Biometrics 62, 728–734
September 2006

DOI: 10.1111/j.1541-0420.2006.00532.x

Improving Estimates of Genetic Maps: A Maximum Likelihood Approach

William C. L. Stewart[1,2,*] and Elizabeth A. Thompson[1]

[1]Department of Statistics, University of Washington, Seattle, Washington 98195, U.S.A.
[2]*Current address:* Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109-2029, U.S.A.
[*]*email:* wstew@umich.edu

Summary. As a result of previous large, multipoint linkage studies there is a substantial amount of existing marker data. Due to the increased sample size, genetic maps estimated from these data could be more accurate than publicly available maps. However, current methods for map estimation are restricted to data sets containing pedigrees with a small number of individuals, or cannot make full use of marker data that are observed at several loci on members of large, extended pedigrees. In this article, a maximum likelihood (ML) method for map estimation that can make full use of the marker data in a large, multipoint linkage study is described. The method is applied to replicate sets of simulated marker data involving seven linked loci, and pedigree structures based on the real multipoint linkage study of Abkevich et al. (2003, *American Journal of Human Genetics* **73,** 1271–1281). The variance of the ML estimate is accurately estimated, and tests of both simple and composite null hypotheses are performed. An efficient procedure for combining map estimates over data sets is also suggested.

Key words: Map estimation; Multipoint linkage analysis; Optimization algorithms; Stochastic approximation.

1. Introduction

Unlike physical maps, which are based on sequence data, genetic maps depend on the stochastic process of inheritance. From one generation to the next, exactly half of a parent's genetic information is passed to its offspring. The genetic material that resides on a parental chromosome tends to be inherited as one contiguous piece. However, any chromosome in an offspring may be a mosaic of the homologous chromosome pair in the corresponding parent. The change points (if any) along the offspring chromosomes are known as *crossovers*. Consider the genetic material at two different markers or *loci* along a single offspring chromosome. The recombination rate between these markers is the probability that an odd number of crossovers occurs in this region. If this probability is less than 1/2, the two markers are said to be *linked*. In this article, it is assumed that crossovers occur along an offspring chromosome according to a Poisson process with rate 1 per Morgan (Haldane, 1919). By convention, the term Morgan denotes a unit of *genetic distance*. A genetic map specifies the order of a set of linked markers or *linkage group* and the genetic distance between each adjacent pair.

This article assumes that genotypic data are observed at some (possibly all) markers of an ordered linkage group on some (possibly all) members of a collection of pedigrees. This reduces genetic map estimation to the estimation of the vector of genetic distances between each pair of adjacent markers. Furthermore, if the crossover process is modeled as a $\frac{1}{2}$-*thinned* point process, then the recombination rate between two markers is an increasing function of the corresponding genetic distance (Mather, 1938). For a characterization of the point processes in this class, which includes the Poisson process, see Yannaros (1988). Hence, maximum likelihood (ML) estimation of the genetic map is equivalent to ML estimation of the corresponding vector of recombination rates, referred to hereafter as the *map*.

Despite these simplifying assumptions, the estimation of genetic maps from arbitrary pedigree data is often complicated. For example, when only some members of an extended pedigree are sampled across several markers, existing methods cannot compute the maximum likelihood estimate (MLE) of the genetic map. For data sets containing extended pedigrees, the program CRIMAP (Lander and Green, 1987) can be used to estimate the map. The resulting estimator is not the MLE, nor is it unbiased. The program avoids computational bottlenecks by ignoring allele frequencies and by ignoring information where there are genotype data that are missing. It implements a method that resembles the expectation maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) and the most recent documentation can be found at `http://linkage.rockefeller.edu/soft/crimap/`. In the context of ML estimation, current methods (Elston and Stewart, 1971; Lathrop, Lalouel, and White, 1986; Lander and Green, 1987; Gudbjartsson et al., 2000) are limited to data sets containing pedigrees with a small number of individuals, or to data sets with a small number of polymorphic loci.

Genetic maps are an integral part of several statistical methods that are commonly used to find disease genes. For example, a large collection of statistical methods known as *multipoint linkage analysis* is often used to locate disease genes relative to a genetic map of DNA markers. These methods typically assume that the marker recombination rates are known, when in fact, they are estimated from data. This leads to map misspecification that can have a negative impact on subsequent inference about the location of disease genes (Halpern and Whittemore, 1999; Daw, Thompson, and Wijsman, 2000). Therefore, improved map estimates should facilitate the discovery of disease genes by reducing map misspecification.

In this article, an ML method for making inference about the genetic map is proposed. It is not limited to the analysis of data sets containing pedigrees with a small number of individuals. In particular, the method can be used to analyze the data that are typically found in large, multipoint linkage studies. Such data sets may contain several thousand meioses with individuals typed across multiple markers of known order, but have missing data. An analysis of these types of data sets has the potential to yield more precise estimates than the published maps of either Marshfield (Broman et al., 1998) or deCODE (Kong et al., 2002). The method combines Markov chain Monte Carlo (MCMC) (Metropolis et al., 1953), Monte Carlo expectation maximization (MCEM) (Guo and Thompson, 1994), and stochastic approximation (SA) (Robbins and Monro, 1951) to find the MLE of the map. An MCMC likelihood ratio estimator is developed for testing both simple and composite null hypotheses, and a procedure for combining map estimates over data sets is suggested. The proposed method is applied to simulated data involving 2201 meioses in 110 pedigrees with as many as seven generations. The map and pedigree structures used to simulate the data are based on the large, multipoint linkage study of Abkevich et al. (2003). For these data, exact computation of the MLE is infeasible. The proposed method is implemented in the program LM_MAP and uses the libraries/structure of the MORGAN 2.7 software. It will be made publicly available through the scheduled release of MORGAN 2.8, which uses the same libraries/structure as MORGAN 2.7. The MORGAN software is maintained at `http://www.stat.washington.edu/thompson/Genepi/MORGAN/Morgan.shtml`.

2. Methods

2.1 *Probability Models for Pedigree Data*

Calculating the probability of data observed on a pedigree is essentially a *missing data* or *latent variable* problem, and computational difficulties arise due to the fact that many members are related. Formally, a *pedigree* is a data structure that specifies the genealogical relationships of a set of individuals. The gender of each individual is also specified. Consider ℓ ordered loci along a chromosome and a collection of pedigrees containing a total of m parent–child transmissions or *meioses*. Define the latent variable S_{ij} as the *meiosis indicator* (Thompson, 2000):

$$S_{ij} = \begin{cases} 1, & \text{if the DNA transmitted in the } i\text{th meiosis at the} \\ & j\text{th locus is the parent's paternal copy,} \\ 0, & \text{otherwise.} \end{cases}$$

Unconditionally, each S_{ij} is either zero or one with probability $\frac{1}{2}$ for all i and j. Group the m meioses into two sets, $\mathcal{F}$ and $\mathcal{M}$. Let the set $\mathcal{F}$ contain the meioses that occur in females (i.e., mothers), and let the set $\mathcal{M}$ be defined similarly for males. Let η_j and μ_j be the female and male recombination rates between loci j and $j+1$, respectively. Then,

$$\eta_j = P(S_{k,j} \neq S_{k,j+1}), \ \forall\, k \in \mathcal{F}$$
$$\mu_j = P(S_{k,j} \neq S_{k,j+1}), \ \forall\, k \in \mathcal{M},$$

for all $j = 1, \ldots, \ell - 1$. Define $\eta \equiv (\eta_1, \ldots, \eta_{\ell-1})$ and $\mu \equiv (\mu_1, \ldots, \mu_{\ell-1})$. Define $\mathbf{S} \equiv \{S_{i,j}\}$ as the matrix of meiosis indicators. Let $\mathbf{G}$ denote the genotypic data observed on some (possibly all) members of each pedigree across some (possibly all) loci. The likelihood of $\phi \equiv (\eta, \mu)$ is

$$L(\phi) = P_\phi(\mathbf{G}) = \sum_{\mathbf{S}} P_\phi(\mathbf{G}, \mathbf{S}). \tag{1}$$

The exact computation of $L(\phi)$ for marker data collected in large linkage studies is often infeasible due to the number of terms involved in the summation in (1). However, in such cases it is possible to compute the joint probability of the meiosis indicators and the data. Under the assumed Poisson process model for crossovers, recombination events in disjoint intervals are independent. Hence, the inheritance vectors $(\mathbf{S}_{\bullet 1}, \ldots, \mathbf{S}_{\bullet \ell})$ are first-order Markov along the chromosome, where $\mathbf{S}_{\bullet j} \equiv (S_{1,j}, \ldots, S_{m,j})^T$ for all $j = 1, \ldots, \ell$. Thus, $P_\phi(\mathbf{S})$ factors into

$$P(\mathbf{S}_{\bullet 1}) \prod_{j=1}^{\ell-1} P_{\phi_j}(\mathbf{S}_{\bullet j+1} \,|\, \mathbf{S}_{\bullet j}). \tag{2}$$

Let $\mathbf{G}_{\bullet j}$ denote the genotypic data at locus j, and assume that in the population an individual's genotype at locus j is independent of that at j', for any two loci j and $j' \neq j$. Under this assumption, the conditional distribution of $\mathbf{G}_{\bullet j}$ given the remaining genotypes and $\mathbf{S}$, only depends on $\mathbf{S}_{\bullet j}$. Thus, the observed data $\mathbf{G}$ and latent variable $\mathbf{S}$ form a hidden Markov model, and $P_\phi(\mathbf{G}, \mathbf{S})$ factors into

$$P(\mathbf{S}_{\bullet,1}) \prod_{j=1}^{\ell} P(\mathbf{G}_{\bullet,j} \,|\, \mathbf{S}_{\bullet,j}) \prod_{j=1}^{\ell-1} P_{\phi_j}(\mathbf{S}_{\bullet j+1} \,|\, \mathbf{S}_{\bullet j}). \tag{3}$$

The expression in (3) can be computed for large data sets containing extended pedigrees and an arbitrary number of polymorphic loci (Thompson, 2000).

2.2 *Finding the MLE*

The combination of two stochastic algorithms, MCEM (Guo and Thompson, 1994) and SA (Robbins and Monro, 1951), is used to estimate the MLE(ϕ). A brief description of EM (Dempster et al., 1977) is given before introducing MCEM. Likewise, a brief description of SA is given before introducing the proposed stochastic optimization hybrid (SOH) algorithm, which exploits the strengths of both MCEM and SA. Let $\phi^{(0)}$ be a point in the interior of the parameter space. To apply the EM algorithm the following recursion is implemented:

E-step: Compute $Q(\phi; \phi^{(n)}) \equiv \mathbf{E}_{\phi^{(n)}}\{\log P_\phi(\mathbf{G}, \mathbf{S}) \,|\, \mathbf{G}\}$,

M-step: Set $\phi^{(n+1)} \equiv \underset{\phi}{\operatorname{argmax}}\, Q(\phi; \phi^{(n)})$.

Under mild conditions (Wu, 1983), the sequence $\{\phi^{(n)}\}$ $n = 0, 1, \ldots$ converges to a local maximum or stationary point. In particular, if $L(\phi)$ is unimodal, convergence to the MLE is assured. While the convergence is only linear, in practice the sequence tends to move very quickly to some small neighborhood of the MLE. Moreover, the ascent property

$$L(\phi^{(n+1)}) \geq L(\phi^{(n)})$$

is guaranteed. However, when data are collected on members of large, extended pedigrees, the E-step is infeasible. In such situations an MCMC estimate of $Q(\phi;\, \phi^{(n)})$ is possible using an MCMC sampler developed by Heath and Thompson (1997). This sampler generates realizations of $\mathbf{S}$ from $P_\phi(\mathbf{S} \,|\, \mathbf{G})$. MCEM is a stochastic version of EM that replaces the deterministic E-step by an MCMC estimate. A serious drawback of the MCEM algorithm is that for a fixed Monte Carlo sample size there is an upper bound on the precision of any MCEM estimator. However, when the Monte Carlo sample size is allowed to increase with each iteration, MCEM tends to possess the ascent property in the early stages of the algorithm with high probability (Caffo, Jank, and Jones, 2005).

Originally, the SA algorithm was invented for the purpose of finding the MLE when the support of the likelihood function is one dimensional. It was extended to the multidimensional case by Nevel'son and Has'minski (1973) and Gu and Li (1998). Gu and Kong (1998) developed further improvements to allow for the joint estimation of the MLE and its variance. In general, all versions of SA exploit the fact that

$$\mathbf{E}_\phi\{\nabla \log P_\phi(\mathbf{G}, \mathbf{S}) \,|\, \mathbf{G}\} = \nabla \log L(\phi).$$

Suppose that $\{{\mathbf{S}_1}^{(n)}, \ldots, {\mathbf{S}_k}^{(n)}\}$ is a set of dependent realizations from $P_{\phi^{(n)}}(\mathbf{S} \,|\, \mathbf{G})$. Define the sequence $\{\phi^{(n)}\}$ by the recursion

$$\phi^{(n+1)} = \phi^{(n)} + \gamma_n \Gamma^{-1} \frac{1}{k} \sum_{t=1}^{k} \nabla \log P_{\phi_{(n)}}\big(\mathbf{G}, {\mathbf{S}_t}^{(n)}\big). \quad (4)$$

Under certain conditions on the sequence $\{\gamma_n \Gamma^{-1}\}$ and the function $\log L(\phi)$, the sequence $\phi^{(n)}$ converges to the MLE (see Nevel'son and Has'minski, 1973 and Younes, 1999 for details). For example, if $\{\gamma_n\}$ is of the form $(c + n\rho)^{-1}$, $c \geq 0$, $\rho > 0$, Γ is positive definite, and $L(\phi)$ is unimodal with a unique maximum, then the sequence defined by (4) converges to the MLE for all $k > 0$. In the current implementation, $\rho = 1$, $k = 10$, and the value of c is random, with distribution depending on the conditional distribution $\mathbf{S}$ given $\mathbf{G}$.

Let $0 < T_1 < \infty$ and $0 \leq T_2 < \infty$ be integers, constant or random with finite variances. The SOH algorithm consists of T_1 MCEM steps, followed by T_2 SA steps. The MLE is approximated by $\phi^{(T_1+T_2)}$, where T_1 and T_2 are random stopping times based on the percent change in successive estimates. Specifically, SA is initiated when the percent change in each component of successive MCEM updates is less than 20%. Similarly, SA is terminated when the percent change in each component of successive SA updates is less than 1%. In practice, MCEM is relatively insensitive to the choice of $\phi^{(0)}$, but may take a long time to converge. By contrast, practical convergence of SA is much more sensitive to the choice of starting position but rapid convergence of SA tends to occur whenever the starting position is close to a local maximum of $L(\phi)$. By combining both algorithms, the SOH algorithm is able to generate reliable estimates of the MLE. In addition, an MCMC estimate of the variance of the MLE follows from the Louis missing information principle (Louis, 1982):

$$-\nabla^2 \log L(\phi) = -\mathbf{E}_\phi\big\{\nabla^2 \log P_\phi(\mathbf{G}, \mathbf{S}) \,|\, \mathbf{G}\big\} - \mathrm{Var}_\phi\{\nabla \log P_\phi(\mathbf{G}, \mathbf{S}) \,|\, \mathbf{G}\}. \quad (5)$$

2.3 *Combining Map Estimates*

A reasonable way to combine map estimates across data sets is to average them with weights that vary inversely in proportion to their variance. Consider the restricted model $P_\tau(\mathbf{G})$ indexed by $\tau \in \{\phi : \eta = \mu\}$. Let $\hat{\tau}$ be the MLE(τ) and define the jth interval as the region between markers j and $j + 1$. Let N_j denote the *effective number of meioses* in the jth interval (Edwards, 1976), and define N_j as the ratio of $\tau_j(1 - \tau_j)$ to $\mathrm{Var}(\hat{\tau}_j)$. An estimate of N_j follows from $\hat{\tau}_j$ and the estimate of $\mathrm{Var}(\hat{\tau}_j)$ based on (5). For a given linkage group, consider a collection of map estimates $(\hat{\tau}^{(1)}, \ldots, \hat{\tau}^{(R)})$, each derived from different data sets $(\mathbf{G}^{(1)}, \ldots, \mathbf{G}^{(R)})$ with $\mathbf{G}^{(r)} \sim P_\tau$, for $r = 1, \ldots, R$. Let $\hat{N}_j^{(r)}$ be the estimated number of effective meioses for the jth interval based on the rth data set. Define the set of weights $\Lambda \equiv \{\lambda_{jr}\}$ by

$$\lambda_{jr} = \frac{\hat{N}_j^{(r)}}{\sum_{s=1}^{R} \hat{N}_j^{(s)}} \quad \forall j, r. \quad (6)$$

For each j, the weights $(\lambda_{j1}, \ldots, \lambda_{jR})$ vary inversely in proportion to the variance of $\hat{\tau}_j^{(r)}$. The composite map estimate $\tilde{\tau} \equiv (\tilde{\tau}_1, \ldots, \tilde{\tau}_{\ell-1})$ defined by

$$\tilde{\tau}_j = \sum_{r=1}^{R} \lambda_{jr} \hat{\tau}_j^{(r)}, \quad j = 1, \ldots, \ell - 1$$

is the weighted average of R map estimates. Furthermore, an estimate of $\hat{N}_j$ relative to the jth interval of the composite map estimate is $\sum \hat{N}_j^{(r)}$, for $r = 1, \ldots, R$. This follows from the weights defined in (6), and from the use of $\hat{N}_j^{(r)}$ as an estimate of $N_j^{(r)}$ in the computation of $\mathrm{Var}(\tilde{\tau}_j)$. In addition, the gender-specific analog of $\hat{N}_j$ follows from the estimation of (5) under the full model. Hence, the construction of a composite gender-specific map estimate is straightforward.

2.4 *Hypothesis Testing*

While it is well known that for certain regions of the human genome the genetic maps of men and women differ (Broman et al., 1998; Matise et al., 2003), there are regions where the pattern and intensity of the differences are less clear. The statistical significance of the difference between male and female maps is rarely tested. For any nested pair of the following parameter spaces, the proposed method can test the corresponding statistical hypothesis:

$$K_0 : \{\phi : \eta = \mu = \tau^*\},$$
$$K_1 : \{\phi : \eta = \eta^*, \mu = \mu^*\},$$
$$K_2 : \{\phi : \eta = \mu\},$$
$$K_3 : \{\phi : \text{unconstrained}\},$$

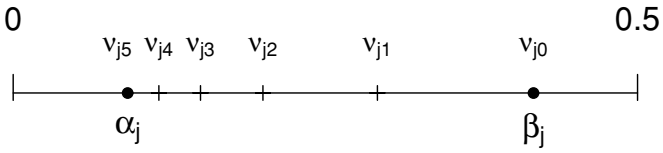

Figure 1. The label $\alpha_j(\beta_j)$ denotes the recombination rate of the jth interval for the sex-averaged map denoted $\alpha(\beta)$. A conceptual example of the jth sequence $\{\nu_{ji} : i = 0, 1, \ldots, 5\}$ of length $D = 5$ that connects α_j and β_j is depicted.

where τ^*, η^*, and μ^* are fixed sex-averaged, female, and male maps, respectively. Models K_0 and K_2 are used when testing a specific sex-averaged map versus a general sex-averaged map. Models K_2 and K_3 are used when testing a general sex-averaged map versus a general sex-specific map.

For ease of exposition, consider again the restricted model $P_\tau(\mathbf{G})$, and let α, β be two sex-averaged maps with $\alpha \neq \beta$. The likelihood ratio formula

$$LR(\alpha, \beta) \equiv \frac{L(\alpha)}{L(\beta)} = \mathbf{E}_\beta \left\{ \frac{P_\alpha(\mathbf{G}, \mathbf{S})}{P_\beta(\mathbf{G}, \mathbf{S})} \,\middle|\, \mathbf{G} \right\} \tag{7}$$

(Guo and Thompson, 1994) leads immediately to MCMC estimation of $LR(\alpha, \beta)$. Let $T^{(\alpha,\beta)}$ be the usual MCMC estimator of $LR(\alpha, \beta)$. When $P_\beta(\mathbf{S} \mid \mathbf{G})$ is close to proportional to $P_\alpha(\mathbf{G}, \mathbf{S})$, accurate estimates of the $LR(\alpha, \beta)$ based on $T^{(\alpha,\beta)}$ are easily obtained. However, when α and β are points of maxima under competing models, the two distributions may not be close to proportional and reliable MCMC estimation of $LR(\alpha, \beta)$ based on $T^{(\alpha,\beta)}$ may be computationally prohibitive.

A simple and effective solution is obtained as follows: Let $\{(\nu_{j0}, \ldots, \nu_{jD})\}$ be a set of $\ell - 1$ monotonic sequences, where each sequence has length D. Let the set satisfy $\alpha_j = \nu_{jD}$, and $\beta_j = \nu_{j0}$ for $j = 1, \ldots, \ell - 1$. Thus, the jth sequence connects the jth components of β and α; D does not depend on j. Moreover, the vector $\nu_{\bullet d} \equiv (\nu_{1,d}, \ldots, \nu_{(\ell-1),d})$ is a sex-averaged map between $\nu_{\bullet 0} \equiv \beta$ and $\nu_{\bullet D} \equiv \alpha$. Define

$$W^{(\alpha,\beta)} = \prod_{d=0}^{D-1} \frac{T^{(d+1,d^*)}}{T^{(d,d^*)}}, \tag{8}$$

where the superscripts $d+1, d$, and d^* denote $\nu_{\bullet(d+1)}, \nu_{\bullet d}$, and $\nu_{\bullet d^*} \equiv \frac{1}{2}(\nu_{\bullet(d+1)} + \nu_{\bullet d})$, respectively. In addition, when $\alpha_j < \beta_j$ the ending terms of the jth sequence are chosen to concentrate near α_j (see Figure 1). The estimator $W^{(\alpha,\beta)}$ is a better estimator of $LR(\alpha, \beta)$ than $T^{(\alpha,\beta)}$ when $P_\beta(\mathbf{S} \mid \mathbf{G})$ is not close to proportional to $P_\alpha(\mathbf{G}, \mathbf{S})$.

If $\alpha = \hat{\tau}$ and $\beta = \tau_0$, then $2\log W^{(\alpha,\beta)}$ is a consistent MCMC estimator of the likelihood ratio test (LRT) statistic for testing K_0 versus K_2. Estimators for testing hypotheses based on any of the other nested models are constructed in a similar fashion.

3. Applications

3.1 *Description of Simulations*

Multiple sets of simulated genotypic data are analyzed to assess the performance of the proposed method. The same sex-averaged map denoted τ_0, and the same pedigree structures are used for all simulations. A single *replicate* contains the data simulated at seven polymorphic markers for each of 1900 individuals comprising 110 pedigrees. The pedigree structures are based on the large, multipoint linkage study of Abkevich et al. (2003). The sex-averaged map is based on the Marshfield map of chromosome 12. For each locus, the number of alleles and their frequencies are based on information taken from the Duke Center for Human Genetics, and 2% of the genotypes in each replicate are masked. A total of 500 replicates are simulated.

To examine the large-scale effects of different levels of missing data, the complete marker data for some individuals in each replicate are masked. The individuals are chosen in a manner consistent with the actual pattern of missing data in the large linkage study of Abkevich et al. (2003). Under Scheme A, the marker data for 38% of the individuals are masked. Under Scheme B, the marker data for an additional 16% are masked. Two different models for marker allele frequencies are considered: Model p_0: assumes that allele frequencies are known; Model $\hat{p}$: estimates allele frequencies from the available data. For each model, the SOH algorithm is used to estimate the map. Additionally, the program CRIMAP is used to estimate the map; CRIMAP does not use any model for allele frequencies.

3.2 *Results*

To assess the quality of the variance estimator based on (5), and to examine the sensitivity of the confidence intervals to the assumption of normality, all 500 replicates are used to estimate the following coverage probabilities: 95%, 90%, 80%, 50% (see Table 1). Specifically, Model p_0 is used to estimate $\hat{\tau}$ for each replicate. Then, the components of $\hat{\tau}$ are used as point estimates to construct confidence intervals for the corresponding components of τ. In general, the estimated coverage of each confidence interval is close to its expected coverage. Inspection of Table 1 across both missing data levels A and B shows that the confidence intervals are slightly conservative.

The quality of an estimator is often assessed through its root mean square error (RMSE). In Table 2, the estimated

Table 1

For each component of τ, the estimated percent coverage of the confidence interval based on all 500 *replicates is shown. The corresponding standard errors are estimated separately for each replicate. Schemes A and B have* 38% *and* 54% *missing data, respectively.*

		Theoretical % coverage			
	Interval	95	90	80	50
Scheme A	1	97.0	93.6	84.8	56.2
	2	98.0	94.2	86.8	53.8
	3	97.0	93.2	82.0	51.6
	4	96.6	93.0	81.8	52.0
	5	96.0	91.2	82.4	51.0
	6	98.0	93.6	85.0	56.2
Scheme B	1	96.8	94.0	86.2	54.6
	2	97.8	94.0	85.0	54.8
	3	96.2	90.8	83.6	53.2
	4	97.6	92.6	79.4	51.2
	5	96.0	91.4	84.0	53.0
	6	97.8	93.2	85.0	53.2

Table 2

For each component of τ and across both levels of missing data, the estimated RMSE $\times 10^4$ for three competing estimators is shown. Schemes A and B have 38% *and* 54% *missing data, respectively. The corresponding estimate of absolute bias (similarly scaled) is shown in parentheses.*

	Interval	Model p_0		Model $\hat{p}$		CRIMAP	
Scheme A	1	165.49	(18.56)	168.00	(16.15)	321.65	(180.85)
	2	90.13	(12.72)	90.62	(13.20)	163.83	(98.93)
	3	97.18	(16.04)	97.78	(15.71)	201.89	(153.75)
	4	68.52	(8.02)	67.88	(6.92)	119.57	(80.37)
	5	88.04	(11.64)	88.00	(11.77)	147.12	(80.43)
	6	102.47	(4.83)	101.56	(6.74)	197.76	(133.82)
Scheme B	1	205.34	(48.38)	209.07	(51.01)	754.33	(536.78)
	2	105.42	(18.11)	104.56	(19.38)	342.61	(250.98)
	3	112.08	(19.48)	111.72	(19.51)	431.62	(383.04)
	4	78.17	(12.81)	79.17	(12.76)	221.02	(178.50)
	5	100.76	(15.63)	102.58	(15.06)	311.38	(215.23)
	6	124.65	(11.50)	127.04	(11.53)	427.46	(315.00)

RMSE(τ) for each component of τ is shown across both levels of missing data using Model p_0, Model $\hat{p}$, and CRIMAP. For these data, the components of the CRIMAP estimator show an increase in their estimated absolute bias relative to the corresponding components of the ML estimators. In fact, the estimated bias of the CRIMAP estimator is negative for all components across both schemes. Furthermore, an increase in the amount of missing data appears to have a larger effect on the CRIMAP estimator than it does on either of the ML estimators. For these data, the ML estimation of the map was insensitive to assumptions about the allele frequencies.

Typically, only one data set is available for any real analysis. Therefore, a single replicate is analyzed at both levels of missing data using Model p_0. The estimated 95% confidence interval for each component of τ and its corresponding estimated number of effective meioses per interval are shown in Table 3. Also shown in Table 3 are the 95% confidence intervals for η and μ. The confidence intervals under Scheme B are wider than those under Scheme A on account of the additional missing data. Inspection of each $\hat{N}_j$ across different schemes shows that these data contain more information than do the data that were used to construct the Marshfield map (188 meioses). By contrast, these data and the data used by deCODE (1257 meioses) may contain similar amounts of information, at least under Scheme A. However, by combining map estimates from different data sets, an increase in accuracy and precision over that of the deCODE map is assured.

An advantage of an ML approach is that hypothesis testing follows immediately from standard likelihood theory. Under Schemes A and B, the empirical distribution functions of the LRT statistics for testing a given (K_0) against an unspecified (K_2) sex-averaged map and the composite null hypothesis (K_2) against a general sex-specific map (K_3) based on 200 replicates are examined (data not shown). For the regions of most interest (i.e., the right-hand tails of the distribution), there is close agreement between each empirical distribution and its limiting chi-squared distribution. As an example, consider testing at the 5% α-level. The estimated size for testing the simple null is 5.5% and 3.5% under Schemes A and B, respectively. Similarly, the estimated size for testing the composite null is 7.5% and 4.0% under Schemes A and B, respectively. Additionally, the power for testing the composite null is estimated

Table 3

The 95% *confidence intervals for the components of τ, η, and μ, along with estimates of the sex-averaged effective number of meioses per interval. Schemes A and B have* 38% *and* 54% *missing data, respectively. All estimates are based on the analysis of a single replicate.*

	τ_o	τ	$\hat{N}_j$	η	μ
Scheme A	0.201	(0.165, 0.245)	389	(0.090, 0.252)	(0.132, 0.316)
	0.083	(0.069, 0.109)	792	(0.031, 0.135)	(0.051, 0.147)
	0.137	(0.117, 0.159)	1055	(0.086, 0.182)	(0.095, 0.191)
	0.059	(0.039, 0.068)	927	(0.017, 0.085)	(0.019, 0.087)
	0.074	(0.061, 0.096)	877	(0.036, 0.123)	(0.041, 0.128)
	0.113	(0.072, 0.114)	717	(0.036, 0.146)	(0.050, 0.146)
Scheme B	0.201	(0.155, 0.244)	313	(0.115, 0.401)	(0.045, 0.270)
	0.083	(0.067, 0.119)	491	(0.053, 0.183)	(0.009, 0.155)
	0.137	(0.120, 0.171)	732	(0.084, 0.202)	(0.083, 0.201)
	0.059	(0.046, 0.078)	828	(0.025, 0.113)	(0.011, 0.079)
	0.074	(0.064, 0.109)	625	(0.045, 0.163)	(0.029, 0.117)
	0.113	(0.105, 0.167)	472	(0.079, 0.258)	(0.042, 0.183)

from 200 replicates simulated under the alternative hypothesis of unequal male and female maps. Specifically, the female genetic map is 1.4 times that of the male genetic map. The estimated power at the 5% α-level is 61%.

4. Discussion

The proposed method makes efficient use of marker data typically found in a large linkage study. This is evidenced by the much lower sampling variance of each ML estimator as compared to the sampling variance of the CRIMAP estimator. The estimated absolute bias for each component of the map is close to zero for each ML estimator (see Table 2), which strongly suggests that the SOH algorithm accurately estimates the MLE. The general agreement between the estimated coverage probability of each confidence interval and its expected value suggests that the variance of the MLE is accurately estimated within each replicate (see Table 1). Moreover, if map estimates are combined over studies, in the manner suggested by (8), improved map estimates are assured.

For most MCMC analyses of genetic data, achieving reliable estimates in a practical amount of time is not trivial. However, the time required for these analyses is not unreasonable. Using an Intel(R) Xeon(TM) CPU 2.80 GHz processor with a Linux operating system, the time required to estimate $(\hat{\tau}, \mathrm{Var}(\hat{\tau}), N_j)$ under Scheme A was approximately 20 minutes. Approximately 10% more time is needed to generate the corresponding sex-specific estimates. The time required to estimate the LRT statistics for the simple and composite null hypotheses is approximately 40 minutes. The time needed for any analysis under Scheme B is roughly 50% greater than the time required for the corresponding analysis under Scheme A. In general, 1000 MCMC realizations correspond to approximately 3.5 minutes of CPU time.

For each component of $\hat{\tau}$, the variance attributable to MCMC error and random fluctuations in the initial map are estimated from independent MCMC analyses of the same replicate. For each component, the variance due to the stochastic nature of the algorithm is an order of magnitude less than the corresponding estimate of statistical variance, which suggests that reliable estimation of the MLE is achieved. Moreover, when analyzing data sets where exact calculation of the MLE is possible, the SOH algorithm converges to the MLE (data not shown). Relative to map estimation, this suggests that the LM sampler (Heath and Thompson, 1997) mixes adequately over the constrained space of inheritance patterns. However, more MCMC is needed to obtain accurate coverage probabilities. A preliminary investigation suggests that the differences between the estimated and expected coverage probabilities (Table 1) do decrease as the amount of MCMC effort increases.

For ease of presentation, considerable attention is given to the problem of making inferences about the sex-averaged map. However, the proposed method also makes analogous inferences about the sex-specific map. In particular, the SOH algorithm accurately estimates $\hat{\eta}$, and $\hat{\mu}$ under the full model, which allows for general η and μ (data not shown). To some extent, this is implicit in the computations used to estimate the size of the LRT for testing the composite null (7.5% and 4.0% at the 5% α-level). When estimating the power—61% at the 5% α-level for a sex-specific ratio of 1.4—the estimation of $\hat{\eta}$ and $\hat{\mu}$ under the full model is also required. Moreover, since the total length of the female genome is estimated at 1.65 times that of the total male genome (Kong et al., 2002; Matise et al., 2003), it is likely that the power to detect a difference between male and female maps in a given region will be high, provided that data sets have a similar amount of information. When the alternative hypothesis is a sex-specific map with a sex-specific ratio of 1.65, the estimated power to detect unequal male and female maps at the 5% α-level is near 1.0 (data not shown). However, due to the computational demands needed for accurate estimation of the LRT statistics, these estimates of size and power are only based on 200 replicates each.

When data are missing, the estimated recombination rate in one interval will depend on the estimated rates in other intervals. Depending on the pattern and the strength of these associations, it may be desirable to analyze many markers jointly. When analyzing many markers simultaneously (possibly hundreds), the likelihood surface may be multimodal. As such, it is important to explore the complete likelihood surface. An option in the current implementation of the algorithm is to choose the MLE from a set of candidate ML estimates, each of which is generated from a separate run of the SOH algorithm. The choice is made based upon a series of likelihood ratio estimates as in (8).

Two implicit assumptions of the proposed method are a known marker order and the absence of genotyping errors. A consequence of the first assumption is that it may be inappropriate to analyze regions of the genome where the order of markers is in question (e.g., regions where markers are separated by little or no recombination). By contrast, the second assumption—that the marker data are observed without error—is almost always violated. Worse yet, undetected genotyping errors tend to result in map expansion (Buetow, 1991; Goldstein, Zhao, and Speed, 1997). To address problems associated with genotyping error, several have developed error detection methods (Ehm, Kimmel, and Cottingham, 1996; Stringham and Boehnke, 1996; Douglas, Skol, and Boehnke, 2002), and genotyping error models have been incorporated into the analysis of pedigree data (Sobel, Papp, and Lange, 2002). Currently, we are developing methods that incorporate genotyping error models into map estimation.

In this article, considerable attention is given to the important problem of map estimation. In fact, we have successfully applied the SOH algorithm to a real data set involving 143 pedigrees and 17 polymorphic loci on chromosome 4 (Sieh et al., 2005). However, the SOH algorithm and the likelihood ratio estimator in (8) are not specific to the problem of map estimation. In principle, the proposed method is applicable to a wide variety of inference problems in structured stochastic systems involving latent variables ("hidden states") and missing data.

Acknowledgements

This work was supported by National Institutes of Health Genome Training Grant HG00035-10 and Grant GM-46255. We are grateful to Dr Ellen Wijsman for providing many helpful comments, and to Dr Victor Abkevich for providing invaluable information regarding the pedigree structures of his

recent linkage study. We also thank the referees for their insightful comments.

References

Abkevich, V., Camp, N. J., Hensel, C. H., et al. (2003). Predisposition locus for major depression at chromosome 12q22-12q23.2. *American Journal of Human Genetics* **73,** 1271–1281.

Broman, K. W., Murray, J. C., Sheffield, V. C., White, R. L., and Weber, J. L. (1998). Comprehensive human genetic map: Individual and sex-specific variation in recombination. *American Journal of Human Genetics* **63,** 861–869.

Buetow, K. H. (1991). Influence of aberrant observations on high-resolution linkage analysis outcomes. *American Journal of Human Genetics* **49,** 985–994.

Caffo, B. S., Jank, W., and Jones, G. L. (2005). Ascent-based Monte Carlo expectation–maximization. *Journal of the Royal Statistical Society, Series B* **67,** 235–251.

Daw, E. W., Thompson, E. A., and Wijsman, E. M. (2000). Bias in multipoint linkage analysis arising from map misspecification. *Genetic Epidemiology* **19,** 366–380.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B* **39,** 1–37.

Douglas, J. A., Skol, A. D., and Boehnke, M. (2002). Probability of detection of genotyping errors and mutations as inheritance inconsistencies in nuclear-family data. *American Journal of Human Genetics* **70,** 487–495.

Edwards, J. H. (1976). The interpretation of lod scores in linkage analysis. *Human Gene Mapping* **3,** 289–293.

Ehm, M. G., Kimmel, M., and Cottingham, R. W. J. (1996). Error detection for genetic data, using likelihood methods. *American Journal of Human Genetics* **58,** 225–234.

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

Goldstein, D. R., Zhao, H., and Speed, T. P. (1997). The effects of genotyping errors and interference on estimation of genetic distance. *Human Heredity* **47,** 86–100.

Gu, M. G. and Kong, F. H. (1998). A stochastic approximation algorithm with Markov chain Monte Carlo method for incomplete data estimation problems. *Proceedings of the National Academy of Sciences of the United States of America* **95,** 7270–7274.

Gu, M. G. and Li, S. (1998). A stochastic approximation algorithm for maximum likelihood estimation with incomplete data. *Canadian Journal of Statistics* **26,** 567–591.

Gudbjartsson, D. F., Jonasson, K., Frigge, M. L., and Kong, A. (2000). Allegro, a new computer program for multipoint linkage analysis. *Nature Genetics* **25,** 12–13.

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

Haldane, J. B. S. (1919). The combination of linkage values and the calculation of distances between the loci of linked factors. *Journal of Genetics* **8,** 229–309.

Halpern, J. and Whittemore, A. S. (1999). Multipoint linkage analysis. A cautionary note. *Human Heredity* **49,** 194–196.

Heath, S. and Thompson, E. A. (1997). MCMC samplers for multilocus analyses on complex pedigrees. *American Journal of Human Genetics* **61,** A278.

Kong, A., Gudbjartsson, D. F., Sainz, J., et al. (2002). A high-resolution recombination map of the human genome. *Nature Genetics* **31,** 241–247.

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proceedings of the National Academy of Sciences of the United States of America* **84,** 2363–2367.

Lathrop, G. M., Lalouel, J. M., and White, R. L. (1986). Construction of human genetic linkage maps: Likelihood calculations for multilocus analysis. *Genetic Epidemiology* **3,** 39–52.

Louis, T. A. (1982). Finding observed information using the EM algorithm. *Journal of the Royal Statistical Society, Series B* **44,** 98–130.

Mather, K. (1938). Crossing-over. *Biological Reviews of the Cambridge Philosophical Society* **13,** 252–292.

Matise, T. C., Sachidanandam, R., Clark, A. G., et al. (2003). A 3.9-centimorgan-resolution human single-nucleotide polymorphism linkage map and screening set. *American Journal of Human Genetics* **73,** 271–284.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equations of state calculations by fast computing machines. *Journal of Chemical Physics* **7,** 277–318.

Nevel'son, M. and Has'minski, R. Z. (1973). An adaptive Robbins–Monro procedure. *Automation and Remote Control* **34,** 1594–1607.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *Annals of Mathematical Statistics* **22,** 400–407.

Sieh, W., Basu, S., Fu, A., Rothstein, J., Scheet, P., Stewart, W., Sung, Y., Thompson, E., and Wijsman, E. (2005). Comparison of marker types and map assumptions using MCMC–based linkage analysis of COGA data. *BioMed Central Genetics* **6**(suppl. 1), S11.

Sobel, E., Papp, J. C., and Lange, K. (2002). Detection and integration of genotyping errors in statistical genetics. *American Journal of Human Genetics* **70,** 496–508.

Stringham, H. M. and Boehnke, M. (1996). Identifying marker typing incompatibilities in linkage analysis. *American Journal of Human Genetics* **59,** 946–950.

Thompson, E. A. (2000). *Statistical inferences from genetic data on pedigrees*. NSF-CBMS Regional Conference Series in Probability and Statistics Volume, 6th edition. Beachwood, Ohio: Institute of Mathematical Statistics.

Wu, C. F. J. (1983). On the convergence properties of the EM algorithm. *Annals of Statistics* **11,** 95–103.

Yannaros, N. (1988). On Cox processes and gamma renewal processes. *Journal of Applied Probability* **25,** 423–427.

Younes, L. (1999). On the convergence of Markovian stochastic algorithms with rapidly decreasing ergodicity rates. *Stochastics and Stochastics Reports* **65,** 177–228.

Received June 2005. *Revised November* 2005.
Accepted November 2005.