Am. J. Hum. Genet. 73:1016–1026, 2003

Accounting for Linkage in Family-Based Tests of Association with Missing Parental Genotypes

Eden R. Martin,[1] Meredyth P. Bass,[1] Elizabeth R. Hauser,[1] and Norman L. Kaplan[2]

[1]Department of Medicine, Duke University Medical Center, Durham, NC; and [2]Biostatistics Branch, National Institute for Environmental Health Sciences, Research Triangle Park, NC

In studies of complex diseases, a common paradigm is to conduct association analysis at markers in regions identified by linkage analysis, to attempt to narrow the region of interest. Family-based tests for association based on parental transmissions to affected offspring are often used in fine-mapping studies. However, for diseases with late onset, parental genotypes are often missing. Without parental genotypes, family-based tests either compare allele frequencies in affected individuals with those in their unaffected siblings or use siblings to infer missing parental genotypes. An example of the latter approach is the score test implemented in the computer program TRANSMIT. The inference of missing parental genotypes in TRANSMIT assumes that transmissions from parents to affected siblings are independent, which is appropriate when there is no linkage. However, using computer simulations, we show that, when the marker and disease locus are linked and the data set consists of families with multiple affected siblings, this assumption leads to a bias in the score statistic under the null hypothesis of no association between the marker and disease alleles. This bias leads to an inflated type I error rate for the score test in regions of linkage. We present a novel test for association in the presence of linkage (APL) that correctly infers missing parental genotypes in regions of linkage by estimating identity-by-descent parameters, to adjust for correlation between parental transmissions to affected siblings. In simulated data, we demonstrate the validity of the APL test under the null hypothesis of no association and show that the test can be more powerful than the pedigree disequilibrium test and family-based association test. As an example, we compare the performance of the tests in a candidate-gene study in families with Parkinson disease.

Introduction

A common strategy in the search for complex disease genes is to identify regions of linkage, often through a genome scan in a large sample of multicase families. However, regions identified through linkage analysis can be large. Often, tests of association are used to further localize the susceptibility locus. With this paradigm, tests of association must allow for linkage between the susceptibility locus and markers. This is not an issue for case-control tests using unrelated individuals, but, if cases are relatives (e.g., affected siblings), then adjustments in the test statistic must be made (Slager and Schaid 2001). For family-based tests of association, linkage between a marker and disease locus must be taken into account when families with multiple affected siblings are included in the analysis. For example, the classic family-based transmission/disequilibrium test (TDT) can be used to test for association in the presence of linkage in family triads (one affected offspring and both parents) (Spielman et al. 1993; Spielman and Ewens 1996). However, for families with more than one affected offspring, the TDT is not valid as a test of association if there is linkage between the disease and marker loci. The loss of validity occurs because, when there is linkage, the transmissions of parental marker alleles to multiple affected offspring are correlated, and this correlation is not accounted for in the variance estimate used in the TDT. Modifications of the TDT have been proposed that correctly account for correlations due to linkage in nuclear families with multiple affected offspring and in extended pedigrees (Martin et al. 1997, 2000; Abecasis et al. 2000; Rabinowitz and Laird 2000).

Linkage between a marker and disease locus can also be a problem in samples of families with multiple affected siblings when testing for association in late-onset diseases, in which parental data are often missing. One approach to the study of late-onset diseases is to compare allele or genotype frequencies in affected and unaffected siblings, and such methods have been proposed that properly test for association in the presence of linkage in sibship data (Horvath and Laird 1998; Monks et al. 1998; Martin et al. 2000). Alternatively, genotypes

Received May 15, 2003; accepted for publication July 28, 2003; electronically published October 9, 2003.

Address for correspondence and reprints: Dr. Eden R. Martin, 595 LaSalle Street, Duke University Medical Center, Box 3468, Durham, NC 27710. E-mail: eden.martin@duke.edu

0002-9297/2003/7305-0005$15.00

of siblings can be used to infer missing parental genotypes (Clayton 1999; Knapp 1999; Weinberg 1999). When there is no linkage between a marker and disease locus, the inference of missing parental genotypes from multiple affected offspring depends only on Mendelian probabilities and allele frequencies. This method of inference is used in the score test implemented in the computer program TRANSMIT (Clayton 1999). The inference of parental genotypes at a marker locus linked to a susceptibility locus is more difficult, because transmissions of parental alleles to multiple affected offspring are correlated.

We demonstrate, using computer simulations, that failure to account for this correlation due to linkage when inferring missing parental data in samples of nuclear families with multiple affected siblings can lead to a bias in the statistic calculated in TRANSMIT. This bias inflates the type I error rate of the score test in TRANSMIT, and this inflation increases with increasing sample size. We present a new test for association in the presence of linkage (APL) that incorporates identity-by-descent (IBD) relationships to adjust for linkage when inferring missing parental genotypes in nuclear families. We use computer simulations to demonstrate the validity of the new APL statistic, and we examine its statistical power and compare the power of the APL test to the power of two alternative methods in nuclear families: the pedigree disequilibrium test (PDT) (Martin et al. 2000) and the family-based association test (FBAT) (Lake et al. 2000). Finally, we show the utility of the APL test in an application to data from a candidate-gene study in a set of families with Parkinson disease (Martin et al. 2001*b*).

Methods

The Difficulty of Parental Genotype Inference in Affected Sib Pairs (ASPs)

Consider the simple situation of a nuclear family with two affected siblings (i.e., a family with one ASP) with marker genotypes $\boldsymbol{G} = (G_1,G_2)$ for sibs 1 and 2. Initially, assume that both parents are typed, with genotypes $\boldsymbol{G}_p = (P_1,P_2)$. Suppose, for simplicity, that the marker is biallelic and focus on allele 1 (although the argument extends to multiple alleles as well). For each ASP family, define X as the number of copies of allele 1 in siblings, with $X = 0,\ldots,4$, and define N_p as the number of copies of allele 1 in parents, with $N_p = 0,\ldots,4$.

If there is no association between alleles at the marker and the disease locus, then the expected value of X conditional on the parents' genotypes is $E(X|\boldsymbol{G}_p) = N_p$. This conditional expectation forms the basis of many of the commonly used family-based tests of association, such as HHRR, TDT, PDT, and TRANSMIT (Terwilliger and Ott 1992; Spielman et al. 1993; Clayton 1999; Martin et al. 2000). It follows that, for the ith family, $T_i = X_i - N_{pi}$ has a mean of 0 under the null hypothesis of no association. If we sample N independent ASP families, then a statistic can be based on $T_\cdot = \sum_{i=1}^{N} T_i$. In constructing the variance of this sum, one must account for correlation between affected siblings that might result from linkage (Martin et al. 1997). Clayton (1999) uses a "robust variance" estimator, which uses an empirical estimate of the variance that treats families, rather than sibs, as the independent units.

There is no difficulty with this approach when full parental genotype data are available. When parental genotypes are missing, however, it becomes necessary to consider all possible parental genotypes that are consistent with the genotypes of the offspring (Clayton 1999). Specifically, if $\mathcal{P}$ denotes the collection of parental genotypes consistent with $\boldsymbol{G}$, then, for the ith family,

$$T_i = X_i - \sum_{j \in \mathcal{P}} \hat{P}(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)N_{pij}$$

where N_{pij} is the number of copies of allele 1 in parents in the jth set of parents in $\mathcal{P}$ and $\hat{P}(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)$ is an estimate of $P(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)$, the probability of the jth set of possible parental genotypes, given the genotypes of the offspring ($\boldsymbol{G}_i$) and the event that both siblings are affected (A).

The value of X_i is observed by counting the number of copies of allele 1 in the affected siblings. N_{pij} is known for any specified set of possible parents. $P(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)$ must be estimated under the null hypothesis of no association between the marker and disease alleles so that T_i has a mean of 0. If we are willing to assume that there is no linkage between a marker and the disease locus, then, under the null hypothesis that there is no association between the marker and disease alleles, $P(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)$ does not depend on disease status and is a function only of mating-type probabilities and Mendelian segregation probabilities. This is the procedure implemented in the TRANSMIT program (Clayton 1999). However, if there is linkage between the marker and the disease locus, then the transmissions to affected siblings are correlated. Failure to account for this correlation leads to biased estimation of the parental genotype frequencies, which results in a biased statistic.

Allowing for Linkage in Parental Genotype Inference in ASPs

We propose the following alternative approach to estimating $P(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)$, an approach that allows for linkage between a marker and the disease locus and provides a valid test of association. The key observation is that,

under the null hypothesis and in the presence of tight linkage between a marker and the disease locus, the only additional parameters that must be considered are the IBD parameters z_0, z_1, and z_2, which are the probabilities that the affected siblings share 0, 1, or 2 alleles IBD, respectively, at the marker locus. To see that this is true, for any set of parental genotypes $\boldsymbol{G}_p$ consistent with offspring genotypes $\boldsymbol{G}$ (dropping subscripts i and j to simplify notation),

$$P(\boldsymbol{G}_p|\boldsymbol{G},A) = \frac{P(\boldsymbol{G}_p|A)P(\boldsymbol{G}|\boldsymbol{G}_p,A)}{P(\boldsymbol{G}|A)}$$

$$= \frac{P(\boldsymbol{G}_p|A)\sum_{k=0}^{2} P(\boldsymbol{G}|\boldsymbol{G}_p,A,IBD = k)P(IBD = k|\boldsymbol{G}_p,A)}{P(\boldsymbol{G}|A)} . \quad (1)$$

If there is tight linkage and no association between marker and disease alleles, then

$$P(\boldsymbol{G}|\boldsymbol{G}_p,A,IBD = k) = P(\boldsymbol{G}|\boldsymbol{G}_p,IBD = k)$$

$$\text{and } P(IBD = k|\boldsymbol{G}_p,A) = P(IBD = k|A) = z_k .$$

Because of tight linkage, we always assume that IBD at the marker is equivalent to IBD at the disease locus. When there is no association, the mating-type probability $P(\boldsymbol{G}_p|A)$ is equal to the unconditional mating-type probability denoted μ_{Gp}. It follows that, under the null hypothesis and the assumption of tight linkage,

$$P(\boldsymbol{G}_p|\boldsymbol{G},A) = \frac{\mu_{Gp}\sum_{k=0}^{2} z_k P(\boldsymbol{G}|\boldsymbol{G}_p,IBD = k)}{P(\boldsymbol{G}|A)} . \quad (2)$$

$P(\boldsymbol{G}|\boldsymbol{G}_p,IBD = k)$ in the numerator is a function of Mendelian segregation probabilities only. $P(\boldsymbol{G}|A)$ can be computed by summing all terms in the numerators over all possible parents, $\boldsymbol{G}_p$, for a given $\boldsymbol{G}$. If there is no linkage between the marker and the disease locus, then $z_2 = z_0 = 1/4$, $z_1 = 1/2$, and the transmissions from parents to each sibling are independent. Thus, to estimate $P(\boldsymbol{G}_p|\boldsymbol{G},A)$, one needs to estimate only the mating-type parameters. This strategy is used in TRANSMIT with the additional assumption of Hardy-Weinberg equilibrium (HWE). If linkage does exist between the marker and disease locus, then the $\{z_i\}$ are unknown and must be estimated jointly with the $\{\mu_{Gp}\}$.

The Expectation Maximization (EM) Algorithm for Parameter Estimation

To estimate the probabilities in equation 2, we need to estimate the mating-type and IBD probabilities. This can be accomplished using the EM algorithm (Dempster et al. 1977). Appendix A shows the steps of the EM algorithm. In the E-step, the expected full data set is computed on the basis of observed data and parameter estimates. Specifically, let the observed data, n_G, be counts of the number of sibships with genotypes $\boldsymbol{G}$ for all $\boldsymbol{G}$. The full data are the partition of n_G, for each $\boldsymbol{G}$, into the joint parental genotype and IBD classes. Then the expected full data counts are the numbers of sibships with genotypes $\boldsymbol{G}$, parental genotypes $\boldsymbol{G}_p$ and $IBD = k$. We denote the expected full data for these classes $\tilde{n}\{\boldsymbol{G},\boldsymbol{G}_p,k\} = n_G P(\boldsymbol{G}_p,IBD = k \mid \boldsymbol{G},A)$. These expected values depend on the unknown mating-type and IBD parameters. Initial estimates of these parameters are plugged into the probability formulas to obtain values for the expected counts. In the M-step, the expected data are used in the equations for the maximum likelihood estimates (MLEs) for mating-type and IBD parameters to provide new parameter estimates. The formulas for these estimates are given in appendix A. These new estimates then replace the initial estimates in the E-step, and the procedure iterates until the estimates converge.

A Novel Statistic for Association in the Presence of Linkage

Once parameter estimates have been calculated, the variance of $T_.$ can be estimated using the general form

$$\hat{Var}(T_.) = \sum_{i=1}^{N} T_i^2 - \boldsymbol{J}'\hat{\boldsymbol{V}}_{mle}\boldsymbol{J} .$$

If we define γ to be a vector of the mating-type and IBD parameters and $g_X(\gamma) = \sum_{i=1}^{N}\sum_{j\in\mathcal{P}} P(\boldsymbol{G}_{pj}|\boldsymbol{G}_i,A)N_{pij}$, then $\boldsymbol{J}$ is a vector of partial derivatives of $g_x(\gamma)$ evaluated at the MLEs for γ. The quantity $\hat{\boldsymbol{V}}_{mle}$ is an estimate of the variance-covariance matrix for the MLEs, which can be computed with standard likelihood theory (Clayton 1999; Shih and Whittemore 2002). Note that the variance estimate is calculated with the family as the independent units and that it thus allows for correlation between siblings as a result of linkage. This variance estimate is similar to the "robust variance" estimator used by Clayton (1999) but takes into account the variance associated with estimation of IBD parameters.

Finally, the statistic that we propose takes the form:

$$\frac{T_.}{\sqrt{\hat{Var}(T_.)}} .$$

Under the null hypothesis, this statistic is asymptotically normal, with a mean of 0 and a variance of 1. We will refer to the test based on this statistic as the "APL test."

Modeling Mating-Type Parameters

For biallelic markers, such as SNPs, the five mating-type parameters can be estimated directly (Weinberg et al. 1998). In this case, the test is not sensitive to deviations from HWE that might be found, for example, in stratified populations. However, if there are several marker alleles (or haplotypes, as discussed below), then HWE must be assumed for our analysis, since there will be a large number of possible mating types.

A simplifying assumption is to suppose HWE in the population. This allows each mating-type parameter μ_{Gp} to be modeled in terms of the allele frequencies, which reduces the number of nuisance parameters that need to be estimated. For example, for a biallelic marker, there are six possible mating types, which gives five free mating-type parameters. If we assume HWE, then the number of free parameters that we need to estimate to model the mating-type parameters is only one, the allele frequency. In general, the number of parameters required when we assume HWE is one less than the number of alleles.

Extensions to Other Family Structures

In practice, many disease data sets will contain a mix of family types. Families may have one or more affected siblings, no unaffected siblings or one or more unaffected siblings, and full or partial parental data. The method outlined above can be extended to accommodate these different family structures and allows the combination of different family types in the test. All families contribute to estimation of the mating-type parameters, but only families with ASPs contribute to estimation of the IBD parameters.

Although some families may have more than two affected siblings, the present article, for the sake of simplicity, considers only cases in which families have one or two affected siblings. For families with a single affected offspring, the contribution to the numerator of the APL statistic is

$$T_i = X_i - \frac{1}{2}N_{pi}$$

if both parents' genotypes are known, and

$$T_i = X_i - \frac{1}{2}\sum_{j\in\Phi}\hat{P}(\boldsymbol{G}_{pj}|G,A)N_{pij}$$

if both parents' genotypes are unknown, where X_i is the number of copies of allele 1 in the single affected offspring in the ith family, N_{pi} is the number of copies of allele 1 in the parents in the ith family, G is the genotype of the affected offspring, and A is the event that the singleton is affected. For singleton families with missing parental data, the expression for $P(G_{pj}|G,A)$ is simpler than the expression in equation (2), because it depends only on the mating-type parameters, which are estimated from the entire collection of families.

For complex diseases, for which any single disease locus is expected to have low penetrances, transmissions from parents to unaffected siblings add little information about association. However, unaffected siblings can be used to improve estimation of mating-type parameters. For families with two affected siblings, additional unaffected siblings contribute only to the calculation of $P(\boldsymbol{G}|\boldsymbol{G}_p,IBD = k)$ in equation (1). For example, for a family with two affected and one unaffected sibling, $\boldsymbol{G}$ denotes the set of genotypes for the three siblings, and the IBD estimates refer only to the ASP. If penetrances are low, then, given the parental genotypes, the transmissions to the unaffected sibling are independent of transmissions to the other siblings. Thus, the probabilities, $\{P(\boldsymbol{G}|\boldsymbol{G}_p,IBD = k)\}$, which were previously derived for ASPs, are simply multiplied by Mendelian transmission probabilities for the unaffected sibling. Estimation of mating-type parameters and IBD can be accomplished with the EM algorithm, as discussed above. This approach extends to additional unaffected siblings in a similar way. The same arguments work for families with one affected sibling.

If some families have partial parental genotype information, then they can be used to aid in parameter estimation for families with missing parental information. If P_1 is missing, we estimate $P(P_1|P_2,\boldsymbol{G},A)$, which will depend on the mating-type and IBD parameters. Estimation proceeds as before, using equation (1) with appropriate accommodations to reflect the fact that one parent, P_2, is known.

Computer Simulations

Computer simulations were used to evaluate type I error and power. The SIMLA computer program (Bass et al. 2002) was used to simulate replicate samples of $N = 250$ and $N = 500$ nuclear families with different numbers of affected and unaffected siblings. A single disease locus was simulated with the model parameters given in table 1. Disease-allele frequency and penetrances were chosen to give prevalences ~0.005. Low penetrances and common allele frequencies were used to reflect a single locus that contributes a small amount of risk, as would be expected in a complex disease. Recessive and multiplicative models were chosen to constrain the relationships between the penetrances. Penetrances were chosen to give four genetic models with a range of genetic effects. To quantify the genetic effect, we computed the recurrence-risk ratio for siblings, λ_s, for each model. The genetic marker was simulated under the assumption of complete linkage to the disease locus. For

Table 1

Genetic Models Used in Simulations

Model of Inheritance	Disease-Marker Allele Frequency	Penetrance			Disease Prevalence	λ_s[a]
		f_0	f_1	f_2		
Recessive:						
RecA	.25	.005	.005	.025	.0063	1.21
RecB	.25	.005	.005	.020	.0059	1.13
RecC	.25	.005	.005	.015	.0056	1.06
RecD	.25	.005	.005	.010	.0053	1.02
Multiplicative:						
MultA	.15	.004	.011	.030	.0064	1.26
MultB	.15	.004	.010	.025	.0060	1.20
MultC	.15	.004	.008	.016	.0053	1.10
MultD	.15	.004	.006	.009	.0046	1.03

[a] Recurrence-risk ratio for siblings.

type I error simulations, there was no association between the disease and marker alleles. Power simulations assumed that the marker and disease alleles were in perfect association, so, in effect, the loci were identical.

Application to Parkinson Disease

To evaluate the utility of the APL test in real data, we applied the APL and other tests to five SNPs in the gene encoding the microtubule-associated protein τ (on 17q21) that we have shown elsewhere to be associated with Parkinson disease (Martin et al. 2001*b*) and that is known to lie in a region linked to Parkinson disease (Scott et al. 2001). For this example, we considered a subset of those families with either two affected siblings (AA families) or two affected siblings and one unaffected sibling (AAU families). Because of the late-onset of Parkinson disease, no parental genotype data were available in these families. The number of fully genotyped families for this example varied between 94 and 120 for the five SNPs. Two of the SNPs genotyped were intronic: one in intron 3 (SNP 3) and one in intron 11 (SNP 11). The other three SNPs lie in exon 9: SNPs 9i, 9ii, and 9iii. The details of SNP genotyping and a description of the sample of families with Parkinson disease are given by Martin et al. (2001*b*).

Results

To examine the impact of testing for association in a region of linkage and the impact of the affected-sibling correlation on the TRANSMIT test statistic, we began by considering an extreme example. This example clearly demonstrates the problem that linkage creates for samples in families with multiple affected siblings and missing parental genotype data. For each of the eight genetic models (table 1), marker data were simulated for 2,000 replicate samples of *N* ASPs with no parental data. Marker and disease alleles were unassociated, but the loci were completely linked. The "robust variance" estimator was used in the calculation of the score statistic from TRANSMIT (-ro flag option in the TRANSMIT software). The results in table 2 show that the type I error rate in the score test can be inflated.

Table 2 also shows the mean and variance over replicate data sets of the TRANSMIT score statistic. Although the program outputs the χ^2 version of the score statistic, we converted the statistic to the normal statistic so that the direction of the deviation from the expected value of 0 under the null hypothesis could be examined. The statistic was computed only for the minor allele (allele frequency .25 for the recessive and allele frequency .15 for the multiplicative models), since the statistic for the other allele would be the same but with the opposite sign. If the test were valid, then the expectation of the statistic would be 0; however, for all of these examples, the estimated means are >0. This reveals a positive bias in the score statistic that occurs because the increased allele sharing among siblings as a result of linkage is ignored. Since the major allele will be shared IBD more frequently than the minor allele, its frequency will be overestimated in parents, resulting in an underestimate of the minor allele in the parents. It is noteworthy that when the alleles have equal frequency, the allele frequency in the parents is correctly estimated, and the score statistic is unbiased.

Table 2

Mean and Variance of Score Test Statistic from TRANSMIT across 2,000 Replicate Data Sets of *N* AA Families without Data from Parents

N and Model of Inheritance	Score Test		
	Mean	Variance	Type I Error[a]
N = 250:			
RecA	1.29	1.05	.25
RecB	.82	1.11	.15
RecC	.38	.94	.06
RecD	.17	.97	.05
MultA	.96	1.06	.17
MultB	.69	1.07	.11
MultC	.39	.98	.06
MultD	.15	.98	.05
N = 500:			
RecA	1.83	1.01	.43
RecB	1.18	1.02	.22
RecC	.62	.96	.09
RecD	.13	1.04	.06
MultA	1.38	1.02	.27
MultB	.99	1.14	.18
MultC	.56	1.03	.08
MultD	.15	1.04	.07

[a] Proportion of data sets with $P \leq .05$.

Table 3

Mean and Variance of Score Test Statistic from TRANSMIT across 2,000 Replicate Data Sets of AAU Families, without Data from Parents, and Combined Samples of AA and AAU Families, without Data from Parents

N and Model of Inheritance	Score Test		
	Mean	Variance	Type I Error[a]
N = 250 AAU:			
RecA	.19	1.02	.051
RecB	.12	.97	.050
MultA	.20	1.04	.057
MultB	.18	1.04	.063
N = 500			
RecA	.22	1.05	.059
RecB	.13	.97	.047
MultA	.27	.95	.048
MultB	.23	1.00	.058
N = 250 AA + 250 AAU[b]:			
RecA	.56	.98	.089
RecB	.37	.99	.068
MultA	.57	1.07	.094
MultB	.45	1.06	.076

[a] Proportion of data sets with $P \leq .05$.

[b] Combined data sets with 250 ASP families and 250 families with two affected siblings and one unaffected sibling, without data from parents.

The bias and increase in type I error are more extreme for models with larger relative recurrence risks for siblings (λ_s) (table 2). For example, for the RecA model, in which $\lambda_s = 1.21$, the type I error of the score test is 0.43 for samples of 500 families. When λ_s is close to 1 (e.g., in the RecD and MultD models), the type I error is close to the nominal level. This occurs because, when λ_s is close to 1, the probability of a sibling of an affected individual being affected is the same as that of a randomly selected member of the population. This means that, even though there is complete linkage between a marker and the disease locus, the transmissions to siblings are, in effect, independent.

An important consequence of the bias incurred by incorrectly assuming no linkage is that the type I error increases with increasing sample size. This can be seen for each model as the sample size increases from 250 to 500 (table 2). Consequently, for large samples, which are required for detection of small contributions to complex traits, type I error can be seriously inflated.

The inclusion of unaffected siblings in the analysis reduces the bias of the score statistic in TRANSMIT because, for alleles with low penetrances, conditional on parental genotypes, transmissions to unaffected siblings are essentially independent of transmissions to any other siblings. Therefore, the parental genotype inference used by TRANSMIT is appropriate for unaffected siblings. Table 3 shows the mean and variance of the score statistic and the type I error for the four models for which the statistic is most biased (RecA, RecB, MultA, and MultB). The bias is considerably smaller—and the type I error estimates are much closer to the nominal level—than in the simulations with only ASPs. However, as noted above, a small bias can inflate the type I error if the sample size is large. We also considered a data set composed of a mixture of families with only an ASP and families with an ASP and one unaffected sibling (table 3). As expected, the bias and its effect on type I error are not as severe as in a data set with only ASPs, but they are somewhat worse than those in a data set in which all families have an unaffected sibling.

The APL statistic that we propose is constructed to be a valid test of association, even when linkage exists. For the ASP data simulated for table 2 (under the null hypothesis), we found that the numerator of the statistic has the correct mean of 0. In fact, when the sample consists only of affected siblings with no parental genotype data, the allele frequency estimated for the parents is always very close to the allele frequency in the sample of affected siblings. Thus, the numerator of the APL statistic is, in effect, 0 for every sample. Furthermore, the same is true when there is association between the marker and disease alleles. Intuitively, it makes sense that if there are no unaffected siblings or parental data, then a test for association based only on affected individuals has no way of obtaining good estimates of allele frequency in the parental populations. Without

Table 4

Mean and Variance of the APL Test across 2,000 Replicate Data Sets of AAU Families, without Data from Parents, and Combined Samples of AA and AAU Families, without Data from Parents

N and Model of Inheritance	APL Test		
	Mean	Variance	Type I Error[a]
N = 250:			
RecA	.01	.98	.048
RecB	.03	.94	.046
MultA	−.02	1.03	.050
MultB	−.04	1.01	.049
N = 500			
RecA	.04	1.01	.050
RecB	.01	.98	.044
MultA	−.032	.94	.040
MultB	−.039	1.00	.049
N = 250 AA + 250 AAU[b]:			
RecA	.03	.98	.053
RecB	.01	.96	.042
MultA	−.04	1.02	.055
MultB	−.03	1.03	.048

[a] Proportion of data sets with $P \leq .05$.

[b] Combined data sets with 250 ASP families and 250 families with two affected and one unaffected siblings, without data from parents.

some "control" data, the frequency estimates in the parents will approximate the frequency in the affected offspring. Thus, for association testing, the extreme example of a sample of ASPs with no parental data provides no information for the APL test.

A more interesting example for studying the properties of the APL test is in samples in which some families include unaffected siblings. Table 4 shows the mean and variance of the APL statistic and the type I error of the APL test for samples of AAU families and combined samples of AA and AAU families for RecA, RecB, MultA, and MultB models, with no association between alleles but with complete linkage between a marker and the disease locus. For simplicity and to make the test more comparable to the score test of TRANSMIT, we have modeled mating-type parameters in the APL test, under the assumption of HWE throughout. For all examples, the mean of the APL statistic was close to 0, the variance was close to 1, and the type I error of the test was close to the nominal level .05.

We next considered power of the APL test when both linkage and association are present. Since the score test in TRANSMIT has been shown to have an inflated type I error rate when there are multiple affected siblings and missing parental data, it will also have inflated power. Thus, it is not appropriate to compare the power of the score test with that of the APL test. Two alternative tests—PDT (Martin et al. 2000, 2001*a*) and the FBAT (Lake et al. 2000)—do, however, maintain the correct type I error as tests of association, even when there is linkage. Thus, we compared the APL test with the PDT and FBAT. For the PDT, we used the PDT-sum version (Martin et al. 2001*a*), and, for the FBAT, we used the empirical-variance estimator (-e flag option in the FBAT program). Figure 1 shows estimates of power for the APL, PDT, and FBAT from simulated data sets of AAU families and of combined AA and AAU families. We considered two significance levels for power calculations: .05 and .001. Power estimates were 1 for all examples in the RecA and MultA models, so they are not shown in figure 1. To make the power simulations more informative, we added two models with intermediate power (RecX with $\lambda_s = 1.04$ and MultX with $\lambda_s = 1.08$).

We see that when all families contain an unaffected sibling (all AAU families), the power of the APL test is greater than or equal to the power of the PDT and FBAT for all but one case—the RecC model with $\alpha = 0.001$, in which the APL has slightly less power than the PDT and FBAT. In the combined data sets with both AAU and AA families, the PDT and FBAT do not use the families without unaffected siblings. The APL test will use information from the entire collection of families. However, the results in figure 1 show that AA families add little information for the APL test. For the recessive models, the APL test with 250 AAU families has power similar to that in the combined data set of 250 AAU and 250 AA families; and, for some examples, use of the combined data actually decreases the power of the APL test. For example, in the RecB model with $\alpha = 0.001$, the addition of 250 AA families decreases the power of the APL test from 0.82 (with 250 AAU families only) to 0.74 (with 250 AAU plus 250 AA families). In the multiplicative models, the APL test is slightly more powerful in the combined data set than in AAU families alone, and the test continues to be more powerful than the PDT and FBAT.

Table 5 shows the results of the APL test, the score test from TRANSMIT, the PDT, and the FBAT in a sample of families with Parkinson disease. The SNPs examined are in the τ gene, lying in a region that has shown evidence of linkage (LOD = 2.62) through a genomic screen in families with Parkinson disease (Scott et al. 2001). The sample analyzed here consists primarily of AAU families, with ~10% of the families having only an ASP (AA). Parental genotypes were missing in all families. As shown above, in regions of linkage, the TRANSMIT statistic may be biased in families with multiple affected siblings and missing parents; thus, the interpretation of the test is unclear in these data. The APL and TRANSMIT tests both show significant *P* values for four of the markers (table 5), with the APL test being somewhat more significant than the TRANSMIT score test. These four markers are known to be in strong linkage disequilibrium (Martin et al. 2001*b*), so it is not correct to interpret these tests as independent of one another; nevertheless, they offer an interesting example for comparison. Note that for each of the five markers, the PDT and FBAT produced identical *P* values, to four decimal places, and these *P* values were considerably larger than those for the APL and TRANSMIT tests.

Discussion

Association analysis can be a useful tool for fine mapping of complex disease genes in regions of linkage. Family-based tests of association that infer missing parental data must take linkage between a marker and disease locus into account when using families with multiple affected siblings. We demonstrated that the score statistic calculated in the computer program TRANSMIT is biased under the null hypothesis of no association in the presence of linkage when ASP data with missing parental genotype information are used. This bias has important implications that can lead to misinterpretation of results in fine-mapping studies. We showed that the bias results in an inflated type I error rate, leading to excessive false-positive results. Furthermore, this inflation increases with increasing sample size. Thus, it is difficult to distinguish this effect from the increase in power that we

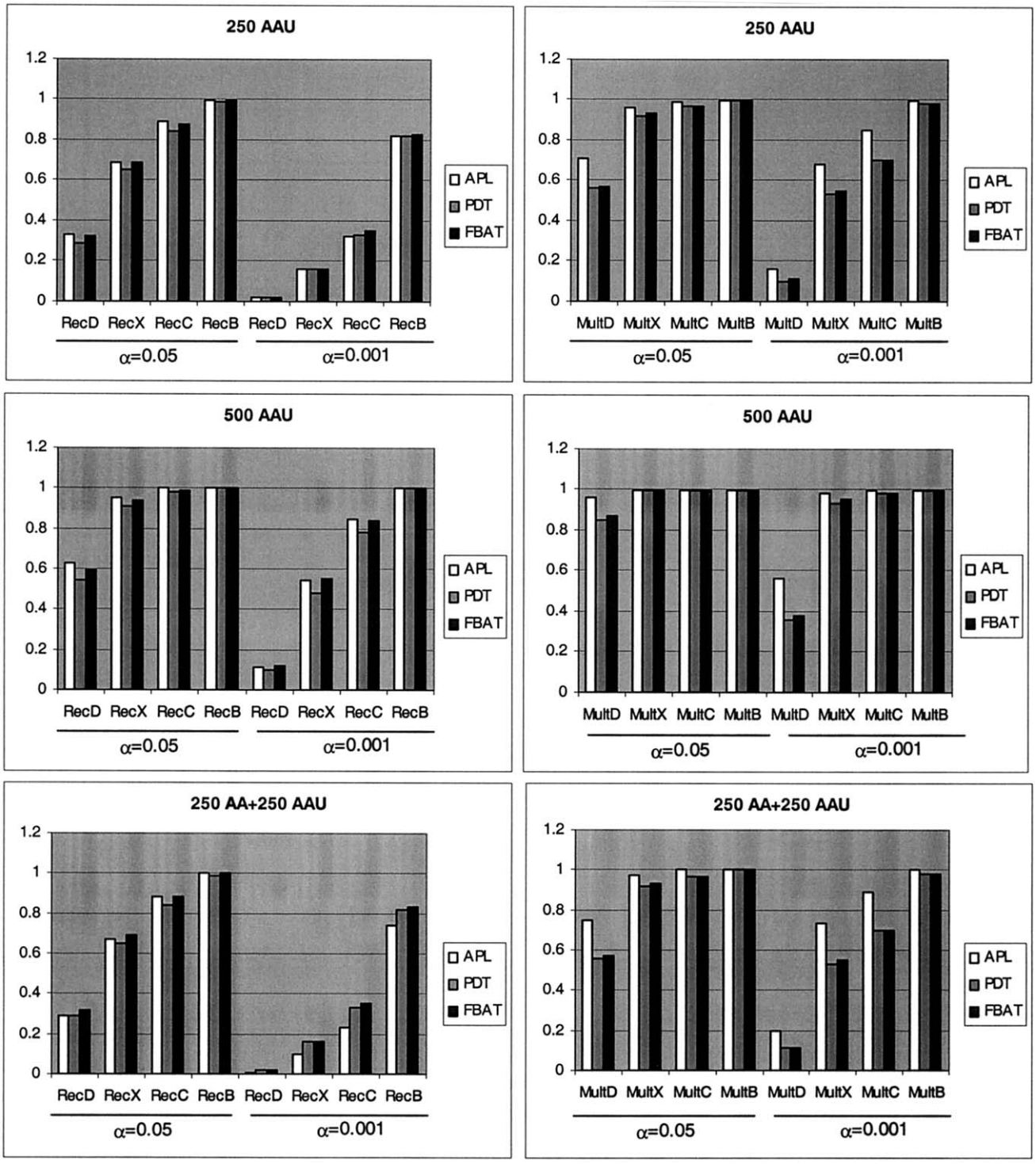

Figure 1 Power of the APL test, PDT, and FBAT over 1,000 replicate data sets of N AAU families, without data from parents, and 1,000 replicate data sets of combined samples of 250 AA and 250 AAU families, without data from parents. Power is computed for significance level (α) of 0.05 and 0.001 for recessive and multiplicative models.

Table 5

***P* Values for Tests of Association in Families with Parkinson Disease for Five SNPs in the τ Gene**

		P			
SNP	N^a	APL	TRANSMIT	PDT	FBAT
3	115	.002	.006	.235	.235
9i	94	.0007	.002	.227	.227
9ii	120	.015	.03	.556	.556
9iii	96	.707	.529	.564	.564
11	115	.004	.007	.278	.278

[a] AA and AAU families, without data from parents, that are genotyped, without regard to marker informativity.

expect with increasing sample size when there is a true association. The bias can also lead to incorrect conclusions in the interpretation of results in stratified data sets—for example, when stratifying families by numbers of affected individuals (e.g., multiplex vs. singleton) or age at onset (e.g., early or late onset). Since the bias due to linkage occurs only when families with multiple affected siblings are used, we would expect to see more false-positive associations in multiplex families than in singleton families. This can lead to the incorrect interpretation that the association is stronger in multiplex families that are "genetically loaded" for disease susceptibility. Similarly, because the bias is limited to families with missing parental data, we may falsely conclude that the association is restricted to families with late-onset disease.

Because testing for association in regions of linkage is a common paradigm for mapping complex disease genes, we proposed an alternative test statistic that infers missing parental genotypes conditional on IBD allele sharing in affected siblings. By conditioning on IBD sharing in affected siblings in the parental genotype inference, the APL test remains valid as a test for association in the presence of linkage. Our simulations demonstrated that the APL test can have more power than the PDT and FBAT. This shows that, in nuclear family data, the APL test may be preferred over the PDT and FBAT. The application to the Parkinson disease data further demonstrates the increase in power of the APL over PDT and FBAT. The APL *P* values are considerably smaller than *P* values from the PDT and FBAT at four of the five SNPs tested and are comparable to those of the TRANSMIT test. The τ gene has been found to be associated with Parkinson disease and other parkinsonian disorders in other samples (Hutton et al. 1998; Baker et al. 1999; Farrer et al. 2002); thus, it is believed that this is a true locus and that the increase in significance represents increased power. Although the APL may have greater power in nuclear families, the test, unlike PDT, is not valid in extended pedigrees, nor does it offer the flexibility to handle quantitative traits as the FBAT does. So each of these tests offers advantages in different situations.

We presented the APL test in terms of general mating-type parameters. These mating-type parameters can be replaced with a single allele-frequency parameter (for a biallelic marker) if we assume HWE, as was done in the TRANSMIT statistic and for the simulations discussed in this article. Use of the full parameterization of mating types allows for deviations from HWE and may be preferable in stratified samples. This flexibility is a strength of the APL test.

Our presentation has been for a single genetic marker; however, the APL test, like the procedure implemented in TRANSMIT, can use multiple markers in a joint haplotype analysis. Clayton (1999) extended the score statistic to test for association between marker haplotypes and a disease allele, but, again, the inference of missing parental genotypes is not valid in regions of linkage when there are multiple affected siblings. Conducting correct haplotype inference in regions of linkage is critical for fine-mapping. The APL test can be extended analogously to the extension of the test in TRANSMIT.

When considering marker haplotypes, even when full parental genotype data are available, there is another level of inference, since we do not directly observe marker haplotypes. For the *i*th family, T_i is defined as before; however, we must sum over possible haplotype-phase configurations for the family. The probabilities for each haplotype-phase configuration depend only on haplotype frequencies in the population under the null hypothesis, and these frequencies can be estimated using the EM algorithm. This approach is taken by Clayton (1999), and it is appropriate if there are no missing parental data and if we are willing to assume HWE. However, if there are missing parental data, it becomes necessary to account for linkage by including parameters for IBD and estimating mating types of haplotypes conditional on IBD. The steps described above for a single locus still apply, but more parameters are required to describe the mating-type probabilities. The primary limitation for the haplotype-based test is that it will likely be necessary to assume HWE unless only a few haplotypes exist in the population. Otherwise the number of haplotype mating-type parameters may become prohibitive. With the development of the haplotype map under way (Casci 2002), we expect to be guided in the selection of a few markers that tag a limited number of haplotypes. In this case, it may be possible to move away from the HWE assumption.

It is noteworthy that if all families have parental data, then the APL test is asymptotically equivalent to the test in TRANSMIT under the null hypothesis. The only difference is that the statistic in TRANSMIT uses a variance estimator that computes deviations around an estimated mean of the numerator, whereas the APL sta-

tistic uses deviations around the null value of 0. The variance estimator used in the TRANSMIT statistic can produce a more powerful test than the APL, but gains will be minimal unless the association is very strong. This estimator could be used in the APL statistic as well.

When testing for association, families with only affected siblings and no parental genotype data are not appropriate for the test in TRANSMIT in regions of linkage, and they are not informative for the PDT or FBAT. Our investigation showed that, although the APL test remains valid when ASP families with no parental data are used, such families provide little information for the test, and their inclusion can even lead to a decrease in power for some models. As discussed earlier, it is essential that family-based tests of association have some "control" data that give information about population allele frequencies for comparison with frequencies in affected individuals. Therefore, collecting families with only affected siblings does not seem to be an efficient use of resources for family-based tests of association, and, when parental data are unavailable, collecting unaffected siblings would strengthen the power of family-based association tests.

Acknowledgments

We are grateful for generous support from National Institute of Mental Health grant R01 MH59528, National Institute on Aging grant R01 AG20135, and the Morris K. Udall Parkinson's Disease Research Center of Excellence grant 5 P50 NS39764–03. We also thank Drs. Richard Morris and William Scott, for their critical review of the manuscript, and Dr. Andrew Allen, for assisting with power calculations for the FBAT.

Appendix A

Steps of the EM Algorithm to Estimate Mating-Type and IBD Parameters

Observed data are the counts $\{n_G\}$, where n_G is the number of ASP families for which siblings have genotypes in the set $\boldsymbol{G}$.

1. Specify initial estimates for mating-type and IBD parameters: μ_{Gp} and z_{Gp}.
2. E-Step: Compute expected full data. The expected count for sibling genotypes $\boldsymbol{G}$, parental genotype $\boldsymbol{G}_p$, and IBD $= k$ is:

$$\tilde{n}\{\boldsymbol{G},\boldsymbol{G}_p,k\} = n_G P(\boldsymbol{G}_p, IBD = k \mid \boldsymbol{G}, A) .$$

3. M-Step: Compute estimates for mating-type and IBD parameters, using MLE formulas:

$$\tilde{\mu}_{Gp} = \frac{\sum_G \sum_k \tilde{n}\{\boldsymbol{G},\boldsymbol{G}_p,IBD = k\}}{n}$$

and

$$\tilde{z}_k = \frac{\sum_G \sum_{Gp} \tilde{n}\{\boldsymbol{G},\boldsymbol{G}_p,IBD = k\}}{n} .$$

4. Replace initial parameter estimates with new estimates from step 3 and repeat steps 1–4 until estimates converge.

Electronic-Database Information

The URL for data presented herein is as follows:

Center for Human Genetics at Duke University, http://wwwchg.duhs.duke.edu/software/index.html (for SIMLA)

References

Abecasis GR, Cookson WOC, Cardon LR (2000) Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8:545–551

Baker M, Litvan I, Houlden H, Adamson J, Dickson D, Perez-Tur J, Hardy J, Lynch T, Bigio E, Hutton M (1999) Asso-

ciation of an extended haplotype in the tau gene with progressive supranuclear palsy. Hum Mol Genet 8:711–715

Bass M, Martin E, Hauser E (2002) Software for simulation studies of complex traits: SIMLA. Am J Hum Genet 71:569

Casci T (2002) Haplotype mapping: shortcut around the block. Nat Rev Genet 3:573

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

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B 39:1–38

Farrer M, Skipper L, Berg M, Bisceglio G, Hanson M, Hardy J, Adam A, Gwinn-Hardy K, Aasly J (2002) The τ H1 haplotype is associated with Parkinson's disease in the Norwegian population. Neurosci Lett 322:83–86

Horvath SM, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. Am J Hum Genet 63:1886–1897

Hutton M, Lendon CL, Rizzu P, Baker M, Froelich S, Houlden H, Pickering-Brown S, et al (1998) Association of missense and 5′-splice-site mutations in τ with the inherited dementia FTDP-17. Nature 393:702–705

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

Lake SL, Blacker D, Laird NM (2000) Family-based tests of association in the presence of linkage. Am J Hum Genet 67:1515–1525

Martin ER, Bass MP, Kaplan NL (2001*a*) Correcting for a potential bias in the pedigree disequilibrium test. Am J Hum Genet 68:1065–1067

Martin ER, Kaplan NL, Weir BS (1997) Tests for linkage and association in nuclear families. Am J Hum Genet 61:439–448

Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. Am J Hum Genet 67:146–154

Martin ER, Scott WK, Nance MA, Watts RL, Hubble JP, Koller WC, Lyons K, et al (2001*b*) Association of single-nucleotide polymorphisms of the τ gene with late-onset Parkinson disease. JAMA 286:2245–2250

Monks SA, Kaplan NL, Weir BS (1998) A comparative study of sibship tests of linkage and/or association. Am J Hum Genet 63:1507–1516

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

Scott WK, Nance MA, Watts RL, Hubble JP, Koller WC, Lyons K, Pahwa R, et al (2001) Complete genomic screen in Parkinson disease: evidence for multiple genes. JAMA 286:2239–2244

Shih MC, Whittemore AS (2002) Tests for genetic association using family data. Genet Epidemiol 22:128–145

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 68:1457–1462

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

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

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

Weinberg C (1999) Allowing for missing parents in genetic studies of case-parent triads. Am J Hum Genet 64:1186–1193

Weinberg CR, Wilcox AJ, Lie RT (1998) A log-linear approach to case-parent-triad data: assessing effects of disease genes that act either directly or through maternal effects and that may be subject to parental imprinting. Am J Hum Genet 62:969–978