STAR Protocols

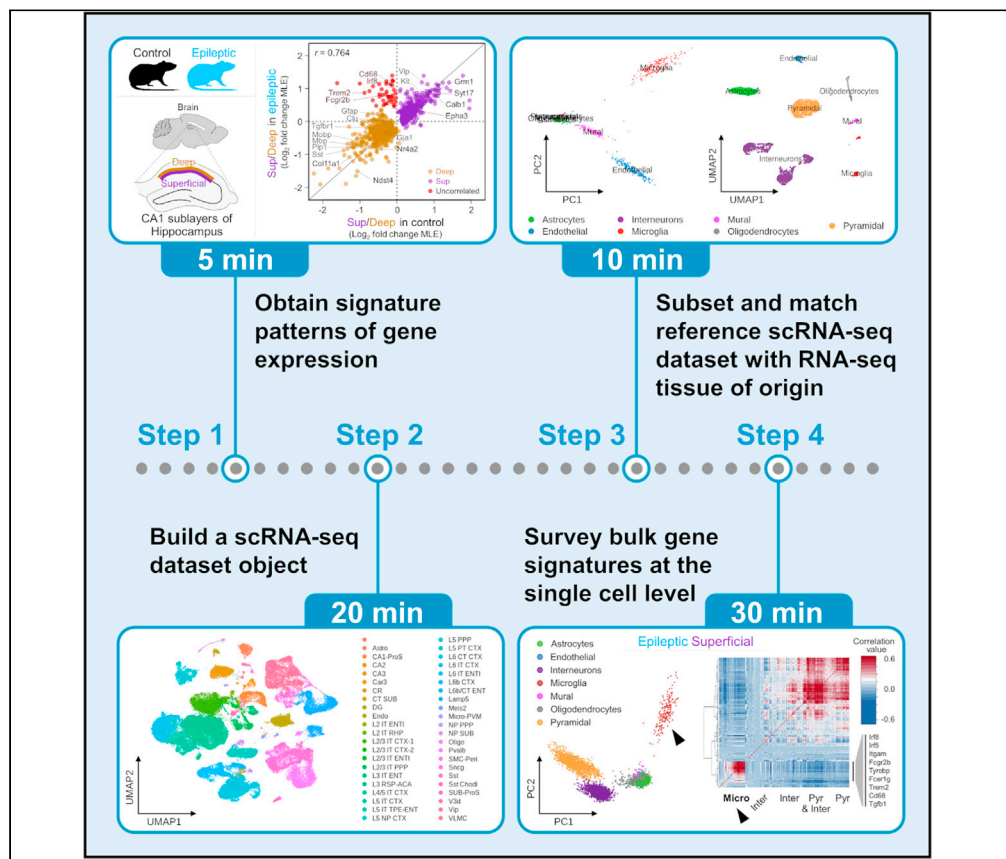**Protocol**

A protocol to extract cell-type-specific signatures from differentially expressed genes in bulk-tissue RNA-seq

Angel
Marquez-Galera,
Liset M. de la Prida,
Jose P.
Lopez-Atalaya

a.marquez@csic.es (A.M.-
G.)
jose.lopezatalaya@csic.es
(J.P.L.-A.)

Highlights

A protocol to explore gene signatures from bulk RNA-seq at the cell-type-specific level

Deconvolution of complex gene signatures from highly heterogeneous tissues

Publicly available single-cell gene expression dataset is retrieved and curated

Gene signatures across brain regions and disease states are surveyed in scRNA-seq data

Bulk-tissue RNA-seq is widely used to dissect variation in gene expression levels across tissues and under different experimental conditions. Here, we introduce a protocol that leverages existing single-cell expression data to deconvolve patterns of cell-type-specific gene expression in differentially expressed gene lists from highly heterogeneous tissue. We apply this protocol to interrogate cell-type-specific gene expression and variation in cell type composition between the distinct sublayers of the hippocampal CA1 region of the brain in a rodent model of epilepsy.

Protocol

A protocol to extract cell-type-specific signatures from differentially expressed genes in bulk-tissue RNA-seq

Angel Marquez-Galera,[1,3,*] Liset M. de la Prida,[2] and Jose P. Lopez-Atalaya[1,4,*]

[1]Instituto de Neurociencias, Consejo Superior de Investigaciones Científicas - Universidad Miguel Hernández, Sant Joan d'Alacant, 03550 Alacant, Spain

[2]Instituto Cajal, Consejo Superior de Investigaciones Científicas, 28002 Madrid, Spain

[3]Technical contact

[4]Lead contact

*Correspondence: a.marquez@csic.es (A.M.-G.), jose.lopezatalaya@csic.es (J.P.L.-A.)
https://doi.org/10.1016/j.xpro.2022.101121

SUMMARY

Bulk-tissue RNA-seq is widely used to dissect variation in gene expression levels across tissues and under different experimental conditions. Here, we introduce a protocol that leverages existing single-cell expression data to deconvolve patterns of cell-type-specific gene expression in differentially expressed gene lists from highly heterogeneous tissue. We apply this protocol to interrogate cell-type-specific gene expression and variation in cell type composition between the distinct sublayers of the hippocampal CA1 region of the brain in a rodent model of epilepsy.
For complete details on the use and execution of this protocol, please refer to Cid et al. (2021).

BEFORE YOU BEGIN

Bulk-tissue RNA-seq is a powerful approach to reveal signature patterns of gene expression across different tissues or disease states. However, it yields an average of gene transcript abundance that reflects the convoluted signal from several sources of variation, such as cell-type-specific gene expression levels. This is particularly relevant for tissues characterized by a highly heterogeneous cell type composition, such as the brain. Notably, many neurological disorders result in profound disturbances in the relative ratio of the different cell types in the affected tissue. This is the case of temporal lobe epilepsy (TLE) that is characterized by neuronal loss, reactive gliosis, and glial scarring in the hippocampus (Blümcke et al., 2013; Rusina et al., 2021). A better understanding of the sources of variation present in bulk-tissue transcriptome profiles may help identifying new targets for novel and more effective therapies. Here, we present a simple and robust protocol that leverages single cell expression data, to explore the contribution of distinct cell types to complex gene signatures obtained by differential gene expression analysis of bulk RNA-seq from brain tissue. We show that linear dimensionality reduction and hierarchical clustering can reveal patterns of cell-type-specific gene expression in gene signatures from bulk RNA-seq data.

This protocol below leverages publicly available datasets of single-cell transcriptomes (Yao et al., 2021) to deconvolve gene signature patterns in our bulk-tissue RNA-seq from the dorsal hippocampal CA1 region of healthy and epileptic rodents. We show that the interrogation of expression levels for significantly differentially expressed genes from bulk RNA-seq in the comparison between superficial and deep sublayers of CA1, unmasked different sets of tightly co-expressed genes that include bona fide markers genes of the distinct cell types (Valero et al., 2015; Cembrowski et al., 2016; Cid et al., 2021). Our analysis also revealed differences of cell type composition across the radial axis of

CA1, including microglia and astrocytes. Moreover, by extending our analysis to the rodent model of temporal lobe epilepsy, we found different transcriptional responses between CA1 sublayers at the single-cell level. This analysis unveiled a module of co-regulated genes in microglia cells that was upregulated in superficial CA1 sublayer, suggesting that reactive gliosis was prominent in this brain region in epilepsy (Cid et al., 2021). These results were reproduced using a different scRNA-seq reference dataset (Cid et al., 2021; Zeisel et al., 2015).

We propose that this protocol can be applied to deconvolve signatures of gene expression from bulk RNA-seq in other brain areas where it has the potential to inform on the contribution of cellular heterogeneity to the extracted gene patterns. We predict this protocol may also uncover valuable insights on the disturbances in cell-type composition that occur in other neurological conditions. For instance, our protocol can be applied to gene lists recovered from differential expression analysis in bulk-tissue transcriptomes in neurodegenerative diseases including Alzheimer's disease, where neuronal death and proliferative reactive gliosis represent key histopathological hallmarks. Finally, we believe this approach could also be extended to other types of tissues or datasets, including single-cell proteomics (SCP), scATAC-seq, scChIP-seq, etc.

Bulk gene expression signatures from highly heterogeneous tissue in cell-type composition

Differential expression analysis is frequently used to extract population-level gene expression signature patterns from bulk-tissue RNA-seq. In our original study, we captured changes in gene expression levels across the radial axis of the hippocampal CA1 area in healthy rats, and in an experimental model of temporal lobe epilepsy (TLE) (Cid et al., 2021). To this aim, we used laser capture microdissection to sample deep and superficial sublayers of the CA1 region of the dorsal hippocampus from healthy and epileptic rats, and performed bulk RNA-seq (LCM-RNA-seq). For a detailed description on sample processing and differential expression analysis refer to Cid et al. (2021).

> *Note:* The results of the differential expression analysis for the different contrasts relevant to run this protocol are publicly available through Mendeley data repository. Create a directory named "Differentially_expressed_gene_tables" and download into this folder, the data tables of differentially expressed genes between deep and superficial sublayers of CA1 region in healthy (Data Table 1, Filename: "Sup_deep_diff_in_controls_FullTable.tsv", https://doi.org/10.17632/p77tj6d88y) and epileptic (Data Table 2, Filename: "Sup_deep_diff_in_epilepsy_FullTable.tsv", https://doi.org/10.17632/jxmg5mwd55) rats. These data are also publicly available through the companion web application to our original publication (Cid et al., 2021): http://lopezatalayalab.in.umh-csic.es/CA1_Sublayers_%26_Epilepsy.

Reference dataset of single-cell type-specific expression profiles

Our approach leverages single-cell type-specific expression data as a reference dataset. The analysis is therefore constrained by the cell populations present in the reference scRNA-seq dataset. In our original study, we leverage scRNA-seq from mouse cortex and hippocampus (Yao et al., 2021) as reference single-cell type-specific expression profiles (Cid et al., 2021). We also performed the protocol using early data generated from mouse somatosensory cortex and the CA1 hippocampal region as scRNA-seq reference (Zeisel et al., 2015).

> ⚠ CRITICAL: The reference scRNA-seq dataset must include all the populations present in the tissue processed to extract bulk-tissue signatures of gene expression (see "limitations" section).

> *Note:* The protocol presented here leverages scRNA-seq data from the Allen Brain Map portal - Mouse Whole Cortex and Hippocampus SMART-seq (2019) with 10x-SMART-seq taxonomy (2020) as reference profiles (Yao et al., 2021). Create a directory named

"AllenBrainMap_MouseCortexAndHippo_SMART-seq" and download into this folder, the single-cell gene expression matrix ("matrix.csv") and cell metadata ("metadata.csv"). These files are publicly available through the Allen Brain Map portal: http://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-smart-seq. The single-cell gene expression matrix ("matrix.csv") and cell metadata ("metadata.csv") can be downloaded from the Allen Brain Map portal, through the link "Table of cell metadata" (Table title: "General"; Column name: "File") and "Gene expression matrix (csv)" (Table title: "Gene Expression"; Column name: "File"), respectively. Raw and processed data in Yao et al. (2021) can also be accessed through NeMO Archive for the BRAIN Initiative Cell Census Network: https://assets.nemoarchive.org/dat-jb2f34y.

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
LCM-RNA-seq	Gene Expression Omnibus (NCBI)	GEO: GSE143555
scRNA-seq data - Mouse Whole Cortex and Hippocampus SMART-seq (2019) with 10x-SMART-seq taxonomy (2020)	Yao et al. (2021); Allen Institute for Brain Science	https://portal.brain-map.org/atlases-and-data/rnaseq/mouse-whole-cortex-and-hippocampus-smart-seq
Data Table 1	Mendeley Data	https://doi.org/10.17632/p77tj6d88y
Data Table 2	Mendeley Data	https://doi.org/10.17632/jxmg5mwd55
Protocol code	Mendeley Data	https://doi.org/10.17632/nkrfxtbrmc
Software and algorithms		
R v4.1.1	R Foundation for Statistical Computing	RRID: SCR_001905
RStudio v2021.09.0-351	RStudio, PBC	RRID: SCR_000432
Seurat v4.0.4	Butler et al. (2018) and Stuart et al. (2019)	RRID: SCR_016341
Other		
Resource website for the LCM-RNA-seq data	Cid et al. (2021)	http://lopezatalayalab.in.umh-csic.es/CA1_Sublayers_&_Epilepsy

MATERIALS AND EQUIPMENT

- Data 1 (Differential gene expression analysis from bulk-tissue RNA-seq – see bulk gene expression signatures from highly heterogeneous tissue in cell-type composition in "before you begin" section).
- Data 2 (scRNA-seq gene expression matrix and metadata – see Reference dataset of single-cell type-specific expression profiles in "before you begin" section).
- R software and required packages. While different versions of the R software and associated packages may work correctly, the authors used R v4.1.1 and the following packages at the indicated versions when writing this protocol:
  - Seurat (v4.0.4)
  - cowplot (v1.1.1)
  - data.table (v1.14.2)
  - ggplot2 (v3.3.5)
  - gplots (v3.1.1)
  - patchwork (v1.1.1)
- Hardware:
  - Memory for reading the single-cell gene expression matrix and building the single-cell object: 64 GB required, 128 GB recommended.
  - Memory for loading single-cell objects once built: 24 GB required, 32 GB recommended.
  - Processor: 1 required, 4 threads are recommended for parallel processing of large datasets.

⏸⏸ **Pause point:** The protocol can be paused at any time. To do so, the user can save an image of the current workspace and R session using the following command within the R or R Studio console:

```
> save.image("Deconvolution_Experiment.Rdata")
```

To reload the workspace image in future R session, use the following command:

```
> load("Deconvolution_Experiment.Rdata")
```

STEP-BY-STEP METHOD DETAILS

Step 1: Load differentially expressed genes (DEGs) between the different sample groups and set a common filter based on Adjusted p value (padj)

⏱ **Timing: 5 min**

This step loads the DEGs between the different contrasts for downstream analysis. DEGs with an padj < 0.1 based on sublayer differences for control and epileptic condition are retained. See bulk gene expression signatures from highly heterogeneous tissue in cell-type composition in the "before you begin" section.

1. Load the DEGs tables into dataframes.
   a. Set the working directory path to the RNA-seq data folder.

```
> path = "~/Differentially_expressed_gene_tables/"
> setwd(path)
```

   b. Read the RNA-seq data from disk.

```
> ctrl_DEGs <- read.delim("Sup_deep_diff_in_controls_FullTable.tsv", stringAsFactors =
FALSE)
> epil_DEGs <- read.delim("Sup_deep_diff_in_epilepsy_FullTable.tsv", stringAsFactors =
FALSE)
```

   Note: The DEGs between CA1 sublayers in control and epileptic condition are available as Data Table 1 and Data Table 2 (see "before you begin" section and "key resources table"). These data are also available in the Resource web application for the LCM-RNA-seq data from Cid et al. (2021): http://lopezatalayalab.in.umh-csic.es/CA1_Sublayers_&_Epilepsy.

2. Remove DEGs table rows with empty gene symbols.

```
> ctrl_DEGs <- ctrl_DEGs[!is.na(ctrl_DEGs$Gene_symbol),]
> epil_DEGs <- epil_DEGs[!is.na(epil_DEGs$Gene_symbol),]
```

3. Filter the DEGs tables by significance (padj < 0.1).

```
> ctrl_DEGs <- ctrl_DEGs[ctrl_DEGs$padj < 0.1,]
> epil_DEGs <- epil_DEGs[epil_DEGs$padj < 0.1,]
```

4. Rank the DEGs tables by fold-change.

```
> ctrl_DEGs <- ctrl_DEGs[order(-ctrl_DEGs$log2FoldChange),]

> epil_DEGs <- epil_DEGs[order(-epil_DEGs$log2FoldChange),]
```

5. Subset DEGs tables based on the layer they are enriched in controls and in epileptics.

```
> ctrl_DEGs_sup <- ctrl_DEGs[ctrl_DEGs$log2FoldChange > 0,]

> ctrl_DEGs_deep <- ctrl_DEGs[ctrl_DEGs$log2FoldChange < 0,]

> epil_DEGs_sup <- epil_DEGs[epil_DEGs$log2FoldChange > 0,]

> epil_DEGs_deep <- epil_DEGs[epil_DEGs$log2FoldChange < 0,]
```

6. Get DEGs lists, extracting gene symbols from DEGs tables.

```
> ctrl_DEGs_sup <- ctrl_DEGs_sup$Gene_symbol

> ctrl_DEGs_deep <- ctrl_DEGs_deep$Gene_symbol

> epil_DEGs_sup <- epil_DEGs_sup$Gene_symbol

> epil_DEGs_deep <- epil_DEGs_deep$Gene_symbol
```

Step 2: Load the single-cell expression dataset and build the single-cell dataset object

⏱ Timing: 20 min

This step loads the original scRNA-seq dataset and builds the single-cell dataset object. See reference dataset of single-cell type-specific expression profiles in the "before you begin" section.

7. Load the single-cell gene expression matrix and cell metadata.
   a. Set the working directory path to a single-cell RNA-seq data folder.

```
> path = "~/AllenBrainMap_MouseCortexAndHippo_SMART-seq/"

> setwd(path)
```

   b. Read the single-cell RNA-seq metadata from disk.

```
> metadata <- read.csv(file = "metadata.csv", row.names = 1)
```

   c. Load the data.table library, to access the fread() function, a very efficient tool to read regular delimited files from disk.

```
> library(data.table)
```

   d. Read the single-cell RNA-seq counts from disk.

```
> counts <- fread("matrix.csv", data.table = FALSE)
```

   e. Set the cell barcodes as rownames.

```
> rownames(counts) <- counts$sample_name
```

    f. Remove the column sample_name to have a count matrix.

```
> counts <- counts[,-!names(counts) %in% c("sample_name")]
```

    g. Set the count matrix as a matrix.

```
> counts <- as.matrix(counts)
```

    h. Transpose the counts matrix to get the correct format to be used as input in Seurat.

```
> transposed_counts <- t(counts)
```

    i. Remove old count objects to free space.

```
> rm(counts)
```

    j. Free memory through garbage collection.

```
> gc()
```

    △ CRITICAL: The gene-barcode matrix from the scRNA-seq dataset must be arranged so that each of the columns represents a barcode and each of the rows represents a gene (related to point 7.h.).

8. Build the Seurat single-cell dataset object.
    a. Load the Seurat library.

```
> library(Seurat)
```

    b. Create the Seurat object with the transposed raw counts and cell metadata without filtering any cell or genes.

```
> sc_data <- CreateSeuratObject(counts = transposed_counts,

  meta.data = metadata, min.cells = 0, min.features = 0,

  project ="AllenBrainMap_MouseCortexHippo_SMART-seq-

  2019_with_10x_SMART-seq-2020-taxonomy")
```

9. Clean unneeded data.
    a. Remove transposed_counts and metadata objects to free space.

```
> rm(transposed_counts)

> rm(metadata)
```

    b. Free memory through garbage collection.

```
> gc()
```

Step 3: Subset scRNA-seq reference dataset to match cell-type populations in the bulk-tissue RNA-seq

    ⊙ Timing: 10 min

This step sets a subset of the scRNA-seq dataset to match the cell types present in the tissue processed for bulk RNA-seq.

10. Explore the different cell metadata categories in the scRNA-seq object to select the categories more related to bulk-tissue RNA-seq data. The categories selected to perform the subset are "region_label" and "subclass_label". See cell numbers in Tables 1 and 2.

Table 1. Cell numbers by "region_label" before filtering step

ACA	AI	AUD	CLA	ENTl	ENTm	GU	HIP	MOp	MOs-FRP	ORB
5122	1536	1486	828	1618	1570	1481	6598	6516	9656	1461
PAR-POST-PRE	PL-ILA	PTLp	RSP	SSp	SSs	SUB-ProS	TEa-PERI-ECT	VIS	VISp	
1636	1452	1539	2007	5577	1864	1608	1602	3395	16421	

Table 2. Cell numbers by "subclass_label" before filtering step

[No label]	Astro	CA1-ProS	CA2	CA3	Car3	CR	CT SUB	DG	Endo	L2 IT ENTl
268	976	1701	21	315	1980	32	173	2469	213	179
L2 IT RHP	L2/3 IT CTX-1	L2/3 IT CTX-2	L2/3 IT ENTl	L2/3 IT PPP	L3 IT ENT	L3 RSP-ACA	L4/5 IT CTX	L5 IT CTX	L5 IT TPE-ENT	L5 NP CTX
375	5959	106	253	1395	577	200	11522	2934	338	2363
L5 PPP	L5 PT CTX	L6 CT CTX	L6 IT CTX	L6 IT ENTl	L6b CTX	L6b/CT ENT	Lamp5	Meis2	Micro-PVM	NP PPP
47	1974	6210	5015	83	2213	693	4755	172	176	150
NP SUB	Oligo	Pvalb	SMC-Peri	Sncg	Sst	Sst Chodl	SUB-ProS	V3d	Vip	VLMC
257	236	4365	198	1491	5258	268	467	1	6436	159

```
> str(sc_data@meta.data)
```

Note: The number of cells can be consulted at any moment using the function table() with the associated category/metadata as argument.

```
> table(sc_data$region_label)

> table(sc_data$subclass_label)
```

Note: Consider removing potential confounding factors (e.g. conditions or treatments) that are unrelated to the bulk-tissue RNA-seq from the initial scRNA-seq dataset.

11. Filter out any cell population from tissue areas, treatments or other groups that are unrelated to the bulk-tissue RNA-seq data. Troubleshooting 1.
   a. We first generate a scRNA-seq reference of the cells-types in the hippocampal CA1 region of the brain. The 74,973 cells in the initial scRNA-seq dataset are filtered to subset cells from hippocampal region and non-neuronal subclasses (Astro, Micro-PVM, Endo, Oligo or VLMC). Then, excitatory cells from CA2, CA3, DG and also a group of cells in a subclass without label ([No label]) are removed. This initial filtering extracts a subset of 5,506 cells.

```
> sc_data <- subset(x = sc_data,
   subset = region_label == "HIP" | subclass_label %in% c("Astro","Micro-PVM","Endo","Oligo","VLMC"))

> sc_data <- subset(x = sc_data,
   cells = colnames(sc_data)[!sc_data$subclass_label %in% c("","CA2","CA3","DG")])
```

   b. Of these, cells labeled by subclass as CA1-ProS, Astro, Lamp5, Vip, Sncg, Sst, Oligo, Endo, Micro-PVM, VLMC, or Pvalb are retained for further analysis, remaining a total of 5,429 cells. See cell numbers in Tables 3 and 4. Troubleshooting 2.

Table 3. Cell numbers by "region_label" after filtering step

ACA	AI	AUD	CLA	ENTl	ENTm	GU	HIP	MOp	MOs-FRP	ORB
13	4	56	11	56	0	87	3707	18	557	48
PAR-POST-PRE	PL-ILA	PTLp	RSP	SSp	SSs	SUB-ProS	TEa-PERI-ECT	VIS	VISp	
7	20	37	15	23	69	0	0	2	699	

Table 4. Cell numbers by "subclass_label" after filtering step

[No label]	Astro	CA1-ProS	CA2	CA3	Car3	CR	CT SUB	DG	Endo	L2 IT ENTl
0	976	1592	0	0	0	0	0	0	213	0
L2 IT RHP	L2/3 IT CTX-1	L2/3 IT CTX-2	L2/3 IT ENTl	L2/3 IT PPP	L3 IT ENT	L3 RSP-ACA	L4/5 IT CTX	L5 IT CTX	L5 IT TPE-ENT	L5 NP CTX
0	0	0	0	0	0	0	0	0	0	0
L5 PPP	L5 PT CTX	L6 CT CTX	L6 IT CTX	L6 IT ENTl	L6b CTX	L6b/CT ENT	Lamp5	Meis2	Micro-PVM	NP PPP
0	0	0	0	0	0	0	864	0	176	0
NP SUB	Oligo	Pvalb	SMC-Peri	Sncg	Sst	Sst Chodl	SUB-ProS	V3d	Vip	VLMC
0	236	69	0	416	266	0	0	0	462	159

```
> sc_data <- subset(x = sc_data,

  cells = colnames(sc_data)[!sc_data$subclass_label %in%

   c("L2 IT RHP","L2/3 IT CTX-1","L2/3 IT CTX-2",

     "L2/3 IT ENTl","L2/3 IT PPP","L5 IT TPE-ENT",

     "L6 CT CTX","L6b CTX","Meis2","NP SUB","Sst Chodl",

     "SUB-ProS")])
```

12. Rename and group the cells in their corresponding major cell populations.
   a. Remove unused levels (labels without associated cells). See cell numbers in Table 5.

Table 5. Cell numbers by "subclass_label" after removing labels without associated cells

Astro	CA1-ProS	Endo	Lamp5	Micro-PVM	Oligo	Pvalb	Sncg	Sst	Vip	VLMC
976	1592	213	864	176	236	69	416	266	462	159

```
> sc_data$subclass_label <- droplevels(x = sc_data$subclass_label)
```

   b. Set the "subclass_label" metadata as the "active.ident" (default identity class) for easy customization.

```
> Idents(sc_data) <- "subclass_label"
```

   c. Rename identity classes to the corresponding major cell population. The remaining 5,429 cells are grouped in their corresponding major cell populations as follows: Astrocytes (Astro) (976 cells), Endothelial (Endo) (213 cells), Interneurons (Lamp5, Pvalb, Sncg, Sst, Vip) (2,077 cells), Microglia (Micro-PVM) (176 cells), Mural (VLMC) (159 cells), Oligodendrocytes (Oligo) (236 cells) and Pyramidal (CA1-ProS) (1,592 cells). See Tables 5 and 6.

Table 6. Cell numbers by "active.ident" after grouping to their major cell-type populations

Astrocytes	Endothelial	Interneurons	Microglia	Mural	Oligodendrocytes	Pyramidal
976	213	2077	176	159	236	1592

```
> sc_data <- RenameIdents(sc_data,

        "Astro" = "Astrocytes",

        "CA1-ProS" = "Pyramidal",

        "Endo" = "Endothelial",

        "Lamp5" = "Interneurons",

        "Micro-PVM" = "Microglia",

        "Oligo" = "Oligodendrocytes",

        "Pvalb" = "Interneurons",

        "Sncg" = "Interneurons",

        "Sst" = "Interneurons",

        "Vip" = "Interneurons",

        "VLMC" = "Mural")
```

d. Reorder alphabetically the identity classes. See cell numbers in Table 6.

```
> Idents(sc_data) <- factor(Idents(sc_data), levels = sort(levels(sc_data)))
```

13. Normalize gene counts.

```
> sc_data <- NormalizeData(sc_data, normalization.method = "LogNormalize", scale.factor =
10000)
```

14. Identify genes that are outliers on a 'mean variability plot'.

```
> sc_data <- FindVariableFeatures(sc_data, selection.method = "vst", nfeatures = 2000)
```

15. Scale and center variable genes.

```
> sc_data <- ScaleData(sc_data)
```

16. Perform dimensionality reduction to summarize and visualize the cells in the low-dimensional space.
    a. Perform linear dimensionality reduction by PCA over the variable genes (default option).

```
> sc_data <- RunPCA(sc_data)
```

b. Estimate the number of principal components (PCs) that are biologically informative by plotting the standard deviations of the PCs for easy identification of an elbow in the graph (Elbow plot). This elbow often corresponds well with the significant dimensions that capture the majority of the variation in the data. Here, the elbow was identified at the 20th PC.

```
> ElbowPlot(sc_data, ndims = 50)
```

c. Perform non-linear dimensionality reduction over biologically informative dimensions from linear dimensionality reduction using stochastic nearest neighbors (tSNE) (van der Maaten and Hinton, 2009) and uniform manifold approximation and projection (UMAP) (McInnes et al., 2018) state-of-the-art techniques.

```
> sc_data <- RunTSNE(sc_data, dims = 1:20)

> sc_data <- RunUMAP(sc_data, dims = 1:20)
```

⏸⏸ Pause point: Save the scRNA-seq object and resume the analysis in future R sessions without repeating the previous steps:

```
> saveRDS(sc_data, "Single-cell_custom_subset.rds")
```

To reload the scRNA-seq object saved in a previous R session, use the following command:

```
> sc_data <- readRDS("Single-cell_custom_subset.rds")
```

17. Create a figure showing dimensionality reduction of major cell types (Figure 1). Use custom labels from the scRNA-seq and the dimensionality reduction techniques applied to the data.
    a. Load the cowplot library, to access the get_legend() function.

```
> library(cowplot)
```

b. Load the patchwork library, to access the area() function.

```
> library(patchwork)
```

c. Load the ggplot2 library, to access the ggsave() function.

```
> library(ggplot2)
```

d. Generate the dimensionality reduction plots.

```
> cols <- c("limegreen", #Astrocytes

            "steelblue", #Endothelial

            "mediumorchid4", #Interneurons

            "firebricks2", #Microglia

            "magenta", #Mural

            "gray52", #Oligodendrocytes

            "tan1") #Pyramidal
> pca_plot <- DimPlot(sc_data, reduction = "pca", pt.size = 0.1, label = T, cols = cols)
> tsne_plot <- DimPlot(sc_data, reduction = "tsne", pt.size = 0.1, label = T, cols = cols)
> umap_plot <- DimPlot(sc_data, reduction = "umap", pt.size = 0.1, label = T, cols = cols)
> legend <- get_legend(umap_plot)
```

e. Define the figure layout.

```
> layout <- c (area(1, 1, 1, 2), #PCA

               area(1, 3, 1, 4), #tSNE
```

```
                    area(1, 5, 1, 6), #UMAP

                    area(1, 7)) #Legend
```

f. Build the figure.

```
> fig1 <- pca_plot + tsne_plot + umap_plot + legend + plot_layout(design = layout) & NoLegend()
```

g. Save the figure to disk.

```
> ggsave(filename = "Fig_1.png", plot = fig1, width = 13, height = 3.75, dpi = 300)
```

Step 4: Perform the deconvolution of cell-type-specific signal in bulk-tissue RNA-seq top DEGs using single-cell RNA-seq reference

⏱ Timing: 10 min

This step subsets top DEGs lists of equal size for the groups of interest and imputes bulk signal to cell-type leveraging on the scRNA-seq subset from step 3.

18. Get the list of genes names from the single-cell object.

```
> gene_names <- rownames(sc_data)
```

19. Intersect the DEGs lists with the list of genes names from the single-cell reference object, preserving the fold change order from point 4.

```
> ctrl_DEGs_sup <- ctrl_DEGs_sup[ctrl_DEGs_sup %in% gene_names]

> ctrl_DEGs_deep <- ctrl_DEGs_deep[ctrl_DEGs_deep %in% gene_names]

> epil_DEGs_sup <- epil_DEGs_sup[epil_DEGs_sup %in% gene_names]

> epil_DEGs_deep <- epil_DEGs_deep[epil_DEGs_deep %in% gene_names]
```

Optional: In the case that the bulk-tissue RNA-seq and reference scRNA-seq dataset are from different species (e.g. rat vs. mouse), ortholog conversion should be performed. Troubleshooting 3.
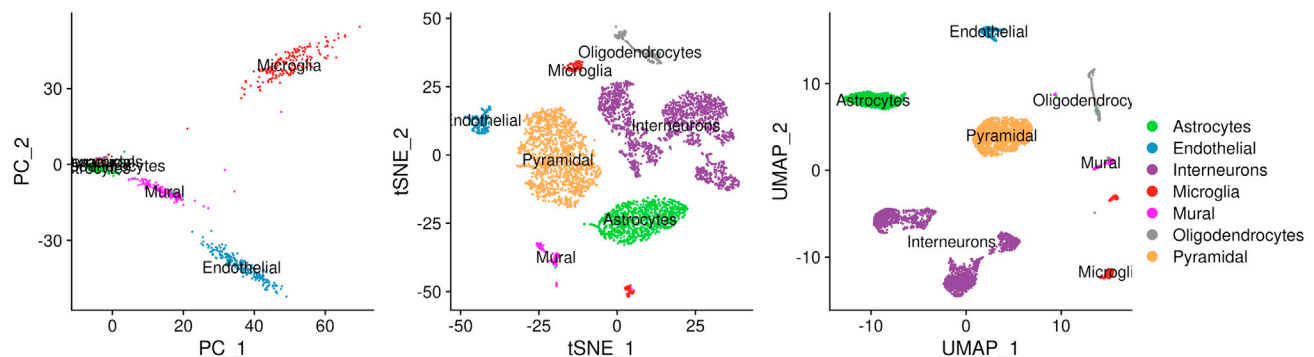

Figure 1. Major cell types present in the reference scRNA-seq dataset
Cell populations in scRNA-seq data from the Allen Brain Map portal (Mouse Whole Cortex and Hippocampus SMART-seq (2019) with 10x-SMART-seq taxonomy (2020) (Yao et al., 2021)) identified as non-neuronal or from CA1 (from "hippocampus" but not NA, "CA2", "CA3" or "DG") were subset and grouped in their corresponding major cell populations as follows: Astrocytes (Astro) (976 cells), Endothelial (Endo) (213 cells), Interneurons (Lamp5, Pvalb, Sncg, Sst, Vip) (2077 cells), Microglia (Micro-PVM) (176 cells), Mural (VLMC) (159 cells), Oligodendrocytes (Oligo) (236 cells) and Pyramidal (CA1-ProS) (1,592 cells). Left, Principal component analysis (PCA); Center, t-distributed stochastic neighbor embedding (tSNE); Right, uniform manifold approximation and projection (UMAP).

20. Capture comparable convoluted gene signatures by using gene sets of equal size that are differentially regulated between the experimental conditions. Subset top 250 differentially regulated genes for each sublayer and condition for further analysis, with the genes sorted in descending order by magnitude of change. Troubleshooting 4.

```
> ctrl_DEGs_sup <- head(ctrl_DEGs_sup,250)

> ctrl_DEGs_deep <- rev(tail(ctrl_DEGs_deep,250))

> epil_DEGs_sup <- head(epil_DEGs_sup,250)

> epil_DEGs_deep <- rev(tail(epil_DEGs_deep,250))
```

Note: In our case, we have 4 top DEGs lists that should be iteratively renamed to "topDEGs_list" in point 21 to generate the plots in Figure 2. Our top DEGs lists are comparable in size, as they contain top 250 DEG genes, with the exception of "epil_DEGs_sup" that contains 230 genes.

21. Set "topDEGs_list" as any of the previous 4 DEGs lists and repeat the following steps (21–25) for all the other lists, renaming the plot output names in points 24 and 25 to avoid overwriting the plots in the different iterations.

```
> topDEGs_list <- ctrl_DEGs_sup     #Repeat steps 21-25 for each DEG list.
```

22. Scale the normalized scRNA-seq data for the top DEGs list.

```
> sc_data <- ScaleData(sc_data, features = topDEGs_list)
```

23. Perform dimensionality reduction of the gene list to summarize and visualize the cells in the low-dimensional space. Troubleshooting 5.
    a. Perform linear dimensionality reduction by PCA over the top DEGs.

```
> sc_data <- RunPCA(sc_data, features = topDEGs_list)
```

   b. Retain genes with variance. This gene list will be used to perform pairwise Pearson correlation coefficients (related to point 25.e.).

```
> topDEGs_list <- rownames(sc_data@reductions[["pca"]]@feature.loadings)
```

24. Generate the dimensionality reduction plots and save them to disk. For Figure 2 we use the PCA plots with labels set to false, to maximize visualization.

```
> pca_plot <- Dimplot(sc_data, reduction = "pca", pt.size = 0.1, label = F, cols = cols)

> legend <- get_legend(pca_plot)

> pca_plot <- pca_plot & NoLegend()

> ggsave(filename = "Fig_2_pca_plot.png", plot = pca_plot, width = 3.75, height = 3.75, dpi = 300)

> ggsave(filename = "Fig_2_legend.png", plot = legend, width = 2, height = 3.75, dpi = 300)
```

25. Perform pairwise correlations and hierarchical clustering for these gene sets (top DEGs) across all cells in the scRNA-seq subset, to capture cell-type-specific gene signatures that were assigned to major cell types in CA1 by using previously identified cell markers.
    a. Define the correlation plot palette.
       i. Create a matrix of 50x10 random values within the range [−1, +1].

```
> random.matrix <- matrix(runif(500, min = -1, max = 1), nrow = 50)
```

    ii. Produce the sample quantiles corresponding to the given probabilities.

```
> quantile.range <- quantile(random.matrix, probs = seq(0, 1, 0.01))
```

    iii. Define the quantiles where the minimum and maximum correlation values were set to the lowest and highest color. In our case, it was set empirically to 35% and 83% respectively, as these values maximized the contrast for the distribution of values.

```
> palette.breaks <- seq(quantile.range["35%"], quantile.range["83%"], 0.06)
```

    iv. Create a color ramp that maps the previous interval.

```
> color.palette <- colorRampPalette(c("#0571b0","#f7f7f7","#ca0020"))(length(palette.
breaks)-1)
```

    b. Import the library gplots to access the enhanced heatmap function heatmap.2().

```
> library(gplots)
```

    c. Define the hierarchical cluster analysis function, with Pearson correlation coefficient as distance matrix and average as agglomeration method.

```
> clustFunction <- function(x)
    hclust(as.dist(1-cor(t(as.matrix(x)), method = "pearson")), method = "average")
```

    d. Define the heatmap correlation function, with Pearson correlation coefficient as numeric matrix and the color palette, breaks and clustering function set above.

```
> heatmapPearson <- function(correlations)
        heatmap.2(x = correlations,
            col = color.palette,
            breaks = palette.breaks,
            trace = "none", symm = T,
            hclustfun = clustFunction)
```

    e. Compute the Pearson correlation coefficient on the logarithm of normalized expression of genes from "topDEGs_list" in all cells from scRNA-seq subset.

```
> correlations_DEGs_log <- cor(method = "pearson",
log2(t(as.matrix(sc_data@assays[["RNA"]]@data[topDEGs_list,]))+1))
```

    f. Generate the gene-gene correlation plot and save it to disk.

```
> pdf(file = "Fig_2_corr_plot.pdf", width = 25, height = 25)
> heatmapPearson(correlations_DEGs_log)
> dev.off()
```

26. Figure 2 shows PCA plots and heatmaps of pairwise correlations generated from the 4 lists of top DEGs in our study.
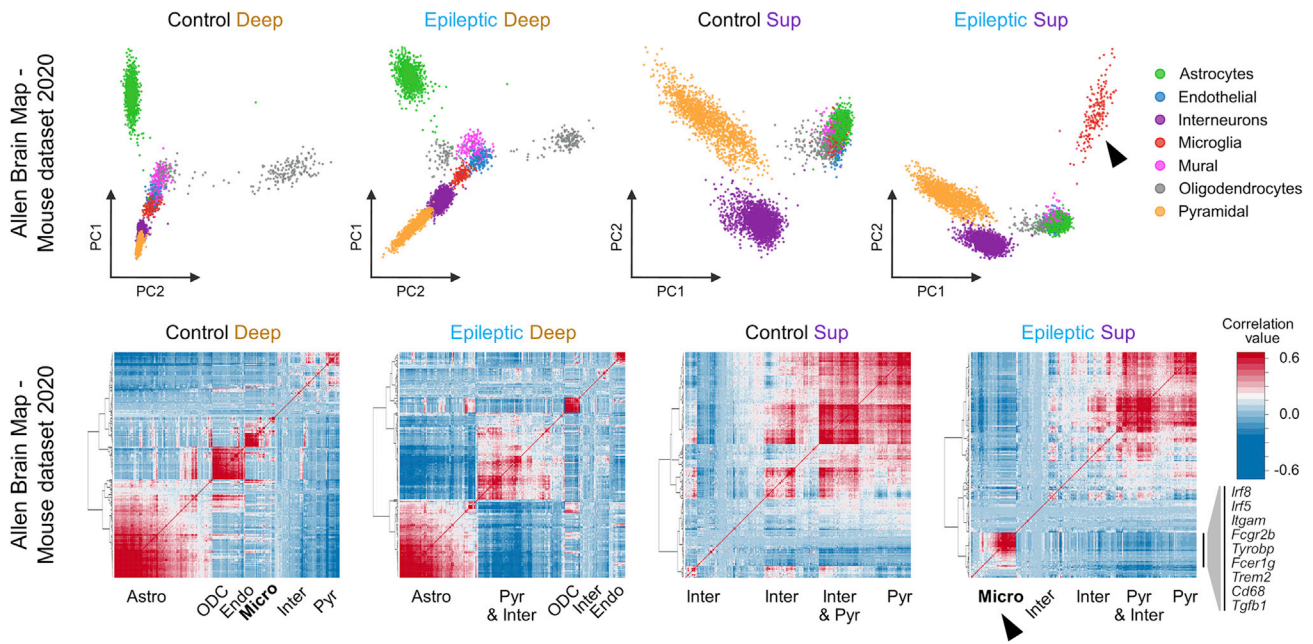
Figure 2. Deconvolution of gene signatures from bulk-tissue RNA-seq reveals strong presence of reactive microglia in superficial CA1 sublayer of the hippocampus in epilepsy

Patterns of cell-type-specific gene expression were identified by deconvolution of bulk-tissue transcriptome profiling of deep and superficial hippocampal CA1 sublayers of control and epileptic animals. This analysis reveals strong presence of reactive microglia in superficial CA1 sublayer in epilepsy. Gene sets were the top 250 DEGs between superficial and deep CA1 sublayers in epileptic and control rats identified in bulk-tissue RNA-seq. For the selected genes, normalized expression in single cells was retrieved from publicly available scRNA-seq data from the Allen Brain Map portal (Mouse Whole Cortex and Hippocampus SMART-seq (2019) with 10x-SMART-seq taxonomy (2020)) (Yao et al., 2021) and single-cells were summarized by linear dimensionality reduction using PCA (top panels). Cells are colored by population membership (step 2). Hierarchical clustering heatmaps of pairwise correlations for all individual cells using scRNA-seq expression data for selected gene sets identified in differential expression analysis with bulk-tissue RNA-seq (top 250 DEGs) (bottom panels). Bona fide markers of distinct cell types are present in clusters of highly correlated genes representing cell-type gene signatures convoluted in the bulk-tissue RNA-seq. Patterns of cell-type-specific gene expression present in the bulk-tissue RNA-seq gene signatures of deep and superficial CA1 sublayers of control and epileptic animals lead to segregation of the individual cells in the corresponding PCAs and in the heatmap of the pairwise correlation matrices. Note the different distribution of distinct cell types in the deep and superficial (Sup) sublayers of the CA1 region in control and epileptic rats. Also note the presence of a strong gene signature of Micro in the Sup CA1 sublayer of epileptic rats (arrowhead). Names of highly correlated genes enriched in Micro are shown (Zeisel et al., 2018). Pyr, pyramidal cells; Inter, interneurons; ODC, oligodendrocytes; Astro, astrocytes; Endo, endothelial cells; Micro, microglia; Mural, mural cells.
Reprinted from Cid et al. (2021), with permission from Elsevier.

Note: Points 24 and 25 generate the plots shown in Figure 2. Rename the output files in points 24 and 25 to avoid overwriting of the plots when different iterations of Step 4 are performed.

Note: A R script with the code of this protocol is publicly available at Mendeley Data repository: https://doi.org/10.17632/nkrfxtbrmc.

EXPECTED OUTCOMES

We show here a reference profile-based deconvolution protocol designed to explore sources of variations such as cell-type-specific gene expression and cell-type composition that are present in signatures of gene expression obtained from bulk-tissue RNA-seq data. We show that linear dimensionality reduction and hierarchical clustering of pairwise correlations of single-cell data for gene lists obtained from differential expression analysis using bulk-tissue RNA-seq, can inform on the contribution of cell types and cellular states to these gene signatures. Anticipated outcomes of this protocol include deconvolution of gene signatures from bulk-tissue RNA-seq in highly heterogeneous tissues, such as brain tissue, for qualitative inference of changes in constituent cell-types (Figure 2). The protocol can also inform on changes in cellular state caused by the disease condition, as in the

case of the presence of reactive microglia in samples of the superficial CA1 sublayer of animals subjected to an experimental model of temporal lobe epilepsy (Figure 2). The gene clusters and correlation values are publicly available through the web application we developed in our original publication (Cid et al., 2021): http://lopezatalayalab.in.umh-csic.es/CA1_Sublayers_&_Epilepsy. Deconvolved signatures in bulk-tissue RNA-seq can be further validated using single-nuclei RNA-seq data from CA1 hippocampal area, available at http://lopezatalayalab.in.umh-csic.es/CA1_SingleNuclei_&_Epilepsy (Cid et al., 2021).

QUANTIFICATION AND STATISTICAL ANALYSIS

We perform principal component analysis (PCA) to interrogate for the presence of strong patterns of cell-type-specific gene expression in the bulk RNA-seq signatures (top DEGs). We also perform pairwise Pearson correlation and hierarchical clustering for the selected gene sets (top DEGs from RNA-seq analysis) across all cells in the reference scRNA-seq dataset. The analysis captures co-regulated gene modules that are assigned to major cell types in CA1 by using previously identified marker genes. First, pairwise Pearson correlation coefficients are calculated for all cells in the reference scRNA-seq dataset using the logarithm of normalized expression for the gene subset of top DEGs from bulk-tissue RNA-seq analysis. Second, hierarchical cluster analysis is performed with the previous gene correlation coefficient as distance matrix and average as agglomeration method. A data visualization tool has been developed to allow users for interactive data exploration and the access to gene correlation values: http://lopezatalayalab.in.umh-csic.es/CA1_Sublayers_&_Epilepsy (go to the tab menu Data visualization/Signature correlation).

LIMITATIONS

Bulk-tissue RNA-seq has been extensively used to extract molecular signatures of complex tissues and disease states. However, one of the major limitations of this technique is that it measures the average of gene expression levels (i.e., transcript and gene abundance estimates) that is the result of cell-type-specific gene expression weighted by cell-type proportions. Efforts aimed to deepen our understanding on the contribution of changes in cell-type composition and variation in cellular states to gene signature patterns in disease may help improving the identification of novel therapeutic targets.

In bulk RNA-seq, changes in gene expression levels between sample groups can result from variations in transcript abundance in the whole tissue. However, in highly heterogeneous tissues, such as the brain, changes in transcripts levels are expected to vary across the different cell types within the tissue, or even, be cell-type-specific, in the most extreme case. Moreover, variation in gene expression levels between experimental conditions can be driven by gene regulatory mechanisms, as well as from changes in the proportions of cell-types. A combination of both sources of variations, cell-type-specific responses and changes in cell-type composition, is present in most prevalent neurological conditions such as epilepsy, Alzheimer's disease, and Parkinson's disease (Cid et al., 2021; Gjoneska et al., 2015; Nido et al., 2020). Here, we present a simple and fast approach that leverage scRNA-seq cell-type-specific gene expression profiles from publicly available datasets, to explore changes in cell proportions and cellular states from bulk-tissue RNA-seq data. To do so, we rely on signature patterns of gene expression extracted from bulk RNA-seq data by performing gene-level differential expression analysis between conditions.

We show that projecting differentially expressed gene lists from bulk RNA-seq in single-cell-type-specific gene profiles from scRNA-seq can inform on changes in cell-type proportions and cellular states. Linear dimensionality reduction by PCA is performed to interrogate for the presence of sources of variation associated to specific cell types. This analysis may reveal strong patterns in the dataset that can be informative of the contribution of distinct cell types to complex signatures of gene expression. We also probed for mutual linearity of all pairs of genes in the bulk-RNA-seq gene signature, across all cells in the scRNA-seq reference. Next, hierarchical clustering on a correlation-based similarity metric is

performed to capture well-defined clusters of genes that are mutually linearly correlated to each other. Finally, the clusters showing strong covariation and intrinsic redundancy are imputed to specific cell types based on the presence of bona fide marker genes. We demonstrate our coarse-grained approach constitutes a fast and simple method to explore the contribution of cell-type composition and cellular states to gene signatures obtained from bulk-tissue RNA-seq. In our original study, this approach revealed that many differentially expressed genes between deep and superficial sublayers of the CA1 area are associated with changes in cell-type composition across the radial axis of this hippocampal region. Notably, we also observed disease-associated gene signatures that could be imputed to a specific cell state, such as reactive microglia (Cid et al., 2021).

A variety of computational methods have been developed to deconvolve bulk RNA-seq data from complex tissues to estimate cell-type composition (Avila Cobos et al., 2018). More recent approaches such as deconvSeq (Du et al., 2019), MuSIC (Wang et al., 2019), DWLS (Tsoucas et al., 2019), Bisque (Jew et al., 2020), and SCDC (Dong et al., 2021) leverage external scRNA-seq data to perform reference profile-based deconvolution of bulk-tissue RNA-seq (Avila Cobos et al., 2020). These methods use gene expression data (count matrices) to obtain a quantitative estimate of the proportions of cell types in bulk RNA-seq samples from complex tissues and disease states.

While this aim is clearly beyond the scope of our work, there are limitations that are common to both approaches. These limitations, that can be due to biological processes underlying complex diseases or arise from technical bias, are related with the fact that deconvolution approaches based on scRNA-seq data are inherently constrained by cell-types present in the single-cell reference. First, diseased tissue can be populated by cell-types that are not normally present under healthy conditions but can infiltrate from external sources such as the bloodstream, during inflammation. Second, scRNA-seq methods often deviate from stereological data on cell-type composition and abundance in tissues of origin. The prevailing view is that sample dissociation methods introduce protocol-specific biases affecting cell proportion estimates in scRNA-seq data (Bakken et al., 2018; Denisenko et al., 2020). Third, a note of caution must be added when using single-nucleus transcriptome (snRNA-seq) profiling as external reference. Single-cell and single-nucleus transcriptome profiles show good general correlation (Bakken et al., 2018; Habib et al., 2017; Lake et al., 2017). However, single-nuclei data have shown to perform poorly in deconvolution of bulk gene expression to estimate cell proportions (Patrick et al., 2020). Moreover, recent data from human microglia have revealed that a significant proportion of genes are depleted in nuclei compared to whole cells (Thrupp et al., 2020). Together, these limitations warn us about the potential impact that the lack of consistency between bulk-tissue RNA-seq data and scRNA-seq reference dataset may have on the results of deconvolutional methods using cell-type-specific gene expression references.

The protocol described here is intended to offer a first glimpse into the sources of intrinsic variations associated with gene expression signatures from bulk-tissue RNA-seq data. It provides a starting point to generate new hypotheses that can be further explored using complementary techniques aimed at validating high-throughput findings at the single-cell level with spatial information in tissue, such as immunofluorescence and single-molecule RNA fluorescence in situ hybridization (Cid et al., 2021).

TROUBLESHOOTING
Problem 1
Unbalanced cell-type representation in the reference single-cell dataset (step 3, point 11).

Potential solution
Add to the subset the underrepresented populations that are ubiquitous to the organ of interest from other areas of the organ in the same dataset, such as non-neuronal populations of CA1 that are more represented in other areas.

Problem 2

Subpopulations with a reduced number of cells can make it difficult to interpret the signature patterns to study (step 3, point 11).

Potential solution

Merge these subpopulations into a parent population or simply remove them from the dataset.

Problem 3

There is no good single-cell reference for the species used in the bulk-tissue RNA-seq experiment, as is the case for rats (step 4, point 19).

Potential solution

Perform deconvolution between different species. To do this, convert the gene names into their orthologues for the species of the scRNA-seq experiment. This can be easily done using R and the functions from biomaRt package or using the Ensembl BioMart web tool: https://www.ensembl.org/biomart/martview/.

Problem 4

Size of DEGs lists can be variable between comparisons, biasing the deconvolution approach based on dimensional reduction by imputing more heterogeneity in longer DEGs lists (step 4, point 20).

Potential solution

Keep the same number of top DEGs between the different contrasts to avoid a list size effect; this implies a reduction of all lists to the smallest one.

Problem 5

Non-linear dimensional reductions on lists of a few genes or dimensional reductions on lists of several hundred or thousands of genes may be useless if they segregate all populations in the different contrasts (step 4, point 23).

Potential solution

Perform deconvolution by gene correlation analysis and hierarchical clustering. It can overcome that limitation, as this technique is less affected by list size than dimensional reductions. Alternatively, you can fix a more stringent threshold of padj, to get a smaller gene list with a more restrictive cell-type signature.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jose P. Lopez-Atalaya. (jose.lopezatalaya@csic.es).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The data tables of differentially expressed genes from LCM-RNA-seq and R code used in this study are publicly available at Mendeley Data repository: R code: https://doi.org/10.17632/nkrfxtbrmc; Data Table 1: https://doi.org/10.17632/p77tj6d88y, data table presented in tab delimited format, of differential gene expression analysis of LCM-RNA-seq between the hippocampal deep and superficial CA1 sublayers in healthy adult rats); Data Table 2: https://doi.org/10.17632/jxmg5mwd55, data table presented in tab delimited format, of differential gene expression analysis of LCM-RNA-seq between the hippocampal deep and superficial CA1 sublayers in epileptic adult rats.

REFERENCES

Avila Cobos, F., Vandesompele, J., Mestdagh, P., and De Preter, K. (2018). Computational deconvolution of transcriptomics data from mixed cell populations. Bioinformatics 34, 1969–1979.

Avila Cobos, F., Alquicira-Hernandez, J., Powell, J.E., Mestdagh, P., and De Preter, K. (2020). Benchmarking of cell type deconvolution pipelines for transcriptomics data. Nat. Commun. 11, 5650.

Bakken, T.E., Hodge, R.D., Miller, J.A., Yao, Z., Nguyen, T.N., Aevermann, B., Barkan, E., Bertagnolli, D., Casper, T., Dee, N., et al. (2018). Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. PLoS One 13, e0209648.

Blümcke, I., Thom, M., Aronica, E., Armstrong, D.D., Bartolomei, F., Bernasconi, A., Bernasconi, N., Bien, C.G., Cendes, F., Coras, R., et al. (2013). International consensus classification of hippocampal sclerosis in temporal lobe epilepsy: a Task Force report from the ILAE Commission on Diagnostic Methods. Epilepsia 54, 1315–1329.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. 36, 411–420.

Cembrowski, M.S., Bachman, J.L., Wang, L., Sugino, K., Shields, B.C., and Spruston, N. (2016). Spatial gene-expression gradients underlie prominent heterogeneity of CA1 pyramidal neurons. Neuron 89, 351–368.

Cid, E., Marquez-Galera, A., Valero, M., Gal, B., Medeiros, D.C., Navarron, C.M., Ballesteros-Esteban, L., Reig-Viader, R., Morales, A.V., Fernandez-Lamo, I., et al. (2021). Sublayer- and cell-type-specific neurodegenerative transcriptional trajectories in hippocampal sclerosis. Cell Rep. 35, 109229.

Denisenko, E., Guo, B.B., Jones, M., Hou, R., de Kock, L., Lassmann, T., Poppe, D., Clément, O., Simmons, R.K., Lister, R., and Forrest, A.R.R. (2020). Systematic assessment of tissue dissociation and storage biases in single-cell and single-nucleus RNA-seq workflows. Genome Biol. 21, 130.

Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., and Jiang, Y. (2021). SCDC: bulk gene expression deconvolution by multiple single-cell

RNA sequencing references. Brief. Bioinform. 22, 416–427.

Du, R., Carey, V., and Weiss, S.T. (2019). deconvSeq: deconvolution of cell mixture distribution in sequencing data. Bioinformatics 35, 5095–5102.

Gjoneska, E., Pfenning, A.R., Mathys, H., Quon, G., Kundaje, A., Tsai, L.H., and Kellis, M. (2015). Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. Nature 518, 365–369.

Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat. Methods 14, 955–958.

Jew, B., Alvarez, M., Rahmani, E., Miao, Z., Ko, A., Garske, K.M., Sul, J.H., Pietiläinen, K.H., Pajukanta, P., and Halperin, E. (2020). Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nat. Commun. 11, 1971.

Lake, B.B., Codeluppi, S., Yung, Y.C., Gao, D., Chun, J., Kharchenko, P.V., Linnarsson, S., and Zhang, K. (2017). A comparative strategy for single-nucleus and single-cell transcriptomes confirms accuracy in predicted cell-type expression from nuclear RNA. Sci. Rep. 7, 6031.

van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. J. Mach. Learn. Res. 9, 2579–2605.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: uniform manifold approximation and projection for dimension reduction. ArXiv, arXiv:1802.03426.

Nido, G.S., Dick, F., Toker, L., Petersen, K., Alves, G., Tysnes, O.B., Jonassen, I., Haugarvoll, K., and Tzoulis, C. (2020). Common gene expression signatures in Parkinson's disease are driven by changes in cell composition. Acta Neuropathol. Commun. 8, 55.

Patrick, E., Taga, M., Ergun, A., Ng, B., Casazza, W., Cimpean, M., Yung, C., Schneider, J.A., Bennett, D.A., Gaiteri, C., et al. (2020). Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. PLoS Comput. Biol. 16, e1008120.

Rusina, E., Bernard, C., and Williamson, A. (2021). The kainic acid models of temporal lobe epilepsy. eNeuro 8, ENEURO.0337-20.2021.

Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive integration of single-cell data. Cell 177, 1888–1902.e21.

Thrupp, N., Sala Frigerio, C., Wolfs, L., Skene, N.G., Fattorelli, N., Poovathingal, S., Fourne, Y., Matthews, P.M., Theys, T., Mancuso, R., et al. (2020). Single-nucleus RNA-seq is not suitable for detection of microglial activation genes in humans. Cell Rep. 32, 108189.

Tsoucas, D., Dong, R., Chen, H., Zhu, Q., Guo, G., and Yuan, G.C. (2019). Accurate estimation of cell-type composition from gene expression data. Nat. Commun. 10, 2975.

Valero, M., Cid, E., Averkin, R.G., Aguilar, J., Sanchez-Aguilera, A., Viney, T.J., Gomez-Dominguez, D., Bellistri, E., and de la Prida, L.M. (2015). Determinants of different deep and superficial CA1 pyramidal cell dynamics during sharp-wave ripples. Nat. Neurosci. 18, 1281–1290.

Wang, X., Park, J., Susztak, K., Zhang, N.R., and Li, M. (2019). Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nat. Commun. 10, 380.

Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell 184, 3222–3241.

Zeisel, A., Muñoz-Manchado, A.B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. (2015). Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. Science 347, 1138–1142.

Zeisel, A., Hochgerner, H., Lönnerberg, P., Johnsson, A., Memic, F., van der Zwan, J., Häring, M., Braun, E., Borm, L.E., La Manno, G., et al. (2018). Molecular architecture of the mouse nervous system. Cell 174, 999–1014.e22.