HHS Public Access

Author manuscript

Genet Epidemiol. Author manuscript; available in PMC 2017 September 01.

Published in final edited form as:

Genet Epidemiol. 2016 September ; 40(6): 502–511. doi:10.1002/gepi.21985.

Family-based Rare Variant Association Analysis: a Fast and Efficient Method of Multivariate Phenotype Association Analysis

Longfei Wang[1], **Sungyoung Lee**[1], **Jungsoo Gim**[2], **Dandi Qiao**[3,4], **Michael Cho**[3,5], **Robert C Elston**[6], **Edwin K Silverman**[3,5], and **Sungho Won**[1,7,*]

[1]Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, 151-742, Korea

[2]Institute of Health and Environment, Seoul National University, Seoul, 151-742, Korea

[3]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA 02115, USA

[4]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA

[5]Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA 02115, USA

[6]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH, 44106, USA

[7]Graduate School of Public Health, Seoul National University, Seoul, 151-742, Korea

Abstract

Motivation—Family-based designs have been repeatedly shown to be powerful in detecting the significant rare variants associated with human diseases. Furthermore, human diseases are often defined by the outcomes of multiple phenotypes, and thus we expect multivariate family-based analyses may be very efficient in detecting associations with rare variants. However, few statistical methods implementing this strategy have been developed for family-based designs. In this report, we describe one such implementation: the multivariate family-based rare variant association tool (*mFARVAT*).

Results—*mFARVAT* is a quasi-likelihood-based score test for rare variant association analysis with multiple phenotypes, and tests both homogeneous and heterogeneous effects of each variant on multiple phenotypes. Simulation results show that the proposed method is generally robust and efficient for various disease models, and we identify some promising candidate genes associated with chronic obstructive pulmonary disease.

Keywords

Family-based design; rare variants; association analysis; multivariate phenotypes

*Corresponding Authors: Sungho Won, Dept. of Public Health Science, Seoul National University, 1 Gwanak-ro Gwanak-gu, Seoul 151-742, Korea, won1@snu.ac.kr, (Tel) +82-2-880-2714.

Availability and Implementation: The software is freely available at http://healthstat.snu.ac.kr/software/mfarvat/, implemented in C++ and supported on Linux and MS Windows.

1 INTRODUCTION

In spite of tens of thousands of genome-wide association studies (GWAS), the so-called missing heritability (Manolio, Collins et al. 2009) reveals that analyses of common variants detect only a limited number of disease susceptibility loci and a substantial amount of causal variants may remain undiscovered by GWAS. Sequencing technology was expected to supply this additional information by obtaining large stretches of DNA spanning the entire genome, and improvements in this technology have enabled genetic association analysis of rare/common causal variants. However, the 'common disease rare variant' hypothesis implies that multiple rare variants can affect disease status and thus the proportion of affected individuals sharing the same causal variants could be very small. Therefore, analyses of rare variants suffer from genetic heterogeneity among affected individuals. In this context, because affected relatives have more chance to share the same causal variants (Shi and Rao 2011), and hence the genetic heterogeneity among affected relatives is expected to be smaller, family-based analyses have been repeatedly addressed as an important strategy.

Genetic association analyses simultaneously test a large number of variants, and stringent significance levels imposed by the multiple testing problem highlight the importance of powerful strategies. In particular, multiple measurements can be obtained from different but related phenotypes, or from repeated measurements of a single phenotype at different time points. Association analyses with multiple phenotypes often lead to substantial improvements in statistical power (Schifano, Li et al. 2013) and such improvements are inversely related to correlations between phenotypes (Lee, Park et al. 2014). Several different methods have been proposed, including the scaled marginal model (Schifano, Li et al. 2013) and the extended Simes procedures for population-based samples (van der Sluis, Posthuma et al. 2013). The statistical power of these methods depends on the relationships between the causal variants and the multiple phenotypes, which are usually unknown (van der Sluis, Posthuma et al. 2013); this property applies to rare variant association analyses. For instance, if the effects of the rare variants on each of the multiple phenotypes are in the same direction, the burden test may be most efficient; but if the multiple genetic effects are heterogeneous, SKAT may be more reasonable (Lee, Wu et al. 2012).

However, phenotypic relatedness between family members complicates parameter estimation, particularly for dichotomous phenotypes. For this situation, very few approaches other than FBAT statistics (Laird, Horvath et al. 2000), which can be used to conduct multivariate genetic association analyses with large families, are available. FBAT statistics preserve robustness against population substructure and have been extended for joint analysis of multiple phenotypes and genotypes (Gray-McGuire, Bochud et al. 2009), and for rare variant association analysis (Yip, De et al. 2011). However, FBAT statistics do not fully use the information in the parental phenotypes, and loss of power can be substantial if the number of founders is relatively large.

Recently, the *FA*mily-based *R*are *V*ariant *A*ssociation *T*est (*FARVAT*) based on quasi-likelihood was proposed (Choi, Lee et al. 2014). *FARVAT* is robust against population substructure, and includes burden, SKAT and SKAT-O statistics for both dichotomous and

quantitative phenotypes. In this report, we extend *FARVAT* to implement the multivariate family-based rare variant association analysis tool (*mFARVAT*). *mFARVAT* includes both homogeneous and heterogeneous approaches, and, in this respect, is similar to skatMeta (Lee, Teslovich et al. 2013). The method can analyze both quantitative and dichotomous phenotypes, and is robust against population substructure if the correlation matrix between individuals can be estimated from large-scale genetic data. *mFARVAT* is implemented in C++, and is computationally fast even for extended families. Furthermore, *mFARVAT* was applied to multiple phenotypes associated with chronic obstructive pulmonary disease (COPD), and some promising results illustrate its practical value.

2 METHODS

For genetic association analyses either prospective or retrospective approaches can be selected and the choice of strategy depends on the sampling scheme. However, it has been shown that even for prospectively selected samples, retrospective analyses can preserve virtually similar statistical power as prospective analyses. Additionally, retrospective strategies are robust against non-normality of phenotypes, and are computationally less intensive (Won and Lange 2013). Therefore, we consider retrospective analysis for both prospectively and retrospectively selected samples, and genetic association is detected by testing the independence of genotype distributions with phenotypes.

2.1 Notation and disease model

Association between M genetic variants and Q phenotypes is examined, and we denote the coded genotype of individual j in family i at variant m and phenotype q by x_{ijm} and y_{ijq}, respectively. We assume there are n families and n_i individuals in family i. Thus, the sample size, N, is $\sum_{i=1}^{n} n_i$. We let

$$\mathbf{X}^m = \begin{bmatrix} x_{11m} \\ \vdots \\ x_{nn_nm} \end{bmatrix}, \mathbf{X} = (\mathbf{X}^1, \ldots, \mathbf{X}^M), \text{ and}$$

$$\boldsymbol{Y}^q = \begin{bmatrix} y_{11q} \\ \vdots \\ y_{nn_nq} \end{bmatrix}, \boldsymbol{Y} = (\boldsymbol{Y}^1, \ldots, \boldsymbol{Y}^Q).$$

We also define

$$\mathbf{X}_{ij} = \begin{bmatrix} x_{ij1} \\ \vdots \\ x_{ijM} \end{bmatrix}, \text{ and } \boldsymbol{Y}_{ij} = \begin{bmatrix} y_{ij1} \\ \vdots \\ y_{ijQ} \end{bmatrix}.$$

The genetic variance-covariance matrix between individuals can be parameterized with the kinship coefficient matrix (KCM), $\boldsymbol{\Phi}$. If we let $\boldsymbol{\pi}_{ij,ij'}$ be the kinship coefficient between

individual *j* and individual *j′* in family *i*, and let d_{ij} be the inbreeding coefficient for individual *j* in family *i*, $\mathbf{\Phi}_i$ is

$$\begin{bmatrix} 1+d_{i1} & 2\pi_{i1,i2} & 2\pi_{i1,i3} & \cdots \\ 2\pi_{i1,i2} & 1+d_{i2} & 2\pi_{i2,i3} & \cdots \\ 2\pi_{i1,i3} & 2\pi_{i2,i3} & 1+d_{i3} & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix},$$

and we define

$$\mathbf{\Phi}=\begin{bmatrix} \mathbf{\Phi}_1 & 0 & 0 & \cdots \\ 0 & \mathbf{\Phi}_2 & 0 & \cdots \\ 0 & 0 & \mathbf{\Phi}_3 & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}.$$

In the presence of population substructure, $\mathbf{\Phi}$ should be replaced with the genetic relationship matrix (GRM) to provide statistically valid results (Thornton, Tang et al. 2012). The variance-covariance matrix between the *M* additively coded markers is denoted by $\mathbf{\Psi}$, and we assume that

$$\mathrm{cov}(\mathbf{X}_{ij}, \mathbf{X}_{i'j'}) \approx 2\pi_{ij,i'j'}\mathrm{var}(\mathbf{X}_{ij})=2\pi_{ij,i'j'}\mathbf{\Psi}.$$

Then we can easily show that

$$\mathrm{var}(\mathrm{vec}(\mathbf{X})) \approx \mathbf{\Psi} \otimes \mathbf{\Phi}.$$

2.2 Choice of offset

It has been shown that the statistical efficiency of test statistics in retrospective analysis can be improved by adjusting phenotypes for relevant covariates (Lange, DeMeo et al. 2002). For our score statistic, we introduced a new parameter μ_{ijq} for phenotype *q* of individual *j* in family *i*, which will be called the offset in the remainder of this report (Won and Lange 2013). We set

$$\boldsymbol{\mu}_{ij}=\begin{bmatrix} \mu_{ij1} \\ \vdots \\ \mu_{ijQ} \end{bmatrix}, \boldsymbol{\mu}=(\boldsymbol{\mu}_{\mathbf{11}}^{\boldsymbol{t}}, \ldots, \boldsymbol{\mu}_{\boldsymbol{nn_n}}^{\boldsymbol{t}})^t, \mathbf{T}_{ij}=\boldsymbol{Y}_{ij}-\boldsymbol{\mu}_{ij}, \mathbf{T}=\boldsymbol{Y}-\boldsymbol{\mu}.$$

Statistical efficiency depends on $\boldsymbol{\mu}$, and thus its elements need to be carefully selected. The offset $\boldsymbol{\mu}$ can be either calculated by the best linear unbiased predictor (BLUP) with covariates, as done for SKAT, or the disease prevalence can be used (Won and Lange 2013). The most efficient $\boldsymbol{\mu}$ will depend on the sampling scheme. If families are randomly selected, BLUP was shown to be most efficient for both dichotomous and quantitative phenotypes

(Won and Lange 2013), while prevalence was recommended to study dichotomous phenotypes if families with a large number of affected family members are selected (Thornton and McPeek 2007, Won and Lange 2013). Therefore, we chose BLUP and prevalence as offsets for quantitative phenotypes and dichotomous phenotypes, respectively.

2.3 Score for quasi-likelihood

We let $\mathbf{e}_{ij}$ be an N dimensional vector in which the $\left(j+\sum_{i'=1}^{i-1} n_{i'}\right)^{\text{th}}$ element is 1 and the others are 0, and $\mathbf{1}_w$ be a column vector with w elements all equal to 1. We denote the effect of rare variant m on phenotype q as β_{mq} which is the regression coefficients of the phenotype on the causal variants. We consider the score statistic and thus β_{mq} is not needed to be estimated. However, the false positive rates can be inflated and the statistic for each β_{mq} has large false negative rates. Therefore, collapsed genotype scores were utilized to prevent these problems. Under the null hypothesis, which is $\beta_{11} = \cdots = \beta_{MQ} = 0$, the best linear unbiased estimator (BLUE) for $E(\mathbf{X}^m)$ (McPeek, Wu et al. 2004) is

$$\mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{X}^m,$$

and if we let $\mathbf{A}=\boldsymbol{\Phi}^{-1}-\boldsymbol{\Phi}^{-1}\mathbf{1}_N(\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}\mathbf{1}_N)^{-1}\mathbf{1}_N^t\boldsymbol{\Phi}^{-1}$, we can define $\mathbf{S}_{ij}^m$ for the individual j in family i by

$$\mathbf{S}_{ij}^m=(\mathbf{T}_{ij}\mathbf{e}_{ij}^t)\boldsymbol{\Phi}\mathbf{A}\mathbf{X}^m.$$

Based on MFQLS (Won, Kim et al. 2015), the score vector for the M variants can be defined by

$$\mathbf{S}=(\mathbf{S}^1,\ldots,\mathbf{S}^M)=\mathbf{T}^t\boldsymbol{\Phi}\mathbf{A}\mathbf{X},$$

and because $\text{var}(\text{vec}(\mathbf{X})) \approx \boldsymbol{\Psi}\otimes\boldsymbol{\Phi}$, the variance-covariance matrix for $\mathbf{S}$ is approximately equal to

$$\text{var}(\text{vec}(\mathbf{S})) \approx \boldsymbol{\Psi}\otimes(\mathbf{T}^t\boldsymbol{\Phi}\mathbf{A}\boldsymbol{\Phi}\mathbf{T}).$$

2.4 Homogeneous mFARVAT

The effects of each causal variant on a phenotype, estimated as the regression coefficients of the phenotype on the causal variants, can be in the same or different directions, and we propose two different statistics for these two scenarios. Our first statistic, homogeneous *mFARVAT*, assumes that effects of each causal variant on the multiple phenotypes are in the same direction, for example, when the phenotypes are highly correlated or longitudinal. For rare variant association analysis, burden tests regress phenotypes on the sum of genotype scores over rare variants. Therefore, association of the Q phenotypes with variant m can be built by testing whether $\beta_{m1} + \ldots + \beta_{mQ} = 0$, and we can provide a statistic based on $\mathbf{1}_Q^t\mathbf{S}$.

The importance of each variant is often different and statistical efficiency can be improved by weighting each variant based on its relative importance (Madsen and Browning 2009). Relative importance is usually expressed by a function of minor allele frequency (MAF). We assume that the weight for variant m is w_m and $\mathbf{W}$ is an $M \times M$ diagonal matrix with diagonal elements w_m; we choose w_m = beta(p_m, a_1, a_2) proposed by Wu et al (Wu, Lee et al. 2011), where p_m is the MAF of variant m and a_1 and a_2 were set to be 1 and 25 respectively. beta(p_m, a_1, a_2) is flexible because it can accommodate a broad range of scenarios by considering different a_1 and a_2, and Wu et al found that the choices of a_1 and a_2 were often efficient. Then the scores for the burden and SKAT tests are, respectively,

$$\frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W 1}_M \mathbf{1}_M^t \mathbf{W S}^t \mathbf{1}_Q,$$

and

$$\frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W W S}^t \mathbf{1}_Q.$$

If we let

$$\mathbf{R}_\rho^{\mathrm{Hom}} = (1-\rho)\mathbf{I}_M + \rho \mathbf{1}_M \mathbf{1}_M^t,$$

scores for burden and SKAT tests can be generalized as

$$MS_\rho^{\mathrm{Hom}} = \frac{1}{\mathbf{1}_Q^t \mathbf{T}^t \mathbf{\Phi A \Phi T} \mathbf{1}_Q} \mathbf{1}_Q^t \mathbf{S W R}_\rho^{\mathrm{Hom}} \mathbf{W S}^t \mathbf{1}_Q,$$

where the optimal choice of ρ depends on the distribution of rare variant effects on the multiple phenotypes.

We denote the eigenvalues of $\mathbf{\Psi}^{1/2} \mathbf{W R}_\rho^{\mathrm{Hom}} \mathbf{W \Psi}^{1/2}$ by $(\lambda_1^\rho, \ldots, \lambda_M^\rho)$. If we let $\chi_{1,m}^2$ be an independent chi-square distribution with a single degree of freedom, we have

$$MS_\rho^{\mathrm{Hom}} \sim \sum_{m=1}^{M} \lambda_m^\rho \chi_{1,m}^2.$$

If we denote the p-value for MS_ρ^{Hom} by pMS_ρ^{Hom}, and let $pmFARVAT_S^{\mathrm{Hom}} = pMS_0^{\mathrm{Hom}}$ and $pmFARVAT_B^{\mathrm{Hom}} = pMS_1^{\mathrm{Hom}}$, the SKAT-O *mFARVAT* (*mFARVAT*$_O$) statistic is defined by

$$mFARVAT_O^{\mathrm{Hom}} = \min\{pMS_0^{\mathrm{Hom}}, pMS_{0.1^2}^{\mathrm{Hom}}, \ldots, pMS_{0.5^2}^{\mathrm{Hom}}, pMS_1^{\mathrm{Hom}}\}.$$

Its p-value will be denoted as $pmFARVAT_O^{\mathrm{Hom}}$ in the remainder of this report, and can be calculated from the numerical algorithm for SKAT-O (Lee, Wu et al. 2012), with a small modification (see Supplementary Text 1 for the detailed algorithm).

2.5 Heterogeneous mFARVAT

The effect of each variant on a phenotype can be heterogeneous in certain situations, and it may be reasonable to consider such effects separately. Therefore, we can provide statistics based on vec(**S**), and, under the null hypothesis $\beta_{11} = \cdots = \beta_{MQ} = 0$, we have

$$\mathrm{E}\{\mathrm{vec}(\mathbf{S})\}=\mathbf{0} \text{ and } \mathrm{var}\{\mathrm{vec}(\mathbf{S})\}=\mathbf{\Psi}\otimes\mathbf{T}^t\mathbf{\Phi}\mathbf{A}\mathbf{\Phi}\mathbf{T}.$$

If we assume that $\mathbf{I}_w$ is a $w \times w$ identity matrix and

$$\mathbf{R}_\rho^{\mathrm{Het}}=(1-\rho)\mathbf{I}_{MQ}+\rho\mathbf{1}_{MQ}\mathbf{1}_{MQ}^t,$$

we define the generalized score by

$$MS_\rho^{\mathrm{Het}}=\mathrm{vec}(\mathbf{S})^t(\mathbf{I}_Q\otimes\mathbf{W})\mathbf{R}_\rho^{\mathrm{Het}}(\mathbf{I}_Q\otimes\mathbf{W})\mathrm{vec}(\mathbf{S}).$$

Then the burden and SKAT tests can be expressed as

$$MS_1^{\mathrm{Het}}=\mathrm{vec}(\mathbf{S})^t(\mathbf{I}_Q\otimes\mathbf{W})\mathbf{1}_{MQ}\mathbf{1}_{MQ}^t(\mathbf{I}_Q\otimes\mathbf{W})\mathrm{vec}(\mathbf{S}),$$

$$MS_0^{\mathrm{Het}}=\mathrm{vec}(\mathbf{S})^t(\mathbf{I}_Q\otimes\mathbf{W})(\mathbf{I}_Q\otimes\mathbf{W})\mathrm{vec}(\mathbf{S}).$$

If we let $(\lambda_1^{'\rho},\ldots,\lambda_{MQ}^{'\rho})$ be the eigenvalues of

$$(\mathbf{\Psi}^{1/2}\otimes(\mathbf{T}^t\mathbf{\Phi}\mathbf{A}\mathbf{\Phi}\mathbf{T})^{1/2})(\mathbf{I}_Q\otimes\mathbf{W})\mathbf{R}_\rho\times(\mathbf{I}_Q\otimes\mathbf{W})(\mathbf{\Psi}^{1/2}\otimes(\mathbf{T}^t\mathbf{\Phi}\mathbf{A}\mathbf{\Phi}\mathbf{T})^{1/2}),$$

then we have

$$MS_\rho^{\mathrm{Het}}\sim\sum_{l=1}^{MQ}\lambda_l^{'\rho}\chi_{1,l}^2 \text{ under } H_0.$$

P-values for MS_ρ^{Het} will be denoted by pMS_ρ^{Het}, and we let $pmFARVAT_S^{\mathrm{Het}}=pMS_0^{\mathrm{Het}}$ and $pmFARVAT_B^{\mathrm{Het}}=pMS_1^{\mathrm{Het}}$. We consider

$$mFARVAT_O^{\mathrm{Het}}=\min\left\{pMS_0^{\mathrm{Het}},pMS_{0.1^2}^{\mathrm{Het}},\ldots,pMS_{0.5^2}^{\mathrm{Het}},pMS_1^{\mathrm{Het}}\right\}.$$

We let the p-value for $mFARVAT_O^{\mathrm{Het}}$ be $pmFARVAT_O^{\mathrm{Het}}$ and the detailed algorithm to calculate the asymptotic p-value is provided in Supplementary Text 2.

2.6 The simulation model

To evaluate *mFARVAT*, we simulated large families that extend three generations and consist of 10 members (see Supplementary Figure 1). 5,000 haplotypes with 50,000 base pairs were generated under a coalescent model using the software COSI (Schaffner, Foo et al. 2005). Each haplotype was generated by setting the mutation rate at 1.5×10^{-8}. Haplotypes were randomly chosen with replacement to build founder genotypes. Nonfounder haplotypes were determined in Mendelian fashion from pairs of parents under the assumption of no recombination. For each simulated haplotype, we defined variants with sample MAFs less than 0.01 as being rare, and 60 rare variants were randomly selected.

Phenotypes were generated under the null and alternative hypotheses, and we considered both quantitative and dichotomous phenotypes. Quantitative phenotypes were defined by summing the phenotypic mean, polygenic effect, main genetic effect and random error, and we assumed there was no environmental effect shared between family members. Phenotypic means were denoted by $\alpha_1,\ldots, \alpha_{Q-1}$ and α_Q. We assumed that $\alpha_1 = 0$, $\alpha_2 = 0.3$ for $Q = 2$, and $\alpha_1 = \alpha_2 = \alpha_3 = 0$, $\alpha_4 = \alpha_5 = 0.3$ for $Q = 5$. The polygenic effects for the Q phenotypes for each founder were independently generated from $MVN(\mathbf{0},\mathbf{\Sigma_B})$, and for nonfounders the average of maternal and paternal polygenic effects were combined with values independently sampled from $MVN(\mathbf{0}, 0.5\mathbf{\Sigma_B})$. Random errors for the Q phenotypes were assumed to be independent, so the random error for phenotype q was independently sampled from $N(\mathbf{0}, \sigma_{\mathbf{E},q}^2)$. If $Q = 2$, we assumed that

$$\mathbf{\Sigma_B} = \begin{bmatrix} 1 & \sqrt{2}c \\ \sqrt{2}c & 2 \end{bmatrix}, \sigma_{E,1}^2 = 1, \sigma_{E,2}^2 = 2,$$

and if $Q = 5$, they were

$$\mathbf{\Sigma_B} = \begin{bmatrix} 1 & c & \sqrt{2}c & \sqrt{2}c & \sqrt{2}c \\ c & 1 & \sqrt{2}c & \sqrt{2}c & \sqrt{2}c \\ \sqrt{2}c & \sqrt{2}c & 2 & 2c & 2c \\ \sqrt{2}c & \sqrt{2}c & 2c & 2 & 2c \\ \sqrt{2}c & \sqrt{2}c & 2c & 2c & 2 \end{bmatrix},$$

$$\sigma_{E,1}^2 = 1, \sigma_{E,2}^2 = 2, \sigma_{E,3}^2 = 3, \sigma_{E,4}^2 = 4, \sigma_{E,5}^2 = 5.$$

For c we chose 0.5 and 0.8.

The genetic effect at variant m for phenotype q was the product of β_{mq} and the number of disease susceptibility alleles. Under the null hypothesis, β_{mq} was assumed to be 0. Under the

alternative hypothesis, if we let h_a^2 be the proportion of variance explained by rare variants, β_{mq} was sampled from $U(0, v_q)$, where

$$v_q = \sqrt{\frac{(\sigma^2_{B,q} + \sigma^2_{E,q})h_a^2}{(1-h_a^2)\sum_{m=1}^{M}\beta_{mq}^2 2p_m(1-p_m)}}.$$

Here $\sigma^2_{\mathbf{B},q}$ indicates the $(q,q)^{\text{th}}$ element of $\mathbf{\Sigma_B}$, and we assumed that $h_a^2 = 0.02$. β_{mq} was generated for both heterogeneous and homogeneous scenarios. For homogeneous scenarios, we assumed that the effects of each rare variant on different phenotypes are similar. For example, the ratios between β_{m1}, …, and β_{mQ} were assumed to be 1:0.9 if Q = 2, and 1:0.9:0.8:0.7:0.6 if Q = 5. For heterogeneous scenario, the effects of each rare variant on phenotypes were independently generated from $U(0, v_q)$.

Simulation of dichotomous phenotypes was performed using the liability threshold model. Once the quantitative phenotypes with genetic effect, polygenic effect and random error were generated, they were transformed to being affected for quantitative phenotypes larger than the threshold, and otherwise were transformed to unaffected. The threshold was chosen to preserve the assumed disease prevalence. We assumed that prevalences of the multiple phenotypes were 0.1 or 0.2 if $Q = 2$, and 0.1, 0.2, 0.2, 0.3, or 0.3 if $Q = 5$. To allow for the ascertainment bias of dichotomous phenotypes in our simulation studies, we assumed that families with at least one affected individual were selected for analysis.

3 RESULTS

3.1 Evaluation of *mFARVAT* with simulated data

To evaluate statistical validity, type-1 error estimates for both dichotomous and quantitative phenotypes were calculated at various significance levels using 20,000 replicates of two hundred extended families, so that each replicate sample contained 2,000 individuals. Supplementary Table 1 shows empirical type-1 error estimates for homogeneous *mFARVAT* (*mFARVAT*^{Hom}) and heterogeneous *mFARVAT* (*mFARVAT*^{Het}) at the 0.05, 0.01, 0.001, and 2.5×10^{-6} significance levels. The estimates are virtually equal to the nominal significance levels for both quantitative and dichotomous phenotypes. Quantile-quantile (QQ) plots in Supplementary Figures 2 and 3 also show consistent results, and we conclude that *mFARVAT*^{Het} and *mFARVAT*^{Hom} are statistically valid.

Empirical power estimates were calculated at the 10^{-4} significance level with correlations 0.5 and 0.8 for quantitative phenotypes (for the underlying quantitative phenotypes in the case of dichotomous phenotypes). We considered two different scenarios, in which either all or half the rare variants were causal, and assumed that 50%, 80% and 100% of causal variants were deleterious, with the rest being protective. Empirical power estimates were calculated with 2,000 replicates for six different statistics: (1) $mFARVAT_O^{\text{Het}}$; (2) $mFARVAT_O^{\text{Hom}}$; (3) $mFARVAT_S^{\text{Het}}$; (4) $mFARVAT_S^{\text{Hom}}$; (5) $mFARVAT_B^{\text{Het}}$; (6) $mFARVAT_B^{\text{Hom}}$. Results are provided in Tables 1–3 and Tables 4–6, which represent respectively scenarios where all or half the rare variants are causal. Notably, each method

performed similarly in both scenarios, although the empirical power estimates improve if causal variants are more abundant.

We first examined the efficiency of the methods. Tables 1–6 confirm that the most efficient method depends on the disease model, which tends to be unknown. For example, when all the rare causal variants have deleterious effects on all phenotypes, burden *mFARVAT* ($mFARVAT_B$) outperforms all other approaches, but if there are variants with deleterious and protective effects, SKAT *mFARVAT* ($mFARVAT_S$) is the most efficient. SKAT-O *mFARVAT* ($mFARVAT_O$) is not always the best, but its empirical power estimates are usually very close to those of the most efficient approach. Therefore, our results are consistent with previous findings that $mFARVAT_O$ is robust and efficient for various disease models (Lee, Wu et al. 2012).

We also compared the performance of $mFARVAT^{Het}$ and $mFARVAT^{Hom}$ using simulated data. Tables 1–6 show that if the effects of each rare variant on phenotypes are heterogeneous, $mFARVAT^{Het}$ performs better than $mFARAVAT^{Hom}$, and *vice versa*. In addition, when the effects of causal variants go in different directions, as in cases where some variants are deleterious while others are protective, the gap between the power of $mFARVAT^{Het}$ and $mFARAVAT^{Hom}$ is larger than in a scenario where such effects are in the same direction. Interestingly, for each method the statistical power difference between 100% and 50% deleterious causal variants seems to be larger for family-based samples than that for population-based designs (Lee, Emond et al. 2012).

Results for dichotomous phenotypes tend to be similar to those for quantitative phenotypes, although statistical power for the former is usually smaller. This difference may be explained by the fact that dichotomous phenotypes were transformed from quantitative phenotypes. Moreover, overall the power is seen to be inversely related to correlations among phenotypes. There is some power loss when c is increased from 0.5 to 0.8. Notably, when more phenotypes are included in the analysis, *mFARVAT* performs more effectively.

Last, we compared the proposed method with univariate analyses using *FARVAT* (Choi, Lee et al. 2014). The minimum p-value adjusted by Bonferroni correction was selected to calculate the power of univariate analyses. We considered two scenarios: multiple phenotypes are associated with variants and only a single phenotype is associated with variants. Results in Tables 1–6 show that for the former scenario multivariate rare variant analyses perform better than univariate analyses. For the latter scenario, univariate rare variant analyses outperform multivariate analyses (see Supplementary Table 2).

3.2 Application to real data

We applied *mFARVAT* to whole-exome sequencing data from the Boston Early-onset COPD Study (Silverman, Chapman et al. 1998). Sequencing was performed at the University of Washington Center for Mendelian Genomics. Quality control was performed using PLINK (Purcell, Neale et al. 2007), vcfTools (Danecek, Auton et al. 2011), and PLINK/SEQ at Brigham and Women's Hospital. Quality control included Mendelian error rates (< 1%), Hardy-Weinberg equilibrium (HWE, $p > 10^{-8}$), and average sequencing depth (> 12). Relatedness of individuals was evaluated by comparing KCM and GRM. Heterozygous/

homozygous genotype ratio, Mendelian errors, proportion of variants in dbSNP and proportion of non-synonymous variants were used to identify outliers. After additionally filtering out samples with missing phenotypes or covariates, 254 samples from 49 families were obtained.

We considered five COPD-related phenotypes: forced expiratory volume in one second pre-bronchodilator (FEVPRE); forced vital capacity post-bronchodilator (FVCPST); forced expiratory flow 25–75% pre-bronchodilator (DPRF2575); FEVPRE divided by FVCPRE (RATIO); and DPRF2575 divided by FVCPRE (F2575RAT). Sex, age, height, and pack-years of cigarette smoking were utilized to estimate BLUP offsets. It should be noted that genotypes were not used to estimate offsets. The correlation structure of the phenotypes is shown in Supplementary Table 3.

We assumed that variants with MAFs less than 5% were rare, and we considered only genes with at least two rare variants and a minor allele count (MAC) of at least four. As a result, 8126 genes and 88,373 rare variants were analyzed. Our statistic requires the correlation matrix between individuals to obtain $\mathbf{\Phi}$. If there exists population substructure, GRM should be utilized for $\mathbf{\Phi}$ and otherwise KCM is adequate. We found no significant population substructure, and KCM was used for $\mathbf{\Phi}$. The Bonferroni-corrected 0.05 genome-wide significance level is 6.15E-6. QQ plots in Supplementary Figures 6 show the statistical validity of our analysis. Manhattan plots are shown in Supplementary Figure 7. The top 10 most significant results from $mFARVAT^{Het}$ and $mFARVAT^{Hom}$ are shown in Table 7. We could not find any genome-wide significant results with association analysis of multiple phenotypes. The most significant result was found for *KRTAP5-9* on chromosome 11, with $mFARVAT^{Het}$ (p-value = 1.00×10^{-4}), but the p-value for *KRTAP5-9* from $mFARVAT^{Hom}$ is 2.72×10^{-4}. The smaller p-value of $mFARVAT^{Het}$ may indicate that effect of each rare variant on the multiple phenotypes is heterogeneous.

4 DISCUSSION

Extended families have complex correlation structure and association analyses using extended families are very complicated, in particular for dichotomous phenotypes. For instance, the unbalanced nature of family-based samples can lead to inflation or deflation of sandwich estimators for the variance-covariance matrix, and results from generalized estimating equation can be invalid (Wang, Lee et al. 2013). An alternative approach is to use a generalized linear mixed model. However, calculating maximum likelihood estimators requires numerical integration, which is computationally very intensive, and approximations to avoid this can introduce serious bias (Gilmour, Anderson et al. 1985, Schall 1991). Therefore in spite of the efficiency of extended families for rare variant association analysis, few methods have been suggested for family-based association analyses. In this report, we propose a new method of family-based analysis of rare variants associated with dichotomous phenotypes, quantitative phenotypes, or both. The proposed method enables multivariate analyses of extended families to detect rare variants. Extensive simulation studies show that *mFARVAT* works well for dichotomous and quantitative phenotypes. Our method is computationally efficient and association analyses at the genome-wide scale are computationally feasible for extended families. In our analyses, an Intel (R) Xeon (R)

E5-2620 0 CPU at 2.00GHz, with a single node and 80 gigabyte memory, required six minutes to analyze the real data on two phenotypes. *mFARVAT* is implemented in C++ and freely downloadable from http://healthstat.snu.ac.kr/software/mfarvat.

However, in spite of the analytical flexibility and efficiency of the method, some limitations still remain. First, GRM should ideally be used as the correlation matrix $\mathbf{\Phi}$ to provide robustness against population substructure; however, proper estimation of GRM requires large-scale common variants. In the absence of such data, the transmission disequilibrium test (Laird, Horvath et al. 2000) is a unique alternative. Second, the proposed statistics are for retrospective designs and power loss is expected if samples are prospectively gathered. It has been shown that appropriate choice of offset minimizes power loss in certain scenarios but further investigation is still necessary. Third, *mFARVAT* cannot be used directly to analyze X-linked variants. The distribution of X-linked genetic variants in the male is different from that in female, and thus different statistics for males and females are required. This issue will be investigated in future work.

Over the last decade, we have recognized that a substantial amount of unidentified genetic risk exists, and much effort has been expended to investigate this risk. Our methods provide an efficient strategy to analyze rare variant associations in family-based samples, and it may increase understanding of heritable diseases.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF), as funded by the Korean Government [NRF-2014S1A2A2028559]; the Industrial Core Technology Development Program, as funded by the Ministry of Trade, Industry and Energy, Korea (MOTIE) [10040176]; the Basic Science Research Program through the NRF, as funded by the Ministry of Education [NRF-2013R1A1A2010437]; National Institutes of Health [P01 HL105339, R01 HL075478, R01 HL113264]. Sequencing for the Boston Early-Onset COPD Study was provided by University of Washington Center for Mendelian Genomics, and was funded by the National Human Genome Research Institute and by the National Heart, Lung and Blood Institute [1U54HG006493 to D.N., J.S. and M.B.]

Appendix

Numerical algorithm to calculate $_{p}mFARVAT_{o}^{Hom}$

If we let

$$\boldsymbol{Z}=\boldsymbol{\Psi}^{1/2}\mathbf{W},\ \text{and}\ \overline{\boldsymbol{Z}}=\boldsymbol{Z}\mathbf{1}_{MQ}(\mathbf{1}_{MQ}^{t}\mathbf{1}_{MQ})^{-1},$$

the projection matrix onto a space spanned by $\overline{\mathbf{Z}}$ becomes

$$\boldsymbol{\Pi}=\overline{\boldsymbol{Z}}(\overline{\boldsymbol{Z}}^{t}\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^{t}.$$

If we let

$$\mathbf{u}=\mathbf{\Psi}^{-1/2}\mathbf{S}^t\mathbf{1}_Q\frac{1}{\sqrt{\mathbf{1}_Q^t\mathbf{T}^t\mathbf{\Phi}\mathbf{A}\mathbf{\Phi}\mathbf{T}\mathbf{1}_Q}},\ \mathbf{u}\sim MVN(0,\mathbf{I}_{MQ}),$$

MS_ρ^{Hom} becomes

$$MS_\rho^{\mathrm{Hom}}=\mathbf{u}^t\mathbf{\Psi}^{\frac{1}{2}}\mathbf{W}\mathbf{R}\mathbf{W}\mathbf{\Psi}^{\frac{1}{2}}\mathbf{u}=(1-\rho)\mathbf{u}^t\boldsymbol{Z}\boldsymbol{Z}^t\mathbf{u}+\rho M^2\mathbf{u}^t\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}^t\mathbf{u}.$$

As was shown by Lee et al (Lee, Wu et al. 2012), if we let

$$\tau(\rho)=M^2\rho\overline{\boldsymbol{Z}}^t\overline{\boldsymbol{Z}}+\frac{(1-\rho)}{\overline{\boldsymbol{Z}}^t\overline{\boldsymbol{Z}}}\overline{\boldsymbol{Z}}^t\boldsymbol{Z}\boldsymbol{Z}^t\overline{\boldsymbol{Z}},$$

we have

$$MS_\rho^{\mathrm{Hom}}=(1-\rho)\mathbf{u}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\boldsymbol{Z}\boldsymbol{Z}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\mathbf{u}+2(1-\rho)\mathbf{u}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\boldsymbol{Z}\boldsymbol{Z}^t\mathbf{\Pi}\mathbf{u}+\tau(\rho)\mathbf{u}^t\mathbf{\Pi}\mathbf{u},$$

where $\mathbf{u}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\mathbf{u}$, $\mathbf{u}^t(\mathbf{I}_{MQ}-\mathbf{\Pi})\mathbf{Z}\mathbf{Z}^t\mathbf{\Pi}\mathbf{u}$ and $\mathbf{u}^t\mathbf{\Pi}\mathbf{u}$ are mutually independent. Therefore, if we let $P_{\min}=\min\{pMS_0^{\mathrm{Hom}},pMS_{0.1^2}^{\mathrm{Hom}},\ldots,pMS_{0.5^2}^{\mathrm{Hom}},pMS_1^{\mathrm{Hom}}\}$, we have

$$\begin{aligned}&P\left(MS_{\rho_0}^{\mathrm{Hom}}\le Q_{\rho_0}(P_{\min}),\ldots,MS_{\rho_L}^{\mathrm{Hom}}\le Q_{\rho_L}(P_{\min})\right)\\&\quad=E\{P(MS_{\rho_0}^{\mathrm{Hom}}\le Q_{\rho_0}(P_{\min}),\ldots,\mathrm{MS}_{\rho_L}^{\mathrm{Hom}}\le Q_{\rho_L}(P_{\min})|\mathbf{u}^t\mathbf{\Pi}\mathbf{u}=\eta)\}.\end{aligned}$$

Conditional probability can be numerically calculated as was suggested by Lee et al (Lee, Emond et al. 2012, Lee, Wu et al. 2012):

$$P(MS_{\rho_0}^{\mathrm{Hom}}\le Q_{\rho_0}(P_{\min}),\ldots,MS_{\rho_L}^{\mathrm{Hom}}\le Q_{\rho_L}(P_{\min})|\mathbf{u}^t\mathbf{\Pi}\mathbf{u}=\eta).$$

Numerical algorithm to calculate $pmFARVAT_o^{Het}$

We assume

$$\boldsymbol{Z}=\mathrm{var}(\mathrm{vec}(\mathbf{S}))^{1/2}(\mathbf{I}_Q\otimes\mathbf{W}),\ \text{and}\ \overline{\boldsymbol{Z}}=\boldsymbol{Z}\mathbf{1}_{MQ}(\mathbf{1}_{MQ}^t\mathbf{1}_{MQ})^{-1}.$$

Then the projection matrix on a space spanned by $\overline{\mathbf{Z}}$ is

$$\mathbf{\Pi}=\overline{\boldsymbol{Z}}(\overline{\boldsymbol{Z}}^t\overline{\boldsymbol{Z}})\overline{\boldsymbol{Z}}^t.$$

If we let

$$\mathbf{u}=\mathrm{var}(\mathrm{vec}(\mathbf{S}))^{-1/2}\mathrm{vec}(\mathbf{S}), \mathbf{u}\sim MVN(0,\mathbf{I}_{MQ}),$$

MS_{ρ}^{Het} becomes

$$MS_{\rho}^{\mathrm{Het}}=\mathbf{u}^{t}\mathrm{var}(\mathrm{vec}(\mathbf{S}))^{\frac{1}{2}}(\mathbf{I}_{Q}\otimes\mathbf{W})\mathrm{var}(\mathrm{vec}(\mathbf{S}))^{\frac{1}{2}}\mathbf{u}=(1-\rho)\mathbf{u}^{t}\boldsymbol{Z}\boldsymbol{Z}^{t}\mathbf{u}+\rho(MQ)^{2}\mathbf{u}^{t}\overline{\boldsymbol{Z}}\overline{\boldsymbol{Z}}^{t}\mathbf{u}.$$

As was suggested by Lee et al (Lee, Wu et al. 2012), if we let

$$\tau(\rho)=(MQ)^{2}\rho\overline{\boldsymbol{Z}}^{t}\overline{\boldsymbol{Z}}+\frac{(1-\rho)}{\overline{\boldsymbol{Z}}^{t}\overline{\boldsymbol{Z}}}\overline{\boldsymbol{Z}}^{t}\boldsymbol{Z}\boldsymbol{Z}^{t}\overline{\boldsymbol{Z}},$$

we have

$$MS_{\rho}^{\mathrm{Het}}=(1-\rho)\mathbf{u}^{t}(\mathbf{I}_{MQ}-\boldsymbol{\Pi})\boldsymbol{Z}\boldsymbol{Z}^{t}(\mathbf{I}_{MQ}-\boldsymbol{\Pi})\mathbf{u}+2(1-\rho)\mathbf{u}^{t}(\mathbf{I}_{MQ}-\boldsymbol{\Pi})\boldsymbol{Z}\boldsymbol{Z}^{t}\boldsymbol{\Pi}\mathbf{u}+\tau(\rho)\mathbf{u}^{t}\boldsymbol{\Pi}\mathbf{u},$$

Therefore, if we let $P_{\min}=\min\{pMS_{0}^{\mathrm{Het}}, pMS_{0.1^{2}}^{\mathrm{Het}},\ldots, pMS_{0.5^{2}}^{\mathrm{Het}}, pMS_{1}^{\mathrm{Het}}\}$, we have

$$\begin{aligned}&P\left(MS_{\rho_0}^{\mathrm{Het}}\le Q_{\rho_0}(P_{\min}),\ldots,MS_{\rho_L}^{\mathrm{Het}}\le Q_{\rho_L}(P_{\min})\right)\\&\quad=E\{P(MS_{\rho_0}^{\mathrm{Het}}\le Q_{\rho_0}(P_{\min}),\ldots,MS_{\rho_L}^{\mathrm{Het}}\le Q_{\rho_L}(P_{\min})|\mathbf{u}^{t}\boldsymbol{\Pi}\mathbf{u}=\eta)\}.\end{aligned}$$

$P(MS_{\rho_0}^{\mathrm{Het}}\le Q_{\rho_0}(P_{\min}),\ldots,MS_{\rho_L}^{\mathrm{Het}}\le Q_{\rho_L}(P_{\min})|\mathbf{u}^{t}\boldsymbol{\Pi}\mathbf{u}=\eta)$ can be calculated as in (Lee, Emond et al. 2012, Lee, Wu et al. 2012).

References

Choi S, Lee S, Cichon S, Nothen MM, Lange C, Park T, Won S. FARVAT: a family-based rare variant association test. Bioinformatics. 2014; 30(22):3197–3205. [PubMed: 25075118]

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Durbin R. G. Genomes Project Analysis. The variant call format and VCFtools. Bioinformatics. 2011; 27(15):2156–2158. [PubMed: 21653522]

Gilmour AR, Anderson RD, Rae A. The analysis of binomial data by a generalized linear mixed model. Biometrika. 1985; 72:539–599.

Gray-McGuire C, Bochud M, Goodloe R, Elston RC. Genetic association tests: a method for the joint analysis of family and case-control data. Hum Genomics. 2009; 4(1):2–20. [PubMed: 19951892]

Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000; 19(Suppl 1):S36–S42. [PubMed: 11055368]

Lange C, DeMeo DL, Laird NM. Power and design considerations for a general class of family-based association tests: quantitative traits. Am J Hum Genet. 2002; 71(6):1330–1341. [PubMed: 12454799]

Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, N. G. E. S. P.-E. L. P. 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(2):224–237. [PubMed: 22863193]

Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. Am J Hum Genet. 2013; 93(1):42–53. [PubMed: 23768515]

Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. Biostatistics. 2012; 13(4):762–775. [PubMed: 22699862]

Lee Y, Park S, Moon S, Lee J, Elston RC, Lee W, Won S. On the analysis of a repeated measure design in genome-wide association analysis. Int J Environ Res Public Health. 2014; 11(12):12283–12303. [PubMed: 25464127]

Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet. 2009; 5(2):e1000384. [PubMed: 19214210]

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM. Finding the missing heritability of complex diseases. Nature. 2009; 461(7265):747–753. [PubMed: 19812666]

McPeek MS, Wu X, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004; 60(2):359–367. [PubMed: 15180661]

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81(3):559–575. [PubMed: 17701901]

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 2005; 15(11):1576–1583. [PubMed: 16251467]

Schall R. Estimation in generalized linear models with random effects. Biometrika. 1991; 78:719–727.

Schifano ED, Li L, Christiani DC, Lin X. Genome-wide association analysis for multiple continuous secondary phenotypes. Am J Hum Genet. 2013; 92(5):744–759. [PubMed: 23643383]

Shi G, Rao DC. Optimum designs for next-generation sequencing to discover rare variants for common complex disease. Genet Epidemiol. 2011; 35(6):572–579. [PubMed: 21618604]

Silverman EK, Chapman HA, Drazen JM, Weiss ST, Rosner B, Campbell EJ, O'Donnell WJ, Reilly JJ, Ginns L, Mentzer S, Wain J, Speizer FE. Genetic epidemiology of severe, early-onset chronic obstructive pulmonary disease. Risk to relatives for airflow obstruction and chronic bronchitis. Am J Respir Crit Care Med. 1998; 157(6 Pt 1):1770–1778. [PubMed: 9620904]

Thornton T, McPeek MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. Am J Hum Genet. 2007; 81(2):321–337. [PubMed: 17668381]

Thornton T, Tang H, Hoffmann TJ, Ochs-Balcom HM, Caan BJ, Risch N. Estimating kinship in admixed populations. Am J Hum Genet. 2012; 91(1):122–138. [PubMed: 22748210]

van der Sluis S, Posthuma D, Dolan CV. TATES: efficient multivariate genotype-phenotype analysis for genome-wide association studies. PLoS Genet. 2013; 9(1):e1003235. [PubMed: 23359524]

Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. Genet Epidemiol. 2013; 37(8):778–786. [PubMed: 24166731]

Won S, Kim W, Lee S, Lee Y, Sung J, Park T. Family-based association analysis: a fast and efficient method of multivariate association analysis with multiple variants. BMC Bioinformatics. 2015; 16:46. [PubMed: 25887481]

Won S, Lange C. A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. Stat Med. 2013

Won S, Lange C. A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. Stat Med. 2013; 32(25):4482–4498. [PubMed: 23740776]

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet. 2011; 89(1):82–93. [PubMed: 21737059]

Yip W, De G, Raby AB, Laird N. Identifying causal rare variants of disease through family-based analysis of Genetics Analysis Workshop 17 data set. BMC Proceedings. 2011

Table 1

Empirical power estimates when all rare variants are causal and 100% of them are deleterious

Empirical power of $mFARVAT_S^{\text{Het}}$, $mFARVAT_B^{\text{Het}}$, $mFARVAT_O^{\text{Het}}$, $mFARVAT_B^{\text{Hom}}$, $mFARVAT_S^{\text{Hom}}$ and $mFARVAT_O^{\text{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes ($Q = 2$ and $Q = 5$) with homogeneous and heterogeneous effects and different correlations ($c = 0.5$ and $c = 0.8$) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			***mFARVAT^{Het}***			***mFARVAT^{Hom}***		
Q	**Type**	***c***	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.208	0.712	0.738	0.331	0.908	0.912	0.337	0.896	0.900
			Hom	0.196	0.766	0.778	0.353	0.928	0.927	0.439	0.915	0.925
		0.8	Het	0.200	0.713	0.723	0.310	0.876	0.875	0.290	0.865	0.859
			Hom	0.201	0.705	0.729	0.333	0.865	0.874	0.373	0.853	0.874
	Q	0.5	Het	0.350	0.987	0.987	0.531	0.998	0.998	0.593	0.999	0.998
			Hom	0.396	0.984	0.979	0.574	0.998	0.998	0.755	0.996	0.997
		0.8	Het	0.251	0.980	0.979	0.490	0.995	0.999	0.486	0.995	0.995
			Hom	0.365	0.977	0.977	0.509	0.996	0.995	0.607	0.996	0.995
5	D	0.5	Het	0.317	0.924	0.934	0.839	1.000	1.000	0.826	1.000	1.000
			Hom	0.315	0.948	0.955	0.868	1.000	1.000	0.947	1.000	1.000
		0.8	Het	0.267	0.887	0.900	0.706	0.991	0.995	0.635	0.990	0.992
			Hom	0.265	0.893	0.914	0.756	0.995	0.995	0.814	0.995	0.995
	Q	0.5	Het	0.540	0.998	0.998	0.952	1.000	1.000	0.973	1.000	1.000
			Hom	0.602	1.000	1.000	0.968	1.000	1.000	0.999	1.000	1.000
		0.8	Het	0.495	0.992	0.993	0.879	1.000	1.000	0.836	1.000	1.000
			Hom	0.525	0.994	0.994	0.890	1.000	1.000	0.957	1.000	1.000

Table 2

Empirical power estimates when all rare variants are causal and 80% of them are deleterious

Empirical power of $mFARVAT_S^{\text{Het}}$, $mFARVAT_B^{\text{Het}}$, $mFARVAT_O^{\text{Het}}$, $mFARVAT_B^{\text{Hom}}$, $mFARVAT_S^{\text{Hom}}$ and $mFARVAT_O^{\text{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes ($Q = 2$ and $Q = 5$) with homogeneous and heterogeneous effects and different correlations ($c = 0.5$ and $c = 0.8$) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			***mFARVAT^{Het}***			***mFARVAT^{Hom}***		
nphe	**Type**	**Cor**	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.129	0.148	0.289	0.231	0.327	0.464	0.135	0.302	0.372
			Hom	0.150	0.194	0.326	0.252	0.389	0.525	0.342	0.356	0.550
		0.8	Het	0.111	0.146	0.270	0.191	0.263	0.422	0.092	0.242	0.316
			Hom	0.133	0.164	0.283	0.212	0.313	0.447	0.264	0.279	0.462
	Q	0.5	Het	0.355	0.414	0.627	0.523	0.592	0.808	0.301	0.581	0.692
			Hom	0.368	0.467	0.670	0.546	0.678	0.854	0.718	0.660	0.892
		0.8	Het	0.331	0.376	0.608	0.451	0.491	0.736	0.190	0.492	0.561
			Hom	0.319	0.407	0.608	0.461	0.564	0.753	0.577	0.536	0.790
5	D	0.5	Het	0.214	0.270	0.479	0.707	0.763	0.903	0.272	0.745	0.764
			Hom	0.228	0.328	0.512	0.750	0.844	0.931	0.887	0.814	0.952
		0.8	Het	0.179	0.215	0.386	0.629	0.590	0.819	0.143	0.562	0.586
			Hom	0.195	0.290	0.453	0.622	0.705	0.855	0.742	0.672	0.881
	Q	0.5	Het	0.577	0.574	0.831	0.962	0.922	0.997	0.459	0.923	0.931
			Hom	0.546	0.643	0.839	0.934	0.961	0.997	0.992	0.954	1.000
		0.8	Het	0.527	0.490	0.765	0.915	0.802	0.972	0.238	0.791	0.804
			Hom	0.477	0.566	0.773	0.865	0.846	0.971	0.953	0.826	0.989

Table 3

Empirical power estimates when all rare variants are causal and 50% of them are deleterious

Empirical power of $mFARVAT_S^{\text{Het}}$, $mFARVAT_B^{\text{Het}}$, $mFARVAT_O^{\text{Het}}$, $mFARVAT_B^{\text{Hom}}$, $mFARVAT_S^{\text{Hom}}$ and $mFARVAT_O^{\text{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes ($Q = 2$ and $Q = 5$) with homogeneous and heterogeneous effects and different correlations ($c = 0.5$ and $c = 0.8$) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			***mFARVAT^{Het}***			***mFARVAT^{Hom}***		
nphe	**Type**	**Cor**	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.038	0.000	0.028	0.069	0.000	0.048	0.020	0.000	0.016
			Hom	0.050	0.000	0.031	0.120	0.003	0.081	0.181	0.002	0.133
		0.8	Het	0.064	0.000	0.038	0.087	0.000	0.068	0.016	0.000	0.010
			Hom	0.052	0.003	0.039	0.100	0.007	0.082	0.127	0.004	0.101
	Q	0.5	Het	0.330	0.000	0.236	0.489	0.001	0.386	0.121	0.001	0.070
			Hom	0.335	0.001	0.240	0.481	0.003	0.405	0.657	0.005	0.566
		0.8	Het	0.341	0.001	0.246	0.431	0.000	0.327	0.103	0.000	0.064
			Hom	0.312	0.002	0.230	0.410	0.008	0.329	0.533	0.007	0.433
5	D	0.5	Het	0.067	0.001	0.038	0.409	0.001	0.320	0.043	0.000	0.024
			Hom	0.065	0.002	0.105	0.499	0.018	0.434	0.763	0.009	0.687
		0.8	Het	0.073	0.001	0.036	0.382	0.000	0.282	0.007	0.000	0.006
			Hom	0.060	0.000	0.044	0.381	0.009	0.305	0.557	0.003	0.445
	Q	0.5	Het	0.529	0.001	0.365	0.944	0.000	0.906	0.043	0.000	0.024
			Hom	0.472	0.001	0.333	0.883	0.018	0.836	0.983	0.012	0.972
		0.8	Het	0.543	0.000	0.371	0.913	0.000	0.866	0.019	0.000	0.012
			Hom	0.411	0.001	0.277	0.817	0.008	0.744	0.918	0.005	0.875

Table 4

Empirical power estimates when half rare variants are causal and 100% of them are deleterious

Empirical power of $mFARVAT_S^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Het}}$, $mFARVAT_O^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Hom}}$, $mFARVAT_S^{\mathrm{Hom}}$ and $mFARVAT_O^{\mathrm{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlation (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			*mFARVAT*^{Het}			*mFARVAT*^{Hom}		
nphe	**Type**	**Cor**	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.207	0.273	0.427	0.340	0.523	0.663	0.219	0.488	0.569
			Hom	0.259	0.298	0.484	0.403	0.578	0.712	0.502	0.536	0.729
		0.8	Het	0.425	0.634	0.788	0.572	0.831	0.910	0.379	0.826	0.863
			Hom	0.488	0.684	0.834	0.678	0.857	0.937	0.812	0.851	0.948
	Q	0.5	Het	0.178	0.254	0.419	0.285	0.422	0.575	0.139	0.382	0.461
			Hom	0.248	0.283	0.462	0.390	0.497	0.647	0.420	0.460	0.631
		0.8	Het	0.400	0.602	0.754	0.521	0.759	0.877	0.251	0.752	0.794
			Hom	0.434	0.646	0.767	0.562	0.781	0.883	0.645	0.768	0.901
5	D	0.5	Het	0.294	0.427	0.630	0.839	0.921	0.977	0.417	0.908	0.910
			Hom	0.375	0.512	0.722	0.886	0.963	0.990	0.952	0.951	0.995
		0.8	Het	0.609	0.807	0.920	0.974	0.997	0.999	0.645	0.997	0.997
			Hom	0.665	0.845	0.944	0.977	0.999	1.000	0.999	0.998	1.000
	Q	0.5	Het	0.266	0.383	0.582	0.729	0.817	0.917	0.276	0.812	0.817
			Hom	0.328	0.464	0.651	0.773	0.867	0.947	0.854	0.835	0.960
		0.8	Het	0.595	0.759	0.901	0.900	0.961	0.991	0.405	0.955	0.956
			Hom	0.631	0.782	0.911	0.919	0.973	0.996	0.963	0.969	0.998

Table 5

Empirical power estimates when half rare variants are causal and 80% of them are deleterious

Empirical power of $mFARVAT_S^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Het}}$, $mFARVAT_O^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Hom}}$, $mFARVAT_S^{\mathrm{Hom}}$ and $mFARVAT_O^{\mathrm{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes (Q = 2 and Q = 5) with homogeneous and heterogeneous effects and different correlation (c = 0.5 and c = 0.8) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			*mFARVAT*^{Het}			*mFARVAT*^{Hom}		
nphe	**Type**	**Cor**	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.128	0.043	0.174	0.215	0.105	0.306	0.098	0.091	0.182
			Hom	0.167	0.055	0.217	0.295	0.165	0.392	0.369	0.132	0.419
		0.8	Het	0.392	0.114	0.453	0.565	0.215	0.640	0.219	0.200	0.384
			Hom	0.413	0.138	0.490	0.601	0.301	0.705	0.751	0.272	0.793
	Q	0.5	Het	0.112	0.045	0.164	0.169	0.079	0.238	0.072	0.062	0.135
			Hom	0.159	0.052	0.203	0.240	0.133	0.327	0.301	0.112	0.348
		0.8	Het	0.375	0.112	0.410	0.469	0.152	0.526	0.137	0.135	0.267
			Hom	0.391	0.145	0.477	0.514	0.245	0.611	0.625	0.223	0.672
5	D	0.5	Het	0.184	0.059	0.245	0.703	0.317	0.769	0.118	0.288	0.363
			Hom	0.254	0.108	0.345	0.773	0.518	0.848	0.907	0.458	0.913
		0.8	Het	0.581	0.152	0.604	0.968	0.469	0.975	0.209	0.452	0.568
			Hom	0.612	0.267	0.696	0.953	0.690	0.977	0.996	0.662	0.997
	Q	0.5	Het	0.237	0.049	0.194	0.612	0.211	0.651	0.062	0.187	0.234
			Hom	0.237	0.090	0.311	0.669	0.353	0.736	0.779	0.292	0.789
		0.8	Het	0.581	0.135	0.568	0.912	0.321	0.927	0.094	0.305	0.352
			Hom	0.573	0.197	0.631	0.875	0.512	0.911	0.943	0.459	0.953

Table 6

Empirical power estimates when half rare variants are causal and 50% of them are deleterious

Empirical power of $mFARVAT_S^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Het}}$, $mFARVAT_O^{\mathrm{Het}}$, $mFARVAT_B^{\mathrm{Hom}}$, $mFARVAT_S^{\mathrm{Hom}}$ and $mFARVAT_O^{\mathrm{Hom}}$ was calculated for dichotomous and quantitative multiple phenotypes ($Q = 2$ and $Q = 5$) with homogeneous and heterogeneous effects and different correlation ($c = 0.5$ and $c = 0.8$) at the 10^{-4} significant level. Empirical power of *FARVAT* was calculated by adopting Bonferroni correction to the minimum p-value of univariate association tests on multiple phenotypes.

				FARVAT			***mFARVAT^{Het}***			***mFARVAT^{Hom}***		
nphe	**Type**	**Cor**	**Eff**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**	**SKAT**	**Burden**	**SKAT-O**
2	D	0.5	Het	0.041	0.000	0.024	0.094	0.000	0.064	0.024	0.000	0.011
			Hom	0.054	0.001	0.036	0.148	0.004	0.113	0.196	0.001	0.139
		0.8	Het	0.343	0.001	0.239	0.500	0.001	0.403	0.127	0.000	0.090
			Hom	0.341	0.004	0.246	0.500	0.004	0.421	0.677	0.004	0.568
	Q	0.5	Het	0.050	0.000	0.036	0.071	0.001	0.050	0.020	0.000	0.012
			Hom	0.062	0.001	0.037	0.118	0.000	0.088	0.139	0.001	0.102
		0.8	Het	0.352	0.001	0.253	0.426	0.000	0.327	0.087	0.000	0.062
			Hom	0.333	0.001	0.232	0.430	0.002	0.356	0.554	0.002	0.448
5	D	0.5	Het	0.079	0.000	0.053	0.420	0.001	0.318	0.017	0.000	0.018
			Hom	0.130	0.000	0.086	0.532	0.014	0.468	0.771	0.009	0.703
		0.8	Het	0.535	0.003	0.367	0.940	0.001	0.891	0.046	0.000	0.031
			Hom	0.508	0.000	0.364	0.900	0.014	0.848	0.977	0.007	0.963
	Q	0.5	Het	0.075	0.010	0.042	0.363	0.000	0.264	0.002	0.000	0.002
			Hom	0.115	0.003	0.075	0.449	0.012	0.389	0.594	0.009	0.505
		0.8	Het	0.562	0.001	0.377	0.901	0.000	0.841	0.017	0.000	0.010
			Hom	0.466	0.003	0.338	0.815	0.014	0.759	0.914	0.010	0.874

Genet Epidemiol. Author manuscript; available in PMC 2017 September 01.

Table 7

***mFARVAT* analysis of COPD-releted phenotypes**

Genes are the top 10 most significant results from $mFARVAT_o^{Het}$ and $mFARVAT_o^{Hom}$.

method	chr	gene	MAC	N. of variants	p-value
Het	11	*KRTAP5-9*	21	3	1.00×10^{-04}
	13	*DIAPH3*	40	7	1.73×10^{-04}
	4	*ENAM*	82	9	3.16×10^{-04}
	2	*SLC8A1*	5	3	3.38×10^{-04}
	3	*MFI2*	32	5	4.30×10^{-04}
	11	*PLEKHA7*	20	9	5.16×10^{-04}
	2	*SLC19A3*	11	4	6.88×10^{-04}
	7	*ZNF736*	8	2	7.94×10^{-04}
	15	*MGA*	49	11	9.08×10^{-04}
	8	*CA1*	7	2	1.18×10^{-03}
Hom	13	*DIAPH3*	40	7	1.25×10^{-04}
	2	*SLC8A1*	5	3	1.80×10^{-04}
	11	*PLEKHA7*	20	9	2.18×10^{-04}
	11	*KRTAP5-9*	21	3	2.72×10^{-04}
	15	*POLG*	58	8	6.28×10^{-04}
	2	*SLC19A3*	11	4	6.37×10^{-04}
	1	*ETV3L*	31	5	6.63×10^{-04}
	7	*ZNF736*	8	2	7.94×10^{-04}
	5	*AFAP1L1*	20	3	7.95×10^{-04}
	3	*ANO10*	32	3	9.57×10^{-04}