ACCOUNTING FOR VARIABILITY IN THE USE OF PERMUTATION TESTING TO DETECT QUANTITATIVE TRAIT LOCI

by

Dan Nettleton
University of Nebraska

and

R. W. Doerge
Purdue University

Technical Report #98-15

Department of Statistics
Purdue University

June 1998
Revised May 1999

Accounting for Variability in the Use of Permutation Testing to Detect Quantitative Trait Loci

Dan Nettleton

Department of Mathematics and Statistics, 917 Oldfather Hall

University of Nebraska, Lincoln, NE 68588-0323

and

R.W. Doerge

Departments of Agronomy and Statistics, 1399 Mathematical Science Building

Purdue University, West Lafayette, IN 47907-1399

Running Head: Variability in Permutation Testing

Corresponding author: Dan Nettleton
Department of Mathematics and Statistics
917 Oldfather Hall
University of Nebraska
Lincoln, NE 68588-0323

Telephone number: (402) 472-7241
Fax number: (402) 472-8466
E-mail: dnettlet@mathstat.unl.edu

Key words: permutation test, QTL mapping
p-value, number of permutations
statistical genetics

Summary

Locating quantitative trait loci (QTL), or genomic regions associated with known molecular markers, is of increasing interest in a wide variety of applications ranging from human genetics to agricultural genetics. The hope of locating QTL (or genes) affecting a quantitative trait is that it will lead to characterization and possible manipulations of these genes. However, the complexity of both statistical and genetic issues surrounding the location of these regions calls into question the asymptotic statistical results supplying the distribution of the test statistics employed. Coupled with the power of current–day computing, permutation theory was reintroduced for the purpose of estimating the distribution of any test statistic used to test for the location of QTL. Permutation techniques have offered an attractive alternative to significance measures based on asymptotic theory. The ideas of permutation testing are extended in this application to include confidence intervals for the thresholds and p-values estimated in permutation testing procedures. The confidence intervals developed account for the Monte Carlo error associated with practical applications of permutation testing, provide tighter controls on type I and type II errors in QTL analyses, and lead to an effective method of determining an efficient permutation sample size.

1 Introduction

Steady advances have been made in the last five years toward developing powerful methodology for locating regions of human, animal, and plant genomes responsible for quantitative variation. These genomic regions are commonly called quantitative trait loci or QTL and serve as fence posts for statistical associations between measurable/observable characteristics (phenotypes) and the genome at hand. The long-term hope in identifying these genomic regions is that they will lead to candidate genes, which in turn may be verified, and eventually sequenced.

Many related statistical techniques can be used to search for associations between genotype and phenotype. Examples include traditional single-marker analysis (Sax, 1923; Soller, Brody, and Genizi, 1976), interval mapping (Lander and Botstein, 1989), regression of phenotype on marker genotype (Haley and Knott, 1992; Martínez and Curnow, 1992; Whittaker, Thompson, and Visscher, 1996), composite interval mapping (Zeng, 1993, 1994), and the MQM methods of Jansen (1993, 1994). For a review of these procedures see Doerge *et al.* (1997). Regardless of the method used, a statistically significant association between a locus (or loci) and a trait is suggested by an unusually large observed test statistic and is typically referred to as a QTL. Following Churchill and Doerge (1994), the location at which any one of these test statistics is computed will be called an analysis point. Determining which analysis points have unusually large test statistics is the goal of a distribution-free technique known as permutation testing.

Permutation testing was originally introduced by Fisher (1935) and first applied to the QTL mapping problem by Churchill and Doerge (1994). The procedure detects significant association between genotype and phenotype by comparing the profile of observed test statis-

tics at all analysis points to its permutation distribution which is estimated empirically. The permutation distribution indicates how the profile of test statistics behaves in the absence of a true association between genotype and phenotype. Determining which, if any, of the test statistics are larger than would be expected because of chance variation alone is possible by comparing the original profile to the distribution of permutation-replicated profiles. The positions corresponding to unusually large test statistics in the original profile are those significantly associated with the quantitative trait.

Churchill and Doerge (1994) discuss how the information about the permutation distribution can be used to assess the evidence for QTL presence at each analysis point. They explain how to determine permutation-based thresholds – values that separate the unusually large test statistics from those whose values are small enough to be considered unassociated with the trait. Two types of thresholds, comparisonwise and experimentwise, are described. Analysis points with test statistics that meet or exceed the selected threshold are considered significantly linked to the trait while those with statistics falling below the threshold are judged to be unassociated with the trait. In practice, both thresholds are estimated from a randomly chosen sample of data permutations because it is typically infeasible from a computational standpoint to consider all possible data permutations directly. This paper focuses on techniques for managing and reporting the variability associated with this permutation sampling in the context of QTL data analysis. We extend the concept of permutation thresholds to permutation p-values and develop confidence intervals for both thresholds and p-values. We suggest the use of these intervals when drawing conclusions from QTL studies that employ permutation techniques. Careful accounting of permutation testing variability through confidence intervals decreases the chance of falsely declaring a significant association between an analysis point and a trait. The presented methodology frees the QTL researcher

from selecting a permutation sample size (*i.e.*, the number of data permutations required), and allows for efficient use of computer resources by automatically determining the minimum number of data permutations necessary to produce unambiguous results.

Figure 1 exhibits typical plots of LOD score (negative log base 10 of the likelihood ratio) versus the genetic map position for each of 12 rice chromosomes. These profiles result from the application of interval mapping (Lander and Botstein, 1989) to data from recombinant inbred lines of rice. (Champoux *et al.*, 1995). Large LOD scores over a genetic region are indicative of potential associations between the genetic region and the quantitative trait of interest – root thickness in micrometers in this case. We discuss these data in greater detail in Section 3 and use permutation testing to determine which of the many peaks in Figure 1 can be considered as significant evidence of association between the corresponding genetic region and the quantitative trait. Churchill and Doerge (1994) highlight the many benefits of a permutation testing analysis for these data, as well as in general. Chief among these benefits is robustness to departures from traditional modeling assumptions and the ability to control overall type I error rate while conducting multiple tests with multiple dependent test statistics. One drawback, however, is the computation time required for such procedures. Thus, our analysis in Section 3.3 focuses on determining the minimum number of permutations needed to make sound claims about the significance of each of the peaks in Figure 1.

2 Methods

2.1 Permutation Thresholds and P-Values

Following Churchill and Doerge (1994), let individuals in the QTL experiment be indexed from 1 to n, and let N denote the number of randomly selected data permutations for which test statistics at all analysis points are recomputed. Test statistics computed from a permutation of the original data will be referred to as permutation-replicated test statistics. There are $n!$ (not necessarily distinct) permutation-replicated test statistics at each analysis point.

In general, a p-value is defined as the maximum probability of obtaining a test statistic at least as extreme as the test statistic observed from the original data under the assumption that the null hypothesis is true. A *comparisonwise permutation p-value* is defined for any particular analysis point as the proportion of the $n!$ permutation-replicated test statistics that match or exceed the original test statistic at that analysis point. Such p-values are the analog of the comparisonwise thresholds discussed by Churchill and Doerge (1994). The level-α comparisonwise threshold for any particular analysis point is defined as the $1-\alpha$ quantile of the $n!$ permutation-replicated test statistics computed at that analysis point. Because it is typically infeasible from a computational standpoint to examine all $n!$ data permutations, the permutation p-values or thresholds are estimated by using a randomly chosen sample of N data permutations in place of all $n!$ data permutations. A small estimated p-value – less than a preselected type I error rate denoted α – is usually considered evidence of an association between the analysis point and the trait. Equivalently, a test statistic exceeding its estimated level-α threshold is usually considered (comparisonwise) significant at level α.

Comparisonwise analyses are necessary when the null distribution of the test statistic

under consideration varies from analysis point to analysis point. A comparisonwise p-value or threshold, however, must be interpreted with caution because its type I error rate applies only to the single locus under consideration. The chance of incorrectly suggesting that some analysis point is linked to the trait is much greater than the specified comparisonwise type I error rate because many analysis points are examined. Since most test statistics used in QTL mapping have null distributions that do not vary with position, Churchill and Doerge (1994) present experimentwise thresholds to remedy the multiple testing problem.

The level-α experimentwise threshold is defined as the $1 - \alpha$ quantile of the $n!$ maximums of the permutation-replicated test statistic profiles. The analog is the *experimentwise permutation p-value*, estimated for any particular analysis point by the proportion of the N maximums of the permutation-replicated test statistic profiles that match or exceed the original test statistic. The chance of any false QTL declaration somewhere in any linkage group is no greater than the selected type I error rate α as long as only analysis points with experimentwise p-values less than or equal to α – or, equivalently, test statistics greater than or equal to the level-α experimentwise threshold – are implicated as QTL.

2.2 Motivation

Point estimates of permutation thresholds or p-values may be extended to include the accuracy with which they are estimated. The effect of permutation sampling variability can be magnified in QTL studies because p-values and/or thresholds are often estimated for many analysis points across a genome. While the chance of an errant decision caused by sampling from all permutations is small at any particular analysis point, the chance of an error at some analysis point grows larger as the number of analysis points increases. Confidence

intervals for p-values and/or thresholds, on the other hand, reflect permutation sampling variability, and are generally more useful than a single point estimate in QTL analyses. We suggest that inferential QTL studies utilizing permutation testing proceed in two stages. First, conclusions in the form of interval estimates for permutation p-values or thresholds should be based on the permutation distribution of all test statistics of interest. Second, the interval estimates should be used to make inferences pertaining to which regions of the genome are significantly associated with the trait of interest. A test statistic at an analysis point may be considered significant at the α level only when interval estimates of the p-value at that analysis point exclude permutation p-values above α. Equivalently, only test statistics exceeding the upper limit of the confidence interval for the level-α threshold should be considered significant at level α.

When a confidence interval for a p-value includes α – or, equivalently, when a confidence interval for a threshold includes a test statistic, the status of the corresponding analysis point is unclear. The permutation sample size N should be increased until the interval for the p-value (threshold) falls either entirely above or entirely below the significance level α (the test statistic). It may not be affordable from a computational standpoint to resolve all such analysis points, but this resolution can be obtained in many cases using relatively few permutations.

2.3 Confidence Intervals for Assessing Comparisonwise Significance

Suppose the profile of test statistics computed from the original data contains k peaks, each of which might suggest significant evidence for QTL presence at k corresponding analysis points.

Let $p_1, \ldots, p_k$ denote the unknown comparisonwise permutation p-values associated with the k analysis points. Let $t_1, \ldots, t_k$ denote the test statistics computed at these k analysis points. For $j = 1, \ldots, N$; let $t_{j1}, \ldots, t_{jk}$ denote permutation replications of $t_1, \ldots, t_k$, respectively, computed with the j^{th} of N permuted data sets. For $j = 1, \ldots, N$ and $\ell = 1, \ldots k$; let

$$X_{j\ell} = \begin{cases} 1 & \text{if } t_\ell \leq t_{j\ell} \\ 0 & \text{if } t_\ell > t_{j\ell}. \end{cases}$$

For $\ell = 1, \ldots, k$; let $\hat{p}_\ell = \frac{1}{N} \sum_{j=1}^{N} X_{j\ell}$. Note that $\hat{p}_\ell$ is the estimated comparisonwise p-value for the ℓ^{th} of the k positions.

For $\ell = 1, \ldots, k$; $N\hat{p}_\ell$ has a binomial distribution with N as the number of trials and p_ℓ as the probability of success. $N\hat{p}_\ell$ counts the number of the N sampled data permutations for which $t_\ell \leq t_{j\ell}$ while p_ℓ is the proportion of all permuted data sets that yield a test statistic greater than or equal to t_ℓ. If the condition $Np_\ell \geq 5$ is satisfied, one may rely on the normal approximation to the binomial distribution so that $100(1-\gamma)\%$ confidence intervals for the k comparisonwise permutation p-values are given by

$$(\hat{p}_\ell - \Phi^{-1}(1-\gamma/2)\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N},\ \ \hat{p}_\ell + \Phi^{-1}(1-\gamma/2)\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N}) \text{ for } \ell = 1, \ldots, k; \quad (1)$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable.

Each individual interval provides some protection against an error associated with sampling from the $n!$ permutations. Simultaneous intervals, however, are needed to account for the possibility of up to k such errors. The intervals in equation (1), when considered together, can be Bonferroni-adjusted to provide at least $100(1-\gamma)\%$ confidence that $p_1, \ldots, p_k$ are all contained in their respective intervals simultaneously. Approximate $100(1-\gamma)\%$ Bonferroni simultaneous confidence intervals for $p_1, \ldots, p_k$ are given by

$$(\hat{p}_\ell - \Phi^{-1}(1-\frac{\gamma}{2k})\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N},\ \ \hat{p}_\ell + \Phi^{-1}(1-\frac{\gamma}{2k})\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N}) \text{ for } \ell = 1, \ldots, k.$$

The condition $Np_\ell \geq 5$ may not be satisfied for any computationally feasible value of N when the permutation p-value p_ℓ is extremely small. In such cases, the permutation p-value is typically much smaller than the significance level α. Because the evidence for QTL presence is likely to be overwhelming for such an analysis point (estimated permutation p-value $<< \alpha$), there is little need for a confidence interval at that particular analysis point. The procedure still provides valid intervals for other analysis points at which the permutation p-values are closer to the selected significance level α. There is evidence that Np_ℓ may be less than 5 when $N\hat{p}_\ell$ is near 5. For such analysis points, the coverage of the comparisonwise p-value confidence interval is likely to be less than the nominal significance level because of the ineffectiveness of the normal approximation to the binomial distribution. Confidence intervals for comparisonwise thresholds can be obtained in place of the intervals for comparisonwise p-values in such cases. We defer the discussion of threshold confidence intervals to the next section.

2.4 Confidence Intervals for Assessing Experimentwise Significance

Suppose the profile of test statistics computed from the original data contains k peaks, each of which might suggest significant evidence for QTL presence at k corresponding analysis points. Let $p_1, \ldots, p_k$ denote the unknown experimentwise permutation p-values associated with the k analysis points. Let $t_1, \ldots, t_k$ denote the test statistics computed at these k analysis points. For $j = 1, \ldots, N$; let M_j denote the maximum of the profile of test statistics computed at all analysis points for the j^{th} of N permuted data sets. For $j = 1, \ldots, N$ and

$\ell = 1, \ldots k$; let

$$X_{j\ell} = \begin{cases} 1 & \text{if } t_\ell \leq M_j \\ 0 & \text{if } t_\ell > M_j. \end{cases}$$

For $\ell = 1, \ldots, k$; let $\hat{p}_\ell = \frac{1}{N}\sum_{j=1}^{N} X_{j\ell}$. Note that $\hat{p}_\ell$ is the estimated experimentwise p-value for the ℓ^{th} of the k positions.

For $\ell = 1, \ldots, k$; $N\hat{p}_\ell$ has a binomial distribution with N as the number of trials and p_ℓ as the probability of success. $N\hat{p}_\ell$ counts the number of the N sampled data permutations for which $t_\ell \leq M_j$ while p_ℓ is the proportion of all permuted data sets that yield a test statistic greater than or equal to t_ℓ. If the condition $Np_\ell \geq 5$ is satisfied, one may rely on the normal approximation to the binomial distribution so that approximate $100(1-\gamma)\%$ confidence intervals for the k experimentwise permutation p-values are given by

$$(\hat{p}_\ell - \Phi^{-1}(1-\gamma/2)\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N},\ \ \hat{p}_\ell + \Phi^{-1}(1-\gamma/2)\sqrt{\hat{p}_\ell(1-\hat{p}_\ell)/N}) \text{ for } \ell = 1, \ldots, k; \quad (2)$$

where $\Phi(\cdot)$ denotes the distribution function of a standard normal random variable.

Confidence intervals for permutation thresholds can be obtained in addition to the intervals for permutation p-values. Let $M_{(1)} \leq \cdots \leq M_{(N)}$ denote the values of $M_1, \ldots, M_N$ ordered from smallest to largest. The level-α experimentwise threshold may be estimated by $M_{(\lceil N\alpha \rceil)}$ where $\lceil x \rceil$ denotes the smallest integer greater than or equal to x. Using standard results on quantile estimation (*e.g.*, Conover, 1980; p.112), an approximate $100(1-\gamma)\%$ confidence interval for the exact experimentwise level-α permutation threshold is $[M_{(L)}, M_{(U)}]$ where

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

This interval is based on the normal approximation to the binomial distribution and is valid

as long as $N\alpha \geq 5$ which suggests that at least $\lceil 5/\alpha \rceil$ data permutations should be considered to estimate a level-α threshold.

We recommend judging the experimentwise significance of the k analysis points of interest by comparing $t_1, \ldots, t_k$ to the interval $[M_{(L)}, M_{(U)}]$. Analysis point ℓ should be considered significantly associated with the trait at experimentwise significance level α if and only if $t_\ell \geq M_{(U)}$. If $t_\ell < M_{(L)}$, it is safe to conclude that analysis point ℓ is not significantly associated with the trait at experimentwise significance level α. If $M_{(L)} \leq t_\ell < M_{(U)}$, the results are inconclusive. When referring to the confidence interval for the experimentwise threshold, there is no need for a Bonferroni correction for multiple testing even though k analysis points are being considered. At this stage of inference $t_1, \ldots, t_k$ are considered fixed, and only a single quantile of the distribution of permutation-replicated profile maximums is being estimated.

The confidence intervals for the permutation p-values could be used to judge significance in place of the procedure above. Analysis point ℓ could be considered significant at experimentwise significance level α if and only if the upper endpoint the confidence interval for its experimentwise permutation p-value is less than or equal to α. If the lower endpoint of the interval is greater than α it is reasonable to conclude that the analysis point is not significantly associated with the trait at experimentwise level of significance α. The results for a given analysis point are inconclusive if the interval contains α.

The main drawback associated with the use of the p-value confidence intervals for determining significance is that these intervals are not simultaneous even though all are computed with respect to the distribution of permutation-replicated profile maximums. In other words, the probability that all k intervals contain their respective permutation p-values is potentially less that $1 - \gamma$ even though each interval has individual coverage probability

approximately $1 - \gamma$. The intervals could be made simultaneous through the common Bonferroni correction, although the resulting procedure for determining significance would be more conservative than the procedure based on $[M_{(L)}, M_{(U)}]$.

2.5 Determination of the Permutation Sample Size

Doerge and Churchill (1996) stated a result that allows one to calculate the minimum number of data permutations required to estimate a level-α threshold with less than a specified level of Monte Carlo resampling error. Inherent in the confidence interval procedures put forth here is the ability to stop permuting once each of the analysis points of interest has been resolved as significant or insignificant. This is a valuable feature since the tolerable amount of Monte Carlo resampling error depends on the true significance of the test statistics at the k analyses worth considering as potential QTL. The status of the k analysis points can be correctly determined despite relatively large Monte Carlo resampling error when all k analysis points are either strongly significant or insignificant. On the other hand, very little Monte Carlo resampling error is acceptable when several analysis points are marginally significant or insignificant. Since it is difficult to know which situation applies to the data at hand before permutation testing begins, a procedure for dynamically determining permutation sample size that is specific to the observed data can substantially improve the efficiency of permutation testing in QTL analyses.

Consider the following procedure for determining the level-α experimentwise significance of the test statistics at k analysis points.

1. Choose $N_{\min} \geq \lceil 5/\alpha \rceil$ as the minimum number of data permutations that will be analyzed.

2. Choose $N_{\max}$ as the maximum number of data permutations that can be analyzed in a reasonable amount of time given the data set at hand and the computing power available. The value $N_{\max}$ should not typically play a role in the analyses. It is specified only to halt the procedure if the desired conclusions cannot be reached after an acceptable number of iterations.

3. Once $N_{\min}$ data permutations have been analyzed, determine whether or not the k test statistics are resolved as significant or insignificant by comparing their values to $M_{(L)}$ and $M_{(U)}$ as described in the previous section.

4. If all k statistics are resolved as significant or insignificant, no additional data permutations need to be considered. If one or more test statistics are unresolved, continue analyzing randomly selected data permutations until all k statistics are resolved or $N = N_{\max}$.

This procedure frees the researcher from specifying an acceptable level of Monte Carlo resampling error or the permutation sample size N. In addition, the question, "Have enough data permutations been considered to yield unambiguous results?", is all but eliminated. The number of permutations required for clear-cut results is automatically determined for the specific data at hand when the procedure terminates before reaching $N_{\max}$. If there is insufficient computing power to resolve all analysis points as significant or insignificant, the confidence interval for the threshold computed with $N_{\max}$ permutations should be reported along with the test statistic or test statistics that fall in this interval.

3 Simulated and Real Data Examples

As demonstration of both the comparisonwise and experimentwise confidence interval methods, we present three different analyses, two of which are centered around the QTL work in root morphology of rice published initially by Champoux *et al.* (1995), and the other being simulated data that are presented and analyzed for the purpose of illustrating the advantages and limitations of this methodology in determining significance of QTL results.

3.1 Single Marker Analysis of Recombinant Inbred Data

Following the single marker analysis initiated by Churchill and Doerge (1994) of a recombinant inbred (F_7) population of rice (*Oryza sativa* L.) (Champoux *et al.*, 1995) derived from a cross between CO39 (maternal) and Moroberekan, we analyze the same data using permutation based confidence intervals. The data consist of 203 recombinant inbred lines, each scored at 147 molecular (RFLP) markers. The focus of the original experiment by Champoux *et al.* was the identification and mapping of QTL associated with root morphology traits. The recombinant inbred lines were the result of a cross between *indica* cultivar CO39, which has a fine root structure, and *japonica*, which has a thick root complex. One of the results of single marker and interval mapping analyses performed by Champoux *et al.* was the key finding that selecting for Moroberekan alleles at marker loci associated with presumed root morphology QTL maybe an adequate course of action for altering the root phenotype. We aim to reanalyze these data using the described confidence intervals, and in keeping with the previous (Churchill and Doerge, 1994; Doerge and Churchill, 1996) analyses, the quantitative trait of interest is root thickness (micrometers).

The reanalysis of these data serves to verify that the proposed confidence intervals for

permutation p-values effectively account for the Monte Carlo error associated with sampling from all data permutations, and to confirm the findings of the original work (Champoux *et al.*, 1995). 1000 data permutations were performed, and respective single marker test statistics (t-tests) were used to estimate the comparisonwise p-value at each marker. 95% confidence intervals for comparisonwise p-values were calculated for each marker and reported when the data permutation resulted in non-zero p-values. The theoretical p-value based on a t-distribution with 201 degrees of freedom was computed for each relevant marker. Since the t-distribution is an appropriate reference distribution in this case, the theoretical t-based p-values serve as good approximations of the exact comparisonwise permutation p-values estimated with the 1000 data permutations. If the confidence intervals for the p-values are working as they should, any t-based p-value will fall within the confidence intervals for the exact permutation p-values with approximate probability 0.95. We found that the approximate 95% confidence intervals for the exact permutation p-values contained their respective t-based p-values for all 44 of the analysis points with non-zero p-value estimates. This performance is perhaps somewhat better than would be expected, but it is important to note there is dependence among the test statistics and among the confidence interval estimates. Each of the remaining 103 analysis points has a t-statistic that was never matched or exceeded by any of its 1000 permutation-replicated t-statistics. Hence the exact permutation p-values for these points are estimated to be zero, and no confidence interval calculation is possible. These 103 analysis points have an average t-based p-value of 0.00009, standard deviation of 0.00036, and a maximum of 0.0026. Based on these findings the zero p-value estimates are realistic in all of the 103 cases.

3.2 Interval Mapping Analysis of Simulated Data

A sample of 100 backcross individuals with four chromosomes and four QTL were simulated. The trait value of the i^{th} individual was determined using $Y_i = 2.50Q_{i1} + 0.75Q_{i2} + 1.00Q_{i3} + 1.00Q_{i4} + \epsilon_i$, where ϵ_i is a standard normal environmental error term, with additive effects defined as $Q_{ij} = 1$ if the i^{th} individual is heterozygous 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 Lander and Botstein (1989). The resulting test statistic profile for each chromosome is depicted in Figure 2. 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, respectively). Examination of the profiles in Figure 2 reveals the only other analysis points with large LOD scores are clearly linked to one of the four analysis points above. Consequently, the subsequent permutation analyses will focus on these four points ($k = 4$).

Only 53 data permutations were required to determine the experimentwise significance of the four analysis points at the 0.10 level. An approximate 95% confidence interval for the 0.90 quantile of the distribution of permutation-replicated profile maximums was determined to be $[1.715, 2.346]$ using the methods outlined previously. The peaks on chromosomes 1 and

2 are judged significant (experimentwise) at the 0.10 level since 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 experimentwise significance at level 0.05, 110 data permutations are sufficient in this case. An approximate 95% confidence interval for the 0.95 quantile of the distribution of permutation-replicated profile maximums was determined to be $[1.940, 2.659]$. The peaks on chromosomes 1 and 2 are thus experimentwise significant at the 0.05 level.

Even 1000 data permutations are insufficient to determine the significance status of all four points when considering 0.01-level experimentwise significance. An approximate 95% confidence interval for the 0.99 quantile was determined to be $[2.801, 3.530]$ using 1,000 randomly selected data permutations. The analysis point on chromosome 1 is clearly significant while the points on chromosome 3 are clearly insignificant. 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 0.99 quantile of the distribution of permutation-replicated profile maximums. Hence, the second analysis point is judged insignificant at the 0.01 experimentwise significance level.

Confidence intervals for experimentwise permutation p-values can be computed for the analysis points 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 analysis point since its test statistic was exceeded by none of the 1094 permutation-replicated profile maximums. If the true experimentwise permutation p-value is actually 0.01 or bigger, the chance of estimating the p-value to be zero based on 1094 data permutation is extremely small (no larger than $0.99^{1094} = 0.0000168$).

We can be quite confident that this analysis point is significantly linked to the trait.

Note that the confidence interval for the second analysis point includes 0.01, suggesting that this analysis point may be significant at the 0.01 level. Examination of the threshold confidence interval suggested insignificance at the 0.01 level using the same 1094 data permutations. Such minor discrepancies are possible when analysis points 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 threshold-based analysis.

3.3 Interval Mapping Analysis of Recombinant Inbred Data

Continuing the prior analysis of the described root morphology data (Champoux *et al.*, 1995), interval mapping (Lander and Botstein, 1989) was used to compute LOD scores at 2 cM increments across each of the 12 rice chromosomes. The resulting LOD profiles are displayed in Figure 1. The procedure described in Section 2.5 was used to determine the smallest permutation sample size needed to obtain the 0.05-level experimentwise significance for each of the many peaks in Figure 1. Initially, 100 permutations were analyzed yielding an estimated experimentwise 0.05-level critical value of 2.644. The associated 95% confidence interval for the exact threshold is $(2.366, 5.793)$.

Based on only these first 100 permutations, it is safe to conclude that most of the peaks in Figure 1 are easily significant at the 0.05 level because most of the corresponding LOD scores exceed 5.793, the upper endpoint of the threshold confidence interval. Peaks whose significance is less certain include the first two peaks on chromosome 2, the middle peak on chromosome 7, the first peak on chromosome 8, and the peaks on chromosome 12. The LOD scores associated with these peaks range from a low of 2.420 on chromosome 7 to a high of

5.707 on chromosome 12.

After 142 permutations, all peaks are excluded from the 95% confidence interval for the 0.05-level threshold. The point and interval estimates for the threshold are 2.934 and $(2.517, 3.569)$, respectively. The middle peak on chromosome 7 falls below the lower endpoint of the confidence interval while all other peaks exceed the upper endpoint of the confidence interval. Thus, it is reasonable to declare all peaks – aside from the middle peak on chromosome 7 – significant at experimentwise level 0.05. This conclusion is correct provided that our confidence interval for the threshold contains the true permutation threshold computed from all possible data permutations. An additional 358 permutations were examined bringing the total number of permutations to 500. Using all 500 permutation yields an interval estimate for the exact permutation threshold of $(2.670, 3.038)$. Our conclusions based on this interval are no different than those obtained with the interval based on 142 permutations because the former interval falls within the latter.

Champoux *et al.* (1995) reported evidence of significant association between genomic region and root thickness on all chromosomes except chromosome 5. Our analysis suggests significant association on all chromosomes because the maximum LOD score on chromosome 5 is 7.746. The discrepancy is the result of the standard for significance used by Champoux *et al.* who only reported significant association if a marker on a chromosome had a LOD score exceeding 4.0 and an F-statistic exceeding 19.22. These values were chosen to assure very small significance levels for individual tests in hopes of appropriately controlling overall type I error. Our analysis allows direct control of the overall type I error rate and is less conservative than the approach used by Champoux *et al.* (1995).

4 Discussion

Permutation testing has become a popular tool for QTL researchers because of its validity and effectiveness in a variety of realistic data analysis settings. The techniques presented in this paper provide an effective way to control for the variability associated with permutation testing in QTL mapping experiments. Without examining confidence intervals like those proposed herein, there is no guarantee that a user-selected permutation sample size will be large enough to sufficiently reduce the ambiguity introduced by sampling from all permutations. At the other extreme, user specified values of N may be larger than necessary, resulting in inefficient use of computer time.

The intervals for p-values and thresholds are approximate confidence intervals because they are based on the normal approximation to the binomial distribution. It is possible to replace the asymptotic intervals by exact intervals based directly on the binomial distribution. Exact intervals, however, are more difficult to compute and provide only small gains in accuracy. The dynamic method of selecting permutation sample size can bias the intervals (asymptotic and exact) to some degree. When the test statistic at an analysis point falls near the boundary of a confidence interval for a threshold, for example, there is a small possibility that the final interval will be shifted artificially away from the test statistic value due to the stopping criterion. Examining a fixed number of permutations before assessing the stopping criterion substantially reduces the chance that such bias will have an important impact on QTL analyses.

It may be possible to develop a less conservative method of determining experimentwise significance for secondary peaks in the profile of test statistics. A set of k analysis points of interest is singled out because the points are associated with the k largest peaks over

the original profile of test statistics. The same basic procedure could be repeated for each permuted data set to gauge the significance of the k peaks. The largest peak in the original profile should be compared with the distribution of permutation-replicated profile maximums as before. The second largest peak in the original profile could be compared with the distribution of permutation-replicated second largest peaks while, perhaps, maintaining an appropriate overall type I error rate. An analogous approach could be used for other lesser peaks. Unfortunately, some judgement is required in determining the k largest peaks in a given profile. It is difficult to know, for example, whether a lesser peak is simply an artifact of close proximity to a larger peak.

We have extended current permutation methodology (Churchill and Doerge, 1994) to include the variability inherent in any resampling procedure by providing confidence intervals for assessing both comparisonwise and experimentwise significance. Permutation based confidence intervals maximize the benefits of the time and resources spent collecting QTL data by improving the quality of the inference made from the data. Such improvements naturally lead to a better understanding of the relationship between QTL and quantitative traits.

5 Acknowledgements

Dan Nettleton acknowledges partial support from the National Research Initiative Competitive Grants Program of the U.S. Department of Agriculture, Award 98-35205-6709. R.W. Doerge acknowledges partial support from the National Research Initiative Competitive Grants Program of the U.S. Department of Agriculture, Award 98-35300-6173.

References

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. Theor. Appl. Genet. **90:** 969-981.

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

Conover, W.J. (1980). Practical Nonparametric Statistics, 3rd ed. Wiley and Sons. New York.

Doerge, R.W. and Churchill , G.A. (1996). Permutation tests for multiple loci affecting a quantitative character. *Genetics* **142:**, 285-294.

Doerge, R.W., Z-B Zeng, Z.-B ,and Weir, B.S. (1997). Statistical issues in the search for genes affecting quantitative traits in experimental populations. *Statistical Science* **12:**, 195-219.

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

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

Jansen, R.C. (1993). Interval mapping of multiple quantitative trait loci. *Genetics* **135:**, 205-211.

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

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

Martínez, O. and Curnow, R.N. (1992). Estimating the locations and the size of the effects of quantitative trait loci using flanking markers. *Theoretical and Applied Genetics* **85:**, 480-488.

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 crosses between inbred lines. *Theoretical and Applied Genetics* **47:**, 35-39.

Whittaker, J.C., Thompson, R., and Visscher, P.M. (1996). On the mapping of QTL by the regression of phenotype on marker-type. *Heredity* **77:**, 23-32.

Zeng, Z-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. *Proceeding of the National Academy of Science* **90:**, 10972-10976.

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

Table 1: Experimentwise permutation p-value estimates and confidence intervals for a backcross sample of 100 individuals.

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 the p-value based on 1094 data permutations

Figure 1: LOD score interval mapping profiles for 12 rice chromosomes based on recombinant inbred line data from the study of Champoux *et al.* (1995). LOD scores are computed every 2 cM across each chromosome beginning at the left marker.

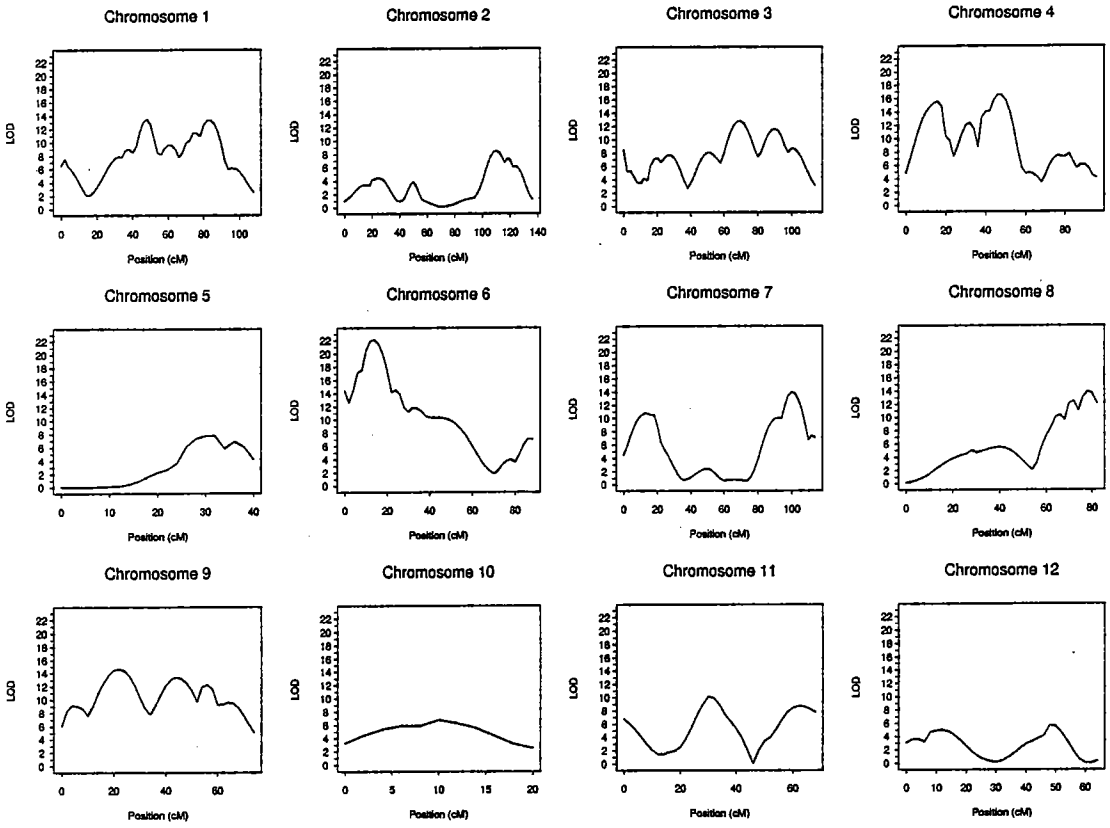

Figure 2: LOD score interval mapping profiles for each of four simulated chromosomes from a backcross experiment comprised of 100 individuals, and total of 46 markers arbitrarily positioned throughout the genome. The first QTL is 62 cM from the left end of chromosome 1, the second QTL is 44 cM from the left end of chromosome 2, the third and fourth QTL are 17 cM and 147 cM, respectively from the left end of chromosome 3, and chromosome 4 contains no QTL.