BIOMETRICS 53, 653–665
June 1997

General Formulas for Obtaining the MLEs and the Asymptotic Variance–Covariance Matrix in Mapping Quantitative Trait Loci When Using the EM Algorithm

Chen-Hung Kao and Zhao-Bang Zeng

Program in Statistical Genetics, Department of Statistics,
North Carolina State University, Raleigh, North Carolina 27695-8203, U.S.A.

SUMMARY

We present in this paper general formulas for deriving the maximum likelihood estimates and the asymptotic variance–covariance matrix of the positions and effects of quantitative trait loci (QTLs) in a finite normal mixture model when the EM algorithm is used for mapping QTLs. The general formulas are based on two matrices **D** and **Q**, where **D** is the genetic design matrix, characterizing the genetic effects of the QTLs, and **Q** is the conditional probability matrix of QTL genotypes given flanking marker genotypes, containing the information on QTL positions. With the general formulas, it is relatively easy to extend QTL mapping analysis to using multiple marker intervals simultaneously for mapping multiple QTLs, for analyzing QTL epistasis, and for estimating the heritability of quantitative traits. Simulations were performed to evaluate the performance of the estimates of the asymptotic variances of QTL positions and effects.

1. Introduction

With the rapid advances in molecular biology, it has become possible to gain fine-scale genetic maps for various organisms by determining the genomic positions of a number of genetic markers (RFLP, isozymes, RAPDs, AFLP, VNTRs, etc.) and to obtain a complete classification of marker genotypes by using codominant markers. These advances greatly facilitate the mapping and analysis of individual quantitative trait loci (QTLs).

Thoday (1960) first proposed the idea of using two markers to bracket a region for testing QTLs. Lander and Botstein (1989) implemented a similar, but much improved, method to use two adjacent markers to test for the existence of a QTL in the interval by performing a likelihood ratio test (LRT) at every position in the interval. This is termed interval mapping (IM). The identification of QTLs by IM has been reported in tomato (Paterson et al., 1991), maize (Stuber et al., 1992), pig (Andersson et al., 1994), and forest trees (Bradshaw and Stettler, 1995). However, IM can bias identification and estimation of QTLs when multiple QTLs are located in the same linkage group (Lander and Botstein, 1989; Haley and Knott, 1992; Jansen, 1993; Zeng, 1994). It is also not efficient to use only two markers at a time for mapping analysis. In view of these problems, QTL mapping combining IM with multiple marker regression analysis is proposed (Jansen, 1993; Zeng, 1993). Zeng (1994) named this combination composite interval mapping (CIM). A great improvement in identification of QTLs by CIM has been reported in mice (Dragani et al., 1995) and in *Drosophila* (Liu et al., 1996).

CIM avoids the use of multiple marker intervals to deal with the problems of mapping multiple QTLs by conditioning a test for a QTL on some linked or unlinked markers that diffuse the effects of other potential QTLs. Ideally, we would like to generalize this analysis to using multiple marker intervals simultaneouly to solve the problem of multiple QTLs, either by a multidimensional search for multiple QTLs or by a one-dimensional search for a QTL in one interval conditioning on the

Key words: Asymptotic variance–covariance matrix; EM algorithm; Epistasis; Gene mapping; General formulas; Heritability; Maximum likelihood; Normal mixture model; Quantitative trait loci.

identified QTLs in other intervals. We are often restricted to using only one marker interval at a time because of the lack of a systematic formulation for deriving the maximum likelihood estimates (MLEs) for the finite normal mixture model when an arbitrary number of intervals are used. Furthermore, most of the current QTL mapping methods, like IM and CIM, provide only point estimates of QTL positions and effects. It is often of importance to know the asymptotic variance–covariance matrix of the estimates. Darvasi et al. (1993) derived the MLEs and the asymptotic variance–covariance matrix of QTL position and effects for IM using the Newton–Raphson method. We present general formulas for deriving the MLEs and the asymptotic variance–covariance matrix using the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977; Meng, 1993, 1994). The reasons for using the EM algorithm are that it is easily implemented and is numerically stable (Meng and Rubin, 1991), especially when multiple marker intervals are considered simultaneously. With the general formulas, it is straightforward to fit multiple putative QTLs in the model to improve the power and precision of mapping, to analyze QTL epistasis, and to estimate the heritability of quantitative traits. Here, we use an F_2 population as an example to explain the general formulas and then apply the general formulas to a backcross population for a simulation study of the performance of the asymptotic variances of QTL positions and effects.

2. Experimental Populations

We consider populations derived from a cross between two parental inbred lines P_1 and P_2, differing substantially in a quantitative trait of interest. Let loci M, with alleles M and m, and N, with alleles N and n, denote two flanking markers for an interval where a putative QTL is being tested. A cross between two parents P_1 and P_2 is performed to produce an F_1 population. The F_1 progeny are all heterozygotes with the same genotype MN/mn. If the F_1 individuals are backcrossed to P_1 or P_2, it produces a backcross population. There are four possible marker genotypes in the backcross population (Table 1). If the F_1 individuals are selfed or intermated, it produces an F_2 population. There are nine observable marker genotypes in the F_2 population (Table 2). Let the unobserved QTL locus Q with alleles Q and q be located in the interval flanked by markers M and N. The distribution of unobserved QTL genotypes can be inferred from the observed flanking marker genotypes according to the recombination frequencies between them. To infer the distribution of QTL genotype, we assume that there is no crossover interference and also that double recombination events within the interval are very rare and can be ignored to simplify the analysis. The conditional probabilities of the QTL genotypes given marker genotypes are given in Table 1 for the backcross population and in Table 2 for the F_2 population. We extract the conditional probabilities in Tables 1 or 2 to form a matrix $\mathbf{Q}$ for each specific population.

3. Genetic Model

Consider a QTL in the F_2 population in which the frequencies of genotypes QQ, Qq, and qq are 1/4, 1/2, and 1/4, respectively. The genetic model for a QTL

$$\mathbf{G} = \begin{bmatrix} G_2 \\ G_1 \\ G_0 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1 & -1/2 \\ 0 & 1/2 \\ -1 & -1/2 \end{bmatrix} \begin{bmatrix} a \\ d \end{bmatrix} = \mathbf{1}_{3\times 1}\mu + \mathbf{D}\mathbf{E} \tag{1}$$

Table 1
Conditional probabilities of a putative QTL given the flanking marker genotypes for a backcross population

Marker genotype	Expected frequency	QTL genotype	
		QQ	Qq
MN/MN	$(1-r_{MN})/2$	1	0
MN/Mn	$r_{MN}/2$	$1-p$	p
MN/mN	$r_{MN}/2$	p	$1-p$
MN/mn	$(1-r_{MN})/2$	0	1

$p = r_{MQ}/r$, where r_{MQ} is the recombination fraction between the left marker M and the putative QTL and r is the recombination fraction between the two flanking markers M and N. The possibility of a double recombination event in the interval is ignored.

Table 2
Conditional probabilities of a putative QTL given the flanking marker genotypes for an F_2 population

Marker genotype	Expected frequency	QTL genotype		
		QQ	Qq	qq
MN/MN	$\frac{(1-r)^2}{4}$	1	0	0
MN/Mn	$\frac{r(1-r)}{2}$	$1-p$	p	0
Mn/Mn	$\frac{r^2}{4}$	$(1-p)^2$	$2p(1-p)$	p^2
MN/mN	$\frac{r(1-r)}{2}$	p	$1-p$	0
MN/mn or Mn/mN	$\frac{(1-r)^2}{2}+\frac{r^2}{2}$	$cp(1-p)$	$1-2cp(1-p)$	$cp(1-p)$
Mn/mn	$\frac{r(1-r)}{2}$	0	$1-p$	p
mN/mN	$\frac{r^2}{4}$	p^2	$2p(1-p)$	$(1-p)^2$
mN/mn	$\frac{r(1-r)}{2}$	0	p	$1-p$
mn/mn	$\frac{(1-r)^2}{4}$	0	0	1

$p = r_{MQ}/r_{MN}$, where r_{MQ} is the recombination fraction between the left marker M and the putative QTL and r_{MN} is the recombination fraction between the two flanking markers M and N. $c = r_{MN}^2/[r_{MN}^2 + (1-r_{MN})^2]$. The possibility of a double recombination event in the interval is ignored.

was proposed to model the relation between a genotypic value G and the genetic parameters μ, a and d. G_2, G_1, and G_0 are the genotypic values of genotypes QQ, Qq, and qq. The unique solutions of the genetic parameters in terms of genotypic values and frequencies are

$$\mu = \frac{G_2}{4} + \frac{G_1}{2} + \frac{G_0}{4}, \qquad a = \frac{G_2 - G_0}{2}, \quad \text{and} \quad d = \frac{2G_1 - G_2 - G_0}{2}.$$

Therefore, the genetic parameter μ is the mean and a and d denote the additive and dominance effects of QTL in the F_2 populaton, respectively. We call $\mathbf{D}$ the genetic design matrix. The first and second columns of $\mathbf{D}$, denoted by D_1 and D_2, represent the status of the additive and dominance parameters of the three different genotypes.

4. Statistical Model for QTL mapping

We assume no epistasis between QTLs, no interference in crossing over, and only one QTL in the testing interval. The analysis of QTL epistasis using this approach will be discussed later. QTL mapping data consist of two parts, y_j $(j = 1, \ldots, n)$ for the quantitative trait value and $\mathbf{X}_j$ $(j = 1, \ldots, n)$ for the genetic markers and other explanatory variables, for example sex or diet. A CIM statistical model based on the genetic model for testing a QTL in a marker interval is proposed as

$$y_j = ax_j^* + dz_j^* + X_j\beta + \epsilon_j, \tag{2}$$

where

$$x_j^* = \begin{cases} 1 & \text{if the QTL is } QQ \\ 0 & \text{if the QTL is } Qq \\ -1 & \text{if the QTL is } qq \end{cases} \quad \text{and} \quad z_j^* = \begin{cases} \frac{1}{2} & \text{if the QTL is } Qq \\ -\frac{1}{2} & \text{otherwise;} \end{cases}$$

y_j is the quantitative trait value of the jth individual; a and d are the additive and dominance effects, respectively, of the putative QTL; X_j, a subset of $\mathbf{X}_j$, may contain some chosen markers and other expanatory variables; β is the partial regression coefficient vector including the mean μ; and ϵ_j is a random error. We assume $\epsilon_j \sim \mathrm{N}(0, \sigma^2)$. The advantages of using X_j in QTL mapping have been discussed by Zeng (1993, 1994). Basically, it could control for the confounding effect of linked QTLs and reduce the residual variance in the analysis.

5. Likelihood of the Statistical Model

Given the data with n individuals, the likelihood function for $\theta = (p, a, d, \beta, \sigma^2)$ is

$$L(\theta \mid \mathbf{Y}, \mathbf{X}) = \prod_{j=1}^{n} \left[\sum_{i=1}^{3} p_{ji} \phi \left(\frac{y_j - \mu_{ji}}{\sigma} \right) \right], \tag{3}$$

where $\phi(\cdot)$ is a standard normal probability density function, $\mu_{j1} = a - d/2 + X_j\beta$, $\mu_{j2} = d/2 + X_j\beta$, and $\mu_{j3} = -a - d/2 + X_j\beta$. Statistically, this is a normal mixture model (Titterington, Smith, and Makov, 1985). The density of each individual is a mixture of three possible normal densities with different means and mixing proportions. The mixing proportions p_{ji}'s, which are functions of the QTL position parameter p, are conditional probabilities of QTL genotypes given marker genotypes. They are given in Tables 1 and 2 for backcross and F_2 populations. To obtain the MLEs of the likelihood, we use the EM algorithm, treating the normal mixture model as an incomplete-data problem.

6. Hypothesis Testing

In QTL mapping, we test whether there is a QTL at a given position within a marker interval. The hypotheses are

H_0: $a = 0$ and $d = 0$ (there is no QTL at that position),

H_1: at least one of them is not 0 (there is a QTL at that position).

To test the hypothesis, we use an LRT statistic $-2\log[\sup_{\Theta_0} L(\theta \mid \mathbf{Y}, \mathbf{X}) / \sup_{\Theta} L(\theta \mid \mathbf{Y}, \mathbf{X})]$, where Θ_0 and Θ are the restricted and unrestricted parameter spaces. Note that the threshold value to reject the null hypothesis can't be simply chosen from a χ^2 distribution because of the violation of regularity conditions of asymptotic theory under H_0 (McLachlan, 1987; Thode, Finch, and Mendell, 1988; Feng and McCulloch, 1994). Also, Lander and Botstein (1989) and Zeng (1994) both suggested that the number and size of intervals should be considered in determining the threshold value since multiple tests are performed in mapping. The hypotheses are usually tested at every position of an interval and for all intervals of the genome to produce a continuous LRT statistic profile. At every position, the position parameter p can be predetermined and only a, d, β, and σ^2 are involved in estimation and testing. If the tests are significant in a chromosome region, the position with the largest LRT statistic is inferred to be the estimate of the QTL position p, and the MLEs at this position are the estimates of a, d, β, and σ^2.

7. General Formulas for Obtaining the MLEs and the Asymptotic Variance–Covariance Matrix

The EM algorithm has been used to obtain MLEs in several studies of QTL mapping analyses (Lander and Botstein, 1989; Carbonell et al., 1992; Jansen, 1992; Zeng, 1994). In these studies, the derivations of MLEs using the EM algorithm vary with the genetic models specified and the populations considered. Also, these studies do not provide the asymptotic variance–covariance matrix for the estimates. In this section, general formulas for the MLEs and the asymptotic variance–covariance matrix of QTL positions and effects are described using an F_2 population as an example. The general formulas, which are based on a genetic design matrix $\mathbf{D}$ and a conditional probability of QTL genotype matrix $\mathbf{Q}$, can apply to different genetic models, experimental populations, and using multiple marker intervals in QTL mapping.

7.1 *General Formulas for MLEs*

The normal mixture model in equation (2) can be treated as an incomplete-data problem (Little and Rubin, 1987) because the genotypes of QTL are unknown. Let

$$g_j(x_j^*, z_j^*) = \begin{cases} p_{j1} & \text{if } x_j^* = 1 \text{ and } z_j^* = -\frac{1}{2} \\ p_{j2} & \text{if } x_j^* = 0 \text{ and } z_j^* = \frac{1}{2} \\ p_{j3} & \text{if } x_j^* = -1 \text{ and } z_j^* = -\frac{1}{2} \end{cases} \tag{4}$$

be the distribution of QTL genotype specified by x_j^* and z_j^*. We treat the unobserved QTL genotypes (x_j^* and z_j^*) as missing data, denoted by $y_{(mis,j)}$, and treat trait (y_j) and selected markers and explanatory variables (X_j) as observed data, denoted by $y_{(obs,j)}$. Then, the combination of $y_{(obs,j)}$ and $y_{(mis,j)}$ is the complete data, denoted by $y_{(com,j)}$. In this setting, we can apply the

EM algorithm to obtain the MLEs. The conditional distribution of observed data, given missing data, can be considered as an independent sample from a population such that

$$y_j \mid (\theta, X_j, x_j^*, z_j^*) \sim \mathrm{N}(x_j^* a + z_j^* d + X_j \beta, \sigma^2).$$

The complete-data density function is thus

$$f(y_{(com,j)} \mid \theta) = f(y_{(obs,j)} \mid \theta, X_j, x_j^*, z_j^*)\ g(x_j^*, z_j^*),$$

the product of densities of conditional observed data and missing data.

At a given position, p can be determined, and the EM algorithm is used for obtaining the MLEs of a, d, β, and σ^2. By definition, the iteration of the $(t+1)$ EM-step is as follows:

E-step: Compute the conditional expected complete-data log-likelihood with respect to the conditional distribution of $\mathbf{Y}_{mis}$ given $\mathbf{Y}_{obs}$ and the current estimated parameter value $\theta^{(t)}$,

$$\begin{aligned} Q(\theta \mid \theta^{(t)}) &= \int \log L(\theta \mid \mathbf{Y}_{com}) f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta = \theta^{(t)})\ d\mathbf{Y}_{mis} \\ &= \int \log \left[\prod_{j=1}^{n} \phi\left(\frac{y_j - \mu_j}{\sigma}\right) g_j(x_j^*, z_j^*) \right] \times f\left(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta = \theta^{(t)}\right) d\mathbf{Y}_{mis}. \end{aligned}$$

Because each observation is independent, the joint probability of n observations $f(\mathbf{Y}_{mis} \mid \mathbf{Y}_{obs}, \theta = \theta^{(t)})$ is the product of n individual probabilities $f(y_{(mis,j)} \mid y_{(obs,j)}, \theta = \theta^{(t)})$, $j = 1, \ldots, n$. By Fubini's Theorem (Ash, 1972), we obtain

$$\begin{aligned} Q\left(\theta \mid \theta^{(t)}\right) &= \sum_{j=1}^{n} \int \log \left[\phi\left(\frac{y_j - \mu_j}{\sigma}\right) g_j(x_j^*, z_j^*) \right] \times f\left(y_{(mis,j)} \mid y_{(obs,j)}, \theta = \theta^{(t)}\right)\ dy_{(mis,j)} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3} \log \left[\phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right)\ p_{ji} \right] \times \frac{p_{ji} \phi\left(\frac{y_j - \mu_{ji}{}^{(t)}}{\sigma^{(t)}}\right)}{\sum_{i=1}^{3} p_{ji} \phi\left(\frac{y_j - \mu_{ji}{}^{(t)}}{\sigma^{(t)}}\right)} \\ &= \sum_{j=1}^{n} \sum_{i=1}^{3} \log \left[\phi\left(\frac{y_j - \mu_{ji}}{\sigma}\right)\ p_{ji} \right] \times \pi_{ji}{}^{(t)}. \end{aligned}$$

A simple application of Bayes' rule on $f(y_{(mis,j)} \mid y_{(obs,j)}, \theta = \theta^{(t)})$ leads to $\pi_{ji} = [p_{ji}\phi((y_j - \mu_{ji})/\sigma)] \ / \ [\Sigma_{i=1}^{3} p_{ji}\phi((y_j - \mu_{ji})/\sigma)]$, which is the posterior probability of the QTL genotype. Equivalently, we can view this as updating the π_{ji} in this step.

M-step: Find $\theta^{(t+1)}$ to maximize the conditional expected log-likelihood $Q(\theta \mid \theta^{(t)})$: By taking the derivatives of $Q(\theta \mid \theta^{(t)})$ with respect to each parameter, the solutions of parameters in closed form are as follows. For a and d,

$$\begin{aligned} a^{(t+1)} &= \frac{\sum_{j=1}^{n} \left[\left(\pi_{j1}^{(t)} - \pi_{j3}^{(t)}\right) \left(y_j - X_j \beta^{(t)}\right) - \frac{1}{2} \left(\pi_{j3}^{(t)} - \pi_{j1}^{(t)}\right) d^{(t)} \right]}{\sum_{j=1}^{n} \left(\pi_{j1}^{(t)} + \pi_{j3}^{(t)}\right)} \\ &= \frac{(\mathbf{Y} - \mathbf{X}\beta^{(t)})' \mathbf{\Pi}^{(t)} D_1 - \mathbf{1}' \mathbf{\Pi}^{(t)} (D_1 \# D_2) d^{(t)}}{\mathbf{1}' \mathbf{\Pi}^{(t)} (D_1 \# D_1)} \end{aligned} \tag{5}$$

$$\begin{aligned} d^{(t+1)} &= \frac{\sum_{j=1}^{n} \frac{1}{2} \left[\left(-\pi_{j1}^{(t)} + \pi_{j2}^{(t)} - \pi_{j3}^{(t)}\right) \left(y_j - X_j \beta^{(t)}\right) - \left(\pi_{j3}^{(t)} - \pi_{j1}^{(t)}\right) a^{(t+1)} \right]}{\frac{1}{4} \sum_{j=1}^{n} \left(\pi_{j1}^{(t)} + \pi_{j2}^{(t)} + \pi_{j3}^{(t)}\right)} \\ &= \frac{(\mathbf{Y} - \mathbf{X}\beta^{(t)})' \mathbf{\Pi}^{(t)} D_2 - \mathbf{1}' \mathbf{\Pi}^{(t)} (D_1 \# D_2) a^{(t+1)}}{\mathbf{1}' \mathbf{\Pi}^{(t)} (D_2 \# D_2)}, \end{aligned} \tag{6}$$

where # denotes Hadamard product, which is the element-by-element product of corresponding elements of two same-order matrices, $\mathbf{\Pi} = \{\pi_{ji}\}_{n \times 3}$. There is a rule in formulating the solutions. In maximizing $Q(\theta \mid \theta^{(t)})$ with respect to a, the corresponding vector D_1 in the genetic design matrix $\mathbf{D}$ plays a central role in the formulation. Likewise, D_2 plays the same role in the formulation for the genetic parameter d. If more genetic parameters are involved in the model, their corresponding

vectors in the genetic design matrix would play a similar role in the formulation. Thus, for simplicity, we can write equations (5) and (6) as

$$\mathbf{E}^{(t+1)} = \mathbf{r}^{(t)} - \mathbf{M}^{(t)}\mathbf{E}^{(t)}, \tag{7}$$

where

$$\mathbf{r} = \begin{bmatrix} \frac{(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{\Pi}D_1}{\mathbf{1}'\mathbf{\Pi}(D_1\#D_1)} \\ \frac{(\mathbf{Y}-\mathbf{X}\beta)'\mathbf{\Pi}D_2}{\mathbf{1}'\mathbf{\Pi}(D_2\#D_2)} \end{bmatrix} \quad \text{and} \quad \mathbf{M} = \begin{bmatrix} 0 & \frac{\mathbf{1}'\mathbf{\Pi}(D_1\#D_2)}{\mathbf{1}'\mathbf{\Pi}(D_1\#D_1)} \\ \frac{\mathbf{1}'\mathbf{\Pi}(D_2\#D_1)}{\mathbf{1}'\mathbf{\Pi}(D_2\#D_2)} & 0 \end{bmatrix}.$$

Note that $\mathbf{M}$ is not a symmetric matrix. For β and σ^2,

$$\beta^{(t+1)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'[\mathbf{Y} - \mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)}] \tag{8}$$

$$\sigma^{2(t+1)} = \frac{1}{n}[(\mathbf{Y} - \mathbf{X}\beta^{(t+1)})'(\mathbf{Y} - \mathbf{X}\beta^{(t+1)}) - 2(\mathbf{Y} - \mathbf{X}\beta^{(t+1)})'\mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)} + \mathbf{E}'^{(t+1)}\mathbf{V}^{(t)}\mathbf{E}^{(t+1)}], \tag{9}$$

where

$$\mathbf{V} = \begin{bmatrix} \mathbf{1}'\mathbf{\Pi}(D_1\#D_1) & \mathbf{1}'\mathbf{\Pi}(D_1\#D_2) \\ \mathbf{1}'\mathbf{\Pi}(D_2\#D_1) & \mathbf{1}'\mathbf{\Pi}(D_2\#D_2) \end{bmatrix}.$$

Note that $\mathbf{V}$ is a symmetric matrix. The E and M steps are iterated until a convergent criterion is satisfied. The converged values of a, d, β, and σ^2 are the MLEs.

To see how to use these general formulas for other genetic models and populations, we take an epistasis model for two unlinked QTLs in a backcross population as an example. The proposed genetic model is

$$\begin{bmatrix} G_{11} \\ G_{10} \\ G_{01} \\ G_{00} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1/2 & 1/2 & 1/4 \\ 1/2 & -1/2 & -1/4 \\ -1/2 & 1/2 & -1/4 \\ -1/2 & -1/2 & 1/4 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ i \end{bmatrix} = \mathbf{1}_{4\times1}\mu + \mathbf{D}_{4\times3}\,\mathbf{E},$$

where G_{11}, G_{10}, G_{01}, and G_{00} are the genotypic values of genotypes $AABB$, $AABb$, $AaBB$, and $AaBb$. The genetic parameters a_1, a_2, and i are the marginal effects of QTL A, B, and the epistatic effect of A and B. The genetic design matrix $\mathbf{D}$ with dimensions 4×3 characterized that the likelihood is a mixture of 4 normals and has 3 genetic parameters (excluding μ) to estimate. Accordingly, $\mathbf{\Pi}$ matrix is an $n \times 4$ matrix.

$$\mathbf{r} = \left\{ \frac{(\mathbf{Y} - \mathbf{X}\beta)'\mathbf{\Pi}\, D_i}{\mathbf{1}'\mathbf{\Pi}(D_i\#D_i)} \right\}_{3\times1} \quad \text{and} \quad \mathbf{M} = \left\{ \frac{\mathbf{1}'\mathbf{\Pi}(D_i\#D_j)}{\mathbf{1}'\mathbf{\Pi}(D_i\#D_i)} \times \delta\ (i \neq j) \right\}_{3\times3},$$

where δ is an indicator variable. $\mathbf{V} = \{\ \mathbf{1}'\mathbf{\Pi}(D_i\#D_j)\ \}_{3\times3}$. In deriving the MLEs, in the E-step the posterior probabilities π_{ji}'s of the four QTL genotypes are updated, and in the M-step the equations (7) to (9) are applied to maximize $Q(\theta \mid \theta^{(t)})$ with $\mathbf{D}$, $\mathbf{\Pi}$, $\mathbf{M}$, $\mathbf{r}$, and $\mathbf{V}$ specified above. When more marker intervals are used in analysis, it is easy to expand the domensions of $\mathbf{\Pi}$, $\mathbf{M}$, $\mathbf{r}$, and $\mathbf{V}$ according to the dimensions of $\mathbf{D}$ for the use of general formulas.

7.2 *General Formulas for the Asymptotic Variance–Covariance Matrix*

The above EM algorithm gives only point estimates of parameters. Additional steps are needed to find the variance–covariance matrix. When the EM algorithm is used, Louis (1982) derived a procedure for obtaining the asymptotic variance–covariance matrix. Meng and Rubin (1991) suggested using the SEM (supplemented EM) algorithm. Here, we adopt Louis's method to derive the asymptotic variance–covariance matrix.

Obtaining the asymptotic variance–covariance matrix is equivalent to extracting the observed information (I_{obs}) in the incomplete-data problem. The complete-data likelihood model in this problem can be regarded as a two-stage hierachical model. First the values of random variables (x_j^*, z_j^*) are sampled by a trinomial experiment to decide QTL genotype, and then a normal variate for that QTL genotype is generated. The random variables (x_j^*, z_j^*) of individual j are $(1, -1/2), (0, 1/2)$, or $(-1, -1/2)$ for QTL genotype QQ, Qq, or qq with probability p_{j1}, p_{j2}, or p_{j3}, respectively. Therefore, the complete-data likelihood is

$$\lambda(\mathbf{Y}_{com} \mid p, a, d, \sigma^2, \beta) = \prod_{j=1}^{n} \left[p_{j1}\phi\left(\frac{y_j - \mu_{j1}}{\sigma}\right)^{-\frac{1}{2}(x_j^*+1)(z_j^*-\frac{1}{2})} p_{j2}\phi\left(\frac{y_j - \mu_{j2}}{\sigma}\right)^{(x_j^*+1)(z_j^*+\frac{1}{2})} \times p_{j3}\phi\left(\frac{y_j - \mu_{j3}}{\sigma}\right)^{\frac{1}{2}(x_j^*-1)(z_j^*-\frac{1}{2})} \right].$$

Louis (1982) showed that the observed information I_{obs} is the difference of complete I_{oc} and missing I_{om} information. That is, $I_{obs}(\theta^* \mid \mathbf{Y}_{obs}) = I_{oc} - I_{om}$, where

$$I_{oc} = \sum_{j=1}^{n} \mathrm{E}\left[-\frac{\partial^2 \log \lambda(y_{(com,j)} \mid \theta)}{\partial \theta^2} \;\middle|\; y_{(obs,j)}, \theta \right]_{\theta=\theta^*}$$

and

$$I_{om} = \sum_{j=1}^{n} \mathrm{E}\left\{ \left[\frac{\partial \log \lambda(y_{(com,j)} \mid \theta)}{\partial \theta}\right] \left[\frac{\partial \log \lambda(y_{(com,j))} \mid \theta)}{\partial \theta}\right]' \mid y_{(obs,j)}, \theta \right\}_{\theta=\theta^*} + \sum_{i \neq j}^{n} \left[\mathrm{E}\left[\frac{\partial \log \lambda(y_{(com,i)} \mid \theta)}{\partial \theta} \mid y_{(obs,i)}, \theta\right]_{\theta=\theta^*} \mathrm{E}\left[\frac{\partial \log \lambda(y_{(com,j)} \mid \theta)}{\partial \theta} \mid y_{(obs,j)}, \theta\right]'_{\theta=\theta^*} \right].$$

θ^* denotes the MLE of θ.

7.2.1 *The Complete-Data Information Matrix.* The complete-data information matrix I_{oc} is obtained as

$$I_{oc} = \frac{1}{\sigma^2} \begin{bmatrix} -\sigma^2 \mathbf{1}'_{n\times 1}(\mathbf{P}^{(2)} \# \mathbf{\Pi})\mathbf{1}_{1\times 3} & 0 & 0 & 0 & 0 \\ 0 & \mathbf{1}'_{n\times 1}\mathbf{\Pi}(D_1 \# D_1) & \mathbf{1}'_{n\times 1}\mathbf{\Pi}(D_1 \# D_2) & \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{\Pi}\#\mathbf{T})D_1 & D_1'\mathbf{\Pi}'\mathbf{X} \\ 0 & \mathbf{1}'_{n\times 1}\mathbf{\Pi}(D_2 \# D_1) & \mathbf{1}'_{n\times 1}\mathbf{\Pi}(D_2 \# D_2) & \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{\Pi}\#\mathbf{T})D_2 & D_2'\mathbf{\Pi}'\mathbf{X} \\ 0 & \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{\Pi}\#\mathbf{T})D_1 & \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{\Pi}\#\mathbf{T})D_2 & \frac{n}{2\sigma^2} & \frac{1}{\sigma^2}\mathbf{1}'_{3\times 1}(\mathbf{\Pi}\#\mathbf{T})'\mathbf{X} \\ 0 & \mathbf{X}'\mathbf{\Pi}D_1 & \mathbf{X}'\mathbf{\Pi}D_2 & \frac{1}{\sigma^2}\mathbf{X}'(\mathbf{\Pi}\#\mathbf{T})\mathbf{1}_{3\times 1} & \mathbf{X}'\mathbf{X} \end{bmatrix}.$$

The derivations of some elements in I_{oc} are shown below.

$$I_{oc\ 11} = -\sum_{j=1}^{n} \mathrm{E}\left[\frac{\partial^2 \log \lambda_j}{\partial p^2} \mid y_{(obs,j)}, \theta\right] = -\sum_{j=1}^{n}\sum_{i=1}^{3} p_{ji}^{(2)}\pi_{ji} = -\mathbf{1}'_{n\times 1}(\mathbf{P}^{(2)} \# \mathbf{\Pi})\mathbf{1}_{3\times 1},$$

where $p_{ji}^{(2)}$ denotes the second derivative of log mixing proportion p_{ji} with respect to p. If p_{ji} equals 0, we assign the corresponding $p_{ji}^{(2)}$ to 0. For an individual j, $p_{ji}^{(2)}$ can be obtained from the corresponding row element of the matrix $\mathbf{Q}^{(2)}$.

$$\mathbf{Q}^{(2)} = \begin{bmatrix} 0 & 0 & 0 \\ -\frac{1}{(1-p)^2} & -\frac{1}{p^2} & 0 \\ -\frac{2}{(1-p)^2} & \frac{-2p^2+2p-1}{p^2(1-p)^2} & -\frac{2}{p^2} \\ -\frac{1}{p^2} & -\frac{1}{(1-p)^2} & 0 \\ \frac{-2p^2+2p-1}{p^2(1-p)^2} & \frac{4c[1+c(-2p^2+2p-1)]}{[1-2cp(1-p)]^2} & \frac{-2p^2+2p-1}{p^2(1-p)^2} \\ 0 & -\frac{1}{(1-p)^2} & -\frac{1}{p^2} \\ -\frac{2}{p^2} & \frac{-2p^2+2p-1}{p^2(1-p)^2} & -\frac{2}{(1-p)^2} \\ 0 & -\frac{1}{p^2} & -\frac{1}{(1-p)^2} \\ 0 & 0 & 0 \end{bmatrix}$$

with elements obtained by taking the second derivative of corresponding log elements in matrix $\mathbf{Q}$ with respect to p. Therefore, $\mathbf{Q}^{(2)}$ follows $\mathbf{Q}$.

$$I_{oc\ 24} = -\sum_{j=1}^{n} \mathrm{E}\left[\frac{\partial^2 \log \lambda_j}{\partial a \partial \sigma^2} \mid y_{(obs,j)}, \theta\right]$$
$$= \frac{1}{\sigma^4} \sum_{j=1}^{n} \left[(y_j - \mu_{j1})\pi_{j1} - (y_j - \mu_{j3})\pi_{j3}\right] = \frac{1}{\sigma^4}\ \mathbf{1}'_{n\times 1}(\mathbf{T}\#\mathbf{\Pi})D_1,$$

where $\mathbf{T} = \{t_{ji}\}_{n\times 3}$. $t_{j1} = y_j - \mu_{j1}$, $t_{j2} = y_j - \mu_{j2}$ and $t_{j3} = y_j - \mu_{j3}$. To relate $\mathbf{T}$ to the genetic design matrix $\mathbf{D}$, we express $\mathbf{T} = (\mathbf{Y} - \mathbf{X}\beta)\otimes\mathbf{1}_{1\times 3} - \mathbf{1}_{n\times 1}\otimes(\mathbf{DE})'$, where $\otimes$ denotes the Kronecker product. Similarly,

$$I_{oc\ 25} = -\sum_{j=1}^{n} \mathrm{E}\left[\frac{\partial^2 \log \lambda_j}{\partial a \partial \beta'} \mid y_{(obs,j)}, \theta\right] = \frac{1}{\sigma^2}\sum_{j=1}^{n} X_j(\pi_{j1} - \pi_{j3}) = \frac{1}{\sigma^2}\ D'_1\mathbf{\Pi}'\mathbf{X}$$

$$I_{oc\ 45} = -\sum_{j=1}^{n} \mathrm{E}\left[\frac{\partial^2 \log \lambda_j}{\partial \sigma^2 \partial \beta'} \mid y_{(obs,j)}, \theta\right] = \frac{1}{\sigma^4}\sum_{j=1}^{n}\sum_{k=1}^{3} X_j[(y_j - \mu_{jk})\pi_{jk}] = \frac{1}{\sigma^4}\ \left[\mathbf{1}'_{3\times 1}(\mathbf{\Pi}\#\mathbf{T})'\mathbf{X}\right]$$

$$I_{oc\ 22} = -\sum_{j=1}^{n} \mathrm{E}\left[\frac{\partial^2 \log \lambda_j}{\partial a^2} \mid y_{(obs,j)}, \theta\right] = \frac{1}{\sigma^2}\sum_{j=1}^{n}[\pi_{j1} + \pi_{j3}] = \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}\mathbf{\Pi}(D_1\#D_1)$$

The other elements of the complete information matrix I_{oc} can be obtained in the same way. If more intervals are involved in the model, it is straightforward to expand the complete information matrix by incorporating corresponding vectors in the genetic design matrix $\mathbf{D}$ and the conditional probability matrix $\mathbf{Q}$ in the formulation.

7.2.2 *The Missing Information Matrix.* The leading element in the missing information matrix is

$$I_{om\ 11} = \sum_{j=1}^{n} \mathrm{E}\left[\left(\frac{\partial \log \lambda_j}{\partial p}\right)\left(\frac{\partial \log \lambda_j}{\partial p}\right)' \mid y_{(obs,j)}, \theta\right]_{\theta=\theta^*}$$
$$+ \sum_{i\neq j}^{n}\left[\mathrm{E}\left[\frac{\partial \log \lambda_i}{\partial p} \mid y_{(obs,i)}, \theta\right]_{\theta=\theta^*} \mathrm{E}\left[\frac{\partial \log \lambda_j}{\partial p} \mid y_{(obs,j)}, \theta\right]'_{\theta=\theta^*}\right]$$
$$= \sum_{j=1}^{n}\sum_{k=1}^{3}\left[p_{jk}^{(1)}\right]^2 \pi_{jk} + \sum_{i\neq j}^{n}\left[\left[\sum_{k=1}^{3} p_{ik}^{(1)}\pi_{ik}\right]\left[\sum_{k=1}^{3} p_{jk}^{(1)}\pi_{jk}\right]\right]$$
$$= \mathbf{1}'_{n\times 1}(\mathbf{P}^{(1)}\#\mathbf{P}^{(1)}\#\mathbf{\Pi})\ \mathbf{1}_{3\times 1} + \sum_{i\neq j}^{n}\left[\left(P_i^{(1)}\#\Pi_i\right)\mathbf{1}_{3\times 1}\right]\left[\left(P_j^{(1)}\#\Pi_j\right)\mathbf{1}_{3\times 1}\right],$$

where $\mathbf{P}^{(1)} = \{p_{ji}^{(1)}\}_{n\times 3}$. Row vector $P_j^{(1)} = (p_{j1}^{(1)}\ \ p_{j2}^{(1)}\ \ p_{j3}^{(1)})$, where $p_{ji}^{(1)}$ denotes the first derivative of $\log p_{ji}$ with respect to p. For those individuals with p_{ji} being 0, we also assign the corresponding $p_{ji}^{(1)}$ to 0. Again, for a given individual j, $p_{ji}^{(1)}$ can be found from the corresponding row element of matrix $\mathbf{Q}^{(1)}$,

$$\mathbf{Q}^{(1)} = \begin{bmatrix} 0 & 0 & 0 \\ -\frac{1}{1-p} & \frac{1}{p} & 0 \\ -\frac{2}{(1-p)} & \frac{1-2p}{p(1-p)} & \frac{2}{p} \\ \frac{1}{p} & -\frac{1}{1-p} & 0 \\ \frac{1-2p}{p(1-p)} & \frac{-2c(1-2p)}{1-2cp(1-p)} & \frac{1-2p}{p(1-p)} \\ 0 & -\frac{1}{1-p} & \frac{1}{p} \\ \frac{2}{p} & \frac{1-2p}{p(1-p)} & -\frac{2}{1-p} \\ 0 & \frac{1}{p} & -\frac{1}{1-p} \\ 0 & 0 & 0 \end{bmatrix}$$

with elements obtained by taking the first derivative of corresponding log elements in matrix $\mathbf{Q}$ with respect to p. Likewise, the other elements in the missing information matrix I_{om} can be derived and expressed in a similar way,

$$I_{om\ 12} = \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{P}^{(1)}\#\mathbf{T}\#\mathbf{\Pi})D_1 + \frac{1}{\sigma^2}\sum_{i\neq j}^{n}\left[\left(P_i^{(1)}\#\Pi_i\right)\mathbf{1}_{3\times 1}\right][(T_j\#\Pi_j)D_1].$$

It can be seen that there is a general trend in the derivation. When the derivative involves p, the formula contains $\mathbf{P}^{(1)}$ and $\mathbf{1}_{3\times 1}$. When the derivative involves a, the formula contains $\mathbf{T}$ and the corresponding column D_1 in the genetic design matrix. Following this rule, we can easily derive the other elements. For example,

$$I_{om\ 13} = \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{P}^{(1)}\#\mathbf{T}\#\mathbf{\Pi})D_2 + \frac{1}{\sigma^2}\sum_{i\neq j}^{n}\left[\left(P_i^{(1)}\#\Pi_i\right)\mathbf{1}_{3\times 1}\right][(T_j\#\Pi_j)D_2].$$

Likewise,

$$I_{om\ 14} = \frac{1}{\sigma^2}\mathbf{1}'_{n\times 1}(\mathbf{P}^{(1)}\#\mathbf{S}\#\mathbf{\Pi})\mathbf{1}_{3\times 1} + \frac{1}{\sigma^2}\sum_{i\neq j}^{n}\left[\left(P_i^{(1)}\#\Pi_i\right)\mathbf{1}_{3\times 1}\right][(S_j\#\Pi_j)\mathbf{1}_{3\times 1}],$$

where $\mathbf{S} = \{s_{ji}\}_{n\times 3}$ with $s_{ji} = t_{ji}^2/(2\sigma^2) - 1/2$ and $S_j = (s_{j1}\ s_{j2}\ s_{j3})$. Thus, when the derivative involves σ^2, the formula contains $\mathbf{S}$ and $\mathbf{1}_{3\times 1}$. Similarly,

$$I_{om\ 15} = \frac{1}{\sigma^2}\ \mathbf{1}'_{3\times 1}(\mathbf{T}\#\mathbf{P}^{(1)}\#\mathbf{\Pi})'\mathbf{X} + \frac{1}{\sigma^2}\sum_{i\neq j}^{n}[\mathbf{1}'_{3\times 1}(T_i\#\mathbf{\Pi}_i)'X_i]\left[(P_j^{(1)}\#\Pi_j)\mathbf{1}_{3\times 1}\right];$$

that is, when the derivative involves β, the formulas contain $(X_j'X_j)$ and $\mathbf{1}_{3\times 1}$. The other elements can be derived and expressed in the similar way.

Both the complete and missing information matrices are characterized by the genetic design matrix $\mathbf{D}$ and the conditional probability matrix $\mathbf{Q}$. As $\mathbf{T}$, $\mathbf{S}$, $\mathbf{\Pi}$, $\mathbf{P}^{(1)}$, and $\mathbf{P}^{(2)}$ follow $\mathbf{D}$ and $\mathbf{Q}$, the formulas for the information matrices can thus be readily used for other different genetic models, experimental designs, and data structures. The observed information can then be obtained by subtracting the missing information from the complete information. The inverse of the observed information matrix gives the asymptotic variance–covariance matrix of QTL position and effects.

8. Simulation Result

Simulations were performed to study the properties of the asymptotic variances of the QTL position and effects using a backcross population. We propose the following genetic model,

$$G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}\mu + \begin{bmatrix} 1/2 \\ -1/2 \end{bmatrix}[a] = \mathbf{1}_{2\times 1}\mu + \mathbf{DE}, \tag{10}$$

for QTL mapping, where G_1 and G_2 denote the genotypic values of the genotypes QQ and Qq. The genetic design matrix $\mathbf{D}$ in this case is a vector and $\mathbf{E}$ is a scalar. The conditional probability matrix $\mathbf{Q}$ is given in Table 1. The MLEs and the asymptotic variance–covarince matrix are obtained using the general formulas.

We assume only one QTL on a chromosome with 16 equally spaced markers. Two marker interval sizes, 5 and 15 cM, are considered. In each case, the QTL position is located either in the middle or at the boundary of the fourth interval. Sample sizes are 200 and 500. Two heritabilities (h^2), 0.1 and 0.2, are considered in each set of simulations. A trait with heritability ($h^2 = 0.1$) means that 10% of the trait variation is controlled by QTL and the remaining 90% is subject to the environment (random error). The genetic parameters μ and a are 0 and 0.7705. IM is used to search at every 1 cM position of the chromosome for the QTL. The threshold value to reject the null hypothesis is 6.9 (about 1.5 in lod score) (Lander and Botstein, 1989). If the estimated position is coincident with the marker, the asymptotic variance of position p is difficult to obtain and only those of μ, a, and β are provided by our approach. By comparing the MSE of the estimates between the results of two different QTL positions, no uniform conclusion on the performance of estimates can be drawn. Also, the pattern of results in the two different positions are similar. Therefore, only the results of the QTL located in the middle of the interval (Tables 3 and 4) are shown for illustration.

Table 3
Simulation result of 100 *replicates with* $h^2 = 0.1$

	Parameter	Mean of estimates	Emp. SD	MSE	Mean of est. asym. SD	SD of est. asym. SD
			size = 15 cM n = 500			
Position	52.5	52.13	5.1574	2646	3.6712	2.1270
μ	0	0.0084	0.1109	1.2233	0.1175	0.0598
a	0.7705	0.7567	0.1109	1.2340	0.0988	0.0089
σ^2	5.343	5.3198	0.3398	11.23	0.3431	0.0209
			size = 15 cM n = 200			
Position	52.5	51.98[a]	11.1591[a]	11856[a]	4.9589[a]	1.3191[a]
μ	0	0.0303[a]	0.1706[a]	2.8530[a]	0.1691[a]	0.0365[a]
a	0.7705	0.7983[a]	0.1583[a]	2.4555[a]	0.1521[a]	0.0095[a]
σ^2	5.343	5.1469[a]	0.4901[a]	26.058[a]	0.5246[a]	0.0495[a]
			size = 5 cM n = 500			
Position	17.5	16.72	3.1045	1015	1.9290	0.5433
μ	0	0.0083	0.1092	1.1876	0.1156	0.0766
a	0.7705	0.7659	0.1117	1.2374	0.0945	0.0030
σ^2	5.343	5.3144	0.3298	10.85	0.3377	0.0206
			size = 5 cM n = 200			
Position	17.5	18.26[b]	5.9000[b]	3432.5[b]	2.6379[b]	0.8488[b]
μ	0	0.0289[b]	0.1684[b]	2.8328[b]	0.1697[b]	0.0954[b]
a	0.7705	0.8037[b]	0.1672[b]	2.8209[b]	0.1465[b]	0.0077[b]
σ^2	5.343	5.1460[b]	0.4909[b]	27.178[b]	0.5177[b]	0.0492[b]

[a] Based on 96 significant replicates.
[b] Based on 98 significant replicates.

We concentrate on the performance of the estimates of the asymptotic standard deviations (ASD). For the cases with $h^2 = 0.1$, the means of the estimates of the ASD for the QTL position significantly underestimate the empirical standard deviations (SD) in all cases (Table 3). For the case with $h^2 = 0.2$, the means of the ASD estimates are very close to the empirical SD. The explanation for the underestimation in $h^2 = 0.1$ cases is that with $h^2 = 0.1$, there is a relatively high chance that the QTL is identified in the wrong interval in the search for the whole chromosome. The percentages of the QTL being localized in the correct interval are 86%, 75%, 69%, and 42% for the cases with interval size 15 cM and $n = 500$, 15 cM and $n = 200$, 5 cM and $n = 500$, and 5 cM and $n = 200$, respectively. Clearly, several replicates in each case have identified QTL in the wrong intervals. This causes a larger empirical SD for the position estimate. However, the estimates of ASD are still calculated based on a 15 (or 5) cM marker interval. Therefore, the mean of the estimated ASD underestimates the SD. When $h^2 = 0.2$ (Table 4), the percentages of the QTL being localized in the correct interval increase to 98%, 94%, 89%, and 69% for the four cases, respectively. Fewer replicates than those for $h^2 = 0.1$ have QTL identified in the wrong interval. The estimated ASD are calculated based on the correct interval so that it gives a good estimation. Therefore, to make the ASD of this approach reliable in QTL mapping, it is very important to localize the QTL in the correct interval. When the QTL is localized in the wrong interval, ASD is underestimated. This is the limitation of this approach.

Among the ASD of the mean μ, QTL effect a, and environmental error σ^2, the ASD of a estimates its empirical SD poorly. In Tables 3 and 4, the range of mean ASD $\pm$ 2 SD fails to cover the empirical SD in several cases. We think that it is because the convergence of the ASD of a is slower. Although the ASD of a underestimates the empirical SD, it is still close to the empirical SD and estimates the sampling variance reasonably well.

9. Discussion

In this paper, we use a normal mixture model to model the relationship between a quantitative trait and the unobserved QTLs using genetic markers, then we present general formulas for deriving the MLEs and the asymptotic variance–covariance matrix of QTL positions and effects

Table 4
Simulation result of 100 *replicates with* $h^2 = 0.2$

	Parameter	Mean of estimates	Emp. SD	MSE	Mean of est. asym. SD	SD of est. asym. SD
		size = 15 cM n = 500				
Position	52.5	52.52	2.5956	666.75	2.6892	0.7498
μ	0	−0.0057	0.0746	0.5547	0.0934	0.0948
a	0.7705	0.7592	0.0746	0.5634	0.0577	0.0082
σ^2	2.3747	2.3668	0.1514	2.2749	0.1622	0.0164
		size = 15 cM n = 200				
Position	52.5	52.09	3.9749	1581	3.9681	0.8074
μ	0	0.0187	0.1182	1.4184	0.1219	0.0189
a	0.7705	0.7703	0.1177	1.3713	0.0853	0.0067
σ^2	2.3747	2.2991	0.2245	5.5564	0.2424	0.0240
		size = 5 cM n = 500				
Position	17.5	17.22	1.5412	234	1.4391	0.3094
μ	0	0.0053	0.0732	0.5336	0.0777	0.0206
a	0.7705	0.7656	0.0759	0.5734	0.0530	0.0025
σ^2	2.3747	2.3631	0.1465	2.1393	0.1510	0.0092
		size = 5 cM n = 200				
Position	17.5	17.73	2.7000	727	2.1410	0.7824
μ	0	0.0183	0.1145	1.3317	0.1235	0.0283
a	0.7705	0.7806	0.1157	1.3369	0.0818	0.0062
σ^2	2.3747	2.2933	0.2191	5.4104	0.2323	0.0236

of the model when the EM algorithm is used in derivation. The general formulas are based on a genetic design matrix $\mathbf{D}$ and a conditional probability matrix $\mathbf{Q}$, where $\mathbf{D}$ characterizes the genetic effects and $\mathbf{Q}$ contains the information on QTL positions. The formulas are general because they apply to different genetic models, experimental designs (e.g., backcross and F_2 populations), and an arbitrary number of marker intervals. The general formulas enable us to consider multiple marker intervals simultaneously in QTL mapping. There are several uses for multiple interval mapping. It can be used for a multidimensional search for multiple QTLs or for searching one interval for a QTL by conditioning on other identified QTLs at given positions to increase the precision and power of mapping, for analyzing QTL epistasis, and for estimating the heritability.

If k intervals are considered jointly in mapping, the dimensions of the genetic design matrix $\mathbf{D}$ augment to $2^k \times k$ for a backcross population and to $3^k \times 2k$ for an F_2 population when epistasis is ignored. Taking a backcross population as an example, if we want to consider three marker intervals (three putative QTLs) simultaneously and use an additive model, the genetic model can be defined as

$$G = \begin{bmatrix} G_{111} \\ G_{110} \\ G_{101} \\ G_{100} \\ G_{011} \\ G_{010} \\ G_{001} \\ G_{000} \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \mu + \begin{bmatrix} 1/2 & 1/2 & 1/2 \\ 1/2 & 1/2 & -1/2 \\ 1/2 & -1/2 & 1/2 \\ 1/2 & -1/2 & -1/2 \\ -1/2 & 1/2 & 1/2 \\ -1/2 & 1/2 & -1/2 \\ -1/2 & -1/2 & 1/2 \\ -1/2 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \mathbf{1}_{8\times 1}\mu + \mathbf{DE}, \tag{11}$$

where G_{111}, G_{110}, G_{101}, G_{100}, G_{011}, G_{010}, G_{001}, and G_{000} denote the genotypic values of QTL genotypes *AABBCC*, *AABBCc*, *AABbCC*, *AABbCc*, *AaBBCC*, *AaBBCc*, *AaBbCC*, and *AaBbCc*, respectively, and a_1, a_2, and a_3 denote the effects of QTLs A, B, and C. The genetic design matrix $\mathbf{D}$ with dimensions 8×3 specifies that the corresponding likelihood is a mixture of 8 normals and has 3 genetic parameters to estimate. To infer the joint conditional probability

matrix **Q** for the three putative QTLs, we use the property that if there is no interference in crossing over, the conditional distributions of the individual putative QTL genotypes given the flanking marker genotypes are independent, irrespective of whether the QTLs are linked or not. This independence property simplifies the inference of **Q** matrix. If epistasis of QTLs B and C is also analyzed, the dimensions of genetic design matrix **D** in equation (11) augment to 8×4. Column 4, which is the product of columns 2 and 3, of the genetic design matrix represents the status of the epistatic parameters of different genotypes. The **Q** matrix is the same as that for the additive model. Kao (1995) applied the general formulas to study digenic epistasis in an F_2 population (the dimensions of the genetic design matrix are 9×8). If higher order of QTL epistasis is considered, the corresponding column vector of D can be generated by the same procedure. By this approach, QTL epistasis can be incorporated in the model easily for analysis. Again, given **D** and **Q** in both cases, the general formulas can be used to obtain the MLEs and the asymptotic variance–covariance matrix for the model. When all the putative QTLs and their possible epistasis are taken into account in the model, the model sum of squares divided by total sum of squares is the estimate of heritability.

It is very important to construct confidence intervals for the estimates of QTL positions and effects. For example, when a particular QTL is to be transferred to a recipient, a confidence interval for the QTL position can give us an idea about how large a chromosome segment around the detected position should be transferred. The asymptotic variances can be used to construct the confidence intervals. For a large sample, the $(1-\alpha)\%$ confidence interval for a position estimate $\hat{p}$ can be approximated by $(\hat{p} - Z_{(1-\alpha/2)}S_{\hat{p}}\ ,\ \hat{p} + Z_{(1-\alpha/2)}S_{\hat{p}})$. Other approaches for constructing the confidence interval of QTL position, such as the LOD support interval (Lander and Botstein, 1989) and bootstrapping, can also be used. By our approach, if the estimated QTL position is right on the marker, there is no position parameter (p) in the model. Its asymptotic variance cannot be provided. In this situation, we might intentionally delete this marker and use the two nearby markers to form a mapping interval. We then perform mapping on this interval and estimate the asymptotic variance. Generally, the asymptotic variance of the QTL position estimate underestimates the sampling variance since there is no guaranty that the QTLs can be localized in the correct intervals in QTL mapping. To localize the QTL in the correct interval, we require reliable experimental designs, mapping analyses and strategies to increase the precision of QTL mapping.

Acknowledgements

The authors are grateful to Dr Nianci Gan and Dr Christopher Basten for their discussion. We are also grateful to the associate editor and the referees for their valuable comments and suggestions. This study was supported in part by grants GM-45344 from the National Institutes of Health and 94-37300-0407 from the U.S. Department of Agriculture, Plant Genome Program.

Résumé

Nous présentons dans ce papier les formules générales permettant de dériver les estimations du maximum de vraisemblance et la matrice de variance–covariance asymptotique des positions et des effets de loci de traits quantitatifs (QTLs) dans un modèle de mélange fini de variables normales, lorsque l'algorithme EM est utilisé pour la localisation des QTLs. Les formules générales sont basées sur les deux matrices **D** et **Q**, **D** étant la matrice du patron génétique, caractérisant les effets génétiques des QTLs, et **Q** la matrice des probabilités conditionnelles des génotypes QTL, étant donné les génotypes avec les marqueurs flanquants, contenant l'information de la position des QTLs. A l'aide des formules générales, il est relativement aisé d'étendre l'analyse de la localisation des QTLs à l'usage d'intervalles de marquage multiples, afin de réaliser simultanément la localisation de QTLs multiples, l'analyse de l'épistasie des QTLs et l'estimation de l'héritabilité des caractéristiques quantitatives. Des simulations ont été réalisées pour évaluer la performance des estimations des variances asymptotiques des effets et des positions des QTLs.

References

Andersson, L., Haley, C. S., Ellegen, H., Knott, S. A., Johansson, M., Andersson, K., Andersson-Ekliund, L., Edfors-Lilja, I., Fredholm, M., Hansson, I., Hakansson, J., and Lundstrom, K. (1994). Genetic mapping of quantitative trait loci for growth and fatness in pigs. *Science* **263,** 1771–1774.

Ash, R. B. (1972). *Real Analysis and Probability.* New York: Academic Press.

Bradshaw, H. D., Jr. and Stettler, R. F. (1995). Molecular genetics of growth and development in *Populus*. IV. Mapping QTLs with large effects on growth, form, and phenology traits in a forest tree. *Genetics* **139,** 964–973.

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

Darvasi, A., Weinreb, A., Minke, V., Weller, J. I., and Soller, M. (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., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society* **39,** 1–38.

Dragani, T. A., Zeng, Z.-B., Canzian, F., Gariboldi, M., Ghilarducci, M. T., Manenti, G., and Pierotti, M. A. (1995). Mapping of body weight loci on mouse chromosome **X**. *Mammalian Genome* **6,** 778–781.

Feng, Z. D. and McCulloch, C. E. (1994). On the likelihood ratio test statistic for the number of components in a normal mixture with unequal variances. *Biometrics* **50,** 1158–1162.

Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait in line crosses using flanking markers. *Heredity* **69,** 315–324.

Horvat, S. and Medrano, J. F. (1995). Interval mapping of *high growth* (*hg*), a major locus that increases weight gain in mice. *Genetics* **139,** 1737–1748.

Jansen, R. C. (1992). A general mixture model for mapping quantitative trait loci by using molecular markers. *Theoretical and Applied Genetics* **85,** 252–260.

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

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.

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

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

Liu, J., Mercer, J. M., Stem, L. F., Gibson, G. C., Zeng, Z.-B., and Laurie, C. C. (1996). Genetic analysis of a morphilogical shape difference in the male genitalia of *Drosophite simulans* and *D. mauritiana*. *Genetics* **142,** 1129–1145.

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of Royal Statistical Society, Series B* **44,** 226–233.

McLachlan, G. J. (1987). On bootstrapping the likelihood ratio test statistic for the number of components in a normal mixture. *Applied Statistics* **36,** 318–324.

Meng, X.-L. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika* **80,** 267–278.

Meng, X.-L. (1994). On the rate of convergence of the ECM algorithm. *The Annals of Statistics* **22,** 326–339.

Meng, X.-L. and Rubin, D. B. (1991). Using EM to obtain asymptotic variance–covariance matrices: The SEM algorithm. *Journal of the American Statistical Association* **86,** 899–909.

Paterson, A. H., Damon, S., Hewitt, J. D., Zamir, D., Rabinowitch, H. D., Lincoln, S. E., Lander, E. S., and Tanksley, S. D. (1991). Mendelian factor underlying quantitative trait in tomato: Comparison across species, generation, and environment. *Genetics* **127,** 181–197.

Stuber, C. W., Lincoln, S. E., Woff, D. W., Helentjaris, T., and Lander, E. S. (1992). Identification of genetic factors contributing to heterosis in a hybrid from two elite maize inbred lines using molecular marker. *Genetics* **132,** 832–839.

Thoday, J. M. (1960). Location of polygenes. *Nature* **191,** 368–370.

Thode, H. C., Jr., Finch, S. J., and Mendell, N. R. (1988). Simulated percentage points for the null distribution of the likelihood ratio test for a mixture of two normal. *Biometrics* **44,** 1195–1201.

Titterington, D. M., Smith, A. F., and Makov, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley and Sons.

Zeng, Z.-B. (1993). Theoretical basis for separation of multiple linked gene effects in mapping of quantitative trait loci. *Proceedings of the National Academy of Science* **90,** 10972–10976.

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

Received May 1996; *revised November* 1996; *accepted November* 1996.