Published in final edited form as:

Genet Epidemiol. 2014 February ; 38(2): 135–143. doi:10.1002/gepi.21787.

A Novel Test for Testing the Optimally Weighted Combination of Rare and Common Variants Based on Data of Parents and Affected Children

Qiuying Sha and Shuanglin Zhang*

Department of Mathematical Sciences, Michigan Technological University, Houghton, MI 49931
USA

Abstract

With the development of sequencing technologies, the direct testing of rare variant associations has become possible. Many statistical methods for detecting associations between rare variants and complex diseases have recently been developed, most of which are population-based methods for unrelated individuals. A limitation of population-based methods is that spurious associations can occur when there is a population structure. For rare variants, this problem can be more serious, since the spectrum of rare variation can be very different in diverse populations, as well as the current nonexistence of methods to control for population stratification in population-based rare variant associations. A solution to the problem of population stratification is to use family-based association tests, which use family members to control for population stratification. In this article, we propose a novel test for Testing the Optimally Weighted combination of variants based on data of Parents and Affected Children (TOW-PAC). TOW-PAC is a family-based association test that tests the combined effect of rare and common variants in a genomic region, and is robust to the directions of the effects of causal variants. Simulation studies confirm that, for rare variant associations, family-based association tests are robust to population stratification while population-based association tests can be seriously confounded by population stratification. The results of power comparisons show that the power of TOW-PAC increases with an increase of the number of affected children in each family and TOW-PAC based on multiple affected children per family is more powerful than TOW based on unrelated individuals.

Keywords

optimal weights; rare variants; association studies; family-based design

Introduction

Since the disease associated common variants identified through genome-wide association studies (GWAS) account for only a small portion of the presumed phenotypic variation (McCarthy *et al.*, 2008) and also due to the development of sequencing technology, there is

---

*Corresponding author.

The authors have no conflict of interests to declare.

increasing interest to detect associations between rare variants and complex traits. Many statistical methods for testing associations of rare variants have been developed. The current popular methods include burden tests (Morgenthaler and Thilly, 2007; Li and Leal, 2008; Madsen and Browning, 2009; Price *et al.*, 2010; Zawistowski *et al.*, 2010), quadratic tests (Neale *et al.*, 2011; Wu *et al.*, 2011; Sha *et al.*, 2012), and combined tests (Derkach *et al.*, 2012; Lee *et al.*, 2012).

Burden tests collapse rare variants in a genomic region into a single burden variable and then regress the phenotype on the burden variable to test for the cumulative effects of rare variants in the region (Lee *et al.*, 2012). These tests implicitly assume that all rare variants are causal and that the directions of the effects are all the same. Quadratic tests include tests with statistics of quadratic forms of the score vector (Neale *et al.*, 2011; Wu *et al.*, 2011; Sha *et al.*, 2012), as well as adaptive weighting methods (Han and Pan, 2010; Hoffmann *et al.*, 2010; Lin and Tang, 2011; Yi and Zhi, 2011; Sha *et al.*, 2013). Quadratic tests are robust to the directions of the effects of causal variants and are less affected by neutral variants than burden tests. Burden tests can only outperform quadratic tests when most of rare variants are causal and the directions of the effects of causal variants are all the same. Combined tests combine information from burden tests, quadratic tests, and possibly other tests aiming to have advantages of burden, quadratic, and possibly other tests.

All the aforementioned methods are population-based methods for unrelated individuals. A limitation of these methods is that it is difficult to control for population stratification and spurious associations can occur. In population-based association studies, it has been long recognized that population stratification can seriously confound association results (Knowler *et al.*, 1988; Lander and Schork, 1994). For rare variants this problem can be more serious, since the spectrum of rare variation can be very different in diverse populations. In common variant association studies, several methods that use a set of genomic markers have been developed to control for population stratification. These methods include the genomic control (GC) approach (Devlin and Roeder, 1999; Devlin *et al.*, 2001; Reich and Goldstein, 2001), the principal component (PC) approach (Chen *et al.*, 2003; Zhang *et al.*, 2003; Zhu *et al.*, 2002; Price *et al.*, 2006; Bauchet *et al.*, 2007), and the mixed linear model (MLM) approach (Kang *et al.*, 2010; Zhang *et al.*, 2010). Mathieson and McVean (2012) demonstrated that rare variants can show a stratification that is systematically different from common variants. They also demonstrated that the commonly used methods such as the GC, PC, and MLM approaches to control for population stratification in population-based common variant associations are not necessarily controlling for population stratification in population-based rare variant associations.

A solution to the problem of population stratification is to use family-based association tests. In family-based data, association information can be partitioned into between-family and within-family information (Fang *et al.*, 2012). Within-family information is robust to population stratification while between-family information can be confounded by population stratification. Recently, two groups of rare variant association tests based on family data have been proposed. The first group, including methods proposed by Fang *et al.*, (2012), Chen *et al.*, (2012), Schaid *et al.*, (2013), and Ouakacha *et al.*, (2013), uses both between-family and within-family information. The second group, including the methods proposed

by Fang et al., (2013), De et al., (2013), and Ionita-Laza et al., (2013), uses within-family information only. The methods in the first group have the same problem as population-based methods, that is, difficulty in controlling for population stratification. The methods in the second group, also called family-based rare variant association tests, are robust to population stratification.

From the statistical power point of view, the power under any type of study design will be improved if rare variants can be enriched in the samples. To enrich rare variants in the samples, families with multiple affected subjects are a good choice (Bodmer and Bonilla, 2008). Based on nuclear families with multiple affected children, we propose a novel test for Testing the Optimally Weighted combination of variants for data of Parents and Affected Children (TOW-PAC). TOW-PAC belongs to family-based rare variant association tests and is robust to the directions of the effects of causal variants. Comparing the existing family-based rare variant association tests (Fang et al., 2013; De et al., 2013; Ionita-Laza et al., 2013) with our proposed TOW-PAC, the method of Fang et al. is only applicable to quantitative traits while TOW-PAC is applicable to qualitative traits; the method of De et al. is a burden test that will lose power when both risk and protective variants are present while TOW-PAC is a quadratic test that is robust to the directions of the effects of causal variants; the method of sequence kernel association test (SKAT) with family-based designs proposed by Ionita-Laza *et al.* is also a quadratic test, but SKAT focuses on trios while TOW-PAC emphasizes on multiple affected children per family. Simulation studies confirm that, for rare variant associations, family-based association tests are robust to population stratification while population-based association tests can be seriously confounded by population stratification. The results of power comparisons show that the power of TOW-PAC increases with an increase of the number of affected children in each family and TOW-PAC based on multiple affected children per family is more powerful than TOW based on unrelated individuals. We also compare the power of TOW-PAC with the powers of other family-based association tests including single-variant PDT (SINGLE), where PDT refers to Pedigree Disequilibrium Test (Martin et al., 2000), multi-variant PDT (MPDT) (Zhang et al., 2007), and Weighted Sum method based on data of Parents and Affected Children (WS-PAC). MPDT essentially compares the numbers of transmitted with the numbers of non-transmitted minor alleles at multiple variants. TOW-PAC, SINGLE, and MPDT are quadratic tests while WS-PAC is a burden test. Our results show that TOW-PAC is consistently more powerful than MPDT and MPDT is consistently more powerful than SINGLE. Power comparisons of TOW-PAC and WS-PAC depend on the percentage of causal variants and the percentage of protective variants.

Method

Consider a sample of n nuclear families with two parents and n_i affected children in the i^{th} family. Suppose that each individual has been genotyped at M variants in a genomic region.

Let $g_{ij} = (g_{ij1}, \dots, g_{ijM})^T$, $g_i^m = (g_{i1}^m, \dots, g_{iM}^m)^T$, and $g_i^f = (g_{i1}^f, \dots, g_{iM}^f)^T$ denote the multi-variant genotypes of the j^{th} child, the mother, and the father in the i^{th} family, respectively, where g_{ijm} is the number of copies of the minor allele of the j^{th} child in i^{th} family at the m^{th} variant site, and (g_{im}^m, g_{im}^f) are similarly defined. We use D_{ij} to represent that the j^{th} child in

the i^{th} family is affected. The probability of genotypes of the affected children, given the genotypes of the parents, is given by

$$= \frac{P(g_{i1}, \dots, g_{in_i} | g_i^m, g_i^f, D_{i1}, \dots, D_{in_i})}{\sum_{g_{i1}^a \in G_i} \dots \sum_{g_{in_i}^a \in G_i} \frac{P(D_{i1} | g_{i1}^a) P(g_{i1}^a | g_i^m, g_i^f) \dots P(D_{in_i} | g_{in_i}^a) P(g_{in_i}^a | g_i^m, g_i^f)}{P(D_{i1} | g_{i1}) \dots P(D_{in_i} | g_{in_i})}}$$

Here, G_i is the set of possible genotypes the parents of the i^{th} family can produce. Choose a baseline genotype and let $r(g)$ be the relative risk of genotype g to the chosen baseline genotype. Then,

$$P(g_{i1}, \dots, g_{in_i} | g_i^m, g_i^f, D_{i1}, \dots, D_{in_i}) = r(g_{i1}) \dots r(g_{in_i}) / \left(\sum_{g^a \in G_i} r(g^a) P(g^a | g_i^m, g_i^f) \right)^{n_i}.$$

Following Schaid (1996), we use log-linear model to model the relative risk, that is, $r(g) =$

$\exp(X^T \beta)$, with X representing the numerical coding of genotype g . Let $X_{ij} = \sum_{m=1}^M w_m g_{ijm}$, where $w_1, \dots, w_M$ are unknown constants. We will find the values of $w_1, \dots, w_M$ later according to some optimal criterion. Under the assumption that the M variants are independent (our proposed test is still valid if this assumption is not true), the conditional likelihood function is given by

$$L(\beta) = \prod_{i=1}^n \exp(\beta \sum_{j=1}^{n_i} X_{ij}) / \left(\prod_{m=1}^M \sum_{g_{im}^a \in G_{im}} \exp(w_m g_{im}^a \beta) \right)^{n_i},$$

where G_{im} is the set of the four possible genotypes at the m^{th} variant the parents of the i^{th} family can produce. Under this likelihood function, the score test statistic to test the null hypothesis $H_0: \beta = 0$ is

$$T(w_1, \dots, w_M) = \frac{w^T u u^T w}{w^T v w}, \quad (1)$$

where $u = \sum_{i=1}^n \sum_{j=1}^{n_i} g_{ij} - \sum_{i=1}^n n_i \bar{g}_i$, $v = \text{diag}(\sum_{i=1}^n n_i \hat{\sigma}_i^2)$, $\bar{g}_i = (\bar{g}_{i1}, \dots, \bar{g}_{iM})^T$, $\hat{\sigma}_i^2 = (\hat{\sigma}_{i1}^2, \dots, \hat{\sigma}_{iM}^2)^T$,
 $\bar{g}_{im} = \frac{1}{4} \sum_{g_{im}^a \in G_{im}} g_{im}^a$, and $\hat{\sigma}_{im}^2 = \frac{1}{4} \left(\sum_{g_{im}^a \in G_{im}} g_{im}^a{}^2 - 4 \bar{g}_{im}^2 \right)$.

The score test statistic $T(w_1, \dots, w_M)$, as a function of w , reaches its maximum when $w = v^{-1}u$. Our proposed test statistic for Testing the Optimally Weighted combination of variants based on data of Parents and Affected Children (TOW-PAC) is the maximum value of $T(w_1, \dots, w_M)$, that is,

$$T_{TOW-PAC} = \max_{w_1, \dots, w_M} T(w_1, \dots, w_M) = u^T v^{-1} u = \sum_{m=1}^M \frac{u_m^2}{v_m}, \quad (2)$$

where $u_m = \sum_{i=1}^n \sum_{j=1}^{n_j} g_{ijm} - \sum_{i=1}^n n_i \bar{g}_{im}$, and $v_m = \sum_{i=1}^n n_i \hat{\sigma}_{im}^2$. Note that $\bar{g}_{im} = E(g_{i1m} | g_{im}^p) = \dots = E(g_{in_i m} | g_{im}^p)$ and $\hat{\sigma}_{im}^2 = \text{var}(g_{i1m} | g_{im}^p) = \dots = \text{var}(g_{in_i m} | g_{im}^p)$, where $g_{im}^p = (g_{im}^f, g_{im}^m)$ is the genotype of the parents of the i^{th} family at the m^{th} variant.

Then, u_m and v_m can be also written as $u_m = \sum_{i=1}^n \sum_{j=1}^{n_j} (g_{ijm} - E(g_{ijm} | g_{im}^p))$ and $v_m = \sum_{i=1}^n n_i \text{var}(g_{i1m} | g_{im}^p)$.

We use a permutation test to evaluate the p-values of the proposed test. For the j^{th} child in the i^{th} family with genotype $g_{ij} = (g_{ij1}, \dots, g_{ijM})^T$, we define another genotype

$g_{ij}^* = (g_{ij1}^*, \dots, g_{ijM}^*)^T$, where g_{ijm}^* is the genotype formed by the two alleles at the m^{th} variant that the parents of the i^{th} family did not transmit to the j^{th} child. If there is only one child in each family, we can replace each g_{ij} with either itself or g_{ij}^* with equal probability in each permutation. For the case of more than one child, as noted by Monks and Kaplan (2000) and Fang *et al.*, (2012), complications arise in the presence of linkage. In this case, children with shared alleles will have similar trait values, even in the absence of association. For these families, we replace each $(g_{i1}, \dots, g_{in_i})$ with either itself or $(g_{i1}^*, \dots, g_{in_i}^*)$ with equal probability (Monks and Kaplan, 2000).

For a simulation study with R replicates, the above procedure will be rather computationally expensive. In our simulation studies, we use the pooling permutation method proposed by Guo and Lin (2009) to evaluate p-values. In the pooling permutation method, permuted samples from all the replicates are pooled together to form a joint sample from the null distribution. Suppose that we have R replicates and we perform B permutations for each replicate. Let $T_{TOW-PAC}^{(b,r)}$ denote the value of $T_{TOW-PAC}$ based on the data of the b^{th} permutation of the r^{th} replicate ($b = 0$ denotes the original data). Then, the p-value of the test in the r^{th} replicate is given by

$$p\text{-value} = \# \{ (b, r_0) : T_{TOW-PAC}^{(b,r_0)} > T_{TOW-PAC}^{(0,r)} ; b=1, \dots, B; r_0=1, \dots, R \} / (BR).$$

Since the permutation samples are pooled across all replicates to form a sample from the null, B can be set to be much smaller than the situation when only one sample is analyzed.

Comparison of Methods

We will compare the performance of the proposed method with three family-based association tests and two population-based association tests. The three family-based

association tests are Multi-variant PDT (MPDT) (Zhang et al., 2007), Single-variant PDT (SINGLE) (Martin et al., 2000), and Weighted Sum method based on data of Parents and Affected Children (WS-PAC). WS-PAC is the score test $T(w_1, \dots, w_M)$ given by equation 1 with the weights $w_1, \dots, w_M$ given by Madsen and Browning (2009). The two population-based association tests are TOW (Sha et al., 2012) and WS (Madsen and Browning, 2009). Using the notation in the method section, the statistics of the three family-based association tests are given by $T_{MPDT} = U^T V^{-1} U$, $T_{SINGLE} = \max(T_1, \dots, T_M)$, and $T_{WS-PAC} = U_w^2 / V_w$,

$$\text{where } U = \sum_{i=1}^n U_i, V = \sum_{i=1}^n U_i U_i^T, U_i = \sum_{j=1}^{n_j} (g_{ij} - \bar{g}_i), T_m = (U^{(m)})^2 / V^{(m)}, U^{(m)} = \sum_{i=1}^n U_i^{(m)},$$

$$V^{(m)} = \sum_{i=1}^n (U_i^{(m)})^2, U_i^{(m)} = \sum_{j=1}^{n_j} (g_{ijm} - \bar{g}_{im}), U_w = \sum_{i=1}^n \sum_{j=1}^{n_j} \sum_{m=1}^M w_m (g_{ijm} - \bar{g}_{im}), \text{ and}$$

$$V_w = \sum_{i=1}^n n_i \sum_{m=1}^M w_m^2 \hat{\sigma}_{im}^2. \text{ For the population-based association tests, denote } y_i \text{ and } X_i = (x_{i1}, \dots, x_{iM})^T \text{ as the trait value (1 for cases and 0 for controls) and genotypic score of the } i^{\text{th}}$$

individual. Then, the statistics of TOW and WS are given by $T_{TOW} = \sum_{m=1}^M u_m^2 / v_m$ and

$$T_{WS} = U^2 / V, \text{ where } u_m = \sum_{i=1}^n (y_i - \bar{y}) x_{im}, v_m = \sum_{i=1}^n (x_{im} - \bar{x}_m)^2, U = \sum_{i=1}^n (y_i - \bar{y}) x_i,$$

$$V = \sum_{i=1}^n (x_i - \bar{x})^2, x_i = \sum_{m=1}^M w_m x_{im}, \text{ and the weights } w_1, \dots, w_M \text{ are those suggested by Madsen and Browning (2009).}$$

Simulation

The empirical Mini-Exome genotype data provided by the genetic analysis workshop 17 (GAW17) is used for simulation studies. This dataset contains genotypes of 697 unrelated individuals on 3,205 genes. The genotypes of the GAW17 data set are extracted from the sequence alignment files provided by the 1,000 Genomes Project for their pilot3 study (<http://www.1000genomes.org>). We choose four genes: ELAVL4 (gene1), MSH4 (gene2), PDE4B (gene3), and ADAMTS4 (gene4) with 10, 20, 30, and 40 variants, respectively. We merge the four genes to form a super gene (Sgene) with 100 variants (86 rare variants (MAF < 0.01) and 14 common variants). We form Sgene for two reasons: (1) the distributions of MAFs in the 100 variants in the Sgene and in the 24,487 variants in all the 3,205 genes are very similar (Sha et al., 2012); (2) there are enough variants in the Sgene to evaluate the performance of the proposed method for different percentages of casual variants. In our simulation studies, we generate genotypes based on the genotypes of 697 individuals in the Sgene. We use the program fastPHASE (Scheet and Stephens, 2006) to infer haplotypic phase for the 697 individuals and calculate haplotype frequencies. To generate the genotype of an individual, we generate two haplotypes according to the haplotype frequencies. To obtain the genotypes of a family, we first generate the genotypes of the parents, and then generate the genotypes of the children from parental haplotypes by random transmission. In type I error evaluation, we consider three sets of simulations: a homogeneous population, a structured population with two subpopulations, and a spatially structured population. For

power comparisons, we only consider the first two sets of simulations. To generate a qualitative disease affection status, we use a liability threshold model based on a continuous phenotype (quantitative trait). An individual is defined to be affected if the individual's phenotype is at least one standard deviation larger than the phenotypic mean. This yields a prevalence of 16% for the simulated disease in the general population. To generate families with k affected children, we follow Feng et al. (2011) and Zhu et al. (2010). Briefly, we keep generating families with k children until we have generated enough families with k affected children. In the following, we describe how to generate a quantitative trait in the three sets of simulations.

Simulation Set 1: A Homogeneous Population

Under the null hypothesis, for each family with k children, let $Y_1 = (y_F, y_M)$ and $Y_2 = (y_1, y_2, \dots, y_k)$ denote the trait values of the parents and the k children in a family, respectively. We generate trait values (Y_1, Y_2) under the assumption that (Y_1, Y_2) follows a multivariate normal distribution with a mean vector of zero and variance-covariance matrix Σ . The variance-covariance matrix Σ has the structure that the parents in each family are independent, and the correlation coefficient between a parent and a child or between two children is a constant ρ (in this study, $\rho = 0.2$).

Under the alternative hypothesis, we choose n_{cau} rare variants ($MAF < 1\%$) as causal variants. The value of n_{cau} is determined by p_{cau} , the percentage of causal variants in rare variants. Let pp denote the percentage of protective variants in causal variants, then the number of protective variants and the number of risk variants are $n_p = n_{cau} \cdot pp$ and $n_r = n_{cau} \cdot (1 - pp)$, respectively. For the j^{th} member in the i^{th} family, let x_{ijk_r} and x_{ijk_p} denote the genotypic scores of the k_r^{th} risk variant and the k_p^{th} protective variant, respectively. Assuming that all causal variants have the same heritability, then the disease model is given by

$$y_{ij} = \sum_{k_r=1}^{n_r} \beta_{k_r} x_{ijk_r} - \sum_{k_p=1}^{n_p} \beta_{k_p} x_{ijk_p} + \varepsilon_{ij}, \text{ where } \beta_{k_r} \text{ and } \beta_{k_p} \text{ are coefficients and their values depend on the total heritability, and } \varepsilon_{ij} \text{ is the trait value under the null hypothesis.}$$

Simulation Set 2: A Structured Population with Two Subpopulations

In this set of simulations, we consider two subpopulations: Asian and African. The Asian subpopulation includes Han Chinese in Beijing (CHB), Chinese in Metropolitan Denver (CHD), and Japanese in Tokyo (JPT), while the African subpopulation includes African ancestry in Southwest USA (ASW), Luhya in Webuye, Kenya (LWK), Maasai in Kinyawa, Kenya (MKK), and Yoruba in Ibadan, Nigeria (YRI). Based on the empirical Mini-Exome genotype data provided by GAW17, we calculated haplotype frequencies at Sgene for each subpopulation. Within each subpopulation, we obtain the genotypes in the same way as Simulation Set 1. Under null hypothesis, the trait values are generated with the same method as Simulation Set 1, except that in the structured population, the trait values of a family in subpopulation 1 are from a multivariate normal distribution with a mean of 0 and those in subpopulation 2 are from a multivariate normal distribution with a mean of μ ($\mu = 1$ in this study). By Ionita-Laza *et al.*, (2007), the differences in allele frequencies and in means of trait values together result in spurious associations.

Under the alternative hypothesis, the trait values are generated in the same way as Simulation Set 1. In our simulation studies, we choose half families from subpopulation 1 and half families from subpopulation 2.

Simulation Set 3: A Spatially Structured Population

We generate genotypes under a spatially structured population using models described in Mathieson and McVean (2012) and Fang *et al.*, (2013). To generate family-based genotypes, we first generate the genotypes of the parents following Mathieson and McVean (2012) and then generate the genotypes of the children by randomly transmitting parental haplotypes. Briefly, we divide the space into $K \times K$ grid squares. We generate parents' genotypes by starting with a number of individuals and their locations on the grid, assuming that the parents of a family are in the same grid, and working backward in time to generate random genealogical events. Each event is either a coalescence of two lineages or a migration of a single lineage from one square to another. The relative rates of coalescence and migration depend on the population-scaled migration rate M and the number and distribution of lineages on the grid (see Supplement Materials or Mathieson and McVean (2012) for details).

To generate quantitative traits under null hypothesis, we first generate trait values of all families using the method described in Simulation Set 1 and denote the generated trait values of the i^{th} family as ε_i , where ε_i is a vector that contains the trait values of all members of the i^{th} family. Let $\phi: [1, n] \rightarrow [1, K] \times [1, K]$ be a function that mapped each family to the grid square from which they originated. Then, for the i^{th} family, we generate the trait values of the i^{th} family by $y_i = \beta R_{\phi(i)} + \varepsilon_i$, where $R_{k,j}$ is the nongenetic risk in grid square k, j and β is a constant. We use the following two models to determine the value of $R_{k,j}$. Model 1: given k_0 and j_0 , $R_{k,j} = 1$ if $k = k_0$ and $j = j_0$; $R_{k,j} = 0$ otherwise. Model 2: given k_0 and j_0 , $R_{k,j} = 0.4e^{-((k-k_0)^2 + (j-j_0)^2)/8}$. In this study, we use the following parameters: $K = 20$, $M = 0.01$, $k_0 = j_0 = 6$, and $\beta = 2$.

Results

In simulation studies, p-values are estimated using a pooling permutation method (Guo and Lin, 2009) in which the permuted samples from all the replicates are pooled together to form a joint sample from the null distribution. For each replicate, we perform 20 permutations. Type I error rates are evaluated using 10,000 replicated samples, while powers are evaluated using 500 replicated samples and a significance level of 0.01.

For evaluating type I error rates, we consider different haplotype structures (different genes), different significance levels, and different population structures. All type I error rates are based on 400 families with two affected children per family. For 10,000 replicated samples, the 95% confidence intervals (CIs) for type I error rates of nominal levels 0.05, 0.01, and 0.001 are (0.046, 0.054), (0.008, 0.012), and (0.0004, 0.0016), respectively. The estimated type I error rates of the five tests in a homogenous population are summarized in Table 1. As shown in this table, all the estimated type I error rates of the five tests are within the 95% CIs, which indicates that all the five tests are valid tests in homogenous populations. The

estimated type I error rates of the five tests in a structured population with two subpopulations and in a spatially structured population are summarized in Table 2 and Table 3, respectively. These two tables show that the estimated type I error rates of the four family-based tests (TOW-PAC, WS-PAC, MPDT, and SINGLE) are all within the 95% CIs while the estimated type I error rates of TOW based on unrelated individuals are all significantly higher than the nominal levels. This indicates that, in structured populations, the four family-based tests are valid tests but the one based on unrelated individuals has inflated type-I error.

For power comparisons, we first compare the powers of TOW-PAC and WS-PAC with different numbers of affected children per family and the powers of TOW and WS based on unrelated individuals while we either fix the total number of individuals or fix the total number of affected individuals in different designs. Results (Figure 1) show that the powers of TOW-PAC and WS-PAC increase with an increase of the number of affected children per family. When there are multiple affected children per family, TOW-PAC and WS-PAC are more powerful than TOW and WS, respectively. Furthermore, we compare the power of Fam-SKAT (Chen et al., 2012), one of mixed linear model approaches, with the power of TOW-PAC. Results (Figure S1) show that, with the same sequencing efforts, powers of both Fam-SKAT and TOW-PAC increase with an increase of the number of affected children per family, but the power of Fam-SKAT increases not as fast as the power of TOW-PAC. As expected, Fam-SKAT is more powerful than TOW-PAC because Fam-SKAT uses both within-family and between-family information while TOW-PAC uses within-family information only.

In the following power comparisons, we compare the powers of the four family-based tests: TOW-PAC, WS-PAC, MPDT, and SINGLE. Power comparisons for power as a function of the percentage of protective variants in a homogenous population are given by Figure 2. As shown by this figure, TOW-PAC is consistently more powerful than MPDT and MPDT is consistently more powerful than SINGLE. TOW-PAC is only less powerful than WS-PAC when the percentage of causal variants is large and the percentage of protective variants is small. Theoretically, TOW-PAC, MPDT, and SINGLE should be robust to the directions of the effects of causal variants. In our simulations, however, the powers of TOW-PAC, MPDT, and SINGLE decrease with increasing percentage of protective variants, although they do not decrease as fast as WS-PAC. As pointed out by Wu *et al.*, (2011) and Sha *et al.*, (2012), decrease in the powers of TOW-PAC, MPDT, and SINGLE in the presence of both risk and protective variants is due to the fact that protective variants lower MAFs in cases and thus make observing rare variants in cases more difficult. The larger decrease in the power of WS-PAC is additionally driven by sensitivity to the direction of the effect due to the aggregation of genotypes.

Figure 3 shows power comparisons for power as a function of the percentage of causal variants in a homogenous population. This figure shows that, TOW-PAC is again consistently more powerful than MPDT and MPDT is again consistently more powerful than SINGLE. Power comparisons between TOW-PAC and WS-PAC depend on the percentage of causal variants and the percentage of protective variants. With increasing percentage of causal variants, the power of WS-PAC increases, the powers of TOW-PAC and MPDT

show little variation, and the power of SINGLE decreases. It is easy to understand that the power increases with increasing percentage of causal variants because larger percentage of causal variants means smaller noise level. The power of SINGLE decreases with increasing percentage of causal variants because larger percentage of causal variants means smaller heritability for each causal variant. Power comparisons of the four family-based association tests based on the structured populations yield the same conclusions (Figures S2 and S3).

Discussion

Motivated by the fact that most of the recently developed statistical methods for detecting associations between rare variants and complex traits are population-based methods for unrelated individuals and the current nonexistence of methods to control for population stratification in population-based rare variant associations (Mathieson and McVean, 2012), we propose a novel family-based association test TOW-PAC to test rare variant associations based on data of parents and affected children. TOW-PAC is a family-based association test and thus can control for population stratification. Our results show that the power of TOW-PAC increases with an increasing number of affected children per family and TOW-PAC based on multiple affected children is more powerful than TOW based on unrelated individuals. Results also show that TOW-PAC is consistently more powerful than MPDT and SINGLE. TOW-PAC is only less powerful than WS-PAC when a large portion of rare variants are causal variants and the directions of the effects of causal variants are the same.

While the quadratic test, TOW-PAC, uses optimally data-driven weights, and is robust to the directions of the effects of causal variants, the burden test, WS-PAC, implicitly assumes that all the rare variants are causal and the directions of the effects are all the same. When the assumption is true, WS-PAC can outperform TOW-PAC; otherwise, TOW-PAC outperforms WS-PAC. To increase the robustness of tests, we can also construct combined tests by combining information from TOW-PAC and WS-PAC (see Appendix A for details). In this study, we assume that parental genotypes are available. In this case, it is easy to know the distribution of offspring genotypes conditional on parental genotypes and thus can calculate $E(g_{ijm}|g_{im}^p)$ and $\text{var}(g_{i1m}|g_{im}^p)$ in equation (2). When parental genotypes are not completely known and unaffected sibships are available, as pointed out by Ionita-Laza et al. (2013), we can use the algorithm developed by Rabinowitz and Laird (2000) that specifies the distribution of offspring genotypes conditional on the sufficient statistic for the parental genotypes to calculate $E(g_{ijm}|g_{im}^p)$ and $\text{var}(g_{i1m}|g_{im}^p)$.

The computation time required for running TOW-PAC depends on the sample size, the number of variants in the genomic region, and the number of permutations. The running time of TOW-PAC with 1000 permutations on the data set with 1000 trios and 100 variants in the genomic region on a laptop with 4 Intel Cores @ 2.00GHz and 4 GB memory is no more than 0.1 s. To perform genome-wide studies, we can first select genomic regions that show evidence of association based on a small number of permutations (e.g. 1,000), and then a large number of permutations are used to test the selected regions.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Research reported in this article was supported by the National Human Genome Research Institute of the National Institutes of Health under Award Number R03 HG006155. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The Genetic Analysis Workshops are supported by NIH grant R01 GM031575 from the National Institute of General Medical Sciences. Preparation of the Genetic Analysis Workshop 17 Simulated Exome Data Set was supported in part by NIH R01 MH059490 and used sequencing data from the 1000 Genomes Project (www.1000genomes.org).

Appendix A

Let $T_{TOW-PAC}$ and T_{WS-PAC} ($p_{TOW-PAC}$ and p_{WS-PAC}) denote the statistics (p-values) of TOW-PAC and WS-PAC, respectively. We can use the following methods to combine information from TOW-PAC and WS-PAC: (1) the statistic of the combined test is given by $T_{com} = -2(\log p_{TOW-PAC} + \log p_{WS-PAC})$ (Derkach *et al.*, 2012); (2) the statistic of the combined test is given by $T_{com} = \min(p_{TOW-PAC}, p_{WS-PAC})$; (3) the statistic of the combined test is given by $T_{com} = \min_{0 \leq \lambda \leq 1} p_{\lambda}$, where p_{λ} is the p-value of the test statistic $T_{\lambda} = \lambda T_{TOW-PAC} + (1-\lambda)T_{WS-PAC}$ (Lee *et al.*, 2012).

References

- Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesyan K, Deka R, Bradley DG, Shriver MD. Measuring European population stratification with microarray genotype data. *Am J Hum Genet.* 2007; 80:948–56. [PubMed: 17436249]
- Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008; 40(6):695–701. [PubMed: 18509313]
- Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genet Epi.* 2012; 37:196–204.
- Chen HS, Zhu X, Zhao H, Zhang SL. Qualitative semi-parametric test for genetic associations in case-control designs under structured populations. *Ann Hum Genet.* 2003; 67:250–264. [PubMed: 12914577]
- De G, Yip W, Ionita-Laza I, Laird N. Rare Variant Analysis for Family-Based Design. *PLoS ONE.* 2013; 8(1):e48495. [PubMed: 23341868]
- Derkach A, Lawless J, Sun L. Robust and powerful tests for rare variants using fisher's method to combine evidence of association from two or more complementary tests. *Genetic Epi.* 2012; 37(1): 110–121.
- Devlin B, Roeder K. Genomic control for association studies. *Biometrics.* 1999; 55:997–1004. [PubMed: 11315092]
- Devlin B, Roeder K, Wasserman L. Genomic control, a new approach to genetic-based association studies. *Theor Popul Biol.* 2001; 60:155–66. [PubMed: 11855950]
- Fang S, Sha Q, Zhang S. Two adaptive weighting methods to test for rare variant associations in family-based designs. *Genet Epidemiol.* 2012; 36:499–507. [PubMed: 22674630]
- Fang S, Zhang S, Sha Q. Detecting association of rare variants by testing an optimally weighted combination of variants for quantitative traits in nuclear families. *Ann Hum Genet.* 2013 revised version was submitted.
- Feng T, Elston R, Zhu X. Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genetic Epidemiol.* 2011; 35:398–409.

- Guo W, Lin S. Generalized linear modeling with regularization for detecting common disease rare haplotype association. *Genet Epi.* 2009; 33:308–316.
- Han F, Pan W. A data-adaptive sum test for disease association with multiple common or rare variants. *Hum Hered.* 2010; 70:42–54. [PubMed: 20413981]
- Hoffmann TJ, Marini NJ, Witte JS. Comprehensive approach to analyzing rare genetic variants. *PLoS ONE.* 2010; 5(11):e13584. [PubMed: 21072163]
- Ionita-Laza I, Lee Seunggeun, Makarov V, Buxbaum JD, Lin X. Family-based association tests for sequence data, and comparisons with population-based association test. *European Journal of Human Genetics.* 2013 advance online publication, 6 February 2013. 10.1038/ejhg.2012.308
- Ionita-Laza I, McQueen MB, Laird NM, Lange C. Genome-wide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. *Am J Hum Genet.* 2007; 81:607–614. [PubMed: 17701906]
- Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010; 42:348–354. [PubMed: 20208533]
- Knowler WC, Williams RC, Pettitt DJ, Steinberg AG. Gm3;5,13,14 and type 2 diabetes mellitus: an association in American Indians with genetic admixture. *Am J Hum Genet.* 1988; 43:520–526. [PubMed: 3177389]
- Lander ES, Schork NJ. Genetic dissection of complex traits. *Science.* 1994; 265:2037–2048. [PubMed: 8091226]
- Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, NHLBI GO Exome Sequencing Project—ESP Lung Project Team, Christiani DC, Wurfel MM, Lin X. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012; 91:224–237. [PubMed: 22863193]
- Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008; 83:311–321. [PubMed: 18691683]
- Lin D-Y, Tang Z-Z. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet.* 2011; 89:354–367. [PubMed: 21885029]
- Madsen BE, Browning SR. A group-wise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Martin ER, Monks SA, Warren LL, Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet.* 2000; 67:146–154. [PubMed: 10825280]
- Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics.* 2012; 44:243–246. [PubMed: 22306651]
- McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JP, Hirschhorn JN. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet.* 2008; 9:356–369. [PubMed: 18398418]
- Morgenthaler S, Thilly WG. A strategy to discover genes that carry multiallelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007; 615:28–56. [PubMed: 17101154]
- Neale BM, Rivas MA, Voight BF, Altshuler D, Devlin B, Orho-Melander M, Kathiresan S, Purcell SM, Roeder K, Daly MJ. Testing for an unusual distribution of rare variants. *PLoS Genetics.* 2011; 7:e1001322. [PubMed: 21408211]
- Ouallacha K, Dastani Z, Li R, Cingolani PE, Spector TD, Hammond CJ, Richards JB, Ciampi A, Greenwood CMT. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epi.* 2013; 37:366–376.
- Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, Wei LJ, Sunyaev SR. Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet.* 2010; 86:832–838. [PubMed: 20471002]
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. PCs analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 2006; 38:904–909. [PubMed: 16862161]

- Rabinowitz D, Laird N. A unified approach to adjusting association tests for population admixture with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered.* 2000; 50:211–223. [PubMed: 10782012]
- Reich DE, Goldstein DB. Detecting associations in case-control studies while correcting for population stratification. *Genetic Epidemiol.* 2001; 20:4–16.
- Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epi.* 1996; 13:423–449.
- Schaid DJ, McDonnell SK, Sinnwell JP, Thibodeau SN. Multiple genetic variant association testing by collapsing and Kernel methods with pedigree or Population Structured Data. *Genet Epidemiol.* 2013; 37(5):409–418. [PubMed: 23650101]
- Scheet P, Stephens M. A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet.* 2006; 78(4): 629–644. [PubMed: 16532393]
- Sha Q, Wang X, Wang X, Zhang S. Detecting association of rare and common variants by testing an optimally weighted combination of variants. *Genet Epidemiol.* 2012; 36:561–571. [PubMed: 22714994]
- Sha Q, Wang S, Zhang S. Adaptive clustering and adaptive weighting methods to detect disease associated rare variants. *Eu J Hum Genet.* 2013; 21(3):332–7.
- Spielman RS, McGinnis RE, Ewens WJ. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet.* 1993; 52:506–516. [PubMed: 8447318]
- Wu M, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare variant association testing for sequencing data using the sequence kernel association test (SKAT). *Am J Hum Genet.* 2011; 89:82–93. [PubMed: 21737059]
- Yi N, Zhi D. Bayesian analysis of rare variants in genetic association studies. *Genet Epi.* 2011; 35:57–69.
- ZawisTOWski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zollner S. Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. *Am J Hum Genet.* 2010; 87:604–617. [PubMed: 21070896]
- Zhang S, Zhu X, Zhao H. On a semi-parameter test to detect associations between quantitative trait and candidate genes using unrelated individuals. *Genetic Epi.* 2003; 24:45–56.
- Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury P, Yu J, Arnett DK, Ordovas JM, Buckler ES. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet.* 2010; 42:355–360. [PubMed: 20208535]
- Zhang Z, Zhang S, Sha Q. A multi-marker test based on family data in genome-wide association study. *BMC Genetics.* 2007; 8:65. [PubMed: 17894890]
- Zhu X, Feng T, Li Y, Lu Q, Elston RC. Detecting rare variants for complex traits using family and unrelated data. *Genet Epidemiol.* 2010; 34:171–187. [PubMed: 19847924]
- Zhu X, Zhang S, Zhao H, Cooper RS. Association mapping, using a mixture model for complex traits. *Genet Epidemiol.* 2002; 23:181–196. [PubMed: 12214310]

Figure 1.

Power comparisons of family-based designs with different numbers of affected children per family and population-based designs. “Same individuals” means that all designs have the same number of individuals. “Same affected individuals” means that all designs have the same number of affected individuals. TOW and WS are based on unrelated individuals. The sample size for TOW and WS is 1000 cases and 1000 controls. TOW-PAC and WS-PAC are based on parents and affected children. With k affected children per family, the sample size (number of families) for TOW-PAC and WS-PAC is $2000/(k+2)$ for the case of “Same individuals” and is $1000/k$ for the case of “Same affected individuals”. pp represents for the percentage of protective variants. The total heritability is 0.03 and the percentage of causal variants is 0.1.

Figure 2.

Power comparisons of the four tests for power as a function of the percentage of protective variants in a homogeneous population. $pcau$ represents for the percentage of causal variants. $herit$ represents for the total heritability of all causal variants. Sample size is 500 families with two children in each family.

Figure 3.

Power comparisons of the four tests for power as a function of the percentage of causal variants in a homogeneous population. pp represents for the percentage of protective variants. $herit$ represents for the total heritability of all causal variants. Sample size is 500 families with two children in each family.

Table 1

Type-I error rates in a homogeneous population.

Gene	alpha	TOW-PAC	WS-PAC	MPDT	SINGLE	TOW
1	0.05	0.0478	0.0476	0.0483	0.0519	0.0516
	0.01	0.0095	0.0099	0.0078	0.0082	0.0099
	0.001	0.0012	0.0007	0.0008	0.0006	0.0012
2	0.05	0.0465	0.0518	0.0474	0.0469	0.0463
	0.01	0.009	0.0099	0.0088	0.0081	0.0098
	0.001	0.0007	0.0006	0.0012	0.0007	0.0016
3	0.05	0.0513	0.0502	0.0513	0.0483	0.0534
	0.01	0.0111	0.0092	0.0101	0.0097	0.0081
	0.001	0.0013	0.0007	0.0007	0.001	0.0015
4	0.05	0.0522	0.0492	0.0505	0.054	0.0471
	0.01	0.0117	0.0114	0.0099	0.0108	0.0093
	0.001	0.0013	0.0012	0.0014	0.0013	0.0006
Sgene	0.05	0.054	0.0489	0.0516	0.0518	0.0539
	0.01	0.0113	0.0093	0.0107	0.0101	0.0106
	0.001	0.0015	0.001	0.0012	0.0012	0.0012

alpha is the nominal significance level.

Table 2

Type-I error rates in a structured population with two subpopulations.

gene	alpha	TOW-PAC	WS-PAC	MPDT	SINGLE	TOW
1	0.05	0.0498	0.0474	0.0492	0.049	0.3098
	0.01	0.0087	0.009	0.0087	0.0095	0.1012
	0.001	0.0008	0.0005	0.0007	0.0005	0.0178
2	0.05	0.0526	0.0488	0.0537	0.0531	0.279
	0.01	0.0113	0.011	0.0103	0.0114	0.0954
	0.001	0.0009	0.0004	0.0014	0.0012	0.0208
3	0.05	0.0476	0.0501	0.0464	0.0525	0.1908
	0.01	0.01	0.0088	0.009	0.0087	0.052
	0.001	0.0014	0.001	0.0013	0.0013	0.008
4	0.05	0.0471	0.0534	0.0513	0.0481	0.1192
	0.01	0.0093	0.0094	0.01	0.0109	0.0254
	0.001	0.0013	0.0015	0.0009	0.0008	0.0026
Sgene	0.05	0.0517	0.0521	0.0536	0.0539	0.317
	0.01	0.0101	0.0113	0.0125	0.0111	0.1212
	0.001	0.0004	0.0009	0.0009	0.0004	0.0306

alpha is the nominal significance level.

Table 3

Type-I error rates in a spatially structured population

Model	L	alpha	TOW-PAC	WS-PAC	MPDT	SINGLE	TOW
1	5	0.05	0.0462	0.0451	0.0476	0.0509	0.3726
		0.01	0.0091	0.009	0.0099	0.0093	0.1897
		0.001	0.0004	0.0012	0.001	0.0009	0.0689
	10	0.05	0.0502	0.0483	0.0503	0.0540	0.5301
		0.01	0.0115	0.0085	0.0120	0.0120	0.3158
2	5	0.001	0.0012	0.0009	0.0012	0.0011	0.1324
		0.05	0.0472	0.0493	0.0457	0.0491	0.3747
		0.01	0.011	0.0093	0.0092	0.0094	0.1942
	10	0.001	0.0013	0.0011	0.0005	0.0013	0.0736
		0.05	0.0486	0.0476	0.0487	0.0519	0.5367
	10	0.01	0.0104	0.0087	0.0093	0.0113	0.3342
		0.001	0.0013	0.0011	0.0011	0.0011	0.1423

alpha is the nominal significance level.