Counterfactual inference for single-cell gene expression analysis

14:04:43, Apr 20, 2021

Author names and affiliations

Yongjin Park[1], Manolis Kellis[2]

1. Department of Pathology and Laboratory Medicine, Department of Statistics, University of British Columbia, Vancouver, BC, Canada
2. Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, United States of America

Contact:

- Yongjin Park: `ypp@stat.ubc.ca`
- Manolis Kellis: `manoli@mit.edu`

Abstract

Finding a causal gene is a fundamental problem in genomic medicine. We present a causal inference framework that prioritizes disease genes by adjusting confounders without prior knowledge of control variables. 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 identified that 215 causal genes are differentially regulated by the disease 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

Problem definition

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 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 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, $\lambda_{gi} \approx \sum_j I\{j \in S_i\} Y_{gj}/|S_i|$, 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 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 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; (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*: 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 for single-cell differential expression analysis

The purpose of our counterfactual confounder adjustment (CoCoA) for single-cell gene expression analysis (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, $\{Y_{gj} : g \in \mathsf{genes}, j \in \mathsf{cells}\}$, we want to estimate two types of pseudo-bulk data: (1) the estimated confounders, $\{\mu_{gi} : g \in \mathsf{genes}, i \in \mathsf{individuals}\}$, and the residual differential effects, $\{\delta_{gi} : g \in \mathsf{genes}, i \in \mathsf{individuals}\}$. 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 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 ($W \to \lambda$) and confounding factors ($X \to \lambda$). The other genes are only affected by confounding factors ($X \to \lambda$); we introduced five confounding variables $\{X_{ik} : i \in \mathsf{individuals}, k \in [5]\}$, 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 ($W \to \lambda$; 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. 1g). 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 up-regulated 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 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 $\times 10^{-7}$).

Simulation experiments

The design of simulation experiments

We evaluated the performance of CoCoA in downstream 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 $\log\lambda$ 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 α required to introduce biases on W and the residual error vector ϵ_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., $\mathbb{V}[X\alpha]/\mathbb{V}[X\alpha + \epsilon_W] = \sigma^2_{X\to W}$. We generate a binary label assignment for an individual i by flipping a coin:

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

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

$$\ln \lambda_{gi} = \underbrace{\tau_g W_i}_{\text{disease effect}} + \underbrace{\sum_k X_{ik}\beta_{k\in[d_C]}}_{\text{confounding effect}} + \underbrace{\sum_{l\in[d_B]} B_{il}\gamma_l}_{\text{batch effect}} + \epsilon_\lambda,$$

where the covariate effect $\beta_g \sim \mathcal{N}(0, \sigma^2_{X\to Y}/5)$, gene-level causal effect $\tau_g \sim \mathcal{N}(0, \sigma^2_{W\to Y})$, and the residuals $\epsilon_\lambda \sim \mathcal{N}(0, 1-\sigma^2_{X\to Y}-\sigma^2_{W\to Y})$. Using the individual-level mean values λ, we stochastically generated 50 cells per individual, multiplying the individual-level average expressions with random sequencing depth (ρ), sampled from $\rho \sim \mathsf{Gamma}(6.25, 6.25)$. For each cell j,

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

Counterfactual adjustment of pseudo-bulk data improves statistical power

We repeated our experiments 30 times for all the different configurations and summarized the performance by the area under the precision-recall curve (AUPRC) and the statistical power when the false discovery rate was controlled at 10% (Fig. 2b-d). 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`. Extensive simulation results can be found in the supplementary website (`https://ypark.github.io/cocoa_paper/result_simulation.html`) along with the source code repository (`https://github.com/ypark/cocoa_paper`).

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

- CoCoA: individual-level disease effects δ_{gi} estimated by CoCoA algorithm (Fig. 1c, step 3).
- Confounder (estimated): the remaining confounder effects μ_{gi}, also esimated by CoCoA algorithm (Fig. 1c, step 2)
- Average: 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 ($\sum_{j \in S_i} Y_{gj}$).

We found that CoCoA effectively adjusts the simulated confounding effects and markedly improves statistical power and AUPRC of differential expression analysis. We also confirmed that the confounder effects that we corrected show no association with disease labels. Unlike different types of covariates (confounding and batch effects) create confusion to an algorithm in selecting factors to be adjusted in a matrix factorization-based analysis, CoCoA performed robustly, showing only a modest reduction of power on large batch effects.

We further investigated whether the bias introduced by confounding effects can be overcome by doubling the number of individuals (Fig. 2e). Because of imperfect confoundedness (50% of variability), we found that larger sample size can improve the accuracy of differential analysis. However, the performance gap between the CoCoA and pseudo-bulk analysis by total expression sustains in almost all cases.

Lastly, we considered a hypothetical scenario where principal components of a gene expression matrix are optimally selected to characterize confounding factors (Fig. 2f-g). CoCoA closely attains the optimal level of statistical power when all the covariates act as a confounding factor (Fig. 2f). However, if there exist other non-confounding factors, CoCoA alone may not reduce all the other types of unwanted variation in a single-cell data matrix (Fig. 2g).

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.[29] Of the total 2,648 PsychENCODE marker genes, we used 1,726 genes expressed in this data set as features (Supplementary Fig. 1). 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 Methods). 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 (Supplementary Fig. 2). We also found that cell types correspond to unique cell clusters after BBKNN (batch-balancing k-Nearest Neighbour)[30] preprocessing (Fig. 3a), showing no apparent bias induced by other demographic and pathological variables (Supplementary Fig. 3). 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.[31]

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 1,444 cells per individual, of which 50.65% cells stem from Ex (N=726 $\pm$ 382 SD), 12.71% cells from In (N=182 $\pm$ 107), 25.72% cells for Oligo (N=380 $\pm$ 252), 3.56% cells from OPC (N= 54 $\pm$ 34), 2.43% cells from Microglia (N= 33 $\pm$ 24), 4.94% cells from Astro (N= 69 $\pm$ 46), 0.1% cells from Endo and Per. We have 726 excitatory neurons per individual (Fig. 3e), of which 52.62% (N=399 $\pm$ 250) cells from the layer 4, 33.1% (N=233 $\pm$ 108) cells from the layer L2/3, 8.26% (N= 50 $\pm$ 28) cells from

the cortical layer L5/6 (CC), 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), 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 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 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[32]) $< 1\%$, we found 1,648 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[33]) 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 procedure did not introduce a systematic bias by shrinking variance on the case or control samples (Supplementary Fig. 6). 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 (Supplementary Fig. 7).

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:

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

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 Supplementary Fig. 8). 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, Supplementary Fig. 8).

The false sign rate (FSR) of these Bayesian estimates of ADE and SE can be controlled by an empirical Bayes procedure, such as `ashr`.[34] Controlling the FSR of ADEs and FDR of the ranksum tests both below 1%, we found a total of 1,330 AD genes (1,669 gene and cell type pairs) and an average of 152 ($\pm$ 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*, sh wing gene expressions up-regulated 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.[35] 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.[36] *BLNK* plays a key regulatory role in well-known microglia-specific *TREM2* signaling pathway[37] and has been proved to be up-regulated with the increase of amyloid β protein.[38] To some degree, conditional genetic analysis suggested that *PTPRJ* is a link to explain pleiotropy between late-onset AD and major depressive disorder.[39]

The full statistics and the visualization of representative examples are shared in the supplementary website (`https://ypark.github.io/cocoa_paper/result_AD.html`).

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

We sought to characterize cell-type-specific mechanisms potentially perturbed by 1,077 ($\pm$ 1,122) significant AD genes found in each cell type (FDR $<$ 20%) using `goseq`[40] 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, up-regulated AD genes found in excitatory neurons are highly enriched in neurodevelopmental pathways, such as "modulation of chemical synaptic transmission," "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 activities, but we noticed that microglial DEGs are highly relevant to Rho GTPase and cadherin binding activities (Fig. 4f).

The full statistics of gene ontology enrichment analysis are available through the supplementary

website (`https://ypark.github.io/cocoa_paper/result_AD_GO.html`).

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[41–43]. 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[44] could pick out non-causal factors in its latent space. Likewise, only if nuisance variables are independently observed, the variational fair autoencoder model[45] 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 overcorrection 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[46]. However, future research can take advantage of more sophisticated modelling of the individual- and cell-level observations[47], 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 improved the downstream data analysis in extensive simulation experiments. We also demonstrated in real-world snRNA-seq data that the CoCoA 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}:

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

We assume the gene and cell parameters, λ, ρ, follow a conjugate prior distribution (Gamma); more precisely, we parameterize the density function: $\mathsf{Gamma}(\theta|a,b) \equiv b^a/\Gamma(a+1)\theta^{a-1}\exp(-b\theta)$. We assume smooth a prior distribution for the ρ and λ parameters, namely $\rho_j, \lambda_{gi} \sim \mathsf{Gamma}(1,1)$. A smaller value for the hyperparameters, such as $\mathsf{Gamma}(10^{-4}, 10^{-4})$, 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.

Remark: For the gene parameter λ_{gi}, if we defined its distribution: $\lambda_{gi} \sim \mathsf{Gamma}(\phi^{-1}, \phi^{-1}/\mu_{gi})$, 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,[48] $q(\lambda) = \mathsf{Gamma}(\lambda|\alpha_\lambda, \beta_\lambda)$ and $q(\rho) = \mathsf{Gamma}(\rho|\alpha_\rho, \beta_\rho)$. We can quickly reach convergence by alternating the following update equations:

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

Here, we first initialize $\mathbb{E}\big[\rho_j\big] = 1$ 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:[18,20,49] (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 (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.[50] 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:[51]

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

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.[30,52,53] 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:

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

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:

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

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

Step 3 (Confounder adjustment)

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

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

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

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

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

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

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

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

See the supplementary method for the derivation of the Gaussian approximation of Gamma distribution.

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 2,648 marker genes on 8 brain cell types.[29] Using 1,726 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 θ_k with the concentration parameter κ, shared across all the cell types:

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

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

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

where $r = \| \sum_j X_j \| / N$. Derivation for the optimization of κ can be found in the previous work on von Mises Fisher mixture model[54].

Acknowledgement

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.

The authors declare no competing interests.

Supplementary methods

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:

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

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

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

Finally, we have

$p(\phi|\alpha, \beta) \approx \mathcal{N}(\phi \big| \ln((\alpha - 1)/\beta), (\alpha - 1)^{-1})$. In our case, we assumed $\lambda \sim \mathsf{Gamma}(1, 1)$ *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 $p(\phi|\alpha, \beta) \approx \mathcal{N}(\phi \big| \ln(\alpha/\beta), \alpha^{-1})$.

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 $\phi_i \sim \mathcal{N}(\phi_i \big| \eta_i, \sigma_i^2)$. We can find another variational distribution $r \equiv \mathcal{N}(\phi_i|\bar{\eta}, \bar{\sigma}^2)$ averaging over all these individual-level posterior distributions by optimizing the following Kullback-Leibler divergence:

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

By Jensen's inequality,

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

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

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

Figures

Figure 1.

Counterfactual confounder adjustment (CoCoA) for single-cell differential gene expression analysis. **(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 seeks to prioritize genes causally-modulated by a disease status, not the genes affecting the predisposition and risk of disease development. **(c)** Overview of CoCoA 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 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 accurately estimates disease-causing effects (δ_g), showing high correlation with true differential effects on causal genes. **(g)** Illustration of CoCoA 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.

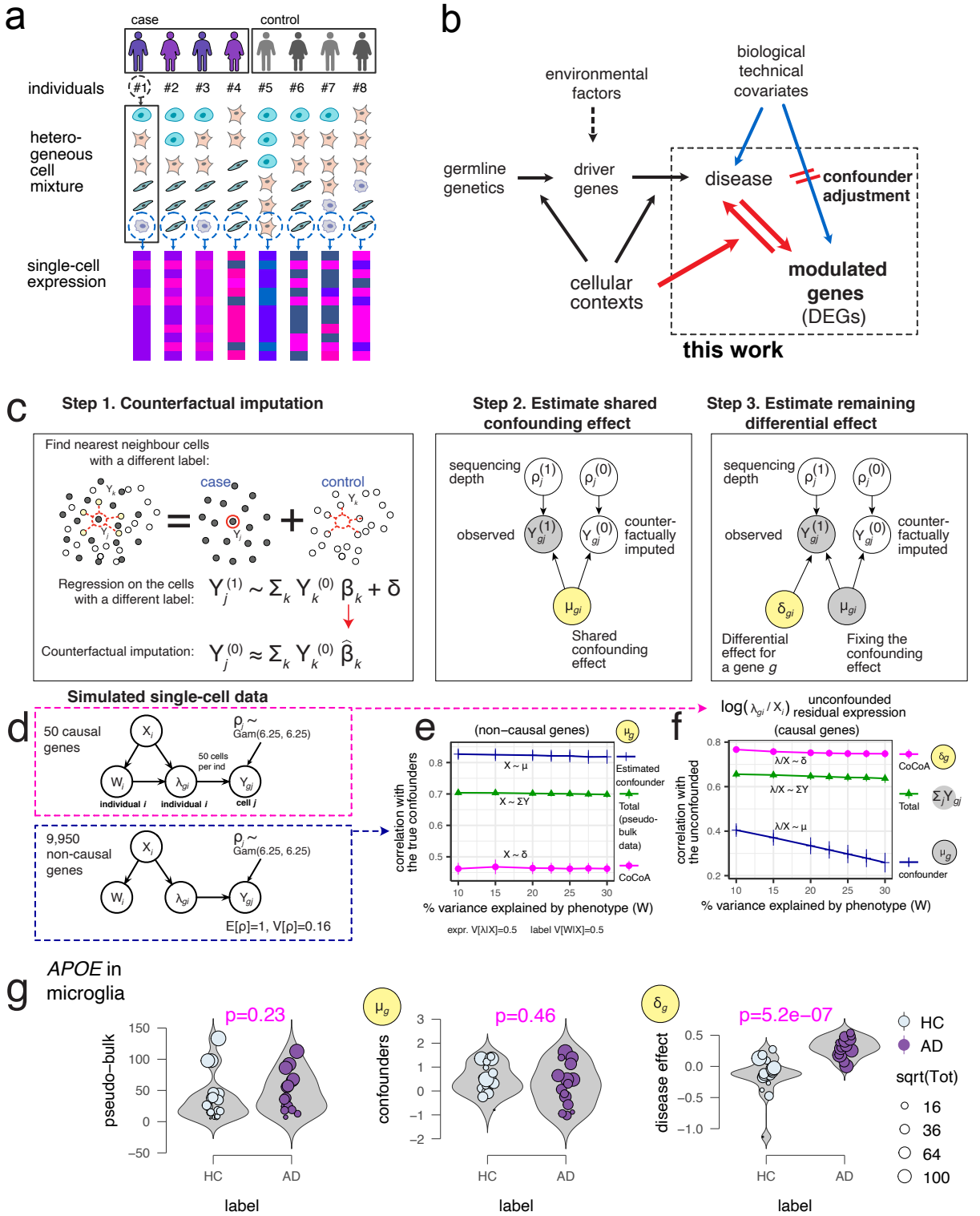

a
case
control
individuals
#1
#2
#3
#4
#5
#6
#7
#8
hetero-
geneous
cell
mixture
single-cell
expression
b
environmental
factors
biological
technical
covariates
germline
genetics
driver
genes
disease
confounder
adjustment
cellular
contexts
modulated
genes
(DEGs)
this work
c
Step 1. Counterfactual imputation
Find nearest neighbour cells
with a different label:
case
control
Regression on the cells
with a different label:
Counterfactual imputation:
Step 2. Estimate shared
confounding effect
sequencing
depth
observed
counter-
factually
imputed
Shared
confounding
effect
Step 3. Estimate remaining
differential effect
Differential
effect for
a gene g
Fixing the
confounding
effect
Simulated single-cell data
d
50 causal
genes
50 cells
per ind
individual i
individual i
cell j
9,950
non-causal
genes
E[ρ]=1, V[ρ]=0.16
e
(non−causal genes)
correlation with
the true confounders
Estimated
confounder
Total
(pseudo-
bulk
data)
CoCoA
% variance explained by phenotype (W)
expr. V[λ|X]=0.5
label V[W|X]=0.5
f
unconfounded
residual expression
(causal genes)
correlation with
the unconfounded
CoCoA
Total
confounder
% variance explained by phenotype (W)
g
APOE in
microglia
p=0.23
p=0.46
p=5.2e−07
pseudo−bulk
confounders
disease effect
HC
AD
label
sqrt(Tot)
16
36
64
100

Figure 2.

Simulation experiments. Extensive simulation experiments confirm that CoCoA 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 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 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. *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; *average*: pseudo-bulk expression averaged over cells within each individual; *confoudner*: the estimated confounding effect μ_{gi} (the step 2 of Fig. 1c). *AUPRC*: area under precision recall curve (numerically integrated by `DescTool` implemented in `R`).[28] **c)** Simulation results when only one covariate confounds disease label assignment and gene expression values, and the other four variables are affecting gene expression as batch effects. **(d)** Can sample size help? Simulation results compare the statistical power of differential expression analysis conducted on two different sizes of simulation studies (N=40 vs. N=80). **(e)** CoCoA avoids over-correction of expression variation and attains nearly optimal performance in differential analysis when there is only one confounding factor and the other four factors correspond to batch effects.

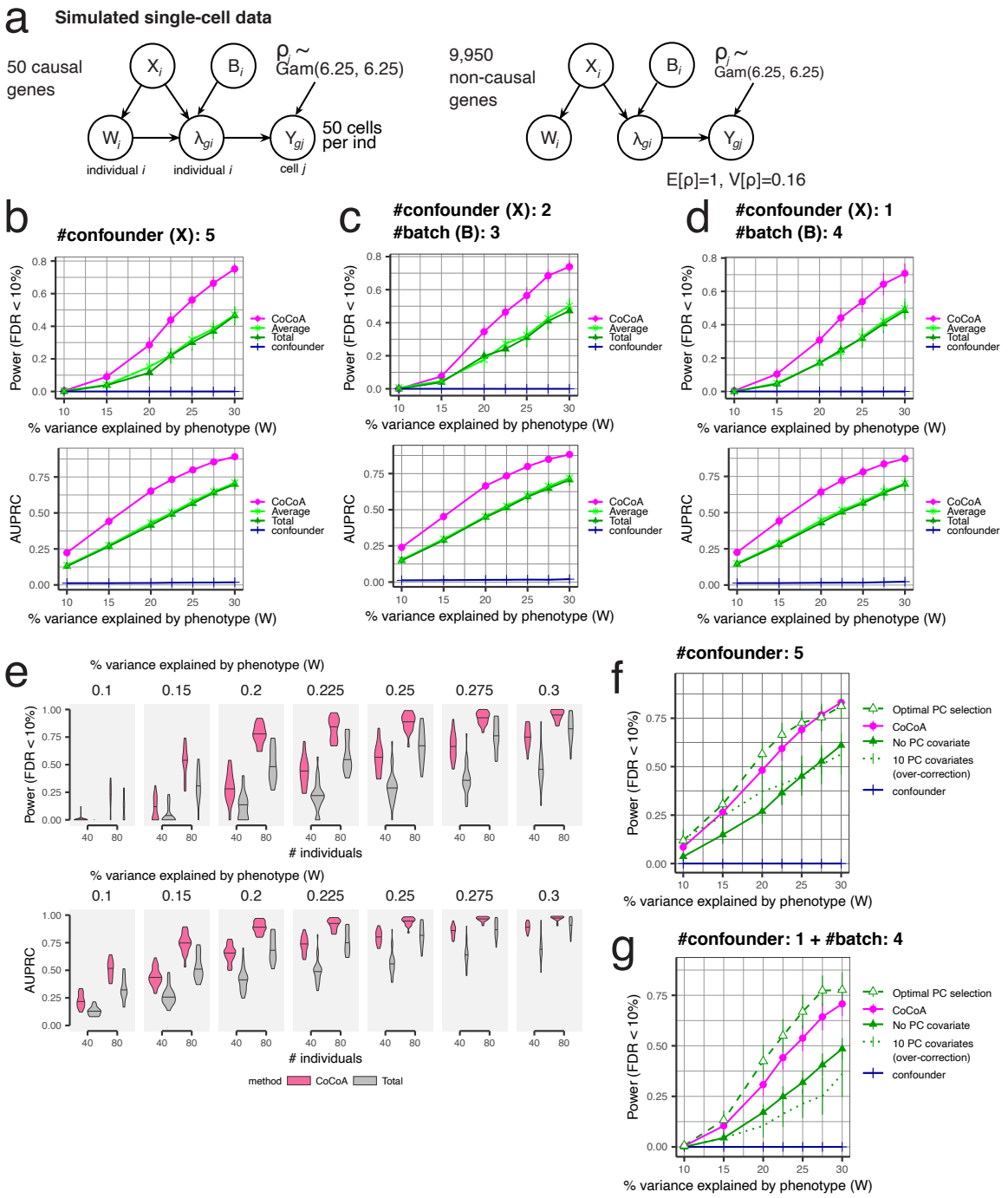

a
Simulated single-cell data
50 causal genes
X_i
B_i
$\rho_j \sim$ Gam(6.25, 6.25)
W_i
λ_{gi}
Y_{gj}
50 cells per ind
individual i
individual i
cell j
9,950 non-causal genes
E[ρ]=1, V[ρ]=0.16
b
#confounder (X): 5
c
#confounder (X): 2
#batch (B): 3
d
#confounder (X): 1
#batch (B): 4
Power (FDR < 10%)
AUPRC
% variance explained by phenotype (W)
CoCoA
Average
Total
confounder
e
individuals
method
CoCoA
Total
f
#confounder: 5
Optimal PC selection
CoCoA
No PC covariate
10 PC covariates (over-correction)
confounder
g
#confounder: 1 + #batch: 4

Figure 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 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.

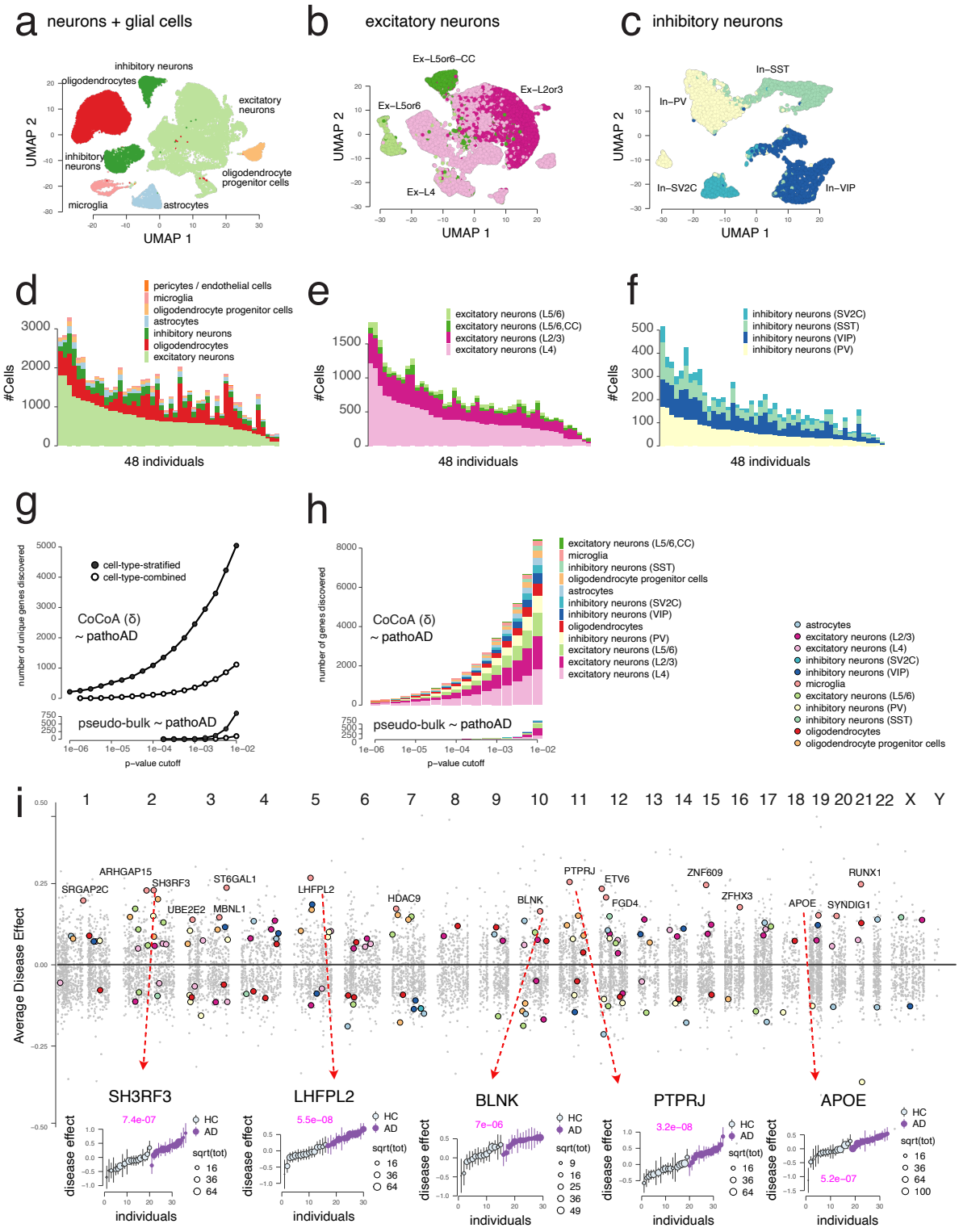
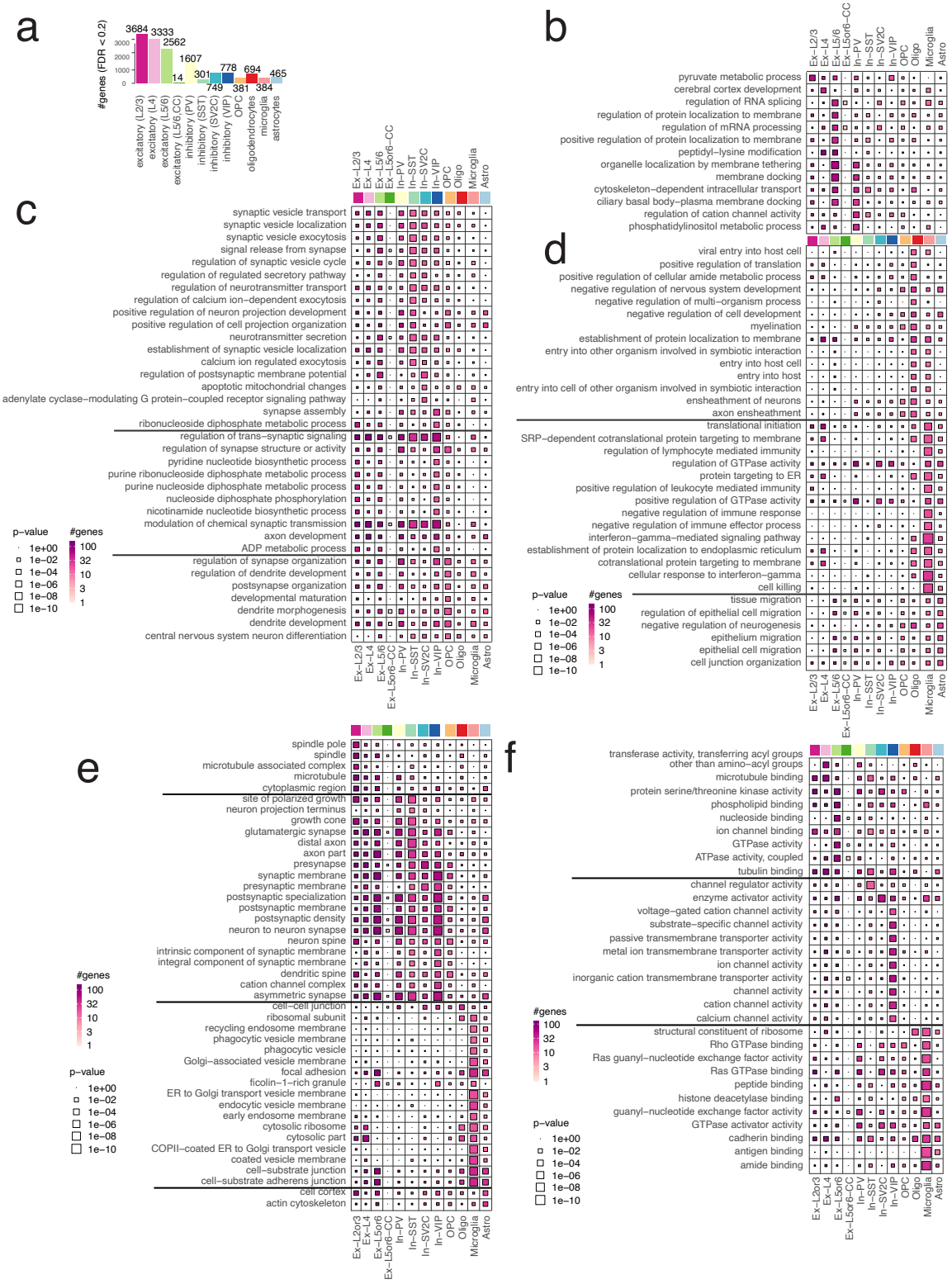

a
neurons + glial cells
inhibitory neurons
oligodendrocytes
excitatory neurons
inhibitory neurons
oligodendrocyte progenitor cells
microglia
astrocytes
UMAP 2
UMAP 1
b
excitatory neurons
Ex−L5or6−CC
Ex−L2or3
Ex−L5or6
Ex−L4
UMAP 2
UMAP 1
c
inhibitory neurons
In−SST
In−PV
In−SV2C
In−VIP
UMAP 2
UMAP 1
d
pericytes / endothelial cells
microglia
oligodendrocyte progenitor cells
astrocytes
inhibitory neurons
oligodendrocytes
excitatory neurons
#Cells
48 individuals
e
excitatory neurons (L5/6)
excitatory neurons (L5/6,CC)
excitatory neurons (L2/3)
excitatory neurons (L4)
#Cells
48 individuals
f
inhibitory neurons (SV2C)
inhibitory neurons (SST)
inhibitory neurons (VIP)
inhibitory neurons (PV)
#Cells
48 individuals
g
cell-type-stratified
cell-type-combined
CoCoA (δ)
~ pathoAD
pseudo-bulk ~ pathoAD
number of unique genes discovered
p−value cutoff
h
CoCoA (δ)
~ pathoAD
pseudo-bulk ~ pathoAD
number of genes discovered
p−value cutoff
excitatory neurons (L5/6,CC)
microglia
inhibitory neurons (SST)
oligodendrocyte progenitor cells
astrocytes
inhibitory neurons (SV2C)
inhibitory neurons (VIP)
oligodendrocytes
inhibitory neurons (PV)
excitatory neurons (L5/6)
excitatory neurons (L2/3)
excitatory neurons (L4)
astrocytes
excitatory neurons (L2/3)
excitatory neurons (L4)
inhibitory neurons (SV2C)
inhibitory neurons (VIP)
microglia
excitatory neurons (L5/6)
inhibitory neurons (PV)
inhibitory neurons (SST)
oligodendrocytes
oligodendrocyte progenitor cells
i
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y
Average Disease Effect
SRGAP2C
ARHGAP15
SH3RF3
UBE2E2
ST6GAL1
MBNL1
LHFPL2
HDAC9
BLNK
PTPRJ
ETV6
FGD4
ZNF609
ZFHX3
APOE
SYNDIG1
RUNX1
SH3RF3
7.4e−07
LHFPL2
5.5e−08
BLNK
7e−06
PTPRJ
3.2e−08
APOE
5.2e−07
disease effect
individuals
HC
AD
sqrt(tot)

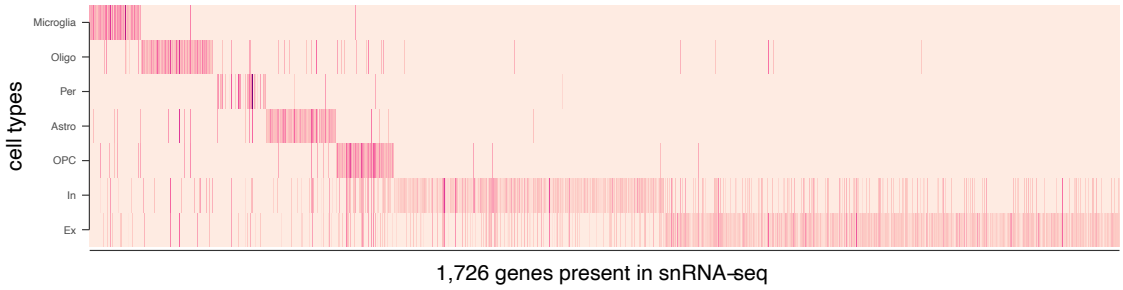
Figure 4.

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.


a
#genes (FDR < 0.2)
3684
3333
2562
1607
14
301
749
778
381
694
384
465
excitatory (L2/3)
excitatory (L4)
excitatory (L5/6)
excitatory (L5/6.CC)
inhibitory (PV)
inhibitory (SST)
inhibitory (SV2C)
inhibitory (VIP)
OPC
oligodendrocytes
microglia
astrocytes
b
pyruvate metabolic process
cerebral cortex development
regulation of RNA splicing
regulation of protein localization to membrane
regulation of mRNA processing
positive regulation of protein localization to membrane
peptidyl−lysine modification
organelle localization by membrane tethering
membrane docking
cytoskeleton−dependent intracellular transport
ciliary basal body−plasma membrane docking
regulation of cation channel activity
phosphatidylinositol metabolic process
c
synaptic vesicle transport
synaptic vesicle localization
synaptic vesicle exocytosis
signal release from synapse
regulation of synaptic vesicle cycle
regulation of regulated secretory pathway
regulation of neurotransmitter transport
regulation of calcium ion−dependent exocytosis
positive regulation of neuron projection development
positive regulation of cell projection organization
neurotransmitter secretion
establishment of synaptic vesicle localization
calcium ion regulated exocytosis
regulation of postsynaptic membrane potential
apoptotic mitochondrial changes
adenylate cyclase−modulating G protein−coupled receptor signaling pathway
synapse assembly
ribonucleoside diphosphate metabolic process
regulation of trans−synaptic signaling
regulation of synapse structure or activity
pyridine nucleotide biosynthetic process
purine ribonucleoside diphosphate metabolic process
purine nucleoside diphosphate metabolic process
nucleoside diphosphate phosphorylation
nicotinamide nucleotide biosynthetic process
modulation of chemical synaptic transmission
axon development
ADP metabolic process
regulation of synapse organization
regulation of dendrite development
postsynapse organization
developmental maturation
dendrite morphogenesis
dendrite development
central nervous system neuron differentiation
d
viral entry into host cell
positive regulation of translation
positive regulation of cellular amide metabolic process
negative regulation of nervous system development
negative regulation of multi−organism process
negative regulation of cell development
myelination
establishment of protein localization to membrane
entry into other organism involved in symbiotic interaction
entry into host cell
entry into host
entry into cell of other organism involved in symbiotic interaction
ensheathment of neurons
axon ensheathment
translational initiation
SRP−dependent cotranslational protein targeting to membrane
regulation of lymphocyte mediated immunity
regulation of GTPase activity
protein targeting to ER
positive regulation of leukocyte mediated immunity
positive regulation of GTPase activity
negative regulation of immune response
negative regulation of immune effector process
interferon−gamma−mediated signaling pathway
establishment of protein localization to endoplasmic reticulum
cotranslational protein targeting to membrane
cellular response to interferon−gamma
cell killing
tissue migration
regulation of epithelial cell migration
negative regulation of neurogenesis
epithelium migration
epithelial cell migration
cell junction organization
e
spindle pole
spindle
microtubule associated complex
microtubule
cytoplasmic region
site of polarized growth
neuron projection terminus
growth cone
glutamatergic synapse
distal axon
axon part
presynapse
synaptic membrane
presynaptic membrane
postsynaptic specialization
postsynaptic membrane
postsynaptic density
neuron to neuron synapse
neuron spine
intrinsic component of synaptic membrane
integral component of synaptic membrane
dendritic spine
cation channel complex
asymmetric synapse
cell−cell junction
ribosomal subunit
recycling endosome membrane
phagocytic vesicle membrane
phagocytic vesicle
Golgi−associated vesicle membrane
focal adhesion
ficolin−1−rich granule
ER to Golgi transport vesicle membrane
endocytic vesicle membrane
early endosome membrane
cytosolic ribosome
cytosolic part
COPII−coated ER to Golgi transport vesicle
coated vesicle membrane
cell−substrate junction
cell−substrate adherens junction
cell cortex
actin cytoskeleton
f
transferase activity, transferring acyl groups other than amino−acyl groups
microtubule binding
protein serine/threonine kinase activity
phospholipid binding
nucleoside binding
ion channel binding
GTPase activity
ATPase activity, coupled
tubulin binding
channel regulator activity
enzyme activator activity
voltage−gated cation channel activity
substrate−specific channel activity
passive transmembrane transporter activity
metal ion transmembrane transporter activity
ion channel activity
inorganic cation transmembrane transporter activity
channel activity
cation channel activity
calcium channel activity
structural constituent of ribosome
Rho GTPase binding
Ras guanyl−nucleotide exchange factor activity
Ras GTPase binding
peptide binding
histone deacetylase binding
guanyl−nucleotide exchange factor activity
GTPase activator activity
cadherin binding
antigen binding
amide binding
Ex−L2/3
Ex−L4
Ex−L5/6
Ex−L5or6−CC
In−PV
In−SST
In−SV2C
In−VIP
OPC
Oligo
Microglia
Astro
p−value
1e+00
1e−02
1e−04
1e−06
1e−08
1e−10
#genes
100
32
10
3
1

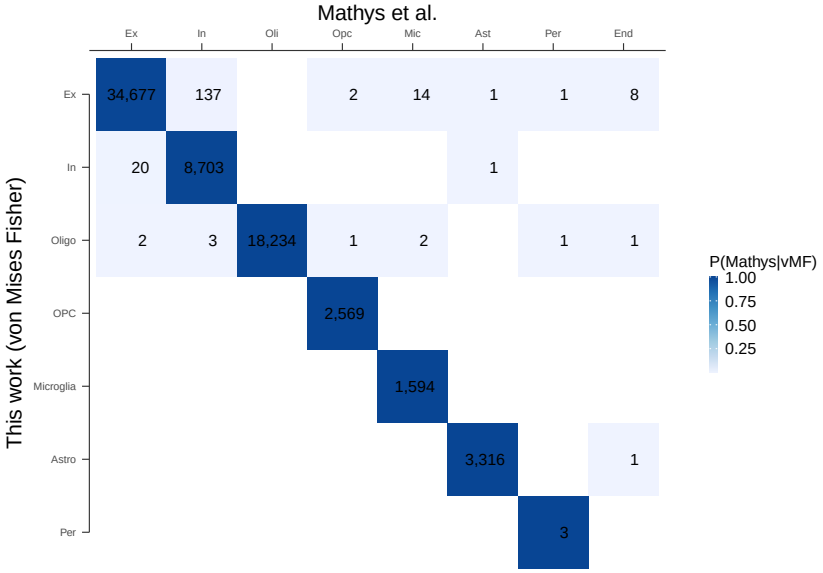
Supplementary Figure 1.

Average cell-type-specific profiles of 1,726 marker genes



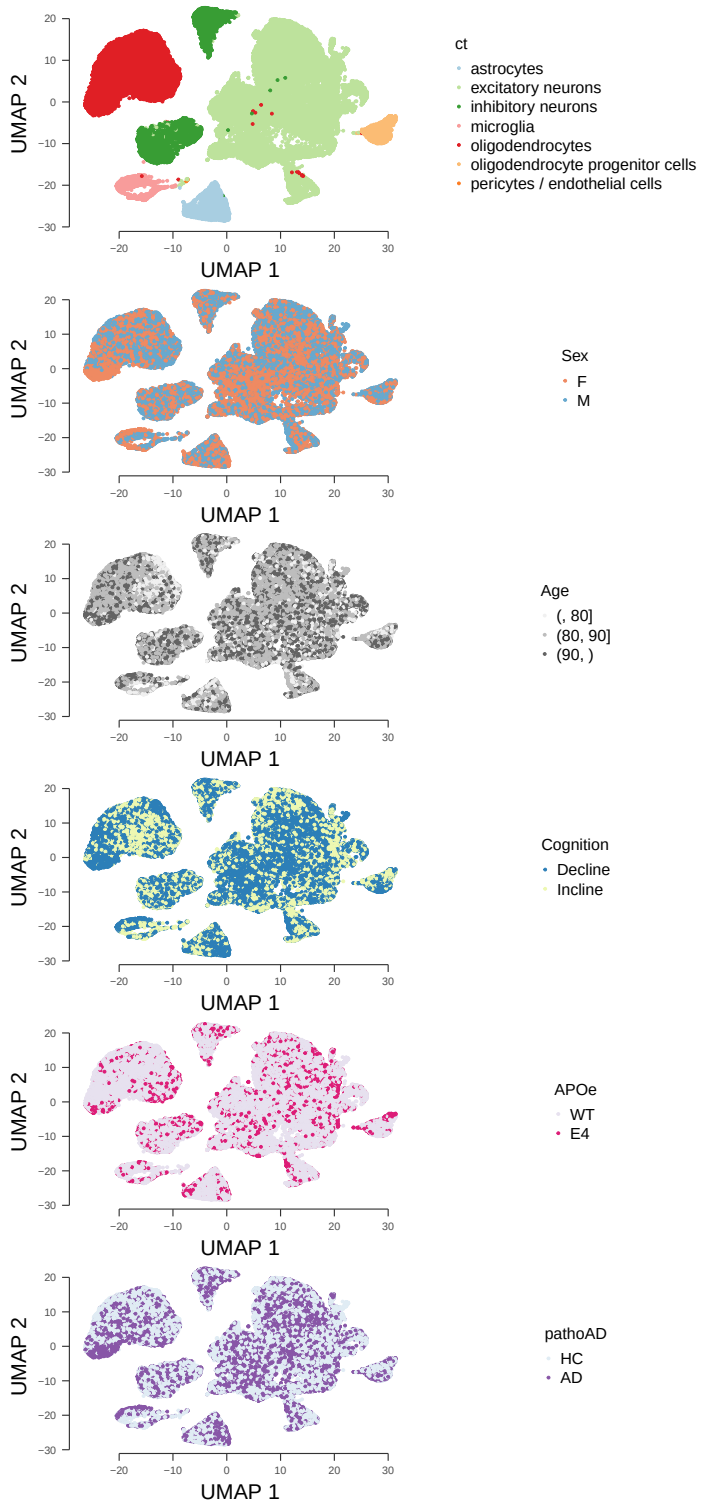
Supplementary Figure 2.

Correspondence with Mathys et al.[22]



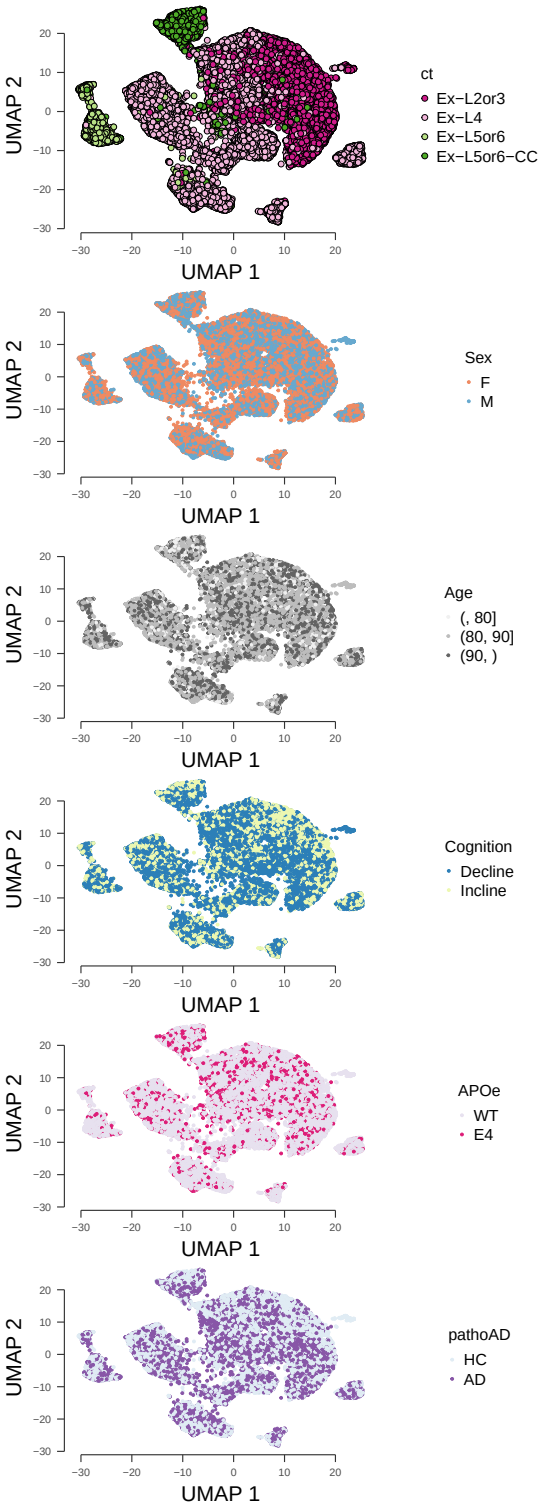
Supplementary Figure 3.

The annotations of the major neuronal and glial cell types are not biased by known biological variables.



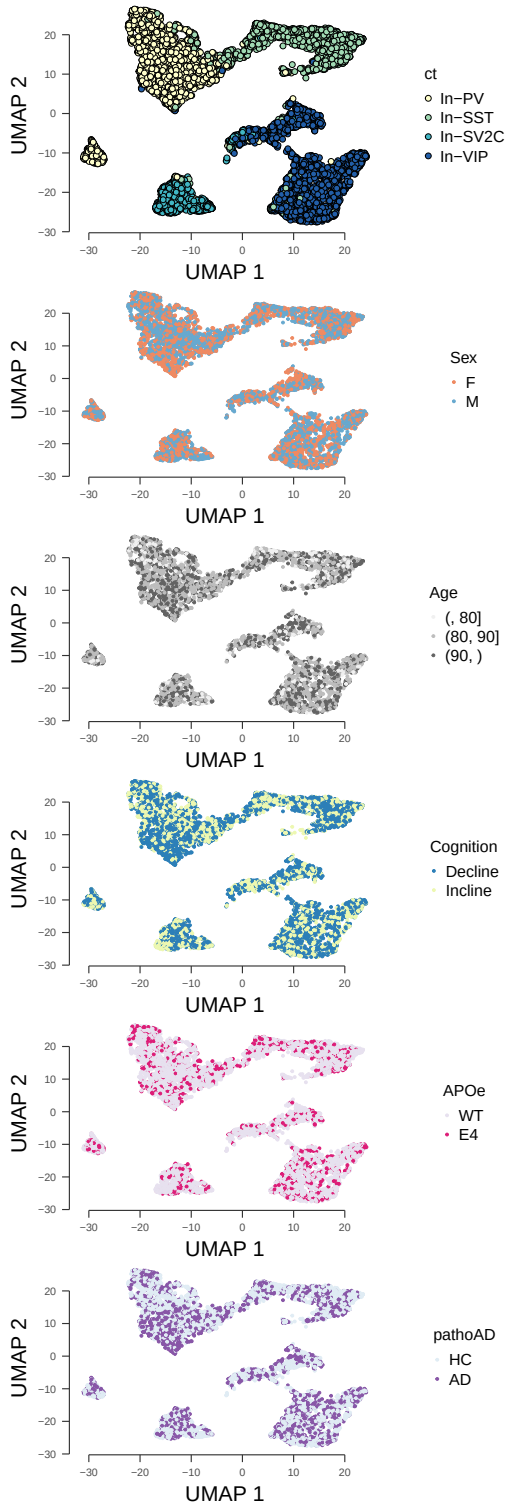
Supplementary Figure 4.

The annotations of the excitatory neuron types are not biased by known biological variables.



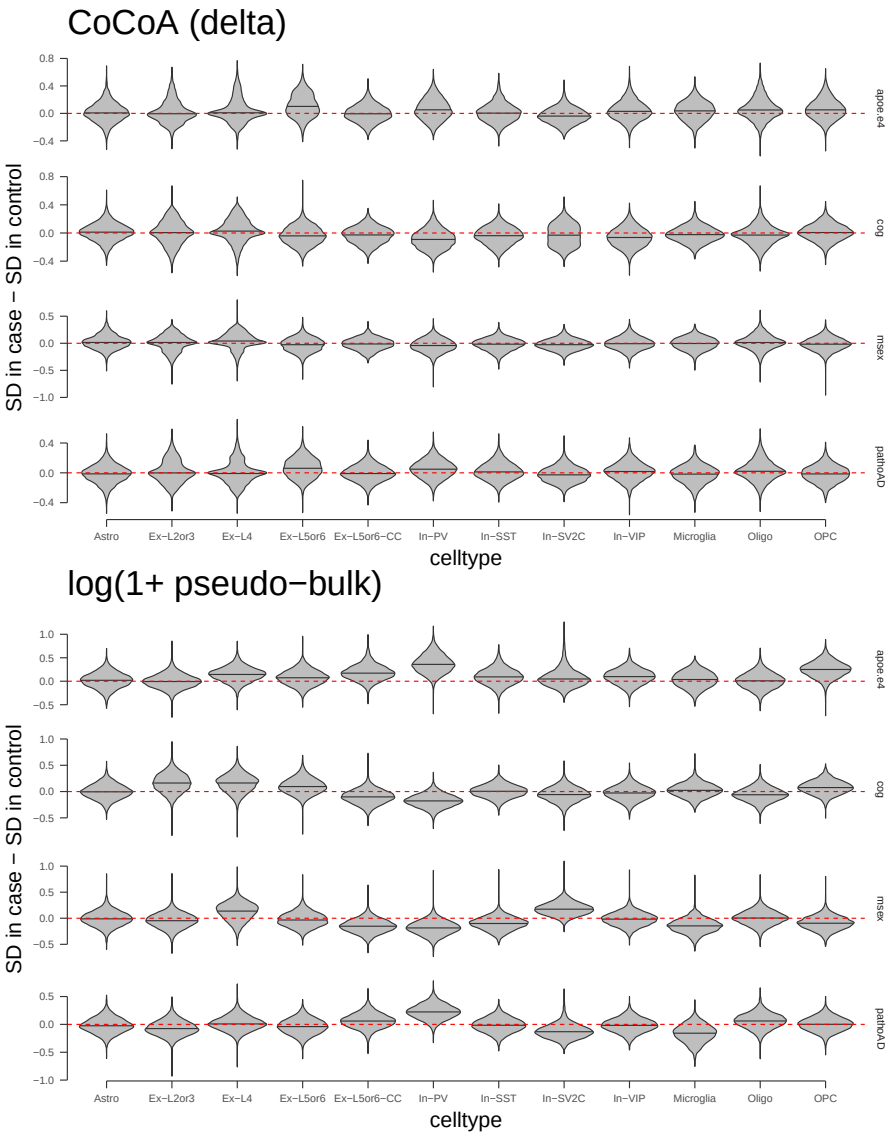
Supplementary Figure 5.

The annotations of the inhibitory neuron types are not biased by known biological variables.



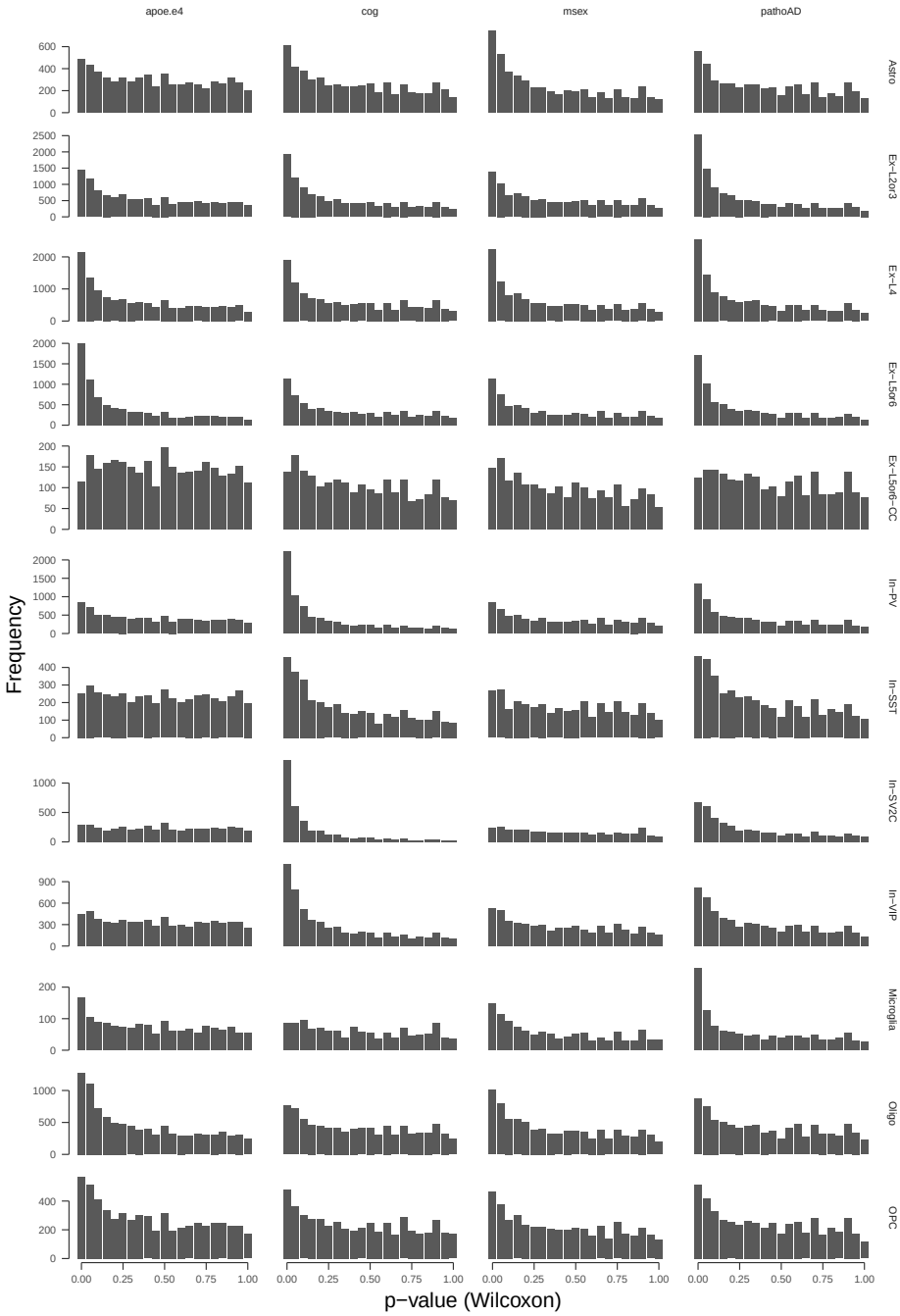
Supplementary Figure 6.

CoCoA algorithm does not create a skewed distribution of variance.



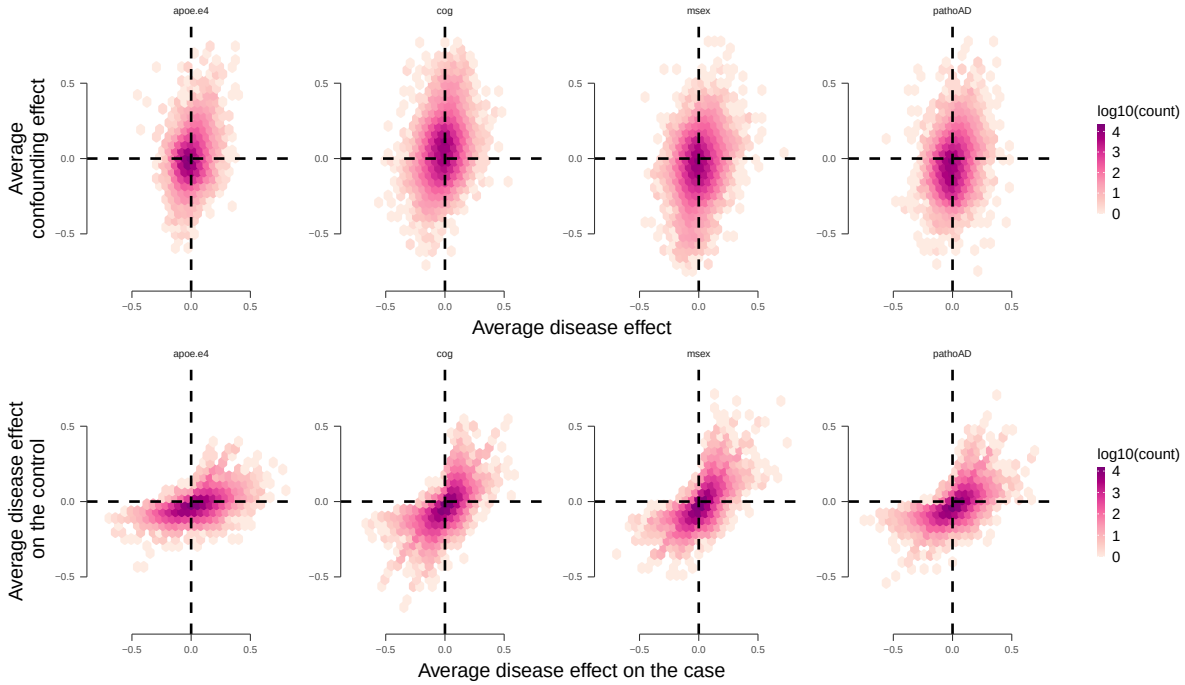
Supplementary Figure 7.

Histogram of p-value distributions



Supplementary Figure 8.

Top: Correlations between the average disease effects (ADE) and gene-level associations with the confounding factors. **Bottom**: Correlations between the average disease effects computed on the disease cohort (ADD) and the average disease effects computed on the control cohort (ADC).



Reference

1. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nat. Biotechnol.* **32,** 381–386 (2014).

2. Ståhl, P. L. *et al.* Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353,** 78–82 (2016).

3. Norman, T. M. *et al.* Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science* **365,** 786–793 (2019).

4. Wijst, M. G. P. van der *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* (2018).

5. Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* **15,** e1008045 (2019).

6. Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 5116–5121 (2001).

7. Irizarry, R. A. *et al.* Multiple-laboratory comparison of microarray platforms. *Nat. Methods* **2,** 345–350 (2005).

8. Irizarry, R. A. *et al.* Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* **4,** 249–264 (2003).

9. Robinson, M. D. & Smyth, G. K. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics* **23,** 2881–2887 (2007).

10. Robinson, M. D. & Smyth, G. K. Small-sample estimation of negative binomial dispersion, with applications to SAGE data. *Biostatistics* **9,** 321–332 (2008).

11. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).

12. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

13. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15,** 1–21 (2014).

14. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15,** 255–261 (2018).

15. Crowell, H. L. *et al.* On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. *bioRxiv* 713412 (2019).

16. Mandric, I. *et al.* Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* **11,** 1–9 (2020).

17. Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *aos* **6,** 34–58 (1978).

18. Rosenbaum, P. R. & Rubin, D. B. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am. Stat.* **39,** 33–38 (1985).

19. Heckman, J. J., Ichimura, H. & Todd, P. E. Matching as an econometric evaluation estimator: Evidence from evaluating a job training programme. *Rev. Econ. Stud.* **64,** 605–654 (1997).

20. Abadie, A. & Imbens, G. W. Large sample properties of matching estimators for average treatment effects. *Econometrica* **74,** 235–267 (2006).

21. Davey Smith, G. & Hemani, G. Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Hum. Mol. Genet.* **23,** R89–98 (2014).

22. Mathys, H. *et al.* Single-cell transcriptomic analysis of alzheimer's disease. *Nature* **570,** 332–337 (2019).

23. VanderWeele, T. J. Bias formulas for sensitivity analysis for direct and indirect effects. *Epidemiology* **21,** 540–551 (2010).

24. Glynn, A. N. The product and difference fallacies for indirect effects: The product and difference fallacies for indirect effects. *Am. J. Pol. Sci.* **56,** 257–269 (2012).

25. Pearl, J. & Mackenzie, D. *The book of why: The new science of cause and effect.* (Basic Books, 2018).

26. VanderWeele, T. J. & Shpitser, I. On the definition of a confounder. *Ann. Stat.* **41,** 196–220 (2013).

27. Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin* **1,** 80–83 (1945).

28. Andri et mult. al., S. DescTools: Tools for descriptive statistics. (2021).

29. Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362,** (2018).

30. Polański, K. *et al.* BBKNN: Fast batch alignment of single cell transcriptomes. *Bioinformatics* (2019).

31. Velmeshev, D. *et al.* Single-cell genomics identifies cell type-specific molecular changes in autism. *Science* **364,** 685–689 (2019).

32. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57,** 289–300 (1995).

33. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. Stat. Theory Appl.* **6,** 65–70 (1979).

34. Stephens, M. False discovery rates: a new deal. *Biostatistics* **18,** 275–294 (2017).

35. Lee, J. H. *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.* **72,** 1043–1051 (2015).

36. Hill-Burns, E. M. *et al.* Identification of genetic modifiers of age-at-onset for familial parkinson's disease. *Hum. Mol. Genet.* **25,** 3849–3862 (2016).

37. Zajkowicz, A. *et al.* The alzheimer's disease-associated TREM2 gene is regulated by p53 tumor suppressor protein. *Neurosci. Lett.* **681,** 62–67 (2018).

38. Sierksma, A. *et al.* Novel alzheimer risk genes determine the microglia response to amyloid-β but not to TAU pathology. *EMBO Mol. Med.* **12,** e10606 (2020).

39. Lutz, M. W., Sprague, D., Barrera, J. & Chiba-Falek, O. Shared genetic etiology underlying alzheimer's disease and major depressive disorder. *Transl. Psychiatry* **10,** 88 (2020).

40. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11,** R14 (2010).

41. Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nat. Biotechnol.* **32,** 896–902 (2014).

42. Gagnon-Bartsch, J. A. & Speed, T. P. Using control genes to correct for unwanted variation in microarray data. *Biostatistics* **13,** 539–552 (2012).

43. Schölkopf, B. *et al.* Modeling confounding by half-sibling regression. *Proc. Natl. Acad. Sci. U. S. A.* **113,** 7391–7398 (2016).

44. Abid, A., Zhang, M. J., Bagaria, V. K. & Zou, J. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nat. Commun.* **9,** 2134 (2018).

45. Louizos, C., Swersky, K., Li, Y., Welling, M. & Zemel, R. The variational fair autoencoder. (2015). at <http://arxiv.org/abs/1511.00830>

46. Townes, F. W., Hicks, S. C., Aryee, M. J. & Irizarry, R. A. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biol.* **20,** 295 (2019).

47. Sarkar, A. & Stephens, M. Separating measurement and expression models clarifies confusion in single cell RNA-seq analysis. *bioRxiv* 2020.04.07.030007 (2020).

48. Jordan, M. I., Jaakkola, T. S. & Saul, L. K. An Introduction to Variational Methods for Graphical Models. *Mach. Learn.* **37,** 183–233 (1999).

49. Hirano, K., Imbens, G. W. & Ridder, G. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* **71,** 1161–1189 (2003).

50. 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>

51. Friedman, J., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33,** 1–22 (2010).

52. Haghverdi, L., Lun, A. T. L., Morgan, M. D. & Marioni, J. C. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat. Biotechnol.* **36,** 421–427 (2018).

53. Hie, B., Bryson, B. & Berger, B. Efficient integration of heterogeneous single-cell transcriptomes using scanorama. *Nat. Biotechnol.* **37,** 685–691 (2019).

54. Banerjee, A., Dhillon, I. S., Ghosh, J. & Sra, S. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Mach. Learn. Res.* **6,** 1345–1382 (2005).