# A New Simple Method for Improving QTL Mapping Under Selective Genotyping

**Hsin-I Lee,* Hsiang-An Ho,* and Chen-Hung Kao*,†,1**
*Institute of Statistical Science, Academia Sinica, Taipei 11529, Taiwan, Republic of China and †Division of Biometry, Institute of Agronomy, National Taiwan University, Taipei 10617, Taiwan, Republic of China

**ABSTRACT** The selective genotyping approach, where only individuals from the high and low extremes of the trait distribution are selected for genotyping and the remaining individuals are not genotyped, has been known as a cost-saving strategy to reduce genotyping work and can still maintain nearly equivalent efficiency to complete genotyping in QTL mapping. We propose a novel and simple statistical method based on the normal mixture model for selective genotyping when both genotyped and ungenotyped individuals are fitted in the model for QTL analysis. Compared to the existing methods, the main feature of our model is that we first provide a simple way for obtaining the distribution of QTL genotypes for the ungenotyped individuals and then use it, rather than the population distribution of QTL genotypes as in the existing methods, to fit the ungenotyped individuals in model construction. Another feature is that the proposed method is developed on the basis of a multiple-QTL model and has a simple estimation procedure similar to that for complete genotyping. As a result, the proposed method has the ability to provide better QTL resolution, analyze QTL epistasis, and tackle multiple QTL problem under selective genotyping. In addition, a truncated normal mixture model based on a multiple-QTL model is developed when only the genotyped individuals are considered in the analysis, so that the two different types of models can be compared and investigated in selective genotyping. The issue in determining threshold values for selective genotyping in QTL mapping is also discussed. Simulation studies are performed to evaluate the proposed methods, compare the different models, and study the QTL mapping properties in selective genotyping. The results show that the proposed method can provide greater QTL detection power and facilitate QTL mapping for selective genotyping. Also, selective genotyping using larger genotyping proportions may provide roughly equivalent power to complete genotyping and that using smaller genotyping proportions has difficulties doing so. The R code of our proposed method is available on http://www.stat.sinica.edu.tw/chkao/.

THE data in the QTL mapping study are usually composed of two parts, phenotypic trait values and marker genotypes, in the individuals, and the cost of producing data includes both phenotyping and genotyping costs. The cost ratio of the phenotyping to genotyping may vary significantly depending on the traits and species in studies. For a fixed budget and time frame in the study, both costs must be considered and properly allocated to make the study optimally cost effective. If the total cost is not of primary concern in QTL experiments, all individuals in the entire sample will be genotyped and phenotyped for QTL analysis. However, QTL experiments are usually conducted under a limited budget, and researchers may not be allowed to genotype and phenotype a large amount of individuals for QTL analysis. It is hence necessary to make a reasonable and effective allocation for the genotyping and phenotyping costs in the experiment. The selective genotyping approach has been known as a cost-effective strategy for reducing genotyping work and still has the ability to maintain efficiency in QTL detection (Lebowitz *et al.* 1987; Lander and Botstein 1989). This approach is intended to select individuals with extreme (high and low) phenotypic values for genotyping and keep the remaining individuals ungenotyped in the entire sample. Later, several statistical methods have been proposed to study QTL mapping under selective genotyping (Darvasi and Soller 1992; Muranty and Goffinet 1997; Henshall and Goddard 1999; Xu and Vogl 2000; Manichaikul *et al.* 2007). They again confirmed that selective genotyping can achieve reasonable power and precision in detecting QTL as compared to complete

Copyright © 2014 by the Genetics Society of America
doi: 10.1534/genetics.114.168385
Manuscript received July 15, 2014; accepted for publication September 16, 2014; published Early Online September 22, 2014.
[1]Corresponding author: Institute of Statistical Science, Academia Sinica, National Taiwan University, Taipei 115529, Taiwan, Republic of China.
E-mail: chkao@stat.sinica.edu.tw

genotyping but at the expense of moderate increase in the number of phenotyped individuals. As a larger number of individuals must be phenotyped first, before the required number of extreme individuals can be collected for genotyping, selective genotyping will be most suitable for the cases in which the phenotyping cost is relatively inexpensive compared with the genotyping cost. Many economically and biologically important traits, such as flowering time, fruit size and shape, crop (meat) yield or quality, growth in height and weight, plant stress and disease resistance, survival time, blood pressure, and body mass index in human, can be obtained at relatively low cost, and QTL mapping of these traits by using selective genotyping can be managed to be more cost effective than that by using complete genotyping. Consequently, although genotyping cost has been dropping recently, selective genotyping has been still widely employed in QTL mapping for improving these traits and understanding their genetic basis in several plant and animal species (Abdel-Haleem *et al.* 2011; Lu *et al.* 2013; Miller *et al.* 2013; Fontanesi *et al.* 2014;). The keyword search using Google Scholar also reveals that selective genotyping remains popular and is more frequently used in genetics studies in recent years. The reasons may be that as marker genotyping becomes cheaper, more researchers are attracted to the QTL mapping analysis, but still face the situation of insufficient budgets to fully cover the expense of complete genotyping, and selective genotyping can become an alternative, cost-effective choice that allows maintenance of similar efficiency to complete genotyping in QTL mapping.

In general, the statistical methods of QTL mapping for selective genotyping can be grouped into two major types according to whether their models take the ungenotyped individuals (ungenotyping data) into account in the analysis. Methods of the first type use only the genotyped individuals (genotyping data) and exclude the ungenotyping data to develop statistical models for QTL analysis (Darvasi and Soller 1992; Henshall and Goddard 1999; Xu and Vogl 2000). Darvasi and Soller (1992) calculated the trait means of the QTL genotypes from the selected tails and constructed a *t*-test for their difference to determine linkage between a marker and a QTL in the backcross population. Henshall and Goddard (1999) used a logistic regression approach to the analysis of selective genotyping data by treating phenotypes as independent variables and genotypes as dependent variables for QTL mapping. Xu and Vogl (2000) developed a truncated normal mixture model for the interval mapping procedure under the framework of selective genotyping in QTL detection. Methods of the second type consider both genotyped and ungenotyped individuals (full data) from selective genotyping for QTL analysis (Muranty and Goffinet 1997; Ronin *et al.* 1998; Xu and Vogl 2000). By analyzing full data, Muranty and Goffinet (1997) adopted a normal mixture model to detect QTL under selective genotyping in the backcross population. Ronin *et al.* (1998) proposed a mixture model for the study of interval mapping of two QTL under selective genotyping. Xu and Vogl (2000) also used a normal mixture model to tackle the issue of QTL mapping under selective genotyping in the $F_2$ population. In their normal mixture models, the mixing proportions for different QTL genotypes of the genotyped individuals are determined by the flanking markers as usual, and those of the ungenotyped individuals are assigned by the population frequencies (*e.g.*, 1/2 and 1/2 in the backcross, and 1/4, 1/2, and 1/4 in the $F_2$ population for a single-QTL model). When comparing the two types of methods in selective genotyping, Xu and Vogl (2000) suggested that whenever possible, the analyses on full data are preferred over those on genotyping data, since the formers tend to provide improved estimations and greater test statistics for QTL parameters.

In this article, we develop a novel and simple statistical method on the basis of a normal mixture model to analyze full data from selective genotyping for QTL detection. Compared to the existing methods (Muranty and Goffinet 1997; Ronin *et al.* 1998; Xu and Vogl 2000), the main novelties of our proposed method are twofold. The first novelty is that we obtain the proportions of QTL genotypes in the ungenotyped individuals by deducting the expected QTL frequencies in the genotyped individuals from their population frequencies. Then these proportions instead of the population QTL frequencies, as in the existing methods, are used to model the ungenotyped individuals in model construction, so that the proposed model can fit better to the data and yield better performance in selective genotyping. The second novelty is that our proposed model is developed on the basis of a multiple-QTL model, as has been done in QTL mapping for complete genotyping (Kao *et al.* 1999). The multiple-QTL model approach can take multiple QTL into account to control more genetic variation for improving QTL detection. As a result, with the novelties, our proposed method can provide better resolution, analyze epistasis, and tackle multiple QTL problems in QTL mapping under selective genotyping. In addition, a truncated normal mixture model based on the multiple-QTL model is developed when only genotyping data are used in the analysis. We then compare the differences in QTL mapping between the two types of selective genotyping methods and investigate their properties under the multiple-QTL framework, as their notable differences blurred in the single-QTL framework may appear in the multiple-QTL framework. The threshold values of selective genotyping for different models are also investigated. Simulation studies are conducted to evaluate the proposed method, compare different models, and study their properties in QTL mapping under selective genotyping.

## The Statistical Methods for Selective Genotyping

In selective genotyping, the selected individuals for genotyping are those with high or low trait values in the two extremes, and the remaining unselected individuals are those with average trait values in the middle. Therefore, the full data generated from selective genotyping can be split into two parts: genotyping data containing phenotypically extreme

individuals with marker genotypes and ungenotyping data containing intermediate individuals without marker genotypes. As described in the Introduction, statistical QTL mapping methods for analyzing selective genotyping data can either take full data or take just genotyping data into account in their models for QTL detection. When considering full data in the analysis, it is important to know that the genotypic frequencies of QTL in the ungenotyping (or genotyping) data are no longer the same as those in the full data, and they will depend on the underlying QTL parameters and population structures. With such facts, we obtain the frequencies of QTL genotypes in the ungenotyped individuals and use them, rather than the population frequencies of QTL genotypes as in the existing methods, to model the ungenotyped individuals and then to propose a new normal mixture model for QTL mapping when full data are used in the analysis. As opposed to our proposed method, the existing methods using normal mixture model will hereinafter be called population frequency-based (PFB) methods. Furthermore, we also develop a truncated normal mixture model for QTL mapping when only genotyping data are used in the analysis. Both types of models are developed on the basis of a multiple-QTL approach, so that they can be compared under multiple-QTL analyses and be used to deal with the multiple-QTL problem. In the following, we first investigate the genotypic distributions in the genotyped and ungenotyped individuals, then present the truncated model, and finally outline our proposed method for selective genotyping.

### *Genotypic distributions in the genotyped and ungenotyped individuals*

In the $F_2$ population, a QTL, $Q$, under consideration has three possible genotypes, $QQ$, $Qq$, and $qq$ with expected frequencies 1/4, 1/2, and 1/4, respectively. Their genotypic values, $G_2$, $G_1$, and $G_0$, can be related to genotypic mean ($\mu$), additive effect ($a$), and dominance effect ($d$) by

$$G = \begin{pmatrix} G_2 \\ G_1 \\ G_0 \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \mu + \begin{pmatrix} 1 & -\frac{1}{2} \\ 0 & \frac{1}{2} \\ -1 & -\frac{1}{2} \end{pmatrix} \begin{pmatrix} a \\ d \end{pmatrix} = 1_3\ \mu + \mathbf{DE}, \tag{1}$$

where $\mathbf{D}$ is known as the genetic design matrix for characterizing the genetic effects of QTL in the vector $\mathbf{E}$. Under such settings, the trait value of an individual, $y_i$, affected by $Q$ may have three possible distributions, *i.e.*, $y_i|QQ \sim N(\mu_1, \sigma^2)$, $y_i|Qq \sim N(\mu_2, \sigma^2)$, or $y_i|qq \sim N(\mu_3, \sigma^2)$, where $\mu_1$, $\mu_2$, and $\mu_3$ are the corresponding genotypic values. If a selective genotyping approach with genotyping proportion $\theta$ ($\theta/2$ from each tail) is conducted on a sample with size $N$, $N \times \theta$ individuals from the two tails will have both trait and marker information, and $N \times (1 - \theta)$ individuals from the middle will have only trait information. Let $T_R$ and $T_L$ be the right and left truncated points so that $P(y_i < T_L) = P(y_i > T_R) = \theta/2$. For $N$ large enough, the expected genotypic frequencies in the ungenotyped individuals are $w_j \times [\Phi((T_R - \mu_j)/\sigma) - \Phi((T_L - \mu_j)/\sigma)]$, $j = 1, 2, 3$, where $\Phi(\cdot)$ denotes a standard normal cumulative distribution function, and $w_1 = 1/4$, $w_2 = 1/2$, and $w_3 = 1/4$ are the expected genotypic frequencies in the whole population. Then, the proportions of the three QTL genotypes in the ungenoytyped individuals can be straightforwardly obtained by

$$\kappa_j = \frac{w_j\left[\Phi\left(\tau_{2j}\right) - \Phi\left(\tau_{1j}\right)\right]}{\sum_{k=1}^{3} w_k[\Phi(\tau_{2k}) - \Phi(\tau_{1k})]}, \quad j = 1, 2, 3, \tag{2}$$

where $\tau_{1j} = (T_L - \mu_j)/\sigma$ and $\tau_{2j} = (T_R - \mu_j)/\sigma$. As $\kappa_j$'s are functions of truncated points ($T_L$ and $T_R$), genetic parameters ($\mu$, $a$, $d$, and $\sigma$), and population frequencies ($w_j$'s), the values of $\kappa_j$'s depend on the factors such as genotyping proportions, heritability, sizes and modes of QTL actions, and population structure. For example, if $a = d = 1$, $h^2 = 0.5$, and $\theta = 0.5$ ($P(y_i < T_L) = P(y_i > T_R) = 0.25$), the two truncated points are $T_L \cong -0.785$ and $T_R \cong 0.873$. The genotypic frequencies of $QQ$, $Qq$, and $qq$ are about 0.017 (0.083), 0.034 (0.166), and 0.199 (0.001) among the genotyped individuals in the left (right) extreme, about 0.150, 0.299, and 0.051 among the ungenotyped individuals, respectively. Then, $\kappa_1 = 0.299$ ($\cong$0.150/0.5), $\kappa_2 = 0.599$ ($\cong$0.299/0.5), and $\kappa_3 = 0.101$ ($\cong$0.051/0.5), respectively, by Equation 2. Similarly, these genotypic proportions are 0.069 (0.332), 0.137 (0.665) , and 0.794 (0.003) for $y_i < T_L$ ($y_i > T_R$), respectively. Essentially, it is worth noting that the expected proportions of the three QTL genotypes in the genotyped and ungenotyped samples are no longer $w_j$'s like those in the whole sample.

Equation 2 can be easily extended to the multiple-QTL case for obtaining the genotypic proportions in the ungenotyped population. For example, in the case of two QTL, $Q_1$ and $Q_2$, there are nine possible QTL genotypes, $Q_1Q_1Q_2Q_2$, $Q_1Q_1Q_2q_2$, $Q_1Q_1q_2q_2$, $Q_1q_1Q_2Q_2$, $Q_1q_1Q_2q_2$, $Q_1q_1q_2q_2$, $q_1q_1Q_2Q_2$, $q_1q_1Q_2q_2$, $q_1q_1q_2q_2$, respectively. Their expected population frequencies are $w_1 = (1 - r)^2/4$, $w_2 = r(1 - r)/2$, $w_3 = r^2/4$, $w_4 = r(1 - r)/2$, $w_5 = (1 - r)^2/2 + r^2/2$, $w_6 = r(1 - r)/2$, $w_7 = r^2/4$, $w_8 = r(1 - r)/2$, $w_9 = (1 - r)^2/4$, respectively, where $r$ is the recombination fraction between the two QTL. For $r = 0.3$, the values of $w_j$'s are ~0.123, 0.105, 0.023, 0.105, 0.290, 0.105, 0.023, 0.105, and 0.123, respectively. If the two QTL have effects ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) and $h^2 = 0.5$, the values of $\kappa_j$'s among the ungenotyped individuals are ~0.040, 0.109, 0.028, 0.129, 0.377, 0.130, 0.029, 0.059 and 0.100, respectively, for $\theta = 0.5$. If the two QTL are unlinked ($r = 0.5$), the values of $w_j$'s are ~0.063, 0.125, 0.063, 0.125, 0.250, 0.125, 0.063, 0.125, and 0.063, respectively, and the values of $\kappa_j$'s are 0.043, 0.124, 0.067, 0.131, 0.270, 0.133, 0.067, 0.106, and 0.059, respectively. The differences between $\kappa_j$'s and $w_i$'s may be minor, but can be significant for some genotypes. In general, greater genotyping proportions, higher heritability, and tighter linkage will cause larger differences between the

values of $w_j$'s and $\kappa_j$'s. Our proposed QTL mapping method for selective genotyping intends to use $\kappa_j$'s rather than $w_j$'s to model the relationship between the trait values and unobservable genotypes in the ungenotyped individuals.

The genetic model for multiple QTL, say $m$ QTL, can be easily obtained from Equation 1 by augmenting the dimensions of the genetic design matrix according to their effects under consideration. Then, on the basis of the genetic model, the statistical model for fitting these $m$ QTL, $Q_1$, $Q_2$, ..., $Q_m$, without epistasis at given positions within the $m$ separate marker intervals, ($M_1$,$N_1$), ($M_2$,$N_2$), ..., ($M_m$, $N_m$), can be written as

$$y_i = \mu + \sum_{k=1}^{m}\left(a_k x_{ik}^* + d_k z_{ik}^*\right) + \epsilon_i, i = 1, 2, \cdots, n, \tag{3}$$

where $x_{ik}^*$'s and $z_{ik}^*$'s are coded variables for the additive and dominance effects, $a_k$'s and $d_k$'s, for $Q_k$'s, $k$ = 1, 2, ..., $m$, $y_i$ is the quantitative trait value of the $i$th individual, and $\varepsilon_i$ is a random error and assumed to follow $N(0, \sigma^2)$. Note that $x_{ik}^*$ and $z_{ik}^*$ associated with $a_k$ and $d_k$ are coded as (1, −1/2), (0, 1/2), and (−1, −1/2) for genotypes $Q_kQ_k$, $Q_kq_k$, and $q_kq_k$, respectively, and the above model can be easily extended to the model with epistasis by introducing the product terms as the terms for epistasis. Under complete genotyping, the likelihood function of the statistical model for the parameters $\Theta$ is a mixture of $3^m$ normals as

$$L(\Theta|Y,X) = \prod_{i=1}^{n}\sum_{j=1}^{3^m} p_{ij} f\left(y_i|\mu_j, \sigma^2\right), \tag{4}$$

where $f(y_i|\mu_j, \sigma^2)$ is a normal p.d.f. with mean $\mu_j$ and variance $\sigma^2$, $\mu_j$'s correspond to the genotypic values of the $3^m$ QTL genotypes, and $p_{ij}$'s are the mixing proportions inferred from their flanking marker genotypes. If the statistical model is applied to analyze the data from selective genotyping, the likelihood will be different and depend on whether all individuals or just genotyping individuals are considered in the model as described below.

### *Model to analyze only genotyping data*

Under selective genotyping, suppose that, among the $n$ individuals, $n_s$ individuals with extreme trait values ($n_s$/2 each from the upper and lower extremes) are selected for marker genotyping, and the remaining $n_u$ ($n_u = n - n_s$) individuals are not genotyped. If only the genotyped individuals from the two extremes are utilized in the analysis, data of this sort are called centrally truncated data and the methods of analyzing truncated data can be applied to the analyses (Cohen 1991). Xu and Vogl (2000) incorporated the truncated model into the mixture structure of interval mapping framework to propose a truncated normal mixture model for QTL analysis. They pointed out that the maximization of the truncated normal mixture likelihood is a challenging task, and they used an EM algorithm to obtain the maximum likelihood estimates (MLE) of the QTL parameters for the model. An investigation of detection of a single QTL under selective genotyping was performed in their analysis. Here, we provide an alternative version of the EM algorithm for obtaining the MLE of the truncated normal mixture model and use it to address more complicated issues involving mapping multiple QTL and analyzing their epistasis. For $n_s$ genotyped individuals, the likelihood function for $\Theta$ is

$$L(\Theta|Y,X) = \prod_{i=1}^{n_s}\sum_{j=1}^{3^m} p_{ij}\,\frac{f\left(y_i|\mu_j, \sigma^2\right)}{U_j}, \tag{5}$$

where $p_{ij}$'s are the conditional probabilities of QTL genotypes given marker genotypes, $f(y_i|\mu_j, \sigma^2)$ is the normal density with mean $\mu_j$ and variance $\sigma^2$, and

$$\begin{aligned} U_j &= \int_{-\infty}^{T_L} f\left(y_i|\mu_j, \sigma^2\right)dy_i + \int_{T_R}^{\infty} f\left(y_i|\mu_j, \sigma^2\right)dy_i \\ &= \Phi\left(\tau_{1j}\right) + 1 - \Phi\left(\tau_{2j}\right) \end{aligned}$$

is the cumulative density with genotypic values greater than $T_R$ and lower than $T_L$. Statistically, the normal density $f(y_i|\mu_j, \sigma^2)$ is standardized by $U_j$ to become a truncated normal density $f(y_i|\mu_j, \sigma^2)/U_j$. The details of the EM algorithm for obtaining the MLE of the parameters in the truncated normal mixture likelihood are described in *Appendix*. In summary, the ($t$ + 1)th iteration of the EM step is given below.

E-step: Update the posterior probabilities of the $3^m$ QTL genotypes,

$$\pi_{ij} = \frac{p_{ij} f\left(y_i|\mu_j, \sigma^2\right)\Big/U_j}{\sum_{k=1}^{3^m} p_{ik} f(y_i|\mu_k, \sigma^2)/U_k}$$

for $j$ = 1, 2, ..., $3^m$, $i$ = 1, 2, ..., $n_s$.

M-step: Find the estimates to maximize the conditional log-likelihood (see *Appendix*). Equivalently, we can obtain the estimates from the following equations. For $\mu$, the QTL effects, and $\sigma^2$, the equations are

$$\mu^{(t+1)} = \frac{1}{n_s}\left[1_{n_s}'\left(\mathbf{Y} - \mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)}\right) + \mathbf{R}_\mu^{(t)}\right], \tag{6}$$

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

$$\begin{aligned} \sigma^{2^{(t+1)}} = \frac{1}{n_s - \mathbf{R}_{\sigma^2}^{(t)}}\Big\{&\left(\mathbf{Y} - \mu^{(t+1)}1_{n_s}\right)'\left(\mathbf{Y} - \mu^{(t+1)}1_{n_s}\right) \\ &- 2\left(\mathbf{Y} - \mu^{(t+1)}1_{n_s}\right)'\mathbf{\Pi}^{(t)}\mathbf{D}\mathbf{E}^{(t+1)} \\ &+ \mathbf{E}^{(t+1)\prime}\mathbf{V}^{(t)}\mathbf{E}^{(t+1)}\Big\}, \end{aligned} \tag{8}$$

where $\mathbf{\Pi} = \{\pi_{ij}\}_{n_s \times 3^m}$ contains the posterior probabilities of QTL genotypes for the $n_s$ genotyped individuals, $\mathbf{E}$ is a $k \times 1$ column vector whose elements denote the QTL

effects (*e.g.*, $\mathbf{E}_{2\times 1} = [a\ d]'$ for a single-QTL model considering the additive and dominance effects),

$$\mathbf{r} = \left\{ \frac{(\mathbf{Y} - \mu 1_{n_s})' \mathbf{\Pi}\, D_i}{1'_{n_s} \mathbf{\Pi}(D_i \# D_i)} \right\}_{k\times 1}, \mathbf{M} = \left\{ \frac{1'_{n_s} \mathbf{\Pi}(D_i \# D_j)}{1'_{n_s} \mathbf{\Pi}(D_i \# D_i)} \times \delta(i \neq j) \right\}_{k\times k},$$

$$\mathbf{R}_\mu = \left[ 1'_{n_s} \mathbf{\Pi}\mathbf{A} \right] \sigma, \mathbf{R}_{\mathrm{E}} = \left\{ A \frac{1'_{n_s} \mathbf{\Pi}(D_i \# \mathbf{A})\sigma}{1'_{n_s} \mathbf{\Pi}(D_i \# D_i)} \right\} A_{k\times 1}, \mathbf{R}_{\sigma^2} = 1'_{n_s} \mathbf{\Pi}\mathbf{B},$$

and $\mathbf{V} = \{1'_{n_s} \mathbf{\Pi}(D_i \# D_j)\}_{k\times k}$. In the above expressions, $\mathbf{R}_\mu$, $\mathbf{R}_{\mathbf{E}}$, and $\mathbf{R}_{\sigma^2}$ are the correction terms to mainly account for the truncated normal distributions, $D_i$ is a $k \times 1$ column vector in the genetic model that associate with the corresponding QTL effect in the $i$th element of $\mathbf{E}$, and $\delta$ is an indicator variable. Also,

$$\mathbf{A} = \left\{ \frac{\phi(\tau_{1j}) - \phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \right\}_{3^m\times 1} \text{ and}$$
$$\mathbf{B} = \left\{ \frac{\tau_{1j}\phi(\tau_{1j}) - \tau_{2j}\phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \right\}_{3^m\times 1},$$

where $\phi(\cdot)$ denotes the normal density function from the derivatives of log $U_j$ (see *Appendix*) and are key components in $\mathbf{R}_\mu$, $\mathbf{R}_{\mathbf{E}}$, and $\mathbf{R}_{\sigma^2}$. The E and M steps are iterated until convergence. The converged values of $\mu$, $\mathbf{E}$, and $\sigma^2$ are the MLEs. We intended to express the solutions of the parameters (Equations 6–8) in the general formulas format designed for complete genotyping (Kao and Zeng 1997), so that the two sets of equations can be compared and investigated under complete and selective genotyping. When all individuals are genotyped for markers and included in the analysis, the correction terms in Equations 6–8 vanish, and the equations reduce to the same equations for complete genotyping.

### *Model to analyze full data*

If all the $n$ individuals, including the $n_s$ genotyping individuals and $n_u$ ungenotyping individuals, are fitted into the statistical model (Equation 3) for QTL analysis, the model likelihood can be written as

$$L(\Theta|Y,X) = \prod_{i=1}^{n_s} \sum_{j=1}^{3^m} p_{ij} f\left(y_i|\mu_j,\sigma^2\right) \times \prod_{i=1}^{n_u} \sum_{j=1}^{3^m} q_j f\left(y_i|\mu_j,\sigma^2\right), \tag{9}$$

where the first and second terms on the right-hand side are the likelihoods for the $n_s$ genotyped and for the $n_u$ ungenotyped individuals, respectively. Both likelihoods are normal mixture densities as the QTL genotypes are not observed. In the likelihood for the $n_s$ genotyped individuals, the mixing proportions, $p_{ij}$'s, of a genotyped individual $i$ are obtained from the conditional probabilities of the QTL genotypes given its flanking marker genotype. In the likelihood for the $n_u$ ungenotyped individuals, each different individual mixture density will be given to the same mixing proportions $q_j$'s. Since there is no marker available to infer $q_j$'s, ideally, we would use $\kappa_j$'s (Equation 2), *i.e.*, the proportions of QTL genotypes in the ungenotyped individuals (see *Genotypic distributions in the genotyped and ungenotyped individuals*), to serve as the role of $q_j$'s in the likelihood from the ungenotyped individuals. However, the values of $\kappa_j$'s depend on the unknown QTL parameters, which will complicate the maximum-likelihood estimation if used directly. To avoid the complication, the PFB methods use $w_j$'s as $q_j$'s in their models, *e.g.*, $q_1 = w_1 = 1/4$, $q_2 = w_2 = 1/2$, and $q_3 = w_3 = 1/4$ for $m = 1$ in the $F_2$ model. Here, we propose the quantities

$$\alpha_j = \frac{w_j - \beta_j}{\sum_{k=1}^{3^m}(w_k - \beta_k)}, \quad j = 1, 2, \cdots, 3^m, \tag{10}$$

for the approximation of $\kappa_j$'s and use $\alpha_j$'s as $q_j$'s in our proposed method. In $\alpha_j$, $\beta_j$ is given as

$$\beta_j = \frac{\sum_{i=1}^{n_s} p_{ij}}{n},$$

which sums up the conditional probabilities of a QTL genotype (indexed by $j$) over the $n_s$ genotyped individuals (indexed by $i$) and then divides the sum by $n$. Therefore, $\beta_j$ is the expected frequency of a QTL genotype among the genotyped individuals in the whole sample. By subtracting $\beta_j$ from its corresponding population frequency $w_j$, *i.e.*, $(w_j - \beta_j)$, we can obtain the expected frequency of a QTL genotype among the ungenotyped individuals in the whole sample. The proposed quantities $\alpha_j$'s in Equation 10 reweigh these subtracted quantities, so that they are summed up to one and can serve as the proportions of QTL genotypes in the ungenotyped individuals for the use of $q_j$'s in Equation 9. Equation 10 can be better understood by the following example: In the $F_2$ population ($w_1 = 0.25$, $w_2 = 0.5$, and $w_3 = 0.25$), if only one QTL coincident with a marker is considered, the expected genotypic frequencies are equivalent to the observed frequencies in the genotyped individuals. Under $\theta = 0.5$, assume that the observed genotypic frequencies are 0.1, 0.2, and 0.2, respectively, in the genotyped individuals, then $\beta_1 = 0.1$, $\beta_2 = 0.2$, and $\beta_3 = 0.2$. Consequently, we can have $\alpha_1 = 0.3$, $\alpha_2 = 0.6$, and $\alpha_3 = 0.1$ by Equation 10. In practice, QTL are usually not coincident with markers, and the expected genotypic frequencies $p_{ij}$'s will be used in obtaining $\beta_j$'s and then $\alpha_j$'s. On very rare occasions, a negative value may occur in $w_j - \beta_j$, and a zero value is suggested as a replacement. Equivalently, we propose

$$\alpha_j = \frac{\max\left\{0, w_j - \beta_j\right\}}{\sum_{k=1}^{3^m}\max\{0, w_k - \beta_k\}}, \quad j = 1, 2, \cdots, 3^m, \tag{11}$$

for practical use of the mixing proportions $q_j$'s in Equation 9. Our proposed model in Equation 9 has two features. First, it

is as simple as the PFB methods in that the mixing proportions are fixed and need not be estimated, so that the estimation procedures are similar to those of the QTL mapping model under complete genotyping. In the parameter estimation, the EM algorithm for complete genotyping (Kao and Zeng 1997) can be directly applied to obtain the MLE for our proposed model. In E step, we update the posterior probabilities of QTL genotypes for the $n_s$ genotyped individuals and $n_u$ ungenotyped individuals,

$$\pi_{s_{ij}} = \frac{p_{ij} f\left(y_i|\mu_j, \sigma^2\right)}{\sum_{k=1}^{3^m} p_{ik} f(y_i|\mu_k, \sigma^2)} \text{ and } \pi_{u_{ij}} = \frac{q_j f\left(y_i|\mu_j, \sigma^2\right)}{\sum_{k=1}^{3^m} q_k f(y_i|\mu_k, \sigma^2)},$$

respectively, at the current estimates of the parameters. In M step, the solutions of the parameters, $\mu$, $\sigma^2$, and QTL effects have the same formulations as Equations 6–8 except that the correction terms, $\mathbf{R}_\mu$, $\mathbf{R}_E$, and $\mathbf{R}_{\sigma^2}$, vanish. Certainly, the posterior probability matrix must be adjusted to $\mathbf{\Pi} = [\{\pi_{s_{ij}}\}_{n_s \times 3^m} \ \{\pi_{u_{ij}}\}_{n_u \times 3^m}]'$ according to the numbers of genotyped and ungenotyped individuals. The E and M steps are iterated until convergence. The converged values of estimates are the MLEs. Second, because $\alpha_j$'s are estimates of the proportions of QTL genotypes in the ungenotyped individuals, our proposed model using $\alpha_j$'s as mixing proportions will fit better to the ungenotyped individuals when compared to the PFB method using $w_j$'s. As a result, the proposed method can be more powerful in QTL detection under selective genotyping as is validated in the simulation study.

## Simulation Result

Simulations were performed to evaluate the performance of our proposed method and to compare it with the currently used methods in QTL mapping under selective genotyping. Assume that a quantitative trait of interest is controlled by two unlinked epistatic QTL, $Q_A$ and $Q_B$, in the $F_2$ population. The two QTL are placed at 52 and 93 cM of two 150-cM chromosomes. Assume $Q_A$ has additive ($a_1$) and dominance effects ($d_1$), and $Q_B$ has only additive effect ($a_2$). Epistasis between QTL is assumed to be present only for the additive by additive effect ($i_{aa}$). Further, assume four scenarios, ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), ($a_1 = 2$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), ($a_1 = 1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), and ($a_1 = -1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), for the four present effects, which can reflect the relative sizes of the two QTL and epistasis. For ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), $Q_A$ and $Q_B$ contribute 76 and 8% to the total genetic variance ($V_{Q\ A}/V_G = 76\%$ and $V_{Q\ B}/V_G = 8\%$), and epistasis contributes 16% to the total genetic variance ($V_I/V_G = 16\%$). Similarly, ($V_{Q\ A}/V_G = 60\%$, $V_{Q\ B}/V_G = 13.3\%$, $V_I/V_G = 26.7\%$) for ($a_1 = 2, d_1 = 1, a_2 = 1$, $i_{aa} = 2$), ($V_{Q\ A}/V_G = 33.3\%$, $V_{Q\ B}/V_G = 22.2\%$, $V_I/V_G = 44.4\%$) for both ($a_1 = 1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) and ($a_1 = -1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$). For each scenario of the effect setting, two kinds of marker maps, 5 and 15 cM, two heritabilities of the quantitative trait, 0.1 and 0.2, and two levels of selective genotyping proportions, 50 and 20%, are considered. The sample size is 200 for selective proportion 50% (the 100/200 design), and it is 500 for selective proportion 20% (the 100/500 design). Three models, the PFB model, the proposed model, and truncated model (Xu and Vogl 2000), are used for selective genotyping analysis. Also, the results of complete genotyping in the 100/100, 200/200, and 500/500 designs are presented for comparison. The number of simulated replicate is 1000. A stepwise selection procedure (Kao *et al.* 1999) was adopted to detect QTL and analyze epistasis. Threshold value of QTL mapping for selective genotyping has been found to be similar to that for complete genotyping (Muranty and Goffinet 1997; Manichaikul *et al.* 2007). We have further confirmed that the threshold values for selective genotyping are similar among the three different methods and among the two different designs based on 10,000 simulation replicates (results not shown). Here, the approximate threshold values for complete genotyping obtained by Gaussian stochastic process (Kao and Ho 2012) are used as those for selective genotyping. The obtained values at 5% level are 9.18 (9.80) and 12.34 (13.35) for one and two degrees of freedom in the 15-cM (5-cM) marker map, respectively.

To shorten the article, the results for the scenario of ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) are reported in detail regarding the power and estimation (Table 1, Table 2, Table 3, and Table 4), and the results of the other scenarios are reported only for power (Table 5). For the ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) scenario, the analyses using the one-QTL model of the three selective genotyping methods are first applied to QTL detection. Under the one-QTL model, the three methods have similar performance in detection power and parameter estimation of QTL effects and positions (results not shown). The powers of the three methods to detect the larger QTL, $Q_A$, are all close to 100%. And their powers to detect the smaller QTL, $Q_B$, in the 15- and 5-cM marker maps are ~16% and ~18% under the 100/200 design and are ~24% and 27% under the 100/500 design. Further analyses using the multiple-QTL model are then followed for all the complete and selective genotyping methods. Table 1 and Table 2 show the results of QTL mapping under complete genotyping (with the 100/100 and 200/200 designs) and selective genotyping (with the 100/200 design) when epistasis is ignored and considered in the 5- and 15-cM marker maps. In general, for all methods and designs, greater power in detection and better quality in estimation can be achieved when the marker map is denser and epistasis is taken into account. The results for considering epistasis in the analyses are described here. In the complete genotyping 100/100 design, the powers of detecting $Q_A$ and $Q_B$ are 93.4% (90.3%) and 23.0% (19.6%), respectively, in the 5-cM (15-cM) marker map. In the complete genotyping 200/200 design, they are 100% (99.9%) and 56.1% (50.7%), respectively, in the 5-cM (15-cM) marker map. In the selective genotyping 100/200 design, the powers of detecting $Q_A$ and $Q_B$ by the PFB method are

**Table 1 Simulation results of mapping epistatic QTL using the complete genotyping and selective genotyping 100/200 design in the $F_2$ population with the 5-cM marker map**

| | Without epistasis[a] | | | | | | | | With epistasis[b] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_A$ | | | | | $Q_B$ | | | $Q_A$ | | | | $Q_B$ | | | |
| Method | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | $i_{aa} = 2$ |
| | | | | | | | | Complete genotyping | | | | | | | | |
| 100/100[c] | 0.805 | 53.4 (16.5) | 3.06 (0.83) | 0.93 (1.41) | 0.114 | 80.0 (41.3) | 1.08 (1.30) | | 0.934 | 53.4 (16.5) | 3.07 (0.83) | 0.95 (1.38) | 0.230 | 85.0 (34.7) | 1.01 (1.10) | 1.89 (1.68) |
| 200/200 | 0.997 | 52.0 (5.2) | 3.01 (0.52) | 1.00 (0.83) | 0.255 | 84.6 (34.5) | 1.13 (0.82) | | 1.000 | 51.9 (5.2) | 3.02 (0.51) | 1.00 (0.82) | 0.561 | 90.5 (22.6) | 1.04 (0.66) | 2.06 (0.98) |
| | | | | | | | | Selective genotyping | | | | | | | | |
| PFB M.[d] | 0.989 | 52.2 (7.1) | 3.02 (0.56) | 1.06 (0.91) | 0.203 | 85.2 (33.4) | 1.17 (0.87) | | 0.998 | 52.2 (7.0) | 3.02 (0.57) | 1.00 (0.90) | 0.436 | 89.9 (26.7) | 1.04 (0.75) | 2.01 (1.15) |
| Truc. M.[e] | 0.991 | 52.4 (7.2) | 3.14 (0.69) | 1.12 (1.00) | 0.231 | 84.4 (34.1) | 1.24 (0.95) | | 1.000 | 52.3 (7.2) | 3.21 (0.71) | 1.00 (1.00) | 0.436 | 89.8 (27.4) | 1.09 (0.82) | 2.14 (1.35) |
| Prop. M.[f] | 0.993 | 52.3 (7.0) | 3.17 (0.62) | 1.09 (0.99) | 0.254 | 85.1 (35.1) | 1.20 (0.96) | | 0.997 | 52.2 (7.0) | 3.20 (0.64) | 1.03 (0.97) | 0.560 | 89.0 (28.5) | 1.00 (0.81) | 2.24 (1.36) |

[a] With epistasis ignored.
[b] With epistasis considered.
[c] The total 100 sampled individuals are genotyping.
[d] The population frequency-based method (see text).
[e] The truncated model.
[f] The proposed method. Numbers in parentheses are standard deviations. The critical values for claiming significance are 9.80 and 13.35 for 1 and 2 degrees of freedom for the test. Two unlinked QTL, $Q_A$ and $Q_B$, are simulated on the chromosomes with 31 5-cM equally spaced markers. $Q_A$ is placed at 52 cM with additive effect 3 and dominance effect 1 ($a_1 = 3$, $d_1 = 1$, $V_A/V_G = 76\%$). $Q_B$ is placed at 93 cM with additive effect 1 ($a_2 = 1$, $V_B/V_G = 8\%$). The mean is zero, and the additive by additive epistatic effect is 2 ($\mu = 0$, $i_{aa} = 2$, $V_I/V_G = 16\%$). The heritability is 0.2 ($h^2 = 0.2$, $\sigma^2 = 25$). The simulated replicates are 1000.

99.8% (99.4%) and 43.6% (38.5%), the powers by the truncated model are 100% (99.5%) and 43.6% (38.1%), and the powers by the proposed method are 99.7% (99.5%) and 56.0% (49.8%) in the 5-cM (15-cM) marker map, respectively. It shows that the proposed method is more powerful in QTL detection than the PFB method and truncated model and that the proposed method in the 100/200 design has the ability to provide similar power to complete genotyping under the two marker maps. All methods for selective genotyping methods provide similar precision and accuracy in the estimation of QTL positions. For example, in the 5-cM marker map, the means of the estimated $Q_A$ and $Q_B$ positions by the PFB method, truncated model, and proposed method are at 52.2 cM (SD 7.0 cM) and 89.9 cM (SD 26.7 cM), 52.3 cM (SD 7.2 cM) and 89.8 cM (SD 27.4 cM), and 52.2 cM (SD 7.0 cM) and 89.0 cM (SD 28.5 cM), respectively. In the estimation of the QTL effects, the three methods generally perform well, as their means of the estimated effects are all very close to the true value. For example, in the 5-cM marker map, the means of the two estimated additive effects are 3.02 (SD 0.57) and 1.04 (SD 0.75) by the PFB method. The means are 3.21 (SD 0.71) and 1.09 (SD 0.82) by the truncated model, and the means are 3.20 (SD 0.64) and 1.00 (SD 0.81) by the proposed method. The epistatic effect between QTL can be also estimated well in selective genotyping. The means of estimated epistatic effects are 2.01 (SD 1.15), 2.14 (SD 1.35), and 2.24 (SD 1.36), respectively. In complete genotyping, their means are 1.89 (SD 1.68) and 2.06 (SD 0.98) for the 100/100 and 200/200 designs, respectively.

Table 3 and Table 4 show the results of QTL mapping under complete genotyping (with the 100/100 and 500/500 designs) and selective genotyping (with the 100/500 design) when epistasis is ignored and considered in the 5- and 15-cM marker maps. Again, for all methods and designs, greater power in detection and better quality in estimation can be achieved in the dense marker maps and after taking epistasis into account. Their results for considering epistasis in the analyses are described here. Except for the cases in the 100/100 design, the powers to detect the larger $Q_A$ are all 100%. The powers to detect the small $Q_B$ vary in value. In complete genotyping 500/500 design, the powers to detect $Q_B$ are 96.5 and 93.1% in 5- and 15-cM marker maps. In selective genotyping 100/500 design, the powers of detecting $Q_B$ are 69.9 and 62.7% by the PFB method, and they are 59.3 and 55.0% by the truncated model in 5- and 15-cM marker maps. By the proposed method, the powers of detecting $Q_B$ are 80.9 and 75.0%, respectively. The proposed method provides greater powers to detect $Q_B$ as compared to the PFB method and the truncated model. Also, the powers in 100/500 design provided by the three methods are all significantly lower than that in the 500/500 design. It may tell us that selective genotyping in the 100/500 design has difficulties maintaining equivalent power to complete genotyping (see *Conclusion and Discussion* for the reason). In parameter estimation, the three methods for selective

**Table 2 Simulation results of mapping epistatic QTL using the complete genotyping and selective genotyping 100/200 design in the $F_2$ population with 15-cM marker map**

| | Without epistasis[a] | | | | | | | | With epistasis[b] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_A$ | | | | $Q_B$ | | | | $Q_A$ | | | | $Q_B$ | | | |
| Method | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | $i_{aa} = 2$ |
| | | | | | | | | Complete genotyping | | | | | | | | |
| 100/100[c] | 0.765 | 53.2 (18.1) | 2.98 (0.86) | 0.92 (1.48) | 0.099 | 80.5 (43.3) | 1.05 (1.27) | | 0.903 | 53.2 (18.0) | 2.99 (0.85) | 0.94 (1.47) | 0.196 | 83.3 (37.8) | 0.96 (1.11) | 1.80 (1.73) |
| 200/200 | 0.978 | 52.0 (9.2) | 2.93 (0.58) | 0.99 (0.91) | 0.237 | 84.4 (36.1) | 1.12 (0.79) | | 0.999 | 51.9 (9.1) | 2.94 (0.57) | 1.00 (0.89) | 0.507 | 88.3 (25.7) | 1.03 (0.70) | 1.95 (1.06) |
| | | | | | | | | Selective genotyping | | | | | | | | |
| PFB M.[d] | 0.969 | 52.0 (10.1) | 2.94 (0.61) | 1.06 (1.00) | 0.196 | 82.8 (36.5) | 1.14 (0.86) | | 0.994 | 51.9 (10.0) | 2.95 (0.60) | 1.01 (0.99) | 0.385 | 86.9 (29.5) | 1.03 (0.77) | 1.87 (1.25) |
| Truc. M.[e] | 0.972 | 52.0 (10.6) | 3.07 (0.72) | 1.10 (1.11) | 0.216 | 82.5 (37.3) | 1.19 (0.95) | | 0.995 | 52.1 (10.5) | 3.15 (0.75) | 0.98 (1.11) | 0.381 | 85.4 (31.5) | 1.05 (0.90) | 1.99 (1.52) |
| Prop. M.[f] | 0.971 | 52.1 (10.3) | 3.08 (0.68) | 1.12 (1.09) | 0.231 | 81.9 (36.7) | 1.18 (0.93) | | 0.995 | 52.1 (10.3) | 3.12 (0.69) | 1.06 (1.07) | 0.498 | 86.7 (31.6) | 0.98 (0.85) | 2.10 (1.42) |

[a] With epistasis ignored.
[b] With epistasis considered.
[c] The total 100 sampled individuals are genotyping.
[d] The population frequency-based method (see text).
[e] The truncated model.
[f] The proposed method. Numbers in the parentheses are standard deviations. The critical values for claiming significance are 9.18 and 12.34 for 1 and 2 degrees of freedom for the test. Two unlinked QTL, $Q_A$ and $Q_B$, are simulated on the chromosomes with 11 15-cM equally spaced markers. $Q_A$ is placed at 52 cM with additive effect 3 and dominance effect 1 ($a_1 = 3$, $d_1 = 1$, $V_A/V_G = 76\%$). $Q_B$ is placed at 93 cM with additive effect 1 ($a_2 = 1$, $V_B/V_G = 8\%$). The mean is zero, and the additive by additive epistatic effect is 2 ($\mu = 0$, $i_{aa} = 2$, $V_I/V_G = 16\%$). The heritability is 0.2 ($h^2 = 0.2$, $\sigma^2 = 25$). The simulated replicates are 1000.

genotyping methods all perform well and provide similar precision and accuracy for the estimates (see Table 3 and Table 4).

Table 5 presents the detection powers obtained by complete genotyping and different selective genotyping methods in the four settings under the cases of different designs, heritabilities, and marker maps. In general, the proposed method is more powerful than the PFB method and the truncated model, especially, for detecting $Q_B$ in the ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) setting. For example, for $h^2 = 0.2$ in the 100/200 (100/500) design, the powers of detecting $Q_B$ by our proposed method are of 0.560 (0.809) and 0.498 (0.750), under 5- and 15-cM marker maps. The powers of detecting $Q_B$ by the PFB method are 0.436 (0.699) and 0.385 (0.627), respectively, and the powers of detecting $Q_B$ by the truncated model are 0.436 (0.593) and 0.381 (0.550), respectively. Besides, in most cases, the (proposed) model fitting full data is more powerful than the truncated model fitting only the genotyped data, which is more evident in the 100/500 design. For example, in the ($a_1 = -1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) setting of the 100/500 design, the powers to detect $Q_A$ and $Q_B$ by the proposed (PFB) method are 0.948 (0.933) and 0.865 (0.833), and the powers by the truncated model are 0.896 and 0.750 under $h^2 = 0.1$ and the 5-cM marker map. Also, the detecting powers obtained by the proposed method in the 100/200 design are roughly close to those obtained by complete genotyping in the 200/200 design. For example, in the case of $h^2 = 0.1$ and the 5-cM marker map in the 100/200 design (200/200 design), the powers to detect $Q_A$ and $Q_B$ are 0.765 and 0.123 (0.817 and 0.103), 0.722 and 0.280 (0.775 and 0.249), 0.659 and 0.510 (0.699 and 0.536), and 0.667 and 0.506 (0.696 and 0.528), respectively, in the ($a_1 = 3$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), ($a_1 = 2$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), ($a_1 = 1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$), and ($a_1 = -1$, $d_1 = 1$, $a_2 = 1$, $i_{aa} = 2$) settings. Similar trends can be observed in the other cases in the 100/200 design as compared to the results from their 200/200 design (the slightly higher power with selective genotyping may be due to simulation error). It shows that our proposed method (in 100/200 design) has a better ability to maintain the similar power to complete genotyping (in 200/200 design) than the other two models (in the 100/200 design). Nevertheless, depending on the (relative) sizes of the QTL, the detecting powers obtained by the selective genotyping methods in the 100/500 design are obviously lower than those obtained by complete genotyping in the 500/500 design (see Table 5).

## Conclusion and Discussion

The approach of selective genotyping has been widely used for the detection and validation of QTL in genetic studies (Lander and Botstein 1989; Manichaikul *et al.* 2007; Vikram *et al.* 2012; Lu *et al.* 2013; Fontanesi *et al.* 2014). Statistical QTL mapping methods developed for selective genotyping can consider either full data or only genotyping data in the

**Table 3 Simulation results of mapping epistatic QTL using the complete genotyping and selective genotyping 100/500 design in the $F_2$ population with 5-cM marker map**

| | Without epistasis[a] | | | | | | | With epistasis[b] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_A$ | | | | $Q_B$ | | | $Q_A$ | | | | $Q_B$ | | | |
| Method | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | $i_{aa} = 2$ |
| Complete genotyping | | | | | | | | | | | | | | | |
| 100/100[c] | 0.805 | 53.4 (16.5) | 3.06 (0.83) | 0.93 (1.41) | 0.114 | 80.0 (41.3) | 1.08 (1.30) | 0.934 | 53.4 (16.5) | 3.07 (0.83) | 0.95 (1.38) | 0.230 | 85.0 (34.7) | 1.01 (1.10) | 1.89 (1.68) |
| 200/200 | 1.000 | 51.8 (1.9) | 2.96 (0.34) | 1.02 (0.49) | 0.557 | 90.9 (21.9) | 1.06 (0.38) | 1.000 | 51.8 (1.8) | 3.00 (0.33) | 1.01 (0.48) | 0.965 | 92.8 (7.8) | 1.01 (0.35) | 2.03 (0.48) |
| Selective genotyping | | | | | | | | | | | | | | | |
| PFB M.[d] | 1.000 | 51.8 (3.9) | 3.01 (0.43) | 1.02 (0.65) | 0.327 | 86.3 (31.3) | 1.10 (0.60) | 1.000 | 51.9 (3.8) | 2.98 (0.43) | 1.04 (0.64) | 0.699 | 91.0 (20.1) | 1.01 (0.54) | 1.96 (0.69) |
| Truc. M.[e] | 1.000 | 51.9 (4.6) | 3.24 (0.68) | 1.33 (0.91) | 0.369 | 85.6 (31.4) | 1.32 (0.79) | 1.000 | 51.9 (4.5) | 3.38 (0.74) | 1.08 (0.91) | 0.593 | 88.2 (24.1) | 1.11 (0.74) | 2.22 (1.28) |
| Prop. M.[f] | 1.000 | 51.8 (3.8) | 3.15 (0.48) | 1.03 (0.68) | 0.378 | 85.8 (31.8) | 1.14 (0.66) | 1.000 | 51.9 (3.8) | 3.18 (0.49) | 1.04 (0.67) | 0.809 | 90.6 (20.3) | 1.02 (0.59) | 2.26 (0.81) |

[a] With epistasis ignored.
[b] With epistasis considered.
[c] The total 100 sampled individuals are genotyping.
[d] The population frequency-based method (see text).
[e] The truncated model.
[f] The proposed method. Numbers in the parentheses are standard deviations. The critical values for claiming significance are 9.80 and 13.35 for 1 and 2 degrees of freedom for the test. Two unlinked QTL, $Q_A$ and $Q_B$, are simulated on the chromosomes with 31 5-cM equally spaced markers. $Q_A$ is placed at 52 cM with additive effect 3 and dominance effect 1 ($a_1 = 3$, $d_1 = 1$, $V_A/V_G = 76\%$). $Q_B$ is placed at 93 cM with additive effect 1 ($a_2 = 1$, $V_B/V_G = 8\%$). The mean is zero, and the additive by additive epistatic effect is 2 ($\mu = 0$, $i_{aa} = 2$, $V_I/V_G = 16\%$). The heritability is 0.2 ($h^2 = 0.2$, $\sigma^2 = 25$). The simulated replicates are 1000.

QTL analysis (Xu and Vogl 2000). When only the genotyping data are used in the analysis, the truncated mixture model can be applied to the QTL analysis. If the full data are considered in the analysis, the mixture model can be readily implemented to the QTL detection. In this article, we proposed a novel mixture model on the basis of the multiple-QTL model for selective genotyping when full data are included in the analysis. Our proposed model attempts to use the proportions of QTL genotypes in the ungenotyped individuals, rather than the population proportions of QTL genotypes as in the PFB model, to fit the ungenotyped individuals in model construction. Consequently, as shown in this article, the proposed method has the ability to provide better resolution for QTL detection, to analyze epistasis, and to deal with multiple QTL problems in selective genotyping. In addition, we provide a version of the multiple-QTL approach for the truncated mixture model, so that QTL mapping under selective genotyping can be compared and investigated among the different models in multiple QTL conditions. Simulation results show that our proposed method is more powerful than the PFB and truncated methods. Notably, under the 100/200 design, our selective genotyping method can produce roughly equivalent power to complete genotyping (the 200/200 design), and the PFB methods may have difficulty doing so. Under the 100/500 design, selective genotyping has difficulty matching complete genotyping (the 500/500 design) in QTL detection power, because the linkage information about QTL from the genotyped individuals is inadequate relative to that from the entire sample. Also, the analysis using full data, such as by using our proposed method, performs better than that using only genotyping data, because additional information from the ungenotyped individuals is incorporated into the analysis (Xu and Vogl 2000).

In the selective genotyping experiment, the individuals with intermediate trait values (in the central part of the trait distribution) contain less information about QTL than the extreme individuals (Lander and Botstein 1989) and are not genotyped for markers. When fitting them into the model, there is no marker to infer the genotypic distributions of the underlying QTL. As a result, the PFB methods use (fixed) genotypic distribution of QTL in the population as their genotypic distributions to model the ungenotyped individuals (Muranty and Goffinet 1997; Xu and Vogl 2000). However, the genotypic distribution of the QTL in these ungenotyped individuals will differ from that in the population and depends on several known and unknown factors. The known factors include the population structure and selection intensity. The unknown factors contain the modes (additive, dominance, epistasis) of QTL action, size of QTL, linkage strength between QTL. That is to say, the distribution of QTL genotype in the ungenotyped individuals is informative about these factors. On the contrary, the genotypic distribution of QTL in the population is not so informative about the factors in modeling the ungenotyped individuals, as it is equivalent to that in the ungenotyped individuals only under the null that

**Table 4 Simulation results of mapping epistatic QTL using the complete genotyping and selective genotyping 100/500 design in the $F_2$ population with 15-cM marker map**

| | Without epistasis[a] | | | | | | | With epistasis[b] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_A$ | | | | $Q_B$ | | | $Q_A$ | | | | $Q_B$ | | | |
| Method | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | Power | Posi. | $a_1 = 3$ | $d_1 = 1$ | Power | Posi. | $a_2 = 1$ | $i_{aa} = 2$ |
| Complete genotyping | | | | | | | | | | | | | | | |
| 100/100[c] | 0.765 | 53.2 (18.1) | 2.98 (0.86) | 0.92 (1.48) | 0.099 | 80.5 (43.3) | 1.05 (1.27) | 0.903 | 53.2 (18.0) | 2.99 (0.85) | 0.94 (1.47) | 0.196 | 83.3 (37.8) | 0.96 (1.11) | 1.80 (1.73) |
| 200/200 | 1.000 | 50.9 (3.3) | 2.90 (0.36) | 0.98 (0.53) | 0.523 | 90.1 (22.1) | 1.04 (0.39) | 1.000 | 50.9 (3.0) | 2.94 (0.35) | 0.97 (0.52) | 0.931 | 91.8 (9.4) | 1.01 (0.36) | 1.96 (0.53) |
| Selective genotyping | | | | | | | | | | | | | | | |
| PFB M.[d] | 1.000 | 50.5 (5.5) | 2.93 (0.46) | 1.00 (0.71) | 0.280 | 85.1 (32.5) | 1.05 (0.62) | 1.000 | 50.6 (5.4) | 2.91 (0.45) | 1.03 (0.70) | 0.627 | 90.7 (22.1) | 0.98 (0.56) | 1.86 (0.73) |
| Truc. M.[e] | 1.000 | 50.6 (5.8) | 3.21 (0.72) | 1.27 (0.97) | 0.347 | 83.8 (33.3) | 1.28 (0.81) | 1.000 | 50.7 (5.7) | 3.36 (0.78) | 1.02 (0.97) | 0.550 | 87.0 (25.2) | 1.11 (0.76) | 2.17 (1.29) |
| Prop. M.[f] | 1.000 | 50.6 (5.6) | 3.05 (0.50) | 1.02 (0.74) | 0.328 | 84.1 (33.1) | 1.09 (0.67) | 1.000 | 50.7 (5.5) | 3.08 (0.51) | 1.04 (0.73) | 0.750 | 90.0 (23.5) | 1.00 (0.60) | 2.12 (0.83) |

[a] With epistasis ignored.
[b] With epistasis considered.
[c] The total 100 sampled individuals are genotyping.
[d] The population frequency-based method (see text).
[e] The truncated model.
[f] The proposed method. Numbers in the parentheses are standard deviations. The critical values for claiming significance are 9.18 and 12.34 for 1 and 2 degrees of freedom for the test. Two unlinked QTL, $Q_A$ and $Q_B$, are simulated on the chromosomes with 11 15-cM equally spaced markers. $Q_A$ is placed at 52 cM with additive effect 3 and dominance effect 1 ($a_1 = 3$, $d_1 = 1$, $V_A/V_G = 76\%$). $Q_B$ is placed at 93 cM with additive effect 1 ($a_2 = 1$, $V_B/V_G = 8\%$). The mean is zero, and the additive by additive epistatic effect is 2 ($\mu = 0$, $i_{aa} = 2$, $V_i/V_G = 16\%$). The heritability is 0.2 ($h^2 = 0.2$, $\sigma^2 = 25$). The simulated replicates are 1000.

there is no QTL. Therefore, it is possible to gain advantage and fit better to the data from selective genotyping if a method can use the distribution of QTL genotypes in the ungenotyped individuals to model the trait values of the ungenotyped individuals in model construction. Apparently, such a consideration is ignored by the PFB methods, but is taken into account in our proposed model (Equation 9). Hence, our proposed method can further improve QTL mapping under selective genotyping. The empirical threshold values for selective genotyping and complete genotyping have been found to be similar (Muranty and Goffinet 1997; Manichaikul *et al.* 2007). We have further identified that the proposed, PFB, and truncated models have similar empirical threshold values, which may be because they are all constructed by assuming the same population genomic structure.

It is challenging to analytically explore how the difference between QTL effects is affecting the efficiency of selective genotyping (relative to complete genotyping). Ronin *et al.* (1998) showed that the power of separating two linked QTL with different direction of effects is higher than that with the same direction under selective genotyping, which is also validated by Kao and Zeng (2010) in complete genotyping. Sen *et al.* (2009) found that when one or more QTL have large effects, the effectiveness can be unpredictable in their information analysis. This implies that selective genotyping may show better efficiency in QTL mapping under some combinations of QTL parameters than under others, which is validated in our simulation study. Selective genotyping might be most convenient and suitable for the cases in which only one trait is targeted for selection and being analyzed in QTL mapping (Darvasi 1997). The multiple-trait analysis by including the targeted trait and its correlated traits in the model at a time can also improve the power of QTL detection due to taking into account the correlated structure of traits (Jiang and Zeng 1995; Muranty *et al.* 1997; Ronin *et al.* 1998). When multiple traits are targeted for selective genotyping experiments, extreme individuals in one trait may not be the extreme individuals in the others, and the issues of which traits should be selected and of defining the criteria of selection have been important and discussed by several researchers (Lin and Ritland 1997; Muranty and Goffinet 1997; Muranty *et al.* 1997; Xu and Vogl 2000). Muranty *et al.* (1997) suggested using a selection index combining all trait values to select the extreme individuals and then treated their index values as new single trait values in the analysis. Our method can be readily implemented to analyze the index values for multiple-trait analysis. For directly taking multiple traits into account in the analyses, the multivariate versions of the PFB model fitting one and two QTL have been developed by Muranty *et al.* (1997) and Ronin *et al.* (1998), respectively, and the multivariate version for the truncated normal mixture model for fitting one QTL has been proposed by Xu and Vogl (2000). The multivariate version of our proposed model for directly analyzing multiple traits has not yet been considered here and can be pursued on the foundation of

**Table 5 Results of detection power obtained by complete genotyping and different selective genotyping methods in different parameter settings**

| Method | (3, 1, 1, 2)[a] | | (2, 1, 1, 2) | | (1, 1, 1, 2) | | (−1, 1, 1, 2) | | (3, 1, 1, 2) | | (2, 1, 1, 2) | | (1, 1, 1, 2) | | (−1, 1, 1, 2) | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ | $Q_A$ | $Q_B$ |
| | $h^2$ = 0.1 and 5-cM marker map in 100/200 design | | | | | | | | 100/200 design for $h^2$ = 0.1 and 15-cM marker map | | | | | | | |
| Comp. G.[b] | 0.817 | 0.103 | 0.775 | 0.249 | 0.699 | 0.536 | 0.696 | 0.528 | 0.801 | 0.116 | 0.751 | 0.264 | 0.669 | 0.529 | 0.654 | 0.519 |
| PFB M.[c] | 0.752 | 0.085 | 0.703 | 0.217 | 0.629 | 0.462 | 0.616 | 0.442 | 0.745 | 0.094 | 0.701 | 0.230 | 0.613 | 0.472 | 0.574 | 0.480 |
| Truc. M.[d] | 0.746 | 0.080 | 0.713 | 0.204 | 0.627 | 0.448 | 0.610 | 0.432 | 0.735 | 0.099 | 0.694 | 0.214 | 0.612 | 0.484 | 0.561 | 0.460 |
| Prop. M.[e] | 0.765 | 0.123 | 0.722 | 0.280 | 0.659 | 0.510 | 0.667 | 0.506 | 0.769 | 0.142 | 0.723 | 0.293 | 0.667 | 0.511 | 0.608 | 0.526 |
| | $h^2$ = 0.2 and 5-cM marker map in 100/200 design | | | | | | | | 100/200 design for $h^2$ = 0.2 and 15-cM marker map | | | | | | | |
| Comp. G. | 1.000 | 0.561 | 0.996 | 0.754 | 0.997 | 0.964 | 0.987 | 0.961 | 0.999 | 0.507 | 0.991 | 0.695 | 0.975 | 0.933 | 0.973 | 0.925 |
| PFB M. | 0.998 | 0.436 | 0.990 | 0.628 | 0.978 | 0.934 | 0.963 | 0.911 | 0.994 | 0.385 | 0.980 | 0.627 | 0.946 | 0.890 | 0.936 | 0.866 |
| Truc. M. | 1.000 | 0.436 | 0.992 | 0.587 | 0.983 | 0.932 | 0.962 | 0.900 | 0.995 | 0.381 | 0.981 | 0.602 | 0.948 | 0.882 | 0.937 | 0.863 |
| Prop. M. | 0.997 | 0.560 | 0.992 | 0.713 | 0.986 | 0.938 | 0.970 | 0.940 | 0.995 | 0.498 | 0.984 | 0.696 | 0.965 | 0.917 | 0.947 | 0.890 |
| | $h^2$ = 0.1 and 5-cM marker map in 100/500 design | | | | | | | | $h^2$ = 0.1 and 15-cM marker map in 100/500 design | | | | | | | |
| Comp. G. | 1.000 | 0.506 | 1.000 | 0.824 | 0.995 | 0.986 | 0.997 | 0.983 | 1.000 | 0.460 | 0.999 | 0.794 | 0.996 | 0.964 | 0.994 | 0.962 |
| PFB M. | 0.976 | 0.201 | 0.959 | 0.495 | 0.926 | 0.839 | 0.933 | 0.833 | 0.966 | 0.215 | 0.946 | 0.484 | 0.893 | 0.807 | 0.904 | 0.802 |
| Truc. M. | 0.970 | 0.172 | 0.957 | 0.421 | 0.909 | 0.811 | 0.896 | 0.750 | 0.963 | 0.169 | 0.943 | 0.422 | 0.875 | 0.771 | 0.863 | 0.722 |
| Prop. M. | 0.977 | 0.252 | 0.964 | 0.545 | 0.939 | 0.855 | 0.948 | 0.865 | 0.968 | 0.268 | 0.952 | 0.540 | 0.904 | 0.818 | 0.922 | 0.832 |
| | $h^2$ = 0.2 and 5-cM marker map in 100/500 design | | | | | | | | $h^2$ = 0.2 and 15-cM marker map in 100/500 design | | | | | | | |
| Comp. G. | 1.000 | 0.965 | 1.000 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.931 | 1.000 | 0.991 | 1.000 | 1.000 | 1.000 | 1.000 |
| PFB M. | 1.000 | 0.699 | 1.000 | 0.933 | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.627 | 1.000 | 0.913 | 1.000 | 0.995 | 1.000 | 0.995 |
| Truc. M. | 1.000 | 0.593 | 1.000 | 0.844 | 1.000 | 0.998 | 0.998 | 0.987 | 1.000 | 0.550 | 1.000 | 0.812 | 0.999 | 0.993 | 0.996 | 0.974 |
| Prop. M. | 1.000 | 0.809 | 1.000 | 0.961 | 1.000 | 0.999 | 1.000 | 1.000 | 1.000 | 0.750 | 1.000 | 0.939 | 1.000 | 0.995 | 1.000 | 0.998 |

[a] $Q_A$ and $Q_B$ are unlinked with effects ($a_1$ = 3, $d_1$ = 1, $a_2$ = 1, $i_{aa}$ = 2) in the $F_2$ population.
[b] Complete genotyping.
[c] The population frequency-based method (see text).
[d] The truncated model.
[e] The proposed method. The simulated replicates are 1000.

multivariate normal mixture model laid by Jiang and Zeng (1995), Muranty *et al.* (1997), Ronin *et al.* (1998), and Xu and Vogl (2000) in future studies.

## Acknowledgments

The authors are grateful to Mr. Po-Ying Chu for the discussions. We are also grateful to the associate editor and the referees for their valuable comments and suggestions. This study was supported partly by grants NSC102-2313-B-001-019 from the National Science Council, Taiwan, Republic of China.

## Literature Cited

Abdel-Haleem, H., G.-J. Lee, and R. H. Boerma, 2011 Identification of QTL for increased fibrous roots in soybean. Theor. Appl. Genet. 122(5): 935–946.

Cohen, A. C., 1991 *Truncated and Censored Samples: Theory and Applications*. Dekker, New York.

Darvasi, A., 1997 Interval-specific congenic strains (ISCS): an experimental design for mapping a QTL into a 1-centimorgan interval. Mamm. Genome 8(3): 163–167.

Darvasi, A., and M. Soller, 1992 Selective genotyping for determination of linkage between a marker locus and a quantitative trait locus. Theor. Appl. Genet. 85: 353–359.

Fontanesi, L., G. Schiavo, G. Galimberti, D. G. Calo, and V. Russo, 2014 A genome wide association study for average daily gain in Italian Large White pigs. J. Anim. Sci. 92: 1385–1394.

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

Jiang, C. C., and Z.-B. Zeng, 1995 Multiple trait analysis of genetic mapping for quantitative trait loci. Genetics 140: 1111–1127.

Kao, C.-H., and H.-A. Ho, 2012 A score-statistic approach for determining threshold values in QTL mapping. Front. Biosci. E4: 2670–2682.

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

Kao, C.-H., and M.-H. Zeng, 2010 An investigation of the power for separating closely linked QTL in experimental populations. Genet. Res. 92: 283–294.

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

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

Lebowitz, R. J., M. Soller, and J. S. Beckmann, 1987 Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. Theor. Appl. Genet. 73: 556–562.

Lin, J. Z., and K. Ritland, 1997 Quantitative trait loci differentiating the outbreeding *Mimulus guttatus* from the inbreeding *M. platycalyx*. Genetics 146: 1115–1121.

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

Lu, X., H. Wang, B. Liu, and J. Xiang, 2013 Three EST-SSR markers associated with QTL for the growth of the clam Meretrix meretrix revealed by selective genotyping. Mar. Biotechnol. 15: 16–25.

Manichaikul, A., A. A. Palmer, Ś. Sen, and K. W. Broman, 2007 Significance thresholds for quantitative trait locus mapping under selective genotyping. Genetics 177: 1963–1966.

Miller, A. R., N. A. Hawkins, C. E. McCollom, and J. A. Kearney, 2013 Mapping genetic modifiers of survival in a mouse model of Dravet syndrome. Genes Brain Behav. 13: 163–172.

Muranty, H., and B. Goffinet, 1997 Selective genotyping for location and estimation of the effect of a quantitative trait locus. Biometrics 53: 629–643.

Muranty, H., B. Goffinet, and F. Santi, 1997 Multitrait and multipopulation QTL search using selective genotyping. Genet. Res. 70: 259–265.

Ronin, Y. I., A. B. Korol, and J. I. Weller, 1998 Selective genotyping to detect quantitative trait loci affecting multiple traits: interval mapping analysis. Theor. Appl. Genet. 97: 1169–1178.

Sen, S., F. Johannes, and K. W. Broman, 2009 Selective genotyping and phenotyping strategies in a complex trait context. Genetics 181: 1613–1626.

Vikram, P., B. P. M. Swamy, S. Dixit, H. Ahmed, M. T. S. Cruz *et al.*, 2012 Bulk segregant analysis: An effective approach for mapping consistent-effect drought grain yield QTLs in rice. Field Crops Res. 134: 185–192.

Xu, S., and C. Vogl, 2000 Maximum likelihood analysis of quantitative trait loci under selective genotyping. Heredity 84: 525–537.

*Communicating editor: S. Sen*

## Appendix

When applying the model in Equation 3 to analyze genotyping data from selective genotyping, the likelihood function is a mixture of truncated normals with different means and mixing proportions. By treating the truncated normal mixture model as an incomplete-data problem (Little and Rubin 1987), we can apply the EM algorithm to obtain the MLEs of QTL parameters. Let $\tau_{1j} = (T_{\mathrm{L}} - \mu_j)/\sigma$ and $\tau_{2j} = (T_{\mathrm{R}} - \mu_j)/\sigma$ be the left and right truncation points after standardizing by their means and standard deviations. Then, we have

$$U_j = \int_{-\infty}^{\tau_{1j}} \phi(x)dx + \int_{\tau_{2j}}^{\infty} \phi(x)dx = \Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j}),$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ denote standard normal p.d.f. and c.d.f., respectively. For illustration, we take an additive model for $m$ QTL in the model as an example. The EM algorithm for obtaining the MLE of the parameters in the likelihood (Equation 5) is described as below. In E step, the conditional expected complete-data log-likelihood

$$\begin{aligned}
Q\left(\Theta|\Theta^{(t)}\right) &= \sum_{i=1}^{n_s}\sum_{j=1}^{3^m} \log\left(\frac{f\left(y_i|\mu_j,\sigma^2\right)}{U_j} p_{ij}\right)\pi_{ij}^{(t)} \\
&= \sum_{i=1}^{n_s}\sum_{j=1}^{3^m}\left[\log f\left(y_i|\mu_j,\sigma^2\right) - \log(U_j) + \log\left(p_{ij}\right)\right]\pi_{ij}^{(t)} \\
&= \sum_{i=1}^{n_s}\sum_{j=1}^{3^m}\left[\log f\left(y_i|\mu_j,\sigma^2\right) + \log\left(p_{ij}\right)\right]\pi_{ij}^{(t)} - \sum_{i=1}^{n_s}\sum_{j=1}^{3^m}\log(U_j)\pi_{ij}^{(t)}
\end{aligned}$$

is computed, where $\pi_{ij}$'s are the posterior probabilities of QTL genotype. The last term in the right-hand side is a correction term for truncation.

The M step is to find $\Theta^{(t+1)}$ that maximize the conditional expected log-likelihood $Q(\Theta|\Theta^{(t)})$. We first obtain the derivatives of log $U_j$ with respect to each parameter as follows. For the effect, say $a_i$, associated with the $i$th column of the genetic design matrix $\mathbf{D}$ ($D_i$), the derivative is

$$\begin{aligned}
\frac{\partial}{\partial a_i}\log(U_j) &= \frac{\partial}{\partial a_i}\log(\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})) = \frac{(\partial/\partial a_i)(\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j}))}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \\
&= \frac{\phi(\tau_{1j})(\partial/\partial a_i)\tau_{1j} - \phi(\tau_{2j})(\partial/\partial a_i)\tau_{2j}}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \\
&= \frac{\phi(\tau_{1j})(\partial/\partial a_i)\left[\left(T_{\mathrm{L}} - \mu_j\right)\big/\sigma\right] - \phi(\tau_{2j})(\partial/\partial a_i)\left[\left(T_{\mathrm{R}} - \mu_j\right)\big/\sigma\right]}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \\
&= \frac{\phi(\tau_{1j})\left(-(1/\sigma)(\partial/\partial a_i)\mu_j\right) - \phi(\tau_{2j})\left(-(1/\sigma)(\partial/\partial a_i)\mu_j\right)}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \\
&= -(1/\sigma)\frac{\phi(\tau_{1j}) - \phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} D_{ij},
\end{aligned}$$

where $D_{ij}$ is the $ij$-entry of design matrix $\mathbf{D}$. For $\mu$ and $\sigma^2$, the derivatives are

$$\frac{\partial}{\partial \mu}\log(U_j) = -\frac{1}{\sigma}\frac{\phi(\tau_{1j}) - \phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})}$$

and

$$\frac{\partial}{\partial\sigma^2}\log(U_j) = \frac{\phi(\tau_{1j})(\partial/\partial\sigma^2)\tau_{1j} - \phi(\tau_{2j})(\partial/\partial\sigma^2)\tau_{2j}}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})}$$

$$= \frac{\phi(\tau_{1j})\left(-(1/2\sigma^2)\left[\left(T_{\mathrm{L}} - \mu_j\right)/\sigma\right]\right) - \phi(\tau_{2j})\left(-(1/2\sigma^2)\left[\left(T_{\mathrm{R}} - \mu_j\right)/\sigma\right]\right)}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})}$$

$$= -\frac{1}{2\sigma^2}\frac{\phi(\tau_{1j})\tau_{1j} - \phi(\tau_{2j})\tau_{2j}}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})},$$

respectively. For simplicity in the following derivations, we define

$$A_j = \frac{\phi(\tau_{1j}) - \phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})} \text{ and } B_j = \frac{\tau_{1j}\phi(\tau_{1j}) - \tau_{2j}\phi(\tau_{2j})}{\Phi(\tau_{1j}) + 1 - \Phi(\tau_{2j})},\ j = 1, 2, \cdots, 3^m.$$

The partial derivatives of $Q(\Theta|\Theta^{(t)})$ with respect of each parameters are described below, For one of the QTL effect, say $a_1$, the derivative is

$$\frac{\partial}{\partial a_1}Q\left(\Theta|\Theta^{(t)}\right) = \frac{1}{\sigma^2}\Big[(\mathbf{Y}-\mu 1_{n_s})'\mathbf{\Pi}^{(t)}D_1 - 1_{n_s}'\mathbf{\Pi}^{(t)}(D_1\#D_1)a_1 - 1_{n_s}'\mathbf{\Pi}^{(t)}(D_1\#D_2)a_2 - \cdots$$
$$- 1_{n_s}'\mathbf{\Pi}^{(t)}(D_1\#D_k)a_k + \sigma 1_{n_s}'\mathbf{\Pi}^{(t)}(\mathbf{A}\#D_1)\Big],$$

where # denotes the element-by-element product of the two same-order matrices, $\mathbf{\Pi} = \{\pi_{ij}\}_{n_s\times 3^m}$, $\mathbf{A} = (A_1, A_2, \cdots, A_{3^m})'$. The derivatives for other effects have the similar expressions. For $\mu$ and $\sigma^2$, the derivatives are

$$\frac{\partial}{\partial\mu}Q\left(\Theta|\Theta^{(t)}\right) = \frac{1}{\sigma^2}1_{n_s}'\left(\mathbf{Y} - \mu 1_{n_s} - \mathbf{\Pi}^{(t)}\mathbf{DE} + \sigma\mathbf{\Pi}^{(t)}\mathbf{A}\right)$$

$$\frac{\partial}{\partial\sigma^2}Q\left(\Theta|\Theta^{(t)}\right) = \frac{1}{2\sigma^4}\left[(\mathbf{Y}-\mu 1_{n_s})'(\mathbf{Y} - \mu 1_{n_s}) - 2(\mathbf{Y}-\mu 1_{n_s})'\mathbf{\Pi}^{(t)}\mathbf{DE} + \mathbf{E}'\mathbf{V}^{(t)}\mathbf{E} - \sigma^2\left(n_s - 1_{n_s}'\mathbf{\Pi}^{(t)}\mathbf{B}\right)\right]$$

where $\mathbf{B} = (B_1, B_2, \cdots, B_{3^m})'$ and

$$\mathbf{V} = \begin{pmatrix} 1_{n_s}'\mathbf{\Pi}(D_1\#D_1) & 1_{n_s}'\mathbf{\Pi}(D_2\#D_1) & \cdots & 1_{n_s}'\mathbf{\Pi}(D_k\#D_1) \\ 1_{n_s}'\mathbf{\Pi}(D_1\#D_2) & 1_{n_s}'\mathbf{\Pi}(D_2\#D_2) & \cdots & 1_{n_s}'\mathbf{\Pi}(D_k\#D_2) \\ \cdot & \cdot & \cdot & \cdot \\ 1_{n_s}'\mathbf{\Pi}(D_1\#D_k) & 1_{n_s}'\mathbf{\Pi}(D_2\#D_k) & \cdots & 1_{n_s}'\mathbf{\Pi}(D_k\#D_k) \end{pmatrix}.$$

The solutions of the above partial derivatives can be obtained in close forms. For the effect $a_1$,

$$a_1^{(t+1)} = \frac{\left(\mathbf{Y}-\mu^{(t)}1_{n_s}\right)'\mathbf{\Pi}^{(t)}D_1 - \cdots - 1_{n_s}'\mathbf{\Pi}^{(t)}(D_1\#D_k)a_k^{(t)} + \sigma^{(t)}1_{n_s}'\mathbf{\Pi}^{(t)}\left(\mathbf{A}^{(t)}\#D_1\right)}{1_{n_s}'\mathbf{\Pi}^{(t)}(D_1\#D_1)}.$$

The solutions for the other effects have similar expressions and can be easily formulated accordingly by operating their corresponding column vectors in $\mathbf{D}$. These solutions for the effects can be arranged together and expressed in matrix form as Equation 7. The solutions for $\mu$ and $\sigma^2$ are

$$\mu^{(t+1)} = \frac{1}{n_{\mathrm{s}}}\left[1_{n_s}'\left(\mathbf{Y} - \mathbf{\Pi}^{(t)}\mathbf{DE}^{(t+1)}\right) + 1_{n_s}'\sigma^{(t)}\mathbf{\Pi}^{(t)}\mathbf{A}^{(t)}\right],$$

$$\sigma^{2^{(t+1)}} = \frac{1}{n_{\mathrm{s}} - 1_{n_s}'\mathbf{\Pi}^{(t)}\mathbf{B}^{(t)}}\left[\left(\mathbf{Y}-\mu^{(t+1)}1_{n_s}\right)'\left(\mathbf{Y} - \mu^{(t+1)}1_{n_s}\right) - 2\left(\mathbf{Y}-\mu^{(t+1)}1_{n_s}\right)'\mathbf{\Pi}^{(t)}\mathbf{DE}^{(t+1)} + \mathbf{E}^{(t+1)'}\mathbf{V}^{(t)}\mathbf{E}^{(t+1)}\right].$$

It is worth pointing out that if other possible effects are considered in the model, the solutions can be still formulated as Equation 7.