Large-Scale Hypothesis Testing for Causal Mediation Effects with Applications in Genome-wide Epigenetic Studies

Zhonghua Liu, Jincheng Shen, Richard Barfield,
Joel Schwartz, Andrea A. Baccarelli and Xihong Lin *

Abstract

In genome-wide epigenetic studies, it is of great scientific interest to assess whether the effect of an exposure on a clinical outcome is mediated through DNA methylations. However, statistical inference for causal mediation effects is challenged by the fact that one needs to test a large number of composite null hypotheses across the whole epigenome. Two popular tests, the Wald-type Sobel's test and the joint significant test are underpowered and thus can miss important scientific discoveries. In this paper, we show that the null distribution of Sobel's test is not the standard normal distribution and the null distribution of the joint significant test is not uniform under the composite null of no mediation effect, especially in finite samples and under the singular point null case that the exposure has no effect on the mediator and the mediator has no effect on the outcome. Our results clearly explain why these two tests are underpowered, and more importantly motivate us to develop a more powerful Divide-Aggregate Composite-null Test (DACT) for the composite null hypothesis of no mediation effect by leveraging epigenome-wide data. We adopted Efron's empirical null framework for assessing statistical significance. We show that the proposed DACT method has improved power, and can well control type I error rate. Our extensive simulation studies showed that the DACT method properly controls the type I error rate and outperforms Sobel's test and the joint significance test for detecting mediation effects. We applied the DACT method to the Normative Aging Study and identified additional DNA methylation CpG sites that might mediate the effect of smoking on lung function. We then performed a comprehensive sensitivity analysis to demonstrate that our mediation data analysis results were robust to unmeasured confounding. We also developed a computationally-efficient R package *DACT* for public use, available at `https://github.com/zhonghualiu/DACT`.

Key words: Causal mediation analysis; Composite null; Divide-aggregate composite-null test; Hypothesis testing; Indirect effects; Genome-wide epigenetic studies; Mediation effects; Proportions of true nulls.

*Zhonghua Liu is Assistant Professor in the Department of Statistics and Actuarial Science at the University of Hong Kong, Jincheng Shen is Assistant Professor in the Department of Population Health Sciences at University of Utah School of Medicine, Richard Barfield is Biostatistician in the Department of Biostatistics and Bioinformatics at Duke University School of Medicine. Joel Schwartz is Professor of Environmental Epidemiology at Harvard T.H. Chan School of Public Health, Andrea A. Baccarelli is Leon Hess Professor of Environmental Health Sciences at Mailman School of Public Health, Columbia University. Xihong Lin is Professor of Biostatistics at Harvard T.H. Chan School of Public Health and Professor of Statistics at Faculty of Arts and Sciences, Harvard University. This work was supported by the National Institutes of Health grants R35-CA197449, P01-CA134294, U01-HG009088, U19-CA203654, R01-HL113338, P30 ES000002, and T32GM074897.

1 Introduction

Cigarette smoking is a well-known risk factor for reduced lung function (Tommola et al. 2016). It is thus of scientific interest to investigate the underlying causal mechanism and epigenetic pathway of the observed association between smoking and lung function. Motivated by the ongoing Normative Aging Genome-Wide Epigenetic Study that will be described in Section 6, we are interested in studying whether the effect of smoking on lung function is mediated by DNA methylations. DNA methylation is a heritable epigenetic mechanism that occurs by the covalent addition of a methyl (CH_3) group to the base cytosine (C) at its 5-position within the CpG dinucleotide. The term CpG refers to the base cytosine (C) linked by a phosphate bond to the base guanine (G) in the DNA nucleotide sequence. Aberrations in the DNA methylations can affect downstream gene expressions and thus have an important role in the etiology of human diseases. There is increasing evidence that epigenetic mechanisms serve to integrate genetic and environmental causes of complex traits and diseases (Liu et al. 2013; Bind et al. 2014). Since DNA methylation is a reversible biological process (Wu and Zhang 2014), mediation analysis results can help discover novel epigenetic pathways as potential therapeutic targets.

Causal mediation analysis is a useful statistical method to answer the scientific question of whether DNA methylation mediates the effect of smoking on lung function. In the causal inference framework, the natural indirect effect (NIE) measures the evidence of mediation effect of an exposure on an outcome through a mediator (Robins and Greenland 1992; Pearl 2001) and is often of primary scientific interest. The classical regression approach to mediation analysis proposed by Baron and Kenny (1986) is a widely used method in social sciences for continuous outcomes and mediators, where the mediation effect is the product of the exposure-mediator and mediator-outcome effects, and is more generally referred to as the product method. This classical product method for mediation analysis is equivalent to the NIE defined in modern causal inference framework when the exposure-mediator interaction is absent (VanderWeele and Vansteelandt 2009; Valeri and VanderWeele 2013).

As the mediation effect is composed of the product of two parameters, MacKinnon et al. (2002) pointed out that the null hypothesis of no mediation effect is composite in the single mediation effect testing settings. Indeed, MacKinnon et al. (2002) found through simulation study that the Wald-type Sobel's test (Sobel 1982) is overly conservative and thus underpowered, and recommended researchers to use the slightly more powerful joint significance test (also known as the MaxP test)

for detecting mediation effects. However, both the Sobel's test and the MaxP test perform poorly in genome-wide epigenetic studies as demonstrated empirically by Barfield et al. (2017). There are three reasons: (1) the association signals are generally weak and sparse with limited sample sizes; (2) the heavy multiple testing burden to be adjusted; (3) the composite null nature of the mediation effect testing that has not been taken into account.

For a variable to serve as a causal mediator in the pathway from an exposure to an outcome, it must satisfy the following two conditions simultaneously: (1) the exposure has an effect on the mediator; (2) the mediator has an effect on the outcome. The null hypothesis of no mediation effect is thus composite and consists of three cases: (1) the exposure has no effect on the mediator, the mediator has an effect on the outcome; (2) the exposure has an effect on the mediator, the mediator has no effect on the outcome; (3) the exposure has no effect on the mediator, and the mediator has no effect on the outcome. This salient feature of the composite null hypothesis imposes great statistical challenges for making inference on the mediation effect, and the uncertainty associated with the three cases under the composite null hypothesis should be taken into account when constructing valid and powerful testing procedures.

One attempt is the MT-Comp method proposed by Huang (2019), which however only works when the sample size is small, as the type I error rate of MT-Comp can be inflated when the sample size is large as stated in the original paper. This is because the MT-Comp method assumes that the association signals (an increasing function of the sample size) of exposure-mediator or/and the mediator-outcome are weak and sparse, which will be violated when the sample size is large. Therefore, it is pressing to develop statistically valid and powerful testing procedures to detect mediation effects that are suitable for general use in large-scale genome-wide epigenetic studies.

The main goal of this paper is to develop a valid and powerful large-scale testing procedure for detecting causal mediation effects by leveraging data from epigenome-wide DNA methylation studies. First, we study the statistical properties of the commonly used tests for causal mediation effects, Sobel's test and the joint significance test. We show that the joint significance test is the likelihood ratio test for the composite null hypothesis of no mediation effect, and derive the null distributions of Sobel's test and the joint significance test. Our results show that they follow non-standard distributions, and both the Sobel test and the MaxP test are conservative in the sense that their actual sizes are always smaller than the nominal significance level for any fixed sample size, and the MaxP test is the likelihood ratio test and is always more powerful than the Sobel's test, but it is still under-powered to detect mediation effects in genome-wide epigenetic studies. We

also studied the powers of these two tests analytically and found that their powers are maximized when the association signals of the exposure-mediator and mediator-outcome are of equal strength. Our results clearly and rigorously explain why these two popular tests are underpowered and thus not suitable for large-scale inference for mediation effects.

To overcome the limitations of Sobel's test and the joint significance test, we propose the Divide-Aggregate Composite-null Test (DACT), which improves the power by leveraging the whole genome DNA methylation data in a way that large-scale mediation effect testing is a blessing rather than a curse. Specifically, genome-wide data allow us to estimate the relative proportions of the three null cases that can be incorporated into the construction of the DACT test statistic as a composite p-value obtained by averaging the case-specific p-values weighted using the estimated case proportions. The DACT statistic follows a uniform distribution on the interval $[0, 1]$ approximately if the exposure-mediator or the mediator-outcome association signals are sparse. It can depart from the uniform distribution when such signals are not sparse. To address this issue, we further propose to use Efron's empirical null framework for inference (Efron 2004), where the empirical null distribution can be consistently estimated using the method developed by Jin and Cai (2007). We also study the statistical properties of the DACT method. We show that the proposed DACT method works well in both simulation studies and real data analysis of the Normative Aging Study (NAS), and outperforms Sobel's test and the MaxP test substantially. We also perform a comprehensive sensitivity analysis to evaluate the robustness of our analysis results with respect to the no unmeasured confounding assumption.

The rest of our paper is organized as follows. In Section 2, we present the regression models for causal mediation analysis, derive the null distributions of Sobel's test and the joint significant test, and then discuss the limitations of these two tests in genome-wide epigenetic studies. In Section 3, we propose the DACT testing procedure and study its statistical properties. In Section 4, we discuss the connections and differences of Sobel's test, the MaxP test and our DACT. In Section 5, we conduct extensive simulation studies to evaluate the type I error rates of DACT along with Sobel's test and the joint significant test, and compare their powers under various alternatives. In Section 6, we apply the DACT method to the Normative Aging Genome-Wide Epigenetic Study to detect the mediation effects of DNA methylation CpG sites in the causal pathway from smoking behavior to lung functions. The paper ends with discussions in Section 7.

2 Causal Mediation Analysis

2.1 Assumptions and Regression Models

Let A denote an exposure, Y a continuous outcome, M a continuous mediator and $\boldsymbol{X}$ additional covariates to adjust for confounding. Baron and Kenny (1986) proposed the following linear structural equation models for the outcome and the mediator

$$Y = \beta_0 + \beta_A A + \beta M + \boldsymbol{\beta}_X^T \boldsymbol{X} + \epsilon_Y, \quad (1)$$

$$M = \gamma_0 + \gamma A + \boldsymbol{\gamma}_X^T \boldsymbol{X} + \epsilon_M, \quad (2)$$

where ϵ_Y and ϵ_M are the error terms with mean zeros and constant variances, which are also uncorrelated under the standard assumptions (1)-(5) stated below in causal mediation analysis (Imai et al. 2010). The constant variance assumption is found to be reasonable when the methylation level is on the M-value scale (Du et al. 2010). It is well-known that the least squares estimation method gives unbiased parameter estimators in models (1) and (2). If the outcome Y is binary and rare, then we can fit the following logistic models using the maximum likelihood estimation (MLE) method

$$logit(Pr(Y = 1|A, M, \boldsymbol{X})) = \beta_0 + \beta_A A + \beta M + \boldsymbol{\beta}_X^T \boldsymbol{X}. \quad (3)$$

Our primary interest is the so called Natural Indirect Effect (NIE) defined by Robins and Greenland (1992) and Pearl (2001), which measures the effect of the exposure on the outcome mediated through the mediator. In the modern causal inference framework, one assumes the following standard identification assumptions for estimating the NIE (VanderWeele and Vansteelandt 2009; Valeri and VanderWeele 2013): (1) There are no unmeasured exposure-outcome confounders given $\boldsymbol{X}$; (2) There are no unmeasured mediator-outcome confounders given $(\boldsymbol{X}, A)$; (3) There are no unmeasured exposure-mediator confounders given $\boldsymbol{X}$; (4) There is no effect of exposure that confounds the mediator-outcome relationship; (5) There is no exposure and mediator interaction on the outcome. Under these standard assumptions, the NIE (mediation effect) can be identified. When both the mediator and the outcome are continuous, the NIE is equal to $\beta\gamma$. When the mediator is continuous, and the outcome is binary and rare, the NIE is approximately equal to $\beta\gamma$ on the log-odds-ratio scale (Valeri and VanderWeele 2013). Graphically, the NIE measures the effect of the causal chain $A \to M \to Y$ as shown in a directed acyclic graph (DAG) in Figure 1. We assume that the covariates (possibly vector-valued) $\boldsymbol{X}$ contain all the confounders. In practice,

there might be unmeasured confounders U omitted from mediation analysis. Sensitivity analysis can be performed to assess the robustness of data analysis results, for example, using the method proposed by Imai et al. (2010) as we will do in Section 6.

Under the assumptions (1)-(5), the causal effect of the exposure A on the mediator M is independent of the causal effect of the mediator M on the outcome Y (Figure 1). We now show this simple but important result, which will be used in Section 3 to simplify the estimation and testing procedure. Specifically, under the assumptions (1)-(5), the joint probability density function of $(Y, M, A, \boldsymbol{X})$ can be factored as $f(Y, M, A, \boldsymbol{X}) = f(Y|M, A, \boldsymbol{X})f(M|A, X)f(A, X)$, where $f(A, X)$ can be discarded because it is ancillary and does not contain any information about the model parameters in equations (1) - (3). Therefore, we only need $f(Y|M, A, \boldsymbol{X})f(M|A, X)$ for the inference of unknown parameters. The $A \to M$ association is contained in $f(M|A, X)$, and the $M \to Y$ association is contained in $f(Y|M, A, \boldsymbol{X})$. Denote the log-likelihood as $\ell(\cdot)$, then we have $\partial\ell(\cdot)/\partial\beta\partial\gamma = 0$. This implies that the two parameters β and γ are independent. This result will be used throughout the whole paper.

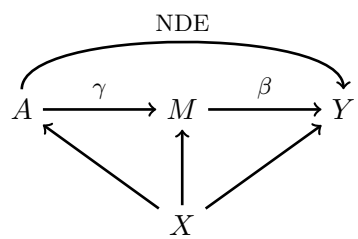

Figure 1: A causal DAG for mediation analysis. A is the exposure, M is the mediator, Y is the outcome and $\boldsymbol{X}$ represents measured confounders. γ is the causal effect of A on M and β is the causal effect of M on Y. NDE stands for Natural Direct Effect.

In genome-wide epigenetic studies, we are interested in assessing whether a particular DNA methylation CpG site lies in the causal pathway from an exposure to a clinical outcome. This can be formulated as the following hypothesis testing problem

$$H_0 : \beta\gamma = 0 \text{ versus } H_1 : \beta\gamma \neq 0. \tag{4}$$

As mentioned in Section 1, the null hypothesis H_0: $\beta\gamma = 0$ is composite and the null parameter space can be decomposed into three disjoint cases,

$$H_0 : \begin{cases} \text{Case 1}: & \beta \neq 0, \gamma = 0; \\ \text{Case 2}: & \beta = 0, \gamma \neq 0; \\ \text{Case 3}: & \beta = 0, \gamma = 0. \end{cases} \tag{5}$$

The fourth case: $\beta \neq 0, \gamma \neq 0$ corresponds to the alternative hypothesis. In practice, we can fit the outcome and mediator regression models and obtain consistent estimates $\widehat{\beta}$ and $\widehat{\gamma}$ for the regression coefficients β and γ respectively. We have the following standard normal approximation

$$\frac{\widehat{\beta} - \beta}{\widehat{\sigma}_\beta} \sim N(0,1), \qquad \frac{\widehat{\gamma} - \gamma}{\widehat{\sigma}_\gamma} \sim N(0,1),$$

where $\widehat{\sigma}_\beta$ and $\widehat{\sigma}_\gamma$ are the estimated standard errors for $\widehat{\beta}$ and $\widehat{\gamma}$ respectively. A consistent point estimator for the mediation effect is $\widehat{\beta}\widehat{\gamma}$. A rejection of the null hypothesis H_0: $\beta\gamma = 0$ suggests the presence of a mediation effect by M.

2.2 The Wald-type Sobel's Test

Using the first-order multivariate delta method, Sobel (1982) obtained the standard error for the product-method estimator $\widehat{\beta}\widehat{\gamma}$ and proposed the following test statistic to detect the mediation effect

$$T_{Sobel} = \frac{\widehat{\beta}\widehat{\gamma}}{\sqrt{\widehat{\gamma}^2\widehat{\sigma}_\beta^2 + \widehat{\beta}^2\widehat{\sigma}_\gamma^2}}.$$

Note that the covariance term between $\widehat{\beta}$ and $\widehat{\gamma}$ was set to zero here because $\widehat{\beta}$ and $\widehat{\gamma}$ are independent of each other. To determine statistical significance, Sobel (1982) used the standard normal distribution as the reference distribution to calculate the p-value of T_{Sobel}. MacKinnon et al. (1998) found that the Sobel's test has low power via simulation studies but did not explain theoretically why the Sobel's test is underpowered.

To provide statistically rigorous guidance for applied researchers on using Sobel' test, we now investigate the statistical properties of Sobel's test and show why it is underpowered. First, we show that under the composite null, Sobel's test is conservative for any finite sample size but has correct type I error rate asymptotically in the null Case 1 and Case 2. While in the null Case 3, Sobel's test is always conservative even asymptotically. The fundamental reason is that the first-order multivariate delta method fails because the gradient is $(0,0)$, and the usual asymptotic normal approximation for the null distribution of Sobel's test is thus incorrect in the null Case 3. Our result explains clearly and rigorously why Sobel's test is underpowered.

For the ease of exposition, we introduce some notation. Denote $Z_\beta = \widehat{\beta}/\widehat{\sigma}_\beta$ and $Z_\gamma = \widehat{\gamma}/\widehat{\sigma}_\gamma$. We write Z_β as $Z_\beta(n)$ and Z_γ as $Z_\gamma(n)$ to emphasize that those two statistics depend on the sample

size n. Direct calculation gives

$$\begin{aligned}\mu_\beta(n) = E\{Z_\beta(n)\} &\approx \sqrt{n}\beta\frac{\sigma_M}{\sigma_Y}\sqrt{1-R^2_{M|A,X}},\\ \mu_\gamma(n) = E\{Z_\gamma(n)\} &\approx \sqrt{n}\gamma\frac{\sigma_A}{\sigma_M}\sqrt{1-R^2_{A|X}},\end{aligned}$$

where σ_A is the standard deviation of exposure A, $R^2_{A|X}$ is the coefficient of determination by regressing exposure A on the covariates X, and $R^2_{M|A,X}$ is the coefficient of determination of the mediator regression model (2). In what follows, $\mu_\gamma(n)$ and $\mu_\beta(n)$ will be referred to as the association signals for the exposure-mediator and mediator-outcome relationships respectively.

It is reasonable to assume that $R^2_{A|X} \neq 1$ and $R^2_{M|A,X} \neq 1$. We then can rewrite T_{Sobel} as

$$T_{Sobel} = \frac{Z_\beta(n)Z_\gamma(n)}{\sqrt{Z^2_\beta(n)+Z^2_\gamma(n)}} = \frac{Z_\gamma(n)}{\sqrt{(Z_\gamma(n)/Z_\beta(n))^2+1}}. \tag{6}$$

This representation of Sobel's test statistic can help us better understand its behavior. In the null Case 1, the size of Sobel's test is strictly smaller than the nominal significance level α for any finite sample size by noting the following result

$$P(|T_{Sobel}| > Z_{1-\alpha/2}) < P(|Z_\gamma(n)| > Z_{1-\alpha/2}) = \alpha,$$

where $Z_{1-\alpha/2}$ denotes the $1-\alpha/2$ percentile of the standard normal distribution. We observe that the conservativeness of Sobel's test in null Case 1 can be alleviated when the sample size goes to infinity. To show this result, without loss of generality, we can assume that $\beta > 0$. Then we have $\mu_\beta(n) \to +\infty$ as the sample size $n \to \infty$. It can be easily seen that $\{Z_\beta(n)\}^{-1}$ converges to zero and $Z_\gamma(n)$ is bounded in probability, therefore the ratio $Z_\gamma(n)/Z_\beta(n)$ converges to zero in probability. Using Slutsky's theorem, T_{Sobel} follows the standard normal distribution asymptotically. Therefore, the Sobel's test has correct size asymptotically, but is conservative for finite sample sizes in Case 1. The same conclusion holds in the null Case 2.

In the null Case 3, the ratio $Z_\gamma(n)/Z_\beta(n)$ is stochastically bounded and in fact follows the standard Cauchy distribution asymptotically. The central limit theorem cannot be applied to the test statistic T_{Sobel} in this case and the asymptotic distribution of T_{Sobel} is not the standard normal, but is normal with mean zero and variance equal to $\frac{1}{4}$ asymptotically. This explains why it is incorrect to use the standard normal distribution as the reference distribution to calculate p-value for Sobel's test. The actual type I error rate is much smaller than the nominal significance level α even asymptotically. The conservativeness of Sobel's test cannot be alleviated in the Case

3 even with increased sample size. We summarize our findings about Sobel's test in Result 1, with proofs provided in the Supplementary Materials.

Result 1 *Sobel's statistic T_{Sobel} for testing the composite null of no mediation effect (4) has the following properties:*

(a) T^2_{Sobel} follows the same distribution as the inverse of the sum of two independent standard Lévy variables (inverse chi-squared random variables with one degree of freedom) asymptotically.

(b) Under the composite null (4), in Cases 1 and 2, T_{Sobel} follows $N(0,1)$ asymptotically; In Case 3, T_{Sobel} follows $N(0,\frac{1}{4})$ or equivalently $4T^2_{Sobel}$ follows χ^2_1 distribution asymptotically.

(c) The power of the Sobel test given the significance level α can be calculated analytically as

$$Power \;=\; \int\int_{\{\frac{1}{x^2}+\frac{1}{y^2}\leq\frac{1}{C_\alpha}\}} \frac{1}{2\pi} e^{-\frac{(x-\mu_\gamma(n))^2}{2}-\frac{(y-\mu_\beta(n))^2}{2}} dxdy, \tag{7}$$

where C_α is the critical value at the significance level α. The power of the Sobel's test is maximized when $|\mu_\beta(n)| = |\mu_\gamma(n)|$ for a fixed NIE strength.

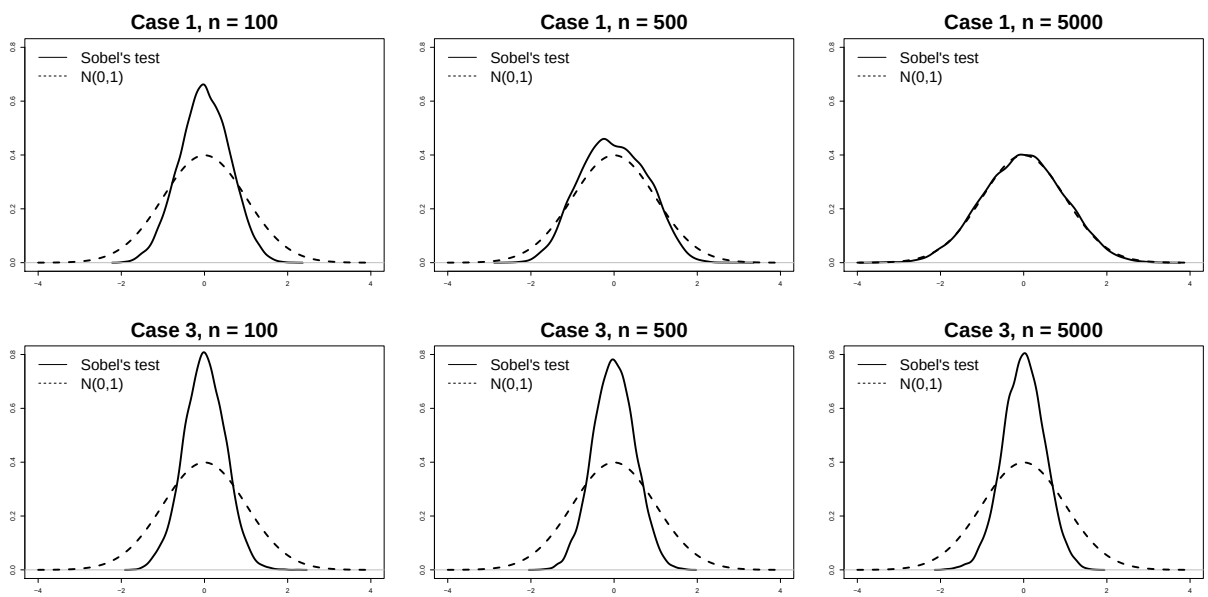

Figure 2: The kernel density estimates (the solid lines) of the probability density functions of T_{Sobel} in the null Case 1 and Case 3 with increasing sample sizes. $\sigma_A/\sigma_M = 1$, $R^2_{A|X} = 0.75$, $\sigma_M/\sigma_Y = 1$, $R^2_{M|A,X} = 0.75$. The upper panel is for null Case 1 ($\beta = 0.2, \gamma = 0$) and the lower panel is for null Case 3 ($\beta = \gamma = 0$) with sample sizes $n = 100, 500, 5000$. In null Case 3, the variance of T_{Sobel} is estimated to be 0.25. The dashed lines represent the probability density functions of the standard normal $N(0,1)$.

Figure 2 shows the empirical distributions of T_{Sobel} in the null Case 1 (upper panels) and Case 3 (lower panels). In the null Case 1, we set $\beta = 0.2, \gamma = 0$; while in the null Case 3, we set $\beta = \gamma = 0$. Sample sizes $n = 100, 500, 5000$ are considered in both Case 1 and Case 3. We first generate random samples for Z_β and Z_γ, and then use the formula (6) to get random samples for T_{Sobel}. The density function of T_{Sobel} was estimated by the kernel density estimator using the R function *density* with its default setting. We then compare the density function plots of T_{Sobel} to the standard normal density function under various scenarios. We found that the normal approximation for the test statistic T_{Sobel} improves as the sample size increases in the null Case 1, but the standard normal approximation fails even with increased sample size in the null Case 3.

2.3 The Joint Significance (MaxP) Test

The joint significance test, also known as the MaxP test (MacKinnon et al. 2002), was developed based on the argument we have already stated in Section 1 that one can claim the presence of mediation effects if the following two conditions are satisfied simultaneously: (1) the exposure has an effect on the mediator; (2) the mediator has an effect on the outcome. Let $p_\beta = 2(1 - \Phi(|Z_\beta|))$ be the p-value for testing H_0: $\beta = 0$, which is uniformly distributed on the interval $[0, 1]$ when $\beta = 0$ holds, and will converge to zero in probability when $\beta \neq 0$. Let $p_\gamma = 2(1 - \Phi(|Z_\gamma|))$ be the p-value for testing H_0: $\gamma = 0$, which is uniformly distributed on the interval $[0, 1]$ when $\gamma = 0$ holds, and will converge to zero in probability when $\gamma \neq 0$. Define MaxP $= \max(p_\beta, p_\gamma)$. Then, the MaxP test declares statistical significance for testing the composite null H_0: $\beta\gamma = 0$ if MaxP $< \alpha$. Intuitively, the MaxP test requires that both p_β and p_γ are significant by rejecting both H_0: $\beta = 0$ and H_0: $\gamma = 0$ individually. This testing procedure has an intuitive appeal and is easy to interpret, and hence has been widely used by applied researchers. MacKinnon et al. (2002) found that the MaxP test is slightly more powerful than Sobel's test using simulation studies, but did not provide any explanation for this empirical observation.

We now show that the MaxP test is conservative for testing H_0: $\beta\gamma = 0$ in all the three null cases for any finite sample size. First, since p_β and p_γ are independent, we have

$$Pr(MaxP < \alpha) = Pr(p_\beta < \alpha)Pr(p_\gamma < \alpha).$$

In Case 1, $Pr(p_\gamma < \alpha | \gamma = 0) = \alpha$ and $Pr(p_\beta < \alpha | \beta \neq 0) < 1$ for any finite sample size, so $Pr(MaxP < \alpha) < \alpha$. Thus, the MaxP test is conservative in Case 1 for any finite sample size.

However, if the sample size goes to infinity, then

$$Pr(p_\beta < \alpha) = Pr(|Z_n(\beta)| > Z_{1-\alpha/2}) \to 1,\ Pr(MaxP < \alpha) \to Pr(p_\gamma < \alpha) = \alpha,\ n \to \infty.$$

Therefore, the MaxP test has correct size and is equivalent to p_γ asymptotically in the null Case 1. Likewise, the MaxP test has correct size and is equivalent to p_β asymptotically in the null Case 2. In the null Case 3, both p_β and p_γ are uniformly distributed. Therefore, $Pr(p_\beta < \alpha) = Pr(p_\gamma < \alpha) = \alpha$, and $Pr(MaxP < \alpha) = \alpha^2 < \alpha$ for any $\alpha \in (0, 1)$ and any sample size. Thus, the MaxP test is always conservative regardless of sample size in the null Case 3. Traditionally, the MaxP test statistic itself is treated as a p-value, which is correct in the null Case 1 and 2 asymptotically, but is incorrect in the null Case 3. In the next section, we will propose a new testing procedure that can greatly improve the power of the MaxP test in large-scale multiple testing settings.

Result 2 states that the MaxP test is the likelihood ratio test (LRT) for the composite null H_0: $\beta\gamma = 0$ and the power of the MaxP test can also be calculated analytically. Its proof is given in the Supplemental Materials.

Result 2 *The joint significance test (MaxP test) has the following properties:*

(a) The MaxP test is the likelihood ratio test for the composite null of no mediation effect.

(b) The exact cumulative distribution function of the MaxP statistic in Case 1 is

$$F_{MaxP}(x) = x\left\{\Phi[\mu_\beta(n) - \Phi^{-1}(1 - \frac{x}{2})] + \Phi[-\mu_\beta(n) - \Phi^{-1}(1 - \frac{x}{2})]\right\};$$

and similarly for Case 2 by changing $\mu_\beta(n)$ to $\mu_\gamma(n)$; MaxP follows $Beta(2, 1)$ exactly in Case 3. The MaxP statistic follows a uniform distribution on [0,1] asymptotically in Case 1 and 2.

(c) The power of the MaxP test given the significance level α can be calculated analytically as

$$Power = \left[\Phi(\mu_\beta(n) - Z_{1-\frac{\alpha}{2}}) + \Phi(-\mu_\beta(n) - Z_{1-\frac{\alpha}{2}})\right]\left[\Phi(\mu_\gamma(n) - Z_{1-\frac{\alpha}{2}}) + \Phi(-\mu_\gamma(n) - Z_{1-\frac{\alpha}{2}})\right], \tag{8}$$

and the power of MaxP test is maximized when $|\mu_\beta(n)| = |\mu_\gamma(n)|$ for a fixed NIE strength.

3 The Divide-Aggregate Composite-null Test (DACT)

3.1 Estimation of the Proportions of the Three Null Cases

In view of the conservativeness of Sobel's test and the MaxP test, we propose in this section the Divide-Aggregate Composite-null Test (DACT) by leveraging information across a large number of

tests in genome-wide epigenetic studies. Suppose that an Oracle knows the true relative proportions of the three null cases, then such information can be incorporated to increase the power of the MaxP test. A single test for mediation effects using either Sobel's test or the MaxP test is challenged by the fact that one does not know which of the three null cases holds. Fortunately, we can obtain such information in modern multiple testing settings, such as in genome-wide epigenetic studies, where a large number of tests across the genome allow us to estimate the relative proportions of the three null cases. It is thus one of the few instances where high-dimensionality is not a curse but rather a blessing if used properly.

Suppose that there are a total of m DNA methylation CpG sites, where m is in the order of hundreds of thousands. For example, there are 484,613 CpG sites in the NAS data set to be described in detail later in Section 6. To identify putative CpG sites lying in the causal pathway from the exposure to the outcome of interest, we need to perform a total of m hypothesis tests to assess the strength of the evidence against the composite null hypothesis H_0: $\beta\gamma = 0$. There are m null hypotheses for the parameter β in the outcome regression model: H_{0j}^{β}: $\beta = 0$, and m null hypotheses for the parameter γ in the mediator regression model: H_{0j}^{γ}: $\gamma = 0$, where $1 \leq j \leq m$. We now define H_j^{β} (the same for H_j^{γ}) as a sequence of (possibly dependent) Bernoulli random variables, where $H_j^{\beta} = 0$ if H_{0j}^{β} is true and $H_j^{\beta} = 1$ if H_{0j}^{β} is false, $1 \leq j \leq m$, a framework proposed by Efron et al. (2001) and later adopted widely (Storey 2002; Genovese and Wasserman 2004).

As shown in Section 2 that β and γ are independent, we have that H_j^{β} is independent of H_j^{γ} for $1 \leq j \leq m$. For each DNA methylation CpG site, we fit the outcome and mediator regression models to obtain p-values $p_{\beta j}$ for testing β and the p-values $p_{\gamma j}$ for testing γ, where $1 \leq j \leq m$. Following Efron et al. (2001), assume that $P(H_j^{\beta} = 0) = \pi_0^{\beta}$ and $P(H_j^{\gamma} = 0) = \pi_0^{\gamma}$, where the parameters π_0^{β} is the proportion of CpG sites that are not associated with the outcome under the outcome models (1) or (3), and π_0^{γ} is the proportion of CpG sites that are not associated with the exposure in the mediator model (2). Since $H_j^{\beta} \perp\!\!\!\perp H_j^{\gamma}$, $1 \leq j \leq m$, then we have

$$\begin{aligned}
&\text{Case 1:} \quad Pr(H_j^{\beta} = 1, H_j^{\gamma} = 0) = (1 - \pi_0^{\beta})\pi_0^{\gamma}, \\
&\text{Case 2:} \quad Pr(H_j^{\beta} = 0, H_j^{\gamma} = 1) = \pi_0^{\beta}(1 - \pi_0^{\gamma}), \\
&\text{Case 3:} \quad Pr(H_j^{\beta} = 0, H_j^{\gamma} = 0) = \pi_0^{\beta}\pi_0^{\gamma}, \\
&\text{Case 4:} \quad Pr(H_j^{\beta} = 1, H_j^{\gamma} = 1) = (1 - \pi_0^{\beta})(1 - \pi_0^{\gamma}),
\end{aligned} \tag{9}$$

where Cases 1-3 together constitute the composite null hypothesis of null mediation effects, and Case 4 represents the alternative of non-null mediation effects.

Under the composite null H_0: $\beta\gamma = 0$, the normalized relative proportions of the three null cases w_1, w_2, w_3 are: $w_1 = \pi_0^\gamma(1-\pi_0^\beta)/c$, $w_2 = \pi_0^\beta(1-\pi_0^\gamma)/c$ and $w_3 = \pi_0^\beta\pi_0^\gamma/c$ respectively, where the normalizing constant $c = \pi_0^\gamma(1-\pi_0^\beta) + \pi_0^\beta(1-\pi_0^\gamma) + \pi_0^\beta\pi_0^\gamma$, and $w_1 + w_2 + w_3 = 1$. In typical epignome-wide association studies (EWAS), both π_0^β and π_0^γ are close to one as in our NAS data set in Section 6. To be more general, we do not impose such sparsity assumption on our method.

We used the method proposed by Jin and Cai (2007), which is referred to as the JC method hereafter, to estimate π_0^β and π_0^γ based on the z-scores for testing $\beta = 0$ and the z-scores for testing $\gamma = 0$ respectively. Suppose that we have m test statistics z-scores $Z_j \sim N(\mu_j, \tau_j^2)$, $1 \le j \le m$, where $\mu_j = \mu_0$ and $\tau_j^2 = \tau_0^2$ under the null. Here, we can set $\mu_0 = 0$ and $\tau_0 = 1$. Jin and Cai (2007) proposed to use the empirical characteristic function and Fourier analysis for estimating the proportion of true nulls. The empirical characteristic function is

$$\psi_m(t) = \frac{1}{m}\sum_{j=1}^{m} \exp(\mathbf{i}tZ_j), \tag{10}$$

where $\mathbf{i} = \sqrt{-1}$. For $r \in (0, 1/2)$, the proportion of true nulls π_0 can be consistently estimated as

$$\widehat{\pi}_0 = \sup_{\{0 \le t \le \sqrt{2r\log(m)}\}} \left\{ \int_{-1}^{1} (1-|\xi|)(\mathrm{Re}(\psi_m(t\xi; Z_1, \ldots, Z_m, m)\exp(-\mathbf{i}\mu_0 t\xi + \tau_0^2 t^2\xi^2/2)))d\xi \right\}, \tag{11}$$

where $\mathrm{Re}(x)$ denotes the real part of the complex number x. Jin and Cai (2007) showed that $\widehat{\pi}_0$ is uniformly consistent over a wide class of parameters for independent and dependent data under regularity conditions.

Kang (2020) also found in a recent simulation study that the JC method outperforms other competitors under practical dependence structures in genomic data. Here, we employ the JC method to estimate π_0^β and π_0^γ separately, and obtain uniformly consistent estimators $\widehat{\pi}_0^\beta$ and $\widehat{\pi}_0^\gamma$. Then w_1, w_2, w_3 are estimated by plugging in $\widehat{\pi}_0^\beta$ and $\widehat{\pi}_0^\gamma$ for the unknown parameters π_0^β and π_0^γ respectively. It is straightforward to show the resulting estimators $\widehat{w}_1, \widehat{w}_2, \widehat{w}_3$ are also consistent under the same regularity conditions of Jin and Cai (2007) using the continuous mapping theorem (van der Vaart 2000, pp. 7).

3.2 Construction of the Divide-Aggregate Composite-null Test (DACT)

We propose in this section the Divide-Aggregate Composite-null Test (DACT) for the composite null of no mediation effect $H_0 : \beta\gamma = 0$. We first consider how to perform mediation effect testing in each of the three null cases as defined in Section 2. In the null Case 1: $\beta \neq 0, \gamma = 0$, we only need to test whether $\gamma = 0$ using the p-value p_γ because $\beta \neq 0$. Similarly, in the null Case 2:

$\beta = 0, \gamma \neq 0$, we only need to test whether $\beta = 0$ using the p-value p_β because $\gamma \neq 0$. While in the null Case 3: $\beta = 0, \gamma = 0$, we need to test whether both β and γ are nonzero. We can reject the null Case 3 if $\max(p_\gamma, p_\beta) < \alpha$ at the significance level α. Intuitively, this requires that both p_β and p_γ are statistically significant. The p-value of the MaxP test can be computed as $(\text{MaxP})^2$ by noting that the MaxP test follows $Beta(2, 1)$ distribution in the null Case 3 as given in the Result 2. Following this logic, we propose the following case-specific p-values for testing mediation effects for the jth CpG site as

$$p = \begin{cases} p_{1j} = p_{\gamma j}, \text{ if Case 1;} \\ p_{2j} = p_{\beta j}, \text{ if Case 2;} \\ p_{3j} = (MaxP_j)^2, \text{ if Case 3.} \end{cases}$$

We now construct the DACT statstistic to test for the composite null of no mediation effect H_0: $\beta\gamma = 0$ by using a composite p-value as a test statistic, which is calculated as follows:

$$\text{DACT}_j = \hat{w}_1 p_{1j} + \hat{w}_2 p_{2j} + \hat{w}_3 p_{3j}. \tag{12}$$

If any of w_1, w_2 and w_3 is close to one, then the DACT statistic follows the uniform distribution on the interval $[0, 1]$ approximately. Based on our empirical observation from the NAS data analysis in Section 6, w_3 is very close to one. However, there are also scenarios when investigators want to conduct a more focused search within a smaller set of epigenetic markers from pre-screening studies, or based on prior knowledge (Cecil et al. 2014). In such circumstances, w_1 or w_2 may be a non-ignorable percentage, and the DACT statistic may depart from the uniform distribution on the interval $[0, 1]$. To make the DACT method applicable to those settings, we need to estimate the empirical null distribution of DACT.

We adopt Efron's empirical null inference framework (Efron 2004) to calibrate the p-values of the DACT statistics by accounting for possible correlations among the tests. Specifically, we transform the DACT statistic using the inverse normal cumulative distribution function (CDF)

$$Z_j^{DACT} = \Phi^{-1}(1 - DACT_j), 1 \leq j \leq m, \tag{13}$$

where $\Phi(\cdot)$ denotes the standard normal CDF. Those m test statistics fall into two categories: 1) null mediation effects; 2) non-null mediation effects. Therefore, the marginal probability density function of Z_j^{DACT} is

$$f(z) = \pi_0^{DACT} f_0(z) + (1 - \pi_0^{DACT}) f_1(z), \tag{14}$$

where π_0^{DACT} denotes the proportion of null mediation effects, $f_0(z)$ denotes the null distribution $N(\delta, \sigma^2)$, and $f_1(z)$ denotes the non-null distribution.

Our goal here is to estimate $f_0(z)$ by estimating δ and σ^2. The empirical characteristic function of Z_j^{DACT} is $\varphi_m(t) = \frac{1}{m}\sum_{j=1}^m \exp(\mathbf{i}tZ_j^{DACT})$. The expected characteristic function is $\varphi(t) = \frac{1}{m}\sum_{j=1}^m \exp(\mathbf{i}t\delta_j - \sigma_j^2 t^2/2)$, which can be decomposed as $\varphi(t) = \varphi_0(t) + \tilde{\varphi}(t)$, where $\varphi_0(t) = \pi_0^{DACT}\exp(\mathbf{i}\delta t\xi - \sigma^2 t^2)$ and $\tilde{\varphi}(t) = (1-\pi_0^{DACT})\mathrm{Ave}_{\{j:(\delta_j,\sigma_j)\neq(\delta,\sigma)\}}\left\{\exp(\mathbf{i}\delta_j t - \sigma_j^2 t^2/2)\right\}$.

Jin and Cai (2007) showed that for all $t \neq 0$,

$$\begin{aligned}\delta &= \delta(\varphi_0;t) = \frac{\mathrm{Re}(\varphi_0(t))\cdot\mathrm{Im}(\varphi_0'(t)) - \mathrm{Re}(\varphi_0'(t))\cdot\mathrm{Im}(\varphi_0(t))}{|\varphi_0(t)|^2},\\ \sigma^2 &= \sigma^2(\varphi_0;t) = -\frac{d|\varphi_0(t)|/dt}{t\varphi_0(t)},\end{aligned}$$

where $\mathrm{Re}(x)$, $\mathrm{Im}(x)$ and $|x|$ denote the real part, the imaginary part and the modulus of the complex number x. For an appropriately chosen large t, $\varphi_m(t) \approx \varphi(t) \approx \varphi_0(t)$, so that the contribution of non-null mediation effects to the empirical characteristic function is negligible. In practice, t is chosen as $\hat{t}(r) = \inf\{t : |\varphi_m(t)| = m^{-r}, 0 \leq t \leq \log(m)\}$, for a given $r \in (0, 1/2)$. One then estimates δ and σ^2 using

$$\widehat{\delta} = \delta(\varphi_m;\hat{t}(r)) \quad \text{and} \quad \widehat{\sigma}^2 = \sigma^2(\varphi_m;\hat{t}(r)), \tag{15}$$

with $r = 0.1$ as recommended by Jin and Cai (2007). The two estimators $\widehat{\delta}$ and $\widehat{\sigma}^2$ have been shown to be uniformly consistent for independent and dependent data under some regularity conditions (Jin and Cai 2007), and hence the empirical null probability density function estimator $\widehat{f}_0$ and the corresponding CDF estimator $\widehat{F}_0$ are both consistent. We then calibrate the p-value of Z_j^{DACT} by

$$p_j = 1 - \Phi\left(\frac{Z_j^{DACT} - \widehat{\delta}}{\widehat{\sigma}}\right). \tag{16}$$

Efron's empirical null framework is really a statement about the nature or the choice of the null distribution, and does not depend on the inference method to be used later for thresholding the test statistics (Schwartzman et al. 2009). If the empirical null is $N(\delta, \sigma^2)$, then any method for controlling family-wise error rate (FWER) can be applied to the normalized z-scores $Z^* = (Z-\delta)/\sigma$ or equivalently the calibrated p-values. The FWER is controlled asymptotically as long as the empirical null distribution can be consistently estimated. The proof is trivial and thus omitted. The same argument also applies to the local and tail area false discovery rate (FDR) control (Efron et al. 2001; Efron 2004, 2010). The local FDR is defined as $fdr = \pi_0^{DACT} f_0(z)/f(z)$ and the tail area FDR is $Fdr = \pi_0^{DACT} F_0(z)/F(z)$, where $F_0(z)$ and $F(z)$ are the corresponding CDFs of $f_0(z)$ and $f(z)$ respectively. The parameter π_0^{DACT} can be consistently estimated using the generic

formula (11) by replacing μ_0, τ_0, Z_j by $\widehat{\delta}, \widehat{\sigma}, Z_j^{DACT}$ respectively. The marginal probability density function $f(z)$ can be consistently estimated using the kernel density estimator $\widehat{f}$ (Wasserman 2006, pp. 133), and the marginal CDF $F(z)$ can be consistently estimated using the empirical CDF $\widehat{F}(z)$ according to the classical Glivenko–Cantelli theorem (van der Vaart 2000, pp. 266) We show in the Supplemental Materials that the (local) FDR can be controlled asymptotically. We summarize our findings about DACT in Result 3.

Result 3 *The proposed DACT has the following properties:*

(a) In Case 1 or Case 2, the DACT is asymptotically equivalent to both the Sobel's test and the MaxP test.

(b) In Case 3, the DACT has the correct size, while both Sobel's test and the MaxP test are conservative for any sample size.

(c) Under regularity conditions of Jin and Cai (2007), $\widehat{\pi}_0^{DACT}, \widehat{f}_0, \widehat{f}$ *are consistent estimators of* e_0, f_0, f *respectively. The local FDR for the* j*th composite null test* H_{0j} *is estimated as* $\widehat{fdr}(Z_j^{DACT}) = \widehat{\pi}_0^{DACT}\widehat{f}_0(Z_j^{DACT})/\widehat{f}(Z_j^{DACT})$*. Then the following procedure controls local FDR asymptotically at a pre-specified level* $q \in [0, 1]$*,*

$$\textit{reject } H_{0j} \textit{ if } \widehat{fdr}(Z_j^{DACT}) \leq q. \tag{17}$$

The same result holds for the tail-area FDR control by replacing $\widehat{f}_0, \widehat{f}$ *by* $\widehat{F}_0, \widehat{F}$ *respectively.*

Remark: The use of the empirical null distribution to correct bias and inflation of the observed p-values in EWAS has been proven useful and effective (van Iterson et al. 2017). If the genomic inflation factor λ of DACT is close to one, then this correction makes little change. However, if none of the three null cases is close to one, for example, when $w_1 = w_2 = w_3 = 1/3$ as shown in Figure 4, then the corrected DACT (calibrated p-value for DACT) using equation (16) performs much better as demonstrated in our simulation studies in Section 5.

4 Comparison of the Three Tests

Our proposed data-adaptive DACT approach leverages information contained in the whole epigenome, and thus has improved power for testing mediation effects. Figure 3 shows that the rejection region of the MaxP test is a proper subset of the rejection region of our DACT method, while the rejection region of Sobel's test is a proper subset of the rejection region of the MaxP test. In other words, our DACT dominates the MaxP test and the MaxP test dominates Sobel's test.

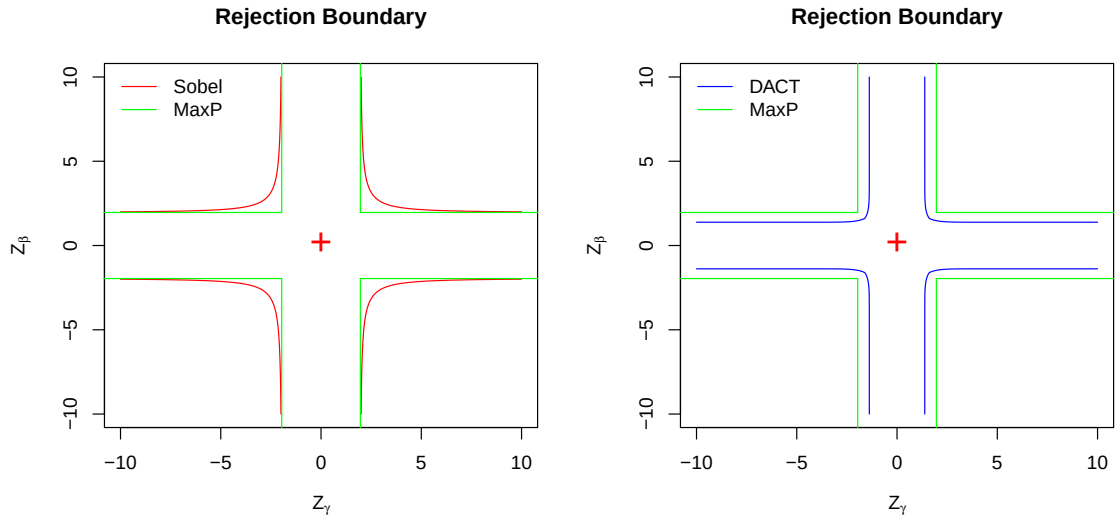

Figure 3: The rejection boundaries of the Sobel's test, the MaxP test and the DACT are plotted at significance level 0.05 on the z-score scale. For DACT, we set $w_1 = w_2 = 0.2$ and $w_3 = 0.6$.

Formally, let's compare the Sobel's test and the MaxP test in finite sample settings. We already know that $|T_{Sobel}| < \min(|Z_\beta|, |Z_\gamma|)$, therefore we have

$$p_{Sobel} = 2(1 - \Phi(|T_{Sobel}|)) > \max(p_\beta, p_\gamma) = MaxP. \tag{18}$$

This result says that the MaxP test is always more significant than Sobel's test at the significance level α. In other words, if the Sobel's test detects a mediation effect, then the MaxP test will do as well, but not vice versa. Therefore, the MaxP test is uniformly more powerful than the Sobel's test for any given significance level. In this regard, the Sobel's test is inadmissible. However, the Sobel's test and the MaxP test are asymptotically equivalent in Case 1 and 2. In Case 1, because T_{Sobel} is asymptotically equivalent to Z_γ and MaxP is asymptotically equivalent to p_γ, therefore the inference using T_{Sobel} is asymptotically equivalent to MaxP. The same conclusion holds in Case 2 as well. In Case 3, the inferences using T_{Sobel} and MaxP are asymptotically different. The asymptotic p-value of T_{Sobel} is calculated using the normal distribution $N(0, 1/4)$, while the asymptotic p-value of the MaxP test is calculated using the Beta distribution $Beta(2, 1)$.

One can also show that the MaxP test based on $MaxP = \max(p_\beta, p_\gamma)$ can be equivalently defined using $MinZ^2 = \min(Z_\beta^2, Z_\gamma^2)$. Both give the same inference. This provides a more clearer relationship of Sobel's test and the MaxP test on the same scale directly using Z_β and Z_γ. Specifically, since $T_{Sobel}^2 = (Z_\beta^{-2} + Z_\gamma^{-2})^{-1}$, both T_{Sobel}^2 and $MinZ^2$ asymptotically follow χ_1^2 in Cases 1

and 2. However, in Case 3, T^2_{Sobel} asymptotically follows $\chi^2_1/4$, while $MinZ^2$ asymptotically follows the distribution of the first order statistic of two independent random variables that follow the χ^2_1 distribution, i.e., the distribution of $\min(S^2_1, S^2_2)$, where S^2_1 and S^2_2 are independent random variables that follow the χ^2_1 distribution. In Case 3, it is straightforward to show that the cumulative distribution function of $MinZ^2$ is

$$Pr(MinZ^2 \leq x) = 1 - [1 - F_{\chi^2_1}(x)]^2,$$

where $F_{\chi^2_1}(x)$ denotes the cumulative distribution function of a central chi-squared random variable with one degree of freedom. Therefore, in Case 3, the Wald-type Sobel's test and the likelihood ratio test equivalent MaxP test have different distributions in both finite and large sample settings. In Section 2, we have shown that the actual sizes of the Sobel's test and the MaxP test are smaller than the pre-specified nominal type I error rate α. Those two tests are thus underpowered because they do not fully spend the allowed amount of type I error α.

5 Simulation Studies

5.1 Type I Error Rates

In this section, we conduct extensive simulation studies to evaluate the type I error rate of the DACT method under the composite null. We include the Sobel's test, the MaxP test and the MT-Comp test (Huang 2019) for comparison. First, the exposure variable A was simulated from a Bernoulli distribution with success probability equal to 0.5. We simulated two continuous covariates X_1 and X_2 from $N(10, 1)$ and $N(5, 1)$ respectively, then the mediator M and the outcome Y were simulated as follows

$$\begin{aligned} Y &= A + \beta M + 0.1X_1 + 0.2X_2 + \epsilon_Y,\ \epsilon_Y \sim N(0, 2), \\ M &= \gamma A + 0.2X_1 + 0.3X_2 + \epsilon_M,\ \epsilon_M \sim N(0, 1), \end{aligned}$$

where (β, γ) take the following three value pairs: $(0.2, 0), (0, 0.2)$ and $(0, 0)$, corresponding to the three cases under the composite null hypothesis (4). The significance levels are: 0.05 and 0.01. We considered three sample sizes: $N = 500, 1000, 2000$. In total, we simulated $100,000$ such datasets for each setting and the type I error rates were estimated as the proportions of rejections among those $100,000$ replicates.

Table 1: *Empirical type I error rates of the four tests: Sobel's test, MaxP test, MT-Comp and our DACT method under three nulls where* (β, γ) *are:* $(0.2, 0), (0, 0.2), (0, 0)$*. The sample sizes are: 500, 1000 and 2000. The significance level* α *are: 0.05, 0.01.*

	β	γ	Sobel		MaxP		MT-Comp		DACT	
Level α			0.05	0.01	0.05	0.01	0.05	0.01	0.05	0.01
N=500	0.2	0	0.005	0.000	0.030	0.004	0.302	0.128	0.050	0.009
	0	0.2	0.005	0.000	0.031	0.004	0.306	0.131	0.051	0.010
	0	0	0.000	0.000	0.003	0.000	0.050	0.010	0.050	0.010
N=1000	0.2	0	0.014	0.000	0.045	0.007	0.458	0.245	0.051	0.010
	0	0.2	0.014	0.000	0.045	0.007	0.455	0.244	0.051	0.010
	0	0	0.000	0.000	0.003	0.000	0.050	0.011	0.050	0.010
N=2000	0.2	0	0.027	0.002	0.049	0.010	0.607	0.405	0.050	0.010
	0	0.2	0.028	0.002	0.050	0.010	0.608	0.402	0.050	0.010
	0	0	0.000	0.000	0.002	0.000	0.050	0.010	0.051	0.010

As shown in Table 1, the type I error rates of the Sobel's test are smaller than the nominal significance levels in all three cases, especially in Case 3. The type I error rates of the MaxP test get closer to the nominal significance levels in Case 1 and Case 2 as the sample size increases. In Case 3, increasing sample size does not change the empirical size of the MaxP test. The type I error rates of the MT-Comp method are inflated in Case 1 and Case 2, and this inflation gets worse when the sample size increases. Huang (2019) also found that the MT-Comp can control type I error rate when the sample size is 500 or smaller. The MT-Comp method has correct size under Case 3 and thus works when the mediation effect signals are sparse with small sample sizes. The type I error rates of the proposed DACT are very close to the nominal levels in all three null cases.

We now consider multiple testing settings where a large number of candidate mediators are tested. Assume the total number of candidate mediators is $300,000$. We vary the relative proportions of w_1, w_2 and $w_3 = 1 - w_1 - w_2$ to assess the performance of our method. We only need to specify (w_1, w_2), and hence consider the following three settings: 1) $w_1 = 0.33, w_2 = 0.33$ which represents the worst-case scenario; 2) $w_1 = 0.05, w_2 = 0.05$; and 3) $w_1 = 0.01, w_2 = 0.01$ which represent average-case scenarios often encountered in genome-wide epigenetic studies. Even in setting 3), there are 3000 mediators associated with the exposure only, and another set of 3000 mediators associated with the outcome only. In a typical epigenome-wide association study, the number of association signals is much smaller. We aim to demonstrate that our methods can perform robustly even in those unfavorable settings. We simulate $300,000$ Z-test statistics $(Z_{\beta j}, Z_{\gamma j})$ where $j = 1, \ldots, 300000$. In Case 1, simulate $Z_{\beta j}$ from $N(\mu_\beta, 1)$ where μ_β drawn from $N(2, 1)$ and simulate $Z_{\gamma j}$ from $N(0, 1)$. In Case 2, simulate $Z_{\beta j}$ from $N(0, 1)$ and $Z_{\gamma j}$ from $N(\mu_\gamma, 1)$ where μ_γ

drawn from $N(2,1)$. In Case 3, simulate $Z_{\beta j}$ from $N(0,1)$ and $Z_{\gamma j}$ from $N(0,1)$.

The QQ (quantile-quantile) plots for the p-values from uncorrected and corrected DACT using the estimated empirical null distribution are summarized in Figure 4. In setting 1), the uncorrected DACT is conservative while the corrected DACT works well. In setting 2), there is a noticeable difference between the uncorrected and corrected DACT methods. In setting 3), there is no noticeable difference between the corrected and uncorrected DACT method because the DACT statistic approximately follow uniform on $[0,1]$, and thus the correction is usually not needed in such settings.

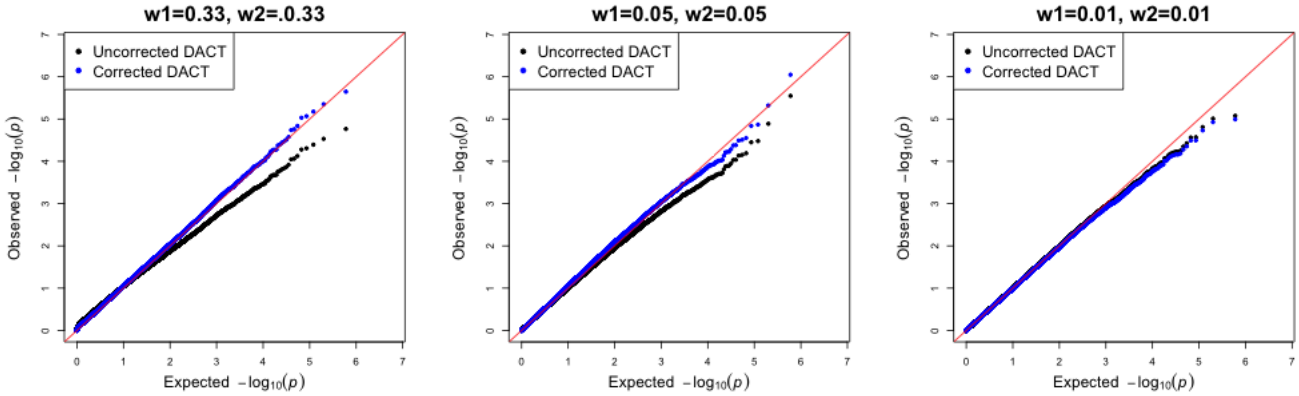

Figure 4: The QQ plots of the p-values for the uncorrected and corrected DACT method in three simulated multiple testing settings. The left-most figure presents the QQ plot of uncorrected and corrected DACT for the worst-case scenario where $w_1 = w_2 = 0.33$; the middle and right-most figures present the QQ plots where $w_1 = w_2 = 0.05$ and $w_1 = w_2 = 0.01$ respectively.

5.2 Power Comparison

Since the MT-Comp method has inflated type I error rates in Case 1 and Case 2, we do not include it for power comparison. The original Sobel and MaxP tests have deflated type I error rates and thus under-powered. At the significance level α, the power of Sobel's test is estimated as the proportion of tests with $p_{Sobel} < \alpha$, where p_{Sobel} is calculated using the standard normal approximation; the power of the MaxP test is estimated as the proportions of tests with $MaxP < \alpha$. These two tests will serve as benchmarks for power comparison with the proposed DACT method.

Table 2: *Power comparisons of the Sobel's test, MaxP test and the DACT test using simulation studies. The sample sizes considered are 800, 1000, 1200. The* $A - M$ *and* $M - Y$ *association effects* (γ, β) *are set to be* $(0.133, 0.3)$ *,* $(0.2, 0.2)$ *and* $(-0.3, 0.133)$ *where* $|\gamma\beta| = 0.04$ *in those three settings.*

(γ, β)	(0.133, 0.3)			(0.2, 0.2)			(-0.3,0.133)		
N	800	1000	1200	800	1000	1200	800	1000	1200
Sobel	0.34	0.45	0.55	0.42	0.60	0.74	0.34	0.46	0.56
MaxP	0.46	0.55	0.62	0.65	0.78	0.87	0.45	0.55	0.63
DACT	0.47	0.55	0.63	0.76	0.87	0.93	0.46	0.56	0.64

We used the same simulation setup as that described in the first paragraph of Section 5.1 except that we simulated data under the alternative hypothesis. Specifically, we set (β, γ) to be the following values: $(0.133, 0.3), (0.2, 0.2), (-0.3, 0.133)$ respectively, where the mediation effect size was set to be 0.04. The sample size N was set to be 500, 1000, or 2000. We generated a total of $10,000$ simulated datasets for each setting and the power was estimated as the proportion of rejections at the significance level 0.05. The results are summarized in Table 2. As expected, the MaxP test is more powerful than Sobel's test and the DACT test is more powerful than the MaxP test.

We found that the power advantage of the DACT test over the MaxP test gets smaller with increasing differences in the magnitudes between β and γ. To investigate this matter, we further performed the following additional simulation studies. First, we set the mediation effect size $\beta\gamma$ to be 0.04. Second, we divided the interval $[0.04, 0.5]$ equally into 400 subintervals specified by 401 grid points γ_j, and set $\beta_j = 0.04/\gamma_j$ where $j = 1, \cdots, 401$. Under each alternative (β_j, γ_j), we performed one million simulations to estimate the powers of Sobel's test, MaxP and DACT. We plotted the power estimates for all 401 grid points for three tests: Sobel, MaxP and DACT. Figure 5 shows that the powers of the three tests are not monotone increasing functions of the mediation effect size $\beta\gamma$, but actually depend on the relative effect sizes of β and γ. The powers of these three tests are all maximized when $|\gamma/\beta| = 1$ and decrease quickly as $|\gamma/\beta|$ deviates away from one. In other words, the powers of Sobel, MaxP and DACT are dictated by the smaller association signal of the $A - M$ and $M - Y$ associations. Those simulation results are in line with our theoretical findings in Results 1 and 2.

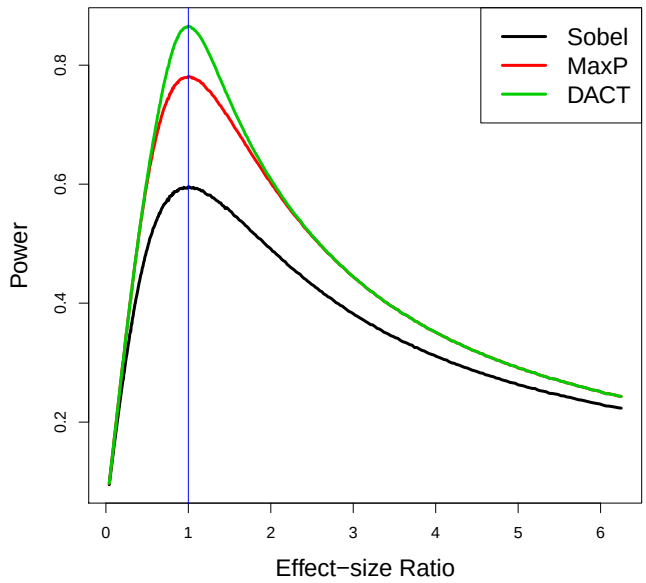

Figure 5: *Power Comparison of the three tests: Sobel, MaxP and DACT using simulation studies. The same mediation effect size is fixed at* 0.04 *with different* β *and* γ *value combinations. The horizontal axis represents the effect size ratio* $|\gamma/\beta|$.

To mimic the real methylation data structure, we performed an additional simulation study by simulating outcomes using the observed DNA methylation M-values of 24,264 CpG sites on chromosome 5 from the NAS data set (See Section 6 for detailed background information), because we found strong mediation effect signals on this chromosome. In this numerical experiment, without loss of generality, we did not include covariates for simplicity. We set the sample size to be 603, the same as in the NAS data. We generated an exposure variable A from a Bernoulli distribution with probability 0.5. We then shifted the mean value of a randomly selected set of 2000 CpG sites among the exposed group ($A = 1$), and simulated the mean shift effect sizes from a uniform distribution on $[-0.6, -0.2]$ mimicking the effect sizes of smoking on the methylation in the NAS data. We generated $Y = \beta_0 + \beta_A A + \sum_{j=1}^{500} \beta_j M_j + \epsilon,\ \epsilon \sim N(0, 1.2)$, where 500 CpG sites were selected based on the most significant associations with the lung function from the analysis of the real NAS data, and the true coefficients $\beta_j, j = 1, \ldots, 500$ were set at the estimated values from the NAS data. In this set-up, the numbers of CpG sites in the four null and alternative cases in (9) were: 500, 2000, 21723 and 41 respectively.

We repeated this numerical experiment 1,000 times and estimated the FDR, and the mean true positive rate (TPP, or average power) (Dudoit and van der Laan 2007), which is defined as the

proportion of mediation signals detected using the FDR threshold at 0.05. We included the MaxP test for comparison as the Sobel's test has been shown to be less powerful than the MaxP test. For the MaxP test and the DACT method, we found that the estimated FDR was 0.042 (DACT) and 0 (MaxP), and the average power using FDR threshold is 0.86 (DACT) and 0.28 (MaxP). Therefore, the MaxP test was overly conservative, and the DACT method had an improved power while controlling for FDR at the nominal level in multiple testing settings.

6 The Normative Aging Genome-Wide Epigenetic Study

Cigarette smoking is an important risk factor for lung diseases (Anthonisen et al. 2002). Smoking behavior has been found to be associated with DNA methylation levels (Breitling et al. 2011; Li et al. 2018), and DNA methylation levels have also been found to be associated with lung functions (Lepeule et al. 2012). It is thus of scientific interest to identify DNA methylation CpG sites that may mediate the effects of smoking on lung functions. Previous research has found two CpG sites (cg05575921, cg24859433) as mediators lying in the causal pathway from smoking to lung functions using underpowered testing procedures (Zhang et al. 2016; Barfield et al. 2017). In this section, we demonstrate that the proposed DACT method has improved power to detect more DNA methylation CpG sites that might mediate the effect of smoking on lung functions.

The Normative Aging Study (NAS) is a prospective cohort study established in Eastern Massachusetts in 1963 by the U.S. Department of Veteran Affairs (Bell et al. 1972). The men were free of known chronic medical conditions at enrollment, and returned for on-sites, follow-up visits every 3-5 years. During these visits, detailed physical examinations were performed, bio-specimens including blood were obtained, and questionnaire data pertaining to diet, smoking status, and additional lifestyle factors that may impact health were collected. The DNA methylation was measured using the Illumina Infinium HumanMethylation450 Beadchip on blood samples collected after an overnight fast (Bibikova et al. 2011). After quality control, methylation Beta-values ranging from 0 (no methylation) to 1 (full methylation) was calculated for each CpG site (Teschendorff et al. 2012). We then use the logit (base 2) function to transform the Beta-values into M-values for statistical analysis because the M-value scale is more statistically valid for regression models as it is approximately homoscedastic (Du et al. 2010). The batch effects were adjusted by the ComBat algorithm (Johnson et al. 2007). In total, we had DNA methylations measured for 484,613 CpG sites on 603 men.

The binary exposure was smoking status (current or former smokers versus never smokers), and

the outcome was the forced expiratory flow at 25%-75% of the Forced Expiratory Vital capacity ($\text{FEF}_{25-75\%}$). We transformed $\text{FEF}_{25-75\%}$ using squared root to achieve better normality. We adjusted for age, height, weight, education history, medication history, blood cell type abundances (Houseman et al. 2012), and five principal components (previously calculated to represent 95% of DNA processing batch effects), all based on our prior work studying DNA methylation in this cohort. We then fit the outcome and mediator linear regression models and obtain p-values for γ (smoking - methylation) and β (methylation - lung function) for each of the 484,613 CpG sites (the QQ plots are given in Figure S1 in the Supplementary Materials). The proportions of nulls for the parameter γ and β were estimated as $0.996, 0.9867$ respectively using the JC method (Jin and Cai 2007). Using equation (9), the proportions of the four cases were estimated as $(0.01423, 0.00416, 0.98155, 0.0006)$. Therefore, the mediation effect signals were very sparse in the NAS data set.

Under the composite null, the relative proportions of the three null cases (after normalization) were estimated as $\hat{w}_1 = 0.014, \hat{w}_2 = 0.004, \hat{w}_3 = 0.982$. We then computed the DACT, performed inverse normal CDF transformation to obtain z-scores. A histogram of the transformed DACT (z-scores) indicates strong normality as shown in Figure 6 (the left sub-figure). The mean δ and standard deviation σ of the null distribution $N(\delta, \sigma^2)$ were estimated as $(\hat{\delta}, \hat{\sigma}) = (-0.053, 0.998)$ using equation (15) in Section 3.2.

The QQ plot in Figure 6 (the middle sub-figure) showed that both Sobel's test and the MaxP test produced seriously deflated p-values and hence were under-powered to detect CpG sites with mediation effects. In contrast, the proposed DACT method performed very well, and its genomic inflation factor was estimated as $\lambda = 1.07$. The volcano plot in Figure 6 (the right sub-figure) showed that those more significant CpG sites also tended to have larger mediation effect sizes, and thus the statistical significance was mainly driven by the large effect sizes rather than small standard errors.

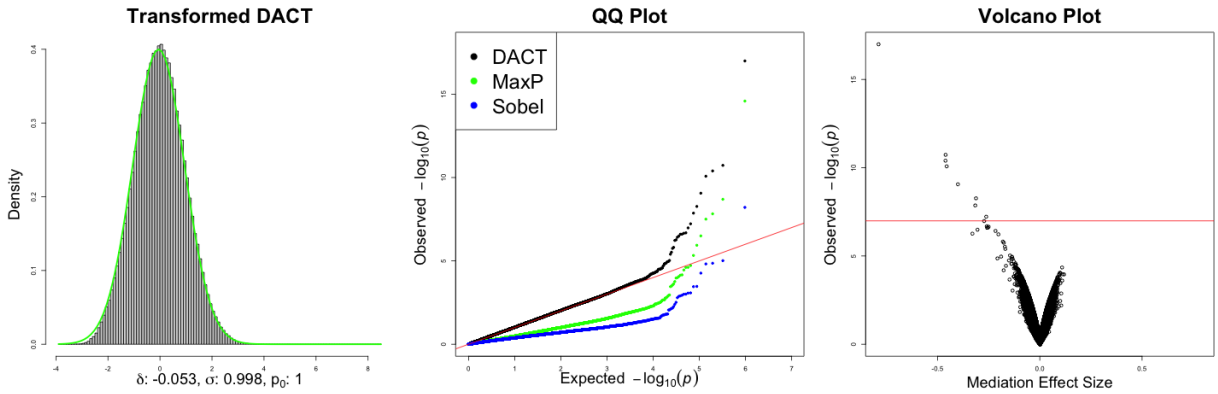

Figure 6: *The left sub-figure is a histogram of the z-scores transformed from the DACT statistics based on the inverse normal cumulative distribution function. The green solid line is the estimated empirical null density function with mean -0.053 and standard deviation 0.998 using equation (15). The middle one is the QQ plot of the Sobel's test, the MaxP test and the corrected DACT method. The right one is the volcano plot for the corrected DACT method, where the horizontal axis represents the mediation effect sizes and the vertical axis represents the corrected p-values of the DACT method on the* $-\log_{10}$ *scale.*

Using the tail FDR threshold at 0.05, we found 19 mediation effect signals summarized in Table S1 in the Supplementary Materials. To save space, we present the most significant top eight CpG sites in Table 3. Those CpG sites are also significant using the more stringent Bonferroni corrected threshold ($0.05/484613 = 1.03 \times 10^{-7}$). A Manhattan plot is also provided in Figure S2 in the Supplementary Materials. In Table 3, the Sobel's test only detected CpG site cg05575921, and the MaxP test detected four CpG sites: cg05575921, cg03636183, cg06126421 and cg21566642. The proposed DACT method further detected additional CpG sites that were missed by the Sobel's test and the MaxP test.

Table 3: *Top hits from causal mediation analysis of the Normative Aging Genome-wide Epigenetic Study. The exposure is smoking status, and the outcome is lung function measure* $FEF_{25-75\%}$*. CHR stands for chromosome number. NIE stands for Natural Indirect Effect (mediation effect). The* p_{NIE} *column is computed using the DACT method after correction.*

CpG Name	CHR	γ	SE_γ	p_γ	β	SE_β	p_β	NIE	p_{Sobel}	p_{NIE}
cg05575921	5	-0.53	0.06	5.93E-16	1.50	0.18	2.60E-15	-0.79	6.19E-09	1.02E-17
cg03636183	19	-0.27	0.04	2.02E-09	1.72	0.27	2.49E-10	-0.46	9.70E-06	1.86E-11
cg06126421	6	-0.37	0.06	6.37E-11	1.23	0.21	1.50E-08	-0.46	1.38E-05	4.01E-11
cg21566642	2	-0.32	0.05	2.59E-11	1.42	0.25	3.14E-08	-0.46	1.52E-05	8.35E-11
cg05951221	2	-0.27	0.04	2.20E-10	1.46	0.28	3.17E-07	-0.40	5.43E-05	8.71E-10
cg14753356	6	-0.15	0.03	2.26E-07	2.02	0.41	1.16E-06	-0.31	3.40E-04	5.42E-09
cg23771366	11	-0.20	0.04	2.55E-08	1.56	0.34	4.72E-06	-0.32	3.51E-04	1.37E-08
cg01940273	2	-0.18	0.04	4.64E-07	1.46	0.34	1.90E-05	-0.26	9.97E-04	5.99E-08

The top CpG site cg05575921 is located in the aryl-hydrocarbon receptor repressor (AHRR) gene on chromosome 5 and has been consistently found to be demethylated among smokers compared to non-smokers (Joubert et al. 2012; Philibert et al. 2012, 2013; Reynolds et al. 2015). It has also been found to be associated with increased lung cancer risk (Fasanelli et al. 2015). Previous mediation analysis using the under-powered MaxP test can also detect this CpG site cg05575921 as an mediator in the pathway from smoking to lung functions (Zhang et al. 2016; Barfield et al. 2017), simply because the p-values for the smoking-methylation and methylation-lung functions associations were both highly significant.

The CpG site cg03636183 in F2RL3 was also found to be a biomarker of smoking exposure (Zhang et al. 2014) and be related to mortality among patients with stable coronary heart disease (Breitling et al. 2012) and increased lung cancer risk (Fasanelli et al. 2015). It has been found that the CpG site cg06126421 in the intergenic region at 6p21.33 to be hypomethylated among smokers compared to non-smokers (Shenker et al. 2012; Elliott et al. 2014). The CpG site cg06126421 was found to be associated with all-cause, cardiovascular, and cancer mortality, for participants with methylation levels in the lowest quartile of this CpG site (Zhang et al. 2016). The CpG sites cg21566642 and cg05951221 located on the same CpG island of chromosome 2 were found to be associated with increased lung cancer risk (Fasanelli et al. 2015). Our analysis suggests that those significant CpG sites might play important biological roles in mediating the effect of smoking on lung functions.

To check for any possible violation of the no unmeasured confounding assumption, we further performed a comprehensive sensitivity analysis to assess the robustness of our mediation analysis results to any unmeasured confounding variables. The idea is that the residual correlation ρ between the two error terms in the mediator and outcome regressions are correlated if the unmeasured confounding assumption is violated and vice versa (Imai et al. 2010). Therefore, the residual correlation ρ can be used to measure the magnitude of confounding bias, where $\rho = 0$ implies no confounding bias. We can hypothetically vary ρ to observe the change to the mediation effect estimates. When $|\rho|$ deviates from zero to some extent, the observed mediation effects could be explained away by the confounding bias. We varied the value of ρ and computed the corresponding value of NIE using the R package *mediation* (Tingley et al. 2013). We found that to explain away the mediation effects of CpG sites cg05575921 and cg03636183 in the causal pathway from smoking to lung function, the confounding bias measured by ρ needs to be at least 0.3, and to explain away the mediation effects of the other CpG sites provided in Table 3, ρ needs to be at least 0.2. Such large

confounding bias is absent in our data analysis, as we found that the residual correlation ρ for all the eight CpG sites are very close to zero with absolute value smaller than 10^{-17}, showing that the confound bias is negligible. Our sensitivity analysis results show that we have adjusted sufficient covariates in the mediation analysis for all the CpG sites in Table 3. Therefore, our mediation analysis results are robust to unmeasured confounding. More detailed sensitivity analysis results are provided in the Supplementary Materials.

7 Discussion

In this paper, we developed a valid and powerful testing procedure for detecting CpG sites that might mediate the effect of an exposure on an outcome in genome-wide epigenetic studies. Despite that the Wald-type Sobel's test and the likelihood ratio test equivalent MaxP test were empirically found to have low power for decades, however, no successful remedy has been proposed to resolve the conservativeness of the two tests. A lack of method development for this problem is incompatible with the increasing need of powerful testing procedures for detecting mediation effects in large-scale epigenetic studies. Testing a large number of composite nulls leverages the two sides of the same coin. On one side, multiple testing correction is a curse and makes it more challenging for inference of mediation effects than the single mediation effect testing problems. But on the other side, multiple testing for mediation effects is a blessing because it enables us to estimate the relative proportions of the three null cases that can be leveraged to improve power.

Understanding the reasons why Sobel's test and the MaxP test are conservative paves the way for developing a more powerful test. We found that the null Case 3 is the singular point in the null parameter space, under which the standard asymptotic arguments all fail. We show that the MaxP test is essentially the likelihood ratio test for the composite null of no mediation effect, but it does not follow the traditional chi-squared distribution with one degree of freedom (on the Z^2 scale) but rather follows Beta distribution $Beta(2, 1)$ in the null Case 3. The Wald-type Sobel's test does not follow the standard normal distribution in the null Case 3 either, instead it follows the normal distribution with mean zero and variance equal to one quarter which can be shown by the not so well-known "super Cauchy phenomenon" (Pillai and Meng 2016). Those important discoveries provide rigorous explanations on why the widely used Sobel's test and the MaxP test are underpowered for inferring the presence of mediation effects in both single test and multiple testing scenarios, more importantly, inspire us to develop the DACT method.

Our contributions are multi-folds. First, we divide the null parameter space into three disjoint

parts and find that the null Case 3 is the culprit of the poor performances of Sobel's test and the MaxP test. Such a decomposition also inspires us to obtain correct case-specific p-values. Second, we leverage the genome-wide data to consistently estimate the relative proportions of the three null cases and then construct the DACT, turning the curse of multiple testing into a blessing. Third, large-scale testing also permits the use of the empirical null distribution for inference. This approach is especially useful when exposure-mediator or/and mediator-outcome association signals are non-sparse. Fourth, the DACT procedure is computationally fast and is scalable for large-scale inference of mediation effects. We also developed an user-friendly R package *DACT* for public use.

Our NAS data analysis findings are of scientific interests. Detection of DNA methylation CpG sites that may mediate the effect of smoking behavior on lung function can help us understand the underlying causal mechanism and pathway of the observed association between smoking and lung function. These identified CpG sites can also be used as intervention targets to reduce the harmful effects of smoking on lung function. Previously, only two CpG sites with strong signals have been found as putative mediators in the causal pathway from smoking to lung function (Barfield et al. 2017). A lack of powerful tests hindered researchers to discover more potential mediators. We applied the newly developed DACT procedure to the Normative Aging Study and identified additional DNA methylation CpG sites that were missed by previous analysis. Our comprehensive sensitivity analysis suggests that the mediation results are robust to unmeasured confounding factors.

The proposed DACT procedure is developed for genome-wide epigenetic studies where we can estimate the relative proportions of the three cases under the composite null hypothesis. Notice that accurate estimation of these proportions is crucial for performing the DACT test, especially when the p-values across the CpG sites are correlated. The JC method for estimating these proportions was found to be accurate and consistent in both sparse and non-sparse settings even for dependent data, and has been adopted in our DACT procedure. It is of future research interest to extend the DACT method to the setting in which there are a large number of exposures, e.g, genetic variants in Genome-Wide Association Studies, as well as univariate or multivariate mediators. When the binary outcome is not rare, the NIE is no longer equal to $\beta\gamma$ even approximately (Gaynor et al. 2019). Testing NIE in those settings is challenging and is of future research direction. Our DACT procedure is not applicable for a single mediation test if the relative proportions of the three null cases cannot be empirically estimated. It is hence of future research interest to develop powerful mediation tests in such settings.

References

Anthonisen, N. R., Connett, J. E., and Murray, R. P. (2002). Smoking and lung function of lung health study participants after 11 years. *American Journal of Respiratory and Critical Care Medicine* **166,** 675–679.

Barfield, R., Shen, J., Just, A. C., Vokonas, P. S., Schwartz, J., Baccarelli, A. A., VanderWeele, T. J., and Lin, X. (2017). Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic Epidemiology* **41,** 824–833.

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

Bell, B., Rose, C. L., and Damon, A. (1972). The normative aging study: an interdisciplinary and longitudinal study of health and aging. *Aging and Human Development* **3,** 5–17.

Bibikova, M., Barnes, B., Tsan, C., Ho, V., Klotzle, B., Le, J. M., Delano, D., Zhang, L., Schroth, G. P., Gunderson, K. L., et al. (2011). High density dna methylation array with single cpg site resolution. *Genomics* **98,** 288–295.

Bind, M.-A., Lepeule, J., Zanobetti, A., Gasparrini, A., Baccarelli, A. A., Coull, B. A., Tarantini, L., Vokonas, P. S., Koutrakis, P., and Schwartz, J. (2014). Air pollution and gene-specific methylation in the normative aging study: association, effect modification, and mediation analysis. *Epigenetics* **9,** 448–458.

Breitling, L. P., Salzmann, K., Rothenbacher, D., Burwinkel, B., and Brenner, H. (2012). Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease. *European Heart Journal* **33,** 2841–2848.

Breitling, L. P., Yang, R., Korn, B., Burwinkel, B., and Brenner, H. (2011). Tobacco-smoking-related differential dna methylation: 27k discovery and replication. *The American Journal of Human Genetics* **88,** 450–457.

Cecil, C. A., Lysenko, L. J., Jaffee, S. R., Pingault, J.-B., Smith, R. G., Relton, C. L., Woodward, G., McArdle, W., Mill, J., and Barker, E. D. (2014). Environmental risk, oxytocin receptor gene (oxtr) methylation and youth callous-unemotional traits: a 13-year longitudinal study. *Molecular Psychiatry* **19,** 1071.

Du, P., Zhang, X., Huang, C.-C., Jafari, N., Kibbe, W. A., Hou, L., and Lin, S. M. (2010). Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587.

Dudoit, S. and van der Laan, M. (2007). *Multiple Testing Procedures with Applications to Genomics.* Springer Series in Statistics. Springer New York.

Efron, B. (2004). Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association* **99,** 96–104.

Efron, B. (2010). Correlated z-values and the accuracy of large-scale statistical estimates. *Journal of the American Statistical Association* **105,** 1042–1055.

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

Elliott, H. R., Tillin, T., McArdle, W. L., Ho, K., Duggirala, A., Frayling, T. M., Davey Smith, G., Hughes, A. D., Chaturvedi, N., and Relton, C. L. (2014). Differences in smoking associated DNA methylation patterns in south asians and europeans. *Clinical Epigenetics* **6:4,** 1–10.

Fasanelli, F., Baglietto, L., Ponzi, E., Guida, F., Campanella, G., Johansson, M., Grankvist, K., Johansson, M., Assumma, M. B., Naccarati, A., et al. (2015). Hypomethylation of smoking-related genes is associated with future lung cancer in four prospective cohorts. *Nature Communications* **6:10192,** 1–9.

Gaynor, S. M., Schwartz, J., and Lin, X. (2019). Mediation analysis for common binary outcomes. *Statistics in Medicine* **38,** 512–529.

Genovese, C. and Wasserman, L. (2004). A stochastic process approach to false discovery control. *Annals of Statistics* pages 1035–1061.

Houseman, E. A., Accomando, W. P., Koestler, D. C., Christensen, B. C., Marsit, C. J., Nelson, H. H., Wiencke, J. K., and Kelsey, K. T. (2012). Dna methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13,** 86.

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

Imai, K., Keele, L., and Yamamoto, T. (2010). Identification, inference and sensitivity analysis for causal mediation effects. *Statistical Science* **25,** 51–71.

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

Johnson, W. E., Li, C., and Rabinovic, A. (2007). Adjusting batch effects in microarray expression data using empirical bayes methods. *Biostatistics* **8,** 118–127.

Joubert, B. R., Håberg, S. E., Nilsen, R. M., Wang, X., Vollset, S. E., Murphy, S. K., Huang, Z., Hoyo, C., Midttun, Ø., Cupul-Uicab, L. A., et al. (2012). 450k epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy. *Environmental Health Perspectives* **120,** 1425.

Kang, J. (2020). Comparison of methods for the proportion of true null hypotheses in microarray studie. *Communications for Statistical Applications and Methods* **27,** 141–148.

Lepeule, J., Baccarelli, A., Tarantini, L., Motta, V., Cantone, L., Litonjua, A. A., Sparrow, D., Vokonas, P. S., and Schwartz, J. (2012). Gene promoter methylation is associated with lung function in the elderly: the normative aging study. *Epigenetics* **7,** 261–269.

Li, S., Wong, E. M., Bui, M., Nguyen, T. L., Joo, J.-H. E., Stone, J., Dite, G. S., Giles, G. G., Saffery, R., Southey, M. C., and Hopper, J. L. (2018). Causal effect of smoking on DNA methylation in peripheral blood: a twin and family study. *Clinical Epigenetics* **10,** 18.

Liu, Y., Aryee, M. J., Padyukov, L., Fallin, M. D., Hesselberg, E., Runarsson, A., Reinius, L., Acevedo, N., Taub, M., Ronninger, M., et al. (2013). Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature Biotechnology* **31,** 142–147.

MacKinnon, D. P., Lockwood, C., and Hoffman, J. (1998). A new method to test for mediation. *Paper presented at the annual meeting of the Society for Prevention Research, Park City, UT* .

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

Pearl, J. (2001). Direct and indirect effects. In *Proceedings of the seventeenth conference on uncertainty in artificial intelligence*, pages 411–420. Morgan Kaufmann Publishers Inc.

Philibert, R. A., Beach, S., Lei, M.-K., and Brody, G. H. (2013). Changes in DNA methylation at the aryl hydrocarbon receptor repressor may be a new biomarker for smoking. *Clinical Epigenetics* **5,** 19.

Philibert, R. A., Beach, S. R., and Brody, G. H. (2012). Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers. *Epigenetics* **7,** 1331–1338.

Pillai, N. S. and Meng, X.-L. (2016). An unexpected encounter with Cauchy and Lévy. *The Annals of Statistics* **44,** 2089–2097.

Reynolds, L. M., Wan, M., Ding, J., Taylor, J. R., Lohman, K., Su, D., Bennett, B. D., Porter, D. K., Gimple, R., Pittman, G. S., et al. (2015). Dna methylation of the aryl hydrocarbon receptor repressor associations with cigarette smoking and subclinical atherosclerosis. *Circulation: Cardiovascular Genetics* **8,** 707–716.

Robins, J. M. and Greenland, S. (1992). Identifiability and exchangeability for direct and indirect effects. *Epidemiology* **3,** 143–155.

Schwartzman, A., Dougherty, R. F., Lee, J., Ghahremani, D., and Taylor, J. E. (2009). Empirical null and false discovery rate analysis in neuroimaging. *Neuroimage* **44,** 71–82.

Shenker, N. S., Polidoro, S., van Veldhoven, K., Sacerdote, C., Ricceri, F., Birrell, M. A., Belvisi, M. G., Brown, R., Vineis, P., and Flanagan, J. M. (2012). Epigenome-wide association study in the european prospective investigation into cancer and nutrition (epic-turin) identifies novel genetic loci associated with smoking. *Human Molecular Genetics* page dds488.

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

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

Teschendorff, A. E., Marabita, F., Lechner, M., Bartlett, T., Tegner, J., Gomez-Cabrero, D., and Beck, S. (2012). A beta-mixture quantile normalization method for correcting probe design bias in illumina infinium 450k dna methylation data. *Bioinformatics* **29,** 189–196.

Tingley, D., Yamamoto, T., Keele, L., and Imai, K. (2013). Mediation: R package for causal mediation analysis (R package version 4.4).

Tommola, M., Ilmarinen, P., Tuomisto, L. E., Haanpää, J., Kankaanranta, T., Niemelä, O., and Kankaanranta, H. (2016). The effect of smoking on lung function: a clinical study of adult-onset asthma. *The European Respiratory Journal* **48,** 1298–1306.

Valeri, L. and VanderWeele, T. J. (2013). Mediation analysis allowing for exposure–mediator interactions and causal interpretation: Theoretical assumptions and implementation with sas and spss macros. *Psychological Methods* **18,** 137.

van der Vaart, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

van Iterson, M., van Zwet, E. W., Consortium, B., and Heijmans, B. T. (2017). Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution. *Genome Biology* **18,** 19.

VanderWeele, T. and Vansteelandt, S. (2009). Conceptual issues concerning mediation, interventions and composition. *Statistics and its Interface* **2,** 457–468.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Texts in Statistics. Springer New York.

Wu, H. and Zhang, Y. (2014). Reversing dna methylation: mechanisms, genomics, and biological functions. *Cell* **156,** 45–68.

Zhang, H., Zheng, Y., Zhang, Z., Gao, T., Joyce, B., Yoon, G., Zhang, W., Schwartz, J., Just, A., Colicino, E., Vokonas, P., Zhao, L., Lv, J., Baccarelli, A., Hou, L., and Liu, L. (2016). Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics* **32,** 3150–3154.

Zhang, Y., Schottker, B., Florath, I., Stock, C., Butterbach, K., Holleczek, B., Mons, U., and Brenner, H. (2016). Smoking-associated DNA methylation biomarkers and their predictive value for all-cause and cardiovascular mortality. *Environmental Health Perspectives* **124,** 67–74.

Zhang, Y., Yang, R., Burwinkel, B., Breitling, L. P., and Brenner, H. (2014). F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environmental Health Perspectives* **122,** 131.