Genetics and Molecular Biology, 35, 3, 622-631 (2012)
Copyright © 2012, Sociedade Brasileira de Genética. Printed in Brazil
www.sbg.org.br

Research Article

Model selection for quantitative trait loci mapping in a full-sib family

Chunfa Tong, Bo Zhang, Huogen Li and Jisen Shi

Key Laboratory of Forest Genetics and Biotechnology of the Ministry of Education,
Nanjing Forestry University, Nanjing, China.

Abstract

Statistical methods for mapping quantitative trait loci (QTLs) in full-sib forest trees, in which the number of alleles and linkage phase can vary from locus to locus, are still not well established. Previous studies assumed that the QTL segregation pattern was fixed throughout the genome in a full-sib family, despite the fact that this pattern can vary among regions of the genome. In this paper, we propose a method for selecting the appropriate model for QTL mapping based on the segregation of different types of markers and QTLs in a full-sib family. The QTL segregation patterns were classified into three types: test cross (1:1 segregation), F_2 cross (1:2:1 segregation) and full cross (1:1:1:1 segregation). Akaike's information criterion (AIC), the Bayesian information criterion (BIC) and the Laplace-empirical criterion (LEC) were used to select the most likely QTL segregation pattern. Simulations were used to evaluate the power of these criteria and the precision of parameter estimates. A Windows-based software was developed to run the selected QTL mapping method. A real example is presented to illustrate QTL mapping in forest trees based on an integrated linkage map with various segregation markers. The implications of this method for accurate QTL mapping in outbred species are discussed.

Key words: full-sib family, interval mapping, model selection, quantitative trait locus.

Received: November 14, 2011; Accepted: April 21, 2012.

Introduction

Genetic mapping of quantitative trait loci (QTLs) based on genetic linkage maps is a powerful tool for unraveling the genetic architecture of quantitative trait variation in plants, animals and humans. Since the seminal publication on interval mapping by Lander and Botstein (1989) there has been a tremendous development of statistical methods and algorithms for QTL mapping. To make interval mapping more useful, Zeng (1993, 1994) and Jansen and Stam (1994) independently proposed so-called composite interval mapping in which partial regression analysis is used to separate the effects of multiple linked QTLs. Zeng and collaborators constructed the framework for multiple interval mapping to simultaneously characterize the underlying QTLs (their number, locations, and main and epistatic effects) for a quantitative trait (Kao *et al.*, 1999; Zeng *et al.*, 1999). Xu and colleagues extended interval mapping to map qualitatively inherited traits, such as binary and categorical traits (Xu and Atchley, 1996; Yi and Xu, 2000; Xu *et al.*, 2005). The principle of interval mapping was established for a pedigree, initiated with two inbred lines, such as the F_2, backcross and recombinant inbred lines. For any two inbred lines, there are only two alleles at each locus and in the F_1 hybrids that transmit gametes to the next generation there is a fixed linkage phase between any two loci. These two features of inbred lines greatly facilitate statistical inference about the QTL location and effects.

In practice, it is difficult or impossible to generate inbred lines for outcrossing species such as forest trees because of their high heterozygosity and long generation intervals. For any two heterozygous individuals, the number of alleles per locus can differ from gene to gene, leading to different segregation patterns when the two individuals are crossed. Wu *et al.* (2002) listed all possible types of marker segregation in a full-sib family derived from two heterozygous lines. For a given heterozygous line, there is uncertainty about the linkage phase between any pair of loci, *i.e.*, diplotype when the two homozygous chromosomes are considered together. Despite these difficulties, various models and methods for linkage analysis in outcrossing species have been developed through the collective efforts of statisticians and geneticists (Grattapaglia and Sederoff, 1994; Maliepaard *et al.*, 1997; Wu *et al.*, 2002). Lu *et al.* (2004) derived a general framework that covers all these approaches and allows for linkage analysis between any types of markers by simultaneously estimating the recombination fraction, parental diplotype and gene order. More recently, Tong *et al.* (2010) described a hidden

Send correspondence to Jisen Shi. Key Laboratory of Forest Genetics and Biotechnology of the Ministry of Education, Nanjing Forestry University, 159 Longpan Road, 210037 Nanjing, Jiangsu Province, China. E-mail: jshi@njfu.edu.cn.

Markov model approach for multilocus linkage analysis and developed a Windows-based software to construct genetic linkage maps with different segregation markers in a full-sib family.

Nevertheless, despite these advances, there has been limited exploration of the modeling and analysis of QTL mapping in outcrossing species. Haley *et al.* (1994) proposed an approach for mapping outcrossing QTLs in an experimental cross with the F_2 type markers. Although this approach was used to detect QTLs in pigs (Andersson *et al.*, 1994), it did not receive widespread acceptance because of its failure to incorporate the linkage phase of the parents and any type of marker segregation. Lin *et al.* (2003) subsequently proposed a general statistical model for simultaneously estimating the QTL-marker linkage phase, QTL location and QTL effects in an outcrossed family. Although some key statistical issues of the latter model have been investigated, there has been no systematic modeling of QTL segregation patterns.

In this article, we propose a method for selecting the appropriate model for mapping QTL intervals in a full-sib family derived from two outcrossing parents by considering all possible patterns of QTL segregation, *i.e.*, test cross (1:1 segregation), F_2 cross (1:2:1 segregation) and full cross (1:1:1:1 segregation). The most likely QTL segregation pattern for a sample was chosen based on model selection criteria such as Akaike's information criterion (AIC; Akaike, 1974), the Bayesian information criterion (BIC; Schwarz, 1978) and the Laplace-empirical criterion (LEC; McLachlan and Pell, 2000). The method capitalizes on all types of marker segregation and provides simultaneous estimates of the QTL segregation pattern, QTL location and QTL effects. Simulations were used to investigate the statistical behavior of this QTL mapping approach. A Windows-based software was developed to implement the statistical model for QTL mapping in outbred species and the usefulness of the method was validated by using an outcrossing forest tree as an example.

Materials and Methods

Segregation pattern

Suppose that two outcrossing lines, P_1 and P_2, are crossed to generate a full-sib family. The number of different alleles at an informative marker locus in the two parents may be 2, 3 or 4. Maliepaard *et al.* (1997) showed that the possible combinations of two parental genotypes at an informative marker locus, *i.e.*, segregation types, were $ab \times aa$, $aa \times ab$, $ab \times ab$, $ab \times cd$, $ao \times ao$, $ab \times ao$ or $ao \times ab$, where a, b, c and d denote different alleles at a marker locus and o denotes the null allele, with the two characters to the left of the crossing symbol representing the marker genotype of P_1 and the two characters on the right representing the marker genotype of P_2. The linkage analysis used to estimate recombination and linkage phase inference between any two markers is well-defined (Wu *et al.*, 2002, 2007; Lu *et al.*, 2004; Tong *et al.*, 2010) and allows the construction of an integrated linkage map that can contain any type of segregation markers. Similarly, a QTL may also have up to four alleles and present different segregation types. However, some of the segregation types, such as $q_1q_1 \times q_1q_2$ and $q_1q_1 \times q_2q_3$, cannot be distinguished from each other because of inadequate information about allelic configurations.

The QTL segregation patterns are generally classified into three types: (1) test cross, in which the segregation type is $q_1q_1 \times q_1q_2$ or $q_1q_2 \times q_1q_1$ that can generate two genotypes, q_1q_1 and q_1q_2 (1:1 segregation), (2) F_2 cross, in which the segregation type is $q_1q_2 \times q_1q_2$ that can generate three genotypes, q_1q_1, q_1q_2 and q_2q_2 (1:2:1 segregation) and (3) full cross, in which the segregation type is $q_1q_2 \times q_3q_4$ that can generate four genotypes, q_1q_3, q_1q_4, q_2q_3, and q_2q_4 (1:1:1:1 segregation). Each of these QTL segregation types reflects different degrees of information and can be discriminated from the others by using appropriate model selection criteria.

Conditional probability

Consider two molecular markers and a putative QTL in the interval of two markers on a chromosome in a diploid full-sib family. We initially assume that there are four alleles for each molecular marker loci or QTL and that the combined genotypes of the two parents at two markers and a QTL are denoted by $a_1q_1a_2$ / $b_1q_2b_2$ and $c_1q_3c_2$ / $d_1q_4d_2$, where the slash is used to segregate the two haplotypes of a genotype. If r is the recombination fraction between the markers, r_1 the recombination fraction between marker 1 and the QTL, and r_2 the recombination fraction between the QTL and marker 2, then we have the relationship $r = r_1 + r_2 - 2r_1r_2$, assuming that there is no interference between two intervals on chromosomes. The frequencies or probabilities of the combined genotypes in the progeny can be easily derived, as shown in Table 1, in which the elements were multiplied by 4. For the other marker and QTL segregation patterns, the probability of marker and QTL genotype can be obtained by first merging the rows of the same marker genotype and then the columns of the same QTL genotype in Table 1. Once the probabilities of all the marker and QTL genotypes have been obtained, the conditional probability of a QTL genotype given the combined genotype of the two markers can be obtained by dividing the probability of the corresponding marker and QTL genotype by the sum of all the probabilities with the same given marker genotype.

Mixed model

For a given QTL segregation pattern, let J be the number of QTL genotypes (J = 2, 3 or 4). Assume that a quantitative trait is distributed as a normal distribution with mean μ_j and variance σ^2 within the jth QTL genotype (j = 1,..., J).

Table 1 - Probabilities (multiplied by 4) of marker and QTL genotypes in the progeny generated by hybridization: $a_1q_1a_2/b_1q_2b_2 \times c_1q_3c_2/d_1q_4d_2$.

Marker genotype	QTL genotype			
	q_1q_3	q_2q_3	q_1q_4	q_2q_4
$a_1c_1\ a_2c_2$	$(1 - r_1)^2\ (1 - r_2)^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1^2\ r_2^2$
$a_1c_1\ b_2c_2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ (1 - r_2)^2$	$r_1(1 - r_1)\ r_2^2$	$r_1^2r_2(1 - r_2)$
$a_1c_1\ a_2d_2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2^2$	$r_1(1 - r_1)\ (1 - r_2)^2$	$r_1^2r_2(1 - r_2)$
$a_1c_1\ b_2d_2$	$(1 - r_1)^2\ r_2^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1^2(1 - r_2)^2$
$b_1c_1\ a_2c_2$	$r_1(1 - r_1)\ (1 - r_2)^2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1^2r_2(1 - r_2)$	$r_1(1 - r_1)\ r_2^2$
$b_1c_1\ b_2c_2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$(1 - r_1)^2\ (1 - r_2)^2$	$r_1^2\ r_2^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$
$b_1c_1\ a_2d_2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$(1 - r_1)^2\ r_2^2$	$r_1^2(1 - r_2)^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$
$b_1c_1\ b_2d_2$	$r_1(1 - r_1)\ r_2^2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1^2r_2(1 - r_2)$	$r_1(1 - r_1)\ (1 - r_2)^2$
$a_1d_1\ a_2c_2$	$r_1(1 - r_1)\ (1 - r_2)^2$	$r_1^2r_2(1 - r_2)$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2^2$
$a_1d_1\ b_2c_2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1^2(1 - r_2)^2$	$(1 - r_1)^2\ r_2^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$
$a_1d_1\ a_2d_2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1^2\ r_2^2$	$(1 - r_1)^2\ (1 - r_2)^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$
$a_1d_1\ b_2d_2$	$r_1(1 - r_1)\ r_2^2$	$r_1^2r_2(1 - r_2)$	$(1 - r_1)^2\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ (1 - r_2)^2$
$b_1d_1\ a_2c_2$	$r_1^2(1 - r_2)^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$(1 - r_1)^2\ r_2^2$
$b_1d_1\ b_2c_2$	$r_1^2r_2(1 - r_2)$	$r_1(1 - r_1)\ (1 - r_2)^2$	$r_1(1 - r_1)\ r_2^2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$
$b_1d_1\ a_2d_2$	$r_1^2r_2(1 - r_2)$	$r_1(1 - r_1)\ r_2^2$	$r_1(1 - r_1)\ (1 - r_2)^2$	$(1 - r_1)^2\ r_2\ (1 - r_2)$
$b_1d_1\ b_2d_2$	$r_1^2\ r_2^2$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$r_1(1 - r_1)\ r_2\ (1 - r_2)$	$(1 - r_1)^2\ (1 - r_2)^2$

The phenotypic value of the *i*th individual, y_i, will then have a mixture of normal distributions:

$$\sum_{j=1}^{J} p_{j|i} N(\mu_j, \sigma^2),$$

where $p_{j|i}$ is the conditional probability of the *j*th QTL genotype given the marker genotype of the *i*th individual.

For a sample of *n* individuals in the full-sib family, the likelihood of the parameter vector, $\theta = (\mu_1,..., \mu_J, \sigma^2)$, for a specific position on the chromosome, can be written as

$$L(\theta) = \prod_{i=1}^{n} \sum_{j=1}^{J} p_{j|i} f(y_i; \mu_j, \sigma^2), \tag{1}$$

where

$$f(y_i; \mu_j, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2} (y_i - \mu_j)^2 \right)$$

is the density function of a normal distribution.

EM algorithm

Under the full model, the maximum-likelihood estimates of the parameters can be obtained with a form of the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977). For iteration *s* + 1, assume that we have estimates of the parameter $\hat{\theta}^{(s)}$. In the E-step, we calculate the conditional mean of the complete data log likelihood, which involves calculating the posterior probability of individual *i* having the *j*th QTL genotype, as

$$p_{j|i}^{*(s+1)} = \frac{p_{j|i} f(y_i; \hat{\mu}_j^{(s)}, \hat{\sigma}^{2(s)})}{\sum_{j=1}^{J} p_{j|i} f(y_i; \hat{\mu}_j^{(s)}, \hat{\sigma}^{2(s)})} \tag{2}$$

In the M-step, we maximize the log likelihood by updating the estimates of μ_j and σ^2 as

$$\hat{\mu}_j^{(s+1)} = \frac{\sum_{i=1}^{n} p_{j|i}^{*(s)} y_i}{\sum_{i=1}^{n} p_{j|i}^{*(s)}}, \ (j = 1, \dots, J) \tag{3}$$

$$\hat{\sigma}^{2(s+1)} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{J} p_{j|i}^{*(s)} (y_i - \hat{\mu}_j^{(s)})^2 \tag{4}$$

The EM algorithm is then initiated by taking

$$\hat{\mu}_j^{(0)} = \frac{\sum_{i=1}^{n} p_{j|i} y_i}{\sum_{i=1}^{n} p_{j|i}} \quad \text{and} \quad \hat{\sigma}^{2(0)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

until the estimates converge, where $\bar{y}$ is the empirical mean of observations.

Hypothesis testing

The null hypothesis of no QTL segregating at the specific position of the chromosome is

$$H_0: \mu_j = \mu_0 \text{ for all } j$$

implying that the distribution of the quantitative phenotype does not depend on the genotype of the putative QTL. The corresponding likelihood function is

$$L_0(\theta_0) = \prod_{i=1}^{n} f(y_i, \mu_0, \sigma_0^2) \tag{5}$$

where μ_0 and σ_0^2 are the mean and variance of the overall population, respectively, and $\theta_0(\mu_0, \sigma_0^2)$ is the parameter vector.

Under the null model, the maximum likelihood of parameters can be directly obtained as

$$\hat{\mu}_0 = \bar{y} \text{ and } \hat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2$$

The test statistic for the above hypothesis can be expressed as the log-likelihood ratio of the full model over the null model:

$$LR = 2\log\left[\frac{L(\hat{\theta})}{L_0(\hat{\theta}_0)}\right] \tag{6}$$

where $\hat{\theta} = (\hat{\mu}_1, \ldots, \hat{\mu}_J, \hat{\sigma}^2)$ and $\hat{\theta}_0 = (\hat{\mu}_0, \hat{\sigma}_0^2)$ are two vectors of the maximum likelihood estimates under the full model and null model, respectively. If a high peak of the *LR* profile exceeds a critical threshold then a QTL that controls the trait is asserted to exist in a marker interval. Because *LR* may not be asymptotically distributed as a chi-square distribution an empirical method for determining the genome-wide threshold can be used by performing permutation tests (Churchill and Doerge, 1994).

Model selection

The purpose of model selection is to identify a model that has a balance between the goodness-of-fit of the data and the complexity of the model. Fisher's maximum likelihood cannot be used as a criterion for model selection because a simpler model has to be a subset of a more complicated model and, hence, the maximum likelihood of the former is always less than that of the latter. Akaike's information criterion (AIC; Akaike, 1974) and the Bayesian information criterion (BIC; Schwarz, 1978) are commonly used for model selection. AIC and BIC are defined as

$$\text{AIC} = -2\log L(\hat{\theta}) + 2d \tag{7}$$

and

$$\text{BIC} = -2\log L(\hat{\theta}) + \log(n)d \tag{8}$$

where $L(\hat{\theta})$ is the maximum likelihood, d the number of parameters to be estimated in the model, and n the sample size. AIC is derived in terms of Kullback and Leibler (1951) information for the true model with respect to the fitted model while BIC is based on an integrated likelihood within a Bayesian framework.

In addition to the above two criteria, the Laplace-Empirical criterion (LEC; McLachlan and Pell, 2000) was expected to be a good choice for model selection. LEC not only contains information on the number of parameters and sample size in a model but also provides *a priori* information on the parameters and information matrix of the log likelihood function. LEC is defined as

$$\begin{aligned}\text{LEC} = &-2\log L(\hat{\theta}) - 2\log\, p(\hat{\theta}) + \\ &\log\left|I_e(\hat{\theta})\right| - d\log(2\pi)\end{aligned} \tag{9}$$

where $p(\hat{\theta})$ is the prior probability density of the estimated parameters and $I_e(\hat{\theta})$ is the observed information matrix, *i.e.*, the negative Hessian matrix of the log likelihood, both evaluated at the maximum likelihood estimate vector $\hat{\theta}$. We assumed, as did Roberts *et al.* (1998), that the estimated parameter μ_j was uniformly distributed over the interval of length $2\hat{\sigma}_0$ for $j = 1, \ldots, J$, that σ^2 was uniformly distributed in the interval $(0, \hat{\sigma}_0^2)$ and that all are independent. The LEC for our QTL mapping model can therefore be written as

$$\begin{aligned}LEC = &-2\log L(\hat{\theta}) + 2J\log 2 + (J+2)\log\hat{\sigma}_0^2 + \\ &\log\left|I_e(\hat{\theta})\right| - (J+1)\log(2\pi)\end{aligned} \tag{10}$$

where J is the number of QTL genotypes for a certain QTL segregation pattern. The appendix (in Supplementary Material) provides the details of each element of the matrix used to calculate the determinant of $I_e(\hat{\theta})$.

The approach described above allowed us to choose the model that was most likely to provide the minimum AIC, BIC or LEC among the three QTL segregation patterns for a specific position on a chromosome. The power of AIC, BIC and LEC was assessed through Monte Carlo simulations.

Monte Carlo simulations

To assess the usefulness of the QTL mapping method and model selection in different QTL segregation patterns in a full-sib family we simulated a 100 cM-long chromosome with six markers evenly spaced along the chromosome. As indicated by Maliepaard *et al.* (1997), the segregation patterns of the six markers were *aa* × *ab*, *ab* × *cd*, *aa* × *ab*, *ab* × *cd*, *ab* × *ab* and *aa* × *ab*, and the linkage phase between two adjacent markers were *r*, *r*, *r*, *c* × *r* and *c*, respectively. One QTL was simulated at position 50 cM and the QTL segregation patterns were assumed to be: (1) $q_1q_1 \times q_1q_2$, (2) $q_1q_2 \times q_1q_2$ or (3) $q_1q_2 \times q_3q_4$, corresponding to the three different QTL segregation patterns.

In the simulation, the effects of the QTL genotypes were set to be $\mu_1 = 15$ and $\mu_2 = 10$ for the test cross segregation pattern, $\mu_1 = 20$, $\mu_2 = 16$ and $\mu_3 = 10$ for the F_2 segregation pattern, and $\mu_1 = 20$, $\mu_2 = 18$, $\mu_3 = 14$ and $\mu_4 = 10$ for the full cross segregation pattern. The heritability of the QTL was set at values of $h^2 = 0.10$, 0.15, 0.20, 0.30 and 0.50. The variance of the environment effect, σ^2, was therefore determined by the variance and the heritability of the assumed QTL and was defined by the relationship

$\sigma^2 = \sigma_q^2(1-h^2)/h^2$. For example, in the test cross segregation pattern, if $h^2 = 0.30$, then $\sigma^2 = 14.6$ because in this case $\sigma_q^2 = 6.25$. For each case of the simulation, we sampled 500 individuals from a full-sib family with 1000 replicates. Model selection criteria such as LEC, AIC and BIC were used to select the best model among the three competing models in this study and the power of these criteria was calculated based on 1000 replicates. The statistical power for each model was obtained by counting the number of runs out of 1000 replicates in which the model selection was correct and the LR value was greater than an empirical threshold. The threshold of the LR for each model was estimated by an additional 1000 simulations with no QTL segregation. Generally, the 0.95 or 0.99 quantile of the 1000 LR values under the null model was used as the empirical threshold.

Software development

We developed a Windows-based software, designated as FsQtlMap, to implement the statistical methods for QTL mapping in a full-sib family. FsQtlMap is written in VC++ 6.0 and runs on Microsoft Windows operating systems, including Windows 2000, 2003, XP, Vista and 7. The software assumes that the segregation pattern in a QTL may be test cross (1:1 segregation), F2 cross (1:2:1 segregation) or full cross (1:1:1:1 segregation) in a full-sib family and uses LEC as a model selection criterion to determine the QTL segregation ratio. The summary of QTL detection and a series of intermediate results are generated and saved in the corresponding files associated with QTL mapping. FsQtlMap uses the free software *gnuplot* to plot LOD (the logarithm of the odds based on 10) profiles along the linkage groups; the plots are generated in enhanced metafile format (EMF) and postscript (PS) format. FsQtlMap also provides a function that runs permutation tests to yield the genome-wide LOD threshold for asserting that a given peak of the profile is a QTL for each of the three QTL models. Further details on the data format and operational procedures are provided in the FsQtlMap manual. The software and its manual can be freely downloaded from http://fgbio.njfu.edu.cn/tong/FsQtlMap/FsQtlMap.htm.

A real example

The applicability of our statistical method for mapping QTLs in a full-sib family was demonstrated for a forest tree, specifically an interspecific F_1 hybrid population between *Populus deltoides* and *Populus euramericana* in Xuchou, Jiangsu Province, China. Ninety-three genotypes randomly selected from the population were used to construct the genetic linkage map based on molecular markers detected by RAPD, AFLP, ISSR, SSR and SNP analysis (Zhang, 2005). The linkage map contained 19 linkage groups and 314 markers, of which 252 segregated in a 1:1 ratio, 7 in a 1:2:1 ratio and 55 in a 1:1:1:1 ratio. The linkage phases of the two parents between any two adjacent markers on the map were also predicted. Our analysis identified QTLs that affected the root number, an adventitious root trait, in all of the 19 linkage groups in the integrated map of *P. deltoides* and *P. euramericana*.

Results

Figure 1 compares the powers for selecting the true model among the three candidate models based on LEC, AIC and BIC. Figure 1a,b indicates that the power of LEC and BIC for selecting the QTL segregation model of test cross and F_2 cross was higher than that of AIC for all the heritabilities, whereas Figure 1c shows the opposite, *i.e.*, that the power of AIC for selecting the QTL segregation model of full cross was higher than that of LEC and BIC. Although BIC showed a slight advantage over LEC for selecting the model of test cross and F_2 cross, it had drastically lower power than LEC for selecting the model of full cross, especially when the heritability of the QTL was ≤ 0.20. Overall, the powers of LEC and BIC were

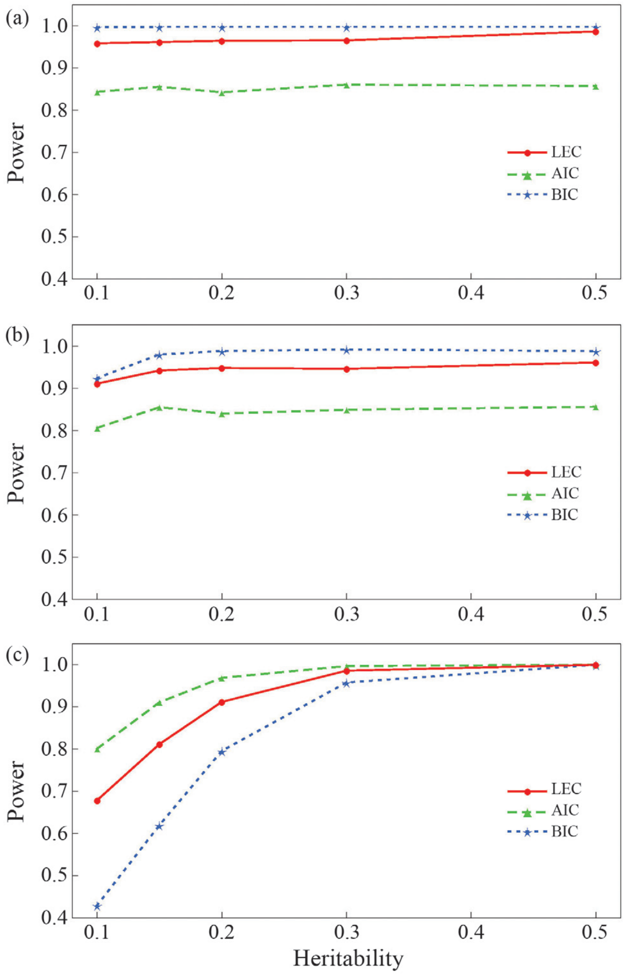

Figure 1 - Comparison of the powers for model selection using different criteria and the QTL segregation pattern associated with (a) test cross, (b) F_2 cross and (c) full cross. Power was estimated as the number of runs out of 1000 replicates in which the correct model was chosen using LEC, AIC and BIC.

almost similar in finding the correct model, probably because these criteria are derived from a Bayesian framework for model selection (McLachlan and Pell, 2000). However, the LEC provides more information of the true model than BIC in that the former not only has *a priori* information of the parameters in the model but also contains information on the negative Hessian matrix of the log likelihood. The result of these simulations suggest that the LEC is the first choice for model selection in mapping QTLs in a full-sib family, a conclusion that agrees well with the findings of model selection theory.

Table 2 provides detailed results on the estimated QTL position, genotypic effects, heritability, and power of model selection using LEC for the three QTL segregation models. The power of model selection increased as the QTL heritability increased and was generally > 90%, except in the case of h^2 = 0.10 and 0.15 for the full cross model. The levels of QTL heritability had a strong effect on the precision of the estimates of the QTL position but had a small effect on the estimates of QTL genotypic effects and QTL heritability. A high QTL heritability can yield estimates of the QTL position that tend towards the true value with a small standard deviation. When the QTL heritability was small, as in the case of h^2 = 0.10, especially for the full cross model, the estimates of QTL position were biased with a standard deviation up to 10.19. The average estimates of QTL genotypic effects and heritability were almost equal to the true values, but the standard deviations decreased as the heritability increased. The precision of the estimates for QTL position, genotypic effects and heritability decreased as the number of parameters in the model increased. The test cross model yielded the most precise estimates of QTL position, genotypic effects and heritability because it had only three parameters (one for residual or environmental variance and two for QTL genotypic effects) while the full cross model yielded less precise estimates with five parameters. This difference can be explained by the fact that the high complexity of the model decreased the precision of the parameter estimates.

The linkage map of *P. deltoides* and *P. euramericana* was scanned with the three QTL segregation models using the interval mapping method. Figure 2 shows the profiles of the log likelihood ratios (LR) generated by each model to detect QTLs that control the adventitious root trait. The critical values determined at the 1% significance level by 1000 permutation tests (Doerge and Churchill, 1996) were 14.71, 21.78 and 27.54 for the test cross, F_2 cross, and full cross models, respectively. For each position on a linkage map, the LEC was used to determine the most likely QTL segregation pattern. Six high peaks (A-F) that exceeded the thresholds were detected in the LR profiles (Figure 2). However, since peak E in Figure 2a and peak F in Figure 2b occurred at the same position in marker interval CG/CTT_440R~ TC/CGT_120, linkage 3 there were only five true QTLs.

Table 3 summarizes the procedure for selecting the most likely QTL segregation pattern for the five positions in Figure 2. According to the LEC, peaks A, C and F were selected to be the significant QTL positions because each of them had the lowest value for LEC and a significant value of LR under the same QTL segregation pattern, whereas peaks B and E were not significant QTL positions since they did not have the lowest values for LEC. However, peak F was close to peak C and they had almost the same

Table 2 - Average estimates for QTL position, effects, heritability and power of model selection (PMS) using LEC for the three QTL segregation patterns based on 1000 simulation replicates. Standard errors are shown in brackets.

QTL segregation pattern	h^2	QTL position	$\hat{u}_1$	$\hat{u}_2$	$\hat{u}_3$	$\hat{u}_4$	$\hat{h}^2$	Power
	0.10	49.91 (5.02)	15.01 (0.51)	10.00 (0.50)			0.102 (0.028)	0.959
	0.15	50.00 (3.60)	15.00 (0.39)	9.99 (0.40)			0.152 (0.031)	0.962
Test cross	0.20	50.00 (2.96)	15.01 (0.33)	10.01 (0.32)			0.202 (0.033)	0.965
	0.30	50.02 (1.98)	15.00 (0.25)	10.00 (0.26)			0.302 (0.036)	0.966
	0.50	50.02 (1.53)	15.00 (0.16)	10.00 (0.17)			0.501 (0.030)	0.987
	0.10	51.36 (7.21)	20.03 (1.17)	15.97 (0.84)	9.98 (1.17)		0.107 (0.030)	0.911
	0.15	50.97 (5.52)	20.01 (0.95)	16.00 (0.65)	10.01 (0.95)		0.155 (0.036)	0.942
F_2 cross	0.20	50.55 (4.43)	19.98 (0.76)	16.00 (0.54)	10.03 (0.78)		0.203 (0.038)	0.948
	0.30	50.13 (2.91)	19.98 (0.57)	15.99 (0.41)	10.00 (0.60)		0.304 (0.043)	0.946
	0.50	50.01 (1.95)	20.01 (0.38)	16.00 (0.26)	10.00 (0.39)		0.503 (0.041)	0.961
	0.10	51.43 (10.19)	20.12 (1.42)	18.58 (1.21)	13.38 (1.22)	9.97 (1.37)	0.121 (0.043)	0.679
	0.15	51.27 (8.27)	20.09 (1.11)	18.31 (1.00)	13.77 (0.99)	9.95 (1.09)	0.169 (0.050)	0.812
Full cross	0.20	50.87 (7.22)	20.02 (0.90)	18.15 (0.88)	13.80 (0.85)	9.98 (0.93)	0.215 (0.054)	0.912
	0.30	50.54 (5.45)	19.97 (0.70)	18.03 (0.69)	13.96 (0.67)	10.00 (0.72)	0.308 (0.059)	0.986
	0.50	50.13 (2.86)	19.98 (0.44)	18.00 (0.42)	14.00 (0.41)	10.03 (0.43)	0.503 (0.046)	1.000

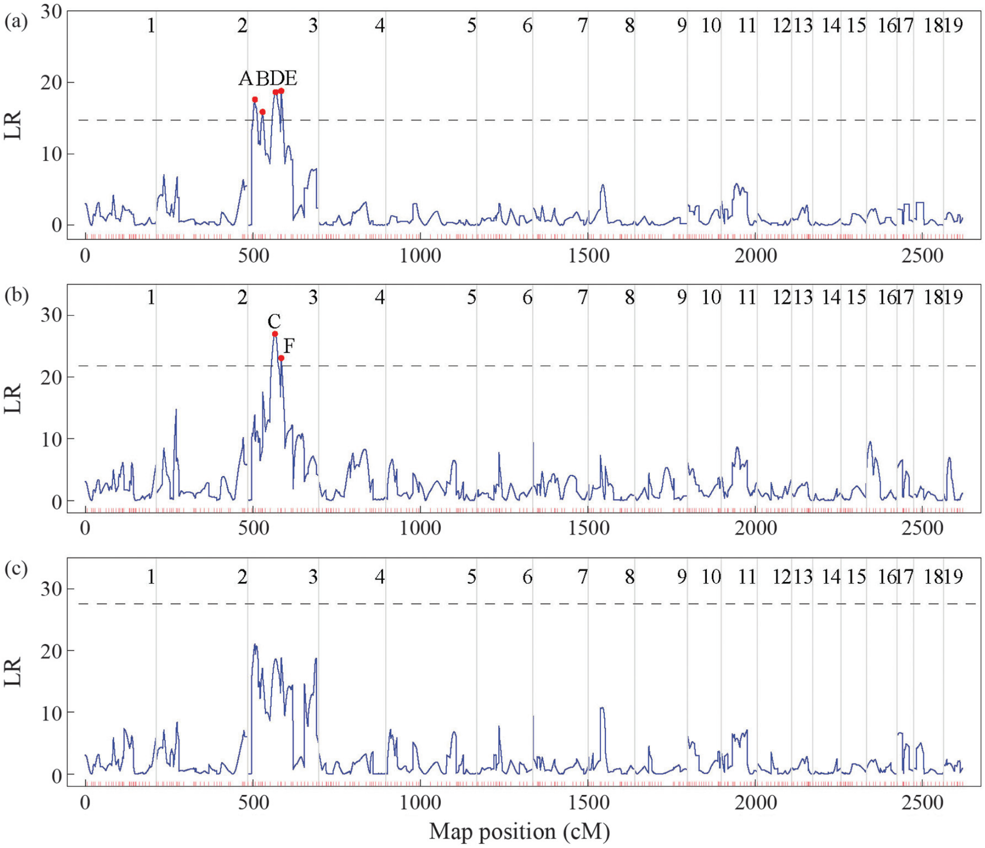

Figure 2 - The profiles of the log likelihood ratios for root number, an adventitious root trait in poplar, across all the 19 linkage groups in the integrated map of *P. deltoids* and *P. euramericana* based on (a) test cross, (b) F_2 cross, and (c) full cross models. The threshold values of the three models for asserting the existence of a QTL at a significance level of p = 0.01 are indicated as horizontal dashed lines. Each short red line at the bottom of the frame indicates a marker position.

genotypic effects so that the former may be considered a ghost QTL (Martinez and Curnow, 1992; Doerge, 2002). Overall, therefore, two QTLs, *i.e.* peaks A and C, were concluded to be the significant QTLs responsible for root number.

Discussion

The efforts of many statistical geneticists in the past two decades mean that genetic linkage maps can now be constructed using different segregation molecular marker data from full-sib families in species such as forest trees in which inbred lines are almost impossible to obtain through traditional self-mating for many generations (Maliepaard *et al.*, 1997; Wu *et al.*, 2002; Lu *et al.*, 2004; Tong *et al.*, 2010). Two softwares, JoinMap (Van Ooijen, 2006) and FsLinkageMap (Tong *et al.*, 2010), are available for constructing an integrated genetic linkage map with predicted linkage phase between any two adjacent markers. Based on such genetic linkage maps in outbred species, we have now proposed a method for selecting the appropriate model for detecting QTLs by considering three QTL segregation patterns, *i.e.*, test cross (1:1 segregation), F_2 cross (1:2:1 segregation) and full cross (1:1:1:1 segregation). Our method has some advantages in the genetic mapping of complex traits by accounting for the biological characteristics of forest trees.

First, our QTL mapping method with model selection procedures allows one to choose the most likely QTL segregation pattern of the three assumed patterns. Like molecular markers, QTL segregation may show different patterns throughout the genome in an outcrossing species. Hence, it is reasonable to incorporate different QTL segregation modes into a statistical model for QTL mapping in a full-sib family. However, MapQTL (Van Ooijen, 2009), the only available software that can be used to detect QTLs with data from a full-sib family, assumes that the QTL segregation is fixed as *ab* × *cd*. This is the case of the full cross pattern in our statistical model. The shortcoming of MapQTL can be illustrated by the real example described above in which QTLs were detected segregating in test cross and F_2 cross patterns. This means that no QTLs would be found if QTL mapping in this example were done with MapQTL.

Second, our QTL mapping method could be done by using genetic linkage maps of outbred species that had been constructed in the past 20 years. For example, in forest trees, many parent-specific linkage maps (Plomion *et al.*,

Table 3 - Results for the detection of QTLs that affect root number, an adventitious root trait in poplar.

High peak	Group	Position (cM)	Interval	Assumed QTL pattern	LR	LEC	Inferred QTL pattern	Effects			Heritability
								$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\mu}_3$	
A	3	20.97	W_19B|P_422	Test cross	17.60^{**}	125.52	Test cross	1.40	1.86		0.185
				F_2 cross	10.40	134.47					
				Full cross	21.07	126.44					
B	3	43.69	G_1158|GT/CTC_765R	Test cross	15.85^{**}	127.18					
				F_2 cross	17.58	126.56					
				Full cross	15.85	127.73					
C	3	80.68	TC/CCG_500|TC/CAG_150	F_2 cross	26.97^{**}	119.58	F_2 cross	1.13	1.55	2.40	0.703
				Test cross	18.47	124.65					
				Full cross	18.47	126.92					
D	3	82.68	TC/CCG_500|TC/CAG_150	Test cross	18.62^{**}	124.48					
				F_2 cross	26.87	119.50					
				Full cross	18.62	126.72					
E	3	99.61	CG/CTT_440R|TC/CGT_120	Test cross	18.79^{**}	124.38					
F				F_2 cross	23.07^{**}	121.68	F_2 cross	1.24	1.54	2.31	0.532
				Full cross	18.79	125.94					

$^{**}p < 0.01$.

1995; Wu *et al.*, 2000; Yin *et al.*, 2002; Shepherd *et al.*, 2003; Gan *et al.*, 2003) have been constructed and QTL mapping studies have also been done with the pseudo-test cross strategy first proposed by Grattapaglia and Sederoff (1994). This method has some limitations in QTL mapping in that the linkage phase between adjacent two markers and possible multiple QTL segregation patterns are not considered. The application of our QTL mapping method to these previous data would be expected to yield better results.

Third, the use of LEC as the criterion for identifying QTL segregation patterns is not only supported by the simulation results but also by the quantity itself. Model selection is an important but very difficult problem that has not been completely resolved for mixed models (McLachlan and Pell, 2000). Although AIC and BIC have been extensively applied to many situations, they were apparently unable to select the correct QTL segregation ratio in our QTL mapping models (Figure 1). Unlike AIC and BIC, LEC contains more information about the model itself. LEC not only contains the number of estimated parameters and the sample size but also the prior probabilities of the estimated parameters and the negative Hessian matrix of the log likelihood. These characteristics indicate that LEC generally has a higher power than AIC and BIC in model selection.

Finally and most importantly, we have developed a Windows-based software (FsQtlMap) to allows the immediate implementation of our QTL mapping strategy. Computer packages for QTL mapping, such as MapMaker/QTL (Lincoln and Lander, 1990) and Windows QTL Cartographer (Wang *et al.*, 2010), are well-established and have been extensively used for inbred lines. In contrast, there are no popular statistical tools for QTL mapping in outbred species such as forest trees. Although MapQTL (Van Ooijen, 2009) has been used for QTL mapping in forest trees by some researchers, its application is limited by the assumption that there is only one QTL segregation pattern in a full-sib family. By incorporating the characteristics of outcross species FsQtlMap provides a much more powerful computing tool for QTL mapping.

Our new QTL mapping method was applied to real data and successfully detected two QTLs that affect adventitious roots in *Populus*. One QTL segregated in an F_2 cross and had much higher heritability. This finding indicates that the rooting capacity of poplars may be controlled by a major gene that can explain ~70% of the phenotypic variance. This conclusion is consistent with that of Han *et al.* (1994).

Acknowledgments

We thank Rongling Wu and two anonymous reviewers for their constructive suggestions and comments on this manuscript. This work was partially supported through a project funded by the Priority Academic Program Development (PAPD) of the Jiangsu Higher Education Institutions and the National Natural Science Foundation of China (grant no. 30872051).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Automat Contr 19:716-723.

Andersson L, Haley CS, Ellegren H, Knott SA, Johansson M, Andersson K, Andersson-Eklund L, Edfors-Lilja I, Fredholm M, Hansson I, *et al.* (1994) Genetic mapping of quantitative trait loci for growth and fatness in pigs. Science 263:1771-1774.

Churchill GA and Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971.

Dempster AP, Laird NM and Rubin DB (1977) Maximum likelihood from incomplete data via EM algorithm. J R Stat Soc Ser B (Methodological) 39:1-38.

Doerge RW (2002) Mapping and analysis of quantitative trait loci in experimental populations. Nat Rev Genet 3:43-52.

Doerge RW and Churchill GA (1996) Permutation tests for multiple loci affecting a quantitative character. Genetics 142:285-294.

Gan S, Shi J, Li M, Wu K, Wu J and Bai J (2003) Moderate-density molecular maps of *Eucalyptus urophylla* S. T. Blake and *E. tereticornis* Smith genomes based on RAPD markers. Genetica 118:59-67.

Grattapaglia D and Sederoff R (1994) Genetic linkage maps of *Eucalyptus grandis* and *Eucalyptus urophylla* using a pseudo-testcross: Mapping strategy and RAPD markers. Genetics 137:1121-1137.

Haley CS, Knott SA and Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195-1207.

Han K, Bradshaw HD, Gordon MP and Han KH (1994) Adventitious root and shoot regeneration *in vitro* is under major gene control in an F2 family of hybrid poplar (*Populus trichocarpa* × *P. deltoides*). Forest Genet 1:139-146.

Jansen RC and Stam P (1994) High resolution of quantitative traits into multiple loci via interval mapping. Genetics 136:1447-1455.

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

Kullback S and Leibler RA (1951) On information and sufficiency. Ann Math Statist 22:79-86.

Lander ES and Botstein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185-199.

Lin M, Lou XY, Chang M and Wu RL (2003) A general statistical framework for mapping quantitative trait loci in nonmodel systems: Issue for characterizing linkage phases. Genetics 165:901-913.

Lincoln SE and Lander ES (1990) Mapping Genes Controlling Quantitative Traits Using MAPMAKER/QTL. Technical Report. Whitehead Institute for Biomedical Research, Cambridge, 46 pp.

Lu Q, Cui YH and Wu RL (2004) A multilocus likelihood approach to joint modeling of linkage, parental diplotype and gene order in a full-sib family. BMC Genetics 5:e20.

Maliepaard C, Jansen J and Van Ooijen JW (1997) Linkage analysis in a full-sib family of an outbreeding plant species: Overview and consequences for applications. Genet Res 70:237-250.

Martinez O and Curnow RN (1992) Estimating the locations and the size of the effects of quantitative trait loci using flanking markers. Theor Appl Genet 85:480-488.

McLachlan G and Pell D (2000) Finite Mixture Models. John Wiley & Sons, New York, 419 pp.

Plomion CD, Malley MO and Durel CE (1995) Genomic analysis in maritime pine (*Pinus pinaster*). Comparison of two RAPD maps using selfed and open-pollinated seeds of the same individual. Theor Appl Genet 90:1028-1034.

Roberts SJ, Husmeier D, Rezek I and Penny W (1998) Bayesian approaches to Gaussian modeling. IEEE Trans Pattern Anal Mach Intell 20:1133-1142.

Schwarz G (1978) Estimating the dimension of a model. Ann Stat 6:461-464.

Shepherd M, Cross M, Dieters MJ and Herry R (2003) Genetic maps for *Pinus elliottii* var. *elliottii* and *P. caribaea* var. *hondurensis* using AFLP and microsatellite markers. Theor Appl Genet 106:1409-1419.

Tong CF, Zhang B and Shi JS (2010) A hidden Markov model approach to multilocus linkage analysis in a full-sib family. Tree Genet Genomes 6:651-662.

Van Ooijen JW (2006) JoinMap 4, Software for the calculation of genetic linkage maps in experimental populations. Kyazma BV, Wageningen, Netherlands.

Van Ooijen JW (2009) MapQTL 6, Software for the mapping of quantitative trait loci in experimental populations of diploid species. Kyazma BV, Wageningen, Netherlands.

Wang S, Basten CJ and Zeng Z-B (2010) Windows QTL Cartographer 2.5. Department of Statistics, North Carolina State University, Raleigh, NC.

Wu RL, Han YF, Hu JJ, Fang JJ, Li L, Li LM and Zeng Z-B (2000) An integrated genetic map of *Populus deltoids* based on amplified fragment length polymorphisms. Theor Appl Genet 100:1249-1256.

Wu RL, Ma CX, Painter I and Zeng Z-B (2002) Simultaneous maximum likelihood estimation of linkage and linkage phases in outcrossing species. Theor Popul Biol 61:349-363.

Wu RL, Ma CX and Casella G (2007) Statistical Genetics of Quantitative Traits: Linkage, Maps and QTL. Springer, New York, 365 pp.

Xu S and Atchley WR (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143:1417-1424.

Xu C, Li Z and Xu S (2005) Joint mapping of quantitative trait loci for multiple binary characters. Genetics 169:1045-1059.

Yi N and Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155:1391-1403.

Yin TM, Zhang XY, Huang MR, Wang MX, Zhuge Q, Tu SM, Zhu LH and Wu RL (2002) Molecular linkage maps of the *Populus* genome. Genome 45:541-555.

Zeng Z-B (1993) Theoretical basis for separation of multiple linked gene effects in 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.

Zeng Z-B, Kao C-H and Basten CJ (1999) Estimating the genetics architecture of quantitative traits. Genet Res 74:279-289.

Internet Resources

gnuplot, http://www.gnuplot.info.

FsQtlMap manual, http://fgbio.njfu.edu.cn/tong/FsQtlMap/FsQtlMap.htm.

Zhang B (2005) Constructing genetic linkage maps and mapping QTLs affecting important traits in poplar. PhD Dissertation, Nanjing Forestry University, Nanjing, China. http://fgbio.njfu.edu.cn/tong/zhang2005.pdf.

Supplementary Material

The following online material is available for this article:

- Appendix: Elements of the information matrix of the log- likelihood.

This material is available as part of the online article from http://www.scielo.br/gmb.

Associate Editor: Everaldo Gonçalves de Barros

Supplementary Material

Appendix: Elements of the information matrix of the log- likelihood

The elements of the information matrix $I_e(\theta)$, denoted by $I_e(\theta)_{j_1 j_2}$, are the negative second partial derivatives of the log likelihood function (Eq. (5)) with respect to μ_1, μ_2, $\cdots$, μ_k, and σ^2, which can be directly obtained as follows:

$$I_e(\theta)_{jj} = -\frac{\partial^2 \ln L(\theta)}{\partial \mu_j^2} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}\left[p_{ij}^*(1-p_{ij}^*)(y_i-\mu_j)^2 - \sigma^2 p_{ij}^* \right], \text{ for } j = 1,2,\cdots,k$$

$$I_e(\theta)_{j_1 j_2} = -\frac{\partial^2 \ln L(\theta)}{\partial \mu_{j_1}\partial \mu_{j_2}} = \frac{1}{\sigma^4}\sum_{i=1}^{n} p_{ij_1}^* p_{ij_2}^* (y_i-\mu_{j_1})(y_i-\mu_{j_2}), \text{ for } j_1, j_2 = 1,2,\cdots,k \text{ and } j_1 \neq j_2$$

$$I_e(\theta)_{k+1,j} = I_e(\theta)_{j,k+1} = -\frac{\partial^2 \ln L(\theta)}{\partial \sigma^2 \partial \mu_j} = \frac{1}{2\sigma^6}\sum_{i=1}^{n} p_{ij}^*(y_i-\mu_j)\left[\sum_{l=1}^{k} p_{il}^*(y_i-\mu_l)^2 - (y_i-\mu_j)^2 \right], \text{ for } j = 1,2,\cdots,k$$

$$I_e(\theta)_{k+1,k+1} = -\frac{\partial^2 \ln L(\theta)}{\partial (\sigma^2)^2} = \frac{n}{2\sigma^4} - \frac{1}{4\sigma^8}\sum_{i=1}^{n}\left[\sum_{l=1}^{k} p_{il}^*(y_i-\mu_l)^4 - \left(\sum_{l=1}^{k} p_{il}^*(y_i-\mu_l)^2 \right)^2 \right]$$

where

$$p_{ij}^* = \frac{p_{ij} f(y_i;\mu_j,\sigma^2)}{\sum_{j=1}^{k} p_{ij} f(y_i;\mu_j,\sigma^2)}$$