

## ARTICLE



# Exploiting family history in aggregation unit-based genetic association tests

Yanbing Wang<sup>1</sup>✉, Han Chen<sup>2,3</sup>, Gina M. Peloso<sup>1</sup>, Anita L. DeStefano<sup>1</sup> and Josée Dupuis<sup>1</sup>✉

© The Author(s), under exclusive licence to European Society of Human Genetics 2021

The development of sequencing technology calls for new powerful methods to detect disease associations and lower the cost of sequencing studies. Family history (FH) contains information on disease status of relatives, adding valuable information about the probands' health problems and risk of diseases. Incorporating data from FH is a cost-effective way to improve statistical evidence in genetic studies, and moreover, overcomes limitations in study designs with insufficient cases or missing genotype information for association analysis. We proposed family history aggregation unit-based test (FHAT) and optimal FHAT (FHAT-O) to exploit available FH for rare variant association analysis. Moreover, we extended liability threshold model of case-control status and FH (LT-FH) method in aggregated unit-based methods and compared that with FHAT and FHAT-O. The computational efficiency and flexibility of the FHAT and FHAT-O were demonstrated through both simulations and applications. We showed that FHAT, FHAT-O, and LT-FH methods offer reasonable control of the type I error unless case/control ratio is unbalanced, in which case they result in smaller inflation than that observed with conventional methods excluding FH. We also demonstrated that FHAT and FHAT-O are more powerful than LT-FH and conventional methods in many scenarios. By applying FHAT and FHAT-O to the analysis of all cause dementia and hypertension using the exome sequencing data from the UK Biobank, we showed that our methods can improve significance for known regions. Furthermore, we replicated the previous associations in all cause dementia and hypertension and detected novel regions through the exome-wide analysis.

*European Journal of Human Genetics* (2022) 30:1355–1362; <https://doi.org/10.1038/s41431-021-00980-0>

## INTRODUCTION

Genome-wide association studies (GWAS) have identified thousands of genetic variants associated with complex diseases at the genome-wide significance level ( $p < 5 \times 10^{-8}$ ). Most of the variants identified by GWAS are common variants with minor allele frequency (MAF)  $\geq 1\%$ , and most of these variants display modest effect sizes and can only explain a small portion of the total heritability of complex diseases. Yet, rare variants (MAF  $< 1\%$ ) are important to uncovering unexplained heritability and discovering novel genes contributing to complex diseases [1–3]. Because standard association approaches testing each variant individually are grossly underpowered for rare variants, aggregation unit-based methods that jointly analyze variants have been proposed to improve power to detect rare variant associations. Aggregation unit-based approaches include, among others, the sequence kernel association test (SKAT) [4], Burden tests [5–7], SKAT-O [8], and aggregated Cauchy association test (ACAT) [9]. However, power of these methods to identify disease regions can be limited by insufficient number of cases in unascertained cohorts.

In genetic association studies, family history (FH) of disease in relatives is often collected in large population cohorts. FH provides an overview of a phenotype within families. Such information typically includes phenotypes of un-genotyped

parents or more distant relatives of probands. FH is related to the genotypes of probands at disease loci based on the Mendelian laws of transmission, and is important in assessing health problems and risk of diseases [10–12]. While collecting cases is expensive, incorporating FH information into standard case-control genetic association analyses is a cost-effective way to potentially increase statistical power [11, 13–15]. Many study designs have limitations for genetic research of late-onset diseases such as Alzheimer's disease (AD), because disease cases may be deceased with unavailable genotype data. The standard statistical association tests in younger cohorts with low prevalence of some late-onset diseases are not powerful to identify genetic regions associated with a trait of interest. In contrast, the incorporation of available information of disease status in the form of FH may increase the sample size in cohorts with limited cases or individuals with unavailable genotypes. Genetic association studies using only cases and controls will greatly benefit by incorporating available FH information to detect associations.

FH cannot be directly incorporated in standard genetic association methods, limiting its use in genetic association testing. FH has been included as a covariate to improve disease prediction [16], or used to infer mode of inheritance to construct statistical tests [17]. However, there are a few reported methods that allow

<sup>1</sup>Department of Biostatistics, School of Public Health, Boston University, Massachusetts, MA 02215, USA. <sup>2</sup>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. <sup>3</sup>Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030, USA. ✉email: yanbing@bu.edu; dupuis@bu.edu

Received: 24 March 2021 Revised: 20 July 2021 Accepted: 4 October 2021

Published online: 25 October 2021

FH to be exploited in genetic association analysis to improve statistical power to detect disease loci. The method developed by Ghosh et al. [13] enables the incorporation of FH as a phenotype into the standard single variant analysis, and the results confirmed that exploiting the information contained in FH substantially boosts power to detect the individual variant at disease loci. Nevertheless, these single variant tests suffer from loss of power to detect rare variant associations. While numerous aggregation unit-based methods to jointly analyze rare variants have been proposed to improve power to detect rare variant associations, aggregation unit-based methods that can directly incorporate FH information are needed.

We developed a new and powerful method of family history aggregation unit-based test (FHAT) that enables the incorporation of FH to enhance the statistical power for rare variant associations. We also developed an optimal unified test FHAT-O to maintain robust power in complex scenarios regardless of directions of genetic effects or the proportion of causal variants. To make the comparison with the recent developed method, liability threshold model of case-control status and FH (LT-FH) [11], we proposed a novel way to utilize LT-FH into aggregation unit-based method for rare variant analysis. We performed an extensive simulation study to evaluate the type I error and power of FHAT and FHAT-O under various scenarios, and illustrated the methods using whole exome sequencing data from the UK Biobank.

**MATERIAL AND METHODS**

**Family history aggregation unit-based test (FHAT)**

We propose a novel approach, FHAT, to incorporate FH information in the aggregation unit-based tests. We assume that there are  $n$  probands with  $m$  observed variants included in the aggregation unit-based test. When we have FH on the relative of the probands, let  $Y_i^p$  denotes the phenotype of the  $i$ th proband;  $Y_i^r$  denotes the phenotype of the relative of the  $i$ th proband, respectively;  $G_i^p$  denotes the genotypes of the  $i$ th proband;  $X_i^p$  denotes covariates for the  $i$ th proband;  $X_i^r$  denotes covariates of the relative of the  $i$ th proband, such as age and ancestral principal components (PCs) that account for population structure. The probability of observing  $(Y_i^p, Y_i^r)$  conditional on  $G_i^p$  can be written as follows (see details in the Supplementary Method):

$$P(Y_i^p, Y_i^r | G_i^p, X_i^p, X_i^r) = P(Y_i^p | G_i^p, X_i^p) P(Y_i^r | G_i^p, Y_i^p, X_i^r) \tag{1}$$

Therefore, the evidence for association can be assessed from two separate analyses for probands and relatives. We assume an additive model and coding the genotypes in  $G_i^p$  as the number of minor alleles. One can also use dominant or recessive models by coding the variants appropriately. Based on  $P(Y_i^p | G_i^p)$ , we first assess the association between probands' genotypes and their disease status using

$$g(E(Y_i^p | G_i^p, X_i^p)) = X_i^p \alpha_p + G_i^p \beta_p \tag{2}$$

where  $g(\cdot)$  is the link function,  $\alpha_p$  is a vector of regression coefficients for covariate effects,  $\beta_p$  is a vector of regression coefficients for the observed genotypes in probands. The model for relatives based on  $P(Y_i^r | G_i^p, Y_i^p)$  is specified as follows:

$$g(E(Y_i^r | G_i^p, Y_i^p, X_i^r)) = X_i^r \alpha_r + G_i^p \beta_r + Y_i^p \lambda_r \tag{3}$$

where  $\lambda_r$  is scalar of regression coefficients for probands' phenotypes for the relatives' model;  $\alpha_r$  is vector of regression coefficients for relatives' covariates;  $\beta_r$  is vector of regression coefficients for  $m$  observed variants in probands. Inclusion of  $Y_i^p$  is necessary to make the analyses of probands and relatives independent based on (1). This relatives' model (3) can analyze FH from unrelated relatives, i.e., single relative per probands or FH from both parents since mothers and fathers are conditional independent. We observe that the two underlying association estimators,  $(\beta_p, \beta_r)$ , have the relationship [18] of  $\beta_r \approx 2\Omega\beta_p$  where  $\Omega$  is the kinship coefficient between probands and their relatives and  $\Omega = \frac{1}{4}$  for first-degree relatives such as parents.

Conventional aggregation unit-based methods evaluate the association between a set of variants and phenotype among probands. One such aggregation unit-based method is called the SKAT [4]. The weighted score

statistic based on the probands' model (2) is

$$Q_{SKAT} = \frac{(Y^p - \hat{\mu}_p)^T G^p W W G^{pT} (Y^p - \hat{\mu}_p)}{\hat{\phi}_p^2}$$

where  $W = \text{diag}(w_1, w_2, \dots, w_m)$  is a pre-specified weight matrix for  $m$  variants;  $G^p$  is a  $n \times m$  genotype matrix with  $(i, j)$ th element corresponding to the additively coded genotype for variant  $j$  of proband  $i$ ;  $\hat{\mu}_p$  is the estimated mean of  $Y^p$  using the null model with only covariates;  $\hat{\phi}_p$  is the estimate of dispersion parameter in the generalized linear model that is related to the variance of the distribution under  $H_0$ : for binary outcomes it is fixed to 1, and for continuous outcomes it is the variance of the random errors. The score statistic can be obtained similarly to evaluate whether genetic variants are associated with disease status using the relatives' phenotypes to replace the probands' phenotypes based on relatives' model (3). The pre-specified weights can be a function of MAF. For example, one can use Wu's weights [4]  $w_j = \text{Beta}(\text{MAF}_j, 1, 25)$  to up-weight the effect of rarer variants.

We propose to combine the score statistics from the two association models for probands and their relatives using a weighted meta-analysis to increase the flexibility of incorporating relatives with different degrees of relatedness (thus different kinship coefficients), as well as different numbers of available relatives for each proband. Meta-analysis is often used in genetic association analysis to increase the power by combining results from multiple studies. Methods to meta-analyze SKAT results have been developed [19]. Meta-analysis of rare variant association tests proposed is based on the study-specific summary statistics, that is, score statistics for each variant and linkage disequilibrium estimates in a region. Because of the genetic relationship between probands and their relatives, we down-weight the scores for relatives by  $2\Omega$  when combining the score statistics in a meta-analysis by assuming the homogeneous genetic effects among probands and their relatives. Specifically, because relative  $k$  of each proband may or may not have phenotype data available, we use  $Y^{Rk}$  to denote the collective phenotype vector for relative  $k$  of all probands (e.g., all mothers), including missing values, with kinship coefficient  $\Omega_k$ . The diagonal matrix  $D(R_k)$  indicates whether corresponding element in  $Y^{Rk}$  for each proband is missing (denoted by 0) or not (denoted by 1). Therefore, relatives with missing phenotype data do not contribute to the test statistic. We fit a single relative model jointly using all relatives' phenotypes and covariates conditional on their probands' phenotypes to get  $\hat{\mu}_{Rk}$ , the estimated mean vector of  $Y^{Rk}$  for relative  $k$  of all probands, as well as the dispersion parameter estimate  $\hat{\phi}_R$  under the null hypothesis of no genetic effects. We assume that all relatives are independent in the model. The general form of FHAT statistics that incorporates FH from relatives is

$$Q_{FHAT} = \left[ \frac{(Y^p - \hat{\mu}_p)^T}{\hat{\phi}_p} + \sum_k \frac{2\Omega_k D(R_k) (Y^{Rk} - \hat{\mu}_{Rk})^T}{\hat{\phi}_R} \right] G^p W W G^{pT} \left[ \frac{(Y^p - \hat{\mu}_p)}{\hat{\phi}_p} + \sum_k \frac{2\Omega_k D(R_k) (Y^{Rk} - \hat{\mu}_{Rk})}{\hat{\phi}_R} \right] \tag{4}$$

Under the null hypothesis,  $Q_{FHAT}$  follows a weighted sum of  $\chi^2$  distributions with 1 degree of freedom,  $Q_{FHAT} \sim \sum_{j=1}^m \lambda_j \chi_{1j}^2$ . The weights  $\lambda_j$  can be estimated from the eigenvalues of  $W G^{pT} (\hat{P} + \sum_k 4\Omega_k^2 D(R_k) \hat{P}_R D(R_k)) G^p W$ , where  $\hat{P}$  and  $\hat{P}_R$  are the projection matrices in probands and relatives  $k$ , respectively, see the Supplementary Method. The  $p$  value can be estimated by Davies' method [20]. The general form can be reduced to

$$Q_{FHAT} = \left[ (Y^p - \hat{\mu}_p)^T + \frac{D(R_m) (Y^{Rm} - \hat{\mu}_{Rm})^T}{2} + \frac{D(R_f) (Y^{Rf} - \hat{\mu}_{Rf})^T}{2} \right] G^p W W G^{pT} \left[ (Y^p - \hat{\mu}_p) + \frac{D(R_m) (Y^{Rm} - \hat{\mu}_{Rm})}{2} + \frac{D(R_f) (Y^{Rf} - \hat{\mu}_{Rf})}{2} \right] \tag{5}$$

for incorporating FH from both parents (with mothers denoted by  $m$  and fathers denoted by  $f$ ) when using logistic regression models for binary trait with the estimates of dispersion parameters fixed to 1 (i.e.,  $\hat{\phi}_p = \hat{\phi}_R = 1$ ), and the kinship coefficients ( $\Omega_m, \Omega_f$ ) fixed to  $\frac{1}{4}$ .

**Optimal FHAT (FHAT-O)**

Using the same framework adopted in FHAT, we develop a FHAT-O statistic based on the optimal unified test SKAT-O [8]. Since SKAT-O combines the

**Table 1.** The empirical type I error rate divided by the significance level.

Alpha	FHAT	SKAT	FHAT-O	SKAT-O	Burden	ACAT-V
<i>Prevalence = 10% in probands</i>						
$2.5 \times 10^{-4}$	1.37	2.43	1.63	2.71	1.2	1.31
$2.5 \times 10^{-6}$	3.16	8.64	3.78	10.92	1.96	2.06
<i>Prevalence = 20% in probands</i>						
$2.5 \times 10^{-4}$	1.01	1.32	1.24	1.63	1.05	1.19
$2.5 \times 10^{-6}$	1.30	2.72	1.98	4.02	1.64	1.36
<i>Prevalence = 50% in probands</i>						
$2.5 \times 10^{-4}$	0.88	0.80	1.10	1.08	0.98	0.88
$2.5 \times 10^{-6}$	0.60	0.54	0.98	0.98	0.96	0.90

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  $2.5 \times 10^{-4}$  and  $2.5 \times 10^{-6}$  using 20 million simulation replicates for prevalence = 5%, 10%, 20%, and 50%. The total sample size of probands was 5000. FHAT, SKAT-LTFH, SKAT, FHAT-O, SKATO-LTFH, SKAT-O, and Burden test all used the same Wu weights with beta (MAF; 1, 25). ACAT-V used the weight of  $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j(1 - MAF_j)}$  to make results comparable. FHAT, SKAT-LTFH, FHAT-O, and SKATO-LTFH analyzed probands and incorporated the FH information, while SKAT, SKAT-O, Burden test, and ACAT-V only included probands. The LTFH phenotype was computed using LT-FH software v2 and then used as the continuous outcome in SKAT and SKAT-O to obtain SKAT-LTFH and SKATO-LTFH.

feature of SKAT and Burden tests, the power is robust in the presence of both different and same directions of causal variant effects.

We first develop a FHAT-Burden, which is a weighted sum of the weighted score statistics in probands, and relatives based on their relationships (Supplementary Method). Then we propose unified test defining as the weighted average of FHAT and FHAT-Burden:

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

where the weight  $\rho$  can be estimated to minimize the  $p$  value using the procedure proposed by Lee et al. [21]. When  $\rho = 1$ ,  $Q_\rho$  reduces to FHAT-Burden, and when  $\rho = 0$ ,  $Q_\rho$  is equivalent to FHAT. The statistic for optimal test FHAT-O that combines the features of FHAT and FHAT-Burden is determined as follows:

$$Q_{FHAT-O} = \min_{0 \leq \rho \leq 1} P_\rho \quad (6)$$

where  $P_\rho$  is the  $p$  value estimated for each given  $\rho$  (more details are in the Supplementary Method).

### Simulation analysis

Simulations were performed to evaluate the FHAT and FHAT-O statistics in terms of empirical type I error and statistical power. We generated 10,000 haplotypes for a 4 kb region on chromosome 19 using HapGen2 software [22]. The data from 1000 genomes project were used as the reference panel to simulate haplotypes. In all simulations, we focused on binary traits because they are more often collected through questionnaire in relatives and we focused on rare variants with  $MAF < 1\%$ . We used the definition from Chen et al. [23] to calculate the genetic effect size. We simulated the probands with both genotypes and phenotypes, and available FH data from both parents. We used LT-FH phenotype in SKAT (SKAT-LTFH) and SKAT-O (SKATO-LTFH) and compared the results to FHAT and FHAT-O, and they were all calculated by combining the FH from relatives (i.e., mothers and fathers) into the analysis. The standard methods (SKAT, SKAT-O, Burden test, and ACAT-V) only used proband data in the analysis. Because mothers and fathers were simulated as independent samples, they were analyzed using a single relatives' model (3) and then FHAT and FHAT-O statistics were calculated using (5) and (6). The type I error and power of FHAT and FHAT-O were compared to SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test, and ACAT-V. Note that ACAT-V is an aggregation unit-based test combining variant-level  $p$  values using ACAT. The detailed description of type I error and power simulations can be found in the Supplementary Method.

### Analysis of whole exome sequencing data in the UK Biobank

The UK Biobank is a large prospective cohort study with information on clinical traits, covariates, and genome-wide genotype data for over 500,000 individuals with age at assessment between 37 and 73 years at baseline (2006–2010). The second tranche of exome sequence data of approximately 4 million coding variants for 200,000 individuals has been recently

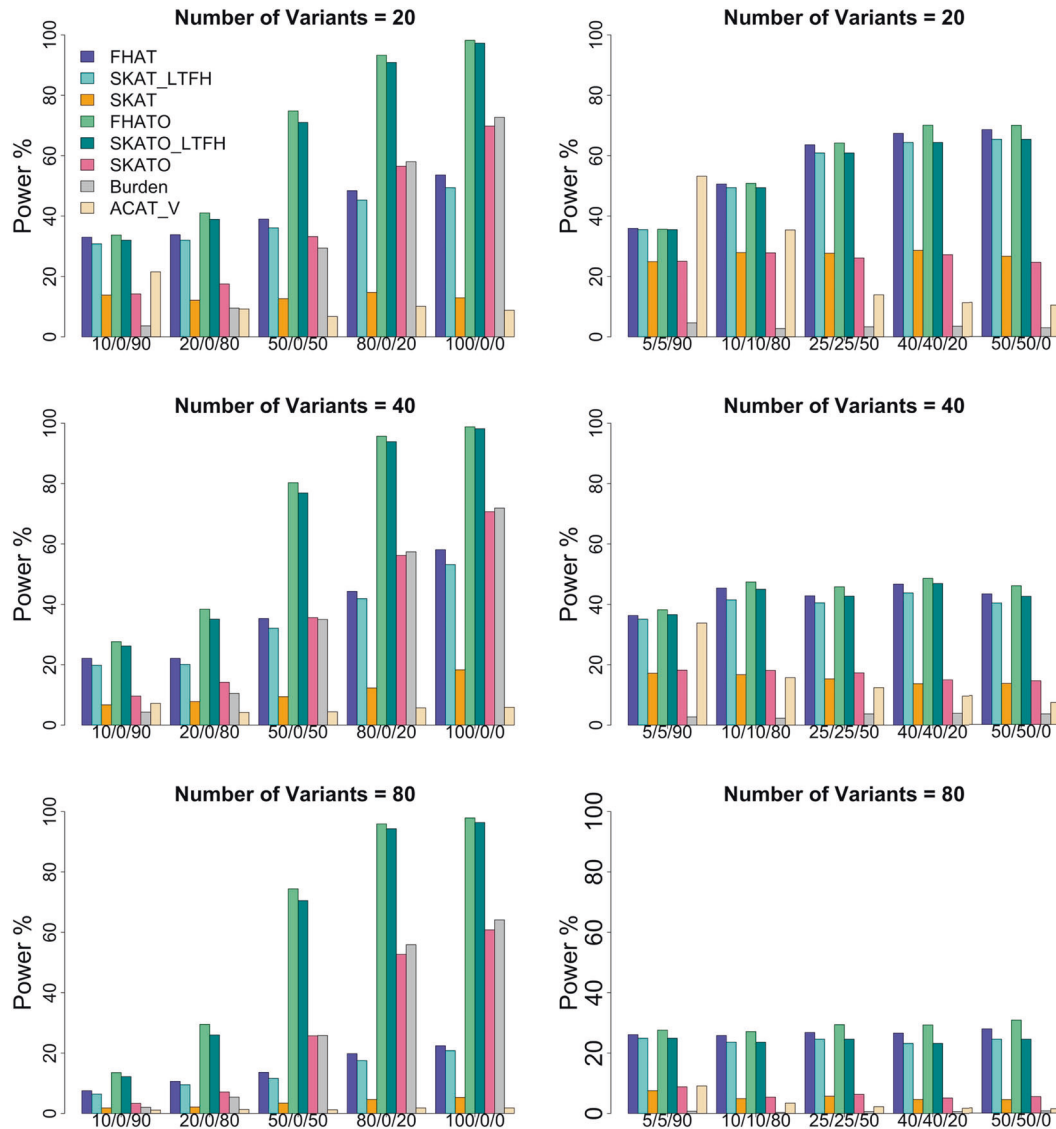
completed in the UK Biobank. FH of all cause dementia and hypertension was collected from questionnaires. Rare variant (with  $MAF < 1\%$ ) gene-based analyses detailed in the Supplementary Method were conducted to analyze all cause dementia and hypertension in the UK Biobank data.

## RESULTS

### Type I error and power

A total of 20 million simulation replicates were first generated to evaluate type I error at various alpha levels for FHAT, FHAT-O, SKAT, SKAT-O, Burden test, and ACAT-V using 5000 probands with available parental history (Table 1). SKAT and SKAT-O have inflated type I error for prevalence = 20%, while the type I error is controlled better in FHAT, FHAT-O. When the disease prevalence is low (i.e., 10%), FHAT and FHAT-O have inflated type I error, especially for exome-wide significance (alpha =  $2.5 \times 10^{-6}$ ), but the inflation is smaller than that was observed with SKAT and SKAT-O after incorporating additional cases in relatives. A slightly deflated type I error was observed in FHAT and SKAT for prevalence = 50%. The conservativeness of SKAT when the prevalence is 50% was also observed in prior publications [4, 8]. Burden test and ACAT-V control the type I error relatively better in some scenarios shown in Table 1. By comparing the type I error of the methods shown in Table 1 to SKAT-LTFH and SKATO-LTFH (Supplementary Table S2), FHAT and FHAT-O yield similar type I error results as in SKAT-LTFH and SKATO-LTFH, respectively. The type I error for the LTFH methods was evaluated at alpha level as low as to  $2.5 \times 10^{-5}$  to reduce the computational cost.

Figure 1 summarizes the power simulation results of FHAT, SKAT-LTFH, SKAT, FHAT-O, SKATO-LTFH, SKAT-O, Burden test, and ACAT-V for disease prevalence = 20% at alpha =  $2.5 \times 10^{-6}$ . Additional power results for prevalence = 50% and other alpha levels can be found in Supplementary Figs. S1–S3. The causal variants in a region were set to have positive effects, or half of the causal variants have positive effects and half of the causal variants have negative effects. In all scenarios, similar patterns are shown in Fig. 1 and Supplementary Fig. S1. Our main findings included: (1) FHAT and FHAT-O are more powerful than SKAT-LTFH and SKATO-LTFH, respectively, under many scenarios when the variants have larger effects on the disease among older people; (2) FHAT and FHAT-O have greatly improved power compared to standard methods that do not incorporate FH in most scenarios except for the scenario when the proportion of causal variants is 10% and half of the causal variants have positive effects and half of the causal variants have negative effects. However, ACAT-V has substantial power loss in many other scenarios; (3) FHAT suffers



**Fig. 1 Empirical power of FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, Burden test, and ACAT-V at exome-wide significance for 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  $\alpha = 2.5 \times 10^{-6}$ . The total sample size of probands was set to 5000. The analyses were restricted to rare variants with  $MAF < 1\%$ . FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, SKAT, SKAT-O, and Burden test all used the same Wu weights with beta ( $MAF_j; 1, 25$ ). ACAT-V used the weights of  $w_{j,ACAT-V} = w_{j,SKAT} \times \sqrt{MAF_j(1 - MAF_j)}$  to make results comparable. FHAT, FHAT-O, SKAT-LTFH, and SKATO-LTFH analyzed probands and incorporated the family history information, while SKAT, SKAT-O, Burden test, and ACAT-V only included probands. The proportion of causal variants was set to 10%, 20%, 50%, 80%, and 100%. The numbers of variants tested in a region considered were: 20, 40, 80.

from a loss of power when the proportion of causal variants is high and the causal variants have effects in the same directions. In contrast, FHAT-O outperforms FHAT in those scenarios, and remains powerful regardless of the directions of genetics effects or number of causal variants.

**Computational cost**

FHAT and FHAT-O and other existing methods (SKAT, SKAT-O, Burden test, and ACAT-V) have lower computational cost compared to SKAT-LTFH and SKATO-LTFH. Table 2 summarizes computation time (in minutes) for all methods for analyzing 1000 regions that contain 30 variants. The computation time of FHAT, FHAT-O, SKAT, SKAT-O, Burden test, and ACAT-V depends on sample size and region size, whereas the running time for SKAT-LTFH and SKATO-LTFH (conducting using the LT-FH software v2

[11]) depends on the number of configurations of probands' disease status and FH.

**Application to the UK Biobank**

We restricted the analysis to 129,670 white individuals who passed all filters and have exome sequencing data, phenotype, and available parental disease status (see details in the Supplementary Method). The age at the first assessment visit for probands is between 38 and 72 with the mothers of probands being between 60 and 105, and the fathers of probands being between 60 and 102. There are 27 dementia cases ( $p = 0.02\%$ ) and 32,773 hypertension cases ( $p = 25.3\%$ ) among probands. While mothers and fathers of probands have similar hypertension prevalence (37,145 hypertension cases in mothers,  $p = 28.6\%$ ; 26,063 hypertension cases in fathers,  $p = 20.1\%$ ), more dementia cases

**Table 2.** Computational time for testing 1000 regions.

Sample size	FHAT	FHAT-O	SKAT	SKAT-O	Burden	ACAT-V	SKAT-LTFH	SKATO-LTFH
200	0.09	0.25	0.12	1.38	0.13	0.04	540.33	544.83
500	0.14	0.33	0.14	1.57	0.16	0.06	536.42	543.09
1000	0.26	0.43	0.24	1.69	0.23	0.09	534.54	541.84
2000	0.53	1.25	0.42	2.56	0.39	0.14	566.45	568.20
5000	1.19	1.89	0.90	5.40	0.81	0.29	551.78	553.74

Each cell summarizes the time (in minutes) that is required to performing the tests on 1000 regions using the methods of FHAT, SKAT-LTFH, SKAT, FHAT-O, SKAT-LTFH, SKAT-O, Burden, and ACAT-V. The regions contain 30 variants.

are observed in the parents (10,654 dementia cases in mothers,  $p = 8.2\%$ ; 5720 dementia cases in fathers,  $p = 4.4\%$ ) compared to probands.

We first evaluated the associations between all cause dementia and hypertension with known regions previously implicated with AD/dementia risk [15] and hypertension [24–27]. We performed the analysis for all unrelated white individuals using FHAT, FHAT-O, SKAT-LTFH, SKATO-LTFH, and other conventional tests (SKAT, SKAT-O, Burden test, and ACAT-V), see results in Table 3. The samples involved in the analyses varied because of missing values in the covariates used for adjustment in the models. FHAT, SKAT-LTFH, FHAT-O, and SKATO-LTFH had improved significance after incorporating parental phenotype information compared to  $p$  values calculated using other conventional tests for majority of genes. SKAT, SKAT-O, and ACAT-V had almost no power to detect some associations for all cause dementia due to low prevalence in probands. The results show that *BCL3* ( $p = 6.8 \times 10^{-5}$  in FHAT,  $p = 2.5 \times 10^{-5}$  in SKAT-LTFH,  $p = 5.9 \times 10^{-5}$  in FHAT-O,  $p = 1.8 \times 10^{-5}$  in SKATO-LTFH) and *TOMM4* ( $p = 3.0 \times 10^{-4}$  in FHAT,  $p = 5.8 \times 10^{-4}$  in SKAT-LTFH,  $p = 3.8 \times 10^{-4}$  in FHAT-O,  $p = 7.7 \times 10^{-4}$  in SKATO-LTFH) were significantly associated with all cause dementia status at a significance level of  $6.3 \times 10^{-3}$  for testing eight genes. At the same significance level, *DBH* ( $p = 1.3 \times 10^{-3}$  in FHAT,  $p = 2.0 \times 10^{-3}$  in SKAT-LTFH,  $p = 2.6 \times 10^{-3}$  in FHAT-O,  $p = 3.3 \times 10^{-3}$  in SKATO-LTFH) was identified for hypertension and which had improved significance compared to the results from conventional methods. Although the tests that incorporate FH demonstrated an improved significance for all eight AD/dementia genes we tested, some  $p$  values for hypertension genes were less significant. This may be due to the fact that the prevalence for hypertension in probands was similar to that in parents, and the associations were diluted by the potential noises that were added when combining the FH from parents.

A comprehensive exome-wide analysis was then conducted. A total of ~18K genes with two or more rare genetic variants meeting our filtering criteria were included. We used models including the same covariates for all cause dementia and hypertension as we did in the known gene analyses. We used  $p < 5.6 \times 10^{-5}$  as the suggestive significance threshold for testing ~18K genes. In the analysis of all cause dementia (Table 4 and Fig. 2), the gene *TREM2* [28] ( $p = 4.1 \times 10^{-9}$ ) with known effects on AD/dementia and late-onset AD achieved a strict exome-wide significance ( $p < 2.8 \times 10^{-6}$ ) using FHAT-O and it was also detected by FHAT ( $p = 5.2 \times 10^{-6}$ ) with a suggestive exome-wide significance. One known AD/dementia gene, *PVR* [29] ( $p = 1.2 \times 10^{-5}$  in FHAT and  $p = 1.8 \times 10^{-5}$  in FHAT-O) was identified with both FHAT and FHAT-O analysis, and *ABCA7* [30] ( $p = 4.1 \times 10^{-5}$ ) with known effects on AD/dementia was identified by FHAT-O. Moreover, three novel genes were found to be significantly associated with all cause dementia using FHAT and FHAT-O (*EFCAB3* with  $p = 4.0 \times 10^{-5}$  in FHAT and  $p = 4.2 \times 10^{-5}$  in FHAT-O, *EMSY* with  $p = 4.4 \times 10^{-5}$  in FHAT and  $p = 2.7 \times 10^{-5}$  in FHAT-O, and *KLC3* with  $p = 1.4 \times 10^{-5}$  in FHAT-O). Because we observed highly inflated results (Fig. 2) from hypertension analysis due to the correlation

among parents' phenotypes, we corrected the analysis by additionally adjusting for the spouse's hypertension status in the parents' model. For the adjusted hypertension analysis (Table 4 and Fig. 2), FHAT identified *GATA5* ( $p = 4.1 \times 10^{-5}$ ), and FHAT-O identified *FGD5* ( $p = 4.3 \times 10^{-5}$ ) and *DDN* ( $p = 4.2 \times 10^{-5}$ ) at a suggestive significance level. Those genes detected by our methods have previously been reported to be associated with hypertension-related trait [31–33].

## DISCUSSION

We proposed two novel approaches, FHAT and FHAT-O, that incorporate FH to increase power to detect rare variant associations in aggregation unit-based analysis. We also offered a novel way to adapt the LT-FH method to analyze rare variants. Because FH of disease is often collected through questionnaires in large cohorts, the added power is at no added cost. We applied our methods to exploit the FH from parents in simulation analysis and using the UK Biobank data, by assuming that the parents are conditionally independent. We analyzed both parents through a single relatives' model, and combined the scores calculated from parents and probands with appropriate weights to calculate the test statistics. Because the probands' analysis is separate from the relatives' analysis, our methods can handle the missingness in FH as presented in (1) and (4), and one can include all probands with or without FH to optimize the usage of data.

The power was evaluated at  $\alpha = 2.5 \times 10^{-6}$  to represent the exome-wide significance for testing 20,000 genes as well as at a suggestive threshold of  $\alpha = 2.5 \times 10^{-5}$ . By assuming that the causal variants in older people have bigger effects compared to younger people, we showed that FHAT and FHAT-O have slightly greater power than SKAT-LTFH, SKATO-LTFH, with greatly reduced computational cost. Compared with SKAT and ACAT-V, FHAT has greater gain in power in most cases. However, FHAT and SKAT are less powerful than Burden test and SKAT-O when there is a high proportion of causal variants, especially when the causal variants all have the positive effects. FHAT-O combines the features of both FHAT and FHAT-Burden, has robust power in many scenarios, and outperforms other methods, as shown in our extensive power simulations. ACAT-V has slightly higher power in some cases where the proportion of causal variants is low, which was expected because only a few genetic variants contribute to the results in ACAT-V, though the score statistic for FHAT and FHAT-O is calculated using a linear combination of squared scores from both causal and non-causal variants. We further demonstrated that our methods have improved significance after incorporating FH from association analyses with all cause dementia and hypertension using genotypes and phenotypes collected from the UK Biobank. We compared results using FHAT, FHAT-O, SKAT-LTFH, and SKATO-LTFH for probands with both genotypes and phenotypes, and their parental history of disease to other methods only using probands. Variants in eight known AD/dementia regions and eight known hypertension regions were selected for the analysis. Using the significance level =  $6.3 \times 10^{-3}$  for testing eight known genes, *BCL3* and *TOMM40* were significantly

**Table 3.** Association analysis for genes previously implicated in all cause dementia and hypertension susceptibility.

Gene	FHAT	SKAT-LTFH	SKAT	FHAT-O	SKATO-LTFH	SKAT-O	Burden	ACAT-V	#variants	cMAC	cMAC in cases
All cause dementia (N = 129,670)											
BCL3	$6.8 \times 10^{-5}$	$2.5 \times 10^{-5}$	0.02	$5.9 \times 10^{-5}$	$1.8 \times 10^{-5}$	0.029	0.11	0.36	65	1157	1
TOMM40	$3.0 \times 10^{-4}$	$5.8 \times 10^{-4}$	1	$3.8 \times 10^{-4}$	$7.7 \times 10^{-4}$	0.85	0.68	0.05	39	809	0
APOE	0.02	0.02	1	0.03	0.04	0.83	0.60	0.06	75	1372	0
PIRRA	0.17	0.15	0.93	0.27	0.25	1.0	0.99	0.88	48	4406	1
BIN1	0.61	0.70	1	0.77	0.88	0.89	0.64	0.75	80	975	0
CR1	0.33	0.26	1	0.51	0.42	0.38	0.21	0.03	310	8500	0
CLU	0.44	0.43	1	0.59	0.59	0.66	0.45	0.79	75	2651	0
MAF	0.50	0.44	1	0.60	0.43	0.88	0.65	0.43	64	930	0
Hypertension (N = 129,206)											
DBH	$1.3 \times 10^{-3}$	$2.0 \times 10^{-3}$	$3.8 \times 10^{-3}$	$2.6 \times 10^{-3}$	$3.3 \times 10^{-3}$	$1.6 \times 10^{-3}$	$3.0 \times 10^{-3}$	0.05	166	7708	1851
SVEP1	0.051	0.09	0.066	0.068	0.10	0.091	0.082	0.36	485	19,123	4707
NR1	0.069	0.06	0.042	0.026	0.03	$6.9 \times 10^{-3}$	$4.0 \times 10^{-3}$	0.06	147	2739	749
REN	0.069	0.12	0.22	0.13	0.22	0.38	0.83	0.45	68	586	145
NPPA	0.21	0.27	0.40	0.28	0.40	0.48	0.33	0.03	31	1200	309
CHDH	0.24	0.25	0.54	0.29	0.33	0.33	0.20	0.75	120	2346	556
NFI	0.28	0.43	0.27	0.44	0.63	0.43	0.51	0.43	325	7826	1944
AGPS	0.36	0.33	0.38	0.54	0.50	0.56	0.98	0.88	172	5748	1449
PABPC4	0.93	0.98	0.91	0.42	0.49	0.11	0.060	0.79	60	283	61

The all cause dementia model adjusted age, sex, and PC11 as covariates. The hypertension model adjusted age, age squared, sex, BMI, PC1-5, PC8, and PC14 as the covariates. Wu weights with beta (MAF<sub>i</sub>, 1, 25), were used. The p values were estimated using Davies' method. The significance threshold is  $\frac{0.05}{8} = 6.3 \times 10^{-3}$ . N is the number of the total samples involved in the analysis. cMAC is the cumulative minor allele counts for the gene we tested. #variants is the total number of variants tested in the gene.

**Table 4.** Whole exome-wide association analysis for all cause dementia and hypertension.

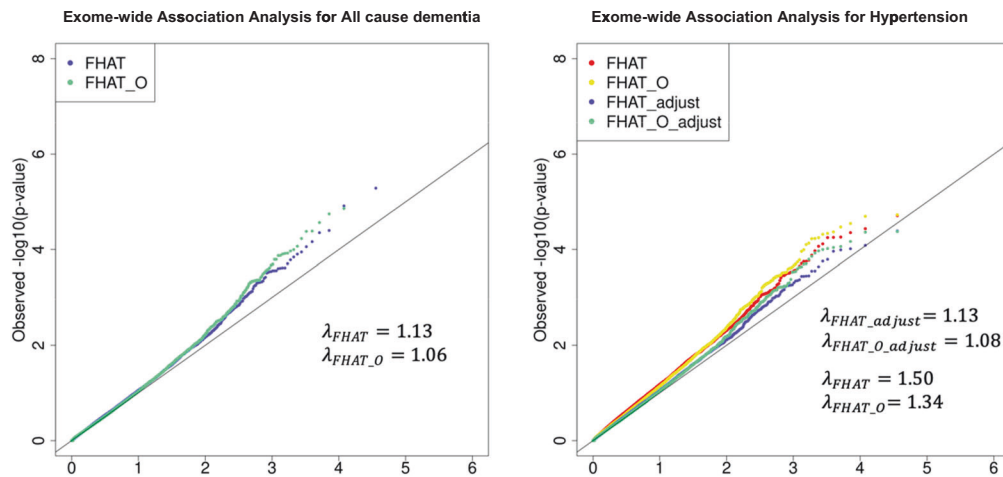
Gene name	FHAT p value	FHAT-O p value	#variants	cumMAC
All cause dementia (N = 129,670)				
TREM2	$5.2 \times 10^{-6}$	$4.1 \times 10^{-9}$	45	4559
PVR	$1.2 \times 10^{-5}$	$1.8 \times 10^{-5}$	75	2068
EFCAB3	$4.0 \times 10^{-5}$	$4.2 \times 10^{-5}$	60	2579
EMSY	$4.4 \times 10^{-5}$	$2.7 \times 10^{-5}$	158	1543
KLC3	$4.8 \times 10^{-4}$	$1.4 \times 10^{-5}$	177	4174
ABCA7	$2.9 \times 10^{-3}$	$4.1 \times 10^{-5}$	487	12,179
Hypertension (N = 129,206)				
GATA5	$4.1 \times 10^{-5}$	$9.1 \times 10^{-5}$	88	5402
FGD5	$2.3 \times 10^{-4}$	$4.3 \times 10^{-5}$	254	5269
DDN	0.016	$4.2 \times 10^{-5}$	107	1621

The exome-wide significance threshold is  $\frac{0.05}{18,000} = 2.8 \times 10^{-6}$  the suggestive exome-wide significance threshold is  $\frac{1}{18,000} = 5.6 \times 10^{-5}$ . cumMAC is the cumulative MAF in the region. #variants is the total number of variants in the gene. N is the number of the total samples involved in the analysis.

associated with all cause dementia, while other known AD/dementia regions had improved significance compared to the methods that do not incorporate FH. Some of the hypertension genes were less significant using our method to incorporate FH, which might be caused by additional noise resulting from a similar hypertension prevalence in probands and their parents. The FHAT and FHAT-O approaches yielded similar conclusions compared to SKAT-LTFH, and SKATO-LTFH, respectively.

We evaluated type I error at various alpha levels and disease prevalence. We did not evaluate the type I error for SKAT-LTFH and SKATO-LTFH at the exome-wide significance ( $\alpha = 2.5 \times 10^{-6}$ ) to limit the computational cost. The type I error of SKAT was previously found to be conservative when the disease prevalence is ~50%, and the Burden test was found to have appropriate type I error when the case-control ratio is balanced [5–7]. However, SKAT, SKAT-O, Burden, and ACAT-V suffer from substantial inflated type I error when the prevalence is low, especially for lower alpha level (i.e.,  $\alpha < 2.5 \times 10^{-4}$ ). In contrast, the FHAT, SKAT-LTFH, FHAT-O, and SKATO-LTFH control the type I error rates relatively better. The type I error is overall well controlled using FHAT and FHAT-O in most scenarios, but a high inflation occurs for  $\alpha = 2.5 \times 10^{-6}$  and prevalence = 10% where the number of cases and controls is unbalanced (Table 1). Unbalanced case-control ratio yields inflated type I error rates because the imbalance invalidates the asymptotic assumption of logistic regression. Saddle point approximation [34–36] method and efficient resampling [37] have been successfully used to calibrate binary phenotype-based logistic mixed models when case-control ratios are extremely unbalanced. In the future, we plan to adopt these cutting-edge methods to properly account for unbalanced case-control ratio.

In the exome-wide association analysis, we used the same covariates (age, sex, PC1-5, PC11) as we did in the known region analysis for all cause dementia. However, as the inflation was observed in our hypertension analysis (Fig. 2), we further adjusted for the spouse's disease status in the parents' model to account for the correlations among parents in addition to the covariates of age, age squared, sex, BMI, PC1-PC5, PC8, and PC14. The FH could be correlated with household effects. In the future, we will extend the current approaches to allow for correlation, as might be induced by household effect, in the analysis. Through the exome-wide analysis using FHAT and FHAT-O, we confirmed previously reported genes (TREM2, PVR, and ABCA7) [28–30] for AD/dementia as well as genes (GATA5, FGD5, DDN) [31–33] related to blood



**Fig. 2** Q-Q plots of whole exome-wide analysis results for all cause dementia and hypertension. The  $p$  values for regions with cumulative minor allele counts  $>20$  were used to generate the Q-Q plots. The left panel is the whole exome-wide analysis results for all cause dementia, where FHAT and FHAT-O were calculated using the model with the same covariates (age, sex, PC1–5, PC11) adjusted in AD/dementia known gene analysis. The right panel is the whole exome-wide analysis results for hypertension, where FHAT and FHAT-O were calculated using the model with the same covariates (age, age squared, sex, body mass index (BMI), PC1-PC5, PC8, and PC14) adjusted in hypertension known gene analysis. FHAT\_adjust and FHAT-O\_adjust were calculated from the adjusted hypertension analysis, where the spouse's hypertension status combining with other previously mentioned covariates were adjusted in the parental analysis.

pressure and hypertension. Moreover, our methods identified three novel regions associated with all cause dementia (*EFCAB3*, *EMSY*, *KLC3*) using a suggestive exome-wide significance threshold. Replication analyses are needed to confirm these findings. While we observed inflated type I error for low prevalence in our simulations, we did not see evidence of large inflation of FHAT and FHAT-O in all cause dementia analysis, as seen from the Q-Q plot (Fig. 2) and genomic control inflation factor (with  $\lambda_{\text{FHAT}} = 1.13$  and  $\lambda_{\text{FHAT-O}} = 1.06$  for all cause dementia analysis). The methods require that all samples are unrelated. The generalized linear mixed model can be used to expand the current methods for related samples, which will allow us to incorporate FH from multiple relatives or handle the consanguineous families.

Although the method development, simulation studies, and UK Biobank analysis described in the paper were focusing on the population samples, our methods can also handle the ascertainment that happens in case-control analysis, because the likelihood can be written as the product of the retrospective proband information, taking ascertainment into consideration:  $P(G_i^P, Y_i^P | Y_i^R, X_i^P, X_i^R) = P(Y_i^R | G_i^P, Y_i^P, X_i^R) P(G_i^P | Y_i^P, X_i^P)$  (Supplementary Method). Equation (1) was derived based on the assumption of independence of the relatives' phenotype and probands' covariates conditional on the relatives' covariates and the strength of the associations in relatives. However, when the proband covariates are believed to have an effect on the relatives' disease status, one can adjust for such covariates in the relatives' model (3) to account for such an effect. There might be a concern about the accuracy about the FH collected from the probands. The reporting bias in FH among relatives would lead to misclassification of relatives' disease status, which might cause biases in effect size estimates. However, the methods we proposed are variance component models that do not rely on effect estimates and only provide statistical significance ( $p$  values) for associations. The misclassification would affect the power of our methods, but would not affect the validity of the test (i.e., type I error) as the misclassification is not related to genotype data under the null hypothesis of no association between genotypes and disease status. We would expect minimal impact on the methods based on the previously published work, where FH was down-weighted based on the accuracy calculated as the correlation of FH recorded among siblings, resulting only in small changes to the association results [11].

In this paper, we demonstrated that FHAT and FHAT-O are computationally efficient compared to SKAT-LTFH and SKATO-LTFH. The significantly reduced computational cost using FHAT and FHAT-O was shown in the analysis time to run 1000 aggregation unit-based tests. Although we focused on binary traits and rare variants, our method can be applied to analyze continuous traits using linear models and common variants. The framework in FHAT is flexible for various settings. While we applied FHAT and FHAT-O for probands with parental disease status available in simulations and the UK Biobank analysis, FHAT can be easily applied to other relative types. We also proposed an extension to FHAT, FHAT-O, to capture the features in SKAT-O, in particular the robustness of the power when all genetic variants have the same direction of effects and the proportion of causal variants is high. The framework can easily be extended to incorporate any other established aggregation unit-based methods. Our methods that allow the incorporation of available FH are innovations compared to traditional rare variant studies that only use cases and controls, which have great potentials to promote genetic association research.

#### DATA AVAILABILITY

The datasets generated during and/or analyzed during the current study are available in the UK Biobank repository (<http://www.ukbiobank.ac.uk/>).

#### CODE AVAILABILITY

The functions for FHAT and FHAT-O are available on the website <http://sites.bu.edu/fhspl/publications/fhat/>.

#### REFERENCES

1. Bodmer W, Bonilla C. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet.* 2008;40:695–701.
2. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev.* 2009;19:212–9.
3. Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, et al. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet.* 2010;11:446–50.
4. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet.* 2011;89:82–93.

5. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or mono-allelic risk for common diseases: a cohort allelic sums test (CAST). *Mutat Res.* 2007;615:28–56.
6. Li B, Leal SM. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet.* 2008;83:311–21.
7. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet.* 2009;5:e1000384.
8. Lee S, Emond MJ, Bamshad MJ, Barnes KC, Rieder MJ, Nickerson DA, et al. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet.* 2012;91:224–37.
9. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies. *Am J Hum Genet.* 2019;104:410–21.
10. So H-C, Kwan JSH, Cherny SS, Sham PC. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. *Am J Hum Genet.* 2011;88:548–65.
11. Hujoel MLA, Gazal S, Loh PR, Patterson N, Price AL. Liability threshold modeling of case-control status and family history of disease increases association power. *Nat Genet.* 2020;52:541–7.
12. Wacholder S, Hartge P, Struwing JP, Pee D, McAdams M, Brody L, et al. The kin-cohort study for estimating penetrance. *Am J Epidemiol.* 1998;148:623–30.
13. Ghosh A, Hartge P, Kraft P, Joshi AD, Ziegler RG, Barrdahl M, et al. Leveraging family history in population-based case-control association studies. *Genet Epidemiol.* 2014;38:114–22.
14. Liu JZ, Erlich Y, Pickrell JK. Case-control association mapping by proxy using family history of disease. *Nat Genet.* 2017;49:325–31.
15. Marioni RE, Harris SE, Zhang Q, McRae AF, Hagenaars SP, Hill WD, et al. GWAS on family history of Alzheimer's disease. *Transl Psychiatry.* 2018;8:99.
16. Gim J, Kim W, Kwak SH, Choi H, Park C, Park KS, et al. Improving disease prediction by incorporating family disease history in risk prediction models with large-scale genetic data. *Genetics.* 2017;207:1147–55.
17. Shi M, Umbach DM, Weinberg CR. Using parental phenotypes in case-parent studies. *Front Genet.* 2015;6:221.
18. Thornton T, McPeck MS. Case-control association testing with related individuals: a more powerful quasi-likelihood score test. *Am J Hum Genet.* 2007;81:321–37.
19. Lee S, Teslovich TM, Boehnke M, Lin X. General framework for meta-analysis of rare variants in sequencing association studies. *Am J Hum Genet.* 2013;91:42–53.
20. Davies RB. The distribution of a linear combination of chi-square random variables. *J R Stat Soc.* 1980;29:323–33.
21. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics.* 2012;13:762–75.
22. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics.* 2011;27:2304–5.
23. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol.* 2013;37:196–204.
24. Mavani G, Kesar V, Devita MV, Rosenstock JL, Michelis MF, Schwimmer JA. Neurofibromatosis type 1-associated hypertension secondary to coarctation of the thoracic aorta. *Clin Kidney J.* 2014;7:394–5.
25. Sun B, Williams JS, Pojoga L, Chamarthi B, Lasky-Su J, Rabuy BA, et al. Renin gene polymorphism: its relationship to hypertension, renin levels and vascular responses. *J Renin Angiotensin Aldosterone Syst.* 2011;12:564–71.
26. Liu C, Kraja AT, Smith JA, Brody JA, Franceschini N, Bis JC, et al. Meta-analysis identifies common and rare variants influencing blood pressure and overlapping with metabolic trait loci. *Nat Genet.* 2016;48:1162–70.
27. Surendran P, Feofanova EV, Lahrouchi N, Natlla L, Karthikeyan S, Cook J, et al. Discovery of rare variants associated with blood pressure regulation through meta-analysis of 1.3 million individuals. *Nat Genet.* 2020;52:1314–32.
28. Gratzke M, Leyns CEG, Holtzman DM. New insights into the role of TREM2 in Alzheimer's disease. *Mol Neurodegener.* 2018;13:66.
29. Jansen IE, Savage JE, Watanabe K, Bryois J, Williams DM, Steinberg S, et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat Genet.* 2019;51:404–13.
30. Hollingworth P, Harold D, Sim R, Gerrish A, Lambert J-C, Carrasquillo MM, et al. Common variants at ABCA7, MS4A6A/MS4A4E, EPHA1, CD33 and CD2AP are associated with Alzheimer's disease. *Nat Genet.* 2011;43:429–35.
31. Messaoudi S, He Y, Gutsol A, Wight A, Hébert RL, Vilmundarson RO, et al. Endothelial Gata5 transcription factor regulates blood pressure. *Nat Commun.* 2015;6:8835.
32. Ehret GB, Ferreira T, Chasman DI, Jackson AU, Schmidt EM, Johnson T, et al. The genetics of blood pressure regulation and its target organs from association studies in 342,415 individuals. *Nat Genet.* 2016;48:1171–84.
33. Giri A, Hellwege JN, Keaton JM, Park J, Qiu C, Warren HR, et al. Trans-ethnic association study of blood pressure determinants in over 750,000 individuals. *Nat Genet.* 2019;51:51–62.
34. Dey R, Schmidt EM, Abecasis GR, Lee S. A fast and accurate algorithm to test for binary phenotypes and its application to PheWAS. *Am J Hum Genet.* 2017;101:37–49.
35. Kuonen D. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika.* 1999;86:929–35.
36. Daniels HE. Saddlepoint approximations in statistics. *Ann Math Stat.* 1954;25:631–50.
37. Lee S, Fuchsberger C, Kim S, Scott L. An efficient resampling method for calibrating single and gene-based rare variant association analysis in case-control studies. *Biostatistics.* 2016;17:1–15.

## ACKNOWLEDGEMENTS

YW, GMP, ALD, and JD acknowledge the grant support (No. 5U01AG058589) from the National Institute of Aging. HC acknowledges the grant support (No. R00 HL130593) from the National Institute of Health. This research was conducted using the UK Biobank Resource under Application Number 42614.

## AUTHOR CONTRIBUTIONS

YW developed the method, analyzed the data, and drafted the manuscript; HC provided critical input in the method development and revised the manuscript for intellectual content; GMP conceived the study, revised the manuscript for intellectual content; ALD revised the manuscript for intellectual content; JD supervised the work, conceived the study, revised the manuscript for intellectual content.

## COMPETING INTERESTS

The authors declare no competing interests.

## ETHICAL APPROVAL

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

## ADDITIONAL INFORMATION

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41431-021-00980-0>.

**Correspondence** and requests for materials should be addressed to Yanbing Wang or José Dupuis.

**Reprints and permission information** is available at <http://www.nature.com/reprints>

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.