Published in final edited form as:

Ann Hum Genet. 2013 November ; 77(6): 524–534. doi:10.1111/ahg.12038.

Detecting Association of Rare Variants by Testing an Optimally Weighted Combination of Variants for Quantitative Traits in General Families

Shurong Fang, Shuanglin Zhang, and Qiuying Sha

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

Shurong Fang: shurongf@mtu.edu; Shuanglin Zhang: shuzhang@mtu.edu; Qiuying Sha: qsha@mtu.edu

Abstract

Although next-generation sequencing technology allows sequencing the whole genome of large groups of individuals, the development of powerful statistical methods for rare variant association studies is still underway. Even though many statistical methods have been developed for mapping rare variants, most of these methods are for unrelated individuals only, whereas family data have been shown to improve power to detect rare variants. The majority of the existing methods for unrelated individuals are essentially testing the effect of a weighted combination of variants with different weighting schemes. The performance of these methods depends on the weights being used. Recently, Sha et al. [Genet Epidemiol, 2012] proposed a test for Testing the effect of an Optimally Weighted combination of variants (TOW) for unrelated individuals. In this article, we extend our previously developed TOW for unrelated individuals to family-based data and propose a novel test for Testing the effect of an Optimally Weighted combination of variants for Family-based designs (TOW-F). The optimal weights are analytically derived. The results of extensive simulation studies show that TOW-F is robust to population stratification in a wide range of population structures, is robust to the direction and magnitude of the effects of causal variants, and is relatively robust to the percentage of neutral variants.

Keywords

rare variants; general families; association studies; quantitative traits; population stratification

Introduction

Both studies of the genetic architecture of several diseases and simulation studies show that complex diseases are caused by both common and rare variants [Pritchard, 2001; Pritchard and Cox, 2002; Walsh and King, 2007; Stratton and Rahman, 2008; Bodmer and Bonilia, 2008; Ng et al., 2009; Teer and Mullikin, 2010]. Next-generation sequencing technology allows sequencing of the whole genome of large groups of individuals, and thus makes directly testing rare variants feasible [Andre's et al., 2007; Metzker, 2010]. Recently, several statistical methods for detecting associations of rare variants have been developed, including

the cohort allelic sums test [Morgenthaler and Thilly, 2007], the combined multivariate and collapsing method [Li and Leal, 2008], the weighted sum statistic [Madsen and Browning, 2009], the variable minor allele frequency threshold method [Price et al., 2010], the cumulative minor-allele test [ZawisTOWski et al., 2010], the adaptive sum test [Han and Pan, 2010], and the step-up method [Hoffmann et al., 2010], among others. These methods for rare variant association studies are essentially testing the effect of a weighted combination of variants. Thus, choosing appropriate weights is critical to the performance of these methods. More recently, Sha et al. [2012] proposed a test for Testing the effect of an Optimally Weighted combination of variants (TOW). The optimal weights are analytically derived and can be calculated from sampled genotypes and phenotypes. Based on the optimal weights, TOW tests the effect of a weighted combination of variants.

All of the aforementioned methods are for unrelated individual. For any type of study design, the statistical power will be improved if rare variants can be enriched in the samples. If one parent has a copy of a rare allele, half of the offspring are expected to carry it, and hence, variants that are rare in the general population could be very common in certain families [Shi and Rao, 2011]. Therefore, family-based designs may play an important role in rare variant association studies. Because of the importance of family-based designs in rare variant association studies, several family-based rare variant association methods for quantitative traits [Fang et al., 2012; Liu and Leal, 2012] and for qualitative traits [Zhu et al., 2010; Feng et al., 2011; Zhu and Xiong, 2012] have been developed recently.

In this article, we extend our previously proposed TOW for unrelated individuals to TOW-F, TOW for Family-based data. We use extensive simulation studies to evaluate the performance of TOW-F. Our simulation results show that TOW-F can control for population stratification in a wide range of population structures including spatially structured populations, in which the commonly used methods such as genomic control (GC) approach [Devlin and Roeder, 1999], principal component (PC) approach [Zhang et al., 2003; Price et al., 2006], and mixed linear model (MLM) approach [Kang et al., 2010] to control for population stratification in common variant associations are not necessarily controlling for population stratification in rare variant associations [Mathieson and McVean, 2012]. TOW-F is robust to the directions of effect of causal variants and is relatively robust to the percentage of neutral variants. Comparing to TOW-F, the two existing family-based rare variant association methods for quantitative traits [Fang et al., 2012; Liu and Leal, 2012] may subject to bias due to population stratification.

Method

For general families, we first define informative individuals. A family member is defined as informative if he/she satisfies the following conditions: (1) he/she has both phenotypes and genotypes, and (2) parental genotypes are available or at least one sibling's genotypes are available. Consider a sample of n general families with n_i informative individuals in the i^{th} family. Suppose that each individual has been genotyped at K variants in a genomic region. Let y_{ij} and x_{ijk} denote the trait value of the j^{th} informative individual in the i^{th} family and genotypic score at the k^{th} variant of the j^{th} informative individual in the i^{th} family. The genotypic score x_{ijk} is the number of copies of the minor allele at the k^{th} variant site. Let

$x_{ij} = \sum_{k=1}^K w_k x_{ijk}$, where the weight function $w = (w_1, \dots, w_K)^T$ is determined later according to some optimal criteria.

We decompose the genotype score x_{ij} into between-family component $x_{i\cdot}$ and within-family component $x_{wij} = x_{ij} - x_{i\cdot}$, where $x_{i\cdot}$ is the average of parental genotypic scores if parental genotypes are available; $x_{i\cdot}$ is the average of sibship's genotypic scores otherwise. Assume a linear model

$$y_{ij} = \alpha + \beta x_{wij} + \delta_{ij} + e_{ij} \text{ or the matrix form } y = \alpha J + X\beta + \delta + e, \quad (1)$$

where $y = (y_{11}, \dots, y_{1n1}, \dots, y_{n1}, \dots, y_{nnn})^T$, $x = (x_{w11}, \dots, x_{w1n1}, \dots, x_{wn1}, \dots, x_{wnnn})^T$, J is a vector with elements 1, δ is a vector for the random effects of familial correlation, and e is a vector for the error. The vector of error e and the random effects δ are assumed normally

distributed and uncorrelated with each other: $\delta \sim N(0, \sigma_G^2 \Phi)$, $e \sim N(0, \sigma_E^2 I)$, where Φ is twice the kinship matrix obtained from family information only, I is the identity matrix, and σ_G^2, σ_E^2 are the corresponding variance component parameters. Let $\sigma^2 = \sigma_E^2$ and $\sigma_C^2 = \sigma_G^2 / \sigma_E^2$. We use the maximum likelihood estimate (MLE) of σ_C^2 under null hypothesis $H_0: \beta = 0$ as an estimate of σ_C^2 and impute the estimated value of σ_C^2 into model (1). Let $\hat{\sigma}_C^2$ denote the MLE under null hypothesis of σ_C^2 , and let $\Sigma = \hat{\sigma}_C^2 \Phi + I$. After imputing the estimated value of σ_C^2 , model (1) becomes

$$y = \alpha J + x\beta + \delta, \quad (2)$$

Where $\varepsilon = \delta + e \sim N(0, \sigma^2 \Sigma)$. Further more, if we let $y_T = \Sigma^{-1/2} y$, $x_T = \Sigma^{-1/2} x$, and $Z = \Sigma^{-1/2} J$, then model (2) is equivalent to

$$y_T = \alpha Z + x_T \beta + \delta_T, \quad (3)$$

where $\delta_T \sim N(0, \sigma^2 I)$.

We have proven in the Appendix that the score test statistic to test the null hypothesis $H_0: \beta = 0$ under model (3) is the same as the score test statistic under the following model

$$y^{\hat{a}} = x^{\hat{a}} \beta + \delta^{\hat{a}}, \quad (4)$$

where $\delta^{\hat{a}} \sim N(0, 1)$, $y^{\hat{a}}$ are the residuals under model $y_T = \alpha Z + \delta_T$, and $x^{\hat{a}}$ are the residuals under model $x_T = \alpha Z + \delta_T$, and the score test statistic under model (3) or model (4) is given by

$$T=U^2/V, \quad (5)$$

where $U=\sum_{i=1}^N x_i^{\hat{a}} y_i^{\hat{a}}$, $V=\hat{\sigma}^2 \sum_{i=1}^N x_i^{\hat{a}2}$, $\hat{\sigma}^2=\frac{1}{N} \sum_{i=1}^N y_i^{\hat{a}2}$, $N=\sum_{i=1}^n n_i$, $x^{\hat{a}}=(x_1^{\hat{a}}, \dots, x_N^{\hat{a}})^T$ and $y^{\hat{a}}=(y_1^{\hat{a}}, \dots, y_N^{\hat{a}})^T$.

Let $X_k=(x_{w11k}, \dots, x_{w1n1k}, \dots, x_{wn1k}, \dots, x_{wnn1k})^T$, $X_{Tk}=\Sigma^{-1/2}X_k$,

$X_{Tk}=\sum^{-\frac{1}{2}}X_k$, $X_k^{\hat{a}}(X_k^{\hat{a}}=X_k-\bar{X}_k)$ are the residuals under model $X_{Tk}=aZ+\delta T$,

$X^{\hat{a}}=(X_1^{\hat{a}}, \dots, X_k^{\hat{a}})$, and $u^T=y^{\hat{a}T}X^{\hat{a}}$. With these notations, we have $x^{\hat{a}}=X^{\hat{a}}w$ and the score

test statistic $T=\frac{w^T u u^T w}{\hat{\sigma}^2 w^T A w}$, where $A=X^{\hat{a}T}X^{\hat{a}}$. One potential problem with the score test T is that for high-dimensional data, it will be problematic to use A to estimate the large covariance matrix. Following Pan [2009] and Sha et al. [2012], we replace A by $A_0=\text{diag}(A)$. Then, we have a test statistic

$$T_0(w)=\frac{w^T u u^T w}{\hat{\sigma}^2 w^T A_0 w}.$$

As a function of w , $T_0(w)$ reaches its maximum when $w=w^o=A_0^{-1}u$ and the maximum value of $T_0(w)$ is $u^T A_0^{-1}u/\hat{\sigma}^2$. We define the statistic to detect association of rare variants by

Testing an Optimally Weighted combination of variants, $\sum_{k=1}^K w_k^0 x_{ijk}$, in Family-based designs (TOW-F) as

$$T_{TOW-F}=u^T A_0^{-1}u=\sum_{k=1}^K \frac{u_k^2}{X_k^{\hat{a}T}X_k^{\hat{a}}}.$$

A permutation procedure is used to evaluate the p-values of TOW-F. Denote $y^{\hat{a}}$ as

$(y_1^{\hat{a}}, \dots, y_N^{\hat{a}})^T$. In each permutation, we randomly shuffle $y_1^{\hat{a}}, \dots, y_N^{\hat{a}}$. By using permutation,

T_{TOW-F} is equivalent to $\sum_{k=1}^K \frac{u_k^2}{\hat{\sigma}^2 X_k^{\hat{a}T}X_k^{\hat{a}}}$.

Comparison of Methods

To evaluate the performance of the proposed approach, we compare the power of TOW-F with that of the following four methods. (a) Weighted Sum Family-Based Association Test (WS-FBAT): in this method, the test statistic to test for the regional association is T given by

equation (5) based on the genotypic score $x_{ij}^{WS}=\sum_{k=1}^K w_k x_{ijk}$, where w_k is the weight of the

k^{th} variant suggested by Madsen and Browning [2009]. WS-FBAT implicitly assumes that all the rare variants are causal and the directions of effects are all the same. (b) Tmax with statistic $T_{\max} = \max\{T_k\}$, where T_k is the score test statistic at the k^{th} variant given by equation (5). Tmax assumes that there is only one causal variant. (c) Adaptive Weighting FBAT (AW-FBAT) proposed by Fang et al. [2012]. In this method, between-family test statistic T_{BFT} (not depend on parental phenotypes) is used to calculate the adaptive weights. The genotypic score across all variants for the j^{th} child in the i^{th} family is calculated as

$$x_{ij}^{AW} = \sum_{k=1}^K T_{AdBFT,k} w_k x_{ijk}, \text{ where } w_k \text{ is the weight of the } k^{th} \text{ variant suggested by Madsen and Browning [2009].}$$

Then the statistic of AW-FBAT is the score test statistic given by equation (5) based on x_{ij}^{AW} . (d) TOW-F1 that is the same as TOW-F but uses x_{ij} to replace x_{wij} . AW-FBAT, TOW-F1, and our proposed TOW-F assume that there are multiple causal variants and the directions of effects of causal variants may be different.

Simulation

The empirical Mini-Exome genotype data provided by the genetic analysis workshop 17 (GAW17) is used for simulation studies to evaluate the performance of TOW-F. The data contain genotypes of 697 unrelated individuals on 3,205 genes. We consider genes ELAVL4, MSH4, PDE4B, and ADAMTS4 with 10, 20, 30, and 40 variants, respectively. In this paper, the four genes are merged to a super gene with 100 variants. In our simulation studies, we generate genotypes based on the genotypes of 697 individuals in the super gene. We generate genotypes based on the super gene because the distributions of MAFs in the 100 variants in the super gene and in the 24,487 variants in all the 3,205 genes are very similar (Figure S1). Among the 100 variants, there are 83 rare variants ($MAF < 0.01$) and 17 common variants. Among 83 rare variants, there are 54 variants with $MAF 1/1394$, 13 variants with $MAF 2/1394$, 7 variants with $MAF 3/1394$, 4 variants with $MAF 4/1394$, 2 variants with $MAF 5/1394$, 2 variants with $MAF 8/1394$, and 1 variant with $MAF 10/1394$. 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. In all simulation scenarios, we consider two family structures. The first one is a nuclear family with two children; the second one is a three generation pedigree as given in Figure 1. This pedigree structure is one of the Centre d'Etude du Polymorphisme Humans (CEPH) pedigrees (Morley et al., 2004).

Simulation Set 1: A Homogeneous Population

The haplotype frequencies at the super gene were calculated using the program fastPHASE [Scheet and Stephens, 2006]. To obtain the genotypes of each funder, we generate two haplotypes according to the haplotype frequencies. Then the genotypes of other members are generated from parental haplotypes by random transmission. Under the null hypothesis, for parents and m children, let $Y_1 = (y_F, y_M)$ and $Y_2 = (y_1, y_2, \dots, y_m)$ denote the trait values of the parents and the m children, respectively. Assume that (Y_1, Y_2) follows a multivariate normal distribution with a mean vector of zero and variance-covariance matrix of

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \text{ where } \Sigma_{11} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \Sigma_{12} = \Sigma_{21}^T = \begin{pmatrix} \rho & \cdots & \rho \\ \rho & \cdots & \rho \end{pmatrix}, \text{ and } \Sigma_{22} = \begin{pmatrix} 1 & \cdots & \rho \\ & \ddots & \\ \rho & \cdots & 1 \end{pmatrix}.$$

This variance-covariance matrix indicates that the parents in each family are independent, and the correlation coefficient between a parent and a child or between two children is constant, ρ . The conditional distribution of $Y_2=(y_1, \dots, y_m)$ given parental trait values

$Y_1=(y_F, \dots, y_M)$ is a normal distribution with mean vector of $\mu_c = \sum_{21} \sum_{11}^{-1} Y_1$ and variance-covariance matrix of $\Sigma_c = \Sigma_{22} - \sum_{21} \sum_{11}^{-1} \sum_{12}$. To generate trait values of all individuals in a pedigree, we first generate the trait value of each founder by using a standard normal distribution. The trait values of other members can be generated by a normal distribution with a mean vector of μ_c and variance-covariance matrix of Σ_c , given the trait values of their parents.

Under the alternative hypothesis, we choose n_{cau} rare variants ($MAF < 1\%$) from Sgene as analyzed causal variants. The value of n_{cau} is determined by the percentage of neutral variants. Let $prot$ denote the percentage of protective variants, then the number of protective variants and the number of risk variants are $n_p = n_{cau} \cdot prot$ and $n = n_{cau} \cdot (1 - prot)$, respectively. For the j^{th} member in the j^{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. Furthermore, we assume that there are 100 unanalyzed causal variants. Let X_{ijl} denote the genotypic scores of the l^{th} unanalyzed causal variant of the j^{th} member in the i^{th} family.

Assume that all analyzed causal variants have the same heritability and all unanalyzed causal variants also have the same heritability. Then the disease model is given by

$$y_{ij} = \sum_{k_r=1}^{n_r} \beta_{k_r} x_{ijk_r} - \sum_{k_p=1}^{n_p} \beta_{k_p} x_{ijk_p} + \sum_{l=1}^{100} \beta_l X_{ijl} + \varepsilon_{ij}, \text{ where } \beta_{k_r}, \beta_{k_p}, \text{ and } \beta_l \text{ are coefficients and their values depend on the total heritability and the heritability of analyzed causal variants, and } \varepsilon_{ij} \text{ is the trait value under the null hypothesis. In this study, we assume that the total heritability is 0.2.}$$

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 the super gene for each subpopulation. Within each subpopulation, we obtain the genotypes in the same way as that in the Simulation Set 1. Under null hypothesis, the trait values are generated similar to that in the Simulation Set 1, except that in the structured population, parental trait values in subpopulation 1 are from the standard normal distribution and those in

subpopulation 2 are from a normal distribution with a mean of μ ($\mu=0.3$ in this study) and variance of 1. 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 that in the Simulation Set 1. In our simulation studies, we choose half individuals from subpopulation 1 and half from subpopulation 2.

Simulation Set 3: A Spatially Structured Population

We generate genotypes and phenotypes under a spatially structured population using models described in Mathieson and McVean [2012]. To generate family-based genotypes, we first generate genotypes of founders and then generate genotypes of other members by randomly transmitting parental haplotypes. We follow Mathieson and McVean [2012] to generate genotypes of founders. Briefly, we divide the space into $K \times K$ grid squares. We generate genotypes by starting with a number of individuals and their locations on the grid and assuming that husband and wife in founders are in the same grid and all founders are in nearby grids. We work 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 founders. Given trait values of founders, trait values of other members can be generated by using the method described in Simulation Set 1. Let n_f denote the number of founders in n general families. Let $\phi: [1, n_f] \rightarrow [1, K] \times [1, K]$ be a function that mapped each founder to the grid square from which they originated. Then, for the i^{th} founder, we generate the trait values of the i^{th} founder by $y_i = \beta R_{\phi(i)} + \varepsilon_i$, where $R_{k,j}$ is the nongenetic risk in grid square k, j , ε_i is a standard normal random number, 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 $k_0 \leq k \leq k_0 + 3$ and $j_0 \leq j \leq j_0 + 3$; $R_{k,j} = 0$ 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 $\beta = 2$.

Table 1, we give a summary of our simulation studies.

Results

In simulation studies, we perform 1,000 permutations to calculate p-values. The type I error rates are evaluated using 1,000 replicated samples. The power is evaluated using 200 replicated samples. In simulations under nuclear families, we consider five tests: TOW-F, WS-FBAT, Tmax, AW-FBAT, and TOW-F1. In simulations under pedigrees, we do not consider AW-FBAT because AW-FBAT is only applicable to nuclear families. For 1,000 replications, the 95% confidence interval of type I error rates is (0.036, 0.064) at the nominal level of 0.05. The estimated type I error rates of all the tests based on the empirical Mini-Exome genotype data provided by GAW17 are listed in Table 2. As shown by Table 2, in a homogeneous population, the estimated type I error rates of all the tests are within the 95%

confidence interval. In a structured population with two subpopulations, the estimated type I error rates of four tests that depend on within-family information only are within the 95% confidence interval, while TOW-F1 that depends on both within-family and between-family information has inflated type I error rates. The estimated type I error rates of all the tests in a spatially structured population are listed in Table 3. As shown by Table 3, the estimated type I error rates of four tests that depend on within-family information only are within the 95% confidence interval. The estimated type I error rates of TOW-F1 are all higher than the upper bound of the 95% confidence interval, which indicates that TOW-F1 has inflated type I error rates.

In power comparisons, we consider three different scenarios: power as a function of heritability, power as a function of the percentage of neutral variants, and power as a function of the percentage of protective variants. In each scenario, we consider different family structures (nuclear family and pedigree) and different population structures (homogeneous and structured populations). Simulation results show that, in all the three scenarios, TOW-F1 is the most powerful one in homogeneous populations because TOW-F1 uses both within-family and between-family information and other tests only use within-family information. In structured populations, we do not include TOW-F1 because TOW-F1 has inflated type I error rates under structured populations. In the following discussion, we only compare the powers of the four tests that can control for population stratification (three tests in simulations under pedigrees).

In the first scenario, we assume that there are 90% neutral variants and all causal variants are risk variants. Figure 2 displays the power comparison for this scenario. It is reasonable that as the heritability rises, the power of each method increases. As shown in Figure 2, in all the cases considered in this scenario, the proposed TOW-F is the most powerful test; WS-FBAT is the least powerful test; Tmax and AW-FBAT have similar powers, and the powers of Tmax and AW-FBAT are in the middle.

In the second scenario, we assume that heritability is 0.02 and all causal variants are risk variants. Power comparisons in this scenario are summarized in Figure 3. This figure shows that, in the case of the percentage of neutral variants being small, WS-FBAT is the most powerful test and, in the case of the percentage of neutral variants being very large, Tmax is the most powerful test. In other cases, TOW-F is the most powerful test. With the increase of the percentage of neutral variants, the powers of TOW-F, AW-FBAT, and WS-FBAT decrease and the power of WS-FBAT decreases much faster than these of TOW-F and AW-FBAT, while the power of Tmax increases. The reason for the powers of TOW-F, AW-FBAT, and WS-FBAT decreasing with the increase of the percentage of neutral variants is that when the percentage of neutral variants increases, the noise to signal ratio increases and thus the powers decrease. The reason for the power of Tmax increasing with the increase of the percentage of neutral variants is that when the percentage of neutral variants increases or the number of causal variants decreases, the heritability for each causal variant increases under the assumption that the total heritability is unchanged and thus the power of Tmax increases.

We compare powers as a function of the percentage of protective variants in the third scenario. In this scenario, we also assume that heritability is 0.02 and there are 90% neutral variants. Power comparisons under this scenario are shown in Figure 4. As shown by Figure 4, in all the cases considered in this scenario, the proposed TOW-F is the most powerful test, WS-FBAT is the least powerful test, and the powers of Tmax and AW-FBAT are in the middle. The powers of TOW-F, Tmax, and AW-FBAT are robust to the percentage of protective variants, while the power of WS-FBAT decreases rapidly as the percentage of protective variants increases from 0% to 50%.

In conclusion, TOW-F1 is the most powerful one in all the cases but it cannot control for population stratification. Among the four tests that can control for population stratification, except in the case of the percentage of neutral variants being small or very large, TOW-F is the most powerful test, and in the case of the percentage of neutral variants being small or very large, TOW-F is the second powerful test. TOW-F is robust to directions of effects of causal variants and is relatively robust to the percentage of neutral variants. With increase of the percentage of neutral variants, the power of WS-FBAT decreases, while the power of Tmax increases.

Discussion

Most of recently developed statistical methods for detecting rare variant association are essentially testing the effect of a weighted combination of variants with different weighting schemes. More recently, Sha et al. [2012] derived the optimal weights and proposed TOW to test the effect of an optimally weighted combination of variants. However, all these methods are designed for unrelated individuals, whereas family data have been shown to improve power [Feng et al., 2011; Zhu et al., 2010]. In this article, we extend TOW for unrelated individuals to TOW-F for family-based designs. Our results show that in most cases, TOW-F is the most powerful test among the four tests that can control for population stratification. Results also show that TOW-F is robust to population stratification, is robust to the directions of the effect of causal variants, and is relatively robust to the percentage of neutral variants. TOW-F also has limitations. Its limitations include (1) it does not use between-family information, thus it may lose power in homogenous populations; (2) it uses permutation test to evaluate p-values, thus it is computationally extensive for genome-wide studies. Furthermore, our proposed TOW-F assumes that there are multiple causal variants and the directions of effects of causal variants may be different. In the case of only one causal variant or very few causal variants, TOW-F may be not as powerful as Tmax. If the directions of effects of causal variants are the same, TOW-F may be not as powerful as WS-FBAT.

In association studies based on unrelated individuals, it has been long recognized that population stratification can seriously confound association results [Knowler et al., 1988; Lander and Schork, 1994]. In common variant association studies, several methods have been developed to control for population stratification. These methods include GC approach [Devlin and Roeder, 1999; Devlin et al., 2001, and Reich and Goldstein, 2001], PC approach [Chen et al., 2003; Zhang et al., 2003; Zhu et al., 2002; Price et al., 2006; Bauchet et al., 2007], and MLM approach [Kang et al., 2010; Zhang et al., 2010]. Recently, several

methods for family-based data that use within-family as well as between-family information and use PC approach to control for population stratification have been proposed for common variant association studies [Zhu et al., 2008; Zhang et al., 2009; Sha et al., 2011] and for rare variant association studies [Fang et al., 2012]. 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 GC, PC, and MLM approaches to control for population stratification in population-based common variant associations cannot control for population stratification in population-based rare variant associations in some population structures. Since PC approach cannot control for population stratification in population-based rare variant associations in some population structures, we cannot expect that using PCs of genotypes at genomic markers to adjust population effects in genotypes of every family member can control for population stratification in family-based rare variant association studies. Thus, we only use within-family information in the proposed TOW-F. Simulation results show that, by using within-family information only, TOW-F can control for population stratification even in spatially structured populations in which GC, PC, and MLM approaches cannot control for population stratification [Mathieson and McVean, 2012]. In this paper, we assume that there are no missing genotypes. If there are missing genotypes, we can first impute missing genotypes using genotype imputation methods for family samples [Burdick et al., 2006; Li et al., 2009].

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

In this appendix, we showed that to test the null hypothesis $H_0: \beta = 0$, the score test statistic under model (3) is the same as the score test statistic under model (4) and the score test statistic is given by equation (5).

The log-likelihood function (up to a constant) under model (3) is

$$\log l = -\frac{N}{2} \log \sigma^2 - \frac{1}{2\sigma^2} (y_T - \sigma Z - \beta x_T)^T (y_T - \sigma Z - \beta x_T).$$

Then the first and second derivatives of $\log l$ with respect to α and β are

$$\frac{\partial \log l}{\partial \beta} = \frac{1}{\sigma^2} (y_T - \alpha Z - \beta x_T)^T x_T, \quad \frac{\partial \log l}{\partial \alpha} = \frac{1}{\sigma^2} (y_T - \alpha Z - \beta x_T)^T Z, \quad \frac{\partial^2 \log l}{\partial \beta^2} = -\frac{1}{\sigma^2} x_T^T x_T,$$

$$\frac{\partial^2 \log l}{\partial \alpha^2} = -\frac{1}{\sigma^2} Z^T Z, \text{ and } \frac{\partial^2 \log l}{\partial \alpha \partial \beta} = -\frac{1}{\sigma^2} x_T^T Z.$$

Under the null hypothesis $H_0: \beta = 0$ the maximum likelihood estimates of α and σ^2 are,

$$\hat{\alpha} = \frac{y_T^T Z}{Z^T Z} \text{ and } \hat{\sigma}^2 = \frac{1}{N} (y_T - \hat{\alpha} Z)^T (y_T - \hat{\alpha} Z), \text{ respectively. Then, the score vector is}$$

$$S = \frac{1}{\hat{\sigma}^2} [0, (y_T - \hat{\alpha} Z)^T x_T]^T = \frac{1}{\hat{\sigma}^2} (0, U)^T.$$

The information matrix is

$$I = \frac{1}{\hat{\sigma}^2} \begin{pmatrix} Z^T Z & Z^T x_T \\ Z^T x_T & x_T^T x_T \end{pmatrix}.$$

The inverse of I is given by

$$I^{-1} = \hat{\sigma}^2 \begin{pmatrix} Z^T Z & Z^T x_T \\ Z^T x_T & x_T^T x_T \end{pmatrix}^{-1} = \frac{\hat{\sigma}^2}{Z^T Z x_T^T x_T - (Z^T x_T)^2} \begin{pmatrix} x_T^T x_T & -Z^T x_T \\ -Z^T x_T & Z^T Z \end{pmatrix}.$$

Then, the score test statistic is

$$T = \frac{U^2}{V},$$

where $V = \frac{\hat{\sigma}^2 (Z^T Z x_T^T x_T - (Z^T x_T)^2)}{Z^T Z}$, $\hat{\sigma}^2 = \frac{1}{N} (y_T - \hat{\alpha} Z)^T (y_T - \hat{\alpha} Z)$, $\hat{\alpha} = \frac{y_T^T Z}{Z^T Z}$, and

$U = (y_T - \hat{\alpha} Z)^T x_T$. Let $P = ZZ^T / Z^T Z$, $y^* = (I - P)y_T$, and $x^* = (I - P)x_T$. It is easy to know that $y^{\hat{a}}$ are the residuals under model $y_T = \alpha Z + \epsilon_T$, and $x^{\hat{a}}$ are the residuals under model $x_T = \alpha Z$

+ ϵ_T . Note that $P^2 = P$. We have $\hat{\sigma}^2 = \frac{1}{N} y^{*T} y^*$, $U = y^{*T} x^*$, and $V = \sigma^2 x^{*T} x^*$. It can be shown that T is also the score test statistic under model (4).

References

- Abecasis GR, Cardon LR, Cookson WO. A general test of association for quantitative traits in nuclear families. *Am J Hum Genet.* 2000; 66:279–292. [PubMed: 10631157]
- Andre's A, Clark A, Shimmin L, Boerwinkle E, Sing C, Hixson JE. Understanding the accuracy of statistical haplotype inference with sequence data of known phase. *Genetic Epi.* 2007; 31:659–671.
- Bauchet M, McEvoy B, Pearson LN, Quillen EE, Sarkisian T, Hovhannesian K, Dekka 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]
- Burdick JT, Chen WM, Abecasis GR, Cheung VG. In silico method for inferring genotypes in pedigrees. *Nat Genet.* 2006; 38:1002–4. [PubMed: 16921375]
- 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]
- 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]
- Feng T, Elston R, Zhu X. Detecting rare and common variants for complex traits: sib pair and odds ratio weighted sum statistics (SPWSS, ORWSS). *Genetic Epidemiol.* 2011; 35:398–409.
- 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):e13584.10.1371/journal.pone.0013584 [PubMed: 21072163]
- 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]
- 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]
- Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet.* 2009; 10: 387–406.
- Liu D, Leal S. A unified framework for detecting rare variant quantitative trait associations in pedigree and unrelated individuals via sequence data. *Hum Hered.* 2012; 73:105–122. [PubMed: 22555759]
- Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009; 5:e1000384. [PubMed: 19214210]
- Mathieson I, McVean G. Differential confounding of rare and common variants in spatially structured populations. *Nature Genetics.* 2012; 44:243–246. [PubMed: 22306651]
- Metzker ML. Sequencing technologies – the next generation. *Nature Reviews Genetics.* 2010; 11:31–46.
- 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]
- Morley M, Molony CM, Weber T, Devlin JL, Ewens KG, Spielman RS, Cheung VG. Genetic analysis of genome-wide variation in human gene expression. *Nature.* 2004; 430:743–747. [PubMed: 15269782]
- Ng SB, Turner EH, Robertson PD. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature Letters.* 2009; 461:272–276.
- Pan W. Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol.* 2009; 33:497–507. [PubMed: 19170135]

- 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]
- Pritchard JK. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet.* 2001; 69:124–137. [PubMed: 11404818]
- Pritchard JK, Cox NJ. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet.* 2002; 11:2417–2423. [PubMed: 12351577]
- Reich DE, Goldstein DB. Detecting associations in case-control studies while correcting for population stratification. *Genetic Epidemiology.* 2001; 20:4–16. [PubMed: 11119293]
- 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, Zhang Z, Zhang S. Joint analysis for genome-wide association studies in family-based designs. *PLoS ONE.* 2011; 6(7):e21957.10.1371/journal.pone.0021957 [PubMed: 21799758]
- Shi G, Rao D. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. *Genetic Epidemiol.* 2011; 35:572–579.
- Stratton MR, Rahman N. The emerging landscape of breast cancer susceptibility. *Nat Genet.* 2008; 40:17–22. [PubMed: 18163131]
- Teer JK, Mullikin JC. Exome sequencing: the sweet spot before whole genomes. *Hum Mol Genet.* 2010; 19(R2):R145–51. [PubMed: 20705737]
- Walsh T, King MC. Ten genes for inherited breast cancer. *Cancer Cell.* 2007; 11:103–105. [PubMed: 17292821]
- 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 L, Pei YF, Li J, Papasian CJ, Deng HW. Univariate/multivariate genome-wide association scans using data from families and unrelated samples. *PLoS ONE.* 200910.1371/journal.pone.0006502
- 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]
- 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, Li S, Cooper R, Elston RC. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet.* 2008; 82:352–365. [PubMed: 18252216]
- 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]
- Zhu Y, Xiong M. Family-based association studies for next-generation sequencing. *Am J Hum Genet.* 2012; 90:1028–1045. [PubMed: 22682329]

Figure 1.
The pedigree structure used in the simulation studies.

Figure 2.

Power comparison as a function of the heritability of analyzed causal variants. The colors red, blue, black, green, and purple represent powers of TOW-F, WS-FBAT, Tmax, AW-FBAT, and TOW-F1 respectively. In this scenario, we assume that the percentage of neutral variants is 90% and all causal variants are risk variants. “nuclear families” means that this set of simulations is based on 800 families with 2 children in each family. “pedigrees” means that this set of simulations is based on 200 pedigrees given in figure 1.

Figure 3.

Power comparison as a function of the percentage of neutral variants. The colors red, blue, black, green, and purple represent powers of TOW-F, WS-FBAT, Tmax, AW-FBAT, and TOW-F1 respectively. In this scenario, we assume that the heritability of analyzed causal variants is 0.02 and all causal variants are risk variants. “nuclear families” means that this set of simulations is based on 800 families with 2 children in each family. “pedigrees” means that this set of simulations is based on 200 pedigrees given in figure 1.

Figure 4.

Power comparison as a function of the percentage of protective variants. The colors red, blue, black, green, and purple represent powers of TOW-F, WS-FBAT, Tmax, AW-FBAT, and TOW-F1 respectively. In this scenario, we assume that the heritability of analyzed causal variants is 0.02 and the percentage of neutral variants is 90%. “nuclear families” means that this set of simulations is based on 800 families with 2 children in each family. “pedigrees” means that this set of simulations is based on 200 pedigrees given in figure 1.

Table 1
Summary of simulation studies

Population structure	Family structure	Simulation scenarios
A homogeneous population: this set of simulations is based on the empirical Mini-Exome genotype data provided by GAW17	Nuclear family	Type I error evaluation for different sample sizes and power comparisons for different values of percentages of neutral variants, and different percentages of protective variants
	Pedigree	
A structured population with two subpopulations (Asian and African): this set of simulations is based on the empirical Mini-Exome genotype data provided by GAW17	Nuclear family	
	Pedigree	
A spatially structured population (Mathieson and McVean, 2012)	Nuclear family	Type I error evaluation for different variants in a region
	Pedigree	

Type I error rates of the five tests at the nominal level of 0.05 in a homogeneous population and a structured population with two subpopulations.

Table 2

	Sample size	TOW-F	WS-FBAT	Tmax	AW-FBAT	TOW-F1
Homogeneous	400 nu-families	0.050	0.054	0.051	0.050	0.049
	600 nu-families	0.039	0.051	0.046	0.056	0.046
	800 nu-families	0.046	0.049	0.039	0.059	0.056
	200 pedigrees	0.058	0.063	0.052		0.037
Structured	400 nu-families	0.044	0.053	0.048	0.062	0.701
	600 nu-families	0.042	0.045	0.049	0.063	0.935
	800 nu-families	0.054	0.047	0.056	0.060	0.858
	200 pedigrees	0.057	0.037	0.038		0.905

Note: “nu-families” means that nuclear families with 2 children in each family. “pedigrees” means pedigrees with the structure given in figure 1.

Table 3
Type I error rates of the five tests at the nominal level of 0.05 in a spatially structured population.

Family Structure	Model	Number of Variants	TOW-F	WS-FBAT	Tmax	AW-FBAT	TOW-F1
Nuclear Families	1	5	0.034	0.047	0.040	0.062	0.080
	1	10	0.049	0.040	0.033	0.063	0.078
	2	5	0.055	0.047	0.048	0.056	0.100
	2	10	0.045	0.057	0.047	0.063	0.103
Pedigrees	1	5	0.055	0.037	0.061		0.093
	1	10	0.056	0.046	0.039		0.094
	2	5	0.057	0.048	0.050		0.120
	2	10	0.062	0.042	0.051		0.100

Note: Simulations of family structure of nuclear families are based on 400 nuclear families with 2 children in each family. Simulations of family structure of pedigrees are based on 200 pedigrees with the structure given in figure 1.