

MCMC Multilocus Lod Scores: Application of a New Approach

Andrew W. George^{a, b} Ellen M. Wijsman^{c, d} Elizabeth A. Thompson^{a, d}^aDepartment of Statistics, University of Washington, Seattle, Wash., ^bProgram in Public Health Genetics, University of Iowa, Iowa City, Iowa, ^cDivision of Medical Genetics, Department of Medicine, University of Washington, ^dDepartment of Biostatistics, University of Washington, Seattle, Wash., USA

Key Words

Alzheimer's disease · Extended pedigrees · Linkage analysis · Markov chain Monte Carlo · Multipoint lod scores

Abstract

On extended pedigrees with extensive missing data, the calculation of multilocus likelihoods for linkage analysis is often beyond the computational bounds of exact methods. Growing interest therefore surrounds the implementation of Monte Carlo estimation methods. In this paper, we demonstrate the speed and accuracy of a new Markov chain Monte Carlo method for the estimation of linkage likelihoods through an analysis of real data from a study of early-onset Alzheimer's disease. For those data sets where comparison with exact analysis is possible, we achieved up to a 100-fold increase in speed. Our approach is implemented in the program *lm_bayes* within the framework of the freely available MORGAN 2.6 package for Monte Carlo genetic analysis (<http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml>).

Copyright © 2005 S. Karger AG, Basel

Introduction

The lod score is widely used for gene mapping. The use of lod scores for linkage analysis of family data was established by Morton (1955), and was defined as the \log_{10} of the ratio of the likelihoods of two models. These models differ only in whether the recombination fraction, θ , between the two loci in the analysis reflects linkage ($\theta < 1/2$) or free recombination ($\theta = 1/2$); all other parameters of the genetic model are held fixed. Data on multiple linked genetic markers are now routinely collected for genetic studies. To use the available information more effectively, multilocus lod scores are often calculated, with the free parameter in the likelihood ratio being the location of one of the loci relative to a fixed map of the remaining loci. LIPED [1] and LINKAGE [2] were the first widely available computer programs that allowed computation of two-point and multipoint lod scores, respectively. However, despite significant advances in computer speed and improved algorithms [3, 4], the calculation of multilocus likelihoods remains a computational bottleneck for data analysis, particularly for larger pedigrees.

Two algorithms exist for the exact calculation of multilocus likelihoods. The Elston-Stewart algorithm [5, 6] can compute likelihoods on large families but is restricted to data at a very few loci. On simple pedigrees, the computational burden of the algorithm increases linearly with family size but exponentially with the number of loci. Conversely, the Lander-Green algorithm [7] calculates

exact likelihoods on many loci jointly, but is restricted to small families. The computational complexity of this algorithm increases linearly with the number of loci and exponentially with family size.

Several programs based on each algorithm have been developed. The earlier programs were based on the Elston-Stewart algorithm [5], with more recent modifications and implementations using allele and genotype recoding schemes [4, 8, 9] or fuzzy inheritance [4] to accelerate the likelihood computations. The Lander-Green algorithm was first implemented in the computer package GENE-HUNTER [10]. Later revisions [11] and other implementations [12, 13] allow faster calculation of multilocus likelihoods on slightly larger families. Despite impressive advances in computations for multipoint analyses, exact likelihood computation for data at a large number of marker loci remains infeasible on families of arbitrary size and complexity.

An attractive alternative to exact computation is the estimation of multilocus likelihoods using Markov chain Monte Carlo (MCMC). MCMC procedures, such as the Metropolis-Hastings algorithm [14] and the Gibbs sampler [15], are methods for generating dependent samples from high-dimensional probability distributions. The samples are realized from a Markov chain that has the distribution of interest as its stationary distribution. These realizations can then be used to construct an estimate of the likelihood.

Several approaches for MCMC estimation of multilocus lod scores have been proposed. The first is due to Lange and Sobel [16], and realizes latent marker genotypes conditional on the observed marker data using a Metropolis algorithm. Given each realized multilocus marker genotype, the exact probability of the trait data is calculated, for each trait model and each hypothesized position of the trait locus. The likelihood estimate, at a given location and trait model, is the average of these probabilities over the successive realizations of the multilocus marker genotypes. Sobel and Lange [17] improved this approach by sampling genetic descent graphs of the markers conditional on the observed marker data, but a feature of this approach remains that the trait data do not enter into the MCMC sampling. This can be quite efficient when the trait itself conveys little information. However, when the trait inheritance pattern is strong, failure to use the trait data in MCMC sampling leads to poor MCMC performance [18].

Lin [19] and Thompson [18] instead adopted the approach of Thompson and Guo [20] for the Monte Carlo estimation of likelihood ratios. Here the latent inheri-

tance patterns at both trait and marker loci are sampled conditional on both trait and marker data, under a fixed genetic model. From a set of realizations from a given MCMC under given parameter values, the complete local likelihood surface relative to that at the simulation parameters can be estimated. However, since there is often strong evidence against a trait locus being tightly linked to a marker, the lod score often has sharp decreases near marker locations. This makes almost impossible the use of the local likelihood-ratio method to estimate accurately the absolute value of the lod score at positions within a multimarker map [18].

George and Thompson [21] proposed a new MCMC method for producing estimates of multilocus lod scores. Their pseudo-Bayesian procedure recovers a likelihood estimate from a set of posterior probabilities of trait locus locations, λ . In a pseudo-Bayes approach [22], a prior distribution is placed on a parameter (in this case, λ), and this parameter is sampled in the MCMC process. However, this prior distribution is simply a tool to enhance MCMC performance: the choice of prior does not affect the likelihood to be estimated. The method of [21] combines the exact computation of trait data probabilities of [17] with the sampling conditional on both trait and marker data of [18], and incorporates several recent advances in MCMC methodology.

In [21] the basic validity of the method and mixing performance of the MCMC was demonstrated through application to data on three loosely linked marker loci on two extended pedigrees of simple structure. Only three markers were used so that the resulting estimated lod scores could be compared with those produced by VITASSE [4]. Although the pedigree structures and genetic markers derived from a real study [23], the trait data were 'invented' to mimic the known reality of linkage in one pedigree and absence of linkage in the other. No trait model was incorporated into these illustrative analyses: the trait genotypes of certain individuals were assumed known. Two estimators of the multipoint lod scores were proposed by [21]. First, a naive estimator of the posterior probability distribution of location was obtained from the empirical distribution of realized values. Second, a Rao-Blackwellized form [24] was obtained. A comparison of the computational efficacy of these two estimators was given by [21], and the second found to be preferable despite the extra computation it entails.

In this paper, we provide an assessment of the methods of [21] through the analysis of real data from a study of Alzheimer's disease. The methods have been further developed to incorporate a variety of qualitative and quan-

titative trait models: here we use a penetrance function incorporating a logistic form for age-of-onset. We use a set of pedigrees of varying size and complexity, and varying in the extent of data and the pattern of missing data. Thus we have a much wider basis for comparison of the new method with exact computational methods. We use both tightly-linked and loosely-linked markers in comparisons with VITESSE [4], and use the full set of 10 markers in comparisons with GENEHUNTER [10]. The latter comparisons are, of course, possible only for the smallest pedigrees. However, on some of the larger pedigrees, knowledge of the underlying disease mutation also allows assessment of the performance of the methods. We provide not only evidence for the speed and accuracy of the method, but show also that MCMC can provide order-of-magnitude speed-up over exact methods, even when the latter are feasible.

Methods

In this section, we describe the data, and the trait model under which multilocus linkage likelihoods are to be estimated. We then provide details of the implementation of the pseudo-Bayes method of [21] that will achieve MCMC estimation of these multipoint linkage likelihoods. Finally, in this section, we discuss the use of the expected complete-data log-likelihood (ECDLL) as a diagnostic to compare MCMC runs which provide divergent likelihood estimates.

The Pedigrees and the Trait

To test the performance of the pseudo-Bayes approach to MCMC estimation of linkage likelihoods, we analyze data from a study of familial Alzheimer's disease (AD). In this data set, the causative presenilin 2 (PS2) mutation has been mapped to chromosome 1 [23] and identified [25]. The data used here includes trait and marker information on six families (HB, HD, R, KS, W, WFL), varying in size (table 1) and pedigree complexity [23, 26, 27]. The pedigrees KS, R, WFL and W have a simple genetic structure, all individuals descending from a single founder couple over 2 to 5 generations. Each of the pedigrees HB and HD has a loop formed from the marriage of cousins. The HB pedigree has a depth of 6 generations, while the 5-generation HD pedigree descends from two original founder couples.

Due to the late onset of AD, many individuals were deceased and unavailable for sampling (table 1). Information was collected on the disease status and age at onset (if affected) or age at last examination (if unaffected). For some individuals, generally those in earlier generations, disease status is unknown. From previous studies [25] it is known that families HB, HD, R, and WFL have a mutation in the PS2 gene on Chromosome 1, but that families KS and W do not carry this mutation. This study was approved by the University of Washington institutional review board, and informed consent was obtained from each participant.

In our analyses, we assume AD is an autosomal dominant disease with age-dependent penetrances based upon the logistic distri-

Table 1. Summary of the AD family data used

Family data	AD data			Marker data n obs.			
	pedigree	size	gen		aff	unaff	unobs
HB	50	6	13	28	9	60.6	27
HD	41	5	14	17	10	52.2	14
R	53	4	17	30	6	50.8	31
KS	53	5	11	36	6	65.5	27
WFL	21	3	6	14	1	63.8	15
W	6	2	4	2	0	59.8	4

Given are the pedigree names, size, and number of generations (gen). For the AD trait, given are the numbers of affected (aff) and unaffected (unaff), the number for which AD status is unknown (unobs), and the mean onset age of affected individuals (onset). Also given is the count of individuals for whom marker data are available.

bution. A frequency of 0.05 is assumed for the disease predisposing allele D . In pedigree j , the probability of individual i of age a_{ij} being affected given disease genotypes DD or Dd is

$$P(Y_{ij} = \text{Affected} | DD) = P(Y_{ij} = \text{Affected} | Dd) = \frac{1}{(1 + e^{-(a_{ij} - \mu_j)/\beta})}$$

where Y_{ij} denotes the trait value for this individual. The parameter μ_j is the sample mean age of the affected individuals in pedigree j and

$$\beta = \sqrt{\frac{3S^2}{\pi}},$$

where $S^2 = 259.7$ is the sample variance of the onset ages of the affected individuals over all the pedigrees.

To account for sporadic cases, the penetrance for the dd genotype is again based on the cumulative logistic distribution with median onset μ taken as 90 years. That is, the probability of individual i from pedigree j and not carrying the disease gene being affected with AD is

$$P(Y_{ij} = \text{Affected} | dd) = \frac{1}{(1 + e^{-(a_{ij} - 90)/\beta})}$$

For the value of β used, the value $\mu = 90$ provides an onset probability of about 15% by age 75 for non-gene carriers.

Since the programs VITESSE [4] and FASTLINK [3] use discrete liability classes to accommodate age-dependent penetrances, we did the same for our MCMC analyses. The penetrance range was divided into 10 intervals, bounded by values (0, 0.1, 0.2, 0.35, 0.5, 0.65, 0.8, 0.85, 0.9, 0.95, 1.0). These values were chosen to ensure a range of liability classes within each pedigree. The actual penetrance value used was the midpoint of the range: for example 0.725 for the interval (0.65, 0.8). A liability class was assigned to each affected individual on the basis of the value of the pedigree-specific logistic curve at that individual's age-of-onset. For unaffected individuals, the same procedure was followed, except that the penetrance intervals were bounded by the values (0, 0.003,

Table 2. The ten chromosome 1 markers used in the multipoint linkage analyses of AD

Index	Marker	Map position cM	Number of alleles
1	D1S306	0.00	12
2	D1S249	5.78	15
3	D1S245	14.22	10
4	D1S237	20.69	13
5	D1S229	27.49	8
6*	D1S479	34.22	11
7	D1S446	49.52	13
8	D1S235	53.66	9
9	D1S180	75.84	11
10	D1S102	90.62	6

Positions are given as distances in Haldane centiMorgans relative to the first marker used.

* Denotes the marker most closely linked to the PS2 locus.

0.008, 0.016, 0.031, 0.056, 0.112, 0.153, 0.222, 0.376, 1.0). Using the logistic curve for the non-gene carriers, the first interval corresponds to individuals under 40, and the last to those over 85.5. Again the midpoint value was the one used, so the maximum affection probability for non-gene carriers (those over 85.5) is 0.688.

The Marker Data and Lod Score Analyses

In our multilocus linkage analyses of the AD family data, we compute map-specific lod scores. That is, the marker map is assumed known. We use marker data on 10 linked genetic markers on chromosome 1 (table 2). These markers are approximately evenly spaced along a 90-cM chromosome segment surrounding marker D1S479, which is closely linked to the PS2 locus. The sex-averaged map is used, and map distances in table 2 have been converted to Haldane centiMorgans, since the Haldane map function is used for our multipoint analyses. Each marker has between 6 and 15 possible alleles.

Data on each family are analyzed separately using different marker subsets (MS) summarized in table 3. The marker D1S479 closest to the PS2 locus was included in all marker subsets. The full set of all 10 markers (MS-A) can only be used in the MCMC analyses or on very small pedigrees. Two 3-marker subsets were chosen to allow comparison of MCMC and exact results for most of the pedigrees. One 3-marker subset typifies tight linkage (MS-T) and the other loose linkage (MS-L). The marker set MS-L was used in previous testing of our method [21], but that study considered only two pedigrees of simple structure and with no trait model, making the unrealistic assumption that genotypes at the trait locus are observable. Here, both the sets MS-T and MS-L were used for comparison because tight linkage is a particular challenge for the MCMC sampling of latent inheritance patterns, and hence for the estima-

Table 3. The marker subsets used in the multipoint linkage analyses of AD

Label	Description	Number of markers	Marker set
MS-A	all markers	10	see table 2
MS-T	tightly linked markers	3	D1S229 D1S479 D1S446
MS-L	loosely linked markers	3	D1S306 D1S479 D1S102
MP-T1	tight linkage: pair 1	2	D1S229 D1S479
MP-T2	tight linkage: pair 2	2	D1S479 D1S446
MP-L1	loose linkage: pair 1	2	D1S306 D1S479
MP-L2	loose linkage: pair 2	2	D1S479 D1S102

tion of multilocus lod scores [18]. A comparison of performance for MS-L and MS-T, under a realistic trait model and on a variety of pedigree structures and missing-data patterns, is therefore of interest. For the larger HB and HD pedigrees, each having a loop, exact computation is possible only for three loci jointly: a marker pair (MP) and the trait locus. For comparison of exact and MCMC computations on these pedigrees, each of the 3-marker subsets MS-T and MS-L was divided into two pairs of adjacent markers (table 3).

All lod scores were estimated at locations within and external to the marker map. For the 10-marker MS-A data, lod scores were obtained at 3 positions within each interval: at 10, 50 and 90% of the map length of the interval. For the 3-marker subsets MS-T and MS-L, 9 equally-spaced positions (in terms of Haldane cM) were used in each interval. Outside the linkage group, lod scores were computed for hypothesized trait locations at recombination fractions 0.05, 0.2, 0.3, 0.4 and 0.45 from the nearest (first or last) marker.

Exact multipoint lod score results were obtained for the same data, model, and hypothesized trait locus positions, using a variety of available software programs. Exact four-locus lod scores were calculated for the KS, R, W and WFL using VITESSE [4]. Exact 11-point lod scores were also computed on the W pedigree and a reduced WFL pedigree using GENEHUNTER [10]. VITESSE could not be used on the HB and HD families, due to their more complex structure. Therefore, FASTLINK [3, 28] was used to compute exact three-locus (2-marker) lod scores for these two pedigrees. Both MCMC and exact computations were performed under Linux using a single Pentium III processor running at 933 MHz.

The MCMC Lod Score Estimator

In this paper we estimate map-specific lod scores by the MCMC method proposed by [21]. A fixed trait and marker model is assumed: we index this model by parameters ψ . The linkage likelihood is to be estimated, under the model ψ , at a set of $K + 1$ discrete alternative hypothesized positions λ of the putative trait locus. We label the positions $\lambda = 0, 1, 2, \dots, K$. The position $\lambda = 0$ denotes an unlinked trait locus and $\lambda = 1, 2, \dots, K$ are K arbitrary fixed positions linked to or within the marker map.

As in other implementations of MCMC on pedigrees [17, 29], the primary latent variables to be sampled are the inheritance vectors [7] at trait and marker loci. We denote by \mathbf{S} the full set of inheritance vectors, and the array only for marker loci by \mathbf{S}_M . If \mathbf{S}_T denotes the inheritance vector for the hypothesized trait locus, $\mathbf{S} = (\mathbf{S}_M, \mathbf{S}_T)$. Correspondingly, let \mathbf{Y}_M be the marker data for the set of

linked markers, and \mathbf{Y}_T be the trait data. The likelihood, $L(\lambda)$ for the trait location λ is the probability of observing data $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$, and the lod score at location $\lambda = x$ is

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

The objective of a pseudo-Bayesian approach is to improve MCMC performance by sampling not only the latent variables \mathbf{S} but also the parameter λ . A prior distribution is placed on the $K + 1$ discrete values of λ . Then, for a fixed value of ψ , MCMC is used to sample n dependent realizations $(\mathbf{S}^{(m)}, \lambda^{(m)})$, $m = 1, \dots, n$, from the joint posterior distribution $\pi_\psi(\mathbf{S}, \lambda|\mathbf{Y})$. Given an estimate of the marginal posterior distribution $\pi_\psi(\lambda|\mathbf{Y})$, the likelihood may be recovered by dividing by the prior distribution on λ :

$$L(\lambda) = P_\psi(\mathbf{Y}|\lambda) \propto \frac{\pi_\psi(\lambda|\mathbf{Y})}{\pi(\lambda)}$$

where $\pi(\lambda)$ is the prior distribution on λ . Any choice of strictly positive $\pi(\lambda)$ is valid: the same lod score estimate (1) is ultimately obtained. Thus the pseudoprior is simply an MCMC tool, chosen to enhance MCMC performance.

Rather than using the naive estimator of the posterior distribution $\pi_\psi(\lambda|\mathbf{Y})$, we use a Rao-Blackwellized form. George and Thompson [21] have shown that a Rao-Blackwellized estimator of $L(x) = P_\psi(\mathbf{Y}|\lambda = x)$ is given by

$$L_{RB}(x) = \sum_{m=1}^n \left(\frac{P_\psi(\mathbf{Y}_T|\mathbf{S}_M^{(m)}, \lambda = x)}{\sum_{x'=0}^K P_\psi(\mathbf{Y}_T|\mathbf{S}_M^{(m)}, \lambda = x') \pi(\lambda = x')} \right). \quad (2)$$

Although this expression appears complex, it is simply a normalized probability of trait data, given the trait locus location and the complete inheritance pattern at all marker loci. It may be evaluated by a single-locus pedigree peeling [5] with transmissions generalized to accommodate the known inheritance at the two adjacent marker locations. In [21], we compared performance of the naive and Rao-Blackwellized forms of the estimator of $L(\lambda)$: here we use only L_{RB} of equation (2).

Each cycle of the MCMC sampling procedure consists of three steps: details are given by [21]. Suppose $(\mathbf{S}^{(m)}, \lambda^{(m)})$ is the present state of the Markov chain. Then a move to the $(m + 1)$ th state is accomplished via the following three steps.

Step 1: Sample a new set of meiosis indicators \mathbf{S} using a sequence of block-Gibbs updates which use either the locus (L) sampler of [30] or the meiosis (M) sampler of [31].

Step 2: Jointly sample a new position λ and set of meiosis indicators \mathbf{S}_T for the trait locus using a Metropolis-Hastings (M-H) algorithm.

Step 3: Propose a restart configuration (\mathbf{S}, λ) using sequential imputation [32] and a M-H algorithm.

After these three steps have been completed, $(\mathbf{S}^{(m+1)}, \lambda^{(m+1)})$ is realized and the process is repeated n times.

Implementation Issues

Specifying the Pseudo-Prior, $\pi(\lambda)$

Note that if $\pi(\lambda)$ were uniform over λ , the posterior distribution $\pi_\psi(\lambda|\mathbf{Y})$ would be proportional to the likelihood. Hence, if we choose the prior on λ to be approximately inversely proportional to the likelihood, then an approximately uniform posterior probability

distribution results. Such approximately uniform posterior sampling of trait positions λ enhances the accuracy of the MCMC estimate of the lod score. To accomplish this, the MCMC procedure is first run using a discrete uniform prior distribution on λ . In this preliminary analysis of the data, trait positions of low likelihood may not even be sampled. However, a very crude estimate of the likelihood $L(\lambda)$ is still obtained using an estimator analogous to equation (2). If $\pi(\lambda)$ is then calculated to be proportional to the inverse of this preliminary estimate of $L(\lambda)$, each trait position should be sampled with approximately equal frequency in the main analysis of the data. The values of $\pi(\lambda)$ were restricted to vary over no more than two orders of magnitude. It is important that no location be accorded a value of $\pi(\lambda)$ that is too extreme, particularly so for the value $\pi(0)$ for the unlinked location.

Determining Run-Length n

In any application of MCMC, it remains a challenge to determine whether an adequate number of realizations have been sampled to ensure accurate estimates. The required lengths of an MCMC run depends on the number of markers and size of the pedigree. The preliminary run must be sufficient to provide a good estimate of the pseudo-prior, so that in the main run the possible trait positions are sampled with approximately equal frequency. The initial length of the main run was double that of the preliminary run.

Many of the commonly used MCMC convergence diagnostics are restricted to continuous variables. Here, we instead determine the main run-length n through a variety of ad hoc approaches. First, using the same starting configuration, we gradually increase the value of n , until the lod score estimates stabilize. Second, we perform parallel MCMC runs using different randomly generated starting configurations, to analyze the variation of lod score estimates across runs and to ensure that the variation among runs of the estimated lod score at each location was within acceptable bounds. For this paper, four parallel independent runs were performed on each component pedigree. Last, we examine plots of the sequence of realizations of trait locations $\lambda^{(m)}$ and inheritance vectors $\{\mathbf{S}_{ij}^{(m)}\}$ for each locus j (including the trait locus T). Systematic patterns of values in these plots may suggest poor MCMC performance and a run-length that is too short. The actual run lengths used are given in the Results section.

Comparing MCMC Runs

A useful tool in assessing the output of an MCMC run is the expected complete-data log-likelihood (ECDLL) conditional on the data \mathbf{Y} [18, 33]. The complete-data likelihood is the probability of both data and latent variables, $P_{\psi,\lambda}(\mathbf{Y}, \mathbf{S})$. In the absence of genetic interference and allelic association among loci, the complete-data log-likelihood may be written

$$\log P_{\psi,\lambda}(\mathbf{Y}, \mathbf{S}) = \log P_\psi(\mathbf{Y}_M|\mathbf{S}_M) + \log P_\psi(\mathbf{Y}_T|\mathbf{S}_T) + \log P_\psi(\mathbf{S}_M) + \log P_\lambda(\mathbf{S}_T|\mathbf{S}_M) \quad (3)$$

where the first three terms depend on those parts of ψ specifying marker model, trait model, and marker map respectively, and the fourth term depends on the trait location λ . The first and third terms can be further partitioned by marker locus and by marker interval respectively [34]. For a given realized $(\mathbf{S}^{(m)}, \lambda^{(m)})$, each term of equation (3) is already computed in performing the MCMC step providing this realization. Thus an MCMC estimate of the expected value of each term conditional on \mathbf{Y} is given by averaging the values over

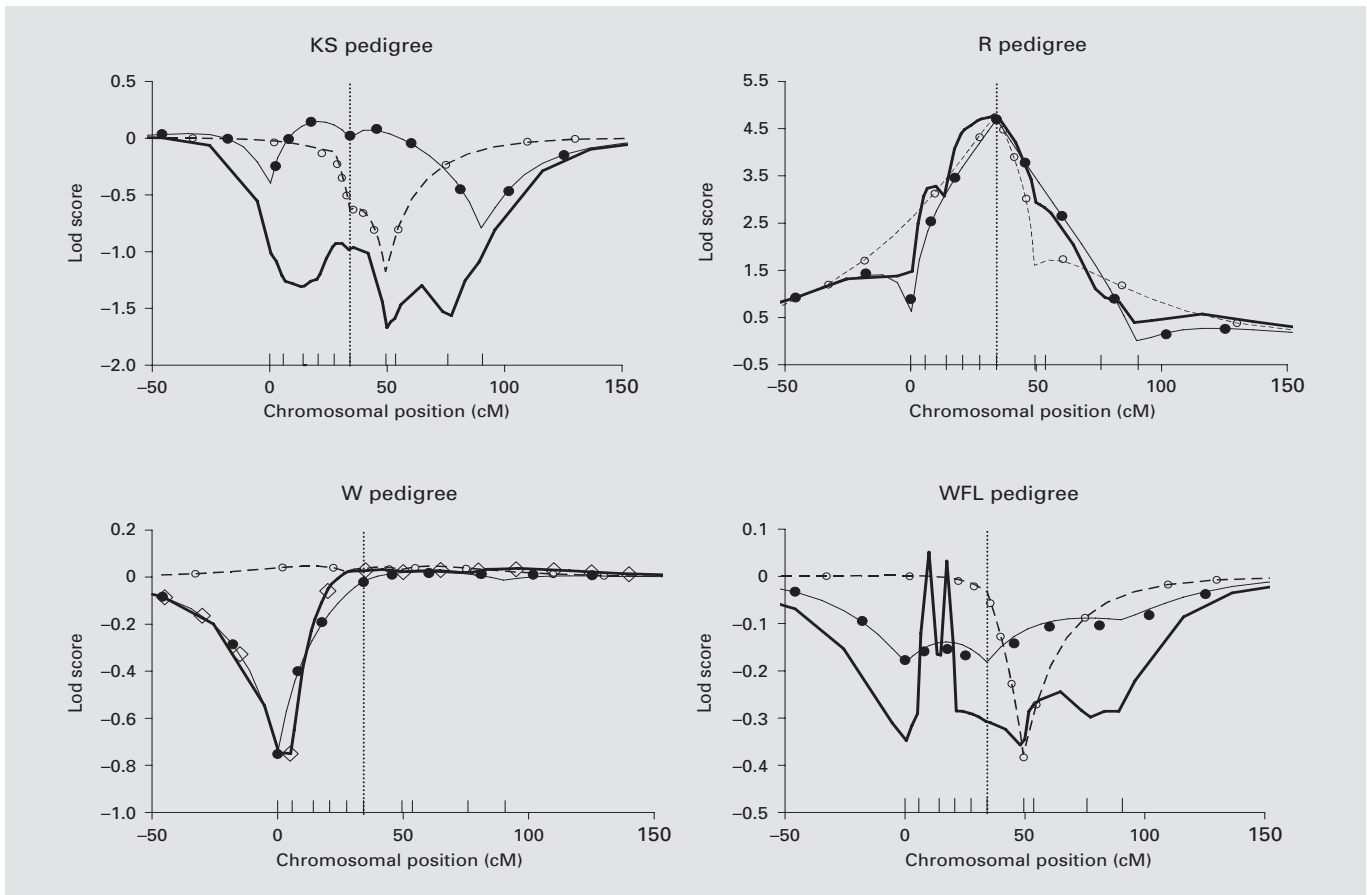


Fig. 1. Exact and estimated lod scores at hypothesized trait locations for the analysis of data on families KS, R, W and WFL. Marker positions are denoted by the short vertical lines on the horizontal axis. The vertical dashed line denotes the known position of the disease mutation. The thin solid line is the lod score curve estimated using *lm_bayes* with loosely linked markers MS-L. The thin dashed line is the lod score curve estimated using *lm_bayes* with markers MS-T. The thick solid line is the lod score curve estimated using *lm_bayes* using all ten markers (MS-A). The closed (open) circles show the exact lod scores calculated by VITESSE using data at marker set MS-L (MS-T), respectively. In the case of the W family, the open diamonds are the exact lod scores calculated by GENEHUNTER using the MS-A marker set.

the n realizations of the MCMC run. Since, for given data \mathbf{Y} , $P_{\psi,\lambda}(\mathbf{S}|\mathbf{Y}) \propto P_{\psi,\lambda}(\mathbf{Y}, \mathbf{S})$, the ECDLL provides an estimate of the relative magnitude of the conditional probability of latent variables \mathbf{S} given data \mathbf{Y} in the part of the space sampled by different MCMC runs. It tells us which runs are sampling in 'better' parts of the space. Our Results section will show an example of use of this ECDLL comparison among runs.

Software

Our method has been implemented in the program *lm_bayes* within the framework of the MORGAN 2.6 package for Monte Carlo genetic analysis (<http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml>). Our MORGAN MCMC programs also routinely provide the ECDLL statistics described above, separating the terms as described in equation (3).

Results

Here we present the results for the multilocus linkage analysis of data on the six AD families. Results are first presented for the analyses of data on families KS, R, WFL and W, which have simple pedigree structures, followed by results for the analysis of data on families HB and HD which have complex pedigree structures.

Exact and Monte Carlo lod scores for a 4-point linkage analysis of data on pedigrees KS, R, W and WFL are shown in figure 1. For clarity, only the results from a single MCMC run are presented since for these pedigrees estimated multipoint lod scores agreed closely across four

Table 4. Run lengths (in units of 1000 MCMC scans) and CPU times (in minutes) for the analysis of data on families KS, R, W and WFL

Pedigree	MS-L			MS-T			MS-A	
	Bayes		VSSE	Bayes		VSSE	Bayes	
	length	time	time	length	time	time	length	time
KS	10:20	12.8	292.9	8:20	15.0	1,156.8	50:100	90.5
R	1.5:3	2.7	62.0	3:7	4.9	41.0	35:70	56.5
W	0.2:0.4	0.1	0.1	0.2:0.5	0.1	0.1	1:2	0.4
WFL	2:4	0.8	0.6	2:5	1.0	0.3	3:5	1.9

Data are analyzed using *lm bayes* (Bayes) and VITESSE (VSSE) for the 3-marker subsets MS-L and MS-T. Analysis of data using all 10 markers (MS-A) was only possible using *lm bayes*. Run lengths are given for both the preliminary and the main run. For example 50:100 denotes a preliminary MCMC run of 50,000 scans to estimate the pseudo-prior, followed by a main run of 100,000 scans.

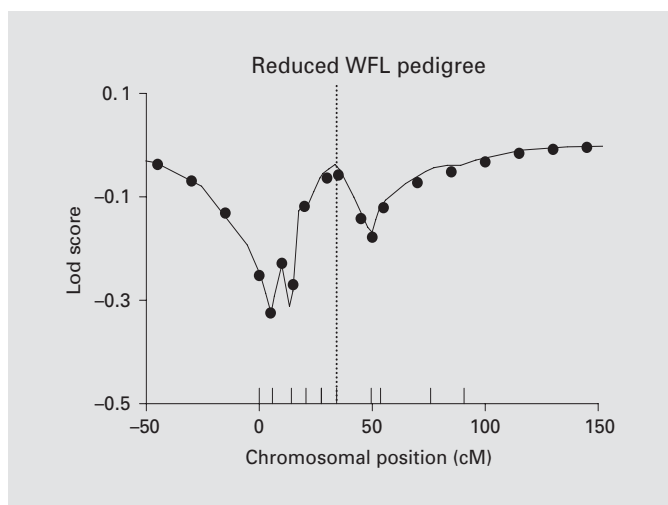


Fig. 2. Exact and estimated 10-marker (MS-A) lod scores at hypothesized trait locations for the analysis of data on the WFL family where two final individuals have been removed. Marker positions are denoted by the short vertical lines on the horizontal axis. The vertical dashed line denotes the known position of the disease mutation. The thick solid line is the lod score curve estimated using *lm bayes*. The circles are exact lod scores calculated by GENEHUNTER.

independent MCMC runs with different starting configurations and would be indistinguishable in the figure. Also, in all cases, exact lod scores were computed for the same hypothesized trait locations as Monte Carlo lod scores were estimated, but for clarity only a subset of the exact computation values are shown in the figures. The lod scores estimated via the MCMC procedure are indistinguishable from the exact lod scores. Further, we do not

see a degradation in the MCMC-based estimates of lod scores using tightly linked markers, although longer MCMC runs are required for MS-T than for MS-L (table 4).

Also shown in figure 1 are the estimated 11-point lod scores (MS-A). Comparison with exact results is only possible for pedigree W. We were also able to compute exact 11-point lod scores on the WFL family using GENEHUNTER if two individuals in the final generation were removed from the analysis. The exact and estimated lod scores on this reduced WFL pedigree are shown in figure 2.

In table 4, run times for the analysis of data on families KS, R, WFL and W are presented. These times are for a single total run-length, including any additional MCMC scans required to stabilize the lod-score estimate, but do not include times for preliminary analyses to tune the sampler. For the smaller families W and WFL, *lm bayes* and VITESSE have comparable run times. However, there is between a 10-fold and 100-fold reduction in run times when *lm bayes*, as opposed to VITESSE, is used for lod score computation on the larger pedigrees KS and R. Note that larger pedigrees require more MCMC scans. With our current software, if multiple pedigrees are analyzed together in a single MCMC run, all will receive the same number of MCMC scans. Currently we therefore recommend including only pedigrees of roughly comparable size within a single run.

Using all 10 marker loci (MS-A), 11-point lod score computation on the nuclear family W is faster with GENEHUNTER which takes 3 s compared to a run time of 25 s for *lm bayes*. However, for the analysis of data on the reduced 3-generation WFL pedigree under marker set

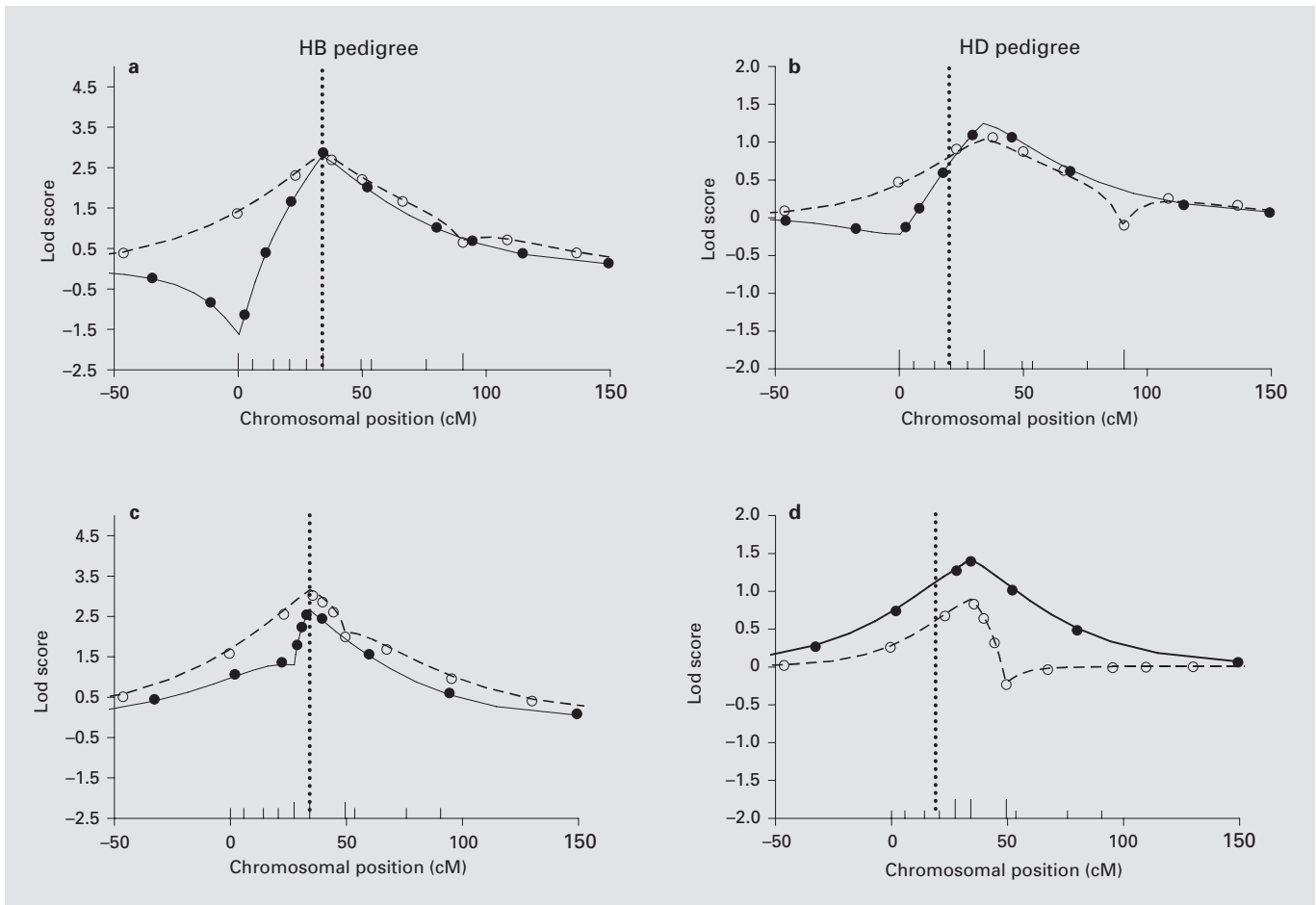


Fig. 3. Exact and estimated 3-point lod scores at hypothesized trait locations for the analysis of data on the HB and HD families. Tick marks show the marker locations, the larger tick marks showing the markers used in these analyses. The vertical dashed line denotes the known position of the disease mutation. In **a** and **b**: the solid line is the lod score curve estimated given by *lm_bayes* using the loosely linked marker pair MP-L1, and the dashed line is the lod score using MP-L2. The closed and open circles show the exact lod scores calculated using FASTLINK with marker pairs MP-L1 and MP-L2, respectively. In **c** and **d**, the details are the same except that the tightly linked marker pairs MP-T1 and MP-T2 are used.

MS-A, *lm_bayes* had a run time of 49.3 min compared to a run time of 85 min for GENEHUNTER.

Analyses of data on families HB and HD are challenging due to the presence of loops. Only 3-point analysis is possible using exact methods. Exact and MCMC lod scores for 3-point linkage analyses on these two pedigrees are shown in figure 3. Once again, there is very little difference between the estimated and exact lod scores. However, *lm_bayes* is considerably faster than FASTLINK for most 3-point analyses (table 5).

For the 11-point MCMC analyses on the HB and HD pedigrees, results are consistent with the known PS2 loca-

tion. Four replicate analyses of the HD family gave very similar lod score estimates (fig. 4), but, for the HB pedigree, lod scores do differ across four replicate analyses. Two of these latter runs gave maximum lod scores of approximately 2.5, near Marker D1S235 (marker 8 at position 54 cM: see table 2) while the other two runs gave maximum lod scores of approximately 3.4, close to marker D1S479. However, in all cases, the correct PS2 gene location had a lod score within one of the maximum value.

To assess the validity of the different lod score estimates, we estimate the expected complete data log likeli-

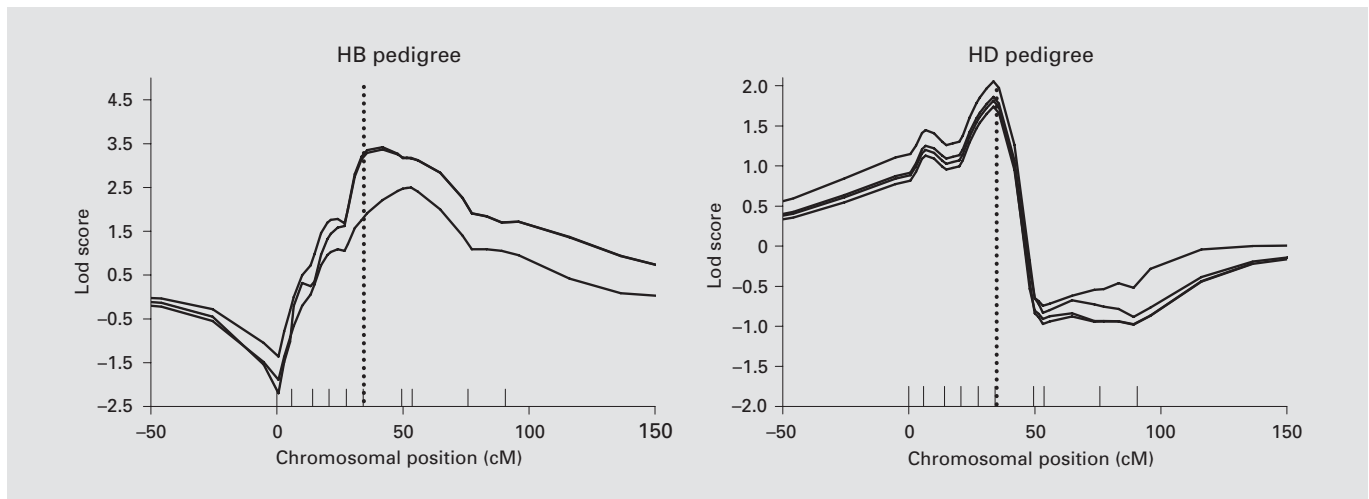


Fig. 4. Results from four independent MCMC analyses of data on the HB and HD families. Each solid line is a lod score curve estimated by *lm_bayes* using all 10 markers (MS-A). Marker positions are denoted by the short vertical lines on the horizontal axis. The vertical dashed line denotes the known position of the disease mutation.

Table 5. Run times (in minutes) for the three-locus (2-marker) linkage analyses of data on families HB and HD

Marker pair	HB pedigree			HD pedigree		
	Bayes		FSTLNK	Bayes		FSTLNK
	length	time	time	length	time	time
MP-L1	20:40	31.2	257.8	8:18	6.7	201.6
MP-L2	8:16	12.1	174.2	20:40	18.5	75.7
MP-T1	60:180	96.4	362.1	300:600	172.3	158.5
MP-T2	30:90	63.9	859.6	50:100	47.9	122.3

Data are analyzed using *lm_bayes* (Bayes) and FASTLINK (FSTLNK) and the loosely and tightly linked marker pairs (MP) of table 3. The notation is as in table 4.

hood or ECDLL (equation (3)) at each hypothesized position of the trait locus within a MCMC run. Across all trait-locus positions, the ECDLL from the two MCMC runs giving maximum lod scores close to 3.4 are up to $2 \log_{10}$ units larger than the ECDLL from MCMC runs giving maximum lod scores of 2.5. This suggests that the MCMC realizations which give maximum estimated lod scores close to 3.4 have been sampled from a part of the space of latent inheritance patterns \mathbf{S} that is two orders of magnitude more probable given the observed data \mathbf{Y} .

Discussion

We have developed an MCMC procedure for the accurate estimation of multilocus likelihoods using data on pedigrees. Our estimation procedure enables linkage analyses to be conducted using data well beyond those for which exact computational methods are feasible. We use joint Gibbs-sampler updating of latent variables across loci and meioses, joint updating of the position and set of meiosis indicators for the trait locus, and M-H restarts and Rao-Blackwellized estimates [21].

We assessed the accuracy of our procedure through comparison with exact results. Unlike Bayesian implementations of MCMC, our method estimates the multilocus lod score curve. Thus, we were able to use well-established exact methods to benchmark the performance characteristics of our approach. The results are highly promising, showing both that the MCMC results agree well with exact results, and that there are substantial gains in the magnitude of the lod score when multiple markers are used in the analysis. To our knowledge, this is the first paper to compare multilocus lod scores estimated using MCMC to exact lod scores using real data on large and challenging pedigrees. Our analyses of the AD family data showed that use of MCMC can provide substantial gains in speed over exact analyses, without compromising accuracy. However, where computationally feasible, exact results should be favored over Monte Carlo estimates.

MCMC is of most practical use for data sets on which exact computation is impossible, or where a speedier preliminary result is desired.

Of the Alzheimer's disease families we analyzed, the HB family proved to be the most challenging. The complex pedigree structure, large number of unobserved individuals in early generations, the pattern of observed data and tightly linked marker loci all contributed to the mixing problems evidenced by the varying lod score estimates shown in figure 4. Difficulties lie in moves between inheritance states in which the disease allele descends via the maternal side of the inbreeding loop and states in which descent is via the paternal side. New joint update strategies for resampling the latent variables have improved mixing but large moves in the latent space are still difficult to achieve.

M-H restarts is potentially a mechanism to allow large moves to be realized from the sample space. As a proposal mechanism, we used sequential imputation [32], which realizes the latent variables sequentially over loci, using only data for previous loci in the defined ordering to contribute to the imputation at each locus. Thus, sequential imputation only partially captures the information from the data that exists jointly among loci. As a result, proposals generated by sequential imputation gen-

erally have very low probabilities of acceptance when conditioning on marker data at many tightly linked loci. Alternative strategies of joint updating of multiple meioses [18, 35] or of relaxation of the proposal probability distributions [36, 37] may provide improved performance, but it remains to be investigated whether the gains outweigh the added computational costs.

MCMC procedures for calculating multilocus likelihoods and probabilities on family data continue to gain in speed and accuracy. By using joint updates that combine exact computation with Monte Carlo sampling, the performance characteristics of MCMC procedures have greatly improved. The MCMC procedures evaluated in this paper permit otherwise impossible or impractical linkage analyses of data, although additional improvements to these MCMC procedures continue to be needed.

Acknowledgments

Research supported in part by NIH grant GM46255. The original data collection was supported by NIH AG05136: we thank Drs. T. Bird and G. Schellenberg and the University of Washington Alzheimer's Disease Center for the use of the data. We are grateful to two referees for their helpful comments.

References

- 1 Ott J: Estimation of the recombination frequency in human pedigrees: Efficient computation of the likelihood for human linkage studies. *Am J Hum Genet* 1974;26:588-597.
- 2 Lathrop GM, Lalouel JM, Julier C, Ott J: Strategies for multilocus linkage analysis in humans. *Proc Natl Acad Sci USA* 1984;81:3443-3446.
- 3 Cottingham RW, Idury RM, Schäffer AA: Faster sequential genetic linkage computations. *Am J Hum Genet* 1993;53:252-263.
- 4 O'Connell JR, Weeks DE: The algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. *Nature Genet* 1995;11:402-408.
- 5 Elston RC, Stewart J: A general model for the analysis of pedigree data. *Hum Hered* 1971;21:523-542.
- 6 Cannings C, Thompson EA, Skolnick MH: Probability functions on complex pedigrees. *Adv Appl Prob* 1978;10:26-61.
- 7 Lander ES, Green P: Construction of multilocus genetic linkage maps in humans. *Proc Natl Acad Sci USA* 1987;84:2363-2367.
- 8 Braverman MS: An algorithm to improve the computational efficiency of genetic linkage analysis. *Computers Biomed Res* 1985;18:24-36.
- 9 Schellenberg GD, Bird TD, Wijsman EM, Orr HT, Anderson L, Nemens E, White JA, Bonneycastle L, Weber JL, Alonso ME, et al: Genetic linkage evidence for a familial Alzheimer's disease locus on chromosome 14. *Science* 1992;258:668-671.
- 10 Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES: Parametric and nonparametric linkage analyses: a unified multipoint approach. *Am J Hum Genet* 1996;58:1347-1363.
- 11 Markianos K, Daly M, Kruglyak L: Efficient multipoint linkage analysis through reduction of the inheritance space. *Am J Hum Genet* 2001;68:963-977.
- 12 Gudbjartsson DF, Jonasson K, Frigge ML, Kong A: Allegro, a new computer program for multipoint linkage analysis. *Nature* 2000;25:12-13.
- 13 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genet* 2002;30:97-101.
- 14 Hastings WK: Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* 1970;57:97-109.
- 15 Geman S, Geman D: Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans Pattern Analysis Machine Intelligence* 1984;6:721-741.
- 16 Lange K, Sobel E: A random walk method for computing genetic location scores. *Am J Hum Genet* 1991;49:1320-1334.
- 17 Sobel E, Lange K: Descent graphs in pedigree analysis: Applications to haplotyping, location scores, and marker-sharing statistics. *Am J Hum Genet* 1996;58:1323-1337.
- 18 Thompson EA: MCMC estimation of multilocus genome sharing and multipoint gene location scores. *Int Stat Rev* 2000;68:53-73.
- 19 Lin S: Multipoint linkage analysis via Metropolis jumping kernels. *Biometrics* 1996;52:1417-1427.
- 20 Thompson EA, Guo SW: Evaluation of likelihood ratios for complex genetic models. *I.M.A. J Math Appl Med Biol* 1991;8:149-169.

- 21 George AW, Thompson EA: Discovering disease genes: multipoint linkage analysis via a new Markov chain Monte Carlo approach. *Stat Science* 2003;18:515–531.
- 22 Geyer CJ, Thompson EA: Annealing Markov chain Monte Carlo with applications to ancestral inference. *J Am Stat Assoc* 1995;90:909–920.
- 23 Levy-Lahad E, Wijsman EM, Nemens E, Anderson L, Goddard KA, Weber JL, Bird TD, Schellenberg GD: Familial Alzheimer's disease locus on Chromosome 1. *Science* 1995;269:970–973.
- 24 Gelfand AE, Smith AFM: Sampling based approaches to calculating marginal densities. *J Am Stat Assoc* 1990;85:398–409.
- 25 Levy-Lahad E, Wasco W, Poorkaj P, Romano DM, Oshima J, Pettingell WH, Yu CE, Jondro PD, Schmidt SD, Wang K, et al: Candidate gene for the Chromosome 1 familial Alzheimer's disease locus. *Science* 1995;269:973–977.
- 26 Bird TD, Lampe TH, Nemens EJ, Miner GW, Sumi SM, Schellenberg GD: Familial Alzheimer's disease in American descendants of the Volga Germans: probably genetic founder effect. *Ann Neurol* 1988;23:25–31.
- 27 Bird TD, Sumi SM, Nemens EJ, Nochlin D, Schellenberg GD, Lampe TH, Sadovnick A, Chiu H, Miner GW, Tinklenberg J: Phenotypic heterogeneity in familial Alzheimer's disease: a study of 24 kindreds. *Ann Neurol* 1989;25:12–25.
- 28 Schäffer AA, Gupta SK, Shriram K, Cottingham RW: Avoiding recomputation in linkage analysis. *Hum Hered* 1994;44:225–237.
- 29 Thompson EA: Monte Carlo likelihood in genetic mapping. *Stat Sci* 1994;9:355–366.
- 30 Heath SC: Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 1997;61:748–760.
- 31 Thompson EA, Heath SC: Estimation of conditional multilocus gene identity among relatives. In *Statistics in Molecular Biology and Genetics: Selected Proceedings of a 1997 Joint AMS-IMS-SIAM Summer Conference on Statistics in Molecular Biology* (edited by F Seilinger-Moiseiwitsch), vol. 33 of IMS Lecture Note–Monograph Series. Institute of Mathematical Statistics, Hayward, CA, 1999, pp 95–113.
- 32 Kong A, Lui JS, Wong WH: Sequential imputations and Bayesian missing data problems. *J Am Stat Assoc* 1994;89:278–288.
- 33 Thompson EA: Monte Carlo likelihood in the genetic mapping of complex traits. *Philosophical Trans R Soc London (Series B)* 1994;344:345–351.
- 34 Thompson EA: *Statistical Inferences from Genetic Data on Pedigrees*, vol. 6 of NSF-CBMS Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, Beachwood, OH, 2000.
- 35 Thomas A, Gutin A, Abkevich V, Bansal A: Multilocus linkage analysis by blocked Gibbs sampling. *Stat Computing* 2000;10:259–269.
- 36 Lin S, Thompson EA, Wijsman EM: An algorithm for Monte Carlo estimation of genotype probabilities on complex pedigrees. *Ann Hum Genet* 1994;58:343–357.
- 37 Sheehan NA, Thomas AW: On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* 1993;49:163–175.