FARVAT: a family-based rare variant association test

Sungkyoung Choi1, Sungyoung Lee1, Sven Cichon2, Markus M. Nöthen2, Christoph Lange3–8, Taesung Park1,9,* and Sungho Won10,*

ABSTRACT

Motivation: Individuals in each family are genetically more homogeneous than unrelated individuals, and family-based designs are often recommended for the analysis of rare variants. However, despite the importance of family-based samples analysis, few statistical methods for rare variant association analysis are available.

Results: In this report, we propose a Family-based Rare Variant Association Test (FARVAT). FARVAT is based on the quasi-likelihood of whole families, and is statistically and computationally efficient for the extended families. FARVAT assumed that families were ascertained with the disease status of family members, and incorporation of the estimated genetic relationship matrix to the proposed method provided robustness under the presence of the population substructure. Depending on the choice of working matrix, our method could be a burden test or a variance component test, and could be extended to the SKAT-O-type statistic. FARVAT was implemented in C+++, and application of the proposed method to schizophrenia data and simulated data for GAW17 illustrated its practical importance.

Availability: The software calculates various statistics for the analysis of related samples, and it is freely downloadable from http://healthstats.snu.ac.kr/software/farvat.

Received on November 21, 2013; revised on June 30, 2014; accepted on July 17, 2014

1 INTRODUCTION

Advances in genotyping technology have enabled researchers to conduct large-scale genetic analyses, and during the last decade, genome-wide association studies have identified >1000 common genetic loci associated with many phenotypes. However, heritabilities for most phenotypes are only partially explained by these significant findings (Manolio et al., 2009), and relatively small proportions of variance explained by common variants have revealed the importance of association analyses with rare variants (Yang et al., 2011).

Contrary to the analysis of common variants, single genetic association analysis with rare variants is often associated with large false-negative results unless sample sizes or effect sizes are very large. Thus, association analysis with the collapsed genotype scores for a set of rare variants has been suggested (Li and Leal, 2008). For instance, minor alleles for all rare variants in a gene or a region are counted, and the disease status is regressed on minor allele counts (MAC). Alternatively, the collapsed amount of variance inflation for rare variants can be compared between affected and unaffected individuals (Neale et al., 2011; Wu et al., 2011). The former is often called a burden test, while the latter is a variance component test. The burden test is statistically more efficient than variance component methods such as C-alpha (Neale et al., 2011) and SKAT (Wu et al., 2011) if most of the rare alleles have similar effects on the disease. However, if rare variants with deleterious and protective effects are combined, the collapsed genotype scores for affected and unaffected individuals are similar, and genetic association analysis with a burden test becomes inefficient, whereas the variance component method become more robust. Both methods can be combined into robust statistical strategies such as the SKAT-O approach (Lee et al., 2012a), which is statistically efficient in both situations.

However, despite these improvements in statistical methods, the high cost of sequencing still prevents large-scale genome-wide rare variant association studies. The common disease rare variant hypothesis assumes genetic heterogeneity between affected individuals, and selecting genetically homogeneous subjects obviously increases the rate of true-positive findings. In particular, family members are genetically more homogeneous than random samples, and rare variant analysis with extended families can lead to identification of more disease-susceptibility variants (Dering et al., 2011; Manolio et al., 2009). For instance, it has been shown that the enrichment of rare alleles in 100 affected sib pairs can be equal to that of 200 cases–control pairs (Shi and Rao, 2011). Therefore, rare variant association analysis with
carefully ascertained families seems to be an efficient strategy, and the development of statistical methods for family-based samples is necessary.

Recently, Family Based Association Tests (FBAT) statistics (Laird et al., 2000) have been extended for application in rare variant association analysis: the burden test (De et al., 2013) and the variance component test (Ionita-Laza et al., 2013) have been proposed. According to the nature of FBAT, these tests are robust against the population substructure and can be combined with rank-based P-values (Van Steen et al., 2005; Won et al., 2009) based on the between-family component (Lange et al., 2003). He et al. (2014) proposed Rare Variant Extensions of the Transmission Disequilibrium Test (RV-TDT) methods, which were extensions of the TDT (Spielman et al., 1993). FBAT and RV-TDT methods were shown to be robust and powerful for exploration of rare variant association in the population substructure. However, even though robustness against the population substructure can be provided, those approaches do not take into account the parental phenotypes, and power loss can be substantial for extended family designs. Alternatively, studies have proposed the functional principal component analysis (FPCA) and pedigree-based combined multivariate and collapsing statistic (PedCMC) tests (Zhu and Xiong, 2012), which are extended Cochran–Armitage tests for family-based samples. These tests use data from the whole family for rare variant association analysis and are expected to be more efficient than FBAT/TDT-type statistics. However, if the effects of rare variants are proportional to MAC or the protective and deleterious variants are mixed in a gene, these approaches can be less efficient.

In this report, we propose a Family-based Rare Variant Association Test (FARVAT). We provide a burden test and a variance component test for extended families, and these approaches are extended to the SKAT-O-type statistic. The proposed method assumes that families are ascertained based on the disease status of family members, and minor allele frequencies (MAFs) between affected and unaffected individuals are compared. MAFs for each rare variant are estimated with the best linear unbiased estimators (McPeek et al., 2004). FARVAT is implemented with C++ and is computationally efficient for the analysis of rare variants with extended families. With extensive simulations, we compared the proposed methods with existing methods (He et al., 2014; Zhu and Xiong, 2012), and results showed that the proposed methods were the most efficient in the considered scenarios. Application of the proposed method to schizophrenia and GAW17 illustrated its practical value in real analyses.

2 METHODS
2.1 Notations and the disease model
We assumed that there are \( n \) families and \( n_i \) individuals in family \( i \), and the total sample size was denoted by \( N = \sum_{i=1}^{n} n_i \). We assumed that genotype data for \( m \) rare variant loci were available. We let \( y_{ij} \) and \( x_{ij} \) be the phenotype and genotype count of an individual \( j \) in a family \( i \) for rare variant \( k \). If we denoted the disease prevalence by \( q \), \( y_{ij} \) was coded as 1 for affected individuals, \( q \) for individuals with missing phenotype and 0 for unaffected individuals. If genotype frequencies of affected and unaffected individuals are compared to detect genetic associations, the statistical efficiency can be improved by modifying the phenotype (Lange and Laird, 2002; Thornton and McPeek, 2007), and we therefore introduced the so-called offset \( \mu_i \) to set \( d_i = y_{ij} - \mu_i \). The disease prevalence \( q \) has often been used as an offset, and if the disease prevalences in males and females are different, the offset should be chosen separately (Thornton et al., 2012). For randomly selected families, the best linear unbiased predictor (BLUP) from the linear mixed model is known to be linear unbiased estimators (McPeek, 2010). We assume that there are \( A \) common variants, and the coded genotype for common variant is denoted by \( a \), \( x_{jk} \) by \( \sigma_{kl} \), \( \sigma_{kl} \) is estimated with the empirical covariance. In the presence of population substructure, \( \Phi \) should be empirically estimated with common variants available at the genome-wide level instead of using the kinship coefficient between individuals (Thornton and McPeek, 2010). We assume that there are \( A \) common variants, and the coded genotype for common variant is denoted by \( a \) for individual \( j \) in family \( i \) at common variant \( a \). If we let \( p_a \) be the MAF of common variant \( a \), \( x_{jk} \) for \( \Phi \) (Thornton and McPeek, 2010) can be estimated by

\[
\phi_{ij} = \frac{1}{A} \sum_{k=1}^{A} \frac{x_{jk}^2 - 2p_a x_{jk} + 2p_a^2}{2p_a (1-p_a)}, \quad i \neq i' \text{ and } j \neq j'
\]

(4)

2.2 Family-based Rare Variant Association Test
For ascertained samples, the disease status can be assumed to be fixed, and the genotype frequencies between affected and unaffected individuals are usually compared. We let \( I_i \) be the \( w \times 1 \) column vector that consisted of 1 and \( I_i \) be the \( w \times w \) identity matrix. If we denoted an MAF of rare variant \( k \) in unaffected individuals by \( p_u \), we assumed (Thornton and McPeek, 2007) that for a constant \( y_a \)

\[
E(\mathbf{X}|Y) = 2p_a I_n + y_a \mathbf{Y}, \quad \text{var}(\mathbf{X}|Y) = \sigma_{kl} \Phi.
\]

(5)
where \(0 < 2p_0 + \gamma_k < 1\). If we let \(V\) be the working variance-covariance matrix, the score for the quasi-likelihood (Thornton and McPeek, 2007) became
\[
T'V^{-1}(X - \hat{E}(X)).
\]
(6)

Recently, we showed that the approximate optimal efficiency for the analysis of common variants is achieved with \(V = I_\lambda\) (Wong and Lange, 2013). For the choice of the offset in \(T\), BLUP and \(q\) have been suggested for randomly selected samples and ascertained samples, respectively (Thornton and McPeek, 2007; Won and Elston, 2008). \(E(X)\) can be estimated with the following best linear unbiased estimator (McPeek et al., 2004):
\[
\hat{E}(X) = I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda' \Phi^{-1} X.
\]
(7)

Therefore, our score based on the quasi-likelihood became
\[
T' (I_\lambda - I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda' \Phi^{-1} X).
\]
(8)

If we let
\[
H = \Phi - I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda'
\]
and \(\Sigma = \begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1m} \\ \vdots & \ddots & \vdots \\ \sigma_{m1} & \cdots & \sigma_{mm} \end{pmatrix}\),
(9)
we have
\[
\text{var} \left( T' (I_\lambda - I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda' \Phi^{-1} X) \right) = \sigma_{kl} T'HT.
\]
and thus the variance-covariance matrix of the score was
\[
\text{var} \left( T' (X - \hat{E}(X)) \right) = (T'HT)\Sigma.
\]
(10)

Therefore, we have
\[
\frac{1}{\sqrt{T'HT}} T' (I_\lambda - I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda' \Phi^{-1} X) \Sigma^{-1/2} \sim MVN(0, I_\Sigma) \text{ under } H_0.
\]
(12)

For rare variant association analysis, the collapsed amount of either rare alleles or variance inflation between affected and unaffected individuals has been compared (Li and Leal, 2008; Neale et al., 2011; Price et al., 2010; Wu et al., 2011). If we let the weight for variant \(k\) be \(w_k\), the null hypothesis for the former was
\[
H^0_k: w_{1k} Y_1 + \ldots + w_{nk} Y_n = 0,
\]
and for the latter was
\[
H^0_k: w_{1k}^2 Y_1^2 + \ldots + w_{nk}^2 Y_n^2 = 0.
\]
For the choice of \(w_k\), \(w_k = 1\) or \(|p_0(1 - p_0)|^{-1/2}\) are often used. If we denoted the \(m \times m\) diagonal matrix, which consists of \(w_k\), by \(W\), the score test for the burden-type test was
\[
\frac{1}{T'HT} T' (X - \hat{E}(X)) W L_n W (X - \hat{E}(X))^T,
\]
and the score test for the C-alpha-type test was
\[
\frac{1}{T'HT} T' (X - \hat{E}(X)) W L_n W (X - \hat{E}(X))^T.
\]
(16)

Both score tests for rare variant analysis could be generalized to
\[
\frac{1}{T'HT} T' (X - \hat{E}(X)) W R W (X - \hat{E}(X))^T.
\]
(17)

and for a given constant \(c \in [0,1]\), \(S_c\) was denoted by
\[
\frac{1}{T'HT} T' (X - \hat{E}(X)) W ((1 - c) L_n + c I_n) W (X - \hat{E}(X))^T.
\]
(18)

We denoted eigenvalues for \(\Sigma^{1/2} W \Sigma^{1/2}\) by \(\lambda_k\). If we let \(\chi^2_k\)'s be independent chi-square distributions with a single degree of freedom, we have
\[
S_c \sim (I_n W Z W L_n) \chi^2_k \text{ under } H^0_c,
\]
(19)

and
\[
S_0 \sim \sum_{k=1}^{m} \lambda_k \chi^2_k \text{ under } H^0_0.
\]
(20)

The \(P\)-values for \(S_1\) and \(S_0\) were, respectively, denoted by \(\text{FARVAT}^1\) and \(\text{FARVAT}_0\), and in particular, \(\text{FARVAT}_0\) can be calculated with the Davies method (Davies, 1980) or the method described by Liu et al. (Lee et al., 2012b; Liu et al., 2009).

### 2.3 Extension of \(S_1\) and \(S_0\) to the SKAT-O-type statistic

The burden test is known to be efficient if all rare variants have either deleterious or protective effects on disease; otherwise, the C-alpha test is more efficient (Neale et al., 2011). A balanced approach for both scenarios can be achieved by the SKAT-O-type statistic (Lee et al., 2012). For \(c_0 = 0 < c_1 < \ldots < c_t = 1\), we denoted the observed value for \(S_c\) by \(s_c\), and their corresponding \(P\)-values were denoted by \(p_c\). Furthermore, we denoted the \((1 - p)^{\text{th}}\) quantile for \(S_c\) by \(Q_c(p_c)\). If we let
\[
p_{\min} = \min \{ p_{c_0}, p_{c_1}, \ldots, p_{c_t} \},
\]
(21)

our final \(P\)-value was obtained by
\[
1 - P(S_{\min} \leq Q_{\min}(p_{\min})), \ldots, S_{c_t} \leq Q_{c_t}(p_{c_t})).
\]
(22)

The numerical calculation of the final \(P\)-value for the independent samples was derived by Lee et al. (2012), and our final \(P\)-values, denoted by \(\text{FARVAT}_0\), were calculated based on their approach with some modification.

If we let \(Z = \Sigma^{1/2} W \text{ and } Z = Z_{1m} (l_{1m}^2)^{-1}\), the projection matrix onto a space spanned by \(Z\) becomes \(\Pi = Z (Z'Z)^{-1} Z'\). If we let
\[
u = \frac{1}{\sqrt{T'HT}} T' (I_\lambda - I_\lambda (I_\lambda' \Phi^{-1} I_\lambda)^{-1} I_\lambda' \Phi^{-1} X) \Sigma^{-1/2},
\]
(23)

\(u \sim MVN(0, I_\nu)\), and \(S_{\nu}\) becomes
\[
S_{\nu} = u' \Sigma^{1/2} W R W \Sigma^{1/2} u = u' Z Z' u.
\]
(24)

As was shown by Lee et al. (2012), if we let
\[
t(c) = \frac{1 - c}{c} Z Z' Z + c m^2 Z' Z,
\]
(25)

we have
\[
S_c = (1 - c) u' \Pi ZZ' (I_n - \Pi) u + 2(1 - c) u' (I_n - \Pi) ZZ' \Pi u + t(c) u' \Pi u,
\]
(26)

where \(u' (I_n - \Pi) ZZ' (I_n - \Pi) u, u' (I_n - \Pi) ZZ' \Pi u\) and \(u' \Pi u\) are mutually independent. Therefore,
\[
P(S_{\nu} \leq Q_{\nu}(p_{\min})), \ldots, S_{c_t} \leq Q_{c_t}(p_{c_t}(\min)) | u' \Pi u = \eta),
\]
(27)

and the following conditional probability can be numerically calculated, as suggested by Lee et al. (2012):
\[
P(S_{\nu} \leq Q_{\nu}(p_{\min})), \ldots, S_{c_t} \leq Q_{c_t}(p_{c_t}(\min)) | u' \Pi u = \eta),
\]
(28)

### 2.4 The simulation model

In our simulation studies, we considered extended families that consisted of 10 individuals, and extended over three generations (see Supplementary Fig. S1). To generate the genotypes for extended families, haplotypes
were simulated with COSI software (Schaffner et al., 2005), based on the coalescent model, and obtained haplotypes were used for founders’ genotypes. In the coalescent model for COSI, we assumed that the mutation rate was 1.5 × 10−8, and 5000 haplotypes with 50 000 bp were generated. m rare variants in a region or all rare variants for which MAFs were <0.01 were randomly selected, and pairs of haplotypes were randomly chosen with replacement to derive the founders’ genotypes. Under the assumption of no recombination, a haplotype from each founder was randomly selected to construct non-founders’ genotypes under the assumption of Mendelian transmission.

The disease status for each individual was generated with the liability threshold model. The underlying liabilities were defined by summing the polygenic effect, common environmental effect, main genetic effect and random error. The polygenic effect p_i was assumed to be 0, and the polygenic effect, common environmental effect and random errors were generated from the normal distribution with mean 0. Variances for the polygenic effect, common environmental effect and random errors were denoted by \sigma^2_g, \sigma^2_e and \sigma^2_r, respectively, and were assumed to be 1. In this setting, the heritability was 1/3. The polygenic average of maternal and paternal polygenic effects was combined with the proportion of variance explained by rare variants, and their main genetic effects for each individual were the product of \rho_i and the number of disease alleles. If we let \hat{h}^2 be the relative proportion of variance explained by rare variants, \rho_i were sampled from U(1.0, v), and v was calculated by

\[ v = \frac{(\sigma^2_g + \sigma^2_e + \sigma^2_r) h_i^2}{(1 - h_i^2) \sum_{k=1}^{m} \rho_k^2 p_k (1 - p_k) } . \]  

(29)

Under the null hypothesis, \hat{h}^2 was set to 0, and \rho_i became 0. Once the underlying liabilities of main genetic effects, polygenic effects, common environmental effects and random errors were generated, they were transformed to being affected if they were larger than the threshold; otherwise, they were considered as unaffected. The threshold was chosen to preserve the assumed prevalence, and the disease prevalence was assumed to be 0.12. Families with more than two affected grandchildren were used for simulation studies, and sampling was repeated until the given numbers of these families were obtained.

Furthermore, the robustness of the proposed statistic under the presence of the population substructure was evaluated with simulated data. We assumed that there were two subpopulations, and each founder was assigned to one of the two subpopulations with 50% probability. Means of liabilities for main genetic effects, polygenic effects, common environmental effects and random errors were generated, they were transformed to being affected if they were larger than the threshold; otherwise, they were considered as unaffected. The threshold was chosen to preserve the assumed prevalence, and the disease prevalence was assumed to be 0.12. Families with more than two affected grandchildren were used for simulation studies, and sampling was repeated until the given numbers of these families were obtained.

The statistical validity of FARVAT, FARVAT_r and FARVAT_e was evaluated under the absence of population substructure, and the results were compared with PedCMC, FPCA (Zhu and Xiong, 2012) and RV-TDT methods (He et al., 2014). RV-TDT methods consist of \text{BRV.T01}, \text{BRV.Hapo.T01}, \text{CMC.T01}, \text{CMC.Hapo.T01}, \text{VT.BRV.Hapo}, \text{VT.CMC.Hapo} and \text{WSS.Hapo}. We generated 50 and 100 extended families in each replicate, and empirical type I error estimates at the 0.05, 0.01 and 0.001 significance levels were calculated with 50 000 replicates. For the proposed methods, 1 and \[ \rho_i(1 - p_i) \] were considered for \( w_i \), and the kinship coefficients were used to build the correlation matrix \( \Phi \). Rare variants for which MAFs were <0.01 were considered for all statistics. In Supplementary Tables S1 and S2, 30 and 100 rare variants were randomly selected, and in Supplementary Table S3, all rare variants in the 30 kb genetic region were considered. These results showed that the empirical type I error estimates for FARVAT, FARVAT_r and FARVAT_e preserved the nominal significance levels. However, CMC.T01, BRV.Hapo.T01, CMC.Hapo.T01, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo and FPCA were usually conservative, and BRV.T01 and PedCMC seemed to be liberal. For PedCMC, genotype scores of individuals with more than or equal to a single rare allele were considered as 1; otherwise, they were 0. If the large number of rare variants is collapsed, its convergence to the chi-square distribution requires very large sample sizes, and genotype scores for all individuals can be 1 in extreme scenarios. Therefore, we could conclude that PedCMC may not be a good choice when the number of rare variants in a gene is very large.

The statistical efficiency of FARVAT, FARVAT_r and FARVAT_e was evaluated with the simulated data, and results were compared with results from PedCMC, FPCA and RV-TDT methods (He et al., 2014; Zhu and Xiong, 2012). We assumed that the relative proportion of variances explained by rare variants \( \hat{h}^2 \) was 0.05. In each replicate, we assumed that all rare variants had either deleterious or protective effects on disease, and the proportions of rare variants with deleterious effects were assumed to be 1, 0.8, 0.6 and 0.5. The numbers of extended families were assumed to be 50 and 100. MAFs for all rare variants were assumed to be <0.01. Thirty rare variants in Figure 1 and 100 rare variants in Figure 2 were selected, and in Figure 3, all rare variants within 30 kb from the generated 1 Mb chromosomes were selected. For the proposed methods, each rare variant was weighed by \( \rho_i(1 - p_i) \) for \( W \). The results in Figures 1, 2 and 3 showed that FARVAT was the most efficient if all rare variants had deleterious effects, but the gap between FARVAT and the second efficient method FARVAT_e was small. However, the power loss of FARVAT was substantial when rare variants with deleterious and protective variants were present in a gene.

If the proportion of rare variants with deleterious effects was 0.5, FARVAT_r was the most efficient, followed by FARVAT_e. PedCMC and FPCA were usually more efficient than RV-TDT methods, but these approaches were not efficient compared with FARVAT_e in the considered scenarios. Therefore, even though the most powerful statistic depended on the disease model, we concluded that FARVAT_e was generally efficient choice under the various disease models.

3 RESULTS

3.1 Simulation studies

3.1.1 Evaluation with simulated data under the absence of population substructure The statistical validity of FARVAT,

FARVAT_r, and FARVAT_e, was evaluated under the absence of population substructure, and the results were compared with PedCMC, FPCA (Zhu and Xiong, 2012) and RV-TDT methods (He et al., 2014). RV-TDT methods consist of \text{BRV.T01}, \text{BRV.Hapo.T01}, \text{CMC.T01}, \text{CMC.Hapo.T01}, \text{VT.BRV.Hapo}, \text{VT.CMC.Hapo} and \text{WSS.Hapo}. We generated 50 and 100 extended families in each replicate, and empirical type I error estimates at the 0.05, 0.01 and 0.001 significance levels were calculated with 50 000 replicates. For the proposed methods, 1 and \[ \rho_i(1 - p_i) \] were considered for \( w_i \), and the kinship coefficients were used to build the correlation matrix \( \Phi \). Rare variants for which MAFs were <0.01 were considered for all statistics. In Supplementary Tables S1 and S2, 30 and 100 rare variants were randomly selected, and in Supplementary Table S3, all rare variants in the 30 kb genetic region were considered. These results showed that the empirical type I error estimates for FARVAT, FARVAT_r and FARVAT_e preserved the nominal significance levels. However, CMC.T01, BRV.Hapo.T01, CMC.Hapo.T01, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo and FPCA were usually conservative, and BRV.T01 and PedCMC seemed to be liberal. For PedCMC, genotype scores of individuals with more than or equal to a single rare allele were considered as 1; otherwise, they were 0. If the large number of rare variants is collapsed, its convergence to the chi-square distribution requires very large sample sizes, and genotype scores for all individuals can be 1 in extreme scenarios. Therefore, we could conclude that PedCMC may not be a good choice when the number of rare variants in a gene is very large.

The statistical efficiency of FARVAT, FARVAT_r and FARVAT_e was evaluated with the simulated data, and results were compared with results from PedCMC, FPCA and RV-TDT methods (He et al., 2014; Zhu and Xiong, 2012). We assumed that the relative proportion of variances explained by rare variants \( \hat{h}^2 \) was 0.05. In each replicate, we assumed that all rare variants had either deleterious or protective effects on disease, and the proportions of rare variants with deleterious effects were assumed to be 1, 0.8, 0.6 and 0.5. The numbers of extended families were assumed to be 50 and 100. MAFs for all rare variants were assumed to be <0.01. Thirty rare variants in Figure 1 and 100 rare variants in Figure 2 were selected, and in Figure 3, all rare variants within 30 kb from the generated 1 Mb chromosomes were selected. For the proposed methods, each rare variant was weighed by \( \rho_i(1 - p_i) \) for \( W \). The results in Figures 1, 2 and 3 showed that FARVAT was the most efficient if all rare variants had deleterious effects, but the gap between FARVAT and the second efficient method FARVAT_e was small. However, the power loss of FARVAT was substantial when rare variants with deleterious and protective variants were present in a gene.

If the proportion of rare variants with deleterious effects was 0.5, FARVAT_r was the most efficient, followed by FARVAT_e. PedCMC and FPCA were usually more efficient than RV-TDT methods, but these approaches were not efficient compared with FARVAT_e in the considered scenarios. Therefore, even though the most powerful statistic depended on the disease model, we concluded that FARVAT_e was generally efficient choice under the various disease models.
3.1.2 Evaluation with simulated data under the presence of population substructure

We assumed that there were two subpopulations, and founders in each family were randomly selected from two different population substructures. Two subpopulations were simulated with the Balding–Nichols model (Balding and Nichols, 1995), and $F_{ST}$ values were assumed to be 0.005, 0.01 and 0.05. To provide robustness against the population substructure, ($\Phi$ was estimated by 20,000 common variants for which MAFs were <0.05 (Thornton and McPeek, 2010), and this was incorporated to the proposed methods. It should be noted that the proposed method was an extension of the $\text{MQLS}_c$ statistic (Thornton and McPeek, 2007) to rare variant analysis, and that $\text{MQLS}_c$ becomes robust under the presence of population substructure if $\Phi$ was estimated with large-scale genomic data (Thornton and McPeek, 2010).

In Supplementary Table S4, we calculated empirical type 1 error estimates from 50,000 replicates at the 0.01, 0.005 and 0.001 significance levels. Our results showed that the empirical type 1 error estimates for FARVAT$_b$ and FARVAT$_o$ preserved the nominal significance levels for the considered $F_{ST}$ values. However, FPCA, PedCMC and RV-TDT methods were usually conservative, and the level of conservativeness was proportional to the amount of $F_{ST}$. FARVAT$_c$ was also conservative, but was less sensitive than FPCA, PedCMC and RV-TDT methods. Furthermore, we evaluated the statistical efficiency under the presence of population substructure with the simulated data. We assumed that $\hat{h}^2_w$ was 0.05, and the empirical power estimates were calculated with 5000 replicates at the 0.001 significance level. As shown in Figure 4, the most efficient approach differed depending on the disease model. For instance, FARVAT$_b$ was the most efficient when all rare variants had deleterious effects, and FARVAT$_o$ was the most efficient when half of the rare variants had deleterious effects. FARVAT$_c$ was usually the second most efficient; however, the power gap between FARVAT$_o$ and the most efficient method was always small. As a result, we concluded that FARVAT$_o$ was generally a robust and efficient choice for various disease models under the presence of population substructure.

3.2 Analysis of GAW17 simulated data

The statistical efficiency of the proposed methods was evaluated with the binary trait in GAW17 simulated data (Almasy et al., 2011). There were 200 replicates in GAW17 simulated data, and each replicate consisted of 209 affected and 488 unaffected individuals distributed in eight extended pedigrees. In 1714 genes, there were 13,784 variants, and MAFs for 10,710 variants were <0.05. In each gene, rare variants for which MAFs were <0.05 were considered for analysis with the proposed methods, and genes in which the number of rare variants was less than or equal to two were excluded from the analysis. To provide the robustness of the proposed methods under the presence of population substructure, the empirical genetic relationship matrix between individuals was estimated with the common variants. The disease status was decided by the underlying liability, and the top 30% of the underlying liability distribution was declared as being affected. In particular, some covariates were related to the...
underlying liability, and the disease prevalence (Thornton and McPeek, 2010) and BLUP from the linear mixed model (Won and Lange, 2013) were used as offsets. For the linear mixed model, we included sex, age, smoking status and 10 principal component scores calculated from the estimated \((\text{Thornton and McPeek}, 2010)\). Among 36 genes related to binary traits, 20 genes consisted of more than one rare variant, and their empirical powers were determined by counting the number of replicates for which \(P\)-values of causal genes were \(5\times10^{-5}\), \(10^{-5}\), and \(10^{-6}\). As shown in Supplementary Tables S5 and S6, most causal genes were not detectable with the proposed methods; however, \(KDR\), \(VEGFA\), \(SIRT1\) and \(VLDLR\) had relatively high rates of detection. By using RV-TDT methods, we could not find any causal genes. In Supplementary Figures S2 and S3, we provided the qq-plots and Manhattan plots of RV-TDT methods, PedCMC, FARVAT\(_b\), FARVAT\(_c\), and FARVAT\(_o\) with the first replicate of GAW17 simulated data. While PedCMC was not conservative, results from the other methods seemed to be valid. As shown in Supplementary Figure S3, we found that \(VEGFA\) was the most significant for FARVAT\(_o\).

### 3.3 Real data analysis

The proposed methods were applied to the genetic association analysis of rare variants in schizophrenia. Thirty-six trios were collected from Germany for which offspring were affected, whereas parents were unaffected. The whole genomes for all individuals were sequenced. There were 10,829,265 bi-allelic variants, and MAFs of 31,860 among them were \(<0.05\). Markers with high missing call rates (>5%) or significant deviation from Hardy–Weinberg equilibrium \((P<1\times10^{-5})\) were excluded, and trios were filtered out if 10% of variants had Mendelian transmission errors. As a result, 9,216,373 common variants and 31,046 rare variants for 105 trios were analyzed with the proposed methods.

Each rare variant was annotated by the SnpEff program (Cingolani et al., 2012) with the UCSC HG19 database. SnpEff 3.2a categorized each variant to four groups: HIGH, MODERATE, LOW and MODIFIER. In our analysis, rare variants assigned to LOW and MODIFIER categories may have little or no effect on protein function, and they were not considered in our analysis. For each gene, the rare variants with HIGH and MODERATE effects were separately analyzed with the proposed methods. In addition, if MAC of all rare variants in each gene were \(<0.05\), the asymptotic convergence of the proposed method to chi-square distribution may not be provided, and \(P\)-values were calculated for genes for which the MAC was \(\geq 0.05\). In total, \(P\)-values were calculated for 13,053 genes. For the proposed methods, the prevalence of schizophrenia was assumed to be 0.0063, and each rare variant was weighted by \(\left[p_k(1 - p_k)^{1/2}\right]\) for \(W\). To provide robustness under the presence of population substructure, the genetic relationship matrix was estimated with common variants, and these data were incorporated into the
proposed methods. We provided the qq-plots of RV-TDT methods, PedCMC, FARVATb, FARVATc and FARVATo. As presented in Figure 5, although RV-TDT methods, PedCMC and FARVATc methods were conservative and FARVATb showed some violations, FARVATo uniquely seems valid. Figure 6 shows the Manhattan plots for the all methods, and the genome-wide significant results from RV-TDT, FARVATb and FARVATo are summarized in Table 1. We found two genome-wide significant genes with WSS.Hapo, FARVATb and FARVATo, and these genome-wide significant genes will be further investigated with replication studies.

4 DISCUSSION

In this article, we proposed burden-type, C-alpha-type and SKAT-O-type statistics for the association analysis of rare variants for binary traits with extended families. The proposed methods were compared with results of PedCMC, FPCA and RV-TDT methods (He et al., 2014; Zhu and Xiong, 2012), and with extensive simulations, we showed that the proposed method was more efficient than existing approaches. In particular, we found that the most efficient statistic among the proposed statistics differed according to the disease model. However, they were usually followed by the SKAT-O-type statistic in such scenarios, and the power differences between the most efficient statistic and the SKAT-O-type statistic were small. Therefore, FARVATo seemed to be a robust choice for the analysis of rare variants in extended families.

Furthermore, the proposed method was very rapid computationally, and the FARVAT software for the proposed methods was implemented with C++ to enhance computational efficiency. The time complexity for the proposed method was $O(m^3 + N^2m + N^3)$, and we found that analysis of the whole genome sequence data for 1000 individuals in the extended

![Fig. 5. QQ-plot of the rare variant association analysis for schizophrenia. The qq-plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, FARVATb, FARVATc and FARVATo. The 95% confidence interval is provided.](https://academic.oup.com/bioinformatics/article-abstract/30/22/3197/2389885)

![Fig. 6. Manhattan plot of the rare variant association analysis for schizophrenia. The Manhattan plots are provided for BRV.T05, BRV.Hapo.T05, CMC.T05, CMC.Hapo.T05, VT.BRV.Hapo, VT.CMC.Hapo, WSS.Hapo, PedCMC, FARVATb, FARVATc and FARVATo. The x-axis indicates the genome in physical position, and y-axis does $-\log_{10}(P\text{-value})$ for all genes. The horizontal line means the threshold for 0.05 genome-wide significance level by Bonferroni correction is 1.74E-05.](https://academic.oup.com/bioinformatics/article-abstract/30/22/3197/2389885)

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Weight</th>
<th>CHR</th>
<th>GENE</th>
<th>$m$</th>
<th>MAC</th>
<th>$P$-value</th>
<th>$q$-value</th>
<th>Aff</th>
<th>Unaff</th>
</tr>
</thead>
<tbody>
<tr>
<td>WSS.Hapo</td>
<td>1</td>
<td>11</td>
<td>Gene1</td>
<td>5</td>
<td>11</td>
<td>5.00E-06</td>
<td>0.01</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FARVATb</td>
<td>$[p(1-p)]^{1/2}$</td>
<td>8</td>
<td>Gene2</td>
<td>25</td>
<td>4</td>
<td>1.67E-05</td>
<td>0.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FARVATc</td>
<td>$[p(1-p)]^{1/2}$</td>
<td>8</td>
<td>Gene2</td>
<td>25</td>
<td>4</td>
<td>1.30E-05</td>
<td>0.04</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notes. The significant results for each method are provided. The numbers of variants for each significant region are provided, and MAC for affected and unaffected individuals is provided. The 0.05 genome-wide significant level adjusted by Bonferroni correction is 1.74E-05, and $q$-values (Benjamini and Hochberg, 1995) are provided.
family design could be conducted within a few hours. FARVAT can handle various input file formats, such as the ped, bed and vcf files, and multithreaded genome-wide association analyses can be conducted. The software calculates various statistics for the analysis of extended families, and it is freely downloadable from http://biostat.cau.ac.kr/farvat/.

However, despite the analytical flexibility of the proposed method, it has some limitations. First, the proposed method could be less efficient if some covariates associated with disease status or phenotypes of interest were continuous. Our recent investigation found that the power improvement of the analysis with phenotypes adjusted by BLUP could be substantial if each family was randomly selected (Won and Lange, 2013). Under certain scenarios, however, power loss may be expected, and the further investigation is necessary. Second, we showed that incorporation of the estimated correlation matrix to the proposed statistics provided sufficient robustness for the proposed method against the presence of population substructure. However, if large-scale common variants were not available or the level of population substructure depended on the genomic location, the proposed adjustment with the estimated correlation matrix did not perform appropriately (Price et al., 2006; Won et al., 2009), and different strategies would be necessary according to the level of population substructure. If large-scale common variants are not available, the FBAT or TDT statistics, based on so-called within-family components, is uniquely robust to population substructure, and the burden-type test for the FBAT statistic or RV-TDT methods can be used (De et al., 2013; He et al., 2014). If the genomic ancestry for each individual differs for some genomic locations, the so-called hybrid-analysis strategy (Won et al., 2009) can be a suitable alternative. The proposed method can simply be extended to the statistic based on the between-family component (Won and Lange, 2013), and its rank-based P-value can be combined with the FBAT burden-type test or SKAT-O-type test.

Advances in genotyping technology will lead to substantial cost reductions for genome sequencing, and it is expected that whole genome sequencing may be feasible for less than a few hundred U.S. dollars in the near future. Importantly, most of human genome consists of rare variants, and thus, we expect that the genetic background for ‘missing heritability’ can be determined by rare variant association analysis (Manolio et al., 2009). However, rare variant association analysis is disrupted by genetic heterogeneity, and in this context, the importance of rare variant analysis with extended families has often been raised (Ionita-Laza et al., 2011). The proposed method enables the analysis of rare variants within extended families, and its application to extended families may provide a breakthrough for the success of genetic association analysis.

Funding: This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education [2013R1A1A010437]; and by NRF grant funded by the Korea government (MSIP) (No. 2012R1A3A2026438).

Conflict of Interest: none declared.

REFERENCES


Cingolani,P. et al. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly, 6, 80–92.


De,G. et al. (2013) Rare variant analysis for family-based design. PLoS One, 8, e48495.


