THE UNIVERSITY OF CHICAGO

METHODS FOR MULTI-OMICS MULTI-CONTEXT INTEGRATIVE ANALYSIS

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE BIOLOGICAL SCIENCES
AND THE PRITZKER SCHOOL OF MEDICINE
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PUBLIC HEALTH SCIENCES

BY
YIHAO LU

CHICAGO, ILLINOIS
JUNE 2024

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to thank my advisor Dr. Lin Chen for her constant and constructive guidance throughout my training. Dr. Chen's dedication to excellence and high-quality research has profoundly impacted me. Her encouragement to explore innovative ideas not only enhanced my work but also inspired me to always pursue excellence. I am greatly thankful for her time, her sharing, and her support. I want to thank my dissertation committee: Drs Donald Hedeker, Brandon Pierce, and Jin Liu. I appreciate the help and input from you. Thank you co-authors and collaborators, especially Dr. Fan Yang, Dr. Meritxell Oliva, and Ke Xu for helpful discussions.

I would like to thank the Susan G. Komen Foundation for funding support (Susan G. Komen® TREND21675016). I also want to express my gratitude for the funding support from the Paul Meier Award.

ABSTRACT

Genome-wide association studies (GWAS) have identified hundreds of thousands of associations between genetic variants and human complex traits/diseases. To functionally annotate the trait/disease-associated variants, extensive efforts are made to study the genetic effects on downstream molecular phenotypes in a wide variety of tissue types and cell types. Genetic effects on functionally related 'omic' traits often co-occur in relevant cellular contexts, such as tissues. In Chapter 2, motivated by the multi-tissue methylation quantitative trait loci (mQTLs) and expression QTLs (eQTLs) analysis of Genotype-Tissue Expression project, we propose X-ING (Cross-INtegrative Genomics) for cross-omics and cross-context integrative analysis. A major innovation of the method is that it models latent association indicators instead of effect sizes and uses multi-view learning to account for major patterns among latent indicators across omics data types and tissue types. This facilitates integrative analysis of different data types of different effect distributions. Moving beyond the integrative association analysis, in Chapter 3 we develop a multi-context multivariable integrative Mendelian randomization framework, mintMR, for mapping expression and molecular traits as joint exposures. The proposed method overcomes the unique challenges in mapping risk genes, and these challenges are under-addressed by conventional Mendelian randomization methods. MintMR improves the estimation of sparse tissue-specific causal effects of multiple genes with a limited number of IVs by simultaneously modeling the latent tissue indicators of disease relevance across multiple gene regions and subsequently improving the estimation of latent disease-relevant probabilities. In Chapter 4, we further expand the framework to study risk genes in specific cell types using deep learning methods. Single-cell RNA sequencing (scRNA-seq) enables the high-throughput profiling of gene expression specific to cell types. We proposed a deep-cellMR method capturing the nonlinear and complex dependencies across cell types and further improving the estimation of cell-type-specific effect of each gene in each cell. The proposed methods in this dissertation can be broadly applied

to map multi-omics QTLs and study risk genes for complex traits and diseases, and they can be applied to many other data types for conducting integrative association and causal analyses.

CHAPTER 1

INTRODUCTION

1.1 GWAS, multi-omics data, QTLs and integrative analysis

Genome-wide association studies (GWAS) have identified a large number of associations between genetic variants and human complex traits/diseases [Loos, 2020, Hindorff, 2009, Witte, 2010, MacArthur et al., 2017, Buniello et al., 2019]. Many of these variants affect complex traits/diseases via effects on gene expression levels and other molecular traits (e.g., DNA methylation) [Umans et al., 2021, Oliva et al., 2023]. However, existing quantitative trait loci (QTLs) explain only a small fraction of those variants [Consortium, 2020]. Complementary approaches and data sources are needed to further explain and understand the functionality of GWAS variants. Mapping genetic variants to their associated molecular traits has become an essential step in the functional annotation of genetic variants [Hormozdiari et al., 2018, Gamazon et al., 2018].

The genetic effects on molecular traits often depend on cellular contexts (e.g., tissues and cell types) [Reuter et al., 2015, Sun et al., 2018, Consortium, 2020, Dries et al., 2021]. Extensive efforts are made to study the genetic effects on downstream molecular traits in a wide variety of tissue types and cell types [Consortium, 2020, Oliva et al., 2023, Bryois et al., 2022]. The GTEx project (v8) [Consortium, 2020] studies the genetic effects on transcriptome across 49 human tissues from 838 donors and has provided a comprehensive list of expression QTLs (eQTLs) by tissue types. Recently, the enhancing GTEx (eGTEx) project [Oliva et al., 2023] sought to complement existing GTEx data with additional molecular traits, including DNA methylation (DNAm). DNAm data from 987 GTEx samples representing nine tissue types were made available, enabling further characterization of the relationships among genetic variants, DNAm, gene expression, and disease in a tissue-specific manner. Most cis-eQTLs co-occur with a methylation QTL (cis-mQTL), suggesting a common causal

variant and shared biological mechanism [Pierce et al., 2018].

In recent years, single-cell RNA sequencing (scRNA-seq) has revolutionized the landscape of genomics by enabling the high-throughput profiling of gene expression specific to cell types and states [Hwang et al., 2018, Tang et al., 2009]. It facilitates single-cell expression quantitative trait loci (sc-eQTL) mapping across different cell types [van der Wijst et al., 2020, Yazar et al., 2022, Oelen et al., 2022, Soskic et al., 2022], revealing how expression levels associated with trait/disease-related genetic variants in specific cell populations [de Vries et al., 2020, Nathan et al., 2022].

The goal of this dissertation is to develop statistical methods and computational tools to conduct integrative association and causal inference analysis to identify genetic effects on multi-omics traits in specific tissues/cell types, and map risk genes in disease-relevant contexts/tissues/cell types by leveraging summary statistics. The integrative analysis using summary statistics from multi-omics and multi-context studies could collectively provide a comprehensive understanding of the dynamic mechanisms underlying human complex diseases and traits.

1.2 Integrative multi-omics multi-context association analysis

The detection of multi-omics QTLs and genetic variants with cascading effects allows the study of causal variants and shared biological mechanisms [Gleason et al., 2020, Ng et al., 2021]. In the analyses of QTLs, it is critical to consider the effect-operating cellular contexts. Many disease-associated genetic and genomic effects are highly context-specific [Consortium, 2020, Umans et al., 2021]. Motivated by the multi-omics multi-tissue QTL analysis, we propose X-ING (Cross-INtegrative Genomics) as a general framework for the cross-integration of summary statistics from multi-omics data each with multiple cellular contexts. X-ING takes as input the summary statistic matrices from multiple data types and models each input statistic as a product of Gaussian and latent binary association status. The model-

ing of latent binary association status matrices across data types is novel, as it allows the cross-integration of different data types of different effect distributions. X-ING captures omics-shared and context-shared association patterns. This is a major innovation compared with existing multi-context/tissue or multi-omics integration methods.

A joint analysis of many tissues, or broadly relevant cellular contexts and conditions, may reveal otherwise hidden and dynamic mechanisms underlying diseases of interest. Previous work integrating and leveraging summary-level data from a variety of cellular contexts, cell/tissue types, and conditions shows improved estimation and detection of disease/trait-relevant effects. Specifically, mash [Urbut et al., 2019] captures shared information across contexts through a mixture model. MT-eQTL [Li et al., 2018] adopts a hierarchical Bayesian model to jointly model eQTL association statistics from multiple tissues. Metasoft [Han and Eskin, 2012] models the effect variation across contexts using a random-effects model.

Existing studies enable multi-tissue integration while challenges arise when integrating effects from multi-omics data. Those effects do not have shared magnitudes and effect distributions. To take advantage of the shared patterns among true nonzero effects, e.g., co-occurring eQTLs and mQTLs. We propose X-ING for multi-omics multi-context integrative analysis. X-ING models the latent binary association status of each statistic to capture shared effect co-occurrence patterns across data types despite different effect distributions.

1.3 Integrative Mendelian randomization for identifying risk genes across human tissue

The proliferation of GWAS results has facilitated studies of using genetic variants to further map risk factors for complex diseases, moving beyond association toward causation. Mendelian Randomization (MR) is one of those causal inference approaches that have achieved many successes in studying the causal relationships between an intermediate phenotype (exposure) and a complex disease (outcome) [Burgess et al., 2013, Bowden et al., 2015, Zhao

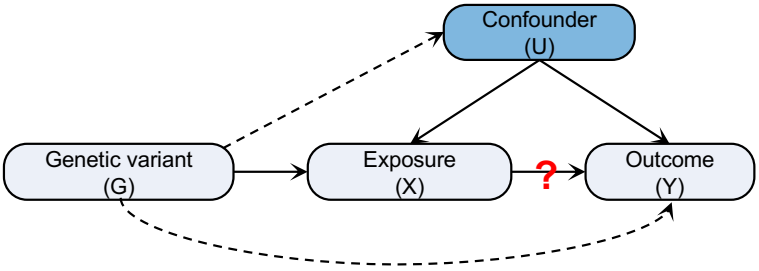


Figure 1.1: An illustration of Mendelian randomization (MR). A genetic variant (SNP) is used as an instrumental variable (IV) to assess a potential causal relationship between exposure and disease outcome. Marginal summary statistics assessing the effect of the SNP on the exposure from one study are integrated with marginal summary statistics from another study assessing the IV-to-outcome effects to infer the effect of the exposure on the outcome.

et al., 2020, Cheng et al., 2022, Wang et al., 2021, Xue et al., 2021, Morrison et al., 2020]. MR treats genetic variants associated with risk exposures as instrumental variables (IVs) to assess the causal effects from exposure on outcome (Fig. 1.1) [Chen et al., 2007, Lawlor et al., 2008, Schadt et al., 2005, Davey Smith and Ebrahim, 2003].

MR and multivariable MR (MVMR) methods imposed assumptions on the validity of IVs. Assumptions are violated when invalid IVs with uncorrelated horizontal pleiotropy (CHP) or correlated horizontal pleiotropy (CHP) are included (Fig. 1.1). Specifically, IVW and MVMR-IVW do not allow invalid IVs [Slob and Burgess, 2020, Burgess and Thompson, 2015]; MR-Egger and MVMR-Egger require Instrument Strength Independent of Direct Effect (InSIDE) assumption [Bowden et al., 2015, Rees et al., 2017]; MVMR-Median assumes the majority of IVs are valid [Grant and Burgess, 2021]; MVMR-Lasso and MVMR-Robust are robust to outliers (few invalid IVs) [Grant and Burgess, 2021]; and MVcML requires plurality condition where the valid IVs form the largest group to give the causal parameter estimate [Lin et al., 2023].

Most existing MR or multivariable MR (MVMR) methods were developed to analyze complex traits as exposure. When studying molecular traits such as gene expression levels as risk exposures for a disease outcome, new challenges arise in MR analyses. First, the number of eQTLs as IVs is limited [Gleason et al., 2020, Consortium, 2020] with cis-eQTLs

being generally correlated [Consortium, 2020]. Second, in multi-tissue MR analysis, the causal effects of genes on diseases are often tissue-specific and sparse, [Hekselman and Yeger-Lotem, 2020, Ongen et al., 2017, Feng et al., 2021] and thus the estimation of tissue-specific causal effects with a limited number of eQTLs/IVs is challenging. Third, eQTL effects as IV-to-exposure effects may not be consistent across two (GWAS and eQTL) samples. It has been reported that eQTL effects can be tissue-specific and may not be shared across all tissues and studies.

We propose a multi-context multivariable integrative Mendelian randomization (mintMR) method for mapping gene expression and molecular traits as risk exposures. For each gene, we perform a multi-tissue MR analysis using eQTLs with non-zero and sign-consistent effects in more than one tissue as IVs, thereby improving the IV consistency. Our method improves the identification of tissue-specific causal effects of all genes by simultaneously modeling the latent disease-relevance tissue indicators for multiple gene regions, jointly learning the major/low-rank patterns of latent indicators/probabilities via multi-view learning techniques, and then using the major patterns to estimate and update the probability of non-zero effects. The joint learning of disease-relevance of latent tissue indicators improves the identification of sparse tissue-specific causal effects for all genes.

1.4 Deep learning Mendelian randomization method for unveiling risk genes in specific cell types

This advancement in scRNA-seq has enabled sc-eQTL mapping. Despite the insights gained from sc-eQTL studies, new challenges arise when mapping risk genes in specific cell types using Mendelian randomization due to the limited power of sc-eQTL data [He et al., 2021, Bryois et al., 2022, Lopes et al., 2022, Young et al., 2021, Jerber et al., 2021]. To improve the power of sc-eQTLs, recent studies integrate sc-eQTL with bulk-tissue eQTL (bk-eQTL) and cell type-specific eQTLs (ct-eQTLs) derived from bulk tissues [Donovan et al., 2020,

Consortium et al., 2020]. To integration of sc-eQTL, ct-eQTLs from bk, with GWAS, an integrative MR faces challenges including heterogeneity in effect sizes [Ding et al., 2024], sparsity of effects of risk genes in specific cell types, and the prevalence of missing values due to the sequencing depth and the non-expression of some genes in specific cells [Zhang et al., 2024, Pool et al., 2023, Hicks et al., 2018].

As risk genes for a disease often show non-zero effects in specific cell types [Boyle et al., 2017, Finucane et al., 2018, Lynall et al., 2022], we propose a joint MR model (deep-cellMR) across multiple genes to estimate the causal effects for each gene in each cell type. We perform the integration of sc-eQTL, ct-eQTLs, with GWAS. We introduce a latent effect indicator of each gene in each cell type and models the low-rank patterns of these latent indicators across genes, cells, and data via deep learning. To handle missingness in sc-eQTL summary statistics, deep-cellMR uses random forest for missing value imputation in latent variables [Stekhoven and Bühlmann, 2012]. Via applying deep multi-view learning methods [Yan et al., 2021, Wang et al., 2015, 2016, Andrew et al., 2013, Ngiam et al., 2011] on the latent indicators across genes, cell types, and data, deep-cellMR captures the underlying nonlinear and complex dependencies and updates the estimation of cell-type-specific effect accounting for those learned patterns.

1.5 Summary

In this work, we develop statistical methods and computational tools to conduct integrative association and causal inference analysis to identify QTLs, and multi-omics QTLs in specific tissues/cell types, and map risk genes in disease-relevant contexts/tissues/cell types by leveraging summary statistics.

The rest of the dissertation is organized as follows. In Chapter 2, we develop an integrative cross-omics and cross-context method and use the method to enhance the power for detecting mQTLs by integrating eQTL statistics. We also study multi-tissue trans-association effects

and integrate spatial differential expression statistics from spatial transcriptomic studies with multi-tissue eQTL statistics from GTEx. In Chapter 3, we propose a multi-context multivariable integrative Mendelian randomization method, mintMR. We apply mintMR to map risk genes for 35 complex traits and diseases. It detects risk genes in disease-relevant tissues while maintaining reasonable controls of genome-wide inflation for the examined traits and diseases. In Chapter 4, we propose a deep-learning-based multi-cell-type MR method and use it to study cell-type-specific effects of risk genes on complex diseases. In Chapter 5, we summarize the presented works and suggest possible directions for future research.

CHAPTER 2

INTEGRATIVE CROSS-OMICS AND CROSS-CONTEXT ANALYSIS ELUCIDATES MOLECULAR LINKS UNDERLYING GENETIC EFFECTS ON COMPLEX TRAITS

2.1 Attributions

Dr. Lin Chen conceived the project. Drs. Lin Chen and Jin Liu contributed to the development of the methods and the writing of the manuscript. Dr. Meritxell Oliva contributed to the data analysis. Drs. Meritxell Oliva and Brandon Pierce provided valuable suggestions to the development of the methods and analyses.

2.2 Introduction

Mapping genetic variants to their associated molecular 'omic' traits has emerged as an essential step in the functional annotation of genetic variants [Hormozdiari et al., 2018, Gamazon et al., 2018]. Rich resources of genetic and genomic data are made available for different molecular omic traits from various cellular contexts (e.g., tissues, cell types) [Reuter et al., 2015, Sun et al., 2018, Consortium, 2020, Dries et al., 2021]. It has been reported that expression quantitative trait loci (eQTLs) and methylation QTLs (mQTLs) often co-occur [Pierce et al., 2018]. The detection of QTLs with cascading effects on multiple omics traits allows the study of functional variants and shared biological mechanisms, while also improving the power and precision for identifying disease/trait-relevant genes influenced by susceptibility loci [Gleason et al., 2020, Ng et al., 2021]. To efficiently leverage the huge amount of existing data, summary statistics provide a convenient way with secured privacy in sharing data from a wide range of studies [MacArthur et al., 2021, Hormozdiari et al., 2018]. The integrative analysis of association summary statistics of genetic effects on multi-omics traits could col-

lectively provide a comprehensive view and offer new insights into the dynamic mechanisms underlying human complex diseases and traits [Subramanian et al., 2020].

In a cross-omics integrative analysis, it is critical to consider the effect-operating cellular contexts. The genetic effects on omic traits depend on cellular contexts (e.g., tissues and cell types). Many disease-associated genetic and genomic effects are highly context-specific [Consortium, 2020, Umans et al., 2021]. For example, studies of common-variant genetic results for schizophrenia (SCZ) showed concerted effects in certain cell types [Skene et al., 2018], while many different cell types play distinct functions in disease etiology. Moreover, most prior studies of genetic effects of molecular traits, i.e., QTL studies, are based on molecular trait measurements from bulk tissues, and those measurements are averaged across many functionally divergent cell types [Blum et al., 2018, Consortium, 2020, Võsa et al., 2021]. There could be multiple tissue types containing relevant cell types [Eraslan et al., 2022]. A joint analysis of many tissues, or broadly relevant cellular contexts and conditions, may reveal otherwise hidden and dynamic mechanisms underlying diseases of interest. Previous work integrating and leveraging summary-level data from a variety of cellular contexts [Battle et al., 2017], cell/tissue types [Shi et al., 2020, Hu et al., 2019], and conditions [Yao et al., 2018, Urbut et al., 2019] show improved estimation and detection of disease/trait relevant effects.

Recently, the enhancing GTEx (eGTEx) project sought to complement existing multi-tissue human transcriptome data with additional molecular traits, including DNA methylation (DNAm) [eGTEx Project., 2017]. DNAm data from 987 GTEx samples representing nine tissue types were made available to further characterize the relationships among inherited genetic variants, DNAm, gene expression, and disease in a tissue-specific manner. Motivated by the multi-tissue mQTL and eQTL analysis, we propose X-ING (Cross-INtegrative Genomics) for cross-omics and cross-context integrative analysis. When integrating effects from multi-omics data, those effects do not share the same magnitudes and effect distri-

butions. Yet there could still be linked or concerted patterns among true nonzero effects, e.g., co-occurring eQTLs and mQTLs [Pierce et al., 2018]. The proposed X-ING method is built on a Bayesian hierarchical model capturing major patterns in the latent association status. A unique feature of X-ING is its ability to simultaneously account for omics-shared and omics-specific context/tissue-shared patterns in the analysis, while also allowing for effect heterogeneity and different levels of sparsity in each context and data type. Although X-ING is motivated by and primarily focuses on multi-tissue eQTL and mQTL analysis, it can be broadly applied to integrate multiple sets of summary statistics from different sources/domains to enhance cross-feature cross-context learning. In this work, X-ING is used to enhance the power for detecting mQTLs by integrating eQTL statistics. Furthermore, we apply X-ING to examine multi-tissue trans-association effects. Our analysis reveals that associations identified by X-ING are enriched in many known disease/trait-relevant tissue types. Additionally, we illustrate how X-ING can integrate spatial differential expression statistics from spatial transcriptomic studies with multi-tissue eQTL statistics from GTEx. The integrative analysis provides new insights into spatially defined molecular mechanisms underlying diseases.

2.3 Methods

2.3.1 Overview

Recently, the GTEx consortium released the single-tissue mQTL summary statistics for nine selected tissue types from 424 GTEx participants. Due to the limited tissue-specific sample sizes, the detection of mQTLs is underpowered compared with existing eQTL maps. Moreover, a large majority of mQTLs lack functionality [Oliva et al., 2023]. It has been reported that eQTLs often co-occur with mQTLs [Pierce et al., 2018]. The joint analysis of associations of a genetic variant to a cis-gene and a cis-CpG site (a tested trio) could enhance

the power and precision in detecting mQTLs, and facilitate the functional interpretations. We develop an integrative association method, X-ING, to jointly analyze multi-tissue mQTL and eQTL statistics, as illustrated in Figure 2.1a-c. The X-ING method takes as input the association statistic matrices of M tested units from L types of omics studies, where each type of omics study has K_ℓ ($\ell = 1, \ldots, L$) cellular contexts, i.e., L matrices each of dimension $M \times K_\ell$. The outputs of X-ING are the posterior mean and the posterior probability of nonzero effect for each input statistic (Figure 2.1d). The posterior probabilities allow flexible inference. For example, we may identify mQTLs with multi-tissue effects (i.e., having effects in two or more DNAm tissues with posterior probabilities >80%), or mQTLs with co-occurring associations to cis-expression (i.e., also with effects in at least one expression tissue) at the 80% posterior probability cutoffs. One may also calculate false discovery rates (FDRs) based on the posterior probabilities [Storey and Tibshirani, 2003, Chen et al., 2007].

In the motivating multi-tissue e/mQTL analysis ($L = 2$), a tested unit i ($= 1, \ldots, M$) is a trio consisting of a single nucleotide polymorphism (SNP), a cis-gene, and a cis-CpG site from different human tissues, and K_ℓ is the number of tissue types in the ℓ-th e/mQTL data. The input association summary statistics are Z-statistics obtained from single-tissue e/mQTL analysis. X-ING models the joint association patterns of Z-statistics (as Z-scores) for tested trios across omics data types and tissues. X-ING assumes those Z-scores from K_ℓ tissues, $\boldsymbol{z}_{i\cdot,\ell}$, following a multivariate normal distribution as follows

$$\boldsymbol{z}_{i\cdot,\ell} \sim \mathcal{N}\left(\widetilde{\boldsymbol{z}}_{i\cdot,\ell} \circ \boldsymbol{\gamma}_{i\cdot,\ell}, \mathbf{R}_\ell\right), \tag{2.1}$$

where $\widetilde{\boldsymbol{z}}_{i\cdot,\ell}$ is a vector of latent genetic association Z-scores (or effect sizes), with $\widetilde{z}_{ij,\ell} \sim \mathcal{N}(\widetilde{z}_{ij,\ell}|0, \sigma^2_{j,\ell})$ in each of the j-th cellular context ($j = 1, \ldots, K_\ell$), $\circ$ denotes the element-wise product of two vectors, $\boldsymbol{\gamma}_{i\cdot,\ell} \in \mathbb{R}^{K_\ell}$ is a vector of latent binary association status, with one denoting the presence of a nonzero effect, and $\mathbf{R}_\ell \in \mathbb{R}^{K_\ell \times K_\ell}$ is a tissue-tissue

correlation matrix (or covariance matrix if effect sizes instead of Z-scores being modeled) among all K_ℓ tissues due to potential sample overlap. The correlation matrix $\mathbf{R}_\ell$ due to sample overlapping (under the null, not of interest) is estimated a priori and is taken as known [Urbut et al., 2019, Gleason et al., 2020]. The latent association status indicator $\gamma_{ij,\ell}$ models sparsity in true nonzero effects. Similar modeling of latent indicators has been used in previous works to model the presence of nonzero Z-scores from multi-tissue eQTL analysis [Li et al., 2018, Urbut et al., 2019], and Z-scores from GWAS [Liu et al., 2017] among others. For different omics data types, effect size distributions are different. Some true nonzero effects are of opposite directions but are co-occurring. The joint modeling of latent indicators for nonzero effects captures shared effect co-occurrence patterns across data types despite different effect distributions – a major innovation of the proposed model. It should be noted that for inference purposes (i.e., detecting nonzero effects being the main goal), the modeling of Z-scores is similar to the modeling of effect sizes and their standard errors [Urbut et al., 2019] and we use Z-scores as an illustration. Each of the L latent Z-score matrices ($\widetilde{\boldsymbol{z}}_{\cdot\cdot,\ell}$) captures the multivariate Gaussian distributions of latent Z-score values from multivariate contexts/tissues within each data type, while the L latent binary matrices ($\boldsymbol{\gamma}_{\cdot\cdot,\ell}$) further capture the major (low-rank) effect-sharing patterns across omics data types and contexts.

A key innovation of X-ING is that it models the patterns of latent binary association status together with effect sizes. This allows the integration of two or more statistic matrices from different data types of various effect distributions and arbitrary structures. In the modeling of latent binary association status $\gamma_{ij,\ell}$, X-ING links (with a logit link) it to a latent low-rank continuous matrix $\boldsymbol{U}_\ell \in \mathbb{R}^{M\times K_\ell}$ to capture the major effect-sharing patterns across omics and across contexts (Figure 2.1d) [Liu et al., 2017],

$$\log \frac{p\left(\gamma_{ij,\ell}=1|\boldsymbol{U}_\ell,\boldsymbol{u}_{0,\ell}\right)}{p\left(\gamma_{ij,\ell}=0|\boldsymbol{U}_\ell,\boldsymbol{u}_{0,\ell}\right)} = U_{ij,\ell} + u_{0j,\ell}, \tag{2.2}$$

where $u_{0j,\ell}$ is the context-specific intercept, controlling the sparsity of nonzero effects in the j-th context of the ℓ-th omics data. The latent matrix $\boldsymbol{U}_\ell$ captures the desired patterns in the data and modulates the latent probability of nonzero associations. When $\boldsymbol{U}_\ell = 0$, there is no borrowing information across contexts/data. Here we propose to capture omics-shared and data-specific context-shared patterns via latent low-rank approximated modulation matrices, $\boldsymbol{U}_\ell = \boldsymbol{U}_{\ell O} + \boldsymbol{U}_{\ell C}, \forall \ell \in \{1, \ldots, L\}$, where $\boldsymbol{U}_{\ell O}$ represents the major (i.e., low-rank with rank p_ℓ) omics-shared data structures for data ℓ, and $\boldsymbol{U}_{\ell C}$ represents the major (rank q_ℓ) data-specific context-shared structures for data ℓ. With a computationally efficient expectation-maximization (EM) algorithm using variational inference, in each iteration, X-ING applies (generalized) canonical correlation analysis (CCA) on the logit-transformed latent probability matrices and retains the top p_ℓ canonical coefficients to obtain $\boldsymbol{U}_{\ell O}$. The number of retained components p_ℓ is determined using parallel analysis [Buja and Eyuboglu, 1992, Franklin et al., 1995]. It then applies principal component analysis (PCA) on each residual matrix after subtracting the $\boldsymbol{U}_{\ell O}$ matrix to estimate the data-specific context-shared matrix $\boldsymbol{U}_{\ell C}$. The number of retained principal components is also determined using parallel analysis. The sequential estimation of $\boldsymbol{U}_{\ell O}$ and $\boldsymbol{U}_{\ell C}$ not only facilitates the computation and also enhances the interpretability of the obtained CCs and PCs within the data context. See Algorithm 1 for details. The performance of X-ING is robust to the choices of p_ℓ and q_ℓ within a reasonable range. X-ING enables the integration of a broader class of test statistic matrices across different (L) data types while capturing the major structures of effects of similar nature from multiple (K_ℓ) contexts.

2.3.2 A starting Bayesian model without the modeling of shared data patterns

Assuming independence among all M tested units/trios, for each trio i ($i = 1, \ldots, M$) in omics data ℓ, we assume its Z-scores from K_ℓ tissues, $\boldsymbol{z}_{i\cdot,\ell}$, following a multivariate normal

distribution as Equation (2.1),

$$\boldsymbol{z}_{i\cdot,\ell} \sim \mathcal{N}\left(\widetilde{\boldsymbol{z}}_{i\cdot,\ell} \circ \boldsymbol{\gamma}_{i\cdot,\ell}, \mathbf{R}_\ell\right), \tag{2.3}$$

where $\widetilde{\boldsymbol{z}}_{i\cdot,\ell}$ and $\boldsymbol{\gamma}_{i\cdot,\ell}$ denote a vector of the latent genetic association Z-score values and latent binary association status, respectively. We further assume that each $\widetilde{z}_{ij,\ell}$ takes a Gaussian prior, $\mathcal{N}(\widetilde{z}_{ij,\ell}|0, \sigma^2_{j,\ell})$, and each $\gamma_{ij,\ell}$ takes a Bernoulli distribution with success probability $\pi_{j,\ell}$, which controls the sparsity of nonzero effects in tissue j for data ℓ. This specification is equivalent to assuming a spike-slab prior for the product, denoted as $\boldsymbol{\eta}_{i\cdot,\ell}$, of $\widetilde{\boldsymbol{z}}_{i\cdot,\ell}$ and $\boldsymbol{\gamma}_{i\cdot,\ell}$ [Yang et al., 2018, Shi et al., 2019], with each $\eta_{ij,\ell}$ distributed as

$$\eta_{ij,\ell} \sim \begin{cases} \mathcal{N}(\eta_{ij,\ell}|0, \sigma^2_{j,\ell}), & \text{if } \gamma_{ij,\ell} = 1, \\ \delta_0(\eta_{ij,\ell}), & \text{if } \gamma_{ij,\ell} = 0. \end{cases}$$

where δ_0 is a Dirac delta function at zero. To promote computational efficiency, we utilize the specification in Equation (2.1). The complete-data likelihood can be written as

$$\begin{aligned} p\left(\boldsymbol{z}, \widetilde{\boldsymbol{z}}, \boldsymbol{\gamma}; \Theta_1\right) &= \prod_{\ell=1}^{L} \left(\prod_{i=1}^{M} p\left(\boldsymbol{z}_{i\cdot,\ell}|\widetilde{\boldsymbol{z}}_{i\cdot,\ell}, \boldsymbol{\gamma}_{i\cdot,\ell}\right) \cdot \prod_{i=1}^{M}\prod_{j=1}^{K_\ell} p\left(\widetilde{z}_{ij,\ell}|\sigma^2_{j,\ell}\right) p\left(\gamma_{ij,\ell}|\pi_{j,\ell}\right) \right), \\ &= \prod_{\ell=1}^{L} \left(\prod_{i=1}^{M} \mathcal{N}\left(\boldsymbol{z}_{i\cdot,\ell}|\widetilde{\boldsymbol{z}}_{i\cdot,\ell} \circ \boldsymbol{\gamma}_{i\cdot,\ell}, \mathbf{R}_\ell\right) \cdot \prod_{i=1}^{M}\prod_{j=1}^{K_\ell} \mathcal{N}\left(\widetilde{z}_{ij,\ell}|0, \sigma^2_{j,\ell}\right) \cdot \pi_{j,\ell}^{\gamma_{ij,\ell}} \left(1-\pi_{j,\ell}\right)^{1-\gamma_{ij,\ell}} \right), \end{aligned} \tag{2.4}$$

where $\Theta_1 = \{\mathbf{R}_\ell, \sigma^2_{j,\ell}, \pi_{j,\ell}, \ell = 1, \ldots, L, j = 1, \ldots, K_\ell\}$ is the collection of model parameters and $\circ$ denotes element-wise product of two vectors. In practice, the tissue-tissue correlation matrix $\mathbf{R}_\ell$ due to sample overlap is often pre-estimated and treated as known [Urbut et al., 2019]. Existing literature often estimates it using a subset of the input Z-statistics that are likely to be from the null distributions (for example, using only SNPs with $|Z| < 5$ in all tissues to estimate the tissue-tissue correlation matrix) [Urbut et al., 2019, Gleason et al.,

2020]. Note that in the starting model (2.4) without modeling shared data patterns, the prior for $\boldsymbol{\gamma}_{\cdot j,\ell}$ is the same for all tested units in the j-th tissue from data ℓ.

To obtain the estimates of parameters in model (2.4), we need to compute the conditional density of latent variables given the observed Z-scores,

$$p(\widetilde{\boldsymbol{z}}, \boldsymbol{\gamma} | \boldsymbol{z}; \Theta_1) = \frac{p(\boldsymbol{z}, \widetilde{\boldsymbol{z}}, \boldsymbol{\gamma}; \Theta_1)}{p(\boldsymbol{z}; \Theta_1)}, \tag{2.5}$$

where $\boldsymbol{z}$, $\widetilde{\boldsymbol{z}}$ and $\boldsymbol{\gamma}$ are the collections of $\boldsymbol{z}_\ell$'s, $\widetilde{\boldsymbol{z}}_\ell$'s and $\boldsymbol{\gamma}_\ell$'s, respectively. To facilitate computation, we could apply an empirical Bayes approach to estimate the conditional density with the specification of distributions for hyperparameters. EM algorithms are usually used to obtain the estimates for models with latent variables. Here, the difficulty of employing standard EM algorithms comes from two folds. First, due to the curse of dimensionality, it would be computationally intensive to evaluate the expectation of high-dimensional latent variables. Second, it would be computationally intractable to integrate out the latent variables with spike-slab priors. To efficiently estimate the parameters in model (2.4), we use an EM algorithm with variational inference.

2.3.3 A Bayesian hierarchical model for the X-ING method

When jointly analyzing summary statistics from multi-omics data each with multivariate cellular contexts, some association patterns are shared between omics data types. For example, eQTLs and mQTLs often co-occur. Some cellular contexts are more correlated than others. The proposed cross-integrative genomics method, X-ING, accounts for those omics-shared and context-shared association patterns. In contrast to existing single-omics methods modeling effect sizes from multiple contexts, the proposed X-ING method jointly analyzes L matrices of Z-statistics and models the latent binary association status, $\{\boldsymbol{\gamma}_\ell\}$, to facilitate the modeling of effect co-occurring patterns from multi-omics data of different nature

and distributions. X-ING enables a broader class of integrative analyses across different (L) data types while also allowing the integration of statistics of similar nature across different (K_ℓ) contexts – a major advantage of X-ING. We develop an EM algorithm with variational inference for the model. In each E-step, we evaluate the posterior distribution of latent variables and obtain the variational parameters. In the M-step, we extract the common features shared among matrices of logit-transformed posterior probabilities of latent binary status by performing a CCA and PCA.

In detail, we build a Bayesian hierarchical framework to model the structured major patterns in the latent association status, $\gamma_{ij,\ell}$. To promote sparsity and model major data patterns across multiple omics and contexts, we modulate the prior probability of the latent status and link it with a latent low-rank matrix $\boldsymbol{U}_\ell \in \mathbb{R}^{M\times K_\ell}$ that captures omics-shared and context-shared major patterns via a logit link, as Equation (2.2).

Via the above modeling, our method efficiently allows the prior probability being specific for each tested unit (pair, trio, etc.) without over-parametrization, $p(\gamma_{ij,\ell}=1|\boldsymbol{U}_\ell, \boldsymbol{u}_{0,\ell}) = \pi_{ij,\ell}$. Moreover, the joint estimation of the low-rank matrix $\boldsymbol{U}_\ell$ across L data types also allows us to capture the shared information across different data types with an unknown extent of information sharing. Combining the prior probability in (2.2) and the Gaussian prior for $\widetilde{\boldsymbol{z}}$ with the multivariate normal distribution for Z-scores in (2.1), the complete-data likelihood for X-ING can be written as

$$p\left(\boldsymbol{z},\widetilde{\boldsymbol{z}},\boldsymbol{\gamma}\right)=\prod_{\ell=1}^{L}\left(\prod_{i=1}^{M}\mathcal{N}\left(\boldsymbol{z}_{i\cdot,\ell}|\widetilde{\boldsymbol{z}}_{i\cdot,\ell}\circ\boldsymbol{\gamma}_{i\cdot,\ell},\mathbf{R}_\ell\right)\cdot\prod_{i=1}^{M}\prod_{j=1}^{K_\ell}\mathcal{N}\left(\widetilde{z}_{ij,\ell}|0,\sigma_{j,\ell}^2\right)\cdot\pi_{ij,\ell}^{\gamma_{ij,\ell}}\left(1-\pi_{ij,\ell}\right)^{1-\gamma_{ij,\ell}}\right) \quad (2.6)$$

Comparing with the starting model (2.4), the X-ING model (2.6) allows each tested unit to have a specific prior that is modulated via a low-rank term, $U_{ij,\ell}$, based on a logit transformation of the latent association indicator, $\gamma_{ij,\ell}$.

When considering an $M\times K_\ell$ matrix of statistics from a single omics data ($L=1$), there exists no data-shared structure and we may regularize the rank of $\boldsymbol{U}_\ell$ using the nuclear

norm [Liu et al., 2017]. If association pattern sharing is limited across K_ℓ tissues or when a larger regularization is imposed, the low-rank matrix $\boldsymbol{U}_\ell$ may become a zero matrix and the prior model (2.2) reduces to a Bernoulli random variable with a shared probability parameter as in model (2.4), i.e., $p(\gamma_{ij,\ell} = 1|\boldsymbol{u}_{0,\ell}) = \pi_{j,\ell}$, indicating only tissue-specific sparse priors being imposed. When jointly estimating the low-rank matrices across L data types, we could estimate the latent low-rank matrix $\boldsymbol{U}_\ell$ as $\boldsymbol{U}_\ell = \boldsymbol{U}_{\ell O} + \boldsymbol{U}_{\ell C}$, where $\boldsymbol{U}_{\ell O}$ captures the omics-shared major patterns across L omics data types, and $\boldsymbol{U}_{\ell C}$ captures data-specific context-shared patterns within data type ℓ across K_ℓ contexts. In detail, we could extract the common latent features, $\boldsymbol{U}_{\ell O}$'s, shared among two or more omics data types by performing a CCA or a generalized canonical correlation analysis (GCCA) on the logit-transformed latent probability matrices using the R package *RGCCA* [Tenenhaus and Tenenhaus, 2011]. The number of retained components p_ℓ is determined using parallel analysis [Buja and Eyuboglu, 1992, Franklin et al., 1995]. Additionally, X-ING further models the major sharing patterns, $\boldsymbol{U}_{\ell C}$, across tissues/contexts within each omics data type by performing PCA on each residual matrix after subtracting the $\boldsymbol{U}_{\ell O}$ matrix. The number of retained principal components is also determined using parallel analysis. The modeling of omics-shared and data-specific patterns may not be uniquely identifiable and they do not need to be. That is, $\boldsymbol{U}_{\ell O}$ and $\boldsymbol{U}_{\ell C}$ do not need to be orthogonal, and their sum $\boldsymbol{U}_\ell$ could still be a good approximation of the logit-transformed probabilities of latent indicators. X-ING performs CCA and PCA sequentially within each iteration and simultaneously accounts for omics-shared and omics-specific context-shared association patterns across data types and contexts. It outputs the posterior mean and probability of non-zero for each input statistic. Additionally, it provides the eigenvectors from PCA and the canonical coefficients from CCA at the final iteration, and these outputs may facilitate the interpretations of the major patterns in the data.

2.3.4 Algorithms for X-ING

In this section, we present the details about the variational EM algorithms for the starting model (2.4) and the X-ING model (2.6) in the main text.

An EM algorithm with variational inference for the starting model

The starting model (2.4) does not model shared data patterns, and the prior for $\boldsymbol{\gamma}_{\cdot j,\ell}$ is the same for all tested units in the j-th tissue from data ℓ. Maximizing the complete-data likelihood with respect to all data types is equivalent to maximizing complete-data likelihood for each data type ℓ separately. For each data type ℓ, we derive a computationally efficient EM algorithm with variational inference.

Let $q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ be an approximation of the posterior $p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})$. The marginal likelihood can be decomposed as

$$\begin{aligned}
\log p(\boldsymbol{z}_\ell; \Theta_1^{(t)}) &= \mathcal{L}_{q,\ell}^{(t)} + \mathrm{KL}\left(q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) || p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})\right) \geq \ \mathcal{L}_{q,\ell}^{(t)}, \\
\mathcal{L}_{q,\ell}^{(t)} &= \mathrm{E}_q \log\left(\frac{p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell; \Theta_1^{(t)})}{q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)}\right), \\
\mathrm{KL}\left(q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) || p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})\right) &= \mathrm{E}_q \log\left(\frac{q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)}{p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})}\right),
\end{aligned} \tag{2.7}$$

where the superscript indicates the estimates being from the t-th step, $\mathcal{L}_{q,\ell}^{(t)}$ is the evidence lower bound (ELBO), and the inequality holds due to Jensen's inequality, i.e., $\mathrm{KL} \geq 0$ with equality holds if and only if $q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ is identical to $p(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell | \boldsymbol{z}_\ell; \Theta_1^{(t)})$ almost surely. To overcome the computational intractability of the ELBO, we use a mean-field variational family. Then $q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ can be factorized as

$$q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) = \prod_{i=1}^{M} \prod_{j=1}^{K_\ell} q(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell}). \tag{2.8}$$

Given mean-field variational family of distributions (2.8), the optimal variational distribution $q^*(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell})$ maximizing the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ has the following form:

$$\log q^*(\tilde{z}_{ij,\ell}, \gamma_{ij,\ell}) = \mathrm{E}_{q(i',j')\neq(i,j)} \log p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) + \text{const}, \tag{2.9}$$

where the expectation is taken with respect to variational q distributions related to all other latent variables $(i', j') \neq (i, j)$. Denote the inverse of context-context covariance matrix as $\mathbf{R}_\ell^{-1} = \boldsymbol{\Lambda}_\ell = \{\lambda_{ij,\ell}\}$.

Based on Equation (2.9), we further separate out (i, j) terms and get

$$\begin{aligned}
\log p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell; \Theta_1^{(t)}) =& -\frac{1}{2}\lambda_{jj,\ell}\left(z_{ij,\ell} - \gamma_{ij,\ell}\widetilde{z}_{ij,\ell}\right)\left(z_{ij,\ell} - \gamma_{ij,\ell}\widetilde{z}_{ij,\ell}\right) \\
& -\frac{1}{2}\sum_{j'\neq j} 2\lambda_{jj',\ell}\left(z_{ij,\ell} - \gamma_{ij,\ell}\widetilde{z}_{ij,\ell}\right)\left(z_{ij',\ell} - \gamma_{ij',\ell}\widetilde{z}_{ij',\ell}\right) \\
& -\frac{1}{2}\sum_{(i',j')\neq(i,j),\ (i',j'')\neq(i,j)} \lambda_{j'j'',\ell}\left(z_{i'j',\ell} - \gamma_{i'j',\ell}\widetilde{z}_{i'j',\ell}\right)\left(z_{i'j'',\ell} - \gamma_{i'j'',\ell}\widetilde{z}_{i'j'',\ell}\right) \\
& -\frac{\widetilde{z}_{ij,\ell}^2}{2\sigma_{j,\ell}^2} - \sum_{(i',j')\neq(i,j)} \frac{\widetilde{z}_{i'j',\ell}^2}{2\sigma_{j',\ell}^2} \\
& +\gamma_{ij,\ell}\log\pi_{j,\ell} + \left(1-\gamma_{ij,\ell}\right)\log\left(1-\pi_{j,\ell}\right) \\
& +\sum_{(i',j')\neq(i,j)} \left\{\gamma_{i'j',\ell}\log\pi_{j',\ell} + \left(1-\gamma_{i'j',\ell}\right)\log\left(1-\pi_{j',\ell}\right)\right\} \\
& +\frac{M}{2}\log|\boldsymbol{\Lambda}| - \frac{M}{2}\sum_j \log\sigma_{j,\ell}^2 + \text{const.}
\end{aligned}$$

To calculate the variational expectation, we retain terms with $\widetilde{z}_{ij,\ell}$:

$$
\begin{aligned}
&\log q^*\left(\widetilde{z}_{ij,\ell} \mid \gamma_{ij,\ell}=1\right) \\
&= \mathrm{E}_{q(i',j')\neq(i,j)}\left[-\frac{1}{2}\left\{\lambda_{jj,\ell}\widetilde{z}^2_{ij,\ell} - 2\lambda_{jj,\ell}z_{ij,\ell}\widetilde{z}_{ij,\ell} + \sum_{j'\neq j} -2\left(\lambda_{jj',\ell}\widetilde{z}_{ij,\ell}\left(z_{ij',\ell} - \gamma_{ij',\ell}\widetilde{z}_{ij',\ell}\right)\right)\right\} - \frac{\widetilde{z}^2_{ij,\ell}}{2\sigma^2_{j,\ell}}\right] \\
&\quad + \text{const} \\
&= \left[\sum_{j'\neq j}\lambda_{jj',\ell}\left(z_{ij',\ell} - \mathrm{E}_q\left(\gamma_{ij',\ell}\widetilde{z}_{ij',\ell}\right)\right) + \lambda_{jj,\ell}z_{ij,\ell}\right]\widetilde{z}_{ij,\ell} - \frac{1}{2}\left(\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}\right)\widetilde{z}^2_{ij,\ell} + \text{ const} \\
&= \left[\sum_{j'=1}^{K_\ell}\lambda_{jj',\ell}z_{ij',\ell} - \sum_{j'\neq j}\lambda_{jj',\ell}\mathrm{E}_q\left(\gamma_{ij',\ell}\widetilde{z}_{ij',\ell}\right)\right]\widetilde{z}_{ij,\ell} - \frac{1}{2}\left(\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}\right)\widetilde{z}^2_{ij,\ell} + \text{ const }\text{,}
\end{aligned}
$$

from which we can see that the posterior of $q^*\left(\widetilde{z}_{ij,\ell} \mid \gamma_{ij,\ell}=1\right) \sim \mathcal{N}\left(\mu_{ij,\ell}, s^2_{ij,\ell}\right)$, where,

$$
\begin{aligned}
s^2_{ij,\ell} &= \frac{1}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}, \\
\mu_{ij,\ell} &= \frac{\sum_{j'=1}^{K_\ell}\lambda_{jj',\ell}z_{ij',\ell} - \sum_{j'\neq j}\lambda_{jj',\ell}\mathrm{E}_q(\gamma_{ij',\ell}\widetilde{z}_{ij',\ell})}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}.
\end{aligned}
$$

Similarly,

$$
\log q^*\left(\widetilde{z}_{ij,\ell} \mid \gamma_{ij,\ell}=0\right) = -\frac{\widetilde{z}^2_{ij,\ell}}{2\sigma^2_{j,\ell}} + \text{const.}
$$

therefore $q^*\left(\widetilde{z}_{ij,\ell} \mid \gamma_{ij,\ell}=0\right) \sim \mathcal{N}\left(0, \sigma^2_{j,\ell}\right)$. Let $\alpha_{ij,\ell} = q(\gamma_{ij,\ell}=1)$, the variational expec-

tation can be written as

$$\begin{aligned}\mathrm{E}_q \log p(\boldsymbol{z}_\ell, \widetilde{\boldsymbol{z}}_{\boldsymbol{\ell}}, \boldsymbol{\gamma}_\ell; \Theta_1^{(t)}) =& -\frac{1}{2}\sum_i \sum_j \sum_{j'} \lambda_{jj',\ell} \mathrm{E}_q \left[\left(z_{ij,\ell} - \gamma_{ij,\ell}\widetilde{z}_{ij,\ell}\right) \left(z_{ij',\ell} - \gamma_{ij',\ell}\widetilde{z}_{ij',\ell}\right) \right] \\ & - \sum_i \sum_j \frac{\mathrm{E}_q\left(\widetilde{z}_{ij,\ell}^2\right)}{2\sigma_{j,\ell}^2} \\ & + \sum_i \sum_j \left[\mathrm{E}_q\left(\gamma_{ij,\ell}\right) \log \pi_{ij,\ell} + \left(1 - \mathrm{E}_q\left(\gamma_{ij,\ell}\right)\right) \log\left(1 - \pi_{ij,\ell}\right) \right] \\ & - \frac{M}{2} \sum_j \log \sigma_{j,\ell}^2 + \text{const},\end{aligned}$$

where

$$\begin{aligned}\mathrm{E}_q\left(\gamma_{ij,\ell}\widetilde{z}_{ij,\ell}\right) &= \alpha_{ij,\ell}\mu_{ij,\ell} \\ \mathrm{E}_q\left(\gamma_{ij,\ell}\gamma_{ij',\ell}\widetilde{z}_{ij,\ell}\widetilde{z}_{ij',\ell}\right) &= \alpha_{ij,\ell}\alpha_{ij',\ell}\mu_{ij,\ell}\mu_{ij',\ell}, \forall j \neq j' \\ \mathrm{E}_q\left(\gamma_{ij,\ell}^2\widetilde{z}_{ij,\ell}^2\right) &= \alpha_{ij,\ell}\left(\mu_{ij,\ell}^2 + s_{ij,\ell}^2\right) \\ \mathrm{E}_q\left(\widetilde{z}_{ij,\ell}^2\right) &= \alpha_{ij,\ell}\left(\mu_{ij,\ell}^2 + s_{ij,\ell}^2\right) + \left(1 - \alpha_{ij,\ell}\right)\sigma_{j,\ell}^2.\end{aligned}$$

Similarly, we have

$$\begin{aligned}\mathrm{E}_q \log q(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell) =& \sum_i \sum_j \left[\alpha_{ij,\ell} \log \alpha_{ij,\ell} + \left(1 - \alpha_{ij,\ell}\right) \log\left(1 - \alpha_{ij,\ell}\right) - \frac{1}{2}\alpha_{ij,\ell} \log \frac{s_{ij,\ell}^2}{\sigma_{j,\ell}^2} \right] \\ & - \frac{M}{2} \sum_j \log \sigma_{j,\ell}^2 + \text{const}.\end{aligned}$$

Then we can obtain $\mathcal{L}_{q,\ell}^{(t)}$ via (2.7). By taking derivative on $\mathcal{L}_{q,\ell}^{(t)}$ with respect to $\alpha_{ij,\ell}$ and equating the derivative to zero, we update $\alpha_{ij,\ell}$ with

$$\alpha_{ij,\ell} = \frac{1}{1 + \exp(-v_{ij,\ell})}, v_{ij,\ell} = \log \frac{\pi_{j,\ell}}{1 - \pi_{j,\ell}} + \frac{1}{2}\left(\log \frac{s_{ij,\ell}^2}{\sigma_{j,\ell}^2} + \frac{\mu_{ij,\ell}^2}{s_{ij,\ell}^2} \right).$$

The optimal q distribution can be written as

$$q^*(\widetilde{z}_{ij,\ell}, \gamma_{ij,\ell}) = \alpha_{ij,\ell}^{\gamma_{ij,\ell}}(1-\alpha_{ij,\ell})^{1-\gamma_{ij,\ell}}\mathcal{N}(\mu_{ij,\ell}, s_{ij,\ell}^2)^{\gamma_{ij,\ell}}\mathcal{N}(0, \sigma_{j,\ell}^2)^{1-\gamma_{ij,\ell}}, \tag{2.10}$$

where $\mu_{ij,\ell}$, and $s_{ij,\ell}^2$ are variational parameters defined as follows

$$\mu_{ij,\ell} = \frac{\sum_{j'=1}^{K_\ell} \lambda_{jj',\ell} z_{ij',\ell} - \sum_{j' \neq j} \lambda_{jj',\ell}\alpha_{ij',\ell}\mu_{ij',\ell}}{\lambda_{jj,\ell} + \frac{1}{\sigma_{j,\ell}^2}},\ s_{ij,\ell}^2 = \frac{1}{\lambda_{jj,\ell} + \frac{1}{\sigma_{j,\ell}^2}}. \tag{2.11}$$

In this way, $q^*(\widetilde{\boldsymbol{z}}_\ell, \boldsymbol{\gamma}_\ell)$ can be obtained analytically and the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ can be evaluated under the variational distribution q^*. This step can be viewed as a generalized E-step within the variational family of distributions (2.8).

In the M-step, by taking partial derivatives of the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ with respect to the model parameters $\Theta_1^{(t)}$ and setting them to zero, we obtain

$$\begin{aligned} \sigma_{j,\ell}^2 &= \frac{\sum_{i=1}^M \alpha_{ij,\ell}(s_{ij,\ell}^2 + \mu_{ij,\ell}^2)}{\sum_{i=1}^M \alpha_{ij,\ell}}, \\ \pi_{j,\ell} &= \frac{\sum_{i=1}^M \alpha_{ij,\ell}}{M}. \end{aligned} \tag{2.12}$$

An EM algorithm with variational inference for X-ING

To account for the major omics-shared and context-shared patterns in the estimation of association probabilities, we estimate the latent low-rank matrix $\boldsymbol{U}_\ell$ for each data type ℓ that modulates the prior probability of the latent status as function (2.2) in the main text. Denote $\Theta_2 = \{\boldsymbol{U}_\ell, \boldsymbol{u}_{0,\ell}, \mathbf{R}_\ell, \sigma_{j,\ell}^2, \pi_{ij,\ell}, \ell = 1, \ldots, L, i = 1, \ldots, M, j = 1, \ldots, K_\ell\}$ as the parameter space for the X-ING model (2.6). Similar to the starting model (2.4), $\mathbf{R}_\ell$ can be pre-estimated and taken as known. To reduce computational complexity, we pre-estimate the intercepts $\boldsymbol{u}_{0,\ell}$'s using the estimated $\pi_{j,\ell}$'s from function (2.12) in the starting model (2.4) with $u_{0j,\ell} = \log\big(\pi_{j,\ell}/(1-\pi_{j,\ell})\big)$.

By taking partial derivatives of the ELBO $\mathcal{L}_{q,\ell}^{(t)}$ with respect to $\pi_{ij,\ell}$'s and setting them to zero, we have $\pi_{ij,\ell} = \alpha_{ij,\ell}$. Then, we have modulation matrices

$$U_{ij,\ell}^{*} = \log\left(\frac{\pi_{ij,\ell}}{1-\pi_{ij,\ell}}\right) - u_{0j,\ell}. \tag{2.13}$$

If no constraints are imposed, there would be an issue of over-parameterization for $\boldsymbol{U}_{\ell}^{*}$'s. Here in the M-step, we apply canonical correlation analysis (CCA) or generalized canonical correlation analysis (GCCA) to the standardized $\boldsymbol{U}_{\ell}^{*}$'s estimated as in formula (2.13), where $\boldsymbol{U}_{\ell}^{*}$'s are standardized with mean zero and unit variance. When $L = 2$, GCCA reduces to the problem of CCA between two latent matrices $\boldsymbol{U}_{1}^{*}$ and $\boldsymbol{U}_{2}^{*}$. CCA/GCCA aims to maximize the pair-wise correlation between linear combinations of $\boldsymbol{U}_{\ell}^{*}$'s. Suppose we have rank p_ℓ ($\leq K_\ell, \forall \ell \in \{1, \ldots, L\}$) approximation for $\boldsymbol{U}_{\ell}^{*}$, the corresponding canonical weight matrices $\boldsymbol{A}_\ell = [\boldsymbol{a}_\ell^1, \ldots, \boldsymbol{a}_\ell^{p_\ell}]$ can be estimated. We choose the number of retained components p_ℓ using parallel analysis (PA) [Buja and Eyuboglu, 1992, Franklin et al., 1995]. Then, for each data type ℓ, the estimated low-rank matrix $\boldsymbol{U}_{\ell O} = \boldsymbol{U}_{\ell}^{*}\boldsymbol{A}_\ell\boldsymbol{A}_\ell^{\dagger}$ has $\text{rank}(\boldsymbol{U}_{\ell O}) = p_\ell$, where † refers to the Moore-Penrose pseudo-inverse. When the dimension of multivariate cellular contexts K_ℓ in each omic data is large, one may also use the regularized GCCA (RGCCA). We use R packages *CCA* and *RGCCA* [Tenenhaus and Tenenhaus, 2011] to perform CCA/GCCA/RGCCA analyses.

To further capture omic-specific patterns for each data type, we perform a PCA on the residual matrix from CCA/GCCA/RGCCA calculated as $\boldsymbol{U}_{\text{res},\ell} = \boldsymbol{U}_{\ell}^{*} - \boldsymbol{U}_{\ell O}$ and get the low-rank approximation matrix $\boldsymbol{U}_{\ell C}$. Specifically, we first standardize the $\boldsymbol{U}_{\ell C}$'s with mean zero and unit variance. We calculate a truncated singular value decomposition (SVD) and keep the top q_ℓ ($\leq K_\ell, \forall \ell \in \{1, \ldots, L\}$) largest singular values to approximate $\boldsymbol{U}_{\ell C}$. We choose the number of components q_ℓ using PA [Buja and Eyuboglu, 1992, Franklin et al., 1995]. Those approximated matrices $\boldsymbol{U}_{\ell C}$'s capture the association patterns shared among

and specific to different cellular contexts (tissues).

We take $\boldsymbol{U}_\ell = \boldsymbol{U}_{\ell O} + \boldsymbol{U}_{\ell C}$ as an estimated modulation matrix capturing both omics-shared and omics-specific context-shared patterns. We update the prior specification $\pi_{ij,\ell}$ as function (2.2). The algorithm for estimating the X-ING model is given in Algorithm 1.

Algorithm 1 An EM algorithm with variational inference for the X-ING method

Input data: for $\ell = 1, \ldots, L$, $\boldsymbol{z}_\ell \in \mathbb{R}^{M \times K_\ell}$, $\mathbf{R}_\ell \in \mathbb{R}^{K_\ell \times K_\ell}$, $p_\ell \in \mathbb{Z}^+$, $q_\ell \in \mathbb{Z}^+$

Initialize parameters: $\alpha_{ij,\ell}, \mu_{ij,\ell}, s^2_{ij,\ell}, \sigma^2_{j,\ell}, U_{ij,\ell}$, $\pi_{ij,\ell} = \alpha_{ij,\ell}$, and specify $u_{0j,\ell}$, for $\ell = 1, \ldots, L$, $j = 1, \ldots, M$, $j = 1, \ldots, K_\ell$. This can be either user-specified or obtained by running the EM algorithm for the starting model (2.4) (by skipping Step 14-20 and setting $\boldsymbol{U}_{\ell O} + \boldsymbol{U}_{\ell C}$ as $\mathbf{0}$ in Step 24-25).

Initialize $\mathcal{L}^{(1)}_{q,\ell} = -\infty$

repeat $t = 2, 3, \ldots$

  E-step:

  for $\ell = 1, \ldots, L$ **do**

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

      for $j = 1, \ldots, K_\ell$ **do**

        $\mu_{ij,\ell} = \frac{\sum_{j'=1}^{K_\ell} \lambda_{jj',\ell} z_{ij',\ell} - \sum_{j' \neq j} \lambda_{jj',\ell} \alpha_{ij',\ell} \mu_{ij',\ell}}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}$,

        $s^2_{ij,\ell} = \frac{1}{\lambda_{jj,\ell} + \frac{1}{\sigma^2_{j,\ell}}}$,

        $v_{ij,\ell} = \log \frac{\pi_{ij,\ell}}{1-\pi_{ij,\ell}} + \frac{1}{2}\Big(\log \frac{s^2_{ij,\ell}}{\sigma^2_{j,\ell}} + \frac{\mu^2_{ij,\ell}}{s^2_{ij,\ell}} \Big)$,

        $\alpha_{ij,\ell} = \frac{1}{1+\exp(-v_{ij,\ell})}$.

  M-step:

  for $\ell = 1, \ldots, L$ **do**

    $\boldsymbol{U}^*_\ell = \Big\{ \log \frac{\alpha_{ij,\ell}}{1-\alpha_{ij,\ell}} - u_{0j,\ell}, 1 \leq i \leq M, 1 \leq j \leq K_\ell \Big\}$,

  Perform a CCA/GCCA/RGCCA on the L standardized $\boldsymbol{U}^*_\ell$'s. Note that CCA applies when $L = 2$, GCCA applies when $L > 2$, and RGCCA is recommended when K_ℓ is large.

  for $\ell = 1, \ldots, L$ **do**

    Get the coefficient matrices $\boldsymbol{A}_\ell$'s. Using the top p_ℓ canonical coefficients to get $\boldsymbol{U}_{\ell O}$: $\boldsymbol{U}_{\ell O} = \boldsymbol{U}^*_\ell \boldsymbol{A}_\ell \boldsymbol{A}^\dagger_\ell$.

    Calculate the residual matrix: $\boldsymbol{U}_{\text{res},\ell} = \boldsymbol{U}^*_\ell - \boldsymbol{U}_{\ell O}$.

    Perform a PCA on each $\boldsymbol{U}_{\text{res},\ell}$ using SVD and keep the top q_ℓ singular values to get the low-rank approximation matrix $\boldsymbol{U}_{\ell C}$ for each omics data type.

  for $\ell = 1, \ldots, L$ **do**

    for $j = 1, \ldots, K_\ell$ **do**

      $\sigma^2_{j,\ell} = \frac{\sum_{i=1}^M \alpha_{ij,\ell}(s^2_{ij,\ell} + \mu^2_{ij,\ell})}{\sum_{i=1}^M \alpha_{ij,\ell}}$,

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

        $\pi_{ij,\ell} = \frac{1}{1+\exp\left(-U_{ij,\ell O} - U_{ij,\ell C} - u_{0j,\ell}\right)}$.

until $\mathcal{L}^{(t)}_{q,\ell} - \mathcal{L}^{(t-1)}_{q,\ell} < \varepsilon$, where ε is a user determined threshold.

Output: for $\ell = 1, \ldots, L$, posterior probability $\boldsymbol{\alpha}_\ell$, posterior mean $\boldsymbol{\mu}_\ell$.

2.3.5 Data processing

Cis-e/mQTL input association statistics

We obtained the single-tissue cis-mQTL and cis-eQTL association statistics from GTEx portal [Consortium, 2020, Oliva et al., 2023]. Those QTL statistics were obtained using FastQTL [Ongen et al., 2016], adjusting for top five genotypic principal components, biological gender, Sequencing platform (Illumina HiSeq 2000 or HiSeq X), Sequencing protocol (polymerase chain reaction, PCR; PCR-based or PCR-free), and a set of variables generated using the method of probabilistic estimation of expression residuals (PEER) [Stegle et al., 2012]. For cis-mQTL analysis integrating eQTL maps, we included both lead and secondary mQTL variants for each CpG site within a 500KB cis-window size. Each CpG site was assigned to a proximal gene with the nearest TSS [Mendioroz et al., 2020, Grand et al., 2021]. We analyzed 204,220 SNP-CpG-gene trios, consisting of 93,681 unique CpG sites and 159,186 unique mQTL variants.

Trans-e/mQTL input association statistics

In our integrative analysis of trans-e/mQTL associations, we obtained the test statistics for GWAS SNPs associated with at least one out of 80 selected diseases/traits ($P < 5 \times 10^{-8}$). In detail, we selected 80 diseases/traits that have more than 100 risk loci. Those diseases/traits were related to brain function (e.g., Alzheimer's disease), artery tissues (e.g., coronary artery disease), heart function (e.g., atrial fibrillation), or cancers (e.g., prostate cancer). Similar to cis-eQTL and cis-mQTL analyses, we adjusted for the same set of covariates when performing trans-eQTL and trans-mQTL analyses, respectively. We tested trans-association for SNP-gene pairs from different chromosomes [Albert et al., 2018, Wright et al., 2014] and obtained the association statistics for trans-eQTLs and trans-mQTLs in 28 and nine GTEx tissues, respectively. The list of tissues and sample sizes was given in Table 2.3-2.4.

Differential expression analysis of disease-relevant genes in brain layers

In the integrative analysis of spatial transcriptomics data with multi-tissue eQTLs, we used LIBD DLPFC data generated using 10x Visium [Maynard et al., 2021] that contained 12 samples from three adult donors. The original study provided manual annotations for the tissue layers based on the cytoarchitecture. For each of the 12 samples, we performed differential expression analysis using beta-Possion GLM model [Vu et al., 2016] to obtain Z-scores of each gene. We compared the expression profiles of spots in a layer with those from the rest of the spots in other layers for each gene and obtained 12 matrices of differential expression statistics with each row being a gene and each column corresponding to a layer. Additionally, we obtained the summary statistics for cis-eQTLs from 13 GTEx brain tissues. For each sample, the two sets of summary statistics were matched through gene names. We conducted a X-ING analysis for each sample. In each analysis, we integrate the spatially differential expression statistics of the sample with 13 sets of GTEx brain eQTL statistics and obtain the posterior probabilities of having cis-association and spatially differential expression.

More specifically, we analyzed 85,944 GWAS SNPs ($P < 5 \times 10^{-8}$) associated with diseases/traits that have more than 100 risk loci. There were 15,244 cis-genes available in GTEx data for those examined GWAS SNPs [Oliva et al., 2023]. For the 85,944 examined GWAS SNPs, we analyzed a total of 1.6 million SNP-gene pairs matched in 13 GTEx brain tissues and differential expression test statistics for genes among 12 DLPFC samples. In analyzing each of the 12 LIBD samples, we applied X-ING on a 1.6M (SNP-gene pairs) by 7 (layers) matrix of Z-scores, and a 1.6M$\times$13 matrix of Z-scores for 13 GTEx brain tissues. Note that 4 of the 12 samples contained only five manually annotated layers (i.e., a 1.6M$\times$5 matrix of Z-scores). At the 90% posterior probability cutoff, we obtained the genes with brain-layer specific expression levels (≥ 1 nonzero spatial differential statistic) and also being in cis-association with disease risk loci (≥ 1 nonzero brain eQTL statistic) in the examined sample. We studied the concerted association and enrichment patterns across 12 samples to

minimize the potential confounding effects due to unknown sample heterogeneity.

2.4 Simulations

2.4.1 Generation of summary statistics in the simulation studies

We generated L matrices of $M \times K_\ell$ association summary statistics using simulated individual-level data. We first simulated predictor variables $\boldsymbol{X}_\ell$, with M omics-specific predictors and a sample size of N_ℓ. Each element of $\boldsymbol{X}_\ell$ was generated independently from a standard normal distribution. Here, we simulated binary association status, $\gamma_{ij,\ell}$ with a given correlation structure using the R package *bindata*. By simulating effect size, $\beta_{ij,\ell} \sim \mathcal{N}(0, \sigma^2_\ell)$, for each data type ℓ, we then considered the following equation to generate response variables,

$$\boldsymbol{y}_{j,\ell} = \boldsymbol{X}_\ell(\boldsymbol{\beta}_{\cdot j,\ell} \circ \boldsymbol{\gamma}_{\cdot j,\ell}) + \boldsymbol{\epsilon}_{\cdot j,\ell}, \tag{2.14}$$

where $\boldsymbol{\epsilon}_{\cdot j,\ell}$ was the error term following $\mathcal{N}(0, \sigma^2_\epsilon)$. In the simulation studies, we controlled the proportion of variation in the response variable, $\boldsymbol{y}_{j,\ell}$, that can be explained by the predictors, $\theta_{j,\ell} = \frac{\text{var}(\boldsymbol{X}_\ell(\boldsymbol{\beta}_{\cdot j,\ell} \circ \boldsymbol{\gamma}_{\cdot j,\ell}))}{\text{var}(\boldsymbol{y}_{j,\ell})}$, where we assumed the same θ_ℓ for all $\theta_{j,\ell}$'s in data ℓ.

In the simulation of two omics data types, i.e., $L = 2$. Each row of the corresponding association status matrices $\boldsymbol{\gamma}_1 \in \mathbb{R}^{M \times K_1}$ and $\boldsymbol{\gamma}_2 \in \mathbb{R}^{M \times K_2}$ was jointly simulated [Leisch et al., 2006] with the correlation matrix $\boldsymbol{\Omega}$ being:

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{W}_1 & \boldsymbol{C} \\ \boldsymbol{C}^\top & \boldsymbol{W}_2 \end{pmatrix}, \tag{2.15}$$

and probability of being 1 as τ_ℓ ($\ell = 1, 2$). Here $\boldsymbol{W}_\ell \in \mathbb{R}^{K_\ell \times K_\ell}$ ($\ell = 1, 2$) was the within-data correlation structure across contexts for the ℓ-th data type, and $\boldsymbol{C} \in \mathbb{R}^{K_1 \times K_2}$ was the between-data correlation matrix. Within each omics data type ℓ, all K_ℓ contexts can be

partitioned into two groups, with $K_\ell^{(1)}$ and $K_\ell^{(2)}$ contexts, respectively, where $K_\ell^{(1)} + K_\ell^{(2)} = K_\ell$. We also considered nonzero within-group context-context correlations for each group of contexts. The correlation matrices $\boldsymbol{W}_\ell$'s and $\boldsymbol{C}$ were specified as

$$\boldsymbol{W}_\ell = \begin{pmatrix} (1-\rho_\ell)\cdot \boldsymbol{I}_{K_\ell^{(1)}} + \rho_\ell \cdot \mathbf{1}_{K_\ell^{(1)}} \mathbf{1}_{K_\ell^{(1)}}^\top & \mathbf{0} \\ \mathbf{0} & (1-\rho_\ell)\cdot \boldsymbol{I}_{K_\ell^{(2)}} + \rho_\ell \cdot \mathbf{1}_{K_\ell^{(2)}} \mathbf{1}_{K_\ell^{(2)}}^\top \end{pmatrix}, \tag{2.16}$$

$$\boldsymbol{C} = \begin{pmatrix} r \cdot \mathbf{1}_{K_1^{(1)}} \mathbf{1}_{K_2^{(1)}}^\top & \mathbf{0} \\ \mathbf{0} & r \cdot \mathbf{1}_{K_1^{(2)}} \mathbf{1}_{K_2^{(2)}}^\top \end{pmatrix},$$

where $\mathbf{1}_{K_\ell^{(1)}}$ and $\mathbf{1}_{K_\ell^{(2)}}$ were column vectors of ones with length $K_\ell^{(1)}$ and $K_\ell^{(2)}$, respectively, ρ_ℓ was the pair-wise correlation coefficient for any two contexts within the ℓ-th data type, and r was the parameter controlling the strength of cross-omics/between-data correlations among shared/similar contexts.

After simulating the individual-level data, we obtained all Z-scores $\{z_{ij,\ell}\}$'s by performing a simple linear regression for each predictor and its simulated response variable. A similar simulation framework could be used to generate Z-scores for three or more data types ($L \geq 3$).

2.4.2 X-ING improves power by borrowing information across different data types and contexts

We conducted simulation studies to evaluate the performance of X-ING in comparison with existing methods in a variety of scenarios. We first simulated individual-level data across L omics data, with K_ℓ contexts for each data ℓ, under various combinations of between- and within-data correlations that represent variations in data-shared and -specific association patterns on M tested units (see Methods). Each set of data has a sample size of N_ℓ, and

the proportion of variation in the simulated response variables that can be explained by the predictors is θ_ℓ. By performing a simple linear regression on each response-predictor pair, we obtained L matrices of $M \times K_\ell$ association statistics as input for the following analyses.

We compared X-ING with three existing methods: the multivariate adaptive shrinkage (mash) [Urbut et al., 2019], multi-tissue eQTL (MT-eQTL) [Li et al., 2018], and Metasoft [Han and Eskin, 2012]. Some of the existing methods were proposed for different purposes, and those methods can be adapted to the integrative analysis of summary data from multiple contexts for comparison purposes. We also included two variations of the X-ING models, X-ING$_{\text{starting}}$ and X-ING$_{\text{single-omics}}$, for comparison and illustration purposes. The X-ING$_{\text{starting}}$ is a starting model without the modeling of major data-shared patterns. It imposed the same prior for $\boldsymbol{\gamma}_{\cdot j,\ell}$ for all tested units in the j-th tissue from data ℓ, i.e., no borrowing information across context and omics. The X-ING$_{\text{single-omics}}$ applies to only one omics data type. It considers shared patterns across contexts but does not allow the joint modeling of two or more omics data of different effect size distributions.

We first evaluated the selection performance (the predicted presence of nonzero effects) using the area under the curve (AUC) of the receiver operating characteristic (ROC) for all methods. We varied the sample size, within-data across-context correlation, between-data correlation, and proportion of tested units that do not have context-shared information. In all scenarios, X-ING outperformed other methods in terms of AUC. Not surprisingly, with increasing sample sizes, the AUC of all six methods increased (Figure 2.2a). All methods except for X-ING$_{\text{starting}}$ (no borrowing information) could gradually gain power as within-data correlation in Data 1 increased (Figure 2.2b), suggesting multivariate integrative methods could improve power by borrowing information across contexts. The performance of the single-omics variation, X-ING$_{\text{single-omics}}$, was similar to that of the existing method for single-omics data, mash [Urbut et al., 2019]. Mash and X-ING$_{\text{single-omics}}$ both allow only cross-context/tissue integration but no borrowing information across data types. The former

models effect sizes and the latter models the latent association status, and their performances are similar. When considering two omics data types, Figure 2.2c showed that only X-ING could gain power as the between-data correlation increases. This is because X-ING borrows information across multi-omics data types. Allowing for multi-omics data integration is a major innovation of X-ING compared with other methods. When the proportion of tested units without context-shared information increased, Figure 2.2d showed that X-ING had higher AUC than other methods, and the AUC of mash and Metasoft reduced substantially.

Since many disease/trait-relevant effects are context-specific and cell-type-specific [Umans et al., 2021], we also evaluated the selection performance of detecting associations that were shared across two to five contexts (Table 2.1, see Methods). X-ING achieved the highest overall and context-specific AUC compared to other methods. In the simulations of sparse associations (i.e., when there are many zero effects in the data), X-ING outperformed competing methods (Figure 2.3). When within- and between-data correlations were both zeros, all methods achieved similar performance as there was no shared information across contexts or data types (Figure 2.4). Figure 2.5 showed that by increasing the number of contexts, X-ING gained improved AUC due to more shared information across contexts. Moreover, X-ING performed better in data with varying percentages of variance explained by the predictors (Figure 2.6). We then evaluated the estimation of effect sizes using the root-mean-square error (RMSE) for both X-ING and mash, since only these two methods provided the posterior mean estimates. As sample size increases, both X-ING and mash gained improved RMSEs. The RMSE of X-ING was smaller than that of mash for data with different sample sizes (Figure 2.7).

Figure 2.9 showed that X-ING was robust to weak correlations among predictors for tested units (i.e., when tested units have moderately dependent effects). We also showed that the performance of X-ING is robust to the choices of retained CCs and PCs in the CCA

Method	Overall AUC			AUC on associations true in two to five contexts		
	$N_1 = 400$	$N_2 = 800$	$N_3 = 1200$	$N_1 = 400$	$N_2 = 800$	$N_3 = 1200$
X-ING	0.823	0.881	0.904	0.856	0.914	0.933
X-ING$_{\text{single-omics}}$	0.797	0.850	0.873	0.819	0.871	0.891
mash	0.787	0.827	0.844	0.792	0.814	0.821
MT-eQTL	0.760	0.816	0.851	0.755	0.808	0.843
X-ING$_{\text{starting}}$	0.769	0.829	0.856	0.769	0.827	0.854
Metasoft	0.702	0.772	0.802	0.745	0.833	0.864

Table 2.1: Comparison of methods on simulated data. We evaluate the AUC of X-ING and competing methods in the analysis of three sets of statistics ($L = 3$). The overall AUC is calculated based on the true association status for all tested effects. We also compare the AUC for context-specific effects that have true nonzero effects in two to five cellular contexts/tissues.

and PCA analysis within a certain range (Figure 2.8). When setting $\boldsymbol{U}_\ell = \mathbf{0}$, the algorithm does not borrow information across data types nor tissue types. When setting p_ℓ or q_ℓ to full ranks, the model is over-parameterized. The low-rank approximation is useful in capturing major patterns in the data and borrowing information across omics data types and contexts.

2.4.3 Sensitivity Analysis

We performed sensitivity analyses to evaluate the robustness of X-ING. We simulated data with varying levels of pairwise correlation for SNPs. We simulated correlated SNPs with a block-diagonal LD matrix. For each gene, we simulated 10 cis-SNPs for each block, with a total of 50 blocks. Within the same block, the pairwise correlation among SNPs varied from 0 to 0.4. This simulation mimics a real data analysis with input statistics being effects for candidate QTLs. The estimation of X-ING model assumes the examined SNPs being independent. The simulation results show that weak correlation among SNPs did not substantially hurt the performance of X-ING (Figure 2.9). In practice, we recommend conducting LD pruning on the input data before applying X-ING. An r^2 threshold of 0.1 is recommended. X-ING was robust to weak correlations among predictors for tested units

(i.e., when tested units have moderately dependent effects).

We then evaluated the choice of the numbers of CCs and PCs retained for low-rank approximation in the X-ING algorithm (Figure 2.8). Simulations were conducted with $K_1 = 40$, $K_2 = 40$, and the proportion of variance in the response variable that can be explained by predictors (θ_ℓ) varying from 0.1 to 0.2. When both p_ℓ and q_ℓ were set to zero, i.e., $\boldsymbol{U}_\ell = \mathbf{0}$, the model reduced to the starting model with fixed context-specific priors and the AUCs were substantially lower compared with those from the suggested #CC and #PC. When setting $\boldsymbol{U}_\ell = \mathbf{0}$, the algorithm does not borrow information across data types nor tissue types. When setting p_ℓ or q_ℓ to full ranks with no constraint imposed on $\boldsymbol{U}_\ell$, the model is over-parameterized. In Figure 2.8, the AUCs are similar within a range near the suggested #CC and #PC by PA, while it starts to decrease when choosing larger numbers than suggested #CC and #PC, reflecting the impact of over-parameterization. Those results suggest that the low-rank approximation is useful in capturing major patterns in the data and borrowing information across omics data types and contexts.

2.5 Data applications

2.5.1 A multi-tissue cis-mQTL analysis integrating eQTL maps

The eGTEx project [eGTEx Project., 2017] generates DNA methylome data on subsets of GTEx samples from nine tissue types to study the genetic regulation of DNAm and expression across human tissues. Due to the limited DNAm tissue sample sizes, the detection of mQTLs is underpowered compared with existing eQTL maps. Moreover, a large majority of mQTLs lack functionality. To examine the genetic association patterns on DNAm together with expression variation in a tissue-specific manner while improving the functionality of the detected mQTLs, we applied X-ING to the cis-mQTL association statistics integrating eQTL maps ($L = 2$) generated on nine tissues representing $N = 367$ and 829 samples, respectively,

from the GTEx project (v8) [Consortium, 2020, Oliva et al., 2023]. The list of tissues and tissue-specific sample sizes for mQTLs and eQTLs are provided in Table 2.3-2.4.

We obtained the sets of mQTLs from single-tissue mQTL analysis of eGTEx [Oliva et al., 2023] including both lead and secondary mQTL variants for each CpG site within a 500 KB cis-window size. We included 93,681 CpG sites and 159,186 unique mQTL variants, forming 204,220 unique SNP-CpG pairs. Each CpG site was assigned to a proximal gene with the nearest transcription start site (TSS) [Mendioroz et al., 2020, Grand et al., 2021], forming 204,220 SNP-CpG-gene trios ($M = 204,220$). We applied X-ING to the cis-eQTL and mQTL association statistics generated on 28 and nine tissues ($K_1 = 28$, $K_2 = 9$). In Figure 2.17, we showed that the major patterns/eigenvectors captured by X-ING can be interpreted as the surrogate variables for tissue-tissue dependence due to similar cell type compositions.

At the 80% posterior probability cutoff (FDR [Storey and Tibshirani, 2003, Chen et al., 2007]=0.031), among the 204,220 analyzed SNP-CpG pairs, we identified a total of 143,801 pairs with nonzero mQTL effects in at least two tissues. Among those 143,801 SNP-CpG pairs (corresponding to 112,162 unique mQTL variants), 79,454 (58,158 unique variants) also exhibited a nonzero association effect to its cis-gene expression in at least one tissue. In other words, more than half of the reported mQTLs also have nonzero associations to their cis-genes, suggesting joint genetic associations to both cis-DNAm and gene expression.

2.5.2 Trans-association enrichment informs disease/trait-relevant tissues

To examine the trans-association enrichment patterns among diseases and traits, we conducted an integrative analysis using multi-tissue inter-chromosomal trans-e/mQTL association statistics ($L = 2$). We first selected 80 diseases/traits from a total of 216 diseases/traits and restricted the analysis to 40,466 GWAS SNPs associated ($P < 5 \times 10^{-8}$) with at least one of those 80 diseases/traits. Those SNPs were generally in weak linkage disequilibrium (LD) (Figure 2.18). We calculated the trans-eQTL association statistics in 28 GTEx tissues

($N \geq 73$; Table 2.3-2.4) and trans-mQTLs statistics in nine GTEx tissues (Table 2.3) using FastQTL [Ongen et al., 2016]. For each of the 80 selected diseases/traits, we applied X-ING to integrate the multi-tissue trans-association statistics for expression ($K_1 = 28$) and DNAm ($K_2 = 9$). At the 80% posterior probability cutoff, there were 644 to 15,490 SNP-gene-CpG site trios out of the examined trios for each selected disease/trait having nonzero trans-expression associations in at least one out of the 28 examined eQTL tissues, or having nonzero trans-methylation associations in at least one out of the nine examined mQTL tissues. We further identified trans-QTL hotspots with nonzero trans-effects on at least five genes/CpG sites. Further analysis showed that disease-associated hotspot regions explained more phenotypic variation compared with trait-associated ones (Figure 2.12).

We examined the enrichment of trans-expression associations across 28 tissue types by evaluating the scaled proportion of SNP-trans-gene pairs with trans-effects. Figure 2.10a showed the heatmap of the scaled proportion of SNP-trans-gene pairs identified with trans-associations for disease/trait-associated SNPs among the 28 tissues. As a comparison, Figure 2.10b showed the corresponding heatmap of the scaled proportion for cis-expression associations. We observed a much stronger enrichment of trans-associations in many known disease/trait-relevant tissue types. For example, we identified the brain amygdala and prostate tissues as being enriched with trans-associations for Alzheimer's disease and prostate cancer, respectively [Prieto del Val et al., 2016, Thibodeau et al., 2015]. The strong enrichment for trans-associations in many disease/trait-relevant tissues and the complementary patterns to cis highlight the potential of leveraging trans-expression associations together with cis in further identifying relevant tissues and cell types. It is consistent with the higher enrichment for heritability in regions surrounding genes with highly tissue-specific expression in disease-relevant tissue [Finucane et al., 2018].

2.5.3 Replication of cis- and trans-associations identified by X-ING

We evaluated the replication rates of SNP-CpG pairs with nonzero effects identified by X-ING. 119,401 out of 204,220 analyzed lead/secondary SNP-CpG pairs were also included in the replication data from FUSION (Finland-United States Investigation of NIDDM Genetics) skeletal muscle study [Taylor et al., 2019]. Among those 119,401 SNP-CpG pairs, 84,255 have nonzero effects in at least two tissues (posterior probability>0.8) in GTEx data. At the P-value threshold of 6×10^{-7} (with Bonferroni correction) [McRae et al., 2018, Min et al., 2021, Qi et al., 2018], 45.04% (53,780 out of 119,401) of the input SNP-CpG pairs were replicated in the FUSION data (without applying X-ING). In contrast, 55.79% (47,010 out of 84,255) of the SNP-CpG pairs identified by X-ING with multi-tissue effects (in two or more tissues) were replicated in FUSION. Moreover, we further categorized the examined mQTLs as 1) single-tissue mQTLs only, 2) single-tissue mQTLs with co-occurring expression associations, 3) multi-tissue mQTLs only, and 4) multi-tissue mQTLs with co-occurring expression associations. Table 2.2a shows the replication rates by tissue type for those four types of mQTLs. Not surprisingly, multi-tissue mQTLs are more likely to be replicated. It is worth noting that mQTLs with co-occurring expression associations have much higher replication rates than those without. Similar replication results were observed in GoDMC data (Table 2.2b). The replication results demonstrate that by integrating multi-omics association studies and borrowing information across data types, X-ING improves the detection, replication, and functional interpretation of mQTLs.

We also evaluated the replication rates for trans-e/m associations identified by X-ING among the nine tissues with both expression and DNAm data, using eQTLGen [Võsa et al., 2021] and GoDMC [Min et al., 2021], respectively, as the replication studies. Among the 282 trans-expression associations identified by X-ING in the whole blood tissue from GTEx, 56 of them (19.9%) were replicated at the 5% FDR level in the whole-blood-sample-based eQTLGen study. In contrast, the proportion of significant trans-expression associations

Tissue	Total # identified SNP-CpG pairs (PP> 0.8)	Type of effects			
		Single-tissue, no expression association	Single-tissue, with expression association	Multi-tissue, no expression association	Multi-tissue, with expression association
Breast Mammary Tissue	13,870	0.08 (18/233)	0.38 (18/47)	0.74 (3,259/4,399)	0.83 (7,619/9,191)
Colon Transverse	83,803	0.17 (1,013/6,047)	0.26 (955/3,710)	0.53 (16,532/31,140)	0.63 (26,949/42,906)
Kidney Cortex	13,690	0.39 (11/28)	0.40 (6/15)	0.64 (3,489/5,445)	0.76 (6,251/8,202)
Lung	74,282	0.09 (427/4,819)	0.19 (464/2,506)	0.54 (14,957/27,733)	0.63 (24,746/39,224)
Muscle Skeletal	11,313	0.44 (24/54)	0.75 (24/32)	0.82 (3,193/3,881)	0.91 (6,672/7,346)
Ovary	61,840	0.14 (922/6,431)	0.24 (1,067/4,047)	0.30 (11,658/21,215)	0.40 (19,660/30,147)
Prostate	51,125	0.12 (258/1,900)	0.27 (273/906)	0.32 (11,386/19,172)	0.42 (20,081/29,147)
Testis	9,631	0.11 (47/501)	0.16 (28/179)	0.33 (2,182/3,272)	0.41 (4,102/5,679)
Whole Blood	29,719	0.08 (34/442)	0.30 (33/114)	0.34 (6,327/10,869)	0.48 (12,638/18,294)

(a)

Tissue	Total # identified SNP-CpG pairs (PP> 0.8)	Type of effects			
		Single-tissue, no expression association	Single-tissue, with expression association	Multi-tissue, no expression association	Multi-tissue, with expression association
Breast Mammary Tissue	24,583	0.07 (31/431)	0.26 (20/78)	0.31 (2,474/8,026)	0.45 (7,259/16,048)
Colon Transverse	143,566	0.18 (1,993/10,792)	0.31 (1,836/5,960)	0.31 (17,060/55,019)	0.41 (29,614/71,795)
Kidney Cortex	24,475	0.18 (10/55)	0.56 (14/25)	0.33 (3,088/9,412)	0.42 (6,255/14,983)
Lung	127,370	0.13 (1,182/8,820)	0.26 (1,003/3,836)	0.31 (15,359/48,930)	0.42 (27,524/65,784)
Muscle Skeletal	19,516	0.21 (16/77)	0.29 (14/48)	0.30 (2,020/6,837)	0.36 (4,572/12,554)
Ovary	105,088	0.13 (1,476/11,079)	0.23 (1,482/6,381)	0.30 (10,895/36,923)	0.40 (20,207/50,705)
Prostate	88,689	0.11 (388/3,421)	0.26 (394/1,513)	0.31 (10,751/34,421)	0.42 (20,708/49,334)
Testis	18,767	0.09 (86/962)	0.30 (96/323)	0.33 (2,123/6,495)	0.41 (4,475/10,987)
Whole Blood	52,558	0.09 (70/767)	0.29 (60/205)	0.34 (6,762/19,717)	0.47 (15,118/31,869)

(b)

Table 2.2: Replication rates in the (a) FUSION and (b) GoDMC data for SNP-CpG associations identified by X-ING (posterior probability>0.8). Proportions and numbers of SNP-CpG pairs identified in GTEx that are also significant in FUSION/GoDMC ($P < 6 \times 10^{-7}$) are listed. Those SNP-CpG pairs are divided into four groups based on the number of tissues and the presence of associations with cis-genes.

among randomly selected SNP-trans-gene pairs in eQTLGen was 0.04%. Twenty-four of the 282 trans-eSNPs were trans-eSNPs in at least one another tissue besides blood and they were all replicated in eQTLGen. Moreover, 139 out of the 282 were also cis-eQTLs in at least one tissue, of which 44 were replicated in eQTLGen. The replication rate showed a 3.8-fold enrichment compared to trans-eSNPs that were not cis-eQTL ($P = 1 \times 10^{-6}$; two-sided Fisher's exact test). We observed similar patterns for trans-methylation associations. At the P-value threshold of 3×10^{-9} (by Bonferroni correction) [Min et al., 2021, Gaunt et al., 2016], 71 trans-methylation associations identified by X-ING were replicated in GoDMC, with a 7.4-fold enrichment ($P < 1 \times 10^{-16}$; two-sided Fisher's exact test) compared to randomly

selected SNP-trans-CpG pairs. Among the 71 replicated trans-methylation associations, 44 of them were identified in at least two tissues in GTEx. Moreover, all 71 replicated trans-mSNPs were also cis-mQTLs in GTEx. Consistent with existing studies for cis-mediated mechanism of trans-associations [Yang et al., 2021, 2017, Pierce et al., 2014], our results suggest that trans-e/m associations with joint trans- and cis-e/mQTL effects are more likely to be replicated.

2.5.4 Tissue-sharing patterns of trans-association and cis-mediated trans-association effects

To further characterize the tissue-sharing patterns of trans-effects mediated by cis-gene/CpG sites, we analyzed the cis- and trans-e/m association effects identified by X-ING in the nine shared tissues. Out of the 19,003 SNP-trans-gene pairs with trans-associations in at least one tissue, we first selected the pairs with cis-eQTLs. There were 7,479 analyzed trios of SNP, cis-gene and trans-gene. Similarly, we selected 13,952 trios of SNP, cis-CpG site and trans-CpG site out of 14,433 SNP-trans-CpG pairs. For each trio, we estimated the indirect effect of a SNP on its trans-gene/CpG site via cis-gene/CpG site and the direct effect.

Figure 2.11a showed that the total effects of 5,934 (31.2%) SNP-trans-gene pairs and the indirect trans-expression association effects through cis-gene of 2,160 (28.8%) SNP-cis-trans trios were shared by magnitude in at least two tissues. Here effects shared by magnitude refer to the effects with the same sign and within a factor of two of the strongest effect across tissues. Our findings are consistent with previous reports of the tissue-sharing patterns of indirect trans-association effects through cis-gene expression [Yang et al., 2021]. Moreover, we found similar effect-sharing patterns for trans-methylation associations (Figure 2.11b). There were 4,705 (32.6%) SNP-trans-CpG site pairs with shared trans-methylation association effects in at least two tissues. 3,436 (24.7%) SNP-cis-trans trios shared similar indirect trans-mQTL effects via cis-CpG site in at least two tissues. Our results suggest that many

trans-associations and cis-mediated trans-association effects are shared in some but not all tissue types. Proper multi-tissue analysis may enhance the power to detect them.

We plotted the negative log base 10 of the mediation P-values versus the percentage of reduction in trans-effects after accounting for putative cis-mediators (Figure 2.13-2.14). The percentage of reduction in trans-effects is also the ratio of indirect effect to total effect and is expected to be in the range of 0 to 1. A negative reduction in trans-effects with a significant mediation P-value suggests a potential false discovery of cis-mediation. Among trios with significant cis-mediated trans-effects (FDR$<$0.05), trios identified as having multi-tissue trans-effects are less likely to have negative reductions in trans-effects, compared to trios identified as having single-tissue trans-effects in both e- and mQTL analyses. Those results suggest that multi-tissue analyses may reduce false discoveries of cis-mediated trans-association effects compared with single-tissue e/mQTL analyses.

2.5.5 Integrating spatial transcriptomic data with multi-tissue eQTLs reveals spatially-defined molecular links underlying SCZ genetics

The X-ING method could also be applied to integrate broader and complementary sets of summary statistics to enhance cross-omics cross-feature learning. Here we apply X-ING to integrate differential expression statistics from spatial transcriptomic data with multi-tissue eQTL statistics from GTEx. We detect the genes in cis-association with SCZ risk loci and also show laminar-specific expression (i.e., differential expression across different brain layers), accounting for the shared and data-specific patterns of the two sets of summary statistics [Maynard et al., 2021]. Additionally, our results reveal the enrichment of laminar-specific expression of these genes in certain brain layers, offering valuable insights into spatially defined mechanisms underlying SCZ genetics.

To examine the laminar-specific variations for genes in cis with SCZ loci, we performed an integrative analysis of spatial differential transcriptomic statistics with multi-tissue eQTL

statistics from GTEx brain tissues [Consortium, 2020] ($L = 2$). We first conducted spatial differential expression analysis [Vu et al., 2016] to obtain test statistics across six layers and white matter (WM) for each of the 12 samples ($K_1 = 7$) from Lieber Institute for Brain Development (LIBD) in human dorsolateral prefrontal cortex (DLPFC). Here we tested whether the expression of a gene in one layer differs from the other layers [Vu et al., 2016]. We obtained the set of brain eQTL statistics from 13 GTEx brain tissues ($K_2 = 13$). In total, we jointly analyzed over 1.6 million SNP-gene pairs ($M = 1.6$ million) matched in the spatial data from LIBD and the GTEx data. For SCZ, we included 8,962 SNP-gene pairs involving 527 SCZ risk SNPs and 3,184 genes in cis (1 MB) with an SCZ SNP. We performed X-ING analysis for each of the 12 samples. At the 90% posterior probability cutoff (FDR = 0.035), we identified genes differentially expressed in each specific layer in each sample and also associated with SCZ risk loci in at least two GTEx brain tissues. Among the 229 genes in cis-association with SCZ loci, a range of 9 to 41 genes exhibited laminar-specific expression for each pair of samples and brain layer. Further examination of these genes revealed that the laminar-specific expression of these SCZ-associated genes was enriched in layer 2 (L2; $P = 0.026$), layer 5 (L5; $P = 0.025$), and WM ($P = 0.070$) (Figure 2.15). The significant enrichment in L2 and L5 were reported by existing studies [Maynard et al., 2021], which demonstrated that SCZ risk genes in L2 and L5 showed decreased expression in SCZ patients. Here we also identified WM being enriched with genes that show spatially differential expression. By performing additional spatial registration of snRNA-seq datasets, Maynard et al. [Maynard et al., 2021] reported preferential expression of oligodendrocyte subtypes in WM, where oligodendrocyte has been reported to contribute to neuropsychiatric disorders such as SCZ and autism spectrum disorder [Raabe et al., 2018, Galvez-Contreras et al., 2020]. Figure 2.16 showed the significance of layer enrichment for differentially expressed genes associated with autism spectrum disorder (ASD) risk loci in at least two GTEx brain tissues. There was an enrichment of differentially expressed genes in L2 ($P = 0.009$), L5

($P = 0.030$), L6 ($P = 0.028$) and WM ($P = 0.020$) for cis-genes associated with ASD risk loci.

2.5.6 X-ING captures biologically meaningful features

In the multi-tissue mQTL (9 tissues) analysis integrating eQTL (28 tissues) maps, we estimated the sample-averaged cell-type fractions using CIBERSORTxNewman et al. [2019] and EpiDISHZheng et al. [2018] from expression and DNA methylation data, respectively. We then calculated the absolute correlations between the eigenvectors from the modulation matrices of X-ING ($\boldsymbol{U}_{\ell C}$'s for eQTL and mQTL data) and sample-averaged cell-type fractions estimated from individual-level data across tissues Oliva et al. [2023]. We showed that the eigenvectors are highly correlated with multiple major cell types (Figure 2.17). In other words, the major patterns/eigenvectors captured by PCA (similarly for CCA) can be interpreted as the surrogate variables for tissue-tissue dependence due to similar cell-type compositions. Similar conclusions have been reported by GTEx and other QTL consortia. In GTEx, PEER factors derived from expression data (similar to PCs) are highly correlated with the enrichment scores of the major cell types estimatedKim-Hellmuth et al. [2020], Consortium [2020].

2.5.7 Disease-specific trans-e/mQTL hotspots explain more phenotypic variation than trait-associated ones

For the 80 selected diseases/traits, at the 80% posterior probability cutoff, there were 644 to 15,490 SNP-gene-CpG site trios out of the examined disease/trait-specific trios with nonzero genetic effects on trans-gene identified in at least one out of the 28 examined eQTL tissues, or having nonzero genetic effects on trans-CpG site in at least one out of the nine examined mQTL tissues. Analyzed SNPs were generally in weak LD (Figure 2.18).

We further studied SNPs with regulatory/association trans-effects in multiple (≥ 5)

genes/CpG sites, i.e., trans-e/mQTL hotspots, to examine their association patterns and contributions to disease/trait heritability. For each disease/trait, we first estimated the SNP-based heritability based on all SNPs, denoted as h^2, using LD score regression [Bulik-Sullivan et al., 2015]. We used genotype data from Caucasian samples in the 1000 Genomes Project as the reference data. Similarly, we re-evaluated the SNP-based heritability, h^2_{r}, after removing T identified trans-hotspots and their neighboring SNPs (within ± 1 MB). Then the percentage of change in heritability per hotspot region was evaluated as

$$\frac{h^2 - h^2_{\mathrm{r}}}{h^2} \cdot \frac{1}{T} \times 100\%.$$

Figure 2.12 shows the violin plots for the percentage of change in heritability per hotspot region. The average percentage of change in heritability per trans-eQTL hotspot region was 1.82% for the 31 examined diseases and 0.84% for the 21 examined traits, and the corresponding average percentage of change attributed to trans-mQTL hotspot regions was 0.73% and 0.36% for diseases and traits, respectively. Disease-associated hotspot regions explained more phenotypic variation compared with trait-associated ones, consistent with their relatively higher contributions to expression difference and their higher fitness burdens Kirsten et al. [2015].

2.6 Discussion

In this work, we propose X-ING as a general framework for the cross-integration of summary statistics from multi-omics data each with multiple cellular contexts. X-ING takes as input the summary statistic matrices from L data types and models each input statistic as a product of Gaussian and latent binary association status. The modeling of L latent binary matrices allows the cross-integration of different data types of different effect distributions, and X-ING captures omics-shared and context-shared association patterns. This is a major

innovation compared with existing multi-context/tissue methods analyzing only one data type at a time. Additionally, X-ING allows for different levels of sparsity in each context, potential sample overlapping, and effect heterogeneity. With simulation studies, we demonstrate that X-ING improves the estimation of association probabilities and effect sizes in various simulated settings by borrowing strengths across different data types and contexts.

We applied X-ING to detect multi-tissue cis-mQTLs integrating eQTL maps, with a focus on cis-mQTLs with co-occurring associations in other omics data and contexts. We examined trans-e/m association patterns across multiple tissues from GTEx, with a focus on the disease/trait-associated SNPs. The cis-mQTLs and trans-e/m associations identified by X-ING were replicable, especially for those with effects identified in multiple tissues or omics data types. The enrichment of trans-associations in tissue types is informative in suggesting the disease/trait relevance of tissues. We also characterized the tissue-sharing patterns of total effects and indirect effects of trans-association through cis-mediators. In another analysis, we illustrate the broader application of X-ING by integrating spatially differential expression statistics from spatial transcriptomic data with multi-tissue eQTL statistics from 13 GTEx brain tissues. We highlighted the spatial heterogeneity in expression variation of many SCZ risk-associated genes and provided new insights into studying the spatially defined mechanisms underlying SCZ genetics.

There are some limitations and caveats of current work. First, the detected joint associations across multi-omics data or in multiple cellular contexts are not evidence of causation. X-ING does not perform colocalization analysis. Though the findings of X-ING may provide insights of potential connected relationships and mechanisms, it should be interpreted as associations. Second, the cross-integrative methods are not powerful in detecting effects and associations that are specific to only one omics data type in only one context. For those omics- and context-specific effects, existing multivariate methods may not improve power and context-specific sample size is still the major limiting factor. Third, X-ING treats the

M tested units as independent in the estimation. When analyzing published disease/trait-associated SNPs or single-tissue QTLs, most of them are uncorrelated or in weak LD. In general, we recommend applying X-ING to tested units with at most moderate dependence. Last but not least, X-ING does not allow missingness in the input statistics, and a naive imputation may facilitate the analysis but may induce biases if there is substantial missingness.

In future work, X-ING can be improved with a more efficient and selective data integration, when the number of available sets of summary statistics is high, e.g., $L \geq 5$ and some $K_\ell \geq 50$. Another potential area of future development is the integration of association statistics with mediation and causal estimates from multiple studies to reduce confounding and spurious associations.

	Expression	Methylation
Breast Mammary Tissue	396	49
Colon Transverse	368	189
Kidney Cortex	73	47
Lung	515	190
Muscle Skeletal	706	42
Ovary	167	140
Prostate	221	105
Testis	322	47
Whole Blood	670	47

Table 2.3: Tissues and e/mQTL analysis sample sizes of the nine tissues with both DNA methylation and expression data from GTEx.

Tissue	Number of samples with expression data
Artery Aorta	387
Artery Coronary	213
Artery Tibial	584
Brain Amygdala	129
Brain Anterior cingulate cortex (BA24)	147
Brain Caudate (basal ganglia)	194
Brain Cerebellar Hemisphere	175
Brain Cerebellum	209
Brain Cortex	205
Brain Frontal Cortex (BA9)	175
Brain Hippocampus	165
Brain Hypothalamus	170
Brain Nucleus accumbens (basal ganglia)	202
Brain Putamen (basal ganglia)	170
Brain Spinal cord (cervical c-1)	126
Brain Substantia nigra	114
Colon Sigmoid	318
Heart Atrial Appendage	372
Heart Left Ventricle	386

Table 2.4: Tissues and tissue sample sizes of the other 19 tissues with only expression data used in cis- and trans-e/mQTL analyses of GTEx data.

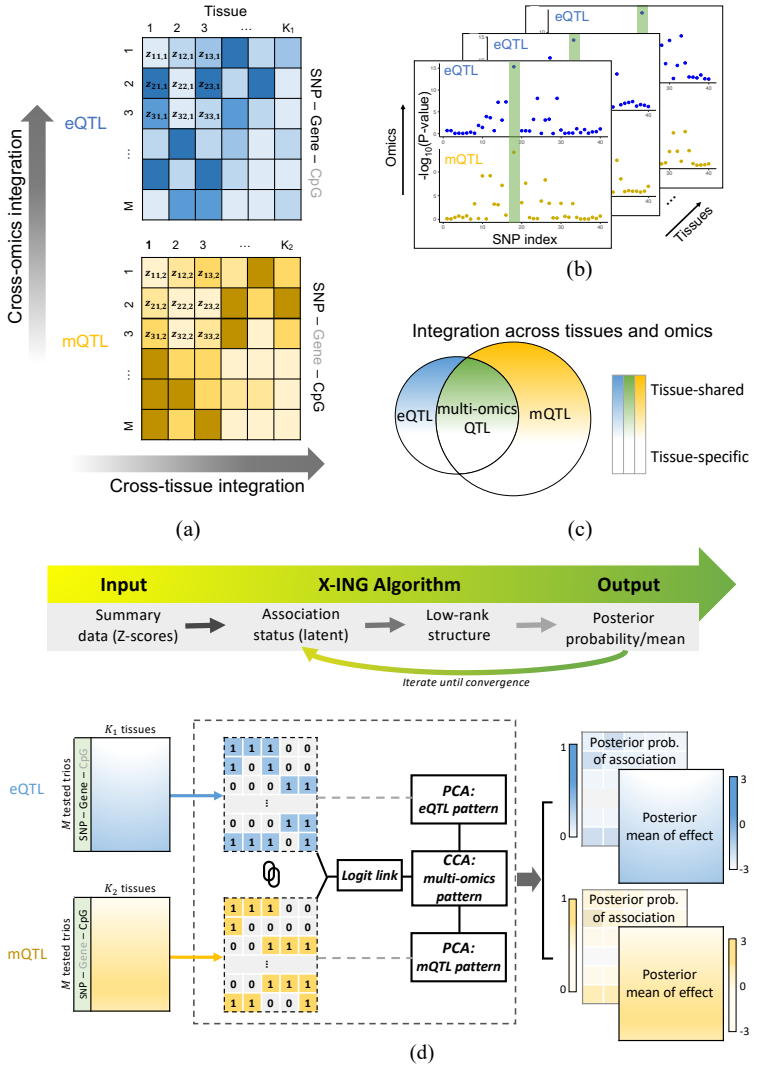


Figure 2.1: Illustrations of the integrative analysis of multi-tissue e/mQTLs and the X-ING algorithm. (a) An illustration of the multi-tissue e/mQTL integrative analysis. A total of M trios are tested, each consisting of a SNP, a cis-gene, and a cis-CpG site. eQTL data are from K_1 tissues and mQTL data are from K_2 tissues. (b) Existing multi-tissue QTL methods analyze each omics data type separately or study the co-occurring patterns of associations tissue by tissue. (c)The X-ING integrative analysis jointly analyzes multi-tissue eQTL and mQTL association statistics, borrows strengths across omics and tissue types, and captures association patterns that are omics-shared or tissue-shared. (d) An illustration of the X-ING algorithm via the multi-tissue e/mQTL analysis: X-ING takes as input $L = 2$ matrices of Z-statistics from eQTL and mQTL studies. It models the latent association status for each input statistic. Via a logit function, X-ING links the latent association status with a continuous modulation matrix for each data type. By performing CCA and PCA on the modulation matrices for e/mQTL data, X-ING captures the low-rank data-shared and data-specific major patterns. X-ING outputs the posterior probability and the posterior mean of association for each input statistic, accounting for those major patterns.

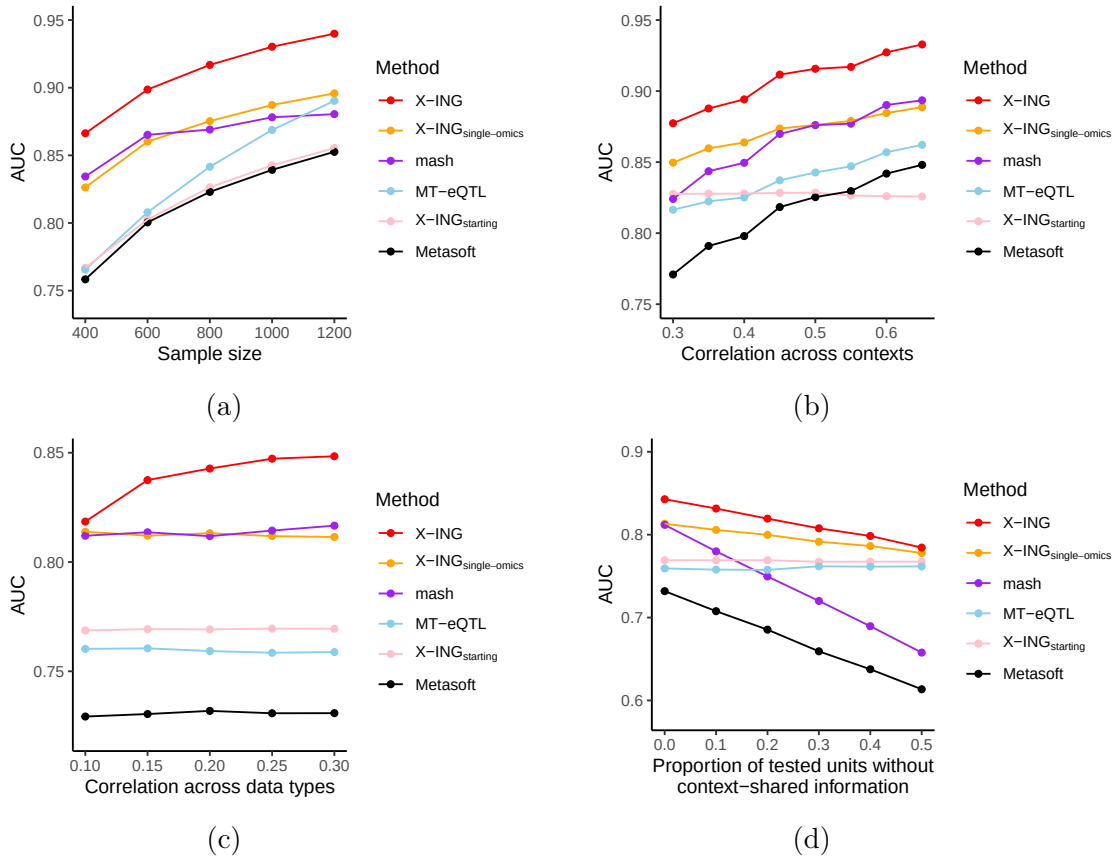


Figure 2.2: Comparison of methods on simulated data. (a) AUC for detecting the presence of nonzero effects on Data 1 with a sample size of Data 1 varying from 400 to 1200. We set $\rho_1 = \rho_2 = 0.5$, $r = 0.3$ and $\theta_1 = \theta_2 = 0.3$. (b) AUC on Data 1 with within-omics cross-context/tissue correlation, ρ_1, varying from 0.3 to 0.65. $N_1 = N_2 = 800$, $r = 0.2$ and $\theta_1 = \theta_2 = 0.3$. (c) AUC on Data 1 with cross-omics tissue-tissue correlation, r, varying from 0.1 to 0.3. $N_1 = N_2 = 400$, $\rho_1 = \rho_2 = 0.4$. (d) AUC on Data 1 with the proportion of tested units that do not have context-shared information varying from 0 to 0.5. $N_1 = N_2 = 400$, $\rho_1 = \rho_2 = 0.4$, $r = 0.2$ and $\theta_1 = \theta_2 = 0.3$.

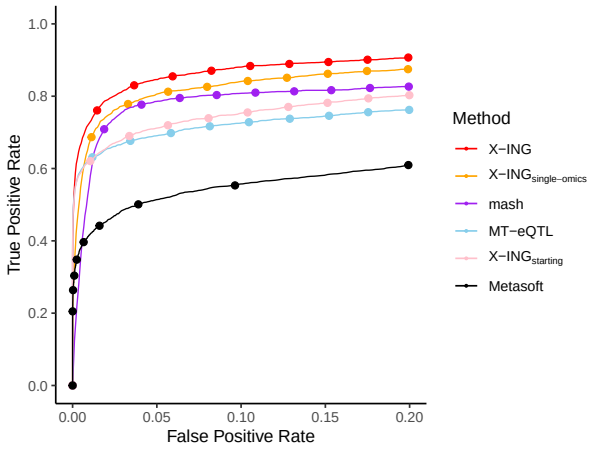


Figure 2.3: ROC curves for detecting nonzero associations in sparse data, with $\tau_\ell = 0.02$. Here $N_1 = N_2 = 1200$, $\rho_1 = \rho_2 = 0.4$, $r = 0.3$ and the proportion of phenotypic variation explained by predictors for each data type was 0.2.

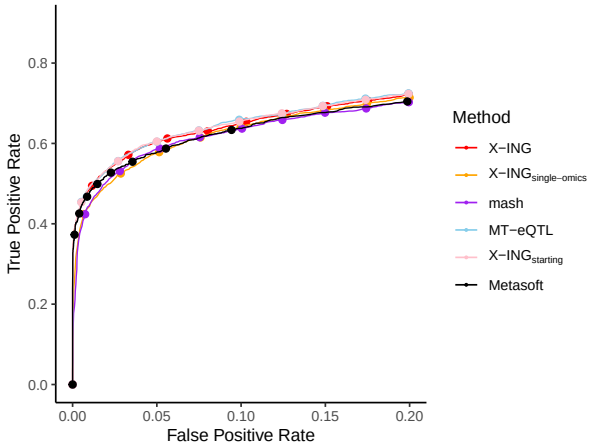


Figure 2.4: ROC curves for detecting nonzero associations when input summary statistics were from independent contexts with no information/effect sharing ($\rho_1 = \rho_2 = r = 0$). Here $N_1 = N_2 = 1200$ and the proportion of phenotypic variation explained by predictors was 0.2. When there was no information/effect sharing, all methods perform similarly.

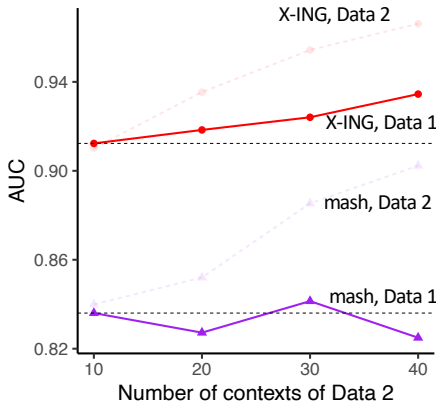


Figure 2.5: Comparison of AUC between X-ING and mash on Data 1 and 2 with varying number of contexts for omics Data 2. K_2 varied from 10 to 40.

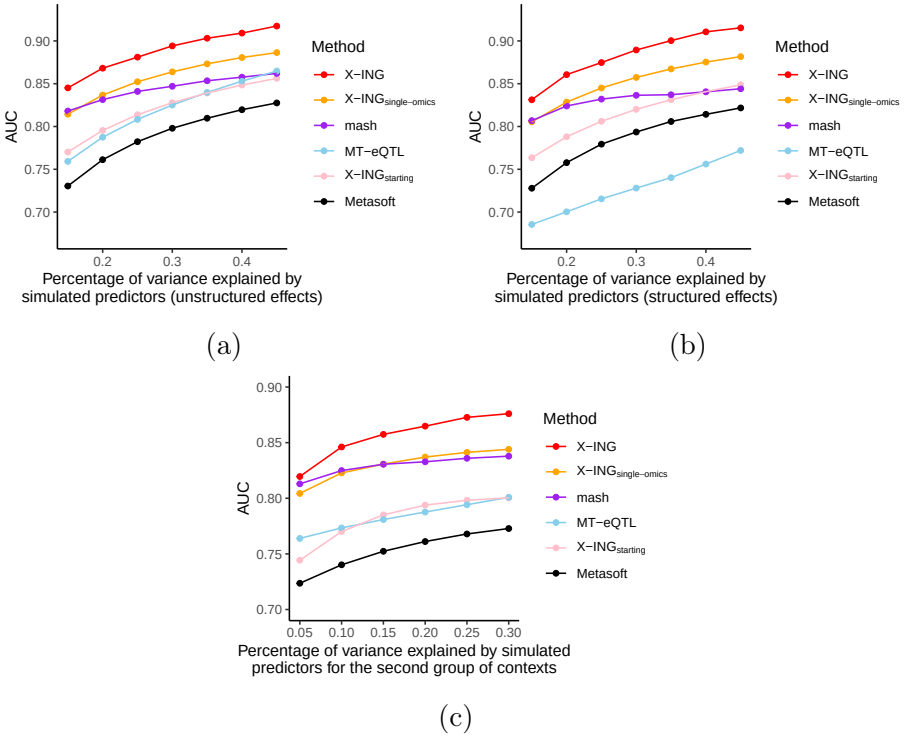


Figure 2.6: Comparison of methods on simulated data. (a) AUC on Data 1 with unstructured effects. The proportion of variation explained by predictors θ_1 varied from 0.15 to 0.45. The simulated true effects were unstructured, i.e., true effects were independently generated. (b) AUC on Data 1 with structured effects. The proportion of phenotypic variation explained by predictors θ_1 varied from 0.15 to 0.45. The simulated true effects were structured, i.e., correlated for those with true non-null association. (c) AUC on Data 1 with unstructured effects. The proportion of phenotypic variation explained by predictors was fixed as 0.2 for the first 7 contexts while that of the left 3 contexts ranged from 0.05 to 0.3.

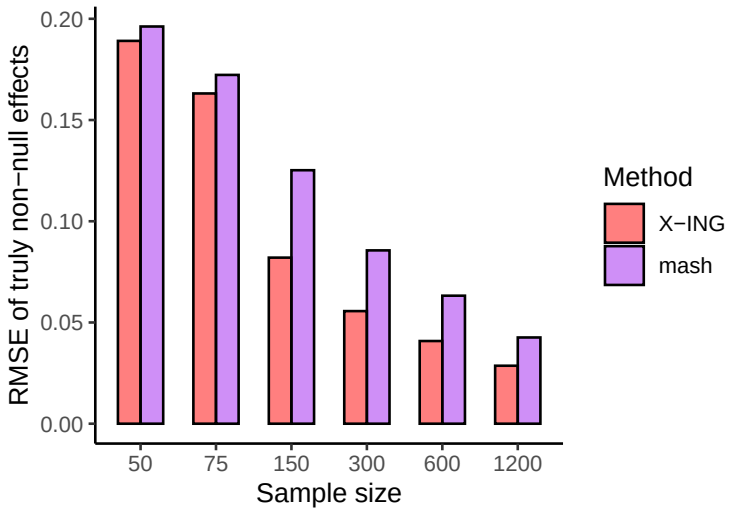
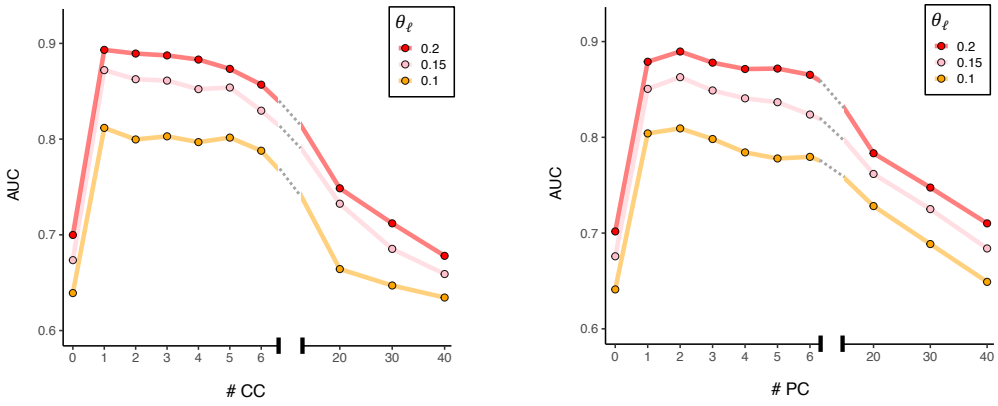


Figure 2.7: Comparison of RMSEs for posterior means estimated by X-ING and mash on true non-null effects. The sample size N_1 varied from 50 to 1200.



(a) (b)

Figure 2.8: Comparison of AUCs using X-ING with different choices of (a) #CC and (b) #PC. The low-rank approximation (i.e., choosing a small but nonzero number of #CC and #PC) is necessary, and X-ING is robust to the choice of #CC and #PC within a certain range.

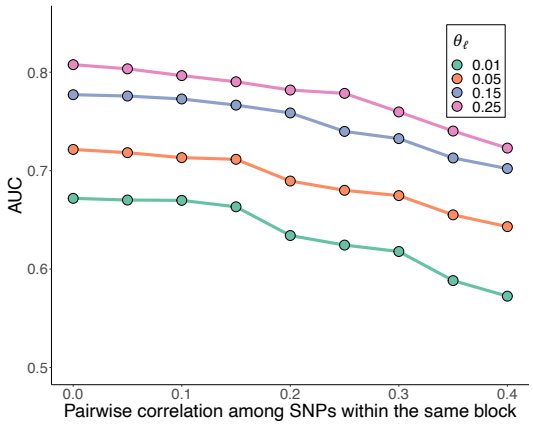


Figure 2.9: Comparison of AUCs using X-ING for detecting nonzero effects. Simulated data were generated with varying levels of pairwise correlation for SNPs within the same block. Here θ_ℓ represents the variance in expression that can be explained by the SNPs.

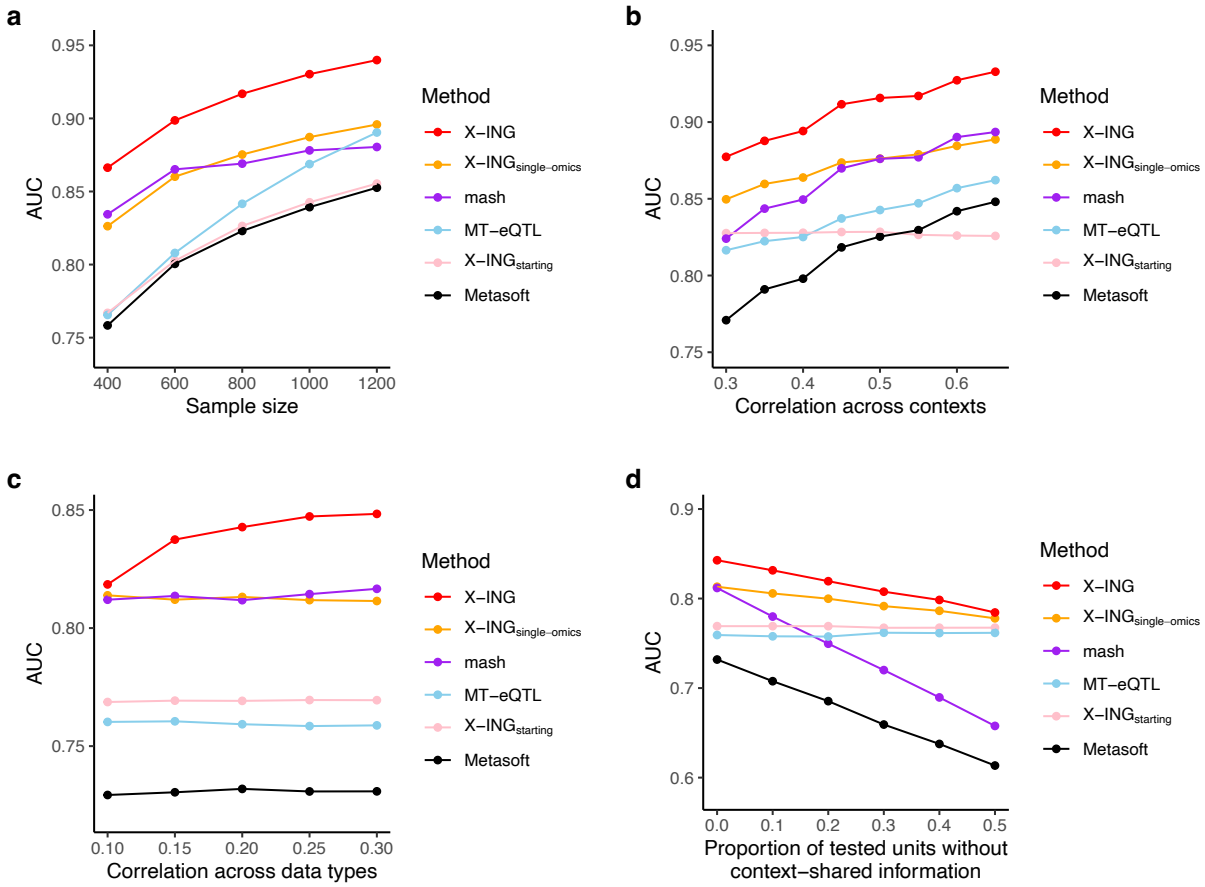
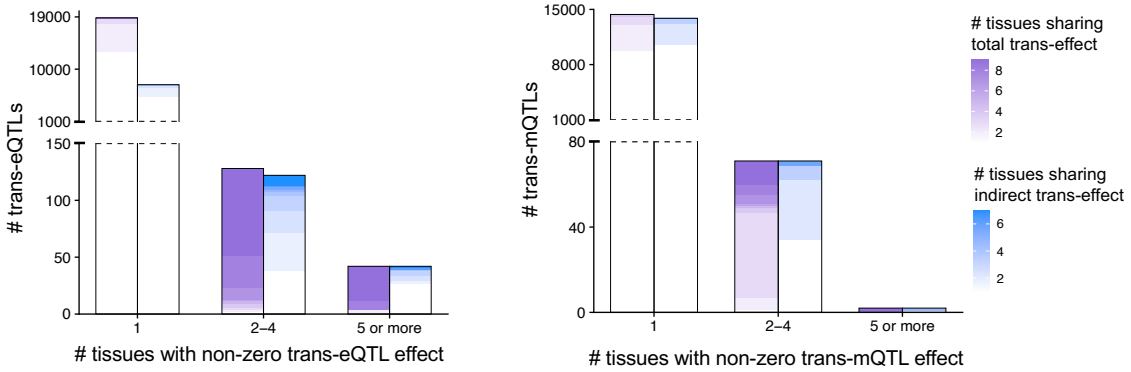
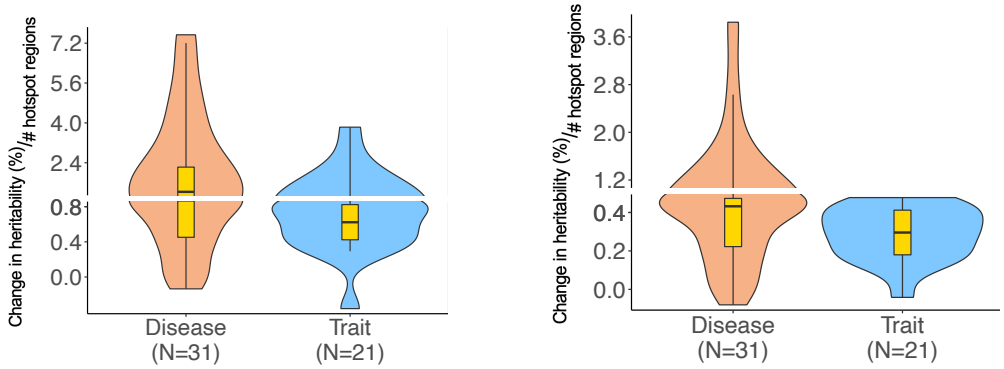


Figure 2.10: Heatmaps showing the scaled proportions of (a) pairs of SNP and trans-gene with trans-association effects and (b) pairs of SNP and cis-gene with cis-association effects identified by X-ING among disease/trait-associated SNPs for 35 selected diseases/traits (y-axis). Red represents an enrichment and blue indicates a depletion of associations. We label the tissue with the highest level of trans-association enrichment for each disease/trait using solid lines. The tissue with the second highest enrichment in trans-association is labeled with dashed lines.



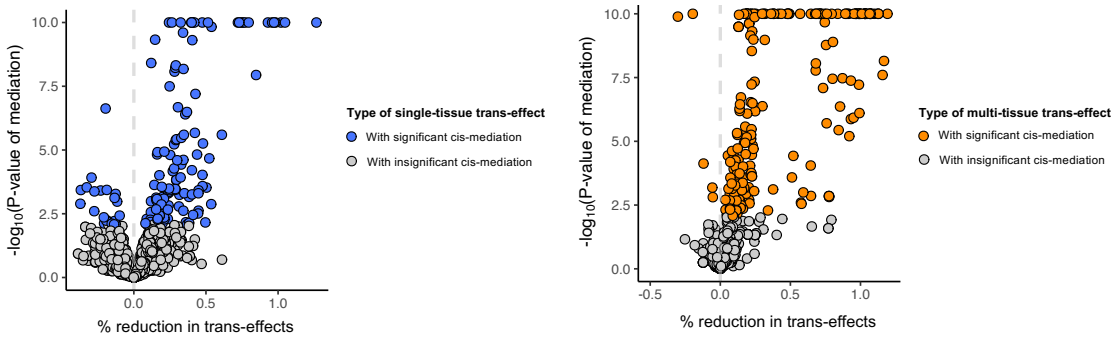
(a) (b)

Figure 2.11: Tissue-sharing patterns of trans-association effects and cis-mediated trans-association effects. (a) Among the 19,003 selected SNP-trans-gene pairs, the total trans-eQTL effects of 5,934 (31.2%; purple) pairs are shared in at least two tissues. Among the 7,479 analyzed trios of SNP, cis-gene and trans-gene, the indirect trans-eQTL effects through cis-gene of 2,160 (28.8%; blue) trios are shared in at least two tissues. (b) Among the 14,433 SNP-trans-CpG pairs, 4,705 (32.6%; purple) pairs have shared trans-mQTL effects in at least two tissues. Among the 13,952 trios of SNP, cis-CpG site and trans-CpG site, 3,436 (24.7%; blue) trios share similar indirect trans-mQTL effects via cis-CpG site in at least two tissues. Note that tissue-shared total/indirect effects refer to the effects with the same sign and within a factor of two of the strongest effect across tissues. The gradient color in each bar represents the number of tissues with shared effects.



(a) (b)

Figure 2.12: (a) The changes in estimated SNP-based heritability per hotspot region for each disease/trait after removing the trans-eQTL hotspot regions. The average change for the 31 examined diseases (orange) was 1.82%. The average change for the 21 examined traits (blue) was 0.84%. The maximum change was 7.54% for diseases and was 3.83% for traits. (b) The changes in heritability attributed to trans-mQTL hotspot regions. The average changes were 0.72% and 0.35% for diseases (orange) and traits (blue), respectively. The maximum change was 3.85% for diseases and was 0.87% for traits.



(a) (b)

Figure 2.13: Reduction in effect of SNP on trans-gene (x-axis) after accounting for cis-gene versus mediation P-values (y-axis; in negative log base of 10). The mediation P-values were calculated for (a) SNP-cis-trans trios identified as having single-tissue trans-effects, and (b) trios identified as having multi-tissue trans-effects. Trios identified as having single-tissue trans-effects with significant mediation effects (FDR<0.05) are in blue. Trios identified as having multi-tissue trans-effects with significant mediation effects (FDR<0.05) are in orange. Trios with insignificant mediation effects are in grey. Here P-values are truncated at 10^{-10}. Reduction percentage in trans-effects is given by $(\beta_{\text{total}} - \beta_{\text{direct}})/\beta_{\text{total}} \times 100\%$, where β_{total} is the total trans-effect, and β_{direct} is the direct trans-effect after adjusting for cis-gene. For trios identified as having single-tissue trans-effects, 23 out of 149 (15.4%) with mediation effect (FDR< 0.05) showed negative reductions, while for trios identified as having multi-tissue trans-effects, 5 out of 226 trios (2.2%) showed negative reductions.

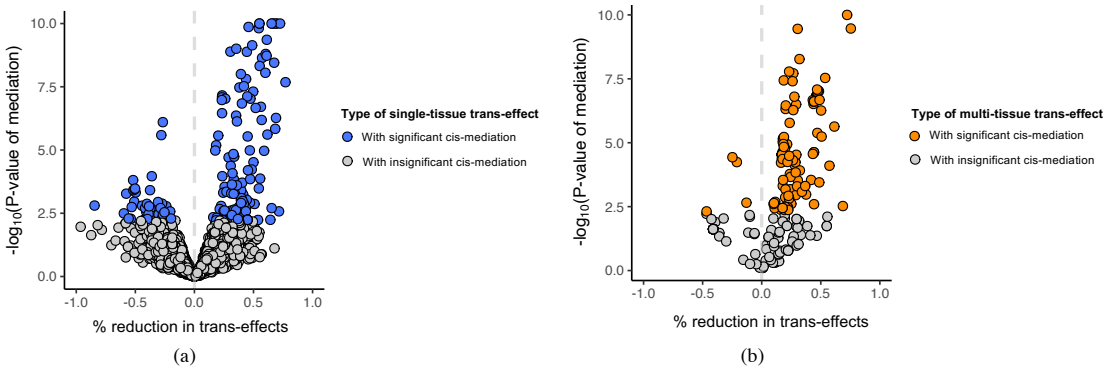


Figure 2.14: Reduction in effect of SNP on trans-CpG site (x-axis) after accounting for cis-CpG site versus mediation P-values (y-axis; in negative log base of 10). The mediation P-values were calculated for (a) SNP-cis-trans trios identified as having tissue-specific trans-effects, and (b) trios identified as having multi-tissue trans-effects. Trios identified as having single-tissue trans-effects and having significant mediation effects (FDR<0.05) are in blue. Trios identified as having multi-tissue trans-mQTLs and having significant mediation effects (FDR<0.05) are in orange. Trios with insignificant mediation effects are in grey. Here P-values are truncated at 10^{-10}. For trios identified as having tissue-specific trans-effects in (a), 41 out of 165 trios (24.8%) with significant mediation effect (FDR< 0.05) showed negative reductions, while for trios identified as having multi-tissue trans-effects, 4 out of 127 trios (3.1%) showed negative reductions.

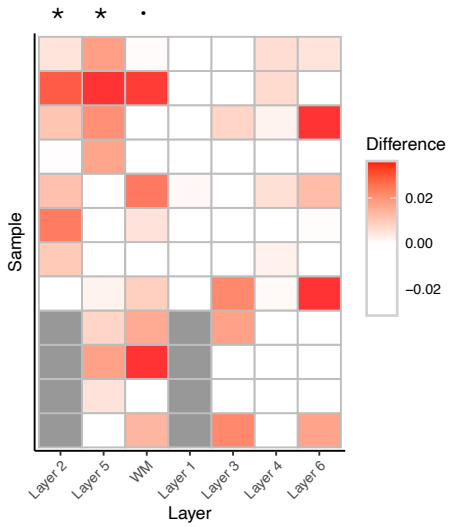


Figure 2.15: Heatmap showing the enrichment of layer-specific differentially expressed genes among disease risk-associated genes, compared with the proportions of layer-specific differentially expressed genes across the genome. The color in each cell indicates the difference between the two proportions for each sample and layer. Red represents an enrichment of differentially expressed genes in a specific sample and layer, and white represents a depletion of differentially expressed genes. Gray cells indicate missing values (no distinct layer information). There is an enrichment of differentially expressed genes in layer 2 ($P = 0.026$), layer 5 ($P = 0.025$) and white matter ($P = 0.070$) for cis-genes associated with SCZ risk loci.

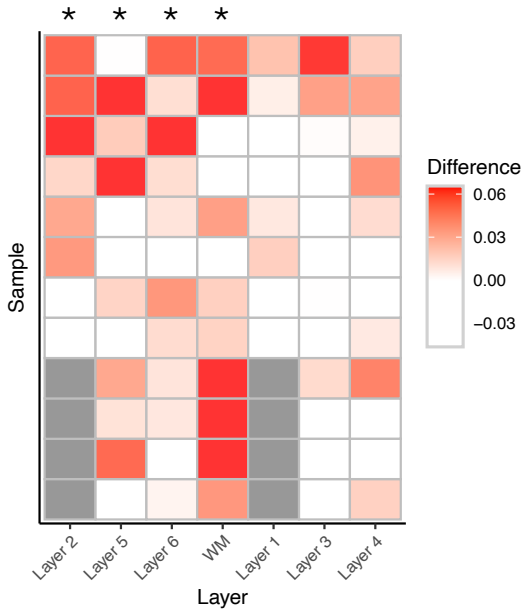


Figure 2.16: Enrichment heatmap of layer-specific differentially expressed genes among ASD risk-associated genes with the proportions of layer-specific differentially expressed genes across the genome. Color in each cell indicates the difference between the two proportions for each sample and layer. Red represents an enrichment of differentially expressed genes in specific samples and layers, and white represents a depletion of differentially expressed genes. Grey cells indicate missing values (no distinct layer information). There is an enrichment of differentially expressed genes in layer 2 ($P = 0.009$), layer 5 ($P = 0.030$), layer 6 ($P = 0.028$) and white matter ($P = 0.020$) for cis-genes associated with ASD risk loci.

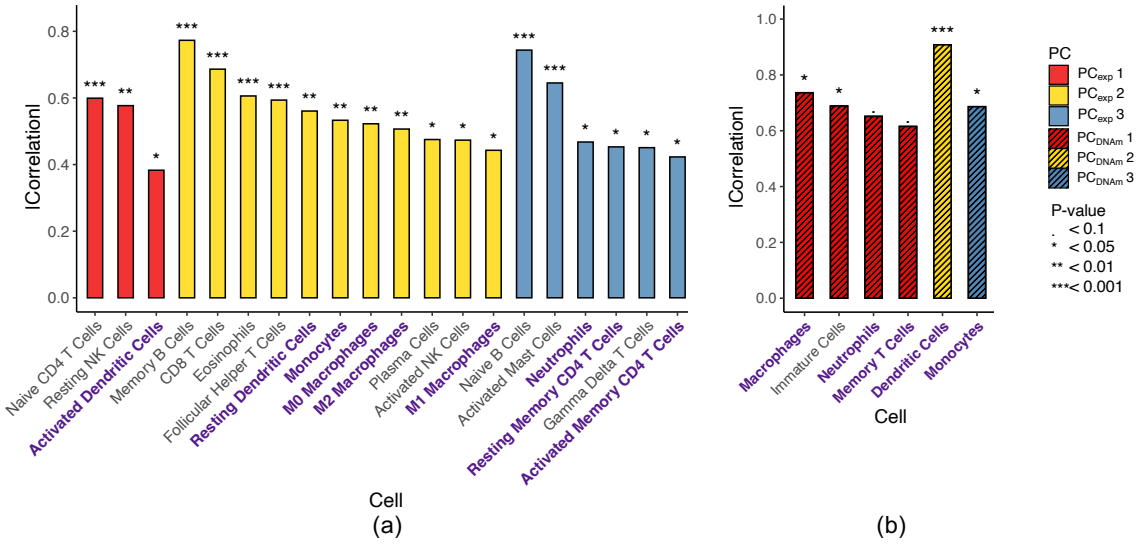


Figure 2.17: Absolute values of correlations between the estimated sample-averaged cell-type fractions for the listed cell types and its most correlated PC. The significance of the correlations is labeled. Cell types that show significant correlations with at least one PC in both eQTL and mQTL data are in purple. The sample-averaged cell-type fractions are derived from (a) expression data using CIBERSORTx and (b) DNA methylation data using EpiDISH.

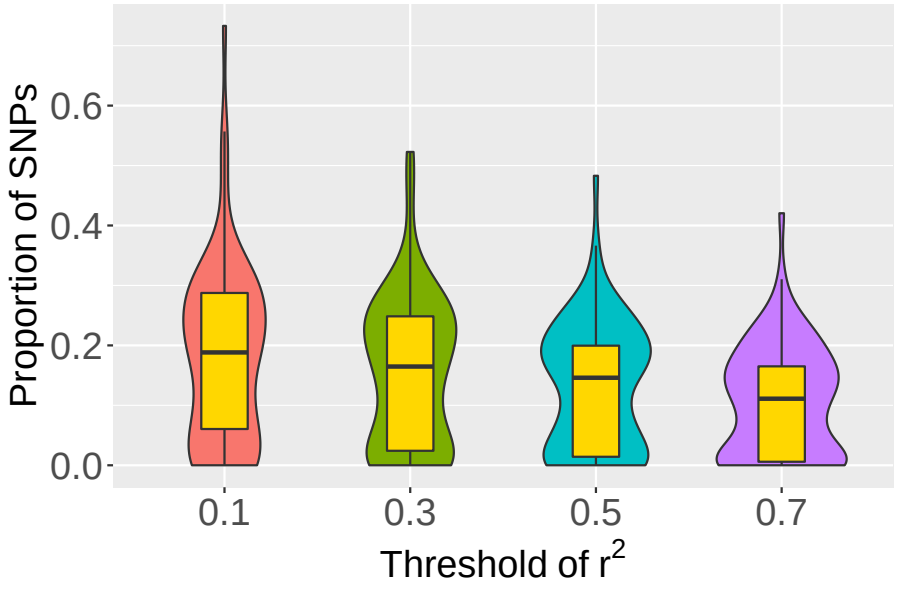


Figure 2.18: Proportion of risk SNPs that are in LD for the 80 diseases/traits. Here the threshold of r^2 varied from 0.1 to 0.7 with fixed window size at 25KB.

CHAPTER 3

AN INTEGRATIVE MULTI-CONTEXT MENDELIAN RANDOMIZATION METHOD FOR IDENTIFYING RISK GENES ACROSS HUMAN TISSUES

3.1 Attributions

Dr. Lin Chen conceived the project. Drs. Lin Chen and Fan Yang contributed to the development of the methods and the writing of the manuscript. Ke Xu contributed to the development of the estimation algorithm. All of the collaborators provided valuable suggestions for the development of the methods and the data analyses.

3.2 Introduction

Mendelian randomization (MR) examines the causal relationships between risk exposures and complex disease outcomes, using genetic variants as instrumental variables (IVs) [Chen et al., 2007, Lawlor et al., 2008, Schadt et al., 2005, Davey Smith and Ebrahim, 2003]. With the rapidly growing availability of summary statistics from genome-wide association studies (GWASs), two-sample MR leveraging two sets of GWAS summary statistics as input has achieved many successes in assessing the causal effects of complex traits as exposures for diseases [Burgess et al., 2013, Bowden et al., 2015, Zhao et al., 2020, Cheng et al., 2022, Wang et al., 2021, Xue et al., 2021, Morrison et al., 2020]. Recently, transcriptome-wide MR (TWMR) considers gene expression as risk exposure and leverages expression quantitative trait loci (eQTLs) and GWAS summary statistics to map risk genes [Gleason et al., 2021, Richardson et al., 2020, Barfield et al., 2018, Zhou et al., 2020]. Unlike transcriptome-wide association studies (TWAS) [Shi et al., 2020, Hu et al., 2019], TWMR focuses on causal assessment. Comparing with colocalization analysis [Oliva et al., 2023, Giambartolomei

et al., 2018, Guo et al., 2015, Foley et al., 2021, Wen et al., 2017], MR offers the flexibility to adjust for known confounders [Sanderson et al., 2019, Anderson et al., 2020], consider joint exposures [Burgess and Thompson, 2015, Rees et al., 2017, Grant and Burgess, 2021, Lin et al., 2023], and allow unmeasured confounders under appropriate assumptions [Cheng et al., 2022, Wang et al., 2021, Xue et al., 2021, Morrison et al., 2020].

While MR offers valuable insights, the application of conventional MR methods in TWMR analysis for mapping risk genes comes with new challenges [Gleason et al., 2021, Yang et al., 2017, Pierce et al., 2018, Verbanck et al., 2018, Gleason et al., 2020]. A notable issue is the limited number of eQTLs as IVs [Gleason et al., 2020, Consortium, 2020], with cis-eQTLs being generally correlated [Consortium, 2020]. Furthermore, disease-associated eQTLs tend to have tissue-specific effects [Umans et al., 2021], while the disease-relevant tissue types are often unknown [Shang et al., 2020, Finucane et al., 2018]. This can lead to inconsistent IV effects across GWAS and eQTL samples, violating core IV assumptions [Burgess et al., 2013, 2015, Pierce and Burgess, 2013]. These issues motivate us to consider multiple tissues simultaneously. Nevertheless, in multi-tissue MR analysis, the causal effects of genes on diseases are often tissue-specific and sparse [Hekselman and Yeger-Lotem, 2020, Ongen et al., 2017, Feng et al., 2021], and thus the estimation of tissue-specific causal effects with a limited number of eQTLs/IVs is challenging.

Recognizing these challenges and opportunities, we propose a multi-context multivariable integrative Mendelian randomization method – mintMR, specifically designed for mapping gene expression and molecular traits as risk exposures. For each gene, we perform a multi-tissue MR analysis using eQTLs with non-zero and sign-consistent effects in more than one tissue as IVs, thereby improving the IV consistency. Our method improves the estimation of tissue-specific causal effects of all genes by simultaneously modeling the latent tissue indicators of disease relevance for multiple gene regions, jointly learning the major/low-rank patterns of latent indicators/probabilities via multi-view learning techniques, and then using

the major patterns to estimate and update the probability of non-zero effects. The rationale is that risk genes for a disease often show non-zero effects in similar or related tissues, [Umans et al., 2021, Finucane et al., 2018] and by jointly learning the major patterns across genes, one can gain improved estimation of tissue-relevance probabilities and further use them to estimate the tissue-specific causal effects for each gene. The joint learning of disease-relevance of latent tissue indicators improves the estimation of sparse tissue-specific causal effects for all genes. Our algorithm iterates between estimating multi-tissue MR models for each gene and jointly learning the latent patterns and probabilities of non-zero causal effects for all genes until the maximum iteration is reached. Our MR framework considers cis gene expression and DNA methylation (DNAm) as joint exposures. Given the frequent co-occurrence of eQTLs and mQTLs, [Pierce et al., 2018] the joint consideration of DNAm with gene expression is crucial for accurately mapping causal genes. If the causal DNAm is associated with gene expression and cis-eQTLs selected as IVs are also associated with DNAm, the DNAm would be a confounder being associated with IV, and its omission could lead to biased causal inference. By jointly assessing the causal effects of gene expression and DNAm, we demonstrate that the proposed method controls genome-wide inflation, improves the power, and offers valuable insights into disease-relevant tissues and mechanisms. Our mintMR approach uniquely tackles challenges in mapping molecular traits as risk exposures via MR, jointly learns the low-rank patterns in the probabilities of disease relevance across many genes, and thereby enhances the estimation of sparse tissue-specific causal effects.

3.3 Methods

3.3.1 A starting model for a single gene region

We start with a multi-tissue MR model for studying the gene expression of a single gene from multiple tissues as the exposure and a complex disease as the outcome. We consider an eQTL

i ($i = 1, ..., I_g$) as an IV for the expression of a gene indexed by g. Let γ_{gik} ($k = 1, ..., K$) denote the true marginal effect of the SNP i on the gene g in tissue k. Let Γ_i^g denote the true marginal association between SNP i and the disease outcome of interest, and the superscript g indicates that the SNP i is an IV for gene g. Denote $\left\{\widehat{\gamma}_{gik}, \widehat{s}_{\gamma_{gik}}\right\}$ as the estimated SNP-gene association and its standard error for SNP i and gene g in tissue k, and $\left\{\widehat{\Gamma}_i^g, \widehat{s}_{\Gamma_i^g}\right\}$ as the estimated effect of SNP i on the outcome and its standard error. We have the model for SNP i:

$$\begin{pmatrix} \widehat{\Gamma}_i^g \\ \widehat{\gamma}_{gi1} \\ \vdots \\ \widehat{\gamma}_{gik} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \Gamma_i^g \\ \gamma_{gi1} \\ \vdots \\ \gamma_{gik} \end{pmatrix}, \widehat{\boldsymbol{S}}_{gi}\boldsymbol{C}\widehat{\boldsymbol{S}}_{gi}\right), \tag{3.1}$$

where $\boldsymbol{C}$ is the tissue-tissue correlation matrix due to sample overlap and is often estimated apriori [Urbut et al., 2019], $\widehat{\boldsymbol{S}}_{gi} = \text{diag}(\widehat{s}_{\Gamma_i^g}, \widehat{s}_{\gamma_{gi1}}, \cdots, \widehat{s}_{\gamma_{giK}})$ is the standard error estimate from GWAS and multi-tissue eQTL studies.

We further assume the true causal relationship between GWAS and eQTL effects, Γ_i^g and γ_{gik}'s, is linear and is given by

$$\Gamma_i^g = \alpha_i^g + \sum_{k=1}^{K} \eta_{gk} \cdot \beta_{gk}\gamma_{gik}, \tag{3.2}$$

where β_{gk} is the causal effect of interest for gene g in tissue k. We introduce η_{gk} as a latent indicator for disease relevance of tissue, and $\eta_{gk} = 1$ if $\beta_{gk} \neq 0$. We assume $\eta_{gk} \sim$ Bernoulli(π_{gk}). The effect of gene expression levels on the disease outcome is often sparse and varies across contexts/tissues/cell types. The effect $\eta_{gk} \cdot \beta_{gk}$ is the direct effect of the gene g in tissue k on the disease outcome not mediated via other exposures (including the gene expression in other tissues). When estimating the latent variables and the causal effects,

the estimated probability of non-zero for the latent indicator can be viewed as a weight on the relevance of tissue types or the proportion of disease-relevant cell types in the current tissues. Without modeling latent disease-relevance tissue indicators, all tissues in the model are weighted equally. Here the true IV-to-exposure effect follows $\gamma_{gik} \sim \mathcal{N}(0, \sigma^2_{\gamma_g})$, and $\alpha_i^g \sim \mathcal{N}(0, \sigma^2_{\alpha^g})$ is the uncorrelated horizontal pleiotropic effect (green arrow in Figure 3.1a) when IV affects outcome not through exposure and IV is not associated with confounder.

Additionally, we consider a multivariable MR (MVMR) framework for a set of L ($l = 1, ..., L$) molecular traits as exposures, each from K_l contexts/tissues. For example, in our motivating application, we jointly consider a gene expression and a CpG site from multiple tissues as the exposures, $L = 2$. Let SNP i be a cis-molecular QTL (xQTLs) for gene g, and $\gamma_{gik,l} \sim \mathcal{N}(0, \sigma^2_{\gamma_{g,l}})$ ($k = 1, ..., K_l$) denote the marginal effect of SNP i on the l-th molecular exposure in tissue k. Extending model (3.2), we assume the following causal relationship holds between the marginal effect of the SNP i on the outcome, i.e., Γ_i^g, and the marginal effects of the SNP i on exposures, i.e., $\gamma_{gik,l}$'s:

$$\Gamma_i^g = \alpha_i^g + \sum_{k=1}^{K_1} \eta_{gk,l} \cdot \beta_{gk,l}\gamma_{gik,l} + \cdots + \sum_{k=1}^{K_L} \eta_{gk,L} \cdot \beta_{gk,L}\gamma_{gik,L}. \tag{3.3}$$

In model (3.3), $\eta_{gk,l} \cdot \beta_{gk,l}$ describes the direct effect of exposure l in tissue k on the outcome not operating through the exposure in other tissues nor through other exposures ($l' \neq l$). Here similar to model (3.2), we assume $\eta_{gk,l} \sim$ Bernoulli($\pi_{gk,l}$) and $\alpha_i^g \sim \mathcal{N}\left(0, \sigma^2_{\alpha^g}\right)$. The MVMR model allows the joint modeling of correlated cis-molecular traits in the gene regions to identify the risk factors and elucidate the mechanisms. In practice, since often there are only a limited number of xQTLs as IVs, the causal effects (and the latent indicators) in the above single-gene models (3.2) and (3.3) may not be statistically identifiable.

3.3.2 The proposed mintMR model for jointly learning the disease-relevance of tissue indicators across G gene-CpG pairs

Common eQTLs are often weakly selected and disease-associated genetic variants typically influence downstream genes with effects being highly context-specific [Umans et al., 2021]. When multiple genes are causally affecting diseases in a pathway or gene set, they often have effects specific to certain disease-associated tissues and cell types. Furthermore, the enrichment of disease-associated gene expression has been successfully used to identify disease-relevant tissues and cell types [Finucane et al., 2018]. These observations motivate us to jointly learn the patterns of disease-relevance indicators/probabilities across many genes, especially considering the sparse nature of disease-relevant causal effects.

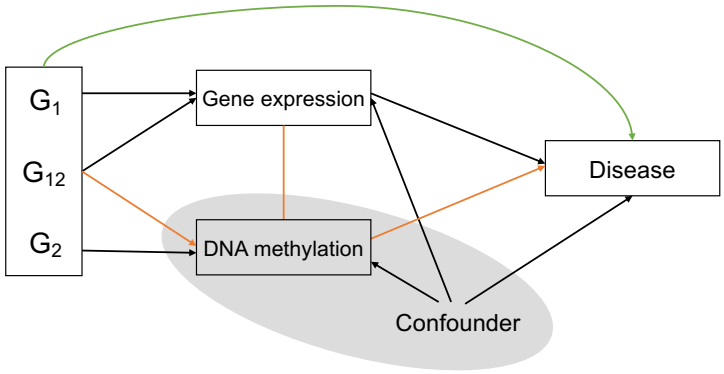
We propose a joint MVMR model across G gene-CpG pairs to estimate the causal effects for each gene and CpG in each tissue and jointly learn the major patterns of latent disease-relevance tissue indicators, particularly in scenarios where these effects are sparse. As illustrated in Figure 3.1b, we consider multi-tissue expression and DNAm of the gene-CpG pairs from the g-th gene region ($g = 1, \cdots, G$) and study their effects on the outcome. While the direct effects $\beta_{gk,l}$'s may vary in magnitude and direction, there could still be concerted patterns among the true non-zero causal effects and their effect operating contexts/tissues. The proposed mintMR model works by iteratively estimating the starting model (3.3) for each gene-CpG pair (one red box in Figure 3.1b) and collectively capturing the low-rank (major) patterns of non-zero causal effects across G gene regions for updating the tissue-relevance probabilities/weights until the maximum iteration is reached. The resulting estimates provide not only the causal effect for each gene and CpG site, but also the estimated probability of disease relevance for each gene-tissue or CpG-tissue pair accounting for shared patterns. A major innovation of our model is the use of multi-view learning methods to capture the low-rank patterns shared across gene regions and omics-data types. The details of the estimation are provided in Algorithm 2.

To learn the low-rank patterns of disease-relevance (non-zero causal effects) across genes, molecular exposures, and tissue types, one may employ multi-view learning strategies such as co-training [Ma et al., 2020], multiple kernel learning [Liu et al., 2023], and canonical correlation analysis (CCA) [Wang et al., 2015, Li et al., 2020]. For each gene-CpG pair, we have model (3.3). We model the latent disease-relevance tissue indicators for all G gene-CpG-tissue trios, assuming the latent indicator $\eta_{gk,l} \sim \text{Bernoulli}(\pi_{gk,l})$. As illustrated in Figure 3.1b, we form L latent disease-relevance indicator matrices for L molecular exposures, $\boldsymbol{\eta}_l = \{\eta_{gk,l}\} \in \mathbb{R}^{G\times K_l}$ for expression and DNAm ($L = 2$) in our motivating application. We introduce a continuous modulation matrix for each exposure l, $\boldsymbol{U}_l = \{U_{gk,l}\} \in \mathbb{R}^{G\times K_l}$, and

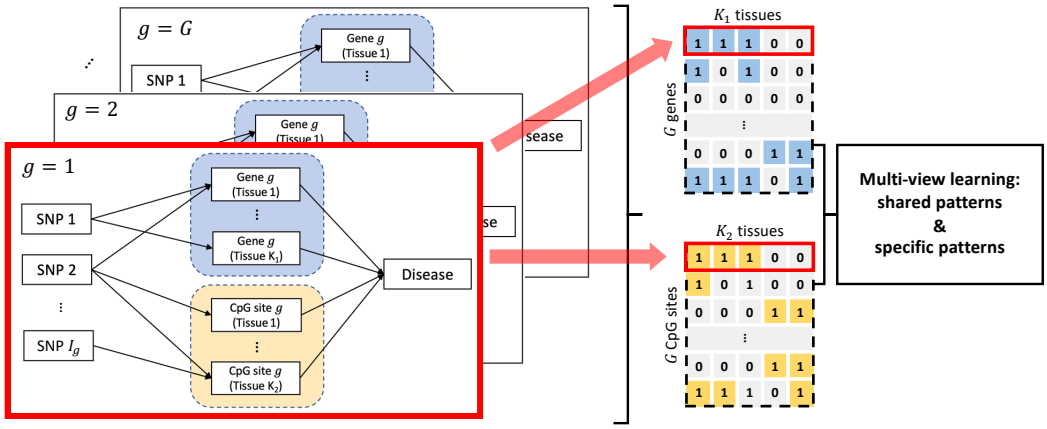
$$U_{gk,l} = \log \frac{\Pr\left(\eta_{gk,l} = 1 \mid \boldsymbol{U}_l, u_{0k,l}\right)}{\Pr\left(\eta_{gk,l} = 0 \mid \boldsymbol{U}_l, u_{0k,l}\right)} - u_{0k,l}. \tag{3.4}$$

Here $\boldsymbol{U}_l$ modulates the probability of the latent binary association status, $u_{0k,l}$ is the tissue-specific intercept, controlling the sparsity of non-zero effects in the k-th tissue of the l-th type of molecular exposure. For each gene ($g = 1, \cdots, G$), we estimate model (3.3) separately and then jointly model the L modulation matrices. We approximate the modulation matrices, $\boldsymbol{U}_l$'s, with low-rank matrices $\widetilde{\boldsymbol{U}}_l$'s capturing the major patterns of disease-relevance across gene regions, molecular exposures, and tissue types. The mintMR model uses these approximated low-rank matrices $\widetilde{\boldsymbol{U}}_l$'s to estimate the disease-relevance probability for each gene/CpG in each tissue without over-parameterization. If there is no pattern shared across gene regions/molecular exposures/tissues, $U_{gk,l} = 0, \forall g, k, l$, and model (3.4) is reduced to $\text{logit}\left(\pi_{gk,l}\right) = u_{0k,l}$, i.e., only tissue-specific prior being imposed for the indicators across exposures.

More specifically, in this work, we capture the major patterns of disease relevance across all genes as the sum of major patterns shared across molecular exposures (expression and



(a)



(b)

Figure 3.1: Illustrations of the multi-context multivariable integrative Mendelian randomization method. (a) The causal diagram of the multivariable MR model. When assessing the effect of gene expression on outcome, if a correlated exposure (e.g., DNA methylation; shaded) is not considered in the model, it will serve as a confounder and bias the inference (orange line). The green line represents the uncorrelated horizontal pleiotropic effect. (b) An illustration of the mintMR framework for analyzing multiple gene-CpG pairs from G gene regions. MintMR takes as input $G \times L$ ($L = 2$ here) sets of IV-to-exposure effects and standard error matrices from multi-tissue eQTL and mQTL studies, respectively. It models the latent status for each causal effect. Via a logit function, mintMR links the latent status of the causal effects with a continuous modulation matrix. By performing multi-view learning on the modulation matrices, mintMR captures the low-rank data-shared and data-specific major patterns and uses them to estimate the disease-relevant probabilities. By iterating between performing MR for each gene region and estimating the disease-relevant probabilities for all genes, mintMR improves the estimation and inference of sparse causal effects for all genes.

DNAm) and major tissue-sharing patterns specific to each data type. We have

$$U_{gk,l} \approx \widetilde{U}_{gk,l} = U^C_{gk,l} + U^R_{gk,l}. \tag{3.5}$$

The matrices $\boldsymbol{U}^C_{\cdot\cdot,l}, l = 1, ..., L$ represent the common major structures shared across the L latent tissue-relevance indicator matrices. We estimate $\boldsymbol{U}^C_{\cdot\cdot,l}$ by applying generalized CCA on the matrices $\{\text{logit}\left(\pi_{gk,l}\right) - u_{0k,l}\}^{G\times K_l}$. Furthermore, the $\boldsymbol{U}^R_{\cdot\cdot,l}, l = 1, ..., L$ matrices capture omics data-specific tissue-sharing patterns. We perform separate principal component analysis (PCA) on each residual matrix $\{\text{logit}\left(\pi_{gk,l}\right) - U^C_{gk,l} - u_{0k,l}\}^{G\times K_l}$ to obtain the low-rank patterns in each omics exposure data type, $\boldsymbol{U}^R_{\cdot\cdot,l}$. Alternative multi-view learning methods could be used to capture different types of desirable data patterns and obtain other approximated matrices [Ma et al., 2020, Liu et al., 2023, Wang et al., 2015, Li et al., 2020]. The proposed mintMR algorithm iterates between estimating the causal effects in the single-gene model (3.3) for each of the G gene regions and jointly learning/estimating the latent disease-relevance indicators/probabilities via Gibbs sampling until the maximum iteration is reached (see Algorithm 2 for details).

Accounting for LD

When studying gene expression and DNAm as joint molecular exposures, the number of e/mQTLs as IVs is generally limited. Applying a stringent LD clumping threshold would lose many IVs and hurt power. Instead of assuming independent IVs as in most existing multivariable MR methods [Lin et al., 2023], we allow IVs to be correlated. Assuming non-overlapping samples, we model the estimated effect sizes by accounting for the correlation

among IVs $i = 1, ..., I_g$:

$$\begin{pmatrix} \widehat{\Gamma}_1^g \\ \vdots \\ \widehat{\Gamma}_{I_g}^g \end{pmatrix} \sim \mathcal{N}\left(\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g}^{-1}\mathbf{\Gamma}^g, \widehat{\mathbf{S}}_{\mathbf{\Gamma}^g}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g}\right), \text{ and}$$
$$\begin{pmatrix} \widehat{\gamma}_{g1k,l} \\ \vdots \\ \widehat{\gamma}_{gI_gk,l} \end{pmatrix} \sim \mathcal{N}\left(\widehat{\mathbf{S}}_{\gamma_{gk,l}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\gamma_{gk,l}}^{-1}\boldsymbol{\gamma}_{gk,l}, \widehat{\mathbf{S}}_{\gamma_{gk,l}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\gamma_{gk,l}}\right), l = 1, ..., L, k = 1, ..., K_l, \tag{3.6}$$

where $\widehat{\mathbf{R}}^g$ is the correlation matrix of the I_g number of IVs for the g-th set of exposures, $\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g} = \text{diag}(\widehat{s}_{\Gamma_1^g}, \cdots, \widehat{s}_{\Gamma_{I_g}^g})$, and $\widehat{\mathbf{S}}_{\gamma_{gk,l}} = \text{diag}(\widehat{s}_{\gamma_{g1k,l}}, \cdots, \widehat{s}_{\gamma_{gI_gk,l}})$. We provided details of the Gibbs sampling algorithm of mintMR accounting for both LD and sample overlap in the following section.

3.3.3 The Gibbs sampling algorithms for mintMR

In this section, we provide the details of the Gibbs sampler used in the mintMR estimation algorithm.

The algorithm for independent SNPs

For the g-th set of exposures, we propose the following Bayesian hierarchical model for independent SNPs and non-overlapping samples.

$$\widehat{\gamma}_{gik,l} \mid \gamma_{gik,l}, \widehat{s}^2_{\Gamma_i^g} \sim \mathcal{N}(\gamma_{gik,l}, \widehat{s}^2_{\gamma_{gik,l}}),\ \widehat{\Gamma}_i^g \mid \Gamma_i^g, \widehat{s}^2_{\Gamma_i^g} \sim \mathcal{N}(\Gamma_i^g, \widehat{s}^2_{\Gamma_i^g}),$$

$$\gamma_{gik,l} \left| \sigma^2_{\gamma_{g,l}} \sim \mathcal{N}\left(0, \sigma^2_{\gamma_{g,l}}\right), \quad \Gamma_i^g \right| \beta_{g\cdot,\cdot}, \gamma_{gi\cdot,\cdot}, \eta_{g\cdot,\cdot}, \sigma^2_{\alpha^g} \sim \mathcal{N}\left(\sum_{l=1}^{L}\sum_{k=1}^{K_l} \eta_{gk,l}\beta_{gk,l}\gamma_{gik,l}, \sigma^2_{\alpha^g}\right),$$

$$\beta_{gk,l} \mid \sigma^2_{\beta_{g,l}} \sim \mathcal{N}\left(0, \sigma^2_{\beta_{g,l}}\right), \quad \alpha_i^g \mid \sigma^2_{\alpha^g} \sim \mathcal{N}\left(0, \sigma^2_{\alpha^g}\right)$$

$$\sigma^2_{\beta_{g,l}} \sim \mathcal{IG}\left(a_\beta, b_\beta\right), \quad \sigma^2_{\gamma_{g,l}} \sim \mathcal{IG}\left(a_\gamma, b_\gamma\right), \quad \sigma^2_{\alpha^g} \sim \mathcal{IG}\left(a_\alpha, b_\alpha\right),$$

$$\eta_{gk,l} \mid \pi_{gk,l} \sim \pi_{gk,l}^{\eta_{gk,l}}(1-\pi_{gk,l})^{1-\eta_{gk,l}}, \quad \pi_{gk,l} \sim \text{Beta}(a_\pi, b_\pi),$$

where $i = 1, 2\cdots, I_g$, $k = 1, 2\cdots, K_l$, and $l = 1, \cdots, L$.

Denote $\widehat{\mathbf{\Gamma}}^g = [\widehat{\Gamma}_1^g, \cdots, \widehat{\Gamma}_{I_g}^g]^\top$, $\widehat{\boldsymbol{\gamma}}_{gk,l} = [\widehat{\gamma}_{g1k,l}, \cdots, \widehat{\gamma}_{gI_gk,l}]^\top$, $\mathbf{\Gamma}^g = [\Gamma_1^g, \cdots, \Gamma_{I_g}^g]^\top$, and $\boldsymbol{\gamma}_{gk,l} = [\gamma_{g1k,l}, \cdots, \gamma_{gI_gk,l}]^\top$, the posterior likelihood is in the form of

$$\begin{aligned}
L(\Theta_g) \propto & \prod_{g=1}^{G} p\left(\mathbf{\Gamma}^g, \boldsymbol{\gamma}_{g1,1}, \cdots, \boldsymbol{\gamma}_{gK_L,L}, \boldsymbol{\beta}_g, \sigma^2_{\mathbf{\Gamma}^g}, \sigma^2_{\alpha^g}, \boldsymbol{\eta}_g, \boldsymbol{\pi}_g \mid \widehat{\mathbf{\Gamma}}^g, \widehat{\boldsymbol{\gamma}}_{g\cdot,\cdot}\right) \\
= & \prod_{g=1}^{G}\left\{\prod_{i=1}^{I_g}\left[p(\widehat{\Gamma}_i^g|\alpha_i^g + \sum_{l=1}^{L}\sum_{k=1}^{K_l}\beta_{gk,l}\eta_{gk,l}\gamma_{gik,l}, \widehat{s}_{\Gamma_i^g})\prod_{l=1}^{L}\prod_{k=1}^{K_l} p(\widehat{\gamma}_{gik,l}|\gamma_{gik,l}, \widehat{s}_{\gamma_{gik,l}})p(\gamma_{gik,l} \mid \sigma^2_{\gamma_{g,l}})\right]\right\} \\
& \left\{\prod_{g=1}^{G} p(\sigma^2_{\alpha_g})\prod_{i=1}^{I_g} p(\alpha_i^g \mid \sigma^2_{\alpha_g})\right\}\left\{\prod_{g=1}^{G}\prod_{l=1}^{L} p(\sigma^2_{\gamma_{g,l}})\right\} \\
& \left\{\prod_{g=1}^{G}\prod_{l=1}^{L} p(\sigma^2_{\beta_{g,l}})\prod_{k=1}^{K_l} p(\beta_{gk,l} \mid \sigma^2_{\beta_{g,l}})\right\}\left\{\prod_{g=1}^{G}\prod_{l=1}^{L}\prod_{k=1}^{K_l} p(\eta_{gk,l} \mid \pi_{gk,l})p(\pi_{gk,l})\right\}.
\end{aligned}$$

Here $p(\widehat{\Gamma}_i^g|\alpha_i^g + \sum_{l=1}^{L}\sum_{k=1}^{K_l}\beta_{gk,l}\eta_{gk,l}\gamma_{gik,l}, \widehat{s}_{\Gamma_i^g}) = p(\widehat{\Gamma}_i^g|\Gamma_i^g, \widehat{s}_{\Gamma_i^g})$.

The conditional posterior distribution of each Γ_i^g given the other parameters in the model

is

$$\Gamma_i^g \mid \widehat{\Gamma}_i^g,\ \widehat{s}_{\Gamma_i^g},\ \gamma_{gi\cdot,\cdot},\ \beta_{g\cdot,\cdot},\ \eta_{g\cdot,\cdot},\ \sigma^2_{\alpha^g} \sim \mathcal{N}\left(\widetilde{\mu}_{gi0}, \widetilde{\sigma}^2_{gi0}\right),$$

where

$$\begin{aligned}
-\frac{1}{2\widetilde{\sigma}^2_{gi0}} &= -\frac{1}{2}\left(\frac{1}{\widehat{s}^2_{\Gamma_i^g}} + \frac{1}{\sigma^2_{\alpha^g}}\right), \\
\frac{\widetilde{\mu}_{gi0}}{\widetilde{\sigma}^2_{gi0}} &= \frac{\widehat{\Gamma}_i^g}{\widehat{s}^2_{\Gamma_i^g}} + \frac{\sum_{l=1}^{L}\sum_{k=1}^{K_l}\eta_{gk,l}\beta_{gk,l}\gamma_{gik,l}}{\sigma^2_{\alpha^g}}.
\end{aligned}$$

The conditional distribution for each element $\gamma_{gik,l}$ comes from a normal distribution with

$$\gamma_{gik,l} \mid \widehat{\gamma}_{gik,l},\ \widehat{s}_{\gamma_{gik,l}},\ \Gamma_i^g,\ \gamma_{gi\cdot,\cdot},\ \beta_{g\cdot,\cdot},\eta_{g\cdot,\cdot},\ \sigma^2_{\gamma_{g,l}},\ \sigma^2_{\alpha^g} \sim \mathcal{N}\left(\widetilde{\mu}_{gik,l}, \widetilde{\sigma}^2_{gik,l}\right),$$

where

$$\begin{aligned}
-\frac{1}{2\widetilde{\sigma}^2_{gik,l}} &= -\frac{1}{2}\left(\frac{1}{\widehat{s}^2_{\gamma_{gik,l}}} + \frac{\eta_{gk,l}\beta^2_{gk,l}}{\sigma^2_{\alpha^g}} + \frac{1}{\sigma^2_{\gamma_{g,l}}}\right), \\
\frac{\widetilde{\mu}_{gik,l}}{\widetilde{\sigma}^2_{gik,l}} &= \frac{\widehat{\gamma}_{gik,l}}{\widehat{s}^2_{\gamma_{gik,l}}} + \frac{\eta_{gk,l}\beta_{gk,l}\left(\Gamma_i^g - \sum_{(k',l')\neq(k,l)}\beta_{gk',l'}\eta_{gk',l'}\gamma_{gik',l'}\right)}{\sigma^2_{\alpha^g}}.
\end{aligned}$$

The conditional posterior distributions of $\beta_{gk,l}$ are from normal distributions,

$$\beta_{gk,l} \mid \Gamma_i^g,\ \gamma_{gi\cdot,\cdot},\ \eta_{g\cdot,\cdot},\ \sigma^2_{\alpha^g},\ \sigma^2_{\beta_{g,l}} \sim (1-\eta_{gk,l})\,\mathcal{N}\left(0, \sigma^2_{\beta_{g,l}}\right) + \eta_{gk,l}\mathcal{N}\left(\mu_{\beta_{gk,l}}, \sigma^2_{\beta_{gk,l}}\right), \quad (3.7)$$

where

$$-\frac{1}{2\sigma^2_{\beta_{gk,l}}} = -\frac{1}{2}\left(\frac{\eta_{gk,l}\sum_{i=1}^{I_g}\gamma^2_{gik,l}}{\sigma^2_{\alpha^g}} + \frac{1}{\sigma^2_{\beta_{g,l}}}\right),$$
$$\frac{\mu_{\beta_{gk,l}}}{\sigma^2_{\beta_{gk,l}}} = \frac{\sum_{i=1}^{I_g}\left(\Gamma^g_i - \sum_{(k',l')\neq(k,l)}\beta_{gk',l'}\eta_{gk',l'}\gamma_{gik',l'}\right)\gamma_{gik,l}}{\sigma^2_{\alpha^g}}.$$

Conditioning on the data and the other parameters in the model, the conditional posterior distribution of $\sigma^2_{\beta_{g,l}}$ is inverse-gamma,

$$\sigma^2_{\beta_{g,l}} \mid \beta_{g\cdot,l}, a_\beta, b_\beta \sim \mathcal{IG}\left(a_\beta + \frac{K_l}{2}, b_\beta + \frac{1}{2}\sum_{k=1}^{K_l}\beta^2_{gk,l}\right). \tag{3.8}$$

The conditional posterior distribution of $\sigma^2_{\gamma_{g,l}}$ is also inverse-gamma,

$$\sigma^2_{\gamma_{g,l}} \mid \gamma_{g\cdot\cdot,l}, a_\gamma, b_\gamma \sim \mathcal{IG}\left(a_\gamma + \frac{I_g K_l}{2}, b_\gamma + \frac{1}{2}\sum_{k=1}^{K_l}\boldsymbol{\gamma}^\top_{gk,l}\boldsymbol{\gamma}_{gk,l}\right). \tag{3.9}$$

The conditional posterior distribution of $\sigma^2_{\alpha^g}$ is also inverse-gamma,

$$\begin{aligned}&\sigma^2_{\alpha^g} \mid \Gamma^g_{\cdot}, \gamma_{g\cdot\cdot,\cdot}, \ \eta_{g\cdot,\cdot}, \ \beta_{g\cdot,\cdot}, a_\alpha, b_\alpha \\ \sim&\mathcal{IG}\left(a_\alpha + \frac{I_g}{2}, b_\alpha + \frac{1}{2}\left(\boldsymbol{\Gamma}^g - \sum_{l=1}^{L}\sum_{k=1}^{K_l}\eta_{gk,l}\beta_{gk,l}\boldsymbol{\gamma}_{gk,l}\right)^\top\left(\boldsymbol{\Gamma}^g - \sum_{l=1}^{L}\sum_{k=1}^{K_l}\eta_{gk,l}\beta_{gk,l}\boldsymbol{\gamma}_{gk,l}\right)\right).\end{aligned} \tag{3.10}$$

The conditional posterior of $\pi_{gk,l}$ is a Beta distribution:

$$\pi_{gk,l} \mid \eta_{gk,l}, a_\pi, b_\pi \sim \text{Beta}\left(a_\pi + \eta_{gk,l}, b_\pi + 1 - \eta_{gk,l}\right). \tag{3.11}$$

The conditional probability of $\boldsymbol{\eta}_{g,l}$ given $\boldsymbol{\Gamma}^g$ can be written using Bayes' theorem:

$$\Pr\left(\eta_{gk,l}=1 \mid \boldsymbol{\Gamma}^g\right)=\frac{\Pr\left(\eta_{gk,l}=1\right) p\left(\boldsymbol{\Gamma}^g \mid \eta_{gk,l}=1\right)}{\Pr\left(\eta_{gk,l}=0\right) p\left(\boldsymbol{\Gamma}^g \mid \eta_{gk,l}=0\right)+\Pr\left(\eta_{gk,l}=1\right) p\left(\boldsymbol{\Gamma}^g \mid \eta_{gk,l}=1\right)} \tag{3.12}$$

where

$$\begin{aligned}
&p\left(\boldsymbol{\Gamma}^g \mid \eta_{gk,l}=1\right)\\
=&\frac{1}{\sigma_{\alpha^g}\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_{\alpha^g}^2}\left(\boldsymbol{\Gamma}^g-\beta_{gk,l}\boldsymbol{\gamma}_{gk,l}-\sum_{(k',l')\neq(k,l)} \eta_{gk',l'}\beta_{gk',l'}\boldsymbol{\gamma}_{gk',l'}\right)^{\top}\left(\boldsymbol{\Gamma}^g-\beta_{gk,l}\boldsymbol{\gamma}_{gk,l}-\sum_{(k',l')\neq(k,l)} \eta_{gk',l'}\beta_{gk',l'}\boldsymbol{\gamma}_{gk',l'}\right)\right\},\\
&p\left(\boldsymbol{\Gamma}^g \mid \eta_{gk,l}=0\right)\\
=&\frac{1}{\sigma_{\alpha^g}\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma_{\alpha^g}^2}\left(\boldsymbol{\Gamma}^g-\sum_{(k',l')\neq(k,l)} \eta_{gk',l'}\beta_{gk',l'}\boldsymbol{\gamma}_{gk',l'}\right)^{\top}\left(\boldsymbol{\Gamma}^g-\sum_{(k',l')\neq(k,l)} \eta_{gk',l'}\beta_{gk',l'}\boldsymbol{\gamma}_{gk',l'}\right)\right\}.
\end{aligned}$$

For the l-th exposure, we apply the CCA (when $L=2$) or GCCA (when $L>2$) to the standardized modulation matrices $\boldsymbol{U}_l=\left\{\text{logit}(\pi_{gk,l})-u_{0k,l}\right\}^{G\times K_l}$, $l=1,\cdots,L$. CCA/GCCA aims to maximize the pair-wise correlation between linear combinations of $\boldsymbol{U}_l$'s. Suppose we have rank-p approximation for $\boldsymbol{U}_l$, the corresponding canonical weight matrices $\boldsymbol{A}_l=\left[\boldsymbol{a}_l^1,\ldots,\boldsymbol{a}_l^p\right]$ can be estimated. Then, for each data type l, the estimated low-rank matrix $\boldsymbol{U}_l^C=\boldsymbol{U}_l\boldsymbol{A}_l\boldsymbol{A}_l^{\dagger}$ with rank $\left(\boldsymbol{U}_l^C\right)=p$, where † refers to the Moore-Penrose pseudo-inverse.

To further capture the low-rank patterns in each molecular exposure data type, we perform a PCA on the residual matrix after subtracting $\boldsymbol{U}_l^C$ from $\boldsymbol{U}_l$. Specifically, we first standardize the $\boldsymbol{U}_l^{\text{res}}$'s with mean zero and unit variance. We calculate a truncated Singular Value Decomposition (SVD) and keep the top q_l largest singular values to approximate $\boldsymbol{U}_l^{\text{res}}$. The rank-$q_l$ approximation of $\boldsymbol{U}_l^{\text{res}}$ is denoted as $\boldsymbol{U}_l^R$. $\boldsymbol{U}_l^R$'s capture the association patterns shared among and specific to different cellular contexts (tissues). Accounting for omic-shared $\left(\boldsymbol{U}_l^C\right)$ and tissue-shared $\left(\boldsymbol{U}_l^R\right)$ patterns, we further update $\pi_{gk,l}$'s with $\pi_{gk,l}=1/\left(1+\exp\left(-U_{gk,l}^C-U_{gk,l}^R-u_{0k,l}\right)\right)$ (Algorithm 2).

The algorithm for correlated SNPs

For correlated SNPs, we consider the following model for the g-th set of exposures:

$$\begin{aligned}\widehat{\boldsymbol{\Gamma}}^g &\sim \mathcal{N}\left(\widehat{\mathbf{S}}_{\boldsymbol{\Gamma}^g}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\Gamma}^g}\boldsymbol{\Gamma}^g, \widehat{\mathbf{S}}_{\boldsymbol{\Gamma}^g}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\boldsymbol{\Gamma}^g}\right), \\ \widehat{\boldsymbol{\gamma}}_{gk,l} &\sim \mathcal{N}\left(\widehat{\mathbf{S}}_{\gamma_{gk,l}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\gamma_{gk,l}}\boldsymbol{\gamma}_{gk,l}, \widehat{\mathbf{S}}_{\gamma_{gk,l}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}_{\gamma_{gk,l}}\right),\end{aligned} \tag{3.13}$$

where $\widehat{\mathbf{R}}^g$ is the correlation matrix of the I_g number of IVs for the g-th set of exposures, $\widehat{\boldsymbol{\Gamma}}^g = [\widehat{\Gamma}^g_1, \cdots, \widehat{\Gamma}^g_{I_g}]^\top$, $\widehat{\boldsymbol{\gamma}}_{gk,l} = [\widehat{\gamma}_{g1k,l}, \cdots, \widehat{\gamma}_{gI_gk,l}]^\top$, $\widehat{\mathbf{S}}_{\boldsymbol{\Gamma}^g} = \text{diag}(\widehat{s}_{\Gamma^g_1}, \cdots, \widehat{s}_{\Gamma^g_{I_g}})$, and $\widehat{\mathbf{S}}_{\gamma_{gk,l}} = \text{diag}(\widehat{s}_{\gamma_{g1k,l}}, \cdots, \widehat{s}_{\gamma_{gI_gk,l}})$. The conditional posterior distribution of each Γ^g_i given the other parameters in the model is

$$\Gamma^g_i \mid \widehat{\Gamma}^g_i,\ \widehat{s}_{\Gamma^g_i},\ \gamma_{gi\cdot,\cdot},\ \beta_{g\cdot,\cdot},\ \eta_{g\cdot,\cdot},\ \sigma^2_{\alpha^g},\ \widehat{R}^g_{\cdot\cdot} \sim \mathcal{N}\left(\widetilde{\mu}_{gi0}, \widetilde{\sigma}^2_{gi0}\right),$$

where

$$\begin{aligned}-\frac{1}{2\widetilde{\sigma}^2_{gi0}} &= -\frac{1}{2}\left(\frac{1}{\widehat{s}^2_{\Gamma^g_i}} + \frac{1}{\sigma^2_{\alpha^g}}\right), \\ \frac{\widetilde{\mu}_{gi0}}{\widetilde{\sigma}^2_{gi0}} &= \frac{\widehat{\Gamma}^g_i}{\widehat{s}^2_{\Gamma^g_i}} - \sum_{j\neq i}\left(\frac{\widehat{R}^g_{ij}\Gamma^g_j}{\widehat{s}_{\Gamma^g_j}}\right)\frac{1}{\widehat{s}_{\Gamma^g_i}} + \frac{\sum_{l=1}^{L}\sum_{k=1}^{K_l}\eta_{gk,l}\beta_{gk,l}\gamma_{gik,l}}{\sigma^2_{\alpha^g}}.\end{aligned}$$

Here $\widehat{\mathbf{R}}^g$ is the estimated correlation matrix among all selected IVs of exposure set g. Conditioning on other parameters, the distribution for each element $\gamma_{gik,l}$ comes from a normal distribution with

$$\gamma_{gik,l} \mid \widehat{\gamma}_{gik,l},\ \widehat{s}_{\gamma_{gik,l}},\ \Gamma^g_i,\ \gamma_{g\cdot k,l},\ \beta_{g\cdot,\cdot},\eta_{g\cdot,\cdot},\ \sigma^2_{\gamma_{g,l}},\ \sigma^2_{\alpha^g},\ \widehat{R}^g_{\cdot\cdot} \sim \mathcal{N}\left(\widetilde{\mu}_{gik,l}, \widetilde{\sigma}^2_{gik,l}\right),$$

where

$$-\frac{1}{2\widetilde{\sigma}^2_{gik,l}} = -\frac{1}{2}\left(\frac{1}{\widehat{s}^2_{\gamma_{gik,l}}} + \frac{\eta_{gk,l}\beta^2_{gk,l}}{\sigma^2_{\alpha^g}} + \frac{1}{\sigma^2_{\gamma_g}}\right),$$

$$\frac{\widetilde{\mu}_{gik,l}}{\widetilde{\sigma}^2_{gik,l}} = \frac{\widehat{\gamma}_{gik,l}}{\widehat{s}^2_{\gamma_{gik,l}}} - \sum_{j\neq i}\left(\frac{\widehat{R}^g_{ij}\gamma_{gjk,l}}{\widehat{s}_{\gamma_{gjk,l}}}\right)\frac{1}{\widehat{s}_{\gamma_{gik,l}}} + \frac{\eta_{gk,l}\beta_{gk,l}\left(\Gamma^g_i - \sum_{(k',l')\neq(k,l)}\beta_{gk',l'}\eta_{gk',l'}\gamma_{gik',l'}\right)}{\sigma^2_{\alpha^g}}.$$

The updates for the remaining parameters are the same as in (3.7)-(3.11).

The algorithm accounting for sample overlap

We further consider sample overlap among tissues, molecular traits, and complex traits. We could rewrite the distribution for the summary statistics in (3.13) as the following matrix normal distribution for Z-score

$$\left[\widehat{\mathbf{S}}^{-1}_{\mathbf{\Gamma}^g}\widehat{\mathbf{\Gamma}}^g,\ \underbrace{\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{g1,1}}\widehat{\boldsymbol{\gamma}}_{g1,1},\ \cdots,\ \widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{gK_1,1}}\widehat{\boldsymbol{\gamma}}_{gK_1,1}}_{\text{Exposure 1, context 1-}K_1},\cdots,\ \underbrace{\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{g1,L}}\widehat{\boldsymbol{\gamma}}_{g1,L},\ \cdots,\ \widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{gK_L,L}}\widehat{\boldsymbol{\gamma}}_{gK_L,L}}_{\text{Exposure L, context 1-}K_L}\right]$$
$$\sim\mathcal{MN}\left(\left[\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\mathbf{\Gamma}^g}\mathbf{\Gamma}^g,\ \widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{g1,1}}\boldsymbol{\gamma}_{g1,1},\ \cdots,\ \widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{gK_L,L}}\boldsymbol{\gamma}_{gK_L,L}\right],\widehat{\mathbf{R}}^g,\ \widehat{\boldsymbol{C}}\right),$$

where $\widehat{\boldsymbol{C}} \in \mathbb{R}^{(1+\sum_{l=1}^L K_l)\times(1+\sum_{l=1}^L K_l)}$ is the correlation matrix that accounts for sample overlap among outcome and the $\sum_{l=1}^L K_l$ contexts of the L exposures. The matrix $\widehat{\boldsymbol{C}}$ can be estimated separately using summary statistics among independent variants with no associations to either exposure or outcome diseases/traits. Equivalently, it can be written as a multivariate normal distribution as:

$$\begin{pmatrix}\widehat{\mathbf{\Gamma}}^g\\ \widehat{\boldsymbol{\gamma}}_{g1,1}\\ \vdots\\ \widehat{\boldsymbol{\gamma}}_{gK_L,L}\end{pmatrix}\sim\mathcal{N}\left(\begin{pmatrix}\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\mathbf{\Gamma}^g}\mathbf{\Gamma}^g\\ \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{g1,1}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{g1,1}}\boldsymbol{\gamma}_{g1,1}\\ \vdots\\ \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{gK_L,L}}\widehat{\mathbf{R}}^g\widehat{\mathbf{S}}^{-1}_{\boldsymbol{\gamma}_{gK_L,L}}\boldsymbol{\gamma}_{gK_L,L}\end{pmatrix},\begin{pmatrix}\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{g1,1}} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{gK_L,L}}\end{pmatrix}\left(\widehat{\boldsymbol{C}}\otimes\widehat{\mathbf{R}}^g\right)\begin{pmatrix}\widehat{\mathbf{S}}_{\mathbf{\Gamma}^g} & \mathbf{0} & \cdots & \mathbf{0}\\ \mathbf{0} & \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{g1,1}} & \cdots & \mathbf{0}\\ \vdots & \vdots & \ddots & \vdots\\ \mathbf{0} & \mathbf{0} & \cdots & \widehat{\mathbf{S}}_{\boldsymbol{\gamma}_{gK_L,L}}\end{pmatrix}\right).$$

We denote $\mathbf{\Lambda} = \{\lambda_{ij}\} = \widehat{\boldsymbol{C}}^{-1}$ $(i, j = 0, 1, \cdots, \sum_{l=1}^{L} K_l)$. For simplicity, in following equations we use $\lambda_{0k,l} = \lambda_{0(k+\sum_{t<l} K_t)}$. It represents the value in $\mathbf{\Lambda}$ for the sample overlap between the outcome and the k-th context of the l-th exposure. We use $\lambda_{kk',ll'} = \lambda_{(k+\sum_{t<l} K_t)(k'+\sum_{t<l'} K_t)}$. It represents the value in $\mathbf{\Lambda}$ for the sample overlap between the k-th context of the l-th exposure and the k'-th context of the l'-th exposure. Through some derivations, the conditional posterior distribution of each Γ_i^g given the other parameters in the model is

$$\Gamma_i^g \mid \widehat{\Gamma}_{\cdot}^g,\ \widehat{\gamma}_{g\cdot\cdot,\cdot},\ \gamma_{g\cdot\cdot,\cdot},\ \widehat{s}_{\Gamma_{\cdot}^g},\ \widehat{s}_{\gamma_{g\cdot\cdot,\cdot}},\ \beta_{g\cdot,\cdot},\ \eta_{g\cdot,\cdot},\ \sigma_{\alpha^g}^2,\ \widehat{R}_{\cdot\cdot}^g,\ \lambda_{\cdot\cdot,\cdot\cdot} \sim \mathcal{N}\left(\widetilde{\mu}_{gi0}, \widetilde{\sigma}_{gi0}^2\right), \quad (3.14)$$

where

$$\begin{aligned}
-\frac{1}{2\widetilde{\sigma}_{gi0}^2} =& -\frac{1}{2}\left(\frac{\lambda_{00}}{\widehat{s}_{\Gamma_i^g}^2} + \frac{1}{\sigma_{\alpha^g}^2}\right) \\
\frac{\widetilde{\mu}_{gi0}}{\widetilde{\sigma}_{gi0}^2} =& \lambda_{00}\left\{\frac{\widehat{\Gamma}_i^g}{\widehat{s}_{\Gamma_i^g}^2} - \sum_{j\neq i}\left(\frac{\widehat{R}_{ij}^g \Gamma_j^g}{\widehat{s}_{\Gamma_j^g}}\right)\frac{1}{\widehat{s}_{\Gamma_i^g}}\right\} + \sum_{l=1}^{L}\sum_{k=1}^{K_l}\frac{\widehat{\gamma}_{gik,l}}{\widehat{s}_{\gamma_{gik,l}}}\cdot\frac{\lambda_{0k,l}}{\widehat{s}_{\Gamma_i^g}} \\
& - \sum_{l=1}^{L}\sum_{k=1}^{K_l}\sum_{j=1}^{I_g}\frac{\widehat{R}_{ij}^g \gamma_{gjk,l}}{\widehat{s}_{\gamma_{gjk,l}}}\cdot\frac{\lambda_{0k,l}}{\widehat{s}_{\Gamma_i^g}} + \frac{\sum_{l=1}^{L}\sum_{k=1}^{K_l}\eta_{gk,l}\beta_{gk,l}\gamma_{gik,l}}{\sigma_{\alpha^g}^2}.
\end{aligned}$$

The conditional distribution for each element $\gamma_{gik,l}$ comes from a normal distribution with

$$\gamma_{gik,l} \mid \widehat{\Gamma}_{\cdot}^g,\ \Gamma_{\cdot}^g,\ \widehat{\gamma}_{g\cdot\cdot,\cdot},\ \widehat{s}_{\Gamma_{\cdot}^g},\ \widehat{s}_{\gamma_{g\cdot\cdot,\cdot}},\ \beta_{g\cdot,\cdot}, \eta_{g\cdot,\cdot},\ \sigma_{\alpha^g}^2,\ \widehat{R}_{\cdot\cdot}^g,\ \lambda_{\cdot\cdot,\cdot\cdot} \sim \mathcal{N}\left(\widetilde{\mu}_{gik,l}, \widetilde{\sigma}_{gik,l}^2\right), \quad (3.15)$$

where

$$
\begin{aligned}
-\frac{1}{2\widetilde{\sigma}^2_{gik,l}} =& -\frac{1}{2}\left(\frac{\lambda_{kk,ll}}{\widehat{s}^2_{\gamma_{gik,l}}} + \frac{\eta_{gk,l}\beta^2_{gk,l}}{\sigma^2_{\alpha^g}} + \frac{1}{\sigma^2_{\boldsymbol{\gamma}_{g,l}}}\right), \\
\frac{\widetilde{\mu}_{gik,l}}{\widetilde{\sigma}^2_{gik,l}} =& \frac{\widehat{\Gamma}^g_i}{\widehat{s}_{\Gamma^g_i}} \cdot \frac{\lambda_{0k,l}}{\widehat{s}_{\gamma_{gik,l}}} + \sum_{l'=1}^{L}\sum_{k'=1}^{K_l} \frac{\widehat{\gamma}_{gik',l'}}{\widehat{s}_{\gamma_{gik',l'}}} \cdot \frac{\lambda_{kk',ll'}}{\widehat{s}_{\gamma_{gik,l}}} \\
& -\sum_{j=1}^{I_g} \frac{\widehat{R}^g_{ij}\Gamma^g_j}{\widehat{s}_{\Gamma^g_j}} \cdot \frac{\lambda_{0k,l}}{\widehat{s}_{\gamma_{gik,l}}} - \sum_{j\neq i}\left(\frac{\widehat{R}^g_{ij}\gamma_{gjk,l}}{\widehat{s}_{\gamma_{gjk,l}}}\right)\frac{\lambda_{kk,ll}}{\widehat{s}_{\gamma_{gik,l}}} - \sum_{(k',l')\neq(k,l)}\sum_{i'=1}^{I_g} \frac{\widehat{R}^g_{ii'}\gamma_{gi'k',l'}}{\widehat{s}_{\gamma_{gi'k',l'}}} \cdot \frac{\lambda_{kk',ll'}}{\widehat{s}_{\gamma_{gik,l}}} \\
& + \frac{\eta_{gk,l}\beta_{gk,l}\left(\Gamma^g_i - \sum_{(k',l')\neq(k,l)}\beta_{gk',l'}\eta_{gk',l'}\gamma_{gik',l'}\right)}{\sigma^2_{\alpha^g}}.
\end{aligned}
$$

The updates for the remaining parameters are the same as in (3.7)-(3.11).

3.4 Simulations

3.4.1 Data generation

In the simulation studies discussed in the main text, we simulated individual-level data for N^y individuals in the GWAS study of outcome and $N^x_{gk,l}$ individuals for tissue k of exposure l in the multi-tissue QTL studies. In most simulations, we set $N^y = 50,000$ and $N^x_{gk,l} = 500$. We simulated an $N^y \times I_g$ genotype matrix $\boldsymbol{Q}_g$ for each gene-CpG pair g with $I_g = 15$. The minor allele frequency (MAF) of each SNP follows Unif$(0.05, 0.5)$. The correlation between SNPs i and j is $r_{ij} = r^{|i-j|}$, where $r = 0$ in most simulations. To be consistent with the prevalent pleiotropy in TWMR analysis [Yang et al., 2017], all generated SNPs have a direct effect on the complex trait not via gene expression or DNA methylation. We also generated a $G \times K_l$ matrix of binary indicators $\boldsymbol{\eta}_l$ for each exposure l, where $\eta_{gk,l} \sim$ Bernoulli$(\pi_{gk,l})$. We simulated the outcome in the GWAS study according to the following data generation

Algorithm 2 The Gibbs sampling algorithm for mintMR model

Input data: $\widehat{\gamma}_{gik,l}$, $\widehat{s}_{\gamma_{gik,l}}$, $\widehat{\Gamma}_i^g$, $\widehat{s}_{\Gamma_i^g}$, $\widehat{\boldsymbol{R}}^g \in \mathbb{R}^{I_g \times I_g}$, $\widehat{\boldsymbol{C}} \in \mathbb{R}^{\sum_l K_l \times \sum_l K_l}$, $p \in \mathbb{Z}^+$, $q_l \in \mathbb{Z}^+$, for $g = 1, \cdots, G$, $l = 1, \cdots, L$, $i = 1, \cdots, I_g$, $k = 1, \cdots, K_l$.
Initialize parameters: $\Gamma_i^g, \gamma_{gik,l}, \sigma^2_{\beta_{g,l}}$, $\sigma^2_{\gamma_{g,l}}$, $\sigma^2_{\alpha^g}, U_{gik,l}$, and specify $u_{0k,l}$, for $g = 1, \cdots, G$, $l = 1, \cdots, L$, $i = 1, \cdots, I_g$, $k = 1, \cdots, K_l$. This can be either user-specified or obtained by running the Gibbs Sampling algorithm for the starting model by skipping steps 16-28.
for each iteration **do**
    for $g = 1$ to G **do**
        for $i = 1$ to I_g **do**
            Sample Γ_i^g using (3.14).
            for $l = 1$ to L **do**
                for $k = 1$ to K_l **do**
                    Sample $\gamma_{gik,l}$ using (3.15).
        for $l = 1$ to L **do**
            for $k = 1$ to K_l **do**
                Sample $\beta_{gk,l}$ using (3.7).
                Sample $\pi_{gk,l}$ using (3.11).
                Sample $\eta_{gk,l}$ using (3.12).
            Sample $\sigma^2_{\beta_{g,l}}$, $\sigma^2_{\gamma_{g,l}}$, $\sigma^2_{\alpha^g}$ using (3.9) and (3.10).
    } Model for single gene region
    for $l = 1$ to L **do**
        $\boldsymbol{U}_l = \left\{ \log \frac{\pi_{gk,l}}{1-\pi_{gk,l}} - u_{0k,l}, 1 \le g \le G,\ 1 \le k \le K_l \right\}$
    Perform a CCA on the L standardized $\boldsymbol{U}_l$'s.
    for $l = 1$ to L **do**
        Get the coefficient matrices $\boldsymbol{A}_l$'s based on the top p canonical coefficients. $\boldsymbol{U}_l^C = \boldsymbol{U}_l \boldsymbol{A}_l \boldsymbol{A}_l^\dagger$.
        Calculate the residual matrix: $\boldsymbol{U}_l^{\text{res}} = \boldsymbol{U}_l - \boldsymbol{U}_l^C$.
        Perform a PCA on each $\boldsymbol{U}_l^{\text{res}}$ using SVD and get the low-rank approximation matrix $\boldsymbol{U}_l^R$.
    } Multi-view learning across G regions
    for $g = 1$ to G **do**
        for $l = 1$ to L **do**
            for $k = 1, \ldots, K_l$ **do**
                $\pi_{gk,l} = \frac{1}{1+\exp\left(-U^C_{gk,l} - U^R_{gk,l} - u_{0k,l}\right)}$.
} Iterate
Until the maximum iteration is reached.

models:

$$X_{gk,l} = \boldsymbol{Q}_g \boldsymbol{\mu}^x_{gk,l} + \mu^{zx}_{gk,l} Z_g + \varepsilon_{gk,l}, \tag{3.16}$$

$$Y = \sum_{g=1}^{G} \sum_{l=1}^{L} \sum_{k=1}^{K_l} \eta_{gk,l} \cdot \beta_{gk,l} X_{gk,l} + \sum_{g=1}^{G} \boldsymbol{Q}_g \boldsymbol{\mu}^y_g + \sum_{g=1}^{G} \mu^{zy}_g Z_g + \varepsilon. \tag{3.17}$$

In model (3.16), the vector $X_{gk,l}$, of length N^y, corresponds to the values of exposure l in tissue k for the gene-CpG pair indexed by g. $\boldsymbol{\mu}^x_{gk,l} = [\mu^x_{g1k,l}, \cdots, \mu^x_{gI_gk,l}]^\top$ is the QTL effect of eSNPs. For the i-th SNP, we sampled $\boldsymbol{\mu}^x_{gi\cdot,\cdot} = [\mu^x_{gi1,1}, \cdots, \mu^x_{giK_L,L}]^\top$ from $\mathcal{N}(0, \boldsymbol{\Sigma}_{\mu_x})$, where the diagonal elements of $\boldsymbol{\Sigma}_{\mu_x}$ are set to 0.3 and the off-diagonal elements are fixed at 0.03, indicating a correlation coefficient of 0.1 across exposures and tissues. Z_g is a vector of a latent confounder sampled from $\mathcal{N}(0, 1)$; $\mu^{zx}_{gk,l} \sim \mathcal{N}(0, 0.1)$ is the effect of the confounder on exposures; $\varepsilon_{gk,l}$ is the error term sampled from $\mathcal{N}(0, \sigma^2_{e_x})$, with errors from different exposures and tissues being correlated with a coefficient of 0.1. In model (3.17), Y is a vector of a continuous trait; $\beta_{gk,l} \sim \mathcal{N}(0, \sigma^2_\beta)$ is the effect of exposure l in tissue k of the g-th gene-CpG pair on the outcome; $\eta_{gk,l}$ is an binary effect indicator following Bernoulli($\pi_{gk,l}$); $\boldsymbol{\mu}^y_g \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\mu_y})$ is the vector of direct effects of SNPs on Y, with $\boldsymbol{\Sigma}_{\mu_y}$ being a diagonal matrix; $\mu^{zy}_g \sim \mathcal{N}(0, 0.1)$ is the effect of the confounder on the outcome; and $\varepsilon \sim \mathcal{N}(0, \sigma^2_{e_y})$.

We generated the QTL data based on the following model:

$$\widetilde{X}_{gk,l} = \widetilde{\boldsymbol{Q}}_g \left(\boldsymbol{\mu}^x_{gk,l} \circ \boldsymbol{\delta}_{gk,l} \right) + \mu^{zx}_{gk,l} \widetilde{Z}_g + \widetilde{\varepsilon}_{gk,l}, \tag{3.18}$$

where $\widetilde{X}_{gk,l}$ is a vector of length $N^x_{gk,l}$, representing the value of exposure l in tissue k from the QTL study; $\widetilde{\boldsymbol{Q}}_g$ is a $N^x_{gk,l} \times I_g$ genotype matrix of the I_g eSNPs in the QTL study; $\boldsymbol{\delta}_{gk,l}$ is a vector of I_g Bernoulli variables for eSNPs, where $\delta_{gk,l} = 1$ indicates that the eSNP shares consistent QTL effects between the k-th tissue of the QTL study and the GWAS

study. In most simulations, $\Pr(\delta_{gk,l} = 1) = 0.8$. The operator $\circ$ is the Hadamard product operator. With the simulated individual-level data, we calculated the marginal QTL and GWAS summary statistics as the input for the MR analyses.

3.4.2 MintMR improves estimation of sparse effects across genes via multi-view learning

We conducted simulation studies to evaluate the performance of mintMR in comparison with existing univariable MR (UVMR) and MVMR methods in various scenarios.

We simulated individual-level data for the GWAS study of outcome and multi-tissue QTL studies for each exposure. We simulated a genotype matrix for each gene-CpG pair g, with all generated SNPs having uncorrelated horizontal pleiotropy (UHP) effects on the simulated outcome not via exposures. We varied the proportion of the variance in the outcome that can be explained by these UHP effects. We then generated matrices of disease-relevance tissue indicators $\boldsymbol{\eta}_l$'s, where $\eta_{gk,l} \sim \text{Bernoulli}(\pi_{gk,l})$. Outcome variables in the GWAS study were simulated according to the data generation models in (3.17). QTL data were simulated based on the model (3.18). With generated individual-level data, we calculated the marginal QTL and GWAS summary statistics as the input for MR analyses.

Most existing MR methods were developed to analyze complex traits as exposure. In TWMR, the number of cis-eQTLs as IVs for gene expression as exposure is generally much smaller than the number of IVs in conventional MR analyses. Our simulation studies show that the limited number of IVs poses a challenge for existing MR methods. We compared mintMR with existing multivariable methods, including MVMR-IVW [Burgess and Thompson, 2015], MVMR-Egger [Rees et al., 2017], MVMR-Lasso [Grant and Burgess, 2021], MVMR-Median [Grant and Burgess, 2021], MVMR-Robust [Grant and Burgess, 2021] and MVcML [Lin et al., 2023]. In addition, we included IVW with cross-tissue IVs and IV effects being estimated based on a meta-analysis of multiple tissue types (termed as "IVW+metaIV"

below) and MR-Egger in the comparison. Among those competing methods, IVW and MVMR-IVW do not allow invalid IVs [Slob and Burgess, 2020, Burgess and Thompson, 2015]; MR-Egger and MVMR-Egger require Instrument Strength Independent of Direct Effect (InSIDE) assumption [Bowden et al., 2015, Rees et al., 2017]; MVMR-Median assumes the majority of IVs are valid [Grant and Burgess, 2021]; MVMR-Lasso and MVMR-Robust are robust to outliers (few invalid IVs) [Grant and Burgess, 2021]; and MVcML requires plurality condition where the valid IVs form the largest group to give the causal parameter estimate [Lin et al., 2023]. All existing UVMR and MVMR methods are developed for using complex traits as exposures. Here we adapted them to TWMR with molecular traits as exposures for comparison purposes. Moreover, we compared the proposed mintMR with its two variations: $\text{mintMR}_{\text{oracle}}$ is a variation of mintMR where the true latent disease-relevance indicator is known, and it provides the optimal performance of mintMR, which in practice cannot be achieved without further information on disease-relevance indicators; and $\text{mintMR}_{\text{single-gene}}$ performs the starting model (3.3) for each single gene region separately without the joint learning of shared patterns, and its comparison to the proposed mintMR illustrates the improvement gained by jointly learning low-rank disease-relevance patterns across multiple gene regions, tissues, and molecular exposures. We applied competing MVMR methods with multiple tissues of both simulated expression and DNA methylation as exposures and applied MR-Egger with a single tissue of gene expression as exposure to evaluate their performance. We presented the comparison of type I error rates and powers of the proposed mintMR versus competing methods at the p-value threshold of 0.05. To evaluate the estimation of effect sizes, we also compared the root-mean-square errors (RMSEs) of all methods.

For each simulation, we generated $G = 50$ pairs of genes and CpGs ($L = 2$) from 5 tissues ($K_l = 5$, $l = 1, 2$), each with 500 samples. We generated 15 IVs for each gene-CpG pair and included IVs with p-value< 0.01 in at least one tissue. We simulated two types

of causal effects of genes on outcomes. In the first setting (Table 3.1a), we simulated genes having effects on outcome in multiple tissues, with effect indicators $\eta_{gk,l}$'s having the same probability ($\pi_{gk,l} = 0.05$) across all tissues. In the second setting (Table 3.1b), 15% of the genes have non-zero effects on outcome in one tissue ($\pi_{gk,l} = 0.15$). In each of the rest tissues, 3% of the genes have non-zero effects with $\pi_{gk',l} = 0.03$. We varied the proportion of the variation in the outcome explained by UHP effects of the $\sum_{g=1}^{G} I_g$ IVs of all G genes from 0.05 to 0.15. As shown in both Table 3.1a and Table 3.1b, when the number of IVs was limited and all IVs had UHP [Morrison et al., 2020], the proposed mintMR model could control type I error rate. Most competing methods, including $\text{mintMR}_{\text{single-gene}}$, suffered from inflated type I error rates. Most of the competing methods showed increases in type I error rates when UHP effects increased. MVMR-Robust had reasonable control of type I error rates but suffered from low power. When the proportion of the variation in the outcome explained by UHP effects increased, the powers of all methods decreased. The proposed mintMR method had comparable power to the oracle method, $\text{mintMR}_{\text{oracle}}$. These simulation results, in particular the comparisons of mintMR with $\text{mintMR}_{\text{oracle}}$ and $\text{mintMR}_{\text{single-gene}}$, suggested that multi-view learning of shared patterns across multiple genes can effectively improve the estimation of latent disease-relevant probabilities, which leads to the improved estimation of the causal effects of interest. In Table 3.2, we showed that mintMR had the smallest RMSE among all the methods. In both settings, the multi-view learning of low-rank patterns of causal effects improved the power and precision when the number of IVs was limited.

In Table 3.3, we compared these methods in different scenarios. In Table 3.3a, we increased the number of IVs from 15, 25, to 100. When the number of IVs increased, all competing methods could better control the type I error rates. When the number of IVs was 100, all MVMR methods had reasonable control of the type I error rates. The power and RMSE (Table 3.4) of all competing MVMR methods were similar to the proposed mintMR and mintMR single-gene version. The univariable MR methods IVW and Egger still had

slightly inflated type I error rates and low power due to the omission of correlated exposures. These competing methods were proposed for analyzing complex traits as exposures, and the number of IVs in conventional MR analyses is usually much larger than the number of cis-QTLs as IVs in TWMR analyses. In other words, while existing MR methods work effectively for complex trait exposures, they may not perform as well in TWMR analyses, and our proposed mintMR was tailored for analyzing molecular traits as exposures from multiple contexts/tissues. In Table 3.3b, we varied the probability of QTL effect sharing. When the probability decreased, eQTL/IV effects became more context/tissue-specific and the consistency of IV effects decreased. Table 3.3b showed that when the consistency of QTL effects across the QTL and GWAS sample decreased, power was reduced for all methods due to the inclusion of many inconsistent IVs. Conversely, the power improved when more QTLs with tissue-shared effects were selected as IVs. This simulation underscores the importance of considering multiple tissues and selecting QTLs with consistent effects across more than one tissue as IVs. In Table 3.3c, we varied the number of tissues for each exposure. When the number of tissues increased, mintMR showed improved power as more IVs were included.

In Table 3.5, we showed that mintMR had the smallest RMSEs when varying the consistency of QTL effect and the number of tissues. In Table 3.6, we increased the sample size from 500 to 10,000 for each tissue type in the presence of UHP. The larger tissue sample size improved the estimation of the IV-to-exposure effects, while also making the impacts of invalid IVs stronger. The performance of competing methods was similar for different sample sizes. In Table 3.7, we varied the causal effect size, the proposed mintMR method controlled the type I error rate and showed improved power compared with other methods. MintMR had the smallest RMSEs on data with varied sample sizes and effect sizes (Table 3.9). In addition, we simulated correlated IVs with genetic correlation up to 0.5. When the IVs were correlated and the numbers of IVs were limited, the proposed mintMR could still control the type I error rate and showed reasonable power (Table 3.8).

3.5 Data analysis: Identifying trait/disease risk-associated genes via mintMR

We applied the proposed mintMR method to map risk genes for 35 complex traits and diseases, including 14 immunological traits, 6 metabolic traits, 2 neurological diseases, 2 cardiovascular traits, 7 psychiatric diseases and traits, and 4 other traits. We used GWAS statistics as the IV-to-outcome statistics. Details of the GWAS statistics can be found in Table S11. We used multi-tissue eQTL and mQTL summary statistics as the IV-to-exposure statistics. For eQTLs, we obtained the summary statistics for blood tissue from the eQTLGen consortium [Võsa et al., 2021] ($N = 31,684$), for muscle tissue ($N = 706$), lung tissue ($N = 515$) and brain cerebellum tissue ($N = 209$) from version 8 of the Genotype-Tissue Expressions (GTEx) project [Consortium, 2020], and for brain dorsolateral prefrontal cortex tissue from the Religious Orders Study and Memory and Aging Project (ROSMAP; $N = 560$) [Bennett et al., 2018]. For mQTLs, we obtained the summary statistics for lung tissue from GTEx [Oliva et al., 2023] ($N = 190$), skeletal muscle tissue from FUSION [Taylor et al., 2019] ($N = 265$), and blood tissue ($N = 1,366$) from Brisbane Systems Genetics Study (BSGS) [McRae et al., 2014, Powell et al., 2012] plus Lothian Birth Cohorts (LBC) [Deary et al., 2012]. For each gene, we selected the proximal CpG site (within 100 KB of TSS) that explained the most variation in expression. For each gene-CpG pair, we selected the cis-eSNPs or mSNPs with non-zero and sign-consistent eQTL or mQTL effects in at least two tissues ($P \leq 0.005$). We performed LD clumping at the r^2 threshold of 0.01. We restricted our analysis to genes with at least 10 IVs overall and at least one IV for each tissue.

We applied mintMR to each of the 35 complex traits and diseases, with an average of 3,440 genes examined for each trait/disease. At the false discovery rate (FDR) of 0.05, we identified the genes and CpG sites showing significant effects in at least two tissues for each examined trait/disease. See Table 3.12 for a list of examined traits/diseases, the number of genes studied, and the number of detected genes and CpG sites. In Table 3.11, we evaluated

the genome-wide inflation factor [Devlin and Roeder, 1999] with and without accounting for DNAm, based on the p-value distributions of gene expression in each tissue. By accounting for the most correlated cis-CpG site, genome-wide inflation is substantially reduced for all examined traits and diseases. An important message from our analysis result is that in mapping the expression of risk genes, cis-DNAm can be a major confounder if not accounted for. Existing studies showed that cis DNAm frequently correlates with cis expression and cis-eQTLs often co-occur with cis-mQTLs [Pierce et al., 2018]. If a cis-e/mQTL or a variant in LD with it is selected as an IV and cis-DNAm is not accounted for, the causal inference can be compromised due to the IVs being correlated with the confounder. In Table 3.10, we showed that mintMR had lower inflation factors than MVMR-Lasso, MVMR-Median, and MVMR-IVW. The inflation factors of mintMR and MVMR-Robust are comparable. Due to the prevalent pleiotropy in TWMR analysis, MVMR-Egger and MVMR-Robust are expected to have lower power than the other examined methods [Lin et al., 2023, Grant and Burgess, 2021]. Simulation showed that MVMR-Egger and MVMR-Robust have much lower power than mintMR when UHP is prevalent. We also note that there is remaining mild inflation in the p-values. It suggests that there are additional factors and potential IV-associated confounders that have not been fully accounted for in the analyses. This could be at least partially due to, for example, secondary cis-CpG sites, and other correlated and co-expressed cis genes in the region. The proposed mintMR model is a multivariable MR framework and it can be applied to jointly consider one or more cis gene expression and multiple CpG sites.

In Figure 3.2a, we showed the quantile-quantile (QQ) plot of negative log base 10 of p-values for gene expression effects on hypertension in the blood tissue. The genome-wide inflation factor decreased from 1.88 to 1.25 after accounting for DNAm. In the 5q31-32 region, we identified four genes (*HSPA4*, *HARS2*, *KIAA0141*, and *ARHGEF37*) showing significant effects on hypertension (FDR< 0.05) without accounting for DNAm. After adjusting for the

most correlated cis-CpG site, only the expression of *HSPA4* still showed a significant effect (Figure 3.2b). *HSPA4* is a member of the heat shock protein 70 family, which is known to be involved in the pathogenesis of hypertension [Mohamed et al., 2012, Rodriguez-Iturbe et al., 2023]. We further conducted a colocalization analysis, and only the gene *HSPA4* showed a high probability of colocalization with hypertension (PP4= 0.95) (Figure 3.2c). Additionally, we examined all the significant genes identified for hypertension at the FDR level of 0.05 in at least two tissues. Out of the 57 identified genes, 49 were analyzed in a TWAS [Gamazon et al., 2019]. Among these, 15 genes (30.6%) were also significant in the TWAS analysis ($P < 0.005$), a proportion much higher than that observed among all genes examined (14.2%). Moreover, 6 out of these 49 genes (12.2%) were supported by colocalization analyses (PP4> 0.7), a much higher proportion than all genes examined (2.3%).

We further conducted pathway analyses on the significant genes and proximal genes correlated with significant CpGs identified for each of the 35 traits and diseases, utilizing the Reactome [Fabregat et al., 2018] and Gene Ontology [Ashburner et al., 2000] database. We detected the significantly enriched biological pathways for each trait/disease, as shown in Figure 3.3. Our results revealed many enriched pathways being shared among related traits, suggesting shared mechanisms. Lipid-related pathways, including lipid localization and transport, are implied for Alzheimer's disease, monocyte count lymphocyte count, and platelet count. As the basic component of cell membranes, lipids play an important role in brain function. Impaired homeostasis of lipids is known to be related to neurologic disorders [Kao et al., 2020, Di Paolo and Kim, 2011, Li et al., 2022]. Monocytes, lymphocytes, and platelets are key components of the immune system [Schluter et al., 2020, Shi and Pamer, 2011, Scherlinger et al., 2023], and the fact that these traits share common enriched pathways with Alzheimer's disease suggests that inflammation and immune response play a significant role in Alzheimer's disease. [Heppner et al., 2015, Heneka et al., 2015]

3.6 Discussion

In this work, we propose an integrative multi-context Mendelian randomization method, mintMR, for addressing unique challenges in TWMR analysis. MintMR performs a multi-tissue MR analysis using QTLs as IVs for each gene region. It improves the estimation of tissue-specific causal effects of all genes by simultaneously modeling the latent disease-relevance context/tissue indicators for multiple gene regions, jointly learning the low-rank patterns of latent indicators/probabilities via multi-view learning techniques, and then using the major patterns to update the probability of non-zero effects. The joint learning of disease-relevance of latent tissue indicators improves the estimation of sparse tissue-specific causal effects for all genes. By selecting cross-tissue QTLs as IVs and considering both gene expression and DNAm as joint exposures, mintMR improves IV consistency and reduces confounding due to correlated cis molecular traits when mapping causal genes. Simulations show that mintMR can control the type I error rates and has good powers in various settings, even when there are a limited number of QTLs as IVs and the causal effects are sparse.

We applied mintMR to map risk genes for 35 complex traits and diseases, leveraging QTL summary statistics from multiple tissues of different studies and GWAS summary statistics. Our results showed a reasonable control of genome-wide inflation for the examined traits and diseases, demonstrating the feasibility of leveraging multi-tissue QTLs and jointly learning disease-relevance probabilities across multiple gene regions in improving causal identification. Our results also suggested DNAm being a major confounder in mapping risk genes. By accounting for cis DNAm, genome-wide inflation for TWMR analyses was substantially reduced. Our analysis and results demonstrated that mintMR could offer valuable insights into disease-relevant tissues and the underlying mechanisms.

There are several limitations of our work. First, mintMR does not allow IV to be associated with unmeasured confounders. As a multivariable MR framework, mintMR allows the adjustment and joint modeling of correlated molecular traits (confounders) as joint ex-

posures. Simulation studies show that mintMR is robust to mild violations of the InSIDE assumption. In the TWMR analysis of 35 traits and diseases, we noted some remaining mild genome-wide inflation after modeling the most correlated cis-CpG sites. In future analyses, additional correlated cis molecular traits, such as secondary cis CpG sites or nearly co-expressed genes, could also be modeled to further reduce genome-wide inflation. Second, we assume linear effects of exposures on outcome. The current mintMR model is not flexible for modeling complex interactions among exposures and interactions with known covariates, such as sex-biased effects.

In future work, mintMR can be extended to allow for correlated horizontal pleiotropy by identifying IVs with such effects. Another area of future development is to improve the modeling of major patterns of disease relevance indicators by adopting other advanced multi-view learning techniques. In this work, we used CCA and PCA to capture omics-shared and tissue-shared patterns in mapping risk genes. Other deep learning and supervised multi-view learning methods could be implemented to promote other desirable patterns among examined genes [Andrew et al., 2013, Yin and Sun, 2019, Wang et al., 2015]. Moreover, the mintMR model could be further expanded to model interaction effects among joint exposures and covariates. These developments will be explored in future works.

	Proportion of the variance in the outcome explained by UHP effects					
	0.05	0.1	0.15	0.05	0.1	0.15
	Power			Type I error rate		
mintMR	0.859	0.786	0.657	0.050	0.049	0.048
$\text{mintMR}_{\text{oracle}}$	0.903	0.842	0.773	0.048	0.050	0.048
$\text{mintMR}_{\text{single-gene}}$	0.718	0.629	0.567	0.072	0.120	0.158
IVW+metaIV	0.351	0.317	0.323	0.149	0.160	0.162
Egger	0.308	0.275	0.262	0.131	0.138	0.141
MVMR-IVW	0.663	0.534	0.473	0.121	0.128	0.134
MVMR-Egger	0.573	0.518	0.443	0.133	0.138	0.138
MVMR-Lasso	0.770	0.730	0.704	0.159	0.214	0.259
MVMR-Median	0.641	0.572	0.519	0.089	0.117	0.132
MVMR-Robust	0.455	0.374	0.315	0.062	0.076	0.080

(a)

	Proportion of the variance in the outcome explained by UHP effects					
	0.05	0.1	0.15	0.05	0.1	0.15
	Power			Type I error rate		
mintMR	0.841	0.794	0.677	0.051	0.050	0.046
$\text{mintMR}_{\text{oracle}}$	0.898	0.830	0.775	0.050	0.048	0.048
$\text{mintMR}_{\text{single-gene}}$	0.691	0.610	0.582	0.074	0.116	0.155
IVW+metaIV	0.362	0.341	0.342	0.146	0.157	0.158
Egger	0.305	0.270	0.273	0.131	0.135	0.139
MVMR-IVW	0.652	0.523	0.461	0.122	0.130	0.134
MVMR-Egger	0.510	0.441	0.384	0.121	0.136	0.141
MVMR-Lasso	0.764	0.710	0.681	0.158	0.210	0.257
MVMR-Median	0.677	0.578	0.500	0.087	0.115	0.127
MVMR-Robust	0.444	0.365	0.295	0.062	0.077	0.080

(b)

Table 3.1: Simulation results evaluating the performance of mintMR versus competing methods when the number of IVs is limited. Two types of causal effects of genes on outcomes are simulated. (a) Genes affect outcomes in multiple tissues, with each gene having an equal probability (5%) of having non-zero effects in any tissue. (b) In one tissue, 15% of the genes have non-zero effects on outcome. In each of the rest tissues, 3% of the genes have non-zero effects. The proportion of variation in outcome explained by UHP effects varies from 0.05 to 0.15. The sample size of the outcome is 50,000 and 500 for exposure. The number of IVs is 15. Two exposures are generated and each exposure has 5 tissues. The causal effects are generated with $\mathcal{N}(0, 0.015)$. Results are underlined for methods unable to control type I error rates (≥ 0.1).

	Proportion of the variance in the outcome explained by UHP effects		
	0.05	0.1	0.15
mintMR	0.027	0.033	0.038
mintMR (oracle)	0.011	0.012	0.013
mintMR (single gene)	0.081	0.092	0.107
IVW+metaIV	0.047	0.053	0.057
Egger	0.163	0.187	0.208
MVMR-IVW	0.044	0.055	0.065
MVMR-Egger	0.061	0.078	0.091
MVMR-Lasso	0.050	0.065	0.077
MVMR-Median	0.051	0.065	0.076
MVMR-Robust	0.044	0.055	0.065
MVcML	0.034	0.038	0.040

(a)

	Proportion of the variance in the outcome explained by UHP effects		
	0.05	0.1	0.15
mintMR	0.027	0.033	0.038
mintMR (oracle)	0.011	0.012	0.014
mintMR (single gene)	0.077	0.089	0.103
IVW+metaIV	0.047	0.052	0.057
Egger	0.160	0.184	0.204
MVMR-IVW	0.044	0.056	0.065
MVMR-Egger	0.066	0.084	0.099
MVMR-Lasso	0.050	0.065	0.077
MVMR-Median	0.051	0.065	0.076
MVMR-Robust	0.044	0.056	0.065
MVcML	0.034	0.038	0.040

(b)

Table 3.2: RMSE comparison of mintMR versus its variants and competing methods when IVs are limited ($I_g = 15$). Two types of causal effects of genes on outcomes are simulated. (a) Genes affect outcomes in multiple tissues, with each gene having an equal probability (5%) of having non-zero effects in any tissue. (b) In one tissue, 15% of the genes have non-zero effects on outcome. In each of the rest tissues, 3% of the genes have non-zero effects. The proportion of the variance in the outcome explained by UHP effect varies from 0.05 to 0.15. The sample size for the outcome is 50,000 and 500 for each exposure. Two exposures are generated and each exposure has 5 tissues. The causal effects are generated from $\mathcal{N}(0, 0.015)$.

	Number of IVs					
	15	25	100	15	25	100
	Power			Type I error rate		
mintMR	0.734	0.822	0.932	0.049	0.055	0.041
$\text{mintMR}_{\text{oracle}}$	0.764	0.861	0.964	0.049	0.051	0.056
$\text{mintMR}_{\text{single-gene}}$	0.547	0.789	0.963	0.115	0.109	0.059
IVW+metaIV	0.351	0.352	0.530	0.145	0.144	0.093
Egger	0.236	0.254	0.220	0.134	0.112	0.074
MVMR-IVW	0.444	0.774	0.969	0.122	0.064	0.055
MVMR-Egger	0.383	0.729	0.898	0.122	0.063	0.055
MVMR-Lasso	0.631	0.783	0.969	0.194	0.068	0.057
MVMR-Median	0.526	0.745	0.920	0.114	0.116	0.100
MVMR-Robust	0.276	0.730	0.961	0.066	0.047	0.051

(a)

	Probability of QTL effect being consistent across samples					
	0.8	0.5	0.2	0.8	0.5	0.2
	Power			Type I error rate		
mintMR	0.867	0.789	0.660	0.053	0.066	0.060
$\text{mintMR}_{\text{oracle}}$	0.898	0.852	0.746	0.049	0.054	0.050
$\text{mintMR}_{\text{single-gene}}$	0.751	0.712	0.527	0.236	0.152	0.194
IVW+metaIV	0.299	0.388	0.399	0.148	0.156	0.168
Egger	0.367	0.289	0.264	0.109	0.110	0.127
MVMR-IVW	0.682	0.572	0.436	0.122	0.106	0.067
MVMR-Egger	0.562	0.495	0.407	0.120	0.098	0.064
MVMR-Lasso	0.793	0.679	0.443	0.188	0.129	0.069
MVMR-Median	0.723	0.671	0.500	0.172	0.112	0.069
MVMR-Robust	0.513	0.432	0.324	0.067	0.052	0.036

(b)

	Number of tissues for each exposure					
	5	10	20	5	10	20
	Power			Type I error rate		
mintMR	0.553	0.713	0.739	0.052	0.037	0.062
$\text{mintMR}_{\text{oracle}}$	0.586	0.810	0.900	0.048	0.050	0.048
$\text{mintMR}_{\text{single-gene}}$	0.497	0.538	0.583	0.174	0.235	0.126
IVW+metaIV	0.321	0.286	0.281	0.146	0.157	0.125
Egger	0.240	0.180	0.192	0.138	0.124	0.105
MVMR-IVW	0.316	0.302	0.487	0.126	0.122	0.081
MVMR-Egger	0.266	0.326	0.404	0.126	0.153	0.084
MVMR-Lasso	0.612	0.364	0.493	0.303	0.145	0.081
MVMR-Median	0.418	0.225	0.480	0.135	0.076	0.103
MVMR-Robust	0.175	0.190	0.407	0.072	0.062	0.060

(c)

Table 3.3: Simulation results evaluating the performance of mintMR versus competing methods. (a) Results when varying the number of IVs. The proportion of variation in outcome explained by UHP effect is 0.1. The causal effects are generated from $\mathcal{N}(0, 0.01)$. The probability of consistency is 0.8. Five tissues are generated for each exposure. (b) Results when decreasing the probability of QTL effect being consistent. The causal effects are generated from $\mathcal{N}(0, 0.02)$. We simulated 15 IVs across 5 tissues for each exposure. (c) Results when the number of tissues increases from 5, 10, to 20. We simulated 15, 25, and 45 IVs correspondingly. The probability of consistency is 0.8. Causal effects are generated from $\mathcal{N}(0, 0.01)$. Results are underlined for methods unable to control type I error rates (≥ 0.1).

	Number of IVs					
	15	25	100	15	25	100
	Proportion of the variance in the outcome explained by UHP effects					
	0.05			0.15		
mintMR	0.032	0.025	0.017	0.037	0.029	0.018
$\text{mintMR}_{\text{oracle}}$	0.015	0.013	0.012	0.017	0.015	0.013
$\text{mintMR}_{\text{single-gene}}$	0.072	0.051	0.034	0.090	0.062	0.036
IVW+metaIV	0.048	0.046	0.028	0.053	0.051	0.030
Egger	0.178	0.142	0.138	0.202	0.158	0.148
MVMR-IVW	0.055	0.029	0.017	0.065	0.033	0.018
MVMR-Egger	0.088	0.033	0.019	0.104	0.038	0.020
MVMR-Lasso	0.063	0.029	0.017	0.076	0.034	0.018
MVMR-Median	0.063	0.034	0.020	0.074	0.040	0.021
MVMR-Robust	0.055	0.029	0.017	0.065	0.033	0.018
MVcML	0.036	0.031	0.022	0.037	0.034	0.025

Table 3.4: RMSE comparison of mintMR versus its variants and competing methods when varying the number of IVs. Two exposures are generated and each exposure has 5 tissues. The total number of IVs varies from 15 to 100. The sample size for the outcome is 50,000 and 500 for each exposure. The causal effects are simulated from $\mathcal{N}(0, 0.01)$.

	Number of tissues			Probability of QTL effect being consistent		
	5	10	15	0.8	0.5	0.2
mintMR	0.041	0.034	0.029	0.032	0.036	0.041
$\text{mintMR}_{\text{oracle}}$	0.019	0.013	0.013	0.014	0.016	0.019
$\text{mintMR}_{\text{single-gene}}$	0.106	0.073	0.051	0.093	0.080	0.076
IVW+metaIV	0.058	0.048	0.036	0.055	0.064	0.065
Egger	0.223	0.199	0.207	0.142	0.179	0.671
MVMR-IVW	0.073	0.072	0.040	0.058	0.060	0.052
MVMR-Egger	0.117	0.091	0.043	0.072	0.072	0.060
MVMR-Lasso	0.088	0.079	0.040	0.065	0.063	0.053
MVMR-Median	0.083	0.080	0.047	0.065	0.067	0.061
MVMR-Robust	0.073	0.072	0.040	0.058	0.060	0.052
MVcML	0.037	0.020	0.013	0.052	0.052	0.042

Table 3.5: RMSE comparison of mintMR versus its variants and competing methods when varying the number of tissues and the probability of having consistent effects in QTL and GWAS data for each IV. Specifically, when varying the number of tissues for each exposure from 5, 10, to 15, we generated 15, 25, and 45 IVs respectively, with the probability of consistent effects fixed at 0.8. When varying the probability of consistent effects from 0.8 to 0.2, we fixed the number of tissues for each exposure as 5. In both sets of simulations, the causal effects are generated from $\mathcal{N}(0, 0.02)$ and the proportion of the variance in the outcome explained by UHP effects is 0.1.

	Number of samples of exposures					
	500	1000	10000	500	1000	10000
	Power			Type I error rate		
mintMR	0.833	0.868	0.872	0.052	0.052	0.049
$\text{mintMR}_{\text{oracle}}$	0.878	0.880	0.922	0.045	0.049	0.052
$\text{mintMR}_{\text{single-gene}}$	0.707	0.758	0.781	0.190	0.232	0.235
IVW+metaIV	0.322	0.296	0.397	0.160	0.149	0.135
Egger	0.256	0.381	0.669	0.129	0.110	0.103
MVMR-IVW	0.653	0.686	0.688	0.127	0.117	0.102
MVMR-Egger	0.527	0.565	0.617	0.142	0.113	0.132
MVMR-Lasso	0.765	0.799	0.790	0.202	0.181	0.136
MVMR-Median	0.686	0.727	0.855	0.158	0.166	0.191
MVMR-Robust	0.462	0.514	0.512	0.058	0.062	0.049
MVcML	0.202	0.309	0.469	0.024	0.043	0.058

Table 3.6: Simulation results evaluating the performance of mintMR versus its variants and competing methods when the number of samples for each exposure varies from 500 to 10,000. Sample size for outcome is fixed at 50,000. Two exposures are generated and each exposure has 5 tissues. The number of IVs is 15. The causal effects are generated from $\mathcal{N}(0, 0.02)$ and the proportion of the variance in the outcome explained by UHP effects is 0.1. Results are underlined for methods unable to control type I error rates (≥ 0.1).

	Variance for generating causal effect					
	0.005	0.01	0.02	0.005	0.01	0.02
	Power			Type I error rate		
mintMR	0.548	0.734	0.815	0.050	0.049	0.050
$\text{mintMR}_{\text{oracle}}$	0.584	0.764	0.865	0.050	0.049	0.050
$\text{mintMR}_{\text{single-gene}}$	0.417	0.547	0.694	0.113	0.115	0.120
IVW+metaIV	0.307	0.351	0.370	0.136	0.145	0.158
Egger	0.214	0.236	0.271	0.131	0.134	0.131
MVMR-IVW	0.296	0.444	0.638	0.122	0.122	0.122
MVMR-Egger	0.276	0.383	0.508	0.124	0.122	0.126
MVMR-Lasso	0.415	0.631	0.825	0.177	0.194	0.211
MVMR-Median	0.382	0.526	0.700	0.117	0.114	0.110
MVMR-Robust	0.178	0.276	0.430	0.070	0.066	0.064

Table 3.7: Simulation results evaluating the performance of mintMR versus its variants and competing methods. Causal effects $\beta_{gk,l}$'s are generated from $\mathcal{N}\left(0, \sigma_\beta^2\right)$ and σ_β^2 varies from 0.005 to 0.02. The proportion of the variance in the outcome explained by UHP effects is 0.1. Results are underlined for methods unable to control type I error rates (≥ 0.1).

	Proportion of the variance in the outcome explained by UHP effects					
	0.05	0.1	0.15	0.05	0.1	0.15
	Power			Type I error rate		
mintMR	0.843	0.786	0.724	0.053	0.050	0.052
$\text{mintMR}_{\text{oracle}}$	0.915	0.869	0.814	0.052	0.053	0.048
$\text{mintMR}_{\text{single-gene}}$	0.644	0.627	<u>0.599</u>	0.062	0.098	<u>0.126</u>
IVW+metaIV	<u>0.399</u>	<u>0.407</u>	<u>0.407</u>	<u>0.216</u>	<u>0.234</u>	<u>0.234</u>
Egger	0.398	0.342	0.308	0.084	0.083	0.085
MVMR-IVW	0.673	0.608	0.538	0.097	0.094	0.090
MVMR-Egger	<u>0.572</u>	<u>0.535</u>	<u>0.490</u>	<u>0.118</u>	<u>0.121</u>	<u>0.120</u>
MVMR-Lasso	<u>0.732</u>	<u>0.710</u>	<u>0.690</u>	<u>0.137</u>	<u>0.153</u>	<u>0.186</u>
MVMR-Median	0.693	0.616	0.532	0.064	0.082	0.092
MVMR-Robust	0.508	0.430	0.373	0.045	0.046	0.044

Table 3.8: Simulation results evaluating the performance of mintMR versus its variants and competing methods when IVs are correlated with genetic correlation up to 0.5. Causal effects are generated from $\mathcal{N}(0, 0.02)$. Results are underlined for methods unable to control type I error rates (≥ 0.1).

Method	Tissue sample size in one tissue			Variance for generating causal effect		
	500	1000	10000	0.005	0.01	0.02
mintMR	0.032	0.025	0.017	0.030	0.032	0.035
$\text{mintMR}_{\text{oracle}}$	0.015	0.013	0.012	0.014	0.015	0.018
$\text{mintMR}_{\text{single-gene}}$	0.072	0.051	0.034	0.064	0.072	0.081
IVW+metaIV	0.048	0.046	0.028	0.042	0.048	0.059
Egger	0.178	0.142	0.138	0.162	0.178	0.206
MVMR-IVW	0.055	0.029	0.017	0.053	0.055	0.060
MVMR-Egger	0.088	0.033	0.019	0.084	0.088	0.094
MVMR-Lasso	0.063	0.029	0.017	0.058	0.063	0.070
MVMR-Median	0.063	0.034	0.020	0.060	0.063	0.068
MVMR-Robust	0.055	0.029	0.017	0.053	0.055	0.060
MVcML	0.036	0.031	0.022	0.034	0.036	0.037

Table 3.9: RMSE comparison of mintMR versus its variants and competing methods when varying the number of samples in each tissue and effect size. The proportion of the variance in the outcome explained by UHP effect is 0.1. When varying sample sizes of exposures from 500 to 10,000, the causal effects are generated from $\mathcal{N}(0, 0.02)$. When varying the variance for generating causal effects (σ^2_β), the sample size of exposure was fixed at 500.

Trait	mintMR	MVMR methods				
		Median	Lasso	IVW	Robust	Egger
Hypertension	1.24	1.79	1.89	1.59	1.84	1.48
Morning or Evening Person	1.37	1.80	1.78	1.58	1.48	1.46
Neuroticism Score	1.40	1.59	1.71	1.58	1.58	1.48
Birth Weight	1.45	1.84	1.83	1.62	1.58	1.51
Body Fat Percentage	1.55	2.13	2.41	1.83	2.01	1.71
Body Mass Index	1.63	2.19	2.49	1.85	2.03	1.74
Height	1.71	2.26	2.54	1.93	2.07	1.75
Depressive Symptoms	1.28	1.57	1.33	1.39	1.21	1.30
Breast Cancer	1.40	1.54	1.43	1.47	1.42	1.36
Eosinophil Count	1.47	1.81	1.71	1.54	1.51	1.40
High Light Scatter Reticulocyte Count	1.47	1.76	1.68	1.54	1.49	1.39
Monocyte Count	1.49	1.77	1.71	1.53	1.49	1.41
Neutrophil Count	1.52	1.85	1.77	1.59	1.53	1.44
Platelet Count	1.58	1.83	1.85	1.61	1.62	1.50
Schizophrenia	1.59	1.80	1.92	1.67	1.70	1.54
Sleep Duration	1.30	1.49	1.33	1.34	1.22	1.25
Chronotype	1.34	1.57	1.45	1.41	1.20	1.27
Low-density Lipoprotein	1.47	1.72	1.59	1.50	1.33	1.39
Sum Eosinophil Basophil Count	1.48	1.82	1.68	1.50	1.48	1.38
Lymphocyte Count	1.49	1.76	1.74	1.60	1.47	1.47
Reticulocyte Count	1.52	1.76	1.73	1.58	1.49	1.43
Asthma	1.53	1.70	1.56	1.54	1.38	1.45
Intelligence	1.54	1.71	1.67	1.56	1.41	1.42
Red Blood Cell Count	1.55	1.84	1.82	1.61	1.52	1.49
Granulocyte Count	1.56	1.88	1.80	1.61	1.53	1.44
White Blood Cell Count	1.57	1.85	1.81	1.64	1.53	1.51
Sum Neutrophil Eosinophil Count	1.58	1.89	1.79	1.62	1.52	1.46
Myeloid White Cell Count	1.59	1.89	1.79	1.60	1.55	1.45
Sum Basophil Neutrophil Count	1.59	1.88	1.76	1.60	1.51	1.43
Intermediate-density Lipoprotein	1.42	1.65	1.45	1.40	1.27	1.30
Insomnia	1.32	1.44	1.27	1.29	1.11	1.22
High-density Lipoprotein	1.39	1.39	1.21	1.26	1.07	1.16
Stroke	1.43	1.59	1.36	1.36	1.22	1.28
Atrial Fibrillation	1.51	1.60	1.44	1.42	1.22	1.34
Alzheimer's Disease	1.41	1.38	1.31	1.32	1.28	1.20

Table 3.10: The averaged genome-wide inflation factors across tissues for the p-values of gene expression adjusting for cis-DNAm on different outcomes based on mintMR and five competing MVMR methods. Results from competing methods are underlined if the method has a higher inflation factor than mintMR.

Class	Trait	Exp. adj. DNAm	Exp no adj.
Immunological	High Light Scatter Reticulocyte Count	1.47	1.98
Immunological	Eosinophil Count	1.47	2.01
Immunological	Sum Eosinophil Basophil Count	1.48	2.05
Immunological	Monocyte Count	1.49	2.08
Immunological	Lymphocyte Count	1.49	2.03
Immunological	Neutrophil Count	1.52	2.09
Immunological	Reticulocyte Count	1.52	1.99
Immunological	Red Blood Cell Count	1.55	2.09
Immunological	Granulocyte Count	1.56	2.08
Immunological	White Blood Cell Count	1.57	2.07
Immunological	Sum Neutrophil Eosinophil Count	1.58	2.02
Immunological	Platelet Count	1.58	2.09
Immunological	Sum Basophil Neutrophil Count	1.59	2.00
Immunological	Myeloid White Cell Count	1.59	2.04
Psychiatric	Depressive Symptoms	1.28	1.76
Psychiatric	Sleep Duration	1.30	1.84
Psychiatric	Chronotype	1.34	1.86
Psychiatric	Morning or Evening Person	1.37	1.93
Psychiatric	Neuroticism Score	1.40	1.87
Psychiatric	Alzheimer's Disease	1.41	1.78
Psychiatric	Schizophrenia	1.59	1.99
Metabolic	High-density Lipoprotein	1.39	1.76
Metabolic	Intermediate-density Lipoprotein	1.42	1.90
Metabolic	Birth Weight	1.45	2.03
Metabolic	Low-density Lipoprotein	1.47	1.94
Metabolic	Body Fat Percentage	1.55	2.15
Metabolic	Body Mass Index	1.63	2.29
Cardiovascular	Hypertension	1.24	1.89
Neurological	Insomnia	1.32	1.75
Neurological	Stroke	1.43	1.82
Cardiovascular	Atrial Fibrillation	1.51	1.87
Neoplasms	Breast Cancer	1.40	1.88
Respiratory	Asthma	1.53	1.97
Cognitive	Intelligence	1.54	1.95
Skeletal	Height	1.71	2.33

Table 3.11: The averaged genome-wide inflation factors across tissues for the p-values of gene expression on different outcomes with and without accounting for DNA methylation, using mintMR.

Category	Trait	# Genes analyzed	# Significant genes	# Significant CpGs
Skeletal	Height	3434	139	59
Metabolic	Body Mass Index	3471	120	55
Immunological	Platelet Count	3418	111	34
Immunological	Lymphocyte Count	3418	110	45
Immunological	Red Blood Cell Count	3418	110	45
Psychiatric	Schizophrenia	3471	110	35
Metabolic	Body Fat Percentage	3471	109	30
Immunological	White Blood Cell Count	3418	106	33
Cardiovascular	Atrial Fibrillation	3486	104	43
Immunological	Granulocyte Count	3418	99	36
Cognitive	Intelligence	3471	99	35
Immunological	Myeloid White Cell Count	3418	98	33
Immunological	Eosinophil Count	3418	96	36
Immunological	Monocyte Count	3419	94	32
Immunological	Sum Neutrophil Eosinophil Count	3418	94	37
Immunological	Neutrophil Count	3418	93	34
Neoplasms	Breast Cancer	3471	93	36
Immunological	Sum Basophil Neutrophil Count	3418	92	34
Neurological	Stroke	3472	91	35
Immunological	Sum Eosinophil Basophil Count	3418	91	34
Respiratory	Asthma	3486	90	48
Immunological	High Light Scatter Reticulocyte Count	3418	88	40
Immunological	Reticulocyte Count	3418	88	32
Metabolic	Birth Weight	3471	87	31
Psychiatric	Neuroticism Score	3471	84	31
Psychiatric	Morning or Evening Person	3471	81	27
Metabolic	Low-density Lipoprotein	3387	79	30
Metabolic	Intermediate-density Lipoprotein	3487	77	32
Psychiatric	Alzheimer's Disease	3458	72	26
Psychiatric	Chronotype	3305	70	19
Psychiatric	Sleep Duration	3463	69	21
Metabolic	High-density Lipoprotein	3471	60	35
Cardiovascular	Hypertension	3471	57	14
Neurological	Insomnia	3467	53	28
Psychiatric	Depressive Symptoms	3393	37	20

Table 3.12: The 35 complex traits/diseases analyzed, their categories, and the number of genes being examined for each trait/disease. The numbers of genes/CpGs showing non-zero effects in at least two tissues for each outcome at the level of FDR<0.05 are listed.

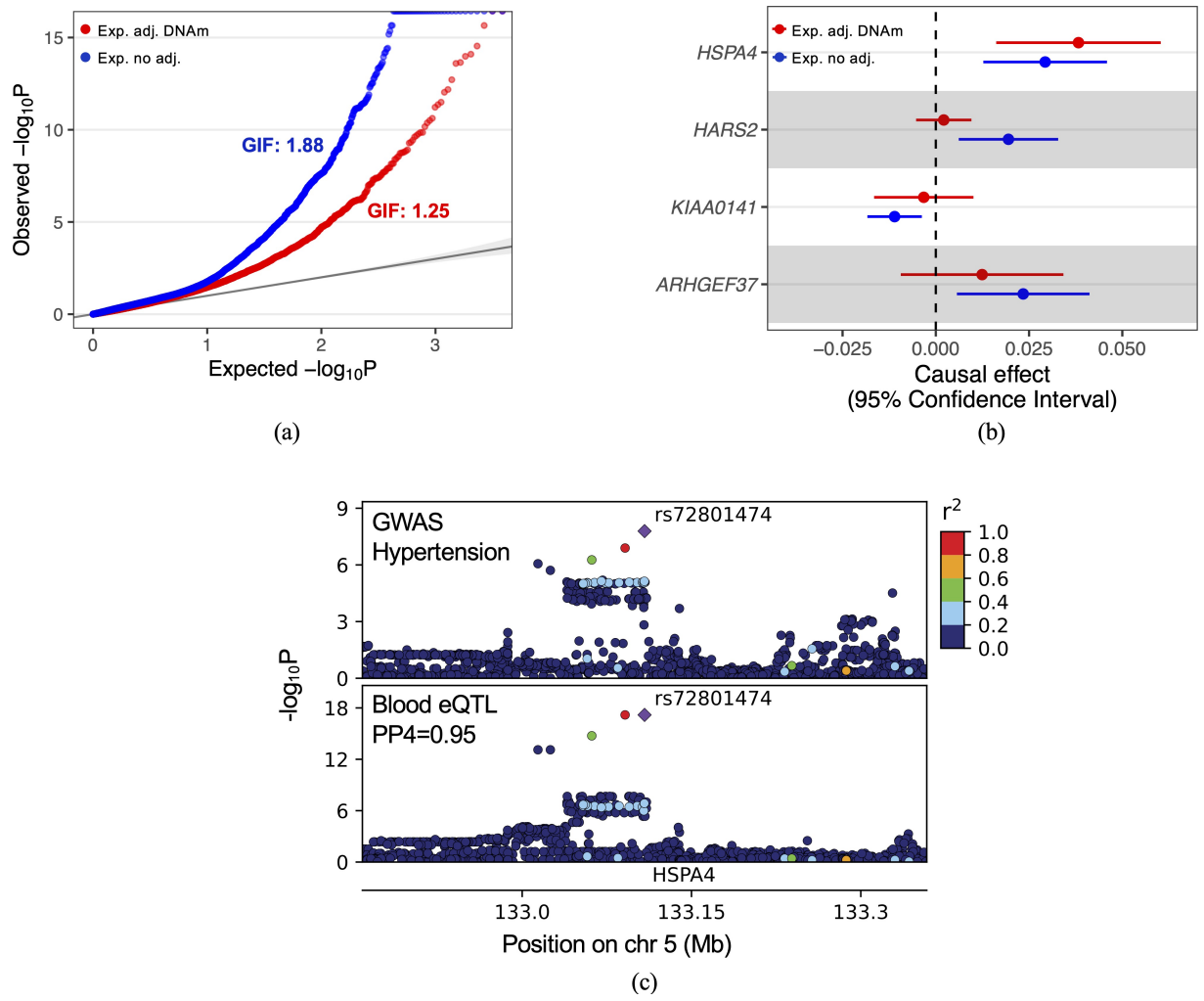
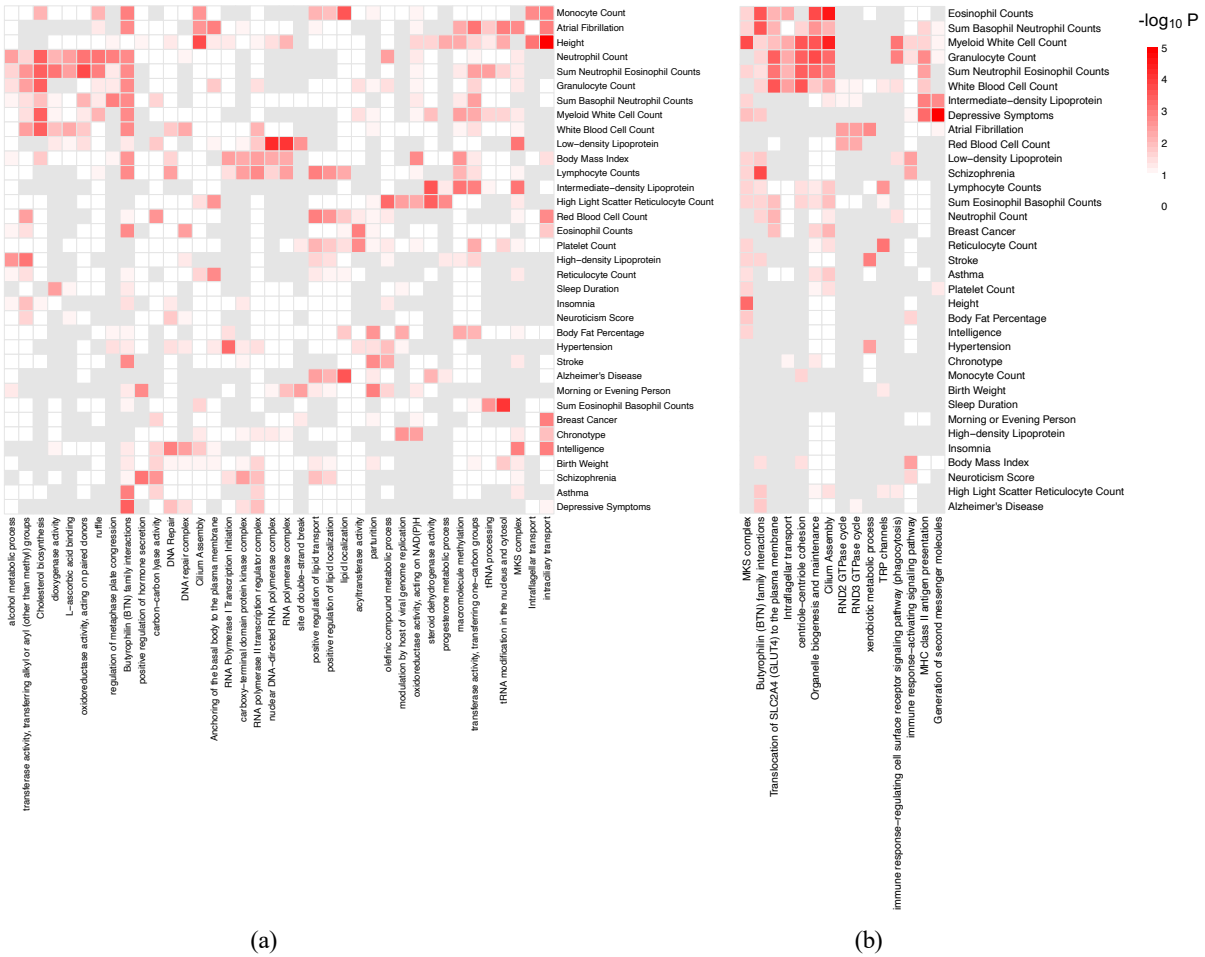


Figure 3.2: (a) A QQ plot of negative log base 10 of p-values for gene expression effects on hypertension. Red points represent the p-values of gene expression adjusting for DNAm. Blue points are the p-values of gene expression without adjusting for DNAm. Genome-wide inflation factors (GIF) for both analyses are shown. (b) The causal effects of four genes on hypertension in the blood tissue in the 5q31-32 region, with and without adjusting for DNAm. Without adjusting for DNAm (blue points and error bars), the four gene expression levels show significant effects on hypertension (FDR<0.05). After adjusting for DNAm (red points and error bars), only the expression of *HSPA4* is significant. (c) Genotype-phenotype association p-values in the *HSPA4* locus for hypertension GWAS (top panel) and eQTL in the blood (bottom panel). The colocalization probability (PP4) of eQTL with GWAS signal is shown. The diamond-shaped point represents the top significant eQTL variant (rs72801474). Linkage disequilibrium between SNPs is assessed by squared Pearson coefficient of correlation (r^2).



(a) (b)

Figure 3.3: The heatmaps of enriched pathways for (a) identified genes affecting complex traits/diseases and (b) proximal genes correlated with the identified significant CpG sites. The p-values of pathway enrichment are calculated based on one-sided Fisher's exact tests without multiple testing adjustments. Pathways with p-values< 0.005 for at least two traits are presented.

CHAPTER 4

UNVEILING RISK GENES IN SPECIFIC CELL TYPES VIA A DEEP LEARNING MENDELIAN RANDOMIZATION METHOD INTEGRATING SINGLE-CELL QTL WITH GWAS

4.1 Attributions

Dr. Lin Chen conceived the project and contributed to the development of the methods and the writing of the manuscript.

4.2 Introduction

In recent years, single-cell RNA sequencing (scRNA-seq) has revolutionized the landscape of genomics by enabling the high-throughput profiling of gene expression at an unprecedented resolution [Hwang et al., 2018, Tang et al., 2009], specific to cell types and states. This advancement has facilitated the emergence of single-cell expression quantitative trait loci (sc-eQTLs) mapping across different cell types [van der Wijst et al., 2020, Yazar et al., 2022, Oelen et al., 2022, Soskic et al., 2022], revealing how expression levels associated with trait/disease-related genetic variants manifest in specific cell populations [de Vries et al., 2020, Nathan et al., 2022]. Despite the significant insights gained from sc-eQTL studies, new challenges arise due to the intrinsic variability in gene expression among individual cells and the heterogeneity of cell types [Carter and Zhao, 2021, Yazar et al., 2022, Perez et al., 2022]. Moreover, the power of sc-eQTL data is currently limited by sample sizes [He et al., 2021, Bryois et al., 2022, Lopes et al., 2022, Young et al., 2021, Jerber et al., 2021]. To improve power, recent studies integrate sc-eQTL with bulk-tissue eQTL (bk-eQTL) and cell type-specific eQTLs (ct-eQTLs) derived from bulk tissues [Donovan et al., 2020, Consortium et al., 2020]. Bk-eQTLs aggregate effects from various cell types and bk-calculated ct-eQTLs

often have consistent effects with sc-eQTL effects in specific contexts [Aguirre-Gamboa et al., 2020].

Existing methods have been proposed to integrate eQTLs with GWAS to map risk genes via Mendelian randomization (MR) [Gleason et al., 2021, Zhou et al., 2020, Richardson et al., 2020], transcriptome-wide association analysis (TWAS) [Hu et al., 2019, Wainberg et al., 2019, Yuan et al., 2020, Shi et al., 2020, Luningham et al., 2020] and colocalization [Giambartolomei et al., 2014, Foley et al., 2021, Pividori et al., 2020, Giambartolomei et al., 2018]. Two-sample MR methods treat SNPs associated with risk exposures as instrumental variables (IVs) to assess the causal effects from exposure on outcome [Chen et al., 2007, Lawlor et al., 2008, Schadt et al., 2005, Davey Smith and Ebrahim, 2003]. MR has been widely used to identify risk factors allowing the presence of unmeasured confounders [Burgess et al., 2013, Bowden et al., 2015, Zhao et al., 2020]. Recently, transcriptome-wide MR methods have been applied to identify risk genes for complex diseases treating eQTLs as IVs [Gleason et al., 2021, Richardson et al., 2020, Barfield et al., 2018, Zhou et al., 2020]. Compared with TWAS, MR infers causal relationships instead of associations, allowing unmeasured confounders if not associated with IVs. Compared with colocalization analysis, recent extensions of MR enjoy relaxed assumptions [Cheng et al., 2022, Wang et al., 2021, Xue et al., 2021, Morrison et al., 2020], and multivariable MR [Burgess and Thompson, 2015, Rees et al., 2017, Grant and Burgess, 2021, Lin et al., 2023, Yihao Lu and Ke Xu and Bowei Kang and Brandon L Pierce and Fan Yang and Lin Chen, 2024] allows the adjustment of known confounders and simultaneously consider multiple cell types. In Chapter 3, we develop a multi-tissue multi-omics MR method for jointly analyzing gene expression and DNA methylation from multiple tissues to identify risk genes, integrating bk-eQTLs and methylation-QTLs from multiple tissues with GWAS. In this Chapter, we propose an MR method integrating sc-eQTLs and ct-eQTLs with genome-wide association study (GWAS) summary statistics to map risk genes in specific cell types. The integration of sc-eQTL,

ct-eQTL with GWAS presents a promising avenue to dissect the effects of gene expression in distinct cell types and to elucidate mechanisms underlying diseases and traits at the cellular level.

The integration of sc-eQTL, ct-eQTLs from bk, with GWAS faces many challenges, including heterogeneity in effect sizes [Ding et al., 2024], sparsity of effects of risk genes in specific cell types, and the prevalence of missing values due to the lack of sequencing in specific cells [Zhang et al., 2024, Pool et al., 2023, Hicks et al., 2018]. Despite these challenges, previous integrative analyses of sc-eQTL and bulk QTL statistics have demonstrated improved power in detecting eQTLs, underscoring the potential of such approaches [Ding et al., 2024]. We propose a deep-learning-based multi-cell-type MR method, deep-cellMR, to simultaneously analyze hundreds of genes for their cell-type-specific effects on disease outcomes by performing MR across multiple cell types for each gene. An innovation of the method is that it introduces a latent effect indicator of each gene in each cell type and models the low-rank patterns of these latent indicators across genes, cells, and data via deep learning. It uses random forest to impute latent variables for missing values [Stekhoven and Bühlmann, 2012]. By applying deep multi-view learning methods [Yan et al., 2021, Wang et al., 2015, 2016, Andrew et al., 2013, Ngiam et al., 2011] on the latent indicators across genes from multiple cell types and data, deep-cellMR captures the nonlinear and complex dependencies and patterns, uses the captured low-rank patterns to estimate the latent probabilities of non-zero, and further improves the estimation of cell-type-specific effect of each gene in each cell. The rationale is that risk genes for a disease often show non-zero effects in specific cell types [Boyle et al., 2017, Finucane et al., 2018, Lynall et al., 2022]. Deep-cellMR maps risk genes in specific cell types, addressing the inherent challenges such as limited sample sizes, cell-type and context-specific mechanisms, sparse effects, complex nonlinear dependence among cell types and data types (sc- vs ct-eQTLs), and substantial missingness in the data.

Through simulations, the deep multi-view learning method we used in deep-cellMR has demonstrated enhanced power and precision compared to other methods. We have applied deep-cellMR to dissect the molecular and cellular pathways implicated in cerebral amyloid angiopathy (CAA), a form of cerebral small vessel disease (CSVD) [Greenberg et al., 2020, Wang et al., 2024], and to explore the shared pathways and mechanisms underlying obesity (OB) and type-2 diabetes (T2D), conditions that exacerbate the risk of both CAA and late-onset Alzheimer's disease (LOAD) [De Felice et al., 2022, Han and Li, 2010]. Our findings not only shed light on the role of microvascular dysfunction in these comorbidities but also pave the way for identifying novel pathophysiological mechanisms and therapeutic targets for CAA and LOAD, marking a significant step forward in our understanding of these complex diseases at a cellular level.

4.3 Methods

4.3.1 The integrative MR framework

The proposed deep-cellMR method simultaneously estimates the cell-type-specific effects of hundreds of genes on the disease outcome across multiple cell types from multiple data. As illustrated in Figure 4.1, for each gene it performs a multi-cell-type MR for estimating cell-type-specific gene effects. By modeling the latent indicators of non-zero effects and extracting the low-rank patterns across genes/cells/data via deep learning, deep-cellMR estimates the disease-relevance probabilities of each gene in each cell type/data accounting for these patterns, and iterates between performing MR and capturing patterns in non-zero effects until the maximum iteration is reached. Specifically, we first introduce the single-gene multi-cell-type MR model. For each gene, it treats gene expression of a single gene from multiple cell types as the exposures, sc-eQTLs as the IVs, and the complex disease of interest as the outcome. In addition to sc-eQTLs, we also integrate ct-eQTL calculated from

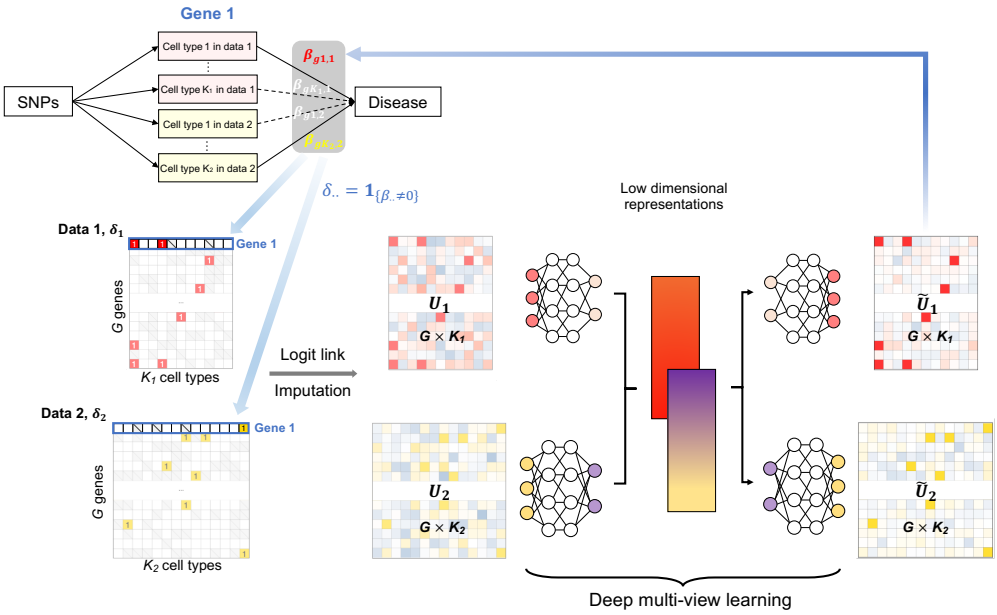


Figure 4.1: Illustration of the deep-cellMR method. For each analyzed gene, deep-cellMR performs a multi-cell-type MR separately. It estimates cell-type-specific gene effects. Deep-cellMR models the latent effect indicators ($\boldsymbol{\delta}_l$) and uses deep multi-view learning methods to capture major patterns in the modulation matrices ($\boldsymbol{U}_l$) of the latent indicators across genes, cells, and data. In the deep multi-view learning step, deep-cellMR uses deep neural networks to find low-dimensional representations of modulation matrices that capture both shared information and unique information across data and uses these representations to reconstruct the modulation matrices via deep neural networks. Deep-cellMR estimates the disease-relevance probabilities of each gene in each cell type/data with the reconstructed modulation matrices. It iterates between performing MR for each single gene and capturing patterns in modulation matrices until the maximum iteration is reached.

bulk tissues as IVs. In general, we may consider eQTLs from a set of L data/studies/tissues, each with K_l cell types ($l = 1, \cdots, L$). For example, in our data applications, we consider sc-eQTLs from Bryois et al. [Bryois et al., 2022] and ct-eQTLs from the Genotype-Tissue Expression (GTEx) project [Consortium, 2020, Donovan et al., 2020, Consortium et al., 2020] as IVs, $L = 2$. The IV-to-exposure effects (here eQTL effects) for each data are measured in multiple cell types.

We consider a SNP i ($i = 1, \cdots, I_g$) as an IV for gene expression indexed by g ($g = 1, \cdots, G$). Let $\gamma_{gik,l} (k = 1, \ldots, K_l; l = 1, \ldots, L)$ denote the true marginal effect of the SNP i on the l-th data of gene g in the cell type k. Let Γ_i^g denote the true marginal association between SNP i and the disease outcome of interest, and the superscript g indicates that the SNP i is an IV for gene g. Denote $\{\widehat{\gamma}_{gik,l}, \widehat{s}_{\gamma_{gik,l}}\}$ as the estimated SNP-gene association and its standard error for SNP i and gene g in cell type k, and $\{\widehat{\Gamma}_i^g, \widehat{s}_{\Gamma_i^g}\}$ as the estimated effect of SNP i on the outcome and its standard error. We have the following MR model for SNP i:

$$\begin{pmatrix} \widehat{\Gamma}_i^g \\ \widehat{\boldsymbol{\gamma}}_{gi,1} \\ \vdots \\ \widehat{\boldsymbol{\gamma}}_{gi,L} \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \Gamma_i^g \\ \boldsymbol{\gamma}_{gi,1} \\ \vdots \\ \boldsymbol{\gamma}_{gi,L} \end{pmatrix}, \widehat{\boldsymbol{S}}_{gi} \boldsymbol{C} \widehat{\boldsymbol{S}}_{gi} \right),$$

where $\boldsymbol{C}$ is the correlation matrix across GWAS and eQTL studies due to potential sample overlap (if there is any) and is often pre-estimated, $\widehat{\boldsymbol{\gamma}}_{gi,l} = [\widehat{\gamma}_{gi1,l}, \cdots, \widehat{\gamma}_{giK_l,l}]^\top$ is a vector of eQTL effects for SNP i and data l of gene g,
$\widehat{\boldsymbol{S}}_{gi} = \text{diag}\left(\widehat{s}_{\Gamma_i^g}, \widehat{s}_{\gamma_{gi1,1}}, \cdots, \widehat{s}_{\gamma_{giK_1,1}}, \cdots, \widehat{s}_{\gamma_{gi1,L}}, \cdots, \widehat{s}_{\gamma_{giK_L,L}}\right)$ is the standard error estimate from GWAS, and (sc- and ct-) eQTL studies.

We further assume the true causal relationship is linear between the marginal effect of the SNP i on the outcome and the marginal effects of the SNP i on data/studies/tissues,

given by

$$\Gamma_i^g = \alpha_i^g + \sum_{k=1}^{K_1} \delta_{gk,l} \cdot \beta_{gk,l}\gamma_{gik,l} + \cdots + \sum_{k=1}^{K_L} \delta_{gk,L} \cdot \beta_{gk,L}\gamma_{gik,L}, \tag{4.1}$$

where $\beta_{gk,l}$ is the causal effect of interest for gene g in cell type k of data l, $\delta_{gk,l} \sim$ Bernoulli $\left(\pi_{gk,l}\right)$ is a latent indicator for disease relevance of cell type for the gene, and $\delta_{gk} = 1$ if $\beta_{gk} \neq 0$, $\delta_{gk,l} \cdot \beta_{gk,l}$ describes the direct effect of gene expression in cell type k on the outcome not operating through other cell types. Here the true IV-to-exposure effect follows $\gamma_{gik,l} \sim \mathcal{N}(0, \sigma^2_{\gamma_g})$, and $\alpha_i^g \sim \mathcal{N}\left(0, \sigma^2_{\alpha^g}\right)$ is the uncorrelated horizontal pleiotropic effect [Morrison et al., 2020] when IV affects outcome not through the exposure in any data/study/tissue and IV is not associated with confounder. In practice, often there are only a very limited number of sc-eQTLs and ct-eQTLs as IVs, and the causal effects are very sparse, as such the latent indicators may not be statistically identifiable. As a result, estimating the effects specific to each cell type for a single gene via MR poses significant challenges.

4.3.2 Deep-cellMR: deep learning MR jointly modeling disease-relevance pattern across cell-types/genes/data-types

Many causal effects and their involved disease-relevant mechanisms are specific to certain cell types [Hekselman and Yeger-Lotem, 2020, Finucane et al., 2018]. Risk genes involved in the same pathway and mechanisms are likely to be expressed and have non-zero effects on diseases in specific cell types. Deep-cellMR jointly models the latent indicators of non-zero effects across multiple genes in multiple cell types from multiple data. Via deep learning of low-rank patterns and accounting for nonlinear and complex dependency between cell types and data types, deep-cellMR efficiently estimates non-zero effect probabilities and improves the estimation of sparse cell-type-specific effects.

In sc-RNAseq data, there are a considerable amount of missing values across various cell types, due to the sequencing depth and the non-expression of some genes in specific cells. We extend the model (4.1) to allow for different numbers of cell types across genes. For the data l, we denote the set of cell types available for gene g as $\mathcal{S}_{g,l}$. The union of analyzed contexts for all G genes is $\mathcal{S}_l = \mathcal{S}_{1,l} \cup \cdots \cup \mathcal{S}_{G,l}$ with $|\mathcal{S}_l| = K_l$. We assume the causal relationship as

$$\Gamma_i^g = \alpha_i^g + \sum_{k \in \mathcal{S}_{g,1}} \delta_{gk,l} \cdot \beta_{gk,l} \gamma_{gik,l} + \cdots + \sum_{k \in \mathcal{S}_{g,L}} \delta_{gk,l} \cdot \beta_{gk,L} \gamma_{gik,L}. \tag{4.2}$$

If there are no missing cell types for all genes, all $S_{g,l}$'s are identical and the model (4.2) reduces to model (4.1).

We propose a joint MR model across G genes to estimate the causal effects for each gene in each cell type and jointly learn the major patterns of latent disease-relevance cell type indicators. As illustrated in Figure 4.1, we consider gene expression in multiple cell types and study their effects on disease. The direct effects $\beta_{gk,l}$'s may have different magnitudes or distributions across data. To capture the shared information among the true non-zero causal effects, the proposed deep-cellMR model works by iteratively estimating an MR model for each gene and collectively capturing the major data-shared and data-specific patterns across G genes via deep learning, and it iterates until the maximum iteration is reached. Specifically, we form latent disease-relevance indicator matrices as $\boldsymbol{\delta}_l = \left\{\delta_{gk,l}\right\}$, where $\delta_{gk,l}$ could be missing if $k \notin \mathcal{S}_{g,l}$. We introduce a continuous modulation matrix for each data l as $\boldsymbol{U}_l = \left\{U_{gk,l}\right\} \in \mathbb{R}^{G \times K_l}$, and

$$U_{gk,l} = \log \frac{\Pr\left(\delta_{gk,l} = 1 | \boldsymbol{U}_l, u_{0k,l}\right)}{\Pr\left(\delta_{gk,l} = 0 | \boldsymbol{U}_l, u_{0k,l}\right)} - u_{0k,l}, \tag{4.3}$$

where $\boldsymbol{U}_l$ is a modulation matrix for the latent disease-relevance indicators, $U_{gk,l}$ is missing if $k \notin \mathcal{S}_{g,l}$. Here $u_{0k,l}$ is the cell type-specific intercept, controlling the sparsity of non-

zero effects in the k-th cell type for the l-th data. As multiple genes that affect diseases in a pathway often have effects specific to certain cell types and similar cell types across data/studies/tissues may provide complementary information, we apply an iterative imputation method (missForest [Stekhoven and Bühlmann, 2012]) based on a random forest to impute missing values in $\boldsymbol{U}_l$. MissForest treats each cell type with missing values in the latent modulation matrices as the response and uses the observed and imputed values of all other cell types as predictors to fit regression trees and averages over the imputed values over many trees. It imputes each cell type one by one and iterates until convergence.

We estimate the cell-type-specific effect on disease outcome for each gene, accounting for the major patterns across multiple genes obtained via deep multi-view learning methods [Wang et al., 2015, Yan et al., 2021, Wang et al., 2016, Andrew et al., 2013]. The proposed deep-cellMR is a flexible framework, and many deep learning algorithms can be used to capture desirable low-rank patterns in estimating the latent disease-relevance (non-zero effect) probabilities, such as split autoencoder (SplitAE) [Ngiam et al., 2011], deep canonically correlated autoencoders (DCCAE) [Wang et al., 2015], and deep variational canonical correlation (DVCCA) [Wang et al., 2016]. In general, these methods can learn low-dimensional representations given multiple modulation matrices, capturing "consensus" patterns shared by modulation matrices of different data. SplitAE and DCCAE involve a reconstruction objective where major patterns are encouraged when learning the representations such that the reconstruction error is minimized. Here the reconstruction represents the process where the low-dimensional latent representations extracted from modulation matrices are used to reconstruct the original modulation matrices through linear transformations or deep neural networks (DNNs). DVCCA extends the latent variable interpretation of linear CCA to nonlinear observation models parameterized by DNNs.

To capture the nonlinear dependence among sc-eQTLs and bk-derived ct-eQTLs, and to account for the patterns in the sparse and weak cell-type-specific effects, we propose to

use DVCCA. Compared to other autoencoder-based deep learning algorithms developed for the same purpose (DCCAE, SplitAE, etc.), DVCCA accounts for the uncertainty in data using a Gaussian observation model. It works well especially when the input data is very noisy. Compared with autoencoder, DVCCA generates many different "noisy" versions of the latent representation and fits DNNs such that these versions reconstruct the original inputs well when the input data is noisy. The methods based on autoencoders might not accurately reconstruct the input when the input is very noisy [Wang et al., 2015] as the network might learn to reconstruct the noise along with the input data. This can lead to the network capturing unwanted noise as part of the learned features, essentially giving noise a representation in the feature space. In other words, the autoencoder may learn to reproduce the input including the noise. Deep CCA (DCCA) extracts nonlinear representations from the input data. However, it does not provide a model for generating samples from the latent space and thus can not be applied in deep-cellMR. DCCAE as a variation of DCCA combines autoencoder and DCCA. The decoder component in DCCAE reconstructs the input. However, in practice, the canonical correlation term often dominates the reconstruction terms in the objective, and therefore the inputs are not reconstructed well [Wang et al., 2016]. In our integrative analysis of sc-eQTLs and ct-eQTLs, we expect that the effects specific to each cell type are both sparse and weak, with the input data being prone to noise. Consequently, methods like SplitAE and DCCAE might exhibit limitations in reconstructing input that contains a substantial amount of noise.

In DVCCA, nonlinear observation models $p_{\boldsymbol{\theta}}\left(\boldsymbol{u}_l|\boldsymbol{z};\boldsymbol{\theta}_l\right)$ are parameterized by weights $(\boldsymbol{\theta}_l)$ of DNNs $\boldsymbol{f}_l$'s, where $\boldsymbol{u}_l$'s are random vectors. The g-th row of the modulation matrix, $\boldsymbol{U}_{g\cdot,l}$'s, are observations of $\boldsymbol{u}_l$. Conditioning on a latent variable $\boldsymbol{z} \in \mathbb{R}^q$, $\boldsymbol{u}_l$'s are independent. Here q is the dimension of representation. In DVCCA, nonlinear observation models $p_{\boldsymbol{\theta}}\left(\boldsymbol{u}_l|\boldsymbol{z};\boldsymbol{\theta}_l\right)$ are parameterized by $\boldsymbol{\theta}_l$, which are weights of DNNs $\boldsymbol{f}_l$'s, i.e.,

$$p_{\boldsymbol{\theta}}(\boldsymbol{u}_l \mid \boldsymbol{z}) = \mathcal{N}\left(\boldsymbol{f}_l\left(\boldsymbol{z};\boldsymbol{\theta}_l\right), \boldsymbol{I}\right). \tag{4.4}$$

Maximizing the lower bound of the marginal likelihood is equivalent to optimizing the following objective function:

$$\begin{aligned}\min_{\boldsymbol{\theta},\boldsymbol{\phi}} -\frac{1}{G}\sum_{g=1}^{G} D_{KL}\left(q_{\boldsymbol{\phi}}\left(\boldsymbol{z}_g \mid \boldsymbol{U}_{g\cdot,l}\right) \| p\left(\boldsymbol{z}_g\right)\right) \\ +\frac{1}{2GT}\sum_{g,t}\left\|\boldsymbol{U}_{g\cdot,1}-\boldsymbol{f}_1\left(\boldsymbol{z}_g^{(t)};\boldsymbol{\theta}_1\right)\right\|^2+\left\|\boldsymbol{U}_{g\cdot,2}-\boldsymbol{f}_2\left(\boldsymbol{z}_g^{(t)};\boldsymbol{\theta}_2\right)\right\|^2,\end{aligned} \tag{4.5}$$

where $D_{KL}\left(q_{\boldsymbol{\phi}}\|p\right)$ denotes the KL divergence between the approximate posterior q_ϕ and the prior q for the latent variables, $\boldsymbol{\phi}$ is the parameters of another DNN. The second term is obtained using Monte Carlo sampling by drawing T samples $\boldsymbol{z}_g^{(t)}$. The reconstruction matrices $\widetilde{\boldsymbol{U}}_{g\cdot}$ from the outputs of $\boldsymbol{f}_l$'s accounts for nonlinear dependency patterns across cell types and data. We use $\widetilde{\boldsymbol{U}}_l$'s as an approximation to the input modulation matrices and update probabilities for latent effect indicators.

Deep-cellMR is a flexible and general MR framework for jointly estimating the causal effects of multiple risk factors across multiple contexts. In the deep learning step, one may also consider SplitAE and DCCAE to capture desirable patterns in other applications and analyses. Here we present these variations. SplitAE can be applied when $L = 2$. It fits feature extraction network $\boldsymbol{f}$ and reconstruction networks $\boldsymbol{p}_l$'s for each data ($l = 1, 2$). The objective is to minimize the sum of reconstruction errors for the two views

$$\min_{\boldsymbol{W}_{\boldsymbol{f}},\boldsymbol{W}_{\boldsymbol{p}_l}} \frac{1}{G}\sum_{g=1}^{G}\left(\left\|\boldsymbol{U}_{g\cdot,1}-\boldsymbol{p}_1\left(\boldsymbol{f}\left(\boldsymbol{U}_{g\cdot,1}\right)\right)\right\|^2+\left\|\boldsymbol{U}_{g\cdot,2}-\boldsymbol{p}_2\left(\boldsymbol{f}\left(\boldsymbol{U}_{g\cdot,2}\right)\right)\right\|^2\right), \tag{4.6}$$

where $\boldsymbol{W}_\cdot$ denotes learnable parameters of the corresponding DNNs. In our framework, we use $\widetilde{\boldsymbol{U}}_l = \boldsymbol{p}_l(\boldsymbol{f}(\boldsymbol{U}_l))$ to update probabilities for latent effect indicators. In Deep CCA, a DNN ($\boldsymbol{f}_l$) is fitted for each data l to extract nonlinear patterns such that the correlation between the transformed matrices is maximized. Based on DCCA, DCCAE adds an autoencoder regularization term to enable the reconstruction from the representations. It optimizes the

combination of an autoencoder objective (reconstruction errors) and the canonical correlation objective:

$$\begin{aligned}\min_{\boldsymbol{W}_{\boldsymbol{f}_l}, \boldsymbol{W}_{\boldsymbol{p}_l}, \boldsymbol{V}_l} & -\frac{1}{G}\operatorname{tr}\left(\boldsymbol{V}_1^\top \boldsymbol{f}_1(\boldsymbol{U}_1)\boldsymbol{f}_2(\boldsymbol{U}_2)^\top \boldsymbol{V}_2\right) \\ & + \frac{\lambda}{G}\sum_{g=1}^{G}\left(\left\|\boldsymbol{U}_{g\cdot,1} - \boldsymbol{p}_1\left(\boldsymbol{f}_1\left(\boldsymbol{U}_{g\cdot,1}\right)\right)\right\|^2 + \left\|\boldsymbol{U}_{g\cdot,2} - \boldsymbol{p}_2\left(\boldsymbol{f}_2\left(\boldsymbol{U}_{g\cdot,2}\right)\right)\right\|^2\right),\end{aligned} \tag{4.7}$$

where λ is a trade-off parameter for the DCCA loss and reconstruction loss. In our framework, we use $\widetilde{\boldsymbol{U}}_l = \boldsymbol{p}_l(\boldsymbol{f}_l(\boldsymbol{U}_l))$ to approximate the input modulation matrices and update the probabilities for latent effect indicators.

The proposed deep-cellMR algorithm iterates between estimating the causal effects in the single-gene model (4.2) for each of the G gene regions, imputing missing values in modulation matrices, and jointly learning the latent disease-relevance probabilities via deep multi-view learning methods (Algorithm 3).

4.4 Results

4.4.1 Deep multi-view learning captures major pattern

We conducted simulation studies to evaluate the performance of the proposed deep-cellMR method with existing Mendelian randomization methods. We simulated latent effect indicator matrices based on modulation matrices obtained from sc-eQTL, ct-eQTL, and GWAS summary statistics. We included 51 genes in six AD-related pathways (details in next section) with available sc-eQTL and ct-eQTL summary statistics in all analyzed cell types. We then introduced missingness to them to construct observed summary statistics as input for deep-cellMR. The data generation process is the same as the process described in Chapter 3.4.1.

We compared deep-cellMR with existing multivariable methods, including MVMR-IVW

Algorithm 3 The Gibbs sampling algorithm for deep-cellMR model

1: Input data: $\widehat{\gamma}_{gik,l}$, $\widehat{s}_{\gamma_{gik,l}}$, $\widehat{\Gamma}_i^g$, $\widehat{s}_{\Gamma_i^g}$, $\widehat{\boldsymbol{R}}^g \in \mathbb{R}^{I_g \times I_g}$, $\widehat{\boldsymbol{C}} \in \mathbb{R}^{\sum_l K_l \times \sum_l K_l}$, $q \in \mathbb{Z}^+$, for $g = 1, \cdots, G$, $l = 1, \cdots, L$, $i = 1, \cdots, I_g$, $k = 1, \cdots, K_l$.
2: Initialize parameters: $\Gamma_i^g, \gamma_{gik,l}, \sigma^2_{\beta_{g,l}}$, $\sigma^2_{\gamma_{g,l}}$, $\sigma^2_{\alpha^g}, U_{gik,l}$, and specify $u_{0k,l}$, for $g = 1, \cdots, G$, $l = 1, \cdots, L$, $i = 1, \cdots, I_g$, $k = 1, \cdots, K_l$. This can be either user-specified or obtained by running the Gibbs Sampling algorithm for a starting model by skipping steps 14-23.
3: **for** each iteration **do**
4: **for** $g = 1$ to G **do**
5: **for** $i = 1$ to I_g **do**
6: Sample Γ_i^g.
7: **for** $l = 1$ to L **do**
8: **for** $k \in \mathcal{S}_{g,l}$ **do**
9: Sample $\gamma_{gik,l}$.
10: **for** $l = 1$ to L **do**
11: **for** $k \in \mathcal{S}_{g,l}$ **do**
12: Sample $\beta_{gk,l}$, $\pi_{gk,l}$, and $\delta_{gk,l}$.
13: Sample $\sigma^2_{\beta_{g,l}}$, $\sigma^2_{\gamma_{g,l}}$, $\sigma^2_{\alpha^g}$.

(lines 4–13: Model for single gene region)

14: **for** $l = 1$ to L **do**
15: $\boldsymbol{U}_l = \left\{ \log \frac{\pi_{gk,l}}{1-\pi_{gk,l}} - u_{0k,l}, 1 \le g \le G,\ 1 \le k \le K_l \right\}$
16: Impute missing value (unavailable cell types for some
17: genes) for $\boldsymbol{U}_l$'s. $[\boldsymbol{U}_1, \cdots, \boldsymbol{U}_L] = \texttt{missForest}([\boldsymbol{U}_1, \cdots, \boldsymbol{U}_L])$.

(lines 16–17: Imputation)

18: Perform deep multi-view learning on imputed $\boldsymbol{U}_l$'s and
19: reconstruct matrices $\widetilde{\boldsymbol{U}}_1, \cdots, \widetilde{\boldsymbol{U}}_L = \texttt{MVL}(\boldsymbol{U}_1, \cdots, \boldsymbol{U}_L)$

(lines 18–19: Fit Deep MVL)

20: **for** $g = 1$ to G **do**
21: **for** $l = 1$ to L **do**
22: **for** $k \in \mathcal{S}_{g,l}$ **do**
23: $\pi_{gk,l} = \frac{1}{1+\exp(-\widetilde{U}_{gk,l} - u_{0k,l})}$.

(lines 14–23: Iterate)

24: **Until** the maximum iteration is reached.

[Burgess and Thompson, 2015], MVMR-Egger [Rees et al., 2017], MVMR-Lasso [Grant and Burgess, 2021], MVMR-Median [Grant and Burgess, 2021] and MVMR-Robust [Grant and Burgess, 2021]. We also included IVW with meta-analyzed IV effects (termed as "IVW+metaIV"). All existing UVMR and MVMR methods are developed for using complex traits as exposures and do not allow for missingness in IV-to-exposure summary statistics. Here we adapted them to TWMR with gene expression in specific cell types as exposures for comparison purposes. For each gene g, we applied the competing methods to examine the cell-type-specific causal effects in the cell types available ($\mathcal{S}_{g,l}$).

For each simulation, we generated two continuous modulation matrices ($L = 2$), each with 51 genes. The number of cell types analyzed was 6 and 12 for these two matrices, respectively. We denote the true underlying matrices obtained from real data as $\boldsymbol{U}_l$ ($l = 1, 2$). We introduced missingness into the modulation matrices completely at random. The latent effect indicator matrices $\boldsymbol{\delta}_l$ were generated with $\delta_{gk,l} \sim \text{Bernoulli}(1/(1+\exp(-U_{gk,l}-u_{0k,l})))$, where $u_{0k,l}$ controls the overall sparsity of non-zero effects ($u_{0k,l} = -4$ for the following simulations). We varied the proportion of missingness for the first data while fixing it as 0.05 for the second data. As shown in Table 4.1a, when the number of IVs was limited and all IVs had UHP, we introduced missingness to 30% pairs of gene and cell type in the data. Deep-cell MR could control type I error rates with varying proportions of variance in the outcome explained by UHP effects while most of the competing methods showed inflated type I error rates. MVMR-Robust showed reasonable controls of type I error rates but suffered from low powers. In Table 4.1b, we increased the proportion of missing values to 0.5 for the first data. Deep-cell MR exhibited robustness to the increased proportion of missing values, controlling type I error rates and demonstrating higher power compared to competing methods. Table 4.2 showed that deep-cellMR controlled the type I error rates for genes without causal effects in different simulations with varying causal effect sizes for non-zero effects. This showed that the deep-cellMR captures only shared patterns among

non-zero effects and would not boost random patterns for noises. In all simulations, we found that the joint modeling of multiple genes in deep-cellMR improved the power and controlled the type I error rates when the number of IVs was limited. The deep-cellMR using DVCCA for multi-view learning showed improved power compared to the variation using CCA (a model similar to mintMR introduced in Chapter 3), indicating the deep learning method further improved the identification of causal effects by capturing nonlinear and complicated dependency among cell types.

4.4.2 Deep-cellMR unveils risk genes in specific cell types

We applied the proposed deep-cellMR method to map risk genes for Alzheimer's disease (AD), small vessel disease (SVD), type 2 diabetes (T2D), and body mass index (BMI). We obtained sc-eQTL summary statistics of eight brain cell types derived from the prefrontal cortex, temporal cortex, and white matter [Bryois et al., 2022]. We also obtained cell-type deconvolution results based on bulk expression. The deconvolution results of GTEx brain cortex, brain cerebellum, lung, heart appendage, and whole blood tissues using CIBERSORT [Donovan et al., 2020] and xCell [Consortium et al., 2020] were analyzed. In total, 14 cell types from those tissues were included. Ct-eQTL summary statistics were then obtained by assessing the interaction effects of genotype and these estimated cell type proportions from different individuals. We analyzed gene sets from inflammatory pathways: nitric oxide (NO), blood-brain barrier (BBB) transport, cytokine-cytokine receptor interaction (termed as 'cytokine pathway' below), TGF-β signaling, VEGF signaling, and TNF signaling. For each gene, we selected the SNPs with non-zero effects in at least one cell type ($P \leq 0.01$). We performed LD clumping at the r^2 threshold of 0.01. We restricted our analysis to genes with at least 25 IVs overall and at least one IV for each cell type.

We applied deep-cellMR with missForest for imputation and DVCCA for multi-view learning to each of the five diseases/traits. The number of genes analyzed for each cell-

	Proportion of the variance in the outcome explained by UHP effects					
	0.05	0.1	0.15	0.05	0.1	0.15
	Power			Type I error rate		
Deep-cellMR	0.835	0.774	0.686	0.051	0.048	0.043
mintMR	0.781	0.691	0.631	0.051	0.055	0.052
IVW+metaIV	0.319	0.314	0.301	0.126	0.124	0.123
Egger	0.248	0.206	0.181	0.106	0.115	0.117
MVMR-IVW	0.632	0.545	0.491	0.081	0.085	0.085
MVMR-Egger	0.559	0.479	0.434	0.079	0.083	0.084
MVMR-Lasso	0.628	0.560	0.522	0.082	0.088	0.097
MVMR-Median	0.619	0.538	0.500	0.076	0.089	0.103
MVMR-Robust	0.518	0.445	0.390	0.048	0.052	0.051

(a)

	Proportion of the variance in the outcome explained by UHP effects					
	0.05	0.1	0.15	0.05	0.1	0.15
	Power			Type I error rate		
Deep-cellMR	0.857	0.792	0.693	0.048	0.044	0.044
mintMR	0.808	0.716	0.648	0.046	0.051	0.049
IVW+metaIV	0.312	0.316	0.297	0.129	0.125	0.124
MR-Egger	0.250	0.204	0.178	0.106	0.113	0.116
MVMR-IVW	0.633	0.575	0.478	0.077	0.079	0.080
MVMR-Egger	0.604	0.517	0.441	0.078	0.083	0.086
MVMR-Lasso	0.644	0.593	0.516	0.077	0.083	0.091
MVMR-Median	0.651	0.600	0.544	0.076	0.098	0.114
MVMR-Robust	0.585	0.472	0.391	0.047	0.048	0.049

(b)

Table 4.1: Simulation results evaluating the performance of Mendelian randomization methods on data with varied missing rates. (a) Results on data with 30% missing gene-cell pairs. (b) Results on data with 50% missing gene-cell pairs. Results are underlined for methods unable to control type I error rates (≥ 0.1).

	Variance of causal effects							
	0.0125	0.01	0.0075	0.005	0.0125	0.01	0.0075	0.005
	Power				Type I error rate			
Deep-cellMR	0.799	0.761	0.689	0.572	0.047	0.049	0.045	0.046
mintMR	0.741	0.724	0.656	0.545	0.049	0.054	0.049	0.045
IVW+metaIV	0.259	0.305	0.261	0.247	0.119	0.124	0.115	0.112
MR-Egger	0.165	0.190	0.151	0.166	0.115	0.115	0.117	0.114
MVMR-IVW	0.577	0.531	0.466	0.441	0.083	0.086	0.083	0.085
MVMR-Egger	0.528	0.486	0.463	0.378	0.084	0.086	0.087	0.089
MVMR-Lasso	0.633	0.615	0.488	0.447	0.139	0.136	0.120	0.113
MVMR-Median	0.565	0.517	0.461	0.367	0.102	0.096	0.100	0.099
MVMR-Robust	0.456	0.436	0.381	0.314	0.052	0.053	0.049	0.048

Table 4.2: Simulation results evaluating the performance of deep-cellMR versus its variants and competing methods on data with varied causal effects. For genes with non-zero effects, $\beta_{gk,l}$'s are generated from $\mathcal{N}\left(0, \sigma_\beta^2\right)$ and σ_β^2 varies from 0.005 to 0.0125. The type I error rates are evaluated for genes without causal effects. The proportion of the variance in the outcome explained by UHP effects is 0.1. Results are underlined for methods unable to control type I error rates (≥ 0.1).

outcome pair varied from 106 (oligodendrocytes and SVD) to 226 (endothelial cells and T2D). At the false discovery rate of 0.05, we identified the genes showing significant effects in each cell type. For Alzheimer's disease, we analyzed an average of 184 genes across the cell types. In Figure 4.2, we showed the negative log base 10 of p-values for gene expression effects on AD among genes in the Cytokine pathway. We found more genes were significant in endothelial cells than in other cell types. A potential interpretation is that the pro-inflammatory cytokines affect AD by altering blood–brain barrier permeability in endothelial cell [Grammas, 2011, Asby et al., 2021, Spangler et al., 2015]. For genes in other pathways, we listed the number of genes with significant effects on AD (FDR<0.05) across different cell types available in the sc-eQTL data. For the NO pathway, 5 out of the 22 (22.73%) analyzed genes showed significant effects in endothelial cells, the highest proportion among all cell types examined. Existing literature showed that NO is continuously released by endothelial cells in the vascular system and vascular pathology plays a crucial role in the progression of Alzheimer's disease [Dubey et al., 2020, Katusic et al., 2023]. In addition, 29.63% of genes

from the BBB pathway were significant in oligodendrocytes, 28.57% of genes from the TGB-β pathway affected AD in inhibitory neurons, 21.21% genes of the TNF pathway affected AD in pericytes, and 13.24% genes of VEGF showed significant effects in excitatory neurons. These identified cell-type-specific effects highlight the importance of mapping risk genes in single-cell resolution. The enrichment of disease-associated gene expression in certain cell types also informs potential underlying cell-type-specific mechanisms of how the pathways affect the disease.

4.4.3 Risk genes in specific cell types inform shared genetic mechanisms underlying AD, SVD and comorbidities

Recent research shows that aging, inflammation, and endothelial dysfunction could contribute to cerebral amyloid angiopathy (CAA), which is a type of cerebral SVD, and further increase the risk of late-onset AD. As known comorbidities, obesity and T2D increase the risk for both CAA and AD. These comorbidities are associated with inflammation and cerebral hypoperfusion. Thus, studying microvascular dysfunction in OB and T2D could potentially reveal new mechanisms behind SVD and AD and identify potential therapeutic targets.

We studied the genes with estimated significant causal effects on SVD/AD (FDR < 0.05) in endothelial cells from the brain and heart, and epithelial cells from the lung. We found that there are many inflammatory genes showing evidence of effects on T2D/BMI and AD/SVD in endothelial cells, indicating shared mechanisms. In detail, in endothelial cells of the brain, we identified a total of 63 genes with significant effects, where 36 of them (57.14%) also showed effects on T2D/BMI. The rate was 30.95% for heart endothelial cells and 42.86% for lung epithelial cells. The high sharing of risk genes especially in the brain suggests these inflammatory genes may affect T2D/BMI in endothelial cells in disease-relevant tissues and further affect the risk of SVD or AD.

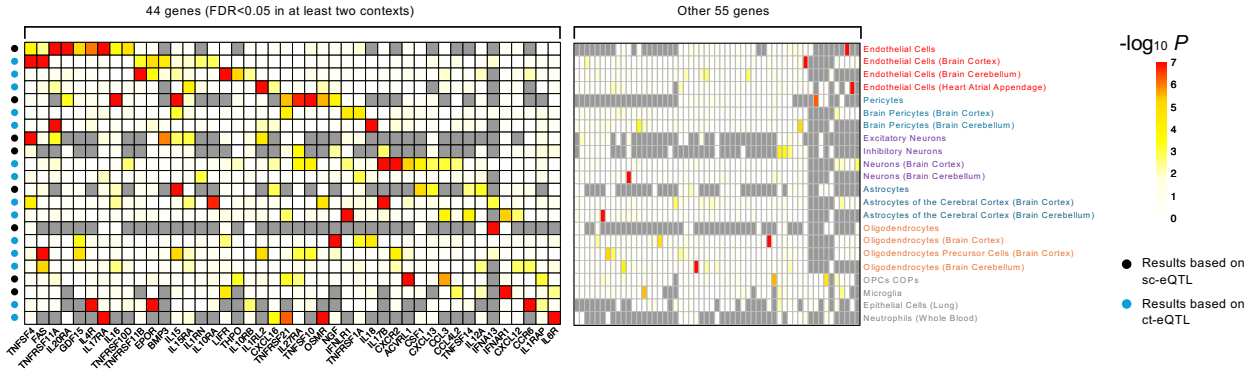


Figure 4.2: Heatmaps of negative log base 10 of P-values for gene expression effects on Alzheimer's disease. The genes are from the cytokine-receptor interaction pathway. The left heatmap shows 44 genes with FDR$<$ 0.05 in at least two contexts out of the 22 contexts analyzed. The right heatmap represents the other 55 genes. Cell types analyzed in both sc-eQTL data (rows with black circles) and ct-eQTL data (rows with blue circles) were grouped.

	NO	Cytokine	BBB	TGF-β	TNF	VEGF
Astrocytes	15% (3/20)	8.51% (4/47)	13.64% (9/66)	21.05% (8/38)	12.5% (5/40)	8% (4/50)
Endothelial cells	**22.73% (5/22)**	14.29% (9/63)	18.18% (12/66)	12.5% (6/48)	9.52% (4/42)	5.36% (3/56)
Excitatory neurons	9.09% (2/22)	10% (4/40)	19.23% (15/78)	10.42% (5/48)	15.79% (6/38)	**13.24% (9/68)**
Inhibitory neurons	10% (2/20)	7.5% (3/40)	17.57% (13/74)	**28.57% (12/42)**	15.15% (5/33)	9.68% (6/62)
Microglia	0% (0/14)	4.92% (3/61)	25% (11/44)	25.71% (9/35)	14.71% (5/34)	4.76% (2/42)
Oligodendrocytes	8.33% (1/12)	4.76% (1/21)	**29.63% (16/54)**	15.62% (5/32)	10.34% (3/29)	5% (2/40)
OPCs/COPs	16.67% (3/18)	13.89% (5/36)	18.75% (12/64)	8.57% (3/35)	3.23% (1/31)	10.87% (5/46)
Pericytes	18.75% (3/16)	**16.07% (9/56)**	22.41% (13/58)	10.26% (4/39)	**21.21% (7/33)**	7.69% (2/26)

Table 4.3: Numbers and proportions of significant genes (FDR < 0.05) for AD among genes from different pathways. For each pathway, the cell type with the highest proportion of significant genes is in bold.

4.5 Discussion

In this work, we propose a deep learning Mendelian randomization method, deep-cellMR, for integrating sc-eQTL and ct-eQTL with GWAS to unveil risk genes in specific cell types. Deep-cellMR performs an MR analysis using sc-eQTLs and ct-eQTLs as IVs for each gene region. It allows the inclusion of different cell types for each gene and jointly models gene sets/pathways. Deep-cellMR improves the estimation of cell-type-specific causal effects of all analyzed genes by modeling the latent non-zero effect indicators. It performs missing value imputation on the latent modulation matrices using random forest and uses deep multi-view learning methods to capture linear and nonlinear patterns of latent indicators/probabilities across genes, cells, and data. Using DVCCA, deep-cellMR handles more complicated dependencies or structures compared to classic CCA. It improves the estimation of sparse cell-type-specific causal effects for all genes.

We applied deep-cellMR to map risk genes for five diseases/traits. We used sc-eQTL summary statistics from eight brain cell types and ct-eQTL summary statistics from brain, heart, lung, and blood. We identified cell-type-specific effects of genes in certain pathways on AD. The sharing of risk genes in disease-relevant cell types also suggests shared mechanisms underlying the analyzed diseases/traits. Our analysis demonstrated that deep-cellMR could identify risk genes in disease-relevant cell types and offer valuable insights into underlying mechanisms.

There are several limitations of our work. First, deep-cellMR trains a deep multi-view learning model in each Gibbs sampling iteration, which may involve training several neural networks. The computation efficiency can be further improved, especially when the number of cell types being jointly analyzed is large. Second, although deep-cellMR can capture linear and nonlinear patterns in the modulation matrices, it assumes a linear causal relationship. The model can be extended to study interactions among exposures and covariates among cell types.

In future work, deep-cellMR can be extended to allow for interactions. For example, the model could be extended to study sex-differentiated genetic regulation across tissues/cell types. Another area of future development is to consider correlated horizontal pleiotropy in the model. Moreover, the deep-cellMR model could be expanded to incorporate additional information when performing deep multi-view learning. For example, a supervised version of the deep learning models could be used to boost desired patterns related to specific mechanisms, supervised by associations to known covariates and/or risk factors.

CHAPTER 5

SUMMARY AND FUTURE DIRECTIONS

5.1 Summary

In this work, we developed several statistical methods for integrative multi-omics multi-context association analyses and Mendelian randomization. The proposed methods take summary statistics as input and jointly analyze the effects of genetic variants on molecular traits or molecular traits on complex diseases/traits across multiple contexts and omics.

We propose X-ING as a general framework for the integration of cross-omics and cross-context summary statistics for association analysis. X-ING takes as input the summary statistic matrices from multiple omics data types. Instead of direct modeling of effect sizes, X-ING models each input statistic as a product of Gaussian and latent binary association status. This allows the integration of different data types, which may have different effect magnitudes and distributions. Via multi-view learning, X-ING captures omics-shared and context-shared association patterns. This is a major innovation compared with existing multi-context/tissue methods analyzing only one data type at a time. X-ING also models sample overlapping across contexts and allows for different levels of sparsity in each context. Through simulation studies, we demonstrate improved performance of estimation of association probabilities in X-ING brought by borrowing strengths across different data types and contexts.

In Chapter 3, we propose mintMR to perform integrative multi-context Mendelian randomization. MintMR addresses unique challenges in TWMR analysis. It performs a multi-tissue MR analysis using QTLs as IVs for each gene region. With a limited number of IVs, mintMR improves the estimation of tissue-specific causal effects of all genes by simultaneously modeling the latent disease-relevance context/tissue indicators for multiple gene regions. MintMR jointly learns the low-rank patterns of latent indicators/probabilities via multi-view learning techniques and then uses the major patterns to update the probability of

non-zero effects. The joint learning of disease-relevance of latent tissue indicators improves the estimation of sparse tissue-specific causal effects for all genes. By selecting cross-tissue QTLs as IVs, mintMR improves IV consistency. MintMR reduces confounding due to correlated cis molecular traits when mapping causal genes by including DNAm as an exposure. Simulations show that mintMR can control the type I error rates and has good powers in various settings, even when there are a limited number of QTLs as IVs and the causal effects are sparse.

To further map risk genes in specific cell types and to elucidate mechanisms underlying diseases and traits at the cellular level, we propose a deep learning Mendelian randomization method, deep-cellMR in Chapter 4. Deep-cellMR integrates sc-eQTL, ct-eQTLs from bulk tissue with GWAS to unveil risk genes in specific cell types. Deep-cellMR performs an MR analysis using sc-eQTLs and ct-eQTLs as IVs for each gene region. It allows the flexible inclusion of different cell types for each gene and jointly models multiple genes. Deep-cellMR improves the estimation of cell-type-specific causal effects of all analyzed genes by modeling the latent effect indicators. It uses deep multi-view learning methods to capture linear or nonlinear patterns of latent indicators/probabilities across genes, cells, and data. Using deep learning models, deep-cellMR handles complicated dependencies or structures and improves the estimation of sparse cell-type-specific causal effects for all genes.

5.2 Future Directions

In this section, we discuss a few potential extensions for current methods.

As the volume of summary statistics increases, X-ING can be improved with a more efficient and selective data integration. When the number of contexts is high (e.g., $K_\ell \geq 50$), or when the number of available sets of summary statistics is high, the computation efficiency of X-ING can be improved. Another potential area of future development is the integration of association statistics with mediation and causal estimates from multiple studies to reduce

confounding and spurious associations. Also, the current X-ING does not allow missing values. Future development of missing value imputation within the framework of X-ING can be explored.

Our integrative multi-context Mendelian randomization method, mintMR, can be extended to allow for correlated horizontal pleiotropy by identifying IVs with such effects. Another area of future development is to model major patterns of disease relevance indicators by adopting other advanced multi-view learning techniques. For example, we may use supervised multi-view learning methods for promoting other desirable patterns among examined genes related to specific known risk factors or mechanisms [Andrew et al., 2013, Yin and Sun, 2019, Wang et al., 2015]. Moreover, the mintMR model could be further expanded to model interaction effects or nonlinear causal effects of exposures on outcome.

Deep-cellMR models the effects of risk genes on complex traits/diseases in cell types. It can be further extended to allow for interactions. For example, the model could be extended to explore sex-differentiated genetic regulation across tissues/cell types. Another area of future development is to consider correlated horizontal pleiotropy in the model. Additionally, the deep-cellMR model could also be expanded to incorporate supervised deep learning methods for boosting specific patterns.

5.3 Conclusions

In conclusion, this dissertation detailed the development of several methods for integrative multi-omics multi-context analysis. The development of these methods was driven by underaddressed challenges in existing literature or underpowered analyses given existing resources. The proposed integrative X-ING association method enables cross-omics and cross-context analysis and simultaneously accounts for omics-shared and omics-specific tissue-shared patterns. We showed improved power, detection, replication, and functional interpretation of QTLs. Our multivariable MR method, mintMR, addresses challenges in transcriptome-wide

MR. It works well on data with limited numbers of IVs and prevalent pleiotropy. MintMR identifies risk genes adjusting for confounding and elucidates the potential pathways. The deep-cellMR method further extends MR to identify risk genes at cell-type resolution by employing novel deep learning methods, capturing more complicated dependency among cell types. As the landscape of genetic and genomic studies in understanding mechanisms underlying complex diseases is developing rapidly, a variety of those studies provides rich resources for integrative analysis. All of our proposed models take summary statistics as input. They can be easily adapted to the changing landscape and integrate summary data from various sources. We developed flexible and efficient software to implement the proposed methods and to facilitate future research on identifying novel biological mechanisms.

REFERENCES

Raúl Aguirre-Gamboa, Niek de Klein, Jennifer di Tommaso, Annique Claringbould, Monique GP van der Wijst, Dylan de Vries, Harm Brugge, Roy Oelen, Urmo Võsa, Maria M Zorro, et al. Deconvolution of bulk blood eQTL effects into immune cell subpopulations. *BMC bioinformatics*, 21:1–23, 2020.

Frank Wolfgang Albert, Joshua S Bloom, Jake Siegel, Laura Day, and Leonid Kruglyak. Genetics of trans-regulatory variation in gene expression. *Elife*, 7:e35471, 2018.

Emma L Anderson, Laura D Howe, Kaitlin H Wade, Yoav Ben-Shlomo, W David Hill, Ian J Deary, Eleanor C Sanderson, Jie Zheng, Roxanna Korologou-Linden, Evie Stergiakouli, et al. Education, intelligence and Alzheimer's disease: evidence from a multivariable two-sample Mendelian randomization study. *International Journal of Epidemiology*, 49(4): 1163–1172, 2020. https://doi.org/10.1093/ije/dyz280.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, volume 28, pages 1247–1255, 2013. https://proceedings.mlr.press/v28/andrew13.html.

Daniel Asby, Delphine Boche, Stuart Allan, Seth Love, and J Scott Miners. Systemic infection exacerbates cerebrovascular dysfunction in Alzheimer's disease. *Brain*, 144(6): 1869–1883, 2021.

Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000. https://doi.org/10.1038/75556.

Richard Barfield, Helian Feng, Alexander Gusev, Lang Wu, Wei Zheng, Bogdan Pasaniuc, and Peter Kraft. Transcriptome-wide association studies accounting for colocalization using Egger regression. *Genetic epidemiology*, 42(5):418–433, 2018. https://doi.org/10.1002/gepi.22131.

Alexis Battle, Christopher D Brown, Barbara E Engelhardt, and Stephen B Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, 2017.

David A Bennett, Aron S Buchman, Patricia A Boyle, Lisa L Barnes, Robert S Wilson, and Julie A Schneider. Religious orders study and rush memory and aging project. *Journal of Alzheimer's Disease*, 64(s1):S161–S189, 2018. https://doi.org/10.3233/JAD-179939.

Amy Blum, Peggy Wang, and Jean C Zenklusen. SnapShot: TCGA-analyzed tumors. *Cell*, 173(2):530, 2018.

Jack Bowden, George Davey Smith, and Stephen Burgess. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology*, 44(2):512–525, 2015. https://doi.org/10.1093/ije/dyv080.

Evan A Boyle, Yang I Li, and Jonathan K Pritchard. An expanded view of complex traits: from polygenic to omnigenic. *Cell*, 169(7):1177–1186, 2017.

Julien Bryois, Daniela Calini, Will Macnair, Lynette Foo, Eduard Urich, Ward Ortmann, Victor Alejandro Iglesias, Suresh Selvaraj, Erik Nutma, Manuel Marzin, et al. Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nature neuroscience*, 25(8):1104–1112, 2022.

Andreas Buja and Nermin Eyuboglu. Remarks on parallel analysis. *Multivariate Behavioral Research*, 27(4):509–540, 1992.

Brendan K Bulik-Sullivan, Po-Ru Loh, Hilary K Finucane, Stephan Ripke, Jian Yang, Nick Patterson, Mark J Daly, Alkes L Price, and Benjamin M Neale. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics*, 47(3):291–295, 2015.

Annalisa Buniello, Jacqueline A L MacArthur, Maria Cerezo, Laura W Harris, James Hayhurst, Cinzia Malangone, Aoife McMahon, Joannella Morales, Edward Mountjoy, Elliot Sollis, et al. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research*, 47(D1):D1005–D1012, 2019.

Stephen Burgess and Simon G Thompson. Multivariable Mendelian randomization: the use of pleiotropic genetic variants to estimate causal effects. *American Journal of Epidemiology*, 181(4):251–260, 2015. https://doi.org/10.1093/aje/kwu283.

Stephen Burgess, Adam Butterworth, and Simon G Thompson. Mendelian randomization analysis with multiple genetic variants using summarized data. *Genetic epidemiology*, 37 (7):658–665, 2013. https://doi.org/10.1002/gepi.21758.

Stephen Burgess, Robert A Scott, Nicholas J Timpson, George Davey Smith, Simon G Thompson, and EPIC-InterAct Consortium. Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *European Journal of Epidemiology*, 30:543–552, 2015. https://doi.org/10.1007/s10654-015-0011-z.

Benjamin Carter and Keji Zhao. The epigenetic basis of cellular heterogeneity. *Nature Reviews Genetics*, 22(4):235–250, 2021.

Lin S Chen, Frank Emmert-Streib, and John D Storey. Harnessing naturally randomized transcription to infer regulatory relationships among genes. *Genome biology*, 8(10):1–13, 2007. https://doi.org/10.1186/gb-2007-8-10-r219.

Qing Cheng, Xiao Zhang, Lin S Chen, and Jin Liu. Mendelian randomization accounting for complex correlated horizontal pleiotropy while elucidating shared genetic etiology. *Nature communications*, 13(1):6490, 2022. https://doi.org/10.1038/s41467-022-34164-1.

GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. https://doi.org/10.1126/science.aaz1776.

GTEx Consortium et al. Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509):eaaz8528, 2020.

George Davey Smith and Shah Ebrahim. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology*, 32(1):1–22, 2003. https://doi.org/10.1093/ije/dyg070.

Fernanda G De Felice, Rafaella A Gonçalves, and Sergio T Ferreira. Impaired insulin signalling and allostatic load in Alzheimer disease. *Nature Reviews Neuroscience*, 23(4): 215–230, 2022.

Dylan H de Vries, Vasiliki Matzaraki, Olivier B Bakker, Harm Brugge, Harm-Jan Westra, Mihai G Netea, Lude Franke, Vinod Kumar, and Monique GP van der Wijst. Integrating GWAS with bulk and single-cell RNA-sequencing reveals a role for LY86 in the anti-Candida host response. *PLoS pathogens*, 16(4):e1008408, 2020.

Ian J Deary, Alan J Gow, Alison Pattie, and John M Starr. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *International Journal of Epidemiology*, 41(6):1576–1584, 2012. https://doi.org/10.1093/ije/dyr197.

Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55 (4):997–1004, 1999. https://doi.org/10.1111/j.0006-341X.1999.00997.x.

Gilbert Di Paolo and Tae-Wan Kim. Linking lipids to Alzheimer's disease: cholesterol and beyond. *Nature Reviews Neuroscience*, 12(5):284–296, 2011. https://doi.org/10.1038/nrn3012.

Ruofan Ding, Xudong Zou, Yangmei Qin, Lihai Gong, Hui Chen, Xuelian Ma, Shouhong Guang, Chen Yu, Gao Wang, and Lei Li. xQTLbiolinks: a comprehensive and scalable tool for integrative analysis of molecular QTLs. *Briefings in Bioinformatics*, 25(1):bbad440, 2024.

Margaret KR Donovan, Agnieszka D'Antonio-Chronowska, Matteo D'Antonio, and Kelly A Frazer. Cellular deconvolution of GTEx tissues powers discovery of disease and cell-type associated regulatory variants. *Nature communications*, 11(1):955, 2020.

Ruben Dries, Jiaji Chen, Natalie Del Rossi, Mohammed Muzamil Khan, Adriana Sistig, and Guo-Cheng Yuan. Advances in spatial transcriptomic data analysis. *Genome research*, 31 (10):1706–1718, 2021.

Harikesh Dubey, Kavita Gulati, and Arunabha Ray. Alzheimer's disease: A contextual link with nitric oxide synthase. *Current Molecular Medicine*, 20(7):505–515, 2020.

eGTEx Project. Enhancing GTEx by bridging the gaps between genotype, gene expression, and disease. *Nature genetics*, 49(12):1664–1670, 2017.

Gökcen Eraslan, Eugene Drokhlyansky, Shankara Anand, Evgenij Fiskin, Ayshwarya Subramanian, Michal Slyper, Jiali Wang, Nicholas Van Wittenberghe, John M Rouhana, Julia Waldman, et al. Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science*, 376(6594):eabl4290, 2022.

Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, et al. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46(D1):D649–D655, 2018. https://doi.org/10.1093/nar/gkx1132.

Helian Feng, Nicholas Mancuso, Alexander Gusev, Arunabha Majumdar, Megan Major, Bogdan Pasaniuc, and Peter Kraft. Leveraging expression from multiple tissues using sparse canonical correlation analysis and aggregate tests improves the power of transcriptome-wide association studies. *PLoS genetics*, 17(4):e1008973, 2021. https://doi.org/10.1371/journal.pgen.1008973.

Hilary K Finucane, Yakir A Reshef, Verneri Anttila, Kamil Slowikowski, Alexander Gusev, Andrea Byrnes, Steven Gazal, Po-Ru Loh, Caleb Lareau, Noam Shoresh, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature genetics*, 50(4):621–629, 2018. https://doi.org/10.1038/s41588-018-0081-4.

Christopher N Foley, James R Staley, Philip G Breen, Benjamin B Sun, Paul DW Kirk, Stephen Burgess, and Joanna MM Howson. A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nature communications*, 12(1):1–18, 2021. https://doi.org/10.1038/s41467-020-20885-8.

Scott B Franklin, David J Gibson, Philip A Robertson, John T Pohlmann, and James S Fralish. Parallel analysis: a method for determining significant principal components. *Journal of Vegetation Science*, 6(1):99–106, 1995.

Alma Y Galvez-Contreras, David Zarate-Lopez, Ana L Torres-Chavez, and Oscar Gonzalez-Perez. Role of oligodendrocytes and myelin in the pathophysiology of autism spectrum disorder. *Brain Sciences*, 10(12):951, 2020.

Eric R Gamazon, Ayellet V Segrè, Martijn Van De Bunt, Xiaoquan Wen, Hualin S Xi, Farhad Hormozdiari, Halit Ongen, Anuar Konkashbaev, Eske M Derks, François Aguet, et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nature genetics*, 50(7):956–967, 2018.

Eric R Gamazon, Aeilko H Zwinderman, Nancy J Cox, Damiaan Denys, and Eske M Derks. Multi-tissue transcriptome analyses identify genetic mechanisms underlying neuropsychiatric traits. *Nature genetics*, 51(6):933–940, 2019. https://doi.org/10.1038/s41588-019-0409-8.

Tom R Gaunt, Hashem A Shihab, Gibran Hemani, Josine L Min, Geoff Woodward, Oliver Lyttleton, Jie Zheng, Aparna Duggirala, Wendy L McArdle, Karen Ho, et al. Systematic identification of genetic influences on methylation across the human life course. *Genome biology*, 17(1):1–14, 2016.

Claudia Giambartolomei, Damjan Vukcevic, Eric E Schadt, Lude Franke, Aroon D Hingorani, Chris Wallace, and Vincent Plagnol. Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics*, 10(5):e1004383, 2014.

Claudia Giambartolomei, Jimmy Zhenli Liu, Wen Zhang, Mads Hauberg, Huwenbo Shi, James Boocock, Joe Pickrell, Andrew E Jaffe, CommonMind Consortium, Bogdan Pasaniuc, et al. A Bayesian framework for multiple trait colocalization from summary association statistics. *Bioinformatics*, 34(15):2538–2545, 2018.

Kevin J Gleason, Fan Yang, Brandon L Pierce, Xin He, and Lin S Chen. Primo: integration of multiple GWAS and omics QTL summary statistics for elucidation of molecular mechanisms of trait-associated SNPs and detection of pleiotropy in complex traits. *Genome biology*, 21(1):1–24, 2020. https://doi.org/10.1186/s13059-020-02125-w.

Kevin J Gleason, Fan Yang, and Lin S Chen. A robust two-sample transcriptome-wide Mendelian randomization method integrating GWAS with multi-tissue eQTL summary statistics. *Genetic epidemiology*, 45(4):353–371, 2021. https://doi.org/10.1002/gepi.22380.

Paula Grammas. Neurovascular dysfunction, inflammation and endothelial activation: implications for the pathogenesis of Alzheimer's disease. *Journal of neuroinflammation*, 8 (1):1–12, 2011.

Ralph S Grand, Lukas Burger, Cathrin Gräwe, Alicia K Michael, Luke Isbel, Daniel Hess, Leslie Hoerner, Vytautas Iesmantavicius, Sevi Durdu, Marco Pregnolato, et al. BANP opens chromatin and activates CpG-island-regulated genes. *Nature*, 596(7870):133–137, 2021.

Andrew J Grant and Stephen Burgess. Pleiotropy robust methods for multivariable Mendelian randomization. *Statistics in Medicine*, 40(26):5813–5830, 2021. https://doi.org/10.1002/sim.9156.

Steven M Greenberg, Brian J Bacskai, Mar Hernandez-Guillamon, Jeremy Pruzin, Reisa Sperling, and Susanne J van Veluw. Cerebral amyloid angiopathy and Alzheimer disease—one peptide, two pathways. *Nature Reviews Neurology*, 16(1):30–42, 2020.

Hui Guo, Mary D Fortune, Oliver S Burren, Ellen Schofield, John A Todd, and Chris Wallace. Integration of disease association and eQTL data using a Bayesian colocalisation approach highlights six candidate causal genes in immune-mediated diseases. *Human Molecular Genetics*, 24(12):3305–3313, 2015. https://doi.org/10.1093/hmg/ddv077.

Buhm Han and Eleazar Eskin. Interpreting meta-analyses of genome-wide association studies. *PLoS genetics*, 8(3):e1002555, 2012.

Weiping Han and Cai Li. Linking type 2 diabetes and Alzheimer's disease. *Proceedings of the National Academy of Sciences*, 107(15):6557–6558, 2010.

Liang He, Yury Loika, Yongjin Park, Genotype Tissue Expression (GTEx) consortium, David A Bennett, Manolis Kellis, Alexander M Kulminski, and Alzheimer's Disease Neuroimaging Initiative. Exome-wide age-of-onset analysis reveals exonic variants in ERN1 and SPPL2C associated with Alzheimer's disease. *Translational psychiatry*, 11(1):146, 2021.

Idan Hekselman and Esti Yeger-Lotem. Mechanisms of tissue and cell-type specificity in heritable traits and diseases. *Nature Reviews Genetics*, 21(3):137–150, 2020.

Michael T Heneka, Douglas T Golenbock, and Eicke Latz. Innate immunity in Alzheimer's disease. *Nature immunology*, 16(3):229–236, 2015. `https://doi.org/10.1038/ni.3102`.

Frank L Heppner, Richard M Ransohoff, and Burkhard Becher. Immune attack: the role of inflammation in Alzheimer disease. *Nature Reviews Neuroscience*, 16(6):358–372, 2015. `https://doi.org/10.1038/nrn3880`.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell RNA-sequencing experiments. *Biostatistics*, 19(4): 562–578, 2018.

Lucia A Hindorff. A catalog of published genome-wide association studies. *http://www. genome. gov/26525384*, 2009.

Farhad Hormozdiari, Steven Gazal, Bryce Van De Geijn, Hilary K Finucane, Chelsea J-T Ju, Po-Ru Loh, Armin Schoech, Yakir Reshef, Xuanyao Liu, Luke O'connor, et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics*, 50(7):1041–1047, 2018.

Yiming Hu, Mo Li, Qiongshi Lu, Haoyi Weng, Jiawei Wang, Seyedeh M Zekavat, Zhaolong Yu, Boyang Li, Jianlei Gu, Sydney Muchnik, et al. A statistical framework for cross-tissue transcriptome-wide association analysis. *Nature genetics*, 51(3):568–576, 2019. `https://doi.org/10.1038/s41588-019-0345-7`.

Byungjin Hwang, Ji Hyun Lee, and Duhee Bang. Single-cell RNA sequencing technologies and bioinformatics pipelines. *Experimental & molecular medicine*, 50(8):1–14, 2018.

Julie Jerber, Daniel D Seaton, Anna SE Cuomo, Natsuhiko Kumasaka, James Haldane, Juliette Steer, Minal Patel, Daniel Pearce, Malin Andersson, Marc Jan Bonder, et al. Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nature genetics*, 53(3):304–312, 2021.

Yu-Chia Kao, Pei-Chuan Ho, Yuan-Kun Tu, I-Ming Jou, and Kuen-Jer Tsai. Lipids and Alzheimer's disease. *International Journal of Molecular Sciences*, 21(4):1505, 2020. https://doi.org/10.3390/ijms21041505.

Zvonimir S Katusic, Livius V d'Uscio, and Tongrong He. Emerging Roles of Endothelial Nitric Oxide in Preservation of Cognitive Health. *Stroke*, 54(3):686–696, 2023.

Sarah Kim-Hellmuth, François Aguet, Meritxell Oliva, Manuel Muñoz-Aguirre, Silva Kasela, Valentin Wucher, Stephane E. Castel, Andrew R. Hamel, Ana Viñuela, Amy L. Roberts, Serghei Mangul, Xiaoquan Wen, Gao Wang, Alvaro N. Barbeira, Diego Garrido-Martín, Brian B. Nadel, Yuxin Zou, Rodrigo Bonazzola, Jie Quan, Andrew Brown, Angel Martinez-Perez, José Manuel Soria, Gad Getz, Emmanouil T. Dermitzakis, Kerrin S. Small, Matthew Stephens, Hualin S. Xi, Hae Kyung Im, Roderic Guigó, Ayellet V. Segrè, Barbara E. Stranger, Kristin G. Ardlie, Tuuli Lappalainen, Shankara Anand, Stacey Gabriel, Gad A. Getz, Aaron Graubert, Kane Hadley, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, ..., Meng Wang, Latarsha J. Carithers, Ping Guan, Susan E. Koester, A. Roger Little, Helen M. Moore, Concepcion R. Nierras, Abhi K. Rao, Jimmie B. Vaught, and Simona Volpi. Cell type–specific genetic regulation of gene expression across human tissues. *Science*, 369(6509):eaaz8528, 2020. doi:10.1126/science.aaz8528.

Holger Kirsten, Hoor Al-Hasani, Lesca Holdt, Arnd Gross, Frank Beutner, Knut Krohn, Katrin Horn, Peter Ahnert, Ralph Burkhardt, Kristin Reiche, et al. Dissecting the genetics of the human transcriptome identifies novel trait-related trans-eQTLs and corroborates the regulatory relevance of non-protein coding loci. *Human Molecular Genetics*, 24(16): 4746–4763, 2015.

Debbie A Lawlor, Roger M Harbord, Jonathan AC Sterne, Nic Timpson, and George Davey Smith. Mendelian randomization: using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine*, 27(8):1133–1163, 2008. https://doi.org/10.1002/sim.3034.

Friedrich Leisch, Andreas Weingessel, and Maintainer Friedrich Leisch. The bindata package. *Citeseer*, 2006.

Gang Li, Depeng Han, Chao Wang, Wenxing Hu, Vince D Calhoun, and Yu-Ping Wang. Application of deep canonically correlated sparse autoencoder for the classification of schizophrenia. *Computer Methods and Programs in Biomedicine*, 183:105073, 2020. https://doi.org/10.1016/j.cmpb.2019.105073.

Gen Li, Andrey A Shabalin, Ivan Rusyn, Fred A Wright, and Andrew B Nobel. An empirical Bayes approach for multiple tissue eQTL analysis. *Biostatistics*, 19(3):391–406, 2018.

Minghui Li, Yan Gao, Dandan Wang, Xiaowen Hu, Jie Jiang, Ying Qing, Xuhan Yang, Gaoping Cui, Pengkun Wang, Juan Zhang, et al. Impaired membrane lipid homeostasis in schizophrenia. *Schizophrenia Bulletin*, 48(5):1125–1135, 2022. https://doi.org/10.1093/schbul/sbac011.

Zhaotong Lin, Haoran Xue, and Wei Pan. Robust multivariable Mendelian randomization based on constrained maximum likelihood. *The American Journal of Human Genetics*, 110(4):592–605, 2023. https://doi.org/10.1016/j.ajhg.2023.02.014.

Jin Liu, Xiang Wan, Chaolong Wang, Chao Yang, Xiaowei Zhou, and Can Yang. LLR: a latent low-rank approach to colocalizing genetic risk variants in multiple GWAS. *Bioinformatics*, 33(24):3878–3886, 2017.

Jiyuan Liu, Xinwang Liu, Yuexiang Yang, Qing Liao, and Yuanqing Xia. Contrastive Multi-View Kernel Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 2023. https://doi.org/10.1109/TPAMI.2023.3253211.

Ruth JF Loos. 15 years of genome-wide association studies and no signs of slowing down. *Nature communications*, 11(1):1–3, 2020.

Katia de Paiva Lopes, Gijsje JL Snijders, Jack Humphrey, Amanda Allan, Marjolein AM Sneeboer, Elisa Navarro, Brian M Schilder, Ricardo A Vialle, Madison Parks, Roy Missall, et al. Genetic analysis of the human microglial transcriptome across brain regions, aging and disease pathologies. *Nature genetics*, 54(1):4–17, 2022.

Justin M Luningham, Junyu Chen, Shizhen Tang, Philip L De Jager, David A Bennett, Aron S Buchman, and Jingjing Yang. Bayesian genome-wide TWAS method to leverage both cis-and trans-eQTL information through summary statistics. *The American Journal of Human Genetics*, 107(4):714–726, 2020.

Mary-Ellen Lynall, Blagoje Soskic, James Hayhurst, Jeremy Schwartzentruber, Daniel F Levey, Gita A Pathak, Renato Polimanti, Joel Gelernter, Murray B Stein, Gosia Trynka, et al. Genetic variants associated with psychiatric disorders are enriched at epigenetically active sites in lymphoid cells. *Nature communications*, 13(1):6102, 2022.

Fan Ma, Deyu Meng, Xuanyi Dong, and Yi Yang. Self-paced multi-view co-training. *Journal of Machine Learning Research*, 21(57):1–38, 2020. https://jmlr.org/papers/v21/18-794.html.

Jacqueline MacArthur, Emily Bowler, Maria Cerezo, Laurent Gil, Peggy Hall, Emma Hastings, Heather Junkins, Aoife McMahon, Annalisa Milano, Joannella Morales, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, 2017.

Jacqueline AL MacArthur, Annalisa Buniello, Laura W Harris, James Hayhurst, Aoife McMahon, Elliot Sollis, Maria Cerezo, Peggy Hall, Elizabeth Lewis, Patricia L Whetzel, et al. Workshop proceedings: GWAS summary statistics standards and sharing. *Cell Genomics*, 1(1):100004, 2021.

Kristen R Maynard, Leonardo Collado-Torres, Lukas M Weber, Cedric Uytingco, Brianna K Barry, Stephen R Williams, Joseph L Catallini, Matthew N Tran, Zachary Besich, Madhavi Tippani, et al. Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nature neuroscience*, 24(3):425–436, 2021.

Allan F McRae, Joseph E Powell, Anjali K Henders, Lisa Bowdler, Gibran Hemani, Sonia Shah, Jodie N Painter, Nicholas G Martin, Peter M Visscher, and Grant W Montgomery. Contribution of genetic variation to transgenerational inheritance of DNA methylation. *Genome biology*, 15:1–10, 2014. https://doi.org/10.1186/gb-2014-15-5-r73.

Allan F McRae, Riccardo E Marioni, Sonia Shah, Jian Yang, Joseph E Powell, Sarah E Harris, Jude Gibson, Anjali K Henders, Lisa Bowdler, Jodie N Painter, et al. Identification of 55,000 replicated DNA methylation QTL. *Scientific Reports*, 8(1):17605, 2018.

Maite Mendioroz, Marta Puebla-Guedea, Jesús Montero-Marín, Amaya Urdánoz-Casado, Idoia Blanco-Luquin, Miren Roldán, Alberto Labarga, and Javier García-Campayo. Telomere length correlates with subtelomeric DNA methylation in long-term mindfulness practitioners. *Scientific Reports*, 10(1):4564, 2020.

Josine L Min, Gibran Hemani, Eilis Hannon, Koen F Dekkers, Juan Castillo-Fernandez, René Luijk, Elena Carnero-Montoro, Daniel J Lawson, Kimberley Burrows, Matthew Suderman, et al. Genomic and phenotypic insights from an atlas of genetic effects on DNA methylation. *Nature genetics*, 53(9):1311–1321, 2021.

Belal A Mohamed, Amal Z Barakat, Wolfram-Hubertus Zimmermann, Reginald E Bittner, Christian Mühlfeld, Mark Hünlich, Wolfgang Engel, Lars S Maier, and Ibrahim M Adham. Targeted disruption of Hspa4 gene leads to cardiac hypertrophy and fibrosis. *Journal of Molecular and Cellular Cardiology*, 53(4):459–468, 2012. https://doi.org/10.1016/j.yjmcc.2012.07.014.

Jean Morrison, Nicholas Knoblauch, Joseph H Marcus, Matthew Stephens, and Xin He. Mendelian randomization accounting for correlated and uncorrelated pleiotropic effects using genome-wide summary statistics. *Nature genetics*, 52(7):740–747, 2020. https://doi.org/10.1038/s41588-020-0631-4.

Aparna Nathan, Samira Asgari, Kazuyoshi Ishigaki, Cristian Valencia, Tiffany Amariuta, Yang Luo, Jessica I Beynor, Yuriy Baglaenko, Sara Suliman, Alkes L Price, et al. Single-cell eQTL models reveal dynamic T cell state dependence of disease loci. *Nature*, 606 (7912):120–128, 2022.

Aaron M Newman, Chloé B Steen, Chih Long Liu, Andrew J Gentles, Aadel A Chaudhuri, Florian Scherer, Michael S Khodadoust, Mohammad S Esfahani, Bogdan A Luca, David Steiner, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019.

Bernard Ng, William Casazza, Nam Hee Kim, Chendi Wang, Farnush Farhadi, Shinya Tasaki, David A Bennett, Philip L De Jager, Christopher Gaiteri, and Sara Mostafavi. Cascading epigenomic analysis for identifying disease genes from the regulatory landscape of GWAS variants. *PLoS genetics*, 17(11):e1009918, 2021.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696, 2011.

Roy Oelen, Dylan H de Vries, Harm Brugge, M Grace Gordon, Martijn Vochteloo, single-cell eQTLGen consortium, BIOS Consortium, Chun J Ye, Harm-Jan Westra, Lude Franke, et al. Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nature communications*, 13(1):3267, 2022.

Meritxell Oliva, Kathryn Demanelis, Yihao Lu, Meytal Chernoff, Farzana Jasmine, Habibul Ahsan, Muhammad G Kibriya, Lin S Chen, and Brandon L Pierce. DNA methylation QTL mapping across diverse human tissues provides molecular links between genetic variation and complex traits. *Nature genetics*, 55(1):112–122, 2023. `https://doi.org/10.1038/s41588-022-01248-z`.

Halit Ongen, Alfonso Buil, Andrew Anand Brown, Emmanouil T Dermitzakis, and Olivier Delaneau. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics*, 32(10):1479–1485, 2016.

Halit Ongen, Andrew A Brown, Olivier Delaneau, Nikolaos I Panousis, Alexandra C Nica, GTEx Consortium, and Emmanouil T Dermitzakis. Estimating the causal tissues for complex traits and diseases. *Nature genetics*, 49(12):1676–1683, 2017. `https://doi.org/10.1038/ng.3981`.

Richard K Perez, M Grace Gordon, Meena Subramaniam, Min Cheol Kim, George C Hartoularos, Sasha Targ, Yang Sun, Anton Ogorodnikov, Raymund Bueno, Andrew Lu, et al. Single-cell RNA-seq reveals cell type–specific molecular and genetic associations to lupus. *Science*, 376(6589):eabf1970, 2022.

Brandon L Pierce and Stephen Burgess. Efficient design for Mendelian randomization studies: subsample and 2-sample instrumental variable estimators. *American Journal of Epidemiology*, 178(7):1177–1184, 2013. `https://doi.org/10.1093/aje/kwt084`.

Brandon L Pierce, Lin Tong, Lin S Chen, Ronald Rahaman, Maria Argos, Farzana Jasmine, Shantanu Roy, Rachelle Paul-Brutus, Harm-Jan Westra, Lude Franke, et al. Mediation analysis demonstrates that trans-eQTLs are often explained by cis-mediation: a genome-wide analysis among 1,800 South Asians. *PLoS genetics*, 10(12):e1004818, 2014.

Brandon L Pierce, Lin Tong, Maria Argos, Kathryn Demanelis, Farzana Jasmine, Muhammad Rakibuz-Zaman, Golam Sarwar, Md Tariqul Islam, Hasan Shahriar, Tariqul Islam, et al. Co-occurring expression and methylation QTLs allow detection of common causal variants and shared biological mechanisms. *Nature communications*, 9(1):804, 2018. `https://doi.org/10.1038/s41467-018-03209-9`.

Milton Pividori, Padma S Rajagopal, Alvaro Barbeira, Yanyu Liang, Owen Melia, Lisa Bastarache, YoSon Park, GTEx Consortium, Xiaoquan Wen, and Hae K Im. PhenomeXcan:

Mapping the genome to the phenome through the transcriptome. *Science advances*, 6(37): eaba2083, 2020.

Allan-Hermann Pool, Helen Poldsam, Sisi Chen, Matt Thomson, and Yuki Oka. Recovery of missing single-cell RNA-sequencing data with optimized transcriptomic references. *Nature methods*, 20(10):1506–1515, 2023.

Joseph E Powell, Anjali K Henders, Allan F McRae, Anthony Caracella, Sara Smith, Margaret J Wright, John B Whitfield, Emmanouil T Dermitzakis, Nicholas G Martin, Peter M Visscher, et al. The Brisbane Systems Genetics Study: genetical genomics meets complex trait genetics. *PloS one*, 7(4):e35430, 2012. https://doi.org/10.1371/journal.pone.0035430.

Laura Prieto del Val, Jose L Cantero, and Mercedes Atienza. Atrophy of amygdala and abnormal memory-related alpha oscillations over posterior cingulate predict conversion to Alzheimer's disease. *Scientific Reports*, 6(1):1–12, 2016.

Ting Qi, Yang Wu, Jian Zeng, Futao Zhang, Angli Xue, Longda Jiang, Zhihong Zhu, Kathryn Kemper, Loic Yengo, Zhili Zheng, et al. Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature communications*, 9(1):2282, 2018.

Florian J Raabe, Sabrina Galinski, Sergi Papiol, Peter G Falkai, Andrea Schmitt, and Moritz J Rossner. Studying and modulating schizophrenia-associated dysfunctions of oligodendrocytes with patient-specific cell systems. *npj Schizophrenia*, 4(1):1–11, 2018.

Jessica MB Rees, Angela M Wood, and Stephen Burgess. Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy. *Statistics in Medicine*, 36(29):4705–4718, 2017. https://doi.org/10.1002/sim.7492.

Jason A Reuter, Damek V Spacek, and Michael P Snyder. High-throughput sequencing technologies. *Mol. Cell*, 58(4):586–597, 2015.

Tom G Richardson, Gibran Hemani, Tom R Gaunt, Caroline L Relton, and George Davey Smith. A transcriptome-wide Mendelian randomization study to uncover tissue-dependent regulatory mechanisms across the human phenome. *Nature communications*, 11(1):185, 2020. https://doi.org/10.1038/s41467-019-13921-9.

Bernardo Rodriguez-Iturbe, Richard J Johnson, Laura Gabriela Sanchez-Lozada, and Hector Pons. HSP70 and Primary Arterial Hypertension. *Biomolecules*, 13(2):272, 2023. https://doi.org/10.3390/biom13020272.

Eleanor Sanderson, George Davey Smith, Frank Windmeijer, and Jack Bowden. An examination of multivariable Mendelian randomization in the single-sample and two-sample summary data settings. *International Journal of Epidemiology*, 48(3):713–727, 2019. https://doi.org/10.1093/ije/dyy262.

Eric E Schadt, John Lamb, Xia Yang, Jun Zhu, Steve Edwards, Debraj GuhaThakurta, Solveig K Sieberts, Stephanie Monks, Marc Reitman, Chunsheng Zhang, et al. An integrative genomics approach to infer causal associations between gene expression and disease. *Nature genetics*, 37(7):710–717, 2005. https://doi.org/10.1038/ng1589.

Marc Scherlinger, Christophe Richez, George C Tsokos, Eric Boilard, and Patrick Blanco. The role of platelets in immune-mediated inflammatory diseases. *Nature Reviews Immunology*, pages 1–16, 2023. https://doi.org/10.1038/s41577-023-00869-7.

Jonas Schluter, Jonathan U Peled, Bradford P Taylor, Kate A Markey, Melody Smith, Ying Taur, Rene Niehus, Anna Staffas, Anqi Dai, Emily Fontana, et al. The gut microbiota is associated with immune cell dynamics in humans. *Nature*, 588(7837):303–307, 2020. https://doi.org/10.1038/s41586-020-2971-8.

Lulu Shang, Jennifer A Smith, and Xiang Zhou. Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. *PLoS genetics*, 16(4): e1008734, 2020. https://doi.org/10.1371/journal.pgen.1008734.

Chao Shi and Eric G Pamer. Monocyte recruitment during infection and inflammation. *Nature Reviews Immunology*, 11(11):762–774, 2011. https://doi.org/10.1038/nri3070.

Xingjie Shi, Yuling Jiao, Yi Yang, Ching-Yu Cheng, Can Yang, Xinyi Lin, and Jin Liu. VIMCO: variational inference for multiple correlated outcomes in genome-wide association studies. *Bioinformatics*, 35(19):3693–3700, 2019.

Xingjie Shi, Xiaoran Chai, Yi Yang, Qing Cheng, Yuling Jiao, Haoyue Chen, Jian Huang, Can Yang, and Jin Liu. A tissue-specific collaborative mixed model for jointly analyzing multiple tissues in transcriptome-wide association studies. *Nucleic Acids Research*, 48(19): e109–e109, 2020. https://doi.org/10.1093/nar/gkaa767.

Nathan G Skene, Julien Bryois, Trygve E Bakken, Gerome Breen, James J Crowley, Héléna A Gaspar, Paola Giusti-Rodriguez, Rebecca D Hodge, Jeremy A Miller, Ana B Muñoz-Manchado, et al. Genetic identification of brain cell types underlying schizophrenia. *Nature genetics*, 50(6):825–833, 2018.

Eric AW Slob and Stephen Burgess. A comparison of robust Mendelian randomization methods using summary data. *Genetic epidemiology*, 44(4):313–329, 2020. https://doi.org/10.1002/gepi.22295.

Blagoje Soskic, Eddie Cano-Gamez, Deborah J Smyth, Kirsty Ambridge, Ziying Ke, Julie C Matte, Lara Bossini-Castillo, Joanna Kaplanis, Lucia Ramirez-Navarro, Anna Lorenc, et al. Immune disease risk variants regulate gene expression dynamics during CD4+ T cell activation. *Nature genetics*, 54(6):817–826, 2022.

Jamie B Spangler, Ignacio Moraga, Juan L Mendoza, and K Christopher Garcia. Insights into cytokine–receptor interactions from cytokine engineering. *Annual review of immunology*, 33:139–167, 2015.

Oliver Stegle, Leopold Parts, Matias Piipari, John Winn, and Richard Durbin. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols*, 7(3):500–507, 2012.

Daniel J Stekhoven and Peter Bühlmann. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

Indhupriya Subramanian, Srikant Verma, Shiva Kumar, Abhay Jere, and Krishanpal Anamika. Multi-omics data integration, interpretation, and its application. *Bioinf. Biol. Insights*, 14:1177932219899051, 2020.

Benjamin B Sun, Joseph C Maranville, James E Peters, David Stacey, James R Staley, James Blackshaw, Stephen Burgess, Tao Jiang, Ellie Paige, Praveen Surendran, et al. Genomic atlas of the human plasma proteome. *Nature*, 558(7708):73–79, 2018.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, et al. mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

D Leland Taylor, Anne U Jackson, Narisu Narisu, Gibran Hemani, Michael R Erdos, Peter S Chines, Amy Swift, Jackie Idol, John P Didion, Ryan P Welch, et al. Integrative analysis of gene expression, DNA methylation, physiological traits, and genetic variation in human skeletal muscle. *Proceedings of the National Academy of Sciences*, 116(22):10883–10888, 2019. `https://doi.org/10.1073/pnas.1814263116`.

Arthur Tenenhaus and Michel Tenenhaus. Regularized generalized canonical correlation analysis. *Psychometrika*, 76(2):257, 2011.

SN Thibodeau, AJ French, SK McDonnell, J Cheville, S Middha, L Tillmans, S Riska, S Baheti, MC Larson, Z Fogarty, et al. Identification of candidate genes for prostate cancer-risk SNPs utilizing a normal prostate tissue eQTL data set. *Nature communications*, 6(1): 1–10, 2015.

Benjamin D Umans, Alexis Battle, and Yoav Gilad. Where are the disease-associated eQTLs? *Trends in Genetics*, 37(2):109–124, 2021.

Sarah M Urbut, Gao Wang, Peter Carbonetto, and Matthew Stephens. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nature genetics*, 51(1):187–195, 2019. `https://doi.org/10.1038/s41588-018-0268-8`.

Monique GP van der Wijst, Dylan H de Vries, Hilde E Groot, Gosia Trynka, Chung-Chau Hon, Marc-Jan Bonder, Oliver Stegle, MC Nawijn, Youssef Idaghdour, Pim van der Harst, et al. The single-cell eQTLGen consortium. *elife*, 9:e52155, 2020.

Marie Verbanck, Chia-Yen Chen, Benjamin Neale, and Ron Do. Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases. *Nature genetics*, 50(5):693–698, 2018. https://doi.org/10.1038/s41588-018-0099-7.

Urmo Võsa, Annique Claringbould, Harm-Jan Westra, Marc Jan Bonder, Patrick Deelen, Biao Zeng, Holger Kirsten, Ashis Saha, Roman Kreuzhuber, Seyhan Yazar, et al. Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. *Nature genetics*, 53(9):1300–1310, 2021. https://doi.org/10.1038/s41588-021-00913-z.

Trung Nghia Vu, Quin F Wills, Krishna R Kalari, Nifang Niu, Liewei Wang, Mattias Rantalainen, and Yudi Pawitan. Beta-Poisson model for single-cell RNA-seq data analyses. *Bioinformatics*, 32(14):2128–2135, 2016.

Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, Alvaro N Barbeira, David A Knowles, David Golan, Raili Ermel, Arno Ruusalepp, Thomas Quertermous, Ke Hao, et al. Opportunities and challenges for transcriptome-wide association studies. *Nature genetics*, 51(4):592–599, 2019.

Jingshu Wang, Qingyuan Zhao, Jack Bowden, Gibran Hemani, George Davey Smith, Dylan S Small, and Nancy R Zhang. Causal inference for heritable phenotypic risk factors using heterogeneous genetic instruments. *PLoS genetics*, 17(6):e1009575, 2021. https://doi.org/10.1371/journal.pgen.1009575.

Litao Wang, Qiong Liu, Dongqi Yue, Jun Liu, and Yi Fu. Cerebral Amyloid Angiopathy: An Undeniable Small Vessel Disease. *Journal of Stroke*, 26(1):1, 2024.

Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083–1092. PMLR, 2015.

Weiran Wang, Xinchen Yan, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

Xiaoquan Wen, Roger Pique-Regi, and Francesca Luca. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization. *PLoS genetics*, 13(3):e1006646, 2017. https://doi.org/10.1371/journal.pgen.1006646.

John S Witte. Genome-wide association studies and beyond. *Annual review of public health*, 31:9–20, 2010.

Fred A Wright, Patrick F Sullivan, Andrew I Brooks, Fei Zou, Wei Sun, Kai Xia, Vered Madar, Rick Jansen, Wonil Chung, Yi-Hui Zhou, et al. Heritability and genomics of gene expression in peripheral blood. *Nature genetics*, 46(5):430–437, 2014.

Haoran Xue, Xiaotong Shen, and Wei Pan. Constrained maximum likelihood-based Mendelian randomization robust to both correlated and uncorrelated pleiotropic effects. *The American Journal of Human Genetics*, 108(7):1251–1269, 2021. https://doi.org/10.1016/j.ajhg.2021.05.014.

Xiaoqiang Yan, Shizhe Hu, Yiqiao Mao, Yangdong Ye, and Hui Yu. Deep multi-view learning methods: A review. *Neurocomputing*, 448:106–129, 2021.

Fan Yang, Jiebiao Wang, Brandon L Pierce, Lin S Chen, François Aguet, Kristin G Ardlie, Beryl B Cummings, Ellen T Gelfand, Gad Getz, Kane Hadley, et al. Identifying cis-mediators for trans-eQTLs across many human tissues using genomic mediation analysis. *Genome research*, 27(11):1859–1871, 2017. https://doi.org/10.1101/gr.216754.116.

Fan Yang, Kevin J Gleason, Jiebiao Wang, Jubao Duan, Xin He, Brandon L Pierce, and Lin S Chen. CCmed: cross-condition mediation analysis for identifying replicable trans-associations mediated by cis-gene expression. *Bioinformatics*, 37(17):2513–2520, 2021.

Yi Yang, Mingwei Dai, Jian Huang, Xinyi Lin, Can Yang, Min Chen, and Jin Liu. LPG: a four-group probabilistic approach to leveraging pleiotropy in genome-wide association studies. *BMC genomics*, 19(1):1–11, 2018.

Chen Yao, George Chen, Ci Song, Joshua Keefe, Michael Mendelson, Tianxiao Huan, Benjamin B Sun, Annika Laser, Joseph C Maranville, Hongsheng Wu, et al. Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nature communications*, 9(1):1–11, 2018.

Seyhan Yazar, Jose Alquicira-Hernandez, Kristof Wing, Anne Senabouth, M Grace Gordon, Stacey Andersen, Qinyi Lu, Antonia Rowson, Thomas RP Taylor, Linda Clarke, et al. Single-cell eQTL mapping identifies cell type–specific genetic control of autoimmune disease. *Science*, 376(6589):eabf3041, 2022.

Yihao Lu and Ke Xu and Bowei Kang and Brandon L Pierce and Fan Yang and Lin Chen. An integrative multi-context mendelian randomization method for identifying risk genes across human tissues. *medRxiv*, 2024. doi:10.1101/2024.03.04.24303731. URL https://www.medrxiv.org/content/early/2024/03/07/2024.03.04.24303731.

Jun Yin and Shiliang Sun. Multiview uncorrelated locality preserving projection. *IEEE transactions on neural networks and learning systems*, 31(9):3442–3455, 2019. https://doi.org/10.1109/TNNLS.2019.2944664.

Adam MH Young, Natsuhiko Kumasaka, Fiona Calvert, Timothy R Hammond, Andrew Knights, Nikolaos Panousis, Jun Sung Park, Jeremy Schwartzentruber, Jimmy Liu, Kousik Kundu, et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nature genetics*, 53(6):861–868, 2021.

Zhongshang Yuan, Huanhuan Zhu, Ping Zeng, Sheng Yang, Shiquan Sun, Can Yang, Jin Liu, and Xiang Zhou. Testing and controlling for horizontal pleiotropy with probabilistic

Mendelian randomization in transcriptome-wide association studies. *Nature communications*, 11(1):3861, 2020.

Wenjuan Zhang, Brandon Huckaby, John Talburt, Sherman Weissman, and Mary Qu Yang. cnnImpute: missing value recovery for single cell RNA sequencing data. *Scientific Reports*, 14(1):3946, 2024.

Qingyuan Zhao, Jingshu Wang, Gibran Hemani, Jack Bowden, Dylan S Small, et al. Statistical inference in two-sample summary-data Mendelian randomization using robust adjusted profile score. *Annals of Statistics*, 48(3):1742–1769, 2020. https://doi.org/10.1214/19-AOS1866.

Shijie C Zheng, Charles E Breeze, Stephan Beck, and Andrew E Teschendorff. Identification of differentially methylated cell types in epigenome-wide association studies. *Nature methods*, 15(12):1059–1066, 2018.

Dan Zhou, Yi Jiang, Xue Zhong, Nancy J Cox, Chunyu Liu, and Eric R Gamazon. A unified framework for joint-tissue transcriptome-wide association and Mendelian randomization analysis. *Nature genetics*, 52(11):1239–1246, 2020. https://doi.org/10.1038/s41588-020-0706-2.