Statistical Methods for Mapping Quantitative Trait Loci From a Dense Set of Markers

Josée Dupuis* and David Siegmund[†]

*Genome Therapeutics Corporation, Waltham, Massachusetts 02453 and [†]Department of Statistics, Stanford University, Stanford, California 94305

Manuscript received April 14, 1997
Accepted for publication September 21, 1998

ABSTRACT

Lander and Botstein introduced statistical methods for searching an entire genome for quantitative trait loci (QTL) in experimental organisms, with emphasis on a backcross design and QTL having only additive effects. We extend their results to intercross and other designs, and we compare the power of the resulting test as a function of the magnitude of the additive and dominance effects, the sample size and intermarker distances. We also compare three methods for constructing confidence regions for a QTL: likelihood regions, Bayesian credible sets, and support regions. We show that with an appropriate evaluation of the coverage probability a support region is approximately a confidence region, and we provide a theroretical explanation of the empirical observation that the size of the support region is proportional to the sample size, not the square root of the sample size, as one might expect from standard statistical theory.

RECENT advances in genetics have led to the identification of genes responsible for certain diseases such as cystic fibrosis, Huntington's disease, breast cancer, and others. Linkage analysis, which is especially effective when the disease or trait of interest exhibits Mendelian inheritance, played an important role in the identification of those genetic loci. When the disease is complex in nature (incomplete penetrance, multiple loci involved, etc.) or quantitative, finding the genetic loci involved in the etiology of the trait can be more difficult. In particular, in human studies, it is difficult to separate environmental and genetic effects. However, with experimental organisms, studies can be designed to provide a similar environment for all individuals, so that the variation in phenotypes can be attributed mainly to genetic factors; and breeding designs can control the nature of the differences in genotype. Studies of experimental organisms can provide useful information for agricultural purposes and/or contribute to our understanding of human disease via animal models. Moreover, it is now feasible to search the entire genome for a gene locus influencing a trait of interest. Statistical methods for mapping quantitative trait loci (QTL) from experimental crosses using a dense set of markers were introduced by Lander and Botstein (1989). Applications have involved (i) tomatoes (Paterson *et al.* 1991) to identify loci influencing traits such as mass per fruit, pH, and soluble solid concentration; (ii) grain yield in maize (Stuber *et al.* 1992); (iii) high blood pressure in rats (Jacob *et al.* 1991); and (iv) fatness and growth rate in pigs (Andersson *et al.* 1994). In their original article, Lander and Botstein suggested statistical tests for general designs, but provided guidelines for declaring statistical significance for the backcross design only. Paterson *et al.* (1991) used these guidelines for intercross designs, but to avoid an increase in the false-positive error rate, they restricted themselves to a 1-d.f. statistic that ignored dominance effects. Churchill and Doerge (1994) proposed use of the permutation distribution to define thresholds for all design types. This method has the advantage that it makes no assumptions on the distribution of the phenotype. However, the thresholds depend on the observed data, so they need to be computed by Monte Carlo for each study; hence the method is less useful for analyzing and comparing different designs.

In this article we propose for intercross and other designs simple approximations that can be used to compare different designs under various conditions or the same design for different sample sizes or marker densities. We also discuss and compare three methods for constructing confidence intervals for a QTL. We assume throughout that markers are equally spaced, that there are no missing data, and except where noted that recombination occurs without interference. While these are artificially simple assumptions, at the cost of some complication they can all be weakened. Rough preliminary calculations suggest that the resulting picture would not change substantially unless the assumptions are radically altered. The sections on QTL detection and confidence regions are independent and can be read in any order.

Corresponding author: Josée Dupuis, Genome Therapeutics Corporation, 100 Beaver St., Waltham, MA 02453-8443.
E-mail: josee.dupuis@genomecorp.com

RESULTS

The model and likelihood ratio statistics: The starting point for our considerations is a cross between two strains that differ substantially in the quantitative trait of interest. The parental lines can be "pure" breeding lines obtained through inbreeding or simply two different strains of the same organism with widely differing mean phenotype. A cross is obtained from the two parental lines, creating the first generation of offspring (generation F_1). The F_1 generation is then allowed to mate together to produce the second generation (F_2), the intercross. We assume that the genotypes of the parental lines are completely different, so that at any marker locus we can label alleles from the strain with the larger mean phenotype as *A*, and alleles from the other strain as *B*. At each locus, each individual of the F_2 generation will have zero, one, or two *A* alleles. A backcross is generated by mating an individual of the F_1 generation to one from the parental line. If the parental line with the smaller mean for the trait is used, the offspring from the backcross will have zero or one *A* alleles at any locus on their genome.

A standard model for quantitative traits (*e.g.*, Kempthorne 1957) in notation suitable for our purposes is the following. Let y_i be the phenotypic value of individual *i*, and let $x_{ij}(d)$ be the number of *A* alleles at locus *d* on the *j*th chromosome. The locus is identified by its genetic distance *d* from one end of the chromosome. If there exists only one QTL on the *j*th chromosome that influences the traits and its location is *q*, the phenotype can be modeled as

$$y_i = \mu + \alpha x_{ij}(q) + \delta 1_{(x_{ij}(q)=1)} + e_{ij}, \tag{1}$$

where μ, α, δ are the phenotypic mean, additive effect, and dominance effect, respectively, and 1_C equals 1 or 0 according to whether the condition *C* is satisfied or not. The e_{ij}'s are residual effects, which include both environmental effects and the genetic effects of QTL on other chromosomes than the *j*th. As we will be considering only a single chromosome at a time, we drop the subscript *j* in what follows. We assume that $x_i(q)$ and e_i are uncorrelated, which would be the case if there is no epistasis and the environmental effect is uncorrelated with the genetic effects. We also assume that the e_i are independent normally distributed random variables with mean 0 and variance σ_e^2. The residual variance σ_e^2 equals the sum of the environmental variance and the genetic variance for those QTL not on the *j*th chromosome. Without the normality assumption the regression-like statistics given below are not exact maximum-log-likelihood ratios, so it is possible that more powerful tests can be found. However, by virtue of the central limit theorem the various approximations to significance level, power, etc. will still be valid in large samples even if the *e*'s are not normally distributed. In fact, for the significance level, it is not necessary to assume any stochastic model for the *y*'s. One can simply regard the *y*'s as fixed numbers and the regression statistic essentially a weighted (by the *y*'s) sum of the *x*'s, to which the central limit theorem applies under an assumption that the empirical behavior of the *y*'s is about what it would be if they were independent and identically distributed observations from a fixed distribution.

For backcross data, because $x_i(q) = 0$ or 1, the additive and dominance effects cannot be estimated separately, and the model reduces to

$$y_i = \mu + \alpha^* x_i(q) + e_i, \tag{2}$$

where the parameter α^* in (2) equals $\alpha + \delta$ from the model (1). This is the model developed by Lander and Botstein (1989), which we review briefly here. Treatment of the full model (1) is shown later in this article.

If one observes the genotype of a marker at a putative trait locus *d*, the maximum-log-likelihood ratio at *d* is given approximately by

$$2 \ln \mathrm{LR}(d) \approx -N \ln(1 - \hat{\alpha}_d^2/4\hat{\sigma}_y^2) \approx \frac{N\hat{\alpha}_d^2}{4\sigma_e^2}, \tag{3}$$

where *N* is the number of typed individuals, $\hat{\alpha}_d$ is the maximum-likelihood estimate of the parameter $\alpha^* = \alpha + \delta$, and $\hat{\sigma}_y^2$ is the maximum-likelihood estimate of the phenotypic variance $\sigma_y^2 = \sigma_e^2 + \alpha^{*2}/4$. It is important to note that both σ_y^2 and σ_e^2 depend on the design and for a backcross differ from the corresponding quantities for an intercross, although this difference is not reflected in the notation. Note also that (3) involves natural logarithms; the marginal asymptotic distribution of (3) at any unlinked locus is χ^2 with 1 d.f. To convert this and subsequent expressions to the LOD scale, one can divide by $2 \ln 10 \approx 4.6$. For the first approximation in (3) we have replaced the empirical variance of $\{x_i(d)\}$, namely $N^{-1}\Sigma_i[x_i(d) - N^{-1}\Sigma_j x_j(d)]^2$, by its asymptotic value of $\frac{1}{4}$; for the second we have approximated the logarithm by the first term of its Taylor expansion and have replaced the estimate $\hat{\sigma}_y$ by the parameter σ_e that it estimates under the hypothesis of no linkage on the *j*th chromosome. Since the trait locus *q* is typically unknown, the log-likelihood ratio is maximized over all marker locations *d* and chromosomes *j*. At each marker, assumed to be a QTL, the log-likelihood ratio is computed exactly. Between markers, Lander and Botstein (1989) suggest the use of "interval mapping," which consists of treating the unobserved marker information as missing data and using the EM algorithm (Dempster *et al.* 1977) to evaluate the log-likelihood ratio at *d* based on the marker information at the flanking markers. A noniterative, regression-based alternative to the EM algorithm was proposed by Haley and Knott (1992) and was shown to give equivalent results provided *N* is sufficiently large.

Detection of linkage in backcrosses: Because the log-likelihood ratio is maximized over the entire genome, it is unclear whether the conventional threshold of LOD = 3.0 [equivalently $2 \ln \mathrm{LR}(d) > 13.8$] to declare statistical significance is appropriate in the present context. To address this issue, Lander and Botstein (1989) proposed the approximation of $N^{1/2}\widehat{\alpha_d}/2\sigma_e$ [*cf.* (3)] by an Ornstein-Uhlenbeck process. This can be justified by the central limit theorem and a straightforward calculation of covariances. For the case of complete marker information (continuous markers), they gave thresholds depending on the length of the genome and the number of chromosomes searched (*cf.* their Proposition 2). For the case of a discrete set of markers evenly distributed over the genome, they obtained thresholds from a simulation study conducted under the assumption of no interference.

For the case of equispaced markers along the genome, Feingold *et al.* (1993) proposed an approximation, which agrees closely with the results from Lander and Botstein's simulations. That approximation is

$$P\{\max_k 2 \ln \mathrm{LR}(k\Delta) > a\} \approx 1 - \exp\{-2C[1 - \Phi(b)] - 2\beta L b \phi(b) \nu(b\{2\beta\Delta\}^{1/2})\}, \tag{4}$$

where $a = b^2$, L is the total length of the genome, C is the number of chromosomes, $\beta = 2\lambda$, λ being the rate of crossovers ($\lambda = 1$ if L is in Morgans and $\lambda = 0.01$ if L is in centimorgans), Δ is the distance between markers in the same units as L, and $\Phi(x)$ and $\phi(x)$ are the standard normal cumulative and density function, respectively. The function ν is a discreteness correction for the distance Δ between markers. The defining expression can be found in Siegmund (1985), p. 82. Often it is adequate to approximate $\nu(x)$ by $\exp(-0.583x)$, which is valid for $x < \sim 2$, while for $x > 2$ the first four terms of the defining infinite series provide a reasonable approximation. For the case of continuous markers $\Delta = 0$, so $\nu = 1$, and (4) is essentially the same as the approximation of Lander and Botstein (1989).

For a backcross design with a QTL located exactly at a marker, Feingold *et al.* (1993) gave as an approximation for the power

$$P\{\max_k 2 \ln \mathrm{LR}(k\Delta) > a\} \approx 1 - \Phi(b - \xi) + \phi(b - \xi)[2\nu/\xi - \nu^2/(b + \xi)^2], \tag{5}$$

where $a = b^2$, $\xi = \{N \ln[1 + (\alpha + \delta)^2/4\sigma_e^2]\}^{1/2}$, and $\nu = \nu(b\{2\beta\Delta\}^{1/2})$, as defined previously. The parameter ξ is the noncentrality parameter of (3) expressed in terms of the parameters of the model (2). The first term in (5) is the probability the process is above the threshold at the QTL; the second is the probability that it is below at the QTL but crosses the threshold at some nearby marker. Unless the markers are closely spaced, the first term by itself is a reasonably good approximation. When the QTL is located between markers, it is necessary to analyze the (correlated) process at the two flanking markers. The more complex approximation, which requires a one-dimensional numerical integration, can be found in Dupuis (1994). The noncentrality parameter at a flanking marker at distance Δ_1 from the QTL is

$$\xi \exp(-\beta\Delta_1), \tag{6}$$

where β and ξ are as defined above.

From (6) and (4) we see the importance of the parameter β, which equals 0.02 for backcross designs, but can assume a larger value for other designs (*e.g.*, recombinant inbred designs). In (4), β multiplies the length of the genome, so a larger value requires a larger threshold to maintain a given false-positive error rate. From (6) we see that it also governs the rate at which the noncentrality parameter decays as a function of the distance from QTL to flanking marker. A large value of β means a rapid falling off in power to detect the QTL as a function of that distance. On the other hand, it also provides the possibility for more precise fine mapping of the QTL location, because a large β leads to a sharper delineation of the "peak" in the process $2 \ln \mathrm{LR}(d)$ that identifies the location of the QTL. We return to these issues below.

The preceding analysis is concerned with the likelihood ratio process observed at the discrete set of marker loci. To mitigate the problems indicated by (6) when the QTL is in the center of a marker interval, Lander and Botstein (1989) suggested the technique of interval mapping, *i.e.*, treating the unobserved intervals between marker loci as missing data and using the EM algorithm to interpolate between the observed data points. Rebai *et al.* (1994, 1995) have used Rice's formula for the expected number of upcrossings of a level by a piecewise smooth Gaussian process to give approximations for the false-positive rates when using interval mapping. The method is analytically tractable when one assumes complete interference, *i.e.*, the recombination probability and map distance in Morgans are equal. Single chromosome simulations performed by these authors and our own whole genome simulations (data not shown) indicate that the approximation is very good when the sample size is reasonably large and markers are not too closely spaced. For dense markers (~1 cM) it is conservative. A modification suitable for small samples can be inferred from Johnstone and Siegmund (1989).

An argument of Siegmund and Worsley (1995) can be adapted to give a simple approximation for the power of an interval mapping test. See appendix a.

Intercrosses: Most previous theoretical analyses have concentrated on backcrosses and consequently have ignored dominance effects. Paterson *et al.* (1991) used the full model (1) to locate QTL in tomatoes in an intercross, and estimated the dominance effects. However, to detect linkage, they used a 1-d.f. statistic that ignores the dominance effects. Here we analyze the

2-d.f. statistic involving both additive and dominance effects.

Consider the likelihood ratio statistic to test the general hypothesis that $\alpha = \delta = 0$ *vs.* the alternative that $\alpha \neq 0$ or $\delta \neq 0$. For intercross data the vectors with coordinates $x_i(d)$ and $1_{(x_i(d)=1)}$ $(i = 1, \ldots, N)$ are asymptotically orthogonal. Therefore, the approximations used to obtain (3) now yield for the log-likelihood ratio at the marker d

$$2 \ln \mathrm{LR}(d) \approx -N \ln\left\{1 - \frac{\widehat{\alpha_d}^2/2 + \hat{\delta}_d^2/4}{\hat{\sigma}_y^2}\right\}$$
$$\approx \left[\left(\frac{N^{1/2}\hat{\alpha}_d}{2^{1/2}\sigma_e}\right)^2 + \left(\frac{N^{1/2}\hat{\delta}_d}{2\sigma_e}\right)^2\right]. \quad (7)$$

To define a significance level, we give an approximation under the hypothesis of no linkage to the distribution of the maximum of (7) over all possible values of d.

Let

$$X_d = \frac{N^{1/2}\hat{\alpha}_d}{2^{1/2}\sigma_e} \quad \text{and} \quad Y_d = \frac{N^{1/2}\hat{\delta}_d}{2\sigma_e}. \quad (8)$$

A straightforward application of the central limit theorem and calculation of covariances shows that when $\alpha = \delta = 0$, for large N, X_d and Y_d are approximately independent Ornstein-Uhlenbeck processes with mean 0 and covariance functions $e^{-2\lambda|t|}$ and $e^{-4\lambda|t|}$, respectively. An approximation to the tail distribution of the maximum of (7) is provided by

$$P\{\max_d 2 \ln \mathrm{LR}(d) \geq a\}$$
$$\approx 1 - \exp\left\{-\left[C + \nu b^2 L\left(\frac{\beta_1 + \beta_2}{2}\right)\right]\exp(-b^2/2)\right\}, \quad (9)$$

where $\beta_1 = 2\lambda$, $\beta_2 = 4\lambda$, $a = b^2$, and $\nu = \nu(b\{\Delta(\beta_1 + \beta_2)\}^{1/2})$. As in the case of (4), this approximation does not take interval mapping into account. It is obtained by a suitable modification of Woodroofe's (1976) argument. For an idealized tomato genome consisting of 12 chromosomes of length 100 cM each and a dense set ($\Delta = 0$) of markers, the 0.05 false-positive threshold obtained from (9) is $a = 19.0$ (LOD = 4.13), in comparison with $a = 14.6$ (LOD = 3.17) for the backcross case. Although smaller thresholds are required when the intermarker distance is greater, for an intercross the conventional LOD = 3 threshold would lead to a false-positive rate greater than 0.05 even for intermarker distances of 25 cM. This stands in contrast to the case of a backcross, where the LOD = 3 threshold is conservative for intermarker distances down to ~1 cM.

Rebai *et al.* (1995) have given an approximation for the false-positive error rate when interval mapping is used. This approximation involves an elliptic integral, to be evaluated numerically, and so is more complicated than the analogous backcross approximation, which can be written in closed form involving only the exponential and inverse tangent functions. In fact, the mathematically correct form of (9) involves similar complications, although extensive numerical calculations show that there is very little difference between the mathematically correct approximation and the more convenient one given above, which is based on replacing the two parameters β_1 and β_2 associated with the two coordinate processes by their average value, $(\beta_1 + \beta_2)/2$. In this spirit one can modify the approximation of Rebai *et al.* (1995) to obtain a closed form approximation that is no more complicated than that obtained for a backcross and gives essentially the same numerical results as the more complicated, mathematically correct approximation.

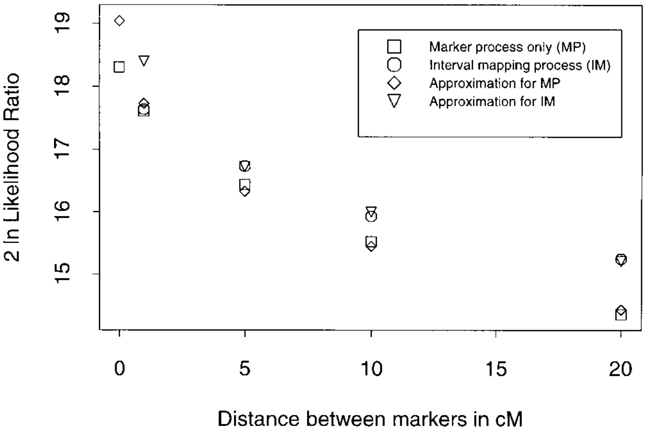

Figure 1.—Thresholds for 350 simulated tomato genomes.

To check the accuracy of (9) and our interval mapping approximation, we simulated thresholds for the log-likelihood ratio based on an intercross sample of $N = 350$ organisms with 12 chromosomes of total length 1200 cM (to approximate the tomato genome). The interval mapping step was performed using an approximation due to Haley and Knott (1992), which is much less computer intensive and gives results almost identical to the EM algorithm for large values of N. Results are shown in Figure 1.

Both approximations are very accurate. As predicted, the process with the interval mapping step requires a higher threshold for a given value of the Type-I error. For smaller N, somewhat different approximations yielding larger thresholds need to be used, since the given approximations do not take into account the variability in the estimate of the variance, σ_y^2. However, when N is large (at least 200), the approximations provide thresholds for the statistic and marker density actually used, which are more appropriate than the conventional LOD = 3.0. In mapping human traits, Lander and Kruglyak (1995) have argued that because the investigator is likely to type more markers around promising loci, the threshold for $\Delta = 0$ should be used in all cases. If we use this threshold, it is not necessary to rationalize the choice of Δ, which should otherwise be an average intermarker distance in the neighborhood of detected linkages, or to concern ourselves about the effect of interval map-

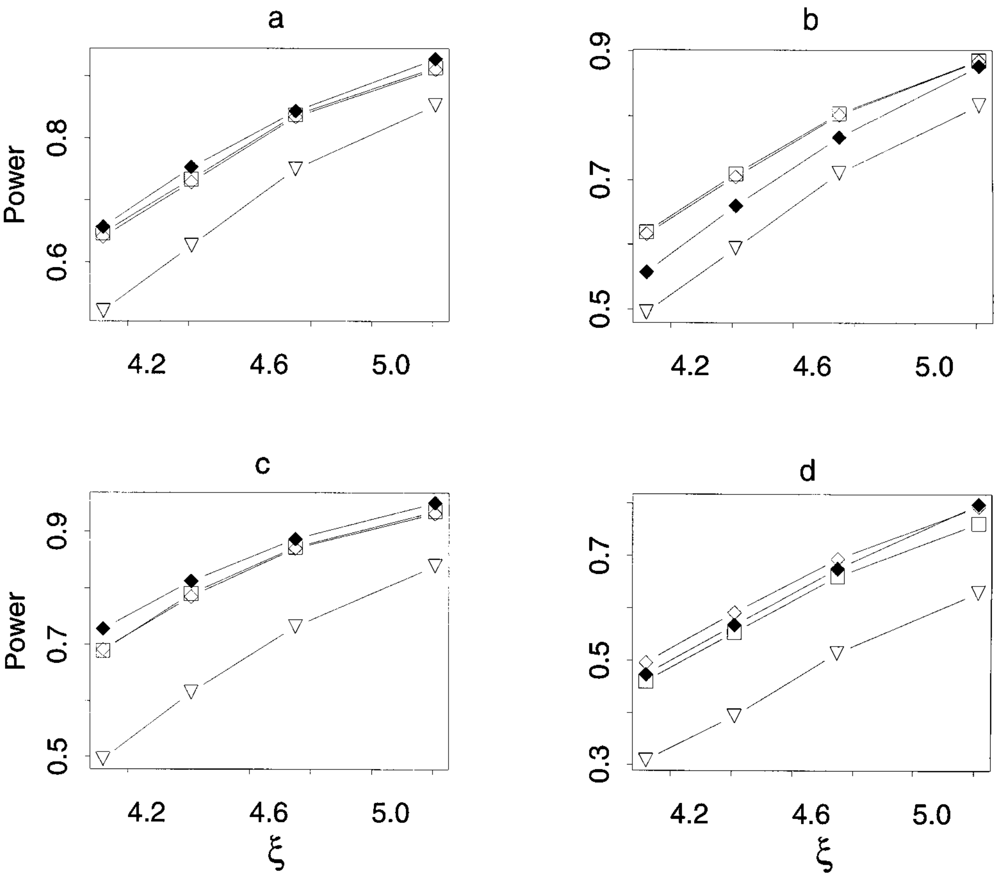

Figure 2.—Power to detect linkage for different map densities, gene locations, and thresholds. In a and b, $\Delta = 5$ cM while $\Delta = 20$ cM in c and d. The trait locus is located at a marker in a and c and mid-markers in b and d. The process without interval mapping is represented by □; the process with interval mapping is represented by ◇ (solid symbols for the theoretical approximation) and ▽ (power for the higher threshold appropriate when $\Delta = 0$).

ping on the false-positive error rate. But insistence on this threshold would noticeably reduce the power of the test, as is shown shortly.

For intercross data the noncentrality parameter for a QTL located at a marker locus is $\xi = \{N \ln[1 + (\alpha^2/2 + \delta^2/4)/\sigma_e^2]\}^{1/2}$. To attribute appropriate parts of the total noncentrality to the two processes in (10), we let $\xi_1 = \xi\alpha/(\alpha^2 + \delta^2/2)^{1/2}$, $\xi_2 = \xi\delta/[2(\alpha^2 + \delta^2/2)]^{1/2}$. If the QTL is located at a marker, the power is approximately

$$P\{\max_k 2 \ln \mathrm{LR}(k\Delta) > a\}$$
$$\approx 1 - \Phi(b - \xi) + \phi(b - \xi)$$
$$\times \left[\frac{1}{2\xi} + \frac{2b^{1/2}\nu}{\xi^{3/2}} - \frac{b^{1/2}\nu^2}{\xi^{1/2}(b + \xi)}\right], \qquad (10)$$

where $\nu = \nu(b\{2\beta\Delta\}^{1/2})$, $\beta = (\beta_1\xi_1^2 + \beta_2\xi_2^2)/\xi^2$. For a QTL between markers, one must as in the backcross case consider the joint distribution at flanking markers. For a marker at distance Δ_1 from the QTL the noncentrality parameters are

$$\xi_1 \exp(-\beta_1\Delta_1) \quad \text{and} \quad \xi_2 \exp(-\beta_2\Delta_1). \qquad (11)$$

See appendix a for an approximation for the power of the interval mapping process.

Using simulations and the theoretical power approximations above, we compare in Figure 2 the power of the marker process with the power of the interval mapping test. We also present the power of the interval mapping test using the more stringent threshold (assuming continuous markers) proposed by Lander and Kruglyak (1995). The power was investigated for a dominant model, so $\delta = \alpha$, and $\xi = 4.12$, 4.41, 4.75, and 5.21, which correspond roughly to powers of 60, 70, 80, and 90% with a continuous map of markers. For recessive ($\delta = -\alpha$) models, the power would be exactly the same. For the same noncentrality values and an additive model ($\delta = 0$), it would be slightly larger. Power under two map densities was estimated ($\Delta = 20$ and 5 cM) and we used $N = 350$ tomato genomes. Each power simulation is based on 1000 replicates. The gain in power from using interval mapping is small, on the order of 2–4%, a result similar to that found by Darvasi *et al.* (1993). The gains anticipated by Lander and Botstein (1986, 1989), who write of interval mapping as providing a "virtual marker" midway between the actual markers, are overly optimistic. Their analysis is marred by their comparison of interval mapping with the marker process at only one of the flanking loci, where a more appropriate comparison would be with the maximum of the process at the two flanking loci. They also neglect the increase in threshold required to maintain a given false-positive error rate for the interval mapping process. The gain in power for interval mapping is largest for the sparse map ($\Delta = 20$ cM), but the gain is only ~3–4%. Using the threshold for a continuous map when in fact a sparse map of markers is used greatly reduces the power (by as much as 20%).

We have made similar computations with similar results for backcross designs.

When the markers only process is used, the theoretical power approximations are very good, so only the simulated values have been included in Figure 2. The approximations are also good for interval mapping except when the intermarker distance is 5 cM and the QTL is midway between markers. In this case the power is underestimated by ~5%. The reason is that the theoretical approximation involves only the probability that the process is above the threshold somewhere in the interval containing the QTL and neglects the probability of detecting the QTL to be in a neighboring interval. This is not a problem when the intermarker interval is large.

Other designs and a comparison of different designs: Many other designs can be handled by similar approximations. To evaluate an appropriate threshold, for the markers only process it is only necessary to know the recombination parameter β (or β_1 and β_2), which depends only on the design, not the mathematical model used for recombination. Although there is no general method to evaluate this parameter, it has been calculated for many different designs. (Some values are given below.) For interval mapping one must know the complete covariance function, which depends on both the design and the model for recombination.

For instance, for recombinant inbred data, which involve the 1-d.f. statistic (3), one can use approximation (4) with $\beta = 0.04$ for recombinants produced by selfing and $\beta = 0.08$ for recombinants produced by recurrent sib mating (as originally suggested by Lander and Botstein 1989). It is only slightly more complicated to incorporate interval mapping. (See Rebai *et al.* 1994 for the case of selfing. A similar formula can be obtained for inbreds produced by recurrent sib mating.) For the advanced intercross designs suggested by Darvasi and Soller (1995) to provide more accurate localization of QTL, for the F_i offspring one can use (9) with $\beta_1 = i\lambda$, $\beta_2 = 2i\lambda$. For reciprocal backcross designs, where half of the offspring are backcrossed to each parental strain, one can use (9) with $\beta_1 = \beta_2 = 0.02$.

In Stuber *et al.* (1992), offspring from a cross of two inbred Maize strains (F_1 generation) were allowed to self twice and then backcrossed to one of the parental lines. A careful examination of that design shows that the maximum LOD for testing the hypothesis of no linkage is approximately [*cf.* (3), (6)]

$$\max_d X^2(d) = \max_d \frac{3N\hat{\alpha}^2(d)}{4\sigma_e^2},$$

where $\hat{\alpha}(d)$ is the maximum-likelihood estimate of the sum of the additive and dominance effects. One can show that under the null hypothesis, $X(d)$ is approximately a Gaussian process with covariance function $R(d) = 1 - \frac{8}{3}\lambda|d| + o(|d|)$ as $d \to 0$. Therefore, approximation (4) can be used with $\beta = \frac{8}{3}\lambda$ to find an appropriate threshold.

Korol *et al.* (1995) have suggested the use of correlated traits as a technique to improve the power of QTL mapping. If the number of traits is t, this would require a t dimensional version of (4) or a $2t$ dimensional version of (9) for the backcross or intercross design, respectively. The appropriate k dimensional approximation ($k = t$ or $2t$) is given by

$$1 - \exp\{-C[1 - F_k(b)] - \beta L 2^{(2-k)/2} \times [\Gamma(k/2)]^{-1} b^k \exp(-b^2/2)\}.$$

Here F_k is the χ^2 distribution with k degrees of freedom, Γ denotes the gamma function, and β would be replaced by $(\beta_1 + \beta_2)/2$ for an intercross design. Corrections for discrete spacing of markers would be exactly as above.

We have used the theory developed above to compare the power of backcross, intercross, and recombinant inbred designs (obtained by recurrent sib mating). Let σ_A^2, σ_D^2, σ_E^2 denote the total additive, dominance, and environmental variances, respectively. Assuming that environmental and genetic effects are uncorrelated and there is no epistasis, we have the usual representation of the phenotypic variance as $\sigma_y^2 = \sigma_A^2 + \sigma_D^2 + \sigma_E^2$. Let $H^2 = (\sigma_A^2 + \sigma_D^2)/\sigma_y^2$ denote the wide sense heritability in the intercross, and put $\rho = \delta/2^{1/2}\alpha$. To reduce the number of different special cases we assume that ρ is the same at all QTL; *i.e.*, they all have the same relative amount of dominance. If we let v^2 be the heritability attributable to the locus of interest, *i.e.*, $v^2 = (\alpha^2/2 + \delta^2/4)/\sigma_y^2 \le H^2$, then the noncentrality parameters of an intercross, backcross, and recombinant inbred design are, respectively, $[-N\ln(1 - v^2)]^{1/2}$, $\{-N\ln[1 - (v^2(1 + 2^{1/2}\rho)^2)/(H^2(1 + 2^{1/2}\rho)^2 + 2(1 - H^2)(1 + \rho^2))]\}^{1/2}$ and $\{-N\ln[1 - 2v^2/(1 + \rho^2 + H^2(1 - \rho^2))]\}^{1/2}$.

Suppose $\rho = 0$. It is easy to see that the noncentrality parameter of the backcross is smallest and that of the recombinant inbred is largest. All three noncentrality parameters are comparable for large H^2, but there can be sizeable differences for small H^2. Because the threshold required for a given significance level is smallest for the backcross and largest for the intercross, one expects to find the backcross the most powerful design when H^2 is large, but not otherwise.

A numerical example is given in Table 1. We have determined for continuous markers sample sizes that give 80% power for values of H^2, v^2, and ρ. Although the exact sample sizes depend on v^2, their relative values are roughly constant throughout a broad range where v^2, the heritability attributable to the QTL, contributes from roughly $\frac{1}{8}$–$\frac{1}{2}$ H^2, so only the intermediate value $v^2 = 0.2\,H^2$ is included in the table. Similarly the relative sample sizes are fairly insensitive to the exact power

TABLE 1

Theoretical sample sizes of intercross, backcross, and recombinant inbred designs necessary to achieve 80% power with dense ($\Delta = 0$) markers

H^2	ν^2	ρ	Intercross	Backcross	Recombinant inbred
0.75	0.15	0.0	139	144	117
		0.2	139	121	118
		−0.2	139	206	118
0.25	0.05	0.0	440	632	264
		0.2	440	430	271
		−0.2	440	1194	271

required. In agreement with the qualitative analysis of the preceding paragraph, for $\rho = 0$ the sample size required by a backcross design is about the same as that of the intercross for $H^2 = 0.75$ but is appreciably larger for $H^2 = 0.25$. For $\rho^2 = 0.04$, the backcross design can require somewhat smaller or much larger sample sizes than the intercross design depending on whether ρ is positive or negative, which in turn depends on the parental strain used for the backcross. Hence with a small amount of dominance, probably too small to be detected in segregation analysis, a backcross design can yield a very misleading picture. The sample sizes required of the recombinant inbred design are smaller than those of the intercross and backcross designs and are insensitive to the values of ρ, at least for the relatively small values considered here.

We have performed similar calculations when the amount of dominance varies across QTL. The sample sizes in the backcross column can change substantially, but the qualitative picture is the same.

This problem with a backcross design could in principle be eliminated by backcrossing to both parental strains and using a 2-d.f. statistic (with $\beta_1 = \beta_2 = 0.02$). One can easily evaluate the noncentrality parameter and see that for small values of H^2 such a reciprocal backcross is less powerful than an intercross design based on an equal number of progeny, but is slightly more powerful than an intercross design based on an equal number of matings (hence presumably half as many progeny). For larger values of H^2, numerical calculations as in Table 1 can help one determine the potential usefulness of such a design.

To simplify the preceding comparison, we have assumed continuously distributed markers. This has the effect of concealing a weakness of the recombinant inbred design, which has a very large recombination parameter ($\beta = 0.08$). A consequence is that if markers are not closely spaced there is a considerable loss of power to detect a QTL located midway between markers. For an example consider the fourth row of Table 1, where the recombinant inbred design is much more powerful than either of the other two. For a $\Delta = 20$-cM map and a QTL midway between markers, the power falls to about 0.73 if we use the sample sizes given in the table with an intercross or backcross design. To achieve this power with a recombinant inbred design, one would need a sample size of ~380, and in this case interval mapping would be mandatory. Otherwise a sample size of ~690 would be required. For a $\Delta = 5$-cM map, the power of a backcross or intercross would fall only to 0.79 for a QTL midway between markers. Now for a recombinant inbred design a sample size of about 291 would be required (300 without interval mapping). To achieve the benefits of a recombinant inbred design, it appears advisable to type markers at no more than 5 cM distance, and closer would be better. A similar caution is applicable to the advanced intercross designs of Darvasi and Soller (1995).

Confidence regions for QTL: A confidence region can be used to identify a chromosomal region in which to concentrate the search for the exact location of a QTL. In this section, three methods of constructing a confidence region around the gene locus are presented and compared. It is perhaps worth noting from the outset that this is not a "regular" estimation problem as the term is used by statisticians. Because the likelihood function has cusps at marker loci, the maximum-likelihood estimate of a QTL may fail to be approximately normally distributed, so one is not justified in using the maximim-likelihood estimator plus or minus two estimated standard errors as an approximate 95% confidence interval. Darvasi *et al.* (1993) in one of their suggestions appear to have assumed incorrectly that the standard statistical theory is applicable. Visscher *et al.* (1996) have suggested a confidence interval based on the unconditional distribution of the maximum-likelihood estimator, which they estimate by bootstrapping. Although their coverage probabilities are shown by a Monte Carlo experiment to be quite close to the specified level, this method does not adapt to the rate of decay of the likelihood function near its maximum and is known to give confidence regions that are unnecessarily large in related "change-point" problems. A numerical example given below suggests that it has the same undesirable feature here. See Siegmund (1988) for a more complete discussion.

Support intervals: Support intervals (*cf.* Conneally *et al.* 1985) provide a method of estimating the location of a trait locus. They are essentially equivalent to the standard statistical technique of inverting the likelihood ratio test to obtain a confidence region. Given a value $x > 0$, a support region includes all the loci q such that

$$2 \ln \mathrm{LR}(q) \geq \max_d 2 \ln \mathrm{LR}(d) - x. \tag{12}$$

Often the 2 is omitted and common logarithms are used. Then one speaks of a LOD support region. The value x in (12) provides an $(x/2 \ln 10)$-LOD support region. With data from a single marker the statistical problem is regular, so a 1-LOD support interval ($x = 4.6$) is approximately a 97% confidence interval (because 4.6 is the 97th percentile of the χ^2 distribution with 1 d.f.; see Ott (1991, p. 67). However, this result does not generalize to genome-wide scans involving reasonably dense markers, where the coverage probability of (12) depends on the density of the map of markers and on the strength of the signal at the trait locus. In fact, there is no exact confidence coefficient that can be assigned to a support region. Through theoretical analysis and a simulation study presented below, we show that a 1-LOD ($x = 4.6$), respectively 1.5-LOD ($x = 6.9$), support interval corresponds roughly to a 90%, respectively 95%, confidence region in the case of a dense map of markers (~1 cM), and provides even greater probability of coverage for sparser maps.

Likelihood methods: A second method to provide a confidence interval for a QTL relies on using likelihood methods for change points (Siegmund 1988; Feingold *et al.* 1993). It is closely related to the support method described above and provides some analytic tools for studying that concept. Unlike the support method, however, for the special case that the trait locus is exactly at a marker location the likelihood method in principle gives an exact confidence region.

Although the actual procedure is based on twice the log-likelihood ratio, our discussion will be simplified notationally by using the asymptotically equivalent $\|Z_d\|^2$, where $Z_d = (X_d, Y_d)$ is defined in (8) [*cf.* also (7)] and $\|Z_d\|^2 = X_d^2 + Y_d^2$. In terms of these variables the acceptance region for the likelihood ratio test of the hypothesis that a QTL is located at q has the form

$$A_q = \{\max_d \|Z_d\|^2 - \|Z_q\|^2 \leq x\}.$$

By sufficiency, the conditional probability of A_q given Z_q does not depend on the unknown parameters α, δ. Hence in principle we can choose $x = x(Z_q)$ such that

$$P(A_q | Z_q) = 1 - \gamma. \tag{13}$$

The set of all values q that are not rejected by this test is a $(1 - \gamma)100\%$ confidence region (Cox and Hinkley 1974).

As the desired conditional probability does not depend on α, δ, it can be evaluated under the hypothesis that these parameters are both zero. The approximation (B1) of appendix b yields as a confidence interval for the QTL those loci q such that

$$P(\max_d \|Z_d\|^2 > (\max_d \|Z_d\|^2)_{\mathrm{obs}} | Z_q) \geq \gamma. \tag{14}$$

The likelihood method works best for very dense sets of markers (~1 cM), as the argument given above is technically correct only when the QTL is at a marker. It can be extended to provide a joint confidence region for the locus and the additive and dominance effects (Dupuis 1994).

By (7) the inequality defining A_q and the inequality in (12) are asymptotically equivalent. The important difference between the likelihood ratio and LOD support methods is that for the former x depends on Z_q and is chosen to make the conditional probability (13) equal to the desired confidence level. For any value x that does not depend on the data, the probability of (12) depends on the values of α and δ. Hence the support region is not a confidence region in the strict sense of the word. However, the similarity between the support regions and the likelihood ratio regions allows us to gain some interesting theoretical insights. For example, under the assumption that the QTL lies at a marker locus and that the distance Δ between markers is small, we can evaluate approximately the probability that a support region does not contain the true QTL, by taking the expectation of (B1) in appendix b with respect to $Z_q = z$. The result of some simple approximations is

$$P(A_q) \approx 1 - 2\nu\{[2\tilde{\beta}\Delta(\xi^2 + x)]^{1/2}\} \times \left[\frac{\xi^2 + x}{\xi^2 + x\xi_2^2/(\xi_1^2 + 2\xi_2^2)}\right]^{3/2} \exp(-x/2), \tag{15}$$

where $\xi = (\xi_1^2 + \xi_2^2)^{1/2}$, $\tilde{\beta} = (\beta\xi_1^2 + 2\beta\xi_2^2)/\xi^2$, $\beta = 0.02$. Numerical calculations based on this approximation suggest, and simulations reported below verify, that for a given value of Δ the coverage probability of the support region is relatively insensitive to the values of ξ and to the relative sizes of the additive and dominance components, at least for values of ξ in the range $4 \leq \xi \leq 10$, where detection of linkage ranges from reasonably likely to virtually certain, so QTL localization is especially important. The coverage probability is an increasing function of the intermarker distance Δ, so a 1.5-LOD support region has ≈95% coverage when $\Delta \approx 1$ cM, while a 1-LOD support region gives similar coverage for $\Delta \approx 20$ cM. Hence for practical purposes a support region is approximately a confidence region, albeit with a different confidence coefficient than that suggested by standard statistical distribution theory.

For problems involving a single parameter, *e.g.*, for backcrosses, recombinant inbreds, or intercrosses where we estimate only α and ignore δ, the factor in square brackets in (15) immediately preceding the exponential would be $[(\xi^2 + x)/\xi^2]^{1/2}$. It is easy to see that at least for comparatively large values of ξ, the coverage probability for a given value of x is relatively insensitive to this change of dimension.

An approximation for the expected size of a support region, which is valid for dense markers ($\sim$1 cM), is given in appendix b. A less precise but more easily interpreted approximation, valid when $\xi \gg x$, is obtained by approximating the normal density in (B2) with mean ξ by a point mass at ξ, then taking two terms of the Taylor series expansion of $\ln[\xi^2/(\xi^2 - x)]$, which yields

$$\beta^{-1}[x/\xi^2 + 0.5x^2/\xi^4 + 2\xi^{-2}(1 - 2\nu(\xi(2\beta\Delta)^{1/2}) + 0.5\nu^2(\xi(2\beta\Delta)^{1/2}))]. \quad (16)$$

This expression is roughly proportional to ξ^{-2}, hence to N^{-1}. In contrast, for regular statistical problems the size of a confidence region is inversely proportional to the square root of the sample size. The fact that confidence regions for a QTL are roughly inversely proportional to the sample size has been observed in the simulations of Darvasi *et al.* (1993) and Visscher *et al.* (1996), although these authors do not provide a theoretical explanation. The approximation (16) also shows, as one might have anticipated, that the average length of a support region is inversely proportional to β, hence to the recombination rate for the design used. Even if we ignore the difference between noncentrality parameters for recombinant inbred and backcross designs, the recombinant inbred design, for which $\beta = 0.08$, will give regions roughly one-fourth the size of those obtained from a backcross, provided the intermarker distances are sufficiently small. In fact, for additive traits recombinant inbreds always have a larger noncentrality parameter than a backcross, so they provide support regions even less than one-fourth as large. In the extreme case of small heritability and a QTL that is responsible for most of the additive variance, the relative size can shrink by another factor of almost 4.

Bayesian credible regions: Given a prior probability for the location of the QTL and for the noncentrality parameters (ξ_1, ξ_2), a set having a posterior probability of $1 - \gamma$ is called a Bayesian credible region. Fisher (1934), in his classical study of ancillarity, showed in effect that under certain conditions Bayesian credible sets are in fact $1 - \gamma$ confidence regions having many desirable properties. Cobb (1978) pointed out that a special class of statistical problems having the required structure are “change-point” problems, which have been studied extensively from this point of view by Zhang (1991). Feingold *et al.* (1993) and Kruglyak and Lander (1995) have noted the similarity between estimating the location of a change-point and estimating the location of a trait locus from data on mapped markers. A consequence of this history is the expectation that a Bayesian credible region for a uniform prior distribution on the location of the QTL will provide satisfactory confidence regions.

A Bayesian credible region B_γ is constructed by including all loci v whose posterior density given the data exceeds c_γ, *i.e.*,

$$B_\gamma = \{v : \pi(v|\mathbf{y},\mathbf{x}) > c_\gamma\}, \quad (17)$$

where c_γ is chosen so that

$$\int_{B_\gamma} \pi(v|\mathbf{y},\mathbf{x})dv = 1 - \gamma.$$

Here $\mathbf{y} = \{y_1, \ldots, y_N\}$, $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ and $\mathbf{x}_i$ is the set of all marker genotypes for individual i. The posterior probability π $(v|\mathbf{y},\mathbf{x})$ is often easy to compute and depends on the prior distribution on the location q and the additive and dominance effects α and δ. If one takes uninformative priors on all parameters,

$$\pi(v \mid \mathbf{y},\mathbf{x}) \cong \frac{\exp(-\frac{1}{4}\|Z_v\|^2)}{\int_0^l \exp(-\frac{1}{4}\|Z_s\|^2)ds}, \quad (18)$$

where $Z_t = (X_t, Y_t)$ was defined previously and can be obtained using least-squares estimates or the interval mapping equivalent. Analogous expressions can be obtained for other priors. We have studied properties of three different priors on the additive and dominance effects, with a uniform prior for the gene location. First a flat prior was implemented. Second, we constructed the confidence sets with uncorrelated normal priors with mean 0 and standard deviation of 4. The mean of 0 is to allow the parameters to be positive or negative and a standard deviation of 4 should be large enough to allow the parameters to vary freely. Finally, since the smallest detectable genetic effect involves a noncentrality of $\sim$4, a uniform mixture of four uncorrelated normal priors with noncentralities of 4 corresponding to dominant ($\delta = \alpha$) and recessive ($\delta = -\alpha$) models and with variance of one was also applied. Results are presented in the next section.

Comparison study: Using simulated tomato genomes, we constructed the likelihood confidence region, the 1.0- and 1.5-LOD support region and the Bayes credible regions, with the three different priors mentioned above. However, only the results from Bayes credible sets with a mixture of normal priors are included in Tables 2 and 3. For each tomato, the crossover process for the chromosome containing the QTL was generated using the Haldane mapping function and the phenotype y_i was assigned the value

$$y_i = \alpha x_i(q) + \delta 1_{(x_i(q)=1)} + e_i,$$

where the e_i’s are normal random variables with mean 0 and variance 1.

We performed the simulations for the dominance model ($\delta = \alpha$), with $\xi = 5$, 7.5, and 10.0. The trait locus was either at a marker, midway between markers, or randomly assigned. We generated 1000 sets of 350 tomatoes and calculated the average size and the probability of covering the true locus given a map with $\Delta = 1$, 5, and 10 cM. Interval mapping was used throughout.

TABLE 2

Average size in centimorgans of simulated confidence intervals

		$\Delta = 1$			$\Delta = 5$			$\Delta = 10$		
ξ	Method	0	r	½	0	r	½	0	r	½
5.0	Likelihood	17.1	16.7	17.0	16.2	16.1	18.7	15.0	18.0	21.3
	1.0-LOD	8.8	8.5	8.8	12.5	12.5	14.5	15.0	17.1	19.7
	Bayes	14.1	13.0	14.0	15.3	14.8	16.9	16.3	18.0	20.9
7.5	Likelihood	4.6	4.7	4.9	5.1	5.7	6.3	5.3	6.4	7.4
	1.0-LOD	3.8	3.9	4.0	6.1	6.7	7.5	8.3	9.5	10.9
	Bayes	6.2	6.2	6.4	7.4	8.0	9.1	9.1	10.5	12.3
10.0	Likelihood	2.4	2.4	2.5	2.7	3.1	3.3	2.2	2.9	3.1
	1.5-LOD	3.1	3.2	3.3	5.2	5.8	6.5	7.3	8.4	9.0
	Bayes	3.6	3.8	4.0	5.0	5.7	6.5	6.8	7.9	8.6

Three locations for the QTL were simulated: 0 for the trait at a marker, ½ for the trait midmarkers, and r for the QTL randomly located between markers.

Both the 1.5-LOD ($x = 6.9$) support regions and the Bayesian credible regions provided at least 95% coverage under all simulated conditions. The support regions gave the smallest confidence regions for dense maps, while the Bayesian credible regions did the same for sparse maps. The coverage probability for the support regions obtained in the simulations is close to that predicted by the approximation (15). The approximate expected size provided by (B2) is close in the case of a dense map, but not otherwise. The likelihood method was conservative; and because it adapts to the observed value of the likelihood ratio statistic at the putative trait locus it resulted in the widest confidence regions for small values of the noncentrality parameter but was equivalent to the support region for the larger values $\xi = 7.5$ and 10. For all methods, the sizes of the intervals were largest when the trait was midmarker. The Bayes credible sets were the widest and they fell short of the desired 95% for large values of ξ and sparse maps, especially when the trait was located at a marker.

The size of the confidence regions is relatively insensitive to the marker density when the distance between markers and the size of the region are roughly commensurate; but when ξ is large, the dense marker map provides substantially smaller regions.

We performed similar simulations for a backcross with essentially the same results (data not shown). The simulations were repeated with fewer tomatoes ($N = 100$) (results not shown). The size of the region was unchanged for all methods, and all methods had the right coverage probability when the locus was located at a marker. The coverage probability was substantially reduced for the case of the likelihood method and the Bayes method when the trait was located midmarkers (≈80% instead of 95%). The LOD support method had a slight drop in confidence coverage (≈90%), but was more robust than the other methods.

We have also simulated support regions under the conditions of Table 2 of Visscher *et al.* (1996), which involved a backcross with no dominance variance and marker spacings of 20 cM. At this intermarker distance 1-LOD ($x = 4.6$) regions had coverage probabilities ranging from 93 to 96% and in all cases gave smaller regions than the 95% bootstrap regions recommended

TABLE 3

Coverage probability of simulated confidence intervals

		$\Delta = 1$			$\Delta = 5$			$\Delta = 10$		
ξ	Method	0	r	½	0	r	½	0	r	½
5.0	Likelihood	94.9	94.6	94.5	92.6	94.2	91.7	91.7	88.7	87.3
	1.5-LOD	95.6	96.1	95.8	96.7	97.5	96.4	98.5	98.0	97.1
	Bayes	96.7	94.5	97.6	95.2	96.2	96.5	97.6	94.7	93.1
7.5	Likelihood	93.8	94.0	93.9	91.6	89.3	85.8	81.9	79.8	78.5
	1.5-LOD	96.2	97.6	97.4	98.5	97.9	96.8	99.3	98.2	97.5
	Bayes	97.4	96.5	99.1	98.0	97.0	98.1	99.3	97.0	93.7
10.0	Likelihood	94.9	93.1	90.2	84.4	78.6	77.8	63.7	55.3	62.0
	1.5-LOD	97.6	97.1	97.2	98.8	97.8	96.4	99.0	97.8	96.7
	Bayes	99.0	96.4	99.8	99.4	96.7	98.3	98.9	96.0	94.8

by Visscher *et al.* (1996), while 1.5-LOD regions had 98–99% coverage probability and about the same expected sizes as the bootstrap regions. For example, for a heritability of 0.05 and a sample size of 500, which yield a noncentrality parameter $\xi = 5.06$, the coverage probability of the 1-LOD region based on 1000 simulations was 96%, and the expected size was 29 cM compared with 96% and 43 cM obtained by Visscher *et al.* (1996) for their bootstrap regions.

Another method to obtain confidence intervals for QTL location has been proposed by Mangin *et al.* (1994). This method amounts to fixing a putative QTL location and testing the hypothesis that there is no QTL between that location and either end of the chromosome. In the statistical literature on change-point analysis Worsley (1986) has discussed a similar idea and has pointed out that if there is another change-point (here QTL on the same chromosome) the method may produce an empty confidence set, since for every putative QTL there is evidence of another somewhere on the chromosome. Of course, the problem of detecting a second, linked QTL given an already detected QTL is itself interesting and important.

DISCUSSION

In this article we have discussed genome scanning methods to detect QTL in experimental genetics. Our goal has been to produce relatively simple approximations for quantities of interest, *e.g.*, the false-positive error rate, power to detect a QTL, and coverage probability of a support region, so that one can easily address questions concerning sample size, marker density, etc., and can compare different designs. Our approximations for significance level and power seem adequate in this regard, but our approximations for the expected size of a support region are good only for dense markers (*e.g.*, $\Delta \approx 1$ cM).

Although in a backcross the conventional LOD = 3 threshold produces false-positive rates <0.05 unless intermarker distances are small, it is anticonservative in an intercross even for intermarker distances as large as 25 cM without interval mapping.

Our approximations are based on the artificial assumption that markers are equally spaced and there are no missing data. If markers are not equally spaced, the approximations (4) and (9) can be modified by averaging the function ν with respect to the distribution of the distances Δ between markers. One can also use the original approximations with an average intermarker distance. (This should be the average distance in the neighborhood of detected QTL if one adds additional markers to promising regions.) Since (4) and (9) are insensitive to minor changes in the assumed value of Δ, one can reasonably expect such refinements to have little practical effect. If we use interval mapping to impute missing marker data, the resulting process is more correlated than would be the case if the data were not missing, so the threshold obtained under the assumption of no missing data is still appropriate and, in fact, slightly conservative.

The assumption of normality is robust in the sense that the regression statistics we use are approximately normally distributed in large samples, so our approximations for significance level and power are valid in large samples. However, it is possible that by using a more appropriate model, *e.g.*, a mixture model if the nonnormality arises from large QTL effects, one can obtain greater power, although large QTL effects will be comparatively easy to detect with a suboptimal procedure.

When using a backcross or intercross, intermarker distances up to ~10 cM are almost as powerful as continuously distributed markers. Except at intermarker distances of ~20 cM or more, or when using a design involving a large recombination rate, *e.g.*, a recombinant inbred design or advanced intercross design, there is little gain in power from interval mapping, which in any event does not provide nearly as much power as more closely spaced markers.

Although intercross designs involve a 2-d.f. statistic and hence a higher threshold than a backcross design, and have larger residual variance, intercross designs are usually more powerful than backcross designs, unless (a) the effect of the gene is large and additive or (b) there is dominance *and* the dominance deviation has the same sign as the additive genetic effect. A backcross design can lose considerable power in the presence of even a small departure from additivity if the incorrect parental strain is used for the backcross. A recombinant inbred design can be more efficient than an intercross, except when dominance effects are large compared to additive effects. Because of the high recombination rate associated with recombinant inbreds, especially those based on recurrent sib mating, power to detect linkage falls off rapidly with intermarker distance when a QTL is located midway between markers. To avoid this loss of power when using an inbred design based on recurrent sib mating, intermarker distances should be no more than 5 cM and preferably should be even less. Similar considerations apply to advanced intercross lines (Darvasi and Soller 1995).

We have also presented three methods of constructing confidence regions for the location of QTL: the likelihood method, Bayes credible sets, and support regions. The support method and the Bayesian credible sets seem roughly comparable in large samples, but the coverage probability of the support method is more robust to changes in the sample size. Both methods are better than the likelihood ratio method, which often has a coverage probability substantially smaller than the nominal level, except for the case of dense markers.

The size of a confidence region depends on the noncentrality parameter and the density of the markers in the neighborhood of the QTL. When the noncentrality parameter is ∼5, which provides power of ∼0.9 for QTL detection, little is gained by having markers more closely spaced than ∼10 cM; but when the noncentrality parameter is 7.5, intermarker distances of 1–5 cM provide shorter confidence regions. A reasonable guideline is to achieve a marker density in the neighborhood of a putative QTL about equal to the expected half length of a support region for a QTL of that strength.

When dominance effects are relatively small and markers sufficiently dense, support regions from recombinant inbred designs are often about one-fourth as large as from intercross designs, which in turn are substantially smaller than from backcross designs. Advanced intercross designs (Darvasi and Soller 1995) are also especially powerful for fine localization of QTL. In almost all cases, however, the size of the confidence regions is on the order of several centimorgans unless the sample size is considerably larger than what is required to detect linkage, so there is a continuing need to develop better designs for fine localization of QTL.

We have not explicitly addressed the complexities associated with identifying multiple, possibly linked, possibly interacting, QTL. For mapping qualitative traits in humans, we have discussed these issues (Dupuis *et al.* 1995), and expect to return to them for QTL mapping. For example, once a linked QTL is located, conditional search removes the effect of that QTL by subtracting its (estimated) genotypic contribution from the phenotypic value to define a new regression model, hence a new log-likelihood ratio statistic, to search for additional QTL. Suppose an intercross design is used and, for simplicity, we use a 1-d.f. statistic to detect a QTL of purely additive effect. Assume also that we know exactly the location of a QTL making contribution v^2 to the heritability. The (asymptotic) correlation function between the new and old processes at each unlinked marker is $(1 - v^2)^{1/2}$, and under the assumption of no epistasis, the noncentrality parameter for the new statistic is larger by the factor $1/(1 - v^2)^{1/2}$. Hence a large QTL effect v^2 is necessary at the detected locus to get a reasonable "gain" from the conditional search, although a large value of v^2 also leads to a new process only weakly correlated with the original search process, which increases the likelihood that conditional search will incur a false-positive error. Of course, there must be another QTL of sufficiently large effect for the gain in noncentrality to be helpful. Rough calculations suggest that suitable combinations of QTL effects will occur relatively rarely.

Similar considerations are relevant to recently suggested multiple regression methods, *e.g.*, Zeng (1994) and Jansen (1994), whereby one searches, for example, a given chromosome or chromosomal arm for a QTL while controlling for QTL on other chromosomes through arbitrarily placed markers. In comparison with conditional search, this method has the potential advantage of controlling the phenotypic variability due to multiple QTL, but at least initially has the disadvantage that the success of the control depends on fortuitously placing the control markers close to true QTL. Straightforward calculations show that the control markers on other chromosomes have no effect on the asymptotic distribution of the log-likelihood ratio process along the currently searched (unlinked) chromosome, although they do reduce the number of degrees of freedom available to estimate the error variance. By considering one chromosome at a time and adding the chromosome-wide false-positive rates, one obtains an asymptotic upper bound on the genome-wide false-positive rate. Because of the independent assortment of chromosomes, this upper bound should not be overly conservative.

The second method discussed by Dupuis *et al.* (1995), simultaneous search, will for the reasons given there rarely be useful in the absence of epistasis. Preliminary calculations suggest it can be very helpful when there is substantial epistasis.

We expect to return to the problem of detecting multiple, possibly linked, QTL in a future article.

This research was partially supported by the National Institutes of Health grant HG-00848 and the National Science Foundation grant DMS 9704324.

LITERATURE CITED

Andersson, L., C. S. Haley, H. Ellegren, S. A. Knott, M. Johansson *et al.*, 1994 Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science **263:** 1771–1774.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics **138:** 963–971.

Cobb, G. W., 1978 The problem of the Nile: conditional solution to a change-point problem. Biometrika **62:** 243–251.

Conneally, P. M., J. H. Edwards, K. K. Kidd, J.-M. Lalouel, N. E. Morton *et al.*, 1985 Report of the Committee on Methods of Linkage Analysis and Reporting. Cytogenet. Cell Genet. **40:** 356–359.

Cox, D. R., and D. V. Hinkley, 1974 *Theoretical Statistics.* Chapman and Hall, London.

Darvasi, A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics **141:** 1199–1207.

Darvasi, A., A. Weinreb, V. Minke, J. I. Weller and M. Soller, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics **134:** 943–951.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Statist. Soc. **B39:** 1–22.

Dupuis, J., 1994 Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data. Ph.D. Thesis, Stanford University, Stanford, CA.

Dupuis, J., P. O. Brown and D. Siegmund, 1995 Statistical methods for linkage analysis of complex traits from high resolution maps of identity by descent. Genetics **140:** 843–856.

Feingold, E., P. O. Brown and D. Siegmund, 1993 Gaussian models for genetic linkage analysis using complete high resolution maps of identity-by-descent. Am. J. Hum. Genet. **53:** 234–251.

Fisher, R. A., 1934 Two new properties of mathematical likelihood. Proc. R. Soc. A **144:** 285–307.

Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.

Jacob, H. J., K. Lindpaintner, S. E. Lincoln, K. Kusumi, R. K. Bunker *et al.*, 1991 Genetic mapping of a gene causing hypertension in the stroke-prone spontaneously hypertensive rat. Cell **67:** 213–224.

Jansen, R. C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics **138:** 871–881.

Johnstone, I. M., and D. Siegmund, 1989 On Hotelling's formula for the volume of tubes and Naiman's inequality. Ann. Statist. **17:** 184–194.

Kempthorne, O., 1957 *An Introduction to Genetic Statistics*. John Wiley and Sons, New York.

Korol, A. B., Y. I. Ronin and V. M. Kirzhner, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes. Genetics **140:** 1137–1147.

Kruglyak, L., and E. S. Lander, 1995 High-resolution mapping of complex traits. Am. J. Hum. Genet. **56:** 1212–1223.

Lander, E. S., and D. Botstein, 1986 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 D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lander, E. S., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.

Mangin, B., B. Goffinet and A. Rebai, 1994 Constructing confidence intervals for QTL location. Genetics **138:** 1301–1308.

Ott, J., 1991 *Analysis of Human Genetic Linkage*, Revised Edition. Johns Hopkins University Press, Baltimore.

Paterson, A. H., S. Damon, J. D. Hewitt, D. Zamir, H. D. Rabinowitch *et al.*, 1991 Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. Genetics **127:** 181–197.

Rebai, A., B. Goffinet and B. Mangin, 1994 Approximate thresholds of interval mapping test for QTL detection. Genetics **138:** 235–240.

Rebai, A., B. Goffinet and B. Mangin, 1995 Comparing power of different methods for QTL detection. Biometrics **51:** 87–99.

Siegmund, D., 1985 *Sequential Analysis: Tests and Confidence Intervals*. Springer-Verlag, New York.

Siegmund, D., 1988 Confidence sets in change-point problems. Int. Statist. Rev. **56:** 31–48.

Siegmund, D., and K. Worsley, 1995 Testing for a signal with unknown location and scale in a stationary Gaussian random field. Ann. Statist. **23:** 608–639.

Stuber, C. W., S. E. Lincoln, D. W. Wolff, T. Helentjaris and E. S. Lander, 1992 Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular markers. Genetics **132:** 823–839.

Visscher, P. M., R. Thompson and C. S. Haley, 1996 Confidence intervals in QTL mapping by bootstrapping. Genetics **143:** 1013–1020.

Woodroofe, M., 1976 Frequentist properties of bayesian sequential tests. Biometrika **63:** 101–110.

Worsley, K. J., 1986 Confidence regions and tests for a change-point in a sequence of exponential family random variables. Biometrika **73:** 91–104.

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

Zhang, H. P., 1991 A study of change-point problems. Ph.D. Thesis, Stanford University, Stanford, CA.

Communicating editor: S. Tavaré

APPENDIX A

Power of interval mapping: We first consider a backcross and suppose there is a single trait locus (on any particular chromosome) at q. Let Z_d denote the signed square root of twice the log-likelihood ratio (incorporating interval mapping), which for large N behaves like a piecewise smooth Gaussian process. We use the basic decomposition

$$P[\max_d Z_d \geq b] = P[Z_q \geq b] + P[Z_q < b, \max_{d\neq q} Z_d \geq b].$$

The first term on the right-hand side is given by $1 - \Phi(b - \xi_q)$, where $\xi_q = E(Z_q)$. To approximate the second term, we assume that if the process exceeds the threshold for some $d \neq q$ it does so at a value of d between the same two flanking markers as q, or in one of the immediately adjoining marker intervals in the case that q is itself a marker. This analysis can be expected to yield reasonable approximations in the case that intermarker intervals are large, when interval mapping is supposed to be most helpful. It may not be effective when the intermarker distances are small, expecially if the noncentrality is also small. We approximate $\max_d Z_d$ by expanding Z_d in two terms of a Taylor series around $d = q$ and using calculus to maximize the resulting expression. See Siegmund and Worsley (1995) for details of this calculation. The final approximation is

$$P[\max_d Z_d \geq b] \approx 1 - \Phi(b - \xi_q) + I_q(\xi_q - b)^{-1}\varphi(b - \xi_q)[1 - (b/\xi_q)^{1/2}], \tag{A1}$$

where I_q equals 2 or 1 according to the trait locus being at a marker or in the interval between markers. This discontinuous behavior at the markers is caused by the discontinuity in the derivative of the interval mapping statistic that occurs at the markers.

The noncentrality parameter ξ_q can be evaluated by a direct computation starting from a suitable explicit representation of the interval mapping statistic. See Rebai *et al.* (1995) for such a representation in a complete interference model; their equation is easily modified for the Haldane model of no interference. We present here an alternative method, which will be easier to apply to intercross designs, where the explicit statistic is much clumsier to manipulate. We begin with the following asymptotically equivalent expression for the square root of (3):

$$\frac{\Sigma[x_i(d) - \tfrac{1}{2}](y_i - \bar{y})}{\sigma_e\{\Sigma[x_i(d) - \tfrac{1}{2}]^2\}^{1/2}}, \tag{A2}$$

where $\bar{y} = N^{-1}\Sigma y_i$. In the case where the locus d lies between flanking markers, we replace the actual marker data, $x_i(d)$, by its conditional expectation given the genotypes of the flanking markers, $E[x_i(d) \mid G_i]$. Taking expectations and using (2), we see from some simple manipulations that the noncentrality is asymptotically equal to

$$[(\alpha + \delta)/\sigma_e]\{\Sigma_i E[E(x_i(q) \mid G_i) - \tfrac{1}{2}]^2\}^{1/2}.$$

To express this explicitly in terms of recombination fractions, let θ_1 (θ_2) denote the recombination fraction between the QTL at q and the marker flanking on the left (right), and θ the recombination fraction between the two flanking markers. Then straightforward calculations yield

$$\xi_q^2 = \xi^2\{(1 - \theta_1 - \theta_2)^2/(1 - \theta) + (\theta_1 - \theta_2)^2/\theta\},$$

where $\xi^2 = N \ln\{1 + [(\alpha + \delta)/2\sigma_e]^2\}$. This reduces to the noncentrality ξ when $\theta_1 = 0$, so $\theta_2 = \theta$. At the midpoint between markers, if we assume the Haldane model of no interference it simplifies to

$$2\xi^2\{\exp(-\beta\Delta)/[1 + \exp(-\beta\Delta)]\}.$$

This always exceeds the parameter (6), although a direct comparison is not really meaningful because the markers only statistic involves the maximum of the process at the two flanking markers.

We can also give as an approximation for the power of the interval mapping process

$$P\Big[\max_d Z_d \geq b\Big] \approx 1 - \Phi(b - \xi_q) + \left[\frac{1}{2\xi_q} + I_q\frac{(b/\xi_q)^{1/2}\{1 - (b/\xi_q)^{1/2}\}}{\xi_q - b}\right] \times \varphi(b - \xi_q). \tag{A3}$$

A more detailed calculation along the lines of that given for a backcross yields an expression for ξ_q, which in general is somewhat complicated. In the special case that q is the midpoint between two markers at distance Δ, the parameter ξ_q is the norm of the vector with coordinates

$$\xi_1\left\{\frac{2\exp(-\beta_1\Delta)}{[1 + \exp(-\beta_1\Delta)]}\right\}^{1/2},$$

$$\xi_2 \exp(-\beta_1\Delta)\left\{\frac{1}{[1 + \exp(-\beta_2\Delta)]} + \frac{2}{[1 + \exp(-\beta_1\Delta)]^2}\right\}^{1/2},$$

where ξ_1, ξ_2, β_1, and β_2 are as defined in the paper.

APPENDIX B

Approximations for the conditional probability of (14) and the expected size of a LOD support region: To approximate the conditional probability of (14), we begin with the following lemma.

Lemma. *Let $Z_t = (Z_{1,t}, Z_{2,t})$ where $Z_{1,t}$ and $Z_{2,t}$ are independent Gaussian processes with covariance functions satisfying*

$$R_i(t) = 1 - \beta_i|t| + o(|t|) \quad as \quad t \to 0.$$

Assume $b \to \infty$, $\Delta \to 0$, and $b\Delta^{1/2}$ is bounded away from 0 and ∞. Let $0 < \|z\|^2 < b^2$ and define t^*, w^* to be the solution of

$$\begin{pmatrix} z_1 \\ z_2 \end{pmatrix} = \begin{pmatrix} b\ R_1(t^*)\cos w^* \\ b\ R_2(t^*)\sin w^* \end{pmatrix}.$$

Assume t^* is contained in $(0,t_1)$ and is bounded away from the upper endpoint $(t_1 > 0)$. Then

$$P\{\max_{0\leq i\Delta\leq l}\|Z_{i\Delta}\| \geq b \mid Z_0 = z\} \sim \frac{\beta\exp[-\tfrac{1}{2}(b^2 - \|z\|^2)]}{|\dot{R}_1(t^*)R_2(t^*)\cos^2 w^* + R_1(t^*)\dot{R}_2(t^*)\sin^2 w^*|} \times \nu[b(2\beta\Delta)^{1/2}],$$

where $\dot{R}_i(t) = dR_i(t)/dt$ and $\beta = \beta_1\cos^2(w^*) + \beta_2\sin^2(w^*)$.

For our particular application, $R_i(t) = \exp(-\beta_i|t|)$. Putting $b^2 = \|z\|^2 + x$ and assuming $|x^{1/2}z_2| \ll |z_1|$, which will be the case with probability close to one unless there is overdominance, we obtain

$$P\{\max_{0\leq i\Delta\leq l}\|Z_{i\Delta}\|^2 > \|z\|^2 + x \mid Z_0 = z\} \approx \frac{[2(\|z\|^2 + x)]^{3/2}\exp(-x/2)}{(z_1^2 + [(z_1^2 + 2z_2^2)^2 + 4z_2^2x]^{1/2})^{3/2}} \times \nu([2\Delta(\beta_1z_1^2 + \beta_2z_2^2)(1 + x/\|z\|^2)]^{1/2}). \tag{B1}$$

A proof of the lemma is given in Dupuis (1994). The false-positive error rate in (9) can be obtained by integration with respect to the distribution of $\|Z_0\|$, although it is easier to give a direct calculation along the same lines as the proof of the lemma.

We can also obtain a rough approximation for the expected size of the support region as follows. First consider the one-dimensional case of a backcross or recombinant inbreds and assume as before that a marker is at the QTL q. Then the expected size of the support region is

$$\Delta\Sigma_k\, P\{Z_{k\Delta}^2 \geq \max Z_{j\Delta}^2 - x\} = \Delta\Sigma_k\int \varphi(z - \xi) \times P^z\,\{Z_{k\Delta}^2 \geq \max Z_{j\Delta}^2 - x\}dz,$$

where P^z denotes probability under the condition that $Z_q = z$. The outcome of substantial calculation along the lines of Siegmund's (1988) Theorem 1 (which contains some minor errors that must be corrected) shows that for large ξ and small Δ, hence in particular for dense markers, the average size of the support region is approximately

$$\beta^{-1}\int\varphi(y - \xi)\,\{\ln[y^2/(y^2 - x)] + 2y^{-2}[1 - 2\nu(y(2\beta\Delta)^{1/2}) + 0.5\nu^2(y(2\beta\Delta)^{1/2})]\}dy. \tag{B2}$$

A similar argument in two dimensions yields a similar expression with β replaced by $\tilde{\beta}$ and the additional factor $(y/\xi)^{1/2}$ multiplying $\varphi(y - \xi)$ to approximate a noncentral χ^2 density.