Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Li and Li, J Biomet Biostat 2011, 2:5
DOI: 10.4172/2155-6180.1000126

Review Article Open Access

A Family-based Association Method for Pedigree Including Half-Sib Data

Yen-Wei Li[1,2] and Yi-Ju Li[2,3]*

[1]Statistics Department, North Carolina State University, Raleigh, NC
[2]Center for Human Genetics, Duke University, Medical Center, Durham, NC
[3]Department of Biostatistics and Bioinformatics, Duke University, Medical Center, Durham, NC

Abstract

Family datasets could provide good resources for association studies as an initial investigation or a replication study. The current family-based association tests analyze data only from full sibships of a nuclear family or extended pedigrees of related nuclear families. In order to fully exert all possible information in the family data, we propose a "Pedigrees with Half-sibs Association Test" (PHAST) to accommodate half-siblings if they are available. PHAST adopts the idea of transmission score from the Pedigree Disequilibrium Test (PDT) to construct the test statistic. The difference is that it utilizes the identity-by-descent (IBD) information of the marker between sibling pairs (full or half sibs) to construct an allelic transmission statistic. If parental genotypes are missing, EM algorithm is used to infer parental genotypes and compute transmission scores for all possible scenarios. The computer simulation results suggested that our new method has valid type I error rates under varied family structures. Our method could have more power than PDT and FBAT when the sample size of half-sibs increases, especially the families without parental genotypes. In conclusion, our method can serve as an alternative method of the existing family-based association tests. Furthermore, it can relax the ascertainment criteria for studying late onset diseases since limited siblings are available.

Keywords: Family-based association; Half-siblings; Identity by descent; Late onset disease

Introduction

Association studies have been a predominant approach for analyzing densely spaced genetic markers (e.g. single nucleotide polymorphisms (SNPs)) for detecting genetic variants that may lead to susceptibility genes or genetic modifiers for the traits of interest (disease or quantitative traits). Since the development of transmission/disequilibrium test (TDT) [1], family-based study design was viewed as a robust approach because of less chances of encountering false positive association results due to population stratification existed in the population samples. However, the development of several prominent association methods such as GENOMIC CONTROL [2,3], STRUCUTRE [4], EIGENSTRAT [5] methods has lessened the concern and changed our view of the case-control study design significantly. As the results, many genome wide association studies (GWAS) utilized easier obtained population-based samples rather than family-based samples.

While case-control design seems to dominate the current practice of the association study, particularly GWAS, familial samples still remain some advantages. First, many familial samples were collected during the era of whole genome linkage scans. These family datasets provide good resources for association studies as either an initial investigation or a replication dataset for other association findings. Second, family data are more robust for detecting genotyping errors (e.g. Mendelian errors) than population-based samples. Third, family data have higher accuracy in inferring haplotypes and confirming the role of rare variants to disease susceptibility. Finally, as the next-generation sequencing is becoming an important approach to pinpoint the causal variant, familial samples have the advantage for the first pass of sequencing effort to discover both common and rare variants. Overall, familial samples will still remain important in genetic studies of human complex diseases.

Many family-based association methods have been developed in the past decade. The TDT was the pioneer of all methods, which tests for association in the presence of linkage using parent-offspring triads. Many extensions of the TDT method were developed for various pedigree structures afterward. The sibling TDT (S-TDT) [6] and several other tests [7,8,9] were designed to accommodate discordant sibpairs without parental genotype data. More general methods such as the pedigree disequilibrium test (PDT) [10], and the family-based association test (FBAT) [11] can accommodate multiple affected offspring.

The current family-based association tests share a characteristic, which is to analyze data only from full sibships of a nuclear family, such as parent-offspring triads, parents and multiple affected full siblings, discordant full sibpairs of large size, or extended pedigrees of related nuclear families. Therefore, full siblings or parents have been the target of ascertainment in genetic research. In this study, we explore a new method that can relax the recruitment criteria such as accommodating half-siblings in addition to full siblings if they are available. We also allow our method to handle multiple affected siblings that may exist from pedigrees recruited for linkage studies.

To accommodate these different family structures, we developed a *Pedigrees with Half-sibs Association Test* (PHAST) to fully exert all possible information in the family data. The development of PHAST was based on the framework of the pedigree disequilibrium test (PDT) but with several new features: (1) applicable to both full siblings or combinations of full and half siblings family data; (2) inferring parental genotypes when they are missing rather than using siblings information to compute test statistics; and (3) accounting for linkage when parental genotypes are missing. Although this new method may have similar properties as the existing methods, it will largely benefit ascertainment

***Corresponding author:** Yi-Ju Li, Center for Human Genetics, DUMC Box 3445, Duke University, Medical Center, Durham, NC27710, Tel: 919-684-0604; Fax: 919-684-0921; E-mail: yiju.li@duke.edu

Received November 03, 2011; **Accepted** November 27, 2011; **Published** December 25, 2011

Citation: Li YW, Li YJ (2011) A Family-based Association Method for Pedigree Including Half-Sib Data. J Biomet Biostat 2:126. doi:10.4172/2155-6180.1000126

process, particularly for aging related diseases when limited siblings are available to ascertain.

Methods

For simplicity, we start with the pedigree structure of affected sibpair (ASP) with parents and concordant half sibpair (HSP) with parents (Figure 1A). Families with more than two affected siblings could be addressed by following the same procedures. We assume that the marker tested is a biallelic marker with a risk allele 1. Our method adopts the concept of allelic transmission presented in PDT but utilizes the identity-by-descent (IBD) information of the marker between a pair of individuals to construct test statistic. For any ASP family, if the two siblings share 2 or 1 IBD at the target marker, we treat the ASP family as a whole unit to compute the allelic transmission score. If the two siblings share 0 IBD at the target marker, we split the ASP family to two independent case-parent trios. The same rules are applied in HSP family except that no 2 IBD sharing between half sibpairs will occur. The mathematical forms of this strategy are described below for different family structures.

Concordant sibpairs with parents

Define a random variable X for allelic transmission score, which is computed as the number of different copies allele 1 transmitted to the two affected sibs minus the number of allele 1 non-transmitted.

For the i^{th} family (F_i),

$$X_i = \sum_{k=0}^{2} X_{ik} P(IBD = k \mid G_p, G_s) I(F_i \notin Trio) + X_{iT} I(F_i \in Trio), \quad (1)$$

where

X_{i0} = (# of allele 1 transmitted to G_{S1}) – (# of allele 1 not transmitted to G_{S1})

+ (# of allele 1 transmitted to G_{S2}) – (# of allele 1 not transmitted to G_{S2}).

X_{i1}, X_{i2} = (# of different allele 1 transmitted in F_i) – (# of allele 1 not transmitted in F_i).

X_{iT} is the number of allele 1 transmitted minus the number of allele 1 non-transmitted for a case-parent trio *(Trio)*. $G_S = (G_{S1}, G_{S2})$ is the marker genotypes for sibling 1 and 2, and G_p is the parental genotypes, that is $G_p = (G_{P1}, G_{P2})$ for ASP family and $G_p = (G_{P1}, G_{P2}, G_{P3})$ for HSP family. $P(IBD = k \mid G_p, G_s)$ is the probability that the two affected siblings share k IBD, k = 0, 1, or 2 (APPENDIX I). Various computer programs can estimate IBD probability. We used MERLIN [12] for estimating IBD probabilities at a particular marker.

The value of X is observed by counting the number of copes of allele 1 in the affected siblings. Under the null hypothesis of no association between the marker and disease alleles, X should have an expected value of 0. To make computation easy for accounting both ASP and HSP units, we looked into the different status of IBD. Under 0 IBD, the two copies of markers of each sibling inherited from totally different source which indicates these two siblings are independent in allelic transmission. Therefore, X_{i0} is formulated to be the summation of two PDT case-parent trios as the total allelic transmission and sharing score for both ASP and HSP. Under 1 IBD, in order to assure the expectation of statistic X to be 0 under the null hypothesis, the number of "different" copies of alleles transmitted must be equal to the number of alleles non-transmitted. For 1 IBD sharing HSP, in genotypes (G_{S1}, G_{S2}), there are three different copies of allele transmitted (*S1* and *S2* sharing one identical copy from the common parent *P2*) and three alleles remain non-transmitted in (G_{P1}, G_{P2}, G_{P3}). However, this is not the case in ASP family. For 1 IBD sharing ASP, there are three different copies of allele transmitted but the number of allele nontransmitted is always one. In order to get correct transmission statistic, we transform the original two parents ASP into a pseudo HSP family (Figure 1A) by repeating one parent *P1* as *P3*. For instance, assume (G_{P1}, G_{P2}) with genotypes (11, 12) and (G_{S1}, G_{S2}) with genotypes (11, 11), if *S1* and *S2* share 1 IBD, the transformed pseudo HSP family is (G_{P1}, G_{P2}, G_{P1}) with genotypes (11, 12, 11) and the corrected allelic transmission score will be X_{i1}= 3 – 2 = 1 (Table 1). Under 2 IBD, which will only occur in ASP unit, it will always be equal between the number of different copies of alleles transmitted and the number of alleles nontransmitted. Therefore, for the 2 IBD sharing ASP family, we can calculate transmission statistic directly without any transformation. In the previous example, if *S1* and *S2* share 2 IBD, G_{S1} and G_{S2} are with the same pair of alleles transmitted from the parents *P1* and *P2*. The different copies allele 1 transmitted are 2, but not 4, and X_{i2}= 2- (# of allele 1 not transmitted in F_i) = 2 – 1 = 1.

Discordant sibpairs with parents or parent-offspring triads

For full and half sibpairs with different disease status (discordance sibpairs, DSP) (Figure 1B,1C) or simple parent-offspring triads, there is only one affected sibling in each family. The random variable X= X_{iT} (the last part of equation (1)) and will follow the PDT framework.

More than two affected siblings

We also formulate PHAST method for taking into account multiple affected siblings (> two affecteds). The same inference principles described for concordance sibpairs are applied. In the example of nuclear family with three affected siblings (Figure 1D), we identify the IBD status between each pair of siblings first. Split the nuclear family into two independent units if sharing 0 IBD, transform nuclear family

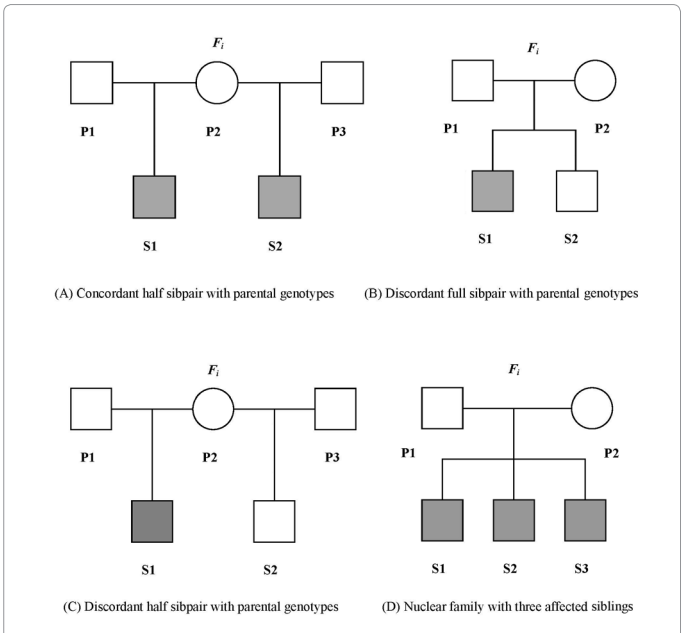

Figure 1: Pedigree Structures.

to pseudo HSP family if sharing 1 IBD, and then count the allelic transmission score number X. The possible X scores for nuclear family with three affected siblings are listed in Table 2.

Missing parents --inferring parental genotypes

When parental genotypes are missing, PDT method does not infer parental genotypes but utilizes information from discordant sibpairs. However, the allelic sharing scores may vary by the number of available sibling samples genotyped. In our proposed method, we took a step further to infer parental genotypes and compute allelic transmission scores for all possible parental genotypes. At the same time, in order to accommodate linkage between a maker and disease locus, the probability of affected siblings sharing k allele IBD, $Z_k = P(IBD = k \mid G_p)$, need to be taken into account. We used the Expectation Maximization (EM) [13] algorithm to estimate the conditional probability $P(G_p \mid G_s)$ and IBD parameters Z_k. The probability of parental genotypes based on offspring genotypes is formulated as

$$P(G_p \mid G_s) = \frac{P(G_p)\sum_{k=0}^{2} P(G_s \mid G_p, IBD = k) Z_k}{P(G_s)}.$$

Then, the expected random variable in family i can be estimated by $X_i = \sum_{j \in C} P(G_{pj} \mid G_{sj}) X_j$. For ASP and HSP without parental genotypes, the value of X_j is calculated by equation (1) consistent with G_{sj} and the jth set of possible parental genotypes G_{pj}. For missing parental genotypes DSP cases, X_j is the number of allele 1 transmitted minus the number of allele 1 non-transmitted for each inferred parental genotypes, as defined in PDT method under the scenario of known parental genotypes.

The PHAST test statistic

We define a general association statistic D_i based on all ASP, HSP, and DSP subunits in pedigree i.

For the ith pedigree, where $i = 1, \ldots, N$, a family-specific score

$$D_i = \sum_{j=1}^{F_{iA}} X_{ij} + \sum_{j=1}^{F_{iH}} X_{ij} + \sum_{j=1}^{F_{iD}} X_{ij},$$

where F_{iA}, F_{iH}, and F_{iD} are the number of ASP, HSP, and DSP families in pedigree i.

Under the null hypothesis, we can derive

$$E(D_i) = \sum_{j=1}^{F_{iA}} E(X_{ij}) + \sum_{j=1}^{F_{iH}} E(X_{ij}) + \sum_{j=1}^{F_{iD}} E(X_{ij}) = 0.$$

Therefore, the PHAST statistic can be written as

$$T = \frac{\sum_{i=1}^{N} D_i}{\sqrt{\sum_{i=1}^{N} D_i^2}},$$

where T follows an asymptotic standard normal distribution under the null hypothesis of no linkage disequilibrium.

Simulation Studies

A series of computer simulations was implemented to study the validity of PHAST. For each simulation, 10,000 replicates were

(G_{P1}, G_{P2})	(G_{S1}, G_{S2})	IBD status	X Statistic
(11, 12)	(11, 11)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 3 – 2 = 1 X_2 = 2 – 1 = 1
(11, 12)	(11, 12)	0 IBD 1 IBD 2 IBD	X_0 = 1 – 1 = 0 X_1 = 2 – 2 = 0 X_2 = none
(11, 12)	(12, 12)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 1 – 2 = -1 X_2 = 1 – 2 = -1
(12, 12)	(11, 11)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = none X_2 = 2 – 0 = 2
(12, 12)	(11, 12)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 2 – 1 = 1 X_2 = none
(12, 12)	(11, 22)	0 IBD 1 IBD 2 IBD	X_0 = 2 – 2 = 0 X_1 = none X_2 = none
(12, 12)	(12, 12)	0 IBD 1 IBD 2 IBD	X_0 = 0 – 0 = 0 X_1 = 1 – 2 = -1 or 2 – 1 = 1 X_2 = 1 – 1 = 0
(12, 12)	(12, 22)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 1 – 2 = -1 X_2 = none
(12, 12)	(22, 22)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = none X_2 = 0 – 2 = -2
(22, 12)	(22, 22)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 0 – 1 = -1 X_2 = 0 – 1 = -1
(22, 12)	(22, 12)	0 IBD 1 IBD 2 IBD	X_0 = 1 – 1 = 0 X_1 = 1 – 1 = 0 X_2 = none
(22, 12)	(12, 12)	0 IBD 1 IBD 2 IBD	X_0 = none X_1 = 1 – 0 = 1 X_2 = 1 – 0 = 1

Table 1: X Statistic in ASP family.

generated to estimate type I errors and statistical power. A nominal significance level of 0.05 was used for all estimates.

We compared PHAST with two alternative methods: FBAT [11] and PDT [10].

Assume a bi-allelic disease locus A with alleles A_1 and A_2 (allele frequencies p_1 and p_2) and a single marker M with alleles M_1 and M_2 (allele frequencies q_1 and q_2). Linkage disequilibrium (LD) between the disease locus and the marker was set as

$$D = P(A_1M_1) - p_1q_1, \quad (2)$$

where $P(A_1M_1)$ is population haplotype frequency for A_1M_1.

To generate simulation data, four population haplotype frequencies for disease locus A and marker M were calculated by $P(A_1M_1) = p_1q_1 + D$, $P(A_1M_2) = p_1q_2 - D$, $P(A_2M_1) = p_2q_1 - D$, and $P(A_2M_2) = p_2q_2 + D$. The haplotypes for parental population were generated based on these frequencies. We assume random mating in the population and form two haplotypes for each offspring by randomly drawing one haplotype from each parent. Genetic markers were simulated under the assumption of complete linkage to the disease locus. Three genetic models (recessive, additive, and dominant) were considered through different disease penetrances. The model parameters are given in Table 3. Disease phenotypes were simulated based on disease locus genotypes and their corresponding penetrances.

The type I error was studied under the null hypothesis of no association between the disease and marker alleles ($D = 0$). We generated replicate samples N= 200 and N =500 of families with different disease models and five types of family structures, ASP, HSP, DSP, discordant half sibpairs (DHSP) with and without parental genotypes, and nuclear family with three affected siblings, in type I error simulations. To evaluate the power and examine the half sibpairs power contribution, three combinations of different types of family structures were used: (1) 200 families with different ratios of DSP to DHSP, (2) 200 families with different ratios of ASP to HSP, and (3) 200 nuclear families with three affected siblings.

Results

Type I error rates

Tables 4 and 5 present the type I error rates for PHAST, PDT, and FBAT tests in 200 HSP, ASP, and DSP with and without parental genotypes for different simulated genetic models. In the cases that parental genotypes are known, Tables 4 and 5 show that type I error estimates for most tests are very close to the nominal significance level of 0.05. For the scenario of concordant sibpairs without parental genotypes (ASP and HSP), since PDT and FBAT cannot address this type of data, no estimates were obtained for both programs, and a nominal level of type I error estimates was obtained for PHAST.

Table 6 shows type I error rates for data simulated from 200 HSP families under different ratios of 1 IBD to 0 IBD families. Both PHAST and PDT are very robust under different ratios of 1 IBD to 0 IBD families. However, FBAT tends to have inflated type I error when the ratio of 1 IBD cases are high and conservative type I error when the 1 IBD ratio is low.

(G_{P1}, G_{P2})	(G_{S1}, G_{S2},G_{S3})	X Statistic
(11, 12)	(11, 11 ,11)	X = 1
	(11, 11, 12)	X = 0
	(11, 12, 12)	X = 0
	(12, 12, 12)	X = -1
(12, 12)	(11, 11 , 11)	X = 2
	(11, 11 , 12)	X = 1
	(11, 11 , 22)	X = 0
	(11, 12 , 12)	X = 1 or 0
	(11, 12 , 22)	X = 0
	(11, 22 , 22)	X = 0
	(12, 12 , 12)	X = 0
	(12, 12 , 22)	X = -1 or 0
	(12, 22 , 22)	X = -1
	(22, 22 , 22)	X = -2
(22, 12)	(22, 22 , 22)	X = -1
	(22, 22 , 12)	X = 0
	(22, 12 , 12)	X = 0
	(12, 12 , 12)	X = 1

Table 2: X Statistic in the nuclear family with three affected siblings.

Parameters used in the simulation study	
Disease allele frequency $P(A_1)$	0.3, 0.5
Single Marker allele frequency $P(M_1)$	0.3, 0.5
Two-locus haplotype frequencies (P_{M1N1},P_{M1N2},P_{M2N1},P_{M2N2})	(0.3, 0.2, 0.1, 0.4)
GRR2 (=P(affected| A_1A_1)/ P(affected| A_2A_2)*	0.15, 0.20
Disease prevalence	0.1
Number of families simulated	200, 500
Number of iterations	10,000

*GRR2: the homozygous genotypic risk ratio

Table 3: Parameters used in the simulation study.

200 HSP		(D= 0, GRR2 = 1.5)			
		With parental genotypes		Without parental genotypes	
Model	Method	P(M)=P(D)=0.3	P(M)=P(D)=0.5	P(M)=P(D)=0.3	P(M)=P(D)=0.5
Recessive	PHAST	0.0507	0.0484	0.0514	0.0429
	PDT	0.0507	0.0484	NA	NA
	FBAT	0.0503	0.0494	NA	NA
Additive	PHAST	0.0477	0.0499	0.0520	0.0427
	PDT	0.0477	0.0499	NA	NA
	FBAT	0.0476	0.0517	NA	NA
Dominant	PHAST	0.0493	0.0464	0.0539	0.0455
	PDT	0.0493	0.0464	NA	NA
	FBAT	0.0489	0.0498	NA	NA

Table 4: Type I error rates for data simulated from 200 concordant half sibpair (HSP) families under different genetic models.

200 ASP		(D= 0, GRR2 = 1.5, P(M)=P(D)=0.3)			
		With parental genotypes		Without parental genotypes	
Model	Method	ASP	DSP	ASP	DSP
Recessive	PHAST	0.0468	0.0496	0.0508	0.0539
	PDT	0.0468	0.0478	NA	0.0492
	FBAT	0.0479	0.0488	NA	0.0492
Additive	PHAST	0.0503	0.0521	0.0576	0.0545
	PDT	0.0503	0.0484	NA	0.0480
	FBAT	0.0513	0.0514	NA	0.0480
Dominant	PHAST	0.0520	0.0497	0.0557	0.0526
	PDT	0.0520	0.0505	NA	0.0461
	FBAT	0.0518	0.0500	NA	0.0461

Table 5: Type I error rates for data simulated from 200 concordant full (ASP) and discordant full sibpair (DSP) families under different genetic models.

200 HSP	(D= 0, GRR2 = 1.5, P(M)=P(D)=0.3)			
# 1 IBD families	# 0 IBD families	PHAST	PDT	FBAT
200	0	0.0450	0.0450	0.1179
150	50	0.0491	0.0491	0.0835
100	100	0.0497	0.0497	0.0499
50	150	0.0503	0.0503	0.0211
0	200	0.0486	0.0486	0.0023

Table 6: Type I error rates for data simulated from 200 concordant half sibpair (HSP) families under different ratios of 1 IBD to 0 IBD families.

Overall, we show that PHAST has correct type I error estimates under different scenario of family structure, and can handle missing parents for concordant sibpairs (full or half sibs). Families with parental genotypes generally show slightly lower type I error rates than those without parents cases. This was expected because the overall sample size is large when parental genotypes are available.

Power estimates

We also carried out simulations for all combinations of genetic models and different pedigree structures to evaluate the statistical power for PHAST method. Because similar power patterns were found in dominant, additive, and recessive models, we only present the results of additive model here.

Generally, power will increase when the degree of linkage disequilibrium (D) between marker and disease locus increases. For the family structure with parental genotypes cases, PHAST has similar statistical power with PDT and FBAT. For example, when D = 0.21 (maximum LD scenario for marker and disease allele frequencies at 0.3 by equation (2)) for 200 HSP with parents, the power of PHAST is 0.602 and powers of PDT and FBAT are 0.602 and 0.606. Although the statistical power under the maximum LD was not strong, this is mostly due to the small size of data and the lack of parental genotypes.

Figure 2 shows the results of power comparison among PHAST, PDT, and FBAT under different ratios of DSP without parents to DHSP without parents. A set of 200 families of difference combinations of DSP and DHSP without parents were simulated for each replicate. It is noted that in this simulation study, PDT and FBAT do not use data from DHSP without parents. Therefore, only PHAST method can utilize the full dataset. The results in Figure 2 show that PHAST has increasing power as the proportion of DHSP without parents increasing while PDT and FBAT have decreased power.

Figure 3 shows the results of power comparison among PHAST, PDT, and FBAT under different ratios of concordant ASP without parents to concordant HSP without parents. Similarly, 200 families were simulated for each replicate. The purpose of this simulation is to show that both PDT and FBAT cannot analyze such data while PHAST can handle them to increase some power. For 200 ASPs, PHAST reaches power 0.323 in the maximum disequilibrium (D = 0.21). In the case of 150 ASP and 50 HSP combination, the power of PHAST is 0.188 when D = 0.21. This implies that the power gains from

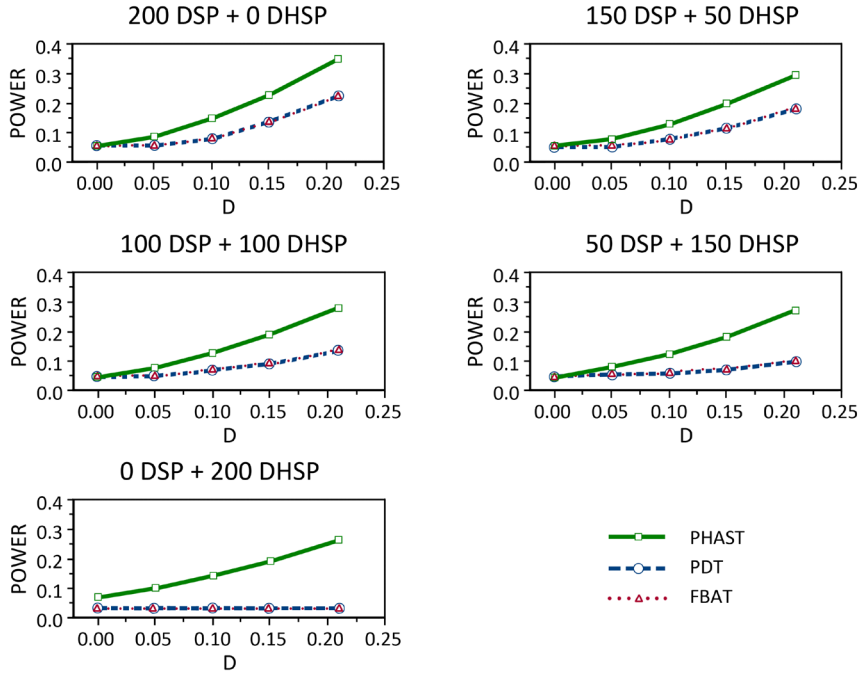

Figure 2: Power comparison among different association methods under different ratios of discordant full sibpairs (DSP) to discordant half sibpairs (DHSP) without parental data in both cases.

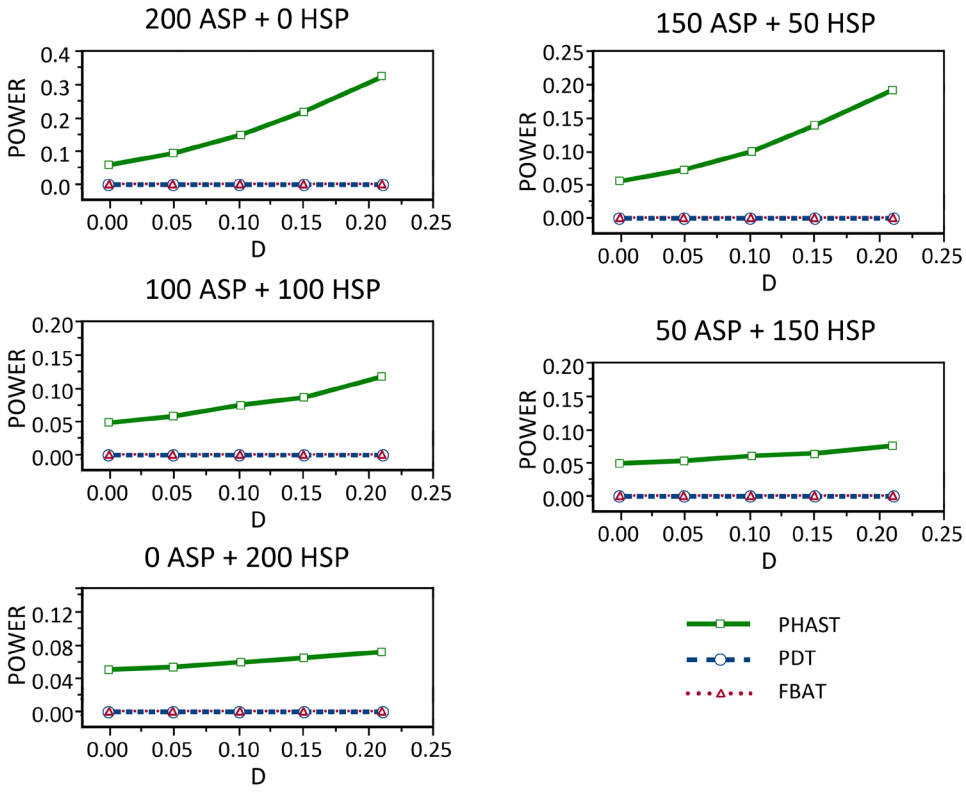

Figure 3: Power comparison among different association methods under different ratios of concordant full sibpairs (ASP) to concordant half sibpairs (HSP) without parental data in both cases.

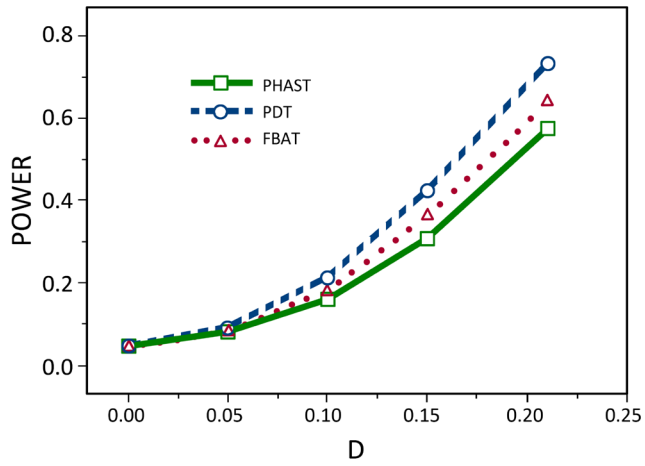

Figure 4: Power comparison among different association methods for 200 known parental genotypes nuclear family with three affected siblings under additive genetic model. Both maker and disease allele frequencies were 0.3. The heritability was 0.1 and GRR2 = 2.0.

HSP without parents are small due to the limited sibling genotypes available for inferring missing parental genotypes in HSP. The above simulation results for either type of family structure -- siblings without parental genotype information, show that PHAST will not reach up to 80% power. However, it shows that PHAST can handle more family types and add to the statistical power. We mentioned earlier that we expanded PHAST statistic to handle more than two affected siblings. Our simulation showed that with additional siblings, power to detect risk effects is improved. Figure 4 shows power among PHAST, PDT, and FBAT for 200 families with three affected siblings and known parental genotypes simulated under the additive model. When D = 0.21, the power of PHAST is 0.576, while the power of PDT and FBAT reach 0.735 and 0.641. However, the type I error of PHAST is closest to the significance level of 0.05 (PHAST: 0.049; PDT: 0.043; FBAT: 0.046). The difference in the analysis strategy between PHAST and the other two methods is that PHAST treated the whole family (two parents+3 affecteds) as a whole unit while PDT and FBAT take families with three affected siblings as three independent trios.

Discussion

In this paper, we proposed the PHAST approach to test for association with the inclusion of half-siblings or more than two affected sibling data. Like other family-based association methods, such as PDT and FBAT, PHAST can also handle the conventional family data structures such as trios, ASP with parents, and DSP. The simulation results showed that PHAST method has correct type I error under varied family structures. We also studied the properties of the PHAST, PDT, and FBAT test statistics as HSP families contain different ratios of 1 IBD to 0 IBD families. It is important to point out that type I error rates in the FBAT test are inflated as the ratio of half sib-pair sharing 1 IBD increases. Therefore, the power of FBAT method could lead to misinterpretation when the data with half-siblings are used.

We compared our method with two alternative methods, PDT and FBAT. We found the PHAST and PDT version to be, in most cases, highly correlated. Thus in most cases, the PHAST will provide similar power to detect an association over current methods and will, for some specific genetic models, offer a gain in power. Simulations revealed that PHAST could have more power than PDT and FBAT when the sample size of half-siblings increases, especially the families without parental genotypes. Therefore, PHAST can be considered as a useful tool for studying late onset disease.

In summary, the PHAST method can serve as an alternative for the PDT method when either full- or half-siblings, or both, are available. Moreover, PHAST is potentially applicable to genetic studies with special features. For instance, for the aging related diseases, available samples to be recruited may be limited and parents are mostly not available. In this case, if the expansion to half-siblings is possible, PHAST will be able to handle this type of data. Thus, we hope that PHAST can be additional tool for researchers to facilitate future studies. Accordingly, this method should be able to expand the scope of the current family-based ascertainment strategy, and hopefully add insights into the genetic association study of human complex disease.

Acknowledgements

We are grateful for generous support from a research grant for American Federal for Aging Research (AFAR), a new investigator grant (NIRG-02-3603) and an investigator initiative research grant (IIRG-05-14708) from Alzheimer's association.

References

1. Spielman RS, McGinnis RE, Ewens WJ (1993) Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). American Journal of Human Genetics 52: 506-516.
2. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55: 997-1004.
3. Bacanu SA, Devlin B, Roeder K (2002) Association studies for quantitative traits in structured populations. Genetic epidemiology 22: 78-93.
4. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
5. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nature 38: 904-909.
6. Spielman RS, Ewens WJ (1998) A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. The American Journal of Human Genetics 62: 450-458.
7. Horvath S, Laird NM (1998) A discordant-sibship test for disequilibrium and linkage: no need for parental data. The American Journal of Human Genetics 63: 1886-1897.
8. Boehnke M, Langefeld CD (1998) Genetic association mapping based on discordant sib pairs: the discordant-alleles test. The American Journal of Human Genetics 62: 950-961.
9. Knapp M (1999) The transmission/disequilibrium test and parental-genotype reconstruction: the reconstruction-combined transmission/disequilibrium test. The American Journal of Human Genetics 64: 861-870.
10. Martin ER, Monks SA, Warren LL, Kaplan NL (2000) A test for linkage and association in general pedigrees: the pedigree disequilibrium test. The American Journal of Human Genetics 67: 146-154.
11. Rabinowitz D, Laird N (2000) A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. Human Heredity 50: 211-223.
12. Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002) Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30: 97-101.
13. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society Series B 39: 1-38.

APPENDIX I

$$P(IBD=k\,|\,G_p,G_s)$$

$$=\frac{P(G_s\,|\,G_p,IBD=k)P(IBD=k\,|\,G_p)}{P(G_s\,|\,G_p)}$$

$$=\frac{P(G_s\,|\,G_p,IBD=k)Z_k}{P(G_s\,|\,G_p)}$$

$$=\frac{P(G_s\,|\,G_p,IBD=k)Z_k}{\sum_{k=0}^{1}P(G_s\,|\,G_p,IBD=k)Z_k}$$

where

$$Z_k=P(IBD=k\,|\,G_p)$$