A Random Model Approach to Mapping Quantitative Trait Loci for Complex Binary Traits in Outbred Populations

Nengjun Yi and Shizhong Xu

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

Manuscript received January 20, 1999
Accepted for publication June 17, 1999

ABSTRACT

Mapping quantitative trait loci (QTL) for complex binary traits is more challenging than for normally distributed traits due to the nonlinear relationship between the observed phenotype and unobservable genetic effects, especially when the mapping population contains multiple outbred families. Because the number of alleles of a QTL depends on the number of founders in an outbred population, it is more appropriate to treat the effect of each allele as a random variable so that a single variance rather than individual allelic effects is estimated and tested. Such a method is called the random model approach. In this study, we develop the random model approach of QTL mapping for binary traits in outbred populations. An EM-algorithm with a Fisher-scoring algorithm embedded in each E-step is adopted here to estimate the genetic variances. A simple Monte Carlo integration technique is used here to calculate the likelihood-ratio test statistic. For the first time we show that QTL of complex binary traits in an outbred population can be scanned along a chromosome for their positions, estimated for their explained variances, and tested for their statistical significance. Application of the method is illustrated using a set of simulated data.

METHODS of QTL mapping for normally distributed quantitative traits are well developed. These methods can be classified into two categories: the fixed model and the random model approaches. Data collected from well-designed crossing experiments, *e.g.*, F_2, backcrossing, or four-way cross, are usually analyzed using the fixed model approach. Usually only a single family in a line cross is analyzed. With these crossing experiments, the parental marker genotypes, the linkage phases of marker loci in the parents, and the number of alleles of putative QTL are known precisely. Under the fixed model, we express the effects of QTL as differences of genotypic means and then estimate and test these QTL effects (Lander and Botstein 1989; Haley and Knott 1992; Jansen 1993; Zeng 1994). In many species, such as large domesticated animals, forest trees, and human beings, we cannot develop inbred lines and manipulate line crosses; instead, we must conduct QTL mapping using data as they exist. Therefore, the fixed model approach of QTL mapping is hard to implement in the unmanipulated outbred populations. The random model approach, which treats the allelic effects of QTL as random variables, requires little knowledge about the number of QTL alleles and marker linkage phases. As a result, the random model approach is more plausible than the fixed model approach for QTL mapping in outbred populations (Xu and Atchley 1995; Xu 1996a). Under the random model, variance of QTL segregation rather than the effects is estimated and tested (Haseman and Elston 1972; Goldgar 1990; Schork 1993; Fulker and Cardon 1994; Xu and Atchley 1995; Grignola *et al.* 1996).

Many characters of biological interest and economical importance that are not inherited in a simple Mendelian fashion vary in a dichotomous or binary form. These traits are called complex binary traits. A complex binary trait is presumably controlled by several genes with its expression modified by environmental effects. They therefore belong to the category of quantitative traits (Falconer and Mackay 1996; Lynch and Walsh 1998). An appealing model for genetic analysis of complex binary data is based on the threshold concept, first used in a genetic context by Wright (1934). In the threshold model, it is postulated that there exists a latent or underlying continuous variable, called the liability, which controls the discrete phenotype. The binary phenotype and the continuous liability are linked through a fixed but unknown threshold. When the value of the liability is above the threshold, an individual shows one phenotype, *e.g.*, affected; otherwise, it will show the other phenotype, *e.g.*, normal. The liability is considered as a regular quantitative trait whose variance can be partitioned into genetic and environmental components. In principle, the existing theory of quantitative genetics developed for continuous traits holds similarly for the liability of a binary trait.

QTL mapping for binary traits is more challenging than for normal traits due to the nonlinear relationship between the observed phenotype and the unobservable genetic effects. Methods of linkage studies for binary

Corresponding author: Shizhong Xu, Department of Botany and Plant Sciences, University of California, Riverside, CA 92521.
E-mail: xu@genetics.ucr.edu

traits are well developed in human populations, and the affected sib-pair method (Olson 1995) is the most popular one of this kind. The method does not depend on the threshold model. As a consequence, it can identify the existence of a QTL but does not provide an estimate of the size (variance) of the QTL. In addition, the affected sib-pair method cannot be applied to plants and animals that typically have large family sizes. Recently, parametric methods of QTL mapping based on a generalized linear model (GLM) have been developed in simple line crosses (Hackett and Weller 1995; Visscher *et al.* 1996; Xu and Atchley 1996a; Rebai 1997). Rao and Xu (1998) extended the methods to four-way crosses. These methods are primarily derived using a single family of line cross.

Combining data from multiple families is deemed to be more useful in outbred populations (Muranty 1996; Xie *et al.* 1998; Xu 1998). For example, animal and plant breeders usually combine data from many half- or full-sib families. The main advantages of QTL mapping using multiple families are the increased power of QTL detection and the broader statistical inference space of the estimated QTL variances. In principle, the fixed model can also be used to analyze data from multiple families where the effect of allelic substitution of the QTL for each parent is estimated and tested. We have developed such a fixed model method and shown that the method is efficient when there are a small number of large families (Yi and Xu 1999). In addition, the fixed model approach is computationally very efficient because of the simplicity of the method. Unfortunately, as the number of families increases, the fixed model approach becomes inefficient and ill conditioned because of the large number of parameters to be estimated and tested. The random model approach, on the other hand, estimates and tests only a few parameters, *i.e.*, a few variance components, and thus is the choice for multiple-family QTL mapping. Such a method, however, has not been available for QTL mapping in binary traits.

The purpose of this research is to develop such a random model approach of QTL mapping for complex binary traits from multiple families of outbred populations. The method is developed on the basis of a generalized linear mixed model (GLMM) or a hierarchical generalized linear model (HGLM) where we treat the effects of QTL and the polygenic effect of the liability as random effects. An EM-algorithm with the Fisher-scoring algorithm embedded in each E-step is adopted to estimate the genetic variances. A simple Monte Carlo integration technique is used to calculate the likelihood-ratio test statistic. Application of the method is illustrated using a set of simulated data.

STATISTICAL METHODS

The threshold model and liability: Let s_{ij} and y_{ij} ($i = 1, \ldots, n$; $j = 1, \ldots, n_i$) be the binary phenotype and the underlying liability, respectively, of the jth individual in the ith full-sib family. The threshold model assumes that there is a fixed threshold in the scale of liability, t, which determines the binary phenotype of an individual by comparing y_{ij} with t. When $y_{ij} > t$, $s_{ij} = 1$, and otherwise, $s_{ij} = 0$. The liability y_{ij} can be treated as a continuous quantitative character and is thus described by the linear model

$$y_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta} + f_i + \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\eta}_i + \varepsilon_{ij}, \qquad (1)$$

where $\boldsymbol{\beta}$ is a vector of fixed effects (including the overall mean), which relates y_{ij} via a known incidence vector $\mathbf{x}_{ij}$; f_i is a family-specific effect, ε_{ij} is the residual effect (including the environmental error) distributed as $N(0, \sigma_\varepsilon^2)$; $\boldsymbol{\eta}_i = (\alpha^{\mathrm{s}}_{i1}, \alpha^{\mathrm{s}}_{i2}, \alpha^{\mathrm{d}}_{i1}, \alpha^{\mathrm{d}}_{i2}, \delta_{i11}, \delta_{i12}, \delta_{i21}, \delta_{i22})^{\mathrm{T}}$ is a vector of the effects of the alleles and the dominance effects of a putative QTL; α^{s}_{ik} ($k = 1, 2$) is the effect of the kth allele in the male parent; α^{d}_{il} ($l = 1, 2$) is the effect of the lth allele in the female parent; δ_{ikl} is the effect of interaction between the kth allele of the male parent and the lth allele of the female parent (dominance deviation); and

$$\mathbf{z}_{ij} = (z^{\mathrm{s}}_{ij1}, 1 - z^{\mathrm{s}}_{ij1}, z^{\mathrm{d}}_{ij1}, 1 - z^{\mathrm{d}}_{ij1}, z^{\mathrm{s}}_{ij1}z^{\mathrm{d}}_{ij1}, z^{\mathrm{s}}_{ij1}(1- z^{\mathrm{d}}_{ij1}), (1 - z^{\mathrm{s}}_{ij1})z^{\mathrm{d}}_{ij1}, (1 - z^{\mathrm{s}}_{ij1})(1 - z^{\mathrm{d}}_{ij1}))^{\mathrm{T}}$$

is a vector of the indicators and defined as $z^{\mathrm{s}}_{ij1} = 1$ if the first allele of the male (female) parent is transmitted to the jth progeny $z^{\mathrm{s}}_{ij1} = 0$ and otherwise. Indicator z^{d}_{ij1} is similarly defined.

We now treat $\boldsymbol{\gamma}_i = (f_i, \boldsymbol{\eta}_i^{\mathrm{T}})^{\mathrm{T}}$ as random effects with a multivariate normal distribution, $\boldsymbol{\gamma}_i \sim N_9(\mathbf{0}, \mathbf{Q})$. Under the assumption of unrelated parents,

$$\mathbf{Q} = \mathrm{diag}(\sigma_f^2, \sigma_\alpha^2, \sigma_\alpha^2, \sigma_\alpha^2, \sigma_\alpha^2, \sigma_\delta^2, \sigma_\delta^2, \sigma_\delta^2, \sigma_\delta^2,),$$

where

$$\sigma_f^2 = \mathrm{Var}(f_i),\ \sigma_\alpha^2 = \mathrm{Var}(\alpha^{\mathrm{s}}_{i1}) = \mathrm{Var}(\alpha^{\mathrm{s}}_{i2}) = \mathrm{Var}(\alpha^{\mathrm{d}}_{i1}) = \mathrm{Var}(\alpha^{\mathrm{d}}_{i2})$$

and

$$\sigma_\delta^2 = \mathrm{Var}(\delta_{i11}) = \mathrm{Var}(\delta_{i12}) = \mathrm{Var}(\delta_{i21}) = \mathrm{Var}(\delta_{i22}).$$

Note that $\boldsymbol{\gamma}_i$ and ε_{ij} are assumed to be mutually independent. Under model (1), the additive and dominance variances of the QTL are defined as

$$V_{\mathrm{a}} = \tfrac{1}{2}[\mathrm{Var}(\alpha^{\mathrm{s}}_{i1}) + \mathrm{Var}(\alpha^{\mathrm{s}}_{i2}) + \mathrm{Var}(\alpha^{\mathrm{d}}_{i1}) + \mathrm{Var}(\alpha^{\mathrm{d}}_{i2})] = 2\sigma_\alpha^2 \quad \text{and} \quad V_{\mathrm{d}} = \sigma_\sigma^2,$$

respectively. The variance of the family-specific effect is $\sigma_f^2 = \tfrac{1}{2}V_{\mathrm{A}} + \tfrac{1}{4}V_{\mathrm{D}} + V_{\mathrm{C}}$, where V_{A} and V_{D} are the polygenic additive and dominance variances (see Table 1), respectively, and V_{C} is the variance of the common environmental effect shared by family members. The residual variance is $\sigma_\varepsilon^2 = \tfrac{1}{2}V_{\mathrm{A}} + \tfrac{3}{4}V_{\mathrm{D}} + \sigma^2_{\mathrm{error}} = \tfrac{1}{2}V_{\mathrm{A}} + \tfrac{3}{4}V_{\mathrm{D}} + 1$ because the error variance is set to unity. The threshold model is overparameterized so that further constraints must be superimposed. As usual, the threshold model

TABLE 1

Description of symbols used in the text domin

Symbol	Description
s_{ij}, y_{ij}	Binary phenotype, liability
β	Vector of fixed effects
f_i	Family-specific effect
η_i	Vector of the effects of the alleles and the dominance effects of a putative QTL
γ_i	Vector of the family-specific effect and the effects of the alleles and the dominance effects of a putative QTL, $\gamma_i = (f_i, \eta_i^T)^T$
$\mathbf{Q}$	Covariance matrix of γ_i
α_{ik}^s, α_{il}^d, δ_{ikl}	Allelic effects of male and female parents, dominance effect
σ_f^2, σ_α^2, σ_δ^2	Variances of f_i, $\alpha_{ik}^s(\alpha_{il}^d)$, and δ_{ikl}
V_A, V_D	Polygenic additive, dominance variances
V_a, V_d	QTL additive, dominance variances, $V_a = 2\sigma_\alpha^2$, $V_d = \sigma_\delta^2$
h_p^2, h_q^2	Polygenic heritability, QTL heritability

is further standardized by setting $\sigma_\varepsilon^2 = 1$ and $t = 0$ (Harville and Mee 1984; McCulloch 1994; Sorensen *et al.* 1995). Under the "standardized threshold model," the vectors of fixed and random effects, β and γ_i, correspond to $\sigma_\varepsilon^{-1}\beta$ and $\sigma_\varepsilon^{-1}\gamma_i$, respectively (Harville and Mee 1984). In subsequent discussion, we use the standardized threshold model.

If the putative QTL is not at a marker locus or even if the QTL is at the marker locus but the marker is not fully informative, the genotype of the QTL is unobservable so that $\mathbf{z}_{ij}$ are missing. We can only infer the distribution of $\mathbf{z}_{ij}$ from observed genotypes of linked markers. Define the probabilities of $z_{ij1}^s = 1$ and $z_{ij1}^d = 1$ conditional on marker information by $p_{ij1}^s = \Pr(z_{ij1}^s = 1|I_M)$ and $p_{ij1}^d = \Pr(z_{ij1}^d = 1|I_M)$, respectively. Let $E(\mathbf{z}_{ij}|I_M)$ be the conditional expectation of $\mathbf{z}_{ij}$ given marker information ($\boldsymbol{I}_M$). The linear model (1) can be approximated by the following heterogeneous residual variance model:

$$y_{ij} = \mathbf{x}_{ij}^T\beta + f_i + E(\mathbf{z}_{ij}|I_M)^T\eta_i + e_{ij}. \quad (2)$$

The residual variance is

$$\begin{aligned}\mathrm{Var}(e_{ij}) &= \mathrm{Var}(y_{ij}|I_M, \beta, f_i, \eta_i)\\ &= \eta_i^T\mathrm{Var}(\mathbf{z}_{ij}|I_M)\eta_i + \sigma_\varepsilon^2 = V_{ij},\end{aligned}$$

where $\sigma_\varepsilon^2 = 1$ (as chosen in the standardized threshold model) and $\eta_i^T\mathrm{Var}(\mathbf{z}_{ij})\eta_i$ is the variance not explained due to the uncertainty of the QTL genotype (Xu 1996b, 1998). Given conditional probabilities p_{ij1}^s and p_{ij1}^d, the expectation and variance matrices of $\boldsymbol{x}_{ij}$ are

$$\begin{aligned}E(\mathbf{z}_{ij}|I_M) = (&p_{ij1}^s, 1 - p_{ij1}^s, p_{ij1}^d, 1 - p_{ij1}^d,\\ &p_{ij1}^s p_{ij1}^d, p_{ij1}^s(1 - p_{ij1}^d), (1 - p_{ij1}^s)p_{ij1}^d,\\ &(1 - p_{ij1}^s)(1 - p_{ij1}^d))^T\end{aligned}$$

and

$$\mathrm{Var}(\mathbf{z}_{ij}) = \begin{pmatrix}\mathrm{Var}(\mathbf{z}_{ij}^\alpha) & \mathrm{Cov}(\mathbf{z}_{ij}^\alpha, \mathbf{z}_{ij}^\delta)\\ \mathrm{Cov}(\mathbf{z}_{ij}^\alpha, \mathbf{z}_{ij}^\delta) & \mathrm{Var}(\mathbf{z}_{ij}^\delta)\end{pmatrix},$$

where $\mathbf{z}_{ij}^\alpha = (z_{ij1}^s, 1 - z_{ij1}^s, z_{ij1}^d, 1 - z_{ij1}^d)^T$, $\mathbf{z}_{ij}^\delta = (z_{ij1}^s z_{ij1}^d, z_{ij1}^s(1 - z_{ij1}^d), (1 - z_{ij1}^s)z_{ij1}^d, (1 - z_{ij1}^s)(1 - z_{ij1}^d))^T$,

$$\mathrm{Var}(\mathbf{z}_{ij}^\alpha) = \begin{pmatrix} p_{ij1}^s(1 - p_{ij1}^s) & -p_{ij1}^s(1 - p_{ij1}^s)\\ -p_{ij1}^s(1 - p_{ij1}^s) & p_{ij1}^s(1 - p_{ij1}^s)\\ 0 & 0\\ 0 & 0\\ 0 & 0\\ 0 & 0\\ p_{ij1}^d(1 - p_{ij1}^d) & -p_{ij1}^d(1 - p_{ij1}^d)\\ -p_{ij1}^d(1 - p_{ij1}^d) & p_{ij1}^d(1 - p_{ij1}^d)\end{pmatrix},$$

$$\mathrm{Cov}(\mathbf{z}_{ij}^\alpha, \mathbf{z}_{ij}^\delta) = \begin{pmatrix} p_{ij1}^s(1 - p_{ij1}^s)p_{ij1}^d & p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ -p_{ij1}^s(1 - p_{ij1}^s)p_{ij1}^d & -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ p_{ij1}^d(1 - p_{ij1}^d)p_{ij1}^s & -p_{ij1}^d(1 - p_{ij1}^d)p_{ij1}^s\\ -p_{ij1}^d(1 - p_{ij1}^d)p_{ij1}^s & p_{ij1}^d(1 - p_{ij1}^d)p_{ij1}^s\\ -p_{ij1}^s(1 - p_{ij1}^s)p_{ij1}^d & -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ p_{ij1}^s(1 - p_{ij1}^s)p_{ij1}^d & p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s) & -p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s)\\ -p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s) & p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s)\end{pmatrix}$$

$$\mathrm{Var}(\mathbf{z}_{ij}^\delta) = \begin{pmatrix} p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s p_{ij1}^d) & -(p_{ij1}^s)^2 p_{ij1}^d(1 - p_{ij1}^d)\\ -(p_{ij1}^s)^2 p_{ij1}^d(1 - p_{ij1}^d) & p_{ij1}^s(1 - p_{ij1}^d)[1 - p_{ij1}^s(1 - p_{ij1}^d)]\\ -(p_{ij1}^d)^2 p_{ij1}^s(1 - p_{ij1}^s) & -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d) & -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)^2\\ -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)^2 & -(p_{ij1}^d)^2 p_{ij1}^s(1 - p_{ij1}^s)\\ -(p_{ij1}^d)^2 p_{ij1}^s(1 - p_{ij1}^s) & -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d) & p_{ij1}^d(1 - p_{ij1}^s)[1 - p_{ij1}^d(1 - p_{ij1}^s)]\\ -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)^2 & -p_{ij1}^d(1 - p_{ij1}^d)\\ -p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s)^2 & -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d)\\ -p_{ij1}^s p_{ij1}^d(1 - p_{ij1}^s)(1 - p_{ij1}^d) & -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)^2\\ -p_{ij1}^s(1 - p_{ij1}^s)(1 - p_{ij1}^d)^2 & -p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s)^2\\ (1 - p_{ij1}^s)(1 - p_{ij1}^d)[1 - (1 - p_{ij1}^s)(1 - p_{ij1}^d)] & -p_{ij1}^d(1 - p_{ij1}^d)(1 - p_{ij1}^s)^2\end{pmatrix}.$$

The conditional probabilities, p_{ij1}^s and p_{ij1}^d, are calculated using the simplified multipoint method proposed in four-way crosses (Rao and Xu 1998). The method requires known marker linkage phases in the parents. When the linkage phases are not known, they must be inferred first from marker genotypes of the parents and

the offspring. If grandparents are also genotyped, the linkage phases can be accurately reconstructed; otherwise, a relatively large number of offspring for each family is required (Knott *et al.* 1996). When family sizes are too small to provide reasonable accuracy of linkage phase inference, alternative models, *e.g.*, the identical-by-descent (IBD)-based random model approach, should be considered, which are discussed later.

Maximum-likelihood estimation of genetic variances: A maximum-likelihood (ML) method is proposed for estimation of the variance components. We first obtain the sample density, $g(s_{ij}|\gamma_i, \beta)$, using the probit relationship between the binary phenotype and the model effects

$$g(s_{ij}|\gamma_i,\beta) = [\Pr(y_{ij} > 0|\gamma_i,\beta)]^{s_{ij}}[1 - \Pr(y_{ij} > 0|\gamma_i,\beta)]^{1-s_{ij}} = \left[\Phi\left(\frac{\mu_{ij}}{\sqrt{V_{ij}}}\right)\right]^{s_{ij}}\left[1 - \Phi\left(\frac{\mu_{ij}}{\sqrt{V_{ij}}}\right)\right]^{1-s_{ij}},$$

where $\Phi(\cdot)$ is the standardized normal distribution function and $\mu_{ij} = \mathbf{x}_{ij}^{\mathrm{T}}\beta + f_i + E(\mathbf{z}_{ij}|I_{\mathrm{M}})^{\mathrm{T}}\eta_i$. For notational simplicity, we denote $P_{ij} = \Pr(y_{ij} > 0|\gamma_i,\beta) = \Phi(\mu_{ij}/\sqrt{V_{ij}})$ in subsequent discussion. Conditional on $\gamma_i = (f_i, \eta_i^{\mathrm{T}})^{\mathrm{T}}$, individuals within the same family are independent. Hence, the joint density for the ith family is

$$\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i,\beta) = \prod_{j=1}^{n_i} P_{ij}^{s_{ij}}(1 - P_{ij})^{1-s_{ij}}.$$

For n independent families, the overall joint density is

$$g(\mathbf{S}|\gamma,\beta) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i,\beta) = \prod_{i=1}^{n}\prod_{j=1}^{n_i} P_{ij}^{s_{ij}}(1 - P_{ij})^{1-s_{ij}},$$

where $\gamma = (\gamma_1^{\mathrm{T}}\ \gamma_2^{\mathrm{T}} \cdots \gamma_n^{\mathrm{T}})^{\mathrm{T}}$ and $\mathbf{S} = (s_{11}\, s_{12} \cdots s_{nn_n})^{\mathrm{T}}$. The likelihood function is obtained by integrating out the random effects γ and thus becomes a function of the data $\mathbf{S}$ and the parameters $\xi = (\beta^{\mathrm{T}}\ \sigma_f^2\ \sigma_\alpha^2\ \sigma_\delta^2)^{\mathrm{T}}$. The log-likelihood function has the form of

$$l(\xi|\mathbf{S}) = \log\int g(\mathbf{S}|\gamma,\beta)\pi(\gamma|\mathbf{Q})d\gamma, \tag{3}$$

where $\mathbf{Q} = \mathrm{diag}(\sigma_f^2\ \sigma_\alpha^2\ \sigma_\alpha^2\ \sigma_\alpha^2\ \sigma_\alpha^2\ \sigma_\delta^2\ \sigma_\delta^2\ \sigma_\delta^2\ \sigma_\delta^2)$, as previously defined,

$$\pi(\gamma|\mathbf{Q}) = \prod_{i=1}^{n}\pi(\gamma_i|\mathbf{Q}),$$

and

$$\pi(\gamma_i|\mathbf{Q}) = (2\pi)^{-2}|\mathbf{Q}|^{-1/2}\exp\{-1/2\gamma_i^{\mathrm{T}}\mathbf{Q}^{-1}\gamma_i\}.$$

Other types of distribution, rather than normal, may be specified for γ_i. The maximum-likelihood estimation (MLE) of ξ is obtained by maximizing the marginal log likelihood $l(\xi|\mathbf{S})$ (Searle *et al.* 1992; Fahrmeir and Tutz 1994).

Direct maximization of (3) is cumbersome due to the high-dimensional integral structure. Instead, we may use an EM-algorithm to maximize (3) indirectly. Define the joint density of the complete data by

$$f(\mathbf{S},\gamma|\beta,\mathbf{Q}) = g(\mathbf{S}|\gamma,\beta)\pi(\gamma|\mathbf{Q}).$$

The EM-algorithm starts from the joint log-likelihood function of the complete data

$$L = \log f(\mathbf{S},\gamma|\beta,\mathbf{Q}) = \sum_{i=1}^{n}\sum_{j=1}^{n_i}[s_{ij}\log P_{ij} + (1 - s_{ij})\log(1 - P_{ij})] + \sum_{i=1}^{n}\log\pi(\gamma_i|\mathbf{Q}). \tag{4}$$

In the $(m + 1)$th cycle of iteration, the E-step consists of computing

$$M(\mathbf{Q}|\mathbf{Q}^{(m)}) = E\{\log(f(\mathbf{S},\gamma|\beta,\mathbf{Q})|\mathbf{S},\mathbf{Q}^{(m)})\},$$

the conditional expectation of the log-likelihood of the complete data given the estimated parameters $\mathbf{Q}^{(m)}$ from the previous cycle of iteration. In the M-step $M(\mathbf{Q}|\mathbf{Q}^{(m)})$ has to be maximized with respect to $\mathbf{Q}$. Differentiation with respect to σ_f^2, σ_α^2, and σ_δ^2, setting derivatives to zero, and solving for σ_f^2, σ_α^2, and σ_δ^2 yield the following updates:

$$\sigma_f^{2(m+1)} = \frac{1}{n}\sum_{i=1}^{n}E(f_i^2|\mathbf{S},\mathbf{Q}^{(m)}) = \frac{1}{n}\sum_{i=1}^{n}[\mathrm{Var}(f_i|\mathbf{S},\mathbf{Q}^{m}) + E^2(f_i|\mathbf{S},\mathbf{Q}^{(m)})],$$

$$\sigma_\alpha^{2(m+1)} = \frac{1}{4n}\sum_{k=1}^{2}\sum_{i=1}^{n}[E((\alpha_{ik}^{\mathrm{s}})^2|\mathbf{S},\mathbf{Q}^{(m)}) + E((\alpha_{ik}^{\mathrm{d}})^2|\mathbf{S},\mathbf{Q}^{(m)})] = \frac{1}{4n}\sum_{k=1}^{2}\sum_{i=1}^{n}[\mathrm{Var}(\alpha_{ik}^{\mathrm{s}}|\mathbf{S},\mathbf{Q}^{(m)}) + E^2(\alpha_{ik}^{\mathrm{s}}|\mathbf{S},\mathbf{Q}^{(m)}) + \mathrm{Var}(\alpha_{ik}^{\mathrm{d}}|\mathbf{S},\mathbf{Q}^{(m)}) + E^2(\alpha_{ik}^{\mathrm{d}}|\mathbf{S},\mathbf{Q}^{(m)})],$$

$$\sigma_\delta^{2(m+1)} = \frac{1}{4n}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{i=1}^{n}E((\delta_{ikl})^2|\mathbf{S},\mathbf{Q}^{(m)}) = \frac{1}{4n}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{i=1}^{n}[\mathrm{Var}(\delta_{ikl}|\mathbf{S},\mathbf{Q}^{(m)}) + E^2(\delta_{ikl}|\mathbf{S},\mathbf{Q}^{(m)})].$$

Although the algorithm is conceptually straightforward, it is difficult to carry out the updates exactly because the posterior means

$$\{E(f_i|\mathbf{S},\mathbf{Q}^{(m)}),\ E(\alpha_{ik}^{\mathrm{s}}|\mathbf{S},\mathbf{Q}^{(m)}),\ E(\alpha_{il}^{\mathrm{d}}|\mathbf{S},\mathbf{Q}^{(m)}),\ E(\delta_{ikl}|\mathbf{S},\mathbf{Q}^{(m)})\}$$

and the posterior variances

$$\{\mathrm{Var}(f_i|\mathbf{S},\mathbf{Q}^{(m)}),\ \mathrm{Var}(\alpha_{ik}^{\mathrm{s}}|\mathbf{S},\mathbf{Q}^{(m)}),\ \mathrm{Var}(\alpha_{il}^{\mathrm{d}}|\mathbf{S},\mathbf{Q}^{(m)}),\ \mathrm{Var}(\delta_{ikl}|\mathbf{S},\mathbf{Q}^{(m)})\}$$

do not have explicit expressions. In principle, the posterior means and variances can be obtained via numerical integration or a Monte Carlo method (*e.g.*, Fahrmeir and Tutz 1994; McCulloch 1994; Chan and Kuk 1997). However, numerical integration and the sampling-based methods can be extremely time consuming

and thus are not feasible for whole-genome scanning for QTL. To ease the computational burden, analytic approximations of the posterior means and the posterior variances are needed. In this study, the posterior means are approximated by the posterior modes $\hat{f}_i$, $\hat{\alpha}^{s}_{ik}$, $\hat{\alpha}^{d}_{il}$, and $\hat{\delta}_{ikl}$ ($i = 1, \cdots, n$; $k, l = 1, 2$) and the posterior variances by the posterior curvatures $\hat{V}^{f}_{i}$, $\hat{V}^{s}_{ik}$, $\hat{V}^{d}_{il}$, and $\hat{V}^{\delta}_{ikl}$ ($i = 1, \cdots, n$; $k, l = 1, 2$). Both the posterior modes and posterior curvatures are obtained by using a Fisher-scoring algorithm, which is described in the appendix.

The resulting EM-algorithm with the Fisher-scoring algorithm embedded in each E-step jointly estimates β, γ, σ^2_f, σ^2_α, and σ^2_δ as follows:

1. Choose starting values $\sigma^{2(0)}_f$, $\sigma^{2(0)}_\alpha$, and $\sigma^{2(0)}_\delta$.
2. Compute posterior mode estimates $\hat{f}_i$, $\hat{\alpha}^{s}_{ik}$, $\hat{\alpha}^{d}_{il}$, and $\hat{\delta}_{ikl}$ ($i = 1, \cdots, n$; $k, l = 1, 2$) and posterior curvature estimates $\hat{V}^{f}_{i}$, $\hat{V}^{s}_{ik}$, $\hat{V}^{d}_{il}$ and $\hat{V}^{\delta}_{ikl}$ ($i = 1, \cdots, n$; $k, l = 1, 2$) via the Fisher-scoring algorithm with variance components replaced by their current estimates $\sigma^{2(m)}_f$, $\sigma^{2(m)}_\alpha$, and $\sigma^{2(m)}_\delta$. The estimated fixed effect $\hat{\beta}$, is also obtained in this step.
3. EM-step: Compute $\sigma^{2(m+1)}_f$, $\sigma^{2(m+1)}_\alpha$, and $\sigma^{2(m+1)}_\delta$ by

$$\sigma^{2(m+1)}_f = \frac{1}{n}\sum_{i=1}^{n}[\hat{V}^{f(m)}_{i} + (\hat{f}^{(m)}_{i})^2],$$

$$\sigma^{2(m+1)}_\alpha = \frac{1}{4n}\sum_{k=1}^{2}\sum_{i=1}^{n}[\hat{V}^{s(m)}_{ik} + (\hat{\alpha}^{s(m)}_{ik})^2 + \hat{V}^{d(m)}_{ik} + (\hat{\alpha}^{d(m)}_{ik})^2],$$

$$\sigma^{2(m+1)}_\delta = \frac{1}{4n}\sum_{k=1}^{2}\sum_{l=1}^{2}\sum_{i=1}^{n}[\hat{V}^{\delta(m)}_{ikl} + (\hat{\delta}^{(m)}_{ikl})^2].$$

4. If convergence is reached, set $\hat{\sigma}^2_f = \sigma^{2(m+1)}_f$, $\hat{\sigma}^2_\alpha = \sigma^{2(m+1)}_\alpha$, and $\hat{\sigma}^2_\delta = \sigma^{2(m+1)}_\delta$; otherwise, increase m by 1 and return to step 1.

Note that under the standardized threshold model, σ^2_f, σ^2_α, and σ^2_δ are actually ratios of variance components. Recall that the value of the residual variance has been set to unity, which leads the above variances to be

$$\sigma^2_f = (\sigma^2_\varepsilon)^{-1}(\tfrac{1}{2}V_A + \tfrac{1}{4}V_D + V_C),$$

where $\sigma^2_\varepsilon = \tfrac{1}{2}V_A + \tfrac{3}{4}V_D + 1$ as given earlier. Let us assume that $V_D = V_C = 0$ so that we have the relationships

$$\sigma^2_f = \frac{\tfrac{1}{2}V_A}{\tfrac{1}{2}V_A + 1}, \quad \sigma^2_\alpha = \frac{\tfrac{1}{2}V_a}{\tfrac{1}{2}V_A + 1}, \quad \sigma^2_\delta = \frac{V_d}{\tfrac{1}{2}V_A + 1},$$

where V_A, V_a, and V_d are additive variance of the polygene and additive and dominance variances of the QTL, respectively. These genetic variances are obtained by using $\hat{\sigma}^2_f$, $\hat{\sigma}^2_\alpha$, and $\hat{\sigma}^2_\delta$ and solving for the above equations:

$$\hat{V}_A = \frac{2\hat{\sigma}^2_f}{1 - \hat{\sigma}^2_f}, \quad \hat{V}_a = 2\left(\frac{\hat{\sigma}^2_f}{1 - \hat{\sigma}^2_f} + 1\right)\hat{\sigma}^2_\alpha,$$

$$\hat{V}_d = \left(\frac{\hat{\sigma}^2_f}{1 - \hat{\sigma}^2_f} + 1\right)\hat{\sigma}^2_\delta. \quad (5)$$

The estimated proportions of variances explained by the polygene and the QTL are

$$\hat{h}^2_p = \frac{\hat{V}_A}{\hat{V}_A + \hat{V}_a + V_d + 1} \quad \text{and} \quad \hat{h}^2_q = \frac{\hat{V}_a}{\hat{V}_A + \hat{V}_a + V_d + 1},$$

respectively.

Tests of hypotheses: Although the additive and dominance effects can be tested individually, we investigate only the overall test of the presence of QTL. The null hypothesis is H_0: $\sigma^2_\alpha = \sigma^2_\delta = 0$. The alternative hypothesis is H_1: $\sigma^2_\alpha \neq 0$ or $\sigma^2_\delta \neq 0$. The likelihood-ratio test statistic is used to test the presence of QTL, which is defined as

$$\Lambda = -2(L_0 - L_1), \quad (6)$$

where L_1 is the log likelihood evaluated under the alternative hypothesis (full model) and L_0 is that evaluated under the null hypothesis (reduced model).

Although the EM-algorithm with the Fisher-scoring embedded in each E-step has eliminated the necessity of numerical integration and provides a convenient way for estimation of variance components, it does not automatically generate the numerical value of the likelihood. To calculate the likelihood value, one still needs numerical integration. Fortunately, it is manageable in this stage because the parameters are replaced by their MLEs without updating. The Monte Carlo numerical integration is applied here due to the high dimensionality of the random effects. The log-likelihood value under the alternative model evaluated at the MLE is

$$L_1 = \sum_{i=1}^{n}\log\int\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i, \hat{\beta})\pi(\gamma_i|\hat{\mathbf{Q}})d\gamma_i.$$

The Monte Carlo approximation of $\int\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i, \hat{\beta})\pi(\gamma_i|\hat{\mathbf{Q}})d\gamma_i$ is obtained by

$$\int\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i, \hat{\beta})\pi(\gamma_i|\hat{\mathbf{Q}})d\gamma_i \approx \frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i^{(k)}, \hat{\beta}),$$

where $\gamma_i^{(k)}$ for $k = 1, \cdots, M$ is a random sample simulated from distribution $\pi(\gamma_i|\hat{\mathbf{Q}})$, and M is the length of the Monte Carlo simulation. The Monte Carlo approximation becomes sufficiently accurate when M is very large. The empirical value of $M = 5000$ (Xu and Atchley 1996b) is adopted in this study.

Under the null hypothesis, model (2) has been reduced to

$$y_{ij} = \mathbf{x}^T_{ij}\beta + f_i + \varepsilon_{ij},$$

where $f_i \sim N(0, \sigma^2_f)$ and $\varepsilon_{ij} \sim N(0, 1)$. The log-likelihood value under the reduced model evaluated at the MLE is

$$L_0 = \sum_{i=1}^{n} \log \int \prod_{j=1}^{n_i} g(s_{ij}|f_i, \hat{\beta})\pi(f_i|\hat{\sigma}_f^2) df_i,$$

where $\hat{\beta}$ and $\hat{\sigma}_f^2$ are the MLE under H_0.

There is a numerical problem with the Monte Carlo approximation; that is, when the family size becomes large, $\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i^{(k)}, \hat{\beta})$ tends to be close to zero, causing computational underflows. This can be circumvented by first calculating

$$C(i,k) = \sum_{j=1}^{n_i} \log g(s_{ij}|\gamma_i^{(k)}, \hat{\beta}) = \frac{1}{M} \sum_{k'=1}^{M} \sum_{j=1}^{n_i} \log g(s_{ij}|\gamma_i^{(k')}, \hat{\beta})$$

and then taking

$$\log\left[\frac{1}{M}\sum_{k=1}^{M}\prod_{j=1}^{n_i} g(s_{ij}|\gamma_i^{(k)}, \hat{\beta})\right] = \log\left[\frac{1}{M}\sum_{k=1}^{M} e^{C(i,k)} + \frac{1}{M}\sum_{k=1}^{M}\sum_{j=1}^{n_i}\log g(s_{ij}|\gamma_i^{(k)}, \hat{\beta})\right].$$

Therefore, the Monte Carlo-evaluated log-likelihood function under the alternative hypothesis is

$$L_1 = \sum_{i=1}^{n} \log\left[\frac{1}{M}\sum_{k=1}^{M} e^{C(i,k)} + \frac{1}{M}\sum_{k=1}^{M}\sum_{n=1}^{n_i}\log g(s_{ij}|\gamma_i^{(k)}, \hat{\beta})\right].$$

The Monte Carlo log-likelihood function under the null hypothesis L_0 is similarly calculated.

SIMULATION STUDIES

Experimental design: Application of the proposed method is illustrated via Monte Carlo simulation experiments. The following properties were examined: the bias, the standard error of parameter estimation, and the statistical power of QTL detection. We simulated a single chromosome segment of length 80 cM with nine evenly spaced codominant markers. In most cases, four equally frequent alleles were simulated at each marker locus. A single QTL residing at position 25 cM and 12 additional independent loci of equal effect (called the polygene) were simulated for the liability. The dominance effect of the polygene was assumed to be absent. Both the allelic effects and the allelic interaction effects (dominance) of the QTL were assumed to be normally distributed and so was the polygenic effect. Value of the liability of each individual took the sum of an overall mean, values of QTL additive and dominance effects, polygenic effect, and a residual error sampled from a standardized normal distribution. The observable binary phenotype was set to be 1 if corresponding liability exceeded 0, and 0 otherwise. The fixed effect contained the population mean of the liability only, which led to a proportion of the trait presence (incidence) of 40%.

To examine the effect of different factors on the performance of the method, we varied each of the following factors successively: (1) the proportion of variance explained by the QTL:

$h_q^2 = 0.10$ ($V_A = 0.80$, $V_a = 0.10$, $V_d = 0.10$),

$h_q^2 = 0.25$ ($V_A = 0.50$, $V_a = 0.25$, $V_d = 0.25$),

$h_q^2 = 0.40$ ($V_A = 0.20$, $V_a = 0.40$, $V_d = 0.40$);

(2) sampling strategy (number of families × family size): 2 × 250, 5 × 100, 10 × 50, 15 × 33, 20 × 25, and 50 × 10; (3) marker polymorphism: two, four, and eight equally frequent alleles at each marker.

Instead of performing simulations under all possible combinations of parametric settings, we simulated a situation in which the central level is chosen for each factor considered above. We referred to this as the "standard setting," which is described as follows: $h_q^2 = 0.25$, 10 families each with 50 sibs and four alleles per marker. When the influence of different levels of a factor on the performance of the method is examined, all other factors are set to the standard levels.

The simulations were repeated 100 times for each parametric setting. The standard error of the parameter estimate was calculated from the standard deviation of the estimates among the 100 replicates. To estimate the statistical power, we ran 1000 additional replicates with no QTL segregating while the other two factors were set at their standard levels. We augmented the polygenic variance such that the total genetic variance of the liability remained unchanged. The statistical power was determined by counting the (proportion) number of runs (over the 100 replicates) that have test statistics greater than a chromosome-wise empirical critical value. The empirical critical value was obtained by choosing the 95th and 99th percentiles of the highest test statistic over the 1000 runs under the null model (no QTL segregating).

Under the standard setting, results of the proposed GLMM approach were then compared with those of the simple LMM analysis, where the binary data were treated as if they were continuous. We modified the random model method of Xu (1998) so that it can be applied to the populations that were used in this study.

Results: The empirical critical values at type I error rates of 0.05 and 0.01 obtained from 1000 replicated simulations were 7.87 and 10.73, respectively. The average likelihood-ratio test statistics and the power estimates over 100 replicated simulations are summarized in Table 2. As expected, the average test statistic increases as the QTL heritability increases. The statistical power also shows the same trend. The sampling strategy also has an effect on the test statistic and the power. When the total number of individuals is fixed, the average test statistic increases as the family size increases. However, when the number of families is too small, *e.g.*, two, the average test statistic tends to decrease, primarily due to sampling error of the parents. There is an optimal sampling strategy, any deviation from which will cause a decrease in the test statistic and the power. Marker polymorphism also plays a role in the test statistic and

TABLE 2

Average test statistics and empirical powers of QTL detection obtained from 100 replicated simulations

	Test statistic	Power %	
		$\alpha = 0.05$	$\alpha = 0.01$
QTL heritability (h_q^2)			
0.10	8.63 (6.16)	55	40
0.25	25.13 (13.97)	92	86
0.40	45.77 (20.38)	100	99
Sampling strategy			
2×250	30.14 (18.34)	91	82
5×100	33.52 (17.33)	96	94
10×50	25.13 (13.97)	92	86
15×33	20.42 (10.98)	92	83
20×25	16.43 (8.18)	85	71
50×10	10.13 (5.09)	74	47
Number of alleles per marker			
Two	21.13 (10.09)	90	84
Four	25.13 (13.97)	92	86
Eight	26.66 (13.31)	96	92

The standard errors of test statistics are given in parentheses.

power. The average test statistic and the statistical power increase as marker polymorphism increases.

The properties of the method were investigated under three levels of proportion of variance explained by the QTL while the other factors were set at their standard levels (see Table 3). As expected, the estimate of position of the QTL is quite accurate when h_q^2 is not too low. A bias toward the center of the chromosome segment was observed when $h_q^2 = 0.10$. The estimation error of the QTL position was also large when h_q^2 is small. Both the polygenic variance and the variances of QTL effects were underestimated by the method. The underestimation can be severe when the true variances are high. There are two approximations in the derivation of the method, one being the mixture of four normal distributions of the residual of the liability approximated by a single normal distribution with heterogeneous variance and the other being the posterior means and variances replaced by the posterior modes and curvatures. These two approximations may explain the bias observed in the estimates of variance components. The bias caused by the first approximation can be reduced by increasing marker density and allelic polymorphism because the uncertainty of the QTL genotype inference can be reduced so that the single normal distribution adequately approximates the mixed distribution. The bias caused by the second approximation can be prevented by using a sampling-based Markov chain Monte Carlo (MCMC) algorithm, which is discussed in the next section. Because most QTL may be in the range of medium to small in size (variance) and the bias is slight when the true QTL heritability is in this range, the bias should not be of major concern for this method.

The sampling strategy was investigated with the other factors set at their standards (see Table 4). The estimate of the QTL position is unbiased when family size is not too small, but larger estimation errors are observed when either too few or too many families are used. The bias in the estimates of variance components also seems to follow the same trend—the least bias at the optimal

TABLE 3

Mean estimates of QTL parameters under different levels of QTL heritability

QTL heritability		Position	V_A	V_a	V_d	h_q^2
0.10	True	25	0.80	0.10	0.10	0.10
	Estimate	29.80 (16.50)	0.71 (0.60)	0.13 (0.12)	0.08 (0.10)	0.10 (0.05)
0.25	True	25	0.50	0.25	0.25	0.25
	Estimate	25.76 (7.38)	0.46 (0.55)	0.26 (0.10)	0.21 (0.24)	0.24 (0.08)
0.40	True	25	0.20	0.40	0.40	0.40
	Estimate	24.68 (3.27)	0.29 (0.45)	0.32 (0.27)	0.32 (0.24)	0.33 (0.09)

The standard errors of estimates obtained by the standard deviation of 100 replicated simulations are given in parentheses.

TABLE 4

Mean estimates of QTL parameters under different sampling strategies (number of families × family size)

Sampling strategy	Position	V_A	V_a	V_d	h_q^2
	25.00[a]	0.50	0.25	0.25	0.25
2 × 250	25.75 (9.27)	0.42 (0.60)	0.21 (0.24)	0.21 (0.26)	0.23 (0.09)
5 × 100	25.23 (5.03)	0.49 (0.53)	0.25 (0.23)	0.23 (0.19)	0.25 (0.09)
10 × 50	25.76 (7.38)	0.46 (0.50)	0.26 (0.22)	0.21 (0.24)	0.24 (0.08)
15 × 33	25.62 (10.08)	0.42 (0.43)	0.21 (0.17)	0.19 (0.15)	0.22 (0.08)
20 × 25	25.02 (7.34)	0.51 (0.41)	0.23 (0.19)	0.18 (0.13)	0.21 (0.07)
50 × 10	27.14 (15.86)	0.49 (0.36)	0.19 (0.13)	0.17 (0.10)	0.20 (0.06)

The standard errors of estimates obtained by the standard deviation of 100 replicated simulations are given in parentheses.

[a] Numbers in this row are the true parametric values.

sampling strategy (5 × 100). The estimation errors of the variance components, however, seem to decrease monotonically as the number of families increases.

Increasing the level of marker polymorphism can improve the estimate of QTL position but has limited effect on the estimation of the variance components (Table 5).

The empirical critical values at type I error rates of 0.05 and 0.01 obtained from 1000 replicated simulations were 8.01 and 14.61, respectively, when treating binary data as normally distributed. In the "standard setting," the power estimates over 100 replicated simulations were 83 and 60%, respectively, lower than 92 and 86%, which were observed under the threshold model. The estimates of the QTL position and its standard error were 25.52 and 11.75, respectively. The heritability estimate of the observed binary trait and its standard error were 0.16 and 0.078, respectively. We converted the heritability estimate of the observed binary trait into that of the liability (Lynch and Walsh 1998, p. 743) and obtained the liability heritability estimate 0.28 and it standard error 0.14. As expected, the proposed method has a higher statistical power and also produces more accurate estimates of the QTL position and the heritability than the method that directly analyzes the binary trait.

The purpose of the simulation experiments is not to exhaustively search for the best design of QTL mapping for binary traits but to illustrate that the proposed method behaves like the existing QTL mapping procedures developed for regular quantitative traits. The conclusion is that the method behaves as expected and works well in the situations examined.

DISCUSSION

Both the random model approach of QTL mapping for normally distributed traits and the fixed model approach of QTL mapping for binary traits are well developed (Fulker and Cardon 1994; Hackett and Weller 1995; Xu and Atchley 1995; Grignola *et al.* 1996; Visscher *et al.* 1996; Xu 1996a; Xu and Atchley 1996a; Rebai 1997; Rao and Xu 1998). Our contribution is to develop the random model approach of QTL mapping for binary data, which is a combination of the two existing approaches. Although neither approach is complicated enough by itself to prevent the use of an exact ML method, the combination becomes cumbersome enough that there does not exist an exact form of ML method due to the lack of analytically and computationally tractable mixing distributions. The method presented here illustrates the use of a GLMM for QTL mapping with some approximations (see simulation studies for the approximations). Results of Monte Carlo simulations show that the approximations are well

TABLE 5

Mean estimates of QTL parameters under different levels of marker polymorphism

Marker alleles	Position	V_A	V_a	V_d	h_q^2
	25.00[a]	0.50	0.25	0.25	0.25
Two	26.57 (8.38)	0.43 (0.40)	0.23 (0.22)	0.20 (0.16)	0.23 (0.07)
Four	25.76 (7.38)	0.46 (0.55)	0.26 (0.22)	0.21 (0.24)	0.24 (0.08)
Eight	25.20 (5.54)	0.58 (0.60)	0.24 (0.20)	0.22 (0.17)	0.23 (0.08)

The standard errors of estimates obtained by the standard deviation of 100 replicated simulations are given in parentheses.

[a] Numbers in this row are the true parametric values.

justified because both the location and the effects of the simulated QTL are well estimated by the approximate method. These minor assumptions have yielded a computationally attractive ML method. With this approximate ML, we can perform genome scanning of QTL for binary traits, just as we do for regular quantitative traits. To the best of our knowledge, this is the first attempt to map QTL for binary data using an approach that is consistent with classical quantitative genetic theory.

Approximations occurred twice in the derivation of the method, one being that the mixture of four normal distributions of the residual in the scale of liability is approximated by a single normal distribution and the other being that the posterior means and posterior variances are replaced by the posterior modes and curvatures. The first approximation can be easily relaxed; *i.e.*, we can directly use the mixed distribution rather than a single distribution, although the exact expression of the Fisher-scoring algorithm is not as clean as it is now. Analytically, the second approximation can also be relaxed, but the price is an unrealistic computational time. The posterior means and posterior variances can be calculated via a sampling-based approach, *e.g.*, the Gibbs sampler. The combination of the EM with the sampling-based approaches is called the Monte Carlo EM (Fahrmeir and Tutz 1994; McCulloch 1994; Chan and Kuk 1997). The Monte Carlo EM is feasible for evaluation of a single point of a genome but unrealistic for the whole genome scanning because many points need to be evaluated. At any single point, there are many EM-steps, each requiring many cycles of the Gibbs sampler (Gibbs chain). If the Gibbs chain is long, the time required in the total genome scanning can be unrealistic. If the chain is too short, on the other hand, the posterior means and posterior variances will deviate from the exact values. At this point, the approximate ML method is perhaps better suited than the Monte Carlo EM. Another exact method is the Bayesian approach of QTL mapping. Instead of performing whole-genome scanning, one can treat the positions of multiple QTL as unknown variables. The number of QTL can even be treated as an unknown variable and be searched simultaneously (Heath 1997; Uimari and Hoeschele 1997; Sillanpaa and Arjas 1998). The full Bayesian treatment must also be implemented via the sampling-based approach, which is of course time consuming. The full Bayesian treatment may show some advantages over the ML approach, but the price is the increased computational burden and a reduced intuitiveness of the genome scanning. Although the full Bayesian approach seems to be the direction of QTL mapping research in the future (Satagopan *et al.* 1996; Heath 1997; Uimari and Hoeschele 1997; Sillanpaa and Arjas 1998), it cannot replace the existing ML approach in practice.

Conventional QTL mapping procedures utilize a single family of line cross. Almost all effort is allocated to the sample of offspring within the single family to ensure that segregation of the QTL alleles in the parents is detectable. The method is heavily dependent on the parents sampled. If the parents do not segregate at a QTL, there is no power to detect it even though the QTL segregates in the population in which these two parents are sampled. One can combine information from multiple families to ensure that segregating parents are sampled (Muranty 1996; Xie *et al.* 1998; Xu 1998). There are two strategies for combining data from multiple families: the fixed model and the random model approaches. Which model should we choose? One should decide whether the parents chosen to form the mapping population are a *random sample* from a hypothetical large population (base population). If they are, one should consider the use of the random model approach, provided that one is interested in understanding the genetic properties of the base population. If the parents of the mapping population are not randomly sampled and one has no desire to understand the base population, but only the mapping population, then the fixed model is more appropriate. Under the random model approach, we are interested in making a statistical inference about the base population. Namely, the estimated QTL variances may reflect the actual genetic variation existing in the population where the experimental units are sampled.

Interestingly, we can use the fixed model approach to solve the random model problem. In this situation, we have a random model in our mind; *i.e.*, we are interested in inferring the statistic(s) drawn from the sample to the base population, but we may first estimate and test the first moment statistics (the effects) as if they were fixed effects and then calculate the variance of the effects by relaxing the fixed assumption. Statistically, this approach is identical to Henderson's method III for variance component estimation (Searle *et al.* 1992). Because the model is a random model, but the statistical method itself is not a real random model approach, we call it the pseudorandom model approach. This approach has been previously examined in QTL mapping using multiple families of line crosses (Xu 1998) and outbred populations (Knott *et al.* 1996). Yi and Xu (1999) recently developed the pseudorandom model approach of QTL mapping for binary data using multiple families. We showed that the pseudorandom model approach is quite efficient and computationally much faster than the true random model approach. There is, though, a complication in the interpretation of the test statistic in the pseudorandom model. Because the QTL effects are tested and the number of QTL effects increases as the number of families increases, the critical value of the test statistic used to declare statistical significance changes accordingly as the number of families changes. This is undesirable because the guidelines for significance declaration suggested by Lander and Kruglyak (1995) cannot be applied. The true random

model approach developed in this study does not have that drawback and is statistically more sound. It should complement, but not replace, the existing pseudorandom model approach.

The QTL mapping procedure presented in this study is based on known marker linkage phases in the parents. Therefore, the inference of marker linkage phases is a prerequisite of the method. There are several ways to deduce the parental haplotypes in outbred populations: (1) track alleles from the parental genotype through the segregating progeny population; (2) use grandparental genotypes; or (3) genotype the parents directly using PCR-based marker technology from parental gametes (Williams 1998). The accuracy of phase inference using method (1) largely depends on the family size. When the family size is small, inference of the parental linkage phases is subject to error, which will likely reduce the power of QTL detection. Therefore, the method presented is practical only for species with large family sizes, say $n_i \geq 20$ for $i = 1, \ldots, n$. To apply this method to QTL mapping in species with small family sizes, *e.g.*, human and some other mammals, we must modify the method by incorporating the IBD approach (Fulker and Cardon 1994). The IBD-based approach requires only the proportion of allelic sharing by two relatives (*e.g.*, sibs) and does not distinguish as to which alleles of the parents are being shared by the sibs. This eliminates the necessity of linkage phase information. Incorporation of the IBD-based approach into the current method is straightforward; one simply replaces the genetic (random) effects of the parents by the genetic effects of the progenies and solves for the genetic effects of the progenies in the Fisher-scoring step with $\mathbf{Q}$ replaced by

$$\mathbf{Q}_i = \begin{pmatrix} \sigma_f^2 & \mathbf{0} \\ \mathbf{0} & \Pi_i \sigma_\alpha^2 \end{pmatrix} = \begin{pmatrix} \sigma_f^2 & 0 & \cdots & 0 \\ 0 & \pi_{11}\sigma_\alpha^2 & \cdots & \pi_{1n_i}\sigma_\alpha^2 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \pi_{1n_i}\sigma_\alpha^2 & \cdots & \pi_{n_in_i}\sigma_\alpha^2 \end{pmatrix},$$

where σ_f^2 is the variance of family-specific effects as defined earlier, σ_α^2 is proportional to the additive genetic variance of the QTL, and Π_i is an $n_i \times n_i$ IBD matrix with the jj' element $\pi_{jj'}$ defined as the IBD proportion shared by sibs j and j' at the QTL. Note that dominance and other effects are assumed absent. The vector of random effects is now defined as $\gamma_i = (f_i\ g_{i1} \cdots g_{in_i})^T$, where g_{ij} is the additive genetic effect of the QTL of the jth sib in the ith family. The IBD matrix at the QTL is unobservable but can be estimated from the IBD matrices of linked markers using the multipoint method (Kruglyak and Lander 1995; Almasy and Blangero 1998). The problem with the IBD-based method is that one needs to invert the matrix $\mathbf{Q_i}$, which further requires inverting matrix Π_i. However, in most situations Π_i is not invertable (not of full rank). Numerically, one must calculate the rank of Π_i, reduce the dimension of γ_i, and reconstruct a full-ranked Π_i with a reduced dimension. The principle of the IBD-based method is not complicated, but implementation of the method deserves further investigation.

In this study, we have assumed that QTL effects are normally distributed. In reality, the number of alleles and the allelic frequencies of a putative QTL in the base population are rarely known, nor are the distributions of the effects of the QTL. However, drawing inferences about the QTL variance via the normal distribution is a natural way to characterize genetic variation in the base population. In addition, normal distribution of the allelic effects is usually a very robust assumption. This has been verified for normal traits by Xu and Atchley (1995), who found that, for data simulated under a biallelic model, the analysis based on the normal distribution provided very accurate estimates of QTL variances.

Although we demonstrate the statistical method of QTL mapping using full-sib families as an example, in principle, families from other types of mating designs can be readily incorporated by modifying model (2) and the conditional probabilities p^{s}_{ijl} and p^{d}_{ijl}. The model considered here assumes only one QTL on the chromosome. In reality, complex binary traits may be controlled by multiple loci. With our random model approach, QTL located on other chromosomes will be absorbed into the polygenic term. If there are multiple QTL in the same chromosome, the estimation tends to be biased because of interference caused by QTL located on the same chromosome but outside the tested region. This problem can be solved by resorting to the concept of composite interval mapping (Jansen 1994; Zeng 1994). Our model can be readily extended to implement composite interval mapping by incorporating markers on other chromosome regions as covariates. These controlled marker effects can be equally treated as random effects and their variances can absorb QTL variances outside the tested region so that bias can be reduced or eliminated (Xu and Atchley 1995).

We thank Dr. Bruce Walsh and an anonymous reviewer for their critical comments on an earlier version of the manuscript. We also thank Dr. Damian Gessler for his helpful comments on the manuscript. This research was supported by the National Institutes of Health Grant GM55321-01 and the U.S. Department of Agriculture National Research Initiative Competitive Grants Program 97-35205-5075.

LITERATURE CITED

Almasy, L., and J. Blangero, 1998 Multipoint quantitative-trait linkage analysis in general pedigrees. Am. J. Hum. Genet. **62:** 1198–1211.

Chan, J. S. K., and A. C. Kuk, 1997 Maximum likelihood estimation for probit-linear mixed models with correlated random effects. Biometrics **88:** 86–97.

Fahrmeir, L., and G. Tutz, 1994 *Multivariate Statistical Modeling Based on Generalized Linear Models*. Springer-Verlag, New York.

Falconer, D. S., and T. F. C. Mackay, 1996 *Introduction to Quantitative Genetics*, Ed. 4. Longman, London.

Fulker, D. W., and L. R. Cardon, 1994 A sib-pair approach to

interval mapping of quantitative trait loci. Am. J. Hum. Genet. **54:** 1092–1103.
Goldgar, D. E., 1990 Multipoint analysis of human quantitative genetic variation. Am. J. Hum. Genet. **47:** 957–967.
Grignola, F. E., I. Hoeschele and B. Tier, 1996 Mapping quantitative trait loci via residual maximum likelihood. I. Methodology. Genet. Sel. Evol. **28:** 479–490.
Haley, C. S., and S. A. Knott, 1992 A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity **69:** 315–324.
Haseman, J. K., and R. C. Elston, 1972 The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:** 3–19.
Hackett, C. A., and J. I. Weller, 1995 Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics **51:** 1252–1263.
Harville, D. A., and R. W. Mee, 1984 A mixed-model procedure for analyzing ordered categorical data. Biometrics **40:** 393–408.
Heath, S. C., 1997 Markov chain Monte Carlo segregation and linkage analysis of oligogenic models. Am. J. Hum. Genet. **61:** 784–760.
Jansen, R. C., 1993 Interval mapping of multiple quantitative trait loci. Genetics **135:** 205–211.
Jansen, R. C., 1994 Controlling the type I and type II errors in mapping quantitative trait loci. Genetics **138:** 871–881.
Knott, S. A., J. M. Elsen and C. S. Haley, 1996 Methods for multiple-marker mapping of quantitative trait loci in half-sib populations. Theor. Appl. Genet. **93:** 71–80.
Kruglyak, E. S., and E. S. Lander, 1995 Complete multipoint sib-pair analysis of qualitative and quantitative traits. Am. J. Hum. Genet. **57:** 439–454.
Lander, E. S., and D. Botstein, 1989 Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics **121:** 185–199.
Lander, E. S., and L. Kruglyak, 1995 Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat. Genet. **11:** 241–247.
Lynch, M., and B. Walsh, 1998 *Genetics and Analysis of Quantitative Traits.* Sinaur Associates, Sunderland, MA.
McCulloch, C. E., 1994 Maximum likelihood variance components estimation for binary data. J. Am. Stat. Assoc. **89:** 330–335.
Muranty, H., 1996 Power of tests for quantitative trait loci detection using full-sib families in different schemes. Heredity **76:** 156–165.
Olson, J. M., 1995 Multipoint linkage analysis using sib pairs: an interval mapping approach for dichotomous outcomes. Am. J. Hum. Genet. **56:** 788–798.
Rao, S., and S. Xu, 1998 Mapping quantitative trait loci for ordered categorical traits in four-way crosses. Heredity **81:** 214–224.
Rebai, A., 1997 Comparison of methods for regression interval mapping in QTL analysis with non-normal traits. Genet. Res. **69:** 69–74.
Satagopan, R. J., B. S. Yandell, M. A. Newton and T. C. Osborn, 1996 A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. Genetics **144:** 805–816.
Schork, N. J., 1993 Extended multipoint identity-by-descent analysis of human quantitative traits: efficiency, power, and modeling considerations. Am. J. Hum. Genet. **53:** 1306–1313.
Searle, S. R., G. Casella and C. E. McCulloch, 1992 *Variance Components.* John Wiley & Sons, New York.
Sillanpaa, M. J., and E. Arjas, 1998 Bayesian mapping of multiple quantitative trait loci from incomplete inbred line cross data. Genetics **148:** 1373-1388.
Sorensen, D. A., D. Erson, D. Gianola and I. Korsgaard, 1995 Bayesian inference in threshold models using Gibbs sampling. Genet. Sel. Evol. **27:** 229–249.
Uimari, P. G., and I. Hoeschele, 1997 Mapping linked quantitative trait loci using Bayesian method analysis and Markov chain Monte Carlo algorithms. Genetics **146:** 735–743.
Visscher, P. M., C. S. Haley and S. A. Knott, 1996 Mapping QTLs for binary traits in backcross and F_2 populations. Genet. Res. **68:** 55–63.
Williams, C. G., 1998 QTL mapping in outbred pedigrees, pp. 81–94 in *Molecular Dissection of Complex Traits*, edited by A. H. Paterson. CRC Press, New York.
Wright, S., 1934 An analysis of variability in number of digits in an inbred strain of guinea pigs. Genetics **19:** 506–536.
Xie, C., D. D. G. Gessler and S. Xu, 1998 Combining different line crosses for mapping quantitative trait loci using the IBD-based variance component method. Genetics **149:** 1139–1146.
Xu, S., 1996a Computation of the full likelihood function for estimating variance at a quantitative trait locus. Genetics **144:** 1951–1960.
Xu, S., 1996b Mapping quantitative trait loci using four-way crosses. Genet. Res. **68:** 175–181.
Xu, S., 1998 Mapping quantitative trait loci using multiple families of line crosses. Genetics **148:** 517–524.
Xu, S., and W. R. Atchley, 1995 A random model approach to interval mapping of quantitative trait loci. Genetics **141:** 1189–1197.
Xu, S., and W. R. Atchley, 1996a Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics **143:** 1417–1424.
Xu, S., and W. R. Atchley, 1996b A Monte-Carlo algorithm for maximum likelihood estimation of variance components. Genet. Sel. Evol. **28:** 329–343.
Yi, N., and S. Xu, 1999 Mapping quantitative trait loci for complex binary traits in outbred populations. Heredity **82:** 668–676.
Zeng, Z. B., 1994 Precision mapping of quantitative trait loci. Genetics **136:** 1457–1468.

Communicating editor: T. F. C. Mackay

APPENDIX

Calculation of the posterior modes and curvatures of random effects using the Fisher-scoring algorithm

We use the Fisher-scoring algorithm to calculate the posterior modes and posterior curvatures of both the random effects γ and the fixed effects β. The MLE of $\theta = (\beta^T, \gamma_1^T, \cdots, \gamma_1^T)^T$ that satisfies $\partial L/\partial\theta$ can be calculated by using the Fisher-scoring algorithm,

$$\theta^{(k+1)} = \theta^{(k)} + \mathbf{F}^{-1}(\theta^{(k)})\mathbf{S}(\theta^{(k)}),$$

where k denotes an iteration index,

$$\mathbf{S}(\theta) = \frac{\partial L}{\partial \theta} = (\mathbf{S}_\beta^T, \mathbf{S}_1^T, \cdots, \mathbf{S}_n^T)^T = \left(\frac{\partial L}{\partial \beta^T}, \frac{\partial L}{\partial \gamma_1^T}, \cdots, \frac{\partial L}{\partial \gamma_n^T}\right)^T$$

is the score function, and

$$\mathbf{F}(\theta) = E[\mathbf{S}(\theta)\mathbf{S}(\theta)^T] = \begin{pmatrix} E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \beta^T}\right) & E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \gamma_i^T}\right) \\ E\left(\frac{\partial L}{\partial \gamma_i}\frac{\partial L}{\partial \beta^T}\right) & E\left(\frac{\partial L}{\partial \gamma_i}\frac{\partial L}{\partial \gamma_i^T}\right) \end{pmatrix}$$

is the Fisher information matrix. The components of the score vector and the Fisher information matrix are

$$\frac{\delta L}{\delta \beta} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{s_{ij} - P_{ij}}{P_{ij}(1 - P_{ij})}\frac{\partial P_{ij}}{\partial \mu} = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{(s_{ij} - P_{ij})\phi(\mu_{ij}/\sqrt{V_{ij}})}{P_{ij}(1 - P_{ij})\sqrt{V_{ij}}}\mathbf{x}_{ij},$$

$$\frac{\partial L}{\partial \gamma_2} = \sum_{j=1}^{n_i} \frac{s_{ij} - P_{ij}}{P_{ij}(1 - P_{ij})}\frac{\partial P_{ij}}{\partial \gamma_1} - \mathbf{Q}^{-1}\gamma_i = \sum_{j=1}^{n_i} \frac{(s_{ij} - P_{ij})\phi(\mu_{ij}/\sqrt{V_{ij}})}{P_{ij}(1 - P_{ij})\sqrt{V_{ij}}}$$

$$\times \left[E(\mathbf{w}_{ij}|I_M) - \frac{\mu_{ij}\mathrm{Var}(\mathbf{w}_{ij}|I_M)\gamma_i}{V_{ij}} \right] - \mathbf{Q}^{-1}\gamma_i,$$

$$E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \beta^T}\right) = \sum_{i=1}^{n}\sum_{j=1}^{n_i} \frac{\left[\phi(\mu_{ij}/\sqrt{V_{ij}})\right]^2}{P_{ij}(1-P_{ij})V_{ij}} x_{ij}x_{ij}^T,$$

$$E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \gamma_i^T}\right) = E\left(\frac{\partial L}{\partial \gamma_i}\frac{\partial L}{\partial \beta^T}\right)^T = \sum_{j=1}^{n_i} \frac{\left[\phi(\mu_{ij}/\sqrt{V_{ij}})\right]^2}{P_{ij}(1-P_{ij})V_{ij}} x_{ij} \times \left[E(w_{ij}|I_M) - \frac{\mu_{ij}\mathrm{Var}(w_{ij}|I_M)\gamma_i}{V_{ij}} \right]$$

$$E\left(\frac{\partial L}{\partial \beta}\frac{\partial L}{\partial \gamma^T}\right) = \sum_{j=1}^{n_i} \frac{\left[\phi(\mu_{ij}/\sqrt{V_{ij}})\right]^2}{P_{ij}(1-P_{ij})V_{ij}} \times \left[E(\mathbf{w}_{ij}|I_M) - \frac{\mu_{ij}\mathrm{Var}(w_{ij}|I_M)\gamma_i}{V_{ij}} \right], \left[E(\mathbf{w}_{ij}|I_M) - \frac{\gamma_{ij}\mathrm{Var}(\mathbf{w}_{ij}|I_M)\gamma_i}{V_{ij}} \right]^T + \mathbf{Q}^{-1},$$

$$E\left(\frac{\partial L}{\partial \gamma_k}\frac{\partial L}{\partial \gamma_l^T}\right) = 0 \quad \text{for } k \neq l,$$

where $\phi(\cdot)$ is the probability density of the standardized normal distribution. Vector $\mathbf{w}_{ij}$ is defined as $\mathbf{w}_{ij} = (1\ z_{ij}^T\mathbf{H})^T$, which leads to

$$E(\mathbf{w}_{ij}|I_M) = (1\ E(\mathbf{z}_{ij}^T|I_M)\mathbf{H})^T$$

and

$$\mathrm{Var}(\mathbf{w}_{ij}|I_M) = \begin{pmatrix} 0 & \mathbf{0}^T \\ \mathbf{0} & \mathbf{H}^T\mathrm{Var}(\mathbf{z}_{ij}|I_M)\mathbf{H} \end{pmatrix},$$

where $\mathbf{0} = (0\ 0\ 0\ 0)^T$.

Let $F_{\beta\beta} = E\partial L/\partial\beta, \partial L/\partial\beta^T\ F_{\beta i} = F_{i\beta}{}^T = E(\partial L/\partial\beta\ \partial L\partial\gamma^T$, and $\mathbf{F}_{ii} = E(\partial L\partial\beta), \partial L/\partial\beta^T$. The Fisher information matrix has the special structure

$$F(\theta) = \begin{pmatrix} \mathbf{F}_{\beta\beta} & \mathbf{F}_{\beta 1} & \mathbf{F}_{\beta 2} & \cdots & \mathbf{F}_{\beta n} \\ \mathbf{F}_{1\beta} & \mathbf{F}_{11} & 0 & \cdots & 0 \\ \mathbf{F}_{2\beta} & 0 & \mathbf{F}_{22} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbf{F}_{n\beta} & 0 & 0 & \cdots & \mathbf{F}_{nn} \end{pmatrix}.$$

Because the lower right part of **F(θ)** is block diagonal, the Fisher-scoring algorithm can be reexpressed more simply as

$$\mathbf{F}_{\beta\beta}^{(k)}\Delta\beta^{(k)} + \sum_{i=1}^{n}\mathbf{F}_{\beta i}^{(k)}\Delta\gamma_i^{(k)} = S_\beta^{(k)},$$

$$\mathbf{F}_{i\beta}^{(k)}\Delta\beta^{(k)} + \mathbf{F}_{ii}^{(k)}\Delta\gamma_i^{(k)} = \mathbf{S}_i^{(k)}, \quad i = 1, \cdots, n,$$

where

$$\Delta\beta^{(k)} = \beta^{(k+1)} - \beta^{(k)} \quad \text{and} \quad \Delta\gamma^{(k)} = \gamma_i^{(k+1)} - \gamma_i^{(k)}.$$

After some transformations, the algorithm

$$\Delta\beta^{(k)} = \left[\mathbf{F}_{\beta\beta}^{(k)} - \sum_{i=1}^{n} F_{\beta i}^{(k)}(F_{ii}^{(k)})^{-1}F_{i\beta}^{(k)}\right]^{-1}\left[S_\beta^{(k)} - \sum_{i=1}^{n}(F_{ii}^{(k)})^{-1}S_i^{(k)}\right]$$

is obtained, where each iteration step implies working off the data twice to obtain first the corrections (Fahrmeir and Tutz 1994), and then

$$\Delta\gamma_i^{(k)} = (F_{ii}^{(k)})^{-1}[S_i^{(k)} - F_{i\beta}^{(k)}\Delta\beta^{(k)}], \quad i = 1, \cdots, n.$$

With the above expressions, inversion of matrix **F(θ)** has been replaced by inversions of many matrices of smaller sizes, *i.e.*, $\mathbf{F}_{\beta\beta}$ and $\mathbf{F}_{ii}$ for $i = 1, \cdots, n$.

A nice property of the Fisher-scoring algorithm is that the variance-covariance matrix of $\hat{\theta}$ can be approximated by the inverse of the Fisher information matrix, *i.e.*, $\mathrm{Var}(\hat{\boldsymbol{\theta}}) \approx \mathbf{F}(\hat{\theta})^{-1}$. Because the resulting estimate $\hat{\theta}$ is a MLE, it follows a multivariate normal distribution if the family sizes n_i are sufficiently large, *i.e.*, $\hat{\theta} \sim N(\boldsymbol{\theta}, \mathbf{F}(\hat{\theta})^{-1})$. As a result, the posterior mode and curvature evaluated at the mode are good approximations of the posterior mean and covariance matrix. The inverse matrix, $\mathbf{F(\theta)^{-1}}$, is obtained using standard formulas for inverting partitioned matrices (Fahrmeir and Tutz 1994). The result is summarized as

$$F(\theta)^{-1} = \begin{pmatrix} \mathbf{V}_{\beta\beta} & \mathbf{V}_{\beta 1} & \mathbf{V}_{\beta 2} & \cdots & V_{\beta n} \\ \mathbf{V}_{1\beta} & \mathbf{V}_{11} & V_{12} & \cdots & V_{1n} \\ \mathbf{V}_{2\beta} & V_{21} & \mathbf{V}_{22} & \cdots & \mathbf{V}_{2n} \\ \vdots & & & & \vdots \\ \mathbf{V}_{n\beta} & \mathbf{V}_{n1} & \mathbf{V}_{n1} & \vdots & V_{nn} \end{pmatrix}.$$

and

$$V_{\beta\beta} = (F_{\beta\beta} - \sum_{i=1}^{n} F_{\beta i}F_{ii}^{-1}F_{i\beta})^{-1},$$

$$V_{\beta i} = V_{i\beta}^T = -V_{\beta\beta}F_{\beta i}F_{ii}^{-1},$$

$$V_{ii} = F_{ii}^{-1} + F_{ii}^{-1}F_{i\beta}V_{\beta\beta}F_{\beta i}F_{ii}^{-1},$$

and

$$V_{ij} = V_{ji}^T = F_{ii}^{-1}F_{i\beta}V_{\beta\beta}F_{\beta j}F_{jj}^{-1} \quad \text{for } i \neq j.$$

According to our experience, the Fisher-scoring algorithm behaves very well with regard to convergence to a local maximum. The algorithm is also relatively fast; *e.g.*, convergence usually took <10 iterations in most situations examined in this study.