION

Although it has long been recognized that quantitative traits often arise from the combined action of multiple Mendelian factors, only recently has it become practical to undertake systematic mapping of such QTLs in experimental organisms (PATERSON *et al.* 1988). While such investigations will by no means be easy, the methodology developed here should increase their accuracy and efficiency. Specifically, by integrating information from genetic markers spaced throughout a genome, the method of *interval mapping* described above allows (i) efficient detection of QTLs while limiting the overall occurrence of false positives; (ii) accurate estimation of phenotypic effects of QTLs; and (iii) localization of QTLs to specific regions (Figure 2). Beyond the increased efficiency due to interval mapping, the strategy of *selective genotyping* can further reduce the number of progeny that must be genotyped in order to detect a QTL: together, the methods lead to a reduction of up to 7-fold in the number of progeny to be genotyped. (Interval mapping with a 40 cM RFLP map leads to a 1.28-fold reduction and selective genotyping of the 5% extremes leads to a 5.5-fold reduction.) Finally, additional savings may be achieved via *progeny testing* and *simultaneous search.* We summarize below the main considerations in designing a cross for genetic dissection of a quantitative trait.

Designing a cross for genetic dissection of a quantitative trait: Strains can be chosen to maximize the chance that they segregate for QTLs having relatively large phenotypic effects, thereby allowing mapping with a manageable number of progeny. The ideal situation occurs when (a) the phenotypic difference D between the strains is large compared to the environmental or within-strain standard deviation σ_E; (b) breeding experiments indicate that the number k of effective factors given by WRIGHT's formula is small; and (c) the strains are the result of selective breeding for the trait.

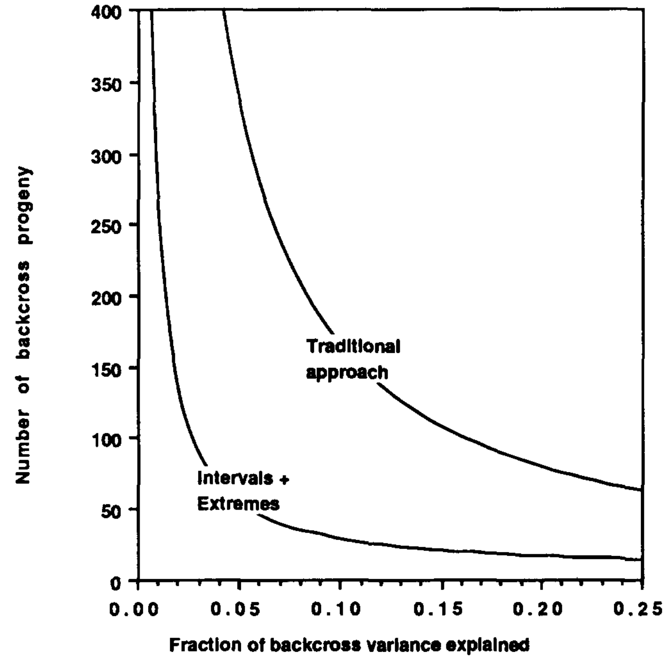

FIGURE 6.—Required progeny size. The number of backcross progeny that must be genotyped to map a QTL, based on the fraction of the backcross variance explained by the segregation of the QTL. The upper curve shows the traditional approach in which all progeny are genotyped and single markers analyzed. In the lower curve, only progeny with 5% most extreme phenotypes are genotyped and interval mapping is used to analyze the data. The calculations are based on use of a complete 20 cM RFLP map, a 50% chance of detection for QTLs in the middle of intervals, and a LOD threshold of 2.5. Note that for a QTL with phenotypic effect δ, the fraction of the backcross variance explained is $\delta^2/16\ \sigma^2_{B_1}$.

Once the strains have been chosen, the experimenter must specify the minimum phenotypic effect δ that the cross will be designed to detect. When using strains resulting from selection, a choice of δ in the range between $\frac{1}{2}(D/k)$ and (D/k) should ensure that QTLs accounting for much of the phenotypic difference will be detected. When using arbitrary strains, the same choice of δ can be used, although the presence of QTLs with this effect is not guaranteed.

The number N of backcross progeny that should be genotyped can then be calculated based on the spacing d between genetic markers in the map, the appropriate threshold T for the LOD score, and the desired probability β of success, assuming either (i) the traditional method of analysis involving single markers and genotyping of all progeny or (ii) interval mapping and selective genotyping of the 5% most extreme progeny. Figure 6 shows N as a function of the fraction of variation v explained by the QTL (where $v = \delta^2/$

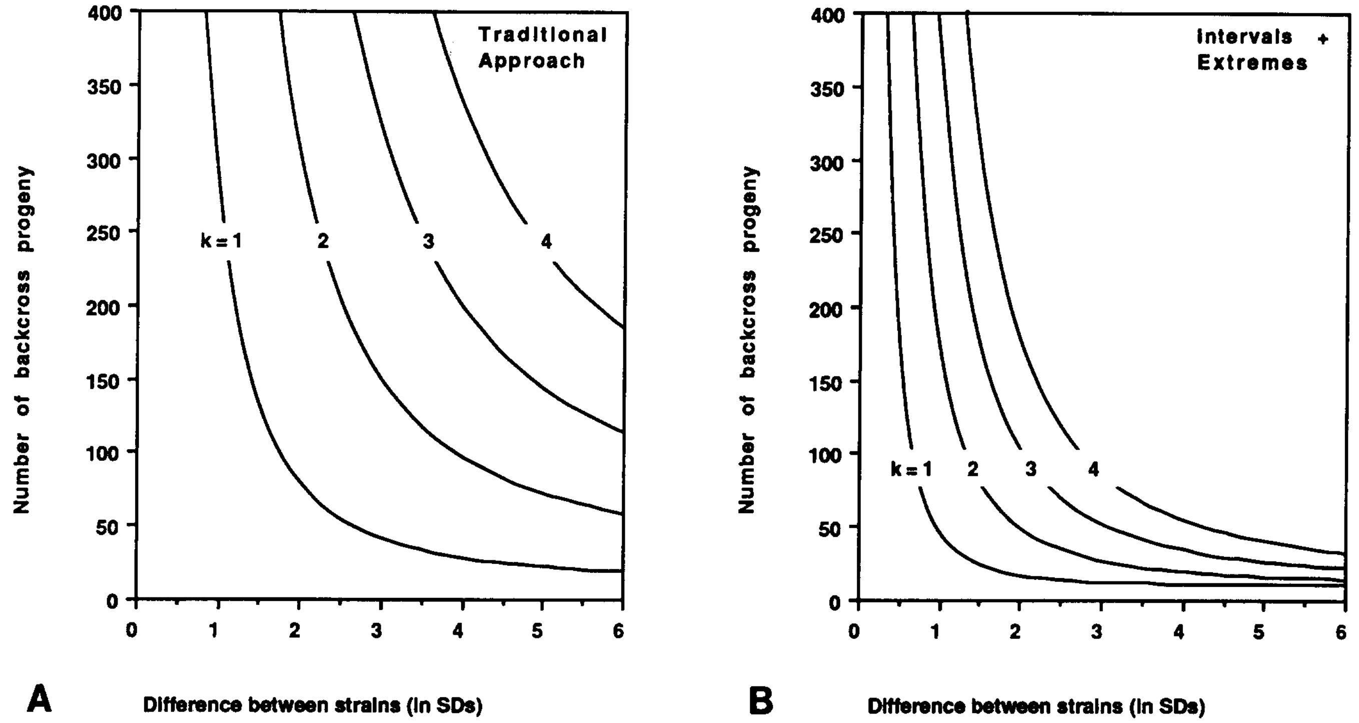

FIGURE 7.—Required progeny size. The number of backcross progeny that must be genotyped to map a QTL, based on the difference D between the strains (measured in environmental standard deviations) and the number k of effective factors. A, The traditional approach: all progeny are genotyped and single markers analyzed. B, Only progeny with 5% most extreme phenotypes are genotyped and interval mapping is used to analyze the data. The calculations are based on QTLs of equal phenotypic effect (D/k), use of a complete 20 cM RFLP map, a 50% chance of detection for QTLs in the middle of intervals, and a LOD threshold of 2.5 (corresponding to a nominal significance level $\alpha' \approx 0.001$). We indicate changes for different assumptions: multiply by 4 to allow for QTLs having half the average effect; multiply by approximately $(1.25)(1 - 2\theta)^2/(1 - \psi)$ to allow for markers every d cM (where θ and ψ are the recombination fractions corresponding to $\frac{1}{2}d$ and d cM, respectively); multiply by approximately 1.50 to allow for a 90% chance of success; multiply by $T/2.5$ to allow for a LOD threshold of T; and multiply by about 0.55 if an F_2 intercross is used instead of a backcross.

$16\sigma_{B1}^2$), while Figure 7, A and B, shows N as a function of the phenotypic difference D between the strains and the number k of effective factors. Together, interval mapping and selective genotyping reduce the number of progeny to be genotyped by up to 7-fold. (Both figures assume that $d = 20$ cM, $T = 2.5$ and $\beta = 0.50$, and Figure 7 assumes that the QTLs have equal phenotypic effects. The figure legend indicates how to modify the results for other values.) As a rule of thumb, it appears practical to map QTLs when the phenotypic difference D measured in environmental standard deviations is on the order of the number k of effective factors segregating.

An example: The Spontaneous Hypertensive rat (SHR) strain (TANASE *et al.* 1970), was derived from the Wistar-Kyoto rat (WKY) strain by selective breeding for high systolic blood pressure followed by inbreeding. Blood pressure in SHR is about 3 standard deviations higher than in WKY, while the number k of effective factors was estimated at about 3. Assuming that the rat genome is about 1500 cM and that a 20 cM RFLP map is available, the appropriate LOD threshold would be about 2.7 (Figure 4). Using the traditional approach, one would need about 325 backcross progeny or about 175 F_2 intercross progeny. With interval mapping, these become about 275 and 145. If it were practical to grow a larger population but genotype only those progeny with the 5% most extreme blood pressures, the number of progeny to genotype could be reduced to about 55 and 30, respectively.

In addition to SHR, a number of other genetically hypertensive strains of rat and mouse have been described, with estimated effective number of factors between 2 and 5 (DEJONG 1984). Study of these strains would elucidate the number and location of the most important genes controlling naturally occurring variation for blood pressure in rodent populations. Such information might shed light on hypertension in humans as well.

Other considerations: In this paper, we have been chiefly concerned with methods for mapping QTLs *per se.* For applications to agricultural breeding programs aimed at introgressing useful QTLs, additional considerations may apply. For example, (i) to avoid QTLs improving a trait of interest but having deleterious pleiotropic effects, one may wish to bias the choice of parental strains in certain ways and to score additional quantitative phenotypes pertinent to agronomic acceptability; and (ii) to minimize the total length of time for the breeding time, one may wish to genotype additional progeny in the hope of finding ones that have retained a fortuitously large proportion of the desired genetic background while gaining *some* of the desired QTLs (PATERSON *et al.* 1988). We will address such breeding considerations more fully elsewhere.

Conclusion: The availability of complete RFLP linkage maps should make it possible to dissect quantitative traits into discrete genetic factors, thereby unifying two historically-separated areas of genetics. Once QTLs have been mapped, isogenic lines can be rapidly constructed differing only in the region of the QTL by using the RFLPs to select for the desired region and against the remainder of the genome (TANKSLEY and RICK 1980; SOLLER and BECKMANN 1983; PATERSON *et al.* 1988). Using such isogenic lines, the fundamental tools of genetics and molecular biology may be brought to bear on the study of a trait—including testing of complementation and epistasis; characterization of physiological and biochemical differences between isogenic lines; isolation of additional alleles via mutagenesis or further selective breeding (at least in favorable systems); and, eventually, molecular cloning of the genes underlying quantitative inheritance.

We are grateful to STEPHEN LINCOLN for invaluable assistance in analyzing the simulated data sets, using the MAPMAKER-QTL computer package. We thank RICHARD ARRATIA, SIMEON BERMAN and DONALD RUBIN for discussions about stochastic processes. We thank ANDY PATERSON and an anonymous referee for helpful comments. This work was supported in part by the National Science Foundation (grant DCB-8611317) and by the System Development Foundation (grant G612).

LITERATURE CITED

ADLER, R. J., 1981 *The Geometry of Random Fields.* Wiley, New York.

BERMAN, S. M. 1982 Sojourns and extremes of stationary processes. Ann. Prob. **10:**1–46.

BOTSTEIN, D., R. L. WHITE, M. SKOLNICK and R. W. DAVIS, 1980 Construction of a genetic map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. **32:** 314–331.

CASTLE, W. E., 1921 An improved method of estimating the number of genetic factors concerned in cases of blending inheritance. Science **54:** 223.

COYNE, J. A., and B. CHARLESWORTH, 1986 Location of an X-linked factor in male hybrids of *Drosophila simulans* and *D. mauritiana.* Heredity **57:** 243–246.

DEJONG, W., 1984 *Handbook of Hypertension, Vol. 4: Experimental and Genetic Models of Hypertension.* Elsevier, New York.

DEMPSTER, A. P., N. M. LAIRD and D. B. RUBIN, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Stat. Soc. **39:** 1–38.

EAST, E. M., 1916 Studies on size inheritance in *Nicotiana.* Genetics **1:** 164–176.

EDWARDS, M. D., C. W. STUBER and J. F. WENDEL, 1987 Molecular-marker-facilitated investigation of quantitative-trait loci in maize. I. Numbers, genomic distribution and types of gene action. Genetics **116:** 113–125.

ELSTON, R. C., and J. STEWART, 1973 The analysis of quantitative traits for simple genetic models from parental, F_1 and backcross data. Genetics **73:** 695–711.

FALCONER, D. S., 1981 *Introduction to Quantitative Genetics.* Longman, London.

FESTING, M. F. W., 1979 *Inbred Strains in Biomedical Research.* Oxford University Press, Oxford.

HALDANE, J. B. S., and C. H. WADDINGTON, 1931 Inbreeding and linkage. Genetics **16:** 357–374.

HARLAN, J. R., 1976 Genetic resources in wild relatives of crops. Crop. Sci. **16:** 329–33.

HESTON, W. E., 1942 Inheritance of susceptibility to spontaneous pulmonary tumors in mice. JNCI **3:** 79–82.

JOHANNSEN, W., 1909 *Elemente der exakten Erblichkeitsliehre.* Fisher, Jena.

KALTER, H., 1954 The inheritance of susceptibility to the teratogenic action of cortisone in mice. Genetics **39:** 185–196.

KENDALL, M., and A. STUART, 1979 *The Advanced Theory of Statistics, Vol. 2.* Griffin, London.

LANDER, E. S., and D. BOTSTEIN, 1986a Strategies for studying heterogeneous genetic traits in humans by using a linkage map of restriction fragment length polymorphisms. Proc. Natl. Acad. Sci. USA **83:** 7353–7357.

LANDER, E. S., and D. BOTSTEIN, 1986b Mapping complex genetic traits in humans: New methods using a complete RFLP linkage map. Cold Spring Harbor Symp. Quant. Biol. **51:** 49–62.

LANDER, E. S., and P. GREEN, 1987 Construction of multilocus genetic linkage maps in humans. Proc. Natl. Acad. Sci. USA **84:** 2363–2367.

LEADBETTER, M. R., G. LINDGREN and H. ROOTZEN, 1983 *Extremes and Related Properties of Random Sequences and Processes.* Springer, New York.

LITTLE, R. J. A., and D. B. RUBIN, 1987 *Statistical Analysis with Missing Data.* Wiley, New York.

MALKINSON, A. M., AND D. S. BEER, 1983 Major effect on susceptibility to urethan-induced pulmonary adenoma by a single gene in BALB/cBy mice. JNCI **70:** 931–936.

MATHER, K., and J. L. JINKS, 1971 *Biometrical Genetics.* Cornell University Press, Ithaca, N.Y.

MORTON, N. E., 1955 Sequential tests for the detection of linkage. Am. J. Hum. Genet. **7:** 277–318.

NILSSON-EHLE, H., 1909 *Kreuzunguntersuchungen an Hafer und Weizen.* Lund.

OTT, J., 1985 *Analysis of Human Genetic Linkage.* Johns Hopkins Press, Baltimore.

PAIGEN, B., A. MORROW, C. BRANDON, D. MITCHELL and P. A. HOLMES, 1985 Variation in susceptibility to atherosclerosis among inbred strains of mouse. Atherosclerosis **57:** 65–73.

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: 721–726.

PROCHAZKA, M., E. H., LEITER, D. V. SERREZE and D. L. COLEMAN, 1987 Three recessive loci required for insulin-dependent diabetes in nonobese diabetic mice. Science **237:** 286–289.

RASMUSSON, J. M., 1933 A contribution to the theory of quantitative character inheritance. Hereditas **18:** 245–261.

RICK, C. M., 1973 Potential genetic resources in tomato: clues from observations in natural habitats. pp. 255–268. In: *Genes, Enzymes and Populations,* Edited by A. M. SRB. Plenum, New York.

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

SHIRE, J. G. M., 1968 Genes, hormones and behavioural variation. pp. 194–205. In: *Genetic and Environmental Influences on Behaviour,* Edited by J. M. THODAY and A. S. PARKS. Oliver & Boyd, Edinburgh.

SOLLER, M., and J. S. BECKMANN, 1983 Genetic polymorphism in varietal identification and genetic improvement. Theor. Appl. Genet. **47:** 179–190.

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:** 35–39.

STEWART, J., and R. C. ELSTON, 1973 Biometrical genetics with one or two loci: the inheritance of physiological characters in mice. Genetics **73:** 675–693.

TANASE, H., Y. SUZUKI, A. OOSHIMA, Y. YAMORI and K. OKAMOTO, 1970 Genetic analysis of blood pressure in spontaneously hypertensive rats. Jpn. Circ. J. **34:** 1197–1212.

TANKSLEY, S. D., and C. M. RICK, 1980 Isozymic gene linkage map of the tomato: Applications in genetic and breeding. Theor. Appl. Genet. **57:** 161–170.

TANKSLEY, S. D., H. MEDINA-FILHO and C. M. RICK, 1982 Use of naturally-occurring enzyme variation to detect and map genes controlling quantitative traits in an interspecific backcross of tomato. Heredity **49:** 11–25.

THODAY, J. M., 1961 Location of polygenes. Nature **191:** 368–370.

WRIGHT, S., 1968 *Evolution and the Genetics of Populations, Vol. 1, Genetic and Biometric Foundations.* University of Chicago Press, Chicago.

Communicating editor: E. THOMPSON

APPENDIX

[A1] To prove Proposition 1, we use the following lemma.

Lemma. *Let $x_1, \cdots, x_n \geq 0$, For $y \geq 0$, let $s_y = \Sigma' x_i$ and $t_y = \Sigma' x_i^2$, where the sum is taken over the terms $x_i \geq y$. If $t_0/s_0 \geq y$, then*

$$s_y \geq \tfrac{1}{2}[y + \sqrt{y^2 - 4(ys_0 - t_0)}] \quad \text{and} \quad t_y \geq t_0 - y(s_0 - s_y).$$

Proof. From the definitions and the non-negativity of the x_i, it is clear that

$$s_y^2 \geq t_y \geq t_0 - y(s_0 - s_y).$$

The constraint on s_y then follows by considering the outer terms and applying the quadratic formula. □

In the context of Proposition 1, suppose that the QTLs in the high strain change the phenotype by $x_1, \cdots, x_n \geq 0$, respectively. Using the notation above, we have $D = s_0$ and $\sigma_G^2 = t_0/16 = D^2/16k$ (because of nonlinkage among QTLs and WRIGHT's formula). Taking $y = \varepsilon(D/k)$, the result then follows from the lemma since $D_\varepsilon = s_y/s_0$, and $V_\varepsilon = t_y/t_0$.

[A2] Suppose that a QTL lies midway between two flanking markers. Let θ be the recombination fraction between the QTL and either marker and $\psi = 2\theta(1 - \theta)$ be the recombination fraction between the two markers (ignoring interference). In meioses in which they have not recombined (a proportion $1 - \psi$ of the total), the flanking markers act as a single virtual marker linked at recombination fraction γ, where γ is the chance that the QTL recombines with both markers given that the markers themselves have not recombined. By contrast, meioses in which the flanking markers have recombined provide zero information about linkage of the QTL. The ELOD for interval mapping is thus $(1 - \psi)$ times the ELOD for a single marker linked at γ which in turn is $(1 - 2\gamma)^2$ times the ELOD for a marker at 0% recombination.
That is,

$$\text{ELOD}_{\text{interval mapping}} = (1 - \psi)(1 - 2\gamma)^2 \text{ ELOD}.$$

Using the relation $\gamma = \theta^2/[(1 - \theta)^2 + \theta^2]$ and simplifying terms, Equation 8a follows.

[A3] In the idealized *dense-map* case, suppose that markers are available at every point along a chromosome. Suppose that there are no QTLs in the genome. For individual i, the phenotype $\phi_i = N(0, 1)$; that is ϕ_i is a random normal variable with mean 0 and variance 1. For individual i, let $x_i(d)$ denote the genotype at a position d cM from the left end of the chromosome ($x_i = 0$ or 1 according to the allele inherited), let $\beta^*(d)$ denote the maximum likelihood estimate of the phenotypic effect of a putative QTL at this position, and let LOD(d) denote the corresponding LOD score. By standard formulas for linear regression,

$$\beta^*(d) = \sum (\phi_i - \phi)(x_i - x)/\sum (x_i - x)^2$$

where ϕ and x are the means of ϕ_i and x_i, respectively. For a large population of size n, the central limit theorem implies that

$$\begin{aligned}
\beta^*(d) &\sim \sum 4\phi_i(x_i - \tfrac{1}{2})/n, \\
v(d) &:= \sqrt{n}\beta^*(d) \sim N(0, 1), \\
\sigma_{\text{res}}^2(d) &= \sum (\phi_i - \beta^*(d)x_i(d))^2 \sim n(1 - \beta^*(d))^2, \\
\sigma_{\text{exp}}^2(d) &\sim n[\beta^*(d)]^2
\end{aligned}$$

and

$$\begin{aligned}
\text{LOD}(d) &\sim \tfrac{1}{2}(\log_{10} e)(\sigma_{\text{exp}}^2(d)/\sigma_{\text{res}}^2(d)) \\
&\sim \tfrac{1}{2}(\log_{10} e)[\sqrt{n}\beta^*(d)]^2,
\end{aligned}$$

where we write $f \sim g$ to denote that $f/g \rightarrow 1$ as $n \rightarrow \infty$ and where ":=" indicates a definition. Thus, LOD(d) is asymptotically proportional to the square of a random normal variable $v(d)$ (which incidentally proves that LOD is proportional to χ^2). More generally, it is not difficult to see that the LOD score follows a stationary normal process--that is, the LOD score at multiple points has a multivariate normal distribution.

Let d_1 and d_2 denote points on the chromosome, let $d = d_1 - d_2$, and let θ be the recombination fraction corresponding to the genetic distance $d = |d_1 - d_2|$. The correlation coefficient between the variables $x_i(d_1)$ and $x_i(d_2)$ is easily seen to be $\rho(x_i(d_1), x_i(d_2)) = 1 - 2\theta$. From the asymptotic expression for $\beta^*(d)$ above, it then follows that

$$r(d) := \rho(v(d_1), v(d_2)) = \sum \sqrt{n}\phi_i(1 - 2\theta) \sim 1 - 2\theta.$$

Assuming HALDANE's map function, $1 - 2\theta = e^{-2d}$.

To summarize, $v(d)$ is a stationary normal process with covariance function $r(d) = e^{-2d}$. Up to rescaling d by a factor of ½, this is the definition of ORENSTEIN-UHLENBECK diffusion and Proposition 2 follows directly [see LEADBETTER, LINDGREN and ROOTSZEN (1983) Theorem 12.2.9 and discussion following]. While only HALDANE's map function yields precisely an ORENSTEIN-UHLENBECK diffusion, the proof of Proposition 2 holds in general. The relevant results in LEADBETTER, LINDGREN and ROOTSZEN (1983) require only that $r(d) \sim 1 - 2d + o(d^2)$ as $d \rightarrow 0$, which holds for all map functions.

These remarks carry over to the situation of mapping QTLs in an F_2 intercross by fitting both an additive and a dominance component. The only substantial difference is that the LOD score now follows a χ^2 process with 2 df. The large deviation theory for such processes has been worked out (Berman 1982). We will discuss its application elsewhere.

[A4] If QTLs with very large effects are segregating, regression analysis is not strictly appropriate (whether in the traditional approach or in the generalization developed in the text) because the phenotypic distribution becomes bimodal. When the phenotypic distribution is bimodal due to the segregation of a QTL with large effects somewhere in the genome, it is no longer possible to use a simple normal distribution as the null hypothesis. (The fit would be so bad that one would always reject the null hypothesis in favor of the presence of a QTL, even at positions unlinked to any QTL.) A good remedy is to use an appropriate null hypothesis, reflecting the fact that the phenotypic distribution may represent the mixture of two normals caused by the segregation of an unlinked QTL. The LOD score for a marker at 0 cM can be redefined as the $\log_{10}$ of

$$L(\hat{a},\ \hat{b},\ \hat{\sigma}^2)/\tfrac{1}{2}[L(\hat{a},\ \hat{b},\ \hat{\sigma}^2) + L(\hat{a},\ -\hat{b},\ \hat{\sigma}^2)]$$

with $L(a,\ b,\ \sigma^2)$ defined in (4). (This ratio measures how much more likely the data are to have been generated by a QTL with the hypothesized effect located at the marker locus than by a QTL with this same effect but unlinked to the marker.) The ELOD can be found by numerical integration over the distribution for ϕ. In the limit of a QTL with large effect, the expression tends to the traditional LOD score for a qualitative trait used in human genetics. For QTLs with small effects, the expression does not differ significantly from the LOD score defined above (since the mixture of the two normal distributions closely resembles a single normal distribution with larger variance).

For the QTLs likely to be encountered in practice, this correction is irrelevant. We have used it in computing the number of progeny required in Figures 5 and 6, however, in order that these graphs exhibit the correct limiting behavior—rather than tending to zero.

[A5] For notational convenience, rescale the phenotype so that its mean in the backcross is 0 and encode the two alternative genotypes by the indicator variable $g = -1$ or 1 (rather than 0 or 1, as in the text). Given a true QTL, let $2b$ be the amount by which substituting an allele increases the phenotype and let σ^2 be the residual variance unexplained by the QTL out of the total backcross variance $\Sigma^2 = \sigma^2 + b^2$. Suppose that a marker is located exactly at the QTL. Conditional on the phenotype ϕ of an individual but unconditional on its genotype x at the marker, the LOD score (comparing the true hypothesis H_1:$(0,\ b,\ \sigma^2)$ to the alternative H_0:$(0,\ 0,\ \Sigma^2)$) is

$$\mathrm{LOD}_\phi = \sum_{g=0,1} \pi(g\,|\,\phi)\ \log_{10}[z(\phi - bg,\ \sigma^2)/z(\phi,\ \Sigma^2)]$$

where $\pi(g = x\,|\,\phi)$ is the probability that the individual has marker genotype x given its phenotype ϕ, given by

$$z(\phi - bg,\ \sigma^2)/[z(\phi - bg,\ \sigma^2) + z(\phi + bg,\ \sigma^2)].$$

As claimed in the text, if b is small, LOD_ϕ is proportional to ϕ^2. Now, the probability distribution for ϕ has density

$$p(\phi) = \tfrac{1}{2}[z(\phi - bg,\ \sigma^2) + z(\phi + bg,\ \sigma^2)].$$

Conditional on the phenotype of a backcross progeny deviating from the mean by $>\mathrm{L}\Sigma$, the LOD score is

$$\mathrm{LOD}_{|\phi| \geq L\Sigma} = \int_{|\phi| \geq L\Sigma} (\mathrm{LOD}_\phi)p(\phi)d\phi.$$

Letting $v = b^2/\Sigma^2$ denote the fraction of variance explained by the QTL, straightforward though tedious integration shows that

$$\begin{aligned} \mathrm{S(L)} &= \mathrm{LOD}_{|\phi| \geq L\Sigma}/\mathrm{LOD}_{|\phi| \geq 0} \\ &\approx Q(L)[1 + 2uLz(L)/Q(L)], \end{aligned} \tag{10}$$

where $u = -v/\log_e(1 - v) \approx (1 - \tfrac{1}{2}v)$ and where the approximation in (10) is $o(v^2)$ for small v. For QTLs with small effects, this reduces to Equation 9.

[A6] Interval mapping can be straightforwardly extended to the case of *multiple* intervals explaining a quantitative phenotype: for m intervals, the bracketed term in Equation 7 becomes a sum with 2^m terms corresponding to the possible joint genotypes at the m putative QTLs. Since simultaneous consideration of multiple QTLs reduces the unexplained variance, it may be somewhat easier to detect linkage to the set of loci than to any one individually (*cf.* LANDER and BOTSTEIN 1986a, b)—although there are possible difficulties in parameter estimation and model identifiability. The subtle issue is the appropriate threshold for simultaneous search for m QTLs. In a genome with no QTLs, how high a LOD score might occur by chance? For any particular choice of putative QTLs, the LOD score is asymptotically distributed as χ^2 with m degrees of freedom. When considering sets of m loci chosen from an entire genome, the LOD score follows a mathematical process known as a χ^2 random field (ADLER 1981)—about which somewhat less is known than the ORENSTEIN-UHLENBECK diffusion. Approximate arguments show that the level of highest excursion of such a χ^2 random field on an entire genome is about m-fold higher than the corresponding level for an ORENSTEIN-UHLENBECK diffusion on the genome. If m QTLs have equal effects, then simultaneous search decreases the number of progeny required to achieve statistical significance by a factor of about $(1 - m\sigma^2)/(1 - \sigma^2)$, where σ^2 is the fraction of variance explained by each. If the QTLs have unequal effects, it may become possible to detect those with smaller effects by first controlling for those with larger effects. We will discuss simultaneous search for QTLs in more detail elsewhere.