Supplementary Materials for

Methods for Mediation Analysis with High-Dimensional DNA Methylation Data: Possible Choices and Comparison

S1 Supporting Information

1. Assumptions for causal mediation analysis

With notation from equations (1) and (2), in order to interpret β_a as the direct effect, $\boldsymbol{\alpha_a^T \beta_m}$ as the global mediation effect, and $\beta_a + \boldsymbol{\alpha_a^T \beta_m}$ as the total effect, we require that:

1) there is no unmeasured confounding of the exposure-mediator association conditional on $\boldsymbol{C}$;

2) there is no unmeasured confounding of the exposure-outcome association conditional on $\boldsymbol{C}$.

3) there is no unmeasured confounding of the mediator-outcome association conditional on $\boldsymbol{C}$; and

4) the confounders of the mediator-outcome association are not affected by $\boldsymbol{A}$, which would make those confounders mediators themselves (1,2).

Our study also assumes there is no exposure-mediator interaction affecting the outcome, which methods for high-dimensional mediation analysis are not generally able to accommodate.

2. Tuning of variance-covariance matrix in simulations

As described in the main text, the residuals of the mediator models are sampled from a multivariate normal distribution with mean $\mathbf{0}_p$ and variance-covariance matrix $\boldsymbol{S}$. To obtain $\boldsymbol{S}$, we begin by computing the true variance-covariance matrix from the real data analysis with DNAm. This matrix is singular due to the fact that there are more CpG sites (2,000) being analyzed than observations (963). To make the matrix non-singular, we first standardize it so that each mediator has variance 1, then add a small penalty, *r*, to every element of the diagonal. Then we standardize a second time. This procedure is necessary because if we do not add such a penalty, the variance-covariance matrix remains non-invertible, so the residuals once sampled will be linearly dependent regardless of the chosen sample size. Moreover, since the size of *r* determines the degree of correlations between mediators—larger values resulting in weaker correlations—we can change the correlations between simulation settings by tweaking *r*. For the baseline correlation level, we set *r* to be 1, which causes the correlations to range from -0.37 to 0.49. For the high-correlation setting, we set *r* to be 0.1, letting the correlations range from -0.68 to 0.89.

3. Method application details

One-at-a-time

We assess the mediators "one-at-a-time" using the standard mediation models proposed by Baron and Kenny (1986) (3). These are analogous to the models established in the main text (i.e., models 1 and 2) except that they treat each mediator separately. We assess the significance of each mediator using the max-P test (or joint significance test), in which the maximum of the exposure-mediator and mediator-outcome association p-values is tested against the significance level (4). In the simulated data analysis, we identify active mediators by thresholding these max p-values so that at most 10% of the "significant" mediators are not true mediators, ensuring that the false discovery proportion (FDP) on that dataset is below 0.10 and the false discovery *rate* (FDR) across all datasets is also below 0.10. In the real data analysis, we use a linear mixed model instead of linear regression so that we can include the appropriate confounding variables as random effects; in particular, age, sex, race, and the estimated proportions of residual non-monocytes are adjusted for as fixed effects and methylation chip and position as random effects (to address potential batch effects). We fit this model with the "lmerTest" package in R. (The mediation model from this set up is the same model used to screen the CpG sites from 402,339 down to 2,000.) To identify CpG sites involved in mediation, we take the max-P values from these models and adjust them using the “qvalue” function from the “qvalue” package, then compare them to a signficance level of 10% to control the FDR.

HIMA

We apply HIMA directly using the “hima” function from the “HIMA” package in R. On the simulated datasets, we apply sure independence screening (5) as recommended by the HIMA authors (6) to reduce the set of mediators to the $n/\log(n)$ mediators mostly strong associated with Y, adjusting for A (screening based on the outcome model). The number $n/\log(n)$ is chosen to encourage dimension reduction while maintaining the accuracy of the screening procedure (5). For FDR correction, we choose a p-value threshold for each dataset to ensure that the false discovery proportion for that dataset is below 0.10.

On the DNAm data from MESA, so that the sure indepdent screening matches the original screening used to arrive at 2,000 CpG sites, we screen on the association of each CpG site with the exposure, low education, using the same linear mixed model used in the mediator model of the one-at-a-time approach. (The HIMA and HDMA authors suggest screening based on either the outcome model or the mediator model is acceptable (6,7)). We identify CpGs as noteworthy based on whether their estimated mediation contribution is not zero. Although HIMA does produce p-values (which we use for thresholding in the simulated data analysis), the p-values are based on the subsequent fitting of an ordinary linear regression after the penalized model has performed feature selection; hence, they can be expected to be overconfident. It is for this reason that throughout our DNAm analysis, we avoid commenting on the “statistical significance” of the mediation contributions and focus only on identifying which ones are “noteworthy.”

HDMA

We apply HDMA using the “hdma” function provided by Gao et al. at their repository (https://github.com/YuzhaoGao/High-dimensional-mediation-analysis-R/blob/master/HDMA.R). Application of the “hdma” function is similar to application of the “hima” function from the “HIMA” package, and we apply HDMA identically to how we apply HIMA as described above.

MedFix

Although code for MedFix is provided in the GitHub repository located at https://github.com/QiZhangStat/highMed, it is designed for the setting where both A and M are high-dimensional (the primary setting for which MedFix is intended). We provide code for implementing MedFix for a single exposure in our package "hdmed" (https://cran.r-project.org/package=hdmed). Although MedFix does not explicitly involve sure independence screening like HIMA or HDMA, we elect to use screening for MedFix as well so that the penalized regression methods (which are very similar) are all applied in the same systematic fashion. Other details on how MedFix was implemented are comparable to HIMA as described above.

PCMA

We apply PCMA using the "mcma_PCA" function in the "SPCMA" R package (https://github.com/zhaoyi1026/spcma). For both the simulated data analysis and DNAm data analysis, we set the number of principal components to be 100. Although it may be preferable in some cases to choose this number to be larger, capturing more of the variance of $\boldsymbol{M}$, it is better for the sake of our comparison to use the same number of principal components in both PCMA and SPCMA, and applying SPCMA with more than 100 principal components would be extremely computationally costly. Further, in the simulated data analysis where SPCMA was not used at all, choosing the maximal number of principal components (the minimum of n and p) resulted in extreme variability in the estimated total indirect effect and did not cause the method to perform better than it did with 100.

SPCMA

We apply SPCMA using the "spcma" function from the "spcma" package (https://github.com/zhaoyi1026/spcma). Analogously to PCMA, we set the number of sparse principal components to be 100. To create sparsity in the principal component loading vectors, SPCMA uses the fused LASSO penalty (8), which in addition to the L1 penalty used by the regular LASSO, penalizes the difference in coefficient effects between adjacent variables, inducing smoothness. The parameter γ represents the ratio of the L1 penalty to the fusion penalty, and can be chosen in advance. We set γ to be 2 so that the L1 penalty is emphasized, as "adjacent" CpG sites could still be far apart in the genome and should not be expected to have similar effects, making the fusion penalty unimportant. However, different choices of γ did not appear to dramatically change our results. We test for the significance of the 100 transformed mediators using the bootstrapping method proposed by the SPCMA authors, with 100 bootstrap samples and bias-corrected confidence intervals (9).

Pathway LASSO

Code for implementing pathway LASSO is provided by the authors in the GitHub repository located at https://github.com/zhaoyi1026/PathwayLasso. Like MedFix, pathway LASSO does not explicitly involve pre-screening the mediators, but we still conduct the pre-screening procedure from HIMA and HDMA so that the penalized regression methods HIMA, HDMA, MedFix, and pathway LASSO are more comparable. This is also beneficial computationally, as it is slow to apply pathway LASSO directly to data with 2,000 mediators and 2,500 or 1,000 observations. Another aspect of pathway LASSO implementation is selecting the tuning parameters. These include ω, which controls a LASSO-like penalty on each $(\boldsymbol{\alpha_a})_j$ and $(\boldsymbol{\beta_m})_j$, ϕ, a convexity parameter, and λ, which controls a complex penalty function including the product terms $(\boldsymbol{\alpha_a})_j(\boldsymbol{\beta_m})_j$. For our analysis, ϕ is fixed at 2, as it is in fMRI study presented by Zhao and Luo (2022), who show that pathway LASSO is not sensitive to the choice of this parameter (10). For the other parameters, we fix the ratio of ω to λ to be 1 (i.e., forcing the parameters to be equal), as this ratio performed best in the Zhao and Luo's simulation study. We then attempt pathway LASSO using 45 different values of λ. In the simulated data analysis, we choose the optimal value of λ based on the observed false discovery proportion, selecting the smallest λ for which fewer than 10% of the selected mediators are true mediators in that simulated dataset. This is necessary because pathway LASSO does not provide a method to test the statistical significance of the effects using p-values, like in HDMA, HIMA, and MedFix, preventing us from using p-values for thresholding. In the DNAm analysis, we use the variable selection stability criterion suggested by the authors (11) with the code provided in the GitHub package. Like in the other penalized regression methods, CpG sites are considered noteworthy if their estimated mediation contribution is not zero.

BSLMM

We apply BSLMM using the "bama" function from the "bama" package in R. Application of BSLMM depends on several parameters: *lm0*, *lm1*, *lma1*, *l*, and *k*. These are, respectively, the scale parameter for the inverse-gamma prior for the small-variance $(\boldsymbol{\alpha_a})_j$ and $(\boldsymbol{\beta_m})_j$, the scale parameter for the inverse-gamma prior for the large-variance $(\boldsymbol{\beta_m})_j$, the scale parameter for the inverse-gamma prior for the large-variance $(\alpha_a)_j$, the scale parameter for the inverse-gamma prior on the other coefficient variances, and the shape parameter for the inverse-gamma priors. In applying BSLMM, we fix *k* at 1 (the default setting) so that the prior mean of each variance is equal to the scale parameter. We also fix *l* at 1 (the default setting). In both the DNAm data analysis and simulated data analysis, we choose *lm1* by taking the variance of the absolute largest 10% of the marginal $(\beta_m)_j$ coefficients (i.e., the coefficient from the one-at-a-time method), and choose *lma1* similarly except with the marginal $(\alpha_a)_j$ coefficients. (When working on simulated data, we know the true variance of the coefficients exactly, but choosing these parameters based on the known truth is not possible in practice and therefore should be avoided in simulations to keep them fair. Part of the difficulty in applying BSLMM on real data is choosing these parameters appropriately, and the results of BSLMM tend to be sensitive to this choice).

HDMM

We apply HDMM using the "PDM_1" function from the "PDM" R package (https://github.com/oliverychen/PDM). The "PDM_1" function computes weights for the first principal direction of mediation (PDM), which are then used to linearly combine the set of mediators into a single, transformed latent mediator. We analyze the transformed mediator using the "mediation" R package, with 2,000 Monte Carlo draws used for the quasi-Bayesian confidence intervals. HDMM cannot directly incorporate settings where *p* is greater than *n* when computing its PDMs, so we repeat our sure independence screening procedure from HIMA, HDMA, MedFix, and pathway LASSO prior to the analysis. (Note that HDMM can be applied when *p* is greater than *n* by using population value decomposition (12). However, population value decomposition is designed for longitudinal settings where each observation contains multiple measurements representing different time points. Our attempts to apply population value decomposition to our data were unsuccessful for this reason.)

HILMA

We apply HILMA using the "hilma" function from the "freebird" R package. We set the tuning method to "uniform" rather than "AIC" as recommended by the HILMA authors, and we standardize the data prior to analysis. (For the simulated data analysis, we multiply the resulting total indirect effect by the standard deviation of *Y*, and divide by the standard deviation of *A*, to project the estimate back to the scale of the original data).

PMED

We apply PMED with code from the paper's supplement, which can be found at https://doi.org/10.1080/01621459.2022.2053136. The authors provide a suite of functions that can be used to implement the results from their study, as well as a wrapper function called "hdMediation" that consolidates their statistical procedure into a single routine. The penalized regression used by PMED depends on a tuning parameter λ, and sequence of candidate parameters must be supplied to the "hdMediation" function. For a sequence, we choose 50 numbers ranging from 0.000001 to 30 that are equally spaced on a logarithmic scale. The function also accepts a matrix of confounding variables: For the simulated data analysis, we supply an n by 1 matrix of 1's to this input (i.e., an intercept term), and for the DNAm analysis, we supply covariates in the same fashion as with BSLMM.

4. MESA study and data processing

Our data were provided by the Multi-Ethnic Study of Atherosclerosis (MESA), a United States population-based longitudinal study on the risk factors and progression of subclinical cardiovascular disease (13). MESA recruitment began in July 2000 and lasted until August 2002, during which period 6,814 participants were recruited from study sites in Forsyth County, North Carolina; New York, New York; Baltimore County, Maryland; St. Paul, Minnesota; and Chicago, Illinois. Ages at recruitment ranged from 45 to 84 years. Multiple examinations since the beginning of the study captured data on clinical information, socio-demographic traits, lifestyle and behavioral characteristics, and other factors. A random subsample of 1,264 MESA participants were selected between April 2010 and February 2012 to have their DNAm measured using the Illumina Infinium HumanMethylation450 Beadchip on purified monocytes. Quality control filters reduced the number of CpG sites from 484,882 to 402,339; in particular, sites were removed if they had "detected" methylation levels in less than 90% of MESA samples at a p-value cutoff of 0.05, were within 10 base pairs of a single nucleotide polymorphism (SNP) based on Illumina annotation, had unreliable probes (i.e., had SNPs with minor allele frequency greater than 0.05 within 2 base pairs or cross reactive probes, recommended by DMRcate (14)), or overlapped with a repetitive element or region; while probes on sex chromosomes, SNPs, and other non-CpG targeting probes were not considered. The raw methylation measurements were transformed into M-values by taking the log-2 ratio of the methylated to unmethylated probe intensities. Further details are provided by Liu et al. (2013) (15).

Our outcome variable was HbA1c measured at Exam 5, which was standardized prior to the analysis. This provides a measure of the average three-month blood sugar level. For our exposure variable, low adult socioeconomic status, we use a binary indicator variable representing the lack of a 4-year college degree (1: less than a 4-year college degree, 0: otherwise). Since 402,339 CpG sites is too many to include in an analysis with only 963 samples, we screened CpG sites in advance to reduce that number further, including only, at most, the 2,000 CpG sites most strongly associated with the exposure. This association was measured for each site by the p-value of the education coefficient from a linear mixed model in which methylation was regressed onto education level (the binary indicator), age, sex, race, and the estimated proportions of residual non-monocytes as fixed effects, and methylation chip and position as random effects. This model is equivalent to the one-at-a-time mediation mediator model described above. For the methods which require or recommend additional screening, we do so with the same model, described in the main text and in the section above.

5. Scalability comparison

We compared the scalability of the methods by assessing their runtime on simulated datasets of two sizes: either 100 observations and 200 mediators or 1,000 observations 2,000 mediators. This was done on a single core of an Intel(R) Xeon(R) Gold 6242R CPU @ 3.10GHz processor. We attempted each method 30 times and report the mean and interquartile range of the runtimes (Table A). Since SCPMA and BSLMM tend to be time-consuming, we approximated their run times by downscaling the appropriate parameters: In particular, since the desired number of principal components in SPCMA is 100, we use only 2 principal components and scale the computing time by 50; and since the desired number of posterior samples in BSLMM is 30,000, we draw only 750 samples and scale the result by 40. Ad hoc experimentation confirmed the run time of these methods were approximately linear with respect to these inputs. We also report the RAM used in a single run (Table B).

Table A. Computation time comparison for high-dimensional mediation analysis methods.

	n = 100, *p* = 200		*n* = 1,000, *p* = 2,000	
Method	Mean	Interquartile Range	Mean	Interquartile Range
BSLMM	39.17s	(38.84s - 39.54s)	40.14m	(39.74m - 40.34m)
HDMA	1.40s	(1.37s - 1.40s)	29.76s	(29.55s - 29.92s)
HDMM	24.85s	(24.80s - 24.89s)	12.36m	(12.33m - 12.37m)
HILMA	24.42s	(24.13s - 24.63s)	40.85m	(38.22m - 40.65m)
HIMA	0.25s	(0.25s - 0.25s)	3.55s	(3.47s - 3.62s)
MEDFIX	0.61s	(0.60s - 0.61s)	7.33s	(7.22s - 7.42s)
PCMA	2.77s	(2.74s - 2.79s)	58.97s	(58.08s - 59.35s)
PLASSO	18.71m	(18.19m - 19.23m)	192.62m	(188.10m - 195.83m)
SPCMA	16.05m	(15.94m - 16.04m)	842.54m	(827.26m - 855.21m)
PMED	0.57s	(0.55s - 0.65s)	14.84s	(14.54s - 15.01s)

Methods were run 30 times each on a single core of an Intel(R) Xeon(R) Gold 6242R CPU @ 3.10GHz processor.

Table B. Memory usage of mediation methods by dataset dimensions.

Method	RAM Usage (Megabytes)	
	n = 100, *p* = 200	*n* = 1,000, *p* = 2,000
BSLMM	114	484
HDMA	238	342
HIMA	233	306
MEDFIX	192	286
PCMA	315	638
PLASSO	72	115
SPCMA	378	1,222
HDMM	359	687
HILMA	106	603
PMED	206	281

References

1. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. Epidemiol Method. 2014 Jan;2(1):95–115.

2. Pearl J. Direct and Indirect Effects. In: Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.; 2001. p. 411–420. (UAI'01).

3. 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.

4. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychol Methods. 2002 Mar;7(1):83–104.

5. Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc [Internet]. 2008 Nov 1;70(5):849–911. Available from: https://doi.org/10.1111/j.1467-9868.2008.00674.x

6. 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 [Internet]. 2016 Oct 15;32(20):3150–4. Available from: https://doi.org/10.1093/bioinformatics/btw351

7. Gao Y, Yang H, Fang R, Zhang Y, Goode EL, Cui Y. Testing Mediation Effects in High-Dimensional Epigenetic Studies. Front Genet. 2019;10:1195.

8. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso. J R Stat Soc Ser B (Statistical Methodol [Internet]. 2005 Feb 1;67(1):91–108. Available from: https://doi.org/10.1111/j.1467-9868.2005.00490.x

9. Zhao Y, Lindquist MA, Caffo BS. Sparse principal component based high-dimensional mediation analysis. Comput Stat Data Anal [Internet]. 2020;142:106835. Available from: https://www.sciencedirect.com/science/article/pii/S0167947319301902

10. Zhao Y, Luo X. Pathway LASSO: pathway estimation and selection with high-dimensional mediators. Stat Interface. 2022;15(1):39–50.

11. Sun W, Wang J, Fang Y. Consistent selection of tuning parameters via variable selection stability. J

Mach Learn Res. 2013;14(1):3419–40.

12. Chén OY, Crainiceanu C, Ogburn EL, Caffo BS, Wager TD, Lindquist MA. High-dimensional multivariate mediation with application to neuroimaging data. Biostatistics. 2018;19(2):121–36.

13. Bild DE, Bluemke DA, Burke GL, Detrano R, Diez Roux A V, Folsom AR, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. Am J Epidemiol. 2002 Nov;156(9):871–81.

14. Peters TJ, Buckley MJ, Statham AL, Pidsley R, Samaras K, V Lord R, et al. De novo identification of differentially methylated regions in the human genome. Epigenetics Chromatin. 2015;8(1):6.

15. Liu Y, Ding J, Reynolds LM, Lohman K, Register TC, De La Fuente A, et al. Methylomics of gene expression in human monocytes. Hum Mol Genet. 2013 Dec;22(24):5065–74.