Letter to the Editor

A Comment on the Simple Regression Method for Interval Mapping

Shizhong Xu

Department of Botany and Plant Sciences, University of California, Riverside, California 92521

Manuscript received July 17, 1995
Accepted for publication September 14, 1995

HALEY and KNOTT (1992) developed a simple regression method for mapping quantitative trait loci (QTLs) using F_2 populations. MARTINEZ and CURNOW (1992) considered the same approach using backcross populations. Studies that compare the regression method with that of LANDER and BOSTEIN's (1989) maximum likelihood find very little difference between the two methods. The simple regression method offers a great advantage over the maximum likelihood method in terms of computing speed. MARTINEZ and CURNOW (1992) state that the difference between the two methods arises solely because the residual term in the linear model does not have a normal distribution, being a random variable from one of several normal distributions. Theoretically, the regression method should suffer from the failure of the assumption of normality within marker genotype due to the segregation of the QTL, but the great majority of information is contained in mean differences between marker genotypes with little coming from the within marker genotype distribution (HALEY and KNOTT 1992). Therefore, the efficiency of the simple regression method is unlikely to be substantially less than the maximum likelihood approach (MARTINEZ and CURNOW 1992).

The parameters of primary interest in QTL mapping are the positions and effects of QTLs. However, estimation of the residual variance is also important for the following reasons: (1) Effects of QTLs are only meaningful when reported relative to the size of the residual standard deviation; (2) The test statistic (t-test) is determined by the ratio of the estimated QTL effect to its standard deviation while the latter is proportional to the residual standard deviation; (3) Constructing confidence intervals about the estimated effects of the QTLs requires knowledge of the residual standard deviation (through the t-test statistic). Although the simple regression method provides good estimates of positions and effects of QTLs, estimation of the residual variance tends to be biased. Unfortunately, there has been no attempt to investigate the bias in the literature. This short note reports the result of theoretical derivation of the bias and proposes a simple method to correct it. The theory is verified from a Monte Carlo simulation experiment.

Consider, for simplicity, a backcross population derived from two inbred parental lines fixed for alternative alleles at all QTLs and marker loci and we want to test for the presence of a QTL flanked by two markers. We can use genotypes of the two markers to predict the genotype of the putative QTL and write the statistical model as

$$y_j = \beta_0 + \beta x_j + \epsilon_j \quad j = 1, \ldots, n \tag{1}$$

where y_j is the trait value of the jth plant, β_0 is the mean of the model, β is the effect of the putative QTL expressed as the difference between homozygotes and heterozygotes, x_j is an indicator variable, taking a value 1 or 0 depending on the genotype of the QTL, and $\epsilon_j \sim N(0, \sigma_\epsilon^2)$ is the residual including random environmental effect and effects of other QTLs not explained by the markers.

Model (1) appears the same but is different from the usual linear model in that the independent variable, x_j, is not directly observable; rather, it has a probability distribution that can be inferred from genotypes of flanking markers and the position being tested for the putative QTL. Disregard the fact that x_j is missing, the mean and variance of y_j conditional on x_j are $E(y_j|x_j) = \beta_0 + \beta x_j$ and $\text{Var}(y_j|x_j) = \sigma_\epsilon^2$, respectively.

If the QTL is not at a marker locus, given the genotypes of flanking markers, x_j is still uncertain and has to be treated as a random variable. Therefore, the mean and variance of y_j conditional on flanking marker genotype are

$$E(y_j|\text{marker}) = \beta_0 + \beta E(x_j|\text{marker})$$

and

$$\text{Var}(y_j|\text{marker}) = \beta^2\text{Var}(x_j|\text{marker}) + \sigma_\epsilon^2$$

respectively, where $E(x_j|\text{marker}) = p_j$, $\text{Var}(x_j|\text{marker}) = p_j(1 - p_j)$ and p_j is the probability of $x_j = 1$ conditional on the flanking marker genotype and the position of the QTL. Because marker genotypes provide information about the QTL through p_j, the above conditional mean and variance can be expressed by

Corresponding author: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: xu@genetics.ucr.edu

TABLE 1

Comparison of regression (RG) and maximum likelihood (ML) analyses of data from a simulated backcross population of size 1000

	20 cM spaced				50 cM spaced			
β	$\hat{\beta}$	cM$_A$	$\hat{\sigma}^{2}_{\epsilon}$*	σ^2_e	$\hat{\beta}$	cM$_A$	$\hat{\sigma}^{2}_{\epsilon}$*	σ^2_e
0.5								
ML	0.489 (0.072)	10.03 (3.558)	1.004 (0.046)		0.498 (0.084)	25.50 (6.185)	1.000 (0.050)	
RG	0.488 (0.072)	10.03 (3.622)	1.017 (0.046)	1.012	0.499 (0.085)	25.30 (5.920)	1.030 (0.048)	1.029
1.0								
ML	0.991 (0.061)	9.86 (2.005)	1.007 (0.048)		1.014 (0.094)	25.17 (3.324)	1.014 (0.094)	
RG	0.991 (0.062)	9.93 (1.981)	1.056 (0.048)	1.049	1.015 (0.096)	25.16 (3.274)	1.015 (0.096)	1.116
2.0								
ML	2.003 (0.071)	9.94 (1.171)	0.999 (0.045)		1.989 (0.075)	24.85 (1.893)	1.010 (0.058)	
RG	2.002 (0.075)	9.90 (1.193)	1.197 (0.054)	1.197	1.996 (0.094)	24.97 (1.936)	1.464 (0.066)	1.462
4.0								
ML	3.994 (0.065)	10.07 (0.807)	1.008 (0.050)		3.991 (0.066)	24.69 (1.643)	1.012 (0.050)	
RG	4.007 (0.076)	9.99 (0.847)	1.785 (0.091)	1.789	4.001 (0.128)	24.69 (1.680)	2.844 (0.133)	2.848
8.0								
ML	7.997 (0.067)	10.08 (0.787)	1.002 (0.045)		7.991 (0.058)	24.98 (1.341)	0.995 (0.038)	
RG	7.987 (0.091)	10.07 (0.820)	4.166 (0.296)	4.156	7.993 (0.167)	24.99 (1.352)	8.349 (0.410)	8.394

The table shows the mean estimates of the gene substitution effect (β) of the QTL and its distance from the first marker (cM$_A$), and the residual variance (with the standard deviation of the estimates over 100 replicates in parentheses). The true QTL location is in the middle of tested interval and the true residual variance is $\sigma^2_\epsilon = 1.0$.

* In the RG method, $\hat{\sigma}^2_\epsilon$ should be expressed by $\hat{\sigma}^2_e$.

$$E(y_j|p_j) = \beta_0 + \beta p_j$$

and

$$\mathrm{Var}(y_j|p_j) = p_j(1 - p_j)\beta^2 + \sigma^2_\epsilon = [p_j(1 - p_j)\gamma + 1]\sigma^2_\epsilon$$

respectively, where $\gamma = \beta^2/\sigma^2_\epsilon$.

MARTINEZ and CURNOW (1992) reformulate model (1) as

$$y_j = \beta_0 + \beta p_j + e_j \qquad (2)$$

and regress y_j on p_j to estimate β. HALEY and KNOTT (1992) also replace the missing QTL genotype by its probability and use the same approach to interval mapping in F_2 populations. Note that the residual, e_j, defined under model (2) is not equivalent to ϵ_j. The variance of e_j is equivalent to the variance of y_j conditional on p_j, *i.e.*,

$$\sigma^2_{ej} = \mathrm{Var}(y_j|p_j) = [p_j(1 - p_j)\gamma + 1]\sigma^2_\epsilon$$

which is not homogenous and has been inflated by the within marker genotype QTL variance. The amount of inflation depends on p_j and γ. When the simple regression analysis is conducted under model (2), the "residual variance" is expected to be

$$\sigma^2_e = \{\gamma E[p_j(1 - p_j)] + 1\}\sigma^2_\epsilon \qquad (3)$$

where the expectation, $E[p_j(1 - p_j)]$, takes the average of $p_j(1 - p_j)$ over j. The actual residual variance is σ^2_ϵ, but the "residual variance" being estimated in the simple regression analysis is σ^2_e. The inflation could be substantial if β is large compared to σ^2_ϵ.

The above derivation was verified by a Monte Carlo simulation experiment. Data were simulated for 100 replicates of 1000 backcross individuals from a cross between two inbred lines. For simplicity, one chromosomal segment was simulated with length of either 20 or 50 cM. The chromosomal segment was flanked by two markers with a QTL located in the middle of the interval. The effect of the QTL was set at 0.5, 1.0, 2.0, 4.0 and 8.0 residual standard deviations. Without loss of generality, the residual standard deviation was set at one ($\sigma^2_\epsilon = 1$). The probability (p_j) of $x_j = 1$ is determined by the position of the QTL and the flanking marker genotypes assuming no interference (MARTINEZ and CURNOW 1992). HALDANE's map function was used to convert the map distance into recombination frequency. When the size of the interval is 20 cM, $E[p_j(1 - p_j)] = 0.0493$, leading to $\sigma^2_e = (0.0493\gamma + 1)\sigma^2_\epsilon = 1 + 0.0493\ \beta^2$, while $\sigma^2_e = 1 + 0.1155\ \beta^2$ if the length of the interval is 50 cM.

Results are listed in Table 1. The estimates of QTL position and effect from the two methods were very similar, which is consistent with HALEY and KNOTT (1992). However, when $\beta > 1.0$, the residual variance was seriously over estimated by the regression method, especially when 50 cM spaced markers were used. For example, when $\beta = 4.0$ and marker space is 50 cM, the mean estimate of $\hat{\sigma}^2_e$ from the regression method is 2.844, while the mean estimate from the ML method is 1.012, very close to one (the true value). The actual proportion of phenotypic variance explained by the QTL was 80%, but this proportion would be reported as 58% from the regression analysis.

In view of the fact that σ^2_ϵ is inflated when the simple regression analysis is used, we can adjust σ_e^2 to obtain an approximately unbiased estimate of σ_ϵ^2 by rearranging (3), *i.e.*,

$$\sigma_\epsilon^2 = \sigma_e^2 - \beta^2 E[p_j(1 - p_j)] \tag{4}$$

The expectation, $E[p_j(1 - p_j)]$, is determined by the position of the QTL relative to the flanking markers and the length of the interval. In a backcross population, there are four possible marker genotype classes, say *AABB, AABb, AaBB* and *AaBb*. Therefore, the expectation, $E[p_j(1 - p_j)]$, is

$$E[p_j(1 - p_j)] = \sum_{i=1}^{4} \Pr(x_j = 1 | M_i) \times [1 - \Pr(x_j = 1 | M_i)] \Pr(M_j)$$

where M_i denotes the *i*th marker genotype class. For example, if the marker genotype is *AaBb*, then $\Pr(M_4) = (1 - r_{12})/2$ and $\Pr(x_j = 1 | M_4) = r_{1q}r_{q2}/(1 - r_{12})$, where r_{1q} and r_{q2} are the recombination fractions of the QTL with the left and the right markers, respectively, and r_{12} denotes the recombination fraction between the two marker loci.

This note does not intend to disqualify the simple regression method in QTL mapping. The effect of any individual QTL being tested is usually small for most polygenic traits, which makes the regression method valid for most situations. However researchers who prefer the regression method need to be aware of the fact that the residual variance estimated from the simple regression method is inflated by part of the variance of the tested QTL. The inflation could be substantial when multiple markers are used simultaneously as covariates (ZENG 1993, 1994; JANSEN 1993, 1994), because the actual residual variance could be small. Therefore, it is necessary to adjust this inflated estimation to obtain an unbiased estimate of the residual variance.

LITERATURE CITED

HALEY, C. S., and S. A. KNOTT, 1992 A simple regression 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 D. BOSTEIN, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

MARTINEZ, O., and R. N. CURNOW, 1992 Estimating the locations and the sizes of the effects of quantitative trait loci using flanking markers. Theor. Appl. Genet. **85:** 480–488.

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.

Communicating editor: Z-B. ZENG