A Class of Tests for Linkage Using Affected Pedigree Members

Alice S. Whittemore, Jerry Halpern

Biometrics, Volume 50, Issue 1 (Mar., 1994), 118-127.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28199403%2950%3A1%3C118%3AACOTFL%3E2.0.CO%3B2-B

BIOMETRICS 50, 118–127
March 1994

A Class of Tests for Linkage Using Affected Pedigree Members

Alice S. Whittemore and Jerry Halpern

Department of Health Research and Policy, Stanford University School of Medicine, Stanford, California 94305, U.S.A.

SUMMARY

We describe a class of nonparametric tests for linkage between a marker and a gene assumed to exist and to govern susceptibility to a disease. The tests are formed by assigning a score to each possible pattern of marker allele sharing (identity-by-descent) among affected pedigree members, and then averaging the scores over all patterns compatible with the observed marker genotype and genealogical relationship of the affected members. Different score functions give different tests. One function, which examines marker allele similarity across pairs of affected pedigree members, gives a test similar to that of Fimmers et al. (1989, in *Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop*, R. C. Elston, M. A. Spence, S. E. Hodge, and J. W. MacCluer (eds), 123–128; City: Alan R. Liss). A second function examines allele similarity across arbitrary subsets, not just pairs, of affected members. The resulting test can be more powerful than the one based solely on pairs of affected members. The approach has several advantages: it does not require knowledge of the mode of disease inheritance; it does not require unambiguous determination of identity-by-descent at the marker; it does not suffer from variability due to chance allele similarity among affected members who are unrelated, such as spouses; it allows marker genotypes of unaffected members to contribute information on allele sharing among the affected; it permits calculation of exact *P*-values. Computational requirements limit the tests to many pedigrees with few (<16) affected members.

1. Introduction

Linkage analysis is used to test the hypothesis that a genetic marker of known location is distant (e.g., on a different chromosome) from a gene presumed to exist and to govern susceptibility to a disease. Several investigators have proposed testing this hypothesis using statistics based on the similarity of marker alleles in pairs of pedigree members affected with the disease (e.g., Day and Simons, 1976; Fimmers et al., 1989, Green and Woodrow, 1977; Neugebauer, Willems, and Baur, 1984; Suarez and Hodge, 1979; Suarez and Van Eerdewegh, 1984; Weeks and Lange, 1988). The statistics do not require knowledge of the mode of disease inheritance, but may require unambiguous determination of identity-by-descent (IBD) among the affected members. Those that do not require such determination (the so-called identity-by-state methods) suffer from variability due to the distribution of marker alleles among affected members who are unrelated, such as spouses. Another limitation is the tests' restriction to pairs (as opposed to arbitrary subsets) of affected relatives. We shall show that such pairwise comparisons can have poor power when several affected pedigree members share a single marker allele IBD.

Section 2 describes a class of tests for linkage that circumvent some of these problems. The tests do not require knowledge of the mode of disease inheritance, nor do they require unambiguous determination of marker IBD relations. Furthermore, their variances are not inflated by the marker distribution in unrelated pedigree members. The tests are limited by the memory and storage needed for complex pedigrees, and thus are most useful for large numbers of simple pedigrees.

The tests are based on the expected number of marker alleles shared IBD among the affected pedigree members, given the marker alleles observed on each member. This number is then compared to the one expected given only the individuals' relationship. More explicitly, a score is assigned to each possible configuration of IBD relations among the set of marker alleles for the affected pedigree members, with high scores assigned to configurations having extensive gene similarity. The contribution to the test statistic from the affected members of that pedigree is then

Key words: Allele; Configuration probability; Genotype; Identity-by-descent; Linkage; Marker.

the difference between the expected score conditional on their observed marker genotypes and the expected score conditional only on their relationship. Thus affected relatives with unusual allele similarity tend to produce large values of the test statistic.

One test in the class evaluates marker allele similarity across arbitrary subsets, not just pairs, of affected pedigree members. We show in Section 3 that this test is more powerful than one based only on pairs of affected members when a single dominant gene governs disease susceptibility. In Section 4 we provide examples and apply the test to data on rheumatoid arthritis and the HLA system.

2. The Tests

Let $\mathcal{A}$ be an ordered set of n pedigree members having a fully specified genealogical relationship $\mathcal{R}$, all of whom are affected with the disease. We observe on each member of $\mathcal{A}$ a pair of alleles at the marker locus. We wish to use these observations to test the hypothesis that the marker is inherited independently of some assumed gene for the disease. To do so, we consider the collection of possible IBD configurations of marker alleles for the n members of $\mathcal{A}$. An IBD configuration, defined precisely in Section 3, specifies which of the members' $2n$ alleles are identical by descent (IBD), i.e., inherited from a common ancestor. We assign to each such configuration ϕ a score $S(\phi)$ that measures the degree of IBD similarity of the configuration. The score is chosen to yield good power against a specific alternative of interest.

Let $x_{i1}x_{i2}$ be the unordered pair of marker alleles observed for the ith individual, $i = 1, \ldots, n$, and call $X = (x_{11}x_{12}x_{21}x_{22} \cdots x_{n1}x_{n2})$ the marker genotype. The score function S induces a score for X:

$$T(X) = \sum_{\phi} P(\phi|X, \mathcal{R})S(\phi). \tag{1}$$

Here Σ_ϕ denotes summation over the set of all IBD configurations possible among the members of $\mathcal{A}$, and $P(\phi|X, \mathcal{R})$ is the probability of ϕ, given X and the individuals' relationship $\mathcal{R}$, under the hypothesis of no linkage. $T(X)$ is the expected marker score for the members of $\mathcal{A}$, given X and $\mathcal{R}$. The weights $P(\phi|X, \mathcal{R})$ depend on the population frequencies of the alleles in X. If the shared alleles in X are rare, then large weights are placed on those IBD configurations ϕ with a lot of IBD similarity, i.e., with high scores $S(\phi)$.

In some instances the marker genotype X completely determines the IBD configuration $\phi = \phi_X$, i.e., $P(\phi|X, \mathcal{R}) = 0$ for $\phi \neq \phi_X$. Then $T(X) = S(\phi_X)$. The literature contains several tests based on score functions $S(\phi_X)$ for sibships (e.g., deVries et al., 1976; Green and Woodrow, 1977; Suarez and Hodge, 1979; Suarez and Van Eerdewegh, 1984; Lange, 1986) and more general pedigrees (Fimmers et al., 1989).

The mean and variance of $T(X)$, under the null hypothesis of no linkage, are

$$E_{\mathcal{R}} = E[T(X)|\mathcal{R}] = \sum_{\phi} P(\phi|\mathcal{R})S(\phi) = E[S(\phi)|\mathcal{R}] \tag{2}$$

and

$$V_{\mathcal{R}} = E[T^2(X)|\mathcal{R}] - E_{\mathcal{R}}^2,$$

where

$$E[T^2(X)|\mathcal{R}] = \sum_{X} T^2(X)P(X|\mathcal{R}). \tag{3}$$

The observed data are genotypes X_{lk} for sets $\mathcal{A}_{lk}$, $k = 1, \ldots, K_l$, where K_l denotes the number of pedigrees in which the affected members have a given relationship $\mathcal{R}_l$, $l = 1, \ldots, L$. We assume that members of different sets are unrelated. The null hypothesis is that, for each relationship $\mathcal{R}_l$, $X_{l1}, \ldots, X_{lK_l}$ is a random sample from the distribution of marker genotypes of individuals with relationship $\mathcal{R}_l$. We shall consider test statistics of the form

$$\mathcal{T} = \frac{T - E(T)}{[V(T)]^{1/2}}, \tag{4}$$

where

$$T = \sum_{l=1}^{L} \sum_{k=1}^{K_l} T(X_{lk})$$

is the sum of the marker scores, and $E(T) = \Sigma_l K_l E_{\mathcal{R}_l}$ and $V(T) = \Sigma_l K_l V_{\mathcal{R}_l}$ denote expectation and variance under the null distributions of the X_{lk}, and $E_{\mathcal{R}_l}$ and $V_{\mathcal{R}_l}$ are given by (2) and (3), respectively.

Liapounov's Central Limit Theorem implies that, under the null hypothesis, $\mathcal{T}$ has a standard Gaussian distribution, asymptotically as $K_l \to \infty$, $l = 1, \ldots, L$ (Loève, 1960, p. 275). When the K_l are small and the Gaussian approximation is suspect, exact significance levels α can be calculated as

$$\alpha = \sum \prod_{l=1}^{L} \prod_{k=1}^{K_l} P(X'_{lk}|\mathcal{R}_l),$$

where the sum is taken over all genotypes $X'_{11}, \ldots, X'_{1K_1}, \ldots, X'_{L1}, \ldots, X'_{LK_L}$ such that $\Sigma_l \Sigma_k T(X'_{lk})$ exceeds the observed T.

Different scoring functions $S(\phi)$ give different tests. The resulting class of tests has several attractive properties. First, as noted above, the tests permit calculation of exact P-values when the number of pedigrees is small and the normal approximation is in doubt.

Second, the tests are unaffected by marker similarity among noninbred unrelated pedigree members. This is easily seen when $\mathcal{A}$ consists only of such individuals. Then in (1), $P(\phi|X, \mathcal{R}) = P(\phi|\mathcal{R}) = 1$ when ϕ is the configuration in which all $2n$ genes are genetically distinct, and $P(\phi|X, \mathcal{R}) = 0$ otherwise. Thus for all X, $T(X)$ equals its expected value and the contribution to the variance of $\mathcal{T}$ from these individuals is zero.

Third, including the marker genotypes of unaffected relatives in the calculation can provide additional information on the IBD configuration of the affected. This can be useful when the genotype of some affected members is partially observed or missing, and in some instances even when X is completely observed. For example, the genotype $X = (a_1, a_2, a_1, a_3)$ of two affected first cousins is consistent with both the configuration ϕ_1 in which the cousins share one gene IBD from a common grandparent, and ϕ_2 in which all four of their genes are genetically distinct. But the genotype (a_1, a_2, a_3, a_4) of their unaffected but related fathers completely determines their IBD configuration as ϕ_2. In such situations (1) becomes

$$T(X;\ Y) = \sum_{\phi} P(\phi|X,\ Y,\ \mathcal{R}^*)S(\phi), \tag{5}$$

where ϕ denotes the IBD configuration only of the affected, Y denotes the marker genotype of such unaffected relatives, and $\mathcal{R}^*$ denotes the relationship among both affected and unaffected. Here the null mean and variance of $T(X; Y)$ are based on the joint distribution of X and Y, given $\mathcal{R}^*$. Amos, Dawson, and Elston (1990) give an algorithm for computing the probabilities $P(\phi|X, Y, \mathcal{R})$.

Fourth, conditioning both $T(X)$ and its null distribution on the genotype of founders and certain unaffected pedigree members can be advantageous. For example, if all four parental alleles are known and distinct, then the genotypes of parents and affected offspring completely determine the offspring's IBD configuration, and one can condition the entire test statistic on the genotypes of the parents. This situation differs from the previous one in that the mean and variance of $T(X; Y)$ now reflect the conditional distribution of X given Y rather than the joint distribution of X and Y, as above. The relatives chosen for conditioning should be such that Y includes no information on the affected's IBD configuration ϕ that is contained in their genotype X but not in their relationship $\mathcal{R}$. This condition is satisfied if

$$P(\phi|Y,\ \mathcal{R}^*) = P(\phi|\mathcal{R}). \tag{6}$$

For example, returning to the two affected first cousins with genotype $X = (a_1, a_2, a_1, a_3)$, one would not condition the mean and variance of $T(X)$ on the genotype $Y = (a_1, a_2, a_3, a_4)$ of their related unaffected fathers, since Y replaces the information in X and the distribution of $T(X)$ conditional on Y is degenerate. In this example (6) is violated because its left side equals 1 when ϕ denotes the absence of any IBD sharing by the cousins, and zero otherwise, whereas its right side is nonzero for all three possible patterns of gene sharing among two first cousins. Requirement (6) implies that the null mean of $T(X)$ is unchanged by conditioning on Y and $\mathcal{R}^*$:

$$E[T(X;\ Y)|Y;\ \mathcal{R}^*] = \sum_X \left[\sum_\phi S(\phi)P(\phi|X,\ Y,\ \mathcal{R}^*)\right]P(X|Y,\ \mathcal{R}^*)$$

$$= \sum_\phi S(\phi)\left[\sum_X P(\phi,\ X|Y,\ \mathcal{R}^*)\right]$$

$$= \sum_\phi S(\phi)P(\phi|\mathcal{R}),$$

in agreement with (2). However, the variance changes since

$$E[T^2(X;\ Y)|Y,\ \mathcal{R}^*] = \sum_X T^2(X;\ Y) \sum_\phi P(X|\phi,\ Y,\ \mathcal{R}^*)P(\phi|\mathcal{R}) \tag{7}$$

differs in general from (3).

This conditioning on Y and $\mathcal{R}^*$ has two advantages. First, it reduces the set of possible marker genotypes X, thereby reducing the calculations needed to compute the variance of T. Second, it makes the test statistic less dependent on the population frequencies of the marker alleles, which often are known only approximately and which may vary across pedigrees arising in different populations. For example, conditional on the unaffected parental genotype $Y = (a_1, a_2, a_3, a_4)$, the expectation (7) for the squared score of the offspring genotype X reduces, after some algebra, to $\Sigma_\phi S^2(\phi)P(\phi|\mathcal{R})$, which is completely independent of the population frequencies of the marker alleles occurring in the family. The probabilities $P(\phi|\mathcal{R})$, i.e., the *condensed identity coefficients* (Karigl, 1982), can be computed as in Whittemore and Halpern (1994).

3. Scoring IBD Configurations

We now introduce two scoring functions $S(\phi)$. Although both are heuristically plausible, neither is uniformly optimal and other choices are possible. To describe them, we first define the configurations that specify IBD relations among the $2n$ alleles of an ordered set of n individuals. To do so, we construct a sequence $s = (s_{11}, s_{12}, \ldots, s_{n1}, s_{n2})$ of $2n$ integers, where s_{i1} and s_{i2} label the paternal and maternal alleles of individual i, and where two alleles get the same label if and only if they are IBD. Thus for two full sibs, $s = (1, 2, 1, 2)$ indicates that they share both their paternal and maternal alleles, whereas $s = (1, 2, 1, 3)$ indicates that they share only the paternal one. The number d of distinct integers in s is the number of genetically distinct alleles among the individuals; we take the integers to be $1, \ldots, d$. Next we identify any two sequences s and s', such as $(1, 2, 1, 2)$ and $(1, 2, 2, 1)$, that differ only in the order of maternal and paternal alleles for one or more individuals. This identification partitions the set of sequences s into equivalence classes, called IBD configurations. We denote an IBD configuration by $\phi = [s_{11}s_{12}\cdots s_{n1}s_{n2}]$, where $(s_{11}, s_{12}, \ldots, s_{n1}, s_{n2})$ is any representative of ϕ. Thus the configuration $\phi = [1212]$ represents the two equivalent sequences $(1, 2, 1, 2)$ and $(1, 2, 2, 1)$. Tables 1 and 2 show, respectively, the three IBD configurations possible among $n = 2$ noninbred individuals and the 16 IBD configurations possible among $n = 3$ noninbred individuals. See Whittemore and Halpern (1994) for detailed discussion of IBD configurations and their properties.

Table 1
The three IBD *configurations and scores for* $n = 2$ *noninbred individuals*

IBD configuration ϕ	$S(\phi)$	$P(\phi\|\mathcal{R})$[a]
[1212]	$\frac{1}{2}$	$\frac{1}{4}$
[1213]	$\frac{1}{4}$	$\frac{1}{2}$
[1234]	0	$\frac{1}{4}$

[a] $\mathcal{R}$ = Full sibs.

We begin with a scoring function S_P based only on IBD similarity among pairs of individuals in $\mathcal{A}$. For an arbitrary IBD configuration $\phi = [s_{11}s_{12}\cdots s_{n1}s_{n2}]$, let

Table 2
The 16 IBD *configurations and sources for* $n = 3$ *noninbred individuals*

IBD configuration ϕ		$S(\phi)$	$S_P(\phi)$	$P(\phi\|\mathcal{R})$[a]	$P((a_1a_2a_1a_2a_1a_2)\|\phi)$
1	[12 12 12]	2	$\frac{1}{2}$	0	—
2	[12 13 12]	$\frac{5}{4}$	$\frac{1}{3}$	0	—
3	[12 13 13]	$\frac{5}{4}$	$\frac{1}{3}$	0	—
4	[12 12 13]	$\frac{5}{4}$	$\frac{1}{3}$	$\frac{1}{8}$	$p_1p_2(p_1 + p_2)$[b]
5	[12 13 23]	$\frac{3}{4}$	$\frac{1}{4}$	0	—
6	[12 12 34]	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{8}$	$4p_1^2p_2^2$
7	[12 13 34]	$\frac{1}{2}$	$\frac{1}{6}$	$\frac{1}{8}$	$2p_1^2p_2^2$
8	[12 34 12]	$\frac{1}{2}$	$\frac{1}{6}$	0	—
9	[12 34 13]	$\frac{1}{2}$	$\frac{1}{6}$	0	—
10	[12 34 34]	$\frac{1}{2}$	$\frac{1}{6}$	0	—
11	[12 13 14]	1	$\frac{1}{4}$	$\frac{1}{8}$	$p_1p_2(p_1^2 + p_2^2)$
12	[12 13 24]	$\frac{1}{2}$	$\frac{1}{6}$	0	—
13	[12 13 45]	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{4}$	$2p_1p_2(p_1p_2^2 + p_1^2p_2)$
14	[12 34 15]	$\frac{1}{4}$	$\frac{1}{12}$	0	—
15	[12 34 35]	$\frac{1}{4}$	$\frac{1}{12}$	$\frac{1}{8}$	$2p_1p_2(p_1p_2^2 + p_1^2p_2)$
16	[12 34 56]	0	0	$\frac{1}{8}$	$8p_1^3p_2^3$

[a] For the ordered set {A, B, D} of Figure 1.
[b] p_i = Population frequency of a_i, $i = 1, 2$.

$$S_P(\phi) = \frac{2}{n(n-1)} \sum_{1 \le i < j \le n} f_{ij}(\phi), \tag{8}$$

where $f_{ij}(\phi)$ is one-fourth the number of alleles shared IBD by the pair of individuals (i, j):

$$f_{ij}(\phi) = \tfrac{1}{4}\left[\delta(s_{i1}, s_{j1}) + \delta(s_{i1}, s_{j2}) + \delta(s_{i2}, s_{j1}) + \delta(s_{i2}, s_{j2})\right],$$

and $\delta(\cdot, \cdot)$ is the Kronecker delta function: $\delta(s, s') = 1$ if $s = s'$ and $\delta(s, s') = 0$ otherwise. The test $\mathcal{T}_P$ obtained by substituting (8) into (1) and (4) is almost identical to one proposed by Fimmers et al. (1989), which uses $[n(n-1)/2]S_P$ as a scoring function.

We shall compare $\mathcal{T}_P$ to a test based on a second function S that gives high scores to IBD configurations in which several individuals share the same single allele IBD. Let $u = (u_1 \ldots, u_n)$, where u_i is either s_{i1} or s_{i2}. For each ϕ there are 2^n such vectors u. Let $h(u)$ be the number of nontrivial permutations of the n symbols $u_1, \ldots, u_n$ that leave u unchanged. We expect $h(u)$ to be large for some of the u when there is extensive identity-by-descent among the n relatives' alleles. Thus we take $S(\phi)$ to be the average value of h:

$$S(\phi) = 2^{-n} \sum_{u} h(u). \tag{9}$$

Tables 1 and 2 show scores for the configurations possible among two and three noninbred individuals, respectively. When $\mathcal{A}$ consists only of $n = 2$ individuals, $S_P(\phi) = S(\phi)$, and from (2), (9), and (8), $E(S)$ is their *kinship coefficient*, i.e., the probability that an allele chosen at random from the first person is IBD to one chosen at random from the second (Jacquard, 1973; Thompson, 1974, 1986; Karigl, 1982). When $n > 2$, however, the two functions differ, as seen for $n = 3$ in columns 2 and 3 of Table 2. $E(S_P)$ is the average kinship coefficient, where the average is taken over all $n(n-1)/2$ pairs.

In some applications, S_P may forfeit power by considering only pairs of members. Consider, for example, K pedigrees, each containing an affected parent and two affected offspring, with all four parental marker alleles known and distinct. We assume that the disease is entirely governed by a

Table 3
IBD *configurations for a noninbred parent and two noninbred offspring*

Configuration ϕ[a]	$P(\phi\|\mathcal{R}, \phi)$[b] Dominant	Recessive	$S(\phi)$	$S_P(\phi)$
[12 13 13]	$\frac{1}{2}[\theta^2 + (1-\theta)^2]$	$\frac{1}{2}[\theta^2 + (1-\theta)^2]$	$\frac{5}{4}$	$\frac{1}{3}$
[12 13 14]	$\frac{1}{2}[\theta^2 + (1-\theta)^2]$	$\theta(1-\theta)$	1	$\frac{1}{4}$
[12 13 23]	$\theta(1-\theta)$	$\frac{1}{2}[\theta^2 + (1-\theta)^2]$	$\frac{3}{4}$	$\frac{1}{4}$
[12 13 24]	$\theta(1-\theta)$	$\theta(1-\theta)$	$\frac{1}{2}$	$\frac{1}{6}$

[a] A = {parent, offspring, offspring}.
[b] $P(\phi|\mathcal{R}, \theta)$ is the probability of ϕ, given the relationship $\mathcal{R}$ of the three affected individuals and assuming a single dominant or recessive disease gene, with probability θ of meiotic recombination between marker and disease loci.

single autosomal gene acting in either a dominant or recessive mode. If dominant, we assume that the affected parent is heterozygous at the disease locus. Under these assumptions, we can get the exact power of the tests obtained by comparing $\mathcal{T}_P$ and $\mathcal{T}$ to a standard Gaussian at significance level $\alpha = .05$. Table 3 shows the four possible IBD configurations for a given pedigree, and their probabilities as functions of the recombination fraction θ between marker and disease loci. The hypothesis of no linkage is that $\theta = \frac{1}{2}$, so that all four configurations are equally likely. As seen in Table 2, S assigns a higher score to ϕ_{11} = [121314] (a configuration in which all three members of $\mathcal{A}$ share the *same* allele IBD) than to ϕ_5 = [121323]. In contrast, S_P assigns them the same score. In this way S is more sensitive than S_P to single allele sharing among three or more individuals. This difference gives $\mathcal{T}$ greater power than $\mathcal{T}_P$ when the disease is governed by a dominant gene (Figure 1A). $\mathcal{T}$ is not always more powerful than $\mathcal{T}_P$, as seen in Figure 1B, when the disease is governed by a recessive gene. Although the situation depicted in Figure 1B (K = 10 matings of a homozygous carrier to a heterozygous carrier) is unlikely to occur for recessive diseases of low frequency, it nevertheless indicates the lack of optimality for either test.

Also shown in Figure 1 is the exact power of the likelihood ratio or *lod score* test, obtained by

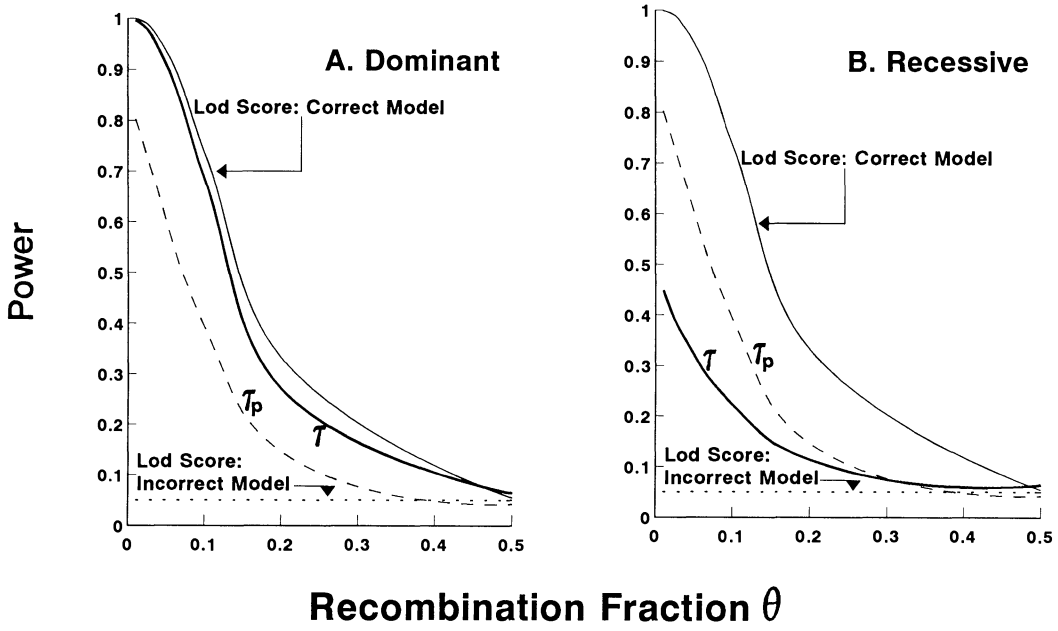

Figure 1. Power of test based on $\mathcal{T}_P$ (---) and $\mathcal{T}$ (——) to detect linkage in K = 10 pedigrees, each containing one affected parent and two affected offspring, vs probability θ of recombination between marker and disease loci. The significance level is $\alpha = .05$ and the disease is assumed to be governed by a single (**A**) dominant or (**B**) recessive gene. Also shown is the power of the likelihood ratio (lod score) test when the mode of disease inheritance is correctly (——) and incorrectly (· · ·) specified.

conditioning on the sibship size and the number of affected individuals, and by both correctly and incorrectly specifying the mode of disease inheritance (dominant vs recessive). The (correctly specified) likelihood ratio test is most powerful, because it uses information about the inheritance mode that is not used by the other tests; but it is quite sensitive to misspecification of the inheritance mode (see also Clerget-Darpoux, Bonaïti-Pellié, and Hochez, 1986).

4. Examples and Application to Data

Example 1. Consider the case of n affected full sibs with unaffected parents, and suppose that all four parental marker alleles are known and distinct. For this situation, several authors (e.g., deVries et al., 1976; Green and Woodrow, 1977; Suarez and Hodge, 1979; Suarez and Van Eerdewegh, 1984; Lange, 1986; Fimmers et al., 1989) have proposed nonparametric test statistics whose distributions are conditioned on the parental genotype Y, as described in Section 2. Green and Woodrow's test uses the "number of gene repeats" among the sibs as a score, say S_{GW}. For example, [121212] contains four repeats (two of the paternal gene and two of the maternal one), and [121213] has three repeats. For a set of sibships each with $n = 2$ affected sibs, it turns out that all three scores S_P, S, and S_{GW} are equal, and they are proportional to other scores discussed in the literature (e.g., deVries et al., 1976; Suarez and Hodge, 1979; Blackwelder and Elston, 1985; Lange, 1986; Fimmers et al., 1989). For sibships having $n = 3$ affected sibs, all three scores give equivalent tests, because each score is obtainable from the others by a linear transformation: $S_{GW}(\phi) = \frac{4}{3}[S(\phi) + 1]$ and $S_P(\phi) = \frac{1}{4.5}[S(\phi) + .25]$; and $\mathcal{T}$ is invariant under such linear transformations. For sibships each having $n > 3$ affected sibs, Suarez and Van Eerdewegh (1984) have shown that a test equivalent to that based on S_P is more powerful than Green and Woodrow's test. We have been unable to prove or disprove the equivalence of the tests based on S and S_P for sibships with $n > 3$ affected sibs.

In this example, the observed marker genotypes of parents and offspring completely determine the latter's IBD configuration ϕ. Furthermore, the null distribution of the offspring genotype, conditional on the observed parental genotype, is simply the distribution of the offspring's IBD configuration, which can be computed using the algorithms described in the preceding paper. More generally, however, IBD relationships cannot be determined unambiguously from the observed genotypes (the so-called identity-by-state relations). In such situations, to compute $T(X)$ in (1) for a given pedigree, we use Bayes' rule to write

$$P(\phi|X, \mathcal{R}) = \frac{P(\phi|\mathcal{R})P(X|\phi)}{\sum_{\phi'} P(\phi'|\mathcal{R})P(X|\phi')}, \tag{10}$$

where we have assumed that X depends on $\mathcal{R}$ only through ϕ. Substituting (10) into (1) gives

$$T(X) = \frac{\sum_{\phi} P(\phi|\mathcal{R})P(X|\phi)S(\phi)}{\sum_{\phi} P(\phi|\mathcal{R})P(X|\phi)}. \tag{11}$$

The expectation (3) needed for the variance is

$$E[T^2(X)|\mathcal{R}] = \sum_{\phi} \left[\sum_{X} T^2(X)P(X|\phi)\right] P(\phi|\mathcal{R}). \tag{12}$$

When conditioning on the genotype of pedigree members, the score $T(X; Y)$ in (5) and its conditional variance are computed using expressions analogous to (11) and (12), but with $P(X|\phi)$ replaced by $P(X|\phi, Y, \mathcal{R}^*)$. Determining the $P(X|\phi)$ or $P(X|\phi, Y, \mathcal{R}^*)$ requires assigning population frequencies to the alleles in X. The following hypothetical example illustrates these computations and shows the advantages of conditioning on the genotypes of unaffected relatives. In practice, one would not attempt to evaluate linkage with such meager data.

Example 2. Figure 2 shows a pedigree whose affected members A, B, and D are two sibs and the grandchild of one of them. Column 6 of Table 2 shows probabilities of A, B, and D's genotype $X = (a_1a_2a_1a_2a_1a_2)$, conditional on each of the seven configurations possible among them. In the absence of information on the marker genotypes of unaffected pedigree members, the test statistic depends on the population frequencies of the alleles a_1 and a_2. Assuming three alleles a_1, a_2, a_3 with frequencies $(p_1, p_2, p_3) = (.05, .75, .2)$, and substituting columns 3, 5, and 6 of Table 2 into (11), (2), and (3), we get $T(X) = .830$, $E[T(X)] = .5$, $V[T(X)] = .025$, and $\mathcal{T} = 2.14$ ($P = .02$). (The exact P-value also is $P = .02$.) The test is sensitive to the allele frequencies. If a_1 and a_2 are both common (say, $p_1 = p_2 = .5$) then $\mathcal{T} = .50$, whereas in the unlikely case that a_1 and a_2 are both very rare (say,

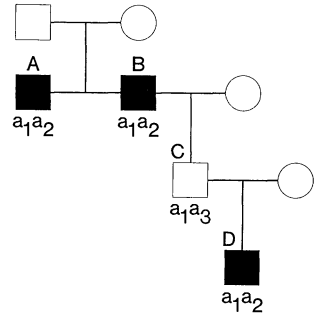

Figure 2. Pedigree with three affected members, A, B, D, indicated by filled symbols. Observed marker genotypes a_ia_j are displayed below symbols.

$p_1 = p_2 = .01$), then $\mathcal{T} = 12.36$. If, on the other hand, the grandchild's unaffected parent C has been typed as $Y = (a_1a_3)$, then grandparent and grandchild must share exactly one allele IBD, namely a_1. Notice that Y satisfies requirement (6) for conditioning, where ϕ and $\mathcal{R}$ refer to {A, B, D} and $\mathcal{R}^*$ is the relationship of {A, B, C, D}. Thus we compute the test statistic conditional on Y and $\mathcal{R}^*$. For allele frequencies (.05, .75, .2), $T(X; Y) = 1.09$, with conditional mean and variance equal to .5 and .108, respectively, giving $\mathcal{T} = 1.80$. In this example the conditional distribution of $\mathcal{T}$ also depends on the allele frequencies. Nevertheless conditioning on Y reduces the sensitivity of $\mathcal{T}$ to these frequencies: for $p_1 = p_2 = .49$, $\mathcal{T} = 1.09$, whereas for $p_1 = p_2 = .01$, $\mathcal{T} = 2.36$.

Example 3. Figure 3 shows a pedigree with multiple incidence of rheumatoid arthritis (RA), studied by Rossen et al. (1982). The pedigree is abbreviated by the deletion of those unaffected members who do not contribute to the test statistic. Rossen et al. obtained HLA haplotypes for this family. Here we test the hypothesis that an assumed RA susceptibility gene is not linked to the HLA DR locus. Figure 3 shows the HLA DR alleles of the members. The observed genotype is $X = (2, 7, 4, 7, 4, 1, 4, 3, 4, 9)$ for the set {B, D, F, G, H} of affected members. We shall condition on the genotype $Y = (2, 9, 2, 7, 3, 1, 4, 7, 2, 9)$ of the set {A, B, C, D, E} consisting of unaffected members A, C, E and affected founders B, D. Note that Y satisfies condition (6), where ϕ denotes the IBD relationship of the affected, even though Y contains genotypes of some of the affected. Given Y, X uniquely determines the affected's IBD configuration $\phi_0 = [1234353637]$, and the conditional distribution of genotype X is that of the corresponding configuration ϕ. The scores for ϕ_0 are $S(\phi_0) = 3.06$ and $S_P(\phi_0) = 1.5$, giving test statistics with exact P-values $P = .25$ and $P = .50$, respectively. For comparison, the P-values based on the Gaussian approximation (4) are $P = .15$ for $\mathcal{T}$ and $P = .35$ for $\mathcal{T}_P$.

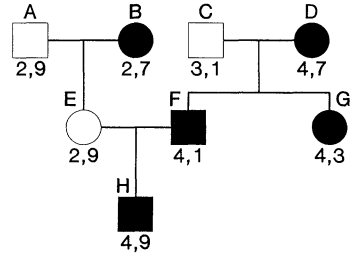

Figure 3. Pedigree with multiple incidence of rheumatoid arthritis (Rossen et al., 1982). Affected members are indicated by filled symbols. Numbers below symbols represent alleles at the HLA DR locus.

5. Discussion

The proposed class of linkage tests shares two desirable features with the "affected relative pair" test of Weeks and Lange (1988): it requires neither knowledge of the mode of disease inheritance, nor definitive determination of IBD relationships at the marker. However the two approaches differ. For a single pedigree, Weeks and Lange's test statistic $T_{WL}(X)$ is simply $S_P(X)$, where X is the observed marker genotype of the affected pedigree members. The null mean and variance of $S_P(X)$

depend on the allele frequencies $p_1, \ldots, p_m$. The mean is

$$E[S_P(\phi)|\mathcal{R}] + \{1 - E[S_P(\phi)|\mathcal{R}]\} \sum_{i=1}^{m} p_i^2,$$

where $\mathcal{R}$ is the relation of the affected members. The variance is inflated by chance marker allele similarity among unrelated affected members. Moreover, unlike the present tests, $S_P(X)$ does not allow the marker genotypes of unaffected relatives to contribute information on gene sharing among the affected.

On the other hand, there are two computational limitations to the present tests. The first is the memory and storage needed to compute the probabilities $P(\phi|\mathcal{R})$ for complex pedigrees, i.e., those with more than 16 affected members. This limitation is discussed by Fimmers et al. (1989) and Whittemore and Halpern (1994). The second is computing time needed for the variance of $T(X)$ when the marker has many alleles. Computing $\Sigma_X T^2(X) P(X|\phi)$ in (12) for a given configuration ϕ may be intractable if the number of genotypes X compatible with ϕ is large. This number is m^d, where m is the number of marker alleles, and d is the number of distinct gene labels in ϕ. We have seen that the number of possible genotypes X is reduced by conditioning on the marker genotypes of unaffected relatives. Nevertheless it may be necessary to use approximations, such as Monte Carlo sampling of genotypes (Geyer and Thompson, 1992). These limitations suggest that the test is most useful when applied to a large number of simple pedigrees. Large numbers of pedigrees also are needed when using the Gaussian approximation to the null distribution of the test statistic.

If the disease-susceptibility gene is highly penetrant, then inclusion of allele sharing among the unaffected may increase a test's power to detect linkage. In principle, such sharing could be included in the present tests by considering IBD configurations for the entire pedigree, and choosing a score function with high values for those configurations showing IBD sharing within two distinct subsets of the pedigree. However, the additional computing needed to process the extended configurations would make the method impractical for large pedigrees.

Acknowledgements

This research was supported by NIH Grants CA 47448 and GM 21215. Thanks are due to Joseph B. Keller for helpful discussions, and to the referees for comments that greatly improved an earlier version of this paper.

Résumé

Nous décrivons une classe de tests non paramétriques pour la liaison entre un marqueur et un gène supposé exister et contrôler la susceptibilité à une maladie. Ces tests sont construits en affectant un score à chaque structure possible de l'allèle marqueur partageant (''identité par descente'') entre les membres au pedigree affecté, et en prenant la moyenne des scores de toutes les structures compatibles entre les génotypes marqueurs observés et la relation généalogique des membres affectés.

Différentes fonctions de score aboutissent à différents tests. Une fonction, qui examine la ressemblance d'allèles marqueurs entre les paires des membres de pedigrees affectés, fournit un test semblable à celui de Fimmers et al. (1989, in *Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop*, R. C. Elston, M. A. Spence, S. E. Hodge, and J. W. MacCluer (eds), 123–128; City: Alan R. Liss). Une seconde fonction examine la ressemblance d'allèles entre des sous-ensembles arbitraires, pas uniquement des paires, de membres affectés. Le test qui en résulte peut être plus puissant que celui fondé uniquement sur les paires de membres affectés. L'approche présente plusieurs avantages: elle ne présuppose pas la connaissance de l'héritabilité du type de maladie; elle ne demande pas une détermination sans ambiguïté de l'identité par descente au marqueur; elle ne souffre pas de la variabilité due au hasard de la ressemblance d'allèles entre les membres affectés qui ne sont pas liés, comme les conjoints; elle permet à des génotypes marqueurs des membres non affectés de contribuer à l'information sur le partage d'allèles entre les affectés; elle permet le calcul exact des P-valeurs. Des nécessités de calcul limitent les tests à de nombreux pedigrees avec peu (<16) de membres affectés.

References

Amos, C. I., Dawson, D. V., and Elston, R. C. (1990). The probability determination of identity-by-descent sharing for pairs of relatives from pedigrees. *American Journal of Human Genetics* **47**, 842–853.

Blackwelder, W. C. and Elston, R. C. (1985). A comparison of sib pair linkage tests for disease susceptibility loci. *Genetic Epidemiology* **2,** 85–97.

Clerget-Darpoux, F., Bonaïti-Pellié, C., and Hochez, J. (1986). Effects of misspecifying genetic parameters in lod score analysis. *Biometrics* **42,** 393–400.

Day, N. E. and Simons, M. J. (1976). Disease susceptibility genes—their identification by multiple case family studies. *Tissue Antigens* **8,** 109–119.

deVries, R. R. P., Fat, R. F. M. L. A., Nijenhuis, L. E., and van Rood, J. J. (1976). HLA-linked genetic control of host response of Mycobacterium leprae. *Lancet* **ii,** 1328–1330.

Fimmers, R., Seuchter, S. A., Neugebauer, M., Knapp, M., and Baur, M. P. (1989). Identity-by-descent analysis using all genotype solutions. In *Multipoint Mapping and Linkage Based on Affected Pedigree Members: Genetic Analysis Workshop* **6,** R. C. Elston, M. A. Spence, S. E. Hodge, and J. W. MacCluer (eds), 123–128. City: Alan R. Liss.

Geyer, C. J. and Thompson, E. A. (1992). Constrained Monte Carlo maximum likelihood for dependent data (with Discussion). *Journal of the Royal Statistical Society, Series B* **00,** 000–000.

Green, J. R. and Woodrow, J. C. (1977). Sibling method for detecting HLA-linked genes in disease. *Tissue Antigens* **9,** 31–35.

Jacquard, A. (1973). *The Genetic Structure of Populations*. Berlin, Heidelberg, New York: Springer.

Karigl, G. (1982). A mathematical approach to multiple genetic relationships. *Theoretical Population Biology* **21,** 379–393.

Lange, K. (1986). A test statistic for the affected sib-set method. *Annals of Human Genetics* **50,** 283–290.

Loève, M. (1960). *Probability Theory*, 2nd edition. Princeton, New Jersey: Van Nostrand.

Neugebauer, M., Willems, J., and Baur, M. P. (1984). Analysis of multilocus pedigree data by computer. In *Ninth International Histocompatibility Workshop and Conference: Histocompatibility Testing 1984*, E. D. Albert, M. P. Baur, and W. R. Mayr (eds), 52–58. Berlin: Springer.

Rossen, R. D., Brewer, E. J., Sharp, R. M., Yunis, E. J., Schanfeld, M. S., Birdsall, H. H., Ferrell, R. E., and Templeton, J. W. (1982). Familial rheumatoid arthritis: A kindred identified through a proband with seronegative juvenile arthritis includes members with seropositive, adult-onset disease. *Human Immunology* **4,** 183–196.

Suarez, B. K. and Hodge, S. E. (1979). A simple method to detect linkage for rare recessive diseases: An application to juvenile diabetes. *Clinical Genetics* **15,** 126–136.

Suarez, B. K. and Van Eerdewegh, P. (1984). A comparison of three affected-sib-pair scoring methods to detect HLA-linked disease susceptibility genes. *American Journal of Medical Genetics* **18,** 135–146.

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

Thompson, E. A. (1986). *Pedigree Analysis in Human Genetics*. Baltimore, Maryland: Johns Hopkins University Press.

Weeks, D. E. and Lange, K. (1988). The affected pedigree member method of linkage analysis. *American Journal of Human Genetics* **42,** 315–326.

Whittemore, A. S. and Halpern, J. (1994). Probability of gene identity by descent: Computation and applications. *Biometrics* **50,** 109–117.

Received July 1991; *revised June and December* 1992; *accepted December* 1992.