Biometrics XX, 1–25 DOI: XX
XX 20xx

Dissecting the colocalized GWAS and eQTLs with mediation analysis for high dimensional exposures and confounders

Qi Zhang*

Department of Mathematics and Statistics, University of New Hampshire, Durham, NH 03824, USA

**email:* qi.zhang2@unh.edu

and

Zhikai Yang

Complex Biosystems Program and Department of Agronomy and Horticulture

University of Nebraska-Lincoln, Lincoln, NE 68583, USA

and

Jinliang Yang

Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Lincoln, NE 68583, USA

Summary:

To leverage the advancements in GWAS and QTL mapping for traits and molecular phenotypes to gain mechanistic understanding of the genetic regulation, biological researchers often investigate the eQTLs that colocalize with QTL or GWAS peaks. Our research is inspired by two such studies. One is in maize that aims to identify the causal SNPs that are responsible for the phenotypic variation and whose effects can be explained by their effects at the transcriptomic level. The other study in mouse focuses on uncovering the cis-driver genes that lead to phenotypic changes through regulating trans-regulated genes. Both studies can be formulated as mediation problems with potentially high-dimensional exposures, confounders and mediators that seek to estimate the overall indirect effect for each exposure. In this paper, we propose MedDiC, a novel procedure to estimate the overall indirect effect based on difference-in-coefficients approach. Our simulation studies find that MedDiC offers valid inference for the indirect effect with higher power, shorter confidence intervals and faster computing time than the competing methods. We apply MedDiC to the two aforementioned motivating datasets, and find the MedDiC yields reproducible outputs across the analysis of closely related traits, and the results are supported by external biological evidence.

Key words: Debiased estimator, GWAS, High-dimensional regression, Integrative Omics, Mediation Analysis

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

1.1 *Integrative analysis of multi-omics data for dissecting the colocalized GWAS and eQTL*

In the last few decades, genome-wide association studies (GWAS) and quantitative trait loci (QTL) mapping have been widely used to identify the genetic variants (e.g., single nucleotide polymorphisms, or SNPs for short) associated with complex traits, including molecular traits such as metabolites and gene expressions. While these studies have provided valuable information on the genetic architecture of these phenotypes and molecular traits, they do not generate any mechanistic hypotheses for these associations, i.e., how a genetic variant affects the phenotype and what genes are involved.

To gain mechanistic understanding of the genetic effects, a common research paradigm is to investigate whether eQTL peaks colocalize with GWAS peaks for a trait of interest. In such cases, some of the genes with eQTL mapped in the region are likely to play a role in regulating the phenotype. Along this line, many works have jointly analyzed the genotype, conventional phenotype, and intermediate molecular phenotypes such as gene expressions to gain further insights into the genetic effects. For example, Tu et al. (2012) focused on the genes with overlapping eQTLs with insulin QTLs, and studied the potential transcriptional regulation mechanism among them. Pang et al. (2019) used regression models to identify the genes associated with the kernel length of maize. However, these studies do not quantify the extent to which the genetic effect on a phenotype could be explained by the expressions of the genes under consideration. Such questions could be addressed using mediation analysis.

1.2 *High-dimensional mediation analysis and its applications in multi-omics data analysis*

Causal mediation analysis partitions the total effect (TE) of the exposures ($\mathbf{Z}$, Figure 1a) on the outcome ($\mathbf{Y}$) into two parts, the indirect effect (IE, or mediation effect) via measured mediating variables ($\mathbf{M}$), and the direct effect (DE) that is not through the given mediators. Since the biologists are interested in explaining the causal mechanism using the

known mediators, investigating the indirect effect (IE) has been the primary interest in most mediation analysis in genomic studies. On the other hand, direct effect (DE) can only be interpreted as the effect that is not mediated by the given set of mediators, and their mechanism may be explained using the other potential mediating variables that are not included in the analysis. In studies that integrate genetic markers, phenotype and gene expressions, the exposure is usually the genetic marker, the outcome is the phenotype such as insulin level, and the mediator is the expressions of a set of genes. Then, the indirect effect is the effect of the genetic marker on the phenotype that is through the changes in the measured gene expressions, and the direct effect is the effect of the genetic marker on the phenotype that is potentially mediated by other unmeasured mediating variables such as the expressions of other genes. The underlying biological assumption is that the gene expressions under consideration contribute to the phenotype rather than being influenced by it.

[Figure 1 about here.]

The classical mediation analysis focuses on the mediation effect of one exposure through one mediator (Figure 1a). In the last decade, mediation analysis for multivariate low-dimensional mediators has been extensively studied in the literature (VanderWeele and Vansteelandt, 2014), and high-dimensional mediation models have also become an active research area, largely driven by scientific challenges in genomics and biomedical imaging.

Most high-dimensional mediation studies focus on mediator selection, which involves identifying mediators with significant mediation effects. This type of research is usually driven by environmental epigenetic studies, where the exposure is the environmental factor, the high-dimensional mediators are epigenetic marks (e.g., DNA-methylation), and the response is a clinical outcome. Commonly, these methods assume that the mediators are not causally related (as shown in Figure 1b). Zhang et al. (2016) proposed HIMA, a high-dimensional mediator selection method for a single exposure. It is based on intersection-union tests

after screening and penalized regression. Many novel mediator selection methods have been developed for specific problems, such as Fang et al. (2021) for testing the significance of a group of mediators with potentially nonlinear mediation effect, Xue et al. (2022) for heterogeneous direct and indirect effects across subgroups, and Sohn et al. (2019) for compositional microbiome mediators. In multi-omics studies, the exposures can be also multivariate or even high-dimensional. For such cases, mediator selection methods have been developed based on the aggregated indirect effects across all exposures for each mediator (Zhong et al., 2019; Zhang, 2021). A related group of works are motivated by brain imaging studies. In these studies, the exposure is usually an environmental/experimental factor or omics variables, and the mediators are usually imaging features that may become more interpretable after some transformations. Chén et al. (2017) introduced PCA-based approaches for high-dimensional mediation analysis that essentially assume the same diagram among the mediators in Figure 1b after a data-driven linear transformation of the mediators. Zhao (2022) decomposed the multivariate exposures and multivariate mediators to orthogonal components and estimate their overall indirect and direct effects.

Zhou et al. (2020); Guo et al. (2022a,b); Huang et al. (2023) studied a different high-dimensional mediation problem. They proposed procedures for the estimation and inference of the overall indirect and direct effects of each exposure on the outcome through all the mediators (Figure 1c). Zhou et al. (2020) analyzed a multi-omics dataset in which the mediators are gene expressions, and the exposure is either one genetic variant or the expression of a driver gene. Huang et al. (2023) used gene expressions as the exposures and proteomics variables. Guo et al. (2022a) was motivated by an environmental epigenetic study on Childhood trauma.

In this paper, we investigate the estimation and inference of the overall indirect effect through high-dimensional mediators for high-dimensional exposures (Figure 1c). Our re-

search is motivated by the following two multi-omics studies that aim to dissect the colocalization of GWAS/QTL and eQTL peaks.

1.3 *Genome-wide mediation analysis of maize data*

In Yang et al. (2022), the authors performed a Genome-Wide Mediation Analysis (GWMA) in maize using MedFix (Zhang, 2021). The primary aim of the study was to investigate the causal chain from genotype to phenotype partially mediated by intermediate molecular processes (i.e., gene expression). The exposures consisted of genotypic data from the Maize Association Panel (MAP), representing the global diversity of public maize inbred lines (Flint-Garcia et al., 2005). The mediators were transcriptomic data from seven tissues (Kremling et al., 2018), and the responses included the metabolomics data and the best linear unbiased prediction (BLUP) for 40 agronomic traits.

After pre-processing, the analysis involved 0.77 million SNPs as potential exposures and the expressions of approximately 12,000 genes as potential mediators. The numbers of subjects with the complete data for each analysis varied by tissue and phenotype, but in all cases, it was slightly less than 300. The authors identified many SNPs with indirect effect (iSNP), SNPs with direct effect (dSNPs) and the mediating genes, and discussed their biological meanings such as pleiotropic mediation and the gene functions.

However, as discussed in their paper, one limitation of this analysis is the low power for identifying the mediating genes and consequently the iSNPs. The relatively small sample size is a contributing factor. Another hypothesis is that some iSNPs may have large overall indirect effect, which is distributed among many mediators, each having a small effect. As a result, some mediators and the iSNPs may remain undetected. To address this issue and to identify iSNPs with higher power, we need a method that estimates the overall indirect effect for each SNP through all potential mediators (as illustrated in Figure 2a).

[Figure 2 about here.]

1.4 *Identifying the cis-drivers of the trans-regulated genes in Islet using a mice f2 cross data for diabetes study*

We studied a comprehensive dataset from a mice f2 cross for diabetes study (Tu et al., 2012; Tian et al., 2016). It contains 2,057 genetic markers, more than 40k transcripts, and multiple phenotypes such as insulin levels. Their research generally aims to identify potential driver genes that regulate the expressions of a broad range of genes and are associated with diabetes. Tu et al. (2012) studied the genes whose eQTLs overlap with the insulin QTLs on two regions of chr2 and chr19. They prioritized these genes based on a gene network model and identified APP as an important gene of the network. They then showed that its expression was negatively correlated with insulin secretion. Tian et al. (2016) identified an islet specific eQTL hotspot at around 75cM on chr2 that regulated a wide range of genes, including both cis- and trans- regulation. Cis-genes are genes whose genomic location are close to the genetic region that they are associated with, while trans-genes are those far away that still have an eQTL association in the region. An eQTL hotspot generally regulates many trans-genes across the genome, which are believed to be co-regulated by common regulator genes such as transcription factors, also cis-regulated at this locus (Pierce et al., 2014). If there is a QTL in the same region, it is possible that the cis-regulator gene influences the phenotype through regulating the trans-genes. Using this data, Keller et al. (2016) narrowed their attention to 129 genes with human homologues associated with T2D and identified Nfatc2, a transcription factor as a candidate cis- regulator of most of these genes.

However, none of their studies directly quantified the effects of the path from the cis-regulator to the trans-genes with shared eQTL and then to the phenotype. High-dimensional mediation can bridge the gap by identifying the driver genes through the inference of overall indirect effects for each exposure (cis-regulated genes; as illustrated in Figure 2b).

1.5 *Contributions and Structure of this paper*

In both examples above, the potential exposures and/or confounders can be high-dimensional. Motivated by this observation, we investigate the estimation and inference of the overall indirect effect for high-dimensional exposures through high-dimensional mediators. Our inference target is similar to that of Zhou et al. (2020); Guo et al. (2022b,a); Huang et al. (2023), but we take a completely different perspective. We propose a novel procedure called **Med**iation analysis via **D**ifference **i**n **C**oefficients (**MedDiC**) based on the difference-in-coefficients approach for mediation (MacKinnon et al., 2002; McGuigan and Langholtz, 1988).

The remaining of the paper are organized as the following. In Section 2, we present the model setup for our high-dimensional mediation problem, and derive the **MedDiC** estimator and its asymptotic distribution. We then present our simulation results in Section 3, followed by the analysis of the maize dataset and the mice dataset discussed earlier in Section 4. Finally, we conclude with discussions and final remarks in Section 5. We remark that the proposed model works in the same way when the exposures and/or the confounders are high-dimensional. For the remaining of the manuscript, we focus our discussions on the setting with high-dimensional exposures without loss of generality.

In this paper, for a matrix A, we use $A_{\mathcal{I},\mathcal{J}}$ to denote its submatrix such that the row and column IDs are restricted to sets $\mathcal{I}$ and $\mathcal{J}$ respectively. When $\mathcal{I} = \{i\}$ and $\mathcal{J} = \{j\}$, we simply write it as A_{ij}. We use $A_{i,:}$ to represent the ith row of A, and $A_{:,j}$ its jth column.

2. Mediation analysis with high-dimensional exposures, confounders, and mediators

For high-dimensional mediation problems, we consider a sample of n subjects from the population, where Y is the length n response vector, Z the $n \times q$ exposure matrix, X the $n \times s$ confounder matrix that includes the intercept, and M is the $n \times p$ mediator matrix.

We assume the following linear mediation model

$$Y = X\tilde{\alpha} + Z\tilde{\beta} + \tilde{\epsilon} \tag{1}$$

$$Y = X\alpha + Z\beta + M\gamma + \epsilon \tag{2}$$

$$M = XA + ZB + \eta \tag{3}$$

where β and $\tilde{\beta}$ are vectors of length q, α and $\tilde{\alpha}$ are vectors of length s, γ is a vector of length p, A and B are $s \times p$ and $q \times p$ matrices, respectively, $\tilde{\epsilon}$, ϵ are length n noise vectors, and η is an $n \times p$ noise matrix. p, q and s are allowed to be larger than n. If the linear model assumption is true, there are

$$\tilde{\alpha} = \alpha + A\gamma, \quad \tilde{\beta} = \beta + B\gamma, \quad \textit{and} \quad \tilde{\epsilon} = \epsilon + \eta\gamma$$

Hence only either (1)-(2) or (2)-(3) are needed for mediation analysis.

The literature of mediation analysis has been mainly concerned with the estimation and inference of the indirect effect, and the two approaches to quantify the indirect effect are the product-of-coefficients and difference-in-coefficients methods. The product-of-coefficients approach focuses on equations (2)-(3), and the parameter representing the indirect effect is $B\gamma$, the product of two coefficients. The difference-in-coefficients approach uses equations (1)-(2), and the indirect effect is represented by the difference between two coefficients, $\tilde{\beta} - \beta$. $B\gamma$ and $\tilde{\beta} - \beta$ are identical for the aforementioned linear mediation models. When the coefficients are estimated using ordinary least squares, the two estimators $\hat{\tilde{\beta}} - \hat{\beta}$ and $\hat{B}\hat{\gamma}$ are also identical (See Web Appendix A for the derivations). However, in more general settings, the two approaches target distinct causal quantities (Pearl, 2012), and which is preferable has been intensively discussed and evaluated in the literature without general consensus (Imai et al., 2010; MacKinnon et al., 2002). High dimensional mediator selection methods usually use the product-of-coefficients approach, because they need to evaluate the links $Z \to M_j$ and $M_j \to Y$ for each mediator M_j. Zhou et al. (2020) also follows the product-of-coefficients approach to estimate and perform inference of the overall mediation

effect. They derived a debiased estimator for $B\gamma$ by plugging in B with its ordinary least square estimator $(Z^TZ)^{-1}(Z^TM)$, and then deriving a debiased estimator of $(Z^TM)\gamma$. Since the ordinary least squares estimator is used, it cannot be directly applied to the setting when either the exposures or the confounders are high-dimensional.

For high-dimensional exposures, we adopt the difference-in-coefficients approach for estimation and inference of the overall indirect effect. Our proposed procedure, called **MedDiC** provides a debiased estimator for the indirect effect, and the standard error is easy to estimate from the data. We note that Guo et al. (2022a) has a similar model setup for the low-dimensional exposures. However, their estimator relied on the partially penalized least squares with no penalty on the direct effect, a pre-selected set of low-dimensional mediators, and the least square estimator of the total effect. Consequently, it cannot be directly applied to high-dimensional exposures. A computationally intensive alternative to developing debiased estimators is to bootstrap the penalized regression models for inference. Huang et al. (2023) presented such a framework using the product-of-coefficients approach.

The next sections detail the derivation of **MedDiC**, and then the implementation.

2.1 *Estimating the indirect effect using the difference of debiased lasso estimators*

We consider the models (1)-(2), and rewrite them as the following

$$Y = Z^\star\tilde{\beta}^\star + \tilde{\epsilon} \tag{4}$$

$$Y = Z^\star\beta^\star + M\gamma + \epsilon \tag{5}$$

where $Z^\star = (X, Z)$, $\tilde{\beta}^\star = (\tilde{\alpha}^T, \tilde{\beta}^T)^T$, $\beta^\star = (\alpha^T, \beta^T)^T$. Thus the exposures are represented by the $s+1$ to $s+q$ columns of $Z^\star$. For the rest of this section, we refer to $Z^\star$ as the exposure matrix, because one may consider a confounder as an exposure whose effect is of no interest. We further assume that for $i = 1, \ldots, n$, $\tilde{\epsilon}_i \overset{iid}{\sim} N(0, \sigma^2_{\tilde{\epsilon}})$ and $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2_{\epsilon})$.

When the mediators and exposures are both potentially high-dimensional, we advocate the

use of difference-in-coefficients approach for mediation. In this framework, the parameter for the overall indirect effect for each exposure is $\tilde{\beta}^\star - \beta^\star$, and we propose to estimate and perform statistical inference for them using $\hat{\tilde{\beta}}^\star - \hat{\beta}^\star$ and its asymptotic distribution, where $\hat{\tilde{\beta}}^\star$ and $\hat{\beta}^\star$ are based on the debiased lasso (Zhang and Zhang, 2014). In the following, we derive this estimator and show that it is asymptotically normal.

Based on the debiased lasso (Zhang and Zhang, 2014), a debiased estimator of $\tilde{\beta}^\star$ is

$$\hat{\tilde{\beta}}^\star = \tilde{R}^T Y - \tilde{W}\hat{\tilde{\beta}}^{\star,init} \tag{6}$$

where $\hat{\tilde{\beta}}^{\star,init}$ is an initial estimator, and $\tilde{R}$ and $\tilde{W}$ are calculated as below. Let $\tilde{e}_j$ be the residual vector after regressing $Z^\star_{:,j}$, the jth column of $Z^\star$ on the other columns of $Z^\star$. Then $\tilde{R}_{:,j} = \frac{\tilde{e}_j}{Z^{\star\,T}_{:,j}\tilde{e}_j}$, and $\tilde{W} = \{\tilde{W}_{jk}\}_{(s+q)\times(s+q)}$ where $\tilde{W}_{jj} = 0$ and $\tilde{W}_{jk} = (Z^\star_{:,k})^T\tilde{R}_{:,j}$ for $j \neq k$.

Plugging in (4) to (6) gives

$$\hat{\tilde{\beta}}^\star = \tilde{\beta}^\star + \tilde{R}^T\tilde{\epsilon} + \tilde{W}(\tilde{\beta}^\star - \hat{\tilde{\beta}}^{\star,init}) \tag{7}$$

Undercertain regularity conditions (Theorem 1 of Zhang and Zhang 2014), the last term in (7) is negligible and there is

$$\sqrt{n}(\hat{\tilde{\beta}}^\star_j - \tilde{\beta}^\star_j) \approx N(0, n\sigma^2_{\tilde{\epsilon}}(\tilde{R}_{:,j})^T\tilde{R}_{:,j})$$

Similarly, a debiased estimator of $((\beta^\star)^T, \gamma^T)^T$ is

$$\begin{pmatrix} \hat{\beta}^\star \\ \hat{\gamma} \end{pmatrix} = \begin{pmatrix} R_Z^T Y - W_{ZZ}\hat{\beta}^{\star,init} - W_{ZM}\hat{\gamma}^{init} \\ R_M^T Y - W_{MZ}\hat{\beta}^{\star,init} - W_{MM}\hat{\gamma}^{init} \end{pmatrix} \tag{8}$$

where $\hat{\beta}^{\star,init}$ and $\hat{\gamma}^{init}$ are initial estimators, and $R = (R_Z, R_M)$ and $W = \{w_{jk}\}_{(s+q+p)\times(s+q+p)} = \begin{pmatrix} W_{ZZ} & W_{ZM} \\ W_{MZ} & W_{MM} \end{pmatrix}$ are defined in a similar fashion after substituting $Z^\star$ with $(Z^\star, M)$. Plugging in (5) into (8), there is

$$\hat{\beta}^\star = \beta^\star + R_Z^T\epsilon + [W_{MZ}(\beta^\star - \hat{\beta}^{\star,init}) + W_{MM}(\gamma - \hat{\gamma}^{init})] \tag{9}$$

Under similar regularity conditions, except the change in covariate matrix, $\hat{\beta}^\star_j$ is asymptoti-

cally normal with

$$\sqrt{n}(\hat{\beta}_j^\star - \beta_j^\star) \approx N(0, n\sigma_\epsilon^2 (R_{:,j})^T R_{:,j})$$

A natural estimator of $\tilde{\beta}_j^\star - \beta_j^\star$ is $\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star$. Combining (7) and (9) yields

$$\hat{\tilde{\beta}}^\star - \hat{\beta}^\star = \tilde{\beta}^\star - \beta^\star + \tilde{R}^T\tilde{\epsilon} - R_Z^T\epsilon + \tilde{W}(\tilde{\beta}^\star - \hat{\tilde{\beta}}^{\star,init}) - [W_{ZZ}(\beta^\star - \hat{\beta}^{\star,init}) + W_{ZM}(\gamma - \hat{\gamma}^{init})]$$

The regularity conditions of the debiased lasso (Zhang and Zhang, 2014) ensure that the two bias terms $\tilde{W}(\tilde{\beta}^\star - \hat{\tilde{\beta}}^{\star,init})$ and $W_{ZZ}(\beta^\star - \hat{\beta}^{\star,init}) + W_{ZM}(\gamma - \hat{\gamma}^{init})$ are negligible. Consequently, the bias of $\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star$ is negligible, and it is also asymptotically normal. Its variance is

$$Var(\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star) = Var(\tilde{R}_{:,j}^T\tilde{\epsilon} - R_{:,j}^T\epsilon) = \sigma_{\tilde{\epsilon}}^2 \tilde{R}_{:,j}^T\tilde{R}_{:,j} + \sigma_\epsilon^2 R_{:,j}^T R_{:,j} - 2\rho_\epsilon\sigma_{\tilde{\epsilon}}\sigma_\epsilon \tilde{R}_{:,j}^T R_{:,j}$$

where $\rho_\epsilon = Corr(\tilde{\epsilon}, \epsilon)$. This variance is essentially the same as the corrected formula of McGuigan-Langholtz standard error (McGuigan and Langholtz, 1988) of the difference-in-coefficients estimator of the indirect effect for univariate exposure and univariate mediator as presented in MacKinnon et al. (2002).

To summarize, if the regularity conditions of debiased lasso (Zhang and Zhang, 2014) are satisfied for both (4) and (5), there is

$$\sqrt{n}[(\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star) - (\tilde{\beta}_j^\star - \beta_j^\star)] \approx N\left(0, n(\sigma_{\tilde{\epsilon}}^2 \tilde{R}_{:,j}^T\tilde{R}_{:,j} + \sigma_\epsilon^2 R_{:,j}^T R_{:,j} - 2\rho_\epsilon\sigma_{\tilde{\epsilon}}\sigma_\epsilon \tilde{R}_{:,j}^T R_{:,j})\right) \tag{10}$$

Here $\sigma_{\tilde{\epsilon}}^2$ and σ_ϵ^2 can be estimated based on the scaled lasso, and ρ_ϵ can be estimated using the empirical correlation of the residuals from the two scaled lasso regressions.

The assumptions for causal interpretation of the direct and indirect effects are similar to what have been presented in the literature (Zhou et al., 2020; VanderWeele and Vansteelandt, 2014; Guo et al., 2022a; Huang et al., 2023). We assume that there are no unmeasured confounding of the exposure-outcome, mediator-outcome or the exposure-mediator relationship, and the exposures do not affect any confounder between mediators and outcome. We also assume that the exposures are not causally related. However, they can be correlated through measured or unmeasured confounders. No assumptions are made on the causal

relationship among the mediators. The most important modeling assumption is the linear model assumption (1)-(3). MedDiC does not require any additional modeling assumptions other than the regularity conditions for the debiased lasso (Zhang and Zhang, 2014). We essentially require that the regression coefficients in (1) and (2) are sparse, and the spectral properties of the design matrix such as the sparse eigenvalues satisfy certain inequalities. Please refer to Sections 3.1, 3.4 and 3.5 of Zhang and Zhang (2014) for the more discussions on the specific conditions and how to check them in practice. We remark that the modeling assumptions needed for the proposed difference-in-coefficients approach are less restrictive than any potential product-of-coefficients approach using similar regression methods, because the product-of-coefficients approach requires additional modeling assumptions for fitting the mediator models (3).

2.2 *Implementation*

Most of the current implementations of the debiased lasso and the scaled lasso are slow or does not accommodate our mediation problem directly. Thus we implement MedDiC from scratch using RCPP. Additionally, we use a scaled adaptive lasso as the initial estimator for the debiased lasso to reduce bias. This algorithm constructs the weights using estimated coefficients from an initial fit, a strategy also used in previous studies (Zhang, 2021). Please refer to the Web Appendix B for details.

3. Simulations

3.1 *Simulation model and methods for comparison*

We generate simulated data using the outcome model and the mediator model (2)-(3). For $i = 1, \ldots, n$, we simulate $X_{i,:} = (X_{i1}, \ldots, X_{is}) \sim N(0, I_s)$, $Z_{i,:} = (Z_{i1}, \ldots, Z_{iq}) \sim N(0, \Sigma)$ where $\Sigma = \left(r^{|j-k|}\right)_{j,k=1}^{q}$ is a toeplitz matrix with entries $\Sigma_{jk} = r^{|j-k|}$. We also simulate $\eta_i \sim N(0, I_p)$ and $\epsilon_i \sim N(0, 1)$. We set $s = 2$, $\alpha = (-0.2, 0.2)^T$, and A to be a matrix

whose rows are random permutations of $\{-0.1, 0.1\}$. Recall that the indirect and the direct effect of exposure j are $B_{j,:}\gamma$, and β_j, respectively, where $B_{j,:}$ is the jth row of matrix B. We simulate β, B and γ such that only four of q exposures have non-zero indirect and/or direct effects. The effect sizes $(B_{j,:}\gamma,\ \beta_j)$ for these four exposures are chosen to represent exposures with different effects: $(-\tau, 0)$ for only indirect effect (complete mediation), $(0, -\tau)$ for only direct effect, (τ, τ) for both effects with the same sign, and $(1.2\tau, -0.8\tau)$ for both effects with different signs. Here τ is the simulation parameter for the signal strength. Please refer to Web Appendix C for details on how γ and B are simulated to satisfy the above requirement.

In this simulation study, we set $n = 300$, $p = 500$, $s = 2$, and explore various signal strength $\tau \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$. Our proposed method is evaluated for low-dimensional ($q = 5$), and high-dimensional ($q = 400$) exposures, with consideration given to both uncorrelated ($r = 0$), and correlated ($r = 0.4$) exposures. For low-dimensional exposures, we examine both $p_m = 1$ and $p_m = 5$, while for high-dimensional exposures, we only consider $p_m = 5$. We repeat each simulation setting for 100 times except for the methods with extremely high computing cost (See Web Appendix D for details). The testing error rates, powers and coverage probabilities are calculated as averages over all exposures in all simulation replicates. For instance, in each of the 100 replicate, there are three exposures with non-zero indirect effect, so the empirical power of detecting the indirect effect is the proportion detected among the 3×100 true non-zero indirect effects.

MedDiC is designed for high-dimensional exposures, and can be applied to low-dimensional exposures. We directly compare MedDiC with ZWZ (Zhou et al., 2020) and GLLZ (Guo et al., 2022a) when the exposures are low-dimensional. When the exposures are high-dimensional, we adapt each of GLLZ and ZWZ in two different ways for benchmarking purpose. The first type of adaptation is to apply GLLZ or ZWZ after marginal screening based on the correlation between the exposure and the outcome (**GLLZ AMS** and **ZWZ AMS**). The other type

of adaptation is applying GLLZ or ZWZ for each of the q exposures separately (**GLLZ US** and **ZWZ US**). These adaptations lead to four competing methods in high-dimensional settings. The bootstrap-based mediateR (Huang et al., 2023) is a computationally intensive method that can potentially be applied to high-dimensional settings. We investigated the computing time and the performance of two versions of mediateR, MedR L2 (with ridge penalty), and medR L1 (with lasso penalty) using the case $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$, and decided not to include them in further simulations due to their poor performance and high computational cost. The estimated computing times for MedR L1 and MedR L2, if included in all simulations, are 10^4 and 10^5 hours, respectively. Following the discussions in Zhou et al. (2020), HIMA and HIMA2 (Perera et al., 2022) are included when there is only one true mediator. Please refer to Web Appendix D for details of these methods.

To account for multiple testing, we apply Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995) at level $\alpha = 0.05$. We compare the empirical FDR and empirical power as defined in Dezeure et al. (2015); Zhang et al. (2016). We also compare the empirical marginal coverage probability and widths of the 95% confidence intervals. The marginal coverage probabilities of the confidence intervals for exposures with zero and nonzero effects are plotted in separate figures, because their behaviors are different in some settings.

3.2 *Inference of indirect effects*

The upper panels of Figure 3 and Web Figure 1 display the empirical FDR and power for detecting significant indirect effects for the cases with low-dimensional exposure and one true mediator. Our findings indicate that HIMA has the highest power. However, HIMA and HIMA2 do not control FDR to the nominal level when the exposures are correlated. Among MedDiC, ZWZ and GLLZ, the three methods that effectively controls FDR, MedDiC demonstrates the highest power. The confidence intervals based on MedDiC are also much shorter than GLLZ and ZWZ, and their coverage probabilities remain comparable and close

to the nominal level (lower panels of Figure 3 and Web Figure 1). MedR L1 and MedR L2 fail to control FDR or return confidence intervals with coverage probabilities close to the nominal level (Web Figure 1).

[Figure 3 about here.]

In applications, there are usually more than one true mediators. Our results for the setting of low-dimensional exposures ($q = 5$) with $p_m = 5$ true mediators (Web Figures 3-4) confirm that MedDiC has higher power and shorter confidence intervals than ZWZ and GLLZ.

Figure 4 and Web Figure 6 depict the comparison of MedDiC with ZWZ AMS, ZWZ US, GLLZ AMS and GLLZ US for high-dimensional exposures. We find that MedDiC maintains the highest power and the shortest confidence intervals. All methods demonstrate reasonable control over FDR and the coverage probabilities of the confidence intervals for true zeros remain near the nominal level. However, the confidence intervals based on ZWZ AMS and GLLZ AMS have low coverage probabilities for the true nonzero indirect effects.

[Figure 4 about here.]

3.3 *Other simulation results*

The primary focus of this paper is the inference for indirect effect, and we do not claim that the results of the direct and total effects from MedDiC is a novel contribution. This is because they are the direct outputs of the debiased lasso applied to (1) and (2). Despite this, we present the inference results of direct effects and total effects in Web Appendix E and Web Figures 7-18, and find that MedDiC outperforms the competing methods in the inference for direct effects.

We also compare the computing times, and our results (Web Appendix F and Web Table 1) suggest that MedDiC is computationally efficient, especially for the high-dimensional applications.

4. Real data examples

4.1 *Mediation-based fine-mapping for flowering time of Zea Mays: high-dimensional exposures*

We apply the proposed method to a subset of the data analyzed in Yang et al. (2022). The outcomes that we are interested in are the best linear unbiased prediction (BLUP) values of days to tassel (**DT**, days after planting until 50% of plants in the row shedding pollen), days to silk (**DS**, days after planting until 50% of plants in the row silking), growing degree days to tassel (**GDT**), and growing degree days to silk (**GDS**). These traits are all closely related to each other because they represent different measures of the maize flowering time. For the potential exposures, we focus on the 10,217 SNPs in the locus chr3:160429526-161478267 where there is a common GWAS peak for the above four traits. The potential mediators are the gene expressions in mature leaves, which is the most relevant tissue to the flowering times in the dataset (Kremling et al., 2018). In particular, we use the 155 genes on chr3 that also have an eQTL peak in this region (Khaipho-Burch et al., 2023). After removing the lines with missing values in the phenotypes and/or gene expression, there are 191 maize inbred lines left. In summary, the dataset has 155 potential mediators, 10,217 exposures, and the sample size is 191. Following the convention in quantitative genetics, we also use the top 20 eigen vectors of the whole genome as the confounders. Since the exposures are high-dimensional, we compare MedDiC with GLLZ US and MedR L1. For GLLZ US, we further add the top 10 eigen vectors of the SNPs in the region as additional confounders.

The scientific goal of this analysis is to identify the causal SNPs that underlie the GWAS association in the region, which is sometimes referred to as fine-mapping (Spain and Barrett, 2015). Raw GWAS outputs usually contain a large number of significant associations within a locus. Fine-mapping methods re-prioritize these SNPs based on statistical and biologically functional evidence to provide the biologists a smaller list of the top potential causal SNPs. In

this aspect, our analysis is mediation-based fine-mapping that incorporates gene expression data. The difference is that our analysis uses the raw data, while most fine-mapping methods are based on GWAS summary statistics as they were developed for human studies.

Web Table 2 reports the numbers of SNPs with each effect type selected by each method from each analysis at significance level $FDR = 0.2$. MedR L1 identifies none. For MedDiC and GLLZ US, the numbers across the four traits shows high variations. In particular, GLLZ US identified no indirect effects for DS and DT, and large numbers of indirect effects for GDS and GDT. This is not expected because the four traits are closely related. It is possible that the noise levels of the four phenotypes are different. Overall, GLLZ US tends to select at least 10 times more SNPs than MedDiC, except the indirect effect for DS and DT. However, this is not an advantage of GLLZ US, because the goal of fine mapping is to provide a smaller list of causal SNPs for downstream interpretations, and 600 to 2,000 SNPs out of 10,217 from one single locus are still too many for most biological interpretations.

An alternative strategy in fine-mapping is to select a fixed number of top SNPs (e.g., the top 100), regardless the statistical significance level. For the rest of this section, we will compare the three methods based on the top 100 SNPs for each effect type.

Since the four traits are related, we can evaluate the reproducibility of each method by overlapping their top SNPs of the same effect type across the four analysis, and larger overlapping sizes are preferred. Table 1 presents the number of overlaps of the top 100 SNPs between each pair of traits. We observe that the top ranked SNPs from MedDiC generally have much larger overlaps than the others, and MedR L1 results rarely overlap.

[Table 1 about here.]

To evaluate the biological relevance of the model outputs, we turn to external biological evidence. In this paragraph, we use the acronyms in Yang et al. (2022) and refer the SNPs with indirect effects as iSNPs. Since the iSNPs are expected to influence the phenotype

through regulating gene expressions of some of the candidate mediating genes, they are more likely to locate in the coding regions or regulatory regions of these mediating genes. The SNPs in this study are predominantly distant from the candidate mediating genes. Therefore, the regulation mechanism between these SNPs and their target genes are likely long-range cis-regulatory activities. Ricci et al. (2019) applied HiChIP and many other high-throughput sequencing methods to profile the widespread distal accessible chromatin regions (dACRs) in the maize genome. These short genomic regions have high potential in transcriptomic regulation activities, and are connected to many genes through the 3D genome structure. According to their argument, chromatin loops that connect dACRs with genes likely contain pairs of long-range cis-regulatory elements and their target genes. An iSNP is considered being supported by HiChIP evidence if there is a HiChIP chromatin loop between it and at least one of the potential mediating genes. Despite that the overall numbers of overlaps are low, there are 4 to 6 of the top 100 iSNP from MedDiC are supported by HiChIP, while the number is only 3 for GLLZ US and 4-9 for MedR L1 (Web Table 3).

4.2 *Identifying cis-driver genes for insulin and islet-specific eQTL hotspot on chr2 of Mus musculus: low-dimensional exposures and high-dimensional confounders*

We analyze the previously mentioned mice f2 cross data for diabetes study (Tu et al., 2012; Tian et al., 2016). There is an islet specific eQTL hotspot and a QTL peak for insulin level colocalized on chr2. We focus on a region from 120Mbps to 170Mbps on chr2, which include the center of the eQTL hotspot and the QTL peak. We define an eQTL as cis- if the distance between the transcription starting site of the gene being regulated and the genetic marker is less than 10Mbps, and the other genes are trans-. This leads to 193 cis-regulated genes as exposures, and 4561 trans-regulated genes as the mediators. There are 2057 genetic markers across the whole genome, which are used as confounders to remove the population structure.

The dataset includes three independent measures of the blood insulin level at sacrifice

(named "10wk","10wkRep", and "rbm", respectively), which provides us a unique opportunity to evaluate statistical methods based the reproducibility of the analysis results across independent measurements of identical biological signal. We conduct three high-dimensional mediation analysis using the same confounders, exposures and mediators, and each of the three measurements of insulin as the outcome. For each analysis, we remove the mice with missing value in the outcome variable. There are 491 mice in total in the dataset with gene expression data, the number of mice in the three analyses are 491, 488 and 490, respectively. This case study is a setting in which the exposures are not high-dimensional ($q < n$), but the confounders are ($s > n$). MedDiC and MedR L1 can be be applied direct. For ZWZ and GLLZ, we use the top 20 eigen vectors of the genotype matrix as the confounders so that $20 + q < n$. All methods return the results at significance level $FDR = 0.1$.

Web Table 4 presents the number of each type of effects in each analysis, and the overlap of the same type of effects between each pair of phenotypes. We find that MedDiC identifies more exposures with indirect effect than the others, while still maintaining similar level of overlaps. More than half of the indirect effects are identified in at least two analysis. GLLZ does not identify any indirect effect and only one to two direct effects.

Furthermore, we compare the methods using external biological evidence. As previously mentioned, exposures with the indirect effects are the cis-regulated genes that regulates the trans-eQTLs, which in turn regulate the outcome. Thus they are more likely to be regulator genes such as transcription factors (TFs) than the exposures with direct and or total effects but no indirect effect. We create a list of 1929 known TFs for *Mus musculus* (See Web Appendix G for details). Table 2 presents the overlap between the model outputs and the TF list. Since the number of transcription factors at the locus under consideration is likely to be small, we do not expect a large overlap. However, the contrasts in Table 2 is clear. MedDiC identifies many known transcription factors among the exposures with indirect

effects, while ZWZ has one and MedR L1 returns none. MedDiC finds no TFs among the exposures with no indirect effect but only direct or total effects, while ZWZ method returns some TFs as significant direct effects. The TFs selected by MedDiC are fairly consistent across all three phenotypes. Especially, MGA, FOXa2, E2F1 and DNMT3B are identified in all three analyses. Many of them are known to be relevant to diabetes. For example, Li et al. (2003) reported that the mutant mice deficient for the E2F1 and E2F2 transcription factors developed progressive pancreatic degeneration and insulin-dependent diabetes, and they could be rescued by transplanting the bone marrow from the wild-type mice with normal E2F1 and E2F2 levels.

[Table 2 about here.]

5. Discussions

Advances in GWAS and QTL mapping and their applications to molecular traits have yielded countless discoveries of genetic associations, and generated rich resources for further dissections of the colocalized GWAS and eQTL to investigate the molecular mechanism of genetic regulation of the phenotypes. Our research is motivated by two such dissection studies. The first motivating example is a genome-wide mediation analysis of maize data, one of whose goals is to detect the causal SNPs responsible for mediation at transcriptomics level. The other is on mice. It focuses on the eQTL hotspot that overlaps with a QTL peak relevant to T2D, and aims at identifying the cis-driver genes of the trans-regulated genes that contributes to the T2D related traits. Both biological research questions can be translated as a mediation problem whose primary goal is the estimation and inference of the overall indirect effect for potentially high-dimensional exposures.

For such mediation problems, we have proposed **Med**iation analysis via **D**ifference **i**n **C**oefficients (**MedDiC**) based on difference-in-coefficients approach of parameterizing the

indirect effect. Our simulation studies have shown that MedDiC consistently yield the highest power and the shortest confidence intervals with valid coverage probabilities and FDR control. MedDiC also provides inference results for the direct and total effects based on the debiased lasso. These inference are mostly valid and at least comparable to the outputs of GLLZ and ZWZ and their high-dimensional adaptations. MedDiC is computationally efficient, especially in high-dimensional settings.

We have applied the proposed MedDiC method to the two motivating datasets, and gained meaningful biological insights. MedDiC yields reproducible results across closely related traits in both analysis of the maize data with high-dimensional exposures and the mouse data with low-dimensional exposures and high-dimensional confounders. In the mouse data example, MedDiC outputs include many known transcription factors that are known to regulate gene expressions, and to be relevant to diabetes.

We have primarily focused on the simple linear structural equations in this paper. We have not considered any models with interactions between exposures and/or mediators, with nonlinear mediation effects, or with non-gaussian responses or mediators. MedDiC can be directly applied to some of these settings and provide statistical inference for indirect effect based on the difference-in-coefficients approach. However, these estimators and their target causal parameters may be different from their counterparts based on the product-of-coefficients approach. Careful studies of MedDiC for any of these models and the justification of their causal interpretations is one of our future interests. MedDiC has utilized the debiased lasso as in Zhang and Zhang (2014). We are aware that there are other debiased estimators for high-dimensional linear models (e.g., Javanmard and Montanari 2014), some of which have better theoretical properties and/or empirical performance in some settings. A comprehensive evaluation of these debiased estimators in the MedDiC framework may further improve the performance of MedDiC. In particular, we are interested in providing better inference for the

direct effects. Another possible extension of the current work is to improve its resolution, and quantify the indirect effect through a small subset of the mediators instead of all of them. This may depend on the positions of this subset in the causal graph among the potential mediators.

Acknowledgements

This research is supported by the startup fund from University of New Hampshire to QZ, and Agriculture and Food Research Initiative Grant number 2019-67013-29167 from the USDA National Institute of Food and Agriculture to JY.

Supplementary Materials

Web Appendix A-G, Web Tables 1-4, and Web Figures 1-18 referenced in Section 2-4 are available with this paper at the Biometrics website on Wiley Online Library.

The code and data for demonstration purpose are available on Github (link) after the paper is accepted.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)* **57,** 289–300.

Chén, O. Y., Crainiceanu, C., Ogburn, E. L., Caffo, B. S., Wager, T. D., and Lindquist, M. A. (2017). High-dimensional multivariate mediation with application to neuroimaging data. *Biostatistics* **19,** 121–136.

Dezeure, R., Bühlmann, P., Meier, L., and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science* pages 533–558.

Fang, R., Yang, H., Gao, Y., Cao, H., Goode, E. L., and Cui, Y. (2021). Gene-based mediation analysis in epigenetic studies. *Briefings in bioinformatics* **22,** bbaa113.

Flint-Garcia, S. A., Thuillet, A.-C., Yu, J., Pressoir, G., Romero, S. M., Mitchell, S. E., Doebley, J., Kresovich, S., Goodman, M. M., and Buckler, E. S. (2005). Maize association population: a high-resolution platform for quantitative trait locus dissection. *The Plant Journal* **44,** 1054–1064.

Guo, X., Li, R., Liu, J., and Zeng, M. (2022a). High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association* **117,** 1110–1121.

Guo, X., Li, R., Liu, J., and Zeng, M. (2022b). Statistical inference for linear mediation models with high-dimensional mediators and application to studying stock reaction to covid-19 pandemic. *Journal of Econometrics* .

Huang, L., Long, J. P., Irajizad, E., Doecke, J. D., Do, K.-A., and Ha, M. J. (2023). A unified mediation analysis framework for integrative cancer proteogenomics with clinical outcomes. *Bioinformatics* **39,** btad023.

Imai, K., Keele, L., and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological methods* **15,** 309.

Javanmard, A. and Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research* **15,** 2869–2909.

Keller, M. P., Paul, P. K., Rabaglia, M. E., Stapleton, D. S., Schueler, K. L., Broman, A. T., Ye, S. I., Leng, N., Brandon, C. J., Neto, E. C., et al. (2016). The transcription factor nfatc2 regulates β-cell proliferation and genes associated with type 2 diabetes in mouse and human islets. *PLoS genetics* **12,** e1006466.

Khaipho-Burch, M., Ferebee, T., Giri, A., Ramstein, G., Monier, B., Yi, E., Romay, M. C., and Buckler, E. S. (2023). Elucidating the patterns of pleiotropy and its biological

relevance in maize. *Plos Genetics* **19,** e1010664.

Kremling, K. A., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., Lu, F., Lorant, A., Bradbury, P. J., and Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* **555,** 520–523.

Li, F. X., Zhu, J. W., Tessem, J. S., Beilke, J., Varella-Garcia, M., Jensen, J., Hogan, C. J., and DeGregori, J. (2003). The development of diabetes in e2f1/e2f2 mutant mice reveals important roles for bone marrow-derived cells in preventing islet cell loss. *Proceedings of the National Academy of Sciences* **100,** 12935–12940.

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.

McGuigan, K. and Langholtz, B. (1988). A note on testing mediation paths using ordinary least-squares regression. *Unpublished note* pages 144–158.

Pang, J., Fu, J., Zong, N., Wang, J., Song, D., Zhang, X., He, C., Fang, T., Zhang, H., Fan, Y., et al. (2019). Kernel size-related genes revealed by an integrated eqtl analysis during early maize kernel development. *The Plant Journal* **98,** 19–32.

Pearl, J. (2012). *The mediation formula: A guide to the assessment of causal pathways in nonlinear models*. Wiley Online Library.

Perera, C., Zhang, H., Zheng, Y., Hou, L., Qu, A., Zheng, C., Xie, K., and Liu, L. (2022). Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC bioinformatics* **23,** 1–14.

Pierce, B. L., Tong, L., Chen, L. S., Rahaman, R., Argos, M., Jasmine, F., Roy, S., Paul-Brutus, R., Westra, H.-J., Franke, L., et al. (2014). Mediation analysis demonstrates that trans-eqtls are often explained by cis-mediation: a genome-wide analysis among 1,800 south asians. *PLoS genetics* **10,** e1004818.

Ricci, W. A., Lu, Z., Ji, L., Marand, A. P., Ethridge, C. L., Murphy, N. G., Noshay, J. M., Galli, M., Mejía-Guerra, M. K., Colomé-Tatché, M., Johannes, F., Rowley, M. J., Corces, V. G., Zhai, J., Scanlon, M. J., Buckler, E. S., Gallavotti, A., Springer, N. M., Schmitz, R. J., and Zhang, X. (2019). Widespread long-range cis-regulatory elements in the maize genome. *Nature Plants* **5,** 1237–1249.

Sohn, M. B., Li, H., et al. (2019). Compositional mediation analysis for microbiome studies. *The Annals of Applied Statistics* **13,** 661–681.

Spain, S. L. and Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human molecular genetics* **24,** R111–R119.

Tian, J., Keller, M. P., Broman, A. T., Kendziorski, C., Yandell, B. S., Attie, A. D., and Broman, K. W. (2016). The dissection of expression quantitative trait locus hotspots. *Genetics* **202,** 1563–1574.

Tu, Z., Keller, M. P., Zhang, C., Rabaglia, M. E., Greenawalt, D. M., Yang, X., Wang, I.-M., Dai, H., Bruss, M. D., Lum, P. Y., et al. (2012). Integrative analysis of a cross-loci regulation network identifies app as a gene regulating insulin secretion from pancreatic islets. *PLoS genetics* **8,** e1003107.

VanderWeele, T. and Vansteelandt, S. (2014). Mediation analysis with multiple mediators. *Epidemiologic methods* **2,** 95–115.

Xue, F., Tang, X., Kim, G., Koenen, K. C., Martin, C. L., Galea, S., Wildman, D., Uddin, M., and Qu, A. (2022). Heterogeneous mediation analysis on epigenomic ptsd and traumatic stress in a predominantly african american cohort. *Journal of the American Statistical Association* pages 1–36.

Yang, Z., Xu, G., Zhang, Q., Obata, T., and Yang, J. (2022). Genome-wide mediation analysis: an empirical study to connect phenotype with genotype via intermediate transcriptomic data in maize. *Genetics* **221,** iyac057.

Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **76,** 217–242.

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

Zhang, Q. (2021). High-dimensional mediation analysis with applications to causal gene identification. *Statistics in Biosciences* pages 1–20.

Zhao, Y. (2022). Mediation analysis with multiple exposures and multiple mediators. *arXiv preprint arXiv:2209.04405* .

Zhong, W., Spracklen, C. N., Mohlke, K. L., Zheng, X., Fine, J., and Li, Y. (2019). Multi-snp mediation intersection-union test. *Bioinformatics* **35,** 4724–4729.

Zhou, R. R., Wang, L., and Zhao, S. D. (2020). Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika* **107,** 573–589.

Received XX 2023. *Revised XX* 20XX.

Accepted XX 20XX.

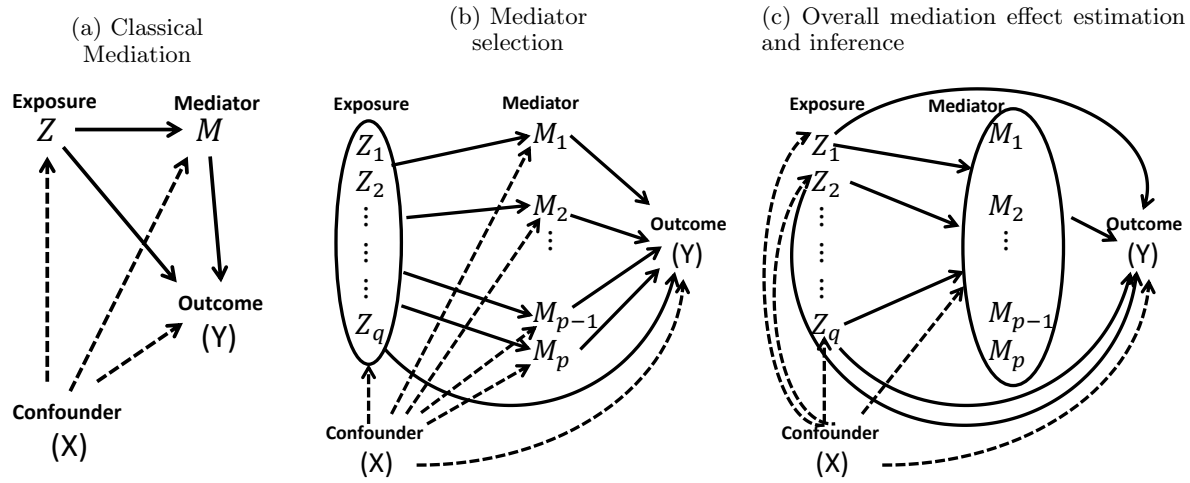
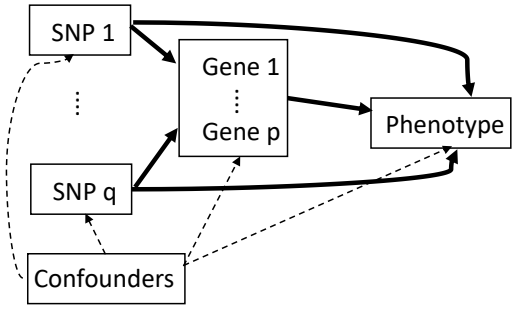


Figure 1: Causal diagrams for different mediation problems: classical mediation analysis (Left), mediator selection (Middle) and estimating the overall mediation effect (Right).

(a) Identifying the iSNPs



(b) Identifying the driver genes

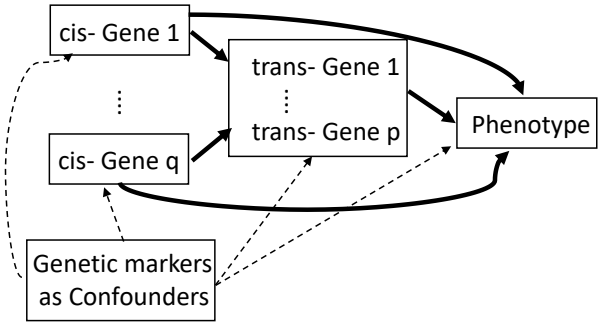


Figure 2: Causal diagrams for the two biological problems: identifying the iSNPs (Left), and identifying the driver genes (Right).

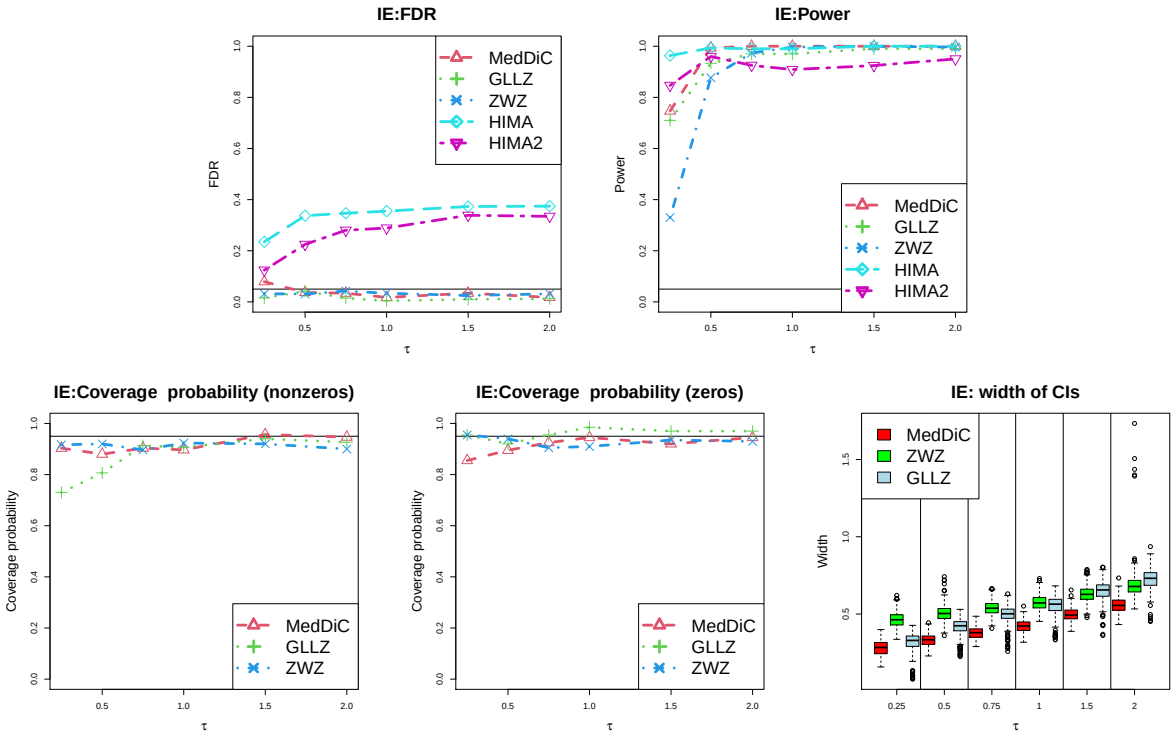


Figure 3: Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low-dimensional ($q = 5$), and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$)

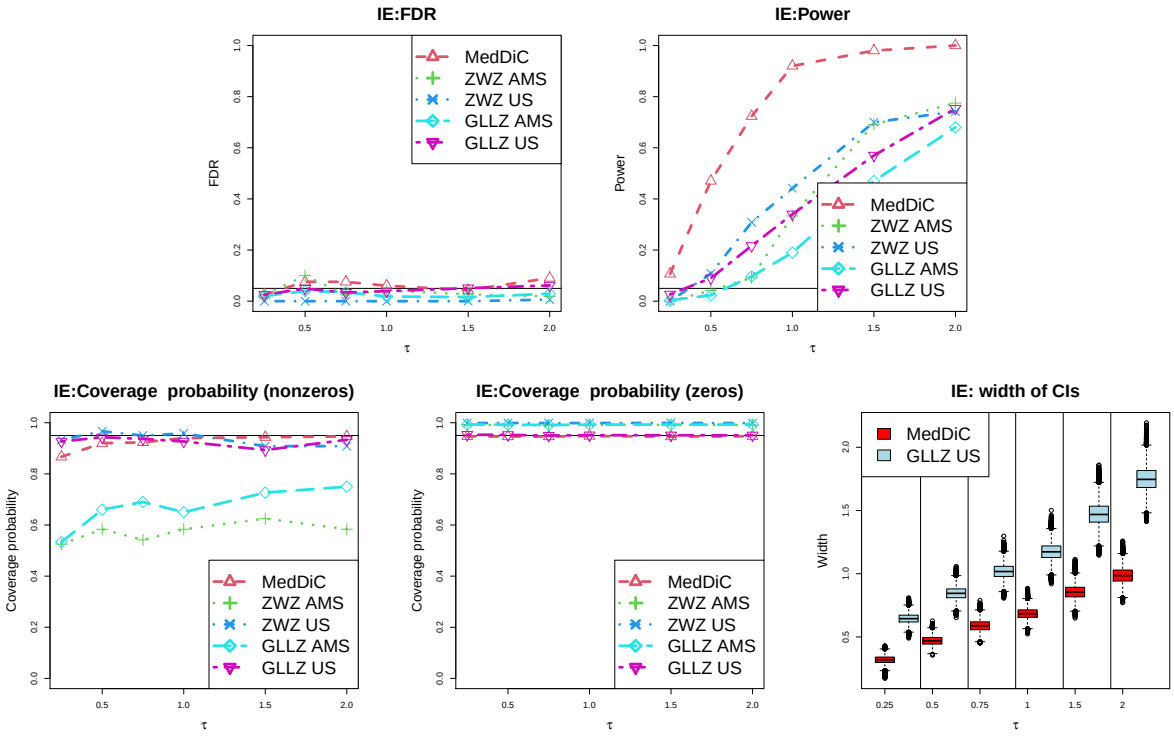


Figure 4: Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high-dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).

Table 1: Results of Maize data: Overlap of the top 100 SNPs.

	DS, DT	DS, GDS	DS, GDT	DT, GDS	DT, GDT	GDS, GDT
IE (MedDiC)	36	55	34	39	60	57
DE (MedDiC)	60	73	63	54	67	68
TE (MedDiC)	70	76	73	59	81	71
IE (GLLZ US)	36	5	2	2	4	49
DE (GLLZ US)	60	12	17	11	35	31
IE (MedR L1)	2	0	2	2	0	2
DE (MedR L1)	0	1	1	0	2	3
TE (MedR L1	1	0	0	1	0	0

Table 2: Results of Mouse data: transcription factors (TFs) in the selected cis-regulated genes with each effect type. GLLZ is excluded from the table because it selects no exposures with indirect effect and 1-2 with direct effect, and there are no TFs. MedR L1 is excluded from the table because there are no TFs among the 15-20 selected exposures with indirect effect for each analysis.

Effect type	Known Transcription Factors		
	10wk	10wkRep	rmb
IE (MedDiC))	mga,zfp106,foxa2, e2f1,dnmt3b,foxs1	mga,foxa2,ncoa6, e2f1,dnmt3b,foxs1	mga,zfp106,e2f1, sall4,ovol2,foxa2, id1,dnmt3b,mgef2
DE/TE no IE (MedDiC)	N/A	N/A	N/A
IE (ZWZ)	N/A	foxa2	foxa2
DE no IE (ZWZ)	foxa2,e2f1,ncoa3	e2f1	e2f1

Supplementary Materials of "Dissecting the colocalized GWAS and eQTLs with mediation analysis for high dimensional exposures and confounders"

Qi Zhang
Department of Mathematics and Statistics,
University of New Hampshire
qi.zhang2@unh.edu

Zhikai Yang
PhD student in Complex Biosystems Program
and
Department of Agronomy and Horticulture,
University of Nebraska-Lincoln

Jinliang Yang
Department of Agronomy and Horticulture,
University of Nebraska-Lincoln

May 29, 2024

Web Appendix A: the equivalence of the difference-in-coefficients approach and the product-of-coefficients approach for least square estimates

Without losing generalities, we consider the following linear meidation models.

$$\begin{aligned} Y &= Z\tilde{\beta} + \tilde{\epsilon} \\ Y &= Z\beta + M\gamma + \epsilon \\ M &= ZB + \eta \end{aligned}$$

Here Z is the matrix of the exposures and confounders combined. If there are q exposures and s confounders, Z is a matrix with $q+s$ columns, and we are only interested in the mediation effects for the exposures represented by the first q columns of Z. We assume that q, s and p, the number of mediators, are all low-dimensional so that the least square estimates of the individual equations work well. Specifically, the least square estimates are

$$\begin{aligned} \hat{\tilde{\beta}} = &\quad (Z^T Z)^{-1} Z^T Y \\ \hat{B} = &\quad (Z^T Z)^{-1} Z^T M \\ \hat{\beta} = &\quad (C_{11} Z^T + C_{12} M^T) Y \\ \hat{\gamma} = &\quad (C_{12}^T Z^T + C_{22} M^T) Y \end{aligned}$$

where

$$\begin{aligned} C_{22} = &\quad (M^T M - M^T Z (Z^T Z)^{-1} Z^T M)^{-1} \\ C_{11} = &\quad (Z^T Z)^{-1} + (Z^T Z)^{-1} Z^T M C_{22} M^T Z (Z^T Z)^{-1} \\ C_{12} = &\quad -(Z^T Z)^{-1} Z^T M C_{22} \end{aligned}$$

The following derivation shows that the difference-in-coefficients estimator based on the least squares is equivalent to the estimator based on the product-of-coefficients approach.

$$
\begin{aligned}
\hat{\tilde{\beta}} - \hat{\beta} &= (Z^T Z)^{-1} Z^T Y - (C_{11} Z^T + C_{12} M^T) Y \\
&= \left\{ (Z^T Z)^{-1} Z^T - (Z^T Z)^{-1} Z^T - (Z^T Z)^{-1} Z^T M C_{22} M^T Z (Z^T Z)^{-1} Z^T + (Z^T Z)^{-1} Z^T M C_{22} M^T \right\} Y \\
&= (Z^T Z)^{-1} Z^T M \left\{ C_{22} M^T - C_{22} M^T Z (Z^T Z)^{-1} Z^T \right\} Y \\
&= (Z^T Z)^{-1} Z^T M \left\{ C_{22} M^T + C_{12}^T Z^T \right\} Y \\
&= \hat{B} \hat{\gamma}
\end{aligned}
$$

This result is reassuring, and enables us to compare the two estimation approaches for the same target parameter. As illustrated in [15] and our study, different penalized estimators yield different asymptotic properties and empirical performance.

Web Appendix B: Implementation of MedDiC

Most of the current implementations of the debiased lasso and the scaled lasso are slow or do not accommodate our mediation problem directly [2, 10, 12]. Thus, we implement MedDiC from scratch using RCPP, leveraging some existing algorithms scattered in various packages such as fastGGM [11] and hdi [2].

Algorithm 1 Scaled adaptive lasso for linear model $y \sim N(X\beta, \sigma^2 I_n)$

Input: $n \times p$ covariate matrix X and length n outcome vector y

Output: $\hat{\beta}$ and $\hat{\sigma}^2$

1: *Initial estimate of β:* Apply the scaled lasso [10] with penalty term $\sum_{j=1}^{p} |\beta_j|$ to regress y on X, and return $\hat{\beta}^{(0)}$, the estimated coefficients.

2: *Construct weights for scaled adaptive lasso:* Let $w = (w_1, \ldots, w_p) \propto 1/(|\hat{\beta}^{(0)}| + \sqrt{var(y)/n})$ such that $\sum_{j=1}^{p} w_j = p$.

3: *Final estimate of β:* Minimize the scaled lasso objective with penalty term $\sum_{j=1}^{p} w_j |\beta_j|$ instead of $\sum_{j=1}^{p} |\beta_j|$, and return $\hat{\beta}$.

4: *Estimate σ^2:* Applied least square to regress y to the selected covariates based on $\hat{\beta}$, and return the mean squared error as $\hat{\sigma}^2$.

We uses a scaled adaptive lasso (Algorithm 1) as the initial estimator for the debiased lasso

to reduce bias. This algorithm constructs the weights using estimated coefficients from an initial fit, a strategy also used in previous studies [14, 16]. We estimate the noise level based on the least square after the scaled lasso, because [9] shows that it is less biased than using the direct output of the original scaled lasso. Although [9] advocates for cross-validation for noise level estimation and [2] suggests using lasso with cross-validation for the initial estimator of the debiased lasso, our exploratory simulation experiments show no clear gain in high dimensional setting to justify the additional computational cost of cross-validation. The score calculations are also based on the scaled lasso. The tuning parameter of the scaled lasso is $\sqrt{2\log(s+q)/n}$ for the models without the mediators, and $\sqrt{2\log(s+q+p)/n}$ for models with mediators. For the low dimensional problems with $q+s < c \cdot n$ where $c \in (0,1)$ is a fixed constant, the ordinary least square is used instead of the debiased lasso for estimating $\tilde{\beta}^\star$. In this study, we set $c = 0.2$.

The proposed MedDiC procedure is summarized as Algrorithm 2 where the equation numbers are referred to the equations in the main manuscript.

Algorithm 2 MedDiC: Mediation analysis using Difference in Coefficients

Input: The length n response vector Y, the $n \times p$ mediator matrix M, $Z^\star = (X, Z)$ where X is the $n \times s$ confounder matrix and Z is the $n \times q$ exposure matrix.

Output: $\hat{\tilde{\beta}}^\star - \hat{\beta}^\star$ the estimated indirect effect, and the standard errors.

1: *Estimate the total effect $\tilde{\beta}^\star$:* Apply Equation (6) where $\tilde{R}$ and $\tilde{W}$ are calculated according to [12], and $\hat{\tilde{\beta}}^{\star,init}$ is based Algorithm 1, and returns $\hat{\tilde{\beta}}^\star$.

2: *Estimate the direct effect $\beta^\star$:* Apply Equation (8) where R and W are calculated according to [12], and $\hat{\beta}^{\star,init}$ is based on Algorithm 1, and returns $\hat{\beta}^\star$.

3: *Estimate the indirect effect $\tilde{\beta}^\star - \beta^\star$:* returns $\hat{\tilde{\beta}}^\star - \hat{\beta}^\star$

4: *Estimate the standard error of indirect effect:* For $j = 1+s, \ldots, q+s$ calculate the standard errors of $\hat{\tilde{\beta}}_j^\star - \hat{\beta}_j^\star$ based on Equation (10).

Web Appendix C: Simulation model setup

We generate simulated data using the outcome model and the mediator model

$$Y = X\alpha + Z\beta + M\gamma + \epsilon \quad (1)$$

$$M = XA + ZB + \eta \quad (2)$$

For $i = 1, \ldots, n$, we simulate $X_{i,:} = (X_{i1}, \ldots, X_{is}) \sim N(0, I_s)$, $Z_{i,:} = (Z_{i1}, \ldots, Z_{iq}) \sim N(0, \Sigma)$ where $\Sigma = \left(r^{|j-k|}\right)_{j,k=1}^{q}$ is a toeplitz matrix with entries $\Sigma_{jk} = r^{|j-k|}$. The cases with $r \neq 0$ represents the scenarios in which the multivariate exposures are correlated through unmeasured confounders. We also simulate $\eta_i \sim N(0, I_p)$ and $\epsilon_i \sim N(0, 1)$. We set $s = 2$, $\alpha = (-0.2, 0.2)^T$, and A to be a matrix whose rows are random permutations of $\{-0.1, 0.1)\}$.

Recall that the indirect and the direct effect of exposure j are $B_{j,:}\gamma$, and β_j, respectively, where $B_{j,:}$ is the jth row of matrix B. We simulate β, B and γ such that only four out of q exposures have non-zero indirect and/or direct effects. The effect sizes $(B_{j,:}\gamma,\ \beta_j)$ for these four exposures are chosen to represent exposures with different effects: $(-\tau, 0)$ for only indirect effect (complete mediation), $(0, -\tau)$ for only direct effect, (τ, τ) for both effects with the same sign, and $(1.2\tau, -0.8\tau)$ for both effects with different signs. Here τ is the simulation parameter for the signal strength.

In the following, we describe in more details how γ and B are simulated to satisfy the above requirement. Let p_m be the number of true mediators. We first simulate γ as a sparse vector with $2p_m$ non-zero elements with value $\sqrt{\tau}$. This choice is made so that the non-zero elements in both B and γ grow as the signal strength τ increases. Half of these $2p_m$ candidate mediators are true mediators, and the other half are referred to as "spurious mediators". They have non-zero coefficients in the outcome model, but zero link to the exposures. Let $C_{true}, C_{spur} \subset \{1, \ldots, p\}$ be their coordinate sets, and $S \subset \{1, \ldots, q\}$ be the subset of exposures with indirect effects. We initialize the $q \times p$ matrix B with 10 nonzero elements at random locations in each row. These non-zero elements are random samples from a uniform distribution between -1 and 1. We then modify B as the following. $B_{S^c, C_{true}}$ is for the association between the exposures with no indirect effect and the true mediators, and $B_{:, C_{spur}}$ is for the association between all exposures and the fake mediators. By definitions of S^c and C_{spur}, their elements are replaced with zeros. $B_{S, C_{true}}$ is the sub-matrix of the association between the exposures with non-zero mediation effects and the

true mediators. We replace the elements in this sub-matrix row-wise with repeating sequence of 1,0,2, allowing each exposure with an indirect effect to influence a distinct subset of true mediators. When $p_m = 1$, however, all exposures with non-zero mediation effect must influence this only true mediator, and $B_{S,C_{true}}$ is filled with the sequence of 1,1,2 instead. We further normalize B row-wise so that the nonzero elements of $B\gamma$ are $-\tau, \tau$, and 1.2τ. We then simulate β as a sparse vector with three nonzero elements $-\tau, \tau, -0.8\tau$. Finally, vectors $B\gamma$ and β are jointly shuffled such that there are only four exposures that have nonzero direct and/or indirect effect, and their values are as specified in the last paragraph.

In this simulation study, we set $n = 300$, $p = 500$, $s = 2$, and explore various signal strength $\tau \in \{0.25, 0.5, 0.75, 1, 1.5, 2\}$. Our proposed method is evaluated for low dimensional exposures $q = 5$, and high dimensional exposures $q = 400$, with consideration given to both uncorrelated exposures ($r = 0$), and correlated exposures ($r = 0.4$). For low dimensional exposures, we examine both $p_m = 1$ and $p_m = 5$, while for high dimensional exposures, we only consider $p_m = 5$. We repeat each simulation setting for 100 times, except for the high dimensional adaptations of [15], which are based on 40 replicates due to their high computational cost. The results for the two versions of mediateR [6] are based on 20 replicates due to the high computational cost and poor performance. The testing error rates, powers, coverage probabilities and the width distribution of the confidence intervals are calculated over all exposures in all simulation replicates. For instance, in each of the 100 simulation replicate, there are three exposures with non-zero indirect effect, so the empirical power of detecting the indirect effect is the proportion detected among the 3×100 true non-zero indirect effects.

Web Appendix D: Methods to be compared in simulations and real data analysis

MedDiC is designed for high dimensional exposures, and it can also be applied to low dimensional exposures. We directly compare MedDiC with two state-of-art methods for the inference of indirect effects with low dimensional exposures. They are ZWZ [15] and GLLZ [3]. As the original code of [15] did not report any results for the inference of direct effect, and that of [3] did not provide any confidence intervals, we modified their code to include these results based on

their asymptotic distributions. We compare these methods in terms of empirical FDR, power, and the marginal coverage probability and widths of 95% confidence intervals for low dimensional exposures.

When the exposures are high dimensional, GLLZ and ZWZ cannot be directly applied, and there is no method for this setting except the proposed MedDiC method and mediateR. For benchmarking purpose, we adapt each of GLLZ and ZWZ in the following two different ways for high dimensional exposures. This results in four methods to be compared with the proposed MedDiC method in high dimensional settings.

(1) After Marginal Screening (**GLLZ AMS** and **ZWZ AMS**): We first apply screening by marginal correlation between the exposures and the outcome, keep the top 20, and then apply GLLZ or ZWZ to the survived exposures. The exposures that are not used in the final model fit return p-value= 1, and a confidence interval $[0, 0]$. Assuming a true effect of 0, we consider this interval to cover the true value in our evaluations.

(2) Univariate Scanning (**GLLZ US** and **ZWZ US**): This is essentially what [15] has used in their real data analysis. Let X_0 be the low dimensional confounder matrix, and U be the top left singular vectors of the high dimensional exposure matrix Z (top five in simulations, and top 20 in real data analysis). For $j = 1, \ldots, q$, GLLZ or ZWZ is applied to exposure matrix $(X_0, U, Z_{:,j})$. It requires to run GLLZ or ZWZ q times for a dataset. We apply GLLZ US to real data analysis, while ZWZ US is excluded due to computational constraints. For simulation studies, we further modify ZWZ US to reduce the computational cost. In our simulations, we only run ZWZ US to a subset of exposures $S_o \subset \{1, \ldots, q\}$ that include all non-zero effects. This subset is chosen based on the simulation model as the following. If $j \in \{1, \ldots, q\}$ is an exposure with non-zero indirect or direct effect, then $\{j-1, j, j+1\} \cap \{1, \ldots, q\} \subset S_0$. Since there are four exposures with non-zero effects in the simulation model, ZWZ US only need to run up to 12 times, rather than 400 times for each simulation replicate. For exposures $j \notin S_o$, ZWZ US return p-value= 1, and $[0, 0]$ as confidence interval. We remark that all true signals are used in fitting the ZWZ model, which is different from ZWZ AMS. Since the computational cost of ZWZ increases roughly linearly with q, this modification allows us to include ZWZ US in simulation-based comparisons at 3% of the computational cost without losing any true signals.

HIMA[13] is one of the most widely recognized mediator selection algorithm, and it has been

recently modified as HIMA2 [8]. Their statistical inference objective differs from ours, and cannot be compared with MedDiC, ZWZ or GLLZ in general settings. As discussed in [15], HIMA's objective aligns with ours when there is only one true mediator. For the low dimensional exposure ($q = 5$) and when $p_m = 1$, we include HIMA and HIMA2 for comparing FDR and power. They are designed for univariate exposure. Thus we apply them to each exposure separately. It should be noted that it requires q separate runs for each simulation replicate. For each exposure, we use the p-value of its most significant mediator as the p-value of the overall mediation effect of this exposure.

The mediateR package [6] proposed a unified inference framework for mediation analysis based on the product-of-coefficients approach. Its estimation is based on penalized generalized linear model and the inference is based on nonparametric bootstrap. In the original paper, this framework was only evaluated for the binary and survival outcomes using ridge penalty. This conceptual framework can be applied the settings with high dimensional mediators, exposures and confounders. We refer to mediateR model with ridge penalty as medR L2 method. We also extend the implementation of mediateR to utilize the lasso penalty for better computational efficiency, and refer to it as medR L1. To accommodate the sparsity penalty, the definition of p-value for mediateR (the paragraph after equation (7) on page 5 of [6]) is modified as $2min(P_L, P_U) + \frac{1}{B}\sum_{b=1}^{B} 1_{\hat{IE}_{X_i}^{(b)}=\Delta}$. Without such modification, the bootstrap replicates whose estimated coefficient is exactly 0 are not counted towards the p-value when $\Delta = 0$. For one simulation setting $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$, we compare their performance and computing time with 500 bootstrap replicates. We decide to exclude MedR L1 and MedR L2 from further simulation studies due to their poor performance (Web Figures 1,7,13) and the high computational cost (See Web Appendix F for details). In particular, if fully included in our simulation studies, the estimated computing times for MedR L1 and MedR L2 are of the order of 10^4 hours and 10^5 hours, respectively . medR L1 is included in the two real data analysis.

To account for multiple testing, we apply Benjamini-Hochberg procedure[1] at level $\alpha = 0.05$ to all methods. At this level, we compare the empirical FDR and empirical power as defined in [14, 2, 13]. We also compare the empirical marginal coverage probability and wdiths for 95% confidence intervals.We plot the marginal coverage probabilities of the confidence intervals for exposures with zero and non-zero effects in separate figures, because their behaviors are different

in some settings.

We recap the methods included in the comparison for each analysis. For the simulations with low dimensional exposures and only one true mediator, we compared MedDiC with ZWZ, GLLZ, HIMA and HIMA2. For the one setting $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$ that has been used for computing time comparisons, MedR L1 and MedR L2 are also included in the comparisons with 20 replicates. For the simulations with low dimensional exposures and five true mediators, we compared MedDiC with ZWZ, GLLZ. For the simulations with high dimensional exposures, we compared MedDiC with ZWZ AMS, ZWZ US, GLLZ AMS, and GLLZ US. The only exception is the comparison in the width of the confidence intervals, ZWZ AMS, GLLZ AMS and ZWZ US are excluded because they return intervals with 0 widths for the screened exposures. For the analysis of the Maize data where the exposure is high dimensional and the number of mediators is slightly smaller than the sample size, MedDiC was compared with GLLZ US and MedR L1. For the analysis of the Mice data with high dimensional mediators, high dimensional confounders, and the number of exposures is about one half of the sample size, MedDiC was compared with GLLZ, ZWZ and MedR L1.

Web Appendix E: Simulation results for the inference of direct effects and total effects

The primary focus of this paper is to infer indirect effect, and we do not claim that inferring the direct and total effects from MedDiC is a novel contribution. This is because they are the direct outputs of the debiased lasso.

Despite this, we present the inference results of direct effects from MedDiC, and compare them with mediateR, ZWZ and GLLZ (and their high dimensional adaptations) in Web Figures 7-12. For low dimensional exposures, MedDiC performs comparably to ZWZ and GLLZ, and the empirical coverage probabilities of non-zero direct effects for all three methods are slightly lower than the nominal level. MedR L1 and MedR L2 have poor coverage probability for the nonzero direct effects. Additionally, MedR L2 also yields poor FDR control and the coverage probability for the zero direct effect. For high dimensional exposures, we find that MedDiC and GLLZ US show the highest power. However, GLLZ US fails to control FDR near the nominal level. The empirical coverage probabilities of non-zero direct effects for all methods are lower than the nominal level, and GLLZ AMS and ZWZ AMS still perform the worst based on this measure. MedDiC yields shorter confidence intervals than GLLZ US.

Web Figure 13-18 reports the inference results of total effects from MedDiC. As noted earlier, they are the direct outputs of the debiased lasso. We do not claim this part as a novel contribution, and only present the simulation results for completeness. ZWZ and GLLZ do not report or estimate the total effects. We find that the inference of the total effect is valid in all scenarios in terms of FDR control and coverage probabilities of the confidence intervals, and the power is reasonable. MedR L1 and MedR L2 return low coverage probability for the nonzero total effects.

Web Appendix F: Simulation results for computational time comparison

We also evaluate the computational cost of MedDiC and compare with ZWZ, GLLZ, HIMA, HIMA2, MedR L1 and MedRL2 on the same computer. This part of the simulation were per-

formed on a laptop with Intel(R) i7-1065G7 CPU @1.30GHz 1.50GHz, 32 GB memory, Win 10 operation system and Microsoft R Open 4.02 for computing. We remark that Microsoft R Open 4.02 is capable of utilizing multiple cores. In Web Table 1, we present the average computational time (in seconds) of MedDiC, ZWZ, HIMA, HIMA2, GLLZ, MedR L1 and MedR L2 in the simulation setting with $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$. We find that MedDiC and GLLZ are much faster than the rest. GLLZ is faster because it does not need to calculate the score matrix for the debiased lasso. We find that our extension of the mediateR method (MedR L1) does reduce the computational cost comparing with the the MedR L2 as implemented in [6]. However, MedR L1 and MedR L2 remain too time consuming to be fully included in all simulation settings. The estimated pure computing times for MedR L1 and MedR L2, if fully included in all simulation settings with 500 bootstrap replicates, are at the order of 10^4 hours and 10^5 hours using the same machine. For the high dimensional setting, We only compare MedDiC and the two high dimensional adaptations of GLLZ, and we find that GLLZ AMS is fast due to the screening step, and GLLZ US is more than 10 times slower than MedDiC because it takes $q = 400$ GLLZ runs. We do not include the other methods, because they are expected to be much slower than in the low dimensional setting as more exposures are included in the model fitting.

We remark that we do not document the computational cost for all our simulation settings or using more replicates largely due to the high computational cost of some of the competing methods. We use parallel computing and multiple computers, and it is difficult to compare the computational times consistently in this setting.

Web Appendix G: Generating the list of known TFs for mouse

To gather a list of known transcription factors for *Mus musculus* , we collect data from three sources [4, 5, 7]. Each list contains over 1300 TFs, with a merged list of 1929 TFs.

References

[1] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[2] R. Dezeure, P. Bühlmann, L. Meier, and N. Meinshausen. High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*, pages 533–558, 2015.

[3] X. Guo, R. Li, J. Liu, and M. Zeng. High-dimensional mediation analysis for selecting dna methylation loci mediating childhood trauma and cortisol stress reactivity. *Journal of the American Statistical Association*, 117(539):1110–1121, 2022.

[4] J. Hammelman, T. Patel, M. Closser, H. Wichterle, and D. Gifford. Ranking reprogramming factors for cell differentiation. *Nature Methods*, 19(7):812–822, 2022.

[5] H. Hu, Y.-R. Miao, L.-H. Jia, Q.-Y. Yu, Q. Zhang, and A.-Y. Guo. Animaltfdb 3.0: a comprehensive resource for annotation and prediction of animal transcription factors. *Nucleic acids research*, 47(D1):D33–D38, 2019.

[6] L. Huang, J. P. Long, E. Irajizad, J. D. Doecke, K.-A. Do, and M. J. Ha. A unified mediation analysis framework for integrative cancer proteogenomics with clinical outcomes. *Bioinformatics*, 39(1):btad023, 2023.

[7] M. Lizio, J. Harshbarger, H. Shimoji, J. Severin, T. Kasukawa, S. Sahin, I. Abugessaisa, S. Fukuda, F. Hori, S. Ishikawa-Kato, et al. Gateways to the fantom5 promoter level mammalian expression atlas. *Genome biology*, 16(1):1–14, 2015.

[8] C. Perera, H. Zhang, Y. Zheng, L. Hou, A. Qu, C. Zheng, K. Xie, and L. Liu. Hima2: high-dimensional mediation analysis and its application in epigenome-wide dna methylation data. *BMC bioinformatics*, 23(1):1–14, 2022.

[9] S. Reid, R. Tibshirani, and J. Friedman. A study of error variance estimation in lasso regression. *Statistica Sinica*, pages 35–67, 2016.

[10] T. Sun and C.-H. Zhang. Scaled sparse linear regression. *Biometrika*, 99(4):879–898, 2012.

[11] T. Wang, Z. Ren, Y. Ding, Z. Fang, Z. Sun, M. L. MacDonald, R. A. Sweet, J. Wang, and W. Chen. Fastggm: an efficient algorithm for the inference of gaussian graphical model in biological networks. *PLoS computational biology*, 12(2):e1004755, 2016.

[12] C.-H. Zhang and S. S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(1):217–242, 2014.

[13] H. Zhang, Y. Zheng, Z. Zhang, T. Gao, B. Joyce, G. Yoon, W. Zhang, J. Schwartz, A. Just, E. Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[14] Q. Zhang. High-dimensional mediation analysis with applications to causal gene identification. *Statistics in Biosciences*, pages 1–20, 2021.

[15] R. R. Zhou, L. Wang, and S. D. Zhao. Estimation and inference for the indirect effect in high-dimensional linear mediation models. *Biometrika*, 107(3):573–589, 2020.

[16] S. Zhou, S. van de Geer, and P. Bühlmann. Adaptive lasso for high dimensional regression and gaussian graphical modeling. *arXiv preprint arXiv:0903.2515*, 2009.

Web Table 1: Average computational time (in seconds) for the simulation settings $(q, p_m, r, \tau) = (5, 1, 0, 0.5)$ and $(400, 5, 0, 0.5)$

(q, p_m, r, τ)	$(5, 1, 0, 0.5)$						
Method	MedDiC	GLLZ	HIMA	HIMA2	ZWZ	MedR L1	MedR L2
Average time	2.8	1.1	68.1	159.5	365.4	2.5×10^3	1.6×10^4

(q, p_m, r, τ)	$(400, 5, 0, 0.5)$		
Method	MedDiC	GLLZ AMS	GLLZ US
Average time	29.6	1.5	471.4

Web Table 2: Results of Maize data: Number of SNPs selected with each type of effect and from each analysis at $FDR = 0.2$.

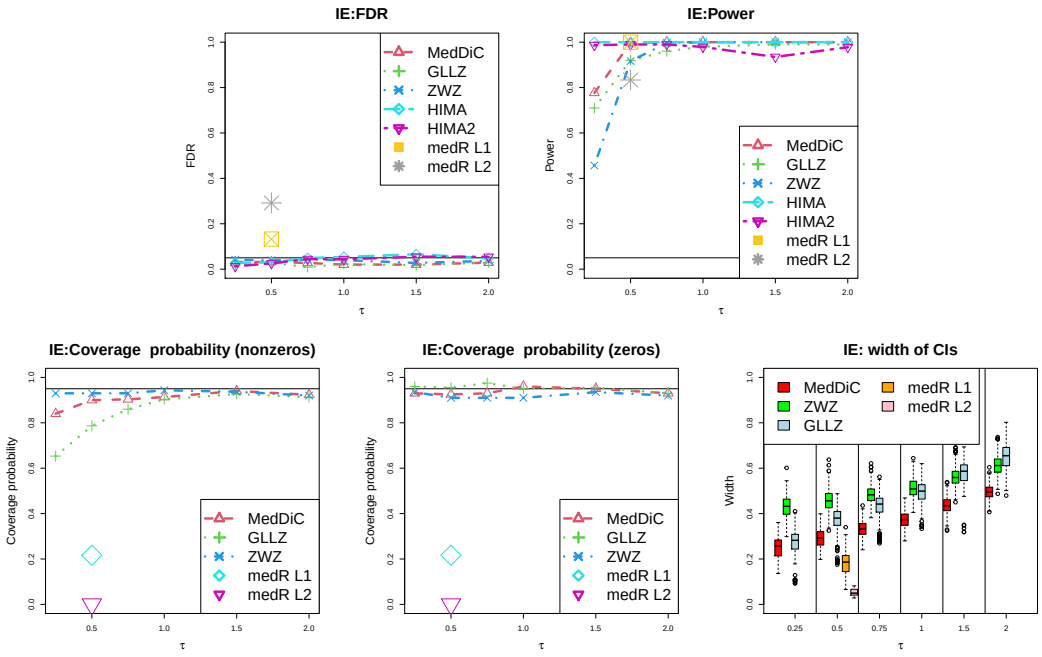
	DS	DT	GDS	GDT
Indirect Effect (MedDiC)	233	551	50	48
Direct Effect (MedDiC)	11	1	31	50
Total Effect (MedDiC)	25	62	37	46
Indirect Effect (GLLZ-US)	0	0	618	721
Direct Effect (GLLZ-US)	1915	2175	2266	1953
Indirect Effect (MedR L1)	0	0	0	0
Direct Effect (MedR L1)	0	0	0	0
Total Effect (MedR L1)	0	0	0	0

Web Table 3: Results of Maize data: Number of iSNPs (top 100) with HiChIP connections to the candidate mediator genes.

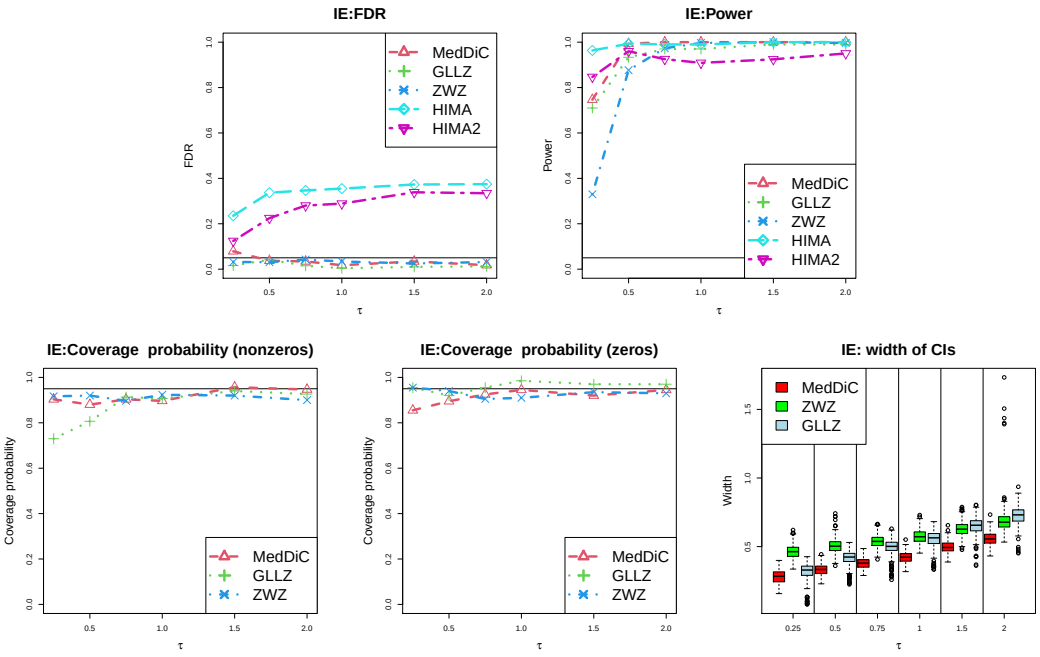
	DS	DT	GDS	GDT
Indirect Effect (MedDiC)	6	4	6	6
Indirect Effect (GLLZ US)	3	3	3	3
Indirect Effect (MedR L1)	9	4	9	7

Web Table 4: Results of Mouse data: Number of cis-regulated genes selected with each type of effect from each analysis, and their overlaps across phenotypes.

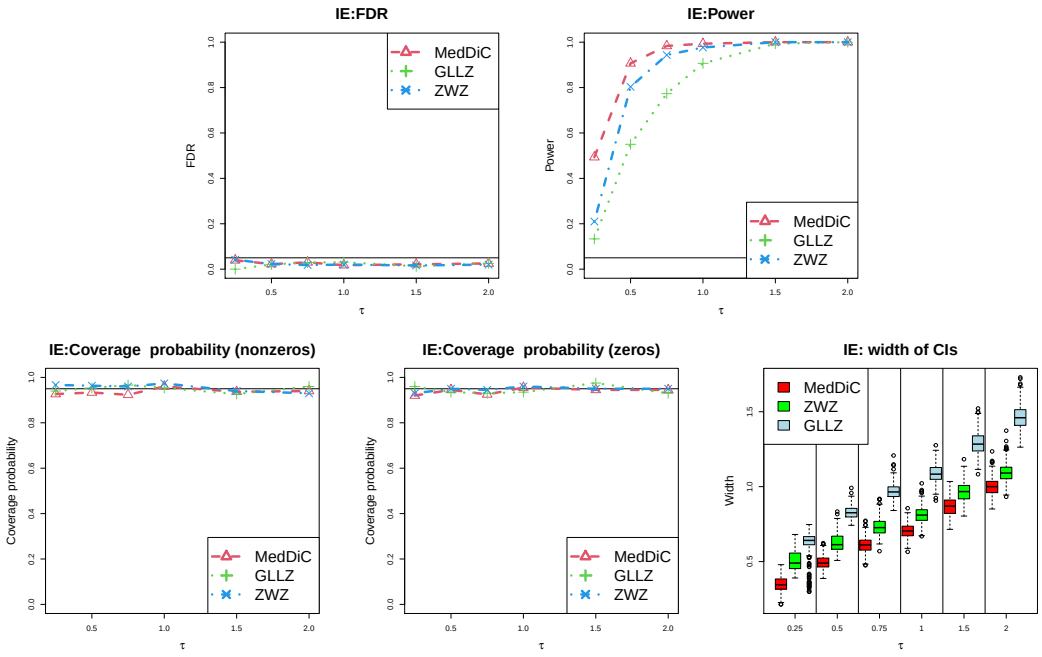
	10wk	10wkRep	rbm	10wk, 10wkRep	10wk, rbm	10wkRep, rbm
Indirect Effect (MedDiC)	40	46	44	30	21	24
Direct Effect (MedDiC)	2	3	5	2	1	1
Total Effect(MedDiC)	22	14	24	12	15	10
Indirect Effect (ZWZ)	13	31	16	12	10	14
Direct Effect (ZWZ)	24	13	15	12	14	11
Indirect Effect (GLLZ)	0	0	0	0	0	0
Direct Effect (GLLZ)	1	1	2	1	1	1
Indirect Effect (MedR L1)	20	15	20	14	16	14
Direct Effect (MedR L1)	0	0	0	0	0	0
Total Effect (MedR L1)	20	15	20	14	16	14



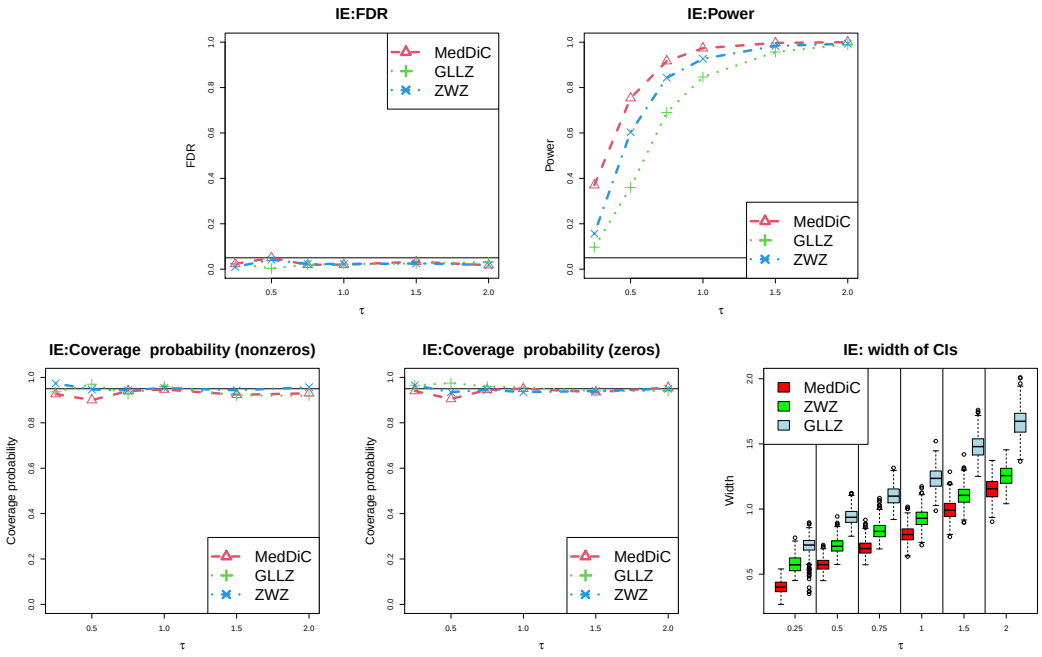
Web Figure 1: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



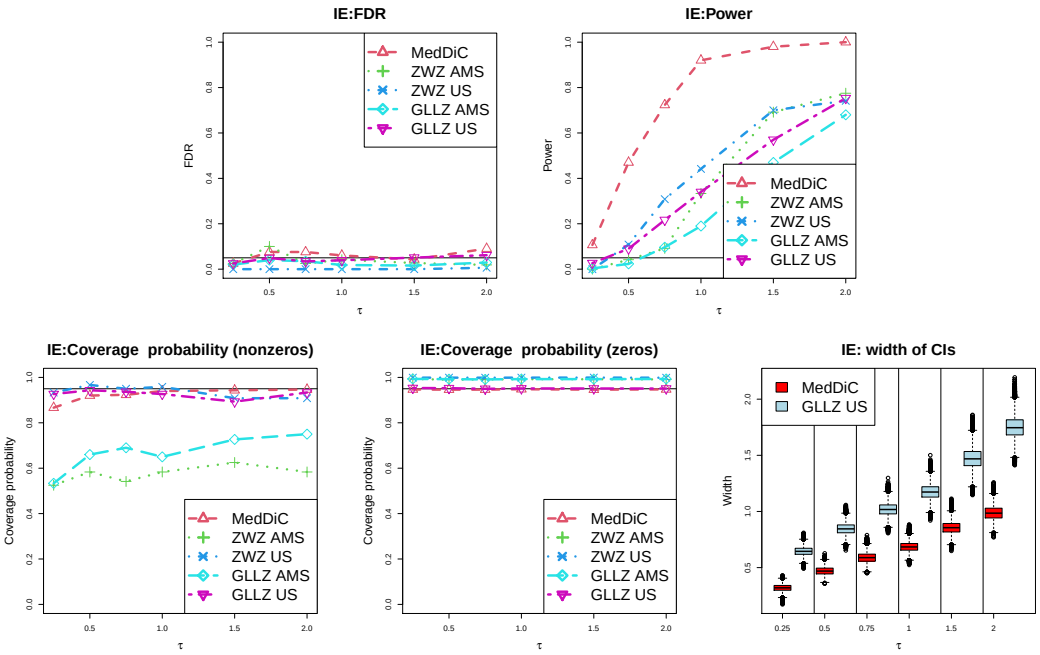
Web Figure 2: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



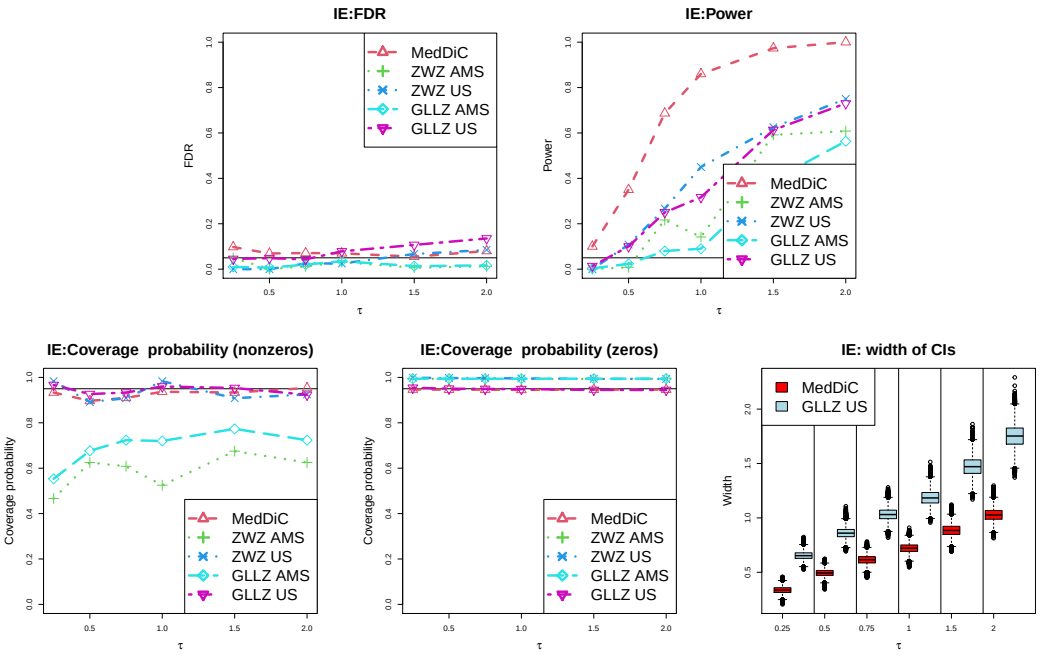
Web Figure 3: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



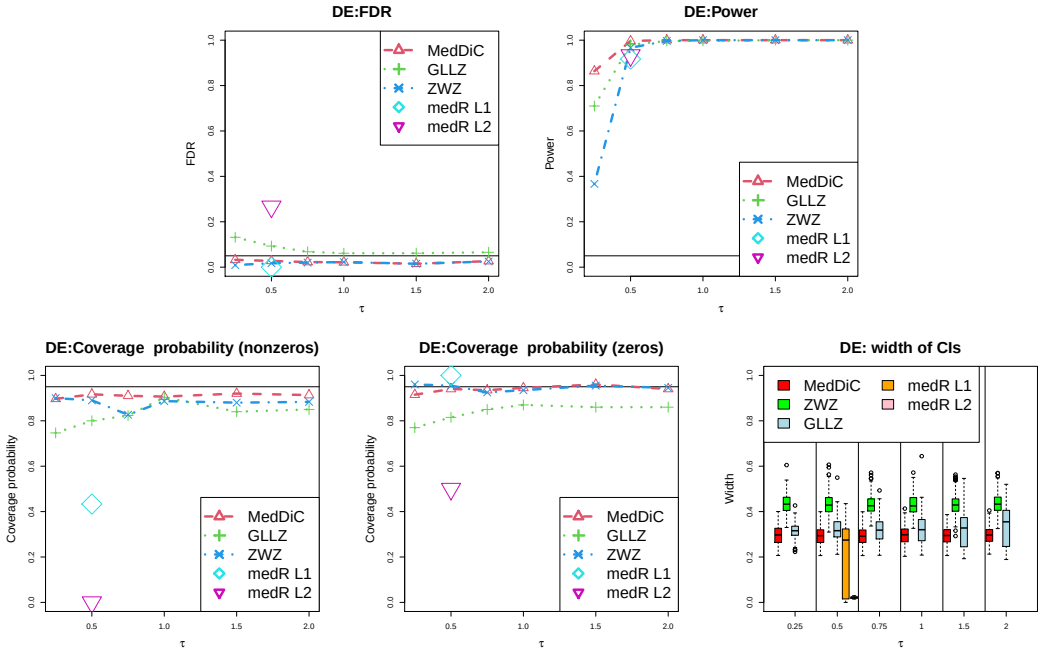
Web Figure 4: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



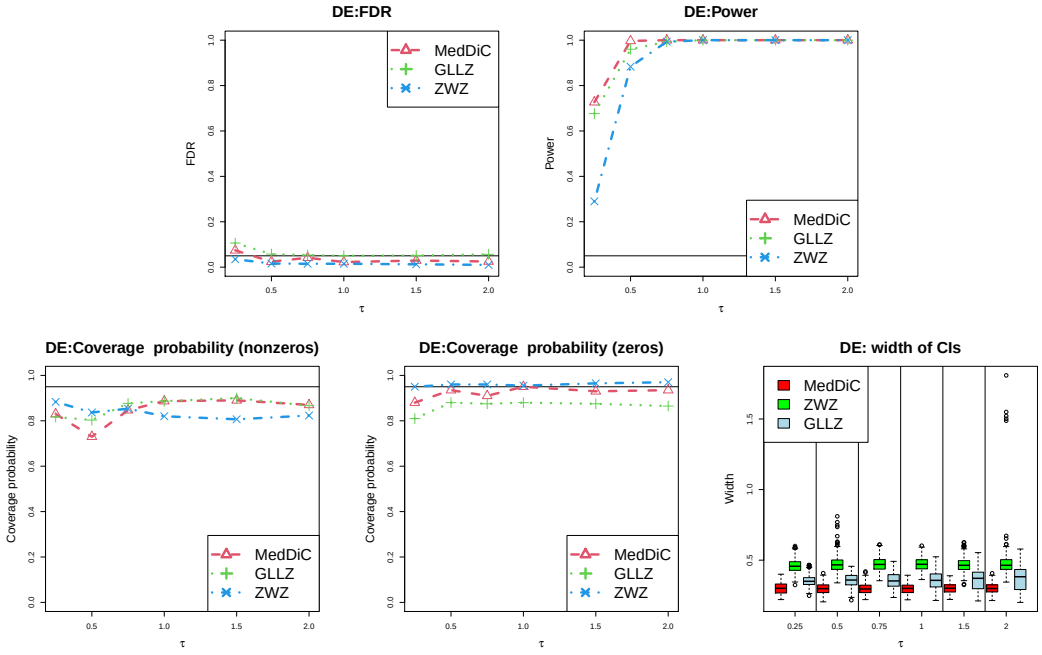
Web Figure 5: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



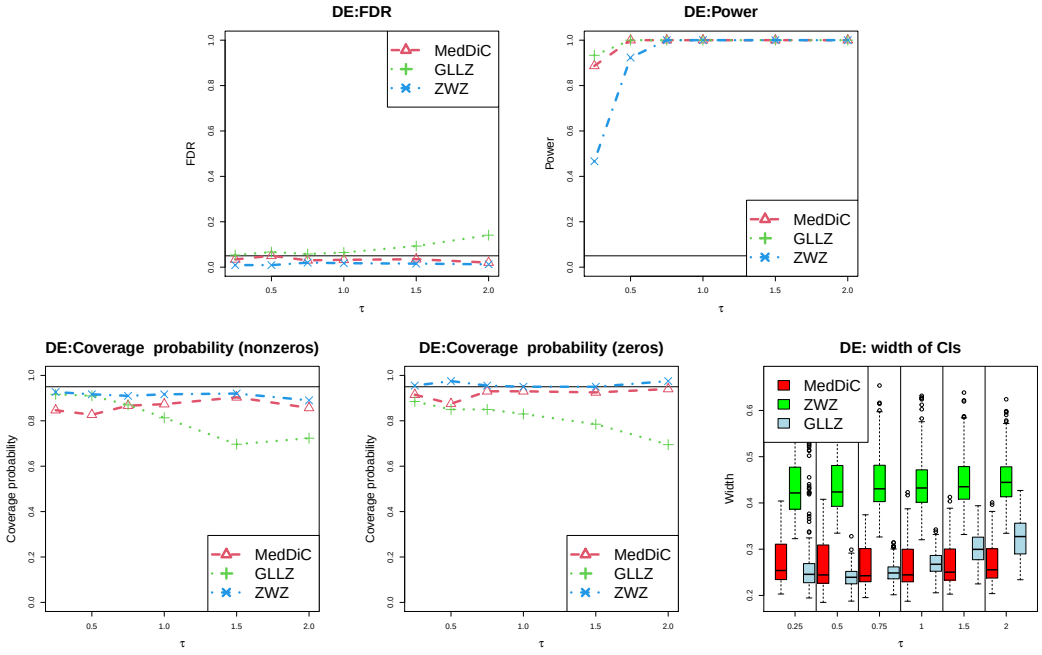
Web Figure 6: Inference results for IE (indirect effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with IE (indirect effect), the marginal coverage probabilities of the confidence intervals for the true nonzero IE (lower left) and the exposures with zero IE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



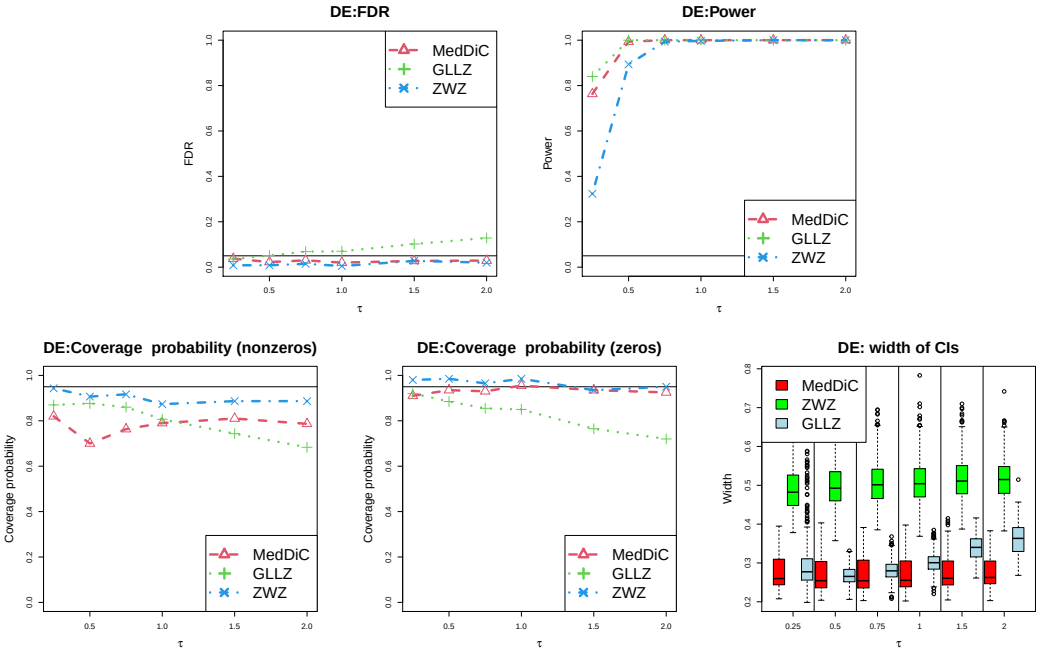
Web Figure 7: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



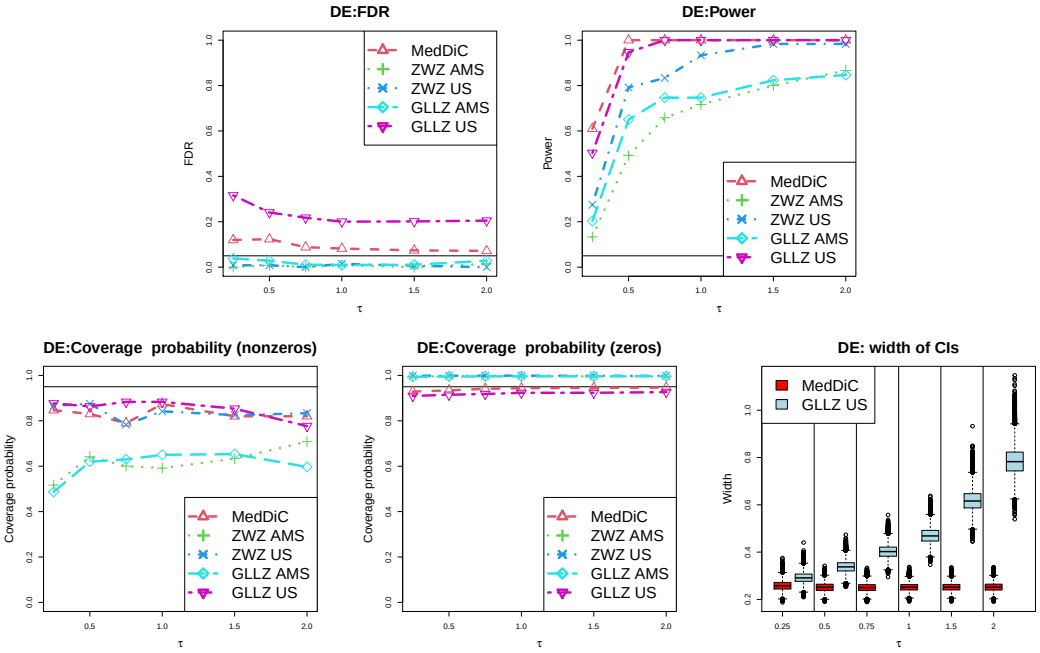
Web Figure 8: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



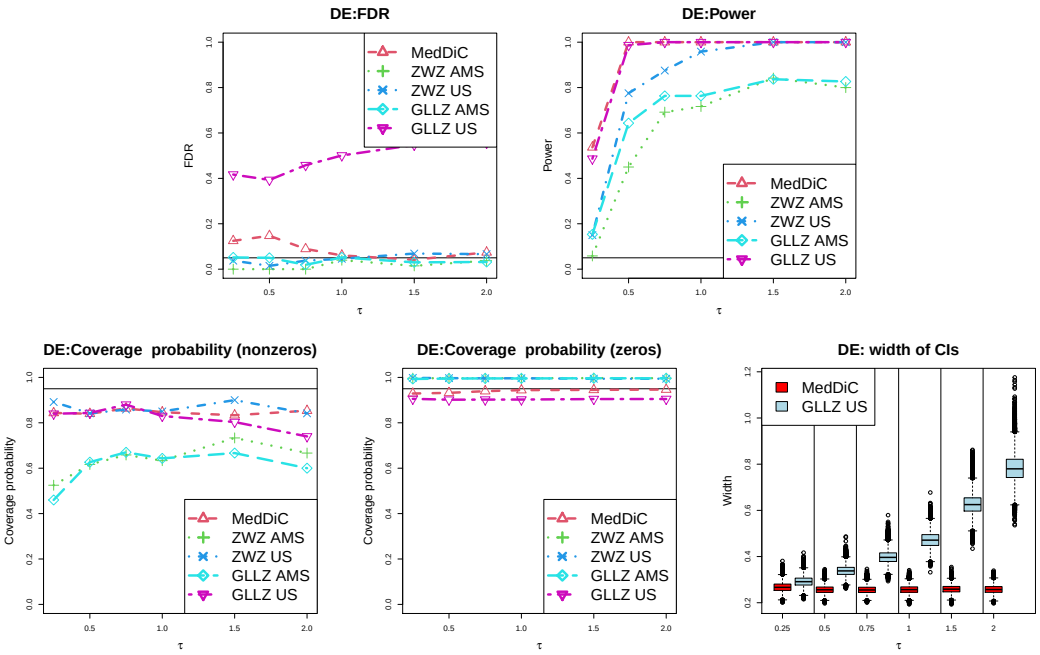
Web Figure 9: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



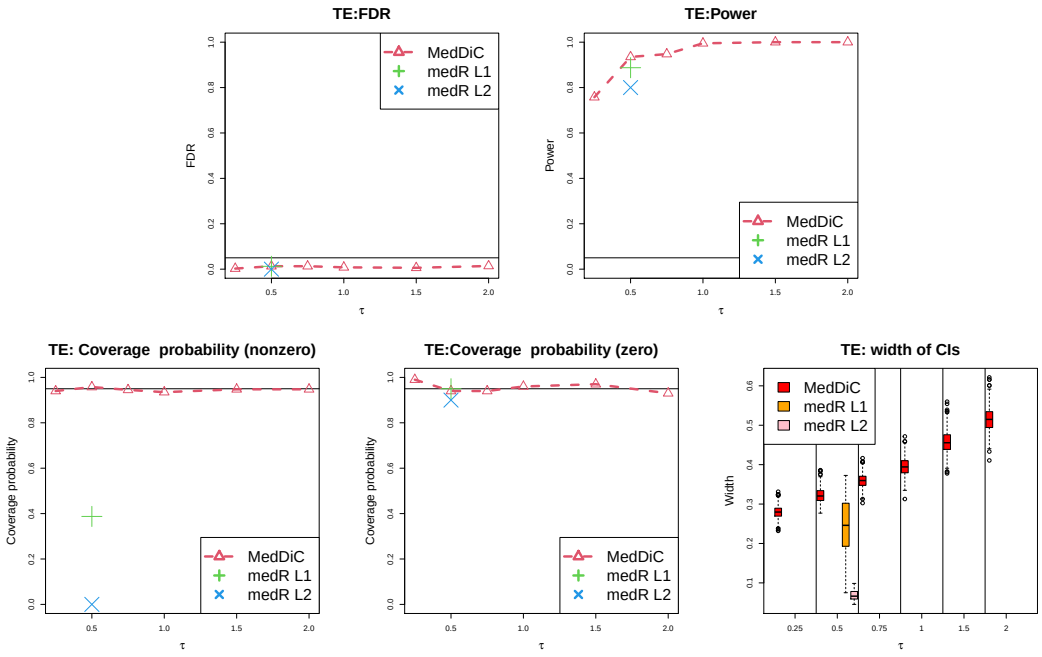
Web Figure 10: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



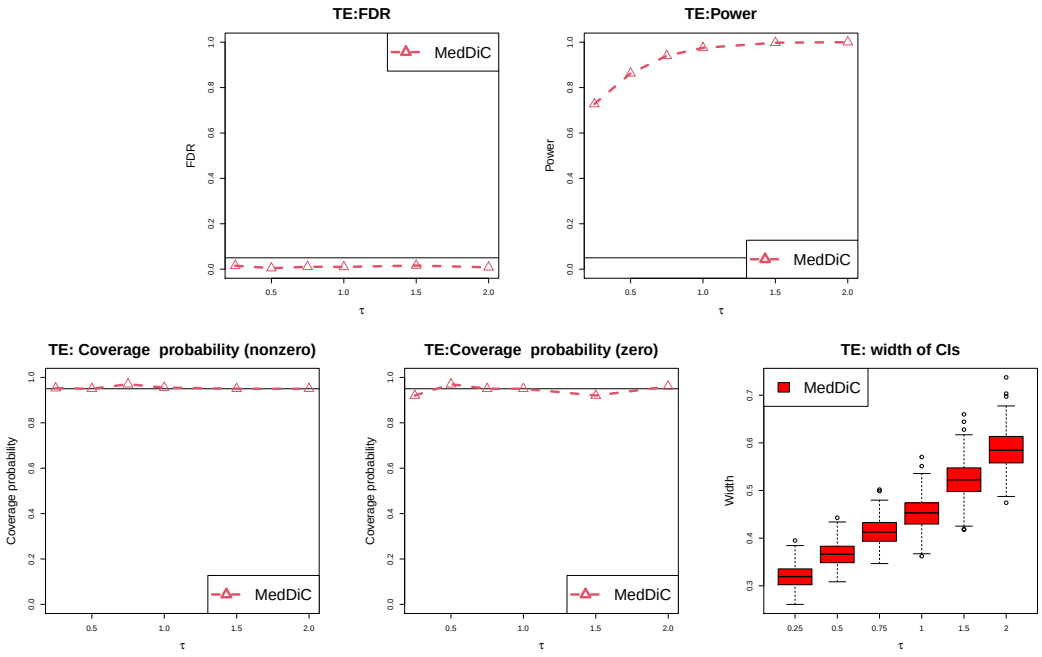
Web Figure 11: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



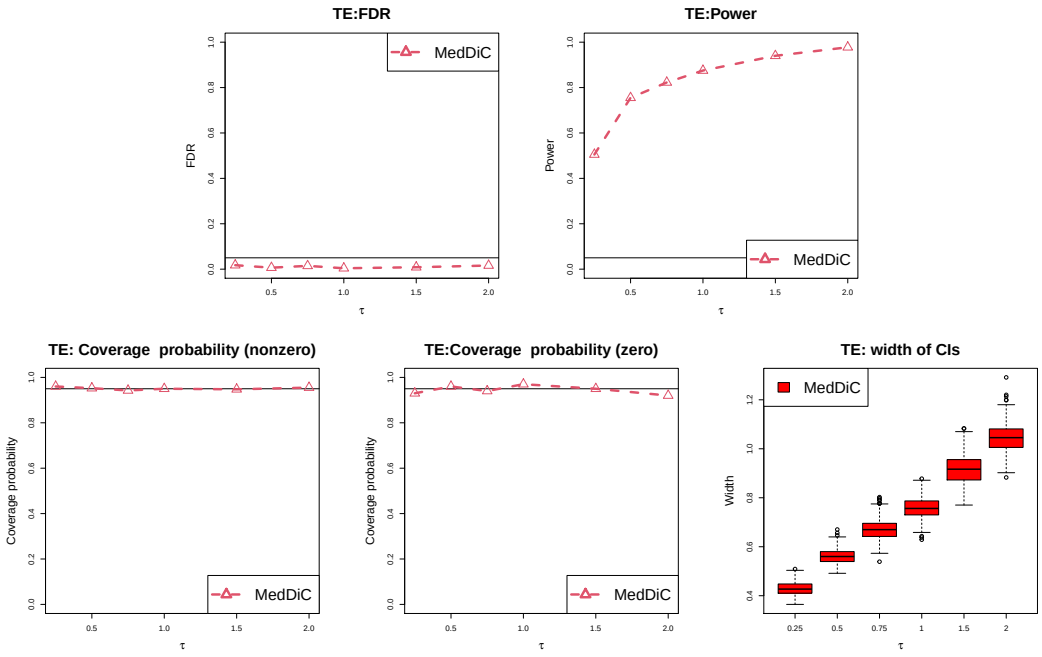
Web Figure 12: Inference results for DE (direct effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with DE (direct effect), the marginal coverage probabilities of the confidence intervals for the true nonzero DE (lower left) and the exposures with zero DE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



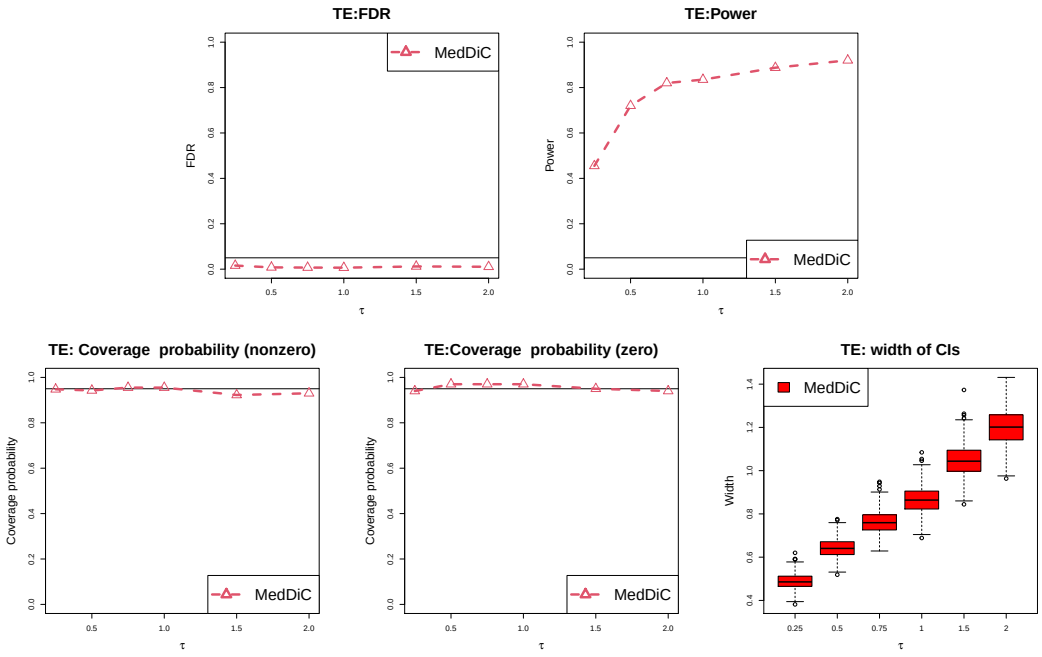
Web Figure 13: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



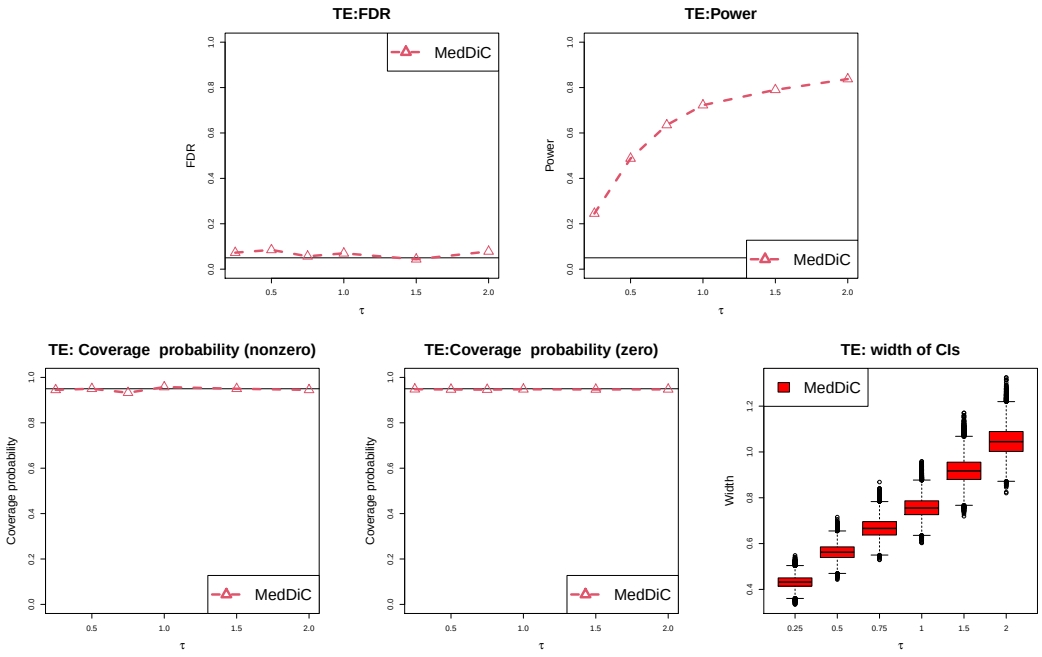
Web Figure 14: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there is one true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



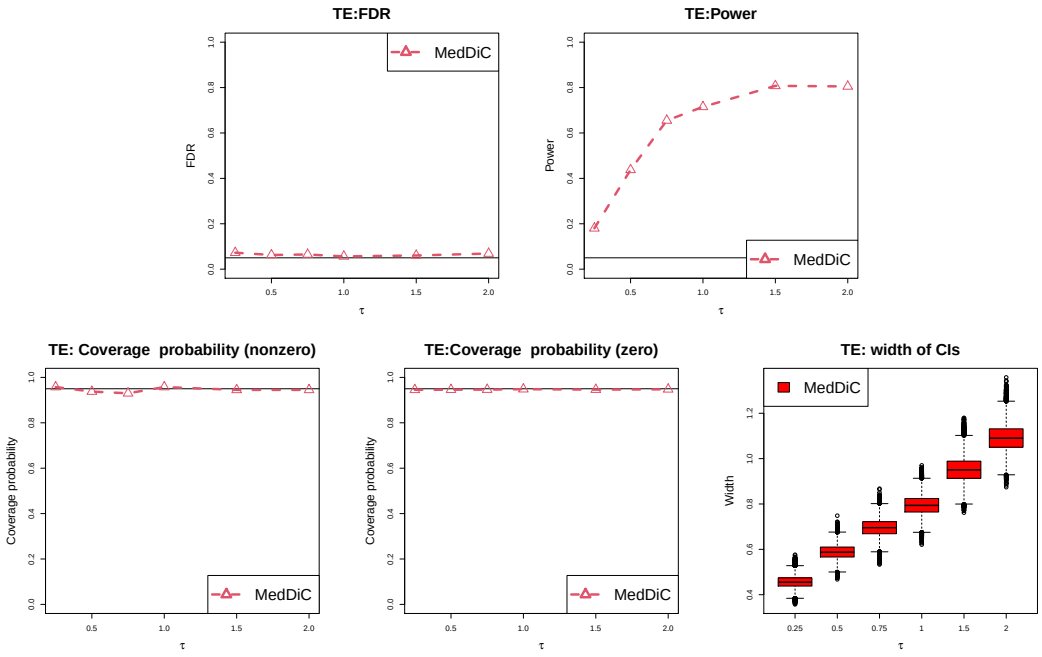
Web Figure 15: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



Web Figure 16: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are low dimensional $q = 5$, and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).



Web Figure 17: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are independent ($r = 0$).



Web Figure 18: Inference results for TE (total effects): Empirical FDR (upper left) and Power (upper right) of detecting exposures with TE (total effect), the marginal coverage probabilities of the confidence intervals for the true nonzero TE (lower left) and the exposures with zero TE (lower middle), and the widths of all confidence intervals (lower right). The exposures are high dimensional($q = 400$), and there are five true mediators among the $p = 500$ candidate mediators that are included in the model. The exposures are correlated ($r = 0.4$).