Statistical Science
2003, Vol. 18, No. 4, 515–531

Discovering Disease Genes: Multipoint Linkage Analysis via a New Markov Chain Monte Carlo Approach

A. W. George and E. A. Thompson

Abstract. Multipoint linkage analyses of data collected on related individuals are often performed as a first step in the discovery of disease genes. Through the dependence in inheritance of genes segregating at several linked loci, multipoint linkage analysis detects and localizes chromosomal regions (called trait loci) which contain disease genes. Our ability to correctly detect and position these trait loci is increased with the analysis of data observed on large pedigrees and multiple genetic markers. However, large pedigrees generally contain substantial missing data and exact calculation of the required multipoint likelihoods quickly becomes intractable. In this paper, we present a new Markov chain Monte Carlo approach to multipoint linkage analysis which greatly extends the range of models and data sets for which analysis is practical. Several advances in Markov chain Monte Carlo theory, namely joint updates of latent variables across loci or meioses, integrated proposals, Metropolis–Hastings restarts via sequential imputation and Rao–Blackwellized estimators, are incorporated into a sampling strategy which mixes well and produces accurate results in real time. The methodology is demonstrated through its application to several data sets originating from a study of early-onset Alzheimer's disease in families of Volga–German ethnic origin.

Key words and phrases: Linkage analysis, joint Gibbs updates, integrated proposals, Metropolis–Hastings restarts, sequential imputation.

1. INTRODUCTION

Within the nucleus of every human cell are 46 chromosomes, long threadlike structures of double-stranded DNA. Organized into homologous pairs (chromosomes of almost identical DNA material), one chromosome is inherited from the father and one chromosome is inherited from the mother. Chromosomes are passed from parent to offspring via a biological process called meiosis. It is during meiosis that gamete cells (sperm and egg) are produced which later combine to form a new offspring. In forming the paternal (maternal) gamete, the father's (mother's) chromosome pair exchange complementary segments of DNA. That is, the DNA of a gamete chromosome switches from being a copy of the parent's paternal (maternal) chromosome to being a copy of the parent's maternal (paternal) chromosome. The exchange points are known as crossovers.

At a single position on a chromosome pair, known as a locus (a very small segment of the DNA), one of several variant DNA types (alleles) may be present. Since chromosomes come in pairs, it follows that there are two alleles at every locus where the unordered pair of alleles represents an individual's genotype at this locus. The segment of DNA passed from parent to offspring is called a gene. An individual may have two copies of the same allele at a locus (e.g., aa), but

A. W. George is Assistant Professor, Department of Biostatistics, University of Iowa, 2190 Westlawn Building, Iowa City, Iowa 52242 (e-mail: andrew-george@uiowa.edu). E. A. Thompson is Professor, Department of Statistics, University of Washington, Padelford C-317, Seattle, Washington 98195.

one allele is associated with a paternally inherited gene and one allele is associated with a maternally inherited gene. Genes segregate (are inherited) according to Mendel's probabilistic rules (Mendel, 1866):

- At a locus, each individual has a gene inherited from the father and a gene inherited from the mother.
- The paternally (maternally) inherited gene is a copy of a randomly chosen one of the father's (mother's) two genes.
- The random choice of genes passed from different parents to a child and from a parent to different children is independent.

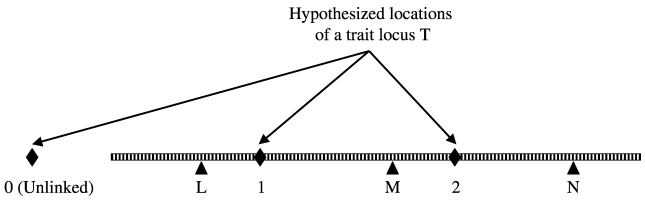
When two or more loci are being considered, the chromosomal locations of the loci become important. If two loci, M and N, are locations on different chromosomes, the loci are said to be unlinked and the genes segregate independently. However, if M and N are locations on the same chromosome, the loci are linked and the genes segregate dependently. The closer two loci are, the stronger the dependence.

A recombination event occurs between two loci if the paternal (maternal) chromosomes of a child at locus M and locus N originate from different parental chromosomes of the father (mother). It is through the recombination frequency that the strength of dependence between two loci is measured. The recombination frequency between loci M and N, denoted by ρ_{MN}, takes on values between 0 and 0.5, where $\rho_{MN} = 0.5$ implies the loci are unlinked and $\rho_{MN} = 0$ implies the loci are completely linked. At unlinked loci the genes are independently inherited.

In a linkage analysis we seek to estimate the recombination fraction between a trait locus of unknown location and genetic markers of known location and to test whether this recombination fraction is significantly different from 0.5 (the trait locus is unlinked to the genetic markers). In this way we are able to detect and localize unknown trait loci. If the observed data are such that the parental origin of the underlying genes can be determined unambiguously, the likelihood used in a linkage analysis is a simple multinomial. However, data collected on large human pedigrees are often sparsely observed with many individuals unavailable for sampling. Thus, calculating probabilities on extended pedigrees is a latent variable or missing data problem.

Traditionally (Elston and Stewart, 1971) multilocus genotypes were used as the latent variables in calculating probabilities of genetic data observed on related individuals. Multilocus genotypes are the set of allele pairs at multiple linked loci where we know the phase. That is, we know which allele in each pair belongs to the paternal chromosome and which allele belongs to the maternal chromosome. Alternative latent variables are meiosis indicators, binary variables which specify the grandparental origins of alleles and hence trace the passage through a pedigree of identical-by-descent (ibd) genes (sometimes called "founder alleles"). Two genes are ibd if they originate from a common founder (an individual for whom we have no parental information) in the pedigree. We will give an example of meiosis indicators and how they specify the passage of ibd genes in Section 2.

Multipoint (use of several linked markers) linkage analysis involves likelihood computations for numerous hypothesized locations of a trait locus on a fixed marker map. In this paper these trait locus locations are discrete and denoted by λ, $\lambda \in \{0, 1, \ldots, K\}$, with $\lambda = 0$ denoting that the trait locus is unlinked, and $1, \ldots, K$ are fixed locations within the marker map. Each λ corresponds to a different set of recombination fractions between a trait locus and genetic markers. For example, suppose we wish to calculate the likelihood for three hypothesized locations of a trait locus given data on three linked genetic markers. Then the sets of recombination fractions associated with each test position are given in Figure 1.



λ	Associated Recombination Frequencies	Relationships
0	ρ_{TL}, ρ_{LM}, ρ_{MN}	$\rho_{TL} = 0.5$
1	ρ_{LT}, ρ_{TM}, ρ_{MN}	$\rho_{LM} = \rho_{LT} + \rho_{TM} - 2\rho_{LT}\rho_{TM}$
2	ρ_{LM}, ρ_{MT}, ρ_{TN}	$\rho_{MN} = \rho_{MT} + \rho_{TN} - 2\rho_{MT}\rho_{TN}$

FIG. 1. *The sets of recombination fractions associated with three different hypothesized locations of a trait locus T with respect to linked genetic markers L, M and N. The hypothesized locations of T are indexed by* 0, 1 *and* 2, *where $\lambda = 0$ is an unlinked position* (*the trait locus is on another chromosome*). *The triangles denote genetic markers, the diamonds denote hypothesized locations of T and ρ_{ij} denotes the recombination fraction between locus i and locus j. In linkage analyses the recombination fraction between marker loci L and M (ρ_{LM}) and between marker loci M and N (ρ_{MN}) is assumed known. We also show the relationships that exist between the recombination fractions when we assume recombinations occur independently across loci intervals.*

The multipoint likelihood is the probability of the observed data given a trait locus at hypothesized position λ, under an assumed genetic model θ. The observed data include information on some individuals for a trait (e.g., disease status) and/or several linked markers (e.g., marker allele types). It is through the genetic model that we specify the relationship between observed data and underlying latent variables, the chromosomal positions of the marker loci and the frequency in the population of the alleles at each locus. Once the $K + 1$ multipoint likelihoods have been computed, lod scores are formed.

The lod score

$$\text{lod}(x) = \log_{10}\left[\frac{P_\theta(\mathbf{Y}|\lambda = x)}{P_\theta(\mathbf{Y}|\lambda = 0)}\right] \tag{1}$$

assesses the support given by observed data $\mathbf{Y}$ for a trait locus in hypothesized trait location x ($H_1 : \lambda = x$) versus an unlinked trait locus ($H_0 : \lambda = 0$). A lod score greater than 3 suggests that there is significant evidence for linkage with the location of the trait locus given by the location of the peak lod score (Morton, 1955).

Two algorithms exist for the exact calculation of multipoint likelihoods: the Lander–Green algorithm and pedigree-peeling algorithms. The Lander–Green (Lander and Green, 1987) algorithm uses meiosis indicators as the latent variables, and exact calculation proceeds almost directly from the Baum algorithm (Baum, 1972). The Baum algorithm is a deterministic procedure for computing the likelihood of a hidden Markov model with discrete-valued latent states. Through the use of a recurrence relation, the calculation of the multipoint likelihood decomposes into a series of sums over the possible latent values of a single locus. Pedigree-peeling algorithms (Elston and Stewart, 1971; Cannings, Thompson and Skolnick, 1978) were originally formulated using multilocus genotypes as the latent variables but can also be implemented using meiosis indicators (Thompson, 2000a). Exact calculation proceeds from a generalization of the Baum algorithm. Lander–Green computations are linear in marker number but exponential in pedigree size. Conversely, pedigree-peeling computations are linear in pedigree size but exponential in marker number, and are also confounded by excessive pedigree complexity. Many multipoint linkage analyses extend beyond the computational boundaries of exact methods.

An attractive alternative to exact computation is Monte Carlo estimation of likelihoods via Markov chain Monte Carlo (MCMC) methods. The first MCMC method for Monte Carlo estimation of multipoint linkage likelihoods was that of Lange and Sobel (1991). Early MCMC methods sampled genotypes (Thompson and Guo, 1991) or meiosis indicators (Thompson, 1994a, b) via single-site updates. Although these methods are easy to implement, ensuring (practical) irreducibility of the realized Markov chain is problematic since the missing data are highly constrained by the observed data and laws of Mendelian segregation. More recently, several innovative schemes for jointly updating blocks of genotypes (Jensen, Kjærulff and Kong, 1995; Jensen and Kong, 1999) or meiosis indicators (Heath, 1997; Thompson and Heath, 1999) have culminated in MCMC methods with improved mixing properties. Generally, meiosis indicators result in a smaller and less constrained latent space for MCMC sampling given data at multiallelic marker loci on sparsely sampled extended pedigrees.

Multipoint linkage analysis can also be conducted within a Bayesian framework (Satagopan, Yandell, Newton and Osborn, 1996; Uimari and Hoeschele, 1997; Daw, Heath and Wijsman, 1999). In this case, not only the latent variables are sampled by MCMC but also the location of the trait locus together with other unfixed parameters of the trait or marker genetic model or map. A Bayesian posterior distribution integrated over these parameters is thus obtained. Although increasingly used as a tool in the analysis of complex traits, a Bayesian posterior distribution alone is often regarded as insufficient, since the integrated posterior lacks the familiarity and interpretability of a linkage likelihood.

In this paper, an MCMC sampler is presented which further advances the Monte Carlo estimation of multipoint likelihoods of genetic data on extended pedigrees. Here we focus on a binary trait, but the methodology can be extended to include ordinal and quantitative traits. The sampler explores the constrained space of latent variables through Metropolis–Hastings (M–H) restarts and joint updates which combine exact single-locus pedigree-peeling calculations with Monte Carlo sampling. The problem is formulated within a pseudo-Bayesian framework where a priori priors no longer mirror belief but are chosen to enhance the performance characteristics of the MCMC method. Hypothesized trait locations λ are sampled from the joint posterior distribution via an M–H acceptance ratio integrated over the latent variables for the trait locus. A Rao–Blackwellized estimator is presented for the Monte Carlo estimation of likelihoods.

2. MULTIPOINT LIKELIHOODS

Suppose observed data are recorded on a trait $Y_T = (Y_{\cdot 0})$ and L linked markers $\mathbf{Y}_M = (Y_{\cdot 1}, Y_{\cdot 2}, \ldots, Y_{\cdot L})$,

where marker loci are ordered along a chromosome $1, 2, \ldots, L$ and $Y_{\cdot j}$ denotes observed data at locus j across (related and unrelated) individuals. Some individuals may provide no observed data, and observed individuals may not have data at all loci: $Y_{\cdot j}$ simply denotes whatever data are available at locus j. Under a known genetic model θ, the multipoint likelihood of λ based on the observed data $\mathbf{Y} = (Y_T, \mathbf{Y}_M)$ is

$$L(\lambda) = P_\theta(\mathbf{Y}|\lambda) = \sum_{\mathbf{S}} P_\theta(\mathbf{Y}, \mathbf{S}|\lambda)$$

$$= \sum_{\mathbf{S}} P_\theta(\mathbf{Y}|\mathbf{S}) P_\theta(\mathbf{S} \mid \lambda) \tag{2}$$

$$= \sum_{\mathbf{S}} \left(\prod_{j=0}^{L} P_\theta(Y_{\cdot j}|S_{\cdot j}) \right) \left(\prod_{i=1}^{m} P(S_{i\cdot}|\lambda) \right),$$

where $\mathbf{S}$ is the array of meiosis indicators S_{ij} ($i = 1, \ldots, m$, $j = 0, 1, \ldots, L$), S_{ij} is 0 or 1 as the ibd inherited gene at meiosis i locus j is the parent's maternal gene, or paternal gene, respectively, $S_{\cdot j}$ is the vector of meiosis indicators at locus j across meioses, $S_{i\cdot}$ is the vector of meiosis indicators at meiosis i across loci and m is the total number of meioses in the pedigree. The latent variables $\mathbf{S}$ trace the unobserved passage of ibd genes through a pedigree.

For example, suppose data are collected on two linked marker loci from 10 related individuals. Eight individuals are observed and two individuals are unobserved. For illustrative purposes only, we assume the marker phase, whether the marker alleles at each locus reside on the paternal chromosome or the maternal chromosome of an individual, is known. Then Figure 2 shows how the binary meiosis indicators $\mathbf{S}$ can be used to trace unobserved ibd genes through a pedigree given the observed data. In fact, $\mathbf{S}$ given in Figure 2 is just one of a possible 1024 configurations consistent with the observed data. If the marker phase were not known, the number of possible $\mathbf{S}$ would be far greater. Here and throughout this paper, we assume founders are noninbred and unrelated. Each founder can pass one of two unique ibd genes to his or her offspring at each locus due to the founders being noninbred. Furthermore, the ibd genes at each locus across founders are unique due to founders being unrelated.

The single-locus probability $P_\theta(Y_{\cdot j}|S_{\cdot j})$ models the relationship between observed data at locus j and underlying latent variables at locus j. Calculation of $P_\theta(Y_{\cdot j}|S_{\cdot j})$ is described in Thompson (1974). Briefly, for genotypic observations,

$$P_\theta(Y_{\cdot j}|S_{\cdot j}) = \sum_{A(j)} \left(\prod_k q_j(a(k)) \right), \tag{3}$$

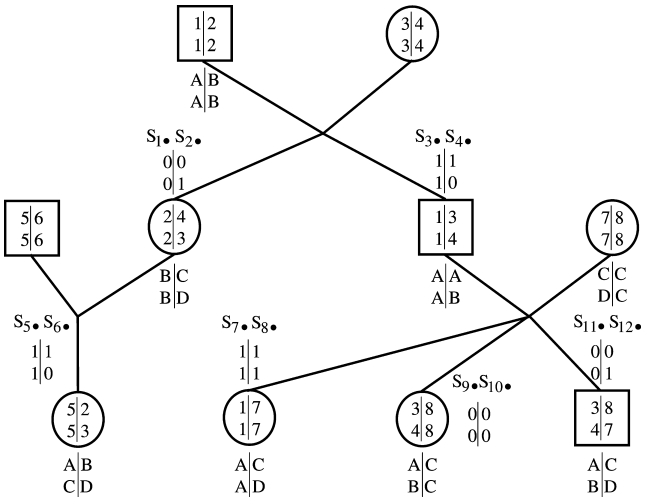


FIG. 2. *A simple pedigree with phase-known marker information recorded at two linked marker loci for eight individuals*; *all other individuals are unobserved. Founders are assumed to be noninbred and unrelated*; *therefore, at each locus each founder can pass one unique ibd gene to his or her offspring. The meiosis indicators* $S_{1\cdot}, \ldots, S_{2\cdot}$ *fully describe the underlying passage of ibd genes through the pedigree. At both loci* (*observed*) *marker alleles are denoted by* A, B, C, D *and ibd genes are denoted by* $1, \ldots, 8$. *The notation* $p|m$ *denotes information inherited from the paternal parent* (p) *and the maternal parent* (m).

where the summation is over all valid assignments $A(j)$ of allelic types at (a marker or trait) locus j to ibd genes, and the probability that a distinct gene k is of allelic type $a(k) \in A(j)$ is the population allele frequency $q_j(a(k))$. Kruglyak, Daly, Reeve-Daly and Lander (1996) present a computationally efficient algorithm for identifying all valid assignments. See Thompson (1974, 2000a) for detailed examples demonstrating the use of (3).

Assuming recombination events in disjoint intervals between loci are independent, the probability $P_\theta(S_{i\cdot}|\lambda)$ of the ith meiosis under θ when the trait locus is in location λ is then

$$P_\theta(S_{i\cdot}|\lambda) = \prod_{j=2}^{L+1} (1 - \rho_{\varphi(j-1;\lambda)})^{1-|\beta|} \cdot (\rho_{\varphi(j-1;\lambda)})^{|\beta|}, \tag{4}$$

where $\beta = (S_{i\varphi(j-1;\lambda)} - S_{i\varphi(j;\lambda)})$, $\varphi(j;\lambda)$ denotes the jth locus in the chromosomal ordering when the trait locus is in location λ and, for ease of notation, $\rho_{\varphi(j-1;\lambda)}$ now is the recombination frequency between loci $\varphi(j-1;\lambda)$ and $\varphi(j;\lambda)$. For example, suppose there are four marker loci with a hypothesized trait locus between markers 3 and 4, so that the locus ordering is 1 2 3 0 4. If $S_{i\cdot} = (1, 0, 1, 1, 0)$, then

this indicates that there are recombinations between markers 1 and 2, between 2 and 3 and between the trait locus and marker 4. Therefore, $P(S_i.|\lambda) = \rho_1\rho_2(1-\rho_3)\rho_4$, where ρ_1, ρ_2, ρ_3 and ρ_4 are the recombination frequencies between loci 1 and 2, 2 and 3, 3 and 0 and 0 and 4, respectively.

Note, for a given $\mathbf{S}$ calculation of the complete-data likelihood, $P_\theta(\mathbf{Y}, \mathbf{S}|\lambda)$, is straightforward using (3) and (4). This feature is fully exploited in MCMC methods. Difficulties surrounding the calculation of (2) stem from the sum over $\mathbf{S}$ because the set of valid $\mathbf{S}$ can be huge and is highly constrained. For the calculation of the lod score (1) it is assumed that only the location λ of the trait locus in (2) is unknown. The genetic model θ is fixed. A strategy for extending our analysis approach to unknown genetic models is discussed in Section 10.

3. A PSEUDO-BAYESIAN PARADIGM FOR LIKELIHOOD ESTIMATION

Previous Monte Carlo estimates of the likelihood curve for the location of a trait locus have been based on sampling latent variables $\mathbf{S}$ from their conditional distribution given $\mathbf{Y}$ under a fixed model (θ, λ). Thompson (1994b) samples from the full conditional distribution $P_{(\theta,\lambda)}(\mathbf{S}|\mathbf{Y})$ to obtain estimates of local likelihood ratios in the neighborhood of λ. Irwin, Cox and Kong (1994) use a sequential imputation distribution and reweight realizations to estimate the location likelihood curve. Lange and Sobel (1991) sample the values of meiosis indicators $\mathbf{S}_M$ at marker loci only, given marker data $\mathbf{Y}_M$, and then compute for each λ and each realization of $\mathbf{S}_M$ the contributions $P_{(\theta,\lambda)}(Y_T|\mathbf{S}_M)$ of trait data Y_T to the multipoint likelihood. Each of these methods has disadvantages in obtaining efficiently an accurate estimate of a multipoint lod score curve from data on an extended pedigree in which many individuals are unobserved.

An alternative approach has been to place a prior distribution on the parameters and genetic model, sampling from the joint posterior distribution $\pi(\mathbf{S}, \lambda, \theta|\mathbf{Y})$ (Heath, 1997; Lee and Thomas, 2000), but these Bayesian methods do not produce an estimate of the lod score curve. Here, since the space of unknown parameters is the discrete one-dimensional λ, it is possible to recover the likelihood from the marginal posterior distribution of λ when sampling from the joint posterior $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$ for the fixed genetic model θ. That is, if a prior distribution $\pi(\lambda)$ is assumed,

$$L(\lambda) = P_\theta(\mathbf{Y}|\lambda) \propto \frac{\pi_\theta(\lambda|\mathbf{Y})}{\pi(\lambda)}. \tag{5}$$

The prior distribution $\pi(\lambda)$ can be chosen arbitrarily, provided it assigns positive mass to each hypothesized trait location, and thus may be chosen to improve performance of the MCMC sampling from $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$ rather than to reflect prior beliefs about trait location. That is, $\pi(\lambda)$ is a pseudo-prior in the sense of Geyer and Thompson (1995). In practice, it is found to be most efficient if $\pi(\lambda)$ is chosen so that the marginal posterior distribution for λ is approximately uniform. The method of choosing such a $\pi(\lambda)$ is deferred to Section 7.

Given N realizations $(\mathbf{S}^{(1)}, \lambda^{(1)}), \ldots, (\mathbf{S}^{(N)}, \lambda^{(N)})$ from the posterior distribution $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$, an estimator of the lod score curve is now constructed. Note that

$$\pi_\theta(\lambda = x|\mathbf{Y}) = \mathrm{E}_{\pi_\theta}\big(I(\lambda = x)|\mathbf{Y}\big), \tag{6}$$

where $I(\lambda = x)$ is an indicator function equal to 1 when $\lambda = x$ and 0 otherwise. Thus, the marginal posterior probability $\pi_\theta(\lambda|\mathbf{Y})$ is most simply estimated by the proportion of realizations $(\lambda^{(n)}; n = 1, \ldots, N)$ that are equal to x.

Combining (5) and (6), the estimate of the marginal posterior for λ is normalized by the prior $\pi(\lambda)$ to obtain, up to a λ-independent constant of proportionality, an unbiased estimator of the likelihood $L(\lambda = x) = P_\theta(\mathbf{Y}|\lambda = x)$:

$$T_N^{\text{crude}}(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{I(\lambda_{(n)} = x)}{\pi(\lambda = x)}. \tag{7}$$

From (1) an estimate of the multipoint lod score is then

$$\log_{10}\left(\frac{T_N^{\text{crude}}(x)}{T_N^{\text{crude}}(0)}\right), \tag{8}$$

where $\lambda = 0$ corresponds to the trait being unlinked to the markers. This simplest estimator (8) disregards the sampled values of $\mathbf{S}$ and uses only the realized values of λ.

4. MCMC SAMPLING

MCMC methods are procedures for drawing dependent samples from high-dimensional probability distributions where the samples form a Markov chain with the distribution of interest as its stationary distribution. Expectations in high-dimensional distributions which preclude exact computation can then be estimated via ergodic averages of the realized samples. The procedures are outlined here: details of the samplers are given in the Appendix.

Dependent realizations $(\mathbf{S}^{(1)}, \lambda^{(1)}), \ldots, (\mathbf{S}^{(n)}, \lambda^{(n)}), \ldots, (\mathbf{S}^{(N)}, \lambda^{(N)})$ are drawn from $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$, where a move from $(\mathbf{S}^{(n)}, \lambda^{(n)})$ to $(\mathbf{S}^{(n+1)}, \lambda^{(n+1)})$ is accomplished via the following steps:

STEP I. Given $(\mathbf{S}^{(n)}, \lambda^{(n)})$, sample $\mathbf{S}^{(n+1)}$ via joint Gibbs steps:

- with probability $1 - p_L$, a whole-meiosis sampler (M-sampler) is used to realize $S_{i\cdot} : i = 1, \ldots, m$, where meioses are updated in random order;
- with probability p_L, a whole-locus sampler (L-sampler) is used to realize $S_{\cdot j} : j = 0, 1, \ldots, L$, where loci are updated in random order.

STEP II. Sample λ via the M–H algorithm using an integrated acceptance probability:

- a new hypothesized trait location λ' is randomly chosen from possible hypothesized trait locations $\{0, 1, \ldots, K\}$, which includes the unlinked hypothesized trait location;
- λ' is accepted with probability $\alpha(\lambda^{(n)}, \lambda')$ given by

$$\alpha(\lambda^{(n)}, \lambda') = \min\left[1, \frac{P_\theta(Y_T|\mathbf{S}_M^{(n+1)}, \lambda')\pi(\lambda')}{P_\theta(Y_T|\mathbf{S}_M^{(n+1)}, \lambda^{(n)})\pi(\lambda^{(n)})}\right],$$

  where each $P_\theta(Y_T|\mathbf{S}_M, \lambda)$ is obtained by single-locus peeling over the trait meiosis indicators S_T;
- if λ' is accepted, then $\lambda^{(n+1)} = \lambda'$ and a new $S_T^{(n+1)}$ is drawn from the full conditional distribution $P_\theta(S_T|\mathbf{S}_M^{(n+1)}, Y_T, \lambda')$; otherwise, $\lambda^{(n+1)} = \lambda^{(n)}$ and $S_T^{(n+1)}$ is unchanged.

These two steps are repeated N times. The M-sampler, L-sampler and integrated acceptance probabilities are described in the Appendix.

5. SEQUENTIAL IMPUTATION FOR STARTS AND RESTARTS

Initially intended as a Monte Carlo technique for estimating multipoint likelihoods, sequential imputation is a useful mechanism for obtaining a starting configuration $\mathbf{S}^{(0)}$ for the Markov chain. Kong, Cox, Frigge and Irwin (1993) and Kong, Liu and Wong (1994) describe sequential imputation as an importance sampling approach where K independent samples $\mathbf{S}^{*k}$ for $k = 1, \ldots, K$ are obtained with associated weights $W(\mathbf{S}^{*k})$ conditioned on the observed data. That is, each $\mathbf{S}^* = (S_{\cdot 0}^*, S_{\cdot 1}^*, \ldots, S_{\cdot L}^*)$ is to be drawn from the joint probability distribution $P^*(\mathbf{S}|\lambda^{(0)}) = P_\theta(\mathbf{Y}, \mathbf{S}|\lambda)(W(\mathbf{S}))^{-1}$, where for notational convenience the k superscript is suppressed. Here λ denotes the initial position of the trait locus normally taken as $\lambda = 0$ (unlinked).

First, the locus in some position h $(h = 1, \ldots, L+1)$ is selected, $S_{\cdot\varphi(h;\lambda)}^*$ is drawn from $P_\theta(S_{\cdot\varphi(h;\lambda)}|Y_{\cdot\varphi(h;\lambda)}, \lambda)$ and $w_h = P_\theta(Y_{\cdot\varphi(h;\lambda)})$ is computed. Second, moving in a forward direction along the chromosome [i.e., for the loci $\varphi(h+1;\lambda), \ldots, \varphi(L+1;\lambda)$ in positions $h+1, \ldots, L+1$], $S_{\cdot\varphi(j;\lambda)}^*$ is drawn from $P_\theta(S_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j-1;\lambda)}^*, Y_{\cdot\varphi(j;\lambda)}, \lambda)$ and $w_j = P_\theta(Y_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j-1;\lambda)}^*, \lambda)$ computed. Last, moving in a backward direction from the initial locus in position h [i.e., for the loci $\varphi(h-1;\lambda), \ldots, \varphi(1;\lambda)$ in positions $h-1, \ldots, 1$], $S_{\cdot\varphi(j;\lambda)}^*$ is drawn from $P_\theta(S_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j+1;\lambda)}^*, Y_{\cdot\varphi(j;\lambda)}, \lambda)$ and $w_j = P_\theta(Y_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j+1;\lambda)}^*, \lambda)$ computed. Therefore, by moving along the chromosome in a sequential manner, a joint sample $\mathbf{S}^*$ is obtained with associated importance weight $W(\mathbf{S}^*) = \prod_{j=1}^{L+1} w_j$.

Sampling $S_{\cdot\varphi(j;\lambda)}$ from $P_\theta(S_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j-1;\lambda)}^*, Y_{\cdot\varphi(j;\lambda)}, \lambda)$ is computationally analogous to an L-sampler step, and computing w_j is computationally equivalent to single-locus pedigree peeling. The $\mathbf{S}^{*k}$ with the largest associated weight among the K independent samples is used as a starting configuration $\mathbf{S}$ for the Markov chain.

Sequential imputation can also be incorporated into the M–H update, allowing the Markov chain to restart in a different part of the parameter space. With $(\mathbf{S}^{(n)}, \lambda^{(n)})$ representing the present state of the Markov chain, a single $\mathbf{S}'$ is drawn from $P^*(\mathbf{S}|\lambda^{(n)})$ via sequential imputation. The restart state $\mathbf{S}'$ is then accepted with probability $\alpha(\mathbf{S}^{(n)}, \mathbf{S}') = \min(1, A)$ with

$$\begin{aligned} A &= \frac{\pi_\theta(\mathbf{S}', \lambda^{(n)}|\mathbf{Y})}{\pi_\theta(\mathbf{S}^{(n)}, \lambda^{(n)}|\mathbf{Y})} \frac{P^*(\mathbf{S}^{(n)}|\lambda^{(n)})}{P^*(\mathbf{S}'|\lambda^{(n)})} \\ &= \frac{P_\theta(\mathbf{Y}|\mathbf{S}', \lambda^{(n)})P_\theta(\mathbf{S}'|\lambda^{(n)})\pi(\lambda^{(n)})}{P_\theta(\mathbf{Y}|\mathbf{S}^{(n)}, \lambda^{(n)})P_\theta(\mathbf{S}^{(n)}|\lambda^{(n)})\pi(\lambda^{(n)})} \\ &\quad \cdot \frac{P_\theta(\mathbf{Y}, \mathbf{S}^{(n)}|\lambda^{(n)})W(\mathbf{S}')}{P_\theta(\mathbf{Y}, \mathbf{S}'|\lambda^{(n)})W(\mathbf{S}^{(n)})} \\ &= \frac{W(\mathbf{S}')}{W(\mathbf{S}^{(n)})}, \end{aligned}$$

where $P^*(\mathbf{S}^{(n)}|\lambda^{(n)})$ is the probability of proposing $\mathbf{S}^{(n)}$ if $\mathbf{S}^{(n)}$ were sampled using sequential imputation, and similarly $P^*(\mathbf{S}'|\lambda^{(n)})$ is the probability of proposing $\mathbf{S}'$ when $\mathbf{S}'$ is sampled using sequential imputation. Thus, the acceptance probability is the ratio of importance weights for $\mathbf{S}'$ and $\mathbf{S}^{(n)}$.

6. RAO–BLACKWELLIZED ESTIMATORS

Rao–Blackwellization is a technique which uses exact computation to construct estimators with reduced Monte Carlo variance (Gelfand and Smith, 1990).

Suppose realizations $(\mathbf{S}^{(n)}, \lambda^{(n)})$, $n = 1, \ldots, N$, have been sampled from the joint posterior distribution $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$. An unbiased Monte Carlo estimate of

$$\tau = \mathrm{E}_{\pi_\theta}\big(g(\mathbf{S}, \lambda)|\mathbf{Y}\big)$$

is

$$T_N = \frac{1}{N}\sum_{n=1}^{N} g\big(\mathbf{S}^{(n)}, \lambda^{(n)}\big). \tag{9}$$

Now, for any function $Z(\mathbf{S}, \lambda)$,

$$\mathrm{E}_{\pi_\theta}\big(\mathrm{E}\big(g(\mathbf{S}, \lambda)|Z(\mathbf{S}, \lambda), \mathbf{Y}\big)|\mathbf{Y}\big) = \mathrm{E}_{\pi_\theta}\big(g(\mathbf{S}_M, \lambda)|\mathbf{Y}\big) = \tau.$$

Thus, if

$$h(Z) = \mathrm{E}\big(g(\mathbf{S}, \lambda)|Z(\mathbf{S}, \lambda), \mathbf{Y}\big)$$

can be computed, an alternative unbiased Monte Carlo estimator of τ is

$$T_N^{\mathrm{RB}} = \frac{1}{N}\sum_{n=1}^{N} h\big(Z\big(\mathbf{S}^{(n)}, \lambda^{(n)}\big)\big). \tag{10}$$

For independent realizations, the reduced variance of the Rao–Blackwellized estimator is assured:

$$\mathrm{var}(T_N^{\mathrm{RB}}) \le \mathrm{var}(T_N).$$

For MCMC realizations, the reduction in variance is not universal, but usually holds (Liu, Wong and Kong, 1994). At the cost of increased computation per realization, a more precise estimator is obtained.

A Rao–Blackwellized estimator can be readily obtained from the crude estimator (7):

$$T_N^{\mathrm{crude}}(x) = \frac{1}{N}\sum_{n=1}^{N} \frac{I(\lambda^{(n)} = x)}{\pi(\lambda = x)}.$$

Taking $Z(\mathbf{S}, \lambda) = \mathbf{S}_M$,

$$\begin{aligned} h(\mathbf{S}_M) &= \mathrm{E}_{\pi_\theta}\left(\frac{I(\lambda^{(n)} = x)}{\pi(\lambda = x)}\Big|\mathbf{S}_M, \mathbf{Y}\right) \\ &= \frac{\pi_\theta(\lambda = x|\mathbf{S}_M, \mathbf{Y})}{\pi(\lambda = x)} \\ &= \frac{P_\theta(Y_T|\mathbf{S}_M, \lambda = x)}{\pi(\lambda = x)} \\ &\quad \cdot \frac{P_\theta(\mathbf{Y}_M|\mathbf{S}_M)P_\theta(\mathbf{S}_M)\pi(\lambda = x)}{P_\theta(\mathbf{Y}_M|\mathbf{S}_M)P_\theta(\mathbf{S}_M)} \\ &= \frac{P_\theta(Y_T|\mathbf{S}_M, \lambda = x)}{P_\theta(Y_T|\mathbf{S}_M)} \\ &= \frac{P_\theta(Y_T|\mathbf{S}_M, \lambda = x)}{\sum_{\lambda'} P_\theta(Y_T|\mathbf{S}_M, \lambda = \lambda')\pi(\lambda = \lambda')}, \end{aligned}$$

which is exactly calculable through single-locus pedigree-peeling computations.

Thus, a Rao–Blackwellized estimator of the multipoint likelihood $P_\theta(\mathbf{Y}|\lambda = x)$ is given by

$$\begin{aligned} T_N^{\mathrm{RB}}(x) &= \frac{1}{N}\sum_{n=1}^{N} h\big(\mathbf{S}_M^{(n)}\big) \\ &= \sum_{n=1}^{N} \frac{P_\theta(Y_T|\mathbf{S}_M^{(n)}, \lambda = x)}{\sum_{\lambda'} P_\theta(Y_T|\mathbf{S}_M^{(n)}, \lambda = \lambda')\pi(\lambda = \lambda')} \end{aligned} \tag{11}$$

and an estimate of the multipoint lod score is

$$\log_{10}\left(\frac{T_N^{\mathrm{RB}}(x)}{T_N^{\mathrm{RB}}(0)}\right). \tag{12}$$

Note that whereas the crude estimator (7) is a function only of the realized $\lambda^{(n)}$ the Rao–Blackwellized estimator (11) is a function only of $\mathbf{S}_M^{(n)}$. For the realized $\mathbf{S}_M^{(n)}$ the contribution to the estimate of the likelihood is computed for each value x of λ.

7. CONSTRUCTION OF A PSEUDO-PRIOR FOR λ

The performance characteristics of the MCMC method will generally be poor if there exist hypothesized trait locations λ with low marginal posterior probability. To improve performance of the MCMC sampling of $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$, a pseudo-prior is placed on λ, such that $\pi(\lambda) \approx (\pi_\theta(\lambda|\mathbf{Y}))^{-1}$, to produce a marginal MCMC sampling distribution for λ which is approximately uniform. Each hypothesized trait location is then sampled with approximately equal frequency, even those locations with very low marginal posterior probabilities.

To construct $\pi(\lambda)$, a preliminary analysis of the genetic data is conducted, obtaining a run of N' realizations using the MCMC method temporarily assigning equal prior probability to each $\lambda \in \{0, 1, \ldots, K\}$. A new $\pi(\lambda)$ is then obtained as the inverse of the estimate of the posterior from the preliminary MCMC run. The estimator used is analogous to the Rao–Blackwellized likelihood estimator (11) of the previous section and uses only the realized $\mathbf{S}_M$ from the preliminary run:

$$\pi(\lambda) = \left(\frac{1}{N'}\sum_{n=1}^{N'} \frac{P_\theta(Y_T|\mathbf{S}_M^{(n)}, \lambda = x)}{\sum_{\lambda'} P_\theta(Y_T|\mathbf{S}_M^{(n)}, \lambda = \lambda')}\right)^{-1}. \tag{13}$$

Some fine-tuning of $\pi(\lambda)$ may be necessary to achieve uniform sampling of λ if some λ have near-zero marginal posterior probability.

8. THE EXAMPLE DATA

Performance characteristics of the proposed MCMC method are explored through the multipoint linkage analysis of data originating from a study of early-onset Alzheimer's disease in families of Volga–German ethnic origin (Levy-Lahad et al., 1995a, b). Several families in this group carry the presenillin PS2 mutation located close to the D1S479 microsatellite marker on Chromosome 1. However, some of these families do not carry this mutation and show no linkage of the disease with markers on Chromosome 1 or on Chromosome 14 (the presenillin PS1 location). Moreover, even in families that do segregate the PS2 mutation, not all affected individuals carry the mutation and not all carriers of the mutation are affected, even at older ages (Wijsman, personal communication). This study thus provides an ideal example of a situation where the answer is known and where we have both families in which the disease is linked to Chromosome 1 and ones in which it is not. For the purposes of our current analysis the disease is treated as dominant. Affected individuals with age of onset larger than a pedigree-specific cutoff are treated as of unknown trait genotype, as also are unaffected individuals.

We have selected two of the larger pedigrees on which to show the performance of our method: one of these (R) segregates the PS2 mutation and the other (KS) does not. The R and KS pedigrees are depicted in Figures 3 and 4, respectively. Note that observed data are available only on the last two generations: due to the late onset of the disease, many individuals of interest are deceased. For the current analysis a subset of the available marker information is used. Ten linked informative microsatellite marker loci on Chromosome 1 were selected. These markers are approximately evenly spaced along a 60cM chromosomal segment surrounding the D1S479 marker. Each marker has between 8 and 12 possible alleles, although only a subset of these is observed in these two pedigrees. These pedigrees and markers are a subset of the data considered by Daw, Heath and Wijsman (1999) in their Bayesian MCMC analysis of genetic heterogeneity in Alzheimer's disease. The map and allele frequency information is taken from that study. For each marker locus, the index, associated name, map position and number of possible segregating alleles are given in Table 1. For simplicity, the sex-averaged genetic map is used in this analysis. The Haldane map function is used to convert genetic distance to recombination frequencies.

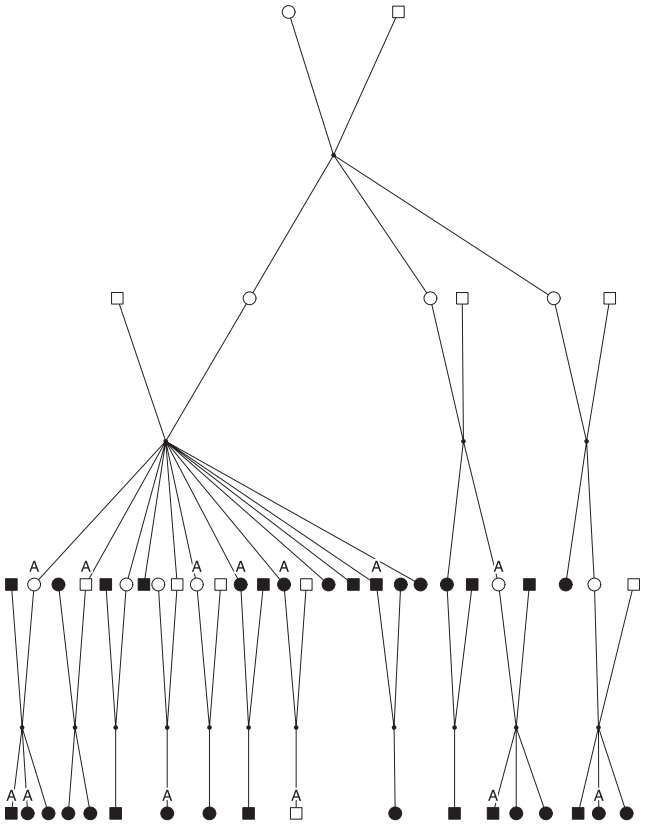


FIG. 3. *The* R *pedigree, containing* 53 *individuals, originates from a genetic study into Alzheimer's disease. The symbol* A *denotes an individual who, for the purposes of this illustrative analysis, was designated as affected. Males are denoted by squares; females are denoted by circles. Shaded squares and circles denote individuals with recorded marker information.*

TABLE 1
Marker indexes, associated names, map positions on Chromosome 1 *and number of alleles at each marker locus used in the multipoint linkage analysis of Alzheimer's disease*

Index	Name	Map position (cM)	Number of alleles
1	D1S306	215.17	12
2	D1S249	220.65	15
3	D1S245	227.81	10
4	D1S237	232.81	13
5	D1S229	237.73	8
6*	D1S479	242.34	11
7	D1S446	252.12	13
8	D1S235	254.64	9
9	D1S180	267.51	11
10	D1S102	275.68	6

*The marker known to be almost completely linked with the trait locus.

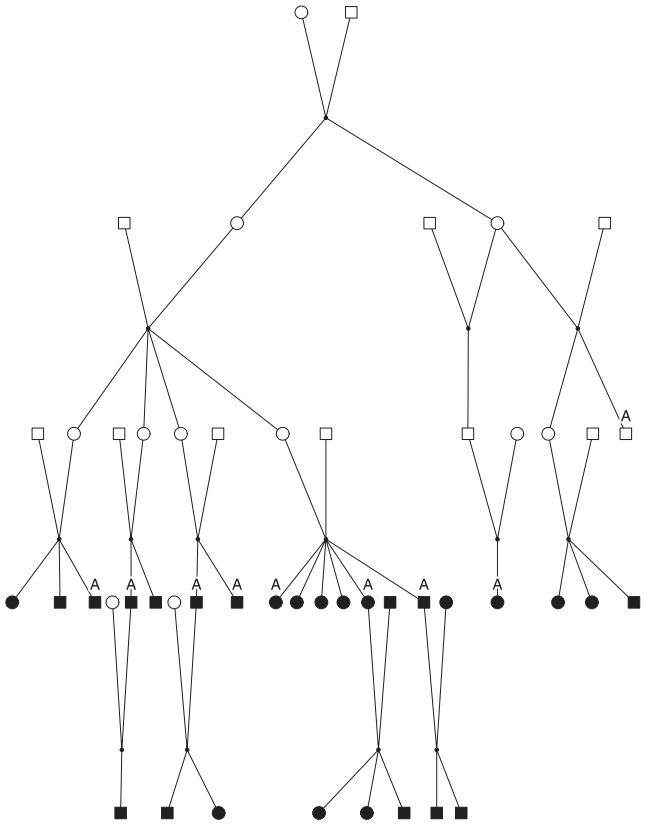


FIG. 4. *The* KS *pedigree, containing* 49 *individuals, originates from a genetic study into Alzheimer's disease. The symbol* A *denotes an individual who, for the purposes of this illustrative analysis, was designated as affected. Males are denoted by squares; females are denoted by circles. Shaded squares and circles denote individuals with recorded marker information.*

9. RESULTS

The accuracy of the crude estimator (8) and the Rao–Blackwellized estimator (12) for multipoint lod score estimation is assessed through the multipoint linkage analysis of data collected on the R and KS pedigrees. For comparison, exact lod scores are obtained via pedigree peeling, but a joint analysis involving all 10 markers is infeasible. Instead, analyses are based on four-locus data, using the three markers D1S306, D1S479 and D1S102. Hypothesized trait locations are placed within each marker interval at proportional genetic distances $0.1, 0.2, \ldots, 0.9$. A trait allele frequency of 0.05 is assumed.

Computations are performed on a Dell Workstation using a single Pentium III processor running at 933 MHz. Exact lod scores at the hypothesized trait locations are computed using the software package VITESSE (O'Connell and Weeks, 1995). The MCMC methods described in previous sections have been implemented using the framework of the MORGAN Version 2.6 package for Monte Carlo genetic analysis (http: // www.stat.washington.edu / thompson / Genepi / MORGAN / Morgan.shtml).

Specifications of the MCMC runs for the analysis of the four-locus data are as follows. A preliminary analysis of the genetic data collected on R and KS pedigrees is conducted for the construction of $\pi(\lambda)$. In this phase the MCMC method is run with a uniform prior placed on λ, for 5000 and 6000 iterations, respectively. Using the pseudo-prior $\pi(\lambda)$ estimated from these preliminary runs (see Section 7), the multipoint lod scores are then estimated from an MCMC run of length 10,000 for the R pedigree and 40,000 for the KS pedigree. All runs are based on an L-sampler proportion of 20% ($p_L = 0.2$).

Figures 5 and 6 show close agreement between the exact lod score curve and the estimated lod score curves for the four-locus analyses. For the R pedigree, exact lod scores are computed in 34.05 minutes, while lod scores estimated via (8) and (12) take only 1.99 and 2.82 minutes, respectively. Similarly, for the KS pedigree, exact lod scores are computed in 168.31 minutes, while lod scores estimated via (8) and (12) take 4.81 and 7.95 minutes, respectively. These MCMC run times include time spent performing a preliminary analysis for the construction of the pseudo-prior $\pi\lambda$.

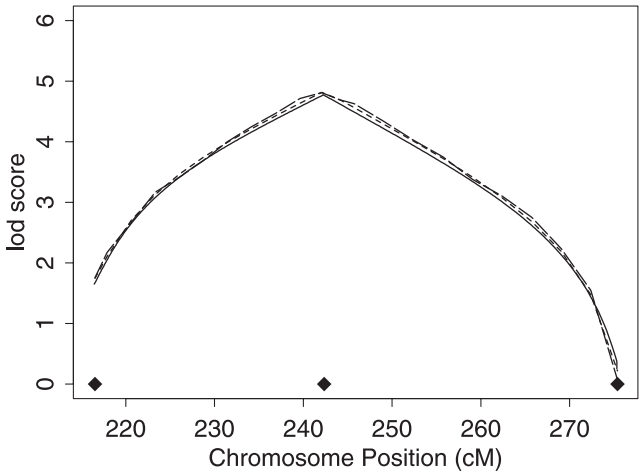


FIG. 5. *Exact and estimated multipoint lod scores at hypothesized trait locations for the analysis of the four-point data collected on the* R *pedigree. Marker positions are denoted by the diamonds. The solid line represents the exact lod score curve computed using pedigree peeling. The long-dashed line represents the lod score curve estimated from an MCMC run using* (8). *The short-dashed line represents the lod score curve estimated from an MCMC run using* (12).

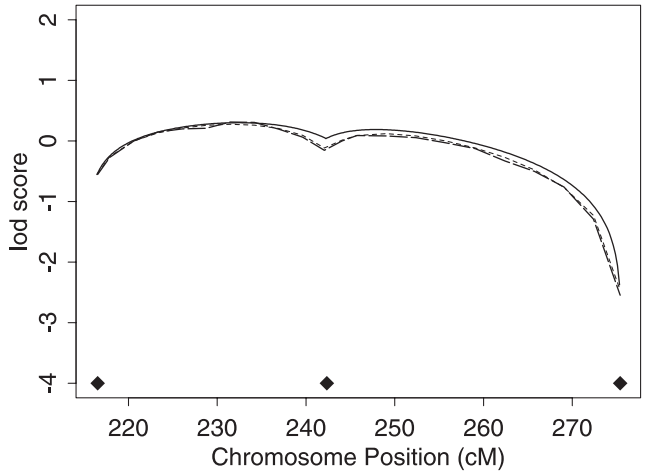


FIG. 6. *Exact and estimated multipoint lod scores at hypothesized trait locations for the analysis of the four-point data collected on the* KS *pedigree. Marker positions are denoted by the diamonds. The solid line represents the exact lod score curve computed using pedigree peeling. The long-dashed line represents the lod score curve estimated from an MCMC run using* (8). *The short-dashed line represents the lod score curve estimated from an MCMC run using* (12).

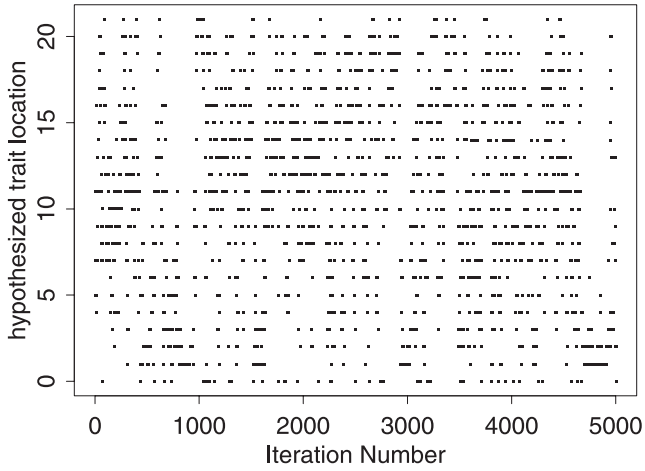


FIG. 7. *Plot of* $\lambda^{(n)}$ *over MCMC iteration number. For clarity only the first* 5000 *iterations, after the preliminary analysis, are shown. Sampled values are obtained from an MCMC run for the analysis of four-point data collected on the* R *pedigree.*

Convergence is diagnosed through the inspection of trace plots of $S_{\cdot j}$, $S_{i\cdot}$ and λ (i.e., plots of realized values of these variables over iteration number). These trace plots also show the good mixing properties of these MCMC runs. For example, Figures 7 and 8 plot sampled values of λ against MCMC iteration number for the analysis of data collected on R and KS pedigrees, respectively. For clarity, each trace plot shows only the first 5000 values of λ not including values sampled in the preliminary analysis. The plots display no trend, and each hypothesized trait location is well sampled with no obvious bias for particular values.

Pointwise Monte Carlo standard errors of the lod scores are estimated using the batch-means method of Hastings (1970) as described in the Appendix. Each run is divided into 20 batches; in no run did the batch means display significant autocorrelation. Monte Carlo standard errors associated with lod scores estimated via (8) are approximately twice as large as standard errors associated with lod scores estimated via (12). In fact, using a run length of 50,000 taking 6.05 minutes for the R pedigree and a run length of 100,000 taking 10.61 minutes for the KS pedigree is required before the standard errors associated with (8) are approximately equal (≈ 0.03) to those for estimator (12) in the runs described above. Using the Rao–Blackwellized estimator (12) requires extra computing time per MCMC iteration, but the reduced standard errors more than compensate for the additional computational cost.

When using the multipoint data for all 10 marker loci, implementation of the MCMC method is essentially the same as described above. A preliminary analysis of the R and KS pedigrees is conducted where 20,000 and 30,000 iterations are performed, respectively. Multipoint lod scores for the R and KS pedigrees are then obtained from an MCMC run of length 200,000 iterations taking 71.41 minutes and 300,000 iterations taking 103.24 minutes, respectively. The resulting lod score curves are shown in Figures 9 and 10.

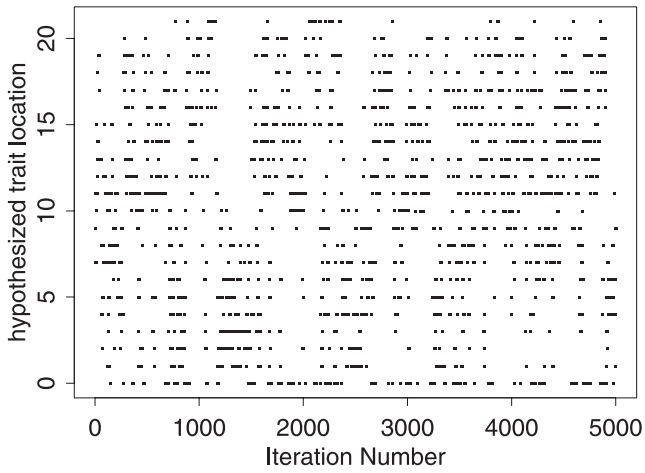


FIG. 8. *Plot of* $\lambda^{(n)}$ *over MCMC iteration number. For clarity only the first* 5000 *iterations, after the preliminary analysis, are shown. Sampled values are obtained from an MCMC run for the analysis of four-point data collected on the* KS *pedigree.*

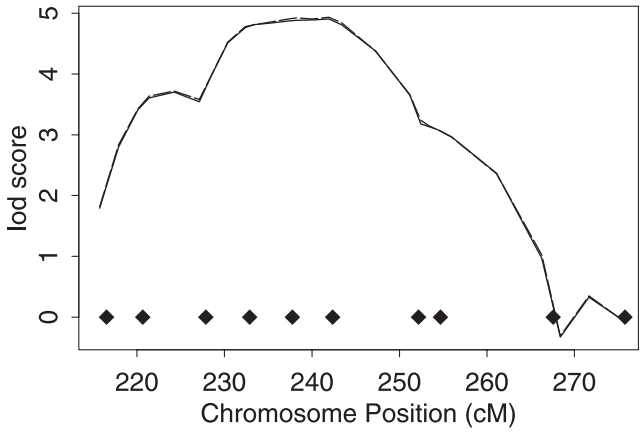


FIG. 9. *Estimated multipoint lod scores at hypothesized trait locations for the analysis of multipoint data collected on the* R *pedigree. Marker positions are denoted by the diamonds. The solid line represents the lod score curve estimated from an MCMC run using* (8). *The dashed line represents the lod score curve estimated from an MCMC run using* (12). *The two curves are practically indistinguishable.*

Hypothesized trait locations are placed within each marker interval at proportional genetic distances 0.1, 0.5, 0.9.

Once again, standard errors associated with lod scores estimated via the crude estimator are approximately twice the standard errors associated with lod scores estimated via the Rao–Blackwellized estimator (which were between 0.03 and 0.09 across λ). Although exact answers are not available for the joint analysis of all 10 markers, a disease gene is correctly

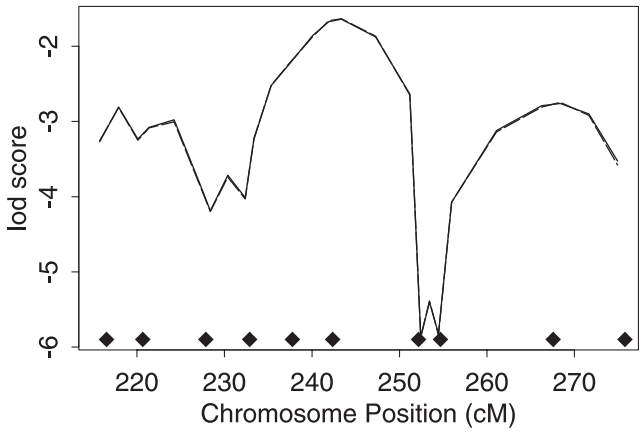


FIG. 10. *Estimated multipoint lod scores at hypothesized trait locations for the analysis of multipoint data collected on the* KS *pedigree. Marker positions are denoted by the diamonds. The solid line represents the lod score curve estimated from an MCMC run using* (8). *The dashed line represents the lod score curve estimated from an MCMC run using* (12). *The two curves are practically indistinguishable.*

detected in the R pedigree and localized to marker D1S479 (see Figure 9). This marker is very close to the presenillin PS2 mutation (Levy-Lahad et al., 1995a). Similarly, the lod score curves depicted in Figure 10 verify the absence of a disease gene in this region of Chromosome 1 segregating in the KS family, in agreement with previous analyses of data on this family.

A reasonable number of restarts are accepted for the four-locus analysis of the R and KS data with acceptance rates of 10% and 5%, respectively. However, sequential imputation failed to propose acceptable restart states for the multipoint analysis of all 10 tightly linked markers.

10. DISCUSSION

Methods to compute exact multipoint lod scores continue to increase in speed and efficiency, allowing exact multipoint linkage analyses of large data sets. However, analyses of data collected on extended pedigrees (more than 40 individuals) with several (more than 3) linked and highly polymorphic markers remain beyond the computational boundaries of exact methods, particularly where a substantial proportion of pedigree members are unobserved. MCMC methods often provide the only viable means of analysis.

Using several recent advances in MCMC methodology, new MCMC procedures for the estimation of multipoint lod scores are presented in this paper. These new MCMC procedures have been implemented within the framework of the MORGAN Version 2.6 package for Monte Carlo genetic analysis (http: // www.stat.washington.edu / thompson / Genepi / MORGAN / Morgan.shtml).

The methodology is demonstrated through its application to multipoint data collected on two large extended pedigrees where data are collected on a disease trait and 10 tightly linked and highly polymorphic markers. Through batch updates of the latent variables, Metropolis–Hastings restarts and integrated proposal distributions, we obtain MCMC runs which mix well over the sample space of trait locations and meiosis indicators. The realized values of these variables result in accurate estimates of the multipoint lod scores.

Rao–Blackwellization is a useful technique for constructing estimators with reduced Monte Carlo variance. At the expense of increased computation per iteration, a more precise estimator is obtained. The Rao–Blackwellized estimator (12) uses trait locus exact likelihood computation conditional on the realized $\mathbf{S}_M$ to estimate multipoint lod scores. The crude

estimator (8) provides zero likelihood estimates, and hence infinite $(-\infty)$ log-likelihoods, at hypothesized trait locations λ not sampled within an MCMC run. The Rao–Blackwellized estimator will always provide finite estimates of the lod score, unless the true likelihood of a location is 0, but still has high variance at locations of small likelihood. More precise estimates of lod scores are obtained if the prior distribution of λ is chosen such that the MCMC run samples the locations approximately uniformly. Since the likelihood for an unlinked trait $L(\lambda = 0)$ enters into every lod score estimate, even better performance might be obtained with a prior which puts increased weight on $\lambda = 0$.

Proposing restarts for multipoint data collected on several tightly linked markers is difficult, as evidenced in this paper. Since the process of realization is sequential over loci, only the data for loci to one side of a given locus contribute to imputation at that locus. The dependence that exists between a locus and its adjacent loci is only partially captured. Improvements suggested by Irwin, Cox and Kong (1994) are to process the loci in an order starting from the locus with the least amount of missing information and splitting highly polymorphic loci into two completely linked (artificial) loci. Implementation of these suggestions may make sequential imputation restarts feasible for our 10-marker example, and improve acceptance rates for the examples with smaller numbers of markers.

Several extensions of the methods of this paper are almost immediate. First, the MCMC method presented here can be extended to detecting and localizing a quantitative trait locus (QTL). Routines to compute the necessary trait likelihoods conditional upon $\mathbf{S}_M$ already exist within the MORGAN package (Heath, 1997). In fact, improved mixing is expected for a QTL because the trait locus places no absolute constraints upon the latent space of meiosis indicators: all configurations of trait genotypes are, in principle, feasible. Second, improvements to the L- and M-samplers can be made through jointly sampling $S_{\cdot j}$ for several loci j and $S_{i\cdot}$ for several meioses i, respectively. However, improved MCMC mixing needs to be weighed against the increased computational burden per iteration. Moreover, the choice of meioses for joint updating is nontrivial (Thompson and Heath, 1999): some proposals in this regard have recently been made by Thompson (2000b) and Thomas, Gutin, Abkevich and Bansal (2000). Third, further investigation is required into other estimators of multipoint lod scores based on realized values of sampled variables. The specific form of Rao–Blackwellized estimator used in this paper is only the simplest of many alternatives. For example, in the context of single-marker lod scores, Jensen and Kong (1999) have proposed an interesting class of estimators that extend the simple MCMC likelihood ratio estimator of a recombination frequency given by Thompson and Guo (1991).

In practice, the parameters associated with a genetic model for a trait are unknown or at best approximately known. Bayesian approaches to linkage analysis treat these parameters as nuisance variables: they are marginalized out of the joint posterior distribution. A similar strategy could be adopted here with regard to the parameters designated θ, but interpretability of the multipoint lod score is then compromised, an issue which limits widespread acceptance of Bayesian linkage analysis. An alternative approach to model uncertainty is through mod scores (Clerget-Darpoux, Bonaïti-Pellié and Hochez, 1986): mod scores are lod scores maximized over the genetic model parameters θ:

$$\mathrm{mod}(\lambda) = \max_{\theta}\bigl(\log P_\theta(\mathbf{Y}|\lambda) - \log P_\theta(\mathbf{Y}|\lambda = 0)\bigr).$$

Used with care and caution, mod scores possess the ability both to detect linkage and to identify an appropriate genetic model. Hodge and Elston (1994) and Liang, Rathouz and Beaty (1996) provide discussion on the calculation, use and limitations of mod scores. For a low-dimensional genetic model θ, for example, four or five parameters of the trait model, MCEM (Guo and Thompson, 1994) provides a method for the maximization of the likelihood for any fixed λ (including $\lambda = 0$) and for the joint maximization of the likelihood with respect to both θ and λ. Thus, determination of the mod score is feasible, although more computationally intensive than determination of the lod score for a fixed value of θ.

APPENDIX

A.1 M-sampler

The M-sampler (Thompson and Heath, 1999; Thompson, 2000a) is a whole-meiosis Gibbs sampler which jointly updates an entire meiosis $S_{i\cdot}$ from the full conditional distribution $P_\theta(S_{i\cdot}|\mathbf{S}_{-i\cdot}, \mathbf{Y}, \lambda)$, where $\mathbf{S}_{-i\cdot} = \{S_{k\cdot}, k \neq i\}$. Calculation of $P_\theta(S_{i\cdot}|\mathbf{S}_{-i\cdot}, \mathbf{Y}, \lambda)$ proceeds from the forward–backward algorithm of Baum, Petrie, Soules and Weiss (1970) which calculates exact probability distributions, under a hidden Markov model (HMM), of latent variables conditional on observed data. For a detailed discussion of the

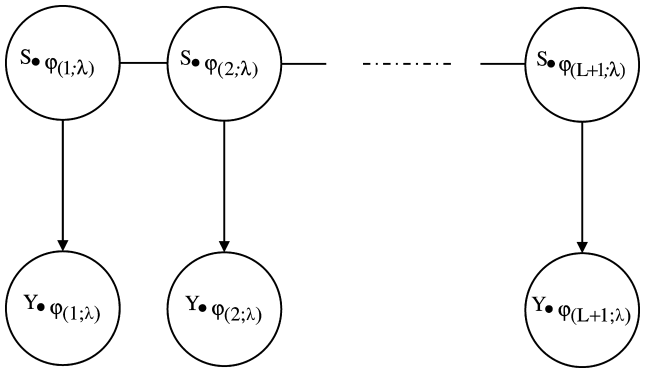


FIG. 11. *Directed acyclic graph of a hidden Markov model for the genetic data, where* $\mathbf{Y} = (Y_{\cdot\varphi(1;\lambda)}, \ldots, Y_{\cdot\varphi(L+1;\lambda)})$ *denotes the observed data and* $\mathbf{S} = (S_{\cdot\varphi(1;\lambda)}, \ldots, S_{\cdot\varphi(L+1;\lambda)})$ *denotes the latent variables. The absence of genetic interference is assumed. Loci are in the order* $\varphi(1;\lambda), \varphi(2;\lambda), \ldots, \varphi(L+1;\lambda)$, *where* $\varphi(j;\lambda)$ *denotes the* j*th element in the list of loci ordered with a trait locus in position* λ.

forward–backward algorithm and its application in genetic analysis, see Thompson (2000b).

Assuming the absence of genetic interference, latent variables $(S_{\cdot\varphi(j;\lambda)}; j = 1, \ldots, L+1)$ follow a first-order Markov chain, where $\varphi(j;\lambda)$ denotes the jth element in the ordered list of loci with the trait locus in hypothesized trait location λ. The observed data $\mathbf{Y}$, partitioned as $(Y_{\cdot\varphi(j;\lambda)}; j = 1, \ldots, L+1)$, can then be modeled as an HMM. Figure 11 shows the direct acyclic graph of an HMM for genetic data where the nodes denote the random variables and the presence of an edge between two nodes indicates a direct dependency between the two (sets of) variables.

Note that, conditional on both $\mathbf{Y}$ and $\{S_{i\cdot}; i \in \mathcal{M}\}$ for any subset $\mathcal{M}$ of the meioses, indicators for the remaining meioses have the same HMM structure. In particular, $(S_{i\varphi(j;\lambda)}; j = 1, \ldots, L+1)$ is a two-state Markov chain ($S_{i\varphi(j;\lambda)} = 0$ or 1) and has the HMM structure conditional upon $\mathbf{S}_{-i\cdot}$ and $\mathbf{Y}$. We give here the M-sampler update for a single meiosis i and denote the current value of $\mathbf{S}$ before this sampling step by $\mathbf{S}^{(n)}$ as if i were the first meiosis to be updated. A full M-sampler scan consists of updating all the meioses successively in random order. Each $S_{i\cdot}$ is updated conditional on the current values $S_{k\cdot}^{(n)}$ or $S_{k\cdot}^{(n+1)}$, depending on whether meiosis k has yet been updated.

The exact calculation of $P_\theta(S_{i\cdot}|\mathbf{S}_{-i\cdot}^{(n)}, \mathbf{Y}, \lambda)$ begins with the computation of the forward probabilities

$$Q_j(s) = P_\theta\big(S_{i\varphi(j;\lambda)} = s|\mathbf{S}_{-i\cdot}^{(n)}, \mathbf{Y}_\cdot(1:j), \lambda\big) \quad \text{for } s = 0, 1,$$

where $\mathbf{Y}_\cdot(1:j) = (Y_{\cdot\varphi(1;\lambda)}, \ldots, Y_{\cdot\varphi(j;\lambda)})$, the data for the first j loci along the chromosome. Now $Q_1(s) \propto P_\theta(Y_{\cdot\varphi(1;\lambda)}|S_{\cdot\varphi(1;\lambda)}, \lambda)$ and, moving along the chromosome in a forward direction, $Q_j(s)$ can be calculated iteratively as

$$\begin{aligned} Q_j(s) \propto P_\theta(Y_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j;\lambda)}, \lambda) \\ \cdot \big(Q_{j-1}(s)(1-\rho_{\varphi(j-1;\lambda)}) \\ + Q_{j-1}(1-s)\rho_{\varphi(j-1;\lambda)}\big), \end{aligned}$$

where $\rho_{\varphi(j-1;\lambda)}$ is the recombination fraction between loci $\varphi(j-1;\lambda)$ and $\varphi(j;\lambda)$. The probabilities $Q_j(s)$ are easily normalized with respect to s, since s takes only the two values 0 or 1.

Once $Q_j(s)$ is available for $j = 1, 2, \ldots, L+1$, the meiosis indicator $S_{i\varphi(j-1;\lambda)}$ is successively sampled back along the chromosome via the backward probability

$$\begin{aligned} P_\theta\big(S_{i\varphi(j-1;\lambda)} = s|\mathbf{S}_{-i\cdot}^{(n)}, \mathbf{S}_i^{(n+1)}(j:L+1), \mathbf{Y}\big) \\ \propto Q_{j-1}(s)\big(|S_{i\varphi(j;\lambda)} - s|\rho_{\varphi(j-1;\lambda)} \\ + (1 - |S_{i\varphi(j;\lambda)} - s|)(1-\rho_{\varphi(j-1;\lambda)})\big), \end{aligned}$$

where $\mathbf{S}_i^{(n+1)}(j:L+1) = \{S_{i\varphi(l;\lambda)}^{(n+1)}, l = j, \ldots, L+1\}$ is the set of previously (back) sampled latent variables. Thus, when the indicators $\{S_{i\varphi(j;\lambda)}; j = L+1, \ldots, 2, 1\}$ at all loci have been successively sampled, a joint realization $\mathbf{S}_{i\cdot}^{(n+1)} = (S_{i0}^{(n+1)}, S_{i1}^{(n+1)}, \ldots, S_{iL}^{(n+1)})$ has been realized from the conditional distribution $P_\theta(S_{i\cdot}|\mathbf{S}_{-i\cdot}^{(n)}, \mathbf{Y}, \lambda)$. Because the components of $S_{i\cdot}$ are updated jointly, the M-sampler does not suffer from problems with mixing when the loci are tightly linked, but does suffer from poor mixing on extended pedigrees.

A.2 L-sampler

The L-sampler (Heath, 1997) is a whole-locus Gibbs sampler, combining Monte Carlo simulation with single-locus peeling, to update jointly the complete set of meiosis indicators at locus in position j conditional on observed phenotypic data $Y_{\cdot\varphi(j;\lambda)}$ and the meiosis indicators at adjacent loci $S_{\cdot\varphi(j-1;\lambda)}$ and $S_{\cdot\varphi(j+1;\lambda)}$. Although implementation of the L-sampler requires pedigrees that are single-locus peelable, the latent variable space associated with a single locus is sufficiently small that large extended pedigrees with multiple marriage loops are computationally feasible.

Each step of the L-sampler draws $S_{\cdot\varphi(j;\lambda)}$ from the full conditional distribution

$$P_\theta\big(S_{\cdot\varphi(j;\lambda)}|\{S_{\cdot j'}: j' = 1, \ldots, L+1, \; j' \neq \varphi(j;\lambda)\}, \mathbf{Y}, \lambda\big),$$

which the conditional independence structure reduces to

$$P_\theta(S_{\cdot\varphi(j;\lambda)}|S_{\cdot\varphi(j-1;\lambda)}, S_{\cdot\varphi(j+1;\lambda)}, Y_{\cdot\varphi(j;\lambda)}, \lambda).$$

For notational convenience, in the following we assume the current step updates locus j ($j \in \{0, 1, 2, \ldots, L\}$), replacing $\varphi(j;\lambda)$ by j, and denote by $j\pm$ the two neighboring loci in the ordering defined by λ. Furthermore, as for the M-sampler, we assume that this locus is the first to be updated, and thus denote the current values by $S^{(n)}_{\cdot j\pm}$. In reality, of course, each $S_{\cdot j}$ is updated successively, conditional on the current values $S^{(n+1)}_{\cdot j'}$ or $S^{(n)}_{\cdot j'}$ depending on whether or not the locus j' has or has not yet been updated.

Pedigree peeling also uses an HMM structure, but there is added complexity due to the pedigree, which is not a linear structure but at best a tree, and due to the fact that there is directionality in the definition of Mendelian transmission probabilities from parents to offspring. Consider a set of meioses $\mathcal{C}$ (a cutset) which together divide the pedigree into two or more disjoint parts. Conditional on the values of the meiosis indicators in $\mathcal{C}$ and allelic types assigned to ibd genes, data observed on the disjoint pedigree segments are independent. As for the M-sampler, we work successively through the data: denote by $Y_{\mathcal{P}j}$ the data at locus j already accumulated from one of the disjoint pedigree partitions $\mathcal{P}$ defined by $\mathcal{C}$. Furthermore, suppose that $\mathcal{C} = \mathcal{C}_1 \cup \mathcal{C}_2$ is such that $\mathcal{P}$ is ancestral to $\mathcal{C}_1$ and descendant to $\mathcal{C}_2$ (see Figure 12). Thus, we can define an R-function (Cannings, Thompson and Skolnick, 1978), analogous to the functions $Q_j(s)$ of the M-sampler:

$$R_{\mathcal{C}j}(s_1, s_2) = P_\theta\big(Y_{\mathcal{P}j}, S_{\mathcal{C}_1 j} = s_1 | S_{\mathcal{C}_2 j} = s_2, S^{(n)}_{\cdot j\pm}, \lambda\big),$$

where s_1 and s_2 are binary vectors of length the number of meioses in $\mathcal{C}_1$ and $\mathcal{C}_2$, respectively.

Consider, for example, a pedigree without loops and an individual who is the only child of his or her parents. In this case, the paternal (maternal) meiosis i_{p} (i_{m}) of the individual is a cutset of size 1, which divides the pedigree into the part connected to the individual through his or her father (mother) from the remainder. More generally, in any pedigree without loops, the two meioses $(i_{\mathrm{p}}, i_{\mathrm{m}})$ of each nonfounder individual divide the pedigree into the ancestral part $\mathcal{P}_1$ and the descendant part $\mathcal{P}_2$. The ancestral part $\mathcal{P}_1$ is connected to the individual via his or her parents and includes these parents, their ancestral relatives, the individual's siblings and all their descendant relatives.

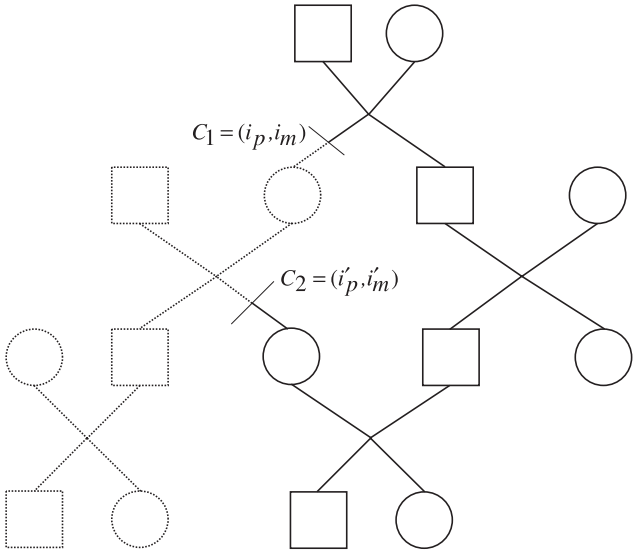


FIG. 12. *A complex pedigree illustrating the use of cutsets for dividing a pedigree into disjoint partitions, ready for computation via the R-function. For this complex pedigree, cuts* $\mathcal{C}_1$ *and* $\mathcal{C}_2$ *divide the pedigree into disjoint parts. Information from the pedigree part denoted by dashed lines has been processed. The pedigree part denoted by solid lines remains to be processed. Each connecting line denotes both the paternal meiosis* (i_p) *and the maternal meiosis* (i_m) *of the offspring individual. The processed part* $\mathcal{P}$ *is descendant to meioses* $\mathcal{C}_1 = (i_{\mathrm{p}}, i_{\mathrm{m}})$ *and ancestral to meioses* $\mathcal{C}_2 = (i'_{\mathrm{p}}, i'_{\mathrm{m}})$.

The descendant part $\mathcal{P}_2$ is connected to the individual via his or her offspring and includes the individual, his or her offspring, their descendant relatives, his or her spouses, their ancestral relatives, and their descendant relatives via their offspring from other marriages. The two relevant R-functions for the cutset $\mathcal{C} = (i_{\mathrm{p}}, i_{\mathrm{m}})$ are then

$$\begin{aligned} &R^{(1)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}}) \\ &\quad = P_\theta\big(Y_{\mathcal{P}_1 j}, S_{\mathcal{C}j} = (s_{\mathrm{p}}, s_{\mathrm{m}}) | S^{(n)}_{\cdot j\pm}, \lambda\big), \\ &R^{(2)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}}) \\ &\quad = P_\theta\big(Y_{\mathcal{P}_2 j} | S_{\mathcal{C}j} = (s_{\mathrm{p}}, s_{\mathrm{m}}), S^{(n)}_{\cdot j\pm}, \lambda\big). \end{aligned} \tag{A.1}$$

Using these R-functions, and the conditional independence provided by the a priori independence of meioses, computation proceeds through the pedigree from cutset to cutset until $\mathcal{P}$ consists of the entire pedigree and the R-function is the probability of $Y_{\cdot j}$ jointly on or conditionally with the meiosis indicators of the final cutset and conditional on $S^{(n)}_{\cdot j\pm}$. The sequence of cutsets $\mathcal{C}$ is known as the peeling sequence. This is analogous to the forward computation of the

M-sampler. Now the meiosis indicators of the final cutset may be realized. Then, using the R-functions computed in the first computation, meiosis indicators of each successive cutset in the reverse of the peeling sequence may be realized conditional on values just realized and on all the data $Y_{\cdot j}$ and $S^{(n)}_{\cdot j\pm}$. This process is analogous to the reverse sampling procedure of the M-sampler.

In general, a single pass through the pedigree suffices to compute the probability $P(Y_{\cdot j}|S^{(n)}_{\cdot j\pm}, \lambda)$ and provide the R-functions used for the realization of $S_{\cdot j}$ from the full conditional distribution $P_\theta(S_{\cdot j}|S^{(n)}_{\cdot j\pm}, Y_{\cdot j}, \lambda)$. Further details of the peeling computation on arbitrary pedigrees are described by Cannings, Thompson and Skolnick (1978), the only conceptual difference to the current case being that Mendelian transmission at the locus in position j is now conditioned on the current inheritance pattern at the two neighboring loci.

In addition to providing a joint realization of $S_{\cdot j}$ from $P_\theta(S_{\cdot j}|S^{(n)}_{\cdot j\pm}, Y_{\cdot j}, \lambda)$, the peeling procedure also provides the marginal probabilities $P_\theta(S_{\mathcal{C}j}|S^{(n)}_{\cdot j\pm}, Y_{\cdot j}, \lambda)$ for the set $\mathcal{C}$ of meioses that are in the final cutset, and can be used to provide this distribution for other cutsets. Consider again a pedigree without loops and let $\mathcal{C} = (i_{\mathrm{p}}, i_{\mathrm{m}})$ be the paternal and maternal meioses of a single individual. A single pass through the pedigree provides either $R^{(1)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})$ or $R^{(2)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})$ of (A.1), and a peeling sequence which passes through the individual in the opposite direction provides the other. Combining these,

$$P_\theta\big(S_{\mathcal{C}j} = (s_{\mathrm{p}}, s_{\mathrm{m}})|S_{\cdot j\pm}, Y_{\cdot j}, \lambda\big) = \frac{R^{(1)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})R^{(2)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})}{\sum_{(s_{\mathrm{p}}, s_{\mathrm{m}})} R^{(1)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})R^{(2)}_{\mathcal{C}j}(s_{\mathrm{p}}, s_{\mathrm{m}})}. \tag{A.2}$$

Since each of s_{p} and s_{m} is 0 or 1, the sum in the denominator of (A.2) has only four terms. Even on a pedigree without loops, several different peeling sequences may be required to obtain $P_\theta(S_{\mathcal{C}j}|S^{(n)}_{\cdot j\pm}, Y_{\cdot j}, \lambda)$ for all the sets $\mathcal{C}$ of interest, but this operation is relatively inexpensive even on complex pedigrees (Thompson, 1981).

The L-sampler is irreducible when the recombination probabilities between adjacent loci are nonzero. Further, since all the components of $S_{\cdot j}$ are updated jointly, extended pedigrees do not cause mixing problems. However, tightly linked markers do result in poor mixing. Only by combining the L-sampler and M-sampler are good Monte Carlo estimates of likelihoods or posterior probabilities obtained (Heath and Thompson, 1997; Thompson, 2000b).

A.3 Updating λ via the M–H Algorithm with an Integrated Acceptance Probability

Sampling λ from $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$ is generally accomplished by updating λ via the M–H algorithm where a proposal state λ' is drawn from some conveniently chosen proposal distribution $q(\lambda|\lambda^{(n)})$ and accepted with probability $\alpha(\lambda^{(n)}, \lambda')$ such that

$$\alpha\big(\lambda^{(n)}, \lambda'\big) = \min\left[1, \frac{\pi_\theta(\mathbf{S}^{(n)}, \lambda'|\mathbf{Y})q(\lambda^{(n)}|\lambda')}{\pi_\theta(\mathbf{S}^{(n)}, \lambda^{(n)}|\mathbf{Y})q(\lambda'|\lambda^{(n)})}\right]. \tag{A.3}$$

However, the realized Markov chain $\lambda^{(n)}$, $n = 1, \ldots, N$, generally mixes poorly over marker intervals. Adjacent marker loci constrain S_T, sometimes strongly if the markers are tightly linked. Therefore, the joint posterior probability of $(\mathbf{S}^{(n)}, \lambda')$ is generally relatively low when compared to the joint posterior probability of $(\mathbf{S}^{(n)}, \lambda^{(n)})$, resulting in a high proportion of λ' being rejected. [Notationally, λ is assumed to have been updated before $\mathbf{S}$, and hence $\mathbf{S}^{(n)}$ is used in (A.3), but the order in which sets of variables are updated within a full MCMC iteration is, in fact, arbitrary.]

An alternate approach, which promotes good mixing, is to draw λ' from some convenient proposal distribution $q(\lambda|\lambda^{(n)})$ and accept with probability $\alpha(\lambda^{(n)}, \lambda')$, where S_T is integrated out of $\pi_\theta(\mathbf{S}, \lambda|\mathbf{Y})$ to give $\pi_\theta(\mathbf{S}_M, \lambda|\mathbf{Y})$. The integrated acceptance probability becomes

$$\alpha\big(\lambda^{(n)}, \lambda'\big) = \min\left[1, \frac{\pi_\theta(\mathbf{S}^{(n)}_M, \lambda'|\mathbf{Y})q(\lambda^{(n)}|\lambda')}{\pi_\theta(\mathbf{S}^{(n)}_M, \lambda^{(n)}|\mathbf{Y})q(\lambda'|\lambda^{(n)})}\right]. \tag{A.4}$$

Since λ' is drawn from a uniform distribution $U[0, K]$, $q(\lambda^{(n)}|\lambda') = q(\lambda'|\lambda^{(n)})$. Furthermore,

$$\pi_\theta(\mathbf{S}_M, \lambda) \propto P_\theta(Y_T|\mathbf{S}_M, \lambda)P_\theta(\mathbf{Y}_M|\mathbf{S}_M)\pi(\mathbf{S}_M)\pi(\lambda),$$

and thus the integrated acceptance probability (A.4) simplifies to

$$\alpha\big(\lambda^{(n)}, \lambda'\big) = \min\left[1, \frac{P_\theta(Y_T|\mathbf{S}^{(n)}_M, \lambda')\pi(\lambda')}{P_\theta(Y_T|\mathbf{S}^{(n)}_M, \lambda^{(n)})\pi(\lambda^{(n)})}\right]. \tag{A.5}$$

The probabilities $P_\theta(S_T|\mathbf{S}_M, \lambda)$ are obtained by peeling over the trait locus at positions $\lambda = \lambda^{(n)}$ and $\lambda = \lambda'$. These conditional probability computations are analogous to single-locus computations.

If λ' is accepted, then $\lambda^{(n+1)} = \lambda'$ and S_T is then sampled from its full conditional distribution $P_\theta(S_T | \mathbf{S}_M^{(n)}, \lambda^{(n+1)}, Y_T)$ which is a single-locus L-sampler step. It is easily seen that this updating of λ conditioning only on $\mathbf{S}_M$ and then resampling S_T from the full conditional is equivalent to a joint Metropolis–Hastings update of (λ, S_T) and hence maintains the sampler's correct equilibrium distribution $\pi_\theta(\lambda, \mathbf{S}|\mathbf{Y})$. Besag, Green, Higdon and Mengersen (1995) discuss in more detail the requirements for validity of sequences of partial conditioning updates. Essentially, any variables being conditioned upon at any stage must currently have the correct joint distribution.

A.4 Calculating Monte Carlo Standard Errors of lod Score Estimators

The method of batch means (Hastings, 1970) is an easily implemented procedure for estimating variances and standard deviations of estimators formed from potentially highly autocorrelated Monte Carlo realizations. Suppose

$$t_f = \frac{\sum_{i=1}^N f(X^{(i)})}{N}$$

is an estimator of $e_f = \mathrm{E}(f(X))$, based on a sequence of dependent realizations $\{X^{(i)}, i = 1, \ldots, N\}$. The standard deviation $\sqrt{\mathrm{var}(t_f)}$ of the estimator is required.

The method of batch means first groups the N realizations into B consecutive and nonoverlapping batches of size M. Denoting the mean of block b by

$$\overline{t}_{f,b} = \sum_{m=1}^{M} \frac{f(X^{((b-1)M+m)})}{M},$$

the batch means $\{\overline{t}_{f,b}, b = 1, \ldots, B\}$ can be treated as independent realizations, provided the batch size is large. Typically, the total run is divided into a small number of batches (≈ 20). Provided there is no significant autocorrelation in the batch means, this small number of large batches leads to accurate estimates of $\mathrm{var}\,(t_f)$. An estimate of $\mathrm{var}(t_f)$ is then given by

$$\sigma_f^2 = \sum_{b=1}^{B} \frac{(\overline{t}_{f,b} - \overline{\overline{t}}_f)^2}{B(B-1)}, \tag{A.6}$$

where the mean of the batch means is

$$\overline{\overline{t}}_f = \sum_{b=1}^{B} \frac{\overline{t}_{f,b}}{B} = \sum_{b=1}^{B} \sum_{m=1}^{M} \frac{X^{((b-1)M+m)}}{MB} = t_f.$$

In the case of the lod score estimators of this paper, the estimator is of the form

$$\log_{10}\left(\frac{t_f}{t_g}\right) = \log_{10} t_f - \log_{10} t_g = \frac{(\log_e t_f - \log_e t_g)}{\log_e 10},$$

where

$$\frac{t_f}{t_g} = \begin{cases} T_N^{\mathrm{crude}}(x)/T_N^{\mathrm{crude}}(0), \\ \quad \text{for the crude estimator,} \\ T_N^{\mathrm{RB}}(x)/T_N^{\mathrm{RB}}(0), \\ \quad \text{for the Rao–Blackwellized estimator.} \end{cases}$$

Note $e_f = \mathrm{E}(t_f) = \mathrm{E}(f(X)$ is the log-likelihood at $\lambda = x$ and $e_g = \mathrm{E}(t_g) = \mathrm{E}(g(X))$ is the log-likelihood for independent segregation at the trait locus ($\lambda = 0$).

Hence, the variance of the lod score estimator is

$$\frac{\mathrm{var}(\log_e t_f) - 2\,\mathrm{cov}(\log_e t_f, \log_e t_g) + \mathrm{var}(\log_e t_g)}{(\log_e 10)^2}.$$

Using the standard delta method, an approximate estimator of the variance of the lod score is

$$\frac{1}{(\log_e 10)^2}\left(\frac{\sigma_f^2}{t_f^2} - 2\frac{\sigma_{fg}}{t_f t_g} + \frac{\sigma_g^2}{t_g^2}\right),$$

where σ_{fg} is the estimator of the covariance term $\mathrm{cov}(t_f, t_g)$. The method of batch means provides estimates of σ_f^2 and σ_g^2 using (A.6), and σ_{fg} is defined analogously by

$$\sigma_{fg} = \sum_{b=1}^{B} \frac{(\overline{t}_{f,b} - \overline{\overline{t}}_f)(\overline{t}_{g,b} - \overline{\overline{t}}_g)}{B(B-1)}.$$

ACKNOWLEDGMENTS

This research was supported by NIH Grant GM-46255. We are grateful to Dr. Ellen Wijsman for helpful discussions regarding the data described in Section 8.

REFERENCES

BAUM, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions on Markov processes. In *Inequalities III* (O. Shisha, ed.) 1–8. Academic Press, New York.

BAUM, L. E., PETRIE, T., SOULES, G. and WEISS, N. (1970). A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Statist.* **41** 164–171.

BESAG, J., GREEN, P., HIGDON, D. and MENGERSEN, K. (1995). Bayesian computation and stochastic systems (with discussion). *Statist. Sci.* **10** 3–66.

CANNINGS, C., THOMPSON, E. A. and SKOLNICK, M. H. (1978). Probability functions on complex pedigrees. *Adv. in Appl. Probab.* **10** 26–61.

CLERGET-DARPOUX, F., BONAÏTI-PELLIÉ, C. and HOCHEZ, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42** 393–399.

DAW, E., HEATH, S. C. and WIJSMAN, E. M. (1999). Multipoint oligogenic analysis of age-of-onset data with applications to Alzheimer's disease pedigrees. *Am. J. Hum. Genet.* **64** 839–851.

ELSTON, R. C. and STEWART, J. (1971). A general model for the analysis of pedigree data. *Human Heredity* **21** 523–542.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* **85** 398–409.

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *J. Amer. Statist. Assoc.* **90** 909–920.

GUO, S. W. and THOMPSON, E. A. (1994). Monte Carlo estimation of mixed models for large complex pedigrees. *Biometrics* **50** 417–432.

HASTINGS, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

HEATH, S. C. (1997). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Amer. J. Human Genetics* **61** 748–760.

HEATH, S. C. and THOMPSON, E. A. (1997). MCMC samplers for multilocus analyses on complex pedigrees. *Amer. J. Human Genetics* **61** A278.

HODGE, S. E. and ELSTON, R. C. (1994). Lods, wrods, and mods: The interpretation of lod scores calculated under different models. *Genetic Epidemiology* **11** 329–342.

IRWIN, M., COX, N. and KONG, A. (1994). Sequential imputation for multilocus linkage analysis. *Proc. Natl. Acad. Sci. U.S.A.* **91** 11,684–11,688.

JENSEN, C. S., KJÆRULFF, U. and KONG, A. (1995). Blocking Gibbs sampling in very large probabilistic expert systems. *Int. J. Human–Computer Studies* **42** 647–666.

JENSEN, C. S. and KONG, A. (1999). Blocking Gibbs sampling for linkage analysis in large pedigrees with many loops. *Amer. J. Human Genetics* **65** 885–901.

KONG, A., COX, N., FRIGGE, M. and IRWIN, M. (1993). Sequential imputation and multipoint linkage analysis. *Genetic Epidemiology* **10** 483–488.

KONG, A., LIU, J. S. and WONG, W. H. (1994). Sequential imputations and Bayesian missing data problems. *J. Amer. Statist. Assoc.* **89** 278–288.

KRUGLYAK, L., DALY, M. J., REEVE-DALY, M. P. and LANDER, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *Amer. J. Human Genetics* **58** 1347–1363.

LANDER, E. S. and GREEN, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.* **84** 2363–2367.

LANGE, K. and SOBEL, E. (1991). A random walk method for computing genetic location scores. *Amer. J. Human Genetics* **49** 1320–1334.

LEE, J. K. and THOMAS, D. C. (2000). Performance of Markov chain–Monte Carlo approaches for mapping genes in oligogenic models with an unknown number of loci. *Amer. J. Human Genetics* **67** 1232–1250.

LEVY-LAHAD, E., WASCO, W., POORKAJ, P., ROMANO, D. M., OSHIMA, J., PETTINGELL, W. H., YU, C. E., JONDRO, P. D., SCHMIDT, S. D., WANG, K. et al. (1995a). Candidate gene for the Chromosome 1 familial Alzheimer's disease locus. *Science* **269** 973–977.

LEVY-LAHAD, E., WIJSMAN, E. M., NEMENS, E., ANDERSON, L., GODDARD, K. A., WEBER, J. L., BIRD, T. D. and SCHELLENBERG, G. D. (1995b). A familial Alzheimer's disease locus on Chromosome 1. *Science* **269** 970–973.

LIANG, K.-Y., RATHOUZ, P. J. and BEATY, T. H. (1996). Determining linkage and mode of inheritance: Mod scores and other methods. *Genetic Epidemiology* **13** 575–593.

LIU, J., WONG, W. H. and KONG, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

MENDEL, G. (1866). Experiments in plant hybridisation. (Mendel's original paper in English translation, with a commentary by R. A. Fisher, J. H. Bennett, ed., was published by Oliver and Boyd, Edinburgh, 1965.)

MORTON, N. E. (1955). Sequential tests for the detection of linkage. *Amer. J. Human Genetics* **7** 277–318.

O'CONNELL, J. R. and WEEKS, D. E. (1995). The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genetics* **11** 402–408.

SATAGOPAN, J. M., YANDELL, B. S., NEWTON, M. A. and OSBORN, T. C. (1996). A Bayesian approach to detect quantitative trait loci using Markov chain Monte Carlo. *Genetics* **144** 805–816.

THOMAS, A., GUTIN, A., ABKEVICH, V. and BANSAL, A. (2000). Multilocus linkage analysis by blocked Gibbs sampling. *Statist. Comput.* **10** 259–269.

THOMPSON, E. A. (1974). Gene identities and multiple relationships. *Biometrics* **30** 667–680.

THOMPSON, E. A. (1981). Pedigree analysis of Hodgkin's disease in a Newfoundland genealogy. *Ann. Human Genetics* **45** 279–292.

THOMPSON, E. A. (1994a). Monte Carlo estimation of multilocus autozygosity probabilities. In *Computing Science and Statistics. Proc. 26th Symposium on the Interface* (J. Sall and A. Lehman, eds.) 498–506. Interface Foundation of North America, Fairfax Station, VA.

THOMPSON, E. A. (1994b). Monte Carlo likelihood in genetic mapping. *Statist. Sci.* **9** 355–366.

THOMPSON, E. A. (2000a). MCMC estimation of multi-locus genome sharing and multipoint gene location scores. *Internat. Statist. Rev.* **68** 53–73.

THOMPSON, E. A. (2000b). *Statistical Inference from Genetic Data on Pedigrees*. IMS, Beachwood, OH.

THOMPSON, E. A. and GUO, S. W. (1991). Evaluation of likelihood ratios for complex genetic models. *IMA J. Math. Appl. Med. Biol.* **8** 149–169.

THOMPSON, E. A. and HEATH, S. C. (1999). Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics* (F. Seillier-Moiseiwitsch, ed.) 95–113. IMS, Hayward, CA.

UIMARI, P. and HOESCHELE, I. (1997). Mapping-linked quantitative trait loci using Bayesian analysis and Markov chain Monte Carlo algorithms. *Genetics* **146** 735–743.