Theor Appl Genet (2001) 102:1099–1102

H.P. Piepho

Significance testing for QTL mapping by marker difference regression

Received: 19 June 2000 / Accepted: 7 August 2000

Abstract Marker difference regression, also known as joint mapping or simple marker regression, regresses differences of means for marker classes ('marker differences') on their expected value, which is a function of the QTL effect and position. The genome is scanned by a likelihood ratio statistic, which conditionally on the QTL position follows a chi-squared distribution. The unconditional distribution is a chi-squared process. This paper proposes a quick method for controlling the genome-wise Type-I error rate. The method is shown by simulation to work satisfactorily for a number of settings.

Keywords Simple marker regression · Joint mapping · Maximum likelihood · Type-I error rate · Generalized least squares

Introduction

Wu and Li (1994) and Kearsey and Hyne (1994) independently suggested a method for quantitative trait locus (QTL) mapping (joint mapping, simple marker regression), which has recently been termed marker difference regression (MDR) by Lynch and Walsh (1998). For further developments see Hyne and Kearsey (1995), Kearsey and Pooni (1996), Wu and Li (1996a,b), and Charmet et al. (1998). The method regresses differences of means for marker classes ('marker differences') on their expected value, which is a simple function of genetic effects and the position of the putative QTL. Parameters may be estimated by maximum likelihood, assuming multivariate normality of marker means. The normality assumption for marker means is an approximation, because the individual observations in a marker class follow a normal mixture distribution. When the population size is not too small (>200, say), the approximation will be rather close. The chromosome is scanned over a grid of putative QTL positions. At each position a likelihood ratio (LR) statistic is computed for the null hypothesis that the chromosome in question has no QTL at the putative position. Conditionally on the QTL position, this test statistic has a χ^2-distribution under the null hypothesis. This is not, however, the appropriate reference distribution for testing the null hypothesis that there is no QTL anywhere on the chromosome, because for this test we need to consider the unconditional distribution of the LR statistic. This distribution is a chi-squared process. Adapting the general results of Davies (1977, 1987) on likelihood ratio testing in non-standard situations, we suggest an approximate unconditional test for controlling the chromosome-wise Type-I error rate in MDR. The performance in finite samples is investigated by simulation.

Communicated by H.C. Becker

H.P. Piepho (✉)
Institut für Nutzpflanzenkunde, Universität-Gesamthochschule Kassel, Nordbahnhofstrasse 1a, 37213 Witzenhausen, Germany
e-mail: piepho@wiz.uni-kassel.de

Theory

The method is exemplified for a backcross (BC_1) population and a single putative QTL. It is equally applicable to other population structures (F_2, doubled haploid lines, etc.) as well as to multiple QTL on the same chromosome. Consider a backcross $m_1m_1.....qq....m_Pm_P \times M_1m_1.....Qq....M_Pm_P$, where M_i, m_i ($i = 1, ..., P$) denote the marker alleles, while Q, q are the QTL alleles. For the i-th marker we compute the difference of marker means, given by

$$y_i = \bar{z}(M_iM_i) - \bar{z}(M_im_i), \quad (1)$$

where $\bar{z}(M_iM_i)$ and $\bar{z}(M_im_i)$ denote the sample means for marker classes M_iM_i and M_im_i, respectively. The marker differences y_i approximately follow a multivariate normal distribution with expectation

$$\mathrm{E}(y_i) = a(1-2r_i), \quad (2)$$

where a is the QTL effect and r_i is the recombination fraction between the i-th marker and the putative QTL.

Under the null hypothesis H_0: $a = 0$ the expectation simplifies to $E(y_j) = 0$. The (co)variances are given by (Wu and Li 1996a)

$$V_{ij} = \begin{cases} \mathrm{var}[\bar{z}(M_iM_i)] + \mathrm{var}[\bar{z}(M_im_i)] & \text{for } i = j \\ (1-2r_{ij})\sqrt{V_{ii}V_{jj}} & \text{for } i \neq j \end{cases}, \quad (3)$$

where $\mathrm{var}[\bar{z}(M_iM_i)]$ and $\mathrm{var}[\bar{z}(M_im_i)]$ are the variances of $\bar{z}(M_iM_i)$ and $\bar{z}(M_im_i)$, and r_{ij} is the recombination fraction between markers i and j. The covariance in eq. 3 is valid asymptotically for large sample sizes (G), i.e., for a large number of individuals. For small G, expression (A1) in Wu and Li (1996a) should be used. Assuming multivariate normality with the first two moments given by eqs. 2 and 3, we may fit the model by maximum likelihood. Let L_0 and $L_1(\theta)$ be the values of the maximized log-likelihoods under the null and alternative hypotheses, respectively, where θ (cM) is the putative QTL position. Then, conditionally on θ, the LR statistic

$$T(\theta) = -2[L_0 - L_1(\theta)] \quad (4)$$

asymptotically (as the number of markers P approaches infinity) has a χ^2 distribution with one degree of freedom. This asymptotic distribution also holds, when (co)variances in eq. 3 are replaced by their sample values, provided the sample size (G) is large. Note that conditional on θ, $T(\theta)$ is equivalent to the difference in the residual weighted sums of squares for the models under the null hypothesis and the alternative hypothesis, where the weighting matrix is the inverse of $\{V_{ij}\}$. To search for QTL, we compute $T(\theta)$ across a fine grid of putative QTL positions θ on the chromosome and determine the maximal value, which is denoted here as $T_{\max}$. Clearly, $T_{\max}$ does not have an asymptotic χ^2 distribution with one degree of freedom. One might think that $T_{\max}$ follows a χ^2 distribution with two degrees of freedom, because there are two unknown parameters under the alternative, i.e., QTL position and effect. Note, however, that one of the two parameters, i.e., the QTL position, is absent under the null hypothesis. For this reason, standard asymptotic maximum likelihood theory does not apply.

It follows from Davies (1977, 1987) that if we reject H_0 when $T_{\max} > C$, an upper bound of the chromosome-wise Type-I error rate can be estimated by

$$\alpha < \Pr(\chi_k^2 > C) + WC^{\frac{1}{2}(k-1)}\exp\left(-\frac{1}{2}C\right)2^{-\frac{1}{2}k}/\Gamma\left(\frac{1}{2}k\right), \quad (5)$$

where $\Gamma(.)$ is the Gamma function and $\Pr(\chi_k^2 > C)$ is the cumulative distribution function of χ^2 with k degrees of freedom. In the case at hand (a BC_1 population) we have $k = 1$, because there is one genetic effect (a) under the alternative, which is fixed to $a = 0$ under the null hypothesis. The term W is computed as

$$W = \left|\sqrt{T(0)} - \sqrt{T(\theta_1)}\right| + \left|\sqrt{T(\theta_1)} - \sqrt{T(\theta_2)}\right| + \ldots + \left|\sqrt{T(\theta_r)} - \sqrt{T(L)}\right|, \quad (6)$$

where θ_1, ..., θ_r are the successive turning points (points of inflection) of $\sqrt{T(\theta)}$, i.e., the values of θ, where the first derivative $\partial\sqrt{T(\theta)}/\partial\theta$ changes sign, and L is the length of the chromosome. For nominal α, a conservative critical value C may be found from eq. 5 by numerical methods. To evaluate eq. 6, it is necessary to find the turning points θ_1, ..., θ_r. If a grid search is done over all θ, the turning points can be determined by pretending that every point on the grid is a turning point. Piepho (in preparation) suggests using a relatively fine grid, e.g., between 1 cM and 2 cM or smaller. To compute W, we simply compute the absolute differences between successive square roots of $T(\theta)$ on the grid and sum these across the chromosome. This will yield the correct result (to the accuracy of the grid), even though only a fraction of grid points will correspond to real turning points. To see this, consider the quantity $w = \left|\sqrt{T(\theta_1)} - \sqrt{T(\theta_2)}\right| + \left|\sqrt{T(\theta_2)} - \sqrt{T(\theta_3)}\right|$ and assume that θ_1 and θ_3 are turning points, while θ_2, a point between θ_1 and θ_3, is not. It is then clear that w equals $\left|\sqrt{T(\theta_1)} - \sqrt{T(\theta_3)}\right|$, thus proving the claim. Upon request, the author can make available a SAS program for computing critical thresholds based on eqs. 5 and 6. Finally, note that eq. 5 can be used to attach a p-value to an observed value for $T_{\max}$, as will be demonstrated using a worked example.

Simulation

We performed a simulation for a BC_1 to demonstrate that Davies' method controls the chromosome-wise Type-I error rate. For comparison, we also included the tests based on χ_1^2 and χ_2^2. The simulation comprises a single chromosome, whose length (L) was set at 60, 100 and 200 centiMorgan (cM). Markers were assumed to be equally spaced. We investigated marker distances (D) of 5, 10, and 20 cM. The empirical Type-I error was assessed at nominal levels $\alpha = 0.01$ and $\alpha = 0.05$. The population size (G) was chosen as 200 and 1,000. The step size for scanning the chromosome was 1 cM. For each setting, we performed 10,000 simulation runs. All simulations were programmed using SAS/IML. Results are shown in Table 1. It is obvious that neither χ_1^2 nor χ_2^2 is the appropriate null distribution; both yield inflated Type-I error rates. The inflation increases with the chromosome length (L) and with marker density. The increase with L is to be expected from eqs. 5 and 6, because W will tend to grow with L. In contrast the quick method of Davies satisfactorily controls the Type-I error rate for different values of L. Comparison of the simulations for sample sizes $G = 200$ and $G = 1{,}000$ shows that the approximation may be slightly on the liberal side in some cases with small sample sizes G but still acceptable for most practical purposes, while with $G = 1{,}000$, all marker distances yield conservative results. Note that in general a relatively large sample size is necessary to obtain good estimates of variance components. Also, the covariance in eq. 3 is based on an approximation, which is valid for large G.

Table 1 Simulated chromosome-wise Type-I error rate for a BC_1 population (10,000 simulations for each setting; step size was 1 cM; simulation for a single chromosome)

Parameters[a]			Davies		χ_1^2		χ_2^2	
L	*D*	G	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$	$\alpha = 0.01$	$\alpha = 0.05$
60	20	200	0.0083	0.0457	0.0420	0.1666	0.0112	0.0572
60	10	200	0.0094	0.0465	0.0534	0.2000	0.0169	0.0715
60	5	200	0.0119	0.0487	0.0668	0.2437	0.0228	0.0883
100	20	200	0.0093	0.0439	0.0595	0.2293	0.0156	0.0825
100	10	200	0.0112	0.0487	0.0790	0.2771	0.0241	0.1033
100	5	200	0.0124	0.0509	0.1039	0.3391	0.0349	0.1381
200	20	200	0.0112	0.0456	0.1057	0.3744	0.0317	0.1418
200	10	200	0.0125	0.0514	0.1402	0.4545	0.0471	0.1859
200	5	200	0.0137	0.0572	0.1823	0.5339	0.0666	0.2403
60	20	1,000	0.0087	0.0405	0.0390	0.1516	0.0112	0.0510
60	10	1,000	0.0087	0.0399	0.0471	0.1896	0.0136	0.0636
60	5	1,000	0.0080	0.0409	0.0592	0.2299	0.0172	0.0814
100	20	1,000	0.0076	0.0408	0.0545	0.2142	0.0156	0.0723
100	10	1,000	0.0066	0.0409	0.0701	0.2755	0.0185	0.0969
100	5	1,000	0.0095	0.0433	0.0933	0.3277	0.0287	0.1273
200	20	1,000	0.0078	0.0426	0.0992	0.3530	0.0280	0.1331
200	10	1,000	0.0095	0.0476	0.1319	0.4382	0.0414	0.1766
200	5	1,000	0.0082	0.0401	0.1553	0.5060	0.0486	0.2072

[a] *L*, Length of chromosome in centiMorgans; *D*, distance of two adjacent markers in centiMorgans; *G*, number of individuals

A worked example

To demonstrate the important computational steps of Davies' method, we simulated a chromosome of length $L = 40$ cM for $G = 200$ BC_1 individuals, assuming absence of any QTL. Markers were spaced $D = 5$ cM apart, and the step size for the grid search over the chromosome was 2 cM (the relatively large step size was chosen to keep the numerical example small for demonstration purposes). The profile of the LR statistic $T(\theta)$ is shown in Table 2. The test statistic, i.e., maximum value for $T(\theta)$ is $T_{\max} = 4.204$. From the profile we computer $W = \left|\sqrt{0.009} - \sqrt{0.101}\right| + \left|\sqrt{0.101} - \sqrt{0.287}\right| + \ldots + \left|\sqrt{3.618} - \sqrt{3.450}\right| = 2.316$. We now evaluate eq. 5 to obtain an approximate p-value for $T_{\max}$. Setting $k = 1$ for a backcross population and inserting the numerical value for W we obtain from eq. 5

$$p < \Pr(\chi_1^2 > T_{\max}) + 2.316 \exp\left(-\frac{1}{2} T_{\max}\right) 2^{-\frac{1}{2}} / \Gamma\left(\frac{1}{2}\right) \quad (7)$$

To evaluate the Gamma function $\Gamma(.)$ and the cumulative distribution function $\Pr(\chi_1^2 > C)$ we use SAS/IML. Inserting $T_{\max} = 4.204$ the p-value becomes $p = 0.153$, which is not significant at the 5% level. Alternatively, we may compute a conservative critical threshold for the desired significance level α by finding the value for C satisfying

$$\alpha = \Pr(\chi_1^2 > C) + 2.317 \exp\left(-\frac{1}{2} C\right) 2^{-\frac{1}{2}} / \Gamma\left(\frac{1}{2}\right) \quad (8)$$

which, again, is obtained from eq. 5. Since eq. 8 is a monotonically decreasing function of C, we may perform a binary search to find the solution. For example with $C = 5$, we find $\alpha = 0.101 > 0.05$, while $C = 15$ yields $\alpha = 0.0006 < 0.05$. Due to decrease of α with increasing C, the value of C yielding $\alpha = 0.05$ must lie in the interval (5, 15), i.e., the pair 5/15 brackets the solution for C. Note that generally for the smaller C-value of the bracket, one must have $\alpha > 0.05$, while for the larger C-value of the bracket $\alpha < 0.05$. The center of the interval lies at $C = 10$, which yields $\alpha = 0.0078 < 0.05$. Thus, we may replace the larger C-value of the bracket by $C = 10$ and proceed with the updated pair 5/10, which is a more narrow bracket of the solution. Repeating this procedure, we obtain the center of the new interval as $C = 7.5$ with $\alpha = 0.028 < 0.05$, so the updated bracket is 5/7.5, and so forth. The process of halfing the interval and updating the bracket is repeated until the bracket becomes arbitrarily small; for example until the width of the bracket becomes smaller than $\varepsilon = 10^{-10}$. This yields the critical value $C = T_{\text{crit}} = 6.364$. Since the test statistic $T_{\max} = 4.204$ does not exceed this critical threshold, it is not significant at $\alpha = 0.05$.

Table 2 Simulated profile of the likelihood ratio statistic $T(\theta)$ for MDR based on a single chromosome in a BC_1 population with $L = 40$, $D = 5$, and $G = 200$, assuming absence of any QTL effect. The step size of the grid search is 2 cM. The QTL position θ is given in cM

θ	$T(\theta)$	θ	$T(\theta)$	θ	$T(\theta)$
0	0.009	14	1.171	28	3.313
2	0.101	16	1.376	30	4.204
4	0.287	18	1.821	32	4.084
6	0.528	20	2.197	34	3.715
8	0.802	22	2.163	36	3.565
10	1.069	24	1.995	38	3.618
12	1.155	26	2.336	40	3.450

Discussion

We have demonstrated the method for a BC_1 population. It is equally valid for other population structures, with a suitable choice of k, the number of degrees of freedom of the LR statistic for a test conditional on the QTL position. For example, with an F_2 population, we have $k = 2$, while for doubled haploids or recombinant inbred lines $k = 1$.

With multiple QTL on the same chromosome, Wu and Li (1996a) suggest a stepwise procedure, by which the number of fitted QTL for a chromosome is increased until a test of a model with b QTL versus a model with $b - 1$ QTL reveals no significance. For a BC_1 population, the LR statistic for this comparison has one degree of freedom, conditional on the QTL position, as the LR statistic for a single QTL. For an F_2, it has two degrees of freedom. The results of Davies (1977, 1987) remain applicable so that we can use eq. 5 to obtain an appropriate critical value. Note that the critical value C for a given chromosome will change with b because C is computed from the LR profile over θ, which also changes with b. The stepwise fitting of QTL on a chromosome using eq. 5 will control the chromosome-wise Type-I error rate in the sense that the probability of detecting at least one false positive is bounded by α.

In many applications, the researcher may want to control not the Type-I error rate for a single chromosome, but the genome-wise Type-I error rate across multiple chromosomes, which is a more stringent requirement. To control the chromosome-wise error rate at nominal level γ, one may use a Bonferroni adjustment, i.e., one simply performs the chromosome-wise tests at level $\alpha = \gamma/n$, where n is the number of chromosomes. Thus, the results presented here for control of the Type-I error rate on a single chromosome can be used also to control the genome-wise Type-I error rate.

Davies' method has been used successfully also with interval mapping (IM) and composite interval mapping (CIM) (Piepho, in preparation). A number of different population structures (backcross, advanced backcross, F_2, advanced intercross lines) have been investigated in simulations, showing conservative control of the Type-I error rate. These findings are in good agreement with the present results for marker difference regression. For some simple population structures such as BC_1 and F_2, closed-form critical thresholds are available for IM (Reba et al. 1994; Dupuis and Siegmund, 1999), which do not depend on the observed data, and they yield similar results as Davies' quick method. Closed-form expressions are not available for many other population structures in the case of IM, and such expressions seem to be generally lacking for other methods such as CIM, MDR, and marker interval mapping (MIM) as proposed by Kao et al. (1999). A great advantage of Davies' quick method is its generality, which makes it applicable to any population structure and to many of the common QTL mapping methods, including MDR, IM, CIM, and MIM. It can also be used with multivariate approaches as those proposed by Jiang and Zeng (1995) and Calinski (2000).

This paper has proposed a simple and versatile method to compute critical values for QTL detection by MDR. The current state of the art for computing critical thresholds is based on computer intensive methods. Hyne and Kearsey (1995) proposed a form of parametric bootstrap to compute appropriate p-values for MDR, which has the same degree of versatility as the present method and should provide adequate control of the chromosome-wise Type-I error rate. Alternatively, one could obtain critical thresholds by permutation methods as currently popular in IM and CIM (Churchill and Doerge 1994). A major advantage of the present method over parametric bootstrap permutation tests, and other Monte Carlo methods is the low computational workload.

Acknowledgements I wish to thank Dr. Kenneth F. Manly (Molecular and Cellular Biology, Roswell Park Cancer Institute, Buffalo, USA) for bringing to my attention the question of appropriate critical thresholds in marker difference regression and for reading an earlier draft of the paper. Thanks are also due to two referees for helpful comments. Support of the Heisenberg Programm of the Deutsche Forschungsgemeinschaft (DFG) is gratefully acknowledged.

References

Calinski T, Kaczmarek Z, Krajewski P, Frova C, Sari-Gorla M (2000) A multivariate approach to the problem of QTL localization. Heredity 84:303–310

Charmet G, Cadalen T, Sourdille P, Bernard M (1998) An extension of the 'marker regression' method to interactive QTL. Mol Breed 4:67–82

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963–971

Davies RB (1977) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64:247–254

Davies RB (1987) Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 74:33–43

Dupuis J, Siegmund D (1999) Statistical methods for mapping quantitative trait loci from a dense set of markers. Genetics 151:373–386

Hyne V, Kearsey MJ (1995) QTL analysis – further uses of marker regression. Theor Appl Genet 91:471–476

Jiang C, Zeng Z-B (1995) Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140:1111–1127

Kao CH, Zeng ZB, Teasdale RD (1999) Multiple interval mapping for quantitative trait loci. Genetics 159:1203–1216

Kearsey MJ, Hyne V (1994) QTL analysis – a simple marker-regression approach. Theor Appl Genet 89:698–702

Kearsey MJ, Pooni HS (1996) The genetical analysis of quantitative traits. Chapman and Hall, London

Lynch M, Walsh B (1998) Genetics and analysis of quantitative traits. Sinauer, Sunderland, Mass.

Reba A, Goffinet B, Mangin B (1994) Approximate thresholds of interval mapping tests for QTL detection. Genetics 138:235–240

Wu W-R, Li W-M (1994) A new approach for mapping quantitative trait loci using complete genetic marker linkage maps. Theor Appl Genet 89:535–539

Wu W-R, Li W-M (1996a) Model fitting and model testing in the method of joint mapping of quantitative trait loci. Theor Appl Genet 92:477–482

Wu W-R, Li W-M (1996b) Joint mapping of quantitative trait loci using F_2 populations. Theor Appl Genet 93:1156–1160