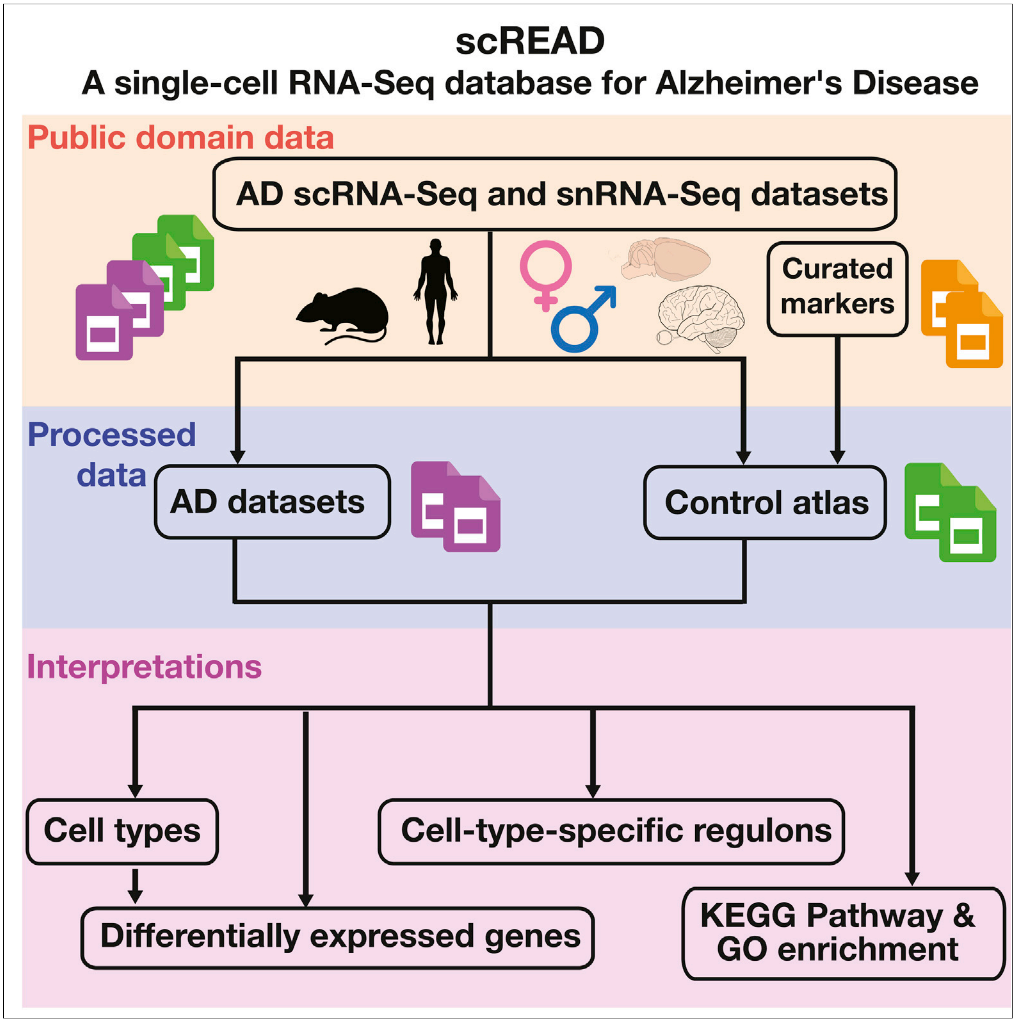
iScience

CellPress
OPEN ACCESS

Article

scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease


scREAD
A single-cell RNA-Seq database for Alzheimer's Disease
Public domain data
AD scRNA-Seq and snRNA-Seq datasets
Curated markers
Processed data
AD datasets
Control atlas
Interpretations
Cell types
Cell-type-specific regulons
Differentially expressed genes
KEGG Pathway & GO enrichment

Jing Jiang, Cankun Wang, Ren Qi, Hongjun Fu, Qin Ma

hongjun.fu@osumc.edu (H.F.)
qin.ma@osumc.edu (Q.M.)

HIGHLIGHTS

First-of-its-kind database dedicated to Alzheimer's disease sc/snRNA-Seq data sets

Control atlas and disease data sets construction for major cell types in the brain

User-friendly web server to provide comprehensive analysis interpretations

Jiang et al., iScience *23*, 101769
November 20, 2020 © 2020 The Authors.
https://doi.org/10.1016/j.isci.2020.101769

iScience

CellPress
OPEN ACCESS

Article

scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease

Jing Jiang,[1,3] Cankun Wang,[1,3] Ren Qi,[1] Hongjun Fu,[2,*] and Qin Ma[1,*]

SUMMARY

Alzheimer's disease (AD) is a progressive neurodegenerative disorder of the brain and the most common form of dementia among the elderly. The single-cell RNA-sequencing (scRNA-Seq) and single-nucleus RNA-sequencing (snRNA-Seq) techniques are extremely useful for dissecting the function/dysfunction of highly heterogeneous cells in the brain at the single-cell level, and the corresponding data analyses can significantly improve our understanding of why particular cells are vulnerable in AD. We developed an integrated database named scREAD (single-cell RNA-Seq database for Alzheimer's disease), which is as far as we know the first database dedicated to the management of all the existing scRNA-Seq and snRNA-Seq data sets from the human postmortem brain tissue with AD and mouse models with AD pathology. scREAD provides comprehensive analysis results for 73 data sets from 10 brain regions, including control atlas construction, cell-type prediction, identification of differentially expressed genes, and identification of cell-type-specific regulons.

INTRODUCTION

Alzheimer's disease (AD) is the most common cause of dementia. Currently, there are an estimated 5.8 million Americans aged 65 yeas or older suffering from AD (Claxton et al., 2015). AD is a slowly progressive brain disease that only after years of brain changes do individuals experience noticeable symptoms, such as difficulty in remembering recent conversations, names or events, and language problems (Shinagawa, 2016). Symptoms occur because neurons in parts of the brain involved in thinking, learning, and memory have been damaged or destroyed, probably by the accumulation of amyloid beta (Aβ) protein and tau protein aggregates and the neuroinflammation (Dolgin, 2018; Mucke, 2009). Unfortunately, there is no effective therapeutics that can cure or alter the disease process (Gao et al., 2016). Furthermore, molecular mechanisms underlying AD, especially the cellular vulnerability, are poorly understood.

Single-cell RNA sequencing (scRNA-Seq) examines the dynamic transcriptomic profile of individual cells with next-generation sequencing technologies and hence provides a higher resolution of cellular differences and a better understanding of the function of an individual cell in the context of its microenvironment (Seweryn et al., 2020; Wang et al., 2020; Wu et al., 2014). For frozen brain samples, using the single-nucleus RNA sequencing (snRNA-Seq) is also an important strategy. It addresses these samples that cannot be readily dissociated into a single-cell suspension and minimizes the alteration of gene expression caused by the procedure of dissociation. Previous studies have demonstrated that AD pathology differs in age, gender, brain regions, and cell types (Ewers et al., 2011; Mucke, 2009; Sala Frigerio et al., 2019). In order to study the cellular heterogeneity of the brain and reveal the complex cellular changes in the AD brain by profiling tens of thousands of individual cells, scRNA-Seq provides an alternative method (Mathys et al., 2017). The scRNA-Seq can reveal complex and rare cell populations, uncover regulatory relationships between genes, and track the trajectories of distinct cell lineages in development (Grubman et al., 2019; Qi et al., 2020). The single-cell view of AD pathology paints a unique cellular-level view of transcriptional alterations associated with AD pathology and significantly improves our understanding of the pathogenesis of AD (Del-Aguila et al., 2019; Mathys et al., 2019).

Here, we developed a database called scREAD (single-cell RNA-Seq database for Alzheimer's disease), which provides comprehensive analysis results of all the existing scRNA-Seq and snRNA-Seq data sets collected from Gene Expression Omnibus (GEO) (Barrett et al., 2013) and Synapse databases. The scREAD

[1]Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

[2]Department of Neuroscience, The Ohio State University, Columbus, OH 43210, USA

[3]These authors contributed equally

*Correspondence: hongjun.fu@osumc.edu (H.F.), qin.ma@osumc.edu (Q.M.)

https://doi.org/10.1016/j.isci.2020.101769

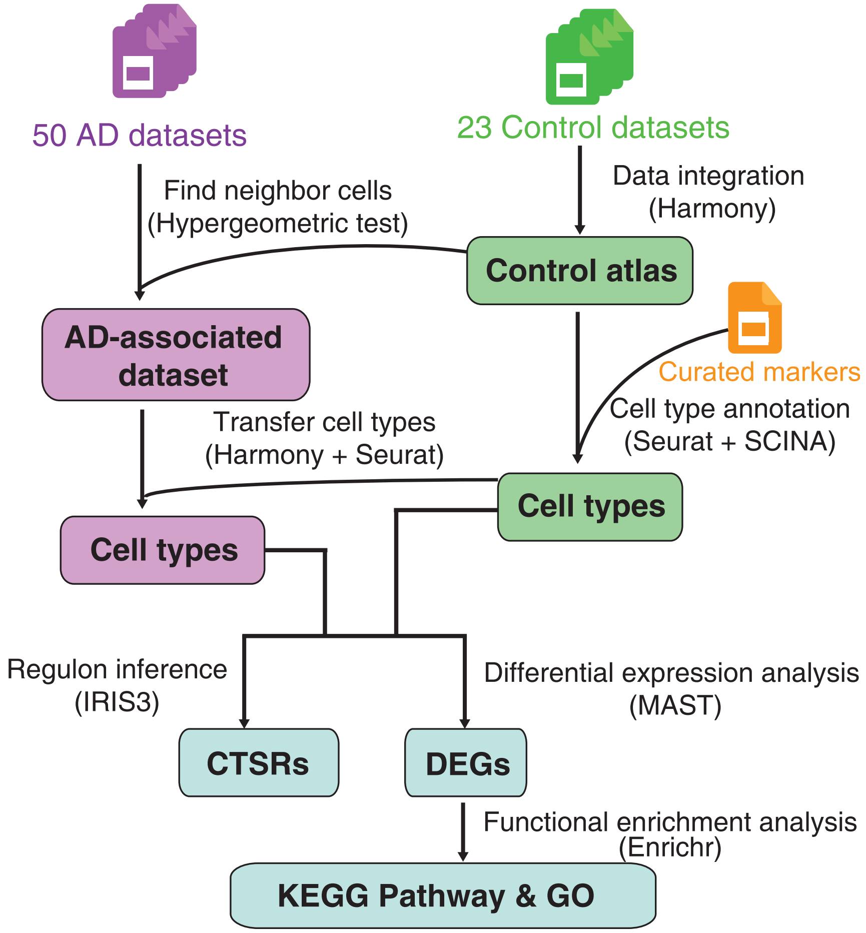


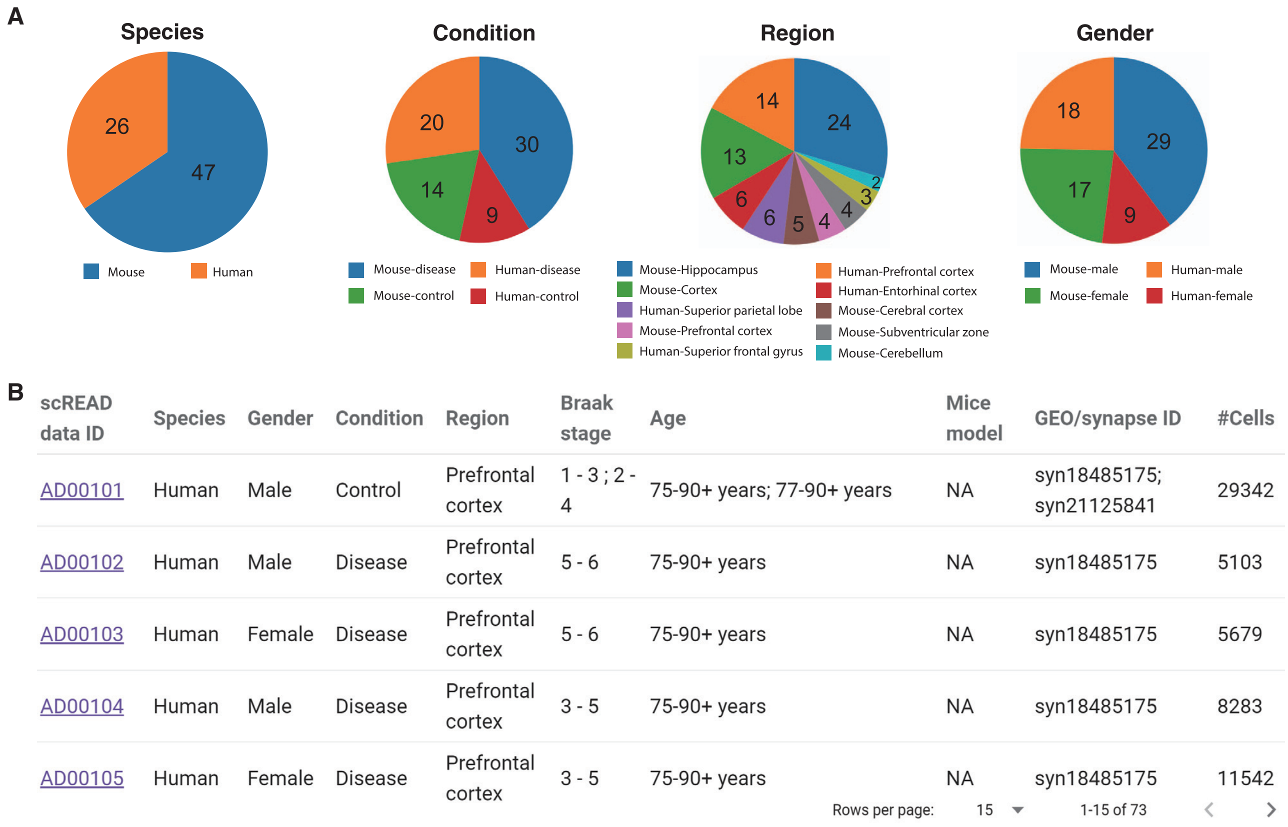
Figure 1. The Workflow of scREAD

has several key features, namely, (i) it is the first-of-its-kind database with a collection of all the 17 existing human and mouse AD scRNA-Seq and snRNA-Seq data sets from the public domain (Table S1); (ii) it re-defines the 17 data sets into 73 data sets, each of which corresponds to a specific species (human or mouse), gender (male or female), brain region (entorhinal cortex, prefrontal cortex, superior frontal gyrus, cortex, cerebellum, subventricular zone, superior parietal lobe, or hippocampus) (Table S2), disease or control, and age stage (7 months, 15 months, or 20 months for mice and 50–100+ years old for human) (Table S3); (iii) it provides comprehensive analysis results for each of the 73 data sets, including but not limited to the construction of control atlas, cell clustering, prediction of cell types, identification of differentially expressed genes (DEGs), and identification of cell-type-specific regulons (CTSRs) in support of the in-depth analysis of heterogeneous regulatory mechanisms; (iv) all these analysis results are visualized through a one-stop and user-friendly interface to free AD biologists from programming burdens (Data S1); (v) the backend workflow enabling all the above computational analyses is freely accessible as stand-alone one-line-command scripts in R (Data S2).

RESULTS

Overview Functionalities of scREAD Database

There are four major functionalities in scREAD: (i) construction of control atlas for different human and mouse brain regions based on the 23 control data sets; (ii) identification of human and mouse disease cell types by projecting the AD data sets onto the control atlases; (iii) identification of DEGs for each cell type among different conditions and functional enrichment analysis of DEGs; and (iv) identification of CTSRs for each cell type among different conditions. These four functions and the schematic workflow of scREAD are shown in Figure 1.



scREAD data ID	Species	Gender	Condition	Region	Braak stage	Age	Mice model	GEO/synapse ID	#Cells
AD00101	Human	Male	Control	Prefrontal cortex	1 - 3 ; 2 - 4	75-90+ years; 77-90+ years	NA	syn18485175; syn21125841	29342
AD00102	Human	Male	Disease	Prefrontal cortex	5 - 6	75-90+ years	NA	syn18485175	5103
AD00103	Human	Female	Disease	Prefrontal cortex	5 - 6	75-90+ years	NA	syn18485175	5679
AD00104	Human	Male	Disease	Prefrontal cortex	3 - 5	75-90+ years	NA	syn18485175	8283
AD00105	Human	Female	Disease	Prefrontal cortex	3 - 5	75-90+ years	NA	syn18485175	11542

Rows per page: 15 ▾ 1-15 of 73 ‹ ›

Figure 2. General Information about scREAD Data sets

(A) General statistical distribution of all the 73 data sets. The pie charts represent four factors of distribution: species, control/disease condition, brain region, and gender from the left side to the right side, respectively. Each pie chart represents one factor, and each color in each pie chart represents one element, and the number represents the number of data sets for each element under each factor for 73 data sets.

(B) General information table on the homepage, which includes nine factors (species, gender, condition, region, Braak stage, age, mice model, GEO/synapse ID, and #cells).

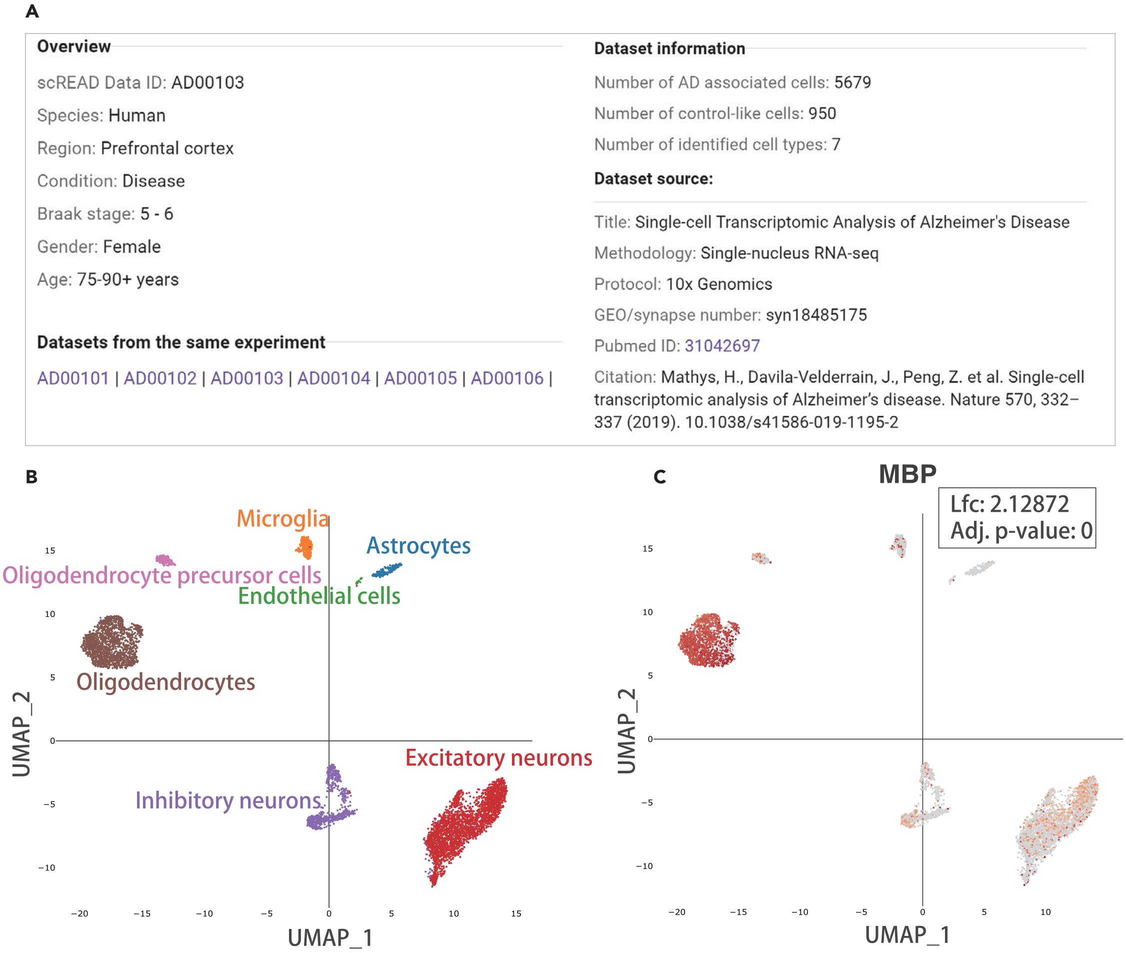
Construction of Control Atlas for Different Human and Mouse Brain Regions

We constructed 23 human and mouse control cell atlases based on 17 scRNA-Seq and snRNA-Seq data sets, which cover 10 brain regions, two genders, and different mouse and human ages, totally 713,640 cells (Figure 2A and Transparent Methods). These 17 data sets were redefined into 73 data sets according to species, gender, brain region, disease or control, and age (Figure 2B). The number of cells and the statistical distribution of these 73 data sets are shown in Figures S1 and S2. Not all the data sets in scREAD are available to download for users; data sets from the GEO database are available to download, but data sets from Synapse are not available to download.

Cell types of these 23 control atlases were assigned using Seurat (Stuart et al., 2019) and Semi-supervised Category Identification and Assignment (SCINA) (Zhang et al., 2019), and the known cell-type marker genes used in this process were collected from literature and PanglaoDB (Franzen et al., 2019) (Table S4 and Transparent Methods). scREAD contains eight major cell types of the human and mouse brain, i.e. astrocytes, endothelial cells, excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, oligodendrocyte precursor cells, and pericytes. These 23 control atlases were then visualized using Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2018) and can be downloaded on the "Browse control atlas" page.

Identification of Human and Mouse Disease Cell Types Based on the Control Atlas

Not all cells collected from AD patient samples are malignant, and there are heterogeneous cells within individual patients, i.e., normal control cells are included. In Granja et al.'s research (Granja et al., 2019),



A

Overview

scREAD Data ID: AD00103

Species: Human

Region: Prefrontal cortex

Condition: Disease

Braak stage: 5 - 6

Gender: Female

Age: 75-90+ years

Datasets from the same experiment

AD00101 | AD00102 | AD00103 | AD00104 | AD00105 | AD00106 |

Dataset information

Number of AD associated cells: 5679

Number of control-like cells: 950

Number of identified cell types: 7

Dataset source:

Title: Single-cell Transcriptomic Analysis of Alzheimer's Disease

Methodology: Single-nucleus RNA-seq

Protocol: 10x Genomics

GEO/synapse number: syn18485175

Pubmed ID: 31042697

Citation: Mathys, H., Davila-Velderrain, J., Peng, Z. et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570, 332–337 (2019). 10.1038/s41586-019-1195-2

Figure 3. Overall Information of an AD Disease Data set (AD00103) and UMAP Plot of the Cell Types and Expression Distribution of Gene *MBP* in This Data set

(A) Overall information on an AD disease data set (AD00103). It includes the information of species, brain region, condition, gender, age, number of control-like and AD-associated cells, data set source, and data sets from the same experiment.

(B) UMAP plot colored by cell type on this AD disease data set. We identified seven cell types, i.e. astrocytes, endothelial cells, excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, and oligodendrocyte precursor cells.

(C) UMAP plot of expression distribution of oligodendrocyte marker gene *MBP* in the same data set. The darker the color is in this UMAP, the higher the expression value of the gene. Adj. p-value: wilcoxon rank sum test, Bonferroni corrected.

they defined these control cells as control-like cells. Here, we applied this concept to AD data sets in our scREAD. These control cells maintain distinct regulatory mechanisms and gene expression patterns compared to AD cells, and they will disturb the accurate identification of AD cell types. Thus, we removed these control cells from AD data sets to identify real AD-associated cells. Using the human and mouse control atlas, we sought to project AD-associated cells onto the control atlas at single-cell resolution to identify human and mouse disease cell types (Transparent Methods). The general information is located at the top of the result page for retrieving an overview of data set description, source, and other data sets from the same experiment in scREAD (Figure 3A). The cell types and subclusters can be visualized interactively on the UMAP plot (Figure 3B) and can be exported in Portable Network Graphics (PNG) format by clicking on the ''save'' button at the right corner. The Adjusted Rand Index or silhouette score is also listed next to the UMAP plots for evaluating the clustering performance (see Transparent Methods) (Lovmar et al., 2005; Steinley et al., 2016). For each gene, the gene expression values are visualized interactively overlaid onto the same UMAP coordinates. For example, *MBP* is the marker gene of oligodendrocyte cell type, and it has higher expression in oligodendrocytes than in other cell types as expected (Figure 3C).

Transcriptional alterations seemed to stem from changes in cell state, with certain cell-type subpopulations more readily captured in AD pathology. To dissect disease-associated cellular subpopulations and cell-type heterogeneity, subclusters were identified for each cell type. For each cell type, we have carried out the subcluster finding analysis for investigating subcluster-specific changes and functional diversity occurring in AD. For different brain regions, enforcing an annotation to the closest cell type is likely to result in misannotation of such regions, but we are aware that subtypes could finely resolve and characterize this problem. Therefore, the subcluster function of our scREAD would provide a more comprehensive cell-type annotation considering cross-region heterogeneity. Due to no standard or consistent annotations available, we have not annotated those subclusters.

Differential Gene Expression and Functional Enrichment Analysis

Differential gene expression analyses include cell-type-specific genes, subcluster-specific genes, and cell-type-pairwise DEGs within one data set or between data sets based on diverse conditions (Table S5) (Monier et al., 2019). All the conditions are in the same species under the same gender, brain region, and age. The DEGs are presented based on the selections of the comparison group and the cell type of interest, and users can drag or type on the panel to apply different parameter cutoffs. The log2 fold-change can be adjusted ranging from 0 to 5, and the adjusted p value range can be adjusted from 10^{-6} to 1. All DEGs are scaled by cell types or conditions and are presented in the tables, allowing users to explore the differential expression of interesting genes across different conditions (Figure 4A).

Functional enrichment analysis is a computational method for inferring knowledge about an input gene set by comparing it to annotated gene sets representing prior biological knowledge. scREAD provides enrichment analysis of the DEGs against Kyoto Encyclopedia of Genes and Genomes pathways and Gene Ontology databases (Figure 4B) (Gene Ontology, 2015; Kanehisa and Goto, 2000). The enrichment analysis is performed and displayed in real time from the DEG list based on the input of the current DEG cutoffs. All of the DEG tables and functional enrichment tables are available to be downloaded by users.

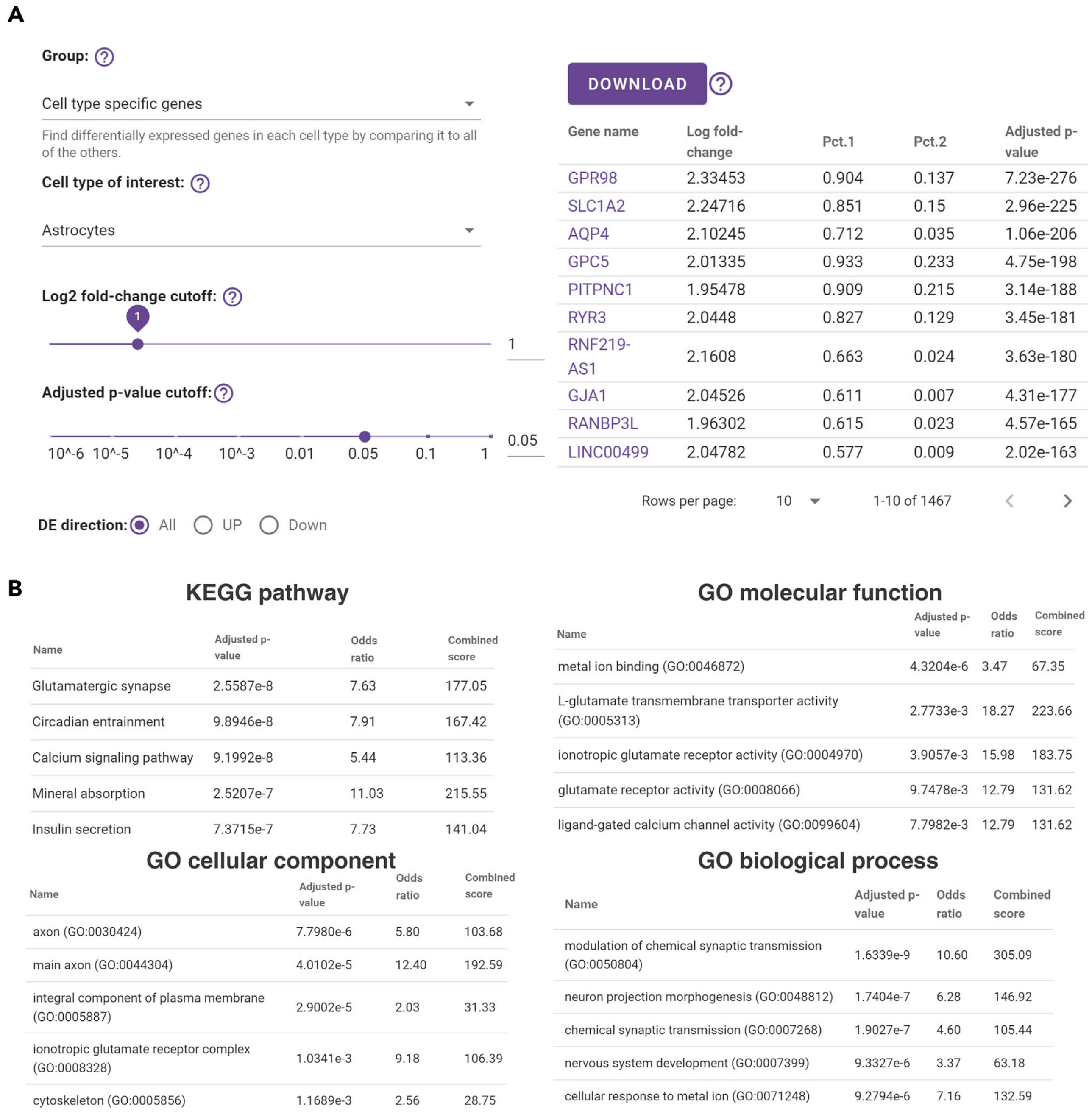
Identification of CTSRs

CTSRs are defined as a group of genes, which receive similar regulatory signals in a specific cell type, hence tending to have similar expression patterns and share conserved motifs in this cell type (Wan et al., 2019; Yang et al., 2019). A successful elucidation of CTSRs will substantially improve the identification of transcriptionally co-regulated gene modules, realistically allowing reliable prediction of global transcription regulation networks encoded in a specific cell type (Ma et al., 2020b; Xie et al., 2020).

scREAD provides both the CTSR result table and the visualization detail information of each CTSR across each cell type for each data set (Figure 5). Taking an AD disease data set (AD00103) as an example, scREAD shows all the identified CTSRs in the table based on the index of cell types and allows users to download this table (Figure 5A). We also display an interactive visualization of all the CTSRs below the result table (Figure 5B). For CT3-R1 (the first regulon in cell type 3), this regulon includes 64 genes co-regulated by the same transcription factor (TF), *MXI1*. CT3-R1 is marked as a CTSR based on a significant regulon specificity score (RSS) of 0.77. Of all the 64 genes, 25 are differentially expressed in CT3 (marked with stars), according to the differential expression analysis using Seurat. Details of each gene and motif can be found by clicking on the gene name and TF logo, respectively. More detailed motif finding results including positions, sequences, and position weight matrix information can be found by clicking the ‘‘Open’’. For each gene in this regulon, the UMAP plot of its expression value across all cell types can be achieved by clicking the ‘‘Display’’ button.

DISCUSSION

In this paper, we described the first release of scREAD, which is as far as we know the first database that collects all existing human and mouse scRNA-Seq and snRNA-Seq data sets with AD pathology and provides a one-stop interactive visualization of the control atlas and analysis results based on these data sets. These data sets have been published and freely accessible in the public domain as of September 22nd, 2020. With the development and application of scRNA-Seq technology, scREAD will continue to be enriched and expanded to be a big database such as SC2disease (Zhao et al., 2020). Furthermore, scREAD allows users to submit a new data set through the submit page to reproduce all the analysis results showcased in scREAD in support of their AD research. We will ask for



Group:

Cell type specific genes

Find differentially expressed genes in each cell type by comparing it to all of the others.

Cell type of interest:

Astrocytes

Log2 fold-change cutoff: 1

Adjusted p-value cutoff: 0.05

DE direction: ● All ○ UP ○ Down

DOWNLOAD

Gene name	Log fold-change	Pct.1	Pct.2	Adjusted p-value
GPR98	2.33453	0.904	0.137	7.23e-276
SLC1A2	2.24716	0.851	0.15	2.96e-225
AQP4	2.10245	0.712	0.035	1.06e-206
GPC5	2.01335	0.933	0.233	4.75e-198
PITPNC1	1.95478	0.909	0.215	3.14e-188
RYR3	2.0448	0.827	0.129	3.45e-181
RNF219-AS1	2.1608	0.663	0.024	3.63e-180
GJA1	2.04526	0.611	0.007	4.31e-177
RANBP3L	1.96302	0.615	0.023	4.57e-165
LINC00499	2.04782	0.577	0.009	2.02e-163

Rows per page: 10 1-10 of 1467

KEGG pathway

Name	Adjusted p-value	Odds ratio	Combined score
Glutamatergic synapse	2.5587e-8	7.63	177.05
Circadian entrainment	9.8946e-8	7.91	167.42
Calcium signaling pathway	9.1992e-8	5.44	113.36
Mineral absorption	2.5207e-7	11.03	215.55
Insulin secretion	7.3715e-7	7.73	141.04

GO molecular function

Name	Adjusted p-value	Odds ratio	Combined score
metal ion binding (GO:0046872)	4.3204e-6	3.47	67.35
L-glutamate transmembrane transporter activity (GO:0005313)	2.7733e-3	18.27	223.66
ionotropic glutamate receptor activity (GO:0004970)	3.9057e-3	15.98	183.75
glutamate receptor activity (GO:0008066)	9.7478e-3	12.79	131.62
ligand-gated calcium channel activity (GO:0099604)	7.7982e-3	12.79	131.62

GO cellular component

Name	Adjusted p-value	Odds ratio	Combined score
axon (GO:0030424)	7.7980e-6	5.80	103.68
main axon (GO:0044304)	4.0102e-5	12.40	192.59
integral component of plasma membrane (GO:0005887)	2.9002e-5	2.03	31.33
ionotropic glutamate receptor complex (GO:0008328)	1.0341e-3	9.18	106.39
cytoskeleton (GO:0005856)	1.1689e-3	2.56	28.75

GO biological process

Name	Adjusted p-value	Odds ratio	Combined score
modulation of chemical synaptic transmission (GO:0050804)	1.6339e-9	10.60	305.09
neuron projection morphogenesis (GO:0048812)	1.7404e-7	6.28	146.92
chemical synaptic transmission (GO:0007268)	1.9027e-7	4.60	105.44
nervous system development (GO:0007399)	9.3327e-6	3.37	63.18
cellular response to metal ion (GO:0071248)	9.2794e-6	7.16	132.59

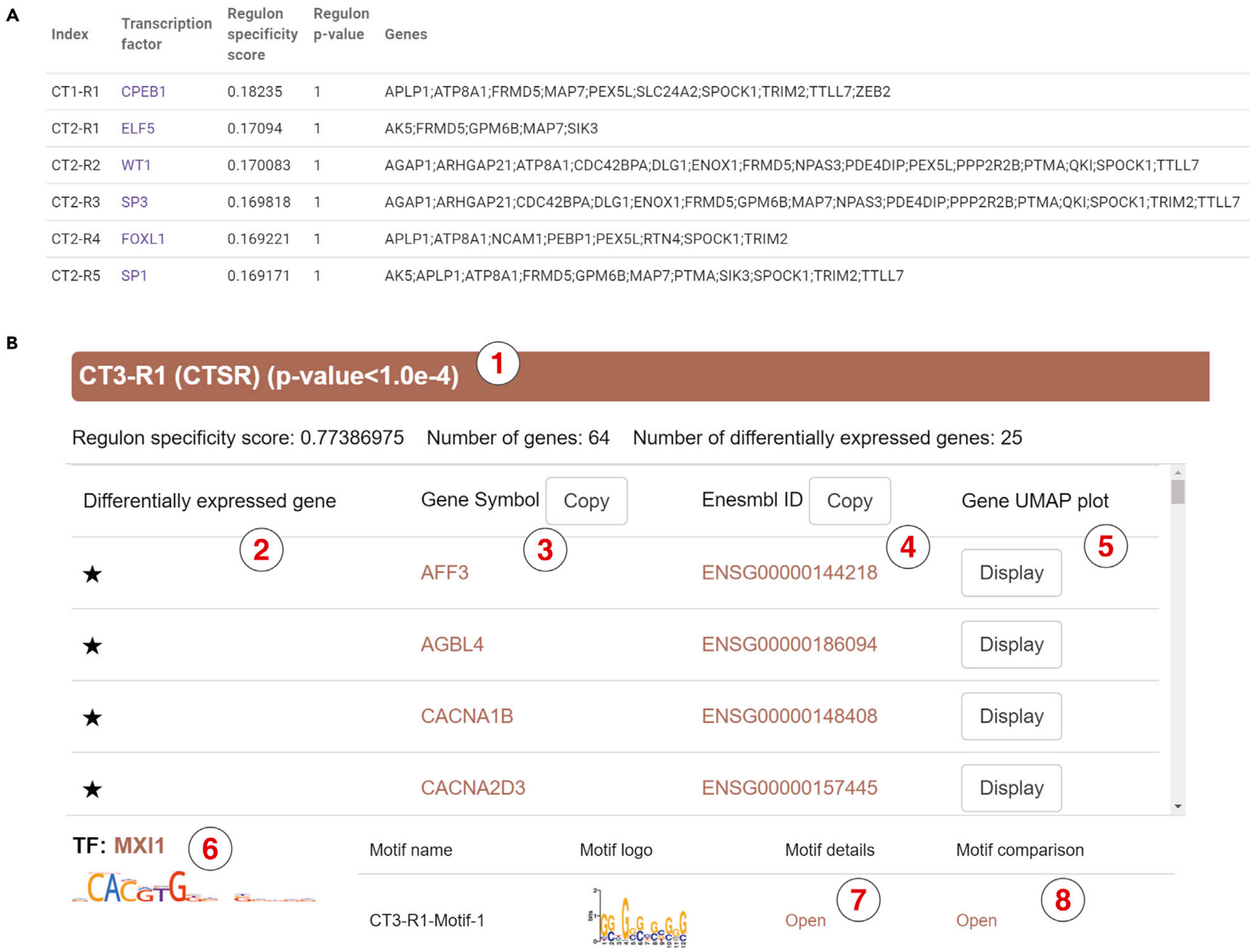
Figure 4. The DEGs and Functional Enrichment Analysis of Selected DEGs on an AD Disease Data set (AD00103)

(A) The DEG panel from the astrocytes cell type using the default parameters (left) and the DEG result table (right). Adjusted p-value: wilcoxon rank sum test, Bonferroni corrected.

(B) The functional enrichment analysis of DEGs. For each functional enrichment analysis, the top five most functional enrichment analysis results are shown. Adjusted p-value: hypergeometric test, Benjamini-Hochberg corrected.

users' permission if we want to store the data uploaded by users into our database. We believe that our database will benefit the AD researchers particularly through studying the data and corresponding analysis results in scREAD.

scREAD provides comprehensive analysis results for those 73 scRNA-Seq data sets collected so far, including the construction of control cell atlas, cell clustering and subclustering, prediction of cell types,



Index	Transcription factor	Regulon specificity score	Regulon p-value	Genes
CT1-R1	CPEB1	0.18235	1	APLP1;ATP8A1;FRMD5;MAP7;PEX5L;SLC24A2;SPOCK1;TRIM2;TTLL7;ZEB2
CT2-R1	ELF5	0.17094	1	AK5;FRMD5;GPM6B;MAP7;SIK3
CT2-R2	WT1	0.170083	1	AGAP1;ARHGAP21;ATP8A1;CDC42BPA;DLG1;ENOX1;FRMD5;NPAS3;PDE4DIP;PEX5L;PPP2R2B;PTMA;QKI;SPOCK1;TTLL7
CT2-R3	SP3	0.169818	1	AGAP1;ARHGAP21;CDC42BPA;DLG1;ENOX1;FRMD5;GPM6B;MAP7;NPAS3;PDE4DIP;PPP2R2B;PTMA;QKI;SPOCK1;TRIM2;TTLL7
CT2-R4	FOXL1	0.169221	1	APLP1;ATP8A1;NCAM1;PEBP1;PEX5L;RTN4;SPOCK1;TRIM2
CT2-R5	SP1	0.169171	1	AK5;APLP1;ATP8A1;FRMD5;GPM6B;MAP7;PTMA;SIK3;SPOCK1;TRIM2;TTLL7

CT3-R1 (CTSR) (p-value<1.0e-4)

Regulon specificity score: 0.77386975 Number of genes: 64 Number of differentially expressed genes: 25

Differentially expressed gene	Gene Symbol Copy	Enesmbl ID Copy	Gene UMAP plot
★	AFF3	ENSG00000144218	Display
★	AGBL4	ENSG00000186094	Display
★	CACNA1B	ENSG00000148408	Display
★	CACNA2D3	ENSG00000157445	Display

TF: MXI1

Motif name	Motif logo	Motif details	Motif comparison
CT3-R1-Motif-1		Open	Open

Figure 5. The CTSR Result Table and Details of the CT3-R1 on an AD Disease Data set (AD00103)

(A) The result table of CTSRs on AD00103.

(B) The details of the top one CTSR of cell type three. (1) A regulon is named as CTn-Rm with n representing the index of cell type and m represents the regulon rank. (2) Asterisks indicate marker genes, that is, the differential expressed gene, identified in each cluster using Seurat. (3) Gene symbols and links to the GeneCards (Human) or the Mouse Genome Informatics (MGI) website. (4) Corresponding gene Ensembl ID columns link to the website. (5) Gene expression UMAP and comparison to the cell types. (6) The corresponding TF with a corresponding link to the HOCOMOCO database. (7) Detailed motif finding results from including positions, sequences, position weight matrix, etc. (8) Motif details linking to the TOMTOM database. Regulon p-value: wilcoxon rank sum test, Bonferroni corrected.

identification of DEGs, and identification of CTSRs. Based on the constructed control cell atlas, we can identify those AD-associated cells at specific brain regions and disease stages. Further analysis of the function/dysfunction of highly heterogeneous cells in the brain at the single-cell level via cell clustering and subclustering, as well as DEG and functional enrichment analysis, can help us understand subcluster-specific changes in the transcriptomic profile and functional diversity occurring in AD. The identification of CTSRs will substantially improve the reliable prediction of global transcription regulation networks encoded in a specific cell type. Thus, scREAD will greatly help the AD community by supporting the in-depth analysis of heterogeneous regulatory mechanisms in AD and identifying the potential therapeutic targets for the prevention and/or treatment of AD.

Limitations of the Study

Currently, scREAD only contains the scRNA-Seq and snRNA-Seq data sets as of September 22nd, 2020, and have not included other omics and spatial transcriptomics data. In the future, we will collect more AD scRNA-Seq and snRNA-Seq data from more brain regions and build up healthy atlas in diverse brain regions of human, mouse and extend to other species. Meanwhile, we will collect AD single-cell omics data, such as scATAC-seq and proteomics data, and achieve more comprehensive analysis results based on single-cell multiple omics data (Li et al., 2020; Ma et al., 2020a). In addition, spatial transcriptomics and *in situ* sequencing have been recently used in studying AD (Chen et al., 2020). The transcriptome-scale spatial gene expression data sets can further provide insights into answering the regional and cellular vulnerability in AD. Thus, we will specifically add the spatial transcriptomics and *in situ* sequencing data sets from human AD and AD-like animals to the current scREAD to enable more functional interpretation. Currently, scREAD only contains nine cell types across 10 human and mouse brain regions; we will provide a more comprehensive cell-type annotation considering more brain regions. In this study, we have only removed the control-like cells in the AD data sets. However, the individuals in the control group may be patients with potential AD, and thus, control samples might also have AD-like cells. Therefore, we will use the same strategy on the control data sets to construct control atlases in the future.

Resource Availability

Lead Contact

Further information and requests for resources should be directed to and will be fulfilled by the Lead Contact, Qin Ma (qin.ma@osumc.edu) or Hongjun Fu (hongjun.fu@osumc.edu).

Materials Availability

This study did not generate new unique data.

Data and Code Availability

All data sets used in this work are available from publicly available sources as cited in the manuscript. scREAD is a one-stop and user-friendly interface and freely available at https://bmbls.bmi.osumc.edu/scread/. The backend workflow can be downloaded from https://github.com/OSU-BMBL/scread/tree/master/script to enable more discovery-driven analyses.

METHODS

All methods can be found in the accompanying Transparent Methods supplemental file.

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.isci.2020.101769.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health [R01 GM131399-01, Q.M.; K01 AG056673, H.F.], the Department of Defense [W81XWH1910309, H.F.], the Alzheimer's Association [AARF-17-505009, H.F.], and the Neuroscience Research Institute Pilot Award and the Chronic Brain Injury Pilot Award from The Ohio State University [H.F.]. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Science Foundation and the National Institutes of Health.

AUTHOR CONTRIBUTIONS

Q.M. and H.F. designed the manuscript contents and experiments. J.J. contributed to data analysis and the initial draft. C.W. developed and implemented the database and the API. Q.R. tested the database and curated part of the data. All authors revised the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 28, 2020
Revised: October 22, 2020
Accepted: October 30, 2020
Published: November 20, 2020

REFERENCES

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al. (2013). NCBI GEO: archive for functional genomics data sets–update. Nucleic Acids Res. *41*, D991–D995.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44.

Chen, W.T., Lu, A., Craessaerts, K., Pavie, B., Sala Frigerio, C., Corthout, N., Qian, X., Lalakova, J., Kuhnemund, M., Voytyuk, I., et al. (2020). Spatial transcriptomics and in situ sequencing to study Alzheimer's disease. Cell *182*, 976–991.

Claxton, A., Baker, L.D., Hanson, A., Trittschuh, E.H., Cholerton, B., Morgan, A., Callaghan, M., Arbuckle, M., Behl, C., and Craft, S. (2015). Long-acting intranasal insulin detemir improves cognition for adults with mild cognitive impairment or early-stage Alzheimer's disease dementia. J. Alzheimer's Dis. *44*, 897–906.

Del-Aguila, J.L., Li, Z., Dube, U., Mihindukulasuriya, K.A., Budde, J.P., Fernandez, M.V., Ibanez, L., Bradley, J., Wang, F., Bergmann, K., et al. (2019). A single-nuclei RNA sequencing study of Mendelian and sporadic AD in the human brain. Alzheimer's Res. Ther. *11*, 71.

Dolgin, E. (2018). Alzheimer's disease is getting easier to spot. Nature *559*, S10–S12.

Ewers, M., Sperling, R.A., Klunk, W.E., Weiner, M.W., and Hampel, H. (2011). Neuroimaging markers for the prediction and early diagnosis of Alzheimer's disease dementia. Trends Neurosciences *34*, 430–442.

Franzen, O., Gan, L.M., and Bjorkegren, J.L.M. (2019). PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database *2019*, baz046.

Gao, L.B., Yu, X.F., Chen, Q., and Zhou, D. (2016). Alzheimer's Disease therapeutics: current and future therapies. Minerva Med. *107*, 108–113.

Gene Ontology, C. (2015). Gene Ontology consortium: going forward. Nucleic Acids Res. *43*, D1049–D1056.

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y., et al. (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nat. Biotechnol. *37*, 1458–1465.

Grubman, A., Chew, G., Ouyang, J.F., Sun, G., Choo, X.Y., McLean, C., Simmons, R.K., Buckberry, S., Vargas-Landin, D.B., Poppe, D., et al. (2019). A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. Nat. Neurosci. *22*, 2087–2097.

Kanehisa, M., and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. *28*, 27–30.

Li, Y., Ma, A., Mathe, E.A., Li, L., Liu, B., and Ma, Q. (2020). Elucidation of biological networks across complex diseases using single-cell omics. Trends Genetics.

Lovmar, L., Ahlford, A., Jonsson, M., and Syvanen, A.C. (2005). Silhouette scores for assessment of SNP genotype clusters. BMC Genomics *6*, 35.

Ma, A., McDermaid, A., Xu, J., Chang, Y., and Ma, Q. (2020a). Integrative methods and practical challenges for single-cell multi-omics. Trends Biotechnol. *38*, 1007–1022.

Ma, A., Wang, C., Chang, Y., Brennan, F.H., McDermaid, A., Liu, B., Zhang, C., Popovich, P.G., and Ma, Q. (2020b). IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. Nucleic Acids Res. *48*, W275–W286.

Mathys, H., Adaikkan, C., Gao, F., Young, J.Z., Manet, E., Hemberg, M., De Jager, P.L., Ransohoff, R.M., Regev, A., and Tsai, L.H. (2017). Temporal tracking of microglia activation in neurodegeneration at single-cell resolution. Cell Rep. *21*, 366–380.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.

Monier, B., McDermaid, A., Wang, C., Zhao, J., Miller, A., Fennell, A., and Ma, Q. (2019). IRIS-EDA: an integrated RNA-Seq interpretation system for gene expression data analysis. PLoS Comput. Biol. *15*, e1006792.

Mucke, L. (2009). Neuroscience: Alzheimer's disease. Nature *461*, 895–897.

Qi, R., Ma, A., Ma, Q., and Zou, Q. (2020). Clustering and classification methods for single-cell RNA-sequencing data. Brief. Bioinform. *21*, 1196–1208.

Sala Frigerio, C., Wolfs, L., Fattorelli, N., Thrupp, N., Voytyuk, I., Schmidt, I., Mancuso, R., Chen, W.T., Woodbury, M.E., Srivastava, G., et al. (2019). The major risk factors for Alzheimer's disease: age, sex, and genes modulate the microglia response to abeta Plaques. Cell Rep. *27*, 1293–1306.e6.

Seweryn, M.T., Pietrzak, M., and Ma, Q. (2020). Application of information theoretical approaches to assess diversity and similarity in single-cell transcriptomics. Comput. Struct. Biotechnol. J. *18*, 1830–1837.

Shinagawa, S. (2016). [Language symptoms of Alzheimer's disease]. Brain and nerve = Shinkei kenkyu no shinpo *68*, 551–557.

Steinley, D., Brusco, M.J., and Hubert, L. (2016). The variance of the adjusted Rand index. Psychol. Methods *21*, 261–272.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902 e1821.

Wan, C., Chang, W., Zhang, Y., Shah, F., Lu, X., Zang, Y., Zhang, A., Cao, S., Fishel, M.L., Ma, Q., et al. (2019). LTMG: a novel statistical modeling of transcriptional expression states in single-cell RNA-Seq data. Nucleic Acids Res. *47*, e111.

Wang, J., Ma, A., Chang, Y., Gong, J., Jiang, Y., Fu, H., Wang, C., Qi, R., Ma, Q., and Xu, D. (2020). scGNN: a novel graph neural network framework for single-cell RNA-Seq analyses. bioRxiv.

Wu, A.R., Neff, N.F., Kalisky, T., Dalerba, P., Treutlein, B., Rothenberg, M.E., Mburu, F.M., Mantalas, G.L., Sim, S., Clarke, M.F., et al. (2014). Quantitative assessment of single-cell RNA-sequencing methods. Nat. Methods *11*, 41–46.

Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., Xu, J., Zhang, C., and Ma, Q. (2020). QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data. Bioinformatics *36*, 1143–1149.

Yang, J., Ma, A., Hoppe, A.D., Wang, C., Li, Y., Zhang, C., Wang, Y., Liu, B., and Ma, Q. (2019). Prediction of regulatory motifs from human Chip-sequencing data using a deep learning framework. Nucleic Acids Res. *47*, 7809–7824.

Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z., et al. (2019). SCINA: a semi-supervised subtyping algorithm of single cells and bulk samples. Genes *10*, 531.

Zhao, T., Lyu, S., Lu, G., Juan, L., Zeng, X., Wei, Z., Hao, J., and Peng, J. (2020). SC2disease: a manually curated database of single-cell transcriptome for human diseases. Nucleic Acids Res. gkaa838.

iScience, Volume *23*

Supplemental Information

scREAD: A Single-Cell RNA-Seq Database for Alzheimer's Disease

Jing Jiang, Cankun Wang, Ren Qi, Hongjun Fu, and Qin Ma

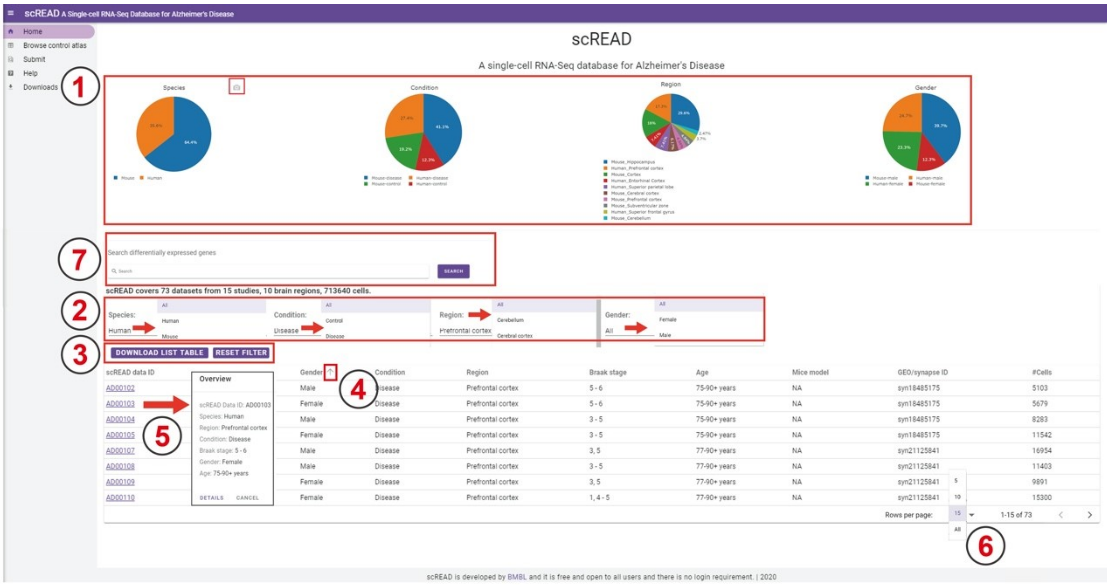
Supplemental Information

Data S1: scREAD server tutorials, Related to Figure 1.

The scREAD server includes six parts:

1. Home
   - Pie charts that reflect ratio distribution in 73 datasets for each of the four factors (species, condition, region, and gender)
   - Search differentially expressed genes
   - Dataset overview
2. Example result illustration
   - A general overview of the dataset including dataset source, and other datasets from the same experiment
   - Interactive UMAP plot for cell types, subclusters, and specific gene expression
   - Differential expression and Gene set enrichment analysis
   - Cell-type-specific regulon inference
2. Browse control atlas
   - 23 control atlases from different brain regions of human and mouse species
3. Submit
   - Submit user's AD scRNA-Seq & snRNA-Seq datasets into scREAD to do the same analysis as shown in our database
4. Download raw and processed datasets

Part 1. Home page



scREAD covers 73 datasets from 15 studies, 10 brain regions, 713640 cells.

Species: Human
Condition: Disease
Region: Prefrontal cortex
Gender: All

scREAD data ID	Gender	Condition	Region	Braak stage	Age	Mice model	GEO/synapse ID	#Cells
AD00102	Male	Disease	Prefrontal cortex	5 - 6	75-90+ years	NA	syn18485175	5103
AD00103	Female	Disease	Prefrontal cortex	5 - 6	75-90+ years	NA	syn18485175	5679
AD00104	Male	Disease	Prefrontal cortex	3 - 5	75-90+ years	NA	syn18485175	8283
AD00105	Female	Disease	Prefrontal cortex	3 - 5	75-90+ years	NA	syn18485175	11542
AD00107	Male	Disease	Prefrontal cortex	3, 5	77-90+ years	NA	syn21125841	16954
AD00108	Male	Disease	Prefrontal cortex	3 - 5	77-90+ years	NA	syn21125841	11403
AD00109	Female	Disease	Prefrontal cortex	3, 5	77-90+ years	NA	syn21125841	9891
AD00110	Female	Disease	Prefrontal cortex	1, 4 - 5	77-90+ years	NA	syn21125841	15300

Overview

scREAD Data ID: AD00103
Species: Human
Region: Prefrontal cortex
Condition: Disease
Braak stage: 5 - 6
Gender: Female
Age: 75-90+ years

scREAD is developed by BMBL and it is free and open to all users and there is no login requirement. | 2020

1. General statistical information of all scRNA-Seq & snRNA-Seq datasets that are covered in scREAD. The pie charts represent four factors of distribution: species, control/disease condition, brain region, and gender.
2. Options to filter presented datasets.
3. Download the current presented table or reset all filters to display all datasets.
4. Each column is sortable by clicking column names.

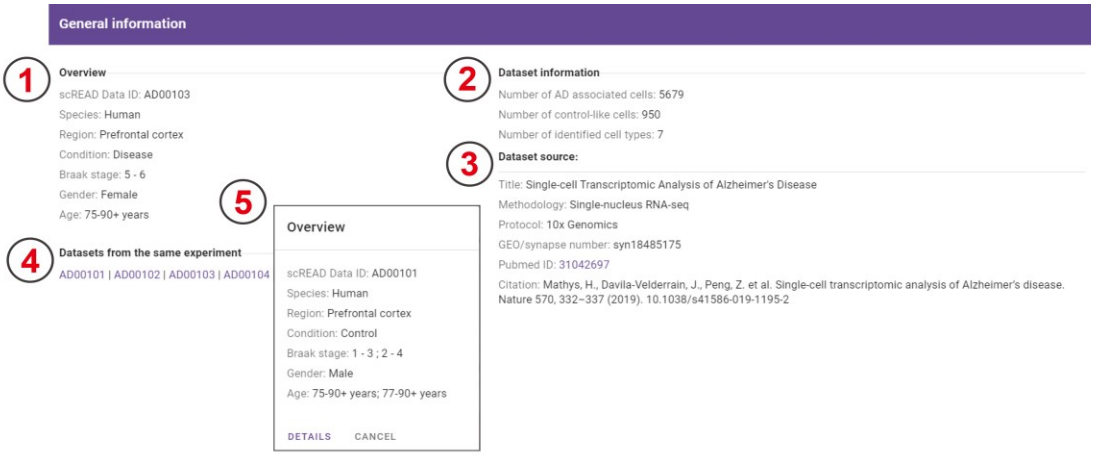
5. A floating dialog for dataset overview will pop up when users click on each row, users can then navigate to the details page.
6. Navigate to different pages or control how many queries on one page.
7. scREAD will return all differential gene expression results queries for a gene.

Part 2. Example result illustration

We used the dataset of AD00103 as an example to show the analysis result. This dataset consists of 6,629 cells isolated from the human AD female prefrontal cortex (Mathys et al., 2019).

This tutorial will guide you through the analysis result page of scREAD in detail.

2.1 General information



General information

Overview

scREAD Data ID: AD00103

Species: Human

Region: Prefrontal cortex

Condition: Disease

Braak stage: 5 - 6

Gender: Female

Age: 75-90+ years

Datasets from the same experiment

AD00101 | AD00102 | AD00103 | AD00104

Overview

scREAD Data ID: AD00101

Species: Human

Region: Prefrontal cortex

Condition: Control

Braak stage: 1 - 3 ; 2 - 4

Gender: Male

Age: 75-90+ years; 77-90+ years

DETAILS CANCEL

Dataset information

Number of AD associated cells: 5679

Number of control-like cells: 950

Number of identified cell types: 7

Dataset source:

Title: Single-cell Transcriptomic Analysis of Alzheimer's Disease

Methodology: Single-nucleus RNA-seq

Protocol: 10x Genomics

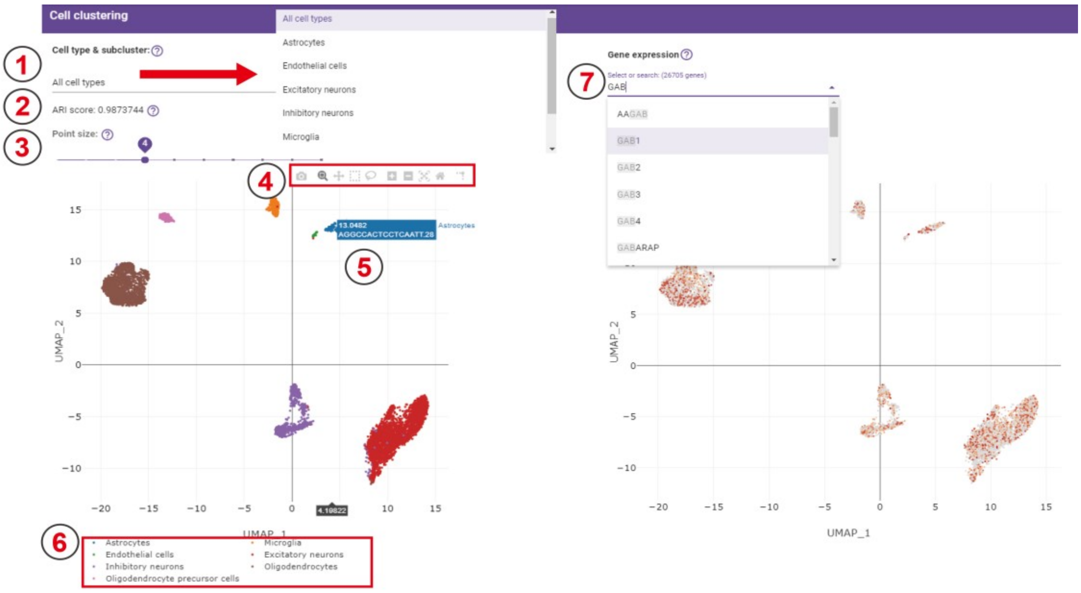
GEO/synapse number: syn18485175

Pubmed ID: 31042697

Citation: Mathys, H., Davila-Velderrain, J., Peng, Z. et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature 570, 332–337 (2019). 10.1038/s41586-019-1195-2

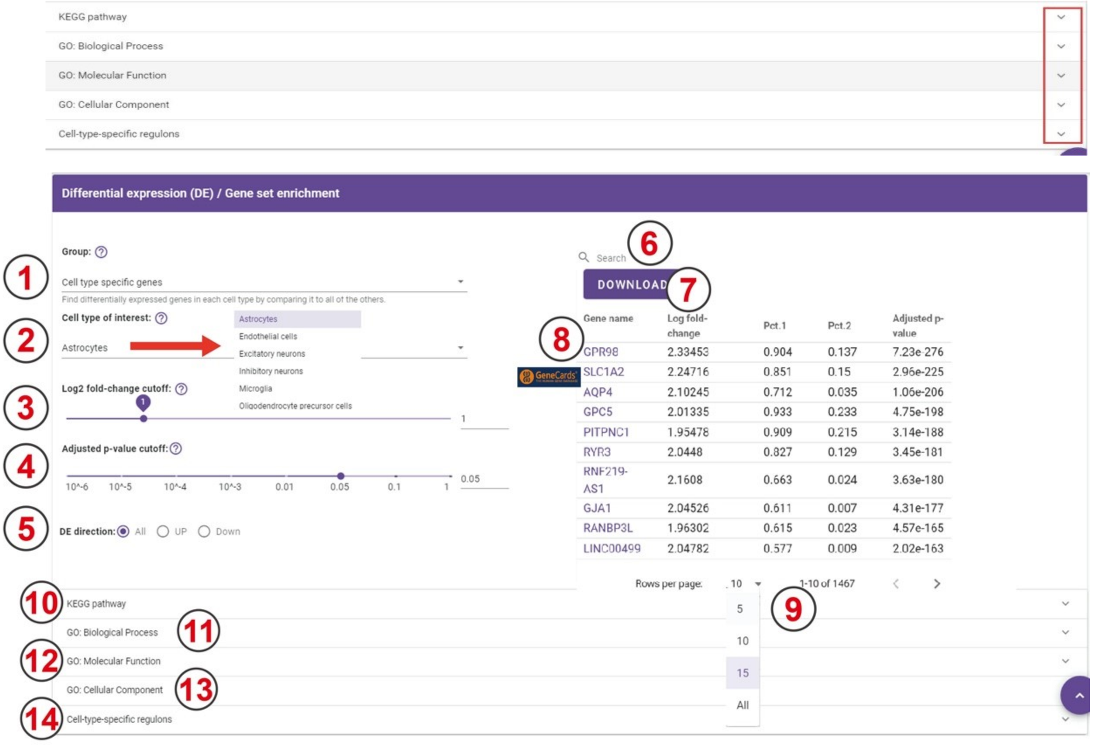
1. Overview of current dataset: 'scREAD Data ID', 'Species', 'Region', 'Condition', 'Braak Stage', 'Gender', and 'Age'.
2. The number of identified cell types, control-like cells, and AD-associated cells.
3. General information of the corresponding research paper for this dataset.
4. All the other datasets that are included in the same experiment or publication.
5. A dialog will appear when users click on the scREAD Data ID, then users can click the 'DETAILS' button to go to the analysis result page of the corresponding dataset.

2.2 Cell clustering



1. Users choose one of these cell types, the following UMAP will change to the UMAP of predicted subclusters for this specific cell type.
2. The ARI score is used to evaluate the performance of our predicted cell types compared with the original cell labels from the original paper. Note: If we don't have the ARI score, it will show a silhouette score instead.
3. A sliding bar is used for controlling the size of each point in the following UMAP. It ranges from 1 to 10, the bigger the number is, the larger the point size is.
4. This function bar contains several quick buttons for graphic operations.
5. Hover cursor on cell points will display cell type, cell name, and the UMAP coordinates.
6. The legend of this UMAP plot.
7. The genes in the drop-down bar are all genes expressed in this dataset, and users can also input the name of genes that they're interested in. The darker the color is in this UMAP, the higher the expression value of the gene.

2.3 Differential expression (DE) / Gene set enrichment



KEGG pathway
GO: Biological Process
GO: Molecular Function
GO: Cellular Component
Cell-type-specific regulons

Differential expression (DE) / Gene set enrichment

Group: ?
Cell type specific genes
Find differentially expressed genes in each cell type by comparing it to all of the others.

Cell type of interest: ?
Astrocytes
Astrocytes
Endothelial cells
Excitatory neurons
Inhibitory neurons
Microglia
Oligodendrocyte precursor cells

Log2 fold-change cutoff: ? 1

Adjusted p-value cutoff: ? 0.05
10^-6 10^-5 10^-4 10^-3 0.01 0.05 0.1 1

DE direction: ● All ○ UP ○ Down

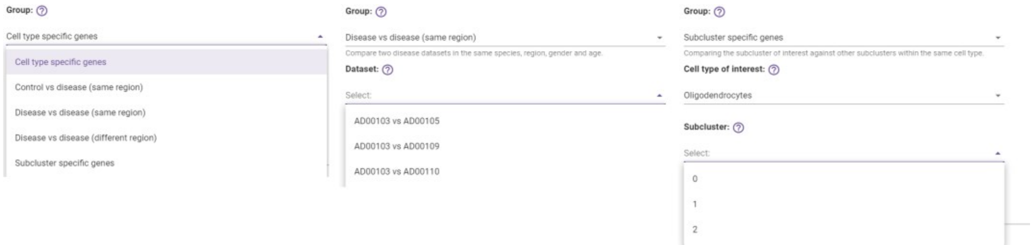
Search
DOWNLOAD

Gene name	Log fold-change	Pct.1	Pct.2	Adjusted p-value
GPR98	2.33453	0.904	0.137	7.23e-276
SLC1A2	2.24716	0.851	0.15	2.96e-225
AQP4	2.10245	0.712	0.035	1.06e-206
GPC5	2.01335	0.933	0.233	4.75e-198
PITPNC1	1.95478	0.909	0.215	3.14e-188
RYR3	2.0448	0.827	0.129	3.45e-181
RNF219-AS1	2.1608	0.663	0.024	3.63e-180
GJA1	2.04526	0.611	0.007	4.31e-177
RANBP3L	1.96302	0.615	0.023	4.57e-165
LINC00499	2.04782	0.577	0.009	2.02e-163

Rows per page: 10 1-10 of 1467
5
10
15
All

KEGG pathway
GO: Biological Process
GO: Molecular Function
GO: Cellular Component
Cell-type-specific regulons

1. DE testing groups for browsing cell-type-specific genes, subcluster specific genes, and DE genes from the cross-dataset comparison.



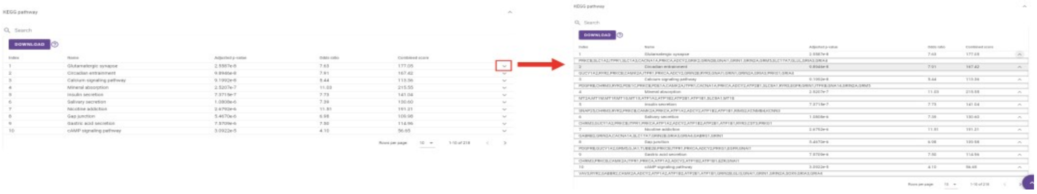
Group: ?
Cell type specific genes
Cell type specific genes
Control vs disease (same region)
Disease vs disease (same region)
Disease vs disease (different region)
Subcluster specific genes

Group: ?
Disease vs disease (same region)
Compare two disease datasets in the same species, region, gender and age.
Dataset: ?
Select:
AD00103 vs AD00105
AD00103 vs AD00109
AD00103 vs AD00110

Group: ?
Subcluster specific genes
Comparing the subcluster of interest against other subclusters within the same cell type.
Cell type of interest: ?
Oligodendrocytes
Subcluster: ?
Select:
0
1
2

2. Choose the cell type of interest in DE testing.
3. The Log2 fold-change ranges from 0 to 5.
4. The Adjusted p-value ranges from 10^-6 to 1.
5. The DE direction can filter by all DE genes, only up-regulated genes, only down-regulated genes.
6. Users can search for genes that they are interested in, and then the following table will return the matching result.
7. Download the currently listed table.
8. GeneCards database is linked to each gene in the table.
9. Set how many rows should the table show.

10. KEGG pathway enrichment analysis result table of the DEGs will appear when users click the inverted triangle, and this table can be downloaded when they click the 'Download' button. When users click the reversed triangle at the end of each row in this table, it shows the genes that are enriched on this pathway, and this table can be downloaded when they click the 'Download' button. Users can also search for a specific item by entering the content they want to search in the search box.



KEGG pathway

Search

DOWNLOAD

Index	Name	Adjusted p-value	Odds ratio	Combined score
1	Glutamatergic synapse	2.5587e-8	7.63	177.05
2	Circadian entrainment	9.8946e-8	7.91	167.42
3	Calcium signaling pathway	9.1992e-8	5.44	113.56
4	Mineral absorption	2.5207e-7	11.03	215.55
5	Insulin secretion	7.3718e-7	7.73	141.04
6	Salivary secretion	1.0808e-6	7.39	130.60
7	Nicotine addiction	2.6792e-6	11.81	191.21
8	Gap junction	5.4670e-6	6.98	109.96
9	Gastric acid secretion	7.3709e-6	7.30	114.96
10	cAMP signaling pathway	3.0922e-5	4.10	56.65

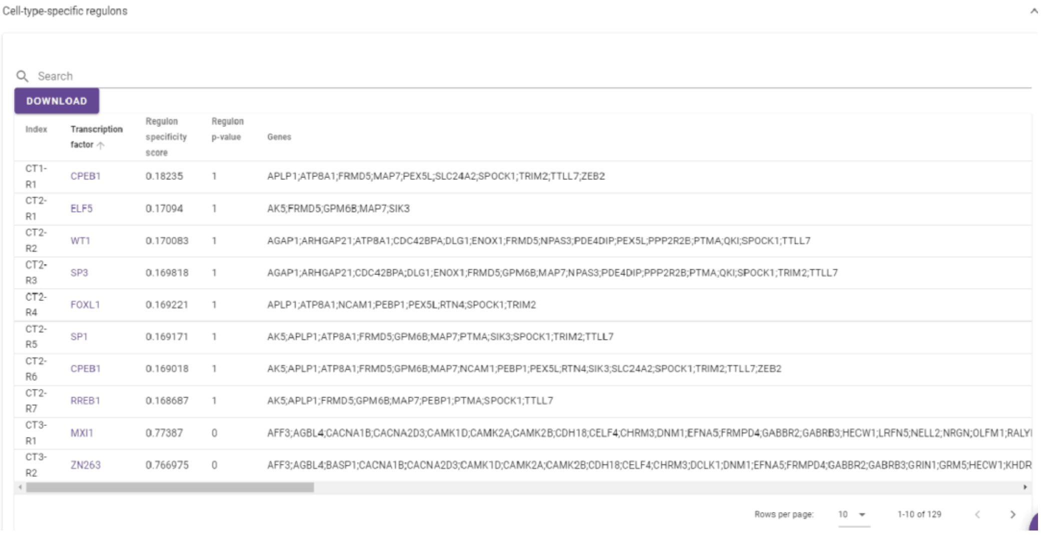
Rows per page: 10 1-10 of 218

11. GO biological process analysis result of the DEGs. Please see entry 10 to know how to use it.

12. GO molecular function analysis result of the DEGs. Please see entry 10 to know how to use it.

13. GO cellular component analysis result of the DEGs. Please see entry 10 to know how to use it.

14. Cell-type-specific regulon analysis result table of this dataset will appear when users click the inverted triangle, and this result only shows up when they choose the 'Cell-type-specific genes' in the 'Group' drop-down bar.



Cell-type-specific regulons

Search

DOWNLOAD

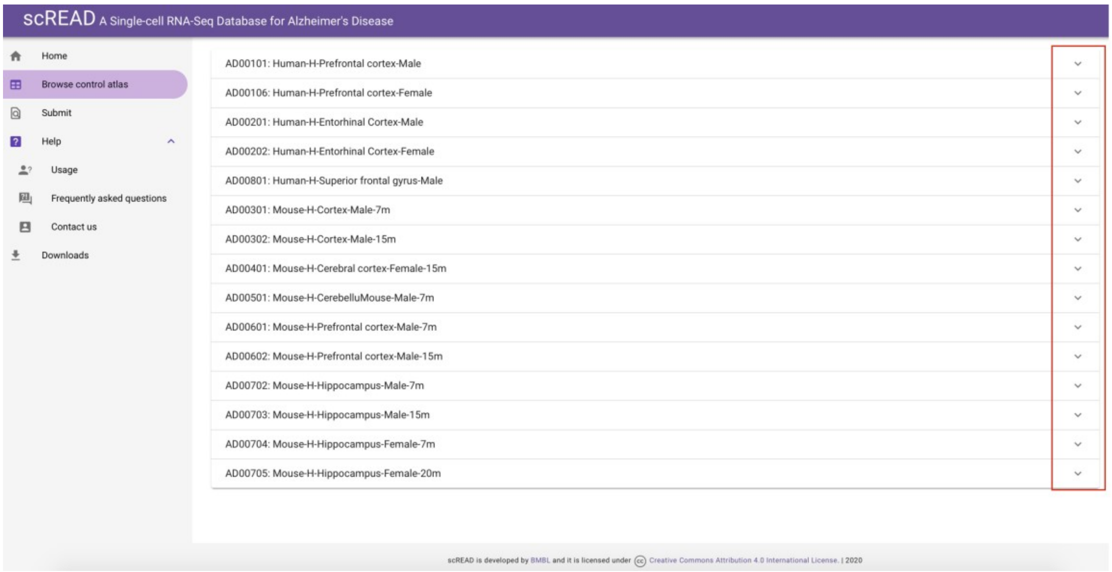
Index	Transcription factor ↑	Regulon specificity score	Regulon p-value	Genes
CT1-R1	CPEB1	0.18235	1	APLP1;ATP8A1;FRMD5;MAP7;PEX5L;SLC24A2;SPOCK1;TRIM2;TTLL7;ZEB2
CT2-R1	ELF5	0.17094	1	AK5;FRMD5;GPM6B;MAP7;SIK3
CT2-R2	WT1	0.170083	1	AGAP1;ARHGAP21;ATP8A1;CDC42BPA;DLG1;ENOX1;FRMD5;NPAS3;PDE4DIP;PEX5L;PPP2R2B;PTMA;QKI;SPOCK1;TTLL7
CT2-R3	SP3	0.169818	1	AGAP1;ARHGAP21;CDC42BPA;DLG1;ENOX1;FRMD5;GPM6B;MAP7;NPAS3;PDE4DIP;PPP2R2B;PTMA;QKI;SPOCK1;TRIM2;TTLL7
CT2-R4	FOXL1	0.169221	1	APLP1;ATP8A1;NCAM1;PEBP1;PEX5L;RTN4;SPOCK1;TRIM2
CT2-R5	SP1	0.169171	1	AK5;APLP1;ATP8A1;FRMD5;GPM6B;MAP7;PTMA;SIK3;SPOCK1;TRIM2;TTLL7
CT2-R6	CPEB1	0.169018	1	AK5;APLP1;ATP8A1;FRMD5;GPM6B;MAP7;NCAM1;PEBP1;PEX5L;RTN4;SIK3;SLC24A2;SPOCK1;TRIM2;TTLL7;ZEB2
CT2-R7	RREB1	0.168687	1	AK5;APLP1;FRMD5;GPM6B;MAP7;PEBP1;PTMA;SPOCK1;TTLL7
CT3-R1	MXI1	0.77387	0	AFF3;AGBL4;CACNA1B;CACNA2D3;CAMK1D;CAMK2A;CAMK2B;CDH18;CELF4;CHRM3;DNM1;EFNA5;FRMPD4;GABBR2;GABRB3;HECW1;LRFN5;NELL2;NRGN;OLFM1;RALYI
CT3-R2	ZN263	0.766975	0	AFF3;AGBL4;BASP1;CACNA1B;CACNA2D3;CAMK1D;CAMK2A;CAMK2B;CDH18;CELF4;CHRM3;DCLK1;DNM1;EFNA5;FRMPD4;GABBR2;GABRB3;GRIN1;GRM5;HECW1;KHDR

Rows per page: 10 1-10 of 129

This is the cell-type-specific regulon result table for each cell type, and this table can be downloaded when users click the 'Download' button.

Part 3. Browse control atlas page

The 'Browse control atlas' page contains all the 23 control atlases that are stored in the scREAD based on different brain regions for different species and different mouse ages.


scREAD A Single-cell RNA-Seq Database for Alzheimer's Disease
Home
Browse control atlas
Submit
Help
Usage
Frequently asked questions
Contact us
Downloads

AD00101: Human-H-Prefrontal cortex-Male

AD00106: Human-H-Prefrontal cortex-Female

AD00201: Human-H-Entorhinal Cortex-Male

AD00202: Human-H-Entorhinal Cortex-Female

AD00801: Human-H-Superior frontal gyrus-Male

AD00301: Mouse-H-Cortex-Male-7m

AD00302: Mouse-H-Cortex-Male-15m

AD00401: Mouse-H-Cerebral cortex-Female-15m

AD00501: Mouse-H-CerebelluMouse-Male-7m

AD00601: Mouse-H-Prefrontal cortex-Male-7m

AD00602: Mouse-H-Prefrontal cortex-Male-15m

AD00702: Mouse-H-Hippocampus-Male-7m

AD00703: Mouse-H-Hippocampus-Male-15m

AD00704: Mouse-H-Hippocampus-Female-7m

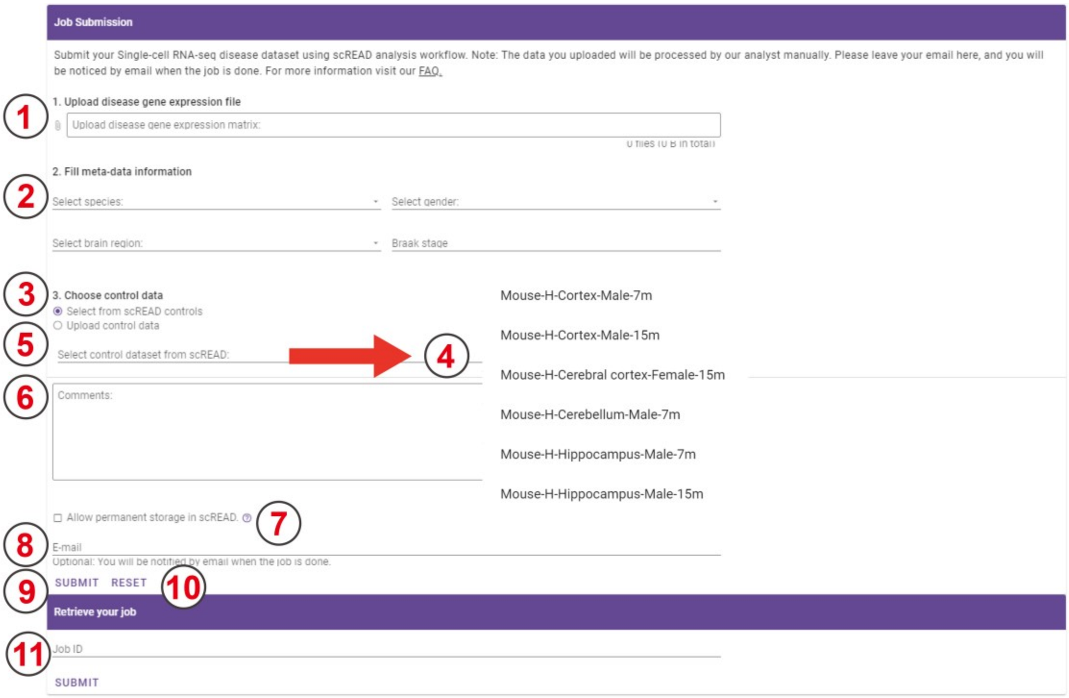
AD00705: Mouse-H-Hippocampus-Female-20m

scREAD is developed by BMBL and it is licensed under Creative Commons Attribution 4.0 International License. | 2020

These are 23 control atlases entries. The default pattern is all the UMAP of control atlases are folded, however, users can click the reverses triangle to unfold the UMAP of each control atlas. The top five control and bottom five atlases are human control atlases, and the rest control atlases are mouse control atlases.

Part 4. Submit page

The submission of a new entry is welcome, and it can be done on the “submit” page. One scRNA-Seq file of AD disease should be uploaded, and one scRNA-Seq file of control can be uploaded or not.



Job Submission

Submit your Single-cell RNA-seq disease dataset using scREAD analysis workflow. Note: The data you uploaded will be processed by our analyst manually. Please leave your email here, and you will be noticed by email when the job is done. For more information visit our FAQ.

1. Upload disease gene expression file

Upload disease gene expression matrix:

0 files (0 B in total)

2. Fill meta-data information

Select species: Select gender:

Select brain region: Braak stage

3. Choose control data

● Select from scREAD controls

○ Upload control data

Select control dataset from scREAD:

Mouse-H-Cortex-Male-7m

Mouse-H-Cortex-Male-15m

Mouse-H-Cerebral cortex-Female-15m

Mouse-H-Cerebellum-Male-7m

Mouse-H-Hippocampus-Male-7m

Mouse-H-Hippocampus-Male-15m

Comments:

□ Allow permanent storage in scREAD.

E-mail

Optional: You will be notified by email when the job is done.

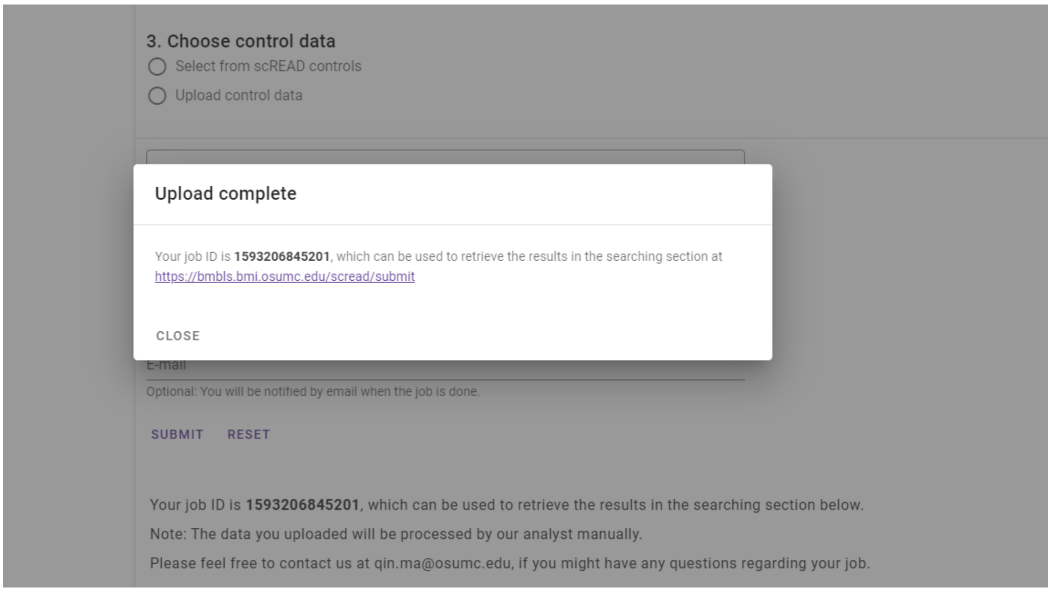
SUBMIT RESET

Retrieve your job

Job ID

SUBMIT

1. Upload your AD scRNA-Seq expression matrix file by selecting the file stored on your computer. Note: This file is required if you want to analyze new data. Note: The format of your uploaded file should be a text format.
2. You can provide species, gender, brain region, and Braak stage these four types of information of your input gene expression dataset to scREAD.
3. You can select one of the control datasets as a reference control atlas to do the downstream analysis by choosing the 'Select from scREAD controls' option.
4. These are all the 23 control files that are stored in scREAD to produce all control atlas.
5. You can also upload your control dataset if you have one to do the comparison within your own paired dataset by selecting 'Upload control data' and then click the bar.
6. If you have any comments about scREAD, we will be appreciative that you can write your comments here.
7. Clicking this option, it means you allow us to store your data in scREAD (both datasets and results) for the future database construction. Be cautious if your data have not been published.
8. An email is not required to submit the job; however, we strongly suggest you provide your email because the data you uploaded will be processed by our analyst manually. So you will be noticed by email when the job is done.
9. Submit the job once everything is ready. If you have provided your email to us you will receive an email after you submit your job successfully. The job ID is in the showed up floating window, which can be used to retrieve the results.



3. Choose control data

○ Select from scREAD controls

○ Upload control data

Upload complete

Your job ID is **1593206845201**, which can be used to retrieve the results in the searching section at https://bmbls.bmi.osumc.edu/scread/submit

CLOSE

E-mail

Optional: You will be notified by email when the job is done.

SUBMIT RESET

Your job ID is **1593206845201**, which can be used to retrieve the results in the searching section below.

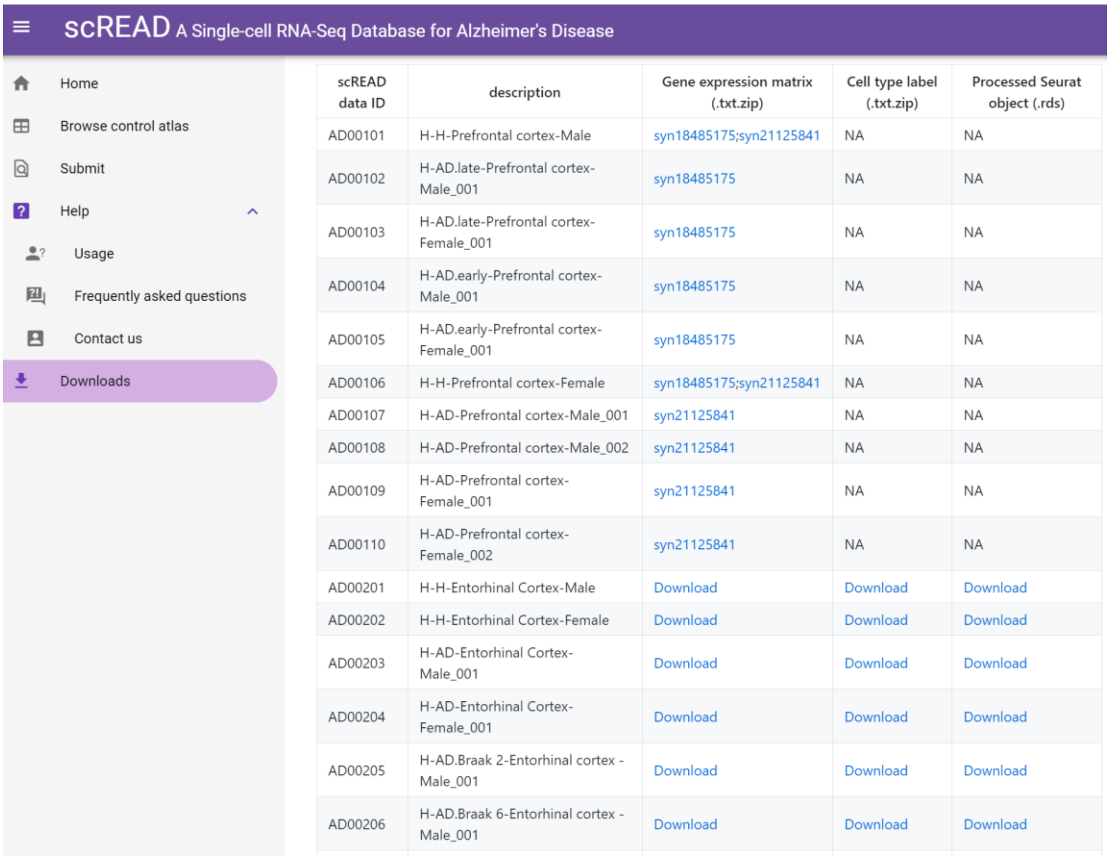
Note: The data you uploaded will be processed by our analyst manually.

Please feel free to contact us at qin.ma@osumc.edu, if you might have any questions regarding your job.

10. You can reset all the input information by clicking this button and restart over again.
11. You can input the job ID here and retrieve the analysis result after the work is done.

Part 5. Download page

Not all the datasets in scREAD are available to download for users. On the “Download” page, datasets that downloaded from the GEO database are available to download, but datasets downloaded from Synapse are not available to download.


scREAD A Single-cell RNA-Seq Database for Alzheimer's Disease
Home
Browse control atlas
Submit
Help
Usage
Frequently asked questions
Contact us
Downloads

scREAD data ID	description	Gene expression matrix (.txt.zip)	Cell type label (.txt.zip)	Processed Seurat object (.rds)
AD00101	H-H-Prefrontal cortex-Male	syn18485175;syn21125841	NA	NA
AD00102	H-AD.late-Prefrontal cortex-Male_001	syn18485175	NA	NA
AD00103	H-AD.late-Prefrontal cortex-Female_001	syn18485175	NA	NA
AD00104	H-AD.early-Prefrontal cortex-Male_001	syn18485175	NA	NA
AD00105	H-AD.early-Prefrontal cortex-Female_001	syn18485175	NA	NA
AD00106	H-H-Prefrontal cortex-Female	syn18485175;syn21125841	NA	NA
AD00107	H-AD-Prefrontal cortex-Male_001	syn21125841	NA	NA
AD00108	H-AD-Prefrontal cortex-Male_002	syn21125841	NA	NA
AD00109	H-AD-Prefrontal cortex-Female_001	syn21125841	NA	NA
AD00110	H-AD-Prefrontal cortex-Female_002	syn21125841	NA	NA
AD00201	H-H-Entorhinal Cortex-Male	Download	Download	Download
AD00202	H-H-Entorhinal Cortex-Female	Download	Download	Download
AD00203	H-AD-Entorhinal Cortex-Male_001	Download	Download	Download
AD00204	H-AD-Entorhinal Cortex-Female_001	Download	Download	Download
AD00205	H-AD.Braak 2-Entorhinal cortex -Male_001	Download	Download	Download
AD00206	H-AD.Braak 6-Entorhinal cortex -Male_001	Download	Download	Download

It provides three files for users to download. 1. The compressed gene expression matrix (.txt.zip); 2. Cell type labels (.txt.zip); 3. Processed Seurat R object (.rds).

Data S2: scREAD workflow tutorials, Related to figure 1.

The data analysis workflow can be downloaded from https://github.com/OSU-BMBL/scread/tree/master/script, the folder contains the following files:

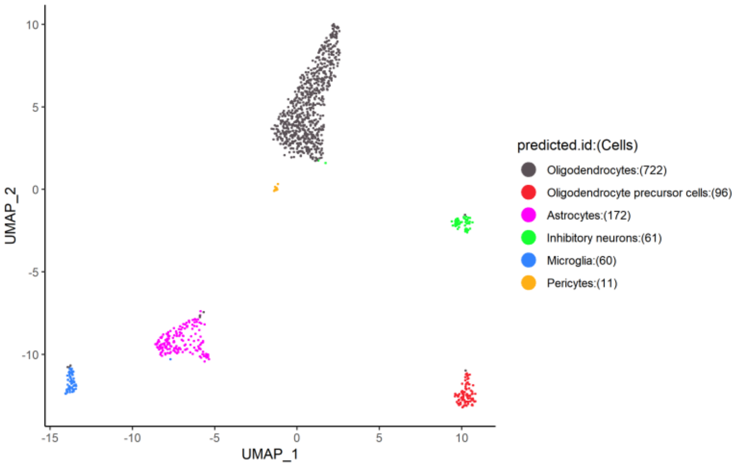
1. custom_marker.csv. A manually created marker gene list file used for identified cell types.
2. functions.R. Visualization functions used in R.
3. build_control_atlas.R: build control cells atlas Seurat object from count matrix file.
4. transfer_cell_type.R: filter out control-like cells in disease dataset
5. run_analysis.R: run analysis workflow, and export tables in scREAD database format.
6. example_control.fst. The example control dataset.
7. example_disease.fst. The example disease dataset.

Build control atlas

1. Goal: Build the control atlas file from the raw gene expression matrix.
2. Prepare your control gene expression data in fst format (https://www.fstpackage.org/), we used fst package to store raw data in scREAD since it provides a fast, easy, and flexible way to serialize data frames. In the data frame, the first column should be gene symbols and other columns as cell labels. Put all code and data in a working directory. (e.g PATH_TO_WD), in this tutorial, we will run example_control.fst.
3. build_control_atlas.R takes three parameters: 1. Working directory path; 2. Control data path. 3. Output data ID

```
4.  cd PATH_TO_WD
5.  Rscript build_control_atlas.R PATH_TO_WD example_control.fst control_example
```

6. The output should contain four files:
   a) control_example.rds. The Seurat R object storing example control data.
   b) control_example_expr.txt. Filtered gene expression matrix.
   c) control_example_cell_label.txt. The first column is the cell name, the second column is the cell type information.
   d) control_example_umap.png. UMAP plot of example control data colored by cell types.

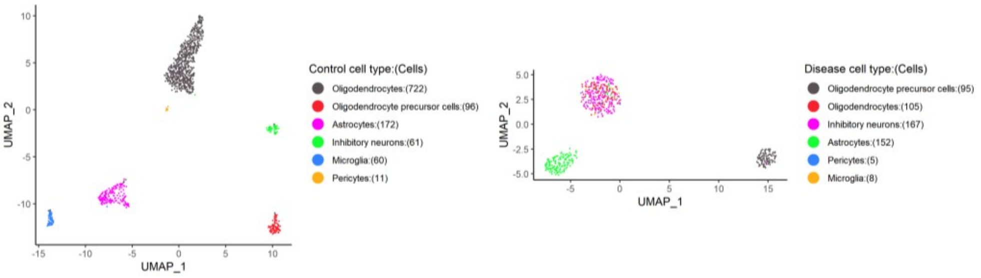


Transfer cell types based on control atlas

1. Goal: Annotate cell type using control atlas as the reference, onto the disease gene expression matrix file.
2. Put all code and data in a working directory. (e.g PATH_TO_WD), after you have generated the control atlas file (control_example.rds).
3. build_control_atlas.R takes four parameters: 1. Working directory path; 2. Control atlas Seurat object file name; 3. Disease gene expression matrix name; 4. Output disease data ID.

7. cd PATH_TO_WD
8. Rscript transfer_cell_type.R PATH_TO_WD control_example.rds example_disease.fst disease_example

9. The output should contain four files:
   e) disease_example.rds. The Seurat R object storing example disease data.
   f) disease_example_expr.txt. Filtered gene expression matrix.
   g) disease_example_cell_label.txt. The first column is the cell name, the second column is the cell type information.
   h) disease_example_umap.png. UMAP plot for both control and disease data colored by cell types.



Run data analysis

1. Goal: Perform analysis between disease and control data
2. Put all code and data in a working directory. (e.g PATH_TO_WD), after you have generated the control atlas file (control_example.rds), and the disease file (disease_example.rds)
3. run_analysis.R takes three parameters: 1. Working directory path; 2. Control Seurat object file name. 3. Disease Seurat object file name.

4. cd PATH_TO_WD
5. Rscript run_analysis.R PATH_TO_WD control_example disease_example

6. The output should be stored in three folders:

a) /de. Differential gene expression analysis results. 1. Cell-type-specific genes; 2. Sub-cluster specific genes; 3. Cell type DE genes between two conditions.

b) /dimension. UMAP coordinates for two datasets.

c) /subcluster_dimension. UMAP coordinates for each sub-clusters in two datasets.

Table S1. The dataset source, Related to Figure 1.

Species	Data_ID	Pubmed_ID
Human	GSE138852	31768052
Human	syn18485175	31042697
Human	GSE147528	https://www.biorxiv.org/content/10.1101/2020.04.04.025825v2
Human	syn21125841	31932797
Human	GSE129308	https://www.biorxiv.org/content/10.1101/2020.05.11.088591v1
Human	GSE146639	https://www.biorxiv.org/content/10.1101/2020.03.18.995332v1
Mouse	GSE98969	28602351
Mouse	GSE103334	29020624
Mouse	GSE130626	31902528
Mouse	GSE141044	31928331
Mouse	GSE140510	31932797
Mouse	GSE140399	31932797
Mouse	GSE143758	32341542
Mouse	GSE147495	32320664
Mouse	GSE150358	32579671
Mouse	GSE142853	32503894
Mouse	GSE142858	32503894

Table S2. The brain regions are covered in scREAD for human and mouse species, Related to Figure 2.

Species	Region	Brodmann area
Human	Entorhinal cortex	NA; NA
Human	Prefrontal cortex	Area 9, Area 46; Area 10
Human	Superior frontal gyrus	Area 8
Human	Superior parietal lobe	NA
Mouse	Cortex	NA
Mouse	Cerebellum	NA
Mouse	Cerebral cortex	NA
Mouse	Hippocampus	NA
Mouse	Prefrontal cortex	NA
Mouse	Subventricular zone	NA

Table S3. The definition of different mouse age stages in scREAD, Related to Figure 2.

Age_Stage	**Range of ages**
2 months	1-2 months
7 months	4-7 months
15 months	10-15 months

Table S4. The marker genes to assign eight major brain cell types, Related to Figure 1.

Cell type	**Genes**
Astrocytes	*GFAP, EAAT1, AQP4, LCN2, GJA1, SLC1A2, FGFR3, NKAIN4*
Endothelial cells	*FLT1, CLDN5, VTN, ITM2A, VWF, FAM167B, BMX, CLEC1B*
Excitatory neurons Pericytes	*SLC17A6, SLC17A7, NRGN, CAMK2A, SATB2, COL5A1, SDK2, NEFM*
Inhibitory neurons	*SLC32A1, GAD1, GAD2, TAC1, PENK, SST, NPY, MYBPC1, PVALB, GABBR2*
Microglia	*IBA-1, P2RY12, CSF1R, CD74, C3, CST3, HEXB, C1QA, CX3CR1, AIF-1*
Oligodendrocytes	*OLIG2, MBP, MOBP, PLP1, MOG, CLDN11, MYRF, GALC, ERMN, MAG*
Oligodendrocyte precursor cells	*VCAN, CSPG4, PDGFRA, SOX10, NEU4, PCDG15, GPR37L1, C1QL1, CDO1, EPN2*
Pericytes	*AMBP, HIGD1B, COX4I2, AOC3, PDE5A, PTH1R, P2RY14, ABCC9, KCNJ8, CD248*

Table S5. The selection of differential gene expression analysis between different conditions (Condition 1 v.s. Condition 2) for diverse cell types in scREAD, Related to Figure 4.

Species	If in the same region	Condition 1	Condition 2
Human	Yes	Disease	Control
Mouse	Yes	Disease	Control
Human	Yes	Disease	Disease
Mouse	Yes	Disease	Disease
Human	No	Disease	Disease
Mouse	No	Disease	Disease

*The comparisons are all in the same gender and age.

Table S6. The computational tools used in scREAD, Related to Figure 1.

Tools	Source code	Version	Language
IRIS3	https://github.com/OSU-BMBL/IRIS3	v1.2.4	R
Seurat	https://github.com/satijalab/seurat	v3.2	R
Harmony	https://github.com/immunogenomics/harmony	v0.1	R/Python
Polychrome	https://github.com/cran/Polychrome	v1.2.5	R
SCINA	https://github.com/jcao89757/SCINA	v1.2.0	R

Table S7. The datasets information of control atlas used in scREAD, Related to Figure 1.

Control atlas	Data_id	Geo/Synapse_id
H-H-Prefrontal cortex-Male	AD00101	syn18485175; syn21125841
H-H-Prefrontal cortex-Female	AD00106	syn18485175; syn21125841
H-H-Entorhinal Cortex-Male	AD00201	GSE138852; GSE147528
H-H-Entorhinal Cortex-Female	AD00202	GSE138852
M-H-Cortex-Male-7m	AD00301	GSE98969; GSE140510
M-H-Cortex-Male-15m	AD00302	GSE140399
M-H-Cerebral cortex-Female-15m	AD00401	GSE147495
M-H-Cerebellum-Male-7m	AD00501	GSE98969
M-H-Prefrontal cortex-Male-7m	AD00601	GSE143758
M-H-Prefrontal cortex-Male-15m	AD00602	GSE143758
M-H-Hippocampus-Male-7m	AD00702	GSE141044
M-H-Hippocampus-Male-15m	AD00703	GSE130626; GSE140399
M-H-Hippocampus-Female-7m	AD00704	GSE141044
M-H-Hippocampus-Female-20m	AD00705	GSE141044
H-H-Superior frontal gyrus-Male	AD00801	GSE147528
M-H-cortex_and_hippocampus-Female-7m_001	AD00901	GSE150358
M-H-cortex_and_hippocampus-Female-7m_002	AD00902	GSE150358
M-H-subventricular_zone_and_hippocampus-Female-7m_001	AD01001	GSE142853
M-H-subventricular_zone_and_hippocampus-Female-7m_002	AD01002	GSE142858
H-H-Prefrontal_cortex-Male_BA9	AD01101	GSE129308
H-H-Prefrontal_cortex-Female_BA9	AD01102	GSE129308
H-H-Superior_parietal_lobe-Male	AD01201	GSE146639
H-H-Superior_parietal_lobe-Female	AD01202	GSE146639

Table S8. The information of disease datasets used in scREAD, Related to Figure 1.

Disease datasets	Data_id	Geo/Synapse_id
H-AD.late-Prefrontal cortex-Male_001	AD00102	syn18485175
H-AD.late-Prefrontal cortex-Female_001	AD00103	syn18485175
H-AD.early-Prefrontal cortex-Male_001	AD00104	syn18485175
H-AD.early-Prefrontal cortex-Female_001	AD00105	syn18485175
H-AD-Prefrontal cortex-Male_001	AD00107	syn21125841
H-AD-Prefrontal cortex-Male_002	AD00108	syn21125841
H-AD-Prefrontal cortex-Female_001	AD00109	syn21125841
H-AD-Prefrontal cortex-Female_002	AD00110	syn21125841
H-AD-Entorhinal Cortex-Male_001	AD00203	GSE138852
H-AD-Entorhinal Cortex-Female_001	AD00204	GSE138852
H-AD.Braak 2-Entorhinal cortex -Male_001	AD00205	GSE147528
H-AD.Braak 6-Entorhinal cortex -Male_001	AD00206	GSE147528
M-AD-Cortex-Male-7m_001	AD00303	GSE98969
M-AD-Cortex-Male-7m_002	AD00304	GSE140510
M-AD-Cortex-Male-7m_003	AD00305	GSE140510
M-AD-Cortex-Male-7m_004	AD00306	GSE140510
M-AD-Cortex-Male-15m_001	AD00307	GSE140399
M-AD-Cortex-Male-15m_002	AD00308	GSE140399
M-AD-Cortex-Male-15m_003	AD00309	GSE140399
M-AD-Cerebral cortex-Female-15m_001	AD00402	GSE147495
M-AD-Cerebral cortex-Female-15m_002	AD00403	GSE147495
M-AD-Cerebral cortex-Male-15m_001	AD00404	GSE147495
M-AD-Cerebral cortex-Male-15m_002	AD00405	GSE147495
M-AD-Cerebellum-Male-7m_001	AD00502	GSE98969
M-AD-Prefrontal cortex-Male-7m_001	AD00603	GSE143758
M-AD-Prefrontal cortex-Male-15m_001	AD00604	GSE143758
M-AD-Hippocampus-Male-7m_001	AD00708	GSE103334
M-AD-Hippocampus-Male-7m_002	AD00709	GSE103334
M-AD-Hippocampus-Male-7m_003	AD00710	GSE141044
M-AD-Hippocampus-Male-15m_001	AD00711	GSE130626
M-AD-Hippocampus-Male-15m_002	AD00712	GSE130626
M-AD-Hippocampus-Male-15m_003	AD00713	GSE130626
M-AD-Hippocampus-Male-15m_006	AD00714	GSE140399
M-AD-Hippocampus-Male-15m_007	AD00715	GSE140399
M-AD-Hippocampus-Male-15m_008	AD00716	GSE140399
M-AD-Hippocampus-Male-20m_002	AD00717	GSE141044
M-AD-Hippocampus-Female-7m_001	AD00718	GSE141044
M-AD-Hippocampus-Female-20m_001	AD00719	GSE141044
H-AD.Braak 2-Superior frontal gyrus-Male_001	AD00802	GSE147528
H-AD.Braak 6-Superior frontal gyrus-Male_001	AD00803	GSE147528

M-AD-cortex_and_hippocampus-Female-7m_001	AD00903	GSE150358
M-AD-cortex_and_hippocampus-Female-7m_002	AD00904	GSE150358
M-AD-subventricular_zone_and_hippocampus-Female-7m_001	AD01003	GSE142853
M-AD-subventricular_zone_and_hippocampus-Female-7m_002	AD01004	GSE142858
H-AD-Prefrontal_cortex_BA9-Male_001	AD01103	GSE129308
H-AD-Prefrontal_cortex_BA9-Female_001	AD01104	GSE129308
H-AD-Superior_parietal_lobe-Male_001	AD01203	GSE146639
H-AD-Superior_parietal_lobe-Female_001	AD01204	GSE146639
H-AD-Superior_parietal_lobe-Male_002	AD01205	GSE146639
H-AD-Superior_parietal_lobe-Female_002	AD01206	GSE146639

Table S9. The definition of AD individuals and AD-like animal models across all datasets used in scREAD, Related to Figure 1.

Data_ID	The pathology of AD	Symptoms
GSE138852	Accumulation of amyloid-beta plaques and tau pathology (Braak stages V and VI)	Dementia
syn18485175	Accumulation of amyloid-beta plaques and tau pathology (Braak stages III-VI)	Mild cognitive impairment and dementia
GSE147528	Accumulation of amyloid-beta plaques and tau pathology (Braak stages II and VI)	Mild cognitive impairment and dementia
syn21125841	Accumulation of amyloid-beta plaques and tau pathology (Braak stages III-V)	Mild cognitive impairment and dementia
GSE129308	Accumulation of tau pathology (Braak stage VI)	Dementia
GSE146639	Accumulation of amyloid-beta plaques in the brain vasculature	Mild cognitive impairment
GSE98969	Parenchymal deposition of amyloid-beta plaques	Severe cognitive dysfunction
GSE103334	Accumulation of amyloid-beta plaques	Cognitive impairment
GSE130626	Severe amyloid-beta pathology	Dementia
GSE141044	Accumulation of amyloid-beta plaques	Cognitive dysfunction
GSE140510	Accumulation of amyloid-beta plaques	Mild cognitive impairment
GSE140399	Accumulation of amyloid-beta plaques	Mild cognitive impairment
GSE143758	Accumulation of amyloid-beta plaques	Cognitive decline
GSE147495	Accumulation of amyloid-beta plaques	Cognitive decline
GSE142853	Accumulation of amyloid-beta plaques	Cognitive decline
GSE142858	Accumulation of amyloid-beta plaques	Cognitive decline
GSE150358	Accumulation of amyloid-beta plaques	Cognitive decline

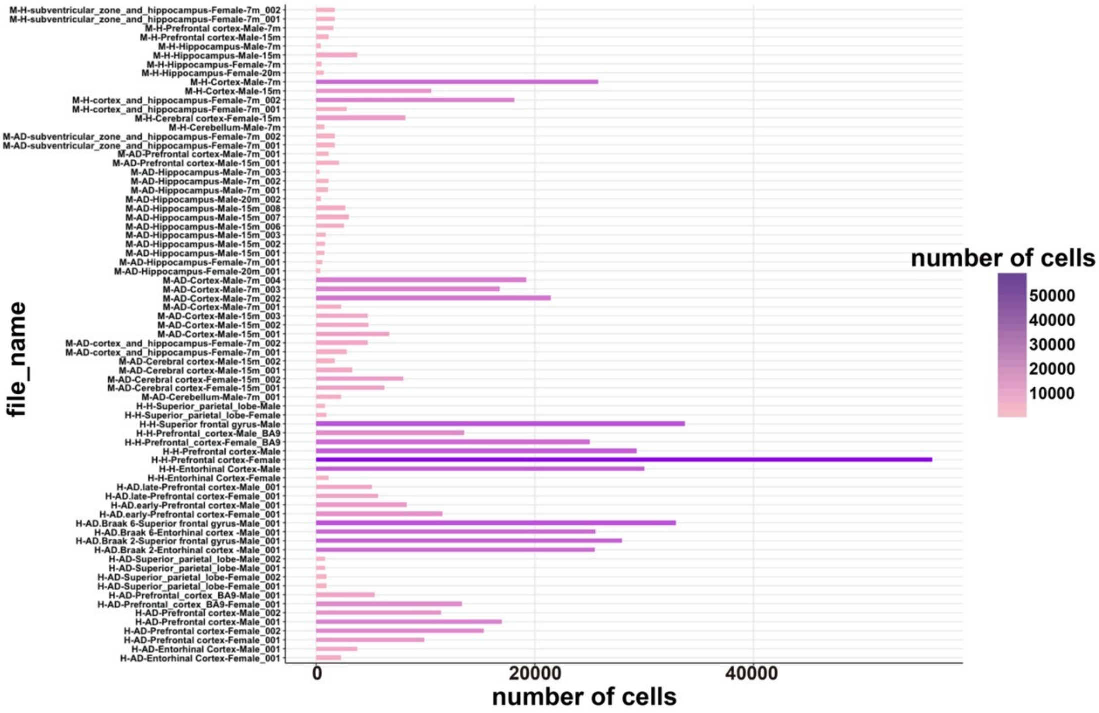


Figure S1. The number of cells in each of the 73 files, Related to Figure 2. The x-axis represents the number of cells of each file, and the y-axis represents the file names of these 73 files. The color intensity of the bar stands for the number of cells, i.e. the darker of the color represents the more cell numbers in the corresponding file.

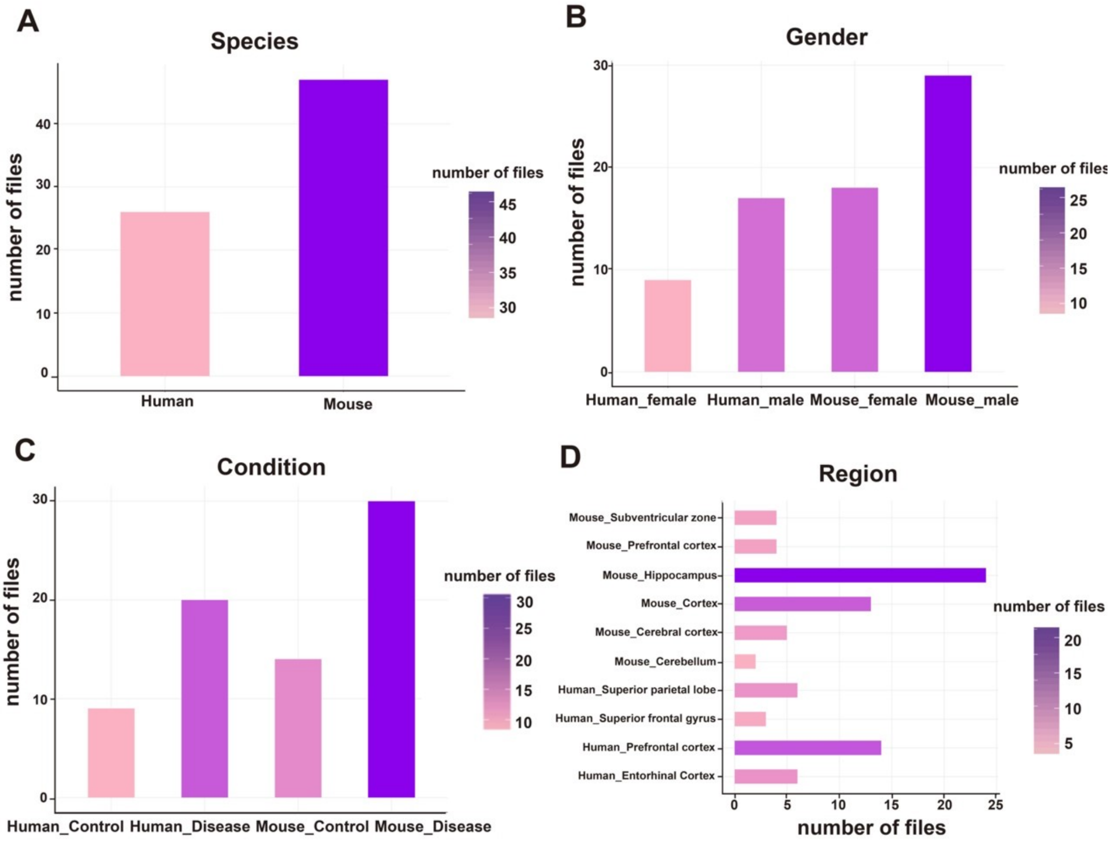


Figure S2. The distribution of the species, gender, condition, and brain region for 73 files, Related to Figure 2. For each panel in this figure, the color of the bar stands for the number of files, the darker the color is the more files in the corresponding factor.

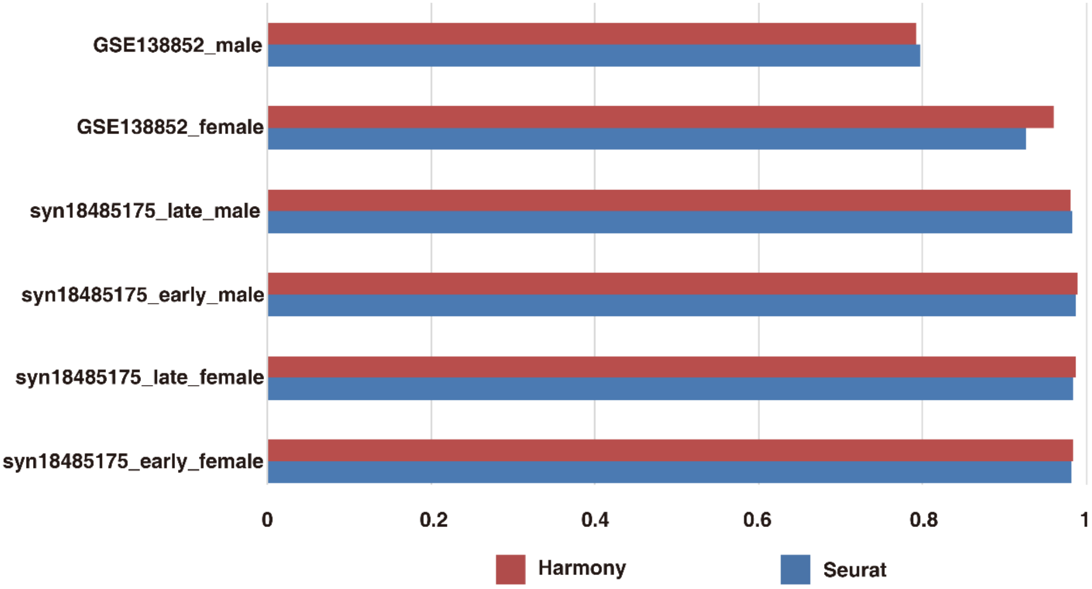


Figure S3. The ARI scores of Harmony and Seurat calculating on six human datasets, Related to Figure 4.

Transparent Methods

Data collection

We manually curated 15 AD related studies, six scRNA-Seq datasets, and 11 snRNA-Seq datasets were retrieved with the following factors well-annotated, i.e., organism, gender, brain region, disease/control, and age information. scREAD redefines the 17 scRNA-Seq & snRNA-Seq datasets into 73 datasets (in total 713,640 cells and nine cell types), each of which corresponds to a specific organism (human or mouse), gender (male or female), brain region (entorhinal cortex, prefrontal cortex, superior frontal gyrus, cortex, cerebellum, cerebral cortex, subventricular zone, superior parietal lobe, or hippocampus), disease or control, and age stage (seven months, 15 months, or 20 months for mice, and 50-100+ years old for human). These datasets have been published and freely accessible in the public domain as of September 22nd, 2020 (Barrett et al., 2013).

Construction of human and mouse control atlas

Human and Mouse control atlases come from the 15 scRNA-Seq & snRNA-Seq studies. Genes detected in less than 3 cells and cells detected in less than 200 genes were filtered out. Principal component analysis (PCA) was performed to obtain a small number of principal components, 25 PCA components were used as input of Uniform Manifold Approximation and Projection (UMAP) (Becht et al., 2018). Initial clustering was performed using Seurat's v3.1.5 SNN graph clustering using the *FindClusters* function with a resolution of 0.8 (Stuart et al., 2019). Seurat is a widely used R toolkit to identify and interpret sources of heterogeneity from single-cell transcriptomic measurements, and to integrate diverse types of single-cell data (Zhang et al., 2019).

SCINA is an R package that leverages prior marker genes information and simultaneously performs cell type clustering and assignment for known cell types (Zhang et al., 2019). Furthermore, SCINA shows good performances among prior-knowledge classifiers when high-quality marker genes are provided (Abdelaal et al., 2019). Each cell was assigned a cell type based on a manually created marker gene list file (Table S4) using SCINA v1.2.0, whereas the cells with unknown labels marked by SCINA were first compared with predicted clusters from Seurat, and then the unknown labels were assigned to the most dominate cell types within the predicted clusters (Zhang et al., 2019).

Evaluation indexes of identified cell types

If benchmark labels are provided from the original study, the identified cell labels will be evaluated by the Adjusted Rand Index (ARI) (Steinley et al., 2016). To calculate ARI, a contingency table is built to summarize the overlaps between the two cell label lists with n elements (cells). Each entry denotes the number of objects in common between the two label lists. The ARI score can be calculated as:

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

where n_{ij} are values from the contingency table, a_i is the sum of the ith row of the contingency table, b_j is the sum of the jth column of the contingency table.

If benchmark labels are not provided from the original study, the predicted cell types will be evaluated by calculating the silhouette score that measures how similar a cell is to its type compared to other clusters (Lovmar et al., 2005). The silhouette ranges from −1 to +1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. The silhouette score can be calculated by:

$$s(i) = \frac{b(i) - a(i)}{max\,\{a(i), b(i)\}} = \begin{cases} 1 - \frac{a(i)}{b(i)}, & if\ a(i) < b(i) \\ 0, & if\ a(i) = b(i) \\ \frac{b(i)}{a(i)} - 1, & if\ a(i) > b(i) \end{cases}$$

where $a(i)$ be the average distance between a sample i and all the rest samples in the same cluster, and $b(i)$ be the smallest average distance of i to all samples.

Identification of human and mouse disease cell types based on the control atlas

Not all cells collected from patient samples are malignant, and there are heterogeneous cells within individual patients, that is, normal healthy cells are included. In Granja *et al.*'s research (Granja et al., 2019), they defined these healthy cells as control-like cells. These control-like cells maintain distinct regulatory mechanisms and gene expression patterns compared to disease cells and will disturb the accurate identification of cancer cell clusters. Thus, the removal of control-like cells from disease data is critical to identify real disease-associated cells. Granja *et al.* used this strategy to remove control-like cells and then identify cancer cells, and we used this strategy in scREAD to identify AD-associated cells. For each of the AD datasets in scREAD, the ratio of the control-like cells out of all the cells in this dataset is about 10%. We tested at Mathys *et al.*'s dataset (Mathys et al., 2019), and found out the ARI scores between with control-like cells and without control-like cells has no significant difference. However, the ARI score of without control-like cells datasets is higher than with control-like cells datasets.

To determine whether cells from disease datasets are control-like, Harmony R package (v1.0) was first used to integrate the disease dataset with its corresponding control atlas. Harmony shows similar ARI scores (Supplementary Figure S3), but it has a significantly shorter run-time compared to other data integration tools (Tran et al., 2020). After the integration, cells were clustered using Seurat's *FindClusters* function with a resolution of 4. A hypergeometric test was performed for each cluster using the number of cells from disease cells and the number of cells from the control atlas. Clusters were considered to be control-like if the hypergeometric test result was significant (p-value < 0.0001, Benjamini-Hochberg adjusted), and the cells from the disease dataset in control-like clusters were removed from the downstream analyses.

For the remaining disease cells, Seurat's *FindTransferAnchors* function was used to find transfer anchors using PCA to project the control-atlas onto the disease dataset. Cell types were transferred using the *TransferData* function with PCA embeddings as the weighting anchors. The subclusters for each cell type were designated using Seurat's *FindClusters* function for all cells in each identified cell type with a resolution of 0.8.

Differential expression and gene set enrichment analysis

MAST is an R package that uses a hurdle model to single-cell RNA-seq data (Finak et al., 2015) and was recommended for single-cell differential expression (DE) testing (Luecken and Theis, 2019; Soneson and Robinson, 2018). Seurat's *FindAllMarkers* and *FindMarkers* functions that utilizes the MAST package were used to run DE testing on normalized gene expression data. Cell-type-specific genes were identified by performing DE testing between the cell type of interest and the average of the remaining cell types. Subcluster-specific genes were identified by performing DE testing between the subcluster of interest and the average of the remaining subclusters from the same cell type. For each cell type, several DE comparisons were performed within two different datasets, categorized from AD versus control, and AD versus AD in the same species under the same gender, brain region, and age. To regress out technical biases from different datasets, the dataset latent variables were added in all cross-dataset DE testing. All of the above-mentioned DE results can be sent to the Enrichr web server in real-time compared to different functional annotation databases to identify the enriched KEGG pathways, Gene Ontology (GO), etc.

Identification of CTSRs

The CTSRs analysis is performed using IRIS3 (Integrative Cell-type-specific Regulon Inference Server from Single-cell RNA-Seq), a highly effective and easy-to-use web server for biologically meaningful CTSR inference from human or mouse scRNA-Seq data (Ma et al., 2020). An empirical p-value of a regulon's RSS can be estimated by comparing it with the RSSs of randomly selected gene sets (having the same number of genes in this regulon through a bootstrap method) in the same cell type for 10,000 times. Regulon p-values will be Bonferroni-adjusted by multiplying the number of all the identified regulons in the exact cell type. Regulons with adjusted p-values less than 0.05 were considered as cell type-specific regulons.

Implementation

scREAD consolidates a variety of web frameworks to provide user-friendly interactive visualizations. The front end was built on top of Nuxt.js (https://nuxtjs.org/) and utilized libraries such as Vuetify (https://vuetifyjs.com/en/) and Plotly.js (https://plotly.com/). Koa.js (https://koajs.com/) serves as the REST API back-end server for data query and custom job submission. All data are stored and managed using a MySQL database. The entire web application is managed by PM2 (https://pm2.keymetrics.io/) and deploys on a Red Hat Enterprise seven Linux system with 28-core Intel Xeon E5–2650 CPU and 64GB RAM. All integrated tools are listed in Table S6.

The browsers that scREAD supported are Google Chrome, Safari, and Firefox. The scREAD is not supported by the Internet Explorer browser.

Supplemental References

Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J.T., and Mahfouz, A. (2019). A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome biology *20*, 194.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic acids research *41*, D991-995.

Becht, E., McInnes, L., Healy, J., Dutertre, C.A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2018). Dimensionality reduction for visualizing single-cell data using UMAP. Nature biotechnology.

Finak, G., McDavid, A., Yajima, M., Deng, J., Gersuk, V., Shalek, A.K., Slichter, C.K., Miller, H.W., McElrath, M.J., Prlic, M.*, et al.* (2015). MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome biology *16*, 278.

Granja, J.M., Klemm, S., McGinnis, L.M., Kathiria, A.S., Mezger, A., Corces, M.R., Parks, B., Gars, E., Liedtke, M., Zheng, G.X.Y.*, et al.* (2019). Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. Nature biotechnology *37*, 1458-1465.

Lovmar, L., Ahlford, A., Jonsson, M., and Syvanen, A.C. (2005). Silhouette scores for assessment of SNP genotype clusters. BMC genomics *6*, 35.

Luecken, M.D., and Theis, F.J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. Molecular systems biology *15*, e8746.

Ma, A., Wang, C., Chang, Y., Brennan, F.H., McDermaid, A., Liu, B., Zhang, C., Popovich, P.G., and Ma, Q. (2020). IRIS3: integrated cell-type-specific regulon inference server from single-cell RNA-Seq. Nucleic acids research *48*, W275-W286.

Soneson, C., and Robinson, M.D. (2018). Bias, robustness and scalability in single-cell differential expression analysis. Nature methods *15*, 255-261.

Steinley, D., Brusco, M.J., and Hubert, L. (2016). The variance of the adjusted Rand index. Psychological methods *21*, 261-272.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell *177*, 1888-1902 e1821.

Tran, H.T.N., Ang, K.S., Chevrier, M., Zhang, X., Lee, N.Y.S., Goh, M., and Chen, J. (2020). A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome biology *21*, 12.

Zhang, Z., Luo, D., Zhong, X., Choi, J.H., Ma, Y., Wang, S., Mahrt, E., Guo, W., Stawiski, E.W., Modrusan, Z.*, et al.* (2019). SCINA: A Semi-Supervised Subtyping Algorithm of Single Cells and Bulk Samples. Genes *10*.