Genet. Res., Camb. (2003), **82**, *pp.* 127–138. *With 3 figures.* © 2003 Cambridge University Press
DOI: 10.1017/S0016672303006414 Printed in the United Kingdom

An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family

SHIZHONG XU[1*], NENGJUN YI[1], DAVID BURKE[2], ANDRZEJ GALECKI[3,4] AND RICHARD A. MILLER[3,4,5]

[1] *Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA*
[2] *Department of Human Genetics, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[3] *Geriatrics Center, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[4] *Institute of Gerontology, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*
[5] *Department of Pathology, University of Michigan School of Medicine, Ann Arbor, MI 48109, USA*

(*Received 30 July 2002 and in revised form 28 July 2003*)

Summary

Many diseases show dichotomous phenotypic variation but do not follow a simple Mendelian pattern of inheritance. Variances of these binary diseases are presumably controlled by multiple loci and environmental variants. A least-squares method has been developed for mapping such complex disease loci by treating the binary phenotypes (0 and 1) as if they were continuous. However, the least-squares method is not recommended because of its ad hoc nature. Maximum Likelihood (ML) and Bayesian methods have also been developed for binary disease mapping by incorporating the discrete nature of the phenotypic distribution. In the ML analysis, the likelihood function is usually maximized using some complicated maximization algorithms (e.g. the Newton–Raphson or the simplex algorithm). Under the threshold model of binary disease, we develop an Expectation Maximization (EM) algorithm to solve for the maximum likelihood estimates (MLEs). The new EM algorithm is developed by treating both the unobserved genotype and the disease liability as missing values. As a result, the EM iteration equations have the same form as the normal equation system in linear regression. The EM algorithm is further modified to take into account sexual dimorphism in the linkage maps. Applying the EM-implemented ML method to a four-way-cross mouse family, we detected two regions on the fourth chromosome that have evidence of QTLs controlling the segregation of fibrosarcoma, a form of connective tissue cancer. The two QTLs explain 50–60% of the variance in the disease liability. We also applied a Bayesian method previously developed (modified to take into account sex-specific maps) to this data set and detected one additional QTL on chromosome 13 that explains another 26% of the variance of the disease liability. All the QTLs detected primarily show dominance effects.

1. Introduction

The specific disease that leads to death in any individual often reflects complex interactions between genetic and non-genetic factors. Many diseases are influenced by polymorphic loci but the inheritance patterns are typically non-Mendelian because of the polygenic influences on their risk, interactions between alleles and environmental variants and the complications of competing risks from other potentially lethal illnesses. Mapping loci that modulate risk of specific diseases is more difficult than mapping simple Mendelian loci. One could take a quantitative trait locus (QTL) mapping approach by treating disease phenotype as a quantitative trait (Visscher *et al.*, 1996). However, many disease phenotypes are measured as binary traits (i.e. disease presence or absence) rather than as continuous quantitative traits. From a theoretical point of view, standard QTL mapping cannot be applied to discrete trait mapping. Non-parametric methods might be used for disease mapping (Kruglyak & Lander, 1995*b*). However, with the non-parametric

* Corresponding author. Department of Botany and Plant Sciences, University of California, Riverside, CA 92521, USA. Tel: +1 909 787 5898. Fax: +1 909 787 4437. e-mail: xu@genetics.ucr.edu

method, it is impossible to quantify the disease penetrance and the proportion of disease variance explained by the genes identified.

McIntyre *et al.* (2001) recently developed a probability model particularly suitable for binary disease mapping. They treated the probabilities of disease incidence conditional on genotypes (called the penetrance) as parameters of interest. Ignoring the mixture distribution of the unobserved genotype, they were able to estimate and test the disease penetrance difference between alternative genotypes. This method explicitly incorporates disease penetrances into the model, which is a great improvement over the simple regression and the non-parametric analyses. The estimation and test are accomplished many times faster than using the generalized linear model (Hackett & Weller, 1995; Xu & Atchley, 1996), which is an ML-based method with solutions achieved using a numerical optimization algorithm. Unfortunately, the high speed of McIntyre *et al.* (2001) comes at the cost of generality, because it cannot incorporate any non-genetic effects (e.g. age or location) into the probability model. By contrast, the generalized linear model is an excellent framework for postulating an unobserved continuous quantitative trait that determines the disease status. This underlying variable is called the liability. Mapping disease traits can now be formulated as mapping disease liability. As a result, the discrete trait is transformed into a continuous liability and disease mapping can be achieved using all statistical techniques applied to QTL mapping.

Early work of the generalized linear model applied to categorical trait mapping can be traced back to Hackett and Weller (1995) and Xu and Atchley (1996), both of which used line-crossing data. Applying the generalized linear model approach, Xu *et al.* (1998) successfully detected loci responsible for Merick disease in chicken. Rao and Xu (1998) extended the method to four-way crosses. All these methods use some special optimization algorithms, such as the simplex algorithm of Nelder and Mead (1965). More recently, Yi and Xu (2000*a*) explored the Bayesian method implemented via the MCMC algorithm, which is a computationally intensive sampling-based method. Yi and Xu (2000*b*) also extended the Bayesian method to outbred populations under the random model framework.

In this study, we adopt the same generalized linear model as Xu & Atchley (1996), but treat the underlying disease liability as a missing value. This has the effect of allowing a formulation of the problem that can be solved using an EM algorithm. In fact, in this case, the EM iteration equations arising from the liability function are identical to the normal equation system in multiple linear regression. This EM algorithm has the advantage of being more intuitive and easier to program than the simplex algorithm (Nelder & Mead, 1965). The EM algorithm also provides an intuitive way to facilitate calculation of the information matrix, and thus the variance–covariance matrix of the estimated parameters. We apply this algorithm to a four-way-cross experiment in laboratory mice. Therefore, the model is described in the context of a four-way-cross design. Although four-way-cross models have been developed by many workers (Knott *et al.*, 1997; Rao & Xu, 1998; Xu, 1996, 1998) based on the linear contrasts of genotypic values, no formal definitions and derivations of the linear contrasts are given by these authors. The linear contrasts are simply linear combinations of the original genetic effects, but they provide a convenient way to perform hypothesis tests. We think that it is essential to provide such information to interested researchers and students in the field.

2. Theory and Methods

(i) *Four-way cross and the threshold model of binary diseases*

The general linear model for genetic mapping of a quantitative trait is

$$y_j = \mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u} + e_j, \tag{1}$$

where y_j ($\forall j = 1,\ldots, n$) is the continuous phenotypic value of the jth individual in a mapping population of size n, $\mathbf{b}$ is a vector of fixed non-genetic effects (e.g. year, location), $\mathbf{X}_j$ is the design matrix for the non-genetic effects, $\mathbf{u}$ is a vector of fixed genetic effects, $\mathbf{Z}_j$ is the design matrix for the genetic effects and e_j is the residual error, assumed to be $N(0, \sigma^2)$.

If we assume that the four-way-cross family under investigation was initiated from $(L_1 \times L_2) \times (L_3 \times L_4)$, where L_i ($\forall i = 1,\ldots, 4$) represents the ith inbred line. The four-way-cross family contains four possible genotypes with three estimable genetic effects (Xu, 1996, 1998). These three effects are the allelic substitution between L_1 and L_2, denoted a_1, the allelic substitution between L_3 and L_4, denoted a_3, and the dominance effect denoted δ_{13}. The genotypic value for individual j may be denoted $g_j = \mathbf{Z}_j\mathbf{u}$, where the vector of genetic effects is $\mathbf{u} = [a_1\ a_3\ \delta_{13}]^{\mathrm{T}}$. Let us denote the four ordered genotypes as L_1L_3, L_1L_4, L_2L_3 and L_2L_4. The corresponding genotypic values for the four genotypes may be denoted by a vector $\mathbf{g} = [g_{13}\ g_{14}\ g_{23}\ g_{24}]^{\mathrm{T}}$. The design matrix is defined as $\mathbf{Z}_j = \mathbf{H}_i$ for $i = 1,\ldots, 4$, where $\mathbf{H}_i$ is the ith row of the following matrix:

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}.$$

Notice that each row of matrix $\mathbf{H}$ corresponds to one of the four possible genotypes. If individual j takes genotype L_1L_3 then its genotypic value will be

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

g_{13} and the design matrix will be $\mathbf{Z}_j = \mathbf{H}_1 = [1\ 1\ 1]$. If individual j takes genotypei L_2L_3 then its genotypic value will be g_{23} and the design matrix will be $\mathbf{Z}_j = \mathbf{H}_3 = [-1\ 1\ -1]$. The $\mathbf{H}$ notation of defining genotypes is a convenient way for deriving the EM algorithm (see Appendix A for the expectation step of the EM algorithm). Estimation of the genetic effects and hypothesis test are conducted using the EM algorithm and the likelihood ratio test statistic.

For binary disease traits, the phenotype is defined as a discrete Bernoulli variable (w) rather than as a continuous quantitative trait (y). For individual j, we define $w_j = 1$ if j is affected and $w_j = 0$ if j is normal. The threshold model serves as a link between the binary disease phenotype and a hypothetical underlying quantitative trait, denoted y_j. The actual connection between w and y is through the following threshold model: $w_j = 1_{(y_j > \mathbf{0})}$ (i.e. $w_j = 1$ if $y_j > 0$, otherwise $w_j = 0$). The binary disease phenotype and the liability have a one-to-one relationship. Mapping genes for w is now converted into a problem of mapping genes for y. As a result, we can take full advantage of the well developed QTL mapping procedures (Lander & Botstein, 1989). It should be emphasized that the residual variance of the liability cannot be estimated because of the unobserved nature of y and thus we are forced to make an assumption of $e_j \sim N(0, 1)$. Models of QTL mapping can be directly adopted here for disease mapping owing to the one-to-one relationship between w and y.

We denote the parameters of the generalized linear model given in Eqn 1 by a vector $\boldsymbol{\theta} = \{\mathbf{b}, \mathbf{u}\}$. We now have the following probit model for the disease trait:

$$\Pr(w_j = 1 | \mathbf{Z}_j, \boldsymbol{\theta}) = \Pr(y_j > 0 | \mathbf{Z}_j, \boldsymbol{\theta}) = 1 - \Phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u}), \tag{2}$$

where $\Phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})$ is the standardized normal distribution function. The probability of w_j is described by the Bernoulli distribution

$$\Pr(w_j | \mathbf{Z}_j, \boldsymbol{\theta}) = [1 - \Phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})]^{w_j} [\Phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})]^{\mathbf{1} - w_j}. \tag{3}$$

Notice that $\mathbf{Z}_j$ is missing, because we cannot normally observe the genotype of the disease locus for individual j. However, the probability $p_{j|i} = \Pr(\mathbf{Z}_j = \mathbf{H}_i | \mathbf{I}_M)$ for $i = 1, \ldots, 4$ is available, where $\mathbf{I}_M$ represents the marker information. This probability is calculated using the multipoint method (Rao & Xu, 1998), which is a special situation of the hidden Markov model for sib analysis when the marker linkage phases are known (Kruglyak & Lander, 1995*a*). The actual probability function for the jth individual is a mixture of four distributions

$$\Pr(w_j | \boldsymbol{\theta}) = \sum_{i=1}^{4} p_{j|i} [\Phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})]^{\mathbf{1} - w_j} \times [1 - \Phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})]^{w_j}. \tag{4}$$

The overall likelihood for the entire mapping population is

$$L(\boldsymbol{\theta}) = \prod_{j=1}^{n} \Pr(w_j | \boldsymbol{\theta}). \tag{5}$$

Notice that the above likelihood function is conditional on a fixed position (λ) of the QTL in question. The maximum likelihood estimate of λ takes the maximum value after we scan the entire genome.

(ii) *EM algorithm for parameter estimation*

Finding the solution of the above likelihood function is not straightforward. Therefore, we developed an EM algorithm that has an attractive iterative form. The EM algorithm is derived from the normal equation system, in which both $\mathbf{Z}_j$ and y_j are assumed to be observed. Under this assumption, the MLE of $\boldsymbol{\theta}$ is

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{Z}_j \\ \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}} y_j \\ \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}} y_j \end{bmatrix}. \tag{6}$$

When $\mathbf{y}$ is observed but $\mathbf{Z}$ is not, as in the usual QTL mapping studies, we can replace all terms involving $\mathbf{Z}$ by the expectations conditional on $\mathbf{y}$ and $\boldsymbol{\theta}$:

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathrm{E}(\mathbf{Z}_j) \\ \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}})\mathbf{X}_j & \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}} y_j \\ \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}}) y_j \end{bmatrix}. \tag{7}$$

When both $\mathbf{Z}$ and $\mathbf{y}$ are missing (the situation we are currently dealing with), we replace all terms involving $\mathbf{Z}$ and $\mathbf{y}$ by the expectations conditional on $\mathbf{w}$ and $\boldsymbol{\theta}$:

$$\begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathrm{E}(\mathbf{Z}_j) \\ \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}})\mathbf{X}_j & \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j) \end{bmatrix}^{-1} \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathrm{E}(y_j) \\ \sum_{j=1}^{n} \mathrm{E}(\mathbf{Z}_j^{\mathrm{T}} y_j) \end{bmatrix}, \tag{8}$$

where

$$\mathrm{E}(y_j) = \mathop{\mathrm{E}}_{Z}[\mathop{\mathrm{E}}_{y}(y_j | \mathbf{Z}_j, w_j, \boldsymbol{\theta})],$$

$$\mathrm{E}(\mathbf{Z}_j^{\mathrm{T}} y_j) = \mathop{\mathrm{E}}_{Z}[\mathbf{Z}_j^{T} \mathop{\mathrm{E}}_{y}(y_j | \mathbf{Z}_j, w_j, \boldsymbol{\theta})]$$

and

$$\mathop{\mathrm{E}}_{y}(y_j | \mathbf{Z}_j, w_j, \boldsymbol{\theta}) = \mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u} + (2w_j - 1)\frac{\phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})}{\Phi[(1 - 2w_j)(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})]}. \tag{9}$$

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

Eqn 9 is the expectation of a truncated normal variable given by Cohen (1991). Eqn 8 is the maximization step of the EM algorithm. The expectation step consists of evaluation of all the expectation terms. Detailed expressions of the various expectations involved in the normal equation are given in Appendix B.

Unlike most other statistical methods for parameter estimation, the EM-implemented ML method does not provide a straightforward way to calculate the variance–covariance matrix of the estimated parameters. The information-based method of Louis (1982) is commonly used in EM estimation and has been applied to QTL mapping by Kao and Zeng (1997) and Luo *et al.* (2003). We extended the method of Luo *et al.* (2003) to the binary disease mapping here in this study to calculate the information matrix. The variance–covariance matrix of the estimated parameters is then approximated by the inverse of the information matrix. The derivation of the information matrix is given in Appendix C.

3. Mapping genes for fibrosarcoma in the mouse

(i) *Experimental design*

Fibrosarcoma is a form of neoplasm of the fibroblasts that form connective tissue. In humans, it occurs rarely with an incidence of less than 1%. It also occurs in animals such as mice, cats and dogs. In the four-way-cross mouse population bred as the progeny of (BALB/c × C57BL/6) F_1 females and (C3H/He × DBA/2) F_1 males, it is diagnosed as the cause of death in 17% of virgin female mice and 5% of virgin male mice. The difference between males and females suggests that, in these mice, the incidence of fibrosarcoma might be influenced by hormone levels and, in fact, the incidence of fibrosarcoma as a cause of death drops to 9% in non-virgin females that have born multiple litters in their first six months of life.

The mice studied were generated from a four-way cross-breeding scheme. These animals were previously described in an analysis of QTL mapping for age-sensitive T-cell subset levels (Jackson *et al.*, 1999). Each of the 267 progeny was subjected to a comprehensive necropsy with histopathology at time of death. All mice were followed until death ($n=267$). 76% of the females died of some form of neoplasia (principally lymphoma, fibrosarcoma or mammary adenocarcinoma), and 42% of the males died of neoplasia (principally hepatocarcinoma, pulmonary adenocarcinoma, lymphoma or fibrosarcoma). We present here an analysis of the genetic influences on fibrosarcoma incidence as an example to demonstrate the utilities of the proposed method. Results of the analyses of other diseases will be reported elsewhere in the future.

The data set contains 96 co-dominant markers distributed among 20 chromosomes for the 267 progeny of the four-way-cross family, designated LAG1. The goal of the QTL analysis was to map loci that influenced the risk of lethal fibrosarcoma. The LAG1 family was derived from the crosses of four inbred lines: BALB/c, C57BL/6, C3H/He and DBA/2, with two rounds of crosses. In the first round, BALB/c was crossed with C57BL/6 to generate a set of F_1 (BALB/c × C57BL/6) females. All the F_1 individuals were genetically identical. Meanwhile, C3H/He was crossed with DBA/2 to generate a set of F_1 (C3H/He × DBA/2) males. Notice that the two sets of F_1 mice might contain different sets of genes. In the second round of crosses, the female F_1 (BALB/c × C57BL/6) mice were crossed with the male F_1 (C3H/He × DBA/2) mice to generate the four-way-cross F_2 progeny (BALB/c × C57BL/6) × (C3H/He × DBA/2). The four-way-cross progeny contained a maximum of four alleles at any given locus, which is identical to an outbred full-sib family except that the linkage phases of markers were uniquely identified in the four-way cross.

(ii) *Method of analysis*

The data were analysed with the ML method developed in this study. The non-genetic effects in the model include the population mean (b_1) and the effect of sex (b_2). The design matrix for the non-genetic effects is coded as $\mathbf{X}_j=[1\ 0]$ for females and $\mathbf{X}_j=[1\ 1]$ for males. For comparison, we also analysed the data using the reversible jump Bayesian method of Yi and Xu (2000*a*), which we modified to incorporate differences in sex-specific recombination fractions (details not shown). Unlike ML, Bayesian mapping does not provide a significance test or power analysis. Therefore, results generated from a Bayesian analysis should be interpreted in a slightly different way.

Most of the genotyped markers are fully informative (i.e. there are four unique alleles in the family). Some loci, however, are not fully informative: they segregate either in the male parents or in the female parents, but not both. Some progenies have missing genotypes. To handle these partially informative and missing markers properly when inferring the probability of disease genotype, we used the multipoint method (using all markers) of Rao and Xu (1998), which has been modified to take into account the sexual difference in recombination fractions (Wu *et al.*, 2002).

We used the maternal map as reference and wrote the paternal map over the maternal map by shrinking or expanding the paternal map distances proportionally. We only evaluate the QTL position along the maternal map. The corresponding position at the paternal map is translated from the maternal map using the procedure shown in the following three-locus example. If the maternal map shows the three loci at 0 cM, 10 cM and 25 cM positions, and the paternal

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

map shows the three loci at 0 cM, 15 cM and 21 cM positions, we know that 15 cM on the paternal map corresponds to 10 cM on the maternal map, and 21 cM on the paternal map corresponds to 25 cM on the maternal map. When we evaluate position 4 cM of the maternal map, the corresponding position on the paternal map is $0+[(15-0)/(10-0)]\times(4-0)=6$ cM. This is called map expanding. By contrast, when we evaluate position 20 cM on the maternal map, the corresponding position on the paternal map is $15+[(21-15)/(25-10)]\times(20-15)=17$ cM. This is called map shrinking. If we want to infer the QTL genotype probability at position 4 cM of the maternal map (6 cM of the paternal map) using the marker at position 0 cM as the left flanking marker (marker M) and the marker at position 10 cM (marker N) as the right flanking marker, the recombination fractions between the QTL and the left hand side markers are $c_M^d=0{\cdot}5\times\{1-\exp[-2\times|4-0|\div 100]\}=0{\cdot}0384$ for the dam and $c_M^s=0{\cdot}5\times\{1-\exp[-2\times|6-0|\div 100]\}=0{\cdot}0565$ for the sire. The recombination fractions between the QTL and the right hand side marker are $c_N^d=0{\cdot}5\times\{1-\exp[-2\times|4-10|\div 100]\}=0{\cdot}0565$ for the dam and $c_N^s=0{\cdot}5\times\{1-\exp[-2\times|6-15|\div 100]\}=0{\cdot}0824$ for the sire. The multipoint method requires a 4×4 transition matrix, which is constructed using the recombination fractions of both sexes (i.e. $\mathbf{T}=\mathbf{T}^s\otimes\mathbf{T}^d$), where

$$\mathbf{T}^s=\begin{bmatrix}1-c^s & c^s\\ c^s & 1-c^s\end{bmatrix},$$

and

$$\mathbf{T}^d=\begin{bmatrix}1-c^d & c^d\\ c^d & 1-c^d\end{bmatrix}.$$

This transition matrix can be found in Wu *et al.* (2002).

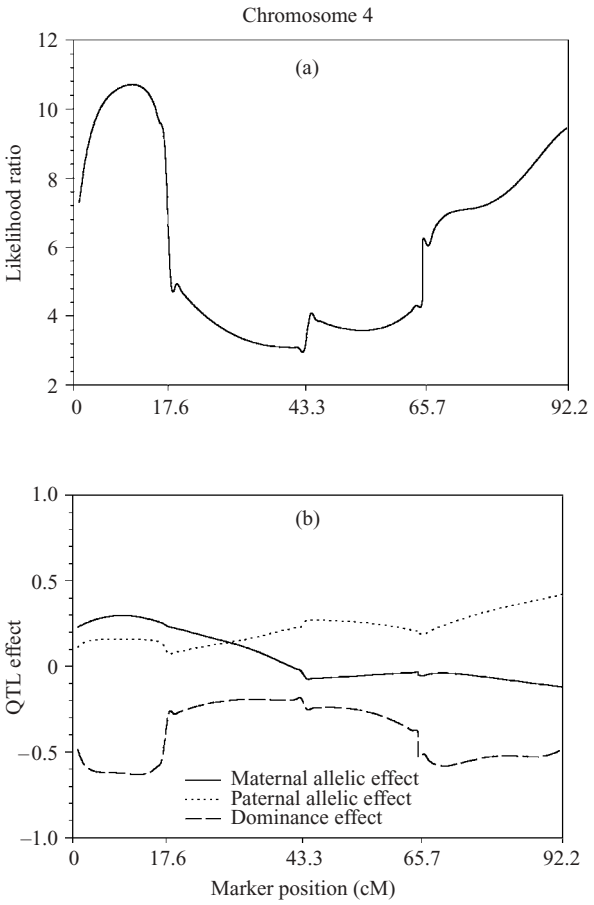

Fig. 1. Likelihood ratio test statistic profile (a) and the QTL effect profiles (b) along chromosome 4 using the maximum likelihood method. (b) The solid line is the estimated effect of allelic substitution from the dam, the dotted line is the estimated effect of allelic substitution from the sire, and the dashed line represents the estimate of the dominance effect. Labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left-hand end of the chromosome.

(iii) *Results*

In the ML analysis, the model only includes a single QTL. So, the entire genome was searched from one position to another with an increment of 1 cM. This type of one-dimensional grid searching has been found to be effective for locating multiple QTLs (Jannick & Jansen, 2001). Most chromosomes show very flat likelihood ratio test statistic profiles (data not shown). The only chromosome that shows evidence of a QTL is chromosome four (Fig. 1a). Two peaks appear in this profile, one at 11 cM and the other at 92 cM (end of the chromosome). The first peak has a likelihood ratio test statistic value of 11 and the second peak has a value of 10, corresponding to p values of 0·0117 and 0·0186, respectively. The QTL effect profiles are depicted in Fig. 1b and the estimated QTL effects are given in Table 1. The first QTL is mainly caused by the dominance effect followed by the paternal effect. The proportion of the total variance of the liability explained by the QTL is 34%. The second QTL is a result of equal contributions of the paternal and dominance effects, together explaining ~30% of the variance of liability. The negative estimate of the sex effect reflects the higher incidence of lethal fibrosarcoma in females than in males.

The Bayesian analysis identified two chromosomes with evidence of QTLs. The QTL intensity profiles (Sillanpaa & Arjas, 1998) and the QTL effects plotted against the chromosomal position are shown in Fig. 2 for chromosome four and in Fig. 3 for chromosome 13. The major peak on chromosome four overlaps with that identified by the ML method, showing the consistency between ML and Bayesian analyses. The second peak of the QTL intensity on chromosome four is close to that of the ML analysis (with a 3 cM deviation). However, the peak is much lower than the first peak. The Bayesian analysis identified an

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

Table 1. *Estimated genetic effects and locations of the identified QTLs. For the QTL positions, in each case the first number represents the chromosome and the second the position within that chromosome* (*e.g.* '*4/11 cM*' *means position 11 cM on chromosome 4*). *For QTL effects, the standard deviations of the maximum likelihood estimates and posterior standard deviations of the Bayesian estimates are given in parentheses*

	Maximum likelihood		Bayesian		
QTL parameter	QTL1	QTL2	QTL1	QTL2	QTL3
Position ($\hat{\lambda}$)	4/11 cM	4/92 cM	4/11 cM	4/89 cM	13/24 cM
Paternal ($\hat{a}_1$)	0·1599 (0·2120)	0·4202 (0·1543)	0·0896 (0·2229)	0·3069 (0·1822)	−0·0032 (0·1928)
Maternal ($\hat{a}_3$)	0·2949 (0·1954)	−0·1199 (0·2001)	0·2664 (0·2025)	−0·0868 (0·2114)	0·2249 (0·2088)
Dominance ($\hat{\delta}_{13}$)	−0·6285 (0·2242)	−0·4853 (0·2518)	−0·3597 (0·2508)	−0·3258 (0·2439)	−0·4746 (0·2322)

additional QTL on chromosome 13 (Fig. 3) which was not detected in the ML analysis. The estimated QTL parameters are listed in Table 1. Similar to that in the ML analysis, the first QTL identified with the Bayesian analysis is mainly caused by the dominance effect. However, the estimated dominance effect is less than that of the ML analysis. The proportion of disease variance explained by this QTL is 25% compared with the 34% estimated with the ML analysis. The second QTL identified by the Bayesian analysis is the result of both the dominance effect and the paternal effect, jointly explaining 21% of the variance of the liability. The third QTL identified is on chromosome 13, explaining 27% of the disease variance.

The probability of being affected by the disease conditional on the genotype is called the penetrance. In our model, the penetrance of the ith genotype is defined as $\Phi(b_1+\mathbf{H}_i\mathbf{u})$ for females and $\Phi(b_1+b_2+\mathbf{H}_i\mathbf{u})$ for males. The estimated penetrances for the four genotypes for each of the two sexes are given in Table 2. There is a large effect of sex apparent in the difference between the estimates of the male and female mice. Consider the penetrances of QTL1 detected by the ML method in the female mice. The penetrance of genotype L_2L_3 is 0·5253 in the female mice, but 0·2399 in the male mice. The effect of the sex of the mouse upon the disease penetrance estimates varies between genotypes and loci.

4. Discussion

One complication in the four-way-cross mapping comes from the sexual differences in the marker map. For the set of markers used in the analysis, the marker orders are the same for both sexes, but the recombination fractions between markers are quite different for the different sexes. Although methods have been developed to construct consensus maps from different sexes or different families (Beavis & Grant, 1991; Bucther & Moran, 2000; Butcher *et al*., 2002), if the sexual difference in the map is real and not due to sampling errors, the consensus map (also called the sex average map) might not be useful to QTL mapping. We actually analysed the data using a consensus map and did not find any evidence of QTLs at all (data not shown). In the four-way-cross experiment of mice, the recombination fractions between markers were estimated separately for different sexes, and thus we have two sex-specific marker maps. There has been no attempt in QTL mapping to incorporate sex-specific maps. The method of Haley *et al*. (1994) was designed to handle situations in which the male and female parents have different marker genotypes, but the recombination fractions between markers must be the same for the male and female parents. Wu *et al*. (2002) and Fann and Ott (1995) developed a method to estimate recombination fractions simultaneously for different sexes. Although these methods are not for QTL mapping, we adopted the method of Wu *et al*. (2002) by using a heterogeneous transition matrix to take into account the sexual dimorphism in the marker map.

Many complex traits show a binary or categorical phenotypic distribution. Mapping loci of such traits requires methods that specifically take into account these phenotypic distributions. The key difference comes from the difference in the conditional distributions under investigation. In traditional disease mapping, the conditional distribution of interest is the distribution of the genotype of putative locus conditional on the disease phenotype. If the genotypic distribution of the 'affected' population is different from that of the 'normal' population, the locus under investigation might be responsible for the variation of the disease trait. In QTL mapping, however, the conditional distribution of interest is the distribution of the phenotype conditional on the genotype of the locus under study. If the phenotypic distribution varies between different genotypes, the locus might be associated with the trait. The difference in the way that the conditional distribution is formulated leads to different level of generality of the methods. The 'conditional phenotype given genotype' model is the quantitative genetics model, and it is more general

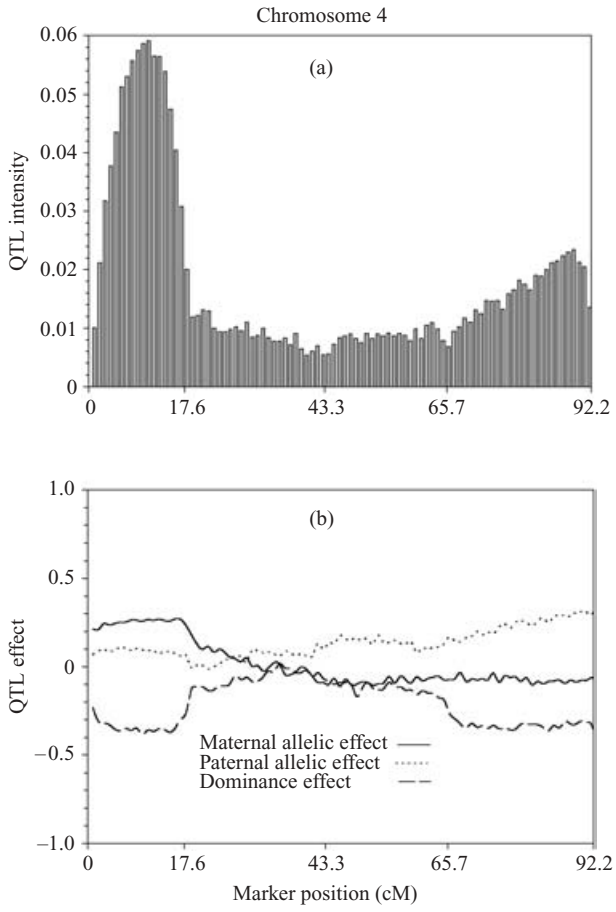

Fig. 2. QTL intensity profile (a) and the QTL effect profiles (b) along chromosome 4 using the Bayesian method. One QTL was supported in this chromosome with highest posterior probability of 0·61. The corresponding posterior probability of two QTLs was 0·25. (b) The solid line is the estimated effect of allelic substitution from the dam, the dotted line is the estimated effect of allelic substitution from the sire, and the dashed line represents the estimate of the dominance effect. Labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left end of the chromosome.

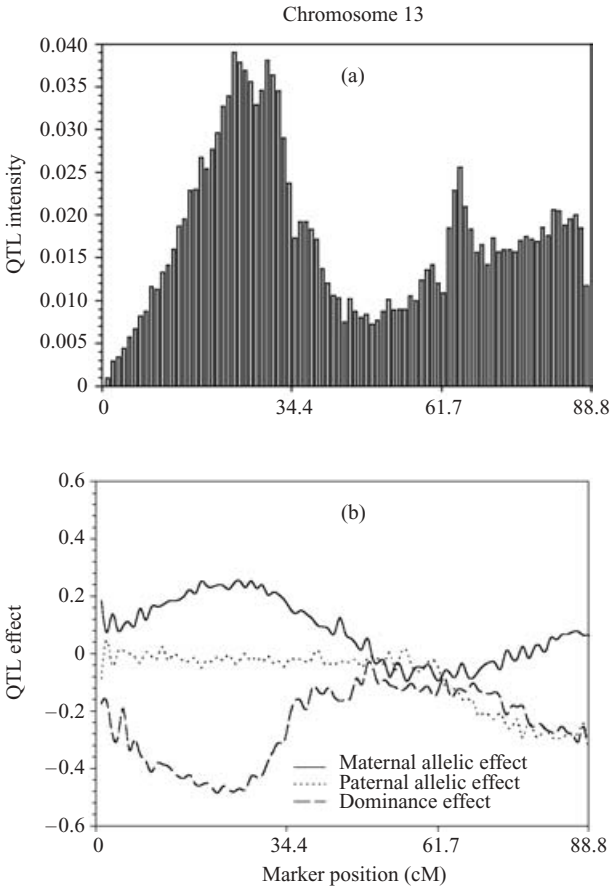

Fig. 3. QTL intensity profile (a) and the QTL effect profiles (b) along chromosome 13 using the Bayesian method. The method supported one QTL with a posterior probability of 0·43. The corresponding probability of two QTLs was 0·31. (b) The solid line is the estimated effect of allelic substitution from the dam, the dotted line is the estimated effect of allelic substitution from the sire, and the dashed line represents the estimate of the dominance effect. Labels of the horizontal axis indicate the marker positions measured in centiMorgans (cM) counted from the left end of the chromosome.

and flexible in terms of handling multiple loci, non-genetic effects, genotype by environment interaction and so on. The 'conditional genotype given phenotype' model is the disease genetics model and it is not as general as the quantitative genetics model. The generalized linear model of disease trait analysis investigated in this study uses the 'conditional phenotype given genotype' model but for disease traits. It bears the attractive flexibility and generality of QTL mapping.

Under the generalized linear model, a disease trait is formulated as a disease liability, which is not much different from a regular quantitative trait except that the liability is simply unobservable. Under the liability model, all quantitative genetics theory can be applied to disease genetics. The unobservable nature of the liability is a typical missing value problem. The EM algorithm of Dempster *et al.* (1977) was specifically designed to solve such a problem, and we took advantage of this particular nature of missing value and formulated the EM algorithm for disease mapping. The linear model given in Eqn 1 provides the theoretical foundation for all genetic mapping. In marker analysis or any typical linear regression analysis, both **X** and **Z** are observable. In QTL mapping, however, **Z** is no longer observable, but the EM algorithm has incorporated the distribution of **Z** into the model by treating **Z** as a missing value so that the usual linear model analysis applies. For disease mapping as we formulated it in this study, in addition to **Z**, the dependent variable **y** is also not observable. There is no reason that we cannot also treat **y** as a missing value. When **y** is indeed treated as a missing value, as shown in this study, the MLE can be obtained using an EM iteration system just as neat as

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

Table 2. *Estimated disease penetrances of the four genotypes for the identified QTLs*

		Maximum likelihood		Bayesian		
Sex	Genotype	QTL1	QTL2	QTL1	QTL2	QTL3
Female	L_1L_3	0·1911	0·1881	0·1966	0·1696	0·1350
	L_1L_4	0·4182	0·6276	0·2523	0·4481	0·2731
	L_2L_3	0·5253	0·2251	0·3769	0·1793	0·4414
	L_2L_4	0·03726	0·0686	0·0587	0·0814	0·0610
Male	L_1L_3	0·0501	0·0489	0·0654	0·0534	0·0392
	L_1L_4	0·1644	0·3283	0·0927	0·2155	0·1036
	L_2L_3	0·2399	0·0636	0·1659	0·0576	0·2105
	L_2L_4	0·0053	0·0120	0·0131	0·0200	0·0138

that of the QTL mapping. Interestingly, we have shown that we can conduct a linear regression analysis with no direct observation on both the independent and dependent variables. The partial information for **Z** is obtained from markers and the partial information for **y** is obtained from the binary disease phenotype.

Bayesian mapping for binary disease traits is also formulated as problem of missing **y**, but **y** was recovered from random sampling (Yi & Xu, 2000*a*). Given the fact that only the expectation of **y** conditional on the disease phenotype is required in the EM and the expectation has an explicit form, the EM algorithm is many times faster than the MCMC-implemented Bayesian method. The ML and Bayesian analyses should coexist because there are many unshared characteristics of the two methods. As well as computing speed, ML analysis allows a significance test, whereas Bayesian analysis does not provide an easy way for significance testing. In the mouse data analysis, the Bayesian analysis appears to identify one additional QTL, indicating that Bayesian mapping might be superior to ML. However, the significance of this putative locus is hard to assess. If this additional QTL is real, the high efficiency of Bayesian analysis must be due to the ability to fit multiple QTLs simultaneously. Although ML can also handle multiple QTLs (Kao *et al.*, 1999), there has been no attempt to do so in four-way-cross experiments. Such a multiple QTL model for four-way crosses will certainly be a welcome project in the future.

Yi and Xu (1999) developed an ML method for binary trait mapping. The method of Rao and Xu (1998) is also ML based. The difference between these methods and the one proposed here is that they ignored the mixture distribution of the QTL genotype and replaced the mixture distribution by a heterogeneous residual variance. They can be classed as quasi-likelihood methods. As a result, they were able to use the Fisher scoring method to find the MLE. A similar Fisher scoring method cannot be used if the mixture distribution is taken into account. The original binary trait mapping procedure developed by Xu and Atchley (1996) was implemented via the simplex algorithm, which directly searches for the MLE. The simplex algorithm, however, is complicated, and users might have to download an existing subroutine without understanding its operation. The EM algorithm developed in this study is visible to users with only preliminary background in linear regression analysis.

The data we analysed happen to be collected from a four-way-cross family. Therefore, the statistical model is described in the context of a four-way cross. A four-way-cross family is similar to a full-sib family derived from the mating of two outbred parents. The only difference between the two different families is that marker linkage phases are usually given in the four-way cross, whereas, in the full-sib family, the linkage phases must be inferred from the data. A full-sib family is the simplest form of an outbred pedigree. Therefore, the method described here has a strong implication in genetic mapping for outbred species (e.g. forest trees and large animals). Compared with the F_2 design derived from the cross of two inbred lines, the four-way cross allows the estimation of one more additive genetic effect. This property might help to harvest more QTLs because those QTLs that do not segregate in one cross but do segregate in the other might still be detected in the four-way cross. The four-way-cross design is also relevant to genetic mapping in multiple line crossing experiments. The diallel crossing design of Rebai and Goffinet (1993) is one form of the multiple line cross. The four-way-cross design differs from the diallel design in that not all the six ($C_4^2 = 4 \times 3 \div 2 = 6$) possible combinations of the four inbred lines are used in mapping. In fact, only two combinations (two F_1 plants each from a different cross) are used in the four-way cross. In addition, the two F_1s from different crosses are further crossed to generate an F_2 family for mapping. If resources permit, one should pursue the diallel cross because

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

it allows more allelic substitution effects to be detected. Current methods for genetic mapping in diallel crosses have only been explored under the least-squares framework (Rebai & Goffinet, 1993). They should be investigated under the ML framework also, and the EM algorithm proposed here provides a clue for the extension. If a crossing experiment involves multiple inbred lines, Bayesian methods have been suggested by Yi and Xu (2002) and Bink *et al.* (2002). The Bayesian methods can handle arbitrarily complicated mating designs but, again, the price is the expensive computing time.

The ML analysis detected two regions with evidence of QTLs and the Bayesian analysis detected one additional QTL on a different chromosome. All the QTLs detected show major contribution from allelic interaction (dominance). This discovery contrasts with the common belief that additive effects usually play a major role in quantitative trait variation. However, the disease trait itself is not a quantitative trait but is formulated as controlled by an underlying quantitative trait. It is possible that most disease traits show the same behaviour of large contribution by dominance effects. More examples are needed to make such a conclusion. These dominance effects would not be detected if the analyses had been done separately for different sexes. This clearly demonstrates the advantage of the four-way crosses over the backcross analyses. The success of mapping genes with dominance effects depends on the lines selected to initiate the crosses. Using four-way crosses for QTL mapping is certainly an efficient way to increase the number of interactions and thus to improve the power of detecting dominance effects.

A commonly used method for disease mapping in human is the transmission disequilibrium test (TDT) method (Monks *et al.*, 1998; Spielman *et al.*, 1993). This method involves a family based association study that is designed to separate the confounding effects between the linkage disequilibrium caused by true linkage and that caused by population mixture or stratifications. The four-way-cross method described here differs from the TDT in the following ways: (1) TDT is an association study whereas the four-way-cross analysis is a linkage-mapping study; (2) TDT deals with multiple censored families (with affected parents and sibs) whereas the four-way-cross analysis deals with a single family with all progeny included in the analysis; (3) TDT uses the conditional distribution of the genotype given the phenotype whereas the four-way cross uses the conditional distribution of the phenotype given the genotype.

De Koning *et al.* (2002) developed a method for mapping imprinted QTLs using F_2 like full-sib families. The four-way-cross family described in this study cannot be used for this purpose. Genomic imprinting is defined as a phenomenon in which the gene expression of progeny depends on the parental origin of the alleles. If the genotypic value of the ordered heterozygote *Aa* is different from that of *aA*, the locus is said to be imprinted. These two heterozygotes cannot be separated in classical F_2 mapping. However, they can be separated if the male and female parents have different genotypes in markers (Haley *et al.*, 1994). In a four-way-cross family, the two parents do have a chance to have different marker genotypes. However, there is no reason to assume that the two parents also have the same QTL genotype. Therefore, it is impossible to detect imprinted QTLs using four-way crosses. The full-sib family investigated by Haley *et al.* (1994) is special because both parents are F_1 hybrids of two outbred populations. There is good reason to assume that QTLs are fixed for the two outbred populations because of divergent selection, although markers might segregate within each population.

The threshold model is an example of the general latent class model. There are other latent class models that can be equally applied here. These models include the logit and log–linear models. We adopted the probit analysis because it has some nice properties. With the probit transformation from the latent variable to the dichotomous phenotype, the latent variable has a normal distribution and, as a result, a QTL mapping approach can be directly adopted here on the latent variable. The other nice property of the probit model is that an EM algorithm can be used to search for the ML solutions, as demonstrated in this study. With the other latent models, alternative non-EM methods must be considered (Galecki *et al.*, 2001; Molenberghs & Goetghebeur, 1997).

Appendix

(A) *Derivation of the four-way-cross model*

The four possible genotypes generated from the four-way cross $(L_1 \times L_2) \times (L_3 \times L_4)$ are L_1L_3, L_1L_4, L_2L_3 and L_2L_4. If we denote the genotypic values of the four genotypes in the above order by a vector $\mathbf{g} = [g_{13}\ g_{23}\ g_{14}\ g_{24}]^T$. Each genotypic effect can be decomposed into the sum of the two allelic effects plus a deviation reflecting the interaction between the two alleles, called dominance effect. Assume that the F_1 derived from $(L_1 \times L_2)$ serves as the male parent (sire) and the F_1 derived from $(L_3 \times L_4)$ serves as the female parent (dam). Let us further define the allelic values of L_1, L_2, L_3 and L_4 by a_1, a_2, a_3 and a_4, respectively. The genotypic value is expressed as

$$g_{pq} = a_p + a_q + \delta_{pq} \ (\forall p = 1, 2; q = 3, 4), \qquad \text{(A1)}$$

where δ_{pq} is the dominance effect. The four-way-cross model (A1) is a special case of the general set-up for association studies developed by Nielsen and Weir (1999). Notice that there are four possible genotypes

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

in the progeny but, after the decomposition, we have eight parameters. Therefore, we must impose some restrictions on the parameter space to make the model estimable. We take the restrictions identical to those used in a 2×2 factorial design (Steel & Torrie, 1980); that is, $a_2 = -a_1$, $a_4 = -a_3$, and $\delta_{13} = \delta_{24} = -\delta_{14} = -\delta_{23}$. We now have only three independent parameters: a_1, a_3 and δ_{13}. Substituting for a_2 with $-a_1$, for a_4 with $-a_3$ and all δ_{pq}s by δ_{13} in Eqn A1, we have

$$\begin{bmatrix} g_{13} \\ g_{14} \\ g_{23} \\ g_{24} \end{bmatrix} = \begin{bmatrix} a_1 + a_3 + \delta_{13} \\ a_1 - a_3 - \delta_{13} \\ -a_1 + a_3 - \delta_{13} \\ -a_1 - a_3 + \delta_{13} \end{bmatrix}. \tag{A2}$$

Define $\mathbf{u} = [a_1\ a_3\ \delta_{13}]^{\mathrm{T}}$ and

$$\mathbf{H} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}.$$

Eqn A2 can thus be expressed as $\mathbf{g} = \mathbf{H}\mathbf{u}$.

(B) *Derivation of the expectation terms in the EM equations*

If we denote the posterior probability of $\mathbf{Z}_j = \mathbf{H}_i$ conditional on both the phenotype and the parameters by $p^*_{j|i}$, which is different from $p_{j|i}$ and is derived as follows. Let

$$\Pr(w_j|\mathbf{H}_i, \boldsymbol{\theta}) = [1 - \Phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})]^{w_j}[\Phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})]^{1-w_j}. \tag{B1}$$

The posterior probability is obtained based on the following Bayes' theorem,

$$p^*_{j|i} = \frac{p_{j|i} \Pr(w_j|\mathbf{H}_i, \boldsymbol{\theta})}{\sum_{k=1}^{4} p_{j|k} \Pr(w_j|\mathbf{H}_k, \boldsymbol{\theta})}. \tag{B2}$$

The expectations involving $\mathbf{Z}$ and $\mathbf{y}$ are

$$\mathrm{E}(\mathbf{Z}_j) = \sum_{i=1}^{4} p^*_{j|i}\mathbf{H}_i, \tag{B3}$$

$$\mathrm{E}(\mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j) = \sum_{i=1}^{4} p^*_{j|i}\mathbf{H}_i^{\mathrm{T}}\mathbf{H}_i, \tag{B4}$$

$$\mathrm{E}(y_j) = \sum_{i=1}^{4} p^*_{j|i} \mathop{\mathrm{E}}_{y}(y_j|\mathbf{H}_i, w_j, \boldsymbol{\theta}), \tag{B5}$$

$$\mathrm{E}(\mathbf{Z}_j^{\mathrm{T}}y_j) = \sum_{i=1}^{4} p^*_{j|i}\mathbf{H}_i^{\mathrm{T}} \mathop{\mathrm{E}}_{y}(y_j|\mathbf{H}_i, w_j, \boldsymbol{\theta}), \tag{B6}$$

$$\begin{aligned} \mathop{\mathrm{E}}_{y}(y_j|\mathbf{H}_i, w_j, \boldsymbol{\theta}) &= \mathop{\mathrm{E}}_{y}(y_j|\mathbf{Z}_j = \mathbf{H}_i, w_j, \boldsymbol{\theta}) \\ &= \mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u} + (2w_j - 1)\frac{\phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})}{\Phi[(1-2w_j)(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})]}, \end{aligned} \tag{B7}$$

where $\phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})$ and $\Phi(\mathbf{X}_j\mathbf{b} + \mathbf{H}_i\mathbf{u})$ are the standardized normal density and probability functions, respectively.

(C) *Variance–covariance matrix of the EM estimates*

Louis's information matrix (Louis, 1982) requires the first and second partial derivatives of the complete data log-likelihood function with respect to the parameters. The complete data likelihood in binary disease mapping is the one when both $\mathbf{Z}$ and $\mathbf{y}$ are assumed to be known. In this case, the (negative) first partial derivative (expressed as a vector) is

$$S(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{y}) = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{Z}_j \\ \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \end{bmatrix} - \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}y_j \\ \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}y_j \end{bmatrix}, \tag{C1}$$

and the (negative) second partial derivative (expressed as a symmetric matrix) is

$$B(\boldsymbol{\theta}, \mathbf{Z}, \mathbf{y}) = \begin{bmatrix} \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{X}_j^{\mathrm{T}}\mathbf{Z}_j \\ \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{X}_j & \sum_{j=1}^{n} \mathbf{Z}_j^{\mathrm{T}}\mathbf{Z}_j \end{bmatrix}. \tag{C2}$$

The observed information matrix (Louis, 1982) evaluated at $\hat{\boldsymbol{\theta}}$ (the MLE of $\boldsymbol{\theta}$) is

$$I(\hat{\boldsymbol{\theta}}) = \mathrm{E}\{B(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\} - \mathrm{E}\{S(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})S^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\}. \tag{C3}$$

The expectations are taken with respect to $\mathbf{Z}$ and $\mathbf{y}$ conditional on $\mathbf{w}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. The variance covariance matrix of $\hat{\boldsymbol{\theta}}$ may be approximated by $\mathrm{Var}(\hat{\boldsymbol{\theta}}) \approx \Gamma^{1}(\hat{\boldsymbol{\theta}})$.

Notice that $\mathrm{E}\{B(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\}$ is not difficult to evaluate, but evaluation of $\mathrm{E}\{S(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})S^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\}$ is complicated. When $\mathbf{y}$ is observed, as in the usual QTL mapping studies, Kao and Zeng (1997) provided explicit expression for $\mathrm{E}\{S(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})S^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\}$, although the formula is lengthy and complicated. Luo *et al.* (2003) proposed to evaluate this expectation by taking advantage of the Monte Carlo technique. They simulated a large number of $\mathbf{Z}$ values and used the Monte Carlo sample mean as the approximate of this expectation. The Monte Carlo method is invoked only after the EM algorithm has converged and only at the positions where QTLs have been detected. Therefore, it does not add significant computational burden to the existing EM algorithm.

The method of Luo *et al.* (2003) is directly adopted in this study by simulating not only $\mathbf{Z}$ but also $\mathbf{y}$ for the Monte Carlo approximation of $\mathrm{E}\{S(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})S^{\mathrm{T}}(\hat{\boldsymbol{\theta}}, \mathbf{Z}, \mathbf{y})\}$. The liability for the jth individual, y_j, is simulated from a truncated normal distribution. We adopt the inverse transformation

https://doi.org/10.1017/S0016672303006414 Published online by Cambridge University Press

method that has an acceptance rate of 100% (Rubinstein, 1981). With this method, we first simulated a variable u from U(0, 1). We then defined $v = 1 - \Phi(\mathbf{X}_j\mathbf{b} + \mathbf{Z}_j\mathbf{u})$. Finally, we took the inverse function of the standardized normal distribution to obtain

$$y_j = w_j\Phi^{-1}[v + u(1-v)] + (1-w_j)\Phi^{-1}(uv). \qquad \text{(C4)}$$

Intrinsic functions for both $\Phi(\cdot)$ and $\Phi^{-1}(\cdot)$ are available in many computer software packages. For example, in the SAS package (SAS Institute, 1999), $\Phi(x)$ is coded as $\Phi(x) = \text{probnorm}(x)$ and $\Phi^{-1}(u)$ is coded as $\Phi^{-1}(u) = \text{probit}(u)$.

We thank two anonymous reviewers for their constructive comments and suggestions on the first two versions of the manuscript. This research was supported by the National Institutes of Health Grants GM55321 (SX) and AG11687 (RAM), and by the USDA National Research Initiative Competitive Grants Program 00–35300–9245 (SX).

References

Beavis, W. D. & Grant, D. (1991). A linkage map based on information from F_2 population of maize (*Zea mays* L.). *Theoretical and Applied Genetics* **82**, 636–644.

Bink, M. C. A. M., Uimari, P., Sillanpaa, M. J., Janns, L. L. G. & Jansen, R. C. (2002). Multiple QTL mapping in related plant populations via a pedigree analysis approach. *Theoretical and Applied Genetics* **104**, 751–762.

Bucther, P. A. & Moran, G. F. (2000). Genetic linkage mapping in *Acacia mangium*. 2. Development of an integrated map from two outbred pedigrees using RFLP and microsatellite loci. *Theoretical and Applied Genetics* **101**, 594–605.

Butcher, P. A., Williams, E. R., Whitaker, D., Ling, S. & Speed, T. P. (2002). Improving linkage analysis in outcrossed forest trees – an example from *Acacia mangium*. *Theoretical and Applied Genetics* **104**, 1185–1191.

Cohen, A. C. (1991). *Truncated and Censored Samples*. New York: Marcel Dekker.

de Koning, D.-J., Bovenhuis, H. & van Arendonk, J. A. M. (2002). On the detection of imprinted quantitative trait loci in experimental crosses of outbred species. *Genetics* **161**, 931–938.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B* **39**, 1–38.

Fann, C. J. & Ott, J. (1995). Parsimonious estimation of sex-specific map distances by stepwise maximum likelihood regression. *Genomics* **29**, 571–575.

Galecki, A., Ten Have, A. T. & Molenberghs, G. (2001). A simple and fast alternative to the EM algorithm for incomplete categorical data and latent class models. *Computational and Statistical Data Analysis* **35**, 265–281.

Hackett, C. A. & Weller, J. I. (1995). Genetic mapping of quantitative trait loci for traits with ordinal distributions. *Biometrics* **51**, 1252–1263.

Haley, C. S., Knott, S. A. & Elsen, J.-M. (1994). Mapping quantitative trait loci in crosses between outbred lines using least squares. *Genetics* **136**, 1195–1207.

Jackson, A. U., Fornes, A., Galecki, A., Miller, M. E. & Burke, D. T. (1999). Multiple-trait quantitative trait loci analysis using a large mouse sibship. *Genetics* **151**, 785–795.

Jannick, J.-L. & Jansen, R. C. (2001). Mapping epistatic quantitative trait loci with one-dimensional genome searches. *Genetics* **157**, 445–454.

Kao, C.-H. & Zeng, Z.-B. (1997). General formulas for obtaining the MLEs and the asymptotic variance–covariance matrix in mapping quantitative trait loci when using the EM algorithm. *Biometrics* **53**, 653–665.

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

Knott, S. A., Neale, D. B., Sewell, M. M. & Haley, C. S. (1997). Multiple marker mapping of quantitative trait loci in an outbred pedigree of loblolly pine. *Theoretical and Applied Genetics* **94**, 810–820.

Kruglyak, L. & Lander, E. S. (1995*a*). Complete multipoint sib-pair analysis of qualitative and quantitative traits. *American Journal of Human Genetics* **57**, 439–454.

Kruglyak, L. & Lander, E. S. (1995*b*). A nonparametric approach for mapping quantitative trait loci. *Genetics* **139**, 1421–1428.

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

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

Luo, L., Mao, C. & Xu, S. (2003). Correcting the bias in estimation of genetic variances contributed by individual QTL. *Genetica* (in press).

McIntyre, L. M., Coffman, C. J. & Doerge, R. W. (2001). Detecting and localization of a single binary trait locus in experimental populations. *Genetical Research* **78**, 79–92.

Molenberghs, G. & Goetghebeur, E. (1997). Simple fitting algorithms for incomplete categorical data. *Journal of the Royal Statistical Society B* **59**, 401–414.

Monks, S. A., Kaplan, N. L. & Weir, B. S. (1998). A comparative study of sibship tests of linkage and/or association. *American Journal of Human Genetics* **63**, 1507–1516.

Nelder, J. A. & Mead, R. (1965). A simplex method for function minimization. *Computer Journal* **7**, 308–313.

Nielsen, D. M. & Weir, B. S. (1999). A classical setting for associations between markers and loci affecting quantitative traits. *Genetical Research* **74**, 271–277.

Rao, S. Q. & Xu, S. (1998). Mapping quantitative trait loci for categorical traits in four-way crosses. *Heredity* **81**, 214–224.

Rebai, A. & Goffinet, B. (1993). Power of tests for QTL detection using replicated progenies derived from a diallel cross. *Theoretical and Applied Genetics* **86**, 1014–1022.

Rubinstein, R. Y. (1981). *Simulation and the Monte Carlo Method*. New York: John Wiley & Sons.

SAS Institute (1999). *SAS/IML User's Guide, Version 8*. Cary: SAS Institute.

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

Spielman, R. S., McGinnis, R. E. & Ewens, W. J. (1993). Transmission test for linkage disequilibrium: the insulin gene region and the insulin-dependent diabetes mellitus (IDDM). *American Journal of Human Genetics* **52**, 506–516.

Steel, R. G. D. & Torrie, J. H. (1980). *Principles and Procedures of Statistics*. New York: McGraw–Hill Book Company.

Visscher, P. M., Haley, C. S. & Knott, S. A. (1996). Mapping QTLs for binary traits in backcross and F_2 populations. *Genetical Research* **68**, 55–63.

Wu, R., Ma, C.-X., Wu, S. S. & Zeng, Z.-B. (2002). Linkage mapping of sex-specific differences. *Genetical Research* **79**, 85–96.

Xu, S. (1996). Mapping quantitative trait loci using four-way crosses. *Genetical Research* **68**, 175–181.

Xu, S. (1998). Iteratively reweighted least squares mapping of quantitative trait loci. *Behavior Genetics* **28**, 341–355.

Xu, S. & Atchley, W. R. (1996). Mapping quantitative trait loci for complex binary diseases using line crosses. *Genetics* **143**, 1417–1424.

Xu, S., Yonash, N., Vallejo, R. L. & Cheng, H. H. (1998). Mapping quantitative trait loci for binary traits using a heterogeneous residual variance model: an application to Marek's disease susceptibility in chickens. *Genetica* **104**, 171–178.

Yi, N. & Xu, S. (1999). Mapping quantitative trait loci for complex binary traits in outbred populations. *Heredity* **82**, 668–676.

Yi, N. & Xu, S. (2000*a*). Bayesian mapping of quantitative trait loci for complex binary traits. *Genetics* **155**, 1391–1403.

Yi, N. & Xu, S. (2000*b*). Bayesian mapping of quantitative trait loci under the identity-by-descent-based variance component model. *Genetics* **156**, 411–422.

Yi, N. & Xu, S. (2002). Linkage analysis of quantitative trait loci in multiple line crosses. *Genetica* **114**, 217–230.