Am. J. Hum. Genet. 62:950–961, 1998

Genetic Association Mapping Based on Discordant Sib Pairs: The Discordant-Alleles Test

Michael Boehnke and Carl D. Langefeld

Department of Biostatistics, University of Michigan, Ann Arbor

Summary

Family-based tests of association provide the opportunity to test for an association between a disease and a genetic marker. Such tests avoid false-positive results produced by population stratification, so that evidence for association may be interpreted as evidence for linkage or causation. Several methods that use family-based controls have been proposed, including the haplotype relative risk, the transmission-disequilibrium test, and affected family–based controls. However, because these methods require genotypes on affected individuals and their parents, they are not ideally suited to the study of late-onset diseases. In this paper, we develop several family-based tests of association that use discordant sib pairs (DSPs) in which one sib is affected with a disease and the other sib is not. These tests are based on statistics that compare counts of alleles or genotypes or that test for symmetry in tables of alleles or genotypes. We describe the use of a permutation framework to assess the significance of these statistics. These DSP-based tests provide the same general advantages as parent-offspring trio–based tests, while being applicable to essentially any disease; they may also be tailored to particular hypotheses regarding the genetic model. We compare the statistical properties of our DSP-based tests by computer simulation and illustrate their use with an application to Alzheimer disease and the apolipoprotein E polymorphism. Our results suggest that the discordant-alleles test, which compares the numbers of nonmatching alleles in DSPs, is the most powerful of the tests we considered, for a wide class of disease models and marker types. Finally, we discuss advantages and disadvantages of the DSP design for genetic association mapping.

Received October 10, 1997; accepted for publication February 6, 1998; electronically published April 7, 1998.

Address for correspondence and reprints: Dr. Michael Boehnke, Department of Biostatistics, School of Public Health, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029. E-mail: boehnke@umich.edu

0002-9297/98/6204-0028$02.00

Introduction

In association-mapping studies, we seek to localize or identify disease genes by searching for dependence between the disease of interest and one or more genetic markers. To date, association-mapping studies have focused primarily on assessing the role of candidate genes, genes believed likely to play a role in disease etiology on the basis of biochemical, physiological, or other available information. In the near future, association mapping may assume substantial or even primary importance in genomewide disease-mapping efforts (Risch and Merikangas 1996).

Classically, association-mapping studies have been used to test for disease-marker associations in a case-control design in which alleles or genotype frequencies in a random sample of affected individuals (cases) are compared with those in a random sample of unaffected individuals (controls). With this case-control design, interpretation of evidence for association can be problematic; while such evidence may reflect linkage or causation, it also may reflect population differences between cases and controls.

To avoid this difficulty, several investigators have developed methods that employ family-based controls (Falk and Rubenstein 1987; Ott 1989; Knapp et al. 1993; Spielman et al. 1993; Thomson 1995; Schaid 1996; Spielman and Ewens 1996). These methods require genotype data on affected individuals and their parents and make use of the transmission information in these parent-offspring trios, instead of requiring an unrelated sample of controls. For these trio-based methods, in the absence of meiotic segregation distortion, positive results reflect evidence for linkage due to either linkage disequilibrium or causation and cannot arise as a result of population differences between cases and controls. These methods have proved useful for early-onset diseases for which parents are routinely available for sampling. However, for late-onset diseases, such as non–insulin dependent diabetes mellitus (NIDDM), Alzheimer disease, heart disease, and many forms of cancer, the requirement for parents cannot always be met and, when met, may result in a sample of cases with unusually

early–onset disease not representative of the disease population as a whole.

In this paper, we describe several tests of association based on a discordant–sib-pair (DSP) design. In this design, we use families with at least one affected and one unaffected sib; parents and additional sibs are not required. Our tests contrast the alleles or genotypes present in the affected and unaffected sibs, either by comparison of counts or tests of symmetry. For several of the tests, dependence of the alleles or genotypes of the DSPs results in test statistics of unknown distribution. To overcome this problem, we use permutation methods (e.g., see Efron and Tibshirani 1993) to assess statistical significance. Because of the matching inherent in the DSP design, each of these tests has the same advantage as the trio-based methods: evidence for association can be taken as evidence for linkage or causation.

To compare the power of our test statistics, we carried out a simulation study in which a genetic disease was partially determined by a genetic locus linked to a genetic marker. Our results demonstrate that the discordant-alleles–test (DAT) statistic, which compares the numbers of nonmatching marker alleles in the affected and unaffected sibs, is the most powerful of the test statistics we considered, for a broad class of genetic models. Similar procedures were proposed by Clarke et al. (1956) and more recently by Curtis (1997). To illustrate the use of our tests, we analyzed 112 northern European Alzheimer disease DSPs genotyped for the apolipoprotein E (ApoE) polymorphism. Our analysis confirms the well-established association between Alzheimer disease and ApoE (Corder et al. 1993, 1994; Saunders et al. 1993), further demonstrating the value of the DSP-based tests. We conclude with a discussion of the advantages and disadvantages of the DSP design and discuss topics for future research.

Methods

Allele- and Genotype-Counting Statistics

We assume that N independent DSPs have been genotyped for a marker with $m \geqslant 2$ alleles. Under the assumption of no disease-marker association, the counts of the alleles in affected and unaffected sibs are expected to be equal. To test for a disease-marker association, we contrast the counts of marker alleles in the affected and unaffected sibs. These counts may be of all alleles present in the sibs or may include only those alleles discordant in the two sibs. These two counting schemes are displayed in table 1. Scheme 1 is simpler and uses all alleles. Scheme 2 uses only those marker alleles that differ within a pair; in so doing, it attempts to reduce the overmatching inherent in the DSP design (see Discussion).

Under either allele-counting scheme, let n_{1j} be the count of marker allele j ($1 \leqslant j \leqslant m$) in the N affected sibs, and let n_{2j} be the corresponding count in the N unaffected sibs. To test for a disease-marker association, we compute the Pearson homogeneity statistic for a $2 \times m$ table:

$$\mathrm{AC}_i = \sum_{j=1}^{m} \frac{(n_{1j} - n_{2j})^2}{n_{1j} + n_{2j}} \ .$$

We call AC_1 the "all-alleles statistic" and AC_2 the "discordant-alleles statistic." Analogous statistics GC_1 and GC_2 compare the counts of all $M = m(m + 1)/2$ genotypes in affected and unaffected sibs and of discordant genotypes in affected and unaffected sibs, respectively.

Table 1

Allele-Counting Schemes for DSPs

Case	Sib Genotypes		Alleles Counted			
			Scheme 1		Scheme 2	
1	11	11	1,1	1,1	…	…
2	11	12	1,1	1,2	1	2
3	11	22	1,1	2,2	1,1	2,2
4	11	23	1,1	2,3	1,1	2,3
5	12	12	1,2	1,2	…	…
6	12	13	1,2	1,3	2	3
7	12	34	1,2	3,4	1,2	3,4

Note.—1, 2, 3, and 4 represent distinct alleles at the marker locus.

Permutation Tests

Although these allele- and genotype-counting statistics directly address the question of disease-marker association, the dependence among the sibling alleles and genotypes violates the assumption of independent observations in the $2 \times m$ (or $2 \times M$) table. Hence, the large-sample distribution of these statistics, under the null hypothesis of no association, is not χ^2 on $m - 1$ (or $M - 1$) df. This problem is easily surmounted by use of a permutation test (e.g., see Efron and Tibshirani 1993).

The basic idea of a permutation test is to choose a family of permutations of the data such that the probability of each permutation is known, under the null hypothesis; usually, all possible permutations are chosen to be equally likely. The distribution of a test statistic is then either determined by evaluating the statistic for all possible permutations of the data or estimated by evaluating the statistic for a random sample of permutations. The P value of the observed test statistic can be estimated by the proportion of permutations for which the permuted-data test statistic is greater than the observed test statistic. When the permuted-data test statistic exactly equals that of the observed data, because of the discrete distributions of our statistics, we may reasonably increment the count by $\frac{1}{2}$ rather than by 1.

In the DSP case, we randomly interchange the affec-

tion statuses of the sibs. Under the null hypothesis of no association, this approach results in permutations of the data that are equally likely. Permutation tests require essentially no assumptions and, in principle, allow estimation of the distribution of any test statistic to any desired level of accuracy; albeit, two applications of the method to the same data may yield slightly different estimates of the P value if a sample of permutations is used instead of all possible permutations. Permutation tests are computationally demanding, particularly if small P values are to be estimated accurately, as is often important in gene-mapping applications. Given current computing power, this disadvantage is more aesthetic than scientific.

Permuting DSP genotype pairs can be done in at least three ways. First, for each permutation, we may switch or not switch the phenotype labels of each DSP independently, with probability $\frac{1}{2}$. Second, we may build the table with entries g_{ij} equal to the number of DSPs in which the affected sib has genotype i and the unaffected sib has genotype j. Then, for each $i < j$, we generate a binomial random variable on $g_{ij} + g_{ji}$ trials and probability of success of $\frac{1}{2}$. If the number of marker alleles is sufficiently small that the number of distinct marker-genotype pairs is small in comparison to the number of DSPs, the second approach is more efficient than the first. Third, if the product

$$\prod_{i=1}^{M-1} \prod_{j=i+1}^{M} (g_{ij} + g_{ji} + 1)$$

is sufficiently small, the P value may be evaluated exactly, by cycling through all sets of genotype counts that maintain the totals $(g_{ij} + g_{ji})$ for all i and j.

Symmetry Statistics

Under the assumption of no disease-marker association, the numbers of each genotype or allele are expected to be equal in the affected and unaffected sibs. Thus, tables of numbers of DSPs in which rows correspond to the genotypes or alleles of affected sibs and the columns to the genotypes or alleles of unaffected sibs are expected to be symmetric. We propose several statistics that exploit this observation.

The simplest of these statistics compares the frequencies of the genotypes in the affected and unaffected sibs. Given no disease-marker association, the expected values $E(g_{ij}) = E(g_{ji})$, for all i and j, suggesting the classic symmetry statistic

$$\mathrm{G_s} = \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} \frac{(g_{ij} - g_{ji})^2}{g_{ij} + g_{ji}}$$

(Bowker 1948), where $M = m(m + 1)/2$ is the number of possible genotypes. For independent DSPs, the genotype-symmetry statistic, G_s, is asymptotically distributed as χ^2 on $M(M - 1)/2$ df; the approximation to this asymptotic distribution generally is accurate if $g_{ij} + g_{ji} > 3$, for all i and j. When this condition is not met, a permutation framework again can be used to assess significance.

Since the number of genotypes M may be large, modest evidence for asymmetry due to a subset of the genotypes may be swamped by the large number of comparisons. A possible solution is to consider alleles rather than genotypes (for another solution, see "Pooling Alleles"). One allele-based symmetry approach is to let a_{ij} be the number of DSPs in which the affected sib has *at least* one copy of allele i and the unaffected sib has *at least* one copy of allele j. Given no disease-marker association, the expected values $E(a_{ij}) = E(a_{ji})$, for all i and j, suggesting the statistic

$$\mathrm{A_s} = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \frac{(a_{ij} - a_{ji})^2}{a_{ij} + a_{ji}} \ .$$

The allele-symmetry statistic, A_s, has the advantage of a smaller number of comparisons, since it is based on alleles rather than genotypes. Because of the double counting of marker heterozygous-homozygous DSPs, the quadruple counting of marker heterozygous-heterozygous DSPs, and the resulting dependent observations, A_s is not distributed asymptotically as χ^2; however, a permutation test can again be applied to assess significance. A potentially useful modification is to weight the contribution of each DSP to sum to 1. Thus, in calculating the weighted allele-symmetry statistic A_{ws}, the four allele pairs of a heterozygous-heterozygous DSP are weighted by $\frac{1}{4}$, and the two allele pairs of a homozygous-heterozygous DSP are weighted by $\frac{1}{2}$.

Pooling Alleles

A prior hypothesis that suggests an association between disease and a specific (set of) marker alleles or genotypes can easily be incorporated into our tests. For example, the prior hypothesis that a single marker allele is associated with disease can be incorporated into all seven statistics by pooling all other marker alleles together. This reduces the effective number of alleles from m to 2. Pooling based on a prior hypothesis regarding a marker genotype can similarly be incorporated into the genotype statistics GC_1, GC_2, or G_s. Pooling increases the power to detect the hypothesized association (see Results). Clearly, if it is hypothesized that more than one allele is associated with disease, alternative pooling strategies, such as the best subset of alleles or an ordering of allele effect, could be employed.

For a marker with a moderate to large number of alleles m, even given no prior hypothesis of association,

Table 2

Characteristics of the Simulated Disease Models

Model	K	f_{DD}	f_{dd}	p	λ_s[a]
Dominant	.05	.20	.025	.074	1.74
		.50	.025	.027	3.23
		.80	.025	.016	4.73
Additive	.05	.20	.025	.143	1.37
		.50	.025	.053	2.12
		.80	.025	.032	2.87
Recessive	.05	.20	.025	.378	1.58
		.50	.025	.229	2.54
		.80	.025	.180	3.45

[a] The sib recurrence-risk ratio λ_s is the recurrence risk to a sib of an affected individual divided by the population prevalence (Risch 1987).

a related allele-pooling strategy may still result in increased power. In particular, we may carry out m tests, one per allele. In test i, alleles $1, 2, \ldots, i-1, i+1, \ldots, m$ are pooled together and are contrasted with allele i to form a two-allele test. As the overall test statistic, we choose the maximum of these m two-allele test statistics (Schaid 1996). This strategy should provide increased power for the case in which a single allele is (primarily) responsible for the disease-marker association. The flexibility of the permutation framework makes the proper accounting for this maximization in assessing significance straightforward, since the same procedure can be carried out for each permutation of the data. This is important, since by ignoring this maximization we could severely overestimate the evidence for association.

Simulation Study

To verify that the permutation-testing framework results in the correct nominal significance level and to evaluate and compare the statistical power for our seven DSP-based association tests, we carried out a set of computer simulations. For each simulation condition, we generated $R = 500$ replicate samples of N DSPs (in most cases, $N = 400$), under a series of one-locus autosomal disease models with reduced penetrance and sporadic cases; we also simulated genotypes for a totally linked codominant genetic marker. For each replicate sample, we calculated the statistics, generated 10,000 permutations of each sample under the null hypothesis of no association, and calculated the resulting 10,000 sets of statistics, to allow us to estimate the P value for each of the tests. For verification of the correct nominal significance level, we generated data with no difference in allele frequencies in affected and unaffected individuals; to assess statistical power, we generated data in which one marker allele was positively associated with disease.

For the disease locus, we assumed two alleles, D and d, with frequencies p and q ($p + q = 1$), respectively, and penetrances $0 \leqslant f_{dd} \leqslant f_{Dd} \leqslant f_{DD} \leqslant 1$, not all equal. We simulated dominant ($f_{dd} < f_{Dd} = f_{DD}$), recessive ($f_{dd} = f_{Dd} < f_{DD}$), and additive [$f_{Dd} = (f_{dd} + f_{DD})/2$] models. For the results reported, we chose a population prevalence $K = q^2 f_{dd} + 2pq f_{Dd} + p^2 f_{DD} = .05$ and an attributable fraction $AF = (K - f_{dd})/K = .50$; the AF is the proportion of disease prevalence that may be ascribed to the presence of the disease-predisposing genotype(s). We chose penetrances $f_{DD} = .20$, .50, and .80 for the predisposing genotype. For the dominant, recessive, and additive models with three free parameters, p, f_{dd}, and f_{DD}, fixing K, AF, and f_{DD} uniquely determined the model. The sporadic rate $f_{dd} = K(1 - AF)$ for all models, and the disease allele frequency $p = 1 - [(f_{DD} - K)/(f_{DD} - f_{dd})]^{1/2}$, $(K - f_{dd})/(f_{DD} - f_{dd})$, and $[(K - f_{dd})/(f_{DD} - f_{dd})]^{1/2}$, for the dominant, additive, and recessive models, respectively. Parameters for the disease models are presented in table 2. Additional analyses were done with $K = .01$ or .10 or with AF = .20 or 1.00.

For the marker locus, we assumed complete linkage to the disease locus (recombination fraction $\theta = 0$). To allow for disease-marker association, we assumed that marker allele 1 was positively associated with disease, with all other marker allele frequencies proportionately reduced in affected individuals. We simulated markers with six codominant alleles and population frequencies .40, .20, .10, .10, .10, and .10 (heterozygosity $H = .76$) and markers with two codominant alleles and frequencies .40 and .60 ($H = .48$), .20 and .80 ($H = .32$), or .10 and .90 ($H = .18$). These simulations allowed us to compare results for two- and multiple-allele markers and to assess the increased power due to a prechosen comparison.

In each simulation case, we set haplotype frequencies to yield an allele-frequency difference of C between randomly selected affected and unaffected individuals. To assess the statistical power of our tests, generally we set $C = .15$; to verify that our permutation-testing framework resulted in appropriate significance levels, we set $C = .00$. For recessive models with $f_{DD} = .20$ and $C = .15$, we excluded the two cases of markers with associated allele frequency of .10, since this combination of parameters would have resulted in negative haplotype frequencies. To assess the impact of sample size and differences in allele frequencies between affected and unaffected individuals, we also analyzed data on 50–2,000 DSPs for the dominant model with prevalence .05 and penetrance .50, a linked marker with allele frequencies .40 and .60, and $C = .05$, .10, or .15.

ApoE and Alzheimer Disease

To illustrate the use of our DSP-based tests for association, we applied them to data on 112 DSPs from

100 unrelated families ascertained for the presence of one or more individuals with Alzheimer disease and typed for the ApoE polymorphism. ApoE has three common alleles, ϵ_2, ϵ_3, and ϵ_4, with frequencies of ~.08, .77, and .15, respectively, in the general population (Utermann et al. 1980). Increased numbers of the ϵ_4 allele result in an increased risk of Alzheimer disease (Corder et al. 1993; Saunders et al. 1993), whereas presence of the ϵ_2 allele appears to reduce Alzheimer disease risk (Corder et al. 1994).

Results

Size and Statistical Power

We began by simulating data for the 12 possible combinations of three disease models (dominant, additive, and recessive with $f_{DD} = .50$) and four genetic marker types (the six-allele marker and each of the three two-allele markers) under the null hypothesis of no disease-marker association. For each combination, we compared the observed distribution of the estimated P values to that expected for a uniform distribution on the interval (0, 1). There was good agreement for all simulation conditions as well as for the data aggregated over all 12 analyses (data not shown).

Next, we generated data, under each of the 52 disease model–genetic marker combinations (nine disease models by six marker types, less two combinations that would have resulted in negative haplotype frequencies), for parameter values that resulted in a difference of $C = .15$ in the frequency of the associated marker in randomly selected affected and unaffected individuals. For each of these 52 simulation conditions, we estimated the power of each of our tests to detect the disease-marker association. To compare the statistical power of the different tests, we estimated the power at significance levels .05, .01, and .001. We then used logistic regression to test for differences in the power estimates among the seven tests.

Table 3 presents power estimates at significance level $\alpha = .001$, for 400 DSPs, for the dominant model with penetrance $f_{DD} = .50$ and for which the allele with population frequency .10, .20, or .40 is associated with disease. For all marker types, the DAT based on AC_2 had the highest power to detect genetic association. This test stood alone as the most powerful test for the six-allele marker, with or without pooling, and was tied to the tests based on AC_1 and A_{ws} for the two-allele markers. For all tests, power was greater for two-allele markers (or, equivalently, for a prechosen comparison), and pooling of alleles tended to be a useful strategy, particularly for the genotype-based statistics. Given a constant allele-frequency difference C between affected and unaffected individuals, evidence for association was stronger when the associated allele was rare or, equivalently, when given a larger relative risk for the disease-marker allele association.

Table 3

Power Estimates for DSP Tests: Dominant Model with $f_{DD} = .50$

Associated Marker Allele Frequency and Test Statistic	Power, by No. of Marker Alleles		
	Six	Six (Pooled)	Two
.40			
AC_1	.718	.812	.918
AC_2	.860	.860	.918
GC_1	.224	.670	.756
GC_2	.276	.502	.710
G_s	.090	.792	.826
A_s	.380	.694	.908
A_{ws}	.598	.812	.918
.20			
AC_1	.854	.918	.986
AC_2	.914	.944	.986
GC_1	.342	.804	.920
GC_2	.428	.834	.962
G_s	.148	.914	.956
A_s	.656	.866	.984
A_{ws}	.768	.914	.986
.10			
AC_1	.980	.992	1.000
AC_2	.988	.994	1.000
GC_1	.636	.970	.996
GC_2	.702	.988	1.000
G_s	.276	.988	1.000
A_s	.930	.990	1.000
A_{ws}	.946	.994	1.000

Note.—Estimated statistical power (based on 500 simulation replicates each) at significance level $\alpha = .001$, for 400 DSPs, with associated marker-allele-frequency difference of $C = .15$ between affected and unaffected individuals. For a detailed description of the statistical tests and genetic models, see Methods.

Repeating the analyses described in table 3 with any combination of genetic model (dominant, additive, or recessive), penetrance value $f_{DD} = .20$, .50, or .80, associated marker allele frequency .10, .20, or .40, or testing at significance level $\alpha = .01$ or .05 resulted in qualitatively the same conclusions. This is illustrated in table 4, which ranks the powers of the various tests (at $\alpha = .001$) separately by disease model and marker type: two-allele marker data or six-allele marker data with or without pooling of alleles, combined over the various marker allele frequencies. Horizontal lines above the test names indicate tests that had power estimates that were indistinguishable at the $\alpha = .05$ level, as tested by logistic regression. Again, the DAT based on AC_2 is at least tied for most powerful in each case considered. For two-allele markers, tests based on AC_1 and A_{ws} are of identical power to AC_2; for six-allele markers without pooling, these tests are next best after the test based on AC_2, whereas, in the six-allele case with pooling, these two

Table 4

Comparison of Statistical Power for the DSP Association Tests

No. of Marker Alleles	Genetic Model	Test Statistic, Ranked by Power						
		1	2	3	4	5	6	7
Six	Dominant	AC_2	AC_1	A_{sw}	A_s	GC_2	GC_1	G_s
	Additive	AC_2	AC_1	A_{sw}	A_s	GC_2	GC_1	G_s
	Recessive	AC_2	AC_1	A_{sw}	A_s	GC_2	GC_1	G_s
Six (pooled)	Dominant	AC_2	G_s	AC_1	A_{sw}	A_s	GC_1	GC_2
	Additive[a]	AC_2	A_{sw}	AC_1	G_s	A_s	GC_1	GC_2
	Recessive	AC_2	A_{sw}	AC_1	G_s	A_s	GC_1	GC_2
Two	Dominant	AC_2 =	AC_1 =	A_{sw}	A_s	G_s	GC_1	GC_2
	Additive	AC_2 =	AC_1 =	A_{sw}	A_s	G_s	GC_2	GC_1
	Recessive	AC_2 =	AC_1 =	A_{sw}	A_s	G_s	GC_1	GC_2

NOTE.—Horizontal lines indicate those tests with power estimates at $\alpha = .001$ that were not significantly different from those at $\alpha = .05$. See Results.

[a] Powers for A_{sw} and AC_1 and for G_s and AC_1 were not distinguishable at the $\alpha = .05$ level, but powers for A_{sw} and G_s were distinguishable.

tests and the test based on G_s rank next. Qualitatively similar results were also obtained for analyses of data with $K = .01$ or .10, AF = .20 or 1.00, or $C = .05$ or .10 (data not shown). The fact that AC_2, AC_1, and A_{ws} result in tests of identical power for two-allele markers in all cases considered, even though their actual statistical values differ, is explained by their identical rankings for the different permutations of the data. We have no explanation of why these ranks are identical for two-allele markers nor of why they are not identical for markers with more than two alleles.

Figure 1 displays the strong effect of sample size and the allele-frequency difference C on the power to detect an association for AC_2, in the case of a dominant disease locus with 50% penetrance and a linked marker with population allele frequencies .40 and .60. Small values of C require very large sample sizes, to achieve even modest power. The sample size required to attain a fixed power appears to increase by a factor of at least x^2 when C decreases by a factor of x.

Effect of Misclassification

Misclassification of genetically predisposed individuals as unaffected will tend to decrease evidence for association. Such genotypic misclassification may occur due to misdiagnosis or if a genetically predisposed individual has not yet progressed to disease or has gone into remission. At its most extreme, this misclassification would be equivalent to selecting sib pairs in which one sib is affected and the other is selected at random with respect to disease phenotype.

To assess the effect of this extreme degree of misclassification, we resimulated data for the 52 genetic models, under this sampling design, and compared allele-frequency differences between the affected and other sibs. For these models, allele-frequency differences for the affected and other sibs (as a proportion of the allele-frequency differences for DSPs) were 91%–93%, 84%–89%, and 77%–86% for the 20%, 50%, and 80% penetrance models, respectively. These results suggest that misclassification will have only modest effect on power, particularly since misclassification generally will not be this extreme. It also suggests that if verifying that a sib truly is unaffected is difficult or expensive, sib pairs in which one sib is known to be affected and the phenotype information regarding the other sib is equivocal might provide a reasonable design. This will be less true if a single locus with high penetrance has strong influence on disease risk.

Application: ApoE and Alzheimer Disease

To further illustrate the use of our DSP-based tests, we applied them to 112 Alzheimer disease DSPs constructed from 100 independent families ascertained for at least one affected individual. In nuclear families with multiple sibs, we included the oldest unaffected sib and a random affected sib; in families with two or more sibships, we selected one DSP per sibship. By choosing the oldest unaffected sib, we hoped to minimize misclassification owing to the unaffected sib later becoming affected.

Table 5 displays the genotype table for the DSPs and the allele counts for the affected and unaffected sibs, under both allele-counting schemes. For example, the

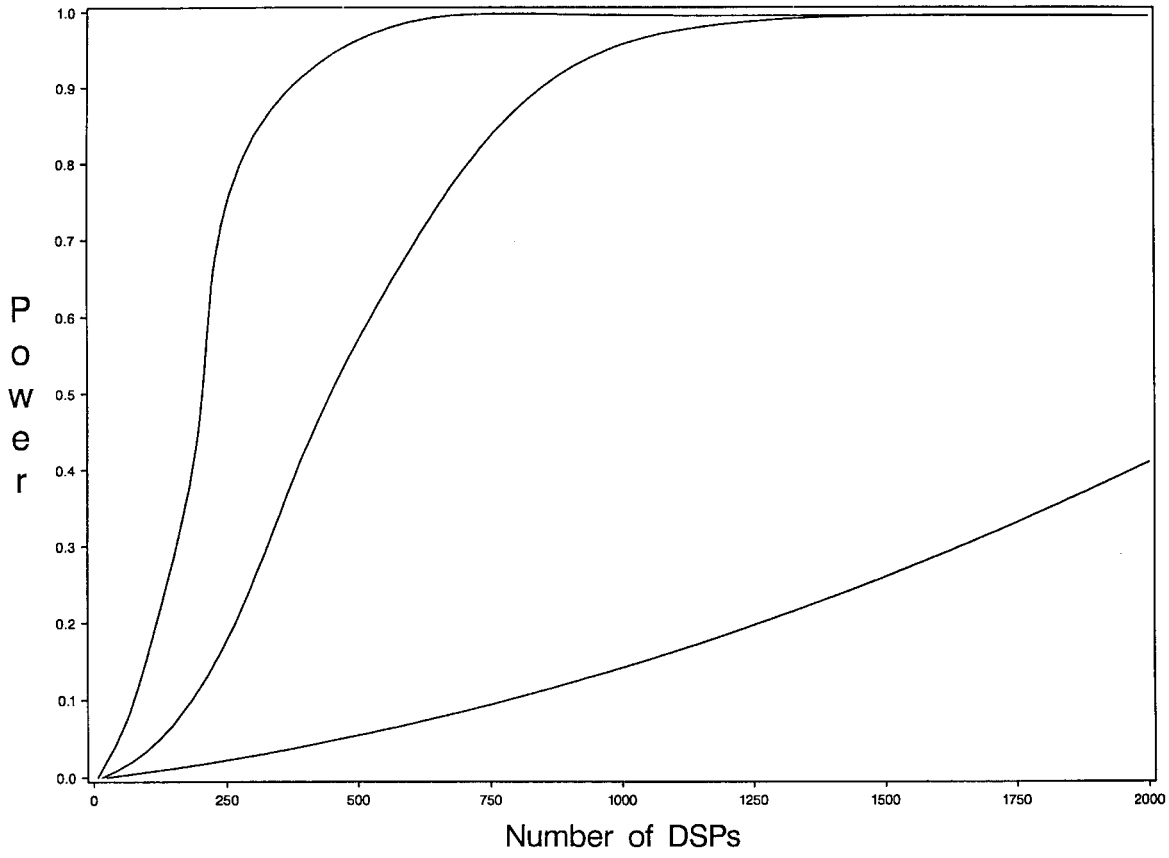

Figure 1 Power of the DAT statistic AC_2. Power curves for associated allele-frequency differences, between unrelated affected and unaffected individuals, of $C = .05$, .10, and .15 (from right to left), for a dominant model with 50% penetrance (sib recurrence-risk ratio $\lambda_s = 3.23$) and a linked two-allele genetic marker with population allele frequencies .40 and .60.

number of ϵ_4 alleles counted for affected sibs, under discordant-alleles scheme 2, was $2 + 23 + 2 \times 2 + 1 + 2 \times 6 + 7 = 49$. As expected, the genotype table appears to be asymmetric. Both sets of counts show a clear excess of ϵ_4 and a deficit of ϵ_2 and ϵ_3 in affected sibs, relative to unaffected sibs (table 5).

Because of the small number of alleles, $m = 3$, and the sparseness of the genotype table, only $2 \times 3 \times 3 \times 2 \times 2 \times 28 \times 7 \times 10 = 141{,}120$ genotype tables are consistent with the observed numbers of genotype pairs. Therefore, we evaluated *P* values by both exact and Monte Carlo methods, the latter based on 100,000,000 permutations of the data (table 6). These *P* values show excellent agreement and demonstrate very strong evidence for association between ApoE and Alzheimer disease. As expected on the basis of our simulation results, the allele-based tests generally performed better than the genotype-based tests. Interestingly, for this example, the tests based on AC_1 and A_{ws} had the smallest *P* values, although not much smaller than those for the tests based on A_s or on AC_2. Pooling alleles ϵ_2 and ϵ_3 resulted in somewhat stronger evidence for association for G_s (probably owing to the smaller number of terms in the genotype-pair table), and slightly less evidence for the allele-counting statistics (probably owing to the loss of information that had previously been provided by allele ϵ_2). Note that, given the sparseness of the genotype table for these data, use of the large-sample χ^2 approximation for the distribution of G_s would be inappropriate.

Despite the strong evidence for association, these data also illustrate the overmatching inherent in the DSP design. Although the ϵ_4 allele frequency of $64/224 \approx .29$ in the unaffected siblings is less than the frequency of $107/224 \approx .48$ in the affected sibs, it still is substantially greater than the general population frequency of .15 (Utermann et al. 1980). Contrasting only those alleles that are discordant between affected and unaffected sibs by use of the DAT results in a much larger difference in the frequency of ϵ_4 in affected ($49/57 \approx .86$) and unaffected ($6/57 \approx .11$) sibs but at the cost of excluding nearly 75% of the data.

Table 5

ApoE Genotype and Allele Counts and Statistics for the Alzheimer Disease DSPs

DSP Genotype Count[a]						
	Unaffected-Sib Genotype					
Affected-Sib Genotype	$\epsilon_2\epsilon_2$	$\epsilon_2\epsilon_3$	$\epsilon_2\epsilon_4$	$\epsilon_3\epsilon_3$	$\epsilon_3\epsilon_4$	$\epsilon_4\epsilon_4$
$\epsilon_2\epsilon_2$	0	0	0	0	0	0
$\epsilon_2\epsilon_3$	0	0	0	0	0	0
$\epsilon_2\epsilon_4$	0	0	1	0	0	0
$\epsilon_3\epsilon_3$	0	1	0	22	4	0
$\epsilon_3\epsilon_4$	0	2	1	23	34	2
$\epsilon_4\epsilon_4$	0	2	1	6	7	6

DSP Allele Count			
	ϵ_2	ϵ_3	ϵ_4
All alleles:[b]			
Affected sibs	1	116	107
Unaffected sibs	8	152	64
Discordant alleles:[c]			
Affected sibs	0	8	49
Unaffected sibs	7	44	6

[a] $G_s = 29.15$; and pooled $G_s = 26.81$.
[b] $AC_1 = 21.09$; and pooled $AC_1 = 17.49$.
[c] $AC_2 = 65.54$; and pooled $AC_2 = 67.24$.

Discussion

Introduction

Genomewide association studies are likely to take on increasing importance in the mapping of genes for complex human diseases (Risch and Merikangas 1996). Emerging genotyping technologies (e.g., see Chee et al. 1996) are likely, in the near future, to enable the genotyping of large numbers of individuals for a dense array of genetic markers quickly and at a manageable cost. Efficient study designs and statistical methods are needed to ensure that the resulting data are put to efficient use.

Parent-offspring trio–based methods (Falk and Rubenstein 1987; Ott 1989; Knapp et al. 1993; Spielman et al. 1993; Thomson 1995; Schaid 1996) have been developed that permit association mapping for diseases such as insulin-dependent diabetes mellitus (IDDM), for which parental genotype information can be obtained. These methods avoid the traditional pitfall of association-mapping studies—that is, false positives due to a poorly matched control sample. In this paper, we have demonstrated that the DSP design, which has this same advantage, can be used for association mapping of late-onset diseases, such as NIDDM, Alzheimer disease, heart disease, and many forms of cancer, for which trio-based methods are not ideally suited.

Advantages and Disadvantages of the DSP Design

The DSP design has a number of advantages. DSPs are generally easy to obtain. Since for most genetic diseases the sib recurrence risk is <.50 and for complex diseases it usually is much lower, most affected individuals provide a DSP. Thus, affected individuals with an unaffected sib are usually more representative of the general disease population than are affected individuals with one or more affected sibs. For late-onset diseases, affected individuals with an unaffected sib also tend to be more representative than affected individuals with living parents. Risch and Zhang (1995) and Rogus and Krolewski (1996) have also demonstrated the utility of the DSP design for linkage mapping of quantitative traits and of qualitative traits with high sib recurrence risks, respectively.

The DSP design has two disadvantages for the purposes of association mapping: misclassification and overmatching. Misclassification may occur because of incorrect diagnosis. It may also occur as a result of reduced penetrance, if a currently unaffected sib develops disease in the future; both of these errors result in a reduction of the power to detect an association. This second sort of misclassification, in which genetically predisposed individuals are classified as nonpredisposed, can be min-

Table 6

Association Test–Statistic Values and *P* Values for ApoE and Alzheimer Disease

Test Statistic	All Alleles			Best Two Alleles		
	Value	Permutation *P* Value[a]	Exact *P* Value	Value	Permutation *P* Value[a]	Exact *P* Value
AC_1	21.09	1×10^{-8}	$.69 \times 10^{-8}$	17.49	3×10^{-8}	3.96×10^{-8}
AC_2	65.54	3×10^{-8}	4.17×10^{-8}	67.24	6×10^{-8}	5.31×10^{-8}
GC_1	22.62	4×10^{-8}	4.00×10^{-8}	18.69	21×10^{-8}	20.25×10^{-8}
GC_2	42.24	89×10^{-8}	84.22×10^{-8}	39.39	260×10^{-8}	254.48×10^{-8}
G_s	29.15	369×10^{-8}	337.23×10^{-8}	26.81	88×10^{-8}	77.74×10^{-8}
A_s	17.97	2×10^{-8}	1.29×10^{-8}	10.29	13×10^{-8}	12.58×10^{-8}
A_{ws}	11.76	1×10^{-8}	$.76 \times 10^{-8}$	10.16	3×10^{-8}	3.96×10^{-8}

[a] Based on 100,000,000 permutations of the data.

imized by appropriate selection of the DSPs. For example, we may select DSPs in which the unaffected sib is relatively old and, if possible, beyond the typical age range of disease onset. We used this approach for analysis of the Alzheimer disease data. In fact, misclassification is unlikely to have any real effect unless high-penetrance alleles play a major role in disease risk (see Results).

Overmatching is less obvious but probably more serious. Because sibs share (disease) genes, allele-frequency differences between affected and unaffected sibs are generally less than between randomly selected affected and unaffected individuals. For example, given a rare, fully penetrant autosomal recessive disease with no sporadic cases and with disease allele frequency p and normal allele frequency q, randomly selected affected and unaffected individuals have disease allele frequencies of 1 and $p/(1 + p)$, so that the difference in disease allele frequencies is nearly 1. In contrast, for DSPs, affected sibs still have a disease allele frequency of 1, whereas unaffected sibs have a disease allele frequency of $\sim\frac{1}{3}$; this difference is $\sim\frac{1}{3}$ less than for randomly selected affected and unaffected individuals. For traits with reduced penetrance and/or sporadic cases, this effect can be more severe. The lone exception is that of a fully penetrant dominant disease, for which the disease-allele-frequency difference in DSPs is essentially the same as that between randomly selected affected and unaffected individuals, which is $\sim\frac{1}{2}$.

To assess more generally the impact of overmatching, we calculated the frequencies $P(1|\mathrm{A})$ and $P(1|\mathrm{U})$ of associated marker allele 1 in randomly selected affected and unaffected individuals and the corresponding probabilities $P(1|\mathrm{A}, \mathrm{DSP})$ and $P(1|\mathrm{U}, \mathrm{DSP})$ for marker allele 1 in the affected and unaffected sibs in a DSP. To do so, we first calculated the corresponding conditional marker-genotype probabilities. For a randomly selected affected individual,

$$P(g_{\mathrm{M}}|\mathrm{A}) = K^{-1} \sum_{g_{\mathrm{D}}} f_{g_{\mathrm{D}}} P(g_{\mathrm{M}}, g_{\mathrm{D}}) \ ,$$

where g_{M} and g_{D} = DD, Dd, or dd are the marker and disease genotypes of the individual, respectively, K is the population prevalence of disease, and f_g is the penetrance of disease genotype g. The joint probability of these genotypes, $P(g_{\mathrm{M}}, g_{\mathrm{D}})$, is easily calculated given the disease-marker haplotype frequencies and the assumption of Hardy-Weinberg equilibrium. The conditional probability of marker genotype g_{M1} for the affected sib in a DSP is

$$P(g_{\mathrm{M1}}|\mathrm{A}, \mathrm{DSP}) = P(\mathrm{DSP})^{-1} \sum_{g_{\mathrm{D1}}} \sum_{g_{\mathrm{D2}}} f_{g_{\mathrm{D1}}} (1 - f_{g_{\mathrm{D2}}}) \times P(g_{\mathrm{D1}}, g_{\mathrm{D2}}) P(g_{\mathrm{M1}}, g_{\mathrm{D1}}) / P(g_{\mathrm{D1}}) \ ,$$

where g_{D1} and g_{D2} are the disease genotypes of the affected and unaffected sibs, respectively. Disease-genotype probabilities $P(g_{\mathrm{D1}})$ are specified by Hardy-Weinberg equilibrium, and the joint probabilities of the disease genotypes $P(g_{\mathrm{D1}}, g_{\mathrm{D2}})$ and phenotypes $P(\mathrm{DSP})$ for the two sibs are calculated easily by conditioning on the number of disease genes the sibs share identical by descent. Analogous expressions hold for an unaffected individual and an unaffected sib in a DSP.

For the disease models we simulated, the ratio of the difference in disease allele frequencies between discordant sibs to that between randomly selected affected and unaffected individuals was 51%–62%; the reduction in differences was smallest for the 80% penetrance models and largest for the 20% penetrance models. This reduction in apparent association can, of course, be overcome by increased sample size. However, since at least a quadratic increase in sample size is required to compensate for a linear decrease in the difference in the allele frequencies, at least a four-fold sample-size increase is required, to compensate for a two-fold reduction in allele-frequency differences.

AC_2 seeks to minimize this overmatching by including only those alleles discordant in the affected and unaffected sibs. The cost of this minimization is a concomitant decrease in sample size. Under the null hypothesis of no association, the expected proportion of alleles actually used can be calculated by conditioning on the number of alleles the sibs share identical by state, as $p_3 + p_4 + p_7 + (p_2 + p_6)$; here, p_i is the probability of genotype pair i (table 1). Each probability p_i may in turn be calculated by conditioning on the number of alleles the sibs share identical by descent. For the marker types considered in our simulation, this proportion was .35 for the six-allele marker and .21, .15, and .09 for the two-allele markers with allele frequencies .40, .20, and .10, respectively. Given an association, disease-discordant sibs will be more marker discordant at a linked marker locus, so more data will be used.

The reduction in allele-frequency differences is not as severe for sib trios with two affected individuals as for DSPs. This was particularly true for the 80% penetrance models, for which the allele-frequency differences between affected and unaffected sibs were at least as great as those between randomly selected affected and unaffected individuals, but it was also true, to a lesser extent, for the 50% and 20% penetrance models. Thus, association-mapping studies carried out in conjunction with affected-sib-pair–based linkage studies may have greater power than those based on a random sample of DSPs. This gain in information must be weighed against the cost of data collection and the possibility of reduced generalizability of the findings.

Alternative Designs

Alternative designs may be used in an effort to achieve appropriate matching between affected and unaffected individuals in an association-mapping study. First, for traits with sufficiently early onset, one can, in principle, choose between parent-offspring trios and DSPs, to carry out family-based tests of association. Because of the overmatching inherent in the DSP design, trio-based methods are generally more efficient at testing for disease-marker association than DSPs; thus, when possible, trio-based methods should be used.

Second, spouse controls may be used. If spouses are well matched, this approach is reasonable. However, if disease prevalence and/or age-at-onset distribution differs in males and females, this approach is less appropriate.

Third, one might seek additional sibs in each family and use trio-based methods on those families for which parental marker genotypes may be inferred with certainty. The problem with this approach is that marker-discordant sibs are more likely to allow inference of parental genotypes. This approach can lead to systematic bias in the evidence for association, with the degree of bias depending on marker allele frequencies (Curtis and Sham 1995).

Assumptions

We made several assumptions in our simulation study. First, we assumed $\theta = 0$ between the disease and marker loci. This assumption is true, or nearly so, for candidate genes and is likely to be nearly true if a genomewide association study is undertaken with a very dense set of genetic markers. For a less dense map, greater distance between disease and marker loci implies attenuation of the degree of association and the need for a larger sample.

Second, we assumed that a single marker allele was positively associated with disease, with all other marker allele frequencies proportionately reduced in affected individuals, and we found our allele-pooling strategy to be effective in this situation. The assumption of a single disease-associated marker allele may be approximately true for isolated populations and/or rare diseases, but it is likely to be an oversimplification for complex diseases, particularly in outbred populations. Given multiple disease-associated marker alleles, the allele-pooling strategy we employed could actually reduce power, as was the case for the allele-counting statistics in the Alzheimer disease example. One might consider other allele-pooling strategies. For example, we might choose the best subset of alleles for the purpose of generating the strongest evidence for association; given m alleles, there are $2^{m-1} - 1$ such subsets. Correcting for multiple tests can still be accomplished in the permutation-testing framework, although this very general approach is likely to have reasonable power only if the number of alleles m is small.

How Many Permutations Are Needed?

In our simulation study, we used only 10,000 data permutations to estimate significance levels. This relatively small number was dictated by our desire to simulate a large number of replicate samples for a large number of different disease models and genetic markers. For an actual data analysis, as in our Alzheimer disease–ApoE example, we would either compute the exact P value or use many more permutations in an effort to estimate the P value to a desired level of accuracy.

A simple approach is to choose a fixed number of permutations R in advance, so that for a particular P value π, the proportional error in estimating π would be no greater than some number k with probability at least $1 - \alpha$. In the case of binomial sampling, this requires that R be no less than $z^2(1 - \pi)/\pi k^2 \approx z^2/\pi k^2$ for π near 0, where z is the $1 - \alpha/2$ point of the standard normal distribution. For example, if $\pi = .0001$, $k = .20$, and $\alpha = .05$, then $z \approx 2$ and $R \approx 1{,}000{,}000$.

A more efficient approach would be to sample sequentially until the estimated P value is determined with sufficient accuracy. This approach is particularly relevant if disease associations are to be tested for many markers. In most such cases, P values will not be near 0 and will require only modest numbers of permutations, for accurate estimation. Besag and Clifford (1991) describe two simple schemes to achieve this end. First, we may choose to generate permutations of the data until S permutations are obtained that yield a larger test-statistic value than that for the actual data. Setting $S = 1/k^2$ for this open-ended procedure results in a proportional error in the estimated P value generally $\leqslant k$; for example, $S = 25$ gives $k \approx .20$. Second, because computing resources are limited and P values will occasionally be small, we may modify the open-ended procedure to stop at some maximum number of permutations R_{max}, even if less than S permutations have resulted in test-statistic values larger than those for the actual data. For either procedure, the estimated P value is the same as that for fixed sampling, and the standard error is approximately the same as that for the fixed-sample estimate of a proportion.

Two-Allele Symmetry Test Statistics

For markers with two alleles, say 1 and 2, we might define three additional allele-symmetry test statistics. The first of these test statistics, A_{2d}, contrasts the numbers of affected and unaffected sibs possessing at least one copy of allele 1. The second, A_{2r}, contrasts the numbers of affected and unaffected sibs possessing two cop-

ies of allele 1. The third, A_{2dr}, is the maximum of A_{2d} and A_{2r}. A_{2d} seems well designed to detect a disease-marker association for a dominant or additive disease allele, A_{2r} to detect a recessive disease allele, and A_{2dr} to be more generally useful. Given a marker with $m > 2$ alleles, each statistic may be computed as the maximum over all m possible two-allele symmetry test statistics. Our simulation results suggested that these statistics did not compare favorably to the others we considered (data not shown).

Possible Extensions and Future Research

We are currently pursuing several extensions to our DSP-based association-mapping methods. First, for sibships with $a > 1$ affected and/or $u > 1$ unaffected sibs, we are attempting to extend our methods to make use of all available sibs rather than just one DSP. One possibility is to base our test on all $a \times u$ possible DSPs in each sibship. A second possibility is to construct for each sibship a weighted DSP, in which the alleles or genotypes present in the a affected and u unaffected sibs are given—for example, weights $1/a$ and $1/u$. A third possibility is to ignore individual DSPs, per se, and to focus on the sibship as a whole. This approach can easily be taken, for example, for AC_1 but not as obviously for AC_2. Any of these three approaches can be incorporated by permutation; instead of permuting disease phenotypes between two sibs, disease phenotypes are permuted among all $a + u$ sibs.

Second, we are seeking to extend our methods to make use of data on multiple tightly linked markers. This becomes particularly relevant once initial evidence for association is obtained, as we seek to fine-map the putative disease gene.

For the sake of simplicity, we have described our DSP-based tests in the context of Pearson χ^2 tests of heterogeneity and symmetry. In fact, for example, each of these test statistics could be replaced by the corresponding likelihood-ratio test statistic. Limited simulation work suggests essentially no differences in size or power when likelihood-ratio statistics are used in the permutation framework. Similarly, one might choose to use a log-linear modeling approach to test for symmetry, quasi-symmetry, or marginal homogeneity (Agresti 1990). The advantage of the current approach is simplicity. The advantage of the log-linear approach is greater flexibility. For example, log-linear models would allow for inclusion of covariates and for tests of specific genetic models such as dominance or recessivity. Standard statistical packages, such as SAS, BMDP, or S-Plus, can be used to compute these and related models. A log-linear model that explicitly models the structure between the four alleles within each DSP may also be a tractable option and is one that we are currently investigating.

Conclusion

Our results demonstrate that the DSP design can be used for association mapping of human genetic diseases. Such an approach is particularly appropriate given a late-onset disease for which the parent-offspring trio–based methods are unusable or inconvenient. Our comparison of several DSP-based test statistics over a broad class of genetic models suggests that the DAT based on a statistic that compares the numbers of nonmatching alleles present in the DSPs was the most powerful test among those we considered. We believe that this test provides a useful approach to association mapping for human genetic diseases.

Acknowledgments

We thank Margaret A. Pericak-Vance, Ann M. Saunders, and their colleagues at the Joseph and Kathleen Bryan Alzheimer Disease Research Center at Duke University, for generously providing the Alzheimer disease family data and ApoE genotypes. We thank David Curtis, for bringing to our attention the study by Clarke et al. (1956). We note that similar work has been carried out independently of us, by Richard Spielman and Warren Ewens. Support from National Institutes of Health (NIH) research grants HG00376 (to M.B.) and NS31153 (to M. A. Pericak-Vance and A. M. Saunders) and from NIH predoctoral training grant HG00040 (to C.D.L.) is gratefully acknowledged.

References

Agresti A (1990) Categorical data analysis. Wiley, New York

Besag J, Clifford P (1991) Sequential Monte Carlo *p*-values. Biometrika 78:301–304

Bowker AH (1948) A test for symmetry in contingency tables. J Am Stat Assoc 43:572–574

Chee M, Yang R, Hubbell E, Berno A, Huang XC, Stern D, Winkler J, et al (1996) Accessing genetic information with high-density DNA arrays. Science 274:610–614

Clarke CA, Edwards JW, Haddock DRW, Howel-Evans AW, McConnell RB, Sheppard PM (1956) ABO blood groups and secretor character in duodenal ulcer. Br Med J:725–731

Corder EH, Saunders AM, Risch NJ, Strittmatter WJ, Schmechel DE, Gaskell PC Jr, Rimmler JB, et al (1994) Protective effect of apolipoprotein E type 2 allele for late onset Alzheimer disease. Nat Genet 7:180–184

Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, Small GW, Roses AD, et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923

Curtis D (1997) Use of siblings as controls in case-control association studies. Ann Hum Genet 61:319–333

Curtis D, Sham PC (1995) A note on the application of the transmission disequilibrium test when a parent is missing. Am J Hum Genet 56:811–812

Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, New York, pp 202–219

Falk CT, Rubenstein P (1987) Haplotype relative risks: an easy

reliable way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Knapp M, Seuchter SA, Baur MP (1993) The haplotype-relative-risk (HRR) method for analysis of association in nuclear families. Am J Hum Genet 52:1085–1093

Ott J (1989) Statistical properties of the haplotype relative risk. Genet Epidemiol 6:127–130

Risch N (1987) Assessing the role of HLA-linked and unlinked determinants of disease. Am J Hum Genet 40:1–14

Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273:1516–1517

Risch N, Zhang H (1995) Extreme discordant sib pairs for mapping quantitative trait loci in humans. Science 268: 1584–1589

Rogus JJ, Krolewski AS (1996) Using discordant sib pairs to map loci for qualitative traits with high sibling recurrence risk. Am J Hum Genet 59:1376–1381

Saunders AM, Strittmatter WJ, Schmechel D, George-Hyslop PH, Pericak-Vance MA, Joo SH, Rosi BL, et al (1993) Association of apolipoprotein E allele epsilon 4 with late-onset familial and sporadic Alzheimer's disease. Neurology 43: 1467–1472

Schaid DJ (1996) General score tests for associations of genetic markers with disease using cases and their parents. Genet Epidemiol 13:423–449

Spielman RS, Ewens WJ (1996) The TDT and other family-based tests for linkage disequilibrium and association. Am J Hum Genet 59:983–989

Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). Am J Hum Genet 52:506–516

Thomson G (1995) Mapping disease genes: family-based association studies. Am J Hum Genet 57:487–498

Utermann G, Langenbeck U, Beisiegel U, Weber W (1980) Genetics of the apolipoprotein E system in man. Am J Hum Genet 32:339–347