A Note on Thresholds for QTL Interval Mapping Tests

R.W. Doerge* and Ahmed Rebaï†

BU-1287-M May 1995

* Biometrics Unit, 337 Warren Hall, Cornell University, Ithaca, NewYork, USA, 14853

† INRA, Centre de Toulouse, Unité de Biometrie et d'Intelligence Artificielle, Auzeville B.P. 27, 31326 Castanet-Tolosan, France

Keywords: threshold values, interval mapping, QTL, molecular marker.

Abstract

The problem of mapping quantitative trait loci (QTL) using genetic marker information is of great interest to the mapping community. There are many statistical methods available for detecting and/or locating QTL, all of which depend on assumptions about the distribution of the quantitative trait values. The distribution of the trait values is affected by sample size, genetic marker density, missing data patterns, environmental noise, etc., all of which affect the distribution of the test statistic used to detect/locate QTL. Failure of the test statistic distribution to follow a standard statistical distribution is the subject of current research. It is necessary to understand the behavior of the test statistic under the null hypothesis so that a critical value may be obtained for the purpose of declaring the presence of a QTL. In this paper we discuss the choices available for obtaining critical values (threshold values) for QTL detection tests using interval mapping procedures. We investigate the effect of deviations from normality of the trait values on the threshold value by comparing analytical approximations to empirical threshold values for simulated backcross and F_2 populations, along with an experimental F_2 population.

Introduction

The mapping of quantitative trait loci (QTL) using information from pairs of linked genetic markers (interval mapping) has received a great deal of attention and has been applied successfully by both plant and animals breeders, as well as geneticists. The basic approach of interval mapping (Lander and Botstein 1989, 1994) has been further generalized by a number of authors (*e.g.* Haley and Knott 1992, Haley *et al.* 1994, Jansen 1994, Jansen and Stam 1994, Rebaï *et al.* 1994a, 1995, Zeng 1993, 1994) to allow the presence of QTL to be tested at every location in a genome for a wide variety of segregating populations by exploiting the full power of high density genetic linkage maps. Recent (Lander and Botstein 1989, 1994, Feingold *et al.* 1993, Dupuis 1994, Rebaï *et al.* 1994b, Churchill and Doerge 1994, Kruglyak and Lander 1995) research on the determination of threshold values used to declare significant QTL has provided the mapping community with both theoretical and empirical threshold values. Each of these efforts recognizes the importance of working with an accurate threshold value, so that progress may continue in the area of QTL detection and location.

The purpose of this paper is to consider the choices (Lander and Botstein 1989, 1994, Feingold *et al.* 1993, Dupuis 1994, Rebaï *et al.* 1994b, Churchill and Doerge 1994, Kruglyak and Lander 1995) available for obtaining threshold values for QTL detection tests via interval mapping, and to discuss their adequacy and practical use. We investigate the effect of deviations from normality of the sample trait values on the threshold value by comparing the analytical approximations and the empirical thresholds based on permutation tests for simulated backcross and F_2 populations, along with an F_2 experimental maize population.

Threshold values

Reliability of QTL detection (control of the false positive rate) is an important problem, which has motivated many simulation based investigations, along with analytical approximations (Feingold *et al.* 1993, Dupuis 1994, Rebaï *et al.* 1994b), as well as empirical methods. In the interval mapping approach a likelihood ratio (or equivalent) test denoted $T(x)$ is performed at every position x (in practice each 1 cM) of a chromosome and a QTL is declared present if the supremum of the test values exceeds a predetermined threshold anywhere on the chromosome or genome. This chromosomewise threshold t is calculated so that for a given per chromosome significance level α we have:

$$\alpha = \Pr(sup_{0 \leq x \leq L} T(x) \geq t)$$

where L is the length of the chromosome in Morgans. A number of approximations have been derived to have analytical equations which permit an easy computation of the threshold t for any significance level α.

Lander and Botstein (1989, 1994) use the asymptotic distribution of the test statistic (LOD score) based on an infinitely dense marker map and the equation (backcross population, single chromosome): $\alpha \approx (1+2Lt)\chi^2(t)$ where χ^2 is the inverse cumulative distribution function of a χ^2 with one degree of freedom.

Feingold *et al.* (1993), Dupuis (1994), and Rebaï *et al.* (1994b) approximations are based on the asymptotic distributional properties of the stochastic process generated by performing the interval mapping test at each position, although the Rebaï *et al.* derivation assumes a finite number of markers (intermediate-map density). Equations for these approaches (for backcross and F_2) are found in Dupuis (1994) and Rebaï *et al.* (1994b).

An empirical approach based on permutation theory (Fisher 1935) de-

veloped by Churchill and Doerge (1994) samples the distribution of the test statistic (under the null hypothesis of no QTL) by shuffling and then analyzing the phenotypic data, under a known fixed genetic map, for the purpose of destroying any genotypic–phenotypic correlation caused by a QTL. This process is repeated numerous times so that the distribution of the test statistic may be randomly sampled and then used to obtain a threshold value. Permutation based methods have the advantage of being distribution free, as they take into account the actual distribution of the trait being studied, and are not limited by experimental design.

Backcross and related populations

Populations where the QTL effect is characterized by a single parameter such as backcross, doubled haploid lines or recombinant inbreds (although there is a slight difference due to recombination) are of interest to the mapping community. In cases such as these, the QTL effect is described by the effect of an allelic substitution. We simulated backcross data (under the null hypothesis of no QTL present) so that the distance between the markers of the chromosome were randomly generated to ensure a length in cM and an average marker density close to the desired one. Based on complete marker and trait data two population sizes of 100 and 200 individuals were considered. Both normally distributed trait data and gamma distributed (Gamma(1,2)) trait data were simulated. The gamma distribution represents the situation of extreme skewing in the trait data, creating a long right tail in the distribution. We calculated chromosomewise threshold values for different chromosome lengths and marker densities at 5% and 1% significance levels using the interval mapping test as described by Lander and Botstein (1989, 1994), based on the

approximation given by Rebaï *et al.* (1994b), and the empirical approach proposed by Churchill and Doerge (1994) with 1,000 permutations.

Results of the comparison are given in Table 1. For normally distributed traits or large sample sizes (so that the convergence of the test statistics is guaranteed) the analytical approximations proposed by Rebaï *et al.* (1994b) for medium marker densities (more than 10 cM), and Lander and Botstein (1989) and Feingold *et al.* (1993) (results not shown) for infinitely dense maps (say less than 10 cM) give threshold values which are very close to those obtained by permutations. The threshold value obtained using the Lander and Botstein's (1989, 1994) proposition 2 for one chromosome is also given in Table 1. The Lander and Botstein threshold provides an upper bound of the actual threshold (as it assumes an infinite information) and gives a conservative test which ensures the type I error to be less than the significance level. *(place Table 1 here)*

F_2 populations

An F_2 population has two parameters which characterize the additive and dominance action of the QTL alleles, unless an additive model is assumed. This characterization makes the covariance of the test process difficult to compute (Dupuis 1994, Kruglyak and Lander 1995). Two analytical approximations are available, one from Dupuis (1994) based on the same approach as that of Feingold *et al.* (1993) and one from Rebaï *et al.* (1994b).

We consider the same situations as in the previous section, comparing thresholds from Churchill and Doerge (1994), Rebaï *et al.* (1994b) and Dupuis (1994). Results are given in Table 2. The Rebaï *et al.* approximation is close to the empirical threshold, with the Dupuis approximation being slightly smaller. *(place Table 2 here)*

Maize F_2 population

We have also computed empirical and approximate threshold values for experimental data from an F_2 maize population with 106 individuals. The estimated length of chromosome 2 is 132.8 cM using 12 RFLP markers. The distribution (cM) of these 12 markers is: {7.8 7.9 9.7 7.1 3.1 9 42.6 5.3 6.5 18.2 15.5}. On average 20% of the marker data are missing, while less than 7% of the trait data are missing. The results are shown (Table 3) for a chromosomewise significance level of 5% (1%). Empirical threshold values are based on 1,000 permutations. *(place Table 3 here)*

Since the empirical threshold values reflect the specifics of the data set it is not surprising that the magnitude of the values is somewhat smaller than both approximations. When compared to simulated F_2 threshold values (Table 2), the empirical threshold values for a real data set are smaller, while the magnitude of the analytical threshold values remains unchanged. The differences between the threshold values as seen in this example are most likely due to the proportion of missing marker data, as well as the environmental specifics of the experiment.

Discussion

Deviations from normality of the trait distribution and sample size are both factors which affect the distribution of the test statistic (in this situation the LOD score), and ultimately affect the threshold level of the interval mapping tests used in QTL detection. When trait distributions deviate from normality and/or the sample sizes are small, approximate values based on the asymptotic distribution properties of the test statistics are not appropriate, and empirical approaches should be used. The results of this paper support the findings that

even if the assumptions do not hold (skewed distribution), the approximations behave quite well. These findings are probably related to the robustness of interval mapping to deviations from normality (Cierco, personal communication to A.R.). In practice, one can see the benefits of using either analytical or empirical methods for obtaining threshold values.

The values obtained by the approximations proposed by Rebaï *et al.* (1994b) are appropriate for intermediate density map (a marker every 10 cM or more), and the others (Lander and Botstein 1989, Feingold *et al.* 1993, Dupuis 1994) appropriate for high density maps (a marker every 10 cM or less). These thresholds (see previous citation) provide stringent values that ensure the type I error to be less than the significance level chosen by the user (conservative). The values obtained are appropriate for the standard interval mapping approach but would be usable, under some conditions, for the multiple QTL approach proposed by Jansen and Stam (1994) (see Jansen 1994) and could be applied after some specific calculations to nonparametric tests of interval mapping (Kruglyak and Lander 1995).

Kruglyak and Lander (1995) recommended that the dense-map threshold always be used, regardless to of the actual density of the map, in order to minimize the false positive rate. However, the use of specific approximations as proposed by Rebaï *et al.* (1994b) will give more appropriate threshold for intermediate dense-maps without the loss in power of the tests consistent with the use of a stringent threshold. The asymptotic approximations based on distributional properties of stochastic processes (Feingold *et al.* 1993, Dupuis 1994, Rebaï *et al.* 1994b, Mangin *et al.* 1994) are with no doubt a powerful tool for further analytical investigations of the threshold problem (especially when mapping multiple QTL), as well as other developments for QTL parameters (location and effect).

The empirical threshold values obtained by permutation theory, while computationally intensive, may be calculated for any experimental design under an unlimited number of experimental situations (*e.g.* sample size, marker density, environmental variance, nonnormal trait distribution, etc.). The number of permutations used in each application of this paper was limited to 1,000. Upwards of 10,000 permutations are more appropriate if an accurate 1% threshold value is desired.

Missing data, either genotypic or phenotypic, greatly influences the quality of the parameter estimates (*e.g.* recombination, additive effects, dominance effects, etc.). Each of the analytical methods discussed in this report are based upon perfect data, no account is made for missing data. While perfect data is a realistic approach through simulation, it is rarely obtainable experimentally. The difference in the magnitude of threshold values (empirical versus analytical) as seen in the maize example is most likely due to the percentage of missing data per marker scored.

The QTL mapping community continues to bring challenging problems to the forefront of QTL research, and while there is no one correct threshold value to use in every situation, it is our long term hope that the comparisons made in this paper will serve as direction to the application and conclusions drawn. Certainly, as the envelop of QTL detection/location is pushed to include multiple QTL, interactions, and fine scale location of QTL, statistical issues relating the relevance of application to the conclusions drawn still await proper statistical attention.

Acknowledgements

We are grateful to Rustica Prograin Génétique (France) for providing the maize F_2 data set. RWD acknowledges financial support from the National Initiative Competitive Grants Program of the United States Department of Agriculture, Award 94-37300-0323.

References

CHURCHILL, G.A. AND DOERGE, R.W. 1994. Empirical threshold values for quantitative trait mapping. *Genetics*, **138,** 963-971.

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.

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

HALEY, C.S. AND KNOTT, S.A. 1992. A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Heredity*, **69,** 315-324.

HALEY, C.S., KNOTT, S.A. AND ELSEN, J.M. 1994. Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics*, **136,** 1195-1207.

JANSEN, R.C. AND STAM, P. 1994. High resolution of quantitative traits into multiple loci via interval mapping. *Genetics*, **136,** 1447-1455.

JANSEN, R.C. 1994. Controlling the type I and type II errors in Mapping Quantitative trait loci. *Genetics*, **138,** 871-881.

KRUGLYAK, L. AND LANDER, E.S. 1995. A nonparametric approach for mapping quantitative trait loci. *Genetics*, **139,** 1421-1428.

LANDER, E.S. AND BOTSTEIN, D. 1989. Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **121,** 185-199.

LANDER, E.S. AND BOTSTEIN, D. 1994. Corrigendum. *Genetics*, **36,** 705.

MANGIN, B., GOFFINET, B. AND REBAÏ, A. 1994. Constructing confidence

intervals for QTL location. *Genetics*, **138,** 1301-1308.

REBAÏ, A., GOFFINET, B. AND MANGIN, B. 1995. Comparing powers of different methods for QTL detection. *Biometrics*, **51,** 87-99.

REBAÏ, A., B. GOFFINET, B. MANGIN, AND D. PERRET, 1994a QTL detection with diallel schemes. in *proceedings of the ninth Eucarpia meeting. Molecular markers in plant breeding*, Wageningen p. 170-177.

REBAÏ, A., GOFFINET, B. AND MANGIN, B. 1994b. Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138,** 235-240.

ZENG, Z-B., 1993. Theoretical basis for separation of multiple linked gene effects in mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA. **90,** 10972-10976.

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

Table 1: Comparison of empirical and approximate threshold values for different marker densities and chromosome lengths in simulated backcross populations.

Sample size	cM [a]	Nb. markers	Empirical [b]	Approx. [c]	LB Approx. [d]
Normally distributed trait					
100	82.3	6	1.45 [e] (2.20[f])	1.47 (2.15)	1.87 (2.65)
200	100.2	6	1.53 (2.26)	1.50 (2.18)	1.97 (2.74)
100	90.6	9	1.58 (2.27)	1.57 (2.26)	1.92 (2.69)
200	106.7	9	1.58 (2.29)	1.61 (2.12)	1.99 (2.77)
100	221.4	11	1.71 (2.25)	1.77 (2.47)	2.34 (3.10)
200	203.4	11	1.56 (2.21)	1.76 (2.45)	2.30 (3.06)
Skewed trait distribution [g]					
100	88.5	6	1.41 (1.95)	1.47 (2.16)	1.91 (2.68)
200	89.4	6	1.40 (1.99)	1.49 (2.17)	1.92 (2.69)

[a] length of chromosome

[b] Churchill and Doerge 1994

[c] Rebaï *et al.* 1994

[d] Lander and Botstein 1989, 1994

[e] 5% threshold value

[f] 1% threshold value

[g] Gamma(1,2)

Table 2: Comparison of empirical and approximate threshold values for different marker densities and chromosome lengths in simulated F_2 populations.

Sample size	cM [a]	Nb. markers	Empirical [b]	Approx. [c]	Dupuis Approx. [d]
Normally distributed trait					
100	107.8	6	2.10 [e] (2.90 [f])	2.12 (2.87)	2.00 (2.69)
200	103.8	6	2.10 (2.80)	2.11 (2.87)	2.00 (2.69)
100	119.3	9	2.30 (3.20)	2.26 (3.02)	2.16 (2.87)
200	97.3	9	2.20 (2.80)	2.24 (2.99)	2.15 (2.86)
100	188.1	11	2.40 (3.30)	2.39 (3.14)	2.25 (2.94)
200	201.3	11	2.50 (3.20)	2.40 (3.15)	2.25 (2.94)
Skewed trait distribution [g]					
100	94.9	6	2.10 (3.20)	2.10 (2.86)	2.00 (2.69)
200	104.7	6	2.10 (2.90)	2.11 (2.87)	2.00 (2.70)

[a] length of chromosome

[b] Churchill and Doerge 1994

[c] Rebaï *et al.* 1994

[d] Dupuis 1993

[e] 5% threshold value

[f] 1% threshold value

[g] Gamma(1,2)

Table 3: Comparison of empirical and approximate threshold values for 12 markers on a single maize chromosome of length 132.8 cM and 106 F_2 individuals.

Empirical [a]	Approx. [b]	Dupuis Approx. [c]
1.90 [d] (2.60 [e])	2.30 (3.06)	2.27 (2.98)

[a] Churchill and Doerge 1994

[b] Rebaï *et al.* 1994

[c] Dupuis 1994

[d] 5% threshold value

[e] 1% threshold value