Genetic Epidemiology 14:1–15 (1997)

Computation of Multilocus Prior Probability of Autozygosity for Complex Inbred Pedigrees

Sun-Wei Guo*

Department of Biostatistics, University of Michigan, Ann Arbor

Homozygosity mapping is a very powerful method for mapping rare recessive diseases in humans. In many applications, it is often desirable to compute prior (or unconditional) multilocus probability of autozygosity for inbred pedigrees. This paper proposes a simple yet powerful method for computing the prior multilocus autozygosity probability for complex inbred pedigrees. The method has an added feature of providing explicit multilocus autozygosity probability in terms of recombination fractions, if desired. An example is presented to illustrate the method. Genet. Epidemiol. 14:1–15, 1997. © 1997 Wiley-Liss, Inc.

Key words: autozygosity; gametogenesis process; Hamming distance; homozygosity by descent; inbreeding; Markov chain; pedigrees

INTRODUCTION

Homozygosity mapping is a very powerful method for mapping rare recessive diseases in humans [Smith, 1953; Lander and Botstein, 1987]. Smith [1953] first pointed out that individuals affected with rare recessive diseases provide information for linkage analysis, even without any marker or phenotypic data on other relatives. The scope of homozygosity mapping was further extended by Lander and Botstein [1987]. For rare recessive diseases, many affected individuals receive two copies of genes at the disease locus identical by descent, or homozygous by descent (HBD), or autozygous, from a recent common ancestor through consanguineous marriages.

The crux of homozygosity mapping is the recognition of the fact that, for those affected individuals receiving, as a result of consanguineous marriage, two copies of mutant alleles HBD at the disease locus, the markers surrounding the disease lo-

Contract grant sponsor: National Institutes of Health; Contract grant number: R29-GM52205.

*Correspondence to: Sun-Wei Guo, Ph.D., Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI 48109-2029.

Received 24 January 1996; Accepted 6 July 1996

cus also tend to be inherited HBD. Thus, by detecting unusually elevated sharing of marker alleles HBD in affected individuals who are offspring of consanguineous marriages, it is feasible to detect the location of the disease gene. However, one may not observe such elevated HBD sharing for markers that are more distant from the disease locus. To ensure a successful mapping by homozygosity, therefore, it is important to type markers at appropriate spacings, for a genetic map that is too sparse will not guarantee successful homozygosity mapping and a map that is too dense would be a waste of resources.

For a given inbred pedigree, if I_D, I_1, and I_2 denote the events that the affected individual is HBD at the disease locus, at the marker locus 1, and at the marker locus 2, respectively, where loci 1 and 2 are two markers flanking the disease locus, then $P(I_1 \cup I_2 \mid I_D)$ is the conditional probability that the individual is HBD at either marker 1 or 2 given he is HBD at the disease locus. If we make a conservative assumption that the disease locus is in the middle of the interval flanked by markers 1 and 2, then $P(I_1 \cup I_2 \mid I_D)$ measures the probability that one will find HBD at either marker 1 or 2 as a function of the genetic distance, θ, between the two markers. Obviously, a genetic map that is too sparse corresponds to large θ, which in turn corresponds to a small value of $P(I_1 \cup I_2 \mid I_D)$. Since $P(I_1 \cup I_2 \mid I_D) \propto P(I_1 I_D) + P(I_D I_2) - P(I_1 I_D I_2)$, one way to determine the map density appropriate for homozygosity mapping is to evaluate the multilocus prior probabilities of homozygosity for given θ. If multiple markers are used in homozygosity mapping, the right map density can also be determined by computing the multilocus prior probabilities of autozygosity. Thus, there is a need for a method for computing such probabilities.

Thompson [1994] proposed a Markov chain Monte Carlo method to *estimate* numerically multilocus autozygosity probability. Kruglyak et al. [1995] recently proposed an algorithm to compute *numerically* the exact autozygosity probability. Both methods can be used to compute or estimate the multilocus autozygosity probability, conditional or unconditional on observed marker information. However, the algorithm of Kruglyak et al. [1995] is limited by the size of the pedigree of interest. More precisely, it is limited by the number of meiosis in the pedigree. Thompson's [1994] method, while versatile, can be quite computer-intensive. More recently, Guo [1996] proposed a method for computing the exact prior autozygosity probability, but the method is also limited to small pedigrees.

In this paper, based on the work of Guo [1995, 1996], a new, simple, yet powerful method for computing the unconditional multilocus autozygosity probability is proposed for complex inbred pedigrees of moderate size. The method has an added feature of providing explicit multilocus autozygosity probability in terms of recombination fractions (not just numerical values), if desired. A numerical example is given to illustrate the proposed method.

METHODS

Set-Up

To make the presentation self-contained, some definitions introduced in Guo [1995] are reviewed. The following assumptions are made throughout the paper: 1) no mutation, translocation, conversion, deletion, or insertion; 2) the founders in the pedigree are biologically unrelated; 3) no sex difference in map length; and 4) no interference. The assumption of no interference implies independent, exponentially distributed intervals with mean 1 (Morgan) between crossovers, which was first noted by Fisher [1949] and leads to Haldane's map function [Haldane, 1919].

Following Guo [1995], we can, for a given pedigree, label the maternal and paternal chromosomes for each individual as 0 and 1, respectively. For founders whose parents are not in the pedigree, the labeling is arbitrary. Define a stochastic process, $g(t)$, as the process taking value 0 or 1 at chromosome location t $(0 \leq t \leq l)$, depending on whether the maternal or paternal gene at t is transmitted, where l is the length of the chromosome (in Morgans). This process, termed the gametogenesis process [Guo, 1995], has been shown to be a time-continuous, two-state Markov chain with the transition probability matrix:

$$P(t) = (p_{ij}(t)) = \frac{1}{2}\begin{pmatrix} 1+e^{-2t} & 1-e^{-2t} \\ 1-e^{-2t} & 1+e^{-2t} \end{pmatrix} = \begin{pmatrix} 1-\theta(t) & \theta(t) \\ \theta(t) & 1-\theta(t) \end{pmatrix}$$

where $\theta(t)$ is the recombination fraction for two loci at distance t Morgans apart, and $p_{ij}(t) = P(g(t) = j \mid g(0) = i)$ for $i, j = 0, 1$. See Guo [1995] for more details about the process.

The gametogenesis process thus defined has the following properties. First, at any point t $(0 \leq t \leq l)$ along the chromosome, $g(t) = C$, where C is a random variable taking values of 0 or 1 with equal probability. Second, all gametogenesis processes for different individuals in a pedigree are stochastically identical and independent. Third, for any two loci t Morgans apart,

$$P(g(t) = C \mid g(0) = C) = 1 - \theta(t)$$

where $\theta(t)$ is the recombination fraction for two loci at distance t Morgans apart.

For convenience, some frequently used terms are defined below. They include the gene-transmission pedigree, and the HBD event and the corresponding event set, $\mathcal{D}$. Only brief descriptions are given here; the reader is referred to Guo [1995] for further details. A *gene-transmission pedigree* is a pedigree in which 1) for each individual in the pedigree, two homologous (one maternal and one paternal) alleles are displayed; and 2) for non-founders, the origins of their genes, or the transmission paths, are marked. For simplicity, sometimes only one allele for some individuals is displayed because of the impossibility of sharing genes HBD for the other allele. For each individual, the maternal and paternal alleles are displayed as left and right, respectively. By convention, an individual either has both parents present in the pedigree or neither.

For a given pedigree and a specific gene HBD event of interest, one can depict an appropriate gene-transmission pedigree. On this basis, n gametogenesis processes, say, $g_1(t), \ldots, g_n(t)$, that are relevant to the event can be identified [Guo, 1995]. For the inbred pedigree in Figure 1a, e.g., if we are interested in the event that individual M (the one at the bottom) is HBD, then there are 18 gametogenesis processes that are relevant (Fig. 1b). The new process $v(t) = (g_1(t), g_2(t), \ldots, g_n(t))$, called the *joint gametogenesis process*, constitutes a random walk on an n-dimensional hypercube $\mathcal{Z}^n = \{(\eta_1, \eta_2, \ldots, \eta_n) : \eta_i = 0 \text{ or } 1\}$ [Donnelly, 1983; Guo, 1995]. The random walk on $\mathcal{Z}^n$ has been shown to be identical to a time-continuous, discrete-state, stochastic process with 2^n states. At any locus t along the genome, $v(t)$ is in a state

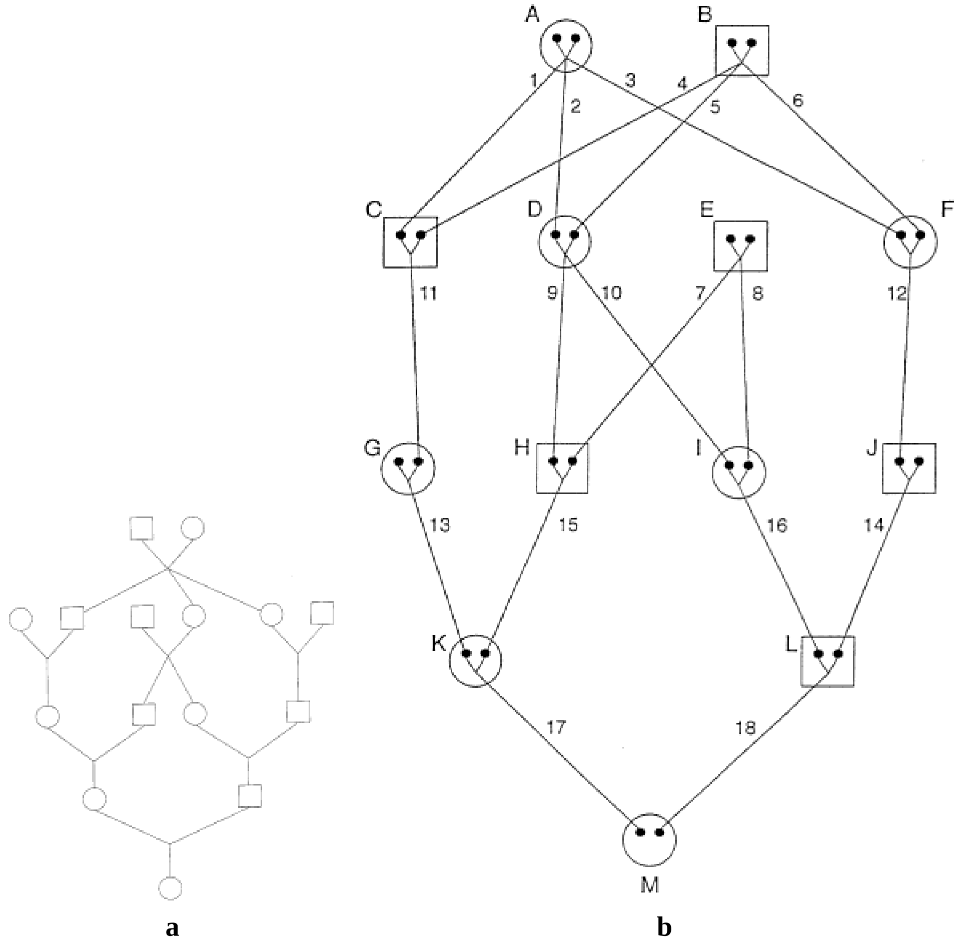

Fig. 1. **a:** An example pedigree. **b:** The gene-transmission pedigree. The numbers are 18 gametogenesis processes.

which corresponds to an element in $\mathcal{Z}^n$, which, in turn, corresponds to a vertex on the cube. Thus, for any $z \in \mathcal{Z}^n$, where $z = (z_1, z_2, \ldots, z_n)$, $v(t) = z$ means $g_i(t) = z_i$ ($i = 1, \ldots, n$). Any particular z corresponds, for the individual of interest (e.g., M in Fig. 1a), to the event that the individual's genes that are either HBD or not. For the pedigree shown in Figure 1b, e.g., $v(t) = (0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0)$ corresponds to the event that individual M receives two copies of genes HBD at locus t from individual A's maternal chromosome. Hence, there are two HBD events: genes HBD and genes not HBD. The collection of those elements in $\mathcal{Z}^n$ that correspond to the event that genes are HBD is called the HBD set, denoted $\mathcal{D}$. The process $v(t)$ has an equilibrium distribution $P(v(t) = z) = \frac{1}{2^n}$ for any $z \in \mathcal{Z}^n$ and for any $0 \leq t \leq l$.

Note that the coordinates of any element z in $\mathcal{Z}^n$ are either 0 or 1. If we define an addition operation $\bigoplus$ for any two elements $z_1, z_2 \in \mathcal{Z}^n$, where $z_i = (z_{i1}, z_{i2}, \ldots, z_{in})$, as

$$z_1 \bigoplus z_2 = (z_{11} + z_{21}, \ldots, z_{1n} + z_{2n}) \pmod 2$$

then $z_{1j} + z_{2j} = 0 \pmod 2$ if $z_{1j} = z_{2j}$, or 1 otherwise. Thus, $z_1 \bigoplus z_2$ is also an element in $\mathcal{Z}^n$. It is easy to see that $\forall \eta \in \mathcal{Z}^n$, $\eta \bigoplus \mathbf{0} = \eta$ and $\eta \bigoplus \eta = \mathbf{0}$, where $\mathbf{0} = (0, 0, \ldots, 0)$.

For convenience, we also can define the Hamming distance [Roman, 1992: p 105] between any two elements $\xi, \eta \in \mathcal{Z}^n$ as

$$H(\xi, \eta) = \sum_{i=1}^{n} |\xi_i - \eta_i|$$

which is the number of positions in which the two elements differ.

For two loci t Morgans apart, because of the independence among the n gametogenesis processes,

$$\begin{aligned} P(v(0) = \xi, v(t) = \eta) &= P(v(0) = \xi) P(v(t) = \eta \mid v(0) = \xi) \\ &= \frac{1}{2^n} [\theta(t)]^{H(\xi,\eta)} [1 - \theta(t)]^{n - H(\xi,\eta)} \qquad (1) \end{aligned}$$

The gene HBD event(s) determine the gene-transmission pedigree, which in turn determines the joint gametogenesis process $v(t)$ and the HBD set $\mathcal{D}$.

Having defined all these terms, we can now proceed to deal with multilocus HBD probability. Without loss of generality, suppose there are $m + 1$ linked loci, $L_1, L_2, \ldots, L_{m+1}$, located, in that order, at $t = 0$, $t = l_1, \ldots, t = l_m$, with $0 < l_1 < \cdots < l_m$. Denote the recombination fractions between loci L_i and L_j $(j > i)$ as θ_{ij} $(1 \leq i, j \leq m + 1)$. Given the HBD set $\mathcal{D}$, the multilocus probability

$$\begin{aligned} & P(v(0) \in \mathcal{D}, v(l_1) \in \mathcal{D}, \ldots, v(l_m) \in \mathcal{D}) \\ &= \sum_{v_1 \in \mathcal{D}} \cdots \sum_{v_{m+1} \in \mathcal{D}} P(v(0) = v_1, \ldots, v(l_m) = v_{m+1}) \\ &= \frac{1}{2^n} \sum_{v_1 \in \mathcal{D}} \cdots \sum_{v_{m+1} \in \mathcal{D}} P(v(l_1) = v_2 \mid v(0) = v_1) \cdots \\ &\qquad\qquad P(v(l_m) = v_{m+1} \mid v(0) = v_1, \ldots, v(l_{m-1}) = v_m) \\ &= \frac{1}{2^n} \sum_{v_1 \in \mathcal{D}} \cdots \sum_{v_{m+1} \in \mathcal{D}} P(v(l_1) = v_2 \mid v(0) = v_1) \cdots \\ &\qquad\qquad P(v(l_m) = v_{m+1} \mid v(l_{m-1}) = v_m) \\ &= \frac{1}{2^n} \sum_{v_1 \in \mathcal{D}} \cdots \sum_{v_{m+1} \in \mathcal{D}} \prod_{i=1}^{m} \theta_{i,i+1}^{H(v_i, v_{i+1})} (1 - \theta_{i,i+1})^{n - H(v_i, v_{i+1})} \qquad (2) \end{aligned}$$

where $v(0), \ldots, v(l_m)$ are the states of the gametogenesis process $v(t)$ at $t = 0, l_1, \ldots, l_m$, and $\theta_{i,i+1} = \frac{1}{2}(1 - e^{-2(l_i - l_{i-1})})$ is the recombination fraction between loci L_i and L_{i+1}.

Let I_i (or N_i) be the event that the individual of interest is (or is not) HBD at the ith locus, $i = 1, 2, \ldots, m+1$. Then, for any event $E = X_1 \cap X_2 \cap \cdots \cap X_{m+1}$, where $X_i = I_i$ or N_i, $P(E)$ can be expressed as a function of $P(I_{j_1} \cap I_{j_2} \cap \cdots \cap I_{j_k})$, where $j_l = 1, \ldots, m+1$ and $1 \leq j_k \leq m+1$, using inclusion/exclusion arguments [Guo, 1996]. That is, the probability that he is HBD at *all* j_k loci. For example,

$$P(N_1 \cap I_2 \cap I_3 \cap I_4) = P(I_2 \cap I_3 \cap I_4) - P(I_1 \cap I_2 \cap I_3 \cap I_4).$$

The above method works fairly well when the inbred pedigree is small and simple [Guo, 1996]. However, if the inbred pedigree contains many meiotic events of interest, the set $\mathcal{D}$ will be large. In this case, the above method will break down, even with the help of symbolic software such as MAPLE [Char et al., 1992].

It is often the case that the HBD set $\mathcal{D}$ can be decomposed into k ($k \geq 0$) disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, so that $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_k$. For the inbred pedigree in Figure 1, e.g., $\mathcal{D}$ consists of 16 subsets (see discussions in the numerical example). Each subset $\mathcal{D}_i$ ($i = 1, \ldots, k$) can be further decomposed into n 1-dimensional subcomponents $\mathcal{D}_{ij}$, $j = 1, \ldots, n$, where $\mathcal{D}_{ij} = \{0\}$, or $\mathcal{D}_{ij} = \{1\}$, or $\mathcal{D}_{ij} = \{0, 1\}$, so that

$$\mathcal{D}_i = \mathcal{D}_{i1} \times \cdots \times \mathcal{D}_{in}$$

where × means that $\mathcal{D}_i$ can be decomposed into several disjoint and mutually exclusive components of lower dimensions. For example, if a subset is $\mathcal{D}_1 = (1, *, *, 0, *, 1)$, say, where $*$ means either 0 or 1, it can be further decomposed into six 1-dimensional subcomponents, i.e., $\mathcal{D}_1 = \{1\} \times \{*\} \times \{*\} \times \{0\} \times \{*\} \times \{1\}$; i.e., jointly and in that order, the six subcomponents $\{1\}$, $\{*\}$, $\{*\}$, $\{0\}$, $\{*\}$, and $\{1\}$ constitute $\mathcal{D}_1$.

The decomposition of the HBD set $\mathcal{D}$ provides an opportunity for a much simpler method for computing the prior autozygosity probability. Before presenting the method, we first present some properties of the Hamming distance measure.

Properties of the Hamming Distance Measure

There are several useful properties for the Hamming distance. They are so obvious that they will be stated without proof.

Property 1: *decomposability.* For any two elements $\eta, \xi \in \mathcal{Z}^n$, η and ξ can be decomposed into two lower-dimensional components, i.e., $\eta = (\eta^{(1)}, \eta^{(2)})$ and $\xi = (\xi^{(1)}, \xi^{(2)})$, where $\eta^{(1)}$ and $\xi^{(1)}$ are elements in an m-dimensional hypercube ($0 \leq m \leq n$), and $\eta^{(2)}$ and $\xi^{(2)}$ are elements in an $(n - m)$-dimensional hypercube. The m-dimensional and $(n - m)$-dimensional hypercubes are disjoint, and, together, constitute $\mathcal{Z}^n$. Then,

$$H(\eta, \xi) = H(\eta^{(1)}, \xi^{(1)}) + H(\eta^{(2)}, \xi^{(2)}).$$

This property can be easily extended to the case where ξ and η are decomposed into k subcomponents ($1 \leq k \leq n$). In particular,

$$H(\eta, \xi) = \sum_{i=1}^{n} H(\eta_i, \xi_i) = \sum_{i=1}^{n} |\eta_i - \xi_i|.$$

Property 2: *permutation invariance.* For any two elements $\xi, \eta \in \mathcal{Z}^n$, the Hamming distance between ξ and η will remain the same if they are transformed under the same permutation.

Property 3: *transformation invariance.* For any three elements $\eta, \xi, \zeta \in \mathcal{Z}^n$,

$$H(\eta, \xi) = H(\eta \oplus \zeta, \xi \oplus \zeta).$$

It is easy to see that there are 2^n elements in $\mathcal{Z}^n$. We call each element a vertex. For each vertex η, there are $\binom{n}{k}$ vertices that have a Hamming distance of k from η, $0 \leq k \leq n$.

Simplifications

Suppose now that the HBD set $\mathcal{D}$ can be decomposed into k disjoint subsets $\mathcal{D}_1, \ldots, \mathcal{D}_k$, so that $\mathcal{D} = \mathcal{D}_1 \cup \mathcal{D}_2 \cup \cdots \cup \mathcal{D}_k$. Suppose also that each subset $\mathcal{D}_i$ ($i = 1, \ldots, k$) is decomposed into n subcomponents $\mathcal{D}_{ij}$, $j = 1, \ldots, n$, where $\mathcal{D}_{ij} = \{0\}$, or $\mathcal{D}_{ij} = \{1\}$, or $\mathcal{D}_{ij} = \{*\}$, where $* = 0, 1$, so that

$$\mathcal{D}_i = \mathcal{D}_{i1} \times \cdots \times \mathcal{D}_{in}$$

where $\mathcal{D}_{i1} \times \cdots \times \mathcal{D}_{in}$ means that, jointly, $\mathcal{D}_{i1}, \ldots, \mathcal{D}_{in}$ constitute $\mathcal{D}_i$.

For two linked loci, Eq. (2) can be rewritten as

$$\begin{aligned}
&\frac{1}{2^n} \sum_{v_1 \in \mathcal{D}} \sum_{v_2 \in \mathcal{D}} \theta_{12}^{H(v_1,v_2)} (1-\theta_{12})^{n-H(v_1,v_2)} \\
&\quad = \frac{1}{2^n} \sum_{(v_{11},\ldots,v_{1n}) \in \mathcal{D}} \; \sum_{(v_{21},\ldots,v_{2n}) \in \mathcal{D}} \theta_{12}^{\sum_{i=1}^{n} H(v_{1i},v_{2i})} (1-\theta_{12})^{\sum_{i=1}^{n}[1-H(v_{1i},v_{2i})]} \\
&\quad = \frac{1}{2^n} \sum_{a,b=1}^{k} \; \sum_{(v_{11},\ldots,v_{1n}) \in \mathcal{D}_a} \; \sum_{(v_{21},\ldots,v_{2n}) \in \mathcal{D}_b} \prod_{i=1}^{n} \theta_{12}^{H(v_{1i},v_{2i})} (1-\theta_{12})^{1-H(v_{1i},v_{2i})} \qquad (3) \\
&\quad = \frac{1}{2^n} \sum_{a,b=1}^{k} \sum_{v_{11} \in \mathcal{D}_{a1}} \cdots \sum_{v_{1n} \in \mathcal{D}_{an}} \sum_{v_{21} \in \mathcal{D}_{b1}} \cdots \sum_{v_{2n} \in \mathcal{D}_{bn}} \prod_{i=1}^{n} \theta_{12}^{v_{1i} \oplus v_{2i}} (1-\theta_{12})^{1-v_{1i} \oplus v_{2i}}. \qquad (4)
\end{aligned}$$

Two points can be made from the above formula. First, we can focus only on subsets in $\mathcal{D}$, instead of elements in $\mathcal{D}$. This would substantially reduce the number of cases to be considered. Second, we can now compute, for a given pair of subsets, the factor $\theta^h(1-\theta)^g$ *component by component*, then take the product of them, and then sum over all possible pairwise *subsets*. The computation can be further reduced by symmetry arguments; e.g., the summation $\sum_{a,b=1}^{k}\sum_{v_1\in\mathcal{D}_a}\sum_{v_2\in\mathcal{D}_b}$ can be reduced to computing $k\sum_{v_1\in\mathcal{D}_1}\sum_{v_2\in\mathcal{D}_1}$ and $2\sum_{a<b}\sum_{v_1\in\mathcal{D}_a}\sum_{v_2\in\mathcal{D}_b}$.

For the two-locus case, there are only four distinctive cases that need to be considered in computing the probability. These are: $(*,*)$, $(i_0,*)$, $(*,j_0)$, and (i_0,j_0), where i_0 and j_0 take either 0 or 1. For the (i_0,j_0) case, the factor is simply $\theta_{12}^{i_0\bigoplus j_0}(1-\theta_{12})^{1-i_0\bigoplus j_0}$, where $i\bigoplus j$ means $i+j$ (mod 2). For the $(*,j_0)$ case, the factor is

$$\theta_{12}^{0\bigoplus j_0}(1-\theta_{12})^{0\bigoplus j_0}+\theta_{12}^{1\bigoplus j_0}(1-\theta_{12})^{1\bigoplus j_0}=1$$

as j_0 is either 0 or 1.

By symmetry, the $(i_0,*)$ case also gives the value 1.

The $(*,*)$ case is also simple. By enumerating all possible values, it has the value 2.

The above formula can be easily extended to more than two loci. For three loci, e.g., we have

$$\begin{aligned}
&P(v(0)\in\mathcal{D},v(l_1)\in\mathcal{D},v(l_2)\in\mathcal{D})\\
&\quad=\frac{1}{2^n}\sum_{v_1\in\mathcal{D}}\sum_{v_2\in\mathcal{D}}\sum_{v_3\in\mathcal{D}}\prod_{1\le r<s\le 3}\theta_{rs}^{H(v_r,v_s)}(1-\theta_{rs})^{n-H(v_r,v_s)}\\
&\quad=\frac{1}{2^n}\sum_{(v_{11},\ldots,v_{1n})\in\mathcal{D}}\sum_{(v_{21},\ldots,v_{2n})\in\mathcal{D}}\sum_{(v_{31},\ldots,v_{3n})\in\mathcal{D}}\prod_{1\le r<s\le 3}\theta_{rs}^{\sum_{i=1}^{n}H(v_{ri},v_{si})}(1-\theta_{rs})^{\sum_{i=1}^{n}[1-H(v_{ri},v_{si})]}\\
&\quad=\frac{1}{2^n}\sum_{a,b,c=1}^{k}\sum_{(v_{11},\ldots,v_{1n})\in\mathcal{D}_a}\sum_{(v_{21},\ldots,v_{2n})\in\mathcal{D}_b}\sum_{(v_{31},\ldots,v_{3n})\in\mathcal{D}_c}\\
&\qquad\prod_{i=1}^{n}\prod_{1\le r<s\le 3}\theta_{rs}^{H(v_{ri},v_{si})}(1-\theta_{rs})^{1-H(v_{ri},v_{si})}\\
&\quad=\frac{1}{2^n}\sum_{a,b,c=1}^{k}\sum_{v_{11}\in\mathcal{D}_{a1}}\cdots\sum_{v_{1n}\in\mathcal{D}_{an}}\sum_{v_{21}\in\mathcal{D}_{b1}}\cdots\sum_{v_{2n}\in\mathcal{D}_{bn}}\sum_{v_{31}\in\mathcal{D}_{c1}}\cdots\sum_{v_{3n}\in\mathcal{D}_{cn}}\\
&\qquad\prod_{i=1}^{n}\prod_{1\le r<s\le 3}\theta_{rs}^{v_{ri}\bigoplus v_{si}}(1-\theta_{rs})^{1-v_{ri}\bigoplus v_{si}}.
\end{aligned}$$

For each component, there are 8 distinct cases, as follows:

		Cases							
		1	2	3	4	5	6	7	8
Subsets	1	i_0	i_0	i_0	$*$	i_0	$*$	$*$	$*$
	2	j_0	j_0	$*$	j_0	$*$	j_0	$*$	$*$
	3	k_0	$*$	k_0	k_0	$*$	$*$	k_0	$*$.

For case 1, it is easy to see that we have a factor of

$$\theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}\,\theta_{23}^{j_0 \oplus k_0}(1 - \theta_{23})^{1 - j_0 \oplus k_0}.$$

For case 2, we have

$$\begin{aligned} &\theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}\,\theta_{23}^{j_0 \oplus 0}(1 - \theta_{23})^{1 - j_0 \oplus 0} \\ &+\ \theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}\,\theta_{23}^{j_0 \oplus 1}(1 - \theta_{23})^{1 - j_0 \oplus 1} \\ &\qquad = \theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}. \end{aligned}$$

For case 3, we have

$$\begin{aligned} &\theta_{12}^{i_0 \oplus 0}(1 - \theta_{12})^{1 - i_0 \oplus 0}\,\theta_{23}^{0 \oplus k_0}(1 - \theta_{23})^{1 - 0 \oplus k_0} \\ &+\ \theta_{12}^{i_0 \oplus 1}(1 - \theta_{12})^{1 - i_0 \oplus 1}\,\theta_{23}^{1 \oplus k_0}(1 - \theta_{23})^{1 \oplus k_0} \\ &\qquad = \theta_{13}^{i_0 \oplus k_0}(1 - \theta_{13})^{1 - i_0 \oplus k_0}. \end{aligned}$$

That is, the contributing factor is completely determined by $i_0 \oplus k_0$ and the recombination fraction between loci 1 and 3.

The corresponding factor for other cases can also be computed. The results are listed in Table I.

For $m + 1$ loci, depending on the number of $*$'s and their locations, there will be 2^{m+1} distinct cases. For a given case, the contributing factor is determined solely by positions of $*$'s, if any, and values of the gametogenesis processes at the different loci. More specifically, if there are exactly m $*$'s, then the contributing factor is 1. This can be easily seen by induction with respect to m. As a corollary, the contributing factor is 2 if there are exactly $m + 1$ $*$'s. If there are less than m $*$'s, the contributing factor is determined by the status of the gametogenesis processes taking non-$*$-values. If the status of the gametogenesis process at a particular locus L_i, say, is f_i ($f_i \neq *$),

TABLE I. Contributing Factors for the Three-Locus Case[†]

	Locus			
Cases	1	2	3	Contribution
1	i_0	j_0	k_0	$\theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}\,\theta_{23}^{j_0 \oplus k_0}(1 - \theta_{23})^{1 - j_0 \oplus k_0}$
2	i_0	j_0	$*$	$\theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}$
3	i_0	$*$	k_0	$\theta_{13}^{i_0 \oplus k_0}(1 - \theta_{13})^{1 - i_0 \oplus k_0}$
4	$*$	j_0	k_0	$\theta_{12}^{i_0 \oplus j_0}(1 - \theta_{12})^{1 - i_0 \oplus j_0}$
5	i_0	$*$	$*$	1
6	$*$	j_0	$*$	1
7	$*$	$*$	k_0	1
8	$*$	$*$	$*$	2

[†] "$*$" denotes either 0 or 1.

and if the next non-$*$-locus is L_j ($i < j$), which takes value f_j, then there is at least a term $\theta_{ij}^{f_i \oplus f_j}(1-\theta_{ij})^{1-f_i \oplus f_j}$, regardless how many $*$'s between loci i and j. Thus, we can ignore $*$'s between any two adjacent non-$*$-valued loci. This can be viewed as if there were no information on recombination at those $*$-valued loci flanked by two non-$*$-valued loci. For example, for a 10-locus component $(*, 1, *, *, 0, *, 0, 1, *, 1)$, the contributing term is $\theta_{25}(1-\theta_{57})\theta_{78}(1-\theta_{8,10})$. Table II presents the results for $m = 3$.

Note that for any two loci L_i and L_j ($i < j$), the recombination fraction θ_{ij} can be expressed in terms of pairwise recombination fractions by Trow's formula

$$1 - 2\theta_{ij} = \prod_{l=i}^{j-1}(1 - 2\theta_{l,l+1})$$

where $\theta_{l,l+1}$ is the recombination fraction between the two adjacent loci l and $l + 1$.

NUMERICAL EXAMPLE

We demonstrate the proposed method using a small yet fairly complex pedigree (Fig. 1a) originally considered by Thompson [1994]. The pedigree was collected in a project on mapping Werner's syndrome [Nakura et al., 1994]. First ascertained as a first cousin marriage, it was later discovered that each parent of the affected proband

TABLE II. Contributing Factors for the Four-Locus Case[†]

	Locus				
Cases	1	2	3	4	Contribution
1	i	j	k	l	$\theta_{12}^{i\oplus j}(1-\theta_{12})^{1-i\oplus j}\theta_{23}^{j\oplus k}(1-\theta_{23})^{1-j\oplus k}\theta_{34}^{k\oplus l}(1-\theta_{34})^{1-k\oplus l}$
2	i	j	k	$*$	$\theta_{12}^{i\oplus j}(1-\theta_{12})^{1-i\oplus j}\theta_{23}^{j\oplus k}(1-\theta_{23})^{1-j\oplus k}$
3	i	j	$*$	k	$\theta_{12}^{i\oplus j}(1-\theta_{12})^{1-i\oplus j}\theta_{24}^{j\oplus l}(1-\theta_{24})^{1-j\oplus l}$
4	i	$*$	k	l	$\theta_{13}^{i\oplus k}(1-\theta_{13})^{1-i\oplus k}\theta_{34}^{k\oplus l}(1-\theta_{34})^{1-k\oplus l}$
5	$*$	j	k	l	$\theta_{23}^{j\oplus k}(1-\theta_{23})^{1-j\oplus k}\theta_{34}^{k\oplus l}(1-\theta_{34})^{1-k\oplus l}$
6	i	j	$*$	$*$	$\theta_{12}^{i\oplus j}(1-\theta_{12})^{1-i\oplus j}$
7	i	$*$	k	$*$	$\theta_{13}^{i\oplus k}(1-\theta_{13})^{1-i\oplus k}$
8	i	$*$	$*$	l	$\theta_{14}^{i\oplus l}(1-\theta_{14})^{1-i\oplus l}$
9	$*$	j	k	$*$	$\theta_{23}^{j\oplus k}(1-\theta_{23})^{1-j\oplus k}$
10	$*$	j	$*$	l	$\theta_{24}^{j\oplus l}(1-\theta_{24})^{1-j\oplus l}$
11	$*$	$*$	k	l	$\theta_{34}^{k\oplus l}(1-\theta_{34})^{1-k\oplus l}$
12	i	$*$	$*$	$*$	1
13	$*$	j	$*$	$*$	1
14	$*$	$*$	k	$*$	1
15	$*$	$*$	$*$	l	1
16	$*$	$*$	$*$	$*$	2

[†] "$*$" denotes either 0 or 1.

(the individual in the bottom) in the pedigree was also the offspring of a first cousin marriage.

Figure 1b is the corresponding gene-transmission pedigree, where the numbers represent $n = 18$ gametogenesis processes of interest. It is easy to see that individual M can have two genes HBD through various paths. In one path a copy of individual A's gene (either maternal or paternal) is transmitted to C, then to G, then to K, and to M; the same (maternal or paternal) gene is also transmitted to D, I, L, and finally, M. We denote this path as ACGKM—ADILM. Thus, as long as $g_1(t) = g_2(t) = g_{10}(t) = g_{11}(t) = g_{16}(t) = g_{17}(t) = g_{18}(t) = 0$ and $g_{13}(t) = 1$, individual M will have two copies of maternal genes from individual A. Other paths are ACGKM—AFJLM, ADHKM—AFJLM, BCGKM—BDILM, BCGKM—BFJLM, BDHKM—BFJLM, EHKM—EILM, and DHKM—DILM. These eight different paths can be represented by 16 distinctive subsets in $\mathcal{D}$, as listed in Table III. Note that each of the 16 subsets represents an HBD event with two copies of either the maternal or paternal allele from individuals A, B, D, or E. Each subset contains 2^{10} elements (subsets 1–12) or 2^{12} elements (subsets 13–16). Thus, there are $12 \cdot 2^{10} + 4 \cdot 2^{12} = 28{,}672$ elements in $\mathcal{D}$. Obviously, using Eq. (2) to compute the multilocus probability of autozygosity will involve a great deal of computations. Also note that at any locus, the HBD probability is $\frac{|\mathcal{D}|}{2^{18}} = 0.10935$, where $|\mathcal{D}| = 28{,}672$, the cardinality of $\mathcal{D}$.

We first demonstrate how to compute a two-locus HBD probability for this pedigree. As shown in Table III, there are 16 subsets in $\mathcal{D}$. By Eq. (3), if the gametogenesis process is in states defined in subsets 1 and 2 at the first and second loci (i.e., $a = 1$ and $b = 2$), respectively, we have

$$\begin{array}{lcccccccccccccccccc} \text{locus 1:} & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 0 & 0 & \ast & 1 & \ast & \ast & 0 & 0 & 0 \\ \text{locus 2:} & 1 & 1 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 0 & 0 & \ast & 1 & \ast & \ast & 0 & 0 & 0 \end{array}$$

TABLE III. Elements in $\mathcal{D}$ for a Complex Inbred Pedigree[†]

	Values of 18 gametogenesis processes																	
Subset	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	0	0	*	*	*	*	*	*	*	0	0	*	1	*	*	0	0	0
2	1	1	*	*	*	*	*	*	*	0	0	*	1	*	*	0	0	0
3	0	*	0	*	*	*	*	*	*	*	0	0	1	0	*	*	0	1
4	1	*	1	*	*	*	*	*	*	*	0	0	1	0	*	*	0	1
5	*	0	0	*	*	*	*	*	0	*	*	0	*	0	0	*	1	1
6	*	1	1	*	*	*	*	*	0	*	*	0	*	0	0	*	1	1
7	*	*	*	0	0	*	*	*	*	1	1	*	1	*	*	0	0	0
8	*	*	*	1	1	*	*	*	*	1	1	*	1	*	*	0	0	0
9	*	*	*	0	*	0	*	*	*	*	1	1	1	0	*	*	0	1
10	*	*	*	1	*	1	*	*	*	*	1	1	1	0	*	*	0	1
11	*	*	*	*	0	0	*	*	1	*	*	1	*	0	0	*	1	1
12	*	*	*	*	1	1	*	*	1	*	*	1	*	0	0	*	1	1
13	*	*	*	*	*	*	0	0	*	*	*	*	*	*	1	1	1	0
14	*	*	*	*	*	*	1	1	*	*	*	*	*	*	1	1	1	0
15	*	*	*	*	*	*	*	*	0	0	*	*	*	*	0	0	1	0
16	*	*	*	*	*	*	*	*	1	1	*	*	*	*	0	0	1	0

[†] "*" denotes either 0 or 1.

where each column represents values of the corresponding gametogenesis processes at loci 1 and 2.

By Eq. (4), we can compute the contributing term for the combination ($a = 1$ and $b = 2$) by computing contributing factors column by column (i.e., component by component) and take the product. By properties 2 and 3 of the Hamming distance measure, however, we know that the above subset combination ($a = 1$ and $b = 2$) is equivalent to

$$\begin{array}{lcccccccccccccccccc} \text{locus 1:} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast \\ \text{locus 2:} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast. \end{array}$$

That is, we have 8 $(0,0)$'s, and 10 $(\ast,\ast)$. This gives a term of $2^{10}(1-\theta_{12})^8$ in Eq. (3).

For $a = 1$ and $b = 16$, say, we have

$$\begin{array}{lcccccccccccccccccc} \text{locus 1:} & 0 & 0 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 0 & 0 & \ast & 1 & \ast & \ast & 0 & 0 & 0 \\ \text{locus 2:} & \ast & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 1 & 1 & \ast & \ast & \ast & \ast & 0 & 0 & 1 & 0. \end{array}$$

This amounts to 2 $(0,0)$'s, 2 $(0,1)$'s, 2 $(1,\ast)$'s, 4 $(0,\ast)$'s, and 8 $(\ast,\ast)$'s, which gives a term of $2^8\theta_{12}^2(1-\theta_{12})^2$. Other combinations can be dealt with in a similar fashion. If we sum over all pairwise combinations, we have

$$\begin{aligned} P(I_1 \cap I_2) = \frac{1}{2^9}\big[&4\theta(1-\theta)^2 + 5\theta^2 + 12\theta^2(1-\theta)^4 + 8\theta(1-\theta) + 16\theta^2(1-\theta)^2 \\ &+ 24(1-\theta)^8 + 32(1-\theta)^6 + 24\theta^2(1-\theta)^6 + 32\theta^6(1-\theta)^4\big] \end{aligned}$$

where I_j (or N_j) denotes the event that individual M is (or not) HBD at locus j, following Thompson's [1994] notation.

The three-locus HBD probability can be computed in a similar fashion. For example, if the gametogenesis process is in states defined by subsets 2, 4, and 9 at first, second, and third loci, respectively, say, we have

$$\begin{array}{lcccccccccccccccccc} \text{locus 1:} & 1 & 1 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 0 & 0 & \ast & 1 & \ast & \ast & 0 & 0 & 0 \\ \text{locus 2:} & 1 & \ast & 1 & \ast & \ast & \ast & \ast & \ast & \ast & \ast & 0 & 0 & 1 & 0 & \ast & \ast & 0 & 1 \\ \text{locus 3:} & \ast & \ast & \ast & 0 & \ast & 0 & \ast & \ast & \ast & \ast & 1 & 1 & 1 & 0 & \ast & \ast & 0 & 1 \end{array}$$

which gives five $(\ast,\ast,\ast)$'s, five $(0,\ast,\ast)$'s, two $(0,0,0)$'s, and one each of $(\ast,0,\ast)$, $(0,0,\ast)$, $(0,1,1)$, $(0,0,1)$, $(\ast,0,1)$, and $(\ast,0,0)$. This yields a term of $2^5\theta_{12}\theta_{23}^2(1-\theta_{12})^4(1-\theta_{23})^4$. Another way to compute this subset trio is to note that there are one $(0,1)$ and four $(0,0)$'s between loci 1 and 2, which gives $\theta_{12}(1-\theta_{12})^4$. In addition, there are two $(0,1)$'s and four $(0,0)$'s between loci 2 and 3, which gives $\theta_{23}(1-\theta_{23})^4$. Moreover, there are five $(\ast,\ast,\ast)$'s, which gives 2^5. Taken together, we have a term of $2^5\theta_{12}\theta_{23}^2(1-\theta_{12})^4(1-\theta_{23})^4$. Other combinations can be similarly calculated.

In general, any subset trio for the three-locus problem will yield $2^{a_{00}}\prod_{1\le i<j\le 3}\theta_{ij}^{a_{ij}}(1-\theta_{ij})^{b_{ij}}$, where $0 \le a_{00}, a_{ij} \le n$, for $1 \le i < j \le 3$. a_{00} is the total number of $(\ast,\ast,\ast)$'s in the combination, a_{01} the number of components (columns) that have different values at the first and second loci (such as $(0,1,\ast)$ and $(1,0,1)$), b_{01} the number of components that have the identical values at the first and second loci (such as $(0, 0, *)$ and $(1, 1, 0)$), a_{02} the number of $(i, *, 1 - i)$'s, where $i = 0$ or 1, and b_{02} the number of $(i, *, i)$'s, and so on. This can be easily extended to $(m + 1)$-locus problem, in which case any $(m + 1)$ subsets would yield $2^{a_{00}} \prod_{1 \leq i < j \leq m+1} \theta_{ij}^{a_{ij}} (1 - \theta_{ij})^{b_{ij}}$, where $0 \leq a_{00} \leq 1$, $0 \leq a_{ij} \leq n$, for $1 \leq i < j \leq m + 1$.

For $\theta_{12} = \theta_{23} = 0.1$, we calculated the *exact* three-locus autozygosity probability and compared it with the results obtained by Thompson [1994] using a Monte Carlo method. The results are listed in Table IV. It can be seen from the table that the Monte Carlo is in general fairly accurate, and has the feature of being as accurate as one wishes at the cost of more computation. However, there are noticeable differences, especially for those HBD configurations that are less likely. In addition, the Monte Carlo method, although versatile, usually takes much more time to run (about 8 hr on a DEC3100 workstation). Moreover, for a different set of parameters, one has to run the Monte Carlo again.

For $\theta_{12} = \theta_{23} = \theta_{34} = 0.1$, we also calculated the exact four-locus autozygosity probability. The results are listed in Table V.

DISCUSSION

This paper proposes a method for computing the multilocus prior probability of autozygosity for complex inbred pedigrees. The crux of the method is to identify loops in the inbred pedigrees, to decompose the HBD set into a small number of disjoint subsets, and then to compute the probability component by component. Because of innumerable configurations in human pedigrees, it is difficult, if not impossible, to devise a general algorithm to identify loops automatically for a complex inbred pedigree. However, for most human pedigrees encountered in practice, the decomposition can be accomplished easily once one understands the principle of the path-counting method [Wright, 1923]. When identification of loops is difficult, as for an extraordinarily complicated pedigree, this method may not work at all. However, it should be pointed out that most human pedigrees encountered in practice are often moderately complex. Thus, the identification of loops in the pedigree is no different from the path-counting method for computing the inbreeding coefficient, i.e., to identify common ancestors and paths (loops) that can lead to gene HBD. Once the loops can

TABLE IV. Three-Locus Prior Autozygosity Probabilities for a Complex Inbred Pedigree

Autozygosity status			Exact	Monte Carlo[a]
N_1	N_2	N_3	.78835130	.7901
N_1	N_2	I_3	.04908464	.0478
N_1	I_2	N_3	.02653601	.0257
N_1	I_2	I_3	.02665301	.0271
I_1	N_2	N_3	.04908464	.0478
I_1	N_2	I_3	.00410438	.0050
I_1	I_2	N_3	.02665302	.0271
I_1	I_2	I_3	.02953297	.0295

[a] From Thompson [1994].

TABLE V. Four-Locus Prior Autozygosity Probabilities for a Complex Inbred Pedigree

Autozygosity status				Probability
N_1	N_2	N_3	N_4	.74200570
N_1	N_2	N_3	I_4	.04634560
N_1	N_2	I_3	N_4	.02484550
N_1	N_2	I_3	I_4	.02423914
N_1	I_2	N_3	N_4	.02484550
N_1	I_2	N_3	I_4	.00169051
N_1	I_2	I_3	N_4	.01282093
N_1	I_2	I_3	I_4	.01383208
I_1	N_2	N_3	N_4	.04634560
I_1	N_2	N_3	I_4	.00273904
I_1	N_2	I_3	N_4	.00169051
I_1	N_2	I_3	I_4	.00241387
I_1	I_2	N_3	N_4	.02423914
I_1	I_2	N_3	I_4	.00241387
I_1	I_2	I_3	N_4	.01383208
I_1	I_2	I_3	I_4	.01570090

be identified, the prior multilocus HBD probability can be computed by the proposed method.

The proposed method, coupled with existing computer power and symbolic software such as MAPLE [Char et al., 1992], can provide explicit solutions to many multilocus gene HBD problems for complex inbred pedigrees. It should be noted that the method can also be used to compute the multilocus prior probability of gene shared *identical by descent* for a group of individuals in complex pedigrees [Guo, 1996].

Because it computes the probability component by component, the method automatically collects various terms and expresses them in a succinct way, thus improving the accuracy.

It should be emphasized that this paper only presents a method for computing the *prior* or *unconditional* HBD probability, i.e., the HBD probability without marker information. The computation of the *posterior* or *conditional* HBD probability given observed marker information would require completely different methods. It should be noted that, although marker data further limit the overall number of legal states, they destroy the simple representation structure inherent in these states, forcing one to consider these states individually. Thus, it may be difficult to apply the proposed method to computing the posterior HBD probability.

Throughout this paper, no interference and no sex difference in map length have been assumed. Neither assumption is of course correct for human genomes. However, the assumption of no interference makes the computation of HBD probability much easier. When there is a weak positive interference while no interference is assumed, the calculated HBD probability will be a slight underestimate of the true one. This is because double crossovers are less likely in actuality, making the sharing of genes more likely. The assumption of no sex difference in map length also is critical for the random walk theory to hold. It is extremely difficult to construct a model allowing for the sex difference while still remaining tractable. Nonetheless, the method presented in this paper is useful for a sex-averaged map.

ACKNOWLEDGMENTS

This research was supported by National Institutes of Health grant R29-GM52205. The author thanks Dr. Kenneth Lange for pointing out Trow's formula and Dr. Ellen Wijsman for discussions on homozygosity mapping and helpful comments on an earlier version of this paper.

REFERENCES

Char BW, Geddes KO, Gonnet GH, Leong BL, Monagan MB, Watt SM (1992): "Maple V. Library Reference Manual." 2nd Ed. New York: Springer-Verlag.

Donnelly KP (1983): The probability that related individuals share some section of the genome identical by descent. Theor Pop Biol 23:34–64.

Fisher RA (1949): "The Theory of Inbreeding." New York: Academic Press.

Guo SW (1995): Proportion of genome shared identical-by-descent by relatives: Concept, computation, and applications. Am J Hum Genet 56:1468–1476.

Guo SW (1996): Gametogenesis processes and multilocus gene identity by descent. Am J Hum Genet 58:408–419.

Haldane JBS (1919): The combination of linkage values and the calculation of distance between the loci of linked factors. J Genet 8:299–309.

Kruglyak L, Daly MJ, Lander ES (1995): Rapid multipoint linkage analysis of recessive traits in nuclear families, including homozygosity mapping. Am J Hum Genet 56:519–527.

Lander ES, Botstein D (1987): Homozygosity mapping: A way to map human recessive traits with the DNA of inbred children. Science 236:1567–1570.

Nakura J, Wijsman EM, Miki T, Kamino K, Yu C-E, Oshima J, Fukuchi K-I, et al. (1994): Homozygosity mapping of the Werner syndrome locus (WRN). Genomics 23:600–608.

Roman S (1992): "Coding and Information Theory." New York: Springer-Verlag.

Smith CAB (1953): Detection of linkage in human genetics. J R Stat Soc B 15:153–192.

Thompson EA (1994): Monte Carlo estimation of multilocus autozygosity probabilities. Proceedings of the 1994 Interface Conference. SAS: Cary, NC.

Wright S (1923): Mendelian analysis of the pure breeds of livestock. I. The measurement of inbreeding and relationship. J Hered 14:339–348.