


RESEARCH ARTICLE

Open Access

Genetic association tests in family samples for multi-category phenotypes



Shuai Wang^{1*} , James B. Meigs^{2,3,4} and Josée Dupuis⁵

Abstract

Background: Advancements in statistical methods and sequencing technology have led to numerous novel discoveries in human genetics in the past two decades. Among phenotypes of interest, most attention has been given to studying genetic associations with continuous or binary traits. Efficient statistical methods have been proposed and are available for both types of traits under different study designs. However, for multinomial categorical traits in related samples, there is a lack of efficient statistical methods and software.

Results: We propose an efficient score test to analyze a multinomial trait in family samples, in the context of genome-wide association/sequencing studies. An alternative Wald statistic is also proposed. We also extend the methodology to be applicable to ordinal traits. We performed extensive simulation studies to evaluate the type-I error of the score test, Wald test compared to the multinomial logistic regression for unrelated samples, under different allele frequency and study designs. We also evaluate the power of these methods. Results show that both the score and Wald tests have a well-controlled type-I error rate, but the multinomial logistic regression has an inflated type-I error rate when applied to family samples. We illustrated the application of the score test with an application to the Framingham Heart Study to uncover genetic variants associated with diabetes, a multi-category phenotype.

Conclusion: Both proposed tests have correct type-I error rate and similar power. However, because the Wald statistics rely on computer-intensive estimation, it is less efficient than the score test in terms of applications to large-scale genetic association studies. We provide computer implementation for both multinomial and ordinal traits.

Keywords: EGEE, Score test, Wald test, Framingham heart study, Family samples, Categorical, Multinomial, Ordinal, GWAS, Sequencing

Background

Genetic association tests for continuous or binary phenotypes have uncovered many susceptibility genes or variants related to diseases. Various methods and efficient software have been developed and used for continuous and binary traits. For family samples, due to the correlation between relatives and violation of the independence assumption of ordinary linear regression, some alternative approaches were proposed. For example, Therneau and colleagues developed an R package (coxme) implementing linear mixed effects model to

evaluate the association between a genetic variant and a continuous trait or survival outcome accounting for correlation present in family samples. Similar extensions to account for familial correlation using mixed effects models have been proposed for gene-based association tests [1]. The progress in family sample designs has been restricted mostly to quantitative traits or binary traits. However, methods are needed to study categorical traits with more than two categories in family samples. For example, the phenotype diabetes has been defined as a four-category (diabetes & obesity, diabetes but no obesity, obesity but no diabetes and no diabetes & no obesity) variable constructed jointly from type 2 diabetes and obesity. Currently, approaches for genetic association

* Correspondence: shuai1107@hotmail.com

¹Pfizer Inc, Global Product Development, Groton, CT 06340, USA
Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

analysis of such multinomial traits are limited. Zhang and colleagues [2] proposed a proportional odds logistic model which allows for the inclusion of covariates. However, it has a few limitations. First, this approach is restricted to nuclear families and cannot handle complex family structures. Second, no software implementation has been made publicly available. Diao and Lin [3] proposed a general framework for linkage and association tests for ordinal traits. Their method utilized adaptive Gaussian quadrature to approximate the maximum log-likelihood and a likelihood ratio test was proposed to test the hypothesis of no association between a genetic variant and an ordinal trait of interest. Again, this approach also has not been widely used due to lack of computer-efficient software and the fact that the likelihood ratio test is computationally intensive. Another possible option is to use the SAS generalized linear mixed models (GLMM) procedure, which can incorporate a kinship matrix. However, in real applications, the current implementation of the GLMM cannot handle extended families due to the computational burden. More recently, Wang and colleagues [4] proposed a Bayesian framework incorporating kinship matrix as a random effect, which however can not be applied to large-scale genetic study because of lack of computational efficiency. Bi and colleagues [5] proposed a computer-efficient framework (POLMM), specifically for ordinal traits. Because it doesn't allow for a user-provided kinship matrix, such as the one estimated from pedigree or using a typical genetic software, this will be a limitation for family-based cohort studies with known relationships. Our proposed method is complementary to these two approaches as it can be applied to family samples without available genome-wide data to compute a GRM, and without the proportional odds assumption. In this paper, we propose a computationally efficient score test based on extended generalized estimating equations (EGEE) for large-scale genetics studies of multi-category phenotypes accounting for familial correlation. We evaluate our approach using simulations and apply it to a genome-wide scan to identify genetic variants associated with diabetes, a four-category phenotype, with the healthy referent category being no diabetes and no obesity and the unhealthiest category, "diabetes" (diabetes and obesity), having a prevalence of at least 25% in several countries [6].

Results

Type-I error

The results of family-based and unrelated samples are summarized in Table 1–2 respectively. Both the score and Wald tests have well-controlled type-I error rates across all MAF scenarios except for rare variants. This conclusion applies to both family-based and unrelated

designs. The multinomial logistic regression, which ignores familial correlation, returns an inflated type-I error rate in the presence of related individuals, although its type-I error rate for unrelated study design is well-controlled. In the application to ordinal trait (Table 3), robust score test preserves the type-I error in all MAF scenarios although the simulated phenotype distribution is highly unbalanced. The Wald test is only very slightly inflated for very rare variants when evaluated at 0.0001. We have also generated QQ-plots (Additional File 3) for the robust score test and the simplified score test for results from all MAF scenarios for both multinomial and ordinal traits when applied to family-based samples. The QQ-plots are consistent with the empirical type-I error summarized in the tables below.

Power evaluation

The results of family-based and unrelated samples are summarized in Tables 4 and 5, respectively. Because we have concluded that multinomial logistic regression leads to inflated type-I error rates, the power rate of multinomial logistic regression is not evaluated for family-based samples (Table 4). The score and Wald tests have approximately the same power rate for each scenario (MAF, study design). The logistic regression using LRT has approximately the same power as the other two approaches in unrelated samples.

Data analysis

Low-frequency (MAF < 0.01) and poorly imputed variants (imputation ratio < 0.3) have been excluded to avoid spurious results. All results are presented in the Manhattan plot (Fig. 1., and Manhattan plots for diabetes, obesity in Additional File 3) and QQ-plot (Fig. 2.). The variants that have reached a genome-wide significance threshold of 5×10^{-8} or a suggestive threshold of 4×10^{-7} (calculated as $1/\text{number of tests} = 1/2542166$) are summarized in Table 6. All variants in Table 6 are located within the *CYP3A43*, *AP3B1* and *LOC105370246* genes. *AP3B1* is known to have variants associated with fasting insulin and HOMA-IR in African Americans without diabetes [7]. The direct association between *LOC105370246* and diabetes or obesity is not known in literature. *CYP3A43* gene encodes a member of the cytochrome P450 superfamily of liver enzymes. Although the direct relationship between *CYP3A43* and diabetes/obesity was not well known, some variants located in *CYP3A4* have been identified in previous studies to be associated with relevant metabolism traits. For instance, one study in 2011 [8] indicated diabetes is associated with a significant decrease in hepatic *CYP3A4* enzymatic activity and protein level. Several studies have demonstrated nonalcoholic fatty liver disease and diabetes are associated with decreased expression of the

Table 1 Simulation results of type-I error for family-based samples

MAF	Robust Score test			Wald test			Logistic regression (LRT)		
	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$
0.01	0.014	0.0020	0.0003	0.012	0.0024	0.00058	0.023	0.0023	0.0006
0.02	0.013	0.0020	0.0003	0.010	0.0011	0.0003	0.021	0.0027	0.0004
0.03	0.012	0.0017	0.0002	0.009	0.0012	0.0002	0.022	0.0025	0.0004
0.04	0.011	0.0014	0.0002	0.007	0.0010	0.0002	0.022	0.0025	0.0006
0.05	0.011	0.0013	0.0002	0.011	0.0008	0.0002	0.021	0.0026	0.0002
0.1	0.011	0.0010	0.0001	0.009	0.0008	0.0001	0.021	0.0024	0.0003
0.2	0.010	0.0010	0.0001	0.010	0.0010	0.0001	0.019	0.0033	0.0004
0.3	0.010	0.0010	0.0001	0.011	0.0013	0.0001	0.021	0.0033	0.0011

protein encoded by this gene in human livers [9, 10]. Two variants on *CYP3A43* were identified to be associated with Ticagrelor levels in individuals with acute coronary syndromes treated with ticagrelor [11] and serum metabolite measurement [12] respectively. Because this gene might have clinical value for treating chronic metabolic diseases such as nonalcoholic fatty liver disease [13], future research efforts targeting this gene area are worthwhile. Additional information about this region might be discovered with targeted sequencing.

For target validation purpose, we have performed two additional GWAS of diabetes and obesity respectively using our approach (Manhattan plots in Additional File 3). The plots confirm that all signals are observed from the combined phenotype and not driven by a single binary trait (diabetes or obesity).

We apply our ordinal approach to the secondary outcome (“ordinal” diabetes) and compare to results obtained from POLMM, an approach for ordinal trait. We observe that the results are similar with small differences (Fig. 3. and 4). Compared to results obtained from the multinomial trait (Fig. 1.), the ordinal trait highlights one region near *DABI* on Chromosome 1 and one region near *LOC107986327* on Chromosome 4 of

potential interest in the search for genes associated with “ordinal” diabetes.

Discussion

The proposed score test offers advantages over the Wald test and the multinomial logistic regression in the following aspects. First, it is more computationally efficient, especially for large-scale genetic studies such as GWAS, or sequencing studies because the iterative Fisher’s scoring algorithm is only applied once under the null hypothesis while the iterative algorithm is implemented for each variant when computing the Wald test statistic. Therefore, for a large-scale genetic study, the Wald test will be less computationally efficient than the score test. We have summarized the computing time for the score and Wald tests in Table 7 for different sample sizes as implemented in R functions using a 3-category multinomial phenotype on a i7-8565u processor with 16GB RAM. Second, the simulation studies show that the type-I error of both the score and Wald tests is well controlled for most scenarios. In contrast, the multinomial logistic regression results in a very inflated type-I error rate for family-based design when the familial correlation is ignored, and therefore it is not recommended

Table 2 Simulation results of type-I error for unrelated samples

MAF	Score test		Wald test		Logistic regression (LRT)	
	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.01$	$\alpha = 0.001$
0.01	0.011	0.0010	0.008	0.0006	0.011	0.0011
0.02	0.010	0.0016	0.010	0.0016	0.012	0.0014
0.03	0.012	0.0012	0.011	0.0010	0.011	0.0014
0.04	0.011	0.0010	0.010	0.0010	0.011	0.0010
0.05	0.010	0.0010	0.006	0.0004	0.010	0.0005
0.1	0.010	0.0010	0.010	0.0008	0.010	0.0010
0.2	0.010	0.0010	0.009	0.0004	0.010	0.0010
0.3	0.009	0.0011	0.009	0.0006	0.010	0.0010

Table 3 Simulation results of type-I error for family-based samples for ordinal traits

MAF	Robust Score test			Wald test		
	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$	$\alpha = 0.01$	$\alpha = 0.001$	$\alpha = 0.0001$
0.01	0.010	0.0008	0.00009	0.012	0.0013	0.00019
0.02	0.009	0.0008	0.00008	0.011	0.0011	0.00013
0.03	0.010	0.0010	0.00009	0.011	0.0012	0.00012
0.04	0.009	0.0009	0.00008	0.010	0.0011	0.00011
0.05	0.010	0.0009	0.00010	0.011	0.0011	0.00014
0.1	0.010	0.0010	0.00009	0.010	0.0011	0.00009
0.2	0.010	0.0009	0.00009	0.010	0.0010	0.00012
0.3	0.009	0.0009	0.00010	0.010	0.0010	0.00012

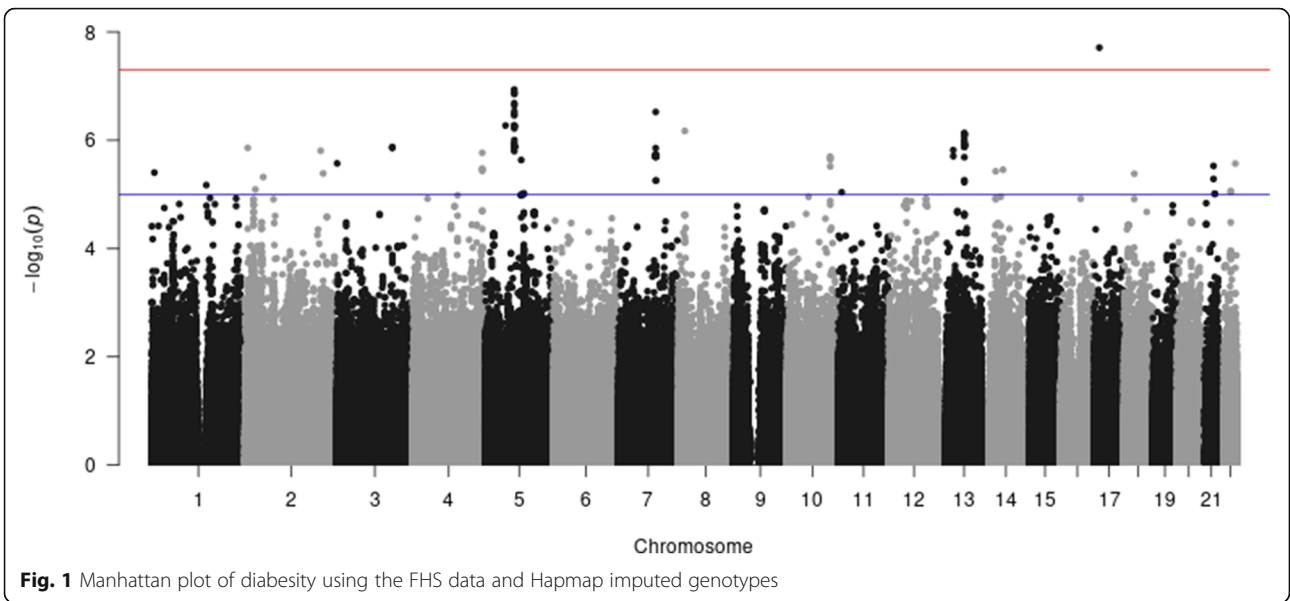
for family-based studies. When the phenotypes are extremely unbalanced, e.g. the allocation ratio of the 4 categories is approximately 2.5:1:10:23, both score and Wald tests can result in slightly inflated type-I error for rare variants in the simulation studies. This conclusion has been noted in most approaches [5]. However, when the phenotype distribution is more balanced, the tests return valid type-I error rates for all MAF scenarios, as demonstrated in the QQ-plot of the FHS data analysis (Fig. 2.). We have observed that the type-I error of ordinal traits is very robust to an unbalanced distribution of the phenotype for all MAF scenarios (Table 3), as indicated by the calculated type-I error rate obtained from 500,000 simulations by treating the simulated phenotype as an ordinal trait. Lastly, the score test has approximately the same power as the Wald test, under the scenarios we evaluated.

Table 4 Power results for family-based samples

MAF	$\alpha=0.01$	$\alpha=0.001$	$\alpha = 5 \times 10^{-8}$
0.01	score	97.2	42.5
	Wald	96.7	29.8
0.02	score	96.5	33.0
	Wald	96.6	24.4
0.03	score	95.5	25.6
	Wald	95.1	20.5
0.04	score	94.9	23.4
	Wald	94.6	17.7
0.05	score	94.3	20.9
	Wald	93.5	15.8
0.1	score	93.0	13.8
	Wald	94.3	11.6
0.2	score	89.4	7.6
	Wald	91.1	8.0
0.3	score	87.4	6.4
	Wald	89.0	6.5

Table 5 Power results for unrelated samples

MAF	$\alpha=0.01$	$\alpha=0.001$	$\alpha = 5 \times 10^{-8}$
0.01	score	95.0	26.5
	Wald	94.1	15.6
	Logistic (LRT)	92.9	11.4
0.02	score	93.3	20.0
	Wald	92.8	14.6
	Logistic (LRT)	91.6	10.6
0.03	score	92.7	15.9
	Wald	92.4	12.7
	Logistic (LRT)	91.2	9.5
0.04	score	92.4	14.1
	Wald	92.0	11.3
	Logistic (LRT)	90.8	8.6
0.05	score	92.0	13.1
	Wald	91.9	10.9
	Logistic (LRT)	90.9	8.2
0.1	score	91.3	10.4
	Wald	91.0	9.5
	Logistic (LRT)	90.3	7.8
0.2	score	89.9	8.1
	Wald	89.7	7.5
	Logistic (LRT)	89.3	6.8
0.3	score	89.2	7.0
	Wald	89.3	6.5
	Logistic (LRT)	89.0	6.3



It is worth noting that the EGEE are simply reduced to the score equations of generalized linear models for a multinomial variable when applied to unrelated samples. Because the same iteratively reweighted least square method is employed under this particular circumstance, the parameter estimates are identical to those obtained using a generalized linear model function for multinomial

variables. This equivalence enhances the applicability of this approach to a general population, regardless of the underlying study design.

The score test can be readily extended to ordinal traits (i.e. categorical traits for which the values are ordered.) in family samples. Due to the nature of the ordinal regression model, fewer regression parameters are estimated. Because

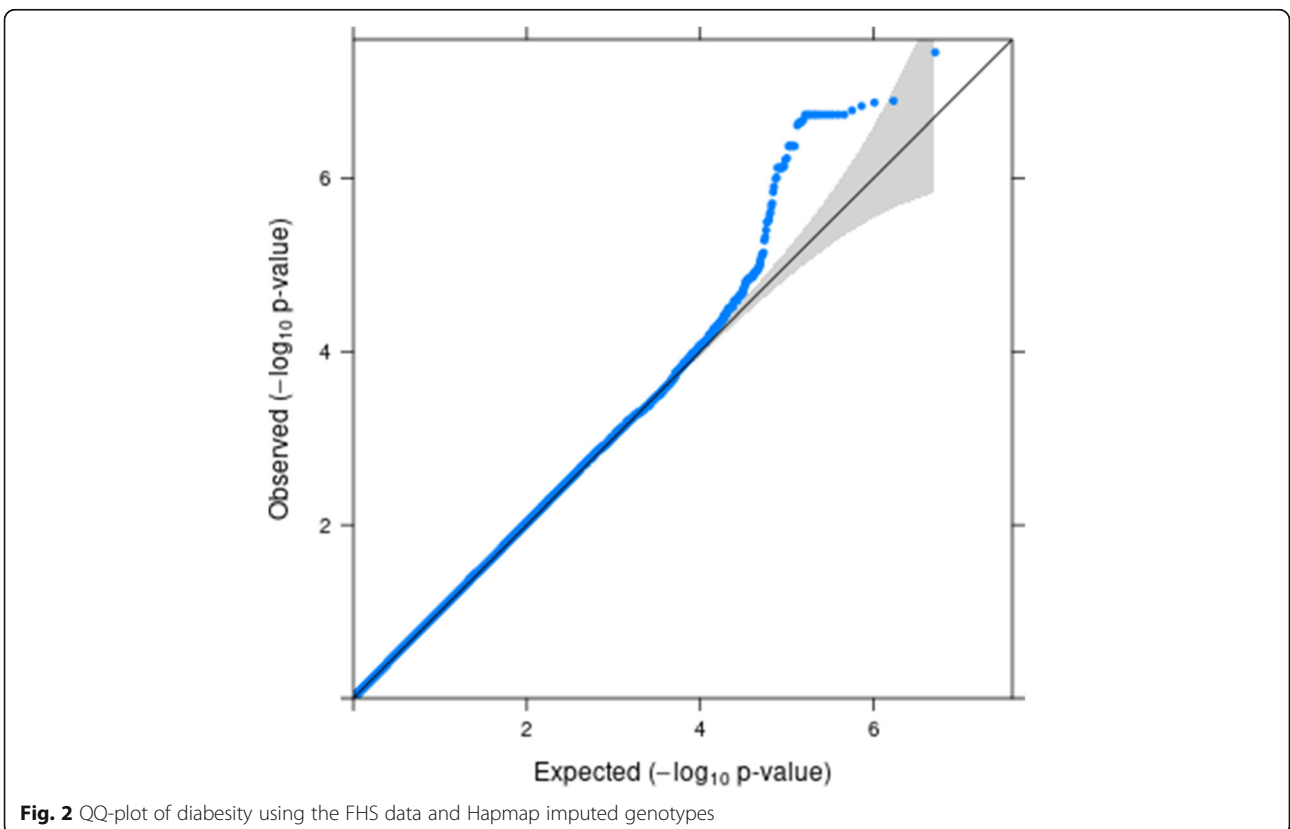


Table 6 Top SNPs and the closest genes

Chr	Lead SNP	p-value	bp (GRCh38)	Loci (closest gene)
5	rs16875172	1.17×10^{-7} to 5.34×10^{-7}	5:77422013–5:77559950	AP3B1
7	rs528144	2.99×10^{-7} to 5.56×10^{-6}	7:99257162–7:99918674	CYP3A43
13	rs1925751	7.34×10^{-7} to 5.97×10^{-6}	13:66763957–13: 66795551	LOC105370246

applications to ordinal traits are a special case of the general framework proposed with reduced complexity, the validity of simulation results should hold when applied to ordinal traits. When $K = 2$, i.e. an ordinal trait with only two categories, the estimates will be the same when using either multinomial or ordinal function, i.e. estimates of a binary logistic regression accounting for familial correlation.

Our proposed approaches have enabled the identification of a few loci associated with diabetes. As discussed, none of the signals were driven solely by one of the two binary traits (diabetes or obesity). Targeted sequencing might reveal more information, by providing a more comprehensive overview of rare and low-frequency variants in that specific regions. We also provide a comparison of our ordinal approach to POLMM for an ordinal trait and found that both approaches have revealed similar regions of association.

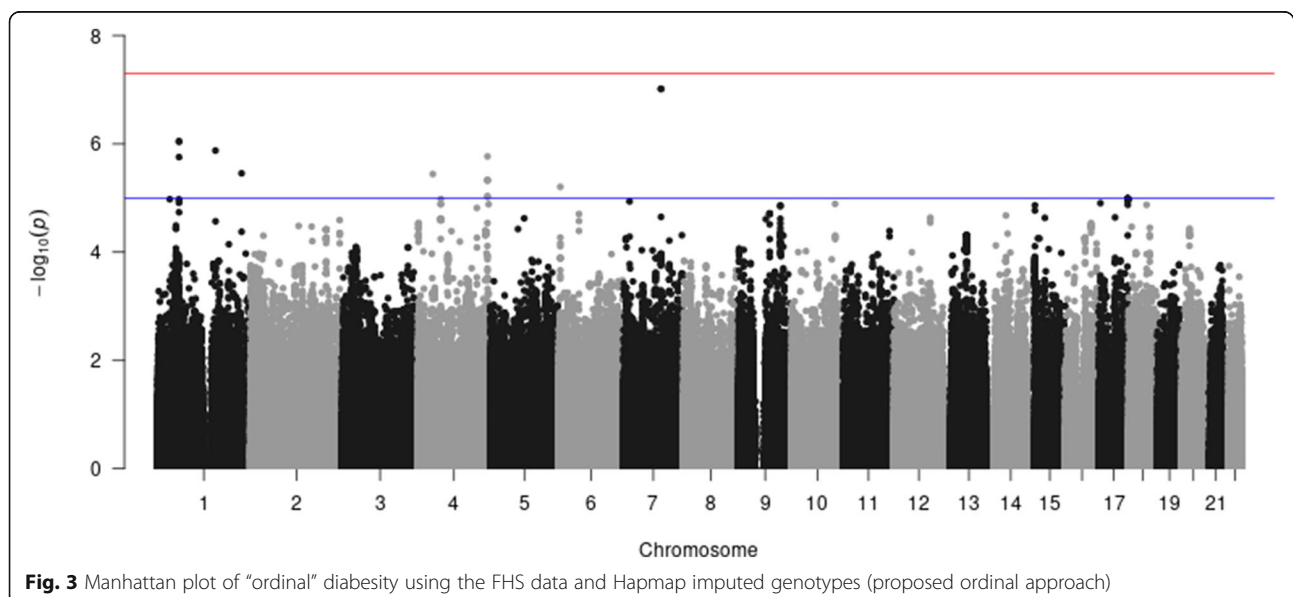
Conclusions

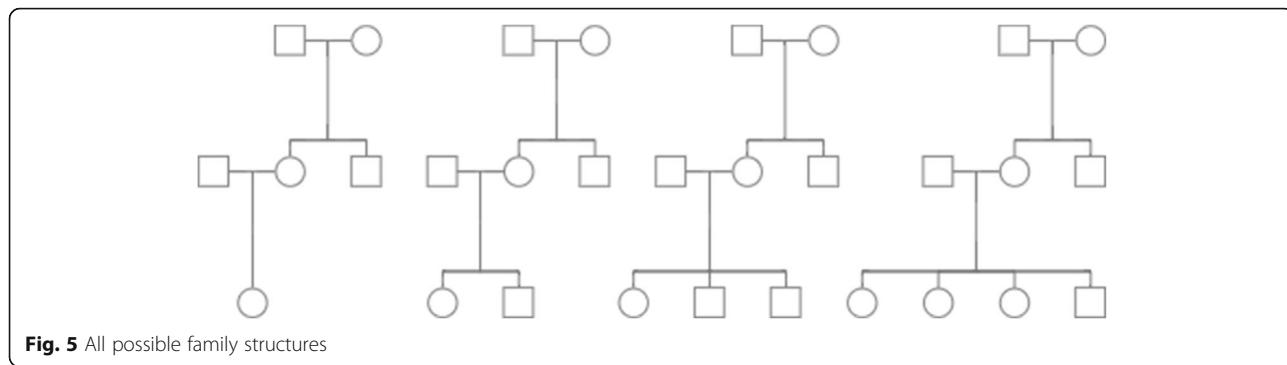
Score tests should be considered for large-scale genetic association testing due to their computational advantage. Because the Wald test also has valid type-I error rates and its computational efficiency is comparable to the

score test (Table 7), if computing resources allow, the Wald test can also be applied for large-scale genetic studies. As illustrated using Framingham heart study data, the proposed score test has enabled the identification of several loci associated with diabetes. One of the drawbacks of the score test is the lack of effect estimates. When only a handful of associated variants are identified from a genetic association study, the effect size and statistical significance of each variant can be estimated using the Wald test. In addition to the multinomial application, we have also provided a computer implementation for ordinal traits in Additional File 2. Although we presented association results from additively coded genetic variants, the application and implementation are not restricted to SNPs, but also applicable to a genetic risk score, weighted-sum gene test [14], and other genetic summary measures.

Methods

Assuming that there are N independent families ($i = 1, \dots, N$), with n_i individuals in family i and a total of $n = \sum_{i=1}^N n_i$ subjects, the basic model for a K -category (multinomial) trait, with the K th level chosen as the reference level, is written as,





$$g(Y_{ij} = k | X_{ij}, G_{ij}) = \alpha_k + \beta_k G_{ij} + X_{ij}^T \gamma_k \quad k = 1, \dots, (K-1)$$

The $n \times 1$ response variable Y has K unordered levels, i.e. $k = 1, \dots, K$, resulting in $(K-1)$ equations; G is the genotype vector of size $n \times 1$; X is the $n \times q$ covariates matrix; $\alpha = (\alpha_1, \dots, \alpha_k, \dots, \alpha_{K-1})^T$ is the intercept vector for the $(K-1)$ equations; $\beta = (\beta_1, \dots, \beta_k, \dots, \beta_{K-1})^T$ is the effect size vector of the genotype in the $(K-1)$ equations; and $\gamma = (\gamma_1, \dots, \gamma_k, \dots, \gamma_{K-1})$ are the parameters of the covariates X , for the $(K-1)$ equations with a dimension of $q \times 1$ for each γ_k . Although there are a variety of choices for the link function g , here we demonstrate with the canonical link function, the general logit, i.e.

$$g(Y_{ij} = k | X_{ij}, G_{ij}) = \log \frac{P(Y_{ij} = k | X_{ij}, G_{ij})}{P(Y_{ij} = K | X_{ij}, G_{ij})}$$

Extended generalized estimating equations (EGEE)

We adopt the idea of EGEE previously proposed [15, 16] to approximate the likelihood using quasi-likelihood, to handle correlated observations. The variance of the response variable Y_{ij} , is defined using $(K-1)$ indicator variables as follows: $z_{ij} = [I(Y_{ij} = 1), \dots, I(Y_{ij} = (K-1))]'$. The expected value of z_{ij} is $E[z_{ij}] = [P(Y_{ij} = 1), \dots, P(Y_{ij} = (K-1))]'$ and the variance of z_{ij} can be derived as:

$$\begin{aligned} \text{var}(z_{ij}) &= \begin{pmatrix} \text{var}(I(Y_{ij} = 1)) & \dots & \text{cov}(I(Y_{ij} = 1), I(Y_{ij} = (K-1))) \\ \vdots & \ddots & \vdots \\ \text{cov}(I(Y_{ij} = (K-1)), I(Y_{ij} = 1)) & \dots & \text{var}(I(Y_{ij} = (K-1))) \end{pmatrix} \\ &= \begin{pmatrix} P(Y_{ij} = 1)(1 - P(Y_{ij} = 1)) & \dots & -P(Y_{ij} = 1)P(Y_{ij} = (K-1)) \\ \vdots & \ddots & \vdots \\ -P(Y_{ij} = (K-1))P(Y_{ij} = 1) & \dots & P(Y_{ij} = (K-1))(1 - P(Y_{ij} = (K-1))) \end{pmatrix} \end{aligned}$$

Let $R = rJ$ where J is a matrix of ones with a dimension of $(K-1)$ by $(K-1)$, and r is an unknown correlation parameter to be estimated with value between -1 and 1 . The implementation of the approach provided in Additional File 1 can also accommodate two-parameter R

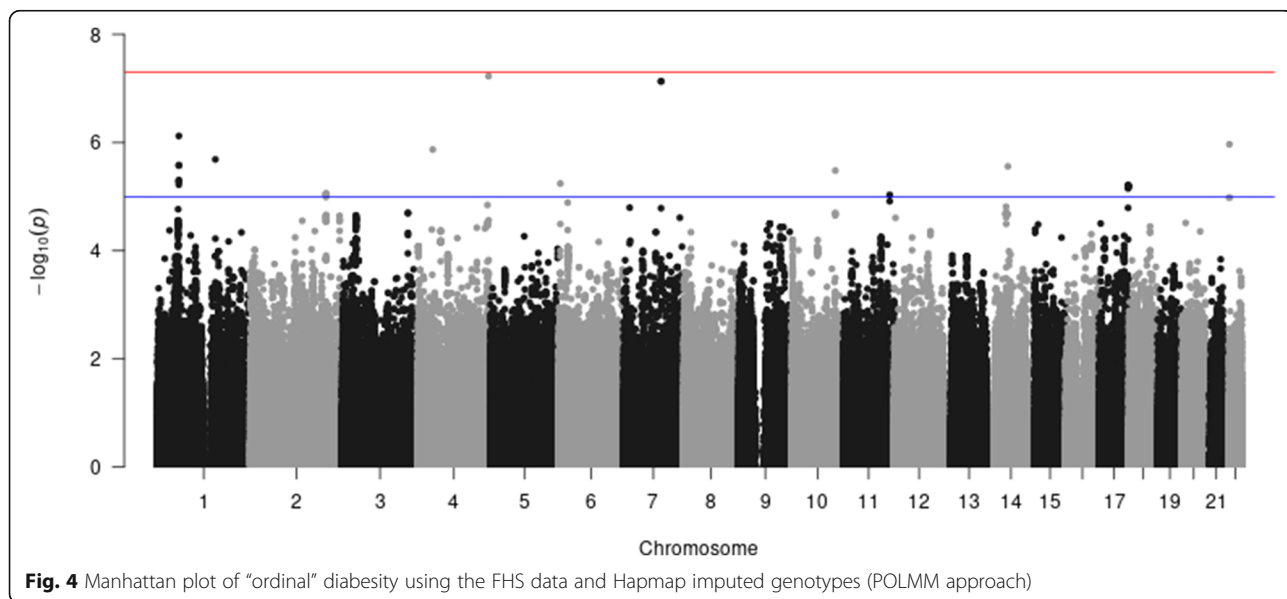


Table 7 Computing time of robust score and Wald tests on a i7-8565u processor with 16GB RAM

Sample size	Robust score test	Wald test
5000 (182 families)	3.09 s (initial) + 1.17 s per SNP	3.44 s per SNP
10,000 (364 families)	4.47 (initial) + 2.45 s per SNP	6.92 s per SNP
20,000 (728 families)	5.47 (initial) + 5.25 s per SNP	10.95 s per SNP

with diagonal elements set to $r1$ and all off-diagonal elements set to $r2$. The matrix R is used to model the correlation between any two individuals in the same family along with the use of relationship matrix, such that $R_i = \Phi_i \otimes R$ (Φ_i is the relationship matrix of the i -th family defined as twice the kinship matrix), similar to how the familial correlation was handled in previous publications [17, 18]. V_i , the overall variance matrix of z_i , for the i -th independent family is constructed as $sd(z_i)R_i sd(z_i)$ with the variance of each subject $var(z_{ij})$ ($j = 1, \dots, n_i$), as derived above,

where

$$sd(z_i) = \begin{pmatrix} sd(z_{i1}) & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & sd(z_{in_i}) \end{pmatrix}$$

and

$$sd(z_{ij}) = \begin{pmatrix} \sqrt{var(I(Y_{ij} = 1))} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sqrt{var(I(Y_{ij} = K-1))} \end{pmatrix} \forall j = 1, \dots, n_i.$$

The following score equations of EGEE [15, 16, 19] are used to estimate the regression parameters $\theta = (\alpha_1, \beta_1, \gamma_1, \dots, \alpha_{K-1}, \beta_{K-1}, \gamma_{K-1})$ and the correlation parameter r .

$$U = \sum_{i=1}^N U_i(\theta, r) = \sum_{i=1}^N \begin{pmatrix} D_i & 0 \\ 0 & F_i \end{pmatrix} \begin{pmatrix} V_i^{-1} & 0 \\ 0 & I \end{pmatrix} \begin{pmatrix} y_i - \mu_i \\ s_i - \sigma_i \end{pmatrix} = 0$$

where the $n_i(K-1) \times (2+q)(K-1)$ matrix D_i is stacked vertically from D_{ij} ($j = 1, \dots, n_i$) and defined as $D_{ij} = \frac{\partial E[z_{ij}]}{\partial \theta} = \left(\frac{\partial P(Y_{ij}=1)}{\partial \theta}, \dots, \frac{\partial P(Y_{ij}=K-1)}{\partial \theta} \right)'$; F_i is the vectorized $\frac{\partial V_i^{-1}}{\partial r}$ with a dimension of $n_i^2(K-1)^2$ by 1, I is an identity matrix with a size of $n_i^2(K-1)^2$ and σ_i is the vectorized version of V_i . Similarly, s_i is vectorized version derived from the following:

$$\begin{pmatrix} e_{i1} \\ \vdots \\ e_{in_i} \end{pmatrix} \begin{pmatrix} e'_{i1} & \dots & e'_{in_i} \end{pmatrix}$$

where $e_{ij} = \left(e_{ij}^1 \dots e_{ij}^{K-1} \right)'$ ($j = 1, \dots, n_i$) and e_{ij}^k ($k = 1, \dots, (K-1)$) is defined as $=I(Y_{ij} = k) - P(Y_{ij} = k)$. Therefore, $E[s_i] = \sigma_i$. Fisher's scoring algorithm is used to update both θ and r from m -th iteration to $(m + 1)$ -th iteration, written as

$$\begin{pmatrix} \theta^{(m+1)} \\ r^{(m+1)} \end{pmatrix} = \begin{pmatrix} \theta^{(m)} \\ r^{(m)} \end{pmatrix} + U^* \left(\theta^{(m)}, r^{(m)} \right)^{-1} \sum_{i=1}^N U_i \left(\theta^{(m)}, r^{(m)} \right)$$

where

$$U^* \left(\theta^{(m)}, r^{(m)} \right) = -E \left[D \sum_{i=1}^N U_i \left(\theta^{(m)}, r^{(m)} \right) \right] = \sum_{i=1}^N \left(\begin{matrix} D_i' V_i^{-1} D_i & 0 \\ F_i' \frac{\partial \sigma_i}{\partial \theta} & F_i' \frac{\partial \sigma_i}{\partial r} \end{matrix} \right) \Bigg|_{\theta = \theta^{(m)}, r = r^{(m)}}$$

and D stands for the first-order derivative with respect to (θ, r) , until the pre-specified convergence criterion is met. Estimates of multinomial logistic regression and $r = 0$ or 0.5 usually work well in terms of starting values.

Note the score equations will be reduced to the following GEE form [20] when applied to N unrelated samples. The coefficients estimation will follow the same iteratively reweighted least square method of generalized linear model [21] for multinomial outcome until a pre-specified convergence criterion is met.

$$U = \sum_{i=1}^N U_i(\theta) = \sum_{i=1}^N D_i' V_i^{-1} (y_i - \mu_i) = 0$$

Robust score test

To determine if a genetic variant is associated with a multi-category phenotype, the following null hypothesis

is tested $H_0: \beta = 0$. We first define the score vectors $U^{(1)}$

$$= \begin{pmatrix} U_{Y_1} \\ \vdots \\ U_{Y_{K-1}} \end{pmatrix}, \quad U^{(2)} = U_\beta = \begin{pmatrix} U_{\beta_1} \\ \vdots \\ U_{\beta_{K-1}} \end{pmatrix}, \quad U^{(3)} = U_r.$$

The score statistic is proposed as follows:

$$s = \left(A(\hat{\theta}_0, \hat{r}_0) U^{main}(\hat{\theta}_0, \hat{r}_0) \right)' \left\{ A(\hat{\theta}_0, \hat{r}_0) \sum_{i=1}^N \left[U_i^{main}(\hat{\theta}_0, \hat{r}_0) U_i^{main}(\hat{\theta}_0, \hat{r}_0)' \right]^{-1} A(\hat{\theta}_0, \hat{r}_0) U^{main}(\hat{\theta}_0, \hat{r}_0) \right\}$$

Where $\hat{\theta}_0, \hat{r}_0$ are parameter estimates under $H_0: \beta = 0$. $U^{main}(\theta, r) = \begin{pmatrix} U^{(1)} \\ U^{(2)} \end{pmatrix} = \sum_{i=1}^N U_i^{main}(\theta, r)$, $A(\theta, r) = (-U_{21}^* U_{11}^{*-1}, I)$ with subscript 2 denoting rows/columns that correspond to β , subscript 1 denoting rows/columns that correspond to $\gamma_1, \dots, \gamma_{K-1}$, and I is an identity matrix of size $(K-1)$.

The score statistic follows a χ_{K-1}^2 asymptotically according to the derivation for bivariate association testing in family samples [17, 22]. One of the major advantages is its robustness to incorrect variance specification. If the variance V_i ($i = 1, \dots, N$) is pre-specified correctly, then $var(U^{main}(\theta, r))$ will equal to U^* restricted to $\beta, \gamma_1, \dots, \gamma_{K-1}$, and the score statistic will be simplified to

$$s = (\mathbf{U}^{(2)}(\hat{\boldsymbol{\theta}}_0, \hat{\mathbf{r}}_0))^T \{V^{(2)}(\hat{\boldsymbol{\theta}}_0, \hat{\mathbf{r}}_0)\}^{-1} \mathbf{U}^{(2)}(\hat{\boldsymbol{\theta}}_0, \hat{\mathbf{r}}_0).$$

where $V^{(2)} = I_{22} - I_{2(-2)} I_{(-2)(-2)}^{-1} I_{(-2)2}$ (The subscript 2 denotes the (K-1) row/columns corresponding to $\beta_1, \dots, \beta_{(K-1)}$; “-” denotes excluding these rows/columns) and $I = \sum_{i=1}^N \mathbf{D}_i' \mathbf{V}_i^{-1} \mathbf{D}_i$.

Wald test

The Wald test is an alternative test with lower computational efficiency when applied to a large-scale genetic study. The Wald test statistic is proposed as follows:

$$w = (\hat{\beta}_1 \dots \hat{\beta}_{(K-1)}) V(\hat{\beta}_1 \dots \hat{\beta}_{(K-1)})^{-1} (\hat{\beta}_1 \dots \hat{\beta}_{(K-1)})'$$

This test statistic follows a χ^2_{K-1} asymptotically. The parameters $\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}}$ are obtained from the score equations with no constraints (i.e. $H_0 \cup H_a$) until the pre-specified convergence criterion is met.

The full variance matrix of all parameters $V(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}})$ is derived as $V(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}}) = \mathbf{U}^*(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}})^{-1} \sum_{i=1}^N \mathbf{U}_i(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}}) \mathbf{U}_i(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}})' (\mathbf{U}^*(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}}))^{-1}$. $V(\hat{\beta}_1 \dots \hat{\beta}_{(K-1)})$ is extracted from $V(\hat{\boldsymbol{\theta}}, \hat{\mathbf{r}})$, a sandwich-type variance estimator [19], with rows and columns corresponding to $(\hat{\beta}_1 \dots \hat{\beta}_{(K-1)})$.

Ordinal traits

Under the same framework, using the statistical theory of ordinal regression, the above score and Wald tests can be easily extended to test the association of a genetic variant with an ordinal trait for a family-based design. More specifically, because $P(Y_{ij} = k)$ can be derived from $P(Y_{ij} = k) = P(Y_{ij} \leq k) - P(Y_{ij} \leq k - 1)$ using proportional cumulative logit models, then the same EGEE equations are used for parameter estimation. However, the dimensions of EGEE equations are reduced and mathematical formulas of the matrix elements are derived differently due to the use of proportional cumulative logit models. A computer implementation for both multinomial and ordinal phenotypes is provided in Additional File 1-2.

Simulations

We conduct type-I error and power simulation studies to evaluate the validity of our score test in assessing the association between single-nucleotide variants (SNVs) with different minor allele frequencies (MAF) and a categorical trait with four categories (“multinomial” trait), and compare the score test to the Wald test and the multinomial logistic regression which does not account for related samples. We then conduct simulations to assess the power of all three approaches.

Type-I error

We compare the type-I error rate of the robust score test to the Wald test as well as multinomial logistic regression (without accounting for related samples) in both family-based and unrelated designs. We simulate a 4-category trait under the null hypothesis that there is no genetic association with the trait, i.e. $H_0: \beta_1 = \dots = \beta_3 = 0$. Eight SNV scenarios with MAF ranging from 0.01 to 0.3 are explored. For each SNV scenario and sample design, 500,000 replicates are simulated and the type-I error rate is defined as the proportion of simulations significant at the threshold of 0.01, 0.001, and 0.0001. For family-based samples, we also have conducted simulations to evaluate the type-I error of robust score and Wald test when applied to ordinal traits, based on 500,000 replicates for each MAF scenario.

Family-based samples: In each replicate, a total of 1000 independent 3-generation families with 2 grandparents who have one son and one daughter (Fig. 5) are simulated. The number of grandchildren (3rd-generation) is randomly determined from a discrete uniform distribution ranging from 1 to 4. Within each of the 1000 families, we simulate additively coded genotypes (0, 1, or 2 minor alleles) of the grandparents under Hardy-Weinberg equilibrium, and the 2nd and 3rd generations’ genotypes are then simulated using random allele dropping. Two covariates (age and sex) are simulated. The sex of the 3rd-generation is randomly assigned, and the covariate of age is simulated in the following way [17]: we start by simulating the age of female offspring (2nd generation) from a continuous uniform distribution ranging from 25 to 50. Her spouse’s age is set to be within 5-year of her age. The male offspring’s ages (2nd generation) are set to be within 5 years of the sister with at least a 1-year gap to exclude twins. Then we simulate the age of the grandparents (1st generation). The grandmother is assumed to be 20 to 45 years older than both offspring (2nd generation), and the grandfather’s age is set to be within 5-year of the grandmother’s age and he must be at least 20 years older than his older offspring. Finally, we simulate the age of the 3rd generation, in such a way that everyone in the 3rd generation is assumed to be 20 to 45 years younger than the mother (2nd generation) and at least 20 years younger than the father (2nd generation). Two continuous traits are simulated from age and sex, based on the following two equations, i.e. age and sex explains around 3 and 0.002% of the total variance of the latent variable u_1 versus 0.8 and 0.01% of the latent variable u_2 :

$$u_1 = 5.6 + 0.025age + 0.5sex + \varepsilon_1;$$

$$u_2 = 30 + 0.04age + 0.2sex + \varepsilon_2;$$

where $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N(0, \boldsymbol{\Sigma}_a \otimes \boldsymbol{\Phi} + \boldsymbol{\Sigma}_e \otimes \mathbf{I})$, the additive covari-

ance matrix is $\Sigma_a = \begin{pmatrix} 4 & 6 \\ 6 & 36 \end{pmatrix}$ and the environmental covariance matrix is $\Sigma_e = \begin{pmatrix} 4 & 6 \\ 6 & 36 \end{pmatrix}$. Φ is the relationship matrix which is a kinship matrix multiplied by 2.

We transform u_1, u_2 to two binary traits using a threshold model with a disease prevalence of 10 and 35%, assuming a disease with a moderate prevalence such as type 2 diabetes (T2D) and a high prevalence such as obesity. The multinomial trait is then defined by these two binary traits as follows: diabetes & obesity, diabetes but no obesity, obesity but no diabetes and no diabetes & no obesity, in adults.

Unrelated samples: In each replicate, we simulate a total of 5000 independent subjects with ages ranging from 18 to 90. A total of 5000 independent additively-coded genotypes are simulated. The sex is randomly assigned (1 = male; 2 = female). We then simulate two continuous traits influenced by age and sex only, based on the following two equations, so that age and sex explain around 3.2 and 0.8% of the total variance of u_1 versus 0.94 and 0.01% of u_2 respectively:

$$u_1 = 5.6 + 0.025age + 0.5sex + \varepsilon_1;$$

$$u_2 = 30 + 0.04age + 0.2sex + \varepsilon_2;$$

where $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim N(0, \Sigma_T \otimes \mathbf{I})$ with $\Sigma_T = \begin{pmatrix} 8 & 12 \\ 12 & 72 \end{pmatrix}$. We transform u_1, u_2 as described in the family design section.

We evaluate the type-I error of the proposed score test and Wald test, and then compare them to the multinomial logistic regression assuming independence among observations (using likelihood ratio test (LRT)).

Power evaluation

We compare the power of the score to the Wald test and multinomial logistic regression under the same allele scenarios and with the same family/unrelated structure as described above. In addition to the effects of age and sex, we also include an additively coded genetic variant g which explains approximately 0.5% of the variance of each continuous trait, i.e.

$$u_1 = 5.6 + 0.025age + 0.5sex + \sqrt{\frac{4 \times 0.01}{2MAF(1-MAF)}}g + \varepsilon_1;$$

$$u_2 = 30 + 0.04age + 0.2sex + \sqrt{\frac{36 \times 0.01}{2MAF(1-MAF)}}g + \varepsilon_2;$$

With this phenotype generation model, both traits are simulated under the alternative hypothesis that there is

an association between the trait and the genetic variant. For each MAF scenario, a total of 5000 replicates are generated. The power rate is then evaluated for 3 different significance thresholds including the commonly used GWAS threshold for each method.

Framingham heart study

The motivation for developing this efficient score test is to make the application to a large-scale genetic study computationally feasible, especially after the cost of whole-genome sequencing has been greatly reduced in recent years.

We apply the robust score test to the Framingham Heart Study (FHS) [17, 23]. A total of 7564 participants from 1315 families are analyzed, after excluding observations with missing values in body mass index (BMI), age, sex, the first 10 principal components (PC) s or T2D status. The primary outcome is diabetes with four categories as defined above. Diabetes is considered a modern epidemic and the largest in human history [24]. However, there are very few papers available regarding genetic association studies on this trait. We analyze the association between diabetes and genotypes from the Framingham SNP Health Association Resource (SHARe) project sponsored by the National Heart, Lung and Blood Institute (NHLBI), adjusting for age, sex, and the first 10 PCs. Genotypes from Affymetrix 550 K genotyping arrays (Affymetrix, Santa Clara, CA, USA), supplemented by the Affymetrix MIPS array, are available on 8481 participants after exclusion for low call rate (< 97%), heterozygosity rate outside of 5 SDs from the mean or excess Mendelian errors (> 1000). Additional SNVs are imputed with the software MACH (Markov Chain-based haplotyper) using the HapMap 2 reference haplotypes [25]. To help understand the GWAS results of diabetes and given the fact that diabetes is jointly constructed from obesity and diabetes, we perform two additional family-based logistic regression analyses using our approach to study the association of diabetes and genotypes, and the association of obesity and genotypes respectively. A secondary outcome treats the diabetes as an ordinal variable with 4 levels of increasing severity. We apply both our ordinal approach and POLMM with derived sparse GRM matrix to the secondary outcome and compare the results.

Abbreviations

GLMM: Generalized linear mixed model; EGEE: Extended generalized estimating equations; SNV: Single-nucleotide variant; MAF: Minor allele frequency; T2D: Type-2 diabetes; FHS: Framingham heart study; BMI: Body mass index; SHARe: SNP Health Association Resource; NHLBI: National Heart, Lung and Blood Institute; MACH: Markov Chain-based haplotyper; POLMM: Proportional Odds Logistic Mixed Model

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12864-021-08107-x>.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgments

We acknowledge the contribution of Achilles Pittsillides to the POLMM analyses. We thank Boston University research computing services and Katia Bulekova for generous assistance with script development and use of the high-performance computing (HPC) cluster.

Authors' contributions

SW analyzed the data. SW, JBM, JD interpreted the data. SW and JD drafted the manuscript. All authors have agreed to the submitted version.

Funding

This work is supported by U.S. National Institutes of Health (NIH) grant U01DK078616, UM1DK078616–13, and 1R01HL151855 awarded to Dr. James B. Meigs. The funding body played no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Availability of data and materials

The FHS dataset that supports the findings of this manuscript is available on dbGap (dbGaP Study Accession: phs000007.v32.p13, https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000007.v32.p13).

Declarations

Ethics approval and consent to participate

The Framingham Heart Study was approved by the Boston University Medical Campus Institutional Review Board and all participants provided written informed consent. Only adult participants' data were analyzed in this manuscript.

Consent for publication

Not applicable.

Competing interests

Authors declare no competing interest.

Author details

¹Pfizer Inc, Global Product Development, Groton, CT 06340, USA. ²Division of General Internal Medicine, Massachusetts General Hospital, Boston, MA 02114, USA. ³Department of Medicine, Harvard Medical School, Boston, MA 02115, USA. ⁴Programs in Metabolism and Medical & Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA 02142, USA. ⁵Department of Biostatistics, Boston University School of Public Health, Boston, MA 02118, USA.

Received: 18 March 2021 Accepted: 19 October 2021

Published online: 04 December 2021

References

1. Therneau T. Mayo Clinic. The Imekin function.
2. Zhang H, Wang X, Ye Y. Detection of genes for ordinal traits in nuclear families and a unified approach for association studies. *Genetics*. 2006; 172(1):693–9. <https://www.ncbi.nlm.nih.gov/pubmed/16219774>. <https://doi.org/10.1534/genetics.105.049122>.
3. Diao G, Lin DY. Variance-components methods for linkage and association analysis of ordinal traits in general pedigrees. *Genetic epidemiology*. 2010; 34(3):232–n/a. <https://www.ncbi.nlm.nih.gov/pubmed/19918762>. <https://doi.org/10.1002/gepi.20453>.
4. Wang X, Philip VM, Ananda G, White CC, Malhotra A, Michalski PJ, et al. A bayesian framework for generalized linear mixed modeling identifies new candidate loci for late-onset alzheimer's disease. *Genetics*. 2018;209(1):51–64. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5937180/>; <https://pubmed.ncbi.nlm.nih.gov/29507048/><https://doi.org/10.1534/genetics.117.300673>.
5. Bi W, Zhou W, Dey R, Mukherjee B, Sampson JN, Lee S. Efficient mixed model approach for large-scale genome-wide association studies of ordinal categorical phenotypes. *The Am J Hum Gen*. 2021;108(5):825–39. <https://www.sciencedirect.com/science/article/pii/S0002929721001038>. <https://doi.org/10.1016/j.ajhg.2021.03.019>.
6. Zimmet PZ. Diabetes and its drivers: The largest epidemic in human history?. 2017;3.
7. Irvin MR, Wineinger NE, Rice TK, Pajewski NM, Kabagambe EK, Gu CC, et al. Genome-wide detection of allele specific copy number variation associated with insulin resistance in african americans from the HyperGEN study. *PLoS One*. 2011;6(8):e24052. <https://doi.org/10.1371/journal.pone.0024052>.
8. Dostalek M, Court MH, Yan B, Akhlaghi F. Significantly reduced cytochrome P450 3A4 expression and activity in liver from humans with diabetes mellitus. *Br J Pharmacol*. 2011;163(5):937–47. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130941/>; <https://pubmed.ncbi.nlm.nih.gov/21323901> <https://doi.org/10.1111/j.1476-5381.2011.01270.x>.
9. Jamwal R, de la Monte, Suzanne M., Ogasawara K, Adusumalli S, Barlock BB, Akhlaghi F. nonalcoholic fatty liver disease and diabetes are associated with decreased CYP3A4 protein expression and activity in human liver. *Mol Pharm*. 2018;15(7):2621–32. <https://doi.org/10.1021/acs.molpharmaceut.8b00159>. <https://doi.org/10.1021/acs.molpharmaceut.8b00159>.
10. Kolwankar D, Vuppalanchi R, Ethell B, Jones DR, Wrighton SA, Hall SD, et al. Association between nonalcoholic hepatic steatosis and hepatic cytochrome P-450 3A activity. *Clin Gastroenterol Hepatol*. 2007;5(3):388–93. <https://doi.org/10.1016/j.cgh.2006.12.021>. <https://doi.org/10.1016/j.cgh.2006.12.021>.
11. Varenhorst C, Eriksson N, Johansson A, et al. Effect of genetic variations on ticagrelor plasma levels and clinical outcomes. *Eur Heart J*. 2015;36(29):1901–12. <https://doi.org/10.1093/eurheartj/ehv116>.
12. Krumsiek J, Suhre K, Evans AM, et al. Mining the unknown: A systems approach to metabolite identification combining genetic and metabolic information. *PLoS Genet*. 2012;8(10):e1003005. <https://doi.org/10.1371/journal.pgen.1003005>.
13. Zanger UM, Schwab M. Cytochrome P450 enzymes in drug metabolism: Regulation of gene expression, enzyme activities, and impact of genetic variation. *Pharmacol Ther*. 2013;138(1):103–41. <https://www.sciencedirect.com/science/article/pii/S0163725813000065>. <https://doi.org/10.1016/j.pharmthera.2012.12.007>.
14. Madsen BE, Browning SR. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet*. 2009;5(2):e1000384. <https://doi.org/10.1371/journal.pgen.1000384>.
15. Hall DB. On the application of extended quasi-likelihood to the clustered data case. *Can J Statistics*. 2001;29(1):77–97. <https://doi.org/10.2307/3316052>. <https://doi.org/10.2307/3316052>.
16. Hall DB, Severini TA. Extended generalized estimating equations for clustered data. *J Am Stat Assoc*. 1998;93(444):1365–75. <https://www.jstor.org/stable/2670052>. <https://doi.org/10.1080/01621459.1998.10473798>.
17. Wang S, Meigs J, & Dupuis, J. Joint association analysis of a binary and a quantitative trait in family studies. 2017;25:130–6.
18. Wang X, Lee S, Zhu X, Redline S, Lin X. GEE-Based SNP set association test for continuous and discrete traits in Family-Based association studies. *Genet Epidemiol*. 2013;37(8):778–86. <https://doi.org/10.1002/gepi.21763>.
19. Liu J, Pei Y, Pappasian CJ, Deng H. Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations. *Genet Epidemiol*. 2009;33(3):217–27. <https://doi.org/10.1002/gepi.20372>.
20. Zeger SL, Liang K-Y, Albert PS. Models for longitudinal data: a generalized estimating equation approach. *Biometrics*. 1988;44(4):1049–60. <https://www.jstor.org/stable/2531734>. <https://doi.org/10.2307/2531734>.
21. Nelder JA, Wedderburn RWM. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*. 1972;135(3):370–84. <https://doi.org/10.2307/2344614>. <https://doi.org/10.2307/2344614>.
22. Davison AC. *Statistical models*: Cambridge University Press; 2003. <https://doi.org/10.1017/CBO9780511815850>.
23. Levy R. The framingham study: The epidemiology of atherosclerotic disease. *JAMA*. 1981;245(5).
24. Farag YMK, Gaballa MR. Diabetes: An overview of a rising epidemic. 2011; 26(1):28–35. <https://doi.org/10.1093/ndt/gfq576>.

25. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol.* 2010;34(8):816–34. <https://doi.org/10.1002/gepi.20533>. <https://doi.org/10.1002/gepi.20533>.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

