Am. J. Hum. Genet. 51:1111–1126, 1992

A Monte Carlo Method for Combined Segregation and Linkage Analysis

Sun Wei Guo* and Elizabeth A. Thompson†

*Department of Biostatistics, University of Michigan, Ann Arbor; and †Department of Statistics, University of Washington, Seattle

Summary

We introduce a Monte Carlo approach to combined segregation and linkage analysis of a quantitative trait observed in an extended pedigree. In conjunction with the Monte Carlo method of likelihood-ratio evaluation proposed by Thompson and Guo, the method provides for estimation and hypothesis testing. The greatest attraction of this approach is its ability to handle complex genetic models and large pedigrees. Two examples illustrate the practicality of the method. One is of simulated data on a large pedigree; the other is a reanalysis of published data previously analyzed by other methods.

Introduction

The past decade has seen enormous success in mapping discrete traits. Notable successes are the localization of genes for Huntington disease, cystic fibrosis, and Duchenne muscular dystrophy. In contrast, progress in mapping quantitative traits has been very slow, despite the fact that many relevant measures of diseases are clinical, physiological, and biological traits that vary continuously among individuals. There is no shortage of data. Advances in biology and molecular genetics have generated so much data that the availability of statistical techniques has become a bottleneck in the process of the mapping of quantitative traits.

The current available techniques for mapping quantitative traits can be grouped into four categories: (1) sib-pair methods (Haseman and Elston 1972), (2) discrete-type linkage analysis (Ott 1991; Thomas and Cortessis 1991), (3) mixed models (Hasstedt 1982), and (4) regressive models (Bonney 1984; Bonney et al. 1988). Although sib-pair methods are fairly robust and have the advantage of no need to make ascertainment corrections, their statistical power is very low, especially when linkage is loose. In addition, they ignore the interdependency among sib pairs from the same nuclear family or same pedigree, resulting in a loss of information inherent in the pedigrees. At best, they can only tell whether there is a linkage and are thus primarily a screening device (Elston 1984).

One commonly used approach to mapping a quantitative trait is to dichotomize the trait and continue the linkage analysis as if the trait were discrete (Ott 1991; Thomas and Cortessis 1991). This approach is not only arbitrary in the choice of the cutoff point but also suffers loss of information. In addition, penetrances must be estimated with respect to this arbitrary cutoff. An alternative method is to assume normal densities as the penetrance functions for a quantitative trait, given the gene frequency and the mean and variance for each genotype (Ott 1974). Since quantitative traits are probably typically controlled by a number of loci acting in concert with environmental effects, the adequacy of these models is questionable.

The regressive models proposed by Bonney (1984; also see Bonney et al. 1988) represent a new development. The model handles the residual variation unaccounted for by the major-gene effects as if it were noise, without specifying its origin. Furthermore, the model assumes a Markovian dependence structure with regard to the residuals among first-degree relatives. By doing so, the model provides both flexibility

Received March 4, 1992; revision received June 26, 1992.

Address for correspondence and reprints: Elizabeth A. Thompson, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195.

0002-9297/92/5105-0021$02.00

in incorporation of covariates and efficiency in computation. However, while simple Markovian dependence may be adequate for modeling nuclear-family data, the dependence structure on an extended pedigree is more complex, and there can be strong correlations even among distant relatives, such as are the result of mitochondrial effects. In addition, the regressive model has difficulty in modeling the data where there are multiple marriages or missing data, both of which are common, and in allowing for ascertainment correction. In a simulation study, Konigsberg et al. (1989) show that, for data generated under a mixed model, the regressive model does not perform as well as the mixed model, in segregation analysis.

Methods for analysis of the mixed model (Morton and MacLean 1974) have been a valuable advance in human quantitative genetic analysis (Ott 1979; Lalouel and Morton 1981; Hasstedt 1982). The model partitions the variation in a quantitative trait into three main sources: (1) major-gene effects, (2) additive polygenic and/or other heritable/nonheritable effects, and (3) the independent random effects of the environment. Although the model is biologically more realistic, computational difficulties have limited its use mainly to segregation analysis, rather than in conjunction with linkage analysis, and primarily to data on nuclear families or small pedigrees (Ott 1979; Hasstedt 1982).

Traditionally, segregation and linkage analyses have been performed separately (Ott 1991). Historically, with limited genetic marker data and computing power, most linkage analyses were carried out only after sufficient information had been gathered to infer a mode of inheritance for the trait. However, segregation analysis can only, at best, demonstrate the presence of major gene(s). It cannot localize them, and it often lacks power to estimate genetic parameters correctly in the presence of multiallelic trait loci or genetic heterogeneity (Risch 1984; Ott 1990). Violation of the distributional assumptions of the mixed model can lead to spurious support for a major gene (MacLean et al. 1975; Go et al. 1978; Eaves 1983). Incorporation of linked markers might potentially improve the robustness of the mixed model. Moreover, linkage of a trait to a genetic marker not only provides unequivocal evidence of the existence of a gene locus controlling the trait but also specifies the region where the gene is located. However, linkage analysis alone cannot elucidate the mode of inheritance (Risch 1984). With the increasing availability of highly polymorphic DNA markers, it is a logical step to combine linkage and segregation analyses (Elston et al. 1989).

When the underlying genetic mechanism is complex, genetic heterogeneity and misspecification of models create difficulties for both segregation and linkage analysis. It is useful to be able to analyze data on large pedigrees, which, in general, are more homogeneous than a pooled sample of many nuclear families. It is also useful to consider more realistic yet more complicated genetic models that can incorporate various heritable/nonheritable random and fixed effects and to develop practical computational methods to accomplish the computation.

In this paper, we propose a Monte Carlo approach to combined segregation and linkage analysis for quantitative traits, which extends our previous work on the Monte Carlo estimation of variance-component models and mixed models (Guo and Thompson 1991, and submitted). The greatest attraction of the approach is that it can handle complex genetic models and data on large pedigrees. In the next section, we describe the method and computational algorithm. The practicality of the approach is then illustrated by two examples. Finally, we discuss the proposed method in relation to other recent work in this area and indicate directions for future research.

Methods

Notation and Assumptions

Consider an n-member pedigree on which a continuous trait y and marker phenotype M are observed. Not all pedigree members need be observed, and marker and trait data need not be available for the same individuals. Suppose trait data are available for k of the pedigree members. For clarity of exposition, we will consider an additive mixed model with a major autosomal locus with two alleles, a polygenic effect, and an independent individual-specific residual effect, without fixed or covariate effects. Extension to include fixed covariate effects and dominance or other heritable or nonheritable random effects is straightforward (Guo and Thompson 1991, and submitted), but in this paper we focus on the inclusion of marker data rather than on complexities of the trait model.

For technical reasons (see Discussion), we consider only a diallelic marker locus. To fix notation, let the two alleles of the major-gene trait locus be D and d, with gene frequencies p and $1 - p$, respectively. Let

the two alleles at the marker locus be B and b, with gene frequencies q and $1 - q$. Let G_j denote the jth individual's combined genotype at trait and marker loci. It is assumed that each of the three genotypes DD, Dd, and dd, denoted as 1, 2, and 3, respectively, makes a specific contribution μ_i ($i = 1, 2, 3$) to the phenotype. It is also assumed that the trait and marker loci are in linkage equilibrium, with each locus in Hardy-Weinberg equilibrium. For a given genotypic configuration G on the pedigree, let $\mathbf{1}_i$ be an indicator vector, with the jth entry equal to 1 or 0, depending on whether the jth individual has genotype i at the quantitative trait locus ($i = 1, 2, 3$). Similarly, we let $\mathbf{1}_F$ be an indicator vector with entry j equal to 1 or 0, depending on whether the jth individual is a founder. For a given major-genotype configuration, the phenotypic model for the quantitative trait y is

$$\mathbf{y} = \mu_1\mathbf{1}_1 + \mu_2\mathbf{1}_2 + \mu_3\mathbf{1}_3 + \mathbf{a} + \mathbf{e}, \qquad (1)$$

where a is a vector of additive genetic (polygenic) effects, and e is a vector of individual environmental effects. Each of the vectors a and e is assumed to be normally distributed with mean 0, e having variance-covariance matrix $\sigma_e^2\mathbf{I}$, and a having variance $\sigma_a^2\mathbf{A}$, where A is the numerator relationship matrix (Henderson 1976). The marker phenotype configuration on the observed individuals of the pedigree will be denoted M.

The EM Framework

There are a total of eight parameters to be estimated: the allele frequencies p and q, the recombination fraction r, major-gene effects μ_i, $i = 1, 2, 3$, and the polygenic and residual variances σ_a^2 and σ_e^2. However, estimation of q within the pedigree analysis is often of secondary interest, as considerable information on the marker may have accumulated. Besides, if the marker is codominant, as is usually the case, q can be easily estimated from observed marker phenotypes. Therefore, we assume that q is known and let $\theta = \{p, \mu_1, \mu_2, \mu_3, \sigma_a^2, \sigma_e^2, r\}$ denote the vector of parameters to be estimated. The likelihood for model (1) is

$$\begin{aligned} L(\theta) = P_\theta(\mathbf{y},\mathbf{M}) &= \sum_{\mathbf{G}}\int_{\mathbf{a}} f_\theta(\mathbf{y}|\mathbf{a},\mathbf{G})P(\mathbf{M}|\mathbf{G})P_\theta(\mathbf{G})dP_\theta(\mathbf{a}) \\ &= \sum_{\mathbf{G}} f_\theta(\mathbf{y}|\mathbf{G})P(\mathbf{M}|\mathbf{G})P_\theta(\mathbf{G}), \qquad (2) \end{aligned}$$

where G is the combined two-locus genotypic configuration on the pedigree, and the sum is over all possible genotypic configurations on the pedigree. The marker penetrance probability, $P(\mathbf{M}|\mathbf{G})$, is either 1 or 0, depending on whether the combined genotype conforms with the marker data. Given any genotypic configuration G, $f_\theta(\mathbf{y}|\mathbf{G})$ is the likelihood for an additive polygenic model (Ott 1979). Also,

$$P_\theta(\mathbf{G}) = \prod_{\text{founders } l} P_\theta(G_l) \prod_{\text{nonfounders } j} P_\theta(G_j|G_{m_j}, G_{f_j}),$$

where m_j and f_j are the parents of j, $P(G_l)$ is the genotypic frequency and is a function of p and q, and $P(G_j|G_{m_j}, G_{f_j})$ is the two-locus transmission probability and is, in general, a function of the recombination fraction r.

A framework for estimation of model (1) is as a "missing data problem," with a and G missing. Thus, formulation of an EM algorithm is appropriate. The form of the EM equations for p, σ_a^2, σ_e^2, and μ_i ($i = 1, 2, 3$) are analogous to those given by Guo and Thompson (submitted):

$$\begin{aligned} p^* &= \frac{\mathrm{E}_\theta(2\mathbf{1}_F'\mathbf{1}_1 + \mathbf{1}_F'\mathbf{1}_2|\mathbf{y},\mathbf{M})}{2\mathbf{1}_F'\mathbf{1}} \\ \mu_i^* &= \frac{\mathrm{E}_\theta[\mathbf{1}_i'(\mathbf{y}-\mathbf{a})|\mathbf{y},\mathbf{M}]}{\mathrm{E}_\theta(\mathbf{1}_i'\mathbf{1}|\mathbf{y},\mathbf{M})} \quad (i = 1,2,3) \\ \sigma_a^{2*} &= \frac{1}{n}\mathrm{E}_\theta(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}|\mathbf{y},\mathbf{M}) \\ \sigma_e^{*2} &= \frac{1}{n}\mathrm{E}_\theta(\mathbf{e}'\mathbf{e}|\mathbf{y},\mathbf{M}). \end{aligned} \qquad (3)$$

The added feature here is the inclusion of the linked marker, with marker phenotypes M, and the estimation of the recombination fraction r.

To obtain the EM equation for r, suppose that (a, G) were observed for all n individuals of the pedigree. Given G, estimation of r would be a matter of counting recombinants. We could restrict attention to those parent pairs in which at least one parent is doubly heterozygous; only these are informative for linkage (e.g., see Ott 1991). Let H_i ($H_i = 0, 1, 2$) be the number of doubly heterozygous parents in the ith parent-offspring trio, and let R_i be the expected number (given G) of recombinant events in segregation from the doubly heterozygous parents to the offspring ($R_i = 0, 1, 2$, $R_i \leqslant H_i$). The values of H_i and R_i for all

possible informative matings are easily determined (Thomas and Cortessis 1991). For example, for a mating $db/dB \times db/DB$, $H_i = 1$, and only the second parent is informative. From this parent, R_i is 0 for a db/db offspring and is 1 for a dB/dB offspring, while for a db/dB offspring the recombination count is not determined, but the expectation is r. For a mating $db/DB \times dB/Db$, $H_i = 2$ and $R_i = 1$ for doubly homozygous or doubly heterozygous offspring (e.g., db/db or db/DB), while for the remainder, such as db/dB, $R_i = 2r^2/[r^2 + (1 - r)^2]$. Given the genetic configuration $\mathbf{G}$ on the pedigree, $H = \Sigma_i H_i$ and $R = \Sigma_i R_i$ are the two sufficient statistics for the recombination fraction r (Thomas and Cortessis 1991), and, if $\mathbf{G}$ were observed, the maximum-likelihood estimate (MLE) of r would be R/H. Hence, the EM equation for r is

$$r^* = \frac{\mathrm{E}_\theta(R|\mathbf{y},\mathbf{M})}{\mathrm{E}_\theta(H|\mathbf{y},\mathbf{M})}. \quad (4)$$

Sex-specific recombination fractions can be estimated with minor modification, by counting segregation in males and in females separately.

Monte Carlo Estimation

Despite the simplicity of the EM framework, implementation is not immediate, since there is no way to valuate explicitly the conditional expectations such as those in equations (3) and (4). Instead, Guo and Thompson (1991, and submitted) have proposed a Monte Carlo EM algorithm, using the Gibbs sampler to (*a*) obtain, at current parameter values θ, realizations of the major genotypes and polygenic values ($\mathbf{a}$, $\mathbf{G}$) given the data ($\mathbf{y}$, $\mathbf{M}$) and hence (*b*) estimate the required conditional expectations. The Gibbs sampler (Geman and Geman 1984; Gelfand and Smith 1990) is an iterative procedure for drawing multiple *dependent* realizations from a distribution known only up to a constant of proportionality. In our case, it provides realizations from the global conditional distribution $P_\theta(\mathbf{y},\mathbf{M}|\mathbf{aG})$, as follows: Beginning from any realization of polygenic values and combined major genotypes, ($\mathbf{a}$, $\mathbf{G}$), that is consistent with phenotypic and marker observations, the polygenic values and genotypes are updated, for each individual in the pedigree in turn (in random order), by sampling from the local conditional distribution at parameter values θ, given observed data (if any) and the polygenic values and genotypes of all other members in the pedigree. This completes one cycle; details will be given in the next section.

The configuration $(\mathbf{a}(t), \mathbf{G}(t))$, $t = 1, 2, \ldots$, obtained at successive cycles is a sample from a Markov chain with stationary distribution $P_\theta(\mathbf{a}, \mathbf{G}|\mathbf{y},\mathbf{M})$. The chain is irreducible, since from any starting major genotype and polygenotype configuration, any set of states ($\mathbf{a}$, $\mathbf{G}$) that has positive probability can be hit in a finite number of steps (Guo and Thompson, submitted). Hence, averages over the Markov chain converge to averages over the stationary distribution, by the ergodic theorem (Breiman 1968, Corollary 6.23 and theorem 7.16). That is, for any integrable function $F(\mathbf{a}, \mathbf{G})$ of the genotypes and for any starting point for the chain,

$$\lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} F(\mathbf{a}(t),\mathbf{G}(t)) \to \mathrm{E}(F) .$$

Thus the Monte Carlo EM algorithm proceeds as follows:

1. First the pedigree and data are input, the pedigree specification is transformed to facilitate Gibbs sampling, the inverse $\mathbf{A}^{-1}$ of the numerator relationship matrix for all members of the pedigree (whether observed or not) is found, and initial estimates of the parameters are set.
2. An initial configuration ($\mathbf{a}$, $\mathbf{G}$) is found by the method of "posterior gene dropping" described below, and a number of cycles of the Gibbs sampler are run. This "burn-in" period (Geyer 1991) reduces the effect of the chosen starting configuration and insures that subsequent realizations are approximately from the required stationary distribution.
3. Each Monte Carlo EM step then proceeds as follows:

3a. E-step: the Gibbs sampler is run at the current parameter values. Every C cycles, a realization of ($\mathbf{a}$, $\mathbf{G}$) for all n members of the pedigree is collected, and the necessary statistics are stored (e.g., $\mathbf{a}^t\mathbf{A}^{-1}\mathbf{A}$, $\mathbf{1}_i^t\mathbf{1}$ and the other functions of $\mathbf{a}$ and $\mathbf{G}$ in the EM eqq. [3] and [4]). (If there are computational costs in using and storing realizations, it may not be computationally efficient to use the realization after every cycle, especially if they are highly dependent; in practice we used $C = 20$.) After N realizations are obtained, the time averages of the statistics are used as Monte Carlo estimates of the expectations in equations (3) and (4).

3b. M-step: the parameter estimates are updated, according to the EM equations (3) and (4).
3c. If desired, the parameter values are printed out, and a likelihood and/or LOD score can be estimated.
4. Step 3 is repeated until the parameter estimates no longer show directional trends over the EM iterations. In the final EM steps, larger Monte Carlo samples can be used to reduce Monte Carlo variation. The parameter estimates are taken as averages over these final EM steps. As an additional check on these estimates, the likelihood surface in the neighborhood of these final values can be explored (Thompson and Guo 1991).

Gibbs Sampler Implementation

To implement the Gibbs sampler, we need, for each individual j, (a) the conditional distribution of his combined major genotype G_j given y_j, M_j, a_j, and the major genotypes of other members in the pedigree and (b) the conditional distribution of his polygenic value, a_j, given y_j, his combined genotype G_j, and the polygenic values of other members in the pedigree. For individual j, conditioning on the major genotypes of all other pedigree members involves only his immediate neighbors: his parents, (if present in the pedigree), his spouse(s) (if any), and his offspring (if any). The genotypes of other pedigree members do not contribute further information. Hence, if we let $G^{(j)}$ denote the genotypes of all pedigree members except j, let G_{fj} and G_{mj} denote the genotypes of the parents of the jth individual, let $\{G_{sj}\}$ denote his spouse's genotype, and let $\{G_{sjl}\}$ denote his offspring's genotype, then

$$\begin{aligned} P_\theta(G_j|\mathbf{G}^{(j)}, y_j, a_j, M_j) &= P_\theta(G_j|\{G_{sj}\},\{G_{sjl}\},G_{fj},G_{mj},y_j,a_j,M_j) \\ &\alpha \left[\prod_{j,l} P_\theta(G_{sjl}|G_j,G_{sj})\right] P_\theta(G_j|G_{fj},G_{mj}) P_\theta(M_j|G_j) f_\theta(y_j|G_j,a_j) \\ &\alpha \left[\prod_{ij,l} P_\theta(G_{sjl}|G_j,G_{sj})\right] P_\theta(G_j|G_{fj},G_{mj}) P_\theta(M_j|G_j) \\ &\quad \exp\left[-\frac{J_d(y_j-\mu_{G_j}-a_j)^2}{2\sigma_e^2}\right] \end{aligned} \tag{5}$$

where $J_d = 1$ or 0, as the trait is or is not observed. If j is a founder, then the combined two-locus segregation probability $P_\theta(G_j|G_{fj}, G_{mj})$ is defined as the population genotypic frequency $P_\theta(G_j)$. If the marker phenotype is unobserved, then $P_\theta(M_j|G_j)| = 1$ for all possible G_j.

Similarly, the polygenic value a_j can be updated, given an observed phenotype y_j (if any), combined genotype G_j, and polygenic values of neighborhood members; details are given by Guo and Thompson (submitted).

Quantitative traits are often affected by covariates such as age and sex. The effects of these covariates can be estimated by appropriate EM equations, as described by Thompson and Shaw (1990). Alternatively, one can, on the basis of current estimates of covariate effects, sample major genotypes and polygenic effects and then use standard regression methods to estimate covariate effects. The latter method is based on the fact that, given major genotypes and polygenic and other heritable random effects, observations on different individuals are independent. Hence, for each realization of major genotypes and polygenic and other heritable random effects, one can treat these realizations as known covariates in a standard regression analysis.

Choice of Starting Realizations

An ergodic Markov chain with sufficient "burn in" to allow convergence (Geyer 1991) will provide realizations from true joint distribution of genotypic effects and genotypes, (**a**, **G**), given the phenotypic data, (**y**, **M**). However, it is important to choose a good starting genotypic configuration in order to avoid unnecessarily prolonged iteration. The observed data on each individual conditionally on his parents provide partial information on the major-gene effects and recombination fraction. This local information can be used in a "posterior gene-dropping" method to provide a sensible starting point, given the current parameter estimates (Guo and Thompson, submitted). Here the procedure is adapted for marker data.

First, the major genes are simulated, from the top of the pedigree down to the bottom, using the information of current estimates of gene frequency, recombination fraction, major-gene effects, and data. For each founder j in the pedigree and each possible two-locus genotype g, we calculate the probability

$$\begin{aligned} P_\theta(G_j=g|y_j,M_j) &\alpha\ P_\theta(G_j=g)P_\theta(M_j|G_j)f_\theta(y_j|G_j=g) \\ &\alpha\ P_\theta(G_j=g)P_\theta(M_j|G_j)\exp\left[-\frac{(y_j-\mu_{G_j})^2}{2(\sigma_a^2+\sigma_e^2)}\right], \end{aligned}$$

normalizing (for each j) the sum over g to 1. Here $P_\theta(G_j = g)$ is just the frequency of combined genotype,

calculated on the assumption of linkage and Hardy-Weinberg equilibria. If y_j is missing, we let $f_\theta(y_j|G_j = g) = 1$, for all g. If M_j is missing, $P(M_j|G_j)$ is set to 1 for all M_j. A genotype is then randomly selected according to the calculated probability. Once all founders are assigned combined genotypes, we can drop the genes to nonfounders. For each nonfounder j, a genotype is generated from the following probability distribution:

$$P_\theta(G_j = g|y_j, G_{fj}, G_{mj}, M_j)$$
$$\alpha\, P(G_j = g|G_{fj}, G_{mj}) P_\theta(M_j|G_j) f_\theta(y_j|G_j = g)$$
$$\alpha\, P(G_j = g|G_{fj}, G_{mj}) P_\theta(M_j|G_j) \exp\left[-\frac{(y_j - \mu_{G_j})^2}{2(\sigma_a^2 + \sigma_e^2)}\right],$$

where f_j and m_j are the parents of j, whose genotypes G_{mj} and G_{fj} are already assigned. With linked markers, however, this "posterior gene-dropping" procedure may not be able to carry through, because it is possible that $P_\theta(G_j = g|y_j, G_{fj}, G_{mj}, M_j) = 0$ for all possible g. This is because some combined genotypes assigned to the parents of, say, the jth individual may not be consistent with his marker phenotype M_j. In practice this is not a problem; if it happens, we restart the "posterior gene-dropping" until all the individuals are assigned combined genotypes that are compatible with their observed marker phenotypes.

Once the major genes at marker and trait loci have been dropped down the pedigree, we drop the polygenic values similarly, conditioning on individual trait values, major genotypes, and already assigned parental polygenic values (Guo and Thompson, submitted).

Estimation of Variance-Covariance Matrix

It is important to estimate the standard errors (SEs) of estimated parameters or to construct confidence intervals for parameter values. We now provide a Monte Carlo estimate of the asymptotic variance-covariance matrix. The method is based on results of Sundberg (1974) and Louis (1982), relating to the analysis of "missing" data. If the observed data are $\mathbf{x}$ and the missing data are $\mathbf{u}$, then

$$-\frac{\partial^2 \log P_\theta(\mathbf{x})}{\partial\theta_i \partial\theta_j} = \mathrm{E}_\theta\left(-\frac{\partial^2 \log P_\theta(\mathbf{u},\mathbf{x})}{\partial\theta_i \partial\theta_j}\Big|\mathbf{x}\right) - \mathrm{cov}_\theta\left(\frac{\partial \log P_\theta(\mathbf{u},\mathbf{x})}{\partial\theta_i}, \frac{\partial \log P_\theta(\mathbf{u},\mathbf{x})}{\partial\theta_j}\Big|\mathbf{x}\right), \quad (6)$$

where θ_i and θ_j are components of θ. In our case, $\mathbf{u} = (\mathbf{a}, \mathbf{G})$ is the "missing data" vector, and the observed data are $\mathbf{x} = (\mathbf{y}, \mathbf{M})$. Each term in the right-hand side of equation (6) can be estimated by the Monte Carlo method, since each consists of conditional expectations of simple functions of $\mathbf{u} = (\mathbf{a}, \mathbf{G})$ given $\mathbf{x} = (\mathbf{y}, \mathbf{M})$. For example, if N realizations $\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \ldots, \mathbf{u}^{(N)}$ are drawn from $P_\theta(\mathbf{u}|\mathbf{x})$, then the first term on the right-hand side can be estimated by

$$-\frac{1}{N}\sum_{l=1}^{N}\frac{\partial^2 \log P_\theta(\mathbf{u}^{(l)},\mathbf{x})}{\partial\theta_i \partial\theta_j},$$

where the Gibbs sampling is run and where derivatives are evaluated at the MLE values of the parameters θ. The other terms are estimated similarly, using the same realizations.

The first and second derivatives of the "complete data" log likelihood, $\log P_\theta(\mathbf{u}, \mathbf{x}) = \log P_\theta(\mathbf{a},\mathbf{G},\mathbf{y},\mathbf{M})$, are easy to evaluate, since $P_\theta(\mathbf{a},\mathbf{G},\mathbf{y},\mathbf{M}) = f_\theta(\mathbf{y}|\mathbf{a}, \mathbf{G}) f_\theta(\mathbf{a}) P_\theta(\mathbf{M}|\mathbf{G}) P_\theta(\mathbf{G})$, and each term has a simple structure, with typically only a subset of the parameters involved in any one component of the model. For example, for the given combined-genotype configuration $\mathbf{G}$, the recombination fraction r appears only in $P_\theta(\mathbf{G})$ and

$$\frac{\partial \log P_\theta(\mathbf{G})}{\partial r} = \sum_{\text{nonfounders } j} \frac{\partial \log P_\theta(G_j|G_{mj}, G_{fj})}{\partial r}. \quad (7)$$

These probabilities and derivatives are given in table 1. Similarly simple formulas can be obtained for the second derivatives.

Thus an estimate of the information matrix is obtained, and it can be inverted to provide an estimate of the variance-covariance matrix of the parameter estimators. As a guide to precision, a nominal 95% confidence interval (± 1.96 estimated SE) can be given for each parameter.

Likelihood-Ratio Evaluation

A general method for Monte Carlo estimation of likelihood ratios was given by Thompson and Guo (1991). For the model (1), the likelihood (2) takes the form

$$L(\theta) = P_\theta(\mathbf{y},\mathbf{M}) = \sum_{\mathbf{G}} f_\theta(\mathbf{y}|\mathbf{G}) P_\theta(\mathbf{M}|\mathbf{G}) P_\theta(\mathbf{G}),$$

where the sum is over all possible combined-genotype configurations in the pedigree. This summation is im-

Table 1

Example of Linkage Segregation Probabilities $P(G_f G_{m_f}, G_{f_f})$ and the First-Order Derivatives of the Logarithm of the Segregation Probabilities

Offspring Genotype	Segregation Probability	First-Order Derivative of Log Segregation Probability
db/db.....	$r(1-r)/4$	$1/r - 1/(1-r)$
db/dB.....	$[r^2 + (1-r)^2]/4$	$2(2r-1)/[r^2 + (1-r)^2]$
dB/dB	$r(1-r)/4$	$1/r - 1/(1-r)$
db/Db.....	$[r^2 + (1-r)^2]/4$	$2(2r-1)/[r^2 + (1-r)^2]$
dB/DB.....	$[r^2 + (1-r)^2]/4$	$2(2r-1)/[r^2 + (1-r)^2]$
Db/Db	$r(1-r)/4$	$1/r - 1/(1-r)$
Db/DB.....	$[r^2 + (1-r)^2]/4$	$2(2r-1)/[r^2 + (1-r)^2]$
DB/DB.....	$r(1-r)/4$	$1/r - 1/(1-r)$
db/DB.....	$r(1-r)/2$	$1/r - 1/(1-r)$
dB/Db.....	$r(1-r)/4$	$1/r - 1/(1-r)$

NOTE.—The parental genotypes are $db/DB \times dB/Db$. The derivative does not exist at $r = 0$.

possible on a large pedigree, because of the prohibitively large number of terms (Ott 1979). However, note that

$$L(\theta_0) = P_{\theta_0}(\mathbf{y},\mathbf{M}) = \frac{P_{\theta_0}(\mathbf{y},\mathbf{M},\mathbf{G})}{P_{\theta_0}(\mathbf{G}|\mathbf{y},\mathbf{M})} = \frac{f_{\theta_0}(\mathbf{y}|\mathbf{G})P_{\theta_0}(\mathbf{M}|\mathbf{G})P_{\theta_0}(\mathbf{G})}{P_{\theta_0}(\mathbf{G}|\mathbf{y},\mathbf{M})}$$

for all G. Hence the likelihood ratio between two parameter values θ and θ_0 can be written in the form

$$\frac{L(\theta)}{L(\theta_0)} = \sum_{\mathbf{G}} \frac{f_\theta(\mathbf{y}|\mathbf{G})}{f_{\theta_0}(\mathbf{y}|\mathbf{G})} \frac{P_\theta(\mathbf{M}|\mathbf{G})}{P_{\theta_0}(\mathbf{M}|\mathbf{G})} \frac{P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{G})} P_{\theta_0}(\mathbf{G}|\mathbf{y},\mathbf{M})$$

$$= \sum_{\mathbf{G}} \frac{f_\theta(\mathbf{y}|\mathbf{G})}{f_{\theta_0}(\mathbf{y}|\mathbf{G})} \frac{P_\theta(\mathbf{G})}{P_{\theta_0}(\mathbf{G})} P_{\theta_0}(\mathbf{G}|\mathbf{y},\mathbf{M})$$

(Thompson and Guo 1991). Thus a Monte Carlo estimate of $L(\theta)/L(\theta_0)$ is

$$\frac{1}{N}\sum_{\mathbf{G}'} \frac{f_\theta(\mathbf{y}|\mathbf{G}')}{f_{\theta_0}(\mathbf{y}|\mathbf{G}')} \frac{P_\theta(\mathbf{G}')}{P_{\theta_0}(\mathbf{G}')}, \tag{8}$$

obtained by sampling N realizations of G′ from $P_{\theta_0}(\mathbf{G}|\mathbf{y},\mathbf{M})$. This is true for any θ and θ_0. For example, if θ_0 is the MLE, we obtained likelihood ratios for other θ values relative to the maximized likelihood.

Note that M does not appear explicitly in the estimator (8); it is involved only through the Gibbs sampling conditional on (y,M). In the preceding equation, the term $P_\theta(\mathbf{M}|\mathbf{G})$ in the numerator cancels with $P_{\theta_0}(\mathbf{M}|\mathbf{G})$ in the denominator, since this marker phenotype penetrance does not depend on the parameters. Nor is a present, although of course realizations of a will be generated alongside those of G. However, $f_\theta(\mathbf{y}|\mathbf{G})$ is a polygenic likelihood involving integration over unobserved a values (eq. [2]). If necessary, this integration may also be replaced by Monte Carlo sampling, but, for a simple polygenic model on a simple pedigree, exact evaluation is possible.

Moreover, any evaluation may in fact be unnecessary. If θ and θ_0 differ only in the recombination fraction r, then $f_\theta(\mathbf{y}|\mathbf{G}) = f_{\theta_0}(\mathbf{y}|\mathbf{G})$ for all G, and these terms also cancel from the likelihood-ratio estimator (8). For linkage analysis, one is often interested in computing the LOD score—a log-likelihood ratio at given values of the other genetic parameters. In this case, a better alternative to running the Gibbs sampler at the MLE is to take $r = .5$ in θ and run the Gibbs sampler at several different θ_0 values differing only in the recombination parameter r_0. Then the estimated LOD score of $r = r_0$ is

$$\mathrm{LOD}(r_0) = -\log_{10}\frac{L(\theta)}{L(\theta_0)} = \log_{10}N - \log_{10}\left[\sum_{\mathbf{G}'} \frac{P_\theta(\mathbf{G}')}{P_{\theta_0}(\mathbf{G}')}\right], \tag{9}$$

with no evaluation of the polygenic likelihood being required. The estimated LOD score curve can then be plotted as a function of r_0.

In the Monte Carlo EM algorithm described above, given the current parameter estimate $\theta^{(k)}$, realizations are obtained from $P_{\theta^{(k)}}(\mathbf{a},\mathbf{G}|\mathbf{y},\mathbf{M})$ and are used to obtain the next parameter estimates, $\theta^{(k+1)}$, say. The realized major genotypes G are realizations from the marginal conditional distribution $P_{\theta^{(k)}}(\mathbf{G}|\mathbf{y},\mathbf{M})$. The same realizations can thus be used to estimate interim LOD scores and likelihood ratios; no additional realizations are needed. However, when satisfactory estimates of the other parameters are obtained, and when the EM procedure is halted, much larger samples should be run at the final estimate and at alternative r values, to provide good estimates of the LOD score curve. These large final samples can also be used to explore likelihood ratios in the neighborhood of the final estimate, to verify that, to within acceptable limits, the MLE has in fact been found.

All the methods above can be used on more than one pedigree; realizations of (a,G) conditional on (y,M) are simply obtained for each, and the required conditional expectations are combined in the EM

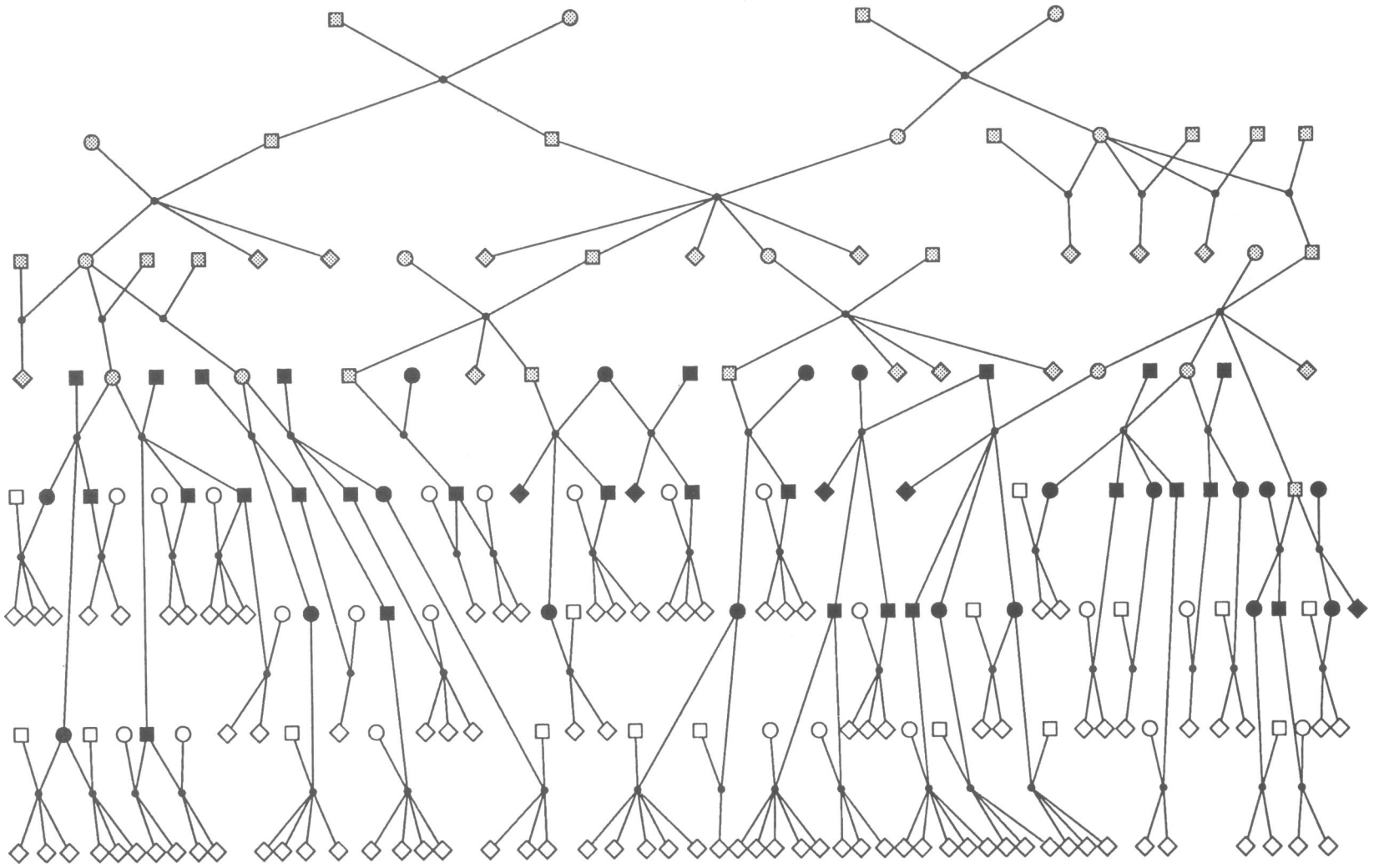

Figure 1 Pedigree structure of the simulated data. The total set of individuals, of grey, black, or white color, constitute the 230-member, six-generation pedigree used in the example in this paper.

equations. For estimation of the information matrix, since the pedigrees are unrelated, the total observed information is simply the sum of the values for the individual pedigrees. The inverse of the observed information matrix is then an estimate of the asymptotic variance-covariance matrix on the total data set. Likewise, the overall LOD score is the sum of the LOD scores on individual pedigrees.

Results

In this section we provide two examples to illustrate the method proposed in the previous section. The programs implementing the method proposed in this paper are written in C. All the computations were carried out on a DECstation 3100. The random-number generator used was the library routine DRAND48. The program PSDRAW (Geyer 1988) was used to draw pedigrees.

Example 1: Simulated Data

We consider one simulated data set on a 230-member, six-generation pedigree (fig. 1). There are 67 founders in the pedigree. Model (1) was used to generate the data; the simulation values are shown in table 2. For simplicity, phenotypic and marker data are observed for each individual. The allele frequency at the marker locus is .5. The same data set, without the marker data, has been used by us elsewhere (Guo and Thompson, submitted).

With starting values $\sigma_a^2 = \sigma_e^2 = 0.5$, $\mu_1 = \mu_2 = 1.0$, $\mu_3 = -1.0$, $r = .25$, and $p = .2$, we performed 200 iterations of Monte Carlo EM. For each EM iteration, 200 Gibbs realizations (**a**,G) were sampled, with 20 cycles of updating of the entire pedigree between each sampling. For the last 10 EM iterations, 1,000 Gibbs realizations were sampled, with 20 cycles between each sampling. Once the final estimates were obtained, 8,000 realizations, with 30 cycles between two consecutive realizations, were drawn and collected to estimate the asymptotic variance-covariance matrix and the LOD scores at various recombination fractions.

Figure 2 shows the LOD score and the parameter estimates against the EM iterations. The Monte Carlo samples used in the EM iterations are not large; figure 2 reflects the continuing random variation in the conditional expectations used for the EM procedure. However, larger samples are unnecessary. Even for this case where the data provide substantial information, the statistical SEs (table 2) are much greater than the SEs in the Monte Carlo sampling. The final estimates, with their estimated SEs and nominal 95% confidence intervals, are shown in table 2. Other starting values for the parameter values led to practically the same results. Table 3 gives the asymptotic variance-covariance matrix of the parameters. Figure 3 shows the final LOD score curve: the estimated maximum LOD score is 10.27. For this particular example, although tight linkage is evident, the estimates of the trait model parameters do not seem to be substantially improved by incorporation of marker data (see Guo and Thompson, submitted). However, both with and without marker data, the strong major-gene effects are correctly inferred. Although the variance-component estimates have higher relative SEs, there is clear evidence of the additive polygenic effect. The high LOD score at small recombination frequencies correctly indicates tight linkage with the marker. In addition, the estimates agree well with the true parameters; in all cases the nominal 95% confidence interval includes the true simulation value.

Table 2

Estimated Parameters, with Their SEs and 95% Confidence Intervals, for Simulated Data

Parameter	True Value	Estimate	SE	95% Confidence Interval
p	.3	.3430	.0471	(.2507, .4353)
μ_1	2.0	2.3400	.2190	(1.9109, 2.7692)
μ_2	.0	.1602	.1805	(−.1936, .5141)
μ_3	−2.0	−2.1133	.1832	(−2.4724, −1.7541)
σ_a^2	.6	.6280	.1557	(0.3227, .9332)
σ_e^2	.2	.1535	.0694	(0.0174, .2895)
r	.1	.0385	.0325	(0.0000, .1021)

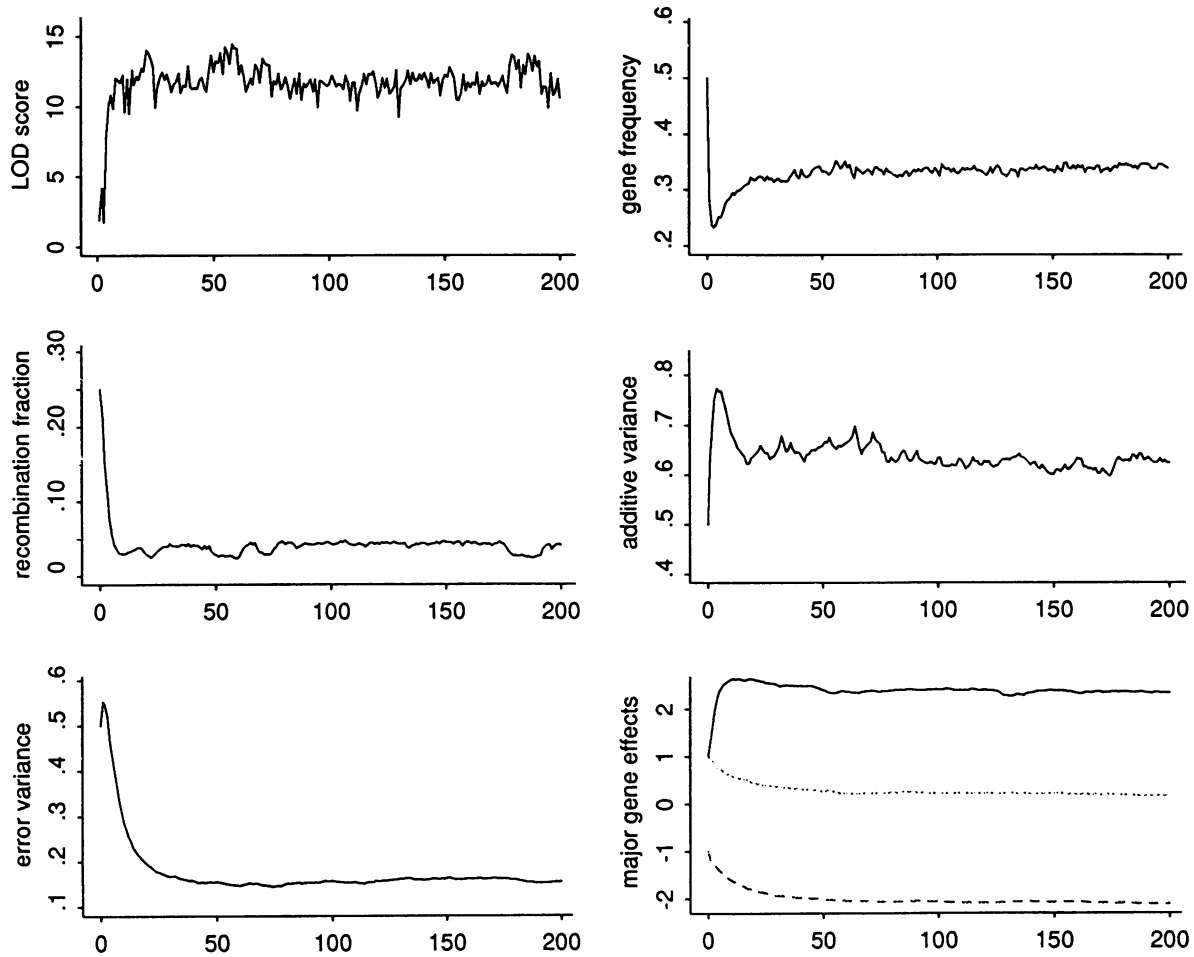

Figure 2 Results for the combined segregation and linkage analysis of the simulated data, plotted against the EM iteration number. The variables plotted are the estimated LOD score, the estimate of gene frequency, the estimate of recombination fraction, the estimate of additive polygenic variance, the estimate of error variance, and the estimate of major-gene effects.

Example 2: Hypercholesterolemia and the LDL Receptor Gene

We reanalyze the data on LDL cholesterol levels and LDL receptor genotypes in a 60-member, five-generation pedigree (Leppert et al. 1986). The pedigree is shown in figure 4. This data set has been extensively studied by several workers; the following analysis is presented to illustrate the methods of this paper, not to draw conclusions about the genetic mechanisms of the disease.

Using the Pedigree Analysis Package (PAP) (Hasstedt 1982), Leppert et al. (1986) carried out a segregation analysis under the assumption of a mixed model. Then they performed a linkage analysis using the parameters obtained from the segregation analysis. They found a maximum LOD score of 7.52 at $\hat{r} = 0$. Using the published data, Bonney et al. (1988) performed a combined segregation and linkage analysis using a regressive model. For ease of computation, they excluded individuals 50 and 59. To avoid bias in the estimate of gene frequency, they assumed a fixed frequency of .001 at the major locus. Also, they assumed a dominant major gene leading to elevated levels of LDL cholesterol. They found a maximum LOD score of 5.9 at $\hat{r} = 0$. Thomas and Cortessis (1991) dichotomized the LDL cholesterol levels and used a Bayesian method to perform a combined segregation and linkage analysis, with no ascertainment correction and assuming a dominant trait. Under various priors, they found that the gene frequency range was .065–

Table 3

Estimated Variance-Covariance Matrix for Simulated Data

p	μ_1	μ_2	μ_3	σ_a^2	σ_e^2	r
2,216.5						
−1,794.3	47,940					
−2,863.1	17,731	32,591				
−2,506.2	9,524.0	25,539	33,578			
−546.84	−3,768.8	5,102.5	−8,386.8	24,253		
−6.6661	2,442.7	−846.45	−1,834.6	−8,173.0	4,817.4	
5.8779	165.69	−39.165	−246.80	−343.39	−27.243	1,053.6

NOTE.—The actual value of each element in the matrix is the shown value times a factor of 10^{-6}.

Table 4

Estimated Parameters, with Their SEs and 95% Confidence Intervals, for the Hypercholesterolemia Data

Parameter	Estimate	SE[a]	95% Confidence Interval
p	.3266	.1066	(.118, .536)
μ_1	378.880	27.133	(325.700, 432.061)
μ_2	157.220	21.851	(114.392, 200.049)
μ_3	94.980	21.106	(53.613, 136.348)
σ_a^2	862.150	847.456	(0, 2,523.163)
σ_e^2	2,933.538	1,122.495	(733.449, 5,133.627)

[a] Since the estimate of the recombination frequency r is 0, estimated SEs cannot be obtained from the estimated information matrix.

.257 and that the range of the posterior mean of recombination fraction was .076–.318.

We performed a combined segregation and linkage analysis using the methods of this paper. Since individual 7 is unobserved and does not have offspring—and thus contributes no information—she was excluded from the analysis. Also, it is evident that the genotypes of individuals 8, 18, and 23 can be inferred from the existing data. Following Leppert et al. (1986), we used model (1) but made no ascertainment correction or any assumption of dominance. Starting values $p = .4$, $\sigma_a^2 = 718.0$, $\sigma_e^2 = 3{,}797.0$, $\mu_1 = 375.6$, $\mu_2 = 139.7$, and $\mu_3 = 95.3$ were obtained by a Monte Carlo EM of the mixed model without marker data. Then we performed Monte Carlo EM for 200 iterations. At each EM iteration a sample of 400 Gibbs realizations were drawn, with 10 cycles between each sampling. For the last 10 iterations, 1,000 Gibbs realizations were sampled, with 20 cycles between each sampling. On the basis of the final estimates, 12,000 realizations were drawn, with 20 cycles between two consecutive samplings. The final estimates, with their SEs and 95% confidence intervals, are listed in table 4. The estimated maximum LOD score is 7.13 (fig. 5). The parameter estimates against the iterations are shown in figure 6.

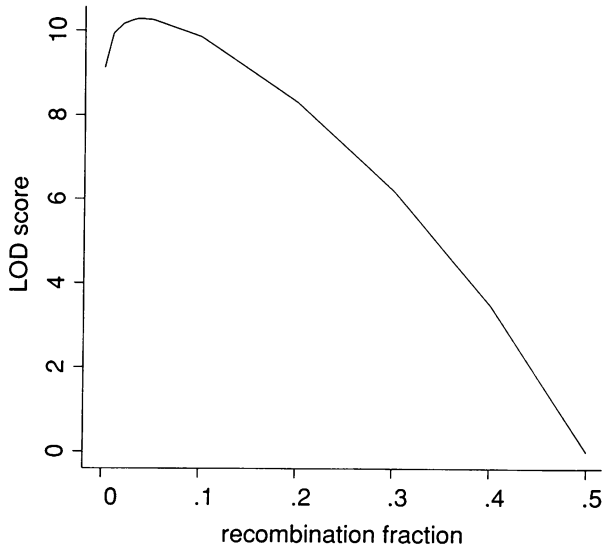

Figure 3 Monte Carlo estimate of the LOD score curve for the simulated data, at final estimates of other genetic parameters.

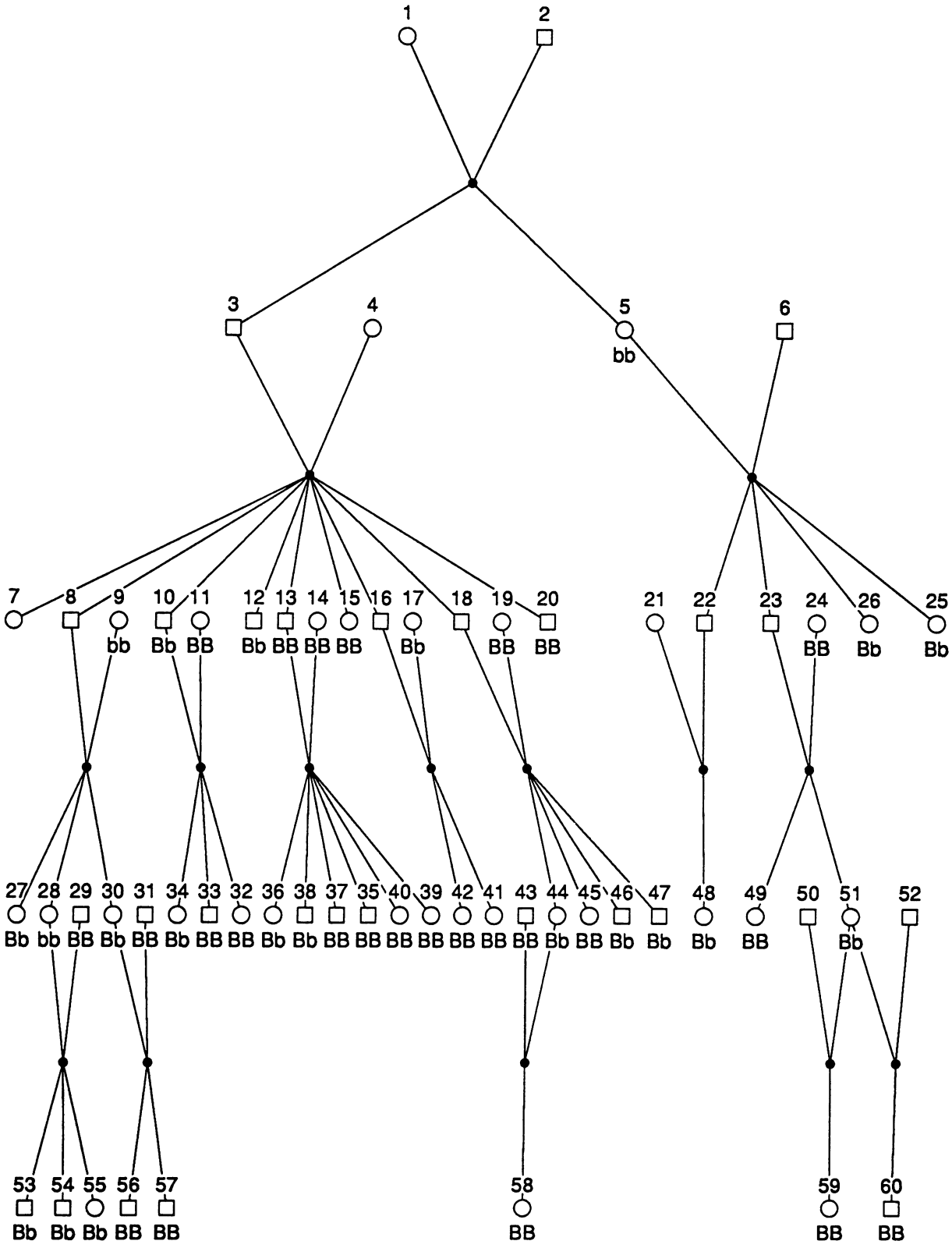

Figure 4 Pedigree used for linkage analysis of hypercholesterolemia and the LDL-receptor gene. The LDL-receptor RFLP genotypes are listed below the symbol for each sampled individual. (From Leppert et al. 1986.)

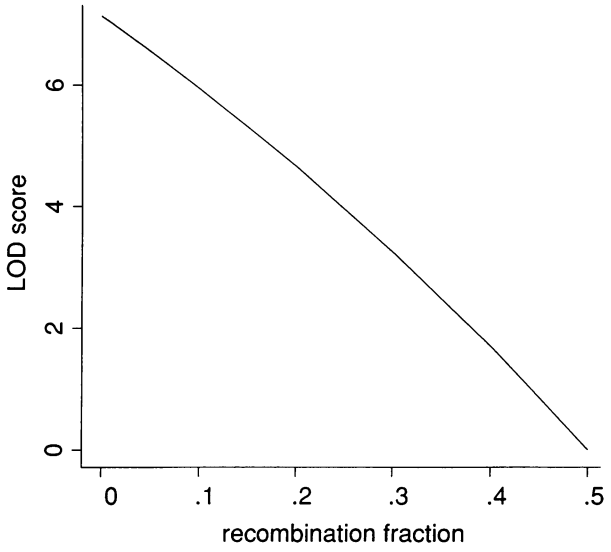

Figure 5 Monte Carlo estimate of LOD score curve for the hypercholesterolemia and LDL data.

The estimate of gene frequency has two shortcomings. First, since no ascertainment correction was made, it cannot be viewed as representing the frequency in the population at large. Second, because of the small number of founders ($n_F = 16$), the estimate has low precision and a wide confidence interval. Generally, since the pedigree size is small ($n = 59$; with 12 individuals not observed, $k = 47$), the information in these data is not great; the likelihood surface is rather flat. Although the presence of the major gene is clear, the magnitudes of the major-gene effects have wide confidence intervals. As usual, the estimates of additive and error variances are even less precise. In fact, the wide confidence intervals for σ_a^2 means that for these data there is no clear evidence of any polygenic effect (table 5).

The results of our analysis of these data are (not surprisingly) consistent with those of previous authors. The current approach provides MLEs of all the

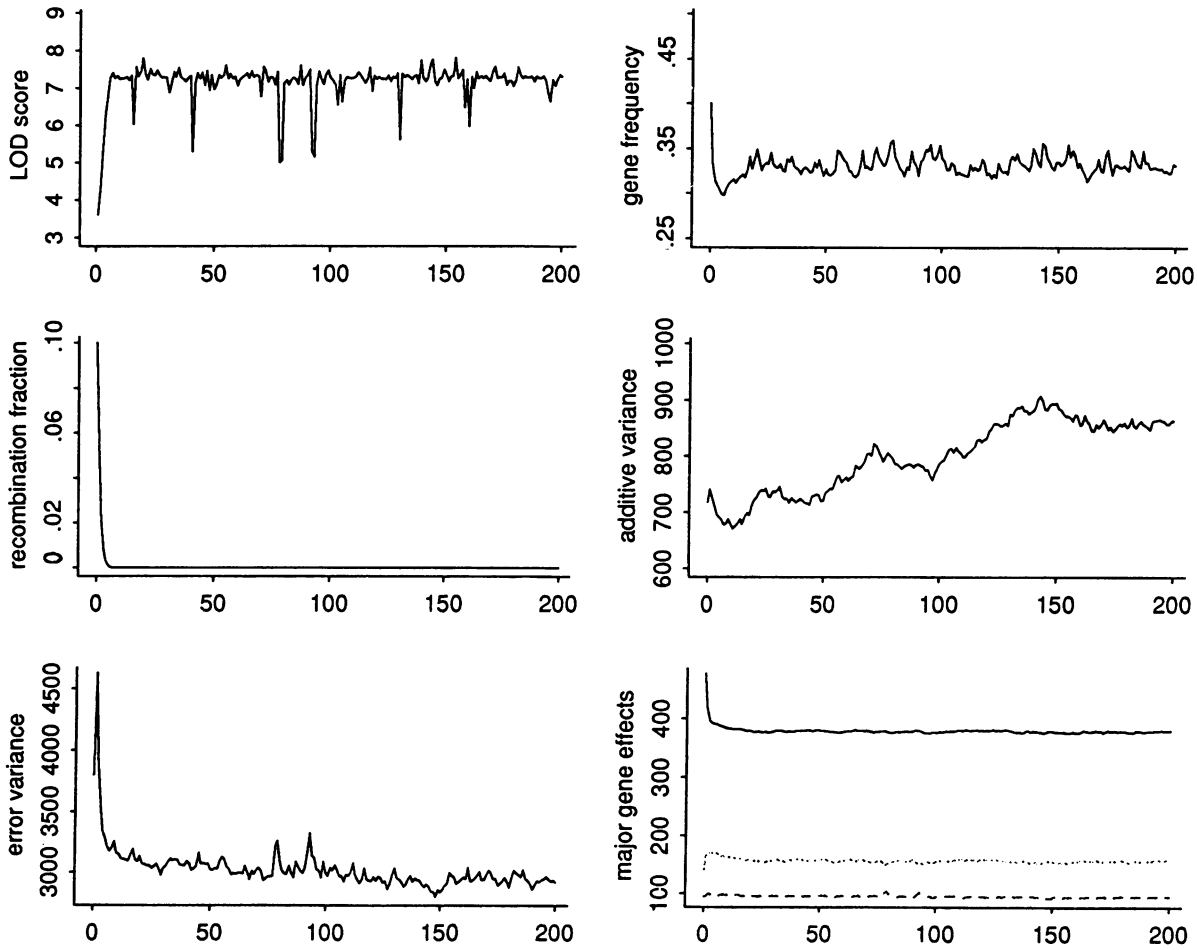

Figure 6 Results for the combined segregation and linkage analysis of the LDL data plotted against EM iteration number. The variables plotted are as in fig. 2.

Table 5

Estimated Variance-Covariance Matrix for Hypercholesterolemia Data

p	μ_1	μ_2	μ_3	σ_a^2	σ_e^2
1.136×10^2					
-6.159×10^1	7.362×10^2				
-7.725×10^1	3.006×10^2	4.775×10^2			
-9.013×10^2	4.157×10^1	6.232×10^1	4.455×10^2		
7.067×10^0	-2.787×10^2	-4.352×10^3	-5.141×10^3	7.182×10^5	
-1.234×10^1	1.740×10^3	6.062×10^3	7.084×10^3	-5.986×10^5	1.260×10^6

parameters in the model, together with SEs or other measures of precision. The procedure for likelihood-ratio evaluation provides a LOD score curve and also permits exploration of the multiparameter likelihood surface.

Discussion

Almost every function in human biology exhibits continuous variation. In human physiology, aspects such as susceptibility or resistance to common diseases such as diabetes and hypertension, predisposition to cancer, and drug and alcohol sensitivity can be measured as quantitative traits. Most complex behavioral and psychological traits are not simply present or absent like many Mendelian characters. Rather, they very continuously and are thus quantitative. On the other hand, a qualitative trait could be made quantitative by focusing both on the age at onset for affected individuals and on the right-censored age at last examination for unaffected individuals (Guo 1991). In light of this, mapping of quantitative trait loci is of great importance. Our approach provides a practical Monte Carlo approach to combined segregation and linkage analysis. The greatest attraction of this approach is its ability to handle complex genetic models and large pedigrees. It is also conceptually simple, numerically stable, and computationally feasible. In a number of examples, this approach works quite well. Using a Monte Carlo EM approach, combined segregation and linkage analysis does not substantially increase the computing time, compared with segregation analysis alone.

Because of the formidable computational problems of complex segregation analysis and increasing computing power, there is greatly increased interest in employing Monte Carlo methods in pedigree analysis. Ott (1989), Ploughman and Boehnke (1989), and Kong et al. (1991) have independently proposed Monte Carlo methods for sampling from the pedigree genotype distribution conditioned on the trait phenotypes observed in the pedigree. Unlike the current approach, those methods require exact probability computations at the trait locus in order to simulate data at a linked marker. Thus they are not feasible for complex models or complex pedigrees. Closer to the present paper is the work of Lange and Matthysse (1989) and Lange and Sobel (1991), who proposed using a Metropolis algorithm to calculate LOD scores and location scores. Similarly, Thomas and Cortessis (1991) propose a Monte Carlo method for two-point linkage analysis that combines a Bayesian perspective and the Gibbs sampler. Thomas and Cortessis' Bayesian method has its appeal in that it builds the prior information on parameters into the analysis. However, there is no consensus on choice of the prior for a complex trait; we prefer exploring the likelihood surface by estimating the likelihood ratios via the Gibbs sampler. The methods of Lange and Matthysse (1989) and Lange and Sobel (1991) are only useful when penetrance functions are known and when LOD or location scores are desired. More generally, all previous methods have been restricted to relatively simple genetic models for the trait. By contrast, the Monte Carlo EM approach permits both estimation of the parameters of complex models, in conjunction with linkage analysis, and exploration of a multiparameter likelihood surface.

EM algorithms have a tendency to converge slowly, and a definitive stopping rule for Monte Carlo EM is no less problematic than it is for deterministic EM. Although larger Monte Carlo samples can be taken as parameter values stabilize, to reduce Monte Carlo error in the estimation of conditional expectations, there are obvious limits to the practicality of this. Overall, we have found that Monte Carlo EM, even with fairly small Monte Carlo samples such as are

described in the Results section, is no slower than a deterministic EM algorithm, in terms of the number of EM iterations required, and is as robust and effective at finding a ballpark estimate. In the examples in the previous section, retrospective examination shows that estimates that are just as good could have been obtained with only 30 or 50 EM iterations; the choice of 200 was designed partly to verify that fact. The procedure for likelihood-ratio evaluation can always be used both to explore the likelihood in the neighborhood of any final estimates and, if it proves necessary, to adjust the final MLEs of parameter values. Also, unlike many cases where deterministic EM is used, the Monte Carlo EM procedures always permit estimation of first and second derivatives of the likelihood surface at any putative final estimates.

The efficiency and validity of alternative methods of the Markov-chain Monte Carlo method is currently an active research area in the statistical literature (Gelfand and Smith 1990; Tierney 1991). For validity, the technical requirement is that of irreducibility (hence ergodicity) of the Markov chain. For the Gibbs sampler employed in the present paper, as also in Thomas and Cortessis (1991), irreducibility is only assured for a diallelic marker locus. Lange and Sobel (1991) point to the same requirement for their Metropolis algorithm. However, irreducibility is not the main barrier in practice. Depending on the marker phenotypes observed in the pedigree, it may in fact obtain for multiallelic markers. Further, it can always be assured by modification of the sampling procedure; one such modification is the rejection sampling method proposed by Sheehan and Thomas (in press).

The greater practical problem is computational efficiency of alternative ergodic sampling schemes (Geyer 1991). If the likelihood has several local maxima that are separated by deep valleys, such as arise in the analysis of location scores, the Metropolis Markov chain can be very slowly mixing or "sticky" (Lange and Sobel 1991). The same is true of the Gibbs sampler when it is used to analyze ancestry of lethal rare recessives in a large complex pedigree (Thompson 1991), for a similar reason. For a quantitative trait, the counterpart "penetrance functions" are normal densities and are thus always positive. This more flexible genotype-phenotype correspondence results in faster mixing for the Markov chain of trait genotypic configurations. However, marker information, with not all individuals observed and/or with tight linkage, is likely to create problems of slow mixing. The occurrence of multiple alleles at marker loci—and the consequent necessity of using rejection sampling or some other method to ensure ergodicity—can only increase these problems. Computational efficiency is an important issue that warrants further investigation.

Finally, it should be pointed out that the approach of the present paper is not limited to the mapping of the single quantitative trait in the framework of mixed models. The same approach can be applied to a variety of gene-mapping problems, such as (*a*) power calculation for linkage analysis and (*b*) combined segregation and linkage analysis for multivariate traits. It can be developed to incorporate genetic heterogeneity among different pedigrees and to handle multiple trait and marker loci. It opens up new ways to tackle complicated genetic models for which analytical methods are often lacking.

Acknowledgments

The authors wish to thank Dr. Charles Geyer for fruitful discussions and for sharing his expert knowledge of computing, Dr. Ellen Wijsman for helpful discussions and comments, Dr. Nuala Sheehan for providing her pedigree neighborhood programs, and two referees for many helpful suggestions and detailed comments. This paper is based on research completed while S.W.G. was a student in the Department of Biostatistics, University of Washington. The research was supported in part by NIH grants NHLBI HL3 0086 (S.W.G.) and P30-HG00209 (S.W.G.) and GM-46255 (E.A.T.).

References

Bonney GE (1984) On the statistical determination of major gene mechanisms in continuous human traits: regressive models. Am J Med Genet 18:731–749

Bonney GE, Lathrop GM, Lalouel J-M (1988) Combined linkage and segregation analysis using regressive models. Am J Hum Genet 43:29–37

Breiman L (1968) Probability. Addison-Wesley, Reading, MA

Eaves LJ (1983) Errors of inference in the detection of major gene effects on psychological test scores. Am J Hum Genet 35:1179–1189

Elston RC (1984) Genetic Analysis Workshop II: sib pair screening tests for linkage. Genet Epidemiol 1:175–178

Elston RC, MacCluer JW, Hodge SE, Spence MA, King RH (1989) Genetic Analysis Workshop 6: linkage analysis based on affected pedigree members. In: Elston RC, Spence MA, Hodge SE, MacCluer JW (eds) Multipoint mapping and linkage based on affected pedigree members. Alan R Liss, New York, pp 93–103

Gelfand AE, Smith AFM (1990) Sampling based approaches

to calculating marginal densities. J Am Stat Assoc 85: 398–409

Geman S, Geman D (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. IEEE Trans Pattern Anal Machine Intelligence 6:721–741

Geyer CJ (1988) Software for calculating gene survival and multigene descent probabilities and for pedigree manipulation and drawing. Tech rep 153, Department of Statistics, University of Washington, Seattle

——— (1991) Markov chain Monte Carlo maximum likelihood. In: Keramidas EM (ed) Computer science and statistics: proceedings of the 23d Symposium on the Interface. Interface Foundation of North America, Fairfax Station, VA, pp 156–163

Go RCP, Elston RC, Kaplan EB (1978) Efficiency and robustness of pedigree segregation analysis. Am J Hum Genet 30:28–37

Guo SW (1991) Monte Carlo methods in quantitative genetics. Ph.D. diss. Department of Biostatistics, University of Washington, Seattle

Guo SW, Thompson EA (1991) Monte Carlo estimation of variance component models. IMA J Math Appl Med Biol 8:171–189

———. Monte Carlo estimation of mixed models for large complex pedigrees (submitted)

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3–19

Hasstedt SJ (1982) A mixed-model likelihood approximation on large pedigrees. Comput Biomed Res 15:295–307

Henderson CR (1976) A simple method for computing the inverse of a numerator relationship matrix used in prediction of breeding values. Biometrics 32:69–83

Kong A, Frigge M, Cox N, Wong WH (1991) Linkage analysis with adjustment for covariates: a method combining peeling with Gibbs sampling. In: MacCluer JW, Chakravarti A, Cox D, Bishop DT, Bale SJ, Skolnick MH (eds) Genetic Analysis Workshop 7: Issues in Gene Mapping and Detection of Major Genes. Cytogenetics and Cell Genetics Supplement. Karger, Basel

Konigsberg LW, Kammerer CM, MacCluer JW (1989) Segregation analysis of quantitative traits in nuclear families: comparison of three program packages. Genet Epidemiol 6:713–726

Lalouel J-M, Morton NE (1981) Complex segregation analysis with pointers. Hum Hered 31:312–321

Lange K, Matthysse S (1989) Simulation of pedigree genotypes by random walks. Am J Hum Genet 45:959–970

Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet 49:1320–1334

Leppert MF, Hasstedt SJ, Holm T, O'Connell P, Wu L, Ash O, Williams RR, et al (1986) A DNA probe for the LDL receptor gene is tightly linked to hypercholesterolemia in a pedigree with early coronary disease. Am J Hum Genet 39:300–306

Louis TA (1982) Finding the observed information matrix using the EM algorithm. J R Stat Soc [B] 44:226–233

MacLean CJ, Morton NE, Lew R (1975) Analysis of family resemblance. IV. Operational characteristics of segregation analysis. Hum Genet 27:365–384

Morton NE, MacLean CJ (1974) Analysis of family resemblance. III. Complex segregation analysis of quantitative traits. Am J Hum Genet 26:489–503

Ott J (1974) Estimation of the recombination fraction in human pedigrees: efficient computation of the likelihood for human linkage studies. Am J Hum Genet 26:588–597

——— (1979) Maximum likelihood estimation by counting methods under polygenic and mixed models in human pedigrees. Am J Hum Genet 31:161–175

——— (1989) Computer simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175–4178

——— (1990) Cutting a Gordian knot in the linkage analysis of complex human traits. Am J Hum Genet 46:219–221

——— (1991) Analysis of human genetics linkage, rev ed. The Johns Hopkins University Press, Baltimore

Ploughman LM, Boehnke M (1989) Estimating the power of a proposed linkage study for a complex genetic trait. Am J Hum Genet 44:543–551

Risch N (1984) Segregation analysis incorporating linkage markers. I. Single-locus models with an application to type I diabetes. Am J Hum Genet 36:363–386

Sheehan NA, Thomas AW. On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. Biometrics (in press)

Sundberg R (1974) Maximum likelihood theory for incomplete data from an exponential family. Scand J Stat 1:49–58

Thomas DC, Cortessis V (1991) A Monte Carlo Bayesian approach to genetic linkage analysis. Tech rep 13, Department of Preventive Medicine, University of Southern California, Los Angeles

Thompson EA (1991) Probabilities on complex pedigrees: the Gibbs sampler approach. In: Keramidas EM (ed) Computer science and statistics: proceedings of the 23d Symposium on the Interface. Interface Foundation of North America, Fairfax Station, VA, pp 321–328

Thompson EA, Guo SW (1991) Evaluation of likelihood ratios for complex genetic models. IMA J of Math Appl Med Biol 8:149–169

Thompson EA, Shaw RG (1990) Pedigree analysis for quantitative traits: variance components without matrix inversion. Biometrics 46:399–413

Tierney L (1991) Markov chains for exploring posterior distributions. Tech rep 560, School of Statistics, University of Minnesota at Minneapolis St Paul, Minneapolis