PLOS Computational Biology

METHODS

Powerful large scale inference in high dimensional mediation analysis

Asmita Roy[1]*, **Xianyang Zhang**[2]

1 Department of Biostatistics/Bloomberg School of Public Health, Johns Hopkins University, Baltimore, Maryland, United States of America, **2** Department of Statistics, Texas A&M University, College Station, Texas, United States of America

* aroy38@jh.edu

Abstract

In genome-wide epigenetic studies, determining how exposures (e.g., Single Nucleotide Polymorphisms) affect outcomes (e.g., gene expression) through intermediate variables, such as DNA methylation, is a key challenge. Mediation analysis provides a framework to identify these causal pathways; however, testing for mediation effects involves a complex composite null hypothesis. Existing methods, such as Sobel's test or the Max-P test, are often underpowered in this context because they rely on null distributions determined under only a subset of the null space and are not optimized for the multiple testing burden inherent in high-dimensional data. To address these limitations, we introduce MLFDR (Mediation Analysis using Local False Discovery Rates), a novel method for high-dimensional mediation analysis. MLFDR leverages local false discovery rates, calculated from the coefficients of structural equation models, to construct an optimal rejection region. We demonstrate theoretically and through simulation that MLFDR asymptotically controls the false discovery rate and achieves superior statistical power compared to recent high-dimensional mediation methods. In real data applications, MLFDR identified 20%–50% more significant mediators than existing methods, demonstrating its ability to uncover biological signals missed by conventional approaches.

Author summary

The paper presents a novel approach to high-dimensional mediation analysis through a local false discovery rate (MLFDR) screening algorithm. It addresses the limitations of traditional methods like Sobel's test and maxP, which are underpowered in high dimensional setting. We extend local FDR principles to composite null hypotheses, and derive a screening rule with a closed-form expression for false discovery proportion. We also show that MLFDR has comparable or better

OPEN ACCESS

Citation: Roy A, Zhang X (2026) Powerful large scale inference in high dimensional mediation analysis. PLoS Comput Biol 22(1): e1013880. https://doi.org/10.1371/journal.pcbi.1013880

Editor: Aakrosh Ratan, University of Virginia, UNITED STATES OF AMERICA

Received: October 3, 2024

Accepted: December 29, 2025

Published: January 14, 2026

Data availability statement: R code for simulations and real data analysis may be obtained at https://github.com/asmita112358/MLFDR as well as the package MLFDR in CRAN.

Funding: This work was supported by National Institute of Health R01GM144351 (Roy & Zhang).

Competing interests: The authors have declared that no competing interests exist.

results than two recently-proposed methods, MDACT [30], HDMT [3] across a wide range of data types and models. We also provide a two-step global latent factor adjustment using surrogate variable analysis [9].

1 Introduction

Mediation analysis serves as a critical tool for deciphering the biological mechanisms underlying genetic associations with diseases identified in Genome-Wide Association Studies (GWAS). By bridging the gap between genetic variants and clinical outcomes, mediation analysis reveals intermediate pathways and elucidates causal relationships. As GWAS continues to uncover a vast number of genetic associations, translating these findings into actionable insights for precision medicine and therapeutic development becomes increasingly important. For instance, cigarette smoking is known to alter DNA methylation and gene expression [12]; concurrently, DNA methylation often regulates gene expression directly [4,14]. Investigating the mediating effect of DNA methylation on gene expression—particularly in the presence of environmental exposures like smoking—is therefore essential. However, these analyses are complicated by high-dimensional outcomes and clinical confounders, such as patient age, which influences both gene expression and DNA methylation heterogeneity [7,19]. This article addresses the statistical challenges inherent in such high-dimensional mediation problems.

Historically, [1] introduced the regression-based definition of mediation analysis, often referred to as the "product of coefficients method," which examines the significance of the product of the exposure-mediator and mediator-outcome coefficients. More recently, the literature has expanded through the "counterfactual framework" [8,15,17,23–26], which provides a causal interpretation for natural direct and indirect effects across various models, including those with non-linearities and binary or survival outcomes.

Let X denote the exposure, M_i the ith mediator, and Y the outcome. Under the product of coefficients approach, mediation analysis tests the null hypothesis $H_{0,i}$: $\alpha_i\beta_i = 0$, where α_i represents the effect of X on M_i, and β_i represents the effect of M_i on Y. This creates a composite null hypothesis comprising three distinct cases: (i) $\alpha_i = 0, \beta_i \neq 0$; (ii) $\alpha_i \neq 0, \beta_i = 0$; or (iii) $\alpha_i = 0, \beta_i = 0$. Assuming no unmeasured confounders, classical tests like Max-P [13] and Sobel's test [18] are known to be conservative under case (iii), as statistical inference is typically derived from distributions determined by cases (i) and (ii). In genome-wide studies, however, the sparse nature of omics data implies that $\alpha_i = 0$ and $\beta_i = 0$ hold for the majority of markers. Recent methods such as JS-mixture (HDMT) [3] and DACT [11] attempt to address this by explicitly modeling the composite nature of the null. JS-mixture improves power by using a mixture-null distribution of maximum p-values, adapting [20]'s procedure to estimate component proportions. DACT estimates the proportions of null α_i and β_i separately to combine case-specific p-values. However, [30] recently demonstrated

that DACT suffers from False Discovery Rate (FDR) inflation under dense alternatives and proposed a modified version (MDACT) that computes the statistic's distribution via numerical integration to improve p-value accuracy.

While JS-mixture and MDACT offer improvements over classical methods, they are not theoretically optimal regarding power. The FDR literature is broadly divided into p-value-based and local FDR-based rejection regions. Local FDR, a Bayesian approach, ranks hypotheses by the posterior probability that a case is null given the observed statistics; this ranking often differs from that based on p-values. [22] demonstrated that, except in cases of symmetric alternatives, local FDR and p-value-based orderings diverge. Furthermore, [22] proved that the local FDR-based oracle procedure is optimal: among all methods controlling the marginal FDR (mFDR), the local FDR approach yields the highest number of rejections. While the power advantage is negligible for symmetric alternatives, it becomes significant when the alternative distribution is asymmetric. Motivated by these theoretical properties, we propose MLFDR, a local FDR-based screening algorithm designed specifically for high-dimensional mediation analysis. Our contributions to the literature are as follows:

1. We extend the concept of local FDR to the composite null hypothesis setting, deriving a screening rule with a closed-form expression for the corresponding false discovery proportion (FDP).
2. We validate the method across a diverse array of data types—including continuous and binary variables, and scenarios with exposure-mediator interactions—demonstrating robust performance across various model specifications. We specifically incorporate Surrogate Variable Analysis (SVA) to adjust for latent confounding and illustrate the method's efficacy in multiple mediator setups with univariate or clinical outcomes.
3. MLFDR offers optimal power improvement over existing methods while maintaining asymptotic FDR control. Extensive simulations confirm its superiority over MDACT and HDMT in terms of power and error rate control.
4. We provide theoretical guarantees for the identifiability and global optimality of our model under relatively mild assumptions, proving FDR control for both the oracle and adaptive procedures.

The remainder of this paper is organized as follows. Sect 2 contains the main results. Specifically, Sect 2.1 outlines the screening procedure for detecting significant mediators. Sect 2.2 presents simulation studies. Sect 2.3 discusses extensions of MLFDR to composite alternatives and latent factor models, which can account for unmeasured confounding and pleiotropy. Sect 3 provides an in-depth analysis of Prostate Cancer data and Lung cancer data from The Cancer Genome Atlas (TCGA), exploring SNP-CpG-gene expression pathways and causal pathways between smoking habits and gene expression, respectively. Sect 5 details the methodology. An R package implementing the method is available at https://github.com/asmita112358/MLFDR as well as in CRAN. Theorems proving the large-sample FDR control of MLFDR are provided in Section E of S1 Text.

Finally, we distinguish our approach from other recent efforts in high-dimensional mediation analysis. [33], [32], and [16] address the multiple mediator problem specifically for survival outcomes. Closer to our framework, [21] and [5] utilize local FDR-based rejection regions; the former approximates the alternative as a mixture of Gaussian distributions, while the latter constructs regions based on p-values. Our work advances this domain in two specific aspects: (i) we incorporate a general prior for the coefficients α and β to estimate the *exact* posterior density for computing the local FDR, rather than relying on approximations; and (ii) we offer theoretical guarantees for the local FDR estimates obtained via the EM algorithm, a property not previously established in this context.

2 Results

2.1 Method overview

This section outlines the workflow of MLFDR; a schematic representation of the framework is provided in Fig 1. Consider a study involving n independent samples. For each testing unit $i = 1, \ldots, m$, we observe an exposure variable X_i, a mediator M_i, and an outcome Y_i. Biologically, these variables may represent distinct contexts: for example, X may denote a patient's smoking history (shared across i), with $\{M_i\}_{i=1}^m$ representing CpG methylation sites and $\{Y_i\}_{i=1}^m$ representing gene

expression levels. Alternatively, the analysis may focus on the functional impact of Single Nucleotide Polymorphisms (X_i) on gene expression (Y_i) as mediated by CpG methylation (M_i) [3].

The mediation model posits that the exposure X_i influences the outcome Y_i through the intermediate variable M_i, rather than solely through a direct relationship. We denote the coefficient for the exposure-mediator relationship ($X_i \to M_i$) as α_i, and the coefficient for the mediator-outcome relationship ($M_i \to Y_i$) as β_i. In Fig 1, solid arrows indicate these direct effects.

We aim to test the composite null hypothesis against the alternative for each unit i:

$$H_{0,i} : \alpha_i\beta_i = 0 \quad \text{versus} \quad H_{1,i} : \alpha_i\beta_i \neq 0, \quad i = 1, 2, \ldots, m. \tag{1}$$

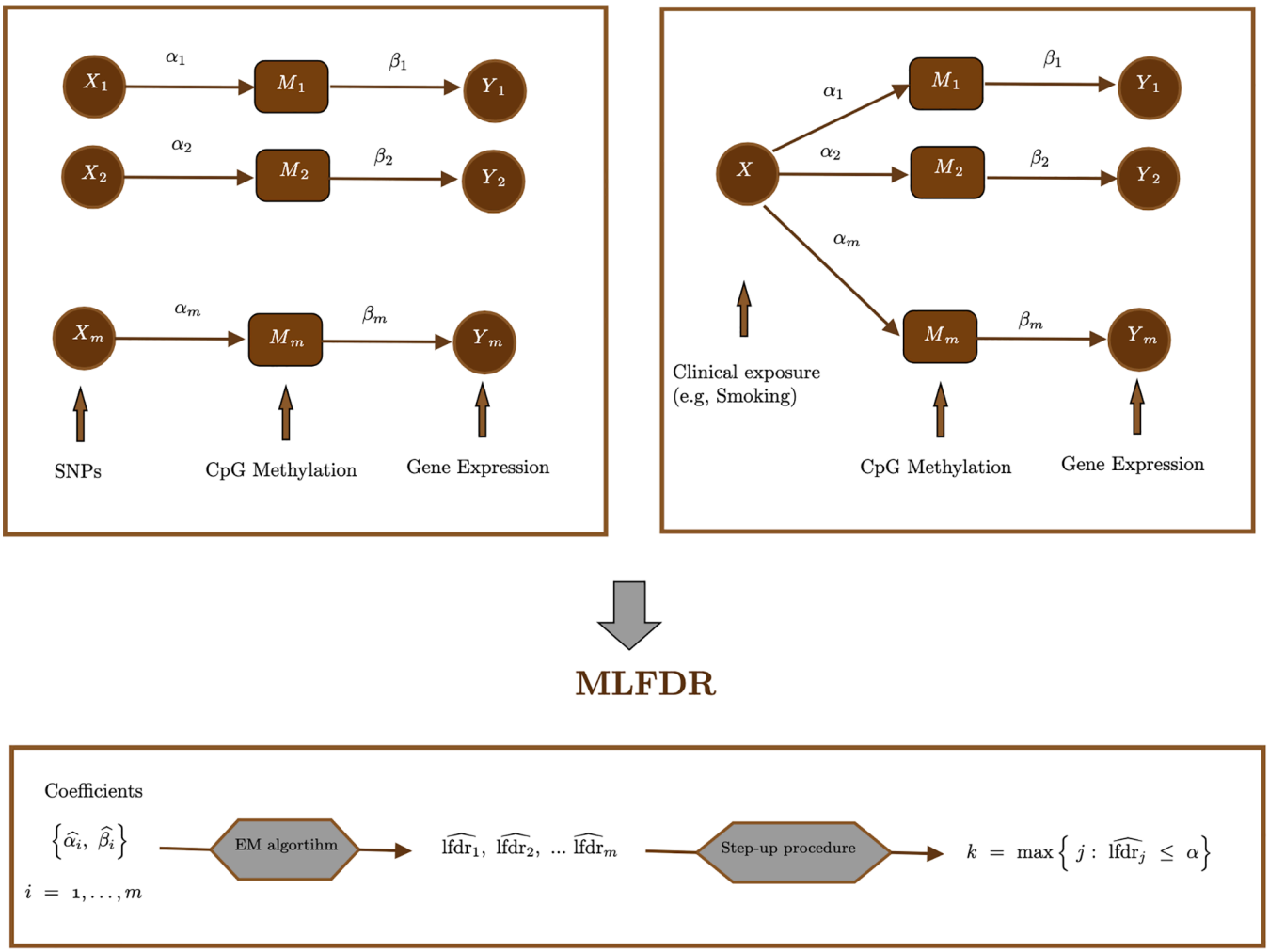

Fig 1. Schematic diagram of MLFDR.

https://doi.org/10.1371/journal.pcbi.1013880.g001

The composite null hypothesis $H_{0,i}$ can be decomposed into three disjoint component nulls, $H_{0,i} = H_{00,i} \cup H_{01,i} \cup H_{10,i}$, defined as:

$$\begin{aligned} H_{00,i} &: \alpha_i = 0 \text{ and } \beta_i = 0, \\ H_{10,i} &: \alpha_i \neq 0 \text{ and } \beta_i = 0, \\ H_{01,i} &: \alpha_i = 0 \text{ and } \beta_i \neq 0, \end{aligned} \tag{2}$$

for $i = 1, 2, \ldots, m$.

We consider a mixture prior for (α_i, β_i), where the probability of each disjoint component nulls H_{00}, H_{10} and H_{01} occur with probability π_{00}, π_{10} and π_{01} respectively. The marginal prior distributions of α_i and β_i, respectively, are degenerate zero under the null and follow a normal prior with an unknown mean and variance under the alternative. The marginal distribution of the least squares coefficient estimates $\{\hat{\alpha}_i, \hat{\beta}_i\}$ given the latent states is computed, and the unknown parameters including the null proportions are estimated using EM algorithm.

Using these estimates, we compute the local false discovery rate (local FDR) for each coefficient pair, denoted as $\widehat{\text{lfdr}}_i$ for $i = 1, \ldots, m$. As the local FDR represents the posterior probability that the i-th hypothesis is null given the observed statistics, a lower value indicates stronger evidence against the null. Consequently, we define the rejection region for the composite null hypothesis as $\widehat{\text{lfdr}}_i \leq \delta$. The threshold δ is determined adaptively using the step-up procedure proposed by [22]. The complete algorithm is detailed in the Methods section.

Additionally, we introduce an extended algorithm (MLFDR2) which can deal with scenarios where the marginal priors of α_i and β_i follow mixture normal distributions under the alternative. A composite alternative leads to a joint distribution of $\{\hat{\alpha}_i, \hat{\beta}_i\}$ with more than 4 mixture components, which can often be computationally burdensome. We introduce a two-step EM alogrithm that estimates the parameters of the marginal distributions of $\{\hat{\alpha}_i\}$ and $\{\hat{\beta}_i\}$ in the first step, then uses these estimates to run another EM algorithm that computes the probabilities of each mixture component.

We also discuss another extension using Surrogate Variable Analysis [10] which can account for unmeasured confounders and pleiotropy in the model. Details are presented in Sect 2.3.

2.2 Simulation studies

We evaluate the performance of MLFDR through extensive simulations under two distinct mixture proportion scenarios: a *dense* alternative and a *sparse* alternative. Following the setups in [3], the latent class probabilities are defined as:

- **Dense alternative:** $(\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}) = (0.4, 0.2, 0.2, 0.2)$.
- **Sparse alternative:** $(\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}) = (0.88, 0.05, 0.05, 0.02)$.

We consider sample sizes of $n \in \{100, 300\}$ and fix the number of mediators at $m = 1000$. The parameter controlling the signal strength of mediation, τ, varies from 0.1 to 1.9 in increments of 0.2. The non-zero coefficients are generated as $\alpha_i = 0.05\tau + h_i$ (under H_{10} and H_{11}) and $\beta_i = -0.5\tau + g_i$ (under H_{01} and H_{11}), where the noise terms follow $h_i \sim N(0, 1/n)$ and $g_i \sim N(0, 4/n)$.

We compare the empirical FDR and power of MLFDR against two competing methods: MDACT [30] and HDMT (JS-mixture) [3]. The simulation settings are detailed below.

1. **Linear Model.** The exposure is univariate with $X \sim \text{Ber}(0.1)$.

$$\begin{aligned} M_i &= X\alpha_i + e_i, \\ Y_i &= M_i\beta_i + X\gamma_i + \epsilon_i, \end{aligned} \tag{3}$$

where $\gamma_i \sim N(1, 0.5)$, and error terms $e_i, \epsilon_i \overset{i.i.d.}{\sim} N(0, 1)$. The results for this setting are summarized in Fig 2.

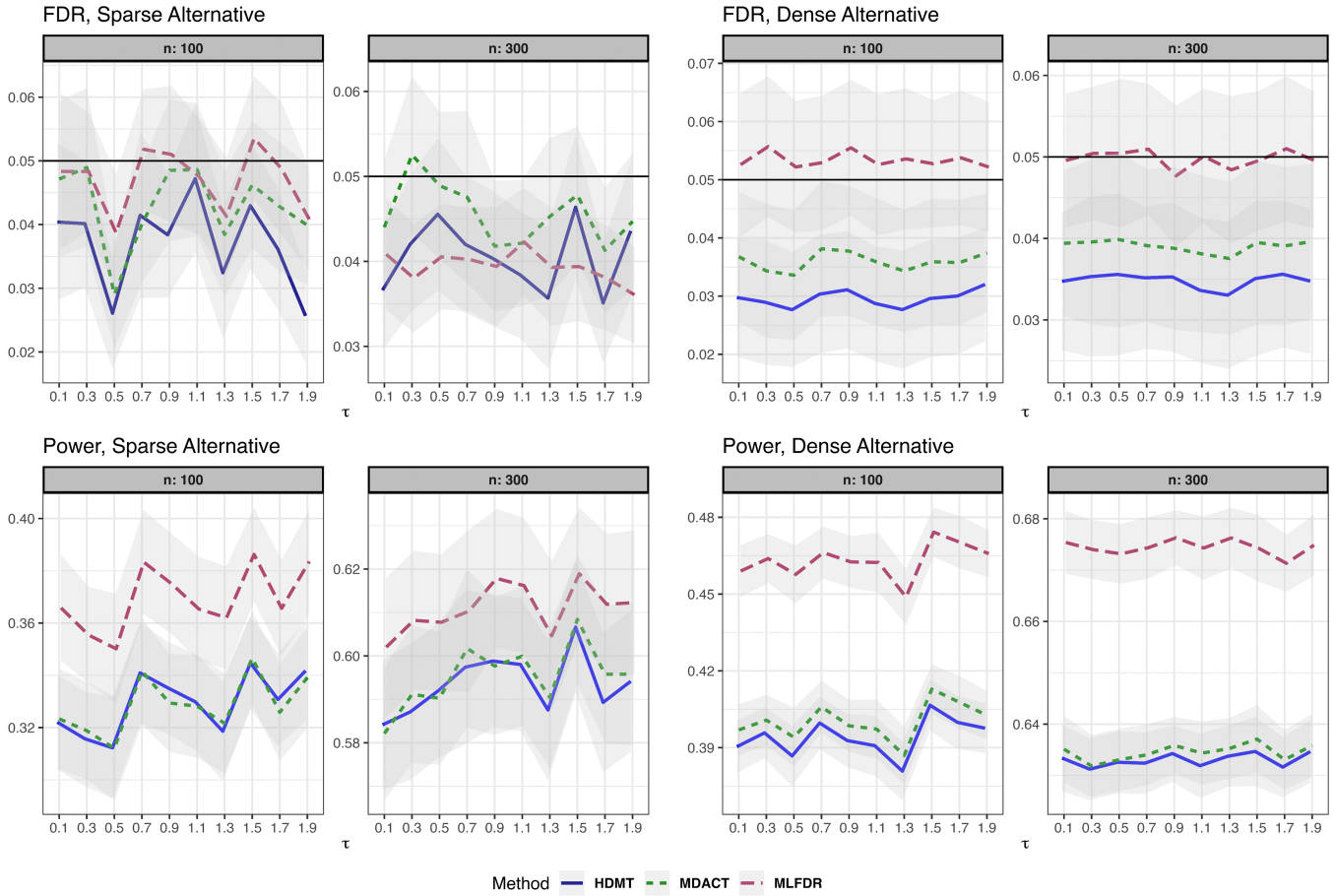

Fig 2. FDR and power comparison for the linear model (Setting 1). Results are displayed for both sparse and dense alternatives. Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g002

2. **Linear Model with measured Confounder.** The exposure is univariate with $X \sim \text{Ber}(0.1)$. We introduce a confounder $Z \sim N(0,1)$:

$$\begin{aligned} M_i &= X\alpha_i + \theta_i Z + e_i, \\ Y_i &= M_i\beta_i + X\gamma_i + \delta_i Z + \epsilon_i, \end{aligned} \tag{4}$$

   where the confounder effects are drawn independently from $\theta_i, \delta_i \sim U(0, 0.5)$. The results are presented in Fig 3.
3. **Binary Outcome.** The exposure is univariate with $X \sim \text{Ber}(0.1)$. The outcome Y_i is binary:

$$\begin{aligned} M_i &= X\alpha_i + e_i, \\ \text{logit}\{\mathbb{P}(Y_i = 1)\} &= M_i\beta_i + X\gamma_i. \end{aligned} \tag{5}$$

   The results are displayed in Fig 4.

Across all settings, the three methods demonstrated satisfactory FDR control. However, MLFDR consistently exhibited the highest power. Specifically, MLFDR achieved an average power improvement of 10.83% over MDACT and 12.23%

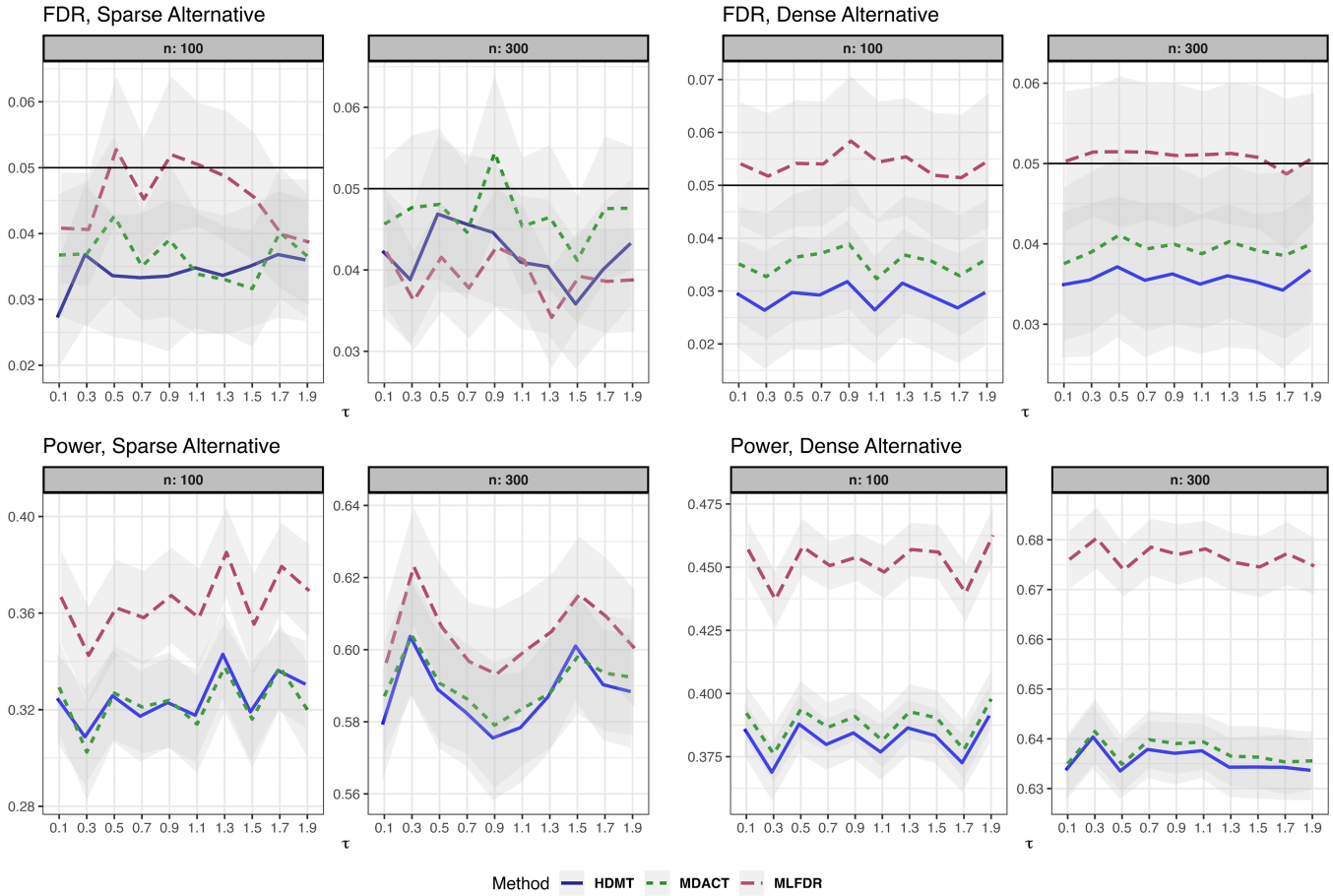

Fig 3. FDR and power comparison for the linear model with measured confounders (Setting 2). Results are displayed for both sparse and dense alternatives. Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g003

over HDMT under dense alternatives. Under sparse alternatives, MLFDR maintained its advantage with an average improvement of 7.47% over MDACT and 8.51% over HDMT.

2.3 Extensions

2.3.1 Latent factors. Unmeasured latent factors may be addressed by surrogate variable analysis [10]. Briefly, surrogate variable analysis considers the following model:

$$M_i = \mu_i + \alpha_i X + \phi_i Z + \sum_{l=1}^{L} \gamma_l^{(i)} g_l + e_i, \tag{6}$$

where X, Z are measured covariates, and $g_1, g_2, \ldots g_L$ are unmeasured latent factors. Surrogate variable analysis produces a set of K mutually orthogonal vectors $\hat{u}_1, \hat{u}_2, \ldots, \hat{u}_K$ (where $K \leq L$), which span the same linear space as the latent factors.

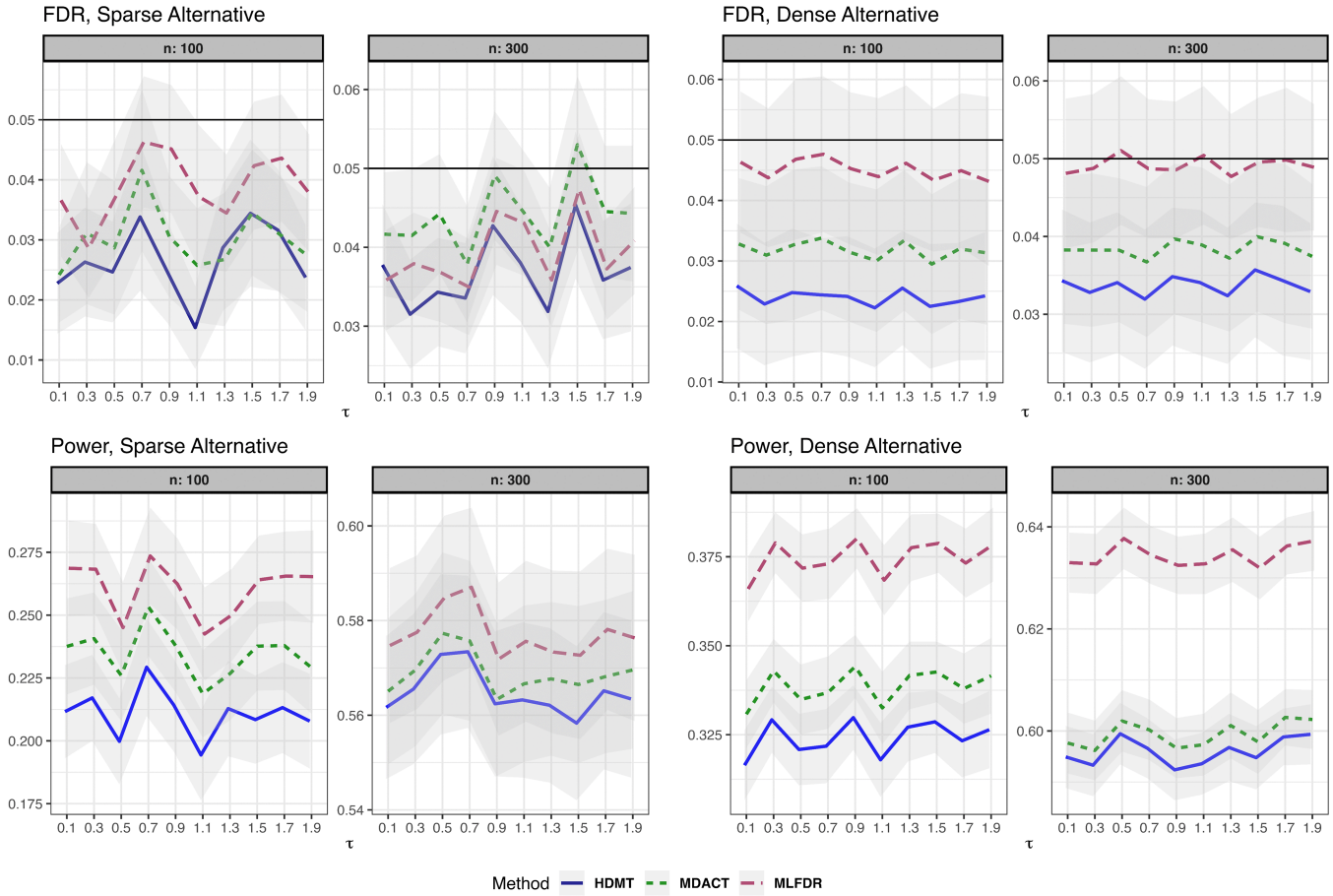

Fig 4. FDR and power comparison for the linear model with binary outcomes (Setting 3). Results are displayed for both sparse and dense alternatives. Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g004

Thus, the original equation may be re-written as:

$$M_i = \mu_i + \alpha_i X + \phi_i Z + \sum_{k=1}^{K} \lambda_k^{(i)} \hat{u}_k + e_i.$$

These estimated factors, collected into a matrix $\hat{\mathbf{U}}_M$, account for unmeasured confounding in the exposure-mediator relationship.

For the mediator-outcome relationship, the latent factors may be modeled as follows:

$$Y_i = \nu_i + \beta_i M_i + \gamma_i X + \delta_i Z + \sum_{l=1}^{L} \eta_l^{(i)} g_l + \epsilon_i. \quad (7)$$

In this setting, the latent terms may account for: 1) unmeasured confounding; 2) measured confounders with unknown relationships to the outcome (e.g., global batch effects); and 3) pleiotropy, where Y_i is influenced by mediators other than M_i.

Algorithm 1 Two-step global factor adjustment for high-dimensional mediation

Require: Outcome matrix $\mathbf{Y} \in \mathbb{R}^{n \times m}$, Mediator matrix $\mathbf{M} \in \mathbb{R}^{n \times m}$, Exposure vector $\mathbf{X} \in \mathbb{R}^{n \times 1}$, Covariates $\mathbf{Z} \in \mathbb{R}^{n \times q}$.

Ensure: Adjusted estimates for $\hat{\alpha}_i$ and $\hat{\beta}_i$ for $i = 1, \ldots, m$.

Step 1: Estimate Latent Factors for Mediators ($\hat{\mathbf{U}}_M$)

Regress $\mathbf{M}$ on $\mathbf{X}$ and $\mathbf{Z}$ to obtain the residual matrix $\hat{\mathbf{R}}_M$:

$$\hat{\mathbf{R}}_M = \mathbf{M} - \left(\mathbf{X}\hat{\mathbf{A}} + \mathbf{Z}\hat{\mathbf{\Phi}}\right)$$

Perform SVA on $\hat{\mathbf{R}}_M$ to extract the top k_M principal components:

$$\hat{\mathbf{U}}_M \leftarrow \mathrm{SVA}(\hat{\mathbf{R}}_M, k_M)$$

Step 2: Estimate Latent Factors for Outcomes ($\hat{\mathbf{U}}_Y$)

Regress $\mathbf{Y}$ on $\mathbf{X}$ and $\mathbf{Z}$ (excluding $\mathbf{M}$) to obtain the residual matrix $\hat{\mathbf{R}}_Y$:

$$\hat{\mathbf{R}}_Y = \mathbf{Y} - \left(\mathbf{X}\hat{\mathbf{\Gamma}} + \mathbf{Z}\hat{\mathbf{\Delta}}\right)$$

Perform SVA on $\hat{\mathbf{R}}_Y$ to extract the top k_Y principal components:

$$\hat{\mathbf{U}}_Y \leftarrow \mathrm{SVA}(\hat{\mathbf{R}}_Y, k_Y)$$

Step 3: Mediation Analysis with Global Adjustment

for $i = 1, \ldots, m$ **do**

Mediator Model: Regress M_i on $\mathbf{X}$, $\mathbf{Z}$, and $\hat{\mathbf{U}}_M$:

$$M_i = \mathbf{X}\alpha_i + \mathbf{Z}\phi_i + \hat{\mathbf{U}}_M\lambda_{M,i} + e_i$$

Outcome Model: Regress Y_i on M_i, $\mathbf{X}$, $\mathbf{Z}$, $\hat{\mathbf{U}}_M$, and $\hat{\mathbf{U}}_Y$:

$$Y_i = M_i\beta_i + \mathbf{X}\gamma_i + \mathbf{Z}\delta_i + \hat{\mathbf{U}}_M\lambda_{Y1,i} + \hat{\mathbf{U}}_Y\lambda_{Y2,i} + \epsilon_i$$

Extract coefficients $\hat{\alpha}_i, \hat{\beta}_i$ and their standard errors for MLFDR.

end for

Direct application of SVA to model (7) would require estimating surrogate variables for each mediator-outcome pair iteratively, leading to a computational bottleneck. To address this, we propose a global factor adjustment. We estimate a second set of surrogate variables, $\hat{\mathbf{U}}_Y$, based on the outcome null model (excluding mediators):

$$Y_i = \nu_i + \gamma_i X + \delta_i Z + \sum_{l=1}^{L} \eta_l^{(i)} g_l + \epsilon_i. \tag{8}$$

We then use the combined set of latent factors $\hat{\mathbf{U}}_M$ (derived from mediators) and $\hat{\mathbf{U}}_Y$ (derived from outcome residuals) to model the mediator-outcome relationship. The validity of using $\hat{\mathbf{U}}_Y$ from (8) relies on the assumption that any single mediator M_i contributes a relatively small amount of variance to the global outcome matrix $\mathbf{Y}$.

The full procedure is summarized in Algorithm 1. We implemented surrogate variable analysis via the Bioconductor package `sva` [9].

We apply this method to two data generating scenarios to demonstrate its performance.

Unknown mediator-exposure interactions.

$$\begin{aligned} M_i &= \alpha_i X + \delta_i Z + e_i, \\ Y_i &= M_i\beta_i + X\gamma_i + \zeta_i Z + X\sum_{j\in S}\theta_j M_j + \epsilon_i, \end{aligned} \tag{9}$$

where S is a randomly selected subset of indices $\{1, \dots, m\}$ with $|S| = 20$, treated as unknown during model fitting. The results are shown in Fig 5.

Unmeasured confounding and pleiotropy.

$$\begin{aligned} M_i &= X\alpha_i + \theta_i Z + 0.4Z_1 + 0.5Z_2 + e_i, \\ Y_i &= M_i\beta_i + X\gamma_i + \delta_i Z - 0.5Z_1 + \sum_{j\in S} M_j\kappa_j + \epsilon_i, \end{aligned} \tag{10}$$

where Z_1 and Z_2 represent unmeasured confounders not included in the model fitting. S is a randomly selected subset of indices $\{1, \dots, m\}$ with $|S| = 20$. The term $\sum_{j\in S} M_j\kappa_j$ represents dense pleiotropy (the effect of other mediators on Y_i), which acts as an additional source of unmeasured variation. The results are shown in Fig 6.

2.3.2 Composite alternatives. In this setting, the coefficients (α, β) follow a Gaussian mixture distribution. The posterior distribution of the estimated coefficients is given by:

$$\begin{aligned} \sqrt{n}\hat{\alpha}_i &\sim p_0 N(0, \sigma_{i1}^2) + p_1 N(\mu_1, \sigma_{i1}^2 + \kappa_1) + p_2 N(\mu_2, \sigma_{i1}^2 + \kappa_2), \\ \sqrt{n}\hat{\beta}_i &\sim q_0 N(0, \sigma_{i2}^2) + q_1 N(\theta_1, \sigma_{i2}^2 + \psi_1) + q_2 N(\theta_2, \sigma_{i2}^2 + \psi_2). \end{aligned}$$

Under this framework, the null hypothesis corresponds to a mixture of 5 bivariate Gaussian distributions, while the alternative hypothesis comprises a mixture of 4 bivariate Gaussian distributions. The parameters are estimated using a two-step EM algorithm, the details of which are provided in Section C of S1 Text.

In our simulations, we set the mixture weights to $(p_0, p_1, p_2) = (0.54, 0.18, 0.28)$ and $(q_0, q_1, q_2) = (0.6, 0.05, 0.35)$. The variance parameters were fixed at $\kappa_1 = 1, \kappa_2 = 2, \psi_1 = 1.5, \psi_2 = 2$. The mean parameters were defined as functions of the mediation signal strength τ: $\mu_1 = 0.05\tau$, $\mu_2 = -0.5\tau$, $\theta_1 = 0.9\tau$, and $\theta_2 = -0.01\tau$, where τ ranges from 0.1 to 1.9 in increments of 0.2.

Fig 7 presents the results for $n \in \{100, 300\}$ with $m = 1000$. All three methods maintained satisfactory FDR control, with MLFDR demonstrating the highest power.

All methods appear to have satisfactory FDR control. MLFDR is uniformly more powerful than HDMT and MDACT in all cases, with better margin of improvement in dense alternatives.

3 Real data analysis

3.1 TCGA prostate cancer data

We apply MLFDR to the Prostate Cancer dataset from The Cancer Genome Atlas (TCGA), previously analyzed by [3]. The study involves mediation analysis for 147 prostate cancer risk SNPs, integrated with DNA methylation and gene expression data from 495 samples. For each risk SNP, we identified CpG methylation probes within a 500 kb window and recorded the gene expression levels for the corresponding probes. This resulted in $m = 69,602$ SNP-CpG-Gene triplets for mediation testing.

In the first stage, we regressed CpG methylation on the SNPs, adjusting for the top 3 principal components (PCs) of genotypes, the top 15 PCs of CpG methylation, age at diagnosis, and pathological stage. From this, we obtained the

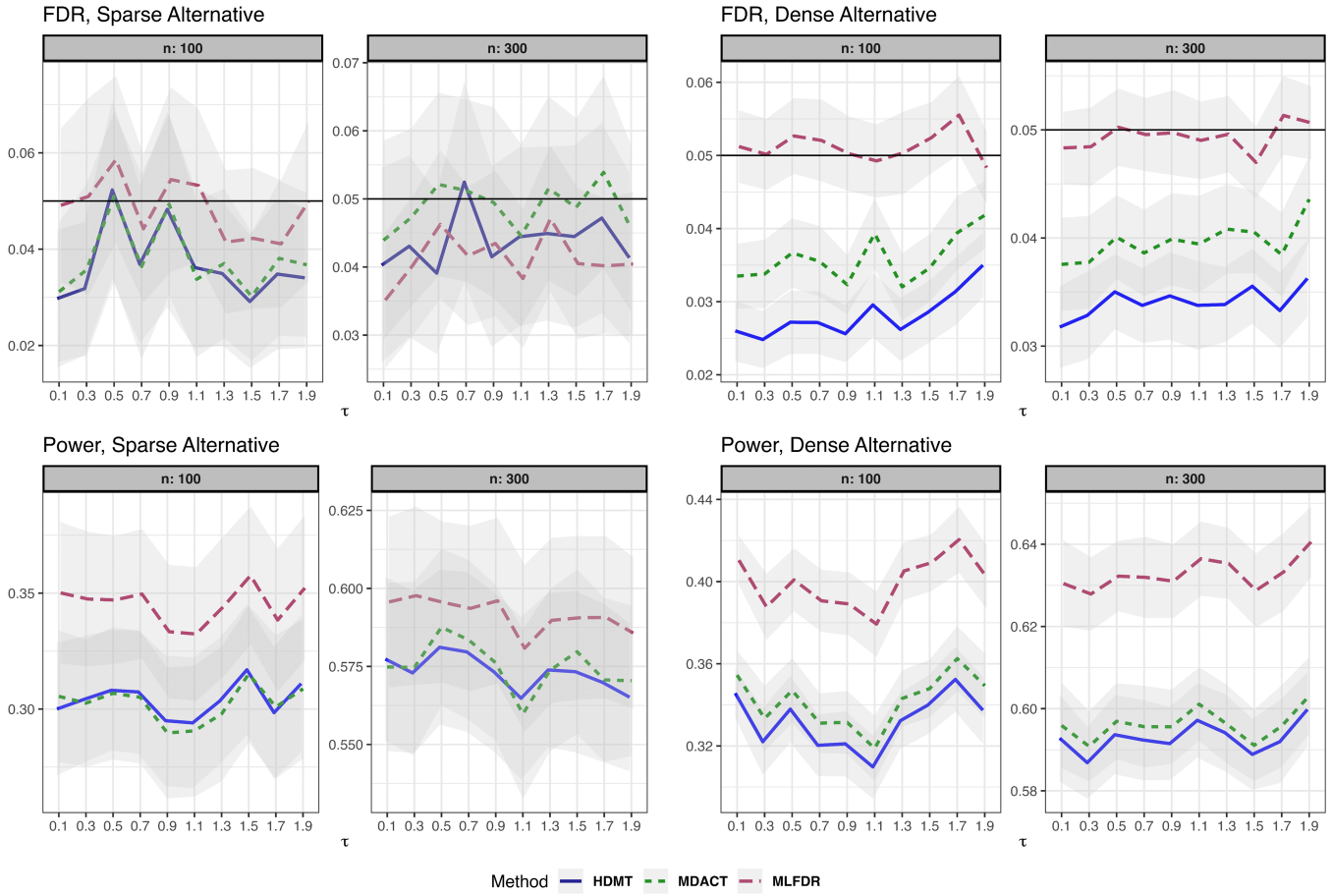

Fig 5. FDR and power comparison for the linear model with unknown mediator-exposure interactions. Results are displayed for both sparse and dense alternatives. Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g005

slope estimates, variances, and p-values for the SNPs. In the second stage, gene expression was regressed on CpG methylation, conditional on the same set of covariates.

The estimated null proportion components $(\pi_{00}, \pi_{10}, \pi_{01})$ were (0.51,0.033,0.41) for HDMT, compared to (0.39,0.004,0.59) obtained via the EM algorithm in MLFDR.

Due to the wide spread of the methylation coefficients (β), we fitted a composite alternative Gaussian mixture model. The number of components, $d_2 = 8$, was selected based on the Akaike Information Criterion (AIC) (Fig 8). Conversely, the SNP coefficients (α) exhibited a narrower range (–0.2 to 0.4) and were adequately modeled using a $d_1 = 2$ component Gaussian mixture.

At an FDR threshold of 0.01, HDMT identified 137 triplets, MLFDR identified 187, and MDACT identified 180. Fig 9 displays a Venn diagram of the overlapping discoveries, along with the number of rejections across FDR cutoffs ranging from 0.001 to 0.05. MLFDR consistently detected more pathways than the competing methods.

Table 1 lists 10 additional pathways identified by MLFDR (but missed by HDMT or MDACT), ranked by local FDR. Notably, six of these triplets involve rs12653946, a known prostate cancer risk variant that influences the expression of *IRX4*, a tumor suppressor gene in the prostate [6]. Another pathway involves rs7767188, a risk SNP associated with prostate cancer through the expression of *TRIM26* [29].

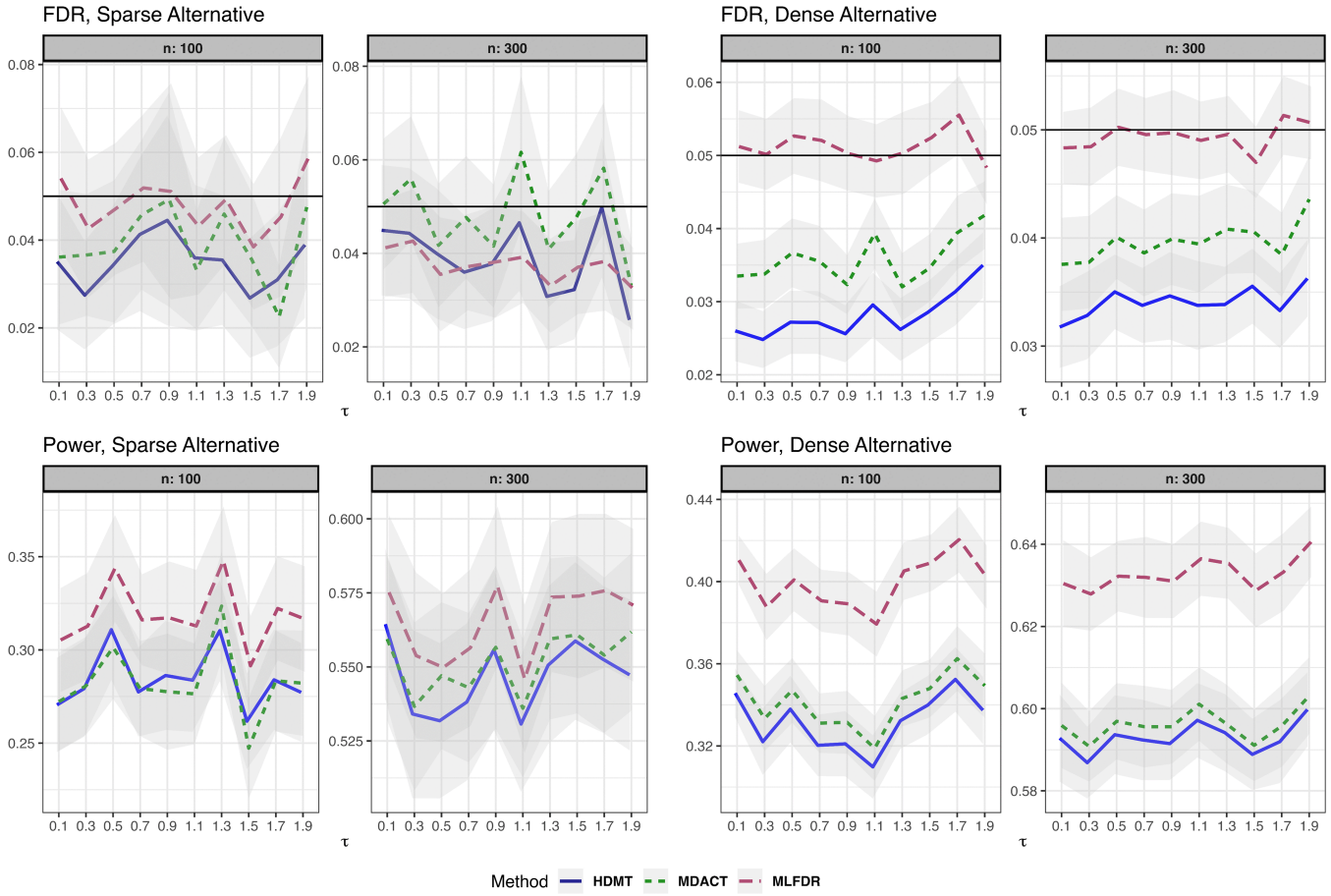

Fig 6. FDR and power comparison for the linear model with unmeasured confounders. Results are displayed for both sparse and dense alternatives. Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g006

These empirical results align with our simulation studies: HDMT yielded the fewest discoveries, followed by MDACT, with MLFDR providing the highest detection rate. We note that the improvement of MLFDR over MDACT is modest in this application. This is likely attributable to the symmetric distribution of α and β; as noted by [22], local FDR-based tests offer limited power gains over p-value-based methods when the alternative distribution is symmetric around the null.

3.2 TCGA lung squamous cell carcinoma

We further extend our analysis to the TCGA Lung Squamous Cell Carcinoma (LUSC) dataset to investigate the mediating role of CpG methylation in the relationship between smoking history and gene expression. The data were acquired using the R package `UCSCXenaTools` [28]. We restricted the analysis to primary tumor samples, resulting in a sample size of $n = 379$ after preprocessing.

Smoking history was quantified by pack-years. Using the publicly available probe map data from the TCGA website, each CpG probe was mapped to the expression profiles of potentially multiple genes; each unique CpG-gene pair was treated as a distinct candidate mediation pathway. The analysis proceeded in two stages. In the first stage, CpG methylation beta values were regressed on smoking history (pack-years). In the second stage, a multiple linear regression was fitted for each gene, including all CpG probes mapped to that gene as predictors. The coefficients and p-values for each

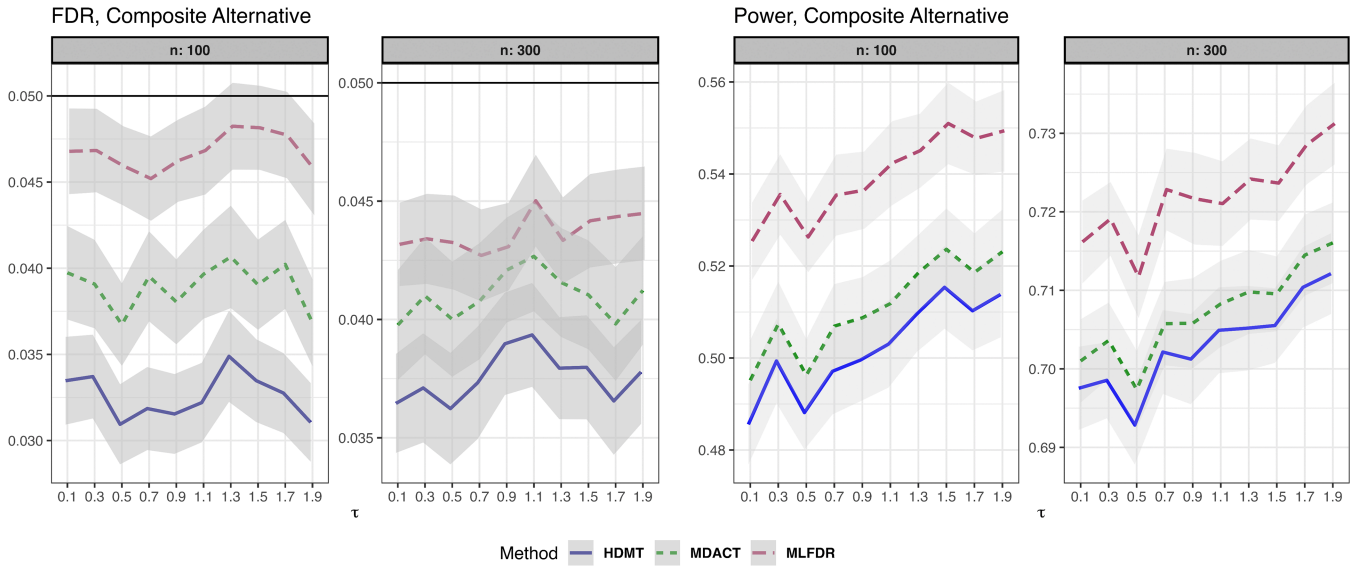

Fig 7. FDR and power comparison for composite alternatives. Results are displayed for varying degrees of mediation (τ). Gray ribbons indicate error margins.

https://doi.org/10.1371/journal.pcbi.1013880.g007

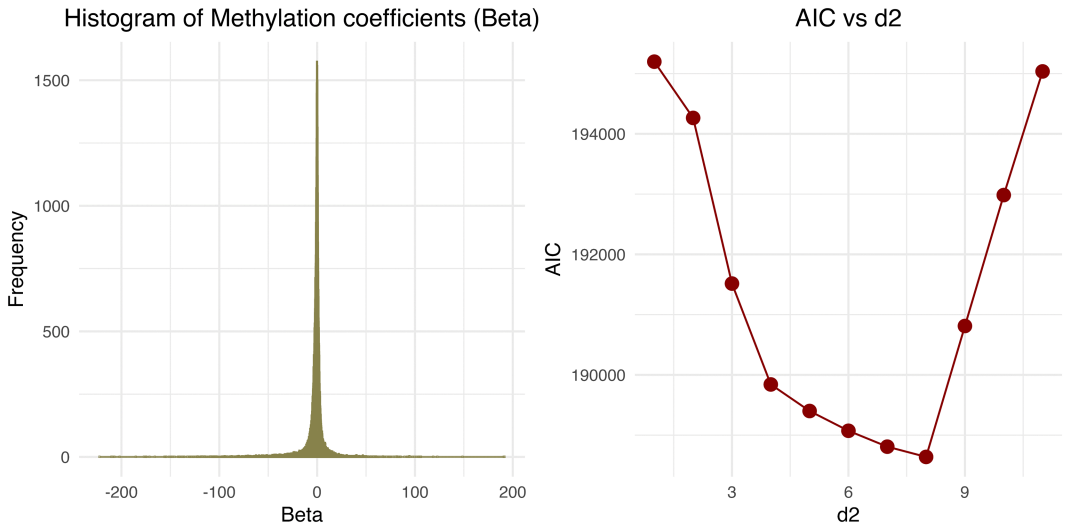

Fig 8. AIC for methylation coefficients when a $d_2 + 1$ component Gaussian mixture model was fit.

https://doi.org/10.1371/journal.pcbi.1013880.g008

specific CpG-gene pair were then extracted to form the mediation hypotheses. All models were adjusted for potential confounders, including sex and the age at initial diagnosis. In total, 319,761 CpG-gene pathways were evaluated.

At an FDR threshold of 0.01, HDMT identified 13 pathways, MDACT identified 25 pathways, and MLFDR identified 44 pathways (Fig 10). The results highlight the increased power of MLFDR in detecting subtle mediation signals. Table 2 details the top 10 additional pathways detected by MLFDR (ranked by local FDR) that were not identified by the competing methods. Several of these findings are supported by existing literature. For instance, [27] discuss the relevance of *WDR66* in lung cancer progression, while [2] highlights the role of *LY6K* as a potential therapeutic target in lung squamous cell carcinoma. [31] links *TCIRG1* to metastatic potential of hepatocellular carcinoma.

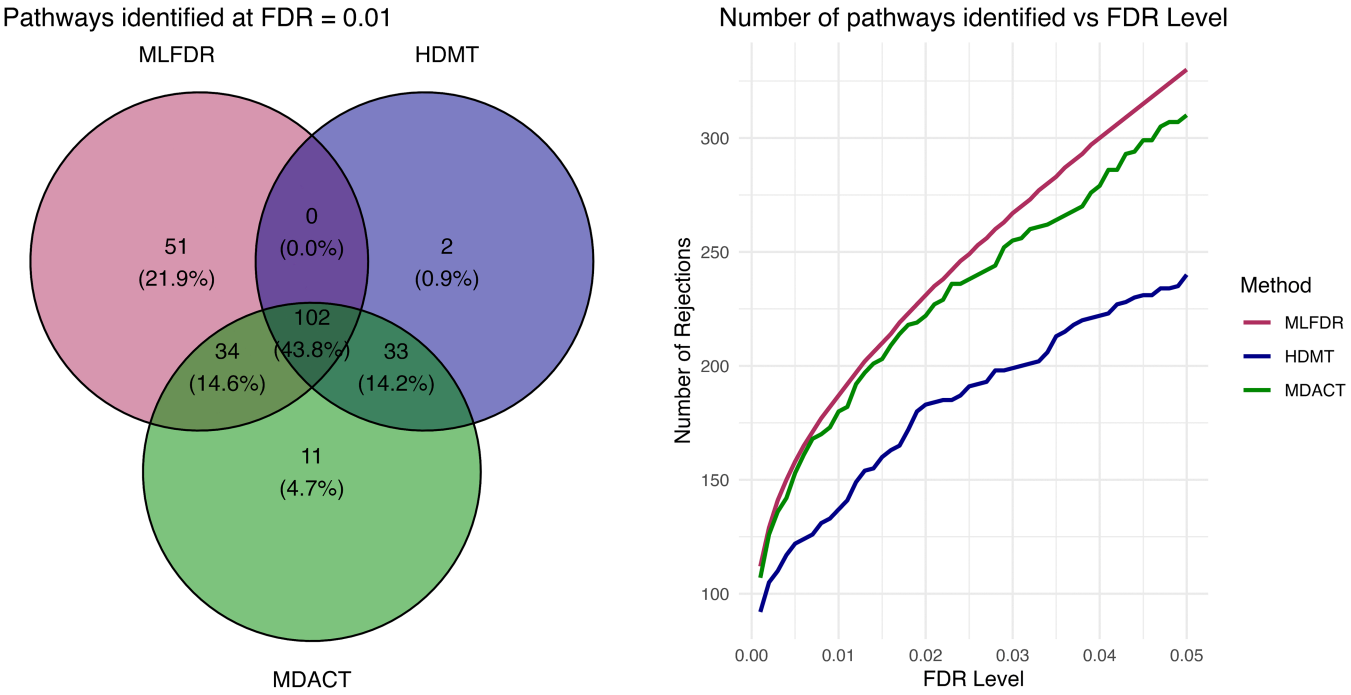

Fig 9. SNP-CpG-Gene triplets identified by MDACT, MLFDR, and HDMT out of 69,602 tests. The Venn diagram (left) corresponds to an FDR cutoff of 0.01.

https://doi.org/10.1371/journal.pcbi.1013880.g009

Table 1. Top 10 additional pathways detected by MLFDR in TCGA Prostate Cancer Data (ranked by p_{max}). These pathways were not detected by HDMT or MDACT.

SNP	CpG Probe	Annotated Gene	Target Gene	p_{max}	Local FDR
rs7767188	cg02749752	TRIM26	TRIM26	0.0017	0.0034
rs12653946	cg12830271	–	IRX4	0.0019	0.0058
rs3096702	cg19609334	TNXB	TNXB	0.0019	0.0060
rs3129859	cg03520342	HLA-DMA	HLA-DMA	0.0019	0.0125
rs12653946	cg00085370	–	IRX4	0.0022	0.0067
rs12653946	cg07144328	–	IRX4	0.0022	0.0063
rs12653946	cg07278634	–	IRX4	0.0028	0.0077
rs12653946	cg06446548	–	NDUFS6	0.0028	0.0288
rs12653946	cg03225093	–	IRX4	0.0029	0.0083
rs5945619	cg10581449	NUDT11	NUDT11	0.0030	0.0077

https://doi.org/10.1371/journal.pcbi.1013880.t001

To evaluate the FDR of the three methods, we selected a subset of the data with $m = 1,000$ tests, and permuted the samples to create a "global null" scenario. For each permutation, the smoking-Cpg and CpG-gene expression model is fit, and the three methods are implemented. At a given permutation, if any method has non-zero rejections, the False Positive rate for that permutation is recorded as 1 for that method. This procedure is repeated over 100 permutations, and the number of rejections is recorded for FDR levels varying from 0.0001 to 0.2. Fig 11 presents the proportion of cases in which each method came up with significant rejections under the global null. HDMT and MLFDR appear to have satisfactory control of the FDR, while MDACT appears to have some FDR inflation at low FDR levels.

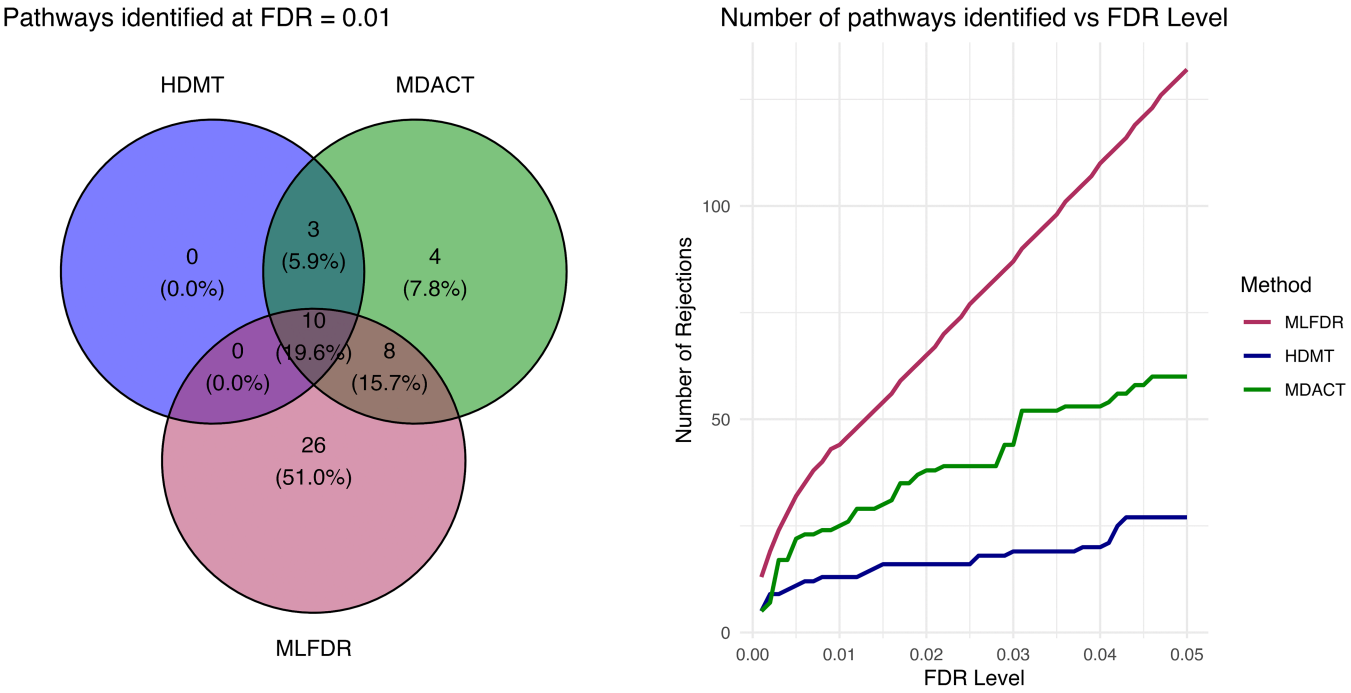

Fig 10. Smoking-CpG-Gene pathways detected by MDACT, MLFDR, and HDMT out of 319,761 tests. The Venn diagram (left) displays overlaps at an FDR cutoff of 0.01, while the plot (right) shows detection counts across varying FDR thresholds.

https://doi.org/10.1371/journal.pcbi.1013880.g010

Table 2. Top 10 additional pathways detected by MLFDR in the TCGA LUSC dataset (ranked by p_{max}). These pathways were not detected by HDMT or MDACT.

Rank	Gene	CpG Probe	p_{max}	Local FDR
1	HOXB6	cg20591728	6.6×10^{-6}	0.0039
2	MYLIP	cg04641165	1.7×10^{-5}	0.0040
3	B3GALT2	cg16712103	1.8×10^{-5}	0.0100
4	ZNF287	cg16964464	2.2×10^{-5}	0.0085
5	TCIRG1	cg20484322	3.6×10^{-5}	0.0131
6	WDR66	cg03560652	3.8×10^{-5}	0.0145
7	ENO3	cg07333510	4.5×10^{-5}	0.0139
8	ADORA2B	cg21501163	5.8×10^{-5}	0.0197
9	GNA14	cg06617692	6.4×10^{-5}	0.0231
10	LY6K	cg16809304	8.1×10^{-5}	0.0148

https://doi.org/10.1371/journal.pcbi.1013880.t002

4 Discussion

We have presented a flexible and powerful screening algorithm for detecting causal pathways in high-dimensional mediation analysis. MLFDR is capable of dealing with a wide array of outcome variables, confounders, and mediation exposure interactions. It can work in tandem with surrogate variable analysis to address complex dependence structures like pleiotropy and unmeasured confounding. Through simulations and theoretical analysis, we have shown that the proposed method is a viable alternative to p-value-based screening methods, which may not produce optimal results depending on the structure of the alternative hypothesis. We have also shown theoretical guarantees of FDR control for MLFDR. Overall, we hope that MLFDR will be used as an alternative to p-value based methods for mediator screening in the future.

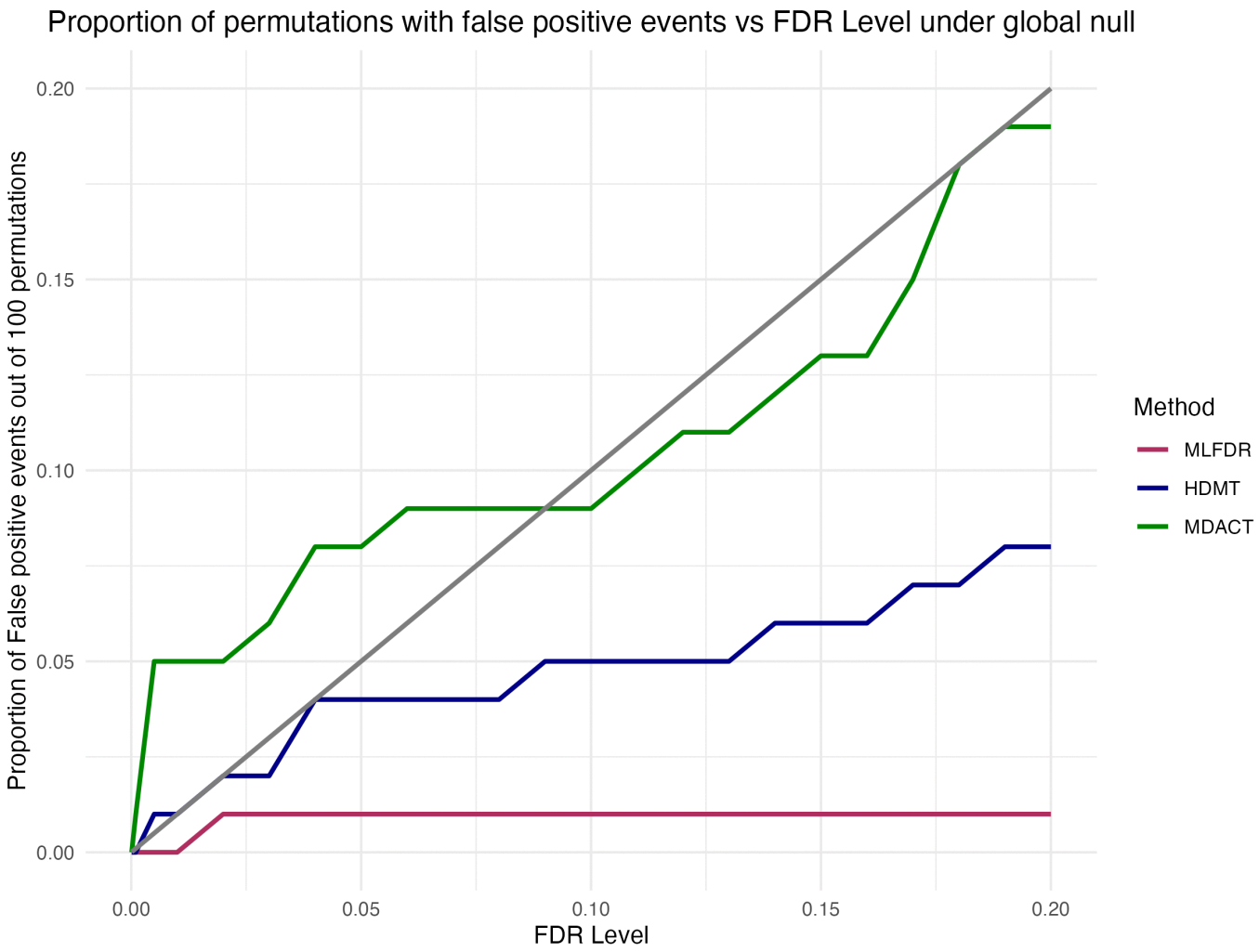

Fig 11. Smoking-CpG-gene pathways detected by all methods under the global null out of $m = 1,000$ tests. The proportion of cases reporting positive detection counts is reported across varying FDR thresholds. The gray solid line indicates target FDR levels.

https://doi.org/10.1371/journal.pcbi.1013880.g011

5 Methods

5.1 The mediation model

Consider a univariate exposure X, a set of mediators $\{M_i\}_{i=1}^m$, and m outcomes $\{Y_i\}_{i=1}^m$. We assume the following structural equation model:

$$\begin{aligned} M_i &= X\alpha_i + e_i, \\ Y_i &= M_i\beta_i + X\gamma_i + \epsilon_i, \end{aligned} \tag{11}$$

where the error terms satisfy $(e_i, \epsilon_i) \perp (X, M_i)$ and are distributed as:

$$(e_i, \epsilon_i) \sim N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{i,a}^2 & 0\\ 0 & \sigma_{i,b}^2\end{pmatrix}\right).$$

In settings with high-dimensional exposures, such as Single Nucleotide Polymorphisms (SNPs), we define the hypothesis based on the Exposure-Mediator-Outcome triplet $\{X_i, M_i, Y_i\}_{i=1}^m$. In such cases, the common exposure X in Eq (11) is replaced by the specific exposure X_i.

To describe the methodology, we focus on the formulation in Model (11). Given n independent samples $\{(Y^j, X^j, M^j)\}_{j=1}^n$, our goal is to test the joint significance of α_i and β_i for $1 \le i \le m$. Let $\mathbf{Y}_i = (Y_i^1, \dots, Y_i^n)'$, $\mathbf{X} = (X^1, \dots, X^n)'$, $\mathbf{M}_i = (M_i^1, \dots, M_i^n)'$, and define the projection matrix $\mathbf{P} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. The ordinary least squares (OLS) estimators are given by:

$$\hat{\alpha}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{M}_i, \quad \hat{\beta}_i = (\mathbf{M}_i'\mathbf{P}\mathbf{M}_i)^{-1}\mathbf{M}_i'\mathbf{P}\mathbf{Y}_i. \tag{12}$$

Conditional on $\mathbf{X}$ and $\mathbf{M}_i$, the estimators follow the distribution:

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}_i - \alpha_i \\ \hat{\beta}_i - \beta_i\end{pmatrix} = \sqrt{n}\begin{pmatrix}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_i \\ (\mathbf{M}_i'\mathbf{P}\mathbf{M}_i)^{-1}\mathbf{M}_i'\mathbf{P}\boldsymbol{\epsilon}_i\end{pmatrix} \overset{d}{=} N\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_{i1}^2 & 0 \\ 0 & \sigma_{i2}^2\end{pmatrix}\right),$$

where $\sigma_{i1}^2 = n\sigma_{i,a}^2(\mathbf{X}'\mathbf{X})^{-1}$ and $\sigma_{i2}^2 = n\sigma_{i,b}^2(\mathbf{M}_i'\mathbf{P}\mathbf{M}_i)^{-1}$. Note that $\hat{\alpha}_i$ and $\hat{\beta}_i$ are independent, as $\mathbb{E}[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}_i\boldsymbol{\epsilon}_i'\mathbf{P}\mathbf{M}_i(\mathbf{M}_i'\mathbf{P}\mathbf{M}_i)^{-1}] = 0$. We test the composite null hypothesis against the alternative:

$$H_{0,i} : \alpha_i\beta_i = 0 \quad \text{versus} \quad H_{11,i} : \alpha_i\beta_i \neq 0, \quad i = 1, 2, \dots, m. \tag{13}$$

Denote by $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})$ the latent vector indicating the underlying truth of the i-th hypothesis, where $\xi_{i1} = \mathbf{1}\{\alpha_i \neq 0\}$ and $\xi_{i2} = \mathbf{1}\{\beta_i \neq 0\}$. The vector $\boldsymbol{\xi}_i$ takes values in the set $\{(0,0), (0,1), (1,0), (1,1)\}$. Let $m_{jk} = \sum_{i=1}^m \mathbf{1}\{\boldsymbol{\xi}_i = (j,k)\}$ represent the count of hypotheses in each state for $0 \le j, k \le 1$.

We assume a prior distribution $P(\boldsymbol{\xi}_i = (j,k)) = \pi_{jk}$, where $\pi_{jk} \ge 0$ and $\sum_{j,k}\pi_{jk} = 1$. Conditional on the latent states, the non-zero effect sizes are modeled as Gaussian:

$$\sqrt{n}\alpha_i \mid (\xi_{i1} = 1) \sim N(\mu, \psi) \quad \text{and} \quad \sqrt{n}\beta_i \mid (\xi_{i2} = 1) \sim N(\theta, \kappa).$$

When $\xi_{i1} = 0$, $\alpha_i = 0$ (and analogously for β_i). Assuming that α_i and β_i are independent conditional on $\boldsymbol{\xi}_i$, the marginal distribution of the coefficient estimates $(\hat{\alpha}_i, \hat{\beta}_i)$ given the latent states is:

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}_i \\ \hat{\beta}_i\end{pmatrix} \Bigg| \boldsymbol{\xi}_i \sim N\left(\begin{pmatrix}\mu\xi_{i1} \\ \theta\xi_{i2}\end{pmatrix}, \begin{pmatrix}\sigma_{i1}^2 + \psi\xi_{i1} & 0 \\ 0 & \sigma_{i2}^2 + \kappa\xi_{i2}\end{pmatrix}\right), \tag{14}$$

where $(\sigma_{i1}^2/n, \sigma_{i2}^2/n)$ denote the variances of the estimates $(\hat{\alpha}_i, \hat{\beta}_i)$ conditional on (α_i, β_i).

We employ an Expectation-Maximization (EM) algorithm to estimate the unknown parameters $\boldsymbol{\Theta} = \{\boldsymbol{\pi}, \mu, \theta, \psi, \kappa\}$; details are provided in Section B of the S1 Text. Although MLFDR is derived under the framework of Model (11), it is applicable to a broader range of settings, including binary outcomes, mediator-exposure interactions, and models with confounders. The extensive simulations presented earlier demonstrate the method's robustness across these varied scenarios.

5.2 Extension: Composite alternative

In the last section we introduced the latent variable $\boldsymbol{\xi}_i = (\xi_{i1}, \xi_{i2})$ to characterize the underlying state of the hypothesis. We now extend this framework to a composite alternative setting, where the non-zero effects are drawn from mixture distributions. Specifically, we assume (α_i, β_i) are generated as follows:

$$\begin{aligned}\sqrt{n}\alpha_i \mid \{\xi_{i1} = u\} &\sim N(\mu_u, \kappa_u), \quad u = 0, 1, \ldots, d_1,\\ \sqrt{n}\beta_i \mid \{\xi_{i2} = v\} &\sim N(\theta_v, \psi_v), \quad v = 0, 1, \ldots, d_2.\end{aligned}$$

Here, the index 0 denotes the null state, such that $\mu_0 = \theta_0 = \kappa_0 = \psi_0 = 0$ (i.e., a degenerate distribution at zero). As before, α_i and β_i are assumed independent conditional on the latent state $\boldsymbol{\xi}_i$.

Marginalizing over the prior distribution, the joint distribution of the estimators $(\hat{\alpha}_i, \hat{\beta}_i)$ follows a Gaussian Mixture Model (GMM):

$$\sqrt{n}\begin{pmatrix}\hat{\alpha}_i\\ \hat{\beta}_i\end{pmatrix} \sim \sum_{u=0}^{d_1}\sum_{v=0}^{d_2} \pi_{uv} N\left(\begin{pmatrix}\mu_u\\ \theta_v\end{pmatrix}, \begin{pmatrix}\sigma_{i1}^2 + \kappa_u & 0\\ 0 & \sigma_{i2}^2 + \psi_v\end{pmatrix}\right), \tag{15}$$

where $\pi_{uv} = P(\xi_{i1} = u, \xi_{i2} = v)$. In this context, the component null hypotheses map to specific index combinations: $H_{01,i}$ corresponds to $\{\xi_{i1} = 0, \xi_{i2} \in \{1, \ldots, d_2\}\}$, while the alternative $H_{11,i}$ corresponds to $\{\xi_{i1} \in \{1, \ldots, d_1\}, \xi_{i2} \in \{1, \ldots, d_2\}\}$.

The corresponding marginal distributions are given by:

$$\begin{aligned}\sqrt{n}\hat{\alpha}_i &\sim \sum_{u=0}^{d_1} \pi_{u\cdot} N(\mu_u, \sigma_{i1}^2 + \kappa_u),\\ \sqrt{n}\hat{\beta}_i &\sim \sum_{v=0}^{d_2} \pi_{\cdot v} N(\theta_v, \sigma_{i2}^2 + \psi_v),\end{aligned} \tag{16}$$

where $\pi_{u\cdot} = \sum_v \pi_{uv}$ and $\pi_{\cdot v} = \sum_u \pi_{uv}$.

Two-step EM algorithm. Directly fitting a $(d_1 + 1)(d_2 + 1)$-component bivariate GMM to estimate the joint probability matrix $\boldsymbol{\pi}$ is computationally intensive. To mitigate this burden, we propose a two-step Expectation-Maximization (EM) algorithm.

In the first step, we estimate the parameters $\{\mu_u, \kappa_u\}_u$ and $\{\theta_v, \psi_v\}_v$ using the univariate marginal distributions described in (Eq 16). While the marginal mixing proportions $\pi_{u\cdot}$ and $\pi_{\cdot v}$ are insufficient to recover the joint distribution π_{uv} (without assuming independence), the moment estimates remain valid.

In the second step, we fix these mean and variance estimates and fit a constrained bivariate GMM to the joint data solely to estimate the joint mixing proportions π_{uv}. This approach significantly reduces the dimensionality of the optimization problem. Details of this algorithm are provided in Section C of S1 Text, and a supporting simulation is presented in Sect 2.3.

This two-step approach is also applicable to the standard case where $d_1 = d_2 = 1$. In Section D of S1 Text, we compare the standard bivariate EM against this two-step variant. The results indicate that while the two-step method offers substantial computational speedups, it incurs a slight reduction in power. Therefore, we recommend using the standard bivariate EM for simple cases ($d_1 = d_2 = 1$) and reserving the two-step EM for complex composite alternatives where computational efficiency is paramount.

5.3 Step-up procedure based on local FDR

[22] demonstrated that for a simple null hypothesis, the oracle local FDR-based rejection region outperforms p-value-based thresholding. Specifically, it achieves a lower marginal False Non-discovery Rate (mFNR) while controlling the marginal False Discovery Rate (mFDR) at the same level. This advantage is particularly pronounced when the alternative distribution is asymmetric about the null. Motivated by these findings, we implement a local FDR-based step-up procedure to identify significant mediation pathways.

We focus our discussion on the Gaussian Mixture Model described in (14). The extension to the more general mixture model in (15) is straightforward. Conditional on the latent variables, the density function of the transformed statistics $\sqrt{n}(\hat{\alpha}_i, \hat{\beta}_i)$ under state $H_{jk,i}$ is given by:

$$\begin{aligned} f_{jk}(\cdot,\cdot) &:= \phi(\cdot,\cdot; \mu\delta_j, \theta\delta_k, \sigma_{i1}^2 + \kappa\delta_j, \sigma_{i2}^2 + \psi\delta_k), \\ f &:= \sum_{j=0}^{1}\sum_{k=0}^{1} \pi_{jk} f_{jk}, \end{aligned} \tag{17}$$

where $\delta_0 = 0$, $\delta_1 = 1$, and $\phi(\cdot,\cdot; \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ denotes the bivariate normal density with mean vector (μ_1, μ_2), variances (σ_1^2, σ_2^2), and zero correlation.

For notational simplicity, let $a_i = \sqrt{n}\hat{\alpha}_i$ and $b_i = \sqrt{n}\hat{\beta}_i$. Under the composite null hypothesis, the joint local FDR is defined as:

$$\text{lfdr}(a_i, b_i) = \frac{\pi_{00} f_{00}(a_i, b_i) + \pi_{10} f_{10}(a_i, b_i) + \pi_{01} f_{01}(a_i, b_i)}{f(a_i, b_i)}. \tag{18}$$

This quantity can be estimated by substituting the parameter estimates obtained from the EM algorithm; we denote this estimate by $\widehat{\text{lfdr}}$. As the local FDR represents the posterior probability of the null hypothesis, a lower value indicates stronger evidence against the null. Consequently, we define the rejection region as $\widehat{\text{lfdr}}(a_i, b_i) \le \delta$, where the threshold δ must be determined to control the error rate.

We assume the cumulative distribution function (CDF) of the joint local FDR is given by:

$$G(t) = \pi_{00} G_{00}(t) + \pi_{10} G_{10}(t) + \pi_{01} G_{01}(t) + \pi_{11} G_{11}(t), \tag{19}$$

where $G_{jk}(t)$ is the conditional CDF of $\text{lfdr}(a_i, b_i)$ under hypothesis H_{jk}:

$$G(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(\text{lfdr}(a_i, b_i) \le t), \tag{20}$$

$$G_{jk}(t) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{P}(\text{lfdr}(a_i, b_i) \le t \mid H_{jk}), \quad j,k \in \{0,1\}. \tag{21}$$

Oracle procedure. The oracle procedure assumes that all parameters $\boldsymbol{\pi} = \{\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}\}$ and densities f_{jk} are known. For a given threshold δ, we define the following counting processes:

$$V_m(\delta) = \sum_{i \in H_{00} \cup H_{10} \cup H_{01}} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\}, \tag{22}$$

$$R_m(\delta) = \sum_{i=1}^{m} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\}, \tag{23}$$

$$P_m(\delta) = \sum_{i \in H_{11}} \mathbf{1}\{\text{lfdr}(a_i, b_i) > \delta\}, \tag{24}$$

$$W_m(\delta) = \sum_{i=1}^{m} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\} \cdot \text{lfdr}(a_i, b_i), \tag{25}$$

where $V_m(\delta)$ represents the number of false rejections, $R_m(\delta)$ the total number of rejections, and $P_m(\delta)$ the number of missed discoveries (false negatives). We can express $V_m(\delta)$ as:

$$V_m(\delta) = \sum_{i=1}^{m} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\} \cdot \mathbf{1}\{\boldsymbol{\xi}_\mathbf{i} \in \{(0,0),(1,0),(0,1)\}\}.$$

Taking the expectation, we obtain:

$$\begin{aligned} E[V_m(\delta)] &= \sum_{i=1}^{m} E\left[\mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\} \cdot \mathbf{1}\{\boldsymbol{\xi}_\mathbf{i} \in \{(0,0),(1,0),(0,1)\}\}\right] \\ &= m_{00} G_{00}(\delta) + m_{10} G_{10}(\delta) + m_{01} G_{01}(\delta). \end{aligned} \tag{26}$$

The mFDR at threshold δ is defined as:

$$\widetilde{Q}(\delta) = \frac{E[V_m(\delta)]}{E[R_m(\delta)]} = \frac{\pi_{00} G_{00}(\delta) + \pi_{10} G_{10}(\delta) + \pi_{01} G_{01}(\delta)}{G(\delta)}. \tag{27}$$

This quantity can be empirically estimated by:

$$Q_m(\delta) = \frac{\sum_{i=1}^{m} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\} \text{lfdr}(a_i, b_i)}{\sum_{i=1}^{m} \mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta\}}. \tag{28}$$

In Section A of the S1 Text, we prove that for a fixed δ, the numerator and denominator of $Q_m(\delta)$ are unbiased estimators of the numerator and denominator of $\widetilde{Q}(\delta)$, respectively. The oracle rejection region for the composite null is defined as $\mathbf{1}\{\text{lfdr}(a_i, b_i) \le \delta_m\}$, where the threshold is selected as:

$$\delta_m = \sup\{t \in (0,1) : Q_m(t) \le \alpha\}. \tag{29}$$

Adaptive procedure. In practical applications, the true parameters $\boldsymbol{\pi} = \{\pi_{00}, \pi_{10}, \pi_{01}, \pi_{11}\}$ and the component densities f_{jk} are unknown. Consequently, the true local FDR values in Equation (28) are inaccessible. To address this, we substitute the unknown quantities with their estimates obtained via the EM algorithm. By plugging in these estimates, we obtain the estimated local FDR, denoted as $\widehat{\text{lfdr}}$, which yields an empirical estimate of $Q_m(\delta)$:

$$\widehat{Q}_m(\delta) = \frac{\sum_{i=1}^{m} \mathbf{1}\{\widehat{\text{lfdr}}(a_i, b_i) \le \delta\} \widehat{\text{lfdr}}(a_i, b_i)}{\sum_{i=1}^{m} \mathbf{1}\{\widehat{\text{lfdr}}(a_i, b_i) \le \delta\}}. \tag{30}$$

Accordingly, the data-adaptive threshold $\hat{\delta}_m$ is defined as:

$$\hat{\delta}_m = \sup\{t \in (0,1) : \widehat{Q}_m(t) \le \alpha\}.$$

Algorithm 2 A data-adaptive procedure for finding the cutoffs in MLFDR.

Require: EM estimates for the parameters $\Theta = \{\pi, \mu, \theta, \psi, \kappa\}$

for $i = 1, \ldots, m$ **do**

  Compute $\widehat{\mathrm{lfdr}}_i = \widehat{\mathrm{lfdr}}(\sqrt{n}\hat{\alpha}_i, \sqrt{n}\hat{\beta}_i)$

end for

Sort the estimated local FDR values in ascending order to obtain the set of order statistics $\Psi = \{\widehat{\mathrm{lfdr}}_{(1)}, \widehat{\mathrm{lfdr}}_{(2)}, \ldots, \widehat{\mathrm{lfdr}}_{(m)}\}$, where ties are broken at random.

For each $k \in \{1, \ldots, m\}$, compute the average estimated local FDR for the top k statistics:

$$\hat{Q}_m(\widehat{\mathrm{lfdr}}_{(k)}) = \frac{1}{k}\sum_{i=1}^{k} \widehat{\mathrm{lfdr}}_{(i)}.$$

Determine the cutoff index $k = \max\left\{j : \frac{1}{j}\sum_{i=1}^{j} \widehat{\mathrm{lfdr}}_{(i)} \le \alpha\right\}$

Reject the null hypotheses corresponding to the smallest k local FDR values, denoted as $H_{(1)}, H_{(2)}, \ldots, H_{(k)}$

The resulting procedure, which rejects the i-th hypothesis if $\widehat{\mathrm{lfdr}}(a_i, b_i) \le \hat{\delta}_m$, is operationally equivalent to the step-up algorithm detailed in Algorithm 2.

Supporting information

S1 Fig. Comparison of empirical FDR and power across different methods for MLFDR and two-step MLFDR.
(TIFF)

S1 Text. Supplementary file detailing the methods and the theory.
(PDF)

Author contributions

Conceptualization: Xianyang Zhang.

Formal analysis: Asmita Roy.

Funding acquisition: Xianyang Zhang.

Methodology: Asmita Roy, Xianyang Zhang.

Software: Asmita Roy.

Supervision: Xianyang Zhang.

Visualization: Asmita Roy.

Writing – original draft: Asmita Roy.

Writing – review & editing: Xianyang Zhang.

References

1. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, strategic, and statistical considerations. J Pers Soc Psychol. 1986;51(6):1173–82. https://doi.org/10.1037//0022-3514.51.6.1173 PMID: 3806354
2. Chen Y, Zhou C, Zhang X, Chen M, Wang M, Zhang L, et al. Construction of a novel radioresistance-related signature for prediction of prognosis, immune microenvironment and anti-tumour drug sensitivity in non-small cell lung cancer. Ann Med. 2025;57(1):2447930. https://doi.org/10.1080/07853890.2024.2447930 PMID: 39797413

3. Dai JY, Stanford JL, LeBlanc M. A multiple-testing procedure for high-dimensional mediation hypotheses. J Am Stat Assoc. 2022;117(537):198–213. https://doi.org/10.1080/01621459.2020.1765785 PMID: 35400115
4. Dhar GA, Saha S, Mitra P, Nag Chaudhuri R. DNA methylation and regulation of gene expression: Guardian of our health. Nucleus (Calcutta). 2021;64(3):259–70. https://doi.org/10.1007/s13237-021-00367-y PMID: 34421129
5. Ding J, Zhu X. Amdp: an adaptive detection procedure for false discovery rate control in high-dimensional mediation analysis. Advances in Neural Information Processing Systems. 2024;36.
6. Nguyen HH, Takata R, Akamatsu S, Shigemizu D, Tsunoda T, Furihata M, et al. IRX4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin D receptor, conferring prostate cancer susceptibility. Hum Mol Genet. 2012;21(9):2076–85. https://doi.org/10.1093/hmg/dds025 PMID: 22323358
7. Harris SE, Riggio V, Evenden L, Gilchrist T, McCafferty S, Murphy L, et al. Age-related gene expression changes, and transcriptome wide association study of physical and cognitive aging traits, in the Lothian Birth Cohort 1936. Aging (Albany NY). 2017;9(12):2489–503. https://doi.org/10.18632/aging.101333 PMID: 29207374
8. Lange T, Hansen JV. Direct and indirect effects in a survival context. Epidemiology. 2011;22(4):575–81. https://doi.org/10.1097/EDE.0b013e31821c680c PMID: 21552129
9. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012;28(6):882–3. https://doi.org/10.1093/bioinformatics/bts034 PMID: 22257669
10. Leek JT, Storey JD. Capturing heterogeneity in gene expression studies by surrogate variable analysis. PLoS Genet. 2007;3(9):1724–35. https://doi.org/10.1371/journal.pgen.0030161 PMID: 17907809
11. Zhonghua L, Shen J, Barfield R, Schwartz J, Baccarelli AA, Lin X. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. Journal of the American Statistical Association. 2022;117:67–81.
12. Maas SCE, Mens MMJ, Kühnel B, van MeursJBJ, UitterlindenAG, Peters A, et al. Smoking-related changes in dna methylation and gene expression are associated with cardio-metabolic traits. Clinical Epigenetics, 12:1–16, 2020.
13. MacKinnon DP, Lockwood CM, Hoffman JM, West SG, Sheets V. A comparison of methods to test mediation and other intervening variable effects. Psychological methods. 2002;7(1):83.
14. Moore LD, Le T, Fan G. DNA methylation and its basic function. Neuropsychopharmacology. 2013;38(1):23–38. https://doi.org/10.1038/npp.2012.112 PMID: 22781841
15. Pearl J. The causal mediation formula–a guide to the assessment of pathways and mechanisms. Prev Sci. 2012;13(4):426–36. https://doi.org/10.1007/s11121-011-0270-1 PMID: 22419385
16. Perera C, Zhang H, Zheng Y, Hou L, Qu A, Zheng C, et al. HIMA2: high-dimensional mediation analysis and its application in epigenome-wide DNA methylation data. BMC Bioinformatics. 2022;23(1):296. https://doi.org/10.1186/s12859-022-04748-1 PMID: 35879655
17. Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992;3(2):143–55. https://doi.org/10.1097/00001648-199203000-00013 PMID: 1576220
18. Sobel ME. Asymptotic confidence intervals for indirect effects in structural equation models. Sociological Methodology. 1982;13:290. https://doi.org/10.2307/270723
19. Somel M, Khaitovich P, Bahn S, Pääbo S, Lachmann M. Gene expression becomes heterogeneous with age. Curr Biol. 2006;16(10):R359-60. https://doi.org/10.1016/j.cub.2006.04.024 PMID: 16713941
20. Storey JD. A direct approach to false discovery rates. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(3):479–98. https://doi.org/10.1111/1467-9868.00346
21. Sun R, McCaw Z, Lin X. Testing a large number of composite null hypotheses using conditionally symmetric multidimensional gaussian mixtures in genome-wide studies. arXiv preprint 2023. https://arxiv.org/abs/2309.12584
22. Sun W, Cai TT. Oracle and adaptive compound decision rules for false discovery rate control. Journal of the American Statistical Association. 2007;102(479):901–12. https://doi.org/10.1198/016214507000000545
23. Tchetgen Tchetgen EJ. On causal mediation analysis with a survival outcome. Int J Biostat. 2011;7(1):Article 33. https://doi.org/10.2202/1557-4679.1351 PMID: 22049268
24. Valeri L, Vanderweele TJ. Mediation analysis allowing for exposure-mediator interactions and causal interpretation: theoretical assumptions and implementation with SAS and SPSS macros. Psychol Methods. 2013;18(2):137–50. https://doi.org/10.1037/a0031034 PMID: 23379553
25. VanderWeele TJ. Mediation analysis: a practitioner's guide. Annu Rev Public Health. 2016;37:17–32. https://doi.org/10.1146/annurev-publhealth-032315-021402 PMID: 26653405
26. Vanderweele TJ, Vansteelandt S. Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol. 2010;172(12):1339–48. https://doi.org/10.1093/aje/kwq332 PMID: 21036955
27. Wang Q, Ma C, Kemmner W. Wdr66 is a novel marker for risk stratification and involved in epithelial-mesenchymal transition of esophageal squamous cell carcinoma. BMC Cancer. 2013;13:137. https://doi.org/10.1186/1471-2407-13-137 PMID: 23514407
28. Wang S, Liu X. The UCSCXenaTools R package: a toolkit for accessing genomics data from UCSC Xena platform, from cancer multi-omics to single-cell RNA-seq. JOSS. 2019;4(40):1627. https://doi.org/10.21105/joss.01627
29. Wu L, Yang Y, Guo X, Shu X-O, Cai Q, Shu X, et al. An integrative multi-omics analysis to identify candidate DNA methylation biomarkers related to prostate cancer risk. Nat Commun. 2020;11(1):3905. https://doi.org/10.1038/s41467-020-17673-9 PMID: 32764609

30. Yang H, Liu Z, Wang R, Lai E-Y, Schwartz J, Baccarelli AA, et al. Causal mediation analysis for integrating exposure, genomic, and phenotype data. Annu Rev Stat Appl. 2025;12:337–60. https://doi.org/10.1146/annurev-statistics-040622-031653 PMID: 41001172

31. Yang HD, Eun JW, Lee K-B, Shen Q, Kim HS, Kim SY, et al. T-cell immune regulator 1 enhances metastasis in hepatocellular carcinoma. Exp Mol Med. 2018;50(1):e420. https://doi.org/10.1038/emm.2017.166 PMID: 29303507

32. Yu Z, Cui Y, Wei T, Ma Y, Luo C. High-dimensional mediation analysis with confounders in survival models. Front Genet. 2021;12:688871. https://doi.org/10.3389/fgene.2021.688871 PMID: 34262599

33. Zhang H, Zheng Y, Hou L, Zheng C, Liu L. Mediation analysis for survival data with high-dimensional mediators. Bioinformatics. 2021;37(21):3815–21. https://doi.org/10.1093/bioinformatics/btab564 PMID: 34343267