Inference on pedigree structure from genome screen data

Running title: Inference on pedigree structure

Mary Sara McPeek

The University of Chicago

Address for correspondence:

Department of Statistics, 5734 S. University Ave., Chicago, IL 60637

phone: (773) 702-7554, fax: (773) 702-9810, email: mcpeek@galton.uchicago.edu

Abstract. The problem of error detection in general inbred and outbred pedigrees on the basis of genome screen data is considered. We develop a novel characterization of pairwise relationships, which is extended to k-wise relationships. Given an arbitrary pedigree specifying the relationship among a set of k individuals, we show how to prune the pedigree so that no information on the genetic relationships is lost and yet no excess meioses remain. We take a likelihood-based approach to inference. Under the assumption of no interference, all the crossover processes in a pedigree can be viewed jointly as a continuous time Markov random walk on the vertices of a hypercube, so a hidden Markov method is a natural approach for likelihood calculation. One strategy to make likelihood calculation feasible is to use aspects of the pedigree structure to find the orbits of the group of symmetries on the hypercube that preserve the information of identity by descent. We describe strategies for accomplishing this for arbitrary pedigrees and give weak sufficient conditions under which the resulting chain has the minimum number of states needed to both contain all the information of the IBD process and to satisfy the Markov property under no interference.

Key words and phrases: crossover process, HMM, IBD, Markov chain, misspecified relationship, pairwise relationship, pedigree error, pedigree graph, relationship estimation, relationship inference

1. Introduction. Genetic linkage analysis is used to locate genetic variants associated with traits of interest. The initial goal is to identify genetic markers whose alleles tend to be co-inherited with the trait within families. This analysis depends on accurate knowledge of the relationships among individuals in the study. If the relationship among individuals is misspecified, this may lead to either reduced power (*e. g.* when the the true relationship among individuals with similar trait values is more distant than what is believed) or false positive evidence for linkage (*e. g.* when the true relationship among individuals with similar trait values is closer than what is believed). The importance of identification of relationship errors in a linkage study is demonstrated by Boehnke and Cox (1997) in an application to non-insulin-dependent diabetes mellitus.

It is common in linkage studies for data to be collected on hundreds (or thousands) of loci throughout the genome, in what is called a genome screen. Relationships among individuals in the study are ascertained by other methods and can be summarized by a pedigree. We consider the problem of using the genome screen data collected for linkage analysis to detect errors in the assumed pedigree. For outbred pairwise relationships, Thompson (1975) considers the special case of unlinked loci, Browning (1998; 2000) assumes that continuous identity by descent information is available, and Zhao and Liang (2001) further assume gamete data. Practical methods for detection of errors in sibling pair relationships from genotype data on linked loci include Göring and Ott (1997), Boehnke and Cox (1997), Ehm and Wagner

(1998), and Olson (1999). Methods for a wider range of common outbred pairwise relationships are given by Thompson and Meagher (1998), McPeek and Sun (2000), Epstein, Duren, and Boehnke (2000), and Sun, Wilder, and McPeek (submitted). To identify errors in pairwise relationships in a complex inbred pedigree, Sun, Abney, and McPeek (2001) use a simple graphical method.

We take a likelihood-based approach to inference and use the MLLR test of McPeek and Sun (2000), extended to k-wise relationships. We give a novel characterization of pairwise relationships, which we extend to k-wise relationships. This characterization allows one to determine which individuals in a pedigree have an impact on the genetic relationship among any given set of individuals, and it is particularly relevant for complex inbred pedigrees. The question of how to automatically generate minimal-state hidden Markov chains to implement the MLLR test for any given pairwise relationship was left as an open problem by McPeek and Sun (2000). In the current work, we describe how to find the hidden Markov model with the minimum number of states for a given k-wise relationship, among those Markov chains that are aggregations of the joint crossover process. This involves finding the orbits of the group of symmetries on a hypercube that preserve certain sets of vertices. Furthermore, we give weak sufficient conditions under which the resulting Markov chain has the minimum number of states needed to both contain all the information of the IBD process and to satisfy the Markov property under no interference. We discuss the practical problems of inference based on genetic data.

2. Likelihood-Based Inference. Let X denote genotype data on a set of k individuals, and suppose that, for each k-wise relationship R, we have a fully specified model for X and can calculate the likelihood $L_R(X)$. In a linkage study, one would typically have a pedigree obtained, for instance, by asking the individuals in the study how they are related. Suppose that the pedigree specifies some relationship R_0 for the k individuals. In performing linkage analysis, one would typically assume that the relationship R_0 is correct unless there are strong indications to the contrary. Thus, one natural approach to pedigree error detection is hypothesis testing with H_0 : true relationship is R_0 vs. H_A : true relationship is not R_0. We choose some subset $\mathcal{R}$ of k-wise relationships and consider the statistic $MLLR = \max_{\{A \in \mathcal{R} \setminus \{R_0\}\}} \log(L_A) - \log(L_{R_0})$ of McPeek and Sun (2000). We obtain an empirical estimate $\hat{F}_0$ of the null distribution F_0 of $MLLR$ by simulation under R_0, using the same map of markers as in the data set. We calculate the p-value associated with $\{MLLR = m\}$ as $2 \min\{\hat{F}_0(m), 1 - \hat{F}_0(m)\}$. If the p-value associated with R_0 is sufficiently small, we reject the null hypothesis. As a point estimate, we could set $\hat{R} = B \in \mathcal{R}$ for some B satisfying $\log(L_B) = \max_{\{A \in \mathcal{R}\}} \log(L_A)$. More useful is a confidence set, which we could define to consist of all relationships in $\mathcal{R}$ for which the p-value is greater than α, for some chosen $\alpha > 0$, in addition to all relationships not included in $\mathcal{R}$. We discuss, in Sections 3 and 4, the space of possible R and, in Sections 5 and 6, the model for X assuming R. Likelihood calculation is discussed in Section 7.

3. Human Pedigrees. The defining characteristics of a pedigree depend on the mating system. For instance, a pedigree for organisms capable of asexual reproduction would follow different rules from one for humans. For humans (or other organisms that reproduce similarly), we define a pedigree to consist of a directed graph P and a function s, where P has nodes $\mathcal{N}(P) \subset \mathcal{Z}$, $|\mathcal{N}(P)| < \infty$, corresponding to individuals in the pedigree, and directed edges $\mathcal{E}(P) \subset \mathcal{N}(P) \times \mathcal{N}(P)$, with $(a, b) \in \mathcal{E}(P)$ precisely when a is a parent of b. Here $s : \mathcal{N}(P) \to \{\text{male, female}\}$ assigns a sex to each individual. (For other ways to represent a pedigree, see Cannings and Thompson (1981), Chap. 1 and Thompson (1986), Chap. 2.) Given $b \in \mathcal{N}(P)$, let $p(b) = \{a \in \mathcal{N}(P) : (a, b) \in \mathcal{E}(P)\}$ be the set of **parents** of b. For $a, b \in \mathcal{N}(P)$ and $k \geq 2$, we define $(a, a_1, \ldots, a_{k-1}, b) \in \mathcal{N}(P)^{k+1}$ to be a **directed path of length** k from a to b if $\{(a, a_1), (a_1, a_2), \ldots, (a_{k-2}, a_{k-1}), (a_{k-1}, b)\} \subset \mathcal{E}(P)$. We define (a, b) to be a **directed path of length 1** from a to b if $a \in p(b)$. We define $\mathcal{A}(b) \subset \mathcal{N}(P)$ to be the set of **ancestors** of b, $\mathcal{A}(b) = \{a \in \mathcal{N}(P) :$ there is a directed path of length $l \geq 1$ from a to $b\}$. In order to be a human pedigree, (P, s) must satisfy the following conditions:

1. For all $b \in \mathcal{N}(P), |p(b)| = 0$, 1 or 2 (each individual has 0, 1, or 2 parents in the pedigree).

2. For all $b \in \mathcal{N}(P)$, if $a_1, a_2 \in p(b)$ with $a_1 \neq a_2$, then $s(a_1) \neq s(a_2)$ (if an individual has two parents in the pedigree, they must have opposite sexes).

3. For all $a \in \mathcal{N}(P), a \notin \mathcal{A}(a)$ (an individual cannot be his or her own ancestor).

Let $\mathcal{P}$ be the set of all **pedigrees**, *i. e.* the set of all (P, s) satisfying the above conditions. If $|p(b)| = 0$ we call $b \in \mathcal{N}(P)$ a **founder** of the pedigree, if $|p(b)| = 1$ we call b a **half founder**, and if $|p(b)| = 2$ we call b a **nonfounder**. (We note that it is conventional to further restrict the definition of a pedigree by disallowing half founders; however, we find this restriction disadvantageous for our purposes.) Let $\mathcal{F}(P) \subset \mathcal{N}(P)$ be the set of founders of P, $\mathcal{H}_m(P) \subset \mathcal{N}(P)$ be the set of half founders with a mother in the pedigree, i.e. with $s(a) =$ female where $p(b) = \{a\}$, let $\mathcal{H}_f(P) \subset \mathcal{N}(P)$ be the set of half founders with a father in the pedigree, and let $\mathcal{NF}(P) \subset \mathcal{N}(P)$ be the set of nonfounders of P. Note that the number of directed edges in the graph is always $2|\mathcal{NF}(P)| + |\mathcal{H}_m(P)| + |\mathcal{H}_f(P)|$. Finally, for the purposes of this study, if a pair of individuals in the pedigree are monozygotic twins, we identify their nodes and treat them as if they were a single individual. The reason is that they are genetically identical (or virtually so).

We define two distinct individuals $a, b \in \mathcal{N}(P)$ to be **unrelated** (with respect to pedigree (P, s)) if they have no common ancestors and neither is an ancestor or descendant of the other, *i. e.* $[\{a\} \cup \mathcal{A}(a)] \cap [\{b\} \cup \mathcal{A}(b)] = \emptyset$. We define a nonfounder $a \in \mathcal{NF}(P)$ to be **outbred** (with respect to pedigree (P, s)) if a's 2 parents are unrelated. In addition, all founders and half founders are considered to be outbred. An individual who is not outbred will be said to be **inbred**. We define a pedigree P to be outbred if a is outbred for all $a \in \mathcal{N}(P)$. A pedigree that is not outbred is said to be inbred.

4. Characterization of Pairwise Relationships, with Extension to k-wise Relationships. The relationships encountered in linkage analysis can range from the very simple, such as sibling and parent-offspring relationships, to the extraordinarily complex. Examples of the latter can be found in the Hutterite data set described in Abney, McPeek, and Ober (2000). This data set involves 806 genotyped individuals related by a 1623-member, 13-generation pedigree with virtually every genotyped individual inbred and most individuals related to one another through multiple lines of descent. Motivated by the richness of relationships in such data and by the need for efficient computational methods to cope with the corresponding pedigrees, we develop below a characterization of k-wise relationships.

Suppose that within a pedigree, we wish to consider the relationship among k chosen individuals. Consider the set $\mathcal{P}_k \subset \mathcal{P} \times \mathcal{Z}^k$ such that every $(P, s, i_1, \ldots, i_k) \in \mathcal{P}_k$ satisfies $\{i_1, \ldots, i_k\} \subset \mathcal{N}(P)$ and $|\{i_1, \ldots, i_k\}| = k$. We first focus on $\mathcal{P}_2$ and define pairwise relationships to be equivalence classes of a particular equivalence relation on $\mathcal{P}_2$. Given $\gamma = (P, s, i, j) \in \mathcal{P}_2$, we define an individual $a \in \mathcal{N}(P) \setminus \{i, j\}$ to be **superfluous** with respect to γ if at least one of the following two conditions holds:

1. $a \notin \mathcal{A}(i) \cup \mathcal{A}(j)$ (a is not an ancestor of i or j).

2. $\mathcal{A}(a) \cap \{i, j\} = \emptyset$ (neither i nor j is an ancestor of a), and there exist $c \in \mathcal{N}(P) \setminus \{i, j\}$ and $d \in \mathcal{N}(P)$ such that for every $e \in \{a\} \cup \mathcal{A}(a)$, for every $l \geq 1$, and for every directed path $q = (q_0, ..., q_l)$ of length l with $q_0 = e$ and $q_l \in \{i, j\}$, we have $c = q_m$ and $d = q_{m+1}$ for some $0 \leq m \leq l - 1$ (every directed path from a or ancestors of a to i or j passes through directed edge (c, d)).

Theorems 1 and 2 in Section 5 justify the terminology "superfluous" in this case. Let $\mathcal{S}(\gamma) \subset \mathcal{N}(P)$ be the set of superfluous nodes with respect to γ.

Given $\gamma_1 = (P_1, s_1, i_1, j_1)$ and $\gamma_2 = (P_2, s_2, i_2, j_2) \in \mathcal{P}_2$, we define $\gamma_1^* = (P_1^*, s_1^*, i_1, j_1)$ and $\gamma_2^* = (P_2^*, s_2^*, i_2, j_2)$ to be the restrictions of γ_1 and γ_2, respectively, to their nonsuperfluous nodes. That is, we define the directed graph P_1^* to have nodes $\mathcal{N}(P_1^*) = \mathcal{N}(P_1) \setminus \mathcal{S}(\gamma_1)$ and directed edges $\mathcal{E}(P_1^*) = \{(a, b) \in \mathcal{N}(P_1^*) \times \mathcal{N}(P_1^*) : (a, b) \in \mathcal{E}(P_1)\}$. Define s_1^* on $\mathcal{N}(P_1^*)$ by $s_1^*(a) = s_1(a)$. We call (P_1^*, s_1^*) the **pruned pedigree** with respect to γ_1. Define (P_2^*, s_2^*) to be the pruned pedigree with respect to γ_2. We say that γ_1 and γ_2 specify the same sex-specific pairwise relationship, and write $\gamma_1 \equiv \gamma_2$ whenever there exists a bijection $g : \mathcal{N}(P_1^*) \to \mathcal{N}(P_2^*)$ such that the following three conditions hold:

1. $g(i_1) = i_2, g(j_1) = j_2$ (the two focal individuals are preserved).

2. $(a, b) \in \mathcal{E}(P_1^*) \Leftrightarrow (g(a), g(b)) \in \mathcal{E}(P_2^*)$ (directed edges are preserved, i.e. the directed graphs P_1^* and P_2^* are isomorphic).

3. For all $a \in \mathcal{N}(P_1^*), s_1^*(a) = s_2^*(g(a))$ (sexes are preserved).

As defined above, "≡" clearly satisfies the requirements of an equivalence relation. We define the set of **sex-specific pairwise relationships** to be the resulting set of equivalence classes. Examples include father-daughter, paternal aunt-niece, and maternal grandmother-grandson. It is usually convenient to further aggregate relationships by removing Condition 3. This has the effect of, for instance, combining the 8 possible avuncular relationships (maternal uncle-niece, paternal aunt-nephew, etc.) into a single class. When Condition 3 is removed, we call the resulting set of equivalence classes the set of **pairwise relationships**. As an example, the pedigree graphs in Figures 1a and b specify the same pairwise relationship for individuals i and j, while that in Figure 1c is different. We say that a pairwise relationship or sex-specific pairwise relationship R is **outbred** if for $\gamma \in R$, the pruned pedigree with respect to γ is outbred. This is clearly a class property. A pairwise relationship that is not outbred is said to be **inbred**.

Note that our definitions of superfluous, pruned pedigree, sex-specific pairwise relationship, pairwise relationship, and inbred and outbred pairwise relationships extend in a straightforward way to k-wise relationships. If $(P, s, i_1, \ldots, i_k) \in \mathcal{P}_k$, then, for instance, in the definition of superfluous, we require $a \in \mathcal{N}(P) \setminus \{i_1, \ldots, i_k\}$, we change Condition 1 to specify $a \notin \cup_{j=1}^{k} \mathcal{A}(i_j)$, and we change Condition 2 to specify $\mathcal{A}(a) \cap \{i_1, \ldots, i_k\} = \emptyset, c \in \mathcal{N}(P) \setminus \{i_1, \ldots, i_k\}$ and $q_l \in \{i_1, \ldots, i_k\}$.

In Theorem 1 of Section 5, we show that if γ_1 and $\gamma_2 \in \mathcal{P}_k$ specify the same k-wise relationship, then they yield the same expanded IBD process on the autosomal chromosomes, where the expanded IBD process is defined below in Section 5. Furthermore, suppose γ^* is the pruned pedigree corresponding to $\gamma \in \mathcal{P}_k$. We show in Theorem 2 of Section 5 that if any directed edge is removed from γ^*, then the resulting IBD process is different. These results justify the definition of superfluous and the characterization of k-wise relationships given above. The additional information on sex given by a sex-specific k-wise relationship is used to determine the IBD process on sex chromosomes (the pairwise case is discussed in Epstein, Duren, and Boehnke (2000)).

Given $(P, s, i, j) \in \mathcal{P}_2$, if i and j are both outbred, we may further aggregate relationships by setting $(P, s, i, j) \equiv (P, s, j, i)$. The IBD process will be invariant to the interchange of i and j in that case.

5. Crossover Process, Mendelian Inheritance at a Single Locus, and Identity States. To each directed edge of a pedigree graph is associated a meiosis, and each meiosis results in an independent realization of the **crossover process**. The crossover process is a binary process $\{C_t\}$ that describes at each point t along the genome whether an offspring inherited from the given parent that parent's maternal ($C_t = 0$) or paternal ($C_t = 1$) DNA. Switches from 0 to 1 or 1 to 0 are called **crossovers**. It is usually assumed that $\{C_t\}$ and $\{1 - C_t\}$ have the same distribution. The restrictions of the crossover process to different chromosomes are assumed to be independent within a meiosis. Crossover processes for different meioses are also independent and will be assumed to be identically distributed. There are special restrictions on the crossover process on the parent's pair of sex chromosomes. In humans, there are two types of sex chromosomes, X and Y. Individuals possessing two X chromosomes are female, and individuals possessing one X and one Y chromosome are male. In females, crossovers between the two X chromosomes are permitted. In males, there is a region that is homologous between X and Y, called the pseudoautosomal region, on which crossovers are permitted, but crossovers are not permitted outside that region.

For a given pedigree (P, s), we can consider the joint crossover process consisting of a component crossover process for each directed edge of the pedigree. For the remainder of this section, we consider a single locus on an autosomal (*i. e.* non-sex) chromosome. Then the joint crossover process results in a random function $V : \mathcal{E}(P) \to \{0,1\}$ where $V(a,b)$ is equal to the value of the crossover process associated with directed edge (a,b), at the given chromosomal location (Donnelly (1983)). Assuming Mendelian inheritance, the distribution of V puts mass $2^{-|\mathcal{E}(P)|}$ on each point of $\{0,1\}^{\mathcal{E}(P)}$. Define the **allele function** to be a random function $\alpha : \mathcal{N}(P) \times \{0,1\} \to \mathcal{Z}$, where $\alpha(a,0)$ gives a's maternal allele and $\alpha(a,1)$ gives a's paternal allele. Define $\mathcal{FA}(P) = [\mathcal{F}(P) \times \{0,1\}] \cup [\mathcal{H}_f(P) \times \{0\}] \cup [\mathcal{H}_m(P) \times \{1\}]$. We refer to the restriction of α to $\mathcal{FA}(P)$ as the assignment of **founder alleles**, and we let $\alpha(\mathcal{FA}(P))$ denote the set of founder alleles. Given V and the assignment of founder alleles, which we will assume to be independent, the function α is completely determined and can be calculated by recursion.

The **identity state** (Gillois (1964); Harris (1964)) for $\gamma = (P,s,i,j) \in \mathcal{P}_2$ at a given locus can be defined as follows: suppose that each element of $\mathcal{FA}(P)$ is assigned a unique founder allele. Then for a given V, the value of $A = (\alpha(i,0), \alpha(i,1), \alpha(j,0), \alpha(j,1))$ is determined from V and the assignment of founder alleles. Each value of A can be mapped to one of the 15 identity states depicted in Figure 2, where each node represents one of $(i,0), (i,1), (j,0)$ and $(j,1)$, and an edge is drawn between nodes a and b whenever $\alpha(a) = \alpha(b)$ (Jacquard (1974), Chap. 6). Since the founder alleles are assumed to be distinct, it is apparent that the identity state depends only on V and not on the assignment of founder alleles. The identity states can be viewed as equivalence classes on the range of V. Then the distribution of V induces a distribution on the identity state for $\gamma \in \mathcal{P}_2$ at the given locus. In Section 6, when we discuss genotype data, we will see that it is usually desirable to combine some of the 15 identity states to yield the 9 **condensed identity states** (Harris (1964); Jacquard (1974), Chap. 6) depicted in Figure 2. These result from identifying elements $v_1, v_2 \in \{0,1\}^{\mathcal{E}(P)}$ when their identity states are the same up to a permutation of $\alpha(i,0)$ and $\alpha(i,1)$ and a permutation of $\alpha(j,0)$ and $\alpha(j,1)$. Let Δ be the distribution on the condensed identity states induced by the distribution of V, as shown in Figure 2. Then Δ can be used to define quantities of interest such as the kinship coefficient for i and j, $\Phi(i,j) = \Delta_1 + (\Delta_3 + \Delta_5 + \Delta_7)/2 + \Delta_8/4$ and the inbreeding coefficients for i and j, $H(i) = \Delta_1 + \Delta_2 + \Delta_3 + \Delta_4$ and $H(j) = \Delta_1 + \Delta_2 + \Delta_5 + \Delta_6$.

The concepts of identity state and condensed identity state extend in a natural way from $\mathcal{P}_2$ to $\mathcal{P}_k$. The details can be found in Thompson (1974). Given $\gamma \in \mathcal{P}_k$, we define the **identity-by-descent (IBD) process** $\{I_t\}$ by $I_t =$ the condensed identity state for γ at location t (in the autosomal portion of the genome). We define the **expanded IBD process** $\{E_t\}$ by $E_t =$ the identity state for γ at location t (in the autosomal portion of the genome).

Theorem 1: If γ_1 and γ_2 specify the same k-wise relationship, then their expanded IBD processes have the same distribution.

It immediately follows that their IBD processes also have the same distribution.

Theorem 2: Suppose $\gamma = (P, s, i_1, \ldots, i_k) \in \mathcal{P}_k$ has no superfluous nodes. Given $A \subset \mathcal{E}(P)$, $A \neq \emptyset$, define P' by $\mathcal{N}(P') = \mathcal{N}(P)$, $\mathcal{E}(P') = \mathcal{E}(P) \setminus A$, and set $\gamma' = (P', s, i_1, \ldots, i_k)$. Then the IBD processes of γ and γ' have different distributions.

It follows that their expanded IBD processes also have different distributions.

Remarks: Theorem 2 provides only a partial converse to Theorem 1. It is possible to have two distinct relationships that yield the same IBD process, *e.g.* half-first-cousin and grand-half-avuncular pairs, but note that neither relationship is obtainable from the other by removal of edges. Theorems 1 and 2 do not depend on the particular choice of model for the crossover process, as long as $\mathrm{Var}(C_t) > 0$ for some t. In particular, the theorems hold in the presence of interference and when $\{C_t\}$ and $\{1 - C_t\}$ have different distributions.

Theorems 1 and 2 extend in a straightforward way to the X chromosome. Given $\gamma_j = (P_j, s_j, i_{j,1}, \ldots, i_{j,k}) \in \mathcal{P}_k$, $j = 1$ or 2, we first eliminate from P_j all directed edges (a, b) for which $s_j(a) = s_j(b) =$ male. Call the resulting pedigree $\tilde{P}_j$. Let $\tilde{\gamma}_j = (\tilde{P}_j, s_j, i_{j,1}, \ldots, i_{j,k})$ and let $(\tilde{P}_j^*, \tilde{s}_j^*)$ be the pruned pedigree for $\tilde{\gamma}_j$. We have: (1) if $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$ specify the same sex-specific k-wise relationship, then the expanded IBD processes for γ_1 and γ_2 on the nonpseudoautosomal X have the same distribution; (2) given γ_1 and γ_2, if $(\tilde{P}_2^*, \tilde{s}_2^*)$ is obtained from $(\tilde{P}_1^*, \tilde{s}_1^*)$ by removal of directed edges, then the IBD processes for γ_1 and γ_2 on the nonpseudoautosomal X have different distributions.

6. Models for Genotype Data. For each genotyped individual, data are typically collected on hundreds (or thousands) of markers throughout the genome. Thus, information on $\{I_t\}$ is obtained only at discrete sites t. Furthermore, different individuals will typically have missing data at different markers for various reasons, such as failure of the experiment to determine the genotype at a marker. Recall that determination of the identity state depends on observation of $(\alpha(i_1, 0), \alpha(i_1, 1)), \ldots, (\alpha(i_k, 0), \alpha(i_k, 1)) \in \alpha(\mathcal{FA}(P))^{2k}$. However, based on a single individual's observed genotype data, when $z_1 \neq z_2$, the two possibilities $\{\alpha(a, 0) = z_1, \alpha(a, 1) = z_2\}$ and $\{\alpha(a, 0) = z_2, \alpha(a, 1) = z_1\}$ cannot be distinguished. Therefore, we define an equivalence relation on $\alpha(\mathcal{FA}(P))^2$ by $(a, b) \equiv (b, a)$ for all $\{a, b\} \subset \alpha(\mathcal{FA}(P))$. Then an individual's observed genotype at the given marker will be one of the equivalence classes under this relation. The difficulty in distinguishing maternally and paternally inherited alleles, together with the assumption that the crossover processes $\{C_t\}$ and $\{1 - C_t\}$ have the same distribution, leads naturally to consideration of the condensed identity states instead of the identity states (see Figure 2 for the pairwise case and Thompson (1974) for the k-wise case). When genotypes are observed for a single individual at two loci, say equivalence class $\{(a_1, a_2), (a_2, a_1)\}$ at marker a and equivalence class $\{(b_1, b_2), (b_2, b_1)\}$ at marker b, the genotype data for the individual do not determine whether a_1 and b_1 were inherited from the same or different parents. This missing information is called **phase**. Note that if one has genotype data on other close relatives of the individual, one may have full or partial information to determine phase and distinguish paternal and maternal inheritance. However, one would need to assume that these relationships were correct in order to use this information. Thus, this approach is less useful for relationship inference than for linkage analysis.

A further complication in real data is that founder alleles are generally not unique. Thus, for example, in the case of a pairwise relationship, the observation $\{\alpha(i,0) = \alpha(i,1) = \alpha(j,0) = \alpha(j,1)\}$ is compatible with all the identity states. In addition there is some rate of genotyping error, so that, in principle, any observation is compatible with any identity state. The rate of genotyping error is generally assumed to be low. However in the pairwise case, for example, genotyping errors cause problems for any relationship for which $\Delta_9 = 0$, for example parent-offspring or monozygotic twin relationships. A genotyping error may cause the observation that the four alleles of individuals i and j are all distinct, resulting in likelihood 0 under any relationship for which $\Delta_9 = 0$, unless genotyping errors are included in the model (see e.g. Broman and Weber (1998)).

In order to calculate the likelihood for the data, we need to specify models for the crossover process $\{C_t\}$, for the assignment of founder alleles, and for genotyping errors. The most widely used model for $\{C_t\}$ is a Poisson process, and failure of $\{C_t\}$ to follow this model is known as **interference**. Note that in linkage studies, distance t along the chromosome is scaled so that the expected number of transitions of the process $\{C_t\}$ in an interval of width s is equal to s. Thus, we need not specify the intensity of the Poisson process or even whether or not it is homogeneous, as long as we assume that the intensity function is bounded (see McPeek and Speed (1995) for details). While the Poisson process model is useful in a wide range of applications, it has long been known to provide a poor fit to data. Although alternative models exist, their use with human data can be computationally quite challenging because of the types of missing information described above. In what follows, we use the Poisson process model and discuss the extension to the Poisson-skip class of models (Lange, Zhao, Speed (1997); Lange (1997), Section 12.5). McPeek and Sun (2000) have performed simulations to investigate the robustness to interference of pairwise relationship inference based on the Poisson process model.

The model for assignment of founder alleles is determined by population genetic assumptions such as Hardy-Weinberg equilibrium and linkage equilibrium, and it requires allele frequency distributions for every marker. In practice, these assumptions may not hold, and accurate estimates of allele frequencies may not be available. The population genetic modeling assumptions certainly have an effect on the analysis, and model misspecification can be problematic. With closer relatives and more informative markers, the impact of such assumptions is diminished.

As long as the rate of genotyping errors is low, they should not have much impact on the analysis except in special cases, such as pairwise relationships with $\Delta_9 = 0$. To deal with this case, Broman and Weber (1998) assume that errors are i.i.d. across loci and meioses. Their model is quite serviceable for relationship inference, even though error rates are known to vary across loci.

X chromosome data can be important in linkage studies. On the non-pseudoautosomal part of the X chromosome, the crossover process $\{C_t\}$ depends on the sexes of the parent and child. For a mother-child meiosis, $\{C_t\}$ behaves as on autosomes. For a father-daughter meiosis $\{C_t\}$ is identically 0 on the non-pseudoautosomal X chromosome, while for a father-son meiosis, no X chromosome is transmitted. When data are available on the X chromosome in addition to the autosomes, relationships such as paternal aunt-niece and maternal aunt-niece are potentially distinguishable (Epstein, Duren, and Boehnke (2000)). Methods for detecting relationship errors in linkage data have generally ignored the Y chromosome. Differences among individuals on the non-pseudoautosomal Y would have arisen exclusively by mutation. Thus, linkage disequilibrium is not expected to decay with distance, and assumptions about the populations from which founder males were drawn are critical for the likelihood calculation. Therefore, the mathematical problem of relationship inference based on non-pseudoautosomal Y data has salient features that distinguish it from the problems of relationship inference considered here.

7. Likelihood Calculation. In light of the probability models and data issues described above, the question arises as to how to calculate the likelihood of the data. Göring and Ott (1997) have calculated the likelihood under the assumption of no interference for the special cases of sibling, half-sibling and unrelated pairs, using the fact that the IBD process $\{I_t\}$ is Markov for these cases. Boehnke and Cox (1997) performed the same calculations more efficiently by using the hidden Markov method (Baum (1972)), with the hidden Markov chain given by restriction of the process $\{I_t\}$ to the marker positions, on which the data provide only partial information. Broman and Weber (1998) extended this method to calculate the likelihood for parent-offspring and monozygotic twins, which have $I_t = 1$ and $I_t = 2$ for all t, respectively, by inclusion of a model for genotyping error in the observation distribution of the hidden Markov model. However, outside of a few special cases in which the IBD process $\{I_t\}$ is either trivial (parent-offspring, unrelated, monozygotic twin) or Markov (sibling, half-sibling, grandparent-grandchild), $\{I_t\}$ will not in general be Markov, even under the assumption of no interference (Donnelly (1983); Feingold (1993)). For instance, $\{I_t\}$ is not Markov for the cases of avuncular and first-cousin relationships. On the other hand, the joint crossover process $\{V_t\}$ will always be Markov under the assumption of no interference (Donnelly (1983)). Thus, one possible approach to likelihood calculation is to apply the hidden Markov method with the hidden Markov chain taken to be the restriction of the process $\{V_t\}$ to the marker positions, and with the observed data viewed as providing partial information on the function $\{I_t\}$ of $\{V_t\}$ at the marker positions (Lander and Green (1987)).

This approach allows for data at a large number of loci, but the computational time is exponential in the number of directed edges in the graph P. The first step in reducing the computational time is to apply the characterization of k-wise relationships in Section 4 and Theorems 1 and 2 of Section 5 to determine which directed edges in the graph can be removed without changing the IBD process $\{I_t\}$. Further dramatic reduction in the state space of the Markov chain can be obtained by expanding on ideas discussed by Donnelly (1983). He observed that $\{V_t\}$ is a Markov random walk on a hypercube. Suppose there is a non-injective map h from the set H of vertices of the hypercube to some finite set (in our case, to the condensed identity states). Consider the group S of symmetries of the hypercube, *i. e.* permutations of H that preserve all the edges of the hypercube. Let $G \subset S$ be the subgroup of symmetries of the hypercube that also preserve values of h, and let O be the set of orbits of G. Define an equivalence relation on H by saying that for $v_1, v_2 \in H$ we have $v_1 \equiv v_2$ when v_1 and v_2 lie in the same element of O. Let $\{A_t\}$ be the process taking values in O, viewed as a function of $\{V_t\}$, defined so that if $V_t = v$, then $A_t =$ the unique element of O containing v. Then $\{A_t\}$ is an irreducible Markov chain with a state space no larger than that of $\{V_t\}$, and often substantially smaller.

In our case, $\{I_t\}$ can be viewed as a function of $\{A_t\}$. Thus, to calculate the likelihood for our observed data, we could apply the hidden Markov method with the hidden Markov chain taken to be the restriction of the process $\{A_t\}$ to the marker positions. To implement this approach, one needs to determine, for each k-wise relationship considered, the orbits O, the transition matrix $P(t)$ for $\{A_t\}$, and the stationary distribution for $\{A_t\}$. Let Q be the matrix of infinitesimal parameters of $\{A_t\}$. Given $v_1, v_2 \in \{0,1\}^{\mathcal{E}(P)}$, define $|v_1 - v_2| = \sum_{k \in \mathcal{E}(P)} |v_1(k) - v_2(k)|$. Given the set of orbits O, the Q matrix is obtained as follows: given k, choose $v \in O_k$. Then $q_{kl} = |\{v' \in O_l : |v - v'| = 1\}|$ for $k \neq l$, $q_{kk} = -\sum_{l \neq k} q_{kl}$. It is not hard to show that this does not depend on the choice of v. From the Q matrix, the transition matrix is obtained as $P(t) = e^{Qt}$. For the restriction of $\{A_t\}$ to the marker positions, we actually prefer to specify the transition matrix in terms of recombination fraction θ rather than t, where $\theta = (1 - e^{-2t})/2$ under the assumption of no interference. In practice, we find that specification of the transition matrix in terms of θ rather than t makes our analysis more robust to the presence of interference, because data are usually obtained on the value of θ between markers, with the value of t between markers estimated from θ using some (incorrectly specified) model for interference. We note that the one-step conditional distribution $P(\theta)$ itself does not depend on assumptions about interference, although the assumption of no interference is used to obtain the Markov property.

We now give sufficient conditions for the Markov process $\{A_t\}$ obtained by the above procedure to have the minimum number of states needed to both contain all the information of the IBD process $\{I_t\}$ and to satisfy the Markov property under no interference. Suppose $\{A_t\}$ is a continuous-time finite state space Markov process, with cardinality of the state space equal to n, and suppose that $\{I_t\}$ is defined by a deterministic function of $\{A_t\}$, $I_t = f(A_t)$, where f is defined on the state space of $\{A_t\}$. Let S be the state space of $\{I_t\}$, and for each $s \in S$, let $n_s = |f^{-1}(s)|$. Following Larget (1998), we define an observable sequence $(\mathbf{y}, \mathbf{t})$ to consist of a finite sequence $\mathbf{y}$ of elements of S and a finite nondecreasing sequence $\mathbf{t}$ of nonnegative real times, where $|\mathbf{y}| = |\mathbf{t}|$ and the first element of $\mathbf{t}$ is always 0. For each observable sequence $(\mathbf{y}, \mathbf{t})$ with $|\mathbf{y}| = |\mathbf{t}| = k$, define

$$Q^{(\mathbf{y},\mathbf{t})} = I^{y_1} e^{Q(t_2-t_1)} I^{y_2} e^{Q(t_3-t_2)} I^{y_3} \cdots I^{y_{k-1}} e^{Q(t_k-t_{k-1})} I^{y_k} \text{ for } k \geq 2$$

and $Q^{(\mathbf{y},\mathbf{t})} = I^{y_1}$ for $k = 1$, where Q is the matrix of infinitesimal parameters of $\{A_t\}$ and I^s is the $n \times n$ diagonal matrix which is the identity on the submatrix where $f(i) = f(j) = s$ and is zero elsewhere. Furthermore, suppose π is the initial distribution of $\{A_t\}$ (which is the stationary distribution in our case) and $\mathbf{1}$ is a vector of ones.

Theorem 3: Suppose $\{A_t\}$ is a continuous-time Markov process with finite state space of size n, and suppose that $\{I_t\}$ is defined by a deterministic function of $\{A_t\}$, $I_t = f(A_t)$, where f is defined on the state space of $\{A_t\}$. Then the following conditions are sufficient to ensure that for any other continuous-time finite state space Markov process $\{B_t\}$ such that $\{I_t\}$ is defined by a deterministic function of $\{B_t\}$, the cardinality of the state space of $\{B_t\}$ is no less than n:

(1) For each $s \in S$, there exist n_s observable sequences $\{(\mathbf{y}, \mathbf{t})_i\}_{i=1}^{n_s}$ such that the vectors

$\{\pi^T Q^{(\mathbf{y},\mathbf{t})_i} I^s\}_{i=1}^{n_s}$ are linearly independent.

(2) For each $s \in S$, there exist n_s observable sequences $\{(\mathbf{y},\mathbf{t})_i\}_{i=1}^{n_s}$ such that the vectors $\{I^s Q^{(\mathbf{y},\mathbf{t})_i} \mathbf{1}\}_{i=1}^{n_s}$ are linearly independent.

Remarks: Theorem 3 is a continuous-time analogue of a result by Gilbert (1959) for discrete-time Markov chains. Conditions (1) and (2) are easily checked once Q is constructed. Later in this section we give some examples to which Theorem 3 applies.

For the following kinds of outbred pairwise relationships, the orbits O, matrix Q, and the stationary distribution are given by Donnelly (1983): ancestor-descendant (i is a gth-generation ancestor of j for $g \geq 1$); half-sib type (i is a μth-generation descendant of a and j is a νth-generation descendant of b, for $\mu, \nu \geq 0$, where a and b are half siblings); cousin-type (i is a μth-generation descendant of a and j is a νth-generation descendant of b, for $\mu, \nu \geq 0$, where a and b are first cousins); uncle-type (j is a μth-generation descendant of a, for $\mu \geq 1$, where a and i are full siblings). The transition matrix $P(\theta)$ can be found in Bishop and Williamson (1990) for the half-sib and grandparent-grandchild relationships, for which $\{A_t\}$ and $\{I_t\}$ are the same, and in McPeek and Sun (2000) for the avuncular and first-cousin relationships, for which $\{A_t\}$ and $\{I_t\}$ are different. We now give $P(\theta)$ for the other pairwise relationship types considered by Donnelly (1983). For the outbred gth-generation ancestor-descendant relationship, in which i is a gth generation ancestor of j, with $g > 1$, the pruned pedigree (P, s) has $\mathcal{N}(P) = \{i, j, a_1, \ldots, a_{g-1}\}, |\mathcal{N}(P)| = g + 1$, and $\mathcal{E}(P) = \{(i, a_1), (a_1, a_2), \ldots, (a_{g-1}, j)\}$. Let v' assign to each $(a, b) \in \mathcal{E}(P)$ the indicator of whether a is male, and let v'_{-1} be the restriction of v' to $\mathcal{E}(P) \setminus \{(i, a_1)\}$. Then $O_k = \{v \in$

$\{0,1\}^{\mathcal{E}(P)} : |v_{-1} - v'_{-1}| = k\}$ and $P(\theta)$ has (k,l)th element

$$P_{kl}(\theta) = \sum_{m=max(0,k-l)}^{min(k,g-l-1)} C_{k,m} C_{g-k-1,l-k+m} \theta^{l-k+2m} (1-\theta)^{g-l+k-2m-1},$$

where $C_{a,b} = a!/[(a-b)!b!]$. Think of this transition matrix as a function of g and call it P_g. Note that when $g = 1$ (parent-offspring), the Markov chain $\{A_t\}$ is trivial, with only a single state, so we can set $P_1 = 1$. Then for the half-sib type relationships, the transition matrix is $P_{\mu+\nu+1} \otimes H$, where H is the matrix for half-sibs given in Bishop and Williamson (1990) and $\otimes$ is Kronecker product. For the cousin-type relationships, the transition matrix is $P_{\mu+\nu+1} \otimes C$ where C is the matrix for cousins given in McPeek and Sun (2000), and for the uncle-type relationships, the transition matrix is $P_\mu \otimes U$ where U is the matrix for the avuncular relationship given in McPeek and Sun (2000). Thus, computation of the likelihood can be accomplished for these types of relationships.

Note, however, that even if we restrict ourselves to outbred pairwise relationships for which the pruned pedigree has no directed paths of length greater than 2, there are 23 distinct non-trivial types of relationships, only 8 of which are covered by the above results. When inbred relationships are permitted, when the lengths of directed paths are allowed to be greater than 2, or when individuals are considered k-wise, the number of possibilities is enormous. Thus, it is desirable to have more general, automatic ways of determining the $\{A_t\}$ process. Note that for the symmetry group S, we have $|S| = 2^d \times d!$, where d is the number of directed edges in the pruned pedigree P. One could find O by considering each of these permutations in turn, deciding whether or not it belongs in G, and then finding the orbits of G. We describe below some shortcuts that make it unnecessary to consider every element of S.

There are certain symmetries in the pedigree structure that can be easily exploited to reduce the size of the state space of the Markov process. These symmetries allow one to find a subgroup of G, resulting in a set of orbits O' that is a finer partition of the state space than O. For instance, the two parental alleles within a founder individual a can be permuted without altering the condensed identity state of γ. In the joint crossover process $\{V_t\}$, this amounts to replacing $V_t(a, b)$ by $1 - V_t(a, b)$ for every child b of a at every locus t. By making use of this symmetry, the state space is reduced from 2^d elements to $|O'| = 2^{d-f}$, where $f = |\mathcal{F}(P)|$. In the context of linkage analysis with moderate-sized outbred pedigrees, this state-space reduction is used by Kruglyak, Daly, Reeve-Daly, Lander (1996). Define a, $b \in \mathcal{F}(P), a \neq b$ to form a founder couple if $\{c : a \in p(c)\} = \{d : b \in p(d)\}$. Another symmetry in the pedigree structure that can be used to reduce the state space is that the permutation of individuals within a founder couple does not alter the condensed identity state of γ, provided that neither individual in the couple is one of the focal individuals $\{i_1, \ldots, i_k\}$. In the crossover process $\{V_t\}$, this amounts to interchanging $V_t(a, c)$ and $V_t(b, c)$ and switching $V_t(c, d)$ to $1 - V_t(c, d)$ for all c with $p(c) = \{a, b\}$, all d with $c \in p(d)$, and all t. In the context of linkage analysis with moderate-sized outbred pedigrees, this type of approach was used by Gudbjartsson, Jonasson, Frigge, and Kong (2000). Let g be the number of founder couples where neither individual is in $\{i_1, \ldots, i_k\}$ and where they have at least one grandchild in the pedigree. For $k = 1, 2, \ldots$, let $c(k)$ be the number of founder couples where neither individual is in $\{i_1, \ldots, i_k\}$ and where they have k children and 0 grandchildren in the pedigree. Let n be the number of founders who are not part of a founder couple or who are in a founder couple in which one of the individuals is in $\{i_1, \ldots, i_k\}$. By making use of the orbits O' of the subgroup of G generated by permutation of individuals within founder couples and permutation of the two alleles within each founder, the state space would be reduced from 2^d elements to $|O'| = 2^{d-3g-n} \prod_k (2^{-3} + 2^{-k-2})^{c(k)}$. For example, for a pair of full sibs, $2^d = 16$ and $|O'| = |O| = 3$, and for an avuncular pair $2^d = 32$ and $|O'| = |O| = 4$. Thus, application of the above two types of symmetry leads to the minimum number of states in these cases. (That these are, indeed, the minimum number of states for these two relationships follows from Theorem 3.) For a pair of first cousins, $2^d = 64, |O'| = 8$, and $|O| = 7$, so application of the above two types of symmetry leads to 1 extra state beyond the minimum, where application of Theorem 3 confirms that 7 is the minimum. For $l > 1$, define the directed path $q = (q_0, \ldots, q_l)$ to be an **isolated branch** of length l if $q_h \in [\mathcal{H}_m(P) \cup \mathcal{H}_f(P)] \setminus \{i_1, \ldots, i_k\}$ and q_h has exactly one offspring for all $0 \leq h \leq l-1$. A further symmetry in the pedigree structure that, when present, can be used to reduce the state space is the permutation of meioses within isolated branches. Given a permutation π on $\{1, \ldots, l\}$, in the crossover process $\{V_t\}$, permutation of meioses within the isolated branch amounts to replacing $V_t(q_{h-1}, q_h)$ by $|1\{s(q_{\pi(h)-2}) \neq s(q_{h-2})\} - V_t(q_{\pi(h)-1}, q_{\pi(h)})|$ for all t and for $h = 1, \ldots, l$, where we define q_{-1} to be the mother of q_0 if $q_0 \in \mathcal{H}_m(P)$ and the father of q_0 if $q_0 \in \mathcal{H}_f(P)$. Let $n(b)$ be the number of isolated branches of length b. By making use of this symmetry, the state space would be reduced from 2^d elements to $|O'| = 2^d \times \prod_{b \geq 2} (\frac{b+1}{2^b})^{n(b)}$. For a pair of third cousins, $2^d = 1024$ and $|O| = 35$. If we make use of all three types of symmetry described above, we obtain $|O'| = 72$.

A brute force approach to finding O is to consider each element of S, determine whether or not it is in G, and then find O from G. When the symmetries described above are used to obtain the set of orbits O', this can be used to find the coarser partition O by consideration of fewer elements of S than would be required by the brute force approach. Recall that if two elements of $\{0,1\}^{\mathcal{E}(P)}$ are in the same orbit of O', then they lead to the same condensed identity state for γ. Thus, to each orbit in O' can be associated a condensed identity state. For each condensed identity state ϕ, let $h(\phi)$ be the number of orbits in O' with condensed identity state ϕ. Let ψ be a condensed identity state such that $h(\psi) \leq h(\phi)$ for all ϕ. Then the set of orbits O can be obtained from the set of orbits O' by consideration of no more than $h(\psi)d! - 1$ symmetries of H. For the case of an outbred pairwise relationship, this can be reduced to $h(\psi)d!2^{-f} - 1$, where f is the number of founders excluding i and j. See Appendix D for details. For example, for first cousins, $d = 6, f = 2$, and $h(\psi) = 2$. Thus, O could be obtained from O' by consideration of 359 elements of S instead of all 46,079 non-identity elements.

8. Discussion. In practical applications, it may be useful to focus attention initially on inference for pairwise relationships, based on genome screen data for the pair. This allows one to easily identify particular directed edges of the pedigree that are likely in error, and it gives useful information for determining plausible alternatives for the local pedigree structure. The pairwise approach can be supplemented by considering multiple individuals jointly, with a set of alternative relationships constructed on the basis of the pairwise results.

Likelihood-based inference on pedigree structure is closely related to likelihood-based inference for linkage mapping, but there are important differences. The goal in linkage mapping is to detect a local change in distribution of the IBD process, whereas in pedigree inference, we are interested in the distribution of the process throughout the entire genome. Use of other genotyped relatives to provide additional information on the IBD process for a set of individuals is important for linkage analysis, but is problematic for pedigree inference because it depends on the accuracy of these relationships.

The likelihood calculations described in Section 7 depend on the fact that $\{V_t\}$ is Markov, which holds under the assumption of no interference. These calculations could be extended to the Poisson-skip class of models for interference (Lange, Zhao, and Speed (1997); Lange (1997), Section 12.5), of which the χ^2 model (Zhao, Speed, and McPeek (1995)) is a special case. For this class of models, $\{V_t\}$ is not Markov, but it can be viewed as hidden Markov, as described by Lange ((1997), Section 12.5).

There are many practical considerations not treated here. One involves the choice of the set $\mathcal{R}$ of relationships over which the likelihood is maximized. Computational feasibility places constraints on the size of $\mathcal{R}$. Important considerations in choosing $\mathcal{R}$ include, first, power to distinguish among relationships based on the data. For instance, common ancestry that is too many generations away will have little impact on the likelihood, even if the IBD process is observed on the entire genome. Second, knowledge of individuals' ages, or of the fact that they were alive simultaneously, combined with knowledge of human generation times, suggests restrictions on the numbers of generations separating a pair of genotyped individuals. For modest-sized outbred pedigrees, Sun, Wilder, and McPeek (submitted) implement pairwise relationship analysis using $\mathcal{R} = \{$monozygotic twins, parent-offspring, full siblings, half siblings plus first cousins, half siblings, grandparent-grandchild, avuncular, first cousin, half-avuncular, half-first-cousin, unrelated$\}$. This set was chosen based on the pedigrees encountered in data and based on the alternative relationships suggested by the estimation of $(\Delta_7, \Delta_8, \Delta_9)$ by the method of McPeek and Sun (2000).

A somewhat simpler approach to pedigree inference was taken by Göring and Ott (1997), who assigned prior probabilities to pairwise relationships, with prior mass 1 on some small finite set $\mathcal{R}$ of relationships and then calculated posterior probabilities for the elements of

$\mathcal{R}$. In Göring and Ott (1997), the reported relationship was full sib, and $\mathcal{R}$ consisted of full sib, half-sib, and unrelated. Aside from the fact that their approach is Bayesian and ours is frequentist, one of the main differences between our approach and theirs is the performance when the true relationship does not lie in $\mathcal{R}$, which is always a possibility. For example, in the case considered by Göring and Ott (1997), some reasonable alternatives not in $\mathcal{R}$ (and which could certainly have an impact on the linkage results if true) are (a) that the sibs are inbred in one of various ways, (b) that they are half-sibs with, say, the same mother and fathers who are related, or (c) that they have some other outbred relationship such as an avuncular relationship. If the true relationship does not lie in $\mathcal{R}$, then by Göring and Ott's (1997) method there is, in principle, no possibility of recognizing this as long as prior mass 1 is assigned to $\mathcal{R}$ and not all likelihoods are 0 for the elements of $\mathcal{R}$. For instance, with the choice of $\mathcal{R}$ and prior distribution used by Göring and Ott (1997), a pedigree error in which a true inbred sib pair is falsely reported as an outbred sib pair would have essentially no chance of being detected. If prior mass less than 1 is assigned to $\mathcal{R}$, but likelihoods are calculated only for the elements of $\mathcal{R}$, then posterior probabilities for the elements of $\mathcal{R}$ are known up to a constant multiple, and one cannot generally construct a confidence set; in particular, one still cannot determine that none of the relationships in $\mathcal{R}$ is in the confidence set (unless all likelihoods are 0 for the elements of $\mathcal{R}$). The previous example of an inbred sib pair being virtually undetectable still applies when prior mass on $\mathcal{R}$ is less than 1. In contrast the Monte-Carlo-based method we use can, in principle, and also sometimes in practice, have power to reject all relationships in $\mathcal{R}$, even though likelihoods are calculated only for relationships in $\mathcal{R}$. In that case, the method could report that the confidence set does not contain any element of $\mathcal{R}$. The trade-off is that our method is more computationally expensive than the Bayesian approach, because simulations are required.

An additional difficulty with the Bayesian approach is that, in general, there is no straightforward choice of prior distribution. One would presumably want to incorporate prior information such as a higher probability for the reported pedigree and information on the frequencies of various types of errors. Rates of non-paternity and inbreeding can depend on the population and the phenotype under study and are not well-known. Problems can arise such as confusion of individuals with similar names and relatedness of individuals thought to be unrelated, which are difficult to quantify. Other sources of error include switched or duplicated samples.

An important set of questions beyond the scope of this paper involves how to perform linkage analysis in light of the pedigree errors detected. In practice, it may be possible to go back and collect additional data that confirm and explain some of the pedigree errors detected (*e. g.* see Epstein, Duren, and Boehnke (2000)). The uncertainty about other parts of the pedigree could, in principle, be incorporated into the analysis.

Sampling of pedigrees for a linkage study is often based on the presence of multiple relatives affected by a trait. Assuming that the trait has genetic determinants, these relatives may be expected to share regions of the genome containing these genetic determinants. This could have an impact on the distributions of their IBD processes, but this ascertainment effect is not expected to be noticeable in practice.

Acknowledgment. Support of NIH HG01645 and DK55889 is gratefully acknowledged.

Appendix A. Proof of Theorem 1. We employ Lemma 1 to show that the identity state for $\gamma = (P, s, i_1, \ldots, i_k)$ at a given locus depends on V only through V^*, where (P^*, s^*) is the pruned pedigree, and V^* is the restriction of V to $\mathcal{E}(P^*)$. We assume throughout that each element of $\mathcal{FA}(P)$ is assigned a unique founder allele.

Lemma 1. Suppose $\{(a_1, j_1), \ldots, (a_n, j_n)\} \in \mathcal{FA}(P^*), |\{(a_1, j_1), \ldots, (a_n, j_n)\}| > 1$, and $\alpha(a_1, j_1) = \alpha(a_2, j_2) = \ldots = \alpha(a_n, j_n)$. Then (1) $\{a_1, \ldots, a_n\} \subset [\mathcal{H}_m(P^*) \cup \mathcal{H}_f(P^*)] \cap \mathcal{NF}(P)$ and (2) there exist $c \in \mathcal{N}(P^*) \setminus \{i_1, \ldots, i_k\}, d \in \mathcal{N}(P^*)$ such that $(c, d) \in \mathcal{E}(P^*)$ and, for every directed path $q = (q_1, \ldots, q_l)$ in P^* with $q_1 \in \{a_1, \ldots, a_n\}, q_l \in \{i_1, \ldots, i_k\}$, we have $q_m = c$ and $q_{m+1} = d$ for some $1 \leq m \leq l - 1$.

To prove Lemma 1, we consider $(a, j) \in \mathcal{FA}(P^*)$ and consider each of the following five possibilities in turn: (i) $a \in \mathcal{F}(P^*) \cap \mathcal{F}(P)$, (ii) $a \in \mathcal{F}(P^*) \cap \mathcal{H}_\phi(P), \phi = m$ or f, (iii) $a \in \mathcal{F}(P^*) \cap \mathcal{NF}(P)$, (iv) $a \in \mathcal{H}_\phi(P^*) \cap \mathcal{H}_\phi(P), \phi = m$ or f, (v) $a \in \mathcal{H}_\phi(P^*) \cap \mathcal{NF}(P), \phi = m$ or f. In each of cases (i) through (iv), we find that if $(c, l) \in \mathcal{FA}(P^*)$ with $(c, l) \neq (a, j)$, then $\alpha(a, j) \neq \alpha(c, l)$. Part (1) of the Lemma follows. Let $D(P) = [\mathcal{N}(P) \times \{0, 1\}] \setminus \mathcal{FA}(P)$, and define the parent function $\beta : D(P) \to \mathcal{N}(P)$ by $\beta(c, 0) = a$, where a is the mother of c, and $\beta(c, 1) = b$, where b is the father of c. In case (v), if we let $\psi = 1\{\phi = f\}$, we find that $\beta(a, 1 - \psi)$ satisfies Condition 2 of superfluous with directed edge (c, d), where either (a) $(c, d) = (\beta(a, 1 - \psi), a)$ or (b) $a \notin \{i_1, \ldots, i_k\}$ and every directed path from a to $\{i_1, \ldots, i_k\}$ passes through (c, d). In case (a), we find the same results as for cases (i)-(iv). For case (b), part (2) of the Lemma follows.

Using Lemma 1, we show that if unique alleles are assigned to the members of $\mathcal{FA}(P)$ then, at a given location t, the identity state E depends on V only through V^*. Consider the function α^* defined on $D(P^*)$ and obtained as follows: assign unique alleles to the members of $\mathcal{FA}(P^*)$, then apply V^* to obtain α^*. Let E^* be the resulting identity state for $A^* = (\alpha^*(i_1, 0), \alpha^*(i_1, 1), \ldots, \alpha^*(i_k, 0), \alpha^*(i_k, 1))$. We show that $E = E^*$. To do this, we construct for each V, an assignment of unique alleles to the members of $\mathcal{FA}(P^*)$ so that application of V^* yields $A^* = A$.

Appendix B. Proof of Theorem 2. Define identity state δ to be "$\leq$" identity state ϵ if the set of edges in the defining graph of δ is a subset of the set of edges in the defining graph of ϵ, with "$=$" holding precisely when δ and ϵ are the same identity state. Given V on $\mathcal{E}(P)$, let V' be the restriction of V to $\mathcal{E}(P')$. Let δ be the identity state for γ based on V, and let ϵ be the identity state for γ' based on V'. It is apparent that $\epsilon \leq \delta$. In order to prove Theorem 2, we need to show that when γ has no superfluous nodes, there is some choice of V for which the inequality $\epsilon \leq \delta$ is strict. It is sufficient to show that given $(a, b) \in \mathcal{E}(P)$, either (i) there exist $e \in \{a\} \cup \mathcal{A}(a)$ and directed paths $q^1 = (q^1_1, ..., q^1_m)$ and $q^2 = (q^2_1, \ldots, q^2_n)$ in P, with $q^1_1 = q^2_1 = e$, $q^1_m, q^2_n \in \{i_1, \ldots, i_k\}$ such that q^1 passes through (a, b), and q^1 and q^2 have no common directed edges or (ii) $\mathcal{F}(P) \cap [\{a\} \cup \mathcal{A}(a)] \setminus \{i_1, \ldots, i_k\} = \emptyset$. To obtain this result, we apply Lemma 2.

Lemma 2. Given $A \subset \mathcal{N}(P)$, $B \subset \mathcal{N}(P)$, $k \geq 3$, and a k-tuple of directed paths from A to B, if there is no directed edge (c, d) through which they all pass, then there exists a disjoint pair of directed paths from A to B.

Proof of Lemma 2. Number the k directed paths 1 to k. Suppose there is no single directed edge through which all k pass. For any $(k-1)$-tuple of these directed paths, describe their intersection by the set of directed edges through which they all pass. Let $\mathcal{E}_{k-1}$ be the union over all $(k-1)$-tuples of these sets of directed edges. Note that $\mathcal{E}_{k-1}$ (which may be empty) is strictly ordered with $(a_1, b_1) < (a_2, b_2)$ if $b_1 \in \mathcal{A}(a_2)$. Write $\mathcal{E}_{k-1} = \{e_1, \ldots, e_l\}$ where $e_\phi < e_{\phi+1}$. Aggregate $(e_1, \ldots, e_l)$ into blocks $(b_1, \ldots, b_o)$ of consecutive directed edges $b_\phi = (e_\psi, \ldots, e_{\psi+\nu})$ such that within each block the same $(k-1)$-tuple of directed paths intersects at all directed edges, but from one block to the next the $(k-1)$-tuple of directed paths changes. Let f_ϕ be the parent in the first directed edge of block b_ϕ, and let l_ϕ be the offspring in the last directed edge of block b_ϕ. Let h map ϕ to the unique ψ such that directed path ψ does not intersect the directed edges of b_ϕ. The proof of Lemma 2 proceeds by induction on k. To show the result for $k = 3$, we construct a disjoint pair of directed paths as follows: one path follows path $h(2)$ to l_3, then follows $h(4)$ to l_5, $h(6)$ to l_7, ..., $h(2[[l/2]])$ to B, while the other path follows path $h(1)$ to l_2, $h(3)$ to l_4, ..., $h(2[[(l-1)/2]]+1)$ to B. At the induction step (going from k to $k+1$), the existence of a disjoint pair of directed paths from A to B is established by first showing the existence of a disjoint pair of directed paths from A to f_1 (note that if f_1 does not exist, then we are done), because the k paths intersecting at b_1 do not have a common intersection up to f_1. Then the existence of a disjoint pair of directed paths, p_1 and p_2, from A to f_m is used to show the existence of a disjoint pair from A to f_{m+1} for $1 \leq m \leq l-1$. This holds because there is no common intersection among the following k directed paths, where we consider each path as a path terminating at f_{m+1}: path $h(m)$, path p_2 from A to f_m followed by one of the $k-1$ paths that has a directed edge in both b_m and b_{m+1}, path p_1 from A to f_m followed by each of the other $k-2$ of the $k-1$ paths that has a directed edge in both b_m and b_{m+1}. Finally, the existence of a pair of disjoint paths, r_1 and r_2, from A to f_l is used to show the existence of a pair of disjoint paths from A to B. The k paths used to show this are: $h(l)$, r_2 followed by one of the $k-1$ paths that intersect at b_l, r_1 followed by each of a set of $k-3$ paths that intersect at b_l, not including the one used with r_2.

To prove Theorem 2 from Lemma 2, we show that given $(a,b) \in \mathcal{E}(P)$, either (i) there exist $e \in \{a\} \cup \mathcal{A}(a)$ and directed paths $q^1 = (q^1_1, ..., q^1_m)$ and $q^2 = (q^2_1, \ldots, q^2_n)$ in P, with $q^1_1 = q^2_1 = e$, $q^1_m, q^2_n \in \{i_1, \ldots, i_k\}$, such that q^1 passes through (a,b) and q^1 and q^2 have no common directed edges or (ii) $\mathcal{F}(P) \cap [\{a\} \cup \mathcal{A}(a)] \setminus \{i_1, \ldots, i_k\} = \emptyset$. Assume (ii) does not hold. To construct q^1 and q^2, choose $f \in \mathcal{F}(P) \cap [\{a\} \cup \mathcal{A}(a)] \setminus \{i_1, \ldots, i_k\}$. Consider the intersection of all directed paths from $\{f\}$ to $\{i_1, \ldots, i_k\}$. This must be empty. Otherwise f would be superfluous. Thus, by Lemma 2, there is a disjoint pair of directed paths r_1 and r_2, from $\{f\}$ to $\{i_1, \ldots, i_k\}$. If either directed path contains (a, b) then we are done, so assume not. Let r_3 be a directed path from $\{a\}$ to $\{i_1, \ldots, i_k\}$ that passes through (a, b). Consider case (i) either r_3 does not intersect r_1 or r_3 does not intersect r_2, and case (ii) r_3 intersects both r_1 and r_2. In case (i), suppose without loss of generality that r_1 and r_3 do not intersect. Then we have a disjoint pair of directed paths, one from $\{a\}$ to $\{i_1, \ldots, i_k\}$ that contains edge (a, b) and one from $\{f\}$ to $\{i_1, \ldots, i_k\}$. In case (ii), suppose, without loss of generality, that r_1 intersects r_3 before r_2 does, i.e., the last node of the first intersecting directed edge of r_1 and r_3 is ancestral to the first node of the first intersecting directed edge of r_2 and r_3. Then set r_3' to follow r_3 from a until the last node of the first intersecting edge of r_1 and r_3, and then to follow r_1. Then r_2 and r_3' are disjoint and, as in case (i), we again have a disjoint pair of paths, one from $\{a\}$ to $\{i_1, \ldots, i_k\}$ that contains edge (a, b) and one from $\{f\}$ to $\{i_1, \ldots, i_k\}$. In either case, choose r_4 from $\{f\}$ to $\{a\}$. If r_4 intersects r_2, let e be the last node of the last edge of intersection. Then consider path r_5 from $\{e\}$ to $\{i_1, \ldots, i_k\}$ which follows r_2, and path r_6 from $\{e\}$ to $\{i_1, \ldots, i_k\}$ which follows r_4 to $\{a\}$ and then r_3 or r_3' in case (i) or (ii), respectively. Then $e \in \mathcal{A}(a) \cup \{a\}$, r_6 passes through (a, b), and r_5 and r_6 are disjoint, as required.

Appendix C. Proof of Theorem 3. Suppose $\{A_t\}$ and $\{I_t\}$ are such that Conditions (1) and (2) of the theorem hold, with observable sequences $\{(\mathbf{y}, \mathbf{t})_{1,i}\}_{i=1}^{n_s}$ and $\{(\mathbf{y}, \mathbf{t})_{2,i}\}_{i=1}^{n_s}$, respectively. Let $\{B_t\}$ be any other continuous-time finite state space Markov process with finite state space $\{\alpha_1, \ldots, \alpha_m\}$, such that $\{I_t\}$ is defined by a deterministic function of $\{B_t\}$, say $I_t = g(B_t)$. For each $s \in S$, let $m_s = |g^{-1}(s)|$, and write $g^{-1}(s) = \{\beta_{s,1}, \ldots, \beta_{s,m_s}\}$. Let ρ and M be the analogues for $\{B_t\}$ of π and Q, which are defined for $\{A_t\}$. Let J^s be the $m \times m_s$ matrix with (i,j)th entry equal to $1_{\alpha_i = \beta_{s,j}}$. Let L be the matrix which is equal to the product of the matrix with rows $\{\pi^T Q^{(\mathbf{y},\mathbf{t})_{1,i}} I^s\}_{i=1}^{n_s}$ and the matrix with columns $\{I^s Q^{(\mathbf{y},\mathbf{t})_{2,j}} \mathbf{1}\}_{j=1}^{n_s}$ Under (1) and (2), L is $n_s \times n_s$ and of full rank. Furthermore, by the Markov property, L is also equal to the product of the matrix with rows $\{\rho^T M^{(\mathbf{y},\mathbf{t})_{1,i}} J^s\}_{i=1}^{n_s}$, which is of dimension $n_s \times m_s$, and the matrix with columns $\{(J^s)^T M^{(\mathbf{y},\mathbf{t})_{2,j}} \mathbf{1}\}_{j=1}^{n_s}$, which is of dimension $m_s \times n_s$. If $m_s < n_s$, one has a contradiction because L could not have rank n_s. Thus, $m = \sum_s m_s \geq \sum_s' n_{s'} = n$.

Appendix D. First note that an element of G is uniquely determined by specifying the image of a single given vertex (call it v) as well as the images of each of the d vertices connected by a single edge to v. To see that the set of orbits O can be obtained from the set of orbits O' by consideration of no more than $h(\psi)d! - 1$ symmetries of H, we choose from each $O'_k \in O'$ a representative element $v_k \in O'_k$. Denote the resulting set of elements E. Given $v \in E$, $O'_l, O'_m \in O'$, $O'_l \neq O'_m$ such that $O'_l \cup O'_m \subset O_k \in O$, then, by Lemma 3, there exists $w \in O'_l$ and $g \in G$ such that $g(v) \in E$ and $g(w) \in O'_m$. Thus if we choose v to have condensed identity state ψ, we need only consider symmetries that map v to any of the $h(\psi)$ elements of E that have identity state ψ.

Lemma 3. Given $v \in E$, $O'_l, O'_m \in O'$, $O'_l \neq O'_m$ such that $O'_l \cup O'_m \subset O_k \in O$, then there exists $w \in O'_l$ and $g \in G$ such that $g(v) \in E$ and $g(w) \in O'_m$.

Proof of Lemma 3. $O'_l \cup O'_m \subset O_k$ implies that there exists $w \in O'_l$ and $g_1 \in G$ such that $g_1(w) \in O'_m$. Let O'_n be the unique element of O' such that $g_1(v) \in O'_n$. Since O' is the set of orbits under some subgroup G' of G, there must exist $g_2 \in G' \subset G$ such that $g_2(g_1(v)) = v_n$ and $g_2(g_1(w)) \in O'_m$. Lemma 3 follows, letting $g = g_2 \circ g_1$.

The reduction to $h(\psi)d!2^{-f} - 1$ for a pairwise relationship when the pruned pedigree is outbred follows from Lemma 4, which states that, for an outbred pruned pedigree $P \in \mathcal{P}_2$, every founder, excluding i and j, has exactly 2 offspring. (Note that this no longer holds for $P \in \mathcal{P}_k, k > 2$.) Thus, for each founder in $\mathcal{F}(P) \setminus \{i, j\}$, there is a pair of edges in $\mathcal{E}(P)$ such that interchange of that founder's 2 alleles toggles the two bits corresponding to these edges, resulting in a mapping that takes each $v_1 \in \{0,1\}^{\mathcal{E}(P)}$ to a $v_2 \in \{0,1\}^{\mathcal{E}(P)}$ such that there exists $v_3 \in \{0,1\}^{\mathcal{E}(P)}$ with $|v_1 - v_3| = |v_1 - v_2| = 1$. Thus, for a given $w \in \{0,1\}^{\mathcal{E}(P)}$, there are f pairs of elements at distance 1 from w such that for each pair there exists $g \in G' \subset G$ such that g interchanges the 2 elements of the pair and preserves w, the other elements at distance 1 from w, and the orbits O' (g will interchange the relevant pair of bits if these bits are equal in w, and g will both interchange and toggle the relevant pair of bits if these bits are unequal in w). As shown above, given v with condensed identity state ψ, it is sufficient to consider all symmetries of H that map v to any of the $h(\psi)$ elements of E with identity state ψ. Suppose v is mapped to $w \in E$. We would ordinarily consider $d!$ possible ways to map elements at distance 1 from v to elements at distance 1 from w. However, as a consequence of Lemma 4, we need not consider those maps that differ only by interchanges of elements within the f pairs at distance 1 from w mentioned above, giving only $d!2^{-f}$ maps to consider.

Lemma 4. For an outbred pruned pedigree $(P, s) \in \mathcal{P}_2$, every founder, excluding i and j, has exactly 2 offspring.

Proof of Lemma 4. If $a \in \mathcal{F}(P) \setminus \{i, j\}$, then a must have at least 2 offspring; otherwise a would be superfluous. For a pair of individuals $a, b \in \mathcal{N}(P)$, write $a \leq b$ whenever $a \in \{b\} \cup \mathcal{A}(b)$. To avoid inbreeding, a could be $\leq$ at most one parent of i and one parent of j. Given $g \geq 1$, suppose that it is established that a could be $\leq$ at most one gth-generation ancestor of i and one gth-generation ancestor of j. Then, to avoid inbreeding, it follows that a could be $\leq$ at most one $(g+1)$th-generation ancestor of i and one $(g+1)$th-generation ancestor of j. Since $\mathcal{F}(P) \setminus \{i, j\} \subset \mathcal{A}(i) \cup \mathcal{A}(j)$, it follows that a can have no more than 2 offspring in P.

References

Abney, M., McPeek, M. S., and Ober C. (2000). Estimation of variance components of quantitative traits in inbred populations. *Am. J. Hum. Genet.* **66**, 629-650.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov processes. *Inequalities* **3**, 1-8.

Bishop, D. T. and Williamson, J. A. (1990). The power of identity-by-state methods for linkage analysis. *Am. J. Hum. Genet.* **46**, 254-265.

Boehnke, M. and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *Am. J. Hum. Genet.* **61**, 423-429.

Broman, K. W. and Weber, J. L. (1998). Estimation of pairwise relationships in the presence of genotyping errors. *Am. J. Hum. Genet.* **63**, 1563-1564.

Browning, S. (1998). Relationship information contained in gamete identity by descent data. *J. Comp. Biol.* **5**, 323-334.

Browning S. (2000). A Monte Carlo approach to calculating probabilities for continuous identity by descent data. *J. Appl. Prob.* **37**, 850-864.

Cannings, C. and Thompson, E. A. (1981). *Genealogical and Genetic Structure.* Cambridge University Press, Cambridge.

Donnelly, K. P. (1983). The probability that related individuals share some section of genome identical by descent. *Theoretical Population Biology* **23**, 34-63.

Ehm, M. G., Wagner, M. (1998). A test statistic to detect errors in sib-pair relationships. *Am. J. Hum. Genet.* **62**, 181-188.

Epstein, M. P., Duren, W. L. and Boehnke, M. (2000). Improved relationship inference for pairs of individuals. *Am. J. Hum. Genet.* **67**, 1219-1231.

Feingold, E. (1993). Markov processes for modeling and analyzing a new genetic mapping method. *J. Appl. Prob.* **30**, 766-779.

Gilbert, E. J. (1959). On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.* **30**, 688-697.

Gillois, M. (1964). La relation d'identité en génétique. *Ann. Inst. Henri Poincaré B* **2**, 1-94.

Göring, H. H. H., Ott, J. (1997). Relationship estimation in affected sib pair analysis of late-onset diseases. *European Journal of Human Genetics* **5**, 69-77.

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

Harris, D. L. (1964). Genotypic covariances between inbred relatives. *Genetics* **50**, 1319-1348.

Jacquard, A. (1974). *The genetic structure of populations*. Springer-Verlag, New York.

Kruglyak, L., Daly, M. J., Reeve-Daly, M. P., Lander, E. S. (1996). Parametric and non-parametric linkage analysis: a unified multipoint approach. *Am. J. Hum. Genet.* **58**, 1347-1363.

Lander, E. S. and Green, P. (1987). Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. USA* **84**, 2363-2367.

Lange, K. (1997). *Mathematical and statistical methods for genetic analysis.* Springer, New York, NY.

Lange, K., Zhao, H., and Speed, T. P. (1997). The Poisson-skip model of crossing-over. *Ann. Appl. Prob.* **7**, 299-313.

Larget, B. (1998). A canonical representation for aggregated Markov processes. *J. Appl. Prob.* **35**, 313-324.

McPeek, M. S., Speed, T. P. (1995). Modeling interference in genetic recombination. *Genetics* **139**, 1031-1044.

McPeek, M. S. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *Am. J. Hum. Genet.* **66**, 1076-1094.

Olson, J. M. (1999). Relationship estimation by Markov-process models in a sib-pair linkage study. *Am. J. Hum. Genet.* **64**, 1464-1472.

Sun, L., Abney, M., McPeek, M. S. (2001). Detection of misspecified relationships in inbred and outbred pedigrees. *Genetic Epidemiology.* **21**, S36-S41.

Sun, L., Wilder, K., McPeek, M. S. (submitted) Enhanced pedigree error detection.

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

Thompson, E. A. (1975). The estimation of pairwise relationships. *Ann. Hum. Genet.* **39**, 173-188.

Thompson, E. A. (1986). *Pedigree analysis in human genetics.* The Johns Hopkins University Press, Baltimore.

Thompson, E. A. and Meagher, T. R. (1998). Genetic linkage in the estimation of pairwise relationship. *Theor. Appl. Genet.* **97**, 857-864.

Zhao, H. and Liang, F. (2001). On relationship inference using gamete identity by descent data. *J. Comp. Biol.* **8**, 191-200.

Zhao, H., Speed, T. P., McPeek, M. S. (1995). Statistical analysis of crossover interference using the chi-square model. *Genetics* **139**, 1045-1056.

Figure Legends:

Figure 1: The pedigrees in Figures 1a and b specify the same relationship for individuals i and j, while that in Figure 1c is different.

Figure 2: The 15 possible identity states for individuals i and j, grouped according to their nine condensed identity states. Edges indicate alleles that are inherited from the same founder.

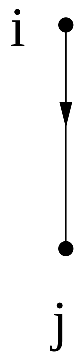


Figure 1a

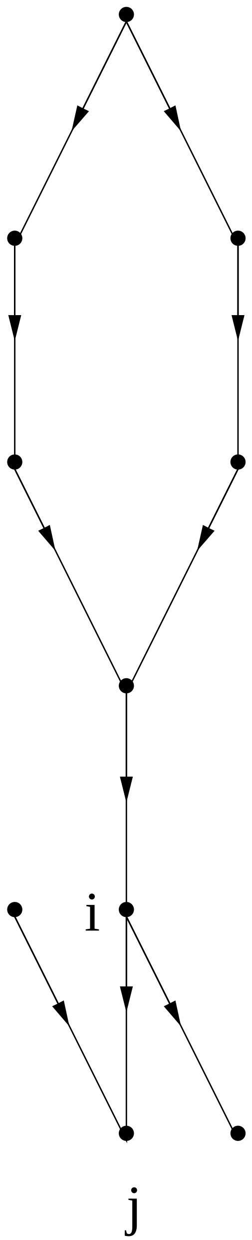


Figure 1b

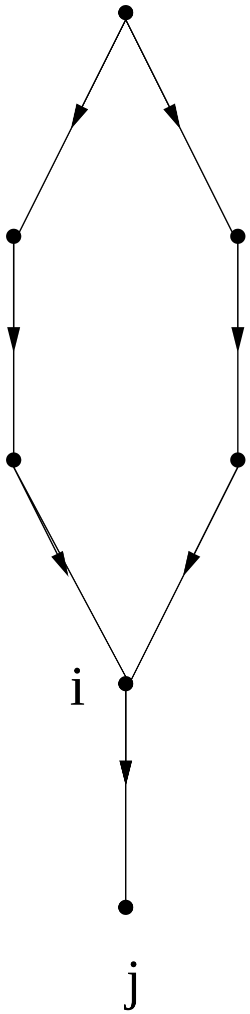
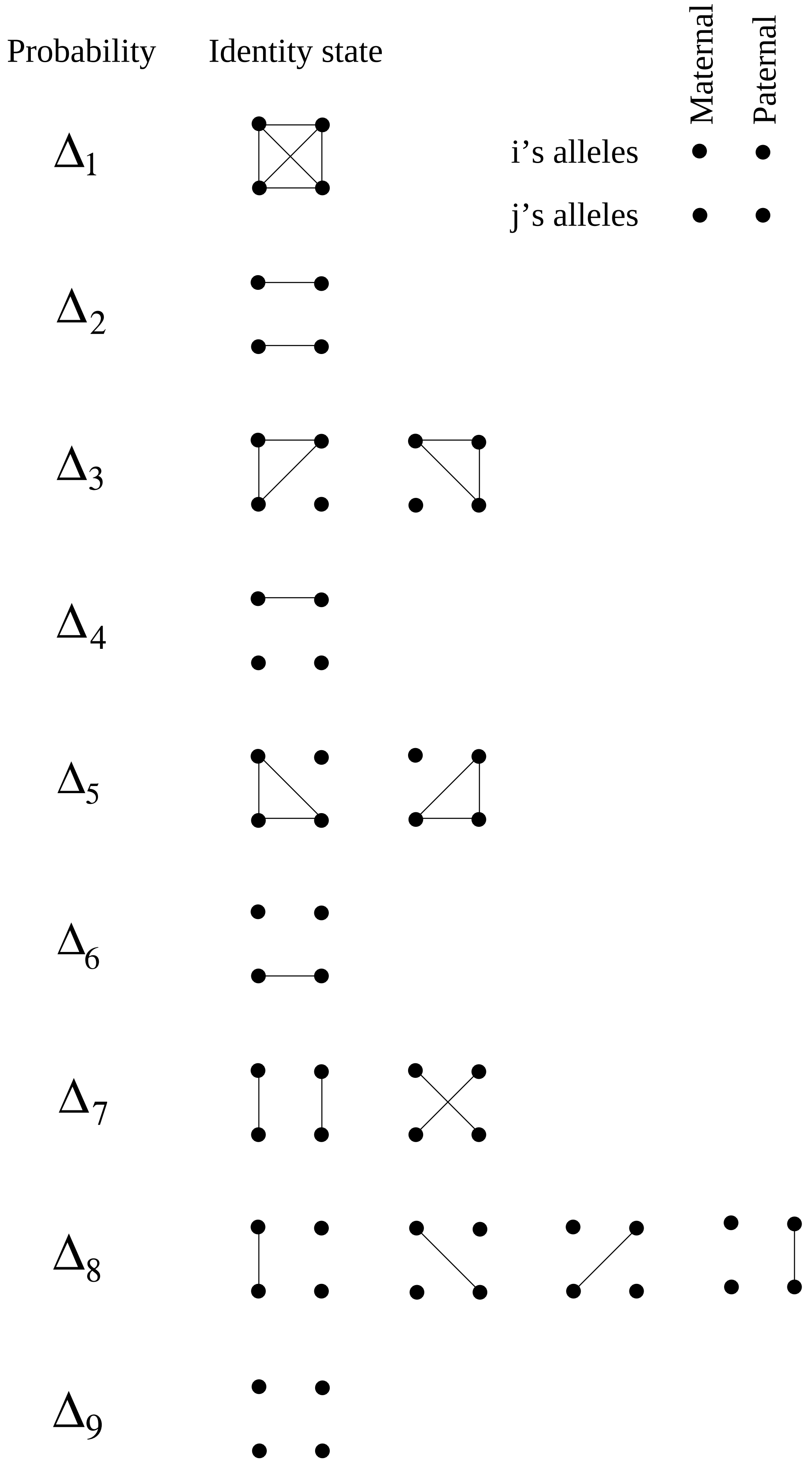


Figure 1c


Probability
Identity state
Maternal
Paternal
i’s alleles
j’s alleles
Δ_1
Δ_2
Δ_3
Δ_4
Δ_5
Δ_6
Δ_7
Δ_8
Δ_9