# The Affected-Pedigree-Member Method of Linkage Analysis[1]

Daniel E. Weeks and Kenneth Lange

Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA

## Summary

**This paper describes a generalization of the affected-sib-pair method of linkage analysis to pedigrees. By substituting identity-by-state relations for identity-by-descent relations, we develop a test statistic for detecting departures from independent segregation of disease and marker phenotypes. The statistic is based on the marker phenotypes of affected pedigree members only. Since it is more striking for distantly affected relatives to share a rare marker allele than a common marker allele, the statistic also includes a weighting factor based on allele frequency. The distributional properties of the statistic are investigated theoretically and by simulation. Part of the theoretical treatment entails generalizing Karigl's multiple-person kinship coefficients. When the test statistic is applied to pedigree data on Huntington disease, the null hypothesis of independent segregation between the marker locus and the disease locus is firmly rejected. In this case, as expected, there is a loss of power when compared with standard lod-score analysis. However, our statistic possesses the advantage of requiring no explicit assumptions about the mode of inheritance of the disease. This point is illustrated by application of the test statistic to data on rheumatoid arthritis.**

## Introduction

The affected-sib-pair method of linkage analysis is designed to detect departures from independent segregation of disease and marker phenotypes (Penrose 1935; Haseman and Elston 1972; Day and Simons 1976; de Vries et al. 1976; Green and Woodrow 1977; Thomson and Bodmer 1977; Fishman et al. 1978; Suarez et al. 1978; Suarez and Hodge 1979). Although it has the advantage of requiring no a priori information about the mode of inheritance of the disease, the method is limited to sibship pairs and usually relies on unambiguous determination of sib identity-by-descent (ibd) relations at the marker locus. Both of these restrictions can be relaxed. We have recently extended the method to multiple affected sibs and have shown how substituting identity-by-state relations circumvents the problem of unambiguous determination of sib-pair ibd relations (Lange 1986*a*, 1986*b*).

In the present paper we generalize the affected-sib-pair method to pedigrees. Because the issue of determining ibd relations between distantly affected relatives is now exacerbated, we retain the substitution of identity-by-state relations. We also modify our previously defined affected-sib-set statistic in two further ways (Lange 1986*b*). The first of these modifications permits computation of the theoretical mean and variance of our new test statistic for each pedigree by taking advantage of the theory of multiple-person ibd relations (Thompson 1974; Karigl 1981, 1982). (Although we can avoid ibd in practice, we cannot in theory.) Our second change is motivated by the simple observation that it is more striking for distantly affected relatives to share a rare marker allele than a common marker allele. Thus we introduce a weighting factor based on allele frequency.

With these modifications we investigate theoretically and by simulation the distributional properties of our proposed new linkage test. The section immediately following defines the test statistic for a single pedigree. To compute its theoretical mean and

Received June 8, 1987; revision received August 24, 1987.

Address for correspondence and reprints: Daniel Weeks, Department of Biomathematics, UCLA School of Medicine, Los Angeles, CA 90024-1766.

[1] Our programs for calculating the test statistic are written in the recursive language PASCAL. The simulation program does not require recursion and is written in FORTRAN. Send Daniel E. Weeks an IBM-compatible floppy disk if you would like copies of the programs, a sample data set, and documentation.

0002-9297/88/4202-0014$02.00

variance then requires a detour into the combinatorics of generalized kinship coefficients. Once this matter is dispatched, we show how to combine test statistics from different pedigrees by means of the Central Limit Theorem. Finally, we apply the test to pedigree data on rheumatoid arthritis and Huntington disease.

## Definition of the Test Statistic for a Single Pedigree

We first define the test statistic for a fixed pedigree. Only those pedigree members who are both affected and typed at the marker locus enter in the definition. Consider two affected individuals numbered $i$ and $j$ and define a random variable $Z_{ij}$ to measure the marker similarity between $i$ and $j$. For the sake of simplicity, we require the various marker alleles to be codominant. Now let $i$ have maternal marker allele $G_{ix}$ and paternal marker allele $G_{iy}$. Likewise, let $j$ have maternal marker allele $G_{jx}$ and paternal marker allele $G_{jy}$. A possible definition of $Z_{ij}$ is

$$\begin{aligned} Z_{ij} = {} & \tfrac{1}{4}\,\delta(G_{ix},G_{jx}) + \tfrac{1}{4}\delta(G_{ix},G_{jy}) \\ & + \tfrac{1}{4}\,\delta(G_{iy},G_{jx}) + \tfrac{1}{4}\delta(G_{iy},G_{jy})\,, \end{aligned} \tag{1}$$

where the Kronecker delta is defined as

$$\delta(G,G') = \begin{cases} 1 & G \text{ and } G' \text{ match in state} \\ 0 & G \text{ and } G' \text{ do not match in state.} \end{cases} \tag{2}$$

A pair $(G,G')$ need not be ibd—i.e., derive from the same ancestral gene—to contribute to $Z_{ij}$. Notice also that definition (1) corresponds to comparing one gene drawn at random from $i$ to one gene drawn at random from $j$. $Z_{ij}$ is the expected probability of a match in state conditional on the observed marker genotypes of $i$ and $j$. Table 1 gives the possible values of $Z_{ij}$.

We now generalize the definition (1) to take into account marker-allele frequencies. Suppose there are $n$ codominant alleles with population frequencies $p_1, \ldots, p_n$. Let $f(p)$ be some function of these frequencies. Then our final definition is

$$\begin{aligned} Z_{ij} = {} & \tfrac{1}{4}\,\delta(G_{ix},G_{jx})f(p_{G_{ix}}) + \tfrac{1}{4}\delta(G_{ix},G_{jy})f(p_{G_{ix}}) \\ & + \tfrac{1}{4}\,\delta(G_{iy},G_{jx})f(p_{G_{iy}}) + \tfrac{1}{4}\delta(G_{iy},G_{jy})f(p_{G_{iy}})\,. \end{aligned} \tag{3}$$

Again $Z_{ij}$ can be interpreted as a conditional expectation. However, the number of matches is no longer simply counted. Each match is weighted by the function $f(p)$ of the frequency of the allele involved. In the usual notation for conditional expectations,

**Table 1**

**Possible Values of the Similarity Statistics $Z_{ij}$ for a Pair of Affecteds**

| Marker Genotype[a] | | |
|---|---|---|
| $i$ | $j$ | $Z_{ij}$ |
| a/a | a/a | 1 |
| a/b | a/a | ½ |
| a/b | a/b | ½ |
| a/b | a/c | ¼ |
| a/b | c/d | 0 |

[a] a, b, c, and d represent different marker alleles.

$$Z_{ij} = E\left[\delta(G_i,G_j)f(p_{G_i}) \,\middle|\, \begin{matrix} \text{observed marker genotypes} \\ \text{of } i \text{ and } j \end{matrix}\right],$$

where $G_i$ and $G_j$ are randomly selected marker genes from $i$ and $j$, respectively. Typical choices for $f(p)$ might be $f(p) = 1$, $f(p) = 1/\sqrt{p}$, and $f(p) = 1/p$; $f(p) = 1$ corresponds to our provisional definition (1).

From the various pair statistics $Z_{ij}$ we form an overall statistic

$$Z = \sum_{i<j} Z_{ij} \tag{4}$$

for the pedigree, where $i < j$ ranges over the set of individuals who are both affected and typed at the marker locus. If there are $r$ such people, the expression (4) for $Z$ has $\binom{r}{2} = [r(r-1)]/2$ terms. As a sample calculation of $Z$ consider the pedigree in figure 1, in which there are three affected people who are typed at the second marker locus. If we take $f(p) = 1/p$, $p_a = ½$, and $p_b = ¼$, the definition (3) yields $Z = Z_{67} + Z_{68} + Z_{78} = \{½[(½)^{-1}]\} + (\{¼[(½)^{-1}]\} + \{¼[(¼)^{-1}]\}) + \{½[(½)^{-1}]\} = 3½$.

Note that $Z$ will often be larger for a recessive disease than for a dominant disease. Indeed, for a rare recessive disease closely linked to a marker, many presenting pedigrees will be inbred, and most affected individuals will have the same identical homozygous genotype. This will lead to most $Z_{ij}$ attaining their maximum value of 1 when $f(p) = 1$.

## Mean and Variance of Z

The mean of $Z_{ij}$ is a simple function of the kinship coefficient $\Phi_{ij}$ of $i$ and $j$ (Crow and Kimura 1970; Jacquard 1974). Recall that $\Phi_{ij}$ is the probability that

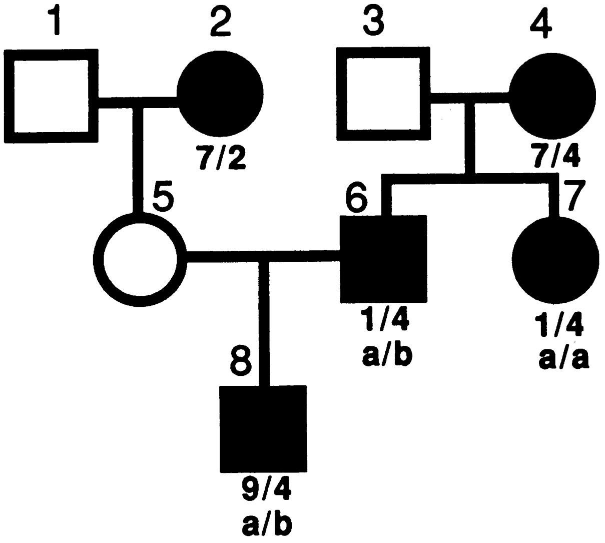

**Figure 1** Sample pedigree with two markers. Affected individuals are denoted by dark squares or circles.

a gene taken at random from $i$ is ibd with a gene at the same locus taken at random from $j$. The mean $E(Z_{ij})$ is easily computed by conditioning on whether the two genes drawn at random from $i$ and $j$ are ibd. If they are, then the two genes necessarily match and they coincide with allele $k$ with probability $p_k$. If they are not ibd, then they are simultaneously allele $k$ with probability $p_k^2$. These observations lead to the conclusion

$$E(Z_{ij}) = \Phi_{ij} \sum_{k=1}^{n} p_k f(p_k) + (1 - \Phi_{ij}) \sum_{k=1}^{n} p_k^2 f(p_k).$$

The overall mean $E(Z)$ can be recovered via

$$E(Z) = \sum_{i<j} E(Z_{ij}) .$$

Finding the variance $\mathrm{Var}(Z) = E(Z^2) - E(Z)^2$ is not so straightforward. Obviously,

$$E(Z^2) = \sum_{\substack{i<j \\ k<l}} E(Z_{ij}Z_{kl}) . \qquad (5)$$

To calculate a typical second moment $E(Z_{ij}Z_{kl})$ in equation (5), we need to extend Karigl's (1982) concept of generalized kinship coefficients. The idea behind our extension is to consider a finite sequence of people with possible repetition of some of them. For each person in this sequence draw at random either his maternally or paternally derived gene at a given locus. If an individual is repeated in the sequence, then sampling is always done with replacement. Once drawn, these genes can be uniquely partitioned into equivalence classes or blocks by assigning two genes to the same block if and only if they are ibd. Our generalized kinship coefficients give the probabilities of the various partitions. Clearly, a generalized kinship coefficient depends neither on the order of its blocks nor on the order of the individuals within a block.

As a concrete illustration of a generalized kinship coefficient, take four individuals $i$, $j$, $k$, and $l$ as in equation (5) and call the genes randomly selected from them $G_i$, $G_j$, $G_k$, and $G_l$, respectively. One possible partition is $(G_i,G_k)$, $(G_j,G_l)$; the corresponding generalized kinship coefficient, $\Phi[(G_i,G_k), (G_j,G_l)]$, is the probability that the four genes drawn fall into two blocks with $G_i = G_k \neq G_j = G_l$, where the equals sign symbolizes ibd. This particular partition is one of 15 different partitions possible for four genes. These partitions are analogous to the 15 detailed identity modes of Jacquard (1974) for two relatives. However, we are here dealing with four randomly drawn genes from four individuals rather than with all four genes from two individuals.

Employing generalized kinship coefficients, one can compute $E(Z_{ij}Z_{kl})$ by conditioning on the possible partitions for the four genes $G_i$, $G_j$, $G_k$, and $G_l$. In fact, because of the obvious conditional independence,

$$\begin{aligned} Z_{ij}Z_{kl} &= E\left[\delta(G_i,G_j)f(p_{G_i}) \,\middle|\, \begin{matrix}\text{observed marker genotypes} \\ \text{of } i \text{ and } j\end{matrix}\right] \\ &\quad \times E\left[\delta(G_k,G_l)f(p_{G_k}) \,\middle|\, \begin{matrix}\text{observed marker} \\ \text{genotypes} \\ \text{of } k \text{ and } l\end{matrix}\right] \\ &= E\left[\delta(G_i,G_j)\delta(G_k,G_l)f(p_{G_i})f(p_{G_k}) \,\middle|\, \begin{matrix}\text{observed marker} \\ \text{genotypes} \\ \text{of } i,\ j,\ k, \text{ and } l\end{matrix}\right], \end{aligned}$$

with the Kronecker deltas above defined as in equation (2) on the basis of identity by state. Taking ordinary expectations now yields the result

$$E(Z_{ij}Z_{kl}) = E[\delta(G_i,G_j)\delta(G_k,G_l)f(p_{G_i})f(p_{G_k})] .$$

If $\sigma$ symbolizes a typical ibd partition of the four genes $G_i$, $G_j$, $G_k$, and $G_l$, then conditioning on $\sigma$ gives

$$
\begin{aligned}
E(Z_{ij}Z_{kl}) = \sum_{\sigma} E[\delta(G_i,G_j)\delta(G_k,G_l)f(p_{G_i})f(p_{G_k})|\sigma]\Phi(\sigma) \\
= [\Sigma p_m f(p_m)^2]\Phi[(G_i,G_j,G_k,G_l)] \\
+ [\Sigma p_m f(p_m)]^2\Phi[(G_i,G_j)(G_k,G_l)] \\
+ [\Sigma p_m^2 f(p_m)^2] \begin{Bmatrix} \Phi[(G_i,G_j,G_k)(G_l)] \\ +\Phi[(G_i,G_j,G_l)(G_k)] \\ +\Phi[(G_i,G_k,G_l)(G_j)] \\ +\Phi[(G_j,G_k,G_l)(G_i)] \\ +\Phi[(G_i,G_l)(G_j,G_k)] \\ +\Phi[(G_i,G_k)(G_j,G_l)] \end{Bmatrix} \\
+ \{[\Sigma p_m^2 f(p_m)][\Sigma p_m f(p_m)]\} \begin{Bmatrix} \Phi[(G_i,G_j)(G_k)(G_l)] \\ +\Phi[(G_i)(G_j)(G_k,G_l)] \end{Bmatrix} \\
+ [\Sigma p_m^3 f(p_m)^2] \begin{Bmatrix} \Phi[(G_i,G_k)(G_j)(G_l)] \\ +\Phi[(G_i,G_l)(G_j)(G_k)] \\ +\Phi[(G_j,G_l)(G_i)(G_k)] \\ +\Phi[(G_j,G_k)(G_i)(G_l)] \end{Bmatrix} \\
+ [\Sigma p_m^2 f(p_m)]^2\,\Phi[(G_i)(G_j)(G_k)(G_l)] \,.
\end{aligned}
\tag{6}
$$

Let us explain in more detail the origin of a few of the 15 terms in equation (6). Take as an example the first term, $[\Sigma p_m f(p_m)^2]\Phi[(G_i,G_j,G_k,G_l)]$. It is derived by conditioning on the partition $\sigma$ having all four genes ibd. Since ibd implies identity by state, in this case there is a single common ancestral gene that turns out to be the $m$th allele with probability $p_m$. As a second example consider the next to the last summand in equation (6),

$$[\Sigma p_m^3 f(p_m)^2]\Phi[(G_j,G_k)(G_i)(G_l)] \,.$$

In this instance $G_j$ and $G_k$ form the only ibd pair of the four genes. Given $G_j = G_k$, $Z_{ij}Z_{kl} \neq 0$ only when $G_i$, $G_j$, $G_k$, and $G_l$ all agree in state. By independence, all four genes coincide with the $m$th allele with probability $p_m^3$.

## Recursive Method for Computing Kinship Coefficients

For equation (6) to be an effective method of computing $E(Z^2)$, we must provide an algorithm for computing the generalized kinship coefficients. This can be accomplished by extending Karigl's (1981) recursive algorithm to deal with an arbitrary number of equivalence classes and an arbitrary number of individuals. Description of the algorithm splits into two parts. The first part specifies boundary conditions, and the second part specifies recurrence rules for moving upward through a pedigree by substituting parents for their offspring. It is a convenient abuse of terminology used below to identify an equivalence class or block of genes with the individuals contributing the genes. A person is a founder if both his ancestors are absent from the pedigree. All founders of a pedigree are assumed to be unrelated and noninbred.

### Boundary Conditions

*Boundary condition 1.*—If a person occurs in three or more blocks, then the generalized kinship coefficient $\Phi = 0$. For example, $\Phi[(G_i)(G_j)(G_k)] = 0$ if $i$, $j$, and $k$ are the same person. This condition is obvious because each person has only two genes.

*Boundary condition 2.*—If two founders occur in the same block, then $\Phi = 0$. This follows because founders are by definition unrelated.

*Boundary condition 3.*—If each block contains only founders and neither boundary condition 1 nor boundary condition 2 holds, then

$$\Phi = (1/2)^{m_1 - m_2} \,,$$

where $m_1$ is the total number of people over all blocks and $m_2$ is the total number of different people. For instance, $\Phi[(G_i,G_j)(G_k)(G_l)] = (1/2)^{4-2} = 1/4$ if $i = j$ and $k = l$ are two distinct founders. To verify this boundary condition imagine choosing one initial gene for each founder. Subsequent gene choices for the same founder must coincide with the initial choice if the genes chosen contribute to the same block. If they contribute to a second block, the genes chosen must differ from the initial choice. Each choice is independent, and the maternally and paternally derived genes are equally likely to be chosen.

### Recurrence Rules

These rules operate by substituting the parents $j$ and $k$ for a person $i$ who is currently a member of one or more blocks of $\Phi$. The person $i$ must be a nonfounder as well as a nonancestor of everyone else involved in $\Phi$. Such a nonancestor exists among the nonfounders; otherwise, someone is his own ancestor. In practice, the members of a pedigree, whether affected or unaffected, can be numbered so that children come after their parents. The current highest-numbered person involved must then be a nonancestor. Also note that application of the three recurrence rules below always preserves or diminishes the number of people involved in the calculation of a generalized kinship coefficient. The number of people is never increased.

*Recurrence rule 1.*—Suppose that $i$ occurs in only one block and in only one copy. To simplify notation suppose that this is the first position of the first block. Then

$$\Phi[(G_i, \ldots)(\;) \ldots (\;)] = \tfrac{1}{2}\{\Phi[(G_j, \ldots)(\;) \ldots (\;)] + \Phi[(G_k, \ldots)(\;) \ldots (\;)]\},$$

where only the position corresponding to $G_i$ undergoes a replacement. For instance,

$$\Phi[(G_i,G_l)] = \tfrac{1}{2}\{\Phi[(G_j,G_l)] + \Phi[(G_k,G_l)]\}$$

is the usual recurrence rule for computing ordinary kinship coefficients (Jacquard 1974). In general, recurrence rule 1 follows because the gene drawn at random from $i$ is equally likely to be a gene drawn at random from either parent $j$ or parent $k$.

Recurrence rule 1 also illustrates why $i$ cannot be an ancestor of someone else involved in $\Phi$ (Karigl 1981). For example, if we use the pedigree in figure 1, then $\Phi[(G_6,G_8)] = \tfrac{1}{4}$, $\Phi[(G_3,G_8)] = \tfrac{1}{8}$, and $\Phi[(G_4,G_8)] = \tfrac{1}{8}$. However, we cannot use recurrence rule 1 to replace person 6 by his parents 3 and 4, since $\tfrac{1}{4} = \Phi[(G_6,G_8)] \neq \tfrac{1}{2}\{\Phi[(G_3,G_8)] + \Phi[(G_4,G_8)]\} = \tfrac{1}{2}(\tfrac{1}{8} + \tfrac{1}{8}) = \tfrac{1}{8}$. The problem here is that if, for instance, the paternal gene of 6 is selected, it is not a random gene from 3 but rather the one actually passed to 6. Only this gene of 3's two possible genes can in turn be passed to 8. The two random experiments of choosing a gene from 3 to pass to 6 and choosing a gene from 3 for kinship comparison are not one and the same. We can avoid this paradox by always operating on a nonancestor $i$.

*Recurrence rule 2.*—Suppose that $i$ occurs in only one block but in $s > 1$ copies. Then, assuming that these occupy the first part of the first block, we have

$$\begin{aligned}\Phi[(G_{i_1},G_{i_2}, \ldots ,G_{i_s}, \ldots)(\;) \ldots (\;)] &= (\tfrac{1}{2})^s\Phi[(G_j, \ldots)(\;) \ldots (\;)] \\ &+ (\tfrac{1}{2})^s\Phi[(G_k, \ldots)(\;) \ldots (\;)] \\ &+ [1 - 2(\tfrac{1}{2})^s]\{\Phi[(G_j,G_k \ldots)(\;) \ldots (\;)]\}.\end{aligned}$$

Here $i_1 = i_2 = \ldots = i_s$ all represent the same individual, $i$. The genes $G_{i_1},G_{i_2}, \ldots , G_{i_s}$ are replaced, respectively, by a single gene of $j$, by a single gene of $k$, or by single genes from each of $j$ and $k$. The proof of this rule is just an application of the binomial distribution. As an illustration,

$$\Phi[(G_{i_1},G_{i_2})(G_l)] = \tfrac{1}{4}\Phi[(G_j)(G_l)] + \tfrac{1}{4}\Phi[(G_k)(G_l)] + \tfrac{1}{2}\Phi[(G_j,G_k)(G_l)]\,.$$

*Recurrence rule 3.*—Suppose that $i$ occurs in two blocks with $s$ copies in one block and $t$ in the other block. Again to simplify notation, suppose that $i$ occupies the first $s$ positions of block 1 and the first $t$ positions of block 2. Then

$$\begin{aligned}\Phi[(G_{i_1},G_{i_2}, \ldots ,G_{i_s}, \ldots)(G_{i_{s+1}},G_{i_{s+2}}, \ldots ,G_{i_{s+t}}, \ldots) \ldots] &= (\tfrac{1}{2})^{s+t}\Phi[(G_j, \ldots)(G_k, \ldots)(\;) \ldots (\;)] \\ &+ (\tfrac{1}{2})^{s+t}\Phi[(G_k, \ldots)(G_j, \ldots)(\;) \ldots (\;)]\,.\end{aligned}$$

For instance,

$$\begin{aligned}\Phi[(G_{i_1})(G_{i_2},G_{i_3})(G_l)] &= \tfrac{1}{8}\Phi[(G_j)(G_k)(G_l)] + \tfrac{1}{8}\Phi[(G_k)(G_j)(G_l)] \\ &= \tfrac{1}{4}\Phi[(G_j)(G_k)(G_l)]\,.\end{aligned}$$

This rule follows because the maternally (equivalently paternally) derived gene of $i$ cannot be present in both blocks. Again the binomial distribution determines the coefficient $(\tfrac{1}{2})^{s+t}$. There is no nontrivial extension of recurrence rule 3 to three or more blocks because of boundary condition 1.

The above boundary conditions and recurrence rules suffice to compute generalized kinship coefficients for any partition of individuals in any pedigree. For our purposes the only partitions of interest are those involving affected pedigree members who are typed at the marker locus. Because of the nature of the recurrence rules, other individuals usually must be included in the pedigree. These individuals serve to specify the correct genetic relationships between the affecteds.

As a sample computation, consider the pedigree in figure 1. To calculate Var($Z$) for the first marker locus, it is necessary to evaluate the generalized kinship coefficient $\Phi[(G_6,G_8)(G_4)(G_2)]$. This is done in steps *a*) through *i*) below.

*a*) By recurrence rule 1, $\Phi[(G_6,G_8)(G_4)(G_2)] = \tfrac{1}{2}\{\Phi[(G_6,G_6)(G_4)\,(G_2)] + \Phi[(G_6,G_5)(G_4)(G_2)]\}$.

*b*) By recurrence rule 2, $\Phi[(G_6,G_6)(G_4)(G_2)] = (\tfrac{1}{2})^2\{\Phi[(G_3)(G_4)(G_2)] + \Phi[(G_4)(G_4)(G_2)]\} + \{1 - [2(\tfrac{1}{2})^2]\}\;\Phi[(G_3,G_4)(G_4)(G_2)]$.

*c*) By boundary condition 3, $\Phi[(G_3)(G_4)(G_2)] = (\tfrac{1}{2})^{(3-3)} = 1$ and $\Phi[(G_4)(G_4)(G_2)] = (\tfrac{1}{2})^{(3-2)} = \tfrac{1}{2}$.

*d*) By boundary condition 2, $\Phi[(G_3,G_4)(G_4)(G_2)] = 0$.

*e*) Now substitute the values from *c*) and *d*) into *b*): $\Phi[(G_6,G_6)(G_4)(G_2)] = (\frac{1}{2})^2(1 + \frac{1}{2}) + \{1 - 2[(\frac{1}{2})^2]\}\,(0) = \frac{3}{8}$.

*f*) By recurrence rule 1, $\Phi[(G_6,G_5)(G_4)(G_2)] = \frac{1}{2}\{\Phi[(G_3,G_5)(G_4)(G_2)] + \Phi[(G_4,G_5)(G_4)(G_2)]\}$.

*g*) By recurrence rule 1 and boundary condition 2, $\Phi[(G_3,G_5)(G_4)(G_2)] = \frac{1}{2}\{\Phi[(G_3,G_1)(G_4)(G_2)] + \Phi[(G_3,G_2)(G_4)(G_2)]\} = 0$ and $\Phi[(G_4,G_5)(G_4)(G_2)] = \frac{1}{2}\{\Phi[(G_4,G_1)(G_4)(G_2)] + \Phi[(G_4,G_2)(G_4)(G_2)]\} = 0$.

*h*) Hence, from *f*) and *g*) one infers $\Phi[(G_6,G_5)(G_4)(G_2)] = 0$.

*i*) Finally, substitution of *e*) and *h*) into *a*) yields $\Phi[(G_6,G_8)(G_4)(G_2)] = \frac{1}{2}\{\Phi[(G_6,G_6)(G_4)(G_2)] + \Phi[(G_6,G_5)(G_4)(G_2)]\} = \frac{1}{2}(\frac{3}{8} + 0) = \frac{3}{16}$.

## Combining *Z* Statistics from Different Families

Let $Z_m$ denote the *Z* statistic for the *m*th pedigree of a finite collection of pedigrees. Then an appropriate test statistic for detecting departures from independent segregation of disease and marker phenotypes is

$$T = \frac{\sum_m w_m[Z_m - \mathrm{E}(Z_m)]}{\sqrt{\sum_m w_m^2 \,\mathrm{Var}(Z_m)}}, \tag{7}$$

where the $w_m$ are positive weights. If $r_m$ is the number of affected and typed individuals in the *m*th pedigree, then a plausible choice for $w_m$ is

$$w_m = \frac{\sqrt{(r_m - 1)}}{\sqrt{\mathrm{Var}(Z_m)}}. \tag{8}$$

In this case the square of the denominator in equation (7) is the total number of affecteds minus the number of pedigrees. This choice of $w_m$ is motivated by Hodge's (1984) result that the information content in an affected sib set of size *r* is ~(*r* − 1) times the information content in an affected sib pair. The choice in equation (8) also seems to represent a good compromise between giving all pedigrees equal weight and overweighting large pedigrees because of the multiple comparisons in definition (4).

As long as the weights are any reasonable function of the $r_m$ and the number of pedigrees is large, then Liapunov's Central Limit Theorem (Renyi 1970) implies that *T* follows approximately a standard normal distribution. A one-sided test based on *T* is appropriate because the $Z_m$ should tend to be inflated when the disease and marker phenotypes do not segregate independently.

### Table 2

**Application of the Test Statistic to Real Data**

| Function | Statistic | *P*-Value |
|---|---|---|
| A. Combined Statistic for 20 Rheumatoid Arthritis Families: Marker Locus HLA-D | | |
| $f(p) = 1$ | 0.365 | .359 |
| $f(p) = 1/\sqrt{p}$ | 2.621 | .004 |
| $f(p) = 1/p$ | 1.908 | .029 |
| B. Combined Statistic for 15 Huntington Disease Families: Marker Locus D4S10 (*Hind*III Polymorphism) | | |
| $f(p) = 1$ | 2.195 | .014 |
| $f(p) = 1/\sqrt{p}$ | 3.035 | .001 |
| $f(p) = 1/p$ | 1.325 | .093 |

## Two Sample Applications

Strom and Moller (1981), Michalski et al. (1982), Rossen et al. (1982), and Grennan et al. (1983) examined the cosegregation of HLA haplotypes and rheumatoid arthritis in a combined total of 20 families. Fourteen of these families contain only affected sibs, five contain affected members typed for HLA in 2 generations, and one (Rossen et al. 1982) contains affected members typed for HLA in 3 generations. For this application we will only use the HLA-D genotypes, making the conservative assumption for our test statistic that any unidentified allele is the null allele. The results of our calculations are shown in table 2, where we see that ignoring allele frequencies (i.e., $f(p) = 1$) results in the test statistic (7) being nonsignificant. The dramatic increase in the value of the statistic when allele frequency is taken into account is due to the fact that rheumatoid arthritis is associated with the HLA-DR4 allele, which has a relatively low frequency of .096. These results indicate that our test is sensitive to the function $f(p)$ used and that association and linkage cannot be fully distinguished.

Our second application uses 15 families affected with Huntington disease (Youngman et al. 1986). The marker is the *Hind*III polymorphism detected by the DNA sequence G8 (locus D4S10). These families are large, with nine of them containing affected relatives at least as distantly related as second cousins. Calculation of the generalized kinship coefficients for

**Table 3**

**Simulation Results for the Test Statistic**

A. Distribution of 1,000 Simulated Test Statistics $T$ for Rheumatoid Arthritis Families: Marker Locus HLA-D

| Function | Mean | Variance | Skewness | Skewness[a] / $\sqrt{(6/N)}$ | Kurtosis | Kurtosis[a] / $\sqrt{(24/N)}$ |
|---|---|---|---|---|---|---|
| $f(p) = 1$ | −0.021 | 0.985 | 0.42 | 5.485 | 0.36 | 2.331 |
| $f(p) = 1/\sqrt{p}$ | −0.055 | 1.064 | 0.60 | 7.707 | 1.24 | 7.981 |
| $f(p) = 1/p$ | 0.025 | 1.026 | 1.23 | 15.869 | 2.54 | 16.416 |

|  | Empirical Upper Fifth Percentile[b] | Empirical Upper First Percentile[b] |
|---|---|---|
| $f(p) = 1$ | 1.745 | 2.664 |
| $f(p) = 1/\sqrt{p}$ | 1.716 | 2.799 |
| $f(p) = 1/p$ | 1.906 | 3.423 |

B. Distribution of 1,000 Simulated Test Statistics $T$ for Huntington disease families: Marker Locus D4S10 (*Hin*dIII Polymorphism)

| Function | Mean | Variance | Skewness | Skewness[a] / $\sqrt{(6/N)}$ | Kurtosis | Kurtosis[a] / $\sqrt{(24/N)}$ |
|---|---|---|---|---|---|---|
| $f(p) = 1$ | 0.052 | 0.957 | 0.15 | 1.996 | −0.08 | −0.547 |
| $f(p) = 1/\sqrt{p}$ | 0.010 | 0.983 | 0.29 | 3.695 | 0.24 | 1.544 |
| $f(p) = 1/p$ | −0.053 | 0.953 | 1.37 | 17.708 | 3.10 | 19.993 |

|  | Empirical Upper Fifth Percentile[b] | Empirical Upper First Percentile[b] |
|---|---|---|
| $f(p) = 1$ | 1.729 | 2.358 |
| $f(p) = 1/\sqrt{p}$ | 1.655 | 2.682 |
| $f(p) = 1/p$ | 1.773 | 3.301 |

[a] $N = 1{,}000$; in theory these values are standard normal variates.

[b] For a standard normal variate, the theoretical upper fifth and first percentiles are 1.645 and 2.326.

some of these families involves recursive stepping through as many as 5 generations. As seen in table 2, the test statistics firmly reject the null hypothesis of independent segregation between the *Hin*dIII polymorphism and the Huntington disease locus. As an informal check on how much power is lost compared with the usual lod-score method, we computed a maximum lod score of 3.88 at a recombination fraction of .016 (Lange et al. 1987) for the affected pedigree members, ignoring the marker phenotypes of the unaffecteds. Converting the maximum lod score to a likelihood-ratio test statistic gives $\chi^2{}_1 = 2 \times \log_e 10 \times 3.88 = 17.85$. This $\chi^2$ with 1 df is highly significant ($P \approx .00002$).

It is noteworthy that the function $f(p) = 1/\sqrt{p}$ maximizes the test statistic in both data sets. This choice represents a compromise between ignoring allele frequency (i.e., $f(p) = 1$) and strongly weighting by allele frequency (i.e., $f(p) = 1/p$). All test statistics were computed on an IBM AT microcomputer, with the more challenging Huntington disease data set taking ~75 min. Since most of this time was spent computing the generalized kinship coefficients, it would have taken little additional time to compute the statistic for additional markers.

## Simulation

Since the Central Limit Theorem does not indicate how fast the distribution of the test statistic $T$ of equation (7) will converge to the standard normal distribution $N(0,1)$, it is of interest to use a simulation approach to examine the validity of our normality approximation. For each of the two data sets examined above, we kept the family structures and the affecteds fixed while simulating the segregation of the appropriate marker locus through the pedigrees independently of the disease phenotypes. The test statistic

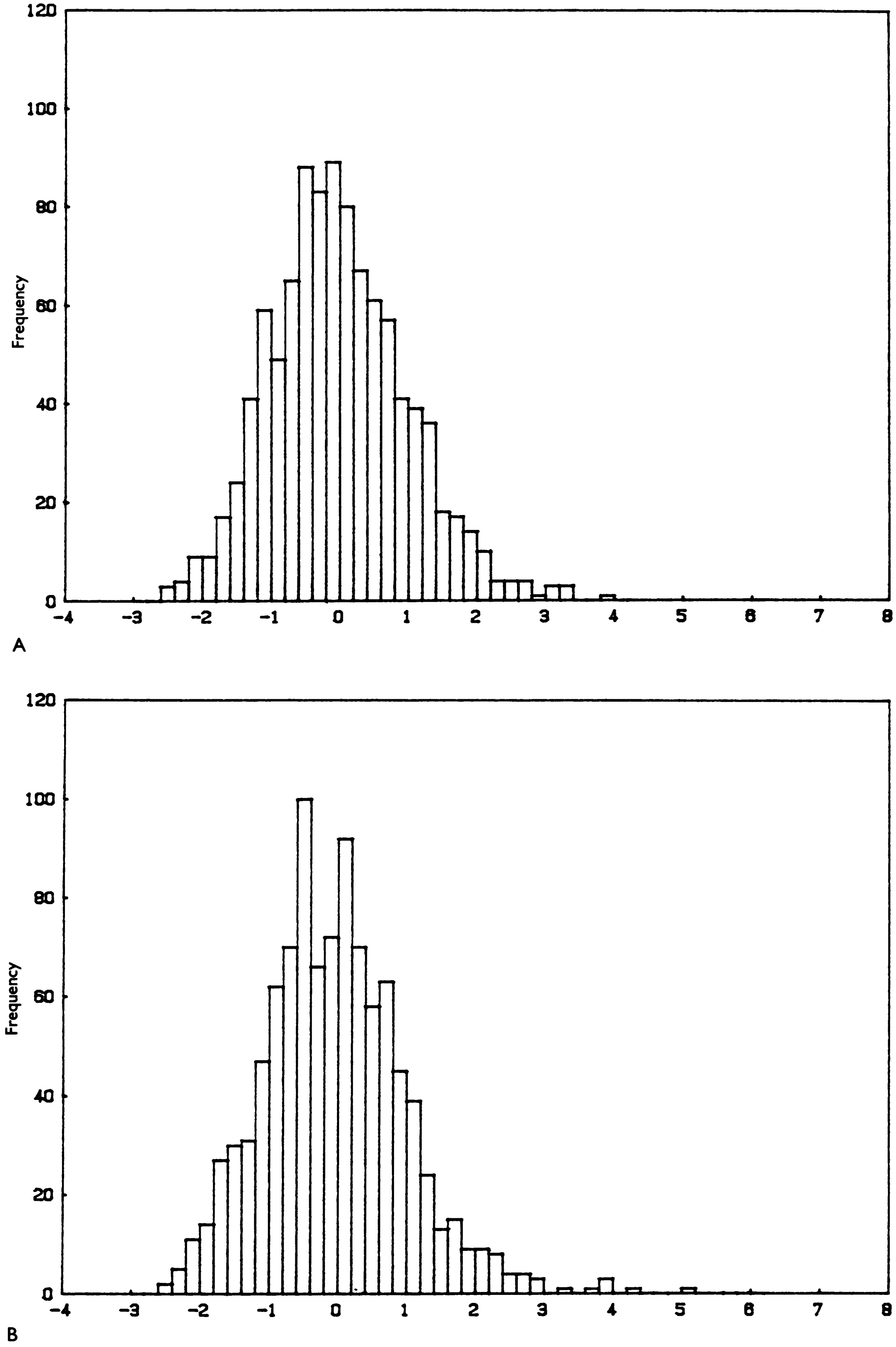

**Figure 2** Histograms of 1,000 simulated test statistics *T*: rheumatoid arthritis families at the HLA-D locus. *A*, $f(p) = 1$; *B*, $f(p) = 1/\sqrt{p}$; *C*, $f(p) = 1/p$.

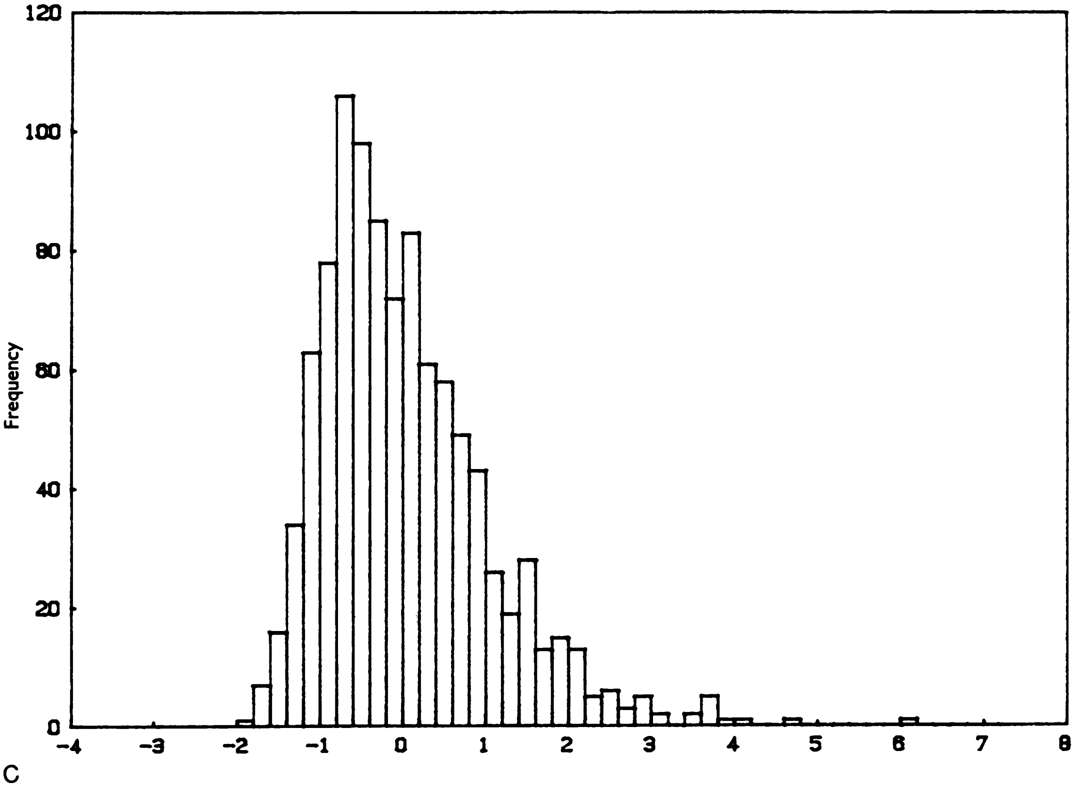

**Figure 2** Continued

was calculated on 1,000 different simulated data sets; the mean, variance, and other pertinent statistics of these simulations are displayed in table 3. Histograms of the simulated $T$ statistics appear in figures 2 and 3. Note that the mean and variance are close to their theoretical values of 0.0 and 1.0 in each case. But the skewness, the kurtosis, and the thickness of the tails of the $T$ distributions all become progressively larger as the influence of allele frequency increases. This effect is sufficiently pronounced for these two data sets that we might be inclined to use the function $f(p) = 1/\sqrt{p}$ instead of the function $f(p) = 1/p$, even ignoring the apparently superior power of the $f(p) = 1/\sqrt{p}$ choice. Our tentative conclusion is that $f(p) = 1/\sqrt{p}$ represents a good compromise between generating a normally distributed test and incorporating important information about allele frequencies. From our point of view, it is important to retain a near-normal distribution of the test statistic. Although simulating for the purpose of estimating a small $P$-value is indeed possible, one cannot achieve much accuracy without considerably more trials than the 1,000 used above.

## Discussion

The test statistic developed here shares the main virtue of the affected-sib-pair method for linkage analysis; namely, it does not rely on questionable assumptions about the mode of inheritance of the disease. The only assumption made under the null hypothesis is independent segregation of disease and marker phenotypes. Furthermore, the test depends only on the marker phenotypes of the affecteds within a pedigree. Other pedigree members need not be typed.

Even when the mode of inheritance is well characterized, our test statistic may prove to be useful. For example, the statistic could serve as a preliminary screen to determine the order in which to type a large number of markers. Instead of typing all available pedigree members at every marker, we could initially type only the affected members at every marker, using our statistic to test for nonindependent segregation of the marker genotypes and the disease phenotypes. If a possible linkage is indicated at any one of the markers, we could then type the remaining unaf-

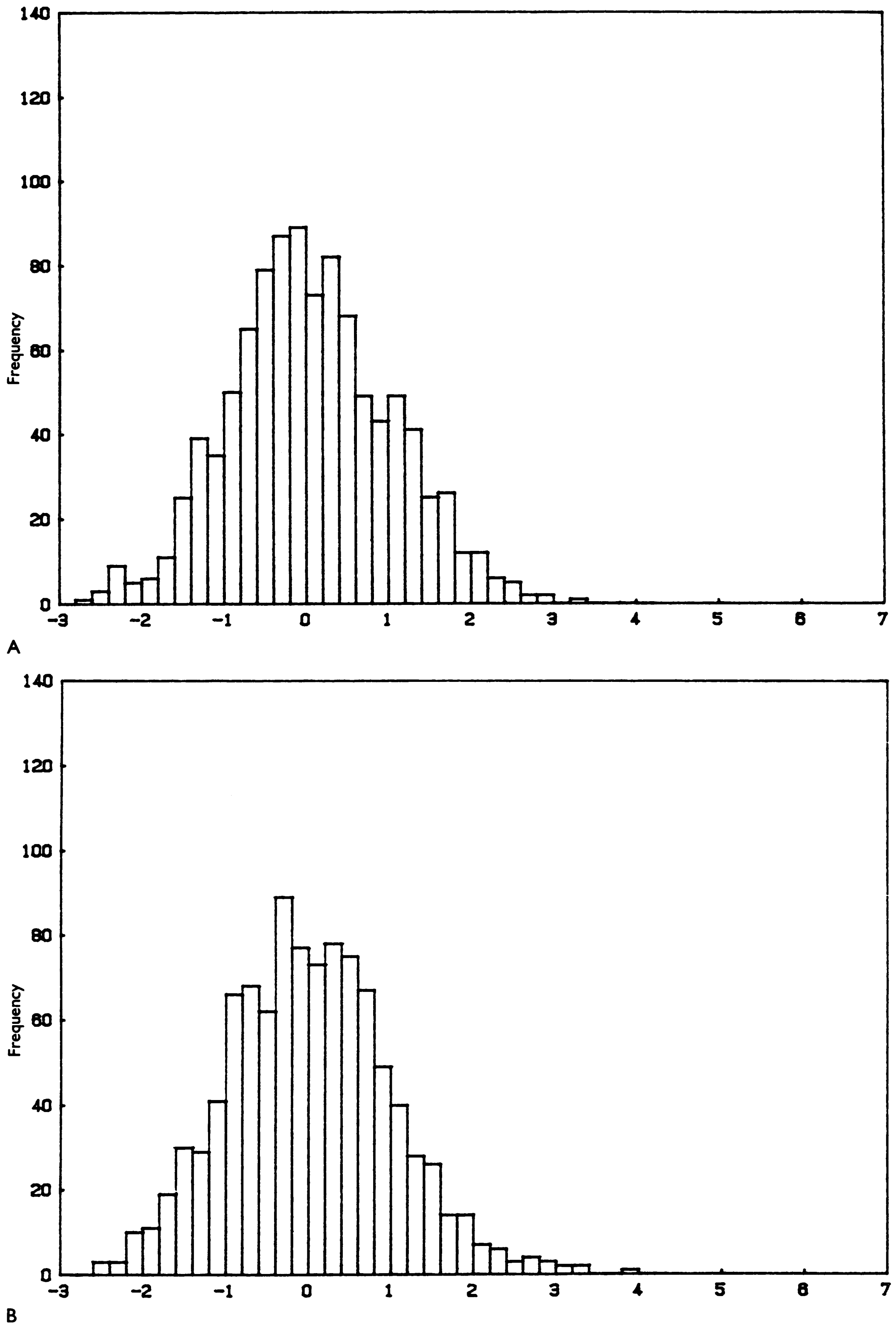

**Figure 3** Histograms of 1,000 simulated test statistics *T*: Huntington disease families at the D4S10 locus. *A*, $f(p) = 1$; *B*, $f(p) = 1/\sqrt{p}$; *C*, $f(p) = 1/p$.

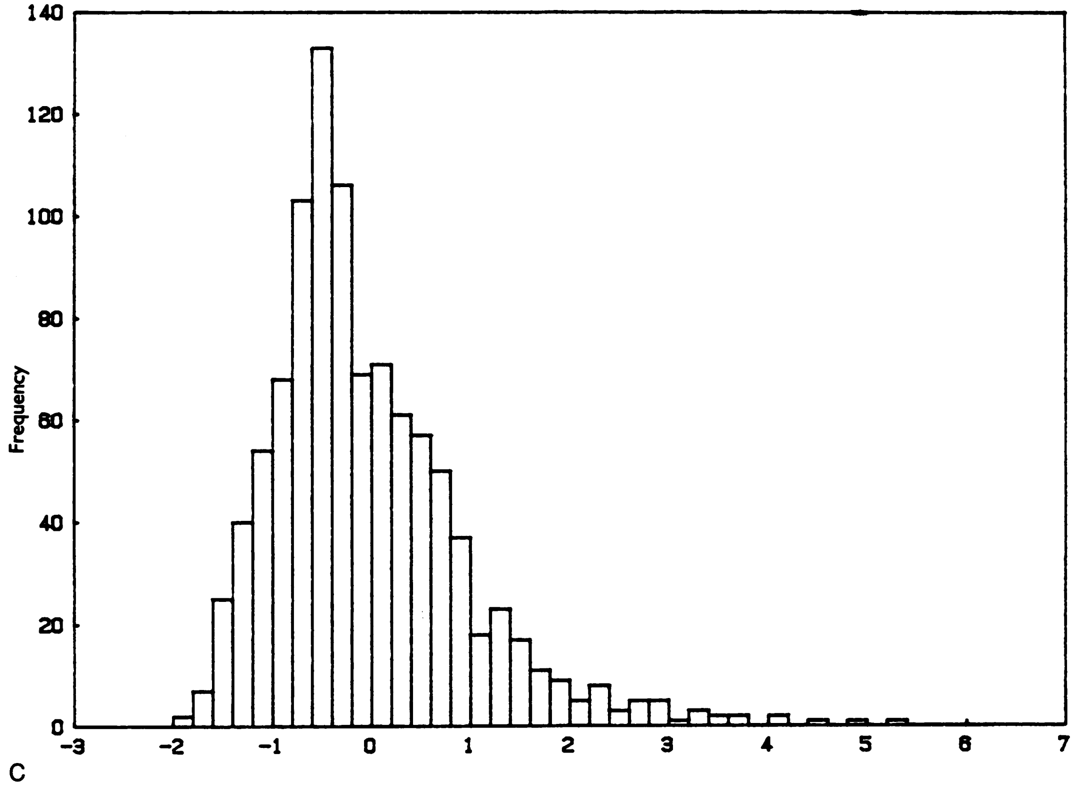

**Figure 3** Continued

fected members at this marker locus and use the more powerful lod-score method. It is relevant to point out here that computing a lod score by using only affected pedigree members can be very difficult, if not impossible, owing to the large number of unknown marker genotypes in unaffecteds that must be considered in the calculations. This computational burden is particularly onerous when the marker is highly polymorphic. In contrast, a highly polymorphic marker presents no difficulty in computing our statistic. The biggest barrier to computing our statistic is the number $r$ of affecteds in a pedigree, since equation (5) has $\binom{r}{2}^2$ terms.

As pointed out above, our test statistic is undoubtedly less powerful than the standard lod-score method when a disease is determined by a well-characterized single locus. There may be less loss of power for recessive diseases if highly inbred pedigrees form a substantial fraction of the data. It would be interesting to know whether fine-tuning the allele-frequency function $f(p)$ or the pedigree weights $w_m$ could increase the power of the statistic. It is clear that the degree of polymorphism at the marker locus critically affects power. In the presence of a highly polymorphic locus, identity-by-state relations closely approximate ibd relations. The test statistic provides no estimate of a recombination fraction, and, like the lod-score method, it can confound linkage and association.

## Acknowledgments

We would like to thank Michael Boehnke, Susan Hodge, and Neil Risch for helpful comments. This research was supported in part by the University of California, Los Angeles; U.S. Public Health Service National Research Service Awards GM-07104 and GM-08185; and National Institutes of Health grant AM 33329.

## References

Crow, J. F., and M. Kimura. 1970. An introduction to population genetics theory. Harper & Row, New York.

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

de Vries, R. R. P., R. F. M. Fat, A. Lai, L. E. Nijenhuis, and J. J. J. Van Rood. 1976. HLA-linked genetic control of host response to *Mycobacterium leprae.* Lancet **2:** 1328–1330.

Fishman, P. M., B. Suarez, S. E. Hodge, and T. Reich. 1978. A robust method for the detection of linkage in familial diseases. Am. J. Hum. Genet. **30:**308–321.

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

Grennan, D. M., P. A. Dyer, R. Clague, W. Dodds, I. Smeaton, and R. Harris. 1983. Family studies in RA—the importance of HLA-DR4 and of genes for autoimmune thyroid disease. Rheumatology **10:**584–589.

Haseman, J. K., and R. C. Elston. 1972. The investigation of linkage between a quantitative trait and a marker locus. Behav. Genet. **2:**3–19.

Hodge, S. E. 1984. The information contained in multiple sibling pairs. Genet. Epidemiol. **1:**109–122.

Jacquard, A. 1974. The genetic structure of populations. Springer, New York.

Karigl, G. 1981. A recursive algorithm for the calculation of identity coefficients. Ann. Hum. Genet. **45:**299–305.

———. 1982. A mathematical approach to multiple genetic relationships. Theor. Popul. Biol. **21:**379–393.

Lange, K. 1986*a*. The affected sib-pair method using identity by state relations. Am. J. Hum. Genet. **39:**148–150.

———. 1986*b*. A test statistic for the affected-sib-set method. Ann. Hum. Genet. **50:**283–290.

Lange, K., M. Boehnke, and D. E. Weeks. 1987. Programs for pedigree analysis. Department of Biomathematics, University of California, Los Angeles.

Michalski, J. P., C. C. McCombs, I. B. DeJesus, and J. L. Anderson. 1982. HLA haplotypes in a family with multiple cases of rheumatoid arthritis. Rheumatology **9:** 451–454.

Penrose, L. S. 1935. The detection of autosomal linkage in data which consist of pairs of brothers and sisters of unspecified parentage. Ann. Eugen. **6:**133–138.

Renyi, A. 1970. Probability theory. North-Holland, Amsterdam.

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

Strom, H., and E. Moller. 1981. HLA and rheumatoid arthritis: a study of five families. Tissue Antigens **18:**92–100.

Suarez, B. K., and S. E. Hodge. 1979. A simple method to detect linkage for rare recessive diseases: an application to juvenile diabetes. Clin. Genet. **15:**126–136.

Suarez B. K., J. P. Rice, and T. Reich. 1978. The generalized sib pair IBD distribution: its use in the detection of linkage. Ann. Hum. Genet. **42:**87–94.

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

Thomson, G., and W. Bodmer. 1977. The genetic analysis of HLA and disease association. Pp. 84–93 *in* J. Dausset and A. Svejgaard, eds. HLA and disease. Munksgaard, Copenhagen.

Youngman, S., M. Sarfarazi, O. W. J. Quarrell, P. M. Conneally, K. Gibbons, P. S. Harper, D. J. Shaw, R. E. Tanzi, M. R. Wallace, and J. F. Gusella. 1986. Studies of a DNA marker (G8) genetically linked to Huntington disease in British families. Hum. Genet. **73:**333–339.