A Powerful Nonparametric Statistical Framework for Family-Based Association Analyses

Ming Li,^{*1} Zihuai He,^{*1} Daniel J. Schaid,^{*} Mario A. Cleves,^{*} Todd G. Nick,^{*} and Qing Lu^{§,2}

^{*}Division of Biostatistics, Department of Pediatrics, University of Arkansas for Medical Sciences, Little Rock, Arkansas 72202,

[†]Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, [‡]Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota 55905, and [§]Department of Epidemiology and Biostatistics, Michigan State University,

East Lansing, Michigan 48824

ORCID ID: 0000-0002-7943-966X (Q.L.)

ABSTRACT Family-based study design is commonly used in genetic research. It has many ideal features, including being robust to population stratification (PS). With the advance of high-throughput technologies and ever-decreasing genotyping cost, it has become common for family studies to examine a large number of variants for their associations with disease phenotypes. The yield from the analysis of these family-based genetic data can be enhanced by adopting computationally efficient and powerful statistical methods. We propose a general framework of a family-based *U*-statistic, referred to as family-*U*, for family-based association studies. Unlike existing parametric-based methods, the proposed method makes no assumption of the underlying disease models and can be applied to various phenotypes (e.g., binary and quantitative phenotypes) and pedigree structures (e.g., nuclear families and extended pedigrees). By using only within-family information, it can offer robust protection against PS. In the absence of PS, it can also utilize additional information (i.e., between-family information) for power improvement. Through simulations, we demonstrated that family-*U* attained higher power over a commonly used method, family-based association tests, under various disease scenarios. We further illustrated the new method with an application to large-scale family data from the Framingham Heart Study. By utilizing additional information (i.e., between-family information), family-*U* confirmed a previous association of *CHRNA5* with nicotine dependence.

KEYWORDS population stratification; pedigree structure; within-family information; between-family information; nicotine dependence

BOTH population-based and family-based designs have been commonly used in genetic association studies. Case-control study is a typical population-based design, by which unrelated cases are recruited and compared with healthy controls. When cases and controls come from different ethnicity backgrounds, they could have distinct ancestry distribution, leading to false-positive association results due to population stratification (PS) (Knowler *et al.* 1988; Freedman *et al.* 2004). Many statistical methods have been proposed to

address the issue of PS (Devlin and Roeder 1999; Pritchard and Rosenberg 1999; Pritchard *et al.* 2000a,b; Satten *et al.* 2001; Chen *et al.* 2003; Price *et al.* 2006; Bauchet *et al.* 2007). However, these methods can infer only the average effect of PS based on a large number of genetic variants, but not the locus-specific PS. Population stratification related to a particular locus may vary and deviate from the average effect. In the presence of locus-specific PS, these approaches may over-adjust or underadjust the PS effect, leading to either low power or inflated type I error (Marchini *et al.* 2004; Qin and Zhu 2012). Unlike population-based studies, family-based studies offer robust protection against PS (Weinberg *et al.* 1998; Cardon and Palmer 2003; Weinberg 2003). In a typical family-based association study, the alleles transmitted to affected individuals are compared with those untransmitted, which provide pseudocontrols with the same genetic ancestry as cases. In addition, family-based studies also allow for the investigation of other hypotheses, e.g., parent-of-origin effects and maternal-fetal interaction effects (Weinberg *et al.* 1998;

Copyright © 2015 by the Genetics Society of America

doi: 10.1534/genetics.115.175174

Manuscript received January 28, 2015; accepted for publication February 23, 2015; published Early Online March 5, 2015.

Supporting information is available online at <http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175174/-DC1>.

The Framingham Heart Study data set used for the analyses was obtained from the National Center for Biotechnology Information database of genotypes and phenotypes (NCBI dbGaP) through accession no. phs000128.v3.p3.

¹These authors contributed equally to this work.

²Corresponding author: Department of Epidemiology and Biostatistics, 909 Fee Rd., Room 601, East Lansing, MI 48824. E-mail: qlu@epi.msu.edu

Sinsheimer *et al.* 2003; Weinberg 2003; Cordell *et al.* 2004). Despite these advantages, family-based studies also have a few disadvantages compared to population-based studies, such as higher requirement of resources (*i.e.*, time and cost) to recruit families and lower statistical power given similar genotyping efforts (Gordon *et al.* 2001, 2004; Cordell and Clayton 2005; Laird and Lange 2006). Because of different strengths and weaknesses between population-based and family-based studies, they should be viewed as complementary rather than competitive strategies in genetic research of complex human diseases (Laird and Lange 2006).

For family-based association studies, the most widely adopted statistical method is the transmission disequilibrium test (TDT) (Spielman *et al.* 1993). The TDT considers the heterozygous parents for an allele that is putatively associated with disease and compares the frequency of the allele being transmitted to affected offspring with that of its alternate. The original TDT method was designed only for family triads with dichotomous phenotypes. It was later extended to accommodate various pedigree structures, as well as quantitative characteristics (Lazzeroni and Lange 1998; Spielman and Ewens 1998; Fulker *et al.* 1999; Martin *et al.* 2000; Rabinowitz and Laird 2000; Lange *et al.* 2003). These extensions have substantially improved the testing power and flexibility of the original TDT, while inheriting the same strength as the TDT (*i.e.*, being robust against PS).

The information in a family-based association study can be generally decomposed into two sources: within-family information and between-family information (Fulker *et al.* 1999). Most conventional family-based methods utilize only within-family information and gain a major advantage over population-based methods for their immunity to PS. However, without using the between-family information, conventional family-based methods could also have a reduced statistical power compared to population-based methods. Statistical methods are greatly needed to utilize both sources of information for more powerful family-based association analysis, especially in the absence of PS. Abecasis *et al.* (2000) proposed a variance-component method to decompose the information into within-family and between-family sources for nuclear families with normally distributed phenotypes. A number of hybrid testing strategies were also suggested to first prioritize SNPs by using between-family information and further test the association by using only within-family information (Van Steen *et al.* 2005; Ionita-Laza *et al.* 2007; Murphy *et al.* 2008). Other work has also suggested unifying two sources of information into a single test statistic by combining the *P*-values of the family-based test with the rank-based *P*-values of the population-based test (Won *et al.* 2009).

In this article, we propose a nonparametric statistical framework, family-*U*, for family-based association analysis. Similar to conventional family-based methods, it can be built on within-family information and generally offers robustness to PS. In the absence of PS, family-*U* is also flexible to integrate the between-family information with the

within-family information to enhance the statistical power of the association test. The proposed family-*U* method has the following properties: (1) It is an entirely nonparametric method without any assumption of the underlying disease model or phenotypic distribution and can be used for analyzing both binary and quantitative phenotypes; (2) it is flexible for all types of pedigree structures, including nuclear families and extended pedigrees; and (3) it utilizes information from all family members for power improvement and offers a convenient way to handle large multigeneration pedigrees. We evaluated the performance of family-*U* and compared it to a commonly used method, family-based association tests (FBAT). The proposed method was further illustrated via an application to the family data from the Framingham Heart Study.

Methods

Notations and hypothesis

We introduce notations, using quantitative phenotypes as an example. The extension to binary phenotypes follows straightforwardly. Suppose we have a study population with N subjects collected from M pedigrees. The i th pedigree comprises n_i subjects, where $\sum_{i=1}^M n_i = N$. We denote $y_{i,j}$ as the phenotype of the j th subject from the i th pedigree; $x_{i,j}$ as its genotype of a particular single-nucleotide polymorphism (SNP), where $x_{i,j} \in \{AA, Aa, aa\}$; $1 \leq i \leq M$; and $1 \leq j \leq n_i$. Let $Y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,n_i})'$ and $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n_i})'$ be the phenotypes and genotypes for the family members in the i th pedigree, $1 \leq i \leq M$. Assuming the SNP is associated with phenotypes, the phenotypic value of the j th individual in pedigree i can be expressed as

$$y_{i,j} = \mu_{AA}I(x_{i,j} = AA) + \mu_{Aa}I(x_{i,j} = Aa) + \mu_{aa}I(x_{i,j} = aa) + \varepsilon_{i,j},$$

where μ_{AA} , μ_{Aa} , and μ_{aa} are the expected phenotypic values for individuals with genotypes AA, Aa, and aa, respectively. $I(\cdot)$ is an indicator function, and $\varepsilon_{i,j}$ is the random error with

$$E(\varepsilon_{i,j}) = 0, \quad \text{for any } 1 \leq i \leq M, 1 \leq j \leq n_i, \\ \text{and}$$

$$= \begin{cases} \sigma^2 & i_1 = i_2, j_1 = j_2 \\ \sigma_1 & \text{if } i_1 = i_2, j_1 \neq j_2; \quad 1 \leq i_1, i_2 \leq M, 1 \leq j_1, j_2 \leq n_i \\ 0 & i_1 \neq i_2. \end{cases}$$

For simplicity, we assume here the covariance is the same for any two members from the same pedigree. In the following section, we discuss alternative ways of modeling relationships between family members (*e.g.*, σ_{PC} for parent-offspring pairs and σ_{CC} for sibling pairs). To examine the association between the phenotypes and genotypes, we can test the following hypothesis:

$$H_0 : \mu_{AA} = \mu_{Aa} = \mu_{aa} \text{ vs. } H_A : \text{they are not all equal.} \quad (1)$$

Family-U statistics

In this section, we propose two types of family-U statistics: a within family-U statistic using within-family information only and an overall family-U statistic using both within-family and between-family information.

Within family-U statistics: We first propose a within family-U statistic based on the within-family information, which is generally robust to PS. For any pedigree i , we first define $S_{i,AA} = \{j | x_{i,j} = AA\}$ as the members of the pedigree i who carry genotype AA and $S_{i,AA}^c = \{j | x_{i,j} \neq AA\}$ as those carrying genotype Aa/aa. Denote $|S_{i,AA}| = n_{i,AA}$ as the number of family members with genotype AA in pedigree i and $|S_{i,AA}^c| = n_i - n_{i,AA}$ as the number of family members with genotype Aa/aa in pedigree i . Further, denote a set of pedigree indexes as $D_{AA} = \{i | \exists 1 \leq j_1, j_2 \leq n_i, \text{ st } x_{i,j_1} = AA, x_{i,j_2} \neq AA\}$, which includes all pedigrees with members carrying genotype AA and genotype Aa/aa. The within family-U statistic can be defined as

$$\begin{aligned} U_{AA}^{\text{Within}} &= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} \frac{1}{n_{i,AA}(n_i - n_{i,AA})} \sum_{j_1 \in S_{i,AA}} \sum_{j_2 \in S_{i,AA}^c} \varphi(y_{i,j_1}, y_{i,j_2}) \\ &= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} (\hat{\mu}_{i,AA} - \hat{\mu}_{i,AA^c}), \end{aligned}$$

where the kernel function is chosen as $\varphi(y_{i,j_1}, y_{i,j_2}) = y_{i,j_1} - y_{i,j_2}$, so that $\hat{\mu}_{i,AA}$ and $\hat{\mu}_{i,AA^c}$ are the estimated sample means for members carrying genotype AA and genotype Aa/aa in the i th pedigree, respectively. The within family-U statistic considers only pedigrees with at least one individual carrying the AA genotype and one individual carrying the Aa/aa genotype. For each of these pedigrees, it calculates the phenotype difference of individuals carrying AA with those of individuals carrying Aa/aa and then averages it across pedigrees. Therefore, individuals carrying genotype AA and those carrying Aa/aa are compared only if they are in the same pedigree. In a similar manner, within family-U statistics can be defined for two other genotypes,

$$\begin{aligned} U_{Aa}^{\text{Within}} &= \frac{1}{|D_{Aa}|} \sum_{i \in D_{Aa}} \frac{1}{n_{i,Aa}(n_i - n_{i,Aa})} \sum_{j_1 \in S_{i,Aa}} \sum_{j_2 \in S_{i,Aa}^c} \varphi(y_{i,j_1}, y_{i,j_2}) \\ &= \frac{1}{|D_{Aa}|} \sum_{i \in D_{Aa}} (\hat{\mu}_{i,Aa} - \hat{\mu}_{i,Aa^c}) \end{aligned}$$

$$\begin{aligned} U_{aa}^{\text{Within}} &= \frac{1}{|D_{aa}|} \sum_{i \in D_{aa}} \frac{1}{n_{i,aa}(n_i - n_{i,aa})} \sum_{j_1 \in S_{i,aa}} \sum_{j_2 \in S_{i,aa}^c} \varphi(y_{i,j_1}, y_{i,j_2}) \\ &= \frac{1}{|D_{aa}|} \sum_{i \in D_{aa}} (\hat{\mu}_{i,aa} - \hat{\mu}_{i,aa^c}). \end{aligned}$$

Each of these U -statistics is expected to follow asymptotically a univariate normal distribution. A multivariate U -

statistic, $U^{\text{Within}} = (U_{AA}^{\text{Within}}, U_{Aa}^{\text{Within}}, U_{aa}^{\text{Within}})$, can thus be formed to test the null hypothesis expressed in Equation 1. We reject the null hypothesis if the within family-U statistic, U^{Within} , significantly deviates from 0.

Overall family-U statistics: In the absence of PS, between-family information can also be used for power improvement. Therefore, we introduce an overall family-U statistic that combines both between-family information and within-family information. We first define $S_{AA} = \{(i,j) | x_{i,j} = AA\}$ and $S_{AA}^c = \{(i,j) | x_{i,j} \neq AA\}$. Let $N_{AA} = |S_{AA}|$ and $N - N_{AA}$ be the number of individuals with genotype AA and genotype Aa/aa in the entire study population, respectively. We consider a two-group U -statistic comparing individuals carrying genotype AA and individuals carrying genotypes Aa/aa,

$$\begin{aligned} U_{AA}^{\text{Overall}} &= \frac{1}{N_{AA}(N - N_{AA})} \sum_{(i_1, j_1) \in S_{AA}} \sum_{(i_2, j_2) \in S_{AA}^c} \varphi(y_{i_1, j_1}, y_{i_2, j_2}), \\ &= (\hat{\mu}_{AA} - \hat{\mu}_{AA^c}), \end{aligned}$$

where $\hat{\mu}_{AA}$ and $\hat{\mu}_{AA^c}$ are the estimated sample means for individuals carrying genotype AA and genotype Aa/aa in the study population. The overall family-U statistic compares individuals carrying genotype AA with those carrying genotypes Aa/aa within a pedigree and across different pedigrees and thus combines both within-family and between-family information for detecting an association. In a similar fashion, two-group U -statistics can be defined for the other genotypes (i.e., Aa and aa),

$$\begin{aligned} U_{Aa}^{\text{Overall}} &= \frac{1}{N_{Aa}(N - N_{Aa})} \sum_{(i_1, j_1) \in S_{Aa}} \sum_{(i_2, j_2) \in S_{Aa}^c} \varphi(y_{i_1, j_1}, y_{i_2, j_2}) \\ &= (\hat{\mu}_{Aa} - \hat{\mu}_{Aa^c}) \end{aligned}$$

$$\begin{aligned} U_{aa}^{\text{Overall}} &= \frac{1}{N_{aa}(N - N_{aa})} \sum_{(i_1, j_1) \in S_{aa}} \sum_{(i_2, j_2) \in S_{aa}^c} \varphi(y_{i_1, j_1}, y_{i_2, j_2}) \\ &= (\hat{\mu}_{aa} - \hat{\mu}_{aa^c}). \end{aligned}$$

Based on these three U -statistics, we form a multivariate U -statistic,

$$U^{\text{Overall}} = (U_{AA}^{\text{Overall}}, U_{Aa}^{\text{Overall}}, U_{aa}^{\text{Overall}}),$$

to test the null hypothesis expressed in Equation 1. We reject the null hypothesis if the overall family-U statistic (U^{Overall}) significantly deviates from 0.

Family-U statistics for binary phenotypes: The method can be simply extended to handle binary phenotypes. For the binary phenotypes, we use a different kernel function,

$$\begin{aligned} \varphi'(y_{i_1, j_1}, y_{i_2, j_2}) &= \frac{\varphi(y_{i_1, j_1}, y_{i_2, j_2}) + 1}{2} \\ &= \begin{cases} 1 & \text{if } y_{i_1, j_1} = 1; y_{i_2, j_2} = 0. \\ 0.5 & \text{if } y_{i_1, j_1} = y_{i_2, j_2} \\ 0 & \text{if } y_{i_1, j_1} = 0; y_{i_2, j_2} = 1. \end{cases} \end{aligned}$$

Both within family- U and overall family- U statistics can then be defined accordingly. By using the above kernel function, the within family- U is an extended stratified Mann–Whitney U considering family correlations, and the overall family- U is equivalent to an extended Mann–Whitney U with the consideration of correlations among family members.

Statistical inference

In [Supporting Information](#), File S1, we show that under the null hypothesis of no association, both within family- U and overall family- U statistics have asymptotically a degenerate multivariate normal distribution with one and only one linear constraint. We illustrate the association test by using the within family- U statistic. The overall family- U test can be derived in the same manner. Recall that we reject the null hypothesis if the within family- U statistic $U^{\text{Within}} = (U_{AA}^{\text{Within}}, U_{Aa}^{\text{Within}}, U_{aa}^{\text{Within}})$ significantly deviates from 0. The variance–covariance matrix of U^{Within} can be empirically estimated (see File S1). The test statistics can thus be defined to evaluate the distance between the random vector, U^{Within} and 0. We considered two distance measurements, Mahalanobis distance and Euclidean distance.

Mahalanobis distance: Naturally, a test statistic can be defined as the Mahalanobis distance of the within family- U statistic, calculated by any two components (e.g., U_{AA}^{Within} and U_{Aa}^{Within}),

$$\|U^{\text{Within}}\|_M^2 = (U_{AA}^{\text{Within}}, U_{Aa}^{\text{Within}}) \sum_{2 \times 2}^{-1} (U_{AA}^{\text{Within}}, U_{Aa}^{\text{Within}})',$$

where $\sum_{2 \times 2}$ is the variance–covariance matrix for $(U_{AA}^{\text{Within}}, U_{Aa}^{\text{Within}})$. The above test statistic has asymptotically a chi-square distribution with 2 d.f.

Euclidean distance: Alternatively, we can form a test statistic by using the Euclidean distance of the within family- U statistic,

$$\|U^{\text{Within}}\|_E^2 = \sum_g \left[\frac{U_g^{\text{Within}}}{\sqrt{\text{Var}(U_g^{\text{Within}})}} \right]^2, \quad g \in \{AA, Aa, aa\}.$$

In File S1, we show that the above test statistic follows asymptotically a mixture of chi-square distributions. Several approximations and exact methods have been proposed to obtain the distribution of a mixture of chi-square distributions (Davies 1980; Liu *et al.* 2009; Duchesne and Lafaye De Micheaux 2010). In this article, we adopt the Liu method that matches three moments of test statistics (Liu *et al.* 2009).

Results

Simulation studies

Simulation studies were conducted to compare the performance of the family- U statistics with a commonly used FBAT

Figure 1 Pedigree structures used in the simulation. Left, a nuclear family with mother, father, and two offspring; right, a three-generation pedigree with three generations and eight family members.

method, with respect to power, type I error, and robustness against PS. FBAT extends the original TDT method in which alleles transmitted to affected offspring are compared with the expected distribution of alleles among offspring (Laird *et al.* 2000; Lake *et al.* 2000). Since it was developed, it has been widely used in family-based association studies of complex diseases, such as asthma, Alzheimer's diseases, and psychiatric disorders (Demeo *et al.* 2002; Saunders *et al.* 2003; Smoller *et al.* 2003, 2005; Lyon *et al.* 2004). In the simulation, FBAT was conducted by using software `fbat204-beta_win` and its default setting, which assumes an additive genetic effect model and a null hypothesis of “no association and no linkage.”

Simulation without population stratification: The first simulation was conducted to evaluate the performance of two methods in the absence of PS. We first simulated a source population with 1 million unrelated individuals. The genotypes of founders were sampled from the source population with an assumption of Hardy–Weinberg equilibrium. The genotypes of offspring were further generated based on the genotypes of parents, assuming Mendelian transmission. Three types of family structures were considered: nuclear families with four family members (*i.e.*, mother, father, and two offspring), three-generation pedigrees with eight family members, and a mixture of nuclear families and three-generation pedigrees (Figure 1). For each pedigree setting, we simulated 1000 replicates to evaluate power and 100,000 replicates to evaluate type I error rates at various significance levels (*i.e.*, 0.05, 0.01, and 0.001). In each replicate, the total number of subjects was fixed at 1440, which corresponded to 360 nuclear families, 180 three-generation pedigrees, and 120 mixed pedigrees (120 nuclear families plus 120 three-generation pedigrees). For simplicity, we assumed the causal SNP had a minor allele frequency of 0.4 in the source population. We considered both quantitative and binary phenotypes and simulated the phenotypic values by assuming an additive genetic model. For quantitative phenotype, we simulated the phenotype as

$$y_{i,j} = \beta_0 + 0.15 \times I(x_{i,j} = Aa) + 0.3 \times I(x_{i,j} = aa) + \varepsilon_{i,j}, \quad (2)$$

Table 1 Performance of family-*U* and FBAT without population stratification

Pedigree	Power/TIE	Quantitative phenotypes			Binary phenotypes		
		Overall family- <i>U</i>	Within family- <i>U</i>	FBAT	Overall family- <i>U</i>	Within family- <i>U</i>	FBAT
Nuclear family ^a	Power	0.884	0.491	0.426	0.933	0.835	0.737
	TIE – 0.05 ^b	0.050	0.051	0.054	0.047	0.046	0.053
	TIE – 0.01	0.0090	0.0110	0.0111	0.010	0.009	0.0123
	TIE – 0.001	0.0013	0.0015	0.0011	0.0004	0.0004	0.0016
Three-generation ^c	Power	0.908	0.732	0.483	0.921	0.842	0.502
	TIE – 0.05	0.053	0.050	0.056	0.048	0.048	0.049
	TIE – 0.01	0.0115	0.0104	0.0095	0.0085	0.0078	0.0111
	TIE – 0.001	0.0015	0.0010	0.0005	0.0006	0.0006	0.0007
Mixture pedigree ^d	Power	0.883	0.629	0.476	0.934	0.825	0.573
	TIE – 0.05	0.056	0.050	0.055	0.054	0.056	0.047
	TIE – 0.01	0.0091	0.0084	0.0118	0.0112	0.0123	0.0099
	TIE – 0.001	0.0010	0.0008	0.0013	0.0009	0.0012	0.0009

^a Nuclear families with mother, father, and two offspring (*i.e.*, quad).

^b Type I error (TIE) at the significance level of 0.05.

^c Three-generation pedigrees with eight family members.

^d A mixture of nuclear families and three-generation pedigrees.

where $\varepsilon_{i,j} \sim N(0, 1)$, and β_0 is the average phenotype in the population.

For binary phenotype, we simulated the phenotype by using

$$\log \frac{P(y_{i,j} = 1)}{P(y_{i,j} = 0)} = b_0 + \log(1.5) \times I(x_{i,j} = Aa) + 2\log(1.5) \times I(x_{i,j} = aa), \quad (3)$$

where b_0 was adjusted to ensure that the disease prevalence was ~5% in the source population. For binary phenotype, we selected families with at least one affected member from the source population.

Simulations with population stratification: The second simulation was conducted to evaluate the performance of the family-*U* statistics and FBAT in the presence of PS. To simulate genotype data, the allele frequency of the ancestral population (p) was sampled from a uniform distribution, Unif [0.1, 0.4]. We assumed the samples came from two subpopulations. The allele frequencies of subpopulations were generated using the Balding–Nichols model (Balding and Nichols 1995) and were sampled from a beta distribution, $\beta(p((1 - F_{ST})/F_{ST}), (1 - p)((1 - F_{ST})/F_{ST}))$, where F_{ST} measured the severity of PS. The genotypes of founders were sampled from the source population with an assumption of Hardy–Weinberg equilibrium. The genotypes of offspring were then generated based on the genotypes of parents, assuming Mendelian transmission. Assuming an additive model, both quantitative and binary phenotypes were simulated using Equations 2 and 3. The two subpopulations had varying baseline parameters, β_0 and b_0 , representing different average phenotypes or disease prevalence. The values of β_0 and b_0 were adjusted to ensure that the average quantitative phenotype was ~0 and the disease prevalence was ~5% prevalence in the combined population. Similar to

a previous study (Won *et al.* 2009), we increased the levels of PS by varying F_{ST} from 0.1 to 0.2.

Alternative ways of modeling family relationship: For the fast computation of large-scale family data, especially those involving multigeneration pedigrees, we assumed a simple covariance structure (*i.e.*, the covariance is the same for any two members from the same family). Alternatively, there were different ways of modeling family structures. For instance, we could model different relationships with different parameters (*e.g.*, σ_{PC} for parent–offspring pairs, σ_{CC} for sibling pairs, and $\sigma_{PP} = 0$ for unrelated parents). Or, if we assumed the correlation between family members was primarily due to their genetic similarity, we could also model covariance as a function of kinship coefficient. In this simulation, we compared two different approaches (*i.e.*, using the same covariance for family members *vs.* using different covariance parameters for family members) of modeling family structure.

Simulation results

In the simulation, we calculated power as the probability of detecting the causal SNP and type I error as the probability of detecting the noncausal SNP. For simplicity, we evaluated the performance of family-*U* based on the Euclidean distance. As we show in Table S1, the performance of family-*U* by using two distance measurements, the Mahalanobis distance and the Euclidean distance, was very similar.

Simulations without population stratification: The performances of three methods, overall family-*U*, within family-*U*, and FBAT, are summarized in Table 1. The results showed that the type I error rates were properly controlled for all three methods at various significance levels. In terms of power, overall family-*U* attained the highest power for all three pedigree structures (*i.e.*, nuclear, three-generation,

Table 2 Performance of family-*U* and FBAT in presence of population stratification

		$F_{st}^a = 0.1$			$F_{st} = 0.2$		
Pedigree	Power/TIE	Overall family- U	Within family- U	FBAT	Overall family- U	Within family- U	FBAT
Quantitative phenotypes							
Nuclear family ^b	Power	0.890	0.464	0.438	0.884	0.492	0.485
	TIE ^c	0.269	0.054	0.047	0.411	0.057	0.055
Three-generation ^d	Power	0.884	0.703	0.435	0.873	0.715	0.445
	TIE	0.242	0.050	0.046	0.406	0.056	0.055
Mixture pedigree ^e	Power	0.880	0.628	0.444	0.884	0.620	0.454
	TIE	0.232	0.048	0.053	0.398	0.037	0.052
Binary phenotypes							
Nuclear family	Power	0.906	0.818	0.699	0.919	0.840	0.732
	TIE	0.154	0.052	0.051	0.248	0.049	0.049
Three-generation	Power	0.880	0.814	0.488	0.878	0.825	0.475
	TIE	0.196	0.052	0.046	0.325	0.050	0.046
Mixture pedigree	Power	0.927	0.835	0.585	0.916	0.813	0.584
	TIE	0.167	0.059	0.056	0.313	0.056	0.053

^a Wright's F_{st} value measuring the level of population stratification.

^b Nuclear families with mother, father, and two offspring (*i.e.*, quad).

^c A mixture of nuclear families and three-generation pedigrees.

^d Three-generation pedigrees with eight family members.

^e Type I error.

and mixed pedigrees) and two types of phenotype (*i.e.*, quantitative and binary phenotypes). The power of within family-*U* was also consistently higher than that of FBAT in the simulated scenarios.

For quantitative phenotypes, the performance of both overall family-*U* and FBAT was relatively stable across different pedigree structures, while the power of within family-*U* was improved as the size of pedigree increased. It is worthwhile to note that FBAT follows the conventional TDT strategy and uses only information from heterozygous parents, while family-*U* can utilize information from homozygous parents if their genotypes are different (*e.g.*, AA and aa). Therefore, compared to FBAT, within family-*U* considered additional information from homozygous parents and information across generations (*e.g.*, grandparents–grandchildren) and thus gained power increase in three-generation pedigrees. Nevertheless, for nuclear families (*e.g.*, mother, father, and two offspring), such information may be limited (*e.g.*, no cross-generation information), and we would expect limited power improvement of family-*U* over FBAT.

For binary phenotypes, the performance of both overall family-*U* and within family-*U* remained the same for different pedigree structures, while the power of FBAT was reduced for three-generation pedigrees. FBAT counts only those affected offspring with heterozygous parents, while family-*U* utilizes information from all family members, no matter whether the affected member was a founder or an offspring. By using all available information from three-generation and mixture pedigrees, family-*U* attains power improvement over FBAT.

Simulations with population stratification: The performances of three methods, overall family-*U*, within family-*U*, and FBAT, in the presence of PS are summarized in Table 2.

Both within family-*U* and FBAT were robust to PS, and their type I error rates were properly controlled at the 5% level under different levels of PS. However, because the between-family information is confounded by population stratification, the overall family-*U* had inflated type I error. The inflation of type I error also increased with the increasing level of PS (*i.e.*, higher value of F_{ST}). In terms of power, the performance of the three methods was consistent with the results of simulation without PS; *i.e.*, overall family-*U* and within family-*U* attained higher power than FBAT.

Alternative ways of modeling family relationship: The correlations between family members can also be modeled differently. In this simulation, we compared a more complex covariance structure to the simple covariance structure used in the methods. In Table 3, we summarize the performance of family-*U* with two covariance structures. When complex family relationship was specified, the type I error was properly controlled at the 5% level. Moreover, the power of family-*U* based on a complex covariance structure is also comparable to that of family-*U* with a simple covariance structure. The results were also consistent for all pedigree structures (*i.e.*, nuclear, three-generation, and mixture pedigrees) and types of phenotype (*i.e.*, quantitative and binary phenotypes). This result suggested that the performance of family-*U* was robust to the specification of family relationships.

Application to the Framingham Heart Study

The Framingham Heart Study (FHS) is a long-term, ongoing prospective cohort study based on residents of the town of Framingham, Massachusetts. The study was initiated in 1948 with 5209 subjects from Framingham and now has a third generation of participants. A detailed description of the recruitment can be found elsewhere (Splansky *et al.*

Table 3 Performance of family-*U* by modeling complex family relationship

		Simple relationship ^a		Complex relationship ^b	
Pedigree	Power/TIE	Overall family- <i>U</i>	Within family- <i>U</i>	Overall family- <i>U</i>	Within family- <i>U</i>
Quantitative phenotypes					
Nuclear families ^c	Power	0.884	0.491	0.882	0.486
	TIE ^d	0.058	0.057	0.058	0.056
Three-generation ^e	Power	0.908	0.732	0.904	0.719
	TIE	0.043	0.052	0.038	0.045
Mixture pedigree ^f	Power	0.883	0.629	0.880	0.627
	TIE	0.045	0.059	0.045	0.057
Binary phenotypes					
Nuclear families	Power	0.933	0.835	0.933	0.854
	TIE	0.052	0.046	0.050	0.051
Three-generation	Power	0.921	0.842	0.914	0.866
	TIE	0.047	0.041	0.041	0.050
Mixture pedigree	Power	0.934	0.825	0.927	0.856
	TIE	0.048	0.060	0.044	0.083

^a Simple relationship by assuming the same correlation between any two family members.

^b Specification of different correlations between parent–offspring pairs, sibling pairs, and grandparent–grandchild pairs. Unrelated parents have a correlation of 0.

^c Nuclear families with mother, father, and two offspring (*i.e.*, quad).

^d Type I error.

^e Three-generation pedigrees with three generations and eight family members.

^f A mixture of nuclear families and three-generation pedigrees.

2007). In the FHS, a wide variety of phenotypes were collected, including systolic blood pressure, total and HDL cholesterol, fasting glucose, cigarette smoking, and body mass index, etc. In 2007, the FHS entered a new phase, conducting genotyping for 9354 subjects from three generations and ~900 pedigrees. Approximately 550,000 SNPs were genotyped mainly by using Affymetrix Human Mapping 500K and 50K Array sets. The data set is available at the National Center for Biotechnology Information database of genotypes and phenotypes (NCBI dbGaP): http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v18.p7.

In this application, we were interested in studying a nicotine dependence (ND)-related phenotype, measured by the number of cigarettes smoked per day. The phenotype of interest was measured in 5197 subjects from 954 pedigrees. The pedigree structures were highly diverse. Among 954 pedigrees, 268 pedigrees were singletons. The remaining 686 pedigrees had an average of 7.2 genotyped members, including 13 pedigrees with >30 members. As an illustration of our proposed method, we focused on 24 SNPs within a genomic region between 50,000 base pairs upstream and downstream of a candidate gene, *CHRNA5*, which had been repeatedly reported for its association with nicotine dependence and lung cancer (Saccone *et al.* 2010; Chen *et al.* 2012; Hartz *et al.* 2012). Both family-*U* and FBAT were applied to evaluate the association between these 24 SNPs and the quantitative phenotype, the number of cigarettes smoked per day. The results are summarized in Figure 2. By using a nominal threshold of 0.05, 4 SNPs were found to be significantly associated with the phenotype by using overall family-*U*. Although none of the SNPs were significant, 3 SNPs reached marginal significance (*P*-values between 0.05 and 0.1) by using within family-*U*, and 1 SNP

was marginally significant (*P*-value = 0.095) by using FBAT. Overall, the results were consistent with the simulation results. Because the majority of the study population is Caucasian and evidence from previous studies indicated PS was not a major issue in the FHS (Cupples *et al.* 2007; Wilk *et al.* 2009), we would expect the overall family-*U* test has more advantage over the within family-*U* test and FBAT for the association analysis.

Discussion

Since the foundational work by Hoeffding (1948), the *U*-statistic has been adopted in a wide variety of research areas, providing statistical inferences that are distribution-free. Several formations of the *U*-statistic were recently adopted in population-based association studies for detecting genetic variants underlying complex human diseases (Schaid *et al.* 2005; Wei *et al.* 2008; Li *et al.* 2011). These *U*-statistic-based methods have shown great promise, especially when underlying phenotype distributions and modes of inheritance are unknown (Li 2012). Despite its advantages, the *U*-statistic has been rarely used for family-based studies. In this article, we introduced a flexible framework of the *U*-statistic, family-*U*, for family-based association analysis. The proposed family-*U* makes no assumption of pedigree structures, modes of inheritance, and phenotype distributions, which makes it useful for the analysis of various types of data. Family-*U* can be built on within-family information (*i.e.*, within family-*U*), which provides robust protection against PS. In the absence of PS, it could also combine both within-family and between-family information (*i.e.*, overall family-*U*) to enhance the power of association analysis. Through simulations, we have shown that both overall family-*U* and within family-*U* attained higher power than

Figure 2 Manhattan plot for 24 SNPs within 50,000 base pairs upstream or downstream of gene *CHRNA5*. Horizontal dashed line shows nominal significance threshold of 0.05.

FBAT. The power improvement of within family-*U* over FBAT is primarily because it makes use of additional information (e.g., cross-generation information and information from homozygous parents) in each pedigree. Overall family-*U* can further improve the power if there is no PS. Recently, Won *et al.* (2009) proposed a strategy to rescale the family-based test statistics by the ranks of their population-based test *P*-values. Such a strategy will generate uniformly distributed *P*-values under the null hypothesis of no association, which maintains the robustness to PS. In the meantime, statistical power can be improved by combining *P*-values of the family-based test with the *P*-values of the rank-based population-based test. Such a strategy can potentially be adapted into the proposed family-*U* framework to integrate within family-*U* and overall family-*U* statistics into a single test statistic.

In this article, the proposed method was illustrated by using a single SNP. However, it can be easily extended to multiple SNPs with the consideration of interactions. For example, nine genotype combinations can be formed by two SNPs. The corresponding multivariate *U*-statistic would be formed by nine univariate *U*-statistics. We have proposed two distance measurements, Mahalanobis distance and Euclidean distance, in family-*U*. When genetic interactions are considered, the dimensionality of multivariate *U*-statistics will increase rapidly with the number of SNPs (e.g., nine genotypes for two biallelic SNPs). It is also worthwhile to note that the test statistic based on the Euclidean distance

can be estimated without any matrix inversion. The Euclidean distance can be computed more efficiently than Mahalanobis distance and thus is more appropriate for high-dimensional data analysis. When a single SNP is tested, Mahalanobis distance is arguably preferred (Table S1), especially when the pedigree structure is simple and the genetic effect is presumed to be additive.

There are certain limitations of the proposed method. First, within family-*U* is robust against PS under the assumption that all members from same pedigree have a similar ancestry background. For data with mixed ethnicities within each pedigree, the covariance structure should be modeled by the genetic relationship matrix (e.g., the identity-by-state matrix) estimated from genetic data to provide the robustness against PS (Kang *et al.* 2010; Thornton and McPeck 2010). Covariate adjustment is also an important and challenging topic in *U*-statistics. Zhu *et al.* (2012) developed a nonparametric covariate-adjusted association test for both population-based and family-based association studies. Recently, we also proposed a projection method for covariate adjustment in the *U*-statistic (Wei *et al.* 2014). These methods can be potentially used for covariate adjustment in family-*U*.

In the empirical data application, we focused on one candidate gene, *CHRNA5*, which was found to be associated with various complex disorders, such as nicotine dependence (Bierut *et al.* 2008; Chen *et al.* 2012; Hartz *et al.*

2012), lung cancer (Amos *et al.* 2008; Hung *et al.* 2008; Wang *et al.* 2009a), chronic obstructive pulmonary disease (Pillai *et al.* 2009), cocaine addiction (Grucza *et al.* 2008), and alcohol dependence (Wang *et al.* 2009b). In our application, none of the SNPs within the gene region was significant. However, four SNPs, rs1504545, rs12902493, rs16969931, and rs3813571, had nominal *P*-values <0.05, all of which were located upstream of the gene. Previous studies indicated that the expression level of *CHRNA5* appears to be regulated by genetic factors upstream of the gene (Wang *et al.* 2009a,b; Smith *et al.* 2011). Although it is biologically plausible that these SNPs may contribute to smoking behavior, further studies are required to validate this finding.

Acknowledgments

This research was supported, in part, by the National Institutes of Health (NIH) National Institute on Drug Abuse under awards K01DA033346 and K05DA015799, the National Institute of Dental and Craniofacial Research under award R03DE022379, the Translational Research Institute through the NIH National Center for Research Resources and the National Center for Advancing Translational Sciences under awards UL1TR000039 and KL2TR000063, the University of Arkansas for Medical Sciences College of Medicine Children's University Medical Group Fund Grant Program, the Arkansas Children's Hospital Research Institute, and the Arkansas Bioscience Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Literature Cited

- Abecasis, G. R., L. R. Cardon, and W. O. Cookson, 2000 A general test of association for quantitative traits in nuclear families. *Am. J. Hum. Genet.* 66: 279–292.
- Amos, C. I., X. Wu, P. Broderick, I. P. Gorlov, J. Gu *et al.*, 2008 Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nat. Genet.* 40: 616–622.
- Balding, D. J., and R. A. Nichols, 1995 A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity. *Genetica* 96: 3–12.
- Bauchet, M., B. McEvoy, L. N. Pearson, E. E. Quillen, T. Sarkisian *et al.*, 2007 Measuring European population stratification with microarray genotype data. *Am. J. Hum. Genet.* 80: 948–956.
- Bierut, L. J., J. A. Stitzel, J. C. Wang, A. L. Hinrichs, R. A. Grucza *et al.*, 2008 Variants in nicotinic receptors and risk for nicotine dependence. *Am. J. Psychiatry* 165: 1163–1171.
- Cardon, L. R., and L. J. Palmer, 2003 Population stratification and spurious allelic association. *Lancet* 361: 598–604.
- Chen, H. S., X. Zhu, H. Zhao, and S. Zhang, 2003 Qualitative semiparametric test for genetic associations in case-control designs under structured populations. *Ann. Hum. Genet.* 67: 250–264.
- Chen, L. S., N. L. Saccone, R. C. Culverhouse, P. M. Bracci, C. H. Chen *et al.*, 2012 Smoking and genetic risk variation across populations of European, Asian, and African American ancestry—a meta-analysis of chromosome 15q25. *Genet. Epidemiol.* 36: 340–351.
- Cordell, H. J., and D. G. Clayton, 2005 Genetic association studies. *Lancet* 366: 1121–1131.
- Cordell, H. J., B. J. Barratt, and D. G. Clayton, 2004 Case/pseudocontrol analysis in genetic association studies: a unified framework for detection of genotype and haplotype associations, gene-gene and gene-environment interactions, and parent-of-origin effects. *Genet. Epidemiol.* 26: 167–185.
- Cupples, L. A., H. T. Arruda, E. J. Benjamin, R. B. D'Agostino, Sr, S. Demissie *et al.*, 2007 The Framingham Heart Study 100K SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med. Genet.* 8(Suppl. 1): S1.
- Davies, R., 1980 The distribution of a linear combination of chi-square random variables. *Appl. Stat.* 29: 323–333.
- DeMeo, D. L., C. Lange, E. K. Silverman, J. M. Senter, J. M. Drazen *et al.*, 2002 Univariate and multivariate family-based association analysis of the IL-13 ARG130GLN polymorphism in the Childhood Asthma Management Program. *Genet. Epidemiol.* 23: 335–348.
- Devlin, B., and K. Roeder, 1999 Genomic control for association studies. *Biometrics* 55: 997–1004.
- Duchesne, P., and P. Lafaye De Micheaux, 2010 Computing the distribution of quadratic forms: further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Comput. Stat. Data Anal.* 54: 858–862.
- Freedman, M. L., D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault *et al.*, 2004 Assessing the impact of population stratification on genetic association studies. *Nat. Genet.* 36: 388–393.
- Fulker, D. W., S. S. Cherny, P. C. Sham, and J. K. Hewitt, 1999 Combined linkage and association sib-pair analysis for quantitative traits. *Am. J. Hum. Genet.* 64: 259–267.
- Gordon, D., S. C. Heath, X. Liu, and J. Ott, 2001 A transmission/disequilibrium test that allows for genotyping errors in the analysis of single-nucleotide polymorphism data. *Am. J. Hum. Genet.* 69: 371–380.
- Gordon, D., C. Haynes, C. Johnnidis, S. B. Patel, A. M. Bowcock *et al.*, 2004 A transmission disequilibrium test for general pedigrees that is robust to the presence of random genotyping errors and any number of untyped parents. *Eur. J. Hum. Genet.* 12: 752–761.
- Grucza, R. A., J. C. Wang, J. A. Stitzel, A. L. Hinrichs, S. F. Saccone *et al.*, 2008 A risk allele for nicotine dependence in *CHRNA5* is a protective allele for cocaine dependence. *Biol. Psychiatry* 64: 922–929.
- Hartz, S. M., S. E. Short, N. L. Saccone, R. Culverhouse, L. Chen *et al.*, 2012 Increased genetic vulnerability to smoking at *CHRNA5* in early-onset smokers. *Arch. Gen. Psychiatry* 69: 854–860.
- Hoeffding, W., 1948 A class of statistics with asymptotically normal distribution. *Ann. Math. Stat.* 19: 293–325.
- Hung, R. J., J. D. McKay, V. Gaborieau, P. Boffetta, M. Hashibe *et al.*, 2008 A susceptibility locus for lung cancer maps to nicotinic acetylcholine receptor subunit genes on 15q25. *Nature* 452: 633–637.
- Ionita-Laza, I., M. B. McQueen, N. M. Laird, and C. Lange, 2007 Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100K scan. *Am. J. Hum. Genet.* 81: 607–614.
- Kang, H. M., J. H. Sul, S. K. Service, N. A. Zaitlen, S. Y. Kong *et al.*, 2010 Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.* 42: 348–354.
- Knowler, W. C., R. C. Williams, D. J. Pettitt, and A. G. Steinberg, 1988 Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am. J. Hum. Genet.* 43: 520–526.
- Laird, N. M., and C. Lange, 2006 Family-based designs in the age of large-scale gene-association studies. *Nat. Rev. Genet.* 7: 385–394.
- Laird, N. M., S. Horvath, and X. Xu, 2000 Implementing a unified approach to family-based tests of association. *Genet. Epidemiol.* 19(Suppl. 1): S36–S42.

- Lake, S. L., D. Blacker, and N. M. Laird, 2000 Family-based tests of association in the presence of linkage. *Am. J. Hum. Genet.* 67: 1515–1525.
- Lange, C., E. K. Silverman, X. Xu, S. T. Weiss, and N. M. Laird, 2003 A multivariate family-based association test using generalized estimating equations: FBAT-GEE. *Biostatistics* 4: 195–206.
- Lazzeroni, L. C., and K. Lange, 1998 A conditional inference framework for extending the transmission/disequilibrium test. *Hum. Hered.* 48: 67–81.
- Li, H., 2012 U-statistics in genetic association studies. *Hum. Genet.* 131: 1395–1401.
- Li, M., C. Ye, W. Fu, R. C. Elston, and Q. Lu, 2011 Detecting genetic interactions for quantitative traits with U-statistics. *Genet. Epidemiol.* 35: 457–468.
- Liu, H., Y. Tang, and H. H. Zhang, 2009 A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput. Stat. Data Anal.* 53: 853–856.
- Lyon, H., C. Lange, S. Lake, E. K. Silverman, A. G. Randolph *et al.*, 2004 IL10 gene polymorphisms are associated with asthma phenotypes in children. *Genet. Epidemiol.* 26: 155–165.
- Marchini, J., L. R. Cardon, M. S. Phillips, and P. Donnelly, 2004 The effects of human population structure on large genetic association studies. *Nat. Genet.* 36: 512–517.
- Martin, E. R., S. A. Monks, L. L. Warren, and N. L. Kaplan, 2000 A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am. J. Hum. Genet.* 67: 146–154.
- Murphy, A., S. T. Weiss, and C. Lange, 2008 Screening and replication using the same data set: testing strategies for family-based studies in which all probands are affected. *PLoS Genet.* 4: e1000197.
- Pillai, S. G., D. Ge, G. Zhu, X. Kong, K. V. Shianna *et al.*, 2009 A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* 5: e1000421.
- Price, A. L., N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick *et al.*, 2006 Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* 38: 904–909.
- Pritchard, J. K., and N. A. Rosenberg, 1999 Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 65: 220–228.
- Pritchard, J. K., M. Stephens, and P. Donnelly, 2000a Inference of population structure using multilocus genotype data. *Genetics* 155: 945–959.
- Pritchard, J. K., M. Stephens, N. A. Rosenberg, and P. Donnelly, 2000b Association mapping in structured populations. *Am. J. Hum. Genet.* 67: 170–181.
- Qin, H., and X. Zhu, 2012 Allowing for population stratification in association analysis. *Methods Mol. Biol.* 850: 399–409.
- Rabinowitz, D., and N. Laird, 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.
- Saccone, N. L., R. C. Culverhouse, T. H. Schwantes-An, D. S. Cannon, X. Chen *et al.*, 2010 Multiple independent loci at chromosome 15q25.1 affect smoking quantity: a meta-analysis and comparison with lung cancer and COPD. *PLoS Genet.* 6: e1001053.
- Satten, G. A., W. D. Flanders, and Q. Yang, 2001 Accounting for unmeasured population substructure in case-control studies of genetic association using a novel latent-class model. *Am. J. Hum. Genet.* 68: 466–477.
- Saunders, A. J., L. Bertram, K. Mullin, A. J. Sampson, K. Latifzai *et al.*, 2003 Genetic association of Alzheimer's disease with multiple polymorphisms in alpha-2-macroglobulin. *Hum. Mol. Genet.* 12: 2765–2776.
- Schaid, D. J., S. K. McDonnell, S. J. Hebring, J. M. Cunningham, and S. N. Thibodeau, 2005 Nonparametric tests of association of multiple genes with human disease. *Am. J. Hum. Genet.* 76: 780–793.
- Sinsheimer, J. S., C. G. Palmer, and J. A. Woodward, 2003 Detecting genotype combinations that increase risk for disease: maternal-fetal genotype incompatibility test. *Genet. Epidemiol.* 24: 1–13.
- Smith, R. M., H. Alachkar, A. C. Papp, D. Wang, D. C. Mash *et al.*, 2011 Nicotinic alpha5 receptor subunit mRNA expression is associated with distant 5' upstream polymorphisms. *Eur. J. Hum. Genet.* 19: 76–83.
- Smoller, J. W., J. F. Rosenbaum, J. Biederman, J. Kennedy, D. Dai *et al.*, 2003 Association of a genetic marker at the corticotropin-releasing hormone locus with behavioral inhibition. *Biol. Psychiatry* 54: 1376–1381.
- Smoller, J. W., L. H. Yamaki, J. A. Fagerness, J. Biederman, S. Racette *et al.*, 2005 The corticotropin-releasing hormone gene and behavioral inhibition in children at risk for panic disorder. *Biol. Psychiatry* 57: 1485–1492.
- Spielman, R. S., and W. J. Ewens, 1998 A sibship test for linkage in the presence of association: the sib transmission/disequilibrium test. *Am. J. Hum. Genet.* 62: 450–458.
- Spielman, R. S., R. E. McGinnis, and W. J. Ewens, 1993 Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am. J. Hum. Genet.* 52: 506–516.
- Splansky, G. L., D. Corey, Q. Yang, L. D. Atwood, L. A. Cupples *et al.*, 2007 The Third Generation Cohort of the National Heart, Lung, and Blood Institute's Framingham Heart Study: design, recruitment, and initial examination. *Am. J. Epidemiol.* 165: 1328–1335.
- Thornton, T., and M. S. McPeck, 2010 ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *Am. J. Hum. Genet.* 86: 172–184.
- Van Steen, K., M. B. McQueen, A. Herbert, B. Raby, H. Lyon *et al.*, 2005 Genomic screening and replication using the same data set in family-based association testing. *Nat. Genet.* 37: 683–691.
- Wang, J. C., C. Cruchaga, N. L. Saccone, S. Bertelsen, P. Liu *et al.*, 2009a Risk for nicotine dependence and lung cancer is conferred by mRNA expression levels and amino acid change in CHRNA5. *Hum. Mol. Genet.* 18: 3125–3135.
- Wang, J. C., R. Grucza, C. Cruchaga, A. L. Hinrichs, S. Bertelsen *et al.*, 2009b Genetic variation in the CHRNA5 gene affects mRNA levels and is associated with risk for alcohol dependence. *Mol. Psychiatry* 14: 501–510.
- Wei, C., M. Li, Z. He, O. Vsevolozhskaya, D. J. Schaid *et al.*, 2014 A Weighted U-statistic for genetic association analyses of sequencing data. *Genet. Epidemiol.* 38: 699–708.
- Wei, Z., M. Li, T. Rebbeck, and H. Li, 2008 U-statistics-based tests for multiple genes in genetic association studies. *Ann. Hum. Genet.* 72: 821–833.
- Weinberg, C. R., 2003 Studying parents and grandparents to assess genetic contributions to early-onset disease. *Am. J. Hum. Genet.* 72: 438–447.
- Weinberg, C. R., A. J. Wilcox, and R. T. Lie, 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.
- Wilk, J. B., T. H. Chen, D. J. Gottlieb, R. E. Walter, M. W. Nagle *et al.*, 2009 A genome-wide association study of pulmonary function measures in the Framingham Heart Study. *PLoS Genet.* 5: e1000429.
- Won, S., J. B. Wilk, R. A. Mathias, C. J. O'Donnell, E. K. Silverman *et al.*, 2009 On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. *PLoS Genet.* 5: e1000741.
- Zhu, W., Y. Jiang, and H. Zhang, 2012 Nonparametric covariate-adjusted association tests based on the generalized Kendall's tau. *J. Am. Stat. Assoc.* 107: 1–11.

Communicating editor: N. Yi

GENETICS

Supporting Information

<http://www.genetics.org/lookup/suppl/doi:10.1534/genetics.115.175174/-/DC1>

A Powerful Nonparametric Statistical Framework for Family-Based Association Analyses

Ming Li, Zihuai He, Daniel J. Schaid, Mario A. Cleves, Todd G. Nick, and Qing Lu

Supplementary Materials

In this section, we give the theoretical details for the proposed test statistics. For simplicity, we assume a quantitative phenotype and 3 genotype groups from a single single-nucleotide polymorphism (SNP). However, it should be noted that our method is also applied to binary phenotypes and more than 3 genotype groups formed by multiple SNPs simultaneously (e.g. 9 genotypes for 2 SNPs). The extension to binary phenotypes and multi-SNP test follows straightforwardly.

1. Asymptotic Distribution of Family-U Test Statistics

In the section 4 of this Supplementary Materials, we show both within family-U and overall family-U follow asymptotically a degenerate multivariate normal distribution with one and only one linear constraint. Therefore, it is easy to show that the Mahalanobis distance, $\|U\|_M^2 = (U_{AA}, U_{Aa}) \Sigma^{-1} (U_{AA}, U_{Aa})'$, follows asymptotically a Chi-square distribution with 2 degrees of freedom.

To study the distribution of Euclidean distance, $\|U\|_E^2 = \sum_g \left[\frac{U_g}{\sqrt{\text{Var}(U_g)}} \right]^2$,

$g \in \{AA, Aa, aa\}$, it should be noted that the variance-covariance matrix of U is a positive semi-definite matrix. Denote $\Sigma_0 = \text{diag}(\Sigma)^{-1/2} \cdot \Sigma \cdot \text{diag}(\Sigma)^{-1/2}$, and there exists an orthogonal matrix P such that $P' \Sigma_0 P = \Lambda = \text{diag}(\lambda_1, \lambda_2, \lambda_3)$ and $\lambda_1, \lambda_2, \lambda_3 \geq 0$.

Let $Z = (Z_1, Z_2, Z_3) = U \cdot \text{diag}(\Sigma)^{-1/2} \cdot P$, then

$$\text{Var}(Z') = P' \cdot \text{diag}(\Sigma)^{-1/2} \cdot \Sigma \cdot \text{diag}(\Sigma)^{-1/2} \cdot P = P' \cdot \Sigma_0 \cdot P = \text{diag}(\lambda_1, \lambda_2, \lambda_3).$$

It follows that (Z_1, Z_2, Z_3) are all independent with $Z_k \sim N(0, \lambda_k); 1 \leq k \leq 3$, where λ_k are the eigenvalues of Σ_0 . In our study, because the rank of Σ_0 is 2, one of the eigenvalues will be 0.

Therefore, the Euclidean distance of U , $\|U\|_E^2$, follows a mixture of Chi-square distributions,

$$\|U\|_E^2 = \sum_g \left[\frac{U_g}{\sqrt{\text{Var}(U_g)}} \right]^2 = [U \cdot \text{diag}(\Sigma)^{-1/2} \cdot P] \cdot [U \cdot \text{diag}(\Sigma)^{-1/2} \cdot P]' = \sum_{k=1}^3 Z_k^2.$$

For any $\lambda_k > 0$, $Z_k^2 = \lambda_k (Z_k / \sqrt{\lambda_k})^2$, where $(Z_k / \sqrt{\lambda_k})^2 \sim \chi^2(1)$. For $\lambda_k = 0$, $Z_k = 0$.

2. Empirical Estimation of the Variance-Covariance Matrix of Family U-statistic

In this section, we provide the estimation of covariance matrix for within family-U. The estimation of overall family-U follows similarly.

We use the same notation as those used in the main text. Suppose we have a study population with N subjects collected from M pedigrees. The i -th pedigree comprises of n_i subjects, where $\sum_{i=1}^M n_i = N$. We denote $y_{i,j}$ as the phenotype of the j -th subject from the i -th pedigree; $x_{i,j}$ as its genotype of a particular SNP, where $x_{i,j} \in \{AA, Aa, aa\}$; $1 \leq i \leq M$, and $1 \leq j \leq n_i$. Let $Y_i = (y_{i,1}, y_{i,2}, \dots, y_{i,n_i})'$ and $X_i = (x_{i,1}, x_{i,2}, \dots, x_{i,n_i})'$ be the phenotypes and genotypes for the family members in the i -th pedigree, $1 \leq i \leq M$. For any pedigree i , we first

define $S_{i,AA} = \{j | x_{i,j} = AA\}$ as the members of pedigree i who carry genotype AA; and

$S_{i,AA}^c = \{j | x_{i,j} \neq AA\}$ as those carrying genotype Aa/aa. Denote $|S_{i,AA}| = n_{i,AA}$ as the number of

family members with genotype AA in pedigree i , and $|S_{i,AA}^c| = n_i - n_{i,AA}$ as the number of

family members with genotype Aa/aa in pedigree i . Further, denote a set of pedigree index as

$D_{AA} = \{i \mid \exists 1 \leq j_1, j_2 \leq n_i, \text{st } x_{i,j_1} = AA, x_{i,j_2} \neq AA\}$, which includes all pedigrees with members carrying genotype AA and genotype Aa/aa.

For quantitative traits with kernel function: $\varphi(y_{i,j_1}, y_{i,j_2}) = y_{i,j_1} - y_{i,j_2}$, we define a multivariate U-Statistic $U^{Within} = (U_{AA}^{Within}, U_{Aa}^{Within}, U_{aa}^{Within})$ as:

$$\begin{aligned} U_{AA}^{Within} &= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} \frac{1}{n_{i,AA}(n_i - n_{i,AA})} \sum_{j_1 \in S_{i,AA}} \sum_{j_2 \in S_{i,AA}^c} \varphi(y_{i,j_1}, y_{i,j_2}) \\ &= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} \left[\sum_{j_1 \in S_{i,AA}} \frac{1}{n_{i,AA}} y_{i,j_1} - \sum_{j_2 \in S_{i,AA}^c} \frac{1}{n_i - n_{i,AA}} y_{i,j_2} \right] \\ &= \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} \frac{I(x_{i,j} = AA)}{n_{i,AA}} - \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} \frac{I(x_{i,j} \neq AA)}{n_i - n_{i,AA}} \\ &= \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} \left[\frac{I(x_{i,j} = AA)}{n_{i,AA}} - \frac{I(x_{i,j} \neq AA)}{n_i - n_{i,AA}} \right] \\ &= \sum_{i=1}^M \gamma'_i(AA) \cdot Y_i \end{aligned}$$

where $c_{i,j}(AA) = \frac{1}{|D_{AA}|} \left[\frac{I(x_{i,j} = AA)}{n_{i,AA}} - \frac{I(x_{i,j} \neq AA)}{n_i - n_{i,AA}} \right]$ and

$$\gamma_i(AA) = \frac{1}{|D_{AA}|} \left[\frac{I(X_{i,j} = AA)}{n_{i,AA}} - \frac{I(X_{i,j} \neq AA)}{n_i - n_{i,AA}} \right].$$

Similarly, we have $U_{Aa}^{Within} = \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} c_{i,j}(Aa) = \sum_{i=1}^M \gamma'_i(Aa) \cdot Y_i$;

$$\text{and } U_{aa}^{Within} = \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} c_{i,j}(aa) = \sum_{i=1}^M \gamma'_i(aa) \cdot Y_i,$$

It is easy to show that each univariate U-statistic is a linear combination of phenotypes from all individuals. Therefore, $Cov(U, U)$ can be obtained by an empirical estimation of $Cov(\vec{Y}, \vec{Y})$.

Here, $\vec{Y}$ is a $N \times 1$ vector with all individuals in sequential order from pedigree 1 to M , denoted by $\vec{Y} = (Y_1', \dots, Y_M')' = (y_{1,1}, \dots, y_{1,n_1}, \dots, y_{M,1}, \dots, y_{M,n_M})'$. Similarly, let

$$\vec{\gamma}(AA) = (\gamma_1'(AA), \dots, \gamma_M'(AA)) = (c_{1,1}(AA), \dots, c_{1,n_1}(AA), \dots, c_{M,1}(AA), \dots, c_{M,n_M}(AA))'$$

$$\vec{\gamma}(Aa) = (\gamma_1'(Aa), \dots, \gamma_M'(Aa)) = (c_{1,1}(Aa), \dots, c_{1,n_1}(Aa), \dots, c_{M,1}(Aa), \dots, c_{M,n_M}(Aa))'$$

$$\vec{\gamma}(aa) = (\gamma_1'(aa), \dots, \gamma_M'(aa)) = (c_{1,1}(aa), \dots, c_{1,n_1}(aa), \dots, c_{M,1}(aa), \dots, c_{M,n_M}(aa))'.$$

$$\text{Then, } Cov(U, U) = \begin{pmatrix} \vec{\gamma}'(AA) \\ \vec{\gamma}'(Aa) \\ \vec{\gamma}'(aa) \end{pmatrix}_{3 \times N} Cov(\vec{Y}, \vec{Y})_{N \times N} \begin{pmatrix} \vec{\gamma}(AA) & \vec{\gamma}(Aa) & \vec{\gamma}(aa) \end{pmatrix}_{N \times 3}.$$

To estimate $Cov(\vec{Y}, \vec{Y})$, we have

$$\hat{\sigma}^2 = \text{var}(y_{i,j}) = \frac{1}{N} \sum_{i=1}^M \sum_{j=1}^{n_i} [y_{i,j} - E(y_{i,j})]^2,$$

where $E(y_{i,j})$ can be empirically estimated by the phenotypic mean among individuals carrying

$$\text{each genotype: } E(y_{i,j}) = \begin{cases} \bar{y}_{AA}, & \text{if } x_{i,j} = AA \\ \bar{y}_{Aa}, & \text{if } x_{i,j} = Aa \\ \bar{y}_{aa}, & \text{if } x_{i,j} = aa \end{cases}.$$

Here, $\bar{y}_{AA}$, $\bar{y}_{Aa}$ and $\bar{y}_{aa}$ represent the sample mean among individuals carrying genotype AA, Aa, and aa, respectively. Furthermore, the covariance among samples from the same family can be estimated as:

$$\hat{\sigma}_1 = Cov(y_{i,j_1}, y_{i,j_2}) = \frac{1}{N_{pair}} \sum_{i=1}^M \sum_{1 \leq j_1 < j_2 \leq n_i} [y_{i,j_1} - E(y_{i,j_1})][y_{i,j_2} - E(y_{i,j_2})];$$

where N_{pair} is the total number of sample pairs within each family, and $N_{pair} = \sum_{i=1}^M \frac{n_i(n_i-1)}{2}$. It

is easy to show that $Cov(\vec{Y}, \vec{Y})$ is a block diagonal matrix with a form of $diag(\Sigma_1, \Sigma_2, \dots, \Sigma_M)$,

where Σ_i is a $n_i \times n_i$ matrix with $\hat{\sigma}^2$ on the diagonal and $\hat{\sigma}_1$ as all other elements.

3. Performance of Test Statistics: Mahalanobis distance Vs. Euclidean distance

Additional simulation was conducted to compare the performance of family-U by using two distance measurements. In the simulation, two genetics models, an additive genetic effect model and a dominant genetic effect model, were studied.

The simulation results are summarized in Table S1. Overall, the results indicated that two distance measurements had similar performance. The type I error rates were properly controlled for both distance measurements. Mahalanobis distance showed slightly advantage under an additive genetic effect model, particularly for within family-U with nuclear families. On the other hand, Euclidean distance performed better under a dominant effect model, except for within family-U with nuclear families.

4. Performance of Family-U and FBAT when Founders Come from Different Ethnic Groups

We conducted an additional simulation to provide insights on the performance of family-U when population admixture exists within pedigrees. For this limited simulation, we simulated 360 nuclear families with mother, father, and two offspring (totally 1,440 individuals). Among these families, we respectively selected 10%, 30% and 50% families and sampled the parents from two sub-populations with different levels of population stratification (i.e. $F_{st}=0.1$ and $F_{st}=0.2$). For the selected families, the phenotypes of parents were simulated according to their corresponding subpopulations, while the phenotypes of offspring were simulated by a summation of allelic effects that varied in their parental subpopulations. In addition to FBAT and within family-U, we also considered a modified within family-U test, within family-U-M test, which excludes the pair comparison between founders.

The simulation results are summarized in Table S2. In the presence of PS, the type I error rates of FBAT were well controlled. For within-family U, the type I error rates were inflated, especially for binary phenotypes. By excluding the founder pairs, the type I error rates of within family-U-M were reduced, but were still inflated, especially when the admixture proportion is high (i.e. $F_{st}=0.2$)

5. Family-U Statistic and its Linear Constraint

5.1 Within family-U Statistic

Without losing generality, we assume that every pedigree includes members with all 3 possible genotypes. The pedigrees that do not have members with certain genotype will not contribute to the within family-U statistic, and therefore can be ignored.

For any pedigree i , we first define $S_{i,AA} = \{j | x_{i,j} = AA\}$ as the members of pedigree i who carry genotype AA; and $S_{i,AA}^c = \{j | x_{i,j} \neq AA\}$ as those carrying genotype Aa/aa. Denote $|S_{i,AA}| = n_{i,AA}$ as the number of family members with genotype AA in pedigree i , and $|S_{i,AA}^c| = n_i - n_{i,AA}$ as the number of family members with genotype Aa/aa in pedigree i . Further, denote a set of pedigree index as $D_{AA} = \{i | \exists 1 \leq j_1, j_2 \leq n_i, st \ x_{i,j_1} = AA, x_{i,j_2} \neq AA\}$. Here, D_{AA} includes all pedigrees with both members carrying AA and members carrying Aa/aa. For multivariate U-Statistic, $U^{Within} = (U_{AA}^{Within}, U_{Aa}^{Within}, U_{aa}^{Within})$, we then have:

$$\begin{aligned}
U_{AA}^{Within} &= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} \frac{1}{n_{i,AA}(n_i - n_{i,AA})} \sum_{j_1 \in S_{i,AA}} \sum_{j_2 \in S_{i,AA}^c} \varphi(\mathcal{Y}_{i,j_1}, \mathcal{Y}_{i,j_2}) \\
&= \frac{1}{|D_{AA}|} \sum_{i \in D_{AA}} \left[\sum_{j_1 \in S_{i,AA}} \frac{1}{n_{i,AA}} \mathcal{Y}_{i,j_1} - \sum_{j_2 \in S_{i,AA}^c} \frac{1}{n_i - n_{i,AA}} \mathcal{Y}_{i,j_2} \right] \\
&= \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} \mathcal{Y}_{i,j} \frac{I(x_{i,j} = AA)}{n_{i,AA}} - \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} \mathcal{Y}_{i,j} \frac{I(x_{i,j} \neq AA)}{n_i - n_{i,AA}} \\
&= \frac{1}{|D_{AA}|} \sum_{i=1}^M \sum_{j=1}^{n_i} \mathcal{Y}_{i,j} \left[\frac{I(x_{i,j} = AA)}{n_{i,AA}} - \frac{I(x_{i,j} \neq AA)}{n_i - n_{i,AA}} \right] \\
&= \sum_{i=1}^M \gamma'_i(AA) \cdot Y_i
\end{aligned}$$

where $\gamma_i(AA) = \frac{I(X_i = AA)}{|D_{AA}|(n_{i,AA})} - \frac{I(X_i \neq AA)}{|D_{AA}|(n_i - n_{i,AA})}$.

Similarly, we have $U_{Aa}^{Within} = \sum_{i=1}^M \gamma'_i(Aa) \cdot Y_i$ and $U_{aa}^{Within} = \sum_{i=1}^M \gamma'_i(aa) \cdot Y_i$,

where $\gamma_i(Aa) = \frac{I(X_i = Aa)}{|D_{Aa}|(n_{i,Aa})} - \frac{I(X_i \neq Aa)}{|D_{Aa}|(n_i - n_{i,Aa})}$, and $\gamma_i(aa) = \frac{I(X_i = aa)}{|D_{aa}|(n_{i,aa})} - \frac{I(X_i \neq aa)}{|D_{aa}|(n_i - n_{i,aa})}$.

5.2 Linear Constraint for Within family-U Statistic

For any given constant $c = (c_{AA}, c_{Aa}, c_{aa})$, the linear combination of univariate U-Statistics

$$c_{AA}U_{AA}^{Within} + c_{Aa}U_{Aa}^{Within} + c_{aa}U_{aa}^{Within} = \sum_{i=1}^M [c_{AA}\gamma'_i(AA) + c_{Aa}\gamma'_i(Aa) + c_{aa}\gamma'_i(aa)] \cdot Y_i = \sum_{i=1}^M \eta'_i \cdot Y_i,$$

where

$$\begin{aligned} \eta'_i &= c_{AA}\gamma'_i(AA) + c_{Aa}\gamma'_i(Aa) + c_{aa}\gamma'_i(aa) \\ &= \left(\frac{c_{AA}}{n_{i,AA}} - \frac{c_{Aa}}{n_i - n_{i,Aa}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right) \frac{I(X_i = AA)}{|D_{AA}|} \\ &\quad + \left(-\frac{c_{AA}}{n_i - n_{i,AA}} + \frac{c_{Aa}}{n_{i,Aa}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right) \frac{I(X_i = Aa)}{|D_{Aa}|} \\ &\quad + \left(-\frac{c_{AA}}{n_i - n_{i,AA}} - \frac{c_{Aa}}{n_i - n_{i,Aa}} + \frac{c_{aa}}{n_{i,aa}} \right) \frac{I(X_i = aa)}{|D_{aa}|} \end{aligned}$$

When $c = c_1 = (\Pi_{i=1}^M n_{i,AA}(n - n_{i,AA}), \Pi_{i=1}^M n_{i,Aa}(n - n_{i,Aa}), \Pi_{i=1}^M n_{i,aa}(n - n_{i,aa}))$, for any pedigree i_0 ,

we have

$$\begin{aligned} \frac{c_{AA}}{n_{i_0,AA}} - \frac{c_{Aa}}{n_{i_0} - n_{i_0,Aa}} - \frac{c_{aa}}{n_{i_0} - n_{i_0,aa}} &= [n_{i_0} - n_{i_0,AA} - n_{i_0,Aa} - n_{i_0,aa}] \Pi_{i \neq i_0} n_{i,AA}(n_i - n_{i,AA}) = 0; \\ -\frac{c_{AA}}{n_{i_0} - n_{i_0,AA}} + \frac{c_{Aa}}{n_{i_0,Aa}} - \frac{c_{aa}}{n_{i_0} - n_{i_0,aa}} &= [-n_{i_0,AA} + n_{i_0} - n_{i_0,Aa} - n_{i_0,aa}] \Pi_{i \neq i_0} n_{i,AA}(n_i - n_{i,AA}) = 0; \\ -\frac{c_{AA}}{n_{i_0} - n_{i_0,AA}} - \frac{c_{Aa}}{n_{i_0} - n_{i_0,Aa}} + \frac{c_{aa}}{n_{i_0,aa}} &= [-n_{i_0,AA} - n_{i_0,Aa} + n_{i_0} - n_{i_0,aa}] \Pi_{i \neq i_0} n_{i,AA}(n_i - n_{i,AA}) = 0. \end{aligned}$$

Therefore $U^{Within} \cdot c'_1 = 0$.

5.3 Asymptotic Distribution of within family-U Statistic

We now give the proof that the within family-U Statistic follows asymptotically a degenerate multivariate normal distribution. We show here, for any $c \neq c_1$, the linear combination of U-Statistics, $c_{AA}U_{AA}^{Within} + c_{Aa}U_{Aa}^{Within} + c_{aa}U_{aa}^{Within}$, follows asymptotically a univariate normal distribution, which is equivalent to show:

$$\lim_{M \rightarrow \infty} \left(\frac{\max_{1 \leq i \leq M} \|\eta'_i\|^2}{\sum_{i=1}^M \|\eta'_i\|^2} \right) = 0.$$

Under mild conditions, the above equation is a corollary of Lindeberg's CLT. The theoretical detail can be found elsewhere (Shao, 2003).

Consider

$$\begin{aligned} \|\eta'_i\|^2 = & \left(\frac{c_{AA}}{n_{i,AA}} - \frac{c_{Aa}}{n_i - n_{i,Aa}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right)^2 \frac{n_{i,AA}}{|D_{AA}|} \\ & + \left(-\frac{c_{AA}}{n_i - n_{i,AA}} + \frac{c_{Aa}}{n_{i,Aa}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right)^2 \frac{n_{i,Aa}}{|D_{Aa}|} \\ & + \left(-\frac{c_{AA}}{n_i - n_{i,AA}} - \frac{c_{Aa}}{n_i - n_{i,Aa}} + \frac{c_{aa}}{n_{i,aa}} \right)^2 \frac{n_{i,aa}}{|D_{aa}|} \end{aligned}$$

Given that $1 < n_{i,AA}, n_{i,Aa}, n_{i,aa} < n$, we define

$$H_i = \begin{pmatrix} \frac{1}{n_{i,\Lambda\Lambda}} & -\frac{1}{n_i - n_{i,\Lambda a}} & -\frac{1}{n_i - n_{i,aa}} \\ -\frac{1}{n_i - n_{i,\Lambda\Lambda}} & \frac{1}{n_{i,\Lambda a}} & -\frac{1}{n_i - n_{i,aa}} \\ -\frac{1}{n_i - n_{i,\Lambda\Lambda}} & -\frac{1}{n_i - n_{i,\Lambda a}} & \frac{1}{n_{i,aa}} \end{pmatrix}; \quad 1 \leq i \leq M.$$

Then, for $H = (H_1, H_2, \dots, H_M)'$, $\text{rank}(H) = 2$. Therefore, for any $c \neq kc_1$, $H_1 \cdot c' \neq 0$.

$$\begin{aligned} \max_{1 \leq i \leq M} \|\eta'_i\|^2 &\leq (|c_{\Lambda\Lambda}| + |c_{\Lambda a}| + |c_{aa}|)^2 \frac{n}{|D_{\Lambda\Lambda}|^2} \\ &\quad + (|c_{\Lambda\Lambda}| + |c_{\Lambda a}| + |c_{aa}|)^2 \frac{n}{|D_{\Lambda a}|^2} \\ &\quad + (|c_{\Lambda\Lambda}| + |c_{\Lambda a}| + |c_{aa}|)^2 \frac{n}{|D_{aa}|^2} \end{aligned} \xrightarrow{as \ M \rightarrow \infty} \sim \frac{n}{|D_{\Lambda\Lambda}|^2} + \frac{n}{|D_{\Lambda a}|^2} + \frac{n}{|D_{aa}|^2}.$$

On the other hand,

$$\begin{aligned} \sum_{i=1}^M \|\eta'_i\|^2 &\geq \left(\frac{c_{\Lambda\Lambda}}{n} - \frac{c_{\Lambda a}}{n_i - n_{i,\Lambda a}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right)^2 \frac{Nf_{\Lambda\Lambda}}{|D_{\Lambda\Lambda}|} \\ &\quad + \left(-\frac{c_{\Lambda\Lambda}}{n_i - n_{i,\Lambda\Lambda}} + \frac{c_{\Lambda a}}{n_{i,\Lambda a}} - \frac{c_{aa}}{n_i - n_{i,aa}} \right)^2 \frac{Nf_{\Lambda a}}{|D_{\Lambda a}|} \\ &\quad + \left(-\frac{c_{\Lambda\Lambda}}{n_i - n_{i,\Lambda\Lambda}} - \frac{c_{\Lambda a}}{n_i - n_{i,\Lambda a}} + \frac{c_{aa}}{n_{i,aa}} \right)^2 \frac{Nf_{aa}}{|D_{aa}|} \end{aligned} \xrightarrow{as \ M, N \rightarrow \infty} \sim \frac{Nf_{\Lambda\Lambda}}{|D_{\Lambda\Lambda}|^2} + \frac{Nf_{\Lambda a}}{|D_{\Lambda a}|^2} + \frac{Nf_{aa}}{|D_{aa}|^2}.$$

It follows that

$$\lim_{M \rightarrow \infty} \left(\frac{\max_{1 \leq i \leq M} \|\eta'_i\|^2}{\sum_{i=1}^M \|\eta'_i\|^2} \right) \leq \lim_{N \rightarrow \infty} \left(\frac{n}{N} \right) = 0.$$

This shows that the within family-U Statistic also follows asymptotically a degenerate multivariate normal distribution with one and only one linear constraint.

5.4 Overall family-U statistic

Following the same notations, we re-write overall family-U in matrix form:

$$\begin{aligned}
U_{AA}^{Overall} &= \frac{1}{N_{AA}(N-N_{AA})} \sum_{(i_1, j_1) \in S_{AA}} \sum_{(i_2, j_2) \in S_{AA}^c} (y_{i_1, j_1} - y_{i_2, j_2}) \\
&= \frac{1}{N_{AA}} \sum_{(i_1, j_1) \in S_{AA}} y_{i_1, j_1} - \frac{1}{N-N_{AA}} \sum_{(i_2, j_2) \in S_{AA}^c} y_{i_2, j_2} \\
&= \frac{1}{N_{AA}} \sum_{i=1}^M \sum_{j=1}^{n_m} y_{i,j} I(x_{i,j} = AA) - \frac{1}{N-N_{AA}} \sum_{i=1}^M \sum_{j=1}^{n_m} y_{i,j} I(x_{i,j} \neq AA) \\
&= \sum_{i=1}^M \sum_{j=1}^{n_i} y_{i,j} \left[\frac{I(x_{i,j} = AA)}{N_{AA}} - \frac{I(x_{i,j} \neq AA)}{N-N_{AA}} \right] \\
&= \sum_{i=1}^M \beta'_i(AA) \cdot Y_i,
\end{aligned}$$

where $\beta_i(AA) = \frac{I(X_i = AA)}{N_{AA}} - \frac{I(X_i \neq AA)}{N-N_{AA}}$ and $I(\cdot)$ is an indicator function.

Similarly, we have $U_{Aa}^{Overall} = \sum_{i=1}^M \beta'_i(Aa) \cdot Y_i$, and $U_{aa}^{Overall} = \sum_{i=1}^M \beta'_i(aa) \cdot Y_i$,

where $\beta_i(Aa) = \frac{I(X_i = Aa)}{N_{Aa}} - \frac{I(X_i \neq Aa)}{N-N_{Aa}}$ and $\beta_i(aa) = \frac{I(X_i = aa)}{N_{aa}} - \frac{I(X_i \neq aa)}{N-N_{aa}}$.

5.5 Linear Constraint for the Overall family-U Statistic

For any given constant, $c=(c_{AA}, c_{Aa}, c_{aa})$, the linear combination of univariate U-Statistics

$$c_{AA}U_{AA}^{Overall} + c_{Aa}U_{Aa}^{Overall} + c_{aa}U_{aa}^{Overall} = \sum_{i=1}^M [c_{AA}\beta'_i(AA) + c_{Aa}\beta'_i(Aa) + c_{aa}\beta'_i(aa)] \cdot Y_m = \sum_{i=1}^M \alpha'_i \cdot Y_i,$$

where

$$\begin{aligned} \alpha'_i &= c_{AA}\beta'_i(AA) + c_{Aa}\beta'_i(Aa) + c_{aa}\beta'_i(aa) \\ &= \left(\frac{c_{AA}}{N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right) I(X_i = AA) \\ &\quad + \left(-\frac{c_{AA}}{N - N_{AA}} + \frac{c_{Aa}}{N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right) I(X_i = Aa) \\ &\quad + \left(-\frac{c_{AA}}{N - N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} + \frac{c_{aa}}{N_{aa}} \right) I(X_i = aa) \end{aligned}$$

When $c=c_0=(N_{AA}(N-N_{AA}), N_{Aa}(N-N_{Aa}), N_{aa}(N-N_{aa}))$, we have

$$\begin{aligned} \frac{c_{AA}}{N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} &= N - N_{AA} - N_{Aa} - N_{aa} = 0 \\ -\frac{c_{AA}}{N - N_{AA}} + \frac{c_{Aa}}{N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} &= -N_{AA} + N - N_{Aa} - N_{aa} = 0; \\ -\frac{c_{AA}}{N - N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} + \frac{c_{aa}}{N_{aa}} &= -N_{AA} - N_{Aa} + N - N_{aa} = 0 \end{aligned}$$

Let f_{AA} , f_{Aa} and f_{aa} be the frequencies of genotypes AA, Aa and aa in the study population respectively. It follows that:

$$f_{AA}(1 - f_{AA})U_{AA}^{Overall} + f_{Aa}(1 - f_{Aa})U_{Aa}^{Overall} + f_{aa}(1 - f_{aa})U_{aa}^{Overall} = 0.$$

5.6 Asymptotic Distribution of the Overall family-U Statistic

Similar to the within family-U statistic, the overall family-U statistic,

$U^{Overall} = (U_{AA}^{Overall}, U_{Aa}^{Overall}, U_{aa}^{Overall})$, also follows asymptotically a degenerate multivariate normal distribution. For this purpose, it is sufficient to show that the linear combination of U-Statistics, $c_{AA}U_{AA}^{Overall} + c_{Aa}U_{Aa}^{Overall} + c_{aa}U_{aa}^{Overall}$, follows asymptotically a univariate normal distribution for any $c = (c_{AA}, c_{Aa}, c_{aa}) \neq c_0$. Given that $Y_i, 1 \leq i \leq M$ are independent, it is also sufficient to show:

$$\lim_{M \rightarrow \infty} \left(\frac{\max_{1 \leq i \leq M} \|\alpha'_i\|^2}{\sum_{i=1}^M \|\alpha'_i\|^2} \right) = 0.$$

Denote $n_{i,AA}, n_{i,Aa}, n_{i,aa}$ as the number of subjects with genotypes AA, Aa, aa in pedigree i , respectively. Therefore,

$$\begin{aligned} \|\alpha'_i\|^2 = & \left(\frac{c_{AA}}{N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 n_{i,AA} \\ & + \left(-\frac{c_{AA}}{N - N_{AA}} + \frac{c_{Aa}}{N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 n_{i,Aa}, \\ & + \left(-\frac{c_{AA}}{N - N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} + \frac{c_{aa}}{N_{aa}} \right)^2 n_{i,aa} \end{aligned}$$

and

$$\begin{aligned}
\sum_{i=1}^M \|\alpha'_i\|^2 &= \left(\frac{c_{AA}}{N_{AA}} - \frac{c_{Aa}}{N - N_{AA}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 N_{AA} \\
&\quad + \left(-\frac{c_{AA}}{N - N_{AA}} + \frac{c_{Aa}}{N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 N_{Aa} \\
&\quad + \left(-\frac{c_{AA}}{N - N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} + \frac{c_{aa}}{N_{aa}} \right)^2 N_{aa} \\
&= \left(c_{AA} \frac{1}{f_{AA}} - c_{Aa} \frac{1}{1 - f_{Aa}} - c_{aa} \frac{1}{1 - f_{aa}} \right)^2 \frac{f_{AA}}{N} \\
&\quad + \left(-c_{AA} \frac{1}{1 - f_{AA}} + c_{Aa} \frac{1}{f_{Aa}} - c_{aa} \frac{1}{1 - f_{aa}} \right)^2 \frac{f_{Aa}}{N} \\
&\quad + \left(-c_{AA} \frac{1}{1 - f_{AA}} - c_{Aa} \frac{1}{1 - f_{Aa}} + c_{aa} \frac{1}{f_{aa}} \right)^2 \frac{f_{aa}}{N}
\end{aligned}$$

On the other hand, assuming the number of subjects in each pedigree is bounded by

$n_i \leq n, 1 \leq i \leq M$, we have

$$\begin{aligned}
\max_{1 \leq i \leq M} \|\alpha'_i\|^2 &\leq \left(\frac{c_{AA}}{N_{AA}} - \frac{c_{Aa}}{N - N_{AA}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 n && \left(\frac{c_{AA}}{f_{AA}} - \frac{c_{Aa}}{1 - f_{Aa}} - \frac{c_{aa}}{1 - f_{aa}} \right)^2 \frac{n}{N^2} \\
&\quad + \left(-\frac{c_{AA}}{N - N_{AA}} + \frac{c_{Aa}}{N_{Aa}} - \frac{c_{aa}}{N - N_{aa}} \right)^2 n && = \quad + \left(-\frac{c_{AA}}{1 - f_{AA}} + \frac{c_{Aa}}{f_{Aa}} - \frac{c_{aa}}{1 - f_{aa}} \right)^2 \frac{n}{N^2} \\
&\quad + \left(-\frac{c_{AA}}{N - N_{AA}} - \frac{c_{Aa}}{N - N_{Aa}} + \frac{c_{aa}}{N_{aa}} \right)^2 n && + \left(-c_{AA} \frac{1}{1 - f_{AA}} - \frac{c_{Aa}}{1 - f_{Aa}} + \frac{c_{aa}}{f_{aa}} \right)^2 \frac{n}{N^2}
\end{aligned}$$

Define

$$H = \begin{pmatrix} \frac{1}{f_{AA}} & -\frac{1}{1 - f_{Aa}} & -\frac{1}{1 - f_{aa}} \\ -\frac{1}{1 - f_{AA}} & \frac{1}{f_{Aa}} & -\frac{1}{1 - f_{aa}} \\ -\frac{1}{1 - f_{AA}} & -\frac{1}{1 - f_{Aa}} & \frac{1}{f_{aa}} \end{pmatrix},$$

and it is easy to know $\text{rank}(H)=2$. Then, for any $c=(c_{AA}, c_{Aa}, c_{aa})$, we have $H \cdot c' \neq 0$ except $c=c_0$. Therefore,

$$\lim_{M \rightarrow \infty} \left(\frac{\max_{1 \leq i \leq M} \|\alpha'_i\|^2}{\sum_{i=1}^M \|\alpha'_i\|^2} \right) \leq \lim_{N \rightarrow \infty} \frac{n/N^2}{1/N} = 0.$$

This proof shows that the U-Statistic $U^{Overall} = (U_{AA}^{Overall}, U_{Aa}^{Overall}, U_{aa}^{Overall})$ follows asymptotically a degenerate multivariate normal distribution with one and only one linear constraint:

$$f_{AA}(1-f_{AA})U_{AA}^{Overall} + f_{Aa}(1-f_{Aa})U_{Aa}^{Overall} + f_{aa}(1-f_{aa})U_{aa}^{Overall} = 0.$$

Table S1. Performance of two distance measurements

Mode of inheritance	Pedigree	Power/T1E	Mahalanobis distance		Euclidean distance	
			Overall family-U	Within family-U	Overall family-U	Within family-U
Additive Effect	Nuclear families ^a	Power T1E ^d	0.996 0.051	0.774 0.060	0.990 0.055	0.689 0.062
	Three-generation ^b	Power T1E	0.990 0.046	0.901 0.040	0.989 0.044	0.890 0.031
	Mixture pedigree ^c	Power T1E	0.984 0.047	0.838 0.049	0.981 0.042	0.794 0.052
Dominant Effect	Nuclear families	Power T1E	0.908 0.048	0.635 0.059	0.923 0.051	0.627 0.061
	Three-generation	Power T1E	0.928 0.047	0.787 0.046	0.936 0.048	0.806 0.041
	Mixture pedigree	Power T1E	0.904 0.043	0.709 0.052	0.910 0.043	0.714 0.058

^a Nuclear families with mother, father, and two offspring (i.e., quad); ^b Three-generation pedigrees with eight family members; ^c A mixture of nuclear families and three-generation pedigrees;

^d Type I error.

Table S2. Performance of Family U and FBAT when founders come from different ethnic groups

Quantitative Phenotypes		Fst=0.1			Fst=0.2		
Proportion of families with admixture founders	Power/ TIE	Within Fam-U	Within Fam-U-M ^c	FBAT	Within Fam-U	Within Fam-U-M	FBAT
10% ^a	Power TIE ^b	0.494 0.069	0.367 0.065	0.440 0.051	0.513 0.064	0.318 0.053	0.465 0.053
30%	Power TIE	0.503 0.075	0.338 0.054	0.463 0.042	0.489 0.070	0.352 0.046	0.459 0.056
50%	Power TIE	0.476 0.062	0.344 0.063	0.468 0.047	0.491 0.099	0.361 0.068	0.477 0.047
Binary Phenotypes		Fst=0.1			Fst=0.2		
Proportion of families with admixture founders	Power/ TIE	Within Fam-U	Within Fam-U-M	FBAT	Within Fam-U	Within Fam-U-M	FBAT
10% ^a	Power TIE	0.465 0.056	0.348 0.048	0.424 0.047	0.509 0.061	0.377 0.056	0.401 0.054
30%	Power TIE	0.485 0.107	0.347 0.076	0.424 0.062	0.492 0.159	0.352 0.097	0.401 0.056
50%	Power TIE	0.467 0.166	0.345 0.093	0.404 0.057	0.484 0.280	0.355 0.184	0.407 0.055

^a 10% of the nuclear families have two founders sampled from two subpopulations with different levels of PS ;

^b Type I Error;

^c The modified within family-U test that excludes comparison between founders.