Family history aggregation unit-based tests to detect rare genetic variant associations with application to the Framingham Heart Study

Authors

Yanbing Wang, Han Chen, Gina M. Peloso, ..., Sudha Seshadri, Anita L. DeStefano, Josée Dupuis

Correspondence

yanbing@bu.edu

Wang et al., 2022, The American Journal of Human Genetics *109*, 738–749
April 7, 2022
https://doi.org/10.1016/j.ajhg.2022.03.001

CellPress

ARTICLE

Family history aggregation unit-based tests to detect rare genetic variant associations with application to the Framingham Heart Study

Yanbing Wang,[1,*] Han Chen,[2,3] Gina M. Peloso,[1] James B. Meigs,[4,5,9] Alexa S. Beiser,[1,6,7] Sudha Seshadri,[6,7,8] Anita L. DeStefano,[1] and Josée Dupuis[1]

Summary

A challenge in standard genetic studies is maintaining good power to detect associations, especially for low prevalent diseases and rare variants. The traditional methods are most powerful when evaluating the association between variants in balanced study designs. Without accounting for family correlation and unbalanced case-control ratio, these analyses could result in inflated type I error. One cost-effective solution to increase statistical power is exploitation of available family history (FH) that contains valuable information about disease heritability. Here, we develop methods to address the aforementioned type I error issues while providing optimal power to analyze aggregates of rare variants by incorporating additional information from FH. With enhanced power in these methods exploiting FH and accounting for relatedness and unbalanced designs, we successfully detect genes with suggestive associations with Alzheimer disease, dementia, and type 2 diabetes by using the exome chip data from the Framingham Heart Study.

Introduction

Family-based study designs have been widely used in genetic association analysis because these designs take advantage of similar environmental exposure and additional quality control checks that can be performed in family members. However, the standard genome-wide association study (GWAS) has inflated type I error if the familial correlation is not appropriately handled. While GWAS that evaluates the variants individually suffers from low power to detect rare variant associations, the aggregation unit-based test assessing the joint effect of variants in a region can increase power to detect association with rare variants. The original aggregation unit-based tests, such as burden tests,[1] sequence kernel association test (SKAT),[2] SKAT-O,[3] and C-alpha,[4] are valid for studies with unrelated samples. Extensions to these existing methods for related individuals include burden test accounting for familial correlation (famBT),[5] family-based SKAT (famSKAT),[5] and MONSTER (minimum p value optimized nuisance parameter score test extended to relatives),[6] where the type I error is well controlled in family studies. Despite the progress made in handling familial correlation, low power to detect association with rare variants remains, especially for studies with a limited number of cases.

Family history (FH) provides an overview of phenotypes among family members and is valuable for genetic association analysis. Although some available methods[7] for rare variant analysis take familial correlation into consideration, there are limited methods that can directly incorporate available FH. CERAMIC[8] has been proposed to optimize the power in family studies with missing genotypes through the use of FH, but the application is restricted to common variants. The family history aggregation unit-based test (FHAT) method was recently developed to incorporate available FH into rare variant association analysis, which demonstrates a power gain when exploiting additional data from relatives (who have available FH).[9] However, no existing methods can appropriately account for correlation among multiple related family members (who have available FH) or related probands or the imbalance in the case-control ratio that often occurs with low prevalent diseases. Here, we develop a robust method, family-based family history aggregation unit-based test (robust-famFHAT) to address these issues. We use generalized linear mixed model (GLMM)[10,11] to model familial correlation with a random effect, and employ the saddle point approximation (SPA) method and efficient resampling (ER) to calibrate the distribution of score statistics for unbalanced case-control phenotypes. We also develop a unified method, robust-famFHAT-O, which outperforms other methods in scenarios regardless of directions of genetic effects or the proportion of causal variants. The adjustment for relatedness allows these methods to preserve the correct type I error in studies containing related samples and prevents the reduction in sample size that occurs when

[1]Department of Biostatistics, School of Public Health, Boston University, Boston, MA 02215, USA; [2]Human Genetics Center, Department of Epidemiology, Human Genetics and Environmental Sciences, School of Public Health, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [3]Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA; [4]Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA 02214, USA; [5]Harvard Medical School, Boston, MA 02215, USA; [6]Framingham Heart Study, Framingham, MA 01701, USA; [7]Department of Neurology, Boston University School of Medicine, Boston, MA 02215, USA; [8]Glenn Biggs Institute for Alzheimer Disease and Neurodegenerative Diseases, University of Texas Health San Antonio, San Antonio, TX 78229, USA; [9]Program in Medical and Population Genetics, Broad Institute, Cambridge, MA 02115, USA
*Correspondence: yanbing@bu.edu
https://doi.org/10.1016/j.ajhg.2022.03.001.

restricting analyses to unrelated samples. Moreover, the robust methods we propose can address inflated type I error for low prevalent diseases through the adjustment of the unbalanced case-control ratios.

The proposed methods can incorporate FH from relatives with different degrees of relationship with the probands. We demonstrate the validity of these approaches to incorporate multiple relatives from complex family structures through a simulation analysis where the type I error is correctly controlled. After incorporating available FH from relatives, our methods are more powerful than other existing methods (i.e., famSKAT and MONSTER). We refer to MONSTER as famSKAT-O in this paper. A recently proposed method incorporating FH, liability threshold model conditional on both case-control status and FH (LT-FH),[12] greatly increases association power in case-control studies. However, LT-FH only incorporates first-degree relative information (i.e., FH from parents or siblings). The simulation analysis shows that the non-robust methods (without unbalanced case-control ratio adjustment) have the correct type I error when case-control ratios are balanced but suffer from the inflation under case-control imbalance. We suggest the use of robust methods that control the type I error well when disease prevalence is low. A gene-based analysis of exome chip data from the Framingham Heart Study (FHS) yields improved significance in rare variant regions known to be associated with Alzheimer disease (AD) and dementia when using famFHAT and famFHAT-O compared to other standard methods that ignore FH. The analysis with robust methods is compared to the results obtained with the non-robust methods. Furthermore, with enhanced power to detect rare variant associations, we identify novel genes with suggestive associations for AD, dementia, and type 2 diabetes (T2D).

Material and methods

Accounting for familial correlation in aggregation unit-based tests

When related samples are present in a study, one can account for the correlation among observations through the GLMM[10,11] with a random effect. Specifically, the GLMM accounting for correlation in n probands can be specified as

$$g\left(E\left(Y_i^P|G_i^P, X_i^P, \delta_i^P\right)\right) = X_i^P\alpha_P + G_i^P\,\beta_P + \delta_i^P,$$

where $g(\bullet)$ is the link function that connects the phenotype mean Y_i^P with the covariates vector X_i^P, the genotype vector G_i^P, and the random effect δ_i^P for i proband. We assume the $n\times1$ random effect vector δ^P that contains each entry δ_i^P follows the distribution of $N\left(0, \sigma_{G_P}^2\Phi_P\right)$, where Φ_P is the $n\times n$ matrix containing twice the kinship estimates obtained from family information or estimated from genotype data when available and $\sigma_{G_P}^2$ is the parameter of variance component.[5] In this model, α_P is a vector of regression coefficients for covariate effects and β_P is a vector of regression coefficients for the observed genotypes in probands. The genetic effect β_{P_j} for variant j is assumed to have an arbitrary distribution with mean zero and a variance of $w_j^2\tau$, where w_j is a pre-specified weight for variant j and τ is a variance component, thus, testing $\beta_P = 0$ is equivalent to test $\tau = 0$ in H_0. To test $\tau = 0$, the corresponding famSKAT score statistic accounting for familial correlation is[5]

$$Q_{famSKAT} = \frac{\left(Y^P - \widehat{\mu}_P\right)^T G^P W W G^{PT}\left(Y^P - \widehat{\mu}_P\right)}{\widehat{\varphi}_P^2},$$

where $\widehat{\varphi}_P$ is the dispersion parameter estimate under H_0 and $\widehat{\varphi}_P = 1$ is fixed for binary traits; $W = diag(w_1, w_2, \ldots, w_m)$ is a pre-specified weight matrix with a size of $m \times m$ for the genotypes of m variants; G^P is the $n\times m$ genotype matrix with (i, v) element corresponding to the additively coded genotype for variant v of proband i; $\widehat{\mu}_P = (\widehat{\mu}_{P1}, \widehat{\mu}_{P2}, \ldots, \widehat{\mu}_{Pn})$ is the estimated mean of $Y^P = (Y_1^P, Y_2^P, \ldots, Y_n^P)$ under H_0 of no genetic effect, where

$$\widehat{\mu}_{Pi} = g\left(E\left(Y_i^P|X_i^P, \delta_i\right)\right) = X_i^P\widehat{\alpha}_P + \widehat{\delta}_{Pi}.$$

On the basis of these assumptions, we can write the variance-covariance matrix under the null hypothesis as

$$\widehat{\Sigma}_{P_0} = \widehat{\sigma}_{G_P}^2\Phi_P + \widehat{H}_P^{-1},$$

where $\widehat{H}_P^{-1} = \widehat{\varphi}_P\boldsymbol{I}$ for continuous traits and $\widehat{H}_P^{-1} = diag\left\{\left(1/\widehat{\mu}_{Pi}(1-\widehat{\mu}_{Pi})\right)\right\}$ for binary traits (supplemental information).

Similarly to the burden test, famBT aggregates the rare variants with pre-specified weights, i.e., $\sum_{j=1}^m w_j g^P{}_{ij}$, and then tests the association between this weighted sum of genotypes with the phenotype:

$$Q_{famBT} = \left[\frac{1}{\widehat{\varphi}_P}\sum_{i=1}^n\left(Y_i^P - \widehat{\mu}_{P_i}\right)\left(\sum_{j=1}^m w_j g^P{}_{ij}\right)\right]^2.$$

The above famSKAT and famBT methods provide a more general framework for the analysis of rare variants for both binary and continuous traits through GLMM, which are equivalent to the variant-set mixed model association tests (SMMAT) proposed in Chen et al.[13]

Incorporating family history in family-based rare variant association tests

We combine the scores accounting for familial correlation to formulate famFHAT and famFHAT-O by using a weighted meta-analysis framework.

The FH could be collected from relatives with different degrees of relationship with the probands. The relatives could be parents, siblings, or more distant relatives, which corresponds to different kinship coefficients. When we have FH from relatives who are from the same type of relationship with the same kinship coefficient (i.e., all siblings with kinship coefficient = 1/4), let $Y_j^{P'}$ and $G_j^{P'}$ denote the vectors of phenotype mean and genotype mean among the closest proband(s) defined on the basis of the known kinship coefficient for relative j and X_j^R be the covariates in relatives. We evaluate the total association evidence through two separate GLMMs for probands and relatives:

$$g\left(E\left(Y_i^P|G_i^P\ ,\ X_i^P\ ,\ \delta_i^P\right)\right) = X_i^P\alpha_P + G_i^P\beta_P + \delta_i^P, \qquad \text{(Equation 1)}$$

$$g\left(E\left(Y_j^R\middle|G_j^{P'}\ ,\ Y_j^{P'}, X_j^R\ ,\ \delta_j^R\right)\right) = X_j^R\alpha_R + G_j^{P'}\beta_R + Y_j^{P'}\lambda_R + \delta_j^R. \qquad \text{(Equation 2)}$$

In the second equation, λ_R is a scalar of the regression coefficient for probands' phenotypes for the relatives' model; α_R is a vector of regression coefficients for relatives' covariates; β_R is a vector of

regression coefficients for the m observed genetic variants in probands. Supposing that all the probands have a single type of relatives R_1 (i.e., all probands have FH from siblings), the formula to calculate the score statistics of famFHAT to test the null hypothesis of no genetic effect is

$$Q_{famFHAT} = \left[\frac{(Y^P - \widehat{\mu}_P)^T}{\widehat{\varphi}_P} G^P W + \frac{\{D(R_1)(Y^{R_1} - \widehat{\mu}_{R_1})\}^T}{\widehat{\varphi}_{R_1}} 2\Omega_1 G^{P'} W\right]$$

$$\left[WG^{P^T}\frac{(Y^P - \widehat{\mu}_P)}{\widehat{\varphi}_P} + 2\Omega_1 WG^{P'^T}\frac{D(R_1)(Y^{R_1} - \widehat{\mu}_{R_1})}{\widehat{\varphi}_{R_1}}\right], \quad \text{(Equation 3)}$$

where $G^{P'}$ is the $n \times m$ genotype matrix with the (j, v) element representing the genotype mean of variant v among the closest probands for relative j; $\widehat{\mu}_{R_1}$ is the estimated mean vector in relatives R_1 of all probands; Ω_1 is the scalar that contains the kinship coefficient between relative and proband; diagonal matrix $D(R_1)$ is defined to indicate the availability of FH in the relative.

A unified test that combines famFHAT and famFHAT-Burden is

$$Q_\rho = (1-\rho)Q_{famFHAT} + \rho Q_{famFHAT-Burden},$$

where

$$Q_{famFHAT-Burden} = \left[\frac{\sum_{i=1}^n (Y_i^P - \widehat{\mu}_{Pi})\left(\sum_{j=1}^m w_j g^P{}_{ij}\right)}{\widehat{\varphi}_P} + \frac{2\Omega_1 \sum_{i=1}^n d_i(R_1)\left(Y_i^{R_1} - \widehat{\mu}_{R_1{}_i}\right)\left(\sum_{j=1}^m w_j g^{P'}{}_{ij}\right)}{\widehat{\varphi}_{R_1}}\right]^2,$$

where $d_i(R_1)$ indicates the availablity of relative's FH for probands i. The value of ρ is selected to minimize the p value, and we can write the famFHAT-O statistic as

$$Q_{famFHAT-O} = \min_{0 \le \rho \le 1} P_\rho.$$

Incorporating data from multiple relatives into analyses

The GLMM can be used to account for correlation in the proband analysis or in the analysis of probands' relatives when we want to combine data from multiple relatives per proband. The famFHAT statistic shown in the previous section can be used when all relatives who have FH are from the same type of relatedness (thus all relatives have the same kinship coefficients). The score in Equation 3 is constructed by combining the scores from probands and relatives with appropriate weights. Because the score for ungenotyped relatives is down-weighted by twice of their kinship coefficient, the scores of relatives with different degrees of relatedness are combined by meta-analysis with relationship-appropriate weights, preventing adjustment for the correlation among relatives with different degrees of relatedness to the probands.

Therefore, we propose an alternative approach to combine the multiple relatives with different degrees of relatedness. We assume that there is a superset of K types of relatives with the sample size of $n \times K$, indexed $k = 1$ to K (e.g., relative $k = 1$ for siblings, $k = 2$ for maternal grandfather, etc.), that includes all possible types of relatives available on any of the probands. We can re-write the relatives' Equation 2 as

$$g\left(E\left(Y_j^R \middle| G_j^P, Y_j^P, X_j^R, \delta_j^R\right)\right) = X_j^R \alpha_R + \tilde{G}_j^{P'} \beta_P + Y_j^{P'} \lambda_R + \delta_j^R, \quad \text{(Equation 4)}$$

where $\tilde{G}_i^{P'}$ is the weighted form of $G_i^{P'}$ with each row j that is down-weighted by twice the kinship coefficient for relative j based on the underlying relationship with proband.[14] Because Equation 2 is equivalent to Equation 4, instead of performing relative-specific analyses and appropriately down-weighting the score for each relative type, we can down-weight the genotypes on the basis of their relationship with probands prior to computing the scores. Thus, the score statistics for probands and relatives can still be obtained through two separate analyses: probands analysis and a single analysis of probands' relatives. The scores from probands and their relatives are combined via meta-analysis to obtain famFHAT and famFHAT-O. Although it is an approximation, we have investigated this alternative approach through a simulation study. This allows the incorporation of FH from multiple relatives with different relatedness in the analysis through a single relatives' model. To address the situation where probands have multiple relatives, we take the following strategies to calculate our statistics in our analyses: (1) for each relative, the genotype average among the closest proband(s) is used in the score calculation, and the phenotype average among the closest proband(s) is used as the covariate in the relatives' model; (2) we down-weight the genotypes in relatives' score calculation and then combine it with probands' score to calculate famFHAT and famFHAT-O. The score statistic that combines the multiple relatives with different relationships is defined as

$$Q_{famFHAT} = \left[\frac{(Y^P - \widehat{\mu}_P)^T}{\widehat{\varphi}_P} G^P W + \sum_k \frac{\{D(R_k)(Y^{R_k} - \widehat{\mu}_{R_k})\}^T}{\widehat{\varphi}_R} \tilde{G}_k^{P'} W\right]$$

$$\left[WG^{P^T}\frac{(Y^P - \widehat{\mu}_P)}{\widehat{\varphi}_P} + \sum_k W\tilde{G}_k^{P'^T}\frac{D(R_k)(Y^{R_k} - \widehat{\mu}_{R_k})}{\widehat{\varphi}_R}\right].$$

Here, we let $\Omega = (\Omega_1, \Omega_2, \ldots, \Omega_K)$denote a vector of length K containing the elements denoting kinship coefficient, and $G_k^{P'}$ be the genotype mean matrix for each relative type k, then $\tilde{G}_k^{P'}$ is obtained by down-weighting by scalar $2\Omega_k$ for relative k. Moreover, we define the $n \times n$ diagonal matrix $D(R_k)$ with i^{th} diagonal element $d_i(R_k)$ to denote the availability of k^{th} relative's FH for proband i. Specifically, $d_i(R_k)$ equals to 0 if the values in Y^{R_k}, the phenotype vector for relative k of probands, are missing, and equals 1 otherwise.

Under the null hypothesis, $\widehat{P}_P = \widehat{\Sigma}_{P_0}{}^{-1} - \widehat{\Sigma}_{P_0}{}^{-1} X_P (X_P^T \widehat{\Sigma}_{P_0}{}^{-1} X_P)^{-1} X_P^T \widehat{\Sigma}_{P_0}{}^{-1}$ and $\widehat{P}_{R_k} = \widehat{\Sigma}_{R_{k_0}}{}^{-1} - \widehat{\Sigma}_{R_{k_0}}{}^{-1} X_{R_k} \left(X_{R_k}^T \widehat{\Sigma}_{R_{k_0}}{}^{-1} X_{R_k}\right)^{-1} X_{R_k}^T \widehat{\Sigma}_{R_{k_0}}{}^{-1}$ are the two projection matrices for proband and relative k, and $Q_{famFHAT}$ asymptotically follows a weighted sum of chi-square distributions with one degree of freedom (df = 1), $\sum_{j=1}^m \lambda_j \chi_{1,j}^2$, where λ_j are the eigenvalues of $W G^{P^T} \widehat{P}_P G^P W + \sum_k W \tilde{G}_k^{P'^T} D(R_k) \widehat{P}_{R_k} D(R_k) \tilde{G}_k^{P'} W$.

Aggregation unit-based test statistics with SPA and ER

We propose robust versions of famFHAT and famFHAT-O to adjust for the case-control imbalance that happens when the disease prevalence is low in unascertained study designs. The normal approximation performs well when the score statistic S_j is within two standard deviations of the mean.[15] However, the skewed distribution of the score statistic when the case-control ratio is extremely unbalanced can result in inflation in type I error rates. SPA is an approach for p value calculation using all cumulants through the cumulant generating function (CGF), which can provide a more accurate

approximation of the score distribution compared to the normal approximation with only two cumulants. However, due to the poor performance for low minor allele count (MAC), the ER is the preferred method that can provide more accurate p values for very rare variants (i.e., MAC $\leq$ 10) by resampling the disease status among individuals containing variants with a low frequency. The details of SPA and ER have been discussed elsewhere.[16,17] Here, we take a similar approach to the one described in Zhou et al.[17] Let $\chi^2_{quantile}$ be the quantile function for the chi-square distribution with df = 1, and the variance for $S^2_{\tilde{P}_j}$ can be adjusted with the following formula:

$$\tilde{V}_j^P = \frac{S^2_{\tilde{P}_j}}{\chi^2_{quantile}\left(1-\tilde{p}_j\right)}, \qquad \text{(Equation 5)}$$

where $\tilde{p}_j$ is the p value estimated with SPA (when MAC > 10) or ER (when MAC $\leq$ 10) for variant j. Given the fact that SPA performs better for single variant tests, a further adjustment is made with burden tests because they can be written as the square of the sum of single variant scores across all samples, i.e., $Q_{P_{Burden}} = (S_{P_{Burden}})^2$ with $S_{P_{Burden}} = \sum_{i=1}^{n}\sum_{j=1}^{m} w_j g^P{}_{ij}\left(Y_i^P - \widehat{u}_{Pi}\right)$. We define $\tilde{r}^P = \min\left(1, \left(w^T \tilde{V}^{P\frac{1}{2}} D \tilde{V}^{P\frac{1}{2}} w\right) / \tilde{V}^P_{Burden}\right)$, where $\tilde{V}^P = (\tilde{V}_1^P, \tilde{V}_2^P, \ldots, \tilde{V}_m^P)$, and $\tilde{V}^P_{Burden} = (\tilde{V}^P_{Burden\ 1}, \tilde{V}^P_{Burden\ 2}, \ldots, \tilde{V}^P_{Burden\ m})$ is the adjusted variance for $S^2_{\tilde{P}_{Burden}}$ using Equation 5, $w = (w_1, w_2, \ldots, w_m)$ is the vector of weights for m variants, and D is the correlation matrix of m variants. In robust-famFHAT and robust-famFHAT-O, we use the calibrated distribution when the case-control ratio is unbalanced to derive an aggregation unit-based test,

$$\tilde{S}_P \sim MVN\left(0, \left(\frac{\tilde{V}^P}{\tilde{r}^P}\right)^{\frac{1}{2}} D \left(\frac{\tilde{V}^P}{\tilde{r}^P}\right)^{\frac{1}{2}}\right).$$

Simulation studies

Validation of incorporating FH from multiple relatives

We first simulate genotypes for 5,000 sibling pairs, and for one of the siblings in the pair, we simulate one offspring. Below the siblings are labeled as "aunt/uncle" and "parent." The offspring in those 5,000 families are treated as probands who have both available phenotypes and genotypes, and the proband in each simulated family has a parent and an aunt/uncle. The model used for the phenotype simulation is

$$\begin{pmatrix} Y^{aunt/uncle} \\ Y^{parent} \\ Y^{proband} \end{pmatrix} = 0.015 \begin{pmatrix} age^{aunt/uncle} \\ age^{parent} \\ age^{proband} \end{pmatrix} + 0.25 \begin{pmatrix} sex^{aunt/uncle} \\ sex^{parent} \\ sex^{proband} \end{pmatrix} + \begin{pmatrix} G^{aunt/uncle}_{causal} \\ G^{parent}_{causal} \\ G^{proband}_{causal} \end{pmatrix} \gamma + \varepsilon,$$

where $\varepsilon \sim MVN(0, \Sigma)$ with $\Sigma = \sigma_g^2 \begin{pmatrix} 1 & 0.5 & 0.25 \\ 0.5 & 1 & 0.5 \\ 0.25 & 0.5 & 1 \end{pmatrix} + \sigma_e^2 \boldsymbol{I}_{3\times 3}$, and $\sigma_g^2 = \sigma_e^2 = 0.5$. In the model, $age^{aunt/uncle}$, age^{parent}, and $age^{proband}$ are vectors of continuous variable randomly selected from ages in UK Biobank data with a range from 38 to 72 for probands and 60 to 105 in relatives; $sex^{proband}$ is a vector of binary variable generated from a Bernoulli distribution with probability for female = 56% in probands; $sex^{aunt/uncle}$ and sex^{parent} are the vectors of binary variables generated from the Bernoulli distribution with probability of having female = male = 50%; G^s_{causal} is the matrix containing the causal genotypes for individual s; and the γ is used to define the genetic effect:[5]

$$\gamma_j = \sqrt{\frac{c}{2MAF_j(1-MAF_j)}}, \qquad \text{(Equation 6)}$$

where MAF_j is the minor allele frequency (MAF) of causal variant j. Let R^2 denote the proportion of variance-explained variants, E denote the vector representing the direction of genetic effects, and D denote the linkage disequilibrium (LD) correlation matrix. We conduct two sets of simulation studies, one to evaluate type I error and a second set of simulations to evaluate power. In the power analysis,

$$c = \frac{R^2}{E^T D E},$$

we fix $R^2 = 1\%$ when all genetic effects are in the same direction and $R^2 = 3\%$ when half of the genetic effects are in the same directions and half of the variants with the opposite directions. We simulate the data by assuming the proportion of causal variants = 20%. We set c equal to zero for the type I error analysis. On the liability scale, we convert the simulated continuous phenotypes to a binary trait with prevalence = 20%. In the model, we adjust for age and sex, and we test the regions containing 40 variants.

In the first analysis, we include all 5,000 probands, the parents from a subset of 2,500 families, and the aunts/uncles from the remaining 2,500 families to create an unrelated relative subset. We calculate the scores for probands and relatives by fitting three separate logistic models:

$$logitP\left(Y^P = 1 \middle| G^P, X^P\right) = X^P \alpha_P + G^P \beta_P,$$

$$logitP\left(Y^{parents} \middle| G^P, X^{parents}\right) = X^{parents} \alpha_{parents} + Y^P \lambda_{parents} + G^P \beta_{parents},$$

$$logitP\left(Y^{aunts/uncles} \middle| G^P, X^{aunts/uncles}\right) = X^{aunts/uncles}\, \alpha_{aunts/uncles} + Y^P\, \lambda_{aunts/uncles} + G^P\, \beta_{aunts/uncles}.$$

Because probands' relatives are unrelated to each other, the FHAT method[9] using the standard logistic model is valid to apply to calculate the scores. We down-weight the relative scores by twice of kinship coefficient (i.e., $^1/_2$ of the score for parents and $^1/_4$ of the score for aunts/uncles) and combine them with the proband score by using a weighted meta-analysis framework to calculate $FHAT_1$, i.e.,

$$Q_{FHAT_1} = \left[\frac{\left(Y^P - \widehat{\mu}_P\right)^T}{\widehat{\varphi}_P} + \frac{\left(Y^{parents} - \widehat{\mu}_{parents}\right)^T}{2\widehat{\varphi}_{parents}} + \frac{\left(Y^{aunts/uncles} - \widehat{\mu}_{aunts/uncles}\right)^T}{4\widehat{\varphi}_{aunts/uncles}}\right] G^P W W G^{P^T} \left[\frac{\left(Y^P - \widehat{\mu}_P\right)}{\widehat{\varphi}_P} + \frac{\left(Y^{parents} - \widehat{\mu}_{parents}\right)}{2\widehat{\varphi}_{parents}} + \frac{\left(Y^{aunts/uncles} - \widehat{\mu}_{aunts/uncles}\right)}{4\widehat{\varphi}_{aunts/uncles}}\right],$$

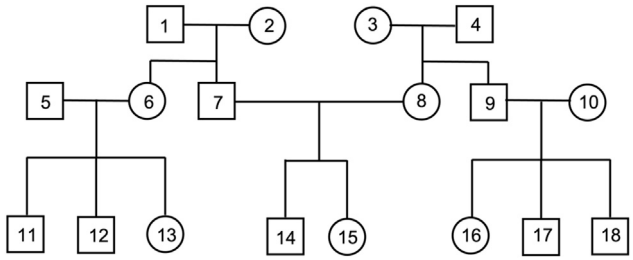

Figure 1. Pedigree used for complex family structure simulation

where G^P is the observed genotype matrix in 5,000 probands; $\widehat{\mu}_P$, $\widehat{\mu}_{parents}$, and $\widehat{\mu}_{aunts/uncles}$ are the estimated means of Y^P, $Y^{parents}$, and $Y^{aunts/uncles}$ under H_0, dispersion parameters are fixed to 1 for binary traits.

In the second analysis, we combine 2,500 parents and 2,500 aunts/uncles into a single relatives' model. Specifically, the standard logistic model for probands and logistic model for relatives (parents and aunts/uncles) is used,

$$logitP\left(Y^P = 1 | G^P, X^P\right) = X^P \alpha_P + G^P \beta_P,$$

$$logitP\left(Y^{relatives} | G^P, X^{relatives}\right) = X^{relatives}\ \alpha_{relatives} + Y^P \lambda_{relatives} + G^P\ \beta_{relatives}.$$

We down-weight the genotypes for parents and aunts/uncles by $^1/_2$ and $^1/_4$, respectively, and we combine the corresponding scores for probands and relatives (parents and aunts/uncles) through meta-analysis to calculate $FHAT_2$,

$$Q_{FHAT_2} = \left[\frac{\left(Y^P - \widehat{\mu}_P\right)^T}{\widehat{\varphi}_P} G^P W + \frac{\left(Y^{relatives} - \widehat{\mu}_{relatives}\right)^T}{\widehat{\varphi}_{relatives}} \tilde{G}^P W\right]$$

$$\left[W G^{P^T} \frac{\left(Y^P - \widehat{\mu}_P\right)}{\widehat{\varphi}_P} + W \tilde{G}^{P^T} \frac{\left(Y^{relatives} - \widehat{\mu}_{relatives}\right)}{\widehat{\varphi}_{relatives}}\right],$$

where $\tilde{G}^P$ is the down-weighted genotype matrix, $\widehat{\mu}_P$ and $\widehat{\mu}_{relatives}$ are the estimated means of Y^P and $Y^{relatives}$ under the null hypothesis, respectively.

In the third analysis, we compare our methods to LT-FH. We calculate LT-FH phenotypes by including FH from 2,500 parents and run a SKAT analysis on the outcome of LT-FH phenotypes (we refer to this method as SKAT-LTFH). Note that the LT-FH approach cannot incorporate the FH from aunts/uncles and it's only applicable for combining FH from the first-degree relatives.

Performance of proposed methods in complex family structure

We simulate 400 families containing 18 family members from three generations (Figure 1) to have probands and relatives with complex family structures to evaluate the performance of the proposed methods combining FH from multiple relatives. We randomly assign two haplotypes simulated from HapGen2[18] to the founders. The offspring are randomly assigned one haplotype without recombination from each parent. The model used to simulate the continuous phenotypes is specified as follows,

$$\begin{pmatrix} Y^{s=1} \\ Y^{s=2} \\ \dots \\ Y^{s=18} \end{pmatrix} = 0.015 \begin{pmatrix} age^{s=1} \\ age^{s=2} \\ \dots \\ age^{s=18} \end{pmatrix} + 0.25 \begin{pmatrix} sex^{s=1} \\ sex^{s=2} \\ \dots \\ sex^{s=18} \end{pmatrix} + \begin{pmatrix} G^{s=1}_{causal} \\ G^{s=2}_{causal} \\ \dots \\ G^{s=18}_{causal} \end{pmatrix} \gamma$$

$$+ \varepsilon, \varepsilon \sim MVN(0, \Sigma),$$

Table 1. Disease prevalence in the FHS

	Probands (with both genotypes and phenotypes)	Relatives (with phenotypes)
AD		
Sample size	3,949	1,744
Disease prevalence	11.3%	14.3%
Dementia		
Sample size	3,949	1,744
Disease prevalence	14.3%	19.6%
T2D		
Sample size	7,356	2,765
Disease prevalence	13.1%	16.8%

where $\Sigma = 0.4\ \Phi + 0.6\ \boldsymbol{I}_{18\times18}$ is the covariance matrix and Φ is twice the kinship matrix representing the relationship among the 18 family members; Y^s, age^s, sex^s are the values for phenotype, age, sex of individual s in the pedigree (Figure 1); G^s_{causal} is the matrix containing the causal genotypes for individual s; and the parameter γ controls the genetic effects as defined in (Equation 6). We fix R^2 to 2% in the power analysis when all causal variants have the same effect directions, and we increase this value to 3% for the scenario where half of the causal variants have positive effects and half of the causal variants have negative effects for a trait. The continuous phenotypes are converted to binary outcomes with the normal approximation to obtain a prespecified prevalence on the liability scale for a binary trait. We randomly select nine probands from each family contributing both genotypes and phenotypes and the remaining nine samples from each family contributing only phenotypes. We repeat this process randomly for each iteration to obtain random and complex relationships among probands and relatives. Therefore, we have 3,600 probands and 3,600 relatives included in the analysis. Only the relatives who have at least one proband are used in the analysis. We analyze the regions with 30 variants by using robust-famFHAT and robust-famFHAT-O, which use both simulated phenotypes and genotypes for probands, whereases only phenotypes are used in the form of FH for relatives by assuming their genotypes are missing. We also apply the adjustment for the case-control imbalance to famSKAT and famSKAT-O, named as robust-famSKAT and robust-famSKAT-O. We compare the results for type I error rates and power for the robust versions to the non-robust versions of the proposed methods (famFHAT, famFHAT-O) and other existing methods that only use proband data.

Application to the Framingham Heart Study

The FHS is a community-based, longitudinal cohort study. The cohort comprises residents of Framingham, Massachusetts, and these residents have undergone up to 32 examinations, performed every 2 years, that have involved detailed history taken by a physician, a physical examination, and laboratory testing.[19] The participants in the cohort's offspring study have completed up to nine examinations, which have taken place approximately every 4 years.[20] Approval was obtained from the Institutional Review Board of the Boston University Medical Campus. All study participants provided written informed consent. The diagnosis of AD

Table 2. Type I error rates for $FHAT_1$ and $FHAT_2$and their 95% confidence intervals

Alpha	Type I error for $FHAT_1$	Type I error for $FHAT_2$
0.05	0.052 (0.048, 0.057)	0.053 (0.049, 0.058)
0.01	0.0094 (0.0076, 0.011)	0.0094 (0.0076, 0.011)
0.001	0.001 (0.00048, 0.0018)	0.0011 (0.00055, 0.0020)
0.005	0.0048 (0.0035, 0.0064)	0.0050 (0.0037, 0.0066)

and related dementia in the FHS is based on criteria for possible, probable, or definite AD from the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer Disease and Related Disorders Association (NINCDS–ADRDA).[21,22] The participants with a history of diabetes treatment such as insulin or an oral hypoglycemic agent were also diagnosed as having diabetes. The diagnosis of diabetes is based on a fasting plasma glucose > 125 mg/dL or non-fasting plasma glucose > 200 mg/dL or Hemoglobin A1c (HbA1c) > 6.5 or history of diabetes treatment such as insulin or an oral hypoglycemic agent.

We apply our methods to analyze exome chip data in the FHS to investigate the regions associated with three phenotypes: AD, dementia, and T2D. In the FHS analysis, we assume that probands are those who have both available genotypes and phenotypes and relatives are those with available phenotypes only (i.e., FH). In our definition of probands, probands can either be cases or controls. The relatives are defined as the phenotyped but not genotyped samples who are related to at least one proband on the basis of the kinship matrix. We use famFHAT and famFHAT-O to combine all FHS relatives without available genotypes but available phenotype data into the analysis to evaluate the association between genes and disease status. We also apply the robust methods for comparison. Note that the inclusion of probands does not depend on the availability of FH from their relatives because the probands' analysis is separated from the relatives' analysis. We classify coding variants for aggregation unit-based testing for exome chip analysis, which include nonsynonymous exonic variants, splicing exonic variants, splicing nonsynonymous exonic variants, exonic splicing stop gained, exonic splicing stop lost, splicing, stop gained, and stop loss.[23] To scan the association across the exome chip data in the FHS, we select variants with MAF < 1% and MAC > 1.

As shown in Table 1, a total of 3,949 probands with available AD/dementia status and exome chip genotypes (p = 11.3% in AD, p = 14.3% in dementia) and 1,744 relatives with available dementia/AD status but missing genotypes are involved in dementia and AD analyses (p = 14.3% in AD, p = 19.6% in dementia). We adjust for age at the time of DNA draw, education status, sex, and the first two principal components (PCs) in AD analysis. Our dementia analysis includes the same covariates as the AD analysis.

The T2D analysis includes 7,356 probands with available T2D status and genotypes (p = 13.1%) as well as 2,765 ungenotyped relatives with available T2D status (p = 16.8%). We adjust the T2D analysis for age at the last exam, body mass index (BMI) at the last exam, sex, PC1, and PC2. We first test three previously reported rare variant gene regions for AD and dementia, *APOE* (MIM: 107741), *SLC24A4* (MIM: 609840), and *INPP5D* (MIM: 601582), and three known genes for T2D, *PSD4* (MIM: 614442), *MRPL46* (MIM: 611851), and *GPD2* (MIM: 138430).[24,25] We then explore novel regions by using all exome chip data.

Results

Simulation studies

Simulation: Validation of incorporating FH from multiple relatives

We assess the type I error rates by performing 10,000 simulation replicates (see Table 2). The power for different scenarios of genetic effect directions using 1,000 replicates is evaluated (Table 3). Figure 2 summarizes the p values computed by $FHAT_1$ and $FHAT_2$ from the first two sets of analyses described in the previous section. Because the results are comparable between the approach analyzing different relative types separately versus the approach combining all relatives in the analysis by down-weighting genotypes proportionally to the coefficient of relationship

Table 3. Power comparison between $FHAT_1$, $FHAT_2$, SKAT, and SKAT-LTFH

Genetic effects (+/−)	Alpha	SKAT analysis	Analysis 1 ($FHAT_1$ with separate analyses for parents and aunts/uncles)	Analysis 2 ($FHAT_2$ with one combined analysis for parents and aunts/uncles)	Analysis 3 (SKAT-LTFH)
(100/0)	2.5×10^{-3}	97.8%	99.2%	99.1%	98.3%
	2.5×10^{-4}	92.9%	96.3%	96.1%	94.6%
	2.5×10^{-5}	84.1%	91.8%	91.5%	89.6%
	2.5×10^{-6}	74.8%	85.1%	84.8%	82.5%
	2.5×10^{-7}	65.9%	76.5%	76.1%	73.9%
(50/50)	2.5×10^{-3}	90.3%	95.3%	95.4%	93.6%
	2.5×10^{-4}	81.5%	89.6%	89.6%	84.2%
	2.5×10^{-5}	72.5%	83.8%	83.8%	79.9%
	2.5×10^{-6}	65.1%	77.3%	77.0%	72.7%
	2.5×10^{-7}	56.4%	70.6%	70.5%	66.4%

+/− indicates the proportion of variants with positive/negative effects. SKAT only includes 5,000 probands in the analysis. FHAT includes 5,000 probands and incorporates FH from 2,500 parents and 2,500 aunts/uncles in the analysis. SKAT-LTFH includes 5,000 probands and incorporates FH from 2,500 parents in the analysis.

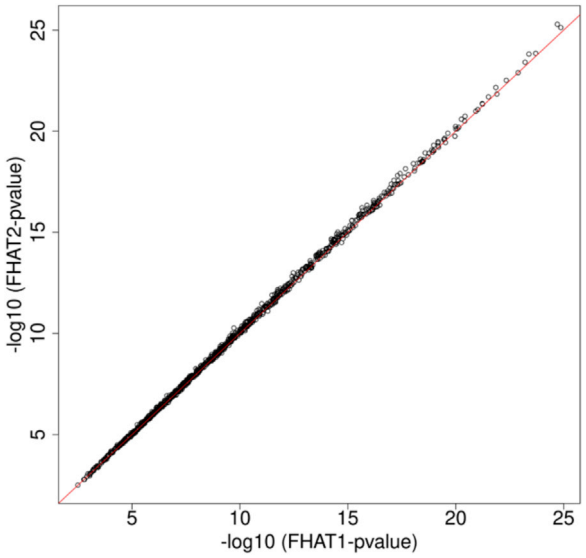

Figure 2. Comparison of −log10(p values) calculated with $FHAT_1$ and $FHAT_2$

with the proband, we conclude that it is feasible to analyze all relatives with various degrees of relationship through a single regression model and down-weight the probands' genotypes in their score calculation.

From the power comparisons between our methods to SKAT-LTFH, we show that our methods have a greater gain in power from the incorporation of FH because our methods are not limited to first-degree relatives and thus can include more data (Table 3). This is most pronounced in the scenarios where half of the variants in the region are risk increasing and the remaining variants are risk decreasing and at the lower alpha levels. In this simulation setting, FHAT incorporates FH from 2,500 parents and 2,500 aunts/uncles, while SKAT-LTFH only incorporates FH from 2,500 parents.

Simulation: Performance of proposed methods in complex family structure

The results summarized in Table 4 show the type I error rates for robust-famFHAT, robust-famFHAT-O, methods without case-control imbalance adjustment (famFHAT and famFHAT-O) with both probands' disease status and FH from relatives, and famSKAT and famSKAT-O with only probands' data for prevalence = 5%, 10%, 20%, and 50%. When the disease prevalence is low (i.e., 5%), methods such as famFHAT, famFHAT-O, famSKAT, and famSKAT-O that ignore the case-control imbalance have inflated type I error, but the type I error is controlled well when using the robust methods (Table 4). The results show that famFHAT, famFHAT-O, and famSKAT control type I error rates well in

Table 4. Type I error rates of robust methods and non-robust methods in related samples

Alpha	robust-famSKAT	famSKAT	robust-famFHAT	famFHAT	robust-famSKAT-O	famSKAT-O	robust-famFHAT-O	famFHAT-O
Prevalence = 50%								
0.1	*0.6*	1.0	*0.8*	1.2	*0.7*	1.0	*0.8*	1.1
0.05	*0.6*	1.1	*0.6*	1.2	*0.7*	1.0	*0.7*	1.1
0.01	*0.4*	0.9	*0.7*	1.2	*0.5*	1.0	*0.7*	1.2
0.005	*0.3*	*0.7*	*0.6*	1.3	*0.4*	0.9	*0.6*	1.4*
Prevalence = 20%								
0.1	*0.7*	1.0	*0.8*	1.1	*0.8*	0.9	*0.8*	1.0
0.05	*0.7*	1.0	*0.7*	1.1	*0.8*	1.0	0.9	1.0
0.01	*0.5*	0.8	*0.6*	1.0	*0.6*	1.1	0.8	1.1
0.005	*0.5*	0.8	*0.5*	1.0	*0.7*	1.0	0.8	1.3
Prevalence = 10%								
0.1	*0.8*	1.0	*0.9*	1.1	*0.9*	1.0	*0.9*	1.0
0.05	*0.8*	1.1	*0.9*	1.1	*0.8*	1.0	*0.9*	1.1
0.01	0.8	1.2	0.9	1.3*	0.9	1.3*	1.0	1.4*
0.005	0.8	1.3*	1.0	1.5*	0.9	1.3*	1.1	1.6*
Prevalence = 5%								
0.1	*0.8*	1.0	0.9	1.1	0.9	1.0	1.0	1.0
0.05	0.9	1.1	1.0	1.2	0.9	1.0	1.0	1.1
0.01	1.1	1.5*	1.1	1.5*	1.1	1.4*	1.1	1.4*
0.005	1.0	1.9*	1.0	1.9*	0.9	1.9*	1.1	1.7*

The number in each cell represents the ratio of type I error and expected significance level (column "alpha"). Type I error was evaluated from the proportion of p values less than or equal to corresponding to each alpha level. All methods used the same Wu weights with beta (MAF_i; 1, 25). The analyses were restricted to rare variants with MAF < 1%. Numbers have an asterisk if they are above the 95% confidence interval and are italic if they are below the 95% confidence interval.

Table 5. Type I error rates of famFHAT, famSKAT, famFHAT-O, famSKAT-O, FHAT, SKAT, FHAT-O, and SKAT-O in family samples

Alpha	famFHAT	famSKAT	famFHAT-O	famSKAT-O	FHAT	FHAT-O	SKAT	SKAT-O
Prevalence = 30%								
0.1	1.1	1.0	1.1	1.0	1.6*	1.5*	1.5*	1.4*
0.05	1.1	1.0	1.1	1.0	1.8*	1.7*	1.6*	1.5*
0.01	1.1	0.9	1.3	1.0	2.3*	2.2*	1.7*	1.8*
0.005	1.1	0.9	1.4	1.1	2.3*	2.5*	1.8*	2.0*
5×10^{-4}	1.2	1.0	1.9*	1.6*	3.2*	4.1*	2.7*	3.5*
Prevalence = 20%								
0.1	1.1	1.0	1.0	0.9	1.6*	1.4*	1.4*	1.3*
0.05	1.1	1.0	1.1	1.0	1.7*	1.6*	1.5*	1.4*
0.01	1.2	1.0	1.3	1.1	2.0*	2.1*	1.8*	1.9*
0.005	1.2	1.0	1.5*	1.2	2.3*	2.5*	2.0*	2.2*
5×10^{-4}	1.3	1.4	1.9*	2.0*	3.3*	4.1*	3.3*	4.1*

The number in each cell represents the ratio of type I error and expected significance level (column "alpha"). Type I error was evaluated from the proportion of p values less than or equal to each corresponding alpha level. All tests used the same Wu weights with beta (MAF_j; 1, 25). The analyses were restricted to rare variants with MAF < 1%. Numbers have an asterisk if they are above the 95% confidence interval.

the presence of familial correlation (Table 5) compared to the methods ignoring the family relatedness. However, a slight inflation in famSKAT-O is observed for the lower alpha level, which is consistent with what was observed in the prior paper.[3]

We estimate the power for 3,600 probands with available FH from 3,600 relatives with different degrees of relationship(Figure 3 and Figure 4). Two prevalences (P = 5% and P = 20%) are cosiderted in the analysis. All power is evaluated for alpha = 2.5×10^{-6} to account for multiple testing in a typical exome analysis testing 20,000 genes. Our methods outperform famSKAT and famSKAT-O after incorporating FH from parents in all scenarios. Overall, the robust methods perform similarly or better than the non-robust methods in a wide range of scenarios. When the proportion of causal variants is larger than or equal to 50%, the robust versions have slightly lower power than their non-robust versions when all variants in the region are risk increasing.

Application to the Framingham Heart Study

The family structures in FHS are very complex, and probands could have one relative, both parents, and/or multiple relatives. The degrees of relatedness are calculated on the basis of known familial relationships with the Kingship2 R package.[26] Table 6 summarizes the types of

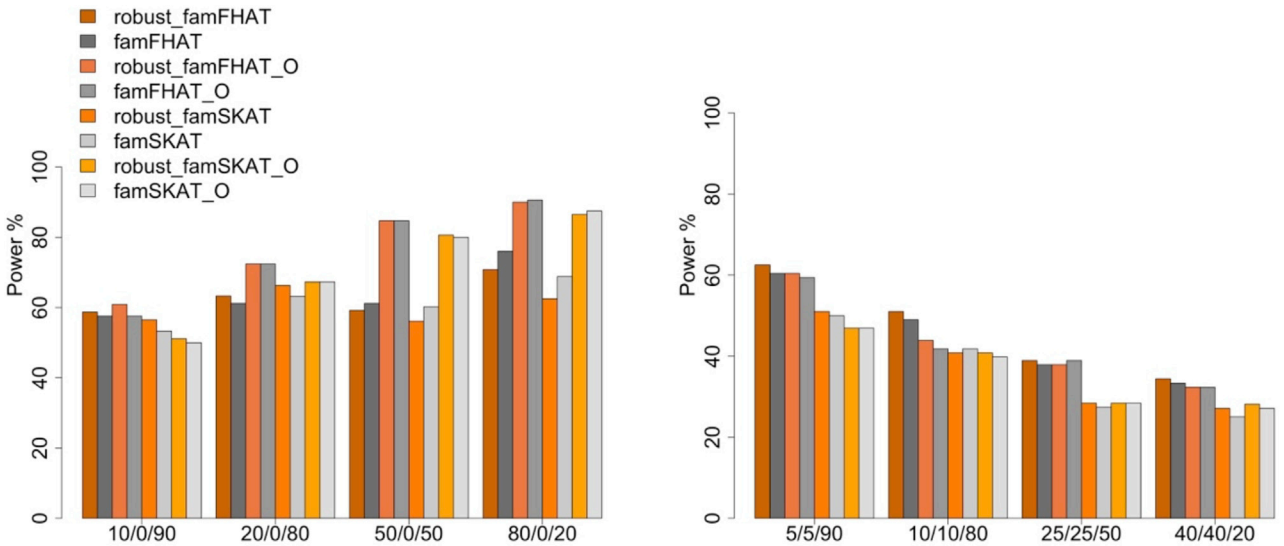

Figure 3. Empirical power of robust methods and non-robust methods in related samples at prevalence = 5%
In each plot, the x axis in the format of +/−/0 indicates the proportion of variants with positive, negative, and no effects. Each bar shows the empirical power evaluated as the proportion of p values less than or equal to alpha = 2.5×10^{-6}.

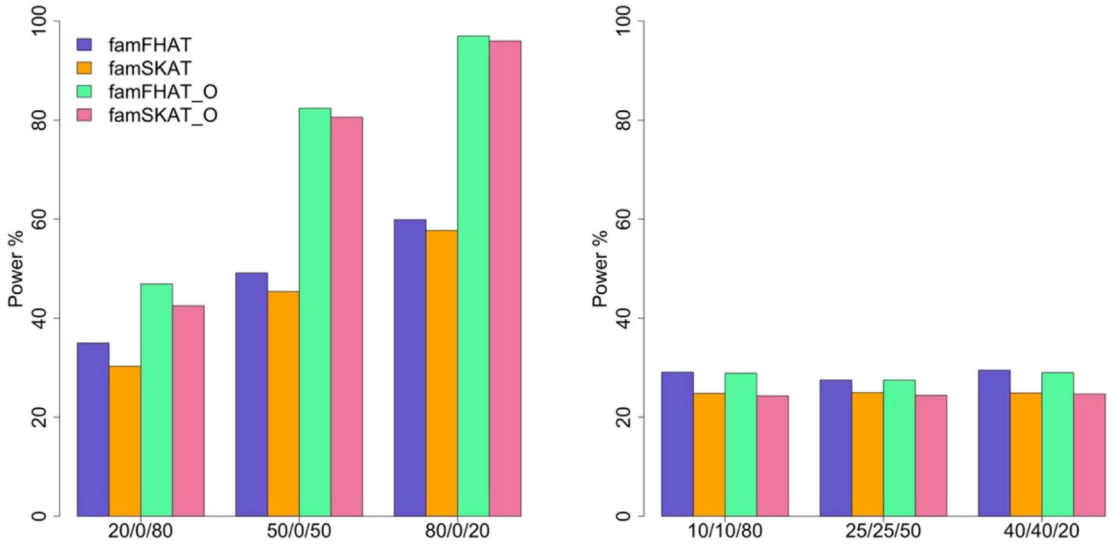

Figure 4. Empirical power of famFHAT, famFHAT-O, famSKAT, and famSKAT-O at prevalence = 20%
In each plot, the x axis in the format of +/−/0 indicates the proportion of variants with positive, negative, and no effects. Each bar shows the empirical power evaluated as the proportion of p values less than or equal to alpha = 2.5×10^{-6}.

relationships between relatives and probands among FHS families.

From investigating associations between previously identified genes for AD/dementia, we identify that *APOE* and *SLC24A4* have improved significance with famFHAT and famFHAT-O compared to famSKAT and famSKAT-O, respectively, for both AD and dementia (Table S1). The gene *INPP5D* has a more significant p value with famFHAT compared to famSKAT for AD after incorporating additional data from disease status from relatives.

When summarizing the results, we restrict the genes with cumulative MAC (cumMAC) > 20 in probands (Table 7, Table 8, and Figure 5). A total of 8,218 genes and 6,831 genes are selected to summarize the results of T2D and AD/dementia, respectively. The top genes in the table are selected via suggestive significance thresholds, which are calculated by 1 divided by the number of tested genes with cumMAC > 20 in probands (i.e., p value < 1.5×10^{-4} for AD/dementia, p value < 1.2×10^{-4} for T2D). Using famFHAT or famFHAT-O, we detect *KIAA0368* (MIM: 616694) for both AD and dementia, and we identify five more genes (*CYP26B1* [MIM: 605207], *CCR5* [MIM: 601373], *ODZ3* [MIM: 610083], *PYGM* [MIM: 608455], and *ZSCAN18*) for AD. Two genes (*PDE8A* [MIM: 602972] and *MAP3K3* [MIM: 602539]) are detected for T2D. *MAP3K3* has been shown to be associated with BMI, as BMI is the significant factor to develop T2D, which might suggest a possible association between *MAP3K3* and T2D. The novel findings on the suggestive cut-offs will need to be validated from replication analysis or functional studies. No genetic associations are detected with these three traits via stricter multiple-testing corrected thresholds (p value < 0.05/6,831 = 7.3×10^{-6} for AD/dementia; p value < 0.05/8,218 = 6.1×10^{-6} for T2D).

Table 6. Relationships between probands and relatives in the FHS

Relationship between proband and relatives	Kinship coefficient	Example type
First degree	0.25	parents/daughter/son
Second degree	0.125	grandparents/grandchildren/ aunt/uncle/niece/nephew
Third degree	0.0625	first cousin
Fourth degree	0.03125	great-great-grandparents/ great-great-grandchildren

As an illustration of robust-famFHAT and robust-famFHAT-O, we apply these methods to compare the results obtained with the non-robust methods. The genomic control inflation factors of robust-famFHAT and robust famFHAT-O ($\lambda_{\text{robust-famFHAT}} = 1.07$ for AD, $\lambda_{\text{robust-famFHAT}} = 1.08$ for dementia, and $\lambda_{\text{robust-famFHAT}} = 0.92$ for T2D, respectively; $\lambda_{\text{robust-famFHAT-O}} = 1.02$ for AD, $\lambda_{\text{robust-famFHAT-O}} = 1.0$ for dementia, and $\lambda_{\text{robust-famFHAT-O}} = 0.95$ for T2D, respectively), which are similar to those obtained with the famFHAT and famFHAT-O (Figure 5).

The p values for genes become insignificant via robust-famFHAT and robust-famFHAT-O compared to the results of the non-robust methods (famFHAT and famFHAT-O) for the T2D analysis. Although there are genes with very similar p values estimated with robust methods and non-robust methods, we find that *PYGM* is not associated with AD anymore with a suggestive significance threshold via the robust methods. However, there are scenarios where the robust-famFHAT yields smaller p values compared to the non-robust methods, such as *ZSCAN18* and *KIAA0368* for AD and dementia. Generally, the robust methods provide more conservative p values in the FHS analysis.

Table 7. Exome chip analysis of AD and dementia

Gene	Number of variants	cumMAC	robust-famFHAT (p value)	robust-famFHAT-O (p value)	famFHAT (p value)	famFHAT-O (p value)
AD						
CYP26B1	3	20	5.1×10^{-4}	4.8×10^{-4}	8.4×10^{-5}	8.8×10^{-5}
CCR5	7	70	4.9×10^{-4}	2.3×10^{-4}	4.4×10^{-4}	9.9×10^{-5}
ODZ3	13	92	2.3×10^{-4}	3.2×10^{-4}	9.7×10^{-5}	1.2×10^{-4}
KIAA0368	9	108	1.6×10^{-4}	6.2×10^{-5}	2.2×10^{-4}	3.6×10^{-5}
PYGM	14	133	4.6×10^{-4}	3.8×10^{-4}	8.4×10^{-5}	7.4×10^{-5}
ZSCAN18	2	38	4.4×10^{-5}	4.1×10^{-5}	7.6×10^{-5}	7.0×10^{-5}
Dementia						
KIAA0368	9	108	5.7×10^{-5}	2.7×10^{-5}	7.1×10^{-5}	2.2×10^{-5}

cumMAC is the cumulative minor allele counts in probands for the gene we tested. Number of variants is the total number of variants tested in the gene. The suggestive significance threshold is $(1/6,831) = 1.5 \times 10^{-4}$.

Discussion

In this work, we proposed powerful approaches, famFHAT and famFHAT-O, to account for correlation among family members when incorporating FH information in aggregation unit-based tests. Our methods are able to incorporate FH from all types of relatives with different degrees of relationship with probands, whereas LT-FH is restricted to the incorporation of FH from parents and siblings. LT-FH incorporates the FH through an adjustment to the probands' phenotype, whereas our proposed methods are able to include the relatives with available FH into the analysis through a meta-analysis framework, which greatly increases the sample sizes. Through a simulation analysis, we showed that our proposed methods have greater power than LT-FH by incorporating additional relatives from more distant relatives (while LT-FH can only incorporate first-degree relatives). The robust versions of those methods allow the incorporation of FH for rare variant analysis with enhanced statistical evidence to detect associations for complex disease when the case-control ratio is unbalanced. We adopted the GLMM to analyze related probands with both genotypes and phenotypes or related relatives with phenotypes. In our analyses, we used the glmmkin function from the R package GMMAT to run the GLMM and used the obtained estimates for score calculations.[13] Analysts can use the relatedness indices between individuals (such as KING or the known kinship matrix) as the relationship matrix. The robust versions were proposed on the basis of cutting-edge approaches of SPA and ER. We showed that our proposed methods are flexible and can be applied to both continuous traits and binary traits by fitting appropriate models. Our methods can also account for the ascertainment that commonly happens in the case-control studies because our basic likelihood function can hold the form of the probability of probands' genotypes and FH conditional on the probands' phenotypes and other covariates of interest, i.e., $P(G^P, Y^R | Y^P, X^P, X^R)$.[9]

We showed that we obtained correct type I error rates, and similar results compared to the score that was calculated by combining the weighted scores from relatives, when combining relatives with different degrees of relationship with the proband into a single model by down-weighting their genotypes before calculating the scores (see the first simulation study). This demonstration allows us to combine available FH from multiple relatives into an aggregation unit-based association test so that the methods are not restricted to FH from a single relative type. We noted that one limitation is that this approach relies on the known relationship, and it cannot be implemented for unknown relationships among relatives via an empirical kinship matrix because the empirical kinship matrix needs to be estimated with observed genotypes from probands and relatives. When the proband has FH from multiple relatives, we took the genotype average and phenotype average among the closest proband in their score calculation. To investigate the performance of the robust methods and non-robust methods when exploiting multiple relatives with different relationships to

Table 8. Exome chip analysis of T2D

Gene	Number of variants	cumMAC	robust-famFHAT (p value)	robust-famFHAT-O (p value)	famFHAT (p value)	famFHAT-O (p value)
PDE8A	4	22	1.9×10^{-2}	1.1×10^{-3}	4.2×10^{-3}	1.1×10^{-4}
MAP3K3	3	29	1.8×10^{-3}	2.2×10^{-4}	8.7×10^{-4}	1.2×10^{-4}

cumMAC is the cumulative minor allele counts in probands for the gene we tested. Number of variants is the total number of variants tested in the gene. The suggestive significance threshold is $(1/8,218) = 1.2 \times 10^{-4}$.

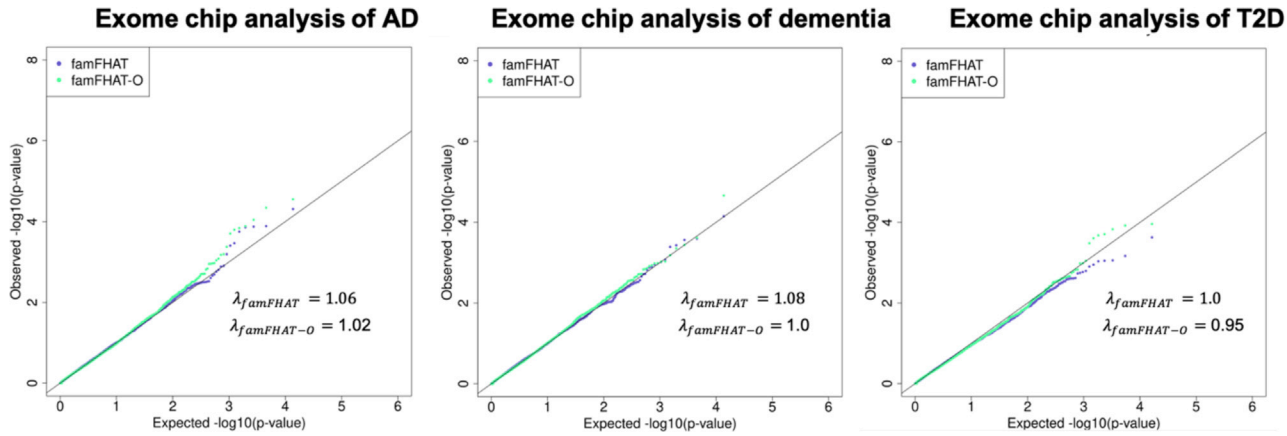

Figure 5. Quantile-quantile plots from FHS exome chip analysis for AD, dementia, and T2D with non-robust methods

the proband, the simulation analysis was conducted by simulating complex family structures with 18 family members from three generations. We showed that famFHAT and famFHAT-O have appropriate type I error rates in the presence of correlation and maintain greater power than famSKAT and famSKAT-O after combining additional phenotype data from relatives. Because famFHAT-O combines the features of famFHAT and famFHAT-Burden, it has the most robust power in all the scenarios we've tested. The correct type I error using robust methods for unbalanced case-control ratios was demonstrated in simulation analyses, and we also showed that non-robust methods have large inflation in type I error when the prevalence is low. For power, the robust methods perform as well or better compared to their non-robust versions that do not account for case-control ratios in most cases.

We applied our methods to analyze AD, dementia, and T2D by using FHS exome chip data. Because the disease prevalence is not too rare ($p > 10\%$), we used famFHAT and famFHAT-O to incorporate all available disease status data from relatives. Three of the known rare variant regions for AD/dementia were shown to have improved significance via our methods compared to the methods without adjustment of FH. By testing all regions with the exome chip data, we found novel regions for AD, dementia, and T2D by using our methods. Compared to the non-robust methods (famFHAT, famFHAT-O, famSKAT, famSKAT-O), the robust methods (robust-famFHAT, robust-famFHAT-O, robust-famSKAT, robust-famSKAT-O) provided more conservative p values for some genes. Because the traits are not too rare in the FHS, the robust and non-robust methods yielded similar values for the inflation factors. While the type I error is well controlled with famFHAT and famFHAT-O in family studies, there is no need to limit the analysis to an unrelated set, thus maintaining a large sample size in most studies. These methods boost power to detect rare variant associations in studies with correlated data by further incorporating FH. With the accurate results provided in the robust methods, our methods will significantly contribute to the identification of trait-associated rare variants.

Data and code availability

The FHS data used in this paper are available on dbGAP (dbGaP: phs000007.v32.p13, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v32.p13). The famFHAT and famFHAT-O functions are available at https://sites.bu.edu/fhspl/publications/fhat/.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.ajhg.2022.03.001.

Acknowledgments

Y.W., G.M.P., A.L.D., and J.D. acknowledge the grant support (5U01AG058589) from the National Institute of Aging. J.B.M. acknowledges the grant support (R01 HL151855) and H.C. acknowledges the grant support (R00 HL130593) from the National Institute of Health. This work was also partially supported by National Institute of Health grant UM1 DK078616.

Declaration of interests

The authors declare no competing interests.

Received: October 21, 2021
Accepted: February 28, 2022
Published: March 21, 2022

Web resources

GMMAT, https://github.com/hanchenphd/GMMAT
OMIM, https://omim.org/

References

1. 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.
2. 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.
3. Lee, S., Wu, M.C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. Biostatistics *13*, 762–775.
4. Neale, B.M., Rivas, M.A., Voight, B.F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S.M., Roeder, K., and Daly, M.J. (2011). Testing for an unusual distribution of rare variants. PLoS Genet. *7*, e1001322.
5. Chen, H., Meigs, J.B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. Genet. Epidemiol. *37*, 196–204.
6. Jiang, D., and McPeek, M.S. (2014). Robust rare variant association testing for quantitative traits in samples with related individuals. Genet. Epidemiol. *38*, 10–20.
7. Schaid, D.J., McDonnell, S.K., Sinnwell, J.P., and Thibodeau, S.N. (2013). Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet. Epidemiol. *37*, 409–418.
8. Zhong, S., Jiang, D., and McPeek, M.S. (2016). CERAMIC: Case-Control Association Testing in Samples with Related Individuals, Based on Retrospective Mixed Model Analysis with Adjustment for Covariates. PLoS Genet. *12*, e1006329.
9. Wang, Y., Chen, H., Peloso, G.M., DeStefano, A.L., and Dupuis, J. (2021). Exploiting family history in aggregation unit-based genetic association tests. Eur. J. Hum. Genet. https://doi.org/10.1038/s41431-021-00980-0.
10. Chen, H., Wang, C., Conomos, M.P., Stilp, A.M., Li, Z., Sofer, T., Szpiro, A.A., Chen, W., Brehm, J.M., Celedón, J.C., et al. (2016). Control for Population structure and relatedness for binary traits in Genetic association studies via logistic mixed models. Am. J. Hum. Genet. *98*, 653–666.
11. Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.
12. Hujoel, M.L.A., Gazal, S., Loh, P.R., Patterson, N., and Price, A.L. (2020). Liability threshold modeling of case-control status and family history of disease increases association power. Nat. Genet. *52*, 541–547.
13. Chen, H., Huffman, J.E., Brody, J.A., Wang, C., Lee, S., Li, Z., Gogarten, S.M., Sofer, T., Bielak, L.F., Bis, J.C., et al. (2019). Efficient Variant Set Mixed Model Association Tests for Continuous and Binary Traits in Large-Scale Whole-Genome Sequencing Studies. Am. J. Hum. Genet. *104*, 260–274.
14. 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.
15. Zhou, W., Nielsen, J.B., Fritsche, L.G., Dey, R., Gabrielsen, M.E., Wolford, B.N., LeFaive, J., VandeHaar, P., Gagliano, S.A., Gifford, A., et al. (2018). Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. Nat. Genet. *50*, 1335–1341.
16. Zhao, Z., Bi, W., Zhou, W., VandeHaar, P., Fritsche, L.G., and Lee, S. (2020). UK Biobank whole-exome sequence binary phenome analysis with robust region-based rare-variant test. Am. J. Hum. Genet. *106*, 3–12.
17. Zhou, W., Zhao, Z., Nielsen, J.B., Fritsche, L.G., LeFaive, J., Gagliano Taliun, S.A., Bi, W., Gabrielsen, M.E., Daly, M.J., Neale, B.M., et al. (2020). Scalable generalized linear mixed model for region-based association tests in large biobanks and cohorts. Nat. Genet. *52*, 634–639.
18. Su, Z., Marchini, J., and Donnelly, P. (2011). HAPGEN2: simulation of multiple disease SNPs. Bioinformatics *27*, 2304–2305.
19. Dawber, T.R., Kannel, W.B., and Lyell, L.P. (1963). An approach to longitudinal studies in a community: the Framingham Study. Ann. N Y Acad. Sci. *107*, 539–556.
20. Kannel, W.B., Feinleib, M., McNamara, P.M., Garrison, R.J., and Castelli, W.P. (2017). An investigation of coronary heart disease in families: The Framingham offspring study. Am. J. Epidemiol. *185*, 1093–1102.
21. Bachman, D.L., Wolf, P.A., Linn, R., Knoefel, J.E., Cobb, J., Belanger, A., D'Agostino, R.B., and White, L.R. (1992). Prevalence of dementia and probable senile dementia of the Alzheimer type in the Framingham Study. Neurology *42*, 115–119.
22. McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., and Stadlan, E.M. (1984). Clinical diagnosis of Alzheimer's disease: report of the NINCDS-ADRDA Work Group under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. Neurology *34*, 939–944.
23. Grove, M.L., Yu, B., Cochran, B.J., Haritunians, T., Bis, J.C., Taylor, K.D., Hansen, M., Borecki, I.B., Cupples, L.A., Fornage, M., et al. (2013). Best practices and joint calling of the HumanExome BeadChip: the CHARGE Consortium. PLoS ONE *8*, e68095.
24. Bellenguez, C., Küçükali, F., Jansen, I., Andrade, V., Moreno-Grau, S., Amin, N., Naj, A.C., Grenier-Boley, B., Campos-Martin, R., Holmans, R.A., et al. (2020). New insights on the genetic etiology of Alzheimer's and related dementia. Preprint at medRxiv. https://doi.org/10.1101/2020.10.01.20200659.
25. Wessel, J., Majarian, T.D., Highland, H.M., Raghavan, S., Szeto, M.D., Hasbani, N.R., de Vries, P.S., Brody, J.A., Sarnowski, C., DiCorpo, D., et al. (2020). Rare non-coding variation identified by large scale whole genome sequencing reveals unexplained heritability of type 2 diabetes. Preprint at medRxiv. https://doi.org/10.1101/2020.11.13.20221812.
26. Sinnwell, J.P., Therneau, T.M., and Schaid, D.J. (2014). The kinship2 R package for pedigree data. Hum. Hered. *78*, 91–93.