Statistics in Molecular Biology
IMS Lecture Notes - Monograph Series (1999) Volume 33

ESTIMATION OF CONDITIONAL MULTILOCUS GENE IDENTITY AMONG RELATIVES

BY ELIZABETH A. THOMPSON AND SIMON C. HEATH

University of Washington

Genetic Analysis Workshop 10 identified five key factors contributing to the resolution of the genetic factors affecting complex traits. These include analysis with multipoint methods, use of extended pedigrees, and selective sampling of pedigrees. By sampling the affected individuals in an extended pedigree, we obtain individuals who have an increased probability of sharing genes identical by descent (IBD) at marker loci that are linked to the trait locus or loci. Given marker data on specified members of a pedigree, the conditional IBD status among relatives can be assessed, but exact computation is often impractical for multiple linked markers on complex pedigrees. The use of Markov chain Monte Carlo (MCMC) methods greatly extends the range of models and data sets for which analysis is computationally feasible. Many forms of MCMC have now been implemented in the context of genetic analysis. Here we propose a new sampler, which takes as latent variables the segregation indicators at marker loci, and jointly updates all indicators corresponding to a given meiosis. The sampler has good mixing properties. Questions of irreducibility are also addressed.

1. Introduction. Relatives share common ancestors. A single gene in such an ancestor may therefore descend via repeated segregations to each of the relatives. Such genes, which are copies of a single ancestral gene within a defined pedigree, are said to be *identical by descent* (IBD). Disregarding mutation, IBD genes must be of like type. It is the sharing of IBD genes that underlies phenotypic similarities among relatives. The probabilities of patterns of gene identity by descent are determined by the pedigree structure, and in turn determine the probability distribution of observed data on individuals of the pedigree.

Genetic linkage is the dependent cosegregation of genes at different loci on the same chromosome. Linkage detection and linkage analysis on the basis of data observed on related individuals require the computation of multilocus probabilities of observed phenotypic data on pedigree structures. Genetic Analysis Workshop 10 identified five key factors contributing to the resolution of the genetic factors affecting complex traits (Wijsman and Amos 1997). These include analysis with multipoint methods, use of extended pedigrees, and selective sampling of pedigrees. Here we consider an approach to linkage detection which uses only data on affected individuals. However, calculation of multilocus probabili-

Work supported in part by NIH grant GM-46255 and NSF grant BIR-9305835.

AMS 1991 *subject classifications.* Primary 62F03 ; secondary 92D10.

Key words and phrases. Markov chain Monte Carlo, meiosis samplers, linkage detection, irreducibility, segregation indicators.

ties on extended pedigrees is computationally intensive, particularly when there are many unobserved individuals. In this paper we present a sampling-based approach for linkage detection which is well suited to sparse data at multiple loci on individuals in an extended complex pedigree.

2. Gene identity and linkage likelihoods. There are many ways to partition linkage likelihoods, the probability $\Pr_\psi(\mathbf{Y})$ of phenotypic data $\mathbf{Y}$ under a genetic linkage model ψ. Let $\mathbf{Y}$ consist of trait data $\mathbf{Y}_T$ and marker data $\mathbf{Y}_M$. The model (genetic map positions and marker alleles frequencies) is assumed known for the data, $\mathbf{Y}_M$, at marker loci. In this paper, we shall focus on the problem of linkage detection, in which no trait specific genetic model for the trait data $\mathbf{Y}_T$ is assumed. However, the development is similar in the case where hypothesized trait loci are explicitly modeled and linkage estimation is the goal of the analysis (Thompson 1994b).

Let B_M denote the pattern of gene IBD at marker loci among observed individuals. The likelihood for the genetic model on the basis of data $\mathbf{Y} = (\mathbf{Y}_T, \mathbf{Y}_M)$ is

$$\begin{aligned} \Pr_\psi(\mathbf{Y}) = \Pr_\psi(\mathbf{Y}_T,\ \mathbf{Y}_M) \quad &\propto \quad \Pr_\psi(\mathbf{Y}_T \mid \mathbf{Y}_M) \\ &= \sum_{B_M} \Pr_\psi(\mathbf{Y}_T \mid B_M)\, \Pr(B_M \mid \mathbf{Y}_M) \end{aligned} \tag{2.1}$$

where the model ψ relates to the trait parameters and loci positions relative to the known marker map. If desired, we may consider also IBD status B_T at putative trait loci, and partition the probability further:

$$\Pr_\psi(\mathbf{Y}_T \mid B_M) = \sum_{B_T} \Pr_\psi(\mathbf{Y}_T \mid B_T)\, \Pr_\psi(B_T \mid B_M).$$

Even where no explicit trait model is assumed, there is an implicit assumption in linkage analysis that a trait is genetically determined. Thus individuals of like phenotype have higher probabilities of sharing genes IBD at trait loci, and hence also at linked marker loci. Thus evidence for linkage is provided by marker data $\mathbf{Y}_M$ that give high posterior probability $\Pr(B_M \mid \mathbf{Y}_M)$ to patterns of gene identity B_M which specify greater than expected gene sharing among affected individuals.

A simple example may clarify this perspective. In homozygosity mapping (Lander and Botstein 1987), data on unrelated inbred affected individuals are used to map rare recessive traits. Since the individuals are unrelated, we may consider separately the IBD pattern for each. An example pedigree is shown in Figure 1; this pedigree resulted from a study of a rare recessive disease (Goddard

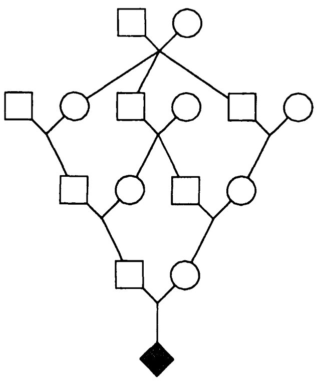

FIG. 1. *Example pedigree, showing the original trait data of a single inbred affected individual.*

et al. 1997). The final individual was ascertained as being affected and the offspring of a marriage between first cousins; it was later recognized that each of his parents was also the child of a first-cousin marriage, as shown. The inbreeding coefficient of the affected individual is $f = 0.109375$.

The IBD patterns of interest are whether the two genes of the inbred affected individual are IBD ($B_T = 1$) or not ($B_T = 0$). Since the trait is a rare recessive

$$\Pr_\psi(\mathbf{Y}_T \mid B_T = 1) \gg \Pr_\psi(\mathbf{Y}_T \mid B_T = 0)$$

and if a marker M is closely linked to the trait locus T

$$\Pr_\psi(B_T = 1 \mid B_M = 1) \gg \Pr(B_T = 1) \ = \ f$$

where $\Pr(B_T = 1)$ is the prior probability of gene identity at a locus implied by the pedigree structure, which in this case is simply the inbreeding coefficient (f) of the affected individual. Finally, if the data $\mathbf{Y}_M$ specify homozygosity of the affected individual at a polymorphic marker locus

$$\Pr(B_M = 1 \mid \mathbf{Y}_M) > \Pr(B_M = 1) \ = \ f.$$

Homozygosity at multiple linked marker loci reinforces the inference that the affected individual is IBD in this segment of the genome. Data on multiple affected individuals, all homozygous in the same genome region, together provide evidence that the hypothesized trait locus is also located in this region.

3. Exact computation of probabilities on pedigrees. In (2.1) the terms $\Pr(B_M \mid \mathbf{Y}_M)$ are the conditional probabilities of marker loci IBD status given marker data. In the current paper we shall not consider explicit trait models, but focus on the IBD information conveyed by marker loci, and the estimation of $\Pr(B_M \mid \mathbf{Y}_M)$. Now

$$\Pr(B_M \mid \mathbf{Y}_M) = \Pr(\mathbf{Y}_M \mid B_M) \Pr(B_M) \ / \ \Pr(\mathbf{Y}_M)$$

and thus exact computation of the conditional probability requires the computation of $\Pr(\mathbf{Y}_M)$ the overall probability of the marker data observed on the pedigree. We consider first, therefore, the evaluation of such probabilities. Since now we consider only marker loci, we drop the subscript M.

Algorithms for the computation of probabilities on pedigrees have followed one of two paradigms. The first, dating to the early days of human linkage analysis (Fisher 1934; Haldane 1934), considers the probability of phenotypic data $\mathbf{Y}$ as the sum over underlying genotypic configurations $\mathbf{G}$:

$$\Pr_\psi(\mathbf{Y}) = \sum_{\mathbf{G}} \Pr_\psi(\mathbf{Y} \mid \mathbf{G}) \Pr_\psi(\mathbf{G}). \tag{3.1}$$

Algorithms for the computation of this sum rely on the conditional independence structure of genotypes on pedigrees, which permits the summation to be performed sequentially through the pedigree structure. The best-known such algorithms derive from the algorithm of Elston and Stewart (1971), and have come to be known as "(pedigree) peeling" (Cannings, Thompson and Skolnick 1978). Generally, peeling algorithms are linear in the size of the pedigree, but exponential in pedigree complexity as measured by the number of interlocking loops. More seriously, they are exponential in the number of alternative (multilocus) genotypes an individual can have. Hence computation rapidly becomes infeasible as the number of loci increases, especially if the loci are multi-allelic.

An alternative approach also dates back to the earliest days of linkage analysis (Sturtevant 1913; Fisher 1922). This method involves direct observation or inference of the segregation events in an experimental cross, and hence scoring of the recombination events. The segregation events can be specified by "segregation indicators" $\mathbf{S} = \{S_{il}, \ \ i = 1, ..., m, \ l = 1, ..., L\}$ where

$$\begin{aligned} S_{il} &= 0 \quad \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's maternal gene} \\ &= 1 \quad \text{if copied gene at segregation } i \text{ locus } l \text{ is parent's paternal gene.} \end{aligned}$$

Here $i = 1, ..., m$ indexes the segregations, and $l = 1, ..., L$ indexes the genetic loci. Where not all segregation events can be precisely inferred, the probability of observed data $\mathbf{Y}$ may again be considered as a sum:

$$\Pr_\psi(\mathbf{Y}) = \sum_{\mathbf{S}} \Pr_\psi(\mathbf{Y} \mid \mathbf{S}) \Pr_\psi(\mathbf{S}). \tag{3.2}$$

Algorithms based on (3.2) rely on the conditional independence structure of the segregation indicators S_{il}, which permits the summation to be performed sequentially along the chromosome ("chromosomal peeling"): such an algorithm was developed by Lander and Green (1987). This approach is ideal for data on experimental crosses, since, in the absence of missing data, computation is linear in the number of loci. However, computation is exponential in the number of meioses which cannot be directly observed.

4. MCMC estimation of probabilities on pedigrees. For multilocus computations, on complex extended pedigrees, with many of the individuals unobserved, exact computation is infeasible with either approach. Therefore, in recent years, alternative Monte Carlo procedures for the summations in (3.1) and (3.2) have been proposed. Most of these proposals have been of Markov chain Monte Carlo (MCMC) algorithms, which rely on the same conditional independence structures as do the exact algorithms. Most of the proposals to date have considered (3.1), the objective being therefore to sample genotypes $\mathbf{G}$ from their conditional distribution given the data $\mathbf{Y}$. The simplest algorithms involve single-site updating via a Metropolis (Lange and Matthysse 1989) or Gibbs (Sheehan *et al.* (1989)] sampler. That is, the update proposal is of the genotype of a single individual at a single locus.

Such algorithms work well on small examples, but do not mix adequately on large pedigrees, especially where there are many unobserved individuals, and/or data at many loci at which multilocus phase is not easily determined. Moreover, for multi-allelic loci, the partial constraints imposed by data may make the single-site updating MCMC methods reducible. There have been numerous proposals to ensure irreducibility of samplers, and to improve mixing. Some examples are the "heating" methods of Sheehan and Thomas (1993) and of Lin *et al.* (1994), the "tunneling" method proposed by Sobel and Lange (1993), the "mode-jumping" method proposed by Lin (1995), and the simulated tempering approach developed by Geyer and Thompson (1995).

Thompson (1994a) proposed use of the alternative paradigm (3.2), in which MCMC sampling is of the segregation indicators $\mathbf{S}$ conditional upon data $\mathbf{Y}$. Where there are many unobserved individuals on a pedigree, especially for multi-allelic loci, the space of segregation indicators is much smaller than the space of genotypes. It is generally much less constrained by data, except where components are fully determined (see section 7). For the estimation of the posterior probabilities of gene IBD patterns at marker loci, it has the added advantage that the IBD pattern B is fully determined by the segregation indicators $\mathbf{S}$. Consider, for example, the segregation pattern on the pedigree shown in Figure 2. The founder genes are labeled $1, \ldots, 2n$ where n is the number of founders, and

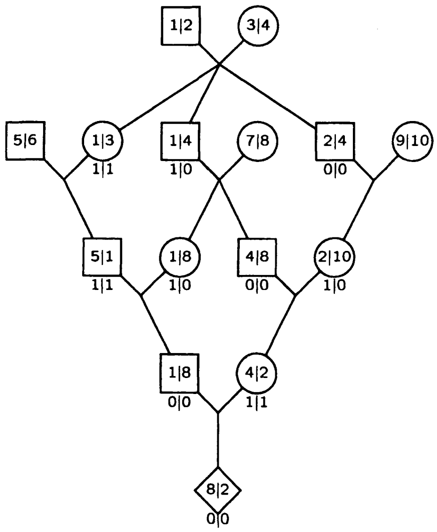

FIG. 2. *The same pedigree structure as Figure 1, showing a particular single-locus realization of the segregation indicators, and the implied gene identities.*

the genes of non-founders are then determined successively by the segregation indicators. In particular, we see than for this realization the final individual does not have two IBD genes. We see also that, although the individual receives his mother's maternal gene (gene "2"), he shares the founder gene "8" IBD with his maternal grandfather.

Thompson (1994b) developed a single-site Metropolis algorithm for sampling the S_{il}, and implemented it in the context of homozygosity mapping where normally the only data are on a single inbred individual in each pedigree. In this case, the segregation-indicator sampler performs much better than the genotypic sampler. However, the single-site updating scheme does not work well when the loci are very tightly linked, since the proposal to update a single locus then often involves the formation of a double-recombinant. Sobel and Lange (1996) also implemented a single-site Metropolis algorithm for $\mathbf{S}$ in a variety of pedigree analysis situations, with similar conclusions as to performance. In this paper, we propose a meiosis-by-meiosis sampler which updates S_{il} jointly for all loci l in a given meiosis i.

5. Implementing the whole-meiosis Gibbs sampler.

For notational convenience we define $S_{*l} = (S_{il}, i = 1, ..., m)$ the vector of segregation indicators at locus l, and $S_{i*} = (S_{il}, l = 1, ..., L)$ the vector of indicators at segregation i.

In order to implement a whole-meiosis Gibbs sampler for S_{i*} we must compute

$$\Pr(S_{i*} \mid \{S_{k*}, k \neq i\}, \mathbf{Y})$$

We suppose that the marker data $\mathbf{Y}$ can be partitioned into data relating to each locus $l = 1, 2, ..., L$, and that the loci are numbered in order along the chromosome. Then $\mathbf{Y} = (Y_1, ..., Y_L)$. Let $Y^{(l)} = (Y_1, ..., Y_l)$, so $\mathbf{Y} = Y^{(L)}$. We suppose also that $\Pr(Y_l \mid S_{*l})$ can be easily computed: we show below how this computation may be done. Now define

$$(5.1) \qquad Q_l(s) = \Pr(S_{il} = s \mid \{S_{k*}, k \neq i\},\ Y^{(l)}) \text{ for } \ s = 0, 1.$$

That is, $Q_l(s)$ is the cumulative probability for the segregation indicator S_{il}, given the data at loci up to and including locus l. Then

$$(5.2) \qquad Q_1(s) \ \propto \ \Pr(Y_1 \mid S_{*1}) \text{ and}$$
$$Q_l(s) \propto \Pr(Y_l \mid S_{*l})\ (Q_{l-1}(s)(1 - \theta_{l-1}) + Q_{l-1}(1 - s)\theta_{l-1}) \text{ for } l = 2, ..., L,$$

where S_{*l} takes the current value at meioses other than i, and the value s for meiosis i, and where θ_{l-1} is the recombination frequency between locus $l - 1$ and locus l. Thus we may compute (5.1) for each l in turn, working forwards sequentially along the chromosome.

Finally we have computed

$$Q_L(s) = \Pr(S_{iL} = s \mid \{S_{k*}, k \neq i\},\ \mathbf{Y} = Y^{(L)})$$

and thus S_{iL} may be sampled from this desired conditional distribution. Suppose S_{ij} has been similarly sampled for $j = l, ..., L$. Then

$$(5.3) \qquad \Pr(S_{i,l-1} = s \mid \{S_{k*}, k \neq i\}, \{S_{ij}, j = l, ..., L\}, \mathbf{Y})$$
$$\propto \ Q_{l-1}(s)\ (\ |S_{il} - s|\theta_{l-1} \ + \ (1 - |S_{il} - s|)(1 - \theta_{l-1}))$$

Thus we may work backwards down the chromosome, sampling each S_{il} in turn ($l = L, ..., 1$), obtaining overall a joint realization of S_{il}, $\ l = 1, ..., L$ from its conditional distribution given $\{S_{k*}, k \neq i\}$ and $\mathbf{Y}$. We note the similarity of this forwards-backwards algorithm (equations (5.2) and (5.3)) along a chromosome to the method of Ploughman and Boehnke (1988), which samples genotypes jointly at a locus by peeling up the pedigree, saving the partial probabilities computed en route, and then sampling down using these partial probabilities. The same method is used in a method to determine feasible genotypic configurations on a pedigree (Heath 1997a), as well as in MCMC genotypic samplers that sample jointly all genotypes at a locus (Kong 1991; Heath 1997b).

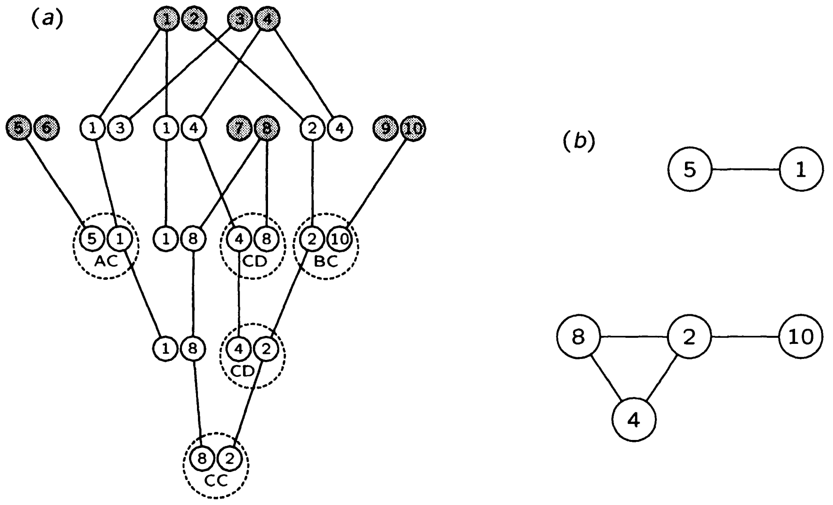

FIG. 3. *(a) The gene pedigree implied by the segregation pattern on the pedigree of figure 2, showing marker data on five individuals, and (b) the gene dependency graph resulting from the segregation pattern of and marker data of (a).*

For completeness we outline briefly the method for efficient computation of the single-locus probability $\Pr(Y_l \mid S_{*l})$, where $S_{*l} = (S_{il},\ i = 1, ..., m)$. We illustrate the calculation with the small but complex pedigree shown in Figure 1. As before, the founder genes in the pedigree are labeled from $1 \ldots 2n$, where n is the number of founders in the pedigree (five in this case). For non-founder individuals, the genes they carry at locus l is determined by S_{*l}, the segregation pattern for that locus (Figure 2). Figure 3(a) shows the same pedigree, but with the individual genes rather than individuals drawn. The ten founder genes (shaded in the figure) have been labeled, and the figure shows the descent of genes to the non-founders for a particular realization of S_{*l}. For this example, five of the individuals are observed; these individuals are marked by the dotted circles on the figure.

Only genes that appear in observed individuals contribute non-trivially to the single-locus probability $\Pr(Y_l|S_{*l})$ so in this example we only have to consider the six genes 1, 2, 4, 5, 8 and 10. However, calculation requires summing over all possible assignments of allelic identities to these six distinct genes (Thompson 1974). A naïve approach for a locus with k alleles would require summing over all k^6 possible combinations of allelic identities. We can improve on this by exploiting the dependence between founder genes. This dependence structure can be shown by a graph whose nodes are the genes that appear in observed individuals. An edge connects two genes if they appear together in an observed individual (Figure 3(b)). If, as in this example, the graph has several components, then each can be considered separately for the purposes of calculating the probability. For codominant markers each component can have at most two possible joint assignments of allelic identities for the genes in the cluster, so the probability calculation becomes trivial (Sobel and Lange 1996). For this example the gene cluster (1, 5) has 2 possible allelic assignments, (A, C) or (C, A), and the cluster (2, 4, 8, 10) has 1 possible assignment, (C, D, C, B). For general loci, the graph of Figure 3(b) defines a conditional independence structure. The desired probability can thus be calculated efficiently by "peeling" the allelic assignment of types to founder genes, in a method analogous to pedigree peeling (Elston and Stewart 1971; Cannings *et al.* 1978) or chromosomal peeling (Lander and Green 1987). Additional details of the calculation are given in Heath and Thompson (in preparation).

Any MCMC sampler needs an initial configuration for the latent variables. In the small examples considered here, values of **S** that are consistent with the data were found by hand. For larger or more complex examples, we may use existing methods for obtaining genotypic configurations **G** that are consistent with observed data **Y** even for highly polymorphic loci on large and complex pedigrees (Heath 1997a). Where the pedigree can be peeled for single-locus data, the initial configuration is from the required equilibrium distribution marginally for each locus. The method produces an ordered genotype for each individual, and this genotypic configuration then provides the implied segregation indicators. These indicators are then necessarily also consistent with the data **Y**. This is not necessarily the best way to obtain a starting configuration **S**, but it is a possible and practical way for which the programs already exist.

6. Performance of the sampler: two examples. For examples, we use the small but complex pedigree (Figure 1), considered in the previous sections. We consider first the case of homozygosity mapping which was the objective of the original study (Goddard *et al.* 1997). Only marker data on the final affected individual were available. This pedigree provides a useful example, since while it is easily analyzed by MCMC methods, exact computation by pedigree peeling is infeasible for more than about four loci, due to the three interlocking loops. Due to the 20 meioses in the pedigree, this example is also close to the limits of feasibility for exact computation using chromosomal peeling.

For homozygosity mapping, the question is of the extent that patches of marker homozygosity imply gene identity by descent in the sampled individual. As output from our sampler, we therefore score the IBD pattern at a set of homozygous marker loci. The single-site Metropolis sampler was previously implemented for this case (Thompson 1994a,b), and we now compare this with the whole-meiosis Gibbs sampler for the same situation. As an example, we con-

TABLE 1

Autozygosity probabilities conditional on homozygosity at five tightly linked loci (recombination = 0.02), as a function of the frequency of the homozygous marker allele. There are 2 states not involving any switches between N (non-IBD) and I (IBD), 8 states involving one switch (e.g. N N N I I), 12 involving two switches (e.g. I I N N I), 8 involving three switches (e.g. N I I N I) and two involving four switches (e.g. N I N I N), for a total of $2^5 = 32$ possible patterns.

pattern of IBD	$q = 1.0$	$q = 0.5$	$q = 0.1$
$N\ N\ N\ N\ N$	0.8389	0.2362	0.00013
$I\ \ I\ \ I\ \ I\ \ I$	0.0644	0.5819	0.9673
1-switch (8)	≈ 0.08	≈ 0.15	≈ 0.03
2-switch (12)	≈ 0.01	≈ 0.03	≈ 0.001
other (10)	≈ 0	≈ 0	≈ 0
single-locus Pr(I)	0.1094	0.1972	0.5512

sider five equally-spaced marker loci (L1 to L5), with recombination frequency 0.02 between adjacent loci. Table 1 shows the results as a function of q, the frequency of the allele for which the observed individual is homozygous.

Figure 4 compares the cumulative IBD probabilities at each locus for the single-site and whole-meiosis samplers in the case when the marker allele frequency is 0.5. To provide a fair comparison, there are five times as many single-site updates as whole meiosis updates. (Each total run is 10,000,000 whole-meiosis updates, or 50,000,000 single-site updates.) Clearly the whole-meiosis sampler has much better mixing, and provides more reliable results. Furthermore, the CPU time for the run using the whole-meiosis sampler was only about 2/3 of that for the single-site sampler: 10 million whole-meiosis Gibbs updates took 328 secs CPU, while 50 million single-site updates took 467 seconds on a DEC Alpha 400M workstation. (The efficiencies of the two programs are quite comparable. Each could be further optimized.)

Over the five loci, as expected, the central locus (L3) of the five homozygous markers has the highest IBD probability, followed by the next two loci (L2, L4), with the end loci (L1, L5) having the lowest IBD probability. For these very tightly linked loci the differences in IBD probability are not large, but they are non-negligible. For both samplers we see very strong correlations among the five loci for the cumulative IBD probability; the five paths track each other closely. This correlation is due to the tight linkage. For unlinked loci, there are no such correlations, even when the sampler is run jointly on the loci from a common starting configuration (results not shown).

We now consider also the case where other individuals are observed on this same pedigree structure; for example, the data of Figure 3, shown again in Figure 5. We see the C allele labeled C_2 must descend from grandparent to parent, to the final individual, while the two D alleles must also be IBD. However, the

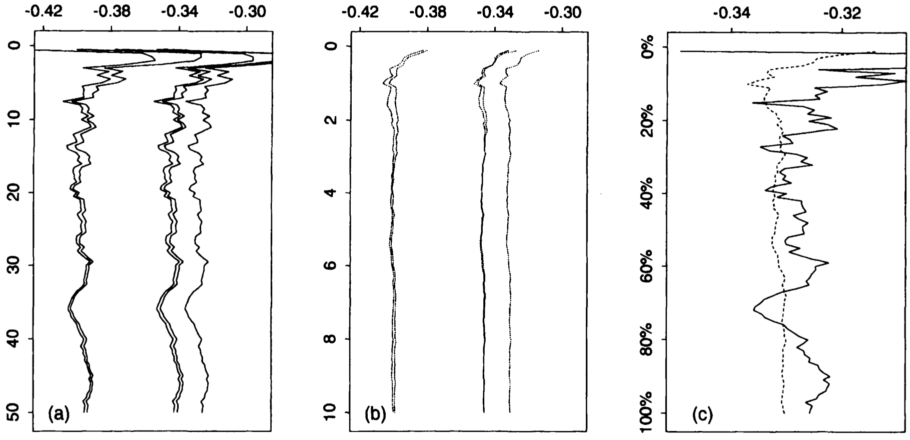

FIG. 4. *Comparison of the single-site Metropolis sampler and the whole-meiosis Gibbs sampler for the example of homozygosity mapping. In each case there are five linked marker loci, with recombination frequency 0.02 between adjacent loci, and the allele frequency of the marker allele for which the final individual is assumed homozygous is 0.5 at each locus. On the horizontal axis of each graph is the* $\log_e$ *of the probability of gene identity by descent between the two haplotypes of the final individual. Plotted are the cumulative estimates over each run, for each locus. (a) Plot for the five linked loci, over a total of 50 million single-site updates. (b) Plot for the five linked loci, over a total of 10 million whole-meiosis Gibbs updates. (c) For easier comparison, the curve for the central one of the five loci, for each of (a) and (b)*

other three C alleles, labeled C_1, C_3 and C_4 in Figure 5, may or may not be IBD to each other or to C_2. In fact, each of the 15 partitions of these four genes is possible, given the data and the pedigree structure. A question of interest might be the posterior probabilities of the patterns of gene identity among these four potentially distinct C alleles.

As above we consider five linked marker loci. To assess the effect of tight linkage, we present results for two different recombination frequencies between adjacent loci; tight linkage ($\theta = 0.02$), and loose linkage ($\theta = 0.1$). We assume the same marker data (Figure 5) at each of the five loci. At each locus, the same allele frequencies are assumed; the allele C has frequency 0.4, and each of the other alleles (A, B, and D) has frequency 0.2. Table 2 gives the results, again as a function of the frequency of allele C. For comparison, we give also the single-locus probability, and also the prior probability of identity patterns among the four genes, given only the pedigree structure.

Each of the probabilities in Table 2 is estimated from 10 million whole-meiosis Gibbs steps. For a problem with 5 loci on this size of pedigree, such a run takes just under 45 minutes CPU on a DEC Alpha 400M workstation. As for the homozygosity probabilities, examination of the cumulative state probabilities

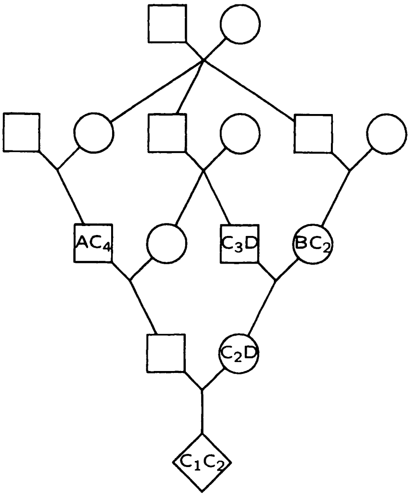

FIG. 5. *The pedigree structure of Figure 1, with the data of Figure 3, and the four possibly distinct C alleles labeled.*

over the run show that the sampler mixes well.

7. Irreducibility: neither necessary nor sufficient. There are many ways to make an MCMC sampler irreducible. Note first that, provided recombination frequencies are strictly positive, irreducibility is a single-locus question. Thus, irreducibility (or otherwise) is the same for a single-segregation-indicator updating sampler as for the whole-meiosis Gibbs sampler. As we have seen, the mixing properties of these two samplers can differ greatly. Irreducibility is not a sufficient criterion for a sampler; in practice, a more important question concerns its mixing properties.

However, there are some interesting features of the irreducibility properties of a sampler based on $\overset{<}{\underset{>}{=}}$, which illuminate the structure of the problem. Note also that, whereas $\between$ constrains the allelic types of genes, $\overset{<}{\underset{>}{=}}$ constrains only which genes must be of like allelic type. Thus there are many examples in which a genotypic sampler is reducible, but in which a segregation-indicator sampler is irreducible. Generally, irreducibility of the MCMC sampler of $\overset{<}{\underset{>}{=}}$ can fail when the allelic types of founder genes are constrained, either directly through founders of observed genotype or through constraints on the number of distinct founder genes.

Reducibility is not a problem for homozygosity mapping when only a single inbred individual is observed in a pedigree (Thompson 1994b). At a given locus l, the two genes are either IBD or not, and the latter state is necessarily

TABLE 2

Probabilities of the fifteen gene identity patterns among the four potentially distinct C alleles The frequency of marker allele C is assumed to be 0.4 at each locus, and the recombination frequency between adjacent loci is either 0.02 or 0.1 as indicated.

pattern	$\theta = 0.02$			$\theta = 0.1$			single locus	
$C_1C_2C_3C_4$	L3	L2,L4	L1,L5	L3	L2,L4	L1,L5	post.	prior
$g\ g\ g\ g$	0.598	0.590	0.561	0.382	0.359	0.287	0.049	0.029
$g\ g\ g\ h$	0.216	0.214	0.207	0.191	0.183	0.155	0.039	0.060
$g\ g\ h\ g$	0.038	0.039	0.046	0.076	0.079	0.091	0.091	0.137
$g\ h\ g\ g$	0.049	0.051	0.059	0.102	0.107	0.121	0.118	0.088
$g\ h\ h\ h$	0.006	0.007	0.010	0.014	0.016	0.021	0.013	0.070
$g\ g\ h\ h$	0.003	0.003	0.004	0.007	0.007	0.008	0.006	0.002
$g\ h\ g\ h$	0.016	0.017	0.019	0.030	0.032	0.035	0.039	0.012
$g\ h\ h\ g$	0.037	0.039	0.046	0.081	0.086	0.100	0.084	0.229
$g\ g\ h\ u$	0.004	0.004	0.005	0.011	0.012	0.015	0.026	0.004
$g\ h\ g\ u$	0.022	0.022	0.026	0.047	0.050	0.059	0.125	0.123
$g\ h\ u\ g$	0.004	0.005	0.008	0.031	0.036	0.057	0.270	0.164
$g\ h\ u\ h$	0.001	0.001	0.001	0.004	0.005	0.009	0.027	0.055
$g\ h\ h\ u$	0.005	0.006	0.008	0.015	0.017	0.023	0.026	0.088
$g\ h\ u\ u$	0.001	0.001	0.001	0.004	0.005	0.009	0.026	0.052
$g\ h\ u\ v$	0.001	0.001	0.001	0.004	0.005	0.010	0.066	0.055

consistent with the data. Sobel and Lange (1996) give an example where the single-site Metropolis sampler of $\{S_{il}\}$ is reducible, due to severe constraints on the types of founder genes. In practice, we are unlikely to have fully observed homozygous founders, except perhaps in crosses among inbred lines. Note that data on descendants can never force an unobserved ancestor to be homozygous, and nor can data on descendants alone force the maternal/paternal origins of genes in an unobserved ancestor. Thus the example of Sobel and Lange (1996) is unlikely to be a practical concern in human genetics.

However, reducibility can also arise from restrictions on the number of founder genes, or number of genes available to segregate to observed descendants. The segregation indicators S_{*l} define a partition of the ordered genes at locus l in observed individuals determining which are identical by descent, and hence must be of like allelic type. Note that if any given partition is consistent with the data, then any finer partition must be so also.

Figure 6(a) shows an example in which there are three full sibs with unobserved parents. The sibs have genotypes AB, AC and BC as shown. Note that any one of the three alleles must be present in each parent, and the other two must be represented once only among the four founder genes. In each case, one pair of sibs share no genes IBD with each of the pair sharing one gene IBD with the third sib. The set of six observed genes at each locus are partitioned into four sets of IBD genes, two partitions size two, and two partitions size one.

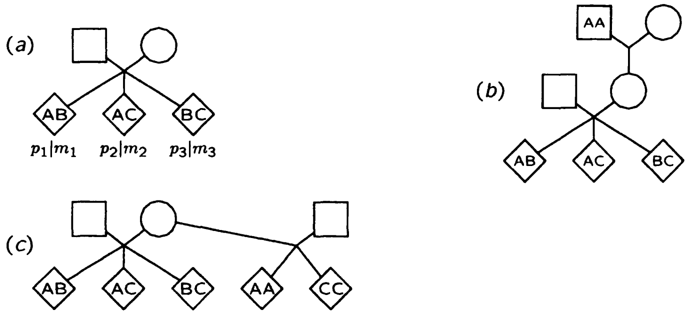

FIG. 6. *Examples of failure of irreducibility. (a) Data on a sibship of size three. (b) An added maternal grandfather (case (b) in Table 3). (c) Added information on maternal genotype (case (c) in Table 3).*

States are compatible with the data if two of the sibs have indicators $(0,1)$ and $(1,0)$, and the third is either $(1,1)$ or $(0,0)$, or if two of the sibs have indicators $(0,0)$ and $(1,1)$, and the third is either $(1,0)$ or $(0,1)$. Of the 64 potential values of the 6 binary segregation indicators, 24 are consistent with the data (Table 3).

If only a single meiosis is updated in a given step, the 24 feasible states fall into two disjoint cycles of length 12. The sampler is reducible. In Table 3, the states are listed in order so that each is obtainable from the previous one by updating a single meiosis. Each feasible state implies an ordered pair of parental genotypes. In Table 3, the parental genotypes are given as implied by the first column of 12 states. For the second column of states, the father's genotype is reversed. Thus, although the space of $\lesseqgtr$ values that are consistent with the phenotypic data is divided into two non-communicating classes, this does not affect MCMC estimation of probabilities. The founder (i.e. parental) genotypes, whose probabilities determine the contribution of a state to any overall probability, are identical on each of the two communicating classes. Thus in this example we see that irreducibility is not necessary in order to obtain correct MCMC probability estimates at a single locus.

We can extend this example. Suppose first the mother's father is known to be of type AA, forcing the mother's paternal gene to be of type A (Figure 6(b)). Then only the 8 states labeled "*b*" in Table 3 are consistent with the data, and these fall into two non-communicating classes, each of size 4. However, again the same set of parental genotypes, and hence the same probability contributions, are implied by each class of states.

Or, we could suppose the existence of other maternal half sibs (e.g. AA and CC) forcing the mother to be of type AC (Figure 6(c)). Then only the 8 states labeled "*c*" are consistent with the data. Now we have four non-communicating classes, each of size two. Again, the same parental genotype combinations are

TABLE 3

The feasible states for the three-sibs example. States are specified as $(p_1, m_1, p_2, m_2, p_3, m_3)$ *the paternal and maternal indicators for each of the three sibs. The data are that the sibs have genotypes AB, AC and BC. For further details, see text.*

sib 1	sib 2	sib 3	sib 1	sib 2	sib 3		parental types	
$p_1, m_1,$	$p_2, m_2,$	p_3, m_3	$p_1, m_1,$	$p_2, m_2,$	p_3, m_3		father	mother
							p m	*p m*
0 0	0 1	1 0	1 0	1 1	0 0		*CA*	*CB*
0 0	0 1	1 1	1 0	1 1	0 1		*BA*	*CB*
1 0	0 1	1 1	0 0	1 1	0 1	*c*	*BA*	*CA*
1 0	0 0	1 1	0 0	1 0	0 1	*c*	*BC*	*CA*
1 0	0 0	0 1	0 0	1 0	1 1		*BC*	*BA*
1 1	0 0	0 1	0 1	1 0	1 1		*AC*	*BA*
1 1	1 0	0 1	0 1	0 0	1 1		*AC*	*BC*
1 1	1 0	0 0	0 1	0 0	1 0		*AB*	*BC*
0 1	1 0	0 0	1 1	0 0	1 0	*b c*	*AB*	*AC*
0 1	1 1	0 0	1 1	0 1	1 0	*b c*	*CB*	*AC*
0 1	1 1	1 0	1 1	0 1	0 0	*b*	*CB*	*AB*
0 0	1 1	1 0	1 0	0 1	0 0	*b*	*CA*	*AB*

implied by each of the four classes. We could further require both the above conditions to be met, reducing the feasible states space to only four states in two classes (those labeled both "*b*" and "*c*"). Yet again the parental genotypes are the same in each class of communicating feasible states. Thus the symmetries of this example mean that irreducibility is unnecessary in obtaining valid MCMC single-locus probability estimates.

However, although single-locus irreducibility implies multi-locus irreducibility, validity of single-locus probabilities does not imply validity of multilocus probabilities in a reducible sampler. (We are indebted to Ken Lange for drawing our attention to this fact.) A simple example will suffice; consider the phenotypic data of Figure 6(b) at each of two very tightly linked loci. If both loci are initialized in the same class of four feasible states, approximately correct probabilities are obtained. If one locus is initialized in each of the two different sets of four states, at least two recombination events are required in the six meioses. For very tight linkage, absolute probabilities are almost negligible, but the relative probabilities bias the sampler towards the states where only these two recombination events are required. In this example, the sampler is biased towards states where C is non-IBD at one locus and B is non-IBD at the other, and away from the states where each parent carries an A allele. In general, it is not easily determined whether irreducibility is necessary.

In examples we have considered, irreducibility fails due to constraints of equality or inequality of segregation indicators from a given parent. There is a strong constraint on the number of available genes, since the parent individual can have at most two distinct genes at a locus. Simultaneous updating of all the meioses from a given individual, or even parental couple, will often be possible, using the computational algorithm of Kruglyak *et al.* (1995), and is a practical way to obtain irreducibility in many cases. However, it is not a universal solution, as is shown by the example of Sobel and Lange (1996) in which two indicators in different sibships are constrained to be unequal.

8. Discussion. As maps become both denser and more precise, there is an increasing demand for multipoint linkage analysis on large and complex pedigrees. Exact computations become infeasible, necessitating the use of Monte Carlo or other approximation methods. The whole-meiosis Gibbs sampler presented here is just one of many possible Monte Carlo algorithms. It is easily implemented, and mixes much better than any single-site MCMC method, particularly when there is tight linkage. In some cases, multi-site genotypic samplers can be implemented. In particular, where pedigree peeling is feasible for each locus separately, a whole-locus Gibbs sampler is possible (Kong 1991; Heath 1997b), and is likewise a great improvement over a single-site updating genotypic scheme. Genotypic samplers work well if there are few missing data on the pedigree, but where there are many unobserved individuals we expect the whole-meiosis sampler to have better performance. For very complex pedigrees, even single-locus (pedigree) peeling is computationally intensive, whereas implementation of the whole-meiosis sampler is almost unaffected by pedigree complexity. For very tight linkage, any sampler will have decreased mixing. However, whereas the performance of a genotypic sampler can be severely adversely affected by tight linkage, the whole-meiosis Gibbs sampler performed well, even for multiple loci at recombination frequencies as low as 0.02.

Exact computational and genotypic sampling methods preclude the inclusion of interference for more than three loci, except on very small pedigrees with few missing data (Lin and Speed 1996), However, a meiosis sampler can include interference in the computation of linkage likelihoods or conditional probabilities of genome sharing. The recombination events within a meiosis may be jointly sampled, or a recombination location may be resampled conditionally upon the locations of other recombinations in the same meiosis. Equations 5.2 and 5.3 become more complex, since dependence in S_{il} extends beyond the adjacent marker loci, but the same approach is computationally feasible.

For almost any sampler, irreducibility is a non-trivial question. Generally, segregation indicators provide fewer constraints on genes. Except where founder genes are constrained in either type or number the sampler will often be irreducible. However, in pedigrees with few founders and for loci with many alleles, it is possible for irreducibility of the whole-meiosis Gibbs sampler to fail. In the examples of section 6, irreducibility is easily established. For homozygosity mapping, irreducibility is trivially satisfied. For the case of the five observed individuals on the same pedigree, one segregation indicator is completely determined; the final individual receives the C_2 allele from his mother's mother. However, parental origins of genes cannot be determined from data on descendants alone; other segregation indicators are freely varying. Where an indicator is completely determined, this reduces the size of the space to be sampled, but of course does not affect the irreducibility of the sampler. (The same is true of determined genotypes in a genotypic sampler.)

This fact could be used to improve the efficiency of the sampler, and to simplify consideration of irreducibility, by conditioning on the segregation from any founder having only one offspring. Such a founder provides no information for linkage, and the conditioning simply determines the ordering of the genes in the founder. This method was used by Thompson (1994b) to improve efficiency of the single-site Metropolis sampler for homozygosity mapping, and has been extended to the whole-meiosis sampler. Similar considerations have been used by Kruglyak el al. (1995) to extend feasibility of exact computational methods using the approach of Lander and Green (1987). Even where a founder has multiple offspring, at a single locus the segregation to a given offspring (say the eldest) may be constrained. However, where a founder has multiple offspring and data at multiple loci are available on descendants, such data can provide partial information on which offspring of the founder are recombinant. It is necessary for the computation to allow for alternate recombination patterns in the meiosis from founder to each offspring, within each linkage group. Similarly to the situation with a reducible MCMC sampler, caution is necessary in ensuring that correct multilocus probabilities are computed.

Where recombination frequencies are strictly positive, irreducibility can be assessed on a single-locus basis. Thus, in particular, irreducibility of the single-site segregation-indicator sampler is the same as for the whole-meiosis Gibbs sampler. More importantly, where a whole-locus genotypic Gibbs sampler is feasible, it is necessarily irreducible. This guarantee of irreducibility, balanced against the greater computational burden and possibly poorer mixing of the genotypic updates, raises the attractive possibility of combining the two samplers. Since a given segregation configuration can be used easily to obtain a genotypic realization, and a genotypic configuration supplies a segregation configuration, the combination of the two samplers is quite practical, wherever single-locus peeling is feasible. In some situations, a sampler interleaving whole-meiosis and whole-locus updates has better mixing properties than either alone (Heath and Thompson 1997). We intend a more detailed study of the whole-meiosis Gibbs sampler, the whole-locus Gibbs sampler, and of samplers which combine the two approaches, to determine the preferred sampler under a variety of pedigree structures and linkage patterns (Heath and Thompson [in preparation]).

Acknowledgments. We are very grateful to Dr. Bernt Guldbrandtsen and to a referee for careful reading of the details of this paper.

REFERENCES

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

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

FISHER, R. A. (1922b). On the systematic location of genes by means of crossover observations. *American Naturalist* **56** 406–411.

FISHER, R. A. (1934). The amount of information supplied by records of families as a function of the linkage in the population sampled. *Ann. Eugen.* **6** 66–70.

GEYER, C. J. and THOMPSON, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.

GODDARD, K. A. B, YU, C-E., OSHIMA, J., MIKI, T., NAKURA, J., PIUSSAN, C., MARTIN, G. M., SCHELLENBERG, G. D., WIJSMAN, E. M. (1996). Towards localization of the Werner syndrome gene by linkage disequilibrium and ancestral haplotyping: lessons learned from analysis of 35 chromosome 8p11.1-21.1 markers. *American Journal of Human Genetics* **58** 1286–1302.

HALDANE, J. B. S. (1934). Methods for the detection of autosomal linkage in man. *Annals of Eugenics* **6** 26–65.

HEATH, S. C. (1997a). Generating consistent genotypic configurations for multi-allelic loci and large complex pedigrees. *Human Heredity* **48** 1–11.

HEATH, S. C. (1997b). Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *American Journal of Human Genetics* **61:** 748–760.

HEATH, S. C. and THOMPSON, E. A. (1997). MCMC Samplers for multilocus analyses on complex pedigrees. *American Journal of Human Genetics* **61:** A278.

KONG, A. (1991). Analysis of pedigree data using methods combining peeling and Gibbs sampling. *Computer Science and Statistics: Proceedings of the 23rd Symposium on the Interface,* (E.M. Keramidas and S. M. Kaufman, eds.) Pp 379–385. Interface Foundation of North America, Fairfax Station, VA.

KRUGLYAK, L., DALY, M. J., and LANDER, E. S. (1995). Rapid multipoint linkage analysis of recessive traits in nuclear families including homozygosity mapping. *American Journal of Human Genetics* **56** 519–527.

LANDER, E. S. and BOTSTEIN, D. (1987). Homozygosity mapping: a way to map human recessive traits with the DNA of inbred children. *Science* **236:** 1567–1570.

LANDER, E. S. and GREEN, P. (1987). Construction of multilocus linkage maps in humans. *Proceedings of the National Academy of Sciences (USA)* **84** 2363–2367.

LANGE, K. and MATTHYSSE, S. (1989). Simulation of pedigree genotypes by random walks. *American Journal of Human Genetics* **45** 959–970.

LIN, S. (1995). A scheme for constructing an irreducible Markov chain for pedigree data. *Biometrics* **55** 318–322.

LIN, S. and SPEED, T. P. (1996) Incorporating crossover interference into pedigree analysis using the χ^2 model. *Human Heredity* **46:** 315–322.

LIN, S., THOMPSON, E. A. and WIJSMAN, E. M. (1994). A faster mixing algorithm for Hastings-Metropolis updates on complex pedigrees. *Annals of Human Genetics* **58** 343–357.

PLOUGHMAN, L. M. and BOEHNKE, M. (1989). Estimating the power of a proposed linkage study for a complex genetic trait. *American Journal of Human Genetics* **44** 543–551.

SOBEL, E. and LANGE, K. (1993). Metropolis sampling in pedigree analysis. *Statistical Methods in Medical Research* **2** 263–282.

SOBEL, E. and LANGE, K. (1996). Descent graphs in pedigree analysis: Applications to haplotyping, location scores and marker-sharing statistics. *American Journal of Human Genetics* **58** 1323–1337.

SHEEHAN, N. A., POSSOLO, A. and THOMPSON, E. A. (1989). Image processing procedures applied to the estimation of genotypes on pedigrees. *American Journal of Human Genetics* **45** (Suppl.) A248.

SHEEHAN, N. A. and THOMAS, A. W. (1993). On the irreducibility of a Markov chain defined on a space of genotype configurations by a sampling scheme. *Biometrics* **49** 163–175.

STURTEVANT, A. H. (1913). The linear association of six sex-linked factors in *Drosophila*, as shown by their mode of association. *J Exp Zool* **14** 43–59.

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

THOMPSON, E. A. (1994a). Monte Carlo estimation of multilocus autozygosity probabilities. *Proceedings of the 1994 Interface Conference* J. Sall and A. Lehman (eds.) Pp. 498–506, Interface Foundation of North America: fairfax Station, VA.

THOMPSON, E. A. (1994b). Monte Carlo likelihood in linkage analysis. *Statistical Science* **9** 355-366.

WIJSMAN, E. M. and AMOS, C. I. (1997). Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: Summary of GAW10 contributions. *Genetic Epidemiology* **14** 719–735.

DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322, SEATTLE WA 98195-4322
THOMPSON@STAT.WASHINGTON.EDU

LABORATORY OF STATISTICAL GENETICS
THE ROCKEFELLER UNIVERSITY
NEW YORK, NY 10021
HEATH@LINKAGE.ROCKEFELLER.EDU