HHS Public Access

Author manuscript

Biometrics. Author manuscript; available in PMC 2022 April 06.

Published in final edited form as:

Biometrics. 2022 March ; 78(1): 364–375. doi:10.1111/biom.13418.

Generalized Multi-SNP Mediation Intersection-Union Test

Wujuan Zhong[1], **Toni Darville**[2], **Xiaojing Zheng**[1,2,*], **Jason Fine**[1,4,*], **Yun Li**[1,3,5,*]

[1]Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2]Department of Pediatrics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[3]Department of Genetics, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[4]Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[5]Department of Computer Science, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

Abstract

To elucidate the molecular mechanisms underlying genetic variants identified from genome-wide association studies (GWAS) for a variety of phenotypic traits encompassing binary, continuous, count, and survival outcomes, we propose a novel and flexible method to test for mediation that can simultaneously accommodate multiple genetic variants and different types of outcome variables. Specifically, we employ the Intersection-union test approach combined with likelihood ratio test to detect mediation effect of multiple genetic variants via some mediator (for example, the expression of a neighboring gene) on outcome. We fit high-dimensional generalized linear mixed models under the mediation framework, separately under the null and alternative hypothesis. We leverage Laplace approximation to compute the marginal likelihood of outcome and use coordinate descent algorithm to estimate corresponding parameters. Our extensive simulations demonstrate the validity of our proposed methods and substantial, up to 97%, power gains over alternative methods. Applications to real data for the study of *Chlamydia trachomatis* infection further showcase advantages of our method. We believe our proposed methods will be of value and general interest in this post-GWAS era to disentangle the potential causal mechanism from DNA to phenotype for new drug discovery and personalized medicine.

1 Introduction

Dissection of mediation pathways underlying genetic association will enhance understanding of disease mechanisms and biomarker development. An example is *Chlamydia trachomatis* infection. Chlamydia is the leading bacterial sexually transmitted infection in the United States (Centers for Disease Control and Prevention, 2019). Infection

*To whom correspondence should be addressed. xiaojinz@email.unc.edu or jfine@email.unc.edu or yunli@med.unc.edu.

is often asymptomatic and after ascending to the upper genital tract may cause severe reproductive morbidities in women. Repeated infection leads to worse disease. Host genetics shapes susceptibility to chlamydia disease and/or reinfection (Bailey et al., 2009; Taylor et al., 2017; Zheng et al., 2018). DNA biomarkers for susceptibility to ascension or risk of reinfection are critically needed for targeted screening for women at high risk of disease and vaccine development. Genome-wide association studies (GWAS) provide candidate loci, but lack mechanistic interpretations. Although expression quantitative trait loci (eQTL) mapping can provide mechanistic hypotheses, GWAS and eQTL both only analyze two sources of data. There is a significant unmet need for simultaneously modeling all three sources of data (namely, genetic variants, gene expression and final outcome) by directly testing the mediation effects of multiple correlated single nucleotide polymorphisms (SNPs) via the expression of some gene (e.g., eGene associated with the eQTL SNP) on chlamydia ascension (binary outcome) and reinfection (time-to-event outcome).

Mediation analysis was firstly proposed by Baron and Kenny to study the association between an independent variable and an outcome by adding an intermediate variable, which is called the mediator (Baron and Kenny, 1986). In genetics and genomics studies, researchers are interested in testing mediation effects of the genetic variant(s), on the outcome through a certain mediator (e.g., the expression level of a neighboring gene). Non-Gaussian outcomes, such as binary, count and time-to-event outcomes (e.g. disease status, time until death), are commonly present in mediation analyses but have been under-studied. Huang et al developed mixed model based methods that can handle binary and time-to-event outcomes, but assume *a priori* that the genetic variants under testing are eQTLs (Huang et al., 2015; Huang, Cai and Kim, 2016).

We have previously proposed a method, SMUT, to assess mediation effect of high-dimensional genetic variants on any continuous outcome (Zhong et al., 2019). To the best of our knowledge, none of the existing methods can jointly test mediation effects of multiple correlated SNPs (not necessarily all eQTLs) on a non-Gaussian outcome. Here, we propose a generalized multi-SNP mediation intersection-union test to evaluate mediation effects of multiple correlated SNPs on a non-Gaussian outcome without prior knowledge of eQTLs. Both SMUT and methods proposed in this work are extensions of Baron and Kenny's framework and leverage intersection-union test (IUT) (Berger and Hsu, 1996) to decompose mediation into two separate regression models. While our earlier SMUT method handles only Gaussian outcome, methods proposed here allow non-Gaussian outcomes by adopting the generalized linear mixed model (GLMM) (McCulloch, Searle and Neuhaus, 2008) or the mixed effects Cox proportional hazards (PH) model (Vaida and Xu, 2000; Pankratz, De Andrade and Therneau, 2005). More details germane to the differences between SMUT and methods proposed here are in Supporting Information Section 1. For presentation brevity, we hereafter refer to our method for a binary or count outcome as SMUT_GLM; while that for a time-to-event outcome as SMUT_PH.

The rest of this article is organized as follows. In Section 2, we present details of our proposed methods SMUT_GLM and SMUT_PH, followed by simulation studies and real data application in Section 3 and Section 4, respectively. Finally, Section 5 concludes the article with some discussions.

2 Methods

2.1 Notation

Without loss of generality, we assume that we have four types of data, namely, genotypes (as the potential causal variables), gene expression measurements (as the mediator, which can be other types of molecular measures such as metabolite levels or protein abundances), phenotypic trait (as the final outcome) and other covariates (e.g. age, gender). Let $\boldsymbol{G}$ be the n by q genotype matrix, where n is the sample size, q is the number of SNPs and G_{ij} is the number of copies of the minor allele for the ith individual at the jth SNP. Let $\boldsymbol{X}$ be the n by p covariate matrix and X_{ij} denote the jth covariate variable for the ith individual. Let $\boldsymbol{M} = (M_1, M_2,\ldots,M_n)^T$ and $\boldsymbol{Y} = (Y_1, Y_2,\ldots, Y_n)^T$ where M_i and Y_i denote the mediator and the outcome for the ith individual, respectively. If Y_i is a binary or count outcome, Y_i is related to the model in (2); if Y_i is a time-to-event outcome, M_i is related to the model in (3) and $M_i = (Z_i, \delta_i)$ where $Z_i = \min(T_i, C_i)$ is the observation time, T_i is the failure time and C_i is the censoring time, and $\delta_i = I(T_i \leq C_i)$ is the failure indicator; $\delta_i = 1$ indicates that the failure is observed and $\delta_i = 0$ indicates that the response is censored. We apologize for abusing notations. Basically, we want to use the same notation Y_i to denote different types of outcomes.

2.2 SMUT_GLM and SMUT_PH model

SMUT_GLM and SMUT_PH model the effects of SNPs on the outcome mediated by the expression level of a single gene via two models, namely a mediator model and an outcome model. We assume the expression level is continuous and consider a linear model for the mediator model (1). As for the outcome model, we fit GLMM if the outcome conditional on SNPs' effects follows an exponential family distribution (2); we fit mixed effects Cox PH model if the outcome is a time-to-event variable (3).

$$M_i = \alpha_1 + \sum_{j=1}^{p} X_{ij} l_j^M + \sum_{j=1}^{q} G_{ij}\beta_j + \epsilon_i \text{ Mediator model} \quad (1)$$

$$g\{E(Y_i \mid \boldsymbol{\gamma})\} = \alpha_2 + M_i\theta + \sum_{j=1}^{p} X_{ij} l_j + \sum_{j=1}^{q} G_{ij}\gamma_j \text{ Exponential Family Outcome model} \quad (2)$$

$$\lambda(t_i) = \lambda_0(t_i)\exp\left(M_i\theta + \sum_{j=1}^{p} X_{ij} t_j + \sum_{j=1}^{q} G_{ij}\gamma_j\right) \text{ Survival Outcome model} \quad (3)$$

Where α_1, α_2 are fixed intercepts; fixed effects $\boldsymbol{\iota}^{\boldsymbol{M}} = \left(\iota_1^M, \iota_2^M, \ldots, \iota_p^M\right)^T$ and $\boldsymbol{\iota} = (\iota_1, \iota_2,\ldots, \iota_p)T$ are vectors of covariates' effects on the mediator and outcome, respectively; random effects $\boldsymbol{\beta} = (\beta_1, \beta_2,\ldots,\beta_q)^T$ is a vector of SNPs' effects on the mediator; fixed effect θ is the mediator's effect on the outcome. The random effects $\boldsymbol{\gamma} = (\gamma_1, \gamma_2,\ldots, \gamma_q)^T$ is a vector of SNPs' effects on the outcome; error terms $\epsilon_1, \epsilon_2, \ldots, \epsilon_n \sim_{i.i.d.} N\left(0, \sigma^2\right)$; g is the link function; $\lambda(t_i)$ is the hazard function; $\lambda_0(t_i)$ is an unspecified baseline hazard function.

We have showed that the hypotheses H_0:$\boldsymbol{\beta}\theta = \boldsymbol{0}$ versus H_1:$\boldsymbol{\beta}\theta = \boldsymbol{0}$ are valid for testing mediation effect in Supporting Information Section 8, where $\boldsymbol{\beta}\theta \neq \boldsymbol{0}$ implies that SNPs exert

mediation effects on the outcome. Following our previous work (Zhong et al., 2019), we employ IUT to decompose the hypothesis testing H_0:$\boldsymbol{\beta\theta} = \mathbf{0}$ versus H_1:$\boldsymbol{\beta\theta} \neq \mathbf{0}$ into two sub-hypotheses $H_{\mathbf{0}}^{\beta}$:$\boldsymbol{\beta} = \mathbf{0}$ versus H_1^{β}:$\boldsymbol{\beta} \neq \mathbf{0}$ and H_0^{θ}:$\theta = \mathbf{0}$ versus H_1^{θ}:$\theta \neq \mathbf{0}$, such that $H_0 = H_0^{\theta} \cup H_0^{\beta}$ and $H_1 = H_1^{\theta} \cap H_1^{\beta}$. Suppose the p values for testing $\boldsymbol{\beta}$ and θ being zero are p_1 and p_2, respectively. Then the p value for testing $\boldsymbol{\beta\theta}$ being zero, using IUT, is the maximum of p_1 and p_2. In the following sections, we provide details regarding how to separately test $\boldsymbol{\beta}$ and θ to obtain p_1 and p_2

2.3 Testing $\boldsymbol{\beta}$ in the mediator model and θ in the outcome model

As in (Zhong et al., 2019), we adopt the widely used SKAT method (Wu et al., 2011) to test $\boldsymbol{\beta}$ in the mediator model to accommodate a potentially large number of correlated SNPs.

Our strategy for testing θ in the outcome model consists of four steps: (1) formulation of the likelihood function based on the nature of the outcome random variable Y, and (2) Laplace approximation of the likelihood function, and (3) application of the coordinate descent algorithm (Fu, 1998; Daubechies, Defrise and De Mol, 2004) to estimate parameters by maximizing the approximated likelihood function, and (4) calculation of the likelihood ratio statistic. These four steps allow us to test the mediator effect θ in the outcome model.

2.3.1 Likelihood function for the outcome model—To reduce the dimensionality of parameters in the outcome model, we adopted a linear mixed model for continuous outcome in our previous work (Zhong et al., 2019). We assume $Y_1, Y_2, \ldots, Y_n$ are independent and identically distributed. When the outcome Y_i (i = 1,2,…,n) conditional on $\boldsymbol{\gamma}$ follows an exponential family distribution, we adopt the GLMM in equation (2).

$$\begin{cases} \gamma_j \sim_{i.i.d.} N\left(0, \sigma_\gamma^2\right) \\ g(\mu_i) = \eta_i = \alpha_2 + M_i\theta + \Sigma_{j=1}^{p} X_{ij}t_j + \Sigma_{j=1}^{q} G_{ij}\gamma_j \\ E(Y_i \mid \boldsymbol{\gamma}) = \mu_i \\ L(\boldsymbol{y} \mid \boldsymbol{\gamma}) = \Pi_{i=1}^{n} \exp\left\{ \frac{y_i\tau_i - b(\tau_i)}{a(\phi)} + C(y_i, \phi) \right\} \end{cases} \tag{4}$$

where τ_i is the canonical parameter; ϕ is the dispersion parameter; $(\boldsymbol{y}|\boldsymbol{\gamma})$ is the likelihood function of the outcome $\boldsymbol{Y}$ conditional on $\boldsymbol{\gamma}$. When the outcome Y_i (i = 1,2,…,n) is a time-to-event variable, we adopt the mixed effects Cox PH model in equation (3).

$$\begin{cases} \gamma_j \sim_{i.i.d.} N\left(0, \sigma_\gamma^2\right) \\ \eta_i = M_i\theta + \Sigma_{j=1}^{p} X_{ij}t_j + \Sigma_{j=1}^{q} G_{ij}\gamma_j \\ \lambda(t_i) = \lambda_0(t_i)\exp\eta_i \\ PL = \Pi_{i=1}^{n} \left(\frac{\exp\eta_i}{\Sigma_{k \in R_i} \exp\eta_k} \right)^{\delta_i} \end{cases} \tag{5}$$

where $R_i = \{k{:}Z_k \geq Z_i\}$ is the risk set and PL is the partial likelihood function conditional on $\boldsymbol{\gamma}$. For the GLMM in (4), $\ell(\boldsymbol{y}|\boldsymbol{\gamma})$ denotes log $L(\boldsymbol{y}|\boldsymbol{\gamma})$ and $L(\boldsymbol{y})$ denotes the likelihood function

of the outcome unconditional on $\boldsymbol{\gamma}$; for the mixed effects Cox PH model in (5), $\ell(\boldsymbol{y}|\boldsymbol{\gamma})$ denotes log PL and $L(\boldsymbol{y})$ denotes the partial likelihood of the outcome unconditional on $\boldsymbol{\gamma}$. We again apologize for abusing notations. Our basic rationale is to employ the same notation $\ell(\boldsymbol{y}|\boldsymbol{\gamma})$ and $L(\boldsymbol{y})$ to denote different log-likelihood and likelihood functions, respectively, for different types of outcomes. Let $f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})$ be the probability density function of $\boldsymbol{\gamma}$, and $f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma}) = \left(2\pi\sigma_\gamma^2\right)^{-\frac{q}{2}}\exp\left(-\frac{1}{2\sigma_\gamma^2}\boldsymbol{\gamma}^T\boldsymbol{\gamma}\right)$. Then we have the following.

$$L(\boldsymbol{y}) = \int_{R^q}\exp\{\ell(\boldsymbol{y}|\boldsymbol{\gamma})\}f_{\boldsymbol{\gamma}}(\boldsymbol{\gamma})d\boldsymbol{\gamma} = \left(2\pi\sigma_\gamma^2\right)^{-\frac{q}{2}}\int_{R^q}\exp\{h(\boldsymbol{\gamma})\}d\boldsymbol{\gamma} \tag{6}$$

where $h\left(\boldsymbol{\gamma}\right) = \ell(\boldsymbol{y}|\boldsymbol{\gamma}) - \frac{1}{2\sigma_\gamma^2}\boldsymbol{\gamma}^T\boldsymbol{\gamma}$. Technical details are in Supporting Information Section 2.1.

2.3.2 Laplace approximation—Laplace's method is widely adopted to approximate the likelihood function (Breslow and Clayton, 1993; Raudenbush, Yang and Yosef, 2000; Pankratz et al., 2005). The integral in equation (6) can be approximated via Laplace's method by taking Taylor expansion to the second order of $h(\boldsymbol{\gamma})$ around its maximum point $\tilde{\boldsymbol{\gamma}}$. After inserting the Taylor expansion into the integral, and taking logarithm, we have the approximated log-likelihood f.

$$\log L(\boldsymbol{y}) \approx f = -\frac{q}{2}\log\sigma_\gamma^2 + h(\tilde{\boldsymbol{\gamma}}) - \frac{1}{2}\log|-h''(\tilde{\boldsymbol{\gamma}})| \tag{7}$$

For the GLMM in (4), we have

$$h''(\boldsymbol{\gamma}) = \frac{\partial^2 h}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = -\left(\boldsymbol{G}^T\boldsymbol{W}\boldsymbol{G} + \sigma_\gamma^{-2}\boldsymbol{I}_q\right) \tag{8}$$

where $\boldsymbol{I}_\mathrm{q}$ is a q by q identity matrix, $\boldsymbol{W} = diag(w_1, w_2, \ldots, w_n$, and w_i is recognizable as GLM (generalized linear model) iterative weight. For the mixed effects Cox PH model in (5), we have

$$h''(\boldsymbol{\gamma}) = \frac{\partial^2 h}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^T} = -\left(\boldsymbol{U} + \sigma_\gamma^{-2}\boldsymbol{I}_q\right) \tag{9}$$

where $\boldsymbol{U} = \left(u_{j_1 j_2}\right)$, $u_{j_1 j_2} = -\frac{\partial^2(\log PL)}{\partial\gamma_{j_1}\partial\gamma_{j_2}}$. More details of Laplace approximation are in Supporting Information Section 2.2.

2.3.3 Coordinate descent algorithm—We apply the coordinate descent algorithm to maximize the approximated log-likelihood in equation (9). Note that $\tilde{\boldsymbol{\gamma}}$ in equation (9) is a function of other parameters $\boldsymbol{\xi} = \left(\alpha_2, \sigma_\gamma^2, \phi, \theta, \iota_1, \iota_2, \ldots, \iota_p\right)$. Instead of taking implicit differentiation of $\tilde{\boldsymbol{\gamma}}$ (Raudenbush et al., 2000), we use the approximation strategy proposed in (Schelldorfer, Meier and Bühlmann, 2014), which regards $\tilde{\boldsymbol{\gamma}}$ as fixed when updating $\boldsymbol{\xi}$. This strategy is computationally convenient and efficient, at little cost of reduced accuracy.

In addition, we take further approximation when taking derivatives of the approximated log-likelihood function f. Specifically, for the GLMM in (4), we assume $\boldsymbol{W}$ in equation (10) varies slowly as a function of $(\mu_1, \mu_2, \ldots \mu_n)^T$ (Breslow and Clayton, 1993). For the mixed effects Cox PH model in (5), we similarly assume that $\boldsymbol{U}$ in equation (11) varies slowly as a function of $(\eta_1, \eta_2, \ldots \eta_n)^T$. Under the assumption, the term $-\frac{1}{2}\log|-h''(\tilde{\gamma})|$ in equation (9) is ignored when taking derivatives of the approximated log-likelihood function over $(\alpha_2, \phi, \theta, \iota_1, \iota_2, \ldots, \iota_p)$. Details of the coordinate descent algorithm are in Supporting Information Section 2.3. Finally, we employ the Newton-Raphson algorithm to sequentially update each parameter.

2.3.4 Likelihood ratio test—We obtain approximated likelihood under the null and the alternative hypothesis separately, denoted by L_0 and L_2 respectively. For GLMM, the likelihood ratio statistic $2(\log L_1 - \log L_0)$ asymptotically follows a chi-square distribution with one degree of freedom, and similarly for the partial likelihood ratio statistics for the survival outcome.

3 Simulation studies

3.1 Simulation settings

To evaluate the performance of SMUT_GLM and SMUT_PH in comparison with alternative methods, we conducted extensive simulations to investigate power and type-I error. Following our previous work (Zhong et al., 2019), we simulated a dataset of 10,000 pseudo-individuals measured at 2,891 SNPs with minor allele frequency (MAF) ≥ 1% in a 1Mb region using the COSI coalescent model (Schaffner et al., 2005) to generate realistic genetic data. The 10,000 pseudo-individuals were constructed by randomly pairing up 20,000 simulated chromosomes without replacement. To evaluate power and type-I error, we generated 500 datasets with 1,000 samples each by sampling without replacement from the entire pool of 10,000 samples simulated above. We randomly selected a set of causal SNPs, which is shared across the 500 simulated datasets, from these 2,891 SNPs. We then classified them into three categories: shared SNPs (*sSNPs*), mediator specific SNPs (*mSNPs*) and outcome specific SNPs (*oSNPs*). The *sSNPs* influence both the mediator and the outcome, while the *mSNPs* and *oSNPs* only contribute to the mediator and outcome, respectively.

We considered two scenarios in terms of causal SNP density: sparse and dense (Table 1). For binary or count outcome, sample size is 1,000 and there are 10 and 500 causal SNPs for sparse and dense scenarios, respectively. For time-to-event outcome, sample size is 200 and there are 10 and 150 causal SNPs for sparse and dense scenarios, respectively. When we fit the model, both the causal and non-causal SNPs (Table 1) are included in the model. Thus, the distribution of coefficients of genetic variants is effectively mis-specified for all the simulations. Covariates matrix $\boldsymbol{X}$ consists of a continuous variable generated from $N(0,1)$ and a binary variable generated from Bernoulli(0.5). We generated the mediator via $M_i = \alpha_1 + (\boldsymbol{G}_i^{sm})^T \boldsymbol{\beta} + (\boldsymbol{X}_i)^T \boldsymbol{\iota}^{\boldsymbol{M}} + \epsilon_i$, where $\boldsymbol{G}_i^{sm}$ denotes the vector of genotype data for the ith individual from *sSNPs* and *mSNPs*, $\boldsymbol{X}_i$ denotes the vector of the covariates for

the ith individual, $\alpha_1 = 1$, $\boldsymbol{\iota}^{\boldsymbol{M}} = (0.5, -0.5)^Y$, $\boldsymbol{\beta} \sim c_\beta N(\mathbf{0}, \boldsymbol{I}_q)$ and c_β is a scalar to scale the SNPs' effects; $\epsilon_i \sim N(0,1)$. We generated the binary or count outcome via $g\{E(Y_i|\boldsymbol{\gamma})\} = \alpha_2 + M_i\theta + \left(\boldsymbol{G}_i^{so}\right)^T\boldsymbol{\gamma} + (\boldsymbol{X}_i)^T\boldsymbol{\iota}$, where $\boldsymbol{G}_i^{so}$ denotes the vector of genotype data for the ith individual from *sSNPs* and *oSNPs*, $\alpha_2 = 0$,, $\boldsymbol{\iota} = 0.5, -0.5)^T$, $\boldsymbol{\gamma} \sim c_\gamma N(0, \boldsymbol{I}_q)$ and $c_\gamma = 0.2$. The link function g was specific to the type of the outcome (Supporting Information Section 2.1). We generated the time-to-event outcome based on Weibull baseline hazard via $t_i = \left[-\frac{\log(v)}{\lambda \exp\left\{M_i\theta + \left(\boldsymbol{G}_i^{so}\right)^T\boldsymbol{\gamma} + (\boldsymbol{X}_i)^T\boldsymbol{\iota}\right\}}\right]^{\frac{1}{\rho}}$ and $c_i \sim Exp(0.001)$, where t_i is failure time and c_i is censoring time, $v \sim Unif(0,1)$, shape $\rho = 1$, scale parameter $\lambda = 0.01$. Note that across the 500 datasets, error terms ϵ were separately simulated for each dataset, but $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ were fixed.

In the simulations, we tested the mediation effects of these SNPs on the binary, count or time-to-event outcome using SMUT_GLM and SMUT_PH, as well as other methods including SMUT, adapted LASSO (Tibshirani, 1996) and adapted Huang et al.'s method. In order to compare the performance of approximations that we adopted, we considered two versions of our method, both treating $\tilde{\boldsymbol{\gamma}}$ as fixed: (1) based on exact derivatives; (2) based on approximated derivatives. For a binary or count outcome, we refer to these two versions as SMUT_GLM exact and SMUT_GLM approxi. For a time-to-event outcome, we refer to the approximated version as SMUT_PH approxi. The exact version of SMUT_PH is not employed because it is hard to derive analytically. SMUT is naively applied to binary and count outcomes by treating them as continuous variables. The adapted LASSO approach adopts SKAT to consider all the genetic variant in the mediator model, while in the outcome model, employs LASSO for variable selection on all genetic variants as well as mediator and covariates, then refits GLM on the selected genetic variants together with mediator and covariates (latter two will be included regardless of LASSO variable selection results), and finally combines p values from the mediator and the refitted outcome model via IUT. The adapted Huang et al.'s method employs SKAT in the mediator model, adopts the original Huang et al.'s method in the outcome model, and then combines p values from the two models via IUT. We use adapted LASSO and SKAT + LASSO exchangeably. Similarly, we use adapted Huang et al. and SKAT + Huang et al. exchangeably. Details of the adapted LASSO and adapted Huang et al.'s method are in Supporting Information Section 3.

To test the robustness and generalizability of the methods, we considered two alternative situations where some assumption is violated. The first situation is the violation of the assumption that coefficients of genetic variants follow a Gaussian distribution. The second situation is when there is an unobserved mediator that is not adjusted in the analysis. Details and results of these two simulation studies are in Supporting Information Section 4.

3.2 Type-I error in simulations

We evaluated the validity of SMUT_GLM and SMUT_PH along with alternative methods in simulations. SMUT_GLM and SMUT_PH exhibited controlled type-I error rates, at α = 0.05 level, regardless of causal SNP density and types of outcome, as shown in Figures 1

and 2 for binary outcome in sparse and dense scenarios respectively, Figures 3 and 4 for time-to-event outcome in sparse and dense scenarios respectively, Web Figures S1 and S2 for count outcome in sparse and dense scenarios respectively. In each figure, the first panel ($c_\beta = 0$) and the leftmost point ($\theta = 0$) in other panels ($c_\beta \neq 0$) all correspond to the null of no mediation of the SNPs through the mediator. SMUT, adapted LASSO and adapted Huang et al.'s method also showed protected type-I error.

3.3 Power in simulations

SMUT_GLM and SMUT_PH demonstrated substantial power gains under both the sparse and dense scenarios. We also observed that the approximated version of SMUT_GLM demonstrated very similar performance when compared with its exact counterpart. For example, for binary outcome and under the scenario of dense causal SNPs when $c_\beta = 0.6$, $\theta = 0.1$, exact SMUT_GLM, approximated SMUT_GLM, SMUT, adapted LASSO and adapted Huang et al. had 97%, 96%, 17%, 54% and 0% power, respectively. Thus, the power gain from the exact SMUT_GLM was 80%, 43% and 97% compared with SMUT, adapted LASSO and adapted Huang et al., respectively. The approximated SMUT_GLM had similar power gains. For time-to-event outcome, under the scenario of dense causal SNPs when $c_\beta = 1$, $\theta = 0.075$, approximated SMUT_PH and adapted LASSO had 69% and 41% power, respectively, leading to a power gain of 28%. In addition, power gains appeared more profound with increasing c_β, likely because adapted LASSO and adapted Huang et al. becomes more conservative as the pleiotropy effect of SNPs on mediator and outcome (measured by c_β) increases.

4 Real data application

We assessed our methods and alternatives in real data from two clinical cohorts, which were designed for the study of chlamydia infection. *Chlamydia trachomatis* can ascend from the cervix to the uterus and fallopian tubes in some women, potentially resulting in pelvic inflammatory disease (PID) and severe reproductive morbidities, including infertility and ectopic pregnancy. Recurrent infection leads to worse disease. We analyzed genotype, gene expression and phenotype data of 200 participants combined from two cohorts, the Anaerobes and Clearance of Endometritis (ACE) cohort and the T cell Response Against Chlamydia (TRAC) cohort (Russell et al., 2015). The Institutional Review Boards for Human Subject Research at the University of Pittsburgh and the University of North Carolina approved the study and all participants provided written informed consent prior to inclusion. Descriptions of the ACE and TRAC cohorts, processing and quality control of genotype and gene expression data, and details of eQTL analysis and mediation analysis of other genes are in Supporting Information Section 6.

4.1 Binary outcome

The outcome of interest is ascending chlamydia infection, among participants who had chlamydia infection at enrollment. The control group is the 71 participants who had chlamydia infection restricted to the cervix, and the case group is the 72 participants with both cervical and endometrial chlamydia infection at enrollment. We analyzed genotype, gene expression and phenotype data from these 143 participants.

Here we presented *SOS1* and *CD151* genes, which were biologically related to the outcome, to illustrate the application of our proposed methods to a binary outcome. Son of sevenless homolog 1 (*SOS1*) is a guanine nucleotide exchange factor that in humans is encoded by the *SOS1* gene. The importance of *SOS1* for chlamydia invasion of host cells has been indicated by multiple biomedical studies (Carabeo et al., 2007; Lane et al., 2008; Hackstadt, 2012; Bastidas et al., 2013; Mehlitz and Rudel, 2013; Elwell, Mirrashidi and Engel, 2016). The *CD151* gene encodes a protein that is known to complex with integrins. It promotes cell adhesion and may regulate integrin trafficking and/or function. It is a member of the tetraspanin family, which are considered as the gateways for infection (Hauck and Meyer, 2003; Hemler, 2008; Hassuna et al., 2009; Join-Lambert et al., 2010; N Monk and J Partridge, 2012; Seu et al., 2017). In addition, SNPs annotation database, RegulomeDB (Boyle et al., 2012), demonstrates that some SNPs in these two genes are eQTLs with experimental evidence. Thus, the presence of mediation effect via the expression of each gene is expected.

For the first gene, *SOS1*, mediation testing encompassed 83 SNPs with MAF ≥ 10% and significant eQTL association (with *SOS1*) at a FDR threshold of 10%, using SMUT_GLM, adapted LASSO and adapted Huang et al.'s method. Both SMUT_GLM and adapted Huang et al.'s method detected significant mediation effects, while adapted LASSO did not (Table 2). For the second gene *CD151*, our mediation (via expression of *CD151*) testing involved 40 SNPs with MAF ≥ 10% and significant eQTL (with *CD151*) at FDR 10%. Only SMUT_GLM showed significant mediation effects of these SNPs through the expression of *CD151* on ascending chlamydia infection (Table 2). Marginal effects of selected SNPs on *SOS1* and *CD151* gene expression and ascending chlamydia infection were visually illustrated in Web Figures S19 and S20 respectively.

4.2 Time-to-event outcome

TRAC participants returned for follow-up visits at 1, 4, 8, and 12 months after enrollment.

The outcome of interest we evaluated here is time to the first incident chlamydia infection. We analyzed genotype, gene expression and time-to-event data from all 181 participants in the TRAC cohort who had both genotype and gene expression data available.

Here we selected *BIRC3* gene, which was biologically related to the outcome, to illustrate the application of our proposed methods to a time-to-event outcome. The gene *BIRC3* encodes for Baculoviral IAP Repeat Containing 3, a E3 ubiquitin-protein ligase regulating NF-kappa-B signaling (Blankenship et al., 2009; Kim et al., 2010; Tan et al., 2013). It acts as an important regulator of pathogen recognition receptor signaling (Bertrand et al., 2009), which can have profound effects on the development of downstream adaptive immune responses (Takeda, Kaisho and Akira, 2003; Palm and Medzhitov, 2009; Kumar, Kawai and Akira, 2011). In addition, biological studies suggested that *BIRC3* may protect mammalian host cells against apoptosis, leading to accommodate chlamydial growth (Bryant et al., 2004; Park, Yoon and Lee, 2004; Paland et al., 2006; Ying et al., 2008). Therefore, mediation effect via the expression of *BIRC3* gene is logical. Our mediation testing involved 4 SNPs with MAF ≥ 10% and eQTL (with *BIRC3*) at FDR 10%, using SMUT_PH, adapted LASSO and adapted Huang et al.'s method. All the methods showed significant mediation effects

through *BIRC3* on incident chlamydia infection (Table 2). Marginal effects of selected SNPs on *BIRC3* gene expression and time to the first incident chlamydia infection were visually illustrated in Web Figures S21.

5 Discussion

Our proposed methods, SMUT_GLM and SMUT_PH, extend our previous work (Zhong et al., 2019) to test mediation effect of multiple correlated genetic variants on a non-Gaussian outcome through a mediator. We adopt a mixed model based approach to handle high dimension of genetic variants and do not apply any variable selection of genetic variants. Our proposed methods are statistically more powerful than alternative methods including SMUT, adapted LASSO and adapted Huang et al.'s method. Analysis and discussions of possible reasons underlying alternative methods' power loss are in Supporting Information Section 5. The approximated version of SMUT_GLM and SMUT_PH are also computationally efficient (Supporting Information Section 7.2).

One limitation of our proposed methods is that we assume the effects of genetic variants follow a Gaussian distribution. This may not be correct when there are non-causal SNPs in the model and in this case, a mixture distribution might be more appropriate. It is reassuring to observe protected type-I error from our simulation studies, which included a large number of non-causal SNPs in all scenarios considered. In addition, supplementary simulation studies (Supporting Information Section 4) further demonstrate controlled type-I error when the effects of genetic variants follow a mixture of two Gaussian distributions. More properly modeling the effects of genetic variants may further increase the statistical power under the alternative hypotheses but due to modeling complexity and subsequently inevitable computational costs, we decide not to further pursue this in our current work.

Our proposed methods can be further extended to handle multiple correlated outcomes for additional power gains as well as to accommodate multiple potentially correlated mediators to jointly assess their mediation effects. Besides, we could adopt nonparametric methods to handle the mediator model and outcome model with more flexibility. Details germane to possible methodological extensions are in Supporting Information Section 7.1. We anticipate our proposed methods will become a powerful tool to bridge the gap in terms of molecular mechanisms between various types of phenotypes and the corresponding associated genetic variant(s) identified in recent literature.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the National Institutes of Health U19 AI084024, U19 AI144181 and R01 AI119164 to TD; U19 AI144181 to XZ; R01 HL129132, R01 HL146500, and U544 HD079124 to YL. YL is also partially supported by U01 DA052713 and R01 GM105785. We thank all participants in ACE and TRAC for agreeing to take part in the studies, and all investigators in these two studies for sharing the data. We thank the editor, the associate editor and three anonymous referees for their helpful comments and suggestions, which significantly improved the manuscript.

Data Availability Statement

The data used in this paper to support our findings are available from the corresponding authors upon reasonable request.

References

Bailey RL, Natividad-Sancho A, Fowler A, Peeling RW, Mabey DC, Whittle HC, et al. (2009) Host genetic contribution to the cellular immune response to Chlamydia trachomatis: Heritability estimate from a Gambian twin study. Drugs of today (Barcelona, Spain: 1998), 45, 45–50.

Baron RM and Kenny DA (1986) The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. Journal of personality and social psychology, 51, 1173–1182. [PubMed: 3806354]

Bastidas RJ, Elwell CA, Engel JN and Valdivia RH (2013) Chlamydial intracellular survival strategies. Cold Spring Harbor perspectives in medicine, 3, a010256. [PubMed: 23637308]

Berger RL and Hsu JC (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. Statistical Science, 11, 283–319.

Bertrand MJM, Doiron K, Labbé K, Korneluk RG, Barker PA and Saleh M (2009) Cellular inhibitors of apoptosis cIAP1 and cIAP2 are required for innate immunity signaling by the pattern recognition receptors NOD1 and NOD2. Immunity, 30, 789–801. [PubMed: 19464198]

Blankenship JW, Varfolomeev E, Goncharov T, Fedorova AV, Kirkpatrick DS, Izrael-Tomasevic A, et al. (2009) Ubiquitin binding modulates IAP antagonist-stimulated proteasomal degradation of c-IAP1 and c-IAP2 1. Biochemical Journal, 417, 149–165.

Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. (2012) Annotation of functional variation in personal genomes using RegulomeDB. Genome research, 22, 1790–1797. [PubMed: 22955989]

Breslow NE and Clayton DG (1993) Approximate Inference in Generalized Linear Mixed Models. Journal of the American Statistical Association, 88, 9.

Bryant PA, Venter D, Robins-Browne R and Curtis N (2004) Chips with everything: DNA microarrays in infectious diseases. The Lancet infectious diseases, 4, 100–111. [PubMed: 14871635]

Carabeo RA, Dooley CA, Grieshaber SS and Hackstadt T (2007) Rac interacts with Abi-1 and WAVE2 to promote an Arp2/3-dependent actin recruitment during chlamydial invasion. Cellular microbiology, 9, 2278–2288. [PubMed: 17501982]

Centers for Disease Control and Prevention. (2019) Sexually Transmitted Disease Surveillance 2018. Atlanta: U.S. Department of Health and Human Services.

Daubechies I, Defrise M and De Mol C (2004) An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 57, 1413–1457.

Elwell C, Mirrashidi K and Engel J (2016) Chlamydia cell biology and pathogenesis. Nature Reviews Microbiology, 14, 385. [PubMed: 27108705]

Fu WJ (1998) Penalized regressions: the bridge versus the lasso. Journal of computational and graphical statistics, 7, 397–416.

Hackstadt T (2012) In: Intracellular Pathogens I: Chlamydiales (eds Tan M and Bavoil P). American Society for Microbiology Press.

Hassuna N, Monk PN, Moseley GW and Partridge LJ (2009) Strategies for targeting tetraspanin proteins. BioDrugs, 23, 341–359. [PubMed: 19894777]

Hauck CR and Meyer TF (2003) 'Small'talk: Opa proteins as mediators of Neisseria--host-cell communication. Current opinion in microbiology, 6, 43–49. [PubMed: 12615218]

Hemler ME (2008) Targeting of tetraspanin proteins—potential benefits and strategies. Nature reviews Drug discovery, 7, 747. [PubMed: 18758472]

Huang Y-T, Cai T and Kim E (2016) Integrative genomic testing of cancer survival using semiparametric linear transformation models. Statistics in medicine, 35, 2831–44. [PubMed: 26887583]

Huang Y-T, Liang L, Moffatt MF, Cookson WOCM and Lin X (2015) iGWAS: Integrative Genome-Wide Association Studies of Genetic and Genomic Data for Disease Susceptibility Using Mediation Analysis. Genetic Epidemiology, 39, 347–356. [PubMed: 25997986]

Join-Lambert O, Morand PC, Carbonnelle E, Coureuil M, Bille E, Bourdoulous S, et al. (2010) Mechanisms of meningeal invasion by a bacterial extracellular pathogen, the example of Neisseria meningitidis. Progress in neurobiology, 91, 130–139. [PubMed: 20026234]

Kim CW, Kim HK, Vo M-T, Lee HH, Kim HJ, Min YJ, et al. (2010) Tristetraprolin controls the stability of cIAP2 mRNA through binding to the 3′ UTR of cIAP2 mRNA. Biochemical and biophysical research communications, 400, 46–52. [PubMed: 20691152]

Kumar H, Kawai T and Akira S (2011) Pathogen recognition by the innate immune system. International reviews of immunology, 30, 16–34. [PubMed: 21235323]

Lane BJ, Mutchler C, Al Khodor S, Grieshaber SS and Carabeo RA (2008) Chlamydial entry involves TARP binding of guanine nucleotide exchange factors. PLoS pathogens, 4, e1000014. [PubMed: 18383626]

McCulloch CE, Searle SR and Neuhaus JM (2008) Generalized, Linear, and Mixed Models, 2nd Edition., 424.

Mehlitz A and Rudel T (2013) Modulation of host signaling and cellular responses by Chlamydia. Cell Communication and Signaling, 11, 90. [PubMed: 24267514]

N Monk P and J Partridge L (2012) Tetraspanins-gateways for infection. Infectious Disorders-Drug Targets (Formerly Current Drug Targets-Infectious Disorders), 12, 4–17.

Paland N, Rajalingam K, Machuy N, Szczepek A, Wehrl W and Rudel T (2006) NF-$κ$B and inhibitor of apoptosis proteins are required for apoptosis resistance of epithelial cells persistently infected with Chlamydophila pneumoniae. Cellular microbiology, 8, 1643–1655. [PubMed: 16984419]

Palm NW and Medzhitov R (2009) Pattern recognition receptors and control of adaptive immunity. Immunological reviews, 227, 221–233. [PubMed: 19120487]

Pankratz VS, De Andrade M and Therneau TM (2005) Random-effects cox proportional hazards model: General variance components methods for time-to-event data. Genetic Epidemiology, 28, 97–109. [PubMed: 15532036]

Park S-M, Yoon J-B and Lee TH (2004) Receptor interacting protein is ubiquitinated by cellular inhibitor of apoptosis proteins (c-IAP1 and c-IAP2) in vitro. FEBS letters, 566, 151–156. [PubMed: 15147886]

Raudenbush SW, Yang ML and Yosef M (2000) Maximum Likelihood for Generalized Linear Models with Nested Random Effects via High-Order, Multivariate Laplace Approximation. Journal of Computational and Graphical Statistics, 9, 141–157.

Russell AN, Zheng X, O'connell CM, Taylor BD, Wiesenfeld HC, Hillier SL, et al. (2015) Analysis of factors driving incident and ascending infection and the role of serum antibody in Chlamydia trachomatis genital tract infection. The Journal of infectious diseases, 213, 523–531. [PubMed: 26347571]

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ and Altshuler D (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Research, 15, 1576–1583. [PubMed: 16251467]

Schelldorfer J, Meier L and Bühlmann P (2014) GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using $\ell 1$ -Penalization. Journal of Computational and Graphical Statistics, 23, 460–477.

Seu L, Tidwell C, Timares L, Duverger A, Wagner FH, Goepfert PA, et al. (2017) CD151 expression is associated with a hyperproliferative T cell phenotype. The Journal of Immunology, 199, 3336–3347. [PubMed: 28954890]

Takeda K, Kaisho T and Akira S (2003) Toll-like receptors. Annual review of immunology, 21, 335–376.

Tan BM, Zammit NW, Yam AO, Slattery R, Walters SN, Malle E, et al. (2013) Baculoviral inhibitors of apoptosis repeat containing (BIRC) proteins fine-tune TNF-induced nuclear factor $κ$B and c-Jun N-terminal kinase signalling in mouse pancreatic beta cells. Diabetologia, 56, 520–532. [PubMed: 23250032]

Taylor BD, Zheng X, Darville T, Zhong W, Konganti K, Abiodun-Ojo O, et al. (2017) Whole-exome sequencing to identify novel biological pathways associated with infertility following pelvic inflammatory disease. Sexually transmitted diseases, 44, 35. [PubMed: 27898568]

Tibshirani R (1996) Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288.

Vaida F and Xu R (2000) Proportional hazards model with random effects. Statistics in medicine, 19, 3309–3324. [PubMed: 11122497]

Wu MC, 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. American Journal of Human Genetics, 89, 82–93. [PubMed: 21737059]

Ying S, Christian JG, Paschen SA and Häcker G (2008) Chlamydia trachomatis can protect host cells against apoptosis in the absence of cellular Inhibitor of Apoptosis Proteins and Mcl-1. Microbes and infection, 10, 97–101. [PubMed: 18069034]

Zheng X, O'Connell CM, Zhong W, Nagarajan UM, Tripathy M, Russell AN, et al. (2018) Discovery of blood transcriptional endotypes in women with pelvic inflammatory disease. The Journal of Immunology, 200, 2941–2956. [PubMed: 29531169]

Zhong W, Spracklen CN, Mohlke KL, Zheng X, Fine J and Li Y (2019) Multi-SNP mediation intersection-union test. Bioinformatics.

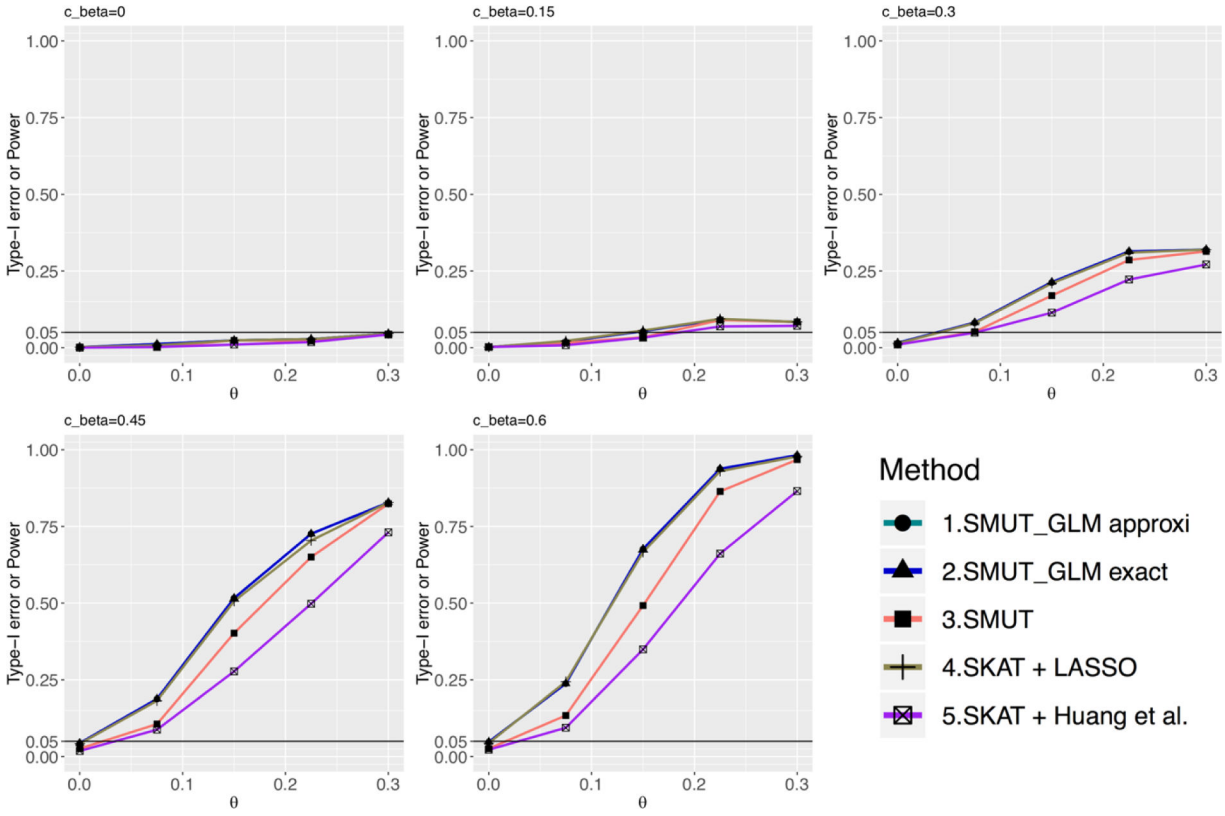

Figure 1.

For binary outcome, power and type-I error under sparse causal SNPs scenario. The x-axis is the true mediator effect (θ) on the outcome. The y-axis is the power or type-I error. Sub-figures vary in c_β value. $c_\beta = 0$ (top-left sub-figure) or $\theta = 0$ (left-most points in each sub-figure) are null settings where y-axis represents the corresponding type-I error. When $c_\beta \neq 0$ and $\theta \neq 0$, it is under alternative hypothesis and y-axis represents the corresponding power. Line for the approximated version of SMUT_GLM is overlapped with the exact version.

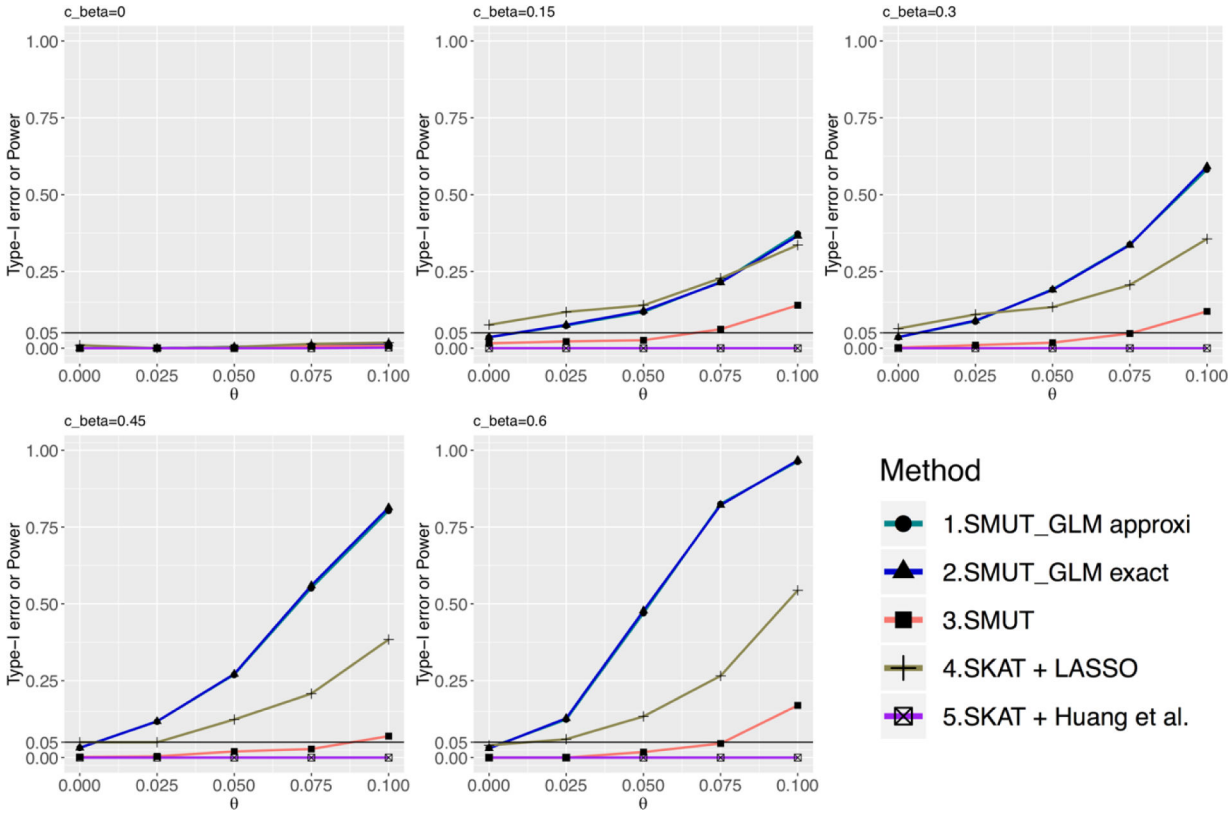

Figure 2.

For binary outcome, power and type-I error under dense causal SNPs scenario. X-axis and y-axis are the same as in Figure 1. Line for the approximated version of SMUT_GLM is overlapped with the exact version.

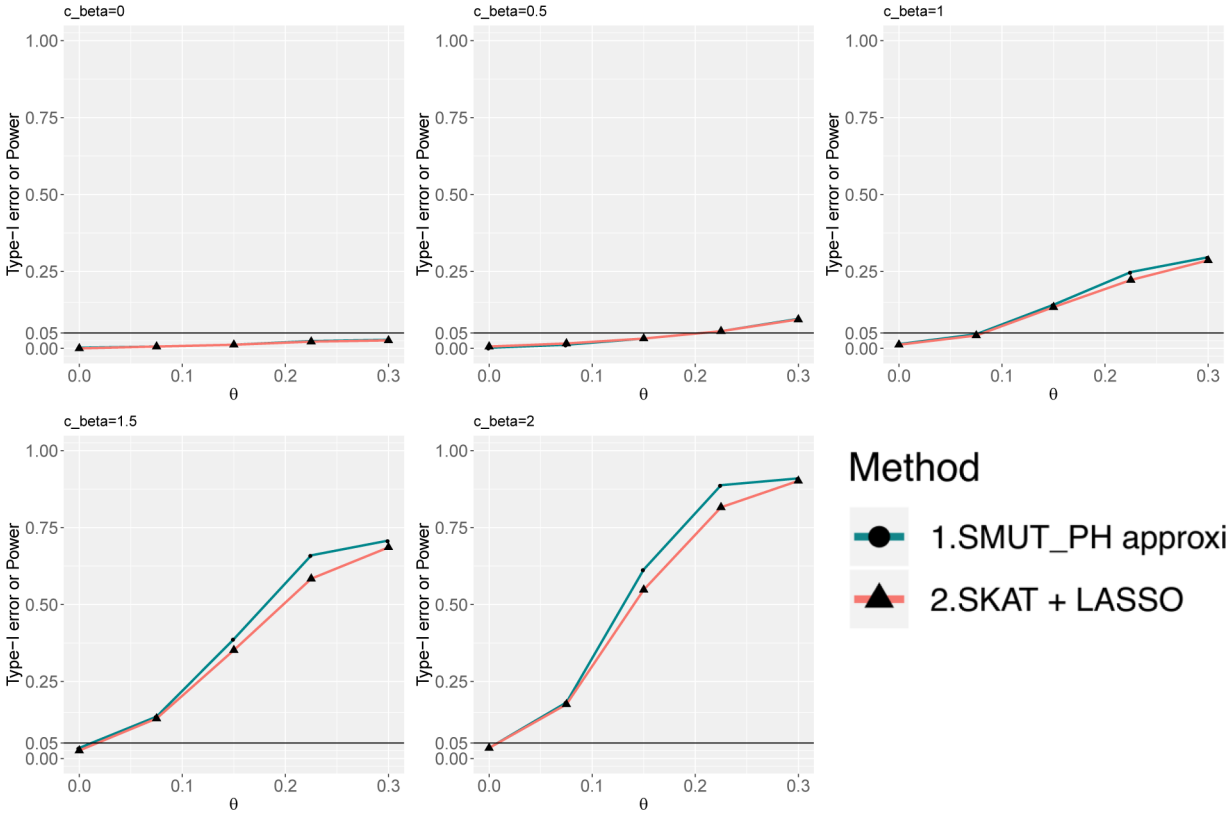

Figure 3.

For time-to-event outcome, power and type-I error under sparse causal SNPs scenario. X-axis and y-axis are the same as in Figure 1.

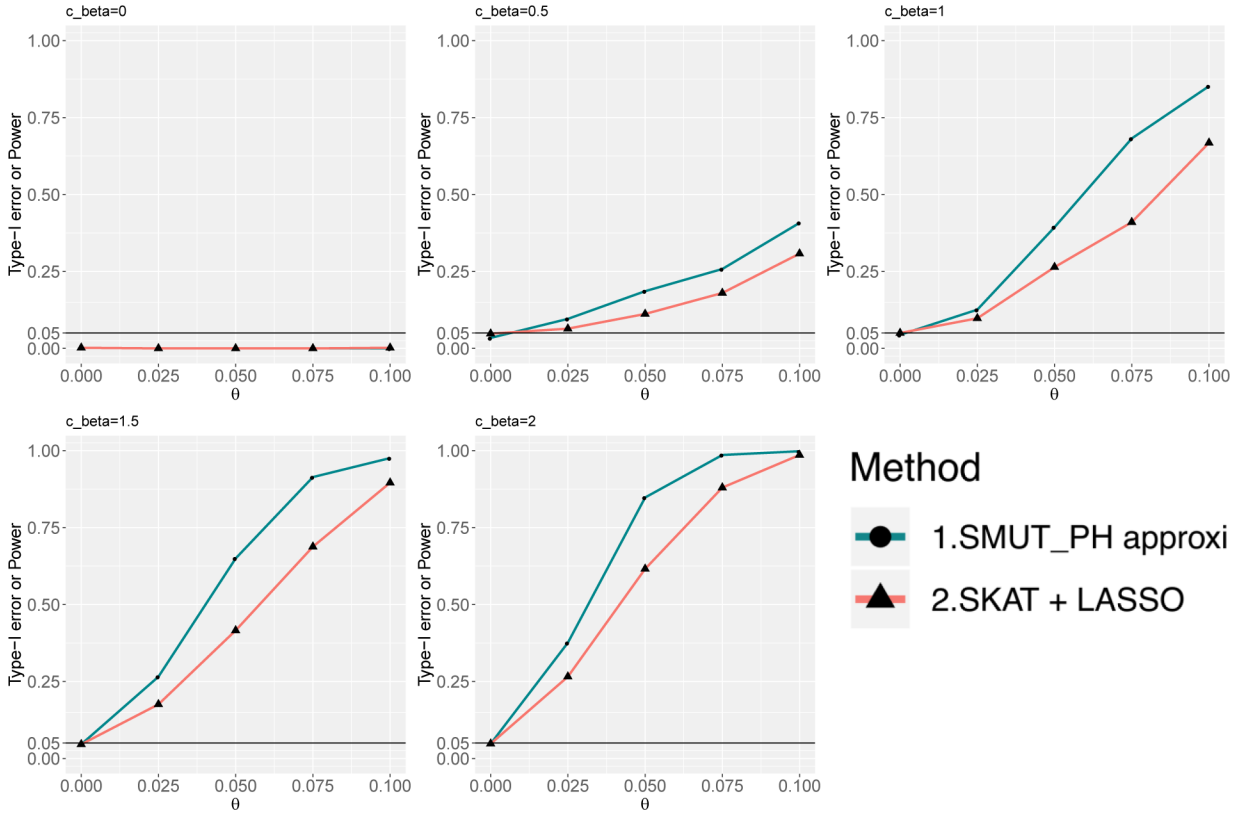

Figure 4.

For time-to-event outcome, power and type-I error under dense causal SNPs scenario. X-axis and y-axis are the same as in Figure 1.

Table 1.

Causal SNP composition in the two simulated scenarios.

Type of outcome	Sample size	Sparse or dense	# causal SNPs	# *sSNPs*	# *mSNPs*	# *oSNPs*	# non-causal SNPs
Binary or Count	1000	Sparse	10	4	3	3	890
		Dense	500	300	100	100	400
Time-to-event	200	Sparse	10	4	3	3	190
		Dense	150	90	30	30	50

Table 2.

Real data application

				***P* values**		
Type of outcome	**Gene**	**Probesets**	**#SNPs**	**SMUT_GLM**	**LASSO**	**Huang et al.**
Binary	*SOS1*	2140519	83	0.0235	0.0691	0.0229
Binary	*CD151*	1940132	40	0.0245	0.1192	0.2289
				SMUT_PH	LASSO	Huang et al.
Time-to-event	*BIRC3*	7210154	4	0.001	0.001	0.002