REPORT

An Evaluation of Power and Type I Error of Single-Nucleotide Polymorphism Transmission/Disequilibrium–Based Statistical Methods under Different Family Structures, Missing Parental Data, and Population Stratification

Kristin K. Nicodemus, Augustin Luna, and Yin Yao Shugart

Researchers conducting family-based association studies have a wide variety of transmission/disequilibrium (TD)–based methods to choose from, but few guidelines exist in the selection of a particular method to apply to available data. Using a simulation study design, we compared the power and type I error of eight popular TD-based methods under different family structures, frequencies of missing parental data, genetic models, and population stratifications. No method was uniformly most powerful under all conditions, but type I error was appropriate for nearly every test statistic under all conditions. Power varied widely across methods, with a 46.5% difference in power observed between the most powerful and the least powerful method when 50% of families consisted of an affected sib pair and one parent genotyped under an additive genetic model and a 35.2% difference when 50% of families consisted of a single affection-discordant sibling pair without parental genotypes available under an additive genetic model. Methods were generally robust to population stratification, although some slightly less so than others. The choice of a TD-based test statistic should be dependent on the predominant family structure ascertained, the frequency of missing parental genotypes, and the assumed genetic model.

The testing of preferential transmission of alleles from parents to affected offspring is a common method of assessing association between genetic markers and disease status. In essence, transmission/disequilibrium (TD)–based methods compare the distribution of alleles transmitted to an affected offspring with the distribution of alleles not transmitted. Therefore, TD-based methods often claim to be robust to confounding from population admixture or stratification because the "case" and "control" alleles come from the same set of parents; therefore, they are expected to have identical genetic backgrounds. The role of population subdivision in the confounding of genetic epidemiologic case-control studies has been controversial.[1–8] Recently, several researchers have shown evidence of population stratification and that it can lead to spurious association, even in populations once thought to be homogeneous, such as Europeans.[9–11] However, choosing an appropriate TD-based method to employ in a particular study may be difficult because no practical guidelines have been provided about which method may be more powerful for a set of families with a particular structure. Although it is feasible to ascertain full parents-child trios when the disease being studied can be diagnosed during childhood (e.g., autism), studies of late-onset diseases (e.g., diabetes) often suffer from incomplete familial ascertainment.

In this report, we focus on the evaluation of various versions of software that have implemented TD-based tests, hoping to provide practical guidelines for users in terms of the choice of study design and the use of statistical methods. Our goal is to clearly illustrate decreases in power when an unwise or invalid test statistic is chosen and to dispel confusion about selection of a TD-based methodology. We sought to assess power and type I error by using a simulation study design to test eight commonly used TD-based methods while varying the genetic model (additive, dominant, or recessive), the family structures, and the frequencies of missing data. We also evaluated TD-based methods under population stratification. The statistical methods considered included the association-in-the-presence-of-linkage statistic (APL)[12]; the family-based association test (FBAT)[13,14]; the pedigree disequilibrium test (PDT)[15]; the sibship disequilibrium test (SDT)[16]; Spielman's TD test (TDT),[17] as implemented in Haploview[18] (hereafter referred to as the "Haploview TDT"); TDTPhase[19]; TRANSMIT,[20] with use of both analytically derived and permutation-based P values (permutation-implementation information is available at Transmit [version 2.5.4] Web site); and the Weinberg log-linear method.[21,22] The Weinberg log-linear method is implemented in a statistical analysis system (SAS)–based program and can be found at Clarice R. Weinberg's Web site. For ease of use of this method, we created an R script (available from K.K.N. at nicodemusk@mail.nih.gov) that

Am. J. Hum. Genet. 2007;80:178–185. © 2006 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8001-0018$15.00

takes a standard-formatted pedigree file (.ped file) as input and creates a count-based output file formatted for use with the SAS program. PDTPhase[19] is an implementation of the PDT,[15] which, when analysis involves a single SNP, performs identically to the PDT (data not shown) and will not be discussed further. Methods were selected on the basis of frequency of use in published applied-data analyses.

To get a baseline estimate of power and type I error for each method, we simulated 1,000 fully genotyped trios. In addition, to determine which methods fared best for late-onset diseases or in studies that have predominantly one parent ascertained, we simulated 1,000 incomplete trios with one parental genotype missing. The assessment of families with two offspring included the following simulation conditions: 50% of families ($n = 500$) were fully genotyped parents–affected child trios, and the remaining 50% of families ($n = 500$) were composed of an affection-discordant sibling pair without parental genotypes, an affection-discordant sibling pair with one parent genotyped, an affected sibling pair without parental genotypes, or an affected sibling pair with one parent genotyped, for a total of 18 simulation conditions. In addition, we simulated 1,000 trios and 1,000 incomplete trios under population stratification.

Simulation of 1,000 replicates per each association-present condition was conducted via SIMLA version 2.2.[23] The minor-allele frequency (MAF) for observed associated SNPs was set at 0.50, except in the simulations under population stratification (discussed below). Type I error was evaluated using a separate set of 1,000 data sets with the same family structure, simulated under the null hypothesis of no linkage and no association by use of Merlin version 1.0.1[24]; for the replicates simulated under population stratification, the MAFs under the null hypothesis were set to be the same as in the associated conditions, thus retaining the difference in MAFs. Power was calculated—with the α level held constant within TD-based methods—as the proportion of data sets showing significant evidence for association divided by the total number of replicates.

To mimic a realistic candidate-gene family-based study with locus heterogeneity, in all associated conditions, we simulated 75% of the families to have association between the observed associated SNP and disease status and 25% of the families to not have association between disease status and the observed SNP. The nonobserved disease-allele frequency was set to 20% for simulations not under population stratification. For the population-stratification simulations, the nonobserved disease-allele frequency was set to 20% for the first population and to 10% for the second population, and observed allele frequencies varied between the populations: population 1 had an observed associated-allele frequency of 50%, and population 2 had an observed associated-allele frequency of 25%. Disease-allele penetrances for the additive conditions were 0, 0.25, and 0.50, for the three possible genotypes; penetrance increased as the number of disease-associated alleles increased. For dominant disease models, the disease-allele carrier penetrance was 0.50, and, for recessive disease models, the disease-allele homozygote penetrance was 0.50. Simulations considering population stratification used the additive model. Although underlying genetic models used to generate data were varied, all analyses were done blind to genetic model, to closely approximate applied analyses in which the genetic model is unknown. The linkage disequilibrium between the unobserved disease allele and the observed associated marker allele was simulated to be incomplete ($D' = 0.50$), to mimic applied association studies.

All methods showed type I error close to 0.05 in fully genotyped trios (table 1), although several methods (FBAT, Haploview TDT, and TRANSMIT) appeared slightly conservative (empirical type I error < 0.04). PDT/PDTPhase was the most powerful method under all genetic models, although with a slight increase in type I error (table 1). TRANSMIT was the least powerful method for fully genotyped trios. All other methods performed equally well under the additive and dominant models. Under the recessive model, all methods performed strongly. Among the three methods that perform association testing of incomplete trios (APL, TDTPhase, and TRANSMIT), we observed difference in power by genetic model (table 1). Under the additive model, APL and TDTPhase showed equivalent power, performing significantly better than TRANSMIT. However, under the dominant model, APL was clearly more powerful than TDTPhase and TRANSMIT, and, under the recessive model, TDTPhase showed much higher power to detect association than did TRANSMIT or APL, although the empirical type I error for TDTPhase was slightly anticonservative (0.068).

Within the condition of families with 50% affected sib pairs with or without a single parent genotyped, TDTPhase consistently showed high power to detect association (table 2). In the condition of 50% affected sib pairs without parental genotypes, both TDTPhase and TRANSMIT with use of the analytically derived P values performed similarly well; after a single pair of parents' genotypes were added, APL, TDTPhase, and TRANSMIT with use of analytically derived P values showed the highest power to detect association. As expected, the log-linear model was more powerful than the Haploview TDT under conditions with 50% affected sib pairs and one parental genotype available, because the log-linear method is able to use the incomplete trio families. Indeed, under a recessive model, the log-linear method is as powerful as APL, TDTPhase, and TRANSMIT. The Haploview TDT, PDT, and FBAT with either variance estimate showed the lowest power to detect association in all affected sib pair conditions. The greatest differences in power for affected sib pair families was 44% under the condition of one parental genotype available with the use of an additive model (in the comparison between TDTPhase and PDT). All methods performed very well under recessive genetic models.

Under conditions of 50% affection-discordant families

Table 1. Power and Type I Error of TD-Based Methods: Full Trios and Incomplete Trios

Conditions	APL		FBAT		PDT/PDTPhase		Haploview TDT		TDTPhase		Transmit: Analytical		Transmit: Permutation		Log-Linear TDT	
	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)
Full trios:																
Additive model	4.9	64.6 (60.1–69.1)	3.9	63.6 (59.1–68.1)	5.5	79.2 (76.0–82.4)	3.9	63.5 (59.2–67.8)	4.5	63.6 (59.1–68.1)	3.8	49.4 (44.5–54.3)	3.7	48.2 (43.3–53.1)	4.5	53.5 (48.6–58.4)
Dominant model	4.9	51.8 (46.9–56.7)	3.9	50.8 (45.9–55.7)	5.5	68.1 (63.8–72.4)	3.9	50.5 (45.6–55.4)	4.5	54.4 (49.5–59.3)	3.8	39.9 (35.2–44.6)	3.7	40.0 (35.3–44.7)	4.5	45.7 (40.8–50.6)
Recessive model	4.9	99.2 (99.0–99.4)	3.9	98.9 (98.8–99.0)	5.5	1.0	3.9	98.9 (98.9–99.0)	4.5	98.9 (98.9–99.0)	3.8	95.3 (94.4–96.2)	3.7	94.8 (94.3–95.3)	4.5	98.0 (97.6–98.4)
Incomplete trios:																
Additive model	4.8	37.4 (32.8–42.0)	...	...	...	...	...	...	6.8	33.2 (28.9–37.5)	5.0	19.6 (16.5–22.7)	4.5	20.3 (17.1–23.5)	...	...
Dominant model	4.8	30.7 (26.5–34.9)	...	...	...	...	...	...	6.8	22.5 (19.1–25.9)	5.0	13.7 (11.4–16.0)	4.5	13.5 (11.2–15.8)	...	...
Recessive model	4.8	66.4 (62.0–70.8)	...	...	...	...	...	...	6.8	90.2 (88.5–91.9)	5.0	69.7 (65.6–73.8)	4.5	69.8 (65.7–73.9)	...	...

Note.—Data are percentages.

Table 2. Power and Type I Error of TD-Based Methods: 50% Fully Genotyped Trios and 50% Affected Sib Pairs

Conditions	APL		FBAT Empirical Variance		FBAT		PDT/PDTPhase		Haploview TDT		TDTPhase		Transmit: Analytical		Transmit: Permutation		Log-Linear TDT	
	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)
Affected sib pairs:																		
Additive model	4.6	52.2 (47.3–57.1)	5.4	39.1 (34.4–43.8)	5.1	37.7 (33.1–42.3)	5.3	28.3 (24.3–32.3)	5.1	37.7 (33.1–42.3)	4.6	60.9 (56.2–65.6)	5.2	57.2 (52.4–62.0)	5.1	50.9 (46.0–55.8)	6.6	34.9 (30.4–39.4)
Dominant model	4.6	34.5 (30.0–38.9)	5.4	33.1 (28.8–37.4)	5.1	31.6 (27.4–35.8)	5.3	23.4 (19.9–26.9)	5.1	31.9 (27.6–36.2)	4.6	46.6 (41.7–51.5)	5.2	45.3 (40.4–50.2)	5.1	39.4 (34.7–44.1)	6.6	28.1 (26.1–30.1)
Recessive model	4.6	96.6 (96.0–97.2)	5.4	92.2 (90.8–93.6)	5.1	92.2 (90.8–93.6)	5.3	87.0 (84.8–89.2)	5.1	92.2 (90.8–93.6)	4.6	98.1 (97.7–98.5)	5.2	96.5 (95.8–97.2)	5.1	88.7 (86.7–90.7)	6.6	89.1 (87.2–91.0)
Affected sib pairs and one parent:																		
Additive model	5.8	73.8 (70.0–77.6)	4.9	41.0 (36.3–45.7)	5.0	41.5 (36.7–46.3)	4.8	29.8 (25.7–33.9)	5.0	41.5 (36.7–46.3)	5.4	76.3 (72.8–79.8)	5.2	72.8 (68.9–76.7)	5.1	68.4 (64.2–72.6)	5.6	57.3 (52.5–62.1)
Dominant model	5.8	52.0 (47.1–56.9)	4.9	31.5 (27.3–35.7)	5.0	31.5 (27.3–35.7)	4.8	22.6 (19.2–26.0)	5.0	31.9 (27.6–36.2)	5.4	60.3 (55.6–65.0)	5.2	57.6 (52.8–62.4)	5.1	52.3 (47.4–57.2)	5.6	42.8 (38.0–47.6)
Recessive model	5.8	99.1 (98.9–99.3)	4.9	92.2 (90.8–93.6)	5.0	92.2 (90.8–93.6)	4.8	86.5 (84.2–88.8)	5.0	92.2 (90.8–93.6)	5.4	99.0 (98.8–99.2)	5.2	98.9 (98.9–99.0)	5.1	93.4 (92.2–94.6)	5.6	97.2 (96.7–97.7)

Note.—Data are percentages.

ascertained, APL and TRANSMIT with use of either analytically derived or permutation-based P values showed the highest power, compared with other test statistics (table 3). FBAT and PDT performed reasonably well under all genetic models when 50% of families consisted of discordant sib pairs without parental genotypes; however, both methods showed less-than-optimal power in the conditions of discordant sibship plus one parental genotype under an additive or dominant model. Interestingly, the opposite trend was observed with the TDTPhase method; this method performed strongly among families with one parental genotype and discordant sib pairs but performed less well when both parental genotypes were missing. The methods giving the lowest power to detect association were the Haploview TDT, the SDT, and the log-linear TDT under conditions of 50% discordant sibships with missing genotype data for both parents. The greatest difference in power was observed for the discordant sibship with no parental data with use of a dominant model: 34.2% (in the comparison between FBAT and the log-linear TDT). The main explanation for why the Haploview TDT and the log-linear TDT performed less well than the other methods is that both of these methods were able to use only 50% of the families in each replicate because the implementation of both methods cannot use families with both parental genotypes missing. The recessive genetic model condition gave the highest power, and the dominant model gave the lowest.

Virtually all methods showed reduced power under the presence of population stratification (table 4). This reduction in power was most likely caused by the reduction in MAFs and the resulting smaller number of informative families, although the loss of power was very moderate for fully genotyped trios (variations ranged from 1.8% higher power for TRANSMIT with permutation-based P values to 6.3% lower power for PDT/PDTPhase) and was only ~10% for incomplete trios (reductions ranged from 7.1% lower power for TRANSMIT with analytical P values to 13.7% lower power for APL) versus simulations conducted without population stratification. As expected, for fully genotyped trios, the type I error was not increased for joint association and linkage methods such as PDT, the log-linear TDT, and the Haploview TDT. Similarly, we did not observe an increase in type I error in association-based methods such as TRANSMIT and TDTPhase. PDT/PDTPhase had the highest power (72.9%) to detect association under population stratification, followed by the Haploview TDT (62.6% power) and TDTPhase (61.7% power). However, with incomplete trios, we observed a slightly anticonservative type I error for the association-based method TDTPhase (0.062) but not for the other association-based method, TRANSMIT (both analytical and permutation-based type I error < 0.04). For incomplete trios, the joint linkage and association method APL (23.7% power) and the association method TDTPhase (25.2% power) showed equivalent power, although APL did not show an inflated type I error rate. Both APL and TDTPhase

showed significantly increased power versus TRANSMIT (12.5% and 11.3% power with use of analytic and permutation-based P values, respectively) for incomplete trios.

Here, we report differences in power across family type, missing-parental-data conditions, genetic models, and population stratification for several commonly used TD-based methods implemented in freely available software. In planning and executing candidate-gene or genomewide association data analyses with the use of family-based association tests, it is important to consider the type of sibship ascertained (concordant or discordant for disease status), the proportion of missing parental genotypes, and the assumed genetic model as key factors in the decision making about appropriate TD-type statistics applied to achieve the highest likelihood of detecting association when association exists. Simulation study designs are useful in assessing the behavior of test statistics under a restricted set of scenarios, but generalization to all possible scenarios must be cautiously applied. However, we believe our recommendations are a helpful addition to the applied researcher's selection of appropriate test statistics, and we are encouraged by the ability of all methods to detect association under locus heterogeneity, incomplete LD, and incomplete penetrance, which are likely to be found in applied candidate-gene association studies.

In our view, the reason for such vast differences in power across TD-based test statistics is twofold. First, there are differences in the number of families that are informative for different methods; second, how each method handles missing parental genotypes varies (table 5). In the simulations with affected sibling pairs and missing parental data, both FBAT and PDT cannot use 50% of the families in the analysis, so the resulting loss in power is not surprising. We also note that the Haploview TDT considers only families with fully genotyped parents, so it used only trios for all conditions in the calculation of the test statistic. A graphical-user-interface–based software for the Windows operating system that implements the full method proposed by Spielman et al.[17] and that can include families with missing parental genotypes is available from the Spielman Lab: TDT & S-TDT Web site. This software also implements the S-TDT[28] and the combined TDT/S-TDT[28] statistic and will perform better than the method implemented in Haploview under conditions of missing parental data and discordant sibships, because it can use a larger number of sibships and can score transmissions to unaffected offspring. In addition, the log-linear method has been implemented only for trio data, although it can assess association in families with some missing parental data.[22,23] The SDT uses only discordant sibships, so the results in table 3 are based on 50% of families. We include these "unfair" comparisons as an illustration for researchers who have to decide which test is most appropriate and to show the magnitude of decreases in power that occur when an unwise or invalid choice of test statistic is made. TRANSMIT infers missing parental genotypes by using sib-

Table 3. Power and Type I Error of TD-Based Methods: 50% Fully Genotyped Trios and 50% Discordant Sib Pairs

Conditions	APL		FBAT		PDT/PDTPhase		SDT		Haploview TDT		TDTPhase		Transmit: Analytical		Transmit: Permutation		Log-Linear TDT	
	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)
Discordant sib pairs:																		
Additive model	4.3	60.7 (56.0–65.4)	4.6	58.3 (53.5–63.1)	4.6	58.6 (53.8–63.4)	4.7	26.5 (22.7–30.3)	3.9	32.3 (28.0–36.6)	3.5	45.9 (41.0–50.8)	4.2	59.4 (64.6–64.1)	4.1	59.0 (54.3–63.7)	4.0	26.5 (22.7–30.3)
Dominant model	4.3	54.2 (49.3–59.1)	4.6	57.9 (63.1–62.7)	4.6	57.1 (52.3–61.9)	4.7	28.5 (24.5–32.5)	3.9	30.4 (26.3–34.5)	3.5	38.6 (34.0–43.2)	4.2	55.3 (50.5–60.1)	4.1	55.7 (50.9–60.5)	4.0	22.7 (19.3–26.1)
Recessive model	4.3	99.4 (99.3–99.5)	4.6	99.5 (99.4–99.6)	4.6	99.5 (99.4–99.6)	4.7	75.7 (72.1–79.3)	3.9	90.5 (88.8–92.2)	3.5	96.8 (96.2–97.4)	4.2	99.5 (99.4–99.6)	4.1	99.5 (99.4–99.6)	4.0	85.1 (82.1–87.6)
Discordant sib pairs and one parent:																		
Additive model	5.1	70.1 (66.0–74.2)	4.4	58.9 (54.2–63.6)	4.3	58.0 (53.2–62.8)	4.6	27.6 (23.7–31.5)	4.5	36.7 (32.1–41.3)	5.4	69.6 (65.5–73.7)	5.1	68.4 (64.2–72.6)	4.8	66.9 (62.6–71.2)	4.5	50.8 (45.9–55.7)
Dominant model	5.1	60.1 (55.4–64.8)	4.4	52.1 (47.2–57.0)	4.3	52.0 (47.1–56.9)	4.6	28.6 (24.6–32.6)	4.5	30.4 (26.3–34.5)	5.4	60.4 (55.7–65.1)	5.1	62.4 (57.8–67.0)	4.8	61.3 (56.7–65.9)	4.5	41.5 (36.7–46.3)
Recessive model	5.1	95.3 (94.4–96.2)	4.4	94.0 (92.9–95.1)	4.3	93.8 (95.7–94.9)	4.6	76.1 (72.5–79.7)	4.5	91.2 (89.6–92.8)	5.4	92.4 (91.0–93.8)	5.1	93.2 (92.0–94.4)	4.8	92.6 (91.3–93.9)	4.5	82.0 (79.1–84.9)

Note.—Data are percentages.

Table 4. Power and Type I Error of TD-Based Methods under Population Stratification and the Additive Model: Fully Genotyped Trios and Incomplete Trios

Condition	APL		FBAT		PDT/PDTPhase		Haploview TDT		TDTPhase		Transmit: Analytical		Transmit: Permutation		Log-Linear TDT	
	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)	Type I Error	Power (95% CI)
Trios	3.9	59.2 (54.5–63.9)	5.2	61.7 (57.1–66.3)	4.5	72.9 (69.0–76.8)	5.5	62.6 (60.3–64.9)	5.0	61.7 (57.1–66.3)	5.2	49.3 (46.8–51.8)	5.4	50.0 (47.5–52.5)	4.4	48.3 (43.4–53.2)
Incomplete trios	3.6	23.7 (21.9–25.5)	...	...	...	...	...	...	6.2	25.2 (23.3–27.1)	3.8	12.5 (11.2–13.4)	3.0	11.3 (10.3–12.3)	...	...

Note.—Data are percentages.

Table 5. Permissible Family Structures, Markers Accepted, and Input Data File Types of TD-Based Methods Evaluated

Method	Permissible Family Structure(s)	Parental Genotypes	Type(s) of Markers Accepted	Data File Type(s)[a]
APL	Parent-child trios, affected sibships, discordant sibships	Complete and incomplete	SNPs	Full LINKAGE pedigree file and data file
FBAT	Parent-child trios, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	Abbreviated pedigree file
FBAT: empirical variance estimate	Parent-child trios, affected sibships, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	Abbreviated pedigree file
PDT/PDTPhase	Parent-child trios, affected sibships, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	PDT: full LINKAGE pedigree file and data file; PDTPhase: abbreviated pedigree file
SDT	Discordant sibships	None	SNPs, multiallelic markers	Abbreviated pedigree file
Haploview TDT	Parent-child trios	Complete	SNPs	Abbreviated pedigree file
TDTPhase	Parent-child trios, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	Abbreviated pedigree file
TRANSMIT: analytical	Parent-child trios, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	Abbreviated pedigree file
TRANSMIT: permutation	Parent-child trios, discordant sibships	Complete and incomplete	SNPs, multiallelic markers	Abbreviated pedigree file
Log-linear TDT	Parent-child trios	Complete and incomplete	SNPs	Count-based input file[b]

[a] Full LINKAGE pedigree file contains the following columns before genotype data: family number, individual number, father's identification (ID) number, mother's ID number, first parental offspring's ID number, next paternal sibling's ID number, next maternal sibling's ID number, sex, proband status, and affection status. The LINKAGE data file is a separate file that lists descriptions of the data contained in the pedigree file. An abbreviated pedigree file contains only the following columns before genotype data: family number, individual number, father's ID number, mother's ID number, sex, and affection status.

[b] An R script is available from K.K.N. (at nicodemusk@mail.nih.gov), to recode abbreviated pedigree files into count-based input files.

ling genotype information, which is appropriate in situations where there are not multiple affected siblings and there is no linkage or when discordant sibships are used.[12] However, it has been shown that the score test used in TRANSMIT has an inflated type I error rate when there are missing parental genotypes and multiple affected siblings and when the MAF is <0.50 because of increased allele sharing between siblings as a result of linkage.[12] The APL statistic handles correlation of transmitted alleles imposed by linkage between affected siblings and thus retains the appropriate type I error rate[12]; therefore, the use of APL may be preferable to TRANSMIT when multiple affected siblings are included in analysis, missing parental genotypes are frequent, and the observed MAF is not equal to 0.50. TDTPhase implements the full likelihood described by Clayton[20] and conducts a likelihood-ratio test between nested models. However, the parental part of the likelihood depends on a population-based model for parental genotype frequencies, which can provide incorrect inference if the population model for parental genotype frequencies is misspecified,[19] especially for a high frequency of missing parental genotypes. The subsequent statistical test is an unconditional logistic regression comparing "case" genotype frequencies with "pseudocontrol" genotype frequencies and is not robust to population stratification unless a permutation procedure that can be time consuming is used,[19] although we observed only a modest increase in type I error (6.2%) in the incomplete-trios condition simulated under population stratification. In addition, the test statistic calculated in TDTPhase is also likely to suffer from the same bias as the score statistic from TRANSMIT, resulting in an inflated type I error rate when multiple affected offspring are ascertained, because it does not take into account correlation between transmissions to affected offspring when linkage is present. Our recommendation is to use APL when a large proportion of families have multiple affected siblings and missing parental genotype data are frequent, to assure appropriate type I error, but to consider APL, TRANSMIT, and TDTPhase as the test statistics of choice when discordant sibships are the predominant ascertained family structure. Under population stratification in fully genotyped trios, we suggest the use of PDT/PDTPhase, TDTPhase, or the Haploview TDT, but, within incomplete trios, we suggest the use of APL to retain adequate power with appropriate type I error rate.

We note that, for the purposes of this study, we employed the joint null hypothesis of no linkage and no association as our null hypothesis of interest, even though some methods are robust to linkage in the detection of association. As discussed by Laird and Lange,[25] the joint null hypothesis may be the null hypothesis of interest when considering a candidate gene that is not under a linkage peak or for genomewide association studies. Therefore, the results reported in the present study cannot be generalized to studies designed to follow up on regions identified as potentially harboring a disease gene from a linkage study.

One condition not considered in the present study is that of informatively missing parental genotypes. When a parental genotype is correlated with its probability of being missing, the distribution of observed genotypes differs from that of the missing genotypes, and thus the missing parental genotypes are informatively missing.[26,27] Informatively missing parental genotypes may be induced through population stratification[26] or association between a disease that creates a higher probability of missingness—for example, an aggressive form of cancer[26] or a debilitating psychiatric disorder, such as schizophrenia—and a genetic marker. Many TD-based tests that allow for missing parental data assume that the distributions of observed and unobserved parental genotypes are the same. Simulation studies have shown that this assumption, if violated, can result in inflated type I error.[26] Two methods have been proposed for use under informative missingness to retain appropriate type I error rates.[26,27] Because informative missingness may be a common feature of family-based genetic association studies, a planned simulation study will consider methods assessed in the present study plus the methods proposed by Allen et al.[26] and Chen,[27] to determine how several TD-based methods fare under varying scenarios of informatively missing parental data.

Replication of candidate-gene TD-based association results has been inconsistent for many diseases and may, in part, be because of differences in power of different TD methods used in applied analyses. Given the disparate power of different TD test statistics, we remain optimistic that some failures to replicate can be reconciled by use of the most appropriate TD-based methodology available, using our simulation study results as a guide to selection of a more appropriate TD-based test statistic, given the family structure, type of sibship ascertained, and genetic model, if known.

Web Resources

The URLs for data presented herein are as follows:

Biowulf Cluster at NIH, http://biowulf.nih.gov/

Clarice R. Weinberg's Web site, http://dir.niehs.nih.gov/dirbb/weinberg/weinberg.htm

Spielman Lab: TDT & S-TDT, http://genomics.med.upenn.edu/spielman/TDT.htm

Transmit (version 2.5.4), http://www-gene.cimr.cam.ac.uk/clayton/software/transmit.txt

References

1. Wacholder S, Rothman N, Caporaso N (2000) Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 92:1151–1158

2. Millikan RC (2001) Re: population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 93:156–157

3. Wacholder S, Rothman N, Caporaso N (2001) Response: re: population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst 93:157–158

4. Thomas DC, Witte JS (2002) Point: population stratification: a problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev 11:505–512

5. Wacholder S, Rothman N, Caporaso N (2002) Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. Cancer Epidemiol Biomarkers Prev 11:513–520

6. Khlat M, Cazes MH, Génin E, Guiguet M (2004) Robustness of case-control studies of genetic factors to population stratification: magnitude of bias and type I error. Cancer Epidemiol Biomarkers Prev 13:1660–1664

7. Heiman GA, Hodge SE, Gorroochurn P, Zhang J, Greenberg DA (2004) Effect of population stratification on case-control association studies. I. Elevation in false positive rates and comparison to confounding risk ratios (a simulation study). Hum Hered 58:30–39

8. Heiman GA, Gorroochurn P, Hodge SE, Greenberg DA (2005) Roubustness of case-control studies to population stratification. Cancer Epidemiol Biomarkers Prev 14:1579–1582

9. Helgason A, Yngvadóttir B, Hrafnkelsson B, Gulcher J, Stefánsson K (2005) An Icelandic example of the impact of population structure on association studies. Nat Genet 37:90–95

10. Campbell CD, Ogburn EL, Lunetta KL, Lyon HN, Freedman ML, Groop LC, Altshuler D, Ardlie KG, Hirschhorn JN (2005) Demonstrating stratification in a European American population. Nat Genet 37:868–872

11. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al (2005) Population structure, differential bias and genomic control in a large-scale, case-control association study. Nat Genet 37:1243–1246

12. Martin ER, Bass MP, Hauser ER, Kaplan NL (2003) Accounting for linkage in family-based tests of association with missing parental genotypes. Am J Hum Genet 73:1016–1026

13. Horvath S, Xu X, Laird NM (2001) The family based association test method: strategies for studying general genotype-phenotype associations. Eur J Hum Genet 9:301–306

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

15. 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

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

17. 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

18. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21:263–265

19. Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25:115–121

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

21. 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

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

23. Bass MP, Martin ER, Hauser ER (2004) Pedigree generation for analysis of genetic linkage and association. Pac Symp Biocomput 9:93–103

24. Abecasis GR, Cherney SS, Cookson WO, Cardon LR (2002) Merlin—rapid analysis of dense genetic maps using sparse gene flow trees. Nat Genet 30:97–101

25. Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7:385–394

26. Allen AS, Rathouz PJ, Satten GA (2003) Informative missingness in genetic association studies: case-parent designs. Am J Hum Genet 72:671–680

27. Chen YH (2004) New approach to association testing in case-parental designs under informative missingness. Genet Epidemiol 27:131–140