STATISTICAL METHODS FOR MAPPING QUANTITATIVE TRAIT LOCI FROM A DENSE SET OF MARKERS

by

Josée Dupuis and David Siegmund

TECHNICAL REPORT NO. 8
April 1997

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Statistical Methods for Mapping Quantitative Trait Loci from a Dense Set of Markers

by

Josée Dupuis
Department of Preventive Medicine and Department of Statistics
Northwestern University, Chicago and Evanston, IL

and

David Siegmund
Department of Statistics
Stanford University

TECHNICAL REPORT NO. 8
April 1997

Prepared Under Grant NIH 1-R01-HG00848
For The National Institutes of Health

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Statistical Methods for Mapping Quantitative Trait Loci from a Dense Set of Markers

Josée Dupuis
Department of Preventive Medicine and Department of Statistics,
Northwestern University, Chicago and Evanston, IL

David Siegmund
Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

LANDER and BOTSTEIN (1989) introduced statistical methods for searching an entire genome for quantitative trait loci in experimental organisms. Their method was developed for a backcross design. We extend their results to intercross and other designs, and study the power of the resulting test through theoretical approximations and simulations. In addition we compare three methods for constructing confidence intervals for a quantitative trait locus: likelihood intervals, Bayesian credible sets and LOD support intervals.
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 genetics 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, with recent advances in genetics, 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 (QTLs) 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 paper, Lander and Botstein suggested statistical tests for general designs, but provided guidelines for declaring statistical significance, in the form of LOD score thresholds, 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 had to restrict themselves to a one degree of freedom statistic that ignored dominance effects. CHURCHILL and DOERGÉ (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; and hence the method is less useful for analyzing and comparing different designs.
In this paper we propose for intercross and other designs simple approximations that can be easily used to compare different designs under various conditions or the same design for different sample sizes or marker densities. We then discuss and compare three methods for constructing confidence intervals for a QTL.

RESULTS

The Model and Detection of Linkage

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. An intercross is generated by crossing 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

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

where $\mu$, $\alpha$, $\delta$ are the phenotypic mean, additive effect and dominance effect, respectively, and $1_C$ equals 1 or 0 according as 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’s on other chromosomes than the $j$th. Since 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 $\sigma_e^2$. The residual variance $\sigma_e^2$ equals the sum of the environmental variance and the genetic variance for those QTL's not on the $j$th chromosome. The assumption of a normal distribution requires that the residual genetic effects be individually small. Without this assumption the regression like statistics given below are not exact maximum LOD scores, 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 $e_i$'s are not normally distributed.

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

$$y_i = \mu + \alpha^* x_i(q) + e_i,$$

where the parameter $\alpha^*$ 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 postponed until the next section.

If one observes the genotype of a marker at a putative trait locus $d$, the maximum LOD score at $d$ is given approximately by

$$\text{LOD}(d) \approx -\frac{N}{2} \log(1 - \frac{\hat{\sigma}_d^2}{4\hat{\sigma}_y^2}) \approx \frac{\log e \, N\hat{\alpha}_d^2}{2 \frac{4\sigma_e^2}{}} ,$$

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 $\sigma_y^2 = \sigma_e^2 + \alpha^* / 4$. It is important to note that both $\sigma_y^2$ and $\sigma_e^2$ 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. To accommodate tradition all logarithms are to the base 10, although this leads to several
appearances of the conversion factor $\log e$ when it is more convenient mathematically to express certain results in terms of natural logarithms (denoted by $\ln$ below). For the first approximation in (3) we have replaced the empirical variance of $\{x_i(d)\}$, namely $N^{-1}\sum_i (x_i(d) - N^{-1}\sum_j x_j(d))^2$, by its asymptotic value of 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 $\sigma_e$ that it estimates under the hypothesis of no linkage on the $j$th chromosome. Since the trait locus $q$ is typically unknown, the LOD score is maximized over all marker locations $d$ and chromosomes $j$. At each marker, assumed to be a QTL, the LOD score is computed exactly. Between markers, LANDER and BOTSTEIN (1989) suggest that “interval mapping” be used to obtain the LOD score, so that the entire genome can be searched for loci influencing the quantitative trait. The interval mapping technique consists of treating the unobserved marker information as missing data and using the EM algorithm (DEMPSTER, LAIRD and RUBIN 1977) to evaluate the maximum LOD score at $d$ based on the marker information at the flanking markers. A noniterative alternative to the EM algorithm was proposed by HALEY and KNOTT (1992) and was shown to give equivalent results provided $N$ is sufficiently large.

Since the LOD score is maximized over the entire genome, it is unclear whether the usual threshold of 3.0 to declare statistical significance is appropriate in the present context. To address this issue, LANDER and BOTSTEIN (1989) proposed the approximation of $N^{1/2}\hat{\sigma}_d/2\sigma_e$ (cf. (3)) by an Ornstein-Uhlenbeck process. This can be justified by the central limit theorem and a straightforward calculation of covariances; and for the case of complete marker information (continuous markers), they gave LOD thresholds depending on the length of the genome and the number of chromosomes searched. See Proposition 2 of the original paper. 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 example, for organisms with 10 chromosomes of 100 centimorgans (cM) in length, the LOD thresholds to declare linkage depend on the density of the map and vary from 2.4 to 3.1. For organisms of genetic length greater than 1000 cM and continuous markers, using the customary bound of 3.0 would result in a probability of a false positive
result greater than the conventional 5%.

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

$$ P\{\max_k \text{LOD}(k\Delta) > \ell\text{od}\} \approx 1 - \exp\{-2C[1 - \Phi(b)] - 2\beta Lb\phi(b)\nu(b\sqrt{2\beta\Delta})\}, \quad (4) $$

where $$ b^2 = 2\ell\text{od}/\log e $$ (cf. (3)), $$ L $$ is the total length of the genome, $$ C $$ is the number of chromosomes, $$ \beta = 2\lambda $$, $$ \lambda $$ being the rate of crossovers ($$ \lambda = 1 $$ if $$ L $$ is in Morgans and $$ \lambda = 0.01 $$ if $$ L $$ is in cM), $$ \Delta $$ 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 $$ \nu $$ is a discreteness correction for the distance $$ \Delta $$ between markers. An exact expression can be found in Siegmund (1985) p.82. More simply, $$ \nu(x) $$ can be approximated by $$ e^{-0.583x} $$ for small values of $$ x $$. For the case of continuous markers $$ \Delta = 0 $$, so $$ \nu = 1 $$.

Remark: Approximation (4) is based on the LOD score calculated at markers, without taking the interval mapping step into account. Slightly higher thresholds are required for the interval mapping statistic. This point is discussed in detail below.

Lander and Botstein's analysis did not consider dominance effects. PATERSON et al. (1991) used the full model (1) to locate QTLs in tomatoes in an intercross, where the dominance effects could be estimated. It is clear that for a given threshold the false positive error rate will be larger since a two degree of freedom statistic is involved. However, the statistical significance of the dominance effects was not assessed, for lack of methodology. Such methodology is developed in the next section.

Thresholds for other designs: Other types of breeding designs, such as the intercross, allow for the estimation of both the additive and dominance effects. The likelihood ratio (equivalently the maximum LOD score) can then be used to test the more general hypothesis of $$ H_0 : \alpha = \delta = 0 $$ versus $$ H_1 : \alpha \neq 0 $$ or $$ \delta \neq 0 $$.

If the residual variability is assumed normally distributed, the maximum likelihood ratio at a putative QTL $$ d $$ can be written as
\[ \frac{\phi\left( \frac{y_i - \hat{\mu}_d - \hat{\alpha}_d x_i(d) - \hat{\delta}_d 1(x_i(d) = 1)}{\hat{\sigma}_d} \right)}{\phi\left( \frac{y_i - \bar{y}}{\hat{\sigma}_y} \right)} \]

where \( \hat{\mu}_d, \hat{\alpha}_d, \hat{\delta}_d, \hat{\sigma}_d^2 \) are the maximum likelihood estimates at the putative QTL \( d \) under \( H_1 \) and \( \hat{\sigma}_y^2 \) is the maximum likelihood estimate of the phenotypic variance of \( y \). Since the gene locus is rarely known, the above (or its logarithm) is maximized over all possible values of \( d \).

For intercross data the vectors \( x_i(d) \) and \( 1(x_i(d) = 1) \) are asymptotically orthogonal. Therefore, the approximations used to obtain (3) now yield

\[ \text{LOD}(d) \approx -\frac{N}{2} \log \left\{ 1 - \frac{\hat{\sigma}_d \hat{\alpha}_d^2 / 2 + \hat{\delta}_d^2 / 4}{\hat{\sigma}_y^2} \right\} \approx \log e \left\{ \left( \frac{N^{1/2} \hat{\alpha}_d}{2^{1/2} \sigma_e} \right)^2 + \left( \frac{N^{1/2} \hat{\delta}_d}{2 \sigma_e} \right)^2 \right\}. \]

To define a significance level, we give an approximation under the hypothesis of no linkage to the distribution of the maximum of (6) over all possible values of \( d \).

Let

\[ X_d = \frac{N^{1/2} \hat{\alpha}_d}{2^{1/2} \sigma_e}, \quad \text{and} \quad Y_d = \frac{N^{1/2} \hat{\delta}_d}{2 \sigma_e}. \]

Under the hypothesis of no linkage \( \alpha = 0, \delta = 0 \). A straightforward application of the central limit theorem and calculation of covariances shows that 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 (6) is provided by

\[ P\left\{ \max_d \text{LOD}(d) \geq \ell \text{od} \right\} \approx 1 - \exp\left\{ -C + b^2 L\left( \frac{\beta_1 + \beta_2}{2} \right) \right\} \exp\left\{ -b^2 / 2 \right\}, \]

where \( \beta_1 = 2\lambda, \beta_2 = 4\lambda \) and \( b^2 = 2\ell \text{od} / \log e \). For the case where the value of \( x_i(d) \) is only known at discrete markers with distance of \( \Delta \) between them, the second term in the exponential in (8) is multiplied by \( \nu \left( b \left\{ \Delta (\beta_1 + \beta_2) \right\} \right)^{1/2} \). As in the case of the approximation (4), this approximation does not take into account the interval mapping step. The implication of this omission will be studied using simulated data.

The first term in the exponential of (8) is to account for the probability that the LOD score exceeds the threshold at the first marker of the chromosomes. The more important second term was obtained by a suitable modification of WOODROOFE's
(1976) argument, which is presented in Appendix A. For an idealized tomato genome consisting of 12 chromosomes of length 100 cM each and a dense set of markers, the 0.05 false positive threshold obtained from (8) is $lod = 4.13 \ (b = 4.36)$, in comparison with $lod = 3.17 \ (b = 3.82)$ for the backcross case. The true false positive rate for a putative 0.05 level test if one were to use the backcross threshold for an intercross with continuous markers would be about 0.3.

![Graph showing LOD scores vs Distance between markers in cM](image)

**Figure 1.** Thresholds for 350 simulated tomato genomes.

To check the accuracy of (8), we simulated thresholds for the LOD score based on an intercross sample of $N = 350$ organisms with 12 chromosomes of total length 1200 cM (to approximate the tomato genome). The Haldane mapping function was used in the simulations and the variance of the error terms was assumed to be unknown. 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.
The discrete approximation for the process at the marker is very accurate. As predicted, the process with the interval mapping step requires a higher threshold for a given value of the type-I error. However, a conservative bound for the interval mapping process is provided by the approximation for the marker process with twice as many markers. For example, for $\Delta = 20$ the interval mapping threshold was $lod = 3.31$, which is between the approximated thresholds using $\Delta = 20$ and $\Delta = 10$, which are 3.12 and 3.36 respectively. For smaller $N$, somewhat larger thresholds need to be used since the approximation does not take into account the variability in the estimate of the variance, $\sigma^2_y$. However, when $N$ is large ($> 200$), the approximation provides a much more appropriate threshold than the conventional 3.0 or the value obtained by assuming only additive effects are present. In mapping human traits, LANDER and KRUGLYAK (1995) have argued that since 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 $\Delta$, which should otherwise be an average intermarker distance in the neighborhood of detected linkages, nor to concern ourselves about the effect of interval mapping on the false positive error rate. But insistence on this threshold would substantially reduce the power of the test, as will be shown shortly.

Most other types of design can be handled by approximations (4) or (8). For instance, for recombinant inbred data, one can use approximation (4) with $\beta = 0.08$ (as initially suggested by LANDER and BOTSTEIN 1989). In STUBER et al. (1992), offspring from a cross of two inbred Maize strains (F1 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))

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

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 $R(d) = 1 - \frac{2}{3}\lambda |d| + o(|d|)$ as $d \to 0$. Therefore, approximation (4) can be used with $\beta = \frac{2}{3}\lambda$ to find an appropriate threshold.
Other designs can be handled similarly.

The Gaussian approximation underlying (8) is similar to the Gaussian approximation arising in allele sharing analysis of affected sib pairs when the degree of recessiveness of a qualitative trait is unknown and one uses the likelihood ratio test to detect linkage (e.g. HOLMANS 1993, FARAWAY 1993). An analysis of this process, which involves some added technicalities because of constraints the parameters analogous to $\alpha$ and $\delta$ must satisfy, is given in Appendix C.

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 $\tau$, this would require a $\tau$ dimensional version of (4) or a $2\tau$ dimensional version of (8) for the backcross or intercross design, respectively. The appropriate $k$ dimensional approximation ($k = \tau$ or $2\tau$) is given by

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

Here $F_k$ is the $\chi^2$ distribution with $k$ degrees of freedom, $\Gamma$ denotes the gamma function, and $\beta$ would be replaced by $(\beta_1 + \beta_2)/2$ for an intercross design. Corrections for discrete spacing of markers would be exactly as above. See SIEGMUND (1985, Chapter 5).

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

$$P\left\{ \max_k \text{LOD}(k\Delta) > \ellod \right\} \approx 1 - \Phi(b - \xi) + \phi(b - \xi)\left[2\nu/\xi - \nu^2/(b + \xi)^2\right], \quad (9)$$

where $b^2 = 2\ellod/\log e$, $\xi = \{N \ln[1 + (\alpha + \delta)^2/4\sigma^2]\}^{1/2}$, and $\nu = \nu\{b(2\beta\Delta)^{1/2}\}$, as defined previously. The parameter $\xi$ is the non-centrality parameter of (3) expressed in terms of the parameters of the model (1). The first term in (9) 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. Often 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. A more complex approximation, which requires a one dimensional numerical integration, can be found in DUPUIS (1994). For intercross data the noncentrality
parameter is \( \xi = \{N \ln[1 + (\alpha^2/2 + \delta^2/4)/\sigma^2]\}^{1/2} \). To attribute appropriate parts of
the total noncentrality to the two processes in (7), we let \( \xi_1 = \xi \alpha/(\alpha^2 + \delta^2/2)^{1/2} \), \( \xi_2 = \xi \delta/[2(\alpha^2 + \delta^2/2)]^{1/2} \). For the case that the QTL is located at a marker, the power is
approximately

\[
P\{\max_k \text{LOD}(k\Delta) > \ell_0 \} \approx 1 - \Phi(b - \xi) + \phi(b - \xi) \left[ \frac{1}{2\xi^2} + \frac{2b^{1/2}\nu}{\xi^{3/2}} - \frac{b^{1/2}\nu^2}{\xi^{1/2}(b + \xi)} \right],
\]

where \( \nu = \nu(b \{2\beta\Delta\}^{1/2}) \), \( \beta = (\beta_1 \xi_1^2 + \beta_2 \xi_2^2)/\xi^2 \).

The approximations given in (9) and (10) are for the process at the markers only, and do not take into account the interval mapping step. For QTL's located midway between markers interval mapping can be expected to utilize the information at flanking markers to provide an increase in power, although the requirement to use a slightly higher threshold to maintain a fixed frequency of false positives removes some of the edge that interval mapping has over using the marker process only. Using simulations and the theoretical power approximations above, we compared the power of the marker process with the power of the interval mapping test with adjusted thresholds. 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 \( \xi = 3.75, 4.76, 5.17 \) and 5.52, which correspond roughly to power of 50%, 80%, 90% and 95% with a continuous map of markers. Exactly the same results would be obtained for a recessive model (\( \delta = -\alpha \)); for an additive model (\( \delta = 0 \)) for the same values of \( \xi \) the power would be slightly larger, with the greatest difference occurring for relatively sparse markers and a QTL midway between markers. Power under two map densities was estimated (\( \Delta = 20 \) cM and 5 cM) and we used \( N = 350 \) tomato genomes. Each power simulation is based on 1000 replicates. From Figure (2) we see that the gain in power from using interval mapping is small, on the order of 2-4%, a result similar to what DARVASI et al. (1993) found. The gains anticipated by LANDER and BOTSTEIN (1986, 1989), who write of interval mapping as providing a "virtual marker" midway between the actual markers, is overly optimistic. Their analysis is marred by their comparison of interval mapping with the marker process at 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 for the interval mapping process, but this is less important. The gain in power for interval mapping is largest for the sparse map ($\Delta = 20 \text{ cM}$). On the other hand, 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%). The approximation suggested in (10) (suitably modified when the QTL is midway between markers) and the simulation results agree closely, with the largest difference being about 3%.

Figure 2.: Power to detect linkage for different map densities, gene locations and thresholds. In Figures (a) and (b), $\Delta = 5 \text{ cM}$ while $\Delta = 20 \text{ cM}$ in Figures (c) and (d). The trait locus is located at a marker in Figures (a) and (c) and mid-markers in (b) and (d). The theoretical approximation is represented by (□), the process without interval mapping by (○), the process with interval mapping by (◊) (simulated threshold for interval mapping) and (▽) (power for the higher threshold appropriate when $\Delta = 0$).
We have also used the theory developed above to compare the power of backcross, intercross and recombinant inbred designs. Let \( \sigma_A^2, \sigma_D^2, \sigma_E^2 \) 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 and put \( \rho^2 = \sigma_D^2/\sigma_A^2 \). It simplifies the discussion to normalize out one parameter by setting \( \sigma_A^2 = 1 \). Then \( \sigma_E^2 = (1 + \rho^2)(1/H^2 - 1) \). To reduce the number of different special cases we also assume that \( \rho^2 = \delta^2/2\alpha^2 \) at all QTL's, i.e., the relative amount of dominance is the same across QTL's. Then the noncentrality parameters of an intercross, backcross, and recombinant inbred design are respectively \([-N \ln(1 - \alpha^2 H^2/2)]^{1/2},
\{-N \ln[1-(\alpha(1+2^{1/2}\rho)/2)^2/(\sigma_E^2+(1+2^{1/2}\rho)^2/2)]\}^{1/2}
\{-N \ln[1-\alpha^2/(\sigma_E^2+2)]\}^{1/2}\).

<table>
<thead>
<tr>
<th>( H^2 )</th>
<th>( N )</th>
<th>( \rho )</th>
<th>Backcross</th>
<th>Recombinant Inbred</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.75</td>
<td>94</td>
<td>0.0</td>
<td>0.69</td>
<td>0.82</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.80</td>
<td>0.81</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.2</td>
<td>0.45</td>
<td>0.81</td>
</tr>
<tr>
<td>0.25</td>
<td>301</td>
<td>0.0</td>
<td>0.47</td>
<td>0.97</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.2</td>
<td>0.72</td>
<td>0.96</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.2</td>
<td>0.18</td>
<td>0.96</td>
</tr>
</tbody>
</table>

Table 1.: Theoretical power of backcross, intercross and recombinant inbred designs. The sample sizes were chosen so that the intercross design has power 0.7 for the indicated parameter values.

A numerical example is given in Table 1. We have determined for continuous markers and an intercross design sample sizes that give 70% power for various values of \( H^2 \) and \( \alpha \). (The noncentrality for an intercross, hence the sample sizes, do not depend on \( \rho \).) For these sample sizes we have computed the power of backcross and recombinant inbred designs for those \( H^2 \) and \( \alpha \), and selected values of \( \rho \). It turns out that the power of the backcross and recombinant inbred designs depends primarily on \( H^2 \), and is relatively insensitive to the exact value of \( \alpha \) in the range \( 1/2 \leq \alpha \leq 1 \),
where the QTL contributes from one eighth to one half \( \sigma^2_A \). Hence \( \alpha \) is not given in the table. For \( \rho = 0 \) the power of the backcross design can be about the same as that of the intercross (70%), but also can be appreciably less. For \( \rho^2 = 0.04 \), the backcross design can be somewhat more powerful or much less powerful than the intercross design depending on whether \( \rho \) is positive or negative. The power of the recombinant inbred design is higher than that of the intercross design and is relatively insensitive to the values of \( \rho \). Hence with a small amount of dominance a backcross design can yield a very misleading picture.

For an intercross design we have emphasized the 2-df statistic that allows one to detect additive and dominance effects simultaneously. A test statistic that deserves some attention is the 1-df statistic designed to detect only an additive effect (e.g., PATERSON et al. 1991). It is easy to see that the threshold would be the same as for a backcross design, and the noncentrality parameter is \([-N \ln(1-\alpha^2H^2/2(1+\rho^2))]^{1/2}\). For the situation detailed in Table 1, where the amount of dominance is very small, the power of this statistic is about 0.8, i.e., larger than the power of the two degree of freedom statistic. If the trait is dominant in the sense that \( \delta = \alpha \) or recessive with \( \delta = -\alpha \), so \(|\rho| = 2^{-1/2} \), the power of the 1-df statistic falls to about 0.50-0.55.

**Confidence Regions for QTLs**

A confidence region can be used to identify a chromosomal region in which to concentrate the search for the exact location of the 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 maximum likelihood estimator plus or minus two estimated standard errors as an approximate 95% confidence interval. DARVASI et al. (1993) 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 is known to give confidence regions that are unnecessarily large in related "change-point" problems; and it appears to have the same undesirable feature here. See SIEGMUND (1988) for a more complete discussion.

**LOD Support Intervals:** LOD support intervals (cf. CONNEALLY et al. 1985) are a standard means of estimating the location of a trait locus. An $x$-LOD support region includes all the loci $q$ such that

$$\text{LOD}(q) \geq \max_d \text{LOD}(d) - x.$$  \hspace{1cm} (11)

In the case of two point linkage analysis (i.e. using the information from a single linked marker), a one-LOD support interval is a better than 95\% confidence interval (cf. OTT 1991, p. 67). However, because of the irregularity mentioned above, this result does not generalize to genome wide scans, where the coverage probability 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 LOD support region. However, through theoretical analysis and a simulation study presented below, we show that a 1-LOD (1.5-LOD) support interval corresponds roughly to a 90\% (95\%) confidence region in the case of dense map of markers ($\sim 1$ 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 LOD support method described above and provides some analytic tools for studying that concept. Unlike the LOD 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 the LOD score, our discussion will be simplified by using the asymptotically equivalent $||Z_d||^2$, where $Z_d = (X_d, Y_d)$ is defined in (7) (cf. also (6)) 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

$$A_q = \{ \max_d ||Z_d||^2 - ||Z_q||^2 \leq c^2 \}.$$
By sufficiency, the conditional probability of $A_q$ given $Z_q$ does not depend on the unknown parameters $\alpha, \delta$. Hence in principle we can choose $c = c(Z_q)$ such that

$$P(A_q|Z_q) = 1 - \gamma. \quad (12)$$

The set of all values $q$ that are not rejected by this test is a $(1 - \gamma)100\%$ confidence region (COX and HINKLEY 1983). It is not actually necessary to solve for $c$ since

$$(\max_d ||Z_d||^2)_\text{obs} - ||Z_q||^2 \leq c^2 \iff \quad (13)$$

$$P\left\{ \max_d ||Z_d||^2 > (\max_d ||Z_d||^2)_\text{obs} | Z_q \right\} \geq P\left\{ \max_d ||Z_d||^2 > c^2 + ||Z_q||^2 | Z_q \right\} = \gamma.$$ 

Since the desired conditional probability does not depend on $\alpha, \delta$, it can be evaluated under the hypothesis that these parameters are both 0. The approximation (B1) of Appendix B in conjunction with equation (13) yields as a confidence interval for the QTL those loci $q$ such that

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

The likelihood method works best for very dense sets of markers ($\sim 1 \text{ cM}$), since the argument given above is technically correct only when the QTL is a marker. It can be extended to provide a joint confidence region for the locus and the additive and dominance effects (DUPUIS 1994).

By (6) the inequality defining $A_q$ and the inequality (11) are asymptotically equivalent if $c^2 = 2x/(\log e)$. The important difference between the likelihood ratio and LOD support method is that for the former $c$ depends on $Z_q$ and is chosen to make the conditional probability (12) equal to the desired confidence level. For any value of $c$ or equivalently $x$ that does not depend on the data, the probability of (11) depends on the values of $\alpha$ and $\delta$. Hence the LOD support region is not a confidence region in the strict sense of the word. However, the similarity between the LOD support regions and the likelihood ratio regions allows us to gain some interesting theoretical insights into LOD support regions. For example, under the assumption that the QTL lies at a marker locus and that the distance $\Delta$ between markers is small, we can take the expectation of (B1) with respect to $Z_q = z$ to evaluate the probability that a LOD
support region does not contain the true QTL. In order to integrate with respect to
the distribution of $Z_q$, we convert to polar coordinates and use the fact that for large
$\xi$ the angular variable is essentially a point mass at $\tan^{-1}(\xi_2/\xi_1)$. The result of some
approximations is

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

(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 $\Delta$ the coverage probability of the LOD support region is relatively insensitive
to the values of $\xi$ and to the relative sizes of the additive and dominance components,
at least for values of $\xi$ in the range $4 \leq \xi \leq 10$ where detection of linkage ranges
from reasonably likely to almost certain, so QTL localization is especially important.
Hence for practical purposes a LOD support region is approximately a confidence
region.

For problems involving a single parameter, e.g., for backcrosses, recombinant in-
breds, or intercrosses where we estimate only $\alpha$ and ignore $\delta$, the factor in square
brackets in (15) immediately preceding the exponential would be $[(\xi^2 + c^2)/\xi^2]^{1/2}$. It
is easy to see that at least for comparatively large values of $\xi$, the coverage probability
for a given value of $c$ is relatively insensitive to this change of dimension—less sensitive
than would be suggested by a comparison of $\chi^2$ quantiles with 1 and 2 degrees
of freedom, which would be appropriate for a regular statistical problem.

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 recombi-
nant inbreds and assume as before that a marker is at the QTL $q$. The expected size
of the LOD support region is

$$\Delta \Sigma_k P\{Z_{k\Delta}^2 \geq \max Z_{j\Delta}^2 - c^2\}$$

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

where $P^z$ denotes probability under the condition that $Z_q = z$. The outcome of
substantial calculation along the lines of SIEGMUND (1988) Theorem 1 (which con-
tains some minor errors that must be corrected) shows that for large $\xi$ and small
\( \Delta \), hence in particular for dense markers, the average size of the support region is approximately

\[
\beta^{-1} \int \phi(x-\xi) \{ \ln[x^2/(x^2-c^2)] + 2x^{-2}[1-2\nu(x(2\beta\Delta)^{1/2}) + 0.5\nu^2(x(2\beta\Delta)^{1/2})] \} dx. \tag{16}
\]

A similar argument in two dimensions yields a similar expression with \( \beta \) replaced by \( \tilde{\beta} \) and the additional factor \((x/\xi)^{1/2}\) multiplying \( \phi(x-\xi) \) to approximate a noncentral \( \chi \) density. A less precise but more easily interpreted approximation is obtained by recognizing that when \( \xi >> c \) the normal density in (16) with mean \( \xi \) can be approximated by a point mass at \( \xi \), so that, after taking two terms of the Taylor series expansion of \( \ln[\xi^2/(\xi^2-c^2)] \), we obtain

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

This expression is roughly proportional to \( \xi^{-2} \), 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 LOD support region is inversely proportional to \( \beta \), 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 1/4th the size of those obtained from a backcross. In fact, for additive traits recombinant inbreds always have a larger noncentrality parameter than a backcross, so they provide support regions even less than 1/4 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 \((\xi_1, \xi_2)\), 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, which is very easy to compute, will provide satisfactory confidence regions for that location.

A Bayesian credible region $B_\gamma$ is constructed by including all loci $v$ whose posterior density given the data exceeds $c_\gamma$, i.e.,

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

where $c_\gamma$ is chosen so that

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

Here $y = \{y_1, \ldots, y_N\}$, $x = \{x_1, \ldots, x_N\}$ and $x_i$ is the set of all marker genotypes for individual $i$. The posterior probability $\pi(v|y, x)$ is often easy to compute and depends on the prior distribution on the location $q$ and the additive and dominance effects $\alpha$ and $\delta$. If one takes uninformative priors on all parameters,

$$\pi(v|y, x) \propto \frac{\exp\left(-\frac{1}{4} ||Z_v||^2\right)}{\int_0^1 \exp\left(-\frac{1}{4} ||Z_s||^2\right) ds},$$

where $Z_t = (X_t, Y_t)$ was defined previously and can be obtained using least squares estimates or the interval mapping equivalent. If one takes a bivariate normal prior on the scaled additive and dominance effect with means $\theta_1$ and $\theta_2$, variance $\eta_1$ and $\eta_2$ and null correlation, then

$$\pi(\tau|y, x) \propto \frac{\exp\left(\frac{(X_\tau + \theta_1/\eta_1)^2}{2(1+1/\eta_1^2)}\right) \exp\left(\frac{(Y_\tau + \theta_2/\eta_2)^2}{2(1+1/\eta_2^2)}\right)}{\int_0^1 \exp\left(\frac{(X_s + \theta_1/\eta_1)^2}{2(1+1/\eta_1^2)}\right) \exp\left(\frac{(Y_s + \theta_2/\eta_2)^2}{2(1+1/\eta_2^2)}\right) ds}. \quad (17)$$

19
Analogous expressions can be obtained for other priors. Properties of three different priors on the additive and dominance effects were studied, 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 additive and dominance effects are around 4, a mixture of uncorrelated normal priors centered at (4, 4), (4, -4), (-4, 4) and (-4, -4) with variance of 1 was also applied. Results are presented in the next section.

**Comparison Study:** Using simulated tomato genomes, we constructed the likelihood confidence set, the 1.5-LOD support interval and the Bayes credible sets, with the three different priors mentioned above. However, only the results from Bayes credible sets with a mixture of normal priors are included in the Figures. 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

\[ y_i = \alpha x_i(q) + \delta \mathbb{1}(x_i(q) = 1) + e_i, \]

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, mid-way 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 based on a map with \( \Delta = 1, 5 \) and 10 cM, assuming the error variance to be unknown. For sparser maps (\( \Delta > 1 \) cM), the likelihood intervals were not constructed since this method is only valid for dense maps, while the Bayesian credible sets and the 1.5-LOD support intervals were computed using the interval mapping procedure. We defined the Bayes credible interval to be the smallest *connected* interval with posterior probability of 95%. Insisting on connected intervals had only a small effect of the size of the confidence regions while increasing the coverage probability substantially. The results are presented in Tables 2-3.
<table>
<thead>
<tr>
<th>$\xi$</th>
<th>Method</th>
<th>$\Delta = 1$</th>
<th></th>
<th>$\Delta = 5$</th>
<th></th>
<th>$\Delta = 10$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0  r  1/2</td>
<td></td>
<td>0  r  1/2</td>
<td></td>
<td>0  r  1/2</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>Lik.</td>
<td>17.2 17.0 18.9</td>
<td></td>
<td></td>
<td></td>
<td>21.8 25.4 26.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>14.0 14.3 15.5</td>
<td></td>
<td>19.1 21.1 22.0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>17.4 17.1 18.6</td>
<td></td>
<td>17.0 18.3 19.0</td>
<td></td>
<td>17.1 19.3 20.1</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>Lik.</td>
<td>5.4  5.8  5.8</td>
<td></td>
<td></td>
<td></td>
<td>10.7 12.8 14.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>5.5  5.8  5.8</td>
<td></td>
<td>8.1  9.4  9.9</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>7.3  7.3  7.3</td>
<td></td>
<td>8.0  9.0  9.3</td>
<td></td>
<td>10.9 11.3 11.9</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>Lik.</td>
<td>3.1  3.3  3.3</td>
<td></td>
<td></td>
<td></td>
<td>7.6  8.8  8.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>3.1  3.2  3.3</td>
<td></td>
<td>5.5  6.1  6.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>4.1  4.7  4.6</td>
<td></td>
<td>6.6  6.8  6.6</td>
<td></td>
<td>10.5 9.4  8.1</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.: Average size in cM of simulated confidence intervals. Three locations for the QTL were simulated: (0) for the trait at a marker, (1/2) for the trait mid-markers and (r) for the QTL randomly located between markers.

<table>
<thead>
<tr>
<th>$\xi$</th>
<th>Method</th>
<th>$\Delta = 1$</th>
<th></th>
<th>$\Delta = 5$</th>
<th></th>
<th>$\Delta = 10$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0  r  1/2</td>
<td></td>
<td>0  r  1/2</td>
<td></td>
<td>0  r  1/2</td>
<td></td>
</tr>
<tr>
<td>5.0</td>
<td>Lik.</td>
<td>95.2 95.6 95.9</td>
<td></td>
<td></td>
<td></td>
<td>98.9 98.3 98.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>95.1 95.6 96.2</td>
<td></td>
<td>97.8 97.0 96.5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>95.7 95.2 95.7</td>
<td></td>
<td>94.6 94.0 93.0</td>
<td></td>
<td>90.4 93.9 93.8</td>
<td></td>
</tr>
<tr>
<td>7.5</td>
<td>Lik.</td>
<td>94.8 96.3 97.4</td>
<td></td>
<td></td>
<td></td>
<td>99.1 99.1 97.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>96.2 96.4 97.9</td>
<td></td>
<td>99.4 98.8 98.6</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>97.2 95.4 95.0</td>
<td></td>
<td>95.8 97.3 96.1</td>
<td></td>
<td>90.1 94.6 97.6</td>
<td></td>
</tr>
<tr>
<td>10.0</td>
<td>Lik.</td>
<td>97.8 98.3 97.8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.5-LOD</td>
<td>98.0 98.1 97.2</td>
<td></td>
<td>98.7 99.3 98.3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Bayes</td>
<td>98.5 97.3 91.0</td>
<td></td>
<td>95.2 97.4 98.5</td>
<td></td>
<td>91.3 94.2 99.5</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.: Coverage probability of simulated confidence intervals.
The 1.5 LOD support interval gave the smallest confidence regions while providing at least 95% coverage under all simulated conditions. The coverage probability and expected size obtained in the simulations was close to that predicted by the approximation (15) when \( \Delta = 1 \). (The mathematical analysis requires the assumption that \( \Delta \approx 0 \).) The likelihood method was conservative; and because it adapts to the observed value of the LOD score 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 LOD support region for the larger values \( \xi = 7.5 \) and 10. For all methods, the size of the intervals were largest when the trait was mid-marker. The Bayes credible sets were the widest and they fell short of the desired 95% for large values of \( \xi \) 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 \( \xi \) is large, the dense marker map provides substantially smaller regions.

The simulations were repeated with fewer tomatoes (\( N = 100 \)) (results not shown), keeping the power constant; the sizes of the confidence sets were increased by up to 50% for the likelihood and Bayes methods. However, the lengths of the 1.5-LOD support interval were unaffected. All three methods had conservative coverage probabilities when the sample size was reduced. For our range of \( \xi \) values, the 1.5-LOD support interval was most robust to small sample sizes, while providing approximate 95% confidence regions for sparse and dense maps.

We have also simulated LOD 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 regions had coverage probabilities ranging from 93% to 96% and in all cases gave smaller regions than the 95% bootstrap regions recommended by VISSCHER et al., 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.

DISCUSSION

LANDER and BOTSTEIN (1989) presented a framework to search an entire genome to locate genetic loci that influence quantitative traits. Their method was developed for data originating from a backcross design. In the present paper, their model has been extended to breeding designs other than the backcross, especially the intercross. Approximations for the power show that an intercross design is substantially more efficient than a backcross, which can yield misleading results in the presence of even small departures from additivity ($\delta = 0$). A recombinant inbred design is more efficient than an intercross, except when dominance effects are large compared to additive effects.

We also presented three methods of constructing confidence intervals for the location of QTLs: the likelihood method, Bayes credible sets and LOD support regions. The LOD support method seems preferable to the other methods, since it yields the smallest confidence regions without regard to the density of the map of markers or the sample size.

When dominance effects are relatively small, 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. In almost all cases, however, the size of the confidence regions is on the order of several cM 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's.

Although the methods of this paper do not explicitly address the complexities associated with identifying multiple, possibly linked QTL's, they can be adapted to deal with cofactors, as discussed, for example, by JANSEN (1994). We have studied similar problems related to mapping qualitative traits in humans (DUPUIS, BROWN and SIEGMUND 1995) and hope to return to this issue in the specific context of mapping QTL's in the future.
LITERATURE CITED


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

KOROL, A.B., RONIN, Y.I. and V.M. KIRZNER, 1995 Interval mapping of quantitative trait loci employing correlated trait complexes, Genetics 140: 1137-1147.


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 N. J. SCHORK, 1994 Genetic Dissection of complex traits, Science 265: 2037-2048.


APPENDIX A

Derivation of approximation (8):

Let $Z = (Z_1, Z_2)$, where $Z_1$ and $Z_2$ are two independent Gaussian processes with covariance functions satisfying

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

For large $b$,

$$P\left\{ \max_{0 \leq i \Delta \leq l} \|Z_{i\Delta}\| \geq b \right\} \approx e^{-\frac{1}{4}b^2} + b^2 t \left( \frac{\beta_1 + \beta_2}{2} \right) \nu \left( b \sqrt{\Delta (\beta_1 + \beta_2)} \right) e^{-\frac{1}{4}b^2}. \quad (A1)$$

Since the LOD score can be approximated by the scaled sum of two independent Ornstein-Uhlenbeck processes, (A1) can be used to determine the threshold needed for a search involving a single chromosome of length $l$. To get a genome wide threshold, equation (A1) and the independent meiotic assortment of chromosomes give approximation (8).

Derivation of approximation (A1)

We can condition on the position where the process last exceeded the value $b$, how far above $b$ the process reached at that position and the angle between $Z_{1,i\Delta}$ and $Z_{2,i\Delta}$ to obtain (A1).

Define $D_i = \{ j : j \geq 1, (i + j)\Delta \leq l \}$, where $l$ and $\Delta$ are fixed. Let $\omega_{i\Delta}$ be the angle between $Z_{1,i\Delta}$ and $Z_{2,i\Delta}$. Then, we can write

$$P\left\{ \max_{0 \leq i \Delta \leq l} \|Z_{i\Delta}\| \geq b \right\} = \sum_{i=0}^{l/\Delta} \int_0^{\infty} \int_{0}^{\infty} P\left\{ \|Z_{i\Delta}\| \in b + dy, \omega_{i\Delta} \in dw \right\} P\left\{ \|Z_{(i+j)\Delta}\| < b : \forall j \in D_i \left\| Z_{i\Delta} \right\| = b + y, \omega_{i\Delta} = w \right\}$$  

(A2)

By the independence of $Z_1$ and $Z_2$ and the Gaussian nature of the processes,

$$P\left\{ \|Z_{i\Delta}\| \in b + dy, \omega_{i\Delta} \in dw \right\} \sim \frac{1}{{\sqrt{2\pi}}} b \phi(b + y) dy dw, \quad \text{for} \quad y = O(1/b). \quad (A3)$$

28
Using an argument similar to SIEGMUND (1985) p. 202, the second term in the integral can be written as

\[ P_{\mu, \sigma}\left\{ \max_{0 \leq (t+j)\Delta < l} S_j < -y \right\}, \]  

(A4)

where \( S_j \) is the sum of \( j \) independent normal random variables with mean \(-\mu\) and variance \( \sigma^2; \mu = b\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w) \) and \( \sigma^2 = 2\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w) \). Substituting (A3) and (A4) into (A2), we obtain

\[
P\left\{ \max_{0 \leq \Delta \leq l} ||Z_\Delta|| \geq b \right\} \sim \frac{l}{\Delta} \int_{-\pi}^{\pi} \frac{1}{\sqrt{2\pi}} b \phi(b) \int_{0}^{\infty} e^{-by} P_{\mu, \sigma}\left\{ \min_{j \geq 0} S_j > y \right\} dydw
\]

\[ \sim b\Delta^{-1}e^{-\frac{1}{2}b^2} (2\pi)^{-1} \int_{-\pi}^{\pi} \mu \nu(2\mu/\sigma) dw \approx b^2 l e^{-\frac{1}{2}b^2} \left( \frac{\beta_1 + \beta_2}{2} \right) \nu\left( b\sqrt{\Delta(\beta_1 + \beta_2)} \right), \]  

(A5)

where the function \( \nu(x) \) is defined on p. 82 of SIEGMUND (1985) and can be approximated by \( \exp(-0.583x) \). The first approximation is obtained by ignoring the term \( \exp(-y^2/2) \), which is negligible when \( y = O(1/b) \) and noticing that except for edge effects the terms involved in the summation in (A.2) are all (approximately) equal. The second approximation follows follows from Corollary 8.45 of SIEGMUND (1985). These two approximations have rigorous mathematical interpretations, provided \( b \to \infty, \Delta \to 0, \) and \( b\Delta^{1/2} \) is bounded away from 0 and \( \infty \). The final approximation is for computational convenience; we have replaced the average indicated by the integral with respect to \( w \) by the integrand evaluated at the average value of \( \mu \) and \( (\mu/\sigma)^2 \). We have evaluated numerically the next to last approximation under a wide variety of conditions and have found that it invariably gives almost exactly the same threshold as the last approximation. Consequently we have chosen the more easily evaluated last expression as our recommended approximation.

APPENDIX B

Approximating the conditional probability in (14)

We first state a lemma with its derivation in order to obtain approximation (14).
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

\[
R_t(t) = 1 - \beta_i|t| + o(|t|) \quad \text{as} \quad t \to 0.
\]

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

\[
\begin{pmatrix}
z_1 \\
z_2
\end{pmatrix} = \begin{pmatrix}
br_1(t^*) \cos w^* \\
br_2(t^*) \sin w^*
\end{pmatrix}.
\]

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

\[
P\left\{ \max_{0 \leq \Delta \leq t} ||Z_{i\Delta}|| \geq b \mid Z_0 = z \right\}
\]

\[
\sim \frac{\beta \exp[-\frac{1}{2}(b^2 - ||z||^2)]}{\left| \dot{R}_1(t^*)R_2(t^*) \cos^2 w^* + R_1(t^*)\dot{R}_2(t^*) \sin^2 w^* \right|} \nu[b(2\beta\Delta)^{\frac{1}{2}}],
\]

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 \( b^2 = ||z||^2 + c^2 \) and assuming \(|cz_2| << |z_1|\), which will be the case with probability close to one unless there is overdominance, we obtain

\[
P\left\{ \max_{0 \leq \Delta \leq t} ||Z_{i\Delta}||^2 > ||z||^2 + c^2 \mid Z_0 = z \right\}
\]

\[
\approx \frac{[2(||z||^2 + c^2)]^{3/2} \exp(-\frac{1}{2}c^2)}{\left( z_1^2 + \sqrt{(z_1^2 + 2z_2^2)^2 + 4z_2^2c^2} \right)^{3/2}} \nu\left(2\Delta(\beta_1 z_1^2 + \beta_2 z_2^2)(1 + c^2/||z||^2)^{\frac{1}{2}}\right). \quad (B1)
\]

**Derivation of the Lemma:** The argument is similar to that found in Appendix A, but the details are substantially more difficult. We first condition on the position where the process last exceeded the value \( b \), how far above \( b \) it reached at that position and the angle between \( Z_{1,t} \) and \( Z_{2,t} \).
Define \( D_{l} = \{ j : j \geq 1, (i + j)\Delta \leq l \} \), where \( l \) and \( \Delta \) are fixed. Let \( \omega_{i\Delta} \) be the angle between \( Z_{1,i\Delta} \) and \( Z_{2,i\Delta} \). Then, we can write
\[
P\{\max_{0\leq i\Delta\leq l} ||Z_{i\Delta}|| > b \mid Z_{0} = z\} = \sum_{i=0}^{l} \int_{\frac{1}{\Delta}}^{\infty} \int P\{||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_{0} = z\} \]
\[
* P\{||Z_{(i+j)\Delta}|| < b \; \forall j \in D_{l} \mid ||Z_{i\Delta}|| = b + y, \omega_{i\Delta} = w, Z_{0} = z\}
\]  
(B2)

The fact that \( Z_{1,t} \) and \( Z_{2,t} \) are independent and normally distributed for fixed \( t \) yields
\[
P\{||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_{0} = z\} = \phi\left(\frac{(b+y) \cos w - x_{1}R_{1}(i\Delta)}{\sqrt{1-R_{1}^{2}(i\Delta)}}\right) \phi\left(\frac{(b+y) \sin w - x_{2}R_{2}(i\Delta)}{\sqrt{1-R_{2}^{2}(i\Delta)}}\right) \frac{(b+y) \, dy \, dw}{\sqrt{1-R_{1}^{2}(i\Delta)\sqrt{1-R_{2}^{2}(i\Delta)}}}
\]

If we expand the above around \( t^{*} \) and \( w^{*} \) where \( t^{*} \) and \( w^{*} \) are defined in the statement of the lemma, we get
\[
P\{||Z_{i\Delta}|| \in b + dy, \omega_{i\Delta} \in dw \mid Z_{0} = z\} \approx \frac{be^{-\frac{1}{2}b^{2} - ||z||^{2}}e^{-\frac{1}{2}b^{2}[t+1+(w-w^{*})^2]^{2}}} {2\pi\sqrt{1-R_{1}^{2}(i\Delta)\sqrt{1-R_{2}^{2}(i\Delta)}}} e^{-by}dydw,
\]  
(B3)

where
\[
a_{1} = \frac{\dot{R}_{1}^{2}(t^{*}) \cos^{2} w^{*} + \dot{R}_{2}^{2}(t^{*}) \sin^{2} w^{*}}{1 - R_{1}^{2}(t^{*}) + \frac{\dot{R}_{2}^{2}(t^{*})^{2}}{1 - R_{2}^{2}(t^{*})}},
\]
\[
a_{2} = \frac{\sin^{2} w^{*}}{1 - R_{1}^{2}(t^{*})} + \frac{\cos^{2} w^{*}}{1 - R_{2}^{2}(t^{*})} - 1,
\]
\[
a_{3} = \cos w^{*} \sin w^{*} \left[\frac{\dot{R}_{1}^{2}(t^{*}) \dot{R}_{2}^{2}(t^{*})}{1 - R_{1}^{2}(t^{*})} + \frac{\dot{R}_{1}^{2}(t^{*})R_{2}^{2}(t^{*})}{1 - R_{2}^{2}(t^{*})}\right].
\]

As was shown in Appendix A,
\[
P\{||Z_{(i+j)\Delta}|| < b \; \forall j \in D_{l} \mid ||Z_{i\Delta}|| = b + y, \omega_{i\Delta} = w, Z_{0} = z\} \approx P_{-\mu,\sigma}\{\max_{0\leq(i+j)\Delta\leq l} S_{j} < -y\},
\]  
(B4)
where $S_j$ is the sum of $j$ independent normal random variables with mean and variance $-\Delta b(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$ and $2\Delta(\beta_1 \cos^2 w + \beta_2 \sin^2 w)$, respectively.

Substituting (B3) and (B4) into (B2) we obtain

\[
P\left\{ \max_t \|Z_t\| \geq b \mid Z_0 = z \right\} 
\approx \sum_{i=0}^{1/\Delta} \int_{\pi}^{\pi} \frac{\beta e^{-\frac{1}{2}(b^2 - ||x||^2)} e^{-\frac{1}{2}b^2[(i\Delta - t^*)^2a_1 + (w-w^*)^2a_2 + 2(w-w^*)(i\Delta - t^*)a_3]}}{2\pi \sqrt{1-R_1^2(i\Delta)} \sqrt{1-R_2^2(i\Delta)}} \times \int_0^\infty e^{-by} P_{\mu, \sigma} \left\{ \min_{0 \leq (i+j)\Delta < t} S_j > y \right\} dy dw
\]

\[
\approx \frac{\beta \exp\left(-\frac{1}{2}(b^2 - ||x||^2)\right)}{R_1(t^*) R_2(t^*) \cos^2 w^* + R_1(t^*) R_2(t^*) \sin^2 w^*} \nu(b[2\beta\Delta]^{1/2}).
\]

To obtain the last line we use Corollary 8.45 of SIEGMUND (1985) to evaluate the inner integral. The summation on $i$, is approximately an integral in the variable $i\Delta$. Because $b >> 0$ the bivariate normal density in $i\Delta$ and $w$ behaves like a delta function concentrating at $t^*, w^*$.

**APPENDIX C**

**Likelihood ratio test for sib pairs**

The tests for sib pairs discussed by HOLMANS (1993) and FARAWAY (1993) have a number of features in common with the two degree of freedom test for linkage in an intercross design. Since they can be studied by the methods of this paper, we discuss them briefly here. Let $X_{i,t}$ be the number out of $N$ sib pairs that share $i$ alleles identical by descent ($i = 0, 1, 2$) at locus $t$. For one pair, on a chromosome containing a trait locus at $\tau$, the probability that $i$ alleles are shared identical by descent at $t$ equals $[1 + (\alpha + \delta) \exp(-\beta|t - \tau|) + \delta \exp(-2\beta|t - \tau|)]/4$, $[1 - \delta \exp(-2\beta|t - \tau|)]/2$, and $[1 - (\alpha + \delta) \exp(-\beta|t - \tau|) + \delta \exp(-2\beta|t - \tau|)]/4$ for $i = 2, 1, 0$, respectively, where $\beta = 0.04$. Genetic interpretations of $\alpha$ and $\delta$ can be derived in terms of the recurrence risk to sibs and to offspring or the additive and dominance variance of the penetrances (cf. RISCH 1990a,b). The exact expressions do not concern us here, except to note that for a wide variety of models involving multiple loci and interactions among
the loci $0 \leq \delta < \alpha$. HOLMANS (1994) calls the likelihood ratio test the "possible triangle test" to reflect this natural genetic constraint. TENG (1996) in her analysis of the HASEMAN-ELSTON (1972) method for detecting linkage of QTL’s in human genetics obtains a parameterization of that problem that leads to effectively the same constraints.

To describe a Gaussian approximation to the likelihood ratio statistic, it is convenient to introduce the notation

$$U_{1,t} = -2(X_{1,t} - N/2)/N^{1/2}, \quad U_{2,t} = 2^{1/2}(X_{2,t} - X_{0,t})/N^{1/2}.$$ 

Direct calculations show that on unlinked chromosomes these two processes have expectation 0, variance 1, and are uncorrelated. Also

$$\text{Cov}[U_{1,s}, U_{1,t}] = \exp[-\beta_1|t-s|], \quad \text{Cov}[U_{2,s}, U_{2,t}] = \exp[-\beta_2|t-s|],$$

where $\beta_1 = 0.08, \beta_2 = 0.04$. On a linked chromosome having a single trait locus at $r$,

$$EU_{1,t} = \mu_1 \exp[-\beta_1|t-r|], \quad EU_{2,t} = \mu_2 \exp[-\beta_2|t-r|], \quad (C1)$$

where $\mu_1 = N^{1/2}\delta, \mu_2 = (N/2)^{1/2}(\alpha + \delta)$. The genetic constraint $0 \leq \delta \leq \alpha$ is equivalent to $0 \leq 2^{1/2}\mu_1 \leq \mu_2$. Thus the triangular constraint noted by Holmans becomes the constraint that $\mu = (\mu_1, \mu_2)$ lie in a wedge in the first quadrant of the $xy$ plane defined by the lines $y = 2^{1/2}x$ and $x = 0$. In the case $\delta = 0$, i.e., an additive model of inheritance, $U_1$ carries no information, so a test to detect linkage is based directly on the maximum over all loci $t$ of

$$U_{2,t}. \quad (C2)$$

For a rare recessive trait, where $\delta \approx \alpha$, the appropriate test is based on the maximum over $t$ of

$$4[X_{2,t} - N/4]/(3N)^{1/2} = [U_{1,t} + 2^{1/2}U_{2,t}]/3^{1/2}. \quad (C3)$$

The statistic (C3) is the projection of the vector $(U_{1,t}, U_{2,t})$ along the line $y = 2^{1/2}x$, which makes an angle $\tan^{-1} 2^{1/2}$ with the positive $x$ axis in the $xy$ plane. The statistic (C2) is obviously the projection of $(U_{1,t}, U_{2,t})$ along the $y$ axis, which makes an angle of $\pi/2$ with the positive $x$ axis.
The log likelihood function is

\[ \mu_1 U_{1,r} + \mu_2 U_{2,r} - (\mu_1^2 + \mu_2^2)/2. \]

If there were no constraints on \( \mu_1, \mu_2 \), the likelihood ratio statistic at the locus \( t \) would be obtained by maximizing this expression with respect to \( \mu_1, \mu_2 \), which yields

\[ [U_{1,t}^2 + U_{2,t}^2]^{1/2}. \]  \hspace{1cm} (C4)

(The square root is introduced to keep (C4) in the same units as (C2) and (C3).)

To incorporate the constraints we use (C4) if the point \( U_{1,t}, U_{2,t} \) lies in the wedge defined by the lines \( y = 2^{1/2}x \) and \( x = 0 \). If the point does not lie in this wedge, we use the larger of (C2) and (C3). In effect the likelihood ratio test incorporating the constraints is based on (C4) unless the data tell us that the mode of inheritance appears to be purely additive or purely recessive; in these extreme cases we use the statistic appropriate for the apparent mode of inheritance.

The false positive rate of the likelihood ratio test is the probability, computed under the assumption \( \alpha = \delta = 0 \) that the statistic described above exceeds the detection threshold \( b \) somewhere throughout the genome. This probability can be evaluated approximately by an argument along the lines given in Appendix A, although the constraint relating \( \alpha \) and \( \delta \) makes the details more complicated. For dense markers and a genome of length \( L \) the approximation is

\[ 1 - \exp[-L \exp(-b^2/2)(C b^2/(2\pi) + b(5\beta_2/6 + \beta_1/6)/(2\pi)^{1/2})], \]

where

\[ C = \int_{\tan^{-1}2^{1/2}}^{\pi/2} (\beta_2 \sin^2 \omega + \beta_1 \cos^2 \omega) d\omega. \]

The term in the exponential involving \( b^2 \) accounts for the probability that (C4) exceeds \( b \) at a point \( t \) where \( (U_{1,t}, U_{2,t}) \) lies inside the wedge, while the terms involving \( b \) account for the probability that \( U_{2,t} \) or \( [U_{1,t} + 2^{1/2}U_{2,t}]/3^{1/2} \) exceeds \( b \) for some value of \( t \) where the two dimensional process \( (U_{1,t}, U_{2,t}) \) is outside the wedge.

The integral is easily shown to equal \((\beta_1 + \beta_2)(\pi/2 - \tan^{-1}2^{1/2})/2 - (\beta_1 - \beta_2)/(3 \times 2^{1/2}) = 0.0275\) for \( \beta_1 = 0.08, \beta_2 = 0.04 \). For a genome of genetic length 3000 cM a
threshold of \( b = 4.3 \) yields a false positive error rate of approximately 0.05. This value has been cited by LANDER and SCHORK (1994).

To approximate the power of the likelihood ratio test we let \( \xi = (\mu_1^2 + \mu_2^2)^{1/2} \) denote the distance of the point \( \mu = (\mu_1, \mu_2) \) from the origin. The point \( \mu \) lies inside the wedge defined by the angles \( \tan^{-1}2^{1/2} \) and \( \pi/2 \) measured from the positive \( x \) axis; and a slightly different approximation to the power is appropriate depending on whether the point is strictly inside the wedge or on one of the edges. In the former case the power is approximately

\[
1 - \Phi(b - \xi) + \phi(b - \xi)\{1/(2\xi) + (b/\xi)^{1/2}[2/\xi - 1/(b + \xi)]\},
\]

(C5)

while in the latter it is

\[
1 - \Phi(b - \xi) + \phi(b - \xi)\{1/4\xi + [(b/\xi)^{1/2} + 1][1/\xi - 1/(2(b + \xi))]\}.
\]

(C6)

These approximations can be derived by arguments similar to those used above. See also FEINGOLD et al. (1993) and SIEGMUND (1985).

Because of the constraint \( 0 \leq \delta < \alpha \) that restricts the two dimensional parameter to lie in the relatively narrow wedge indicated above, one suspects that it may be possible to choose an appropriate direction somewhere in the middle of the wedge and construct a one dimensional statistic that performs about as well as the two dimensional likelihood ratio statistic. See FEINGOLD and SIEGMUND (1997) for a complete discussion.