Supplementary Data to "Multi-SNP Mediation Intersection-Union Test"

Table of contents:

Supplementary Text

Section 1. Decomposition of null hypothesis using intersection-union test (IUT)

We test the null hypothesis $\mathrm{H}_0: \beta\theta = 0$. If we have only one genetic variant, then $\beta\theta$ would be a scalar and the classic methods for testing mediation effects, such as the Sobel test, under the framework of Baron and Kenny can be applied. Since we focus on the joint (from multiple genetic variants) mediation effects, $\beta\theta$ is thus a vector in our setup. The null hypothesis again is $\mathrm{H}_0: \beta\theta = 0$, versus the alternative hypothesis $\mathrm{H}_1: \beta\theta \neq 0$. The hypothesis is divided into two sub-hypotheses, $\mathrm{H}_0^{\theta}: \theta = 0$ versus $\mathrm{H}_1^{\theta}: \theta \neq 0$ and $\mathrm{H}_0^{\beta}: \beta = 0$ versus $\mathrm{H}_1^{\beta}: \beta \neq 0$. Thus, we have

$$\mathrm{H}_0 = \mathrm{H}_0^{\theta} \cup \mathrm{H}_0^{\beta} \quad (1)$$

$$\mathrm{H}_1 = \mathrm{H}_1^{\theta} \cap \mathrm{H}_1^{\beta} \quad (2)$$

This can be conveniently solved by the intersection-union test (IUT). Suppose the *p* value for testing H_0^{θ} versus H_1^{θ} is p_1; and the *p* value for testing H_0^{β} versus H_1^{β} is p_2. Then the *p* value for testing the overall H_0 versus H_1 applying IUT is the maximum of p_1 and

p_2. In the following sections, we use the SMUT strategy to test θ and β separately to obtain p_1 and p_2.

Section 2. Testing θ in the Outcome Model

The outcome model is also high dimensional with multiple genetic effects and the mediator. Classic regression models tend to fail for such models. As a solution, we employ the following mixed effects model to reduce the dimension of parameters.

$$\begin{cases} \gamma_j \sim_{i.i.d.} N(\mu_\gamma, \sigma_\gamma^2) \\ \epsilon_i \sim_{i.i.d.} N(0, \sigma_\epsilon^2) \\ Y_i|(\gamma_1, \dots, \gamma_q, G) = \alpha_1 + M_i\theta + \Sigma_{j=1}^{q} G_{ij}\gamma_j + \epsilon_i \end{cases} \quad (3)$$

We first write out the log-likelihood function for model (3) and then derive the Rao's score statistic (Radhakrishna Rao and Bartlett, 1948; Engle, 1984) for testing θ. Next, we apply Expectation–maximization (EM) algorithm to obtain maximum likelihood estimate (MLE) under the null hypothesis (Dempster *et al.*, 1977; McCulloch *et al.*, 2008). Finally, the score statistic is evaluated at MLE.

The log-likelihood for outcome Y is

$$\ell_Y \coloneqq -\frac{1}{2}\log(\det(2\pi V))$$

$$-\frac{1}{2}\left(Y - \alpha_1 1_n - M\theta - G1_q\mu_\gamma\right)^T V^{-1}\left(Y - \alpha_1 1_n - M\theta - G1_q\mu_\gamma\right) \quad (4)$$

where $V \coloneqq Cov(Y) = \sigma_\gamma^2 GG^T + \sigma_\epsilon^2 I$ and $1_k \coloneqq (1,1,\dots,1)^T$ is a vector of k copies of 1.

The Rao's score statistic for testing θ is

$$SC(\theta) = \frac{\left[\frac{\partial \ell_Y}{\partial \theta}\right]^2}{Fisher(\theta)} \tag{5}$$

where $Fisher(\theta) = E\left(-\frac{\partial^2 \ell_Y}{\partial \theta^2}\right) - E\left(-\frac{\partial^2 \ell_Y}{\partial \theta \partial \xi}\right)^T \left[E\left(-\frac{\partial^2 \ell_Y}{\partial \xi \partial \xi^T}\right)\right]^{-1} E\left(-\frac{\partial^2 \ell_Y}{\partial \theta \partial \xi}\right), \xi = \left(\alpha_1, \mu_\gamma, \sigma_\gamma^2, \sigma_\epsilon^2\right)^T$

Derivations can be found in (McCulloch *et al.*, 2008). The following are the first and second derivatives of ℓ_Y.

First derivatives

$$\frac{\partial \ell_Y}{\partial \theta} = \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)^T V^{-1} M \tag{6}$$

$$\frac{\partial \ell_Y}{\partial \alpha_1} = \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)^T V^{-1} 1_n \tag{7}$$

$$\frac{\partial \ell_Y}{\partial \mu_\gamma} = \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)^T V^{-1} G 1_q \tag{8}$$

$$\begin{aligned} \frac{\partial^2 \ell_Y}{\partial \sigma_\epsilon^2} = -\frac{1}{2}\Big[& tr(V^{-1}) \\ & - \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)^T V^{-2} \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)\Big] \end{aligned} \tag{9}$$

$$\begin{aligned} \frac{\partial^2 \ell_Y}{\partial \sigma_\gamma^2} = -\frac{1}{2}\Big[& tr(V^{-1} G G^T) \\ & - \left(Y - \alpha_1 1_n - M\theta - G1_q \mu_\gamma\right)^T V^{-1} G G^T V^{-1} \left(Y - \alpha_1 1_n - M\theta \right. \\ & \left. - G1_q \mu_\gamma\right)\Big] \end{aligned} \tag{10}$$

Expected value of second derivatives

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \theta^2}\right) = M^T V^{-1} M \tag{11}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \alpha_1^2}\right) = 1_n^T V^{-1} 1_n \tag{12}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \mu_\gamma^2}\right) = \left(G1_q\right)^T V^{-1} G1_q \tag{13}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \left(\sigma_\gamma^2\right)^2}\right) = \frac{1}{2} tr(V^{-1} G G^T V^{-1} G G^T) \tag{14}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial (\sigma_\epsilon^2)^2}\right) = \frac{1}{2} tr(V^{-2}) \tag{15}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \theta \partial \alpha_1}\right) = 1_n^T V^{-1} M \tag{16}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \theta \partial \mu_\gamma}\right) = \left(G1_q\right)^T V^{-1} M \tag{17}$$

$$E\left(-\frac{\partial^2 \ell_{Y,Z}}{\partial \theta \partial \sigma_\gamma^2}\right) = 0 \tag{18}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \theta \partial \sigma_\epsilon^2}\right) = 0 \tag{19}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \alpha_1 \partial \mu_\gamma}\right) = \left(G1_q\right)^T V^{-1} 1_n \tag{20}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \alpha_1 \partial \sigma_\gamma^2}\right) = E\left(-\frac{\partial^2 \ell_Y}{\partial \alpha_1 \partial \sigma_\epsilon^2}\right) = 0 \tag{21}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \mu \partial \sigma_\gamma^2}\right) = E\left(-\frac{\partial^2 \ell_Y}{\partial \mu \partial \sigma_\epsilon^2}\right) = 0 \tag{22}$$

$$E\left(-\frac{\partial^2 \ell_Y}{\partial \sigma_\gamma^2 \partial \sigma_\epsilon^2}\right) = \frac{1}{2} tr(V^{-2} G G^T) \tag{23}$$

Under the null hypothesis $\theta = 0$, this score statistic $SC(\theta)$ asymptotically follows a Chi-squared distribution with one degree of freedom when MLE under the null is plugged in. This assumes at least some of the direct effects $\gamma_j (j = 1,2,\dots,q)$ are nonzero. When there are no direct effects, the variance component σ_γ^2 is on the boundary. The

asymptotic Chi-square distribution works well in simulations (Supplementary Fig. S1 and S2).

We leverage the EM algorithm to obtain MLE under the null. When applying EM algorithm to mixed effects model, random effects γ are treated as missing data. The complete data comprise the observed outcome data and random effects. The log-likelihood for complete data (Y, γ) is

$$LL(Y,\gamma|G;\xi) \coloneqq \log[p(Y,\gamma|G;\xi)] = \log[p(Y|\gamma,G;\xi)] + \log[p(\gamma|G;\xi)] \tag{24}$$

$$\begin{aligned} &= -\frac{n}{2}\log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}(Y-\alpha_1-G\gamma)^T(Y-\alpha_1-G\gamma) - \frac{q}{2}\log\left(2\pi\sigma_\gamma^2\right) \\ &\quad -\frac{1}{2\sigma_\gamma^2}\left(\gamma-\mu_\gamma\right)^T\left(\gamma-\mu_\gamma\right) \end{aligned} \tag{25}$$

where $\xi = \left(\alpha_1, \mu_\gamma, \sigma_\gamma^2, \sigma_\epsilon^2\right)^T$

Derivations for E-step and M-step can be found in (McCulloch *et al.*, 2008).

E-step of EM algorithm is

$$\hat{\eta}^{(t)} = E(\gamma|Y) = 1_q\mu_\gamma^{(t)} + \sigma_\gamma^{2(t)}G^T{V^{(t)}}^{-1}\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right) \tag{26}$$

$$\begin{aligned} &E\left(\left(\gamma-\mu_\gamma\right)^T\left(\gamma-\mu_\gamma\right)|Y\right) \\ &= q\sigma_\gamma^{2(t)} + \sigma_\gamma^{4(t)}\left\{\left(Y-\alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)^T {V^{(t)}}^{-1}GG^T{V^{(t)}}^{-1}\left(Y-\alpha_1^{(t)}1_n\right.\right. \\ &\quad \left.\left. - G1_q\mu_\gamma^{(t)}\right) - tr(G^TV^{-1}G)\right\} \end{aligned} \tag{27}$$

$$E((Y-\alpha_1-G\gamma)^T(Y-\alpha_1-G\gamma)|Y) = E(\epsilon^T\epsilon|Y) \tag{28}$$

$$= n\sigma_\epsilon^{2(t)} + \sigma_\epsilon^{4(t)}\left\{\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)^T V^{(t)^{-1}}V^{(t)^{-1}}\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)\right.$$

$$\left. - tr(V^{-1})\right\}$$

M-step of EM algorithm is

$$\alpha_1^{(t+1)} = E\left(\frac{1}{n}\Sigma_{i=1}^n\left[Y_i - \Sigma_{j=1}^q G_{ij}\gamma_j\right]|Y\right) = \frac{1}{n}\Sigma_{i=1}^n\left[Y_i - \Sigma_{j=1}^q G_{ij}\hat{\eta}_j^{(t)}\right] \quad (29)$$

$$\mu_\gamma^{(t+1)} = E\left(\frac{1}{q}\Sigma_{j=1}^q\gamma_j|Y\right) = \frac{1}{q}\Sigma_{j=1}^q\hat{\eta}_j^{(t)} \quad (30)$$

$$\sigma_\gamma^{2(t+1)} = E\left(\frac{1}{q}(\gamma - \mu_\gamma)^T(\gamma - \mu_\gamma)|Y\right)$$

$$= \sigma_\gamma^{2(t)} + \frac{1}{q}\sigma_\gamma^{4(t)}\left\{\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)^T V^{(t)^{-1}}GG^TV^{(t)^{-1}}\left(Y - \alpha_1^{(t)}1_n\right.\right. \quad (31)$$

$$\left.\left. - G1_q\mu_\gamma^{(t)}\right) - tr(G^TV^{-1}G)\right\}$$

$$\sigma_\epsilon^{2(t+1)} = E\left(\frac{1}{n}(Y - \alpha_1 - G\gamma)^T(Y - \alpha_1 - G\gamma)\middle|Y\right)$$

$$= \sigma_\epsilon^{2(t)} + \frac{1}{n}\sigma_\epsilon^{4(t)}\left\{\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)^T V^{(t)^{-1}}V^{(t)^{-1}}\left(Y - \alpha_1^{(t)}1_n - G1_q\mu_\gamma^{(t)}\right)\right. \quad (32)$$

$$\left. - tr(V^{-1})\right\}$$

Convergence criterion for EM algorithm is

$$\max\left(\left|\sigma_\gamma^{2(t+1)} - \sigma_\gamma^{2(t)}\right|, \left|\sigma_\epsilon^{2(t+1)} - \sigma_\epsilon^{2(t)}\right|\right) \leq 1 \times 10^{-6} \quad (33)$$

If convergence is not reached, iteration stops when the number of iterations exceeds a pre-specified large number.

As for the starting values of EM algorithm, the intercept α_1 is randomly generated from uniform distribution $Unif(-1,1)$. And μ_γ is also randomly generated from uniform

distribution $Unif(-1,1)$. The variance components σ_γ^2 and σ_ϵ^2 are independently generated from uniform distribution $Unif(0,1)$.

Section 3. Robustness with Alternative Testing Strategies

As aforementioned, the true causal SNPs were drawn from common (MAF ≥ 1%) SNPs and by default all common SNPs were simultaneously modeled and tested. Thus the set of testing SNPs include all the causal SNPs. Alternatively, we considered two other testing strategies: (1) eQTL SNPs only; and (2) SNPs with MAF ≥ 5% only. Under (1), our observations above regarding Type-I error and power remained largely the same: namely SMUT remained valid and more powerful than alternative methods (Supplementary Fig. S3 and S4). In addition, adapted Huang et al. was more powerful using testing strategy (1) than testing all common SNPs in the default setting in most scenarios. For example, with sparse causal SNPs and $c_\beta = 0.05$, $\theta = 0.15$, adapted Huang et al. had 25% and 96% power using the default and testing strategy (1) while SMUT had 36% and 97% power (Supplementary Fig. S5).

Because SMUT and adapted Huang et al. had protected Type-I error, we evaluated only their performance under alternative setting (2). Using testing strategy (2) where only SNPs with MAF ≥ 5% were tested, both SMUT and adapted Huang et al. had inflated Type-I error (Supplementary Fig. S6 and S7). This might be due to the violation of confounding assumptions for mediation analysis (VanderWeele, 2016), because shared SNPs became mediator-outcome confounders when absent in models.

Section 4. All the graphs are generated using the R package ggplot2 (Wickham, 2016) and RColorBrewer (Harrower and Brewer, 2003).

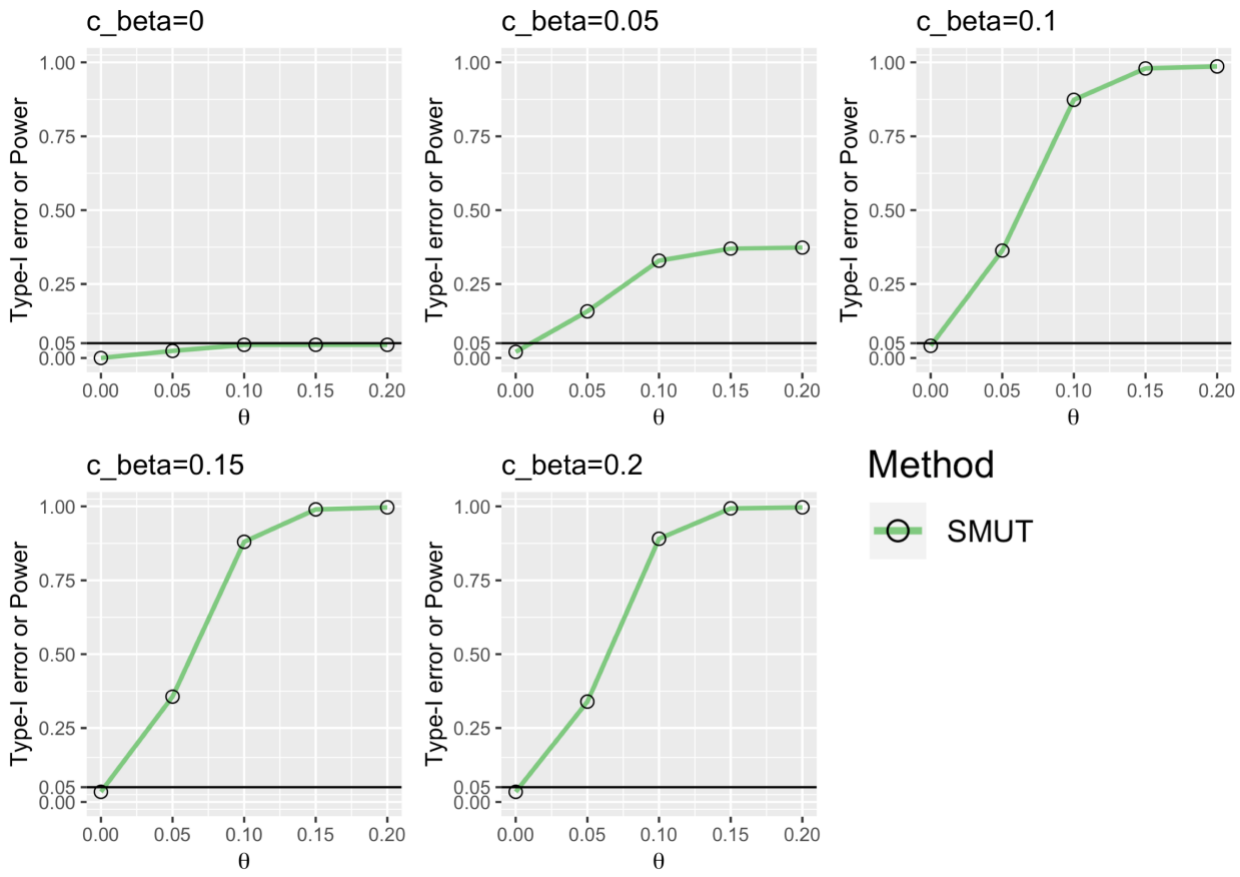

Fig. S1. Power and Type-I error under sparse causal SNPs scenario and no direct effects (γ_j is zero, $j = 1, 2, \ldots, q$). X-axis and y-axis are the same as in **Figure 2**. The candidate SNPs tested in the mediator and outcome model are the 2,891 SNPs with MAF ≥ 1%.

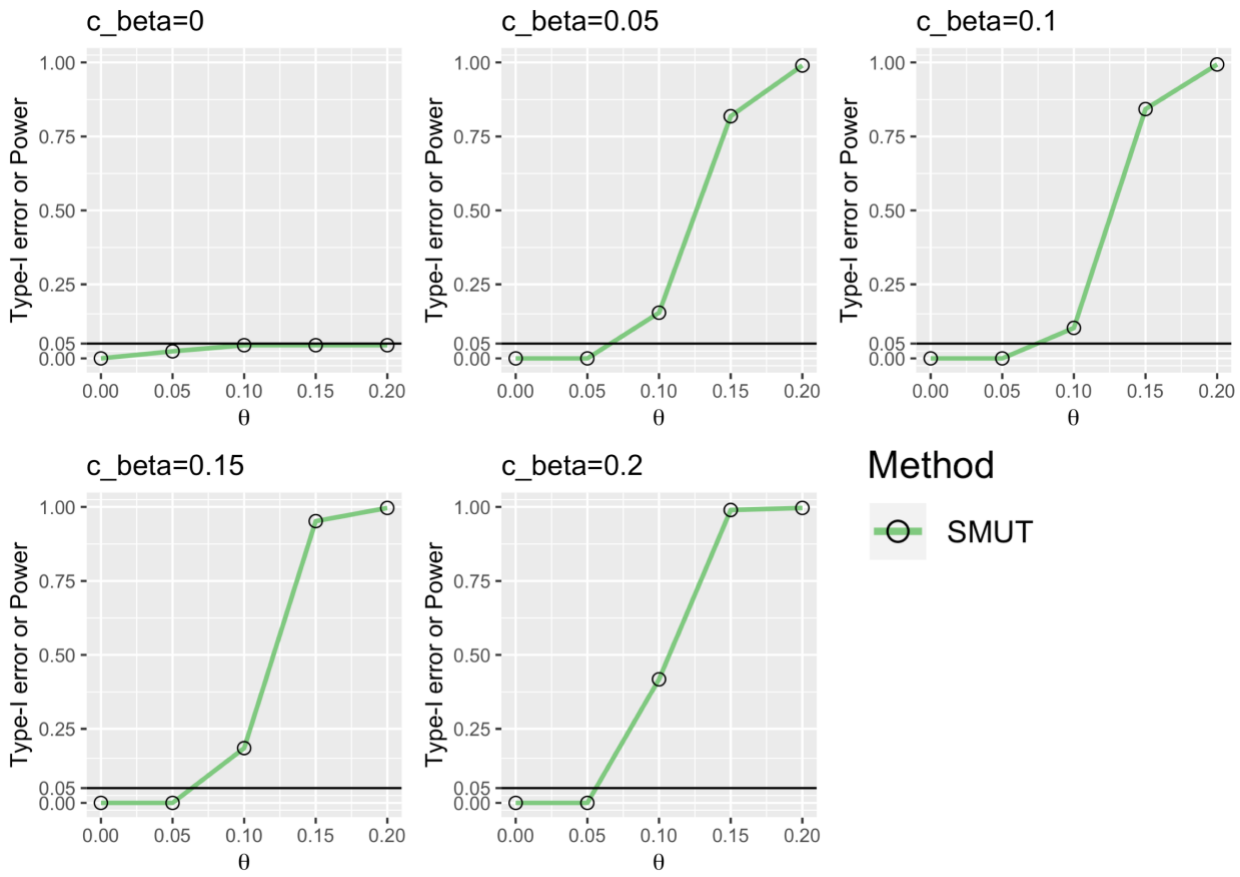

Fig. S2. Power and Type-I error under dense causal SNPs scenario and no direct effects (γ_j is zero, $j = 1, 2, \ldots, q$). X-axis and y-axis are the same as in **Figure 2**. The candidate SNPs tested in the mediator and outcome model are the 2,891 SNPs with MAF ≥ 1%.

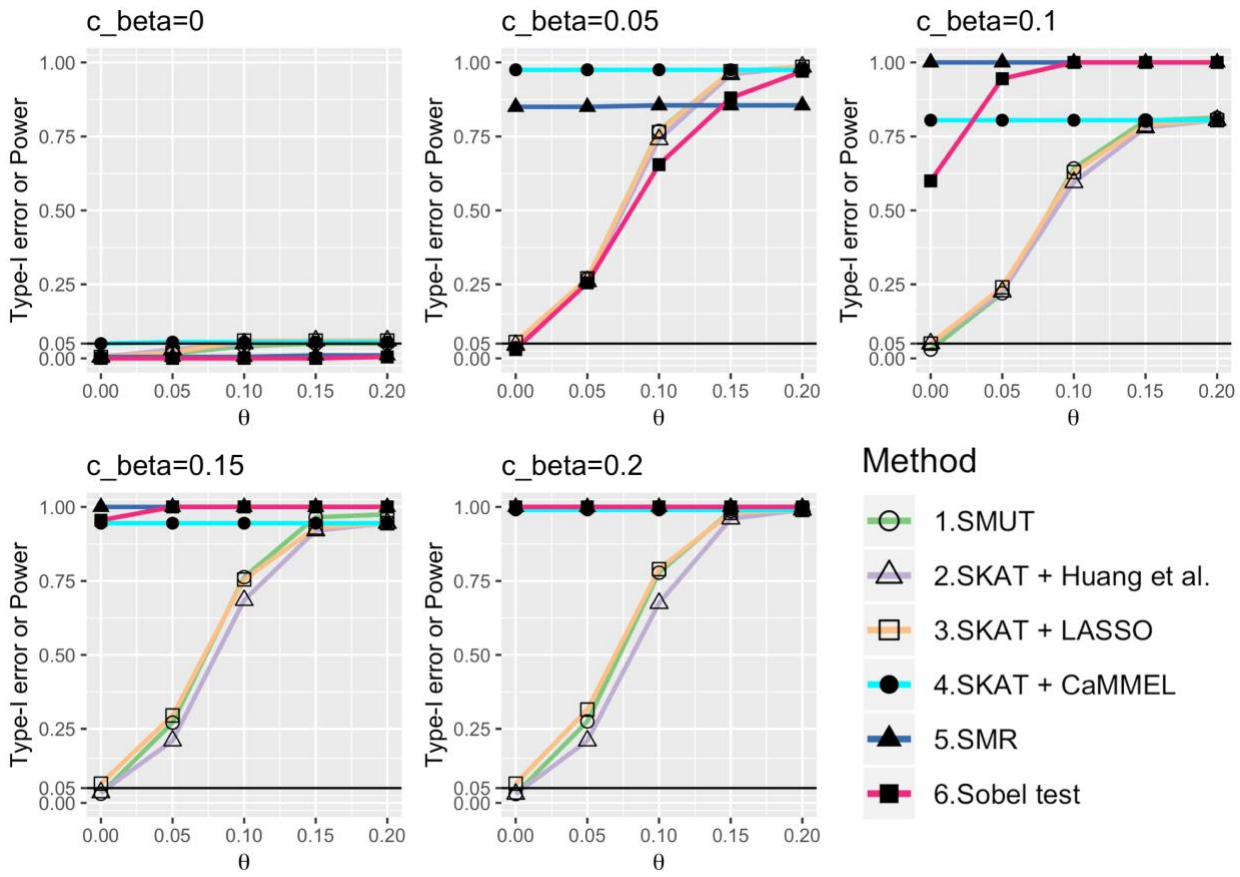

Fig. S3. Power and Type-I error under sparse causal SNPs scenario and alternative setting (1) testing eQTL SNPs only. X-axis and y-axis are the same as in **Figure 2**.

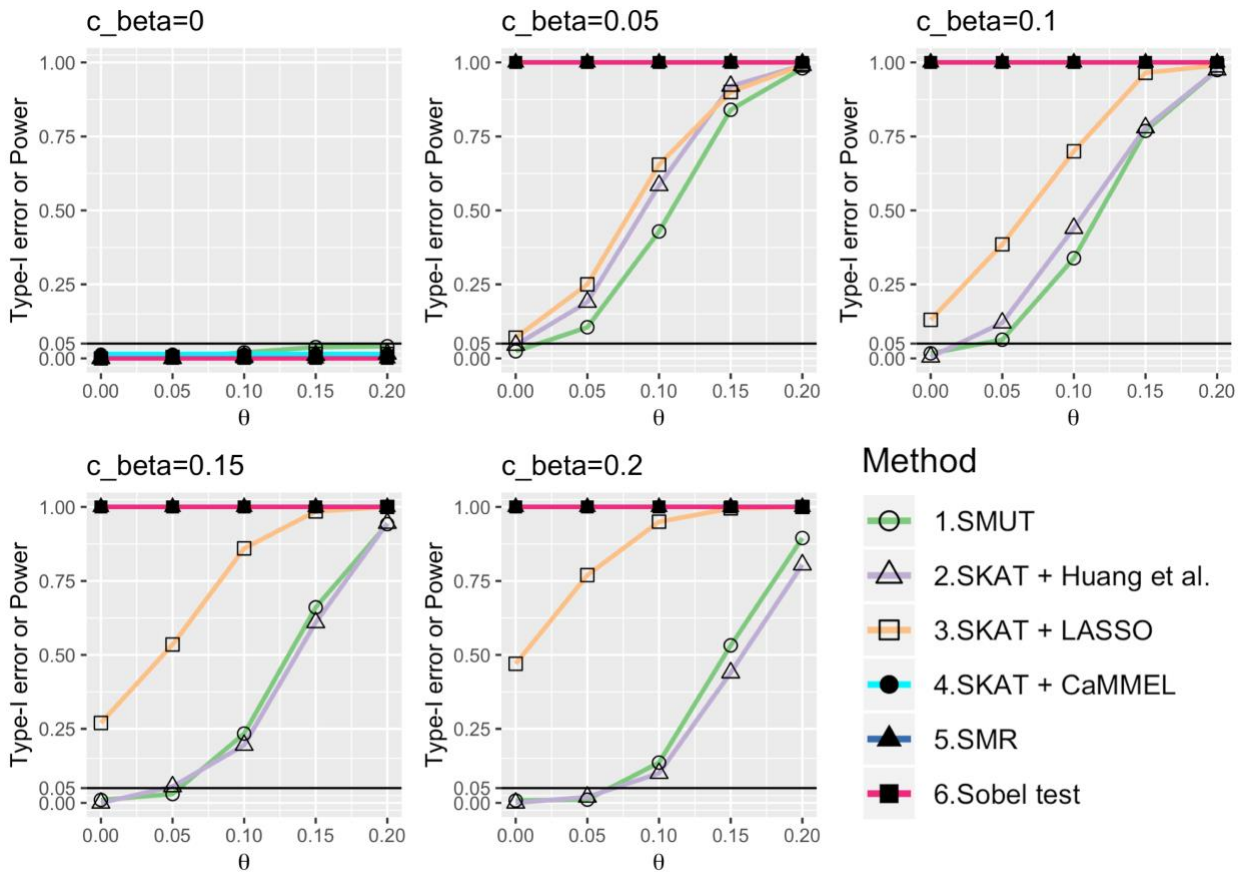

Fig. S4. Power and Type-I error under dense causal SNPs scenario and alternative setting (1) testing eQTL SNPs only. X-axis and y-axis are the same as in **Figure 2**.

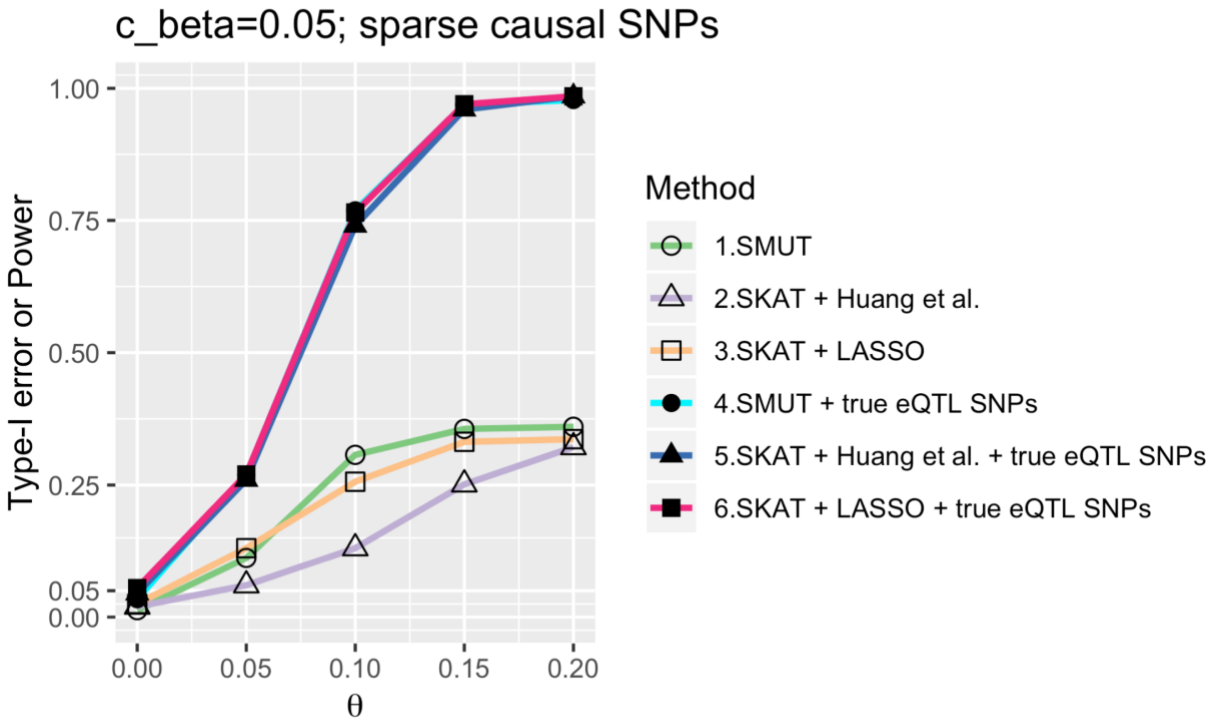

Fig. S5. **Example of power gain when the true eQTL SNPs are known.** Under this situation, with the knowledge on eQTL SNPs helps increase power for SMUT, adaptive LASSO and adaptive Huang et al.'s method.

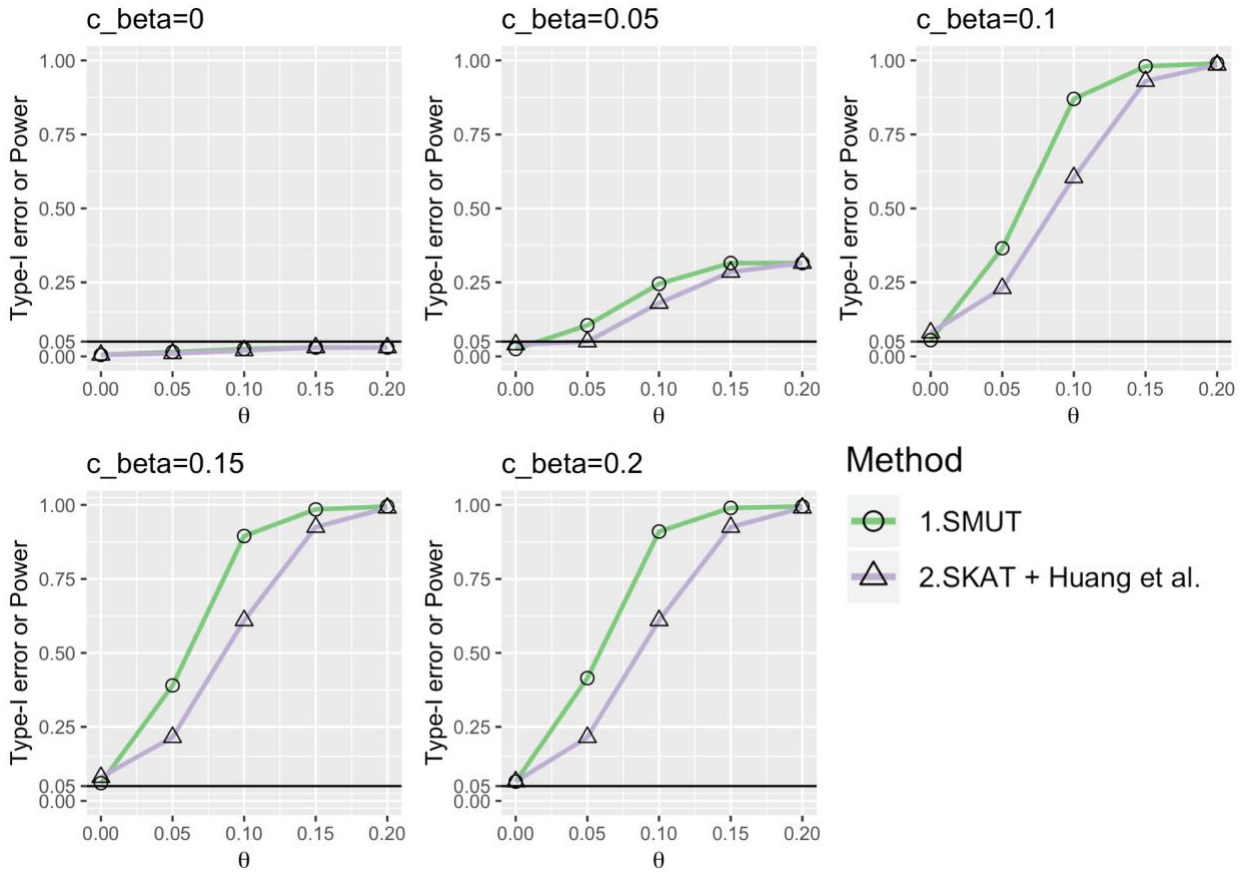

Fig. S6. Power and Type-I error under sparse causal SNPs scenario and alternative setting (2) testing SNPs with MAF ≥ 5% only. X-axis and y-axis are the same as in **Figure 2**.

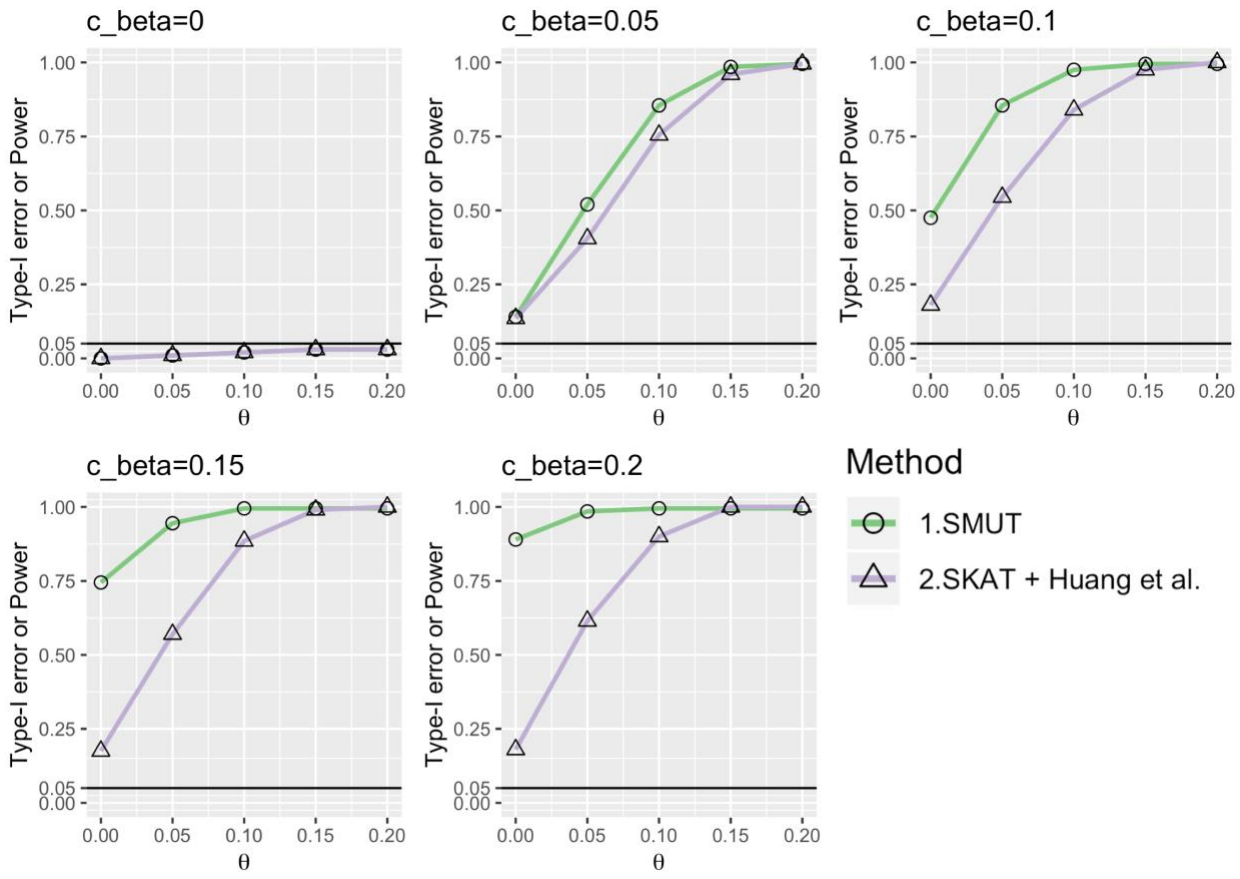

Fig. S7. Power and Type-I error under dense causal SNPs scenario and alternative setting (2) testing SNPs with MAF ≥ 5% only. X-axis and y-axis are the same as in **Figure 2**.

References

Dempster,A.P. *et al.* (1977) Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc. Ser. B*, **39**, 1–38.

Engle,R.F. (1984) Wald, likelihood ratio, and Lagrange multiplier tests in econometrics. *Handb. Econom.*, **2**, 775–826.

Harrower,M. and Brewer,C.A. (2003) ColorBrewer. org: an online tool for selecting colour schemes for maps. *Cartogr. J.*, **40**, 27–37.

McCulloch,C.E. *et al.* (2008) Generalized, Linear, and Mixed Models, 2nd Edition. 424.

Radhakrishna Rao,C. and Bartlett,M.S. (1948) Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math. Proc. Cambridge Philos. Soc.*, **44**, 50.

VanderWeele,T.J. (2016) Mediation Analysis: A Practitioner's Guide. *Annu. Rev. Public Health*, **37**, 17–32.

Wickham,H. (2016) ggplot2: elegant graphics for data analysis Springer.