Conference on Applied Statistics in Agriculture 1999 - 11th Annual Conference Proceedings

A COMPUTATIONALLY EFFICIENT METHOD FOR DETERMINING SIGNIFICANCE IN INTERVAL MAPPING OF QUANTITATIVE TRAIT LOCI

Dan Nettleton

Recommended Citation
Nettleton, Dan (1999). "A COMPUTATIONALLY EFFICIENT METHOD FOR DETERMINING SIGNIFICANCE IN INTERVAL MAPPING OF QUANTITATIVE TRAIT LOCI," *Conference on Applied Statistics in Agriculture*. https://doi.org/10.4148/2475-7772.1265

A COMPUTATIONALLY EFFICIENT METHOD FOR DETERMINING SIGNIFICANCE IN INTERVAL MAPPING OF QUANTITATIVE TRAIT LOCI

Dan Nettleton
917 Oldfather Hall
Department of Mathematics and Statistics
University of Nebraska-Lincoln 68588-0323

ABSTRACT

This paper provides a brief introduction to the mapping of quantitative trait loci (QTL). An example on mapping QTL for root thickness in rice is presented to illustrate popular statistical methods used in QTL mapping. Interval mapping is used in conjunction with permutation testing techniques to detect significant associations between genetic positions and quantitative traits while controlling overall type I error rate. A review of a recent technique that can greatly reduce the computational expense of permutation testing in QTL mapping is discussed. Theory is provided for an extension of recent results that may lead to more powerful methods of QTL mapping through permutation testing.

1. Introduction

Quantitative trait loci (QTL) are regions of the genome that affect quantitative characteristics of plants and animals. Researchers have attempted to locate QTL for a variety of traits in many organisms in recent years. A few examples, arbitrarily selected from many, include efforts to map the genetic regions affecting fruit weight in tomatoes (Paterson et al., 1988, 1991), body fat in pigs (Andersson et al., 1994), sugar-cane-borer resistance in maize (Bohn et al., 1996), stinging behavior and body size in honey bees (Hunt, 1998), milk production in dairy cattle (Arranz, Coppieters, and Georges, 1998), and IQ in humans (Plomin, McClearn, and Smith, 1994). This paper will use the work of Champoux et al. (1995) on mapping QTL for root thickness in rice as a context for discussing some important computational and statistical issues in QTL mapping. The focus will be on permutation testing – developed by Fisher (1935) and first applied to QTL mapping by Churchill and Doerge (1994).

Permutation testing is a computationally intensive method that provides a means of determining the significance of test statistics used in various QTL mapping procedures. This intuitive method is popular in QTL mapping for many reasons. It is, for example, robust against departures from standard QTL modelling assumptions that are often in doubt but difficult to check. It provides natural controls on overall type I error rate when multiple dependent hypothesis tests are considered simultaneously. Permutation testing "automatically reflects the characteristics of the particular experiment to which it is applied" (Churchill and Doerge, 1994). Unfortunately, permutation testing can be very computationally expensive. Nettleton and Doerge (2000) presented a method for minimizing the computational cost of permutation testing in QTL mapping studies. This method is reviewed in Section 5.

Section 4 describes the use of permutation testing in interval mapping of QTL. Interval mapping, a statistical method for locating QTL developed by Lander and Botstein (1989), is widely used for initial genome scans. This method allows all positions throughout the genome to be tested for association with the quantitative trait of interest. It is an extension of single-marker-analysis techniques (Sax,1923; Soller, Brody, and Genizi, 1976) that test for association at only the limited number of genetic positions for which explicit genotype information is available. In interval mapping, test statistics known as LOD scores are computed for closely spaced positions throughout the genome. Large LOD scores are evidence of association between the corresponding genetic positions and the trait of interest. The LOD scores are typically plotted against genetic position to provide a visual assessment of the nature of association between the genome and the trait. Permutation testing can be used to determine the statistical significance of each LOD score. More detail on interval mapping is presented in Section 3.

Figure 1 shows plots of LOD score versus genetic position for 4 simulated chromosomes. Figure 2 shows plots of LOD score versus genetic position for the 12 chromosomes of rice using data on root thickness from Champoux et al. (1995). This paper will illustrate how to make efficient use of permutation testing to determine the significance of each LOD score in Figures 1 and 2 using the method developed by Nettleton and Doerge (2000). The simulated data is analyzed in Section 6. Analysis of the rice data is presented in Section 7. An extension to the work of Nettleton and Doerge (2000) will be discussed in Section 8. The next Section provides background information concerning the study of Champoux et al. (1995) and contains a minimal discussion of genetic issues necessary for understanding the remainder of the paper.

2. Mapping QTL for Root Thickness in Rice

Champoux et al. (1995) discuss evidence suggesting that rice plants with thick roots generally have better drought tolerance than plants with thin root systems. Thus, the development of varieties of rice with thick roots is desired for regions where rice is grown without the guarantee of sufficient soil moisture. Mapping QTL for root thickness is a first step toward understanding the genetic architecture of root thickness and eventually developing lines of rice with better drought tolerance characteristics.

Champoux et al. (1995) studied 203 recombinant inbred rice lines derived from a cross between two rice lines, *indica* cultivar CO39 and *japonica* cultivar Moroberekan. The CO39 cultivar has a thin root system and is susceptible to drought while Moroberekan has a thick root system and is drought resistant. The genetic make up of any given recombinant inbred line is a unique mixture of genetic material from CO39 and Moroberekan, i.e., genetic material from the CO39 parent is interspersed with genetic material from the Moroberekan parent throughout the genome. At any given locus (genetic position), some of the 203 recombinant inbred lines will have genetic material from the CO39 parent and all other lines will have genetic material from the Moroberekan parent. A locus exhibits an association with root thickness if the recombinant inbred lines with CO39 genetic material at the locus have significantly different root thickness than the lines with Moroberekan genetic material at the locus. Because of the wide difference in root thickness between the parental lines, at least one and probably several genetic positions are expected to exhibit an association with root thickness in the sample of 203 recombinant inbred lines.

Identifying the loci that are significantly associated with the trait is complicated by the fact that the type of genetic material (genotype) is observed only at certain loci known as markers. Fortunately, the genotypes of markers, together with a genetic map, provide enough information about genotypes at non-marker loci to allow tests of association at all loci. The details of genetic maps will not be discussed in this paper. To understand subsequent sections of this paper, it is necessary to know that:

1. Only the genotypes of marker loci are observed.

2. There is a spatial dependence in genotype across any chromosome.

3. This dependence can be exploited to determine the conditional probability of each parental genotype at any particular position, given the genotypes of the markers flanking that position.

4. Genetic distances are often reported in centiMorgans (cM). The dependence of genotypes at genetic positions increases as the distance in cM between the positions decreases.

The genotypes of the 203 recombinant inbred lines were recorded at 123 markers spread over all 12 chromosomes. This information, along with a measurement of root thickness (in micrometers) for each line, constitute the data available for QTL mapping. The next section describes how interval mapping can be used to identify associations between loci and trait with such data.

3. Interval Mapping

For any particular locus $\mathbf{X}$ suppose that

$$Y_i \sim N(\mu_a^{\mathbf{X}}, \sigma^2) \text{ when } X_i = a \text{ and } Y_i \sim N(\mu_b^{\mathbf{X}}, \sigma^2) \text{ when } X_i = b, \tag{1}$$

where $\mu_a^{\mathbf{X}}$, $\mu_b^{\mathbf{X}}$, and σ^2 are unknown parameters, Y_i denotes the quantitative trait measurement associated with the ith line, and X_i denotes the genotype of the ith line at locus $\mathbf{X}$ ($i = 1, \ldots, n$). For the rice data discussed in the previous section, $n = 203$, Y_i is the root thickness measurement for the ith line, and we could choose a to represent the CO39 genotype and b to represent the Moroberekan genotype. The parameter $\mu_a^{\mathbf{X}}$ ($\mu_b^{\mathbf{X}}$) would represent the mean root thickness for lines with CO39 (Moroberekan) genotype at locus $\mathbf{X}$. The parameter σ^2 represents the common root thickness variance within each of the two groups of lines.

A test of $H_0^{\mathbf{X}} : \mu_a^{\mathbf{X}} = \mu_b^{\mathbf{X}}$ against $H_1 : \mu_a^{\mathbf{X}} \neq \mu_b^{\mathbf{X}}$ can be used to test locus $\mathbf{X}$ for association with the trait of interest. The null hypothesis indicates that locus $\mathbf{X}$ is unassociated with the trait since the trait distributions are the same for lines with either genotype under the null hypothesis. Locus $\mathbf{X}$ is associated with the trait under the alternative hypothesis because the distribution of the trait depends on genotype at locus $\mathbf{X}$ according to the alternative hypothesis. Interval mapping consists of conducting multiple tests of this form at loci closely spaced throughout the entire genome.

The test statistic typically used in interval mapping is the negative base-ten log of the likelihood ratio. This quantity is called a LOD score and is related to the usual likelihood ratio test statistic

$(-2\ln\Lambda)$ as follows.

$$\text{LOD} = \log_{10}\frac{\sup_{H_1^X} L(\mu_c^X, \mu_m^X, \sigma^2)}{\sup_{H_0^X} L(\mu_c^X, \mu_m^X, \sigma^2)} = -2\ln\Lambda/\ln 100,$$

where $L(\mu_c^X, \mu_m^X, \sigma^2)$ denotes the likelihood function. The LOD score is equivalent to a two-sample t statistic when $\mathbf{X}$ is a marker locus. The likelihood becomes a special normal mixture likelihood when $\mathbf{X}$ is not a marker. Maximization of this mixture likelihood under the alternative hypothesis requires the EM algorithm or a similar iterative procedure. Details of the computation of LOD scores via the EM algorithm are provided by Carbonell, et al. (1992) and Nettleton and Praestgaard (1998).

Nettleton (1999) studied the likelihood ratio test of $H_0^{\mathbf{X}}$ versus $H_1^{\mathbf{X}}$ for an arbitrary number of genotype-class means. The asymptotic null distribution of the likelihood ratio test was shown to be chi-square with degrees of freedom equal to one less than the number of genotype-class means. Thus, the asymptotic distribution of the LOD score for the case considered here is that of a single-degree-of-freedom chi-square random variable divided by $\ln 100$ (or, equivalently, multiplied by $1/2\log_{10} e$) as claimed by Lander and Botstein (1989). This asymptotic result is useful for evaluating the significance of a single locus. In interval mapping, however, we are interested in determining the significance of hundreds of loci using hundreds of correlated LOD scores which means that issues of multiple testing must be addressed. Using the chi-square critical value for each of hundreds of tests will almost certainly lead to an unacceptable number of type I errors (false declarations of association between locus and trait).

Let $\mathbf{X}_1, \ldots, \mathbf{X}_p$ denote the loci at which LOD scores are computed in an interval mapping procedure. Let LOD_j denote the LOD score computed at locus $\mathbf{X}_j$ $(j = 1, \ldots, p)$. To control the rate at which the interval mapping procedure will yield one or more false positive QTL declarations, we seek constants $q_1, \ldots, q_p$ such that

$$P_{H_0^*}[\text{LOD}_1 > q_1 \text{ or } \text{LOD}_2 > q_2 \text{ or } \cdots \text{ or } \text{LOD}_p > q_p] \leq \alpha,$$

where $H_0^* : \mu_a^{\mathbf{X}_j} = \mu_b^{\mathbf{X}_j}$ for all $j = 1, \ldots, p$. The overall type I error rate will be no larger than α if a locus $\mathbf{X}_j$ is declared associated with the trait only when $\text{LOD}_j > q_j$. It seems reasonable to require $q_1 = \cdots = q_p$ because the asymptotic null distribution of each LOD score is the same (chi-square with one degree of freedom for the recombinant inbred lines studied here). Hence, we seek a constant q such that

$$P_{H_0^*}[\max_{1\leq j\leq p} \text{LOD}_j > q] \leq \alpha.$$

It is difficult to determine the asymptotic distribution of $\max_{1\leq j\leq p}\text{LOD}_j$ under H_0^* because of the complex dependence structure among the LOD scores. Lander and Botstein (1989) offered an approximate value of q for the *dense-map* case in which the spacing of consecutive markers approaches zero. Rebaï, Goffinet, and Mangin (1994) and Dupuis (1994) provided other analytical methods for estimating q. Churchill and Doerge (1994) offered permutation testing as an alternative means of estimating the critical value q that is free of many assumptions required by asymptotic approaches. Doerge (1998) discussed some benefits of the permutation technique relative to approaches based on asymptotic approximations. The next section provides a brief explanation of permutation testing in the context of interval mapping.

4. Permutation Testing in Interval Mapping

The distribution of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ under H_0^* can be approximated by the permutation distribution of $\max_{1\leq j\leq p} \mathrm{LOD}_j$. The permutation distribution of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ can be determined, in theory, by computing the value of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ for $n!$ permuted data sets. A permuted data set is obtained by randomly assigning the observed trait values $Y_1, \ldots, Y_n$ to the n lines while holding the genotype information for each line fixed at the observed values. There are $n!$ permuted data sets since there are $n!$ ways to assign the observed trait values to the lines. One of the $n!$ permuted data sets is the original data set where each trait value is assigned to its own line. We may choose q to be the $1-\alpha$ quantile of the $n!$ values of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ that arise from the analysis of the $n!$ permuted data sets. Specifically, let $q = M_{\lfloor n!\alpha \rfloor +1}$ where $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x and $M_1 \geq M_2 \geq \cdots \geq M_{n!}$ denote the values of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ computed from the $n!$ permuted data sets and ordered from largest to smallest.

Choosing q in this manner yields a testing procedure which maintains appropriate type I error rate for arbitrary sample size and marker spacings. The normality assumption in (1) can be relaxed to

$$P[Y_i \leq y] = F(y) \text{ when } X_i = a \text{ and } P[Y_i \leq y] = F(y - \delta^{\mathbf{X}}) \text{ when } X_i = b,$$

where F is any distribution function and $\delta^{\mathbf{X}}$ is an unknown parameter that equals 0 under the null hypothesis. The parameter $\delta^{\mathbf{X}}$ represents the shift in distribution between the two genotype classes at locus $\mathbf{X}$. For the special case of normal distributions with common scale assumed in (1), $\delta^{\mathbf{X}} = \mu_b^{\mathbf{X}} - \mu_a^{\mathbf{X}}$. Churchill and Doerge (1994) discussed other desirable theoretical properties of the permutation testing procedure.

To see that the permutation test will indeed maintain the proper type I error rate, first note that any permutation of the trait values is equally likely under the null hypothesis. Thus, given the observed trait values and genotype information associated with the lines, the conditional probability that the observed value of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ will exceed q is given by the proportion of permuted data sets for which $\max_{1\leq j\leq p} \mathrm{LOD}_j > q$. This proportion is less than or equal to α by the definition of q. Multiplying the conditional probability by the joint density of the trait values and genotype information under H_0^* and integrating over all possible trait values and genotype information yields $P_{H_0^*}[\max_{1\leq j\leq p} \mathrm{LOD}_j > q] \leq \alpha$.

Unfortunately, it is typically infeasible to determine q in the manner described above. Recall that LOD scores for non-marker loci must be computed using numerical methods. Thus, it can be quite time consuming to compute LOD scores at loci $\mathbf{X}_1, \ldots, \mathbf{X}_p$ for a single data set, and it is almost always impossible to compute these LOD scores for all $n!$ permuted data sets. Instead, $\max_{1\leq j\leq p} \mathrm{LOD}_j$ is computed for a simple random sample of N permuted data sets. The $1-\alpha$ quantile of the N sampled values of $\max_{1\leq j\leq p} \mathrm{LOD}_j$, $\hat{q}$, serves as an estimate of q.

5. Determining Permutation Sample Size

Nettleton and Doerge (2000) developed methods of accounting for the variability associated with sampling from all possible permutations when using interval mapping and permutation testing to detect QTL. They describe confidence intervals for permutation p-values and critical values and explain how to use such confidence intervals to dynamically determine an appropriate value of N. This section presents some of the main results in Nettleton and Doerge (2000) that will be demonstrated in Sections 6 and 7 for the simulated data set and the rice data set discussed previously. See Nettleton and Doerge (2000) for detail on the derivation of each result.

A permutation p-value is defined, for any locus $\mathbf{X}$, as the proportion of permuted data sets for which $\max_{1\leq j\leq p} \mathrm{LOD}_j$ matches or exceeds the observed LOD score at locus $\mathbf{X}$. This p-value provides a measure of significance that accounts for testing multiple loci over the entire genome. Nettleton and Doerge call such a p-value an *experimentwise* permutation p-value to distinguish it from a p-value that is not adjusted for multiple testing. Let $p^{\mathbf{X}}$ denote the permutation p-value at locus $\mathbf{X}$, and let $\hat{p}^{\mathbf{X}}$ denote the proportion of the N sampled permuted data sets for which $\max_{1\leq j\leq p} \mathrm{LOD}_j$ matches or exceeds the observed LOD score at locus $\mathbf{X}$. The proportion $\hat{p}^{\mathbf{X}}$ is an unbiased estimate of $p^{\mathbf{X}}$, and a $100(1-\gamma)\%$ confidence interval for $p^{\mathbf{X}}$ is

$$(\hat{p}^{\mathbf{X}} - \Phi^{-1}(1-\frac{\gamma}{2})\sqrt{\hat{p}^{\mathbf{X}}(1-\hat{p}^{\mathbf{X}})/N},\ \ \hat{p}^{\mathbf{X}} + \Phi^{-1}(1-\frac{\gamma}{2})\sqrt{\hat{p}^{\mathbf{X}}(1-\hat{p}^{\mathbf{X}})/N}), \tag{2}$$

where Φ^{-1} denote the inverse of the standard normal cumulative distribution function. The interval is based on the normal approximation to the binomial distribution and will have coverage close to nominal when $Np^{\mathbf{X}} \geq 5$. Such a confidence interval is useful for reporting a measure of significance that accounts for multiple testing and the effect of sampling from $n!$ data permutations.

A confidence interval for the permutation critical value q is also needed to reflect variation in the estimated critical value $\hat{q}$ discussed in Section 4. Let $M_1 \geq M_2 \geq \cdots \geq M_N$ denote the values of $\max_{1\leq j\leq p} \mathrm{LOD}_j$ computed from the N sampled permuted data sets and ordered from largest to smallest. An approximate $100(1-\gamma)\%$ confidence interval for the level-α critical value q is $[M_L, M_U]$ where

$$L = \lfloor N\alpha + \Phi^{-1}(1-\frac{\gamma}{2})\sqrt{N(1-\alpha)\alpha}\rfloor \text{ and } U = \lfloor N\alpha - \Phi^{-1}(1-\frac{\gamma}{2})\sqrt{N(1-\alpha)\alpha}\rfloor.$$

This interval is based on the normal approximation to the binomial distribution and will provide appropriate coverage when $N\alpha \geq 5$.

Nettleton and Doerge (2000) recommended using this interval as a guide in determining an appropriate permutation sample size. Initially, a $100(1-\gamma)\%$ confidence interval for the α-level critical value is computed using at least $\lceil 5/\alpha \rceil$ permutations so that $N\alpha$ will be greater than or equal to 5. The LOD scores associated with peaks in the plot of LOD score versus genetic position are compared to this interval. The α-level significance status of each peak can be determined if each corresponding LOD score falls outside the confidence interval. Those LOD scores that fall above the upper endpoint of the interval can be judged significant at the α level while those falling below the lower endpoint of the confidence interval should not be considered statistically significant. These significance decisions will be the same as the decisions that would be made if all $n!$ data permutations could be considered – provided the $100(1-\gamma)\%$ confidence interval contains the exact permutation critical value q. Additional permutations should be considered when one or more of the LOD scores associated with peaks in the plot of LOD score versus genetic position fall inside the confidence interval for the critical value. Nettleton and Doerge recommended considering additional permutations until all the LOD scores associated with peaks fall outside the confidence interval for the critical value or until time and/or computational limits are reached.

Without considering a confidence interval for the permutation critical value, prespecified values of N may be too small to determine the significance of peaks at the desired level or, on the other hand, may be much larger than necessary – resulting in inefficient use of computing resources. The examples in Sections 6 and 7 will show that the significance of most peaks can be resolved using relatively few permutations when the procedure of Nettleton and Doerge (2000) is implemented. There can be considerable time savings over permutation procedures using earlier recommendations in the literature which called for the analysis of at least 1000 data permutations. The significance of some loci, however, may be in question even after considering as many as 2000 permuted data sets. The work of Nettleton and Doerge (2000) provides a means of estimating significance when the computational demands are too great to determine significance at a specified level.

6. Analysis of a Simulated Data Set

A sample of 100 lines with four chromosomes and four QTL were simulated. The trait value of the i^{th} line was determined using $Y_i = 2.50Q_{i1} + 0.75Q_{i2} + 1.00Q_{i3} + 1.00Q_{i4} + \varepsilon_i$ where ε_i is a standard normal environmental error term, with additive effects defined as $Q_{ij} = 1$ if the i^{th} line has genotype a at the j^{th} QTL and 0 otherwise. Chromosomes 1, 2, 3 and 4 are 102, 130, 169, and 70 cM in length, respectively. A total of 46 markers are arbitrarily positioned throughout the genome – 11 on chromosome 1, 13 on chromosome 2, 15 on chromosome 3, and 7 on chromosome 4. QTL 1 is 62 cM from the left end of chromosome 1. QTL 2 is 44 cM from the left end of chromosome 2. Chromosome 3 has QTL 3 and QTL 4 at 17 and 147 cM, respectively from the left. No QTL are present on chromosome 4.

LOD scores were computed every 1 cM on each chromosome as described in Section 3. The resulting plot of LOD score versus genetic position for each chromosome is depicted in Figure 1. Major peaks occur at 66 cM on chromosome 1 (LOD = 12.808), 24 cM on chromosome 2 (LOD=2.894), and at 22 and 152 cM on chromosome 3 (LOD=1.640 and LOD=1.013). Examination of the plots in Figure 1 reveals that the only other loci with large LOD scores are clearly linked to one of the four loci above. Consequently, the subsequent permutation analyses will focus on these four loci only.

Only 53 data permutations were required to determine the experimentwise significance of the four loci at the 0.10 level. An approximate 95% confidence interval for the 0.10-level permutation critical value was determined to be $[1.715, 2.346]$ using the methods outlined previously. The peaks on chromosomes 1 and 2 are judged significant at the 0.10 level because 2.346 is less than the LOD scores 12.808 and 2.894. The peaks on chromosome 3, on the other hand, fall short of experimentwise significance at level 0.10 since 1.640 and 1.012 are less than 1.715. If the goal is to determine significance at level 0.05, 110 data permutations are sufficient in this case. An approximate 95% confidence interval for the 0.05-level permutation critical value was determined to be $[1.940, 2.659]$. The peaks on chromosomes 1 and 2 are thus significant at the 0.05 level.

Even 1000 data permutations are insufficient to determine the significance status of all four positions when considering 0.01-level experimentwise significance. An approximate 95% confidence interval for the 0.01-level critical value was determined to be $[2.801, 3.530]$ using 1,000 randomly selected data permutations. The locus on chromosome 1 is clearly significant while the loci on chromosome 3 are clearly not significant. The status of the point on chromosome 2, however, is uncertain. Considering 94 additional randomly chosen data permutations yielded $[2.897, 3.530]$ as an approximate 95% confidence interval for the 0.01-level critical value. Hence, the second locus is judged insignificant at the 0.01 significance level.

Confidence intervals for permutation p-values can be computed for the loci of interest on chromosomes 2 and 3 using equation (2). Point estimates and approximate 95% confidence intervals are displayed in Table 1. The p-value estimates are based on the 1094 data permutations used to estimate the 0.01-level experimentwise threshold. No interval is provided for the first locus since its test statistic was exceeded by none of the 1094 values of $\max \mathrm{LOD}$. If the true permutation p-value is actually 0.01 or bigger, the chance of estimating the p-value to be zero based on 1094 data permutations is extremely small (no larger than $0.99^{1094} = 0.0000168$). We can be quite confident that this locus is significantly linked to the trait.

Note that the confidence interval for the second locus includes 0.01, suggesting that this position may be significant at the 0.01 level. Examination of the critical-value confidence interval suggested insignificance at the 0.01 level using the same 1094 data permutations. Such minor discrepancies are possible when LOD scores are near the borderline. If we consider 334 additional data permutations, the confidence interval for the p-value becomes $[0.0101, 0.0235]$, bringing it into agreement with the critical-value-based analysis.

7. Analysis of the Rice Data

Nettleton and Doerge (2000) conducted interval mapping for the rice data of Champoux et al. (1995). LOD scores were computed at 2 cM increments across each of the 12 chromosomes. The plots of LOD score versus genetic position are provided in Figure 2. Nettleton and Doerge showed that only 142 permutations were needed to establish the 0.05-level significance status for all peaks. The 0.05-level critical value estimate and the corresponding confidence interval are 2.934 and $[2.517, 3.569]$, respectively. The middle peak on chromosome 7 (LOD $= 2.420$) is the only peak that falls short of significance at the 0.05 level. The other peaks have LOD scores that exceed 3.569. Note that there is very little overlap between the confidence intervals for the 0.05-level critical value in the rice data and the simulated data discussed in Section 6. The critical value appears less stringent for the simulated data. There are fewer opportunities for type I error in the simulated data because of the small genome size relative to rice. Thus, a smaller critical value for the simulated data seems appropriate. Such issues are automatically accounted for by the permutation testing procedure.

Determining significance of all peaks in Figure 2 at the 0.01 level is more computationally challenging. At least 500 permutations should be considered according to the method discussed in Section 5. The 95% confidence interval for the 0.01-level critical value based on 500 permutations is $[3.355, 5.793]$. Most of the peaks that were declared significant at the 0.05 level can be judged significant at the 0.01 level after considering only 500 permutations. The significance status of some peaks on chromosomes 2, 8, and 12 are in doubt based on this interval because their corresponding LOD scores fall between 3.355 and 5.793.

An additional 1500 permutations were considered, bringing the total to $N = 2000$. The corresponding 95% confidence interval for the 0.01-level critical value is $[3.627, 4.348]$. All peaks fall below 3.627 or above 4.348 except the second peak on chromosome 2 (LOD $= 3.875$) and the peak 4 cM from the left end of chromosome 12 (LOD $= 3.665$). (It may be reasonable to consider this latter peak as part of the larger peak occurring at 12 cM from the left of chromosome 12, but it will be considered as a separate peak here for illustration purposes.) Because these LOD scores fall between 3.627 and 4.348, their significance at the 0.01 level is uncertain. A pair of Bonferroni-adjusted confidence intervals that will jointly contain the exact permutation p-values for the two loci with 95% confidence are $(0.0039, 0.0131)$ for the peak on chromosome 2 and $(0.0077, 0.0193)$ for the peak on chromosome 12. Both loci appear significant at the 0.02 level. Resolving the significance of these peaks at the 0.01 level would require more computing power or more patience.

8. An Extension

It may be possible to develop a procedure for controlling overall type I error rate that is less conservative than the methods presented in Sections 3 and 4. In Section 3 we sought constants $q_1, \ldots, q_p$ such that

$$P_{H_0^*}[\text{LOD}_1 > q_1 \text{ or } \text{LOD}_2 > q_2 \text{ or } \cdots \text{ or } \text{LOD}_p > q_p] \leq \alpha$$

to control the rate at which the interval mapping procedure would yield one or more false positive QTL declarations. In practice, only loci whose LOD scores are associated with peaks in the plots of LOD score versus genetic position are considered as candidate QTL. Hence, the rate of false positive QTL declarations will be appropriately controlled if we can find $q_1, \ldots, q_k$ such that

$$P_{H_0^*}[\text{LOD}_{(1)} > q_1 \text{ or } \text{LOD}_{(2)} > q_2 \text{ or } \cdots \text{ or } \text{LOD}_{(k)} > q_k] \leq \alpha, \tag{3}$$

where $\text{LOD}_{(1)} \geq \cdots \geq \text{LOD}_{(k)}$ denote the LOD scores associated with the peaks in the plots of LOD score versus genetic position.

Let $\text{LOD}_{(1)\ell} \geq \cdots \geq \text{LOD}_{(k)\ell}$ denote the LOD scores associated with the k largest peaks in the plots of LOD score versus genetic position for the ℓth permuted data set ($\ell = 1, \ldots, n!$). Let $\mathbf{V}_1$ denote the $n!$-dimensional vector whose ℓth component is $V_{1\ell} = \text{LOD}_{(1)\ell}$. Let $\mathbf{V}_2$ denote the $n!$-dimensional vector whose ℓth component is

$$V_{2\ell} = \begin{cases} \text{LOD}_{(2)\ell} & \text{if } \text{LOD}_{(1)} > V_{1\ell} \\ V_{1\ell} & \text{if } \text{LOD}_{(1)} \leq V_{1\ell} \end{cases} .$$

Similarly, for $j = 3, \ldots, k$; let $\mathbf{V}_j$ denote the $n!$-dimensional vector whose ℓth component is

$$V_{j\ell} = \begin{cases} \text{LOD}_{(j)\ell} & \text{if } \text{LOD}_{(j-1)} > V_{j-1,\ell} \\ V_{j-1,\ell} & \text{if } \text{LOD}_{(j-1)} \leq V_{j-1,\ell} \end{cases} .$$

THEOREM. *For $j = 1, \ldots k$; let $V_{j(1)} \geq V_{j(2)} \geq \cdots \geq V_{j(n!)}$ denote the components of $\mathbf{V}_j$ ordered from largest to smallest. Equation (3) will be satisfied with $q_j = V_{j(\lfloor n!\alpha \rfloor + 1)}$ for $j = 1, \ldots, k$.*

PROOF. By the definition of $\mathbf{V}_1, \ldots, \mathbf{V}_k$; $\text{LOD}_{(j)} \leq V_{j\ell}$ implies that $\text{LOD}_{(j')} \leq V_{j'\ell}$ for $1 \leq j \leq j' \leq k$. Thus, the number of components of $\mathbf{V}_j$ that match or exceed $\text{LOD}_{(j)}$ is less than or equal to the number of components of $\mathbf{V}_{j'}$ match or exceed $\text{LOD}_{(j')}$ for any $1 \leq j \leq j' \leq k$. By the definition of q_j, $\text{LOD}_{(j)} > q_j$ if and only if $\lfloor \alpha n! \rfloor$ or fewer components of $\mathbf{V}_j$ match or exceed $\text{LOD}_{(j)}$. It follows that $\text{LOD}_{(j')} > q_{j'}$ implies $\text{LOD}_{(j)} > q_j$ for all $1 \leq j \leq j' \leq k$. Hence, $\text{LOD}_{(1)} > q_1$ or $\text{LOD}_{(2)} > q_2$ or $\cdots$ or $\text{LOD}_{(k)} > q_k$ is equivalent to $\text{LOD}_{(1)} > q_1$. Note that $\text{LOD}_{(1)} = \max_{1 \leq j \leq p} \text{LOD}_j$ and $q_1 = q$ as defined in Section 4. Thus,

$$P_{H_0^*}[\text{LOD}_{(1)} > q_1] = P_{H_0^*}[\max_{1 \leq j \leq p} \text{LOD}_j > q] \leq \alpha,$$

as demonstrated in Section 4. □

The proposed procedure will give a generally less conservative assessment of the significance of secondary peaks because $q1 \geq q_2 \geq \cdots \geq q_k$. This is a simple consequence of the construction of the vectors $\mathbf{V}_1, \ldots, \mathbf{V}_k$ which ensures that $V_{j\ell} \geq V_{j'\ell}$ for all $\ell = 1, \ldots, n!$ and $1 \leq j \leq j' \leq k$. A simpler procedure would measure the significance of the locus associated with the jth largest peak by comparing $\text{LOD}_{(j)}$ to the $1 - \alpha$ quantile of the distribution of the permutation-replicated statistics $\text{LOD}_{(j)1}, \ldots, \text{LOD}_{(j)n!}$. However, this procedure would not guarantee the specified overall type I error rate and could lead to an undesirable situation in which the largest peak is judged insignificant while lesser peaks are declared significant.

There are a few obstacles to implementation of the proposed procedure. The same computational challenges discussed in Section 4 will not permit a direct determination of $q_1, \ldots, q_k$. Rather, $q_1, \ldots, q_k$ must be estimated as described in Sections 4 and 5. Simultaneous estimation issues arise because k quantiles are estimated instead of one. Some judgement is required in determining the k largest peaks in a given plot of LOD score versus genetic position. It is difficult to know, for example, whether a lesser peak is simply an artifact of close proximity to a larger peak. A algorithm that appropriately selects the k largest peaks for each permuted data set is needed. A conservative procedure would define a peak as any locus with a LOD score greater than the LOD scores of the two flanking positions.

9. Summary

Scientists studying a wide variety of organisms have attempted to locate regions of the genome that affect quantitative characteristics. The goal of many such attempts is to understand the genetic mechanisms responsible for quantitative variation with the hope of using this understanding to develop genetically superior lines for improved agricultural production. The study of Champoux et al. (1995) on root morphology in rice is one such example.

Interval mapping is a statistical technique that can be used to scan a genome in search of loci that are associated with a quantitative trait of interest. Permutation testing is a computationally intensive procedure that can be used in conjunction with interval mapping to evaluate the strength of evidence for association between locus and trait. Recent results of Nettleton and Doerge (2000) can significantly reduce the computational expense of permutation testing in QTL mapping problems and allow more researchers to take advantage of the benefits that permutation testing holds over competing methods for determining statistical significance.

It may be possible to develop less conservative means of determining significance in interval mapping via permutation testing. The previous section provides the theoretical background in support of a new permutation testing procedure that will maintain appropriate type I error rate despite potentially lower critical values for judging the significance of secondary peaks in plots of LOD score versus genetic position. Some obstacles must be overcome before such a procedure can be implemented, but if they can be overcome, there is promise for increased power with little added computational expense.

ACKNOWLEDGEMENT

The author acknowledges support from the National Research Initiative Competitive Grants Program of the U.S. Department of Agriculture, Award 98-35205-6709, and would like to thank an anonymous referee whose numerous suggestions improved the readability of this paper.

REFERENCES

Andersson, L., Haley, C. S., Ellegren, H., Knott, S. A., Johansson, M., Andersson, K., Andersson-Eklund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., Håkansson, J., Lundström, K. (1994). Genetic Mapping of Quantitative Trait Loci for Growth and Fatness in Pigs. *Science* **263**, 1771-1774.

Arranz, J. J., Coppieters, W., Georges, M. (1998). A QTL affecting milk yield and composition maps to bovine chromosome 20: a confirmation, *Animal genetics* **29**, 107-115.

Bohn, M., Khairallah, M. M., González-de-León, D., Hoisington, D. A., Utz, H. F., Deutsch, J. A., Jewell, D. C., Mihm, J. A., and Melchinger, A. E. (1996). QTL Mapping in Tropical Maize: I. Genomic Regions Affecting Leaf Feeding Resistance to Sugarcane Borer and Other Traits. *Crop Science* **36**, 1352-1361.

Carbonell, E. A., Gerig, T. M., Balansard, E., and Asins, M. J. (1992). Interval mapping in the analysis of nonadditive quantitative trait loci. *Biometrics* **48**, 305-315.

Champoux, M.C., Wang, G., Sarkarung, S., Mackill, D.J., O'Toole, J.C., Huang, N., McCouch, S.R. (1995). Locating genes associated with root morphology and drought avoidance in rice via linkage to molecular markers. Theoretical and Applied Genetics **90**, 969-981.

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

Doerge, R. W. (1998). Statistical threshold values for locating quantitative trait loci. *Proceedings of the Kansas State University Conference on Applied Statistics in Agriculture*. 84-95.

Doerge, R. W. and Rebaï, A. (1996). Significance thresholds for QTL interval mapping tests. *Heredity* **76**, 459-464.

Dupuis, J. (1994). *Statistical problems associated with mapping complex and quantitative traits from genomic mismatch scanning data.* Ph. D. Thesis of Stanford University, Department of Statistics, USA.

Fisher, R. A. (1935). The Design of Experiments, 3rd ed. Oliver and Boyd, London.

Hunt, G. J. (1998). Quantitative trait loci for honey bee stinging behavior and body size. *Genetics* **148**, 1203-1214.

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

Nettleton, D. (1999). Order-restricted hypothesis testing in a variation of the normal mixture model. To appear in *The Canadian Journal of Statistics.*

Nettleton, D. and Doerge, R. W. (2000). Accounting for variability in the use of permutation testing to detect quantitative trait loci. To appear in *Biometrics.*

Nettleton, D. and Praestgaard J. (1998). Interval mapping of quantitative trait loci through order restricted inference. *Biometrics* **54**, 74-87.

Paterson, A. H., Damon, S., Hewitt, J. D., Zamir, D., Rabinowitch, H. D., Lincoln, S. E., Lander, E. S., and Tanksley, S. D. (1991). Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics* **127**, 181-197.

Paterson, A. H., Lander, E. S., Hewitt, J. D., Peterson, S., Lincoln, S. E., and Tanksley, S. D. (1988). Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature* **335**, 721-726.

Plomin, R., McClearn, G. E., and Smith, D. L. (1994). DNA Markers Associated with High Versus Low IQ: The IQ Quantitative Trait Loci (QTL) Project. *Behavior Genetics* **24**, 107-118.

Rebaï, A., Goffinet, B., and Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics* **138**, 235-240.

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

Soller, M., Brody, T., and Genizi, A. (1976). On the power of experimental designs for the detection of linkage between marker loci and quantitative loci in crossses between inbred lines. *Theoretical and Applied Genetics* **47**, 35-39.

Table 1: Permutation p-value estimates and confidence intervals for a simulated sample of 100 lines.

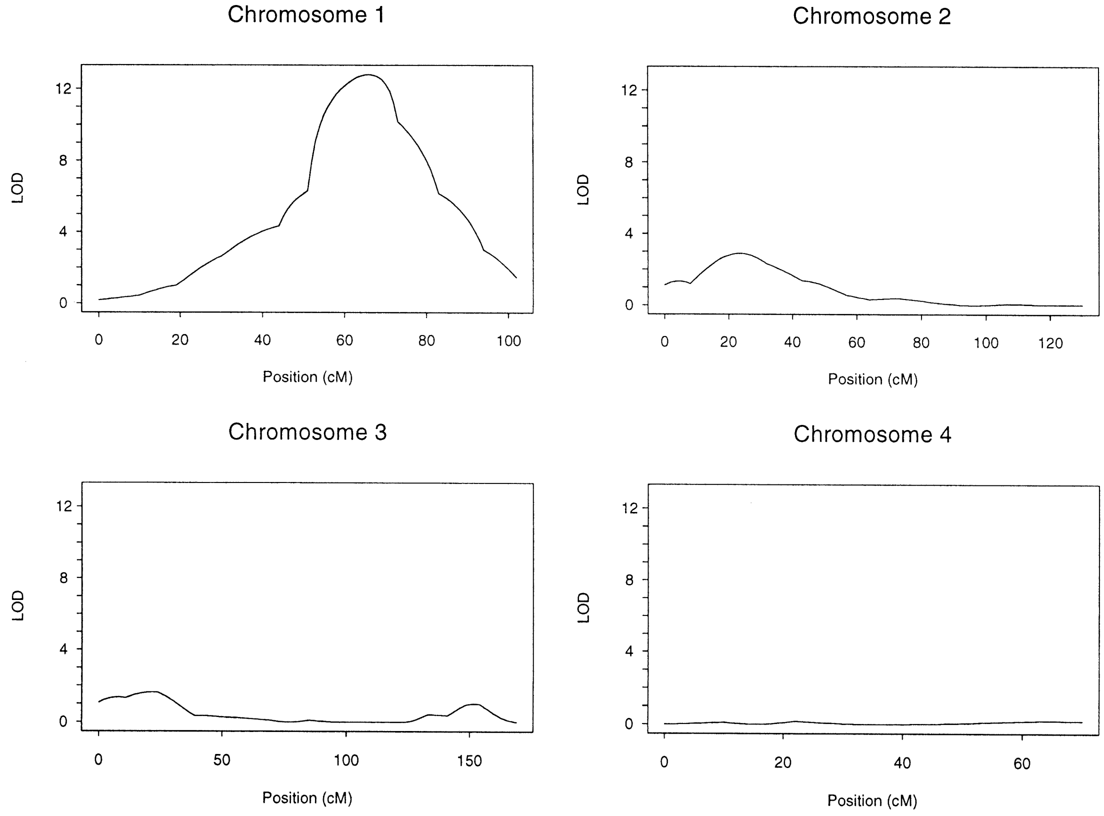
Chromosome	Position[a]	LOD	P-Value[b]	Confidence Interval[c]
1	66	12.808	0.0000	***
2	24	2.894	0.01645	[0.009,0.024]
3	22	1.640	0.2112	[0.187, 0.235]
3	152	1.013	0.6417	[0.613,0.670]

[a] Distance in cM from the first marker on the chromosome

[b] Estimated experimentwise permutation p-value based on 1094 data permutations

[c] Approximate 95% confidence interval for the p-value based on 1094 data permutations

Figure 1: Plots of LOD scores at 1 cM intervals for each of four simulated chromosomes for 100 lines and a total of 46 markers arbitrarily positioned throughout the genome. The first QTL is 62 cM from the left end of chromosome 1. The second QTL is 44 cM from the left end of chromosome 2. The third and fourth QTL are 17 cM and 147 cM, respectively, from the left end of chromosome 3. Chromosome 4 contains no QTL.

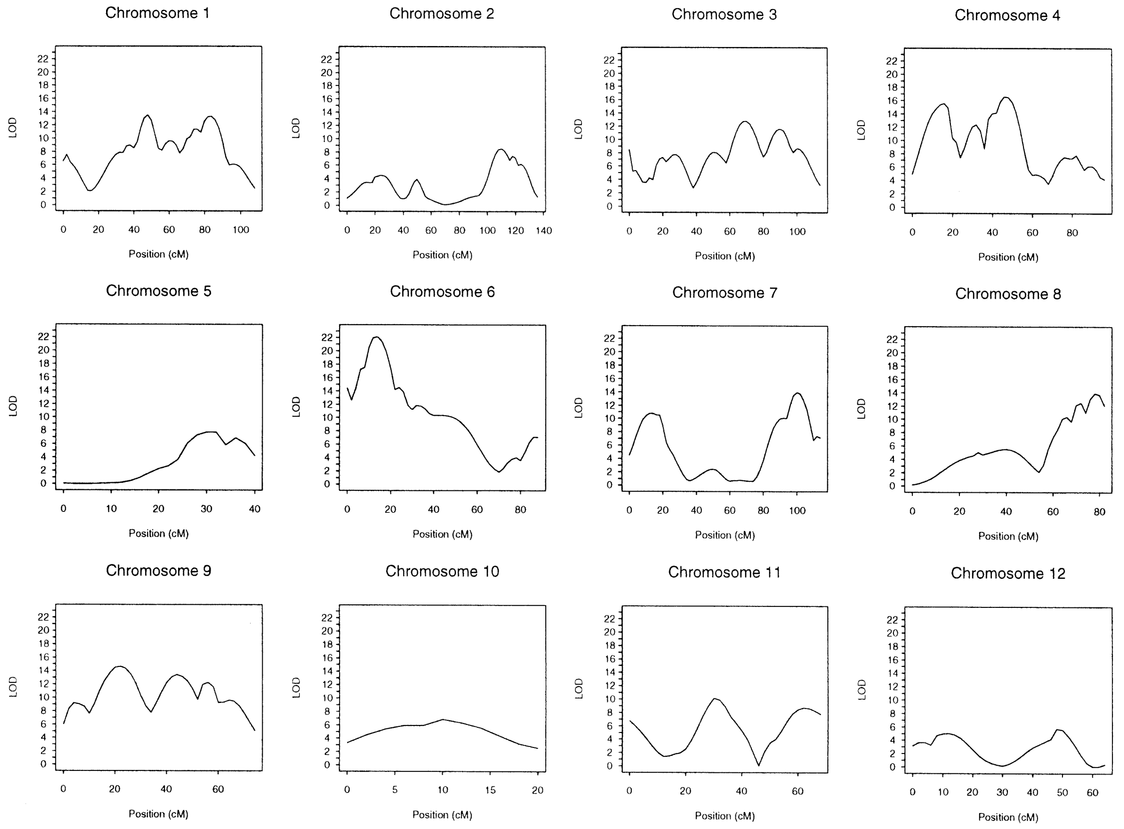

Chromosome 1

Chromosome 2

Chromosome 3

Chromosome 4

Figure 2: Plots of LOD scores at 2 cM intervals for 12 rice chromosomes based on recombinant inbred line data from the study of Champoux et al. (1995).


Chromosome 1
LOD
Position (cM)
Chromosome 2
LOD
Position (cM)
Chromosome 3
LOD
Position (cM)
Chromosome 4
LOD
Position (cM)
Chromosome 5
LOD
Position (cM)
Chromosome 6
LOD
Position (cM)
Chromosome 7
LOD
Position (cM)
Chromosome 8
LOD
Position (cM)
Chromosome 9
LOD
Position (cM)
Chromosome 10
LOD
Position (cM)
Chromosome 11
LOD
Position (cM)
Chromosome 12
LOD
Position (cM)