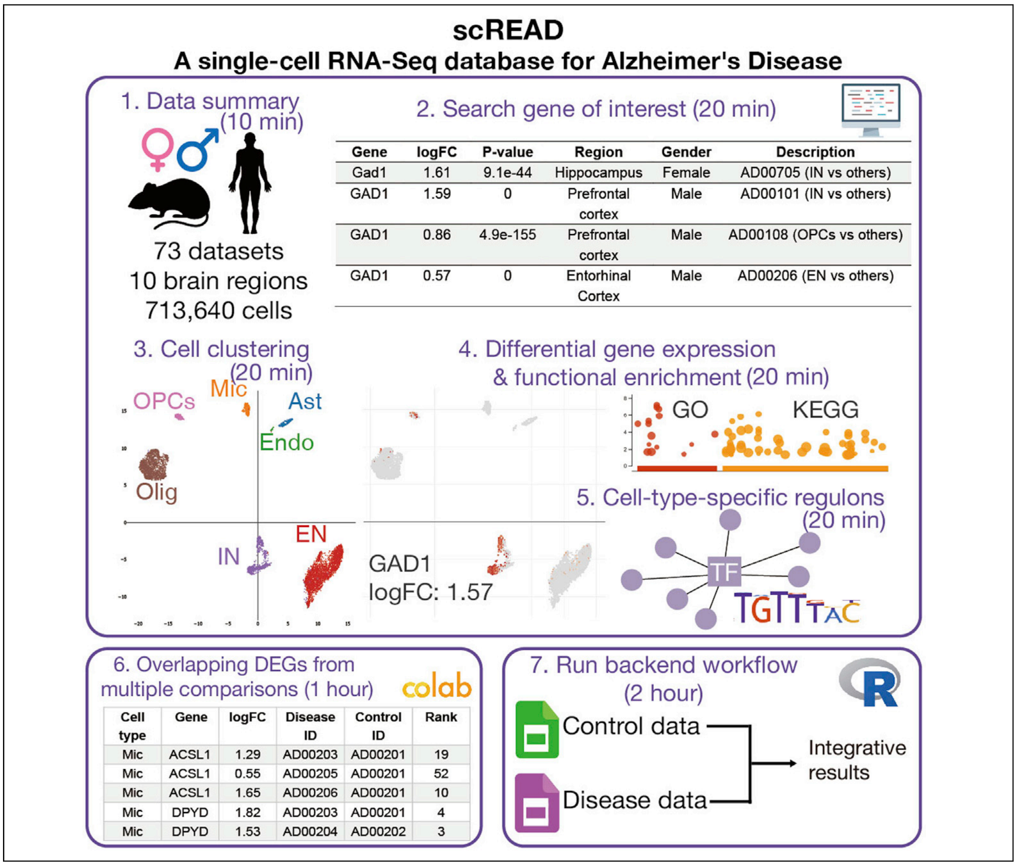
STAR Protocols

CellPress
OPEN ACCESS

Protocol

Use of scREAD to explore and analyze single-cell and single-nucleus RNA-seq data for Alzheimer's disease


scREAD
A single-cell RNA-Seq database for Alzheimer's Disease
1. Data summary (10 min)
73 datasets
10 brain regions
713,640 cells
2. Search gene of interest (20 min)
3. Cell clustering (20 min)
OPCs
Mic
Ast
Endo
Olig
IN
EN
-20
-15
-10
-5
0
5
10
15
4. Differential gene expression & functional enrichment (20 min)
GAD1
logFC: 1.57
GO
KEGG
5. Cell-type-specific regulons (20 min)
TF
TGTTTAC
6. Overlapping DEGs from multiple comparisons (1 hour)
colab
7. Run backend workflow (2 hour)
Control data
Disease data
Integrative results
R

Gene	logFC	P-value	Region	Gender	Description
Gad1	1.61	9.1e-44	Hippocampus	Female	AD00705 (IN vs others)
GAD1	1.59	0	Prefrontal cortex	Male	AD00101 (IN vs others)
GAD1	0.86	4.9e-155	Prefrontal cortex	Male	AD00108 (OPCs vs others)
GAD1	0.57	0	Entorhinal Cortex	Male	AD00206 (EN vs others)

Cell type	Gene	logFC	Disease ID	Control ID	Rank
Mic	ACSL1	1.29	AD00203	AD00201	19
Mic	ACSL1	0.55	AD00205	AD00201	52
Mic	ACSL1	1.65	AD00206	AD00201	10
Mic	DPYD	1.82	AD00203	AD00201	4
Mic	DPYD	1.53	AD00204	AD00202	3

Single-cell RNA-sequencing (scRNA-seq) and single-nucleus RNA-sequencing (snRNA-seq) studies have provided remarkable insights into understanding the molecular pathogenesis of Alzheimer's disease. We recently developed scREAD, a database to provide comprehensive analyses of all the existing AD scRNA-seq and snRNA-seq data from the public domain. Here, we report protocols for using the scREAD web interface and running the backend workflow locally. Our protocols enable custom analyses of AD single-cell and single-nucleus gene expression profiles.

Cankun Wang, Yujia Xiang, Hongjun Fu, Qin Ma

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

Highlights

scREAD protocol for Alzheimer's disease sc/snRNA-seq analysis and interpretation

Step-by-step tutorial to run the scREAD workflow on users' local computers

Comprehensive demonstrations of cell type annotation and other downstream analysis

Wang et al., STAR Protocols 2, 100513
June 18, 2021 © 2021 The Author(s).
https://doi.org/10.1016/j.xpro.2021.100513

Protocol

Use of scREAD to explore and analyze single-cell and single-nucleus RNA-seq data for Alzheimer's disease

Cankun Wang,[1,3] Yujia Xiang,[1] Hongjun Fu,[2,*] and Qin Ma[1,4,*]

[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]Technical contact

[4]Lead contact

*Correspondence: hongjun.fu@osumc.edu (H.F.), qin.ma@osumc.edu (Q.M.)
https://doi.org/10.1016/j.xpro.2021.100513

SUMMARY

Single-cell RNA-sequencing (scRNA-seq) and single-nucleus RNA-sequencing (snRNA-seq) studies have provided remarkable insights into understanding the molecular pathogenesis of Alzheimer's disease. We recently developed scREAD, a database to provide comprehensive analyses of all the existing AD scRNA-seq and snRNA-seq data from the public domain. Here, we report protocols for using the scREAD web interface and running the backend workflow locally. Our protocols enable custom analyses of AD single-cell and single-nucleus gene expression profiles. For complete details on the use and execution of this protocol, please refer to Jiang et al. (2020).

BEFORE YOU BEGIN

Preparing to use the scREAD website

⏱ Timing: 5 min

1. We recommend using Chrome, Safari, Microsoft Edge, or Firefox web browser to access scREAD (https://bmbls.bmi.osumc.edu/scread/). Microsoft Internet Explorer is not supported.
2. If you would like to submit data to scREAD, prepare raw scRNA-seq or snRNA-seq gene expression data in text format, in which rows represent genes, and columns represent cells.

Preparing local environment to run the scREAD workflow

⏱ Timing: 30 min

3. Install R, version 3.6 or greater. The required R dependencies are listed in the key resources table section below.

4. Raw scRNA-seq or snRNA-seq expression data mainly has three formats:
    a. A single *.txt*, *.tsv* or *.csv* formatted gene expression matrix, in which each row represents a feature (gene), and each column represents a cell.

Feature	Cell1	Cell2	Cell3	Cell4	Cell$_n$
Feature 1	0	2	7	3	…
Feature 2	0	4	6	0	…
Feature 3	2	0	2	0	…

b. A hierarchical data format (hdf5) feature-barcode matrix, generally named as *filtered_feature_bc_matrix.h5* in the 10× Genomics CellRanger output folder.

c. The Hdf5 format consists of a matrix and metadata. The matrix contains barcodes (barcode sequences), data (Nonzero UMI counts in column-major order), indices (0-based row index), indptr (0-based column index) and shape (a tuple contains matrix dimensions, (# row, # column); and the metadata contain diverse attributes.

The Hdf5 file hierarchy example (https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/advanced/h5_matrices) :

```
(root)
 └── matrix [HDF5 group]
     ├── barcodes
     ├── data
     ├── indices
     ├── indptr
     ├── shape
     └── features [HDF5 group]
         ├ _all_tag_keys
         ├ target_sets [for Targeted Gene Expression]
         |… └ [target set name]
         ├ feature_type
         ├ genome
         ├ id
         ├ name
         ├ pattern [Feature Barcode only]
         ├ read [Feature Barcode only]
         └ sequence [Feature Barcode only]
```

d. The three gzip files recording information of barcodes (*barcodes.tsv*), features (*genes.tsv*), and gene expressions (*matrix.mtx*) in the 10× Genomics CellRanger output folder.

Barcodes file (Barcodes.tsv)

Barcode
AAACATACAAAACG-1
AAACATACAAAAGC-1
AAACATACAAACAG-1
AAACATACAAACGA-1
…

Features (genes.tsv)

The first column represents the Ensembl gene id, and the second column represents the gene symbol.

Ensembl_id	Gene_symbol
ENSG00000243485	MIR1302-10
ENSG00000237613	FAM138A
ENSG00000186092	OR4F5
ENSG00000238009	RP11-34P13.7
…	…

Gene expression matrix (matrix.mtx)

i. The first column "gene id index" represents the row number of genes in the *genes.tsv*.
ii. The second column "cell id index" represents the row number of cell-barcode in the *barcodes.tsv*.
iii. The third column represents the total Unique Molecular Identifier (UMI) count in each cell and gene combination.

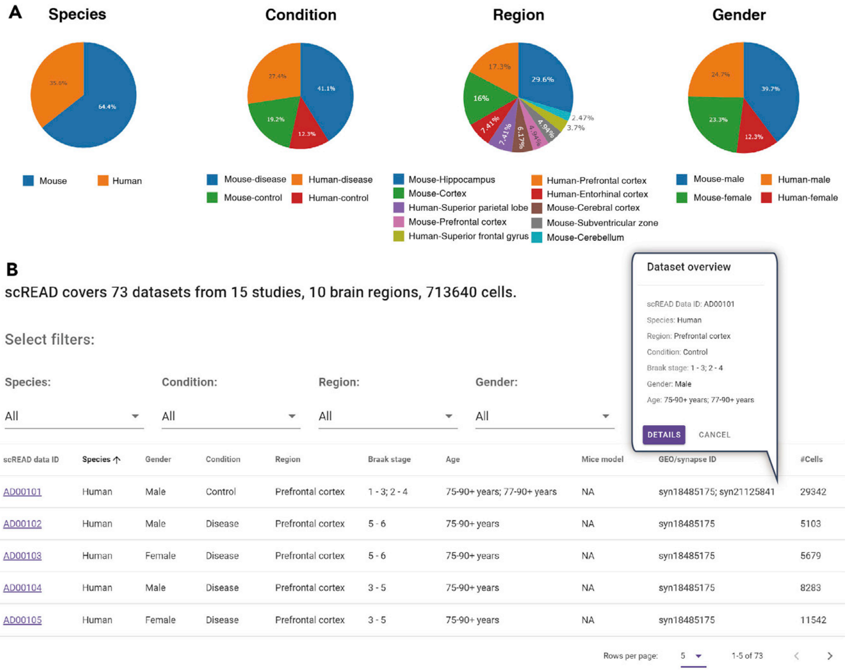
Gene id index	Cell id index	UMI count
498	1	1
5423	1	6
6374	1	3
12932	1	1
…	…	…

scREAD maintenance and sustainability plan

Currently, scREAD only contains the scRNA-seq and snRNA-seq data sets as of September 22nd, 2020. We will routinely collect more publicly available AD scRNA-seq and snRNA-seq data. You may also contact us through emails and let us know about your dataset of interest to be added in scREAD. We will continue to develop scREAD to support integrative analysis of single-cell multimodal omics data and spatial transcriptomics data. We appreciate any suggestions or feedback from you.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
AD scRNA-seq data (Grubman et al., 2019)	GEO	GSE138852
scREAD database	(Jiang et al., 2020)	https://bmbls.bmi.osumc.edu/scread/
All code for the scREAD protocols	GitHub	https://github.com/OSU-BMBL/scread-protocol
The interactive tutorial for optional step: calculating overlapping DEGs from multiple comparisons to find overlapping genes	Google Colab	https://colab.research.google.com/drive/1llnXa6jD4yc7RGJc0EWDfy5NNoXT1qye?usp=sharing
Software and algorithms		
Chrome	Google	https://www.google.com/chrome/
Safari	Apple	https://www.apple.com/safari/
Microsoft Edge	Microsoft	https://www.microsoft.com/en-us/edge
Firefox	Mozilla	https://www.mozilla.org/en-US/
R (>=3.6)	R Project	https://www.r-project.org/
IRIS3 (v1.2.4)	(Ma et al., 2020)	https://bmbl.bmi.osumc.edu/iris3/
Seurat (v3.2)	(Stuart et al., 2019)	https://satijalab.org/seurat
Harmony (v0.1)	(Korsunsky et al., 2019)	https://github.com/immunogenomics/harmony
Polychrome (v1.2.5)	R CRAN	https://cran.r-project.org/web/packages/Polychrome
SCINA (v1.2.0)	(Zhang et al., 2019)	https://github.com/jcao89757/SCINA
RColorBrewer (v1.1-2)	R CRAN	https://cran.r-project.org/web/packages/RColorBrewer
ggplot2 (v3.3.2)	R CRAN	https://ggplot2.tidyverse.org/
tidyverse (v1.3.0)	R CRAN	https://www.tidyverse.org/
cowplot (v1.0.0)	(Wilke et al., 2021)	https://github.com/wilkelab/cowplot/tree/1.1.1
RMySQL (v0.10.21)	R CRAN	https://cran.r-project.org/web/packages/RMySQL
future (v1.21.0)	R CRAN	https://cran.r-project.org/web/packages/future
MAST (v1.16.0)	(Finak et al., 2015)	https://www.bioconductor.org/packages/release/bioc/html/MAST.html



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

Select filters:

Species: All
Condition: All
Region: All
Gender: All

Dataset 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

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: 5 1-5 of 73

Figure 1. Overview of the scREAD homepage

(A) 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 color in each pie chart represents one element, and the number represents the distribution ratio for each element under each factor for 73 datasets.

(B) The table shows the general information of all 73 datasets. Users can select filters and the table will be updated accordingly. Clicking a row in the table will pop up the dataset overview panel, and users can navigate to the dataset details page through the link.

STEP-BY-STEP METHOD DETAILS

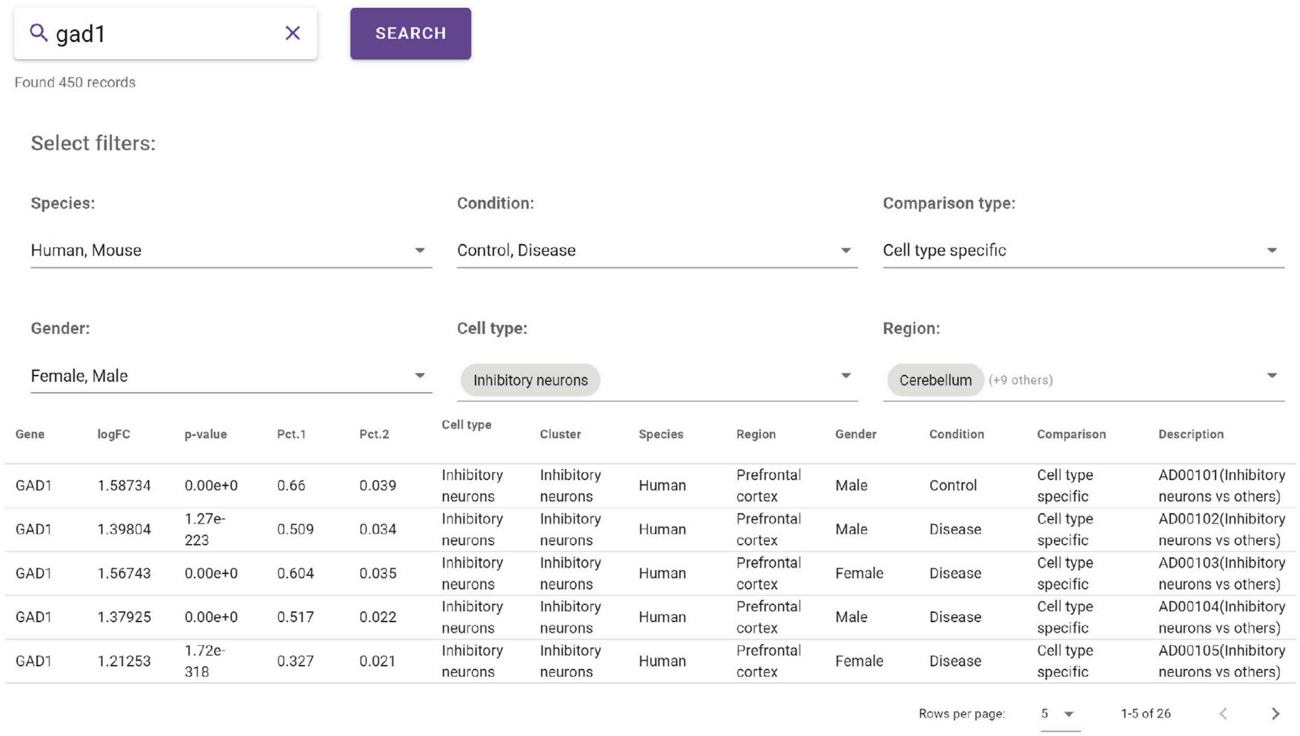
Checking AD studies summary statistics

Timing: 10 min

This section allows one to browse the overall summary statistics, including species, disease condition, brain regions, and gender. The researcher can then make an informed decision of which dataset of interest one should navigate to.

The current release of the scREAD collected datasets from 15 studies in total. Based on the metadata provided from the original papers, we constructed the original samples into 73 datasets, 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), disease or control, and age stage (7 months, 15 months, or 20 months for mice, and 50–100+ years old for human).

1. Navigate to https://bmbls.bmi.osumc.edu/scread/, and the dataset summary should be listed (Figures 1A and 1B).
2. You can either click on one of the pie charts or select filters from the dataset summary table, including species, sample condition, brain region, and gender. The content of the table will be updated accordingly.
3. You can click any row in the table, and a dataset overview panel will pop. You can further navigate to the dataset details page through the link.



gad1

Found 450 records

Select filters:

Species: Human, Mouse

Condition: Control, Disease

Comparison type: Cell type specific

Gender: Female, Male

Cell type: Inhibitory neurons

Region: Cerebellum (+9 others)

Gene	logFC	p-value	Pct.1	Pct.2	Cell type	Cluster	Species	Region	Gender	Condition	Comparison	Description
GAD1	1.58734	0.00e+0	0.66	0.039	Inhibitory neurons	Inhibitory neurons	Human	Prefrontal cortex	Male	Control	Cell type specific	AD00101(Inhibitory neurons vs others)
GAD1	1.39804	1.27e-223	0.509	0.034	Inhibitory neurons	Inhibitory neurons	Human	Prefrontal cortex	Male	Disease	Cell type specific	AD00102(Inhibitory neurons vs others)
GAD1	1.56743	0.00e+0	0.604	0.035	Inhibitory neurons	Inhibitory neurons	Human	Prefrontal cortex	Female	Disease	Cell type specific	AD00103(Inhibitory neurons vs others)
GAD1	1.37925	0.00e+0	0.517	0.022	Inhibitory neurons	Inhibitory neurons	Human	Prefrontal cortex	Male	Disease	Cell type specific	AD00104(Inhibitory neurons vs others)
GAD1	1.21253	1.72e-318	0.327	0.021	Inhibitory neurons	Inhibitory neurons	Human	Prefrontal cortex	Female	Disease	Cell type specific	AD00105(Inhibitory neurons vs others)

Rows per page: 5 1-5 of 26

Figure 2. Example DGE analysis searching result of the *GAD1* gene, a marker gene of inhibitory neurons
Users can select filters and the table will be updated accordingly.

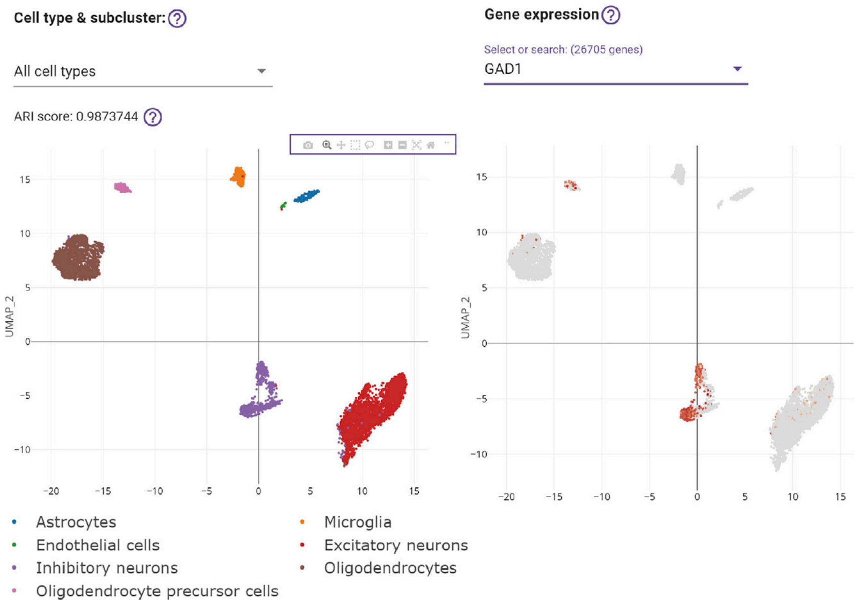
Searching genes of interest from the differential gene expression (DGE) analysis results

⏲ Timing: 20 min

This section allows one to search a gene of interest from DGE analysis results across multiple comparisons. For detailed gene information, the researcher can check from the link to the dataset ID of interest (Figure 2).

4. To search genes of interest from DGE analysis results across multiple datasets, type the gene symbol in the search box, and click on the search button. You can select filters to specify dataset sources or decide which comparison types should be displayed.
5. The query results are returned from multiple DGE analyses, including cell-type-specific genes, subcluster specific genes, AD vs control differentially expressed genes (DEGs), or AD vs AD DEGs.
   a. Cell-type-specific genes were identified by performing DGE analysis between the cell type of interest and the average of the remaining cell types.
   b. Subcluster-specific genes were identified by performing DGE analysis between the subcluster of interest and the average of the remaining subclusters from the same cell type.
   c. AD vs control DGE analysis was performed within the same cell type, brain region, and gender. For example, Male-AD-Prefrontal cortex vs Male-Control-Prefrontal cortex.
   d. AD vs AD DGE analysis was performed within the sample cell type while based on different sample conditions. For example, one AD vs AD comparison can be Male-AD-Prefrontal cortex vs Male-AD-Entorhinal Cortex or Male-AD-Prefrontal cortex vs Female-AD-Prefrontal cortex.

Note: The 'multiple comparison types' in the comparison type selection box include disease vs control performed within the same dataset and disease vs disease performed from two



Cell type & subcluster:
All cell types
ARI score: 0.9873744

Gene expression
Select or search: (26705 genes)
GAD1

Figure 3. Cell clustering and gene expression result from the scREAD homepage
Cell clustering results including UMAP plot colored by cell types or subclusters (left), and searching gene expression using the same UMAP coordinates (right). The darker the color is in this UMAP, the higher the expression value of the gene.

different datasets. A positive log foldchange (FC) value indicates the gene expressions are higher in the first group. The log FC is returned in the natural logarithm.

Checking cell clustering results

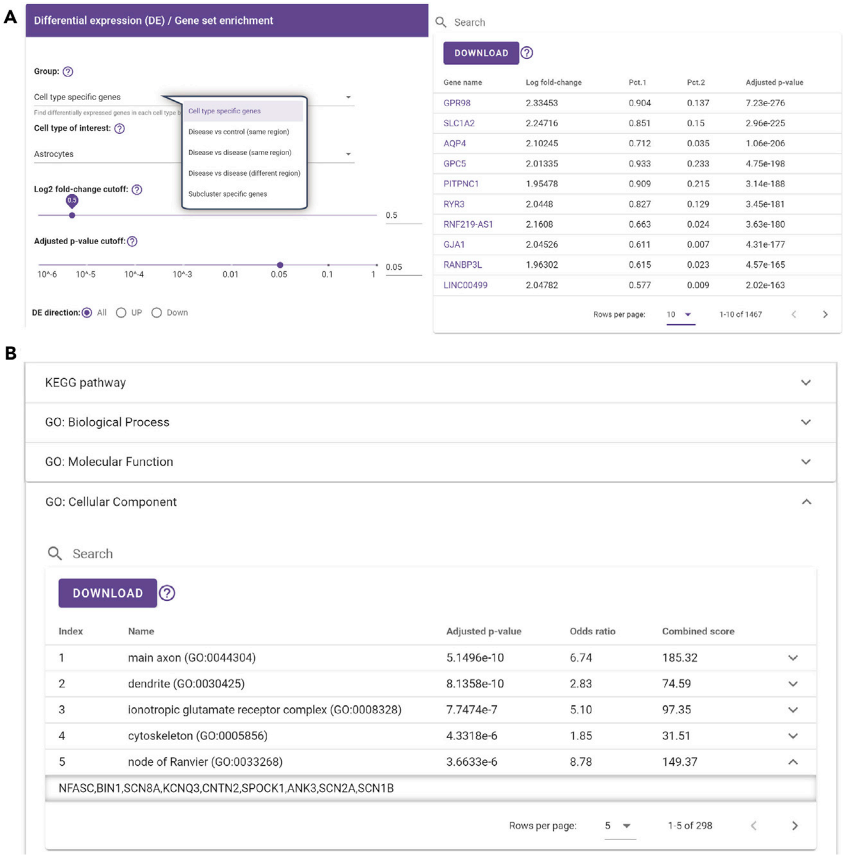
Timing: 20 min

In this section, we used a dataset from scREAD as an example to show the analysis result (ID: AD00103), https://bmbls.bmi.osumc.edu/scread/AD00103. This dataset consists of 6,629 cells isolated from a human AD female prefrontal cortex sample (Mathys et al., 2019). As we know, not all cells collected from AD patient samples are malignant, and some healthy cells may be included in the cell populations, which were defined as healthy-like cells in Granja et al.'s study (Granja et al., 2019). We applied this concept to all AD datasets in scREAD and defined these healthy-like cells as control-like cells. These control-like 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 disease datasets and identify AD-associated cells. Here, scREAD filtered out 950 control-like cells and kept 5,679 AD-associated cells for the downstream analysis. See the quantification and statistical analysis section for more details about how scREAD filtered out control-like cells.

6. Checking cell clustering results (Figure 3).
   a. By default, all the cell types will be selected and the corresponding Adjusted Rand Index (ARI) will be displayed. The ARI score is used to evaluate the similarity of our predicted cell types compared with the original cell labels in the original paper.

   Note: The ARI score will not be displayed when the cell labels were not provided from the original paper, and a silhouette score will be displayed instead. Meanwhile, the ARI score will be hidden if users did not select all cell types.

   b. Choose one of these cell types, the following Uniform Manifold Approximation and Projection (UMAP)(Becht et al., 2019) will change to the UMAP of predicted subclusters for this specific cell type.



Differential expression (DE) / Gene set enrichment

Group: ⓘ

Cell type specific genes

Find differentially expressed genes in each cell type b

- Cell type specific genes
- Disease vs control (same region)
- Disease vs disease (same region)
- Disease vs disease (different region)
- Subcluster specific genes

Cell type of interest: ⓘ

Astrocytes

Log2 fold-change cutoff: ⓘ 0.5

Adjusted p-value cutoff: ⓘ 0.05

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

KEGG pathway

GO: Biological Process

GO: Molecular Function

GO: Cellular Component

Search

DOWNLOAD ⓘ

Index	Name	Adjusted p-value	Odds ratio	Combined score
1	main axon (GO:0044304)	5.1496e-10	6.74	185.32
2	dendrite (GO:0030425)	8.1358e-10	2.83	74.59
3	ionotropic glutamate receptor complex (GO:0008328)	7.7474e-7	5.10	97.35
4	cytoskeleton (GO:0005856)	4.3318e-6	1.85	31.51
5	node of Ranvier (GO:0033268)	3.6633e-6	8.78	149.37

NFASC,BIN1,SCN8A,KCNQ3,CNTN2,SPOCK1,ANK3,SCN2A,SCN1B

Rows per page: 5 1-5 of 298

Figure 4. Differential gene expression (DGE), and functional gene set enrichment based on DEGs

(A) Differential gene expression analysis panel, the comparison groups include cell-type-specific genes, subcluster specific genes, and DEGs from the cross-dataset comparison.

(B) KEGG pathway, GO biological process, molecular function, and cellular component analysis using the DEGs from (A), an enriched genes example were displayed on the 5th GO cellular component term.

c. A sliding bar is used for controlling the size of each point in the following UMAP. It ranges from 1 to 10, i.e., the bigger the number is, the larger the point size is.
d. This function bar contains several quick buttons for graphic operations.
e. Hovering the cursor on cell points will display cell type, cell name, and the UMAP coordinates.
f. The legend of this UMAP plot will be displayed based on the genes selected in the drop-down bar. The darker the color is in this UMAP, the higher the expression value of the gene.

⚠ CRITICAL: Rendering gene expression scatter plot can be slow due to network speed or a large number of cells data need to process, please be patient while scREAD is fetching data from the backend server.

Checking differential expression (DE) results and performing functional enrichment analysis

⏱ Timing: 20 min

In this section, we used the same example data from checking cell clustering results to illustrate DGE analysis results and to perform functional gene set enrichment analysis based on DEGs.

7. First, apply the necessary filtering criteria from the DGE analysis results panel (Figure 4A).

8. DGE analysis groups for browsing cell-type-specific genes, subcluster specific genes, and DE genes from the cross-dataset comparison.
   a. Choose the cell type of interest in DGE analysis.
   b. Choose the log fold-change ranges. (default = 0.5; ranges from 0 to 5).
   c. The adjusted p value ranges. (default = 0.05; ranges from 10^−6 to 1).
   d. The DE direction can filter by all DE genes, only up-regulated genes, only down-regulated genes (default = 'all').
   e. You can search for genes that you are interested in, and then the following table will return the matching result.
   f. Download the currently listed table.
   g. GeneCards database (https://www.genecards.org/) is linked to each gene in the table.

   Note: Adjusting any parameters above will immediately affect the displayed DEGs table.

9. Performing functional gene set enrichment (Figure 4B)
   a. KEGG pathway, GO biological process, GO molecular function, and GO cellular component analysis using the DEGs from above, an example of enriched genes are displayed on the 4th GO cellular component term. You can also search for a specific item by entering the content that you want to search in the search box.

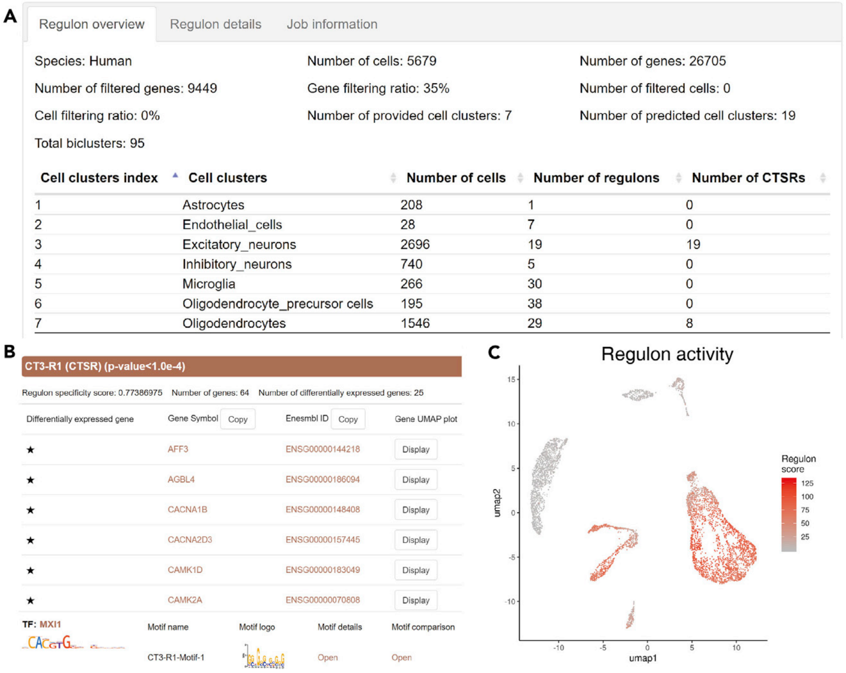
   ⚠ CRITICAL: The functional gene set analysis results are calculated in real-time by sending the DEGs to the Enrichr (Kuleshov et al., 2016) web server (https://maayanlab.cloud/Enrichr/). Thus, changing DEG log FC or p value cutoffs can significantly change the results of enrichment analysis. Considering Enrichr does not provide options to submit a custom background and the results could be potentially misleading (Timmons et al., 2015). We also provide a link to another enrichment analysis tool, g: Profiler (https://biit.cs.ut.ee/gprofiler/), which allows users to submit custom background gene sets.

Identifying cell-type-specific regulons

Timing: 20 min

In this section, we describe the process of identifying cell-type-specific regulons (CTSRs) using IRIS3.

10. In the DE section, when you select the "Cell-type-specific genes" item in the "Group" select box, cell-type-specific regulon analysis will be performed. CTSRs results are displayed at the bottom of the screen. Click on the "Cell-type-specific regulons" bar, detailed CTSRs information will be shown.
11. The CTSRs are displayed in a reactive table in the DE section, each row shows that for each cell type, a set of genes are regulated by a specific transcription factor (TF). Clicking on the "Regulon overview" panel, a table in the panel summarizes the overall cell number and regulon number in each cell cluster (Figure 5A).
12. You can navigate into the IRIS3 to see the detailed results of this job by clicking the 'Open cell-type-specific regulon result page in the new tab' button. In the table, the index number will be given to represent CTSRs (Figure 5B).
    a. Both gene compositions of regulons and their expression values across different cell types can be intuitively displayed in a heatmap. Regulons are ranked in increasing order of the empirical p values of regulon specificity scores (RSS) as described above, and a regulon is named as CTn-Rm with 'n' representing the index of cell type and 'm' represents the regulon rank. Due to the space limitation, only the top ten regulons and their corresponding genes are showcased in the heatmap, and the component genes of each regulon are indicated as green rectangles. The heatmap records the log-transformed expression level of each top-ten-regulon-covered gene across all cells.
    b. Regulon results are separately showcased in each cell type. Click on the "CT#" button to switch to see results in other cell types. A scatter plot shows the distribution of the RSS of



A

Regulon overview | Regulon details | Job information

Species: Human | Number of cells: 5679 | Number of genes: 26705
Number of filtered genes: 9449 | Gene filtering ratio: 35% | Number of filtered cells: 0
Cell filtering ratio: 0% | Number of provided cell clusters: 7 | Number of predicted cell clusters: 19
Total biclusters: 95

Cell clusters index	Cell clusters	Number of cells	Number of regulons	Number of CTSRs
1	Astrocytes	208	1	0
2	Endothelial_cells	28	7	0
3	Excitatory_neurons	2696	19	19
4	Inhibitory_neurons	740	5	0
5	Microglia	266	30	0
6	Oligodendrocyte_precursor cells	195	38	0
7	Oligodendrocytes	1546	29	8

B

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
★	CAMK1D	ENSG00000183049	Display
★	CAMK2A	ENSG00000070808	Display

TF: MXI1	Motif name	Motif logo	Motif details	Motif comparison
	CT3-R1-Motif-1		Open	Open

Figure 5. Cell-type-specific regulon (CTSR) analysis results from IRIS3

(A) Overview table of identified CTSRs.

(B) An example of CTSR (p value < 0.05), stars indicate differential expressed gene within the cell type, the corresponding matched TF is linked to the HOCOMOCO database.

(C) UMAP plot colored by regulon activities in each cell.

each regulon. CTSRs are ranked top and marked as blue dots with their representative TF names, and insignificant regulons are marked as grey dots. For each regulon, genes and the corresponding TF are presented (Figure 5B) with several actions that link to showing heatmap, functional gene set enrichment analysis, and regulon activities in the UMAP plot (Figure 6C). A more detailed interpretation of each regulon can be found on the IRIS3 website, https://bmbl.bmi.osumc.edu/iris3/tutorial.php#3example&q=2.

⚠ CRITICAL: The CTSRs results are only available when you selected the cell-type-specific option in the DEG group box.

Optional step: calculating overlapping DEGs from multiple comparisons

⏲ Timing: 1 h

In this section, we provide a workflow for calculating overlapping DEGs from multiple comparisons. Suppose we have *m* AD vs control comparisons from a cell type of interest in a specific brain region. For each comparison, we select top *t* DEGs based on the ranked log FC. We define an "overlapping gene" as the gene that appears at least *n* times in *m* comparisons ($n \leq m$). $t, n$are parameters set by the users.

Here, we provide two approaches, you can follow the code example below on your local R environment; For users wishing to avoid setup procedures, you can use the following link from Google Colab, which is an interactive computational environment that combines live code, visualizations, and explanatory text, https://colab.research.google.com/drive/1IlnXa6jD4yc7RGJc0EWDfy5NNoXT1qye?usp=sharing.

13. If you wish to perform the calculation in your local computer, first, load the R packages, scREAD data, and predefined functions in your R local environment:

```
library(tidyverse)
library(RVenn)
library(rlist)
library(knitr)
tryCatch({
  load(
    url(
      'https://bmbl.bmi.osumc.edu/downloadFiles/scread/protocol/scre
ad_db.rdata'
    )
  )
}, error = {
  load(
    url(
      'https://github.com/OSU-BMBL/scread-
protocol/raw/master/overlapping_genes/scread_db.rdata'
    )
  )
})
```

14. To calculate overlapping genes, these parameters are needed
    a. The number of genes to be selected in each AD vs control DEG results (default = 100)
    b. Species (default = Human)
    c. Brain region (e.g., Entorhinal Cortex)
    d. DE direction (e.g., up)
    e. Overlap threshold (For example, A gene is an overlapping gene if A should at least appear 3 times in total 4 comparisons, here the threshold is 3)
15. We can then process some of our metadata:

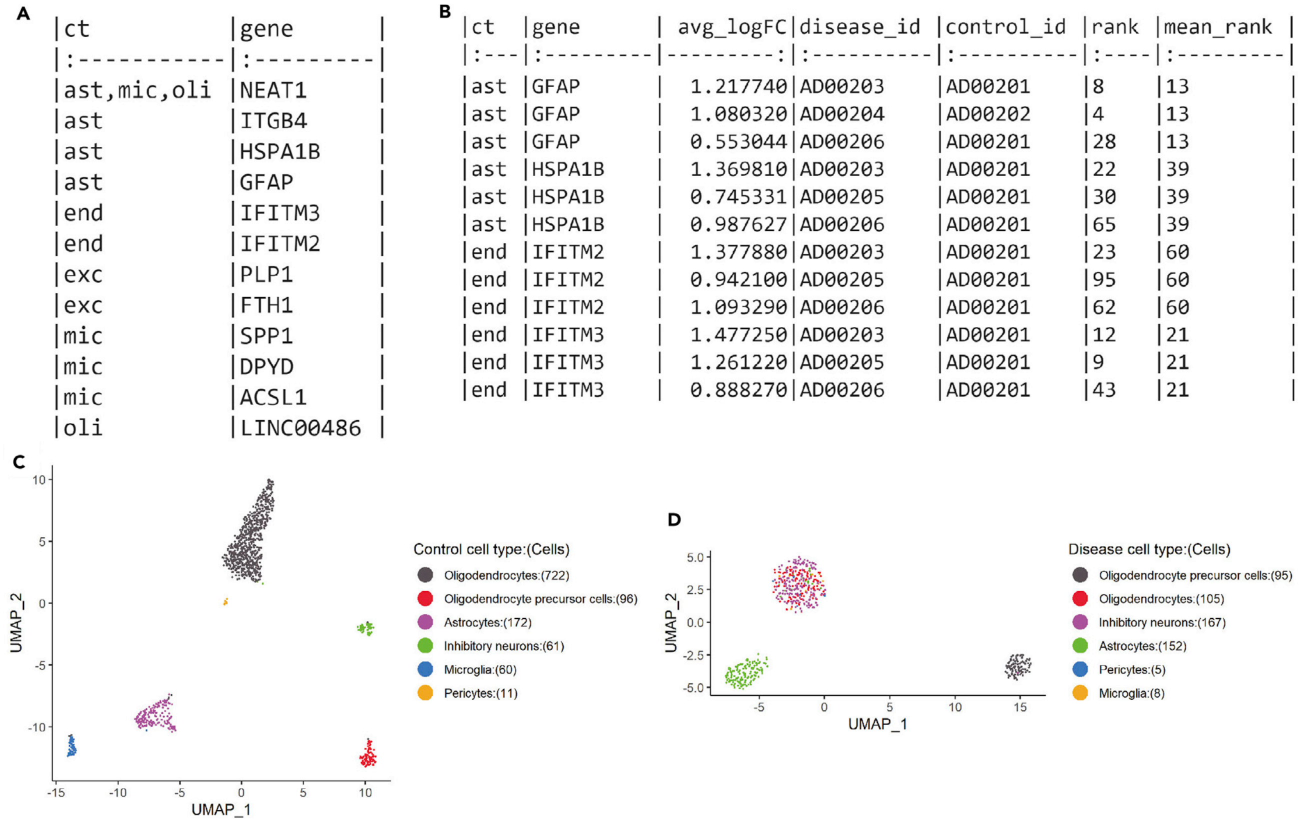
```
REGION_LIST <- sort(unique(dataset$region))
CT_LIST <- sort(unique(cell_type_meta$cell_type))
CT_SHORT_LIST <- CT_LIST
CT_SHORT_LIST[CT_LIST=="Oligodendrocyte precursor cells"] <-
"opc"
CT_SHORT_LIST <- tolower(substr(CT_SHORT_LIST, 1, 3))
```

16. Below are the necessary settings to calculate the overlapping genes.
    a. We use the top 100 DE genes in each AD vs control comparison:

```
TOP <- 100
```

    b. Species should be either 'Human' or 'Mouse':

```
this_species <- 'Human'
```



```
|ct          |gene      |
|:-----------|:---------|
|ast,mic,oli |NEAT1     |
|ast         |ITGB4     |
|ast         |HSPA1B    |
|ast         |GFAP      |
|end         |IFITM3    |
|end         |IFITM2    |
|exc         |PLP1      |
|exc         |FTH1      |
|mic         |SPP1      |
|mic         |DPYD      |
|mic         |ACSL1     |
|oli         |LINC00486 |
```

```
|ct  |gene      | avg_logFC|disease_id |control_id |rank |mean_rank |
|:---|:---------|---------:|:----------|:----------|:----|:---------|
|ast |GFAP      |  1.217740|AD00203    |AD00201    |8    |13        |
|ast |GFAP      |  1.080320|AD00204    |AD00202    |4    |13        |
|ast |GFAP      |  0.553044|AD00206    |AD00201    |28   |13        |
|ast |HSPA1B    |  1.369810|AD00203    |AD00201    |22   |39        |
|ast |HSPA1B    |  0.745331|AD00205    |AD00201    |30   |39        |
|ast |HSPA1B    |  0.987627|AD00206    |AD00201    |65   |39        |
|end |IFITM2    |  1.377880|AD00203    |AD00201    |23   |60        |
|end |IFITM2    |  0.942100|AD00205    |AD00201    |95   |60        |
|end |IFITM2    |  1.093290|AD00206    |AD00201    |62   |60        |
|end |IFITM3    |  1.477250|AD00203    |AD00201    |12   |21        |
|end |IFITM3    |  1.261220|AD00205    |AD00201    |9    |21        |
|end |IFITM3    |  0.888270|AD00206    |AD00201    |43   |21        |
```

Figure 6. Overlapping genes and annotated cell type (ct) from the example dataset

(A) The up-regulated overlapping genes in Human Entorhinal Cortex Astrocytes (ast).

(B) The log FC, dataset source, and rankings from the overlapping genes table in (A).

(C) A UMAP plot of the control dataset example with six cell types was annotated from the scREAD workflow.

(D) A UMAP plot of the disease dataset example with six cell types was transferred from the example reference control dataset.

c. Specify our brain region of interest, here we selected the 5th brain region in REGION_LIST variable, i.e., Entorhinal Cortex':

```
this_region <- REGION_LIST[5]
```

d. DE direction should either 'up' or 'down', 'up' means we select DE genes that are expressed higher in the disease dataset (the first group):

```
this_direction <- 'up'
```

e. The OVERLAP_THRES should be manually defined based on your interest and the total number of comparisons in scREAD. For example, scREAD have 4 total AD vs control datasets comparisons, we set the threshold to 3, meaning that we want to find overlapping genes that are at least appeared in 3 comparisons:

```
OVERLAP_THRES <- 3
```

f. Now, we can calculate the overlapping genes based on the parameters above, the results are stored in a list variable:

```
result <- calc_overlap_list()
```

g. Two tables can be generated by accessing the result variable:
   i. The overlapping genes in the selected brain region (Figure 6A)

```
print(result$list)
```

   ii. The detailed information, including rankings, log FC, dataset source information from the overlapping genes (Figure 6B)

```
print(result$rank)
```

Optional section 7: Running the scREAD backend analysis workflow locally

Timing: 2 h

In this section, we present how to run the scREAD workflow to process a custom dataset. The workflow can be used in the Unix command-line environment with R installed.

17. Download the scREAD workflow and an example dataset from https://github.com/OSU-BMBL/scread-protocol/tree/master/workflow, the folder should contain the following files (**5 min**):
    a. custom_marker.csv: A manually created marker gene list file used for identified cell types.
    b. functions.R: Visualization functions used in R.
    c. build_control_atlas.R: build control cells atlas Seurat object from count matrix file.
    d. transfer_cell_type.R: filter out control-like cells in disease dataset.
    e. run_analysis.R: run analysis workflow, and export tables in the scREAD database format.
    f. example_control.csv. The example control dataset.
    g. example_disease.csv. The example disease dataset.

18. Build the control atlas file from the raw gene expression matrix (**5 min**).
    a. Prepare your control gene expression data. 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 protocol, we will run example_control.csv.
    b. build_control_atlas.R takes three parameters:
       i. Working directory path.
       ii. Control data path.
       iii. Output data ID.
    c. Next, run the following command, remember to change PATH_TO_WD to your working directory path:

```
cd PATH_TO_WD
Rscript build_control_atlas.R PATH_TO_WD example_control.csv
control_example
```

    d. The expected output for this step contains four files:
       i. control_example.rds: The Seurat R object storing example control data.
       ii. control_example_expr.txt: Filtered gene expression matrix.
       iii. control_example_cell_label.txt: The first column is the cell name, the second column is the cell type information.
       iv. control_example_umap.png: UMAP plot of example control data colored by cell types (Figure 6C).

19. Transfer cell types based on control atlas, the goal of this step is to annotate cell type using the control atlas as the reference, onto the disease gene expression matrix file (~**5 min**).
    a. 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).
    b. transfer_cell_type.R takes four parameters:
        i. Working directory path.
        ii. Control atlas Seurat object file name.
        iii. Disease gene expression matrix name.
        iv. Output disease data ID.
    c. Next, run the following command

```
cd PATH_TO_WD
Rscript transfer_cell_type.R PATH_TO_WD control_example.rds example_dise
ase.csv disease_example
```

    d. The expected output for this step contains four files:
        i. disease_example.rds: The Seurat R object storing example disease data.
        ii. disease_example_expr.txt: Filtered gene expression matrix.
        iii. disease_example_cell_label.txt: The first column is the cell name, the second column is the cell type information.
        iv. disease_example_umap.png: UMAP plot for disease data colored by cell types (Figure 6D).

20. Run data analysis, the goal of this step is to identify cell-type-specific genes, DEGs from two example datasets (**60 min**).
    a. 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)
    b. run_analysis.R takes three parameters:
        i. Working directory path.
        ii. Control Seurat object file name.
        iii. Disease Seurat object file name.
    c. Next, run the following command:

```
cd PATH_TO_WD
Rscript run_analysis.R PATH_TO_WD control_example disease_example
```

    d. The expected output for this step contains three folders:
        i. /de. Differential gene expression analysis results.
            Cell-type-specific genes.
            Sub-cluster specific genes.
            DEGs between two conditions.
        ii. /dimension: UMAP coordinates for two datasets.
        iii. /subcluster_dimension: UMAP coordinates for each sub-clusters in two datasets.

21. Identify CTSRs using IRIS3 (**2 h**).
    a. Navigate to https://bmbl.bmi.osumc.edu/iris3/submit.php, submit two jobs for the two example datasets:
        i. upload control_example_expr.txt and control_example_cell_label.txt
        ii. upload disease_example_expr.txt and disease_example_cell_label.txt
    b. The expected output for IRIS3 contains these files:
        i. Lists
            CTSR gene list
            Marker gene list

Gene module list
Motif list
Transcription factor list

ii. Tables

Predicted cell types
Bulk ATAC peak enrichment
TAD association

iii. Figures (only display on the website)

CTSR active UMAP
CTSR gene heatmap
Trajectory
UMAP

⚠ CRITICAL: You need to open the advanced options tab in the IRIS3 submission page to upload a custom cell label.

EXPECTED OUTCOMES

Web server results

scREAD provides comprehensive analysis results for the selected AD scRNA-seq or snRNA-seq datasets. These summaries and integrated analysis of differentially expressed genes (DEGs) are presented in a graphical and tabular format (Figure 1).

Figure 2A shows the dataset details page that outlines all the necessary information regarding dataset metadata, source, and related data IDs from the same study. Figure 3 annotates cell clustering results including UMAP plot colored by cell types (left), and gene expression using the same UMAP coordinates (right). The darker the color is in this UMAP, the higher the expression value of the gene.

Figure 4Ashows DEGs after the user selects the comparison group and the cell type of interest. Users can sort the table by log FC or p values. The DEGs results table can be downloaded as a tab-separated value file for further analysis. The DEGs are filtered based on the threshold set by the user, and the functional gene set enrichment analysis for the KEGG pathways, Gene Ontology terms are calculated using the current DEGs as input (Figure 4B). If the user selects ‘cell-type-specific genes’ in the group select box, the identified CTSRs along with RSS will be displayed at the bottom of the screen. The user can also navigate to the IRIS3 website to browse the regulons for more details.

Figure 5 shows the identified CTSR analysis results from IRIS3, including the table of all identified regulons (Figure 5A), and detailed information from CT3-R1 (cell type: 3, regulon 1), the stars indicate differential expressed gene within the cell type, the corresponding matched TF is linked to the HOCOMOCO (Kulakovskiy et al., 2018) database (Figure 5B). The regulon activities are visualized in a UMAP plot (Figure 5C).

Overlapping DEGs results

After calculating overlapping DEGs from the same cell type across datasets using the default settings, two tables for overlapping genes will be generated:

The overlapping genes in the selected region, using the default settings (Figure 6A), the fourth row in the table below can be interpreted as: ‘For all AD vs control datasets comparisons in Human Entorhinal Cortex Astrocytes (ast), the GFAP gene ranked top 100 by log FC values in at least 3 comparisons’.

The ranking information from the overlapping genes. Using the GFAP gene as an example, we found GFAP is an overlapping gene in Human Entorhinal Cortex Astrocytes. The GFAP ranked top 50 in 3 comparisons of 4 total comparisons, the mean rank of the gene is 13, and the average log-FC of GFAP in each comparison are also listed (Figure 6B).

Backend workflow results

The backend workflow generated is a series of tables containing information on the intermediate results and final analysis results, including cell-type-specific genes, subcluster-specific genes, and DEGs between two datasets. The following descriptions of output files use the example dataset given as the filename, while the names of your files will differ depending on your filename settings.

The control atlas output should contain four files:

File 1: control_example.rds: The Seurat R object storing example control data.
File 2: control_example_expr.txt: Filtered gene expression matrix.
File 3: control_example_cell_label.txt: The first column is the cell name, the second column is the cell type information.
File 4: control_example_umap.png: UMAP plot of the example control data. The UMAP plot visualizes scRNA-seq data in two-dimensional spaces. Each dot represents a cell and cells are colored by annotated cell types. Cells within the cell types should co-localize on the UMAP plot (Figure 6C).

The disease dataset output should contain four files:

File 1 : disease_example.rds: The Seurat R object storing example disease data.
File 2: disease_example_expr.txt: Filtered gene expression matrix.
File 3: disease_example_cell_label.txt: The first column is the cell name, the second column is the cell type information.
File 4: disease_example_umap.png: UMAP plot of the example disease data. The UMAP plot visualizes scRNA-seq data in two-dimensional spaces. Each dot represents a cell and cells are colored by annotated cell types. Cells within the cell types should co-localize on the UMAP plot (Figure 6D).

The output tables should be stored in three folders:

Folder 1: /de. Differential gene expression analysis results.
  Cell-type-specific genes.
  Sub-cluster specific genes.
  DEGs between two conditions.
Folder 2: /dimension: UMAP coordinates for two datasets.
Folder 3: /subcluster_dimension: UMAP coordinates for each sub-clusters in two datasets.

QUANTIFICATION AND STATISTICAL ANALYSIS

To determine whether cells from disease datasets are control-like, the Harmony R package (v1.0) was first used to integrate the disease dataset with its corresponding control atlas. 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.

Differential gene expression analysis was performed using MAST (Finak et al., 2015). Seurat's FindAllMarkers and FindMarkers functions that utilize the MAST package were used to run DGE analysis on normalized gene expression data. Cell-type-specific genes were identified by performing DGE analysis between the cell type of interest and the average of the remaining cell types. Subcluster-specific genes were identified by performing DGE analysis between the subcluster of interest and the average of the remaining subclusters from the same cell type. For each cell type, several DGE analysis was 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 DGE analyses.

Functional enrichment analysis was performed using the Enrichr web server. The p value was computed using a standard statistical method used by most enrichment analysis tools: Fisher's exact test or the hypergeometric test. This is a binomial proportion test that assumes a binomial distribution and independence for the probability of any gene belonging to any set.

CTSRs were identified using IRIS3. The RSS for a cell type was calculated according to the entropy of regulon activity score (RAS) of cells within the cell type compared to other cell types. An RSS ranges from 0 to 1, with a higher value representing greater specificity of a regulon in the cell type. 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, 10 000 times. Regulon P-values are Bonferroni-adjusted by multiplying the number of regulons in the exact cell type. Regulons with adjusted P-values < 0.05 (by default) are considered CTSRs.

LIMITATIONS

Limitation 1

Although scREAD is continuing to collect all publicly available AD scRNA-seq & snRNA-seq data, you may still find your dataset of interest was not included in scREAD. This may be due to the policy that we are not allowed to put some datasets in scREAD.

Limitation 2

scREAD uses a semi-supervised cell type annotation method, thus, only eight major cell types from our marker genes list can be annotated, i.e., astrocytes, endothelial cells, excitatory neurons, inhibitory neurons, microglia, oligodendrocytes, oligodendrocyte precursor cells, and pericytes. 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 be finely resolved 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 the subclusters.

TROUBLESHOOTING

Problem 1

The content on the scREAD webpage cannot be displayed correctly in checking AD studies summary statistics.

Potential solution

Please make sure to use one of the modern browsers, including Chrome, Firefox, Microsoft Edge, or Safari. Internet Explorer is not supported in scREAD.

Problem 2

Uploading incorrect file format to the scREAD online analysis workflow (https://bmbls.bmi.osumc.edu/scread/submit).

Potential solution

scREAD online analysis workflow needs to provide a gene expression matrix in text format, in which each row represents a gene, each column represents a cell.

Problem 3

Difficulty in selecting parameters at DEGs at the DE results in checking differential expression (DE) results and performing functional enrichment analysis.

Potential solution

Adjusting any parameters in the DE section will immediately affect the displayed DEGs table. A less-stringent p value is recommended to be used to select a large preliminary list of genes, then all the genes in the list are ranked by FC, and finally, an FC cutoff is applied to determine the final set of DEGs (Zhao et al., 2018). If you would like to use more stringent criteria for the enrichment analysis result, please increase the log fold-change values or decrease the p values and vice versa.

Problem 4

File system permission errors occur while installing R packages in section 7.

Potential solution

In some computers or high-performance computing (HPC) systems, your user account may not have the necessary permission to install new packages, we suggest contacting your administrator to install these dependencies.

Specifically, if you have trouble compiling Harmony, please refer to this solution: https://github.com/immunogenomics/harmony/issues/10.

Problem 5

Using different input file format files while running workflow in section 7.

Potential solution

The protocol in section 7 provided examples with CSV file extension. Please change the file reading function from read.csv to the following formats accordingly:

HDF5 file: https://rdrr.io/cran/Seurat/man/Read10X_h5.html
Three gzip files recording information of barcodes (barcodes.tsv), features (genes.tsv), and gene expressions (matrix.mtx): https://rdrr.io/cran/Seurat/man/Read10X.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Qin Ma (qin.ma@osumc.edu).

Materials availability

This study did not generate any new materials.

Data and code availability

scREAD is freely available at https://bmbls.bmi.osumc.edu/scread/. All code for the scREAD protocols is freely available on GitHub: https://github.com/OSU-BMBL/scread-protocol. The interactive tutorial for optional step: calculating overlapping DEGs from multiple comparisons to find overlapping genes is freely available in Google Colab: https://colab.research.google.com/drive/1lInXa6jD4yc7RGJc0EWDfy5NNoXT1qye?usp=sharing.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (R01 GM131399-01, Q.M.; K01 AG056673, H.F.; and R56 AG066782-01, 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. C.W. contributed to data analysis and the initial draft. C.W. developed and implemented the code in the manuscript and Google Colab. Y.X. tested the database and workflow. All authors revised the final manuscript.

DECLARATION OF INTERESTS

The authors declare no competing interests.

REFERENCES

Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I.W.H., Ng, L.G., Ginhoux, F., and Newell, E.W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. Nat. Biotechnol. *37*, 38–44. https://doi.org/10.1038/nbt.4314.

Wilke, C., Fox, S.J., Bates, T., Manalo, K., Lang, B., Barrett, M., Stoiber, M., Philipp, A., Denney, B., Hesselberth, J., et al. (2021). wilkelab/cowplot: 1.1.1. Zenodo. https://doi.org/10.5281/zenodo.4411966.

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 Biol. *16*, 278. https://doi.org/10.1186/s13059-015-0844-5.

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. https://doi.org/10.1038/s41587-019-0332-7.

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. https://doi.org/10.1038/s41593-019-0539-4.

Jiang, J., Wang, C., Qi, R., Fu, H., and Ma, Q. (2020). scREAD: a single-cell RNA-seq database for Alzheimer's disease. iScience *23*, 101769. https://doi.org/10.1016/j.isci.2020.101769.

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.-R., and Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. Nat. Methods *16*, 1289–1296. https://doi.org/10.1038/s41592-019-0619-0.

Kulakovskiy, I.V., Vorontsov, I.E., Yevshin, I.S., Sharipov, R.N., Fedorova, A.D., Rumynskiy, E.I., Medvedeva, Y.A., Magana-Mora, A., Bajic, V.B., Papatsenko, D.A., et al. (2018). HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. Nucleic Acids Res. *46*, D252–D259. https://doi.org/10.1093/nar/gkx1106.

Kuleshov, M.V., Jones, M.R., Rouillard, A.D., Fernandez, N.F., Duan, Q., Wang, Z., Koplev, S., Jenkins, S.L., Jagodnik, K.M., Lachmann, A., et al. (2016). Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res. *44*, W90–W97. https://doi.org/10.1093/nar/gkw377.

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 Res. *48*, W275–W286. https://doi.org/10.1093/nar/gkaa394.

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. https://doi.org/10.1038/s41586-019-1195-2.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell *177*, 1888–1902.e21. https://doi.org/10.1016/j.cell.2019.05.031.

Timmons, J.A., Szkop, K.J., and Gallagher, I.J. (2015). Multiple sources of bias confound functional enrichment analysis of global -omics data. Genome Biol. *16*. https://doi.org/10.1186/s13059-015-0761-7.

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. https://doi.org/10.3390/genes10070531.

Zhao, B., Erwin, A., and Xue, B. (2018). How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances. Genomics *110*, 67–73. https://doi.org/10.1016/j.ygeno.2017.08.007.