frontiers in Genetics

ORIGINAL RESEARCH
published: 19 June 2019
doi: 10.3389/fgene.2019.00572

metaFARVAT: An Efficient Tool for Meta-Analysis of Family-Based, Case-Control, and Population-Based Rare Variant Association Studies

Longfei Wang[1], Sungyoung Lee[2], Dandi Qiao[3], Michael H. Cho[3,4], Edwin K. Silverman[3,4], Christoph Lange[3,5] and Sungho Won[1,6,7]*

[1] Interdisciplinary Program in Bioinformatics, Seoul National University, Seoul, South Korea, [2] Center for Precision Medicine, Seoul National University Hospital, Seoul, South Korea, [3] Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, United States, [4] Division of Pulmonary and Critical Care Medicine, Brigham and Women's Hospital, Boston, MA, United States, [5] Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, United States, [6] Department of Public Health Sciences, Seoul National University, Seoul, South Korea, [7] Institute of Health and Environment, Seoul National University, Seoul, South Korea

Family-based designs have been shown to be powerful in detecting the significant rare variants associated with human diseases. However, very few significant results have been found owing to relatively small sample sizes and the fact that statistical analyses often suffer from high false-negative error rates. These limitations can be avoided by combining results from multiple studies via meta-analysis. However, statistical methods for meta-analysis with rare variants are limited for family-based samples. In this report, we propose a tool for the meta-analysis of family-based rare variant associations, metaFARVAT. metaFARVAT is based on a quasi-likelihood score for each variant. These scores are combined to generate burden test, variable-threshold test, sequence kernel association test (SKAT), and optimal SKAT statistics. The proposed method tests homogeneous and heterogeneous effects of variants among different studies and can be applied to both quantitative and dichotomous phenotypes. Simulation results demonstrated the robustness and efficiency of the proposed method in different scenarios. By applying metaFARVAT to data from a family-based study and a case-control study, we identified a few promising candidate genes, including *DLEC1,* which is associated with chronic obstructive pulmonary disease.

Keywords: meta-analysis, family-based design, rare variant association analysis, metaFARVAT, chronic obstructive pulmonary disease

OPEN ACCESS

Edited by:
Mariza De Andrade,
Mayo Clinic, United States

Reviewed by:
Karim Oualkacha,
Université du Québec à Montréal, Canada
Alexandre Bureau,
Laval University, Canada

***Correspondence:**
Sungho Won
sunghow@gmail.com

Specialty section:
This article was submitted to Statistical Genetics and Methodology, a section of the journal Frontiers in Genetics

Received: *26 October 2018*
Accepted: *31 May 2019*
Published: *19 June 2019*

Citation:
Wang L, Lee S, Qiao D, Cho MH, Silverman EK, Lange C and Won S (2019) metaFARVAT: An Efficient Tool for Meta-Analysis of Family-Based, Case-Control, and Population-Based Rare Variant Association Studies. Front. Genet. 10:572. doi: 10.3389/fgene.2019.00572

INTRODUCTION

In recent decades, genome-wide association studies (GWAS) have identified tens of thousands of common variants associated with various complex diseases. However, in spite of their success in discovering disease susceptibility loci (DSL), the DSL identified by GWAS only partially explain disease heritability, and rare variants have been implicated as one contributor to this missing heritability (Manolio et al., 2009). Recent improvements in sequencing technology have enabled rare variant association analyses, and various methods have been proposed for rare variant association studies, such as the combined multivariate and collapsing (CMC) method, burden test, variable-threshold test (VT) (Price et al., 2010), sequence kernel association test (SKAT) (Wu et al., 2011), and optimal SKAT (SKAT-O) (Lee et al., 2012b).

Multiple rare variants can affect disease status, and thus, association analyses with rare variants suffer from genetic heterogeneity among affected individuals. In families, Mendelian transmission results in family members sharing the same alleles, and therefore, affected relatives have a greater chance of being affected by the same disease-causing single-nucleotide polymorphisms (SNPs) than unrelated subjects. For instance, the probability of sibling pairs sharing rare alleles can be calculated (Ionita-Laza et al., 2011). Therefore, family-based analyses have been generally recognized as an important strategy for rare variant association studies. We proposed FARVAT (Choi et al., 2014) statistics based on quasi-likelihood. This includes burden, SKAT, and SKAT-O statistics for both dichotomous and quantitative phenotypes, and we have shown that they are robust against population substructure and outperform other existing rare variant association tests for family samples (Wang et al., 2016).

Aggregation of association signals across multiple genetic variants was expected to provide sufficient statistical power for rare variant analyses and to identify various DSL. However, very few genome-wide significant results have been found because of relatively small sample sizes. When the sample size is small, statistical analyses suffer from high false-negative error rates, and this limitation can be avoided by combining results from multiple studies via mega- or meta-analysis. Mega-analysis assumes that subjects' genotypes and phenotypes from different studies are available, and these are pooled for genetic association analyses. Meta-analysis directly utilizes test statistics from separate studies and combines them into a single test statistic. The choice between mega- and meta-analysis depends on the heterogeneity among studies and the availability of individual genotype and phenotype data from all studies. In particular, if there are systematic differences in phenotype diagnosis or sequencing technology, meta-analysis is often preferred. Otherwise, mega-analysis is recommended if genotypes and phenotypes are available. Recently, several meta-analysis methods for rare variant association tests have been proposed, such as MASS (Tang and Lin, 2013, 2014), RAREMETAL (Feng et al., 2014; Liu et al., 2014), seqMeta (Chen et al., 2014), and MetaSKAT (Lee et al., 2013). However, the available statistical methods for family-based samples or dichotomous phenotypes are limited, and it is moreover worthwhile to provide a method that can be applied to both quantitative and dichotomous phenotypes under homogeneous and heterogeneous disease models.

In this study, we proposed a new meta-analysis method for family-based, population-based, and case-control rare variant association tests, metaFARVAT. metaFARVAT generates a quasi-likelihood score for each variant and combines them to generate burden, VT, SKAT, and SKAT-O statistics. metaFARVAT can assume homogeneous or heterogeneous effects of variants among different studies and can be applied to both quantitative and dichotomous phenotypes. We evaluated the statistical validity of metaFARVAT using simulated data and compared its estimated power with those of RAREMETAL and seqMeta under various scenarios. Furthermore, metaFARVAT was applied to identify rare variants for chronic obstructive pulmonary disease (COPD) using whole-exome sequencing (WES) data from family-based samples from the Boston Early-Onset COPD Study (EOCOPD) and case-control samples from the COPDGene study.

METHODS

Notations and Disease Model

We assume that there are K studies available and that each study is of either a population-based, case-control, or family-based design. It is assumed that N_k subjects are available in study k. We assume that there are M rare variants in a gene, and the minor allele count of variant m for subject i in study k is coded by x_{imk}. Traits can be either quantitative or dichotomous, and y_{ik} indicates a phenotype of subject i in study k. Their vectors are denoted by:

$$\mathbf{X}_k^j = \begin{bmatrix} x_{1jk} \\ \vdots \\ x_{N_k jk} \end{bmatrix},\ \mathbf{X}_k = \left(\mathbf{X}_k^1, \cdots, \mathbf{X}_k^M\right),\ \mathbf{Y}_k = \begin{bmatrix} y_{1k} \\ \vdots \\ y_{N_k k} \end{bmatrix}.$$

In some cases, rare variants may be observed only in a subset of studies. If variant m is missing or monomorphic in study k, we assume that $\mathbf{X}_k^m$ is $\mathbf{0}$, and its variance and covariance with $\mathbf{X}_k^{m'}$ $\left(m \neq m'\right)$ are 0. If variant m is missing for all studies, then it should be removed from the analysis.

Parental genotypes are transmitted to offspring under Mendelian transmission, and thus our test statistics consider the genetic correlation between family members. The genetic variance-covariance matrix among family members can be specified by a kinship coefficient matrix, $\mathbf{\Phi}_k$. If we let $\pi_{i,i',k}$ be the kinship coefficient between subject i and subject i' for study k, and d_{ik} be the inbreeding coefficient for subject i, $\mathbf{\Phi}_k$ is defined by:

$$\begin{bmatrix} 1+d_{1,k} & 2\pi_{1,2,k} & 2\pi_{1,3,k} & \cdots \\ 2\pi_{1,2,k} & 1+d_{2,k} & 2\pi_{2,3,k} & \cdots \\ 2\pi_{1,3,k} & 2\pi_{2,3,k} & 1+d_{3,k} & \ddots \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix}$$

Under the presence of population substructure, the genetic relationship matrix (GRM) can be estimated with large-scale genotyping data and should alternatively replace $\mathbf{\Phi}_k$ (Thornton et al., 2012).

Last, meta-analysis of rare variant association analyses with multiple studies requires two different types of weights. First, when multiple studies are combined, each study has different features, such as sample size and disease diagnosis, and such differences can be handled with an *a priori* specified weight for each study. We assume that the statistics for study k are weighted by v_k, and their $K \times K$ dimensional diagonal matrix is denoted by $\mathbf{W}_B$. Second, rare variants have different gene annotations, genomic coordinates, and functional characterization, and various annotation tools have been proposed to choose important features based on their biological properties. We denote the weight for rare variant m by w_m, and we let $\mathbf{W}_W$ be their $M \times M$ dimensional diagonal matrix.

Choices of Offset

We introduce the offset μ_{ik} for subject i at study k to improve the efficiency of the proposed score test (Lange et al., 2002). We set:

$$\boldsymbol{\mu}_k = \begin{bmatrix} \mu_{1k} \\ \vdots \\ \mu_{N_k k} \end{bmatrix}, \ \boldsymbol{\mu} = \left(\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\right), \ \mathbf{T}_k = \mathbf{Y}_k - \boldsymbol{\mu}_k.$$

The most efficient choice of $\boldsymbol{\mu}$ may depend on the sampling scheme, and either the best linear unbiased predictor (BLUP) with covariates or the prevalence were shown to be the most efficient (Won and Lange, 2013). If families are randomly selected, BLUP was shown to be the most efficient (Won and Lange, 2013); otherwise, the prevalence is recommended for dichotomous phenotypes (Thornton and McPeek, 2007; Won and Lange, 2013). In this report, we focus on randomly selected families, and we incorporate BLUP from the linear mixed model for $\boldsymbol{\mu}$. Under the null hypothesis, the linear mixed model (George and Elston, 1987) for a quantitative phenotype is given by:

$$\mathbf{Y} = \mathbf{Z}\alpha + \mathbf{B} + \mathbf{E},\ \mathbf{B} \sim MVN\left(0, \sigma_b^2 \Phi\right)\ and\ \mathbf{E} \sim MVN\left(0, \sigma_e^2 \mathbf{I}_N\right),$$

where $\mathbf{B}$ and $\mathbf{E}$ indicate the polygenetic random effect and random error, respectively. Then, incorporation of BLUP as an offset gives:

$$\mathbf{T} = \mathbf{Y} - \boldsymbol{\mu} = \left(\mathbf{I} - \mathbf{Z}\left(\mathbf{Z^tH^{-1}Z}\right)^{-1} \mathbf{Z^tH^{-1}} - \hat{\sigma}_b^2 \Phi \mathbf{P}\right) \mathbf{Y},$$

where $\mathbf{H} = \hat{\sigma}_b^2\Phi + \hat{\sigma}_e^2\mathbf{I}_N$, and $\mathbf{P} = \mathbf{H^{-1}} - \mathbf{H^{-1}Z}\left(\mathbf{Z^tH^{-1}Z}\right)\mathbf{^{-1}Z^tH^{-1}}$. For a dichotomous phenotype, use of the generalized linear mixed model might be considered an appropriate approach, but we estimated $\mathbf{T}$ in the same way for quantitative phenotypes when individuals were randomly selected because of its superior statistical power (Won and Lange, 2013).

Score for Quasi-Likelihood

We let $\mathbf{1}$ w be the $w \times 1$ column vector, of which the elements are one. The score based on quasi-likelihood for variant m in study k is defined by:

$$u_{m,k} = \mathbf{T}_k^t \left(\mathbf{I}_{N_k} - \mathbf{1}_{N_k}\left(\mathbf{1}_{N_k}^t \Phi_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t \Phi_k^{-1}\right) \mathbf{X}_k^m.$$

If we denote the covariance between $x_{m,k}$ and $x_{m',k}$ by $\sigma_{mm',k}$, then $cov\left(\mathbf{X}_k^m, \mathbf{X}_k^{m'}\right) = \sigma_{mm',k}\Phi_k$, and $\sigma_{mm',k}$ is estimated by the empirical covariance. We let:

$$\boldsymbol{\Sigma_k} = \begin{bmatrix} \sigma_{11,k} & \cdots & \sigma_{1M,k} \\ \vdots & \ddots & \vdots \\ \sigma_{M1,k} & \cdots & \sigma_{MM,k} \end{bmatrix}.$$

If we let $\mathbf{H}_k = \Phi_k - \mathbf{1}_{N_k}\left(\mathbf{1}_{N_k}^t \Phi_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t$, the variance-covariance matrix of $u_{m,k}$(Choi et al., 2014) was shown to be:

$$var\begin{bmatrix} \mathbf{T}_k^t\left(\mathbf{X}_k^1 - \hat{E}\left(\mathbf{X}_k^1\right)\right) \\ \vdots \\ \mathbf{T}_k^t\left(\mathbf{X}_k^M - \hat{E}\left(\mathbf{X}_k^M\right)\right) \end{bmatrix} = \left(\mathbf{T}_k^t \mathbf{H}_k \mathbf{T}_k\right) \boldsymbol{\Sigma}_k.$$

The score vector of rare variants in study k can be defined by:

$$\mathbf{U}_k = \frac{1}{\sqrt{\mathbf{T}_k^t \mathbf{H}_k \mathbf{T}_k}} \mathbf{T}_k^t \left(\mathbf{I}_{N_k} - \mathbf{1}_{N_k}\left(\mathbf{1}_{N_k}^t \Phi_k^{-1} \mathbf{1}_{N_k}\right)^{-1} \mathbf{1}_{N_k}^t \Phi_k^{-1}\right) \mathbf{X}_k.$$

The score statistic tests whether the coded genotypes are linearly independent from the phenotypes; for dichotomous phenotypes, it is equivalent to comparing the minor allele frequencies (MAFs) between cases and controls.

Homogeneous Model

The homogeneous model assumes that the effect sizes of each variant are expected to be similar among different studies, and thus the proposed scores for each study can be collapsed across studies as follows:

$$\mathbf{U}^{Hom} \equiv \sum_k \nu_k \mathbf{U}_k^t,\ \boldsymbol{\Sigma}^{Hom} \equiv var\left(\mathbf{U}^{Hom}\right) = \sum_k \nu_k^2 \boldsymbol{\Sigma}_k.$$

Here, we set ν_k to be one. However, the proposed statistics are sometimes unavailable, and the appropriate choice can vary according to the available information. For instance, if standardized test statistics and sample sizes are available, then the inverse function to the square root of the sample size can be utilized.

Rare variant association analysis can be categorized into burden and variance-component tests (Li and Leal, 2008; Price et al., 2010; Neale et al., 2011; Wu et al., 2011). The burden test is known to be the most powerful if all rare variants have either deleterious or protective effects on disease; otherwise, the variance-component test is more efficient (Neale et al., 2011). If we let χ_1^2 be a chi-square distribution with a single degree of freedom, the burden test for a homogeneous model becomes:

$$S_{burden}^{Hom} = \frac{\left(\mathbf{U}^{Hom}\right)^t \mathbf{W}_W \mathbf{1}_M \mathbf{1}_M^t \mathbf{W}_W \mathbf{U}^{Hom}}{\mathbf{1}_M^t \mathbf{W}_W \boldsymbol{\Sigma}^{Hom} \mathbf{W}_W \mathbf{1}_M} \sim \chi_1^2 \text{ under } H_0.$$

Variance component tests use the collapsed squared scores (Neale et al., 2011; Wu et al., 2011) and can be expressed by:

$$S_{SKAT}^{Hom} = \left(\mathbf{U}^{Hom}\right)^t \mathbf{W}_W \mathbf{I}_M \mathbf{W}_W \mathbf{U}^{Hom}.$$

We denote eigenvalues for $\left(\boldsymbol{\Sigma}^{Hom}\right)^{1/2} \mathbf{W}_W \mathbf{W}_W \left(\boldsymbol{\Sigma}^{Hom}\right)^{1/2}$ by λ_l. If we let $\chi_{1,l}^2$ be an independent chi-square distribution with a single degree of freedom, the variance component test for the homogeneous model follows:

$$S_{SKAT}^{Hom} \sim \sum_{l=1}^{M} \lambda_l \chi_{1,l}^2 \text{ under } H_0.$$

A balanced approach for both scenarios can be achieved by the SKAT-O type statistic (Lee et al., 2012a). For a certain c between 0 and 1, we consider:

$$\left(\mathbf{U}^{Hom}\right)^t \mathbf{W}_W \left((1-c)\,\mathbf{I}_M + c\mathbf{1}_M\mathbf{1}_M^t\right) \mathbf{W}_W \mathbf{U}^{Hom}.$$

If we let its p-value be pS_c^{Hom}, the SKAT-O type statistic for $c_0 = 0 < c_1 < \ldots < c_L = 1$ is defined by:

$$S_{SKATO}^{Hom} = p_{\min}^{Hom} = \min\left\{pS_0^{Hom}, pS_{0.01}^{Hom}, pS_{0.04}^{Hom}, pS_{0.09}^{Hom}, pS_{0.16}^{Hom}, pS_{0.25}^{Hom}, pS_1^{Hom}\right\}.$$

Its p-value can be calculated with the numerical algorithm for the FARVAT statistic (Choi et al., 2014).

Last, rare variant association analysis utilizes rare variants, but the definition of a rare variant is not clear. VT approaches are very useful in such scenarios. We assume that rare variants are sorted in ascending order of overall MAF. We let $\mathbf{1}_{(m)}$ be an M-dimensional column vector whose 1st,…, mth elements are 1 and the others are 0. If we let:

$$U_{(m)}^{Hom} = \sum_{k=1}^{K} \nu_k \mathbf{1}_{(m)}^t \mathbf{W}_W \mathbf{U}_k^t = \mathbf{1}_{(m\ )}^t \mathbf{W}_W \mathbf{U}^{Hom},$$

then the covariance between $U_{(m)}^{Hom}$ and $U_{(m')}^{Hom}$is:

$$\mathbf{1}_{(m)}^t \mathbf{W}_W \mathbf{\Sigma}^{Hom} \mathbf{W}_W \mathbf{1}_{(m')}.$$

Therefore, we let:

$$T_{(m)}^{Hom} = \frac{U_{(m)}^{Hom}}{\sqrt{\mathbf{1}_{(m)}^t \mathbf{W}_W \Sigma^{Hom} \mathbf{W}_W \mathbf{1}_{(m)}}}.$$

If we denote the realization of $T_{(m)}^{Hom}$ by $t_{(\mathrm{m})}$ and let $t_{(|\max|)} = \max\{\ |t_{(1)}|,\ldots,\ |t_{(M)}|\}$, the p-value for the VT method can be calculated by:

$$1 - P\left(|T_{(1)}^{Hom}| > t_{(|\max|)}, \cdots, |T_{(M)}^{Hom}| > t_{(|max|)}\right).$$

Here, $\left(\boldsymbol{T}_{(\mathbf{1})}^{\boldsymbol{Hom}},\cdots,\boldsymbol{T}_{(\boldsymbol{M})}^{\boldsymbol{Hom}}\right)^{\boldsymbol{t}}$ follows the multivariate normal distribution with mean 0 and the following correlation matrix:

$$\Psi^{\boldsymbol{Hom}} = \left(\Psi_{\boldsymbol{mm'}}^{\boldsymbol{Hom}}\right)_{\boldsymbol{M\times M}},$$

where $\Psi_{\boldsymbol{mm'}}^{Hom} = \dfrac{\mathbf{1}_{(\boldsymbol{m})}^{\boldsymbol{t}} \mathbf{W}_W \Sigma^{\boldsymbol{Hom}} \mathbf{W}_W \mathbf{1}_{(\boldsymbol{m'})}}{\sqrt{\left(\mathbf{1}_{(\boldsymbol{m})}^{\boldsymbol{t}} \mathbf{W}_W \Sigma^{\boldsymbol{Hom}} \mathbf{W}_W \mathbf{1}_{(\boldsymbol{m})}\right)\left(\mathbf{1}_{(\boldsymbol{m'})}^{\boldsymbol{t}} \mathbf{W}_W \Sigma^{\boldsymbol{Hom}} \mathbf{W}_W \mathbf{1}_{(\boldsymbol{m'})}\ \right)}}.$

Heterogeneous Model

As in the homogeneous model, we propose burden and variance component tests for the heterogeneous model. The heterogeneous model assumes that the effects of specific variants are heterogeneous among studies. If we let $E\left(u_{m,k}\right) = \beta_{mk}$, the null hypothesis can be expressed by $\beta_{11} = \ldots = \beta_{MK} = 0$, and we consider the following score vector and its variance matrix:

$$\begin{aligned}\mathbf{U}^{Het} \equiv \left(\nu_1\mathbf{U}_1 \ldots \nu_K\mathbf{U}_K\right)^t,\ \Sigma^{Het} &\equiv var\left(vec\left(\mathbf{U}\right)\right) \\ &= \sum_k \left(\nu_k^2 \Sigma_k \otimes \mathbf{e}_{kk}\right),\end{aligned}$$

where $\mathbf{e}_{kk}$ is a $K \times K$ dimensional matrix whose (k, k) element is 1 and the others are 0. Then, the burden test can be expressed as:

$$S_{burden}^{Het} = \frac{\mathbf{U}^{Het}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{1}_{MK}\mathbf{1}_{MK}^t\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{U}^{Het^t}}{\mathbf{1}_{MK}^t\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\Sigma^{Het}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{1}_{MK}} \sim \chi_1^2 \text{ under} H_0.$$

We let:

$$\begin{aligned}\mathbf{R}_c^{Het} &= (1-c)\,\mathbf{I}_{MK} + c\mathbf{1}_{MK}\mathbf{1}_{MK}^t,\ S_c^{Het} \\ &= \mathbf{U}^{Het}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{R}_c^{Het}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{U}^{Het^t},\end{aligned}$$

and we let $\left(\lambda_1^c, \ldots, \lambda_{MK}^c\right)$ be the eigenvalues of:

$$\sum_k \left(\Sigma_k \otimes \mathbf{e}_{kk}\right)\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{I}_{MK}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\sum_k \left(\Sigma_k \otimes \mathbf{e}_{kk}\right).$$

Then, S_c^{Het} follows:

$$S_c^{Het} \sim \sum_{l=1}^{MK} \lambda_l^c \chi_{1,l}^2 \text{ under } H_0.$$

Therefore, the variance component test is defined by:

$$\begin{aligned}S_{SKAT}^{Het} &= \mathbf{U}^{Het}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{I}_{MK}\left(\mathbf{W}_W \otimes \mathbf{W}_B\right)\mathbf{U}^{Het^t} \\ &= S_0^{Het} \sim \sum_{l=1}^{MK} \lambda_l^0 \chi_{1,l}^2 \text{ under } H_0.\end{aligned}$$

If we denote the p-value for S_c^{Het} by pS_c^{Het}, the SKAT-O-type statistic is defined by:

$$S_{SKATO}^{Het} = p_{\min}^{Het} = min\left\{pS_0^{Het}, pS_{0.01}^{Het}, pS_{0.04}^{Het}, pS_{0.09}^{Het}, pS_{0.16}^{Het}, pS_{0.25}^{Het}, pS_1^{Het}\right\},$$

and its p-value is also obtained by the numerical algorithm for the FARVAT statistic (Choi et al., 2014).

Simulation Model

The performance of metaFARVAT was evaluated via extensive simulation studies. metaFARVAT can be applied to population-based and case-control designs by calculating GRM among samples. Therefore, we only focused on family-based designs in our simulation studies and considered unbalanced families consisting of trios, nuclear families, and extended families with three generations; the family structures that we considered are presented in **Supplementary Figure 1**. The families for our simulations were randomly selected from these different family structures. To generate rare variants, 1,200 haplotypes with 50,000 base pairs were generated under a coalescent model using the software COSI (Schaffner et al., 2005). Each haplotype was generated by setting the mutation rate to 1.5×10^{-8}, and haplotypes were randomly chosen with replacement to build founder genotypes. We defined variants with MAFs < 0.01 as being rare, and 60 rare variants were randomly selected from their haplotypes. Then, non-founder haplotypes were chosen from their parents' haplotypes in Mendelian fashion under the assumption of no recombination.

Phenotypes were generated under the null and alternative hypotheses. Simulation of dichotomous phenotypes was performed using the liability threshold model. Once the quantitative phenotypes were generated, they were transformed into case-control status for dichotomous phenotypes. If quantitative phenotypes were larger than the threshold, they were considered affected and otherwise were considered unaffected. The threshold was chosen to preserve the assumed disease prevalence of 0.1. If the disease prevalence is misspecified, loss of statistical power is expected; however, it has been shown with simulation studies that the effect of misspecification is not very substantial (Won and Lange, 2013). To allow for the ascertainment bias of dichotomous phenotypes in our simulation studies, we assumed that families with at least one affected subject were selected for analysis.

Quantitative phenotypes were defined by summing the phenotypic mean, genetic effect, polygenic effect, main genetic effect, and random error, and we assumed there was no environmental effect shared between family members. The phenotypic mean was denoted by $\alpha = 0.3$. The polygenic effect for each founder was independently generated from $N(0, {\sigma_g}^2 = 1)$, and for non-founders, the average of maternal and paternal polygenic effects was combined with values independently sampled from $N(0, 0.5{\sigma_g}^2)$. Random error was independently sampled from $N(0, {\sigma_e}^2 = 1)$. Therefore, the heritability of the simulated trait is 0.5. The genetic effect at variant m in study k was the product of β_{mk} and the number of disease susceptibility alleles. To evaluate the type-1 error estimates, β_{mk} was assumed to be 0. To evaluate the statistical power estimates, if we let ${h_a}^2$ be the proportion of variance explained by rare variants, β_{mk} values were iteratively sampled with a two-step approach. $\beta_{mk}^{(0)}$ were first sampled from $U(0,1)$. Then, if we let:

$$\nu_k = \sqrt{\frac{\left(\sigma_g^2 + \sigma_e^2\right) h_a^2}{\left(1-h_a^2\right) \sum_{m=1}^{M} \left[\left(\beta_{mk}^{(0)}\right)^2 2p_m \left(1-p_m\right)\right]}},$$

β_{mk} values were sampled from the uniform distribution $U(0, \nu_k)$. This procedure was repeated until ν_k converged. We assumed that ${h_a}^2 = 0.01$. β_{mk} was generated from heterogeneous or homogeneous scenarios. For homogeneous scenarios, we assumed that the effects of each rare variant were in the same direction in all studies. For heterogeneous scenarios, the signs (±) of β_{mk} values sampled from $U(0,\nu_k)$ were chosen randomly.

Application to COPD Data

We considered previously reported WES data from Boston Early-Onset COPD Study (EOCOPD) families and COPDGene case-control subjects for meta-analysis (Qiao et al., 2016). Details of the EOCOPD study have been described previously (Silverman et al., 1998). The EOCOPD data are derived from an extended pedigree-based design. Probands were 53 years old or younger with prebronchodilator forced expiratory volume in 1 s (FEV_1) of $\leq 40\%$, physician-diagnosed COPD, and without severe alpha-1 antitrypsin deficiency. All first-degree relatives, older second-degree relatives, and additional affected family members were enrolled. There were 49 pedigrees with at least two affected family members selected for WES. COPDGene was a multi-center study of smokers with and without COPD and included African-Americans and non-Hispanic whites (Regan et al., 2010). The COPDGene participants, consisting of 10,192 smokers, had at least 10 pack years of smoking, and their ages were between 45 and 80 years. From the COPDGene study, 204 COPD subjects with GOLD (Global Initiative for Chronic Obstructive Lung Disease) spirometry grades 3–4 (post-bronchodilator $\text{FEV}_1 < 50\%$ and ratio of FEV_1 to forced vital capacity (FEV_1/FVC) < 0.7), as well as 195 controls with normal spirometry (frequency-matched to COPD cases on pack-years of cigarette smoking), were chosen for WES.

Sequencing for both cohorts was performed at the University of Washington (Seattle, WA), using Nimblegen V2 capture (Roche NimbleGen, Inc., Madison, WI), and the Illumina platform (Illumina, Inc., San Diego, CA). Participants selected from the COPDGene cohort were sequenced via the NHLBI Exome Sequencing Program, and EOCOPD subjects were sequenced as part of the Center for Mendelian Genomics. Quality control (QC) filtering for both data sets was performed by the method of Qiao et al. (2016) and filtered out variants with Mendelian errors (for family-based data), call rate <99%, Hardy-Weinberg equilibrium p-value $<10^{-8}$, and average sequencing depth <12, as well as excluding subjects with pedigree, racial, or sex mismatches. After QC, there were 303 individuals from 49 families and 124,288 variants in the EOCOPD data set, and there were 394 unrelated individuals and 108,443 variants in the COPDGene data set. For rare variant analyses, we assumed that variants with MAFs $< 5\%$ in dbSNP were rare, and in both studies, we separately filtered out singleton variants or genes with minor allele counts (MACs) <10. Finally, 88,737 rare variants in 13,935 genes were analyzed in the EOCOPD data set, and 24,846 rare variants in 10,550 genes were tested in the COPDGene data set. For both EOCOPD and COPDGene data, GRMs were estimated for variants with MAFs >5% and were incorporated as variance-covariance matrices of genotypes to adjust for population substructure. Effects of covariates for binary phenotypes were adjusted by using the BLUP as an offset. First, we fitted the linear mixed model with adjustments for age, sex, and pack-years of smoking as covariates, and then BLUP was set as the offset for the proposed methods. A description of the two datasets is provided in **Supplementary Table 4**.

TABLE 1 | Type-1 error estimates from simulation study with dichotomous phenotypes.

			Dichotomous			
	# Studies	Significance level	Burden	SKAT	SKAT-O	VT
Hom	3	0.1	0.0960	0.0950	0.0953	0.1100
		0.01	0.0103	0.0099	0.0100	0.0116
		10^{-3}	0.0009	0.0012	0.0014	0.0017
		10^{-4}	0.0001	0.0001	0.0001	0.0004
	6	0.1	0.1002	0.0953	0.0957	0.1018
		0.01	0.0094	0.0085	0.0088	0.0106
		10^{-3}	0.0008	0.0009	0.0008	0.0011
		10^{-4}	0.0001	0.0000	0.0000	0.0001
	9	0.1	0.1000	0.1015	0.1025	0.1018
		0.01	0.0096	0.0098	0.0093	0.0110
		10^{-3}	0.0007	0.0009	0.0007	0.0015
		10^{-4}	0.0001	0.0000	0.0000	0.0001
Het	3	0.1	0.0987	0.1006	0.0981	–
		0.01	0.0100	0.0091	0.0094	–
		10^{-3}	0.0008	0.0008	0.0013	–
		10^{-4}	0.0001	0.0002	0.0002	–
	6	0.1	0.1036	0.0986	0.0985	–
		0.01	0.0094	0.0106	0.0105	–
		10^{-3}	0.0008	0.0014	0.0012	–
		10^{-4}	0.0001	0.0003	0.0002	–
	9	0.1	0.1041	0.1026	0.1046	–
		0.01	0.0107	0.0095	0.0107	–
		10^{-3}	0.0009	0.0011	0.0009	–
		10^{-4}	0.0001	0.0002	0.0001	–

The empirical type-1 error was estimated for the proposed methods with 20,000 replicates at the 0.1, 0.01, 10^{-3}, and 10^{-4} significance levels for dichotomous phenotypes. We assumed that the number of rare variants is 60, and that their minor allele frequencies <0.01. Both homogeneous (Hom) and heterogeneous (Het) models were considered.

RESULTS

Evaluation of metaFARVAT 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 200 unbalanced families. For each replicate, we performed three different meta-analyses, including 3, 6, and 9 studies. **Table 1** shows empirical type-1 error estimates for homogeneous metaFARVAT (metaFARVAT$^{\mathrm{Hom}}$) and heterogeneous metaFARVAT (metaFARVAT$^{\mathrm{Het}}$) at the 0.1, 0.01, 10^{-3}, and 10^{-4} significance levels with dichotomous phenotypes. Estimates of type-1 error rates were virtually equal to nominal significance levels. However, VT type metaFARVAT$^{\mathrm{Hom}}$ showed inflation, especially when there were three studies, and if the number of rare variants is small, it is not recommended. Quantile-quantile (QQ) plots in **Supplementary Figures 2–4** also show consistent results. Therefore, we conclude that the proposed metaFARVAT$^{\mathrm{Hom}}$ and metaFARVAT$^{\mathrm{Het}}$ are statistically valid.

Secondly, empirical power estimates for dichotomous phenotypes were calculated at the 2.5×10^{-6} significance level, showing the changes in power under different scenarios. Empirical power estimates were calculated with 2,000 replicates for seven different statistics: burden, SKAT, SKAT-O, and VT type statistics for metaFARVAT$^{\mathrm{Hom}}$ and burden, SKAT, and SKAT-O type statistics for metaFARVAT$^{\mathrm{Het}}$. Results are provided in **Tables 2**, **3** for homogeneous and heterogeneous scenarios, respectively. In addition, we compared the proposed methods with two meta-analysis methods based on the use of p-values across studies: the minimum p-value method and Fisher's method. If we let p_k be the p-value from the kth study ($k = 1,2,\ldots,K$), the minimum p-value and Fisher's method can be obtained by:

$$\mathrm{minP} = \min\left(p_k\right) \sim \mathrm{Beta}(1,K), \mathrm{Fisher} = -2\sum_{k=1}^{K}\ln p_k \sim \chi^2\left(df = 2K\right) \ \mathrm{under} H_0.$$

According to our results, the minimum p-value approach usually performed the least efficiently, especially when there were equal numbers of protective and deleterious rare variants in the targeted gene. Moreover, the power of the minimum p-value approach was not much improved by including more studies in the meta-analysis. The Fisher approach always performed better than the minimum p-value approach but was less powerful than the metaFARVAT method, regardless of the scenario. Statistical power estimates depend on the scenarios, and both tables show that the best power estimates are usually found from metaFARVAT$^{\mathrm{Hom}}$ and metaFARVAT$^{\mathrm{Het}}$ under homogeneous and heterogeneous scenarios, respectively. Statistical power estimates also depend on the proportion of rare variants with deleterious or protective effects on the phenotypes, which is often unknown. For example, when all rare causal variants had deleterious effects on the phenotype, burden, and VT type metaFARVAT outperformed all other approaches, but if there were variants with deleterious and protective effects, SKAT-type metaFARVAT was the most efficient. SKAT-O metaFARVAT was not always the most powerful, but its empirical power estimates were usually very close to those of the most efficient approach.

The proposed methods can be applied to quantitative phenotypes, and results for quantitative phenotypes are provided in **Supplementary Tables 1–3** and **Supplementary Figures 5–7**. For quantitative phenotypes, we compared our method with RAREMETAL and seqMeta, since these two methods can only be applied to quantitative phenotypes. Both approaches performed better than the proposed approach for homogeneous scenarios. However, RAREMETAL does not provide the SKAT-O type statistic and seqMeta does not provide the VT type statistic. seqMeta performed better than RAREMETAL in most scenarios and was similar to metaFARVAT$^{\mathrm{Hom}}$ under homogeneous scenarios. The SKAT-O type statistic in seqMeta did not perform well when there were as many protective variants as deleterious variants in the gene. metaFARVAT$^{\mathrm{Het}}$ outperformed other methods when the effects of each rare variant differed among studies and when there were variants with deleterious and protective effects within a gene.

TABLE 2 | Empirical power estimates for dichotomous phenotype for homogeneous variants among studies.

		3 studies				6 studies				9 studies			
±	**Method**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**
60/0	Fisher	0.1990				0.6495				0.8940			
	minP	0.0315				0.0610				0.0715			
	Hom	0.0195	0.3590	0.3660	0.3915	0.1690	0.9265	0.9150	0.9240	0.4920	0.9975	0.9945	0.9965
	Het	0.0115	0.3390	0.4160	–	0.0750	0.9095	0.9330	–	0.1865	0.9930	0.9960	–
48/12	Fisher	0.0270				0.1060				0.2400			
	minP	0.0060				0.0070				0.0070			
	Hom	0.0105	0.0335	0.0670	0.0450	0.1105	0.2290	0.3720	0.2665	0.4000	0.5355	0.7565	0.5720
	Het	0.0045	0.0310	0.0720	–	0.0225	0.2080	0.3305	–	0.0760	0.4825	0.6325	–
30/30	Fisher	0.0000				0.0015				0.0035			
	minP	0.0000				0.0000				0.0000			
	Hom	0.0050	0.0000	0.0025	0.0010	0.0555	0.0000	0.0270	0.0000	0.2615	0.0000	0.1650	0.0065
	Het	0.0000	0.0000	0.0005	–	0.0020	0.0000	0.0015	–	0.0120	0.0000	0.0090	–
30/0	Fisher	0.0440				0.2090				0.4520			
	minP	0.0090				0.0170				0.0205			
	Hom	0.0140	0.0725	0.1145	0.0900	0.1790	0.4260	0.5760	0.4785	0.5515	0.7970	0.9125	0.8220
	Het	0.0070	0.0605	0.1290	–	0.0555	0.3905	0.5545	–	0.1410	0.7545	0.8590	–
24/6	Fisher	0.0075				0.0365				0.0895			
	minP	0.0020				0.0020				0.0015			
	Hom	0.0095	0.0045	0.0215	0.0085	0.1285	0.0465	0.1980	0.0610	0.4480	0.1440	0.5480	0.1765
	Het	0.0025	0.0035	0.0240	–	0.0225	0.0340	0.1215	–	0.0630	0.1105	0.2890	–
15/15	Fisher	0.0000				0.0030				0.0045			
	minP	0.0000				0.0010				0.0005			
	Hom	0.0020	0.0000	0.0005	0.0010	0.0550	0.0000	0.0270	0.0025	0.2700	0.0000	0.1650	0.0060
	Het	0.0000	0.0000	0.0000	–	0.0025	0.0000	0.0025	–	0.0090	0.0000	0.0030	–

Empirical power of burden, SKAT, SKAT-O, and VT type of metaFARVAT^{Hom} and metaFARVAT^{Het} was calculated for dichotomous phenotypes with homogeneous effects at the 2.5 × 10^{-6} significance level.

Application to COPD Data

To identify rare variants associated with COPD, we separately conducted rare variant analyses with EOCOPD and COPDGene data. Manhattan and QQ plots are provided in **Supplementary Figure 8**. According to the results, there were no exome-wide significant genes. We also conducted meta-analysis with metaFARVAT$^{\text{Hom}}$ and metaFARVAT$^{\text{Het}}$. For both statistics, ν_1 and ν_2 were set to 1. The QQ plots in **Supplementary Figure 9** show that SKAT-O type metaFARVAT$^{\text{Het}}$ and metaFARVAT$^{\text{Hom}}$ preserved the nominal significance level. However, VT type metaFARVAT exhibited some inflation, and its results are therefore not included in **Table 4**. Manhattan plots are provided in **Supplementary Figure 10**. The Bonferroni-corrected 0.05 genome-wide significance level was 6.76×10^{-6} and is indicated by a solid blue line. **Table 4** shows that *DLEC1* achieved genome-wide significance under both methods, and *ZNF441* was implicated with potentially significant results (p-value $<10^{-4}$) by metaFARVAT$^{\text{Hom}}$ SKAT-O.

DISCUSSION

Family-based association methods are robust against population substructure, and because of genetic homogeneity among family members, they are often utilized for rare variant association analyses. Multiple approaches have been proposed, and Tang and Lin (2015) provided a comprehensive overview of the statistical methods for meta-analysis of sequencing studies for discovering rare variant associations. According to their overview, RAREMETAL (Feng et al., 2014; Liu et al., 2014) and seqMeta (Chen et al., 2014) can be applied to family-based samples. However, these methods can consider only homogeneous effects with quantitative phenotypes, and no statistical methods for dichotomous phenotypes with family-based samples have been proposed.

In this study, we proposed a new meta-analysis method for family-based rare variant association analyses with dichotomous phenotypes, which can test both homogeneous and heterogeneous effects of variants in different studies. metaFARVAT can also be applied to quantitative phenotypes, and is able to combine all study designs, including family-based, case-control, and population-based designs. Furthermore, the proposed method was applied to a meta-analysis of EOCOPD and COPDGene data, and *DLEC1* was found to be genome-wide significant. *DLEC1* is a protein-coding gene encoding a cilia and flagella-associated protein. This gene has been implicated in several cancers but has not been previously associated with COPD. However, cilia-associated genes have been previously implicated in COPD (Tilley et al., 2015).

TABLE 3 | Empirical power estimates for dichotomous phenotype for heterogeneous variants among studies.

		3 studies				**6 studies**				**9 studies**			
±	**Method**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**	**SKAT**	**Burden**	**SKAT-O**	**VT**
48/12	Fisher	0.0240				0.1040				0.2235			
	minP	0.0065				0.0070				0.0105			
	Hom	0.0040	0.0340	0.0425	0.0460	0.0130	0.2170	0.2450	0.2555	0.0420	0.5200	0.5305	0.5680
	Het	0.0080	0.0325	0.0590	–	0.0270	0.1865	0.3180	–	0.0715	0.4555	0.6115	–
30/30	Fisher	0.0000				0.0030				0.0020			
	minP	0.0000				0.0000				0.0000			
	Hom	0.0005	0.0000	0.0000	0.0005	0.0005	0.0000	0.0005	0.0000	0.0015	0.0000	0.0015	0.0000
	Het	0.0005	0.0000	0.0005	–	0.0050	0.0000	0.0030	–	0.0090	0.0000	0.0070	–
30/0	Fisher	0.0460				0.2220				0.4690			
	minP	0.0115				0.0145				0.0160			
	Hom	0.0065	0.0670	0.0945	0.0880	0.0400	0.4385	0.4595	0.4730	0.1185	0.7880	0.7875	0.7980
	Het	0.0060	0.0570	0.1340	–	0.0510	0.3930	0.5580	–	0.1370	0.7425	0.8380	–
24/6	Fisher	0.0095				0.0325				0.0850			
	minP	0.0030				0.0035				0.0030			
	Hom	0.0020	0.0070	0.0115	0.0120	0.0045	0.0470	0.0680	0.0655	0.0125	0.1520	0.1875	0.1900
	Het	0.0040	0.0065	0.0270	–	0.0215	0.0335	0.1230	–	0.0590	0.1185	0.2825	–
15/15	Fisher	0.0010				0.0015				0.0020			
	minP	0.0005				0.0010				0.0005			
	Hom	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	0.0005
	Het	0.0015	0.0000	0.0015	–	0.0030	0.0000	0.0010	–	0.0095	0.0000	0.0060	–

Empirical power of burden, SKAT, SKAT-O, and VT type of metaFARVAT^{Hom} and metaFARVAT^{Het} was calculated for dichotomous phenotypes with heterogeneous effects at the 2.5 × 10^{-6} significance level.

TABLE 4 | The candidate genes found by meta-analysis in COPD studies.

Method	**Data**	**Gene**	**Sample size**	**Chromosome**	**Start**	**End**	**# Rare variants**	**MAC**	**P_B**	**P_S**	***c***	**P_O**
metaFARVAT ^{Hom}	EOCOPD & COPDGene	*DLEC1*	697	3	38,080,978	38,163,785	9	66	1.21e-05	1.02e-04	0.25	5.24e-06
		ZNF441	697	19	11,890,983	11,892,255	2	24	9.88e-05	3.46e-04	1	2.13e-05
metaFARVAT ^{Het}	EOCOPD & COPDGene	*DLEC1*	697	3	38,080,978	38,163,785	15	66	8.03e-06	9.54e-04	0.16	5.43e-06
FARVAT	EOCOPD	*DLEC1*	303	3	38,080,978	38,163,785	9	28	3.70e-03	0.147	1	7.24e-03
		ZNF441	303	19	11,890,983	11,892,255	2	13	1.12e-03	3.83e-03	1	1.37e-03
FARVAT	COPDGene	*DLEC1*	394	3	38,080,978	38,163,785	6	38	5.80e-04	7.96e-04	0.25	3.53e-04
		ZNF441	394	19	11,890,983	11,892,255	1	11	3.09e-02	3.09e-02	1	3.09e-02

The definition of the acronyms in ***Table 4****: (1) #rare variants, the number of rare variants in the gene; (2) MAC, minor allele counts; (3) P_B, the p-value of burden type test; (4) P_S, the p-value of SKAT type test; (5) c, the parameter used for SKAT-O; (6) P_O, the p-value of SKAT-O type test.*

Despite the robustness and efficiency of the proposed method, there are still some limitations of the developed method. First, VT methods sort rare variants according to their MAFs and search the optimal threshold for rare variants. This approach is useful when it is not clear how to define rare variants. However, we found that type-1 error can be inflated if the number of rare variants is too small, and it is computationally intensive if there are a large number of variants to investigate. This problem can be solved by using a permutation method, and further investigation of this approach is necessary. Secondly, sufficiently large samples are necessary to guarantee that SKAT-O follows the assumed asymptotic distribution of the SKAT-O approach under the null hypothesis. Therefore, the SKAT-O type metaFARVAT also has this limitation when it is applied to a dichotomous phenotype with a small sample size. Thirdly, the proposed method cannot be applied to X- or Y-linked genes because the distributions of variants in X and Y chromosomes are different in males and females. Such an improvement will be considered in our future work. Lastly, in the simulation studies, we considered a limited number of rare variants and excluded noise variants. However, in practice, it is not known which rare variants are causal and which represent noise. Extensive simulations are thus necessary in our future work.

Despite the importance of rare variant analyses with family-based samples, this field of study has suffered over the last decades from a lack of statistical methods. In this study, we proposed new methods for family-based samples, enabling such statistical analyses.

AVAILABILITY AND IMPLEMENTATION

The R package for metaFARVAT can be downloaded from http://healthstat.snu.ac.kr/software/metaFARVAT/.

AUTHOR CONTRIBUTIONS

LW and SW conceived and designed the experiments, performed the experiments, analyzed and interpreted the data, and drafted the manuscript. SL maintains the software homepage. DQ, MC, ES, and CL edited the manuscript. All authors read and approved the final version of the manuscript.

FUNDING

This research was supported by the National Research Foundation of Korea (2017M3A9F3046543); by the Industrial Core Technology Development Program (20000134) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea); by National Institutes of Health (NIH) grants U01 HL089856, U01 HL089897, and R01 HL113264; and by National Heart, Lung, and Blood Institute (NHLBI) grant K01 HL129039 (DQ). Sequencing for the Boston Early-Onset COPD Study was provided by the University of Washington Center for Mendelian Genomics and was funded by the National Human Genome Research Institute and by the NHLBI (1U54HG006493). MC has received support from R01HL113264, R01HL137927, and R01HL135142. The COPDGene study (NCT00608764) is supported by NHLBI grants R01 U01 HL089897 and U01 HL089856 and by the COPD Foundation through contributions made to an Industry Advisory Board composed of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, GlaxoSmithKline, Siemens, and Sunovion. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH. The funding bodies had no role in the design of the study; the collection, analysis, and interpretation of data; or in writing the manuscript.

ACKNOWLEDGMENTS

Sequencing for the Boston Early-Onset COPD Study was provided by the University of Washington Center for Mendelian Genomics (UW CMG) and was funded by the National Human Genome Research Institute and the National Heart, Lung, and Blood Institute grant 1U54HG006493. Sequencing for the COPDGene subjects was provided the NHLBI Exome Sequencing Program (ESP) at the University of Washington.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2019.00572/full#supplementary-material

REFERENCES

Chen, H., Lumley, T., Brody, J., Heard-Costa, N. L., Fox, C. S., Cupples, L. A., et al. (2014). Sequence kernel association test for survival traits. *Genet. Epidemiol.* 38, 191–197. doi: 10.1002/gepi.21791

Choi, S., Lee, S., Cichon, S., Nothen, M. M., Lange, C., Park, T., et al. (2014). FARVAT: a family-based rare variant association test. *Bioinformatics* 30, 3197–3205. doi: 10.1093/bioinformatics/btu496

Feng, S., Liu, D., Zhan, X., Wing, M. K., and Abecasis, G. R. (2014). RAREMETAL: fast and powerful meta-analysis for rare variants. *Bioinformatics* 30, 2828–2829. doi: 10.1093/bioinformatics/btu367

George, V. T., and Elston, R. C. (1987). Testing the association between polymorphic markers and quantitative traits in pedigrees. *Genet. Epidemiol.* 4, 193–201. doi: 10.1002/gepi.1370040304

Ionita-Laza, I., Makarov, V., Yoon, S., Raby, B., Buxbaum, J., Nicolae, D. L., et al. (2011). Finding disease variants in Mendelian disorders by using sequence data: methods and applications. *Am. J. Hum. Genet.* 89, 701–712. doi: 10.1016/j.ajhg.2011.11.003

Lange, C., DeMeo, D. L., and Laird, N. M. (2002). Power and design considerations for a general class of family-based association tests: quantitative traits. *Am. J. Hum. Genet.* 71, 1330–1341. doi: 10.1086/344696

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., et al. (2012a). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am. J. Hum. Genet.* 91, 224–237. doi: 10.1016/j.ajhg.2012.06.007

Lee, S., Teslovich, T. M., Boehnke, M., and Lin, X. (2013). General framework for meta-analysis of rare variants in sequencing association studies. *Am. J. Hum. Genet.* 93, 42–53. doi: 10.1016/j.ajhg.2013.05.010

Lee, S., Wu, M. C., and Lin, X. (2012b). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 13, 762–775. doi: 10.1093/biostatistics/kxs014

Li, B., and Leal, S. M. (2008). Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am. J. Hum. Genet.* 83, 311–321. doi: 10.1016/j.ajhg.2008.06.024

Liu, D. J., Peloso, G. M., Zhan, X., Holmen, O. L., Zawistowski, M., Feng, S., et al. (2014). Meta-analysis of gene-level tests for rare variant association. *Nat. Genet.* 46, 200–204. doi: 10.1038/ng.2852

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., et al. (2011). Testing for an unusual distribution of rare variants. *PLoS Genet.* 7:e1001322. doi: 10.1371/journal.pgen.1001322

Price, A. L., Kryukov, G. V., de Bakker, P. I., Purcell, S. M., Staples, J., Wei, L. J., et al. (2010). Pooled association tests for rare variants in exon-resequencing studies. *Am. J. Hum. Genet.* 86, 832–838. doi: 10.1016/j.ajhg.2010.04.005

Qiao, D., Lange, C., Beaty, T. H., Crapo, J. D., Bames, K. C., Bamshad, M., et al. (2016). Exome sequencing analysis in severe, early-onset chronic obstructive pulmonary disease. *Am. J. Respir. Crit. Care Med.* 193, 1353–1363. doi: 10.1164/rccm.201506-1223OC

Regan, E. A., Hokanson, J. E., Murphy, J. R., Make, B., Lynch, D. A., Beaty, T. H., et al. (2010). Genetic epidemiology of COPD (COPDGene) study design. *COPD* 7, 32–43. doi: 10.3109/15412550903499522

Schaffner, S. F., Foo, C., Gabriel, S., Reich, D., Daly, M. J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15, 1576–1583. doi: 10.1101/gr.3709305

Silverman, E. K., Chapman, H. A., Drazen, J. M., Weiss, S. T., Rosner, B., Campbell, E. J., et al. (1998). 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.* 157, 1770–1778. doi: 10.1164/ajrccm.157.6.9706014

Tang, Z. Z., and Lin, D. Y. (2013). MASS: meta-analysis of score statistics for sequencing studies. *Bioinformatics* 29, 1803–1805. doi: 10.1093/bioinformatics/btt280

Tang, Z. Z., and Lin, D. Y. (2014). Meta-analysis of sequencing studies with heterogeneous genetic associations. *Genet. Epidemiol.* 38, 389–401. doi: 10.1002/gepi.21798

Tang, Z. Z., and Lin, D. Y. (2015). Meta-analysis for discovering rare-variant associations: statistical methods and software programs. *Am. J. Hum. Genet.* 97, 35–53. doi: 10.1016/j.ajhg.2015.05.001

Thornton, T., and McPeek, M. S. (2007). Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am. J. Hum. Genet.* 81, 321–337. doi: 10.1086/519497

Thornton, T., Tang, H., Hoffmann, T. J., Ochs-Balcom, H. M., Caan, B. J., and Risch, N. (2012). Estimating kinship in admixed populations. *Am. J. Hum. Genet.* 91, 122–138. doi: 10.1016/j.ajhg.2012.05.024

Tilley, A. E., Walters, M. S., Shaykhiev, R., and Crystal, R. G. (2015). Cilia dysfunction in lung disease. *Annu. Rev. Physiol.* 77, 379–406. doi: 10.1146/annurev-physiol-021014-071931

Wang, L., Choi, S., Lee, S., Park, T., and Won, S. (2016). Comparing family-based rare variant association tests for dichotomous phenotypes. *BMC Proc.* 10 (Suppl. 7), 181–186. doi: 10.1186/s12919-016-0027-8

Won, S., and Lange, C. (2013). A general framework for robust and efficient association analysis in family-based designs: quantitative and dichotomous phenotypes. *Stat. Med.* 32, 4482–4498. doi: 10.1002/sim.5865

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.* 89, 82–93. doi: 10.1016/j.ajhg.2011.05.029

Conflict of Interest Statement: In the past 3 years, ES received honoraria from Novartis for Continuing Medical Education Seminars and grant and travel support from GlaxoSmithKline. MC has received grant support from GSK.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.