LRcell: detecting the source of differential expression at the sub-cell type level from bulk RNA-seq data.

Wenjing Ma (wenjing.ma@emory.edu)[1], Sumeet Sharma (sumeet.sharma@emory.edu)[2], Peng Jin (peng.jin@emory.edu)[3], Shannon L. Gourley (shannon.l.gourley@emory.edu)[4], Zhaohui Qin (zhaohui.qin@emory.edu)[1,5,*]

* Correspondence: zhaohui.qin@emory.edu

[1] Department of Computer Science, Emory University, 400 Dowman Drive, Atlanta, GA 30322, USA

[2] Graduate Program in Neuroscience, Emory University, 1462 Clifton Road NE, Atlanta, GA 30322, USA

[3] Department of Human Genetics, Emory University, 1365 Clifton Road, Atlanta, GA 30322, USA

[4] Department of Pediatrics, School of Medicine, Emory University, 100 Woodruff Circle, Atlanta, GA 30322, USA; Yerkes National Primate Research Center, Atlanta, GA 30322, USA

[5] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, 1518 Clifton Road NE, Atlanta, GA 30322, USA

Abstract

The rapid proliferation of single-cell RNA-sequencing (scRNA-seq) datasets have revealed cell heterogeneity at unprecedented scales. Several deconvolution methods have been developed to decompose bulk experiments to reveal cell type contributions. However, these methods lack power in identifying the accurate cell type composition when having a considerable amount of sub-cell types in the reference dataset. Here, we present *LRcell*, a R Bioconductor package (DOI: https://doi.org/doi:10.18129/B9.bioc.LRcell) aiming to identify specific sub-cell type(s) that drives the changes observed in a bulk RNA-seq differential gene expression experiment. In addition, *LRcell* provides pre-embedded marker genes computed from putative single-cell RNA-seq experiments as options to execute the analyses.

Keywords: Cell-type enrichment, Cell marker genes, Differential gene expression

Background

Finding genes that are differentially expressed (DE) between experimental conditions is a powerful approach to understand the molecular basis of phenotypic variation. However, most tissues are consisted of tens or even hundreds of diverse sub-cell types and DE may only occur in a small subset of these sub-cell types, which are relevant to the experimental condition. Bulk RNA-seq data alone is unable to reveal the sub-cell types that drives the DE.

The rapid development and proliferation of single cell technologies resulted in massive accumulation of single cell transcriptomics data (scRNA-seq) from diverse tissue types. These data reveal substantial variations in transcriptional regulation among different cell types and offer an unprecedented close-up view of the modifications underlying important biological processes including which cell types drive DE [1]. However, steep cost and complicated protocols prevent the wide-spread adoption of scRNA-seq.

In this study, we propose a novel computational tool named *LRcell.* Given the result from a bulk RNA-seq DE study, the goal of *LRcell* is to delineate which sub-cell type(s) of the tissue underwent substantial changes between the two experimental conditions. Exploiting cell type-specific marker genes identified from generic scRNA-seq available from publicly-available data repositories, *LRcell* achieves the goal by surveying the enrichment of marker genes across all sub-cell types in the tissue (Figure 1). Thus, no scRNA-seq experiment matching the bulk RNA-seq experimental condition is needed. When applying *LRcell* to a diverse panel of bulk RNA-seq DE experiments, we successfully identify known cell types involved in the pathogenesis of psychiatric disorders as well as produce testable new hypotheses that have the potential to produce fresh new biological insights.

Results and Discussion

In this work, we collect and curate a compendium of marker genes from multiple published scRNA-seq datasets. We then conduct *LRcell* analysis on multiple bulk RNA-seq DE experiments to demonstrate its utility.

3.1 Marker gene collection and sources

Similar to a collection of gene set for Gene Set Enrichment Analysis (GSEA) [2], *LRcell* requires a compendium of high-quality cell type marker genes. Currently, *LRcell* package provides users with multiple pre-loaded marker gene sets from human blood, human brain and mouse brain (Figure 2.A), computed from scRNA-seq datasets. Additionally, *LRcell* package offers external cell markers collected by Molecular Signatures Database (MSigDB) [3]. The external makers all originate from human species including midbrain, cord blood, ovary and skeletal muscle. We store all cell-type specific marker gene sets into another R Bioconductor ExperimentHub package named *LRcellTypeMarkers*. Additional marker gene sets are being tested and will be added to the collection.

3.2 Microglia highly enriched in Neurodegenerative Dementia

In a recent neurodegenerative dementia study, Swarup and colleagues contrasted TPR50 mice expressing tau mutant with wild type mice using bulk RNA-seq in order to identify gene networks mediating dementia [4]. To identify the cell type(s) most involved in the condition, we apply *LRcell* to the DE gene list using pre-embedded marker genes from adult mouse Frontal Cortex (FC) region [5]. From *LRcell* result, we observed that Microglia show up as highly significant (Figure 2.B) which is concordant with previous studies [6]. Additionally, the FC_11-3.unknown and FC_11-4.unknown sub cell types also show high level of significance. No annotation is available for these two cell clusters in the original publication.

However, pair-wise comparison of marker genes among all cell clusters reveal that these two unknown cell clusters have considerable overlaps with the FC_11-1. Microglia cell type (Figure S1. A), which explains the pattern we observe.

3.3 CD16+ Monocytes highly enriched in Post-Traumatic Stress Disorder (PTSD)

In a recent study, Breen and colleagues conducted a bulk whole-transcriptome study using peripheral blood leukocytes collected from U.S. Marines, among which some developed PTSD post-deployment [7]. Using this dataset, we generate a list of DE genes that show significant difference between the PTSD group and the control group at the pre-deployment timepoint.

Using human marker genes derived from a single cell transcriptomic study on PBMC [8], *LRcell* analysis finds that cells annotated as CD16+ non-classical Monocytes shows up as the most significant among all cell types in PBMC (Figure 2. C). Our finding makes biological sense because as stated in previous studies [9], heterogeneity exists in monocytes distinguished by CD16 surface proteins and non-classical monocytes have been validated to regulate immune responses in trauma [10,11].

3.4 Marker genes from different region or timepoints

To apply *LRcell*, an important question is that which marker gene sets to use, i.e., how to select single cell RNA-seq data where the source of the tissue match the tissue type profiled in the bulk transcriptomic study. This is particularly important for complex tissues such as brain. To address this issue, we use the neurodegenerative dementia study [4] as an example, which contains information from four brain regions: cortex, hippocampus (HC), cerebellum

(CB) and brain stem. Brain stem is excluded from our analysis due to the lack of marker gene information from that region of the brain.

To understand how marker genes vary across brain regions, we first define marker genes in all regions of the brain to explore their spatial pattern (Figure S1. B; Figure S1. C). We observe that glia cells, such as Astrocytes, from different regions have higher number of overlapping marker genes which indicates the homogeneity of glia cells across the brain . In contrast, neurons and interneurons share very few marker genes across different brain regions. We then apply pre-embedded adult mouse brain marker genes from FC, HC and CB to bulk DE genes obtained from cortex, HC and CB respectively (Figure S2. A-H). We observe that Microglia cells are highly enriched in all three brain regions whereas Astrocytes are particularly highly enriched in CB (Figure S2. B, D, H). Especially when applying CB marker genes to CB bulk DE experiment, we notice that one sub-cell type of Astrocytes is highly enriched compared to others. Our observations demonstrate that selected cell types are heterogeneous spatially, meaning marker genes are highly specific not only for the cell type, but also which region the cell belongs to. Because of this finding, it is highly desirable to run *LRcell* using marker genes of cell types located in closely matched brain regions.

We are also curious whether marker genes curated from scRNA-seq experiments conducted on non-normal samples is acceptable as the reference. To address this question, we use data from the HIV vaccine study [8], we observe that the expression of cell-type specific marker genes is mostly consistent across different timepoints within the same cell type (such as CD8 cells), and distinct across different cell types (Figure 2.D; Figure 2.E). For example, B sub-cell types share a considerable number of marker genes across timepoints, while sharing fewer with other cell types. Thus, although default marker genes used in *LRcell* are

collected from control samples, we believe that marker genes identified from non-normal samples are acceptable when scRNA-seq data from normal samples are not available.

3.5 Comparison to GSEA

GSEA is a powerful tool to determine whether a pre-defined gene set show concordant shift in expression when comparing two biological conditions. One could potentially replace *LRcell* with GSEA to identify DE-driving cell types, by treating cell-type specific marker genes as pre-defined gene sets. To compare performance of the two methods, we repeat the analyses done in section 3.1 and 3.2 using GSEA. The GSEA result from the neurodegenerative dementia mouse model (Figure S3A) yields several equally significant cell types including: Astrocyte, Endothelial, Microglia, Mural, Oligodendrocyte and Polydendrocyte. The tied significances lead to difficulties in determining which cell type(s) plays a key role in the dementia. Similar pattern is observed in the GSEA result on PTSD study (Figure S4. A) which shows that Monocytes, Dendritic cells and some T sub-cell types are equally enriched. Based on the above observations, we conclude that *LRcell* is more effective than GSEA to identify sub cell types that are most impacted by the condition change in bulk DE studies.

3.5 Comparison to MuSiC

A plethora of deconvolution methods has been developed to infer the proportions of different sub cell types from bulk transcriptomic data [1][12][13][14][15]. By deconvoluting individual bulk samples and comparing their proportions, one can identify both cell-type proportion and gene expression changes. Hence it is of interest to compare *LRcell* with such a bulk deconvolution strategy. For comparison, we apply MuSiC [13], a deconvolution method, to the mouse neurodegenerative dementia study [4] on both conditions. Because some layers

of neurons are predicted to have almost zero proportion (Figure S3. B) when using all 81 sub-cell types, we curate the original sub-clusters into 15 major cell types in order to achieve a better representation. Despite this, MuSiC does not detect significant difference in Microglia or Astrocyte in terms of their proportions between the two conditions. When applied to the PTSD study, using the original cell cluster annotations, MuSiC shows that most of the T sub-cell types have zero proportion and the proportion of CD14+ Monocytes is up to 60% (Figure S4. B). In contrast, *LRcell* produces more sensible results because it is not limited by the number of cell types as it can detect the subtle differences among sub-cell types.

3.5 Discussions

Detecting transcriptional activity changes at the individual cell type level, especially their modifications in disease samples, is crucial for understanding the mechanisms of diseases development. Deconvolution-based computational tools have been developed to dissect cell type proportions from bulk gene expression profiles. The computationally-inferred proportions can then be used for comparison. However, such a strategy is numerically unstable and unable to handle large number of sub-cell types. In this study, we propose a novel alternative strategy named *LRcell*, *LRcell* conducts enrichment analysis of cell-type specific marker genes among the top (or bottom) DE genes identified by bulk transcriptome studies. Cell types show the most enrichment are likely to play an important role in the condition alteration. When applying to real datasets, we found *LRcell* can successfully identify the involvement of the Microglia and Astrocytes in neurodegenerative dementia and rare Monocytes in PTSD.

In spirit, *LRcell* operates similarly as GSEA. But *LRcell* is much more sensitive to minor differences in marker genes of sub-cell types, which indicates its potential to detect changes

in sub-cell types caused by disease conditions. Additionally, when compared to existing bulk deconvolution methods, *LRcell* is more stable in its ability to handle the similarities among sub-cell types. Thus, *LRcell* enables researchers to glean new biological insights from the bulk transcriptomics experiments with no need of matching single cell experiment.

To enable easy comparison, currently, we select a fix number of 100 marker genes from each sub-cell type. Understandably, the number of marker genes for different cell types varies, it is desirable to allow flexibility in choosing the number of marker genes based on the transcriptomic patterns across cell types. However, different numbers of marker genes post challenge for conducting enrichment analyses fairly across all cell types. This will be investigated in our future studies.

LRcell currently provides embedded marker genes from human blood, human brain and mouse brain calculated from scRNA-seq experiments along with markers from 66 cell types in four tissues adopted from MSigDB. We are working to include more tissue types in the future releases of *LRcell* which will make it more widely applicable.

Conclusions

In summary, we develop *LRcell*, an R Bioconductor package for identifying sub-cell type(s) that drive the changes observed in bulk comparative transcriptomic studies, taking advantages of newly emerged scRNA-seq data. The rationale of *LRcell* is that we believe marker genes of the modifying cell types tend to be enriched toward the top (or bottom) of the DE gene lists. We conduct comprehensive surveys applying *LRcell* across various experimental conditions, and successfully identify cell types that play important roles in neurodegenerative dementia and PTSD. Hence, we believe *LRcell* can provide researchers

important and new biological insights in terms of the source of the biological changes at the sub-cell type level, without the need of conducting costly and laborious scRNA-seq experiments.

Methods

Basic Assumptions

The goal of *LRcell* is to identify the most affected sub-cell type(s) during the transition of experimental conditions using only bulk transcriptomic data. Based on the assumptions that cell-type specific marker genes of key cell types tend to be over-represented among the significant DE genes in bulk transcriptomic studies, *LRcell* is able to discover which cell type(s) is involved in certain disease or condition change. In recent years, computational methods have been developed to deconvolve bulk RNA-seq data to delineate cell-type proportion changes, which could be borrowed to answer the same question. However, whenever there are ten or more sub-cell types, the results from deconvolution methods become unreliable. In contrast, *LRcell* enables comparison across many more cell types which is important for complex tissues such as brain.

The development of *LRcell* is inspired by LRpath [16], which is designed for linking experimental changes to biological pathways or a pre-defined gene set. In *LRcell*, we utilize cell-type specific marker genes as gene sets and calculate the enrichment of each cell type when comparing two biological conditions. We believe that the most enriched cell type(s) is highly likely to play an important role in the experimental condition change.

scRNA-seq Data Preprocessing

In this study, we include marker genes from mouse whole brain, human prefrontal cortex and human PBMC, along with 66 cell types' markers from four tissues adopted from MSigDB.

For each scRNA-seq dataset, we first retrieved raw read count matrix. Next, we filtered out low-quality cells and genes and apply column-wise normalization and log transformation on the data.

The mouse whole brain scRNA-seq dataset [5] produced using the Drop-seq technology [17], contains nine brain regions from adult mice. The data provided has already been pre-filtered by the authors. For cell types other than neurons, we directly utilized the information provided on the study website. For neurons and interneurons, we curated the sub-cell types following the original study.

The human prefrontal cortex scRNA-seq dataset [18], produced by 10X Genomics Chromium, is derived from the prefrontal cortex (pFC) region (specifically BA9). The dataset contains two conditions: healthy controls and major depressive disorder (MDD). We split the data matrix into two parts and filtered out cells expressing less than 10 genes and genes expressed in less than 10 cells respectively. We also filtered out mitochondrial, ribosomal genes, and genes from annotation clusters (Astros_1, Mix_1, Mix_2, Mix_3, Mix_4, Mix_5, Inhib_4_SST).

The PBMC dataset [8], generated by CITE-seq technology [19], is derived from an HIV vaccine trial study which involves eight volunteers at three timepoints: immediately before, three days and seven days after the vaccine. The study contains 161,764 cells in total. To accelerate the marker gene selection, we separated the count matrix according to the time label and filtered out low-quality cells and genes (mitochondrial, ribosomal genes and those expressed in less than 1,000 cells). The cluster annotated as “Doublet” was filtered out.

Marker Gene Selection

After obtaining the log-normalized gene expression matrix along with high-quality sub-cell type clusters, we calculated the enrichment scores for each cluster (sub-cell type) using the

marker gene selection method described in Marques et al[20]. The cluster-specific gene enrichment is defined as the average gene expression levels of cells in that cluster divided by the average gene expression levels in all cells. The enrichment score is adjusted by introducing a penalty representing the fraction of cells in that cluster expressing the marker gene. Combined together, this score allows the identification of genes with high, cluster-specific expression to be selected as marker genes. The description below is adapted from the original publication.

Suppose there are a total of M genes, L different clusters each with N_j cells, and the total number of cells is N. Let $E = \{E_{ijk}\}$ represent the gene by cell read count matrix. Here $i = 1, \dots, M, j = 1, \dots, L, k = 1, \dots, N_j$ and $N = \sum_{j=1}^{L} N_j$. The overall average expression of the ith gene across all cells is defined as:

$$\bar{E}_{i..} = \frac{1}{N} \sum_{j=1}^{L} \sum_{k=1}^{N_j} E_{ijk}$$

The average expression of gene *i* in the *j*th cluster is defined as:

$$\bar{E}_{ij.} = \frac{1}{N_j} \sum_{k=1}^{N_j} E_{ijk}$$

The enrichment for gene *i* in the *j*th cluster as:

$$Enrichment_{i,j} = \frac{\bar{E}_{ij.}}{\bar{E}_{i.}}$$

Next, we consider the proportion expressing the gene *i* in the *j*th cluster as:

$$Prop_{i,j} = \frac{1}{N_j} \sum_{k=1}^{N_j} \mathrm{I}(E_{ijk} > 0)$$

The $I(\cdot)$ is an indicator function.

The enrichment score for gene *i* in the *j*th cluster is computed as:

$$Score_{i,j} = Enrichment_{i,j} \times (Prop_{i,j})^{power}$$

where “power” is a hyperparameter to be tuned manually to control the penalization for the cell cluster proportion term. The power parameter is set to 1 throughout this study. After calculating the weighed gene enrichment scores in each cluster, we ranked genes based on the scores and selected the top 100 genes as the marker genes for each cluster.

MSigDB Marker Genes

We downloaded cell marker gene sets from MSigDB category C8 – cell type signature gene sets. Since not all tissue types are suitable for *LRcell*, we applied the following criteria to select tissues: 1), non-fetal tissues; , 2), have more than 8 (sub-) cell types; 3), minimum number of marker gene greater than 50; 4), median number of marker genes greater than 80. In the end, four tissue types, the midbrain, cord blood, ovary and skeletal muscle remain.

Bulk RNA-seq Data Preprocessing

The raw count of mouse bulk RNA-seq study on AD was downloaded from GEO (Accession number: GSE90693). DE analysis is performed using DESeq2 [21] to obtain DEGs in each brain region.

The raw count of bulk RNA-seq study on PTSD was downloaded from Recount2 [22]. We extracted out the experiment contrasting PTSD cases and healthy controls with timepoint of pre-employment and performed DESeq2 to obtain DEGs.

LRcell Analysis

LRcell is inspired by LRpath, which was designed for identifying sets of predefined gene sets that show enrichment with differentially expressed transcripts in microarray experiments. *LRcell* uses logistic regression or linear regression to assess whether marker genes (as defined in the **Marker Gene Selection** sub-section above) of a specific cell type are more likely to be

DE genes in a particular comparison study. The linear regression option is added in order to handle the continuous enrichment scores of marker genes. Users can choose accordingly. To facilitate our analysis, we assume that the major sub cell types of the tissue their marker genes are known *a priori*.

We apply LRcell to each cell type independently. The required input includes a list of DE genes ranked by the level of significance, and a set of marker genes for each cell type. Then *LRcell* runs a logistic regression as

$$log\frac{\theta}{1-\theta} = \alpha + \beta x$$

and

$$\theta = P(Y = 1)$$

In which $Y = 1$ denotes that gene is a marker gene and $Y = 0$ otherwise. Hence θ represents the chance that the gene is a marker gene. We use $-\log(p - \text{value})$ as the explanatory variable x. Whether a specific cell type is involved in the experimental condition change is evaluated by testing the null hypothesis that $\beta = 0$ against the alternative that $\beta \neq 0$ using the Wald test. Typically, we run *LRcell* on all sub cell types found in the tissue to see which sub cell type(s) drives the changes.

Similar to logistic regression, linear regression directly performs

$$Y = \alpha + \beta x$$

where Y indicates the enrichment scores of genes. Same as logistic regression, the p-value can be obtained from testing the null hypothesis that $\beta = 0$ against the alternative that $\beta \neq 0$ using the t-test.

Once the p-values are obtained, we calculate False Discover Rate (FDR) using p.adjust() function in R to adjust p-values with Benhamini-Hochberg method.

Input and Output

LRcell requires two inputs: (1) a ranked list of genes with differential expression (DE) p-values in a bulk RNA-seq experiment; (2) sets of marker genes from all sub-cell types of the bulk tissue acquired from scRNA-seq datasets *a priori* or from MSigDB C8—cell type signature gene sets. For those cell markers derived from scRNA-seq datasets, we offer choices for users to choose between species as human or mouse and the region indicates the specific brain region or PBMC. For MSigDB cell markers, we store the marker genes into the *LRcellTypeMarkers* packages which can be easily downloaded. When running LRcell() function, the Logistic Regression (LR) option is set as the default, while users can also set the method option as LiR if Linear Regression is desired. For linear regression, an enrichment score is needed as input for each gene whereas gene sets are sufficient for logistic regression. For MSigDB cell type signature gene set, LR option is recommended as there is no enrichment score information available. For customized input, i.e., a scRNA-seq data, we offer a LRcell_gene_enriched_scores function which takes the read counts matrix and cell annotation as input to generate enrichment scores for genes in each cell type. For further subsetting, get_markergenes can be used for generating marker genes for more specific sub cell types.

The output is a list of significance p-value (or FDR), one for each sub cell type. For visualization, *LRcell* produces Manhattan plot, which can be drawn through plot_manhattan_enrich function. We also provided a plot (plot_marker_dist function) indicating where certain cell-type specific marker genes locate on the bulk DE genes. The bulk DE genes are better sorted using -log10(p-value) × sign(log2FoldChange) which could potentially give information on both up-/down-regulated directions. More detailed information about *LRcell* is available at http://bioconductor.org/packages/release/bioc/html/LRcell.html.

LRcell requires an R version beyond 4.1 and a prerequisite installation of BiocManager.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The datasets analyzed during the current study are available in the Gene Expression Omnibus (GEO) with following accession numbers: mouse whole brain [5] (GSE116470), human prefrontal cortex [18] (GSE144136), human PBMC [8] (GSE164378), neurodegenerative dementia study [4] (GSE90696) and PTSD study [7] (GSE64814). The R package is freely available on Bioconductor (https://doi.org/doi:10.18129/B9.bioc.LRcell) and the external marker genes are stored in another R package named *LRcellTypeMarkers* on Bioconductor (https://doi.org/doi:10.18129/B9.bioc.LRcellTypeMarkers).

Competing interests

The authors declare that they have no competing interests.

Author's contributions

ZQ and WM conceived the idea and supervised the study. WM realized the method and developed the R package. WM analyzed the real data and conducted the computational experiments. SS provided guidance on validating biological experiments and tested the package. WM and ZQ summarized the results and wrote the manuscript. All authors have read and approved the manuscript.

References

1. Newman AM, Steen CB, Liu CL, Gentles AJ, Chaudhuri AA, Scherer F, et al. Determining cell type abundance and expression from bulk tissues with digital cytometry. Nat Biotechnol. 2019;37:773–82.

2. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS. 2005;102:15545–50.

3. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst. 2015;1:417–25.

4. Swarup V, Hinz FI, Rexach JE, Noguchi K, Toyoshiba H, Oda A, et al. Identification of evolutionarily conserved gene networks mediating neurodegenerative dementia. Nature medicine. Nature Publishing Group; 2019;25:152–64.

5. Saunders A, Macosko EZ, Wysoker A, Goldman M, Krienen FM, Rivera H de, et al. Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain. Cell. 2018;174:1015-1030.e16.

6. Perry VH, Nicoll JA, Holmes C. Microglia in neurodegenerative disease. Nature Reviews Neurology. Nature Publishing Group; 2010;6:193.

7. Breen MS, Maihofer AX, Glatt SJ, Tylee DS, Chandler SD, Tsuang MT, et al. Gene networks specific for innate immunity define post-traumatic stress disorder. Molecular psychiatry. Nature Publishing Group; 2015;20:1538–45.

8. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184:3573-3587.e29.

9. Ong S-M, Teng K, Newell E, Chen H, Chen J, Loy T, et al. A novel, five-marker alternative to CD16–CD14 gating to identify the three human monocyte subsets. Frontiers in immunology. Frontiers; 2019;10:1761.

10. Kratofil RM, Kubes P, Deniset JF. Monocyte Conversion During Inflammation and Injury. Arterioscler Thromb Vasc Biol. 2017;37:35–42.

11. Kuan P-F, Yang X, Clouston S, Ren X, Kotov R, Waszczuk M, et al. Cell type-specific gene expression patterns associated with posttraumatic stress disorder in World Trade Center responders. Transl Psychiatry. 2019;9:1.

12. Gong T, Szustakowski JD. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. Bioinformatics. 2013;29:1083–5.

13. Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat Commun. 2019;10:1–9.

14. Baron M, Veres A, Wolock SL, Faust AL, Gaujoux R, Vetere A, et al. A single-cell transcriptomic map of the human and mouse pancreas reveals inter-and intra-cell population structure. Cell systems. Elsevier; 2016;3:346-360. e4.

15. Tsoucas D, Dong R, Chen H, Zhu Q, Guo G, Yuan G-C. Accurate estimation of cell-type composition from gene expression data. Nat Commun. 2019;10:1–9.

16. Sartor MA, Leikauf GD, Medvedovic M. LRpath: a logistic regression approach for identifying enriched biological groups in gene expression data. Bioinformatics. 2009;25:211–7.

17. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. Elsevier; 2015;161:1202–14.

18. Nagy C, Maitra M, Tanti A, Suderman M, Théroux J-F, Davoli MA, et al. Single-nucleus transcriptomics of the prefrontal cortex in major depressive disorder implicates oligodendrocyte precursor cells and excitatory neurons. Nature neuroscience. Nature Publishing Group; 2020;23:771–81.

19. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nature Methods. 2017;14:865–8.

20. Marques S, Zeisel A, Codeluppi S, van Bruggen D, Falcão AM, Xiao L, et al. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science. American Association for the Advancement of Science; 2016;352:1326–9.

21. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

22. Collado-Torres L, Nellore A, Kammers K, Ellis SE, Taub MA, Hansen KD, et al. Reproducible RNA-seq analysis using recount2. Nat Biotechnol. 2017;35:319–21.

Figures

Figure 1 Overview of *LRcell* workflow. (A) describes a bulk RNA-seq experiment design with control and disease conditions. (B) describes a marker gene list derived from scRNA-seq experiment through LRcell or prior knowledge. (C) (upper panel) illustrates how the distribution of marker genes on bulk DE genes may affect the regression; (lower panel) visualizes the FDR rate for each cell type in which the grey cells and dark blue cells are most likely involved in the condition change.

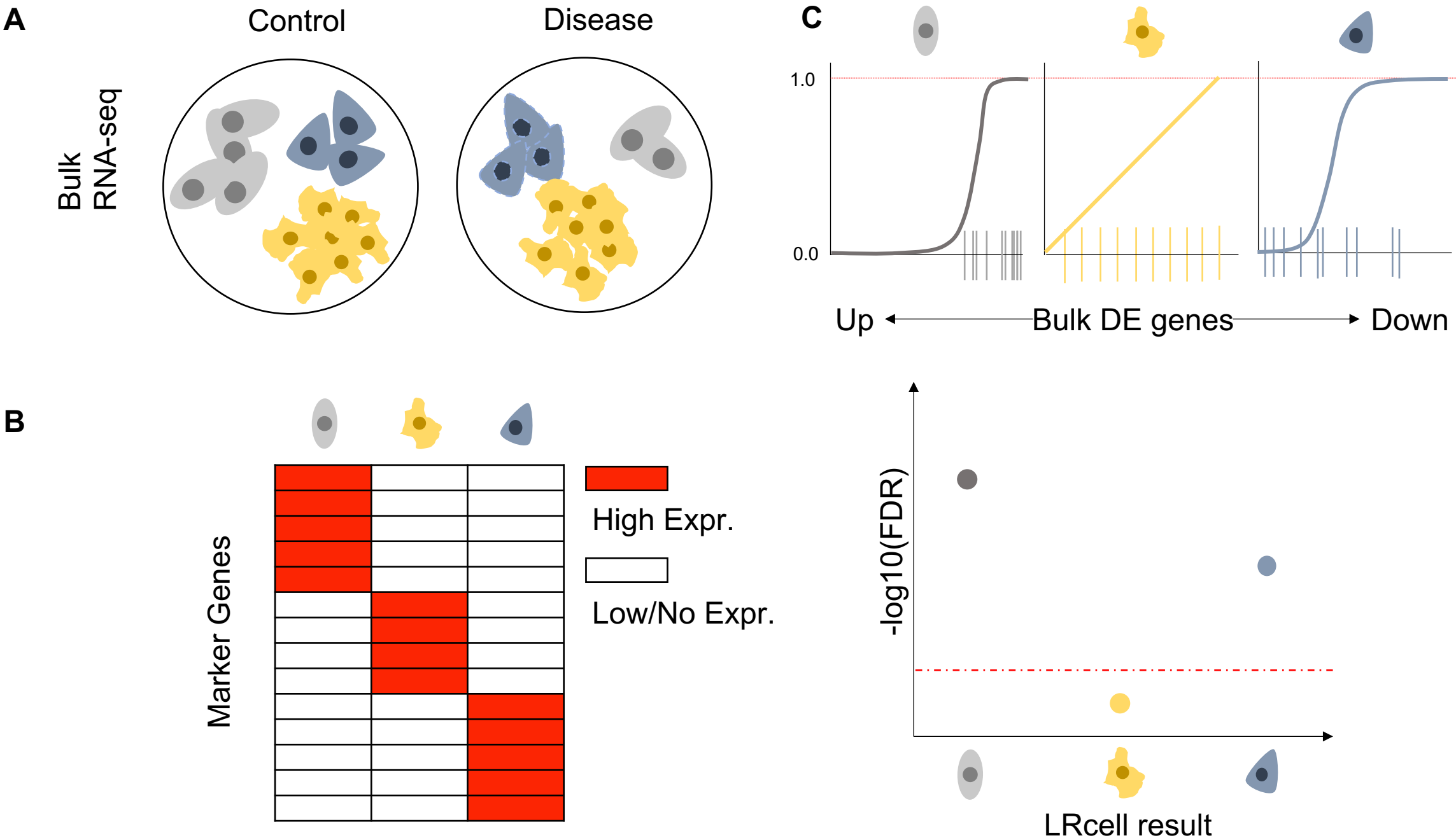
Figure 2 *LRcell* datasets and *applying LRcell* to real cases. (A) demonstrates the marker genes included in *LRcell*. (B) visualzes the *LRcell* result of applying mouse brain FC marker genes to FC region from bulk neurodegenerative dementia study. Microglia is shown as the most significant cell type. (C) visualizes the *LRcell* result of applying human PBMC marker genes to bulk PTSD study. CD16+ Monocytes is shown as the most significant cell type. (D) (E) show the cell-type-specific marker genes overlap between different time points derived from human PBMC. Top 100 marker genes for each cell type are selected.

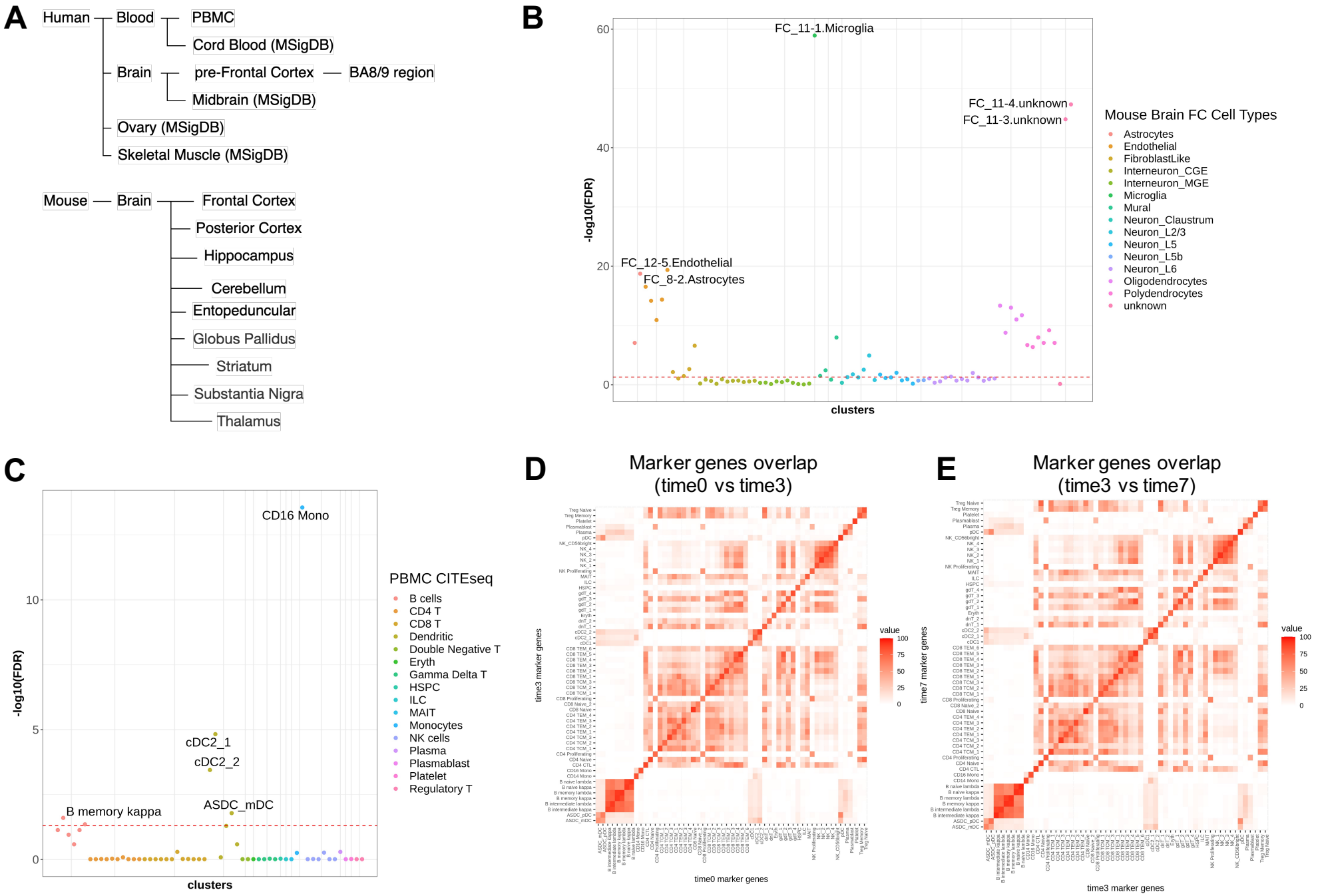
Figure S1: Marker genes overlap between different brain regions. Top 100 marker genes are selected for each cell type, thus, the maximum overlap in figures is 100. (A) describes the marker genes overlap within Frontal Cortex (FC) region derived from mouse whole brain scRNA-seq dataset. The highlighted area describes the overlap between FC_11-3.unknown, FC_11-4.unknwon and FC_11-1.Microglia as an illustration for the similarity between these three sub-cell types. (B) describes the marker genes overlap between FC and Cerebellum (CB). (C) describes the marker genes overlap between FC and Hippocampus (HC). As shown in (B)(C), the overlaps are lower than in (A).

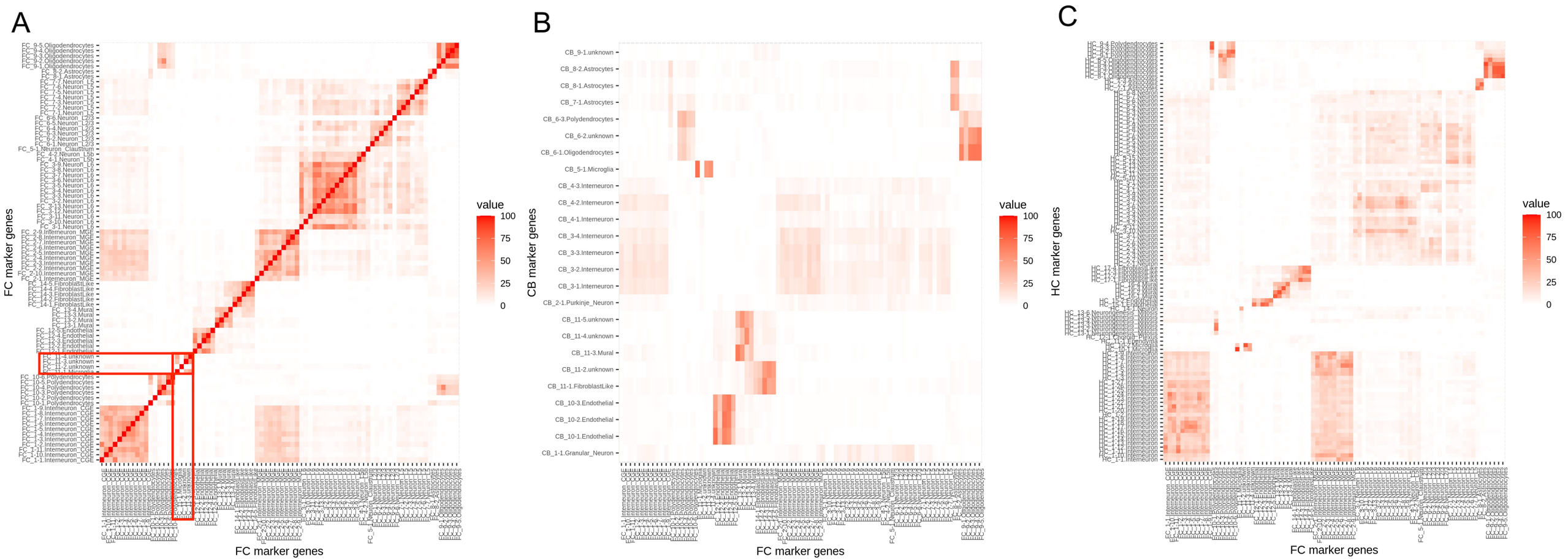
Figure S2. *LRcell* analysis on different combinations of marker genes and bulk DE genes. (A) applies FC marker genes on HC; (B) applies FC markers on CB; (C) applies HC markers on HC; (D) applies HC markers on CB; (E) applies HC markers on FC; (F) applies CB markers on FC; (G) applies CB markers on HC; (H) applies CB markers on CB. Compared to (B) and (D), which analyze CB bulk DE genes, (H) has a higher resolution as the FDR ranges from 0 to 12.

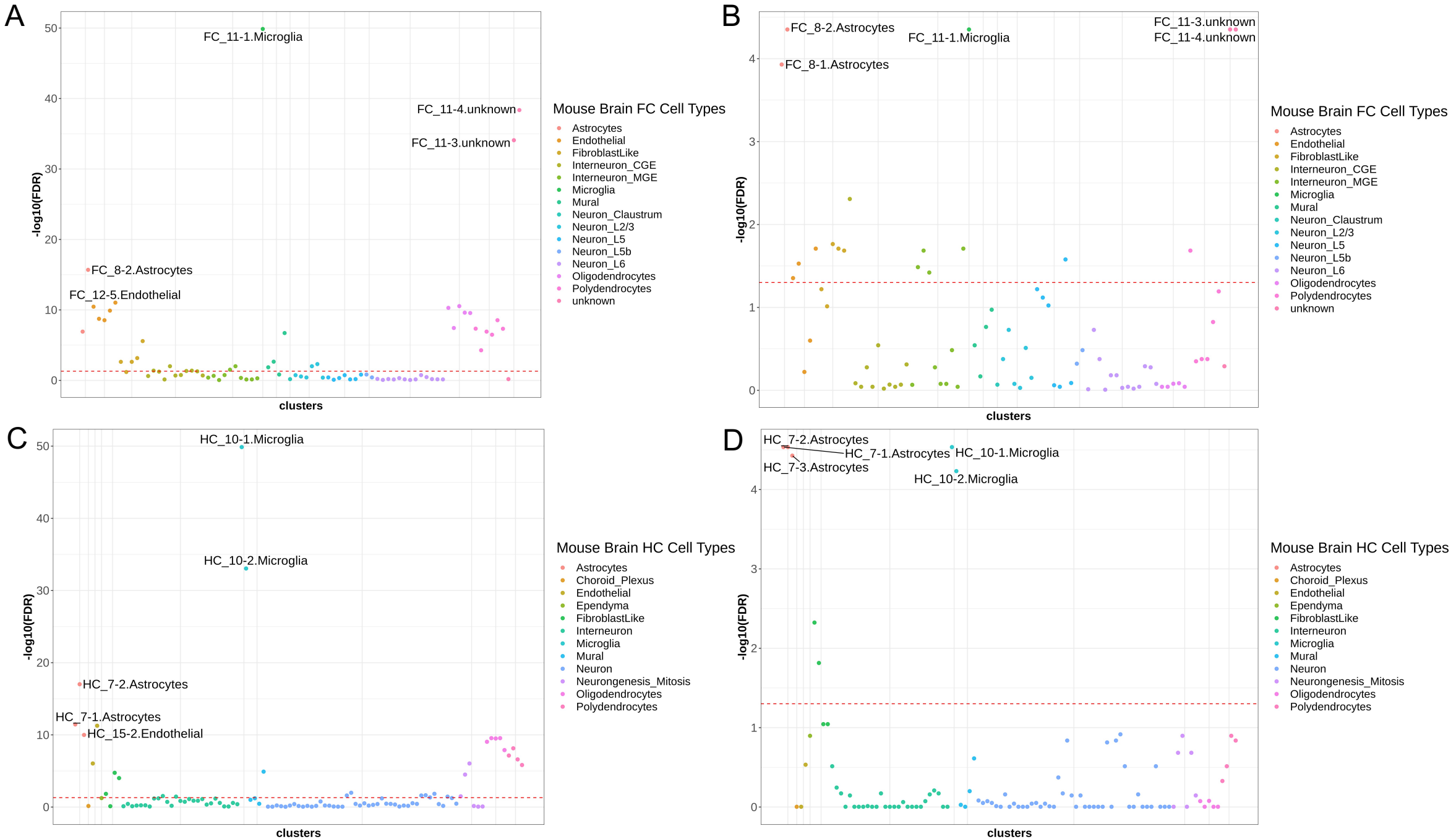
Figure S3: GSEA and MuSiC comparison on neurodegenerative dementia study. (A) shows the GSEA enrichment when using the same marker genes for LRcell (mouse brain FC) as input. Several dots are sharing the same significances which is difficult to interpret. (B) shows the proportions calculated from MuSiC between control and disease. Each box contains 17 individuals. The x-axis is ordered by the t-test significance between two conditions.

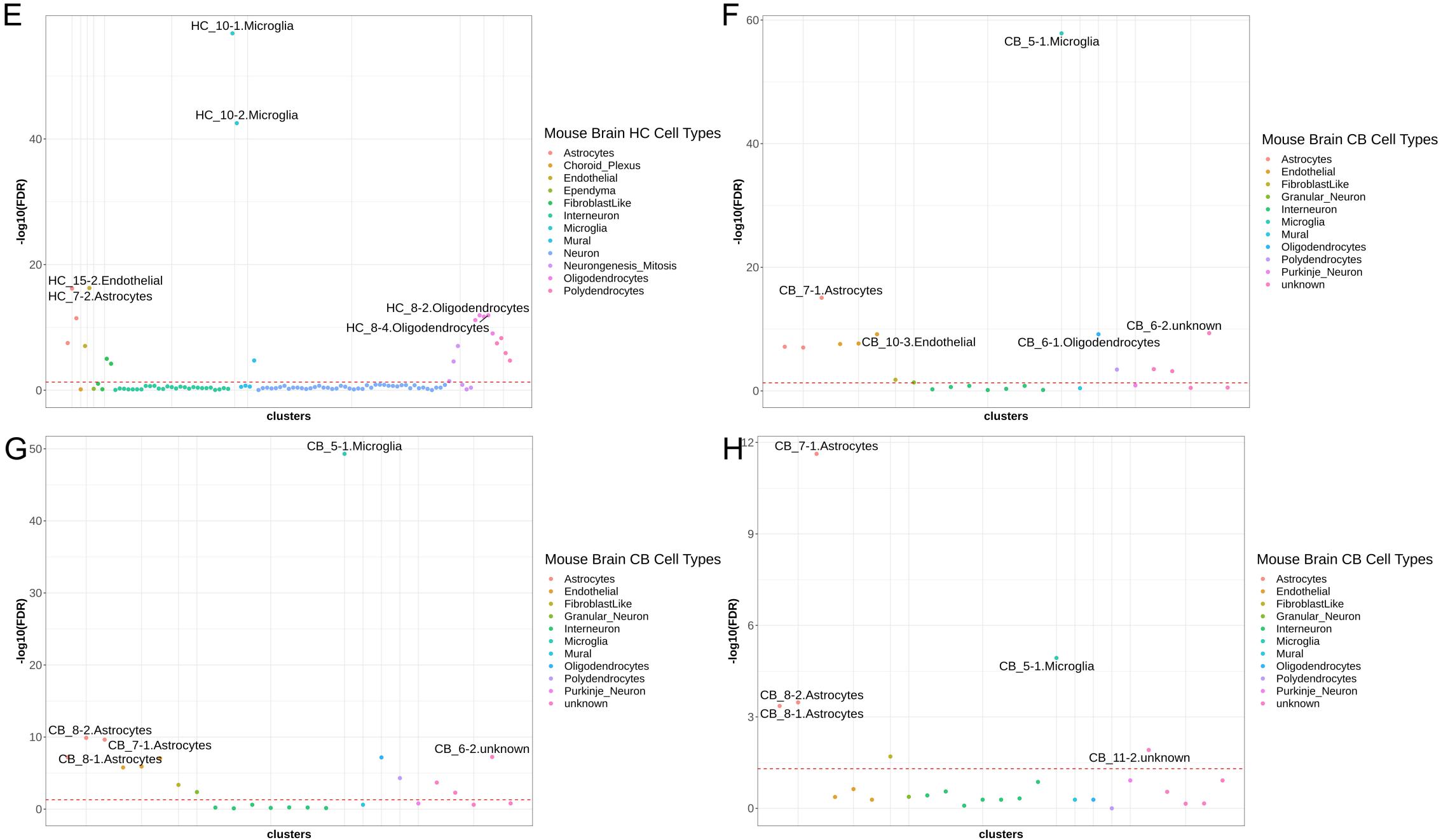
Figure S4: GSEA and MuSiC comparison on PTSD study. (A) shows the GSEA enrichment when using the same marker genes for LRcell (human PBMC) as input. Similarly, several dots are sharing the same significances. (B) shows the proportions calculated from MuSiC between control and disease. The x-axis is ordered by the t-test significance between two conditions. As shown in the figures, MuSiC failed to identify all cell type proportions as there are plenty of 0s starting from ASCD_pDC cell type and the CD14+ Monocytes are wrongly inferred.

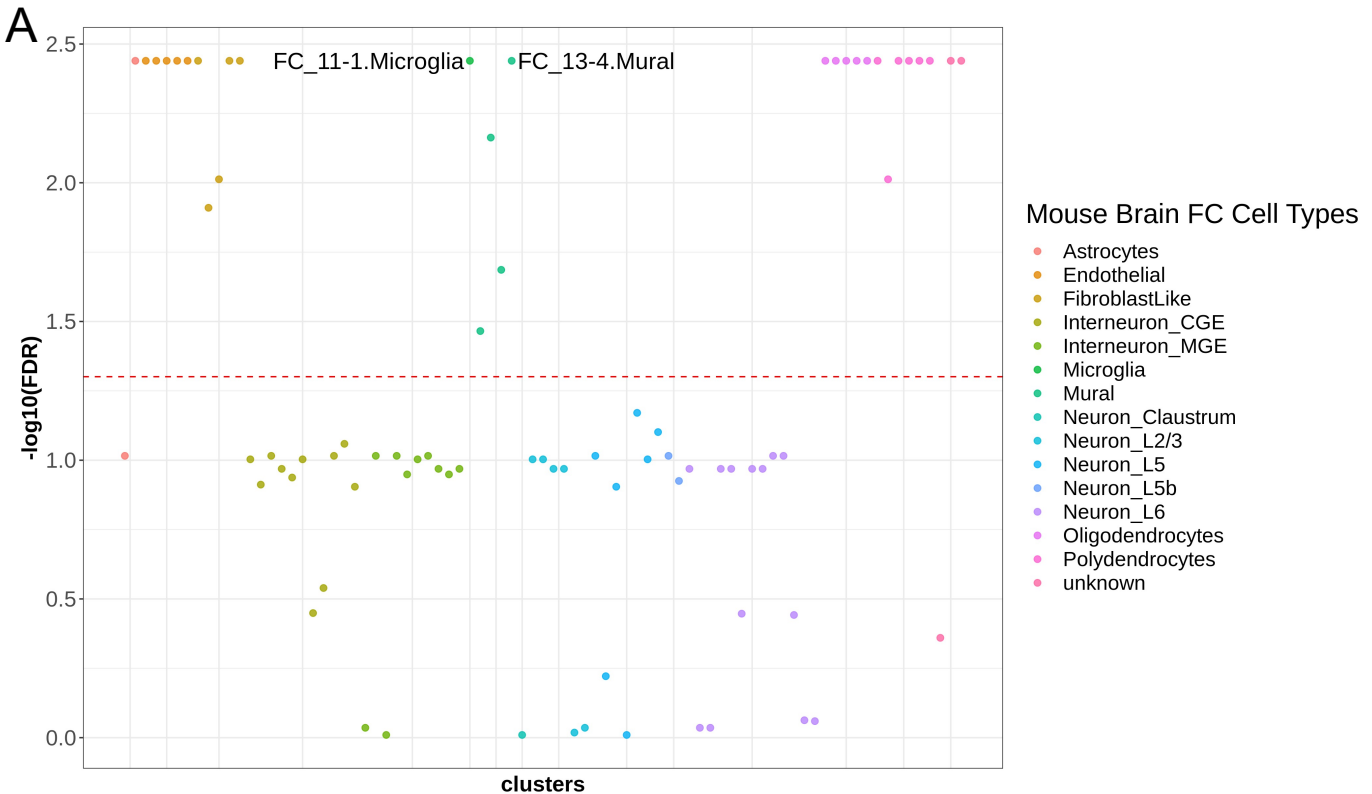

A
Control
Disease
Bulk
RNA-seq
C
1.0
0.0
Up
Bulk DE genes
Down
B
Marker Genes
High Expr.
Low/No Expr.
-log10(FDR)
LRcell result

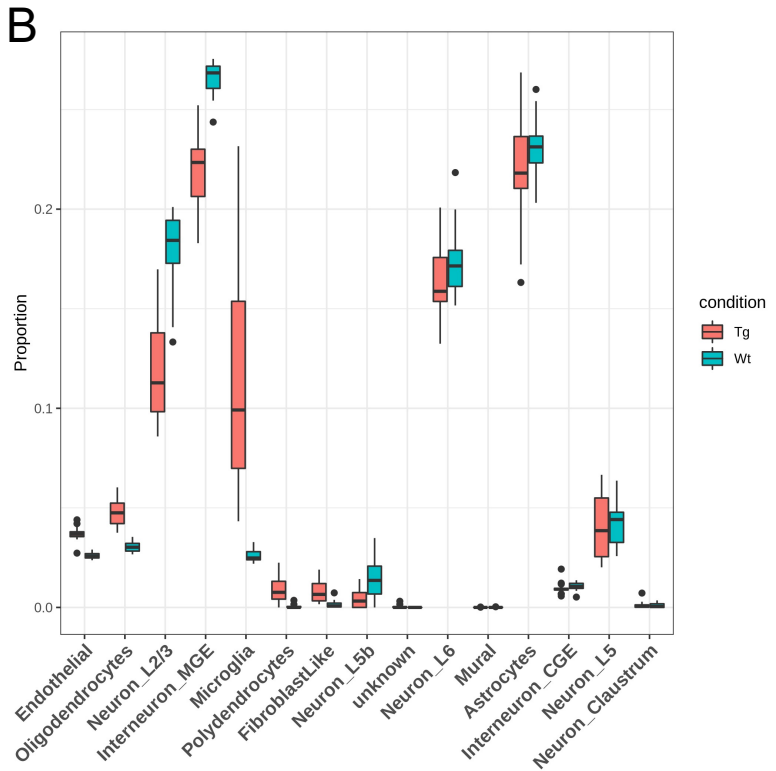

A
Human
Blood
PBMC
Cord Blood (MSigDB)
Brain
pre-Frontal Cortex
BA8/9 region
Midbrain (MSigDB)
Ovary (MSigDB)
Skeletal Muscle (MSigDB)
Mouse
Brain
Frontal Cortex
Posterior Cortex
Hippocampus
Cerebellum
Entopeduncular
Globus Pallidus
Striatum
Substantia Nigra
Thalamus
B
FC_11-1.Microglia
FC_11-4.unknown
FC_11-3.unknown
FC_12-5.Endothelial
FC_8-2.Astrocytes
-log10(FDR)
clusters
Mouse Brain FC Cell Types
Astrocytes
Endothelial
FibroblastLike
Interneuron_CGE
Interneuron_MGE
Microglia
Mural
Neuron_Claustrum
Neuron_L2/3
Neuron_L5
Neuron_L5b
Neuron_L6
Oligodendrocytes
Polydendrocytes
unknown
C
CD16 Mono
cDC2_1
cDC2_2
ASDC_mDC
B memory kappa
-log10(FDR)
clusters
PBMC CITEseq
B cells
CD4 T
CD8 T
Dendritic
Double Negative T
Eryth
Gamma Delta T
HSPC
ILC
MAIT
Monocytes
NK cells
Plasma
Plasmablast
Platelet
Regulatory T
D
Marker genes overlap
(time0 vs time3)
time3 marker genes
time0 marker genes
value
E
Marker genes overlap
(time3 vs time7)
time7 marker genes
time3 marker genes
value

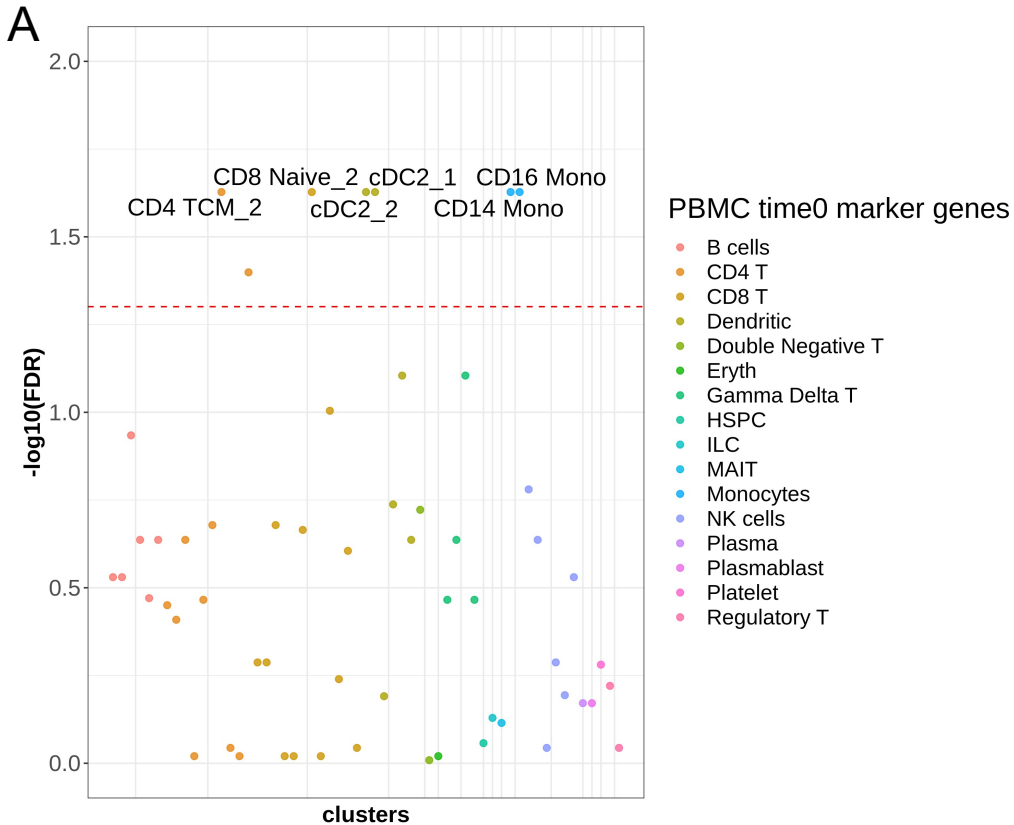

A
FC marker genes
FC marker genes
value
100
75
50
25
0
B
CB marker genes
FC marker genes
value
100
75
50
25
0
C
HC marker genes
FC marker genes
value
100
75
50
25
0

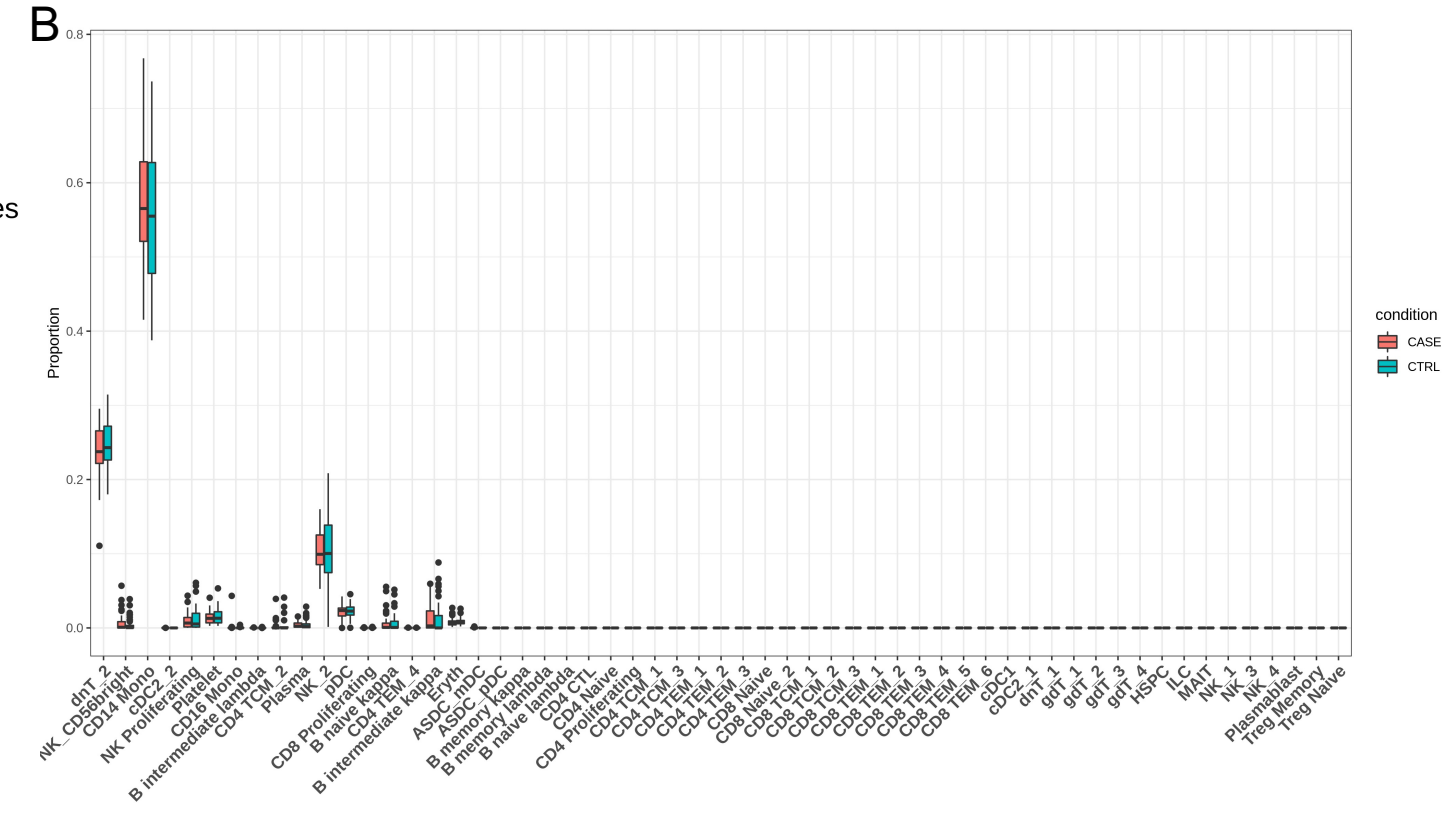

A
FC_11-1.Microglia
FC_11-4.unknown
FC_11-3.unknown
FC_8-2.Astrocytes
FC_12-5.Endothelial
-log10(FDR)
clusters
Mouse Brain FC Cell Types
Astrocytes
Endothelial
FibroblastLike
Interneuron_CGE
Interneuron_MGE
Microglia
Mural
Neuron_Claustrum
Neuron_L2/3
Neuron_L5
Neuron_L5b
Neuron_L6
Oligodendrocytes
Polydendrocytes
unknown
B
FC_8-2.Astrocytes
FC_11-1.Microglia
FC_11-3.unknown
FC_11-4.unknown
FC_8-1.Astrocytes
-log10(FDR)
clusters
Mouse Brain FC Cell Types
Astrocytes
Endothelial
FibroblastLike
Interneuron_CGE
Interneuron_MGE
Microglia
Mural
Neuron_Claustrum
Neuron_L2/3
Neuron_L5
Neuron_L5b
Neuron_L6
Oligodendrocytes
Polydendrocytes
unknown
C
HC_10-1.Microglia
HC_10-2.Microglia
HC_7-2.Astrocytes
HC_7-1.Astrocytes
HC_15-2.Endothelial
-log10(FDR)
clusters
Mouse Brain HC Cell Types
Astrocytes
Choroid_Plexus
Endothelial
Ependyma
FibroblastLike
Interneuron
Microglia
Mural
Neuron
Neurongenesis_Mitosis
Oligodendrocytes
Polydendrocytes
D
HC_7-2.Astrocytes
HC_7-1.Astrocytes
HC_10-1.Microglia
HC_7-3.Astrocytes
HC_10-2.Microglia
-log10(FDR)
clusters
Mouse Brain HC Cell Types
Astrocytes
Choroid_Plexus
Endothelial
Ependyma
FibroblastLike
Interneuron
Microglia
Mural
Neuron
Neurongenesis_Mitosis
Oligodendrocytes
Polydendrocytes


E
HC_10-1.Microglia
HC_10-2.Microglia
HC_15-2.Endothelial
HC_7-2.Astrocytes
HC_8-2.Oligodendrocytes
HC_8-4.Oligodendrocytes
-log10(FDR)
clusters
Mouse Brain HC Cell Types
Astrocytes
Choroid_Plexus
Endothelial
Ependyma
FibroblastLike
Interneuron
Microglia
Mural
Neuron
Neurongenesis_Mitosis
Oligodendrocytes
Polydendrocytes
F
CB_5-1.Microglia
CB_7-1.Astrocytes
CB_10-3.Endothelial
CB_6-1.Oligodendrocytes
CB_6-2.unknown
-log10(FDR)
clusters
Mouse Brain CB Cell Types
Astrocytes
Endothelial
FibroblastLike
Granular_Neuron
Interneuron
Microglia
Mural
Oligodendrocytes
Polydendrocytes
Purkinje_Neuron
unknown
G
CB_5-1.Microglia
CB_8-2.Astrocytes
CB_7-1.Astrocytes
CB_8-1.Astrocytes
CB_6-2.unknown
-log10(FDR)
clusters
Mouse Brain CB Cell Types
Astrocytes
Endothelial
FibroblastLike
Granular_Neuron
Interneuron
Microglia
Mural
Oligodendrocytes
Polydendrocytes
Purkinje_Neuron
unknown
H
CB_7-1.Astrocytes
CB_5-1.Microglia
CB_8-2.Astrocytes
CB_8-1.Astrocytes
CB_11-2.unknown
-log10(FDR)
clusters
Mouse Brain CB Cell Types
Astrocytes
Endothelial
FibroblastLike
Granular_Neuron
Interneuron
Microglia
Mural
Oligodendrocytes
Polydendrocytes
Purkinje_Neuron
unknown


A
FC_11-1.Microglia
FC_13-4.Mural
-log10(FDR)
clusters
2.5
2.0
1.5
1.0
0.5
0.0
Mouse Brain FC Cell Types
Astrocytes
Endothelial
FibroblastLike
Interneuron_CGE
Interneuron_MGE
Microglia
Mural
Neuron_Claustrum
Neuron_L2/3
Neuron_L5
Neuron_L5b
Neuron_L6
Oligodendrocytes
Polydendrocytes
unknown


B
Proportion
0.2
0.1
0.0
condition
Tg
Wt
Endothelial
Oligodendrocytes
Neuron_L2/3
Interneuron_MGE
Microglia
Polydendrocytes
FibroblastLike
Neuron_L5b
unknown
Neuron_L6
Mural
Astrocytes
Interneuron_CGE
Neuron_L5
Neuron_Claustrum


A
CD8 Naive_2
cDC2_1
CD16 Mono
CD4 TCM_2
cDC2_2
CD14 Mono
-log10(FDR)
clusters
PBMC time0 marker genes
B cells
CD4 T
CD8 T
Dendritic
Double Negative T
Eryth
Gamma Delta T
HSPC
ILC
MAIT
Monocytes
NK cells
Plasma
Plasmablast
Platelet
Regulatory T


B
Proportion
condition
CASE
CTRL