Nonparametric simulation-based statistics for detecting linkage in general pedigrees

Sean Davis[*,‡], Mark Schroeder[‡], Lynn R. Goldin[†], and Daniel E. Weeks[‡,§]

*University of Pittsburgh School of Medicine, Pittsburgh

‡Department of Human Genetics, University of Pittsburgh, Pittsburgh

†Clinical Neurogenetics Branch, National Institutes of Mental Health, Bethesda

§Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford

Address correspondence to:

Daniel E. Weeks, Ph.D.

The Wellcome Trust Centre	University of Pittsburgh
for Human Genetics	Department of Human Genetics
University of Oxford	Crabtree Hall, Room A310
Windmill Road	130 DeSoto Street
Oxford OX3 7BN	Pittsburgh, PA 15261
United Kingdom	USA
(+44) 1865 742 441	(1-412) 624-3066
FAX: (+44) 1865 742 196	FAX: (1-412) 624-3020
daniel.weeks@well.ox.ac.uk	dweeks@watson.hgen.pitt.edu

Running head: Simulation-based tests for linkage

Keywords: linkage analysis, nonparametric analysis, simulation methods, affected-pedigree-member method, power analysis

Summary

We present here four nonparametric statistics for linkage analysis that test whether or not pairs of affected relatives share marker alleles more often than expected. These statistics are based on simulating the null distribution of a given statistic conditional on the unaffecteds' marker genotypes. Each statistic uses a different measure of marker sharing: the SimAPM statistic uses the Affected-Pedigree-Member measure based on identity-by-state (IBS) sharing. The SimKIN measure is 1.0 for identity-by-descent (IBD) sharing, 0.0 for no IBD sharing, and the kinship coefficient when the IBD status is ambiguous. The simulation-based IBD (SimIBD) statistic uses a recursive algorithm to determine the probability of two affecteds sharing a specific allele IBD. The SimISO statistic is identical to SimIBD, except that it also measures marker similarity between unaffected pairs. We evaluated our statistics on data simulated under different two-locus disease models, comparing our results to those obtained with several other nonparametric statistics. Using IBD information produces dramatic increases in power over the SimAPM method, which uses only IBS information. The power of our best statistic in most cases meets or exceeds the power of the other nonparametric statistics. Furthermore, our statistics perform comparisons between all affected relative pairs within general pedigrees, and are not restricted to sib pairs or nuclear families.

Introduction

As researchers seek to map increasingly complex inherited disorders, the power of methods to detect the underlying genes becomes increasingly important. Further, because many of these conditions do not have a clear pattern of inheritance, much attention has been focused on nonparametric methods (Lander and Schork 1994). These nonparametric methods make no assumptions about the mode of inheritance of a given genetic trait and have been valuable in mapping susceptibility genes associated with many complex genetic diseases for which the mode of inheritance is unclear (Davies et al. 1994; Hashimoto et al. 1994).

The affected-pedigree-member (APM) method (Weeks and Lange 1988; Weeks and Lange 1992) is a nonparametric method that uses identity-by-state (IBS) status to test for linkage as indicated by increased marker similarity between affected relative pairs. (Two alleles are IBS if the alleles are identical, regardless of ancestral origin.) Because the actual null distribution for the APM statistic may be slightly skewed while the asymptotic null distribution is symmetric, the significance of the observed APM statistic with respect to the actual distribution may differ from that based on the asymptotic null distribution. This difference has been observed (Weeks and Lange 1988; Weeks and Lange 1992); use of the asymptotic distribution commonly leads to an increased false-positive rate compared to that based on the actual distribution. To combat these problems, we routinely recommend using simulation to determine the p-value based on the empirical null distribution of the APM statistic. By using a simulated null distribution, one need not make any assumptions about the distribution of the test statistic or its convergence to an asymptotic distribution. Therefore, each of our new statistics (referred to as "simulation-based statistics") relies on simulation-based techniques to determine the empirical p-value. Similar simulation techniques to determine empirical threshold values for quantitative trait mapping have recently been developed (Churchill and Doerge 1994).

Our first simulation-based statistic, called SimAPM, was the result of combining the simulation program FASTSLINK (Ott 1989; Weeks et al. 1990; Cottingham Jr. et al. 1993) with the APM method to create a new software package, which first calculates the APM statistic from the observed data in a pedigree. Then, the affecteds are untyped at the marker locus and many replicates of new marker data for the affecteds are simulated *conditional on the unaffecteds' genotypes*. An APM statistic is evaluated for each of these new replicates; each value forms a point in the empirical null distribution. The p-value is then found by determining the location of the observed statistic in the empirical null distribution. Note that SimAPM performs simulations *conditional on the unaffecteds' genotypes*, while the original APM simulation program completely ignores unaffecteds' genotypes. Kong and others have argued that, in general terms, "the conditional p-value is more appropriate statistically because it is more relevant to the observed data" (Cox and Hinkley 1974; Kong et al. 1992, p. 1419). Others [notably, Whittemore and Halpern (1994)] have noted that conditioning on the unaffecteds' genotypes has the added benefit of reducing the sensitivity of the statistics to misspecified allele frequencies. All our simulation-based statistics presented here simulate conditional on the set of all typed unaffecteds. Conditioning on some other subset of the pedigree could, however, prove to be more powerful.

After completing the SimAPM program, we realized that we had an extremely flexible tool for examining any desired measurement of marker similarity between individuals. A natural extension of the SimAPM method is to measure whether or not *affected relative pairs* share alleles *identical-by-descent* (IBD). (Two alleles are IBD if the alleles are of the same ancestral origin.) This fills a void that exists between the APM statistic, which examines *IBS* allele sharing between *affected relative pairs*, and several statistics (*i.e.*, the Haseman-Elston (H-E) and Affected-Sib-Pair (ASP) statistics) that examine *IBD* information between *affected sib pairs* (Amos et al. 1989; Blackwelder and Elston 1985; Haseman and Elston 1972; SAGE 1994). Since the APM statistic uses IBS information in lieu of IBD information (and so only uses marker information on the affected individuals), it completely ignores unaffecteds' genotypes; this information is used by many affected-sib-pair methods. However, the APM statistic does assay marker sharing between distantly related individuals in a pedigree — something that affected-sib-pair methods, by definition, do not.

We have created three new statistics that, unlike the SimAPM statistic that measures only IBS status, use genotype information from all pedigree members to estimate IBD status for affected relative pairs. The SimKIN statistic uses the kinship coefficient as the similarity measure when the IBD state is ambiguous, 1.0 when allele sharing is definitely IBD, and 0.0 when alleles are not IBD. The simulation-based IBD statistic (SimIBD) uses a recursive algorithm to determine the probability of two affecteds sharing a specific allele IBD. The SimISO statistic is identical to the SimIBD statistic, except that it also assays the marker sharing between all pairs of unaffected relatives within each pedigree. The SimIBD and

SimISO statistics provide a more precise estimate of IBD sharing than the SimKIN statistic and, therefore, give higher power in most cases.

We evaluated the power of our simulation-based statistics by examining data simulated under eight different two-locus disease models (Goldin and Weeks 1993; Martinez and Goldin 1990). We compared the power of our statistic to the power of the H-E (Haseman and Elston 1972), ASP (Blackwelder and Elston 1985; SAGE 1994), and extended relative pair analysis (ERPA) (Curtis and Sham 1994) statistics. We also investigated the false-positive rate for the H-E, ASP, SimIBD, and SimAPM statistics.

Methods

The simulated statistic

Our implementation of the simulation-based statistics involves the following general steps:

(1) Calculation and storage of the statistic on the original pedigree data; this is the *observed statistic*;

(2) Simulation of many replicates containing new marker genotypes for all typed affecteds conditional on any known marker genotypes of unaffected individuals; these new marker genotypes are simulated ignoring disease status;

(3) Calculation of the statistic based on the simulated data for each replicate, yielding the null distribution in which the marker is

completely unlinked to the disease (because disease status is ignored when simulating the marker data); and

(4) Determination of the empirical p-value based on where the observed statistic falls within the simulated null distribution.

The simulation in step (2) is done using FASTSLINK (Ott 1989; Weeks et al. 1990; Cottingham Jr. et al. 1993) to simulate genotypes *conditional* on other typing information contained in a pedigree. Our program modifies LINKAGE input files automatically to instruct FASTSLINK to conditionally simulate marker genotypes for only the affecteds. The marker genotypes of the unaffected individuals are not altered. Note that the simulations are relatively fast because we are simulating marker genotypes at a single locus (the marker locus) and not at two loci (the marker and disease).

Approximation of the IBD probabilities

Our four simulation-based statistics use the following general approach. Assume a pedigree of n members whose marker genotypes are described by $G=\{1_1, 1_2, \ldots, n_1, n_2\}$ where i_1 and i_2 are arbitrarily ordered alleles of individual i. We denote two alleles i_x and j_y that are IBD by $i_x \equiv j_y$. Thus, we define $P(i_x \equiv j_y \mid G)$ as the probability allele i_x is IBD to allele j_y given the marker genotypes. This general approach is as follows:

(1) Choose two affecteds i and j, who are not a parent-child pair.

(2) Assign a value to α_{i_x,j_y} as a measure of $P(i_x \equiv j_y \mid G)$:

A) For the SimAPM statistic, α_{i_x,j_y}=1.0 when i_x and j_y are IBS and 0.0 when i_x and j_y are not IBS (Table 1). That is, the estimated $P(i_x \equiv j_y \mid G)$ is simply the probability that i_x and j_y are IBS.

B) For the SimKIN statistic, the value of α_{i_x,j_y} is defined according to IBD status rather than IBS status (Table 1). If two alleles are IBD, we assign a value of 1.0; if they are not IBD, we assign a value of 0.0. If the IBD status is ambiguous ("Maybe IBD" in Table 1), we use the kinship coefficient Φ_{ij} as the value of the statistic. Recall that the kinship coefficient is defined as the *unconditional* probability that an allele picked at random from each affected will be shared IBD; the kinship coefficient depends solely on the pedigree structure (*i.e.*, the relationship of the two people to each other) (Jacquard 1974).

C) For the SimIBD and SimISO statistics, we set α_{i_x,j_y} to 1.0 if the alleles are IBD and to 0.0 if the alleles are not IBD (Table 1); if the IBD status is ambiguous ("Maybe IBD"), we set α_{i_x,j_y} to the value returned by the recursive algorithm described below.

(3) Compute Z_{ij}, the pairwise statistic for affecteds i and j, by the following formula where p_a= population frequency of allele i_a and p_b= population frequency of allele i_b and p_a=p_b (because i_a and i_b are IBS):

$$Z_{ij} = \sum_{a=1}^{2} \sum_{b=1}^{2} f(p_a) \;\; \alpha_{i_a, j_b}$$

where $f(p_a) = \frac{1}{\sqrt{p_a}}$ as in the original APM statistic (Weeks and Lange 1988).

(4) Sum the pairwise statistics over all pairs of affecteds (i, j) to arrive at the value of the statistic for the pedigree p:

$$Z_p = \frac{1}{\sqrt{r - 1}} \sum_{i=1}^{r} \sum_{j=i}^{r} Z_{ij}$$

where r is the number of affected individuals in the pedigree.

(5) Sum the pedigree statistics Z_p across the M pedigrees to obtain the overall similarity statistic Z for the whole data set:

$$Z = \sum_{p=1}^{M} Z_p$$

The SimISO statistic includes a similarity measure between unaffected pairs, called the unaffected pair statistic, in addition to the similarity measure between affected pairs used by our other statistics (SimIBD, SimAPM, and SimKIN statistics). The unaffected pair statistic is calculated using the same recursive algorithm, described below, that is used by SimIBD. To find the null distribution of the unaffected pair statistic, we simulated the marker genotypes of unaffected individuals conditional on marker information for the affected individuals. The affected and unaffected pair statistics were combined into a single statistic as follows:

$$Z_p = \frac{1}{\sqrt{r-1}} \sum_{i=1}^{r} \sum_{j=i}^{r} Z_{ij,\mathrm{affected}} + \frac{1}{\sqrt{s-1}} \sum_{i=1}^{s} \sum_{j=i}^{s} Z_{ij,\mathrm{unaffected}}$$

where p is the pedigree number, r is the number of affected individuals, s is the number of unaffected individuals, and $Z_{ij,\mathrm{affected}}$ and $Z_{ij,\mathrm{unaffected}}$ are the pairwise statistics for each affected and unaffected pair, respectively.

A recursive algorithm for computing $P(i_x \equiv j_y \mid G)$

Our recursive algorithm will find $P(i_x \equiv j_y \mid G)$ exactly if all ancestors of i and j are fully-typed, but approximates it otherwise. The recursion algorithm correctly handles inbreeding loops.

Amos et al. (1990) developed an algorithm based on the Elston-Stewart algorithm for computing the probability that a pair of relatives share zero, one, or two alleles IBD on pedigrees without inbreeding loops. We wanted to use the algorithm of Amos et al. (1990) to check our recursive algorithm, but were not able to since the computer program implementing their algorithm has not yet been completed. Note also that the probability that two relatives share a *specific* allele IBD is not the same as the probability they share k alleles IBD.

Our recursive algorithm is defined in terms of three boundary conditions and two recurrence rules. Before describing these, we first define some terminology: A “founder” is a person who is a “root node” of the pedigree, *i.e.,* whose parents are absent from the pedigree drawing (*eg.*, members 1, 2, and 3 in Figure 1). We assume that founders are non-inbred and unrelated to each other.

Boundary conditions 1 and 2 (below) apply to terms in the recurrence rules that represent relationships between founders. Boundary condition 3 holds for any individual in the pedigree. These three rules are given as follows:

Boundary condition 1: If X and Y are two distinct individual founders, then:

$\alpha_{X_i,Y_i} = 0$ if $X \neq Y$ (because X and Y are unrelated by assumption)

Boundary condition 2: If X_1 and X_2 are the two different alleles of a founder X, then:

$\alpha_{X_1,X_2} = 0$ (because X is non-inbred by assumption)

Boundary condition 3: If X_i is a single allele of person X, then:

$\alpha_{X_i,X_i} = 1$

The following two recurrence rules replace each allele of the child B with her parents' alleles F_1, F_2, M_1, and M_2, where F stands for father and M stands for the mother. We require that person B is not an ancestor of person A; this condition is easily met by choosing A and B in the following recurrence rules so that B is always the lower of the two individuals in the pedigree.

Recurrence rule 1: Given two individuals A and B with $A \neq B$, then

if allele A_1 is not IBS to B_1, then

$\alpha_{A_1,B_1} = 0$,

otherwise

$$\alpha_{A_1,B_1} = \alpha_{A_1,F_1}\,\omega(F_1 \equiv B_1) + \alpha_{A_1,F_2}\,\omega(F_2 \equiv B_1)$$

$$+ \alpha_{A_1,M_1}\,\omega(M_1 \equiv B_1) + \alpha_{A_1,M_2}\,\omega(M_2 \equiv B_1)$$

where $\omega(\ )$ are defined in Table 2 for fully typed parents and child. Recurrence rule 1 transforms the problem of finding the probability of allele 1 of person A being IBD to allele 1 of person B to four new problems — finding the probabilities of IBD sharing between allele 1 of person A and each allele of the parents of person B. In other words, if R represents the marker data on the pedigree:

$$\begin{aligned} P(A_1 \equiv B_1 \mid R) = \quad & P(A_1 \equiv B_1 \mid F_1 \equiv B_1, R)\, P(F_1 \equiv B_1 \mid R) \\ + \quad & P(A_1 \equiv B_1 \mid F_2 \equiv B_1, R)\, P(F_2 \equiv B_1 \mid R) \\ + \quad & P(A_1 \equiv B_1 \mid M_1 \equiv B_1, R)\, P(M_1 \equiv B_1 \mid R) \\ + \quad & P(A_1 \equiv B_1 \mid M_2 \equiv B_1, R)\, P(M_2 \equiv B_1 \mid R) \end{aligned}$$

Recurrence rule 1 follows by the appropriate definition of α and ω. For example, $\alpha_{A_1,F_1} = P(A_1 \equiv B_1 \mid F_1 \equiv B_1, R) = P(A_1 \equiv F_1 \mid R)$ and $\omega(F_1 \equiv B_1) = P(F_1 \equiv B_1 \mid R)$. By recursively applying recurrence rule 1, we can move up through the pedigree, “expanding” all the IBD sharing information of lower generations onto each higher generation until we have reached all founders.

Recurrence rule 2: If B_1 and B_2 are two alleles in the *same* person, then

if allele B_1 is not IBS to B_2, then

$$\alpha_{B_1,B_2} = 0,$$

otherwise

$$\alpha_{B_1,B_2} = \alpha_{M_1,F_1}\,\omega(M_1 \equiv B_2, F_1 \equiv B_1) + \alpha_{M_1,F_2}\,\omega(M_1 \equiv B_2, F_2 \equiv B_1)$$

$$+ \alpha_{M_2,F_1}\,\omega(M_2 \equiv B_2, F_1 \equiv B_1) + \alpha_{M_2,F_2}\,\omega(M_2 \equiv B_2, F_2 \equiv B_1)$$

where $\omega(\,)$ are given in Table 3 for fully typed parents and children. Recurrence rule 2 is a special case of recurrence rule 1. If there were no loops in the pedigree, we could simply define the boundary conditions to stop when we reached this condition, as B_1 and B_2 could not be IBD even if they were IBS. However, if there is a loop, then it is possible for $B_1 \equiv B_2$ if B_1 and B_2 are IBS. Because B_1 and B_2 are IBS and cannot be distinguished from one another, we can assume that B_2 is inherited from the mother and B_1 is inherited from the father. Recurrence rule 2 allows the recursion to continue by simultaneously expanding B_2 onto the mother and B_1 onto the father.

Note that the weighting factors are evaluated by simple allele matching arguments in the parent-child triads (Tables 2 and 3). For example, consider person 10 in Figure 1. Person 10, the child, has genotype B=(2/3); the mother has M=(2/3); the father has F=(2/3). To find the weighting vector for allele B_1=2, we would use case 6 of Table 2. Allele B_1 can come from either the mother (*i.e.*, $\omega(M_1 \equiv B_1)=0.5$) or the father (*i.e.*, $\omega(F_1 \equiv B_1)=0.5$) with equal probability because allele $B_2=3$ can be inherited from either the mother or the father. However, if the father had genotype F=(2/4) instead of (2/3), then the allele $B_1=2$ must have come from the father (*i.e.*, $\omega(F_1 \equiv B_1)=1$) because allele $B_2=3$ would be from the mother. In the latter case, we would apply case 7 from Table 2. Similar arguments apply for other cases in Table 2 and Table 3.

Untyped individuals in any path connecting a pair of affecteds introduce uncertainty as to the pair's IBD status. The traditional way (in lod score analyses) to deal with such uncertainty is to loop over all possible genotypes of the untyped individuals, adding in a partial statistic for each genotype. This summation may be accelerated by using genotype elimination algorithms (Lange 1987; O'Connell et al. 1995) to limit the summation to consider only those genotypes that are consistent with everyone's observed phenotypes. However, the number of genotypes can still increase dramatically as the proportion of untyped people in the pedigree increases. When we tried such an approach, this resulted in a very slow program, since these time-consuming summations had to be carried for every replicate of simulated data.

We now present a simple algorithm for determining the general weighting factors for recurrence rule 1 that can be used for all cases. Given specific allele B_1 in the child $B=(B_1/B_2)$ with mother $M=(M_1/M_2)$ and father $F=(F_1/F_2)$, we first determine $P(B_1 \in (M_1,M_2) \mid R)$ and $P(B_1 \in (F_1,F_2) \mid R)$, where $B_1 \in (M_1,M_2)$ means that allele B_1 in the child was inherited from the mother. These two probabilities $\{P(B_1 \in (M_1,M_2) \mid R), P(B_1 \in (F_1,F_2) \mid R)\}$ can assume only three pairs of values, {0, 1}, {1, 0}, or {0.5, 0.5} when the child and parents are fully typed. If the triad of child B and parents M and F is fully typed, these are easily determined by simple allele matching; i.e., $P(B_1 \in (M_1,M_2) \mid R) = P(B_1 \in$

$(M_1,M_2) \mid M,F,B)$ and $P(B_1 \in (F_1,F_2) \mid R) = P(B_1 \in (F_1,F_2) \mid M,F,B)$. Then, we determine $P(B_1 \equiv M_1 \mid B_1 \in (M_1,M_2),M,F,B)$ and $P(B_1 \equiv M_2 \mid B_1 \in (M_1,M_2),M,F,B)$; again, the same three pairs, {0, 1}, {1, 0}, or {0.5, 0.5}, are the only values possible. Analogous reasoning holds for the corresponding probabilities in the father. To find $\omega(M_1 \equiv B_1)$, we observe that when child and parents are fully typed,

$$\omega(M_1 \equiv B_1) = P(B_1 \equiv M_1 \mid R)$$
$$= P(B_1 \equiv M_1 \mid B_1 \in (M_1,M_2),M,F,B)\ P(B_1 \in (M_1,M_2) \mid M,F,B)$$

Using the same reasoning gives $\omega(M_2 \equiv B_1)$, $\omega(F_1 \equiv B_1)$, $\omega(F_2 \equiv B_1)$. Weighting factors for B_2 are analogous to those shown but with B_2 substituted for B_1. If the child and parents are not fully typed, $P(B_1 \in (M_1,M_2) \mid R) \neq P(B_1 \in (M_1,M_2) \mid M,F,B)$ and $P(B_1 \in (F_1,F_2) \mid R) \neq P(B_1 \in (F_1,F_2) \mid M,F,B)$. However, because calculating the left hand terms conditional on R is non-trivial, we use the right hand terms as an approximation when the triad of child and parents is not fully typed.

As a specific example for the fully typed case, assume a child B with genotype B=(2/3), mother M=(2/3), and father F=(2/3). We wish to find the weighting factors between the specific allele B_1 between the child B and the parents M and F. Then, $P(B_1 \in (M_1,M_2) \mid M, F, B) = 0.5$ and $P(B_1 \in (F_1,F_2) \mid M, F, B) = 0.5$ because in the child-parent triplet, we cannot determine the origin of either of the child's alleles by allele matching. The probabilities $P(B_1 \equiv M_1 \mid B_1 \in (M_1,M_2),M,F,B)=1.0$ and $P(B_1 \equiv M_2 \mid B_1 \in (M_1,M_2),M,F,B) = 0$. For the father, $P(B_1 \equiv F_1 \mid B_1 \in (F_1,F_2),M,F,B)=1.0$ and $P(B_1 \equiv F_2 \mid B_1 \in (F_1,F_2),M,F,B) = 0$. With these probabilities known, $\omega(M_1 \equiv B_1) = 0.5$, $\omega(M_2 \equiv B_1)= 0$, $\omega(F_1 \equiv B_1) = 0.5$,

$\omega(F_2 \equiv B_1) = 0$. Note that exactly the same logic was used to derive case 6 in Table 2.

We now present an example with an untyped parent. We use the same example as above but with the father F untyped . To find the weighting function for B_1, $P(B_1 \in (M_1,M_2) \mid M, F, B) = 0.5$ and $P(B_1 \in (F_1,F_2) \mid M, F, B) = 0.5$ as for the fully typed case, since mother M can supply allele B_1 and father F could also supply B_1 (i.e., the genotype of the father could include the allele 2). Furthermore, $P(M_1 \equiv B_1 \mid B_1 \in (M_1,M_2),M,F,B) = 1.0$ and $P(M_2 \equiv B_1 \mid B_1 \in (M_1,M_2),M,F,B) = 0$, the same as above. However, unlike in the fully typed case, we cannot tell whether B_1 is F_1 or F_2, so $P(B_1 \equiv F_1 \mid B_1 \in (F_1,F_2),M,F,B) = 0.5$ and $P(B_1 \equiv F_2 \mid B_1 \in (F_1,F_2),M,F,B) = 0.5$. Using these probabilities, the weighting functions for this case are: $\omega(M_1 \equiv B_1) = 0.5$, $\omega(M_2 \equiv B_1) = 0$, $\omega(F_1 \equiv B_1) = 0.25$, $\omega(F_2 \equiv B_1) = 0.25$.

For the weighting functions used for recursion rule 2, similar logic can be applied. As above, assume that child B has genotype $B=(B_1/B_2)$ with mother $M=(M_1/M_2)$ and father $F=(F_1/F_2)$. Because B_1 and B_2 are IBS, when using the weighting functions for recurrence rule 2, we can assume that B_1 and B_2 are ordered such that $B_1 \in (F_1,F_2)$ and $B_2 \in (M_1,M_2)$. The general weighting functions $\omega(\,)$ for recurrence rule 2 are then given by:

$$\omega(M_1 \equiv B_2, F_1 \equiv B_1) = P(M_1 \equiv B_2, F_1 \equiv B_1 \mid R, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))$$

As for the weighting functions for recurrence rule 1, if the child and both parents are fully typed, knowing M, F, and B (a subset of R) allows exact computation of the right hand side. Therefore,

$$\omega(M_1{\equiv}B_2, F_1{\equiv}B_1) = P(M_1{\equiv}B_2, F_1{\equiv}B_1 \mid M, F, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))$$
$$= P(M_1{\equiv}B_2 \mid F_1{\equiv}B_1, M, F, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))*$$
$$P(F_1{\equiv}B_1 \mid M, F, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))$$
$$= P(M_1{\equiv}B_2 \mid M, F, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2)) *$$
$$P(F_1{\equiv}B_1 \mid M, F, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))$$

Removing the $F_1{\equiv}B_1$ is valid because, to find $P(M_1{\equiv}B_2 \mid M, B, B_1 \in (F_1,F_2), B_2 \in (M_1,M_2))$, knowledge of the specific paternal genotype is not necessary since we assume that the parental origin of B_2 is maternal; therefore, $F_1{\equiv}B_1$ is not relevant. The weighting functions $\omega(M_1{\equiv}B_2, F_2{\equiv}B_1)$, $\omega(M_2{\equiv}B_2, F_1{\equiv}B_1)$, $\omega(M_2{\equiv}B_2, F_2{\equiv}B_1)$ are analogous to $\omega(M_1{\equiv}B_2, F_1{\equiv}B_1)$. As for the weighting functions presented for recursion rule 1, these weighting functions provide exact answers when the triad of child and parents is fully typed and is an approximation when full typing information is not available for every member of the triad.

As an example of the possible time savings using our approximation over using genotype elimination to supply genotypes for untyped individuals, the time required to analyze 20 families of the structure given in Figure 2 with the top two generations untyped was just over 10 minutes using our approximation. Using genotype elimination and summing over the resulting list of genotypes required over 6 hours. (Both times are system time plus user time on a Sun SPARCstation 514MP.) Furthermore, because it is conservative in most cases, it is unlikely to result in artificially inflated power of the statistics. Finally, since all individuals untyped in the observed pedigree are left untyped in the simulation used to generate the null distribution, any bias introduced in calculating observed statistic will likely be duplicated when calculating points in the null distribution.

Example applications of the recursive algorithm

We illustrate the recursive algorithm by applying it to the pedigree given in Figure 1. Individuals 10 and 7 are affected and are IBS at alleles 1 and 2, respectively. Therefore, we wish to find $\alpha_{10_1,7_2}$. We first apply recursion rule 1 to replace 10 with her parents:

$$\begin{aligned}\alpha_{10_1,7_2} &= \alpha_{7_2,10_1}\\ &= \alpha_{7_2,9_1}\,\omega(10_1\equiv 9_1) + \alpha_{7_2,9_2}\,\omega(10_1\equiv 9_2) + \alpha_{7_2,8_1}\,\omega(10_1\equiv 8_1)\\ &\quad + \alpha_{7_2,8_2}\,\omega(10_1\equiv 8_2)\\ &= \alpha_{7_2,9_1}\,(1/2) + \alpha_{7_2,9_2}\,(0) + \alpha_{7_2,8_1}\,(1/2) + \alpha_{7_2,8_2}\,(0)\\ &= 1/2\ \alpha_{7_2,9_1} + 1/2\ \alpha_{7_2,8_1}\end{aligned}$$

According to Table 2, the weighting factors for each term are: $\omega(10_1\equiv 9_1)=0$, $\omega(10_1\equiv 9_2)=1$, $\omega(10_1\equiv 8_1)=0$, and $\omega(10_1\equiv 8_2)=0$. Again, we apply recursion rule 1 to replace 7 with his parents:

$$\begin{aligned}\alpha_{10_1,7_2} &= 1/2\ [\alpha_{9_1,6_1}\,\omega(7_2\equiv 6_1) + \alpha_{9_1,6_2}\,\omega(7_2\equiv 6_2) + \alpha_{9_1,4_1}\,\omega(7_2\equiv 4_1)\\ &\quad + \alpha_{9_1,4_2}\,\omega(7_2\equiv 4_2)] + 1/2\ [\alpha_{8_1,6_1}\,\omega(7_2\equiv 6_1) + \alpha_{8_1,6_2}\,\omega(7_2\equiv 6_2)\\ &\quad + \alpha_{8_1,4_1}\,\omega(7_2\equiv 4_1) + \alpha_{8_1,4_2}\,\omega(7_2\equiv 4_2)]\\ &= 1/2\ [\alpha_{9_1,6_1}\,(0) + \alpha_{9_1,6_2}\,(1) + \alpha_{9_1,4_1}\,(0) + \alpha_{9_1,4_2}\,(0)]\\ &\quad + 1/2\ [\alpha_{8_1,6_1}\,(0) + \alpha_{8_1,6_2}\,(1) + \alpha_{8_1,4_1}\,(0) + \alpha_{8_1,4_2}\,(0)]\\ &= 1/2\ \alpha_{9_1,6_2} + 1/2\ \alpha_{8_1,6_2}\end{aligned}$$

The weighting factors for each term are: $\omega(7_2 \equiv 6_1)=0$, $\omega(7_2 \equiv 6_2)=1$, $\omega(7_2 \equiv 4_1)=0$, and $\omega(7_2 \equiv 4_2)=0$. Also, because individuals 6 and 9 are different founders, we have $\alpha_{6_2,9_1} = 0$ by boundary condition 1. We then apply recursion rule 1 to replace 8 with her parents:

$$
\begin{aligned}
\alpha_{10_1,7_2} &= 1/2\ \alpha_{6_2,8_1} \\
&= 1/2\ [\alpha_{6_2,6_1}\ \omega(8_1 \equiv 6_1) + \alpha_{6_2,6_2}\ \omega(8_1 \equiv 6_2) + \alpha_{6_2,4_1}\ \omega(8_1 \equiv 4_1) \\
&\quad + \alpha_{6_2,4_2}\ \omega(8_1 \equiv 4_2)] \\
&= 1/2\ [\alpha_{6_2,6_1}\ (0) + \alpha_{6_2,6_2}\ (1) + \alpha_{6_2,4_1}\ (0) + \alpha_{6_2,4_2}\ (0)] \\
&= 1/2\ \alpha_{6_2,6_2}
\end{aligned}
$$

Boundary condition 3 gives $\alpha_{6_2,6_2}=1$. Therefore,

$$\alpha_{10_1,7_2} = 1/2.$$

which can be inferred as follows: 7_2 is IBD to 6_2, and 10_1 has a 50% chance of being IBD to 6_2. For more complex pedigrees, it may be difficult to compute $\alpha_{i,j}$ by directly examining the pedigree.

As an example of the application of the recursion algorithm when untyped individuals are present in the path between the two affected individuals, we wish to find $\alpha_{11_1,7_2}$, the probability that the underlined allele in person 11 (see figure 1) is IBD to the underlined allele in person 7. We start by applying recursion 1 to replace individual 7 with his parents.

$$
\begin{aligned}
\alpha_{11_1,7_2} &= \alpha_{11_1,6_1}\ \omega(7_2 \equiv 6_1) + \alpha_{11_1,6_2}\ \omega(7_2 \equiv 6_2) + \alpha_{11_1,4_1}\ \omega(7_2 \equiv 4_1) \\
&\quad + \alpha_{11_1,4_2}\ \omega(7_2 \equiv 4_2) \\
&= \alpha_{11_1,6_1}\ (0) + \alpha_{11_1,6_2}\ (1) + \alpha_{11_1,4_1}\ (0) + \alpha_{11_1,4_2}\ (0)
\end{aligned}
$$

According to Table 2, the weighting factors for each term are: $\omega(7_2 \equiv 6_1)=0$, $\omega(7_2 \equiv 6_2)=1$, $\omega(7_2 \equiv 4_1)=0$, and $\omega(7_2 \equiv 4_2)=0$. Therefore, we again apply recursion rule 1 to replace 11 with her parents:

$$
\begin{aligned}
\alpha_{11_1,7_2} = \alpha_{11_1,6_2} = \alpha_{6_2,11_1} \\
&= \alpha_{6_2,3_1}\,\omega(11_1 \equiv 3_1) + \alpha_{6_2,3_2}\,\omega(11_1 \equiv 3_2) + \alpha_{6_2,5_1}\,\omega(11_1 \equiv 5_1) \\
&+ \alpha_{6_2,5_2}\,\omega(11_1 \equiv 5_2) \\
&= \alpha_{6_2,3_1}\,(0) + \alpha_{6_2,3_2}\,(0) + \alpha_{6_2,5_1}\,(1/2) + \alpha_{6_2,5_2}\,(1/2) \\
&= 1/2\ \alpha_{6_2,5_1} + 1/2\ \alpha_{6_2,5_2}
\end{aligned}
$$

By inspection, $\omega(11_1 \equiv 3_1)=0$ and $\omega(11_1 \equiv 3_2)=0$ because individual 3 cannot pass the allele 2 on to her child. The specific allele of origin from individual 5 is unknown, so $\omega(11_1 \equiv 5_1) = \omega(11_1 \equiv 5_2) = 1/2$. Though these weighting functions are not given specifically in Table 2, they are easily found by the algorithms outlined above or by inspection. The next step is to again apply recursion rule 1 to replace person 5 in the term $\alpha_{6_2,5_1}$ with his parents.

$$
\begin{aligned}
\alpha_{6_2,5_1} &= \alpha_{6_2,1_1}\,\omega(5_1 \equiv 1_1) + \alpha_{6_2,1_2}\,\omega(5_1 \equiv 1_2) + \alpha_{6_2,2_1}\,\omega(5_1 \equiv 2_1) \\
&+ \alpha_{6_2,2_2}\,\omega(5_1 \equiv 2_2) \\
&= \alpha_{6_2,1_1}\,(1/4) + \alpha_{6_2,1_2}\,(1/4) + \alpha_{6_2,2_1}\,(1/4) + \alpha_{6_2,2_2}\,(1/4)
\end{aligned}
$$

Because we cannot discern parental origin of alleles inherited from persons 1 or 2, each of the four weighting functions has the value 1/4 . Performing two more iterations of recursion rule 1 gives:

$$
\begin{aligned}
\alpha_{11_1,7_2} &= (1/16)\ \alpha_{1_1,1_1} + (1/16)\ \alpha_{1_2,1_2} + (1/16)\ \alpha_{2_1,2_1} + (1/16)\ \alpha_{2_2,2_2} \\
&= 1/4
\end{aligned}
$$

The final result is obtained by appealing four times to boundary condition 3.

Comparison of linkage methods

Data originally simulated by Goldin and Weeks (1993) using the method of Martinez and Goldin (1990) were analyzed in order to evaluate our new statistics. Martinez and Goldin simulated their data assuming a two-locus disease model with a single marker (with four equally frequent alleles) linked at θ=0.05 to the first disease locus. Models involving two dominant loci (DD), two recessive loci (RR), and a dominant and a recessive locus (DR and RD) represent the epistatic models. A model with additive penetrance (AD) was also analyzed. The parameters used in the simulation of these five epistatic models correspond to those of unipolar and bipolar affective disorders and predict a population prevalence of 7% and a recurrence risk in first degree relatives of 25%-30%. In addition, three models included heterogeneity at levels of 50% (H50), 25% (H25), and 10% (H10), respectively, of families linked to the marker. The population prevalence for these models was 2% and disease could be due to either of the two dominant loci (90% penetrant). A simulated pedigree was selected to be in the data set if 2 or more affecteds were present in each of the three sibships (pedigree structure shown in Figure 2). For each model, the simulated data consisted of 50 replicates of 20 families for a total of 1,000 families per model and 8,000 families overall. We analyzed the simulated data using the SimAPM, SimKIN, SimIBD, and SimISO statistics. For each replicate in each disease model, we generated 1,000 simulated replicates to construct the empirical null distribution. Therefore, simulation of 400,000 replicates was required (50 data sets * 8 disease models * 1,000 replicates).

To evaluate further the power of our statistics, we analyzed the same simulated data using three other popular nonparametric statistics for comparison. The Haseman-Elston (H-E) statistic is a method of detecting linkage that uses a regression approach (Amos et al. 1989; Haseman and Elston 1972). With a simple dichotomous trait, this method reduces to testing for a difference in IBD sharing between concordant and discordant pairs. The affected sib pair method (ASP) tests whether the mean proportion of IBD sharing is greater than 0.50 (Blackwelder and Elston 1985). Another nonparametric method, extended relative pair analysis (ERPA), uses the MLINK program to derive risk probabilities of affected pairs sharing 0, 1 or 2 marker alleles IBD and then compares these probabilities with those observed using a chi-squared statistic (Curtis and Sham 1994). ERPA assays IBD information between *all* affected relative pairs, whereas H-E and ASP assay IBD sharing only within a given sibship. Indeed, to analyze the simulated data using H-E and ASP, it is necessary to split the pedigree into three sibships, which are then (incorrectly) treated as if they are independent sibships.

We investigated how our statistics performed on partially typed pedigrees by generating eight new partially typed data sets per each original set. For each data set, we systematically chose people for whom we would assume that marker genotype information was missing. In one set, the grandparents was assumed to be untyped. Six sets had, in addition to two untyped grandparents, the genotypes of 1 to 6 random members of the second generation missing. In the final set, all unaffected individuals except parents and grandparents were assumed to be untyped; this set represents a minimal genotyping strategy where only affecteds and their connecting relatives are genotyped. Analyses of these eight data sets were performed using the H-E, ASP, and SimIBD statistics (another 3,200,000 simulations from 50 replicates * 8 disease models * 8 levels of incomplete marker typing * 1,000 replicates).

To test whether or not the theoretical p-values normally used for the H-E and ASP statistics are accurate, we adapted our conditional simulation-based methodology to these two statistics and analyzed the fully typed simulated data using empirical p-values instead of theoretical p-values.

Examination of false-positive rates

We examined the false-positive rates for the ASP, H-E, SimIBD, and SimAPM statistics. We simulated 1,000 replicates of 20 families each (pedigree structure given in figure 2) with the marker locus unlinked to the disease locus in all of the families. We then analyzed these data, calculating the false-positive rate as the observed proportion of replicates that were significant at $p \leq 0.05$ or $p \leq 0.01$. The expected proportion of significant statistics is equal to the p-value at which the significance is assessed. A false-positive rate greater than the expected proportion of significant statistics would be considered anti-conservative, while a false-positive rate less than the expected proportion would be considered conservative.

Results

Power of the simulation-based statistics

Comparison of the power of the simulation-based statistics, the SimAPM, SimKIN, SimIBD, and SimISO statistics, as well as of the original APM statistic (data from Goldin and Weeks 1993) show that the SimIBD and SimISO statistics provide much higher power than the other statistics (Figure 3). For most models, the SimIBD and SimISO statistics showed identical power to detect linkage. The SimIBD and SimISO statistics do particularly well under the RR and H50 models with power of 100% in both cases. The H10 model represented the greatest challenge for all statistics as would be expected since only 10% of families are linked to the disease locus. Both the APM and SimAPM statistics performed relatively poorly in comparison to the other statistics.

Power of several nonparametric statistics

Comparison of the power of the SimIBD and SimISO statistics with those of the ERPA, ASP, and H-E statistics when the entire pedigree is typed yielded the results in Figure 4. Note that the values reported for the H-E, ASP, and ERPA statistics are based on theoretical p-values. Under the AD model, SimIBD had the highest power, while the SimISO statistic outperformed other statistics under the H25 model. The H-E statistic detected linkage more often than other methods under the RD model. Under the remaining models, the power of the SimIBD, SimAPM, and H-E statistics was similar and was higher than the ASP and ERPA statistics, which performed poorly for most models. All five statistics exhibited decreasing power with increasing heterogeneity and had the lowest power for the H10 model of inheritance.

Though powers of the SimIBD, SimISO, and H-E statistics are comparable, the H-E statistic performs quite well for 5 of the 8 models. The H-E statistic might be elevated relative to the other statistics because it benefits from the sibships of size four where both unaffected and affected sibs are typed. The power of the H-E statistic decreases significantly when all unaffecteds who are not parents are untyped, resulting, in the worst case, under the H50 model, in a power difference of 84% (data not shown). However, the power of the ASP and SimIBD statistics is maintained on the same partially-typed data sets. These findings for the H-E and ASP statistics support a similar finding by Demenais and Amos (1989).

The effect of incomplete pedigree typing on the power of the ASP, H-E, and SimIBD statistics is shown in Figure 5. The H-E statistic maintains its power well relative to the other statistics as the proportion of untyped pedigree members increases. Both the ASP and H-E statistics lose power in a nearly linear fashion with increasing numbers of untyped people. SimIBD loses more power as the number of untyped people increases than either the H-E or ASP statistics do. Note that our scheme for choosing people in the pedigrees that are untyped always leaves the two third-generation sibships fully typed. As the H-E statistic seems to benefit from these large, fully typed sibships, it is not surprising that it performs so well. On extended pedigrees with small sibships, we expect that SimIBD may perform better than the other sibship-based statistics, as SimIBD considers marker sharing between all pairs of relatives within each pedigree.

Analysis of Episodic Ataxia Pedigrees

While our nonparametric methods would normally be applied mainly to complex diseases with uncertain modes of inheritance, for the purposes of comparison, we apply them here to a simple autosomal dominant trait, episodic ataxia (EA). EA is a familial disorder that presents as attacks of generalized ataxia with normal or nearly normal neurologic findings between attacks (Litt et al. 1994). We analyzed two EA pedigrees (of 29 and 63 members, pedigrees 3 and 4 in the original reference) typed for 7 markers using two point lod scores and the SimIBD, SimISO, ASP, and H-E statistics (Table 4). The order of the markers, all on the short arm of chromosome 12, is the same as that given in table 4; the distance between each marker is also given in Table 4.

According to these data, the marker S99 is the most closely linked marker to the disease (maximum LOD score=10.79, θ=0.001). The H-E statistic found linkage ($p \leq 0.05$) between all 7 markers and the disease locus; thus, a 10 cM region on one side and a 1 cM region on the other side of the S99 marker showed statistically significant evidence for linkage to the disease. The ASP statistic failed to find linkage ($p \leq 0.05$) at the CACNL1A1, KCNA5, S99, and S93 markers. Note that the regions found by the ASP are not contiguous and that the most closely linked marker, S99, and the 3 cM region surrouding it were not found to be statistically significant. The SimIBD statistic found linkage between markers S100, S99, KCNA5, and S93 and the disease locus; SimISO (adding an unaffected-unaffected pair comparison) found linkage between the same loci. With both the SimIBD and SimISO statistics, a region comprising approximately 3 cM around the S99 marker had statistically significant evidence for linkage. In general, however, for the EA pedigrees, the differences in the power between the SimISO and SimIBD methods were small.

Simulation-based H-E and ASP statistics

We adapted our conditional simulation-based methodology to the H-E and ASP statistics and then analyzed the fully-typed simulated data (data not shown). We found that the powers of the H-E and ASP statistics were very similar for the original and simulation-based statistics, though the powers tended to be slightly higher for the theoretical statistics than for the simulation-based statistics.

False-positive rates

Measurement of false-positive rates (Table 5) revealed that the proportion of replicates that were significant for the SimIBD statistic (*i.e.*, the false-positive rate) was lower than expected. The ASP, H-E, and SimAPM methods had false-positive rates that slightly exceeded the expected rates. Therefore, the SimIBD statistic was slightly conservative while the other statistics were slightly anticonservative.

Timing Results

Computation of the simulation-based statistics is more time-consuming than computation of the (non simulation-based) H-E and ASP statistics. For example, on our Sun SPARCstation 514MP, analysis of one replicate (20 families with 16 members each) took the ASP and H-E statistics about 1.1 seconds; SimIBD 4 minutes, 23 seconds; SimISO 13 minutes; and ERPA 14 minutes, 50 seconds. Calculation of the SimIBD statistic when the top two generations were untyped took just over 10 minutes; with genotype elimination on the same data took 6 hours, 22 minutes, 27 seconds. (All reported times are system time plus user time.) With larger incompletely typed pedigrees, using genotype elimination would be impractical or impossible.

Discussion

Simulation-based statistics

Comparison of the results of the SimIBD, SimISO, SimAPM, and SimKIN statistics allows some qualitative and quantitative determination of the contributions of IBS information, pedigree structure, and pedigree typing information to the power of SimIBD.

The APM and SimAPM methods use IBS information to approximate the underlying IBD information, and so the difference in power between the SimAPM and SimIBD statistics loosely represents the power lost by using IBS information instead of IBD information. In the simulated data sets, the average difference between the power of SimIBD and the SimAPM statistic is 49%, showing how crucial it is to use IBD information if it is available.

The SimKIN statistic measures the genetic relationship between two affecteds and is simply a function of the pedigree structure. Again, the power of this statistic can provide insight into the power of the SimIBD statistic because the SimIBD statistic uses a more exact measure of IBD sharing than the SimKIN statistic does. In the simulated data, the average difference between power of the SimKIN and SimIBD statistics is 30%. These results provide a measure of how well the new recursive algorithm used in SimIBD determines the probability of IBD sharing, as the SimKIN statistic ignores marker genotypes when the IBD status is ambiguous.

We also investigated genotype elimination using the method of Lange (1987) as a solution to the problem of low power at high levels of incomplete marker typing. However, we found that the inclusion of a genotype elimination algorithm in our SimIBD statistic did not increase power very much (data not shown). In addition, computing the SimIBD statistic on all possible genotype vectors at even modest levels of incomplete typing was too time-consuming.

We have presented a family of statistics that rely on simulating marker genotypes for affected individuals conditional on the set of all typed unaffected individuals. However, there are other choices of sets upon which to condition. In fact, the simulation could be done using nothing more than population frequencies of alleles and Hardy-Weinberg equilibrium; we refer to this method as unconditional simulation. We have implemented a version of SimIBD that uses the unconditional simulation method and have found that, at least for the complex modes of inheritance presented here, the conditional simulation method affords better power for both the fully- and partially-typed data sets. (See Table 6 for the fully-typed results.) Furthermore, a conditional simulation method reduces the effects of misspecification of allele frequencies. However, on different pedigree structures where there are a large number of unaffected individuals in the chain connecting two affected individuals, the unconditional simulation method is more powerful than the conditional simulation method. Conditioning the simulations on sets other than the set of all typed unaffecteds may prove more powerful; we plan to investigate which set is optimal.

Other nonparametric statistics

The SimIBD statistic attempts to determine the IBD status of all pairs of affected relatives as does ERPA, while the H-E and ASP statistics use only pairs of sibs. We might expect the power of the SimIBD statistics and ERPA (which is the most computationally intensive of the statistics) to be higher than that for either the H-E or ASP statistics. While ERPA (average power over all models, fully typed pedigrees, 69%; Figure 4) exceeds the power of the ASP statistic (average power, 67%), it falls short of the other statistics in most cases. One possible reason for

ERPA's lack of power may be that the distribution of its raw statistic may be poorly approximated by the chi-squared distribution used by ERPA (Curtis and Sham 1994). The ASP statistic had the worse power in most cases than the SimIBD, ERPA, or H-E statistics. Goldin and Weeks (1993) also found the ASP statistic to have less power than the H-E statistic. The SimIBD statistic (average power, 76%) performs better than the ERPA and ASP statistics, but does not surpass the H-E statistic (average power, 78%) when only the theoretical p-values are used to define power. When the simulation-based empirical p-values are used, the ASP statistic is again much less powerful (average power, 65%, data not shown) than other statistics but does nearly agree with the results found for the original statistic based on theoretical p-values. The results from the simulation-based H-E statistic were also consistent with the original, theoretical results.

We expected that the SimISO statistic might have less power than the SimIBD statistic if there were incomplete penetrances, since comparisons between unaffecteds involves two sub-groups of unaffecteds: a group of phenotypically unaffected (yet genetically susceptible) individuals and a group of unaffected individuals who are truly not genetically susceptible. Comparisons between a pair of unaffecteds with one from each group could weaken the evidence for linkage, as such pairs should be less marker similar than expected. However, while SimISO performed worse than SimIBD for the least penetrant model (AD) as expected, overall SimISO had the same average power over all models as SimIBD. This indicates that inclusion of the unaffected comparison may be beneficial.

Finally, the H-E statistic assays marker similarity between all sets of sib pairs including unaffected-affected pairs. Because we condition on unaffecteds' genotypes when simulating genotypes for affecteds (and vice-versa when making comparisons between unaffected pairs), we are unable to perform a similar measure of the unaffected-affected pairs with our present statistic. [Note that Ward has incorporated the unaffected-affected pair comparison into an extension of the APM method (Ward 1993).]

In terms of optimal study design, it is important to genotype only those individuals who contribute the most information. For complex traits, affected individuals contain the majority of the information regarding linkage. Our study indicates that it is important to type not only the affected individuals, but any intervening pedigree members who may aid in the determination of IBD sharing between pairs of affecteds. This was also pointed out by Risch (1990). Note that the powers of the ASP and SimIBD statistics are nearly unchanged when *non-intervening* unaffecteds are not genotyped; accordingly, it may be unnecessary to type all members of a pedigree in order to have maximum power. In contrast, the H-E statistic, according to our results, requires genotyping of both unaffected and affected sibs to maintain power.

Summary

We have presented here four new simulation-based statistics — the SimAPM, SimKIN, SimIBD, and SimISO statistics. We have shown that the incorporation of IBD information into measurement of allele sharing between affected relative pairs produces a large increase in power over the APM method (which uses only IBS information). Also, the use of a simulation-based approach has proven to be a powerful and flexible technique for defining and implementing linkage methods. The power of our new methods to detect linkage compares favorably to the power of the H-E statistic and exceeds the powers of ASP and ERPA statistics. Furthermore, our statistics perform comparisons between all affected relative pairs within general pedigrees.

Future enhancements and research

The original APM statistic is sensitive to misspecification of allele frequencies because the expected values under the null hypothesis of no linkage are derived using the population allele frequencies (Babron et al. 1993; Van Eerdewegh et al. 1993). Our new simulation-based statistics should be less sensitive to allele frequency misspecification than the APM method because their expected values are derived conditional on the observed marker genotypes of the unaffecteds. For example, if a pedigree is completely typed and all the affecteds are offspring of typed unaffecteds, then the SimIBD statistic is completely independent of the marker allele frequencies. As the pedigree becomes more untyped, the sensitivity to misspecified marker allele frequencies will increase in proportion to the percentage of untyped people. We plan to investigate the sensitivity of the SimIBD, SimISO, and SimKIN statistics to misspecified allele frequencies. Also, note that the power of the multilocus APM statistic can be significantly higher than that of the single locus statistic (Weeks and Lange 1992). We plan to extend these simulation-based statistics to allow multilocus linkage analysis.

Our simulation-based statistic could be computationally time-consuming on large extended pedigrees, so we plan to link it to our parallelized simulation program (SLINK-PVM) — written using PVM (Beguelin et al. 1992), a parallel processing library — that accelerates the computations by simultaneously using a large number of networked workstations . Even so, simulations using SLINK-

PVM may still be too slow on complex pedigree structures. For those cases, we plan to use a random walk approach for simulation (Lange and Matthysse 1989; Lange and Sobel 1991). The time required for simulation using a random walk is "essentially independent of the number of markers and the structure of the original pedigrees" (Lange and Sobel 1991, p. 20). We will also investigate the effects of conditioning on sets of individuals within the pedigree besides the set of all typed unaffecteds when performing simulations.

Program availability

Our computer program implementing the statistics described here is available by anonymous ftp from watson.hgen.pitt.edu.

Acknowledgments

We wish to thank Chris Aston for technical assistance, Carol Haynes for her LINKAGE to SAGE conversion programs, David Curtis for the ERPA software, John Mulvihill for his helpful comments, and Eric Litt for providing the episodic ataxia family data. Please note that our “SimKIN” is honored to (inadvertantly) share its name with the “SIMKIN” program written by A.W.F. Edwards in 1967 for simulation of gene flow through pedigrees. Portions of the results of this paper were obtained by using the program package, S.A.G.E., which is supported by a U.S. Public Health Service Resource Grant (1 P41 RR03655) from the Division of Research Resources. This work was supported by NIH grant HG00719, the University of Pittsburgh, the Wellcome Trust Centre for Human Genetics at the University of Oxford, the Association Française Contre Les Myopathies (AFM), and the W.M. Keck Center for Advanced Training in Computational Biology at the University of Pittsburgh, Carnegie Mellon University, and the Pittsburgh Supercomputing Center.

References

Amos CI, Dawson DV, Elston RC (1990) The probabilistic determination of identity-by-descent sharing for pairs of relatives from pedigrees. American Journal of Human Genetics 47:842-853

Amos CI, Elston RC, Wilson AF, Bailey-Wilson JE (1989) A more powerful robust sib-pair test of linkage for quantitative traits. Genet Epidemiol 6:435-449

Babron MC, Martinez M, Bonaiti-Pellie C, Clerget-Darpoux F (1993) Linkage detection by the affected-pedigree-member method: what is really being tested? Genetic Epidemiology 10:389-94

Beguelin A, Dongarra J, Geist G, Jiand W, Manchek R, Moor B, Sunderam B (1992) Parallel virtual machine system. Oak Ridge National Laboratory, Oak Ridge

Blackwelder WC, Elston RC (1985) A comparison of sib-pair linkage tests for disease susceptibility loci. Genetic Epidemiology 2:85-98

Churchill GA, Doerge RW (1994) Empirical threshold values for quantitative trait mapping. Genetics 138:963-971

Cottingham Jr. RW, Idury RM, Schaffer AA (1993) Faster sequential genetic linkage computations. Am J Hum Genet 53:252-263

Cox NJ, Hinkley DV (1974) Theoretical Statistics. Chapman & Hall, New York

Curtis D, Sham PC (1994) Using risk calculation to implement an extended relative pair analysis. Annals of Human Genetics 58:151-162

Davies JL, Kawaguchi Y, Bennett ST, Copeman JB, Cordell HJ, Pritchard LE, Reed PW, Gough SC, Jenkins SC, Palmer SM et al (1994) A genome-wide search for human type 1 diabetes susceptibility genes. Nature 371:130-136

Demenais FM, Amos CI (1989) Power of the sib pair and lod score method for linkage analysis of quantitave traits. In: Elston R, Spence A, Hodge S, MacCluer J (eds) Multipoint mapping and linkage based upon affected members: Genetic Analysis Workshop 6. Alan R. Liss, pp 201-206

Goldin LR, Weeks DE (1993) Two-locus models of disease: comparison of likelihood and nonparametric linkage methods. American Journal of Human Genetics 53:908-915

Haseman JK, Elston RC (1972) The investigation of linkage between a quantitative trait and a marker locus. Behav Genet 2:3-19

Hashimoto L, Habita C, Beressi JP, Delepine M, Besse C, CambonThomsen A, Deschamps I, Rotter JI, Djoulah S, James MR et al (1994) Genetic mapping of a susceptibility locus for insulin-dependent diabetes mellitus on chromosome 11q. Nature 371:161-164

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

Kong A, Frigge M, Irwin M, Cox N (1992) Importance Sampling. I. Computing Multimodel p Values in Linkage Analysis. American Journal of Human Genetics 51:1413-1429

Lander ES, Schork NJ (1994) Genetic dissection of complex traits. Science 265:2037-2048

Lange K (1987) An algorithm for automatic genotype elimination. American Journal of Human Genetics 40:250-256

Lange K, Matthysse S (1989) Simulation of pedigree genotypes by random walks. Am J Hum Genet 45:959-970

Lange K, Sobel E (1991) A random walk method for computing genetic location scores. Am J Hum Genet 49:1320-1334

Litt M, Kramer P, Browne D, Gancher S, Brunt ERP, Root D, Phromchotikul T, Dubay C, Nutt J (1994) A gene for episodic ataxia/myokomia maps to chromosome 12p13. Annals of Human Genetics 55:702-709

Martinez M, Goldin LR (1990) Power of the linkage test for a heterogeneous disorder due to two independent inherited causes: a simulation study. Genetic Epidemiology 7:219-320

O'Connell JR, Weeks DE (1995) The VITESSE algorithm for rapid multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 11:402-408

Ott J (1989) Computer-simulation methods in human linkage analysis. Proc Natl Acad Sci USA 86:4175-4178

Risch N (1990) Linkage Strategies for Genetically Complex Traits. III. The Effect of Marker Polymorphism on Analysis of Affected Relative Pairs. American Journal of Human Genetics 46:242-253

SAGE (1994) Statistical Analysis for Genetic Epidemiology, Release 2.2. Department of Biometry and Genetics, LSU Medical Center, New Orleans

Schroeder M, Brown DL, Weeks DE (1994) Improved programs for the affected-pedigree-member method of linkage analysis. Genetic Epidemiology 11:69-74

Van Eerdewegh P, Hampe CL, Suarez BK, Reich T (1993) Alzheimer disease: a piscatorial trek. Genet Epidemiol 10:395-400

Ward PJ (1993) Some developments on the affected-pedigree-member method of linkage analysis. American Journal of Human Genetics 52:1200-1215

Weeks DE, Lange K (1988) The affected-pedigree-member method of linkage analysis. American Journal of Human Genetics 42:315-326

Weeks DE, Lange K (1992) A multilocus extension of the affected-pedigree-member method of linkage analysis. Am J Hum Genet 50:859-868

Weeks DE, Ott J, Lathrop GM (1990) SLINK: a general simulation program for linkage analysis. American Journal of Human Genetics 47:A204

Whittemore AS, Halpern J (1994) A class of tests for linkage using affected pedigree members. Biometrics 50:118-127

Figure Legends

Figure 1. Example pedigree for illustrating recursive computation of the probability that two alleles are IBD.

Figure 2. Pedigree structure of the simulated data from Goldin and Weeks (1993).

Figure 3. Power at $p \leq 0.05$ for the SimIBD, SimISO, SimKIN, SimAPM, and APM statistics when applied to the fully typed simulated data. Refer to the text for model descriptions.

Figure 4. Power at $p \leq 0.05$ for the SimIBD, SimISO, ASP, H-E, and ERPA statistics when applied to the fully typed simulated data. Data for the H-E and ASP statistics are based on theoretical p-values.

Figure 5. Power at $p \leq 0.05$ for SimIBD, ASP, and H-E statistics when applied to simulated data with various levels of incomplete marker typing (AxB on the X-axis signifies missing marker information for A grandparents and B random members of the second generation and proceeds from 0x0 to 2x0, 2x1, 2x2, …, 2x6). Data for the H-E and ASP statistics are based on theoretical p-values.

Figure 1.

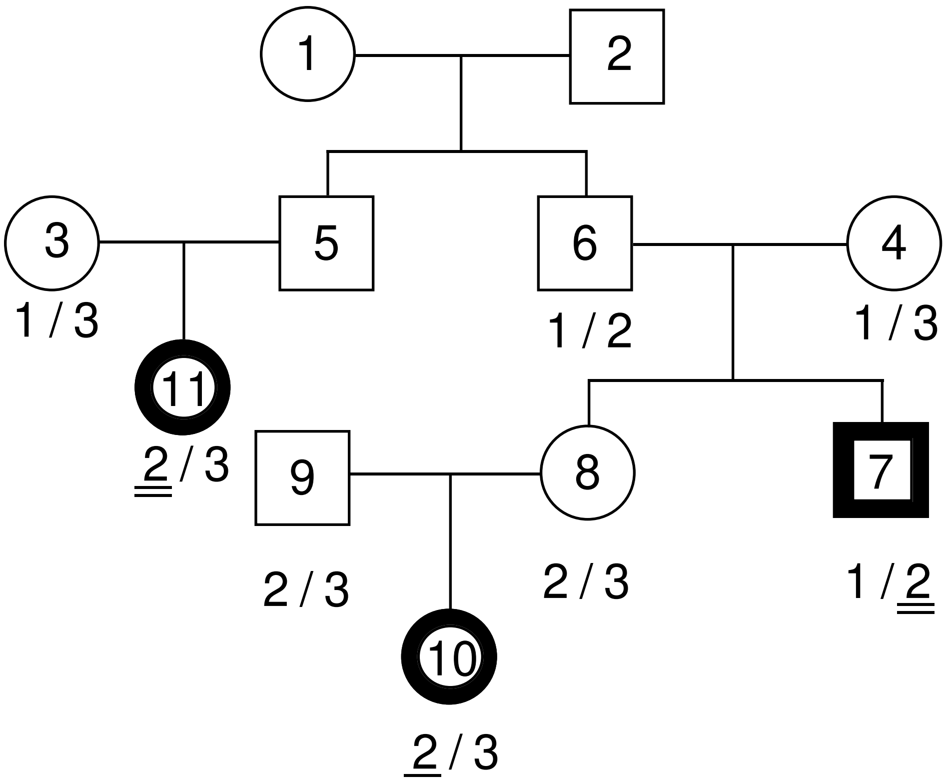


Figure 2.

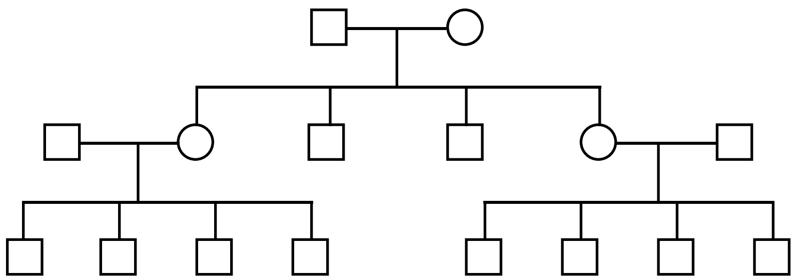


Figure 3.

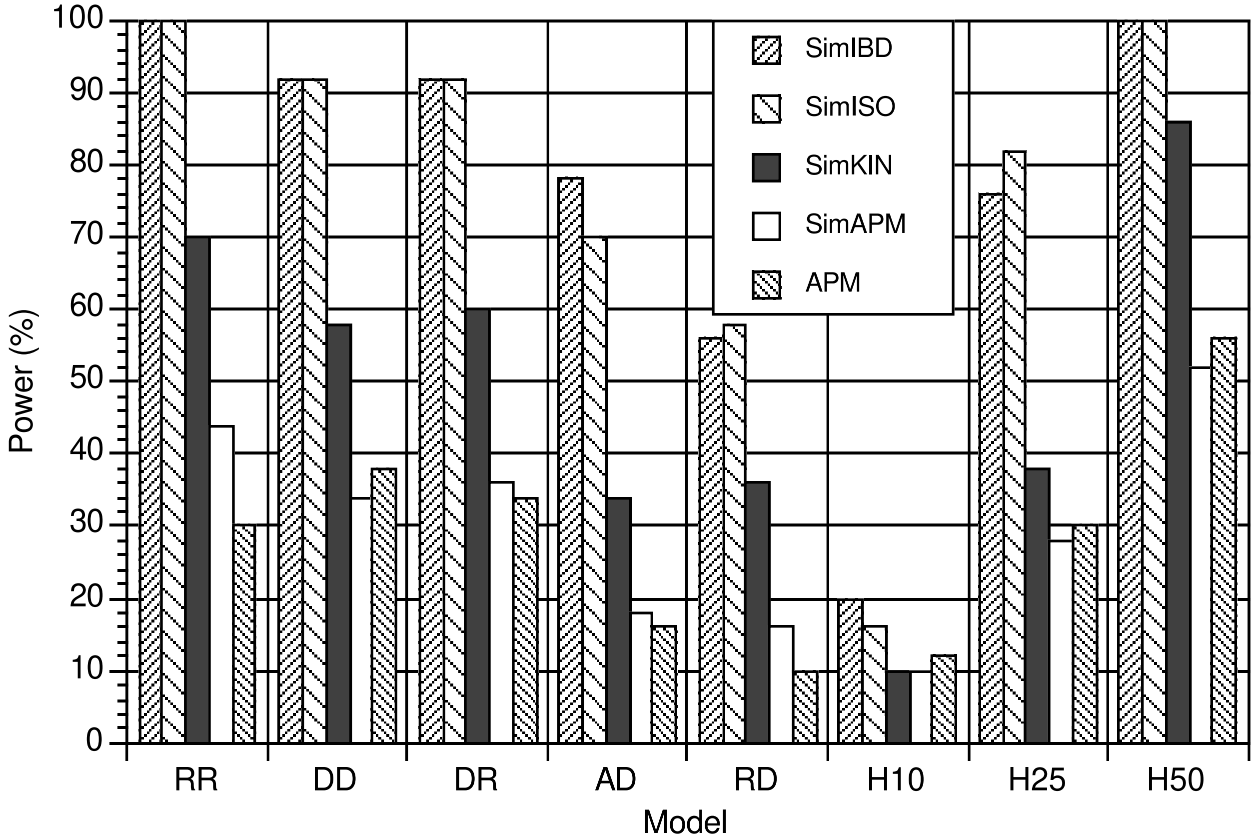
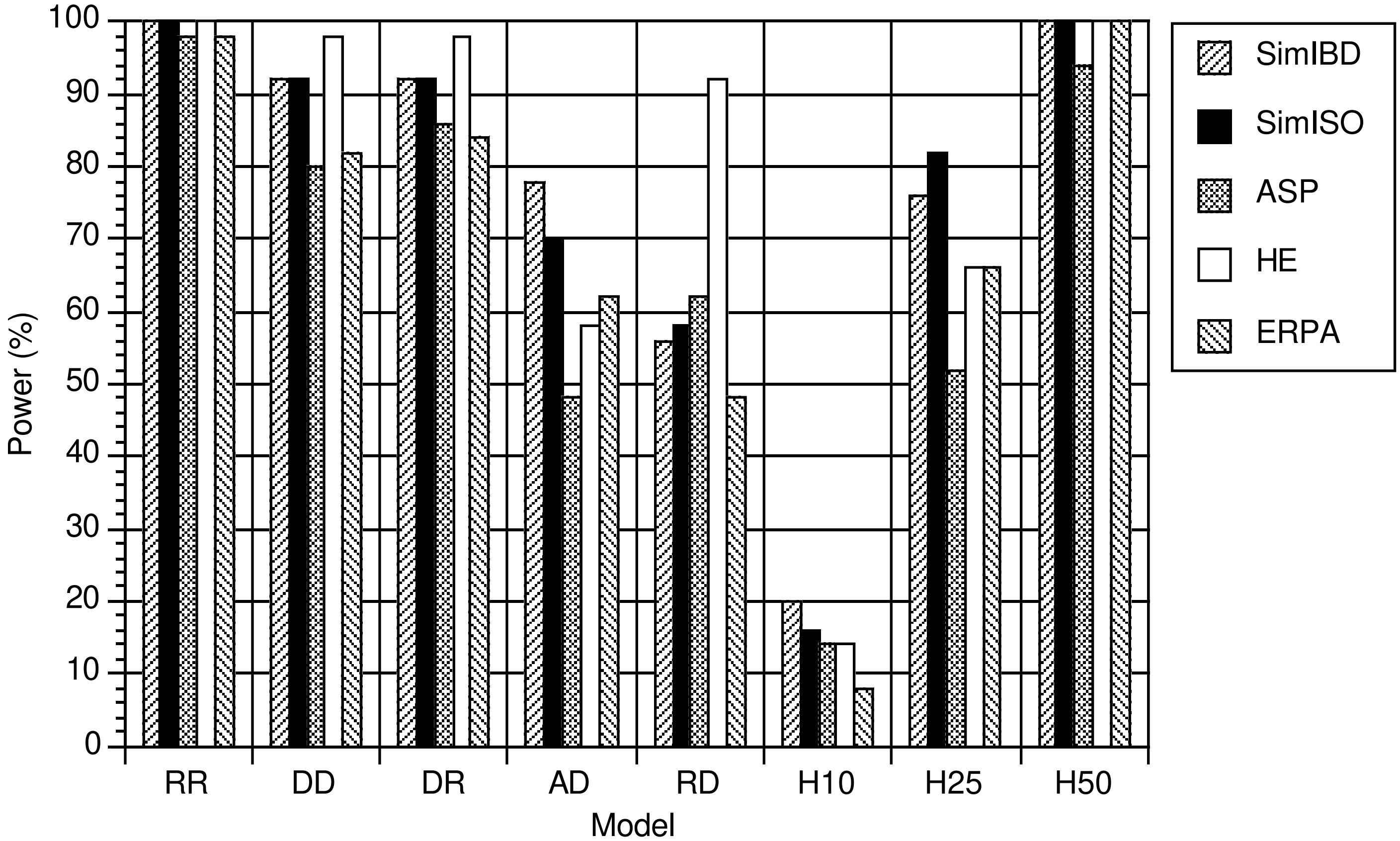

SimIBD
SimISO
SimKIN
SimAPM
APM
Power (%)
100
90
80
70
60
50
40
30
20
10
0
RR
DD
DR
AD
RD
H10
H25
H50
Model

Figure 4.



100
90
80
70
60
50
40
30
20
10
0
Power (%)
RR
DD
DR
AD
RD
H10
H25
H50
Model
SimIBD
SimISO
ASP
HE
ERPA

Figure 5.

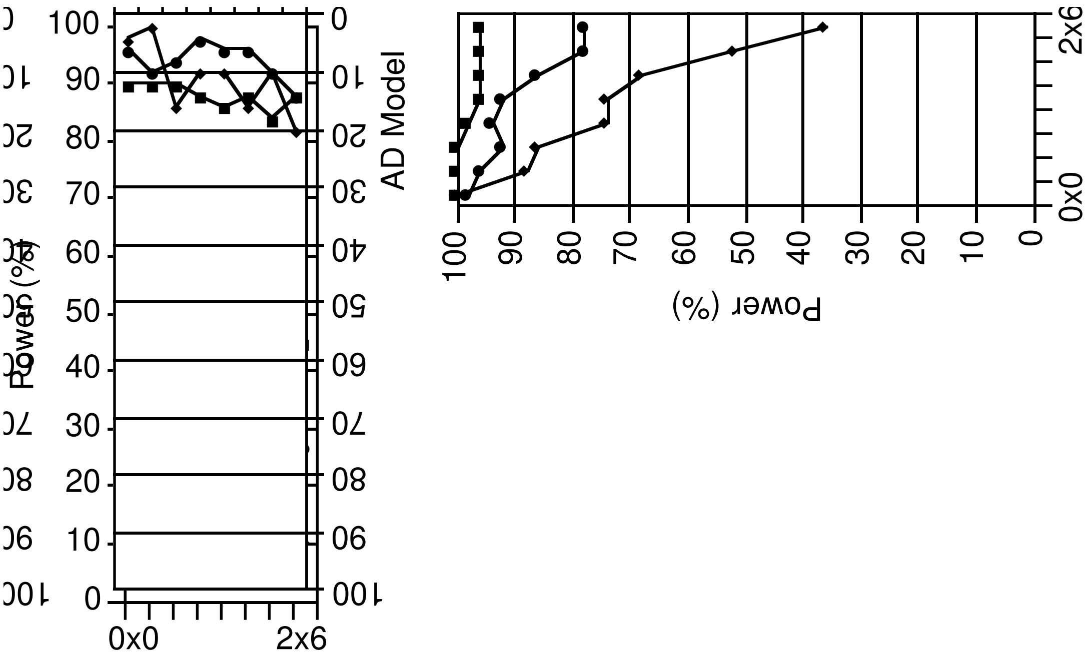


Table 1

Values of the SimAPM, SimKIN, and SimIBD similarity measures for various IBD conditions

IBD STATUS OF ALLELES I_X AND J_Y	SIMAPM VALUE	SIMKIN VALUE	SIMIBD VALUE
Not IBD	0.0 or 1.0[a]	0.0	0.0
Maybe IBD	1.0	Φ_{ij}	$P(i_x \equiv j_y \mid G)$[b]
IBD	1.0	1.0	1.0

[a]The SimAPM value is 0.0 if the pair of alleles are not IBS and 1.0 if they are IBS. A pair of alleles can be IBS but not IBD.

[b]The SimIBD value is only an approximation when incomplete marker typing is present.

Table 2

Weighting factors used in recursion rule 1 when both parents are fully typed

	CHILD TYPE ($\underline{B}_1B_2$)	MATING TYPE (F_1F_2 X M_1M_2)	VECTOR OF WEIGHTS FOR RULE 1 $\omega(F_1 \equiv B_1), \omega(F_2 \equiv B_1), \omega(M_1 \equiv B_1), \omega(M_2 \equiv B_1)$
1)	$\underline{i}i$	ii x ii	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$
2)	$\underline{i}i$	ii x ij	$\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0$
3)	$\underline{i}i$	ik x ij	$\frac{1}{2}, 0, \frac{1}{2}, 0$
4)	$\underline{i}j$	ii x jj	$\frac{1}{2}, \frac{1}{2}, 0, 0$
5)	$\underline{i}j$	ii x ij	$\frac{1}{2}, \frac{1}{2}, 0, 0$
6)	$\underline{i}j$	ij x ij	$\frac{1}{2}, 0, \frac{1}{2}, 0$
7)	$\underline{i}j$	ik x ij	1, 0, 0, 0
8)	$\underline{i}k$	ij x kl	1, 0, 0, 0

NOTE — Alleles i, j, k, and l differ in state.

Table 3

Weighting factors used in recursion rule 2 when both parents are fully typed

	CHILD TYPE (B)	MATING TYPE (F X M)	VECTOR OF WEIGHTS FOR RULE 2 $\omega(F_1 \equiv B_1, M_1 \equiv B_2), \omega(F_2 \equiv B_1, M_1 \equiv B_2), \omega(F_1 \equiv B_1, M_2 \equiv B_2), \omega(F_2 \equiv B_1, M_2 \equiv B_2)$
1)	ii	ii x ii	$\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$
2)	ii	ii x ij	$\frac{1}{2}, \frac{1}{2}, 0, 0$
3)	ii	ij x ii	$\frac{1}{2}, 0, \frac{1}{2}, 0$
4)	ii	ij x ij	1, 0, 0, 0
5)	ii	ik x ij	1, 0, 0, 0

NOTE — Alleles i, j, and k differ in state.

Table 4

Results of Analysis of Episodic Ataxia Pedigrees

Marker	Distance between Loci (cM)	Max two-point LOD score (estimated θ)	ASP	H-E	SIMIBD	SIMISO
S91		4.47 (0.142)	0.034 *	0.022 *	0.375	0.251
	0.01					
S100		3.88 (0.104)	0.029 *	0.012 *	0.000 **	0.001 **
	0.01					
CACNL1A1		2.79 (0.009)	0.253	0.002 **	0.127	0.138
	0.06					
pY21/1		8.63 (0.019)	0.007 **	0.000 **	0.064	0.091
	0.01					
KCNA5		8.38 (0.025)	0.133	0.006 **	0.000 **	0.000 **
	0.01					
S99		10.79 (0.001)	0.072	0.001 **	0.002 **	0.004 **
	0.01					
S93		6.46 (0.046)	0.053	0.000 **	0.032 *	0.023 *

NOTE — * indicates significance at $p \leq 0.05$.
** indicates significance at $p \leq 0.01$.

Table 5

False-positive rates

NAME OF STATISTIC	EXPECTED FALSE POSITIVE RATE	OBSERVED FALSE-POSITIVE RATE
SimIBD	0.05	0.046
	0.01	0.009
SimAPM	0.05	0.060
	0.01	0.015
ASP	0.05	0.058
	0.01	0.016
H-E	0.05	0.061
	0.01	0.016

Table 6

Comparison of the power of the SimIBD statistic to detect linkage using the conditional simulation approach to that using the unconditional simulation approach (on fully-typed data).

Approach	RR	DD	DR	AD	RD	H10	H25	H50
Conditional	100	92	92	78	56	20	76	100
Unconditional	100	88	88	70	56	14	70	100