7-25-2022

# HIMA2: High-dimensional mediation analysis and its application in epigenome-wide DNA methylation data

Chamila Perera

Haixiang Zhang

Yinan Zheng

Lifang Hou

Annie Qu

*See next page for additional authors*

Authors
Chamila Perera, Haixiang Zhang, Yinan Zheng, Lifang Hou, Annie Qu, Cheng Zheng, Ke Xie, and Lei Liu

BMC Bioinformatics

RESEARCH

Open Access

# HIMA2: high-dimensional mediation analysis and its application in epigenome-wide DNA methylation data

Chamila Perera[1], Haixiang Zhang[2], Yinan Zheng[3], Lifang Hou[3], Annie Qu[4], Cheng Zheng[5], Ke Xie[1] and Lei Liu[1*]

*Correspondence:
lei.liu@wustl.edu

[1] Division of Biostatistics, Washington University in St. Louis, St. Louis, MO 63110, USA
[2] Center for Applied Mathematics, Tianjin University, Tianjin 300072, China
[3] Department of Preventive Medicine, Northwestern University, Chicago, IL 60611, USA
[4] Department of Statistics, University of California, Irvine, CA 92697, USA
[5] Department of Biostatistics, University of Nebraska Medical Center, Omaha, NE 68198, USA

## Abstract

Mediation analysis plays a major role in identifying significant mediators in the pathway between environmental exposures and health outcomes. With advanced data collection technology for large-scale studies, there has been growing research interest in developing methodology for high-dimensional mediation analysis. In this paper we present HIMA2, an extension of the HIMA method (Zhang in Bioinformatics 32:3150–3154, 2016). First, the proposed HIMA2 reduces the dimension of mediators to a manageable level based on the sure independence screening (SIS) method (Fan in J R Stat Soc Ser B 70:849–911, 2008). Second, a de-biased Lasso procedure is implemented for estimating regression parameters. Third, we use a multiple-testing procedure to accurately control the false discovery rate (FDR) when testing high-dimensional mediation hypotheses. We demonstrate its practical performance using Monte Carlo simulation studies and apply our method to identify DNA methylation markers which mediate the pathway from smoking to reduced lung function in the Coronary Artery Risk Development in Young Adults (CARDIA) Study.

**Keywords:** Variable selection, Joint significant test, Epigenetics, Causality

## Introduction

Mediation analysis explores the underlying mechanism by which an independent variable (e.g., exposure or treatment) influences the dependent variable (e.g., health outcome) through a mediator variable [1]. Mediation analysis has been playing a major role in many areas, e.g., social studies, economics, and health sciences [2]. More recently, with the advancement of large-scale data collection techniques, there has been substantial interest in developing methodology for high-dimensional mediation analysis in omics and imaging studies. An incomplete list of publications include [2–22]. For example, Derkach et al. [11] considered a latent variable model for high-dimensional mediation analysis. Huang et al. [12] presented a hypothesis test of the mediation effect in a causal mediation model with high-dimensional continuous mediators. Dai et al. [22] developed a multiple-testing procedure that accurately controls the false discovery rate (FDR) when testing high-dimensional mediation hypotheses.

Our motivating example comes from the DNA methylation (DNAm) research of the Coronary Artery Risk Development in Young Adults (CARDIA) Study [23]. In the DNAm process, methyl groups are added to DNA at binding sites referred to as cytosine-phosphate-guanine (CpG) islands, which inhibits the binding of transcription factors to DNA and results in changes (typically down regulation) to the expression of genes [24]. The platform Illumina MethylationEPIC Beadchip array is used to measure DNAm levels of roughly 850 K probes, which are ultra high-dimensional. Such high-dimensional DNAm markers may mediate pathways linking environmental exposures with health outcomes. Our objective is to explore the mediating role from high dimensional DNAm markers on the relationship between smoking and lung function in the CARDIA study.

In this paper, we propose an improved estimation and inference procedure for the high-dimensional mediation model, extending the work of Zhang et al. [3]. Our method includes three major steps: First, to tackle the ultra-high dimensionality of the DNAm markers, we screen out potentially a large number of mediators using a series of marginal mediation effect pathways (exposure → mediator → outcome). Second, we adopt the de-biased Lasso method [25] to estimate the high dimensional regression coefficients (mediator → outcome). Third, we employ a joint significance test with a mixture of null distributions to accurately control the FDR for large-scale multiple tests [22].

The remainder of this paper is structured as follows. In "Methodology" Section, we propose a three-step inference procedure for mediation effects in the high-dimensional regression model. In "Simulation studies" Section, we evaluate the performance of our method via numerical simulations. In "Application" Section, an application to the CARDIA study is provided. Finally, some discussion and concluding remarks are presented in "Conclusion and remarks" Section.

## Methodology

Denote the exposure as $X$, baseline covariates to be adjusted for as $\boldsymbol{Z} = (Z_1, \ldots, Z_q)^T$, where the superscript $T$ denotes the transpose of a vector or a matrix. We adopt the following counterfactual framework for the vector of potential mediators $\boldsymbol{M}(x) = (M_1(x), \ldots, M_p(x))^T$ under exposure level $x$, and counterfactual $Y(x, \boldsymbol{m})$ under exposure level $x$ and mediators level $\boldsymbol{m}$, to perform the mediation analysis [26]:

$$Y(x, \boldsymbol{m}) = \gamma x + \boldsymbol{\beta}^T \boldsymbol{m} + \boldsymbol{\eta}^T \boldsymbol{Z} + \varepsilon \tag{1'}$$

$$M_j(x) = \alpha_j x + \boldsymbol{\delta}_j^T \boldsymbol{Z} + e_j \text{ for } j = 1, \ldots, p \tag{2'}$$

where $\gamma$ is the direct effect of exposure on the outcome; $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the regression parameter vector relating the mediators to the outcome; $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_p)^T$ is the parameter vector relating the exposure to mediators; $\boldsymbol{\eta}$ and $\boldsymbol{\delta_j}$ are vectors of regression coefficients for the covariates; and $\varepsilon$ and $e_j$ are error terms in Models (1') and (2'), respectively. Note there are $p$ submodels in Model (2'), one for each mediator. We allow the correlation between the error terms, i.e., $\boldsymbol{e} = (e_1, \ldots, e_p)^T \sim N(0, \Sigma_e)$, where $\Sigma_e$ is a positive definite covariance matrix.

A few causal assumptions that are needed for the identification of natural direct effect (NDE) and natural indirect effects (NIE) are listed below [41–42]:

A1. Stable unit treatment value assumption (SUTVA) for both the mediators and the outcome. This assumption means that there is no multiple versions of exposures and there is no interference between individuals, which implies that $\boldsymbol{M}(x)$ and $Y(x, \boldsymbol{m})$ are well defined.

A2. Consistency for the mediators and the outcome. That is, there are no measurement errors in the mediators and thus the observed variables satisfy $\boldsymbol{M} = \boldsymbol{M}(X)$ and $Y = Y(X, \boldsymbol{M})$.

A3. Sequential ignorability: This assumption contains 4 parts:

(A3.1) $X \perp Y(x, \boldsymbol{m})|\boldsymbol{Z}$, i.e., no unmeasured confounding between exposure and the potential outcome;

(A3.2) $\boldsymbol{M} \perp Y(x, \boldsymbol{m})|X, \boldsymbol{Z}$, i.e., no unmeasured confounding between mediators and the potential outcome;

(A3.3) $X \perp \boldsymbol{M}(x)|\boldsymbol{Z}$, i.e., no unmeasured confounding between exposure and the potential mediators;

(A3.4) $\boldsymbol{M}\left(x'\right) \perp Y(x, \boldsymbol{m})|\boldsymbol{Z}$, i.e., no exposure-induced confounding between mediators and the potential outcome. In other words, the potential mediators under any intervention level $\boldsymbol{m}$ are independent of potential outcomes under any intervention $x$ and mediator level $x'$ given covariate $\boldsymbol{Z}$.

A4. No direct causal relationship between mediators. We do not allow one mediator to be the cause of another, but we do allow them to have shared common causes.

Under A1-A3, we have direct effect $NDE = E[Y(1, \boldsymbol{M}(0)) - Y(0, \boldsymbol{M}(0))] = \gamma$, indirect effect $NIE = E[Y(1, \boldsymbol{M}(1)) - Y(1, \boldsymbol{M}(0))] = \sum_{j=1}^{p} \alpha_j \beta_j$. Under the additional assumption A4, we can decompose the indirect effect into sum of indirect effects through each mediator $M_j$, $NIE_j = \alpha_j \beta_j$. Also we obtain the structural equation model for the observed outcome as in previous literature [3] to assess the mediation effects of high-dimensional mediators:

$$Y = \gamma X + \boldsymbol{\beta}^{\boldsymbol{T}} \boldsymbol{M} + \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{Z} + \varepsilon, \tag{1}$$

$$M_j = \alpha_j X + \boldsymbol{\delta}_{\boldsymbol{j}}^{\boldsymbol{T}} \boldsymbol{Z} + e_j \text{ for } j = 1, \ldots, p, \tag{2}$$

Our goal is to estimate and test the mediation effects $\alpha_j \beta_j$ jointly for $j = 1, \ldots, p$. An illustration of mediation analyses with single mediator and high dimensional mediators is given in Fig. 1.

As shown in Fig. 1, we do allow these mediators to share common unmeasured causes. These assumptions are in line with the underlying biologic procedures. Smoking could induce biochemical alterations to the DNAs, which lead to methylation changes. Such change in a certain CpG site is unlikely to *directly* cause the methylation alternation of other CpG sites. Rather, such dependency is most likely to be indirect, for example, by regulating gene expressions in related pathways that in turn modify other CpGs, or several CpGs are modified by common unmeasured causes (e.g., inflammatory response).

Details of our proposed approach are given below.

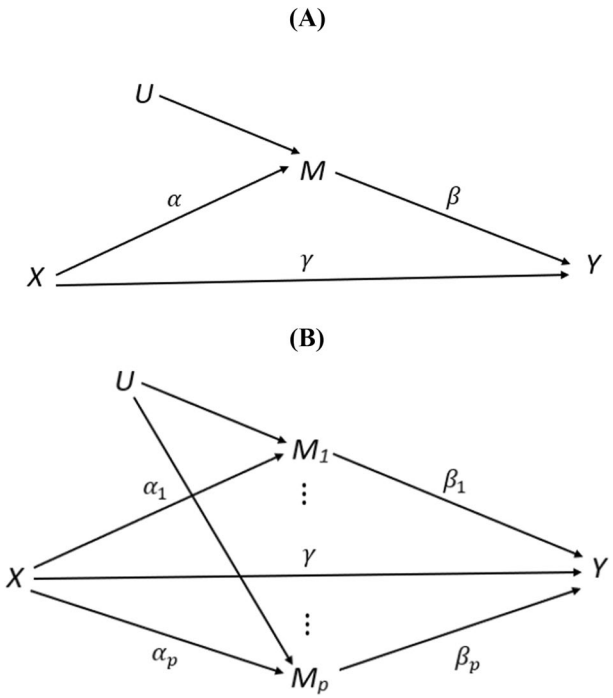


**Fig. 1** Mediation analysis of **A** a single mediator; **B** high dimensional mediators, plotted similarly to [3]. An arrow from $X$ to $U$ is possible though omitted to avoid the complexity in interpreting $\alpha$ as the total effect

*Step 1*: (*Screening of Mediators*). For $j = 1, \ldots, p$, we consider a series of marginal models:

$$Y = \gamma X + \beta_j M_j + \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{Z} + \varepsilon \tag{3}$$

$$M_j = \alpha_j X + \boldsymbol{\delta}_{\boldsymbol{j}}^{\boldsymbol{T}} \boldsymbol{Z} + e_j \tag{4}$$

Along the lines of the sure independence screening (SIS) method [27], we select a subset $\mathcal{D} = \{j : M_j$ is among the top $d = [2n/\log(n)]$ largest effect$\left|\widehat{\alpha}_j \widehat{\beta}_j\right|$, for$j = 1, \ldots, p\}$, where $\widehat{\alpha}_j$ and $\widehat{\beta}_j$ are ordinary least square (OLS) estimators based on marginal models (3) and (4), respectively.

All $M_j$'s are scaled with mean zero and unit variance before performing this screening procedure. The key advantage of Step 1 is that the product term $\widehat{\alpha}_j \widehat{\beta}_j$ could roughly describe the mediated effect of the $j$ th mediator. Therefore, the selected subset $\mathcal{D}$ contains true mediators with a large probability.

*Step 2*: (*De-biased Lasso*). We consider the following submodel based on the selected set $\mathcal{D}$,

$$Y = \gamma X + \boldsymbol{\beta}_{\mathcal{D}}^{\boldsymbol{T}} \boldsymbol{M}_{\mathcal{D}} + \boldsymbol{\eta}^{\boldsymbol{T}} \boldsymbol{Z} + \varepsilon \tag{5}$$

where $\boldsymbol{\beta}_{\mathcal{D}}$ and $\boldsymbol{M}_{\mathcal{D}}$ denote sub-vectors of $\boldsymbol{\beta}$ and $\boldsymbol{M}$ with index belonging to $\mathcal{D}$ respectively, and $\boldsymbol{\beta}_{\mathcal{D}}$ is estimated using the de-biased Lasso method, with estimator $\widehat{\beta}_j$ and its standard error $\widehat{\sigma}_{\beta_j}$ obtained in [25]. The corresponding p-values are given as:

$$P_{\beta_j} = 2\left\{1 - \Phi\left(\left|\hat{\beta}_j\right|/\hat{\sigma}_{\beta_j}\right)\right\}, \text{ for } j\mathcal{D} \tag{6}$$

where $\Phi(\cdot)$ is the cumulative distribution function of $N(0,1)$. De-biased Lasso in Step 2 is necessary as the ordinary least square will yield inefficient estimates (with reduced power), because the dimension of survived mediators after Step 1 is still relatively large.

*Step 3:* (*Joint Significance Test*). We consider the multiple testing problem for $j\epsilon\mathcal{D}$ as follows:

$$H_{0j} : \alpha_j = 0 \text{ or } \beta_j = 0,$$

with corresponding p-value

$$P_j = \max\left(P_{\alpha_j}, P_{\beta_j}\right) \tag{7}$$

where $P_{\beta_j}$ is given in (6), $P_{\alpha_j} = 2\left\{1 - \Phi\left(\left|\widehat{\alpha}_j\right|/\widehat{\sigma}_{\alpha_j}\right)\right\}$, $\widehat{\alpha}_j$ and $\widehat{\sigma}_{\alpha_j}$ are OLS estimators. Zhang et al. [3] considered the joint significant test (termed "JS-uniform"), which assumes that $P_j$ follows a uniform distribution. However, although $P_{\alpha_j}$ and $P_{\beta_j}$ are each uniformly distributed, their maximum is not. As a result, the significance rule using the uniform null distribution for $P_j$ results in a valid but overly conservative test [28]. In this paper, we will adopt the "JS-mixture" approach to accurately control the FDR [22] (Sect. 2.3).

The multiple testing problem (7) is equivalent to the union of the following three disjoint component null hypotheses,

$$H_{00,j} : \alpha_j = 0 \text{ or } \beta_j = 0,$$

$$H_{01,j} : \alpha_j = 0 \text{ or } \beta_j \neq 0,$$

$$H_{10,j} : \alpha_j \neq 0 \text{ or } \beta_j = 0.$$

That is, $P_j$ is a 3-component mixture distribution instead of the uniform distribution. Dai et al. [22] proposed the following estimated FDR for testing mediation:

$$\widehat{FDR}(t) = \frac{\hat{\pi}_{01}t + \hat{\pi}_{10}t + \hat{\pi}_{00}t^2}{\max\{1, R(t)\}/d} \tag{8}$$

where $\widehat{\pi}_{01}, \widehat{\pi}_{10}$ and $\widehat{\pi}_{00}$ are the estimates of proportions $H_{01,j}, H_{10,j}$ and $H_{00,j}$, respectively, and $R(t) = V_{00}(t) + V_{01}(t) + V_{10}(t) + V_{11}(t)$, where $V_{00}(t) = \#\{P_j \leq t | H_{00}\}$, $V_{01}(t) = \#\{P_j \leq t | H_{01}\}, V_{10}(t) = \#\{P_j \leq t | H_{10}\}, V_{11}(t) = \#\{P_j \leq t | H_{11}\}$ for $t \in [0,1]$.

We define the significant threshold for $P_j$ as $\widehat{t}_b = \sup\left\{t : \widehat{FDR}(t) \leq b\right\}$, to control the FDR at level $b$. Then $\widehat{S} = \left\{j : P_j \leq \widehat{t}_b, j \in \mathcal{D}\right\}$ gives the estimated index set of significant mediators.

We can obtain $\widehat{\pi}_{01}, \widehat{\pi}_{10}, \widehat{\pi}_{00}$ and $\widehat{t}_b$ using the R package HDMT [22].

Compared to the estimation and inference method in [3] (termed "HIMA"), our new method (termed "HIMA2") has the following three advantages. First, HIMA only considers $\beta$ (mediator $\rightarrow$ outcome) for screening in Step 1, while HIMA2 considers

the indirect effect of $\alpha\beta$. Therefore, the mediation-based screening method in HIMA2 addresses the indirect effect more accurately than HIMA. Second, HIMA uses the minimax concave penalty (MCP; [29]) technique to estimate the effect $\beta$, which can only provide p-values for selected mediators in Step 2. That is, $P_{\beta_j}$ is set to 1 for those not selected, which results in poor estimate of $P_j$ in Eq. (7). In contrast, de-biased Lasso in HIMA2 yields p-values for all $\beta_j$'s in $\mathcal{D}$, which gives more appropriate estimate of $P_{\beta_j}$. Third, HIMA adopts a naive joint significance rule assuming a uniform null distribution for the maximum p-value calculation in Step 3, which may result in a valid but overly conservative test with lower power.

## Simulation studies

In this section we assess our proposed method using simulation studies. For Model (1), we generate the exposure $X$ from $N(0,2)$; covariates $Z=(Z_1,Z_2)^T$, where $Z_1$ and $Z_2$ are independently generated from $N(0,2)$. We set $\gamma=0.5$, $\delta=(0.3,0.3)^T$ and $\eta=(0.5,0.5)^T$; $\beta_1=0.20$, $\beta_2=0.25$, $\beta_3=0.15$, $\beta_4=0.30$, $\beta_5=0.35$, $\beta_6=0.10$, and $\beta_j=0$ for all other $j$'s; $\alpha_1=0.20, \alpha_2=0.25, \alpha_3=0.15, \alpha_4=0.30, \alpha_5=0.35, \alpha_7=0.10$, and $\alpha_j=0$ for all other $j$'s. Therefore, we have: (i) $\alpha_j\beta_j\neq 0$ for $j=1,\ldots,5$; (ii) $\alpha_j=0$ but $\beta_j\neq 0$ for $j=6$; (iii) $\alpha_j\neq 0$ but $\beta_j=0$ for $j=7$; and (iv) $\alpha_j=0$ and $\beta_j=0$ for $j>7$. The error terms $e=\left(e_1,\ldots,e_p\right)^T$ are generated from $N(0,\Sigma_e)$, where $\Sigma_e=\left(\rho^{|j-j'|}\right)_{j,j'}$ and $\varepsilon$ is generated from $N(0,1)$. All the simulations are based on 500 replications with 16 factorial settings: $p=1000,5000$, $n=300$, 600, and $\rho=0,0.25,0.5,0.75$.

We compare the performance of HIMA2 with HIMA in Table 1, which provides the estimated biases (Bias) given by the sample mean of the estimates minus the true value, and the mean-square error (MSE) of the estimates. Table 1 shows that both HIMA2 and HIMA are unbiased, however, HIMA2 has smaller MSEs than HIMA for significant mediators. MSEs for both HIMA and HIMA2 decrease as the sample size increases. Of note, the results for $j>8$ ($\alpha_j=0$ and $\beta_j=0$) are close to those of $j=8$ and thus omitted.

We also present the estimated FDR and power of mediation effects testing in Tables 2 and 3, where the nominal level is 0.05. The results indicate that both HIMA2 and HIMA can achieve valid FDR control. Furthermore, HIMA2 is more powerful than HIMA in selecting significant mediators, though the differences become smaller when sample size increases. We also note that as the correlation among the mediators becomes larger, both methods suffer in terms of power.

Per suggestion from a reviewer, we compare our method to HDMA [30], which was developed along the lines of HIMA but adopts the de-biased Lasso method in Step 2. However, no multiple testing adjustment was used in HDMA for inference. As a result, HDMA suffers from poor FDR control albeit with higher power as shown in Tables 2 and 3.

Per suggestion from a reviewer, similar to our real data analysis, we also consider a setting with 2 significant mediators, i.e.: $\beta_1=0.15$, $\beta_2=0.3$, $\beta_3=0.1$, $\beta_4=0$, and $\beta_j=0$ for all other $j$'s; $\alpha_1=0.15, \alpha_2=0.3, \alpha_3=0, \alpha_4=0.1$, and $\alpha_j=0$ for all other $j$ 's. As shown in the supplementary materials (Additional file 1: Tables S1, S2 and S3), we

**Table 1** Bias (MSE) for mediation effect estimates

| | | $\rho = 0$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 300$ | $\alpha_1\beta_1$ | 7.21E−04 (1.72E−04) | −7.07E−03 (4.23E−04) | −8.95E−03 (2.80E−04) | −8.90E−03 (1.97E−04) | −2.23E−02 (8.19E−04) | −2.25E−02 (7.14E−04) |
| | $\alpha_2\beta_2$ | −2.44E−04 (2.80E−04) | −7.27E−03 (5.10E−04) | −1.30E−02 (4.28E−04) | −1.34E−02 (4.12E−04) | −2.71E−02 (1.22E−03) | −2.90E−02 (1.09E−03) |
| | $\alpha_3\beta_3$ | 9.45E−04 (1.00E−04) | −8.32E−03 (2.84E−04) | −7.54E−03 (1.91E−04) | −4.17E−03 (1.17E−04) | −1.43E−02 (3.63E−04) | −1.40E−02 (3.00E−04) |
| | $\alpha_4\beta_4$ | −2.22E−03 (4.86E−04) | −1.16E−02 (7.75E−04) | −1.95E−02 (8.11E−04) | −1.87E−02 (6.97E−04) | −3.27E−02 (1.52E−03) | −3.89E−02 (1.77E−03) |
| | $\alpha_5\beta_5$ | −4.82E−03 (5.89E−04) | −1.63E−02 (1.03E−03) | −2.66E−02 (1.23E−03) | −2.51E−02 (1.07E−03) | −4.48E−02 (2.66E−03) | −5.21E−02 (3.04E−03) |
| | $\alpha_6\beta_6$ | 6.06E−05 (1.05E−05) | −3.76E−05 (8.92E−06) | 4.81E−05 (7.21E−06) | 5.10E−05 (4.32E−06) | 3.56E−05 (1.62E−06) | 2.88E−05 (1.57E−06) |
| | $\alpha_7\beta_7$ | 2.41E−03 (3.71E−05) | 3.21E−04 (5.91E−06) | 3.75E−04 (8.43E−06) | 1.13E−03 (1.81E−05) | 5.43E−05 (1.85E−06) | 6.02E−05 (1.60E−06) |
| | $\alpha_8\beta_8$ | 1.11E−04 (1.08E−06) | 1.17E−05 (1.09E−07) | 3.62E−05 (2.72E−07) | 4.58E−05 (4.02E−07) | −1.03E−05 (4.31E−08) | −5.98E−06 (3.04E−08) |
| $n = 600$ | $\alpha_1\beta_1$ | 1.46E−03 (1.01E−04) | 2.88E−03 (1.64E−04) | −4.34E−03 (1.06E−04) | −6.78E−03 (1.18E−04) | −9.43E−03 (2.22E−04) | −1.49E−02 (2.89E−04) |
| | $\alpha_2\beta_2$ | 1.00E−03 (1.59E−04) | 3.47E−03 (2.45E−04) | −7.38E−03 (2.00E−04) | −1.01E−02 (2.23E−04) | −1.32E−02 (3.40E−04) | −2.26E−02 (6.03E−04) |
| | $\alpha_3\beta_3$ | 8.87E−04 (4.72E−05) | −2.49E−04 (1.38E−04) | −3.13E−03 (6.45E−05) | −3.22E−03 (5.42E−05) | −7.51E−03 (1.67E−04) | −9.34E−03 (1.40E−04) |
| | $\alpha_4\beta_4$ | 7.51E−04 (1.97E−04) | 4.32E−03 (2.97E−04) | −1.03E−02 (2.94E−04) | −1.44E−02 (4.12E−04) | −1.90E−02 (5.96E−04) | −3.22E−02 (1.18E−03) |
| | $\alpha_5\beta_5$ | −1.03E−03 (2.80E−04) | 2.39E−03 (4.03E−04) | −1.59E−02 (5.36E−04) | −2.12E−02 (7.05E−04) | −2.89E−02 (1.13E−03) | −4.57E−02 (2.27E−03) |
| | $\alpha_6\beta_6$ | 1.15E−05 (4.16E−06) | −9.97E−05 (4.28E−06) | −3.24E−05 (3.21E−06) | 1.59E−04 (2.27E−06) | 1.43E−04 (2.06E−06) | 1.37E−04 (1.45E−06) |
| | $\alpha_7\beta_7$ | 2.20E−03 (2.09E−05) | 1.29E−04 (1.32E−06) | 2.99E−04 (7.03E−06) | 6.68E−04 (1.07E−05) | 4.24E−05 (1.28E−06) | 1.01E−04 (1.29E−06) |
| | $\alpha_8\beta_8$ | −4.17E−06 (4.54E−07) | 3.83E−06 (7.98E−08) | −4.66E−06 (2.06E−07) | −6.36E−06 (1.62E−07) | −4.77E−06 (9.88E−09) | 5.34E−06 (4.52E−08) |
| | | $\rho = 0.25$ | | | | | |
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 300$ | $\alpha_1\beta_1$ | 4.46E−04 (1.82E−04) | −7.65E−03 (4.95E−04) | −7.79E−03 (2.49E−04) | −6.92E−03 (1.95E−04) | −1.85E−02 (6.84E−04) | −1.84E−02 (5.13E−04) |
| | $\alpha_2\beta_2$ | 3.22E−04 (3.08E−04) | −6.70E−03 (5.90E−04) | −1.07E−02 (3.88E−04) | −1.07E−02 (3.65E−04) | −2.17E−02 (9.31E−04) | −2.58E−02 (8.66E−04) |
| | $\alpha_3\beta_3$ | 5.90E−04 (1.09E−04) | −7.05E−03 (2.83E−04) | −4.87E−03 (1.40E−04) | −3.934E−03 (1.01E−04) | −1.33E−02 (3.34E−04) | −1.22E−02 (2.38E−04) |
| | $\alpha_4\beta_4$ | −2.44E−03 (4.26E−04) | −1.04E−02 (7.13E−04) | −1.74E−02 (7.17E−04) | −1.81E−02 (6.81E−04) | −3.33E−02 (1.69E−03) | −3.79E−02 (1.74E−03) |
| | $\alpha_5\beta_5$ | −3.57E−03 (5.93E−04) | −1.53E−02 (1.02E−03) | −2.34E−02 (1.15E−03) | −2.25E−02 (9.91E−04) | −4.21E−02 (2.39E−03) | −5.08E−02 (2.94E−03) |
| | $\alpha_6\beta_6$ | 1.77E−04 (9.38E−06) | 1.56E−04 (7.85E−06) | 2.24E−04 (7.31E−06) | −4.44E−05 (4.70E−06) | 2.45E−05 (2.84E−06) | −1.12E−05 (2.35E−06) |
| | $\alpha_7\beta_7$ | 3.63E−03 (4.74E−05) | 4.27E−04 (1.09E−05) | 7.49E−04 (1.07E−05) | 1.89E−03 (2.29E−05) | 1.84E−04 (3.29E−06) | 2.20E−04 (3.13E−06) |
| | $\alpha_8\beta_8$ | 7.33E−05 (9.05E−07) | −2.12E−05 (1.81E−07) | −9.48E−06 (2.56E−07) | 3.63E−05 (4.97E−07) | −1.24E−05 (3.90E−08) | 6.43E−07 (2.70E−08) |

**Table 1** (continued)

| | | $\rho = 0.25$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 600$ | $\alpha_1\beta_1$ | 8.96E−04 (9.43E−05) | 2.08E−03 (1.75E−04) | −3.94E−03 (1.07E−04) | −6.44E−03 (1.24E−04) | −8.81E−03 (1.98E−04) | −1.44E−02 (2.69E−04) |
| | $\alpha_2\beta_2$ | 3.74E−04 (1.50E−04) | 2.77E−03 (2.21E−04) | −6.33E−03 (1.95E−04) | −9.28E−03 (2.15E−04) | −1.32E−02 (3.50E−04) | −2.20E−02 (5.88E−04) |
| | $\alpha_3\beta_3$ | 1.41E−03 (5.39E−05) | 1.33E−04 (1.31E−04) | −1.94E−03 (5.17E−05) | −3.45E−03 (5.67E−05) | −6.69E−03 (1.41E−04) | −8.19E−03 (1.06E−04) |
| | $\alpha_4\beta_4$ | 3.66E−04 (2.13E−04) | 3.32E−03 (2.95E−04) | −8.58E−03 (2.91E−04) | −1.35E−02 (3.75E−04) | −1.93E−02 (5.98E−04) | −3.14E−02 (1.13E−03) |
| | $\alpha_5\beta_5$ | 4.79E−04 (3.29E−04) | 3.09E−03 (4.39E−04) | −1.15E−02 (4.74E−04) | −1.95E−02 (6.63E−04) | −2.81E−02 (1.11E−03) | −4.37E−02 (2.12E−03) |
| | $\alpha_6\beta_6$ | −2.74E−05 (3.95E−06) | −1.16E−04 (4.39E−06) | −2.32E−05 (3.27E−06) | −1.33E−04 (2.97E−06) | −1.96E−04 (2.85E−06) | −6.82E−05 (1.90E−06) |
| | $\alpha_7\beta_7$ | 3.17E−03 (2.65E−05) | 3.37E−04 (4.72E−06) | 1.13E−03 (8.65E−06) | 2.24E−03 (1.90E−05) | 1.23E−04 (1.92E−06) | 2.37E−04 (2.55E−06) |
| | $\alpha_8\beta_8$ | 1.62E−05 (3.35E−07) | −5.68E−07 (4.13E−08) | −6.73E−07 (2.26E−07) | 2.72E−05 (1.54E−07) | 1.69E−05 (6.54E−08) | 1.30E−05 (2.72E−08) |
| | | $\rho = 0.50$ | | | | | |
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 300$ | $\alpha_1\beta_1$ | 3.76E−03 (2.34E−04) | −3.46E−03 (4.90E−04) | −1.78E−03 (2.13E−04) | −5.14E−03 (1.98E−04) | −1.63E−02 (6.36E−04) | −1.46E−02 (3.74E−04) |
| | $\alpha_2\beta_2$ | 2.82E−03 (3.66E−04) | −5.79E−03 (8.16E−04) | −4.11E−03 (3.79E−04) | −7.97E−03 (3.72E−04) | −2.14E−02 (1.13E−03) | −2.18E−02 (7.77E−04) |
| | $\alpha_3\beta_3$ | 2.25E−03 (1.42E−04) | −6.91E−03 (3.36E−04) | −9.13E−04 (1.19E−04) | −3.27E−03 (1.19E−04) | −1.32E−02 (3.33E−04) | −8.64E−03 (1.59E−04) |
| | $\alpha_4\beta_4$ | 1.03E−03 (5.02E−04) | −8.62E−03 (9.89E−04) | −7.04E−03 (5.47E−04) | −1.15E−02 (5.83E−04) | −2.90E−02 (1.65E−03) | −2.97E−02 (1.26E−03) |
| | $\alpha_5\beta_5$ | 6.66E−03 (7.53E−04) | −5.38E−03 (1.01E−03) | −7.22E−03 (7.94E−04) | −1.27E−02 (7.87E−04) | −3.52E−02 (2.06E−03) | −4.02E−02 (2.24E−03) |
| | $\alpha_6\beta_6$ | −3.47E−05 (8.66E−06) | −2.58E−04 (7.30E−06) | 9.43E−05 (7.46E−06) | 6.55E−05 (6.88E−06) | 6.48E−05 (6.34E−06) | 7.46E−05 (5.69E−06) |
| | $\alpha_7\beta_7$ | 4.69E−03 (5.74E−05) | 9.18E−04 (1.63E−05) | 2.09E−03 (2.36E−05) | 3.81E−03 (4.57E−05) | 5.83E−04 (9.04E−06) | 1.06E−03 (1.21E−05) |
| | $\alpha_8\beta_8$ | 1.82E−05 (1.91E−06) | −3.56E−05 (5.92E−07) | −1.17E−05 (7.80E−07) | −1.28E−05 (4.19E−07) | −2.13E−05 (1.67E−07) | 1.08E−05 (1.89E−07) |

**Table 1** (continued)

| | | $\rho = 0.50$ | | | | | |
|---|---|---|---|---|---|---|---|
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 600$ | $\alpha_1\beta_1$ | 2.44E−03 (1.08E−04) | 2.14E−03 (1.74E−04) | 4.14E−04 (1.01E−04) | −3.50E−03 (1.01E−04) | −6.87E−03 (1.99E−04) | −1.07E−02 (1.95E−04) |
| | $\alpha_2\beta_2$ | 3.41E−03 (2.02E−04) | 4.48E−03 (2.75E−04) | 7.19E−04 (1.71E−04) | −5.51E−03 (2.26E−04) | −9.64E−03 (3.32E−04) | −1.68E−02 (4.46E−04) |
| | $\alpha_3\beta_3$ | 1.76E−03 (7.22E−05) | −3.66E−03 (1.91E−04) | 6.19E−04 (6.90E−05) | −1.24E−03 (6.31E−05) | −6.00E−03 (1.74E−04) | −5.34E−03 (7.98E−05) |
| | $\alpha_4\beta_4$ | 4.34E−03 (2.65E−04) | 4.77E−03 (3.36E−04) | 2.98E−04 (2.42E−04) | −9.92E−03 (3.48E−04) | −1.59E−02 (5.71E−04) | −2.48E−02 (8.05E−04) |
| | $\alpha_5\beta_5$ | 8.04E−03 (4.34E−04) | 9.29E−03 (5.06E−04) | 1.81E−03 (3.96E−04) | −9.00E−03 (4.37E−04) | −1.82E−02 (7.82E−04) | −3.04E−02 (1.23E−03) |
| | $\alpha_6\beta_6$ | 1.10E−04 (4.60E−06) | 5.09E−05 (3.85E−06) | 9.57E−05 (4.21E−06) | −1.66E−04 (3.28E−06) | −1.12E−04 (2.79E−06) | −7.62E−05 (2.46E−06) |
| | $\alpha_7\beta_7$ | 2.91E−03 (3.12E−05) | 8.64E−04 (1.22E−05) | 2.17E−03 (2.10E−05) | 3.23E−03 (2.66E−05) | 6.05E−04 (6.67E−06) | 1.23E−03 (9.06E−06) |
| | $\alpha_8\beta_8$ | 6.18E−06 (4.08E−07) | −1.44E−06 (1.23E−07) | −2.07E−06 (4.32E−07) | 2.24E−05 (2.45E−07) | 1.13E−05 (5.31E−08) | 5.00E−06 (1.28E−07) |
| | | $\rho = 0.75$ | | | | | |
| | | $p = 1000$ | | | $p = 5000$ | | |
| | | **HIMA2** | **HIMA** | **HDMA** | **HIMA2** | **HIMA** | **HDMA** |
| $n = 300$ | $\alpha_1\beta_1$ | 7.39E−03 (3.47E−04) | −2.75E−03 (7.68E−04) | 4.38E−03 (3.12E−04) | 2.65E−03 (2.98E−04) | −8.93E−03 (7.81E−04) | −4.19E−03 (2.82E−04) |
| | $\alpha_2\beta_2$ | 8.32E−03 (7.65E−04) | −1.43E−02 (2.29E−03) | 5.21E−03 (6.63E−04) | 2.17E−03 (5.13E−04) | −2.20E−02 (2.15E−03) | −8.00E−03 (6.11E−04) |
| | $\alpha_3\beta_3$ | 4.03E−03 (2.70E−04) | −1.05E−02 (5.41E−04) | 2.83E−03 (2.53E−04) | 8.24E−05 (2.15E−04) | −1.32E−02 (5.05E−04) | −3.20E−03 (2.06E−04) |
| | $\alpha_4\beta_4$ | 1.07E−02 (1.07E−03) | −6.93E−03 (2.81E−03) | 5.83E−03 (9.77E−04) | 1.03E−03 (7.32E−04) | −2.05E−02 (2.85E−03) | −1.21E−02 (9.51E−04) |
| | $\alpha_5\beta_5$ | 1.74E−02 (1.50E−03) | −1.48E−02 (4.46E−03) | 1.08E−02 (1.33E−03) | 7.51E−03 (1.09E−03) | −3.15E−02 (4.61E−03) | −1.28E−02 (1.26E−03) |
| | $\alpha_6\beta_6$ | 2.06E−05 (8.13E−06) | −1.05E−04 (6.97E−06) | 4.44E−05 (8.01E−06) | 9.89E−06 (7.72E−06) | −1.27E−04 (6.67E−06) | 6.18E−05 (7.25E−06) |
| | $\alpha_7\beta_7$ | 6.19E−03 (1.20E−04) | 1.66E−03 (4.07E−05) | 4.97E−03 (1.08E−04) | 4.63E−03 (8.72E−05) | 9.38E−04 (1.99E−05) | 2.55E−03 (7.09E−05) |
| | $\alpha_8\beta_8$ | −1.15E−04 (5.76E−06) | −8.81E−06 (8.67E−07) | −1.04E−04 (5.11E−06) | −1.78E−05 (2.51E−06) | −5.95E−06 (1.02E−06) | −6.78E−05 (2.77E−06) |
| $n = 600$ | $\alpha_1\beta_1$ | 4.81E−03 (1.64E−04) | 8.99E−04 (4.05E−04) | 3.88E−03 (1.57E−04) | 1.75E−03 (1.52E−04) | −2.42E−03 (3.86E−04) | −3.55E−03 (1.46E−04) |
| | $\alpha_2\beta_2$ | 8.29E−03 (4.00E−04) | −2.07E−03 (1.30E−03) | 6.75E−03 (3.79E−04) | 3.43E−03 (3.20E−04) | −8.50E−03 (1.30E−03) | −4.32E−03 (3.05E−04) |
| | $\alpha_3\beta_3$ | 2.88E−03 (1.17E−04) | −1.15E−02 (4.25E−04) | 2.38E−03 (1.15E−04) | 6.43E−04 (1.17E−04) | −1.38E−02 (4.22E−04) | −2.27E−03 (1.18E−04) |
| | $\alpha_4\beta_4$ | 1.06E−02 (5.68E−04) | 6.82E−03 (1.20E−03) | 9.09E−03 (5.32E−04) | 2.99E−03 (4.38E−04) | −4.25E−03 (1.26E−03) | −7.83E−03 (4.99E−04) |
| | $\alpha_5\beta_5$ | 1.57E−02 (8.59E−04) | 6.22E−03 (1.70E−03) | 1.26E−02 (7.55E−04) | 6.31E−03 (6.40E−04) | −8.25E−03 (1.96E−03) | −9.78E−03 (7.13E−04) |
| | $\alpha_6\beta_6$ | −9.35E−05 (3.98E−06) | −9.84E−06 (2.76E−06) | −1.16E−04 (3.92E−06) | −1.13E−04 (3.75E−06) | −6.49E−05 (2.45E−06) | −9.40E−05 (3.77E−06) |
| | $\alpha_7\beta_7$ | 4.63E−03 (6.31E−05) | 1.03E−03 (1.79E−05) | 3.94E−03 (5.85E−05) | 4.57E−03 (5.64E−05) | 1.13E−03 (1.85E−05) | 2.41E−03 (3.99E−05) |
| | $\alpha_8\beta_8$ | 2.92E−05 (1.56E−06) | −1.46E−05 (2.31E−07) | −2.62E−05 (1.51E−06) | 9.12E−07 (1.19E−06) | −1.37E−05 (7.91E−08) | −2.86E−05 (1.11E−06) |

**Table 2** FDR at significance level 0.05

| | $\rho = 0$ | | | | $\rho = 0.25$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $p = 1000$ | | $p = 5000$ | | $p = 1000$ | | $p = 5000$ | |
| **Method** | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ |
| HIMA2 | 0.0110 | 0.0030 | 0.0634 | 0.0214 | 0.0094 | 0.0053 | 0.0569 | 0.0202 |
| HIMA | 0.0225 | 0.0149 | 0.0316 | 0.0316 | 0.0244 | 0.0238 | 0.0320 | 0.0301 |
| HDMA | 0.2067 | 0.2553 | 0.2994 | 0.3739 | 0.1880 | 0.2299 | 0.2712 | 0.3678 |
| | $\rho = 0.50$ | | | | $\rho = 0.75$ | | | |
| | $p = 1000$ | | $p = 5000$ | | $p = 1000$ | | $p = 5000$ | |
| **Method** | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ |
| HIMA2 | 0.0099 | 0.0039 | 0.0351 | 0.0129 | 0.0055 | 0.0026 | 0.0097 | 0.0025 |
| HIMA | 0.0322 | 0.0253 | 0.0339 | 0.0281 | 0.0325 | 0.0232 | 0.0306 | 0.0327 |
| HDMA | 0.1482 | 0.1764 | 0.2533 | 0.3174 | 0.0990 | 0.1211 | 0.1740 | 0.1816 |

**Table 3** Power at significance level 0.05

| | $\rho = 0$ | | | | $\rho = 0.25$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $p = 1000$ | | $p = 5000$ | | $p = 1000$ | | $p = 5000$ | |
| **Method** | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ |
| HIMA2 | 0.8640 | 0.9608 | 0.8024 | 0.9392 | 0.8440 | 0.9512 | 0.8076 | 0.9364 |
| HIMA | 0.7760 | 0.9464 | 0.6192 | 0.8872 | 0.7800 | 0.9480 | 0.6472 | 0.9020 |
| HDMA | 0.8928 | 0.9848 | 0.7680 | 0.9496 | 0.9032 | 0.9880 | 0.8236 | 0.9652 |
| | $\rho = 0.50$ | | | | $\rho = 0.75$ | | | |
| | $p = 1000$ | | $p = 5000$ | | $p = 1000$ | | $p = 5000$ | |
| **Method** | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ | $n = 300$ | $n = 600$ |
| HIMA2 | 0.7996 | 0.9180 | 0.7596 | 0.9096 | 0.6612 | 0.8200 | 0.6416 | 0.8072 |
| HIMA | 0.7672 | 0.9252 | 0.6412 | 0.8876 | 0.5860 | 0.7584 | 0.5244 | 0.7232 |
| HDMA | 0.9052 | 0.9816 | 0.8136 | 0.9560 | 0.8436 | 0.9452 | 0.7900 | 0.9208 |

observe similar results to those in Tables 1, 2 and 3. We note that the results from HIMA and HIMA2 are more close to each other when the correlation is high ($\rho = 0.75$).

Per suggestion from a reviewer, we use the standardized coefficient estimates in the SIS step, but the results are close to those without standardization (results available upon request).

Finally, we notice that in Tables 2 and Additional file 1: Table S2, the FDR of HIMA2 decreases with sample size. This also happens with HIMA, though to a less magnitude.

## Application

We apply our method to the Coronary Artery Risk Development in Young Adults (CARDIA) Study, an ongoing longitudinal cohort examining the development and determinants of clinical and subclinical cardiovascular disease and their risk factors [23]. A group of 5115 black and white men and women aged 18–30 years were enrolled in 1985–6 from 4 study centers: Birmingham, AL; Chicago, IL; Minneapolis, MN; and Oakland, CA. They were followed-up during 1987–1988 (Year 2), 1990–1991 (Year 5),

1992–1993 (Year 7), 1995–1996 (Year 10), 2000–2001 (Year 15), 2005–2006 (Year 20), 2010–2011 (Year 25), and 2015–2016 (Year 30).

We are interested in investigating how the DNA methylation (DNAm) markers mediate the relation between smoking and lung function. Due to budget limitation, 1200 individuals from the CARDIA participants at Year 15 were randomly selected for DNAm profiling using the Illumina MethylationEPIC Beadchip ($p = \sim 850{,}000$ sites). The R package Enmix [31] was used to perform quality control, background correction, dye bias correction, quantile normalization (by probe types), and extreme outliers removal. Eventually, the DNAm measurements were obtained for a total of 1042 blood samples, which are treated as mediators in this study. The FEV1 (forced expiratory volume in 1 s) measured at Year 20 is considered as the lung function outcome. The number of cigarette packs/year in Year 10 is the exposure variable. We are interested in building the mediation pathway in sequence: smoking at Year 10 → High dimensional DNAm markers at Year 15 → lung function at Year 20.

Our analysis adjusts for age, height, weight, study center, gender, and race in Models (1) and (2). Additionally, we estimated the proportions of CD4 + T lymphocytes, CD8 + T lymphocytes, B lymphocytes, natural killer cells, monocytes, and granulocytes using [32], which are also adjusted in the models. To account for experimental batch effects and other technical biases, we derive surrogate variables from intensity data for non-negative internal control probes using principal components (PCs) analysis [31]. The top eight PCs, explaining 95.06% of the variation across the non-negative internal control probes, are also adjusted as covariates in the model. All the covariates are measured at Year 10.

After screening in Step 1, the average of the absolute values of correlation among CpGs is 0.25 (max 0.93). In Table 4, we present the summary results on selected mediators. For FDR < 0.05, HIMA2 identifies 2 CpGs: cg26331243 and cg19862839 as mediators. CpG cg26331243 is located in the body region of gene *CCDC33*, which is differentially expressed for tobacco smoke exposure [33, 34]. *CCDC33* is also linked to susceptibility to lung function disorders, e.g., pneumococcal meningitis [35] and SARS-CoV-2 infection [36]. Therefore, it is plausible that cg26331243 plays a role in regulating the expression of *CCDC33*, which in turn mediates the pathway from smoking to lung function.

CpG cg19862839 is located in the body region of gene *TBX4*. Growing evidence has indicated that *TBX4* variants are associated with a wide spectrum of lung disorders [37, 38]. Patients with mutations in *TBX4* may also be more susceptible to cigarette smoking [39]. Therefore, we speculate that cg19862839 could participate in regulating the expression of *TBX4*, which also acts as a mediator between smoking and lung function.

**Table 4** Summary of selected CpGs with mediation effects subject to FDR < 0.05

| CpGs | Chromosome[1] | Position[a] | Proximal gene target[b] | $\widehat{\alpha}_k\,(SE)$ | $\widehat{\beta}_k\,(SE)$ | $\widehat{\alpha}_k\widehat{\beta}_k$ | FDR |
|---|---|---|---|---|---|---|---|
| cg26331243 | chr15 | 74,550,946 | CCDC33 | −0.081 (0.016) | 0.084 (0.027) | −0.0067 | 0.0345 |
| cg19862839 | chr17 | 59,543,726 | TBX4 | −0.082 (0.024) | 0.059 (0.020) | −0.0049 | 0.0397 |

[a] Genome assembly GRCh37 (hg19)

[b] Based on UCSC RefGene

In comparison, HIMA only identifies cg26331243 as a mediator with FDR < 0.05. Therefore, the proposed HIMA2 has better power to identify CpGs in high dimensional mediation analysis.

Finally, we note that cg05575921, which was identified in the normative aging study (NAS) [3], is not a significant mediator in the CARDIA study. In CARDIA, the estimate of $\alpha$ (from smoking to DNAm) is highly significant for cg05575921. However, the estimate of $\beta$ (from DNAm to FEV1) is not significant. This may be due to that participants in CARDIA were much younger (mean age 45 at Year 20, range 38–55) than NAS (mean age 74, range 55–100), when the lung function of CARDIA participants are more homogenous. Therefore, the association between DNAm to lung function at Year 20 may not be significant in CARDIA.

In the current analysis, there is a 5-year gap between the exposure and the mediator. A reviewer raised the concern on treatment-induced-mediator-outcome confounding. The life-course smoking trajectories for the majority of individuals were relatively stable before age 40–45, which corresponds to the Year 10–15 of our study cohort [40]. Although DNA methylation is modifiable by smoking, it is still a relatively stable biomarker over time [41]. Short-term exposure-induced covariates within a 5-year gap (if any) are unlikely to produce biologically functional changes in DNA methylation for us to detect as mediators.

## Conclusion and remarks

In this paper we proposed an improved method HIMA2 for high dimensional mediation analysis, which was shown to have better performance than HIMA [3] by numerical studies. We applied HIMA2 to the identification and testing of the DNA methylation mediating effects in the CARDIA study. Our method is relatively simple to implement, and can be widely used in high-dimensional mediation analyses.

Our method can be extended in several directions. First, we will consider how to address the correlation among DNA methylation markers to improve the inferential results, as shown in the Simulation Studies that both HIMA and HIMA2 lose power for high correlation. Second, it is of interest to incorporate the interaction terms between the exposure and the mediators in our model, i.e., the high dimensional moderated mediation analysis. Third, there has been an increasing interest and development in longitudinal studies of DNA methylation. We can also consider repeated measures of DNA methylation markers as mediators in our future research.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04748-1.

**Additional file 1. Table S1:** Bias (MSE) for mediation effect estimates. **Table S2:** FDR at significance level 0.05. **Table S3:** Power at significance level 0.05.

**Acknowledgements**
None.

**Author contributions**
LL conceived and designed the study. CP and KX analyzed the data and wrote the manuscript. HZ, YZ, AQ, CZ and LF guided analyses, provided advice, and critically reviewed the manuscript. All authors read and approved the final manuscript.

**Funding**
This research was partly supported by NIH/NIA R21 AG 063370, R21 AG068955, and NIH/NCATS UL1 TR002345. The Coronary Artery Risk Development in Young Adults Study (CARDIA) is supported by contracts HHSN268201800003I, HHSN268201800004I, HHSN268201800005I, HHSN268201800006I, and HHSN268201800007I from the National Heart, Lung, and Blood Institute (NHLBI). This manuscript has been reviewed by CARDIA for scientific content. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Availability of data and materials**
R package, source code, and simulation study are available at https://github.com/joyfulstones/HIMA2.

## Declarations

**Ethical approval and consent to participate**
All CARDIA participants provided written informed consent, with institutional review board approval at each field center (the University of Alabama at Birmingham, Northwestern University, University of Minnesota, and Kaiser Permanente). All methods were performed in accordance with the relevant guidelines and regulations (for example- Declarations of Helsinki).

**Consent to publish**
Not applicable.

**Competing interests**
The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 27 September 2021 Accepted: 23 May 2022
Published online: 25 July 2022

**References**

1. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research – conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173–82.
2. MacKinnon DP. Introduction to statistical mediation analysis. New York: Erlbaum; 2008.
3. Zhang H, Zheng Y, Zhang Z, Gao T, Joyce B, Yoon G, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. Bioinformatics. 2016;32(20):3150–4.
4. Valeri L, Reese SL, Zhao S, Page CM, Nystad W, Coull BA, London SJ. Misclassified exposure in epigenetic mediation analyses. Does DNA methylation mediate effects of smoking on birthweight? Epigenomics. 2017;9(3):253–65.
5. Fang R, Yang H, Gao Y, Cao H, Goode EL, Cui Y. Gene-based mediation analysis in epigenetic studies. Brief Bioinform. 2020. https://doi.org/10.1093/bib/bbaa113.
6. Zhang J, Wei Z, Chen J. A distance-based approach for testing the mediation effect of the human microbiome. Bioinformatics. 2018;34(11):1875–83.
7. Sohn MB, Li H. Compositional mediation analysis for microbiome studies. Ann Appl Stat. 2019;13(1):661–81.
8. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics. 2017;19(2):121–36.
9. Zhao Y, Lindquist MA, Caffo BS. Sparse principal component based high-dimensional mediation analysis. Comput Stat Data Anal. 2020;142:106835.
10. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y. Testing mediation effects in high-dimensional epigenetic studies. Front Genet. 2019. https://doi.org/10.3389/fgene.2019.01195.
11. Derkach A, Pfeiffer RM, Chen TH, Sampson JN. High dimensional mediation analysis with latent variables. Biometrics. 2019;75(3):745–56.
12. Huang YT, Pan WC. Hypothesis test of mediation effect in causal mediation mode with high-dimensional continuous mediators. Biometrics. 2016;72(2):402–13.
13. Zhang, Q. High dimensional mediation analysis with applications to causal gene identification. bioRxiv. Doi: https://doi.org/10.1101/497826 (2019)
14. Djordjilović V, Page CM, Gran JM, Nøst TH, Sandanger TM, Veierød MB, Thoresen M. Global test for high-dimensional mediation: testing groups of potential mediators. Stat Med. 2019;38:3346–60.
15. Zhang H, Chen J, Li Z, Liu L. Testing for mediation effect with application to human microbiome data. Stat Biosci. 2019. https://doi.org/10.1007/s12561-019-09253-3.
16. Zhang H, Chen J, Feng Y, Wang C, Li H, Liu L. Mediation effect selection in high-dimensional and compositional microbiome data. Stat Med. 2021;40(4):885–96.
17. Wang C, Hu J, Blaser MJ, Li H. Estimating and testing the microbial causal mediation effect with high-dimensional and compositional microbiome data. Bioinformatics. 2020;36:347–55.
18. Liu Z, Shen J, Barfield R, Schwartz J, Baccarelli AA, Lin X. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. J Am Stat Assoc. 2021. https://doi.org/10.1080/01621459.2021.1914634.
19. Loh WW, Moerkerke B, Loeys T, Vansteelandt S. Non-linear mediation analysis with high-dimensional mediators whose causal structure is unknown. Biometrics. 2021. https://doi.org/10.1111/biom.13402.
20. Zhou RR, Wang L, Zhao SD. Estimation and inference for the indirect effect in high-dimensional linear mediation models. Biometrika. 2020;107(3):573–89.

21. Shi CA, Li L. Testing mediation effects using logic of Boolean matrices. J Am Stat Assoc. 2021. https://doi.org/10.1080/01621459.2021.1895177.
22. Dai JY, Stanford JL, LeBlanc M. A multiple-testing procedure for high-dimensional mediation hypotheses. J Am Stat Assoc. 2021. https://doi.org/10.1080/01621459.2020.1765785.
23. Friedman GD, Cutter GR, Donahue RP, Hughes GH, Hulley SB, Jacobs DR Jr, et al. CARDIA: study design, recruitment, and some characteristics of the examined subjects. J Clin Epidemiol. 1998;41(11):1105–16.
24. Tate PH, Bird AP. Effects of DNA methylation on DNA-binding proteins and gene expression. Curr Opin Genet Dev. 1993;3(2):226–31 (**PMID: 8504247**).
25. Fang EX, Ning Y, Liu H. Testing and confidence intervals for high dimensional proportional hazards models. J R Stat Soc Series B (Statistical Methodology). 2016;79(5):1415–37.
26. Tsai PC, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin Epigenet. 2018;10:126. https://doi.org/10.1186/s13148-018-0558-0.
27. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B. 2008;70:849–911.
28. Huang YT. Joint significance tests for mediation effects of socioeconomic adversity on adiposity via epigenetics. Ann Appl Stat. 2018;12(3):1535–57.
29. Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.
30. Gao Y, Yang H, Fang R, Zhang Y, Goode E, Cui Y. Testing mediation effects in high-dimensional epigenetic studies. Front Genet. 2019. https://doi.org/10.3389/fgene.2019.01195.
31 Xu Z, Niu L, Li L, Taylor JA. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. Nucleic Acids Res. 2016;44(3):e20 (**PMID: 26384415; PMCID: PMC4756845**).
32 Houseman EA, Accomando WP, Koestler DC, Christensen BC, Marsit CJ, Nelson HH, Wiencke JK, Kelsey KT. DNA methylation arrays as surrogate measures of cell mixture distribution. BMC Bioinform. 2012;13:86 (**PMCID: PMC3532182**).
33 Beane J, Sebastiani P, Liu G, Brody JS, Lenburg ME, Spira A. Reversible and permanent effects of tobacco smoke exposure on airway epithelial gene expression. Genome Biol. 2007;8(9):R201 (**PMID: 17894889; PMCID: PMC2375039**).
34. Gower AC, Steiling K, Brothers JF 2nd, Lenburg ME, Spira A. Transcriptomic studies of the airway field of injury associated with smoking-related lung disease. Proc Am Thorac Soc. 2011;8(2):173–9.
35. Lees JA, Ferwerda B, Kremer PHC, et al. Joint sequencing of human and pathogen genomes reveals the genetics of pneumococcal meningitis. Nat Commun. 2019;10:2176.
36. Vastrad B, Vastrad C, Tengli A. Bioinformatics analyses of significant genes, related pathways, and candidate diagnostic biomarkers and molecular targets in SARS-CoV-2/COVID-19. Gene Rep. 2020;21:100956.
37. Haarman MG, Kerstjens-Frederikse WS, Berger RMF. TBX4 variants and pulmonary diseases: getting out of the "Box." Curr Opin Pulm Med. 2020;26(3):277–84.
38. Xie T, Liang J, Liu N, et al. Transcription factor TBX4 regulates myofibroblast accumulation and lung fibrosis. J Clin Investig. 2016;126(8):3063–79.
39. Maurac A, Lardenois É, Eyries M, et al. T-box protein 4 mutation causing pulmonary arterial hypertension and lung disease. Eur Respir J. 2019;54:1900388.
40. Mathew AR, et al. Life-course smoking trajectories and risk for emphysema in middle age: the CARDIA lung study. Am J Respir Crit Care Med. 2019;199:237–40. https://doi.org/10.1164/rccm.201808-1568LE.
41. Tsai PC, et al. Smoking induces coordinated DNA methylation and gene expression changes in adipose tissue with consequences for metabolic health. Clin Epigene. 2018;10:26. https://doi.org/10.1186/s13148-018-0558-0.


## Publisher's Note