Counterfactual inference for single-cell gene expression analysis

04:40:31 PM, Jan 12, 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, USA

Contact:

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

Abstract

Finding a causal gene from case-control studies is a classic and fundamental problem in genomics. To date, we still ask which genes are differentially regulated by a disease with single-cell sequencing data, but in a cell-type-specific way. Here, we present a causal inference framework that effectively adjusts confounding effects, not requiring prior knowledge of control genes or cells. We demonstrate that our causal inference algorithm substantially improves statistical power in simulations and real-world data analysis of 70k brain cells, collected for dissecting Alzheimer's disease (AD) mechanisms. We identified that 377 causal genes are differentially regulated by the disease in various brain cell types, including highly-relevant AD genes with a proper cell type annotation, such

as *DGKD* in neurons, *SNCA* in microglia, *PIAS* in oligodendrocyte progenitor cells, and *FGFR2* in astrocytes. Causal genes in different cell types also enrich distinctive pathways, highlighting multiple components of the disease progressions.

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 create "batch" effects, although the underlying individu-

als are mostly exchangeable. The number of individuals determines the statistical power of DEG discovery in single-cell data. Although we may obtain gene expressions in millions of cells, we often have tens or a hundred phenotyped individuals. 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 across heterogeneous individuals. We establish our causal claims in differential expression analysis based on Rubin's potential outcome framework.[17] Our method is inspired by the seminary work of econometric data analysis.[18,19] We highlight 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

CoCoA for single-cell differential expression analysis

We obtain confounder-adjusted pseudo-bulk data to improve the power of differential expression analysis. We term our approach "CoCoA," short for Counterfactual confounder adjustment. The CoCoA algorithm consists of three steps (Fig. 1A): (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.

As an example, we illustrate the effectiveness of our approach in the case of *APOE* gene measured in microglia samples (Fig. 1B). In pseudo-bulk (PB) data analysis with 40 individuals, it appears that the average expression values are only mildly up-regulated in disease subjects (Wilcoxon $p = 0.13$) even though *APOE* over-expression is one of the most frequently-observed factors of Alzheimer's disease (AD). Interestingly, we found that the shared confounding factor between the case and control shows a marginally-significant correlation with Sex (female versus male; $p = 0.016$). After adjusting the confounders on the data, we recover a significant correlation of *APOE* gene expressions with AD status ($p = 0.0024$). Note that we had not exploited the prior knowledge regarding the sex-bias of *APOE* expression,[20] but it was merely a serendipitous discovery of our causal analysis. One may ask whether increasing the sample size can help. We disagree that since we view the underlying problem is largely a causal inference problem, not a statistical issue.[21]

To evaluate our approach more extensively, we simulated single-cell data matrix (Y) of 5,000 genes with 40 individuals (Figure 1C). We stochastically generated 50 cells per individual, multiplying

the individual-level average expressions with random sequencing depth (ρ). Of the 5,000 genes, we randomly selected 50 causal genes and varied the effect size of disease status (W), so that the disease label explains from 10% to 40% of total expression variation. For simplicity, we fixed the variation of covariates to 50%. Due to the confoundedness of the disease W and expression Y by the covariate X, the covariate X also accounts of the variability of expressions from 10% to 30%.

We repeated our experiments 30 times for all the different simulation settings and summarized the performance by F1 score of the precision-recall curve (Fig. 1D). We compared CoCoA with three different PB methods:

1. Bayesian quantification: fitting Poisson-Gamma hierarchical model without confounding factor adjustment (i.e., δ without the step 1 and step 2).
2. Average: arithmetic mean of cell-level expressions within each individual
3. Total: summation of cell-level expressions within each individual

As expected, the performance of the association methods on PB data degrades without confounder adjustment, and the gap between CoCoA and the other PB methods widens as we introduce a higher degree of confoundedness (Fig. 1D). We also measured statistical power at false discovery rate (FDR) 10%, and the same trends were observed (Supplementary Fig. 1A,B). Moreover, we generated null data without a causal gene and confirmed that an empirical FDR method (Benjamini-Hochberg[22]) controls actual FDR at the desired levels (Supplementary Fig. 1C).

Case study: Finding genes causally-affected by Alzheimer's disease

We reanalyzed published single-nuclei RNA-seq (snRNA-seq) data of 48 individuals in postmortem brain samples.[23] 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 8 samples.

Cell type annotation for 70,634 cells

We directly annotated the cell types of these 70,634 cells using the list of cell-type-specific genes provided by PsychENCODE project.[24] Of the total 2,648 PsychENCODE marker genes, we used 1,726 genes expressed in this data set as features without using an additional unsupervised manifold learning method. 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 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 strongly agrees with the original paper's cell type annotation (Fig. 2B).

We notice a wide spectrum of cell-type variability across 48 individuals, both in terms of number of cells and proportions (Fig. 2A). The Mathys *et al.* data contain on average 1,472 cells per individual, of which 50.9% cells stem from Ex (N=745 $\pm$ 388 SD), 12.23% cells from In (N=179 $\pm$ 103), 3.54% cells from OPC (N= 54 $\pm$ 35), 25.22% cells for Oligo (N=380 $\pm$ 252), 2.61% cells from Microglia (N= 36 $\pm$ 25), 4.95% cells from Astro (N= 71 $\pm$ 47), 0.33% cells from Endo (N= 4), and 0.34% cells for Per (N= 4).

Cell-type-sorted data improve statistical power and interpretation

Next, 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-sorted 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-sorted data analysis will improve statistical power and render better biological interpretations in genomics analysis.

Using these cell type annotations, we constructed cell-type-sorted PB 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 PB data by merging all the cells in each individual, ignoring the cell type annotations, and carried out the same association analysis (Fig. 2C). We found that cell-type-sorting substantially boost statistical power for several AD GWAS genes[25] in relevant cell types, including *APOE* gene (p=0.003 in microglia versus p=0.25 in bulk), *PICALM* (p=3.16 $\times 10^{-5}$ in astrocyte versus p=0.36 in bulk), *IQCK* (p=0.01 in OPC versus p=0.33 in bulk), and *ADAM10* (p=0.002 in inhibitory neurons versus p=0.27 in bulk). We carried out the same type of analysis with the PB data aggregated by total and average profiles and found that cell-type sorting substantially improves gene-level associations (Supplementary Fig. 2A and 2B). However, most cell-type-specific discoveries were skewed toward abundant cell types without the confounder adjustment steps. For instance, we may only discover genes associated in excitatory neurons in associations with total expressions (Supplementary Fig. 2B), whereas many genes are significantly associated in glial cell types after the confounder adjustment (Fig. 2C).

Disease status modulates the cell-type-specific gene expressions

This work aims to identify causal genes modulated by the disease in a relevant cell type (Fig. 2D). Under this causal assumption, the stratification procedure for cell types provides a legitimate procedure to control cell-type biases that may impact identifying DEGs. Here, we reject a simplified view that a complex (polygenic) disease can bring systematic changes in cellular environments, and the disease phenotype is solely mediated by cell type fractions. We modestly consider that cell type information is useful to infer cellular environments of disease progression and other related mechanisms, but not beyond that. We found no apparent correlation between the disease and cell types (Supplementary Fig. 2C; Wilcoxon p-value $> .4$). It is hard to consider the cell type fractions act as a mediator variable; thus, we think there is little possibility of "mediation fallacy.[21,26,27]" Moreover, in many cell-type marker genes, the confounding effects estimated by CoCoA (μ_g in Fig. 1A) are marginally-associated with AD (Supplementary Fig. 2D), which provides another reason for justifying the cell type stratification steps.

377 genes are differentially-regulated with AD

Using our CoCoA pipeline (Methods, Fig. 1A), we found that a total of 377 genes (2.72% of total 13,848 genes) are significantly associated with AD, controlling FDR at 1%, which include 123 genes found in excitatory neurons, 36 found in inhibitory neurons, 86 found in oligodendrocytes, 58 found in OPC, 24 found in microglia, and 57 found in Astrocytes. However, we found no transcriptome-wide significant associations without the confounder correction. We found the directions of association effect sizes remain mostly unchanged with or without CoCoA (Supplementary Fig. 3A). However, we were able to confirm that CoCoA identified unwanted variation of gene-level effects correlated with the marginal associations (Supplementary Fig. 3B). Removing these unwanted effects, we made the associations with residual effects (δ) orthogonal with respect to the associations with putative confounding factors (Supplementary Fig. 3C). CoCoA also qualitatively better-calibrates the p-value histogram for insignificant p-values (Supplementary Fig. 4).

Of the 377 genes, we found 28 associations stand significant at stringent family-wise error rate 1% (Fig. 3B, Tab. 1), which break down into 5 genes found in excitatory neurons, 4 found in inhibitory neurons, 6 found in oligodendrocytes, 6 found in OPC, 2 found in microglia, and 5 found in astrocytes. We highlight the following four examples (Fig. 3A): We found *DGKD*, a subtype of DGK (diacylglycerol kinase), strongly associated with AD in excitatory neurons. It is well-known that DKG genes are highly expressed in brain[28] and regulate the metabolism of lipid neurotransmitters[29]. Interestingly, the concentration of diacynglycerol (DAG) molecules in brain is found to be elevated among subjects suffered from mild cognitive impairment[30], AD patients[31], and in the frontal cortex of subjects with Parkinson's disease (PD) and Lewy body dementia (LBD)[32].

SNCA found in the microglia analysis is particularly relevant to AD phenotypes. In some studies, soluble α-synuclein (α-Syn) levels (protein) worked better as a biomarker for cognitive impairment than amyloid-β and tau proteins among AD brains without LBD phenotypes[33]. Several genetic association studies successfully demonstrate that genetic variants located in the *SNCA* region are robustly associated with AD and LBD pathology[34,35]. In the epigenetic study of immune cells (leukocytes), both mRNA expression and DNA methylation of *SNCA* were strongly associated with AD[36].

Another gene, *PIAS2*, found in OPC, also corroborates the causality of *SNCA* and related pathways. *PIAS2* regulates post-translational modification of α-Syn in the postmortem brains of PD and LBD.[37,38] Moreover, along with PIAS1, this gene is thought to play an essential role in autoimmune, neurological disorders, such as multiple sclerosis[39]. In rat hippocampus, PIAS1 expression was increased by amyloid-β induction[40].

We also found considerable experimental evidence for *FGFR2* (fibroblast growth factor receptor 2) associated with neurodegeneration, manifested in astrocytes and neurons. *FGFR* genes, including *FGFR2*, are crucial for differentiation and maturation to astrocyte[41–43]. When neurons are damaged, FGF pathways turn on and help attract other glial cells for phagocytosis[44]. A recent study suggested that down-regulated *Negr1-Fgfr2* activity exhibits cellular phenotypes of autism spectrum disorders in mouse brains[45].

For the readers interested in other significant genes (FDR $< 1\%$), we refer to Supplementary Fig. 5 for their distributions in the genomic axis. Among those, we also highlight top five genes found in each cell type (Supplementary Fig. 6 to 11).

Lastly, we sought to characterize biological mechanisms potentially perturbed by the significant AD genes (FDR $< 5\%$) using gene ontology (GO) annotations of *Homo Sapiens* genes downloaded from `biomaRt`[46] database (accessed at 18:56 UTC December 23, 2020). We tested gene set enrichment analysis of DEGs up-regulated by AD (956 genes; Fig. 4A, Supplementary Tab. 1) and DEGs down-regulated by AD (1,044 genes; Fig. 4B, Supplementary Tab. 2) by multiple hypergeometric tests in each cell type independently. To avoid numerical artefacts induced by GO keywords' size, we focused on GO keywords with at least five genes, but no more than 500 genes.

Gene set analysis shows that DEGs identified in different cell types indeed influence on markedly different biological mechanisms. By visual inspection, we can identify cell-type-specific modules of the enriched GO keywords (Fig. 4). For instance, up-regulated AD genes found in excitatory neurons are highly enriched in neurodevelopmental pathways and relevant cellular components, such as 'neuron projection,' 'chemical synaptic transmission,' 'axon,' and 'dendrite.' However, microglial DEGs are mostly found in immune-related activities, such as 'viral transcription' and inflammatory response,' and synaptic pruning. Cell-cell communication, such as 'focal adhesion,' and signalling

pathways are widely over-represented across many cell types. Additionally, astrocyte-specific DEGs are found in energy and metabolism-related categories.

We acknowledge that it is hard to assess the overall impact of DEGs to the enriched pathways solely from gene set analysis. Nonetheless, we expect our list of cell-type-specific DEGs to provide an entry point for further integrative investigations, e.g., graph-diffusion analysis[47–49] on cell-type-specific network data[50].

Discussion

Here, we addressed a subset of a causal inference problem that emerges in disease studies (the red arrows in Fig. 5). We sought to characterize and estimate the average causal effect of genes between the case and control individuals from observational single-cell data. Using CoCoA, we expect a researcher to prioritize disease genes causally-modulated by disease status, not a causal driver gene located in the upstream of disease in the causal model (Fig. 5). To causally identify the upstream driver genes, we would need to conduct mediation analysis, such as Mendelian randomization[51], leveraging genetic information as an instrumental variable[52].

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), genomics community has been using so-called control genes and samples to extract factors shared in both control and discovery data[53–55]. 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[56] could pick out non-causal factors in its latent space. Likewise, only if nuisance variables are independently observed, variational fair autoencoder model[57] project cells onto unconfounded ("fair") latent space.

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[58]. However, future research can take advantages of

more sophisticated modelling of both individual- and cell-level observations[59], 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 CoCoA approach was necessary to reveal both well-established and novel causal genes in AD. To the best of our knowledge, our work is the first application of counterfactual inference to single-cell genomics. 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,[60] $q(\lambda) = \mathsf{Gamma}(\lambda|\alpha_\lambda, \beta_\lambda)$ and $q(\rho) = \mathsf{Gamma}(\rho|\alpha_\rho, \beta_\rho)$. We can quickly reach convergence by alternating the following update equations:

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

Here, we first initialize $\mathbb{E}\left[\rho_j\right] = 1$ 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:[17,19,61] (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.[62] 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:[63]

$$\mathbb{E}\left[\ln Y_{gj}^{(w)}\right] \approx \sum_{j'=1}^{k} F_{gj'}^{(1-w)} \beta_{j'} + \epsilon_g$$

Given the optimized coefficients, we predict the potential outcome $\hat{Y}_{gj}^{(1-w)} \leftarrow \exp\left(\sum_{j'}^{k} F_{gj'}^{(1-w)} \hat{\beta}_{j'}\right)$, ignoring the residual errors (ϵ_g). We also considered a non-parametric imputation method which takes weighted average over the matched cells.[64–66] 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, dented by μ_g for a gene g:

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

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

We estimate the posterior mean of μ_g by variational Bayes by alternating the following update equations until convergence:

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

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

Step 3 (Confounder adjustment)

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

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

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

$$\mathbb{E}\big[\delta_g\big] \leftarrow \frac{1 + \sum_j Y_{gj}}{1 + \hat{\mu}_g \sum_j \mathbb{E}\big[\rho_j\big]}, \quad \mathbb{E}_q\big[\rho_j\big] \leftarrow \frac{1 + \sum_g Y_{gj} + 1}{1 + \sum_g \hat{\mu}_g \mathbb{E}\big[\delta_g\big]}.$$

Since the δ_g variable follows Gamma distribution, we also have $\mathbb{E}_q\big[\ln \delta_g\big] = \psi(1 + \sum_j Y_{gj}) - \log(1 +$

$\hat{\mu}_g \sum_j \mathbb{E}[\rho_j]$) and calibrate the standard deviation by bootstrapping cells in the steps 2 and 3.

Simulation by generalized linear model

We generate single-cell data matrix in two causal trails: Firstly, $X \to W$, we simulate disease assignment W affected by "pretreatment'' covariate X; then, we generate observed counting data with the covariate X and assignment W ($X, W \to \bar{Y} \to Y$). We introduce $\bar{Y}$ to represent individual-level mean values. In simulation experiments, we mainly changed the following three parameters:

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

For 40 individuals ($i \in [40]$), we begin the simulation by generating a real-valued covariate X_i from the standard normal distribution. For disease assignment W_i, we constitute $\mathsf{logit}(W_i) = X_i\alpha + \epsilon_W$, using $\alpha \sim \mathcal{N}(0, \sigma^2_{X\to W})$ and $\epsilon_W \sim \mathcal{N}(0, 1 - \sigma^2_{X\to W})$. For the mean value, we put together X and W: $\log \bar{Y}_{gi} = X_i\beta_g + W_i\tau_g + \epsilon_Y$, where the covariate effect $\beta_g \sim \mathcal{N}(0, \sigma^2_{X\to Y})$, gene-level causal effect $\tau_g \sim \mathcal{N}(0, \sigma^2_{W\to Y})$, and the residuals $\epsilon \sim \mathcal{N}(0, 1 - \sigma^2_{X\to Y} - \sigma^2_{W\to Y})$. Finally, we generate cell-level count data by $Y_{gj} \sim \mathsf{Poisson}\left(Y_{gj} | \bar{Y}_{gi}\rho_j\right)$ with equally-distributed sequencing depth $\rho_j \sim \mathsf{Gamma}(6.25, 6.25)$.

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.[24] Using 1,726 genes present in our data, we construct a normalized vector $\mathbf{m}_j$ for each gene 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[67].

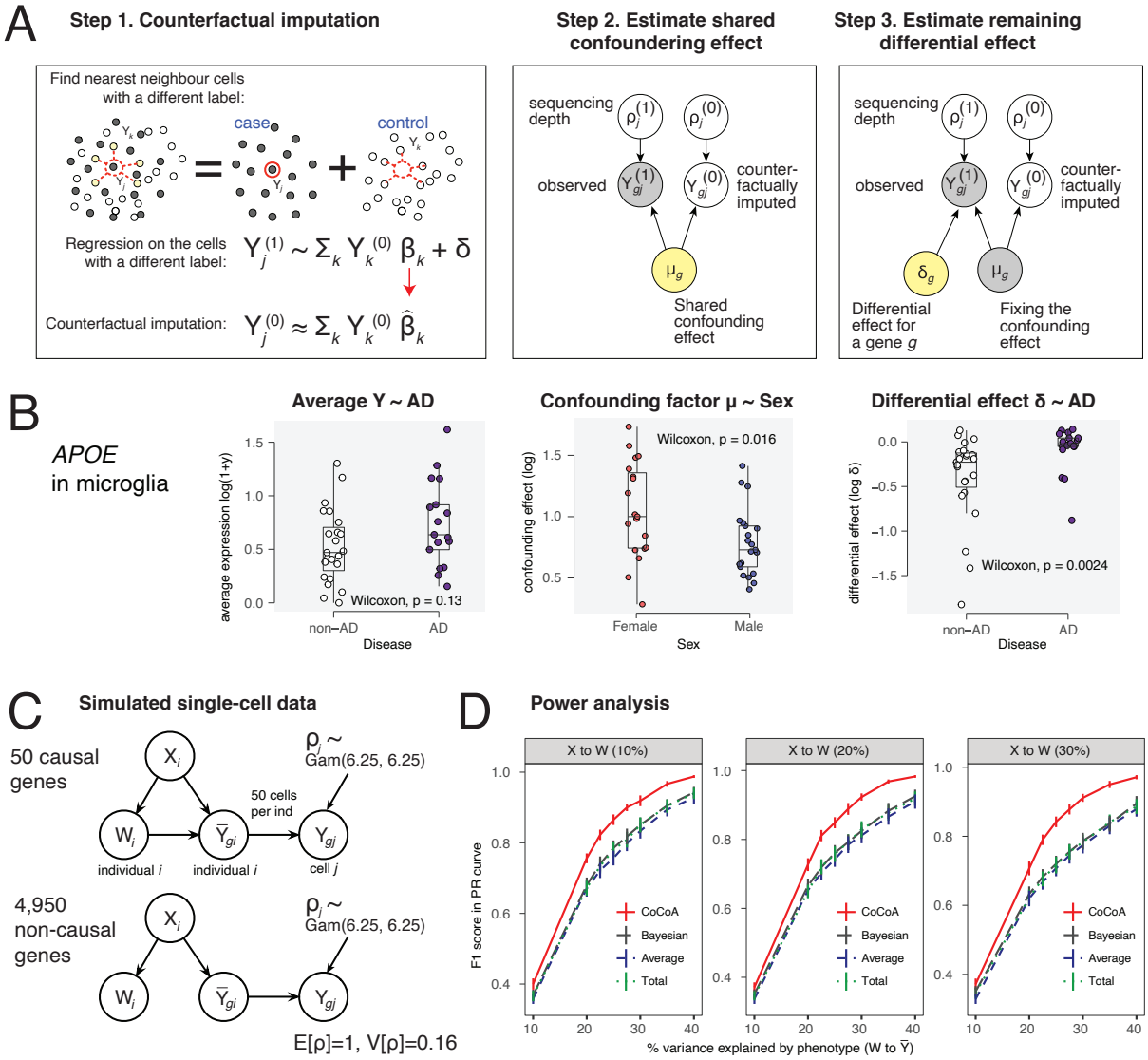
Acknowledgement

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.

The authors declare no competing interests.

Figures

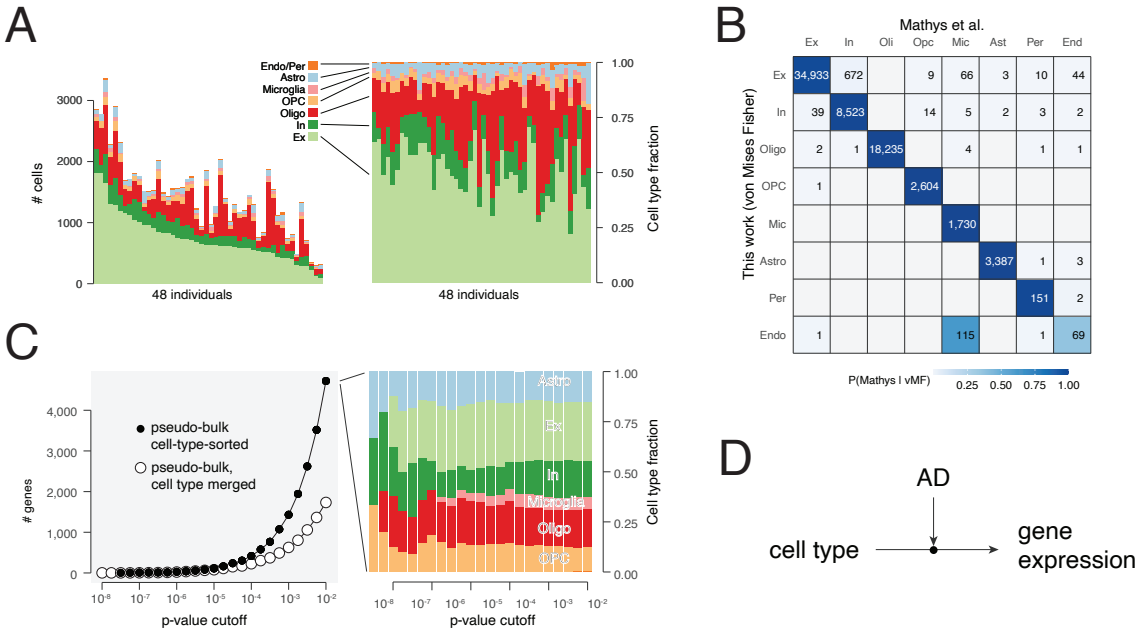
Figure 1.



Counterfactual confounder adjustment (CoCoA) on single-cell expression data. (A) 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. **(B)** Illustration of CoCoA on the *APOE* example. AD: Alzheimer's disease. μ: shared confounding effect; δ: residual differential effect. **(C)** Diagrams for causal and non-causal genes in simulation

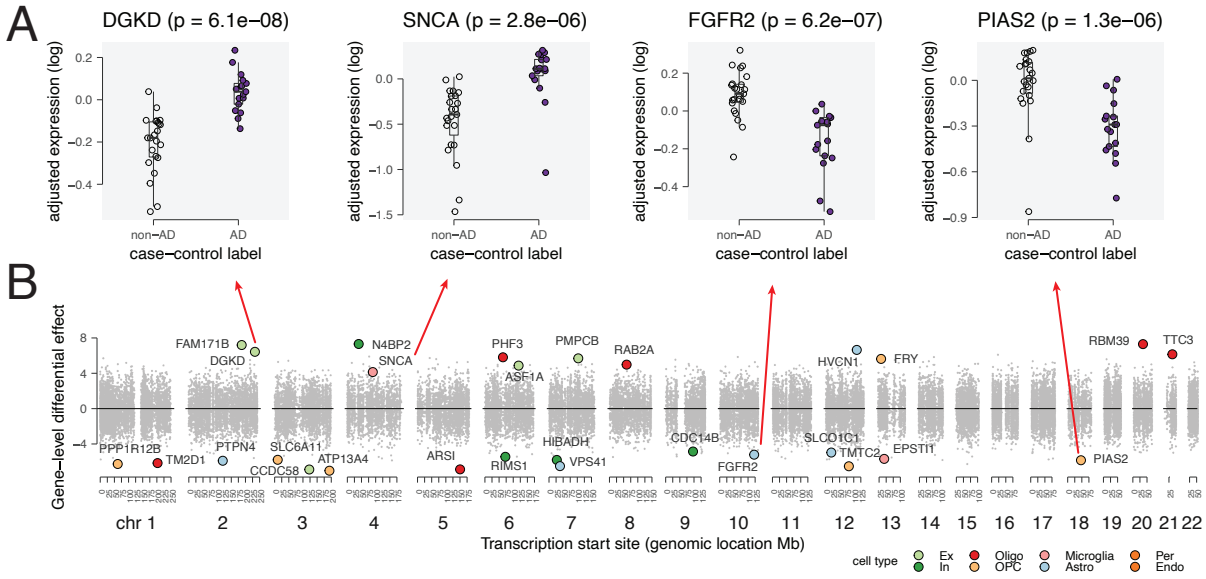
studies (see Methods for details). **(D)** Power analysis comparing CoCoA with other pseudo-bulk methods. *Bayesian*: Bayesian inference for individual-level gene quantification without confounder adjustment. *Average*: average expression across cells within each individual. *Total*: total expression summed over cells within each individual.

Figure 2.



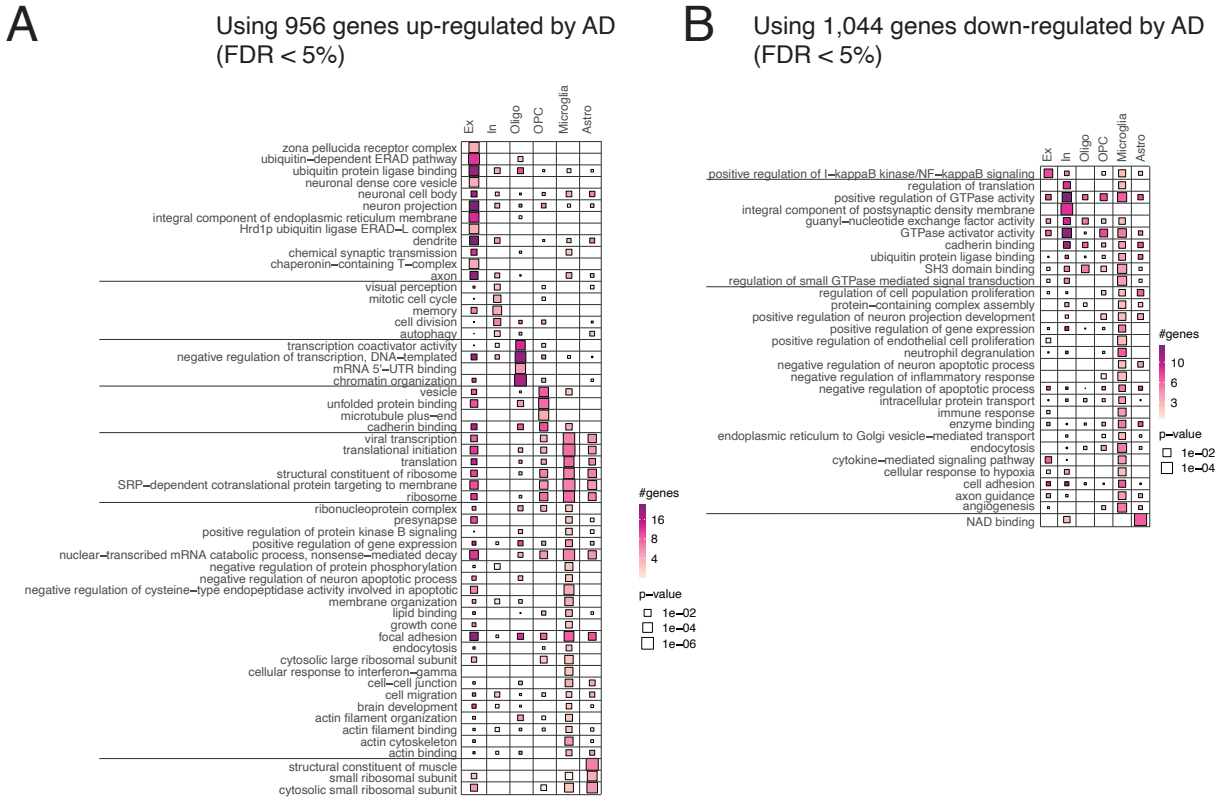
Cell-type-sorted approach improves statistical power and interpretation of differential expression analysis. (A) Cell type decomposition across 48 individuals. *Ex*: expiatory neurons; *In*: inhibitory neurons; *Oligo*: oligodendrocytes; *OPC*: oligodendrocyte progenitor cells; *Mic*: microglia; *Astro*: astrocytes; *Per*: pericytes; *Endo*: endothelial cells. **(B)** Cell type annotations of our von Mises-Fisher mixture model agree with the annotations of the original work[23]. **(C)** Cell-type-sorted CoCoA approach improves statistical power and identifies genes in a wide spectrum of the cell types. **(D)** A causal structural model for cell-type-sorted differential expression analysis.

Figure 3.



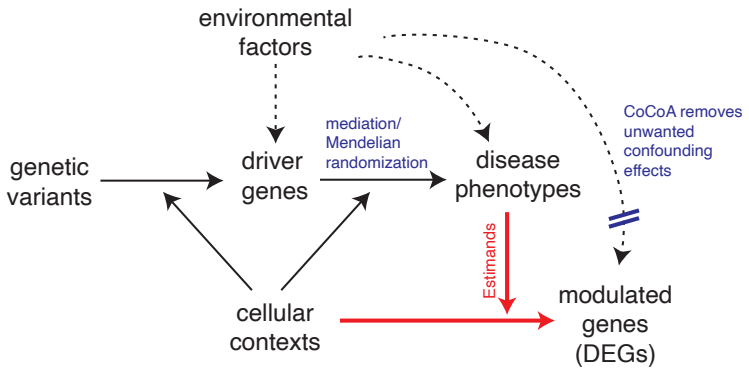
CoCoA identifies causal AD genes from snRNA-seq data. (A) We highlight four genes transcriptome-wide-significantly associated with AD. **(B)** Genomic locations for transcriptome-wide significant 28 genes (FWER $< 1\%$), coloured by cell types.

Figure 4.



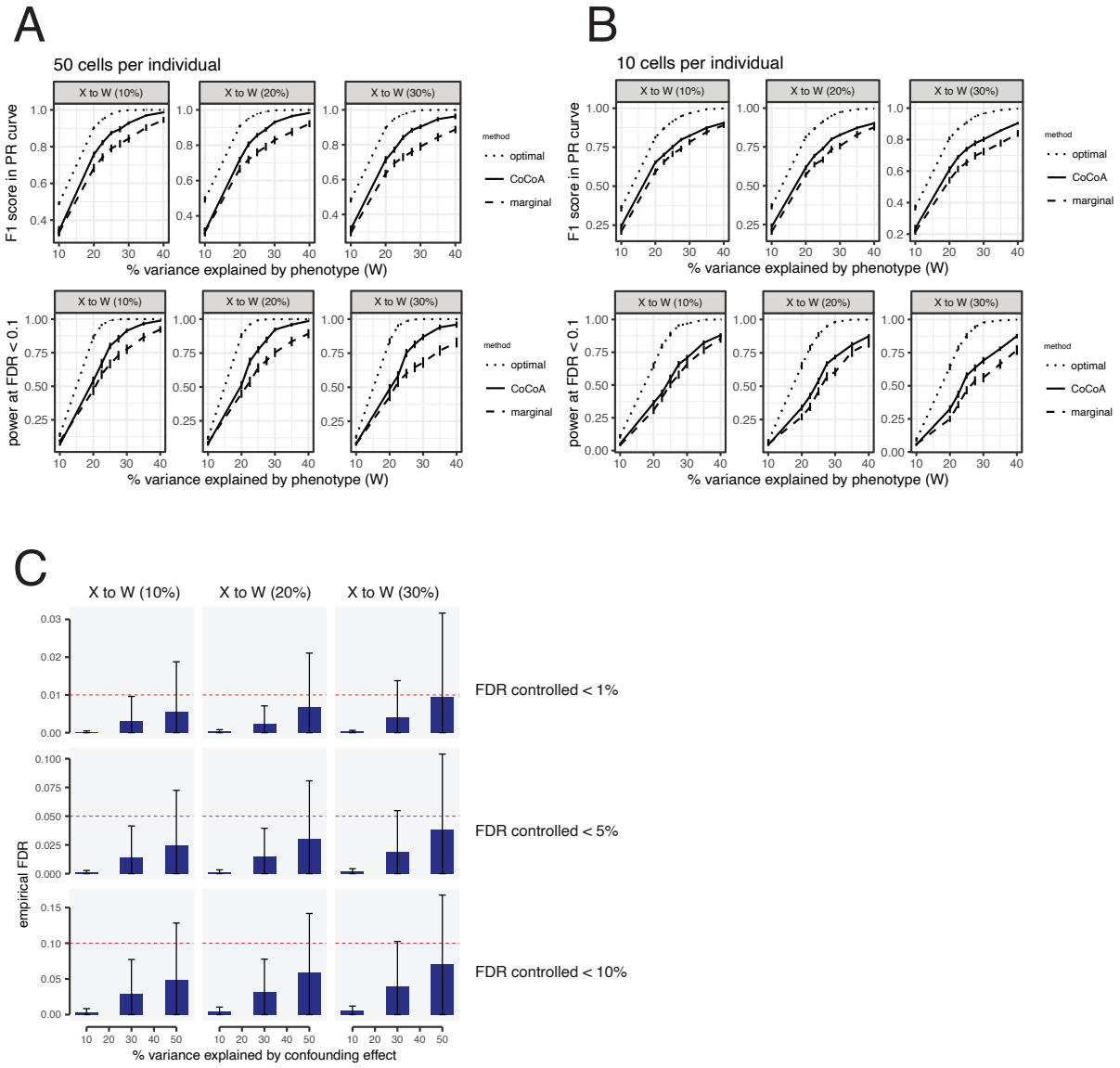
Cell-type-specific DEGs differentially enrich gene ontology (GO) keywords. *Ex*: expiatory neurons; *In*: inhibitory neurons; *Oligo*: oligodendrocytes; *OPC*: oligodendrocyte progenitor cells; *Mic*: microglia; *Astro*: astrocytes; *Per*: pericytes; *Endo*: endothelial cells. Each box is scaled proportional to the level of enrichment significance, and the colour gradient marks the number genes overlapped in each keyword and cell type. **(A)** Top GO keywords enriched by up-regulated AD genes. **(B)** Top GO keywords enriched by down-regulated AD genes.

Figure 5.



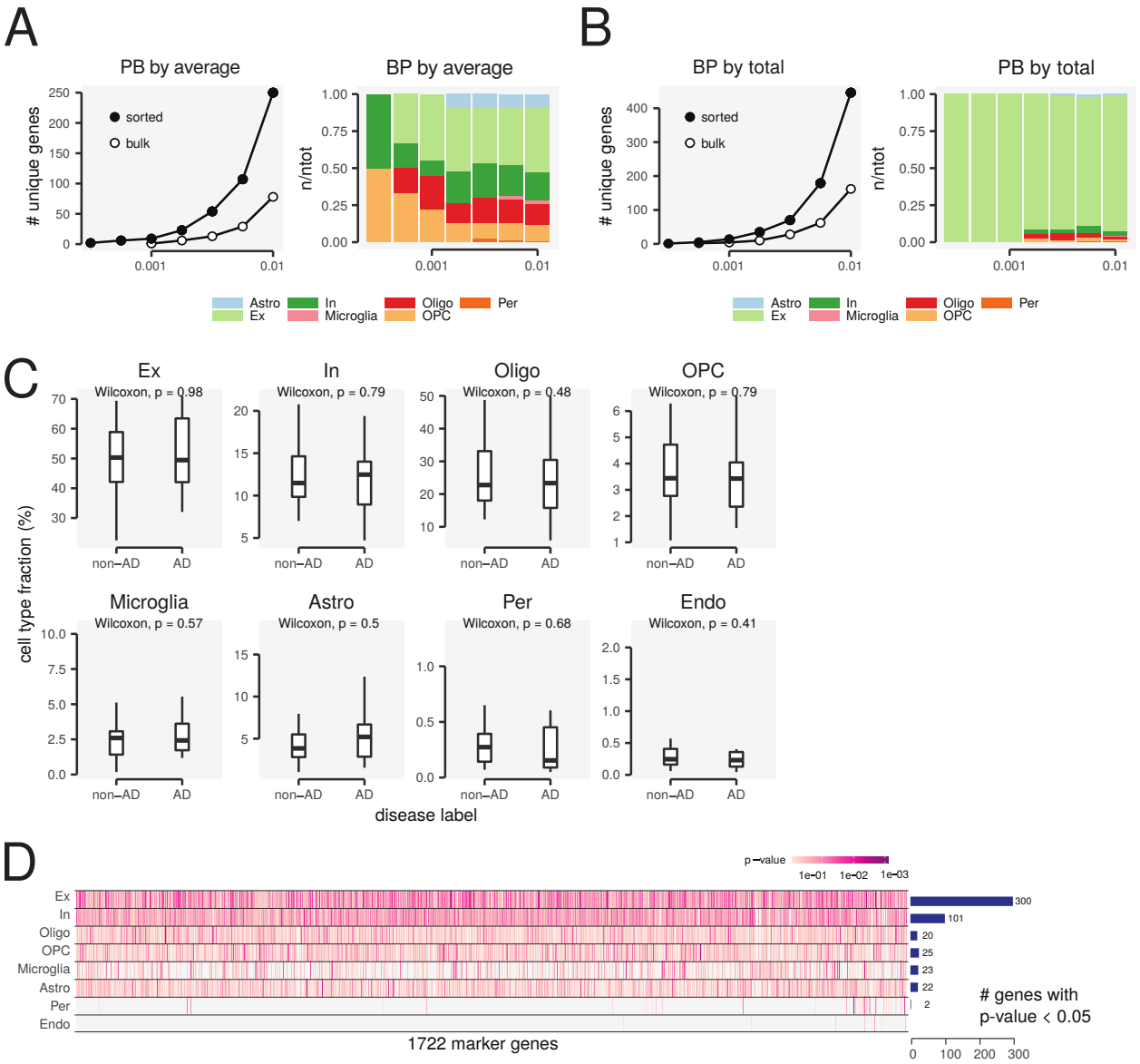
An overall picture of causal inference for dissecting disease mechanisms using genomics data

Supplementary Figure 1.



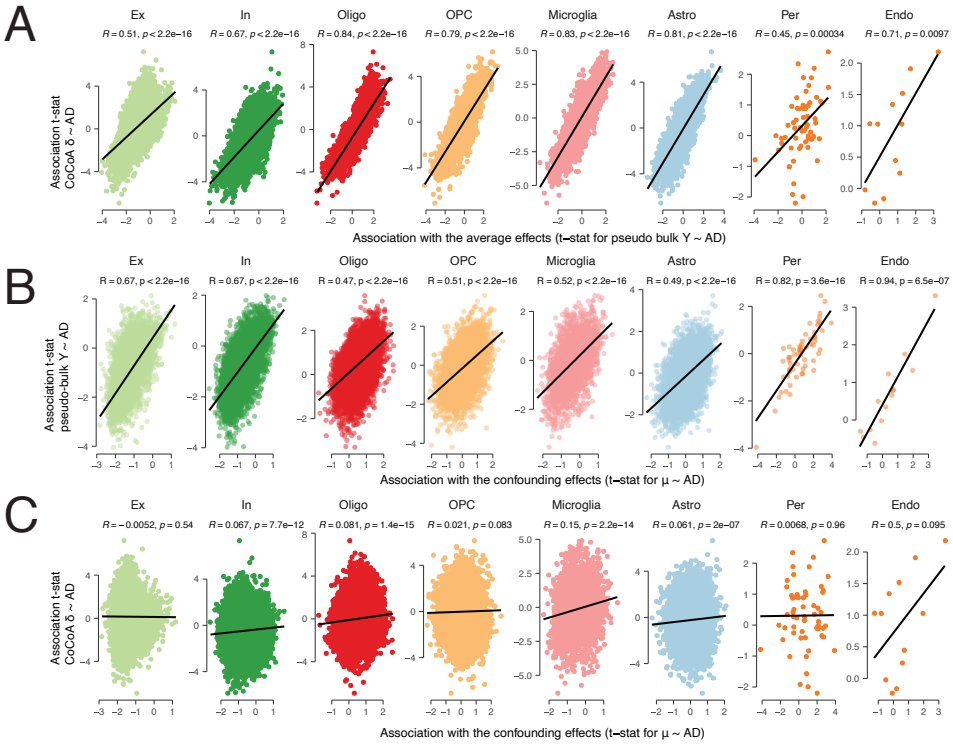
CoCoA improves statistical power. (A) Simulation results with 50 cells per individual. Error bars indicate standard errors for 30 experiments repeated with different pseudo-random seeds. **(B)** Simulation results with 10 cells per individual. **(C)** Empirical false discovery rate calibration.

Supplementary Figure 2.



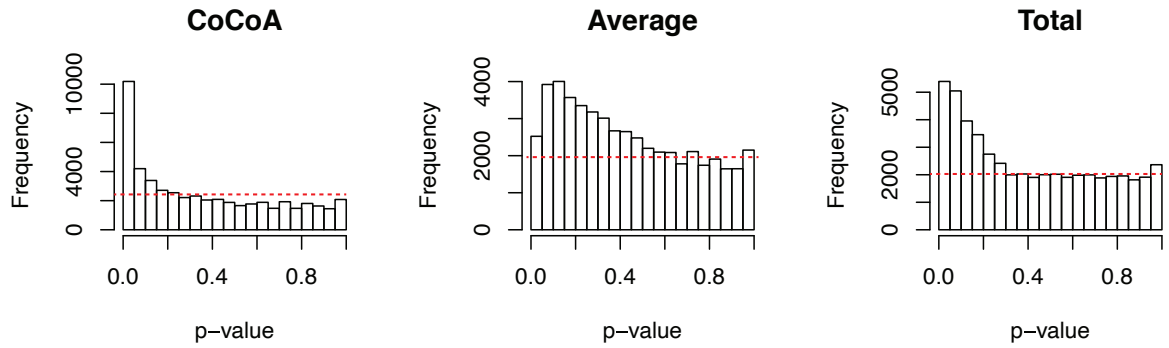
Cell-type-sorted approach improves statistical power and interpretation of differential expression analysis. **(A)** Associations of AD with pseudo-bulk data aggregated by an average value for each gene. The number of genes discovered with different p-value thresholds (left) and the genes's proportion of cell types (right). **(B)** Associations of AD with pseudo-bulk data aggregated by a total count for each gene. The number of genes discovered with different p-value thresholds (left) and the genes's proportion of cell types (right). **(C)** Marginal associations between AD and cell type fractions across 48 individuals. **(D)** Associations of AD with pseudo-bulk data aggregated by an average of each marker gene.

Supplementary Figure 3.



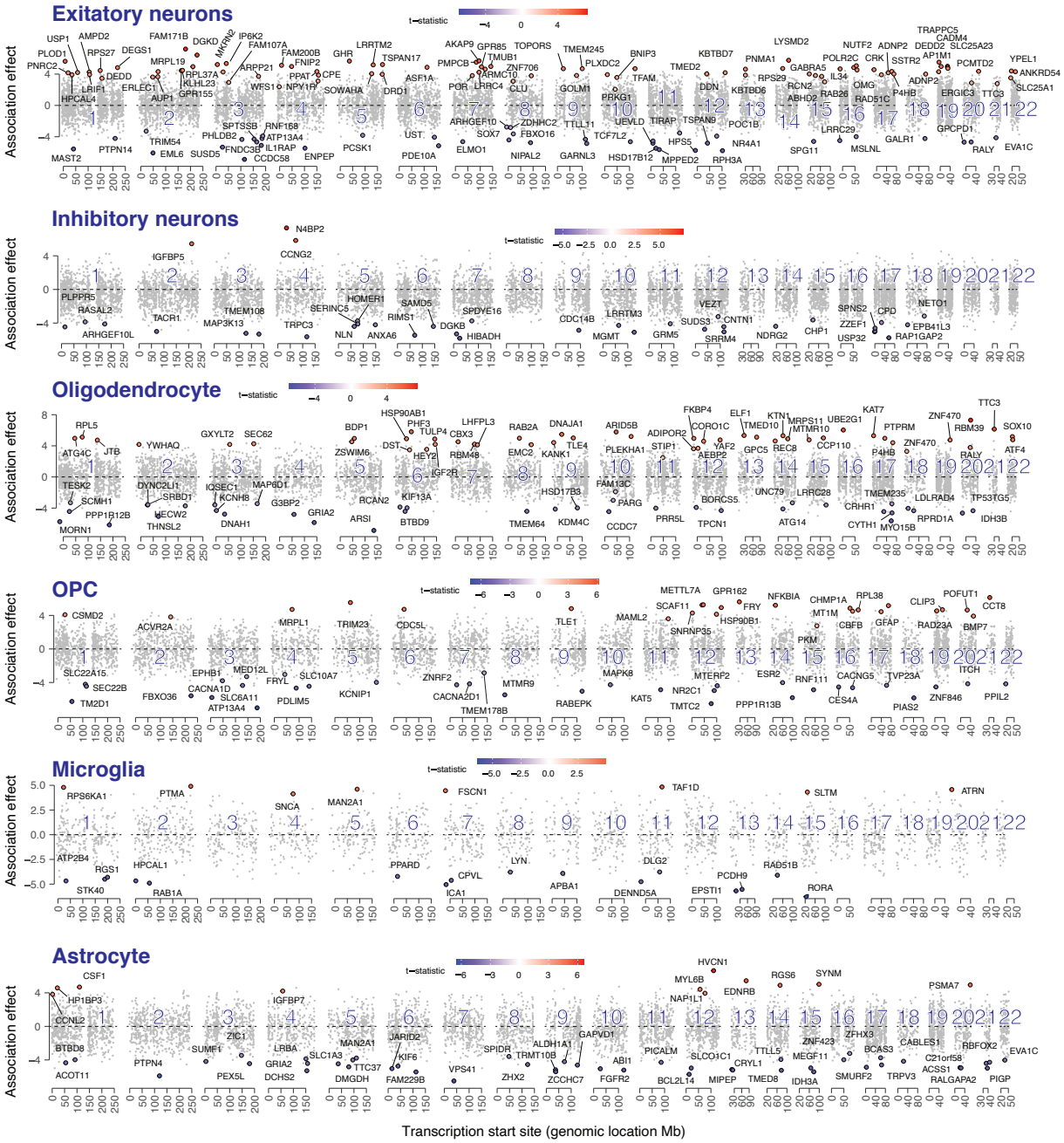
CoCoA improves the statistical power of causal gene discovery in Alzheimer's disease case-control study. (A) The effect size directions calculated on CoCoA and PB (average) data agree globally in all the cell types. **(B)** The effect sizes calculated on gene-level confounding factors and PB (average) data are globally correlated in all the cell types. **(C)** The effect size directions calculated on CoCoA and confounding factors are nearly orthogonal in most of the cell types.

Supplementary Figure 4.



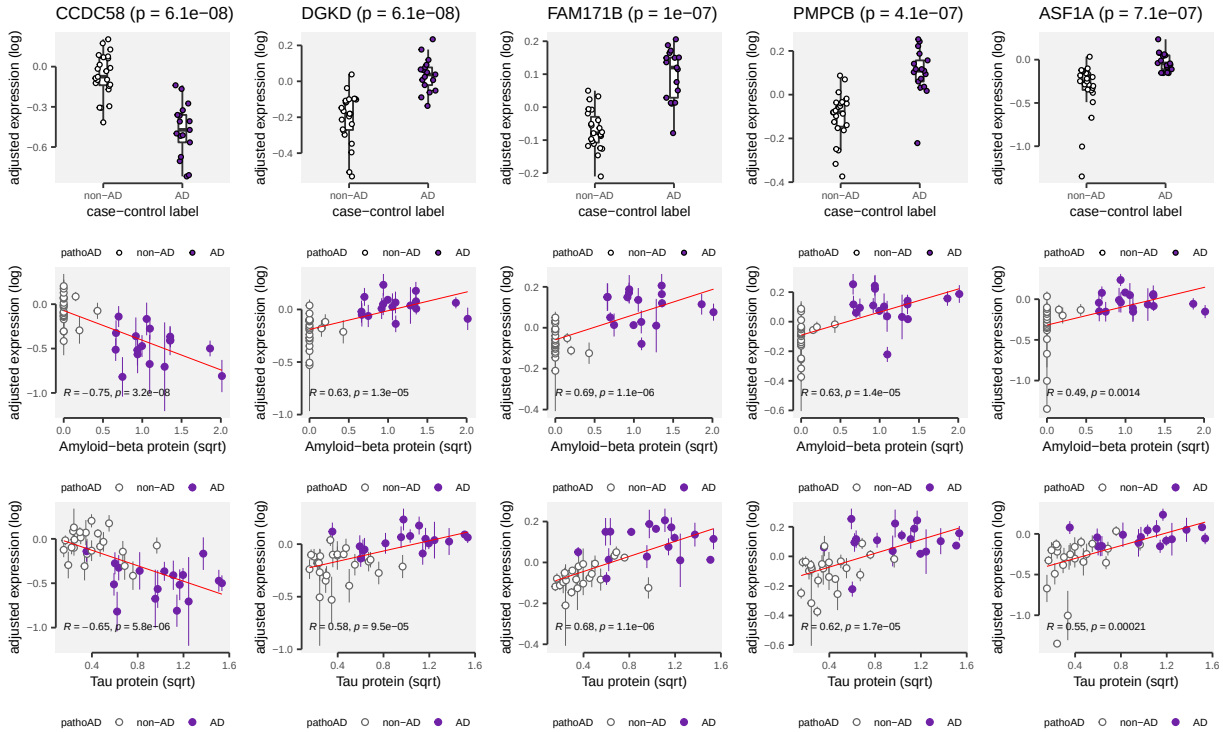
The histograms of p-values for cell-type-specific associations of AD with gene expressions estimated by three different pseudo-bulk methods. *CoCoA*: δ_g values after counterfactual confounder adjustment (Fig. 1). *Average*: each gene's average expression across cells within each individual. *Average*: each gene's total count across cells within each individual.

Supplementary Figure 5.

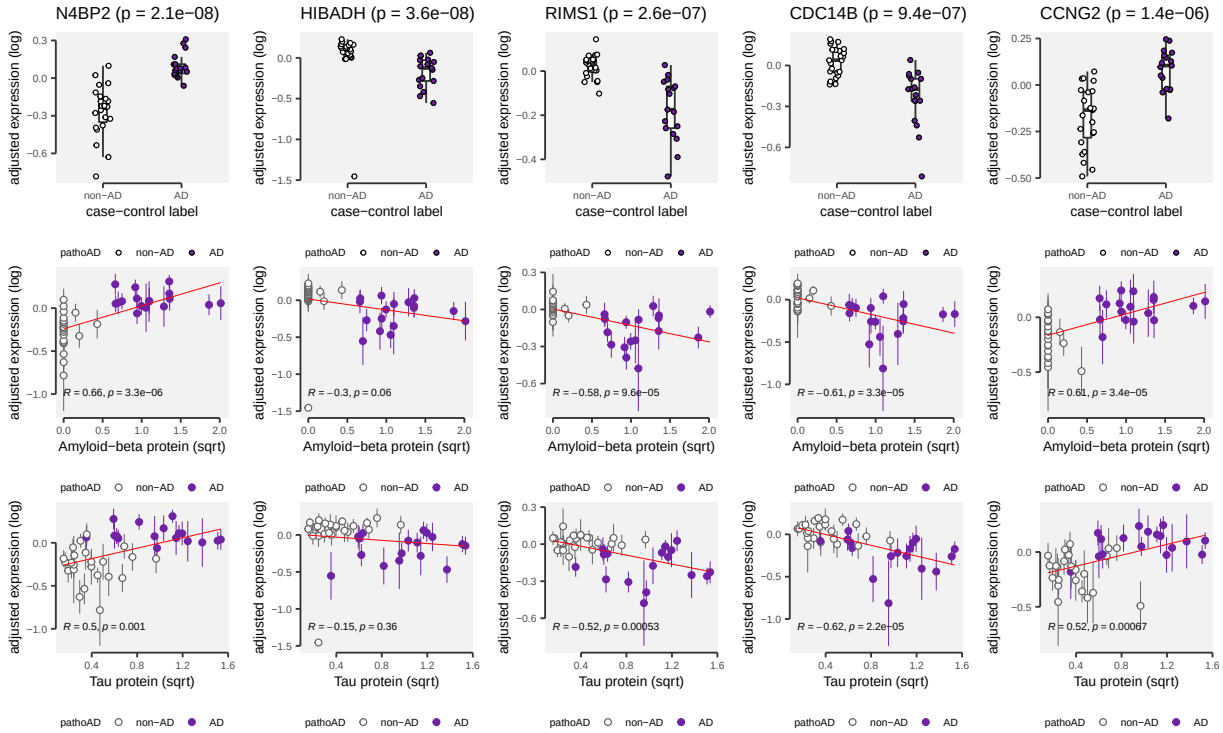


Cell-type-specific genes regulated by AD (FDR < 1%). *x-axis*: transcription start site for the genes. *y-axis*: association t-statistics. *colour*: t-statistics.

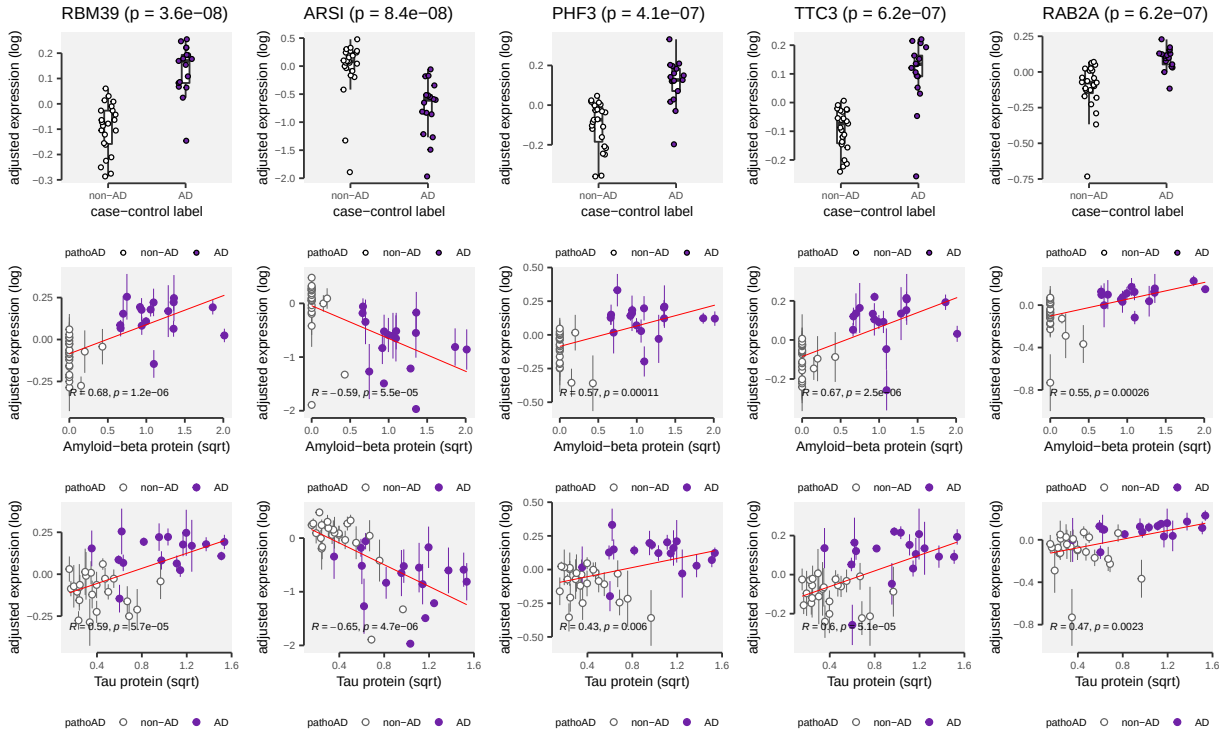
Supplementary Figure 6.



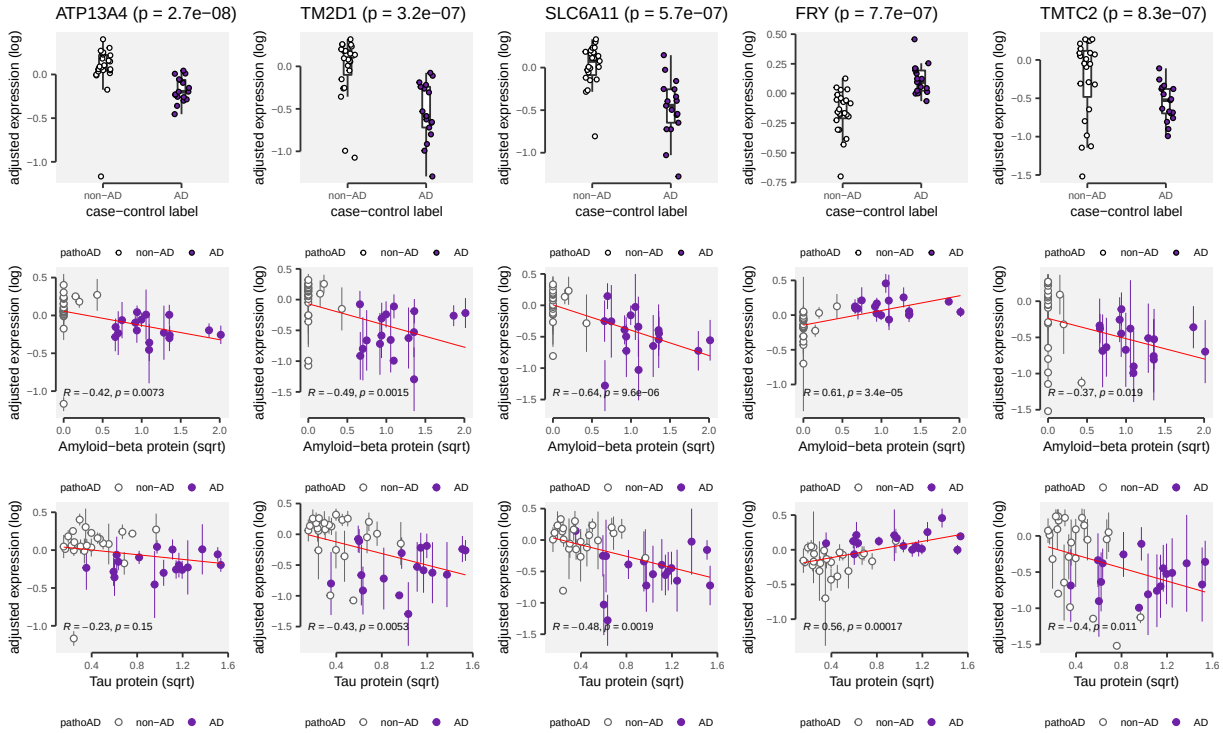
Supplementary Figure 7.



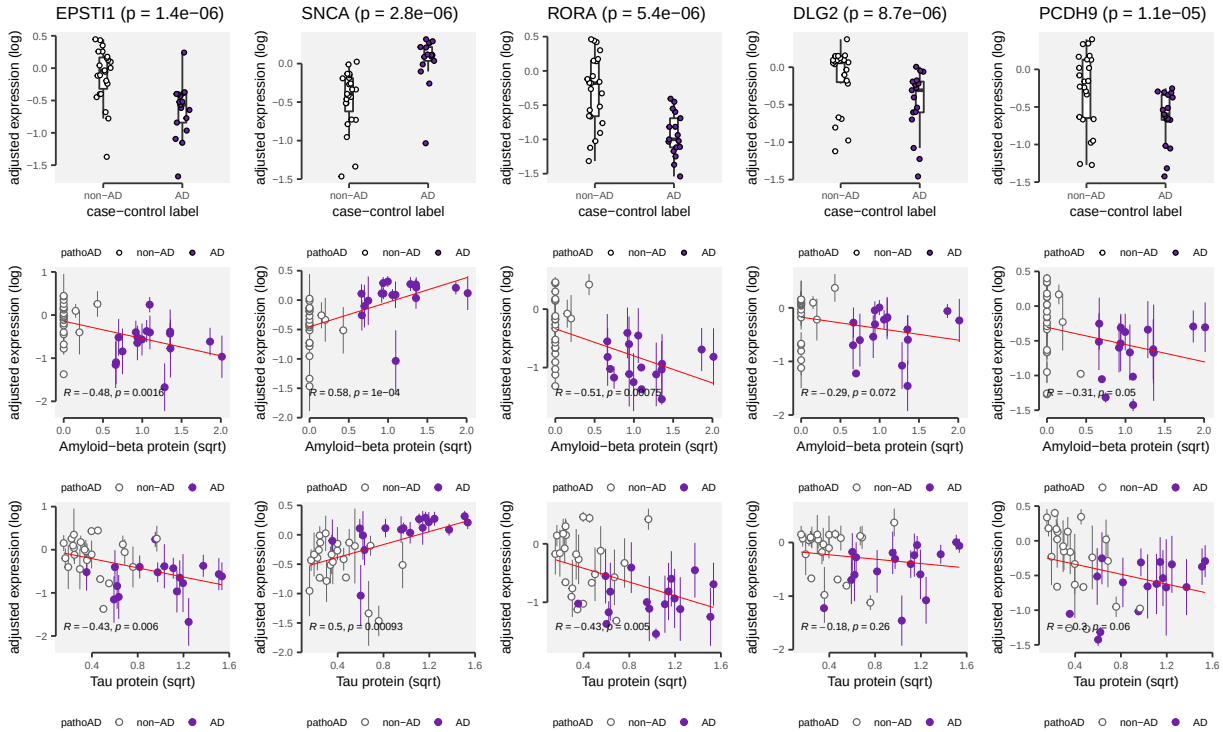
Supplementary Figure 8.



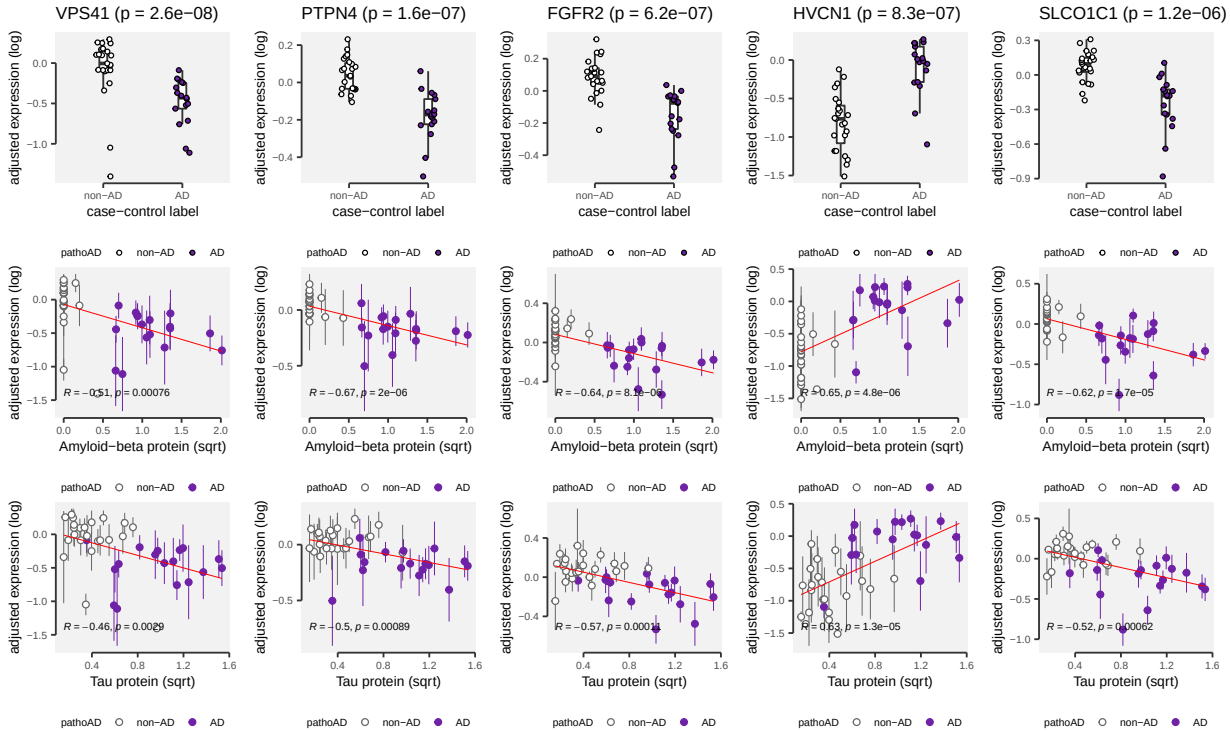
Supplementary Figure 9.



Supplementary Figure 10.



Supplementary Figure 11.



Tables

Table 1.

Top cell-type-specific genes regulated by AD (FWER < 1%). *Chr*: chromosome name; *TSS*: transcription start site (Mb). *Ex*: expiatory neurons; *In*: inhibitory neurons; *Oligo*: oligodendrocytes; *OPC*: oligodendrocyte progenitor cells; *Mic*: microglia; *Astro*: astrocytes; *Per*: pericytes; *Endo*: endothelial cells. The first numbers for the cell types denote Wilcoxon p-value, and the second numbers in the parenthesis mark directional effects (t-statistics between AD and non-AD; positive = up-regulation, negative = down-regulation in AD).

gene	Chr	TSS	Ex	In	Oligo	OPC	Microglia	Astro
TM2D1	1	61	6.3e-01 (0.61)	1.2e-01 (-1.57)	5.7e-01 (-0.27)	3.2e-07 (-6.27)		2.4e-02 (2.87)
PPP1R12B	1	202	4.8e-01 (0.04)	8.1e-01 (-0.08)	6.2e-07 (-6.16)	2.9e-02 (-2.26)	8.3e-01 (0.73)	7.2e-02 (-1.56)
PTPN4	2	119	9.1e-01 (0.4)	7.5e-02 (1.54)	5.7e-01 (-0.42)	9.1e-01 (-0.26)		1.6e-07 (-5.9)
FAM171B	2	186	1.0e-07 (7.2)	2.2e-01 (1.66)	6.2e-02 (2.34)	3.0e-01 (1.37)		6.1e-01 (0.94)
DGKD	2	233	6.1e-08 (6.43)	1.0e-01 (-2.03)	7.5e-01 (0.03)	2.5e-02 (-1.79)		2.5e-01 (-1.03)
SLC6A11	3	10	1.4e-01 (0.49)	7.6e-01 (0.11)		5.7e-07 (-5.79)		4.8e-01 (0.38)
CCDC58	3	122	6.1e-08 (-6.89)	1.6e-01 (-1.57)				
ATP13A4	3	193	5.4e-05 (-4.22)	1.9e-01 (-0.9)	5.8e-03 (-2.55)	2.7e-08 (-7.03)		8.1e-01 (-0.18)
N4BP2	4	40	7.5e-04 (3.77)	2.1e-08 (7.31)	1.1e-01 (2.1)	9.2e-01 (0.24)		2.3e-01 (1.32)
SNCA	4	89	6.9e-03 (2.29)	4.2e-01 (1.07)	1.1e-01 (0.6)	3.9e-03 (1.97)	2.8e-06 (4.14)	
ARSI	5	150	3.1e-01 (-1.01)		8.4e-08 (-6.88)			
PHF3	6	63	1.6e-02 (2.3)	3.4e-01 (0.55)	4.1e-07 (5.81)	8.4e-01 (-0.42)	6.7e-01 (0.64)	1.3e-01 (1.5)
RIMS1	6	72	9.0e-02 (-1.58)	2.6e-07 (-5.45)	2.9e-01 (-0.47)	3.9e-01 (-1.82)		2.9e-01 (0.57)
ASF1A	6	118	7.1e-07 (4.87)	8.6e-01 (-0.03)	3.1e-01 (1.08)	3.7e-01 (0.78)		
HIBADH	7	27	8.3e-01 (-0.66)	3.6e-08 (-5.81)	2.7e-02 (-2.18)	2.5e-01 (-1.45)		3.4e-02 (-2.12)
VPS41	7	38	2.4e-02 (-2.08)	1.0e-01 (-1.24)	1.1e-01 (-0.83)	8.9e-02 (1.22)	6.1e-01 (0.4)	2.6e-08 (-6.51)
PMPCB	7	103	4.1e-07 (5.69)	8.1e-01 (-0.34)	4.4e-03 (2.81)	2.8e-01 (0.54)		3.6e-02 (1.28)
RAB2A	8	60	3.6e-03 (2.74)	5.9e-01 (-0.8)	6.2e-07 (4.98)	4.3e-01 (0.44)	6.8e-04 (4.18)	5.5e-01 (-0.48)
CDC14B	9	96	7.5e-02 (-1.8)	9.4e-07 (-4.85)	2.9e-01 (-0.81)	7.0e-01 (-0.52)	6.3e-02 (1.54)	2.4e-03 (-2.37)
FGFR2	10	121	6.6e-02 (-1.78)	6.9e-03 (-3.02)	5.1e-02 (-2.26)			6.2e-07 (-5.21)
SLCO1C1	12	20	9.2e-01 (-0.1)					1.2e-06 (-4.97)
TMTC2	12	82	2.3e-01 (-1.1)	4.5e-02 (-1.98)	7.6e-03 (2.03)	8.3e-07 (-6.52)		3.3e-01 (0.91)
HVCN1	12	110	4.5e-01 (-0.52)					8.3e-07 (6.64)
FRY	13	32	5.0e-01 (0.82)	2.2e-01 (1.49)	2.0e-01 (1.09)	7.7e-07 (5.63)		6.7e-01 (-0.46)
EPSTI1	13	42	3.1e-01 (-1.03)	3.6e-02 (-2.32)			1.4e-06 (-5.69)	
PIAS2	18	46	1.4e-02 (-2.06)	1.3e-01 (-1.26)	5.6e-01 (-0.17)	1.3e-06 (-5.84)	9.3e-01 (0)	1.1e-01 (-1.28)
RBM39	20	35	4.2e-04 (3.76)	5.9e-01 (-0.87)	3.6e-08 (7.3)	8.3e-01 (0.35)	1.2e-03 (-3.78)	2.4e-02 (-1.91)

(continued)

gene	Chr	TSS	Ex	In	Oligo	OPC	Microglia	Astro
TTC3	21	37	4.9e-05 (2.78)	4.2e-01 (0.9)	6.2e-07 (6.15)	1.2e-01 (0.62)	3.6e-02 (2.21)	8.4e-04 (3.78)

Supplementary Table 1.

Gene ontology (GO) enriched by up-regulated genes (FDR < 5%). *Ex*: expiatory neurons; *In*: inhibitory neurons; *Oligo*: oligodendrocytes; *OPC*: oligodendrocyte progenitor cells; *Mic*: microglia; *Astro*: astrocytes; *Per*: pericytes; *Endo*: endothelial cells. *p-value*: hypergeometric p-value. *q-value*: adjusted p-value by Benjamni-Hochberg procedure.

cell type	GO	p-value	q-value
Ex	chaperonin-containing T-complex	1.2e-04	1.8e-02
Ex	dendrite	4.8e-04	5.5e-02
Ex	Hrd1p ubiquitin ligase ERAD-L complex	1.2e-04	1.8e-02
Ex	integral component of endoplasmic reticulum membrane	6.5e-05	1.8e-02
Ex	neuron projection	1.3e-04	1.8e-02
Ex	neuronal dense core vesicle	1.2e-04	1.8e-02
Ex	ubiquitin protein ligase binding	5.6e-05	6.8e-02
Ex	ubiquitin-dependent ERAD pathway	6.5e-06	2.9e-02
Ex	zona pellucida receptor complex	3.5e-05	1.8e-02
In	autophagy	4.6e-02	9.8e-02
In	cell division	7.9e-03	9.8e-02
In	cellular protein localization	2.5e-02	9.8e-02
In	chromosome organization	6.8e-03	9.8e-02
In	DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest	2.8e-02	9.8e-02
In	ER overload response	1.1e-03	9.8e-02
In	eye photoreceptor cell development	1.6e-03	9.8e-02
In	intrinsic apoptotic signaling pathway in response to DNA damage by p53 class mediator	9.1e-03	9.8e-02
In	memory	5.5e-04	9.8e-02
In	mitotic cell cycle	5.2e-03	9.8e-02
In	negative regulation of interleukin-6 production	1.0e-02	9.8e-02
In	negative regulation of intrinsic apoptotic signaling pathway	5.3e-03	9.8e-02
In	negative regulation of phagocytosis	1.8e-03	9.8e-02
In	negative regulation of protein phosphorylation	4.9e-02	9.8e-02
In	negative regulation of tumor necrosis factor production	1.5e-02	9.8e-02
In	neuron apoptotic process	1.5e-02	9.8e-02
In	positive regulation of release of cytochrome c from mitochondria	4.8e-03	9.8e-02
In	positive regulation of vascular associated smooth muscle cell proliferation	9.1e-03	9.8e-02
In	Ras protein signal transduction	4.4e-02	9.8e-02
In	regulation of DNA replication	4.0e-03	9.8e-02
In	response to ischemia	2.6e-02	9.8e-02

(continued)

cell type	GO	p-value	q-value
In	response to nutrient	3.9e-02	9.8e-02
In	sister chromatid cohesion	3.2e-03	9.8e-02
In	visual perception	3.9e-02	9.8e-02
Oligo	chromatin organization	9.2e-07	4.1e-03
Oligo	mRNA 5'-UTR binding	3.7e-05	4.4e-02
Oligo	negative regulation of transcription, DNA-templated	1.1e-05	2.5e-02
Oligo	transcription coactivator activity	1.4e-04	8.6e-02
OPC	microtubule plus-end	4.6e-05	3.7e-02
OPC	unfolded protein binding	3.5e-05	4.3e-02
OPC	vesicle	1.5e-04	5.9e-02
Microglia	actin binding	3.3e-02	7.8e-02
Microglia	actin cytoskeleton	1.5e-03	8.7e-02
Microglia	actin filament binding	2.9e-02	7.8e-02
Microglia	actin filament organization	1.0e-02	7.5e-02
Microglia	adult walking behavior	5.8e-03	7.5e-02
Microglia	axon	5.1e-02	9.2e-02
Microglia	brain development	6.0e-02	8.7e-02
Microglia	brush border	1.7e-02	8.7e-02
Microglia	cadherin binding	2.1e-02	7.8e-02
Microglia	cell migration	5.0e-02	8.1e-02
Microglia	cell motility	6.2e-03	7.5e-02
Microglia	cell-cell junction	3.8e-03	8.7e-02
Microglia	cellular response to ethanol	9.3e-04	7.5e-02
Microglia	cellular response to interferon-gamma	2.1e-03	7.5e-02
Microglia	cellular response to oxidative stress	4.3e-02	7.8e-02
Microglia	chemical synaptic transmission	4.2e-02	7.8e-02
Microglia	cysteine-type endopeptidase inhibitor activity involved in apoptotic process	1.9e-03	7.8e-02
Microglia	cytosolic large ribosomal subunit	9.7e-04	8.7e-02
Microglia	cytosolic small ribosomal subunit	4.7e-04	8.7e-02
Microglia	endocytosis	4.5e-02	7.8e-02
Microglia	endothelial cell migration	4.7e-03	7.5e-02
Microglia	ephrin receptor signaling pathway	3.3e-02	7.5e-02
Microglia	establishment or maintenance of cell polarity	1.3e-02	7.5e-02
Microglia	excitatory postsynaptic potential	1.6e-02	7.5e-02
Microglia	fat cell differentiation	2.6e-02	7.5e-02
Microglia	Fc-gamma receptor signaling pathway involved in phagocytosis	2.5e-02	7.5e-02
Microglia	focal adhesion	8.8e-05	3.5e-02
Microglia	growth cone	1.5e-02	8.7e-02
Microglia	hydrolase activity, acting on glycosyl bonds	2.7e-02	7.8e-02

(continued)

cell type	GO	p-value	q-value
Microglia	immunological synapse	7.9e-03	8.7e-02
Microglia	lipid binding	1.6e-02	7.8e-02
Microglia	long-term synaptic potentiation	1.3e-02	7.5e-02
Microglia	membrane organization	1.1e-03	7.5e-02
Microglia	microvillus	1.7e-02	8.7e-02
Microglia	motor activity	5.8e-02	9.0e-02
Microglia	negative regulation of cell migration	4.6e-02	8.0e-02
Microglia	negative regulation of cysteine-type endopeptidase activity involved in apoptotic process	5.6e-05	4.2e-02
Microglia	negative regulation of gene expression, epigenetic	3.2e-03	7.5e-02
Microglia	negative regulation of neuron apoptotic process	1.1e-02	7.5e-02
Microglia	negative regulation of protein kinase B signaling	9.3e-03	7.5e-02
Microglia	negative regulation of protein phosphorylation	2.5e-03	7.5e-02
Microglia	negative regulation of release of cytochrome c from mitochondria	1.9e-03	7.5e-02
Microglia	neuromuscular synaptic transmission	1.7e-03	7.5e-02
Microglia	neuron-neuron synaptic transmission	5.1e-04	7.5e-02
Microglia	neuronal cell body	5.4e-02	9.4e-02
Microglia	nuclear-transcribed mRNA catabolic process, nonsense-mediated decay	3.0e-06	3.3e-03
Microglia	PDZ domain binding	3.7e-02	7.8e-02
Microglia	platelet-derived growth factor receptor signaling pathway	6.2e-03	7.5e-02
Microglia	positive regulation of gene expression	1.1e-02	7.5e-02
Microglia	positive regulation of lamellipodium assembly	3.2e-03	7.5e-02
Microglia	positive regulation of peptidyl-serine phosphorylation	3.0e-02	7.5e-02
Microglia	positive regulation of protein kinase B signaling	1.7e-02	7.5e-02
Microglia	presynapse	1.8e-02	8.7e-02
Microglia	receptor internalization	8.8e-03	7.5e-02
Microglia	regulation of neuron projection development	8.8e-03	7.5e-02
Microglia	regulation of small GTPase mediated signal transduction	7.4e-02	9.9e-02
Microglia	response to cocaine	9.7e-03	7.5e-02
Microglia	response to ethanol	5.7e-02	8.6e-02
Microglia	response to oxidative stress	6.3e-02	9.0e-02
Microglia	ribonucleoprotein complex	1.6e-02	8.7e-02
Microglia	ribosomal small subunit biogenesis	2.6e-03	7.5e-02
Microglia	ribosome	4.3e-06	3.5e-03
Microglia	small ribosomal subunit	5.5e-03	8.7e-02
Microglia	SRP-dependent cotranslational protein targeting to membrane	5.7e-07	1.3e-03
Microglia	structural constituent of cytoskeleton	3.4e-02	7.8e-02
Microglia	structural constituent of ribosome	1.7e-05	2.1e-02
Microglia	synapse assembly	2.1e-02	7.5e-02

(continued)

cell type	GO	p-value	q-value
Microglia	synaptic vesicle endocytosis	7.5e-03	7.5e-02
Microglia	synaptic vesicle transport	1.9e-03	7.5e-02
Microglia	terminal bouton	1.4e-02	8.7e-02
Microglia	translation	5.2e-05	4.2e-02
Microglia	translational initiation	5.6e-07	1.3e-03
Microglia	vesicle	2.5e-02	8.7e-02
Microglia	viral transcription	2.3e-06	3.3e-03
Astro	cytosolic small ribosomal subunit	1.5e-05	1.2e-02
Astro	small ribosomal subunit	7.4e-05	3.0e-02
Astro	structural constituent of muscle	3.9e-07	4.8e-04

Supplementary Table 2.

Gene ontology (GO) enriched by down-regulated genes (FDR < 5%). *Ex*: expiatory neurons; *In*: inhibitory neurons; *Oligo*: oligodendrocytes; *OPC*: oligodendrocyte progenitor cells; *Mic*: microglia; *Astro*: astrocytes; *Per*: pericytes; *Endo*: endothelial cells. *p-value*: hypergeometric p-value. *q-value*: adjusted p-value by Benjamni-Hochberg procedure.

cell type	GO	p-value	q-value
In	integral component of postsynaptic density membrane	4.1e-05	3.3e-02
Microglia	angiogenesis	2.7e-03	9.3e-02
Microglia	apoptotic signaling pathway	3.0e-02	9.3e-02
Microglia	axon ensheathment	3.3e-04	9.3e-02
Microglia	axon guidance	1.4e-02	9.3e-02
Microglia	cadherin binding	4.3e-02	9.6e-02
Microglia	cell adhesion	2.9e-02	9.3e-02
Microglia	cell death	1.1e-02	9.3e-02
Microglia	cellular calcium ion homeostasis	6.1e-02	9.9e-02
Microglia	cellular iron ion homeostasis	1.9e-02	9.3e-02
Microglia	cellular response to amino acid stimulus	1.9e-02	9.3e-02
Microglia	cellular response to glucose starvation	1.7e-02	9.3e-02
Microglia	cellular response to hypoxia	1.1e-02	9.3e-02
Microglia	cellular response to interleukin-1	2.2e-02	9.3e-02
Microglia	cellular response to nerve growth factor stimulus	1.3e-02	9.3e-02
Microglia	cerebral cortex development	3.8e-02	9.3e-02
Microglia	cytokine-mediated signaling pathway	1.1e-02	9.3e-02
Microglia	DNA-dependent ATPase activity	1.1e-02	9.6e-02
Microglia	endocytosis	3.0e-03	9.3e-02
Microglia	endoplasmic reticulum to Golgi vesicle-mediated transport	3.7e-02	9.3e-02
Microglia	enzyme binding	1.7e-02	9.6e-02
Microglia	gamma-tubulin binding	5.8e-03	9.6e-02
Microglia	glucose transmembrane transport	3.3e-03	9.3e-02
Microglia	GTPase activator activity	5.0e-03	9.6e-02
Microglia	guanyl-nucleotide exchange factor activity	5.2e-02	9.8e-02
Microglia	immune response	1.0e-02	9.3e-02
Microglia	intracellular protein transport	4.3e-02	9.3e-02
Microglia	intracellular receptor signaling pathway	1.3e-02	9.3e-02
Microglia	mRNA 3'-UTR binding	4.3e-02	9.6e-02
Microglia	multicellular organism growth	4.9e-02	9.5e-02
Microglia	muscle cell differentiation	2.0e-03	9.3e-02
Microglia	NAD+ ADP-ribosyltransferase activity	6.3e-03	9.6e-02

(continued)

cell type	GO	p-value	q-value
Microglia	negative regulation of apoptotic process	5.0e-02	9.5e-02
Microglia	negative regulation of bone mineralization	1.2e-03	9.3e-02
Microglia	negative regulation of cytoplasmic translation	3.3e-04	9.3e-02
Microglia	negative regulation of fat cell differentiation	1.3e-02	9.3e-02
Microglia	negative regulation of inflammatory response	4.9e-03	9.3e-02
Microglia	negative regulation of neuron apoptotic process	2.0e-02	9.3e-02
Microglia	negative regulation of neuron projection development	3.6e-02	9.3e-02
Microglia	neural crest cell migration	1.5e-02	9.3e-02
Microglia	neurotransmitter secretion	2.0e-02	9.3e-02
Microglia	neutrophil degranulation	8.0e-03	9.3e-02
Microglia	nuclear receptor activity	2.1e-02	9.6e-02
Microglia	O-glycan processing	1.9e-02	9.3e-02
Microglia	phospholipid biosynthetic process	2.5e-02	9.3e-02
Microglia	phosphorylation-dependent protein binding	2.2e-04	9.6e-02
Microglia	platelet activation	5.6e-02	9.6e-02
Microglia	positive regulation of blood vessel endothelial cell migration	8.9e-03	9.3e-02
Microglia	positive regulation of endothelial cell migration	2.9e-02	9.3e-02
Microglia	positive regulation of endothelial cell proliferation	2.9e-03	9.3e-02
Microglia	positive regulation of gene expression	2.7e-02	9.3e-02
Microglia	positive regulation of GTPase activity	2.2e-03	9.3e-02
Microglia	positive regulation of I-kappaB kinase/NF-kappaB signaling	2.8e-02	9.3e-02
Microglia	positive regulation of insulin secretion involved in cellular response to glucose stimulus	8.9e-03	9.3e-02
Microglia	positive regulation of membrane protein ectodomain proteolysis	1.4e-03	9.3e-02
Microglia	positive regulation of neuron projection development	1.6e-02	9.3e-02
Microglia	positive regulation of phosphatidylinositol 3-kinase signaling	5.5e-02	9.6e-02
Microglia	positive regulation of tyrosine phosphorylation of STAT protein	1.1e-02	9.3e-02
Microglia	positive regulation of vascular endothelial growth factor production	4.9e-03	9.3e-02
Microglia	protein ADP-ribosylation	4.9e-03	9.3e-02
Microglia	protein localization to endoplasmic reticulum	3.3e-04	9.3e-02
Microglia	protein-containing complex assembly	1.9e-02	9.3e-02
Microglia	receptor clustering	4.9e-03	9.3e-02
Microglia	regulation of cell population proliferation	5.0e-02	9.5e-02
Microglia	regulation of insulin secretion	2.8e-02	9.3e-02
Microglia	regulation of protein phosphorylation	2.7e-02	9.3e-02
Microglia	regulation of protein stability	4.6e-02	9.3e-02
Microglia	regulation of release of sequestered calcium ion into cytosol	1.7e-03	9.3e-02
Microglia	regulation of small GTPase mediated signal transduction	2.3e-03	9.3e-02
Microglia	regulation of translation	3.1e-02	9.3e-02

(continued)

cell type	GO	p-value	q-value
Microglia	retrograde vesicle-mediated transport, Golgi to endoplasmic reticulum	5.3e-02	9.6e-02
Microglia	ribosome binding	1.9e-02	9.6e-02
Microglia	scaffold protein binding	2.5e-02	9.6e-02
Microglia	SH3 domain binding	2.3e-03	9.6e-02
Microglia	steroid hormone receptor activity	2.0e-02	9.6e-02
Microglia	transcription coactivator binding	3.6e-03	9.6e-02
Microglia	translation factor activity, RNA binding	3.6e-03	9.6e-02
Microglia	translation regulator activity	4.0e-03	9.6e-02
Microglia	translation repressor activity, mRNA regulatory element binding	2.3e-03	9.6e-02
Microglia	transport across blood-brain barrier	5.1e-02	9.5e-02
Microglia	ubiquitin protein ligase binding	3.8e-02	9.6e-02
Astro	NAD binding	1.7e-05	2.1e-02

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* (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. 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).

18. 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).

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

20. Cacciottolo, M. *et al.* The APOE4 allele shows opposite sex bias in microbleeds and alzheimer's disease of humans and mice. *Neurobiol. Aging* **37,** 47–57 (2016).

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

22. 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).

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

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

25. Kunkle, B. W. *et al.* Genetic meta-analysis of diagnosed alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nat. Genet.* **51,** 414–430 (2019).

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

27. 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).

28. Goto, K. & Kondo, H. Molecular cloning and expression of a 90-kDa diacylglycerol kinase that predominantly localizes in neurons. *Proc. Natl. Acad. Sci. U. S. A.* **90,** 7598–7602 (1993).

29. Shirai, Y. & Saito, N. Diacylglycerol kinase as a possible therapeutic target for neuronal diseases. *J. Biomed. Sci.* **21,** 28 (2014).

30. Wood, P. L. *et al.* Targeted lipidomics of fontal cortex and plasma diacylglycerols (DAG) in mild cognitive impairment and alzheimer's disease: Validation of DAG accumulation early in the pathophysiology of alzheimer's disease. *J. Alzheimers. Dis.* **48,** 537–546 (2015).

31. Chan, R. B. *et al.* Comparative lipidomic analysis of mouse and human brain with alzheimer disease. *J. Biol. Chem.* **287,** 2678–2688 (2012).

32. Wood, P. L., Tippireddy, S., Feriante, J. & Woltjer, R. L. Augmented frontal cortex diacylglycerol levels in parkinson's disease and lewy body disease. *PLoS One* **13,** e0191815 (2018).

33. Larson, M. E. *et al.* Soluble α-synuclein is a novel modulator of alzheimer's disease pathophysiology. *J. Neurosci.* **32,** 10253–10266 (2012).

34. Linnertz, C. *et al.* The genetic contributions of SNCA and LRRK2 genes to lewy body pathology in alzheimer's disease. *Hum. Mol. Genet.* **23,** 4814–4821 (2014).

35. Wang, Q., Tian, Q., Song, X., Liu, Y. & Li, W. SNCA gene polymorphism may contribute to an increased risk of alzheimer's disease. *J. Clin. Lab. Anal.* **30,** 1092–1099 (2016).

36. Yoshino, Y. *et al.* Elevated mRNA expression and low methylation of SNCA in japanese alzheimer's disease subjects. *J. Alzheimers. Dis.* **54,** 1349–1357 (2016).

37. Kim, Y. M. *et al.* Proteasome inhibition induces α-synuclein SUMOylation and aggregate formation. *J. Neurol. Sci.* **307,** 157–161 (2011).

38. Rott, R. *et al.* SUMOylation and ubiquitination reciprocally regulate α-synuclein degradation

and pathological aggregation. *Proc. Natl. Acad. Sci. U. S. A.* **114,** 13176–13181 (2017).

39. Kouchaki, E., Nikoueinejad, H., Akbari, H., Azimi, S. & Behnam, M. The investigation of relevancy between PIAS1 and PIAS2 gene expression and disease severity of multiple sclerosis. *J. Immunoassay Immunochem.* **40,** 396–406 (2019).

40. Tao, C. C., Hsu, W. L., Ma, Y. L., Cheng, S. J. & Lee, E. H. Y. Epigenetic regulation of HDAC1 SUMOylation as an endogenous neuroprotection against Aβ toxicity in a mouse model of alzheimer's disease. *Cell Death Differ.* **24,** 597–614 (2017).

41. Kang, K. & Song, M.-R. Diverse FGF receptor signaling controls astrocyte specification and proliferation. *Biochem. Biophys. Res. Commun.* **395,** 324–329 (2010).

42. Savchenko, E. *et al.* FGF family members differentially regulate maturation and proliferation of stem cell-derived astrocytes. *Sci. Rep.* **9,** 9610 (2019).

43. Kang, W. *et al.* Astrocyte activation is suppressed in both normal and injured brain by FGF signaling. *Proc. Natl. Acad. Sci. U. S. A.* **111,** E2987–95 (2014).

44. Noda, M. *et al.* FGF-2 released from degenerating neurons exerts microglial-induced neuroprotection via FGFR3-ERK signaling pathway. *J. Neuroinflammation* **11,** 76 (2014).

45. Szczurkowska, J. *et al.* NEGR1 and FGFR2 cooperatively regulate cortical development and core behaviours related to autism disorders in mice. *Brain* **141,** 2772–2794 (2018).

46. Smedley, D. *et al.* BioMart–biological queries made easy. *BMC Genomics* **10,** 22 (2009).

47. Köhler, S., Bauer, S., Horn, D. & Robinson, P. N. Walking the interactome for prioritization of candidate disease genes. *Am. J. Hum. Genet.* **82,** 949–958 (2008).

48. Kim, S. S. *et al.* Genes with high network connectivity are enriched for disease heritability. *Am. J. Hum. Genet.* **104,** 896–913 (2019).

49. Quan, Y., Zhang, Q.-Y., Lv, B.-M., Xu, R.-F. & Zhang, H.-Y. Genome-wide pathogenesis interpretation using a heat diffusion-based systems genetics method and implications for gene function annotation. *Mol Genet Genomic Med* **8,** e1456 (2020).

50. Mohammadi, S., Davila-Velderrain, J. & Kellis, M. Reconstruction of Cell-type-Specific inter-

actomes at Single-Cell resolution. *Cell Syst* **9,** 559–568.e4 (2019).

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

52. Angrist, J. D., Imbens, G. W. & Rubin, D. B. Identification of causal effects using instrumental variables. *J. Am. Stat. Assoc.* **91,** 444–455 (1996).

53. 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).

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

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

56. 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).

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

58. 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).

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

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

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

62. 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>

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

64. 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).

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

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

67. 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).