Methods for Large-scale Single Mediator Hypothesis Testing: Possible Choices and Comparisons

Jiacong Du[1], Xiang Zhou[1], Wei Hao[1], Yongmei Liu[2],
Jennifer A. Smith[3] and Bhramar Mukherjee[1]

[1] Department of Biostatistics, University of Michigan, Ann Arbor, MI

[2] Department of Medicine, Divisions of Cardiology and Neurology,
Duke University Medical Center, Durham, NC

[3] Department of Epidemiology, University of Michigan, Ann Arbor, MI

Abstract

Mediation hypothesis testing for a large number of mediators is challenging due to the composite structure of the null hypothesis, $H_0 : \alpha\beta = 0$ (α: effect of the exposure on the mediator after adjusting for confounders; β: effect of the mediator on the outcome after adjusting for exposure and confounders). In this paper, we reviewed three classes of methods for multiple mediation hypothesis testing. In addition to these existing methods, we developed the Sobel-comp method, which uses a corrected mixture reference distribution for Sobel's test statistic. We performed extensive simulation studies to compare all six methods in terms of the false positive rates under the null hypothesis and the true positive rates under the alternative hypothesis. We found that the class of methods which uses a mixture reference distribution could best maintain the false positive rates at the nominal level under the null hypothesis and had the greatest true positive rates under the alternative hypothesis. We applied all methods to study the mediation mechanism of DNA methylation sites in the pathway from adult socioeconomic status to glycated hemoglobin level using data from the Multi-Ethnic Study of Atherosclerosis (MESA). We also provide guidelines for choosing the optimal mediation hypothesis testing method in practice. (word count: 196)

Keywords: Agnostic mediation analysis; Composite null hypothesis; Indirect effect; Mediation effect; Multiple hypothesis testing.

1 Introduction

Mediation analysis is often used to identify potential mechanistic pathways of the effect of an exposure on an outcome. It becomes increasingly popular in recent decades in epidemiology [1, 2, 3, 4, 5]. With the advances of high-throughput technologies in genomics studies, mediation analysis often requires analyzing a large number of potential mediators [6, 7]. These agnostic explorations of high-dimensional mediators allow researchers to investigate molecular traits associated with complex diseases that may be a result of socioeconomic inequalities, environmental pollution, or other exogenous factors. In particular, molecular epidemiological research has frequently considered the mediating role of DNA methylation (DNAm), and mounting studies have identified methylation differences at CpG sites as important mediators for various diseases such as cancer [8, 9, 10], cardiovascular disease [11] and diabetes [12].

Suppose there is a total number of J mediators potentially mediating the effect of an exposure X on the outcome Y. Let M_j denote the j-th mediator where $j \in \{1, 2, ..., J\}$. To identify which M_j's are truly in the mediating pathways, one can jointly model $M_1, M_2, ..., M_J$ [13, 14, 15]. However, the computational burden may be too great and the solution may not be robust for large J but modest sample sizes. Therefore, people often use the traditional univariate mediation analysis which examines one mediator at a time. This is often performed based on the parametric models proposed by Baron & Kenny [16]. For $j \in \{1, 2, ..., J\}$, the two regression models involved in a mediation analysis with continuous outcome and continuous mediators are:

$$Y = \beta_{0,j} + \beta_{X,j} X + \beta_j M_j + \boldsymbol{\beta}_{C,j}^\top \boldsymbol{C} + \epsilon_{Y,j} \tag{1}$$

$$M_j = \alpha_{0,j} + \alpha_j X + \boldsymbol{\alpha}_{C,j}^\top \boldsymbol{C} + \epsilon_{M,j} \tag{2}$$

where $\boldsymbol{C}$ is the set of potential confounders and $\epsilon_{Y,j} \sim N(0, \sigma_{Y,j}^2)$ and $\epsilon_{M,j} \sim N(0, \sigma_{M,j}^2)$ are indepen-
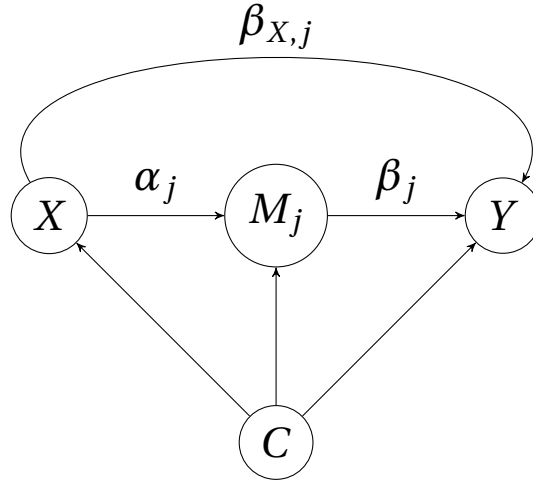
Figure 1: A causal diagram for mediation analysis. For $j = 1, 2, ..., J$, X is the exposure, M_j is the j-th mediator, Y is the outcome, C is the set of confounders. α_j is the effect of X on M_j after adjusting for C. β_j is the effect of M_j on Y after adjusting for (X, C). $\beta_{X,j}$ is the direct effect of X on Y after adjusting for M_j and C.

dent. The counterfactual framework, also called the potential outcome framework [1, 17], is often used to define a causal mediation effect with certain accompanying assumptions [18, 19, 20], including 1) no unmeasured confounders for the exposure-outcome relationship conditional on C; 2) no unmeasured confounders for the mediator-outcome relationship conditional on (X, C); 3) no unmeasured confounders for the exposure-mediator relationship conditional on C; 4) no mediator-outcome confounders affected by X. Under the counterfactual framework with assumptions 1-4, the effect of the exposure on the outcome is decomposed into *direct effect* and *indirect effect* (also called *mediation effect*). In addition, if there is no exposure-mediator interaction affecting the outcome, the causal estimate of the indirect effect is the same as the classical product estimate proposed by Baron & Kenny [16]. A causal diagram for illustrating the role of the j-th mediator is presented in **Figure 1**.

To test whether M_j is mediating the effect of X on Y, the underlying null and alternative hypotheses can be stated as:

$$H_{0,j} : \alpha_j \beta_j = 0 \quad vs. \quad H_{1,j} : \alpha_j \beta_j \neq 0, \quad for\ j = 1, 2, ..., J$$

Since $H_{0,1}, ..., H_{0,J}$ are tested in a similar manner, we drop the subscript j for now.

The first class of methods contains Sobel's test [21] and the MaxP test [22]. The null hypothesis involving the product of parameters is composite [23] and consists of three cases, namely, 1) H_{01} : $\alpha = 0, \beta \neq 0$; 2) $H_{10} : \alpha \neq 0, \beta = 0$; and 3) $H_{00} : \alpha = \beta = 0$. Since Sobel's test and the MaxP test do not consider this composite structure, they are conservative [23, 24], especially in high-dimensional settings where the majority of mediators are likely to have no mediation effect.

Many recent studies have developed univariate analysis methods to produce calibrated p-values that consider the composite null structure. Huang *et al.* proposed the joint significance test under the composite null hypothesis (JT-comp) that uses the product of two normally distributed variables as the test statistic [25]. Dai *et al.* developed a procedure for high-dimensional mediation hypotheses testing (HDMT) which considered the correct reference distribution for the MaxP statistic [26]. A common feature for these two methods is to weight the reference distribution under H_{01}, H_{10}, H_{00} to form a mixture null distribution corresponding to the test statistic. We group these two methods into the second class.

The third class contains the Divide-Aggregate Composite-null Test (DACT) method proposed by Liu *et al.* In contrast to the second class which form a mixture reference distribution, this method constructs a composite test statistic using the three p-values obtained under H_{01}, H_{10} and H_{00} [24].

However, no study has numerically compared the testing performance of the above-mentioned methods. It remains unclear how these methods would be affected by various factors with high-dimensional mediators, in particular, by the sample size, the proportion of $H_{01}, H_{10}, H_{00}, H_1$ being true, the variation of non-zero α and β across J tests, and the R^2 in the data generating models, i.e. models (1) and (2). Our contribution in this paper is twofold. First, in addition to the existing methods, we develop a new method, called Sobel-comp, which is a variant of HDMT. Sobel-comp uses a corrected mixture reference distribution for Sobel's test statistic utilizing the composite structure of

the null. Second, we perform extensive simulation studies to compare all six methods in three classes in terms of false positive rates under the null hypothesis and true positive rates under the alternative hypothesis.

This paper is organized as follows: In Section 2.1, we first describe the five existing mediation hypothesis testing methods, including Sobel's test, MaxP, JT-comp, HDMT, and DACT, and discuss their potential advantages and limitations. We then propose our new method, Sobel-comp. In Section 2.2, we describe the simulation setup to compare the testing performance of the six methods. In Section 2.3, we describe the analyzing procedure for studying the mediation mechanism of DNAm in the pathway from adult socioeconomic status (SES) to glycated hemoglobin (HbA1c) level using data from the Multi-Ethnic Study of Atherosclerosis (MESA). Numerical results are presented in Section 3. We summarize the key strengths and limitations of each method and provide recommendations for applying these methods in practical settings in Section 4.

2 Methods and Materials

2.1 Methods for mediation hypothesis testing

Mediation hypothesis testing methods are often based on the Wald test statistics obtained from models (1) and (2). Denote Z_β and Z_α as the test statistics for testing $\beta = 0$ in model (1) and for testing $\alpha = 0$ in model (2), respectively. Under the null hypothesis, we have:

$$Z_\alpha = \frac{\widehat{\alpha} - \alpha}{\widehat{\sigma}_\alpha} \sim N(0,1); Z_\beta = \frac{\widehat{\beta} - \beta}{\widehat{\sigma}_\beta} \sim N(0,1)$$

where $\widehat{\alpha}$ and $\widehat{\beta}$ are the maximum likelihood estimates for α and β, respectively. $\widehat{\sigma}_\alpha$ and $\widehat{\sigma}_\beta$ are the estimated standard error of $\widehat{\alpha}$ and $\widehat{\beta}$, respectively. Let the two-sided p-value for Z_α be p_α and for Z_β be p_β.

2.1.1 Sobel's test

Sobel's test statistic [21] uses the first-order multivariate delta method to find the standard error of $\hat{\alpha}\hat{\beta}$. Since $\hat{\alpha}$ and $\hat{\beta}$ derived from models (1) and (2) are independent [21, 27], Sobel's test statistic is defined as:

$$T_{sobel} = \frac{\hat{\beta}\hat{\alpha}}{\sqrt{\widehat{\beta^2\sigma_\alpha^2} + \widehat{\alpha^2\sigma_\beta^2}}} = \frac{Z_\alpha}{\sqrt{1 + (Z_\alpha/Z_\beta)^2}} \tag{3}$$

T_{Sobel} is compared to $N(0,1)$ to determine the p-value. The null distribution of $N(0,1)$ is asymptotically correct under H_{01} and H_{10}, but is incorrect under H_{00}, since the multivariate delta method fails at $(\alpha, \beta) = (0,0)$. Liu *et al.* proved that T_{Sobel} under H_{00} asymptotically follows $N(0,1/4)$ [24]. As a result, Sobel's test, which incorrectly uses $N(0,1)$ under H_{00} as the reference distribution, yields larger p-values than the truth, and thus is conservative.

2.1.2 MaxP test

The MaxP test, also called the joint significance test [22], has been developed based on the idea that if we want to reject H_0 at level t, we should reject two separate hypothesis tests of $\alpha = 0$ and $\beta = 0$ at level t simultaneously. The MaxP test statistic is defined as:

$$p_{max} = max(p_\alpha, p_\beta) \tag{4}$$

p_{max} is compared to $U(0,1)$ to determine the p-value. Equivalently, p_{max} is determined by the smaller $|Z_\alpha|$ or $|Z_\beta|$. Since $min(|Z_\alpha|, |Z_\beta|) > |T_{Sobel}|$ in a finite sample, the MaxP p-value is always smaller than that from Sobel's test and thus is more powerful. However, similar to Sobel's test, the reference distribution of $U(0,1)$ is incorrect under H_{00}. Since $P(p_{max} < t) = P(p_\alpha < t) \cdot P(p_\beta < t) = t^2$, the correct reference distribution for p_{max} under H_{00} is $Beta(2,1)$ [24, 26]. Since the p-value under H_{00} determined by $U(0,1)$ will be larger than that by $Beta(2,1)$, the MaxP test is conservative.

6

2.1.3 Joint significance test under the composite null hypothesis (JT-comp)

We now resume to use the subscript j corresponding to the j-th hypothesis test for $j = 1, 2, ..., J$. The test statistic for JT-comp is the product of two normally distributed random variables, $Z_{\alpha,j} Z_{\beta,j}$ [25]. Unlike Sobel's test and the MaxP test, JT-comp distinguishes the null distributions for its test statistic under $H_{01,j}, H_{10,j}$ and $H_{00,j}$ to obtain case-specific p-values. Specifically, let $w_{01,j}, w_{10,j}, w_{00,j}$ be the probability of $H_{01,j}, H_{10,j}$ and $H_{00,j}$ being true, respectively. Denote $F(t)$ as the two-sided tail probability of the standard normal product distribution evaluated at t. Under $H_{00,j}$, since $Z_{\alpha,j} \sim N(0,1)$ and $Z_{\beta,j} \sim N(0,1)$, the case-specific p-value is $F(Z_{\alpha,j} Z_{\beta,j})$. Under $H_{01,j}$, $Z_{\alpha,j} \sim N(0,1)$ and $Z_{\beta,j} \sim N(\mu_{\beta,j}, 1)$, where $\mu_{\beta,j} = \beta_j / \widehat{\sigma}_{\beta,j} \neq 0$. Huang et $al.$ further assumes that $\mu_{\beta,j}$ follows a symmetric distribution with mean 0 and variance $\delta_{\beta,j}^2$, e.g. $\mu_{\beta,j} \sim N(0, \delta_{\beta,j}^2)$. By integrating out $\mu_{\beta,j}$, the p-value under $H_{01,j}$ is obtained by using the same $F(\cdot)$ function as if under $H_{00,j}$, but only differs by a scaling factor of $1/\sqrt{1 + \delta_{\beta,j}^2}$. That is, the p-value under $H_{01,j}$ is $F(Z_{\alpha,j} Z_{\beta,j} / \sqrt{1 + \delta_{\beta,j}^2})$. Similarly, the p-value under $H_{10,j}$ is $F(Z_{\alpha,j} Z_{\beta,j} / \sqrt{1 + \delta_{\alpha,j}^2})$, where $\delta_{\alpha,j}^2$ is the assumed variance of the mean of $Z_{\alpha,j}$ under $H_{10,j}$. The final composite p-value is aggregated as:

$$p_{JT-comp,j} = w_{01,j} F\left(\frac{Z_{\alpha,j} Z_{\beta,j}}{\sqrt{1 + \delta_{\beta,j}^2}}\right) + w_{10,j} F\left(\frac{Z_{\alpha,j} Z_{\beta,j}}{\sqrt{1 + \delta_{\alpha,j}^2}}\right) + w_{00,j} F\left(Z_{\alpha,j} Z_{\beta,j}\right)$$

$p_{JT-comp,j}$ is then approximated by Taylor series:

$$\hat{p}_{JT-comp,j} = F\left(\frac{Z_{\alpha,j} Z_{\beta,j}}{\sqrt{Var(Z_{\beta,j})}}\right) + F\left(\frac{Z_{\alpha,j} Z_{\beta,j}}{\sqrt{Var(Z_{\alpha,j})}}\right) - F(Z_{\alpha,j} Z_{\beta,j}) \tag{5}$$

where $Var(Z_{\beta,j}) = 1 + w_{01,j} \delta_{\beta,j}^2$ and $Var(Z_{\alpha,j}) = 1 + w_{10,j} \delta_{\alpha,j}^2$. Sample variances of $Z_{\alpha,j}$ and $Z_{\beta,j}$ across all J tests are used to estimate $Var(Z_{\alpha,j})$ and $Var(Z_{\beta,j})$. The advantage of using the approximation $\hat{p}_{JT-comp}$ is to avoid estimating $w_{01,j}, w_{10,j}, w_{00,j}$. Since the reference distribution of $Z_{\alpha,j} Z_{\beta,j}$ is correct under $H_{01,j}, H_{10,j}$ and $H_{00,j}$, JT-comp is more powerful than Sobel's and MaxP tests.

7

However, the accuracy of $p_{JT-comp,j}$ approximated by $\hat{p}_{JT-comp,j}$ depends on the residual error from Taylor series expansion in (5). The residual error relative to the p-value becomes larger when the p-value becomes smaller [25], suggesting that JT-comp cannot maintain the family-wise-error-rate (FWER) at small significance thresholds. A good approximation requires that $\delta_{\alpha,j}^2$ and $\delta_{\beta,j}^2$ are close to 0. Namely, the approximation works well when $\mu_{\alpha,j}$ is concentrated near zero ($\mu_{\beta,j}$ is similar). Since $\mu_{\alpha,j} = \alpha_j/\hat{\sigma}_{\alpha,j}$, this condition is violated in cases such as having large α_j; or having a large sample size so that $\hat{\sigma}_{\alpha,j}$ is small. A practical suggestion given by Huang *et al.* is to check whether the sample variance of $Z_{\alpha,j}$ and $Z_{\beta,j}$ are less than 1.5. Since JT-comp only works well for small $\delta_{\alpha,j}^2$ and $\delta_{\beta,j}^2$, its applicability is limited to the settings with small samples and small α_j's and β_j's.

2.1.4 High dimensional mediation testing (HDMT)

Another method which uses the correct reference distribution is HDMT [26]. Let $\pi_{01}, \pi_{10}, \pi_{00}$ be the proportion of $(\alpha_j = 0, \beta_j \neq 0), (\alpha_j \neq 0, \beta_j = 0)$ and $(\alpha_j = \beta_j = 0)$ among all J tests. The test statistic for the HDMT method is the MaxP statistic. Under $H_{01,j}$ and $H_{10,j}$, $p_{max,j} \sim U(0,1)$ asymptotically. Under $H_{00,j}$, $p_{max,j} \sim Beta(2,1)$. The asymptotic reference distribution for $p_{max,j}$ is:

$$(\hat{\pi}_{01} + \hat{\pi}_{10})U(0,1) + \hat{\pi}_{00}Beta(2,1)$$

where $\hat{\pi}_{01}, \hat{\pi}_{10}$ and $\hat{\pi}_{00}$ are obtained by non-parametric methods for estimating the proportion of nulls[28]. It is worth mentioning that HDMT further proposes improving the power under finite samples. Under $H_{01,j}$, the p-value determined by $U(0,1)$ is accurate asymptotically when the power of rejecting $\beta_j = 0$ goes to 1. Namely, $P(p_{\beta,j} < t|H_{01,j}) \xrightarrow{n\to\infty} 1$ for any $t > 0$. However, this condition is difficult to hold when t is extremely small in a finite sample, resulting in a noticeably larger p-value than the truth. In such cases, HDMT uses the Grenander estimator to estimate $P(p_{\beta,j} < t|H_{01,j})$ and $P(p_{\alpha,j} < t|H_{10,j})$.

Overall, since the mixture null distribution of $p_{max,j}$ statistic is asymptotically correct, HDMT is robust to any choices of $\pi_{01}, \pi_{10}, \pi_{00}$. However, since the rejection rule of HDMT is determined by empirically estimating the significance thresholds and false discovery rates, it is difficult to compare it with other methods in terms of p-values. We make the following modifications to obtain p-values from HDMT using the asymptotic mixture reference distribution:

$$p_{HDMT,j} = (\hat{\pi}_{01} + \hat{\pi}_{10}) p_{max,j} + \hat{\pi}_{00} p_{max,j}^2$$

With finite samples, we estimate $P(p_{\alpha,j} < p_{max,j} | H_{10,j})$ and $P(p_{\beta,j} < p_{max,j} | H_{01,j})$ by the Grenander estimator as described in [26]. The adjusted p-value is:

$$\tilde{p}_{HDMT,j} = \hat{\pi}_{01} p_{max,j} \hat{P}(p_{\beta,j} < p_{max,j} | H_{01,j}) + \hat{\pi}_{10} p_{max,j} \hat{P}(p_{\alpha,j} < p_{max,j} | H_{10,j}) + \hat{\pi}_{00} p_{max,j}^2$$

2.1.5 Divide-Aggregate Composite-null Test (DACT)

The test statistic for DACT is constructed as a composite p-value obtained by averaging the three case-specific p-values weighted by $\pi_{01}, \pi_{10}, \pi_{00}$, respectively [24]. Under $H_{01,j}$, the p-value is $p_{\alpha,j}$ since β_j is known to be non-zero. Similarly, the p-value under $H_{10,j}$ is $p_{\beta,j}$. Under $H_{00,j}$, the p-value is $p_{max,j}^2$ using the MaxP statistic, which follows $Beta(2,1)$. The DACT test statistic is defined as:

$$DACT_j = \hat{\pi}_{01} p_{\alpha,j} + \hat{\pi}_{10} p_{\beta,j} + \hat{\pi}_{00} p_{max,j}^2 \tag{6}$$

where $\hat{\pi}_{01}, \hat{\pi}_{10}$ and $\hat{\pi}_{00}$ are obtained based on the empirical characteristic function and Fourier analysis [29]. If any of $\hat{\pi}_{00}, \hat{\pi}_{10}, \hat{\pi}_{01}$ is close to 1, DACT then follows $U(0,1)$ approximately. Otherwise, the DACT statistic deviates from $U(0,1)$. Under this scenario, the DACT method adapts Efron's empirical null framework [30] to estimate the null distribution of the transformed DACT statistic using inverse standard normal distribution function. The final p-value is calibrated using the empirical null distribution.

9

However, the reference distribution for the DACT test statistic has not been established. When none of $\pi_{00}, \pi_{10}, \pi_{01}$ is close to 1, although Efron's method has been adapted as a remedy to estimate the null distribution of the transformed DACT statistic, it remains unclear how close the estimation is to the truth. In fact, the cumulative distribution function for the DACT statistic is complicated, because the third term $p_{max,j}^2$ in (6) depends on the larger of the first two terms such that the three terms are dependent. Therefore, DACT should be used cautiously when $\pi_{00}, \pi_{01}, \pi_{10}$ are all far from 1, for example, when they are all 1/3 say.

2.1.6   A new variant of HDMT: Sobel-comp

We propose a variant of HDMT using Sobel's test statistic, called Sobel-comp. Under $H_{01,j}$ and $H_{10,j}$, $T_{sobel,j} \sim N(0,1)$ asymptotically. Under $H_{00,j}$, $T_{sobel,j} \sim N(0,1/4)$ asymptotically. The asymptotic reference distribution for $T_{sobel,j}$ is:

$$(\hat{\pi}_{01} + \hat{\pi}_{10}) N(0,1) + \hat{\pi}_{00} N(0,1/4)$$

where $\hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{00}$ are obtained from the HDMT method. When $|Z_{\beta,j}| > |Z_{\alpha,j}|$, the p-value for HDMT under $H_{00,j}$ is identical no matter how large $|Z_{\beta,j}|$ is. Therefore, the HDMT method loses power since a stronger effect of the mediator on the outcome does not increase the power to detect the mediation effect if the exposure has a relatively weak effect on the mediator. In contrast, the p-value for Sobel-comp under $H_{00,j}$ decreases as $|Z_{\beta,j}|$ increases. In particular,

Proposition 1. Suppose $|Z_{\beta,j}| > |Z_{\alpha,j}| \geq 0$. The case-specific p-value under $H_{00,j}$ from Sobel-comp is smaller than that from HDMT if $|Z_{\beta,j}| > max\left(|Z_{\alpha,j}|, \left\{4\left(\Phi^{-1}\left(2\Phi(|Z_{\alpha,j}|)^2\right)\right)^{-2} - Z_{\alpha,j}^{-2}\right\}^{-1/2}\right)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable.

Proposition 1 is also true when we interchange $|Z_{\beta,j}|$ and $|Z_{\alpha,j}|$. The proof of *Proposition 1* is provided in the supplementary materials. However, in addition to the conditions in *Proposition 1*,

Sobel-comp requires π_{00} close to 1 to be more powerful than HDMT. On the other hand, unlike HDMT which can estimate $P(p_{\alpha,j} < p_{max,j}|H_{10,j})$ and $P(p_{\beta,j} < p_{max,j}|H_{01,j})$ to further increase power with finite samples, it is difficult to extend Sobel-comp using similar technique because $Z_{\alpha,j}$ and $Z_{\beta,j}$ in the Sobel's statistic are not separable.

2.2 Simulation setup

We evaluate the performance of Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT in terms of false positive rate (FPR) under the null hypothesis and true positive rate (TPR) under the alternative hypothesis in simulation scenarios by varying 1) the proportion of the null and the alternative components, denoted as $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$; 2) the sample size n; 3) the variation of the non-zero parameters α, β across mediators; and 4) R^2 in the data-generating models. We assess the mediation effect of $J = 100,000$ mediators (denoted as M_j where $j \in \{1, 2, ..., J\}$) from the exposure (X) to the outcome (Y). For the j-th pair of models, we first generate the exposure $X \sim N(0, 1)$ and then generate M_j and Y from:

$$M_j = \alpha_j X + \epsilon_{M_j} \tag{7}$$

$$Y = \beta_j M_j + \beta_X X + \epsilon_Y \tag{8}$$

where $\epsilon_{M_j} \sim N(0, \sigma^2_{M_j})$, $\epsilon_Y \sim N(0, \sigma^2_Y)$ and $\beta_X = 1$. For J pairs of models, with probability π_{00}, $\alpha_j = \beta_j = 0$; with probability π_{01}, $\alpha_j = 0, \beta_j \sim N(0, \tau^2)$; with probability π_{10}, $\alpha_j \sim N(0, 5\tau^2), \beta_j = 0$; and with probability π_{11}, $\alpha_j \sim N(0, 5\tau^2), \beta_j \sim N(0, \tau^2)$. The parameter τ controls the dispersion of the non-zero coefficients.

To evaluate the FPR for the six methods under the composite null hypothesis, π_{11} is set as 0. We construct four classes of scenarios (**Table 1**).

	$\pi_{11}, \pi_{01}, \pi_{10}, \pi_{00}$	Sample size	τ	R^2
Sparse null 1	$0, 0.001, 0.001, 0.998$	$(200, 500, 1000)$	$(0.1, 0.3, 0.7)$	Not controlled
Dense null 1	$0, 0.33, 0.33, 0.34$	$(200, 500, 1000)$	$(0.1, 0.3, 0.7)$	Not controlled
Sparse null 2	$0, 0.001, 0.001, 0.998$	$(200, 500, 1000)$	(0.3)	$(0.1, 0.15, 0.2)$
Dense null 2	$0, 0.33, 0.33, 0.34$	$(200, 500, 1000)$	(0.3)	$(0.1, 0.15, 0.2)$

Table 1: Simulation scenarios for comparing false positive rates . With probability π_{01}, $\alpha_j = 0$ and $\beta_j \sim N(0, \tau^2)$; with probability $\pi_{10}, \alpha_j \sim N(0, 5\tau^2)$ and $\beta_j = 0$; with probability $\pi_{00}, \alpha_j = \beta_j = 0$. The last column refers to the R^2 in the data generating models.

In Sparse&Dense null 1 scenarios, $\delta_{M_j} = \delta_Y = 1$. In contrast to Sparse&Dense null 1 scenarios where R^2 varies across mediators, Sparse&Dense null 2 scenarios control R^2 at the same level. We calculate the FPR at the nominal significance levels of $10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}$, and 5×10^{-7}, where 5×10^{-7} corresponds to controlling the overall FWER at 0.05. Under the null hypothesis, the FPR given a significance level is calculated as the proportion of p-values among 100,000 tests below this level. We repeat this process 2,000 times and average FPRs over 2,000 replicates.

For power comparison, we follow the same data generation process described above except that we also simulate data under the alternative hypothesis. We have four classes of scenarios in **Table 2** .

	$\pi_{11}, \pi_{01}, \pi_{10}, \pi_{00}$	Sample size	τ	R^2
Sparse alternative 1	$0.001, 0.001, 0.001, 0.997$	$(200, 500, 1000)$	$(0.1, 0.3, 0.7)$	Not controlled
Dense alternative 1	$0.2, 0.2, 0.2, 0.4$	$(200, 500, 1000)$	$(0.1, 0.3, 0.7)$	Not controlled
Sparse alternative 2	$0.001, 0.001, 0.001, 0.997$	$(200, 500, 1000)$	(0.3)	$(0.1, 0.15, 0.2)$
Dense alternative 2	$0.2, 0.2, 0.2, 0.4$	$(200, 500, 1000)$	(0.3)	$(0.1, 0.15, 0.2)$

Table 2: Simulation scenarios for comparing true positive rates. With probability $\pi_{11}, \alpha_j \sim N(0, 5\tau^2), \beta_j \sim N(0, \tau^2)$; with probability π_{01}, $\alpha_j = 0$ and $\beta_j \sim N(0, \tau^2)$; with probability $\pi_{10}, \alpha_j \sim N(0, 5\tau^2)$ and $\beta_j = 0$; with probability $\pi_{00}, \alpha_j = \beta_j = 0$. The last column refers to the R^2 in the data generating models.

Under the control of the true FDR at 0.05, we evaluate the TPR for each method by calculating the number of observed rejections under which the alternative hypothesis is true to the total number true non-null signals. Calculating the true FDR is possible in simulation studies since the underlying truth is known. We repeat the process 200 times, and the TPR is averaged over all 200 replicates. We

use existing R software and packages to implement JT-comp (thttp://www.stat.sinica.edu.tw/ythuang/JT-Comp.zip), DACT [24] and HDMT [26].

2.3 Data example using MESA: study design and methods

We apply all six methods (Sobel's test, the MaxP test, JT-comp, HDMT, Sobel-comp and DACT) to study the mediation mechanism of DNA methylation levels at CpG sites in the pathway from adult SES to HbA1c using data from MESA [31]. Our exposure, adult SES, defined by educational attainment, is a risk factor for cardiovascular disease and diabetes [32, 33]. Our outcome, HbA1c, which reflects the three-month average blood sugar level, is a critical measurement in the diagnosis of diabetes [34] and is a known risk factor for cardiovascular disease [35, 36, 37]. We assume that the effect direction is from educational attainment to HbA1c level since the exposure has remained unchanged during the study and was collected before measuring HbA1c. Moreover, previous research has reported potential causality between educational attainment and type 2 diabetes [38]. In addition, educational attainment is associated with DNAm [39], and DNAm is also associated with HbA1c [40]. It is thus of interest to identify DNAm sites that mediate the effect of educational attainment on HbA1c.

Since correlated mediators may lead to inflated Type I error rates and spurious signals, we selected a subset of 228,088 potentially mediating CpGs that were, at most, only weakly correlated with one another. We provide details for processing MESA data in the supplementary materials. For each CpG site, we obtained $z_{\alpha,j}$ and $z_{\beta,j}$ from linear mixed models for testing $\alpha_j = 0$ (effect of the exposure on the j-th mediator) and $\beta_j = 0$ (effect of the j-th mediator on the outcome). In both models, we adjusted for age, sex and race as potential confounders and adjusted for the estimated proportions of residual non-monocytes (neutrophils, B cells, T cells, and natural killer cells) to account for potential contamination by non-monocyte cell types. We included the methylation chip and position as random

effects to account for potential batch effects. In addition, we adjusted for the exposure in the outcome-mediator model. We applied the six mediation methods to the selected 228,088 CpGs, and obtained p-values from each method for testing the mediation effect. CpG sites with a significant mediation effect are determined by the p-value threshold of 2.19×10^{-7}, which corresponds to controlling FWER at 0.05.

3 Results

3.1 Simulation results

3.1.1 False positive rates under the composite null hypothesis

In **Table S1**, we present FPR from six methods under the *Sparse null 1* scenario, where $(\pi_{01}, \pi_{10}, \pi_{00}) = (0.001, 0.001, 0.998)$, the sample size $n \in (200, 500, 1000)$ and $\tau \in (0.1, 0.3, 0.7)$. To better illustrate the distributions of p-values, we provide QQ plots from one replication in **Figure 2**. For all nine cases, Sobel's test is the most conservative test, followed by the MaxP test. P-values from both tests are uniformly larger than the expected p-values due to large π_{00}. R package DACT fails in certain cases, e.g. when $\tau = 0.7$ or when $n = 1000$. When $n = 200$ and $\tau = 0.1$, the FPRs from HDMT and Sobel-comp are close to expected values at the cut-off higher than 10^{-6}, but are inflated at a lower cut-off. In comparison, p-values from JT-comp and DACT are greatly inflated, especially when the cut-off is lower than 10^{-6}. At the cut-off of 5×10^{-7}, the ratio of the FPR to the corresponding cut-off for JT-comp, DACT, Sobel-comp, and HDMT is 15.4, 1.5, 2.1 and 21.9, respectively. When increasing n from 200 to 1000 with $\tau = 0.1$, the FPR for JT-comp dramatically increases. In comparison, Sobel-comp is less inflated and HDMT almost keeps the same level of FPR. Similar trends are observed with an increasing τ.

When the non-zero coefficients are dense in the *Dense null 1* scenario (**Figure 3** and **Table S2**),

14

HDMT is the only method that maintains the FPR at the nominal level in all scenarios, and is robust to the change of n or τ.

In **Tables S3 and S4**, we present the FPR for the *Sparse&Dense null 2* scenarios, where $R^2 \in (0.1, 0.15, 0.2)$ is controlled across J tests. Overall, the impact of R^2 is similar to τ in the *Sparse&Dense null 1* scenario for all methods except DACT. In the *Sparse null 2* scenario, increasing R^2 ameliorates the inflated FPR for DACT. However, there is no clear trend of how the sample size impacts DACT. With a fixed R^2, DACT has the largest FPR when $n = 500$, but has smaller ones when $n = 200$ and $n = 1000$.

Figure 2: QQ plots for p-values from Sobel's test, the MaxP test, JT-comp, HDMT, Sobel-comp and DACT under the *Sparse null 1* scenario. n is the sample size. The total number of mediators is 100,000. For $j = 1, 2, ..., 100,000$, with probability $\pi_{01} = 0.001$, $\alpha_j = 0$ and $\beta_j \sim N(0, \tau^2)$; with probability $\pi_{10} = 0.001$, $\alpha_j \sim N(0, 5\tau^2)$ and $\beta_j = 0$; with probability $\pi_{00} = 0.998$, $\alpha_j = \beta_j = 0$.

Figure 3: QQ plots for p-values from Sobel's test, the MaxP test, JT-comp, HDMT, Sobel-comp and DACT under the *Dense null 1* scenario. n is the sample size. The total number of mediators is 100,000. For $j = 1, 2, ..., 100,000$, with probability $\pi_{01} = 0.33$, $\alpha_j = 0$ and $\beta_j \sim N(0, \tau^2)$; with probability $\pi_{10} = 0.33$, $\alpha_j \sim N(0, 5\tau^2)$ and $\beta_j = 0$; with probability $\pi_{00} = 0.34$, $\alpha_j = \beta_j = 0$.

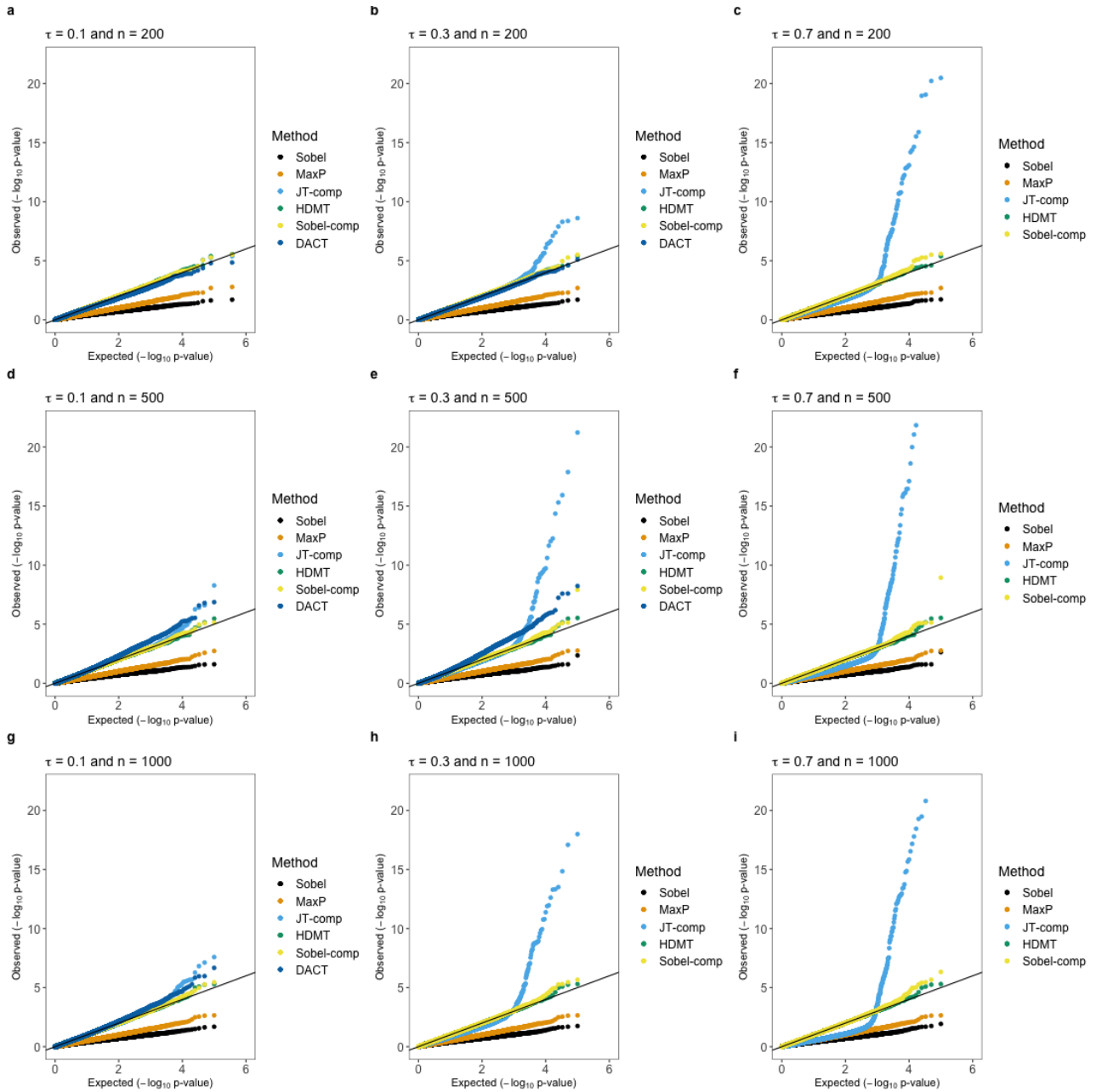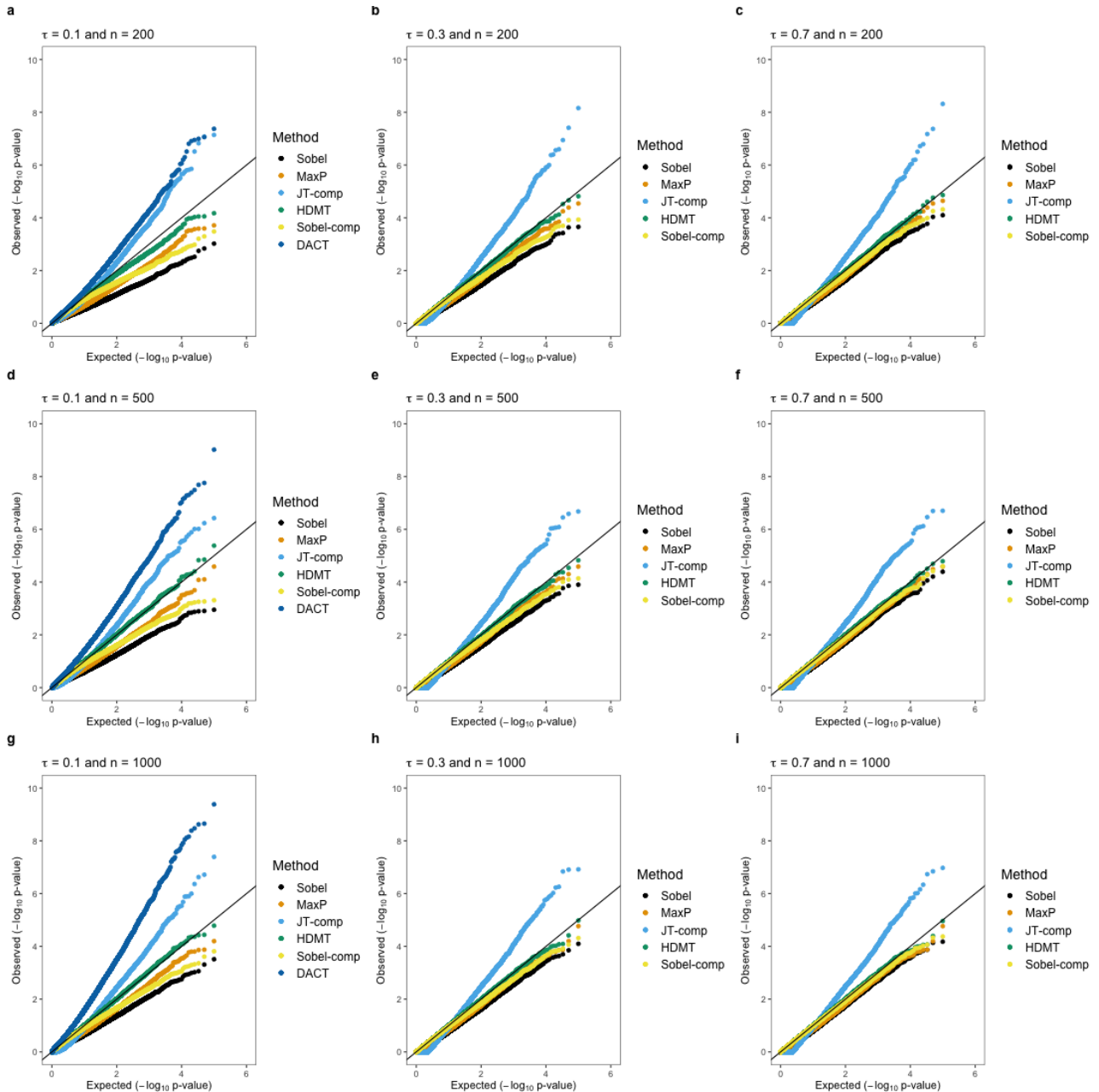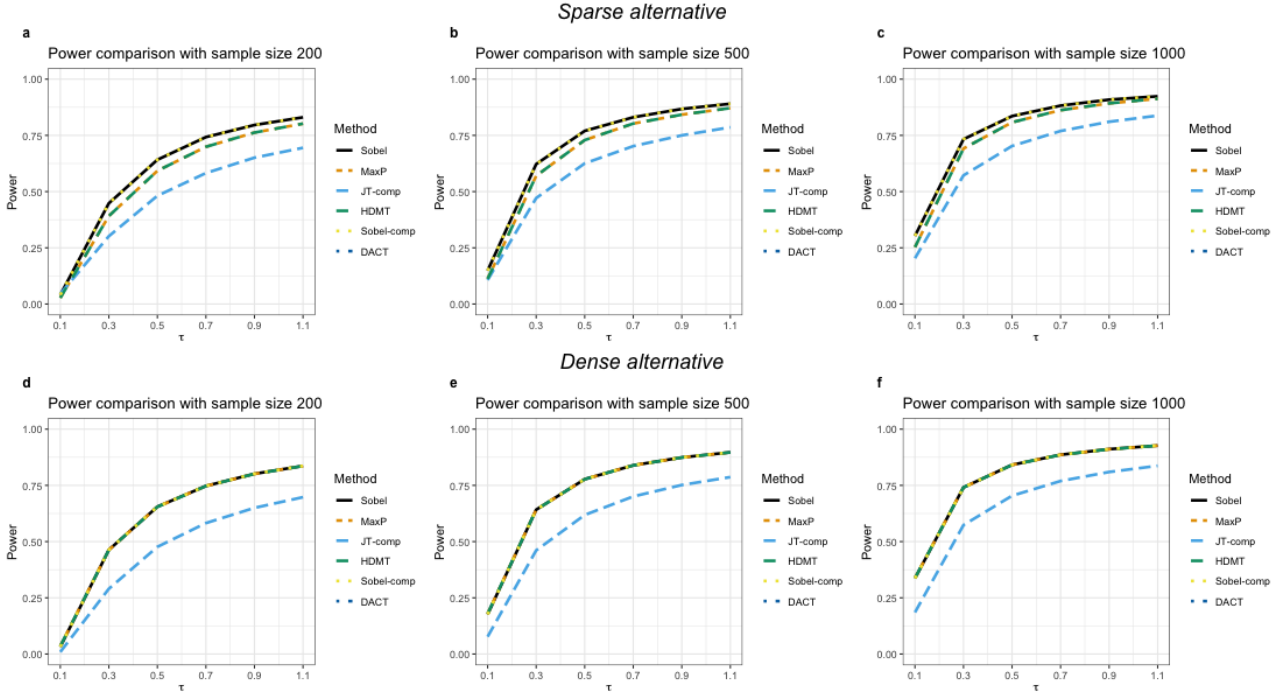### 3.1.2 True positive rates under the alternative hypothesis

Results of the TPRs using *FDR* < 0.05 in *Sparse and Dense alternative 1* scenarios are shown in

Figure 4. DACT fails when $\tau > 0.1$. Under the *Sparse alternative 1* scenario, JT-comp has lower TPR

17

than the four other methods in most simulation scenarios, except when τ is small (e.g. $\tau = 0.1$) and the sample size is small (e.g. $n = 200$). Sobel's test and Sobel-comp have the highest TPRs, closely followed by HDMT and MaxP. The TPR increases for all methods when the sample size increases. Sobel's test and Sobel-comp perform the same because the rank of the weighted composite p-values is unchanged and so are the MaxP test and HDMT. Under the *Dense alternative 1* scenario, the TPR of Sobel's test, MaxP, HDMT and Sobel-comp is the same under the control of FDR. JT-comp has the lowest TPR among all methods.

Results for the average TPR using FDR<0.05 in *Sparse and Dense alternative 2* scenarios are shown in **Figure S1**. Under the *Sparse alternative 2* scenario, JT-comp, Sobel's test and Sobel-comp have the highest TPRs, followed by the other three methods. The TPR for each methods first increases as R^2 increases and stays the same afterward. All methods have increasing TPR as n increases. The TPR for each method is nearly the same with the change of R^2 when $n = 1000$. Under the *Dense alternative 2* case, all methods have similar TPR and the trend with varying n and R^2 is similar to the *Sparse alternative 2* scenario.

Figure 4: The average true positive rate over 200 replicates when controlling the true false discovery rate (FDR) at 0.05 for Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT under the *Spase and Dense alternative 1* scenarios. The total number of mediators is 100,000. n is the sample size. For $j = 1, 2, ..., 100,000$, with probability π_{11}, $\alpha_j \sim N(0, 5\tau^2)$, $\beta_j \sim N(0, \tau^2)$; with probability π_{01}, $\alpha_j = 0$ and $\beta_j \sim N(0, \tau^2)$; with probability $\pi_{10}, \alpha_j \sim N(0, 5\tau^2)$ and $\beta_j = 0$; with probability $\pi_{00}, \alpha_j = \beta_j = 0$. Under the *Sparse alternative 1* scenario, $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$ are set as $0.001, 0.001, 0.001, 0.997$ and under the *Dense alternative 1* scenario, $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$ are set as $0.2, 0.2, 0.2, 0.4$.

3.2 Results from MESA

In **Figure 5**, we present the QQ plot for p-values of all 228,088 CpGs from six methods, including Sobel's test, the MaxP test, JT-comp, HDMT, Sobel-comp and DACT. As expected, p-values from Sobel's test and the MaxP test were deflated, potentially due to a large number of zero α_j and β_j. JT-comp identified two significant CpGs, HDMT identified two significant CpGs, and DACT identified four significant CpGs (**Table S5**). Two CpG sites, cg10508317 and cg01288337, were significant using all three methods (**Table 3**). In contrast, Sobel-comp detected no significant mediation effects probably because $\hat{\pi}_{00}$ is far from 1 ($\hat{\pi}_{00} = 0.872, \hat{\pi}_{01} = 0.032, \hat{\pi}_{10} = 0.052$).

The CpG site cg10508317 in the SOCS3 gene on chromosome 17 encodes a protein that is involved

in the signaling pathways of key hormones such as insulin [41]. It has been found that increased

SOCS3 expression is associated with insulin resistance [41], which is directly related to HbA1c. The

CpG site cg01288337 is in the RIN3 gene on chromosome 14. The RIN3 gene encodes a member

of the RIN family of Ras interaction-interference proteins and is next to the SLC24A4 gene. Recent

studies showed that SLC24A4/RIN3 is significantly associated with brain glucose metabolism in

humans [42] and SLC24A4 knockout mice revealed brain glucose hypometabolism [43].

Figure 5: QQ plot for 6 mediation hypothesis testing methods, including Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT with 963 observations. The outcome is the continuous HbA1c level, the exposure is the binary adult SES, and the mediators are 228,088 CpG sites. In the mediator-exposure model, we adjusted for age, sex, race and residual white blood cell types (neutrophils, B cells, T cells, and natural killer cells), and included the methylation chip and position as random effects to account for potential batch effects. In addition, we adjusted for the exposure in the outcome-mediator model.
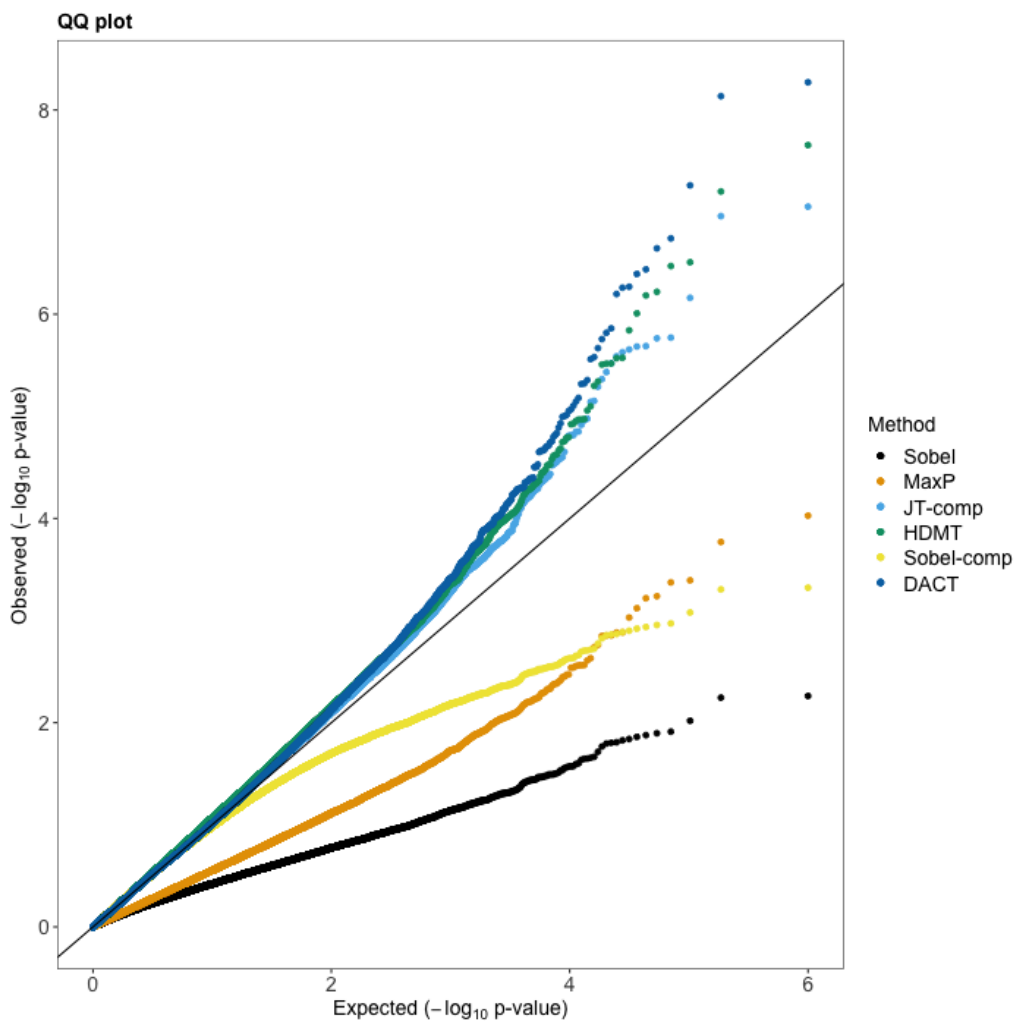
Table 3 : Two mediation pathways identified by JT-comp, HDMT and DACT after controlling the FWER at 0.05. The exposure is adult SES and the outcome is HbA1c. The total number of mediators is 228,088. In the mediator-exposure model, we adjusted for age, sex, race and residual white blood cell types (neutrophils, B cells, T cells, and natural killer cells), and included the methylation chip and position as random effects to account for potential batch effects. In addition, we adjusted for the exposure in the outcome-mediator model. The estimated mediation effect is $\hat{\alpha}\hat{\beta}$ and the proportion of mediation effect is provided in the parenthesis. The 95% confidence interval (CI) is calculated based on 1,000 bootstrap samples.

CpG	Chr	Gene	UCSC RefGene Group	$\hat{\alpha}$	$\hat{\beta}$	Mediation effect (proportion)	95% CI	$p_{JT-comp}$	p_{HDMT}	p_{DACT}
cg10508317	17	SOCS3	Body	-9.92E-02	-1.78E-01	0.018 (0.18)	(0.009,0.029)	8.84E-08	6.28E-08	5.35E-09
cg01288337	14	RIN3	Body	5.55E-02	3.06E-01	0.017 (0.17)	(0.009,0.028)	1.09E-07	2.21E-08	7.32E-09

4 Discussion

We reviewed and compared the testing performance of six mediation methods (Sobel's test, MaxP, JT-comp, HDMT, DACT and Sobel-comp). Our study indicates that the methods which use the mixture reference distribution (HDMT, Sobel-comp) can better control false positive rates and yield larger true positive rates. However, there is no uniform dominance of one method over the others across all simulation scenarios. The performance of the methods differs according to values of $\pi_{00}, \pi_{01}, \pi_{10}, \pi_{11}$, the sample size and the strength of independent variables explaining the variation of the dependent variable in the two models (1) and (2), as captured by the variance of non-zero α, β or R^2 in models (7) and (8).

Under the null hypothesis, the distribution of p-values is strongly affected by the three proportions, $\pi_{00}, \pi_{01}, \pi_{10}$, for all methods except HDMT. Our simulation studies show that HDMT is the only method that controls FPR when non-zero coefficients are dense, i.e. when π_{01} and π_{10} are large. On the other hand, when non-zero coefficients are sparse, Sobel-comp performs similar to HDMT. In comparison, JT-comp maintains the nominal level of FPR only when the sample size is small and the variances of non-zero α and β are small (or R^2 is small). The application of JT-comp is limited to sparse settings with small samples and relatively weak signals.

Under the alternative hypothesis with sparse signals, all methods perform similar with a small sample size n and small τ. As n and τ increase, Sobel-comp is most powerful method with the greatest TPR, followed by HDMT. Under the dense settings, Sobel-comp has the same TPR as HDMT. In practice, we recommend to first estimate $\pi_{01}, \pi_{10}, \pi_{00}$ using R package HDMT [26] and then choose the method based on $\hat{\pi}_{01}, \hat{\pi}_{10}, \hat{\pi}_{00}$. Sobel-comp is preferred when $\hat{\pi}_{01}$ and $\hat{\pi}_{10}$ are close to 0. Otherwise, HDMT is preferred. Although we do not provide strict guidelines, our simulation studies show that when $\pi_{01} = \pi_{10} = \pi_{11} = 0.001$, Sobel-comp is the most powerful method in almost all scenarios. We summarize key features, advantages and limitations for all the six methods in **Table 4** and provide a decision tree for choosing an appropriate method in **Figure 6**.

A common limitation for all six methods is that none of them work when mediators are correlated. Presented with correlated mediators, univariate mediation analysis does not adjust for all the mediator-outcome confounders affected by the exposure, resulting in a violation of assumption 4 mentioned in Section 1. In this case, it is necessary to extend the mediation analysis models to jointly account for multiple correlated mediators [13, 44]. For computational reasons, we only explore a range of parameters. Parameter values beyond this range combined with correlated mediators are of interest for future analysis.

The two significant CpGs we identified in the SOCS3 and RIN3 genes from MESA add to a growing body of literature for the mediating role of DNA methylation between socioeconomic status and disease risk factors associated with HbA1c [13, 45]. However, a limitation of our analysis is that our mediator (methylation) and outcome (HbA1c) were measured concurrently. Therefore, we identify statistical mediation but are unable to formally evaluate causal mediation. More studies are needed to fully understand the underlying biological mechanisms that link socioeconomic disadvantage to HbA1c-associated diseases.

Table 4: A summary of methods with key advantages and limitations. For J tests, $\pi_{11}, \pi_{01}, \pi_{10}, \pi_{00}$ are the true proportion of $(\alpha \neq 0, \beta \neq 0)$, $(\alpha = 0, \beta \neq 0)$, $(\alpha \neq 0, \beta = 0)$, and $(\alpha = \beta = 0)$, respectively. τ^2 is the variance for the non-zero α, β, respectively. p_α and p_β are the two-sided p-value for Z_α and Z_β, respectively.

Method	First Author Year (Reference No.)	R package (package name)	Test statistic	Reference Distribution	Advantages	Limitations
Sobel's test	Sobel, 1982 [21]	✗	$\dfrac{Z_\alpha}{\sqrt{1+(Z_\alpha/Z_\beta)^2}}$	$N(0,1)$	Protect the false positive rate (FPR) at the nominal level under the null hypothesis. Robust for any value of $\pi_{00}, \pi_{01}, \pi_{10}$ and τ.	Conservative since the multivariate delta method fails when α and β are zero.
MaxP	Mackinnon, 2002 [22]	✗	$max(p_\alpha, p_\beta)$	$U(0,1)$	Protect the FPR at the nominal level under the null hypothesis. Robust for any value of $\pi_{00}, \pi_{01}, \pi_{10}$ and τ. Uniformly more powerful than Sobel's test.	Conservative because of using the incorrect reference distribution when α and β are zero.
JT-comp	Huang, 2019 [25]	✓	$Z_\alpha Z_\beta$	H_{00}: Standard product normal distribution. H_{01}, H_{10}: Normal product distribution with non-zero mean.	Correct mixture reference distribution for $Z_\alpha Z_\beta$. No need to estimate $\pi_{00}, \pi_{01}, \pi_{10}$. Keeping the FPR close to the nominal level when the sample size and τ are small.	Inflated FPR at a small significance threshold. Inflated FPR when the sample size or τ increases. Only works with small samples and relatively weak signals.
HDMT	Dai, 2020 [26]	✓(HDMT)	$max(p_\alpha, p_\beta)$	H_{00} : $Beta(2,1)$. $H_{01}, H_{10} : U(0,1)$.	Correct mixture reference distribution for the MaxP test statistic. Maintaining the FPR close to the nominal level for any value of $\pi_{00}, \pi_{01}, \pi_{10}$ and τ. More powerful than Sobel's test and the MaxP test. Provides finite-sample size adjustment for p-values to increase power.	Power lose when p_β is much smaller than p_α and vice versa.
Sobel-comp	-	✗	$\dfrac{Z_\alpha}{\sqrt{1+(Z_\alpha/Z_\beta)^2}}$	H_{00} : $N(0,1/4)$. $H_{01}, H_{10} : N(0,1)$	Correct mixture reference distribution for Sobel's test statistic. Maintaining the FPR close to the nominal level and more powerful than HDMT when π_{01} and π_{10} are close to 0.	Conservative if π_{01} or π_{10} is far from 0 due to the use of the asymptotic reference distribution under H_{01} and H_{10}.
DACT	Liu, 2020 [24]	✓(DACT)	$\hat{\pi}_{01} p_\alpha + \hat{\pi}_{10} p_\beta + \hat{\pi}_{00} p_{max}^2$	$U(0,1)$ approximately.	Weights the case-specific p-values to construct a composite test statistic to accommodate the composite nature of the null hypothesis.	Exact reference distribution of $DACT$ statistic is not established. Approximation of the reference distribution is often inaccurate, causing the FPR deviating from the nominal level.

Figure 6: Decision tree for choosing the optimal mediation hypothesis testing method based on the simulation studies for the normally-distributed outcome. $\pi_{11}, \pi_{01}, \pi_{10}, \pi_{00}$ are the proportion of $(\alpha \neq 0, \beta \neq 0), (\alpha = 0, \beta \neq 0), (\alpha \neq 0, \beta = 0)$, and $(\alpha = \beta = 0)$, respectively.

(word count: 3984)

Acknowledgments

Conflict of Interest

References

1. VanderWeele, T. *Explanation in causal inference: methods for mediation and interaction* (Oxford University Press, 2015).

2. Pierce, B. L. *et al.* Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS genetics* **10,** e1004818 (2014).

3. Huang, Y.-T., Liang, L., Moffatt, M. F., Cookson, W. O. & Lin, X. iGWAS: Integrative genome-wide association studies of genetic and genomic data for disease susceptibility using mediation analysis. *Genetic epidemiology* **39,** 347–356 (2015).

4. Yang, F. *et al.* Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome research* **27,** 1859–1871 (2017).

5. Chen, Z. *et al.* From tobacco smoking to cancer mutational signature: a mediation analysis strategy to explore the role of epigenetic changes. *BMC cancer* **20,** 1–11 (2020).

6. Zhang, H. *et al.* Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32,** 3150–3154 (2016).

7. Zeng, P., Shao, Z. & Zhou, X. Statistical methods for mediation analysis in the era of high-throughput genomics: Current successes and future challenges. *Computational and Structural Biotechnology Journal* (2021).

8. Kulis, M. & Esteller, M. DNA methylation and cancer. *Advances in genetics* **70,** 27–56 (2010).

9. VanderWeele, T. J. *et al.* Genetic variants on 15q25. 1, smoking, and lung cancer: an assessment of mediation and interaction. *American journal of epidemiology* **175,** 1013–1020 (2012).

10. Wu, D. *et al.* Mediation analysis of alcohol consumption, DNA methylation, and epithelial ovarian cancer. *Journal of human genetics* **63,** 339–348 (2018).

11. Richardson, T. G. *et al.* Mendelian randomization analysis identifies CpG sites as putative mediators for genetic influences on cardiovascular disease risk. *The American Journal of Human Genetics* **101,** 590–602 (2017).

12. Grant, C. D. *et al.* A longitudinal study of DNA methylation as a potential mediator of age-related diabetes risk. *Geroscience* **39,** 475–489 (2017).

13. Song, Y. *et al.* Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics* **76,** 700–710 (2020).

14. Chén, O. Y. *et al.* High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19,** 121–136 (2018).

15. Huang, Y.-T. Variance component tests of multivariate mediation effects under composite null hypotheses. *Biometrics* **75,** 1191–1204 (2019).

16. Baron, R. M. & Kenny, D. A. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology* **51,** 1173 (1986).

17. Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics,* 34–58 (1978).

18. Valeri, L. & VanderWeele, T. J. Mediation analysis allowing for exposure–mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. *Psychological methods* **18,** 137 (2013).

19. Richiardi, L., Bellocco, R. & Zugna, D. Mediation analysis in epidemiology: methods, interpretation and bias. *International journal of epidemiology* **42,** 1511–1519 (2013).

20. Valente, M. J., Pelham III, W. E., Smyth, H. & MacKinnon, D. P. Confounding in statistical mediation analysis: What it is and how to address it. *Journal of counseling psychology* **64,** 659 (2017).

21. Sobel, M. E. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology* **13,** 290–312 (1982).

22. MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G. & Sheets, V. A comparison of methods to test mediation and other intervening variable effects. *Psychological methods* **7,** 83 (2002).

23. Barfield, R. *et al.* Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic epidemiology* **41,** 824–833 (2017).

24. Liu, Z. *et al.* Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies. *medRxiv* (2020).

25. Huang, Y.-T. *et al.* Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics* **13,** 60–84 (2019).

26. Dai, J. Y., Stanford, J. L. & LeBlanc, M. A multiple-testing procedure for high-dimensional mediation hypotheses. *Journal of the American Statistical Association,* 1–16 (2020).

27. MacKinnon, D. P., Warsi, G. & Dwyer, J. H. A simulation study of mediated effect measures. *Multivariate behavioral research* **30,** 41–62 (1995).

28. Storey, J. D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **64,** 479–498 (2002).

29. Jin, J. & Cai, T. T. Estimating the null and the proportion of nonnull effects in large-scale multiple comparisons. *Journal of the American Statistical Association* **102,** 495–506 (2007).

30. Efron, B., Tibshirani, R., Storey, J. D. & Tusher, V. Empirical Bayes analysis of a microarray experiment. *Journal of the American statistical association* **96,** 1151–1160 (2001).

31. Bild, D. E. *et al.* Multi-ethnic study of atherosclerosis: objectives and design. *American journal of epidemiology* **156,** 871–881 (2002).

32. Whitaker, S. M. *et al.* The association between educational attainment and diabetes among men in the United States. *American journal of men's health* **8,** 349–356 (2014).

33. Telfair, J. & Shelton, T. L. Educational attainment as a social determinant of health. *North Carolina medical journal* **73,** 358–365 (2012).

34. Organization, W. H. *et al. Use of glycated haemoglobin (HbA1c) in diagnosis of diabetes mellitus: abbreviated report of a WHO consultation* tech. rep. (World Health Organization, 2011).

35. Singer, D. E., Nathan, D. M., Anderson, K. M., Wilson, P. W. & Evans, J. C. Association of HbA1c with prevalent cardiovascular disease in the original cohort of the Framingham Heart Study. *Diabetes* **41,** 202–208 (1992).

36. Sakurai, M. *et al.* HbA1c and the risks for all-cause and cardiovascular mortality in the general Japanese population: NIPPON DATA90. *Diabetes care* **36,** 3759–3765 (2013).

37. Yeung, S. L. A., Luo, S. & Schooling, C. M. The impact of glycated hemoglobin (HbA1c) on cardiovascular disease risk: a Mendelian randomization study using UK Biobank. *Diabetes Care* **41,** 1991–1997 (2018).

38. Liang, J. *et al.* Educational attainment protects against type 2 diabetes independently of cognitive performance: A Mendelian randomization study. *Acta Diabetologica* **58,** 567–574 (2021).

39. Van Dongen, J. *et al.* DNA methylation signatures of educational attainment. *npj Science of Learning* **3,** 1–14 (2018).

40. Chen, Z. *et al.* DNA methylation mediates development of HbA1c-associated complications in type 1 diabetes. *Nature metabolism* **2,** 744–762 (2020).

41. Pedroso, J. A., Ramos-Lobo, A. M. & Donato, J. SOCS3 as a future target to treat metabolic disorders. *Hormones* **18,** 127–136 (2019).

42. Stage, E. *et al.* The effect of the top 20 Alzheimer disease risk genes on gray-matter density and FDG PET brain metabolism. *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring* **5,** 53–66 (2016).

43. Li, X.-F. & Lytton, J. An essential role for the K+-dependent Na+/Ca2+-exchanger, NCKX4, in melanocortin-4-receptor-dependent satiety. *Journal of Biological Chemistry* **289,** 25445–25459 (2014).

44. Song, Y. *et al.* Bayesian Hierarchical Models for High-dimensional Mediation Analysis with Coordinated Selection of Correlated Mediators. *arXiv preprint arXiv:2009.11409* (2020).

45. Giurgescu, C. *et al.* Neighborhood environment and DNA methylation: implications for cardiovascular disease risk. *Journal of Urban Health* **96,** 23–34 (2019).

46. Liu, Y. *et al.* Methylomics of gene expression in human monocytes. *Human molecular genetics* **22,** 5065–5074 (2013).

Supplementary Materials

Proof of proposition 1

Proof. In this proof, we drop the subscript j since the statement is true for all $j \in \{1, 2, ..., J\}$. When $|Z_\beta| > |Z_\alpha|$, under H_{00},

$$T_{sobel} = \frac{|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}} \sim N(0, 1/4)$$

$$p_{max} = 2\Phi(|Z_\alpha|) \sim Beta(2, 1)$$

The p-value for Sobel-comp under H_{00} is:

$$p_{sobel}^{00} = 2 \int_{\frac{|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}}}^{\infty} (2\pi \frac{1}{4})^{-1/2} exp(-2u^2) du$$

$$= 2 \int_{\frac{2|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}}}^{\infty} (2\pi)^{-1/2} exp(-1/2v^2) dv$$

$$= 2\Phi\left(\frac{2|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}}\right)$$

The p-value for HDMT under H_{00} is:

$$p_{max}^2 = 4\Phi(|Z_\alpha|)^2$$

Notice that p_{sobel}^{00} is a decreasing function of $|Z_\beta|$, and

$$p_{sobel}^{00} < p_{max}^2$$

$$\iff 2\Phi\left(\frac{2|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}}\right) < 4\Phi(|Z_\alpha|)^2$$

$$\iff \frac{2|Z_\alpha|}{\sqrt{1 + Z_\alpha^2 / Z_\beta^2}} > \Phi^{-1}\left(2\Phi(|Z_\beta|)^2\right)$$

$$\iff |Z_\beta| > \left\{4\left(\Phi^{-1}\left(2\Phi(|Z_\alpha|)^2\right)\right)^{-2} - Z_\alpha^{-2}\right\}^{-1/2}$$

Note that $\left\{4\left(\Phi^{-1}\left(2\Phi(|Z_\alpha|)^2\right)\right)^{-2} - Z_\alpha^{-2}\right\} > 0$. Therefore, a sufficient condition for $p_{sobel}^{00} < p_{max}^2$ is $|Z_\beta| > max\left(|Z_\alpha|, \left\{4\left(\Phi^{-1}\left(2\Phi(|Z_\alpha|)^2\right)\right)^{-2} - Z_\alpha^{-2}\right\}^{-1/2}\right)$.

$\square$

Detailed description of MESA data

MESA is a population-based longitudinal study designed to investigate the predictors and progression of subclinical cardiovascular disease in a cohort of 6,814 participants [31]. Clinical, socio-demographic, lifestyle and behavior, laboratory, nutrition and medication data have been collected at multiple examinations beginning in 2000-2002. We used participants' educational attainment based on their highest degree at MESA Exam 1 as a measure of adult SES (less than a 4-year college degree as 1 vs. with a 4-year college degree or higher as 0). DNAm levels were measured using the Illumina Infinium HumanMethylation450 Beadchip on purified monocytes from a random subsample of 1,264 non-Hispanic white, African-American, and Hispanic MESA participants between April 2010 and February 2012 (corresponding to MESA Exam 5). A total of 402,339 CpGs remained after quality control and filtering, including: "detected" methylation levels in <90% of MESA samples using a detection p-value cut-off of 0.05, overlap with a repetitive element or region, presence of SNPs within 10 base pairs according to Illumina annotation, non-reliable probes recommended by DMRcate (having SNPs with minor allele frequency > 0.05 within 2 base pairs or cross reactive probes), probes on sex chromosomes, SNPs, and other non-CpG targeting probes. Additional details about the data collection and processing procedures can be found in [46]. We used HbA1c measured at Exam 5 as the outcome. Our analysis focused on the participants taking no insulin or oral hypoglycemic medication. After removing missing values, a total of 963 individuals remained for analysis.

For the j-th CpG site, where $j = 1, 2, ..., 402339$, we obtained $z_{\alpha,j}$ and $z_{\beta,j}$ from linear mixed models for testing $\alpha_j = 0$ (effect of the exposure on the j-th mediator) and $\beta_j = 0$ (effect of the j-th mediator on the outcome). In both models, we adjusted for age, sex and race as potential confounders and adjusted for the estimated proportions of residual non-monocytes (neutrophils, B cells, T cells, and natural killer cells) to account for potential contamination by non-monocyte cell types. We included the methylation chip and position as random effects to account for potential batch effects. In addition,

we adjusted for the exposure in the outcome-mediator model.

Before performing mediation analysis, since correlated mediators may lead to inflated Type I error rates and spurious signals, we selected a subset of 228,088 potentially mediating CpGs that were, at most, only weakly correlated with one another (correlation coefficient ≤ 0.3). More specifically, we first calculated the correlation matrix for all CpGs on each chromosome. Then we found the mediator with the smallest MaxP p-value. Next, we identified and removed the group of mediators which were correlated with this mediator with correlation coefficient larger than 0.3. We repeated the previous steps until all elements in the correlation matrix were less than or equal to 0.3.

Figure S1: The average true positive rate over 200 replicates when controlling the true false discovery rate (FDR) at 0.05 for Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT under the *Spase and Dense alternative 2* scenarios. The total number of mediators is 100,000. n is the sample size. For $j = 1, 2, ..., 100,000$, with probability π_{11}, $\alpha_j \sim N(0, 5 \times 0.3^2)$, $\beta_j \sim N(0, 0.3^2)$; with probability π_{01}, $\alpha_j = 0$ and $\beta_j \sim N(0, 0.3^2)$; with probability π_{10}, $\alpha_j \sim N(0, 5 \times 0.3^2)$ and $\beta_j = 0$; with probability π_{00}, $\alpha_j = \beta_j = 0$. R^2 is controlled in the j-th mediator-exposure and outcome-mediator models. Under the *Sparse alternative 2* scenario, $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$ are set as $0.001, 0.001, 0.001, 0.997$ and under the *Dense alternative 2* scenario, $\pi_{11}, \pi_{10}, \pi_{01}, \pi_{00}$ are set as $0.2, 0.2, 0.2, 0.4$.

Table S1: The mean and standard deviation of the ratio of the false positive rates to the nominal significance level based on 2,000 replications using Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT under the *Sparse null 1* case. n is the sample size. The total number of mediators is 100,000. With probability $\pi_{01} = 0.001$, $\alpha = 0$ and $\beta \sim N(0, \tau^2)$; with probability $\pi_{10} = 0.001$, $\alpha \sim N(0, 5\tau^2)$ and $\beta = 0$; with probability $\pi_{00} = 0.998$, $\alpha = \beta = 0$.

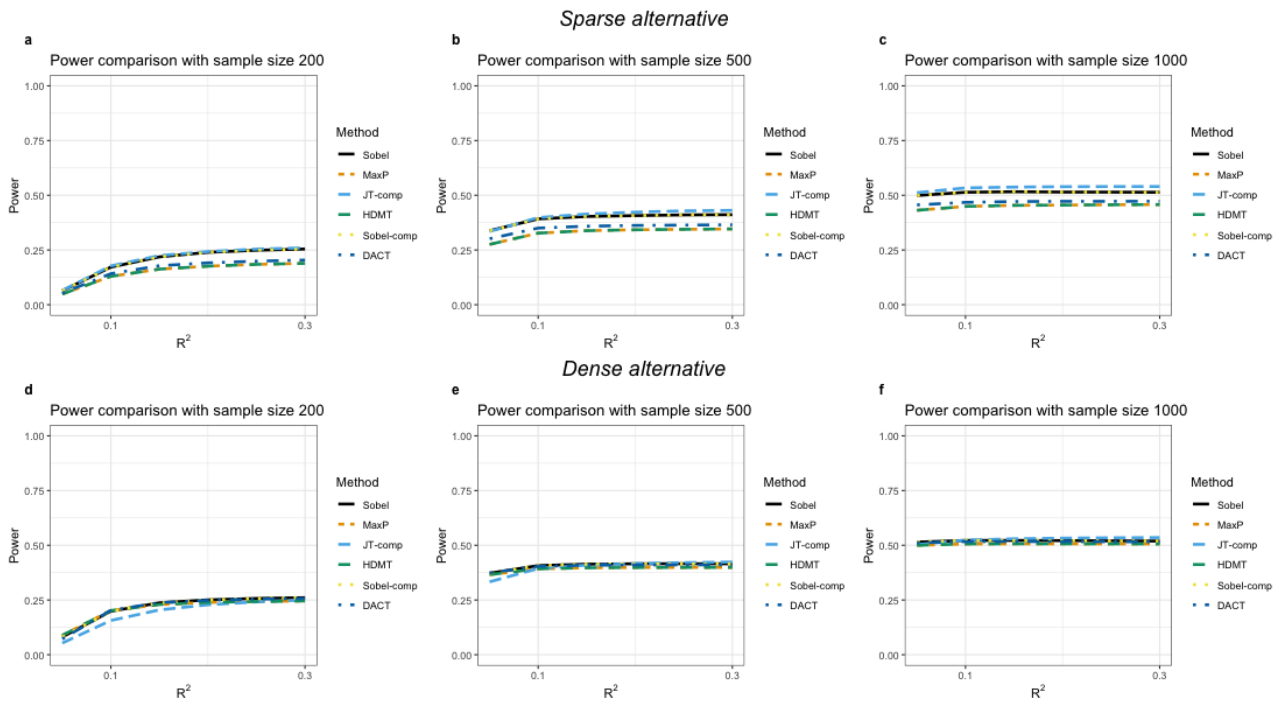| Cut-off	Sobel	MaxP	JT-comp	HDMT	Sobel-comp	DACT
			$n = 200, \tau = 0.1$			
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.11 (0.10)	1.06 (0.11)	0.90 (0.34)	1.46 (1.78)
10^{-4}	0.00 (0.00)	0.00 (0.01)	1.49 (0.38)	1.14 (0.36)	0.83 (0.67)	2.85 (4.10)
10^{-5}	0.00 (0.00)	0.00 (0.00)	3.06 (1.67)	1.26 (1.12)	1.03 (1.30)	6.36 (10.29)
10^{-6}	0.00 (0.00)	0.00 (0.00)	10.02 (10.09)	1.47 (3.94)	1.75 (4.53)	16.09 (30.78)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	15.40 (17.71)	1.53 (5.68)	2.18 (6.85)	21.94 (45.68)
			$n = 200, \tau = 0.3$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.26 (0.11)	1.07 (0.12)	0.91 (0.36)	1.60 (1.82)
10^{-4}	0.00 (0.00)	0.00 (0.01)	3.40 (0.57)	1.18 (0.38)	0.95 (0.80)	3.21 (4.29)
10^{-5}	0.00 (0.00)	0.00 (0.02)	17.27 (4.13)	1.39 (1.21)	1.67 (1.96)	7.29 (11.02)
10^{-6}	0.00 (0.00)	0.00 (0.00)	108.76 (33.03)	2.05 (4.62)	4.87 (8.36)	18.62 (31.79)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	191.40 (62.55)	2.35 (7.04)	7.34 (14.11)	25.11 (45.02)
			$n = 200, \tau = 0.7$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.10 (0.09)	1.02 (0.16)	0.93 (0.37)	NA (NA)
10^{-4}	0.00 (0.01)	0.00 (0.02)	6.79 (0.73)	1.03 (0.47)	1.10 (0.90)	NA (NA)
10^{-5}	0.00 (0.02)	0.00 (0.05)	52.94 (6.71)	1.15 (1.26)	2.44 (2.76)	NA (NA)
10^{-6}	0.00 (0.00)	0.01 (0.32)	426.53 (60.62)	1.76 (4.53)	9.17 (13.32)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	801.93 (115.88)	2.20 (7.07)	14.30 (22.35)	NA (NA)
			$n = 500, \tau = 0.1$			
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.14 (0.10)	1.05 (0.11)	0.97 (0.29)	1.13 (1.62)
10^{-4}	0.00 (0.00)	0.00 (0.01)	2.08 (0.45)	1.11 (0.34)	1.04 (0.62)	2.27 (3.75)
10^{-5}	0.00 (0.00)	0.00 (0.03)	7.50 (2.70)	1.26 (1.12)	1.63 (1.58)	5.74 (10.89)
10^{-6}	0.00 (0.00)	0.00 (0.00)	38.94 (19.27)	1.65 (4.09)	4.13 (6.85)	20.82 (61.49)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	65.99 (35.02)	1.96 (6.15)	5.92 (11.29)	33.65 (114.42)
			$n = 500, \tau = 0.3$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.18 (0.10)	1.03 (0.14)	1.01 (0.30)	NA (NA)
10^{-4}	0.00 (0.01)	0.00 (0.01)	5.43 (0.68)	1.08 (0.41)	1.29 (0.76)	NA (NA)
10^{-5}	0.00 (0.00)	0.00 (0.04)	38.43 (5.85)	1.28 (1.23)	3.01 (2.48)	NA (NA)
10^{-6}	0.00 (0.00)	0.00 (0.00)	292.64 (50.35)	2.05 (4.64)	11.13 (12.48)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	542.37 (97.17)	2.48 (7.01)	17.43 (21.55)	NA (NA)
			$n = 500, \tau = 0.7$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.03 (0.09)	1.03 (0.15)	1.03 (0.31)	NA (NA)
10^{-4}	0.00 (0.01)	0.00 (0.01)	8.10 (0.82)	1.07 (0.43)	1.45 (0.83)	NA (NA)
10^{-5}	0.00 (0.03)	0.00 (0.04)	67.27 (7.26)	1.30 (1.26)	3.88 (3.04)	NA (NA)
10^{-6}	0.00 (0.00)	0.00 (0.00)	564.57 (66.14)	2.16 (4.78)	16.38 (16.36)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	1072.33 (128.19)	2.71 (7.44)	25.90 (27.85)	NA (NA)
			$n = 1000, \tau = 0.1$			
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.19 (0.11)	1.04 (0.11)	1.00 (0.26)	NA (NA)
10^{-4}	0.00 (0.00)	0.00 (0.01)	2.94 (0.54)	1.11 (0.33)	1.15 (0.61)	NA (NA)
10^{-5}	0.00 (0.00)	0.00 (0.03)	14.52 (3.82)	1.35 (1.17)	2.12 (1.79)	NA (NA)
10^{-6}	0.00 (0.00)	0.00 (0.00)	90.51 (29.89)	1.91 (4.45)	6.72 (9.08)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	160.19 (56.83)	2.32 (6.92)	9.97 (15.33)	NA (NA)
			$n = 1000, \tau = 0.3$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.09 (0.09)	1.03 (0.14)	1.05 (0.28)	NA (NA)
10^{-4}	0.00 (0.01)	0.00 (0.01)	6.66 (0.73)	1.07 (0.40)	1.43 (0.73)	NA (NA)
10^{-5}	0.00 (0.02)	0.00 (0.04)	51.77 (6.56)	1.35 (1.26)	3.63 (2.66)	NA (NA)
10^{-6}	0.00 (0.00)	0.00 (0.22)	417.38 (58.22)	2.11 (4.70)	14.93 (14.85)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	783.81 (113.24)	2.65 (7.48)	23.61 (25.70)	NA (NA)
			$n = 1000, \tau = 0.7$			
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.06 (0.09)	1.03 (0.14)	1.07 (0.28)	NA (NA)
10^{-4}	0.00 (0.01)	0.00 (0.01)	8.84 (0.84)	1.08 (0.41)	1.56 (0.79)	NA (NA)
10^{-5}	0.00 (0.04)	0.00 (0.04)	74.83 (7.54)	1.39 (1.28)	4.39 (3.07)	NA (NA)
10^{-6}	0.00 (0.00)	0.00 (0.22)	638.24 (68.79)	2.28 (4.89)	19.36 (17.50)	NA (NA)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	1218.12 (133.97)	2.90 (7.77)	31.27 (30.74)	NA (NA)

Table S2: The mean and standard deviation of the ratio of the false positive rates to the nominal significance level based on 2,000 replications using Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT under the *Dense null 1* case. n is the sample size. The total number of mediators is 100,000. With probability $\pi_{01} = 0.33$, $\alpha = 0$ and $\beta \sim N(0, \tau^2)$; with probability $\pi_{10} = 0.33$, $\alpha \sim N(0, 5\tau^2)$ and $\beta = 0$; with probability $\pi_{00} = 0.34$, $\alpha = \beta = 0$.

Cut-off	Sobel	MaxP	JT-comp	HDMT	Sobel-comp	DACT
			$n = 200, \tau = 0.1$			
10^{-3}	0.01 (0.01)	0.14 (0.04)	3.48 (0.18)	1.08 (0.10)	0.04 (0.02)	7.16 (2.17)
10^{-4}	0.00 (0.01)	0.11 (0.11)	7.92 (0.89)	1.16 (0.34)	0.01 (0.03)	18.32 (7.32)
10^{-5}	0.00 (0.02)	0.12 (0.34)	18.47 (4.25)	1.29 (1.16)	0.00 (0.06)	49.82 (25.09)
10^{-6}	0.00 (0.00)	0.08 (0.86)	42.73 (20.33)	1.66 (4.07)	0.00 (0.00)	142.94 (88.76)
5×10^{-7}	0.00 (0.00)	0.09 (1.34)	54.96 (32.89)	1.86 (6.01)	0.00 (0.00)	196.51 (130.30)
			$n = 200, \tau = 0.3$			
10^{-3}	0.13 (0.04)	0.46 (0.07)	4.60 (0.21)	0.86 (0.10)	0.27 (0.05)	NA (NA)
10^{-4}	0.09 (0.09)	0.47 (0.22)	11.52 (1.08)	0.86 (0.30)	0.18 (0.13)	NA (NA)
10^{-5}	0.07 (0.25)	0.55 (0.74)	29.43 (5.53)	0.94 (0.97)	0.14 (0.38)	NA (NA)
10^{-6}	0.04 (0.63)	0.53 (2.35)	76.89 (27.70)	1.02 (3.23)	0.09 (0.94)	NA (NA)
5×10^{-7}	0.05 (1.00)	0.52 (3.31)	101.87 (44.87)	1.01 (4.60)	0.08 (1.26)	NA (NA)
			$n = 200, \tau = 0.7$			
10^{-3}	0.35 (0.06)	0.63 (0.08)	4.75 (0.21)	1.03 (0.10)	0.59 (0.08)	NA (NA)
10^{-4}	0.30 (0.17)	0.69 (0.26)	12.03 (1.10)	1.12 (0.33)	0.51 (0.22)	NA (NA)
10^{-5}	0.30 (0.55)	0.87 (0.92)	31.19 (5.67)	1.33 (1.15)	0.50 (0.70)	NA (NA)
10^{-6}	0.26 (1.58)	0.91 (3.11)	82.34 (28.33)	1.53 (4.02)	0.41 (1.97)	NA (NA)
5×10^{-7}	0.28 (2.35)	0.88 (4.29)	109.64 (46.28)	1.54 (5.77)	0.40 (2.80)	NA (NA)
			$n = 500, \tau = 0.1$			
10^{-3}	0.03 (0.02)	0.24 (0.05)	4.16 (0.20)	1.04 (0.10)	0.09 (0.03)	13.78 (2.37)
10^{-4}	0.01 (0.03)	0.21 (0.14)	9.99 (0.98)	1.08 (0.34)	0.04 (0.06)	41.14 (9.28)
10^{-5}	0.00 (0.06)	0.17 (0.41)	24.32 (4.87)	1.12 (1.08)	0.01 (0.12)	127.45 (36.18)
10^{-6}	0.00 (0.00)	0.14 (1.18)	60.01 (24.74)	1.08 (3.25)	0.00 (0.00)	406.34 (144.92)
5×10^{-7}	0.00 (0.00)	0.10 (1.41)	79.58 (40.30)	1.18 (4.88)	0.00 (0.00)	578.58 (222.45)
			$n = 500, \tau = 0.3$			
10^{-3}	0.22 (0.05)	0.51 (0.07)	4.69 (0.21)	0.88 (0.09)	0.40 (0.06)	NA (NA)
10^{-4}	0.15 (0.13)	0.51 (0.22)	11.78 (1.06)	0.87 (0.30)	0.28 (0.17)	NA (NA)
10^{-5}	0.11 (0.33)	0.48 (0.70)	30.04 (5.44)	0.84 (0.91)	0.19 (0.44)	NA (NA)
10^{-6}	0.06 (0.80)	0.56 (2.40)	77.33 (27.88)	0.88 (3.03)	0.12 (1.11)	NA (NA)
5×10^{-7}	0.09 (1.34)	0.54 (3.30)	103.02 (45.89)	0.91 (4.36)	0.11 (1.48)	NA (NA)
			$n = 500, \tau = 0.7$			
10^{-3}	0.42 (0.06)	0.62 (0.08)	4.75 (0.21)	0.98 (0.10)	0.68 (0.08)	NA (NA)
10^{-4}	0.36 (0.19)	0.64 (0.25)	11.98 (1.06)	1.01 (0.32)	0.60 (0.25)	NA (NA)
10^{-5}	0.31 (0.56)	0.63 (0.80)	30.81 (5.53)	1.03 (1.01)	0.51 (0.71)	NA (NA)
10^{-6}	0.30 (1.74)	0.70 (2.64)	79.39 (28.47)	1.04 (3.29)	0.47 (2.19)	NA (NA)
5×10^{-7}	0.21 (2.04)	0.71 (3.76)	106.34 (46.60)	1.11 (4.71)	0.42 (2.87)	NA (NA)
			$n = 1000, \tau = 0.1$			
10^{-3}	0.07 (0.03)	0.33 (0.06)	4.44 (0.20)	1.03 (0.10)	0.17 (0.04)	NA (NA)
10^{-4}	0.03 (0.06)	0.29 (0.17)	10.91 (1.03)	1.05 (0.32)	0.08 (0.09)	NA (NA)
10^{-5}	0.02 (0.13)	0.26 (0.50)	27.16 (5.29)	1.09 (1.03)	0.04 (0.21)	NA (NA)
10^{-6}	0.02 (0.39)	0.24 (1.55)	68.32 (26.56)	1.12 (3.39)	0.03 (0.50)	NA (NA)
5×10^{-7}	0.02 (0.63)	0.22 (2.09)	90.70 (43.41)	1.17 (4.90)	0.03 (0.77)	NA (NA)
			$n = 1000, \tau = 0.3$			
10^{-3}	0.29 (0.05)	0.55 (0.07)	4.73 (0.21)	0.91 (0.09)	0.51 (0.07)	NA (NA)
10^{-4}	0.23 (0.15)	0.55 (0.23)	11.89 (1.09)	0.91 (0.31)	0.40 (0.20)	NA (NA)
10^{-5}	0.18 (0.41)	0.55 (0.74)	30.29 (5.61)	0.91 (0.95)	0.30 (0.55)	NA (NA)
10^{-6}	0.11 (1.04)	0.53 (2.29)	77.89 (28.14)	0.95 (3.06)	0.19 (1.37)	NA (NA)
5×10^{-7}	0.14 (1.67)	0.52 (3.25)	103.73 (46.64)	0.84 (4.06)	0.18 (1.89)	NA (NA)
			$n = 1000, \tau = 0.7$			
10^{-3}	0.47 (0.07)	0.63 (0.08)	4.76 (0.21)	0.98 (0.10)	0.76 (0.09)	NA (NA)
10^{-4}	0.43 (0.20)	0.64 (0.25)	12.00 (1.10)	1.00 (0.32)	0.69 (0.26)	NA (NA)
10^{-5}	0.39 (0.61)	0.65 (0.81)	30.60 (5.65)	1.03 (1.02)	0.63 (0.79)	NA (NA)
10^{-6}	0.32 (1.76)	0.64 (2.53)	79.30 (28.23)	1.06 (3.27)	0.55 (2.32)	NA (NA)
5×10^{-7}	0.31 (2.47)	0.69 (3.71)	104.99 (46.75)	0.99 (4.48)	0.46 (3.00)	NA (NA)

Table S3: The mean and standard deviation of the ratio of the false positive rates to the nominal significance level based on 2,000 replications using Sobel's test, MaxP, JT-comp, HDMT, Sobel-comp and DACT under the *Sparse null 2* case. n is the sample size. The total number of mediators is 100,000. With probability $\pi_{01} = 0.001$, $\alpha = 0$ and $\beta \sim N(0, 0.3^2)$; with probability $\pi_{10} = 0.001$, $\alpha \sim N(0, 5 \times 0.3^2)$ and $\beta = 0$; with probability $\pi_{00} = 0.998$, $\alpha = \beta = 0$. R^2 is controlled in the mediator-exposure and outcome-mediator models.

Cut-off	Sobel	MaxP	JT-comp	HDMT	Sobel-comp	DACT
$n = 200, R^2 = 0.1$						
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.20 (0.11)	1.06 (0.12)	0.90 (0.35)	1.58 (1.83)
10^{-4}	0.00 (0.00)	0.00 (0.01)	2.02 (0.46)	1.16 (0.38)	0.89 (0.76)	3.21 (4.37)
10^{-5}	0.00 (0.00)	0.00 (0.03)	5.11 (2.36)	1.36 (1.21)	1.37 (1.73)	7.76 (12.62)
10^{-6}	0.00 (0.00)	0.00 (0.00)	16.45 (13.27)	1.85 (4.48)	2.99 (6.13)	22.41 (54.88)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	23.71 (22.35)	2.17 (6.83)	4.17 (9.82)	31.85 (91.44)
$n = 200, R^2 = 0.15$						
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.27 (0.11)	1.06 (0.12)	0.91 (0.36)	1.59 (1.83)
10^{-4}	0.00 (0.00)	0.00 (0.01)	2.76 (0.55)	1.17 (0.38)	0.94 (0.79)	3.22 (4.38)
10^{-5}	0.00 (0.00)	0.00 (0.04)	9.97 (3.44)	1.39 (1.22)	1.65 (2.00)	7.80 (13.07)
10^{-6}	0.00 (0.00)	0.00 (0.22)	42.82 (22.12)	1.98 (4.60)	4.44 (7.88)	23.25 (70.65)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	67.14 (38.99)	2.34 (7.02)	6.38 (12.50)	34.41 (129.76)
$n = 200, R^2 = 0.2$						
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.32 (0.11)	1.06 (0.12)	0.91 (0.36)	1.61 (1.84)
10^{-4}	0.00 (0.00)	0.00 (0.01)	3.46 (0.60)	1.17 (0.38)	0.97 (0.81)	3.24 (4.38)
10^{-5}	0.00 (0.00)	0.00 (0.04)	15.43 (4.35)	1.40 (1.22)	1.84 (2.18)	7.57 (11.98)
10^{-6}	0.00 (0.00)	0.00 (0.22)	79.90 (31.26)	2.00 (4.62)	5.64 (9.37)	20.57 (52.19)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	130.40 (55.40)	2.38 (7.10)	8.26 (15.14)	29.00 (93.88)
$n = 500, R^2 = 0.1$						
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.30 (0.11)	1.05 (0.11)	1.00 (0.29)	1.27 (1.63)
10^{-4}	0.00 (0.00)	0.00 (0.01)	3.61 (0.59)	1.14 (0.35)	1.20 (0.69)	2.85 (4.24)
10^{-5}	0.00 (0.00)	0.00 (0.03)	17.37 (4.26)	1.35 (1.16)	2.50 (2.12)	10.53 (23.75)
10^{-6}	0.00 (0.00)	0.00 (0.00)	93.96 (31.59)	1.93 (4.28)	8.26 (10.03)	66.33 (223.61)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	156.65 (57.45)	2.40 (6.80)	12.51 (17.02)	124.35 (446.18)
$n = 500, R^2 = 0.15$						
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.34 (0.11)	1.05 (0.11)	1.00 (0.30)	1.23 (1.66)
10^{-4}	0.00 (0.00)	0.00 (0.01)	4.63 (0.66)	1.14 (0.35)	1.24 (0.72)	2.57 (4.04)
10^{-5}	0.00 (0.00)	0.00 (0.03)	27.08 (5.19)	1.37 (1.18)	2.78 (2.29)	7.53 (17.31)
10^{-6}	0.00 (0.00)	0.00 (0.00)	173.38 (41.99)	2.07 (4.48)	9.87 (11.31)	35.27 (148.51)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	301.01 (79.72)	2.57 (7.07)	15.22 (19.09)	62.57 (295.70)
$n = 500, R^2 = 0.2$						
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.34 (0.10)	1.06 (0.11)	1.01 (0.30)	1.23 (1.67)
10^{-4}	0.00 (0.00)	0.00 (0.01)	5.37 (0.69)	1.15 (0.35)	1.28 (0.74)	2.47 (3.96)
10^{-5}	0.00 (0.00)	0.00 (0.03)	34.99 (5.75)	1.39 (1.19)	2.94 (2.40)	6.14 (12.83)
10^{-6}	0.00 (0.00)	0.00 (0.00)	244.80 (48.87)	2.11 (4.56)	11.00 (12.19)	20.89 (88.39)
5×10^{-7}	0.00 (0.00)	0.00 (0.00)	439.05 (94.04)	2.65 (7.18)	17.01 (20.63)	33.05 (173.35)
$n = 1000, R^2 = 0.1$						
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.34 (0.10)	1.05 (0.11)	1.03 (0.27)	1.14 (1.59)
10^{-4}	0.00 (0.00)	0.00 (0.01)	5.13 (0.69)	1.13 (0.34)	1.32 (0.68)	2.36 (3.86)
10^{-5}	0.00 (0.02)	0.00 (0.04)	32.49 (5.64)	1.40 (1.20)	3.01 (2.30)	6.91 (16.64)
10^{-6}	0.00 (0.00)	0.00 (0.22)	221.03 (46.60)	2.05 (4.61)	11.15 (12.16)	33.20 (145.14)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	392.65 (88.03)	2.55 (7.27)	17.08 (20.52)	59.07 (289.05)
$n = 1000, R^2 = 0.15$						
10^{-3}	0.00 (0.00)	0.00 (0.00)	1.32 (0.10)	1.05 (0.11)	1.04 (0.27)	1.13 (1.62)
10^{-4}	0.00 (0.01)	0.00 (0.01)	6.07 (0.74)	1.14 (0.34)	1.36 (0.70)	2.22 (3.74)
10^{-5}	0.00 (0.02)	0.00 (0.04)	43.05 (6.25)	1.43 (1.21)	3.30 (2.45)	5.10 (10.19)
10^{-6}	0.00 (0.00)	0.00 (0.22)	321.66 (55.73)	2.15 (4.71)	12.68 (13.18)	14.34 (51.36)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	588.52 (105.67)	2.69 (7.47)	19.84 (22.59)	20.69 (95.92)
$n = 1000, R^2 = 0.2$						
10^{-3}	0.00 (0.00)	0.00 (0.01)	1.28 (0.09)	1.05 (0.11)	1.04 (0.27)	1.15 (1.63)
10^{-4}	0.00 (0.01)	0.00 (0.01)	6.73 (0.76)	1.14 (0.34)	1.39 (0.71)	2.26 (3.75)
10^{-5}	0.00 (0.02)	0.00 (0.04)	50.70 (6.65)	1.45 (1.22)	3.48 (2.57)	5.01 (9.17)
10^{-6}	0.00 (0.00)	0.00 (0.22)	397.30 (59.78)	2.27 (4.83)	13.80 (14.02)	12.86 (26.81)
5×10^{-7}	0.00 (0.00)	0.01 (0.45)	737.90 (116.72)	2.82 (7.62)	21.51 (23.94)	17.11 (37.74)

Table S4: The mean and standard deviation of the ratio of the false positive rates to the nominal significance level based on 2000 replications using Sobel's test, MaxP test, JT-comp, HDMT, Sobel-comp and DACT under the *Sparse null 2* case. n is the sample size. The total number of mediators is 100,000. With probability $\pi_{01} = 0.33$, $\alpha = 0$ and $\beta \sim N(0, 0.3^2)$; with probability $\pi_{10} = 0.33$, $\alpha \sim N(0, 5 \times 0.3^2)$ and $\beta = 0$; with probability $\pi_{00} = 0.34$, $\alpha = \beta = 0$. R^2 is controlled in the mediator-exposure and outcome-mediator models.

Cut-off	Sobel	MaxP	JT-comp	HDMT	Sobel-comp	DACT
$n = 200, R^2 = 0.1$						
10^{-3}	0.01 (0.01)	0.37 (0.06)	0.78 (0.09)	1.12 (0.11)	0.06 (0.03)	1.06 (0.50)
10^{-4}	0.00 (0.01)	0.35 (0.19)	0.33 (0.18)	1.26 (0.36)	0.01 (0.03)	1.32 (0.94)
10^{-5}	0.00 (0.00)	0.37 (0.61)	0.12 (0.35)	1.51 (1.24)	0.00 (0.02)	1.73 (1.98)
10^{-6}	0.00 (0.00)	0.35 (1.87)	0.05 (0.67)	1.92 (4.37)	0.00 (0.00)	2.39 (5.39)
5×10^{-7}	0.00 (0.00)	0.27 (2.31)	0.02 (0.63)	2.05 (6.39)	0.00 (0.00)	2.79 (8.10)
$n = 200, R^2 = 0.15$						
10^{-3}	0.04 (0.02)	0.42 (0.07)	0.69 (0.09)	1.13 (0.11)	0.14 (0.05)	3.18 (1.06)
10^{-4}	0.01 (0.03)	0.45 (0.21)	0.23 (0.15)	1.27 (0.35)	0.03 (0.06)	6.05 (2.76)
10^{-5}	0.00 (0.04)	0.54 (0.75)	0.06 (0.24)	1.51 (1.22)	0.01 (0.08)	12.24 (7.61)
10^{-6}	0.00 (0.00)	0.66 (2.53)	0.01 (0.32)	1.96 (4.40)	0.00 (0.00)	25.58 (23.17)
5×10^{-7}	0.00 (0.00)	0.64 (3.58)	0.01 (0.45)	2.24 (6.74)	0.00 (0.00)	31.95 (33.88)
$n = 200, R^2 = 0.2$						
10^{-3}	0.09 (0.03)	0.44 (0.07)	0.63 (0.08)	1.13 (0.11)	0.23 (0.06)	5.58 (1.47)
10^{-4}	0.03 (0.05)	0.47 (0.21)	0.17 (0.13)	1.27 (0.35)	0.08 (0.09)	12.54 (4.42)
10^{-5}	0.01 (0.08)	0.60 (0.79)	0.04 (0.19)	1.52 (1.22)	0.03 (0.16)	29.89 (13.90)
10^{-6}	0.00 (0.00)	0.76 (2.72)	0.00 (0.22)	1.96 (4.39)	0.00 (0.00)	74.23 (45.75)
5×10^{-7}	0.00 (0.00)	0.78 (3.92)	0.00 (0.00)	2.21 (6.73)	0.00 (0.00)	99.05 (67.99)
$n = 500, R^2 = 0.1$						
10^{-3}	0.09 (0.03)	0.40 (0.06)	0.54 (0.07)	1.05 (0.10)	0.23 (0.05)	6.12 (1.56)
10^{-4}	0.02 (0.05)	0.40 (0.20)	0.11 (0.11)	1.09 (0.33)	0.07 (0.09)	13.78 (4.71)
10^{-5}	0.01 (0.09)	0.41 (0.65)	0.02 (0.14)	1.17 (1.07)	0.02 (0.13)	32.60 (14.54)
10^{-6}	0.00 (0.22)	0.38 (1.94)	0.00 (0.22)	1.15 (3.43)	0.01 (0.32)	78.78 (47.59)
5×10^{-7}	0.00 (0.00)	0.43 (2.90)	0.00 (0.00)	1.16 (4.89)	0.01 (0.45)	103.77 (69.89)
$n = 500, R^2 = 0.15$						
10^{-3}	0.16 (0.04)	0.43 (0.07)	0.49 (0.07)	1.05 (0.10)	0.34 (0.06)	10.31 (1.98)
10^{-4}	0.07 (0.09)	0.43 (0.20)	0.08 (0.09)	1.09 (0.33)	0.18 (0.13)	26.86 (6.84)
10^{-5}	0.03 (0.17)	0.43 (0.66)	0.01 (0.10)	1.16 (1.07)	0.07 (0.27)	72.74 (23.79)
10^{-6}	0.02 (0.45)	0.40 (1.97)	0.00 (0.00)	1.15 (3.43)	0.03 (0.55)	202.01 (85.94)
5×10^{-7}	0.01 (0.45)	0.45 (2.97)	0.00 (0.00)	1.14 (4.81)	0.04 (0.89)	275.85 (127.94)
$n = 500, R^2 = 0.2$						
10^{-3}	0.21 (0.05)	0.46 (0.07)	0.46 (0.07)	1.05 (0.10)	0.41 (0.07)	13.65 (2.14)
10^{-4}	0.13 (0.11)	0.45 (0.21)	0.06 (0.08)	1.09 (0.33)	0.26 (0.16)	38.47 (7.96)
10^{-5}	0.06 (0.26)	0.45 (0.67)	0.01 (0.09)	1.17 (1.06)	0.13 (0.36)	111.80 (29.46)
10^{-6}	0.04 (0.59)	0.43 (2.07)	0.00 (0.00)	1.15 (3.42)	0.06 (0.77)	333.32 (112.87)
5×10^{-7}	0.04 (0.89)	0.48 (3.13)	0.00 (0.00)	1.17 (4.90)	0.07 (1.18)	464.10 (171.83)
$n = 1000, R^2 = 0.1$						
10^{-3}	0.18 (0.04)	0.45 (0.07)	0.45 (0.07)	1.03 (0.10)	0.38 (0.06)	12.49 (2.16)
10^{-4}	0.11 (0.10)	0.43 (0.21)	0.06 (0.08)	1.06 (0.33)	0.22 (0.15)	34.29 (7.83)
10^{-5}	0.05 (0.22)	0.42 (0.65)	0.00 (0.07)	1.12 (1.06)	0.11 (0.34)	97.28 (27.95)
10^{-6}	0.01 (0.32)	0.42 (2.02)	0.00 (0.00)	1.19 (3.48)	0.04 (0.59)	282.24 (105.72)
5×10^{-7}	0.00 (0.00)	0.48 (3.06)	0.00 (0.00)	1.28 (5.10)	0.00 (0.00)	391.28 (159.65)
$n = 1000, R^2 = 0.15$						
10^{-3}	0.24 (0.05)	0.48 (0.07)	0.42 (0.06)	1.04 (0.10)	0.46 (0.07)	16.73 (2.26)
10^{-4}	0.17 (0.13)	0.47 (0.22)	0.05 (0.07)	1.06 (0.33)	0.33 (0.18)	49.76 (8.79)
10^{-5}	0.11 (0.33)	0.47 (0.68)	0.00 (0.05)	1.13 (1.06)	0.21 (0.46)	152.08 (33.59)
10^{-6}	0.05 (0.71)	0.47 (2.13)	0.00 (0.00)	1.24 (3.52)	0.12 (1.09)	474.34 (135.54)
5×10^{-7}	0.03 (0.77)	0.53 (3.21)	0.00 (0.00)	1.23 (4.93)	0.10 (1.41)	669.21 (208.14)
$n = 1000, R^2 = 0.2$						
10^{-3}	0.28 (0.05)	0.51 (0.07)	0.40 (0.06)	1.04 (0.10)	0.51 (0.07)	19.59 (2.19)
10^{-4}	0.22 (0.15)	0.50 (0.22)	0.04 (0.06)	1.07 (0.34)	0.39 (0.20)	60.79 (8.92)
10^{-5}	0.16 (0.39)	0.50 (0.70)	0.00 (0.03)	1.12 (1.04)	0.29 (0.54)	193.25 (35.62)
10^{-6}	0.12 (1.07)	0.49 (2.17)	0.00 (0.00)	1.22 (3.43)	0.23 (1.50)	626.61 (146.57)
5×10^{-7}	0.09 (1.34)	0.56 (3.30)	0.00 (0.00)	1.21 (4.89)	0.18 (1.89)	896.07 (228.65)

Table S5 : Four mediation pathways identified by DACT after controlling the FWER at 0.05. The exposure is adult SES and the outcome is HbA1c. The total number of mediators is 228,088. In the mediator-exposure model, we adjusted for age, sex, race and residual white blood cell types (neutrophils, B cells, T cells, and natural killer cells), and included the methylation chip and position as random effects to account for potential batch effects. In addition, we adjusted for the exposure in the outcome-mediator model. The estimated mediation effect is $\hat{\alpha}\hat{\beta}$ and the proportion of mediation effect is provided in the parenthesis. The 95% confidence interval (CI) is calculated based on 1,000 bootstrap samples.

CpG	Chr	Gene	UCSC RefGene Group	$\hat{\alpha}$	$\hat{\beta}$	Mediation effect (proportion)	95% CI	p_{DACT}
cg10508317	17	SOCS3	Body	-9.92E-02	-1.78E-01	0.018 (0.18)	(0.009,0.029)	5.35E-09
cg01288337	14	RIN3	Body	5.55E-02	3.06E-01	0.017 (0.17)	(0.009,0.028)	7.32E-09
cg10244976	16	LMF1	Body	-1.19E-01	-1.28E-01	0.015 (0.15)	(0.007,0.025)	5.47E-08
cg21263566	1	TLCD4	Body	9.00E-02	1.59E-01	0.014 (0.14)	(0.006,0.024)	1.81E-07