Biometrics 73, 876–884
September 2017

DOI: 10.1111/biom.12643

# Sequence Robust Association Test for Familial Data

**Wei Dai,[1,†] Ming Yang,[1,†] Chaolong Wang,[2] and Tianxi Cai[1,*]**

[1]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, Massachusetts, U.S.A.
[2]Computational and Systems Biology, Genome Institute of Singapore, Singapore
[†]Wei Dai and Ming Yang contributed equally to this work.
[*]*email:* tcai@hsph.harvard.edu

Summary. Genome-wide association studies (GWAS) and next generation sequencing studies (NGSS) are often performed in family studies to improve power in identifying genetic variants that are associated with clinical phenotypes. Efficient analysis of genome-wide studies with familial data is challenging due to the difficulty in modeling shared but unmeasured genetic and/or environmental factors that cause dependencies among family members. Existing genetic association testing procedures for family studies largely rely on generalized estimating equations (GEE) or linear mixed-effects (LME) models. These procedures may fail to properly control for type I errors when the imposed model assumptions fail. In this article, we propose the Sequence Robust Association Test (SRAT), a fully rank-based, flexible approach that tests for association between a set of genetic variants and an outcome, while accounting for within-family correlation and adjusting for covariates. Comparing to existing methods, SRAT has the advantages of allowing for unknown correlation structures and weaker assumptions about the outcome distribution. We provide theoretical justifications for SRAT and show that SRAT includes the well-known Wilcoxon rank sum test as a special case. Extensive simulation studies suggest that SRAT provides better protection against type I error rate inflation, and could be much more powerful for settings with skewed outcome distribution than existing methods. For illustration, we also apply SRAT to the familial data from the Framingham Heart Study and Offspring Study to examine the association between an inflammatory marker and a few sets of genetic variants.

Key words: Genetic association testing; Marginal model; Non-parametric transformation model; Perturbation resampling; Robust procedures.

## 1. Introduction

Genome-wide association studies (GWAS) and next generation sequencing studies (NGSS) have emerged as popular tools for identifying genetic variants associated with complex phenotypes (McCarthy et al., 2008; Koboldt et al., 2013). Such studies have successfully identified variants associated with susceptibility to diseases such as breast cancer (Easton et al., 2007; Hunter et al., 2007), prostate cancer (Gudmundsson et al., 2007; Thomas et al., 2008), and type II diabetes (Scott et al., 2007; Sladek et al., 2007). Standard analysis of genome-wide studies typically employs individual-marker based testing procedures (Nyholt, 2004; Lin, 2005; Moskvina and Schmidt, 2008; Gao et al., 2010). However, when multiple variants are related to the phenotype simultaneously, a single marker analysis may have low statistical power and ultimately may not prove effective (Vo et al., 2007).

To improve statistical power, a wide range of marker-set based testing procedures have been proposed and widely used in the statistical genetics community. A good review of these methods can be found in Lee et al. (2014). For example, Wu et al. (2010,2011) proposed the Sequence Kernel Association Test (SKAT), based on a variance component score statistic, for continuous and discrete phenotypes. These variance component tests improve power by both leveraging the correlation among the markers and combining signals from multiple markers. Additionally, SKAT and some related methods can be used to detect signals from rare variants by assigning larger weights to variants with lower minor allele frequency (MAF).

Although traditional genetic studies often collect data from unrelated individuals, datasets from families have become increasingly available in recent years(Chen et al., 2016). Family studies are potentially more powerful for gene discovery than standard population-based studies, since family members serve as better controls for each other due to their shared genetic background and environmental exposures. However, genomic association analysis of familial data is challenging due to complex yet unknown within-family dependencies among the outcomes. To test for genomic associations with familial data, SKAT has recently been extended to incorporate within family correlation by using generalized estimating equations (GEE-SKAT) for continuous/binary outcomes (Wang et al., 2013), linear mixed-effects (LME-SKAT) models for continuous outcomes (Schifano et al., 2012; Chen et al., 2013). Recently,Chen et al. (2016) proposed logistic mixed model-based testing for single variants with binary outcomes. All these methods make strong model assumptions regarding the outcome distribution and/or correlation structure between family members. However, in practice, it is difficult, if not impossible, to model outcome distributions or their within-family correlation structure accurately, particularly in the presence of unobservable shared environmental and genetic factors. Examples of such unobserved shared factors include dietary intake patterns, physical activity habits,

as well as genetic information not captured by the observed variant such as insertion–deletion polymorphism or untyped single nucleotide polymorphisms (SNPs). Mis-specification in the correlation structure or outcome distribution may result in incorrect type I error rate for both GEE-SKAT and LME-SKAT. In addition, all these methods are sensitive to outliers or skewness in the outcome distribution.

To overcome these challenges, we propose a flexible rank regression-based marker-set association testing procedure, named Sequence Robust Association Test (SRAT), which allows for continuous, binary, or ordinal outcomes while accounting for within family correlation and adjusting for potential covariates. SRAT is derived under a marginal nonparametric transformation (NPT) model with robust variance calculation, which adjusts for correlation without necessitating the specification of error distributions or underlying correlation structure. Since SRAT is fully rank-based, the $p$-value from SRAT is invariant to monotone transformation of the response variable.

The remainder of this manuscript is organized as follows. In Section 2, we describe the proposed testing procedure and also demonstrate that SRAT includes the well-known Wilcoxon rank sum test as a special case. We present simulation results in Section 3, comparing our approach to existing methods under a variety of settings. In Section 4, we apply SRAT to a familial genetic dataset from the Framingham Heart Study and Offspring Study to examine the associations between the C-reactive protein (CRP) level and a list of candidate genes. We conclude with brief discussions in Section 5.

## 2. Methods

Our goal is to test for the association between phenotype $Y$ and a p-dimensional genomic marker set $\mathbf{G} = (G_1, G_2, \ldots, G_p)^\top$ adjusting for some potential covariates $\mathbf{X} = (X_1, X_2, \ldots, X_q)^\top$ with familial data, where $Y$ is allowed to be continuous, binary or ordinal. For the $j$th subject in the $i$th family, we observe $\mathbf{D}_{ij} = \{Y_{ij}, \mathbf{X}_{ij} = (X_{ij,1}, X_{ij,2}, \ldots, X_{ij,q})^\top, \mathbf{G}_{ij} = (G_{ij,1}, G_{ij,2}, \ldots, G_{ij,p})^\top\}$ and assume that data from $n$ families are available for analysis with complete data consisting of $\mathscr{D} = \{\mathbf{D}_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, n\}$ from $N = \sum_{i=1}^{n} m_i$ individuals, where $m_i$ is the number of available subjects in the $i$th family.

### 2.1. *NPT Model*

To provide a unified framework for phenotypes of various type, we let $Y_{ij}^*$ denote the underlying continuous phenotype such that $Y_{ij} = Y_{ij}^*$ when $Y$ is continuous; $Y_{ij} = I(Y_{ij}^* \geq 1)$ when $Y$ is binary, and $Y_{ij} = \sum_{k=1}^{K} I(Y_{ij}^* \geq k)$ for ordinal outcomes taking values of 1, ..., $K$. Thus for a binary or ordinal outcome $Y$, $Y_{ij}$ can be viewed as a thresholded version of the underlying continuous phenotype $Y_{ij}^*$. Such an approach to modeling binary or ordinal outcomes has been adopted in genetic literature (McIntosh et al., 2006; Zuk et al., 2012). To relate $\mathbf{G}_{ij}$ and $\mathbf{X}_{ij}$ to $Y_{ij}$, we consider a marginal NPT model

$$H(Y_{ij}^*) = \boldsymbol{\alpha}^\top \mathbf{X}_{ij} + \boldsymbol{\beta}^\top \mathbf{G}_{ij} + \epsilon_{ij},\ j = 1, \ldots, m_i,\ i = 1, \ldots, n, \quad (1)$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_q)^\top$ is the unknown vector of covariate effects, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ is the unknown effect vector of the genomic markers, the errors $\{\epsilon_{ij}\}$ are independent of $\{\mathbf{X}_{ij}, \mathbf{G}_{ij}, j = 1, \ldots, m_i, i = 1, \ldots, n\}$ but potentially correlated within family with a common unknown marginal distribution $g(\cdot)$, and $H(\cdot)$ is unspecified but assumed to be smooth and strictly increasing. Since both $g(\cdot)$ and $H(\cdot)$ are unspecified, $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\beta}^\top)^\top$ is only identifiable up to a scalar. From model (1), we have

$$\begin{aligned} P(Y_{ij}^* \geq y \mid \mathbf{D_{ij}}) &= P\{H(Y_{ij}^*) \geq H(y) \mid \mathbf{D_{ij}}\} \\ &= \bar{g}\{H(y) - \boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\beta}^\top \mathbf{G}_{ij}\}, \end{aligned}$$

where $\bar{g} = 1 - g$. Thus, $P(Y_{ij} = 1 \mid \mathbf{D_{ij}}) = P(H(Y_{ij}^*) \geq H(1) \mid \mathbf{D_{ij}}) = \bar{g}\{H(1) - \boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\beta}^\top \mathbf{G}_{ij}\}$ with $H(y)$ only defined at $y = 1$ for binary outcomes; and $P(Y_{ij} \geq k \mid \mathbf{D_{ij}}) = P(H(Y_{ij}^*) \geq H(k) \mid \mathbf{D_{ij}}) = \bar{g}\{H(k) - \boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\beta}^\top \mathbf{G}_{ij}\}$ with $H(y)$ only defined at $y \in \{1, \ldots, K\}$ for ordinal outcomes. The marginal model (1) can also be motivated by an NPT mixed-effects model:

$$H(Y_{ij}^*) = \boldsymbol{\alpha}^\top \mathbf{X}_{ij} + \boldsymbol{\beta}^\top \mathbf{G}_{ij} + a_{ij} + \mathcal{E}_{ij},\ j = 1, \ldots, m_i,\ i = 1, \ldots, n, \quad (2)$$

where $\{\mathcal{E}_{ij}\}$ are independent and identically distributed (iid) errors with an unknown common distribution $F_e(\cdot)$, and the random effect $a_{ij}$ captures the effect of unobservable genetic and environmental factors shared by the $i$th family, and is assumed independent of $\mathbf{G}_{ij}$, $\mathbf{X}_{ij}$, and $\mathcal{E}_{ij}$ with a common but unknown marginal distribution function $F_A(.)$. Then (2) can be viewed as a special case of (1) with error distribution $g(x) = \int F_e(x + a) dF_A(a)$.

When $m_i = 1$, (1) reduces to the NPT model studied in Han (1987). For this uncorrelated case, the maximum rank correlation estimator, $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta} \in \Omega} L(\boldsymbol{\alpha}, \boldsymbol{\beta})$, was shown to be consistent and asymptotically normal under mild regularity conditions (Sherman, 1993), where

$$\begin{aligned} L(\boldsymbol{\alpha}, \boldsymbol{\beta}) = n^{-2} \sum_{i,j,i',j'} & I(Y_{ij} > Y_{i'j'}) \\ & \times I(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} + \boldsymbol{\beta}^\top \mathbf{G}_{ij} > \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'} + \boldsymbol{\beta}^\top \mathbf{G}_{i'j'}), \end{aligned} \quad (3)$$

the maximization is constrained within the parameter space $\Omega = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_1 = 1\}$ since $\boldsymbol{\theta}$ is only identifiable up to a scalar, and $\|\cdot\|_d$ denotes the $L_d$ norm. Here and throughout, we let $\sum_{i,j,i',j'} = \sum_{i=1}^{n} \sum_{j=1}^{m_i} \sum_{i'=1}^{n} \sum_{j'=1}^{m_{i'}}$, $\sum_{i,j} = \sum_{i=1}^{n} \sum_{j=1}^{m_i}$, and $\sum_{i',j'} = \sum_{i'=1}^{n} \sum_{j'=1}^{m_i}$ for notational conciseness. In the presence of correlation, we mimic the GEE approach by imposing an independence working assumption and continue to employ $\widehat{\boldsymbol{\theta}}$ to estimate $\boldsymbol{\theta}$.

### 2.2. *Test Statistic of SRAT*

Under model (1), evaluating whether $\mathbf{G}$ influence $Y$ after adjusting for $\mathbf{X}$, corresponds to testing the null hypothesis $H_0 : \boldsymbol{\beta} = 0$, that is, $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. To test $H_0$, we mimic the variance component tests and impose a working assumption that $\beta_1, \beta_2, \ldots, \beta_p$ follow some common distribution with mean 0 and variance $\tau^2$. Then testing $H_0 : \boldsymbol{\beta} = \mathbf{0}$ is equivalent to testing $H_0 : \tau = 0$, which can be performed

based on a pseudo score statistic corresponding to (3). Specifically, to derive the pseudo-score statistic, we first consider a smoothed version of $L(\boldsymbol{\alpha}, \boldsymbol{\beta})$, namely

$$L_{sm}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = n^{-2} \sum_{i,j,i',j'} I(Y_{ij} > Y_{i'j'}) \mathbb{K}_h \times \{\boldsymbol{\alpha}^\top (\mathbf{X}_{ij} - \mathbf{X}_{i'j'}) + \boldsymbol{\beta}^\top (\mathbf{G}_{ij} - \mathbf{G}_{i'j'})\}, \quad (4)$$

as in Ma and Huang (2007), where $\mathbb{K}_h(x) = \mathbb{K}(x/h)$, $\mathbb{K}$ is a smooth cumulative distribution function with $\mathbb{K}(x) = 1 - \mathbb{K}(-x)$, and $h$ is a bandwidth of order $n^{-\nu}$ with $\nu \in [1/5, 1/3)$ which ensures the consistency of both the score function and its corresponding derivative functions (Pagan and Ullah, 1999). Then for a given $\boldsymbol{\alpha}$ and writing $\boldsymbol{\beta} = \tau \mathbf{b}$ where $b_1, \ldots, b_p$ are with iid with zero mean and unit variance, we may derive a test statistic for $H_0 : \tau = 0$ as

$$Q(\boldsymbol{\alpha}) = E[\{\partial L_{sm}(\boldsymbol{\alpha}, \tau \mathbf{b})/\partial \tau\}^2 \mid \mathscr{D} = \mathbf{S}(\boldsymbol{\alpha})^\top \mathbf{S}(\boldsymbol{\alpha}),$$

where

$$\mathbf{S}(\boldsymbol{\alpha}) = n^{-2} \sum_{i,j,i',j'} I(Y_{ij} > Y_{i'j'})(\mathbf{G}_{ij} - \mathbf{G}_{i'j'}) K_h(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'}) = n^{-1} \sum_{i,j} R_{ij}(\boldsymbol{\alpha}) \mathbf{G}_{ij}. \quad (5)$$

$R_{ij}(\boldsymbol{\alpha}) = n^{-1} \sum_{i',j'} \text{sign}(Y_{ij} - Y_{i'j'}) K_h(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'})$ and $K_h(x) = \partial \mathbb{K}_h(x)/\partial x$. The intuition behind $\mathbf{S}(\boldsymbol{\alpha})$ being a valid pseudo-score vector is that, under $H_0$, $Y_{ij}$ and $\mathbf{G}_{ij}$ are independent given $\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}$ and thus $E\{I(Y_{ij} > Y_{i'j'})(\mathbf{G}_{ij} - \mathbf{G}_{i'j'}) \mid \boldsymbol{\alpha}_0^\top \mathbf{X}_{ij} = \boldsymbol{\alpha}_0^\top \mathbf{X}_{i'j'}\} = \mathbf{0}$, where $\boldsymbol{\alpha}_0$ is the true value of $\boldsymbol{\alpha}$. One can view $\mathbf{S}(\boldsymbol{\alpha}_0)$ as a scaled estimator of this expectation, which should be close to $\mathbf{0}$ under $H_0$. Since $\boldsymbol{\alpha}_0$ is unknown, we estimate it under the null as $\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \mathbf{0})$ and construct the final test statistic as $\widehat{Q} = n \|\mathbf{S}(\widehat{\boldsymbol{\alpha}})\|^2 = \hat{\mathbb{R}}^\top (\mathbb{G}\mathbb{G}^\top) \hat{\mathbb{R}}$, where $\|\cdot\|$ denotes the $L_2$ norm, $\hat{\mathbb{R}}_{N\times 1} = [R_{ij}(\widehat{\boldsymbol{\alpha}})]_{i=1,\ldots,n; j=1,\ldots,m_i}$ and $\mathbb{G}_{N\times p} = [\mathbf{G}_{ij}]_{i=1,\ldots,n; j=1,\ldots,m_i}$. Similar to GEE-SKAT and LME-SKAT, we can easily incorporate weights in the test statistic by replacing $\mathbb{G}\mathbb{G}^\top$ in $\widehat{Q}$ with $\mathbb{G}\mathbb{W}\mathbb{G}^\top$, where $\mathbb{W} = \text{diag}\{w_1, \ldots, w_p\}$ and $w_k$ is a pre-specified weight based on prior knowledge. For example, $w_k$ can be chosen as a function of MAF to upweight rare variants (Wu et al., 2011). For simplicity of the presentation, we focus on the unweighted statistic.

In the absence of covariates, the score vector (5) reduces to

$$n^{-1} \sum_{i,j} \left\{ n^{-1} \sum_{i',j'} \text{sign}(Y_{ij} - Y_{i'j'}) \right\} \mathbf{G}_{ij} \propto N^{-1} \sum_{i,j} (R_{ij} - \bar{R}) \mathbf{G}_{ij},$$

where $\bar{R} = N^{-1} \sum_{i=1}^{n} \sum_{j=1}^{m_i} R_{ij}$, and $R_{ij}$ is the rank of $Y_{ij}$ among $\{Y_{i'j'}, j' = 1, \ldots, m_{i'}, i' = 1, \ldots, n\}$. When all participants are unrelated (i.e., $m_i = 1$) and $G_{ij}$ is a binary variable, $S(\boldsymbol{\alpha})$ reduces to the Mann–Whitney U statistic and thus SRAT includes the Wilcoxon rank sum test as a special case.

### 2.3. *Perturbation Resampling*

In the Appendix, we show that $\mathbf{S}(\widehat{\boldsymbol{\alpha}}) \to \mathbf{0}$ in probability under $H_0$. Using similar arguments as given in the Appendix and Sherman (1993), one may also show that $\sqrt{n}\mathbf{S}(\widehat{\boldsymbol{\alpha}})$ is asymptotically normal with mean $\mathbf{0}$ and some covariance matrix $\boldsymbol{\Sigma}$ under $H_0$. Hence, the test statistic $\widehat{Q}$ follows a mixture of $\chi^2$ distribution with mixing constants being the eigenvalues of $\boldsymbol{\Sigma}$. In practice, empirically estimating $\boldsymbol{\Sigma}$ is challenging due to the unknown transformation function and error distribution. Here, we employ a perturbation resampling method to approximate the null distribution of $\sqrt{n}\mathbf{S}(\widehat{\boldsymbol{\alpha}})$ and $\widehat{Q}$ similar to the resampling procedure proposed in Jin et al. (2001). Specifically, let $\mathbf{V} = (V_1, \ldots, V_n)^\top$ be a vector of iid positive random variables that are independent of the data with $\text{E}(V_i) = 1$ and $\text{Var}(V_i) = 1$. We perturb the pseudo-score statistic $\mathbf{S}(\widehat{\boldsymbol{\alpha}})$ as $\mathbf{S}^*(\widehat{\boldsymbol{\alpha}}^*)$, where

$$\mathbf{S}^*(\boldsymbol{\alpha}) = \frac{1}{(\sum_{i=1}^{n} V_i)^2} \sum_{i,j,i',j'} V_i V_{i'} I(Y_{ij} > Y_{i'j'})(\mathbf{G}_{ij} - \mathbf{G}_{i'j'}) K_h \times (\boldsymbol{\alpha}^\top \mathbf{X}_{ij} - \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'})$$

and $\widehat{\boldsymbol{\alpha}}^* = \arg\max_{\boldsymbol{\alpha}} L^*(\boldsymbol{\alpha}, \mathbf{0})$, and $L^*(\boldsymbol{\alpha}, \mathbf{0}) = \sum_{i,j,i',j'} V_i V_{i'} I(Y_{ij} > Y_{i'j'}) I(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} > \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'})$.

Following similar arguments as in Jin et al. (2001) and Cai and Cheng (2008), one may show that the unconditional distribution of $\sqrt{n}\mathbf{S}(\widehat{\boldsymbol{\alpha}})$ can be approximated by the conditional distribution of $\sqrt{n}\left\{\mathbf{S}^*(\widehat{\boldsymbol{\alpha}}^*) - \mathbf{S}(\widehat{\boldsymbol{\alpha}})\right\}$ given the observed data $\mathscr{D}$. In practice, to obtain the above approximation, one may generate a large number, $\mathcal{B}$, of realizations of $\mathbf{V}$, and calculate $\mathbf{S}^*(\widehat{\boldsymbol{\alpha}}^*)$ for each realized sample. The null distribution of $\widehat{Q}$ can be approximated by the empirical distribution of $\widehat{Q}^* = n\|\mathbf{S}^*(\widehat{\boldsymbol{\alpha}}^*) - \mathbf{S}(\widehat{\boldsymbol{\alpha}})\|^2$. The $p$-value for testing $H_0$ may be directly obtained as $p^* = \mathcal{B}^{-1} \sum_{b=1}^{\mathcal{B}} I(\widehat{Q}_b^* > \widehat{Q})$, where $\widehat{Q}_b^*$ is the $b$th realization of $\widehat{Q}^*$. However, for very small $p$-values, a large of number of perturbations is required to obtain accurate $p$-values. Alternatively, one can first obtain the variance–covariance matrix estimate $\widehat{\boldsymbol{\Sigma}}$ through perturbation resampling and then apply algorithms (Liu et al., 2009; Kuonen, 1999) to approximate the distribution of $\|N(0, \widehat{\boldsymbol{\Sigma}})\|^2$. Specifically, a saddlepoint approximation can be used to approximate the distribution of $\|N(0, \widehat{\boldsymbol{\Sigma}})\|^2$ based on the eigenvalues of $\widehat{\boldsymbol{\Sigma}}$. Such an approach is computationally more feasible than calculating $p$-value based on $p^*$, since only a few hundreds repetitions of perturbation are needed to get a stable estimate of $\boldsymbol{\Sigma}$. Throughout, we use Kuonen's saddlepoint approximation method for our numerical studies due to its simplicity, stability, and accuracy.

## 3. Simulation Studies

We present results from extensive simulation studies evaluating the performance of SRAT and comparing it to existing methods with respect to empirical size and power under a variety of settings. We first describe the data generating process in Section 3.1 and then present some of the main results for continuous phenotype with multiple causal variants in

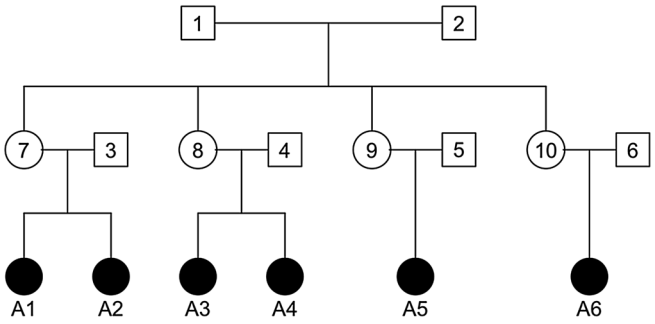


**Figure 1.** Family structure for simulations. The squares represent the founders and the circles represent subjects generated from the founders. The black circles represent subjects used in the analysis.

Section 3.2. Additional details on the simulation results for binary phenotype as well as for the case without covariates (Web Appendix A) and sensitivity analysis of SRAT on bandwidth (Web Appendix B) can be found in the Supplementary Materials.

### 3.1. *Family Structure and Data Generation*

We simulate the familial data on the genotypes and phenotypes under the family structure shown in Figure 1. Each family consists of six individuals of the same generation, coming from four nuclear families. Subjects from the same nuclear families are siblings, while subjects from different nuclear families are first-cousins. For each replication, we first generate complete underlying data for $n = 500$ families (i.e., $N = 3000$ subjects) and then randomly sample a subset of 2000 subjects to form $\mathcal{D}$ to allow for varying family sizes. Throughout, we use 1000 simulation replications to compute the empirical size and 500 replications to evaluate the empirical power.

Genotypes are generated based on the linkage disequilibrium structure of the ASAH1 gene. We use HAPGEN2 (Su et al., 2011) and the CEU sample of the International HapMap Project (Altschuler et al., 2005) to generate haplotypes for subjects 1–6 in Figure 1 at each of the 93 SNPs. Then haplotypes of subjects 7–10 are randomly generated from their parents, subjects 1 and 2. Subsequently, we generate the genotypes of the subjects in the study population (black circles in Figure 1) from their parents, respectively. Since the genomic regions are relatively small, recombination would be extremely rare and hence not considered in the creation of the offspring genotypes. Based on the Illumina 500 K platform, we included a total of $p = 14$ tagged SNPs to form the SNP-set. For comparisons, we also apply existing methods including LME-SKAT (for continuous $Y$ only), GEE-SKAT, and a minimum $p$-value approach that combines results from single variant tests, denoted by GEE-MinP, to evaluate the significance of the association. For GEE-MinP, we calculate the effective number of markers based on Gao et al. (2010). Two of these 14 SNPs have MAF lower than 5% (MAF = 0.6 and 3.0%). To improve power for detecting signals from rare variants, we also consider weighted versions of SRAT, GEE-SKAT, and LME-SKAT. Specifically, for weighted analysis, we use the Beta(1, 5) weights with $w_k = 5(1 - \mathrm{MAF}_k)^4$ to weight up SNPs with lower MAF. All existing methods can incorporate covariates by calculating the test statistics based on conditional residual under their corresponding null models.

Across all settings, we generated data using a random effects set up as in (2) with random effects $a_{ij} = \epsilon_{ij}^G + \epsilon_{ij}^E$, where the shared genetic factor $\boldsymbol{\epsilon}_i^G = (\epsilon_{i1}^G, \epsilon_{i2}^G, \ldots, \epsilon_{i6}^G)^\top$ is generated from $N(\mathbf{0}, \sigma_G^2 \boldsymbol{\Phi})$, $\boldsymbol{\Phi}$ is the $6 \times 6$ kinship matrix (Lange, 1997), and the environmental factor $\boldsymbol{\epsilon}_i^E = (Z_{i1}, Z_{i1}, Z_{i2}, Z_{i2}, Z_{i3}, Z_{i4})^\top$ with $Z_{ik}, k = 1, 2, 3, 4$ generated from independent $N(0, \sigma_E^2)$.

### 3.2. *Summary of Simulation Results*

We first considered continuous phenotype in the presence of covariates, where we generated $q = 4$ covariates with $X_{ij,1} \sim \mathrm{Bernoulli}(0.5)$, $X_{ij,2} \sim 2X_{ij,1} - 1 + \boldsymbol{\gamma}_2^\top \mathbf{G}_{ij} + N(0, 1)$, $X_{ij,3} \sim 2X_{ij,1} - 1 + \boldsymbol{\gamma}_3^\top \mathbf{G}_{ij} + N(0, 1)$, and $X_{ij,4} = X_{ij,2} X_{ij,3} + \boldsymbol{\gamma}_4^\top \mathbf{G}_{ij}$. We let $\boldsymbol{\gamma}_2 = \boldsymbol{\gamma}_3 = \boldsymbol{\gamma}_4 = \mathbf{0}$ for the setting with $\mathbf{X}$ independent of $\mathbf{G}$; and $\boldsymbol{\gamma}_2 = (-1, 0, 0, \ldots, 0)^\top$, $\boldsymbol{\gamma}_3 = (0, -1, 0, \ldots, 0)^\top$, $\boldsymbol{\gamma}_4 = (1, 1, 1, \ldots, 1)^\top/14$ for the case of $\mathbf{X}$ dependent on $\mathbf{G}$. Given $\mathbf{X}$ and $\mathbf{G}$, we then generated correlated $Y_{ij}$ from exponential (mean=$e^{\eta_{ij}}$), where $\eta_{ij} = \boldsymbol{\alpha}^\top \mathbf{X}_{ij} + \boldsymbol{\beta}^\top \mathbf{G}_{ij} + a_{ij}$, $\boldsymbol{\alpha} = (1.0, 0.0, 0.5, -0.5)^\top$, and $\beta_k = \tau |\log_{10}(\mathrm{MAF}_k)|$ with $\tau$=0 for the null model and $\tau = 0.01, 0.02, \ldots, 0.05$ for the alternative models.

Figure 2a shows the QQ plots of the $p$-values for all four methods with and without weighting when $\mathbf{X}$ and $\mathbf{G}$ are independent of each other. The empirical distribution of the $p$-values from SRAT matches the theoretical distribution. Both LME-SKAT and GEE-SKAT appear to have mild bias in estimating the null distribution of their test statistics due to model mis-specification. The type I errors of GEE-MinP are substantially inflated. As shown in Figure 2b, SRAT is substantially more powerful than existing methods, despite the fact that existing methods also have inflated type I errors. In Figure 3, we compare the empirical distribution of the $p$-value to its theoretical distribution under the null for the case when $\mathbf{X}$ is dependent on $\mathbf{G}$. In this setting, all three existing methods have substantially inflated type I errors due to model mis-specification while SRAT is able to attain desired type I error rates. These results demonstrate the robustness of SRAT and potential power gain over existing methods when the outcome has a skewed distribution.

For binary phenotype, we generated $(\mathbf{X}_{ij}, \mathbf{G}_{ij}, \eta_{ij})$ under the same models as described above and then generated $Y_{ij} \sim \mathrm{Bernoulli}(\pi_{ij})$, where $\pi_{ij} = 1 - \exp\{-\exp(\eta_{ij})\}$. Details of the results can be found in Figure 1 of the Supplementary Materials. In general, we find that when $\mathbf{X}$ is independent of $\mathbf{G}$, both SRAT and GEE-SKAT maintain type I errors well and even GEE-MinP attains reasonable size. Under the alternative, SRAT and GEE-SKAT achieve similar power while GEE-MinP has slightly lower power. It is not surprising that SRAT and GEE-SKAT attain similar power for the binary case since the rank and the actual outcome values are essentially the same. On the other hand, when $\mathbf{X}$ and $\mathbf{G}$ are correlated, SRAT again attains correct type I errors while both GEE-SKAT and GEE-MinP have substantially inflated type I errors similar to the case for continuous phenotype.

We also performed additional simulations under the simple setting in the absence of covariates for both continuous and

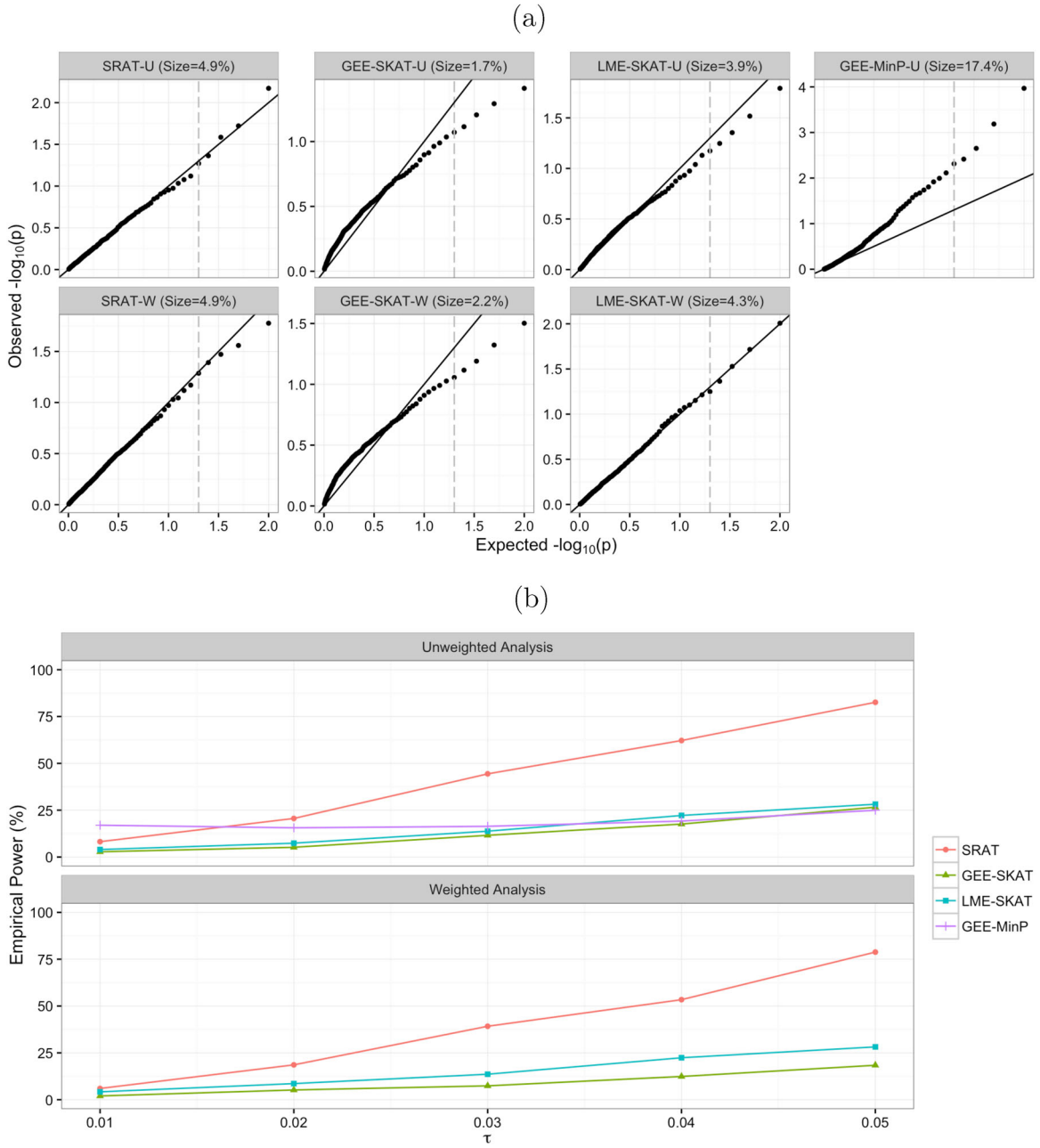


**Figure 2.** Continuous phenotype with uncorrelated **X** and **G**. (a) Empirical size (-U denotes unweighted analysis; -W denotes weighted analysis). (b) Empirical power.

binary phenotypes. We considered both the case with signals from multiple variants as specified above and with each of the SNP being the single causal variant at a time. Details of the simulation settings are summarized in Web Appendix A and results are shown in Figure 2 of the Supplementary Materials. All methods except for GEE-MinP with continuous outcome do a reasonable job in controlling type I error. For the multiple causal variants setting, the relative performance in terms of power has a similar pattern as described above. Under the setting with single causal variants, our method still outperforms other existing methods for continuous phenotype and performs similarly to GEE-SKAT and GEE-MinP for binary outcome.

## 4. Application

To illustrate our proposed method, we apply SRAT, GEE-SKAT, and LME-SKAT to data from the Framingham Offspring Study (FOS) (Feinleib et al., 1975). The FOS was initiated in 1971 and enrolled 5124 adult offspring and the spouses of participants in the Framingham Study (Kannel and McGee, 1979). These study participants have been followed over time to assess various clinical outcomes and

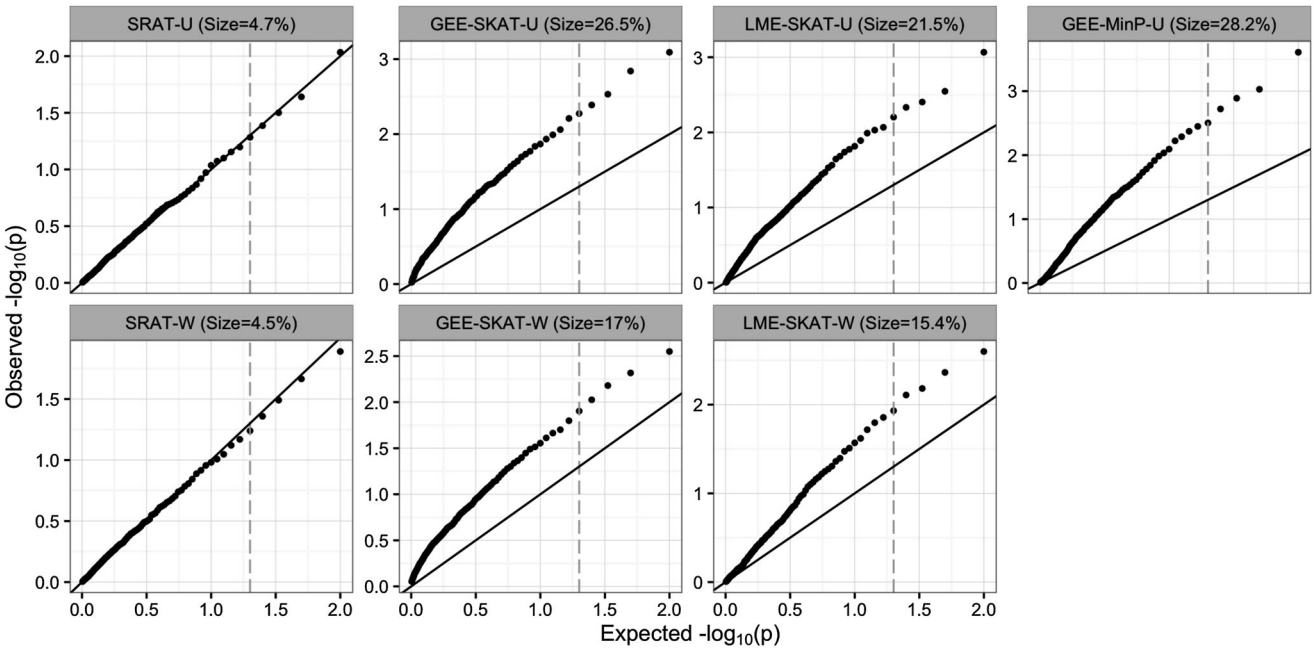


**Figure 3.** Continuous phenotype with correlated **X** and **G**. -U denotes unweighted analysis; -W denotes weighted analysis.

measure a wide range of biological and genetic markers. This study has provided valuable resources for investigating epidemiological and genetic risk factors of cardiovascular diseases (CVD). As an example, we applied the three SNP-set tests to identity genes that might be associated with C-reactive protein (CRP), an inflammatory marker that has been sought for use in CVD risk stratification and preventive decision making (Ridker et al., 2000). We extracted genotype information from the Framingham SNP Health Association Resource (SHARe) data through dbGaP (access number: phs000007.v3.p2), which contains 6923 individuals genotyped on the Affymetrix 500 K SNP array. We used the genotype data to create the full kinship matrix that is needed for implementing LME-SKAT.

After merging the genotype data from Framingham SHARe with the phenotype data from FOS, our final dataset consists of 2492 participants from FOS, including 1153 men and 1339 women who had data on both genotype and the CRP outcome. Since the synthesis of CRP is mainly regulated by interleukin-6 (IL6) and other inflammatory cytokines and is also produced locally in atherosclerotic lesions by smooth muscle cells (SMCs) lymphocytes and monocytic cells (Paffen and Moniek, 2006; Shrivastava et al., 2015), we considered a total of 11 candidate genes including IL6, interleukin-1beta (IL1B), Transforming Growth Factor Beta 2 (TGFB2), and other genes that relate to inflammatory cytokines and SMCs. For each gene, we perform testing based on SRAT, GEE-SKAT and LME-SKAT without weighting and with Beta(1, 5) weight. We also included results on the methods based on log-transformed CRP (logCRP) to illustrate the impact of transformation on GEE-SKAT and LME-SKAT. The analyses are conducted separately for men and women, adjusting for age and two leading principal components (PCs) accounting for population stratification. Due to the complicated family structure in FS and FOS, we used the LASER method (Wang et al., 2015) to identify ancestry PCs.

Figure 4 shows the results for the aforementioned 11 genes. For example, for the IL1B gene, we included six SNPs (rs315920, rs4251961, rs2637988, rs4251984, rs4251985, and rs928940) that are available from the Framingham SHARe data. The IL1B has been previously reported as associated with higher CRP levels in patients with CVD (Latkovskis et al., 2004). We found a significant association between CRP and the IL1B gene in men based on the unweighted SRAT ($p$-value = 0.045). However, this association is not detected using the GEE-SKAT ($p$-value = 0.866) or LME-SKAT ($p$-value = 0.920). The discrepancy in the $p$-values could in part be attributed to the skewness in the distribution of CRP. With logCRP, the $p$-values of GEE-SKAT and LME-SKAT become smaller but are still greater than the $p$-value of SRAT, which is invariant to any monotone transformation of the phenotype. For the IL6 gene ($p = 3$), LME-SKAT has a much smaller $p$-value (0.001) compared to the $p$-values of GEE-SKAT (0.027) and SRAT (0.053) in the unweighted analysis in men. However, the $p$-values of these three tests become very similar after the log transformation of the CRP outcome. For the TGFB2 gene ($p = 18$), SRAT detects a significant association in both unweighted (0.034) and weighted (0.010) analysis among women. However, this association was not detected by GEE-SKAT or LME-SKAT.

## 5. Discussion

One main advantage of the proposed SRAT procedure is its robustness, with key features being its scale invariance with respect to outcome and flexibility in incorporating the unknown correlation structure. Our numerical results suggest that our procedure is not only more robust and less susceptible to inflated type I errors due to mis-specification in

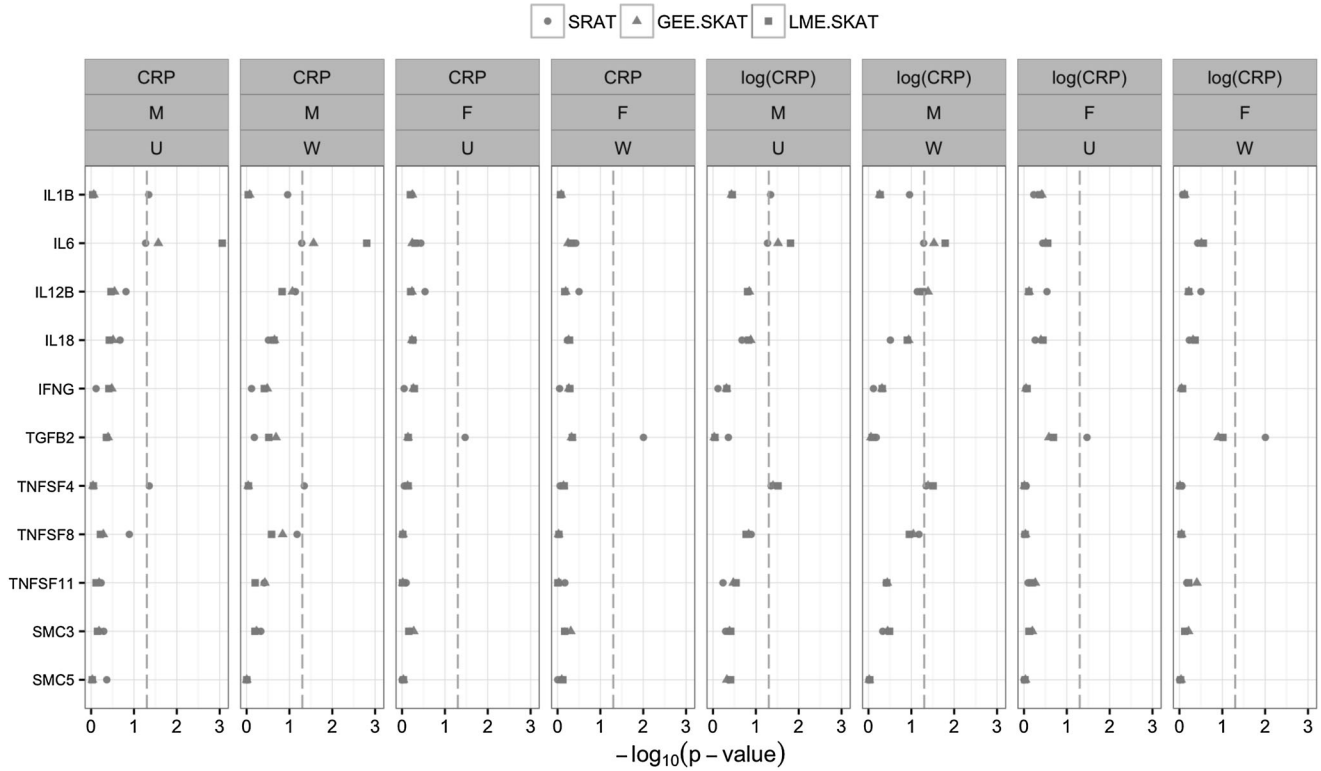


**Figure 4.** Results from testing genetic association between candidate genes and the CRP or logCRP in the Framingham cohort based on SRAT, GEE-SKAT, and LME-SKAT for male and female subjects separately. For all methods, we considered both unweighted (U) analysis and weighted (W) analysis with Beta(1,5) weights.

the outcome distribution or correlation structure, but also more powerful than existing methods when the outcome has a skewed distribution. Although some simple transformation can potentially be applied to the outcome prior to the analysis for GEE-SKAT or LME-SKAT, it is not always clear what transformation is appropriate. It is interesting to note that for binary outcomes, GEE-SKAT is fairly robust to the misspecification in the link function and attains similar power as SRAT under such settings in the absence of covariates depending on the SNPs. This is in part due to the robustness of logistic regression under link violation as suggested in Li and Duan (1989) and Eguchi and Copas (2002), as well as the fact that the ranks of the outcomes are essentially the same as the actual outcome values in this setting.

Unlike other parametric and semi-parametric testing procedures, the family correlations are treated as nuisance parameters in calculating SRAT. Although we use the working independence assumption, we effectively employ the "sandwich" estimator in estimating the variance of the score vector and our inference procedure does not assume independence. The proposed SRAT test requires smoothing and hence depends on the bandwidth parameter $h$. Our sensitivity analysis, shown in Web-Appendix B of the Supplementary Materials, indicates that the performance of SRAT is not sensitive to the choice of $h$ provided that it is in the correct range.

Although in this case SRAT is motivated to analyze familial data, it can also be used to analyze data from unrelated individuals (i.e., $m_i = 1$) and is expected to be more robust and powerful than SKAT under settings where the outcome distribution is skewed. The proposed testing procedure assumes linear genetic effects. The incorporation of nonlinear effects can be achieved by implicitly specifying the genetic effects through a kernel machine regression framework similar to those considered in Liu et al. (2007), which warrants further research. Similar to other gene-set analysis methods, SRAT also relies on user-defined group structure to form the marker set. Various knowledge bases such as gene structure, recombination hotspots, protein–protein interaction networks and pathway information can be used to form marker sets. Data adaptive approach to forming such groups warrants future research.

Due to the need of resampling for $p$-value calculation and the rank-based estimation under the null model, the proposed method is more computationally intensive compared to other competing methods in the presence of covariates, although for the simple case without covariates, the test statistic can be much simplified resulting in a substantially faster procedure even when compared to existing methods. However, similar to other score-type tests, the null model only needs to be fit once and the model fit results can be used repeatedly to test the association of the phenotype with a large number of SNP-sets. In addition, we only rely on resampling to estimate the covariance matrix of $\mathbf{S}(\widehat{\boldsymbol{\alpha}})$ and subsequently employ saddlepoint approximation to calculate the $p$-value. Accordingly, only a few hundreds replications of resampling

are needed for each $p$-value calculation to achieve a reasonable accuracy in approximation. Software package for implementing SRAT is available upon request and will be made publicly available at CRAN and/or GitHub. The core algorithm was written in C++ and then integrated with R via the Rcpp and RcppArmadillo packages.

## 6. Supplementary Materials

Web Appendices and Figures referenced in Sections 3 and 5 are available with this article at the *Biometrics* website on Wiley Online Library.

## Acknowledgements

Dai and Yang contributed equally to this work. The Framingham Offspring Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Offspring Study and does not necessarily reflect the opinions or views of the Framingham Offspring Study, Boston University, or the NHLBI. This research was in part supported by grants U54 HG007963, R01 HL089778 , and P01 CA134294 from the National Institute of Health.

## References

Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J., Donnelly, P., et al. (2005). A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Cai, T. and Cheng, S. (2008). Robust combination of multiple diagnostic tests for classifying censored event times. *Biostatistics* **9**, 216–233.

Chen, H., Meigs, J. B., and Dupuis, J. (2013). Sequence kernel association test for quantitative traits in family samples. *Genetic Epidemiology* **37**, 196–204.

Chen, H., Wang, C., Conomos, M., Stilp, A., Li, Z., Sofer, T., et al. (2016). Control for population structure and relatedness for binary traits in genetic association studies via logistic mixed models. *American Journal of Human Genetics* **98**, 653–666.

Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D. P., Thompson, D., Ballinger, D. G., et al. (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093.

Eguchi, S. and Copas, J. (2002). A class of logistic-type discriminant functions. *Biometrika* **89**, 1–22.

Feinleib, M., Kannel, W. B., Garrison, R. J., McNamara, P. M., and Castelli, W. P. (1975). The framingham offspring study. Design and preliminary data. *Preventive Medicine* **4**, 518–525.

Gao, X., Becker, L. C., Becker, D. M., Starmer, J. D., and Province, M. A. (2010). Avoiding the high bonferroni penalty in genome-wide association studies. *Genetic Epidemiology* **34**, 100–105.

Gudmundsson, J., Sulem, P., Manolescu, A., Amundadottir, L. T., Gudbjartsson, D., and Helgason, A. (2007). Genome-wide association study identifies a second prostate cancer susceptibility variate at 8q24. *Nature Genetics* **39**, 631–637.

Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *Journal of Econometrics* **35**, 303–316.

Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E., et al. (2007). A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nature Genetics* **39**, 870–874.

Jin, Z., Ying, Z., and Wei, L. (2001). A simple resampling method by perturbing the minimand. *Biometrika* **88**, 381–390.

Kannel, W. B. and McGee, D. L. (1979). Diabetes and cardiovascular disease: The framingham study. *Jama* **241**, 2035–2038.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next generation sequencing revolution and its impact on genomics. *Cell* **155**, 27–38.

Kuonen, D. (1999). Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929–935.

Lange, K. (1997). *Mathematical and Statistical Methods for Genetic Analysis*. New York: Springer-Verlag.

Latkovskis, G., Licis, N., and Kalnins, U. (2004). C-reactive protein levels and common polymorphisms of the interleukin-1 gene cluster and interleukin-6 gene in patients with coronary heart disease. *European Journal of Immunogenetics* **31**, 207–213.

Lee, S., Abecasis, G., Boehnke, M., and Lin, X. (2014). Rare-variant association analysis: Study designs and statistical tests. *American Journal of Human Genetics* **95**, 5–23.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17**, 1009–1052.

Lin, D. Y. (2005). An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* **21**, 781–787.

Liu, D., Lin, X., and Ghosh, D. (2007). Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. *Biometrics* **63**, 1079–1088.

Liu, H., Tang, Y., and Zhang, H. H. (2009). A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Computational Statistics and Data Analysis* **53**, 853–856.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. *Biometrics* **63**, 751–757.

McCarthy, M. I., Abecasis, G. R., Cardon, L. R., Goldstein, D. B., Little, J., Ioannidis, J. P. A., et al. (2008). Genome-wide association studies for complex traits: Consensus, uncertainty and challenges. *Nature Reviews Genetics* **9**, 356–569.

McIntosh, A. M., Job, D. E., Moorhead, W. J., Harrison, L. K., Whalley, H. C., Johnstone, E. C., et al. (2006). Genetic liability to schizophrenia or bipolar disorder and its relationship to brain structure. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics* **141**, 76–83.

Moskvina, V. and Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology* **32**, 567–573.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics* **74**, 765–769.

Paffen, E. and Moniek, P. (2006). C-reactive protein in atherosclerosis: A causal factor? *cardiovascular Research* **71**, 30–39.

Pagan, A. and Ullah, A. (1999). *Nonparametric Econometrics*. Cambridge, UK: Cambridge University Press.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*. Hayward, California: Institute of Mathematical Statistics.

Ridker, P. M., Hennekens, C. H., Buring, J. E., and Rifai, N. (2000). C-reactive protein and other markers of inflammation in the prediction of cardivascular disease in women. *New England Journal of Medicine* **342**, 836–843.

Schifano, E. D., Epstein, M. P., Bielak, L. F., Jhun, M. A., Kardia, S. L. R., Peyser, P. A., et al. (2012). Snp set association analysis for familial data. *Genetic Epidemiology* **36**, 797–810.

Scott, L. J., Mohlke, K. L., Bonnycastle, L. L., Willer, C. J., Li, Y., Duren, W. L., et al. (2007). A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345.

Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica* **61**, 123–137.

Shrivastava, A. K., Singh, H. V., Raizada, A., and Singh, S. K. (2015). C-reactive protein, inflammation and coronary heart disease. *The Egyptian Heart Journal* **67**, 89–97.

Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., et al. (2007). Agenome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885.

Su, Z., Marchini, J., and Donnelly, P. (2011). Statistical significance for genomewide studies. *Proceedings of National Academy of Sciences* **100**, 9440–9445.

Thomas, G., Jocobs, K. B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., et al. (2008). Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genetics* **40**, 310–315.

Vo, T. M., Phan, J. H., Huynh, K. N., and Wang, M. D. (2007). Reproducibility of differential gene detection across multiple microarray studies. In *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, 4231–4234. IEEE.

Wang, C., Zhan, X., Liang, L., Abecasis, G. R., and Lin, X. (2015). Improved ancestry estimation for both genotyping and sequencing data using projection Procrustes analysis and genotype imputation. *American Journal of Human Genetics* **96**, 926–937.

Wang, X., Lee, S., Zhu, X., Redline, S., and Lin, X. (2013). GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genetic Epidemiology* **37**, 778–786.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., et al. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *The American Journal of Human Genetics* **86**, 929–942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics* **89**, 82–93.

Zuk, O., Hechter, E., Sunyaev, S. R., and Lander, E. S. (2012). The mystery of missing heritability: Genetic interactions create phantom heritability. *Proceedings of the National Academy of Sciences* **109**, 1193–1198.

*Received November* 2015. *Revised September* 2016. *Accepted November* 2016.

## Appendix

To establish the asymptotic properties of $\mathbf{S}(\widehat{\boldsymbol{\alpha}})$, we require all assumptions given in Sherman (1993). Additionally, we assume that each family potentially has $M$ members and we let $\delta_{ij}$ be a binary indicator denoting whether the $j$th member of the $i$th family is observed. We assume that the underlying data $\mathscr{D}_0 = \{\overline{\mathbf{D}}_i = (\mathbf{D}_{i1}^\top, ...., \mathbf{D}_{iM}^\top)^\top, i = 1, \ldots, n\}$ consist of iid random vectors and the underlying data also follow the NPT model:

$$H(Y_{ij}) = \boldsymbol{\alpha}_0^\top \mathbf{X}_{ij} + \boldsymbol{\beta}_0^\top \mathbf{G}_{ij} + \epsilon_{ij},\ j = 1, \ldots, M, i = 1, \ldots, n.$$

We further assume that missing is completely at random with $P(\delta_{ij} = 1|\overline{\mathbf{D}}_i) = \pi$ and $\mathbf{D}_{ij}$'s have the same marginal distribution. Under the NPT model and $H_0$,

$$\begin{aligned} P(Y_{ij} \le y|\mathbf{X}_i, \mathbf{G}_i) &= P(Y_{ij} \le y|\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}) \\ &= g\{H(y) - \boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}\} \equiv F_{\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}}(y). \end{aligned}$$

Then, we may write $L(\boldsymbol{\alpha}, \mathbf{0}) = n^{-2} \sum_{i=1}^n \sum_{j=1}^M \sum_{i'=1}^n \sum_{j'=1}^M \delta_{ij}\delta_{i'j'} I(Y_{ij} > Y_{i'j'}) I(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} > \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'})$ and it follows similar arguments as given in Sherman (1993) that $\widehat{\boldsymbol{\alpha}} = \arg\max_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \mathbf{0}) \to \boldsymbol{\alpha}_0 = \arg\max_{\boldsymbol{\alpha}} P(Y_{ij} > Y_{i'j'}, \boldsymbol{\alpha}^\top \mathbf{X}_{ij} > \boldsymbol{\alpha}^\top \mathbf{X}_{i'j'})$ in probability.

$$\begin{aligned} \mathbf{S}(\boldsymbol{\alpha}) = \int \frac{1}{n} \sum_{i=1}^n \\ \times \left\{ \sum_{j=1}^M \delta_{ij} \mathbf{G}_{ij} \text{sign}(Y_{ij} - y) K_h(\boldsymbol{\alpha}^\top \mathbf{X}_{ij} - x) \right\} \widehat{F}_{\boldsymbol{\alpha}}(dx, dy), \end{aligned}$$

where

$$\widehat{F}_{\boldsymbol{\alpha}}(x, y) = \frac{1}{n} \sum_{i'=1}^n \left\{ \sum_{j'=1}^M \delta_{i'j'} I(\boldsymbol{\alpha}^\top \mathbf{X}_{i'j'} \le x, Y_{i'j'} \le y) \right\}.$$

It follows from a uniform law of large numbers (Pollard, 1990) that $\sup_{\boldsymbol{\alpha},x,y} |\widehat{F}_{\boldsymbol{\alpha}}(x, y) - F_{\boldsymbol{\alpha}}(x, y)| \xrightarrow{p} 0$, where $F_{\boldsymbol{\alpha}}(x, y) = \pi M \int_{-\infty}^x f_{\boldsymbol{\alpha}}(v) F_{\boldsymbol{\alpha}}(y|v) dv$, $f_{\boldsymbol{\alpha}}(\cdot)$ is the marginal density function of $\boldsymbol{\alpha}^\top \mathbf{X}_{ij}$, and $F_{\boldsymbol{\alpha}}(y|v) = P(Y_{ij} \le y|\boldsymbol{\alpha}^\top \mathbf{X}_{ij} = v)$. This, together with the uniform consistency of non-parametric kernel estimators (Pagan and Ullah, 1999) implies that uniformly in $\boldsymbol{\alpha}$,

$$\begin{aligned} \mathbf{S}(\boldsymbol{\alpha}) &= n^{-1}\pi M \sum_{i=1}^n \sum_{j=1}^M \int \delta_{ij} \mathbf{G}_{ij} \text{sign}(Y_{ij} - y) K_h \\ &\quad \times (\boldsymbol{\alpha}^\top \mathbf{X}_{ij} - x)\ F_{\boldsymbol{\alpha}}(dy|x) f_{\boldsymbol{\alpha}}(x) dx + o_p(1) \\ &= n^{-1}\pi M \sum_{i=1}^n \sum_{j=1}^M \delta_{ij} \mathbf{G}_{ij} \left\{ 2F_{\boldsymbol{\alpha}}(Y_{ij}|\boldsymbol{\alpha}^\top \mathbf{X}_{ij}) - 1 \right\} \\ &\quad \times f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top \mathbf{X}_{ij}) + o_p(1) \\ &\xrightarrow{p} \mathbf{s}(\boldsymbol{\alpha}) \equiv \pi^2 M^2 E \left[ \mathbf{G}_{ij} \left\{ 2F_{\boldsymbol{\alpha}}(Y_{ij}|\boldsymbol{\alpha}^\top \mathbf{X}_{ij}) - 1 \right\} \right. \\ &\quad \left. \times f_{\boldsymbol{\alpha}}(\boldsymbol{\alpha}^\top \mathbf{X}_{ij}) \right]. \end{aligned} \tag{A1}$$

From (A1) and the fact that $\mathbf{G}_{ij} \perp Y_{ij}$ given $\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}$, we have $F_{\boldsymbol{\alpha}_0}(Y_{ij}|\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}) = F_{\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}}(Y_{ij})$ following a Uniform(0,1) distribution and

$$\begin{aligned} \mathbf{S}(\boldsymbol{\alpha}_0) \xrightarrow{p} \mathbf{s}(\boldsymbol{\alpha}_0) = \pi^2 M^2 E \left[ E(\mathbf{G}_{ij} \mid \boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}) E \right. \\ \left. \times \left\{ 2F_{\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}}(Y_{ij}) - 1 \right\} f_{\boldsymbol{\alpha}_0}(\boldsymbol{\alpha}_0^\top \mathbf{X}_{ij}) \right] = \mathbf{0}. \end{aligned}$$

It follows from $\widehat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_0$ and the uniform convergence of $\mathbf{S}(\boldsymbol{\alpha}) \to s(\boldsymbol{\alpha})$ in probability that $\mathbf{S}(\widehat{\boldsymbol{\alpha}}) \xrightarrow{p} \mathbf{0}$.