

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.

Data S2: scREAD workflow tutorials.

Table S1: The dataset source.

Table S2: The brain regions are covered in scREAD for human and mouse species.

Table S3: The definition of different mouse age stages in scREAD.

Table S4: The marker genes to assign eight major cell types in the brain.

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

Table S6: The computational tools used in scREAD.

Table S7: The datasets information on control atlas used in scREAD.

Table S8: The information of disease datasets used in scREAD.

Table S9. The definition of AD individuals and AD-like animal models across all datasets used in scREAD.

Figure S1: The number of cells in each of the 73 files.

Figure S2: The distribution of the species, gender, condition, and brain region for 73 files.

Figure S3: The ARI scores of Harmony and Seurat calculating on six human datasets.

Transparent Methods

Supplemental References

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

The screenshot shows the scREAD server interface. At the top, there are four pie charts representing the distribution of 73 datasets across four factors: Species (Human/Mouse), Condition (Disease/Control), Region (Prefrontal cortex, Entorhinal cortex, Cerebellum), and Gender (Male/Female). Below the charts is a search bar for differentially expressed genes. A navigation menu on the left includes Home, Browse control atlas, Submit, Help, and Downloads. The main content area features a filter bar with dropdown menus for Species (Human/Mouse), Condition (Disease/Control), Region (Prefrontal cortex/Entorhinal cortex/Cerebellum), and Gender (All/Male/Female). Below the filters is a table of datasets with columns for scREAD data ID, Overview, Gender, Condition, Region, Brain stage, Age, Misc model, GEO/ncbi ID, and #Cells. A 'DOWNLOAD LIST TABLE' button is visible. The table shows 10 rows of data, with the first row highlighted. A footer note states: 'scREAD is developed by SMEL 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.

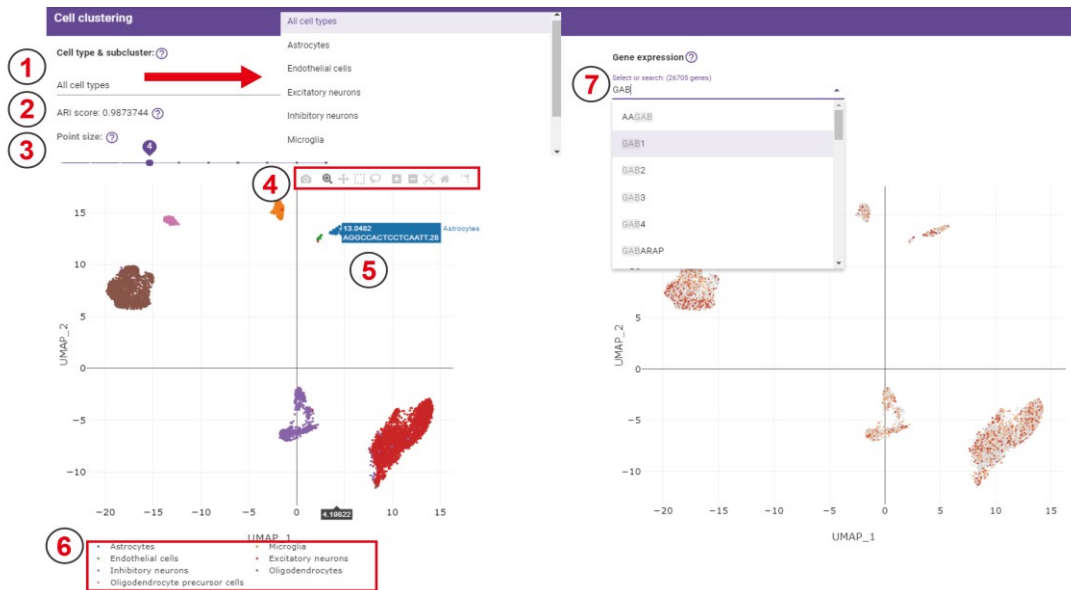
2.1 General information

The screenshot displays the 'General information' page for dataset AD00103. It is organized into several sections, each marked with a red circle and a number:

- 1 Overview:** scREAD Data ID: AD00103, Species: Human, Region: Prefrontal cortex, Condition: Disease, Braak stage: 5 - 6, Gender: Female, Age: 75-90+ years.
- 2 Dataset information:** Number of AD associated cells: 5679, Number of control-like cells: 950, Number of identified cell types: 7.
- 3 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.
- 4 Datasets from the same experiment:** AD00101 | AD00102 | AD00103 | AD00104.
- 5 Overview (floating dialog):** 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. Includes 'DETAILS' and 'CANCEL' buttons.

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

The screenshot shows the 'Differential expression (DE) / Gene set enrichment' tool interface. It includes a sidebar with navigation options (1-14), a main control panel with filters for group, cell type of interest, cutoffs, and direction, a search bar (6), a 'DOWNLOAD' button (7), a gene list table (8), a 'Rows per page' dropdown (9), and a bottom sidebar with additional filters (10-13).

| 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 | 2.95e-225 |
| AQP4 | 2.10245 | 0.712 | 0.035 | 1.05e-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 |

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

The screenshot shows the 'DE testing groups' selection interface. It includes three columns of options: 'Cell type specific genes', 'Disease vs disease (same region)', and 'Subcluster specific genes'. The 'Cell type specific genes' column is selected, and the 'Oligodendrocytes' cell type is chosen under 'Cell type of interest'.

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.

- 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.

| Index | Name | Enrichment score | Adjusted p-value | Download |
|-------|--------------------------|------------------|------------------|----------|
| 1 | Cholesterol biosynthesis | 2.5817e-05 | 7.81 | Download |
| 2 | Cholesterol homeostasis | 6.191e-05 | 7.81 | Download |
| 3 | Cholesterol metabolism | 0.000102 | 7.81 | Download |
| 4 | Cholesterol transport | 0.000102 | 7.81 | Download |
| 5 | Cholesterol synthesis | 0.000102 | 7.81 | Download |
| 6 | Cholesterol transport | 0.000102 | 7.81 | Download |
| 7 | Cholesterol synthesis | 0.000102 | 7.81 | Download |
| 8 | Cholesterol transport | 0.000102 | 7.81 | Download |
| 9 | Cholesterol synthesis | 0.000102 | 7.81 | Download |
| 10 | Cholesterol transport | 0.000102 | 7.81 | Download |

- GO biological process analysis result of the DEGs. Please see entry 10 to know how to use it.
- GO molecular function analysis result of the DEGs. Please see entry 10 to know how to use it.
- GO cellular component analysis result of the DEGs. Please see entry 10 to know how to use it.
- 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.

| Index | Transcription factor | Regulon specificity score | Regulon p-value | Genes |
|--------|----------------------|---------------------------|-----------------|---|
| CT1-R1 | CPEB1 | 0.18235 | 1 | APLP1,ATP8A1,FRMD5,MAP7,PEX5L,SLC24A2,SPOCK1,TRIM2,TLL7,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,PKI,SPOCK1,TLL7 |
| CT2-R3 | SP3 | 0.169818 | 1 | AGAP1,ARHGAP21,CDC42BPA,DLG1,ENOX1,FRMD5,GPM6B,MAP7,NPAS3,PDE4DIP,PPP2R2B,PTMA,PKI,SPOCK1,TRIM2,TLL7 |
| CT2-R4 | FOXO1 | 0.169221 | 1 | APLP1,ATP8A1,FRMD5,MAP7,PEX5L,RTN4,SPOCK1,TRIM2 |
| CT2-R5 | SP1 | 0.169171 | 1 | AK5,APLP1,ATP8A1,FRMD5,GPM6B,MAP7,PTMA,SIK3,SPOCK1,TRIM2,TLL7 |
| CT2-R6 | CPEB1 | 0.169018 | 1 | AK5,APLP1,ATP8A1,FRMD5,GPM6B,MAP7,NCAM1,PEBP1,PEX5L,RTN4,SIK3,SLC24A2,SPOCK1,TRIM2,TLL7,ZEB2 |
| CT2-R7 | RREB1 | 0.168687 | 1 | AK5,APLP1,FRMD5,GPM6B,MAP7,PEBP1,PTMA,SPOCK1,TLL7 |
| CT3-R1 | MXI1 | 0.77387 | 0 | AFF3,AGBL4,CACNA1B,CACNA2D3,CAMK1D,CAMK2A,CAMK2B,CDH18,CELF4,CHRM3,DNM1,EFNA5,FRMPD4,GABBR2,GABBR3,HECW1,LRFN5,NELL2,NRGN,DLFM1,RALY1 |
| CT3-R2 | ZNF263 | 0.766975 | 0 | AFF3,AGBL4,BASP1,CACNA1B,CACNA2D3,CAMK1D,CAMK2A,CAMK2B,CDH18,CELF4,CHRM3,DCLK1,DNM1,EFNA5,FRMPD4,GABBR2,GABBR3,GRIN1,GRM5,HECW1,KHDR |

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 | |
|---|---|
| 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-Cerebellum-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 | ▼ |

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.

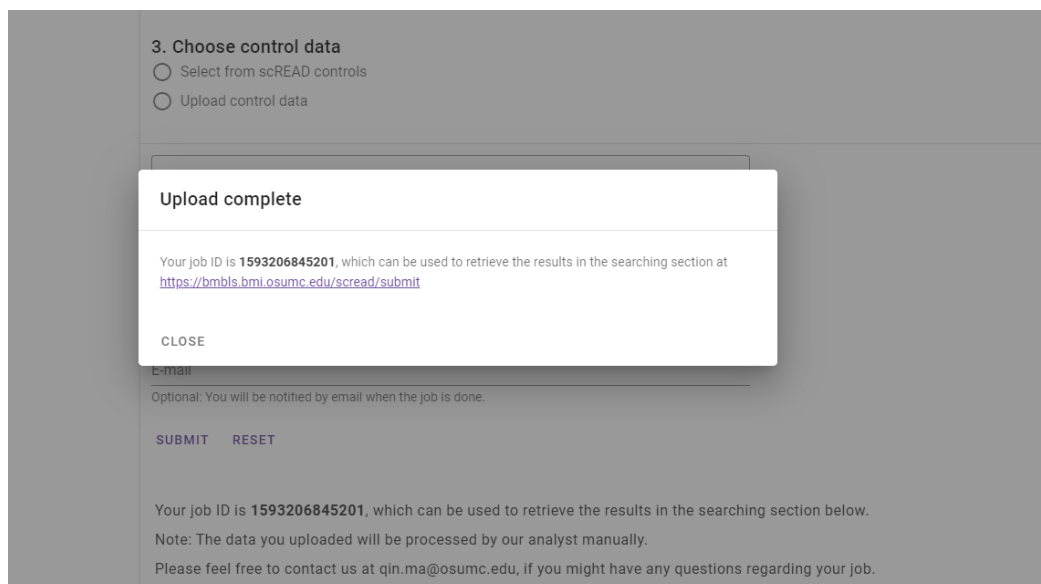
Job Submission

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 notified by email when the job is done. For more information visit our [FAQ](#).

1. Upload disease gene expression file

U TIES (U B (N 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
5. Select control dataset from scREAD: ➔ 4. _____
 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
6. Comments: _____
7. Allow permanent storage in scREAD. @
8. E-mail: _____
Optional: You will be notified by email when the job is done.
9. **SUBMIT** **RESET** 10. **Retrieve your job**
11. 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.



The screenshot shows a web interface for submitting data to scREAD. At the top, there is a section titled "3. Choose control data" with two radio button options: "Select from scREAD controls" and "Upload control data". Below this is a large empty rectangular box. A white modal window is overlaid on the interface, titled "Upload complete". The modal contains the text: "Your job ID is 1593206845201, which can be used to retrieve the results in the searching section at <https://bmbpls.bmi.osumc.edu/scread/submit>". Below the text is a "CLOSE" button. Underneath the modal, there is an "E-mail" section with the text "Optional: You will be notified by email when the job is done." and two buttons: "SUBMIT" and "RESET". At the bottom of the interface, there is a footer area with the text: "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 | | | | | |
|---|----------------|---|---|----------------------------|--------------------------------|
| | scREAD data ID | description | Gene expression matrix (.txt.zip) | Cell type label (.txt.zip) | Processed Seurat object (.rds) |
| Home | AD00101 | H-H-Prefrontal cortex-Male | syn18485175;syn21125841 | NA | NA |
| Browse control atlas | AD00102 | H-AD.late-Prefrontal cortex-Male_001 | syn18485175 | NA | NA |
| Submit | AD00103 | H-AD.late-Prefrontal cortex-Female_001 | syn18485175 | NA | NA |
| Help | AD00104 | H-AD.early-Prefrontal cortex-Male_001 | syn18485175 | NA | NA |
| Usage | AD00105 | H-AD.early-Prefrontal cortex-Female_001 | syn18485175 | NA | NA |
| Frequently asked questions | AD00106 | H-H-Prefrontal cortex-Female | syn18485175;syn21125841 | NA | NA |
| Contact us | AD00107 | H-AD-Prefrontal cortex-Male_001 | syn21125841 | NA | NA |
| Downloads | 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.

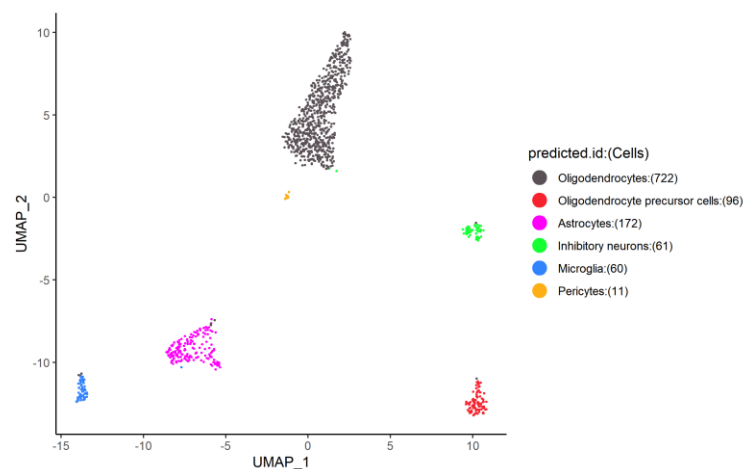
The data analysis workflow can be downloaded from

<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

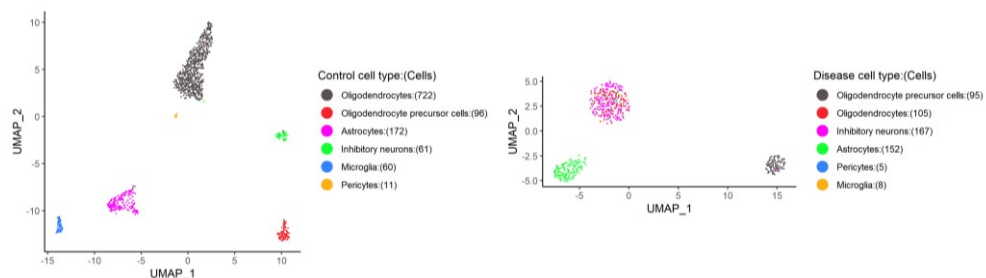
1. Goal: Annotate cell type using control atlas as the reference, onto the disease gene expression matrix file.
2. Put all code and data in a working directory. (e.g PATH_TO_WD), after you have generated the control atlas file (control_example.rds).
3. build_control_atlas.R takes four parameters: 1. 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 | <i>GFAP, EAAT1, AQP4, LCN2, GJA1, SLC1A2, FGFR3, NKAIN4</i> |
| Endothelial cells | <i>FLT1, CLDN5, VTN, ITM2A, VWF, FAM167B, BMX, CLEC1B</i> |
| Excitatory neurons | <i>SLC17A6, SLC17A7, NRG1, CAMK2A, SATB2, COL5A1,</i> |
| Pericytes | <i>SDK2, NEFM</i> |
| Inhibitory neurons | <i>SLC32A1, GAD1, GAD2, TAC1, PENK, SST, NPY, MYBPC1,</i> |
| | <i>PVALB, GABBR2</i> |
| Microglia | <i>IBA-1, P2RY12, CSF1R, CD74, C3, CST3, HEXB, C1QA,</i> |
| | <i>CX3CR1, AIF-1</i> |
| Oligodendrocytes | <i>OLIG2, MBP, MOBP, PLP1, MOG, CLDN11, MYRF, GALC,</i> |
| | <i>ERMN, MAG</i> |
| Oligodendrocyte precursor cells | <i>VCAN, CSPG4, PDGFRA, SOX10, NEU4, PCDG15, GPR37L1,</i> |
| | <i>C1QL1, CDO1, EPN2</i> |
| Pericytes | <i>AMBP, HIGD1B, COX4I2, AOC3, PDE5A, PTH1R, P2RY14,</i> |
| | <i>ABCC9, KCNJ8, CD248</i> |

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 impairment and dementia | cognitive and |
| GSE147528 | Accumulation of amyloid-beta plaques and tau pathology (Braak stages II and VI) | Mild impairment and dementia | cognitive and |
| syn21125841 | Accumulation of amyloid-beta plaques and tau pathology (Braak stages III-V) | Mild impairment and dementia | cognitive and |
| GSE129308 | Accumulation of tau pathology (Braak stage VI) | Dementia | |
| GSE146639 | Accumulation of amyloid-beta plaques in the brain vasculature | Mild impairment | cognitive |
| GSE98969 | Parenchymal deposition of amyloid-beta plaques | Severe dysfunction | cognitive |
| 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 impairment | cognitive |
| GSE140399 | Accumulation of amyloid-beta plaques | Mild impairment | cognitive |
| 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 | |

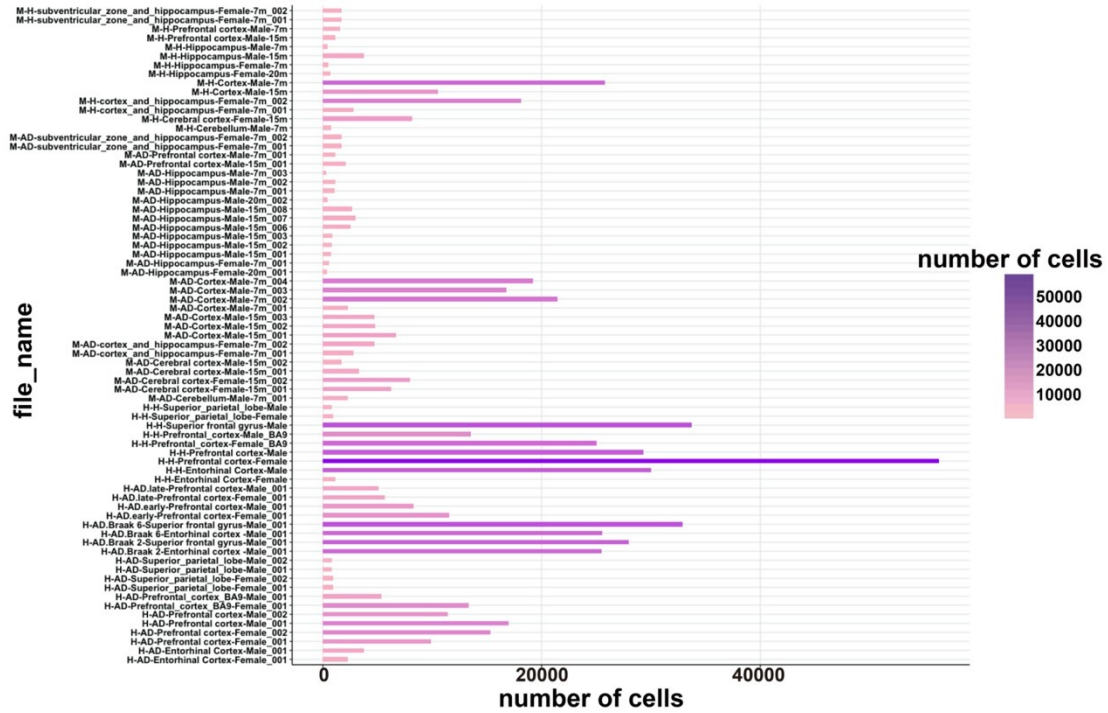


Figure S1. The number of cells in each of the 73 files, Related to Figure 2. The x-axis represents the number of cells of each file, and the y-axis represents the file names of these 73 files. The color intensity of the bar stands for the number of cells, i.e. the darker of the color represents the more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

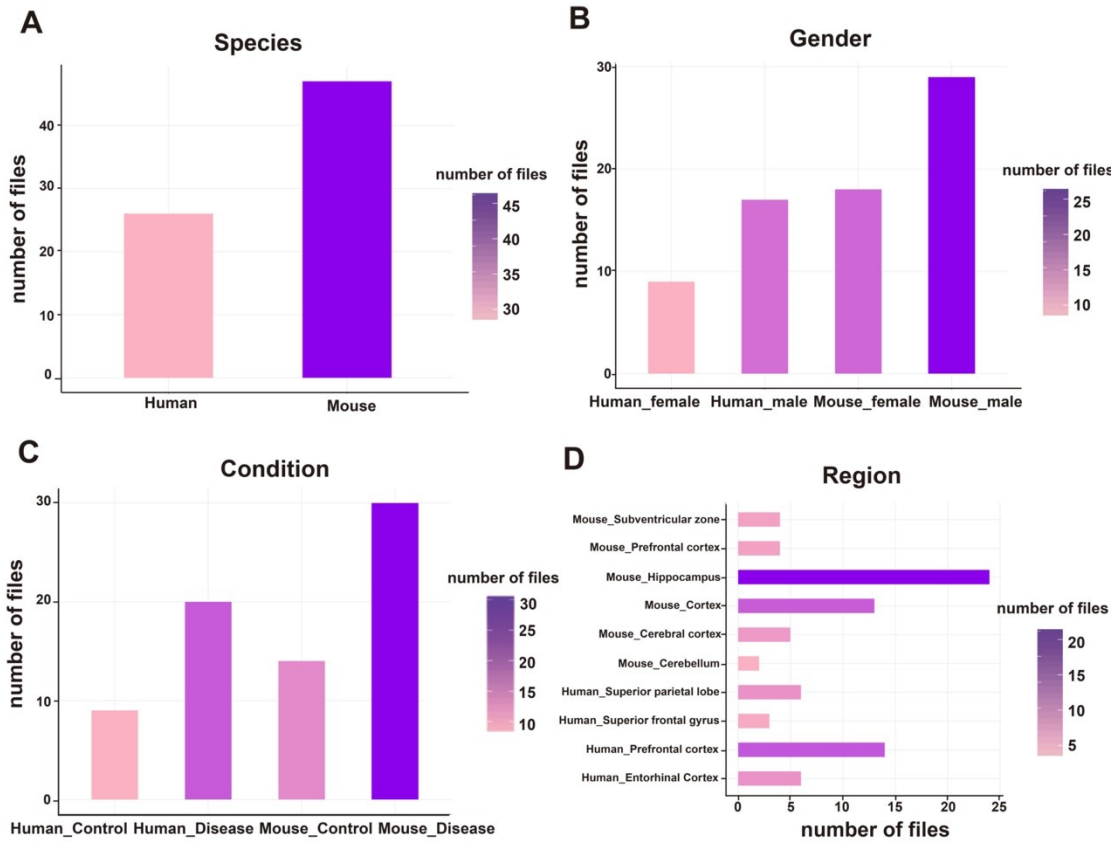


Figure S2. The distribution of the species, gender, condition, and brain region for 73 files, Related to Figure 2. For each panel in this figure, the color of the bar stands for the number of files, the darker the color is the more files in the corresponding factor.

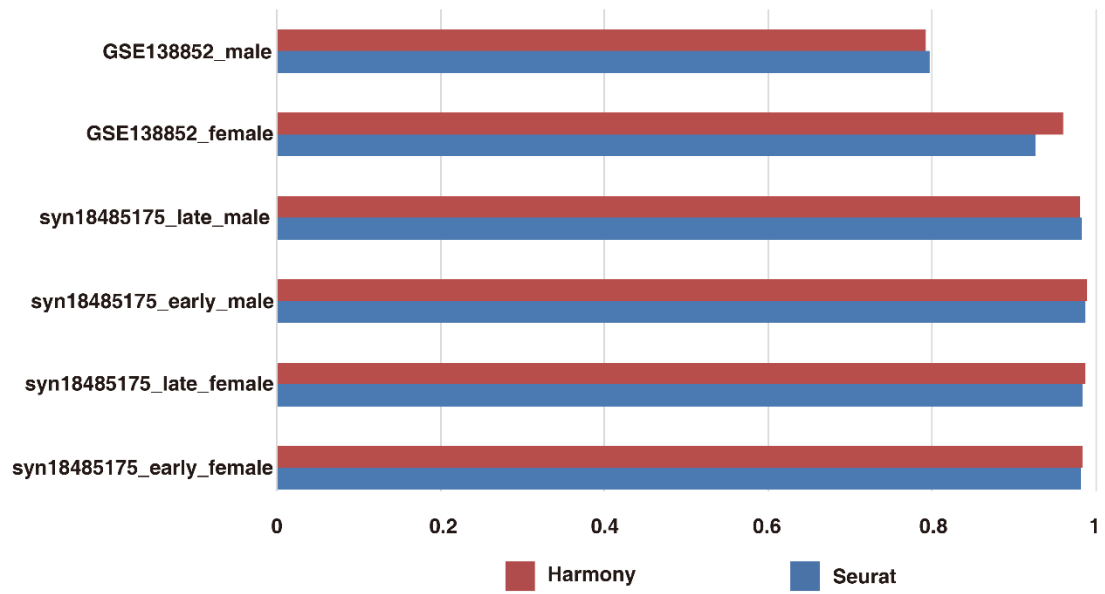


Figure S3. The ARI scores of Harmony and Seurat calculating on six human datasets, Related to Figure 4.

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:

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

$$ARI = \frac{\frac{1}{2} [\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}] - \frac{[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}]}{\binom{n}{2}}}{\binom{n}{2}}$$

where n_{ij} are values from the contingency table, a_i is the sum of the i th row of the contingency table, b_j is the sum of the j th 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:

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

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; Sonesson 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*.