ORIGINAL ARTICLE

Multimarker omnibus tests by leveraging individual marker

summary statistics from large biobanks

Angela M Zigarelli[01] | Hanna M Venera[02] | Brody A Receveur[3] | Jack M Wolf[4] | JasonWestra[5] | Nathan L Tintle[6]

[1]Department of Mathematics and Statistics,University of Massachusetts Amherst, Massachusetts, USA

[2]Division of Biostatistics, University of Michigan, Michigan, USA

[3]Department of Statistics, George Mason University, Virginia, USA

[4]Division of Biostatistics, University of Minnesota, Minnesota, USA

[5]Department of Math, Computer Science,

and Statistics, Dordt University, Iowa, USA

[6]Department of Population Health Nursing

Sciences, University of Illinois Chicago,

Illinois, USA

Correspondence

Nathan Tintle Email: ntintle@uic.edu

Summary

As biobanks become increasingly popular, access to genotypic and phenotypic data continues to increase in the form of pre-computed summary statistics (PCSS). Widespread accessibility of PCSS alleviate many issues related to biobank data, including that of data privacy and confidentiality, as well as high computational costs. However, questions remain about how to maximally leverage PCSS for downstream statistical analyses. Here we present a novel method for testing the association of an arbitrary number of SNVs on a linear combination of phenotypes after adjusting for covariates for common multi-marker tests (e.g., SKAT, SKAT-O) without access to individual patient-level data (IPD). We validate exact formulas for each method, and demonstrate their accuracy through simulation studies and an application to fatty acid phenotypic data from the Framingham Heart Study.

KEYWORDS:

Genetic Data Banks, Genetic Markers, Genetic Privacy, Genotype-Phenotype Associations, Statistical

Data Analysis

1   INTRODUCTION

The availability and accessibility of data is an exciting prospect for advancements in science – especially in the area of biomedicalresearch. However, the utilization of individual patient-level data (IPD) raises issues of data privacy and security, and in the biobank era, issues with computational cost and data processing time (Heatherly 2016; Huppertz & Holzinger 2014).

With the popularization of biobanks, there has been an explosion in the magnitude of available genotypic and phenotypic data, to facilitate the connection between genetics and human health. In an effort to make biobank data more accessible and usable, projects like GeneAtlas and PheWeb provide pre-computed summary statistics (PCSS) eliminating the need for some

---

[0]Hanna Venera and Angela Zigarelli contributed equally to this manuscript.

---

researchers to access IPD (Neale 2018; PheWeb 2018). Most often these projects have computed simple linear regression

results between many combinations of genotypes and phenotypes and then made summary statistics like β and $SE(\beta)$ publiclyavailable (Canela-Xandri, Rawlik, & Tenesa 2018; Sudlow et al. 2015). Though this process alleviates the issue of data privacy and confidentiality, as well as high computational costs, questions are raised about how useful PCSS can be given that the scopeof PCSS is often merely results from simple linear regression models.

In an effort to more fully leverage PCSS in downstream analyses, several groups have published methods that utilize GWAS PCSS to perform meta-analyses, as well as multivariate methods to investigate relationships between phenotypes (Cichonska et al. 2016; Liu & Lin 2018; Ray & Boehnke 2018; Stephens 2013; van der Sluis, Posthuma, & Dolan 2013; Vuckovic, Gasparini, Soranzo, & Iotchkova 2015). More recently, Gasdaska et al. (2019) developed a method utilizing PCSS from simplelinear regression models to regress linear combinations of phenotypes against genotypes, with an extension by Wolf, Westra, and Tintle (2021) for the multiplicative phenotype case. Subsequently, Wolf et al. (2020) expanded the approach to be able to use PCSS to adjust for covariates not included in the simple linear regression models on which the PCSS are based. Other recent developments in this area include Svishcheva, Belonogova, Zorkoltseva, and Axenovich (2019) who presented an approach which leverages simple linear regression PCSS to perform gene-based (multi-marker) tests on a single phenotype. Despite these advances, significant gaps in the

literature remain. In particular, methods for conducting multi-marker tests of association uti- lizing complex (e.g., linear/multiplicative combinations) phenotypes are lacking. Furthermore, post-hoc covariate adjustment utilizing PCSS is important and, yet, is lacking for most methods utilizing PCSS to date. While metaSKAT (Lee, Teslovich, TM, and Lin 2013) utilizes PCSS in a multi-marker framework, this approach requires pre-specification of phenotypes and covariates before PCSS are computed, limiting post-hoc research exploration on different sets of phenotypes and/or covariates. Further- more, multiSKAT (D, L, M, and Lee 2019) tests for a general multivariate phenotype limiting power for research hypotheses involving pre-specified, multivariate phenotypes.

To address these gaps, we present a method that calculates various multi-marker tests (e.g., SKAT; SKAT-O), for a linear combination of phenotypes using only PCSS from simple linear regression models of a single phenotype on a single genotype. Thus, our method allows phenotypes, genotypes, and covariates from different analyses to be combined using researcher specified linear combinations of phenotypes, genotype sets and covariates. We provide an analytic framework for our methodology which is validated through simulation. Finally, we apply these methods on real data from the Framingham Heart Study.

2 METHODS

The following sections outline first a method to calculate the F-Test statistic, and second, a method to calculate rare variant tests. Both methods use only PCSS inputs, by which we mean only the estimates and standard errors from a simple linear regression model of a single phenotype on a single genetic marker (e.g., single nucleotide variant; SNV).

2.1 Notation and Assumptions

Throughout this paper we use the column vectors $\boldsymbol{y}_1, \dots, \boldsymbol{y}_m$ to represent a collection of vectors of $n \times 1$ measurements on m

phenotypes such that $\boldsymbol{y}_j = [y_{1j}, \dots, y_{nj}]$ is a vector of n measurements on the jth phenotype. We define $\boldsymbol{y} = \sum_{j=1}^{m} \boldsymbol{\tau}_j \boldsymbol{y}_j$ as n

measurements across a weighted linear combination of m phenotypes where $\boldsymbol{\tau}$ is an $m \times 1$ vector of weights corresponding to each phenotype. We use $\boldsymbol{X} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_p]$ to denote an $n \times p$ design matrix of n individuals on p variables including k SNVs for $p > 1$. We use $\overline{\boldsymbol{y}}$ to represent the $m \times 1$ vector of means for each phenotype and $\overline{\boldsymbol{x}}$ to represent the $p \times 1$ vector of means for the design matrix. We also define $\mathrm{cov}(\boldsymbol{X})$ as the empirical covariance matrix of $\boldsymbol{X}$ such that the i, j^{th} entry in the matrix represents the covariance of $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Moreover, we assume that $\sum_{j=1}^{m} \boldsymbol{\tau}_j \boldsymbol{y}_j = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$ which is a standard assumption for both the F and rare variant tests used to derive the test statistics' sampling distributions. We also assume to have PCSS including the means for each phenotype, genotype, and covariate, as well as the full covariance matrix, i.e. $\mathrm{cov}(\boldsymbol{y}_1, \dots, \boldsymbol{y}_m, \boldsymbol{X})$, which we will leverage to perform the omnibus tests of interest. These statistics can be aggregated from a variety of sources including single-marker GWAS results. For example, the sample covariance of a given SNV and a phenotype can be calculated by diving the GWAS simple linear regression slope coefficient by the sample variance of the SNV (Wolf et al., 2021). Figure 1 proposes one framework for collecting and compiling this information in practice.

Though we will assume to have a known covariance matrix of the phenotypes for our methods to produce calculations identical to IPD, it may be approximated using correlations of GWAS test statistics for each phenotype by techniques proposed by Kim, Bai, and Pan (2015); Zhu et al. (2015).

2.2 F-Tests for a Linear Model

The F-test can be used to evaluate the combined effect of several linear predictors on a response. Consider the design matrix $\widetilde{\boldsymbol{X}}$ that contains a subset of q selected attributes from the design matrix for $q < p$. For example, $\widetilde{\boldsymbol{X}}$ could be a collection of covariates that does not include the genetic markers of interest. Then, the test has null hypothesis $\boldsymbol{y} = \widetilde{\boldsymbol{X}}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$ for $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$ and has test statistic:

$$F = \frac{(SSR_R - SSR_F)/(p-q)}{SSR_F/(n-p)}$$

(1)

which follows an $F_{p-q,n-p}$ distribution under the null hypothesis. We define SSR_F as the sum of squared residuals for the full model $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ with p predictors including the intercept, and SSR_R as the sum of squared residuals for the reduced model $\widehat{\boldsymbol{y}} = \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{\gamma}}$. The sum of squared residuals (SSR) can be expressed as follows (see Appendix for details):

$$SSR = \boldsymbol{y}^T\boldsymbol{y} - (\boldsymbol{X}^T\boldsymbol{y})^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \tag{2}$$

In order to express SSR in terms of PCSS, we use the following from Wolf et al. (2020):

$$\boldsymbol{y}^T\boldsymbol{y} = \sum_{h=1}^{m}\sum_{j=1}^{m}\tau_h\tau_j\left(\mathrm{cov}(\boldsymbol{y}_h,\boldsymbol{y}_j)(n-1) + \bar{\boldsymbol{y}}_h\bar{\boldsymbol{y}}_j n\right) \tag{3}$$

$$\boldsymbol{X}^T\boldsymbol{y} = (n-1)\left[\mathrm{cov}(\boldsymbol{x}_1,\boldsymbol{y}),\dots,\mathrm{cov}(\boldsymbol{x}_p,\boldsymbol{y})\right]^T + n\bar{\boldsymbol{x}}\bar{\boldsymbol{y}}^T\boldsymbol{\tau} \tag{4}$$

$$\boldsymbol{X}^T\boldsymbol{X} = (n-1)\,\mathrm{cov}(\boldsymbol{X}) + n\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^T \tag{5}$$

where $\mathrm{cov}(\boldsymbol{x}_i,\boldsymbol{y}) = \tau_1\mathrm{cov}(\boldsymbol{x}_1,\boldsymbol{y}_j) + \cdots + \tau_m\mathrm{cov}(\boldsymbol{x}_p,\boldsymbol{y}_j)$. We can then evaluate $\boldsymbol{y}^T\boldsymbol{y}$, $\boldsymbol{X}^T\boldsymbol{y}$, and $\boldsymbol{X}^T\boldsymbol{X}$ using the appropriate subset of $\boldsymbol{X}$ to calculate F from SSR and compare it to its null distribution to obtain a *p*-value under the null hypothesis.

2.3 Rare Variant Tests

Rare variant tests including burden, SKAT, and SKAT-O also test the null hypothesis that a collection of features have no effect on a response, are useful in rarer genetic cases where the F-Test may be less efficient (Lee et al. 2012; Li & SM 2008; Wu et al. 2011). The following sections detail how to implement these tests using only PCSS inputs.

Consider the partition of our covariate space $\boldsymbol{X} = [\boldsymbol{G}, \boldsymbol{Z}]$ where $\boldsymbol{G} = [\boldsymbol{g}_1, \dots, \boldsymbol{g}_k]$ is an $n \times k$ genotype matrix, $\boldsymbol{g}_j$ is an $n \times 1$ column vector of minor allele counts at the jth variant, $\boldsymbol{Z} = [\boldsymbol{z}_1, \dots, \boldsymbol{z}_c]$ is an $n \times c$ matrix of covariates, and $\boldsymbol{z}_l$ is an $n \times 1$ vector for the lth covariate. Similar to the F-test, the null hypothesis of the

rare variant tests assumes that $\boldsymbol{y} = \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$ and $\boldsymbol{\varepsilon} \sim N(\boldsymbol{0}, \sigma^2 I_n)$ The test statistics can be obtained by first calculating a site-specific score statistic for each SNV in $\boldsymbol{G}$ and then combining the score statistics into one statistic. We use $\bar{\boldsymbol{g}}$ to represent the $k \times 1$ vector of means for each SNV and $\bar{\boldsymbol{z}}$ to represent the $c \times 1$ vector of means for each covariate. For the null model $\boldsymbol{y} = \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$, let $\hat{\boldsymbol{y}} = \boldsymbol{Z\hat{\alpha}}$ be the $n \times 1$ vector of the fitted values.

2.3.1 Score Statistic Framework

A site score statistic that captures how much the minor alleles at variant j empirically contribute to increases in a continuousphenotype as referenced in Wu et al. (2011) is

$$S_j = \boldsymbol{g}_j^T (\boldsymbol{y} - \hat{\boldsymbol{y}})/\hat{\sigma}^2 \tag{6}$$

where $\hat{\sigma}^2$ is the unbiased estimator of σ^2 under the null model. This can be expressed using PCSS inputs as follows (see Appendix for derivation):

$$S_j = \frac{(n-1)}{\hat{\sigma}^2} \left(\mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y}) - \sum_{l=1}^{c} \hat{\alpha}_l \, \mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{z}_l) \right) \tag{7}$$

where

$$\hat{\boldsymbol{\alpha}} = (\boldsymbol{Z}^T \boldsymbol{Z})^{-1} \boldsymbol{Z}^T \boldsymbol{y} \tag{8}$$

is the vector of slope coefficients under the covariate-only model, which can be computed by calculating $\boldsymbol{Z}^T \boldsymbol{y}$ and $\boldsymbol{Z}^T \boldsymbol{Z}$ using Equations 4 and 5 above. We note that $\hat{\sigma}^2 = SSR/(n-q)$ using SSR from the covariate only model. If there is no covariate adjustment in the null model, the expression can be simplified to

$$S_j = (n-1) \, \mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y})/\hat{\sigma}^2.$$

2.3.2 Calculating the Unified Score Test Statistic

The site-specific score statistics can then be aggregated to assess the overall contribution of all k SNVs with $Q_{SKAT} = \sum_{j=1}^{k} w_j S_j^2$ and $Q_{Burden} = (\sum_{j=1}^{k} w_j S_j)^2$ where $\boldsymbol{w}$ is a vector of weights (Lee et al. 2012; Li & SM 2008; Wu et al. 2011). A common set of weights $\boldsymbol{w}$ are the suggested weights as proposed in Wu et al. (2011) given as the square root of the density of a beta(2, 25) distribution evaluated at MAF_j The unified score statistic merges Q_{SKAT} and Q_{Burden} to appeal to advantages of both test statistics, and may be written for the entire set of variants as referenced in Lee et al. (2012):

$$Q_\rho = \rho Q_{Burden} + (1-\rho) Q_{SKAT} \tag{9}$$

for some weight $\rho \in [0,1]$. The optimized SKAT (SKAT-O) test statistic is a special case of the unified score Q_ρ such that the weight ρ yields the minimum p-value. The weight ρ^* can be calculated numerically or analytically as seen is Lee et al. (2012), and the SKAT-O test statistic is written as

$$Q_{SKAT-O} = \rho^* Q_{Burden} + (1-\rho^*) Q_{SKAT} \tag{10}$$

We note that burden and SKAT test statistics are also special cases of the unified score statistic Q_ρ that occur when ρ equalsone and zero, respectively.

The test statistic's null distribution follows a mixture of independent $\chi^2(1)$ distribution such that under the null, $Q_\rho \sim \sum_{j=1}^{k} \lambda_j \chi_j^2(1)$ (Lee et al. 2013). We derive an expression for the weights for this null distribution Q_ρ in terms of PCSS.

Treating this as a meta-analysis for one cohort and following the framework proposed by Lee et al. (2013), we reference the following expression where $\boldsymbol{\Phi}$ is the $k \times k$ between-variant relationship matrix (Lee et al. 2013):

$$\boldsymbol{\Phi} = (\boldsymbol{G}^T\boldsymbol{G} - (\boldsymbol{Z}^T\boldsymbol{G})^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{G})/\hat{\sigma}^2 \tag{11}$$

We find that $\boldsymbol{Z}^T\boldsymbol{G}$ can be expressed using PCSS as

$$\boldsymbol{Z}^T\boldsymbol{G} = (n-1)\text{cov}(\mathbf{Z}, \mathbf{G}) + n\bar{\mathbf{z}}\bar{\mathbf{g}}^T \tag{12}$$

where $\text{cov}(\boldsymbol{Z,\,G})$ represents the $c \times k$ covariance matrix of $\boldsymbol{Z}$ and $\boldsymbol{G}$ such that $\text{cov}(\boldsymbol{Z,\ G})_{(i,j)} = \text{cov}(\boldsymbol{z}_i, \boldsymbol{g}_j)$. Thus, $\boldsymbol{\Phi}$ and $\boldsymbol{Z}^T\boldsymbol{G}$

can be expressed using only PCSS inputs, and we solve for the matrix $\boldsymbol{\Phi}_\rho$ which is defined as Lee et al. (2013):

$$\boldsymbol{\Phi}_\rho = \boldsymbol{L}_\rho^T\boldsymbol{W}\boldsymbol{\Phi}\boldsymbol{W}\boldsymbol{L}_\rho \tag{13}$$

where $\boldsymbol{L}_\rho$ is the Cholesky decomposition of the $k \times k$ compound symmetric matrix $\boldsymbol{R}_\rho = (1-\rho)\boldsymbol{I} + \rho\boldsymbol{1}\boldsymbol{1}^T$ and $\boldsymbol{W}$ is a diagonal matrix with weights $w_1, w_2, \ldots, w_k$ corresponding to the column order of $\boldsymbol{G}$ (Lee et al., 2013). Then the weights $\lambda_1, \lambda_2, \ldots, \lambda_k$ are the nonzero eigenvalues of $\boldsymbol{\Phi}_\rho$. To find the optimal ρ^*, we calculate $\mathbf{Q}_\rho$ for each ρ over a grid and use the corresponding weights to calculate the associated p-values using Davies method. Then, we choose ρ^* such that the *p*-value is minimized to find the SKAT-O score statistic Q_{SKAT-O} (Lee et al. 2013).

2.4   Simulation

To evaluate the accuracy of our proposed method we performed a large simulation study. We simulated three correlated SNVs across 2,000 subjects and generated minor allele frequency (MAF) using a beta(1, 4) (divided by four in order to simulate rare variants). The correlation matrix was sampled uniformly from the set of all possible 3×3 correlation matrices given these MAFs. Subjects' age and sex were generated independently from a Poisson distribution and a Bernoulli distribution, respectively. Each subject had three phenotypes: $\boldsymbol{y}_1$, $\boldsymbol{y}_2$, and $\boldsymbol{y}_3$ which were generated from a multiple linear regression with coefficients age, sex, and

the three SNVs. Error terms were generated from a multivariate normal distribution with mean $\mathbf{0}$ and covariance Σ where $\boldsymbol{\Sigma}$ was randomly generated in each simulation using correlations for each pair of phenotypes drawn from a uniform(-0.5, 0.5) distribution, to simulate positive and negative relationships between phenotypes, and error variances held fixed at one for eachphenotype. Coefficients for the SNVs and covariates were generated to simulate both negative and positive linear relationshipswith the phenotypes. We held the intercept at a constant value of zero across all simulations to represent an arbitrary intercept. We then considered the linear combination $\tau_1 \boldsymbol{y}_1 + \tau_2 \boldsymbol{y}_2 + \tau_3 \boldsymbol{y}_3$ as the primary response for all of our analyses where arbitrary weights τ_j were independently generated from a uniform$(0.5, 5)$ distribution.

2.4.1   F-Test Simulation

To assess our F-Test method's accuracy compared to an IPD-computed F-Test, we calculated *SSR* using the simulated IPD andPCSS for the full model using all predictors, and a reduced model excluding both SNV1 and SNV2. We then performed a nested F-Test using only PCSS, calculated the F-Test statistic and the corresponding p-value for the linear combination of phenotypes, and evaluated bias in our methods by comparing our result to the F-Test statistic using IPD.

2.4.2   Rare Variant Test Simulation

To assess our rare variant tests, we calculated the test statistic $\boldsymbol{Q}_{p^*}$ that combined all $\boldsymbol{S}_j$ across the region using the simulated IPD and PCSS using the suggested weights as proposed in Wu et al. (2011). To assess the potential effect of covariates in the model, we also calculated a *p*-value using

a standard SKAT-O test on the IPD. We note that by calculating the SKAT-O score statistic, SKAT and burden are trivialextensions.

2.5   Real Data Example

Several groups have investigated the relationship between SNPs and fatty acid levels in the blood using data from the Framingham Heart Study (Kalsbeek et al. 2018; Tintle et al. 2015; Veenstra et al. 2017; Wolf et al. 2020). We applied our method to an unrelated subset of Generation 3 and Offspring cohorts from the Framingham Heart Study using the sequence data on 1,212 subjects across 20 genes with common variants associated with omega 3 fatty acids as indicated by the GWAS catalog. We tested if the rare variants in these genes had an effect on the omega 3 index (EPA + DHA) adjusting for age and sex. We used dbGene(NLM n.d.) for the start and end points of each gene referencing human genome build 38, and all variants in the minor allele frequency window between 0.01 and 0.05. When examining the nested F-test, we used all variants with minor allele frequency greater than 0.05 for the SKAT-O test, and an absolute correlation less than 0.9 compared to all other SNPs in the subset.

After calculating basic summary statistics for our phenotypes (EPA, DHA, age, and sex) and all genotypes, we used the

summary statistics to calculate the SKAT-O score statistics and p-value. We then compared our results to those found using IPD and the meta analysis null distribution. For the nested F-test, we compared a model with all SNPs, age, and sex to a model with only age and sex.

3   RESULTS

The following sections detail our results from our simulation and real data application to the Framingham Heart Study.

3.1   Simulation F-Test Statistic Results

We performed two nested F-Tests with 3 SNVs: one full model adjusting for age and sex, and one full model adjusting only for sex. Table 1 shows the accuracy of our results for both nested F-Test simulations averaged across 100,000 iterations. We show the precision of our method with respect to calculated test statistics and the associated p-values in Figures 2 and 3. In short, our method to describe covariate-adjusted models with multiple SNVs on a linear combination of phenotypes with a nested F-Test proved to be exact to rounding

errors. As expected, due to the exactness of the method, simulation shows maintenance of the type I error rate and power as compared to the IPD method (detailed results not shown).

3.2   Simulation Rare Variant Test Statistic Results

We tested the full model with three SNVs adjusting for age and sex, and the reduced model removing age. Table 2 shows the flexibility of our results for the SKAT-O simulations averaged across 100,000 iterations, as well as results for SKAT and burden test for completeness, noting that these results are trivial after showing the accuracy of the SKAT-O test. We show the precision of our method with respect to calculated test statistics and the associated p-values for each of the three tests in Figures 4 and 5 for the SKAT-O test. Similarly, Figures 6 and 7 for SKAT and Figures 8 and 9 for burden tests are included in the appendix, as they are trivial extensions. Again, our method to describe covariate adjusted models with multiple SNVs with a SKAT-O test statistic proved to be exact to rounding errors. As expected, due to the exactness of the method, simulation shows maintenance of the type I error rate and power as compared to the IPD method (detailed results not shown).

3.3   Real Data Results

As indicated in Tables C1 and C2, the results from our PCSS methods using the Generation 3 and Offspring cohorts from the Framingham Heart Study are equal to IPD methods to rounding errors. No gene p-value was less than the Bonferroni corrected cut off of 0.0025. Small bias is due to rounding error and missing data and the bias is similar across a varied number of different SNVs. CONCLUSIONS

We have developed and demonstrated exact methods for F-Tests and common rare variant tests on a linear combination of continuous phenotypes regressed on an arbitrary number of SNVs and covariates using only PCSS. We provided the mathematical framework behind these methods, and validated them through simulation and a real data application using the Framingham heart study.

Since the method we have developed is mathematically exact, limitations due to methodological approximations are limited.There are, however, some limitations worth noting. For example, the rare variant tests we have derived here are only robust forcontinuous phenotypes. Extensions of the method as shown to binary or other categorical variables may not be robust and is anarea of future research. We note that while we simulated the variance of the SNVs under the assumption of Hardy Weinberg Equilibrium using the minor allele frequency. In practice, if the variance of the SNVs is known, the HWE assumption becomes unnecessary.

We have applied the method to a well-studied genome wide dataset (Framingham) and shown that the method provides exactsolutions in this context as well. However, future research is needed in order to more fully vet the method on real data to explore its robustness to larger amounts of missing data, skewness and outliers, and situations that arise when some PCSS are not available for the cohort of interest.

Finally, we note that estimating the covariance matrix of the phenotypes using the correlation matrix of slope coefficients andeach individual genotype has been shown by Kim et al. (2015); Zhu et al. (2015) and can be utilized if needed. Earlier work on combined phenotypes (Gasdaska et al. 2019) using this approach showed unbiased behavior with some power loss due to anincrease in variability. In the case where correlations or values are estimated and investigators wish to evaluate the sensitivity of their findings to these estimated quantities, we recommend re-running the method using different estimates of these quantities to qualitatively understand the sensitivity of their findings to the estimated values. Future simulation studies should continue to explore the sensitivity of this approach and other PCSS-based methods to estimated and incorrect PCSS inputs.

ACKNOWLEDGMENTS

Data availability

The data that support the findings of this study are openly available in dbGaP at

```
https://www.ncbi.nlm.nih.gov/ projects/gap/cgi-
bin/study.cgi?study_id=phs000007.v32.p13
```
[phs000007].

Author Contributions

JW(1) (Jack Wolf), JW(2) (Jason Westra) and NT conceived of the study. HV, AZ and BR participated in initial discussions and explorations of an approach and relevant literature review, with HV and AZ finalizing mathematical derivations with JW(1). AZ led development and implementation of the simulation with input from JW(2). JW(2) led the real data analysis. HV led manuscript development, AZ created figures and with additional substantive input from all co-authors. All co-authors read andapproved the final manuscript.

Conflict of Interest

The authors declare no potential conflict of interests.

References

Canela-Xandri, O., Rawlik, K., & Tenesa, A. (2018). An atlas of genetic associations in uk biobank. *Nature Genetics*, *50*, 1593–1599.

Cichonska, A., Rousu, J., Marttinen, P., Kangas, A., Soininen, P., Lehtimäki, T., . . . Pirinen, M. (2016). metacca: sum- mary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis.*Bioinformatics*, *32*(13), 1981—1989.

D, D., L, S., M, B., & Lee, S. (2019). Multi-skat: General framework to test for rare-variant association with multiple phenotypes.

Genetic Epidemiology, *43*(1), 4–23.

Gasdaska, A., Friend, D., Chen, R., Westra, J., Zawitowski, M., Lindsey, W., & Tintle, N. (2019). Leveraging summary statistics to make inferences about complex phenotypes in large biobanks. *Pacific Symposium on Biocomputing*, *24*, 391–402.

Heatherly, R. (2016). Privacy and security within biobanking: The role of information technology. *Journal of Law, Medicine Ethics*, *44*(1), 156–160.

Huppertz, B., & Holzinger, A. (2014). Biobanks – a source of large biological data sets: Open problems and future challenges.In A. Holzinger & I. Jurisica (Eds.), *Interactive Knowledge Discovery and Data mining in Biomedical Informatics.* Berlin:Springer. ISBN 978-3-662-43967-8.

Kalsbeek, A., Veenstra, J., Westra, J., Disselkoen, C., Koch, K., McKenzie, K. A., . . . Tintle, N. L. (2018). A genome-wide asso- ciation study of red-blood cell fatty acids and ratios incorporating dietary covariates: Framingham heart study offspring cohort. *PLoS One*, *13*(4), e0194882.

Kim, J., Bai, Y., & Pan, W. (2015). An adaptive association test for multiple phenotypes with gwas summary statistics. *GeneticEpidemiology*, *39*(8), 651–663.

Lee, S., Emond, M., Bamshad, M., Barnes, K., Rieder, M., & Nickerson, D. (2012). Optimal unified approach for rare-variant

association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal ofHuman Genetics*, *91*(2), 224–237.

Lee, S., Teslovich, TM, M., Boehnke, & Lin, X. (2013). General framework for meta-analysis of rare variants in sequencingassociation studies. *American Journal of Human Genetics*, *93*(1), 42–53.

Li, B., & SM, L. (2008). Methods for detecting associations with rare variants for common diseases: Application to analysis ofsequence data. *American Journal of Human Genetics*, *83*(3), 311–321.

Liu, Z., & Lin, X. (2018). Multiple phenotype association tests using summary statistics in genome-wide association studies.

Biometrics, *74*(1), 165–175.

Neale, B. M. (2018). *Biobank gwas.* Retrieved from

`http://www.nealelab.is/uk-`

   `biobank/`

NLM. (n.d.). Dbgene. Retrieved from

`https://www.ncbi.nlm.nih.gov/gene`

*Pheweb.* (2018). Retrieved from `https://pheweb.sph.umich.edu/`

Ray, D., & Boehnke, M. (2018). Methods for meta-analysis of multiple traits using gwas summary statistics. *GeneticEpidemiology*, *42*(2), 134–145.

Stephens, M. (2013). A unified framework for association analysis with multiple related phenotypes. *PLOS ONE*, *14*(3), e0213951.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., & et al. (2015). Uk biobank: An open access resource foridentifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, *12*(3), e1001779.

Svishcheva, G., Belonogova, N., Zorkoltseva, K. A., I, & Axenovich, T. (2019). Gene-based association tests using gwas summary statistics. *Bioinformatics*, *35*(19), 3701–3708.

Tintle, N. L., Pottala, J. V., Lacey, S., Ramachandrane, V., Westra, J., Rogers, A., . . . Sheareri, G. C. (2015). A genome-wide association study of saturated, mono- and polyunsaturated red blood cell fatty acids in the framingham heart offspring study. *Prostaglandins, Leukotrienes and Essential Fatty Acids*, *94*, 65-72.

van der Sluis, S., Posthuma, D., & Dolan, C. (2013). Tates: Efficient multivariate genotype-phenotype analysis for genome-wideassociation studies. *PLOS Genetics*, *9*, e1003235.

Veenstra, J., Kalsbeek, A., Westra, J., Disselkoen, C., Smith, C. E., & Tintle, N. (2017). Genome-wide interaction study of omega-3 pufas and other fatty acids on inflammatory biomarkers of cardiovascular health in the framingham heart study.*Nutrients*, *9*(8), 900.

Vuckovic, D., Gasparini, P., Soranzo, N., & Iotchkova, V. (2015). Multimeta: an r package for meta-analyzing multi-phenotypegenome-wide association studies. *Bioinformatics*, *31*(16), 2754–2756.

Wolf, J., Barnard, M., Xia, X., Ryder, N., Westra, J., & Tintle, N. (2020). Computationally efficient, exact, covariate-adjusted genetic principal component analysis by leveraging individual marker summary statistics from large biobanks. *PacificSymposium on Biocomputing*, *25*, 719–730.

Wolf, J., Westra, J., & Tintle, N. (2021). Using summary statistics to model multiplicative combinations of initially analyzedphenotypes with a flexible choice of covariates. *Frontiers in Genetics*, *12*.

Wu, M., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing

   data with thesequence kernel association test. *American Journal of Human Genetics*, *89*(1), 82–93.

Zhu, X., Feng, T., Tayo, B., Liang, J., Young, J., Franceschini, N., … Rotimi, C. (2015). Meta-analysis of

   correlated traits viasummary statistics from gwass with an application in hypertension. *American*

   Journal of Human Genetics, *96*(1), 21–36.

TABLES AND FIGURES

FIGURE 1 Workflow for an omnibus test using only pre-computed summary statistics.[1]
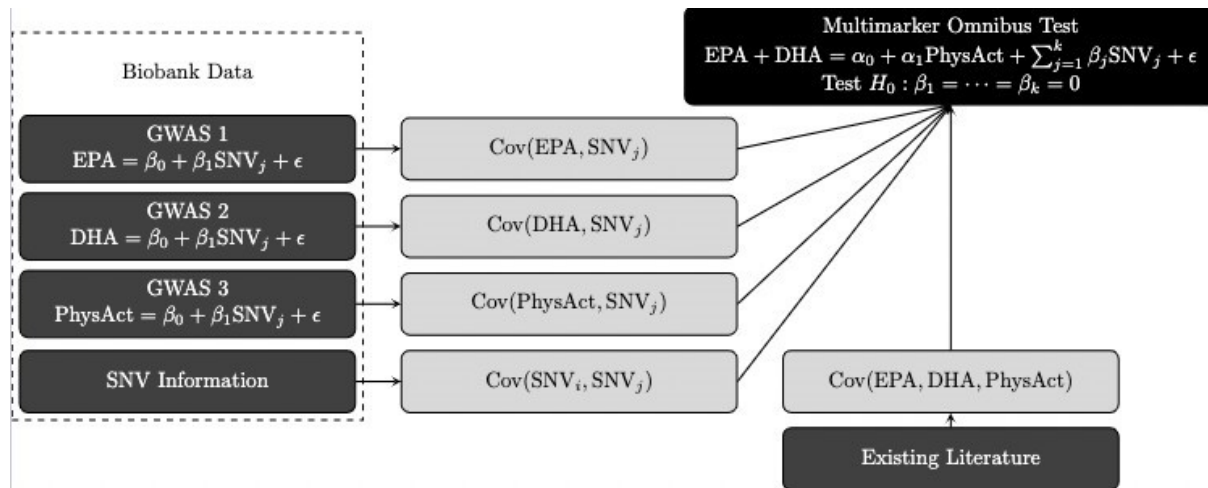
TABLE 1 The accuracy of our methods to estimate the nested F-Test statistic for the reduced model adjusting for covariatesusing PCSS. Errors were minimal with low variance in all cases.

Omnibus Statistic	Full Model	Reduced Model	Mean Error	Variance
F-Test Statistic	SNV1, SNV2, SNV3, Age, Sex	Age, Sex	3.59×10^{-16}	9.48×10^{-27}
P-Value	SNV1, SNV2, SNV3, Age, Sex	Age, Sex		
	SNV1, SNV2, SNV3, Sex		1.03×10^{-17}	7.68×10^{-28}
F-Test Statistic		Sex		

-3.82×10^{-17} 4.13×10^{-27}

P-Value	SNV1, SNV2, SNV3, Sex	Sex	3.48×10^{-17}	4.15×10^{-28}

FIGURE 2 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to a nested F-Test adjusted for age and sex (example 1). These figures illustrateour method's high accuracy on simulated data.

Example 1: Comparison of PCSS and IPD nested F-test statistics and p-values
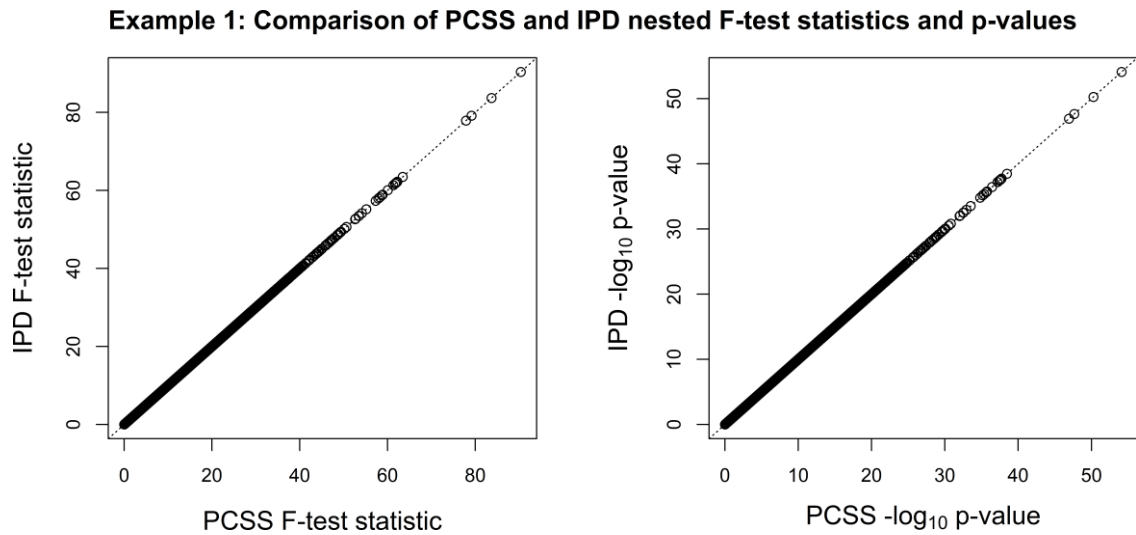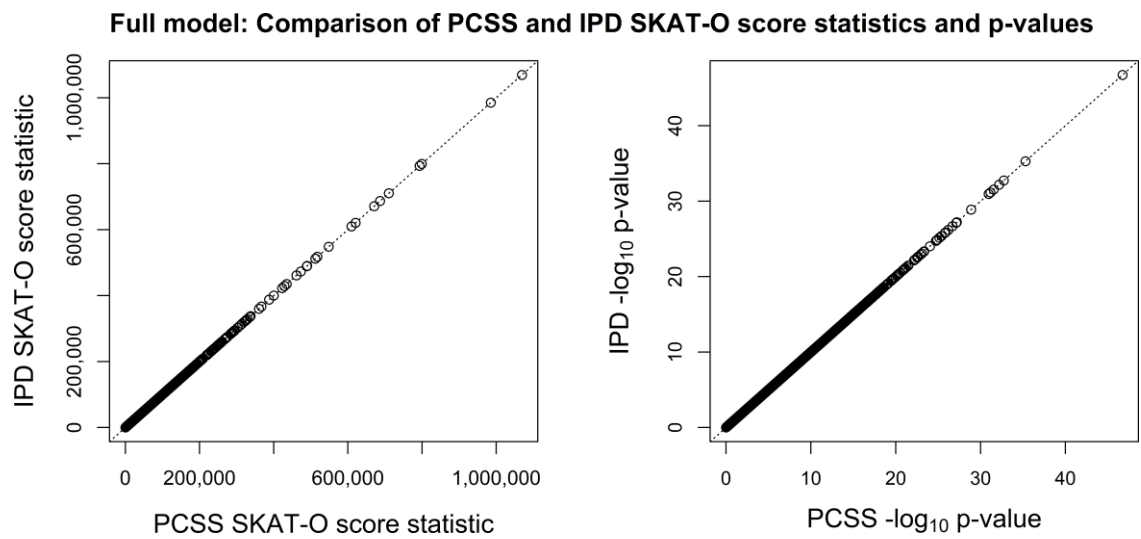
FIGURE 3 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to a nested F-Test adjusted for sex (example 2). These figures illustrate our method's high accuracy on simulated data.

Example 2: Comparison of PCSS and IPD nested F-test statistics and p-values
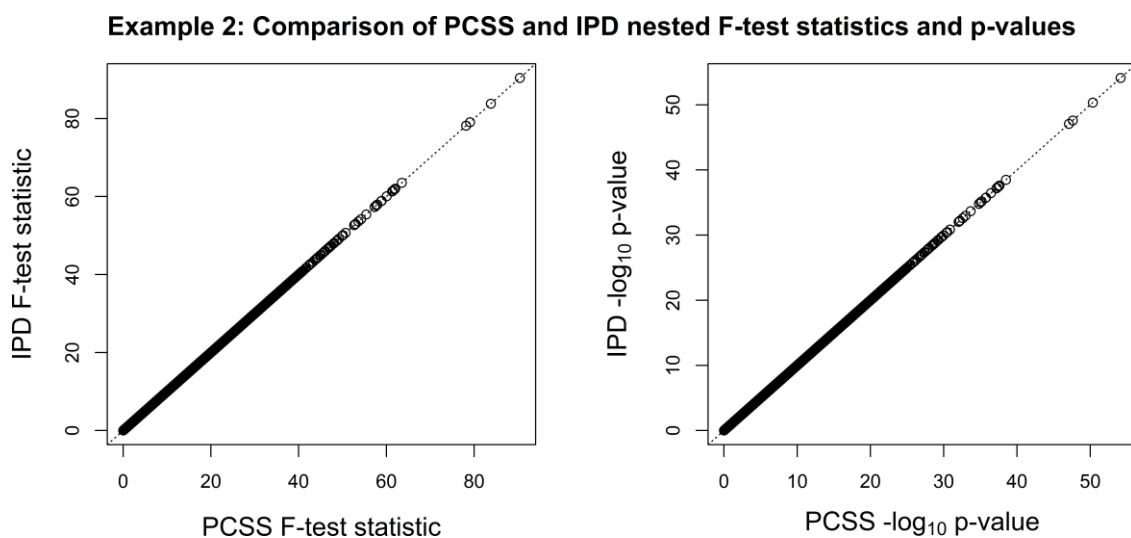
TABLE 2 The accuracy of our methods to estimate the rare variant test statistics for SKAT-O, SKAT, and burden tests using PCSS are depicted below. Errors were minimal with low variance in all cases.

Model (No Age)	Full Model (Age) Omnibus Statistic	Mean Error	Reduced Model (No Age) Variance
SKAT-O Score Statistic	7.84×10^{-14} 4.05×10^{-22}	4.40×10^{-22}	-1.62×10^{-14}
SKAT-O P-Value	3.23×10^{-18} 1.96×10^{-31}	2.87×10^{-31}	3.01×10^{-18}
SKAT Score Statistic	6.23×10^{-14} 1.40×10^{-22}	2.08×10^{-22}	4.33×10^{-14}
SKAT P-Value	3.56×10^{-18} 1.90×10^{-31}	2.44×10^{-31}	8.46×10^{-19}
Burden Score Statistic	1.09×10^{-13} 6.78×10^{-22}	5.63×10^{-22}	-5.11×10^{-14}
Burden P-Value	-7.02×10^{-19} 5.49×10^{-31}	7.70×10^{-31}	1.60×10^{-18}

FIGURE 4 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the SKAT-O test for the full model, featuring three SNVs, age, and sex. These figures illustrate our method's high accuracy on simulated data.

Full model: Comparison of PCSS and IPD SKAT-O score statistics and p-values

FIGURE 5 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the SKAT-O test for the reduced model, featuring three SNVs and sex. Thesefigures illustrate our method's high accuracy on simulated data.

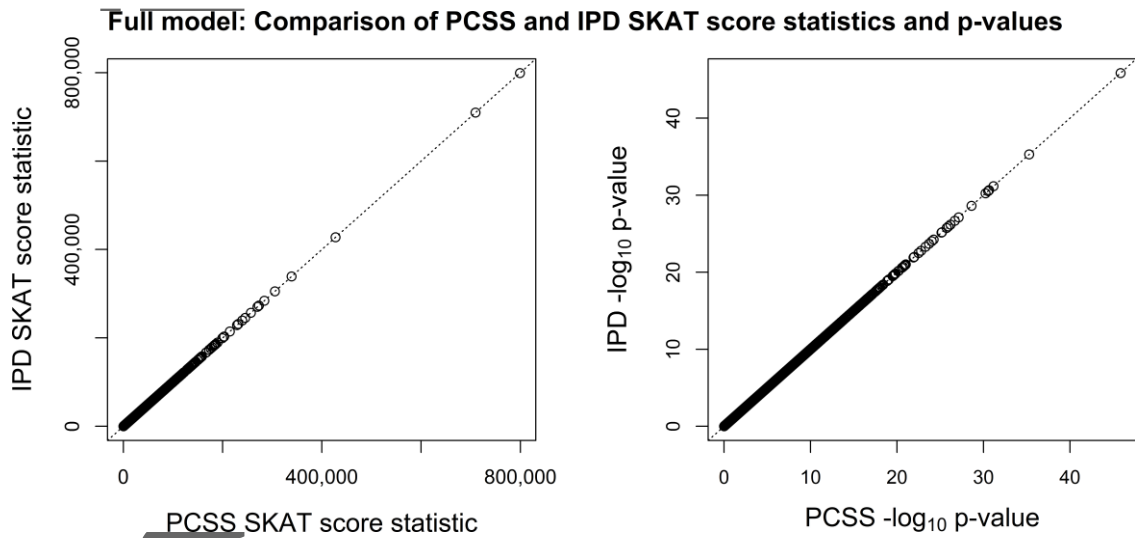**Reduced model: Comparison of PCSS and IPD SKAT-O score statistics and p-values**
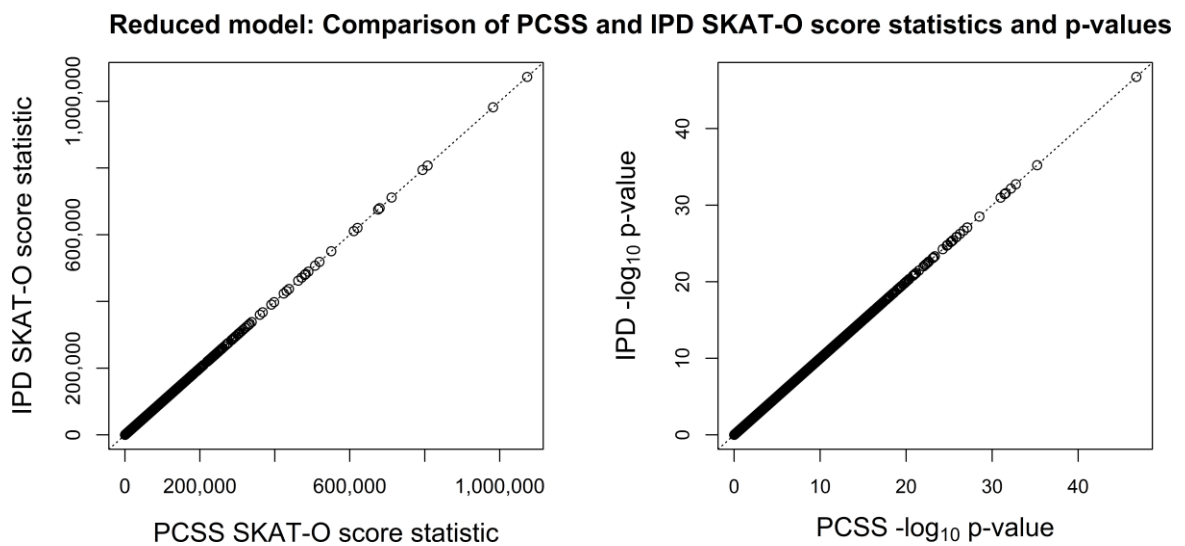
FIGURE 6 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the SKAT test for the full model, featuring three SNVs, age, and sex. Thesefigures illustrate our method's high accuracy on simulated data.
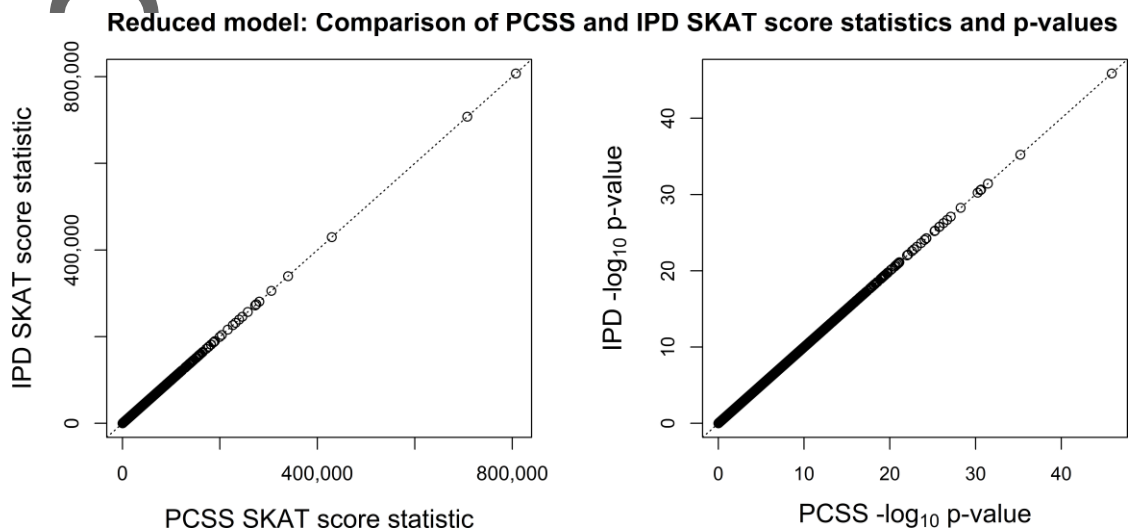

Full model: Comparison of PCSS and IPD SKAT score statistics and p-values

FIGURE 7 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the SKAT test for the reduced model, featuring three SNVs and sex. Thesefigures illustrate our method's high accuracy on simulated data.

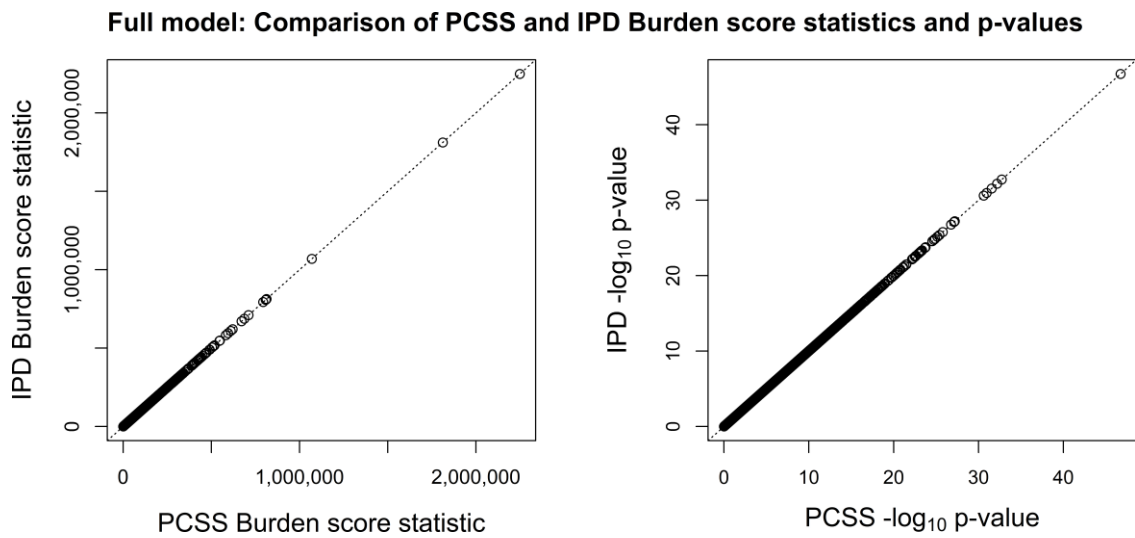Reduced model: Comparison of PCSS and IPD SKAT score statistics and p-values

FIGURE 8 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the Burden test for the full model, featuring three SNVs, age, and sex. Thesefigures illustrate our method's high accuracy on simulated data.

**Full model: Comparison of PCSS and IPD Burden score statistics and p-values**

FIGURE 9 The below plots depict the difference between score statistics (left) and log transformed *p*-values (right) on our method using PCSS compared to IPD as applied to the Burden test for the reduced model, featuring three SNVs and sex. Thesefigures illustrate our method's high accuracy on simulated data.

Reduced model: Comparison of PCSS and IPD Burden score statistics and p-values
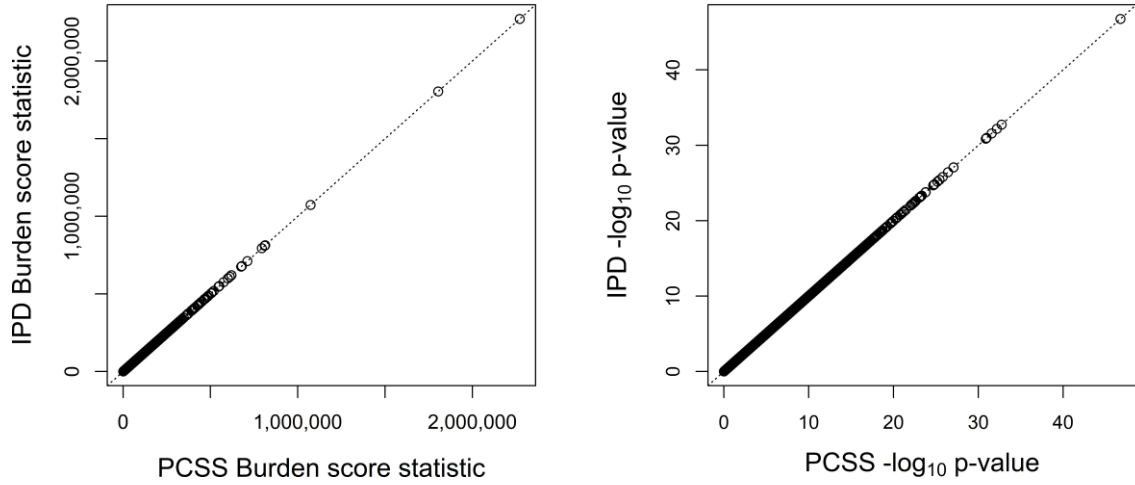
APPENDIX

A SSR DERIVATION

For a linear model of the form $\widehat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$ fit via least squares, we can simplify the sum of squared residuals such that it can be computed using only summary statistics by the following:

$$SSR = (\boldsymbol{y} - \boldsymbol{X}\hat{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})$$

$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}\hat{\beta} + \hat{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\beta$$

$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$
$$+ \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}^T\boldsymbol{y}$$

$$= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

$$= (\boldsymbol{y}^T\boldsymbol{y}) - (\boldsymbol{X}^T\boldsymbol{y})^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

B S_j DERIVATION

Given a null model $\boldsymbol{y} = \boldsymbol{Z\alpha} + \boldsymbol{\varepsilon}$ and least squares estimate $\hat{\boldsymbol{y}} = \boldsymbol{Z\hat{\alpha}}$, the contribution to the score statistic from any one SNV,

$\boldsymbol{g}_j$, can be expressed as: [2]

$$S_j = \boldsymbol{g}_j^T (\boldsymbol{y} - \hat{\boldsymbol{y}})/\hat{\sigma}$$

$$= \sum_{i=1}^{n} g_{ij}(y_i - \hat{y})/\hat{\sigma}$$

$$= \left((n-1)\mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y} - \hat{\boldsymbol{y}}) + \bar{g}_j \sum_{i=1}^{n} y_i - \hat{y}_i\right)/\hat{\sigma}$$

$$= (n-1)\mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y} - \hat{\boldsymbol{y}})/\hat{\sigma}$$

$$= (n-1)\left(\mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y}) - \mathrm{cov}(\boldsymbol{g}_j, \hat{\boldsymbol{y}})\right)/\hat{\sigma}$$

$$= (n-1)\left(\mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{y}) - \sum_{l=1}^{c} \hat{\alpha}_l \mathrm{cov}(\boldsymbol{g}_j, \boldsymbol{z}_l)\right)/\hat{\sigma}$$

C REAL DATA RESULTS

TABLE C1 The accuracy of our methods to perform nested F-Tests using PCSS from the Framingham Heart Study.

		With Covariate Adjustment (Age)				Without Covariate Adjustment (No Age)	
Gene SNVs	Num SNVs	F Stat	P-Value	Bias F†	Bias P-Value†	F Stat	P-Value
ADRA1D	51	0.664	0.967	-2.08×10^{-10}	1.01×10^{-10}	0.638	0.978
AHI1	74	1.364	0.025	-1.24×10^{-11}	3.41×10^{-12}	1.501	0.005
AMIGO2	4	0.987	0.413	-2.34×10^{-11}	1.28×10^{-11}	1.083	0.363
CCDC141	219	1.045	0.329	5.86×10^{-12}	-1.94×10^{-11}	1.058	0.287
CD96	70	1.007	0.464	-2.06×10^{-11}	4.67×10^{-11}	1.164	0.174
CR1L	77	0.853	0.811	-6.31×10^{-12}	1.15×10^{-11}	0.878	0.763
DSPP	68	1.090	0.293	-2.83×10^{-11}	5.18×10^{-11}	1.088	0.297
ELOVL1	3	0.474	0.700	1.55×10^{-11}	-1.09×10^{-11}	0.626	0.598
ELOVL2	27	1.569	0.032	1.58×10^{-11}	-3.07×10^{-12}	1.621	0.024
EPHA2	37	0.972	0.518	-7.15×10^{-12}	1.23×10^{-11}	1.044	0.398
FADS1	3	4.292	0.005	-3.59×10^{-11}	2.51×10^{-13}	4.550	0.004
FADS2	63	1.223	0.117	-1.44×10^{-11}	1.35×10^{-11}	1.261	0.086
FADS3	21	1.479	0.075	-1.11×10^{-11}	3.88×10^{-12}	1.646	0.033
FFAR4	24	0.619	0.924	8.86×10^{-12}	-6.05×10^{-12}	0.703	0.852
FILNC1	81	1.159	0.165	3.54×10^{-11}	-4.85×10^{-11}	1.202	0.114

Gene	Num SNVs	F Stat	P-Value	Bias	Bias	F Stat	P-Value
LRRC3B	92	1.208	0.096	1.35×10^{-11}	-1.31×10^{-11}	1.257	0.057
ME1	111	1.023	0.421	3.20×10^{-11}	-8.74×10^{-11}	1.021	0.427
POLR1D	47	1.145	0.236	-5.45×10^{-11}	7.26×10^{-11}	1.141	0.241
PTGS2	7	1.697	0.106	2.57×10^{-11}	-6.14×10^{-12}	1.756	0.093
SCFD1	63	1.104	0.273	-1.03×10^{-11}	1.75×10^{-11}	1.077	0.322
TMEM258	2	2.239	0.107	-1.21×10^{-10}	1.29×10^{-11}	2.657	0.071
WDR70	343	0.901	0.872	-1.13×10^{-10}	2.88×10^{-10}	0.880	0.917

[†]Bias was calculated as the difference of the PCSS estimate and the individual patient-level data (IPD) estimate

TABLE C2 The accuracy of our methods to estimate SKAT-O test statistics using PCSS from the Framingham Heart Study. SKAT-O test statistics have been scaled by

10^8 for readability.

	With Covariate Adjustment (Age)			Without Covariate Adjustment (No Age)		
Gene	Num SNVs	F Stat	P-Value	Bias F[‡]	Bias P-Value[‡]	F Stat
	P-Value	Bias F1	Bias P-Value1			
ADRA1D	58	64.49	0.196	3.89×10^{-04}	-6.99×10^{-15}	
	69.04	0.172	-1.70×10^{-04}	3.00×10^{-15}		
AHI1	419	95.40	0.233	5.89×10^{-04}	-2.00×10^{-14}	
	100.88	0.176	-2.00×10^{-04}	6.00×10^{-15}		
AMIGO2	3	0.43	0.368	2.40×10^{-06}	-6.00×10^{-15}	
	0.40	0.375	-7.97×10^{-07}	2.00×10^{-15}		
CCDC141	328	57.80	0.437	3.40×10^{-04}	-4.80×10^{-14}	
	52.61	0.527	-1.10×10^{-04}	1.70×10^{-14}		
CD96	110	3.22	0.891	5.30×10^{-05}	-7.99×10^{-15}	
	58.61	0.553	-1.39×10^{-04}	3.00×10^{-15}		
CR1L	182	24.84	0.477	1.50×10^{-04}	-2.30×10^{-14}	
	26.28	0.407	-5.01×10^{-05}	6.00×10^{-15}		
DSPP	55	0.84	0.837	-3.40×10^{-06}	6.00×10^{-15}	
	0.38	0.889	-6.03×10^{-07}		9.99×10^{-16}	
ELOVL1	5	0.57	0.417	2.60×10^{-06}	-3.00×10^{-15}	
	0.89	0.429	-1.80×10^{-06}	6.99×10^{-15}		

Gene					
ELOVL2	67	14.92	0.203	9.01×10^{-05}	-3.90×10^{-14}
		13.84	0.251	-3.00×10^{-05}	1.30×10^{-14}
ELOVL3	3	0.40	0.521	2.50×10^{-06}	-1.40×10^{-14}
		0.52	0.390	-1.30×10^{-06}	6.99×10^{-15}
EPHA2	51	13.91	0.056	8.01×10^{-05}	-1.13×10^{-14}
		15.07	0.041	-3.00×10^{-05}	2.60×10^{-15}
FADS1	12	1.85	0.569	1.10×10^{-05}	-2.20×10^{-14}
		1.05	0.525	-2.00×10^{-06}	9.99×10^{-16}
FADS2	89	16.35	0.276	1.00×10^{-04}	-2.80×10^{-14}
		18.22	0.171	-4.01×10^{-05}	6.99×10^{-15}
FADS3	10	3.42	0.116	2.10×10^{-05}	-9.99×10^{-15}
		3.04	0.073	-6.97×10^{-06}	3.79×10^{-15}
FFAR4	41	16.73	0.511	7.01×10^{-05}	-2.00×10^{-15}
		6.00	0.562	-1.20×10^{-05}	7.99×10^{-15}
FILNC1	98	21.16	0.173	1.30×10^{-04}	-2.50×10^{-14}
		20.54	0.165	-5.01×10^{-05}	7.99×10^{-15}
LRRC3B	176	966.21	0.119	6.90×10^{-03}	-6.99×10^{-15}
		1270.36	0.068	-2.99×10^{-03}	1.80×10^{-15}
ME1	316	58.06	0.450	3.50×10^{-04}	-3.50×10^{-14}
		54.61	0.477	-1.11×10^{-04}	1.40×10^{-14}
POLR1D	54	26.39	0.062	1.40×10^{-04}	-4.91×10^{-15}
		59.55	0.091	-1.20×10^{-04}	2.19×10^{-15}
PTGS2	10	40.39	0.023	2.40×10^{-04}	-2.30×10^{-15}
		44.28	0.016	-8.96×10^{-05}	5.00×10^{-16}
SCFD1	270	90.74	0.061	5.19×10^{-04}	-7.00×10^{-15}
		75.25	0.106	-1.50×10^{-04}	4.00×10^{-15}
TMEM258	3	0.19	0.624	1.70×10^{-06}	-9.99×10^{-15}
		0.70	0.339	-1.19×10^{-06}	9.99×10^{-16}
WDR70	455	18.01	0.838	1.30×10^{-04}	-3.00×10^{-15}
		70.15	0.720	-1.40×10^{-04}	2.70×10^{-14}

[‡]Bias was calculated as the difference of the PCSS estimate and the individual patient-level data (IPD) estimate.