STATISTICAL THRESHOLD VALUES FOR LOCATING QUANTITATIVE TRAIT LOCI

R. W. Doerge

Recommended Citation

STATISTICAL THRESHOLD VALUES FOR LOCATING QUANTITATIVE TRAIT LOCI

R.W. Doerge

Departments of Agronomy and Statistics, 1399 Mathematical Science Building

Purdue University, West Lafayette, IN 47907-1399

Abstract

The detection and location of quantitative trait loci (QTL) that control quantitative characters is a problem of great interest to the genetic mapping community. Interval mapping has proved to be a useful tool in locating QTL, but has recently been challenged by faster, more sophisticated regression methods (*e.g.*, composite interval mapping). Regardless of the method used to locate QTL, the distribution of the test statistic (LOD score or likelihood ratio test) is unknown. Due to the quantitative trait values following a mixture distribution rather than a single distribution, the asymptotic distribution of the test statistic is not from a standard family, such as chi-square. The purpose of this work is to introduce interval mapping, discuss the distribution of the resulting test statistic, and then present empirical threshold values for the declaration of major QTL, as well as minor QTL. Empirical threshold values are obtained by permuting the actual experimental trait data, under a fixed and known genetic map, for the purpose of representing the distribution of the test statistic under the null hypothesis of no QTL effect. Not only is a permutation test statistically justified in this case, the test reflects the specifics of the experimental situation under investigation (*i.e.*, sample size, marker density, skewing, etc.), and may be used in a conditional sense to derive thresholds for minor QTL once a major effect has been determined.

1   Introduction

Quantitative trait mapping employs statistical genetics in conjunction with recent advances in biotechnology toward the goal of identifying statistically significant regions of various genomes associated with known diseases, resistances, and a variety of other quantitative traits. The specific genomic regions are referred to as quantitative trait loci (QTL), and the procedure by which the regions are identified known as QTL mapping (Lander and Botstein 1989). In human genetics the identification of QTL leads potentially to our understanding of the underlying genetic architecture of a disease, while in agricultural crops, the identified regions may lead to the introgression of certain regions (QTL) responsible for disease resistance into the same or closely related species. Regardless of the genome under investigation, most of the associations made between genomes and measurable traits are performed using statistical testing, and herein are the problems that will be discussed.

Standard interval mapping (Lander and Botstein 1989) techniques rely on a previously estimated genetic map from the same experimental data. Intervals of adjacent genetic markers are considered in turn, as follows: at prespecified increments (usually 2 cM), the process travels along the known distance between two genetic markers testing whether or not there is a QTL linked (associated) to the current position. In this paper, we will concentrate on one rather large issue, namely the criterion by which a statistical test is declared significant within a QTL mapping framework. QTL mapping, like many multiple testing situations, depends on a large number of tests, most of which are not independent. Making the situation even more difficult is our inability to assess the dependence structure of the problem, other than at the the extremes of the spectrum (*i.e.*, sparse and dense genetic maps). It is standard practice in QTL experiments to have upwards of 300 markers. If we assume 10 linkage groups (*e.g.*, which may be thought of as chromosomes) of approximate length 200 cM, having on average 30 evenly spaced genetic markers per group, each interval will be approximately 7 cM. If we test for a QTL every 2 cM within each interval, 5 tests (including at the markers) will be performed for each interval. In total, 123 dependent tests per linkage

group will be made, making an overall total of 1230 tests across the genome. Clearly, there are problems with multiple testing. In addition to the issues involved in multiple testing, there are issues pertaining to the distribution of the test statistic under an appropriate null hypothesis. Quite often the regularity conditions that ensure an asymptotic chi-square distribution for the likelihood ratio test statistic (or in the case of Lander and Botstein interval mapping, a LOD score) are not satisfied (Ghosh and Sen 1985; Hartigan 1985; Feng 1990). The difficulties in determining an appropriate asymptotic distribution are not surprising considering that the most popular hypothesis test compares a mixture distribution under the alternative hypothesis to a non-mixture distribution under the null hypothesis. Such situations are well cited in the literature (*e.g.*, Hartigan 1985; Ghosh and Sen 1985) and it is well known that the usual asymptotic arguments do not apply in this setting as the hypotheses are not appropriately nested. Amplifying these problems even further are issues of finite sample sizes, and distributional properties of the quantitative trait that may cause one to question the reliability of any asymptotic approximation. Certainly, asymptotic theories (Lander and Botstein 1989; Feingold *et al*, 1993; Dupuis 1994; Rebaï *et al.* 1994; Kruglyak and Lander 1995) have been presented for the determination of critical (threshold) values used to declare significant QTL. However these theories do little to incorporate the realities of the experimental situations.

The problem of determining appropriate threshold values is complicated by many experimental factors that vary, and that can, in effect, influence the distribution of the test statistic. Sample size, genome size, marker density, patterns of genetic marker data, missing data (both phenotypic and genotypic) are among the factors that often control the distribution of the test statistic. The purpose of this paper is to review a permutation theory method for estimating threshold values that is currently being used by the QTL community. Permutation theory is used to incorporate both the statistical issues, as well as the experimental issues for the purpose of developing an empirically based threshold that can be used in a variety of applications across all QTL experiments.

2 Threshold Values

2.1 Bonferroni Critical Values

Many researchers rely on a Bonferroni correction for multiple comparisons when testing for QTL. As such, if m independent tests with significance level α are conducted, the probability that at least one test shows a false positive is

$$1 - (1 - \alpha)^m$$

In the previously described situation of 300 markers on 10 linkage groups, setting $\alpha = 0.05$ for each test involved in say, interval mapping, would provide a probability of 1.0 for at least one false positive. QTL mapping at the markers (known as single marker QTL analysis) or using interval mapping under a Bonferroni correction for multiple testing guarantees false positives. Making matters worse, the Bonferroni correction is only appropriate for unlinked (independent) markers. Typically, the whole reason for interval mapping is to exploit the known (dependent) relationships between the genetic markers.

2.2 Monte Carlo Critical Values

Resampling techniques have been employed as a more robust means for obtaining overall significance levels in QTL mapping. Resampling phenotypes (the measurable characteristic per individual) and/or genotypes (the observable state of a genetic marker) across individuals provides a means of assessing statistical significance for any QTL mapping procedure by developing the distribution of the test statistic from each continual resampling. The key issue in resampling is whether to perform the resampling with or without replacement. When resampling is performed with replacement, this technique is known as bootstrapping (Efron 1979). When resampling is without replacement, the method is known as permuting (Fisher 1935; Good 1994). We discuss the implications of the bootstrap on QTL mapping here, while deferring the discussion of permutations until the next section.

Since a bootstrap sample is derived based upon a resampling procedure that allows individuals to be selected numerous times, when applied to QTL mapping the underlying genetic structure is destroyed. The estimated genetic map that acts as the foundation for interval mapping analyses relies on the dependence structure between genetic markers. When resampling is performed with replacement, the dependence structure changes with each resampling, thus changing the distribution of the test statistic for each bootstrap sample. Essentially, a bootstrap sample destroys all underlying statistical and genetic specificities of the experiment (*e.g.*, sample size, marker density, genetic map, environmental effects). Bootstrap tests in the QTL mapping situation are asymptotically equivalent to permutation tests. However in finite samples, the bootstrap cannot be guaranteed to provide a conservative type I error rate, nor can the bootstrap be guaranteed to be most powerful.

2.3   Permutation Critical Values

Permutation theory is a Monte Carlo resampling procedure that involves sampling without replacement. Originating from Fisher (1935) permutation involves the repeated shuffling of the dependent variable and the generation of a random sample of the test statistic from an appropriate null distribution. The procedure is statistically valid when used in conjunction with likelihood or regression based test statistics and for any distribution of the dependent variable. As stated by R.A. Fisher, permutations are time consuming, yet extremely useful.

> "Actually, the statistician does not carry out this very tedious process but his conclusions have no justification beyond the fact they could have been arrived at by this very elementary method."
>
> R.A. Fisher, 1936 on permutation tests.

With the state of modern computing as it is, permutation theory is now a viable means of assessing irregularities of any test statistic distribution that otherwise would be intractable. Permutation testing lends itself well to parallel processing, meaning that with multiple processors <u>many</u> permutations and analyses can be performed simultaneously. In the most

Applied Statistics in Agriculture

futuristic case, 1000 permutations can be accomplished in the same (wall clock) time as one permutation.

As mentioned previously, data for quantitative trait mapping consists of a set of genetic marker genotypes measured on each individual together with trait measurements on each individual. To detect QTL effects in the genome, statistical tests may be carried out at each marker, or for intervals between ordered markers. If there is a QTL effect at a specific location in the genome there will be an association between the trait values and the analysis point linked to that location. This association is typically detected by a range of statistical tests (t-test, regression, likelihood ratio, LOD score). If, in fact, there is no QTL effect linked to a test location in the genome, any apparent association between the trait values and the position is likely to be attributable to random chance.

In the QTL application, permutation of the trait values involves simply the reassignment of trait measurements to individuals while retaining the underlying genetic structure (*i.e.*, the genetic map). If the original data indicate that there are QTL effects, shuffling the trait measurements across individuals would effectively destroy any association between the trait measurements and the testing position linked to the QTL. If in fact, there are no QTL effects linked to the testing position in the genome, random reassignment of the trait values changes nothing. When we compute the value of the appropriate test statistic at each testing position in the shuffled data sets, we are basically sampling from the null distribution corresponding to the hypothesis of no associations between the trait values and the genetic map. Because the genetic map, as well as the trait values, is not altered during the permutation process this estimated distribution of the test statistic will take into account the particulars of the experiment at hand.

The details of permutation as applied directly to the QTL problem are as follows. Individuals in an experiment are indexed from $1, \ldots, n$. The data are shuffled by computing a random permutation of the indices and assigning the ith trait value to the individual whose index is given by the ith element of the permutation. The new (shuffled) data set is then

analyzed for QTL effects. The resulting test statistics are stored and the entire process (shuffling and analysis) repeated N times. At the end of this procedure, test statistics for N QTL analyses will have been stored. Essentially, the stored analyses represent a random sample of test statistics from the null distribution. When the N test statistics are ordered (from smallest to largest), in a particular fashion, various threshold values can be determined that correspond to the desired type I error rate. Churchill and Doerge (1994) suggested three types of critical (threshold) values to be estimated from the stored results. The first threshold is comparisonwise, and can be estimated separately for each analysis point in the genome. The comparisonwise thresholds provide $100(1-\alpha)$ critical values for the test at each analysis point. The second type of threshold is the chromosomewise threshold value, one for each chromosome or linkage group being analyzed. Based upon the N permuted data analyses, the maximum test statistic from each analysis (for each chromosome) is saved and ordered for the purpose of obtaining $100(1-\alpha)$ critical values for each chromosome. Finally, the third type of critical value suggested by Churchill and Doerge (1994) is an experimentwise threshold that provides an overall $100(1-\alpha)$ critical value which is valid simultaneously for all analysis points. To obtain an experimentwise threshold the maximum test statistic from each of the N permuted data sets are ordered, and then the threshold determined. Results from the original QTL analysis can be compared to any of these critical values to determine the statistical significance and thus to detect QTL effects for single traits.

Doerge and Churchill (1996) gave two extensions of the permutation based method for estimating empirical threshold values. Conditional empirical thresholds and residual empirical thresholds were introduced for the purpose of yielding critical values to be used to construct tests for the presence of minor QTL effects while accounting for effects of known major QTL (Churchill and Doerge 1994).

2.4   Confidence Intervals for Permutation P-Values

The most recent work on threshold values as applied to QTL mapping has been by Nettleton and Doerge (1998) and includes confidence intervals for the thresholds and p-values estimated in the permutation testing procedures described previously. The confidence intervals that Nettleton and Doerge propose account for the Monte Carlo error associated with the permutation testing, as well as providing tighter control on type I and type II errors in QTL analyses. An added benefit of the work on confidence intervals is to provide for an effective means for determining an efficient permutation sample size (N).

3   Data Analysis and Results

Permutation threshold values have become the standard in the QTL literature. Many publically available software packages have incorporated the ideas presented here, making it easy for researchers to apply these techniques to their experimental data. Comparisons have been made between theoretical (asymptotic based) thresholds and empirical (permutation) thresholds. Generally, these comparisons have been made on both simulated and real experimental data. Since simulated data are generally perfect, with no missing data, understood quantitative trait distributions, and large sample sizes, evidence for a large difference between theoretical and empirical thresholds is not strong. However, when real data are analyzed, the difference between empirical and theoretical thresholds widens as the effects of experimental specifics become more obvious.

Doerge and Rebaï (1996) presented a comparison between empirical and theoretical (Dupuis 1994; Rebaï *et al.* 1994) thresholds using an actual QTL mapping data set from an F_2 maize population consisting of 106 individuals. The QTL analysis was specific to chromosome 2 (estimated length 132.8 cM), with 12 RFLP markers. The distribution (cM) of these 12 markers was: {7.8 7.9 9.7 7.1 3.1 9.0 42.6 5.3 6.5 18.2 15.5}, where each value in this array represents the interval between adjacent markers. On average 20% of the marker data

were missing (*i.e.*, individuals with at least one missing marker), while less than 7% of the trait data were missing (*i.e.*, individuals with a missing trait measurement, irrespective of marker status). The distribution of the trait values appeared to follow a normal distribution. Thresholds for a chromosomewise (*i.e.*, specific to only one chromosome) significance level of 5% were calculated using methods of Churchill and Doerge (1994), Rebaï *et al.* (1994) and Dupuis (1994). Empirical threshold values were based on 1,000 permutations of the actual phenotypic measurements. In this specific example, where there was a group of individuals with missing trait measurements, the missing trait measurements were permuted along with the other trait measurements for the purpose of retaining the specifics of the experiment.

Since the empirical threshold values reflect the specifics of the data set it is not surprising that the magnitude of the threshold was somewhat smaller than both approximations. Using a LOD score as the test statistic, for a significance level of 5% the permutation threshold was 1.90, as compared to 2.30 when calculated using Rebaï, and 2.27 using Dupuis. When compared to threshold values based on a similar simulated F_2 population, the empirical threshold values for a real data set were smaller, while the magnitude of the analytical threshold values remained unchanged (*i.e.*, the analytic threshold values were the same for both real and simulated data). The differences between the threshold values as seen in this example using real data are most likely due to the proportion of missing marker data, as well as the environmental specifics of the experiment. For significance levels of 5% and 1%, the empirical permutation method performed at the specified significance levels, while the analytical methods of both Rebaï *et al.* and Dupuis provided conservative lower bounds (*i.e.*, less than α).

4   Summary

There are many benefits to each of the previously mentioned approaches to threshold values, and certainly there are other theoretical methods by which thresholds can be approximated. While no one threshold value can be exact and true, each threshold, if used in an informed

manner may provide an appropriate value against which to compare test statistics for significant QTL. The main differences between threshold values that are not empirically based are that simulation based thresholds and analytic thresholds are model based, and do little to account for effects of missing data patterns and multiple QTL interactions. Permutation testing, on the other hand, provides an easy to use method for estimating threshold values that are statistically sound, are robust to departures from standard assumptions and are tailored specifically for the experiment at hand.

5 Cited Literature

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

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

Doerge, R.W. and A. Rebaï. 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.

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

Efron, B. 1979. Bootstrap method: another look at the jackknife. Ann. Stat. 7:1-26.

Feng, Z. 1990. Statistical Inference Using Maximum Likelihood Estimation and the Generalized Likelihood Ratio Under Nonstandard Conditions. Cornell University. Ph.D. Dissertation.

Fisher, R.A. 1935. *The Design of Experiments*, Ed. 3, Oliver and Boyd Ltd., London.

Ghosh. J.K. and P.K. Sen. 1985. On the asymptotic performance of the log likelihood ratio statistic for the mixture model and related results. Proc. of the Berkeley Conf.. Vol. II. Eds. Lucien M. LeCam and Richard A. Olshen. pp.789–807.

Good, P. 1994. *Permutation Tests: A Practical Guide to Resampling for Testing Hypotheses.* Springer-Verlag, New York.

Hartigan, J.A. 1985. A failure of likelihood asymptotics for normal distributions. Proc. of the Berkeley Conf., Vol. II. Eds. Lucien M. LeCam and Richard A. Olshen. pp.807–810.

Kruglyak, L. and E.S. Lander. 1995. A nonparametric approach for mapping quantitative trait loci. Genetics 139:1421-1428.

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

Nettleton, D. and R.W. Doerge. 1998. Accounting for variability in the use of permutation testing to detect quantitative trait loci. Biometrics (accepted).

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