MULTIPOINT LINKAGE ANALYSIS USING AFFECTED RELATIVE PAIRS AND PARTIALLY INFORMATIVE MARKERS

by

Jun Teng and David Siegmund

TECHNICAL REPORT NO. 9
July 1997

Prepared Under Grant NIH 2-R01-HG00848
For The National Institutes of Health

Approved for public release; distribution unlimited

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Multipoint Linkage Analysis Using Affected Relative Pairs and Partially Informative Markers

by

Jun Teng
Department of Statistics
Stanford University

and

David Siegmund
Department of Statistics
Stanford University

TECHNICAL REPORT NO. 9
July 1997

Prepared Under Grant NIH 2-R01-HG00848
For The National Institutes of Health

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
Multipoint Linkage Analysis Using Affected Relative Pairs and Partially Informative Markers

Jun Teng and David Siegmund
Department of Statistics, Stanford University, Stanford, CA

ABSTRACT

Linkage analysis based on identity by descent of affected relative pairs is complicated by the problem that markers are often insufficiently informative allowing unambiguous determination of identity by descent. Assuming data from a set of mapped, partially informative markers, we evaluate the effectiveness of multipoint methods, as a function of the polymorphism and density of the markers, to use the correlation between nearby markers to remove this ambiguity. For the special case of pairs of half siblings whose parents are also typed, a combination of analysis and simulation is used to obtain insight into the problem of setting thresholds to control the false positive error rate. Approximations are given for the power, and guidelines are developed to help describe the trade-offs between marker density and polymorphism.

1 Introduction

Many diseases are more common in relatives of affected individuals than in the general population, and genetic factors are believed to be an important influence. One commonly used method to discover the location of the genes causing predispositions
to inherited diseases is to compare the inheritance pattern of the trait with the inheritance patterns of a set of genetic markers. If a marker is close to genes giving rise to the disease, cosegregation of the disease and the marker is expected to occur within families.

An approach that has gained popularity recently is to investigate the extent of identity by descent (IBD) of pairs of affected relatives. One can detect the linkage between a marker and disease genes by showing the amount of IBD at the markers is significantly larger than what is expected under random segregation. Although the null distribution of IBD at a single marker is easy to calculate, to search for the disease genes along the whole genome, many markers need to be tested simultaneously, and the significance level should be adjusted to this multiple testing procedure.

Our point of departure is the work of Feingold (1993) and Feingold, Brown and Siegmund (1993), who describe stochastic processes to model the identity by descent of different types of relative pairs, and derive approximations for p-values and power for these processes. Their work focuses on the idealized situation that a continuous specification of the regions of IBD, or exact specification of IBD at a set of equally spaced markers, is available. See also Lander and Botstein (1989), Lander and Schork (1994) and Lander and Kruglyak (1995) for related results.

Although future laboratory methods may realize this ideal situation, in current practice one cannot always unambiguously determine whether two relatives are IBD at every marker along the genome. Specifically, if two relatives have the same alleles at a genetic marker, one cannot tell whether these two alleles are actually inherited from a common ancestor, especially when this is a common allele. The methods proposed to cope with this important practical difficulty basically fall into two categories. One approach is to focus solely on whether affected relatives happen to show the same alleles at a locus (i.e., are identical by state (IBS)), which is compared to
Mendelian expectation (Weeks and Lange 1988, 1992). Another approach amounts to inferring IBD sharing on the basis of the marker data, which usually involves the use of genotype information for additional members of the pedigree to resolve whether or not alleles shared IBS are actually IBD (Risch 1990c).

Most theoretical research has concentrated on studying individual markers one at a time and fails to take advantage of information available from multiple markers. To exploit fully the power of a genetic map, a complete multipoint analysis using the information from all genetic markers to infer the probability distribution of the IBD status at each point along the genome is discussed by Lander and Kruglyak (1995), who implemented their analysis in the program MAPMAKER/SIBS. See also Kruglyak et al. (1996). The purpose of this paper is to assess the value of these methods.

We begin in Section 2 with a brief review of the models of Feingold (1993) and Feingold, Brown and Siegmund (1993). In Section 3 we describe our model of incompletely informative markers and propose a multipoint score statistic to test the hypothesis of no linkage. We give two correlation inequalities, which allow one to choose a conservative threshold to control the false positive error rate. We are only able to prove the stronger of these inequalities in the mathematically simplest case that our sample consists entirely of grandparent/grandchild or half sibling pairs, although we conjecture that it is true more generally. The inequalities suggest that in practice the thresholds for detection of linkage when there is complete information will be reasonable also in the case of incomplete information. This suggestion is supported by a simulation experiment. Section 4 is concerned with measuring the loss of power due to incomplete information and the success of multipoint analysis in recovering that information. In particular it is shown (Figures 4,5) that for markers having three equally likely alleles single point analysis loses substantial information
regardless of marker density, while for multipoint analysis the loss of information due
to marker spacing and incomplete polymorphism is relatively insignificant at less than
5 centimorgan (cM) intermarker distances, but is substantial at greater than 10 cM
distances. Proofs of the first correlation inequality and other technical results are
given in two Appendices.

2 The Ideal Situation

Since our goal is to understand the limitations imposed by partially informative mark-
ers in terms of the theory developed for completely informative markers, we begin
with a brief review of that theory. To simplify the exposition we suppose that our
data consist of a large number of affected grandparent-grandchild pairs or half sibling
pairs. Similar results apply with some modifications to pedigrees containing other
groups of affected relatives. In particular, the Gaussian approximations given below
for the false positive error rate and the power are quite generally valid (cf. Feingold,

We adopt the statistical model described by Feingold (1993). For the $j$th pair, let
$X^j(t)$ be 1 or 0 according as the pair is or is not IBD at the marker locus $t$, specified
by its position in cM from one end of the chromosome. Let $p_0 (= 1/2)$ denote the
probability of IBD at an arbitrary locus. Suppose there is at most one trait locus $\tau$ on
the chromosome. At $\tau$ the pairs have an excess probability of sharing an allele IBD,
say $p_\alpha = (1 + \alpha)/2$. Risch (1990a,b) presents models that allow one to evaluate $\alpha$ in
terms of allele frequencies and penetrances. Under the Haldane model specifying that
crossovers occur according to a homogeneous Poisson process, for $t$ moving in both
directions away from $\tau$, $X^j(t)$ is Markov chain on the states 0 and 1. The transition
rates can be described by the $Q$-matrix
\[
\begin{pmatrix}
-\lambda & \lambda \\
\lambda & -\lambda
\end{pmatrix},
\]
where $\lambda$ is the crossover rate per unit length. For grandparent-grandchild pairs $\lambda = 0.01/$cM; for half sibling pairs $\lambda = 0.02$.

Let $N$ be the total number of affected pairs in our sample. Let $X_t = \sum_j X_j^2(t)$ be the total number of pairs identical by descent at marker locus locus $t$, and put
\[
Z_t = (X_t - Np_0)/[Np_0(1 - p_0)]^{1/2}.
\]

It follows for large $N$ from the central limit theorem that on an unlinked chromosome $Z_t$ is approximately a stationary mean 0 Gaussian process with Ornstein-Uhlenbeck covariance function given by $\text{cov}(Z_{s+t}, Z_s) = \exp(-\beta|t|)$. Here $\beta = 2\lambda = 0.02$ for grandparent-grandchild pairs, 0.04 for half sibling pairs. On a chromosome with one trait locus at $\tau$ the expected value of $Z_t$ is
\[
E(Z_t) = \xi \exp(-\beta|t - \tau|),
\]
where $\xi = N^{1/2}(p_\alpha - p_0)/[p_0(1 - p_0)]^{1/2} = N^{1/2}\alpha$. The log-likelihood function of the observed Gaussian process $\{Z_t, 0 \leq t \leq \ell\}$ on a chromosome of length $\ell$ as a function of the unknown parameters $\tau, \xi$ equals
\[
\xi Z_\tau - \xi^2/2
\]
(Feingold, Brown and Siegmund (1993)). Hence the log-likelihood ratio statistic for testing the null hypothesis of no linkage, $\xi = 0$, against the one-sided alternative that $\xi > 0$ is
\[
\max_{1 \leq \xi \leq C} \max_{0 \leq t \leq \ell} Z_t.
\]
The first maximum is over all $C$ chromosomes, and the second is over all marker locations on each chromosome. We have suppressed the chromosome index $c$ in our description of the statistic $Z_t$ for convenience.
The false positive error rate is \( P_0\{\max_{\tau \leq t \leq t_e} \max_{0 \leq \xi < t} Z_{\tau} > b \} \), where the subscript 0 denotes that the probability is evaluated under the null hypothesis. The approximation suggested by Feingold, Brown and Siegmund (1993) in the case that the identity by descent information is available at a set of equally spaced markers \( i \Delta \) is

\[
P_0\{\max_{1 \leq \tau \leq \xi} \max_{0 \leq i \Delta \leq t} Z_{i \Delta} > b \} \approx 1 - \exp\{-C[1 - \Phi(b)] - \beta(\Sigma \ell_c)b\phi(b)\nu(b(2\beta \Delta)^{1/2})\}.
\]  

(2)

Here \( \Phi \) and \( \phi \) are the standard normal distribution and density functions, \( \Delta \) is the distance between adjacent markers and \( \nu(x) \) is a special function that can be evaluated numerically. It is reasonably well approximated by \( \exp(-0.583x) \) for \( 0 < x < 2.5 \). For larger \( x \) four terms of the defining infinite series provide a good approximation (Siegmund 1985, p. 82).

An approximation for the power depends on the location of the trait locus with respect to flanking marker loci. For the case where there is zero recombination between the trait locus and its closest marker, a simple approximation is given by

\[
P_\xi\{\max_{0 \leq i \Delta \leq \xi} Z_{i \Delta} > b \} \approx 1 - \Phi(b - \xi) + \phi(b - \xi)[2\xi^{-1}\nu - (\xi + b)^{-1}\nu^2],
\]

where \( \xi = E(Z_{\tau}) \) and \( \nu = \nu(b(2\beta \Delta)^{1/2}) \). The first term is the probability that \( Z_{\tau} \geq b \), while the second term gives the (substantially smaller) probability that \( Z_{\tau} < b \), but \( Z_{i \Delta} \geq b \) at some nearby marker different from the trait locus.

Dupuis (1994) gives an approximation for the case that the trait locus \( \tau \) is not exactly at a marker. The derivation requires that one condition on the process \( Z \) at the markers flanking the trait locus. The resulting approximation involves a one-dimensional integral to be evaluated numerically. We omit the final expression.

Remarks. (i) The situation is slightly more complicated for sibling pairs, which can be identical by descent on 0, 1, or 2 chromosomes. Let \( X_t \) denote the total number of alleles shared identical by descent by \( N \) sib pairs at the marker locus \( t \). When
penetrances are additive within loci (the only case we consider in this paper) and \( N = 1, P(X_r = i) = (1 + \alpha)/4, 1/2 \) or \((1 - \alpha)/4\) for \( i = 2, 1 \) or 0, respectively. The “mean sharing” statistic equals
\[
(X_t - N)/(N/2)^{1/2}.
\]

Its expectation at the locus \( t \) is of the form (1) with \( \xi = (N/2)^{1/2}\alpha \); also \( \beta = 0.04 \), and the approximating Gaussian process has covariance function \( \exp(-\beta|t|) \) (cf. Feingold, Brown and Siegmund 1993). (ii) In general there can be more than a single trait predisposing locus, although we have assumed that there is no more than one on any one chromosome. Strictly speaking what we have called the likelihood function, is a marginal likelihood for the data of a single chromosome. For methods that search for a single locus at a time, this distinction is unimportant. (iii) If the markers are not equally spaced, one can often replace (2) by a slightly more complicated approximation involving the average marker density. For example, suppose the position of the \( i \)th marker on chromosome \( c \) is located at \( K_c(i\Delta) \), where \( K_c \) is a differentiable function satisfying \( K_c(0) = 0, K_c(\ell_c) = \ell_c \). Then (1) continues to hold provided we replace \( \Sigma_c \ell_c \nu[\Delta(2\beta\Delta)^{1/2}] \) by \( \Sigma_c \int_0^{\ell_c} K'_c(x) \nu[\Delta(2\beta\Delta K'_c(x))^{1/2}]dx \). (iv) Our test statistic involves the process at the markers only because we think that interval mapping (Lander and Botstein 1989) to interpolate between markers does not appreciably increase the power. For a detailed discussion, see Dupuis and Siegmund (1997). With some minor technical complications our analysis can be extended to incorporate interval mapping.

3 Partially Informative Markers

It is convenient to introduce the following notation. We consider a chromosome of length \( \ell \) cM’s containing at most one trait susceptibility gene at locus \( \tau \). Assume
there are $M$ markers at loci $t_1, t_2, \ldots, t_M$ on the chromosome. For each pedigree $j$, let $X^j(t)$ be the number of alleles shared IBD by the affected relative pairs at locus $t$, and let $G^j(t_i)$ denote the genotype information for all typed members of the pedigree at marker $t_i$. Each pedigree may consist of only the affected relative pairs or may include other additional relatives. Let $G^j = (G^j(t_1), G^j(t_2), \ldots, G^j(t_M))$ be the total genotype information for the $j^{th}$ pedigree.

Throughout we assume a random mating population, so allele frequencies are in Hardy-Weinberg equilibrium and genotypes at marker loci are in linkage equilibrium with genotypes at trait predisposing loci.

If only affecteds are genotyped, the information necessary to infer IBD status is knowledge of marker allele frequencies. A preferable situation is that other members of the pedigree can also be genotyped. In the case of grandparent-grandchild pairs, the appropriate pedigree members are the mate of the affected grandparent, the intervening parent, and the mate of that parent. In the case of half sibs they are the common parent of the half sibs and the mates of that parent. See Risch (1990c) for a complete discussion of these and other situations for single marker analysis.

3.1 Grandparent/Grandchild Pairs; Half Sibling Pairs

We consider first the simple cases of grandparent-grandchild or half sibling pairs. The likelihood function for a pedigree is the probability of its observed marker data. So for the $j^{th}$ family, the likelihood is $(dP_a/dP_0)(G^j)$. For the marker data augmented by the ideal IBD information the likelihood is $(dP_a/dP_0)(G^j, X^j(\cdot))$. If there is a single trait locus at $\tau$, then $(dP_a/dP_0)(G^j, X^j(\cdot)) = (1 + \alpha)^{X^j(\tau)}(1 - \alpha)^{1-X^j(\tau)}$. (In fact, if only $X^j(\tau)$ were observed, this would be obvious. Since the entire process $X^j(\cdot)$ is generated from $X^j(\tau)$ by the recombination process, which does not involve $\alpha$, the likelihood function would not change if the entire process $X^j(\cdot)$ were observed.
Finally by the assumption of linkage equilibrium there is no additional information about \( \alpha \) in the actual marker genotypes once the complete IBD information is known.) Consequently

\[
(dP_\alpha/dP_0)(G^j) = E_0[(dP_\alpha/dP_0)(G^j, X^j(\cdot))|G^j]
\]

\[
= (1 + \alpha)P_0(X^j(\tau) = 1|G^j) + (1 - \alpha)P_0(X^j(\tau) = 0|G^j),
\]

from which it follows by differentiation with respect to \( \alpha \) that the efficient score at \( \alpha = 0 \) is

\[
\frac{\partial}{\partial \alpha} \sum_j \log[(dP_\alpha/dP_0(G^j))|_{\alpha=0} = \sum_j [2P_0(X^j(\tau) = 1|G^j) - 1].
\]

For simplicity, we introduce the notation

\[
Y^j(t) = P_0(X^j(t) = 1|G^j).
\]

If the trait location \( \tau \) were known, the score test for the null hypothesis \( \alpha = 0 \) would be

\[
Z_2(\tau) = \sum_j (Y^j(\tau) - E_0[X^j(\tau)])/(N^{1/2}\sigma_Y(\tau)), \tag{3}
\]

where \( N \) is total number of pedigrees and \( \sigma_Y^2(t) \) is the variance of \( Y^j(t) \) under \( H_0 \).

When the trait locus is unknown, a test statistic for global search is

\[
\max_{1 \leq i \leq M} Z_2(t_i).
\]

Compared to the statistic used by Feingold (1993), we have replaced the exact IBD count, \( X^j(t_i) \), at each marker locus by its estimator \( Y^j(t_i) \), equal to the conditional expectation of \( X^j(t_i) \) given all marker genotypes under the hypothesis \( \alpha = 0 \) of no linkage. For this test statistic, we have to re-examine the proper threshold \( b \) for declaring linkage. If we were to use single point analysis, so at each marker
we condition only on the observations at that marker, the consequence would be to add noise to the ideal IBD process at each marker. This would necessitate a larger threshold than in the ideal case. The impact of multipoint analysis, where we condition on observations at all markers, is to smooth the single point process. The problem is to understand the total impact of adding noise and smoothing. The theorem given below shows that the multipoint process is actually more correlated than the original process, so the threshold for the ideal process is a conservative bound on the threshold for the multipoint process.

This analysis can be made precise by regarding $Z_2(t)$ as a Gaussian process. Gaussian models have the limitation of needing relatively large numbers of affected relative pairs to be valid, but they allow the rich literature of normal theory to be applied and give useful insight into the problem at hand.

Our statistic $Z_2(t)$ was standardized, so marginally it has approximately a standard normal distribution under the null hypothesis. To find the threshold defining the critical region for this statistic, one needs to find the correlation function for $Z_2(t)$. For the ideal case that the exact IBD is known, the correlation function is $\exp(-\beta|t|)$, where $\beta = .02$ for grandparent-grandchild pairs and $= 0.04$ for half sibling pairs. When the exact information is not available and the IBD status has to be estimated from the marker genotypes, the calculation of the covariance function becomes much more difficult if not impossible. So instead of calculating the exact covariance function, we establish an inequality between the covariance functions of our new statistic and the one based on exact IBD. Based on this inequality, we can show that the thresholds for the new process will be smaller than those based on the exact IBD. To state this result it will be convenient to let $Z_1(t)$ denote the complete information process $Z_t$ defined in Section 2.

**Theorem.** For the limiting Gaussian processes $Z_1$ and $Z_2$, under the null hypothesis
that $\alpha = 0$,

$$P_0\{\max_{1 \leq i \leq M} Z_1(t_i) \geq b\} \geq P_0\{\max_{1 \leq i \leq M} Z_2(t_i) \geq b\}.$$  

The following three results will be used in the proof of the theorem.

**Proposition 1.** ("Slepian's Inequality")

Let $\{\zeta_1(t)\}$ and $\{\zeta_2(t)\}$ be normal processes (possessing continuous sample functions but not necessarily stationary). Suppose that they are standardized so that $E(\zeta_1(t)) = E(\zeta_2(t)) = 0$, and $E(\zeta_1(t)^2) = E(\zeta_2(t)^2) = 1$, and write $\rho_1(t,s)$ and $\rho_2(t,s)$ for their correlation functions. Suppose that for some $c > 0$ we have $\rho_1(t,s) \geq \rho_2(t,s)$ when $0 \leq t, s \leq c$. Then the maxima $M_i(t) = \max_{s \leq t} \zeta_i(s)$ satisfy

$$P\{M_1(T) \leq u\} \geq P\{M_2(T) \leq u\}$$

when $0 \leq T \leq c$.

A proof of Proposition 1 can be found in Leadbetter, Lindgren and Rootzén (1983) p.156.

**Lemma 1.** Let $Z_1(t_i)$ and $Z_2(t_i)$ be defined as before. Then under the null hypothesis, for all $i,j$

$$\text{cov}(Z_1(t_i), Z_1(t_j)) \leq \text{cov}(Z_2(t_i), Z_2(t_j)).$$

Because the distribution of $\{X^k(t_i), Y^k(t_i), G^k(t_i), 1 \leq i \leq M\}$ does not depend on the particular family $k$ chosen, it is convenient to drop the superscript $k$ in the following arguments.

**Lemma 2.** Under the hypothesis $\alpha = 0$

$$E_0[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \geq E_0[Y(t_i)|X(t_i) = \delta]$$

(4)
where $\delta = 0$ or 1.

The theorem is an immediate consequence of the Proposition and Lemma 1, which in turn follows from Lemma 2. The two lemmas are proved in Appendix A.

**Remark:** The standard deviation $\sigma_Y(t)$ is in general difficult to compute exactly. In practice it can be replaced by 
$$\hat{\sigma}_Y(t) = \left[ \sum_{k=1}^{n} (Y^k(t) - p_0)^2/n \right]^{1/2},$$
where $p_0 = E_0(Y^k(t)) = E_0(X^k(t)) (= 0.5$ for grandparent/grandchild and for half siblings). The asymptotic properties of the process $Z_2(t)$ do not change, although the success of the asymptotic theory to describe small to moderate sample behavior may be affected.

The theorem gives an upper bound on the total false positive rate for multipoint analysis and consequently allows us to derive a conservative threshold for the test. We have also compared the thresholds for the ideal process and for the multipoint process in a Monte Carlo experiment. We assume each marker is either completely informative or noninformative. This corresponds to the case that other relevant relatives are genotyped. (See Risch 1990c.) The simulation involves $N = 100$ affected pairs, equally spaced markers $\Delta$ cM apart, with each marker having probability 0.5 to be informative. There are 100,000 repetitions of the Monte Carlo experiment. The 95th quantile of the test statistic is presented in Table 1 for half sibling pairs. In all the calculations we assume a human genome consisting of 23 chromosomes of length 140 cM each.

### 3.2 Other Relative Pairs

The proof of the theorem given above depends very heavily on the facts that the IBD process for a single pair is a two state Markov chain. This presupposes the no interference model for crossovers and is limited to the two cases of grandparent-grandchild and half sibling pairs. Although we suspect a similar inequality is true for other relative pairs including siblings, and perhaps for allele sharing statistics
involving more than two affecteds in a pedigree, we are unable to prove it for these cases.

Here we give a weaker inequality that holds under much more general conditions. For sibling pairs, under the hypothesis of additivity of penetrances within loci, the likelihood function of the ideal data for the kth pair is
\[(1 - \alpha)^{X_{2}^k(\tau)}X_{0}^k(\tau),\]
where \(X_{m}^k(\tau)\) is the indicator that the kth sib pair have inherited \(m\) alleles identical by descent at the trait locus \(\tau\). By the same argument as above the likelihood for the observed data is
\[(dP_\alpha/dP_0)(G^k) = 1 + \alpha[Y_{2}^k(\tau) - Y_{0}^k(\tau)],\]
where \(Y_{m}^k(\tau) = E_0[X_{m}^k(\tau)|G^k].\) A similar representation for other unilineal relative pairs is obtained in Appendix B by a similar but somewhat more complicated argument. In both cases the score statistic is again of the form (3), where an exact IBD count is replaced by its conditional expectation given the marker data, and the statistic is renormalized to have unit variance under the hypothesis of no linkage. The following argument applies to any statistic of this form, whether or not the original statistic is the score statistic for a particular genetic model. To study the p-value of the statistic, we once again use the normal approximation to give an (asymptotic) upper bound by establishing an inequality between the correlation of our new statistic and the one based on the exact IBD process. In the following proposition, all the expectations are taken under the null hypothesis. We omit the subscript 0 for convenience.

**Proposition 2.** Let
\[R(t) = \text{corr}(X(s), X(s + t)).\]
be the correlation function for an ideal IBD statistic. Then
\[\text{cov}(Z_2(t_i), Z_2(t_j)) \geq R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1),\]  
(5)
where
\[c_{i,j} = [\sigma^2_X(t_i) + \sigma^2_X(t_j)]/[\sigma^2_Y(t_i) + \sigma^2_Y(t_j)].\]
Proof. By the properties of conditional expectations and the independence of different pedigrees

\[
\text{cov}(Y(t_i), Y(t_j)) = \frac{1}{2}[\sigma^2_Y(t_i) + \sigma^2_Y(t_j)] - \frac{1}{2}E(Y(t_i) - Y(t_j))^2
\]

\[
\geq \frac{1}{2}[\sigma^2_Y(t_i) + \sigma^2_Y(t_j)] - \frac{1}{2}E(X(t_i) - X(t_j))^2
\]

\[
= \frac{1}{2}[\sigma^2_Y(t_i) + \sigma^2_Y(t_j)] - \frac{1}{2}(1 - R(t_j - t_i))[\sigma^2_X(t_i) + \sigma^2_X(t_j)].
\]

Therefore,

\[
\text{cov}(Z_2(t_i), Z_2(t_j)) = \text{corr}(Y(t_i), Y(t_j))
\]

\[
= \frac{\text{cov}(Y(t_i), Y(t_j))}{\sigma_Y(t_i)\sigma_Y(t_j)}
\]

\[
\geq \frac{\text{cov}(Y(t_i), Y(t_j))}{[(\sigma^2_Y(t_i) + \sigma^2_Y(t_j))/2]}
\]

\[
\geq 1 + (1 - R(t_j - t_i))c_{i,j}
\]

\[
= R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1).
\]

Feingold, Brown and Siegmund (1993) have shown in a number of examples that the covariance function \(R(t)\) of the exact IBD process satisfies \(R(t) = 1 - \beta|t| + o(t)\) as \(t \to 0\), where \(\beta\) is a parameter measuring the rate of recombination for the relative pairs involved. For example, when genetic distance \(t\) is measured cM, for aunt-niece pairs, \(\beta = 0.05\). The right hand side of the preceding display can be written in the form

\[
R(t_j - t_i) - (1 - R(t_j - t_i))(c_{i,j} - 1)
\]

\[
= (1 - \beta|t_j - t_i|) - \beta|t_j - t_i|(c_{i,j} - 1) + o(|t_j - t_i|)
\]

\[
= 1 - c_{i,j} \beta|t_j - t_i| + o(|t_j - t_i|).
\]

Remarks: The constants \(c_{i,j}\) are always greater or equal to 1, and equality holds when the exact IBD status can be recovered from the data. This can be achieved when the
markers are 100 percent polymorphic or they are infinitely dense. Based on this upper bound for the covariance function, we see the basic structure of the covariance – linear decay near \( t = 0 \) – does not change. For the single point analysis, i.e. making inference about the IBD configuration based on the single marker alone, there is a substantial change in the covariance structure. Let \( Z_3(t_i) \) denote the appropriately normalized sum over \( k \) of \( E_0[X^k(t_i)|G^k(t_i)] \), which only involves the marker information at \( t_i \). If, for example, all the markers have the same degree of polymorphism, then

\[
cov(Z_3(t_i), Z_3(t_j)) = \begin{cases} 
1 & \text{if } t_i = t_j \\
f(1 - \beta |t_i - t_j| + o(|t_i - t_j|) & \text{if } t_i \neq t_j
\end{cases}
\]

where \( f \) is a constant less than 1. Because the process \( Z_3 \) is substantially less correlated for single point analysis, we may need a significantly larger threshold to control the total false positive rate. In addition, its expectation is smaller when there is linkage, as we shall see below.

If the markers are equally spaced at \( \Delta \) cM apart, and have the same degree of polymorphism, then \( \sigma_X^2(t_i) \) would be approximately equal across markers, say \( \sigma_X^2(t_i) \approx \sigma_X^2 \), and hence \( c_{i,j} \approx c =: \sigma_X^2 / \sigma_Y^2 \). Then using Proposition 2, Slepian’s inequality, and (2), we obtain as an approximate upper bound for the p-value the right hand side of (2) with \( \beta \) replaced by \( c\beta \). If \( \sigma_X^2(t_i) \) is not constant, but can be assumed to be given by a differentiable function of \( t_i \), one can use an average value of \( c(t_i) = \sigma_X^2(t_i) / \sigma_Y^2(t_i) \) exactly as in the remark at the end of Section 2, where we discussed the possibility that the spacing between markers is not constant.

The threshold based on this upper bound is larger than the threshold derived for the exact IBD case. However, since a small perturbation of the threshold will result in a large change of the p-value, the discrepancy between these two thresholds tends to be small unless we have very large value of \( c \). For example, for sib pairs \( \beta = 0.04 \), and for a 5-cM map the 0.05 threshold obtained from (2) for the IBD process is \( b = 3.73 \). For
$c = 1.5$, which we shall see below (Table 2) corresponds to having two equally likely alleles at each marker, the upper bound is $b = 3.76$. Thus even with the relatively simple argument of Proposition 2 we find evidence that thresholds based on ideal IBD data are approximately correct for multipoint analysis of partially informative markers. Note, however, that the upper bound obtained by this argument is larger for less informative markers, whereas the simulations in Tables 1 and 2 show that multipoint analysis of less informative markers leads to a more correlated process, hence a lower threshold.

4 Efficiency

In the previous sections, we described multipoint analysis for affected relative pairs, and derived upper bounds for p-values. Here we try to evaluate the effectiveness of multipoint analysis in recovering the information about the status of identity by descent for different degrees of informativeness and density of the markers.

We first calculate the expectation of $Z_2(t)$ under the alternative hypothesis. We have

$$E_\alpha[Z_2(t)] = \frac{N^{1/2}}{\sigma_Y(t)}E_\alpha[Y(t) - \mu_0],$$

where $\mu_0 = p_0$ for unilineal relatives and $= 1$ for the mean sharing statistic for sibling pairs; Hence by the representations of the likelihood ratio derived in Sections 3.1 and 3.2 (cf. also Appendix B)

$$E_\alpha[Y(t)] - \mu_0 = E_0[Y(t) - \mu_0][(dP_\alpha/dP_0)(G)]$$

$$= p_0^{-1} \alpha \text{Cov}_0[Y(t), Y(\tau)]$$

in the case of unilineal relative pairs and

$$= \alpha \text{Cov}_0[Y(t), Y(\tau)]$$

16
in the case of sibs. Hence

\[ E_a[Z_2(t)] = \xi \frac{\sigma_Y(\tau)}{\sigma_X(\tau)} \text{corr}_0(Y(t), Y(\tau)); \]  

(6)

in particular

\[ E_a[Z_2(\tau)] = \xi \sigma_Y(\tau)/\sigma_X(\tau), \]

where \( \xi \) is the noncentrality parameter defined by \( \xi = [EX_aX(\tau) - E_0X(\tau)]/\sigma_X(\tau) \).

The preceding calculations suggest, in the special case where the trait locus \( \tau \) is exactly at one of the marker loci, that if instead of using the statistic \( Z_1 \) based on the exact IBD, we use the statistic \( Z_2 \) based on the estimated IBD, we would have approximately the same power if we have \( \sigma^2_X(\tau)/\sigma^2_Y(\tau) \) as many observations. A rough analysis goes as follows. If we only test for one candidate gene locus \( \tau \), then this is the usual definition of asymptotic relative efficiency of \( Z_2 \) with respect to \( Z_1 \), which says to obtain the same power, asymptotically \( \sigma^2_X(\tau)/\sigma^2_Y(\tau) \) as many observations are required with \( Z_2 \) as with \( Z_1 \). However, we are in a more complicated situation. Our tests are based on \( \max_{1 \leq i \leq M} Z_j(t_i) \geq b \) for \( j = 1, 2 \), so the usual asymptotic arguments are changed in two ways. First although marginally \( Z_1(t) \) and \( Z_2(t) \) are the same under the null hypothesis, they have different covariance functions, so different thresholds \( b \) are required for a given false positive rate. As we noticed earlier, there are usually minor changes in \( b \), so this difference plays an almost insignificant role. Secondly, the power of the test can be written as

\[ P(\max_{1 \leq i \leq M} Z_j(t_i) \geq b) = P(Z_j(\tau) \geq b) + P(\max_{1 \leq i \leq M} Z_j(t_i) \geq b, Z_j(\tau) < b). \]  

(7)

The first term accounts for the probability that the statistic will exceed the threshold at the trait locus \( \tau \) and is a simple function of the noncentrality parameter \( E_a[Z_j(\tau)] \). The second term is the probability that the process is below the threshold at the trait locus \( \tau \) but, because of random variation, exceeds the threshold at some nearby locus.
This depends not only on the noncentrality parameter, but also on the covariance function. However, the second term makes a relatively minor contribution to the power unless the markers are closely spaced and the power is low to moderate. Hence the ratio of the square of the noncentrality parameters, or, equivalently, $\sigma^2_Y(\tau)/\sigma^2_X(\tau)$, can still be used as a rough measure of relative efficiency in our situation.

The ratio $\sigma^2_Y(t)/\sigma^2_X(t)$ can also be interpreted as the relative amount of information recovered by the multipoint analysis at the locus $t$. Kruglyak and Lander (1995) suggested defining the information of the multipoint analysis as $1 - \text{var}[X(t)|G]/\sigma^2_X(t)$, but observed that this measure can take on negative values. In view of the identity $\sigma^2_X(t) = \sigma^2_Y(t) + E[\text{var}(X(t)|G)]$, one sees that the two definitions are closely related, but ours cannot take on negative values.

Some numerical examples are given below for half-siblings. We assume that the markers are equally spaced and each marker has the same number of alleles, all of which occur with equal frequency. Also in calculating the noncentrality, we assume the chromosome is infinitely long so that we can ignore end effects. Figure 1 gives the relative efficiency for the case that other relevant relatives are also typed, while Figure 2 gives the relative efficiency for the case that only half siblings are available. The loss of efficiency compared to fully informative markers is quite substantial over a broad range of marker informativeness, unless a 1 cM map is used. When other relatives are typed, the loss of efficiency is comparatively modest with three or more equally likely alleles and marker densities up to 5 cM, and with six or more equally likely alleles and marker densities up to 10 cM.

For the more interesting case that the trait locus is not exactly at one of the marker loci, there is no adequate measure of efficiency. A power calculation along the lines of (7) must be modified by considering the joint distribution of $Z_2(t_i), Z_2(t_{i+1})$ at the two markers flanking a trait locus, and in particular the maximum of these two
variables. We have used the approximation mentioned in Section 2 for ideal (completely informative) data to derive approximations for the power. We use a bivariate normal approximation for the joint distribution of $Z_2(t_i), Z_2(t_{i+1})$, with mean values given by (6) and correlation calculated numerically. We approximate the conditional probability given $Z_2(t_i), Z_2(t_{i+1})$ that $Z_2(t_j) \geq b$ at some other marker $t_j$ by the formula for the ideal case. In principle, $\text{corr}_\alpha(Z_2(t_i), Z_2(t_{i+1}))$ depends on $\alpha$, but for small to moderate values of $\alpha$, it is approximately $\text{corr}_0(Z_2(t_i), Z_2(t_{i+1}))$; and the latter is what we used in the numerical examples.

Numerical parameters for performing these approximations have been evaluated by simulation under the assumptions that all markers have the same number of equally likely alleles. The values are given in Table 2. The first entry is the correlation between $Z_2(t_i), Z_2(t_{i+1})$, the second is the factor $\sigma_Y(\tau)/\sigma_X(\tau)$ by which the noncentrality $\xi$ is reduced when the trait locus is itself a marker locus, and the third entry is the more complicated correction factor given in (6) when the trait locus is midway between two marker loci. As expected the correlations decrease and the noncentrality factors increase with the number of alleles.

We have simulated a simple case to check the suggested approximations. The trait locus is either at a marker locus or midway between marker loci. We assume each marker is either fully informative or noninformative with equal probability. This corresponds to the case that parents of the half sibling pairs are typed and there are approximately three equally likely alleles. The value of $\alpha$ is taken to be 0.25. The results given in Figures 3 and 4 are sample sizes necessary to achieve 90% power under various conditions. In the ideal case with dense markers, to have power of 90% to detect linkage, we need a value of $\xi = N^{1/2} \alpha \approx 5$, hence a sample size of about $N = 400$. The sample sizes for the ideal case are calculated using the approximations described in Section 2. The theoretical power calculations for the case of partially
informative markers are based on the nominal thresholds, while the simulated sample sizes are based on simulated thresholds (cf. Table 1). Since the nominal thresholds are larger than the simulated thresholds, the sample sizes based on the theoretical approximations are generally slightly larger than those based on simulation.

From the simulations we see that the theoretical approximations are reasonable. The advantage of using multipoint methods is quite substantial. The greatest percentage gains occur when the intermarker distances are relatively small, when multipoint analysis can be almost as efficient as completely informative markers. For completely informative markers the loss of efficiency as a function of intermarker distance is relatively insignificant up to a distance is about 10 cM; but for partially informative markers a comparable loss of efficiency shows up at about 5 cM.

Additional calculations based on the parameters in Table 2 show a greater loss of efficiency compared to completely informative markers occurs when a trait locus is at a marker locus than when it is midway between marker loci. The loss is very roughly \((1/m) \times 100\%\) for \(3 \leq m \leq 20\) equally likely alleles.

5 Discussion

We have studied an idealized model of multipoint gene mapping using affected relative pairs and partially informative markers, with non affected relatives available for marker typing. Thresholds to control the false positive error rate, derived under the assumption of completely informative markers, are still appropriate. Compared to single point analysis, multipoint analysis is substantially more efficient, with the greatest gains in efficiency when the markers are relatively closely spaced. Compared to the ideal situation of completely informative markers the efficiency of multipoint analysis is a complicated function of the marker informativeness and map density. For moderately informative markers (three equally likely alleles) and map densities
up to about 5 cM, there is only a modest loss of efficiency, which increases rapidly at greater intermarker distances.

When unaffected relatives are unavailable for marker typing, e.g., for late onset diseases where parents may already be deceased, one can in principle use allele frequencies to infer identity by descent status from marker data. Although most of our analytic results apply to this case as well, and the picture provided by Figure 2 of the value of multipoint analysis is almost certainly qualitatively correct, some preliminary simulations indicate that the detailed picture is more complicated. Although the complete information statistic at each marker is symmetrically distributed (under the hypothesis of no linkage), the multipoint statistic has a skewed distribution when intermarker distances are large or the information content of the markers is low. Consequently the process is not approximately Gaussian unless the sample size is very large, so the threshold appropriate for completely informative markers can be anticonservative unless one makes a correction for skewness. When markers are closely spaced and highly informative, this problem seems to be less bothersome.

The approximations of Section 2 for the ideal case of completely informative markers do not require the assumption of no crossover interference, while this assumption plays a very important role in our analysis of incompletely informative markers. It would be interesting to know how a multipoint algorithm based on the no interference assumption performs when interference is present.

REFERENCES


loci from a dense set of markers, submitted for publication.


APPENDIX A

**Proof of Lemma 2.**

Without loss of generality, we may assume $j > i$. To prove (4), we first derive for $\delta' = 0$ or 1 the inequality

$$E_0[Y(t_i) | X(t_i) = \delta', X(t_{j-1}) = \delta', X(t_j) = 1] \geq E_0[Y(t_i) | X(t_i) = \delta, X(t_{j-1}) = \delta'],$$

(A1)

which is equivalent to

$$E_0[Y(t_i) | X(t_i) = \delta', X(t_{j-1}) = \delta', X(t_j) = 1] \geq E_0[Y(t_i) | X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0].$$

(A2)
We have

\[ E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \]

\[ -E_0[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0] \]

\[ = \sum_G P(X(t_i) = 1|G)[P(G|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1) \]

\[ -P(G|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 0)] \]

\[ = \sum_G P(X(t_i) = 1|G)P(G_1, \cdots, G_{j-1}|X(t_i) = \delta, X(t_{j-1}) = \delta') \]

\[ [P(G_j, \cdots, G_M|X(t_j) = 1) - P(G_j, \cdots, G_M|X(t_j) = 0)], \]

where to simplify the notation we have written \( G_k \) for \( G(t_k), \ k = 1, \cdots, M \). This last expression can be rewritten to equal the average of

\[ \sum_{G_j, \cdots, G_M} P(X(t_i) = 1|G)[P(G_j, \cdots, G_M|X(t_j) = 1) \]

\[ -P(G_j, \cdots, G_M|X(t_j) = 0)] \]  

(A3)

with respect to the conditional distribution of \( P(G_1, \cdots, G_{j-1}|X(t_i) = \delta, X(t_{j-1}) = \delta') \). Hence it suffices to show that (A3) is non-negative for every \( G_1, \cdots, G_{j-1} \).

To prove this non-negativity, we use induction. We first prove (A3) is non-negative when \( i = j \). Since

\[ P(X(t_j) = 1|G) - P(X(t_j) = 1|G_1, \cdots, G_{j-1}) \]

\[ = \frac{P(X(t_j) = 1, G)}{P(G)} - \frac{P(X(t_j) = 1, G_1, \cdots, G_{j-1})}{P(G_1, \cdots, G_{j-1})} \]

\[ = \frac{P(X(t_j) = 1, G_1, \cdots, G_{j-1})P(G_j, \cdots, G_M|X(t_j) = 1)}{P(G_1, \cdots, G_{j-1})} \]

\[ - \frac{P(X(t_j) = 1, G_1, \cdots, G_{j-1})}{P(G_1, \cdots, G_{j-1})} \]

\[ = P(X(t_j) = 1, G_1, \cdots, G_{j-1}) \]

\[ \frac{P(G_j, \cdots, G_M|X(t_j) = 1) - P(G)}{P(G)P(G_1, \cdots, G_{j-1})} \]

24
\[ \begin{align*}
\frac{P(X(t_j) = 1, G_1, \cdots, G_{j-1})}{P(G)P(G_1, \cdots, G_{j-1})} & \times [P(G_1, \cdots, G_{j-1})P(G_j, \cdots, G_M|X(t_j) = 1) \\
& - P(G_1, \cdots, G_{j-1}, X(t_j) = 1)P(G_j, \cdots, G_M|X(t_j) = 1) \\
& - P(G_1, \cdots, G_{j-1}, X(t_j) = 0)P(G_j, \cdots, G_M|X(t_j) = 0)] \\
= & \frac{P(X(t_j) = 1, G_1), \cdots, G_{j-1})P(X(t_j) = 0, G_1, \cdots, G_{j-1})}{P(G)P(G_1, \cdots, G_{j-1})} \\
& \times [P(G_j, \cdots, G_M|X(t_{j-1}) = 1) - P(G_j, \cdots, G_M|X(t_{j-1}) = 0)] \\
\end{align*} \]

we have

\[ \sum_{G_j, \cdots, G_M} \frac{(P(X(t_j) = 1|G) - P(X(t_j) = 1|G_1, \cdots, G_{j-1})}{P(G)P(G_1, \cdots, G_{j-1})} \\
[P(G_j, \cdots, G_M|X(t_{j-1}) = 1) - P(G_j, \cdots, G_M|X(t_{j-1}) = 0)] \\
\geq 0. \]

But

\[ \sum_{G_j, \cdots, G_M} \frac{P(X(t_j) = 1|G_1, \cdots, G_{j-1})}{P(G_1, \cdots, G_{j-1})} \\
[P(G_j, \cdots, G_M|X(t_{j-1}) = 1) - P(G_j, \cdots, G_M|X(t_{j-1}) = 0)] \\
= 0. \]

So (A3) is non-negative when \( i = j \).

We now assume that (A3) is non-negative for some value \( i \leq j \) and prove that it remains so for \( i - 1 \). We have

\[ \sum_{G_j, \cdots, G_M} \frac{P(X(t_{i-1}) = 1|G)[P(G_j, \cdots, G_M|X(t_j) = 1) \\
- P(G_j, \cdots, G_M|X(t_j) = 0)]}{P(G_1, \cdots, G_{j-1})} \]
\[\begin{align*}
&= \sum_{G_j, \ldots, G_M} P(X(t_i) = 1|G_1, \ldots, G_{i-1}) \\
&\quad \times \frac{1}{P(X(t_i) = \epsilon|G_1, \ldots, G_{i-1})} \left[ P(G_j, \ldots, G_M|X(t_j) = 1) - P(G_j, \ldots, G_M|X(t_j) = 0) \right] \\
&= P(X(t_{i-1}) = 1|G_1, \ldots, G_{i-1}) \\
&\quad \times \sum_{G_j, \ldots, G_M} \left[ P(X(t_i) = 1|G_1, \ldots, G_{i-1}) \frac{P(X(t_i) = 1|X(t_{i-1}) = 1)}{P(X(t_i) = 1|G_1, \ldots, G_{i-1})} \\
&\quad \quad \quad + P(X(t_i) = 0|G_1) \frac{P(X(t_i) = 0|X(t_{i-1}) = 1)}{P(X(t_i) = 0|G_1, \ldots, G_{i-1})} \right] \\
&\quad \quad \quad \left[ P(G_j, \ldots, G_M|X(t_j) = 1) - P(G_j, \ldots, G_M|X(t_j) = 0) \right] \\
&\geq 0.
\end{align*}\]

Hence (A3) is non-negative, and consequently inequalities (A1) and (A2) are true.

To complete the proof of (4), we again use an induction argument. As a direct consequence of inequality (A1), (4) holds with \( j = i + 1 \). Assuming (4) holds for \( j - 1 \), we will prove it holds for \( j \). We have

\[\begin{align*}
E[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \\
&= \sum_{\delta' = 0}^1 E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta', X(t_j) = 1] \\
&\quad \times P(X(t_{j-1}) = \delta'|X(t_i) = \delta, X(t_j) = 1) \\
&\geq \sum_{\delta' = 0}^1 E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta'] P(X(t_{j-1}) = \delta'|X(t_i) = \delta, X(t_j) = 1) \\
&= \sum_{\delta' = 0}^1 E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = \delta'] P(X(t_{j-1}) = \delta'|X(t_i) = \delta) \\
&= 26
\]
\[ + E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0] (P(X(t_{j-1}) = 0|X(t_i) = \delta, X(t_j) = 1) \\
- P(X(t_{j-1}) = 0|X(t_i) = \delta)) \\
+ E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] (P(X(t_{j-1}) = 1|X(t_i) = \delta, X(t_j) = 1) \\
- P(X(t_{j-1}) = 1|X(t_i) = \delta)) \]

\[ = E[Y(t_i)|X(t_i) = \delta] \\
+ (E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] - E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0]) \\
\cdot (P(X(t_{j-1}) = 1|X(t_i) = \delta, X(t_j) = 1) - P(X(t_{j-1}) = 1|X(t_i) = \delta)) \]

By the induction assumption

\[ E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 1] - E[Y(t_i)|X(t_i) = \delta, X(t_{j-1}) = 0] \geq 0, \]

and by straightforward calculation

\[ P(X(t_{j-1}) = 1|X(t_i) = \delta, X(t_j) = 1) - P(X(t_{j-1}) = 1|X(t_i) = \delta) \geq 0. \]

Hence

\[ E[Y(t_i)|X(t_i) = \delta, X(t_j) = 1] \geq E[Y(t_i)|X(t_i) = \delta], \]

which completes the proof of Lemma 2.

We now use Lemma 2 to prove Lemma 1.

**Proof of Lemma 1.**

\[
\text{cov}(Y(t_i), Y(t_j)) \\
= (1/2)E[(Y(t_i) - 1/2)Y(t_j)] + (1/2)E[(Y(t_j) - 1/2)Y(t_i)] \\
= (1/2)E[(Y(t_i) - 1/2)E[X(t_j)|G]] + (1/2)E[(Y(t_j) - 1/2)E[X(t_i)|G]] \\
= (1/2)E[(Y(t_i) - 1/2)X(t_j)] + (1/2)E[(Y(t_j) - 1/2)X(t_i)]
\]
However,

\[ E[(Y(t_i) - 1/2)X(t_j)] \]
\[ = \sum_{\delta=0}^{1} E[Y(t_i) - 1/2|X(t_i) = \delta, X(t_j) = 1]P(X(t_i) = \delta, X(t_j) = 1) \]
\[ \geq \sum_{\delta=0}^{1} E[Y(t_i) - 1/2|X(t_i) = \delta]P(X(t_i) = \delta, X(t_j) = 1) \]
\[ = \sum_{\delta=0}^{1} E[(Y(t_i) - 1/2)I_{X(t_i)=\delta}]P(X(t_j) = 1|X(t_i) = \delta) \]
\[ = E[(Y(t_i) - 1/2)X(t_j)]|P(X(t_j) = 1|X(t_i) = 1) - P(X(t_j) = 1|X(t_i) = 0)] \]
\[ = E[(Y(t_i) - 1/2)Y(t_j)][(1 + \exp(-\beta|t_i - t_j|))/2 - (1 - \exp(-\beta|t_i - t_j|))/2] \]
\[ = \sigma_Y^2(t_i) \exp(-\beta|t_i - t_j|), \]

where \( \beta = 2\lambda \) for grandparent-grandchild pairs and \( = 4\lambda \) for half siblings.

Similarly,

\[ E[(Y(t_j) - 1/2)X(t_i)] \geq \sigma_Y^2(t_j) \exp(-\beta|t_i - t_j|). \]

Therefore,

\[ \text{cov}(Y(t_i), Y(t_j)) \geq \exp(-\beta|t_i - t_j|)(\sigma_Y^2(t_i) + \sigma_Y^2(t_j))/2 \]
\[ \geq \exp(-\beta|t_i - t_j|)\sigma_Y(t_i)\sigma_Y(t_j). \tag{A4} \]

Since

\[ \text{cov}(Z_1(t_i), Z_1(t_j)) = N\text{cov}(X(t_i), X(t_j))/(1/4) \]
\[ = N\exp(-\beta|t_i - t_j|), \]

and

\[ \text{cov}(Z_2(t_i), Z_2(t_j)) = N\text{cov}(Y(t_i), Y(t_j))/(\sigma_Y(t_i)\sigma_Y(t_j)), \]

Lemma 1 follows from (A4).
APPENDIX B

For unilineal relative pairs other than grandparent-grandchild and half siblings, e.g., for aunt/niece pairs, cousins, or second cousins, the IBD process $X(t)$ is no longer a Markov chain. We can still use a Markov chain model; but the underlying chain is not the IBD process itself as in the grandparent/grandchild or half-sibling case. Here we use the inheritance vector $v^j(t) = (p_1, m_1, p_2, m_2, \ldots, p_n, m_n)$ (Kruglyak et al., 1996) as the underlying chain. The coordinates of $v^j(t)$ describe the outcome of the paternal and maternal meiosis giving rise to the n non-founders in the pedigree j. Specifically, $p_i = p_i^j(t) = 0$ or 1 according to whether the grandpaternal or grandmaternal allele was transmitted in the paternal meiosis giving rise to the i-th non-founder; $m_i = m_i^j(t)$ carries the same information for the corresponding maternal meiosis. So $v^j(t)$ completely specifies the inheritance pattern at each point $t$. Let $v^j = (v^j(t_1), v^j(t_2), \ldots, v^j(t_m))$. Now the likelihood of the observed genotypes for the j-th pedigree can be written as

$$P_\alpha(G^j) = \sum_{v^j} P_\alpha(G^j, v^j)$$

$$= \sum_{v^j} P_0(G^j|v^j)P_\alpha(v^j|X^j(\tau) = 1)P_\alpha(X^j(\tau) = 1)$$

$$+ \sum_{v^j} P_0(G^j|v^j)P_\alpha(v^j|X^j(\tau) = 0)P_\alpha(X^j(\tau) = 0).$$

We consider the specific alternative hypothesis described by Feingold (1993), which assigns alternative probabilities to the states of the underlying chain. She sets $P_\alpha(X^j(\tau) = 1)$ equal to $p_\alpha = p_0 + \alpha(1 - p_0)$, and for any vector $i$ describing specific meiotic outcomes, she assigns probabilities $P(v^j(\tau) = i) = a_i$ for all $i$ such that $v^j(\tau) = i \Rightarrow X^j(\tau) = 1$ and $P(v^j(\tau) = i) = b_i$ for $i$ such that $v^j(\tau) = i \Rightarrow X^j(\tau) = 0$, where $\sum a_i = p_\alpha$ and $\sum b_i = 1 - p_\alpha$. She makes the further assumption that $a_i = a$ for all $i$ and $b_i = b$ for all $i$, and obtains the important result that the distribution of
$v^j$ given $X^j(\tau)$ is independent of $\alpha$. The likelihood becomes,

$$P_\alpha(G^j) = \sum_{v^j} P_0(G^j|v^j)P_0(v^j|X^j(\tau) = 1)p_\alpha$$

$$+ \sum_{v^j} P_0(G^j|v^j)P_0(v^j|X^j(\tau) = 0)(1 - p_\alpha)$$

$$= P_0(G^j|X^j(\tau) = 1)p_\alpha + P_0(G^j|X^j(\tau) = 0)(1 - p_\alpha).$$

The subscript 0 indicates that the probability is taken under the null hypothesis of random segregation. The calculation of the likelihood proceeds exactly as in Sections 3.1 and 3.2 to yield

$$(dP_\alpha/dP_0)(G^j) = p_0^{-1}p_\alpha Y^j(\tau) + (1 - p_0)^{-1}(1 - p_\alpha)(1 - Y^j(\tau)).$$

If the trait location $\tau$ is known, the score test for the null hypothesis $\alpha = 0$ is

$$Z_2(\tau) = \sum_j (Y^j(\tau) - p_0)/(N^{1/2}\sigma Y(\tau)).$$

and the test statistic for the global search when the trait locus is unknown is

$$\max_{1 \leq i \leq M} Z_2(t_i).$$

Again we obtain a statistic similar to the one used by Feingold (1993), but the exact number of IBD is replaced by its estimator.
Table 1
Thresholds for Type I Error of 0.05 for Half Siblings

The distance between markers is $\Delta$; $m$ is the number of equally likely alleles at each marker.

<table>
<thead>
<tr>
<th>$\Delta$</th>
<th>Simulated</th>
<th>Nominal</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$m=2$ $m=5$</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>3.72 3.80</td>
<td>4.08</td>
</tr>
<tr>
<td>1</td>
<td>3.57 3.67</td>
<td>3.91</td>
</tr>
<tr>
<td>5</td>
<td>3.50 3.53</td>
<td>3.73</td>
</tr>
<tr>
<td>10</td>
<td>3.36 3.39</td>
<td>3.60</td>
</tr>
<tr>
<td>20</td>
<td></td>
<td>3.46</td>
</tr>
</tbody>
</table>
Table 2

Correlations and Correction Factors for Multipoint Analysis

The distance between markers is $\Delta$. Correlation is the correlation between adjacent markers on an unlinked chromosome. The columns headed Marker and Midmarker give the factor multiplying $\xi$ in equation (6) for the case when the trait locus is at a marker locus and half way between two marker loci, respectively.

<table>
<thead>
<tr>
<th>No. alleles</th>
<th>$\Delta$</th>
<th>Correlation</th>
<th>Marker</th>
<th>Midmarker</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>1</td>
<td>0.989</td>
<td>0.961</td>
<td>0.955</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.938</td>
<td>0.825</td>
<td>0.800</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.852</td>
<td>0.715</td>
<td>0.649</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.667</td>
<td>0.611</td>
<td>0.470</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0.979</td>
<td>0.985</td>
<td>0.974</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.900</td>
<td>0.926</td>
<td>0.875</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.792</td>
<td>0.866</td>
<td>0.761</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.592</td>
<td>0.791</td>
<td>0.582</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>0.974</td>
<td>0.991</td>
<td>0.978</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.873</td>
<td>0.955</td>
<td>0.890</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.756</td>
<td>0.915</td>
<td>0.788</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.553</td>
<td>0.864</td>
<td>0.619</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.965</td>
<td>0.997</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.838</td>
<td>0.987</td>
<td>0.903</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.701</td>
<td>0.975</td>
<td>0.813</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.485</td>
<td>0.958</td>
<td>0.658</td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>0.962</td>
<td>0.999</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.826</td>
<td>0.994</td>
<td>0.903</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.683</td>
<td>0.989</td>
<td>0.816</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.464</td>
<td>0.981</td>
<td>0.665</td>
</tr>
<tr>
<td>$\infty$</td>
<td>1</td>
<td>0.961</td>
<td>1.000</td>
<td>0.980</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.819</td>
<td>1.000</td>
<td>0.905</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>0.670</td>
<td>1.000</td>
<td>0.818</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.449</td>
<td>1.000</td>
<td>0.670</td>
</tr>
</tbody>
</table>
Figure Legends

Figure 1. Relative efficiency when other relatives are typed as a function of the number of equally likely alleles at each marker locus: (i) Intermarker distance of 20 cM ■, (ii) distance of 10 cM □, (iii) distance of 5 cM ♦, and (iv) distance of 1 cM ◊.

Figure 2. Relative efficiency when only half siblings are typed as a function of the number of equally likely alleles at each marker locus: (i) Intermarker distance of 20 cM ■, (ii) distance of 10 cM □, (iii) distance of 5 cM ♦, and (iv) distance of 1 cM ◊.

Figure 3. Approximate sample sizes when the trait locus is at a marker as a function of the intermarker distance: (i) Completely informative markers ■, (ii) Monte Carlo estimate □, (iii) Approximation ♦, (iv) Single point analysis ◊.

Figure 4. Approximate sample sizes when the trait locus is midway between markers as a function of intermarker distance: (i) Completely informative markers ■, (ii) Monte Carlo estimate □, (iii) Approximation ♦, (iv) Single point analysis ◊.
Figure 1