

Evolutionary-based Association Analysis Using Haplotype Data

Howard Seltman,¹ Kathryn Roeder,¹ and B. Devlin^{2*}

¹Department of Statistics, Carnegie Mellon University, Pittsburgh, Pennsylvania

²Department of Psychiatry, University of Pittsburgh, Pittsburgh, Pennsylvania

Association studies, both family-based and population-based, can be powerful means of detecting disease-liability alleles. To increase the information of the test, various researchers have proposed targeting haplotypes. The larger number of haplotypes, however, relative to alleles at individual loci, could 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, much as is done with cladistic-based association analysis. First suggested by Templeton et al. ([1987] *Genetics* 117:343–351), such analyses use the evolutionary relationships among haplotypes to produce a limited set of hypothesis tests and to increase the interpretability of these tests. To more fully utilize the information contained in the evolutionary relationships among haplotypes and in the sample, we propose generalized linear models (GLM) for the analysis of data from family-based and population-based studies. These models fully account for haplotype phase ambiguity and allow for covariates. The models are encoded into a software package (the Evolutionary-Based Haplotype Analysis Package, EHAP), which also provides for various kinds of exploratory data analysis. The exploratory analyses, such as error checking, estimation of haplotype frequencies, and tools for building cladograms, should facilitate the implementation of cladistic-based association analysis with haplotypes. *Genet Epidemiol* 25:48–58, 2003. © 2003 Wiley-Liss, Inc.

Key words: cladistic analysis; coalescent; family-based study; measured haplotype analysis; score test

Grant sponsor: National Institutes of Health; Grant numbers: MH057781, CA-54852-07; DA011922; Grant Sponsor: National Science Foundation; Grant number: DMS-9803433.

*Correspondence to: Bernie Devlin, Department of Psychiatry, University of Pittsburgh School of Medicine, 3811 O'Hara St., Pittsburgh, PA 15213. E-mail: devlinbj@msx.upmc.edu

Received for publication 19 September 2002; Revision accepted 16 December 2002

Published online in Wiley InterScience (www.interscience.wiley.com).

DOI: 10.1002/gepi.10246

INTRODUCTION

The evolutionary history of a sample of haplotypes can be represented by a coalescent process [Kingman, 1982], with disease liability mutations embedded in the evolution of the distinct haplotype forms. Noting this fact, Templeton et al. [1987] conjectured that the power to discover liability alleles is often enhanced by using a cladogram to guide the search for an association between a trait and one or more clusters of similar haplotypes. Although this conjecture has been verified in both in empirical studies [Havililad et al., 1997; Keavney et al., 1998; Soubrier et al., 2002] and in simulations [Seltman et al., 2001], cladistic analysis depends on the accurate reconstruction of evolutionary relationships. The recent discovery that haplotype blocks are often conserved, even across ethnic groups [Gabriel et al., 2002], supports the cladistic approach. Indeed, haplotype-based methods that exploit

the historical development of distinct haplotype forms appear to offer a very promising approach to disease gene mapping.

Support for the analysis of haplotypes, rather than single genetic markers, can also be derived directly from a functional viewpoint. Evidence is mounting that multiple mutations within a gene, occurring on the same chromosome, can have a large effect on the phenotype [e.g., Hollox et al., 2001; Clark et al., 1998; Tavtigian et al., 2001]. These large haplotype effects are the natural biological consequence of the induced change in the protein product. The total effect of such wholesale changes cannot be detected by sequential tests for effects in a series of polymorphisms. Moreover, even if the causal single nucleotide polymorphisms (SNPs) are not measured, they are embedded in the haplotypes formed by markers spanning the region, making haplotype analysis more sensitive than a series of single SNP association tests.

Recently, numerous methods for analyzing haplotype data have appeared, most of them targeted at distinct types of study designs. For instance, for family-based designs in which parents and their affected offspring are sampled, various transmission disequilibrium tests (TDT) exist [Clayton, 1999; Clayton and Jones, 1999; Seltman et al., 2001]. For population-based designs, particularly case-control studies, a likelihood ratio test is often used [e.g., Sham, 1998]. Alternatively, a score approach, which also models the effect of environmental covariates, can be applied to both quantitative and qualitative traits [Schaid et al., 2002].

In practice, even when family data are collected, haplotype-based methods must incorporate the effect of ambiguity in haplotype phase. This uncertainty can be handled by treating the phase as “missing data” [Clayton, 1999; Schaid et al., 2002] and estimating the frequency distribution of the haplotypes by likelihood-based methods. From this comes the probability that the observed multilocus genotype resolves into each pair of haplotypes consistent with the genotype. These probabilities are used to weight the likelihood contribution for each consistent configuration. Regardless of the planned statistical analysis, the more family members genotyped, the greater the restrictions on the haplotypes consistent with the multilocus genotypes, and hence the greater the power of the analysis to detect significant associations between haplotypes and phenotypes.

For a small stretch of DNA, the number of common haplotypes in the population is often quite small [Gabriel et al., 2002]. Still, the number of *potential* haplotypes grows exponentially in the number of markers. Proposed methods for analyzing haplotype data differ in how they cope with the large number of possible haplotypic effects. Approaches range from directed comparisons to omnibus tests [e.g., Clayton, 1999], with the accompanying differences in interpretability. Ideally there would be some way of focusing tests on particular haplotypes or sets of haplotypes, thereby increasing both the power of the test and the interpretability. For example, Clayton and Jones [1999] suggested limiting the number of parameters by using a random-effects model.

The cladistic-based association analysis pioneered by Templeton et al. [1987] is another option. It has two notable features: the

distribution of phenotypes within the cladogram directs the search for causal polymorphisms, while the evolutionary relationships can be used to direct analyses to reduce the number of comparisons, thereby potentially increasing the power. Methods were developed for population-based samples on which either quantitative or qualitative outcomes were measured [Templeton et al., 1987, 1988, 1992; Templeton and Sing, 1993; Templeton, 1995]. Seltman et al. [2001] extended cladistic-based analyses to family-based designs [Spielman et al., 1993], with a single affected offspring sampled from each family to produce the evolutionary tree transmission disequilibrium test (ET-TDT). Unlike the other cladistic procedures mentioned above, this method formally incorporated haplotype uncertainty into the test statistic. Simulations demonstrated that ET-TDT did indeed enhance interpretability and sometimes power over a TDT-based analysis.

Genetic studies, however, do not always fall definitively into family-based or population-based designs. Families consisting primarily of parents and offspring or large sibships are generally analyzed using family-based likelihood models such as the TDT, and families consisting primarily of singletons or small sibships without parents are generally analyzed using population-based likelihood models. By directly allowing for missing data, it is possible to analyze a broad range of study samples with either approach. How such data are analyzed follows from what assumptions the investigators are willing to entertain. For instance, if the population is likely to be relatively homogeneous, then a population-based analysis could be more appealing than a family-based analysis, in the sense that it is likely to be more powerful with little risk of false positives due to population substructure [e.g., Bacanu et al., 2000].

In this article, we set up notation and generalized linear models that facilitate cladistic analysis of both family-based and population-based study designs for quantitative and qualitative responses, including case-control study designs. As in ET-TDT, the tests fully incorporate haplotype uncertainty. The methods also allow for environmental covariates and multiple siblings per family. We also introduce software that permits exploratory data analysis, which should be helpful for directing inferences, data processing, and cleaning.

METHODS AND RESULTS

HAPLOTYPES, CLADISTIC ANALYSES, AND SOFTWARE IMPLEMENTATION

In the ideal scenario for evolutionary-based association analysis, the history of a sample of haplotypes is represented by a simple coalescent process. From an ancestral haplotype, mutations lead to the observed diversity of haplotypes, and the history of these distinct forms can be summarized in a cladogram. For instance, of the 11 haplotypes (A–K) depicted in Figure 1, those connected by an edge (line) differ by a single mutation. Given that haplotype A is the ancestral form, the evolutionary history is apparent. Suppose a candidate gene contains polymorphisms having a direct impact on liability. These polymorphisms may or may not be genotyped; regardless of whether they are, they will be embedded in the evolutionary history of haplotypes comprising the cladogram. For instance, if a

deleterious mutation occurred after the evolution of haplotype B and before haplotype C, then this mutation will be embedded in the clade represented by haplotypes (C,E,F,G). This emphasizes the major advantage of measured haplotype analysis, which is to direct the search for variants having a direct impact on liability through the pattern of association found over related haplotypes.

We next present a worked example of cladistic analysis and a description of new software to implement those analyses, named the Evolutionary-Based Haplotype Analysis Package (EHAP). In this simulated example, we generate a population from which 200 case and 200 control individuals are sampled and two additional covariates measured. Nine SNPs within a candidate gene are genotyped on this sample. The response variable is binary (case/control), and a logit model is fit to the data. The simulation mimics the evolutionary scenario described above, in which a liability mutation occurs on haplotypes

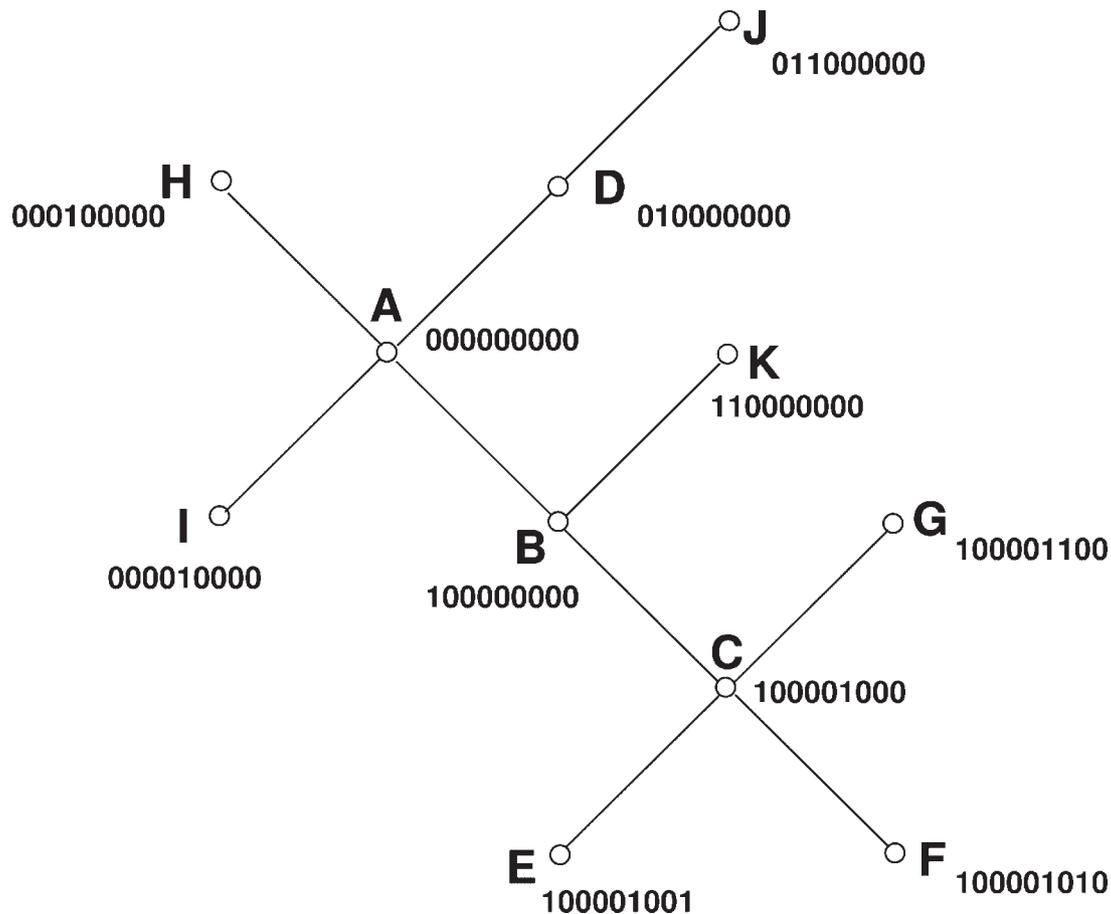


Fig. 1. Cladogram of haplotypes A–K.

(C,E,F,G). For the example, we take the cladogram depicted in Figure 1 as given. After presenting the example, the workings of EHAP will be described in more detail, from the initial input of genotypic data to the production (or input) of the cladogram, and finally to the association analysis.

To perform a cladistic analysis, a cladogram is divided into subgroups (clades) by an algorithm developed by Templeton et al. [1987]. The individual haplotypes occurring as leaves (terminal nodes) on the tree represent 0-step clades. The 1-step clades are produced by moving backward one mutational step from the 0-step clades toward internal nodes. Repeating this procedure, the 1-step clades cluster to produce the 2-step clades, and so forth. For example, in Figure 1, 0-step clades are E, F, G, H, I, J , and K ; 1-step clades are (C, E, F, G) , (B, K) , (A, H, I) , and (D, J) ; 2-step clades are (B, C, E, F, G, K) and (A, H, I, D, J) ; and the 3-step clade is the entire cladogram.

Rather than perform an omnibus test to determine whether the effects of the haplotypes in the cladogram differ significantly, a series of 1 degree-of-freedom (df) tests is performed. Here we describe the sequential testing procedure known as the *cladogram-collapsing algorithm* in the context of Figure 1. At each step in the algorithm, a “full model” is required. The full model is the same within each step, but changes between steps, conditional upon the results of the previous step. Let β_j denote the parameter associated with the phenotypic effect of the j 'th node/haplotype.

0-step. Begin with a full model with distinct parameters at all 11 nodes of the cladogram. Test whether any of the 0-step clades has an effect equal to the nearest internal node. For example, compare the effects of nodes E and C : $H_0: \beta_E = \beta_C$. Each of the seven tests has 1 df, and they are performed independently (Table 1).

1-step. Next, conditional on the previous results, the new full model sets the appropriate parameters as equal. For instance, in this example, no tests are rejected at the 0-step stage; we set $\beta_A = \beta_H = \beta_I = \beta_B = \beta_K$, $\beta_C = \beta_E = \beta_F = \beta_G$, and $\beta_D = \beta_J$. The resulting model has four distinct parameters, β_A , β_B , β_C , β_D . Two tests are performed: one comparing $\beta_B = \beta_C$, and other comparing $\beta_A = \beta_D$; the former hypothesis is rejected (Table 1).

r -step. In general, at the r 'th step, the new full model sets the appropriate parameters as equal. Tests are performed for the equality of the parameters associated with the r -step clades within the

TABLE I. Cladistic Analysis of Simulated Case-Control Data From Cladogram in Figure 1, Using a Likelihood Ratio Test and Logistic Regression^a

	Model	Parms	P-value
0-step	$H_1: A, B, C, D, E, F, G, H, I, J, K$	11	
	$H_0: E=C; A, B, D, F, G, H, I, J, K$	10	0.585
	$H_0: F=C; A, B, D, E, G, H, I, J, K$	10	0.709
	$H_0: G=C; A, B, D, E, F, H, I, J, K$	10	0.636
	$H_0: H=A; B, C, D, E, F, G, I, J, K$	10	1.000
	$H_0: I=A; B, C, D, E, F, G, H, J, K$	10	0.045
	$H_0: J=D; A, B, C, E, F, G, H, I, K$	10	0.499
	$H_0: K=B; A, C, D, E, F, G, H, I, J$	10	0.101
1-step	$H_1: (C, E, F, G), (B, K), (A, H, I), (D, J)$	4	
	$H_0: (C, E, F, G)=(B, K); (A, H, I), (D, J)$	3	0.000000002
	$H_0: (A, H, I)=(D, J); (C, E, F, G), (B, K)$	3	0.082
2-step	$H_1: (A, D, H, I, J), (B, K), (C, E, F, G)$	3	
	$H_0: (A, D, H, I, J)=(B, K); (C, E, F, G)$	2	0.817

^aIn this shorthand, A refers to parameter associated with haplotype A (or effect of A on response variable), B the parameter associated with haplotype B , and so on; “ $A=B; C, D, \dots, K$ ” refers to a test of equality of parameters for haplotypes A and B in a model for which each other specified haplotype, C, D, \dots, K , has a corresponding parameter fitted in GLM model. A set of haplotypes in parentheses, (A, B, C) , indicates the model will be constrained so that haplotypic effect of each node in the set is equal. Parms indicates number of parameters in GLM model for haplotypes under each set of constraints. In this example, the model reduces from 11 parameters to two parameters: (A, B, D, H, I, J, K) and (C, E, F, G) .

$(r+1)$ -step clades. In the example, at the 2-step stage, a single test comparing $\beta_B = \beta_A$ is conducted (Table 1).

To avoid false positives from multiple testing, a Bonferroni correction should be taken to account for the $(M-1)$ tests conducted for a cladogram with M nodes. By default, EHAP implements Bonferroni corrections for its tests. For small samples, a permutation test is preferable to the asymptotic procedures described; for details, see Seltman et al. [2001].

In the example, it follows that in the process of moving through the tests defined by this 11-node cladogram, 10 one-degree-of-freedom tests are conducted, rather than a single omnibus 10-degrees-of-freedom-test. At its conclusion, EHAP produces estimates of the parameters for a model with an intercept term, a single parameter describing the clade C haplotype effects ($\beta_C = 1.19$) and the coefficients for the nongenetic covariates; the coefficient for clade A is set to zero, to ensure identifiability of the model. Overall, we conclude that there is an association with the trait under investigation (P -value = 0.000000002), and the effect is attributed to clade (C, E, F, G) . As the next step in the analysis, it would be desirable to

sequence the haplotypes to discover whether any polymorphisms are associated almost exclusively with this clade. Alternatively, there could be a protective variant associated with the opposite clade.

EHAP

In previous research [Seltman et al., 2001] and in this sequel, we lay out a theoretical framework that uses evolutionary relationships among haplotypes to determine whether certain haplotypes are associated with a liability to disease. Prior to any association analyses, however, certain challenges must be met. For example, due to haplotype uncertainty, some haplotypes could be consistent with observed multilocus genotypes, yet they are either very rare or absent from both the sample and the population from which the sample was drawn. Analysts will have to make a decision about what to do with these haplotypes. One possibility, which we favor, is to treat the haplotypes as if they were absent from the population and, when necessary, discard the data from those few individuals whose multilocus genotypes can only be explained by those rare haplotypes. After establishing the set of extant haplotypes, the parsimonious set of evolutionary relationships among them can be a network instead of a cladogram. The cycles of the network must be broken, and the network simplified to a cladogram before association analysis. Such challenges can be overcome, but they require much from the genetic analyst.

EHAP, a freely available software package, contains an evolving set of tools to address these challenges. At the present time, EHAP is capable of performing all analyses to be described here, as well as ET-TDT analyses described previously [Seltman et al., 2001], and it contains a variety of tools and functions to facilitate evolutionary-based analyses.

EHAP takes data in "prelinkage" format [Ott, 1991], which is standard now for programs such as GeneHunter [Kruglyak et al., 1996]. To infer the set of haplotypes present in the sample, EHAP will produce, for each individual in the sample, the list of all haplotype pairs consistent with the individual's multilocus genotype and that of any other family members who are genotyped. EHAP uses the Lange-Gordia algorithm, as described by O'Connell [2000], which assumes no contemporary recombination. These results, together with an implementation of the EM algorithm, yield

maximum likelihood estimates of haplotype frequencies. Unlike most other software, there are no restrictions on the type of locus comprising the haplotype. The user can eliminate rare haplotypes; EHAP will report which families are inconsistent with the limited set of possible haplotypes. In this way, EHAP also provides general checks for genotype errors because it provides this information for any set of haplotypes, including all possible haplotypes. At the present time, the algorithm for determining haplotypes compatible with multilocus genotypes is not optimized, so large problems can tax both the user's patience and the computer's memory. Optimization of the kind proposed by Niu et al. [2002] is planned for future releases.

On the basis of the set of haplotypes in the sample, or any subset thereof supplied by the user, EHAP produces a distance matrix in terms of weighted differences between haplotypes (default or user-supplied weighting scheme). It then uses the distance matrix to graphically connect similar haplotypes. "Similarity" is defined either using a default standard of connecting haplotypes differing by a single mutation (Fig. 1) or a user-defined criterion. Users can manipulate the graphic for optimal presentation. For our example, the default for EHAP reveals Figure 1, but with a loop that connects haplotypes A, B, K, D. Incorporated into EHAP are algorithms by Crandall and Templeton [1993] for breaking loops in graphs to produce cladograms. Consistent with evolutionary theory that older haplotypes should occupy internal nodes in cladograms, Crandall and Templeton [1993] argued that three key features can be used to establish which edges to break: older haplotypes tend to be more common in the population than are more recently derived haplotypes; older haplotypes also tend to have more unambiguous descendants; and older haplotypes tend to be more centrally located within a network. By these criteria, EHAP breaks the edge between *D* and *K* to reveal an unrooted tree. In future releases of EHAP, additional algorithms will be provided for the purpose of choosing the cladogram; for example, we plan to incorporate recent results about aging haplotypes on the basis of additional molecular variants embedded in the haplotypes [Slatkin and Rannala, 1997]. Users can also input any cladogram, such as one obtained from other software. This feature allows the user to evaluate the robustness of the results by exploring a set of cladograms. By appropriate construction of pseudo-cladograms, the user can also contrast the

effects of haplotypes of interest, even without an evolutionary hypothesis.

In a future release of EHAP, we plan to incorporate the logic of median-joining networks [Bandelt et al., 1999] to fill in any haplotypes required to produce a single cladogram/network. For the present time, users are referred to the Network Analysis software of Bandelt et al. [1999]. Nonetheless, if portions of the cladogram are vastly different, and intervening haplotypes are largely absent, then it would likely be best to analyze the clades separately. EHAP has this option.

Results from the association analysis are presented in pseudographical form, with various statistics (parameter estimates, test statistics, and P -values) interspersed with graphical representation of the results and stage of the clade-collapsing algorithm. Results of haplotype phase and EM estimates of haplotype frequencies are written to files, thus expediting additional analyses of the data.

LIKELIHOOD

Consider a sample of N nuclear families, sibships, or singletons, indexed by i , each with n_i sampled siblings. (Note that we can consider a population sample of size N to be a sample of N families of size 1.) We assume for the moment that the sample is drawn randomly from a population; later, we discuss how nonrandom sampling works into the analysis.

Our objective is to formulate a probability model for the phenotypes and genotypes observed within each family. A natural model is based on the link between an individual's haplotypes and resulting phenotype. The ambiguity due to haplotype phase uncertainty is then accounted for via the missing-data principle. For the i th family, let P_i denote the set of haplotype configurations mutually consistent with the parents and all the sampled offspring, i.e., each entry $l \in P_i$ denotes a set of haplotype pairs consistent with the measured genotypes and satisfying the laws of inheritance. Let $L_{i(l)}$ be the joint likelihood for the phenotype and the l 'th haplotype configuration for the i 'th family. Under certain assumptions about missing data (i.e., data missing at random), L_i can be computed directly from $L_{i(l)}$.

Due to correlation in genotypes and phenotypes among family members, $L_{i(l)}$ cannot be obtained directly by taking the product of the likelihood contributions obtained from each

sibling $L_{ij(l)}$, $j = 1, \dots, n_i$. Consequently, for experimental designs that ascertain multiple siblings per family ($n_i > 1$), test statistics must account for the correlation among siblings. At least two analytically convenient options present themselves. With the *composite likelihood score approach*, the "likelihood of the family" is computed by taking the product of likelihoods over family members $j = 1, \dots, n_i$ as if they were independent: $L_{i(l)} \equiv \prod_j L_{ij(l)}$. An empirical calculation of the variance of the score accounts for the ignored correlation and yields a test of appropriate size [cf. Clayton, 1999]. With the *biometric likelihood ratio approach*, the correlation among siblings is incorporated directly into the joint likelihood [cf. Abecasis et al., 2000]. This approach is potentially more powerful, but has two drawbacks. The model only applies naturally when the phenotypes are assumed to follow the normal distribution. Moreover, this approach is less robust to misspecification of the correlation matrix. Other options for modeling correlation among family members are also available [e.g., Slager and Schaid, 2001], but we do not pursue them here.

POPULATION-BASED SAMPLES

A population-based sample consists primarily of one or more siblings drawn from each of n unrelated families. Multilocus genotypes and phenotypes are measured on the generation under investigation. If another generation is sampled for some of the families, their genotypes can be used to limit P_i . Phenotypes may be quantitative, qualitative (affected/unaffected), or count data.

Let H_{ij} denote the pair of haplotypes possessed by the j 'th individual in the i 'th family, and let Y_{ij} denote the corresponding phenotype. A set of environmental covariates such as sex and age, denoted by $(C_{ij1}, \dots, C_{ijm})$, might be measured as well.

Suppose there are K distinct haplotypes in the population. Let ψ_k , $k = 0, \dots, K - 1$ denote the relative frequency of haplotype k in the population under investigation. Assuming Hardy Weinberg equilibrium,

$$Pr[H = (a, b)] = \begin{cases} 2\psi_a\psi_b & \text{if } a \neq b \\ \psi_a^2 & \text{if } a = b. \end{cases} \quad (1)$$

For the j 'th offspring in the i 'th family, let X_{ijr} = number of "r" haplotypes minus one. The distribution of the phenotype Y_{ij} , given the haplotype, is assumed to be described by a generalized linear model (GLM) with link

function $g(\mu_{ij}) = \eta_{ij} = \sum_r \beta_r X_{ijr} + \boldsymbol{\alpha}^T \mathbf{C}_{ij}$, where $\mathbf{C} = (1, C_{ij1}, \dots, C_{ijm})^T$. We model the phenotype using a generalized linear model (GLM),

$$f(y_{ij}|H_{ij} = l) = \exp\{[y_{ij}\eta_{ij} - b(\eta_{ij})]/a(\phi) + c(y_{ij}, \phi)\} \quad (2)$$

[McCullagh and Nelder, 1983].

These two features define the full likelihood model for the j 'th subject within family i , and it factors into two components, each of which depends on a distinct set of parameters:

$$\begin{aligned} Lik(\boldsymbol{\psi}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \phi; H_{ij}, Y_{ij}) &= Lik(\boldsymbol{\psi}; H) \times Lik(\boldsymbol{\alpha}, \boldsymbol{\beta}, \phi; Y, H) \\ L &= L^{(H)} \times L^{(Y|H)}. \end{aligned} \quad (3)$$

If haplotypes were directly observable, then inferences would be based solely on $L^{(Y|H)}$, the conditional likelihood of Y , given H . Because of haplotype uncertainty, it is convenient to base the likelihood on the joint probability of the phenotype and genotype. For $n_i = 1, \dots, L_{i(l)}$ is the product of (1) and (2). The desired likelihood can be obtained directly by summing $L_{i(l)}$ over all haplotypes consistent with the observed genotype: $\sum_{l \in P_i} L_{i(l)}$ [Little and Rubin, 1987]. It follows that $\log L_i = \log \left[\sum_{l \in P_i} \Pr[H_i = l] f(y_i|H_i = l) \right]$, and the full loglikelihood is of the form

$$\ell = \sum_i \log \left[\sum_{l \in P_i} L_{i(l)} \right]. \quad (4)$$

For $n_i > 1$, the approach is similar, but one can account for correlation among family members by either taking the composite likelihood score approach or the biometric likelihood ratio approach, both of which are developed in detail shortly.

Throughout the preceding text, we assumed that the sample of families is drawn randomly. This assumption does not hold for many designs, such as case-control studies. With a case-control study, the phenotype is fixed by design, and the response is H . To utilize the GLM model, we implement the traditional approach of moving from a retrospective likelihood to a prospective likelihood (i.e., switching the conditioning from $H|Y$ to $Y|H$). Under certain sampling schemes, and provided no inferences are drawn about the intercept term in the likelihood, this widely used approach has no adverse inferential consequences [e.g., Breslow and Day, 1980; Roeder et al., 1996]. One complication for this implementation is that the estimated distribution of haplotypes will be obtained from a population rich in cases; however, because this is the population for which we wish

to impute missing data, this seems appropriate. For some sampling schemes, such as pedigrees selected to be dense in affected individuals, parameter estimates can be biased. This bias does not affect the validity of the test.

FAMILY-BASED SAMPLES

A family-based sample is a set of independent nuclear families with zero, one, or two parents and one or more siblings sampled. Multilocus genotypes are measured on available parents and offspring, while the phenotypes usually are sampled for offspring only. Phenotypes may be quantitative, qualitative (affected/unaffected), or count data.

For this study design, haplotypes are observed over two generations, both of which directly enter the likelihood. Let H represent the haplotypes of the parents (PG) and offspring (CG). The likelihood contribution for the i 'th family from the genetic observations is

$$L_i^{(H)} = Pr(PG_i) \prod_{j=1}^{n_i} Pr(CG_{ij}|PG_i). \quad (5)$$

The first term is calculated using (1), and the second is computed using Mendelian probabilities.

To obtain $L_{ij(l)}^{(Y|H)}$, define N_{ijr} for the j 'th offspring in the i 'th family as the number of " r " haplotypes minus one. Label the analogous terms for the maternal and paternal parent as N_{iMr} and N_{iPr} , respectively. Define covariates for the model using the quantities analogous to those used in Abecasis et al. [2000]: $Z_{ir} = (N_{iMr} + N_{iPr})/2$, and $X_{ijr} = N_{ijr} - Z_{ir}$. The former is the expectation of each N_{ijr} conditional on the average parental haplotype count, and the latter is the deviation from this expectation for offspring j . In particular, Z measures between family associations, while X measures within family associations. Interest is focused on the coefficients of X_{ijr} . Incorporating Z_{ir} into the model reduces the chance of detecting spurious associations due to population substructure [see Abecasis et al., 2000].

Given H_{ij} , the model for the phenotype of the j 'th child is assumed to be described by a GLM with link function $g(\mu_{ij}) = \eta_{ij} = \sum_r \{\beta_r X_{ijr} + \gamma_r Z_{ir}\} + \boldsymbol{\alpha}^T \mathbf{C}_{ij}$, where $\mathbf{C}_{ij} = (1, C_{ij1}, \dots, C_{ijm})^T$. If there is no population substructure, family-based analyses are usually inefficient relative to population-based analyses [Bacanu et al., 2000]. One source of lost efficiency can be seen by considering a family constellation in which both parents and the

child are homozygous for a “liability haplotype,” labeled r . In this case, $X_{ijr}=0$, and this family contributes no evidence regarding the existence of a liability allele in the region. Indeed, all evidence of an effect due to the r 'th haplotype obtained from this family is absorbed into the term γ_r . The coefficient for γ_r may be highly significant, but this effect is potentially confounded with population substructure [Abecasis et al., 2000], and hence the effect is not tested in a family-based study.

As in the population-based samples, when there is no haplotype uncertainty, the likelihood factors into two components, as in (3). To account for haplotype uncertainty, inferences depend upon the haplotype frequencies in the population, using (5) rather than (1) in (3).

When the response is affected/unaffected, it is illustrative to contrast this design and method of analysis with the traditional TDT. For the latter, each offspring is, by design, affected by the disease, and the response variable is transmission/nontransmission. Hence the TDT likelihood is based on $Pr(CG|PG)$, assuming the child is affected. With the GLM approach, both affected and unaffected siblings are readily incorporated into the analysis, transmission status is a covariate, and status is the response. Population substructure is taken into account by incorporating the extra covariate Z_{ir} , which conditions on the parental genotypes. Because the GLM treatment requires some affected and unaffected responses, it is not applicable to data from affected individuals only. Nonetheless, such data can be analyzed by the ET-TDT model [Seltman et al., 2001], which is implemented in EHAP.

INFERENCES

Cladogram analysis. First consider population-based designs. As part of the cladogram collapsing algorithm, we wish to test if an external clade T has the same β as an internal clade, S . If at the current stage of the cladogram-collapsing algorithm there are $R+1$ clades, $L^{(Y|H)}$ is parameterized by $R+1$ β 's measuring haplotype effects, but only R are identifiable. To ensure identifiability, set $\beta_S=0$. In addition, let $\beta_T=\delta$. Under the null hypothesis, $\delta=0$, but under the alternative hypothesis, δ is unconstrained.

As the cladogram collapses, N_{ijr} collapses and the corresponding covariate X_{ijr} is redefined. As one moves through the steps of the cladogram-collapsing algorithm, different nodes (clades) will take on the parameter δ , and different clades will

be constrained to have a common effect. Parameters $(\delta, \beta_1^u, \dots, \beta_{R-1}^u)$ correspond to the haplotypes clustered into $R+1$ clades, where each node within a clade is constrained to have an identical β . Define $\theta = (\beta_0^u = \delta, \beta_1^u, \dots, \beta_{R-1}^u, \alpha)$ to include all the parameters in $L^{(Y|H)}$ except ϕ . We exclude ϕ because it is a constant for binomial and Poisson models; for the normal model, ϕ is the sample variance, which is orthogonal to the mean parameters. Hence we do not need to include it in the score vector below.

In family-based designs, we follow the same basic algorithm, except that we also have a γ_r corresponding to each clade. We expand θ to include γ . To ensure identifiability, we set $\gamma_0 = 0$. As the cladogram collapses, again N_{ijr} collapses, and the corresponding covariate X_{ijr} is redefined according to the clades. Although the vector β is updated to β^u , the vector γ does not adjust as the dimension of the problem changes. Specifically, X is redefined relative to a Z that has been collapsed into the appropriate clades, but the Z entered as a covariate in the model is unchanged as the cladogram collapsing algorithm progresses. In this way, X represents a deviation in a child's genotype relative to the expectation given the parents, all based on the existing collapsed cladogram. The purpose of this covariate is to explain effects due to population substructure. The covariates associated with Z are not tested, and hence the partition defined by the original cladogram is unchanged as one proceeds through the cladogram-collapsing procedure.

Composite likelihood score approach. For the i 'th family, recall that P_i represents the sets of haplotypes consistent with the observed data. To compute the composite likelihood for $l \in P_i$, multiply the individually obtained likelihoods. Formulas for the composite likelihood and the corresponding score and information matrix are provided in Equations (6–8), respectively, in the Appendix. For $l \in P_i$ write $\mathbf{u}_{i(l)} = \partial l_{i(l)}^{(Y)} / \partial \theta$. The score for the i 'th family is obtained by taking a weighted average of the partial score obtained for each consistent scenario: $\mathbf{u}_{P_i} = \sum_{l \in P_i} w_{i(l)} \mathbf{u}_{i(l)}$, where $w_{i(l)} = L_{i(l)} / \sum_{l \in P_i} L_{i(l)}$. (Note: $L_{i(l)}$ is the full composite likelihood for the i 'th family evaluated for $l \in P_i$.) The total score vector, \mathbf{u} , is obtained by summing the contributions over all families: $\mathbf{u} = \sum_i \mathbf{u}_{P_i}$.

For inferences, the natural partitioning of θ is (δ, λ) , where δ is the parameter of interest, and $\lambda = (\beta_1^u, \dots, \beta_{R-1}^u, \alpha, \gamma)$; \mathbf{u} is similarly partitioned,

$\mathbf{u} = (u_\delta, \mathbf{u}_\lambda)^T$. The other parameters in the model, $(\boldsymbol{\psi}, \phi)$, constitute the additional set of nuisance parameters.

The matrix of negative second derivatives of $\log L^{(Y|H)}$ with respect to $\boldsymbol{\theta}$ for a single haplotype configuration is denoted by $\mathbf{J}_{(l)}$ (see Appendix). As described in Clayton [1999], the corresponding term for the entire set of legal configurations for the i 'th family \mathcal{P}_i is $\mathbf{J}_{\mathcal{P}_i} = \sum_{l \in \mathcal{P}_i} w_{i(l)} \mathbf{J}_{i(l)} - \{ \sum_{l \in \mathcal{P}_i} w_{i(l)} \mathbf{u}_{i(l)} (\mathbf{u}_{i(l)})^T - \mathbf{u}_{\mathcal{P}_i} (\mathbf{u}_{\mathcal{P}_i})^T \}$.

Finally, $\mathbf{J} = \sum_i \mathbf{J}_{\mathcal{P}_i}$. A robust estimate of 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 also partitioned by $(\delta, \boldsymbol{\lambda})$.

The standard arguments used to obtain the score test apply here, but with a slight modification because \mathbf{V} is not identical to \mathbf{J} . This difference occurs because we used the composite likelihood rather than the true likelihood (see Clayton [1999] for another application of this principle). Define $(\hat{\boldsymbol{\lambda}}_0, \hat{\phi}_0, \hat{\boldsymbol{\psi}}_0)$ as the maximum likelihood estimators for $(\boldsymbol{\lambda}, \phi, \boldsymbol{\psi})$ when $\delta = 0$. It can be shown that the score test for testing $\delta = 0$ in the presence of the nuisance parameter $\boldsymbol{\lambda}$ is $\mathbf{u}_\delta^2(\delta = 0, \hat{\boldsymbol{\lambda}}_0, \hat{\phi}_0, \hat{\boldsymbol{\psi}}_0) / \hat{\mathbf{V}}_{\delta, \delta}$, where $\hat{\mathbf{V}}_{\delta, \delta} = \mathbf{V}_{\delta\delta} + \mathbf{J}_{\delta\lambda} \mathbf{J}_{\lambda\lambda}^{-1} \mathbf{V}_{\lambda\lambda} (\mathbf{J}_{\lambda\lambda}^{-1})^T (\mathbf{J}_{\delta\lambda})^T - \mathbf{J}_{\delta\lambda} \mathbf{J}_{\lambda\lambda}^{-1} \mathbf{V}_{\lambda\delta} - \mathbf{V}_{\delta\lambda} (\mathbf{J}_{\lambda\lambda}^{-1})^T (\mathbf{J}_{\delta\lambda})^T$. The latter term is also evaluated at $(\delta = 0, \hat{\boldsymbol{\lambda}}_0, \hat{\phi}_0, \hat{\boldsymbol{\psi}}_0)$. For large samples, the score test is a one-degree-of-freedom test which is distributed as a χ_1^2 under the null hypothesis. For small samples, a permutation test can be performed. For details concerning how to obtain an overall P -value for the series of tests performed, see Seltman et al. [2001].

Biometric likelihood ratio approach. With this approach, the cladogram again defines a sequence of tests, but the tests are now based on a likelihood that incorporates the covariance among siblings. In addition, we utilize likelihood ratio tests instead of score tests; however, the corresponding score tests should yield similar results.

Following the additive biometric model [Falconer, 1989], it can be assumed that the covariance within sibships is explained by three terms: the additive genetic variance (σ_a^2), familial effects due to shared environment (σ_s^2), and residual variability (σ_e^2). With this model, the variance-covariance terms for siblings within a family is

$$\Omega_{ijl} = \begin{cases} \sigma_a^2 + \sigma_s^2 + \sigma_e^2 & \text{if } j = j' \\ 1/2\sigma_a^2 + \sigma_s^2 & \text{if } j \neq j'. \end{cases}$$

The biometric model is a natural choice only for quantitative traits that are approximately normally distributed. Using the covariance terms

above and assuming that the vector of traits measured for the i 'th family is normally distributed, it follows that the likelihood contribution $L_i^{(Y|H)}$ from the family as a whole is of the form of a multivariate normal distribution, with siblings having equicorrelation.

The full loglikelihood is given in (4), and the likelihood ratio test is $2[\ell(\hat{\delta}, \hat{\boldsymbol{\lambda}}, \hat{\phi}, \hat{\boldsymbol{\psi}}) - \ell(0, \hat{\boldsymbol{\lambda}}_0, \hat{\phi}_0, \hat{\boldsymbol{\psi}}_0)]$, where $(\hat{\boldsymbol{\lambda}}_0, \hat{\phi}_0, \hat{\boldsymbol{\psi}}_0)$ are the constrained maximum likelihood estimates for the nuisance parameters when $\delta = 0$. For large samples, this statistic is approximately χ_1^2 under the null hypothesis.

DISCUSSION

For a sample consisting of trios of parents and their affected offspring, and data consisting of multilocus genotypes spanning a candidate gene, we previously introduced ET-TDT [Seltman et al., 2001] to assess the association between disease status and haplotype transmission. ET-TDT used the evolutionary relationships among haplotypes to structure tests of significance, while also accounting for haplotype uncertainty. In this article, we generalize this approach via GLM models for both family-based and population-based samples. These models, which account for haplotype uncertainty but also covariates and the correlation among nuclear family members, are implemented in the software package EHAP. EHAP also contains accessory tools to facilitate evolutionary-based analyses of haplotypes.

In theory, association analyses organized by the evolutionary relationships among haplotypes are expected to increase power in some settings [Templeton et al., 1987]. More importantly, they are more interpretable than standard analyses because they more often identify the haplotype or constellation of haplotypes bearing liability alleles [Templeton et al., 1987; Seltman et al., 2001]. Still, evolutionary-based analyses are not without caveats.

When a relatively large number of polymorphisms is genotyped over a small genomic region, and many of the variants are not in absolute disequilibrium, then the evolutionary relationships among the resulting haplotypes will be specified by a complex cladogram or network, with numerous nodes represented by only one or a few sampled haplotypes. Such a sparsely populated cladogram is not likely to be optimal for statistical analysis. Instead, an optimal solution might be to choose a subset of polymorphisms to

use for cladistic analysis, with the subset chosen to represent the common haplotypes covering a region. This problem is similar to the determination of haplotype blocks for genomic regions, and various solutions have been proposed [Gabriel et al., 2002; Zhang et al., 2002]. The other “ancillary” polymorphisms could be used to bolster inference about the relationships among haplotypes (i.e., the cladogram), as described in Methods and Results under EHAP.

The fundamental organizing structure of these analyses is the cladogram, which presumably depicts the evolutionary relationships among haplotypes. It should be recognized that determining the true evolutionary relationships among haplotypes can be challenging, even with the best of data. Several processes are assumed to be negligible, key among them that recombination and gene conversion in the region are rare and thus have no material impact on the haplotype distribution. Templeton et al. [1987] outlined algorithms to check these assumptions. Crandall and Templeton [1993] also provided algorithms to bolster inference about cladograms, and these algorithms are incorporated into EHAP. Nonetheless, such algorithms are not sufficient to guarantee “correct” evolutionary inference. Interestingly, Seltman et al. [2001] explored the impact of basing analyses on the wrong cladogram. They found that while power and interpretability are diminished, type I error is unaffected. On the other hand, when the cladogram does depict the evolutionary relationships among haplotypes, power and especially interpretability are enhanced substantially. In the latter setting, our versatile software package EHAP should prove to be a useful tool for genetic epidemiologists who are searching for the genetic and environmental basis of complex disease.

ACKNOWLEDGMENTS

We thank the editor and referees for helpful comments, and Shawn Wood for expert programming of portions of EHAP.

APPENDIX

SCORES AND INFORMATION MATRICES

Population-based design. Let \mathcal{H}_i represent the one or more haplotypes corresponding to β_i^u for $i \in \{0, \dots, R-1\}$, 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 β^u for $i \in \{0, \dots, K-1\}$. 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 $\mathcal{H}^{(-1)}(i) = 0$, for all $i \in T$.

Consider a single observation j within the i 'th family, and suppose l , which denotes haplotype pair $H = (a, b)$, is consistent with the observed multilocus genotype. For all of the quantities computed here, we drop the subscript (l) for notational convenience. One needs to compute the terms below for all $l \in \mathcal{P}_i$ for each (i, j) .

Define $\mathbf{X}_{ij} = [X_{ij1}^u, \dots, X_{ijR-1}^u, 1, C_{ij1}, \dots, C_{ijm}]^T$, where $X_{ijk}^u = I\{\mathcal{H}^{-1}(a) = k\} + I\{\mathcal{H}^{-1}(b) = k\} - 1$ for the l 'th configuration for individual j in family i , and $I\{a = b\}$ is an indicator function that is one when $a = b$ and zero otherwise. Note that \mathbf{X}_{ij} is the design matrix corresponding to θ : $\eta_{ij} = \theta^T \mathbf{X}_{ij}$.

Now for the i 'th family, the composite likelihood for the i 'th family is

$$L_{i(l)}^{(Y^H)} = \sum_{j=1}^{n_i} \exp\{[y_{ij}\eta_{ij} - b(\eta_{ij})]/a(\phi) + c(y_{ij}, \phi)\}, \quad (6)$$

$$\mathbf{u}_{i(l)} \sum_{j=1}^{n_i} \mathbf{X}_{ij}^T \{[y_{ij} - b'(\eta_{ij})]/a(\phi)\}, \quad (7)$$

and

$$\mathbf{J}_{i(l)} = \sum_{j=1}^{n_i} b''(\eta_{ij}) \mathbf{X}_{ij} \mathbf{X}_{ij}^T / a(\phi). \quad (8)$$

Family-based design. Let X_{ijk}^u be defined analogously to the population-based covariates, so that it conforms to the cladogram's form. It follows that \mathbf{X}_{ij}^u is naturally defined as $[X_{ij1}^u, \dots, X_{ijR-1}^u, Z_{ij1}, \dots, Z_{ijK-1}, 1, C_{ij1}, \dots, C_{ijm}]^T$. Definitions of the score and information matrices can be obtained readily, as in the population-based case.

ELECTRONIC DATABASE INFORMATION

EHAP, a software package capable of performing all the analyses described here, can be downloaded from <http://wpicr.wpic.pitt.edu/WPICCompGen/>. The Network Analysis software of Bandelt et al. [1999] can be found at <http://www.fluxus-engineering.com/>.

REFERENCES

- Abecasis GR, Cardon LR, Cookson WO. 2000. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet* 66:279–92.
- Bacanu S-A, Devlin B, Roeder K. 2000. The power of genomic control. *Am J Hum Genet* 66:933–44.
- Bandelt HJ, Forster P, Rohlf A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol* 16:37–48.
- Breslow NE, Day NE. 1980. *Statistical methods in cancer research volume 1—the analysis of case-control studies*. Lyon: International Agency for Research on Cancer.
- Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengaard J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E, Sing CF. 1998. Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 63:595–612.
- Clayton DG. 1999. A generalization of the transmission/disequilibrium test for uncertain-haplotype transmission. *Am J Hum Genet* 65:1170–7.
- Clayton DG, Jones H. 1999. Transmission/disequilibrium tests for extended marker haplotypes. *Am J Hum Genet* 65:1161–9.
- Crandall KA, Templeton AR. 1993. Empirical tests of some predictions from coalescent theory with applications to intraspecific phylogeny construction. *Genetics* 134:959–69.
- Falconer D. 1989. *An introduction to quantitative genetics*. Essex, UK: Longman Group, Ltd.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero SN, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander ES, Daly MJ, Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–9.
- Haviland MB, Ferrell RE, Sing CF. 1997. Association between common alleles of the low-density lipoprotein receptor gene region and interindividual variation in plasma lipid and apolipoprotein levels in a population-based sample from Rochester, Minnesota. *Hum Genet* 99:108–14.
- Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM. 2001. Lactase haplotype diversity in the Old World. *Am J Hum Genet* 68:160–72.
- Keavney B, McKensize 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* 7:1745–51.
- Kingman JFC. 1982. On the genealogy of large populations. *Appl Probability Trust* 82:27–43.
- Kruglyak L, Daly MJ, Reeve-Daly MP, Lander ES. 1996. Parametric and nonparametric linkage analysis: a unified multipoint approach. *Am J Hum Genet* 58:1347–63.
- Little RJA, Rubin DB. 1987. *Statistical analysis with missing data*. New York: John Wiley and Sons.
- McCullagh P, Nelder JA. 1983. *Generalized linear models*. London: Chapman and Hall.
- Niu T, Qin S, Xu X, Liu J. 2002. Bayesian haplotype inference for multiple linked single nucleotide polymorphisms. *Am J Hum Genet* 70:157–69.
- O'Connell JR. 2000. Zero-recombinant haplotyping: applications to fine mapping using SNPs. *Genet Epidemiol* 19:64–70.
- Ott J. 1991. *Analysis of human genetic linkage*. Baltimore: Johns Hopkins University Press.
- Roeder K, Carroll RJ, Lindsay BG. 1996. A nonparametric maximum likelihood approach to case-control studies with errors in covariables. *J Am Stat Assoc* 91:722–32.
- Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. 2002. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 70:425–34.
- Seltman H, Roeder K, Devlin B. 2001. Transmission/disequilibrium test meets measured haplotype analysis: family-based association analysis guided by evolution of haplotypes. *Am J Hum Genet* 68:1250–63.
- Sham P. 1998. *Statistics in human genetics*. London: Arnold.
- Slager SL, Schaid DJ. 2001. Evaluation of candidate genes in case-control studies: a statistical method to account for related subjects. *Am J Hum Genet* 68:1457–62.
- Slatkin M, Rannala R. 1997. Estimating the age of alleles by use of intraallelic variability. *Am J Hum Genet* 60:447–58.
- Soubrier F, Martin S, Alonso A, Visvikis S, Tired L, Matsuda F, Lathrop GM, Farrall M. 2002. High-resolution genetic mapping of the ACE-linked QTL influencing circulating ACE activity. *Eur J Hum Genet* 10:553–61.
- Spielman RS, McGinnis RE, Ewens WJ. 1993. The transmission test for linkage disequilibrium: the insulin gene and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* 52:506–16.
- Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Cam NJ, et al. 2001. A candidate prostate cancer susceptibility gene at chromosome 17p. *Nat Genet* 27:172–80.
- 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 apolipoprotein E locus. *Genetics* 140:403–9.
- 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* 117:343–51.
- 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* 132:619–33.
- Templeton AR, Sing CF, Kessler A, Humphries S. 1988. A cladistic analysis of phenotypic associations with haplotypes inferred from restriction endonuclease mapping. II. The analysis of natural populations. *Genetics* 120:1145–54.
- 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* 134:659–69.
- Zhang K, Deng M, Chen T, Waterman MS, Sun F. 2002. A dynamic programming algorithm for haplotype block partitioning. *Proc Natl Acad Sci USA* 99:7335–9.