Am. J. Hum. Genet. 68:1250–1263, 2001

Transmission/Disequilibrium Test Meets Measured Haplotype Analysis: Family-Based Association Analysis Guided by Evolution of Haplotypes

Howard Seltman,[1] Kathryn Roeder,[1] and B. Devlin[2]

[1]Department of Statistics, Carnegie Mellon University, and [2]Department of Psychiatry, University of Pittsburgh, Pittsburgh

Family data teamed with the transmission/disequilibrium test (TDT), which simultaneously evaluates linkage and association, is a powerful means of detecting disease-liability alleles. To increase the information provided by the test, various researchers have proposed TDT-based methods for haplotype transmission. Haplotypes indeed produce more-definitive transmissions than do the alleles comprising them, and this tends to increase power. However, the larger number of haplotypes, relative to alleles at individual loci, tends to decrease power, because of the additional degrees of freedom required for the test. An optimal strategy would focus the test on particular haplotypes or groups of haplotypes. In this report we develop such an approach by combining the theory of TDT with that of measured haplotype analysis (MHA). MHA uses the evolutionary relationships among haplotypes to produce a limited set of hypothesis tests and to increase the interpretability of these tests. The theory of our approach, called the "evolutionary tree" (ET)–TDT, is developed for two cases: when haplotype transmission is certain and when it is not. Simulations show the ET-TDT can be more powerful than other proposed methods under reasonable conditions. More importantly, our results show that, when multiple polymorphisms are found within the gene, the ET-TDT can be useful for determining which polymorphisms affect liability.

Introduction

Linkage and association between disease status and marker alleles can help pinpoint a liability locus that affects a complex disease or phenotype. To circumvent spurious associations arising from population heterogeneity, Falk and Rubinstein (1987) proposed using the alleles transmitted from parents to their affected offspring as case observations and using untransmitted alleles as control observations. From their insight evolved the transmission/disequilibrium test (TDT) (Spielman et al. 1993). For families containing affected offspring, such as affected sib pairs with parents, the TDT uses the distribution of marker alleles within and among families to test for linkage and association while controlling for population heterogeneity (Ewens and Spielman 1995). The power of the TDT in this setting has been amply demonstrated by the original analysis of insulin-dependent diabetes mellitus and a 5′ flanking polymorphism of the insulin locus (Spielman et al. 1993) and by subsequent power analyses (e.g., Risch and Merikangas 1996; Knapp 1999). For these reasons, the TDT and allied tests have become a favorite tool for analysis of genetic linkage and of association in complex diseases.

A stringent requirement of the original TDT is the definitive transmission of alleles from parents to offspring. Therefore, for a single marker, at least one parent must be heterozygous. Even then, transmissions may not be obvious when parents and offspring are all heterozygous for the same biallelic marker. To increase definitive transmissions, several authors have proposed TDT tests using haplotypes (e.g., Lazzeroni and Lange 1998; Merriman et al. 1998; Clayton and Jones 1999; Clayton 1999; Rabinowitz and Laird 2000; Zhao et al. 2000). In all but the most extreme case of absolute association, transmissions from parents to offspring are more informative for haplotypes than for single markers. One trade-off, however, is the increase in the degrees of freedom of the test: in general, for M realized haplotypes, the tests follow a χ^2 distribution, having $M - 1$ df under the null hypothesis of no linkage or association.

Because of the trade-off between the informativeness of genetic transmission and the increase in degrees of freedom, it is unclear, a priori, whether single-locus or haplotype-based approaches are best. Ideally there would be some way of focusing TDT-type tests on particular haplotypes or sets of haplotypes, thereby increasing both the power of the test and the informativeness of transmissions. There are many ways to group haplotypes. For example, Clayton and Jones (1999) suggest limiting the number of parameters by using a random-effects model.

Received January 12, 2001; accepted March 2, 2001; electronically published April 10, 2001.

Address for correspondence and reprints: Dr. Bernie Devlin, Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara Street, Pittsburgh, PA 15213. E-mail: devlinbj@msx.upmc.edu

0002-9297/2001/6805-0019$02.00

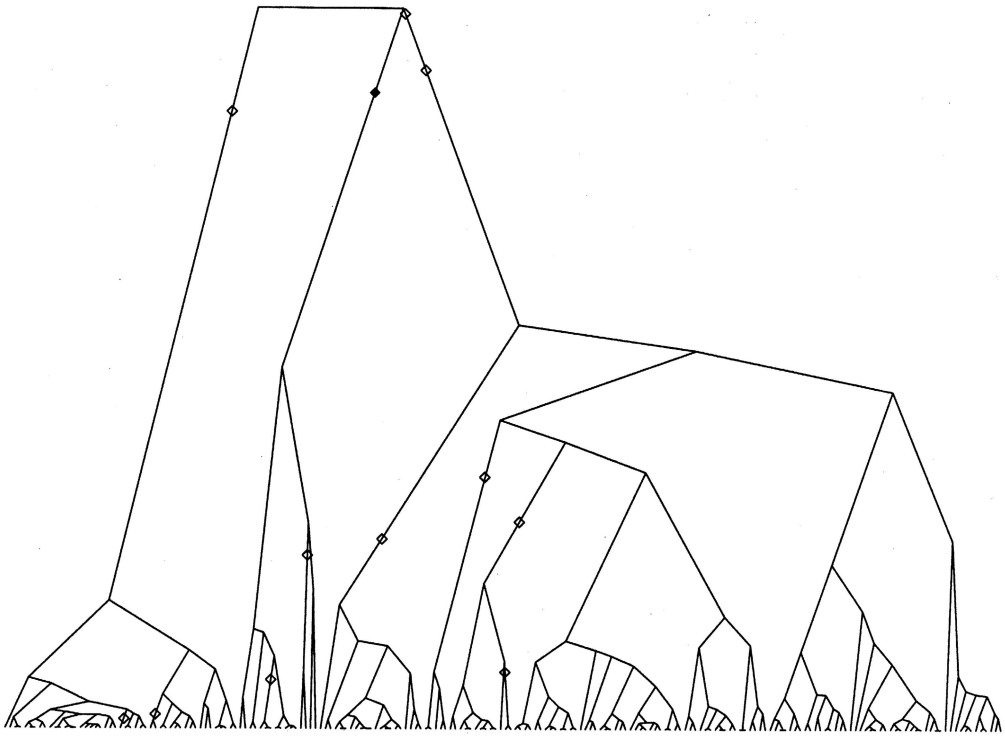

Figure 1 Genealogy of a sample of haplotypes, showing time intervals between coalescent events. Superimposed on this evolutionary process are the mutations that lead to the polymorphisms under study (*unfilled diamonds*) and the disease polymorphism itself (*filled diamond*). Notice that a haplotype possesses each mutation that preceded it in the evolutionary process.

A relatively unexplored method of grouping haplotypes, at least for TDT-type analyses, is to exploit the evolutionary relationships among them. Notably, for both random (Templeton et al. 1987, 1988, 1992; Templeton and Sing 1993; Hallman et al. 1994) and case-control (Templeton 1995) samples, measured haplotype analysis (MHA) has been proposed and used as a means of grouping haplotypes on the basis of evolutionary relationships. To our knowledge, this option has not yet been explored in the TDT setting. The guiding framework of MHA is the cladogram, an unrooted tree of haplotypes. This tree presumably mirrors, as closely as is possible, the mutational process by which the current array of haplotypes evolved.

To guide the grouping of haplotypes, we transfer the ideas of MHA to family-based studies. After a cladogram has been established for the observed haplotypes, one seeks clusters of haplotypes on the cladogram (i.e., clades) that share a common predisposition to disease. This approach potentially increases the power of the TDT test by presenting a reduced number of haplotype groupings for use in testing the transmission-equilibrium null hypothesis. The haplotypes within a clade that are associated with a statistically significant increase in transmission rates are likely to contain functional DNA sequence variations that have a direct impact on the phenotype. These functional variations may involve the polymorphisms defining the haplotype, or they may simply be unmeasured polymorphisms in tight linkage disequilibrium with these haplotypes. Thus, an important advantage of MHA is that the structure of the "associated" portion of the cladogram directs the search for undetected, causal polymorphisms, as well as providing clues to the potential causal effects of the polymorphisms comprising the haplotypes.

In this report, we develop the evolutionary tree (ET)–TDT in two settings: when haplotypes can be determined with certainty and when they cannot. For the latter setting, we adapt many of the results of Clayton (1999). Clayton's methods use the population haplotype frequencies to estimate the missing haplotypes via a partial-likelihood argument. The approach has been shown by simulations to be insensitive to violations of modeling assumptions (Cervino and Hill 2000).

Background

The evolutionary history of a sample of haplotypes can be represented by a coalescent process joining individuals/haplotypes to most-recent common ancestors (MRCA)

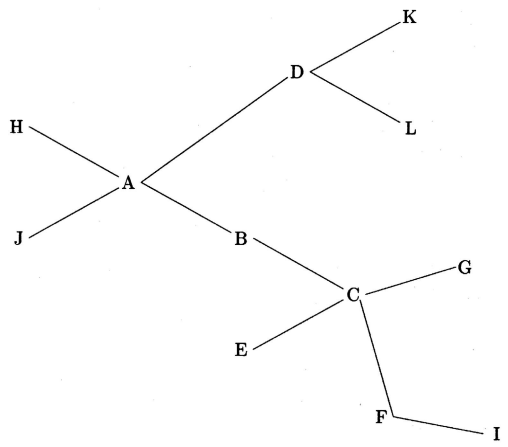

Figure 2 Cladogram of haplotypes from the genealogy depicted in figure 1. The founder is labeled as A, and, working from the root of the tree onward, each new form of haplotype is labeled in the order of occurrence of marker mutations (see open diamonds in fig. 1) to obtain haplotypes B–L.

in the distant past (Kingman 1982); see Hudson (1990) for a general description of the "coalescent." Templeton et al. (1987) note that the disease mutations are embedded in the coalescent process. Figure 1 provides a heuristic example. The nodes at the bottom of the figure represent the current sample of haplotypes, and each node in the tree indicates the MRCA of the lineages beneath them. Superimposed on this evolutionary process, the population of haplotypes experiences mutations that lead to the polymorphisms under study (open diamonds) and the disease polymorphism itself (closed diamond). A haplotype possesses each mutation that preceded it in the evolutionary process.

In this particular example (fig. 1), assuming the causal mutation is not measured, 11 distinct haplotypes are observed in the sample: the MRCA of all the haplotypes (founder) and the 10 new haplotypes created by the 11 depicted mutations. Label the founder as A, and, working from the root of the tree onward, label each new haplotype in the order of occurrence of marker mutations to obtain haplotypes B–L; notice that B is not observed in the extant population. With this one exception, each observed haplotype can be connected to another that differs by a single mutation. Three of the haplotypes (A, H, and J) have the disease mutation embedded in their history, but the remaining seven do not. If there were no other disease mutations in this chromosomal region, these seven haplotypes would share a common probability of being associated with a disease outcome, and the three haplotypes bearing the disease mutation would share a different common probability. Notice the scenario would become more complex if the third marker mutation from the founder were not measured: in this case, D merges with A, and some of these haplotypes do not have the disease mutation; hence, on average, the relative risk of this haplotype is lower than that of the other two mutation-bearing haplotypes (H and J).

If the time at which the mutational events occurred is ignored, the remaining information contained in the rooted tree (fig. 1) emerges as an unrooted tree called a cladogram (fig. 2), with edges representing mutations that result in new haplotypes. Such a cladogram can be reconstructed from a sample of haplotypes, using the method of maximum parsimony, as implemented in the computer program PAUP (Swofford 1998). The parsimony algorithm finds the unrooted tree that connects the observed haplotypes, while using the minimum number of mutations. Although the method of parsimony does not use the frequency of the haplotypes in the sample, it has been found to be a robust and effective tool for reconstructing evolutionary relationships (Swofford et al. 1996).

To understand the impact of apolipoprotein B (apo B) polymorphisms on cholesterol, Hallman et al. (1994) studied the evolutionary relationships among a set of apo B haplotypes (table 1), as determined by maximum parsimony (fig. 3). We use these data and the cladogram throughout this report. Each haplotype in the cladogram is separated by a single mutational step, and, with the exception of haplotype K, uncommon haplotypes appear at the external nodes. The five loci measured in this study are in tight linkage disequilibrium (see table 5 of Hallman et al. [1994]), suggesting that recombination in this region is rare.

The cladogram in figure 3 has been divided into subgroups (clades) by the algorithm of Templeton et al. (1987). The individual haplotypes (C, D, F, G, H, and J), occurring as leaves (terminal nodes) on the tree, rep-

Table 1

Haplotype Structures and Relative Frequencies from Hallman et al. (1994)

	Marker					
Haplotype[a]	1	2	3	4	5	Frequency
A	1	1	1	1	1	.180
B	0	0	1	1	1	.214
C	1	1	0	1	0	.194
D	0	0	1	0	1	.100
E	1	1	0	1	1	.277
F	0	1	0	1	1	.006
G	0	0	0	1	1	.014
H	1	1	1	0	1	.006
J	1	1	0	0	1	.004
K	0	1	1	1	1	.006

[a] The frequency of haplotype K was set at .006, and haplotype I was removed from the table for reasons explained in the original report.

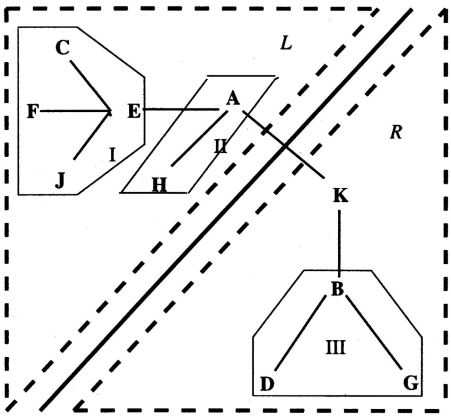

Figure 3 Cladogram of haplotypes from Hallman et al. (1994), as presented in table 1. Subdivisions of the cladogram were produced by the algorithm of Templeton et al. (1987). Zero-step clades are labeled A–K; one-step clades are labeled I, II, and III; and two-step clades are labeled L and R.

resent zero-step clades. Clades I, II, and III represent one-step clades, produced by moving backward one mutational step from the zero-step clades toward internal nodes to produce clusters of haplotypes. When this procedure is repeated, the one-step clades cluster to produce the two-step clades L and R. In the final step, these two clusters are combined to form a single three-step clade.

We adapt the cladistic approach for family-based association analysis. Our adaptation, the ET-TDT, provides some additional power, compared with standard tests for evolutionary scenarios similar to that depicted in figures 1 and 2. As illustrated later in this report, there are some evolutionary scenarios for which one loses some power. Nonetheless, like MHA itself, the most important feature of ET-TDT is its ability to guide the search for mutations that directly affect liability to disease.

For the ET-TDT, suppose the cladogram has M nodes, each having a distinct β_k that measures the relative risk. Instead of conducting an omnibus test (TDT_{all}) with $M - 1$ df, the cladogram delineates a set of sensible comparisons for equality among haplotypes. We will call the rules that describe the set of tests the "cladogram-collapsing algorithm"—even though it is the parameters associated with a clade, not the nodes of the cladogram, that are collapsed to a common value. The objective of the cladistic approach is to find the most parsimonious model that is consistent with the data. If the disease is not linked to or associated with any of the measured haplotypes, then the cladogram-collapsing algorithm should declare that every haplotype has the same relative risk.

We describe the cladogram-collapsing algorithm in the context of the cladogram in figure 3. At any step in the algorithm, a "full model" is required. This is the model from which a score test, which will be described in the following sections, is derived. The full model is the same within each step, but it changes between steps, depending on results in the previous step.

Zero-Step Clade

Begin with a full model that has distinct parameters at nodes A–K ($M = 9$; fig. 3). Test whether any of the zero-step clades has relative risk equal to the nearest internal node; for example, test $\beta_C = \beta_E$. Each of the six tests so defined has 1 df, and each is performed independently. For example, suppose that all the nodes in clade I, except C, have equal βs, and suppose that both of the nodes in clade II have equal βs and that the nodes (B and G) in clade III have equal βs.

One-Step Clade

Next, with conditions based on the previous results, the new full model sets the appropriate parameters as equal (a reduction to $M = 6$ in our example). Test for equality of the parameters associated with the one-step clades within the two-step clades (L and R). In our example, clade (J, E, F) is compared with clade (A, H), and clade (G, B) is compared with clade K.

Two-Step Clade

The new full model constrains βs to be equal within any two-step clade remaining from the previous step. For example, assuming the null hypothesis has not been rejected in either of the two tests in the previous step, $M = 4$. Next, test the equality of the parameters associated with the two-step clades; that is, consider collapsing (J, E, F, A, H) with (G, B, K). If this reduction is not rejected, then the final cladogram has three distinct sets of parameters: a distinct β is associated with each of the following: clades C, D, and (A, B, E, F, G, H, J, K). Notice that in the process of moving through the cladogram, nine 1-df tests were conducted, rather than a single 9-df test.

ET-TDT Analysis with No Missing Data

We initially develop the method assuming no data are missing, which is our shorthand for completely specified transmission of haplotypes from parents to offspring, and then extend it to the typical situation in which parental genotypes and/or haplotype-phase information are missing. The first step of the analysis is to construct a cladogram from the parental haplotypes. The next step is to conduct a series of hypothesis tests to discover the

most parsimonious set of parameters consistent with the cladogram and the data.

Assume that the data consist of parent/affected offspring trios, with known haplotype phases. Suppose that M distinct haplotypes are observed in the sample. Define n_{ij} as the number of heterozygous parents with the haplotype pair (i, j). The transmission data may be recorded as $x_{i/j}$, the number of times that a parent with haplotypes (i, j) transmits i to the affected offspring. By conditioning on the n_{ij}s, it follows that $X_{i/j} \sim$ binomial(n_{ij},$p_{i/j}$), and all $M(M-1)/2$ binomials are independent of each other. The $p_{i/j}$s can be parameterized by $M-1$ free parameters,

$$p_{i/j} = \frac{e^{\beta_i}}{e^{\beta_i} + e^{\beta_j}} ,$$

with the Bradley-Terry model (also known as the "gamete-competition model"; Sinsheimer et al. [2000]). For biallelic markers, this parameterization is equivalent to Terwilliger and Ott's (1992) haplotype relative-risk parameterization.

When no data are missing, the full likelihood from each family trio is expressed as the probability of observing the parental haplotypes multiplied by the conditional probability of the offspring's haplotypes, given the parental haplotypes: $L^{(F)} = L^{(P)} \times L^{(C)}$. If the corresponding log-likelihood contributions are denoted by l, the log likelihood decomposes additively: $l^{(F)} = l^{(P)} + l^{(C)}$. As noted above, for each family, $L^{(C)}$ corresponds to a Bernoulli random variable that depends only upon β. Pooling the data from all mating types results in a likelihood composed of a product of independent binomial random variables.

The parental likelihood, $L^{(P)}$, is obtained assuming that the haplotypes are approximately in Hardy-Weinberg equilibrium with haplotype relative frequencies ψ_i. Thus, the probability that a parent drawn at random from the population with genotype $g = (i,j)$ is

$$\Pr[g = (i,j)] = \begin{cases} \psi_i^2 & \text{if } i = j \\ 2\psi_i\psi_j & \text{if } i \neq j \end{cases} .$$

In a population of cases, the corresponding haplotype relative frequency is $\psi_i^* = \psi_i e^{\beta_i} / \sum_j \psi_j e^{\beta_j}$. Notice that the likelihood contribution obtained from observing the haplotypes of the parents, $L^{(P)}$, also contributes information about β, but this information is confounded with the haplotype frequencies in the population. Thus, when no data are missing, inferences robust to population stratification are traditionally based on $L^{(C)}$ only. Detailed expressions for the likelihood of a trio are given in Appendix A.

In this analysis, we assume that (1) the relative risk of a genotype is adequately modeled by the product of haplotype effects and (2) the cladogram can be correctly inferred. Implicit in the second assumption is the requirement that recombinations between markers are rare, so that the history of the haplotypes can be described by a mutation tree. This assumption is not unduly restrictive; in general, only if the region is in tight linkage will there be substantial association between haplotypes and an embedded disorder-associated mutation. In our simulations, we we examine the robustness of our inferences to violations of the second assumption.

With the cladistic approach, external clades are compared with internal clades to determine whether the two have equal risk. Let S and T represent the internal and external clades, respectively. To test $\beta_S = \beta_T$, we develop a score test, which is described in Appendix B. To test this hypothesis, we require estimates of the remaining βs, which are obtained by maximizing the likelihood under this constraint.

For a cladogram with M nodes, we suggest conducting $M-1$ score tests, each with 1 df, following the cladogram-collapsing algorithm. (Recall that it is the parameters, not the haplotypes themselves, that are collapsed into common clades. A parent with two distinct haplotypes within a common clade is still considered to be heterozygous.) The procedure terminates when no further reductions are possible for a given significance level of $\alpha/(M-1)$. If the cladogram-collapsing procedure results in a cladogram that possesses $K > 1$ distinct parameters associated with the clades, then conclude that a disease-susceptibility locus is likely to be embedded within the region under study and that the deleterious allele is likely to be associated with the clade(s) exhibiting higher estimated levels of risk. We note, however, that these conclusions are only valid if the sample size for each haplotype is sufficiently large. Uncommon haplotypes may invalidate the asymptotic results; see the section entitled "ET-TDT Analysis with Missing Data" for a discussion of the estimation of the cladogram and the treatment of uncommon haplotypes. For small samples, we recommend that a permutation-type test be conducted to obtain P values (e.g., Morris et al. 1997). Generate N data sets, each using the following algorithm:

1. Generate an offspring from each parent pair under the null hypothesis.
2. Repeat the set of cladistic hypothesis tests outlined above, and let $p^* = \min\{p_1, \ldots, p_{M-1}\}$, where p_i is the P value associated with the ith test performed in the cladogram-collapsing algorithm.
3. Compute the empirical P value by counting the fraction of simulated realizations, $\{p_1^*, \ldots, p_N^*\}$, smaller than the realized p_{obs}^*: $P = \frac{1}{N}\sum_j I\{p_j^* < p_{\text{obs}}^*\}$.

ET-TDT Analysis with Missing Data

We now extend the methods to account for ambiguity of haplotypes, which we refer to as the "missing-data" case. We assume that, from each family, one affected offspring is sampled along with parents and/or some siblings (parents and siblings may be unaffected or of unknown status). Multiplex families can be analyzed with our methods, but they substantially complicate the likelihood (Clayton 1999). For our purposes, then, additional siblings of unknown or unaffected status will be used only to help infer the missing information about the haplotypes of the parents.

To generalize the ET-TDT for missing data, we follow Clayton (1999) by formulating a partial-likelihood model. In this instance, the full likelihood no longer factors into two components: the likelihood attributable to parental haplotypes, $L^{(P)}$, versus the likelihood of transmission of haplotypes to affected offspring, $L^{(C)}$. Although the primary inferences will be based upon $L^{(C)}$, multiple configurations of haplotypes can be consistent with the observed data for a trio. $L^{(P)}$ is now needed to weight the likelihood over this set of plausible haplotype configurations. Clayton's partial-likelihood approach represents a robust and powerful compromise between using the full likelihood and using only those families with no missing data (Cervino and Hill 2000). Although this method has some potential for bias under extreme population admixture, it is clearly superior to either of the two options just mentioned, both of which are biased.

For incomplete data, the ith family may be consistent with a set of possible haplotypes and transmissions, each giving a different likelihood contribution. If the observed data from the ith family are consistent with a set of scenarios $\mathcal{P}_i$, the score is obtained as a weighted average of particular scenarios consistent with $\mathcal{P}_i$. Detailed expressions for the score test are given in Appendix C.

Summary of the Test Procedure

1. Estimate the cladogram and collapse sparse nodes into neighboring nodes so that no node is uncommon. M refers to the number of nodes remaining after this collapse.

2. Organize the cladogram by clades, using the Templeton et al. algorithm.

3. Perform a score test for each potential collapse determined by the cladogram-collapsing algorithm.

   (a) Estimate the constrained-nuisance parameters using the expectation-maximization (EM) algorithm (Appendix D).

   (b) Compute the score test (Appendix C).

   (c) If $P > \alpha/(M - 1)$, assume the parameters are equal for the two nodes (or clades).

4. Use the permutation procedure described in the section entitled "ET-TDT Analysis with No Missing Data," to obtain a P value, with one modification. For each iteration of the permutation procedure, randomly select one consistent setting for the parents' haplotypes from set $\mathcal{P}$ for each family with incomplete data. Simulate the affected offspring, using the selected parental haplotypes.

Constructing a Haplotype Cladogram

For some of the families, the haplotypes will be directly observable using the genotype-elimination algorithm (Lange 1997). More of the haplotypes can be inferred with a high degree of accuracy by use of Clark's (1990) algorithm, which infers uncertain haplotypes from the set of observed haplotypes. We suggest that this collection of haplotypes be used to build an unrooted cladogram, using PAUP (Swofford 1998).

Next, the set of families with unresolved haplotype configurations should be examined to determine whether their multilocus genotypes are consistent with any pair of haplotypes, each of which deviates from the nodes in the cladogram by, at most, one step. Include all of these nodes in the cladogram. Because these unambiguous haplotypes have not been directly observed in the population, it seems safe to assume they are uncommon. Hence, this somewhat arbitrary rule will have little or no impact on the analysis. At this point, the multilocus genotypes from virtually all families should be consistent with subsets of haplotypes that constitute the cladogram, and families that do not meet this condition should be removed from the analysis. Assuming the data are highly informative, the data set now has two features: (1) the deduced cladogram includes all but the most uncommon haplotypes from the population; and (2) all the families are consistent with at least one subset of haplotypes in this cladogram. If the data are only weakly informative, the problem should be recast, as described in the Discussion, or haplotype analysis should be avoided.

Assuming that the cladogram includes the full set of haplotypes, set $\beta = 0$ and estimate ψ, using the EM algorithm described in Appendix D. Collapse any nodes that have very small expectation under the null hypothesis—for example, any with $2n\psi_k < 10$. This rule differs from the cladogram-collapsing algorithm in that the pooled haplotypes are treated as a single haplotype. If the cladogram is extremely complex, a liberal amount of pooling will reduce the degrees of freedom in the test and could enhance the power.

Two types of variability can cause errors in the estimated evolutionary history of a set of haplotypes: sampling error and evolutionary error. In sampling error, the observed frequencies of the haplotypes differ from the

frequencies in the population. Variability in haplotype frequencies has little influence on the cladogram built by the method of parsimony, because parsimony is affected by haplotype frequencies only in the weakest sense. The most obvious impact occurs when rare haplotypes are missing from the sample. Rare haplotypes typically are located at the external nodes of the cladogram; however, in this case, they have little practical consequence, because they should be pooled with the nearest internal node. If, by chance, an uncommon internal node is not sampled, then parsimony typically imputes the missing node to complete the graph.

Evolutionary error presents a much thornier problem, which will be described in more detail in the Discussion. At least two processes cause parsimony to inconsistently reconstruct the true evolutionary relationships among haplotypes: (1) a recombinational event early in history that disrupts tightly linked alleles; and (2) convergent evolution, a situation that may be expected if markers or the disease gene undergo frequent, recurrent mutation. Our methods condition on the reconstructed cladogram and therefore assume it is the true cladogram, because it would be difficult or impossible to account for evolutionary error with standard statistical methods. In the simulations in the next section, we investigate the impact of evolutionary error.

Simulations

Data were generated using the haplotypes and relative frequencies given in table 1. To test a range of evolutionary scenarios, we considered cases in which all the haplotypes within a clade associated with the disease share equal risk level (eq), other cases in which there is a gradient of risk levels within a clade (neq), incorrect cladograms resulting from recombination (rec), and errors in the reconstruction of the cladogram (clad). A total of 14 models were explored, each named by the symbols denoting the haplotypes associated with an increased relative risk. For model "null," none of the haplotypes was associated with the disorder. The next seven models mimicked an evolutionary history with a single ancestral mutation and a correct cladogram: for model C, only haplotype C was associated with the disorder ($e^{\beta_C} = 1.7$); for model BD-eq, haplotypes B and D were associated with the disorder ($e^{\beta_B} = e^{\beta_D} = 1.6$); for model CE-eq, haplotypes C and E were associated with the disorder ($e^{\beta_C} = e^{\beta_E} = 1.6$); for model KBDG-eq, all the haplotypes in clade *R* were associated with the disorder ($e^{\beta_K} = e^{\beta_D} = e^{\beta_B} = e^{\beta_G} = 1.6$); for model BD-neq, $e^{\beta_B} = 1.6$ and $e^{\beta_D} = 1.9$; for model CE-neq, $e^{\beta_C} = 1.7$ and $e^{\beta_E} = 1.5$; and for model KBDG-neq, $e^{\beta_K} = 1.3$, $e^{\beta_B} = 1.5$, $e^{\beta_D} = 2$, and $e^{\beta_G} = 2$. Two models were motivated by two distinct ancestral mutations: in model CBG-eq, haplotypes C, B, and G were associated with the disorder ($e^{\beta_C} = 1.4$ and $e^{\beta_B} = e^{\beta_G} = 1.6$); and in model CBG-neq, $e^{\beta_C} = 1.4$, $e^{\beta_B} = 1.6$, and $e^{\beta_G} = 2.0$. Three models mimicked the case in which one haplotype was created by recombination rather than a mutation: for model DH-rec, haplotype H was generated by recombination between A and D, and D was associated with the disorder ($e^{\beta_D} = 2.0$ and $e^{\beta_H} = 2.5$); for model EFG-rec, haplotype G was generated by recombination between B and E, and E was associated with the disorder ($e^{\beta_E} = 1.7$, $e^{\beta_F} = 3.0$, and $e^{\beta_G} = 4.0$); and, for model BK-rec, haplotype K was generated by recombination between A and B, and B was associated with the disorder ($e^{\beta_B} = 1.8$ and $e^{\beta_K} = 4.0$). Finally, model KBDG-clad simulated an incorrect cladogram; haplotype D should have been placed on the cladogram as a leaf attached to A instead of B, and $e^{\beta_K} = 1.3$, $e^{\beta_B} = 1.5$, and $e^{\beta_D} = e^{\beta_G} = 2.0$. Each data set consisted of 400 simplex families; one half of the families had two genotyped parents, and the other half had only one genotyped parent.

Because we had missing data, we used Clayton's TRANSMIT program to perform TDT tests for power comparisons (Clayton 1999). We used two versions of the TDT test: the TDT_{all} and the max-TDT test (which tests X vs. not X for all designated haplotypes, using a Bonferroni correction). With TRANSMIT we analyzed only those nodes with haplotype frequencies of ≥ 0.04, to avoid small cell counts. To compute the ET-TDT test, the cladogram was taken to be given as in figure 3 but with the following nodes collapsed and relabeled to avoid analyzing any haplotypes with extremely small counts N = (K, B, and G), P = (F, E, and J), and Q = (A and H).

Two hundred data sets were generated under each of the conditions considered. With the ET-TDT, we consider the null hypothesis to be rejected if $p^* < 0.05/4$. In a power comparison, the ET-TDT was somewhat more powerful than the TDT_{all} test for many scenarios; and in some cases, the difference was substantial (table 2). The relative power of the tests varied greatly by cladogram structure. For model KBDG-eq, where the entire clade has a common relative risk, ET-TDT had a slight edge in the power comparison. However, ET-TDT did not outperform the TDT_{all} in two conditions where only a portion of the clade was associated with the disease (BD-eq and CE-eq). Other results were much more promising than expected. For instance, although KBDG-neq was comparatively harder to analyze successfully than KBDG-eq, in models BD-eq and CE-eq, creating a gradient of relative risks did not affect the relative performance of the ET-TDT. Even for models CBG-neq, DH-rec, and BK-rec—which were challenging because they had a configuration of haplotype effects that were not well supported by the cladogram—the power was greater for ET-TDT. This

Table 2

Power of ET-TDT versus Two TDT Approaches Described by Clayton (1999)

Model	ET-TDT	TDT_{all}	max-TDT
Null	4.0	6.5	3.8
C	82.0	69.0	74.5
BD-eq	62.5	71.0	61.5
CE-eq	69.5	73.5	63.5
KBDG-eq	81.5	78.5	68.5
BD-neq	75.5	81.0	70.5
CE-neq	68.0	73.5	61.0
KBDG-neq	80.5	85.0	75.5
CBG-eq	82.0	65.5	56.5
CBG-neq	76.5	63.0	54.5
DH-rec	77.5	73.0	78.8
EFG-rec	82.0	91.0	73.5
BK-rec	86.0	84.5	84.5
KBDG-clad	95.0	80.5	68.0

NOTE.—Each model is named by the symbols denoting the haplotypes associated with an increased relative risk with additional nomenclature indicating cases in which all the haplotypes within a clade associated with the disease share equal risk level (eq); cases in which there is a gradient of risk levels within a clade (neq); incorrect cladograms due to recombination (rec); and errors in the reconstruction of the cladogram (clad)

robustness was not universal, however, as demonstrated by the lower power observed for model EFG-rec.

Compared with max-TDT, ET-TDT was considerably more powerful for most models. The most surprising result occurred for model C, where ET-TDT exhibited a substantial advantage in power. In addition, whenever more than one haplotype was associated with the disease, the TDT_{all} test was more powerful than the max-TDT test, suggesting that the X-vs.-not-X collapsing of haplotypes is nonoptimal for the type of scenarios often envisioned to occur in cladistic analysis.

More noteworthy than the power differential, however, is the difference in interpretability. Let X–Y indicate that clades X and Y are significantly different. Consider model KBDG-eq and fully informative data: of 200 simulations, the cladogram collapsed to the correct model (CPQ–DN) 157 times and to the null model 37 times. Of the remaining six simulations, the model collapsed twice to (CPQ–D–N) and once each to (C–PQ–DN), (CP–QDN), (CP–Q–DN), and (CPQN–D). Thus, in nearly every case for which the method had power to detect a deviation from the null, the resulting collapsed cladogram guides the researcher to search for deleterious mutations within the appropriate clade of haplotypes.

With max-TDT, the output indicates only which haplotypes are declared significantly different from the others. For example, we might find (B), (B, E), (B, C, E), (A, B), or any other subset of haplotypes are significant. Upon closer inspection, if several haplotypes are significant, it can be determined which appear to be deleterious and which are protective. However, in our simulation, we find 19 different patterns of significant haplotypes from data generated under case 6, and no pattern is dominant. Consequently, had two different research teams investigated the same region, they would have been likely to report different results.

Table 3 summarizes the results of the ET-TDT for the 14 models investigated. Although many collapses of the cladogram are possible, the results are summarized into categories. A fraction of the outcomes result in an essentially correct cladogram, which is subdivided into two levels, excellent or underinclusive. An outcome is excellent if it is either the true cladogram or if haplotypes with similar relative risk are pooled together. An outcome is underinclusive if at least one of the haplotypes with enhanced relative risk is identified, and others are grouped with null haplotypes. The remaining, significant outcomes are subdivided into two categories: informative and noninformative. Informative outcomes provide some guide for the discovery of liability alleles. For example, consider model CBG-eq. The correct cladogram, classified as excellent, is C–PQ–N–D. The outcome CPQ–N–D is underinclusive, because it failed to detect the enhanced risk associated with haplotype C. Outcome C–PQ–ND is considered informative, even though haplotype D is incorrectly grouped with null haplotypes. Finally, outcome CPQN–D is considered noninformative, because it provides no guidance about which haplotypes are more likely to possess enhanced risk. For many models, ET-TDT performs admirably (table 3), essentially finding the correct cladogram for a majority of the cases. Some models, such as DH-rec, are inherently difficult to interpret. For example, because haplotype H is rare, it is pooled with A, and therefore the majority of the outcomes are classified as underinclusive.

Discussion

We present a method, which we call "ET-TDT," to test for excess transmission of haplotypes from parents to affected children. The ET-TDT uses the framework of the standard TDT for individual polymorphisms (Spielman et al. 1993) and extends it to the transmission of entire haplotypes. Several tests for excess haplotype transmission are compatible with the TDT framework (Wilson 1997; Lazzeroni and Lange 1998; Merriman et al. 1998; Clayton 1999; Clayton and Jones 1999; Dudbridge et al. 2000; Rabinowitz and Laird 2000; Zhao et al. 2000). The ET-TDT is distinguished by its use of the evolutionary relationships among the observed haplotypes, as depicted by the cladogram, to produce an

Table 3

Summary of Collapses of the Cladograms for ET-TDT Based on the Simulations Reported in Table 2

	Correct Cladogram		Other		
Model	Excellent[a]	Underinclusive[b]	Informative[c]	Noninformative	Null
C	156	0	7	1	36
BD-eq	121	2	1	1	75
CE-eq	130	4	1	2	61
KBDG-eq	121	0	3	0	75
BD-neq	143	4	3	1	49
CE-neq	127	8	1	0	64
KBDG-neq	144	19	0	1	39
CBG-eq	14	94	49	0	36
CBG-neq	17	69	47	20	47
DH-rec	0	143	2	10	45
EFG-rec	2	94	50	18	36
BK-rec	106	0	51	16	27
KBDG-clad	88	44	58	0	10

[a] "Excellent" indicates that the outcome is the true cladogram or that haplotypes with similar relative risk are pooled together.

[b] "Underinclusive" indicates that at least one of the haplotypes with enhanced relative risk is identified, whereas others are grouped with null haplotypes.

[c] "Informative" indicates that the outcome provides some guide for the discovery of liability alleles.

organized series of tests of differential transmission. The goal of these tests is to group—on the basis of their evolutionary relationships—haplotypes that have indistinguishable transmission rates. Readers will recognize immediately that the ET-TDT also inherits the concepts of MHA (Templeton et al. 1987, 1988, 1992; Templeton and Sing 1993; Hallman et al. 1994; Templeton 1995).

The key assumption of MHA is that "if an undetected mutation with a phenotypic effect occurred at some point in the evolutionary history of the population, it would be embedded within the same historical structure represented by the cladogram" (Templeton et al. 1987; see fig. 1 and 2 in the present report). MHA has two notable features: (1) the distribution of phenotypes within the cladogram directs the search for causal polymorphisms, and (2) the evolutionary relationships can be used to direct nested analyses to reduce the number of comparisons, thereby increasing the power. MHA has enjoyed success in identifying genes affecting liability for coronary heart disease and other quantitative phenotypes (e.g., Keavney et al. 1998), but it can also be used in conjunction with standard case-control analyses (Templeton 1995).

The sine qua non of MHA is the cladogram structure, which is usually constructed using maximum parsimony or maximum likelihood. Programs to develop these cladograms are readily available (e.g., PAUP; Swofford 1998). Conditional on the cladogram, the procedures for MHA are elaborated in detail in various reports (Templeton et al. 1987, 1988, 1992; Templeton and Sing 1993; Hallman et al. 1994; Templeton 1995). Those articles describe how (1) to adjust for cladogram uncertainty and the possibility of recombinant haplotypes, (2) to assess phenotypic associations of individuals who represent two haplotypes, (3) and to perform permutation tests when parametric assumptions fail.

Because ET-TDT also relies on the cladogram, the caveats for MHA extend to ET-TDT. Foremost among these caveats, the estimated cladogram should reflect the true evolutionary relationships among haplotypes to the extent possible. Thus, recombination and gene conversion in the region should be rare. When recombination is common, linkage disequilibrium will be small, and the evolution of haplotypes will not be reflected by the cladogram. As was recommended by Templeton et al. (1987), researchers will want to evaluate the cladogram to determine whether this key assumption is plausible. If it is not, the genomic region should be examined to determine, a priori, whether subdivisions of the region yield results congruent with the assumption. There is no statistical penalty for this exploration, because it involves no hypothesis test. In the same way, recurrent disease mutations violate the key assumption of ET-TDT, because the cladogram will not represent the evolutionary process. In this case, however, we see no easy solution.

Haplotype uncertainty also has an impact on ET-TDT. For haplotype uncertainty, our recommendations follow Templeton et al. (1992), with some adaptations of our own. We propose an algorithm, using Clark's (1990) method and the developing cladogram structure, to infer parental haplotypes. When this algorithm is

used, only families presenting with highly unusual haplotypes should be eliminated from the analysis. To account for families whose multilocus genotypes are compatible with various sets of haplotypes, we develop a partial-likelihood model similar to that introduced by Clayton (1999).

The simulations described in the present report evaluated the effect of haplotype and cladogram uncertainty on the performance of the ET-TDT. Although uncertainty clearly affects power, the test remains valid. For the scenarios we examined, the decrease in power in the face of uncertainty is not large. In fact, often ET-TDT has greater power despite the uncertainty (table 2). In addition, ET-TDT offers a distinct advantage over other methods, namely, its potential for greater interpretability and deeper inference. Ultimately, to identify liability (or protective) alleles, researchers want to know what haplotypes carry alleles imparting greater disease liability. Results of our simulation show that other inferential methods, although powerful, often produce misleading information about which haplotypes determine liability. By contrast, when the evolutionary model is correct, ET-TDT often yields the correct information (table 3). Thus ET-TDT should be a valuable aid for causal inference.

Acknowledgments

This research was supported by National Institute of Health grants MH57881 and DA11922 and National Science Foundation grants DMS9803433 and DMS9819950. We thank Dan Weeks for useful discussions about this project.

Appendix A

The Likelihood

Consider a trio of father, mother, and child ($pq \times rs \rightarrow pr$) in which the child has the disease of interest. The penetrance for a child of multilocus genotype pr is $\Pr(\mathrm{CD} = 1|\mathrm{CG} = pr) \propto \theta_p\theta_r$. For any given trio with parental genotype $pq \times rs$, the probability of disease in the child is

$$\Pr(\mathrm{CD} = 1|\mathrm{PG} = pq \times rs) = \sum_g \{\Pr(\mathrm{CG} = g|\mathrm{PG}) \Pr(\mathrm{CD} = 1|PG,\ \mathrm{CG} = \mathrm{g})\} \propto (\theta_p + \theta_q)(\theta_r + \theta_s) \ .$$

The marginal probability of disease in the child is

$$\Pr(\mathrm{CD} = 1) = \sum_i \sum_j \Pr(\mathrm{CG} = ij) \Pr(\mathrm{CD} = 1|\mathrm{CG} = ij) \propto \left(\sum_{i=0}^{K-1} \psi_i\theta_i\right)^2 .$$

Combining these equations gives

$$L^{(P)} \propto \Pr(\mathrm{PG} = pq \times rs|\mathrm{CD} = 1) \propto \frac{\psi_p\psi_q\psi_r\psi_s(\theta_p + \theta_q)(\theta_r + \theta_s)}{\left(\sum_{i=0}^{K-1} \psi_i\theta_i\right)^2} \ .$$

The conditional probability of child pr, given that the child has the disease and the parents are $pq \times rs$, is

$$L^{(C)} \propto \Pr(\mathrm{CG} = pr|\mathrm{PG} = pq \times rs,\ \mathrm{CD} = 1) \propto \frac{\theta_p\theta_r}{(\theta_p + \theta_q)(\theta_r + \theta_s)} \ .$$

Therefore, the full likelihood is

$$L^{(F)} = L^{(P)}L^{(C)} = \psi_p^*\psi_q\psi_r^*\psi_s \ .$$

Appendix B

Inference from $L^{(C)}$

For the first step of the cladogram-collapsing algorithm, we wish to test whether an external node, *T,* has the same risk as an internal node, *S*. $L^{(C)}$ is parameterized by M parameters, but only $M - 1$ are identifiable. To ensure identifiability, set $\beta_S = 0$. In addition, let $\beta_T = \delta$. Under the reduced model $\delta = 0$, but, under the full model, δ is unconstrained. Split the vector β into two components: $\beta = (\delta,\boldsymbol{\eta}^u)$, where $\boldsymbol{\eta}$ consists of the remaining set of βs. As one moves through the steps of the cladogram-collapsing algorithm, different nodes (clades) will take on the parameter δ.

Define $\mathbf{u}_{(i)} = \partial l^{(C)}/\partial\beta_i$, and let $\mathbf{u} = \sum_{i=1}^{n}\mathbf{u}_{(i)}$ be the score vector, where i indexes the trios in the study. Similarly, let $\mathbf{J}$ denote the matrix of negative second derivatives of $\ell^{(C)}$. For the purpose of conducting inferences, $\mathbf{u}$ and $\mathbf{J}$ are naturally partitioned by $(\delta,\boldsymbol{\eta})$. To test $\delta = 0$ when $\boldsymbol{\eta}$ is unknown, we require an estimate of these nuisance parameters, which is obtained by maximizing the likelihood when $\delta = 0$. Call this constrained maximum $\hat{\boldsymbol{\eta}}_0$.

The score for δ evaluated at $(\delta = 0,\hat{\boldsymbol{\eta}}_0)$ is approximately equal to $\mathbf{u}_\delta(0,\hat{\boldsymbol{\eta}}_0) \approx \mathbf{u}_\delta - \mathbf{J}_{\delta\eta}\mathbf{J}_{\eta\eta}^{-1}\mathbf{u}_\eta$. The score test, which is asymptotically χ_1^2 under the null hypothesis, is of the form $\mathbf{u}_\delta^2(0,\hat{\boldsymbol{\eta}}_0)/\tilde{\mathbf{V}}_{\delta\delta}$, where $\tilde{\mathbf{V}}_{\delta\delta} = \mathbf{J}_{\delta\delta} - \mathbf{J}_{\delta\eta}\mathbf{J}_{\eta\eta}^{-1}\mathbf{J}_{\eta\delta}$ is also evaluated at $(\delta = 0,\hat{\boldsymbol{\eta}}_0)$.

This test procedure reflects the formulas for the first step of the algorithm before any collapsing has occurred. In general, we want to test whether $\beta_T \equiv \beta_0^u \equiv \delta = 0$, in which δ is the parameter for haplotype(s) T, and the β parameter(s) for haplotype(s) S is (are) fixed at 0. Parameters $(\beta_1^u, \ldots, \beta_{R-1}^u)$ correspond to the remaining haplotypes clustered into $R - 1$ clades, where each node within a clade is constrained to have identical relative risk. Let $\mathcal{H}_i$ represent the one or more haplotypes corresponding to β_i^u for $i \in \{0, \ldots, R - 1\}$, so that $\beta_{\mathcal{H}_i}$ may represent several parameters all constrained to be equal. Conversely, let $\mathcal{H}^{-1}(i)$ represent the position of β_i in $\boldsymbol{\beta}^u$ for $i \in \{0, \ldots, M - 1\}$. For convenience, define $\mathcal{H}^{-1}(i) = -1$ if $i \in S$. Note that $\mathcal{H}_0$ is equal to the haplotype(s) T, and $\mathcal{H}^{-1}(i) = 0$, $\forall i \in T$.

If a trio has haplotypes, $pq \times rs \rightarrow pr$, then, using $\theta_i = e^{\beta_i}$,

$$\ell^{(C)} = \log(\theta_p\theta_r) - \log(\theta_p + \theta_q) - \log(\theta_r + \theta_s) = \log(\theta^u_{\mathcal{H}^{-1}(p)}\theta^u_{\mathcal{H}^{-1}(r)}) - \log(\theta^u_{\mathcal{H}^{-1}(p)} + \theta^u_{\mathcal{H}^{-1}(q)}) - \log(\theta^u_{\mathcal{H}^{-1}(r)} + \theta^u_{\mathcal{H}^{-1}(s)})$$

and $\mathbf{u}^{(C)}$, the score for the child likelihood is of the form

$$u^{(C)}_{\beta_i^u} = \frac{\partial\ell^{(C)}}{\partial\beta_i^u} = T_i - E_i\ ,$$

where T_i is the number of times haplotype(s) $\mathcal{H}_i$ is (are) transmitted to the child from among the four parental haplotypes, and E_i is the expected number of times haplotype(s) $\mathcal{H}_i$ is (are) transmitted to the child.

$$E_i = \frac{\Delta_{i\mathcal{H}^{-1}(p)}\theta^u_{\mathcal{H}^{-1}(p)} + \Delta_{i\mathcal{H}^{-1}(q)}\theta^u_{\mathcal{H}^{-1}(q)}}{\theta^u_{\mathcal{H}^{-1}(p)} + \theta^u_{\mathcal{H}^{-1}(q)}} + \frac{\Delta_{i\mathcal{H}^{-1}(r)}\theta^u_{\mathcal{H}^{-1}(r)} + \Delta_{i\mathcal{H}^{-1}(s)}\theta^u_{\mathcal{H}^{-1}(s)}}{\theta^u_{\mathcal{H}^{-1}(r)} + \theta^u_{\mathcal{H}^{-1}(s)}}\ , \tag{B1}$$

where $\theta^u_{-1} \equiv 1$ and $\Delta_{xy} = 0$ if $x \neq y$ and $\Delta_{xy} = 1$ if $x = y$. In summary, for $i \in \{0, \ldots, R - 1\}$, the contribution of any given trio to $u_i^{(C)}$ is $T_i - E_i$ where both E_i and T_i are 0–2.

To compute

$$\mathbf{J}^{(C)}_{\beta^u\beta^u} = \left[-\frac{\partial^2\ell^{(C)}}{\partial\beta_i^u\beta_j^u}\right],$$

notice

$$-\frac{\partial^2\ell^{(C)}}{\partial\beta_i^u\beta_j^u} = -\frac{\partial u^{(C)}_{\beta_i^u}}{\partial\beta_j^u} = -\frac{\partial E_i}{\partial\beta_j^u}\ .$$

This quantity can be determined by summing the derivatives of the separate contributions to E_i resulting from each parent. Note that the only three possibilities are that one parent's contribution to E_i is 0, 1, or of the form

$$\frac{\theta_x^u}{\theta_x^u + \theta_y^u} ,$$

the last occurring when $\mathcal{H}^{-1}(x) \neq \mathcal{H}^{-1}(y)$ and either $\mathcal{H}^{-1}(x)$ or $\mathcal{H}^{-1}(y)$ is equal to i.

For $i,j \in \{0, \ldots, R-1\}$ the contribution(s) of any given trio to

$$-\frac{\partial u_{\beta_i}^{(C)}}{\partial \beta_j^u}$$

is (are) given here for the two haplotypes of either parent. This is then repeated for the other parent. Find $i = \mathcal{H}^{-1}(p)$ and $j = \mathcal{H}^{-1}(q)$. If $i = j$, no change is made to J. If $i \neq j$, calculate

$$v = \frac{\theta_i^u \theta_j^u}{(\theta_i^u + \theta_j^u)^2} ,$$

where $\theta_i^u = 1$ if p∈S and $\theta_j^u = 1$ if q∈S. If neither $p \in S$ nor $q \in S$, add v to J_{ii} and J_{jj} and subtract v from J_{ij} and J_{ji}. If one of the haplotypes is in S and the other corresponds to θ_k^u, the only change made to $\mathbf{J}$ is to add

$$\frac{\theta_k^u}{(1 + \theta_k^u)^2}$$

to J_{kk}.

Appendix C

Inference with Missing Data

With missing data, we require an estimate of the haplotype frequencies in the population. Define $\psi_i = e^{\gamma_i} / \sum_j e^{\gamma_i}$. We set $\gamma_1 = 0$ to ensure identifiability. As in the Appendix B, the dimension of the parameter space changes as $\boldsymbol{\beta}^u$ and $\boldsymbol{\eta}^u$ are defined by the cladogram collapsing algorithm. The set of nuisance parameters in the model is expanded to $\boldsymbol{\lambda} = (\boldsymbol{\eta}^u, \boldsymbol{\gamma})$. For $j \in \mathcal{P}_i$, write

$$\mathbf{u}_{(j)} = \begin{pmatrix} \mathbf{u}_{\beta(j)}^{(C)} \\ \mathbf{u}_{\gamma(j)}^{(P)} \end{pmatrix} = \begin{pmatrix} \dfrac{\partial l_{(j)}^{(C)}}{\partial \boldsymbol{\beta}^u} \\ \dfrac{\partial l_{(j)}^{(P)}}{\partial \boldsymbol{\gamma}} \end{pmatrix} .$$

The total partial score for the ith family is obtained by taking a weighted average of the partial score obtained for each consistent scenario:

$$\mathbf{u}_{\mathcal{P}_i} = \frac{\sum_{j \in \mathcal{P}_i} L_{(j)}^{(F)} \mathbf{u}_{(j)}}{\sum_{j \in \mathcal{P}_i} L_{(j)}^{(F)}} .$$

The total partial-score vector, $\mathbf{u}$, is obtained by summing the contributions over all families. For inferences, the natural partitioning is $\mathbf{u} = (u_\delta, \mathbf{u}_\lambda)^T$. The matrix of negative second derivatives for a single configuration is

$$\mathbf{J}_{(j)} = \begin{bmatrix} \mathbf{J}_{\beta^u \beta^u(j)}^{(C)} & 0 \\ \mathbf{J}_{\gamma \beta^u(j)}^{(P)} & \mathbf{J}_{\gamma\gamma(j)}^{(P)} \end{bmatrix} .$$

As described by Clayton (1999), the corresponding term for the entire set of allowable configurations for the ith subject $\mathcal{P}_i$ is

$$\mathbf{J}_{\mathcal{P}_i} = \frac{\sum_{j \in \mathcal{P}_i} L_{(j)}^{(F)} \mathbf{J}_{(j)}}{\sum L_{(j)}^{(F)}} - \left\{ \frac{\sum_{j \in \mathcal{P}_i} L_{(j)}^{(F)} \mathbf{u}_{(j)} (\mathbf{u}_{(j)})^T}{\sum_{j \in \mathcal{P}_i} L_{(j)}^{(F)}} - \mathbf{u}_{\mathcal{P}_i} (\mathbf{u}_{\mathcal{P}_i})^T \right\} .$$

Finally, $\mathbf{J} = \sum_i \mathbf{J}_{\mathcal{P}_i}$. The variance of $\mathbf{u}_{\mathcal{P}_i}$ can be computed empirically using $\mathbf{V} = \sum_i \mathbf{u}_{\mathcal{P}_i} (\mathbf{u}_{\mathcal{P}_i})^T - \frac{1}{N} \mathbf{u}\mathbf{u}^T$. For inferences, $\mathbf{J}$ and $\mathbf{V}$ are partitioned by $(\delta, \boldsymbol{\lambda})$.

The standard arguments used to obtain the score test for complete data also apply for partial-likelihood models. It can be shown that the score test for testing $\delta = 0$ in the presence of the nuisance parameter $\boldsymbol{\lambda}$ is

$$\mathbf{u}_\delta^2(0, \hat{\boldsymbol{\lambda}}_0) / \tilde{\mathbf{V}}_{\delta\delta} ,$$

where

$$\tilde{\mathbf{V}}_{\delta\delta} = \mathbf{V}_{\delta\delta} + \mathbf{J}_{\delta\lambda} \mathbf{J}_{\lambda\lambda}^{-1} \mathbf{V}_{\lambda\lambda} (\mathbf{J}_{\lambda\lambda}^{-1})^T (\mathbf{J}_{\delta\lambda})^T - \mathbf{J}_{\delta\lambda} \mathbf{J}_{\lambda\lambda}^{-1} \mathbf{V}_{\lambda\delta} - \mathbf{V}_{\delta\lambda} (\mathbf{J}_{\lambda\lambda}^{-1})^T (\mathbf{J}_{\delta\lambda})^T .$$

The latter term is also evaluated at $(\delta = 0, \hat{\boldsymbol{\lambda}}_0)$. As before, the score test is a 1-df test, which is asymptotically distributed as a χ^2_1 under the null hypothesis.

Expressions for $\mathbf{u}^{(C)}$ and $\mathbf{J}^{(C)}$ are given in Appendix B. The score for the parental likelihood $\mathbf{u}^{(P)}$, is of the form $u^{(P)}_{\gamma_i} = N_i - 2\psi_i - 2\psi^*_i$, where N_i is the number of parental haplotypes equal to i. To summarize, for $i \in \{1, \ldots, M-1\}$, the contribution of any given trio to $u^{(P)}_i$ is $N_i - 2\psi_i - 2\psi^*_i$ where N_i is 0–4.

The remaining terms,

$$\mathbf{J}^{(P)}_{\gamma\gamma} = \left[-\frac{\partial^2 \ell^{(P)}}{\partial\gamma_i\gamma_j}\right],$$

are simple to compute.

$$\mathbf{J}^{(P)}_{\gamma_i\gamma_j} = -\frac{\partial \mathbf{u}^{(P)}_{\gamma_i}}{\partial\gamma_j} = 2\left[\Delta_{ij}(\psi_i + \psi^*_i) - (\psi_i\psi_j + \psi^*_i\psi^*_j)\right];$$

that is, regardless of the haplotypes present, each trio contributes $2[\Delta_{ij}(\psi_i + \psi^*_i) - (\psi_i\psi_j + \psi^*_i\psi^*_j)]$ to $\mathbf{J}^{(P)}_{\gamma_i\gamma_j}$, for $i,j \in \{1, \ldots, K-1\}$. Similarly,

$$\mathbf{J}^{(P)}_{\gamma\beta^u} = \left[-\frac{\partial^2 \ell^{(P)}}{\partial\gamma_i\beta^u_j}\right],$$

and

$$\mathbf{J}^{(P)}_{\gamma_i\beta^u_j} = -\frac{\partial^2 \ell^{(P)}}{\partial\gamma_i\beta^u_j} = -\frac{\partial u^{(P)}_{\gamma_i}}{\partial\beta^u_j} = \frac{\partial}{\partial\beta^u_j}(2\psi^*_i)$$

$$= 2\frac{I(i \in \mathcal{H}_j)\psi_i\theta_i}{\sum_{k=0}^{K-1}\psi_k\theta_k} - 2\frac{\left[\sum_{k=0}^{K-1} I(k \in \mathcal{H}_j)\psi_k\theta_k(\psi_i\theta_i)\right]}{\left(\sum_{k=0}^{K-1}\psi_k\theta_k\right)^2},$$

for $i \in \{1, \ldots, M-1\}$, $j \in \{0, \ldots, R-1\}$, for each trio, regardless of the haplotypes present.

Appendix D

Estimating Equations

To estimate $\boldsymbol{\lambda}$ the objective is to simultaneously solve for $\boldsymbol{\eta}^u$ and $\boldsymbol{\gamma}$ such that

$$\mathbf{u}^{(C)}_{\eta} = 0 \quad \text{and} \quad \mathbf{u}^{(P)}_{\gamma} = 0\,. \tag{D1}$$

Both equations depend upon $\boldsymbol{\gamma}$ and $\boldsymbol{\eta}^u$ in the form of weights, $L^{(F)}$. With an iterative, EM-algorithm approach, these weights are taken to be fixed at the solution obtained in the previous iteration. For the ith family, let $j \in \mathcal{P}_i$ index one consistent haplotype configuration. Let (p_j,q_j) and (r_j,s_j) denote the parental haplotypes and let (p_j,r_j) denote the offspring's haplotypes for this configuration. Here, we describe the EM algorithm for obtaining the solutions to equation D1.

To estimate $\boldsymbol{\gamma}$: let $N^{(j)}_{ki}$ be the number of times haplotype k is observed in the parents, within the jth consistent configuration. Set $\psi_k = \psi_k\theta^{(m)}_k / \sum_j \psi_j\theta^{(m)}_j$, where $\theta^{(m)}$ is the estimate obtained from the mth step of the algorithm. Given $(\delta = 0, \hat{\boldsymbol{\eta}}^{(m)}, \hat{\boldsymbol{\gamma}}^{(m)})$, also obtained from the mth step of the algorithm, solve:

$$\sum_i \frac{\sum_{j\in\mathcal{P}_i} \psi^{*(m)}_{p_k}\psi^{(m)}_{q_k}\psi^{*(m)}_{r_j}\psi^{(m)}_{s_j}(N^{(j)}_{ki} - 2\psi_k - 2\psi^*_k)}{\sum_{j\in\mathcal{P}_i} \psi^{*(m)}_{p_j}\psi^{(m)}_{q_j}\psi^{*(m)}_{r_j}\psi^{(m)}_{s_j}} = 0\,,$$

for $k = 1, \ldots, H$.

To estimate $\boldsymbol{\eta}$: Let $T^{(j)}_k$ be the number of times haplotype k is transmitted to the offspring, within the jth consistent configuration. For every node except those involved in the hypothesis test solve:

$$\sum_i \frac{\sum_{j\in\mathcal{P}_i} \psi^{*(m)}_{p_j}\psi^{(m)}_{q_j}\psi^{*(m)}_{r_j}\psi^{(m)}_{s_j}(T^{(j)}_k - E^{(j)}_k)}{\sum_{j\in\mathcal{P}_i} \psi^{*(m)}_{p_j}\psi^{(m)}_{q_j}\psi^{*(m)}_{r_j}\psi^{(m)}_{s_j}} = 0\,,$$

where $E^{(j)}_k$ is the expected value (at the mth step of the algorithm) under configuration j; see equation B1, Appendix B.

Electronic-Database Information

Accession numbers and URLs for data in this article are as follows:

Computational Genetics, http://wpicr.wpic.pitt.edu/WPICCompGen/, or K.R.'s Web site, http://www.stat.cmu.edu/~roeder/ (for a program to perform the ET-TDT analyses).

References

Cervino AC, Hill AV (2000) Comparison of tests for association and linkage in incomplete families. Am J Hum Genet 67:120–132

Clark AG (1990) Inferences of haplotypes from PCR-amplified samples of diploid populations. Mol Biol Evol 7:111–122

Clayton DG (1999) A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. Am J Hum Genet 65:1170–1177

Clayton DG, Jones H (1999) Transmission/disequilibrium tests for extended marker haplotypes. Am J Hum Genet 65:1161–1169

Dudbridge F, Koeleman BPC, Todd JA, Clayton DG (2000) Unbiased application of the transmission/disequilibrium test to multilocus haplotypes. Am J Hum Genet 66:2009–2012

Ewens WJ, Spielman RS (1995) The transmission/disequilib-

rium test: history subdivision and admixture. Am J Hum Genet 57:455–464

Falk CT, Rubinstein P (1987) Haplotype relative risk: an easy way to construct a proper control sample for risk calculations. Ann Hum Genet 51:227–233

Hallman DM, Visvikis S, Steinmetz J, Boerwinkle E (1994). The effect of variation in the apolipoprotein B gene on plasma lipid and apolipoprotein B levels. Ann Hum Genet 58:35–64

Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D and Antonovics J (eds) Oxford surveys in evolutionary biology. Vol 7. Oxford University Press, New York, pp 1–44

Keavney B, McKenszie CA, Connell JM, Julier C, Ratcliffe PJ, Sobel E, Lathrop M, Farrall M (1998) Measured haplotype analysis of the angiotensin-I converting enzyme gene. Hum Mol Genet 7:1745–1751

Kingman JFC (1982) On the genealogy of large populations. J Appl Prob 19A:27–43

Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction–combined disequilibrium test. Am J Hum Genet 64: 861–870

Lazzeroni LC, Lange K (1998) A conditional inference framework for extending the transmission/disequilibrium test. Hum Hered 48:67–81

Lange K (1997) Mathematical and statistical methods for genetic analysis. Springer-Verlag, New York, pp 105–107

Merriman TR, Eaves IA, Twells RC, Merriman ME, Danoy PA, Muxworthy CE, Hunter KM, Cox RD, Cucca F, McKinney PA, Shield JP, Baum JD, Tuomilehto J, Tuomilehto-Wolf E, Ionesco-Tirgoviste C, Joner G, Thorsby E, Undlien DE, Pociot F, Nerup J, Ronningen KS, Bain SC, Todd JA (1998) Transmission of haplotypes of microsatellite markers rather than single marker alleles in the mapping of a putative type 1 diabetes susceptibility gene (IDDM6). Hum Mol Genet 7:517–524

Morris AP, Curnow RN, Whittaker JC (1997) Randomization tests of disease-marker associations. Ann Hum Genet 61: 49–60

Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Hum Hered 50:211–223

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

Sinsheimer JS, Blangero J, Lange K (2000) Gamete-competition models. Am J Hum Genet 66:1168–1172

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

Swofford DL (1998) PAUP: phylogenetic analysis using parsimony, release 4.0b1. Sinauer Associates, Sunderland, MA

Swofford DL, Olsen GJ, Waddell PJ and Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz C, Mabel BK (eds). Molecular systematics, 2nd ed. Sinauer Associates, Sunderland, MA, pp 407–514

Templeton AR (1995) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping or DNA sequencing. V. Analysis of case/control sampling designs: Alzheimer's disease and the apoprotein E locus. Genetics 140:403–409

Templeton AR, Boerwinkle E, Sing CF (1987) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. I. Basic theory and an analysis of alcohol dehydrogenase activity in *Drosophila*. Genetics 117:343–351

Templeton AR, Crandall KA, Sing CF (1992) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping and DNA sequence data. III. Cladogram estimation. Genetics 132:619–633

Templeton AR, Sing CF, Kessling A, Humphries S (1988) A cladistic analysis of phenotype associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. Genetics 120:1145–1154

Templeton AR, Sing CF (1993) A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. IV. Nested analyses with cladogram uncertainty and recombination. Genetics 134:659–669

Terwilliger JD, Ott J (1992) A haplotype-based "haplotype relative risk" approach to detecting allelic associations. Hum Hered 42:337–346

Wilson SR (1997) On extending the transmission/disequilibrium test (TDT). Ann Hum Genet 61:151–161

Zhao H, Zhang S, Merikangas KR, Trixler M, Wildenauer DB, Sun F, Kidd KK (2000) Transmission/disequilibrium tests using multiple tightly linked markers. Am J Hum Genet 67:936–946