Heredity **80** (1998) 364–373

Received 11 April 1997

Further investigation on the regression method of mapping quantitative trait loci

SHIZHONG XU*

Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, U.S.A.

The simple regression method of mapping quantitative trait loci (QTL) is further investigated in comparison with the mixture model maximum likelihood method under high heritabilities, dominant and missing markers. No significant difference between the two methods is detected in terms of errors of parameter estimation and statistical powers, with the exception that the estimation of residual variance provided by the regression method is confounded with part of the QTL variance. The test statistic profiles show some difference between the two methods, but the difference is only detectable at the micro level. An alternative method, referred to as iteratively reweighted least squares, is proposed, which can correct the deficiency of parameter confounding in the regression method yet retains the properties of simplicity and rapidity of the ordinary regression method. Like the existing regression method, the weighted least squares method can be useful in QTL mapping in conjunction with the permutation tests and construction of confidence intervals by bootstrapping.

Keywords: linear regression, maximum likelihood, QTL, weighted least squares.

Introduction

Lander & Botstein (1989) presented an exact maximum likelihood method (ML) for mapping quantitative trait loci (QTL) in line crossing experiments. When the putative position is off the markers, the QTL genotype is actually not observed, so the model involves missing data. Solutions of the exact maximum likelihood method involving missing data are usually obtained using the (Expectation–Maximization) EM algorithm (Dempster *et al.*, 1977), which requires many cycles of iterations. Haley & Knott (1992) discovered that the ML can be well approximated by the simple regression method (REG). The authors conducted extensive computer simulations, showing no detectable difference between ML and REG in the range of parameters considered in the simulation experiment. A similar argument is also found in Martinez & Curnow (1992). As a consequence, the simple regression method has become widely accepted, especially in European countries, because of its simplicity and convenience of use relative to the ML.

Given that two methods are available for QTL mapping, which method should be chosen for real data analyses? For analysis of a single data set, it does not matter which one is used because the two methods will generate almost identical results. Some researchers may want to avoid the word 'approximation' and choose ML, and others may prefer simplicity and thus choose REG. Xu (1995) recently found that the residual variance estimated by the REG method contains part of the QTL variance caused by the uncertainty of QTL genotype. This observation may alert users of the REG that the explanation of the residual variance should be treated with caution. However, the REG method is computationally so superior to the ML that it may become the choice for multiple data analyses, such as the permutation tests (Churchill & Doerge, 1994) and the bootstrap construction of confidence intervals (Visscher *et al.*, 1996). These nonparametric methods involve thousands of analyses of the (resampled) same data set and could be prohibitive for ML if the data set and genome size are large.

The purposes of this paper are: (i) to investigate further the difference between the REG and ML methods via simulation studies in situations with high heritabilities and dominant and/or missing markers; and (ii) to improve the existing regression method so that the pure environmental variance can be separated from the residual variance, yet the property of high computing speed is retained.

*E-mail: xu@genetics.ucr.edu

Statistical methods

Linear model

Let y_j be the phenotypic value of an F_2 individual that can be described by the following linear model:

$$y_j = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} + z_j \alpha + w_j \delta + \varepsilon_j, \tag{1}$$

where $\mathbf{x}_j$ is a known vector, $\boldsymbol{\beta}$ is a vector of unknown fixed effects, α and δ are, respectively, the average effect of allelic substitution and the dominance effect of a putative QTL, and ε_j is the residual error with $N(0,\sigma_\varepsilon^2)$. Note that for a single-QTL model the residual error is purely caused by uncontrollable environmental noise. The independent variables, z_j and w_j, are defined as:

$$z_j = \begin{cases} +1 & \text{for } Q_1Q_1 \\ 0 & \text{for } Q_1Q_2 \\ -1 & \text{for } Q_2Q_2 \end{cases}$$

and

$$w_j = \begin{cases} +1 & \text{for } Q_1Q_2 \\ -1 & \text{for } Q_1Q_1 \text{ or } Q_2Q_2, \end{cases}$$

where Q_1Q_1, Q_2Q_2 and Q_1Q_2 are, respectively, the genotypes of the two parental lines and the F_1 hybrid. Because the genotype of a QTL is not observable if the QTL is not at a marker, z_j and w_j are usually missing. However, the conditional distribution of z and w can be inferred from the genotypes of linked markers. Let $p_{(kl)j}$ be the conditional probability that the individual is of genotype Q_kQ_l, given marker information. Given the conditional probabilities, y_j is considered to be sampled from a mixture of three distributions with means of μ_{11}, μ_{12} and μ_{22} and a common variance σ_ε^2, where:

$$\mu_{11} = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} + \alpha - \delta, \; \mu_{12} = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} + \delta$$

and $\mu_{22} = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} - \alpha - \delta$.

Statistical tests and parameter estimation are conducted through one of the three methods described below.

Maximum likelihood method (ML)

The likelihood function is:

$$L(\boldsymbol{\theta} \mid \mathbf{y}) = \prod_{j=1}^{n} [p_{(11)j}\phi_{11}(y_j) + p_{(12)j}\phi_{12}(y_j) + p_{(22)j}\,\phi_{22}(y_j)], \tag{2}$$

where $\phi_{kl}(y_j)$ is the normal probability density for those individuals with genotype Q_kQ_l. It is well known that the maximum likelihood solution for the unknown parameters, $\boldsymbol{\theta} = [\boldsymbol{\beta}\ \alpha\ \delta\ \sigma_\varepsilon^2]^{\mathrm{T}}$, can be solved via the EM algorithm (Dempster *et al.*, 1977). To test the hypothesis that no QTLs are segregating, i.e. H_0: $\alpha = \delta = 0$, the following likelihood ratio test statistic is applied:

$$\Lambda = -2\,\{\log_e[L(\boldsymbol{\theta}_0 \mid \boldsymbol{y})] - \log_e[L(\boldsymbol{\theta} \mid \boldsymbol{y})]\}, \tag{3}$$

where $\boldsymbol{\theta}_0$ is different from $\boldsymbol{\theta}$ by introducing two constraints, $\alpha = 0$ and $\delta = 0$.

Simple regression method (REG)

The regression method of QTL mapping developed by Haley & Knott (1992) and Martinez & Curnow (1992) is an approximation of the ML method. These authors approximate the mixture of three distributions by a single distribution so that the ML solution can be obtained by a simple regression approach. The approximate single model is:

$$y_j = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} + E(z_j \mid I_M)\,\alpha + E(w_j \mid I_M)\,\delta + e_j, \tag{4}$$

where I_M denotes marker information and:

$$E(z_j \mid I_M) = (+1)\,p_{(11)j} + (0)\,p_{(12)j} + (-1)\,p_{(22)j}$$

and:

$$E(w_j \mid I_M) = (+1)\,p_{(12)j} + (-1)\,[p_{(11)j} + p_{(22)j}].$$

Note that the residual e_j is different from that given earlier. This single model has a mean of:

$$E(y_j) = \mathbf{x}_j^{\mathrm{T}} \boldsymbol{\beta} + E(z_j \mid I_M)\,\alpha + E(w_j \mid I_M)\,\delta$$

and a variance of:

$$\mathrm{Var}(y_j) = \mathrm{Var}(e_j) = \sigma_e^2.$$

The unknown parameters are solved using the ordinary least squares method (Haley & Knott, 1992). Under the assumption that y_j is normal, the least squares solutions are identical to the maximum likelihood estimators if the likelihood function is defined by:

$$L(\boldsymbol{\theta} \mid \boldsymbol{y}) = \prod_{j=1}^{n} \phi(y_j). \tag{5}$$

Two assumptions of the ML are violated by the regression analysis. One is the normal distribution of y_j and the other is the homogeneous residual variance. Violation of the normal distribution is not a problem with the regression method because estimation of the parameter does not depend on a normal distribution. Although the hypothesis test depends on the normal assumption, the *t*- or *F*-tests are usually very robust. Heterogeneous residual variance may cause a slight problem in the regression analysis (Xu, 1995), but is not likely to change the results qualitatively relative to the true ML analysis (Haley & Knott, 1992). The difference between the true ML and the regression method comes from the difference in the estimation of the residual variance. The regression method generally provides a residual variance estimation that contains part of the QTL variance not explained because of the uncertainty of QTL genotype (Xu, 1995). The *F*-value can be used as the test statistic for the simple regression method. However, to compare this method with the ML, the test statistic, originally used by Haley & Knott (1992), is adopted here:

$$\Lambda = n \log_e(\mathrm{RSS}_{\mathrm{reduced}}/\mathrm{RSS}_{\mathrm{full}}),$$

where $\mathrm{RSS}_{\mathrm{full}}$ is the residual sum of squares of the full model and $\mathrm{RSS}_{\mathrm{reduced}}$ is that of the reduced model. This test statistic can be compared with that given in eqn (3) because they are very similar under the null hypothesis (see Table 5).

Iteratively reweighted least squares method (IRWLS)

To retain the advantages of both the regression method and the ML method, a weighted regression method is investigated here. The mixture model is still approximated by a single model (eqn 4), but the residual variance is further partitioned into several components:

$$\mathrm{Var}(e_j) = \mathrm{Var}(z_j \mid I_\mathrm{M})\alpha^2 + \mathrm{Var}(w_j \mid I_\mathrm{M})\delta^2 + 2\mathrm{Cov}(z_j w_j \mid I_\mathrm{M})\alpha\delta + \sigma_\varepsilon^2, \quad (6)$$

where $\mathrm{Var}(z_j \mid I_\mathrm{M})\alpha^2$ is part of the QTL variance not explained because of the uncertainty of z_j, $\mathrm{Var}(w_j \mid I_\mathrm{M})\delta^2$ is part of the QTL variance not explained because of the uncertainty of w_j, and $2\mathrm{Cov}(z_j\ w_j \mid I_\mathrm{M})\alpha\delta$ is because of the uncertainty of both z_j and w_j. All three additional components in the residual will vanish if the genotype of the QTL is actually observed, i.e. $\mathrm{Var}(z_j \mid I_\mathrm{M}) = \mathrm{Var}(w_j \mid I_\mathrm{M}) = \mathrm{Cov}(z_j\ w_j \mid I_\mathrm{M}) = 0$. These additional components are computed as follows:

$$\mathrm{Var}(z_j \mid I_\mathrm{M}) = p_{(11)j}[1-p_{(11)j}] + p_{(22)j}[1-p_{(22)j}] + 2p_{(11)j}p_{(22)j},$$

$$\mathrm{Var}(w_j \mid I_\mathrm{M}) = 4p_{(12)j}[1-p_{(12)j}]$$

and:

$$\mathrm{Cov}(z_j\ w_j \mid I_\mathrm{M}) = p_{(22)j}[1-p_{(22)j}] - p_{(11)j}[1-p_{(11)j}] + p_{(12)j}p_{(22)j} - p_{(12)j}p_{(11)j}.$$

Let **y** be an $n \times 1$ vector of the data. The model can be expressed in matrix notation as:

$$\mathbf{y} = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{Z}\,\alpha + \mathbf{W}\,\delta + \mathbf{e} \quad (7)$$

where **Z** is an $n \times 1$ vector with the jth element equal to $\mathrm{E}(z_j \mid I_\mathrm{M})$, **W** is an $n \times 1$ vector with the jth element equal to $\mathrm{E}(w_j \mid I_\mathrm{M})$, and **e** is an $n \times 1$ vector of residuals. The expectation and variance matrix of the model are:

$$\mathrm{E}(\mathbf{y} \mid I_\mathrm{M}) = \mathbf{X}\,\boldsymbol{\beta} + \mathbf{Z}\,\alpha + \mathbf{W}\,\delta$$

and:

$$\mathrm{Var}(\mathbf{y} \mid I_\mathrm{M}) = \mathrm{Var}(\mathbf{e}) = \mathbf{R}\sigma_\varepsilon^2,$$

where **R** is a diagonal matrix with the jjth element equal to:

$$R_{jj} = \mathrm{Var}(z_j \mid I_\mathrm{M})\,\lambda_\alpha + \mathrm{Var}(w_j \mid I_\mathrm{M})\,\lambda_\delta + 2\mathrm{Cov}(z_j w_j \mid I_\mathrm{M})\,\lambda_{\alpha\delta} + 1 \quad (8)$$

and:

$$\lambda_\alpha = \alpha^2/\sigma_\varepsilon^2,\ \lambda_\delta = \delta^2/\sigma_\varepsilon^2 \text{ and } \lambda_{\alpha\delta} = \alpha\delta/\sigma_\varepsilon^2.$$

The likelihood function is:

$$\mathrm{L}(\boldsymbol{\theta} \mid \boldsymbol{y}) = (\sigma_\varepsilon^2)^{-1/2n}\,|\,\mathbf{R}\,|^{-1/2}\,\mathrm{Exp} \times \left\{ -\frac{1}{2\sigma_\varepsilon^2}(\boldsymbol{y} - \boldsymbol{X\beta} - \mathbf{Z}\alpha - \mathbf{W}\delta)^\mathrm{T}\mathbf{R}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\alpha - \mathbf{W}\delta) \right\}. \quad (9)$$

The ML solution can be solved via a weighted least squares approach which is described below.

Given an initial guess of the values of λ_α, λ_δ and $\lambda_{\alpha\delta}$, matrix **R** is treated as known. Under the pretence of known **R**, the solution of $\boldsymbol{\theta}$ can be easily obtained via the weighted regression analysis:

$$\begin{bmatrix} \beta \\ \alpha \\ \delta \end{bmatrix} = \begin{bmatrix} \mathbf{X}^\mathrm{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}^\mathrm{T}\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{Z}^\mathrm{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Z} & \mathbf{Z}^\mathrm{T}\mathbf{R}^{-1}\mathbf{W} \\ \mathbf{W}^\mathrm{T}\mathbf{R}^{-1}\mathbf{X} & \mathbf{W}^\mathrm{T}\mathbf{R}^{-1}\mathbf{Z} & \mathbf{W}^\mathrm{T}\mathbf{R}^{-1}\mathbf{W} \end{bmatrix}^{-1} \times \begin{bmatrix} \mathbf{X}^\mathrm{T}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^\mathrm{T}\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{W}^\mathrm{T}\mathbf{R}^{-1}\mathbf{y} \end{bmatrix} \quad (10)$$

and:

$$\sigma_\varepsilon^2 = \frac{1}{n}(\boldsymbol{y} - \mathbf{X}\beta - \mathbf{Z}\alpha - \mathbf{W}\delta)^\mathrm{T}\mathbf{R}^{-1}(\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{Z}\alpha - \mathbf{W}\delta). \quad (11)$$

Because **R** depends on unknown parameters, it must be updated by the estimates of α, δ and σ_ε^2, and the estimation is then repeated until convergence. This algorithm is extremely fast — only two to three cycles of iteration are required, in contrast to 80–100 iterations in the EM algorithm at the same accuracy. The likelihood ratio test statistic, Λ, is applied to the weighted regression analysis.

Dominant and missing markers

The missing marker problem can be solved easily. A missing marker should be skipped over and the nearest nonmissing markers are picked up. Dominant markers provide partial information which is extracted by using a hidden Markov model algorithm. Details of the hidden Markov model are found in Lander & Green (1987) and Kruglyak *et al.* (1995).

Simulation studies

Eleven equally spaced markers were simulated on a single chromosome segment of length 100 cM. A single QTL was located at position 25 cM. The population size (number of F_2 individuals) was set at 300. Under the null model, the QTL was assigned a value of zero for both the additive and dominance effects. Simulations were repeated 1000 times and the 95 and 99 percentiles of the test statistics were chosen as the empirical critical values for power calculation. Under the alternative model, a nonzero additive effect was simulated while the dominance effect was still set to zero. Simulations were repeated 100 times. Empirical power was calculated by counting the number of runs in which test statistics were greater than the empirical critical values. In all simulations, the variance of the environmental effect was set at $\sigma_\varepsilon^2 = 1.0$.

Each data set was analysed using the three methods: the exact maximum likelihood method (ML), the simple linear regression analysis (REG) and the iteratively reweighted least squares method (IRWLS). Powers and estimation errors of the three methods were compared, based on averages of 100 runs.

Factors considered include the size of the QTL effect, measured by the average effect of gene substitution (α), and the amount of marker information. The average effect of gene substitution was examined at three levels: $\alpha = 0.324$ leading to $h^2 = 0.05$; $\alpha = 0.820$ resulting in $h^2 = 0.25$ and $\alpha = 1.155$ corresponding to $h^2 = 0.40$. The amount of marker information was investigated in four situations: (i) all markers codominant and no missing markers, the highest level of marker information content; (ii) 50 per cent loci in the F_1 parent randomly set to dominant and no missing markers in the offspring; (iii) 50 per cent loci in the F_2 offspring randomly set to missing values; and (iv) 50 per cent loci in the parent dominant and 50 per cent loci in the offspring missing, the lowest level of marker information content.

Average values of the estimated parameters and their standard deviations calculated based on 100 replicated simulations are listed in Tables 1–4. The

Table 1 Comparison of three methods of QTL mapping via Monte Carlo simulations. All markers are codominant and there are no missing values. Parametric values not listed in the table are: QTL position ($\mathrm{cM_A}$) = 25 cM, $\delta = 0$ and $\sigma_\varepsilon^2 = 1.0$. Results are averages of 100 replicated simulations with the standard deviations over the replicates given in parentheses

			Estimate				
α	h^2		$\hat{\alpha}$	$\hat{\delta}$	$\hat{h}^2$	$\mathrm{c\hat{M}_A}$	$\hat{\sigma}_\varepsilon^2$
0.324	0.05	ML	0.345(0.095)	−0.005(0.084)	0.062(0.030)	25.16(9.31)	0.981(0.079)
		REG	0.347(0.098)	−0.002(0.086)	0.062(0.031)	25.45(10.6)	0.988(0.078)
		IRWLS	0.347(0.096)	−0.005(0.085)	0.063(0.031)	25.36(9.65)	0.980(0.079)
0.820	0.25	ML	0.851(0.089)	0.001(0.065)	0.269(0.047)	25.07(3.11)	0.997(0.091)
		REG	0.852(0.089)	0.001(0.066)	0.263(0.044)	25.28(3.14)	1.027(0.089)
		IRWLS	0.851(0.088)	0.001(0.065)	0.268(0.046)	24.97(2.87)	0.998(0.091)
1.155	0.40	ML	1.17(0.068)	0.000(0.072)	0.407(0.043)	24.77(1.91)	1.005(0.078)
		REG	1.17(0.089)	0.000(0.071)	0.393(0.042)	24.79(2.08)	1.065(0.079)
		IRWLS	1.17(0.089)	−0.001(0.071)	0.406(0.045)	24.75(1.95)	1.008(0.079)

IRWLS, iteratively reweighted least squares method.
REG, simple regression method.
ML, maximum likelihood method.

Table 2 Comparison of three methods of QTL mapping via Monte Carlo simulations. There are 50 per cent dominant markers with no missing values. Parametric values not listed in the table are: QTL position (cM_A) = 25 cM, $\delta = 0$ and $\sigma^2_\varepsilon = 1.0$. Results are averages of 100 replicated simulations with the standard deviations over the replicates given in parentheses

			Estimate				
α	h^2		$\hat{\alpha}$	$\hat{\delta}$	$\hat{h}^2$	$c\hat{M}_A$	$\hat{\sigma}^2_\varepsilon$
0.324	0.05	ML	0.343(0.084)	−0.006(0.069)	0.061(0.027)	26.81(15.19)	0.975(0.086)
		REG	0.342(0.084)	−0.004(0.071)	0.060(0.027)	26.44(15.32)	0.980(0.086)
		IRWLS	0.344(0.084)	−0.007(0.068)	0.061(0.027)	26.05(15.06)	0.975(0.086)
0.820	0.25	ML	0.850(0.082)	−0.003(0.065)	0.268(0.041)	25.43(3.32)	0.994(0.086)
		REG	0.853(0.087)	−0.003(0.068)	0.261(0.041)	25.34(3.63)	1.036(0.088)
		IRWLS	0.851(0.084)	−0.003(0.067)	0.269(0.042)	25.41(3.41)	0.994(0.085)
1.155	0.40	ML	1.19(0.087)	0.008(0.065)	0.418(0.045)	24.87(2.03)	0.989(0.080)
		REG	1.20(0.099)	0.008(0.070)	0.403(0.045)	24.94(1.93)	1.070(0.083)
		IRWLS	1.19(0.096)	0.009(0.071)	0.419(0.049)	24.97(2.10)	0.989(0.083)

three methods show virtually no difference with regard to parametric estimation of the additive effect (α), dominance effect (δ) and the location of the QTL (cM_A), which is consistent with Haley & Knott (1992) for the comparison of ML and REG. Another observation is that when both marker information content and the heritability are low, estimation of the QTL position tends to be biased towards the centre of the chromosome for all three methods. This bias occurs because, with smaller QTL effects and less marker information, some of the QTL peaks found may represent, not the simulated QTL but, a Type I error. The position of these Type I errors tends to be randomly distributed along the linkage group; thus the mean position of Type I errors is at the centre of the chromosome and their joint effect, along with some real QTL, is to move the estimated position over all simulated replicates towards the centre of the chromosome. The last, and important, observation is that the simulations verify the theoretical prediction that the simple regression provides a confounded estimation of the true residual variance and part of the QTL variance. The level of confounding increases as the marker information content decreases (from Table 1 to Table 4). The confounding, however, no longer exists in the IRWLS method (see the comparison with ML).

The empirical critical values based on 1000 repeated simulations are given in Table 5, showing very little difference between the three methods.

Table 3 Comparison of three methods of QTL mapping via Monte Carlo simulations. All markers are codominant and there are, on average, 50 per cent missing markers. Parametric values not listed in the table are: QTL position (cM_A) = 25 cM, $\delta = 0$ and $\sigma^2_\varepsilon = 1.0$. Results are averages of 100 replicated simulations with the standard deviations over the replicates given in parentheses

			Estimate				
α	h^2		$\hat{\alpha}$	$\hat{\delta}$	$\hat{h}^2$	$c\hat{M}_A$	$\hat{\sigma}^2_\varepsilon$
0.324	0.05	ML	0.355(0.099)	−0.011(0.091)	0.066(0.032)	27.95(17.53)	0.989(0.078)
		REG	0.357(0.098)	−0.009(0.092)	0.065(0.031)	27.98(16.56)	1.003(0.079)
		IRWLS	0.355(0.099)	−0.011(0.092)	0.065(0.032)	28.41(17.48)	0.989(0.079)
0.820	0.25	ML	0.870(0.096)	−0.010(0.068)	0.278(0.049)	25.41(3.71)	0.991(0.094)
		REG	0.875(0.095)	−0.013(0.073)	0.267(0.044)	25.46(3.91)	1.060(0.091)
		IRWLS	0.871(0.095)	−0.012(0.072)	0.279(0.048)	25.31(3.63)	0.991(0.093)
1.155	0.40	ML	1.19(0.097)	−0.002(0.062)	0.413(0.051)	24.95(2.52)	1.013(0.099)
		REG	1.20(0.101)	−0.003(0.072)	0.386(0.044)	25.17(2.53)	1.150(0.096)
		IRWLS	1.19(0.099)	−0.003(0.071)	0.412(0.051)	25.04(2.59)	1.017(0.098)

Table 4 Comparison of three methods of QTL mapping via Monte Carlo simulations. There are, on average, 50 per cent dominant and 50 per cent missing markers. Parametric values not listed in the table are: QTL position (cM_A) = 25 cM, $\delta = 0$, and $\sigma_\varepsilon^2 = 1.0$. Results are averages of 100 replicated simulations with the standard deviations over the replicates given in parentheses

			Estimate				
α	h^2		$\hat{\alpha}$	$\hat{\delta}$	$\hat{h}^2$	$c\hat{M}_A$	$\hat{\sigma}_\varepsilon^2$
0.324	0.05	ML	0.345(0.115)	−0.010(0.108)	0.066(0.039)	30.24(20.06)	0.973(0.089)
		REG	0.345(0.116)	−0.007(0.105)	0.063(0.037)	29.81(18.28)	0.997(0.083)
		IRWLS	0.342(0.115)	−0.008(0.104)	0.064(0.037)	30.01(19.66)	0.975(0.086)
0.820	0.25	ML	0.877(0.114)	−0.008(0.095)	0.284(0.059)	25.00(4.68)	0.983(0.089)
		REG	0.883(0.117)	−0.005(0.101)	0.268(0.053)	24.70(4.93)	1.078(0.087)
		IRWLS	0.877(0.116)	−0.008(0.099)	0.284(0.059)	25.09(4.66)	0.985(0.089)
1.155	0.40	ML	1.214(0.103)	−0.005(0.060)	0.430(0.049)	24.72(3.82)	0.977(0.090)
		REG	1.242(0.116)	0.003(0.079)	0.403(0.046)	24.85(3.62)	1.146(0.106)
		IRWLS	1.227(0.112)	0.005(0.077)	0.437(0.057)	24.76(3.85)	0.970(0.100)

These critical values, however, are different across different levels of marker information contents. The highest critical values occur when all markers are codominant and there is no missing marker. These empirical critical values are then used to compute the empirical statistical powers for the three methods (see Table 6). Again, the three methods have virtually identical statistical powers.

To view the details of the comparison of the three methods, the likelihood ratio test statistics of the three methods are plotted against the chromosome position. Figure 1 shows the likelihood ratio profiles (average of 100 runs) at three levels of heritability in the situation where 50 per cent of the marker loci in the offspring are missing. The IRWLS method is nearly indistinguishable from the ML method, and both methods have higher testing signals than the REG method. Figure 1(a–c) also shows that the difference between ML (IRWLS) and REG increases as the heritability increases. When the heritability is fixed at 0.25, the likelihood ratio profiles (average of 100 runs) of the three methods are compared at each of the four levels of marker information content. Again, IRWLS and ML are virtually identical but both are different from that of the simple regression method. When all markers are codominant and there is no missing marker, the test statistics of the three methods are identical at marker loci but different off the markers. The ML(IRWLS) curves shows significant discontinuity at marker loci (Fig. 2a). When 50 per cent of the marker loci are dominant and there is no missing marker, the discontinuity of the ML (IRWLS) still exists but becomes less obvious (Fig. 2b). The test statistics of the ML (IRWLS) at the marker loci are now different from those of the REG. As the marker information content decreases, the discontinuity of ML (IRWLS) disappears (Fig. 2c,d).

In conclusion, ML and IRWLS show no difference but both differ from REG. However, the difference is only detectable at the micro level. The advantage of ML and IRWLS over the REG is that

Table 5 Empirical critical values of the test statistic for testing the presence of a QTL on a chromosome of length 100 cM

	95 per cent			99 per cent		
	ML	REG	IRWLS	ML	REG	IRWLS
Codominant markers	10.52	10.49	10.60	15.85	15.79	15.95
Dominant markers	9.98	9.92	10.04	13.50	13.41	13.63
Missing markers	9.73	9.70	9.77	12.99	12.75	13.07
Missing and dominant	9.75	9.96	9.77	13.08	14.10	13.07

Table 6 Empirical powers of three methods for QTL detection under various situations. α is the Type I error rate

		$\alpha = 0.05$			$\alpha = 0.01$		
Marker	h^2	ML	REG	IRWLS	ML	REG	IRWLS
Codominant	0.05	0.82	0.82	0.82	0.62	0.62	0.62
	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	0.40	1.00	1.00	1.00	1.00	1.00	1.00
Dominant(D)	0.05	0.81	0.80	0.80	0.53	0.52	0.52
	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	0.40	1.00	1.00	1.00	1.00	1.00	1.00
Missing (M)	0.05	0.79	0.79	0.79	0.49	0.48	0.49
	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	0.40	1.00	1.00	1.00	1.00	1.00	1.00
Missing and dominant	0.05	0.68	0.67	0.68	0.38	0.37	0.38
	0.25	1.00	1.00	1.00	1.00	1.00	1.00
	0.40	1.00	1.00	1.00	1.00	1.00	1.00

they provide a true estimate of σ_ε^2. The ML, however, is many times slower than REG because many cycles of iterations (≈ 80) are required for the EM algorithm to converge. In contrast, the IRWLS algorithm only requires two to three cycles of iterations to converge, about two or three times slower than the REG but 30–40 times faster than the ML. Of course, the comparisons in computing speed are based on the algorithms adopted here in this particular research. If other algorithms had been used, such as the Newton–Raphson iteration for the ML and the regression on marker-type algorithm (Whittaker *et al.*, 1996) for the REG, the comparisons would produce quantitatively different results, but the conclusion is not anticipated to change qualitatively.

Discussion

In an earlier paper (Xu, 1995) it was pointed out that estimation of the residual variance with the simple regression method is confounded by part of the QTL variance. A simple way was also provided to separate the confounding variances in a backcross design. However, simply correcting the estimated residual variance does not necessarily correct the difference in test statistic between the REG and ML. The improved regression method (IRWLS) corrects both deficiencies yet retains the simplicity and rapidity of the regression method. With the current improvement, the regression method can now be safely applied to all data analyses without any concerns.

The (revised) regression method is particularly useful for permutation tests (Churchill & Doerge, 1994) and construction of confidence intervals by bootstrapping (Visscher *et al.*, 1996) because thousands of analyses of resampled data sets are required. In addition to its simplicity and speed allowing resampling and permutation, the regression method has another major strength that makes it very valuable for use on real data: it can be used to fit relatively complex models and thus include multiple or interacting QTL effects. The weighted regression method retains this strength of regression. If the distribution of residual error is known, the ML is optimal. In some situations, the distribution is unknown and normality is only an approximation, so the ML is also an approximate method. In contrast, REG and IRWLS are independent of the distribution of the residual error. Combined with the permutation test, the regression methods are actually nonparametric methods which may be applied to a wider range of data.

The significant discontinuity of the likelihood ratio profiles at fully informative markers is a drawback of the ML and IRWLS compared with the REG. The peaks within marker intervals have a clear pattern, that is they all face in the direction where the true QTL resides. The strong discontinuity is analogous with linkage analysis (of markers), where the likelihood ratio of zero recombination can show very strong discontinuity (to minus infinity) at a marker once one or more recombination events have been observed, because the probability that the two markers are fully linked is zero. The difference between quantitative change and qualitative change can also explain the discontinuity. When the putative QTL position is off the markers, all three genotypes of the QTL are possible so that the population actually has a mixture of three distributions, no matter how likely a particular genotype is (e.g. 0.999).

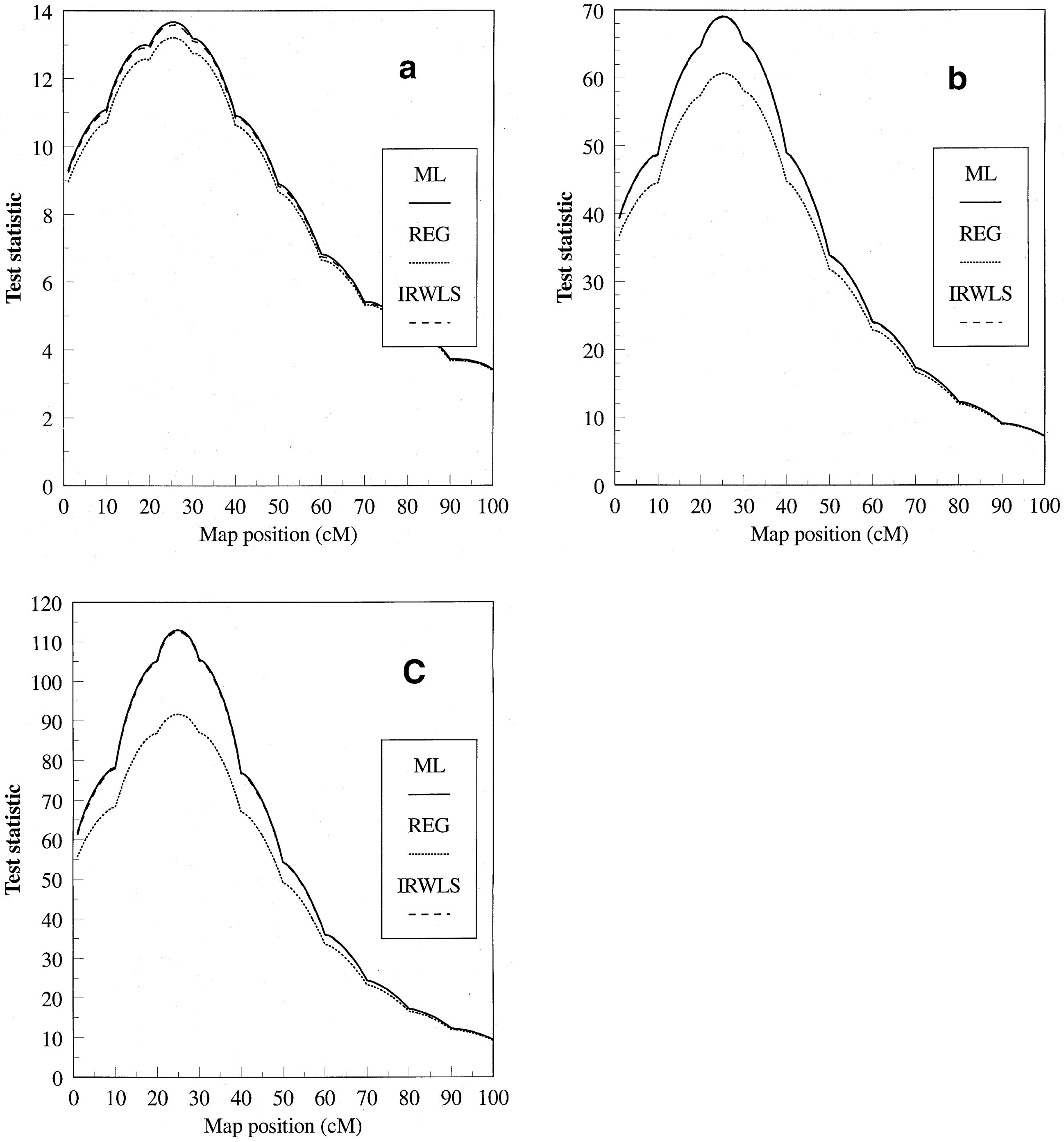

Fig. 1 Comparison of the likelihood ratio profiles (test statistics) of three methods, maximum likelihood (ML), simple regression (REG) and iteratively reweighted least squares (IRWLS). Eleven codominant markers (with a 50 per cent chance of missing) are equally spaced along a chromosome of 100 cM. A single QTL resides at position 25 cM. (a) Variation explained by the QTL is 0.05; (b) variation explained by the QTL is 0.25; (c) variation explained by the QTL is 0.40.

When the putative position moves to a marker locus, the genotype is actually observed so that the population has a single distribution. The observed genotype has occurred with probability 1.0. The difference between 0.999 and 1.0 is a qualitative change, whereas the change from 0.998 to 0.999 is a quanti-

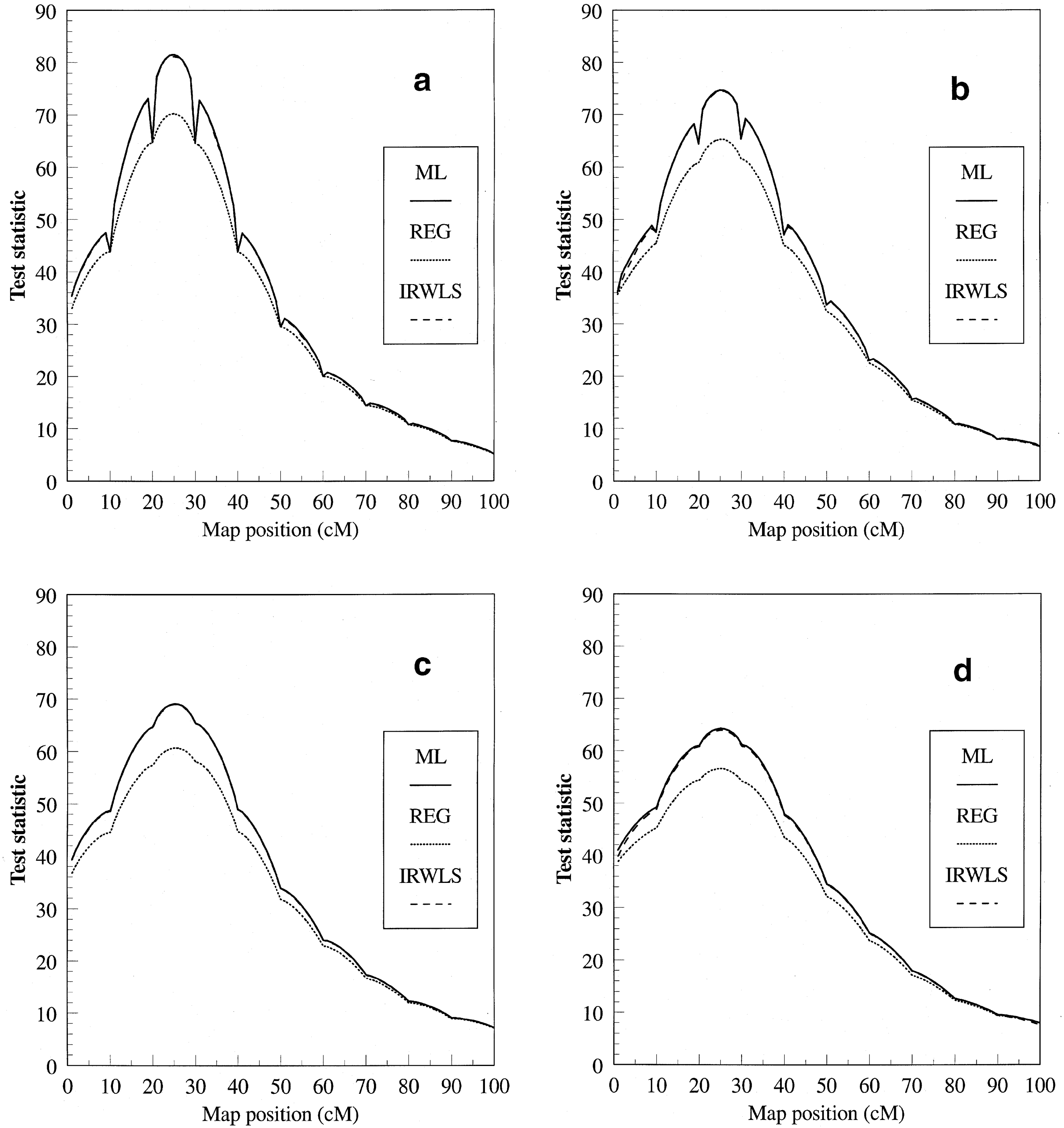

Fig. 2 Comparison of the likelihood ratio profiles (test statistics) of three methods, maximum likelihood (ML), simple regression (REG) and iteratively reweighted least squares (IRWLS). Eleven markers are equally spaced on a chromosome of 100 cM. A single QTL explaining 25 per cent of the phenotypic variation resides at position 25 cM. (a) All markers codominant and no missing markers; (b) 50 per cent of the markers dominant and no missing markers; (c) 50 per cent of the markers missing; (d) 50 per cent dominant and 50 per cent missing markers.

tative change. The ML and IRWLS methods are extremely sensitive to the qualitative change, whereas the REG method does not distinguish between the two types of change.

It should be noted that the test statistic for the weighted least squares method (IRWLS) cannot be chosen as the reduction of the weighted residual sum of squares. This is in contrast to the simple regression method, where the QTL location is chosen at the position with the minimum residual sum of squares. The residual sum of squares for the IRWLS method is:

$$\mathrm{RSS} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\alpha-\mathbf{W}\delta)^{\mathrm{T}}\mathbf{R}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\alpha-\mathbf{W}\delta)$$

which can be made as small as possible by increasing the values of the diagonal elements of **R**. The diagonal elements, however, are proportional to the uncertainty of the genotype of a putative position, i.e. the variance of the independent variables, z_j and w_j, as seen in eqn (8). The uncertainty, nonetheless, takes its maximum value at a position with minimum information content, in the middle of an interval. Therefore, the estimated QTL position will be biased towards the centre of an interval if RSS is used as the test statistic. Therefore, the likelihood ratio has been chosen as the test statistic in this paper. However, other test statistics might be more appropriate, and this deserves further investigation.

Acknowledgements

The author thanks two anonymous reviewers for their helpful comments and criticisms on an earlier version of the manuscript. The reviewers provided elegant explanations for the bias in QTL position estimation and the discontinuity of the test statistic profiles in the ML, which have been incorporated in the revision. This research was supported by the National Institutes of Health Grant GM55321–01 and the USDA National Research Initiative Competitive Grants Program 95–37205–2313.

References

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

DEMPSTER, A. P., LAIRD, N. M. AND RIBIN, D. B. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B.*, **39**, 1–38.

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.

KRUGLYAK, L., DALY, M. J. AND LANDER, E. S. 1995. Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. *Am. J. Hum. Genet.*, **56**, 519–527.

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 GREEN, P. 1987. Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.*, **84**, 2363–2367.

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

VISSCHER, P. M., THOMPSON, R. AND HALEY, C. S. 1996. Confidence intervals in QTL mapping by bootstrapping. *Genetics*, **143**, 1013–1020.

WHITTAKER, J. C., THOMPSON, R. AND VISSCHER, P. M. 1996. On the mapping of QTL by regression of phenotype on marker type. *Heredity*, **77**, 23–32.

XU, S. 1995. A comment on the simple regression method for interval mapping. *Genetics*, **141**, 1657–1659.