# On the Differences Between Maximum Likelihood and Regression Interval Mapping in the Analysis of Quantitative Trait Loci

**Chen-Hung Kao**

*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China*

Manuscript received July 26, 1999
Accepted for publication May 30, 2000

ABSTRACT

The differences between maximum-likelihood (ML) and regression (REG) interval mapping in the analysis of quantitative trait loci (QTL) are investigated analytically and numerically by simulation. The analytical investigation is based on the comparison of the solution sets of the ML and REG methods in the estimation of QTL parameters. Their differences are found to relate to the similarity between the conditional posterior and conditional probabilities of QTL genotypes and depend on several factors, such as the proportion of variance explained by QTL, relative QTL position in an interval, interval size, difference between the sizes of QTL, epistasis, and linkage between QTL. The differences in mean squared error (MSE) of the estimates, likelihood-ratio test (LRT) statistics in testing parameters, and power of QTL detection between the two methods become larger as (1) the proportion of variance explained by QTL becomes higher, (2) the QTL locations are positioned toward the middle of intervals, (3) the QTL are located in wider marker intervals, (4) epistasis between QTL is stronger, (5) the difference between QTL effects becomes larger, and (6) the positions of QTL get closer in QTL mapping. The REG method is biased in the estimation of the proportion of variance explained by QTL, and it may have a serious problem in detecting closely linked QTL when compared to the ML method. In general, the differences between the two methods may be minor, but can be significant when QTL interact or are closely linked. The ML method tends to be more powerful and to give estimates with smaller MSEs and larger LRT statistics. This implies that ML interval mapping can be more accurate, precise, and powerful than REG interval mapping. The REG method is faster in computation, especially when the number of QTL considered in the model is large. Recognizing the factors affecting the differences between REG and ML interval mapping can help an efficient strategy, using both methods in QTL mapping to be outlined.

SINCE Lander and Botstein (1989) initiated an interval mapping method for systematically searching the entire genome for quantitative trait loci (QTL) using molecular genetic marker data, many efforts have been made to enhance the precision, accuracy, and power of QTL mapping. They include the extension of the statistical model from one-QTL to multiple-QTL (Jansen 1993; Zeng 1994; Kao *et al.* 1999), incorporation of random effects in the model (Hoeschele and VanRaden 1993a,b), ease of computation (Haley and Knott 1992; Martinez and Curnow 1992; Xu 1998a,b), generalization to different experimental designs (Carbonell *et al.* 1992; Jiang and Zeng 1997; Song *et al.* 1999; Zeng *et al.* 2000) and to multiple and categorical trait analyses (Hackett and Weller 1995; Jiang and Zeng 1995; Henshall and Goddard 1999), the use of permutation tests, and Bayesian estimation (Doerge and Churchill 1996; Satagopan *et al.* 1996; Sillanpaa and Arjas 1999) in QTL mapping.

The likelihood of the interval mapping model is generally a finite normal mixture (Lander and Botstein 1989; Jansen 1993; Zeng 1994; Kao *et al.* 1999). In the computation of the maximum-likelihood estimates (MLE) of the finite normal mixture model, the iterative expectation-maximization (EM) algorithm (Dempster *et al.* 1977) is broadly applicable as Newton-Raphson and Fisher's score methods may turn out to be complicated. When the number of QTL considered in the model increases, the numbers of mixture components and parameters in the likelihood increase dramatically. As a result, maximization of the likelihood through the EM algorithm could become difficult to obtain; moreover, when mapping the entire genome for QTL, the search needs to be performed at every position of the genome. Therefore, the ML estimation by the EM algorithm is often regarded to be complex in analysis and computationally expensive for QTL mapping (Haley and Knott 1992; Satagopan *et al.* 1996; Xu 1998a,b). In view of these difficulties, regression (REG) interval mapping, which regresses the quantitative trait value on the conditional expected genotypic value, was proposed to approximate ML interval mapping to save computation time at one or multiple genomic positions (Haley and Knott 1992; Martinez and Curnow 1992). Although REG interval mapping lacks some attractive properties, such as consistency and asymptotic efficiency, as compared to ML interval mapping in statistical inference,

*Author e-mail:* chkao@stat.sinica.edu.tw

TABLE 1

**Conditional probabilities of a putative QTL given flanking marker genotypes for a backcross population**

| Marker type | Marker genotype | QTL genotype | |
|---|---|---|---|
| | | *QQ* | *Qq* |
| 1 | *MN*/*MN* | $\frac{(1-r_1)(1-r_2)}{1-r}$ | $\frac{r_1r_2}{1-r}$ |
| 2 | *MN*/*Mn* | $\frac{(1-r_1)r_2}{r}$ | $\frac{r_1(1-r_2)}{r}$ |
| 3 | *MN*/*mN* | $\frac{r_1(1-r_2)}{r}$ | $\frac{(1-r_1)r_2}{r}$ |
| 4 | *MN*/*mn* | $\frac{r_1r_2}{1-r}$ | $\frac{(1-r_1)(1-r_2)}{1-r}$ |

$r$ is the recombination fraction between the two flanking markers $M$ and $N$. $r_1$ ($r_2$) is the recombination fraction between the putative QTL and the left (right) marker $M$ ($N$).

and may suffer from the lack of interpretability in terms of the genetic model (Haley and Knott 1992; Jansen 1993), it is often claimed that the two approaches provide virtually similar or identical estimates and test statistics in QTL mapping (Haley and Knott 1992; Xu 1998a,b). As a consequence, the REG method has been widely accepted and applied to QTL mapping studies by many researchers (Haley *et al.* 1994; Whittaker *et al.* 1996; Xu 1996, 1998a,b; Goffinet and Mangin 1998; Lebreton *et al.* 1998; Dupuis and Siegmund 1999; Rebai and Goffinet 2000).

Although REG may approximate ML interval mapping well in some cases as shown by Haley and Knott (1992) and Xu (1998a,b), their differences in the estimation of QTL parameters could be significant for practical QTL mapping as shown in this article. Unfortunately, there are few attempts to investigate these differences in the literature. Xu (1995) pointed out that the estimation of residual variance by REG interval mapping is biased. In this article, the differences between the two approaches in the estimation of and testing for QTL parameters due to several factors, such as heritability, size of interval, relative QTL position in an interval, the difference between QTL effects, epistasis, and linkage between QTL, are investigated both analytically and numerically by simulation. With the understanding of the factors affecting the differences between the two methods, a more efficient, precise, and powerful strategy using both methods can be explored in QTL mapping. The QTL mapping properties under these factors are also investigated and discussed.

## MAXIMUM-LIKELIHOOD INTERVAL MAPPING

The differences between the ML and REG interval mappings can be illustrated by investigating the differences between their estimators of mean, genetic effects, and residual variance. To simplify the explanation of their differences, a one-QTL model for a backcross population is first used as an example. Their differences under a multiple-QTL model are discussed later. The one-QTL ML mapping model can be written as

$$y_i = \mu + ax_i^* + \varepsilon_i, \quad i = 1, 2, \ldots, n, \qquad (1)$$

where $y_i$ is the quantitative trait value of individual $i$, $\mu$ is the mean, $a$ is the effect of QTL Q, $x_i^*$, taking a value $\frac{1}{2}$ ($-\frac{1}{2}$) for homozygote *QQ* (heterozygote *Qq*), denotes the genotype of Q, and $\varepsilon_i$ is the environmental deviation and is assumed to follow $N(0, \sigma^2)$. Although the genotype of Q for an individual is usually unobserved and could be *QQ* or *Qq*, its distribution can be inferred from its flanking marker genotypes. Suppose the flanking markers are $M$ and $N$. Then, there are four types of marker genotypes, type 1 *MN*/*MN*, type 2 *MN*/*Mn*, type 3 *MN*/*mN*, and type 4 *MN*/*mn*, as shown in Table 1. Given the four marker genotypes, the conditional probabilities for QTL genotypes *QQ* and *Qq*, denoted by $p_{i1}$ and $p_{i2}$, respectively, at a position between the markers can be calculated based on Haldane's mapping function (Haldane 1919), and they are listed in Table 1.

Since the QTL genotype $x_i^*$ could be homozygote ($\frac{1}{2}$) or heterozygote ($-\frac{1}{2}$) for an individual, the likelihood is then a normal mixture with mixing proportions equivalent to the conditional probabilities $p_{i1}$ and $p_{i2}$. For $n$ individuals in the sample, the likelihood of the model in Equation 1 is

$$L(\theta|Y,X) = \prod_{i=1}^{n}\left[\sum_{j=1}^{2} p_{ij}N(\mu_{ij},\sigma^2)\right], \qquad (2)$$

where $\theta$ denotes parameters ($p_{ij}$, $\mu$, $a$, $\sigma^2$), $Y$ and $X$ denote the trait value and marker genotypes, $N(\mu_{ij}, \sigma^2)$ denotes the normal density function with mean $\mu_{ij}$ and variance $\sigma^2$, and

$$\mu_{i1} = \mu + \frac{a}{2} \quad \text{and} \quad \mu_{i2} = \mu - \frac{a}{2}$$

are genotypic values of *QQ* and *Qq*. In estimation, this normal mixture model can be treated as an incomplete data problem (Little and Rubin 1987), which regards the QTL genotype $x_i^*$ as missing data and trait $y_i$ and markers $X_i$ as observed data, and the EM algorithm can be implemented to maximize the likelihood and obtain the MLE.

Let the probability distribution of missing data $x_i^*$ as

$$g(x_i^*) = \begin{cases} p_{i1} & \text{if } x_i^* = 1/2 \\ p_{i2} & \text{if } x_i^* = -1/2. \end{cases}$$

The conditional distribution of the observed data, $y_i$ and $X_i$, given the missing data $x_i^*$ can be considered as an independent sample from a population such that $y_i|(\theta, X_i, x_i^*) \sim N(\mu + ax_i^*, \sigma^2)$, and the EM algorithm can be used to obtain the MLE. At a given position, $p_{ij}$'s can be determined. By the definition of the EM algorithm, the iteration of the EM-step for obtaining $\mu$, *a, and* $\sigma^2$ *proceeds as follows:*

**E-step:** The posterior probabilities of the QTL genotypes $x_i^*$'s of the $n$ individuals are updated as

$$\pi_{i1} = p(x_i^* = \frac{1}{2}|y_i,X_i,\theta) = \frac{p_{i1}N(\mu_{i1},\sigma^2)}{p_{i1}N(\mu_{i1},\sigma^2) + p_{i2}N(\mu_{i2},\sigma^2)}$$

and

$$\pi_{i2} = p(x_i^* = \frac{1}{2}|y_i,X_i,\theta) = \frac{p_{i2}N(\mu_{i2},\sigma^2)}{p_{i1}N(\mu_{i1},\sigma^2) + p_{i2}N(\mu_{i2},\sigma^2)}$$

for $i = 1, 2, \ldots, n$ (see Kao and Zeng 1997 for the detailed procedure of derivation). Note that $\pi_{i1}$ for individual $i$ is a function of $p_{i1}$, $p_{i2}$, $y_i$, $\mu$, $a$, and $\sigma^2$. It is important to clarify the relationship between the conditional probability $p_{i1}$ and conditional posterior probability $\pi_{i1}$ of the QTL genotype in the comparison of ML and REG interval mappings. It is shown later that the more similarity between $\pi_{i1}$ and $p_{i1}$ for each $i$, the better the approximation of REG to ML interval mapping. Note that $p_{i1} = \pi_{i1}$ if $p_{i1} = 1$ or $p_{i1} = 0$ (*i.e.*, the QTL is located at the marker) or $a = 0$. If $p_{i1} \approx 1$ or $p_{i2} \approx 0$ or $a \approx 0$, then $\pi_{i1} \approx p_{i1}$.

**M-step:** Find $\mu$, $a$, and $\sigma^2$ to satisfy

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_i^* a) \tag{3}$$

$$a = \frac{\sum_{i=1}^{n} w_i^*(y_i - \mu)}{\sum_{i=1}^{n} z_i^*} \tag{4}$$

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \mu)^2 - 2(y_i - \mu)w_i^* a + z_i^* a^2\right], \tag{5}$$

where

$$w_i^* = E(x_i^*|y_i,X_i,\theta) = \frac{1}{2}\pi_{i1} - \frac{1}{2}\pi_{i2} = \pi_{i1} - \frac{1}{2}$$

and

$$z_i^* = E(x_i^{*2}|y_i,X_i,\theta) = \left(\frac{1}{2}\right)^2\pi_{i1} + \left(-\frac{1}{2}\right)^2\pi_{i2} = \frac{1}{4}$$

are the conditional posterior expectations of $x_i^*$ and $x_i^{*2}$ given $y_i$ and $X_i$, respectively. In each iteration, new estimates of $\mu$, $a$, and $\sigma^2$ are obtained in the M-step. These new estimates are then used to obtain new $\pi_{i1}$'s and $\pi_{i2}$'s for the next iteration. The converged values in the iteration are the MLE. A disadvantage of the EM algorithm was that it did not provide the estimates of the covariance matrix of the MLE. However, this disadvantage can be easily removed by using appropriate methods, such as by Louis (1982) and Meng and Rubin (1991), associated with the EM algorithm. The null hypothesis, $H_0$, $a = 0$, for the existence of a QTL is tested by a likelihood-ratio test (LRT), $-2\log_e(L_0/L_1)$, where $L_0$ and $L_1$ are the maximum likelihoods under $H_0$ and no restriction. The larger the LRT statistic at a testing position, the more likely the existence of a QTL at that position.

## REGRESSION INTERVAL MAPPING

Haley and Knott (1992) commented on ML interval mapping that the iterative procedure of the EM algorithm in obtaining the MLE can be complex and computationally slow to converge. Therefore, they developed REG interval mapping to approximate ML interval mapping for mapping QTL. They claimed that the REG method can ease the computation and produce very similar results as those obtained by the ML method. The one-QTL REG interval mapping model for a backcross population can be formulated as

$$y_i = \mu + aw_i + \varepsilon_i, \tag{6}$$

where $\mu$, $a$, and $\varepsilon_i$ have the same definitions as the model in Equation 1, and

$$w_i = E(x_i^*|X_i) = \frac{1}{2}p_{i1} - \frac{1}{2}p_{i2} = p_{i1} - \frac{1}{2}$$

is the conditional expectation of the QTL genotype given the flanking marker genotype. By treating $w_i$ as fixed, the model is a regression model, and this method is called REG interval mapping in QTL mapping. In estimation, both least-squares and maximum-likelihood techniques can be implemented to estimate $\mu$, $a$, and $\sigma^2$ in Equation 6. Least-squares estimates (LSE) of $\mu$ and $a$ are the solutions of

$$\mu = \frac{1}{n}\sum_{i=1}^{n}(y_i - w_i a) \tag{7}$$

$$a = \frac{\sum_{i=1}^{n} w_i(y_i - \mu)}{\sum_{i=1}^{n} z_i}, \tag{8}$$

where

$$z_i = w_i^2 = \left(p_{i1} - \frac{1}{2}\right)^2 = \frac{1}{4} - p_{i1}(1 - p_{i1}).$$

Note that the estimates of the regression model will fail if $z_i = 0$ ($p_{i1} = p_{i2} = 1|2$) for every $i$. However, this situation will not occur because $p_{i1} \neq p_{i2}$ for individuals with type 1 (*MN*/*MN*) and type 4 (*MN*/*mn*) flanking marker genotypes in the backcross population. The LSE of $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n} [(y_i - \mu)^2 - 2(y_i - \mu) w_i a + z_i a^2], \qquad (9)$$

where $n - 2$ is the degree of freedom for the residual sum of squares. The likelihood of the REG mapping model is a normal density

$$L_{\text{REG}}(\theta | Y, X) = \prod_{i=1}^{n} N(\mu + a w_i, \sigma^2) \qquad (10)$$

rather than a normal mixture density. The mixing proportion of $p_{ij}$'s in the ML mapping likelihood (Equation 2) is blended into $w_i$ of Equation 10. If the maximum-likelihood principle is used in estimation, the MLE of $\mu$ and $a$ for maximizing Equation 10 are the same as Equations 7 and 8. The MLE of $\sigma^2$ has a divisor $n$ instead of $n - 2$ in Equation 9.

## DIFFERENCES BETWEEN ML AND REG INTERVAL MAPPING

By comparing the solution sets between ML and REG interval mappings, it can be seen that the two solution sets have similar expressions, but different contents. In the REG method, the conditional expectations of QTL genotype, $w_i$ and $z_i$, are used in estimation. In the ML method, the conditional posterior expectations of QTL genotype, $w_i$* and $z_i$*, play the same role in estimation. The conditional expectations consider only the conditional probabilities of QTL genotypes $p_{i1}$'s, and the conditional posterior expectations consider the posterior probabilities $\pi_{i1}$'s. It can be seen that the posterior probability $\pi_{i1}$ also utilizes phenotypic information as well as marker information. Intuitively, the ML method can provide better estimates than the REG method because $\pi_{i1}$ is more informative than $p_{i1}$. Analytically, the differences between ML and REG interval mapping in estimation will depend on the differences between the two kinds of expectations. The two kinds of expectations are equivalent if and only if $\pi_{i1} = p_{i1}$ and $p_{i1} = 1$ (or $p_{i1} = 0$) for each $i$ (the QTL is located at a marker). How good the approximation of REG to ML interval mapping is depends on the similarity between $p_{i1}$'s and $\pi_{i1}$'s. Investigating the factors affecting the similarity between $\pi_{i1}$ and $p_{i1}$ can lead to identifying the differences between the two methods. These factors include (1) proportion of variance explained by a QTL (size of a QTL), (2) the relative QTL position within an interval, and (3) the size of the interval flanking the QTL.

**Proportion of variance explained by a QTL:** If the proportion of variance explained by a QTL is small, the ratio of QTL effect $a$ to the environmental deviation $\sigma$ ($a/\sigma$) will be small. Consequently, the densities of normal mixture components are about the same for different genotypes, *i.e.*,

$$N(\mu_{i1}, \sigma^2) \approx N(\mu_{i2}, \sigma^2) \approx N(\mu, \sigma^2),$$

and $p_{i1} \approx \pi_{i1}$. The extreme case is $a = 0$ ($\mu_{i1} = \mu_{i2} = \mu$) and $p_{i1} = \pi_{i1}$. Therefore, REG mapping can approximate ML mapping well when the proportion is low (the QTL effect $a$ is small). When the proportion is high, QTL effect $a$ becomes relatively large when compared with the environmental deviation $\sigma$, and the difference between the two normal densities can become significant. As a result, the approximation of REG to ML interval mapping may not be good for QTL with large effect.

**Relative QTL location in an interval:** If a QTL is located on the boundary of a marker interval, $p_{i1}$ is close to 1 or 0, and the conditional and posterior probabilities will be similar ($p_{i1} \approx \pi_{i1}$). When the QTL position shifts from the boundary toward the middle of an interval, $p_{i1}$ and $\pi_{i1}$ become more dissimilar to each other. If the QTL is located in the middle, individuals with type 2 or 3 flanking marker genotype have $p_{i1} = 0.5$, and $p_{i1}$ and $\pi_{i1}$ will be the most dissimilar. Consequently, the approximation of REG to ML mapping will be better when the QTL is located near the boundary, but it becomes poor as the location moves toward the middle of an interval.

**Interval size:** There are four types of flanking marker genotypes (Table 1). Types 1 and 4 are nonrecombinant, and types 2 and 3 are recombinant. Given a position in an interval, the conditional probability $p_{i1}$ for *QQ* will be closer to 1 or 0, *i.e.*, $p_{i1}$ can be closer to $\pi_{i1}$, for nonrecombinant individuals than recombinant individuals. As there are more nonrecombinant flanking genotypes in a narrow interval than in a wider interval, the approximation of REG to ML mapping consequently is better for a QTL located in a narrow interval than in a wider interval. Therefore, if QTL are located in the dense marker region, the differences between the two methods will be minor.

## MULTIPLE-QTL MODEL

For the one-QTL model, it has been shown that the approximation of REG to ML interval mapping depends on the similarity between the conditional probability $\pi_{i1}$ and conditional posterior probability $\pi_{i1}$ for each $i$. The same argument also applies to the multiple-QTL model. When multiple, say, $m$ QTL are considered, the multiple interval mapping (MIM; Kao *et al.* 1999) model can be written as

$$y_i = \mu + \sum_{j=1}^{m} a_j x_{ij}^* + \sum_{j \neq k}^{m} \delta_{jk} (I_{jk} x_{ij}^* x_{ik}^*) + \varepsilon_i, \qquad (11)$$

where $x_{ij}$* denotes the genotype of QTL $Q_j$, $a_j$ and $I_{jk}$ are the main and epistatic effects, $\delta_{jk}$ is an indicator variable for indicating whether the epistasis between $Q_j$ and $Q_k$ is present or not, and $\varepsilon_i$ is the environmental deviation.

For $m$ QTL, there are $2^m$ possible QTL genotypes; hence there are $2^m$ corresponding genotypic values, $\mu_{ij}$'s, with probabilities $p_{ij}$'s, $j = 1, 2, \ldots, 2^m$. The likelihood of the multiple-QTL model is then a $2^m$ normal mixture

$$L(\theta|Y,X) = \prod_{i=1}^{n}\left[\sum_{j=1}^{2m} p_{ij}\,N(\mu_{ij},\sigma^2)\right]. \qquad (12)$$

It seems that the derivation of the MLE of $\mu$, $a_1$, $a_2$, . . . , $a_m$, $I_{jk}$, and $\sigma^2$ and their asymptotic variance-covariance matrix is tedious as the number of QTL considered increases in the model. However, this tedious estimation problem can be easily solved by the general formulas proposed by Kao and Zeng (1997) by expanding the genetic design matrix **D** and conditional probability matrix **Q** according to the number and positions of testing QTL. These two matrices play the same role as the matrix of independent variable $X$ in regression. Given $X$ matrix in multiple regression, the estimates of regression coefficients and the asymptotic variance-covariance matrix can be easily obtained by formulas $\beta = (X'X)^{-1} X'Y$ and $V = (X'X)^{-1}\sigma^2$. Given these two matrices **D** and **Q** in the multiple-QTL mapping model, the derivation of the MLE and the asymptotic variance-covariance matrix can be systematically obtained by the general formulas. Given the testing QTL positions, $p_{ij}$'s can be determined. According to the general formulas, in the E-step, the $2^m$ posterior probabilities of QTL genotypes for $n$ individuals,

$$\pi_{ij} = \frac{p_{ij}\,N(\mu_{ij},\sigma^2)}{\sum_{j=1}^{2^m} p_{ij}\,N(\mu_{ij},\sigma^2)}; \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, 2^m,$$

are updated. In the M-step, the solutions of the parameter estimation are in the closed form as shown in Kao and Zeng (1997). The asymptotic variances of QTL positions and effects can also be obtained using the general formulas.

The REG interval mapping model for taking $m$ QTL into account can be written as

$$y_i = \mu + \sum_{j=1}^{m} a_j w_{ij} + \sum_{j\neq k}^{m} \delta_{jk}(I_{jk} w_{ij} w_{ik}) + \varepsilon_i.$$

In the model, $w_{ij}$ is the conditional expectation of $Q_j$ given its flanking markers. The LSE of $\mu$, $a_1$, $a_2$, . . . , $a_m$, $I_{jk}$, and $\sigma^2$ as well as their asymptotic variances can be obtained using the standard least-squares technique.

**Differences due to QTL effects, epistasis, and linkage:** By the same argument, it is required that $p_{ij} \approx \pi_{ij}$ for each $i$ and $j$ for REG interval mapping to approximate ML interval mapping well in the multiple-QTL model. Besides the factors, such as the proportion explained by QTL, the relative QTL position in an interval, and interval sizes, discussed in the previous section, the relative sizes of genotypic values of the $2^m$ possible genotypes, $\mu_{ij}$'s, can affect the approximation of REG to ML interval mapping in the multiple-QTL model. If $\mu_{ij}$'s are dissimilar to each other (more disperse), $\pi_{ij}$'s can be more dissimilar to $p_{ij}$'s, and the differences between the two methods can become large. The difference between the sizes of QTL effects and the strength of epistasis between QTL seems to be an appropriate measure to quantify the dissimilarity between the $2^m$ genotypic values. If QTL effects differ from each other significantly or epistasis between QTL is strong, $\pi_{ij}$'s tend to be dissimilar to $p_{ij}$'s. Consequently, the differences between REG and ML interval mapping will be larger if QTL effects differ significantly or the interaction between QTL is strong.

If QTL are linked, they are correlated. Their correlation is $1 - 2r$, where $r$ is the recombination fraction between QTL. As QTL (predictors) in the model are correlated, the effects of collinearity, on modeling such as imprecise estimation and losing power in testing for individual parameters, will occur (Neter *et al.* 1990). When detecting closely linked QTL using the multiple-QTL model, the correlations between the conditional expectations of QTL, $w_{ij}$'s, in the REG model tend to be higher than those between the conditional posterior expectations, $w_{ij}^*$'s, in the ML model. As a result, the REG method tends to give estimates with large SD and be less powerful in testing for closely linked QTL (to separate closely linked QTL). The closer they are, the worse the approximation of the REG to ML method will be.

## SIMULATION STUDIES

Simulations were performed to verify the effects of the above factors, such as the proportion of variance explained by QTL, interval size, QTL position, the difference between QTL effects, epistasis, and linkage, on the approximation of REG to ML interval mapping. Assume two unlinked epistatic QTL with effects ($a_1 = 1$, $a_2 = 1$, $I_{12} = 1$) that affected a quantitative trait of interest in a backcross population (epistasis contributes 11.11% of the total genetic variation). For simplicity, 10 equally spaced marker intervals were simulated for each chromosome. Four proportions of variance explained by QTL ($h^2$'s), 0.01, 0.1, 0.3, and 0.5, and three different interval sizes, 10, 20, and 40 cM, are simulated. The relative QTL positions are placed in the middle or on the boundary of a marker interval (1, 2, and 4 cM away from the left marker of the three different spaced intervals, respectively). When investigating the effect of epistasis, the main and epistatic QTL effects are further set at ($a_1 = 1$, $a_2 = 1$, $I_{12} = 2$) or ($a_1 = 1$, $a_2 = 1$, $I_{12} = 3$), and the QTL are placed in the middle of 40-cM intervals. Together the QTL contribute 50% of the

TABLE 2

**Comparison of maximum likelihood and regression interval mapping of simulated data ($h^2 = 0.5$)**

| | | Spacing | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 10 cM | | | 20 cM | | | 40 cM | | |
| | | Mean | SD | MSE | Mean | SD | MSE | Mean | SD | MSE |
| $\mu = 0$ | ML | 0.004 | 0.056 | 0.003 | 0.005 | 0.057 | 0.003 | 0.004 | 0.063 | 0.004 |
| | REG | 0.004 | 0.057 | 0.003 | 0.005 | 0.060 | 0.004 | 0.004 | 0.064 | 0.004 |
| $a_1 = 1$ | ML | 1.008 | 0.117 | 0.014 | 1.001 | 0.121 | 0.015 | 0.993 | 0.142 | 0.020 |
| | REG | 1.006 | 0.121 | 0.015 | 1.000 | 0.125 | 0.016 | 0.993 | 0.156 | 0.024 |
| $a_2 = 1$ | ML | 0.996 | 0.114 | 0.013 | 0.997 | 0.126 | 0.016 | 0.998 | 0.138 | 0.019 |
| | REG | 0.997 | 0.115 | 0.013 | 0.997 | 0.132 | 0.017 | 0.998 | 0.154 | 0.024 |
| $I_{12} = 1$ | ML | 1.011 | 0.230 | 0.053 | 0.991 | 0.250 | 0.063 | 0.994 | 0.274 | 0.075 |
| | REG | 1.008 | 0.241 | 0.058 | 0.996 | 0.286 | 0.082 | 0.999 | 0.357 | 0.127 |
| $\sigma^2 = 0.563$ | ML | 0.551 | 0.071 | 0.005 | 0.549 | 0.076 | 0.006 | 0.552 | 0.085 | 0.007 |
| | REG | 0.614 | 0.074 | 0.008 | 0.670 | 0.086 | 0.019 | 0.782 | 0.099 | 0.058 |
| $h^2 = 0.5$ | ML | 0.511 | 0.044 | 0.002 | 0.510 | 0.049 | 0.003 | 0.509 | 0.061 | 0.004 |
| | REG | 0.452 | 0.043 | 0.004 | 0.399 | 0.048 | 0.013 | 0.302 | 0.051 | 0.042 |
| $\mathrm{LRT}_1$ | ML | 80.0 | 16.1 | | 69.4 | 15.0 | | 50.9 | 12.8 | |
| | REG | 75.4 | 15.5 | | 62.2 | 14.2 | | 42.1 | 11.7 | |
| $\mathrm{LRT}_2$ | ML | 79.5 | 14.9 | | 69.2 | 14.2 | | 51.3 | 12.6 | |
| | REG | 74.8 | 14.2 | | 62.0 | 13.2 | | 42.4 | 11.3 | |
| LRT | ML | 125.8 | 16.5 | | 110.5 | 15.8 | | 82.5 | 15.1 | |
| | REG | 120.5 | 16.1 | | 102.4 | 15.5 | | 72.6 | 14.4 | |

For each combination of simulated parameters, 500 replicates, each with sample size 200, were analyzed with QTL located in the middle of the marker interval. LRT is the likelihood-ratio test for $H_0$: $a_1 = 0$, $a_2 = 0$, and $I_{12} = 0$. $\mathrm{LRT}_1$ is the likelihood-ratio test for $H_0$: $a_1 = 0$, $I_{12} = 0$, and $a_2 \neq 0$. $\mathrm{LRT}_2$ is the likelihood-ratio test for $H_0$: $a_2 = 0$, $I_{12} = 0$, and $a_1 \neq 0$. $h^2$, the proportion of variance explained by QTL.

quantitative trait variation (the percentages of epistatic variance in the total genetic variance are 33.33 and 52.94%, respectively). When investigating the effect of the difference between QTL effects on the approximation, five unlinked QTL are placed in the middle of 10- or 40-cM intervals with effects ($a_1 = 1$, $a_2 = 1$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$), ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = 2$, $a_5 = 1$), or ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = -1$, $a_5 = 1$), respectively, and together contribute 50% of the trait variation. When investigating the effect of linkage, two QTL are placed in two neighboring 40-cM intervals and are 10, 20, 30, or 40 cM apart from each other (5, 10, 15, and 20 cM from the marker between them). Their effects are set at ($a_1 = 1$, $a_2 = -1$) without epistasis or ($a_1 = 1$, $a_2 = -1$, $I_{12} = 1$) with epistasis, and the heritability is assumed to be 0.1, 0.3, or 0.5 for each case. The sample size is 200, and 500 replicates were simulated for all cases.

For simplicity of comparison, the QTL positions are assumed to be known, and the simulation is performed at the positions. When calculating the power of separating closely linked QTL, a successful separation requires the partial LRT statistic for each QTL $> \chi^2_{1,0.05/10}$ ($\chi^2_{2,0.05/10}$ for the epistasis case). Means of the estimated parameter values, their standard deviations (SDs), and mean squared errors (MSEs), as well as the LRT statistics, are recorded. MSE is used to evaluate the approximation of REG to ML interval mapping and the performance of the two methods under various cases. MSE, which is defined as

$$E(\hat{\theta} - \theta)^2 = \mathrm{Var}(\hat{\theta}) + [E(\hat{\theta}) - \theta]^2$$
$$= \mathrm{Var}(\hat{\theta}) + (\mathrm{Bias}\ \hat{\theta})^2,$$

incorporates two components, one measuring the variability of the estimate (precision), and the other measuring its bias (accuracy). A good method needs to control both variance and bias in estimation.

As expected, if $h^2$ is low ($h^2 = 0.01$), the two methods provide almost identical means and SDs of the estimates for $\mu$, $a_1$, $a_2$, $I_{12}$, $\sigma^2$, and $h^2$, and LRT statistics. These results for $h^2 = 0.01$ correspond with the findings of Haley and Knott (1992) and Xu (1995, 1998a,b). When $h^2$ becomes higher ($h^2 > 0.1$), their differences due to the factors of interval size and QTL position become observable, but minor (the ML method generally has a smaller MSE and larger LRT statistic). To shorten the article, only part of the results are presented, and the investigation focuses on the factors of linkage, different QTL sizes, and epistasis under the multiple-QTL model.

**Proportion of variance explained by QTL, interval size, and QTL position:** The means of the estimated main and epistatic effects by the ML and REG methods are almost identical and very close to the true values for various $h^2$'s, interval sizes, and QTL positions. How-

TABLE 3

**Comparison of maximum likelihood and regression interval mapping of simulated data under different strengths of epistasis**

| | | Effect ratio | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 1:1:1 | | | 1:1:2 | | | 1:1:3 | | |
| | | Mean | SD | MSE | Mean | SD | MSE | Mean | SD | MSE |
| $\mu = 0$ | ML | 0.004 | 0.063 | 0.004 | 0.005 | 0.072 | 0.005 | 0.006 | 0.087 | 0.008 |
| | REG | 0.004 | 0.064 | 0.004 | 0.004 | 0.075 | 0.006 | 0.005 | 0.091 | 0.008 |
| $a_1 = 1$ | ML | 0.993 | 0.142 | 0.020 | 0.993 | 0.160 | 0.026 | 0.994 | 0.187 | 0.035 |
| | REG | 0.993 | 0.156 | 0.024 | 0.992 | 0.182 | 0.033 | 0.990 | 0.219 | 0.048 |
| $a_2 = 1$ | ML | 0.998 | 0.138 | 0.019 | 1.001 | 0.158 | 0.025 | 1.002 | 0.186 | 0.035 |
| | REG | 0.998 | 0.154 | 0.024 | 1.000 | 0.179 | 0.032 | 1.002 | 0.215 | 0.046 |
| $I_{12}$ | ML | 0.994 | 0.274 | 0.075 | 1.993 | 0.331 | 0.110 | 2.984 | 0.398 | 0.159 |
| | REG | 0.999 | 0.357 | 0.127 | 2.000 | 0.418 | 0.175 | 3.000 | 0.504 | 0.254 |
| $\sigma^2$ | ML | 0.552 | 0.085 | 0.007 | 0.735 | 0.110 | 0.012 | 1.042 | 0.156 | 0.025 |
| | REG | 0.782 | 0.099 | 0.058 | 1.081 | 0.139 | 0.129 | 1.577 | 0.199 | 0.304 |
| $h^2 = 0.5$ | ML | 0.509 | 0.061 | 0.004 | 0.510 | 0.056 | 0.003 | 0.510 | 0.058 | 0.003 |
| | REG | 0.302 | 0.051 | 0.042 | 0.277 | 0.053 | 0.053 | 0.255 | 0.053 | 0.063 |
| LRT | ML | 82.5 | 15.1 | | 80.3 | 16.3 | | 75.2 | 16.0 | |
| | REG | 72.6 | 14.4 | | 65.5 | 14.8 | | 59.5 | 14.4 | |

For each combination of simulated parameters, 500 replicates, each with sample size 200, were analyzed with QTL located in the middle of a 40-cM marker interval. $\sigma^2 = 0.563$ for effect 1:1:1; $\sigma^2 = 0.75$ for effect 1:1:2; $\sigma^2 = 1.063$ for effect 1:1:3. $I_{12} = 1$, 2, and 3 for the three effect ratios, respectively. $h^2$, the proportion of variance explained by QTL.

ever, the ML method tends to provide estimates with smaller SD (MSE) and larger LRT statistics when compared to the REG method. For example, the MSEs of $\hat{a}_1$ by the REG method are 0.178, 0.050, and 0.024 for $h^2 = 0.1$, 0.3, and 0.5, respectively, and the MSEs by the ML method are 0.175, 0.047, and 0.020, respectively (only the result for $h^2 = 0.5$ and QTL located in the middle of the intervals is shown in Table 2). There is a similar pattern for other estimates. The estimates of $\sigma^2$ and $h^2$ by the REG method are biased, and the estimates by the ML method are (asymptotically) unbiased. For example, the $\hat{h}^2$ by the REG method is 0.072 (SD 0.033), 0.187 (SD 0.047), and 0.302 (SD 0.051) for $h^2 = 0.1$, 0.3, or 0.5, respectively, and the $\hat{h}^2$ by the ML method is 0.128 (SD 0.055), 0.316 (SD 0.070), and 0.509 (SD 0.061), respectively. The bias of the REG method in estimating $\sigma^2$ and $h^2$ becomes obvious as $h^2$ becomes large. Also, the ML method gives larger LRT statistics than the REG method in all cases. The difference in mean LRT statistics between the two methods is negligible: 0.4 ($15.2 - 14.8 = 0.4$ for 40-cM marker spacing) for $h^2 = 0.1$ (results not shown), but 9.9 ($82.5 - 72.6 = 9.9$ for 40-cM marker spacing) for $h^2 = 0.5$. Therefore, the difference in the LRT statistic becomes larger as $h^2$ becomes higher. Similar patterns of difference in MSE and LRT statistics, caused by the change of $h^2$, can be observed for other interval sizes and QTL positions.

**Epistasis:** The means, SDs, and MSEs of the estimates as well as the mean LRT statistics for effect ratios 1:1:1, 1:1:2, and 1:1:3 are listed in Table 3. As interaction becomes stronger, the MSEs of the estimates by both methods become larger, and their differences in MSE and LRT statistics become larger. For example, the MSEs of $\hat{I}_{12}$ by the ML method are 0.075, 0.110, and 0.159 for the three ratios, respectively, and they are 0.127, 0.175, and 0.254 by the REG method, respectively. A similar trend can be observed for other estimates. The means of LRT statistics for the three ratios are 82.5, 80.3, and 75.2 for the ML method, respectively, and they are 72.6, 65.5, and 59.5 for the REG method, respectively. The bias of the REG method in the estimation of $\sigma^2$ and $h^2$ also becomes much more serious as interaction between QTL gets stronger. The $\hat{h}^2$ by the REG method is 0.302 (SD 0.051), 0.277 (SD 0.053), and 0.255 (SD 0.053) for the three ratios, respectively ($h^2 = 0.5$). The ML method, however, can estimate $h^2$ and other parameters well for all ratios.

**Difference between QTL effects:** Table 4 shows the means, SDs, and MSEs of the estimates as well as the mean LRT statistics for QTL effects ($a_1 = 1$, $a_2 = 1$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$), ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = 2$, $a_5 = 1$), and ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = -1$, $a_5 = 1$). When QTL effects are of the same size, the mean LRT statistic of the ML method is 1.8 ($81.5 - 79.7$) larger than that of the REG method. If there are some relatively large and small QTL, their differences in LRT statistics are 3.7 ($84.1 - 80.4$) and 5.8 ($84.7 - 78.9$), respectively, for the other two cases. Also, the estimates by the REG method tend to have larger MSEs, and the $\hat{h}^2$ and $\hat{\sigma}^2$ by the REG method are biased. For the case ($a_1 = 1$, $a_2 =$

**TABLE 4**

**Comparison of maximum likelihood and regression interval mapping of simulated data under different relative sizes of QTL effects**

| | | 40-cM interval, $h^2 = 0.5$ | | | | | | 10-cM interval, $h^2 = 0.1$ | |
|---|---|---|---|---|---|---|---|---|---|
| | | ($a_1 = a_4 = 1$) | | ($a_1 = 4$, $a_4 = 2$) | | ($a_1 = 4$, $a_4 = -1$) | | ($a_1 = a_4 = 1$) | |
| | | Mean | MSE | Mean | MSE | Mean | MSE | Mean | MSE |
| $\mu = 0$ | ML | −0.001 | 0.009 | −0.002 | 0.041 | 0.002 | 0.035 | −0.011 | 0.022 |
| | REG | 0.000 | 0.009 | −0.003 | 0.040 | −0.001 | 0.037 | −0.011 | 0.022 |
| $a_1$ | ML | 1.012 | 0.051 | 3.981 | 0.176 | 3.984 | 0.210 | 1.000 | 0.094 |
| | REG | 1.013 | 0.055 | 4.033 | 0.223 | 3.981 | 0.210 | 1.000 | 0.100 |
| $a_2 = 1$ | ML | 0.994 | 0.050 | 1.024 | 0.249 | 1.014 | 0.210 | 1.011 | 0.094 |
| | REG | 0.993 | 0.056 | 0.978 | 0.273 | 1.020 | 0.225 | 1.011 | 0.091 |
| $a_3 = 1$ | ML | 0.997 | 0.052 | 0.992 | 0.244 | 0.990 | 0.212 | 1.005 | 0.101 |
| | REG | 0.993 | 0.052 | 0.985 | 0.251 | 0.988 | 0.219 | 1.005 | 0.101 |
| $a_4$ | ML | 1.002 | 0.046 | 2.010 | 0.213 | −0.993 | 0.181 | 1.010 | 0.107 |
| | REG | 1.006 | 0.048 | 2.013 | 0.227 | −0.980 | 0.195 | 1.010 | 0.107 |
| $a_5 = 1$ | ML | 0.996 | 0.049 | 0.993 | 0.238 | 1.002 | 0.202 | 0.995 | 0.110 |
| | REG | 0.996 | 0.053 | 0.986 | 0.265 | 1.003 | 0.219 | 0.993 | 0.111 |
| $\sigma^2$ | ML | 1.165 | 0.214 | 5.396 | 0.901 | 4.731 | 0.576 | 11.15 | 1.01 |
| | REG | 1.661 | 0.217 | 7.640 | 4.361 | 6.690 | 3.467 | 11.2 | 0.919 |
| $h^2$ | ML | 0.530 | 0.069 | 0.527 | 0.005 | 0.524 | 0.004 | 0.112 | 0.001 |
| | REG | 0.327 | 0.032 | 0.329 | 0.032 | 0.324 | 0.032 | 0.099 | 0.001 |
| LRT | ML | 81.5 (14.0) | | 84.1 (14.2) | | 84.7 (14.4) | | 52.3 (13.6) | |
| | REG | 79.7 (14.1) | | 80.4 (14.4) | | 78.9 (14.4) | | 52.2 (13.5) | |

For each combination of simulated parameters, 500 replicates, each with sample size 200, were analyzed with QTL located in the middle of the marker interval. $h^2$, the proportion of variance explained by QTL. $\sigma^2 = 1.25$ for ($a_1 = 1$, $a_2 = 1$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$) and $h^2 = 0.5$. $\sigma^2 = 5.75$ for ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = 2$, $a_5 = 1$) and $h^2 = 0.5$. $\sigma^2 = 5.00$ for ($a_1 = 4$, $a_2 = 1$, $a_3 = 1$, $a_4 = -1$, $a_5 = 1$) and $h^2 = 0.5$. $\sigma^2 = 11.25$ for ($a_1 = 1$, $a_2 = 1$, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$) and $h^2 = 0.1$.

1, $a_3 = 1$, $a_4 = 1$, $a_5 = 1$, and $h^2 = 0.1$) and 10-cM intervals, the difference in mean LRT statistic between the two methods is at a very micro level ($52.3 - 52.2 = 0.08$).

**Linkage:** The powers of separating 10-, 20-, 30-, and 40-cM-apart QTL are 97.2, 99.0, 99.6, and 99.8% for the ML method, respectively, and are 22.0, 60.0, 91.6, and 98.4% for the REG method, respectively (Table 5 and Figure 1). Also, the difference in MSE between the two methods becomes larger as QTL get closer. The MSE ratios of $\hat{a}_1$ for the two methods are 4.52 (0.226/0.050), 8.67 (0.091/0.011), 4.00 (0.056/0.014), and 2.57 (0.036/0.011), respectively (Figure 1c). The estimated $h^2$ by the REG method is seriously biased. The means of $\hat{h}^2$ by the REG method are 0.037, 0.070, 0.112, and 0.162 for 10-, 20-, 30-, and 40-cM-apart QTL, respectively ($h^2 = 0.5$). The ML method, however, can estimate $h^2$ well (see also Figure 1c). If $h^2 = 0.3$ or 0.1, the power, the MSE ratio of $\hat{a}_1$, and $\hat{h}^2$ for the two methods are shown in Figure 1, a–c. If the linked QTL show epistasis, the advantage gained by the ML method becomes even more significant (Figure 1d).

## CONCLUSION AND DISCUSSION

In this article, the differences in QTL parameter estimation and testing for the existence of QTL between the ML and REG methods are investigated both analytically and numerically. It is found that the REG method tends to give estimates with larger MSE and smaller LRT statistics in testing parameters, and it is less powerful in QTL detection when compared with the ML method. Also, the REG method is biased in estimating the residual variance and the proportion of total variance explained by QTL. Therefore, ML interval mapping is more accurate, precise, and powerful than REG interval mapping in QTL mapping. The differences in power, MSE, and LRT statistics between the two methods depend on factors such as size of QTL effect, interval size, relative QTL position in an interval, difference between QTL effects, epistasis, and linkage between QTL, as shown in the article. Their differences in general may be minor, but can be significant in certain situations. The differences become larger as the proportion explained by QTL becomes higher, marker interval becomes wider, QTL position moves from boundary to middle of an interval, the difference of QTL effects is larger, epistasis becomes stronger, and the QTL positions are closer. Especially, the REG method may have a serious problem in detecting closely linked QTL when compared with the ML method. As shown in Table 5 and Figure 1, the difference in detecting closely linked (10–20 cM apart) QTL with opposite effects is quite

**TABLE 5**

**Comparison of maximum likelihood and regression interval mapping of simulated data under different strengths of linkage**

| | | Distance between QTL | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 10 cM | | 20 cM | | 30 cM | | 40 cM | |
| | | Mean | MSE | Mean | MSE | Mean | MSE | Mean | MSE |
| $\mu = 0$ | ML | 0.001 | 0.001 | 0.000 | 0.001 | 0.000 | 0.002 | 0.000 | 0.002 |
| | REG | 0.000 | 0.008 | −0.001 | 0.002 | 0.001 | 0.002 | 0.001 | 0.002 |
| $a_1 = 1$ | ML | 0.974 | 0.050 | 1.002 | 0.011 | 0.994 | 0.014 | 0.996 | 0.011 |
| | REG | 1.002 | 0.226 | 1.005 | 0.091 | 1.005 | 0.056 | 0.999 | 0.036 |
| $a_2 = -1$ | ML | −0.972 | 0.051 | −1.002 | 0.011 | −0.994 | 0.013 | −0.995 | 0.014 |
| | REG | −1.002 | 0.228 | −1.005 | 0.090 | −1.005 | 0.057 | −0.998 | 0.036 |
| $h^2 = 0.5$ | ML | 0.496 | 0.007 | 0.508 | 0.005 | 0.507 | 0.006 | 0.510 | 0.006 |
| | REG | 0.037 | 0.217 | 0.070 | 0.186 | 0.112 | 0.152 | 0.162 | 0.116 |
| LRT | ML | 42.0 (16.1) | | 32.7 (11.0) | | 36.6 (11.0) | | 46.0 (12.4) | |
| | REG | 6.7 (5.1) | | 13.5 (7.1) | | 21.6 (9.0) | | 30.6 (10.5) | |
| $\text{LRT}_1$ | ML | 40.9 (16.1) | | 31.2 (10.8) | | 33.7 (10.6) | | 39.9 (11.4) | |
| | REG | 7.7 (5.4) | | 14.7 (7.2) | | 24.0 (9.3) | | 35.6 (11.4) | |
| $\text{LRT}_2$ | ML | 40.9 (16.1) | | 31.3 (11.0) | | 33.8 (10.9) | | 40.0 (11.7) | |
| | REG | 6.7 (5.1) | | 13.4 (6.9) | | 21.5 (9.0) | | 30.5 (10.5) | |
| Power (%) | ML | 97.2 | | 99.0 | | 99.6 | | 99.8 | |
| | REG | 22.0 | | 60.0 | | 91.6 | | 98.4 | |

For each combination of simulated parameters, 500 replicates, each with sample size 200, were analyzed with two QTL contributing 50% of the total genetic variance and located in the middle of the marker interval. LRT is the likelihood ratio test for $H_0$: $a_1 = 0$ and $a_2 = 0$. $\text{LRT}_1$ is the likelihood ratio test for $H_0$: $a_1 = 0$ and $a_2 \neq 0$. $\text{LRT}_2$ is the likelihood ratio test for $H_0$: $a_2 = 0$ and $a_1 \neq 0$. Power, percentage of replicates with $\text{LRT}_1 > 7.88$ and $\text{LRT}_2 > 7.88$. Numbers in parentheses denote standard deviation. $h^2$, the proportion of variance explained by QTL.

significant (power 0.22 *vs.* 0.97 for 10-cM-apart QTL and $h^2 = 0.5$; power 0.60 *vs.* 0.99 for 20-cM-apart QTL and $h^2 = 0.5$; power 0.09 *vs.* 0.54 for 10-cM-apart QTL and $h^2 = 0.3$; power 0.34 *vs.* 0.61 for 20-cM-apart QTL and $h^2 = 0.3$). In addition, the REG method is seriously biased in estimating the proportion of variance explained by QTL (Figure 1b), and it gives the estimates of the effects with much larger MSEs (Figure 1c). The problem of the REG method in detecting closely linked QTL becomes worse if epistasis is present (Figure 1d).

It was often pointed out that there is no significant difference in the estimation of QTL parameter and statistical power of QTL detection between the REG and ML methods with the exception that the estimate of residual variance by the REG method is biased (Haley and Knott 1992; Xu 1995, 1998a,b). These findings were mostly done by simulation and concentrated on the comparison of mean estimate (accuracy) of QTL effect for low heritability ($h^2 = 0$, 0.008, 0.03, and 0.111 in Haley and Knott 1992) and a QTL positioned in a narrow interval (interval size 10 cM in Xu 1998a,b) for a one-QTL model. Therefore, their differences due to factors such as different QTL sizes, epistasis, and linkage between QTL had not been identified and needed to be checked using the multiple-QTL model. When multiple QTL are considered simultaneously in the model, the differences in power and estimation between the two methods can be significant for these factors as shown in this article. Using the multiple-QTL model of the ML approach, it has been found that quantitative traits with a somewhat medium to high heritability might be affected by several linked and unlinked QTL, having different sizes, directions, and interaction of effects (Kao *et al.* 1999; Weber *et al.* 1999; Zeng *et al.* 2000). Also, the linkage map may have wide marker intervals (Grattapaglia *et al.* 1996; Satagopan *et al.* 1996; Li *et al.* 1997; Kao *et al.* 1999). As a result, the REG method can be significantly different from the ML method and thus be problematic in practical QTL mapping.

The cost in computation per iteration in the EM algorithm is generally not very expensive (McLachlan and Krishnan 1997). If the model is extended to fit five QTL, the ML method needs ~18 iterations to converge and is <10 times slower than the REG method for the 40-cM interval case (the REG method takes ~65 sec and the ML method takes ~608 sec to finish the computation of 500 replicates). Therefore, the ML method should not be regarded as formidably expensive in computation as the computer technology is advancing. Xu (1998a,b) proposed the iteratively reweighted least-squares (IRWLS) method to correct the bias of the REG method in estimating residual variance. The estimates by both REG and IRWLS methods tend to have larger SD than those by the ML method (Tables 1–4 in Xu 1998a; Table 7 in Xu 1998b). As the MLEs have the

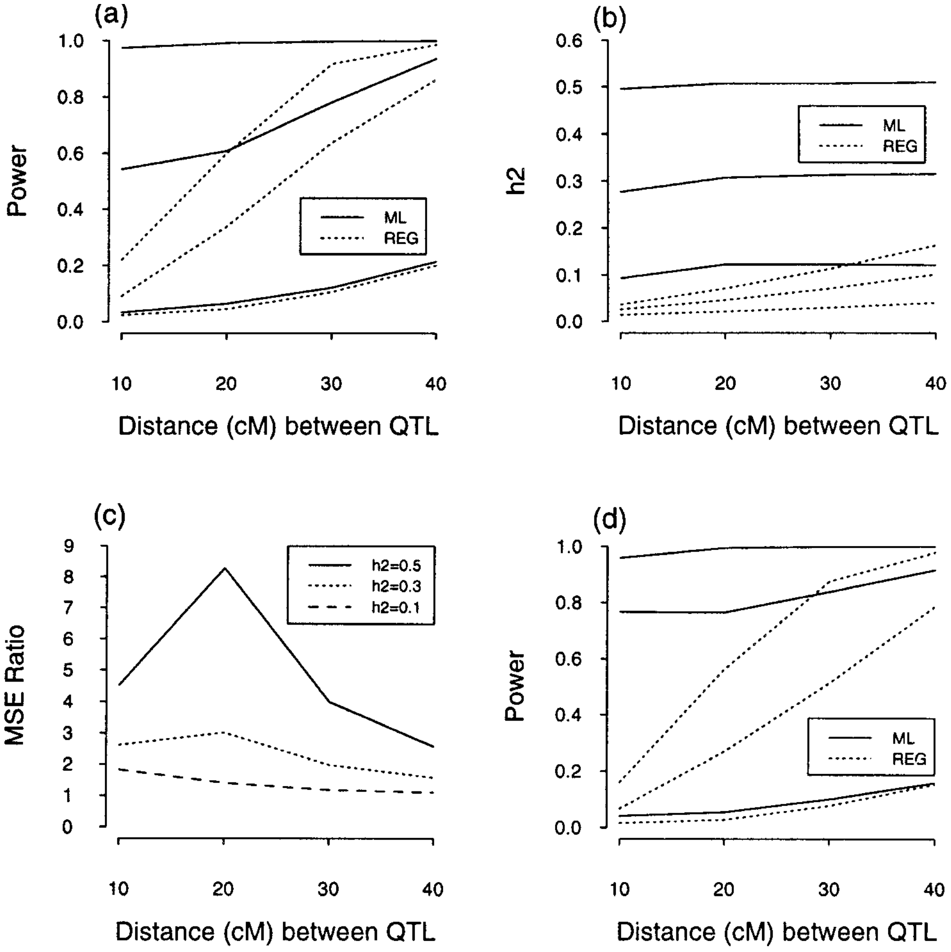

FIGURE 1.—The power, estimate of proportion of variance explained by QTL ($h^2$), and MSE of effect estimate by the ML and REG methods in the analysis of two linked QTL without ($a_1 = 1$, $a_2 = -1$) and with ($a_1 = 1$, $a_2 = -1$, $I_{12} = 1$) epistasis under different genetic distances and proportions of variance explained by QTL ($h^2$). (a) Power of separating two linked QTL with no epistasis. The solid lines from bottom to top denote the power by the ML method for $h^2 = 0.1$, 0.3, and 0.5, respectively. The dotted lines from bottom to top denote the power by the REG method for $h^2 = 0.1$, 0.3, and 0.5, respectively. (b) Estimate of $h^2$. The solid lines from bottom to top denote the estimate of $h^2$ by the ML method for $h^2 = 0.1$, 0.3, and 0.5, respectively. The dotted lines from bottom to top denote the estimate of $h^2$ by the REG method for $h^2 = 0.1$, 0.3, and 0.5, respectively. (c) Ratio of MSE. The solid, dotted, and dashed lines from bottom to top denote the MSE ratio of $\hat{a}_1$ by the REG method and ML method for $h^2 = 0.1$, 0.3, and 0.5, respectively. (d) Power of separating two linked QTL with epistasis. The solid lines from bottom to top denote the power by the ML method for $h^2 = 0.1$, 0.3, and 0.5, respectively. The dotted lines from bottom to top denote the power by the REG method for $h^2 = 0.1$, 0.3, and 0.5, respectively.

property of asymptotical efficiency, it should not be surprising that the ML method has the ability to provide the smallest SD among estimates (CASELLA and BERGER 1990). Therefore, ML interval mapping is not only a more powerful but also a more precise method in QTL mapping.

The distributions of most quantitative traits approximate more or less close to normal or can be scaled to normal through simple transformation (FALCONER and MACKAY 1996). Therefore, when mapping QTL, the likelihood is generally modeled as a normal mixture (Equation 2). When applying the EM algorithm to the estimation of a normal mixture model, the estimating Equations 3, 4, and 5 depend on the conditional posterior probabilities of QTL genotypes $\pi_{ij}$'s, which take the distribution of the residual error into account using normal density. The estimation of the REG method depends on the conditional probabilities $p_{ij}$'s, which ignore the distribution of residual error, and the IRWLS method takes only the second moment of residual error into account whatever the underlying residual error distribution is. This is also the reason why the ML method can be better than the REG and IRWLS methods. If the residual error does not follow normal distribution, the mixture model in Equation 2 should take its specific form into account to model the relation between the quantitative trait and QTL in estimation. In practice, although most of the residual errors are normally distributed and the use of the normal mixture model should be safe in most situations, it is important to examine the pattern of residuals, which is a requisite procedure in model selection, to ensure that the final QTL mapping model is appropriate.

The QTL mapping result will be used as a base for follow-up operations, such as marker-assisted selection or gene transfer, on QTL for trait improvement. To ensure the validity of trait improvement, the quality of QTL mapping should be more important than the ease of computation. Researchers using the REG method for mapping QTL need to be concerned with the factors affecting its approximation to the ML method in practice. For example, if there are wide marker intervals along the genome (known data structure), or the QTL effects are not sure to be equally small, or the QTL are linked with epistasis (unknown QTL parameters), the REG method may perform poorly when compared to the ML method. Then, after the analysis of the REG method, there is a need to further use the ML method to finalize the QTL mapping result. As far as computation is concerned, it is suggested that researchers may use the REG method as an initial procedure to obtain preliminary results and further use the ML method as a final procedure to obtain the conclusive results of QTL mapping.

This study was supported by grants NCS89-2313-B-001-006 from the National Science Council, Taiwan, Republic of China.

## LITERATURE CITED

Carbonell, E. A., T. M. Gerig, E. Balansard and M. J. Asins, 1992 Interval mapping in the analysis of nonadditive quantitative trait loci. Biometrics **48:** 305–315.

Casella, G., and R. Berger, 1990 *Statistical Inference*. Wadsworth, Belmont, CA.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. J. R. Stat. Soc. **39:** 1–38.

Doerge, R. W., and G. A. Churchill, 1996 Permutation test for multiple loci affecting a quantitative character. Genetics **142:** 284–294.

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

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*. Longman Group, London.

Goffinet, B., and B. Mangin, 1998 Comparing methods to detect more than one QTL on a chromosome. Theor. Appl. Genet. **96:** 628–633.

Grattapaglia, D., F. L. G. Bertolucci, R. Penchel and R. R. Sederoff, 1996 Genetic mapping of quantitative trait loci controlling growth and wood quality traits in *Eucalyptus grandis* using a maternal half-sib family and RAPD markers. Genetics **144:** 1205–1214.

Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.

Haldane, J. B. S., 1919 The combination of linkage values and the calculation of distances between the loci of linked factors. J. Genet. **8:** 299–309.

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.

Haley, C. S., S. A. Knott and J.-M. Elsen, 1994 Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics **136:** 1195–1207.

Henshall, J. M., and M. E. Goddard, 1999 Multiple-trait mapping of quantitative trait loci after selective genotyping using logistic regression. Genetics **151:** 885–894.

Hoeschele, I., and P. VanRaden, 1993a Bayesian analysis of linkage between genetic markers and quantitative trait loci. I. Prior knowledge. Theor. Appl. Genet. **85:** 953–960.

Hoeschele, I., and P. VanRaden, 1993b Bayesian analysis of linkage between genetic markers and quantitative trait loci. II. Combining prior knowledge with experimental evidence. Theor. Appl. Genet. **85:** 946–952.

Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.

Jiang, C., and Z.-B. Zeng, 1997 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics **140:** 1111–1127.

Kao, C.-H., and Z.-B. Zeng, 1997 General formulas for obtaining the MLE and the asymptotic variance-covariance matrix in mapping quantitative trait loci when using the EM algorithm. Biometrics **53:** 359–371.

Kao, C.-H., Z.-B. Zeng and R. D. Teasdale, 1999 Multiple interval mapping for quantitative trait loci. Genetics **152:** 1203–1216.

Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.

Lebreton, C. M., P. M. Visscher, C. S. Haley, A. Semikhodskii and S. A. Quarrie, 1998 A nonparametric bootstrap method for testing close linkage *vs.* pleiotropy of coincident quantitative trait loci. Genetics **150:** 931–943.

Li, Z., S. R. M. Pinson, W. D. Park, A. H. Paterson and J. W. Stansel, 1997 Epistasis for three grain yield components in rice (*Oryza sativa L.*). Genetics **145:** 453–465.

Little, R. J. A., and D. B. Rubin, 1987 *Statistical Analysis With Missing Data*. John Wiley, New York.

Louis, T. A., 1982 Finding the observed information matrix when using the EM algorithm. J. R. Stat. Soc. Ser. B **44:** 226–233.

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.

McLachlan, G. F., and T. Krishnan, 1997 *The EM Algorithm and Extensions*. John Wiley, New York.

Meng, X.-L., and B. Rubin, 1991 Using EM to obtain asymptotic variance-covariance matrix: the SEM algorithm. J. Am. Stat. Assoc. **86:** 899–909.

Neter, J., W. Wasserman and M. H. Kutner, 1990 *Applied Linear Statistical Model*. Richard D. Irwin, Tokyo.

Rebai, A., and B. Goffinet, 2000 More about quantitative trait locus mapping with diallel designs. Genet. Res. **75:** 243–247.

Satagopan, J. M., B. S. Yandell, M. A. Newton and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144:** 805–816.

Sillanpaa, M. J., and E. Arjas, 1999 Bayesian mapping of multiple quantitative trait loci from incomplete outbred offspring data. Genetics **151:** 1605–1619.

Song, J. Z., M. Soller and A. Genizi, 1999 The full-sib intercross line (FSIL): a QTL mapping design for outcross species. Genet. Res. **77:** 61–73.

Weber, K., R. Eisman, L. Morey, A. Patty, J. Sparks *et al.*, 1999 An analysis of polygenes affecting wing shape on Chromosome 3 in Drosophila melanogaster. Genetics **153:** 773–786.

Whittaker, J. C., R. Thompson and P. M. Visscher, 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.

Xu, S., 1996 Mapping quantitative trait loci using four-way crosses. Genet. Res. **68:** 175–181.

Xu, S., 1998a Further investigation on the regression method of mapping quantitative trait loci. Heredity **80:** 364–373.

Xu, S., 1998b Iteratively reweighted least squares mapping for quantitative trait loci. Behav. Genet. **28:** 341–355.

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

Zeng, Z.-B., J. Liu, L. F. Stam, C.-H. Kao, J. M. Mercer *et al.*, 2000 Genetic architecture of a morphological shape difference between two Drosophila species. Genetics **154:** 299–310.

Communicating editor: Z-B. Zeng