

Significance thresholds for QTL interval mapping tests

R. W. DOERGE* & AHMED REBAÏ†

*Departments of Statistics and Agronomy, 1399 Mathematical Sciences Building, Purdue University, West Lafayette, IN 47907-1399, U.S.A. and †INRA, Centre de Toulouse, Unité de Biometrie et d'Intelligence Artificielle, Auzeville B.P. 27, 31326 Castanet-Tolosan, France

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. In order to declare a significant QTL effect it is necessary first to understand the behaviour of the test statistic under the null hypothesis so that a critical value may be employed. In this paper we discuss the choices available for obtaining critical values (threshold values) used in locating QTL via interval mapping procedures. We investigate threshold values obtained by different means (analytical approximations and empirical) for the same level of significance (type I error rate) under a normality assumption (null hypothesis of no QTL). In addition, we explore the effect of deviations from normality of the trait values on the threshold value by comparing analytical approximations and empirical threshold values for simulated backcross and F_2 experiments, along with an actual experimental F_2 data set.

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

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 animal breeders, as well as geneticists. The basic approach of interval mapping (Lander & Botstein, 1989, 1994) has been further generalized by a number of authors (e.g. Haley & Knott, 1992; Zeng, 1993, 1994; Haley *et al.*, 1994; Jansen, 1994; Jansen & Stam, 1994; Rebaï *et al.*, 1994a, 1995) 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 research (Lander & Botstein, 1989, 1994; Feingold *et al.*, 1993; Churchill & Doerge, 1994; Dupuis, 1994;

Rebaï *et al.*, 1994b; Kruglyak & Lander, 1995) 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 & Botstein, 1989, 1994; Feingold *et al.*, 1993; Churchill & Doerge, 1994; Dupuis, 1994; Rebaï *et al.*, 1994b; Kruglyak & Lander, 1995) available for obtaining threshold values for QTL detection tests via interval mapping, and to discuss their adequacy and practical use. A review of interval mapping using LOD score test statistics can be found in Lander & Botstein (1989). We investigate the effect of deviations from normality of the sample trait values on the threshold value using LOD scores by comparing the analytical approximations and the empirical thresholds based on permutation tests for

*Correspondence.

simulated backcross and F_2 populations, along with an F_2 experimental maize population.

Threshold values

Proper detection of significant QTL (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 (Churchill & Doerge, 1994). 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. A 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 . We will compare the performance of different analytical threshold values (Lander & Botstein, 1989, 1994; Feingold *et al.*, 1993; Dupuis, 1994; Rebaï *et al.*, 1994b) with each other, as well as with the empirical threshold values of Churchill & Doerge (1994).

Lander & 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 $\chi^2(t)$ is the inverse cumulative distribution function of a χ^2 with one degree of freedom.

The approximations of Feingold *et al.* (1993), Dupuis (1994) and Rebaï *et al.* (1994b) 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) developed by Churchill & Doerge (1994), samples the distribution of the test statistic (under the null hypothesis of no QTL) by shuffling the original phenotypic data under a known fixed genetic map, for the purpose of destroying any genotypic-phenotypic correlation caused by a QTL,

and then analysing the (new) phenotypic data. This resampling process (without replacement) is repeated numerous times so that the distribution of the test statistic (under the null hypothesis of no QTL) may be randomly sampled and then used to obtain a threshold value at a user-specified significance level α . Permutation-based methods have the advantage of being distribution free, and thus they do not depend upon a model to describe the actual trait being studied, and are not limited by experimental design.

The accuracy of each threshold method may be evaluated by comparing each respective threshold value to the maximum test statistic resulting from the analysis of many resamplings from the null distribution of no QTL, and then calculating the proportion of times the maximum test statistic is greater than the threshold value under investigation. Since threshold values are calculated according to some specified significance level, α , the proportion of times the maximum test statistic is greater than the calculated threshold should be close to α . Analytical methods may be compared directly because they provide the proportion of maximum test statistics greater than the respective calculated threshold value. Intuitively, because the empirical threshold value is developed from a specified number of resamplings from the null distribution of the trait values, the proportion of times the maximum test statistic from each analysis is greater than the estimated empirical threshold value will be, by definition, less than or equal to α .

Backcross and related populations

Populations where the QTL effect is characterized by a single parameter (variance held constant) such as backcross, doubled haploid lines or recombinant inbreds (although there is a slight difference arising from a differential effect of 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 a genetic map and phenotypic measurements for each backcross experimental design presented in Table 1. Genetic markers were generated so that the distances between markers on a single chromosome were randomly simulated to ensure a length in cM and average marker density close to the one desired. Phenotypic trait values were simulated, under the null hypothesis of no QTL, from a standard normal distribution as well as from a gamma distribution (Gamma(1,2)). The normal distribution represents the 'perfect' data

Table 1 Comparison of empirical and approximate threshold values for different marker densities and chromosome lengths in simulated backcross experiments using the LOD score test statistic

n^a	cM ^b	m^c	Empirical ^d	$\hat{\alpha}_E^e$	R. approx. ^f	$\hat{\alpha}_R^g$	L.B. approx. ^h	$\hat{\alpha}_{LB}^i$
Normally distributed trait								
100	82.3	6	1.45 ^j (2.20 ^k)	0.05	1.47 (2.15)	0.05	1.87 (2.65)	0.02
200	100.2	6	1.53 (2.26)	0.05	1.50 (2.18)	0.05	1.97 (2.74)	0.02
100	90.6	9	1.58 (2.27)	0.05	1.57 (2.26)	0.05	1.92 (2.69)	0.03
200	106.7	9	1.58 (2.29)	0.05	1.61 (2.12)	0.05	1.99 (2.77)	0.02
100	221.4	11	1.71 (2.25)	0.05	1.77 (2.47)	0.05	2.34 (3.10)	0.01
200	203.4	11	1.56 (2.21)	0.05	1.76 (2.45)	0.04	2.30 (3.06)	0.05
Skewed trait distribution ^l								
100	88.5	6	1.41 (1.95)	0.05	1.47 (2.16)	0.04	1.91 (2.68)	0.01
200	89.4	6	1.40 (1.99)	0.05	1.49 (2.17)	0.04	1.92 (2.69)	0.01

^aSample size, ^blength of chromosome, ^cnumber of genetic markers, ^dChurchill & Doerge, 1994, ^e5% error rate using empirical threshold value based on 1000 samples from the null distribution, ^fRebäi *et al.*, 1994, ^g5% error rate using Rebäi *et al.*'s analytical threshold value based on 1000 samples from the null distribution, ^hLander & Botstein, 1989, 1994, ⁱ5% error rate using Lander & Botstein's analytical threshold value based on 1000 samples from the null distribution, ^j5% threshold value, ^k1% threshold value, ^lGamma (1,2).

situation having no outside effects from sample size, missing data, segregation distortion, etc. The gamma distribution represents some of these previously noted effects by presenting extreme skewing in the trait data, thus creating a long right tail in the distribution. Based on complete marker and trait data two sample sizes of 100 and 200 individuals were considered under an increasing number (6, 9, 11) of genetic markers. Chromosomewise threshold values were calculated for different chromosome lengths and marker densities at 5 per cent and 1 per cent significance levels using the LOD score test statistic in an interval mapping procedure described by Lander & Botstein (1989, 1994). Significance threshold values for each of the backcross experimental situations were calculated based on three methods: (i) the empirical approach proposed by Churchill & Doerge (1994) with 1,000 permutations, (ii) the analytical approximation given by Rebäi *et al.* (1994b) and (iii) the analytical threshold of Lander and Botstein (1989, 1994).

Results of the comparisons are given in Table 1. For normally distributed traits or a large sample size the analytical approximations proposed by Rebäi *et al.* (1994b) for medium marker densities (more than 10 cM), and by Lander & 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 in magnitude to those obtained by permutation (empirical threshold values). When 1000 random samples from the null

distribution are analysed and the maximum test statistic from each of the analyses compared with each of the threshold values being investigated (Table 1), the proportion of times the test statistic is greater than the respective threshold value provides an error rate for the method. The Rebäi *et al.* approximation achieves the appropriate significance level ($\alpha = 0.05$) in all but the last simulation, and as one would expect the empirical threshold value performs at the specified significance level. 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 that the type I error rate is less than the significance level.

In the limited simulation for skewed backcross data drawn from a gamma distribution, the Rebäi *et al.* threshold provides a conservative threshold value, and as mentioned previously the Lander and Botstein threshold provides a conservative upper bound on the actual threshold value. By definition, the empirical threshold method performs at the specified significance level.

F₂ populations

An F₂ 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 & 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). The empirical approach (Churchill & Doerge, 1994) remains as defined previously. We consider the same simulation set up as in the previous section under an F_2 experimental design for the purpose of comparing thresholds from (i) Churchill & Doerge (1994), (ii) Rebaï *et al.* (1994b) and (iii) Dupuis (1994).

Results for the F_2 simulation are given in Table 2. For normally distributed trait data the Rebaï *et al.* approximate threshold value is close to the empirical threshold, with the Dupuis approximation being slightly smaller in magnitude. When 1000 samples from the null distribution are analysed, and the maximum test statistic compared with the actual threshold value of each method, Table 2 shows the empirical threshold method achieving a type I error rate less than or equal to the specified significance level, and that both the Rebaï *et al.* and Dupuis thresholds fail to achieve the appropriate lower bound for all situations. Failure to achieve the appropriate lower bound means that the proportion of times 1000 random samples from the null distribution produce a test statistic larger than the respective threshold value is greater than the specified significance level used to calculate the analytical threshold value. When skewed-trait data (gamma distributed) are considered, both the empirical thresholds and the Rebaï *et al.* analytical thresholds perform well, however Dupuis's analytical threshold

error rate fails to achieve the specified significance level in each case.

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 per cent of the marker data are missing, whereas less than 7 per cent of the trait data are missing. The distribution of the trait values appears to follow a normal distribution. The results are shown (Table 3) for a chromosomewise significance level of 5 per cent (1 per cent). Empirical threshold values are based on 1000 permutations of the actual phenotypic traits.

Because 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, whereas the magnitude of the analytical threshold values remains unchanged. The differences between the threshold values as seen in this example most likely result from the proportion of missing marker data, as well as the environmental specifics of the experiment. For significance levels of 5 per cent and 1 per

Table 2 Comparison of empirical and approximate threshold values for different marker densities and chromosome lengths in simulated F_2 experiments using the LOD score test statistics

n^a	cM ^b	m^c	Empirical ^d	$\hat{\alpha}_E^e$	R. approx. ^f	$\hat{\alpha}_R^g$	D. approx. ^h	$\hat{\alpha}_D^i$
Normally distributed trait								
100	107.8	6	2.10 ^j (2.90 ^k)	0.04	2.12 (2.87)	0.04	2.00 (2.69)	0.06
200	103.8	6	2.10 (2.80)	0.05	2.11 (2.87)	0.05	2.00 (2.69)	0.06
100	119.3	9	2.30 (3.20)	0.05	2.26 (3.02)	0.06	2.16 (2.87)	0.06
200	97.3	9	2.20 (2.80)	0.04	2.24 (2.99)	0.04	2.15 (2.86)	0.05
100	188.1	11	2.40 (3.30)	0.05	2.39 (3.14)	0.06	2.25 (2.94)	0.07
200	201.3	11	2.50 (3.20)	0.04	2.40 (3.15)	0.05	2.25 (2.94)	0.08
Skewed trait distribution ^l								
100	94.9	6	2.10 (3.20)	0.04	2.10 (2.86)	0.04	2.00 (2.69)	0.05
200	104.7	6	2.10 (2.90)	0.05	2.11 (2.87)	0.05	2.00 (2.70)	0.06

^aSample size, ^blength of chromosome, ^cnumber of genetic markers, ^dChurchill & Doerge, 1994, ^e5% error rate using empirical threshold value based on 1000 samples from the null distribution, ^fRebaï *et al.*, 1994, ^g5% error rate using Rebaï *et al.*'s analytical threshold value based on 1000 samples from the null distribution, ^hDupuis, 1993, ⁱ5% error rate using Dupuis's threshold value based on 1000 samples from the null distribution, ^j5% threshold value, ^k1% threshold value, ^lGamma (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 using LOD score test statistics

Empirical ^a	$\hat{\alpha}_E^b$	R. approx. ^c	$\hat{\alpha}_R^d$	D. approx. ^e	$\hat{\alpha}_D^f$
1.90 ^g (2.60 ^h)	0.05 (0.01 ⁱ)	2.30 (3.06)	0.02 (0.005)	2.27 (2.98)	0.03 (0.005)

^aChurchill & Doerge, 1994, ^b5% error rate using empirical threshold value based on 1000 samples from the null distribution, ^cRabaï *et al.*, 1994, ^d5% error rate using Rabaï *et al.*'s threshold value based on 1000 samples from the null distribution, ^eDupuis, 1994, ^f5% error rate using Dupuis's threshold value based on 1000 samples from the null distribution, ^g5% threshold value, ^h1% threshold value, ⁱ1% error rate based on 1000 samples from the null distribution.

cent, Table 3 shows that the empirical method performs at the specified significance levels, while the analytical methods of both Rabaï *et al.* and Dupuis provide conservative lower bounds.

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 test 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 may not be appropriate, and empirical approaches should be considered. The results of this paper support the findings that even if the assumptions do not hold (skewed distribution), the analytical approximations behave relatively well. These findings are most likely related to the robustness of interval mapping to deviations from normality (C. 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 Rabaï *et al.* (1994b) are most appropriate for intermediate density maps (a marker every 10 cM or more), and the others (Lander & 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) should provide stringent values that ensure that the type I error rate is 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 condi-

tions, for the multiple QTL approach proposed by Jansen & Stam (1994) (see Jansen, 1994) and could be applied after some specific calculations to nonparametric tests of interval mapping (Kruglyak & Lander, 1995). Further work in each of these areas is needed in order to specify the specifics of each application.

Kruglyak & Lander (1995) recommended that the dense-map threshold always be used, regardless of the actual density of the map, in order to minimize the false positive rate. However, the use of specific approximations as proposed by Rabaï *et al.* (1994b) will give more appropriate thresholds for intermediate density 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; Mangin *et al.*, 1994; Rabaï *et al.*, 1994b) are without 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 of the actual phenotypic trait data, although computationally intensive, may be calculated for any experimental design under an unlimited number of experimental situations (e.g. sample size, marker density, environmental variance, non-normal trait distribution, etc.). No restrictions are placed on the marker density of the genetic map, and as the experimental trait values are permuted across the individuals in the sample, the specifics of the experiment are retained. The number of permutations used in each application of this paper was limited to 1000. Upwards of 10 000 permutations are more appropriate if an accurate 1 per cent threshold value is desired.

Missing data, either genotypic or phenotypic, greatly influence the quality of the parameter estimates (e.g. recombination, additive effects, dominance effects, etc.). Each of the analytical calculations discussed in this work is based upon perfect data; no account is made for missing data. Although perfect data represent a realistic approach through simulation, they are rarely obtainable experimentally. The difference in the magnitude of threshold values (empirical vs. analytical) as seen in the maize example most likely results from the percentage of missing data per marker scored, and is certainly worthy of further investigation.

The QTL mapping community continues to bring challenging problems to the forefront of QTL research, and although 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 a direction to the application and conclusions drawn. Certainly, as the scope of QTL detection/location is extended to include multiple QTL, interactions, and fine scale location of QTL, 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 F₂ data set. R.W.D. 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, Stanford University, Department of Statistics.

FEINGOLD, E., BROWN, P. O. AND 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*, 3rd edn. Oliver and Boyd, 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. 1994. Controlling the type I and type II errors in mapping quantitative trait loci. *Genetics*, **138**, 871–881.

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

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. 1994b. Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235–240.

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

REBAÏ, A., GOFFINET, B., MANGIN, B. AND PERRET, D. 1994a. QTL detection with diallel schemes. In: *Proceedings of the Ninth Eucarpia Meeting. Molecular markers in Plant Breeding*, pp. 170–177. Wageningen.

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

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