Mapping Quantitative Trait Loci in F_2 Incorporating Phenotypes of F_3 Progeny

Yuan-Ming Zhang and Shizhong Xu[1]

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

Manuscript received October 6, 2003
Accepted for publication December 31, 2003

ABSTRACT

In plants and laboratory animals, QTL mapping is commonly performed using F_2 or BC individuals derived from the cross of two inbred lines. Typical QTL mapping statistics assume that each F_2 individual is genotyped for the markers and phenotyped for the trait. For plant traits with low heritability, it has been suggested to use the average phenotypic values of F_3 progeny derived from selfing F_2 plants in place of the F_2 phenotype itself. All F_3 progeny derived from the same F_2 plant belong to the same $F_{2:3}$ family, denoted by $F_{2:3}$. If the size of each $F_{2:3}$ family (the number of F_3 progeny) is sufficiently large, the average value of the family will represent the genotypic value of the F_2 plant, and thus the power of QTL mapping may be significantly increased. The strategy of using F_2 marker genotypes and F_3 average phenotypes for QTL mapping in plants is quite similar to the daughter design of QTL mapping in dairy cattle. We study the fundamental principle of the plant version of the daughter design and develop a new statistical method to map QTL under this $F_{2:3}$ strategy. We also propose to combine both the F_2 phenotypes and the $F_{2:3}$ average phenotypes to further increase the power of QTL mapping. The statistical method developed in this study differs from published ones in that the new method fully takes advantage of the mixture distribution for $F_{2:3}$ families of heterozygous F_2 plants. Incorporation of this new information has significantly increased the statistical power of QTL detection relative to the classical F_2 design, even if only a single F_3 progeny is collected from each $F_{2:3}$ family. The mixture model is developed on the basis of a single-QTL model and implemented via the EM algorithm. Substantial computer simulation was conducted to demonstrate the improved efficiency of the mixture model. Extension of the mixture model to multiple QTL analysis is developed using a Bayesian approach. The computer program performing the Bayesian analysis of the simulated data is available to users for real data analysis.

IN classical quantitative genetics, if a trait has a low heritability, one can take the family mean as the unit of phenotypic measurement and select the parents with high average performance on the basis of the family mean (Mather and Jinks 1982; Gai *et al.* 2003). The reason is that the family-mean-based heritability can be significantly increased by increasing the number of progeny. This idea has been applied to genetic mapping for low heritability traits in animals by using the daughter design (Weller *et al.* 1990; Bovenhuis and Weller 1994; Mackinnon and Weller 1995; Ron *et al.* 2001), where the phenotypic value of the sire has been replaced by the mean phenotypic value of the daughters. By increasing the number of daughters tested, the mean phenotypic value may represent the true genotypic value of the sire, and thus using the mean value for genetic analysis can reduce the residual error variance. In such a design of experiment, progeny do not have to be typed for markers, leading to substantial cost saving. In addition, some traits, *e.g.*, endosperm traits, can be measured only in tissues controlled by genotypes of progeny (Xu *et al.* 2003). Mapping QTL for such traits requires a special technique to handle the distribution of mean phenotypic values.

Plant geneticists have developed the plant version of the daughter design by replacing the phenotypic value of an F_2 plant by the mean of F_3 progeny, called the $F_{2:3}$ design (Austin and Lee 1996; Fisch *et al.* 1996). One can arbitrarily increase the number of generations from 3 to y, leading to an $F_{2:y}$ design. It is even possible to genotype plants in generation x and phenotype plants in generation y to conduct QTL mapping. This design is called the $F_{x:y}$ design for $y \geq x$ (Fisch *et al.* 1996; Jiang and Zeng 1997; Chapman *et al.* 2003). This generalized design actually covers the special case of $F_{y:y}$. As $y \to \infty$, $F_{y:y}$ design becomes the recombinant inbred line (RIL) design (Knapp 1991). If $y = 2$, $F_{y:y}$ becomes the typical F_2 design of the experiment. For more information about the generalized design, one may consult with Fisch *et al.* (1996) and Jiang and Zeng (1997).

Statistical methods for F_2 and RIL designs have been well developed (Lander and Botstein 1989; Zeng 1993, 1994; Xu 1998a). The current method for the $F_{2:3}$ design is adopted from $F_{2:2}$ by simply replacing the F_2 phenotype by the average value of the F_3 progeny (Zhang *et al.* 2003). This simple treatment has ignored an important factor in the distribution of the residual error and has not been acknowledged in the quantita-

[1]*Corresponding author:* Department of Botany and Plant Sciences, 3126 Batchelor Hall, University of California, Riverside, California 92521-0124. E-mail: xu@genetics.ucr.edu

tive trait locus (QTL) mapping literature. Consider a particular locus of an F_2 plant. If the F_2 is homozygous, the residual variance of the mean phenotypic values of m F_3 progeny is simply σ^2/m, where σ^2 is the residual variance based on a single plant. However, if the F_2 is heterozygous, the residual error of the mean of m F_3 plants becomes a mixture of many distributions (Fisch *et al.* 1996; Jiang and Zeng 1997; Zhang *et al.* 2003). For example, if $m = 1$, the F_3 may take one of three [$= (m + 1)(m + 2)/2$] possible genotypes. Therefore, the residual error is a mixture of three normal distributions. If $m = 2$, the average of F_3 may take a mixture of six normal distributions. This mixed nature of distribution has not been investigated in QTL mapping studies. Incorporation of this mixture into the QTL model may significantly increase the efficiency.

If both F_2 and $F_{2:3}$ measurements are available, there is no reason to leave the F_2 phenotypes out and simply take the $F_{2:3}$ averages. Separate analyses of F_2 and $F_{2:3}$ will waste valuable information. No methods have been developed to combine the two data sets for a consensus analysis.

Multiple-interval mapping (MIM; Kao *et al.* 1999) of quantitative trait loci is now the state-of-the-art method for QTL mapping. However, it is difficult to implement MIM under the maximum-likelihood framework. The Bayesian method implemented via the Markov chain Monte Carlo (MCMC) algorithm (Hastings 1970; Geman and Geman 1984; Green 1995; Satagopan *et al.* 1996) is specialized to handle complicated models (Xu 2003) and thus it is the ideal tool for mapping multiple QTL.

In this study, we first investigate the theory of mixture distribution in the $F_{2:3}$ design. We then develop a method to combine both F_2 and $F_{2:3}$ populations in a single model by taking advantage of the mixture distribution. Finally, we develop a Bayesian method to implement a multiple-QTL mapping strategy.

STATISTICAL METHODS

Genetic model of $F_{2:3}$: Consider a quantitative trait locus Q in an F_2 population. Let us define the three possible genotypes by Q_1Q_1, Q_1Q_2, and Q_2Q_2, respectively. The values of the three genotypes in an F_2 population are defined as a, d, and $-a$. It is more convenient to use standard regression notation by letting $b_1 = a$ and $b_2 = d$. The phenotypic value of an individual j may be described by the following linear model,

$$y_j = b_0 + x_{1j}b_1 + x_{2j}b_2 + \varepsilon_j, \tag{1}$$

where b_0 is the population mean, ε_j is the residual error with a $N(0, \sigma^2)$ distribution, and

$$x_{1j} = \begin{cases} +1 & \text{for } Q_1Q_1 \\ 0 & \text{for } Q_1Q_2 \\ -1 & \text{for } Q_2Q_2 \end{cases} \quad \text{and} \quad x_{2j} = \begin{cases} 0 & \text{for } Q_1Q_1 \\ 1 & \text{for } Q_1Q_2. \\ 0 & \text{for } Q_2Q_2 \end{cases}$$

If the genotype of individual j is observed, we can proceed with the usual regression analysis to estimate and test the genetic effects of the QTL.

Consider the situation where the phenotypic value is not measured from the F_2 plant; rather, it is the average of m_j F_3 plants derived by selfing the F_2 individual. The model is modified as

$$\bar{y}_j = b_0 + \bar{x}_{1j}b_1 + \bar{x}_{2j}b_2 + \bar{\varepsilon}_j, \tag{2}$$

where

$$\bar{y}_j = m_j^{-1}\sum_{k=1}^{m_j} y_{jk},\ \bar{x}_{1j} = m_j^{-1}\sum_{k=1}^{m_j} x_{1jk},\ \bar{x}_{2j} = m_j^{-1}\sum_{k=1}^{m_j} x_{2jk},\ \bar{\varepsilon}_j = m_j^{-1}\sum_{k=1}^{m_j} \varepsilon_{jk},$$

and the variable with additional subscript k indicates the corresponding variable for the kth progeny of the jth F_2 plant. The residual error now follows a $N(0, \sigma^2/m_j)$ distribution.

Maximum likelihood: If genotypes of all the F_3 progeny are observed, $\bar{x}_{1j}$ and $\bar{x}_{2j}$ will be known and the log-likelihood value under the complete data situation will be

$$L_c(\boldsymbol{\theta}) = -\frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{n} m_j(\bar{y}_j - b_0 - \bar{x}_{1j}b_1 - \bar{x}_{2j}b_2)^2, \tag{3}$$

where $\boldsymbol{\theta} = [b_0\ b_1\ b_2\ \sigma^2]^T$ is the vector of parameters. The maximum-likelihood estimate (MLE) under this complete data likelihood will be easily obtained by the standard regression analysis. When the QTL genotypes are not observable but inferred from marker information, the likelihood function is defined differently. Let

$$f(\bar{y}_j|\boldsymbol{\theta}, \bar{x}_{1j}, \bar{x}_{2j}) = (2\pi\sigma^2)^{-1/2}m_j^{1/2} \times \exp\left[-\frac{m_j}{2\sigma^2}(\bar{y}_j - b_0 - \bar{x}_{1j}b_1 - \bar{x}_{2j}b_2)^2\right] \tag{4}$$

be the conditional density of $\bar{y}_j$; the log-likelihood function defined under the missing-value situation is

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n}\ln\{E[f(\bar{y}_j|\boldsymbol{\theta}, \bar{x}_{1j}, \bar{x}_{2j})]\}. \tag{5}$$

The asymptotically orthogonal variables $\bar{x}_{1j}$ and $\bar{x}_{2j}$ are missing and the expectation-maximization (EM) algorithm (Dempster *et al.* 1977) can be used to obtain the MLE, as shown below,

$$\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^n m_j & \sum_{j=1}^n m_j E(\bar{x}_{1j}) & \sum_{j=1}^n m_j E(\bar{x}_{2j}) \\ \sum_{j=1}^n m_j E(\bar{x}_{1j}) & \sum_{j=1}^n m_j E(\bar{x}_{1j}^2) & \sum_{j=1}^n m_j E(\bar{x}_{1j}\bar{x}_{2j}) \\ \sum_{j=1}^n m_j E(\bar{x}_{2j}) & \sum_{j=1}^n m_j E(\bar{x}_{1j}\bar{x}_{2j}) & \sum_{j=1}^n m_j E(\bar{x}_{2j}^2) \end{bmatrix}^{-1} \times \begin{bmatrix} \sum_{j=1}^n m_j\bar{y}_j \\ \sum_{j=1}^n m_j E(\bar{x}_{1j})\bar{y}_j \\ \sum_{j=1}^n m_j E(\bar{x}_{2j})\bar{y}_j \end{bmatrix} \tag{6}$$

and

$$\sigma^2 = \frac{1}{n}\sum_{j=1}^{n} m_j E[(\bar{y}_j - b_0 - \bar{x}_{1j}b_1 - \bar{x}_{2j}b_2)^2]. \tag{7}$$

The expectation shown in Equation 7 can be further expressed as

$$E[(\bar{y}_j - b_0 - \bar{x}_{1j}b_1 - \bar{x}_{2j}b_2)^2] = (\bar{y}_j - b_0)^2 + b_1^2E(\bar{x}_{1j}^2) + b_2^2E(\bar{x}_{2j}^2) - 2(\bar{y}_j - b_0)[b_1E(\bar{x}_{1j}) + b_2E(\bar{x}_{2j})] + 2b_1b_2E(\bar{x}_{1j}\bar{x}_{2j}). \quad (8)$$

The E-step is to obtain the expectations involving $\bar{x}_{ij}$ ($i = 1, 2$; $j = 1, \ldots, n$) in Equations 6 and 7 and the M-step is to update the estimates of the parameters using Equations 6 and 7.

The current method of calculating the expectations simply adopts the F_2 mapping procedure by replacing the average genotypic indicators of the F_3 plants by the corresponding genotype of the F_2 parent, *i.e.*, substituting $\bar{x}_{1j}$ and $\bar{x}_{2j}$ by the corresponding x_{1j} and x_{2j} of the F_2 parent. This *ad hoc* method has ignored an important feature of the $F_{2:3}$ design. If the F_2 plant is homozygous at the QTL considered, all the F_3 progeny should be homozygous also and the relationships $\bar{x}_{1j} = x_{1j}$ and $\bar{x}_{2j} = x_{2j}$ indeed apply. However, if the F_2 plant is heterozygous, half of the F_3 progeny remain heterozygous, but half of them will be split into the two homozygous classes. This nature should be incorporated into the calculation of the expectations, but so far it has not been acknowledged in the literature. We now propose a couple of methods to take this into account.

Exact EM algorithm: Let us consider progeny of the jth F_2 plant and thus the subscript j can be ignored as needed. Denote the numbers of the three possible genotypes of the F_3 progeny by m_{11}, m_{12}, and m_{22}, respectively, where $m_{11} + m_{12} + m_{22} = m_j$. This leads to

$$\bar{x}_{1j} = m_j^{-1}(m_{11} - m_{22}) \quad \text{and} \quad \bar{x}_{2j} = m_j^{-1}m_{12}. \quad (9)$$

If the F_2 plant is Q_1Q_1, then all the F_3 progeny will be Q_1Q_1, leading to $m_{11} = m_j$, and thus $\bar{x}_{1j} = m_j^{-1}(m_j - 0) = 1$ and $\bar{x}_{2j} = m_j^{-1} \times 0 = 0$. If the F_2 plant is Q_2Q_2, then all the F_3 progeny will be Q_2Q_2, leading to $m_{22} = m_j$, and thus $\bar{x}_{1j} = m_j^{-1}(0 - m_j) = -1$ and $\bar{x}_{2j} = m_j^{-1} \times 0 = 0$. However, if the F_2 plant is Q_1Q_2, the general formula in Equations 6 and 7 must apply. Therefore, conditional on Q_1Q_2 of the F_2 parent, these m's follow a multinomial distribution with a probability

$$p(m_{11}, m_{12}, m_{22}) = \frac{m_j!}{m_{11}!m_{12}!m_{22}!}\left(\frac{1}{4}\right)^{m_{11}+m_{22}}\left(\frac{1}{2}\right)^{m_{12}}. \quad (10)$$

As a result, the distribution of $\bar{y}_j$ is a mixture of several normal distributions, *i.e.*,

$$f(\bar{y}_j|\boldsymbol{\theta}) = p_{11}\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j) + p_{12}\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta}) + p_{22}\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j), \quad (11)$$

where $\phi(t_1, t_2, t_3)$ represents the normal density of variable t_1 with mean t_2 and variance t_3; p_{11}, p_{12}, and p_{22} are the probabilities of the three QTL genotypes for the F_2 parent inferred from marker information; and

$$\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta}) = \sum_{\Omega} p(m_{11}, m_{12}, m_{22})\phi[\bar{y}_j, b_0 + m_j^{-1}(m_{11} - m_{22})b_1 + m_j^{-1}m_{12}b_2, \sigma^2/m_j], \quad (12)$$

where Ω defines the domain of all possible values of the m's subject to the restriction of $m_{11} + m_{12} + m_{22} = m_j$. The posterior probabilities of these QTL genotypes are calculated as

$$p_{11}^* = \frac{p_{11}\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j)}{p_{11}\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j) + p_{12}\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta}) + p_{22}\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j)} \quad (13a)$$

$$p_{12}^* = \frac{p_{12}\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta})}{p_{11}\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j) + p_{12}\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta}) + p_{22}\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j)} \quad (13b)$$

and

$$p_{22}^* = \frac{p_{22}\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j)}{p_{11}\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j) + p_{12}\phi_{\text{mix}}(\bar{y}_j, \boldsymbol{\theta}) + p_{22}\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j)}. \quad (13c)$$

In addition, we need to calculate the posterior probabilities of $\bar{x}_{1j}$ and $\bar{x}_{2j}$ for the F_3 progeny of the heterozygous F_2 plant, as shown below:

$$p^*(m_{11}, m_{12}, m_{22}) = \frac{p(m_{11}, m_{12}, m_{22})\phi[\bar{y}_j, b_0 + m_j^{-1}(m_{11} - m_{22})b_1 + m_j^{-1}m_{12}b_2, \sigma^2/m_j]}{\sum_{\Omega}p(m_{11}, m_{12}, m_{22})\phi[\bar{y}_j, b_0 + m_j^{-1}(m_{11} - m_{22})b_1 + m_j^{-1}m_{12}b_2, \sigma^2/m_j]}. \quad (14)$$

This posterior probability is used to calculate various expectations, as shown below,

$$\begin{aligned} E(\bar{x}_{1j}) &= p_{11}^* - p_{22}^* + p_{12}^*m_j^{-1}E(m_{11} - m_{22}) \\ E(\bar{x}_{2j}) &= p_{12}^*m_j^{-1}E(m_{12}) \\ E(\bar{x}_{1j}^2) &= p_{11}^* + p_{22}^* + p_{12}^*m_j^{-2}E(m_{11} - m_{22})^2 \\ E(\bar{x}_{2j}^2) &= p_{12}^*m_j^{-2}E(m_{12}^2) \\ E(\bar{x}_{1j}\bar{x}_{2j}) &= p_{12}^*m_j^{-2}E[m_{12}(m_{11} - m_{22})], \end{aligned} \quad (15)$$

where

$$\begin{aligned} E(m_{11} - m_{22}) &= \sum_{\Omega}p^*(m_{11}, m_{12}, m_{22})(m_{11} - m_{22}) \\ E(m_{11} - m_{22})^2 &= \sum_{\Omega}p^*(m_{11}, m_{12}, m_{22})(m_{11} - m_{22})^2 \\ E(m_{12}) &= \sum_{\Omega}p^*(m_{11}, m_{12}, m_{22})m_{12} \\ E(m_{12}^2) &= \sum_{\Omega}p^*(m_{11}, m_{12}, m_{22})m_{12}^2 \\ E[m_{12}(m_{11} - m_{22})] &= \sum_{\Omega}p^*(m_{11}, m_{12}, m_{22})m_{12}(m_{11} - m_{22}). \end{aligned} \quad (16)$$

This exact EM algorithm is recommended for small m_j, say $m_j \leq 5$. Note that the number of possible partitionings of m_j is $(m_j + 1)(m_j + 2)/2$, which can be very large as m_j increases.

Approximate EM algorithm: When m_j is large, $\bar{x}_{1j} = m_j^{-1}(m_{11} - m_{22})$ and $\bar{x}_{2j} = m_j^{-1}m_{12}$ derived from the heterozygous F_2 may be approximated by a joint bivariate

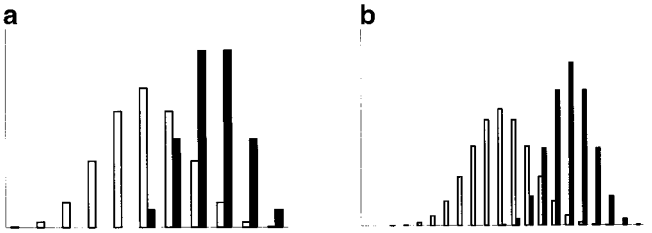

FIGURE 1.—Distributions of $\bar{x}_{1j}$ (open bars) and $\bar{x}_{2j}$ (solid bars) for F_3 progeny derived from Q_1Q_2 parents (a) when the number of F_3 progeny per F_2 plant is 5 and (b) when the number of F_3 progeny per F_2 plant is 10.

normal distribution (Figure 1). In this approximation, we simply replace all distributions related to $p(m_{11}, m_{12}, m_{22})$ and $p^*(m_{11}, m_{12}, m_{22})$ in the exact method by the corresponding joint bivariate normal, $p(m_j^{-1}(m_{11} - m_{22}), m_j^{-1}m_{12})$ and $p^*(m_j^{-1}(m_{11} - m_{22}), m_j^{-1}m_{12})$. The joint prior normal $p(m_j^{-1}(m_{11} - m_{22}), m_j^{-1}m_{12})$ has the following expectation c and variance-covariance matrix $\mathbf{C}$,

$$c = \begin{pmatrix} 0 \\ 1/2 \end{pmatrix} \quad \text{and} \quad \mathbf{C} = \begin{pmatrix} 1/(2m_j) & 0 \\ 0 & 1/(4m_j) \end{pmatrix}. \tag{17}$$

The joint posterior normal $p^*(m_j^{-1}(m_{11} - m_{22}), m_j^{-1}m_{12})$ has the following expectation and variance-covariance matrix,

$$\boldsymbol{\mu}^* = (\mathbf{V} + \mathbf{C}^{-1})^{-1}(\mathbf{C}^{-1}c + m_j\mathbf{b}^{\mathrm{T}}\bar{y}_j^*/\sigma^2) \tag{18a}$$

$$\mathbf{V}^* = (\mathbf{V} + \mathbf{C}^{-1})^{-1} \tag{18b}$$

(SORENSEN and GIANOLA 2002), where $\mathbf{b} = (b_1 \ b_2)$, $\mathbf{V} = m_j\mathbf{b}^{\mathrm{T}}\mathbf{b}/\sigma^2$, and $\bar{y}_j^* = \bar{y}_j - b_0$. Given the above joint normal distribution, we can replace $\phi_{\mathrm{mix}}(\bar{y}_j, \boldsymbol{\theta})$ in the exact method by a single normal distribution,

$$\phi_{\mathrm{mix}}(\bar{y}_j, \boldsymbol{\theta}) = \phi[\bar{y}_j, b_0 + b_2/2, (b_1^2/2 + b_2^2/4 + \sigma^2)/m_j] \tag{19}$$

(ZHANG *et al.* 2003), which is required to calculate p_{11}^*, p_{12}^*, and p_{22}^*. The expectations involving m_{11}, m_{12}, and m_{22} in the exact method are substituted by

$$\begin{aligned} E(m_{11} - m_{22}) &= m_j\mu_1^* \\ E(m_{11} - m_{22})^2 &= m_j^2[V_{11}^* + (\mu_1^*)^2] \\ E(m_{12}) &= m_j\mu_2^* \\ E(m_{12}^2) &= m_j^2[V_{22}^* + (\mu_2^*)^2] \\ E[m_{12}(m_{11} - m_{22})] &= m_j^2(V_{12}^* + \mu_1^*\mu_2^*), \end{aligned} \tag{20}$$

where μ_i^* and V_{ij}^* $(i, j = 1, 2)$ are the elements of vector $\boldsymbol{\mu}^*$ and matrix $\mathbf{V}^*$, respectively. We now demonstrate that the normal approximation has substantially speeded up the computation with negligible loss in power.

Joint analysis of F_2 and $F_{2:3}$: When both the F_2 phenotype (y_j) and the average value of the F_3 progeny ($\bar{y}_j$) are available, the two sources of data can be combined in a joint analysis. Generally speaking, there may be different means and environmental variances for the two different populations because other QTL not included in the model (collectively called the polygene) may cause the differences in the mean and variance (JANSEN *et al.* 1998; XU 1998b). The method developed can handle this special feature. Let $f(\bar{y}_j|\boldsymbol{\theta}, \bar{x}_{1j}, \bar{x}_{2j})$ be the probability density of the mean of F_3 progeny as defined in Equation 4 but now the residual variance is denoted by σ_3^2 for an individual F_3 plant. The corresponding probability density of F_2 phenotype that takes into account the different mean and variance is

$$f(y_j|\boldsymbol{\theta}, x_{1j}, x_{2j}) = (2\pi\sigma_2^2)^{-1/2} \times \exp\left[-\frac{1}{2\sigma_2^2}(y_j - b_0 - b - x_{1j}b_1 - x_{2j}b_2)^2\right], \tag{21}$$

where b is the difference between the means of populations F_2 and F_3, and σ_2^2 is the environmental variance of the F_2 plants. The joint density of the F_2 and F_3 conditional on the genotypes is

$$f(y_j, \bar{y}_j|\boldsymbol{\theta}, x_{1j}, x_{2j}, \bar{x}_{1j}, \bar{x}_{2j}) = f(y_j|\boldsymbol{\theta}, x_{1j}, x_{2j})f(\bar{y}_j|\boldsymbol{\theta}, \bar{x}_{1j}, \bar{x}_{2j}). \tag{22}$$

The log-likelihood function with the missing genotypes integrated out is

$$L(\boldsymbol{\theta}) = \sum_{j=1}^{n} \ln\{E[f(y_j, \bar{y}_j|\boldsymbol{\theta}, x_{1j}, x_{2j}, \bar{x}_{1j}, \bar{x}_{2j})]\}, \tag{23}$$

where the expectation is taken with respect to both the F_2 genotype and the genotypes of the F_3 progeny derived from the heterozygous F_2 parent. For the joint analysis, the EM equations are

$$\begin{pmatrix} b_0 \\ b \\ b_1 \\ b_2 \end{pmatrix} = \begin{bmatrix} n+\rho\sum_{j=1}^{n} m_j & n & \sum_{j=1}^{n}[E(x_{1j})+\rho m_j E(\bar{x}_{1j})] \\ n & n & \sum_{j=1}^{n} E(x_{1j}) \\ \sum_{j=1}^{n}[E(x_{1j})+\rho m_j E(\bar{x}_{1j})] & \sum_{j=1}^{n} E(x_{1j}) & \sum_{j=1}^{n}[E(x_{1j}^2)+\rho m_j E(\bar{x}_{1j}^2)] \\ \sum_{j=1}^{n}[E(x_{2j})+\rho m_j E(\bar{x}_{2j})] & \sum_{j=1}^{n} E(x_{2j}) & \sum_{j=1}^{n}[E(x_{1j}x_{2j})+\rho m_j E(\bar{x}_{1j}\bar{x}_{2j})] \end{bmatrix}$$

$$\begin{bmatrix} \sum_{j=1}^{n}[E(x_{2j})+\rho m_j E(\bar{x}_{2j})] \\ \sum_{j=1}^{n} E(x_{2j}) \\ \sum_{j=1}^{n}[E(x_{1j}x_{2j})+\rho m_j E(\bar{x}_{1j}\bar{x}_{2j})] \\ \sum_{j=1}^{n}[E(x_{2j}^2)+\rho m_j E(\bar{x}_{2j}^2)] \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n}(y_j+\rho m_j\bar{y}_j) \\ \sum_{j=1}^{n} y_j \\ \sum_{j=1}^{n}[E(x_{1j})y_j+\rho m_j E(\bar{x}_{1j})\bar{y}_j] \\ \sum_{j=1}^{n}[E(x_{2j})y_j+\rho m_j E(\bar{x}_{2j})\bar{y}_j] \end{bmatrix} \tag{24}$$

and

$$\sigma_2^2 = \sum_{j=1}^{n} E(y_j - \hat{y}_j)^2/n \tag{25a}$$

$$\sigma_3^2 = \sum_{j=1}^{n} m_j E(\bar{y}_j - \hat{\bar{y}}_j)^2/n, \tag{25b}$$

where $\rho = \sigma_2^2/\sigma_3^2$, $\hat{y}_j = b_0 + b + x_{1j}b_1 + x_{2j}b_2$, and $\hat{\bar{y}}_j = b_0 + \bar{x}_{1j}b_1 + \bar{x}_{2j}b_2$.

In addition, the joint analysis differs from the $F_{2:3}$ analysis in that the posterior probabilities of the QTL genotypes for the F_2 parents should incorporate the phenotypic values of both the parents and the progeny.

The following terms need to be modified in developing the new posterior probabilities. The mixture distribution $\phi_{mix}(\bar{y}_j, \boldsymbol{\theta})$ in Equation 13 should be replaced by

$$\phi_{12}(y_j, \bar{y}_j, \boldsymbol{\theta}) = \phi_{mix}(\bar{y}_j, \boldsymbol{\theta})\phi(y_j, b_0 + b + b_2, \sigma_2^2). \quad (26)$$

Furthermore, we need to replace $\phi(\bar{y}_j, b_0 + b_1, \sigma^2/m_j)$ by

$$\phi_{11}(y_j, \bar{y}_j, \boldsymbol{\theta}) = \phi(\bar{y}_j, b_0 + b_1, \sigma_3^2/m_j)\phi(y_j, b_0 + b + b_1, \sigma_2^2) \quad (27)$$

and $\phi(\bar{y}_j, b_0 - b_1, \sigma^2/m_j)$ by

$$\phi_{22}(y_j, \bar{y}_j, \boldsymbol{\theta}) = \phi(\bar{y}_j, b_0 - b_1, \sigma_3^2/m_j)\phi(y_j, b_0 + b - b_1, \sigma_2^2). \quad (28)$$

The mixture model described so far applies to a single-QTL model. The single-QTL model has been applied to mapping multiple QTL using the one-dimensional search. Multiple peaks in the likelihood-ratio profile indicate multiple QTL. This single-QTL model is somehow *ad hoc* and it is presented here to demonstrate the theory and the advantage of the improved method. Because the single-QTL model is handled via the EM algorithm, extensive simulation is possible to validate the improved method. In practice, a multiple-QTL model should be used. However, it is hard to implement the multiple-QTL model using the EM algorithm. We decide to adopt a Bayesian method for multiple-QTL mapping. The probability model developed in the EM is largely applicable to the Bayesian method. So, with little additional effort, we can develop a Bayesian method to implement multiple-QTL mapping.

Mapping multiple QTL: The major difference between ML and Bayesian methods is that quantities involving unobserved variables are replaced by the posterior expectations in the ML, whereas in Bayesian analysis, the unobserved variables are replaced by values simulated from their posterior distributions. Conditional on simulated values of all other variables, one can concentrate on a model that contains only a single variable. All the probability theory developed early in the single-QTL model applies here to the Bayesian analysis as a result of the simulation.

The multiple-QTL models for an F_2 plant and the mean of F_3 progeny may be described as

$$y_j = b_0 + b + \sum_{l=1}^{P}(x_{l1j}b_{l1} + x_{l2j}b_{l2}) + \varepsilon_j \quad (29a)$$

$$\bar{y}_j = b_0 + \sum_{l=1}^{P}(\bar{x}_{l1j}b_{l1} + \bar{x}_{l2j}b_{l2}) + \bar{\varepsilon}_j, \quad (29b)$$

where P is the total number of QTL included in the models, l indexes the QTL, $\varepsilon_j \sim N(0, \sigma_2^2)$, and $\bar{\varepsilon}_j \sim N(0, \sigma_3^2/m_j)$ are the residual errors. The total number of QTL, P, is unknown and usually treated as a parameter that can be inferred from the data (Green 1995; Stephens and Fisch 1998; Yi and Xu 2001; Xu 2002). In this study, we treated P as the maximum (or potential) number of QTL that the model can resolve. We assume that there is one QTL in every marker interval, and thus P is identical to the number of intervals throughout the genome. Of course, most of the intervals will contain no QTL. Traditionally, one needs a special model selection mechanism to select the set of intervals containing significant QTL and leave those null intervals out of the model. In this study, however, we take a different approach that can avoid all problems associated with model selection by allowing the estimated QTL effects of the null intervals to be close to zero. So, an interval with no QTL is the same as an interval having a QTL but with a zero effect. If an interval contains more than one QTL, assuming one QTL for the interval is still the best we can do because there may not be enough information to resolve for more than one QTL in any interval.

The Bayesian method is implemented via the MCMC by iteratively simulating the missing values and all other unknown variables. Once the missing values are simulated, the model can be handled in a straightforward way. The missing values in the model are the genotypes of the QTL, *i.e.*, the x's variables in the model. Once they are given, we can sample the unknown parameters from their conditional posterior distribution. Methods used to sample these x's variables are quite standard from the posterior distributions described in earlier sections. What we need here is to express the multiple-QTL model into a single-QTL model by fixing the genotypes and the effects of all other QTL that are not currently under investigation. For example, when we update the lth QTL effects and the genotypes, we can rewrite the above models as

$$y_j^* = x_{l1j}b_{l1} + x_{l2j}b_{l2} + \varepsilon_j \quad (30a)$$

$$\bar{y}_j^* = \bar{x}_{l1j}b_{l1} + \bar{x}_{l2j}b_{l2} + \bar{\varepsilon}_j, \quad (30b)$$

where

$$y_j^* = y_j - b_0 - b - \sum_{k\neq l}^{P}(x_{k1j}b_{k1} + x_{k2j}b_{k2})$$

and

$$\bar{y}_j^* = \bar{y}_j - b_0 - \sum_{k\neq l}^{P}(\bar{x}_{k1j}b_{k1} + \bar{x}_{k2j}b_{k2}).$$

Upon substituting y_j and $\bar{y}_j$ by y_j^* and $\bar{y}_j^*$, respectively, the posterior distribution of the effects of the lth QTL is simply a joint normal distribution.

Bayesian methods for QTL mapping have been well developed (Hoeschele and Vanraden 1993; Satagopan *et al.* 1996; Thaller and Hoeschele 1996; Uimari and Hoeschele 1997; Sillanpaa and Arjas 1998, 1999; Stephens and Fisch 1998; Sorensen and Gianola 2002; Yi and Xu 2001; Xu 2002, 2003) and the MCMC algorithms used are fairly standard. Hence, no detailed description is necessary. Two characteristics of our Bayesian method are worth mentioning. One is the setup of prior distribution for the QTL effects and the other is the restriction superimposed on the sampled

position of the QTL within the interval, both of which are discussed as follows.

The prior distribution for a QTL effect is normal with mean zero and a specific variance. For instance, the prior for the additive effect of QTL l is $b_{l1} \sim N(0, \sigma^2_{l1})$, where σ^2_{l1} varies from one locus to another. Similarly, the dominance effect has a prior of $b_{l2} \sim N(0, \sigma^2_{l2})$. The conditional posterior distribution of b_{l1} is also normal but with mean

$$E(b_{l1}) = \left[\sum_{j=1}^{n}(x^2_{l1j} + \rho m_j \bar{x}^2_{l1j}) + \sigma^2_2/\sigma^2_{l1}\right]^{-1}\left[\sum_{j=1}^{n}(x_{l1j}y^*_j + \rho m_j \bar{x}_{l1j}\bar{y}^*_j)\right]$$

and variance

$$\mathrm{Var}(b_{l1}) = \left[\sum_{j=1}^{n}(x^2_{l1j} + \rho m_j \bar{x}^2_{l1j}) + \sigma^2_2/\sigma^2_{l1}\right]^{-1}\sigma^2_2,$$

where $\rho = \sigma^2_2/\sigma^2_3$,

$$y^*_j = y_j - b_0 - b - \sum_{k\neq l}^{P} x_{k1j}b_{k1} - \sum_{k=1}^{P} x_{k2j}b_{k2}$$

and

$$\bar{y}^*_j = \bar{y}_j - b_0 - \sum_{k\neq l}^{P} \bar{x}_{k1j}b_{k1} - \sum_{k=1}^{P} \bar{x}_{k2j}b_{k2}.$$

The additive effect b_{l1} can be sampled from the above conditional posterior distribution. The dominance effect b_{l2} is sampled similarly.

The remaining question is how to choose σ^2_{l1}. In typical Bayesian regression analysis, σ^2_{l1} is considered as the parameter of a prior distribution and thus it is a constant in the MCMC process. In this study, we consider σ^2_{l1} as another variable described by its own prior distribution. We choose a vague prior for σ^2_{l1}, *i.e.*, $p(\sigma^2_{l1}) \propto 1/\sigma^2_{l1}$, which leads to a posterior distribution proportional to an inverted chi square. Therefore, σ^2_{l1} can be sampled from a scaled inverted chi-square distribution, *i.e.*, $\sigma^2_{l1} = b^2_{l1}/\chi^2_1$, where χ^2_1 is a random variable sampled from a chi-square distribution with 1 d.f. This is the key step in our Bayesian analysis for handling a large number of model effects but requiring no model selection. If b_{l1} is small, σ^2_{l1} may be even smaller, leading to a posterior distribution of b_{l1} with even smaller mean and variance. This is the so-called shrinkage estimate. This method was first used by Xu (2003) for marker analysis.

When each interval is assumed to have a QTL, the position of the QTL must be sampled within this interval. We used a random walk approach to sample the new position around the current position. We first sample a new position for the lth QTL, called the proposal position and denoted by

$$\lambda^*_j = \lambda_j + \delta, \tag{31}$$

where λ_j is the current position and δ is sampled from $U(-d, d)$, where d is a small positive number (tuning parameter) usually taking 1 cM. This new position is accepted with a probability $\min(1, \alpha)$, where

$$\alpha = \frac{\prod_{j=1}^{n} p(y_j, \bar{y}_j|\lambda^*_j, \ldots)}{\prod_{j=1}^{n} p(y_j, \bar{y}_j|\lambda_j, \ldots)} \times \frac{q(\lambda_j|\lambda^*_j)}{q(\lambda^*_j|\lambda_j)}. \tag{32}$$

If the new position is accepted, the current position is replaced by λ^*_j and the QTL genotypes are updated according to the conditional posterior distribution at the new position. If the new position is rejected, the current position is carried on to the next cycle, but the genotypes of the old position are still subject to updating. The proposal probability for the new position is $q(\lambda^*_j|\lambda_j)$, which is usually cancelled out with its reverse partner $q(\lambda_j|\lambda^*_j)$ in a typical Bayesian mapping procedure using model selection (Xu 2002). However, in the situation where the new QTL position is always restricted to the particular interval in which the old position occurs, the two proposal probabilities often are not symmetrical because the new position may frequently hit the boundaries. Therefore, we take the proposal probabilities with the following modifications,

$$q(\lambda_j|\lambda^*_j) = \begin{cases} \dfrac{1}{d + \lambda^*_j - l_1}, & \text{if } \lambda^*_j < l_1 + d \\ \dfrac{1}{d + l_2 - \lambda^*_j}, & \text{if } \lambda^*_j > l_2 - d \\ \dfrac{1}{2d}, & \text{otherwise} \end{cases}$$

and

$$q(\lambda^*_j|\lambda_j) = \begin{cases} \dfrac{1}{d + \lambda_j - l_1}, & \text{if } \lambda_j < l_1 + d \\ \dfrac{1}{d + l_2 - \lambda_j}, & \text{if } \lambda_j > l_2 - d \\ \dfrac{1}{2d}, & \text{otherwise,} \end{cases}$$

where l_1 and l_2 are the left and right positions of the interval in question.

SIMULATION STUDIES

Single-QTL analysis: The purpose of the simulation studies is to demonstrate that (1) the correct $F_{2:3}$ analysis using the new EM algorithm is more efficient than the currently adopted F_2 method, (2) the normal approximation of the EM algorithm is as efficient as the exact EM algorithm in practice, and (3) joint analysis of F_2 and $F_{2:3}$ can substantially increase the power as opposed to either analysis.

Eleven equally spaced markers were simulated on a single-chromosome segment of length 100 cM. A single QTL was located at position 25 cM. Under the null model, the QTL was assigned a value of zero for both the additive and dominance effects. All simulations were replicated 100 times. The quick method developed by

Piepho (2001) was used to determine the critical value for power calculation, the empirical type I error was <5%, although the targeted type I error was set at 5% (Piepho 2001). Under the alternative model, nonzero additive and dominance effects were simulated. Empirical power was calculated by counting the number of runs in which test statistics were greater than the critical values. In all simulations, the environmental error variance on the individual plant basis was set at $\sigma^2 = 1$, and the overall means for F_2 and $F_{2:3}$ were the same. The conditional probabilities of QTL genotypes given marker information were calculated on the basis of the multipoint method (Rao and Xu 1998). The simulation studies were by no means exhaustive. As long as we can demonstrate the superiority of the new method over the existing *ad hoc* method, we will make a recommendation in favor of the new method.

To demonstrate the first objective of the simulation experiments, we simulated 100 $F_{2:3}$ families each with five plants. Only $\bar{y}_j$ was recorded and used in the analysis. Each data set was analyzed under two methods: (1) the old method (currently available), where the mixture distribution of the F_3 average derived from the heterozygous F_2 parents was completely ignored (the adopted F_2 mapping procedure), and (2) the exact EM method developed in this study, where the mixture distribution of the F_3 progeny derived from the heterozygous F_2 parents was fully taken into account. The two methods were compared under three levels of the QTL size measured as the proportion of the phenotypic variance (individual plant basis) explained by the QTL (also called the QTL heritability): $h^2 = 0.05$, 0.10, and 0.15. The corresponding additive and dominance effects under each level of h^2 are given in Table 1. For example, when $h^2 = 0.15$, we set $a = 0.420$ and $d = 0.594$, leading to a total genetic variance of $0.5 \times 0.420^2 + 0.25 \times 0.594^2 = 0.1764$. The heritability is $h^2 = 0.1764/(0.1764 + 1.0) = 0.15$. The heritability is actually expressed on the individual F_2 plant basis. One complication from the multiple-generation QTL mapping problem is that different generations usually have different genetic variances and thus different definitions of heritability. Here we simply defined the genetic variance and the environmental residual variance on the single-plant basis. This will eliminate such complication. The results are listed in Table 1, where the old method is called the "adopted F_2" method and the new method is called the "exact EM" method. The two methods differ in the following aspects: (1) the estimates of the dominance effects are severely biased for the adopted F_2 method, whereas the biases are largely corrected by using the correct model, the exact EM method; (2) using the correct model can significantly increase the statistical power of QTL detecting compared with the old method. The two differences clearly demonstrate the superiority of the new method over the old one. We also simulated several

TABLE 1

Comparisons of the "adopted F_2" (old) method with the "exact EM" (new) method for QTL mapping in an $F_{2:3}$ design

					Estimates				
a	d	h^2	Method	Power (%)	cM̂	$\hat{a}$	$\hat{d}$	$\hat{h}^2$	$\hat{\sigma}^2$	LOD
0.229	0.324	0.05	Adopted F_2	63	25.96 (13.71)	0.2323 (0.0618)	0.1637 (0.1368)	0.040 (0.021)	0.9450 (0.1450)	3.80 (1.84)
			Exact EM	86	26.43 (13.82)	0.2236 (0.0662)	0.3403 (0.2656)	0.074 (0.036)	0.9132 (0.1461)	5.45 (2.75)
0.333	0.471	0.10	Adopted F_2	96	25.64 (8.51)	0.3375 (0.0619)	0.2449 (0.1041)	0.073 (0.023)	0.9732 (0.1323)	6.54 (2.00)
			Exact EM	100	25.53 (6.83)	0.3316 (0.0654)	0.4598 (0.2121)	0.116 (0.041)	0.9273 (0.1320)	9.10 (2.69)
0.420	0.594	0.15	Adopted F_2	100	24.91 (3.58)	0.4166 (0.0646)	0.3024 (0.0949)	0.106 (0.026)	0.9818 (0.1413)	8.87 (2.20)
			Exact EM	100	25.08 (3.54)	0.4075 (0.0633)	0.5522 (0.1736)	0.161 (0.040)	0.9265 (0.1356)	12.32 (2.90)

a, additive effect; d, dominance effect; h^2, the proportion of phenotypic variance explained by the QTL; and cM̂, for the estimated position of the QTL. The true QTL position is at 25 cM of the simulated chromosome of 100 cM. The residual variance on the individual plant basis is $\sigma^2 = 1$. The estimated parameters were obtained from the averages of 100 replicated simulations with the standard deviations among the replicates listed after the estimated values (in parentheses). There are 100 $F_{2:3}$ families each with five progeny.

TABLE 2

Comparisons of the exact EM method with the approx EM method for joint analysis of F_2 and F_3 generations

			Estimates				
a	h^2	Method	$\hat{a}$	$\hat{d}$	$\hat{h}^2$	$\hat{cM}$	$\hat{\sigma}^2$
0.324	0.05	Exact EM	0.3286 (0.0403)	−0.0137 (0.0717)	0.054 (0.014)	25.37 (3.52)	0.9874 (0.0730)
		Approx EM	0.3267 (0.0389)	0.0062 (0.0934)	0.055 (0.013)	25.43 (4.07)	0.9740 (0.0728)
0.594	0.15	Exact EM	0.5940 (0.0432)	−0.0038 (0.0730)	0.158 (0.023)	24.99 (1.54)	0.9548 (0.0753)
		Approx EM	0.5831 (0.0453)	−0.0119 (0.0816)	0.154 (0.023)	24.91 (1.58)	0.9477 (0.0752)
0.816	0.25	Exact EM	0.8029 (0.0444)	−0.0061 (0.0701)	0.257 (0.028)	25.04 (1.38)	0.9411 (0.0820)
		Approx EM	0.8110 (0.0453)	−0.0081 (0.0718)	0.259 (0.028)	24.89 (1.16)	0.9493 (0.0737)

The true dominance effect is absent in the simulation. There are 200 F_2 plants each having five F_3 progeny. Standard deviations are in parentheses.

other situations and did not find a single case where the new method is not better than the old method.

To demonstrate the second objective of the simulation experiments, we compared the efficiencies of two methods developed in this study: the exact EM method described in the previous paragraph and the "approx EM" method in which the multinomial distribution of the average F_3 genotype derived from the heterozygous F_2 parents has been replaced by the approximate normal distribution. The number of F_2 plants was set at 200 and each F_2 parent had five F_3 progeny. Other setups are identical to the first simulation experiment (Table 1). The two methods are compared under three levels of heritability: 0.05, 0.15, and 0.25. The results of joint analysis for both y_j and $\bar{y}_j$ are listed in Table 2. No significant differences were found for the two methods. The goodness of normal approximation depends on the family size (the number of F_3 progeny per F_2 parent). We simulated a situation where only one F_3 progeny was used from each F_2 plant. The results are given in Table 3, where we do see some disadvantage of the approximate method. We recommend that the approximate method be used if the family size is $m_j \geq 5$ because there is no apparent information loss when $m_j = 5$ and beyond this the exact method becomes extremely time consuming. For example, when $m_j = 5$, a single run takes 30 min for the exact method but it takes only 10 min for the approximate method.

In all subsequent analyses, we used the exact EM for $m_j \leq 5$ and the approximate EM method for $m_j > 5$. Table 3 shows the effect of m_j on the new method under the joint analysis. The results show the general behavior of QTL mapping; *i.e.*, as m_j increases, the result becomes better (judged by the decrease in the standard deviation). We also investigated the impact of the sample size (number of F_2 plants) on the results of the joint mapping (see Table 4). Again, the observed trend is consistent with what was expected; *i.e.*, the method performs better as the sample increases.

The last objective of the simulation is to demonstrate

TABLE 3

Effects of the number of F_3 progeny per F_2 plant (m_j) on the joint analysis of F_2 and F_3 generations for QTL mapping

				Estimates					
a	h^2	m_j	Power (%)	$\hat{a}$	$\hat{d}$	$\hat{h}^2$	$\hat{cM}$	$\hat{\sigma}^2$	LOD
0.324	0.05	1	77	0.3224 (0.0646)	−0.0184 (0.1326)	0.057 (0.020)	25.87 (13.93)	0.9746 (0.0778)	4.27 (1.81)
		2	95	0.3270 (0.0637)	0.0113 (0.1374)	0.058 (0.019)	28.15 (12.09)	0.9733 (0.0709)	6.43 (2.41)
		3	100	0.3289 (0.0492)	−0.0179 (0.1102)	0.057 (0.016)	24.53 (5.32)	0.9691 (0.0731)	8.37 (2.66)
		4	100	0.3256 (0.0491)	0.0069 (0.1016)	0.055 (0.016)	25.15 (3.78)	0.9740 (0.0701)	9.76 (2.96)
		5	100	0.3199 (0.0412)	0.0011 (0.0993)	0.053 (0.014)	24.84 (3.55)	0.9788 (0.0730)	11.19 (3.41)
		10	100	0.3223 (0.0310)	0.0095 (0.0723)	0.053 (0.010)	24.99 (2.60)	0.9726 (0.0766)	18.44 (3.43)
0.816	0.25	1	100	0.7870 (0.0610)	0.0227 (0.1031)	0.248 (0.034)	24.7 (2.17)	0.9526 (0.0772)	22.17 (3.53)
		2	100	0.7860 (0.0573)	0.0046 (0.0989)	0.250 (0.030)	25.11 (1.69)	0.9404 (0.0731)	30.37 (4.30)
		3	100	0.7965 (0.0491)	−0.0012 (0.0883)	0.252 (0.030)	24.74 (1.36)	0.9536 (0.0807)	37.78 (5.11)
		4	100	0.8010 (0.0425)	−0.0169 (0.0850)	0.256 (0.025)	25.03 (1.34)	0.9402 (0.0669)	43.79 (4.59)
		5	100	0.8043 (0.0410)	0.0091 (0.0815)	0.255 (0.025)	24.97 (1.42)	0.9548 (0.0792)	48.38 (5.17)
		10	100	0.8097 (0.0318)	0.0016 (0.0666)	0.258 (0.022)	25.06 (1.09)	0.9512 (0.0714)	66.91 (5.68)

Dominance effect is absent in the simulation. Standard deviations are in parentheses.

TABLE 4

Comparisons of results of various sample sizes for joint analysis of F_2 and F_3 generations

a	h^2	Sample size	Estimates						
			Power	$\hat{a}$	$\hat{d}$	$\hat{h}^2$	cM̂	$\hat{\sigma}^2$	LOD
0.324	0.05	100	94	0.3323 (0.0614)	−0.0125 (0.1576)	0.063 (0.023)	25.58 (7.22)	0.9586 (0.0914)	7.36 (2.76)
		150	100	0.3218 (0.0442)	−0.0128 (0.1169)	0.054 (0.014)	25.57 (5.16)	0.9794 (0.0767)	10.23 (2.66)
		200	100	0.3207 (0.0408)	−0.0187 (0.0939)	0.053 (0.013)	25.35 (4.10)	0.9750 (0.0917)	13.50 (3.42)
		300	100	0.3216 (0.0302)	−0.0006 (0.0864)	0.052 (0.009)	25.52 (2.68)	0.9844 (0.0557)	20.02 (3.62)
0.816	0.25	100	100	0.8004 (0.0611)	−0.0009 (0.1207)	0.256 (0.033)	24.77 (1.82)	0.9459 (0.0900)	29.50 (4.17)
		150	100	0.8055 (0.0491)	0.0023 (0.1056)	0.261 (0.033)	24.74 (1.73)	0.9337 (0.0904)	45.17 (5.44)
		200	100	0.8096 (0.0493)	0.0157 (0.0786)	0.258 (0.028)	24.99 (1.02)	0.9527 (0.0685)	59.91 (6.22)
		300	100	0.8064 (0.0319)	0.0035 (0.0667)	0.253 (0.019)	25.00 (1.02)	0.9638 (0.0626)	88.41 (6.92)

Dominance effect of the QTL is absent in the simulation. Standard deviations are in parentheses.

the superiority of the joint analysis over either analysis. Although there appears to be no need for such a demonstration because the joint analysis in fact uses the whole sample, whereas both the "F_2 only" and the "$F_{2:3}$ only" analyses use different subsamples of the data, it is important to show the superiority of the $F_{2:3}$ only analysis over the F_2 only analysis. The comparisons were conducted under two levels of h^2 (0.05, 0.15) and three levels of m_j (1, 3, 5). The results are given in Table 5, where the joint analysis is indeed better than either analysis. Interestingly, the $F_{2:3}$ only analysis is always better than the F_2 only analysis, even in the situation where $m_j = 1$. It is understandable that $F_{2:3}$ is superior over F_2 for $m_j \geq 2$ because the $F_{2:3}$ analysis is equivalent to the analysis with a large sample size, mn, for $m_j = m$, $\forall\ j = 1, \ldots, n$. The fact that $F_{2:3}$ is superior over F_2 when $m_j = 1$ cannot be explained if we used the old method (adopted F_2 method). This can be explained only by the further segregation of the F_3 progeny derived from heterozygous F_2 parents. The new method developed in this study has captured this information and thus shows the increase of statistical power.

Multiple-QTL analysis: Having demonstrated the superiority of the new model that takes into account the further segregation of F_3 progeny under heterozygous F_2 parents over the adopted F_2 method, we now implement the Bayesian method that also takes into consideration this special feature to map multiple QTL. Eleven equally spaced markers were simulated on a single chromosome of length 100 cM. Three QTL were simulated with heritabilities of 0.05, 0.15, and 0.25 and locations at position 25, 55, and 85 cM, respectively. The overall means of F_2 and $F_{2:3}$ populations were set at 4 and 0, respectively. The residual variances of both populations were set at a value of 1.0. We simulated 100 F_2 plants, each with 20 F_3 progeny. The proposed MCMC sampler was run for 24,000 cycles in each of the MCMC analyses. The first 4000 samples (burn-in period) were discarded. To reduce serial correlation in the samples, we saved one observation in every 20 cycles of the simulation so that the total number of observations kept in the sample was 1000. The simulation was repeated 20 times. We first compared results of QTL mapping, using only F_3 progeny. Two models were compared, the adopted F_2 analysis, which ignores the further segregation of the F_3 progeny of the heterozygous F_2 parents, and the $F_{2:3}$ model that takes into consideration this special feature. The means and standard deviations of the parameters of interest across the 20 replications are given in Table 6. The three QTL were detected by both models in every single replicate. However, the $F_{2:3}$ model performed consistently better than the adopted F_2 method. The standard deviations of the estimated QTL positions are smaller for the $F_{2:3}$ model than for the adopted F_2 model. The estimated additive effects also show the same trend. The estimated dominance effects are seriously biased (downward) for the adopted F_2 model. As a result, the adopted F_2 model has smaller standard deviations due to scaling effect (standard deviation is proportional to the mean). Both models give a biased estimate of the residual error variance.

For the same setup of the experiment, we analyzed three different data sets using the multiple-QTL Bayesian method. The first data set contains only the 100 F_2 plants, called F_2 only, the second contains only the $F_{2:3}$ progeny (excluding the F_2 parents), called $F_{2:3}$ only, and the third contains both the F_2 parents and the $F_{2:3}$ progeny, called "both F_2 and $F_{2:3}$." Results of 20 repeated simulations are summarized in Table 7. We found that the result of the $F_{2:3}$ only analysis was consistently better than that of the F_2 only analysis. Combined analysis of F_2 and $F_{2:3}$ further improved the efficiency. The result of the F_2 only analysis shows that a sample size of 100 F_2 plants is not sufficient to generate a reliable result, but 100 $F_{2:3}$ families each with 20 progeny are sufficient.

We have done a further simulation using 200 F_2 plants each with 10 $F_{2:3}$ progeny. The QTL intensity profiles from a single replicated simulation experiment are shown in Figure 2. This single experiment also demonstrates the progressive improvement using $F_{2:3}$ and com-

TABLE 5

Comparison of results using data of the F_2 only, the $F_{2:3}$ only, and the both F_2 and $F_{2:3}$, respectively, in QTL interval mapping

Plant no. per family	a	h^2	Population	Estimates						
				Power	$\hat{a}$	$\hat{d}$	$\hat{h}^2$	cM̂	$\hat{\sigma}^2$	LOD
1	0.324	0.05	F_2	37	0.3240 (0.1365)	0.0176 (0.2116)	0.070 (0.033)	31.59 (21.69)	0.9582 (0.1617)	2.85 (1.45)
			$F_{2:3}$	45	0.3197 (0.1459)	−0.0239 (0.2109)	0.074 (0.038)	29.43 (20.07)	0.9039 (0.2076)	2.88 (1.69)
			$F_2 + F_{2:3}$	68	0.3255 (0.0916)	−0.0030 (0.1545)	0.061 (0.027)	25.10 (10.48)	0.9681 (0.1155)	4.37 (2.31)
	0.594	0.15	F_2	98	0.5959 (0.1033)	0.0217 (0.1436)	0.162 (0.046)	24.97 (5.00)	0.9646 (0.0924)	7.09 (2.19)
			$F_{2:3}$	98	0.6106 (0.1150)	−0.0010 (0.1650)	0.182 (0.063)	25.95 (4.79)	0.9020 (0.1188)	7.30 (2.60)
			$F_2 + F_{2:3}$	100	0.6015 (0.0760)	0.0281 (0.1197)	0.165 (0.038)	25.20 (2.92)	0.9483 (0.0757)	13.25 (3.48)
3	0.324	0.05	F_2	41	0.3328 (0.1393)	−0.0395 (0.2065)	0.071 (0.032)	29.45 (18.63)	0.9729 (0.1405)	2.99 (1.45)
			$F_{2:3}$	99	0.3325 (0.0578)	0.0099 (0.0998)	0.056 (0.016)	24.81 (5.54)	0.9495 (0.0930)	6.75 (2.31)
			$F_2 + F_{2:3}$	100	0.3330 (0.0457)	−0.0105 (0.1133)	0.057 (0.015)	24.34 (4.86)	0.9806 (0.0701)	8.39 (2.49)
	0.594	0.15	F_2	98	0.5847 (0.1073)	0.0114 (0.1667)	0.157 (0.049)	25.12 (5.75)	0.9824 (0.1074)	6.86 (2.29)
			$F_{2:3}$	100	0.5810 (0.0628)	−0.0074 (0.0920)	0.158 (0.037)	25.13 (2.88)	0.9186 (0.1095)	16.56 (3.84)
			$F_2 + F_{2:3}$	100	0.5795 (0.0525)	−0.0054 (0.0990)	0.151 (0.027)	25.20 (2.26)	0.9669 (0.0688)	22.14 (4.49)
5	0.324	0.05	F_2	42	0.3473 (0.1151)	−0.0047 (0.1802)	0.072 (0.029)	29.44 (16.19)	0.9621 (0.1684)	2.96 (1.29)
			$F_{2:3}$	100	0.3271 (0.0430)	−0.0100 (0.0636)	0.055 (0.023)	25.09 (4.36)	0.9405 (0.0995)	10.01 (2.51)
			$F_2 + F_{2:3}$	100	0.3294 (0.0407)	−0.0149 (0.0889)	0.056 (0.013)	25.25 (4.23)	0.9724 (0.0686)	11.76 (2.80)
	0.594	0.15	F_2	97	0.6051 (0.1244)	−0.0065 (0.1705)	0.167 (0.047)	25.05 (6.49)	0.9728 (0.1350)	7.21 (2.26)
			$F_{2:3}$	100	0.5913 (0.0486)	−0.0015 (0.0650)	0.160 (0.023)	25.10 (1.77)	0.9210 (0.1075)	24.42 (3.59)
			$F_2 + F_{2:3}$	100	0.5915 (0.0414)	−0.0095 (0.0923)	0.155 (0.020)	25.15 (1.81)	0.9684 (0.0711)	30.53 (4.00)

Dominance effect of the QTL was absent in the simulation. The overall means of F_2 and $F_{2:3}$ were the same in this simulation. Standard deviations are in parentheses.

TABLE 6

Comparisons of the adopted F_2 model and the $F_{2:3}$ model for mapping multiple QTL in an $F_{2:3}$ design using the Bayesian method

				Estimates			
a	d	h^2	Method	$c\hat{M}$	$\hat{a}$	$\hat{d}$	$\hat{\sigma}^2$
0.4264	0.0000	0.05	Adopted F_2	23.65 (5.74)	0.3316 (0.1336)	−0.0002 (0.0035)	1.2683 (0.1368)
			$F_{2:3}$ model	24.35 (4.60)	0.3268 (0.1149)	0.0058 (0.0348)	1.2364 (0.1756)
0.0000	1.0445	0.15	Adopted F_2	54.25 (3.58)	−0.0094 (0.0354)	0.4421 (0.1704)	
			$F_{2:3}$ model	54.50 (2.73)	0.0082 (0.0219)	1.0901 (0.2447)	
0.6742	0.9535	0.25	Adopted F_2	84.75 (2.53)	0.6122 (0.1985)	0.8347 (0.2848)	
			$F_{2:3}$ model	85.40 (2.15)	0.6495 (0.0734)	0.9821 (0.3528)	

The true positions of the three QTL are 25, 55, and 85 cM, respectively. The true residual variance on the individual plant basis is $\sigma^2 = 1$. Standard deviations are in parentheses.

bined F_2 and $F_{2:3}$ over F_2 only analysis. This number is what we expect to see in real data analysis.

DISCUSSION

We developed a new method to map QTL using the $F_{2:3}$ design of experiments. The current method has ignored the mixture distribution of the average phenotypic value of the F_3 progeny derived from heterozygous F_2 plants. This adopted F_2 method is inferior to the new method developed here because of the information loss. We also developed a method to combine F_2 and $F_{2:3}$ data to perform a joint mapping. Such a combined analysis has never been proposed. The important novel contribution of this study actually comes from both the conceptual contribution of the mixture distribution of the F_3 progeny derived from the heterozygous F_2 plants and the technical solution to take into account this mixture distribution. Simulation experiments showed substantial increase in both the statistical power and the efficiency of parameter estimation. The method is a plant version of the "daughter design" developed by animal breeders (Weller *et al.* 1990; Bovenhuis and Weller 1994; Mackinnon and Weller 1995; Ron *et al.* 2001). We learned that the daughter design can be more efficient than the traditional analysis because the "average value" of the daughters may be a better measurement of the breeding value of the sire. However, the animal breeders should also learn from this study that if the variance of the "average daughter" can be incorporated into the analysis, additional power increase can be achieved. This variance is important in power increase. Let us take the $F_{2:3}$ design as an example. In the F_3 progeny, half of the plants derived from the heterozygous F_2 will be fixed to either homozygous genotype. Overall, the F_3 progeny will have more extreme genotypes than the intermediate (heterozygous) genotypes. This has increased the additive genetic variance compared with that of the F_2 population. Therefore, the $F_{2:3}$ design is more powerful than the F_2 design, even if only one progeny is used per F_2 plant. This additional information has not been captured in the existing method. The new method can be extended to an $F_{2:y}$ design for $y > 3$. In this case, the heterozygous individual in generation y has been further decreased to $(1/2)^{y-1}$ and thus the genetic variance in generation y is even greater than that in generation 3. For example, if $y = 4$, one can expect to have a ratio of (7/16):(2/16):(7/16) for the three genotypes in the F_4 population. The expected genetic variance in such a population is of course larger than that of the F_3 generation, which has a ratio of (3/8):(2/8):(3/8). Further extension to $y \to \infty$ leads to the RIL design, which has a ratio of (1/2):(0):(1/2). The genetic variance in the RIL population has reached its upper limit. However, the adopted F_2 method will lose half of the sample size because the heterozygous markers will not be informative in the analysis. The new method will recover much more information by taking this information into account. Of course, the optimal method is to genotype RIL lines also, which is $F_{\infty:\infty}$, a true RIL design.

Understanding the mixture distribution allows us to develop an exact EM algorithm for solving the MLE. We were able to define $\bar{x}_{1j} = m_j^{-1}(m_{11} - m_{22})$ and $\bar{x}_{2j} = m_j^{-1}m_{12}$ and further take advantage of the multinomial distribution of the m's. Although the exact EM is not suitable for large m_j, the approximate method by assuming joint normal distribution of $\bar{x}_{1j}$ and $\bar{x}_{2j}$ has significantly reduced the computing burden with almost no loss in power and efficiency. This approximation is different from the adopted F_2 method, which is based on a wrong model and completely ignored the mixture distribution. We acknowledge that the mixture distribution can be approximated by a heterogeneous residual variance model (Zhang *et al.* 2003), which is for segregation analysis, not for QTL mapping. We modified the heterogeneous residual variance model of Zhang *et al.* (2003) to map QTL using the general approach developed by Xu (1996). The result was almost identical to the result reported here.

TABLE 7

Comparison of results using data of the F_2 only, the $F_{2:3}$ only, and the both F_2 and $F_{2:3}$ for mapping multiple QTL

Population	a	d	h^2	Estimates					
				b_0	b	$\hat{cM}$	$\hat{a}$	$\hat{d}$	$\hat{\sigma}^2$
F_2									
QTL1	0.4264	0.0000	0.05	—		26.75 (10.87)	0.1727 (0.1950)	0.0149 (0.0401)	
QTL2	0.0000	1.0445	0.15	—	4.3427 (0.2473)	57.70 (4.95)	0.0487 (0.1249)	0.6801 (0.4712)	1.1583 (0.2403)
QTL3	0.6742	0.9535	0.25	—		83.85 (4.94)	0.3605 (0.2006)	0.4507 (0.3891)	
$F_{2:3}$									
QTL1	0.4264	0.0000	0.05		—	24.35 (4.60)	0.3268 (0.1149)	0.0058 (0.0348)	
QTL2	0.0000	1.0445	0.15	−0.0083 (0.0494)	—	54.50 (2.73)	0.0082 (0.0219)	1.0901 (0.2447)	1.2364 (0.1756)
QTL3	0.6742	0.9535	0.25		—	85.40 (2.15)	0.6495 (0.0734)	0.9821 (0.3528)	
$F_2 + F_{2:3}$									
QTL1	0.4264	0.0000	0.05			24.35 (4.68)	0.3211 (0.0730)	−0.0038 (0.0169)	$\hat{\sigma}^2_2$: 1.0278 (0.1592)
QTL2	0.0000	1.0445	0.15	−0.0044 (0.0413)	4.0042 (0.1136)	53.65 (2.29)	0.0013 (0.0065)	1.0259 (0.1452)	$\hat{\sigma}^2_3$: 1.1484 (0.1919)
QTL3	0.6742	0.9535	0.25			84.55 (1.36)	0.6685 (0.0567)	1.0018 (0.1459)	

The true positions of the three QTL are 25, 55, and 85 cM, respectively. The true values of b_0 and b are 0 and 4, respectively. Standard deviations are in parentheses.

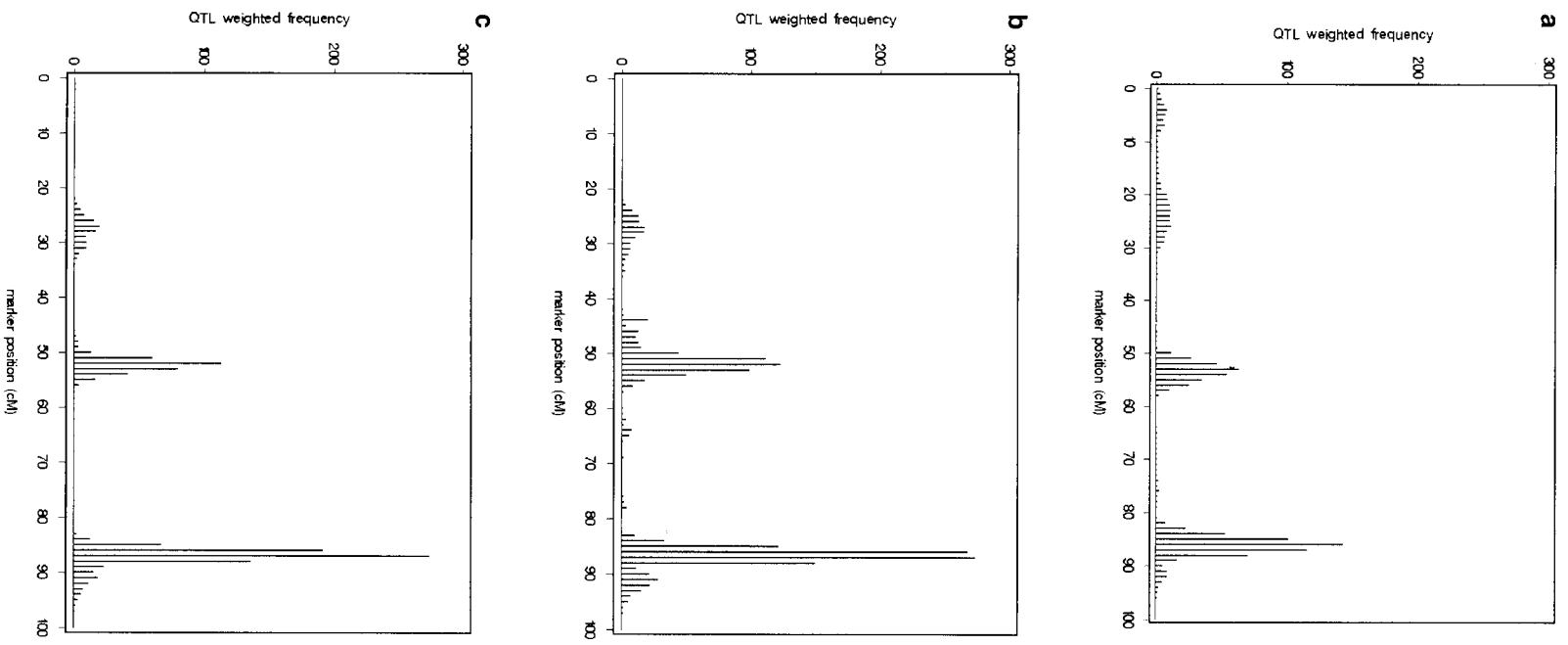

FIGURE 2.—QTL intensity profiles of the multiple-QTL Bayesian analysis from a single simulation experiment of 200 F_2 plants each with 10 F_3 progeny. (a) F_2 only; (b) $F_{2:3}$ only; and (c) both F_2 and $F_{2:3}$. The true positions of the three simulated QTL are 25, 55, and 85 cM along the chromosome and the corresponding effects (expressed as variance explained by QTL) are 0.05, 0.15, and 0.25, respectively.

The mixture model of the $F_{2:3}$ progeny under heterozygous F_2 parents was evaluated largely under the single-QTL model implemented via the EM algorithm. Our EM analysis serves as a tool only to evaluate the improved model, rather than as a tool to map QTL in practice. In actual QTL-mapping experiments, the number of QTL is most likely greater than one and is usually unknown. Therefore, a multiple-QTL model must be adopted in real data analysis. The EM algorithm is not sufficient to cope with the multiple-QTL model. The Bayesian method implemented via the MCMC algorithm is the ideal tool for this. Therefore, we developed a Bayesian method for actual implementation of the mixture distribution of the $F_{2:3}$ progeny. To avoid all problems encountered in model selection for the multiple-QTL model, we adopted a model-selection-free approach (Xu 2003) by including all potential QTL effects. The oversaturated model was then handled in a shrinkage approach. This new approach of QTL mapping may represent a new direction in developing QTL-mapping methodology.

We are greatly indebted to two anonymous reviewers for their comments on the first version of the manuscript. This research was supported by the National Institutes of Health grant R01-GM55321 and the United States Department of Agriculture National Research Initiative Competitive Grants Program 00-35300-9245 to S.X.

LITERATURE CITED

Austin, D. F., and M. Lee, 1996 Comparative mapping in $F_{2:3}$ and $F_{6:7}$ generations of quantitative trait loci for grain yield and yield component in maize. Theor. Appl. Genet. **92:** 817–826.

Bovenhuis, H., and J. I. Weller, 1994 Mapping and analysis of dairy cattle quantitative traits loci by maximum likelihood methodology using milk protein genes as genetic markers. Genetics **136:** 267–280.

Chapman, A., V. R. Pantalone, A. Ustun, F. L. Allen, D. Landau-Ellis *et al.*, 2003 Quantitative trait loci for agronomic and seed quality traits in an F_2 and $F_{4:6}$ soybean population. Euphytica **129:** 387–393.

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

Fisch, R. D., M. Ragot and G. Gay, 1996 A generalization of the mixture model in the mapping of quantitative trait loci for progeny from a biparental cross of inbred lines. Genetics **143:** 571–577.

Gai, J. Y., Y. M. Zhang and J. K. Wang, 2003 *The Genetic System of Quantitative Traits in Plants.* Chinese Science Press, Beijing.

Geman, S., and D. Geman, 1984 Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. IEEE Trans. Patt. Anal. Machine Intell. **6:** 721–741.

Green, P. J., 1995 Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika **82:** 711–732.

Hastings, W. K., 1970 Monte Carlo sampling methods using Markov chains and their applications. Biometrika **57:** 97–109.

Hoeschele, I., and P. M. Vanraden, 1993 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., D. L. Johnson and J. A. M. V. Arendonk, 1998 A mixture model approach to the mapping of quantitative trait loci in complex populations with an application to multiple cattle families. Genetics **148:** 391–399.

Jiang, C. J., and Z-B. Zeng, 1997 Mapping quantitative trait loci with dominant and missing markers in various crosses from two inbred lines. Genetica **101:** 47–58.

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

Knapp, S. J., 1991 Using molecular markers to map multiple quantitative trait loci: models for backcross, recombinant inbred, and doubled haploid progeny. Theor. Appl. Genet. **81:** 333–338.

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

Mackinnon, M. J., and J. I. Weller, 1995 Methodology and accuracy of estimation of quantitative trait loci parameters in a half-sib design using maximum likelihood. Genetics **141:** 755–770.

Mather, K., and J. L. Jinks, 1982 *Biometrical Genetics*, Ed. 2. Chapman & Hall, London.

Piepho, H. P., 2001 A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics **157:** 425–432.

Rao, S., and S. Xu, 1998 Mapping quantitative trait loci for ordered categorical traits in four-way crosses. Heredity **81:** 214–224.

Ron, M., D. Kliger, E. Feldmesser, E. Seroussi, E. Rzra *et al.*, 2001 Multiple quantitative trait locus analysis of bovine chromosome 6 in the Israeli Holstein population by a daughter design. Genetics **159:** 727–735.

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

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

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

Sorensen, D., and D. Gianola, 2002 *Likelihood, Bayesian, and MCMC Methods in Quantitative Genetics.* Springer-Verlag, New York.

Stephens, D. A., and R. D. Fisch, 1998 Bayesian analysis of quantitative trait locus data using reversible jump Markov chain Monte Carlo. Biometrics **54:** 1334–1347.

Thaller, G., and I. Hoeschele, 1996 A Monte Carlo method for Bayesian analysis of linkage between single markers and quantitative trait loci.1. Methodology. Theor. Appl. Genet. **93:** 1161–1166.

Uimari, P., and I. Hoeschele, 1997 Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. Genetics **146:** 735–743.

Weller, J. I., Y. Kashi and M. Soller, 1990 Power of "daughter" and "granddaughter" designs for genetic mapping of quantitative traits in dairy cattle using genetic markers. J. Dairy Sci. **73:** 2525–2537.

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

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

Xu, S., 1998b Mapping quantitative trait loci using multiple families of line crosses. Genetics **148:** 517–524.

Xu, S., 2002 *QTL Analysis in Plants.* Humana Press, Totowa, NJ.

Xu, S., 2003 Estimating polygenic effects using markers of the entire genome. Genetics **163:** 789–801.

Xu, C., X. He and S. Xu, 2003 Mapping quantitative trait loci underlying triploid endosperm traits. Heredity **90:** 228–235.

Yi, N., and S. Xu, 2001 Bayesian mapping of quantitative trait loci under complicated mating designs. Genetics **157:** 1759–1771.

Zeng, Z-B., 1993 Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc. Natl. Acad. Sci. USA **90:** 10972–10976.

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

Zhang, T., Y. Yuan, J. Yu, W. Z. Guo and R. J. Kohel, 2003 Molecular tagging of a major QTL for fiber strong in upland cotton and its marker-assisted selection. Theor. Appl. Genet. **106:** 262–268.

Zhang, Y. M., J. Y. Gai and Y. H. Yang, 2003 The EIM algorithm in the joint segregation analysis of quantitative traits. Genet. Res. **81:** 157–163.

Communicating editor: J. B. Walsh