Park and Kellis *Genome Biology* (2021) 22:228
https://doi.org/10.1186/s13059-021-02438-4

Genome Biology

METHOD Open Access

CoCoA-diff: counterfactual inference for single-cell gene expression analysis

Yongjin P. Park[1,2*] and Manolis Kellis[3,4]

* Correspondence: ypp@stat.ubc.ca
[1]Department of Pathology and Laboratory Medicine, Department of Statistics, University of British Columbia, Vancouver, BC, Canada
[2]Department of Molecular Oncology, BC Cancer, Vancouver, BC, Canada
Full list of author information is available at the end of the article

Abstract

Finding a causal gene is a fundamental problem in genomic medicine. We present a causal inference framework, CoCoA-diff, that prioritizes disease genes by adjusting confounders without prior knowledge of control variables in single-cell RNA-seq data. We demonstrate that our method substantially improves statistical power in simulations and real-world data analysis of 70k brain cells collected for dissecting Alzheimer's disease. We identify 215 differentially regulated causal genes in various cell types, including highly relevant genes with a proper cell type context. Genes found in different types enrich distinctive pathways, implicating the importance of cell types in understanding multifaceted disease mechanisms.

Keywords: Causal inference, Single-cell RNA-seq, Counterfactual inference, Alzheimer's disease

Backgrounds

Single-cell RNA-seq is a scalable approach to measure thousands of gene expression values in hundreds of thousands of cells, sampled from a hundred individuals. As technology becomes mature and economical, single-cell sequencing methods have been used to solve a variety of biological and medical problems, and many large-scale data sets are becoming available to research communities. Unlike previous bulk RNA-seq, single-cell RNA-seq analysis quantifies gene expression changes from a large number of cells, and researchers dare to ask unprecedented questions, which had not been feasible in bulk data analysis. Only a subset of such examples includes cell-level developmental trajectory analysis [1], spatial transcriptomics [2], regulatory network reconstruction with perturbation [3], and variance quantitative trait analysis [4, 5].

Interestingly, some research questions hitherto remain fundamentally attractive since gene expression microarrays [6–8] and bulk RNA-seq [9–13] era. Differential expression analysis is such a classical problem. For case-control studies, knowing differentially expressed genes (DEGs) is often of research and clinical interest. Our primary interest also centres on developing a statistical method for differential expression analysis between different groups of individuals, not between cells. The underlying statistical problem is straightforward. However, finding DEGs from case-control single-cell

data poses several challenges in practice. This work seeks to identify and propose an algorithmic approach that resolves two of those challenges from a causal inference perspective.

Firstly, cells are not independently and identically distributed. Instead, cells belong to a particular individual, hierarchically organized, and naturally create "batch" effects (Fig. 1a). Cells belonging to the same individual are necessarily affected by the same biological and technical factors. The number of individuals essentially determines the statistical power of DEG discovery in single-cell data. Along the same line, a benchmark comparison demonstrates that existing bulk RNA-seq methods on pseudo-bulk data (using the individual-level aggregate of cells of a particular cell type) still perform decently while correctly controlling false discovery rates [14, 15]. Likewise, for genetic analysis (expression quantitative trait loci), the statistical power of eQTL discovery is primarily determined by the degree of genetic variation across individuals rather than the number of cells per individual [16]. Nonetheless, differential expression analysis of single-cell RNA-seq is a state-of-the-art and unbiased approach to characterize cell-type-specific transcriptomic changes.

Another challenge stems from the study design of case-control data analysis. In contrast to randomized control trials, most studies are observational, and we have incomplete knowledge of a disease assignment mechanism. Investigators usually cannot make an intervention for practical and ethical reasons. Considering that many complex disease phenotypes occur at the late onset of a lifetime, finding a suitable set of covariates for causal inference is often infeasible as well. Matrix factorization or latent variable modelling can be used to characterize technical covariates or batch effects. However, it is difficult to identify which principal axes of variation capture confounding effects, independently from unknown disease-causing mechanisms. A latent variable model of a single-cell count matrix is frequently used for clustering and cell type annotations, and the resulting latent factors are more suitable for the characterization of intercellular heterogeneity than inter-individual variability.

We present a novel application of a causal inference method as a straightforward approach to improve the statistical power in case-control single-cell analysis while adjusting for unwanted confounding effects existing across heterogeneous individuals. We establish our causal claims in differential expression analysis based on Rubin's potential outcome framework [17, 18]. Our method is inspired by the seminary work of outcome regression analysis by a matching algorithm [19, 20]. We highlight that our causal inference approach is beneficial in the analysis of disease case-control studies, especially when meta-data for covariates are scarcely available, and covariates may influence both disease status and gene expressions simultaneously. With respect to the underlying causal structural model (disease to gene expression), we seek to identify genes that are differentially expressed as a result of disease.

Results

Overview of our causal inference approach

Definition of causal genes

Here, we ask whether a gene is causally affecting or affected by a disease variable but not affected by other technical and biological covariates, which may confound the

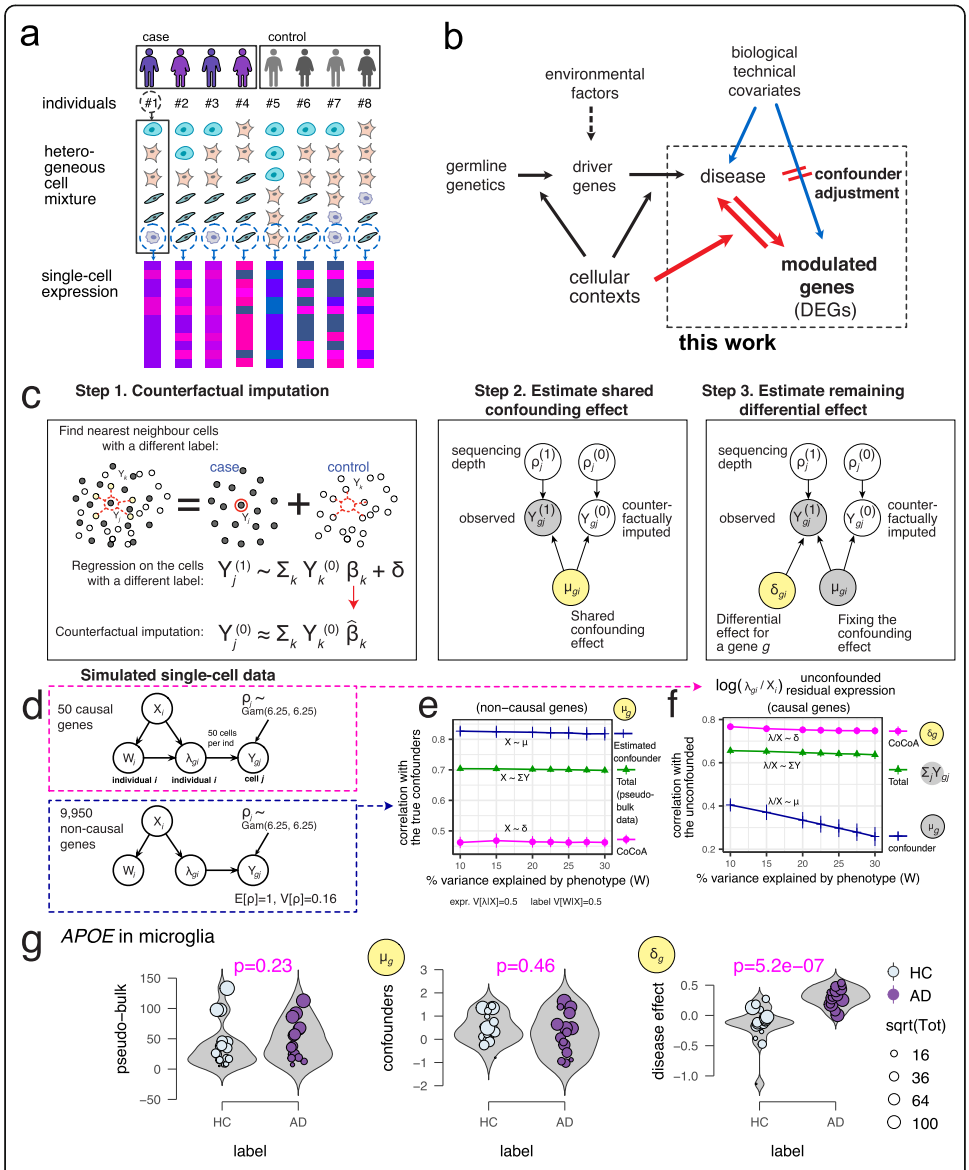

Fig. 1 Counterfactual confounder adjustment for single-cell differential gene expression analysis (CoCoA-diff). **a** Hierarchical (nested) structure of single-cell gene expression data. We have tens of individuals for one case-control study. Each individual (*i*) contains a heterogeneous mixture of multiple cell types. Single-cell technology measure a thousand genes on each cell (*j*). **b** This work seeks to address a specific causal inference problem of genomics research. We seek to prioritize genes causally modulated by a disease status, not the genes affecting the predisposition and risk of disease development. **c** Overview of CoCoA-diff approach (see Methods for details). *Y*: gene expression matrix. $Y^{(0)}$: counterfactual data with disease $W = 0$, $Y^{(1)}$: counterfactual data with disease $W = 1$. β: Poisson regression coefficient. δ: residual effect. ρ: sequencing depth. μ: shared confounding effect. **d** Data generation scheme for simulation studies. We simulate 50 causal and 9,950 non-causal genes with or without disease-causing mechanisms (an edge between *W* and λ). W_i: disease label assignment for an individual *i*. X_i: confounding effects for an individual *i*. λ_{gi}: unobserved gene expression for a gene *g* of an individual *i* as a function of *X* and *W*. Y_{gj}: realization of cell-level gene expression of a gene *g* with a cell *j*-specific sequencing depth ρ_j (stochastically sampled from Gamma distribution). Here, we simulated five different *X* variables. **e** CoCoA-diff accurately estimates shared confounder variables (μ_g), showing a significantly higher level of correlation with true confounding effects on non-causal genes than a pseudo-bulk analysis. **f** CoCoA-diff accurately estimates disease-causing effects (δ_g), showing high correlation with true differential effects on causal genes. **g** Illustration of CoCoA-diff approach on the *APOE* in microglia example. HC, health control. AD, Alzheimer's disease. μ, shared confounding effect; δ, residual differential effect. For a clear visualization, we omitted samples (individuals) with zero read count observed on *APOE* gene in the microglial cell type

disease status and gene expressions. In this work, a causal gene is defined as a gene affecting or being affected by a disease status independent of other confounding variables. Although many differentially expressed genes can be considered a result of disease status for most late-onset disorders, we also acknowledge that aberrant changes on a handful of genes can initiate disease phenotypes. To distinguish causal vs. anti-causal mechanisms, we would need additional perturbation experiments. Alternatively, driver genes can be characterized by mediation analysis using genetic variants as an instrumental variable (Mendelian randomization) [21].

Moreover, concerning cell types and states, we need to assume that cell type fractions are not a mediating factor between the disease and gene expression variables. We found a negligible correlation between cell-type proportions and observed disease status in the study of Alzheimer's disease [22]. Under this causal assumption, the stratification procedure for cell types provides a legitimate strategy to control cell-type biases that may impact on identifying DEGs. We think there is almost no chance of a "mediation fallacy [23–25]."

Differential analysis on pseudo-bulk expression profiles

We are interested in comparing pseudo-bulk gene expression profiles stratified within each cell type and individual between the case and control samples. Letting Y_{gj} be a gene expression of a gene g on a cell j and S_i be a set of cell indexes for an individual $i \in [n]$, we can create a pseudo-bulk expression by aggregating all the expression vectors. We will use λ_{gi} to generally refer to a pseudo-bulk estimate of a gene g on an individual i. For instance, we could take an average, $\lambda_{gi} \approx \sum_j |I\{j \in S_i\} Y_{gj} / \mid S_i \mid$, or take the total count, $\lambda_{gi} \approx \sum_j |I_{j \in S_i} Y_{gj}$. Given the estimate of the λ values across n individuals, $\{\lambda_{gi} : i \in [n]\}$, we can construct a hypothesis test that seeks to reject a null hypothesis that the distributions of pseudo-bulk profiles are the same among the case and control individuals (Wilcoxon's test).

Potential outcome framework for single-cell differential expression analysis

In observational data, where the label assignment is not controlled, data matrices of raw $\{Y_{gj}\}$ and pseudo-bulk count $\{\lambda_{gi}\}$ can become confounded with the disease label assignment by unknown biological and technical covariates (Fig. 1b). Such confounding factors obfuscate actual disease-specific effects with other effects of unknown covariates and may lead to false discoveries and dampen the statistical power of differential expression analysis. Rubin's potential outcome framework [17, 18] seeks to separate the actual disease (or treatment) effects from other effects by asking the following counterfactual questions:

- What would be a gene expression if an individual had not been exposed to a disease?
- What would be a gene expression if an individual had been exposed a disease?

In our pseudo-bulk analysis context, we are interested in estimating the following quantities:

- $\lambda_{gi}^{(0)}$: What would be the pseudo-bulk expression of a gene g if an individual i had not been exposed to a disease?
- $\lambda_{gi}^{(1)}$: What would be the pseudo-bulk expression of a gene g if an individual i had been exposed to a disease?

In a binary case-control study, we observe one of the values for each individual while the other side is left unobserved (denoted by the "?" mark). Letting $W_i \in \{0, 1\}$ be a disease label assignment variable for an individual i, only a part of potential gene expressions are made directly observable from data:

$$\lambda_{gi}^{(0)} = \begin{cases} \lambda_{gi}, & W_i = 0 \\ ?, & W_i = 1 \end{cases}, \quad \lambda_{gi}^{(1)} = \begin{cases} \lambda_{gi}, & W_i = 1 \\ ?, & W_i = 0 \end{cases}$$

At the cell level ($\forall j \in S_i$), we have the same structure:

$$Y_{gj}^{(0)} = \begin{cases} Y_{gj}, & W_i = 0, j \in S_i \\ ?, & \text{otherwise} \end{cases}, \quad Y_{gi}^{(1)} = \begin{cases} Y_{gi}, & W_i = 1, j \in S_i \\ ?, & \text{otherwise} \end{cases}$$

If both sides of the potential expression, $\{Y_{gj}^{(0)}, Y_{gj}^{(1)}\}$, were known, we would be able to estimate the disease effect on a gene g for each individual by comparing pseudo-bulk profiles ($\lambda_{gi}^{(0)}$ vs. $\lambda_{gi}^{(1)}$) constructed from the potential single-cell gene expressions. The ultimate goal of causal inference in Rubin's potential outcome framework is to impute the missing part of potential outcomes since a comparison between the case and control becomes straightforward on the imputed data.

The definition of a confounding variable and causal assumptions

We define that a variable can confound a disease label (W) and gene expressions (λ and Y) if (1) it is associated with the disease and gene expression variables and (2) it is still associated with the expression even after conditioning on the disease label [26]. Unless we adjust/stratify a sufficient set of confounding variables, gene expression changes observed between the case and control samples are not necessarily the causal effect of disease mechanisms.

It is crucial to state casual assumptions to proceed with our causal inference:

- *Stable individual disease effect*: We assume that the potential expressions of an individual i, namely $\lambda_{gi}^{(0)}$ and $\lambda_{gi}^{(1)}$, are not affected by the expressions of other individuals $\{i' \in [n] : i' \neq i\}$.
- *Conditional independence of the potential expression and disease exposure* (conditional ignorability [17, 18]): For a non-causal gene, by definition, gene expressions are independent of disease status. Therefore, we do not need an assumption on this matter. However, for a causal gene, we assume that potential (counterfactual) gene expressions are independent of a disease label conditioning on a sufficient set of confounding variables. In other words, genes differentially regulated for a diseased individual would not have been aberrantly expressed if this individual had not developed the disease.
- *Overlap of confounding effects between the case and control*: Within a stratum of individuals, homogeneous with respect to confounding variables, we have both the

case and control subjects with non-zero probability. In the single-cell analysis, we assume disease and non-disease cells simultaneously exist in a homogeneous group of cells stratified by confounding factors.

CoCoA-diff for single-cell differential expression analysis

The purpose of our counterfactual confounder adjustment for differential single-cell gene expression analysis (CoCoA-diff) (Fig. 1c) is to impute the missing part of potential outcomes of single-cell profiles (step 1), propagate the imputed results to the pseudo-bulk estimation, and decompose the total pseudo-bulk profiles into the confounding (step 2) and differential effects (step 3). Using a single-cell gene expression matrix, $\{Y_{gj} : g \in \text{genes}, j \in \text{cells}\}$, we want to estimate two types of pseudo-bulk data: (1) the estimated confounders, $\{\mu_{gi} : g \in \text{genes}, i \in \text{individuals}\}$, and (2) the residual differential effects, $\{\delta_{gi} : g \in \text{genes}, i \in \text{individuals}\}$. In other words, we want to estimate the decomposition of pseudo-bulk data, such as $\lambda_{gi} = \mu_{gi}\delta_{gi}$. Briefly, the algorithm proceeds as follows: (1) we seek to estimate (or impute) counterfactual measurement of single cells' expression by matching cells in a particular condition with neighbouring cells in the opposite conditions. The distance between cells was calculated on the top principal component axes. (2) Having paired sets of observed and counterfactual single-cell data, we estimate the mean expression of genes shared across two opposite conditions in Bayesian posterior inference. We treat them as putative confounding factors. (3) While holding the estimated confounding effects fixed, we measure the conditional (or residual) mean effect on the observed cells. We refer the readers to Materials and Methods for technical details.

To demonstrate how CoCoA-diff actually works, we simulated a single-cell data matrix consisting of 10,000 genes and 40 individuals (Fig. 1d). Each individual contains 50 cells on average; 50 of the 10,000 genes are causally affected by disease labels ($W \rightarrow \lambda$) and confounding factors ($X \rightarrow \lambda$). The other genes are only affected by confounding factors ($X \rightarrow \lambda$); we introduced five confounding variables $\{X_{ik} : i \in \text{individuals}, k \in [5]\}$, and a linear combination of these X variables introduces biases on W_i and λ_{gi}. Here, we set the variance of λ explained by confounding variables X to 0.5 and the variance of disease label W explained by the same confounding variables to 0.5, but we varied the true disease variability between 0.1 to 0.3 on 50 causal genes ($W \rightarrow \lambda$; the x-axes of Fig. 1e, f). As expected, on non-causal genes (Fig. 1e), we found a strong correlation between the estimated confounding effects (μ_{gi}) and true confounding effects (the linear combination of X_{ik} variables), which is far greater than the correlation with the estimated differential effects (δ_{gi}). We also observed that the unconfounded pseudo-bulk data (removing the effects of X variables) are correlated with the estimated differential effects (δ_{gi}), consistently not affected by the change of disease variability (Fig. 1f).

As an example, we demonstrate the effectiveness of our approach in the case of *APOE* gene measured in microglia samples (Fig. 1 g). For better visualization, we removed an individual with only a single read was observed on the *APOE* gene. In pseudo-bulk data analysis with 39 individuals, it appears that the total expression values are only mildly upregulated in disease subjects (Wilcoxon p = 0.23) even though *APOE* over-expression is one of the most frequently observed hallmark of Alzheimer's disease (AD). We also found that the shared confounding factors across the case and control

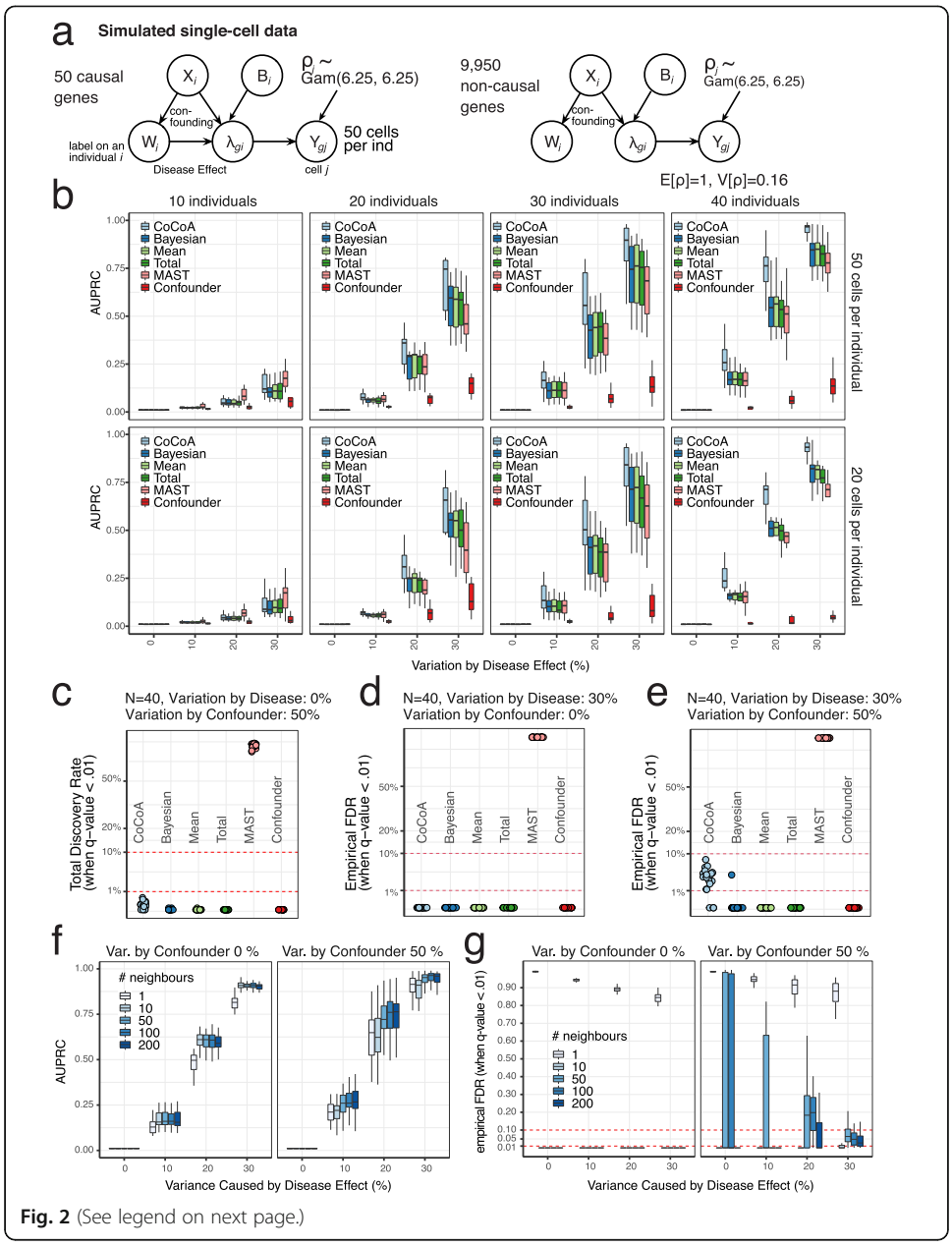

Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Simulation experiments. Extensive simulation experiments confirm that CoCoA-diff effectively adjusts existing confounding effects and improves statistical power of differential expression analysis. **a** Data generation scheme for simulation experiments. We simulate 50 causal and 9950 non-causal genes with or without disease-causing mechanisms (an edge between W and λ). W_i: disease label assignment for an individual i. X_i: confounding effects for an individual i. λ_{gi}: unobserved gene expression for a gene g of an individual i as a function of X and W. Y_{gj}: realization of cell-level gene expression of a gene g with a cell j-specific sequencing depth ρ_j (stochastically sampled from Gamma distribution). Here, we simulated total five covariates consisting of confounding (X) and batch effect variables (B). **b** Simulation results when all the five covariates are confounding disease label assignment and gene expression values, accounting for 50% of mean expression variation ($\sigma^2_{X,B\to Y}$). Different subpanels correspond to different configurations of the number of individuals and cells per individual. *Y-axis* (AUPRC): area under precision recall curve (numerically integrated by DescTool [28] implemented in R); *x-axis*: the proportion of variation contributed by the disease label ($\sigma^2_{W\to Y}$). The following methods were considered: *CoCoA*: Wilcoxon's ranksum test using individual-specific confounder-adjusted gene expression values δ_{gi} (the step 3 of Fig. 1c); *Total*: pseudo-bulk expression aggregated within each individual; *Bayesian*: Bayesian estimate of pseudo-bulk expression averaged over cells within each individual; *Mean*: pseudo-bulk expression averaged over cells within each individual; *MAST*: Model-based Analysis of Single-cell Transcriptomics [29] implemented in R (cell-level differential expression analysis); *Confoudner*: the estimated confounding effect μ_{gi} (the step 2 of Fig. 1c). **c** Total discovery rates of the differential expression methods when there were no disease effect. The fraction of positive discovery when multiple hypothesis-adjusted q-values were empirically calibrated by qvalue [30, 31] package controlled at 1% (*y*-axis). **d** Empirical false discovery rates of the differential expression methods when there were no confounding effect, but the 30% of individual-level expression variation is attributed to the disease effect ($W\to\lambda$; $\sigma^2_{W\to Y}$) on 50 causal genes. *Y-axis*: empirical false discovery rate, the frequency of the non-causal among genes with the estimated q-value below 0.01. **e** Empirical false discovery rates of the differential expression methods when there were substantial confounding effects on gene expressions ($\sigma^2_{X,B\to Y}$) and the 30% of individual-level expression variation is attributed to the disease effect ($W\to\lambda$; $\sigma^2_{W\to Y}$) on 50 causal genes. *Y-axis*: empirical false discovery rate (the frequency of the non-causal among genes with the estimated q-value below 0.01); *x-axis*: different methods. **f** The performance of the CoCoA method with different settings of the k-NN parameters in the first matching step. *Y-axis* (AUPRC): area under precision recall curve (numerically integrated by DescTool [28] implemented in R); *x-axis*: the proportion of variation contributed by the disease label ($\sigma^2_{W\to Y}$). Variation by confounder: $\sigma^2_{X,B\to Y}$. **g** Empirical false discovery rates for the same experiments in **f** with different settings of the k-NN Parameter. Empirical false discovery rate: the frequency of the non-causal among genes with the estimated q value below 0.01

exhibit almost no apparent correlation with the disease label (p = 0.46). After adjusting the confounders on the data, we recover a significant correlation of *APOE* gene expressions with AD status (p = 5.2×10^{-7}).

Simulation experiments

The design of simulation experiments

We evaluated the performance of our approach in differential expression analysis using a simulated single-cell data matrix (Y) of 10,000 genes with 40 individuals with 50 causal and 9,950 non-causal genes (Fig. 2a). We introduced two types of total five covariates on the individual-level expressions λ_{gi}, explicitly designating confounding variables X and non-confounding batch effect B. By definition, confounding factors X_{ik} affect the label assignment W_i and the individual-level mean values λ_{gi}. For a causal gene g, λ_{gi} values are determined by the disease label W_i, the confounders X_{ik}, and the batch effect variables B_{il}; for a non-causal gene g, there is no contribution from the disease variable. In each simulation experiment, we specify the following parameters:

- $\sigma^2_{X\to W}$: the variance of W explained by the covariate X.
- $\sigma^2_{W\to Y}$: the variance of $\log\lambda$ explained by the disease assignment W.

- $\sigma^2_{X,B\to Y}$: the variance of logλ explained by the covariates X and B.
- d_C: the number of confounding variables (from 1 to 5).
- d_B: The number of non-confounding batch effects (from 0 to 4).

For each individual $i \in [n]$, we first sample covariates $X_{ik} \sim \mathcal{N}(0,1)$ for $k \in [d_C]$ and $B_{il} \sim \mathcal{N}(0,1)$ for $l \in [d_B]$. Given the X matrix, we sample the parameter vector $\boldsymbol{\alpha}$ required to introduce biases on W and the residual error vector $\boldsymbol{\epsilon}_W$ from isotropic Gaussian distributions and adjusted the scale of the error vector to have the simulation proportion of variance matched with the prescribed σ^2 value, i.e. $\mathbb{V}[X\alpha]/\mathbb{V}[X\alpha + \varepsilon_W] = \sigma^2_{X\to W}$. We generate a binary label assignment for an individual i by flipping a coin:

$$W_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp\left(-\sum_k X_{ik}\alpha_k - \epsilon_{W,i}\right)}\right)$$

Combining these values, we construct the mean values of a gene expression g for an individual i by a generalized linear model:

$$\ln\lambda_{gi} = \underbrace{\tau_g W_i}_{\text{disease effect}} + \underbrace{\sum_{k=1}^{d_C} X_{ik}\beta_{kg}}_{\text{confounding effect}} + \underbrace{\sum_{l=1}^{d_B} B_{il}\gamma_{lg}}_{\text{batch effect}} + \epsilon_\lambda,$$

where the covariate effect $\beta_{kg} \sim \mathcal{N}\left(0, \sigma^2_{X,B\to Y}/(d_C+d_B)\right), \gamma_{lg} \sim \mathcal{N}\left(0, \sigma^2_{X,B\to Y}/(d_C+d_B)\right)$, gene-level causal effect $\tau_g \sim \mathcal{N}\left(0, \sigma^2_{W\to Y}\right)$, and the residuals $\epsilon_\lambda \sim \mathcal{N}\left(0, 1-\sigma^2_{X\to Y}-\sigma^2_{W\to Y}\right)$. Using the individual-level mean values λ, we stochastically generated cell-level expressions by multiplying the individual-level average expressions with random sequencing depth (ρ), sampled from $\rho \sim \text{Gamma}(6.25, 6.25)$. For each cell j,

$$Y_{gj} \sim \text{Poisson}\left(\lambda_{gi}\rho_j\right).$$

Once we have estimated different types of pseudo-bulk data, we ranked genes based on Wilcoxon's rank-sum test [27] implemented in R and constructed receiver-operating and precision-recall curves to calculate the power and AUPRC using DescTool [28] implemented in R.

Here, we show and compare the performance of differential analysis conducted on the five different pseudo-bulk data:

- CoCoA: individual-level disease effects δ_{gi} estimated by CoCoA-diff algorithm (Fig. 1c, step 3).
- Confounder: the confounder effects μ_{gi} estimated by CoCoA-diff algorithm (Fig. 1c, step 2)
- Bayesian: Bayesian estimate of pseudo-bulk expression averaged over cells within each individual ($\mu_{gi}\delta_{gi}$ combined, not decomposed).
- Mean: arithmetic mean of cell-level expressions within each individual ($|S_i|^{-1}\sum_{j\in S_i} Y_{gj}$).
- Total: summation of cell-level expressions within each individual $\left(\sum_{j\in S_i} Y_{gj}\right)$.

In addition, we considered a cell-level differential expression method although such cell-level model estimation/hypothesis test violates exchangeability assumptions across different individuals (Fig. 1a).

- MAST: Model-based Analysis of Single-cell Transcriptomics [29] implemented in R

Counterfactual adjustment of pseudo-bulk data improves statistical power

We repeated our experiments 20 times for all the different configurations and summarized the performance by the area under the precision-recall curve (AUPRC), varying the gene expression variance caused by disease ($0 \le \sigma^2_{W \to Y} \le 0.3$) and the number of individuals (from 10 to 40), also considering a different number of cells per individual (20 and 50). Since our method primarily focuses on adjusting confounding factors (X variables in Fig. 2a), we highlight the results, where all the five covariates act as a confounder (Fig. 2b). However, we generally reached qualitatively a similar conclusion in further experiments, where batch effect variables (B variables in Fig. 2a) coexist with confounding effects (see Fig. S2 for the details). The performance gap between the CoCoA-diff and other pseudo-bulk analysis methods persists in almost all cases, regardless of the sample size of individuals and cells. As expected, causal genes are located at the bottom of the list ranked by confounding effects, yielding poor AUPRC scores. The cell-level DEG analysis (MAST) performed better than other pseudo-bulk methods only if the number of individuals is few (N = 10). Considering that model fitting based on cell-level data generally demands higher computational costs, individual-level pseudo-bulk analysis is better suited for DEG analysis if the data come with sufficient sample size (individuals) and case-control labels were assigned at an individual level.

As demonstrated by previous analysis [14], we also confirmed that pseudo-bulk methods effectively control type I errors (Fig. 2c–e). However, a cell-level test often produces an inflated p-value histogram; thus, a subsequent empirical FDR estimation method, such as qvalue [30, 31], ends up drawing a decision boundary at a wrong p value cutoff. Even when we included no causal effect, the cell-level method (MAST) predicted that a high fraction of genes are differentially expressed, whereas the other pseudo-bulk-based methods, including CoCoA-diff, made almost no discoveries (Fig. 2c). As long as we keep the contributions from confounding effects low, all the pseudo-bulk methods conservatively (and correctly) control type I errors (Fig. 2d). We found that CoCoA-diff might loosely control type I errors, higher than desired by an empirical false discovery rate (eFDR) calibration method when the simulated data were contaminated by confounding factors (Fig. 2e). We define eFDR as the fraction of positive discovery when multiple hypothesis-adjusted q values were empirically calibrated by qvalue [30, 31] package controlled at 1%.

In some sense, the loosened type I error control can arise due to the suboptimal choice of the k-Nearest Neighbour parameter in the cell-cell matching step. We evaluated the performance of the CoCoA-diff methods with different settings of the k-NN parameters (Fig. 2f) and calculated eFDR (Fig. 2g). When confounding effects are absent ($\sigma^2_{X,B \to Y} = 0$), the k parameter does not affect the AUPRC performance and eFDR for $k > 1$. However, a sufficient number of k-nearest neighbours are required for large confounding effects ($\sigma^2_{X,B \to Y} = 0.5$); in our experiments, the AUPRC scores

saturated after $k \geq 50$ (Fig. 2f) and the eFDR levels decreased when a larger k-nearest neighbours were used to control confounding effects.

We further conducted simulation experiments, where the conditional ignorability [17, 18] assumption no longer holds due to feedback loops on causal genes and leaking causal effects (see Fig. S1a). In particular, we infused the first principal component (PC) of disease effects on causal genes and reintroduced collider effects shared between the causal and non-causal genes. Unfortunately, unlike confounding variables, adjusting a collider between multiple variables creates spurious associations between them [32]. Moreover, having such a collider variable creates substantial challenges in eFDR calibrations for all the methods (Fig. S1b). However, in terms of gene ranking tasks, CoCoA-diff-adjusted pseudo-bulk analysis still outperforms other methods, consistently across many different settings (Fig. S2c).

Case study: finding cell-type-specific causal genes in Alzheimer's disease

We reanalyzed published single-nuclei RNA-seq (snRNA-seq) data of 48 individuals in postmortem brain samples [22]. To our knowledge, this is one of the largest snRNA-seq data on case-control disease studies. Of the 48 individuals, we included 40 individuals for differential expression analysis because we found no case-control disease labels on the reaming eight individuals.

Cell type annotations of 70,634 cells

We directly annotated the cell types of these 70,634 cells using the list of cell-type-specific genes provided by the PsychENCODE project [33]. Of the total 2648 PsychENCODE marker genes, we used 1726 genes expressed in this data set as features (Fig. S3). We identified the eight cell-type clusters of cells while estimating a mixture of von Mises-Fisher distributions, measuring cells' likelihood to centroids by angular distance (see the "Methods" section). We found that this gene-to-cell-type membership information was sufficient enough to distinguish eight cell types. These eight cell types include expiatory (Ex) and inhibitory neurons (In), oligodendrocytes (Oligo), oligodendrocyte progenitor cells (OPC), microglia, astrocytes (Astro), pericytes (Per), and endothelial cells (Endo). We found that our annotation almost perfectly agrees with the original paper's cell type annotation (Fig. S4). We also found that cell types correspond to unique cell clusters after BBKNN (batch-balancing k-Nearest Neighbour) [34] pre-processing (Fig. 3a), showing no apparent bias induced by other demographic and pathological variables (Fig. S5). Moreover, we further dissected four different cortical layer-specific cell types for excitatory neurons (Fig. 3b) and four different subtypes for inhibitory neurons (Fig. 3c) using a refined set of marker genes provided by previous single-nucleus analysis [35].

We notice a wide spectrum of cell-type variability across 48 individuals, both in terms of the number of cells and proportions (Fig. 3d). The Mathys et al. data contain on average 1444 cells per individual, of which 50.65% cells stem from Ex (N = 726 ± 382 SD), 12.71% cells from In (N = 182 ± 107), 25.72% cells for Oligo (N = 380 ± 252), 3.56% cells from OPC (N = 54 ± 34), 2.43% cells from Microglia (N = 33 ± 24), 4.94% cells from Astro (N = 69 ± 46), and 0.1% cells from Endo and Per. We have 726 excitatory neurons per individual (Fig. 3e), of which 52.62% (N = 399 ± 250) cells from the

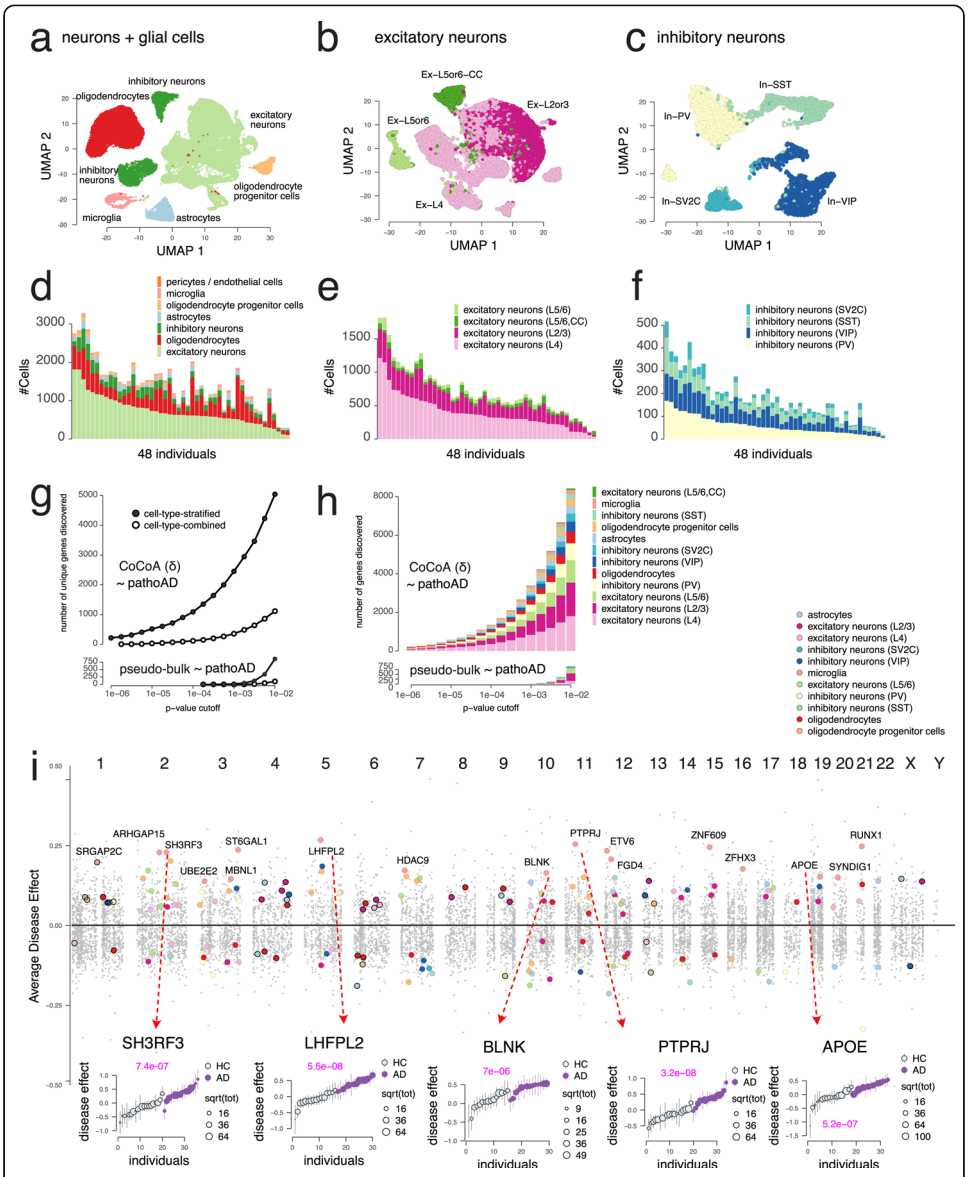

Fig. 3 The case study of Alzheimer's disease snRNA-seq profiles [22]. Cell-type-stratified approach improves statistical power and interpretation of differential expression analysis. **a** UMAP projection of the major cell types. **b** UMAP projection of the excitatory neuron subtypes. **c** UMAP projection of the inhibitory neuron subtypes. **d** Cell type decomposition of the major cell types across 48 individuals. **e** Cell type decomposition of the excitatory neurons across 48 individuals. **f** Cell type decomposition of the inhibitory neurons across 48 individuals. **g**, **h** Cell-type-stratified CoCoA-diff approach improves statistical power and identifies genes in a wide spectrum of the cell types. **i** A genomic view of genes strongly modulated by AD pathology (coloured). *Top panel*: genomic location of cell-type-specific causal genes; *bottom panels*: five representative examples of the most significant genes

layer 4, 33.1% (N = 233 ± 108) cells from the layer L2/3, 8.26% (N = 50 ± 28) cells from the cortical layer L5/6 (CC), and 6.03% (N = 44 ± 31) cells from the layer 5/6. We have 182 inhibitory neurons per individual (Fig. 3c), consisting of 32.9% (N = 59 ± 34) cells from inhibitory neurons (VIP), 31.72% (N = 59 ± 40) cells from inhibitory neurons (PV), 24.51% (N = 43 ± 26) cells from inhibitory neurons (SST), and 11.1% (N = 22 ± 17) cells from inhibitory neurons (SV2C).

Cell type stratification improves the statistical power and interpretation of differential expression analysis

We investigated the impact of such a high level of cell-type heterogeneity on subsequent differential expression analysis. Tissue-level bulk RNA-seq analysis data can be arguably thought of as an aggregate of single-cell-level expressions. If genes were similarly affected by the disease phenotype in most cell types, we would expect the bulk-level associations to be similar, and cell-type-stratified analysis would benefit less than more of stochasticity—fewer cells per individual. On the other hand, if most disease-responsive genes act through a cell-type-specific mechanism, cell-type-stratified data analysis will improve statistical power and render better biological interpretations in genomics analysis.

Using these cell type annotations, we constructed cell-type-stratified pseudo-bulk data for all the genes and individuals in each cell type independently, treated them as a gene expression matrix, and tested associations of genes with AD status. We also constructed the pseudo-bulk profiles by combining all the cells in each individual, ignoring the cell type annotations, and carried out the same association analysis. It is clearly shown that the number of discoveries (unique genes) dramatically increase with cell-type-specific stratification steps in both studies using CoCoA-diff and total expression profiles (Fig. 3g). Considering the variety of cell types in each p value cutoff, such cell-type-specific mechanisms are likely to remain hidden in bulk, combined differential analysis but better revealed after taking into account cell type heterogeneity (Fig. 3h).

Disease status modulates the cell-type-specific gene expressions

215 genes are differentially regulated with AD pathology

We prioritized genes based on testing a hypothesis that the pseudo-bulk profiles processed by CoCoA-diff are differentially ranked by AD pathology (Wilcoxon's ranksum test) [27]. We conservatively adjusted putative confounding effects with (100-nearest neighbour search) in a spectral space constructed by 50 principal components. Controlling the false discovery rate (FDR [36]) < 1%, we found 1648 genes (11.68% of 14,106), consisting of 672 genes found in Ex-L4, 522 in Ex-L2/3, 297 in Ex-L5/6, 210 in In-PV, 98 in Oligo, 84 in Microglia, 80 in Astro, 57 in In-VIP, 49 in OPC, 11 in In-SST, and 4 in In-SV2C. Controlling family-wise error rate (FWER [37]) at 1%, we found a total of 215 genes (1.52%), which consist of 55 genes found in Ex-L4, 39 in Ex-L2/3, 28 in Ex-L5/6, 28 in Oligo 24 in In-PV, 19 in Microglia, 19 in Astro, 9 in OPC, 7 in In-VIP, and 3 in In-SST.

We confirmed that the CoCoA-diff procedure did not introduce a systematic bias by shrinking variance on the case or control samples (Fig. S8). We tested our method on four different phenotypes using twelve cell-type-specifically confounder-adjusted profiles and cell-type-sorted pseudo-bulk data. Moreover, visual inspection of the p value distributions for Wilcoxon's tests suggests no apparent inflation/deflation in our multiple hypothesis testing (Fig. S9).

In addition to the non-parametric ranksum test, we propose a model-based Wald statistic for an individual-level test (for each gene g and an individual i), namely $Z_{gi} = \mathbb{E}[\ln \delta_{gi}]/\mathbb{V}[\ln \delta_{gi}]$, and the group-wise average disease effect size (ADE) and standard error (SE) for each gene g:

$$\mathrm{ADE}_g = \frac{\sum_{i=1}^{n} \mathbb{E}[\ln \delta_{gi}]/\omega_i}{\sum_i 1/\omega_i}, \mathrm{SE}_g = \sqrt{\frac{1}{\sum_i 1/\omega_i}},$$

where $\omega_i = 1/\mathbb{V}[\ln \delta_{gi}]$ for brevity (the method). We found that gene-level ADE values are marginally independent of average confounding effects (the top panels of Fig. S10). However, we confirmed that average disease effects on the disease samples (ADD) generally align well with the average disease effects computed on the control samples (the bottom, Fig. S10).

The false sign rate (FSR) of these Bayesian estimates of ADE and SE can be controlled by an empirical Bayes procedure, such as ashr [38]. Controlling the FSR of ADEs and FDR of the ranksum tests both below 1%, we found a total of 1330 AD genes (1669 gene and cell type pairs) and an average of 152 (± 206) genes per cell type. Of them, we highlighted 182 genes sampled at most 20 genes within each cell type (Fig. 3i) and annotated 17 genes specifically acted in the microglia. We found multiple lines of independent evidence to corroborate the causal role of these genes.

Of these top AD genes found in microglial cells, we highlight five genes, including *APOE*, showing gene expressions upregulated clearly among the AD individuals (the bottom panels of Fig. 3i). *SH3RF3* gene has been found significantly associated with the age at onset of AD in the family-based genome-wide association studies [39]. Interestingly, regarding Parkinson's disease, another neurodegenerative disorder, genetic variants located in *LHFPL2* have been associated with accelerated onset of the disease by 8 to 12 years [40]. *BLNK* plays a key regulatory role in well-known microglia-specific *TREM2* signalling pathway [41] and has been proved to be upregulated with the increase of amyloid β protein [42]. To some degree, conditional genetic analysis suggested that *PTPRJ* is a link to explain pleiotropy between late-onset AD and major depressive disorder [43].

Gene ontology analysis characterizes a variety of cell-type-specific pathways in AD

We sought to characterize cell-type-specific mechanisms potentially perturbed by average 1077 (± 1122) significant AD genes found in each cell type (FDR < 20%) using goseq [44] package. Gene ontology (GO) enrichment analysis shows that DEGs identified in different cell types indeed influence markedly different biological mechanisms. By visual inspection, we can identify cell-type-specific modules of the enriched GO terms in the biological process category (Fig. 4b–d). For instance, upregulated AD genes found in excitatory neurons are highly enriched in neurodevelopmental pathways, such as "modulation of chemical synaptic transmission" and "regulation of trans-synaptic signalling." However, microglial DEGs are mostly found in immune-related activities, such as "interferon-gamma-mediated signalling pathway" and "regulation of lymphocyte-mediated immunity," and oligodendrocyte DEGs enrich terms reflect the functional role of the cell type such as "myelination" and "axon ensheathment." For the GO terms in the cellular component category, DEGs found in neurons over-represent synapse and axon, but glial cell-type-specific DEGs highlight cell-cell junction and focal adhesion (Fig. 4e). DEGs found in neurons generally participate in ion channel

Fig. 4 Genes modulated by AD pathology highlight disease mechanisms in relevant cell types in gene ontology enrichment analysis. *Ex*, expiatory neurons; *In*, inhibitory neurons; *Oligo*, oligodendrocytes; *OPC*, oligodendrocyte progenitor cells; *Mic*, microglia; *Astro*, astrocytes. Each box is scaled proportional to the level of enrichment significance (*p* value), and the colour gradient marks the number genes overlapped in each keyword and cell type. **a** The number of significant genes modulated by AD pathology. **b**–**d** Top gene ontology terms in biological process over-represented by cell-type-specific AD genes. **e** Top gene ontology terms in cellular component over-represented by cell-type-specific AD genes. **f** Top gene ontology terms in molecular function over-represented by cell-type-specific AD genes

activities, but we noticed that microglial DEGs are highly relevant to Rho GTPase and cadherin binding activities (Fig. 4f).

Discussion

We addressed a subset of a causal inference problem that emerges in disease studies. We sought to characterize and estimate the average causal effect of genes between the

case and control individuals from observational single-cell data. Delineating confounding and non-causal factors from causal effects is a crucial step to many genomics problems. Not to be trapped in circular reasoning (identifiability issue), the genomics community has been using so-called control genes and samples to extract factors shared in both control and discovery data [45–47]. One of the steps in our algorithm enjoys a similar idea, but there is no need to prescribe control cells or genes for our purposes. Along the same line, only if control features were known a priori, contrastive principal component analysis [48] could pick out non-causal factors in its latent space. Likewise, only if nuisance variables are independently observed, the variational fair autoencoder model [49] project cells onto unconfounded ("fair") latent space.

Our method builds on the outcome regression facilitate by a matching algorithm [19, 20]. Like most existing single-cell analysis pipelines, finding reliable k-nearest neighbour cells is a crucial step. If some cells in one condition were poorly matched with other cells in the opposite condition, failing to capture a shared component of confounding effects, our analysis might not work as expected. However, we want to emphasize that a failure of the matching step does not lead to an over-correction of pseudo-bulk data. It is important to understand and reliably quantify to what degree a cell-cell matching procedure can address the intrinsic and another technical variability of a single-cell RNA-seq data matrix.

A sparsity of single-cell data still casts a wide range of modelling questions. As we only consider the average effect within each individual, and we take a simple model that is just enough to capture our estimands. We ignored the notion of zero-inflation since we treat single-cell data as a count matrix, not being transformed by logarithm [50]. However, future research can take advantage of more sophisticated modelling of the individual- and cell-level observations [51], perhaps involving latent variables for representational learning.

Conclusions

We present a causal inference method that identifies and removes putative confounding effects from single-cell RNA-seq data so that the subsequent differential expression analysis can become unbiased and gain more statistical power. We have empirically shown that CoCoA-diff improved the downstream data analysis in extensive simulation experiments. We also demonstrated in real-world snRNA-seq data that the CoCoA-diff approach was necessary to reveal both well-established and novel causal genes in AD. Our work is the first application of counterfactual inference to single-cell genomics to the best of our knowledge. We expect that many existing inference methods and models can be reformulated in the same causal inference framework. More broadly, we believe that causal inference methods can improve the interpretation of genomics analysis and ultimately benefit translation researches.

Methods

Preliminary modelling of single-cell RNA-seq counting data

Individual-level gene expression quantification

We describe the single-cell RNA-seq data-generating process in a Poisson-Gamma hierarchical model. For each individual, we measure thousands of gene expression on

nearly a thousand cells. Here, we denote each individual, gene, and cell by index i, g, j, respectively. We model the expression count Y_{gj} of a gene g in a cell j follows Poisson distribution with the composite rate parameter, $\lambda_{gi}\rho_j$, where λ_{gi} quantifies the gene's mean activity in the corresponding individual i, and ρ_j accounts for the sequencing depth of a cell j. More precisely, we define the likelihood of Y_{gj}:

$$f(Y_{gj};\lambda_{gi}\rho_j) = \frac{(\lambda_{gi}\rho_j)^{Y_{gj}}}{\Gamma(Y_{gj}+1)}\exp(-\lambda_{gi}\rho_j).$$

We assume the gene and cell parameters, λ, ρ, follow a conjugate prior distribution (Gamma); more precisely, we parameterize the density function: $\text{Gamma}(\theta|a,b) \equiv b^a/\Gamma(a)\theta^{a-1}\exp(-b\theta)$. We assume smooth a prior distribution for the ρ and λ parameters, namely $\rho_j, \lambda_{gi} \sim \text{Gamma}(1,1)$. A smaller value for the hyperparameters, such as $\text{Gamma}(10^{-4}, 10^{-4})$, could encourage the effect of prior distributions vanish; however, we found it often results in numerically unstable posterior estimation when RNA-seq samples are shallowly sampled.

For the gene parameter λ_{gi}, if we defined its distribution: $\lambda_{gi} \sim \text{Gamma}(\phi^{-1}, \phi^{-1}/\mu_{gi})$, we would have $\mathbb{E}[\lambda] = \mu$ and $\mathbb{V}[\lambda] = \mu^2\phi$. Integrating out the uncertainty over λ, we derive the following negative binomial model:

$$f(Y_{gj};\mu_{gi}\rho_j,\phi) = \frac{\Gamma(Y_{gj}+\phi^{-1})}{\Gamma(Y_{gj}+1)\Gamma(\phi^{-1})}\left(\frac{1}{1+\mu_{gi}\rho_j\phi}\right)^{1/\phi}\left(\frac{\mu_{gi}\rho_j}{\mu_{gi}\rho_j+1/\phi}\right)^{Y_{gj}},$$

which preserves the characteristic quadratic relationship between the mean and variance: $\mathbb{V}[Y] = \mathbb{E}[Y] + \mathbb{E}[Y]^2\phi$.

Variational Bayes for parametric inference

We estimate the posterior distribution of λ_{gi} and ρ_j by minimizing Kullback-Leibler divergence between the joint likelihood $\mathcal{L} \equiv \prod_{gj} f(Y_{gj};\lambda_{gi},\rho_j)f(\lambda)f(\rho)$ and the fully factored variational distributions [52], $q(\lambda) = \text{Gamma}(\lambda|\alpha_\lambda,\beta_\lambda)$ and $q(\rho) = \text{Gamma}(\rho|\alpha_\rho,\beta_\rho)$. We can quickly reach convergence by alternating the following update equations:

$$\mathbb{E}_q[\lambda_{gi}] \leftarrow \frac{\sum_j Y_{gj}+1}{\sum_j \mathbb{E}[\rho_j]+1}, \quad \mathbb{E}_q[\rho_j] \leftarrow \frac{\sum_g Y_{gj}+1}{\sum_g \mathbb{E}[\lambda_{gi}]+1}.$$

Here, we first initialize $\mathbb{E}[\rho_j] = 1$ for all j, and add pseudo-count 1 on both numerators and denominators because of the prior distribution of ρ and λ.

Counterfactual confounder adjustment for differential expression analysis

Step 1: Imputation of potential outcomes by Poisson regression)

We assume binary treatment assignment and denote disease assignment (or nature's treatment) by $W \in \{0,1\}$. We denote an individual have suffered from a disease by $W = 1$ and the healthy one by $W = 0$. For clarity, we introduce the potential outcome notations to the gene expression variables. Let $Y_{gj}^{(w)}$ be gene expression of a gene g in a cell

j if this expression value was observed from an individual with a disease label $W = w$. For a disease individual, $Y^{(1)}$ is the same as observed Y value, but $Y^{(0)}$ is unknown, requiring counterfactual inference; for the opposite case, a healthy individual, $Y^{(0)}$ is observed, but $Y^{(1)}$ is counterfactual. To proceed, we assume the following causal assumptions [18, 20, 53]: (1) The disease assignment mechanism (W) is unconfounded with potential outcomes $Y^{(0)}$, $Y^{(1)}$, conditioning on some covariates X. (2) There is sufficient overlap between the case and control cells with respect to the covariates X. In other words, in almost every $X = x$, we have $0 < P(W = 1 | X = x) < 1$.

How do we find the counterfactual $Y^{(1-w)}$ for the observed $Y^{(w)}$? We construct feature vectors for potential outcome prediction by searching k-nearest neighbours (k-NN) from the cells belonging to the opposite conditions. To avoid the curse of dimensionality, we first perform spectral decomposition of single-cell data matrix and efficiently search k-NN on the spectral domain with hierarchical hashing algorithm [54]. Using these counterfactually matched cells, we construct feature matrix with each element $F_{gk}^{(1-w)} = \log(1 + Y_{gk}^{(1-w)})$ and quickly estimate regression coefficients β's in the Poisson regression by coordinate-wise descent method [55]:

$$\text{Poisson}\left(Y_{gj}^{(w)} \middle| \exp\left\{\sum_{j'=1}^{k} F_{gj'}^{(1-w)}\beta_{j'} + \beta_0 + \epsilon\right\}\right)$$

where β_0 captures the intercept term.

Given the optimized coefficients, we predict the potential outcome $\hat{Y}_{gj}^{(1-w)} \leftarrow \exp\left(\sum_{j'}^{k} F_{gj'}^{(1-w)}\hat{\beta}_{j'} + \hat{\beta}_0\right)$, ignoring the residual errors (ϵ). We also considered a non-parametric imputation method which takes weighted average over the matched cells [34, 56, 57]. Although such non-parametric methods are frequently used in single-cell data analysis, we found that Poisson regression yields more robust performance with fewer neighbouring cells than the other kNN-based imputation methods.

Step 2: Identification of potential confounding effects)

After the matching followed by the regression, we have observed $Y_{gj}^{(w)}$ and counterfactual $\hat{Y}_{gj}^{(1-w)}$. By construction, one of them carry disease-relevant effects unlike the other one. However, both of them can provide disease-invariant information that implicate potential confounding effects, denoted by μ_{gi} for a gene g and individual i:

$$\mathcal{L}' = \prod_j \prod_g \prod_{w=0}^{1} \text{Poisson}\left(Y_{gj}^{(w)} | \mu_{gi}\rho_j^{(w)}\right),$$

where we introduce the conditional-specific sequencing depth parameters $\rho^{(w)}$. However, note that μ_{gi} is shared and label-invariant.

We estimate the posterior mean of μ_{gi} by variational Bayes by alternating the following update equations until convergence:

$$\mathbb{E}_q[\mu_{gi}] = \frac{1 + \sum_{w'=0}^{1} \sum_j Y_{gj}^{(w')}}{1 + \sum_{w'=0}^{1} \sum_j \mathbb{E}_q[\rho_j^{(w')}]}, \quad \mathbb{E}_q\left[\rho_j^{(w)}\right] = \frac{1 + \sum_g Y_{gj}^{(w)}}{1 + \sum_g \mathbb{E}_q[\mu_{gi}]},$$

for all $w \in \{0, 1\}$.

Step 3: Confounder adjustment)

While fixing the value $\hat{\mu}_{gi}$ to its (variational) posterior mean $\mathbb{E}_q[\mu_{gi}]$, we redeem the confounder-adjusted mean parameters δ_{gi}, by maximizing the data likelihood:

$$\mathcal{L}'' = \prod_j \prod_g \text{Poisson}\left(Y_{gj} | \hat{\mu}_{gi} \delta_{gi} \rho_j\right)$$

Again, the posterior distributions are found by alternating the following update equations:

$$\mathbb{E}[\delta_{gi}] \leftarrow \frac{1 + \sum_j Y_{gj}}{1 + \hat{\mu}_{gi} \sum_j \mathbb{E}[\rho_j]}, \quad \mathbb{E}_q[\rho_j] \leftarrow \frac{1 + \sum_g Y_{gj} + 1}{1 + \sum_g \hat{\mu}_{gi} \mathbb{E}[\delta_{gi}]}.$$

Since the δ_{gi} variable follows Gamma distribution, we also have

$$\mathbb{E}_q[\ln \delta_{gi}] = \psi\left(1 + \sum_j Y_{gj}\right) - \log\left(1 + \hat{\mu}_{gi} \sum_j \mathbb{E}[\rho_j]\right),$$

where $\psi(\cdot)$ is the digamma function, and approximate its variance,

$$\mathbb{V}[\ln \delta_{gi}] = \left(\sum_j Y_{gj}\right)^{-1}.$$

See the following derivations of the Gaussian approximation of Gamma distribution.

Technical details

Local Gaussian approximation of Gamma distribution

We approximate the distribution of $\ln\lambda$ by constructing a local quadratic approximation of the original log-probability density function:

$$\ln p(\lambda | \alpha, \beta) = (\alpha - 1) \ln \lambda - \beta\lambda + \alpha \ln \beta - \ln\Gamma(\alpha)$$

Letting $\phi = \ln \lambda$, we can rewrite the above as:

$$\mathcal{L} = (\alpha - 1)\phi - \beta e^{\phi} + \alpha \ln \beta - \ln\Gamma(\alpha)$$

At some $\hat{\phi}$, we can find a quadratic form:

$$\mathcal{L} \approx -\frac{1}{2} \beta e^{\hat{\phi}} \left(\phi - \left[\hat{\phi} + \frac{\alpha - 1 - \beta e^{\hat{\phi}}}{\beta e^{\hat{\phi}}}\right]\right)^2 + \text{const.}$$

Setting $e^{\hat{\phi}} = (\alpha - 1)/\beta$ (the mode of Gamma distribution), we have

$$\mathcal{L} \approx -\frac{1}{2}(\alpha-1)\left(\phi-\hat{\phi}\right)^2 + \text{const}$$

Finally, we have

$p(\phi|\alpha,\beta) \approx \mathcal{N}|(\phi|\ \ln((\alpha-1)/\beta),(\alpha-1)^{-1})$. In our case, we assumed $\lambda \sim \text{Gamma}(1,1)$ a priori and only derived approximate Gaussian whenever we have at least 1 read per individual; therefore, $\alpha > 1$. However, if $0 < \alpha \leq 1$, we can approximate the Gaussian at $\lambda = \alpha/\beta$, and this results in $p(\phi|\alpha,\beta) \approx \mathcal{N}|(\phi|\ \ln(\alpha/\beta),\alpha^{-1})$.

Derivation of average disease effect across individuals (meta-analysis)

From the above, we derived the posterior distribution of $\phi_i(\equiv \ln\lambda_i)$ variables. Let $\eta_i = \ln((\alpha_i - 1)/\beta)$ and $\sigma_i^2 = (\alpha_i-1)^{-1}$. Then we have $\phi_i \sim \mathcal{N}(\phi_i|\eta_i,\sigma_i^2)$. We can find another variational distribution $r \equiv \mathcal{N}(\phi_i|\overline{\eta},\overline{\sigma}^2)$ averaging over all these individual-level posterior distributions by optimizing the following Kullback-Leibler divergence:

$$D = \ln \int d\phi_i \frac{q(\phi_i|\eta_i,\sigma_i^2)}{r(\phi_i|\overline{\mu},\overline{\sigma}^2)} r(\phi_i|\overline{\mu},\overline{\sigma}^2)$$

By Jensen's inequality,

$$D \geq -\sum_{i=1}^{n} \mathbb{E}_r\left[\frac{1}{2}\ln\sigma_i^2 + \frac{1}{2\sigma_i^2}(\phi_i-\eta_i)^2\right] + \mathbb{E}_r\left[\frac{1}{2}\ln\bar{\sigma}^2 + \frac{1}{2\bar{\sigma}^2}(\phi_i-\bar{\eta})^2\right] + \text{const.}$$

Optimizing this with respect to $\overline{\eta}$ and $\overline{\sigma}$, we have:

$$\overline{\eta} = \frac{\sum_i \eta_i/\sigma_i^2}{\sum_i 1/\sigma_i^2}, \quad \overline{\sigma}^2 = \frac{1}{\sum_i 1/\sigma_i^2}.$$

Cell type annotation by constrained mixture of von Mises-Fisher

We classify a cell type of 70,634 cells based on the prior knowledge of cell-type-specific 2648 marker genes on 8 brain cell types [33]. Using 1726 genes present in our data, we construct a normalized vector $\mathbf{m}_j$ for each cell with the dimensionality $d = 1726$ and $\|\mathbf{m}_j\| = 1$. Additionally, we define a label matrix L to designate the activities of the marker genes to the relevant cell types. Each element L_{gk} takes 1 if and only if a gene $g \in [d]$ is active on a $k \in [8]$ cell type; otherwise, we set $L_{jk} = 0$. We assume that each normalized vector $\mathbf{m}_j$ follows von Mises-Fisher (vMF) distribution with cell type k-specific mean vector $\boldsymbol{\theta}_k$ with the concentration parameter κ, shared across all the cell types:

$$\mathcal{L} = C(\kappa)\prod_{j=1}^{n}\prod_{k=1}^{K}\left[\ \exp\left(\kappa\mathbf{m}_j^{\top}\boldsymbol{\theta}_k\right)\right]^{z_{jk}},$$

where $n = 70634$ for the cells and $K = 8$ for the cell types. Here, we introduce z_{jk}, an indicator variable to mark the assignment of a cell j to a cell type k. Our goal is to estimate the posterior probability of $z_{jk} = 1$ by stochastic expectation maximization (EM) algorithm. In the E-step, we simply sample the latent membership z_{jk} from the discrete

distribution proportional to $\exp(\mathbf{m}_j^\top \boldsymbol{\theta}_k)$. In the M-step, we maximize the mean and concentration parameters with the cell type constraints L:

$$\boldsymbol{\theta}_k \leftarrow \frac{\sum_j z_{jk}\mathbf{m}_j \circ \mathbf{l}_k}{\| \sum_j z_{jk}\mathbf{m}_j \circ \mathbf{l}_k \|}, \quad \kappa \leftarrow \frac{rd - r^3}{1 - r^2}$$

where $r = \| \sum_j \mathbf{m}_j \| / n$. Derivation for the optimization of κ can be found in the previous work on von Mises Fisher mixture model [58].

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-021-02438-4.

Additional file 1: Figs. S1 to S10. with the figure legend texts.

Additional file 2. Review history

Acknowledgements

We thank Liang He, Matthew Lincoln, and Tomokazu Sumida for biological inspiration, which later led to adopting von Mises-Fisher model for cell type annotations. We also thank Abhishek Sarkar for helpful discussion on individual-level quantification model for single-cell RNA-seq data. We owe a debt of gratitude to anonymous reviewers for constructive feedback that significantly improved the exposition of this manuscript.

Review history

The review history is available as Additional file 2.

Peer review information

Barbara Cheifet was the primary editor of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Authors' contributions

YPP and MK conceived the concept and designed the study. MK directed and coordinated data acquisition. YPP implemented software and scripts, and conducted the statistical analysis. YPP and MK wrote the manuscript.

Authors' information

Authors' Twitter handle:
YPP: @ypp_lab, MK: @manoliskellis

Funding

This work was supported by NIH grants U01NS110453, U24-HG009446, and U01-RFA-HG009088 (MK). We also acknowledge generous supports from the BC Cancer Foundation, Project ID 1NSRG048 (YPP).

Availability of data and materials

We made the C++ source code of binary programs used in simulation and data analysis available in the following public repository, https://ypark.github.io/mmutil/ under MIT License. We also deposited a compressed tarball in Zenodo under the following accession [59]: https://doi.org/10.5281/zenodo.5106691 A full list of differential expression analysis, analysis pipeline (GNU Makefile), and the vignettes of simulation experiments are available in the separate repository: https://ypark.github.io/cocoa_paper/ [60].
The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.synapse.org). Study data were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago.

Declarations

Ethics approval and consent to participate

Data collection was supported through funding by NIA grants RF1AG57473 (single nucleus RNAseq) and the Illinois Department of Public Health (ROSMAP). The Religious Orders Study and Rush Memory and Aging Project were approved by an IRB of Rush University Medical Center.

Competing interests

The authors declare no competing interests.

Author details

[1]Department of Pathology and Laboratory Medicine, Department of Statistics, University of British Columbia, Vancouver, BC, Canada. [2]Department of Molecular Oncology, BC Cancer, Vancouver, BC, Canada. [3]Computer Science

and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Broad Institute of MIT and Harvard, Cambridge, MA, USA.

Received: 8 January 2021 Accepted: 16 July 2021
Published online: 17 August 2021

References

1. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6 https://doi.org/10.1038/nbt.2859.
2. Ståhl PL, Salmén F, Vickovic S, Lundmark A, Navarro JF, Magnusson J, et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science. 2016;353(6294):78–82 https://doi.org/10.1126/science.aaf2403.
3. Norman TM, Horlbeck MA, Replogle JM, Ge AY, Xu A, Jost M, et al. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. Science. 2019;365(6455):786–93 https://doi.org/10.1126/science.aax4438.
4. van der Wijst MGP, et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet. 2018.
5. Sarkar AK, Tung PY, Blischak JD, Burnett JE, Li YI, Stephens M, et al. Discovery and characterization of variance QTLs in human induced pluripotent stem cells. PLoS Genet. 2019;15(4):e1008045 https://doi.org/10.1371/journal.pgen.1008045.
6. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A. 2001;98(9):5116–21 https://doi.org/10.1073/pnas.091062498.
7. Irizarry RA, Warren D, Spencer F, Kim IF, Biswal S, Frank BC, et al. Multiple-laboratory comparison of microarray platforms. Nat Methods. 2005;2(5):345–50 https://doi.org/10.1038/nmeth756.
8. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics. 2003;4(2):249–64 https://doi.org/10.1093/biostatistics/4.2.249.
9. Robinson MD, Smyth GK. Moderated statistical tests for assessing differences in tag abundance. Bioinformatics. 2007;23(21):2881–7 https://doi.org/10.1093/bioinformatics/btm453.
10. Robinson MD, Smyth GK. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. Biostatistics. 2008;9(2):321–32 https://doi.org/10.1093/biostatistics/kxm030.
11. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106 https://doi.org/10.1186/gb-2010-11-10-r106.
12. Law CW, Chen Y, Shi W, Smyth GK. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. 2014;15:R29.
13. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:1–21.
14. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61 https://doi.org/10.1038/nmeth.4612.
15. Crowell HL, et al. On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. bioRxiv. 2019:713412.
16. Mandric I, et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. Nat Commun. 2020;11:1–9.
17. Rubin DB. Bayesian inference for causal effects: the role of randomization. aos. 1978;6:34–58.
18. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. Am Stat. 1985;39:33–8.
19. Heckman JJ, Ichimura H, Todd PE. Matching as an econometric evaluation estimator: evidence from evaluating a job training programme. Rev Econ Stud. 1997;64(4):605–54 https://doi.org/10.2307/2971733.
20. Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. Econometrica. 2006;74(1):235–67 https://doi.org/10.1111/j.1468-0262.2006.00655.x.
21. Davey Smith G, Hemani G. Mendelian randomization: genetic anchors for causal inference in epidemiological studies. Hum Mol Genet. 2014;23(R1):R89–98 https://doi.org/10.1093/hmg/ddu328.
22. Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019;570(7761):332–7 https://doi.org/10.1038/s41586-019-1195-2.
23. VanderWeele TJ. Bias formulas for sensitivity analysis for direct and indirect effects. Epidemiology. 2010;21(4):540–51 https://doi.org/10.1097/EDE.0b013e3181df191c.
24. Glynn AN. The product and difference fallacies for indirect effects: the product and difference fallacies for indirect effects. Am J Polit Sci. 2012;56(1):257–69 https://doi.org/10.1111/j.1540-5907.2011.00543.x.
25. Pearl, J. & Mackenzie, D. The book of why: the new science of cause and effect. 2018. Basic Books.
26. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013;41(1):196–220 https://doi.org/10.1214/12-aos1058.
27. Wilcoxon F. Individual comparisons by ranking methods. Biom Bull. 1945;1(6):80–3 https://doi.org/10.2307/3001968.
28. Andri et mult. al., S. DescTools: tools for descriptive statistics. 2021.
29. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):278 https://doi.org/10.1186/s13059-015-0844-5.
30. Storey JD. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol. 2002;64(3):479–98 https://doi.org/10.1111/1467-9868.00346.
31. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci. 2003;100(16):9440–5 https://doi.org/10.1073/pnas.1530509100.
32. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. Int J Epidemiol. 2010;39(2):417–20 https://doi.org/10.1093/ije/dyp334.

33. Gandal MJ, Zhang P, Hadjimichael E, Walker RL, Chen C, Liu S, et al. Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science. 2018;362(6420):eaat8127 https://doi.org/10.1126/science.aat8127.
34. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park JE. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2019; https://doi.org/10.1093/bioinformatics/btz625.
35. Velmeshev D, Schirmer L, Jung D, Haeussler M, Perez Y, Mayer S, et al. Single-cell genomics identifies cell type-specific molecular changes in autism. Science. 2019;364(6441):685–9 https://doi.org/10.1126/science.aav8130.
36. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol. 1995;57:289–300.
37. Holm S. A simple sequentially rejective multiple test procedure. Scand Stat Theory Appl. 1979;6:65–70.
38. Stephens M. False discovery rates: a new deal. Biostatistics. 2017;18(2):275–94 https://doi.org/10.1093/biostatistics/kxw041.
39. Lee JH, Cheng R, Vardarajan B, Lantigua R, Reyes-Dumeyer D, Ortmann W, et al. Genetic modifiers of age at onset in carriers of the G206A mutation in PSEN1 with familial Alzheimer disease among caribbean hispanics. JAMA Neurol. 2015;72(9):1043–51 https://doi.org/10.1001/jamaneurol.2015.1424.
40. Hill-Burns EM, Ross OA, Wissemann WT, Soto-Ortolaza AI, Zareparsi S, Siuda J, et al. Identification of genetic modifiers of age-at-onset for familial Parkinson's disease. Hum Mol Genet. 2016;25(17):3849–62 https://doi.org/10.1093/hmg/ddw206.
41. Zajkowicz A, Gdowicz-Kłosok A, Krześniak M, Janus P, Łasut B, Rusin M. The Alzheimer's disease-associated TREM2 gene is regulated by p53 tumor suppressor protein. Neurosci Lett. 2018;681:62–7 https://doi.org/10.1016/j.neulet.2018.05.037.
42. Sierksma A, et al. Novel Alzheimer risk genes determine the microglia response to amyloid-β but not to TAU pathology. EMBO Mol Med. 2020;12:e10606.
43. Lutz MW, Sprague D, Barrera J, Chiba-Falek O. Shared genetic etiology underlying Alzheimer's disease and major depressive disorder. Transl Psychiatry. 2020;10(1):88 https://doi.org/10.1038/s41398-020-0769-y.
44. Young MD, Wakefield MJ, Smyth GK, Oshlack A. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. 2010;11(2):R14 https://doi.org/10.1186/gb-2010-11-2-r14.
45. Risso D, Ngai J, Speed TP, Dudoit S. Normalization of RNA-seq data using factor analysis of control genes or samples. Nat Biotechnol. 2014;32(9):896–902 https://doi.org/10.1038/nbt.2931.
46. Gagnon-Bartsch JA, Speed TP. Using control genes to correct for unwanted variation in microarray data. Biostatistics. 2012;13(3):539–52 https://doi.org/10.1093/biostatistics/kxr034.
47. Schölkopf B, Hogg DW, Wang D, Foreman-Mackey D, Janzing D, Simon-Gabriel CJ, et al. Modeling confounding by half-sibling regression. Proc Natl Acad Sci U S A. 2016;113(27):7391–8 https://doi.org/10.1073/pnas.1511656113.
48. Abid A, Zhang MJ, Bagaria VK, Zou J. Exploring patterns enriched in a dataset with contrastive principal component analysis. Nat Commun. 2018;9(1):2134 https://doi.org/10.1038/s41467-018-04608-8.
49. Louizos, C., Swersky, K., Li, Y., Welling, M., Zemel, R. The variational fair autoencoder. 2015. at http://arxiv.org/abs/1511.00830.
50. Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome Biol. 2019;20(1):295 https://doi.org/10.1186/s13059-019-1861-6.
51. Sarkar A, Stephens M. Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. bioRxiv. 2020:2020.04.07.030007.
52. Jordan MI, Jaakkola TS, Saul LK. An introduction to variational methods for graphical models. Mach Learn. 1999;37(2):183–233 https://doi.org/10.1023/A:1007665907178.
53. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica. 2003;71(4):1161–89 https://doi.org/10.1111/1468-0262.00442.
54. Malkov, Y. A. & Yashunin, D. A. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. 2016. at http://arxiv.org/abs/1603.09320.
55. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1–22.
56. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7 https://doi.org/10.1038/nbt.4091.
57. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. Nat Biotechnol. 2019;37(6):685–91 https://doi.org/10.1038/s41587-019-0113-3.
58. Banerjee A, Dhillon IS, Ghosh J, Sra S. Clustering on the unit hypersphere using von Mises-Fisher distributions. J Mach Learn Res. 2005;6:1345–82.
59. Park, Y. Matrix Market Utility for single-cell sequencing data analysis. 2021. Zenodo. https://doi.org/10.5281/zenodo.5106691.
60. Park, Y. CoCoA-diff: counterfactual inference for single-cell gene expression analysis. Source code. 2021. https://github.com/ypark/cocoa_paper.

Publisher's Note