Causal differential expression analysis under unmeasured confounders with causarray

Jin-Hong Du[1,2], **Maya Shen**[1], **Hansruedi Mathys**[3], **and Kathryn Roeder**[1,4,✉]

[1]Department of Statistics and Data Science, Carnegie Mellon University
[2]Machine Learning Department, Carnegie Mellon University
[3]Department of Neurobiology, University of Pittsburgh
[4]Computational Biology Department, Carnegie Mellon University

Advances in single-cell sequencing and CRISPR technologies have enabled detailed case-control comparisons and experimental perturbations at single-cell resolution. However, uncovering causal relationships in observational genomic data remains challenging due to selection bias and inadequate adjustment for unmeasured confounders, particularly in heterogeneous datasets. To address these challenges, we introduce causarray, a doubly robust causal inference framework for analyzing array-based genomic data at both bulk-cell and single-cell levels. causarray integrates a generalized confounder adjustment method to account for unmeasured confounders and employs semiparametric inference with flexible machine learning techniques to ensure robust statistical estimation of treatment effects. Benchmarking results show that causarray robustly separates treatment effects from confounders while preserving biological signals across diverse settings. We also apply causarray to two single-cell genomic studies: (1) an in vivo Perturb-seq study of autism risk genes in developing mouse brains and (2) a case-control study of Alzheimer's disease using three human brain transcriptomic datasets. In these applications, causarray identifies clustered causal effects of multiple autism risk genes and consistent causally affected genes across Alzheimer's disease datasets, uncovering biologically relevant pathways directly linked to neuronal development and synaptic functions that are critical for understanding disease pathology.

Keywords: causal inference, confounder adjustment, counterfactual, double robustness, differential expression analysis

Introduction

The advent of genomic research has transformed our understanding of biological processes and disease mechanisms. Advances in single-cell RNA sequencing (scRNA-seq) have driven this rapid progress, offering unprecedented insights into gene expression patterns at the cellular level (1). The high resolution provided by scRNA-seq data is essential to elucidate cellular heterogeneity and its implications for health and disease (2–4). However, fully harnessing the potential of these data requires robust analytical frameworks capable of moving beyond association to unravel complex causal relationships at single-cell resolution (5–7). The fundamental difference between association and causation is that association assesses correlations between treatments and outcomes, whereas causal inference aims to quantify the effect of a treatment on an outcome. A popular framework for causal inference is the *potential outcomes* framework, which estimates what would have happened if a different treatment had been assigned, the *counterfactual* (7, 8). To understand the inner workings and mechanisms of biological processes and diseases for the purpose of treatments, precision medicine, genomic medicine and more, causal inferences will be required (9, 10).

One of the primary challenges in leveraging scRNA-seq data for causal inference is its inherent hierarchical organization and heterogeneity (6, 7, 11). Cells derived from the same individual are not independent observations; they share biological factors, such as correlated variability and technical factors, including batch effects introduced during storage and sequencing. These dependencies violate the assumption of independent and identically distributed (i.i.d.) samples, complicating statistical analyses and rendering traditional methods inadequate for handling heterogeneous data with unwanted variations (12, 13). Furthermore, most genomic studies are observational in nature. Unlike randomized controlled trials, observational studies lack complete knowledge of the disease or treatment assignment mechanism, leading to potential biases in counterfactual estimation.

CRISPR perturbation experiments, a more recent but rapidly expanding area, offer a new set of challenging analysis scenarios (14–16). For this experimental setting, perturbed cells are contrasted with cells that receive a non-targeting perturbation. While there is some randomness in the treatment assignment, it is not entirely random: continuous unmeasured confounders such as variability in cell size or differential drug exposure can result in biased causal estimates. Additionally, when such experiments are performed in vivo, the possibility of confounding increases (17), further justifying the need for robust causal inference analysis.

Existing methods for causal inference, such as CoCoA-diff (6) and CINEMA-OT (11), rely on simple matching techniques that assume the causal structure is transferable between treatment and control groups. However, this assumption breaks down when covariate distributions differ significantly across groups, leading to biased estimates. Moreover, even after controlling for observed confounders, unmeasured confounders can undermine the validity of causal conclusions (18, 19). Other methods like surrogate variable analysis (SVA) (20) and RUV (13) aim to address confounding and unwanted variation via linear models that assume additive relationships between covariates and outcomes. While effective for certain bulk RNA-seq datasets, these approaches often fail to capture the sparsity, zero inflation, and overdispersion inherent in single-cell genomic data (18, 21). Tackling these challenges requires integrating robust confounder adjustment with flexible modeling techniques to ensure valid causal inference in complex genomic data.

In response to these challenges, we introduce a new framework for applying causal inference in genomic studies. Our approach leverages generalized factor models tailored to count data to account for unmeasured confounders, ensuring robust adjustment for unmeasured confounders while preserving biological signals. It further relies on the potential outcomes framework and employs a doubly robust estimation procedure, which combines outcome and propensity score models to ensure reliable statistical inference even if one model is misspecified (22, 23). This framework effectively addresses biases introduced by both observed and unobserved confounders, making it particularly well-suited for analyzing complex genomic data at both bulk and single-cell levels (Fig. 1a). By integrating advanced statistical and machine learning techniques with a causal inference framework, our method enables a range of downstream analyses, including accurate estimation of counterfactual distributions, causal gene detection, and conditional treatment effect analysis. This approach not only improves the interpretability and precision of genomic analyses but also uncovers critical insights into gene expression dynamics under disease or perturbation conditions, advancing our understanding of underlying biological mechanisms.

We demonstrate the effectiveness of causarray through benchmarking on several simulated datasets, comparing its performance with existing single-cell-level perturbation analysis methods and pseudo-bulk-level differential expression (DE) analysis methods. Next, we apply causarray to two single-cell genomic studies: a Perturb-seq study investigating autism spectrum disorder/neurodevelopmental disorder (ASD/ND) genes in developing mouse brains and a case-control study of Alzheimer's disease using human brain transcriptomic datasets. For the Alzheimer's disease analysis, we validate our findings across three independent datasets, showcasing the robustness and reproducibility of causarray in identifying causally affected genes and uncovering biologically meaningful pathways. These applications highlight the potential of causarray to advance our understanding of complex disease mechanisms through rigorous causal inference.

Results

Doubly-robust counterfactual imputation and inference

Our objective is to determine whether a gene is causally affected by a "treatment" variable after controlling for other technical and biological covariates, which may affect the treatment and outcome variables. Here, we use the term treatment generally; in the narrow sense, it can mean genetic and/or chemical perturbations (17, 24), such as CRISPR-CAS9, and, more broadly, it can mean the phenotype of a disease (6). We acknowledge that while many differentially expressed genes can be considered a result of disease status, for most late-onset disorders, a smaller fraction of genes could have initiated disease phenotypes. Our method aims to determine the direct effects of treatments on modulated gene expression outcomes.

In observational data, the response variable can be confounded by measured and unmeasured biological and technical covariates, making it difficult to separate the treatment effect from other unknown covariates. As a consequence, it is challenging to draw causal inferences; even tests of association may lead to an excess of false discoveries and/or low power. Fortunately, the potential outcomes framework (22, 23) formulates general causal problems in a way that allows for the treatment effect to be separated from the effects of other variables. However, even this framework is challenged by unmeasured covariates. Before introducing our method for estimating unmeasured confounders, we first outline the general potential outcomes framework.

Consider a study in which Y is the response variable and A is the binary treatment variable for an observation. In the potential outcomes framework, $Y(a)$ is the outcome that we would have observed if we set the treatment to $A = a$. Naturally, we can only observe one of the two potential outcomes for each observation, so

$$Y = \mathbb{1}\{A=1\}Y(1) + \mathbb{1}\{A=0\}Y(0),$$

In the context of a case-control study of a disease, this would answer the question: What is the expected difference in gene expression if an individual had the disease (case, $A = 1$) versus if they did not (control, $A = 0$)?

Doubly robust methods provide a powerful tool for estimating potential outcomes in observational studies where randomization is not possible (22, 23). Specifically, we estimate two key quantities: (1) $\mu_a(X)$, the mean response of the outcome variable conditional on treatment $A = a$ and covariates $X = x$, and (2) $\pi_a(X)$, the propensity score, which is defined as the probability of receiving treatment $A = a$ given covariates X, i.e., $\pi_a(X) = \mathbb{P}(A = a \mid X)$. Using these estimates, we compute potential outcomes as

$$\widehat{Y}(a) = \frac{\mathbb{1}\{A=a\}}{\widehat{\pi}_a(X)}(Y - \widehat{\mu}_a(X)) + \widehat{\mu}_a(X).$$

The doubly robust estimator's name comes from the fact that it provides a consistent estimate as long as *either* the outcome model, $\mu_a(X)$, or the propensity score model, $\pi_a(X)$, is correctly specified. Given this estimate, we can easily perform downstream inference tasks such as computing log fold change (LFC) (Methods), and testing for causal effects on gene expressions (Fig. 1a). An advantage of this approach is that counterfactual imputation denoises/balances gene expression under two different conditions. Additionally, having access to estimated potential outcomes facilitates downstream analyses such as estimating causal effects conditional on measured confounders like age.

A key step in these types of analyses is estimating unmeasured confounders. To adjust for confounding, factor models were popularized in surrogate variable analysis literature and have since been widely adopted in bulk gene expression studies (20). Recently, we extended this approach to single-cell RNA-seq data using generalized linear models that better

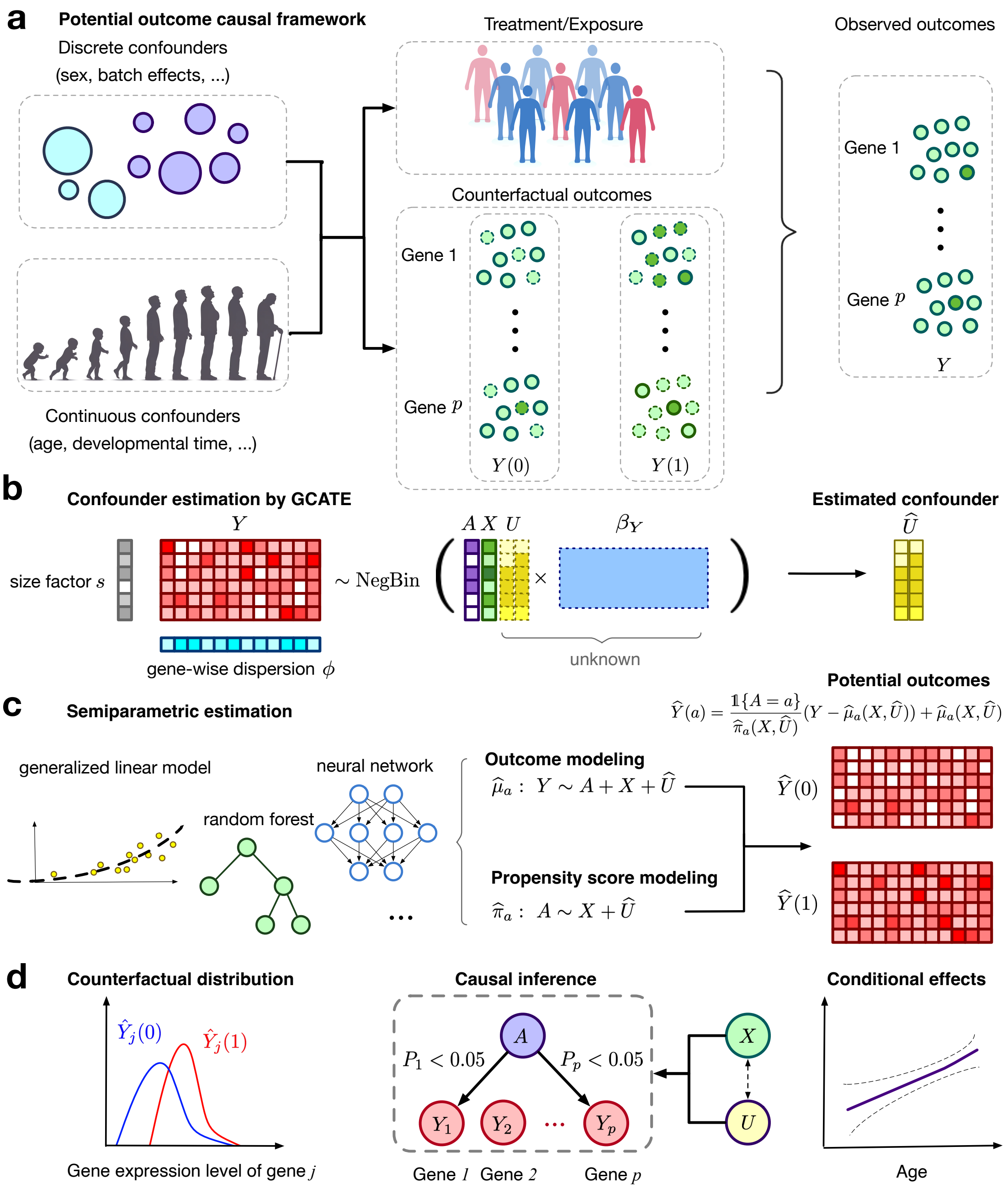

Fig. 1. Overview of the proposed causarray method. a, Illustration of the data generation process for pseudo-bulk and single-cell data. **b,** The gene expression matrix, Y, is linked to the treatment, A, measured covariates, X, and confounding variables, U, via a GLM model. The cell-wise size factor, s, and gene-wise dispersion parameter, ϕ, are estimated from the data, and the unmeasured confounder U is estimated by $\widehat{U}$ through the augmented GCATE method. **c,** Generalized linear models and flexible machine learning methods including random forest and neural network can be applied for outcome modeling ($\mathbb{E}[Y \mid A=a, X, \widehat{U}] = \widehat{\mu}_a(X,\widehat{U})$) and propensity modeling ($\mathbb{P}(A=a \mid X, U) = \widehat{\pi}_a(X,\widehat{U})$) The estimated outcome and propensity score functions give rise to the estimated potential outcomes for each cell and each gene. **d,** Downstream analysis includes contrasting the estimated counterfactual distributions, performing causal inference, and estimating the conditional average treatment effects.

accommodate pseudobulk and single-cell outcome variables (18). Using this generalized factor analysis approach, we estimate unmeasured confounders U alongside potential outcomes (Fig. 1b-c), enabling direct estimation of downstream quantities such as LFC (Fig. 1d).

Simulation study demonstrates the advantages of causarray

We evaluate the performance of causarray in two simulated settings (Appendix S3). In the first setting, we generate simulated pseudo-bulk data, while in the second, we generate simulated single-cell data using the Splatter simulator (25), which explicitly models the hierarchical Gamma-Poisson processes underlying scRNA-seq data and captures multi-faceted variability. Each dataset consists of 100-300 cells, approximately 2,000 genes, 1-2 covariates, and 4 unmeasured confounders.

To benchmark causarray, we compare it with several existing methods designed for differential expression (DE) testing, both with and without confounder adjustment (Fig. 2a). For methods that do not account for unmeasured confounders,

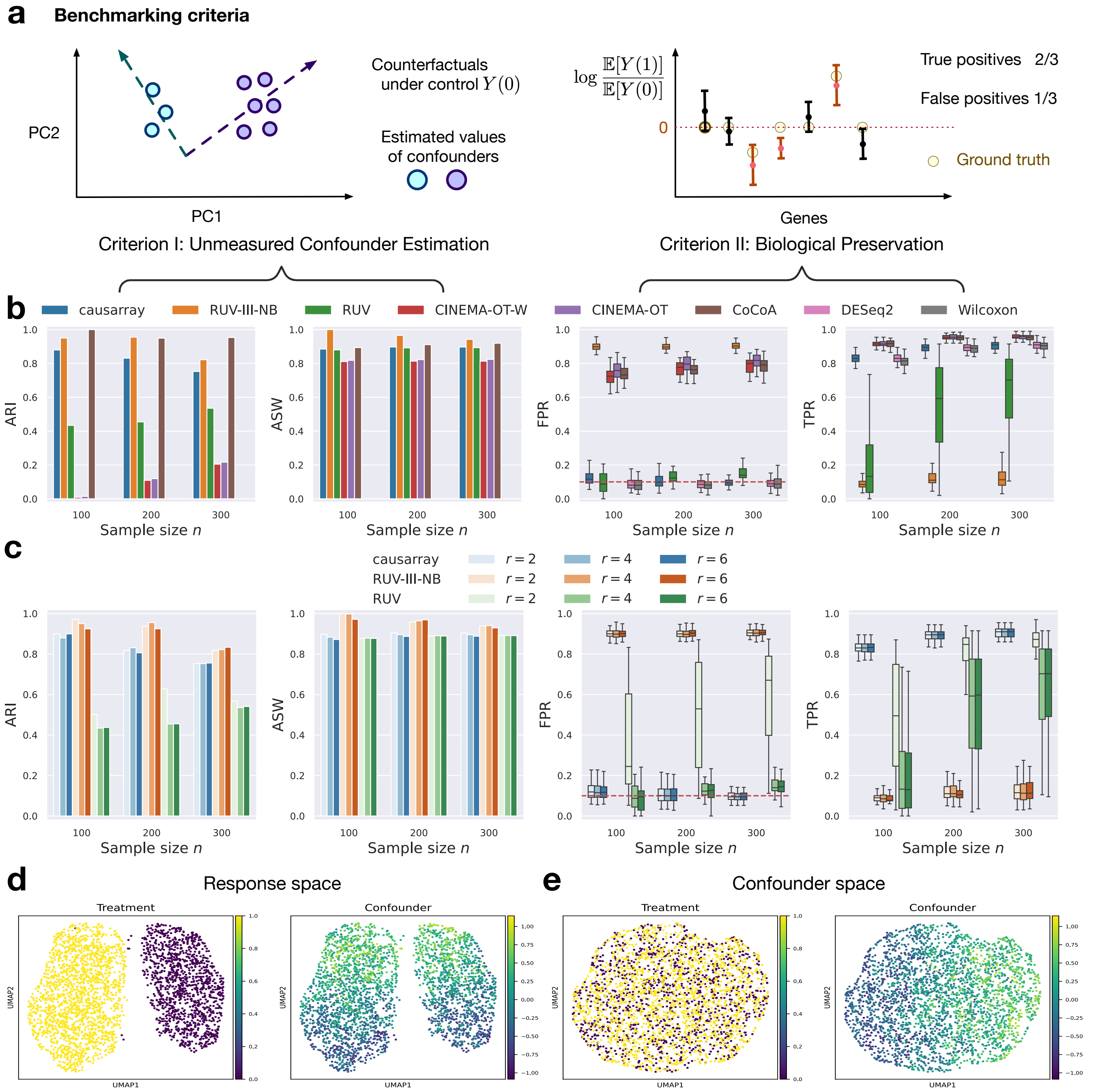

Fig. 2. Benchmarking of causarray against other methods for single-cell differential expression testing on synthetic expression data with unmeasured confounders. **a,** The analysis pipeline produces a confounder adjustment and a statistic for DE testing. We illustrate two types of criteria used for benchmarking confounder adjustment and DE methods in simulation for bulk simulations (**b**-**e**) and single-cell simulations (Fig. S1). **b,** Performance comparison of causarray and other methods with a well-specified number of latent factors ($r = 4$). Bar plots show median ARI and ASW scores for confounder estimation, while box plots display FPR and TPR for biological signal preservation. The top and bottom hinges represent the top and bottom quartiles, and whiskers extend from the hinge to the largest or smallest value no further than 1.5 times the interquartile range from the hinge. The center indicates the median. **c,** Robustness analysis of causarray, RUV-III-NB, and RUV under varying numbers of latent factors ($r = 2, 4, 6$). Bar plots show ARI and ASW scores for confounder estimation, while box plots display FPR and TPR for DE testing. **d-e,** causarray disentangles the treatment effects and unmeasured confounding effects in the response and confounder spaces. UMAP projection of (**d**) expression data Y colored by the values of treatment A (purple for control $A = 0$ and yellow for treated $A = 1$) and unmeasured continuous confounder U; and (**e**) estimated potential outcome under control $Y(0)$ colored by the values of treatment A and continuous confounder U.

we include the Wilcoxon rank-sum test and DESeq2 (26). In the presence of measured covariates, both regress the gene expression counts with respect to the covariates using the Poisson or negative binomial generalized linear model, respectively. The input to the Wilcoxon rank sum test is the deviance residuals. For confounder-adjusted methods, we consider CoCoA-diff (6), CINEMA-OT (11), CINEMA-OT-W (11), RUV (12), and RUV-III-NB (13), where recommended DE test methods are subsequently applied with estimated confounders. A short summary of each of these benchmarking

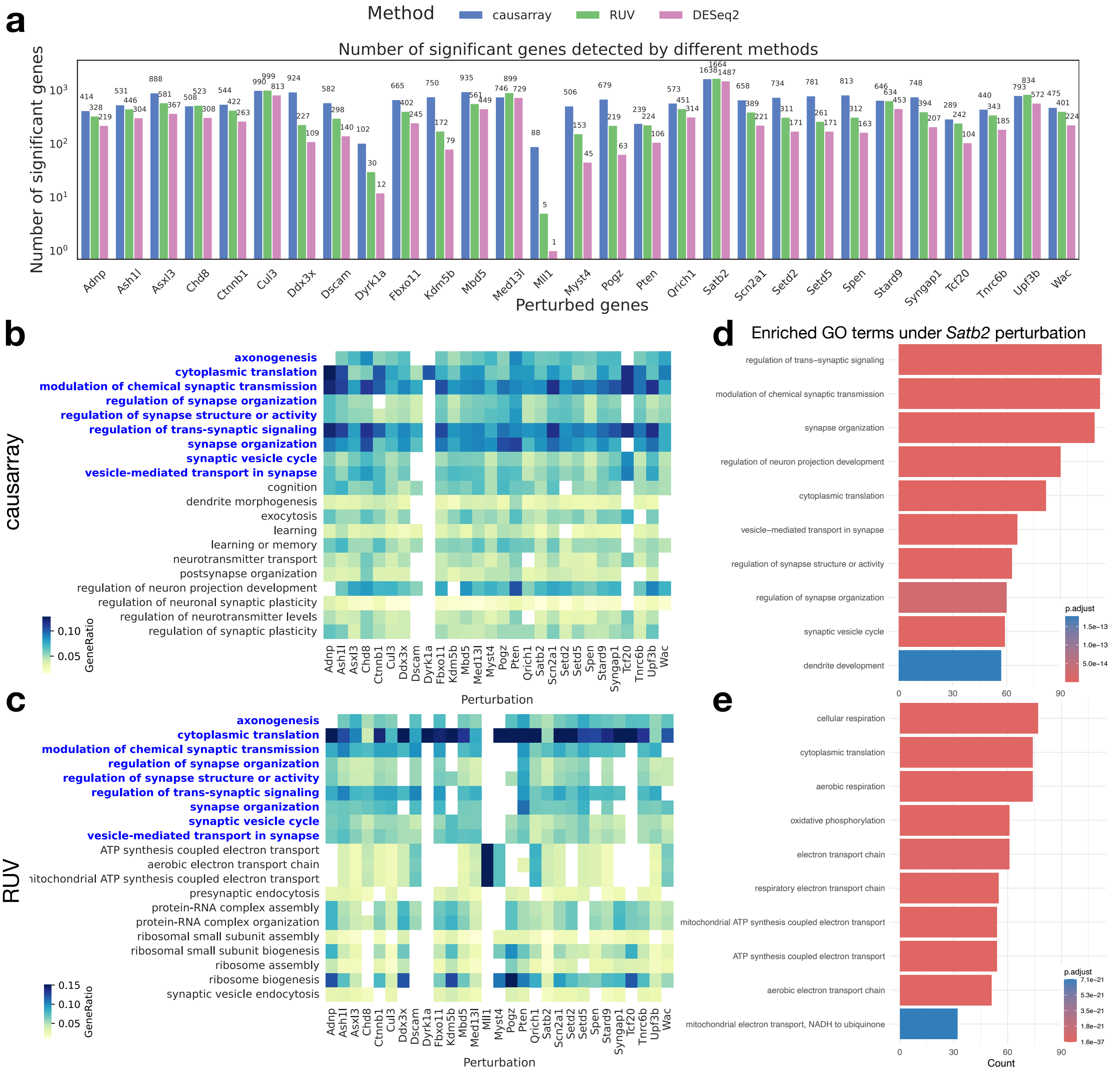

Fig. 3. Statistical test results of the effects of CRISPR perturbation on gene expression in excitatory neuron data. a, Number of significant genes detected under all perturbations using three different methods. The detection threshold for significant genes is FDR< 0.1 for all methods. **b-c,** Heatmaps of GO terms enriched (adjusted P value < 0.05, $q < 0.2$) in discoveries from causarray and RUV, respectively, where the common GO terms are highlighted in blue. Only the top 20 GO terms that have the most occurrences in all perturbations are displayed. **d-e,** Barplots of GO terms enriched in discoveries under *Satb2* perturbation from causarray and RUV, respectively.

comparison methods can be found in Methods.

To assess the performance of unmeasured confounder adjustment procedures, we use two metrics: adjusted Rand index (ARI) and average silhouette width (ASW). More specifically, we use ARI to quantify the alignment between estimated and true unmeasured confounders and ASW to evaluate cell type separation in the control response space. A higher ARI value indicates better coherence and a higher ASW value reflects better preservation of biological signals after removing confounding effects. Additionally, to assess the performance of DE testing, we use two metrics: false positive rate (FPR) and true positive rate (TPR) (Methods).

We first evaluate how sample size and confounding levels influence the performance of DE testing across methods. Among all tested approaches, only causarray, RUV, Wilcoxon, and DESeq2 effectively control FPR across all settings (Fig. 2b and Fig. S1ab). causarray maintains FPR close to the nominal level of 0.1 across all sample sizes and confounding levels, while RUV-III-NB, CINEMA-OT-W, CINEMA-OT, and CoCoA-diff exhibit inflated FPRs exceeding 0.5 in most cases. Notably, causarray achieves the highest TPRs across all scenarios, with values ranging from approximately 0.8 to 0.9 depending on sample sizes and confounding levels (Fig. 2b and Fig. S1ab). This is significantly higher than RUV-III-NB and

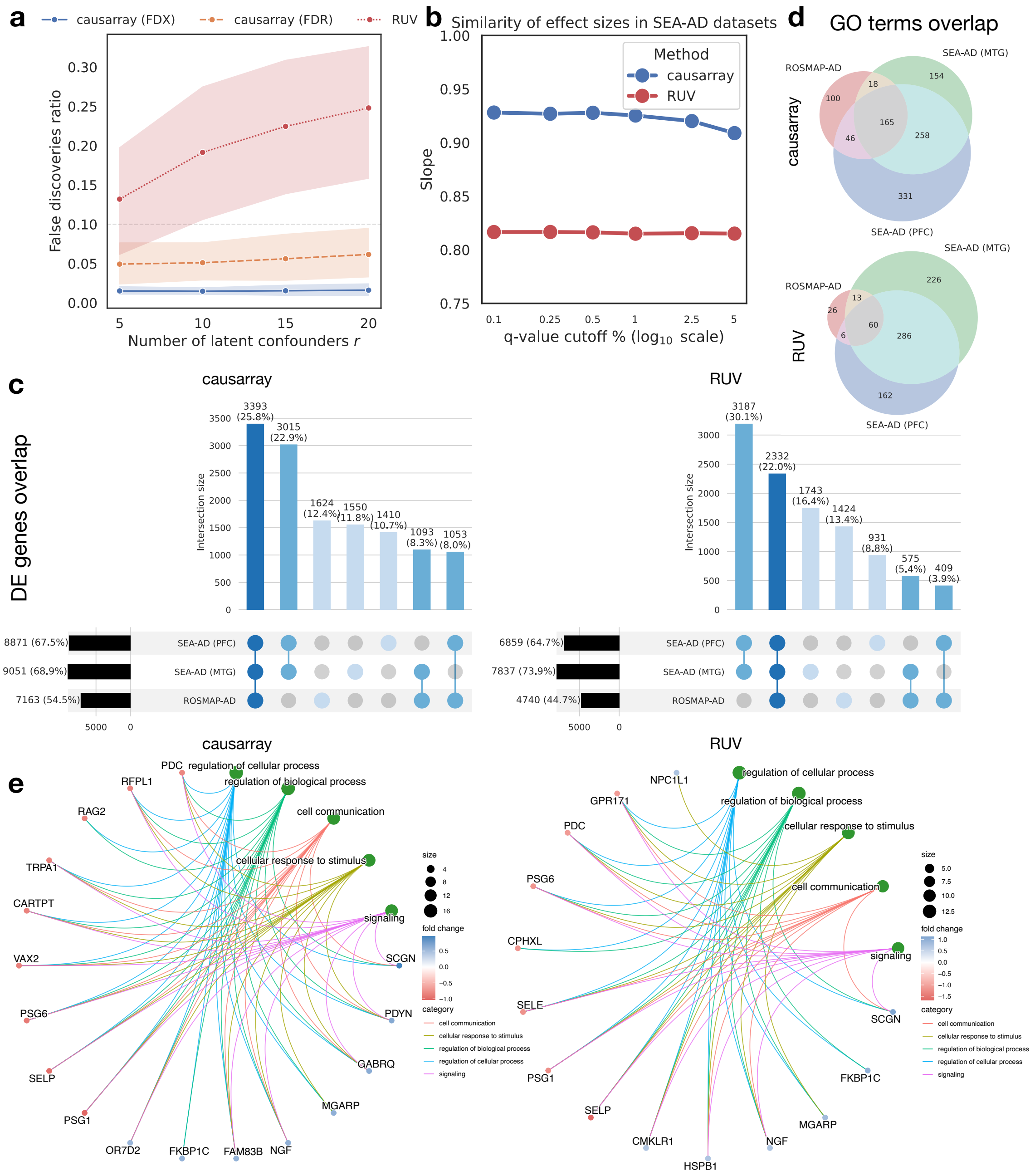

Fig. 4. Comparison of DE genes discovered by causarray and RUV on excitatory neurons for Alzheimer's disease. a, The ratio of false discoveries to all 15586 genes of DE test results with permuted disease labels on the ROSMAP-AD dataset. Three methods, causarray with FDX control, causarray with FDR control, and RUV with FDR control, are compared. **b,** The similarity of estimated effect sizes on SEA-AD MTG and PFC datasets. The slope is estimated from linear regression of effect sizes on the PFC dataset against those on the MTG dataset. **c,** DE genes by causarray and RUV over 15586 genes (adjusted P value < 0.1). **d,** Venn diagram of associated GO terms from causarray and RUV (adjusted P value < 0.05, $q < 0.2$). **e,** Considering only the top 50 positively regulated and the top 50 negatively regulated DE genes from causarray and RUV, we map them to the top 5 biological processes (the green nodes).

CoCoA-diff, which achieve TPRs below 0.5 in most settings, particularly for smaller sample sizes or higher confounding levels. These results highlight causarray's ability to balance sensitivity and specificity effectively.

In terms of unmeasured confounder adjustment, causarray, RUV-III-NB, and CoCoA-diff achieve both ARI and ASW scores consistently above 0.7 across all sample sizes in both bulk and single-cell data (Fig. 2b, Fig. S1ab), outperforming RUV, CINEMA-OT-W, CINEMA-OT, which show ARI scores below 0.5 in most cases. Furthermore, causarray effectively disentangles treatment effects from unmeasured confounding effects. In the response space (Fig. 2d), treatment groups are distinctly separated with minimal overlap, while variations within groups reflect unmeasured confounders. In the confounder space (Fig. 2e), causarray produces a uniform mixing of treatment groups while accurately reconstructing continuous confounder values.

Finally, we assess the robustness of causarray, RUV-III-NB, and RUV under varying numbers of latent factors (Fig. 2c and Fig. S1c). Among these methods, only causarray consis-

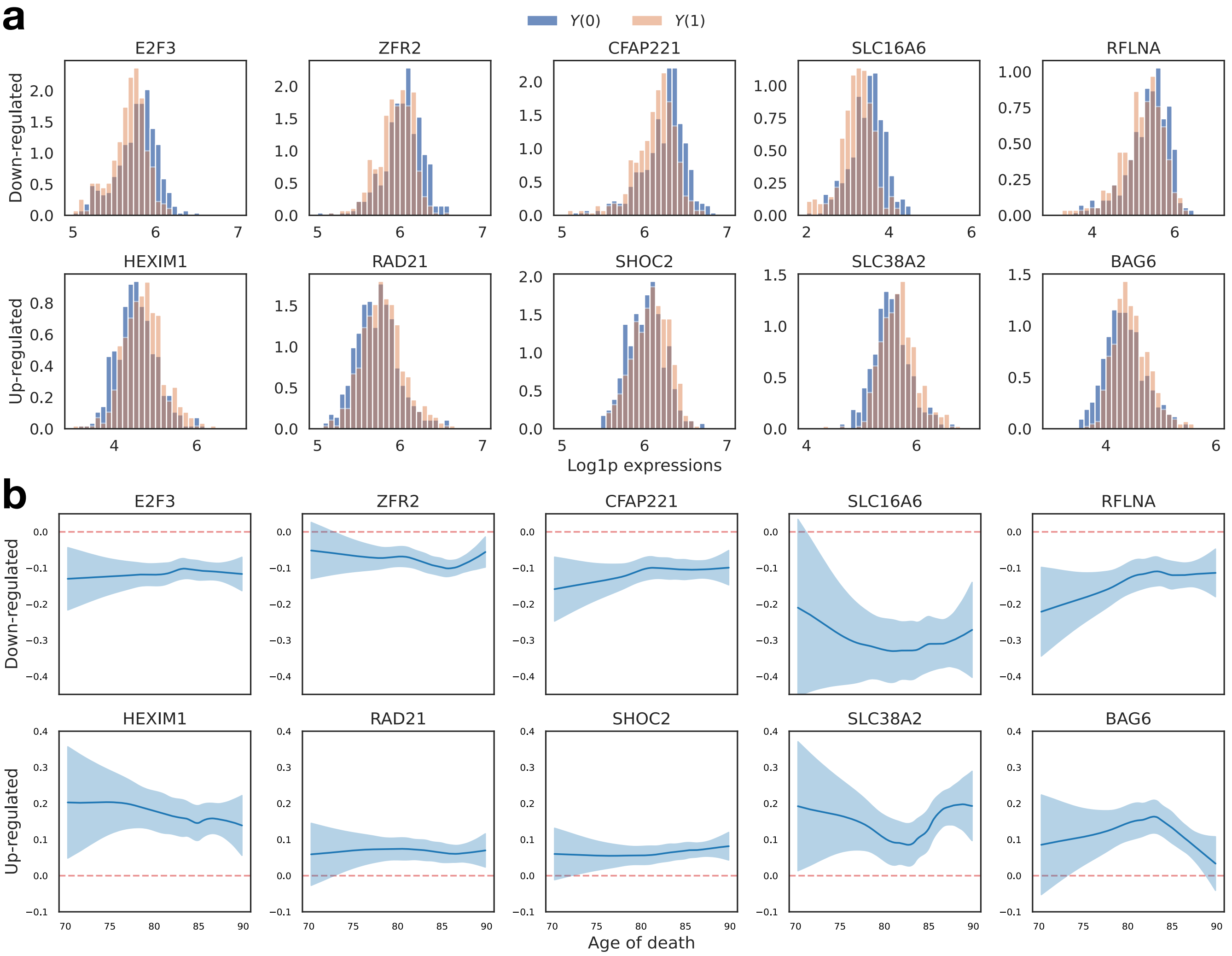

Fig. 5. Results of DE analysis of 10 selected genes by causarray. The top 5 up-regulated and top 5 down-regulated genes in estimated LFCs (adjusted P value < 0.05) are visualized. **a,** Estimated counterfactual distributions. The values are shown in the log scale after adding one pseudo-count. **b,** Estimated log-fold change of treatment effects, conditional on age for selected genes. The center lines represent the mean of the locally estimated scatter plot smoothing (LOESS) regression, and the shaded area represents a 95% confidence interval at each value of age.

tently controls FPR at nominal levels of 0.1 regardless of the number of factors or sample size. In contrast, RUV-III-NB exhibits inflated median FPRs exceeding 0.2 when more factors are included (e.g., $r = 6$). While RUV-III-NB performs well in terms of ARI (above 0.8) and ASW (above 0.7), its DE testing performance is inferior to RUV due to poor FPR control under certain conditions. Based on these findings, we proceed with causarray and RUV for real data analysis.

causarray applied to an in vivo Perturb-seq study reveals causal effects of ASD/ND genes

An integrative analysis of multiple single perturbations. Autism spectrum disorders and neurodevelopmental delay (ASD/ND) represent a complex group of conditions that have been extensively studied using genetic approaches. To investigate the underlying mechanisms of these disorders, researchers have employed scalable genetic screening with CRISPR-Cas9 technology (17). Frameshift mutations were introduced in the developing mouse neocortex in utero, followed by single-cell transcriptomic analysis of perturbed cells from the early postnatal brain (17). These in vivo single-cell Perturb-seq data allow for the investigation of causal effects of a panel of ASD/ND risk genes. We analyze the transcriptome of cortical projection neurons (excitatory neurons) perturbed by one risk gene or a non-targeting control perturbation, which serves as a negative control.

Unmeasured confounders, such as batch effects and unwanted variation, are likely present in this dataset due to the batch design being highly correlated with perturbation conditions (Fig. S2ab). Additionally, the heterogeneity of single cells assessed in vivo introduces further complexity. These confounding factors may reduce statistical power for gene-level differential expression (DE) tests, as noted in the original study (17), which instead focused on gene module-level effects. To address this limitation, we apply causarray to incorporate unmeasured confounder adjustment and conduct a more granular analysis at the single-gene level. This approach enables us to uncover nuanced genetic interactions and causal effects that may provide deeper insights into the etiology of ASD/ND.

Functional analysis. Gene module-level analyses have been shown to provide greater statistical power for detecting biologically meaningful perturbation effects when fewer cells are available (17). The original study adopted this approach but relied on a linear model rather than a negative binomial model, potentially limiting its ability to detect broader signals at the individual gene level. Here, we compare causarray with RUV and DESeq2 (without confounder adjustment) to identify significant genes and enriched gene ontology (GO) terms associated with various perturbations.

In terms of significant gene detection, causarray identifies a comparable number of significant genes to RUV across most perturbations, while DESeq2 consistently detects fewer significant genes (Fig. 3a). The variation in significant detections across different perturbed genes suggests distinct biological impacts of each knockout. Functional analysis focuses on enriched GO terms on the DE genes under each perturbation condition where discrepancies arise between causarray and other methods. Genes identified by causarray are enriched for biologically relevant GO terms with clear clustering patterns (Fig. 3b-c, Fig. S2c). In contrast, RUV shows less distinct clustering and enrichment patterns.

Notably, while RUV identifies GO terms related to ribosome processes previously implicated in ASD studies (27), these findings remain controversial. Some argue that dysregulation in translation processes and ribosomal proteins may reflect secondary changes triggered by expression alterations in synaptic genes rather than direct causal effects (28). In contrast, GO terms identified by causarray align more closely with the expected causal effects of ASD/ND gene perturbations (29, 30).

To further validate these findings, we examine the perturbation condition for *Satb2*, which yields the largest number of significant genes identified by both methods (adjusted P value < 0.1). *Satb2* is known to play critical roles in neuronal development, synaptic function, and cognitive processes (31, 32). Using causarray, we detect enrichment for GO terms directly related to neuronal function and development, such as "regulation of neuron projection development," "regulation of synapse structure or activity," and "synapse organization" (Fig. 3d). These findings are consistent with *Satb2*'s established roles in neuronal development and synaptic plasticity (33, 34). On the other hand, RUV identifies enrichment for terms related to mitochondrial function and energy metabolism, such as "mitochondrial electron transport," "cellular respiration," and "ATP synthesis" (Fig. 3e). While these processes are important for general cellular function, they are less directly relevant to *Satb2*'s primary biological roles.

Overall, this analysis demonstrates that causarray provides greater specificity in detecting biologically meaningful causal effects of gene perturbations. Its ability to disentangle confounding influences while preserving relevant biological signals highlights its effectiveness in analyzing complex genomic datasets.

causarray reveals causally affected genes of Alzheimer's disease in a case-control study

An integrative analysis of excitatory neurons. We analyze three Alzheimer's disease (AD) single-nucleus RNA sequencing (snRNA-seq) datasets: a transcriptomic atlas from the Religious Orders Study and Memory and Aging Project (ROSMAP) (35) and two datasets from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium (36), which include samples from the middle temporal gyrus (MTG) and prefrontal cortex (PFC). Our objective is to compare the performance of causarray and RUV in pseudo-bulk DE tests of AD in excitatory neurons.

To evaluate the validity, we perform a permutation experiment on the ROSMAP-AD dataset by permuting phenotypic labels. Ideally, no significant discoveries should be made under this null scenario. However, RUV produces a large number of false discoveries, with its performance deteriorating as the number of latent factors increases. In contrast, causarray effectively controls the false discovery rate (FDR), producing minimal false positives (Fig. 4a). Additionally, we assess coherence across datasets by examining effect sizes in SEA-AD (MTG) and SEA-AD (PFC). Effect sizes estimated by causarray exhibit higher consistency across varying q-value cutoffs compared to RUV (Fig. 4b, Fig. S3b). When inspecting DE genes across all three AD datasets, causarray identifies more consistent discoveries than RUV (Fig. 4c), highlighting its robustness in detecting causally affected genes.

Functional analysis. We further compare functional enrichment results between causarray and RUV using gene ontology (GO) terms associated with DE genes. Across the three datasets, causarray identifies 165 common GO terms, significantly more than the 60 identified by RUV (Fig. 4d). Both methods detect GO terms relevant to neuronal development and synaptic functions, which are critical for understanding AD pathology. However, causarray shows distinct enrichment in categories such as "positive regulation of cell development" and "negative regulation of cell cycle', reflecting its increased sensitivity to synaptic and neurotransmission-related processes. In contrast, RUV's results exhibit more dataset-specific enrichments, such as biosynthetic processes in SEA-AD (PFC), apoptotic processes in SEA-AD (MTG), and catabolic processes in ROSMAP-AD (Fig. S3c). These findings suggest that causarray captures more generalizable biological signals across datasets.

Both methods identify overlapping top functional categories related to key biological processes associated with AD pathology (Fig. S3e). However, causarray associates a larger number of genes with these categories, identifying 3393 DE genes compared to 3187 for RUV (Fig. 4c). Additionally, causarray reveals 165 common GO terms across the three datasets, significantly more than the 60 identified by RUV (Fig. 4d). The visualization of the discovered networks, as defined as the top 5 GO terms and associated genes included in the top 100 DE gene discoveries, further highlights the enhanced sensitivity and comprehensiveness of causarray. Specifically, the causarray network contains 17 gene nodes and 81 edges, compared to 14 gene nodes and 57 edges in the RUV network (Fig. 4e).

This greater interconnectedness in the larger causarray network suggests a more intricate and informative representation of underlying biological relationships, emphasizing its ability to capture broader and more relevant genetic factors associated with AD pathology.

Counterfactual analysis. The counterfactual framework employed by causarray enables downstream analyses that directly utilize estimated potential outcomes. By examining counterfactual distributions for significant genes (Fig. 5a), we observe distinct shifts in expression levels between treatment ($Y(1)$) and control ($Y(0)$) groups. Downregulated genes show a shift toward lower expression levels under disease conditions, while upregulated genes exhibit increased expression. Conditional average treatment effects (CATEs) reveal age-dependent trends for these genes (Fig. 5b). For example, upregulated genes such as *SLC16A6* and *RFLNA* show stronger effects at extreme ends of the age distribution, while others like *SLC38A2* and *BAG6* display nuanced changes across the aging spectrum.

These findings align with prior studies highlighting the roles of specific genes in aging-related processes. For instance, *ZFR2*, *RFLNA*, *BAG6*, and *RAD21* have been implicated in chromatin remodeling, synaptic plasticity, and cellular stress responses critical for aging and neurodegeneration (37–40). While nonparametric fitted curves may exaggerate age effects due to uncertainty bands, significant trends observed for key genes underscore their potential relevance in AD pathology. Overall, these results demonstrate that causarray provides nuanced insights into age-dependent gene regulation mechanisms while maintaining robust control over confounding influences.

Discussion

The rapid growth of high-throughput single-cell technologies has created an urgent need for robust causal inference frameworks capable of disentangling treatment effects from confounding influences. Existing methods, such as CINEMA-OT (11), have advanced the field by separating confounder and treatment signals and providing per-cell treatment-effect estimates. However, these methods rely on the assumption of no unmeasured confounders, which is often violated in observational studies and in vivo experiments. Additionally, many confounder adjustment methods, such as RUV (12), depend on linear model assumptions that do not directly model count data or provide robust differential expression testing at the gene level. Addressing these limitations, causarray introduces a doubly robust framework that integrates generalized confounder adjustment with semiparametric inference to enable reliable and interpretable causal analysis.

causarray directly models count data using generalized linear models for unmeasured confounder estimation, overcoming a key limitation of RUV in DE analysis. Unlike CINEMA-OT (11) and CoCoA-diff (6), which rely on optimal transport or matching techniques, causarray employs a doubly robust framework that combines flexible machine learning models with semiparametric inference. This approach enhances stability and interpretability while enabling valid statistical inference of treatment effects. Benchmarking results demonstrate that causarray outperforms existing methods in disentangling treatment effects from confounding influences across diverse experimental settings, maintaining superior control over false positive rates while achieving higher true positive rates.

In an in vivo Perturb-seq study of ASD/ND genes, causarray uncovered gene-level perturbation effects that were missed by prior module-based analyses. It identified biologically relevant pathways linked to neuronal development and synaptic functions for multiple autism risk genes. Similarly, in a case-control study of Alzheimer's disease using three human brain transcriptomic datasets, causarray revealed consistent causal gene expression changes across datasets and highlighted key biological processes such as synaptic signaling and cell development. These findings underscore the ability of causarray to provide biologically meaningful insights across diverse contexts.

Despite its strengths, causarray has certain limitations. Its performance depends on the accurate estimation of unmeasured confounders, which may vary with dataset complexity and experimental design. Furthermore, while causarray provides robust DE testing, its integration with advanced spatial or trajectory analysis frameworks remains unexplored (41, 42). Future research could focus on extending causarray to incorporate prior biological knowledge or extrapolate to unseen perturbation-cell pairs, similar to emerging methods like CPA (43). Such advancements would further enhance its applicability in single-cell causal inference.

Methods

Counterfactual

Potential outcomes framework. Let $O=(A,W,Y)\in\{0,1\}\times\mathbb{R}^{d_W}\times\mathbb{R}^p$ be a tuple of random vectors, where A is the binary treatment variable (e.g., presence or absence of a disease or perturbation), W is the vector of covariates (e.g., biological or technical factors influencing both treatment and outcome), and Y is the observed outcomes, defined as $Y=AY(1)+(1-A)Y(0)$, where $Y(1)$ and $Y(0)$ are the potential outcomes under treatment and control, respectively.

The potential outcomes framework assumes that for each individual or observation, there exist two potential outcomes: one if the individual receives the treatment ($Y(1)$) and one if they do not ($Y(0)$). However, only one of these outcomes can be observed for each individual, depending on whether they were treated ($A=1$) or not ($A=0$). This framework allows us to define causal effects in terms of these unobservable potential outcomes.

To estimate causal effects, we rely on the following key assumptions:

Assumption 1 (Consistency) The observed response is consistent such that $Y(a)=Y\mid A=a$.

Assumption 2 (Positivity) The propensity score $\pi_a(W):=\mathbb{P}(A=a\mid W)\in(\epsilon,1-\epsilon)$ for some $\epsilon\in(0,1/2)$.

Assumption 3 (No unmeasured confounders) $A\perp\!\!\!\perp Y(a)\mid W$, for all $a\in\{0,1\}$.

Under these assumptions (Assumptions 1–3), the observed outcome Y is conditionally independent of the treatment A, given the covariates W. This allows us to estimate the expected potential outcome for gene j under treatment ($a=1$) or control ($a=0$) as:

$$\mathbb{E}[Y_j(a)]=\psi_j(W,a):=\mathbb{E}[\mu_j(W,a)],$$

where $\mu_j(W,a)=\mathbb{E}[Y_j\mid W,A=a]$ is a regression function that models the relationship between covariates, treatment, and outcomes.

Suppose we have a dataset $\mathcal{D}=\{O_1,\ldots,O_n\}$ consisting of i.i.d. samples from the same distribution as O. Let $\mathbb{P}_n$ denote the empirical measure over $\mathcal{D}$, defined as:

$$\mathbb{P}_n f(O)=n^{-1}\sum_{i=1}^{n}f(O_i),$$

for any measurable function f. This represents the sample average of a function evaluated on all observations in the dataset.

A naive plug-in estimator for ψ_j can then be constructed by replacing the true regression function $\mu_j(W,a)$ with its estimated counterpart $\widehat{\mu}_j(W,a)$ and using sample averages to approximate expectations. The resulting estimator is:

$$\widehat{\psi}_j^{\mathrm{PI}}=\mathbb{P}_n[\widehat{\mu}_j(W,a)]=n^{-1}\sum_{i=1}^{n}\widehat{\mu}_j(W_i,a).$$

This plug-in estimator provides an estimate of the expected potential outcome by averaging predictions from the estimated regression model over all observations in the dataset.

While Assumptions 1–3 are foundational for causal inference, violations of the no unmeasured confounders assumption (Assumption 3) are common in real-world applications (18, 19). For instance, in single-cell transcriptomic studies, technical factors such as batch effects or biological heterogeneity (e.g., cell size or cell cycle stage) may act as unmeasured confounders. These unmeasured variables can bias estimates of causal effects by introducing spurious associations between treatment and outcome. Addressing this limitation motivates the need for methods that explicitly model and adjust for unmeasured confounders.

The probabilistic modeling of confounders. To account for unmeasured confounders, we propose an improved version of the GCATE method (18), which identifies potential unmeasured confounders under generalized linear models (GLMs). This approach extends traditional confounder adjustment methods by incorporating more flexible nonlinear models that better capture the unique characteristics of genomic count data, such as zero-inflation (an excess of zero counts) and overdispersion (greater variability than expected under standard Poisson assumptions). These enhancements allow for more accurate modeling of gene expression data, addressing limitations of simpler linear models in high-dimensional genomic analyses.

For the ith observation (e.g., a single cell or sample) and the jth gene, we model the adjusted expression $\mu_{ij}=Y_{ij}/s_j$, where Y_{ij} is the observed expression level, and s_j is the size factor for the jth gene. The size factor accounts for differences in sequencing depth or library size across samples, ensuring that comparisons are not biased by technical variability. We assume that μ_{ij} follows an exponential family distribution, which is a flexible class of probability distributions commonly used in GLMs. The density of μ_{ij} is given by:

$$p(\mu_{ij}\mid\theta_{ij})=h(\mu_{ij})\exp\left(\mu_{ij}\theta_{ij}-A(\theta_{ij})\right),$$

where θ_{ij} is the natural parameter that determines the mean and variance of μ_{ij}, $h(\mu_{ij})$ is a known base measure, and $A(\theta_{ij})$ is the log-partition function, which ensures that the density integrates to 1.

In matrix form, we model the natural parameters

$$\mathbf{\Theta}=(\theta_{ij})_{1\leq i\leq n,1\leq j\leq p},$$

as a decomposition into two components:

$$\mathbf{\Theta}=\widetilde{\boldsymbol{X}}\boldsymbol{B}^\top+\boldsymbol{U}\mathbf{\Gamma}^\top.$$

Here, $\widetilde{\boldsymbol{X}}=[\boldsymbol{X},\boldsymbol{A}]\in\mathbb{R}^{n\times(d+1)}$ combines observed covariates $\boldsymbol{X}$ (e.g., biological or technical factors) with treatment indicators $\boldsymbol{A}$, where n is the number of observations, and d is the dimension of $\boldsymbol{X}$; $\boldsymbol{B}\in\mathbb{R}^{p\times(d+1)}$ represents unknown regression coefficients for the effects of covariates and treatments on gene expression; $\boldsymbol{U}\in\mathbb{R}^{n\times r}$ represents latent variables capturing unmeasured confounders, where r is the number of latent factors; and $\mathbf{\Gamma}\in\mathbb{R}^{p\times r}$ represents unknown coefficients linking unmeasured confounders to gene expression.

This decomposition assumes that gene expression levels are influenced by both observed covariates ($\widetilde{\boldsymbol{X}}$) and unmeasured confounders ($\boldsymbol{U}$). The term $\widetilde{\boldsymbol{X}}\boldsymbol{B}^\top$ captures the effects of observed covariates and treatments, while $\boldsymbol{U}\boldsymbol{\Gamma}^\top$ captures the effects of unmeasured confounders.

To estimate these unknown quantities ($\boldsymbol{B}$, $\boldsymbol{U}$, $\boldsymbol{\Gamma}$), we employ methods detailed in Appendix S1. This includes techniques for estimating latent factors ($\widehat{\boldsymbol{U}}$) and extending the framework to handle multiple treatments. Once these quantities are estimated, we treat $\boldsymbol{W} = [\boldsymbol{X}, \widehat{\boldsymbol{U}}] \in \mathbb{R}^{d+r}$ as the complete set of confounding covariates—combining both observed covariates ($\boldsymbol{X}$) and estimated unmeasured confounders ($\widehat{\boldsymbol{U}}$).

With this expanded set of covariates, we perform doubly robust estimation and inference as described in subsequent sections. This approach ensures that treatment effects are estimated while accounting for both observed and unmeasured confounding influences, improving robustness and reliability in causal inference.

Doubly robust estimation. Throughout the paper, we consider the log fold change (LFC) as the target estimand:

$$\tau_j := \log(\mathbb{E}[Y_j(1)]/\mathbb{E}[Y_j(0)]),$$

which quantifies the relative change in expected gene expression levels between treatment ($A = 1$) and control ($A = 0$) conditions for gene j. Extensions to other estimands are provided in Appendix S2.

The doubly robust estimation framework is a widely used approach that is agnostic to the underlying data-generating process. It provides valid estimation and inference results as long as either the conditional mean model (μ_j) or the propensity score model (π) is correctly specified. This robustness property ensures reliable causal effect estimation even in the presence of potential misspecification of one of the models.

More specifically, a one-step estimator $\widehat{\tau}_j$ of the estimand τ_j admits a linear expansion:

$$\widehat{\tau}_j - \tau_j = \frac{1}{n}\sum_{i=1}^n \eta_j(O_i; \pi, \mu_j) + o_{\mathbb{P}}(n^{-1/2}),$$

where $\eta_j(O_i; \pi, \mu_j)$ is the influence function of τ_j, which quantifies how individual observations contribute to the overall estimate. Here, $\pi(\boldsymbol{W}) = \mathbb{P}(A = a \mid \boldsymbol{W})$ is the propensity score model, and $\mu_j(\boldsymbol{W}, a) = \mathbb{E}[Y_j \mid W, A = a]$ is the outcome model for gene j. See Appendix S2 for detailed derivations of these functions.

To estimate the nuisance functions μ_j's (outcome models) and π (propensity score model), we use flexible statistical machine learning methods. Specifically, for outcome models μ_j, we employ generalized linear models (GLMs) with a negative binomial likelihood and log link function. This choice accounts for over-dispersion in count data while ensuring computational efficiency given the high dimensionality of genomic data. For the propensity score model π, we provide two built-in options: (i) logistic regression and (ii) random forests. In our experiments, random forests are configured with 1,000 trees, a minimum leaf size of 3, and a maximum tree depth of 11. Extrapolated cross-validation (ECV) (44) is used to select hyperparameters by minimizing the estimated mean squared error. Users can also supply alternative estimates for these nuisance functions if desired.

To perform inference, we first compute the estimated influence function values $\widehat{\eta}_j(O_i; \widehat{\pi}, \widehat{\mu}_j)$ and use them to estimate the variance for gene j:

$$\widehat{\sigma}_j^2 = \frac{\sqrt{n}}{n-1}\sum_{i=1}^n \widehat{\eta}_j(O_i; \widehat{\pi}, \widehat{\mu}_j)^2.$$

Using these quantities, a t-statistic for gene j can be computed as:

$$T_j = \frac{\widehat{\tau}_j - \tau_j}{\widehat{\sigma}_j}.$$

This statistic enables hypothesis testing and confidence interval construction for causal effects on gene expression.

False discovery rate control. Genomic studies often involve testing thousands of hypotheses simultaneously, making it crucial to control statistical Type-I errors. Two widely recognized error rate metrics are the Family-Wise Error Rate (FWER) and the False Discovery Rate (FDR), each suited to different contexts. Consider p hypothesis tests, let $\mathcal{S} \subset \{1, \ldots, p\}$ denote the set of discoveries, and $\mathcal{H}_0 \subset \{1, \ldots, p\}$ denote the set of true null hypotheses. The false discovery proportion (FDP) is defined as the ratio of false positives to total discoveries:

$$\text{FDP} = \frac{|\mathcal{S} \cap \mathcal{H}_0|}{|\mathcal{S}| \vee 1}.$$

The FWER controls the probability of making at least one false discovery:

$$\text{FWER} := \mathbb{P}(\text{FDP} > 0) \leq \alpha,$$

where $\alpha \in (0, 1)$ is a predefined significance level. This stringent control is particularly useful in scenarios where even a single false positive is unacceptable. However, FWER control often leads to reduced statistical power, especially in high-dimensional settings with many hypotheses, potentially overlooking true effects.

In contrast, FDR control provides a more balanced approach by controlling the expected proportion of false discoveries among all discoveries:

$$\text{FDR} := \mathbb{E}[\text{FDP}] \leq \alpha.$$

This approach enhances power in multiple testing scenarios and has become the standard for differential expression analysis in genomics due to its ability to identify more significant features while maintaining a low proportion of false positives (45). Importantly, FDR controls the *expected* proportion of false discoveries across repeated experiments but does not guarantee bounds on FDP in any single experiment. This distinction becomes critical in genomic studies where test statistics are often highly dependent, leading to variability in FDP across experiments.

To address limitations of standard FDR procedures, such as their inability to capture FDP variability in a single experiment, alternative error control metrics like False Discovery Exceedance (FDX) have been proposed:

$$\text{FDX} := \mathbb{P}(\text{FDP} \geq c) \leq \alpha,$$

for a threshold $c \in (0,1)$. FDX provides stricter control by limiting the probability that FDP exceeds a predefined threshold c. This makes it particularly useful in applications where minimizing false positives is critical or when restricting analysis to a small subset of discoveries is desired.

To ensure robust error rate control tailored to genomic applications, causarray implements two complementary strategies for FDR control: (i) Benjamini–Hochberg (BH) Procedure: The BH procedure (45) is applied directly to P-values obtained from the doubly robust estimation framework. BH controls the FDR under independence or specific positive dependence structures among test statistics. (ii) Gaussian Multiplier Bootstrap: For tighter control of FDP variability, particularly when test statistics are highly dependent, causarray incorporates a Gaussian multiplier bootstrap approach (Algorithm S2). This method simulates null distributions to estimate FDP more accurately and provides robust FDR control even under complex dependence structures (7).

The choice between BH and Gaussian multiplier bootstrap depends on the dependency structure among test statistics. While BH is computationally efficient and widely used, it may not adequately control FDR under strong dependencies. The Gaussian multiplier bootstrap, on the other hand, accounts for complex dependency structures and provides more accurate bounds on FDP variability. Additionally, incorporating FDX offers an extra layer of conservatism for applications where minimizing false positives is critical. By offering these complementary strategies, causarray ensures robust error rate control tailored to diverse genomic applications while balancing power and error control.

Data simulation and analysis

We consider two simulation settings. In the first simulation, we generate cells from zero-inflated Poisson distributions. In the second simulation, we use a specialized single-cell simulator Splatter (25) to generate cells with batch effects. Both simulations include 1 observed covariate and 4 unmeasured confounders. The details of the simulation are provided in Appendix S3.

Benchmarking methods. To evaluate the performance of differential expression (DE) testing, we compare causarray with several established methods, both with and without confounder adjustment. These methods are grouped into two categories based on whether they account for unmeasured confounders.

Methods without confounder adjustment include:

- Wilcoxon rank-sum test: This nonparametric test is applied to deviance residuals obtained by regressing gene expression counts on measured covariates using a negative binomial generalized linear model (GLM). The deviance residuals serve as input for the test, which does not explicitly account for unmeasured confounders.
- DESeq2 (26): This widely used method fits a negative binomial GLM to gene expression counts and adjusts for measured covariates. However, it does not account for unmeasured confounders, which may bias results in the presence of hidden variation.

Methods with confounder adjustment include:

- CoCoA-diff (R package `mmutilR 1.0.5`) (6): Designed for individual-level case-control studies, CoCoA-diff prioritizes disease genes by adjusting for confounders estimated from parametric models. After adjusting for these confounders, the Wilcoxon rank-sum test is applied to the adjusted residuals, as recommended in the original paper.
- CINEMA-OT (Python package `cinemaot 0.0.3`) (11): CINEMA-OT separates confounding sources of variation from perturbation effects using optimal transport matching to estimate counterfactual cell pairs. Similar to CoCoA-diff, the Wilcoxon rank-sum test is applied to the adjusted residuals of CINEMA-OT.
- RUV-III-NB (R package `ruvIIInb 0.8.2.0`) (13): This method normalizes gene expression data using pseudo-replicates and a negative binomial model to remove unwanted variation induced by library size differences. The Kruskal-Wallis test (equivalent to the Wilcoxon test for two-group comparisons) is then applied to log-percentile adjusted counts, as suggested by the authors. However, RUV-III-NB does not directly adjust for library size and its ability to control FDR remains unclear, as it was not demonstrated in their experiments.
- RUV (R package `ruv 0.9.7.1`) (12): RUVr is used to estimate unmeasured confounders, which are then incorporated into DESeq2 for statistical inference based on both observed and estimated covariates. Before running RUV, we successively use the functions `calcNormFactors`, `estimateGLMCommonDisp`, `estimateGLMTagwiseDisp`, and `glmFit` of edgeR package (4.0.16) (46) to extract residuals not explained by observed covariates and treatments.

This comprehensive benchmarking enables a thorough evaluation of each method's ability to address unmeasured confounder estimation and perform robust statistical inference in simulated data settings.

Evaluation metrics. To compare the performance of different methods, we use four evaluation metrics, focusing on two aspects: confounder estimation and biological signal preservation. DESeq2 and Wilcoxon are excluded from confounder estimation evaluation as they do not estimate unmeasured confounders or counterfactuals.

The performance of confounder estimation is assessed using two clustering-based metrics: Adjusted Rand Index (ARI) and Average Silhouette Width (ASW) (47). These metrics evaluate the quality of mixing in response and confounder spaces, respectively. Formally, measures the similarity between the clustering results based on the estimated control responses $Y(0)$ and the true cell-type labels of the same samples. It adjusts for similarities that occur by chance:

$$\mathrm{ARI} = \frac{\sum_{ij}\binom{n_{ij}}{2} - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}}{\frac{1}{2}[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - [\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}]/\binom{n}{2}},$$

where n is the total number of samples, n_{ij} is the number of samples in both cluster i and partition j, a_i is the sum over rows in the contingency table, and b_j is the sum over columns. Higher ARI values indicate better conservation of cell identity based on estimated counterfactuals compared to true labels. ARI ranges from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating random clustering. On the other hand, ASW quantifies how well each sample fits within its assigned cluster compared to other clusters. It is defined as:

$$\mathrm{ASW} = \frac{1}{n}\sum_{i=1}^{n} \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

where $a(i)$ is the average dissimilarity of sample i to all other samples within its cluster, and $b(i)$ is the average dissimilarity to samples in the nearest neighboring cluster. ASW values range from -1 to 1, with higher values indicating better-defined clusters (47). For both metrics, median scores are scaled between 0 and 1 across methods within each simulation setup. For these two metrics, we use the implementations from the `scib` (`1.1.5`) package (47).

To evaluate biological signal preservation, we use False Positive Rate (FPR) and True Positive Rate (TPR), which are standard metrics derived from confusion matrices: PR quantifies the proportion of false positives among all true negatives:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}},$$

where FP and TN are false positives and true negatives, respectively. A lower FPR indicates fewer false discoveries relative to true negatives. Also known as sensitivity or recall, TPR measures the proportion of true positives among all actual positives:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}},$$

where TP and FN are true positives and false negatives, respectively. A higher TPR indicates better detection of true signals. These metrics provide complementary insights: FPR evaluates specificity by penalizing false discoveries, while TPR assesses sensitivity by rewarding correct detections. Together, they measure how well a method balances identifying true signals while avoiding false discoveries.

Single-cell Perturb-Seq dataset

We utilize the Perturb-Seq dataset from (17), which enables high-resolution transcriptomic profiling of genetic perturbations in excitatory neurons. This scalable platform systematically investigates gene functions across diverse cell types and perturbation conditions, providing critical insights into neurodevelopmental processes (17). We focus on excitatory neurons of the dataset, a key population implicated in neurodevelopmental disorders such as autism spectrum disorders and neurodevelopmental delay, with perturbations targeting genes involved in neuronal development and synaptic function (17).

For preprocessing, we filter out cells with perturbations measured in fewer than 50 cells and genes expressed in fewer than 50 cells, resulting in a dataset containing 2926 cells under 30 perturbation conditions. The GFP (Green Fluorescent Protein) condition is used as a negative control to benchmark the effects of other perturbations by providing a baseline for comparison in downstream analyses. After filtering lowly expressed genes with a maximum count of fewer than 10, we retain 3221 genes.

The batch design is highly correlated with perturbation conditions; therefore, it is not included as a covariate in the model for testing. Instead, only the intercept is included as a covariate. For propensity score estimation, we incorporate the logarithm of library sizes as an additional covariate to account for technical variability and use GLM as the propensity score model.

Single-nucleus Alzheimer's disease dataset

This study integrates data from three single-nucleus RNA sequencing (snRNA-seq) datasets to investigate Alzheimer's disease (AD): the ROSMAP-AD dataset (35) and two datasets from the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium (36), covering the middle temporal gyrus (MTG) and prefrontal cortex (PFC). These datasets provide complementary insights into AD pathology across different brain regions and donor cohorts.

The ROSMAP-AD dataset is derived from a single-nucleus transcriptomic atlas of the aged human prefrontal cortex, including 2.3 million cells from postmortem brain samples of 427 individuals with varying degrees of AD pathology and cognitive impairment (35). To ensure balanced representation across subjects, we perform stratified down-sampling of 300 cells per subject, focusing on excitatory neurons while excluding two rare subtypes ('Exc RELN CHD7' and 'Exc NRGN'). This preprocessing results in a dataset with 124997 cells and 33538 genes.

Next, we create pseudo-bulk gene expression profiles by aggregating gene expression counts across cells for each subject. Genes expressed in fewer than 10 subjects are filtered out, resulting in a final dataset of 427 samples and 26,106 genes. Binary treatment is defined based on the variable 'age_first_ad_dx', which approximates the "age at the time of onset of Alzheimer's dementia." Covariates included in the analysis are 'msex' (biological sex), 'pmi' (postmortem interval), and 'age_death' (age at death). Missing values for 'pmi' are imputed using the median of observed values.

The SEA-AD data are obtained from a multimodal cell atlas of AD developed by the Seattle Alzheimer's Disease Brain Cell Atlas (SEA-AD) consortium (36). This resource includes snRNA-seq datasets from two brain regions: the middle temporal gyrus (MTG) and prefrontal cortex (PFC), covering 84 donors with varying AD pathologies.

For both MTG and PFC datasets, we perform stratified down-sampling of 300 cells per subject, focusing on excitatory neurons. Pseudo-bulk gene expression profiles are created by aggregating counts across cells for each subject. Genes expressed in fewer than 40 subjects are filtered out, resulting in final datasets with: 80 samples and 24,621 genes for MTG and 80 samples and 25,361 genes for PFC. Covariates included in the analysis are 'sex', 'pmi', and 'Age_at_death'. These variables account for biological and technical variability across donors.

To enable comparative analyses across the three datasets (ROSMAP-AD, SEA-AD MTG, and SEA-AD PFC), we restrict the analysis to 15586 common genes that are expressed in all three datasets. Genes with a maximum expression count below 10 among subjects are excluded to ensure robust comparisons.

CODE AVAILABILITY

The code for reproducing the results in the paper and the causarray package can be accessed at https://github.com/jaydu1/causarray.

DATA AVAILABILITY

All datasets used in this paper are previously published and freely available, except the metadata for donors from the ROSMAP cohort. The Perturb-seq dataset is available through the Broad single cell portal as txt files. The gene expression count matrices of ROSMAP-AD datasets (35) can be obtained from supplementary website, which have been deidentified to protect confidentiality - the mapping to ROSMAP IDs and complete metadata can be found on Synapse as Seurat objects (rds files). The SEA-AD datasets of nuclei-by-gene matrices with counts and normalized expression values from the snRNA-seq assay (36) are available through the Open Data Registry in an AWS bucket (sea-ad-single-cell-profiling) as AnnData objects (h5ad files).

ACKNOWLEDGEMENTS

This work used the Bridges-2 system at the Pittsburgh Supercomputing Center (PSC) through allocation MTH230011P from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program. This project was funded by the National Institute of Mental Health (NIMH) grant R01MH123184.

Bibliography

1. Valentine Svensson, Roser Vento-Tormo, and Sarah A Teichmann. Exponential scaling of single-cell rna-seq in the past decade. *Nature protocols*, 13(4):599–604, 2018.
2. Itay Tirosh, Benjamin Izar, Sanjay M Prakadan, Marc H Wadsworth, Daniel Treacy, John J Trombetta, Asaf Rotem, Christopher Rodman, Christine Lian, George Murphy, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell rna-seq. *Science*, 352(6282):189–196, 2016.
3. Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular systems biology*, 15(6):e8746, 2019.
4. Editorial. A focus on single-cell omics. *Nat Rev Genet*, 24(8):485, Aug 2023. doi: 10.1038/s41576-023-00628-3.
5. David Lähnemann, Johannes Köster, Ewa Szczurek, Davis J McCarthy, Stephanie C Hicks, Mark D Robinson, Catalina A Vallejos, Kieran R Campbell, Niko Beerenwinkel, Ahmed Mahfouz, et al. Eleven grand challenges in single-cell data science. *Genome biology*, 21:1–35, 2020.
6. Yongjin P Park and Manolis Kellis. Cocoa-diff: counterfactual inference for single-cell gene expression analysis. *Genome Biology*, 22(1):1–23, 2021.
7. Jin-Hong Du, Zhenghao Zeng, Edward H Kennedy, Larry Wasserman, and Kathryn Roeder. Causal inference for genomic data with multiple heterogeneous outcomes. *arXiv preprint arXiv:2404.09119*, 2024.
8. Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.
9. Jay Shendure, Gregory M Findlay, and Matthew W Snyder. Genomic medicine-progress, pitfalls, and promise. *Cell*, 177(1):45–57, Mar 2019. doi: 10.1016/j.cell.2019.02.003.
10. Pedro Sanchez, Jeremy P Voisey, Tian Xia, Hannah I Watson, Alison Q O'Neil, and Sotirios A Tsaftaris. Causal machine learning for healthcare and precision medicine. *R Soc Open Sci*, 9(8):220638, Aug 2022. doi: 10.1098/rsos.220638.
11. Mingze Dong, Bao Wang, Jessica Wei, Antonio H de O. Fonseca, Curtis J Perry, Alexander Frey, Feriel Ouerghi, Ellen F Foxman, Jeffrey J Ishizuka, Rahul M Dhodapkar, et al. Causal identification of single-cell experimental perturbation effects with cinema-ot. *Nature Methods*, pages 1–11, 2023.
12. Davide Risso, John Ngai, Terence P Speed, and Sandrine Dudoit. Normalization of rna-seq data using factor analysis of control genes or samples. *Nat Biotechnol*, 32(9):896–902, Sep 2014. doi: 10.1038/nbt.2931.
13. Agus Salim, Ramyar Molania, Jianan Wang, Alysha De Livera, Rachel Thijssen, and Terence P Speed. Ruv-iii-nb: Normalization of single cell rna-seq data. *Nucleic Acids Research*, 50(16):e96–e96, 2022.
14. Martin Kampmann. Crispr-based functional genomics for neurological disease. *Nat Rev Neurol*, 16(9):465–480, Sep 2020. doi: 10.1038/s41582-020-0373-z.
15. Derek Hong and Lilia M Iakoucheva. Therapeutic strategies for autism: targeting three levels of the central dogma of molecular biology. *Transl Psychiatry*, 13(1):58, Feb 2023. doi: 10.1038/s41398-023-02356-y.
16. Junyun Cheng, Gaole Lin, Tianhao Wang, Yunzhu Wang, Wenbo Guo, Jie Liao, Penghui Yang, Jie Chen, Xin Shao, Xiaoyan Lu, Ling Zhu, Yi Wang, and Xiaohui Fan. Massively parallel CRISPR-based genetic perturbation screening at single-cell resolution. *Adv Sci (Weinh)*, 10(4):e2204484, Feb 2023. doi: 10.1002/advs.202204484.
17. Xin Jin, Sean K Simmons, Amy Guo, Ashwin S Shetty, Michelle Ko, Lan Nguyen, Vahbiz Jokhi, Elise Robinson, Paul Oyler, Nathan Curry, Giulio Deangeli, Simona Lodato, Joshua Z Levin, Aviv Regev, Feng Zhang, and Paola Arlotta. In vivo perturb-seq reveals neuronal and glial abnormalities associated with autism risk genes. *Science*, 370(6520), Nov 2020. doi: 10.1126/science.aaz6063.
18. Jin-Hong Du, Larry Wasserman, and Kathryn Roeder. Simultaneous inference for generalized linear models with unmeasured confounders. *arXiv preprint arXiv:2309.07261*, 2023.
19. Jin-Hong Du, Kathryn Roeder, and Larry Wasserman. Assumption-lean post-integrated inference with negative control outcomes. *arXiv preprint arXiv:2410.04996*, 2024.
20. Jeffrey T Leek and John D Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet*, 3(9):1724–35, Sep 2007. doi: 10.1371/journal.pgen.0030161.
21. Abhishek Sarkar and Matthew Stephens. Separating measurement and expression models clarifies confusion in single-cell rna sequencing analysis. *Nature genetics*, 53(6):770–777, 2021.
22. James M Robins, Andrea Rotnitzky, and Lue Ping Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866, 1994.
23. Daniel O Scharfstein, Andrea Rotnitzky, and James M Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models. *Journal of the American Statistical Association*, 94(448):1096–1120, 1999.
24. José L McFaline-Figueroa, Sanjay Srivatsan, Andrew J Hill, Molly Gasperini, Dana L Jackson, Lauren Saunders, Silvia Domcke, Samuel G Regalado, Paul Lazarchuck, Sarai Alvarez, et al. Multiplex single-cell chemical genomics reveals the kinase dependence of the response to targeted therapy. *Cell Genomics*, 4(2), 2024.
25. Luke Zappia, Belinda Phipson, and Alicia Oshlack. Splatter: simulation of single-cell RNA sequencing data. *Genome biology*, 18(1):174, 2017.
26. Michael Love, Simon Anders, and Wolfgang Huber. Differential analysis of count data–the deseq2 package. *Genome Biol*, 15(550):10–1186, 2014.
27. Michael V Lombardo. Ribosomal protein genes in post-mortem cortical tissue and ipsc-derived neural progenitor cells are commonly upregulated in expression in autism. *Mol Psychiatry*, 26(5):1432–1435, May 2021. doi: 10.1038/s41380-020-0773-x.
28. Karina Griesi-Oliveira and Maria Rita Passos-Bueno. Reply to lombardo, 2020: An additional route of investigation: what are the mechanisms controlling ribosomal protein genes dysregulation in autistic neuronal cells? *Mol Psychiatry*, 26(5):1436–1437, May 2021. doi: 10.1038/s41380-020-0792-7.
29. Matthew A Lalli, Denis Avey, Joseph D Dougherty, Jeffrey Milbrandt, and Robi D Mitra. High-throughput single-cell functional elucidation of neurodevelopmental disease–associated genes reveals convergent mechanisms altering neuronal differentiation. *Genome research*, 30(9):1317–1331, 2020.
30. Jack M Fu, F Kyle Satterstrom, Minshi Peng, Harrison Brand, Ryan L Collins, Shan Dong, Brie Wamsley, Lambertus Klei, Lily Wang, Stephanie P Hao, Christine R Stevens, Caroline Cusick, Mehrtash Babadi, Eric Banks, Brett Collins, Sheila Dodge, Stacey B Gabriel, Laura Gauthier, Samuel K Lee, Lindsay Liang, Alicia Ljungdahl, Behrang Mahjani, Laura Sloofman, Andrey N Smirnov, Mafalda Barbosa, Catalina Betancur, Alfredo Brusco, Brian H Y Chung, Edwin H Cook, Michael L Cuccaro, Enrico Domenici, Giovanni Battista Ferrero, J Jay Gargus, Gail E Herman, Irva Hertz-Picciotto, Patricia Maciel, Dara S Manoach, Maria Rita Passos-Bueno, Antonio M Persico, Alessandra Renieri, James S Sutcliffe, Flora Tassone, Elisabetta Trabetti, Gabriele Campos, Simona Cardaropoli, Diana Carli, Marcus C Y Chan, Chiara Fallerini, Elisa Giorgio, Ana Cristina Girardi, Emily Hansen-Kiss, So Lun Lee, Carla Lintas, Yunin Ludena, Rachel Nguyen, Lisa Pavinato, Margaret Pericak-Vance, Isaac N Pessah, Rebecca J Schmidt, Moyra Smith, Claudia I S Costa, Slavica Trajkova, Jaqueline Y T Wang, Mullin H C Yu, Autism Sequencing Consortium (ASC), Broad Institute Center for Common Disease Genomics (Broad-CCDG), iPSYCH-BROAD Consortium, David J Cutler, Silvia De Rubeis, Joseph D Buxbaum, Mark J Daly, Bernie Devlin, Kathryn Roeder, Stephan J Sanders, and Michael E Talkowski. Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. *Nat Genet*, 54(9):1320–1331, Sep 2022. doi: 10.1038/s41588-022-01104-0.
31. Lei Zhang, Ning-Ning Song, Qiong Zhang, Wan-Ying Mei, Chun-Hui He, Pengcheng Ma, Ying Huang, Jia-Yin Chen, Bingyu Mao, Bing Lang, et al. Satb2 is required for the regionalization of retrosplenial cortex. *Cell Death & Differentiation*, 27(5):1604–1617, 2020.
32. Nico Wahl, Sergio Espeso-Gil, Paola Chietera, Amelie Nagel, Aodán Laighneach, Derek W Morris, Prashanth Rajarajan, Schahram Akbarian, Georg Dechant, and Galina Apostolova. Satb2 organizes the 3d genome architecture of cognition in cortical neurons. *Molecular Cell*, 84(4):621–639, 2024.
33. Clemens Jaitner, Chethan Reddy, Andreas Abentung, Nigel Whittle, Dietmar Rieder, Andrea Delekate, Martin Korte, Gaurav Jain, Andre Fischer, Farahnaz Sananbenesi, et al. Satb2 determines mirna expression and long-term memory in the adult central nervous system. *Elife*, 5:e17361, 2016.

34. Qiufang Guo, Yaqiong Wang, Qing Wang, Yanyan Qian, Yinmo Jiang, Xinran Dong, Huiyao Chen, Xiang Chen, Xiuyun Liu, Sha Yu, et al. In the developing cerebral cortex: axonogenesis, synapse formation, and synaptic plasticity are regulated by satb2 target genes. *Pediatric Research*, 93(6):1519–1527, 2023.
35. Hansruedi Mathys, Zhuyu Peng, Carles A Boix, Matheus B Victor, Noelle Leary, Sudhagar Babu, Ghada Abdelhady, Xueqiao Jiang, Ayesha P Ng, Kimia Ghafari, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to alzheimer's disease pathology. *Cell*, 186(20):4365–4385, 2023.
36. Mariano I Gabitto, Kyle J Travaglini, Victoria M Rachleff, Eitan S Kaplan, Brian Long, Jeanelle Ariza, Yi Ding, Joseph T Mahoney, Nick Dee, Jeff Goldy, et al. Integrated multimodal cell atlas of alzheimer's disease. *Nature Neuroscience*, pages 1–18, 2024.
37. Ming-Hui Lee, Yao-Hsiang Shih, Sing-Ru Lin, Jean-Yun Chang, Yu-Hao Lin, Chun-I Sze, Yu-Min Kuo, and Nan-Shan Chang. Zfra restores memory deficits in alzheimer's disease triple-transgenic mice by blocking aggregation of trappc6aδ, sh3glb2, tau, and amyloid β, and inflammatory nf-κb activation. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 3(2):189–204, 2017.
38. Kan He, Jian Zhang, Justin Liu, Yandi Cui, Leyna G Liu, Shoudong Ye, Qian Ban, Ruolan Pan, and Dahai Liu. Functional genomics study of protein inhibitor of activated stat1 in mouse hippocampal neuronal cells revealed by rna sequencing. *Aging (Albany NY)*, 13(6): 9011, 2021.
39. Yasar Arfat T Kasu, Akshaya Arva, Jess Johnson, Christin Sajan, Jasmin Manzano, Andrew Hennes, Jacy Haynes, and Christopher S Brower. Bag6 prevents the aggregation of neurodegeneration-associated fragments of tdp43. *Iscience*, 25(5), 2022.
40. Raffaella Nativio, Yemin Lan, Greg Donahue, Oksana Shcherbakova, Noah Barnett, Katelyn R Titus, Harshini Chandrashekar, Jennifer E Phillips-Cremins, Nancy M Bonini, and Shelley L Berger. The chromatin conformation landscape of alzheimer's disease. *bioRxiv*, pages 2024–04, 2024.
41. Wenbin Zhou and Jin-Hong Du. Distance-preserving spatial representations in genomic data. *arXiv preprint arXiv:2408.00911*, 2024.
42. Jin-Hong Du, Tianyu Chen, Ming Gao, and Jingshu Wang. Joint trajectory inference for single-cell genomics using deep learning with a mixture prior. *Proceedings of the National Academy of Sciences*, 121(37):e2316256121, 2024.
43. Mohammad Lotfollahi, Anna Klimovskaia Susmelj, Carlo De Donno, Leon Hetzel, Yuge Ji, Ignacio L Ibarra, Sanjay R Srivatsan, Mohsen Naghipourfar, Riza M Daza, Beth Martin, et al. Predicting cellular responses to complex perturbations in high-throughput screens. *Molecular systems biology*, 19(6):e11517, 2023.
44. Jin-Hong Du, Pratik Patil, Kathryn Roeder, and Arun Kumar Kuchibhotla. Extrapolated cross-validation for randomized ensembles. *Journal of Computational and Graphical Statistics*, pages 1–12, 2024.
45. Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.
46. Yunshun Chen, Lizhong Chen, Aaron TL Lun, Pedro L Baldoni, and Gordon K Smyth. edger 4.0: powerful differential analysis of sequencing data with expanded functionality and improved support for small counts and larger datasets. *bioRxiv*, pages 2024–01, 2024.
47. Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022.
48. Yingxin Lin, Shila Ghazanfar, Kevin YX Wang, Johann A Gagnon-Bartsch, Kitty K Lo, Xianbin Su, Ze-Guang Han, John T Ormerod, Terence P Speed, Pengyi Yang, et al. scmerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell rna-seq datasets. *Proceedings of the National Academy of Sciences*, 116(20):9775–9784, 2019.
49. Edward H Kennedy, Shreya Kangovi, and Nandita Mitra. Estimating scaled treatment effects with multiple outcomes. *Statistical methods in medical research*, 28(4):1094–1104, 2019.

Supplementary Note S1: Confounder estimation

Comparison with reference-based confounder adjustment methods

Another approach to adjust for the unmeasured confounders is to utilize the information from negative control genes. This includes scMerge (48), RUV-III-NB (13) and RUVSeq (12) etc. These methods require users to specify a set of negative control genes, such as housekeeping genes, which are assumed to be solely due to unwanted variation between the two cells. The approach necessitates strong prior knowledge to accurately identify negative control genes, which may not always be available, especially in less well-characterized biological systems. This reliance on prior knowledge can limit the applicability of the method in novel or poorly understood contexts.

Algorithm

To estimate the unmeasured confounders, we employ an improved version of GCATE (18). Suppose (X_i, A_i, Y_i) for $i = 1, \ldots, n$ are n independently and identically distributed samples coming from the same distribution as $(X, A, Y) \in \mathbb{R}^d \times \mathbb{R}^a \times \mathbb{R}^p$. Here, A consists of a treatments and can be both continuous and discrete for the purpose of confounder estimation. Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}, \boldsymbol{A} \in \mathbb{R}^{n\times a}, \boldsymbol{Y} \in \mathbb{R}^{n\times p}$ denote the design matrix, treatment matrix and gene expression matrix, respectively. To account for different library sizes, we model the mean of the size-normalized counts

$$\mu_{ij} = \frac{Y_{ij}}{s_i},$$

which is assumed to follow a negative binomial distribution. Technically, μ_{ij}'s should be non-negative integers; however, the likelihood-based approaches work seamlessly even when they are non-negative real numbers. Here s_i is the size factor of cell i, which will be specified later. We assume the conditional mean is characterized by a generalized linear model

$$\log \mu_{ij} \sim A_i + X_i + U_i,$$

and its dispersion parameter ϕ is predetermined.

The adjusted expression μ_{ij} of the ith observation and the jth gene has the density:

$$p(\mu_{ij} \mid \theta_{ij}) = h(\mu_{ij}) \exp\left(\mu_{ij}\theta_{ij} - A(\theta_{ij})\right),$$

where θ_{ij} is the natural parameter. In matrix form, the natural parameters decompose as

$$\boldsymbol{\Theta} = \widetilde{\boldsymbol{X}}\boldsymbol{B}^\top + \boldsymbol{U}\boldsymbol{\Gamma}^\top,$$

where $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{A}] \in \mathbb{R}^{n\times(d+a)}$, $\boldsymbol{B} \in \mathbb{R}^{p\times(d+a)}$, $\boldsymbol{U} \in \mathbb{R}^{n\times r}$, and $\boldsymbol{\Gamma} \in \mathbb{R}^{p\times r}$ are unknown. Note that μ_{ij}'s are conditionally independent given the natural parameter $\boldsymbol{\Theta}$. With this notation, the procedure of unmeasured confounder estimation is summarized in Algorithm S1, and the details of the method are described below.

Estimation of size factors. We follow the procedure in (26) to compute the size factors s_i for $i = 1, \ldots, n$. We start by calculating the geometric Mean for each gene j:

$$g_j = \exp\left(\frac{\sum_i \log(Y_{ij}) \mathbb{1}\{Y_{ij} > 0\}}{\sum_i \mathbb{1}\{Y_{ij} > 0\}}\right).$$

Next, for each sample i, compute the initial size factors:

$$d_i = \exp\left(\operatorname*{median}_{j:Y_{ij}>0}\{\log(Y_{ij}) - \log(g_j)\}\right).$$

Finally, we normalize these size factors to have a geometric mean of 1 across all samples:

$$s_i = \frac{d_i}{(\prod_i d_i)^{1/n}}. \qquad \textbf{(S1)}$$

The size factors can then be used to normalize gene expression data, adjusting for differences in sequencing depth and other systematic biases across samples. The normalization ensures that observed differences in expression levels reflect true biological variation rather than technical artifacts.

Algorithm S1 Unmeasured confounder estimation

```
Input: A data matrix $\boldsymbol{Y} \in \mathbb{R}^{n\times p}$, a design matrix $\widetilde{\boldsymbol{X}} = [\boldsymbol{X}, \boldsymbol{A}] \in \mathbb{R}^{n\times(d+a)}$, a natural number $r \geq 1$ (the number of latent
    factors), a constant $C = 2\times 10^3$ for the norm constraint
1: (Estimation of size factors) Compute $\boldsymbol{s} \in \mathbb{R}^n$ according to Eq. (S1).
2: (Estimation of dispersion parameters) Compute $\boldsymbol{\phi} \in \mathbb{R}^p$ according to Eq. (S2).
3: (Estimation of marginal effects $\boldsymbol{F}$ and uncorrelated latent components $\boldsymbol{W}\boldsymbol{\Gamma}^\top$) Solve optimization problem Eq. (S3) to
   obtain $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top$ and the initial estimate of the natural parameter matrix $\widehat{\boldsymbol{\Theta}}_0 = \widetilde{\boldsymbol{X}}\widehat{\boldsymbol{F}}^\top + \widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top$ by alternative maximization:

   $$\begin{aligned} \widehat{\boldsymbol{F}}, \widehat{\boldsymbol{W}}_0, \widehat{\boldsymbol{\Gamma}}_0 \in & \mathop{\mathrm{argmin}}_{\boldsymbol{F}\in\mathbb{R}^{p\times(d+a)}, \boldsymbol{W}\in\mathbb{R}^{n\times r}, \boldsymbol{\Gamma}\in\mathbb{R}^{p\times r}} \mathcal{L}(\widetilde{\boldsymbol{X}}\boldsymbol{F}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top) \\ \text{subject to} \quad & \widetilde{\boldsymbol{X}}\boldsymbol{F}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top \in \mathcal{B}_C^{n\times p}, \qquad \mathcal{P}_{\widetilde{\boldsymbol{X}}}\boldsymbol{W} = \boldsymbol{0}. \end{aligned} \tag{S3}$$

4: (Estimation of latent coefficients $\boldsymbol{\Gamma}$) Set $\widehat{\boldsymbol{W}} := \sqrt{n}\boldsymbol{Q}\boldsymbol{\Sigma}^{1/2}$ and $\widehat{\boldsymbol{\Gamma}} := \sqrt{p}\boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$, where $\widehat{\boldsymbol{W}}_0\widehat{\boldsymbol{\Gamma}}_0^\top = \sqrt{np}\boldsymbol{Q}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the
   condensed SVD with $\boldsymbol{Q} \in \mathbb{R}^{n\times r}$, $\boldsymbol{\Sigma} \in \mathbb{R}^{r\times r}$, $\boldsymbol{V} \in \mathbb{R}^{p\times r}$.
5: (Estimation of direct effects $\boldsymbol{B}$ and latent factors $\boldsymbol{U}$) Solve optimization problem Eq. (S4) to obtain $(\widehat{\boldsymbol{B}}, \widehat{\boldsymbol{U}})$:

   $$\begin{aligned} \widehat{\boldsymbol{B}}, \widehat{\boldsymbol{U}} = & \mathop{\mathrm{argmin}}_{\boldsymbol{B}\in\mathbb{R}^{p\times(d+a)}, \boldsymbol{U}\in\mathbb{R}^{p\times r}} \mathcal{L}(\widetilde{\boldsymbol{X}}\boldsymbol{B}^\top + \boldsymbol{U}\widehat{\boldsymbol{\Gamma}}^\top) + \sum_{j=1}^p \lambda_j \|\boldsymbol{B}_{\cdot j}\|_1 \\ \text{subject to} \quad & \widetilde{\boldsymbol{X}}\boldsymbol{B}^\top + \boldsymbol{U}\widehat{\boldsymbol{\Gamma}}^\top \in \mathcal{B}_C^{n\times p}, \qquad \mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}\boldsymbol{B} = \boldsymbol{0}. \end{aligned} \tag{S4}$$

Output: Return the estimated confounders $\widehat{\boldsymbol{U}}$.
```

Estimation of dispersion parameters. To estimate the dispersion parameter, we first fit generalized linear models (GLMs) on the data and obtain the estimated mean expression of gene j, denoted as $\widehat{\nu}_j$ for $j = 1, \dots, p$. Note that when μ_{ij} comes from a Negative Binomial distribution, its variance is given by

$$\mathrm{Var}(\mu_{ij} \mid \theta_{ij}) = \nu(1 + \alpha_j \nu),$$

where $\nu = \mathbb{E}[\mu_{ij} \mid \theta_{ij}]$ is the conditional mean while α_j is the dispersion parameter of the NB1 form. In the form of exponential family parameterized by the parameter ϕ_j, α_j is the reciprocal of ϕ_j, namely, $\alpha_j = 1/\phi_j$. By methods of moments, we can solve the following equation to obtain an estimator $\widehat{\phi}_j$ for ϕ_j:

$$\frac{1}{n}\sum_{i=1}^n (y_{ij} - \widehat{\nu}_j)^2 = \widehat{\nu}_j(1 + \alpha\widehat{\nu}_j).$$

Finally, we clip $\widehat{\alpha}_j$ to be in $[10^{-2}, 10^2]$ and set $\widehat{\phi}_j = 1/\widehat{\alpha}_j$. The estimated dispersion parameter has a close-form expression:

$$\phi_j = \min\left\{\max\left\{\frac{\widehat{\nu}_j^2}{\frac{1}{n}\sum_{i=1}^n (y_{ij} - \widehat{\nu}_j)^2 - \widehat{\nu}_j}, 0.01\right\}, 100\right\}. \tag{S2}$$

Estimation of marginal effects by joint likelihood estimation. The negative log-likelihood function of the data is given by

$$\mathcal{L}(\boldsymbol{\Theta}) = \mathcal{L}(\boldsymbol{B}, \boldsymbol{U}, \boldsymbol{\Gamma}) = -\frac{1}{n}\sum_{i=1}^n\sum_{j=1}^p \left(\mu_{ij}\theta_{ij} - A(\theta_{ij}) + \log\begin{pmatrix}\mu_{ij} + \phi_j - 1 \\ \mu_{ij}\end{pmatrix}\right).$$

Although this is a nonconvex optimization problem, an alternative descent algorithm as in (18) can be employed to solve it efficiently. By rewriting $\boldsymbol{\Theta} = \widetilde{\boldsymbol{X}}\boldsymbol{B}^\top + \boldsymbol{Z}\boldsymbol{\Gamma}^\top$ as $\boldsymbol{\Theta} = \widetilde{\boldsymbol{X}}\boldsymbol{F}^\top + \boldsymbol{W}\boldsymbol{\Gamma}^\top$ with $\mathcal{P}_{\widetilde{\boldsymbol{X}}}\boldsymbol{W} = \boldsymbol{0}$, we can disentangle the marginal effects and the uncorrelated latent components. This is correspond to step 3 of Algorithm S1. Each entry of the estimated natural parameter matrix is constrained within the Euclidean ball $\mathcal{B}_C$ with radius C ($C = 2\times 10^3$ by default).

Before alternative maximization, we compute deviance residuals $\boldsymbol{R}$ from the NB GLM fits with offsets $\log\boldsymbol{s}$ and dispersion parameters ϕ, and initialize the uncorrelated confounders by $\boldsymbol{W} = \mathcal{P}_{\widetilde{\boldsymbol{X}}}^{\perp}\boldsymbol{U}_{\boldsymbol{R}}$ where $\boldsymbol{U}_{\boldsymbol{R}} \in \mathbb{R}^{n\times r}$ contains the first r left singular vectors of $\boldsymbol{R}$. Here, the projection $\mathcal{P}_{\widetilde{\boldsymbol{X}}}^{\perp}$ ensures that $\boldsymbol{W}$ is uncorrelated with $\boldsymbol{X}$. Then, we initialize the marginal effects $\boldsymbol{F}$ and latent coefficient $\boldsymbol{\Gamma}$ by solving GLMs with covariates $[\widetilde{\boldsymbol{X}}, \boldsymbol{W}]$. In particular, when the intercept is included in the covariates, the initial value of $\boldsymbol{W}$ also has zero means per column.

Estimation of latent coefficients. Because the (uncorrelated) latent factors are identifiable only up to scaling and rotations, we rescle the estimate at step 4 of Algorithm S1. This ensures the eigenvalues of $\widehat{\boldsymbol{W}}$ and $\widehat{\boldsymbol{\Gamma}}$ have the same order, making the alternative optimization more stable.

Estimation of confounding effects by adaptive penalization. The last step is to jointly recover the direct effects and the unmeasured confounders. This is done by imposing orthogonality between $\widehat{\boldsymbol{B}}$ and $\widehat{\boldsymbol{\Gamma}}$, as well as imposing sparsity on $\widehat{\boldsymbol{B}}$. The former ensures the gene-wise effects of the observed covariates and the unmeasured confounders are uncorrelated, while the latter aims to reveal signals from noisy measurements.

The direct effect $\boldsymbol{B}$ is initialized as $\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\widehat{\boldsymbol{F}}$. Then, Initialize $\boldsymbol{Z}$ and $\boldsymbol{\Gamma}$ using the SVD of the matrix $\boldsymbol{X}\widehat{\boldsymbol{F}}^{\top}\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}} + \widehat{\boldsymbol{W}}\widehat{\boldsymbol{\Gamma}}^{\top} = \boldsymbol{U}'\boldsymbol{\Sigma}'\boldsymbol{V}'^{\top}$. Let $\boldsymbol{Z} = (\boldsymbol{U}'\boldsymbol{\Sigma}'^{1/2})_{1:r}$ and $\boldsymbol{\Gamma} = (\boldsymbol{V}'\boldsymbol{\Sigma}'^{1/2})_{1:r}$ be the initialized values.

To account for different scales of the effects induced by different treatment conditions, we propose to use the adaptive lasso to induce sparsity of effects from multiple treatments. More specifically, the regularization parameters are set as $\lambda_j = \lambda/\|(\mathcal{P}_{\widehat{\boldsymbol{\Gamma}}}^{\perp}\widehat{\boldsymbol{F}})_{\cdot j}\|_1$ for $j = 1, \ldots, p$ in optimization problem S4.

Because of regularization, the estimate $\widehat{\boldsymbol{B}}$ is typically biased towards zero, so we don't use it for downstream analysis. It is possible to perform inference with additional debiasing procedure (18). However, we use a more flexible semiparametric inference method, as described below.

Determine the number of latent factors r

To determine the number of unmeasured confounders r, one can use the joint-likelihood-based information criterion (JIC) (18). The JIC value is the sum of deviance and a penalty on model complexity:

$$\mathrm{JIC}(\widehat{\boldsymbol{\Theta}}^{(r)}) = -2\sum_{i=1}^{n}\sum_{j=1}^{p}\log p(\mu_{ij} \mid \widehat{\theta}_{ij}^{(r)}) + c_{\mathrm{JIC}} \cdot \frac{(d+a+r)\log(n\wedge p)}{n\wedge p},$$

where $\widehat{\boldsymbol{\Theta}}^{(r)}$ is the estimated natural parameter matrix with r unmeasured confounders and $d+a$ observed covariates, and $c_{\mathrm{JIC}} > 0$ is a universal constant set to be 1 by default.

Supplementary Note S2: Doubly robust inference

Target estimands

For semiparametric inference, a target estimand is a distributional functional of the observed random variables. For example, we can consider the average treatment effects (ATE), the standardized average treatment effect (SATE), the average treatment effect in levels or fold change (FC), and the log fold change (LFC). Below, we define these estimands:

- ATE: $\tau_j^{\mathrm{ATE}} = \mathbb{E}[Y_j(1) - Y_j(0)]$.
- SATE: $\tau_j^{\mathrm{SATE}} = \mathbb{E}[Y_j(1) - Y_j(0)]/\sqrt{\mathrm{Var}(Y_j(0))}$.
- ATE in levels: $\tau_j^{\mathrm{FC}} = \mathbb{E}[Y_j(1) - Y_j(0)]/\mathbb{E}[Y_j(0)]$.
- LFC: $\tau_j^{\mathrm{LFC}} = \log(\mathbb{E}[Y_j(1)]/\mathbb{E}[Y_j(0)])$.

Here, we use Y_j to denote the random variable of the jth outcome and $(Y_j(0), Y_j(1))$ to denote its potential outcomes. Next, we present the corresponding influence functions under the identification assumptions, Assumptions 1–3. Before we present the influence functions, we introduce the uncentered influence function for $\mathbb{E}[Y_j(a)]$ and $\mathbb{E}[Y_j(0)^2]$:

$$\phi\_ja(O;\pi_a,\mu_{ja}) = \frac{\mathbb{1}\{A=a\}}{\pi_a(W)}(Y_j - \mu_{ja}(W)) + \mu_{ja}(W), \qquad a = 0,1$$

$$\phi_{j2}(O;\pi_0,\mu_{j2}) = \frac{\mathbb{1}\{A=0\}}{\pi_0(W)}(Y_j^2 - \mu_{j2}(W)) + \mu_{j2}(W),$$

where $\mu_{ja}(W) = \mathbb{E}[Y_j \mid W, A = a]$ for $a = 0,1$ and $\mu_{j2}(W) = \mathbb{E}[Y_j^2 \mid W, A = 0]$. Note that the (centered) influence function of $\mathbb{E}[Y(a)]$ is given by $\phi_{ja}(O;\pi_a,\mu_{ja}) - \mathbb{E}[Y_j(a)]$. It follows that

$$\eta_j^{\mathrm{ATE}}(O;\pi,\mu_j) = \phi_{j1} - \phi_{j0} - \tau_j^{\mathrm{ATE}}.$$

The efficient centered influence function of τ_j^{SATE} is given by

$$\eta_j^{\mathrm{SATE}}(O;\pi,\mu_j) = \frac{\phi_{j1} - \phi_{j0}}{\sqrt{\mathbb{V}[Y_j(0)]}} - \tau_j^{\mathrm{SATE}}\left[\frac{\phi_{j2} + \mathbb{E}[Y_j(0)^2] - 2\mathbb{E}[Y_j(0)]\phi_{j0}}{2\mathbb{V}[Y_j(0)]}\right].$$

See for example, Kennedy et al. (49, Equation (6)) and Du et al. (7, Equation (4.3)). Similarly, the efficient influence function of τ_j^{FC} is given by

$$\eta_j^{\mathrm{FC}}(O;\pi,\mu_j) = \frac{\phi_{j1} - \phi_{j0}}{\mathbb{E}[Y_j(0)]} - \frac{\tau_j^{\mathrm{FC}}\phi_{j0}}{\mathbb{E}[Y_j(0)]}$$

$$\eta_j^{\mathrm{LFC}}(O;\pi,\mu_j) = \frac{\phi_{j1}}{\mathbb{E}[Y_j(1)]} - \frac{\phi_{j0}}{\mathbb{E}[Y_j(0)]}.$$

In the current paper, we restrict our focus to LFC; however, our implementation also allows the computation and inference using other estimands listed above. When computing the LFCs, we use the size-normalized counts Y_{ij}/s_i adjusted by the size factors s_i in place of the raw count Y_{ij}. This is akin to taking a weighted average of the sample to estimate ATE (and, subsequently, LFC). Otherwise, the effect will be driven by cells with large library sizes.

CATE and VTE

Under standard identification assumptions of consistency, conditional exchangeability, and positivity as in Assumptions 1–3, the conditional average treatment effect (CATE) is identified by $\tau_j(w) = \mu_{j1}(w) - \mu_{j0}(w)$. This also applies to conditional log-fold change.

When one is only interested in the conditional effects in a subset of variable $\mathcal{S} \subset [d_W + a]$, the DR-learner readily accommodates runtime confounding through the decomposition $\tau_{\mathcal{S}}(w) = \mathbb{E}[\phi(O) \mid W_{\mathcal{S}} = w_{\mathcal{S}}]$. This decomposition implies that one may estimate $\tau_{\mathcal{S}}(w)$ by regressing $\phi(O)$ on $W_{\mathcal{S}}$, i.e. modifying the final regression step of the DR-learner.

Multiple testing adjustment by multiplier bootstrap

Algorithm S2 Multiple testing on standardized treatment effects

```
Input: The estimated influence function values $\widehat{\eta}_{ij}$, the estimated variance $\widehat{\sigma}_j^2$ for $i = 1,\dots,n$ and $j = 1,\dots,p$. The FDP exceedance threshold $c$, the FDP exceedance probability $\alpha$, and the number of bootstrap samples $B$. The threshold $\widetilde{c}$ to exclude genes with small variation.
Initialize the iteration number $\ell = 1$, the candidate set $\mathcal{A}_1 = \{j \in [p] \mid \widehat{\sigma}_j^2 \geq \widetilde{c}\}$, the set of discoveries $\mathcal{V}_1 = \varnothing$, and the maximal statistic of $M_1 = \max_{j\in\mathcal{A}_1} |t_j|$.
while not converge do
    Let $\boldsymbol{D}_{n\ell} = \mathrm{diag}((\widehat{\sigma}_j)_{j\in\mathcal{A}_\ell})$ be the diagonal matrix of the estimated standard deviations and $\widehat{\boldsymbol{\eta}}_{i\ell} = (\widehat{\eta}_{ij})_{j\in\mathcal{A}_\ell}$ be the vector of estimated influence function values at iteration $\ell$.
    Draw multiplier bootstrap samples $\boldsymbol{g}_\ell^{(b)} = (\sqrt{n}\boldsymbol{D}_{n\ell})^{-1}\sum_{i=1}^n \varepsilon_{i\ell}^{(b)} \widehat{\boldsymbol{\eta}}_{i\ell}$, where $\varepsilon_{i\ell}^{(b)}$'s are independent samples from $\mathcal{N}(0,1)$ for $i = 1,\dots,n$ and $b = 1,\dots,B$.
    Compute the maximal statistic $M_\ell = \max_{j\in\mathcal{A}_\ell} |t_j|$.
    Estimate the upper $\alpha$-quantile of $M_\ell$ under $H_0^{(\ell)} : \tau_j^* = 0,\ \forall\, j \in \mathcal{A}_\ell$ by
        $$\widehat{q}_\ell(\alpha) = \inf\left\{x \,\middle|\, \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\{\|\boldsymbol{g}_\ell^{(b)}\|_\infty \leq x\} \geq 1-\alpha\right\}$$
    Set $j_\ell = \mathrm{argmax}_{j\in\mathcal{A}_\ell} |t_j|$ and $\mathcal{A}_{\ell+1} = \mathcal{A}_\ell \setminus \{j_\ell\}$.
    if $M_\ell > \widehat{q}_\ell(\alpha)$ then
        Set $\mathcal{V}_{\ell+1} = \mathcal{V}_\ell \cup \{j_\ell\}$.
    else
        Declare the standardized treatment effects in $\mathcal{A}_\ell$ are not significant stop the step-down process.
    end if
    $\ell \leftarrow \ell + 1$.
end while
Augmentation: Set $\mathcal{V}$ to be the union of $\mathcal{V}_\ell$ and the $\lfloor |\mathcal{V}_\ell| \cdot c/(1-c) \rfloor$ elements from $\mathcal{A}_\ell$ with largest magnitudes of $t_j$.
Output: The set of discoveries $\mathcal{V}$.
```

Supplementary Note S3: Data simulation and analysis

Bulk expression simulation details

The bulk expression data are generated using a Poisson distribution with a zero-inflation component. The setup involves generating a latent signal matrix influenced by random noise and specific parameters. The data generation process is described in Algorithm S3 in detail. For experimental results in Fig. 2, we set $d = 2$ and $r^* = 1$, and vary $n \in \{100, 200, 300\}$. For causarray, RUV, and RUV-III-NB, we provide the number of latent factors in $r \in \{2, 4, 6\}$. Because the simulated data consists of 3 cell types, which may be explained with 3 additional degrees of freedom, the best possible choice of the number of latent factors would be $r = 4$.

Algorithm S3 Data generation process for pseudo-bulk gene expressions.

```
Input: Number of subjects $n$, number of covariates $d$, number of latent factors $r_0$, number of cells per subject $m = 10$, number of genes $p = 2000$, number of significant genes $s = 100$, and zero-inflation probability $\psi = 0.1$.
  (Signals) The $p$-dimensional signal is derived from multiplying the signal strength by a Beta distributed vector, modified by a random sign flip:
      $$\beta_j \sim 0.5 \times \mathrm{Beta}(1, 0.1) \times (2 \times \mathrm{Bernoulli}(0.5) - 1), \qquad j = 1, \ldots, s,$$
    and $\beta_j \equiv 0$ for $j = s + 1, \ldots, p$.
  (Cell types) The 3 cell types are generated with means $\{-0.5, 0, 0.5\}$ and standard deviations drawn from $\mathrm{Uniform}(0.5, 1)$. For $n$ subjects, the cell type assignment is randomly sampled from $\mathrm{Categorical}(3)$ and the cell-type specific means and scales are stored as $n$-dimensional vectors $\boldsymbol{\mu}_{\mathrm{ct}}$ and $\boldsymbol{\sigma}_{\mathrm{ct}}$.
  (Covariates) Sample $d$ observed covariates $\boldsymbol{W}_{\cdot j} \sim 0.5\boldsymbol{\sigma}_{\mathrm{ct}} \times \mathcal{N}_n(\boldsymbol{\mu}_{\mathrm{ct}}, \mathbf{1}_n)$ for $j = 1, \ldots, d$, and unobserved covariates $\boldsymbol{W}_{\cdot j} \sim 0.25\boldsymbol{\sigma}_{\mathrm{ct}} \times \mathcal{N}_n(\boldsymbol{\mu}_{\mathrm{ct}}, \mathbf{1}_n)$ for $j = d + 1, \ldots, d + r_0$.
  (Treatments) Sample $\boldsymbol{A} \sim \mathrm{Bernoulli}(\mathrm{Logistic}(\boldsymbol{W}\boldsymbol{\alpha}))$ where $\boldsymbol{\alpha} \sim \mathcal{N}_d(\mathbf{0}_{d+r_0}, (4(d + r_0))^{-1/2}\mathbf{1}_{d+r_0})$.
  (Coefficient matrix) Sample $b_{0j} \sim \mathrm{Beta}(2, 1)$ and $\boldsymbol{B}_{\cdot j} \sim \mathcal{N}_d(\mathbf{0}_{d+r_0}, (4(d + r_0))^{-1/2}\mathbf{1}_{d+r_0})$ for $j = 1, \ldots, p$.
  (Natural parameters) Let $\boldsymbol{\Theta} = \mathbf{1}\boldsymbol{b}_0^\top + \boldsymbol{W}\boldsymbol{B}^\top + \boldsymbol{A}\boldsymbol{\beta}^\top$.
  (Single-cell observations) Let $\boldsymbol{Y}^{\mathrm{sc}} \in \mathbb{R}^{n \times p \times m}$ with $\boldsymbol{Y}^{\mathrm{sc}}_{\cdot\cdot\ell} \sim \mathrm{Bernoulli}((1 - \psi) \times \mathbf{1}_{n \times p}) \times \mathrm{Poisson}(\exp(\boldsymbol{\Theta}))$ for $\ell = 1, \ldots, m$.
  (Bulk observations) Let $\boldsymbol{Y} \in \mathbb{R}^{n \times p}$ with $\boldsymbol{Y} = \sum_{\ell=1}^m \boldsymbol{Y}^{\mathrm{sc}}_{\cdot\cdot\ell}$.
Output: Covariates $\boldsymbol{W}$, treatment $\boldsymbol{A}$, single-cell gene expression $\boldsymbol{Y}^{\mathrm{sc}}$, and bulk gene expression $\boldsymbol{Y}$.
```

Single-cell expression simulation details

The single-cell expression data are generated by Splatter (25). Splatter explicitly models the hierarchical Gamma-Poisson processes that give rise to data observed in scRNA-seq experiments and can model the multiple-faceted variability. The data is generated from `splatSimulate` function from Splatter (1.26.0) package (25). When calling this function, the treatment effects are simulated with the parameters: `group.prob = c(0.5, 0.5), method = "groups", de.prob=0.05, de.facLoc=1., de.facScale=0.5, de.downProb=0.5`; the dropout effects are simulated with the parameters: `dropout.type="experiment", dropout.mid=20, dropout.shape=0.001`; the batch effects are simulated with the parameters: `batch.facLoc=noise, batch.facScale=0.5`; while all the other parameters are the same as returned by the function `newSplatParams`. For experimental results in Fig. S1, we generate $d = 1$ covariates and $r = 4$ unmeasured confounders. We first generate $(d + r + 1)/2$ batches with equal sample sizes, which account for $d + r$ degrees of freedom. To simulate varying confounding levels, we set `noise` in $\{0.1, 0.2, 0.3\}$.

Supplementary Note S4: Extra results

Simulation

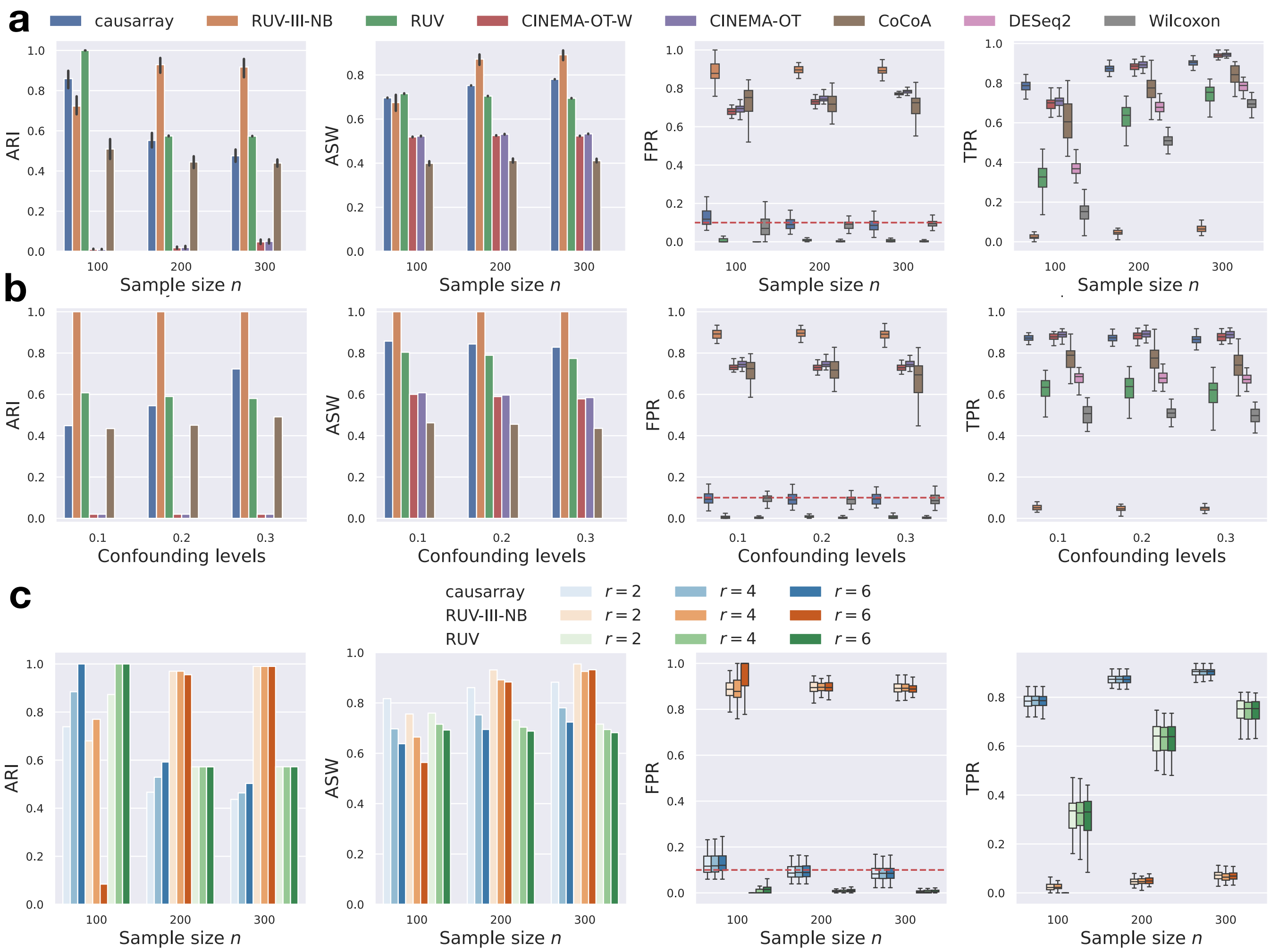

Fig. S1. Benchmarking of causarray against other methods for single-cell differential expression testing on synthetic single-cell expression data under unmeasured confounders. a, Bar plots and box plots of different validation metrics for causarray and other methods with $r = 4$ latent factors and a moderate confounding level. Bar plots (ARI, adjusted Rand index, and ASW, average silhouette width) indicate the median performance of confounder estimation. Box plots (FPR, false positive rate, and TPR, true positive rate) indicate the performance of biological signal preservation. The top and bottom hinges represent the top and bottom quartiles, and whiskers extend from the hinge to the largest or smallest value no further than 1.5 times the interquartile range from the hinge. The median is used as the center. **b,** Bar plots and box plots of different validation metrics for causarray and other methods with varying confounding effects. **c,** Bar plots and box plots of different validation metrics for RUV, RUV-III-NB, and causarray, with varying numbers of latent factors.

Perturb-seq data

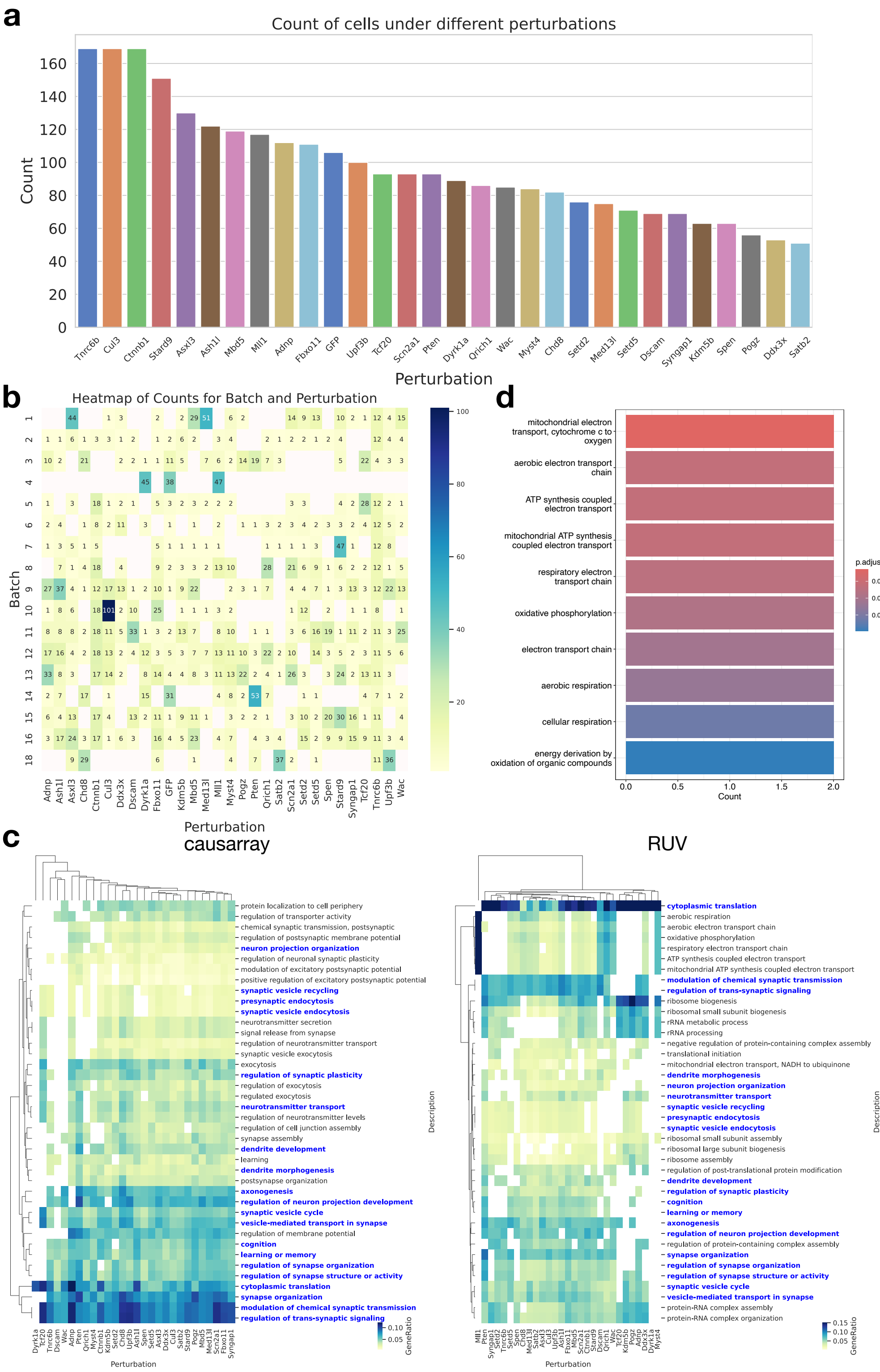

Fig. S2. Additional results on the Perturb-seq dataset. **a,** Barplot of the number of cells in each perturbation. **b,** Heatmap of the number of cells in each batch and perturbation. The batch design and the perturbation assignment of the Perturb-seq dataset are highly correlated. **c,** Clustermaps of GO terms enriched in discoveries (FDR< 0.1) from causarray and RUV, respectively, where the common GO terms are highlighted in blue. Only the top 40 GO terms that have the most occurrences in all perturbations are displayed. **d,** Barplot of GO terms enriched in discoveries under *Mll1* perturbation from RUV.

Alzheimer's data

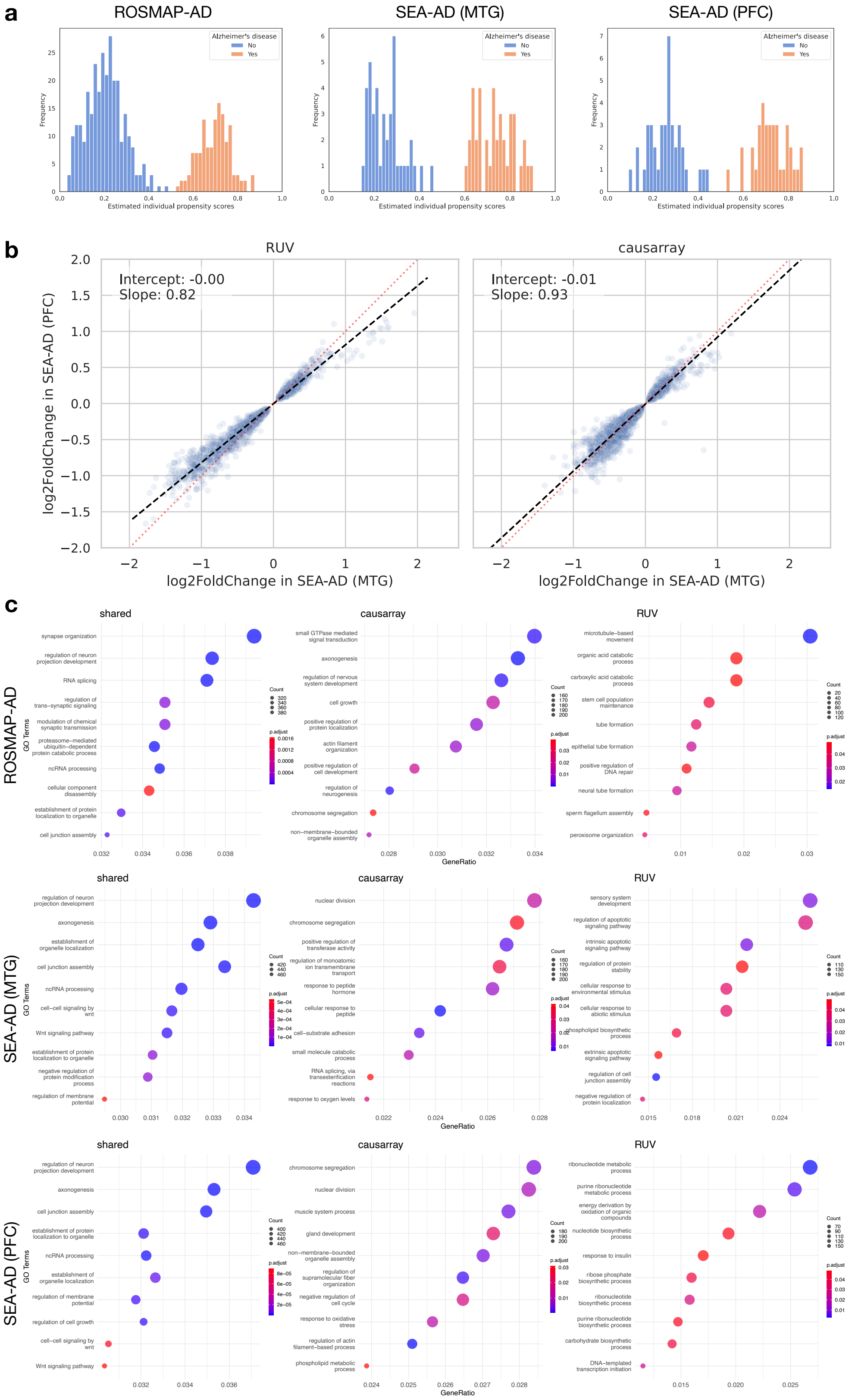

Fig. S3. Extra experimental results in AD datasets. a, Histogram of estimated propensity score in three AD datasets. **b,** Estimated effect sizes of DE genes (FDR < 0.001) in SEA-AD datasets. The black dashed line represents the fitted linear regression model, and the red dotted line represents the line $y = x$. **c,** Top gene ontology terms of the shared and distinct discoveries by causarray and RUV.