Multiple Interval Mapping for Quantitative Trait Loci

Chen-Hung Kao,* Zhao-Bang Zeng† and Robert D. Teasdale‡

*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China, †Program in Statistical Genetics, Department of Statistics, North Carolina State University, Raleigh, North Carolina 27695-8203 and ‡ForBio Research Pty Ltd., Toowong, Queensland 4066, Australia

Manuscript received December 5, 1997
Accepted for publication March 24, 1999

ABSTRACT

A new statistical method for mapping quantitative trait loci (QTL), called multiple interval mapping (MIM), is presented. It uses multiple marker intervals simultaneously to fit multiple putative QTL directly in the model for mapping QTL. The MIM model is based on Cockerham's model for interpreting genetic parameters and the method of maximum likelihood for estimating genetic parameters. With the MIM approach, the precision and power of QTL mapping could be improved. Also, epistasis between QTL, genotypic values of individuals, and heritabilities of quantitative traits can be readily estimated and analyzed. Using the MIM model, a stepwise selection procedure with likelihood ratio test statistic as a criterion is proposed to identify QTL. This MIM method was applied to a mapping data set of radiata pine on three traits: brown cone number, tree diameter, and branch quality scores. Based on the MIM result, seven, six, and five QTL were detected for the three traits, respectively. The detected QTL individually contributed from ~1 to 27% of the total genetic variation. Significant epistasis between four pairs of QTL in two traits was detected, and the four pairs of QTL contributed ~10.38 and 14.14% of the total genetic variation. The asymptotic variances of QTL positions and effects were also provided to construct the confidence intervals. The estimated heritabilities were 0.5606, 0.5226, and 0.3630 for the three traits, respectively. With the estimated QTL effects and positions, the best strategy of marker-assisted selection for trait improvement for a specific purpose and requirement can be explored. The MIM FORTRAN program is available on the worldwide web (http://www.stat.sinica.edu.tw/~chkao/).

THE basic principle of using genetic markers to study quantitative trait loci (QTL) is well established (Sax 1923; Thoday 1960; Jayakar 1970; Lander and Botstein 1989; Carbonell *et al.* 1992; Haley and Knott 1992; Jansen 1993; Zeng 1993, 1994). Sax (1923) first used pattern and pigment markers in beans to analyze genes affecting seed size by investigating the segregation ratio of F_2 progeny of different crosses. Thoday (1960) proposed the idea of using two markers to bracket a region for detecting QTL. The basic idea of Sax and Thoday for detecting the association of a QTL with a marker rests on the comparisons of trait means of different marker (chromosomal segment) classes. These methods, such as *t*-test and simple and multiple regressions, directly analyze markers.

In recent years, the advent of fine-scale molecular genetic marker maps for various organisms by molecular biology techniques has greatly facilitated the systematic mapping and analysis of individual QTL. Lander and Botstein (1989) proposed a much-improved method, named interval mapping (IM), for mapping QTL. They used one marker interval at a time to construct a putative QTL for testing by performing a likelihood ratio test (LRT) at every position in the interval. With a fine-scale genetic marker map throughout the genome, IM can be performed at any position covered by markers to produce a continuous LRT statistical profile along chromosomes. The position with the significantly largest LRT statistic in a chromosome region is an estimate of QTL position. It has been shown that IM has more power and requires fewer progeny than the methods for direct analysis of markers (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1994). Haley and Knott (1992) proposed a regression version of interval mapping to approximate IM. Although Haley and Knott's method could save time in computation and produce similar results to those obtained by IM, the estimate of the residual variance is biased, and the power of QTL detection can be affected (Xu 1995).

The approach of IM considers one QTL at a time in the model for QTL mapping. Therefore, IM can bias identification and estimation of QTL when multiple QTL are located in the same linkage group (Lander and Botstein 1989; Haley and Knott 1992; Zeng 1994). To deal with multiple QTL problems, Jansen (1993) and Zeng (1993, 1994) independently proposed the idea of combining IM with multiple regression analysis in mapping. Zeng named this combination "composite interval mapping" (CIM). The approach of CIM is that, when testing for the putative QTL in an interval, one uses other markers as covariates to control for other QTL and to reduce the residual variance such

Corresponding author: Chen-Hung Kao, Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China. E-mail: chkao@stat.sinica.edu.tw

that the test can be improved. The model of CIM includes one QTL and markers. Hoeschele and Vanranden (1993a,b), Satagopan *et al.* (1996), and Sillanpaa and Arjas (1998) used a Bayesian approach in estimation and to identify QTL. Doerge and Churchill (1996) used permutation tests for QTL detection. Mapping for QTL controlling binary trait and ordinal categorical traits is presented by Hackett and Weller (1995) and Xu and Atchley (1996). In human and animal genetics, the mixed model, including random effect, has been applied to QTL mapping (Xu and Atchley 1995; Grignola *et al.* 1996a,b).

Ideally, we would extend the current QTL mapping models to a multiple QTL model for mapping multiple QTL in a way that QTL can be directly controlled in the model to further improve QTL mapping. In this article, a new QTL mapping method named multiple interval mapping (MIM) was developed. MIM uses multiple marker intervals simultaneously to construct multiple putative QTL in the model for QTL mapping. Therefore, when compared with the current methods such as IM and CIM, MIM tends to be more powerful and precise in detecting QTL as shown by the example in this article. In addition, MIM can readily search for and analyze epistatic QTL and estimate the individual genotypic value and the heritabilities of quantitative traits. On the basis of the MIM result, genetic variance components contributed by individual QTL were also estimated, and marker-assisted selection can be performed.

GENETIC MODEL

Consider m QTL, Q_1, Q_2, · · ·, and Q_m, in a backcross population in which there are two genotypes, Q_jQ_j and Q_jq_j, each with one-half frequency for a QTL, say Q_j. For m QTL, there are 2^m possible different QTL genotypes in the population. Cockerham's genetic model (C-H. Kao and Z-B. Zeng, unpublished results) is used to define the genetic parameters and model the relation between the genotypic value and the genetic parameters. If only up to digenic epistasis is considered, the relation between the genotypic value of individual i, G_i, and the genetic parameters can be expressed in the equation

$$G_i = \mu + \sum_{j=1}^{m} a_j x_{ij} + \sum_{j<k}^{m} w_{jk} (x_{ij}x_{ik}), \quad i = 1, \cdots, 2^m, \qquad (1)$$

where x_{ij} is coded as $\frac{1}{2}$ or $-\frac{1}{2}$ if the genotype of Q_j is Q_jQ_j or Q_jq_j, respectively, a_j is the corresponding main effect of Q_j, and w_{jk} is the epistatic effect between Q_j and Q_k. The main advantage of Cockerham's model is that it possesses an orthogonal property in modeling genetic parameters.

To assist with explaining the estimation of the genetic effects in the MIM model (Equation 3), the genetic model in Equation 1 is expressed in matrix notation as Equation 2 (Scheme 1). In Equation 2, the column vector **G** contains the genotypic values of the 2^m possible genotypes. The subscripts of G (1 or 0) denote the homozygote or heterozygote of the QTL in the order of the first, second, third, · · ·, and mth QTL, respectively. The first m columns in the genetic design matrix **D** are the coefficients associated with the main effects of the m QTL, and the last $m(m-1)/2$ columns represent the coefficients of the epistatic effects among them. Vector E contains the QTL main and epistatic effects. If there is no epistasis between some QTL, some of the columns for epistasis should be dropped out from matrix **D**. If higher-order epistasis is considered, the dimension of matrix **D** is easy to expand accordingly. The matrix **D** plays an important role in estimation of genetic parameters in the MIM model.

STATISTICAL MODEL OF MIM

Multiple interval mapping: Assume m QTL, Q_1, Q_2, · · ·, and Q_m, located at positions p_1, p_2, · · ·, p_m in m different marker intervals, I_1, I_2, · · ·, I_m, along the genome, control a quantitative trait y. Among the m QTL, some may show epistasis and some may not. The quantitative trait value for an individual, i, can be related

$$\begin{bmatrix} G_{\overbrace{11\cdots11}^{m}} \\ G_{11\cdots10} \\ G_{11\cdots01} \\ G_{11\cdots00} \\ \vdots \\ G_{00\cdots11} \\ G_{00\cdots10} \\ G_{00\cdots01} \\ G_{00\cdots00} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ \vdots \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} \overbrace{\frac{1}{2}\ \frac{1}{2}\ \cdots\ \frac{1}{2}\ \frac{1}{2}}^{m} & \overbrace{\frac{1}{4}\ \cdots\ \cdots\ \cdots\ \frac{1}{4}}^{m(m-1)/2} \\ \frac{1}{2}\ \frac{1}{2}\ \cdots\ \frac{1}{2}\ -\frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ \frac{1}{4} \\ \frac{1}{2}\ \frac{1}{2}\ \cdots\ -\frac{1}{2}\ \frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ -\frac{1}{4} \\ \frac{1}{2}\ \frac{1}{2}\ \cdots\ -\frac{1}{2}\ -\frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ \frac{1}{4} \\ \vdots\ \vdots\ \cdots\ \vdots\ \vdots & \vdots\ \cdots\ \cdots\ \cdots\ \vdots \\ -\frac{1}{2}\ -\frac{1}{2}\ \cdots\ \frac{1}{2}\ \frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ \frac{1}{4} \\ -\frac{1}{2}\ -\frac{1}{2}\ \cdots\ \frac{1}{2}\ -\frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ -\frac{1}{4} \\ -\frac{1}{2}\ -\frac{1}{2}\ \cdots\ -\frac{1}{2}\ \frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ -\frac{1}{4} \\ -\frac{1}{2}\ -\frac{1}{2}\ \cdots\ -\frac{1}{2}\ -\frac{1}{2} & \frac{1}{4}\ \cdots\ \cdots\ \cdots\ \frac{1}{4} \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ \vdots \\ a_m \\ w_{12} \\ w_{13} \\ \vdots \\ \vdots \\ \vdots \\ w_{(m-1)m} \end{bmatrix} = 1_{2^m\times1}\mu + \mathbf{D}_{2^m\times m(m+1)/2}\mathbf{E}_{m(m+1)/2\times1}. \qquad (2)$$

Scheme 1

to the m putative QTL by the model

$$y_i = \mu + \sum_{j=1}^{m} a_j x_{ij}^* + \sum_{j \neq k}^{m} \delta_{jk}(w_{jk} x_{ij}^* x_{ik}^*) + \varepsilon_i, \qquad (3)$$

where μ is the mean, x_{ij}^* is the coded variable for the genotype of Q_j, a_j and w_{jk} have the same definitions as those in the genetic model in Equation 1, δ_{jk} is an indicator variable for epistasis between Q_j and Q_k, and ε_i is assumed to follow $N(0, \sigma^2)$. Indicator variable δ_{jk} takes value one if Q_j and Q_k interact; otherwise its value is zero. In this model, the first summation is for the main effects of the m QTL, the second summation is for their possible epistasis, and ε_i is the environmental deviation. This is termed the MIM model because multiple (m) marker intervals are simultaneously used to construct multiple (m) putative QTL in the model for QTL mapping. If QTL genotypes are known, the model tells that the quantitative trait value is the sum of the QTL main effects, their possible epistatic effects, and environmental deviation, and the MIM model is a regression model. However, the putative QTL genotypes denoted by x_{ij}^*'s are usually not observed because QTL could be located in the intervals. Given observed flanking marker genotypes, the conditional distributions of QTL genotypes, x_{ij}^*'s, for QTL at specific positions, p_j's, can be inferred based on Haldane's mapping function (Haldane 1919) assuming no crossover interference (Table 1 in Kao and Zeng 1997), and the MIM model is then a normal mixture model. For each Q_j, its conditional probabilities are extracted to form a matrix $\mathbf{Q}_j$ (note that Q denotes QTL and $\mathbf{Q}$ denotes the conditional probability matrix; see Kao and Zeng 1997). The conditional probability matrices, $\mathbf{Q}_j$'s, $j = 1, 2, \cdots, m$, play an important role in estimation of the QTL positions in the intervals.

The MIM model is a multiple QTL model and its likelihood is a finite normal mixture. There are two problems that need to be solved for the MIM model. The first is that of parameter estimation of the finite normal mixture model. As m becomes large, the derivation of the maximum-likelihood estimates (MLEs) of the QTL effects and positions in estimation quickly becomes unwieldy. To handle the estimation problem, the general formulas derived by Kao and Zeng (1997) are used to obtain the MLEs in parameter estimation. The second problem is how to find QTL to fit into the MIM model. To select QTL for the MIM model, a stepwise model selection procedure is proposed in strategy of qtl mapping.

LIKELIHOOD OF THE MIM MODEL

In the MIM model, the genotype of each putative QTL, Q_j in interval I_j, is not observed, but its distribution can be inferred from the flanking markers of I_j based on the recombination frequency between them. For every QTL in the backcross population, the conditional probabilities of the QTL genotypes, given different flanking marker genotypes, can be found in Table 1 of Kao and Zeng (1997). To infer the joint conditional probability of the genotype of the m putative QTL, we use the property that if there is no crossing-over interference, the conditional distributions of the individual putative QTL genotypes, given the flanking marker genotypes, are independent. That is,

$$\text{prob}(Q_1, Q_2, \cdots, Q_m \mid I_1, I_2, \cdots, I_m) = \prod_{i=1}^{m} \text{prob}(Q_i \mid I_i).$$

The joint conditional probability of the m QTL is the product of the marginal conditional probabilities of individual QTL. We refer to p_{ij}, $j = 1, 2, \cdots, 2^m$, as the conditional probabilities of 2^m possible QTL genotypes (note that p_j's denote QTL positions and p_{ij}'s denote the conditional probabilities). If multiple putative QTL within a single marker interval are considered, the individual and joint conditional probabilities of QTL genotypes can be also inferred directly or by a Markov chain procedure (Jiang and Zeng 1997) assuming no interference.

Given a sample with size n, the likelihood function of the MIM model for $\theta = (p_1, p_2, \cdots, p_m, a_1, \cdots, a_m, \cdots, w_{jk}, \cdots, \sigma^2)$ is

$$L(\theta \mid Y, X) = \prod_{i=1}^{n} \left[\sum_{j=1}^{2^m} p_{ij} \phi\left(\frac{y_i - \mu_{ij}}{\sigma}\right) \right], \qquad (4)$$

where $\phi(\cdot)$ is a standard normal probability density function, μ_{ij}'s correspond to the genotypic values of the 2^m different QTL genotypes in Equation 1, and p_{ij}'s containing information on QTL positions are the corresponding joint conditional probabilities. Statistically, this is a normal mixture model. The density of each individual is a mixture of 2^m possible normal densities with different means μ_{ij}'s and mixing proportions p_{ij}'s. To obtain the MLEs and the asymptotic variance-covariance matrix of the model, the general formulas of Kao and Zeng (1997), based on the expectation and maximization (EM) algorithm (Dempster *et al.* 1977), are used for parameter estimation.

PARAMETER ESTIMATION

The likelihood of the MIM model is a finite normal mixture. In parameter estimation, the finite normal mixture model can be treated as an *incomplete-data* problem (Little and Rubin 1987) by regarding the trait and markers as *observed data* and the QTL as *missing data*. The EM algorithm can be used for obtaining the MLEs of the genetic parameters, and Louis's (1982) method can be implemented to obtain the variance-covariance matrix.

In the MIM model, when only one putative QTL ($m = 1$) is considered in a backcross population, the likelihood is a mixture of two normals (like IM and CIM), and four parameters need to be estimated. The derivation of

the MLEs for the one putative QTL model using the EM algorithm has been provided (Carbonell *et al.* 1992; Zeng 1994). When arbitrary m putative QTL are considered, the likelihood is a mixture of 2^m normals, and at least $2m + 2$ parameters (including mean μ, QTL positions and effects, environment variance, and epistasis) need to be estimated. The number of mixture components and parameters increases dramatically as the number of putative QTL taken into account in the model increases. Taking $m = 10$ as an example, the likelihood is a mixture of 1024 normals with more than 22 parameters to estimate. Therefore, one of the main difficulties with the MIM model is that the derivation of the MLEs quickly becomes unwieldy if the number of putative QTL is large, and an efficient and systematic method for parameter estimation of the MIM model is needed to avoid rederivation for each m. Here, we use the general formulas provided by Kao and Zeng (1997) for deriving the MLEs and the asymptotic variance-covariance matrix of the parameters as the estimation method of MIM. The general formulas are based on two matrices, **D** and **Q**. The matrix **D** is the genetic design matrix that characterizes the genetic parameters of the QTL effects, and the matrix **Q** is the conditional probability matrix that contains the information on QTL positions. Given the two matrices, the MLEs and the asymptotic variance-covariance matrix can be systematically obtained.

To apply the general formulas to MIM, the genetic design matrix **D** of the MIM model has the same first m columns as those in Equation 2 for indicating the m main QTL effects and has some or none of the last $m(m-1)/2$ columns for specifying epistasis. We refer to **D** as a $2^m \times k$ matrix, where k is the column dimension. There are m individual conditional probability matrices, $\mathbf{Q}_1, \mathbf{Q}_2, \cdots$, and $\mathbf{Q}_m$ for the m QTL. The components of the conditional probability matrix $\mathbf{Q}_j$ of QTL Q_j in the interval I_j with flanking markers M_j and N_j can be found in Table 1 of Kao and Zeng (1997). For each interval, there are four possible flanking marker genotypes. Totally, there are 4^m possible flanking marker genotypes for m intervals. The joint conditional probability matrix **Q** then has dimension $4^m \times 2^m$ and can be obtained by $\mathbf{Q} = \mathbf{Q}_1 \otimes \mathbf{Q}_2 \otimes \cdots \otimes \mathbf{Q}_m$, where $\otimes$ denotes the Kronecker product. The 2^m mixing proportions of any individual i, p_{ij}'s, can be found to be one of the 4^m rows in **Q** according to its flanking marker genotype. Given the matrices **D** and **Q**, the MLEs and the asymptotic variance-covariance matrix can be readily obtained by the general formulas.

Note that, at the tested positions $p_1, p_2, \cdots$, and p_m, the mixing proportions p_{ij}'s in the likelihood are fixed and need not be estimated. For obtaining the MLEs of mean, environmental variance, and marginal and epistatic effects, the general equations formulate the iteration of the $(t + 1)$ EM step as follows:

E step: Update the posterior probabilities of the 2^m possible QTL genotypes for each individual i,

$$\pi_{ij}^{(t+1)} = \frac{p_{ij}\phi((y_i - \mu_{ij}^{(t)})/\sigma^{(t)})}{\sum_{j=1}^{2^m} p_{ij}\phi((y_i - \mu_{ij}^{(t)})/\sigma^{(t)})}; \quad i = 1, 2, \cdots, n, \quad j = 1, 2, \cdots, 2^m. \tag{5}$$

M step: Find $\theta^{(t+1)}$, which satisfies the solutions

$$\mathrm{E}^{(t+1)} = \mathrm{r}^{(t)} - \mathrm{M}^{(t)}\mathrm{E}^{(t)} \tag{6}$$

$$\mu^{(t+1)} = \frac{1}{n}1'[\mathrm{Y} - \boldsymbol{\Pi}^{(t)}\mathrm{D}\mathrm{E}^{(t+1)}] \tag{7}$$

$$\sigma^{2(t+1)} = \frac{1}{n}[(\mathrm{Y} - 1\mu^{(t+1)})'(\mathrm{Y} - 1\mu^{(t+1)}) - 2(\mathrm{Y} - 1\mu^{(t+1)})'\boldsymbol{\Pi}^{(t)}\mathrm{D}\mathrm{E}^{(t+1)} + \mathrm{E}'^{(t+1)}\mathrm{V}^{(t)}\mathrm{E}^{(t+1)}], \tag{8}$$

where $\boldsymbol{\Pi} = \{\pi_{ij}\}_{n\times 2^m}$, $\mathrm{V} = \{1'\boldsymbol{\Pi}(D_i \# D_j)\}_{k\times k}$, $\mathrm{r} = \{(\mathrm{Y} - \mathrm{X}\beta)'\Pi D_i / 1'\Pi(D_i \# D_i)\}_{k\times 1}$, and $\mathrm{M} = \{1'\Pi(D_i \# D_j)/1'\Pi(D_i \# D_i) \times \delta(i \neq j)\}_{k\times k}$. $D_i(D_j)$ is the ith(jth) column of the genetic design matrix **D**. The notation $\delta(i \neq j)$ is an indicator variable that takes value 1 if $i \neq j$, and 0 otherwise, and # denotes Hadamard product, which is the element-by-element product of corresponding elements of two same order matrices. For more detailed procedures of the derivation see Kao and Zeng (1997). The E and M steps are iterated until a convergent criterion is satisfied. The converged values are the MLEs. The asymptotic variance-covariance matrix can also be obtained using the general formulas. The general formulas can be easily applied to obtaining the MLEs and evaluating the likelihoods for different genetic models and population structures by setting up the corresponding genetic design matrix **D** and conditional QTL genotype probability matrices $\mathbf{Q}_j$'s. Through comparisons of the likelihoods, hypotheses about the parameters of QTL can be tested by the LRT.

STRATEGY OF QTL MAPPING

For the MIM approach, the second problem that needs to be considered is how to search for QTL to fit into the MIM model. It is quite common that genetic marker data, *e.g.*, rice (Li *et al.* 1997), pine (Aitkin *et al.* 1997), and eucalyptus (Grattapaglia *et al.* 1996), contain more than 100 markers in several linkage groups to cover most of the genome. A QTL is potentially located in any position of each interval. To detect QTL using the MIM model, model selection procedures are considered because all possible subset selection is not feasible. There are at least three basic model selection techniques, forward, backward, and stepwise selections, for exploring the relationship between the independent and dependent variables (Draper and Smith 1981; Kleinbaum *et al.* 1988; Miller 1990). Several selection criteria, such as Akaike information crite-

rion (AIC; Akaike 1974), cross-validation (Stone 1974), predictive sample reuse (Geisser 1975), Baysian information criterion (BIC; Schwarz 1978), minimum posterior predictive loss (Gelfand and Ghosh 1998), or LRT statistic for selection of variables can be incorporated with model selection techniques to determine the final model. In QTL mapping, any criterion used has to take the genetic marker data structure, such as genome size and distribution of markers, into account. There have been studies on the connection of the LRT statistic to the data structure (see below). So far, however, these related studies lack other criteria. The stepwise selection technique with the LRT statistic as a criterion is adopted for identifying QTL here.

Critical value for claiming QTL detection: When using the LRT statistic as a criterion in model selection for QTL detection, it is very important to determine the appropriate critical value for claiming QTL detection such that correct statistical inference about QTL parameters can be made. Lander and Botstein (1989) suggested using the Bonferroni argument for the sparse-map case and Orenstein-Uhlenbeck diffusion for the dense-map case to determine the critical value. Generally, it has been pointed out that the critical value might need to be adjusted for the number and size of interval, different levels of heritability, different number of multiple linked or unlinked QTL, and linked QTL in the same or opposite direction of effects (Lander and Botstein 1989; Jansen 1993; Zeng 1994). Visscher and Haley (1996) suggested that the critical value should be reduced after a QTL of large effect has been detected. However, most of this information is not available before mapping, and consequently the answers to most of the above questions remain unknown. Churchill and Doerge (1994) therefore suggested using a permutation test for determining an appropriate critical value for specific data sets.

The above considerations on critical value are for the single-QTL model. For a multiple-QTL model, a model selection procedure is required to determine the final model. If stepwise selection is used, the final model is selected from a sequence of nested tests, and the significance level of the sequence will depend on the unknown true model (Atkinson 1980; Terasvirta and Mellin 1986). Therefore, the critical value of the multiple-QTL model depends not only on the above considerations but also on the unknown true model, and the choice of critical value for claiming QTL detection becomes even more complicated for MIM. We are not sure currently what the appropriate critical value is for the MIM model. In practice, the critical value from IM or CIM based on the Bonferroni argument may be used until the complicated issue of choosing the significance level for the multiple-QTL model is solved.

Stepwise selection procedure: The stepwise selection begins with no QTL ($m = 0$). QTL are then added or deleted one by one in the model. Alternatively, a group of QTL can be added or deleted together. The testing hypotheses for adding or deleting one additional QTL Q_i are

$$\begin{aligned} &H_0\text{: } a_i = 0 \\ &H_1\text{: } a_i \neq 0, \end{aligned} \tag{9}$$

given other, say, k QTL in the model. In hypotheses 9, a_i denotes the effect of Q_i. A LRT statistic

$$\text{LRT} = -2 \log \frac{L_0}{L_1}$$

is used for testing the hypotheses, where L_0 and L_1 are the likelihoods of the MIM models with k and $k + 1$ QTL, respectively. If a group of QTL is tested, the hypothesis testing would contain several QTL effects. The stepwise model selection procedure proceeds as follows:

Step 1: Significant values for entry (SVE) and staying (SVS) of a LRT statistic are specified for adding and dropping a QTL in the MIM model. Note that SVE and SVS could be different in model selection.

Step 2: For each position on the genome covered by markers, the LRT statistic reflecting the contribution of the putative QTL to quantitative trait variation is calculated ($m = 1$; IM). If there are positions with LRT statistics larger than SVE, the position with the largest value will be selected and added first in the model. When $m = 1$, it is important to note the shape of the likelihood profile and the direction of effect change along the genome for further mapping. Note that quite often no position is found with the LRT statistic larger than SVE when $m = 1$ because individual QTL contribute little to the trait variation. Two alternative approaches are proposed to prevent the procedure from stopping at a very early stage.

First, when $m = 1$, the position with the highest LRT statistic is automatically included in the model to initiate the procedure. In our experience, when only one QTL is considered in the model ($m = 1$), it is quite often found that the LRT statistic of a QTL could be less than SVE. But, when multiple QTL (if any) are accumulated in the model ($m > 1$), the partial LRT statistics of individual QTL might become significant because more genetic variation is removed from residual variation by taking multiple QTL into account.

Second, chunkwise selection (Kleinbaum *et al.* 1988) can be used. For closely linked QTL with opposite effects, more than one QTL may be selected in the model as a chunk to effectively reduce the genetic residue in the model. If only one of them is selected, its contribution may not be significant because the effect is canceled out due to failure to consider the others. When $m = 1$ in the MIM model, the chromosome region with significant change in the directions of effect could suggest that linked QTL with opposite effects are present. Also, epistatic QTL can constitute a chunk. If QTL interact, they may not be significant if only one of them is consid-

ered, but they could be significant if they are considered together. Note that the critical value should be higher for chunkwise selection because more parameters are tested. Chunkwise selection allows the incorporation of prior knowledge and preference into the model selection procedure.

Step 3: After the first k QTL are added to the model, the MIM model with $m = k + 1$ QTL is considered. The position that produces the most significant partial LRT statistic at the SVE level is added into the model. After the $k + 1$ QTL are fitted to the model, stepwise selection checks all the QTL and deletes any QTL that does not produce a significant partial LRT statistic at the SVS level. Note that a QTL that enters at an early stage may become superfluous at a later stage in stepwise selection procedure. By the same argument, chunkwise selection ($m = k + l$, $l > 1$) can be implemented. The stepwise process ends when none of the other positions has a partial LRT statistic significant at the SVE level.

Separating linked QTL: The evidence of multiple-linked QTL clustering in a region could be suggested by the shape of the likelihood profile, for example, a likelihood profile with a wide range of significant multiple peaks, or by significant change in the direction of estimated QTL effects on a chromosome region. To separate closely linked QTL in a certain chromosome region, we can compare the likelihood of the multiple-QTL model with that of a single-QTL model in this region for separation.

Analyses of epistasis: For a backcross population, it can be shown that if epistasis is present and ignored in mapping, the estimates of main effects of epistatic QTL are asymptotically unbiased whether epistasis between QTL is considered in the model or not, and the power of the test for detecting epistatic QTL could be low (appendix). Therefore, when mapping QTL without considering epistasis in a backcross population, the positions and effects of the identified QTL could still be unbiased. For l QTL being tested, there are $k = l(l - 1)/2$ possible digenic epistases. For each pair of QTL Q_i and Q_j, the hypotheses for testing their epistatic effect w_{ij} are

$$\begin{aligned} &H_0: w_{ij} = 0 \\ &H_1: w_{ij} \neq 0, \end{aligned} \qquad (10)$$

given the l QTL in the MIM model. Again, the LRT is used to test the hypotheses. The hypotheses in Equation 10 can also be used to identify QTL with no main effect but interacting with other QTL. To choose the critical value for epistasis detection, a Bonferroni argument can be used. The critical value for rejection of H_0 is suggested as $\chi^2_{1,\alpha/k}$, where α is the overall significance level.

Fine tuning the estimates of QTL positions and effects: In the above procedures, the estimates of QTL effects and positions were obtained individually. Therefore, the model likelihood might not be at the maximum, and the model is not the final model. To obtain the MLEs of the positions and effects, a multidimensional search around the regions of the identified QTL is suggested. By doing this, QTL estimates can be fine tuned and the final model can be determined. With estimates of QTL positions and effects, other composite genetic parameters (*e.g.*, heritability and variance components) of a quantitative trait can be estimated and response to selection can be predicted.

Construction of the confidence interval for QTL positions and effects: It is important to construct the confidence interval (C.I.) for QTL effects and positions. For example, when a particular QTL is to be transferred to a recipient, a C.I. of QTL position estimate can give us an idea about how large a chromosome segment is around the detected position to be transferred. There are several approaches to constructing a C.I. of the QTL positions and effects, including lod support interval (Lander and Botstein 1989), bootstrapping, using asymptotic standard deviation (ASD; Darvasi *et al.* 1993; Kao and Zeng 1997), and the methods by Dupuis and Siegmund (1999). Darvasi *et al.* (1993) and Kao and Zeng (1997) suggested using $(\hat{p} - Z_{(1-\alpha/2)}S_{\hat{p}}, \hat{p} + Z_{(1-\alpha/2)}S_{\hat{p}})$, where $\hat{p}$ and $S_{\hat{p}}$ are the estimates of QTL position and its standard deviation, to construct a C.I.

Estimation of variance components and heritability: When the final model is determined, the variance components and the heritability of the quantitative trait can be estimated. The ratio V_G/V_p, denoted by h_b^2, is called the heritability of a quantitative trait in the broad sense, where V_G and V_p are the genetic and phenotypic variances. The genetic variance V_G can be estimated by the sum of squares of the final model, and the phenotypic variance V_p can be estimated by the total sum of squares. The estimate of h_b^2 can be approximated by the coefficient of determination R^2 of the MIM model

$$\hat{h}_b^2 = \frac{\hat{V}_G}{\hat{V}_p} = \frac{\text{Model sum of squares}}{\text{Total sum of squares}} = R^2.$$

To estimate the genetic variance components, for example, the total genetic variance contributed by m QTL in the backcross population by Equation 1 is

$$V_G = \sum_{i=1}^{m}\frac{a_i^2}{4} + 2\sum_{i<j}^{m} D_{ij}a_ia_j + \sum_{i<j}^{m}\delta_{ij}\left(\frac{1}{16} - D_{ij}^2\right)w_{ij}^2, \qquad (11)$$

where D_{ij} is the gametic linkage disequilibrium coefficient between Q_i and Q_j (Weir 1996). The coefficient D_{ij} is equivalent to $(1 - 2r_{ij})/4$, where r_{ij} is the recombination fraction between two QTL. In Equation 11, the first term is the genetic variance contributed by QTL Q_i. The second term, $D_{ij}a_ia_j$, is the genetic covariance between two QTL due to linkage disequilibrium. The last term is contributed by epistasis. The genetic variance component contributed by Q_i is defined by $a_i^2/4$. However, the estimated genetic component by $\hat{a}_i^2/4$ is biased, and this bias can be corrected by $[\hat{a}_i^2 + \text{Var}(\hat{a}_i)]/4$. The genetic

covariance between Q_i and Q_j is defined by $2Da_ia_j$. By the same argument, the estimated genetic covariance by $2\hat{D}\hat{a}_i\hat{a}_j$ is also biased and can be corrected by $2\hat{D}[\hat{a}_i\hat{a}_j + \mathrm{Cov}(\hat{a}_i, \hat{a}_j)]$ under the assumption that the effect and location of QTL are independent. Other genetic components can also be estimated in the same way. For an F_2 population or a backcross population with segregation distortion, the partition of genetic variance into components is presented by C-H. Kao and Z-B. Zeng (unpublished results).

Estimation of individual genotypic value and marker-assisted selection: In plant or animal breeding, individuals with high genotypic values or favorable genotypes are usually selected for producing progeny. With the estimated QTL effects and positions, the genotypic values of individuals can be estimated by Equation 1 and the favorable QTL genotypes can be determined for selection. To select individuals with large trait values, genotype *AA* (*Aa*) of nonepistatic QTL with positive (negative) effects is preferred. For QTL with epistasis, their epistatic effects must be considered in selecting the best combination of genotypes. If QTL controlling different traits are closely linked or at the same positions, traits are genetically correlated. Selecting individuals for improvement of one trait will affect the other trait due to linkage or pleiotropy. In practice, selecting individuals with the desired character for one trait will frequently accompany an undesired character for other traits. By considering circumstances such as genetic correlation between traits, the distances between markers and QTL, and the effects of QTL, the best strategies of marker-assisted selection for (multiple) trait improvement under specific purposes and requirements can be explored.

DATA ANALYSIS

Radiata pine: Radiata pine is one of the most widely planted forestry species in the Southern Hemisphere. Two elite parents were crossed to produce 134 progeny. For each progeny, random amplified polymorphic DNA (RAPD) markers were generated, and traits measured included annual brown cone number at eight years of age, diameter of stem at breast height, and branch quality score. The cone number per tree, which varied from 0 to 45, was transformed to approximate a normal distribution using a square root transformation. The quality of branches of a tree were scored on a scale from 1 (poorest) to 6 (best). The mean of several branch quality scores denoted the branch quality of a tree. A pseudo-testcross strategy is used to construct a linkage map for each parent, and then a backcross model can be used for mapping QTL for each parent separately (Grattapaglia and Sederoff 1994; Grattapaglia *et al.* 1996). The analysis reported here is on one parent. A genetic marker map was constructed using MapMaker/EXP (Lincoln *et al.* 1993). The RAPD marker data contained 120 markers in 12 linkage groups and covered ~1679.3 cM. The average spacing of the 107 marker intervals was 13.5 cM.

As mentioned in strategy of qtl mapping, the choice of critical value is a very complicated issue for the multiple-QTL model. The value depends on the marker data structure and several unknown QTL parameters (true model). In data analysis, a critical value from IM based on Bonferroni argument is used to evaluate and illustrate the MIM approach. The SVE and SVS of the LRT statistic for claiming a QTL detection at the overall $\alpha = 0.05$ level were chosen as 12.12 ($\chi^2_{1,0.05/107} \approx \chi^2_{1,0.0005}$). For QTL selected as a chunk, the overall $\alpha = 0.05$ level was chosen as $\chi^2_{k,0.05/107}$, where k is the number of tested parameters in the chunk.

QTL detection: For trait DBH, when $m = 1$, there is no position along the genome with an LRT statistic higher than SVE. The position with the largest LRT statistic (7.85; $R^2 = 0.0639$) was found at position [12,5,0] (0 cM away from the left marker of the fifth marker interval on the twelfth linkage group). The chromosome region between C1M3 (the third marker of the first linkage group) and C1M7 showed opposite direction of effects. At C1M3, the effect was positive ($P = 0.57$), while at C1M4 and C1M5, the effects were negative ($P = 0.0253$ and 0.4181, respectively). The genetic distance between C1M3 and C1M4 is 74.8 cM. It could suggest that there are two closely linked QTL with opposite directions of effects in this region. If only one QTL ($m = 1$) is fitted in the model for search, the effect can be canceled out by opposing QTL effects. QTL will be out of detection as shown by the LRT statistic profile of IM in Figure 1. Therefore, on linkage group 1, the MIM model with $m = 2$ selected two candidate QTL, at positions [1,3,63] and [1,4,0], as a chunk. The partial LRT statistic for fitting the two QTL in the model was 13.13 (SVE and SVS for two parameters are $\chi^2_{2,0.0005} = 15.2$), and the model R^2 was 0.2104. Although the LRT statistic was less than SVE, the two QTL were selected as a chunk to initiate the stepwise selection process.

The procedure restarted at $m = 2$ by fitting two QTL with effects of opposite directions at [1,3,63] and [1,4,0]. The partial LRT statistics were 8.034 and 8.458 for the two QTL, with estimated effects 65.65 and −73.48, respectively. Given QTL at [1,3,63] and [1,4,0] in the model, a QTL at [10,5,12] with partial LRT statistic 12.83 was selected into the model ($m = 3$). The partial LRT statistics became 14.89, 15.42, and 12.83, which were all larger than the SVS of 12.12, for the three QTL. The model R^2 was 0.3202. Given these three QTL in the model, the largest partial LRT statistic 7.40 was found at position [2,2,0]. A chunkwise selection for epistatic QTL was attempted. If the candidate QTL at [2,2,0] and [12,5,12] with epistasis were selected as a chunk ($m = 5$ and one epistasis, $k = 6$), the partial LRT statistic of the chunk would be 24.76 (compared with

TABLE 1

Summary of QTL detected by MIM in Radiata pine

Linkage group	Quantitative trait loci								
	Cone number			Tree diameter			Branch score		
	Position	Effect	LRT	Position	Effect	LRT	Position	Effect	LRT
1	[1, 1, 3][a]	−0.5745	9.64	[1, 3, 61]	81.05	24.00	[1, 4, 11]	0.5273	10.37
	(4.60)[b]	(0.1796)		(1.96)	(8.48)		(7.49)	(0.1734)	
	([1, 1, 0], [1, 2, 3])[c]			([1, 3, 48], [1, 3, 66])			([1, 3, 47], [1, 5, 7])		
				[1, 4, 0]	−92.99	24.00			
				NA	(8.89)				
				([1, 4, 0], [1, 4, 1])					
2	[2, 6, 0][d]	0.5228	14.85	[2, 2, 0][f]	14.71	25.91	[2, 1, 0]	−0.4597	15.62
	NA	(0.1965)		NA	(4.49)		NA	(0.1647)	
	([2, 5, 1], [2, 6, 10])			([2, 1, 2], [2, 2, 9])			([2, 1, 0], [2, 2, 7])		
5	[5, 10, 0][d,e]	0.4537	23.77	[5, 5, 0][g]	7.16	9.48			
	NA	0.1756		NA	(4.46)				
	([5, 8, 1], [5, 10, 7])			([5, 4, 7], [5, 5, 16])					
6	[6, 4, 18]	0.8505	15.71						
	(1.63)	(0.1719)							
	([6, 3, 7], [6, 5, 5])								
10	[10, 5, 7]	1.2679	24.56	[10, 5, 9][g]	15.92	18.90			
	(1.01)	(0.2361)		NA	(4.70)				
	([10, 4, 46], [10, 5, 14])			([10, 4, 34], [10, 6, 4])					
	[10, 9, 0]	−0.9656	25.06						
	NA	(0.2230)							
	([10, 8, 2], [10, 9, 10])								
11							[11, 4, 21]	−1.3144	27.39
							(1.95)	(0.2317)	
							([11, 4, 8.5], [11, 4, 25])		
							[11, 6, 0]	1.1122	20.47
							NA	(0.2361)	
							([11, 5, 8], [11, 5, 12]		
12	[12, 3, 2][e]	−0.8178	25.03	[12, 5, 9][f]	−8.41	17.90	[12, 5, 0]	0.5085	10.36
	(1.76)	(0.1874)		NA	(4.49)		NA	(0.1597)	
	([12, 2, 6], [12, 3, 6])			([12, 4, 1], [12, 6, 3])			([12, 2, 8]), [12, 5, 11])		

NA, not available.

[a] [1, 1, 3] denotes the QTL at 3 cM away from the left marker of the first interval on linkage group 1.

[b] Asymptotic standard deviation.

[c] Lod support interval. ([1, 1, 0], [1, 2, 3]) denotes the interval with lower bound at [1, 1, 0] and upper bound at [1, 2, 3].

[d] QTL interact with epistatic effect −0.9783 (LRT = 10.22).

[e] QTL interact with epistatic effect −1.0800 (LRT = 4.48).

[f] QTL interact with epistatic effect 39.54 (LRT = 15.23).

[g] QTL interact with epistatic effect 22.64 (LRT = 4.84).

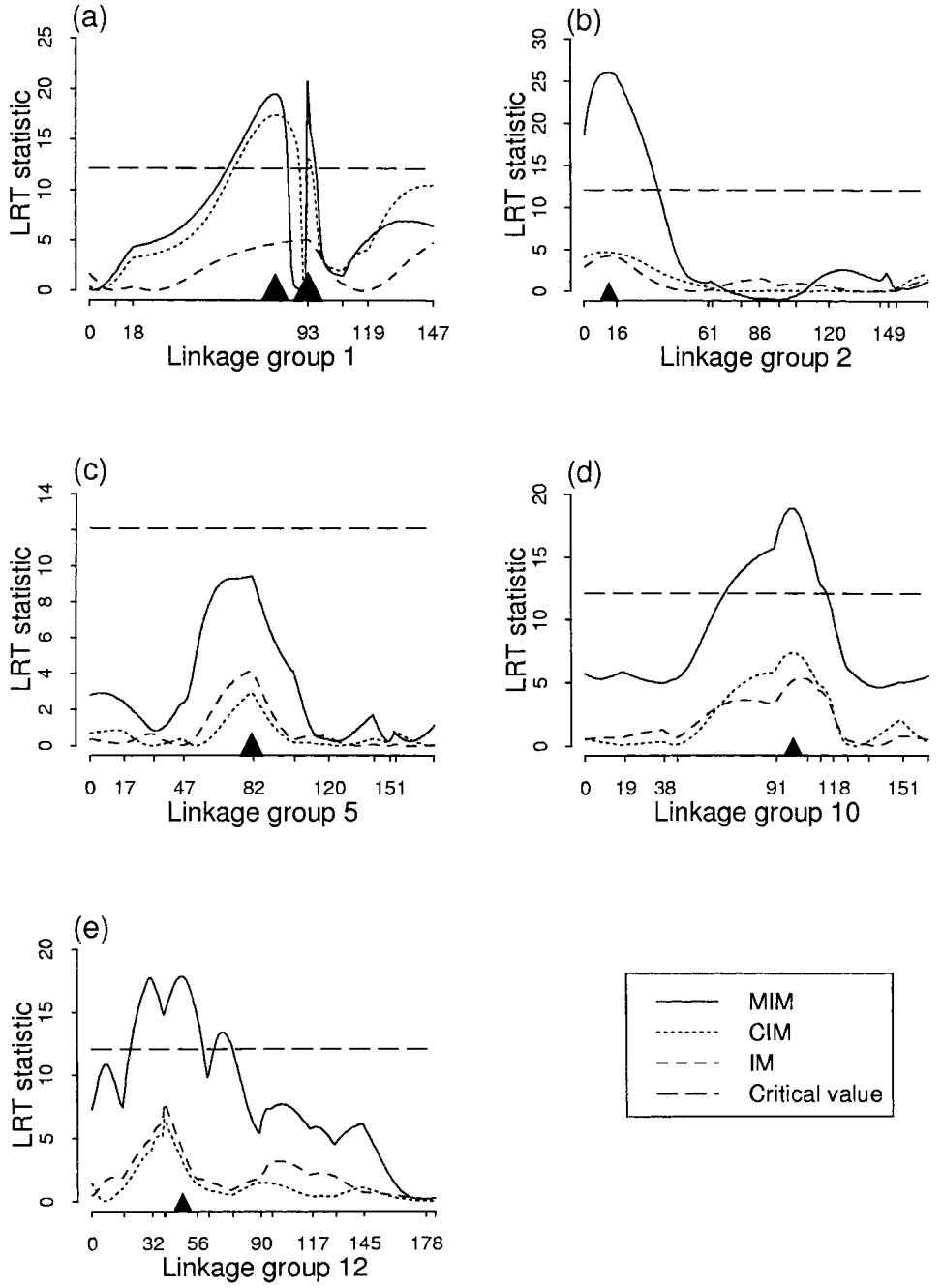

Figure 1.—Results of QTL mapping. (a–e) The solid triangles denote the QTL positions localized by MIM. The size of triangle reflects the size of QTL effect.

$\chi^2_{3,0.0005} = 17.73$). The partial LRT statistics were 23.48, 24.39, and 8.76 for the three preselected QTL at [1,3,63], [1,4,0], and [10,5,12], respectively. The QTL at [10,5,12] became nonsignificant and, therefore, was dropped from the model. Given the four QTL [1,3,63], [1,4,0], [2,2,0], and [12,5,12] in the MIM model, no other single position had a partial LRT statistic >8.76. The chunkwise selection was implemented again to find epistatic QTL. When the candidate QTL at [5,5,0] and [10,5,12] with epistasis were considered as the third chunk, the partial LRT statistic was 19.85. Adding these two epistatic QTL into the model ($m = 6$ and two epistasis, $k = 8$), the partial LRT statistics were 19.48, 20.69, and 26.91 for QTL at positions [1,3,63], [1,4,0], and the first chunk of QTL, respectively. Given the six QTL, no other QTL were identified.

Fine tuning the estimates of QTL position and effect: Two epistatic pairs were identified as described above; no other epistatic interaction between QTL was found. No QTL without main effect but interacting with the identified QTL were found. A multidimensional search around the detected QTL was used to fine tune the estimates of QTL parameters. The locations changed to [1,3,61], [1,4,0], [2,2,0], [5,5,0], [10,5,9], and [12,5,9]. The estimated QTL effects are shown in Table 1. QTL at positions [1,3,61], [2,2,0], [10,5,9], and [5,5,0] had

positive effects, and QTL at positions [1,4,0] and [12,5,9] had negative effects. The effects of QTL at positions [1,3,61] and [1,4,0] were larger when compared with others. The model R^2 was 0.5226. Therefore, six identified QTL were conclusively identified in QTL mapping for the diameter trait. The partial LRT statistic profiles for each QTL are shown in Figure 1.

Epistasis: The estimated epistatic effect between QTL at positions [2,2,0] and [12,5,9] was 39.54 (partial LRT statistic 15.23), and the epistatic effect between QTL at [5,5,0] and [10,5,9] was 22.64 (partial LRT statistic 4.84). Figure 2 shows how the QTL interact. Figure 2a shows that the effect of QTL ($G_{BB} - G_{Bb}$) at position [12,5,9] was positive in the background of homozygote QTL (AA) at position [2,2,0], but it was negative in the heterozygote background (Aa). Figure 2b shows that the QTL at position [10,5,9] had a large effect in the background of homozygote QTL (AA) at [5,5,0], but it had a small effect in the heterozygote background (Aa).

Heritability and variance components: The broad sense heritability for tree diameter can be estimated by the R^2 value of the final MIM model. The R^2 of the model including six QTL and two epistases was 0.5226. QTL at positions [2,2,0], [5,5,0], [10,5,12], and [12,5,9] contributed ~4.50, 1.36, 5.25, and 1.76% of the total genetic variance, respectively. The percentage of genetic variance contributed by the two linked QTL separated by 13.8 cM on the first linkage group was 76.75%. There was a negative genetic covariance between the two linked QTL. Two epistatic pairs contributed ~10.38% to the total genetic variance.

QTL mapping for cone number and branch quality: QTL mapping was also performed on the traits of cone number and branch score. The mapping results are listed in Table 1. For cone number, seven QTL were identified (although the QTL at [1,1,3] was not significant with partial LRT statistic 9.44, we considered it as a candidate QTL). Epistasis was found between two QTL pairs using chunkwise selection. The model R^2 value of the MIM model fitted to the seven QTL and their epistasis was 0.5606. The two linked QTL, separated by 27.6 cM on linkage group 10, contributed 29.93% of the genetic variance. The other five QTL contributed ~55.93% of the total genetic variance. Epistasis contributed 14.14%. For branch quality, five QTL were identified (we also considered the two QTL with partial LRT statistic values 10.37 and 10.36 at [1,4,11] and [12,5,0] as candidate QTL). No epistasis was found for QTL controlling branch score. The model R^2 was 0.3630. Two linked QTL, separated by 19.6 cM on linkage group 11, contributed 48.69% of the genetic variance. The remaining three QTL contributed from ~11 to 27% of the total genetic variance.

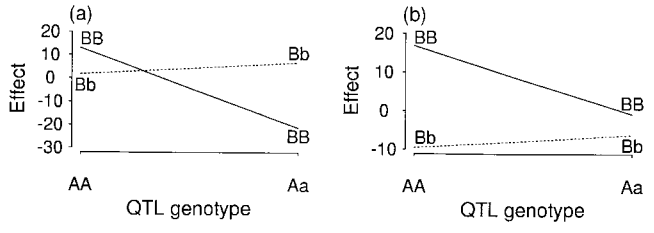

Figure 2.—Epistasis between QTL controlling tree diameter. (a) Epistasis between QTL at positions [2,2,0] and [12,5,9]. A and B denote QTL at [2,2,0] and [12,5,9], respectively. (b) Epistasis between QTL at positions [5,5,0] and [10,5,9]. A and B denote QTL at [5,5,0] and [10,5,9], respectively.

Confidence intervals of QTL positions and effects: The lod support interval and the ASD of QTL effect and position are listed in Table 1. Out of the 18 QTL detected for three traits, 9 ([2,6,0], [5,10,0], and [10,9,0] for cone number; [1,4,0], [2,2,0], and [5,5,0] for tree diameter; [2,1,0], [11,6,0], and [12,5,0] for branch score) of them were localized at the markers, and 2 ([10,5,9] and [12,5,9]) had negative ASD. Therefore, the ASD of these QTL position estimates were not available for constructing C.I.'s. The asymmetric lod support intervals are typical in this case. For example, the diameter QTL at [5,5,0] has an asymmetric lod support interval ([5,4,7], [5,5,16]). In general, the interval constructed by ASD is much narrower than the lod support interval. For example, C.I.'s constructed using four times ASD were 6.52 and 7.04 cM for the cone QTL at [6,4,18] and [12,3,2], and the lod support intervals are 59.6 cM and 14.6 cM, respectively.

Marker-assisted selection: Individuals with favorable QTL genotypes are selected as parents to produce progeny. Trees carrying all the favorable QTL genotypes were not found for each trait in the sample. Therefore, only a subset of the detected QTL was considered in selection. For tree diameter, three trees were found to carry favorable genotypes and two trees were found to carry unfavorable genotypes (consider epistasis) of the five QTL (out of the six detected QTL) at positions [1,4,0], [2,2,0], [5,5,0], [10,5,9], and [12,5,9]. The observed trait means for the two groups were 232.38 and 163.05 mm, respectively, through selection of these five diameter QTL. The estimated genotypic values of the two groups were 233.84 and 160.06 mm (Table 2). The observed and estimated values of performing selection for the other two traits on the sample based on four and five QTL are also shown in Table 2. The mapping results in Table 1 also allow us to estimate the genotypic values of certain genotypes. For example, if trees carrying all six favorable diameter QTL were selected with epistasis taken into consideration, the estimated tree diameter for those trees would be 314.17 mm and the estimated cone number would be 8.22. If trees carry all seven favorable QTL (epistasis considered) for reducing cone number, the estimated cone number would be 0.33 and the estimated tree diameter would be 196.45 mm. Consequently, the improvement of tree diameter would cause simultaneous increase in cone number,

TABLE 2

Comparison of the predicted and observed means of the selected populations

			Subpopulation					
	No. of QTL		Select decrease		Select increase		Unselected population	
Trait	Detected	Applied	Observed	Predicted	Observed	Predicted	Observed	Predicted
CN[a]	7	4[b]	0.83 (0.43)	0.90	13.61 (·)	17.03	9.93 (8.29)	7.83
DBH[c]	6	5[d]	163.05 (35.64)	160.06[g]	232.38 (20.00)	233.84[g]	197.75 (34.49)	197.68
BS[e]	5	3[f]	1.24 (·)	2.20	5.01 (0.82)	5.13	3.70 (1.07)	3.66

CN, DBH, and BS denote cone number, diameter, and branch score, respectively. Numbers in parentheses are standard deviations.

[a] Numbers of individual trees are 3, 1, and 113 in the three subpopulations.
[b] Select cone QTL at [2, 6, 0], [5, 10, 0], [10, 5, 7], and [10, 9, 0].
[c] Numbers of individuals are 2, 3, and 129 in the three populations.
[d] Select diameter QTL at [1, 4, 0], [2, 2, 0], [5, 5, 0], [10, 5, 9], and [12, 5, 9].
[e] Numbers of individuals are 1, 5, and 128 in the three populations.
[f] Select branch score QTL at [11, 4, 21], [11, 6, 0], and [12, 5, 0].
[g] Assume that the QTL at [1, 3, 61] and [1, 4, 0] have coupling phase.

which is a reflection of the positive genetic correlation between the two traits. Generally, the estimated and observed results were quite close based on the MIM result as found in this sample.

DISCUSSION

A new QTL mapping approach named MIM is proposed. It uses multiple-marker intervals simultaneously to construct multiple QTL in the model for QTL mapping. The MIM model is based on Cockerham's model (C-H. Kao and Z-B. Zeng, unpublished results) for defining genetic parameters and on the general formulas of Kao and Zeng (1997) for statistical estimation. Using the MIM model, stepwise and chunkwise selections with the LRT statistic as a selection criterion are proposed to identify QTL, to separate linked QTL, and to analyze epistasis between QTL. The asymptotic standard deviations of the estimated QTL positions and effects can be obtained for constructing the C.I.s. With the estimated QTL effects and positions provided by MIM, the variance components of QTL, the heritability of a quantitative trait, and the genotypic values of individuals can be estimated, and marker-assisted selection can be performed for trait improvement. Experimental data on three traits on radiata pine were analyzed to illustrate the potential power and benefit of MIM in comparison with the current methods, such as IM and CIM. While a backcross MIM model was used here as an example, the MIM model can be easily extended to an F_2 population (C-H. Kao and Z-B. Zeng, unpublished results).

The MIM model is a multiple-QTL model. When the multiple-QTL model is considered, the likelihood is a finite normal mixture and becomes increasingly intractable in maximization as the number of QTL fitted into the model increases (Haley and Knott 1992; Satagopan *et al.* 1996). We used the method of maximum likelihood in estimation by applying the general formulas of Kao and Zeng (1997) to maximize likelihood and obtain MLEs as well as the variance-covariance matrix of the MLEs. The MLEs have some attractive properties, such as invariance, consistency, and asymptomatic efficiency, in statistical inference. If prior information of parameters is available, the Bayesian approach, such as in Satagopan *et al.* (1996) and Sillanpaa and Arjas (1998), can be used to incorporate the prior information in estimation. By specifying prior density of parameters, they used Markov chain Monte Carlo to evaluate the posterior density and to output empirical distribution of QTL parameters for QTL mapping. The MIM approach of using the multiple-QTL model in QTL mapping distinguishes itself from the current approaches, such as IM and CIM, by the ability to use multiple-marker intervals simultaneously to search the chromosome region between markers for QTL. As a result, the MIM method may provide greater power and precision for QTL mapping. However, it should be noted that the significance level of the multiple-QTL model depends on the marker data structure and the unknown true model, and the critical value for claiming QTL detection becomes a complicated issue for MIM (see strategy of qtl mapping). In the example, we used an *ad hoc* critical value. This value is appropriate for the one-QTL model, but it may not be appropriate for MIM. Although the MIM method claims more QTL detection than the current methods in data analysis, it is not appropriate to conclude that the MIM method is better until the complicated problem of assessing the

appropriate critical value for the multiple-QTL model has been solved.

Under the *ad hoc* critical value, MIM detected six QTL for tree diameter and CIM detected only two of them on the first linkage group in this example. IM failed to detect any QTL (Figure 1). The major reason for this difference is that CIM is not capable of controlling the two detected linked QTL simultaneously in further mapping. As a result, only the QTL at position [1,4,0] is controlled, but it does not contribute substantially to reducing the genetic variation because its effect has been canceled out by ignoring the linked QTL with opposite effects at position [1,3,61]. Accordingly, most of the genetic variance (76.75%) contributed by the two linked QTL becomes part of the genetic residue, making the other four QTL undetectable. This shows the beauty of MIM, which allows the current detected QTL being fitted directly into the model to search for the next QTL. Consequently, more QTL were detected by MIM than the current methods in this example.

In the data analyses, MIM localized two linked QTL with large opposite directions of effect in the third interval of linkage group 1 (Figure 1a). They contributed 76.75% of the total genetic variance. The size of this interval was 74.8 cM, so it is suggested that more markers should be added to this interval to permit further investigation. Two linked QTL, one controlling diameter and another controlling cone number, were detected in the same fifth interval of linkage group 10 (Table 1). The estimated locations are 2 cM apart. Further investigation is needed to check if they are the same (pleiotropic) or different (closely linked) QTL. The likelihood profile of linkage group 12 in Figure 1e is a result of conditioning on the other five unlinked QTL. It shows multiple significant peaks, which could suggest multiple-linked QTL on the same linkage group. However, after further investigating the linkage group, there was no evidence of multiple QTL given the peak at position [12,5,9] and the other five detected QTL. It is therefore concluded that there is only one QTL at position [12,5,9] on linkage group 12.

Another benefit derived by MIM is that epistasis can be readily incorporated in the model for analysis or searching for epistatic QTL. When taking both main and epistatic effects into account in searching for QTL, the critical value for hypothesis testing needs to be adjusted for the extra degree of freedom for epistasis. It is interesting to know that the estimated main effects of linked QTL are asymptotically unbiased in the backcross population (appendix), but they are biased in the F_2 population if epistasis is present and ignored in mapping (Kao 1995). This is because the coded variables for main and epistatic effects in Cockerham's model are still orthogonal under linkage disequilibrium for the backcross population but not for the F_2 population. This asymptotic unbiasedness property ensures that QTL mapping could first be performed without taking epistasis into account without causing a problem in the backcross population. For tree diameter and cone number, respectively, epistasis contributed 10.38 and 14.14% of the total genetic variance. Therefore, epistasis should be generally considered in searching for QTL and marker-assisted selection. For example, in Figure 2a, the best combination of QTL genotype at positions [2,2,0] and [12,5,9] was *AABB*, which had an estimated genotypic value of 13.2. If epistasis was ignored, genotype *AABb*, with estimated genotypic value 1.75, would be selected. The benefit of taking epistasis into account was reflected in the mapping result in Table 1 and in grouping genotypes in Table 2.

It has been 76 yr since Sax (1923) associated seed coat characters with seed size in beans. The QTL mapping model has evolved from using marker analyses, *e.g.*, *t*-test, simple or multiple regression, to one-QTL model (IM and CIM), and further to the multiple-QTL model, such as the MIM approach. In practice, the detected QTL will be used for selecting parents with desired genotypes for producing progeny or gene transfer to achieve the ultimate goal of trait improvement in later generations. QTL have to be mapped as precisely as possible to ensure good quality of the follow-up operation on QTL. Therefore, precision and unbiasedness in estimating the parameters of QTL should be more important than the ease of computation and implementation in QTL mapping. The computation burden of the multiple-QTL model is heavy when compared with the one-QTL model. However, the gain of doing so, as shown in this article, could be significant. Although further work is needed to establish a theoretical basis for determining an appropriate criterion of model selection in QTL mapping under MIM, MIM has the potentiality to be more powerful and more precise in QTL mapping by directly conditioning putative QTL and incorporating possible epistasis in the model. Thus, more genetic variation can be controlled in the model. With the estimates of QTL parameters, other composite genetic parameters, such as the genetic variance components and heritabilities, can also be estimated. In addition, based on the MIM results, genotypic values of individuals can be estimated to allow desired genotypes to be selected in marker-assisted selection under various requirements (*e.g.*, cost, efficiency, and trait correlations).

An initial version of the MIM program source code (written in Fortran 77 language) is available on the worldwide web (http://www.stat.sinica.edu.tw/~chkao/). A more user-friendly package can be developed based on this program. Using the MIM program, we implemented stepwise and chunkwise selections with the LRT statistic as a selection criterion to search for QTL in data analyses. In analyzing the data, we chose the two linked QTL with opposite direction of effect on the first linkage as a starting point to initiate the selection process, and six QTL were found for tree diameter. We

also tried another possible starting point at [12,5,0] to initiate the process and obtained the same final model. This final model obtained by model selection might not be optimal. Even though the optimal model was obtained, there is no guarantee that it is the true model (the estimated QTL are the true QTL) for limited sample size. Ultimately, the reliability of the identified QTL will depend on further experiments to assess the validity of QTL. There is no single criterion that plays the role of a panacea in the model selection problem. Other model selection techniques and criteria could also be implemented. It is a very important task to explore and automate the model selection procedures of the MIM approach for general use in the QTL mapping community.

We are greatly indebted to Dr. Chung-I Wu and three anonymous reviewers for their comments and criticisms. Chen-Hung Kao is grateful to Corinna Lange for her suggestions. C-H.K. was supported by grants NSC87-2313-B-324-001 and NSC88-2313-B-324-001 from the National Science Council, Taiwan, Republic of China; Z-B.Z. was funded by GM-45344 from the National Institutes of Health and no. 9600645 from the United States Department of Agriculture Plant Genome Program.

LITERATURE CITED

Aitken, K. S., G. Smail, J. Drenth, Y. Li, C.-H. Kao *et al.*, 1997 Detection of quantitative trait loci (QTL) for cone production in *Pinus radiata*, pp. 337–341 in *IUFRO '97 Genetics of Radiata Pine*, edited by R. D. Burdon and J. M. Moore. Proceedings of NZFRI-IUFRO Conference, 1–4 December, and Workshop, 5 December, Rotorua, New Zealand, FRI Bulletin no. 203.

Akaike, H., 1974 A new look at the statistical model identification. IEEE Trans. Auto. Control 19: 716–723.

Atkinson, A. C., 1980 A note on the generalized information criterion for the choice of a model. Biometrika 67(2): 413–418.

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.

Churchill, G. A., and R. W. Doerge, 1994 Empirical threshold values for quantitative trait mapping. Genetics 138: 967–971.

Darvasi, A., A. Weinreb, V. Minke, J. I. Weller and M. Soller, 1993 Detecting marker-QTL linkage and estimating QTL gene effect and map location using a saturated genetic map. Genetics 134: 943–951.

Dempster, A. P., N. M. Laird and D. B. Rubin, 1977 Maximum likelihood from incomplete data via the EM algorithm. Journal 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.

Draper, N. R., and H. Smith, 1981 *Applied Regression Analysis*, Ed. 2. John Wiley & Sons, New York.

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

Geisser, S., 1975 The predictive sample reuse method with applications. J. Am. Stat. Assoc. 70: 320–328.

Gelfand, A. E., and S. K. Ghosh, 1998 Model choice: a minimum posterior predictive loss approach. Biometrika 85: 1–11.

Grattapaglia, D., and R. R. Sederoff, 1994 Genetic linkage map of *Eucalyptus grandis* and *Eucalyptus urophylla* using a Pseudo-Testcross: mapping strategy and RAPD markers. Genetics 137: 1121–1137.

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.

Grignola, F. E., I. Hoeschele and B. Tier, 1996a Mapping quantitative trait loci via residual maximum likelihood: II. A simulation study. Genet. Sel. Evol. 28: 479–490.

Grignola, F. E., I. Hoeschele, Q. Zhang and G. Thaller, 1996b Mapping quantitative trait loci via residual maximum likelihood: I. Methodology. Genet. Sel. Evol. 28: 491–504.

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.

Hoeschele, I., and P. Vanranden, 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. Vanranden, 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.

Jansen, R. C., and P. Stam, 1994 High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136: 1447–1455.

Jayakar, S. D., 1970 On the detection and estimation of linkage between a locus influencing a quantitative trait character and a marker locus. Biometrics 26: 451–464.

Jiang, C., and Z-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers. Genetica 101: 47–58.

Kao, C.-H., 1995 Statistical methods for locating the positions and analyzing epistasis of multiple quantitative trait genes using molecular marker information. Ph.D. Thesis, North Carolina State University, Raleigh.

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

Kleinbaum, D. G., L. L. Kupper and K. E. Muller, 1988 *Applied Regression Analysis and Other Multivariate Methods.* PWS-KENT Publishing Company, Boston.

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

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.

Lincoln, S. E., M. J. Daly and E. S. Lander, 1993 *Constructing Genetic Linkage Maps with Mapmaker/Exp Version 3.0.* The Whitehead Institute, Cambridge, MA.

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.

Miller, A. J., 1990 *Subset Selection in Regression.* Chapman and Hall, London.

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.

Sax, K., 1923 The association of size difference with seed-coat pattern and pigmentation in *Phaseolus vulgaris.* Genetics 8: 552–560.

Schwarz, G., 1978 Estimating the dimension of a model. Ann. Stat. 6: 461–464.

Sillanpaa, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics 148: 1373–1388.

Stone, M., 1974 Cross-validation choice and assessment of statistical predictions. I. R. Stat. Soc. B 36: 111–147.

Terasvirta, T., and I. Mellin, 1986 Model selection criteria and model selection tests in regression models. Scand. J. Stat. 13: 159–171.

Thoday, J. M., 1960 Location of polygenes. Nature 191: 368–370.

Visscher, P. M., and C. S. Haley, 1996 Detection of the putative quantitative trait loci in line crosses under infinitesimal genetic models. Theor. Appl. Genet. 93: 691–702.
Weir, B. S., 1996 *Genetic Data Analysis II.* Sinauer Associates, Sunderland, MA.
Xu, S., 1995 A comment on the simple regression method for interval mapping. Genetics 141: 1657–1659.
Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics 141: 1189–1197.
Xu, S., and W. R. Atchley, 1996 Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143: 1417–1424.
Zeng, Z-B., 1993 Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. Proc. Natl. Acad. Sci. USA 90: 10972–10976.
Zeng, Z-B., 1994 Precision mapping of quantitative trait loci. Genetics 136: 1457–1468.

Communicating editor: C.-I Wu

APPENDIX: THE PROBLEMS OF IGNORING EPISTASIS IN QTL MAPPING

To simplify the argument, consider the situation where the test positions for QTL are located precisely at the marker position. If only two epistatic QTL, A (x_1) and B (x_2), control a quantitative trait y, the single-marker regression coefficient of y on one of the QTL, say x_1, is given by $b_{yx_1} = \mathrm{Cov}(y, x_1)/V(x_1)$, where $\mathrm{Cov}(y, x_1)$ is the covariance between the trait and QTL A and $V(x_1)$ is the variance of QTL A. Assuming that there is no covariance between environmental deviation and QTL, it is easy to show that

$$\begin{aligned}\mathrm{Cov}(y,x_1) &= \mathrm{Cov}(\mu + a_1x_1 + a_2x_2 + wx_1x_x + \varepsilon_i,x_1)\\ &= a_1V(x_1) + a_2\mathrm{Cov}(x_1,x_2) + w\,\mathrm{Cov}(x_1x_2,x_1)\\ &= \frac{a_1}{4} + \frac{1-2r}{4}a_2 \qquad (12)\end{aligned}$$

because $\mathrm{Cov}(x_1x_2, x_1) = 0$ under Cockerham's model (C-H. Kao and Z-B. Zeng, unpublished results). The single-marker regression coefficient is $b_{yx_1} = a_1 + (1 - 2r)a_2$ because $V(x_1) = 1/4$. The main effect of the linked QTL B is involved in the estimation, but epistatic effect w is not involved. If both QTL A and B are considered in the model, the partial regression coefficient $b_{yx_1.x_2}$ becomes

$$\begin{aligned}b_{yx_1.x_2} &= \frac{\sigma_{yx_1.x_2}}{\sigma^2_{x_2}} = \sigma_{yx_1} - \frac{\sigma_{yx_2}\sigma_{x_1x_2}}{\sigma^2_{x_2}}\\ &= \frac{(1-(1-2r)^2)/4\,a_1}{(1-(1-2r)^2)/4} = a_1, \qquad (13)\end{aligned}$$

where $\sigma_{yx_1.x_2}$ and $\sigma^2_{x_1.x_2}$ denote the conditional covariance of trait y and QTL x_1 on QTL x_2 and conditional variance of x_1 on x_2 (Zeng 1993). In the same way, the partial regression coefficient $b_{yx_2.x_1}$ for QTL B is a_2. That is, the partial regression coefficients are asymptotically unbiased for main effects of QTL and not affected by epistasis if epistasis is present but ignored in the backcross population. In an F_2 population, the partial regressions would be affected by epistasis if epistatic QTL are linked (Kao 1995). However, if epistatic effect is not fitted into the model, the genetic variance contributed by epistasis is not controlled and becomes part of the residue in the model, and the power of detection could be low. This conclusion can be applied to mapping QTL.