Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Andrea N. Lane Date

Statistical Methods for Mediation Analysis of Omics Data

By

Andrea N. Lane
Doctor of Philosophy

Biostatistics and Bioinformatics

Hao Wu, Ph.D.
Advisor

David Benkeser, Ph.D.
Committee Member

Karen Conneely, Ph.D.
Committee Member

Amita Manatunga, Ph.D.
Committee Member

Accepted:

Kimberly Jacob Arriola, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Statistical Methods for Mediation Analysis of Omics Data

By

Andrea N. Lane
B.S.P.H., UNC Chapel Hill, NC, 2016
B.A., UNC Chapel Hill, NC, 2016

Advisor: Hao Wu, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2022

Abstract

Statistical Methods for Mediation Analysis of Omics Data
By Andrea N. Lane

Epigenome-wide association studies (EWAS) have identified associations between epigenetic modifications (e.g. DNA methylation) and both exposures (e.g. smoking status) and certain health outcomes (e.g. lung function). These associations naturally lead to an interest in studying DNA methylation as a potential mediator between an exposure and outcome. To this point, EWAS mediation studies adopted the canonical mediation analysis method and ignored a very important aspect of the data: the sample complexity. The samples being studied (such as blood) are comprised of a mix of cell types. Distinct cell types are known to present distinct methylation profiles and play unique mediation roles in disease pathogenesis.

In this dissertation, we develop novel statistical methods to study the cell type-specific mediating effects from population level EWAS data. In the first project, we present a novel statistical method called TOols for the Analysis of heterogeneouS Tissues – Mediation with a Continuous outcome (TOAST-MC) to detect this cell-type-specific mediation effect with a continuous outcome. Our method extends the traditional mediation models by treating the unobserved cell type-specific methylation as missing data. We then derive an EM-algorithm for parameter estimations and perform a bootstrap test of the indirect effect.

In the second project, we develop a procedure called TOAST-MB that can handle both a continuous and a binary outcome. The method utilizes a Bayesian model framework to obtain a marginal posterior distribution of the indirect effect for each cell type. Posterior samples are obtained via Hamiltonian Monte Carlo MCMC.

In the third project, we conduct a series of simulation studies to compare the performance of three methods of high dimensional mediation analysis: HIgh dimensional Mediation Analysis (HIMA), Divide-Aggregate Composite-null Test (DACT), and BAyesian Mediation Analysis (BAMA). We then apply the three methods to a dataset from the Grady Trauma Project, in which we analyze the role of DNA methylation as a mediator between smoking and weight.

The statistical methods and tools developed in this dissertation help to better analyze EWAS data and can potentially aid in the discovery of novel diagnostic biomarkers and therapeutic targets.

Statistical Methods for Mediation Analysis of Omics Data

By

Andrea N. Lane
B.S.P.H., UNC Chapel Hill, NC, 2016
B.A., UNC Chapel Hill, NC, 2016

Advisor: Hao Wu, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics and Bioinformatics
2022

Acknowledgments

I am so grateful for the time I have had in the Emory Department of Biostatistics and Bioinformatics. I have crossed paths with so many kind and supportive people during the last six years, and for that I cannot adequately express my gratitude. I came to Emory hoping to find a supportive, collaborative, engaging community, and my experience went above and beyond my expectations.

To my advisor, Dr. Hao Wu, thank you for your generosity of time and spirit. Thank you for believing in me when I lacked confidence, for challenging me to grow as a researcher, and for sharing your passion for your work. Your energy and optimism are contagious, and these qualities helped me make it to the finish line. It was truly a pleasure to work with you.

To my committee members, Dr. David Benkeser, Dr. Karen Conneely, and Dr. Amita Manatunga, I thank each of you for playing a unique role in my PhD journey. David, thank you for challenging me to look at the big picture and conduct rigorous scientific research. Thank you for always being willing to chat, whether it is about mediation, collaborative problems, or career advice. Amita, thank you for your guidance and mentorship over the course of my time at Emory. You supported me through highs and lows and challenged me to work hard in all phases of the program. Karen, thank you for offering your time, advice, and resources to help me complete this dissertation.

To my mentors and collaborators, Dr. Renee' Moore, Dr. Anke Huels, and Dr. Lance Waller, thank you for supporting me and helping me grow. Renee', you made me a better scholar, collaborator, professional, teacher, colleague, and person. I give you so much credit for helping me to transition from being a student to a professional. You challenged me when I needed to be challenged, and listened when I needed a kind

ear. I am so glad I had the opportunity to work with you in so many different capacities, and I am even more grateful that our collaboration does not end here. Anke, thank you for offering me the experience to collaborate with a multidisciplinary team. The skills I have learned from working with you will undoubtedly benefit me going forward. Thank you for your kindness, humor, and advice. Lance, thank you for offering encouragement and advice at every step during this PhD journey. Thank you for giving me opportunities to teach and serve in the department; these opportunities have been some of my favorite experiences here.

To the department staff, Mary Abosi, Angela Guinyard, Melissa Sherrer, Joy Hearn, and Bob Waggoner, thank you for always being there to help (and chat). When I think of the supportive environment at Emory, I think of you first and foremost. You all went out of your way to ensure I had everything I needed to succeed in this program. Mary, thank you for your friendship. You always checked in on me throughout my time here, even in the midst of a pandemic. I loved that the first thing I would do when I got to school was head straight to your desk to say hello.

To my QUEST colleagues, Josh and Raphiel, thank you first and foremost for getting my through the qualifying exams. I would not have made it without you, and my memories of that time even have a glimmer of happiness thanks to you. You both inspire me to do great work and achieve great things. I am so glad we have continued to support each other and I hope that QUEST lives on forever.

To my peers in the biostatistics department, particularly my BIOS Ladies, thank you for the many shared laughs (and meals). Hannah, thank you for all of the discussions about anything and everything, and specifically those that occurred in the aisles at the N Druid Hills Target.

To my "Day 1s", Nancy, Elliot, Alexia, and Emily, thank you for being amazing friends through the highs and lows of this PhD journey. Nancy, thank you for going above and beyond as a mentor and friend. You always offered a listening ear, advice, and encouragement at just the right time. Elliot, this dissertation would probably not exist without you. Thank you for the late night Facetime calls to discuss anything from EM algorithm to reality TV drama. Even though you live many miles away, you have always been there for me any time I needed anything. Alexia, thank you for the many game nights, nights out, and words of encouragement. Your love of math and statistics inspires me. Emily, I am so thankful we were on this journey together in our parallel biostatistics lives. I am so grateful for our long distance friendship and cannot wait to see what the future holds for each of us in this new chapter.

Music has been a big part of my life, so I would like to acknowledge a few songs that will always remind me of my time at Emory: "Vines" by Hippo Campus, "Sedona" by Houndmouth, "Dog Days Are Over" by Florence + The Machine, and the Father of the Bride album by Vampire Weekend.

To my family, thank you for loving me for me, regardless of anything I might accomplish. Thank you for supporting and challenging me throughout this journey, and for reminding me that life is much bigger than graduate school. I am so happy to be moving back to North Carolina to be closer to you all.

Finally, to my sweet dog, Greta, thank you for coming into my life at just the right time. I love that no matter how great or not-so-great my day has been, you are there to remind me that life is just about treating yourself and enjoying the sunshine.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The omics revolution

In the last twenty years, developments in high-throughput technologies have dramatically reduced the cost of genetic sequencing [14]. These technological advancements have led to a new era of biological discovery known as the "omics revolution" [3]. Several areas of "omics" have emerged, including genomics, epigenomics, transcriptomics, proteomics, and metabolomics. Genomics refers to the examination of variability in the genome; epigenomics refers to epigenetic modifications of DNA; transcriptomics refers to the assessment of gene expression via mRNA; proteomics refers to the analysis of proteins; and metabolomics refers to variation in metabolites [3]. In public health research, the emergence of these new types of data present exciting opportunities to achieve a better understanding of the biological mechanisms underlying complex diseases. Omics data also present unique statistical challenges that must be addressed to effectively answer research questions about how these processes are related to health outcomes.

Research questions (and therefore statistical methods) have largely been focused on

finding associations between different types of omics and health outcomes. Because Genome-Wide Association Studies (GWAS) preceded other types of omics analyses, statistical focus was first on developing methods to analyze genomic data. These methods primarily sought to address the inflated rate of false discoveries that can result from multiple testing [4]. Because each type of omics data presents unique statistical challenges due to differing characteristics and correlation structures, new statistical methods have emerged to assess associations for each area of omics, largely building on the foundation laid by genomics analysis methods [11].

In this dissertation, we primarily focus on applications in epigenomics, which refers to the analysis of variability in epigenetic modifications to DNA [3]. Epigenetics is comprised of changes that occur to DNA that do not alter the genetic sequence itself, but rather, can alter the way the genetic material functions [49]. Epigenetic mechanisms include histone modifications and DNA methylation. Histone modifications refer to the alteration of chromatin structure, which can change gene expression by affecting the access to DNA for transcription [5]. Most Epigenome-Wide Association Studies (EWAS), however, have focused on DNA methylation (DNAm). DNA methylation is the epigenetic phenomenon in which a methyl group attaches to a cytosine nucleotide [49], as shown in figure 1.1. Most often, this occurs on a cytosine nucleotide that precedes a guanine nucleotide, and the pair is referred to as a CpG site. In EWAS, associations are assessed between patterns of DNAm at several CpG sites (often as high as 400-800K sites) and various health outcomes. Figure 1.2 illustrates the general workflow for a case-control EWAS.

EWAS have resulted in the discovery of relationships between DNA methylation and many health-related factors, including diet, air pollution, tobacco smoke, reproductive conditions, asthma, neurological disorders, and cancer [16]. While some of these,

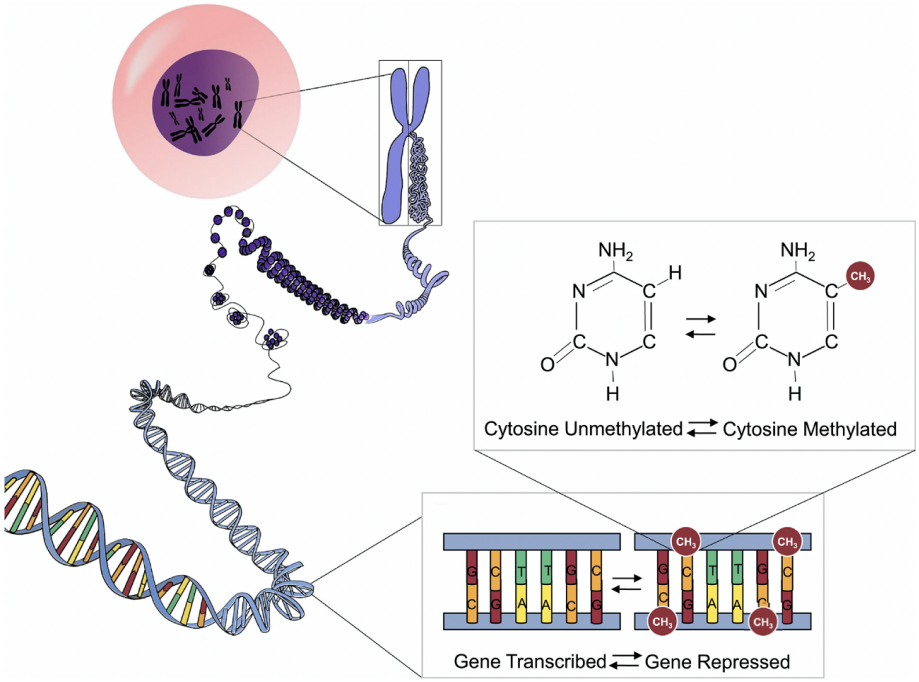

Figure 1.1: DNA methylation [25]

like diet, air pollution, and tobacco smoke, can be categorized as exposures, others, such as asthma and cancer, can be categorized as outcomes. Categorizing these associations in this way has led to increased interest in moving beyond association analyses to investigate causal relationships with DNA methylation.

1.2 Omics mediation

More recently, interest has arisen in multiple areas of omics to move beyond association analyses and investigate causal relationships [53, 55]. Investigating causal relationships with DNAm is particularly intriguing because methylation is a dynamic process that can change over time [49]. More specifically, researchers have become more interested in studying DNAm as a mediator between an exposure and an outcome.

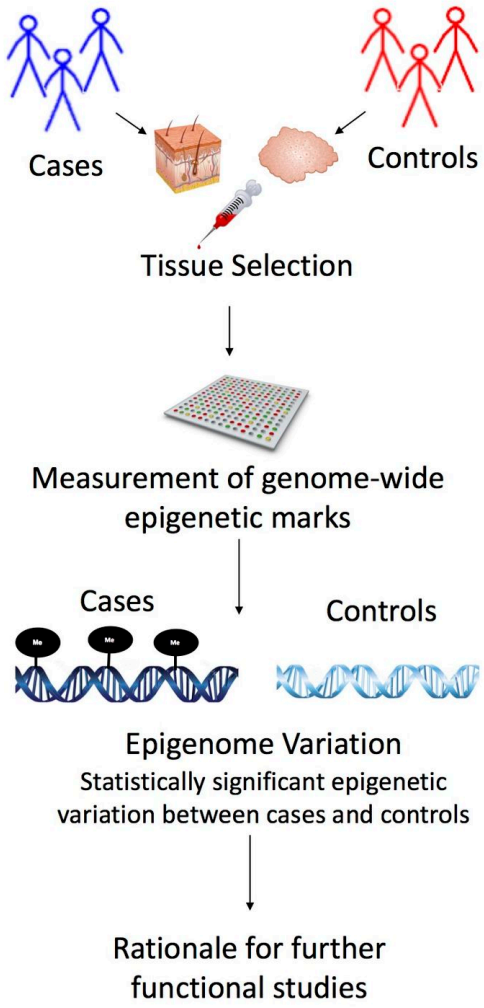

Figure 1.2: Case-control EWAS workflow [1]

Mediation analysis is generally defined as the investigation of how one variable, known as the mediating variable, operates along the causal pathway between an exposure and an outcome. A diagram of a simple mediation model is shown in figure 1.3. Mediation analysis has been heavily used in the social sciences since the seminal Baron and Kenny publication in 1986 [7]. More recently, however, the field of causal inference has transformed mediation analysis by defining causal effects and specifying the necessary assumptions to draw causal conclusions [51, 59, 44, 72]. As the statistical approaches to mediation have modernized, mediation analysis has expanded into several scientific fields, including omics studies.

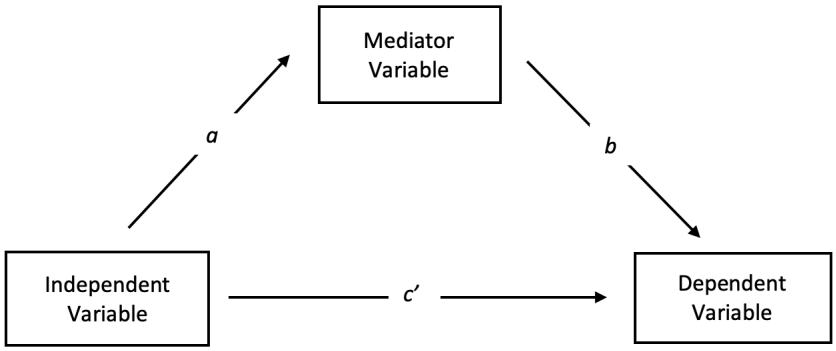

Figure 1.3: A mediation model [2]

Omics, and specifically DNAm, has been investigated as a mediator between exposures and outcomes including age and diabetes risk [26], genetic risk and rheumatoid arthritis [40], prenatal adversity and metabolic disease [70], and genetic variants and multiple sclerosis [35]. These analyses face three main challenges that motivate the work in this dissertation: 1) cell type heterogeneity, 2) assessing causality, and 3) high-dimensionality.

1.3 Outline

In this dissertation, we present novel statistical methods that seek to address the methodological challenges in EWAS mediation. Chapters 2 and 3 primarily focus on the first two challenges: cell type heterogeneity and assessing causality. In chapter two, we present a novel method called TOols for the Analysis of heterogeneouS Tissues - Mediation with a Continuous outcome (TOAST-MC). TOAST-MC detects mediators at the cell type-specific level by using EM algorithm and a bootstrap procedure. In chapter 3, we present TOols for the Analysis of heterogeneouS Tissues - Mediation with a Binary outcome (TOAST-MB). Here we utilize a Bayesian framework to detect cell type-specific mediators when the outcome of interest is binary. In

chapter 4, we address the third challenge of EWAS mediation, high-dimensionality. In this chapter, we present an overview of current high-dimensional mediation statistical methods and then compare three of those methods in a simulation study. In the final chapter, we present a discussion of current challenges and future research plans in omics mediation.

Chapter 2

Detecting cell type-specific mediation effects from bulk omics data with continuous outcomes

2.1 Introduction

In recent years, health researchers have utilized Epigenome-Wide Association Studies (EWAS) to elucidate factors associated with various diseases that cannot be explained by genetics alone [57]. Specifically, EWAS typically measure DNA methylation (DNAm), an epigenetic mechanism in which a methyl group attaches to a cytosine nucleotide, potentially impacting gene expression [49]. DNA methylation has now been shown to be associated with several exposures and diseases, including diet, air pollution, tobacco smoke, reproductive conditions, asthma, neurological disorders, and cancer [16]. In fact, a recent catalogue of EWAS identified over 1 million associations from 342 peer-reviewed publications [8].

Because EWAS have found associations between DNA methylation and exposures

(e.g., smoking), and between DNA methylation and health outcomes (e.g., lung cancer), interest has naturally arisen in studying DNA methylation as a mediator between an exposure and an outcome [16]. For example, DNA methylation has been investigated as a medatior between age and diabetes risk [26], genetic risk and rheumatoid arthritis [40], prenatal adversity and metabolic disease [70], and genetic variants and multiple sclerosis [35]. But EWAS mediation analyses face three key methodological challenges: 1) accounting for cell type heterogeneity, 2) assessing causality, and 3) dealing with high dimensionality.

DNA methylation is often measured at the bulk tissue level, meaning multiple cell types are mixed together. Figure 2.1 illustrates the structure of this bulk level methylation data. For example, samples may include blood, tumor, or brain tissues; each of these tissue types are comprised of different cell types, each of which may present unique methylation profiles [58]. A single sample can contain millions of cells and while the technology to sort cell types does exist, these processes are expensive and laborious. An ongoing challenge of DNAm mediation analyses, and DNAm association studies in general, is not only how to properly account for distinct cell types, but also how to identify biologically meaningful cell type-specific signals. Despite the differential functions of unique cell types, initial attention was primarily aimed at accounting for cell type proportions as a confounding factor rather than a primary source of scientific interest [31], [83], [69], [48]. In DNAm association studies, cell type proportions (or their principal components) are often included in statistical models to control for the confounding effect.

More recently, however, scientific interest has grown in identifying cell type-specific effects in DNAm association analyses. For example, Chan et al. sought to identify cell type-specific effects in major depressive disorder [12]. Accordingly, statistical methods

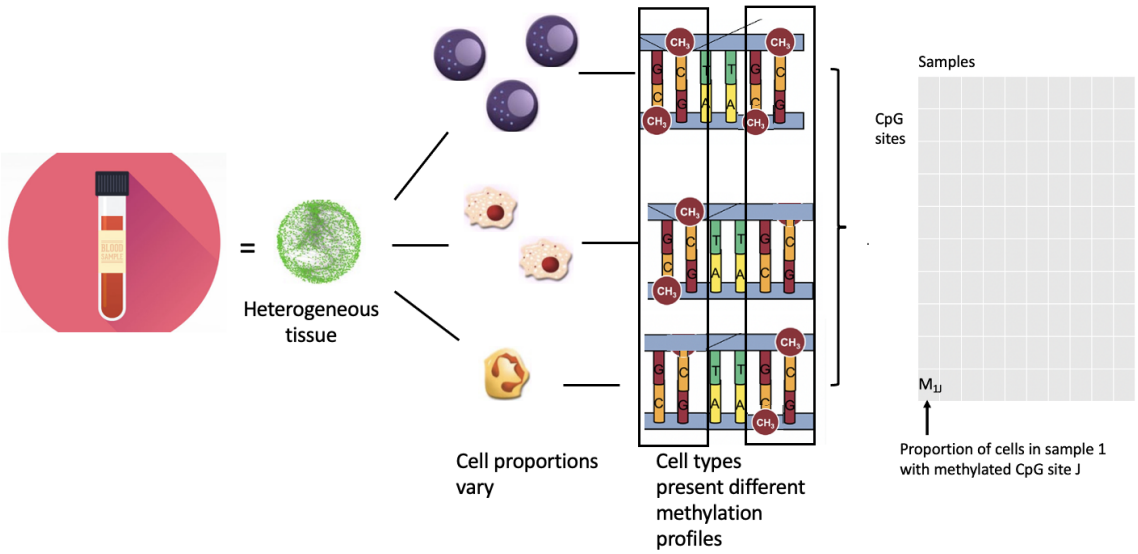

Figure 2.1: Bulk level omics data collection [61, 25]

have been developed to identify cell type-specific effects in DNAm association studies. TOols for the Analysis of heterogeneouS Tissues (TOAST) utilizes a linear model to dissect cell type-specific signals and will be discussed in more detail in the methods section [39]. HIgh REsoluation (HIRE) employs the Expectation-Maximization algorithm to detect cell type-specific effects [42], and Tensor Composition Analysis (TCA) estimates the individual cell type methylation values to assess associations [56].

Mediation analysis seeks to assess the effect of an exposure on an outcome that operates through a third variable, known as the mediator. Baron and Kenny proposed the product of coefficients approach that is often used for such analyses [7]. More recently, causal inference principles have been proposed to clarify the assumptions and methods necessary to assess a causal mediation effect. Assessing mediation is inherently a causal question, so analyses should take these principles into account when studying DNA methylation as a mediator [51]. The causal inference approach to mediation analysis involves three steps: 1) effect definition, 2) effect identification, and 3) effect estimation.

Effect definition refers to the process of selecting which specific causal mediation effect is of interest based on the research question [51]. The total effect is defined as the effect of the exposure on the outcome. This includes both the effect of the exposure on the outcome that operates through the mediator and the remaining effect of the exposure on the outcome that does not go through the mediator. The total effect can be decomposed into the sum of the natural indirect effect and the natural direct effect. The natural indirect effect refers to the effect of the exposure on the outcome that operates through the mediator, and this is what is most often of interest in mediation analyses. The natural direct effect describes the effect of the exposure on the outcome that does not operate through the mediator [71]. Another class of effects called interventional effects include the controlled direct effect, which is the effect of the exposure on the outcome if the mediator were set to a specific value. In this work, we do not focus on interventional effects because we are interested in mediation from an explanatory perspective rather than an interventional perspective [51].

The second step of the causal mediation analysis process is effect identification, which involves careful consideration of particular assumptions to determine if the effect of interest can be learned from the data [51]. In this work, the natural indirect is of primary interest, so we will focus on the assumptions needed to identify the natural indirect effect. Identifying a natural indirect effect in mediation studies relies on the following assumptions: 1) no unmeasured confounding of the exposure-outcome relationship, 2) no unmeasured confounding of the mediator-outcome relationship, 3) no unmeasured confounding of the exposure-mediator relationship, and 4) no mediator-outcome confounder that is affected by the exposure [71]. Identification also inherently relies on the assumption of appropriate temporality, i.e., the exposure precedes

the mediator and the mediator precedes the outcome.

The third step in causal mediation analysis, effect estimation, refers to the process of selecting a model for the mediator and the outcome and deriving the effect(s) of interest from those models. The natural indirect effect is defined in counterfactual notation as $E[Y(e, M(1)] - E[Y(e, M(0)]$, given an outcome Y, a binary exposure e that takes values 0 or 1, and a mediator M. The counterfactual notation $M(1)$ refers to the value the mediator M will naturally take when the exposure is set to $e = 1$. Similarly, $Y(e, M(1)$ refers to the value of the outcome if the exposure were set to level e and the mediator takes the value it would take if $e = 1$. If main effects models are used, as they are in this work, the natural indirect effect coincides with the traditional Baron and Kenny definition of the indirect effect [71].

Developing statistical methods for mediation in the context of EWAS is an ongoing and rich area of research. Of primary focus has been the problem of high dimensionality, given that DNAm data can include hundreds of thousands of CpG sites. For example, Illumina Infinium microarray measures the methylation levels for between 400,000 and 800,000 CpG sites. Some approaches have been to analyze DNAm mediation at the region or gene level, or to employ dimension reduction techniques [76], [19], [21], [22], [79]. However, less attention has gone toward the problem of cell type heterogeneity and identifying cell type-specific mediation effects.

Figure 2.2 depicts the problem of identifying cell type-specific mediation effects. The primary statistical challenge is that the individual cell type methylation values are unobserved. Therefore, this can be treated as a latent class mediator problem, where the observed bulk data is comprised of a weighted sum of the unobserved components. Additionally, decomposing the bulk data transitions the mediation model

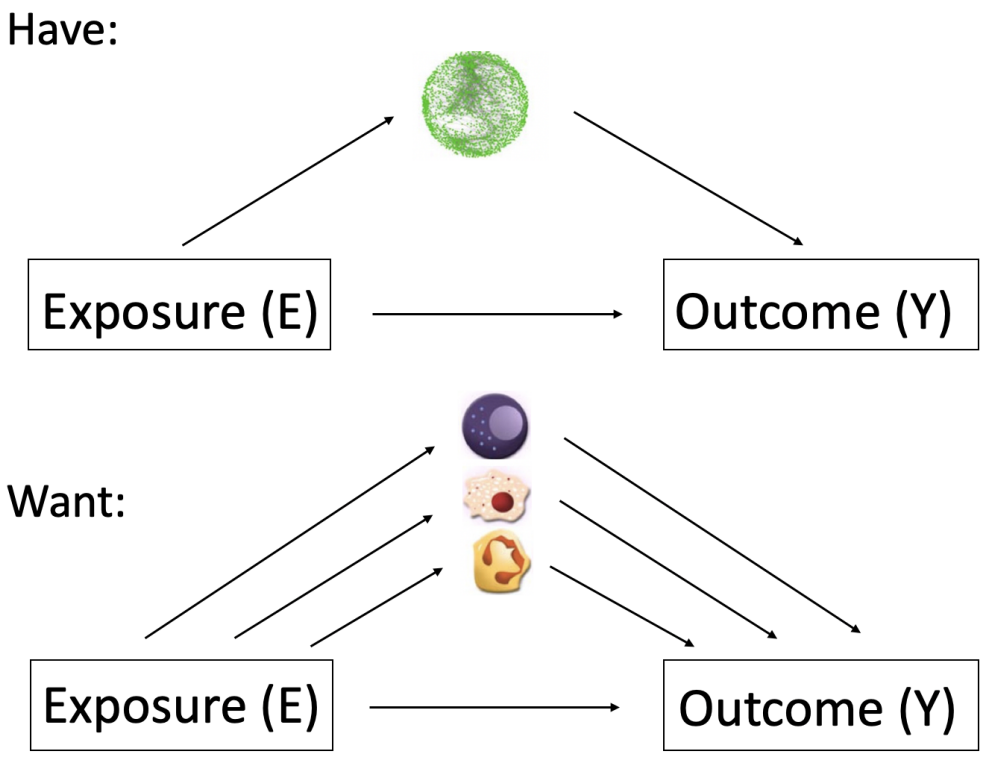

Figure 2.2: Observed data vs. desired cell type-specific mediators [61]

from a single to a multiple mediator model. Luo et. al proposed Mediation in a Cell-type-Specific Fashion (MICS), which uses an inverse linear regression approach to identify mediating cell types [43]. One could also utilize TCA in a two step procedure where the first step is to obtain estimates for the unobserved cell type-specific methylation values, and the second step is to plug the estimates into the mediation models.

We present a novel statistical method called TOAST-MC (TOols for the Analysis of heterogeneouS Tissues - Mediation with a Continuous outcome), which builds on the previously described TOAST method, to detect a cell type-specific mediation effect between an arbitrary exposure and a continuous outcome. The method utilizes the Expectation-Maximization (EM) algorithm. We treat the cell type-specific methylation values as missing data, and the EM algorithm offers a way to obtain parameter estimates in the presence of missing data.

In this paper, we first introduce TOAST-MC, which is comprised of three primary steps to obtain an estimate of the indirect effect and assess statistical significance. We then present results from an extensive simulation study. Finally, we apply the TOAST-MC method to a subset of the Grady Trauma Project dataset.

2.2 Methods

2.2.1 Notation and Models

Assume we have a sample of N individuals in which we have measured the following quantities: an exposure, E, which can be either binary or continuous, a continuous outcome, Y, a matrix of L binary or continuous covariates $\boldsymbol{X_{N\times L}}$, and DNA methylation beta value matrix $\boldsymbol{M}$. The matrix $\boldsymbol{M}$ has dimension $N \times J$, where J is the total number of CpG sites.

DNA methylation values are often measured at the bulk level, meaning the observed values are a weighted sum of the cell type-specific methylation values and the cell type proportions. The cell type proportion matrix may be known but is often estimated using reference-based or reference-free methods [39]. Denote the cell type proportion matrix for K distinct cell types as $\boldsymbol{P_{N\times K}}$. Denoting the (unobserved) cell type-specific methylation value as m_k, the observed methylation vector M is therefore written as the following weighted sum for each subject i and CpG site j:

$$M_{ij} = \sum_{k=1}^{K} m_{ijk} p_{ik} \tag{2.1}$$

We utilize the multiple mediator models proposed by VanderWeele and Vansteelandt

[72] . Although we use separate models for each CpG site, we omit the site index j for simplicity. The **mediator model** assesses the relationship between the mediator and the exposure, and there is a separate model for each mediator (e.g., cell type). The **outcome model** assesses the relationship between the outcome and mediator, accounting for the relationship between the outcome and the exposure.

$$E[m_k|E, X] = \beta_{0k} + \beta_{1k}E + \sum_{l=1}^{L} \beta_{2lk}x_l \tag{2.2}$$

$$E[Y|E, M, X] = \theta_0 + \theta_1 E + \sum_{l=1}^{L} \theta_{2l}x_l + \sum_{k=1}^{K} \theta_{3k}m_k \tag{2.3}$$

Using these models, we can then define the natural indirect effect, the natural direct effect, and the total effect. The natural indirect effect, or the effect of the exposure on the outcome that operates through the mediator, is defined as $E[Y(e, M(1)] - E[Y(e, M(0)] = \beta_{1k}\theta_{2k}$ for each cell type k. Note that we assume no interaction among the cell types to decompose the natural indirect effect from the multiple mediator model into distinct components. The natural direct effect, or the effect of the exposure on the outcome that does not operate through the mediator, is defined as $E[Y(1, M(0))] - E[Y(0, M(0))] = \theta_1$. The total effect is the sum of the natural indirect effect and the natural direct effect, $\beta_{1k}\theta_{2k} + \theta_1$. TOAST-MC obtains an estimate and significance test of the natural indirect effect, as that effect is of primary scientific interest in EWAS mediation analysis.

The statistical challenge of cell type-specific mediation analysis is that the cell type-specific methylation values, m_k, are unobserved. Because we will use EM algorithm to obtain parameter estimates, we establish the following classifications for each component of the models:

- Observed quantities: exposure E, continuous outcome Y, number of cell types K, cell type proportions $\boldsymbol{P}$, and bulk level methylation M
- Unobserved quantities: cell type-specific methylation m_k
- quantity of interest: indirect effect for each cell type $\beta_{1k}\theta_{2k}$

We present a three-step procedure (TOAST-MC) to detect a cell type-specific mediation effect in bulk omics data:

1. Utilize TOAST to analyze the cell type-specific exposure-mediator relationship. Use the results to select a subgroup of CpG sites with which to move forward in step 2 and 3.

2. Use EM algorithm to analyze the cell type-specific mediation effect, one CpG site at a time, in the subgroup of CpG sites obtained in step 1. Obtain parameter estimates and calculate the observed indirect effect.

3. Apply a bootstrap procedure to test the significance of the indirect effect in each cell type and CpG site.

Step 1: Use TOAST to analyze exposure-mediator relationship

TOAST provides a computationally efficient method to assess a cell type-specific relationship between DNAm and an exposure of interest [39]. To reduce the number of CpG sites under consideration in steps 2 and 3, the TOAST-MC procedure begins by analyzing the exposure-mediator relationship.

The TOAST model utilizes the relationship between the unobserved cell type-specific methylation values, the observed bulk data, and the cell type proportions to assess cell type-specific associations with a linear model. The model regresses the observed

methylation matrix M on the cell type proportions and the proportion-exposure interaction term. We can see how the model is derived by substituting the expectation of the observed data as follows:

$$E[M_i|E_i, P_i] = \sum_{k=1}^{K} p_{ik} E[m_{ik}] = \sum_{k=1}^{K} (p_{ik}\beta_{0k} + p_{ik}\beta_1 e_i + \sum_{l=1}^{L} p_{ik}\beta_{2lk} x_l) \tag{2.4}$$

Step 2: Use EM algorithm to assess cell type-specific mediation effects

In the second step, TOAST-MC utilizes the Expectation-Maximization algorithm to detect mediating cell types. The EM algorithm allows us to obtain parameter estimates in the presence of missing data [18]. We obtain the complete data log-likelihood by augmenting the missing data to observed data. EM then iterates between the expectation step and the maximization step. The unobserved cell type-specific methylation levels are considered missing data, while the outcome, exposure, bulk methylation, and cell type proportions constitute the observed data. Rewriting equations 1.1, 1.2, and 1.3, we present the components to be used in EM:

$$m_k \sim N(\beta_{0k} + \beta_{1k} E + \sum_{l=1}^{L} \beta_{2lk} x_l, \tau_k^2) \tag{2.5}$$

$$Y \sim N(\theta_0 + \theta_1 E + \sum_{l=1}^{L} \theta_{2l} x_l + \sum_{k=1}^{K} \theta_{3k} m_k, \gamma^2) \tag{2.6}$$

$$M \sim N(\sum_{k=1}^{K} m_k p_k, \sigma^2) \tag{2.7}$$

After augmenting the missing data $\boldsymbol{\mathcal{M}} = \{m_k : 1 \leq k \leq K\}$ to the observed data, the complete-data log-likelihood function has a tractable form:

$$l_c(\mathbf{\Theta}|\mathbf{Y},\mathbf{E},\boldsymbol{M},\boldsymbol{\mathcal{M}},\boldsymbol{P}) = \sum_{i=1}^{n}\big[-\frac{1}{2}log\sigma^2 - \frac{(M_i - \sum_k m_{ik}p_{ik})^2}{2\sigma^2} - \frac{1}{2}\sum_k log\tau_k^2 - \sum_k \frac{(m_{ik} - \beta_0^k - \beta_1^k E_i - \sum_l \beta_l^k x_{il})^2}{2\tau_{ik}^2} - \frac{1}{2}log\gamma^2 - \frac{(Y_i - \theta_0 - \theta_1 E_i - \sum_l \theta_l x_l - \sum_k \theta_2^k m_{ik})^2}{2\gamma^2}\big].$$

Here $\mathbf{\Theta} = \{\boldsymbol{\beta},\boldsymbol{\theta},\tau_k^2,\sigma^2,\gamma^2 : 1 \le k \le K\}$. We will estimate $\boldsymbol{\mathcal{M}}$ in the E-step and $\mathbf{\Theta}$ in the M-step.

E-step

Taking the expectation of the complete-data log-likelihood, we have:

$$\sum_{i=1}^{n} -\frac{1}{2}log\sigma^{2(t)} - \frac{E[(M_i - \sum_k m_{ik}p_{ik})^2|Y_i,M_i,E_i,\Theta]}{2\sigma^{2(t)}} - \frac{1}{2}\sum_k log\tau_k^{2(t)} - \sum_k \frac{E[(m_{ik} - \beta_0^k - \beta_1^k E_i - \sum_l \beta_l^k x_{il})^2|Y_i,M_i,E_i,x_{il},\Theta]}{2\tau_k^{2(t)}} - \frac{1}{2}log\gamma^{2(t)} - \frac{1}{2\gamma^{2(t)}}E[(Y_i - \theta_0 - \theta_1 E_i - \sum_l \theta_l x_{il} - \sum_k \theta_2^k m_{ik})^2|Y_i,M_i,E_i,x_{il},\Theta]$$

To calculate these expectations, we need $f(m_{ik}|Y_i,E_i,M_i,x_{il},\Theta)$

$$f(m_{ik}|Y_i,E_i,M_i,x_{il},\Theta) \propto f(Y_i|E_i,M_{ik},x_{il},;\Theta)f(M_i|m_{ik},p_{ik};\Theta)f(m_{ik}|E_i,x_{il},;\Theta)$$

$$\propto e^{\frac{-(y_i-\theta_0^{(t)}-\theta_1^{(t)}E_i-\sum_l\theta_l^{(t)}x_{il}-m_i^T\theta_2^{(t)})^2}{2\gamma^{(t)2}}} e^{\frac{-(M_i-m_i^T\pi_i)^2}{2\sigma^{(t)2}}} e^{-\frac{1}{2}(m_i-\beta_0^{(t)}-\beta_1^{(t)}E_i-\sum_l\beta_l^{(t)}x_{il})^T\Sigma^{(t)-1}(m_i-\beta_0^{(t)}-\beta_1^{(t)}E_i-\sum_l\beta_l^{(t)}x_{il})}$$

where $m_i = (M_{i1},...,M_{iK})^T$, β_0, β_1, β_l, π_i and θ_2 are k-length vectors, and $\Sigma = diag(\tau_1^2...\tau_k^2)$ We can combine these in this form:

$$\propto e^{-\frac{1}{2}\left[m_i^T(\frac{\theta_2^{(t)}\theta_2^{(t)T}}{\gamma^{(t)2}}+\frac{p_ip_i^T}{\sigma^{(t)2}}+\Sigma^{(t)-1})m_i-2(\frac{(y_i-\theta_0^{(t)}-\theta_1^{(t)}E_i-\sum_l\theta_l^{(t)}x_{il})\theta_2^{(t)T}}{\gamma^{(t)2}}+\frac{M_ip_i^T}{\sigma^{(t)2}}+(\beta_0^{(t)}+\beta_1^{(t)}E_i+\sum_l\beta_l^{(t)}x_{il})^T\Sigma^{(t)-1})m_i\right]}$$

Therefore, the conditional distribution of m_i is $N(\mu_{i*}^{(t)}, \Sigma_{i*}^{(t)})$, where $\Sigma_{i*}^{(t)-1} := \frac{\theta_2^{(t)}\theta_2^{(t)T}}{\gamma^{(t)2}} + \frac{p_ip_i^T}{\sigma^{(t)2}} + \Sigma^{(t)-1}$ and $\mu_{i*}^{(t)}\Sigma_{i*}^{(t)-1} := \frac{(y_i-\theta_0^{(t)}-\theta_1^{(t)}E_i-\sum_l\theta_l^{(t)}x_{il})\theta_2^{(t)T}}{\gamma^{(t)2}} + \frac{M_ip_i^T}{\sigma^{(t)2}} + (\beta_0^{(t)} + \beta_1^{(t)}E_i + \sum_l\beta_l^{(t)}x_{il})^T\Sigma^{(t)-1}$. Note that μ_{i*} is a k-length vector and Σ_{i*} is a $k \times k$ matrix.

So then

$$E[(M_i - \sum_k m_{ik}p_{ki})^2|Y_i, M_i, E_i, x_{il}, p_i, \Theta] = p_i^T\Sigma_{i*}^{(t)}p_i + (M_i - \mu_{i*}^{(t)T}p_i)^2$$

$$\begin{aligned}&E[(Y_i - \theta_0 - \theta_1E_i - \sum_k \theta_2^k m_{ik})^2|Y_i, M_i, E_i, x_{il}, p_i, \Theta]\\ &= \theta_2^T\Sigma_{i*}^{(t)}\theta_2 + (Y_i - \theta_0 - \theta_1E_i - \sum_l \theta_l x_{il} - \mu_{i*}^{(t)T}\theta_2)^2\end{aligned}$$

and

$$\begin{aligned}&E[(m_i - \beta_0 - \beta_1E_i - \sum_l \beta_l x_{il})^T\Sigma^{-1}(m_i - \beta_0 - \beta_1E_i - \sum_l \beta_l x_{il})|Y_i, M_i, E_i, x_{il}, p_i, \Theta]\\ &= (\mu_{i*}^{(t)} - \beta_0 - \beta_1E_i - \sum_l \beta_l^{(t)}x_{il})^T\Sigma^{-1}(\mu_{i*}^{(t)} - \beta_0 - \beta_1E_i - \sum_l \beta_l x_{il}) + \sum_k \frac{\sigma_{i*,kk}^{(t)2}}{\sigma_k^2}\end{aligned}$$

where $\sigma_{i*,kk}^{(t)2}$ is the kth diagonal element of $\Sigma_{i*}^{(t)}$

M-step

Using the general form of derivative of quadratic form, $\frac{d}{dx}x^TAx = x^T(A + A^T)$, we have

$$\begin{aligned}
\frac{dE[l_c]}{d\beta_0^k} = \sum_{i=1}^{n} -(\mu_{i*}^{(t)} - \beta_0^k - \beta_1^k E_i - \sum_l \beta_l^k x_{il})^T(\Sigma^{-1} + \Sigma^{-1T}) := 0 \\
\implies \sum_{i=1}^{n} \mu_*^{(t)} - \sum_{i=1}^{n} \beta_0^k - \sum_{i=1}^{n} \beta_1^k E_i - \sum_{i=1}^{n}\sum_l \beta_l^k x_{il} = 0 \\
\implies n\beta_0^k = \sum_{i=1}^{n}(\mu_{i*}^{(t)} - \beta_1^k E_i - \sum_l \beta_l^k x_{il}) \\
\implies \beta_0^{k(t+1)} = \frac{\sum_{i=1}^{n}(\mu_{i*}^{(t)} - \beta_1^{k(t)} E_i - \sum_l \beta_l^{k(t)} x_{il})}{n}
\end{aligned}$$

Similarly, we have

$$\beta_1^{k(t+1)} = \frac{\sum_{i=1}^{n}(\mu_{i*}^{(t)} - \beta_0^{k(t+1)} - \sum_l \beta_l^{k(t)} x_{il})E_i}{\sum_{i=1}^{n} E_i^2}$$

$$\beta_l^{k(t+1)} = \frac{\sum_{i=1}^{n}(\mu_{i*}^{(t)} - \beta_0^{k(t+1)} - \beta_1^{k(t+1)} E_i - \sum_{s=1}^{l-1} \beta_s^{k(t+1)} x_{is} - \sum_{s=l+1}^{L} \beta_s^{k(t)} x_{is})x_{il}}{\sum_{i=1}^{n} x_{il}^2}$$

$$\theta_0^{(t+1)} = \frac{\sum_{i=1}^{n}(Y_i - \theta_1^{(t)} E_i - \sum_l \theta_l^{(t)} x_{il} - \sum_k (\mu_{i*}^{k(t)} \theta_2^{k(t)}))}{n}$$

$$\theta_1^{(t+1)} = \frac{\sum_{i=1}^{n}(Y_i - \theta_0^{(t+1)} - \sum_l \theta_l^{(t)} x_{il} - \sum_k \mu_{i*}^{k(t)} \theta_2^{k(t)})E_i}{\sum_{i=1}^{n} E_i^2}$$

$$\theta_l^{(t+1)} = \frac{\sum_{i=1}^{n}(Y_i - \theta_0^{(t+1)} - \theta_1^{(t+1)} E_i - \sum_{s=1}^{l-1} \theta_s^{(t+1)} x_{is} - \sum_{s=l+1}^{L} \theta_s^{(t)} x_{is} - \sum_k \mu_{i*}^{k(t)} \theta_2^{k(t)})x_{il}}{\sum_{i=1}^{n} x_{il}^2}$$

$$\frac{dE[l_c]}{d\theta_2} = \sum_{i=1}^{n} \frac{-1}{2\gamma^{2(t)}} \left[\theta_2^T(\Sigma_{i*}^{(t)} + \Sigma_{i*}^{(t)T}) - 2(Y_i - \theta_0 - \theta_1 E_i - \sum_l \theta_l x_{il})\mu_{i*}^{(t)} + 2\theta_2^T \mu_{i*}^{(t)} \mu_{i*}^{(t)T})\right] := 0$$

$$\implies \theta_2^{T(t+1)} = \sum_{i=1}^{n} (Y_i - \theta_0^{(t+1)} - \theta_1^{(t+1)} E_i - \sum_l \theta_l x_{il}) \mu_{i*}^{(t)} (\sum_{i=1}^{n} \Sigma_{i*}^{(t)} + \mu_{i*}^{(t)} \mu_{i*}^{(t)T})^{-1}$$

$$\sigma^{2(t+1)} = \frac{\sum_{i=1}^{n} p_i^T \Sigma_*^{(t)} p_i + (M_i - \mu_{i*}^{(t)T} p_i)^2}{n}$$

$$\gamma^{2(t+1)} = \frac{\sum_{i=1}^{n} \theta_2^{(t+1)T} \Sigma_{i*}^{(t)} \theta_2^{(t+1)} + (Y_i - \theta_0^{(t+1)} - \theta_1^{(t+1)} E_i - \sum_l \theta_l^{(t+1)} x_{il} - \mu_{i*}^{(t)T} \theta_2^{(t+1)})^2}{n}$$

$$\tau_k^{2(t+1)} = \frac{\sum_{i=1}^{n} [(\mu_{i*}^{(t)} - \beta_0^{k(t+1)} - \beta_1^{k(t+1)} E_i - \sum_l \beta_l^{k(t+1)} x_{il})^2 + \sigma_{i*,kk}^{2(t+1)}]}{n}$$

We will calculate the observed log-likelihood to assess convergence.

$$\begin{aligned} Y &\sim N(\theta_0 + \theta_1 E + \sum_l \theta_l x_l + \sum_k \theta_2^k (\beta_0^k + \beta_1^k E + \sum_l \beta_l^k x_l), \sum_k \theta_2^{k2} \tau_k^2 + \gamma^2) \\ M &\sim N(\sum_k p_k (\beta_0^k + \beta_1^k E + \sum_l \beta_l^k x_l), \sum_k p_k^2 \tau_k^2 + \sigma^2) \\ l_o(\Theta | Y, M, P) &= \sum_{i=1}^{n} logN(Y) + logN(M) \end{aligned}$$

where $N(Y)$ and $N(M)$ indicate the normal density evaluated with the above means and variances. Note that p_k depends on i, so the variance of M is different for each i, while the variance of Y is the same for each i.

Step 3: Use a bootstrap procedure to test the significance of the indirect effect in each cell type and CpG site

After obtaining parameter estimate and an estimate of the indirect effect, we utilize a bootstrap procedure to generate a standard error estimate and p-value for the indirect effect. Bootstrap procedures are generally preferred to the Sobel method and the joint significance test as they are more powerful while maintaining type 1 error control [27]. For the bootstrap procedure, we sample N rows of the observed data with replacement, obtain parameter estimates with the resampled data using the EM algorithm, and calculate the estimated indirect effect. A p-value is calculated as the minimum of the number of bootstrapped indirect effect estimates that are greater than zero and the number that are less than zero. To account for multiple testing with multiple cell types and CpG sites, we use the Benjamini-Yekutieli adjustment; this conservative approach accounts for possible dependence among the cell types and CpG sites [10].

2.2.2 Simulation Study

We performed simulation studies to assess the performance of TOAST-MC. We generated cell type-specific methylation values from a beta distribution using mean and standard deviation parameters from purified human blood cell methylation data [58]. Cell type proportions were generated from a dirichlet distribution. We generated 2, 3, and 4 distinct cell types with mean proportions as follows:

- 2 cells: 0.4, 0.6
- 3 cells: 0.2, 0.3, 0.5
- 4 cells: 0.1, 0.2, 0.3, 0.4

The bulk level methylation values were then generated by equation 2.7 with $\sigma^2 = 0.03$.

In the absence of confounders, the exposure was binary with 50% of the sample exposed and 50% unexposed. One mediating cell was designated in each case, and the exposure-mediator effect was induced by adding $\beta_1 \sim Unif(0.1, 0.5)$ to the cell type-specific methylation value in the mediating cell type. The mediator-outcome effect $\theta_2 \sim Unif(0.1, 0.5)$ was multiplied by the cell type-specific methylation in the mediating cell type. The outcome was then generated by adding a direct effect generated from $Unif(0.01, 0.05)$ and random noise ($\gamma = 0.05$).

We generated three types of confounders that should be controlled in mediation analyses: an exposure-mediator confounder, a mediator-outcome confounder, and an exposure-outcome confounder. For confounders related to the exposure, the confounder x was generated from $Unif(0.3, 0.7)$ and the binary exposure was then generated from a binomial distribution with $p = 0.4 + 0.5x$. The mediator and outcome confounder effect sizes were drawn from $Unif(0.01, 0.05)$.

We compared the performance of TOAST-MC to TCA and MICS. Tensor Composition Analysis (TCA) is not specifically a mediation method, but it can be used to generate estimates of the cell type-specific methylation values. These estimates can then be plugged in to the mediation models. Coefficient estimates are obtained through least squares regression, and the estimate of the indirect effect is assessed with the same bootstrap procedure used in TOAST-MC. Mediation in a Cell-type Specific fashion (MICS) utilizes an inverse regression approach to detect cell type-specific mediating CpG sites. Unlike TOAST-MC and TCA, MICS does not provide an estimate of the indirect effect. Therefore, we only included MICS in the ROC curve comparison.

In the simulation study, we sought to assess two key aspects of the TOAST-MC

method compared to TCA and MICS: 1) estimation bias, and 2) ability to detect significant cell types.

2.3 Results

2.3.1 Simulation Study

Figure 2.3 shows the estimation bias of the indirect effect for TOAST-MC and MICS. While TOAST-MC tends to accurately estimate the indirect effect, TCA consistently underestimates the indirect effect. This is not surprising given that the TCA estimates vary in accuracy among the cell types. Cell types with larger proportion tend to be more accurate while cell types with a lower proportion tend to have less accurate TCA estimates. TCA estimates are also less stable when the true methylation value is close to 0 or 1, or when the variance of the cell type-specific methylation is small. Given that TCA underestimates the indirect effect, we would expect to see that TCA is less powerful in detecting significant mediating cell types.

Table 2.1 shows the number of TOAST-MC discoveries for each cell case. The cell type proportion increases going left to right across the table. Given this, we see that power improves as the cell type proportion increases. Power appears to depend more on the cell type proportion than on the number of cell types. The cell 1 proportion in the 2 cell case is 0.4, which has 89 true discoveries. In the 4 cell case, cell 4 has proportion 0.4, and the result is very similar at 86 discoveries. This is also evident in the 3 cell case, cell 1, and 4 cell case, cell 2, which both have proportion 0.2, and have 72 and 69 discoveries, respectively.

Figures 2.4 and 2.5 compare the performance of TOAST-MC, TCA, and MICS in detecting mediating cell types. We present ROC curves in which discoveries in non-

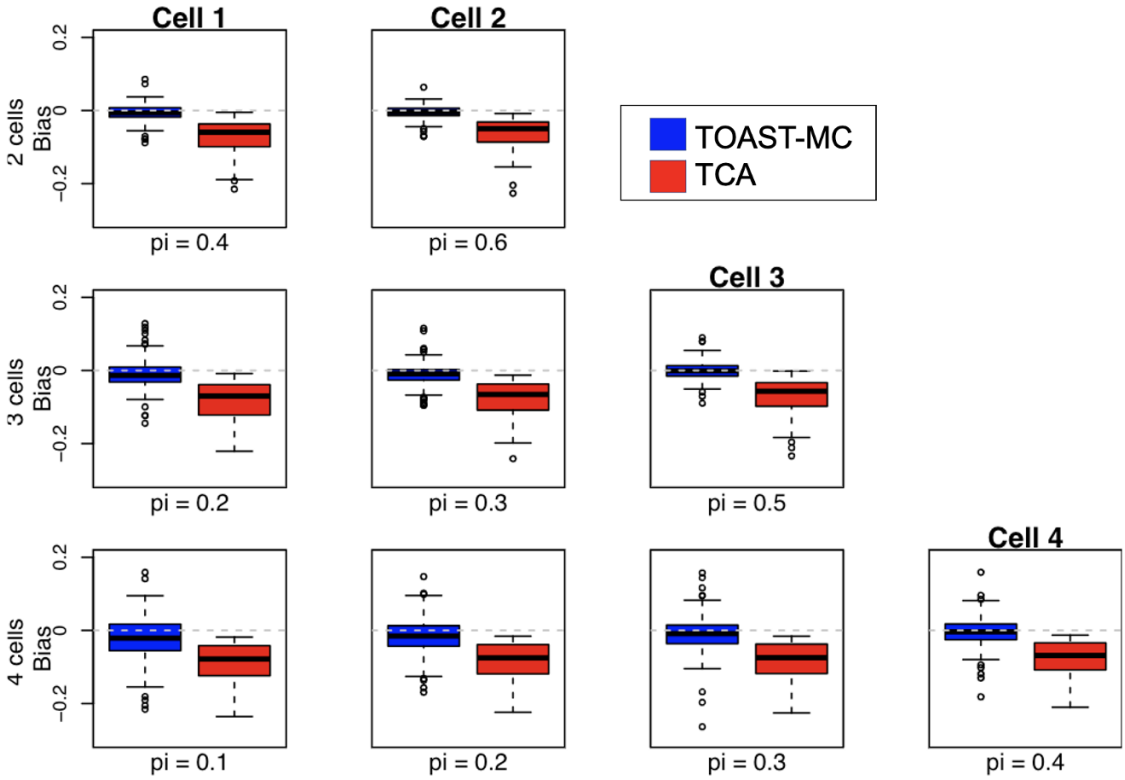

Figure 2.3: Bias in the estimate of the natural indirect effect: Each plot represents a unique simulation setting. The row indicates how many cell types are present in that simulation, and the column indicates which cell type is the correct mediating cell type. The boxplots display the distribution of the difference between the true indirect effect and the estimated indirect effect over 100 simulation replicates. TOAST-MC is shown on the left side of each plot in blue and TCA is shown on the right side in red. Moving from left to right, the proportion of the mediating cell type increases. Note that MICS is not included because it does not give an estimate of the indirect effect.

mediating cell types are categorized as false positives. The sample size in figure 2.4 is 500 and in 2.5 the sample size is 1,000. TOAST-MC consistently has higher AUC than TCA and MICS, indicating superior performance in detecting mediating cell types. As expected, the performance generally improves as the mediating cell type proportion increases (moving from left to right in the figure). Figure 2.6 compares the performance of the three methods in the presence of confounding. Again, TOAST-MC outperforms TCA and MICS in its ability to detect mediating cell types in the presence of an exposure-outcome confounder, a mediator-outcome confounder, or an exposure-mediator confounder.

Number of cells	Mediator	Cell 1	Cell 2	Cell 3	Cell 4
2 cells	Cell 1	**89**	12		
	Cell 2	7	**93**		
3 cells	Cell 1	**72**	7	16	
	Cell 2	11	**81**	10	
	Cell 3	10	16	**95**	
4 cells	Cell 1	**56**	19	14	23
	Cell 2	19	**69**	12	12
	Cell 3	19	14	**80**	19
	Cell 4	16	16	20	**86**

Table 2.1: Simulation results: number of times each cell type was identified as a mediator by TOAST-MC. The row indicates which cell type is the true mediator. Therefore, the diagonal entries (bolded) are true positives and off diagonal elements are false positives. Numbers are out of 100 simulation replicates.

2.3.2 Real Data Analysis

To further evaluate the performance of TOAST-MC, we analyzed a dataset from the Grady Trauma Project, which assesses the influence of various factors, including DNAm, on response to traumatic events [24]. The cohort is predominantly comprised of African American individuals of low socioeconomic status in Atlanta, GA. Data were collected via interviews in waiting rooms of primary care or obstetrical-gynecological clinics. Clinical and life experience data, as well as blood samples, were

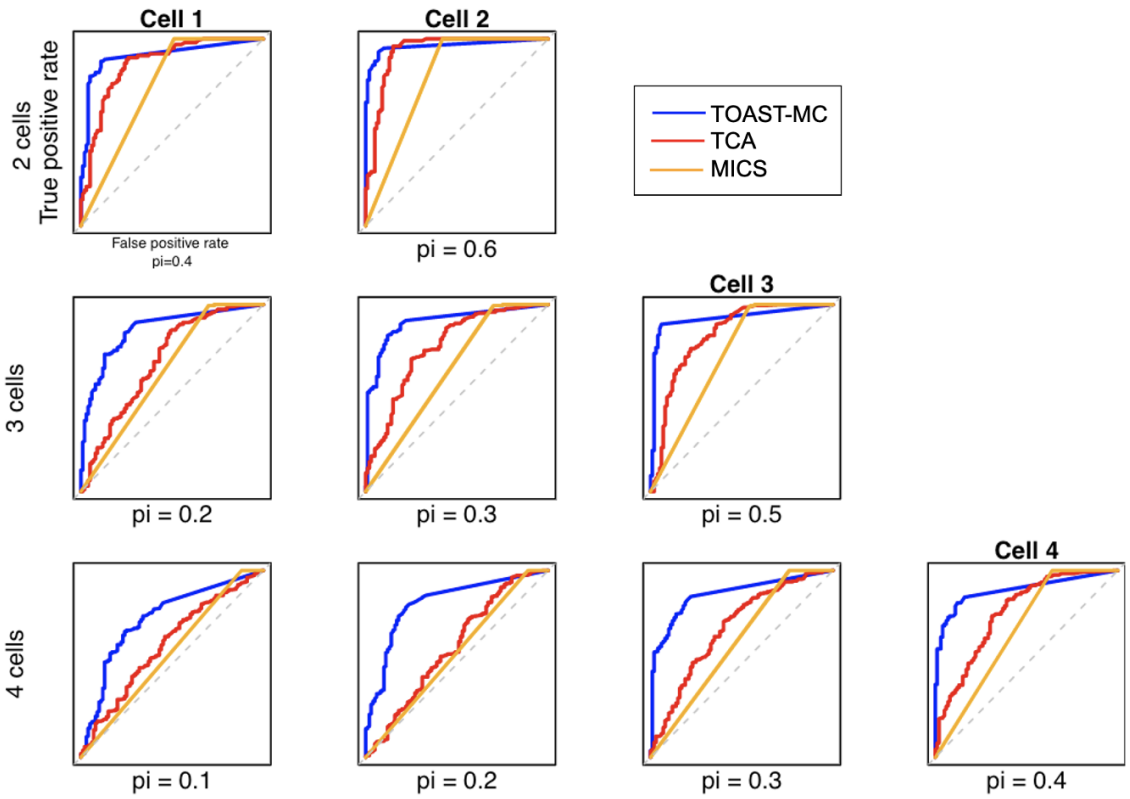

Figure 2.4: ROC curves for TOAST-MC, TCA, and MICS simulation study, N=500: Each plot represents a unique simulation setting. The row indicates how many cell types are present in that simulation, and the column indicates which cell type is the correct mediating cell type. Detections in the true mediating cell type are classified as true positives while detections in the non-mediating cell type(s) are classified as false positives. The proportion of the mediating cell type increases from left to right.

collected. We analyzed 679 individuals for whom EPIC array DNA methylation data were available. Binary smoking status was the exposure of interest and weight (kg) was the continuous outcome. Both smoking status and weight have been shown to be associated with epigenetics [20], [37], [73], [17], [33], [46], [77]. Additionally, cell type-specific associations have been found for both smoking and weight, providing motivation for studying cell type-specific DNAm in a mediation context [75], [9], [52], [68]. Specifically, neutrophil cells have been found to be associated with both smoking and obesity in African American males [75], [9]. Sex and age were controlled for as confounding variables. Observations with missing values for smoking status, weight,

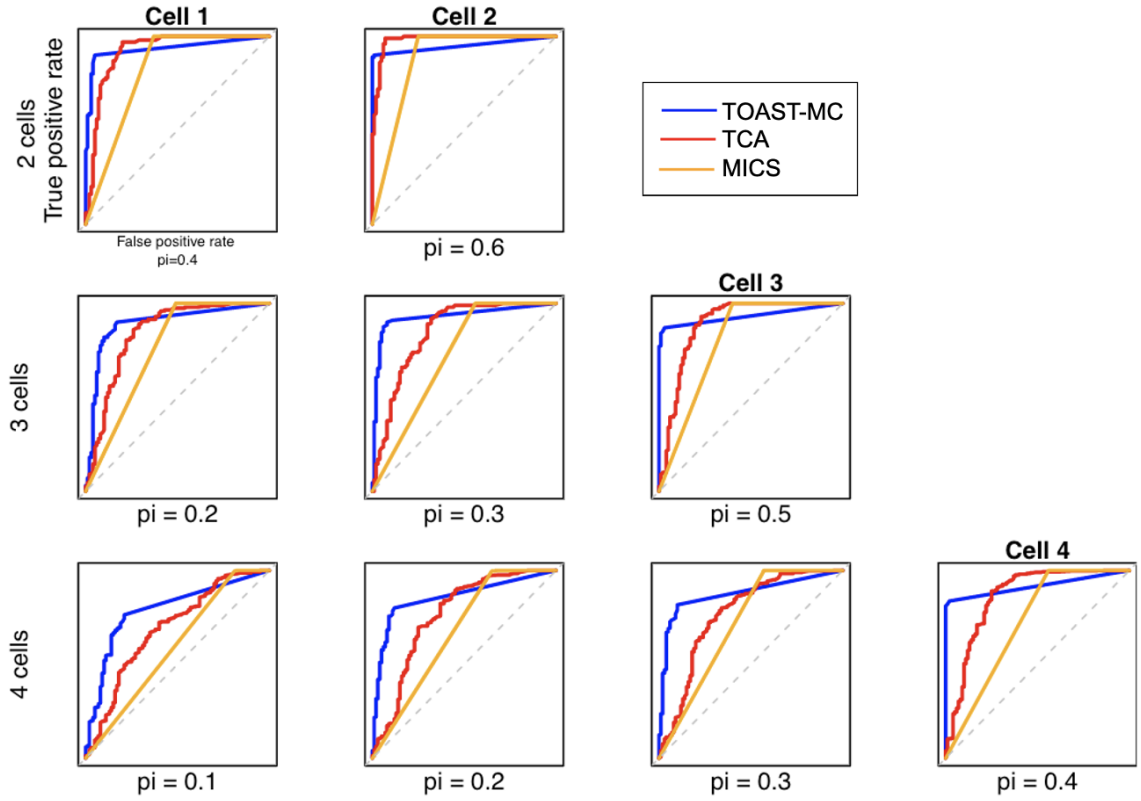

Figure 2.5: ROC curves for TOAST-MC, TCA, and MICS simulation study, N=1000: Each plot represents a unique simulation setting. The row indicates how many cell types are present in that simulation, and the column indicates which cell type is the correct mediating cell type. Detections in the true mediating cell type are classified as true positives while detections in the non-mediating cell type(s) are classified as false positives. The proportion of the mediating cell type increases from left to right.

sex, or age were excluded. Cell type proportions were estimated using publicly available reference data and Robust Partial Correlation (RPF) method implemented in the R package EpiDish [69]. Six cell types were ascertained: CD8T, CD4T, Natural Killer (NK), Neutrophill, B cells, and monocyte cells.

Of the 679 individuals, 273 (40%) were smokers and 406 (60%) were nonsmokers. The mean weight in nonsmokers was 96.7 kg (SD=26.1) and 90.0 (25.1) in smokers (p¡0.001, two sample t-test). 79.8% of nonsmokers were female with a mean age of 40.8 (SD=12.7), while 62.3% of smokers were female with a mean age of 44.5 (SD=11.0).

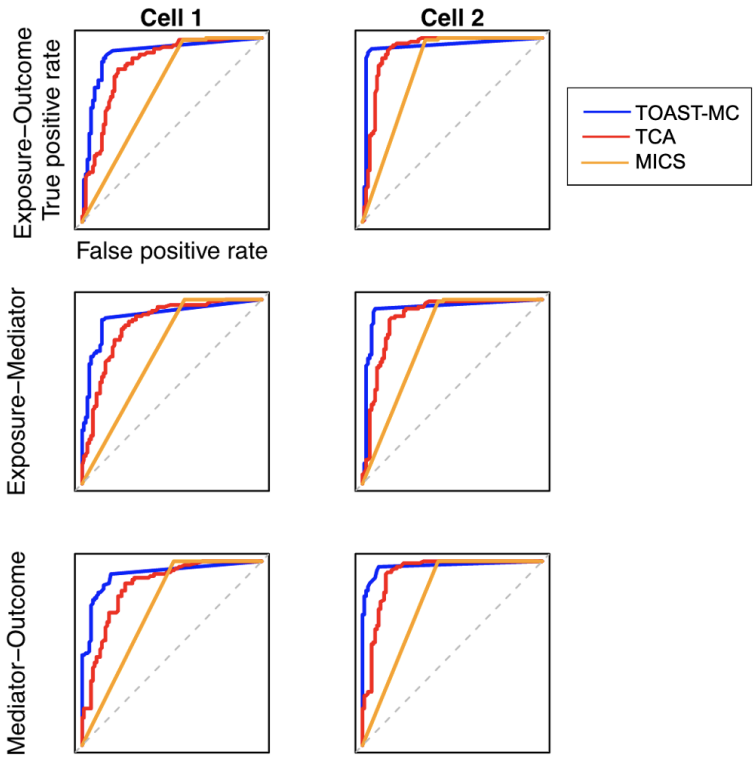

Figure 2.6: ROC curves for performance in the presence of confounding: The row indicates the type of confounder that is present, covering the three primary assumptions to identify the natural indirect effect. The column indicates which cell type is the mediating cell type. Mean cell type proportions are 0.4 and 0.6 for cell 1 and cell 2, respectively.

Mean cell type proportions were as follows: CD8T - 0.11, CD4T - 0.17, NK - 0.05, B cell - 0.07, Monocyte - 0.08, Neutrophill - 0.51.

	CD8T	CD4T	NK	B cell	Mono	Neu
TOAST-MC	185	304	294	351	194	140
TCA	0	0	0	0	0	1
MICS	236	274	25	5	44	1

Table 2.2: Grady Trauma Project data: number of CpG sites detected as mediators for each cell type

In accordance with the first step of TOAST-MC, we used TOAST to select a subset of 1522 CpG sites (out of 819708) using a p-value threshold of $1e-5$. Table 2.2

shows the number of significant CpG sites detected as significant mediators for each cell type. TCA only detects one site in one cell type, indicating very low power for this two step procedure. MICS detects the most CpG sites in CD8T cells, while TOAST-MC detects more CpG sites in every other cell type. A particular result of interest is in neutrophill cells, in which TOAST-MC detects 140 sites while TCA and MICS only detect one site each. In this dataset, neutrophills are present in the highest proportion at around 50%. Given this, we would hope to find a mediating effect specifically in this cell type, which TOAST-MC does.

We also conducted gene and pathyway analyses on the detected CpG sites using Enrichr [13, 36, 74]. The AHRR gene, which is well known to be associated with smoking, is detected in every cell type. Other genes detected include F2RL3, PRSS23, GFI1, LRP5, MYO1G, GNG12, ANPEP, RARA, C5orf62, and ITPK1, which have all been previously shown to be associated with smoking [38]. In neutrophills, GFI1 and F2RL3 were detected, which have been previously shown to be associated with smoking [9]. GFI1 and F2RL3 were also detected in monocytes, which aligns with previous results [9], [68]. Genes detected for B cells, NK cells, and CD4T cells are involved in energy/metabolism pathways which, given that the outcome of interest was weight, provides further evidence of the biological plausibility of the TOAST-MC results. Although we would be hesitant to assert that we adequately controlled for confounding to justify a causal interpretation, these results demonstrate the power of TOAST-MC in exploratory cell type-specific mediation analysis.

2.4 Discussion

We present a novel method, TOAST-MC, to detect a cell type-specific mediation effect and estimate the natural indirect effect. The method involves three steps: first,

we select a subset of CpG sites with TOAST; second, we employ the EM algorithm to estimate the indirect effect for each cell type in each CpG site in the subset; third, we use a bootstrap procedure to assess statistical significance of each cell type and CpG site. TOAST-MC outperforms competing methods TCA and MICS in simulation studies. When applied to a real dataset, TOAST-MC detects more CpG sites in each cell type, and the results are biologically plausible. Although we focus on DNA methylation data, this method can be applied to any bulk omics data in which there is scientific interest in a cell type-specific effect. Even more generally, this method can be framed as a way to detect latent mediators when only a linear combination of those mediators is observed.

To make a causal conclusion from a mediation analysis, one must make strict assumptions regarding the control of confounding. Moreover, in the TOAST-MC bootstrap procedure, we separate the natural indirect effect from the multiple mediator models into individual components. To do this, we assume that the mediators (in this case, cell types) do not interact. We also assume independence among the CpG sites since we analyze them separately in steps two and three of TOAST-MC. Because of these strict assumptions, we recommend using TOAST-MC primarily for exploratory analysis of cell type-specific mediation effects. Much rigor must be employed in the data analysis to justify these assumptions to utilize a causal interpretation of the natural indirect effect. However, we maintain that this method is an important innovation in DNA methylation mediation analysis that can lead to important biological discoveries.

In this method, we also assume that the number of cell types and their proportions are known, and that they operate independently. In reality, cell types have a hierarchical structure, and the cell types identified likely have more specific sub-types. Because of the statistical challenge of identifying cell type-specific signals when cell type-specific

data are unavailable, the method (and any deconvolution method at present) works best for major cell types and up to about 8 different cell types. With too many cell subtypes, signals are weaker and therefore statistical power is limited. Large sample sizes are needed to overcome this challenge.

More research is needed to improve cell type-specific methods. Most notably, a computationally efficient method that jointly analyzes CpG sites and cell types remains an important future research goal. TOAST-MC also requires a continuous outcome, and many DNA methylation mediation analyses have a binary or count outcome of interest. Finally, more work is needed to relax the modeling assumptions of mediation analyses in many areas of omics data to allow for more flexible models, including interaction terms and non-parametric models.

Chapter 3

Detecting cell type-specific mediation effects from bulk omics data with binary outcomes

3.1 Introduction

Epigenome-Wide Association Studies (EWAS) have become a valuable way to study the effects of differential DNA methylation (DNAm) on a variety of diseases. EWAS commonly utilize case-control study designs in which participants are recruited based on the disease of interest [57]. A key advantage of this study design is that it is more cost-effective than longitudinal cohort studies and many case-control cohorts already exist [57]. However, a drawback is that it can be difficult to draw causal conclusions based on a retrospective study design because of the challenges of controlling for all confounding factors. Because case-control design remains the primary study design used in EWAS mediation, it is important for statistical methods to handle a binary outcome in a mediation context.

In mediation analysis, the primary effect of interest is often the natural indirect effect, or the effect of the exposure on the outcome that operates through the mediator [72]. When the outcome is binary, the natural indirect effect is defined on the odds ratio scale as $OR^{NIE} = exp((\theta_2\beta_1)(a - a^*))$, where a and a^* are two levels of the exposure, and θ_2 and β_1 are based on the following mediation models (where E is the exposure, Y is the outcome, and M is the mediator) [71]:

$$E[M|E] = \beta_0 + \beta_1 E \tag{3.1}$$

$$logit[P(Y = 1|E, M)] = \theta_0 + \theta_1 E + \theta_2 M \tag{3.2}$$

Note that the mediator is assumed to be continuous here, as is the case with DNAm mediation. Mediation analysis with a binary outcome relies on an important assumption that the outcome is rare. A prevalence of 10% is typically used as a cutoff. This is because the natural indirect effect is defined on the odds ratio scale and the odds ratio only approximates the risk ratio when the outcome is rare. When the outcome is common, using logistic regression can be a conservative approach to detecting mediation [71]. To address this, a log-linear model can be used to model the outcome when it is binary and common [71]. Moreover, in a case-control study design, it is recommended to fit the mediator model only for the control subjects. This is due to the oversampling of the cases in this study design; if the outcome is rare, then fitting the mediator model only in control subjects will approximate what would be obtained in a cohort study [71].

It is also important to establish temporal relationships to study mediation, and this is particularly true in a case-control study design. One of the primary challenges

of DNAm studies has been establishing causality because of the mutable nature of methylation patterns [57]. Therefore, one must carefully consider temporal ordering when a causal interpretation is desired [71].

Nonlinearities present particular statistical challenges that must be addressed in the case of a binary outcome. In the case of the continuous outcome, we employed EM algorithm to obtain maximum-likelihood estimates. We assumed linear models for the outcome and multiple mediator models, which resulted in closed form solutions for the M-step equations. With a binary outcome, however, we no longer have closed form solutions for the M-step, so optimization algorithms must be employed to obtain estimates. This presents computational challenges that may prove too burdensome in the development of a user-friendly method.

Currently, there is a lack of methods to analyze a cell type-specific mediation effect with a binary outcome. In this paper, we present TOols for the Analysis of heterogeneouS Tissues - Mediation with a Binary outcome (TOAST-MB), which utilizes a Bayesian framework to detect a cell type-specific mediation effect. We first describe the models and the three-step TOAST-MB procedure. We then describe a simulation study to compare the performance of TOAST-MB to TCA and MICS, and apply the three methods to a dataset from the Grady Trauma Project.

3.2 Methods

3.2.1 Notation and Models

Assume we have a sample of N individuals in which we have measured the following quantities: an exposure, E, which can be either binary or continuous, a binary outcome, Y, a matrix of L binary or continuous covariates $\boldsymbol{X_{N\times L}}$, and DNA methylation

beta value matrix $\boldsymbol{M}$. The matrix $\boldsymbol{M}$ has dimension $N \times J$, where J is the total number of CpG sites.

DNA methylation values are often measured at the bulk level, meaning the observed values are a weighted sum of the cell type-specific methylation values and the cell type proportions. The cell type proportion matrix may be known but is often estimated using reference-based or reference-free methods [39]. Denote the cell type proportion matrix for K distinct cell types as $\boldsymbol{P_{N \times K}}$. Denoting the (unobserved) cell type-specific methylation value as m_k, the observed methylation vector M is therefore written as the following weighted sum for each subject i and CpG site j:

$$M_{ij} = \sum_{k=1}^{K} m_{ijk} p_{ik} \tag{3.3}$$

We utilize the multiple mediator models proposed by VanderWeele and Vansteelandt [72] . Although we use separate models for each CpG site, we omit the site index j for simplicity. The **mediator model** assesses the relationship between the mediator and the exposure, and there is a separate model for each mediator (e.g., cell type). The **outcome model** assesses the relationship between the outcome and mediator, accounting for the relationship between the outcome and the exposure.

$$E[m_k|E, X] = \beta_{0k} + \beta_{1k}E + \sum_{l=1}^{L} \beta_{2lk} x_l \tag{3.4}$$

$$logit[P(Y = 1|E, M, X)] = \theta_0 + \theta_1 E + \sum_{l=1}^{L} \theta_{2l} x_l + \sum_{k=1}^{K} \theta_{3k} m_k \tag{3.5}$$

Using these models, we can then define the natural indirect effect, the natural direct effect, and the total effect. The natural indirect effect, or the effect of the exposure

on the outcome that operates through the mediator, is defined on the odds ratio scale as $e^{\beta_1^k \theta_2^k}$ for each cell type k. The natural direct effect, or the effect of the exposure on the outcome that does not operate through the mediator, is defined on the odds ratio scale as e^{θ_1}. For a continuous exposure, the natural indirect effect and natural direct effect are defined on the odds ratio scale as $e^{\beta_1^k \theta_2^k (e-e^*)}$ and $e^{\theta_1 (e-e^*)}$, respectively, where e and e^* are two values of the exposure. The total effect in this case is the product of the natural direct effect and the natural indirect effect. TOAST-MB obtains an estimate and significance test of the natural indirect effect, as that effect is of primary scientific interest in EWAS mediation analysis.

The statistical challenge of cell type-specific mediation analysis is that the cell type-specific methylation values, m_k, are unobserved. Because we will use a Bayesian hierarchical model to obtain posterior distributions of the parameters, we summarize the model components as follows:

- Observed quantities: exposure E, binary outcome Y, number of cell types K, cell type proportions $\boldsymbol{P}$, and bulk level methylation M
- Unobserved quantities: cell type-specific methylation m_k
- quantity of interest: indirect effect for each cell type $e^{\beta_1^k \theta_2^k}$

We present a three-step procedure called TOAST-MB to detect a cell type-specific mediation effect in bulk omics data:

1. Utilize TOAST to analyze the cell type-specific exposure-mediator relationship. Use the results to select a subgroup of CpG sites with which to move forward in step 2 and 3.

2. Use Bayesian hierarchical modeling to fit the three models, one CpG site at a time, in the subgroup of CpG sites obtained in step 1. Obtain posterior distributions for

the indirect effect in each cell type.

3. Categorize each cell type as mediating or non-mediating using posterior probabilities.

Step 1: Use TOAST to analyze exposure-mediator relationship

TOAST provides a computationally efficient method to assess a cell type-specific relationship between DNAm and an exposure of interest [39]. To reduce the number of CpG sites under consideration in steps 2 and 3, the TOAST-MC procedure begins by analyzing the exposure-mediator relationship.

The TOAST model utilizes the relationship between the unobserved cell type-specific methylation values, the observed bulk data, and the cell type proportions to assess cell type-specific associations with a linear model. The model regresses the observed methylation matrix M on the cell type proportions and the proportion-exposure interaction term. We can see how the model is derived by substituting the expectation of the observed data as follows:

$$E[M_i|E_i, P_i] = \sum_{k=1}^{K} p_{ik} E[m_{ik}] = \sum_{k=1}^{K} (p_{ik}\beta_{0k} + p_{ik}\beta_1 e_i + \sum_{l=1}^{L} p_{ik}\beta_{2lk} x_l) \tag{3.6}$$

Step 2: Use Bayesian hierarchical modeling to fit mediation models

Given the models above, assume the following distributions:

$$m_k \sim N(\beta_{0k} + \beta_{1k} E + \sum_{l=1}^{L} \beta_{2lk} x_l, \tau_k^2) \tag{3.7}$$

$$Y \sim Bern(\frac{1}{1+\exp(-(\theta_0+\theta_1 E+\sum_{l=1}^{L}\theta_{2l}x_l+\sum_{k=1}^{K}\theta_{3k}m_k))}) \tag{3.8}$$

$$M \sim N(\sum_{k=1}^{K} m_k p_k, \sigma^2) \tag{3.9}$$

Prior specification

We specify prior distributions for the coefficient parameters β_{0k}, β_{1k}, β_{2kl}, θ_0, θ_1, θ_{2l}, and θ_{3k}, as well as the variance parameters σ^2 and τ_k^2. Although each individual model is not high dimensional, we fit each CpG site independently. To avoid falsely detecting cell types as mediators due to the number of CpG sites tested, we assume small variances in the coefficient priors and a mean of 0; thus, we require strong evidence to arrive at a non-zero effect. The priors for the coefficient parameters are as follows:

$$\theta_0 \sim N(0, 0.1)$$

$$\theta_1 \sim N(0, 0.1)$$

$$\theta_{2l} \sim N(0, 0.1)$$

$$\theta_{3k} \sim N(0, 2)$$

$$\beta_{0k} \sim N(0, 0.1)$$

$$\beta_{1k} \sim N(0, 1)$$

$$\beta_{2lk} \sim N(0, 0.1)$$

For the variance parameters, we specify a $gamma(2, 1)$ as recommended in the Stan User's Guide [66]. Posterior samples are obtained via Markov chain Monte Carlo (MCMC) sampling in Stan [67]. Specifically, Stan uses the No-U-Turn variant of Hamiltonian Monte Carlo (HMC), which has been shown to be efficient and robust [28].

Step 3: Categorize each cell type as mediating or non-mediating

Because we obtain a marginal posterior distribution for β_{1k} and θ_{2k}, we can obtain a posterior distribution for the odds ratio of the natural indirect effect for each cell type, $\exp(\beta_{1k}\theta_{2k})$. We classify a cell type as a mediator if $\min(P(\exp(\beta_{1k}\theta_{2k}) < 1), P(\exp(\beta_{1k}\theta_{2k}) > 1) < p$. The user can select the probability threshold p, where a lower value represents a more conservative threshold. In the following simulation study and analysis of Grady Trauma Project data, we use $p = 0.05$.

3.2.2 Simulation Study

We performed simulation studies to assess the performance of TOAST-MB. We generated cell type-specific methylation values from a normal distribution with mean 0.5 and standard deviation 0.2. Cell type proportions were generated from a dirichlet distribution. We generated 2, 3, and 4 distinct cell types with mean proportions as follows:

- 2 cells: 0.4, 0.6
- 3 cells: 0.2, 0.3, 0.5
- 4 cells: 0.1, 0.2, 0.3, 0.4

The proportions in the 4 cell type case were varied to investigate the effect of the distribution of cell type proportions. The bulk level methylation values were then generated by equation 3.9 with $\sigma^2 = 0.03$, reflecting the variance seen in bulk methylation data. In the absence of confounders, the exposure was binary with 50% of the sample exposed and 50% unexposed. The sample size was 1000. One mediating cell was designated in each case, and the exposure-mediator effect was induced by adding $\beta_1 = 0.3$ to the cell type-specific methylation value in the mediating cell type. The mediator-outcome effect $\theta_2 = 3$ was multiplied by the cell type-specific methylation in the mediating cell type. The outcome was then generated using a bernoulli distribution with probability of success $1/[1+exp(-(-0.5+O_a+DE\times E))]$, where O_a refers to the mediating cell type-specific methylation after inducing the mediator-outcome effect, and DE refers to the direct effect, generated from $Unif(0.01, 0.05)$.

We generated three types of confounders that should be controlled in mediation analyses: an exposure-mediator confounder, a mediator-outcome confounder, and an exposure-outcome confounder. For confounders related to the exposure, the confounder x was generated from $Unif(0.3, 0.7)$ and the binary exposure was then generated from a binomial distribution with $p = 0.4 + 0.5x$. The mediator and outcome confounder effect sizes were drawn from $Unif(0.01, 0.05)$.

We compared the performance of TOAST-MB to TCA and MICS. Tensor Composition Analysis (TCA) is not specifically a mediation method, but it can be used to generate estimates of the cell type-specific methylation values. These estimates can

then be plugged in to the mediation models. Coefficient estimates are obtained with a logistic regression model, and the estimate of the indirect effect is assessed with a bootstrap procedure. Mediation in a Cell-type Specific fashion (MICS) utilizes an inverse regression approach to detect cell type-specific mediating CpG sites.

3.3 Results

3.3.1 Simulation Study

First, we assessed the performance of the Bayesian hierarchical model via trace plots. An example of the trace plots are shown in figure 3.1. This figure shows the case where there are two cell types and cell type 1 is the mediating cell type. The two parameters shown, β_2 and θ_3 are the parameters used to estimate the indirect effect. We see that the four chains converge around the true values.

Figure 3.2 presents ROC curves to demonstrate the comparative performance of TOAST-MB, TCA and MICS in detecting a mediating cell type. In each case, one cell type is a mediator. TOAST-MB consistently outperforms TCA and MICS. The difference is most pronounced when the mediating cell type proportion is low. As the mediating cell type proportion increases, MICS and TCA improve in their ability to detect the mediating cell type, while TOAST-MB slightly declines; however, TOAST-MB continues to outperform TCA and MICS in each case. Table 3.1 displays these simulation results for the four cell type case using a posterior probability cutoff of 0.05, which represents the cell types that would be selected as mediators. TOAST-MB consistently shows higher power in its ability to detect the mediating cell type compared to TCA and MICS. Additionally, using this probability cutoff, we see that the proportion of false positives for each case does not exceed 0.1. TCA, on the other hand, exhibits lower power and higher false discovery rate compared to

TOAST-MB. MICS appears to be conservative with zero false discoveries and highest power of 0.25.

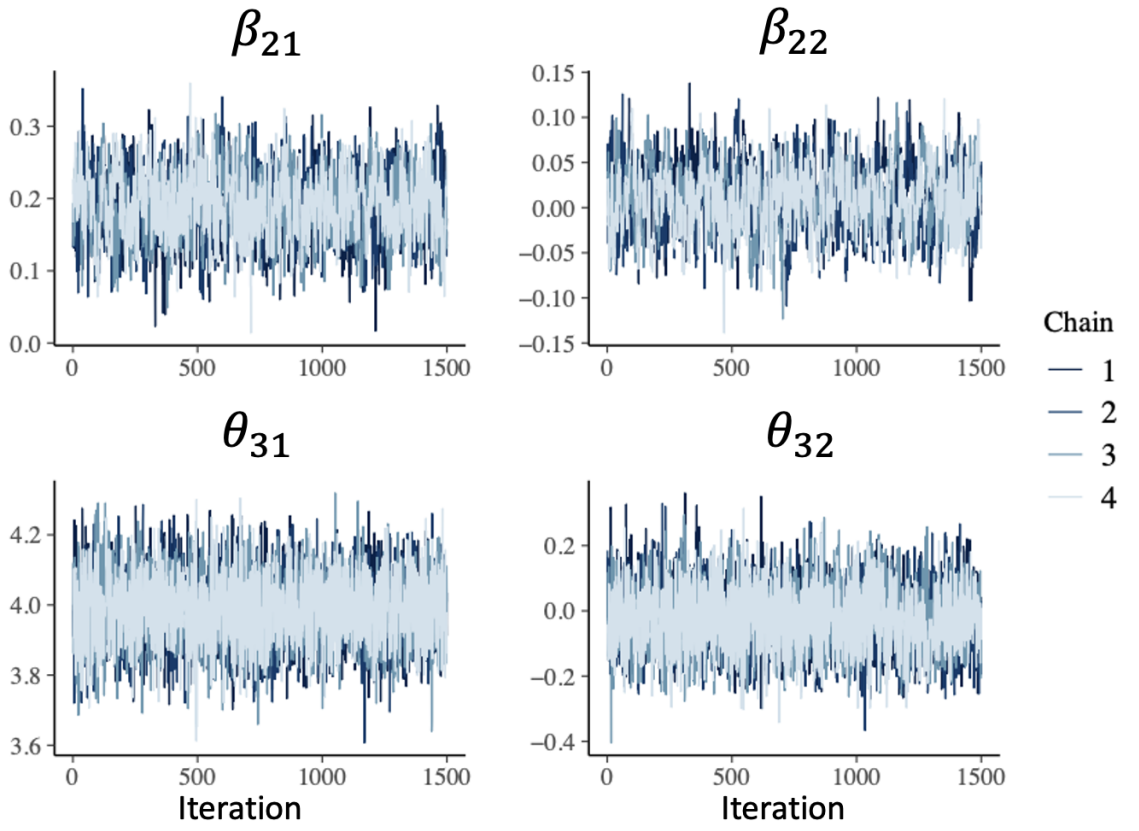

Figure 3.1: Trace plots for the parameters used to calculate the indirect effect. In this case, there are 2 cell types, where the first cell type is the mediating cell type. The true values are as follows: $\beta_{21} = 0.2$, $\beta_{22} = 0$, $\theta_{31} = 4$, $\theta_{32} = 0$

We also assessed the performance of TOAST-MB, TCA, and MICS in the presence of confounders. The key assumptions in causal mediation analysis relate to controlling for exposure-outcome, exposure-mediator, and mediator-outcome confounders to identify the natural indirect effect. In the presence of a single confounder, in this case an exposure-mediator confounder, as shown in figure 3.3, TOAST-MB continues to outperform TCA and MICS. We see the same pattern in that TOAST-MB performance declines as the mediating cell type proportion increases, while TCA and

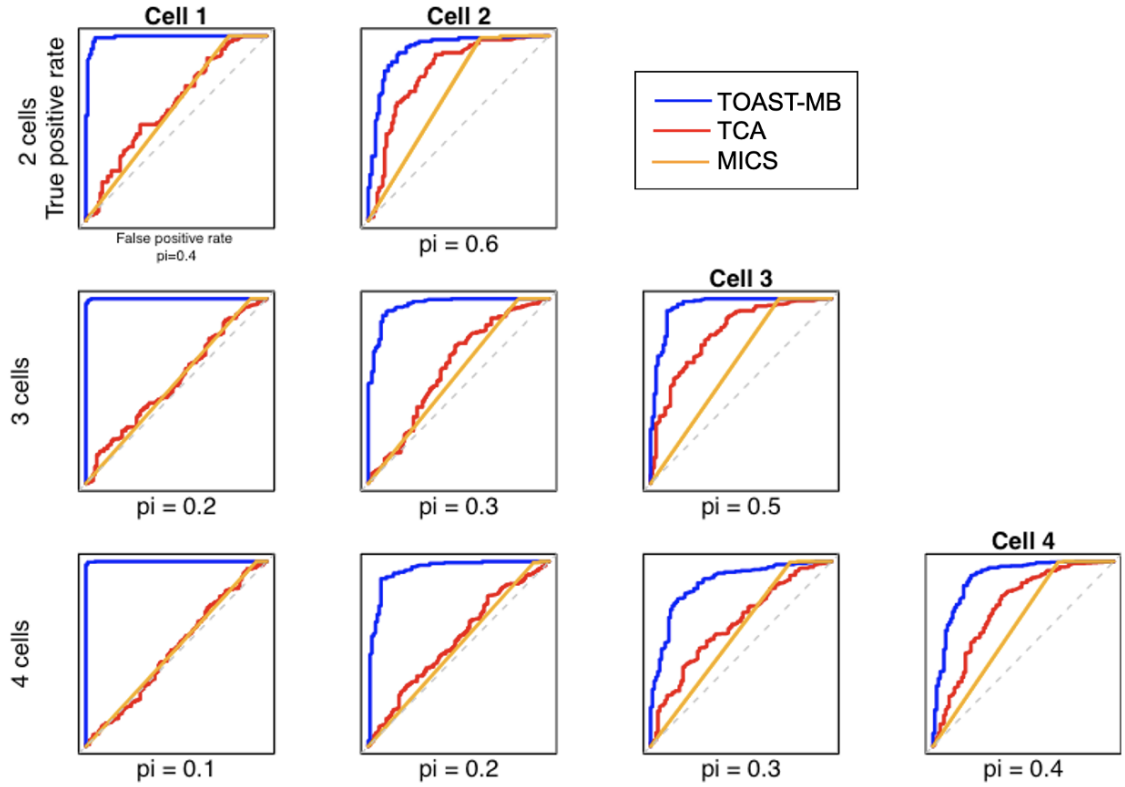

Figure 3.2: ROC curves for TOAST-MB, TCA, and MICS simulation study: Each plot represents a unique simulation setting. The row indicates how many cell types are present in that simulation, and the column indicates which cell type is the correct mediating cell type. Detections in the true mediating cell type are classified as true positives while detections in the non-mediating cell type(s) are classified as false positives. The proportion of the mediating cell type increases from left to right.

MICS improve, though again in each case, TOAST-MB outperforms TCA and MICS. Figure 3.4 shows the performance of TOAST-MB in the presence of all three types of confounders at once. TOAST-MB continues to outperform TCA and MICS in this case. [h!] Additionally, we assessed the effect of the cell type proportions on the TOAST-MB performance. The results of this analysis are shown in figure 3.5. We found that in all three methods, the performance varies based on the distribution of the cell type proportions. In the top row, the proportions gradually increase; as seen in previous results, TOAST-MB performance decreases as the proportion increases, but MICS and TCA improve as the proportion improves. The second row shows a

	Mediator	Cell 1	Cell 2	Cell 3	Cell 4
TOAST-MB	Cell 1	**85**	0	0	0
	Cell 2	2	**55**	0	0
	Cell 3	7	2	**35**	1
	Cell 4	5	1	1	**54**
TCA	Cell 1	**11**	5	2	7
	Cell 2	5	**6**	3	7
	Cell 3	3	8	**20**	8
	Cell 4	7	4	4	**34**
MICS	Cell 1	**1**	0	0	0
	Cell 2	0	**5**	0	0
	Cell 3	0	0	**17**	0
	Cell 4	0	0	0	**25**

Table 3.1: Simulation results: number of times each cell type was identified as a mediator in the 4 cell type case. The row indicates which cell type is the true mediator. Therefore, the diagonal entries (bolded) are true positives and off diagonal elements are false positives. Numbers are out of 100 simulation replicates.

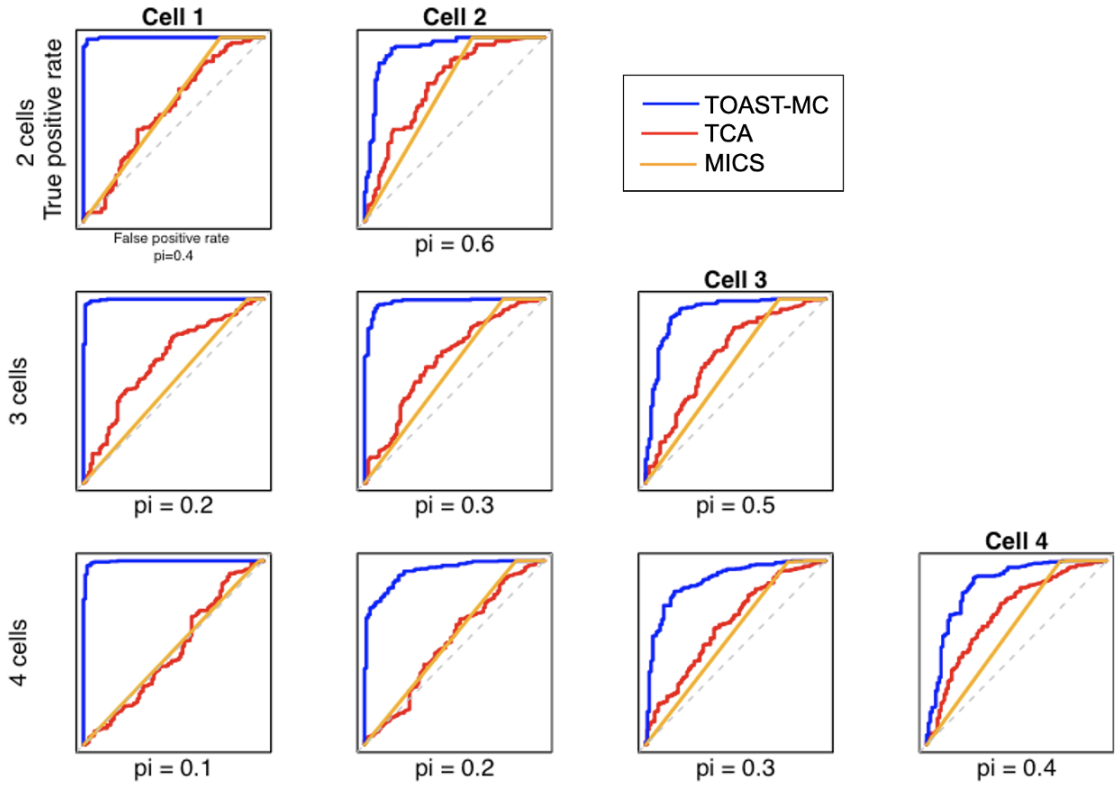

Figure 3.3: ROC curves to demonstrate the performance of TOAST-MB, TCA, and MICS in the presence of a single confounder

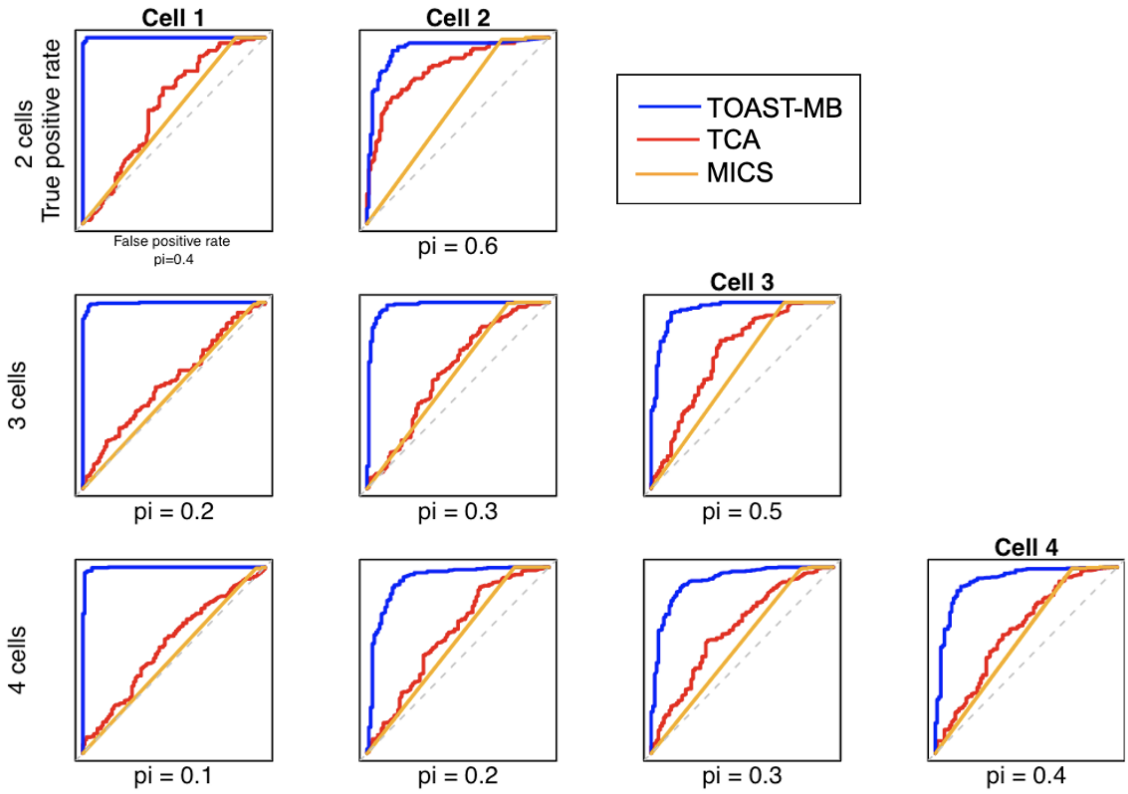

Figure 3.4: ROC curves to demonstrate the performance of TOAST-MB, TCA, and MICS in the presence of all three types of confounding: exposure-mediator, mediator-outcome, and exposure-outcome

more extreme case in which one cell type has a very large proportion compared to the other three. In this case, the performance of TOAST-MB is consistent in the three cell types with proportion 0.1, and performs similarly when the cell type proportion is 0.7. TCA and MICS, however, perform quite poorly when the cell type proportion is 0.1, and improve when the cell type proportion is 0.7. The bottom row shows the case where all cell types are distributed equally, and performance is consistent across cell types for all methods. In each case, regardless of the distribution of the cell type proportions, TOAST-MB outperforms TCA and MICS.

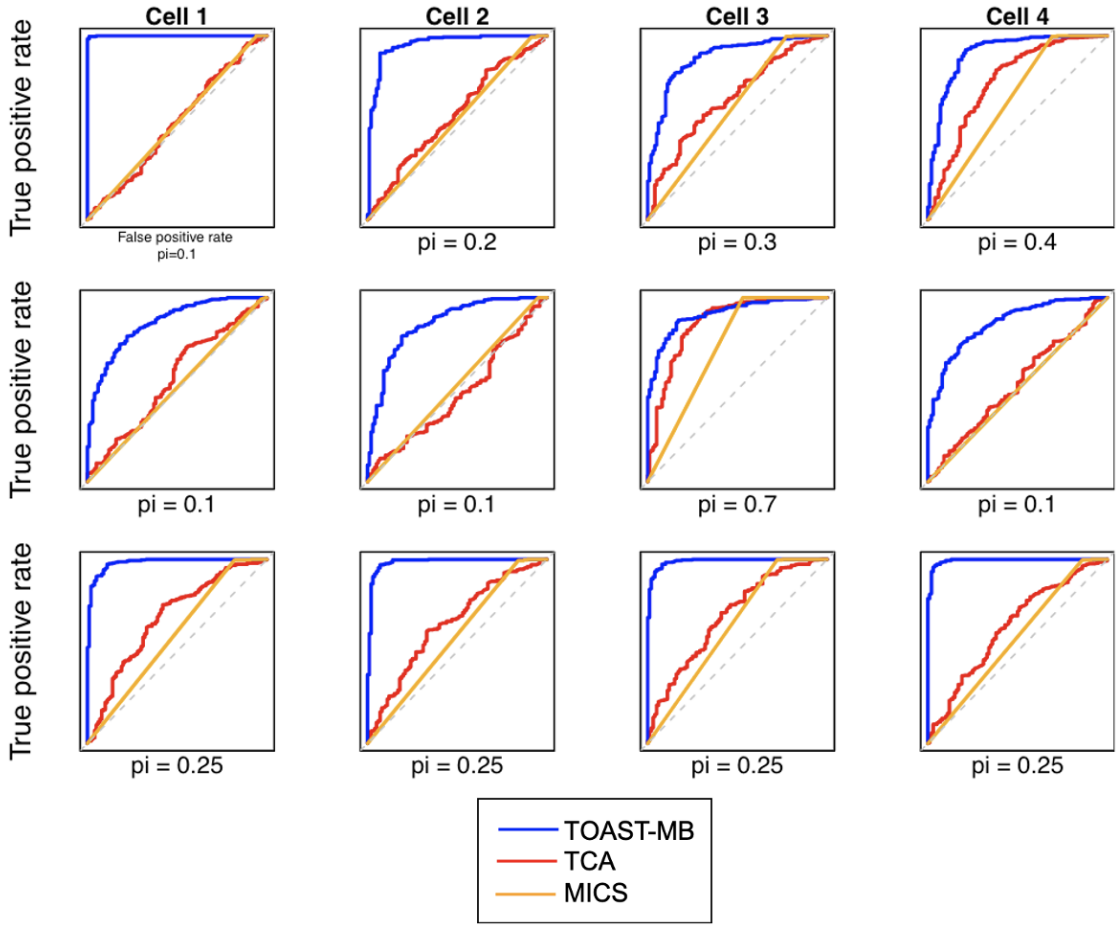

Figure 3.5: ROC curves to demonstrate the performance of TOAST-MB in the 4 cell type case when the cell type proportions are varied. Each row shows a different cell type proportion setting.

3.3.2 Real Data Analysis

To further evaluate the performance of TOAST-MB, we analyzed a dataset from the Grady Trauma Project, which assesses the influence of various factors, including DNAm, on response to traumatic events [24]. The cohort is predominantly comprised of African American individuals of low socioeconomic status in Atlanta, GA. Data were collected via interviews in waiting rooms of primary care or obstetrical-gynecological clinics. Clinical and life experience data, as well as blood samples, were collected. The exposure of interest was trauma as measured by the Trauma Events Inventory (TEI), a continuous measure of self-reported trauma experience [24]. The outcome was binary PTSD status as measured by the Clinician-Administered Post-traumatic Stress Disorder Scale (CAPS) for DSM-IV, or the modified PTSD Symptomatic Scale (PSS) when CAPS data were unavailable [32]. Sex and age were included as covariates. After removing subjects with missing data for trauma, PTSD, sex, or age, we analyzed a set of 660 individuals for whom EPIC array DNA methylation data were available. Cell type proportions were estimated using publicly available reference data and Robust Partial Correlation (RPF) method implemented in the R package EpiDish [69]. Six cell types were ascertained: CD8T, CD4T, Natural Killer (NK), Neutrophill, B cells, and monocyte cells.

The relationship between trauma exposure and PTSD is not yet well understood. Although experiencing a traumatic event is common, the prevalence of PTSD in the general population is estimated to be about 7% [50]. Although some predisposition to PTSD can be explained by genetics, epigenetics, and specifically DNA methylation, is thought to play a role in the causal pathway from exposure to a traumatic event to PTSD [50, 34, 60, 47]. Two key areas of interest in this relationship are the immune system and the central nervous system, both of which provide motivation for cell type-specific mediation analysis [50].

In the sample of 660 study participants, 166 (25%) were PTSD cases and 494 (75%) were controls. The TEI ranges from 0 to 16, with a higher value indicating more trauma exposure. The overall mean TEI score in the sample was 5.2 (SD=3.2); the mean TEI in cases and controls was 6.9 (3.2) and 4.6 (3.1) respectively. In a logistic regression model of PTSD regressed on trauma exposure, sex, and age, the odds of PTSD increased by 1.28 per unit increase in trauma exposure ($p < 0.001$). Mean cell type proportions were as follows: CD8T - 0.11, CD4T - 0.17, NK - 0.05, B cell - 0.07, Monocyte - 0.08, Neutrophill - 0.51.

We first analyzed the cell type-specific relationship between trauma exposure and DNA methylation with TOAST. Using a p-value threshold of 0.0001, 1269 CpG sites out of a total of 819708 were selected to analyze with TOAST-MB. Table 3.2 displays the number of CpG sites detected as mediators by TOAST-MB, TCA, and MICS. TCA fails to detect any CpG sites. For each cell type except CD8T, TOAST-MB detects more CpG sites than MICS. Figure 3.6 contains Manhattan plots for each cell type indicating the chromosome for each CpG site selected with a mediating cell type. The chromosomes with significant CpG sites differ for each cell type, which provides further justification for the cell type-specific analysis. Corresponding to table 3.2, CD4T, B, and neutrophill cells are detected as mediators in the most CpG sites. All six cell types are involved in the immune system, which is a primary system of interest in the relationship between trauma exposure and PTSD [50].

Gene and pathway analyses with Enrichr were conducted for each cell type to further evaluate the TOAST-MB results [13, 36, 74]. The *FKBP* gene, detected in monocyte cells, has been shown to be involved in the relationship between DNA methylation and stress-related disorders [50]. Methylation in *SATB1*, detected in neutrophill cells,

	CD8T	CD4T	NK	B cell	Mono	Neu
TOAST-MB	13	335	48	152	44	133
TCA	0	0	0	0	0	0
MICS	44	47	27	35	18	5

Table 3.2: Grady Trauma Project data: number of CpG sites detected as mediators for each cell type

has been shown to be associated with suicide [34]. Methylation in *SLC6A*, detected in CD8T cells, has been shown to be associated with PTSD [34]. *ANXA* has previously been shown to be differentially methylated in PTSD cases and controls, and this gene was detected in CD4T cells [34]. The *DOCK* gene was also detected in CD4T cells, and methylation in this gene has been shown in multiple studies (including an analysis of Grady Trauma Project data) to be associated with PTSD [50]. The *DLG4* gene is also implicated in psychiatric disorders, and this gene was detected in neutrophill cells [47].

In the Enrichr analysis, the genes differentially methylated in natural killer, B cells, monocytes, and neutrophills are involved in the central nervouse system. Those found in monocytes and neutrophills specifically are related to depression, acute stress disorder, PTSD, and increased blood pressure and heart rate. These connections provide evidence that the TOAST-MB results are biologically plausible and valuable for exploratory analysis of cell type-specific mediation.

3.4 Discussion

In this chapter, we presented a novel method called TOAST-MB to detect a cell type-specific mediation effect in bulk omics data. The method involves three steps. First, we select a subset of CpG sites by analyzing the cell type-specific exposure-mediator relationship with TOAST. Second, we employ Bayesian hierarchical modeling to fit

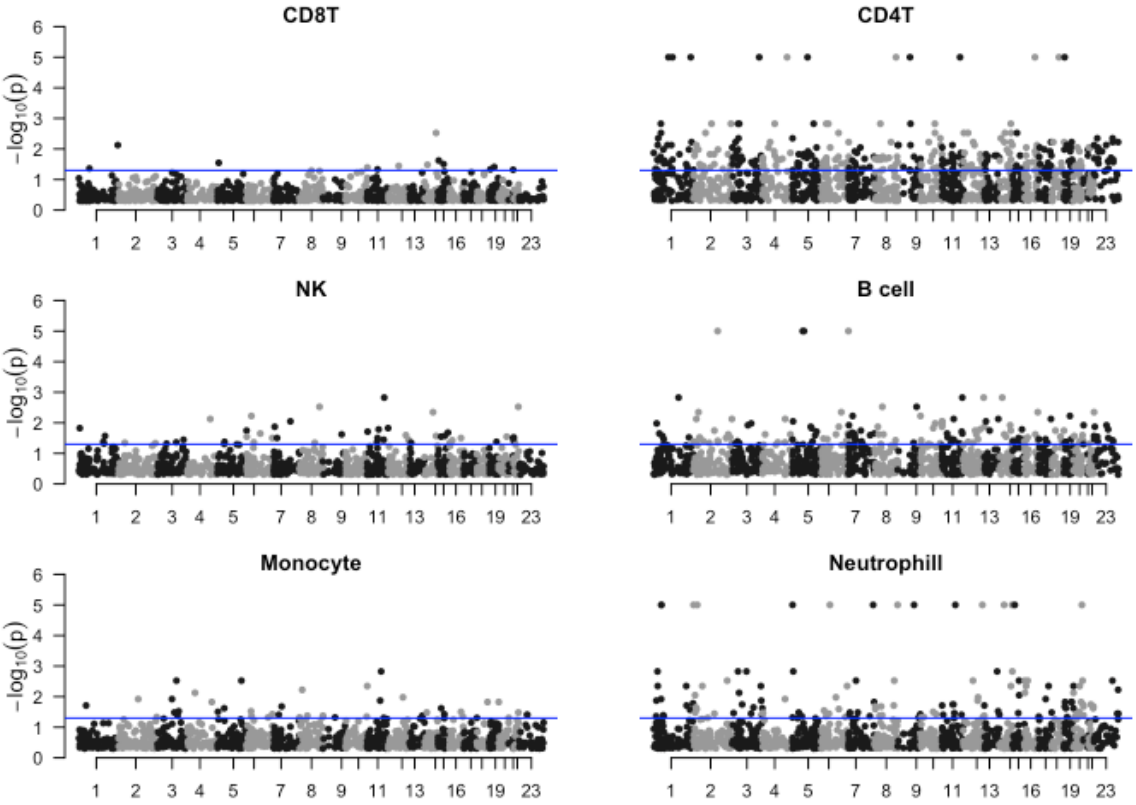

Figure 3.6: Manhattan plots for each cell type analyzed in the Grady Trauma Project dataset. The subset of 1269 CpG sites selected by TOAST are shown. The x axis indicates the chromosome for each CpG site. Note that here p refers to the posterior probability $\min(P(\exp(\beta_{1k}\theta_{2k}) < 1), P(\exp(\beta_{1k}\theta_{2k}) > 1)$ The blue line indicates the threshold for the posterior probability at which a cell type was selected as a mediator $(-log_{10}(0.05))$

the mediation models of interest using MCMC sampling; from this, we obtain a marginal posterior distribution of the natural indirect effect on the odds ratio scale. Third, we use this posterior distribution and calculate a posterior probability to classify each cell type as a mediator or non-mediator for each CpG site. For identification of the natural indirect effect, we rely on the assumptions outlined in the causal inference literature: namely, we assume that exposure-mediator, mediator-outcome, and exposure-outcome confounders are controlled for in the analysis. Moreover, to decompose the natural indirect effect from the multiple mediator model into individual effects for each cell type, we assume there is no interaction among the mediators (cell types). Finally, because we analyze one CpG site at a time, we assume independence among CpG sites.

Through simulation, we showed that TOAST-MB is more powerful than competing methods while controlling for false discoveries. This remains true when we add exposure-mediator, mediator-outcome, and exposure-outcome confounders. In analysis of the Grady Trauma Project dataset, TOAST-MB detected more CpG sites than competing methods. Using gene and pathway analyses, we found that the results are biologically plausible and meaningful in the context of trauma exposure and PTSD. However, we are hesitant to claim that the assumptions regarding confounding are met in this analysis by merely controlling for sex and age.

The proposed method faces some limitations. Most consequentially, the method relies on a number of assumptions, some of which may not be biologically plausible. Assumptions related to confounding required for a causal interpretation are subject to scrutiny on a case by case basis. As is true for any causal analysis, care must be taken to consider if these assumptions are met before employing a causal interpretation of results. That said, mediation is inherently a causal question, so these

assumptions must be accounted for to the extent possible in the data collection phase. Other assumptions inherent in TOAST-MB include that the cell types do not interact and there is independence among CpG sites. It is likely reasonable to assume that cell types do not operate completely independently of one another, and CpG sites, particularly close to one another, are known to be correlated [80]. Still, TOAST-MB provides a powerful method for exploratory analysis of cell type-specific mediation.

One advantage of TOAST-MB is that it can easily be extended to accommodate different types of outcomes. Although we focus on the binary outcome case, one could specify a linear model for a continuous outcome or a poisson model for a count outcome. Using Stan for MCMC sampling makes it straightforward to adjust the model structure depending on the research question. A more difficult hurdle lies in the correct model specification more generally. A key advantage of causal mediation over the traditional Baron and Kenny approach is that the effects of interest are defined outside of, rather than based, on the specified model [51]. This gives the researcher flexibility in the choice of the model. For example, it may be of interest to add an interaction term for the exposure and mediator, and indeed much of the causal inference literature encourages such a term and defines the natural indirect effect accordingly [72]. More work is needed to better understand cell type-specific biological mechanisms to determine if such a term is needed, and then statistical methods should accommodate the appropriate models. Currently, most statistical methods in this realm (including TOAST-MB) employ main effects models, which may not match biological reality.

Accommodating the high dimensionality of omics data also remains a challenge in cell type-specific mediation and the field of causal genomics more broadly. High dimensionality is a challenge from both the statistical and the computational per-

spective. To relax the assumption of independence among CpG sites, we would want to include all CpG sites in a single model rather than fitting the mediation models separately for each site. However, separating and defining causal effects for each cell type and CpG site would require careful consideration of how the sites and cell types interact. It is logical to think that the mediation effect of one cell type in one CpG site would not be independent of its effect in a second CpG site. In such a case, developing a robust statistical model that accurately reflects these interactions would be difficult. One approach might be to combine CpG sites into methylation regions and assess cell type-specific regional effects.

The application of causal inference to omics data offers many opportunities for future research. Mediation analysis in particular has become a strong area of interest, with applications in many omics realms including methylation, metabolomics, and microbiome data. Each of these present unique statistical challenges. Additionally, as more studies begin to collect multiple types of omics data, there is a need for mediation methods that can parse through the different types of omics mediation effects. Although this may be theoretically and computationally difficult, this area of research could likely lead to valuable discoveries to better understand biological mechanisms of disease.

Chapter 4

A comparison of high dimensional mediation methods

4.1 Introduction

Mediation analysis has grown in popularity over the last several years, and is now an active area of research in both methodological and applied settings. In fact, the number of Google scholar entries with "mediation analysis" in the title or abstract has grown exponentially over the last ten years [51]. The goal of mediation analysis is to assess the role that a factor of interest (called the mediator) plays in the intermediate path between an exposure and an outcome. Modern mediation analysis uses the causal inference counterfactual framework, which requires particular assumptions to identify a causal mediation effect [59, 44, 71].

As high-throughput omics data have become more accessible, researchers have naturally become interested in how different types of omics may mediate the relationship between an exposure and an outcome [16]. Assessing these relationships could revolutionize the way that many health conditions are understood and treated. But these

data are high-dimensional, and traditional approaches to mediation analysis are inadequate. Although omics data may have hundreds of thousands of components, sample sizes for these studies are typically in the hundreds. Fitting linear models in this case will then be inadequate because the number of predictors vastly outnumbers the number of observations. One could assess each mediator individually, but this fails to account for potential correlations among the mediators. Several methods have been developed in recent years to assess high-dimensional mediation models. Broadly, these methods fall into four categories: 1) methods that utilize dimension reduction or group mediators, 2) methods based on the composite null, 3) penalized regression methods, and 4) Bayesian methods.

In this work, we present a simulation study comparing three methods of high-dimensional mediation analysis. Although these methods can be employed in a range of high-dimensional omics contexts, we focus on the setting in which DNA methylation serves as the mediator of interest. DNA methylation has been studied as a mediator between multiple exposures and outcomes, including age and diabetes risk [26], genetics and rheumatoid arthritis [40], and prenatal adversity and metabolic disease [70]. We then apply the three methods to a dataset from the Grady Trauma Project to assess DNA methylation as a possible mediator between smoking and weight. The paper is organized as follows: first, we present a general overview of mediation models with multiple mediators, which serves as a foundation for omics mediation analysis. We then present a categorical overview of current methods for high-dimensional mediation. Next, we present the simulation study comparing three methods (HIMA, DACT, and BAMA) and apply the three methods to data from the Grady Trauma Project. We conclude with a summary and future directions in high-dimensional mediation.

4.1.1 General overview of mediation with multiple mediators

VanderWeele presents the following framework to analyze the joint effect of multiple mediators [72]. Suppose we have an exposure E, a continuous outcome Y, a set of covariates $\mathbf{X}$, and a set of K mediators $\mathbf{M} = (M^{(1)}, \ldots, M(K))$, and that the following regressions are specified:

$$E[M^{(i)}|e, x] = \beta_0 + \beta_1 e + \beta_2' X \tag{4.1}$$

$$E[Y|\mathbf{m}, e, \mathbf{x}] = \theta_0 + \theta_1 e + \theta_2^{(1)} m^{(1)} + \ldots + \theta_2^{(K)} m^{(K)} + \theta_4' X \tag{4.2}$$

Then the controlled direct effect, or the effect of E on Y not mediated through $\mathbf{M}$, is given by $\theta_1(e - e^*)$. The natural direct effect, or the effect of E on Y if M takes the value it would naturally take based on the value of E, is given by $\theta_1(e - e^*)$. The natural indirect effect, or the effect of E on Y that is mediated through M, is given by $[\beta_1^{(1)}\theta_2^{(1)} + \ldots + \beta_1^{(K)}\theta_2^{(K)}](e - e^*)$. e and e^* represent two levels of the exposure, so in the case of a binary exposure, this term is 1. Although the controlled direct effect and the natural direct effect coincide in this case, they diverge when other models are specified, e.g., when an interaction between the exposure and mediator is used. In the case of a binary outcome, a logistic or log-linear model can be specified for the outcome model and the effects are defined on the odds ratio scale [72].

This approach differs from assessing mediators one at a time because it includes all mediators in the outcome model, equation 4.2. If mediators do not affect one another, the multiple mediator approach is equivalent to assessing mediators one at a time using the single mediator model approach. However, in omics applications, the nature of the relationship among mediators is often not well understood, so making

the assumption that mediators do not affect one another may be too strong.

Testing the indirect effect

Multiple procedures have been proposed to assess statistical significance of the natural indirect effect, $\beta_1^k\theta_2^k$. The two most common methods are the Sobel test and the joint significance test [78]. The Sobel test constructs a test statistic directly for the product $\beta_1^k\theta_2^k$ by using the delta method to obtain an estimate of its standard error [62]. The joint significance test obtains p-values for β_1^k and θ_2^k from fitting the mediator and outcome models, respectively, and defines the mediation p-value as the maximum of the two p-values [15]. Resampling methods such as bootstrap and permutation tests have also been proposed to test the indirect effect [54]. These methods have been compared in simulations, though not always in the context of multiple mediators [27, 6, 45]. The Sobel test has been found to achieve lower statistical power compared to other methods. Resampling methods, namely bootstrap procedures, have been shown to be more powerful than the Sobel test and the joint significance test, and are generally recommended [6, 27].

4.1.2 General overview of proposed high-dimensional mediation methods

Dimension reduction or grouped mediators methods

Some proposed methods of high-dimensional mediation analysis have first reduced the number of mediators through either dimension reduction or grouping. Huang and Pan proposed transforming the mediators using Principal Component Analysis [30]. This method benefits from the fact that it addresses correlation among the mediators, which is often the case in DNAm. However, using principal components as mediators makes biological interpretation of results more difficult. Derkach et al. employ a

similar approach in which they assume that the high-dimensional group of mediators comes from an underlying group of latent factors [19]. Another method, proposed by Fang et al. called g-HMA, assesses a mediation effect based on genes rather than individual CpG sites [22]. A similar approach might be to assess methylation regions (DMRs) as mediators. These methods may be effective approaches to high-dimensional mediation but depend on the researcher's interest in the biological unit of the mediator.

Composite null methods

Other proposed methods are based on the composite nature of the null hypothesis in mediation analysis. When the interest lies in hypothesis testing of the natural indirect effect (as is often the case in mediation analysis), and when main effects linear models are used to model the mediator(s) and the outcome, the null hypothesis takes the form $H_0 : \beta_1^k \theta_2^k = 0$. Here β_1^k represents the coefficient of the exposure in the mediator model and θ_2^k represents the coefficient of the mediator in the outcome model. This null hypothesis is a composite null because any of the following conditions can be satisfied for the null to hold:

$$\begin{cases} \beta_1^k = 0, \theta_2^k \neq 0 \\ \beta_1^k \neq 0, \theta_2^k = 0 \\ \beta_1^k = 0, \theta_2^k = 0 \end{cases}$$

The methods in this category analyze one mediator at a time but pool information across mediators to account for the composite null hypothesis. For example, Huang et al. developed the JT-comp method, which uses single mediator models to obtain p-values z_a and z_b from the mediator and outcome models, respectively, then assesses the distribution of these p-values under the three null scenarios. JT-comp then estimates the probability of each null scenario and uses those estimates to generate an adjusted

p-value for each mediator [29]. JT-comp requires several assumptions that may not hold in real data, particularly when the sample size is greater than ≈500, which may lead to inflated type 1 error [78]. Similarly, DACT analyzes mediators one at a time, but directly obtains estimates of the proportions of each component of the composite null hypothesis to generate a calibrated p-value [41]. A third method in this category, JS-mixture, directly constructs the null distribution but does so by estimating the proportions of the three null components, similarly to JT-comp and DACT. These methods may benefit from higher power compared to methods that do not directly address the composite null hypothesis. However, assessing mediators one at a time does not account for potential correlation among the mediators.

Penalized regression methods

The third type of high-dimensional mediation analysis method employs penalized regression to identify true mediators. Penalized regression methods, such as LASSO, add a penalty term to the objective function to be minimized; by doing so, penalized regression methods can handle regression scenarios in which the number of explanatory variables is greater than the sample size ($p > n$) [82]. LASSO shrinks some coefficients to 0 and therefore can serve as a variable selection procedure [82]. This technique can be useful in high-dimensional mediation problems to analyze all mediators jointly. Because the number of mediators is often much larger than the sample size in omics mediation analysise ($p >> n$), a screening procedure is often employed in these methods as a first step prior to penalized regression [79, 81].

Zhang et al. proposed the method HIMA, which analyzes a set of high-dimensional mediators in three steps [79]. In the first step, HIMA employs Sure Independence Screening to reduce the number of mediators from ultra high-dimensional to high-dimensional. Then, HIMA uses penalized regression to perform variable selection.

In the third step, HIMA evaluates the remaining candidate mediators in a multiple mediator model and performs the joint significance test with a Bonferroni multiple comparisons adjustment.

Another method in this category is Pathway Lasso [81]. Pathway Lasso adds two penalty terms, one aimed at shrinking estimates for the product $\beta_1\theta_2$, and one aimed at shrinking β_1 and θ_2 individually. This method does not use an initial screening procedure. In the original paper presenting Pathway Lasso, it is applied to a neuroimaging dataset with 76 potential mediators. Therefore, it is unclear whether or not this method can handle high-dimensional mediation. For example, in the context of DNA methylation, the number of potential mediators can be as high as $400,000$ or more.

Bayesian methods

Finally, Song et al. have proposed a series of Bayesian methods for high-dimensional mediation analysis [63]. The set of methods, called BAMA, use the same general framework but differ in the structure of the prior distributions. BAMA applies a sparsity assumption that only a small proportion of the potential mediators are true mediators. The first iteration of BAMA, which will be used in the simulation study, models all mediators jointly by applying a Bayesian sparse linear mixed model prior. This assumes that all mediator effects follow a two-component normal mixture model in which one component has a large variance and one has a small variance. BAMA uses an MCMC sampling procedure to obtain posterior samples. BAMA produces a posterior inclusion probability (PIP), defined as the probability that both β_1 and θ_2 belong to to the normal component with a large variance, and are therefore active mediators. The user can specify a PIP threshold at which to select true mediators. Song et al. have proposed some extensions to BAMA to jointly model the β_1 and θ_2

effects and explicitly model the correlation structure among mediators [64, 65].

4.2 Methods

4.2.1 Selected high-dimensional mediation methods

We conducted a comprehensive simulation study to compare the performance of HIMA, DACT, and BAMA. These three methods were chosen based on the following considerations: 1) to achieve representation from each category of methods, 2) availability of an R package to use for analysis, and 3) popularity of use (based on Google Scholar citations). The first category of methods, dimension reduction or group-based methods, was omitted because the interpretation of the results differs from methods that test individual CpG sites.

Figure 4.1 illustrates the three main steps of HIMA. The HIMA method and corresponding R package is self-contained and uses the Bonferroni correction for multiple comparisons [79]. Figure 4.2 illustrates the DACT method. The DACT function requires two vectors of p-values as inputs, one from the mediator models and one from the outcome models. The user is responsible for correcting for multiple testing. Equations 4.1 and 4.2 were fit for each CpG site individually and the p-values for β_1 and θ_2 were used for DACT inputs. Following the procedure used by Liu et al. in the real data analysis, q-values were computed to correct for multiple testing [41]. FDR thresholds of 0.05 and 0.1 were compared, with the lower value being the more conservative case. BAMA produces a posterior inclusion probability (PIP) for each CpG site, and the user can decide a threshold to use to classify a site as a mediator [63]. PIP thresholds of 0.1, 0.3, and 0.5 were compared. Note that the higher value of PIP represents the more conservative threshold. Table 4.1 summarizes the characteristics of the three methods.

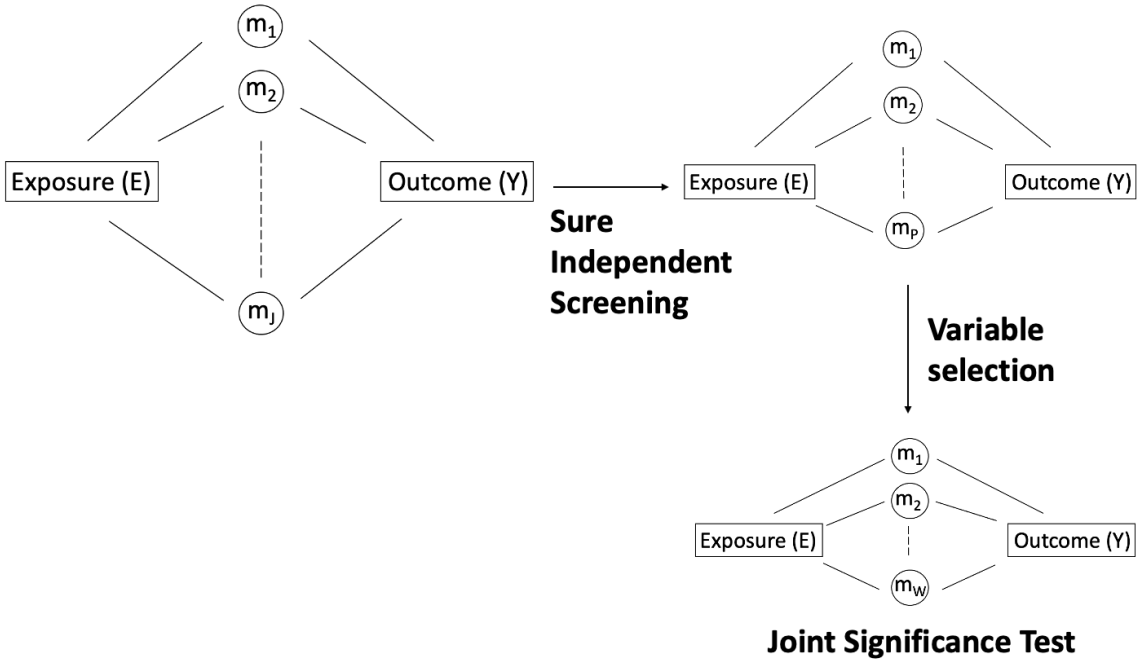

Figure 4.1: HIMA first uses sure independence screening to reduce the dimension of the mediators from ultra high to high dimensional. Then it utilizes a penalized regression method to further reduce the dimension of the mediators. Once the set of mediators is sufficiently reduced ($p < n$), HIMA fits the multiple mediator models and uses the joint significance test with the Bonferroni multiple testing correction.

Method	Category	Input	Summary
HIMA	Penalized regression	Full methylation data	1. Sure independence screening 2. penalized regression 3. joint significance test
DACT	Composite null	vector of p-values from mediator models and vector of p-values from outcome models	Estimates proportions of each null case and calculates new p-value accordingly
BAMA	Bayesian	Subset of methylation data based on marginal analysis	Assumes each mediator comes from a normal mixture and calculates a probability that each site is a mediator

Table 4.1: Summary of the three selected methods

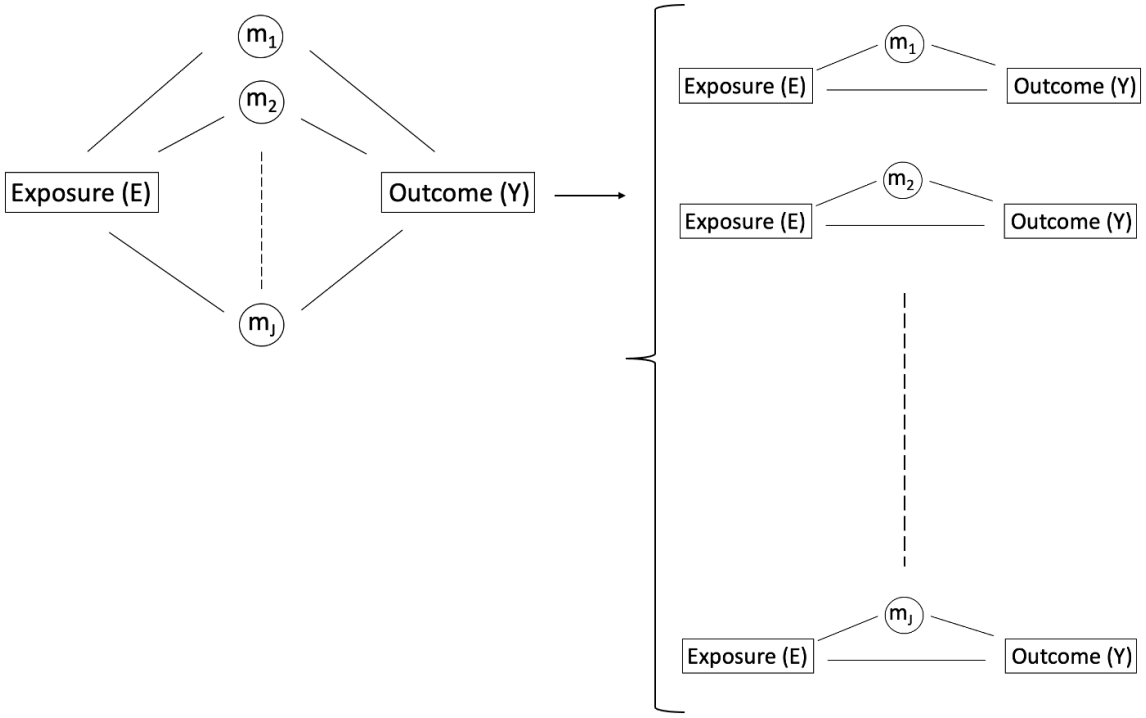

Figure 4.2: DACT requires the user to fit individual models for each CpG site and input two vectors of p-values, one from the mediator model and one from the outcome model. DACT then pools information from these p-values to estimate the proportion of each null hypothesis case and combine the given p-values into a single value using the estimated proportions.

4.2.2 Simulation

To capture the correlation and variance structure of methylation data, an open source dataset from a multi-ethnic cohort was used in the simulation [23]. The methylation matrix of beta values contained 1,219 total subjects and 386,362 CpG sites. Random samples of subjects and CpG sites were taken to compare different settings. Sample sizes were 200, 500, and 1000, and the number of CpG sites were $1,000$ or $100,000$. The proportion of CpG sites designated as mediating sites was compared between 0.1 and 0.001. The sampled subjects and CpG sites were fixed for each setting, but the sampled mediating sites varied for each simulation replicate. In the case of 1,000 CpG sites, to ensure a mediator-exposure effect would be evident, CpG sites were sampled

from those with mean methylation value between 0.4 and 0.6. A random covariate was generated from Unif$(0.3, 0.7)$. A binary exposure was generated with 50% of subjects classified as exposed and 50% classified as unexposed. A direct effect was drawn from Unif$(0.01, 0.05)$. The exposure-mediator effect was induced using equation 4.1 with $\beta_1 = 0.4$ and random noise added from N$(0, 0.01)$. The methylation values were then bounded at 0.01 and 0.99. A continuous outcome was generated by adding θ_2 in the mediating CpG sites and then summing over CpG sites and adding the direct effect and random noise N$(0, 0.01)$. Two values of θ_2 were compared: 0.4 and 2.

Each simulation setting specified a number of CpG sites and a proportion that would serve as mediating CpG sites. BAMA is computationally intensive and it is not feasible to run on the entire set of CpG sites. Therefore, BAMA was included in the simulations for 1,000 sites but not for 100,000 sites. Proportions of CpG sites were compared because 1,000 sites could represent a selected subset of CpG sites following marginal mediation analysis, and the true biological proportion of mediating sites is unknown. False discovery rate (FDR) and power were compared for the three methods. Because each simulation contained mediating sites and non-mediating sites, FDR was calculated as the number of significant discoveries in non-mediating sites divided by the total number of significant discoveries. Power was calculated as the mean proportion of mediating sites that were detected.

4.3 Results

4.3.1 Simulation

False Discovery Rate

Figure 4.3 shows the distribution of the FDR for each method over 100 simulation replicates by sample size and proportion of mediators with 1,000 CpG sites. 1,000 sites mimics the case where a subset of CpG sites is selected to be used in the high-dimensional mediation method. This is particularly relevant for BAMA, which requires significantly more computation time than HIMA or DACT. In the case where the proportion of mediating sites is 0.001, meaning there is only 1 mediating site, FDR is largely controlled by both methods for N=200 and N=500. For N=1000, we see more variability in the FDR for BAMA and DACT, though the median FDR is still below 0.1 for each method except BAMA-0.1. The bottom row shows the FDR when the proportion of mediating sites is 0.1, meaning there are 100 mediating sites in this case. We see large variability in the FDR of BAMA-0.1, but this is not surprising given that 0.1 is the most liberal PIP threshold of the three BAMA cases. In each BAMA case, the median FDR is below 0.1. For DACT, we see similar variability regardless of sample size, though the median indicates FDR inflation for sample sizes of 200 and 500. HIMA appears to control FDR in each case.

Figure 4.4 shows the distribution of the FDR for DACT and HIMA with 100,000 CpG sites. BAMA is omitted because of the increased computation time. This case is perhaps more relevant for DACT and HIMA than the 1,000 CpG site case because it is feasible to run these methods on the full set of CpG sites in a given dataset. When the proportion of mediating CpG sites is 0.001 (100 sites), FDR is controlled in each sample size and method, as indicated by the median. However, these is more variation in FDR in the N=500 case, and some outliers in the other sample size settings. When the proportion of mediators is 0.1 (10,000 sites), there is more variability in FDR for

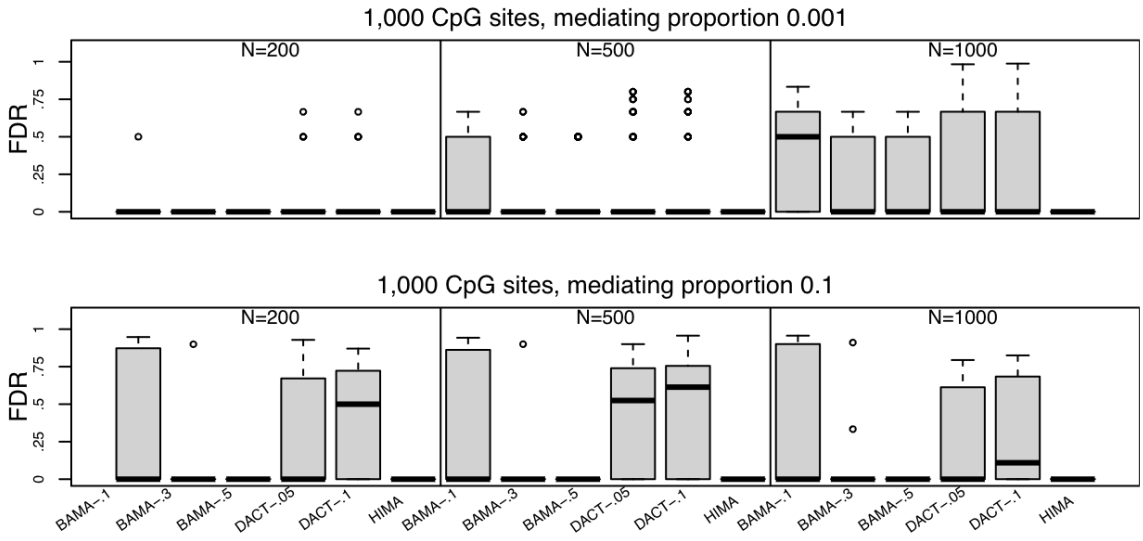

Figure 4.3: FDR for 1,000 CpG sites. The top row shows the FDR with a mediating site proportion of 0.001, meaning in this case there is 1 mediating CpG site. The bottom row shows the FDR with a mediating site proportion of 0.1, meaning there are 100 mediating CpG sites. FDR is calculated as the number of discoveries in non-mediating sites divided by the total number of discoveries. The boxplots show the distribution of FDR for 100 simulation replicates.

DACT. The median FDR for DACT increases as the sample size increases. The 0.1 FDR threshold is more liberal, so it is not surprising that FDR is more inflated in this case. HIMA appears to control in the larger CpG site setting as well.

Power

Figure 4.5 shows the distribution of power, as calculated by the proportion of mediating sites detected, in the 1,000 CpG site case. Because in the top row there is only 1 mediating site, the site is either detected or not, which explains the shape of the BAMA and DACT boxplots. For N=200 and N=500, BAMA and DACT have median power of 1, and means above 0.65 for every case except BAMA-0.5, which has mean 0.54 for both sample sizes. The mean power decreases for BAMA and DACT in the N=1000 case (as does the median except for BAMA-0.1 and BAMA-0.3). This could be a result of the variability inherent in selecting a single CpG site as the mediating site. Because a random sample of both subjects and CpG sites were taken from the

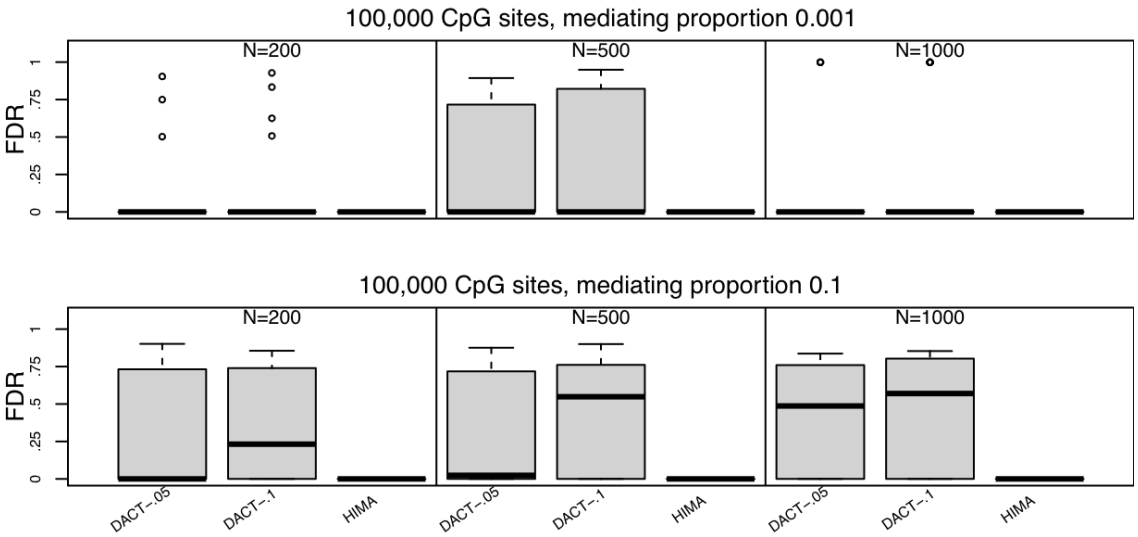

Figure 4.4: FDR for 100,000 CpG sites. The top row shows the FDR with a mediating site proportion of 0.001, meaning in this case there are 100 mediating CpG site. The bottom row shows the FDR with a mediating site proportion of 0.1, meaning there are 10,000 mediating CpG sites. BAMA is omitted because of the higher computation time required. FDR is calculated as the number of discoveries in non-mediating sites divided by the total number of discoveries. The boxplots show the distribution of FDR for 100 simulation replicates.

full dataset, and only one CpG site is a mediator in this case, the methods could be sensitive to characteristics (such as mean or variance) of the CpG site selected. The bottom row of 4.5 shows the distribution of power when the proportion of mediating sites is 0.1 (100 sites). BAMA demonstrates low power, and the power decreases as the PIP threshold increases, as expected. DACT has the highest median power for both 0.05 and 0.1 thresholds, though FDR was shown to be inflated in some of these cases (primarily for DACT=0.1). HIMA fails to detect any CpG sites in any case.

Figure 4.6 shows the distribution of power for the 100,000 CpG site case. When the proportion is 0.001 (100 mediating sites), both methods have very low power, though DACT appears to occasionally be better than HIMA. This case is likely the most biologically realistic setting for DACT and HIMA. Although FDR was controlled for both methods in this case, the power is quite low. The second row contains power dis-

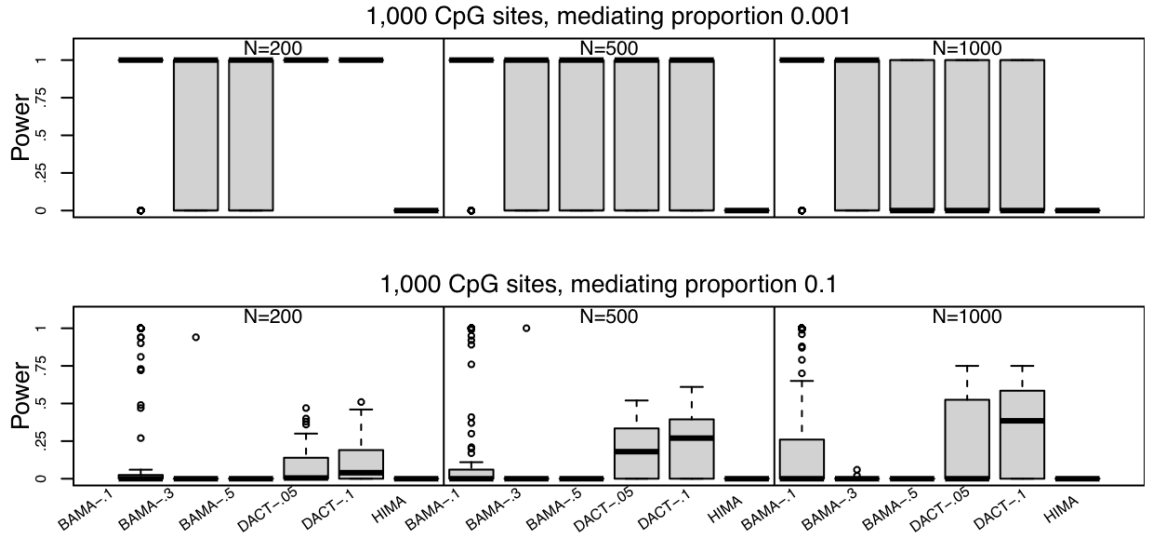

Figure 4.5: Power for 1,000 CpG sites. The top row shows the power with a mediating site proportion of 0.001, meaning in this case there is 1 mediating CpG site. The bottom row shows the power with a mediating site proportion of 0.1, meaning there are 100 mediating CpG sites. Power was calculated as the fraction of mediating sites detected. The boxplots show the distribution of power for 100 simulation replicates.

tributions for proportion 0.1 (10,000 mediating sites). Although DACT shows higher power than HIMA, the FDR was inflated in this case, particularly for DACT-0.1. For DACT-.05, the mean power was 0.07, 0.08, and 0.12 for sample sizes 200, 500, and 1000, respectively.

Computation

Table 4.2 shows the computation time for each method and each simulation setting. For the 1,000 CpG site case, the run time for HIMA ranges from 1.7-3.2 seconds depending on the sample size. DACT computation time is similar but slightly lower in this case, ranging from 1.8-2.2 seconds. BAMA has the highest computation time, taking 2.8 mins for $N = 200$, 6.2 mins for $N = 500$, and 12.0 mins for $N = 1,000$. Because of the increased computation time for BAMA, it is recommended by Song et al. to use with a subset of CpG sites after marginal mediation analysis. In the 100,000 CpG site case, HIMA requires between 2.5 and 3.5 mins to run depending on

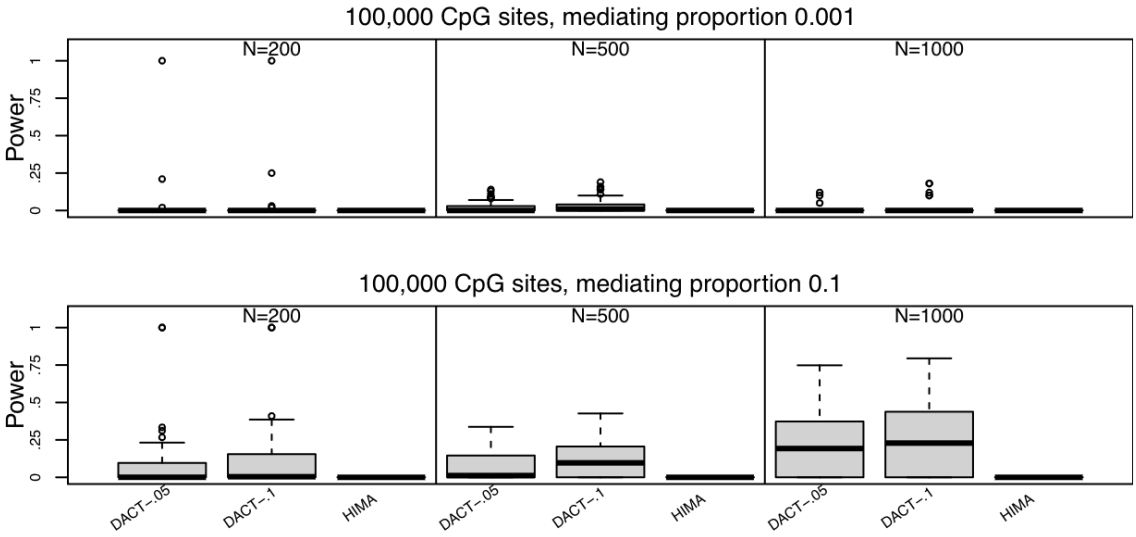

Figure 4.6: Power for 100,000 CpG sites. The top row shows the power with a mediating site proportion of 0.001, meaning in this case there are 100 mediating CpG site. The bottom row shows the power with a mediating site proportion of 0.1, meaning there are 10,000 mediating CpG sites. BAMA is omitted because of the higher computation time required. Power was calculated as the fraction of mediating sites detected. The boxplots show the distribution of power for 100 simulation replicates.

sample size, and DACT requires between 2.9 and 3.6 minutes.

Method	Number of CpG sites	Sample size	Computation time
HIMA	1,000	200	1.7 sec
		500	2.1 sec
		1,000	3.2 sec
	100,000	200	2.5 min
		500	2.8 min
		1,000	3.5 min
DACT	1,000	200	1.8 sec
		500	1.9 sec
		1,000	2.2 sec
	100,000	200	2.9 min
		500	3.1 min
		1,000	3.6 min
BAMA	1,000	200	2.8 min
		500	6.2 min
		1,000	12.0 min

Table 4.2: Simulation results: computation time for each method based on the number of CpG sites and the sample size

4.3.2 Real Data Analysis

To further evaluate the performance of TOAST-MC, we analyzed a dataset from the Grady Trauma Project, which assesses the influence of various factors, including DNAm, on response to traumatic events [24]. The cohort is predominantly comprised of African American individuals of low socioeconomic status in Atlanta, GA. Data were collected via interviews in waiting rooms of primary care or obstetrical-gynecological clinics. Clinical and life experience data, as well as blood samples, were collected. We analyzed 679 individuals for whom EPIC array DNA methylation data were available. Binary smoking status was the exposure of interest and weight (kg) was the continuous outcome. Both smoking status and weight have been shown to be associated with epigenetics [20], [37], [73], [17], [33], [46], [77]. Sex and age were controlled for as confounding variables. Observations with missing values for smoking status, weight, sex, or age were excluded. Cell type proportions were estimated using publicly available reference data and Robust Partial Correlation (RPF) method implemented in the R package EpiDish [69], and cell type proportions were included as covariates.

Of the 679 individuals, 273 (40%) were smokers and 406 (60%) were nonsmokers. The mean weight in nonsmokers was 96.7 kg (SD=26.1) and 90.0 (25.1) in smokers (p¡0.001, two sample t-test). 79.8% of nonsmokers were female with a mean age of 40.8 (SD=12.7), while 62.3% of smokers were female with a mean age of 44.5 (SD=11.0). HIMA and DACT were used to analyze the full set of 761,950 CpG sites. HIMA did not detect any significant CpG sites. Based on the simulation results for FDR, the 0.05 FDR threshold was used for DACT. Four CpG sites were detected, as shown in table 4.3. Most notably, two sites on the AHRR gene were detected, and this is a gene known to be associated with smoking [38]. Site cg05575921 was also detected in the DACT analysis of the mediating effect of DNA methylation between smoking

and lung function in the Normative Aging Study [41]. The computation time for DACT was 29 minutes. A subset of 834 CpG sites was selected based on a marginal p-value of < 0.001 in the mediator model to be analyzed with BAMA. BAMA did not detect any CpG sites at the 0.1 PIP threshold (and, therefore, neither the 0.3 nor 0.5 thresholds).

CpG	CHR	Gene
cg05575921	5	AHRR
cg17287155	5	AHRR
cg17739917	17	RARA
cg07390844	18	TSHZ1

Table 4.3: Grady Trauma Project data: CpG sites detected by DACT

4.4 Discussion

In this chapter, we presented an overview of high-dimensional mediation methods and compared the performance of three methods. HIMA performs two screening steps to reduce the number of mediators, and then uses the joint significance test with the Bonferroni correction to classify significant mediators. DACT analyzes mediators one at a time but pools information across the p-values to estimate proportions of each null hypothesis case, and then recalculates an overall p-value using those proportions. BAMA assumes sparsity and utilizes continuous shrinkage priors to select true mediators in a Bayesian framework.

In the simulation study, all methods controlled FDR fairly well. The most biologically realistic settings are 1,000 sites with mediator proportion 0.1, and 100,000 sites with mediator proportion 0.001. Although DACT shows FDR inflation in some cases, it does not appear to be inflated in the more biologically realistic cases. BAMA showed

inflated FDR when using the 0.1 PIP threshold. This is not surprising as that is the most liberal of the three thresholds used. All three methods show very low power in general, with the exception being the case with 1,000 CpG sites with proportion 0.001 in smaller samples. HIMA fails to detect a CpG site in any case. In the analysis of the Grady Trauma Project data, only DACT detects significant CpG sites. Because it shows the highest power in the simulation study and controls FDR in most cases, DACT appears to be the best choice for high-dimensional mediation at this time. This is further evidenced by its ability to detect sites in the real data. Additionally, the sites detected in the Grady Trauma Project data are on the AHRR gene, which is well understood to be associated with smoking.

The comparison of these methods demonstrate the challenge of high-dimensional mediation, specifically with DNA methylation data, and motivate the need for more statistical methods development in this area. A primary challenge of DNA methylation data is that beta values lie between 0 and 1, so the variance is limited. Upon close inspection of the original simulation study for HIMA, this setting was not accounted for. Therefore, HIMA may still be a useful method in other high-dimensional omics cases that do not face this limitation. Moreover, the Bonferroni correction is known to be conservative; perhaps if another approach to multiple testing were used, HIMA may see improved overall performance. DACT has not been previously tested in settings where some CpG sites are considered mediators while others are not, and FDR and power are calculated accordingly. In the original simulation study for DACT, type 1 error and power were assessed separately, so the set of CpG sites were either all non-mediating or all mediating. This may impact the difference in FDR and power reported in the original study as opposed to the simulation presented here. Similarly, the results seen here (namely, low power) are consistent with those reported in the original BAMA simulation study.

While more research is needed to develop more powerful statistical methods for high-dimensional mediation, these analyses can also stand to benefit from larger samples sizes in the data collection phase. Particularly for DNA methylation studies, where the effect size is limited by the range of the beta values, an increase in sample size can help to increase the power to detect significant mediators.

While we omitted grouping or dimension reduction-based methods from the simulation study, these methods represent an important collection of approaches to high-dimensional that may be more powerful than the CpG site approaches. Using principal components as mediators suffers from difficult interpretation, but may help in addressing the research question of the presence or absence of an overall mediating effect. Gene-based methods, or those that use DNA methylation regions, combine the signals from multiple CpG sites, which may be a more powerful initial approach than looking at CpG sites individually. Once a particular gene or region is identified, the researcher then could look at individual CpG sites without facing the challenge of such a high dimension.

The application of causal inference principles to high-dimensional mediation also remains an important area of future work. Although more attention has been drawn to the assumptions regarding control of confounding factors, more work is needed to better understand the causal structure of omics mediation problems. For example, although CpG sites near each other are known to sometimes be correlated, the causal structure of this relationship is unknown. More clarity on these relationships can help to more accurately model and assess the role of omics data as a mediator between an exposure and outcome.

Although challenges remain, omics mediation represents a promising area of future biological discovery. A better understanding of the mediating role of epigenetics, the microbiome, proteomics, metabolomics, or a combination thereof, between an exposure and outcome may transform the approach to the treatment of many diseases.

Chapter 5

Discussion

Recent technological innovations in sequencing techniques have led to a tremendous increase in the availability of different kinds of omics data. These data can help to better understand the biological mechanisms by which complex diseases originate. In this dissertation, we have presented three chapters exploring statistical methods to better answer research questions related to omics mediation. In the second chapter, we presented TOAST-MC, a procedure based on EM algorithm to identify mediating cell types in the relationship between an arbitrary exposure and a continuous outcome. Then, to have a unified method, we presented TOAST-MB in chapter three. TOAST-MB utilizes a Bayesian framework to identify mediating cell types between an arbitrary exposure and a binary outcome. This framework can easily be extended to accommodate other types of outcomes, including count outcomes. In the fourth chapter, we compared methods of high-dimensional mediation. These methods do not explore cell type-specific relationships, but attempt to jointly analyze CpG sites to better account for the correlation among the mediators.

The key innovation of this dissertation is providing an approach to detect cell type-specific mediation effects between an exposure and an outcome. Although we focus

on the application to DNA methylation, the method can be extended to other types of omics data. Interest in omics data mediation analysis has increased in recent years, but these studies have faced three main challenges: 1) accounting for cell type heterogeneity, 2) assessing causality, and 3) dealing with high-dimensionality. In the first two projects, we sought to address the first two challenges. We utilized the causal inference framework of mediation to develop methods to assess cell type-specific mediators. More generally, these methods can be viewed as methods to deal with mediation problems where the mediators of interest are observed at the bulk level rather than the individual level. In this way, we contribute not only to innovations in omics mediation, but in statistical mediation analysis more generally.

This work faces a few limitations which provide opportunities for interesting future research. First, cell type-specific mediation methods, including TOAST-MC, TOAST-MB, and MICS, analyze CpG sites one at a time. This assumes that CpG sites operate independently, which is known to be untrue. Specifically, CpG sites close to each other in genomic location are understood to be correlated [23]. To account for this correlation, CpG sites could either be analyzed jointly or grouped (e.g., DNA methylation regions) before analysis. The primary challenges with analyzing CpG sites jointly in this context are 1) computational time, and 2) clarifying the causal structure of multiple cell types and multiple CpG sites in the same model. By analyzing multiple sites in the same model, one would need to be careful about defining and modeling the relationship between the same cell type on different sites. Even without analyzing CpG sites jointly, these methods face computational challenges. DNA methylation data often contain hundreds of thousands of CpG sites. More work is needed to efficiently analyze this many sites and the cell type-specific effects within them.

The fourth chapter, which compares three methods of high-dimensional mediation analysis, displays the challenge of developing a powerful method that controls FDR in the high-dimensional setting. One might naturally wonder why these methods detect so few CpG sites as significant, while the cell type-specific methods presented detect tens or hundreds in each cell type. One possible explanation is that separating the effects into specific cell types may elucidate signals that, when combined at the bulk level, cancel out. If this is the case, cell type-specific analysis provides a great opportunity for future discovery. However, more work is needed to compare these methods and the statistical mechanisms by which the cell type-specific methods detect so many more signals than the general high-dimensional mediation methods.

The structure of DNA methylation data presents unique challenges that may not be present in other types of omics data. Specifically, beta values for DNA methylation are bounded between 0 and 1. This limits the variability of the data and possible effect sizes. Statistically, the small variance makes it more difficult to develop powerful methods to detect small differences between subjects. Because of this, it is important to increase the sample size as much as possible in these studies during the study design phase, and limit missing data as much as possible. Additionally, more biological insight into what constitutes a meanginful difference in methylation level can help in the development of statistical methods and power analyses in the future. Focusing on these aspects during study design can help to add statistical power and lead to more biological discoveries in DNA methylation mediation.

This work is motivated by the fact that EWAS data have historically been collected at the bulk level, where different cell types are mixed together. More recent technology has allowed for the advent of single cell omics data. If single cell omics data are available, mediation models can be fit more directly, though the challenges of

high-dimensionality and assessing causality would remain. However, single cell omics is currently difficult to obtain for a population level study because of its high cost. Still, as single cell data become easier to obtain, statistical methods will be needed to analyze them in a mediation context. Even more beneficial would be methods that can jointly analyze bulk and single cell data to reap the benefits of both accessibility and biological accuracy.

A better understanding and application of causal inference principles is vital to the success of omics mediation analysis in the future. At present, there remains much to learn about the relationships between omics processes, exposures, and health outcomes. Causal interpretations rely on several assumptions related to the control of confounding factors. As knowledge grows about these processes, researchers will be better able to control for as many confounding factors as possible, thus leading to a better understanding of the causal processes by which an exposure leads to a disease. The need to control for confounders in the exposure-mediator, mediator-outcome, and exposure-outcome relationships presents a further argument for increasing sample sizes in omics mediation studies. One must have a sufficient size to appropriately model the necessary confounding factors. Similarly, more biological understanding is needed to design statistical models that accurately represent biological processes. For example, in the context of DNA methylation mediation, a better understanding is needed regarding the temporal relationship and biological interactions between certain exposures and methylation. Mediation methods in the causal inference literature can accommodate interaction and non-linear terms, but a better biological understanding is needed to justify more complex models.

Finally, more work is needed to enhance the reproducibility of EWAS and omics mediation studies broadly. Developing powerful statistical methods can help with

this, as more powerful methods that detect specific mechanisms may help researchers identify similar results across studies. Significant technical variability remains a challenge in omics studies more generally; however, as sequencing technology continues to improve, so will the reproducibility of these analyses.

Facing these biological and statistical challenges can lead to important discoveries and a better understanding of the mechanisms by which health outcomes occur. In this work, we have taken an important step forward by presenting a method to detect cell type-specific mediation effects. A better understanding of these mechanisms could lead to great innovations in how many health outcomes are treated and, ultimately, prevented.

Bibliography

[1] Epigenome-wide association study, Mar 2022. URL `https://en.wikipedia.org/wiki/Epigenome-wide_association_study`.

[2] Mediation (statistics), Apr 2022. URL `https://en.wikipedia.org/wiki/Mediation_(statistics)`.

[3] Signe Altmäe, Francisco J Esteban, Anneli Stavreus-Evers, Carlos Simón, Linda Giudice, Bruce A Lessey, Jose A Horcajadas, Nick S Macklon, Thomas D'Hooghe, Cristina Campoy, et al. Guidelines for the design, analysis and interpretation of 'omics' data: focus on human endometrium. *Human reproduction update*, 20(1): 12–28, 2014.

[4] David J Balding. A tutorial on statistical methods for population association studies. *Nature reviews genetics*, 7(10):781–791, 2006.

[5] Andrew J Bannister and Tony Kouzarides. Regulation of chromatin by histone modifications. *Cell research*, 21(3):381–395, 2011.

[6] Richard Barfield, Jincheng Shen, Allan C Just, Pantel S Vokonas, Joel Schwartz, Andrea A Baccarelli, Tyler J VanderWeele, and Xihong Lin. Testing for the indirect effect under the null for genome-wide mediation analyses. *Genetic epidemiology*, 41(8):824–833, 2017.

[7] Reuben M Baron and David A Kenny. The moderator–mediator variable dis-

tinction in social psychological research: Conceptual, strategic, and statistical considerations. *Journal of personality and social psychology*, 51(6):1173, 1986.

[8] Thomas Battram, Paul Yousefi, Gemma Crawford, Claire Prince, Mahsa Sheikhali Babaei, Gemma Sharp, Charlie Hatcher, María Jesús Vega-Salas, Sahar Khodabakhsh, Oliver Whitehurst, et al. The ewas catalog: a database of epigenome-wide association studies. *Wellcome Open Research*, 7 (41):41, 2022.

[9] Mario Bauer, Beate Fink, Loreen Thürmann, Markus Eszlinger, Gunda Herberth, and Irina Lehmann. Tobacco smoking differently influences cell types of the innate and adaptive immune system—indications from cpg site methylation. *Clinical epigenetics*, 8(1):1–12, 2016.

[10] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001.

[11] Marc Chadeau-Hyam, Gianluca Campanella, Thibaut Jombart, Leonardo Bottolo, Lutzen Portengen, Paolo Vineis, Benoit Liquet, and Roel CH Vermeulen. Deciphering the complex: Methodological overview of statistical models to derive omics-based biomarkers. *Environmental and molecular mutagenesis*, 54(7): 542–557, 2013.

[12] Robin F Chan, Gustavo Turecki, Andrey A Shabalin, Jerry Guintivano, Min Zhao, Lin Y Xie, Gerard van Grootheest, Zachary A Kaminsky, Brian Dean, Brenda WJH Penninx, et al. Cell-type-specific methylome-wide association studies implicate neurodegenerative processes and neuroimmune communication in major depressive disorder. *bioRxiv*, page 432088, 2018.

[13] Edward Y Chen, Christopher M Tan, Yan Kou, Qiaonan Duan, Zichen Wang, Gabriela Vaz Meirelles, Neil R Clark, and Avi Ma'ayan. Enrichr: interactive and

collaborative html5 gene list enrichment analysis tool. *BMC bioinformatics*, 14 (1):1–14, 2013.

[14] Jared M Churko, Gary L Mantalas, Michael P Snyder, and Joseph C Wu. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. *Circulation research*, 112(12):1613–1623, 2013.

[15] P Cohen, J Cohen, SG West, and LS Aiken. Applied multiple regression/correlation analysis for the behavioral sciences . hillsdale, nj: Erlbaum. *INTELLIGENCE AND ASSESSMENT*, 531, 1983.

[16] Victoria K Cortessis, Duncan C Thomas, A Joan Levine, Carrie V Breton, Thomas M Mack, Kimberly D Siegmund, Robert W Haile, and Peter W Laird. Environmental epigenetics: prospects for studying epigenetic mediation of exposure–response relationships. *Human genetics*, 131(10):1565–1589, 2012.

[17] Ellen W Demerath, Weihua Guan, Megan L Grove, Stella Aslibekyan, Michael Mendelson, Yi-Hui Zhou, Åsa K Hedman, Johanna K Sandling, Li-An Li, Marguerite R Irvin, et al. Epigenome-wide association study (ewas) of bmi, bmi change and waist circumference in african american adults identifies multiple replicated loci. *Human molecular genetics*, 24(15):4464–4479, 2015.

[18] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[19] Andriy Derkach, Ruth M Pfeiffer, Ting-Huei Chen, and Joshua N Sampson. High dimensional mediation analysis with latent variables. *Biometrics*, 75(3):745–756, 2019.

[20] Katherine J Dick, Christopher P Nelson, Loukia Tsaprouni, Johanna K Sandling, Dylan Aïssi, Simone Wahl, Eshwar Meduri, Pierre-Emmanuel Morange, France Gagnon, Harald Grallert, et al. Dna methylation and body-mass index: a genome-wide analysis. *The Lancet*, 383(9933):1990–1998, 2014.

[21] Vera Djordjilović, Christian M Page, Jon Michael Gran, Therese H Nøst, Torkjel M Sandanger, Marit B Veierød, and Magne Thoresen. Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in medicine*, 38(18):3346–3360, 2019.

[22] Ruiling Fang, Haitao Yang, Yuzhao Gao, Hongyan Cao, Ellen L Goode, and Yuehua Cui. Gene-based mediation analysis in epigenetic studies. *Briefings in bioinformatics*, 22(3):bbaa113, 2021.

[23] Evan Gatev, Nicole Gladish, Sara Mostafavi, and Michael S Kobor. Comeback: Dna methylation array data analysis for co-methylated regions. *Bioinformatics*, 36(9):2675–2683, 2020.

[24] Charles F Gillespie, Bekh Bradley, Kristie Mercer, Alicia K Smith, Karen Conneely, Mark Gapen, Tamara Weiss, Ann C Schwartz, Joseph F Cubells, and Kerry J Ressler. Trauma exposure and stress-related disorders in inner city primary care patients. *General hospital psychiatry*, 31(6):505–514, 2009.

[25] Shannon L Gillespie, Lynda R Hardy, and Cindy M Anderson. Patterns of dna methylation as an indicator of biological aging: State of the science and future directions in precision health promotion. *Nursing outlook*, 67(4):337–344, 2019.

[26] Crystal D Grant, Nadereh Jafari, Lifang Hou, Yun Li, James D Stewart, Guosheng Zhang, Archana Lamichhane, JoAnn E Manson, Andrea A Baccarelli, Eric A Whitsel, et al. A longitudinal study of dna methylation as a potential mediator of age-related diabetes risk. *Geroscience*, 39(5):475–489, 2017.

[27] Andrew F Hayes and Michael Scharkow. The relative trustworthiness of inferential tests of the indirect effect in statistical mediation analysis: does method really matter? *Psychological science*, 24(10):1918–1927, 2013.

[28] Matthew D Hoffman, Andrew Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1): 1593–1623, 2014.

[29] Yen-Tsung Huang. Genome-wide analyses of sparse mediation effects under composite null hypotheses. *The Annals of Applied Statistics*, 13(1):60–84, 2019.

[30] Yen-Tsung Huang and Wen-Chi Pan. Hypothesis test of mediation effect in causal mediation model with high-dimensional continuous mediators. *Biometrics*, 72(2):402–413, 2016.

[31] A. E. Jaffe and R. A. Irizarry. Accounting for cellular heterogeneity is critical in epigenome-wide association studies. *Genome biology*, 15(2):1–9, 2014.

[32] Seyma Katrinli, Yuanchao Zheng, Aarti Gautam, Rasha Hammamieh, Ruoting Yang, Suresh Venkateswaran, Varun Kilaru, Adriana Lori, Rebecca Hinrichs, Abigail Powers, et al. Ptsd is associated with increased dna methylation across regions of hla-dpb1 and spatc1l. *Brain, Behavior, and Immunity*, 91:429–436, 2021.

[33] Samuel T Keating and Assam El-Osta. Epigenetics and metabolism. *Circulation research*, 116(4):715–736, 2015.

[34] Torsten Klengel, Julius Pape, Elisabeth B Binder, and Divya Mehta. The role of dna methylation in stress-related psychiatric disorders. *Neuropharmacology*, 80:115–132, 2014.

[35] Lara Kular, Yun Liu, Sabrina Ruhrmann, Galina Zheleznyakova, Francesco Marabita, David Gomez-Cabrero, Tojo James, Ewoud Ewing, Magdalena Lindén, Bartosz Górnikiewicz, et al. Dna methylation as a mediator of hla-drb1* 15: 01 and a protective variant in multiple sclerosis. *Nature communications*, 9 (1):1–15, 2018.

[36] Maxim V Kuleshov, Matthew R Jones, Andrew D Rouillard, Nicolas F Fernandez, Qiaonan Duan, Zichen Wang, Simon Koplev, Sherry L Jenkins, Kathleen M Jagodnik, Alexander Lachmann, et al. Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. *Nucleic acids research*, 44(W1): W90–W97, 2016.

[37] Ken WK Lee and Zdenka Pausova. Cigarette smoking and dna methylation. *Frontiers in genetics*, 4:132, 2013.

[38] Shuai Li, Ee Ming Wong, Minh Bui, Tuong L Nguyen, Ji-Hoon Eric Joo, Jennifer Stone, Gillian S Dite, Graham G Giles, Richard Saffery, Melissa C Southey, et al. Causal effect of smoking on dna methylation in peripheral blood: a twin and family study. *Clinical epigenetics*, 10(1):1–12, 2018.

[39] Ziyi Li, Zhijin Wu, Peng Jin, and Hao Wu. Dissecting differential signals in high-throughput data from complex tissues. *Bioinformatics*, 35(20):3898–3905, 2019.

[40] Yun Liu, Martin J Aryee, Leonid Padyukov, M Daniele Fallin, Espen Hesselberg, Arni Runarsson, Lovisa Reinius, Nathalie Acevedo, Margaret Taub, Marcus Ronninger, et al. Epigenome-wide association data implicate dna methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nature biotechnology*, 31(2): 142–147, 2013.

[41] Zhonghua Liu, Jincheng Shen, Richard Barfield, Joel Schwartz, Andrea A Bac-

carelli, and Xihong Lin. Large-scale hypothesis testing for causal mediation effects with applications in genome-wide epigenetic studies. *Journal of the American Statistical Association*, pages 1–15, 2021.

[42] Xiangyu Luo, Can Yang, and Yingying Wei. Detection of cell-type-specific risk-cpg sites in epigenome-wide association studies. *Nature communications*, 10(1): 1–12, 2019.

[43] Xiangyu Luo, Joel Schwartz, Andrea Baccarelli, and Zhonghua Liu. Testing cell-type-specific mediation effects in genome-wide epigenetic studies. *Briefings in Bioinformatics*, 22(3):bbaa131, 2021.

[44] David P MacKinnon. *Introduction to statistical mediation analysis*. Routledge, 2012.

[45] David P MacKinnon, Chondra M Lockwood, Jeanne M Hoffman, Stephen G West, and Virgil Sheets. A comparison of methods to test mediation and other intervening variable effects. *Psychological methods*, 7(1):83, 2002.

[46] J Alfredo Martínez, Fermín I Milagro, Kate J Claycombe, and Kevin L Schalinske. Epigenetics in adipose tissue, obesity, weight loss, and diabetes. *Advances in nutrition*, 5(1):71–81, 2014.

[47] Natalie Matosin, Cristiana Cruceanu, and Elisabeth B Binder. Preclinical and clinical evidence of dna methylation changes in response to trauma and chronic stress. *Chronic Stress*, 1:2470547017710764, 2017.

[48] Kevin McGregor, Sasha Bernatsky, Ines Colmegna, Marie Hudson, Tomi Pastinen, Aurélie Labbe, and Celia MT Greenwood. An evaluation of methods correcting for cell-type heterogeneity in dna methylation studies. *Genome biology*, 17(1):1–17, 2016.

[49] Lisa D Moore, Thuc Le, and Guoping Fan. Dna methylation and its basic function. *Neuropsychopharmacology*, 38(1):23–38, 2013.

[50] Filomene G Morrison, Mark W Miller, Mark W Logue, Michele Assef, and Erika J Wolf. Dna methylation correlates of ptsd: Recent findings and technical challenges. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, 90: 223–234, 2019.

[51] Trang Quynh Nguyen, Ian Schmid, and Elizabeth A Stuart. Clarifying causal mediation analysis for the applied researcher: Defining effects based on what we want to learn. *Psychological methods*, 26(2):255, 2021.

[52] Kasper Mønsted Pedersen, Yunus Çolak, Christina Ellervik, Hans Carl Hasselbalch, Stig Egil Bojesen, and Børge Grønne Nordestgaard. Smoking and increased white and red blood cells: a mendelian randomization approach in the copenhagen general population study. *Arteriosclerosis, thrombosis, and vascular biology*, 39(5):965–977, 2019.

[53] Eleonora Porcu, Jennifer Sjaarda, Kaido Lepik, Cristian Carmeli, Liza Darrous, Jonathan Sulc, Ninon Mounier, and Zoltán Kutalik. Causal inference methods to integrate omics and complex traits. *Cold Spring Harbor Perspectives in Medicine*, 11(5):a040493, 2021.

[54] Kristopher J Preacher and Andrew F Hayes. Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior research methods*, 40(3):879–891, 2008.

[55] Huaizhen Qin, Tianhua Niu, and Jinying Zhao. Identifying multi-omics causers and causal pathways for complex traits. *Frontiers in genetics*, 10:110, 2019.

[56] Elior Rahmani, Regev Schweiger, Brooke Rhead, Lindsey A Criswell, Lisa F Barcellos, Eleazar Eskin, Saharon Rosset, Sriram Sankararaman, and Eran Halperin.

Cell-type-specific resolution epigenetics without the need for cell sorting or single-cell biology. *Nature communications*, 10(1):1–11, 2019.

[57] Vardhman K Rakyan, Thomas A Down, David J Balding, and Stephan Beck. Epigenome-wide association studies for common human diseases. *Nature Reviews Genetics*, 12(8):529–541, 2011.

[58] Lovisa E Reinius, Nathalie Acevedo, Maaike Joerink, Göran Pershagen, Sven-Erik Dahlén, Dario Greco, Cilla Söderhäll, Annika Scheynius, and Juha Kere. Differential dna methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. *PloS one*, 7(7):e41361, 2012.

[59] James M Robins and Sander Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, pages 143–155, 1992.

[60] Bart PF Rutten, Eric Vermetten, Christiaan H Vinkers, Gianluca Ursini, Nikolaos P Daskalakis, Ehsan Pishva, Laurence de Nijs, Lotte C Houtepen, Lars Eijssen, Andrew E Jaffe, et al. Longitudinal analyses of the dna methylome in deployed military servicemen identify susceptibility loci for post-traumatic stress disorder. *Molecular psychiatry*, 23(5):1145–1156, 2018.

[61] Shai S Shen-Orr and Renaud Gaujoux. Computational deconvolution: extracting cell type-specific information from heterogeneous samples. *Current opinion in immunology*, 25(5):571–578, 2013.

[62] Michael E Sobel. Asymptotic confidence intervals for indirect effects in structural equation models. *Sociological methodology*, 13:290–312, 1982.

[63] Yanyi Song, Xiang Zhou, Min Zhang, Wei Zhao, Yongmei Liu, Sharon LR Kardia, Ana V Diez Roux, Belinda L Needham, Jennifer A Smith, and Bhramar Mukherjee. Bayesian shrinkage estimation of high dimensional causal mediation effects in omics studies. *Biometrics*, 76(3):700–710, 2020.

[64] Yanyi Song, Xiang Zhou, Jian Kang, Max T Aung, Min Zhang, Wei Zhao, Belinda L Needham, Sharon LR Kardia, Yongmei Liu, John D Meeker, et al. Bayesian sparse mediation analysis with targeted penalization of natural indirect effects. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 2021.

[65] Yanyi Song, Xiang Zhou, Jian Kang, Max T Aung, Min Zhang, Wei Zhao, Belinda L Needham, Sharon LR Kardia, Yongmei Liu, John D Meeker, et al. Bayesian hierarchical models for high-dimensional mediation analysis with coordinated selection of correlated mediators. *Statistics in medicine*, 40(27):6038–6056, 2021.

[66] Stan Development Team. *Stan Modeling Language User's Guide and Reference Manual, Version 1.0*. 2012. URL `http://mc-stan.org/`.

[67] Stan Development Team. RStan: the R interface to Stan, 2018. URL `http://mc-stan.org/7`. R package version 2.17.3.

[68] Dan Su, Xuting Wang, Michelle R Campbell, Devin K Porter, Gary S Pittman, Brian D Bennett, Ma Wan, Neal A Englert, Christopher L Crowl, Ryan N Gimple, et al. Distinct epigenetic effects of tobacco smoking in whole blood and among leukocyte subtypes. *PloS one*, 11(12):e0166486, 2016.

[69] Andrew E Teschendorff, Charles E Breeze, Shijie C Zheng, and Stephan Beck. A comparison of reference-based algorithms for correcting cell-type heterogeneity in epigenome-wide association studies. *BMC bioinformatics*, 18(1):1–14, 2017.

[70] Elmar W Tobi, Roderick C Slieker, René Luijk, Koen F Dekkers, Aryeh D Stein, Kate M Xu, Biobank based Integrative Omics Studies Consortium, P Eline Slagboom, Erik W van Zwet, LH Lumey, et al. Dna methylation as a mediator of

the association between prenatal adversity and risk factors for metabolic disease in adulthood. *Science advances*, 4(1):eaao4364, 2018.

[71] Tyler VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

[72] Tyler VanderWeele and Stijn Vansteelandt. Mediation analysis with multiple mediators. *Epidemiologic methods*, 2(1):95–115, 2014.

[73] Simone Wahl, Alexander Drong, Benjamin Lehne, Marie Loh, William R Scott, Sonja Kunze, Pei-Chien Tsai, Janina S Ried, Weihua Zhang, Youwen Yang, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature*, 541(7635):81–86, 2017.

[74] Zhuorui Xie, Allison Bailey, Maxim V Kuleshov, Daniel JB Clarke, John E Evangelista, Sherry L Jenkins, Alexander Lachmann, Megan L Wojciechowicz, Eryk Kropiwnicki, Kathleen M Jagodnik, et al. Gene set knowledge discovery with enrichr. *Current protocols*, 1(3):e90, 2021.

[75] Xiaojing Xu, Shaoyong Su, Xin Wang, Vernon Barnes, Carmen De Miguel, Dennis Ownby, Jennifer Pollock, Harold Snieder, and Weiqin Chen. Obesity is associated with more activated neutrophils in african american male youth. *International journal of obesity*, 39(1):26–32, 2015.

[76] Qi Yan, Erick Forno, Juan C. Celedón, and Wei Chen. A region-based method for causal mediation analysis of dna methylation data. *Epigenetics*, pages 1–11, 2021.

[77] Sonja Zeilinger, Brigitte Kühnel, Norman Klopp, Hansjörg Baurecht, Anja Kleinschmidt, Christian Gieger, Stephan Weidinger, Eva Lattka, Jerzy Adamski, Annette Peters, et al. Tobacco smoking leads to extensive genome-wide changes in dna methylation. *PloS one*, 8(5):e63812, 2013.

[78] Ping Zeng, Zhonghe Shao, and Xiang Zhou. Statistical methods for mediation analysis in the era of high-throughput genomics: current successes and future challenges. *Computational and structural biotechnology journal*, 19:3209–3224, 2021.

[79] Haixiang Zhang, Yinan Zheng, Zhou Zhang, Tao Gao, Brian Joyce, Grace Yoon, Wei Zhang, Joel Schwartz, Allan Just, Elena Colicino, et al. Estimating and testing high-dimensional mediation effects in epigenetic studies. *Bioinformatics*, 32(20):3150–3154, 2016.

[80] Weiwei Zhang, Tim D Spector, Panos Deloukas, Jordana T Bell, and Barbara E Engelhardt. Predicting genome-wide dna methylation using methylation marks, genomic position, and dna regulatory elements. *Genome biology*, 16(1):1–20, 2015.

[81] Yi Zhao and Xi Luo. Pathway lasso: estimate and select sparse mediation pathways with high dimensional mediators. *arXiv preprint arXiv:1603.07749*, 2016.

[82] Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

[83] James Zou, Christoph Lippert, David Heckerman, Martin Aryee, and Jennifer Listgarten. Epigenome-wide association studies without the need for cell-type composition. *Nature methods*, 11(3):309–311, 2014.