

Introduction to Data Mining Publicly Available Single-Cell RNA-Sequencing Datasets

BMB Seminar

John Klement

June 4th, 2021

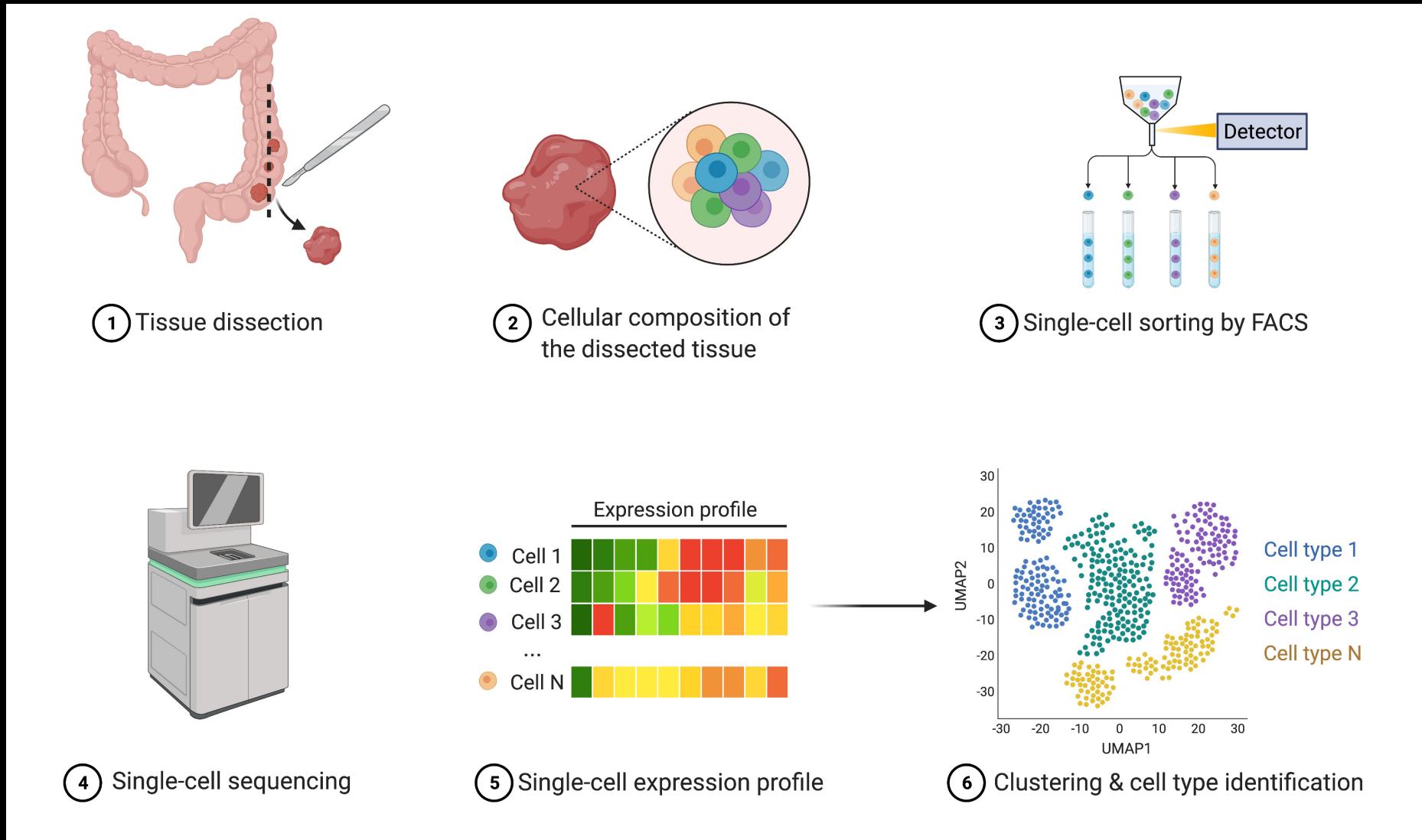
Outline

- High Level
 - How scRNAseq works
 - Recent ‘Add-ons’ to single-cell analysis
 - Sources of data
 - Seurat Platform
- Low level
 - Walkthrough using Colon HCA scRNAseq data
 - Acquiring Data
 - Preprocessing
 - Analysis
 - Generating Figures

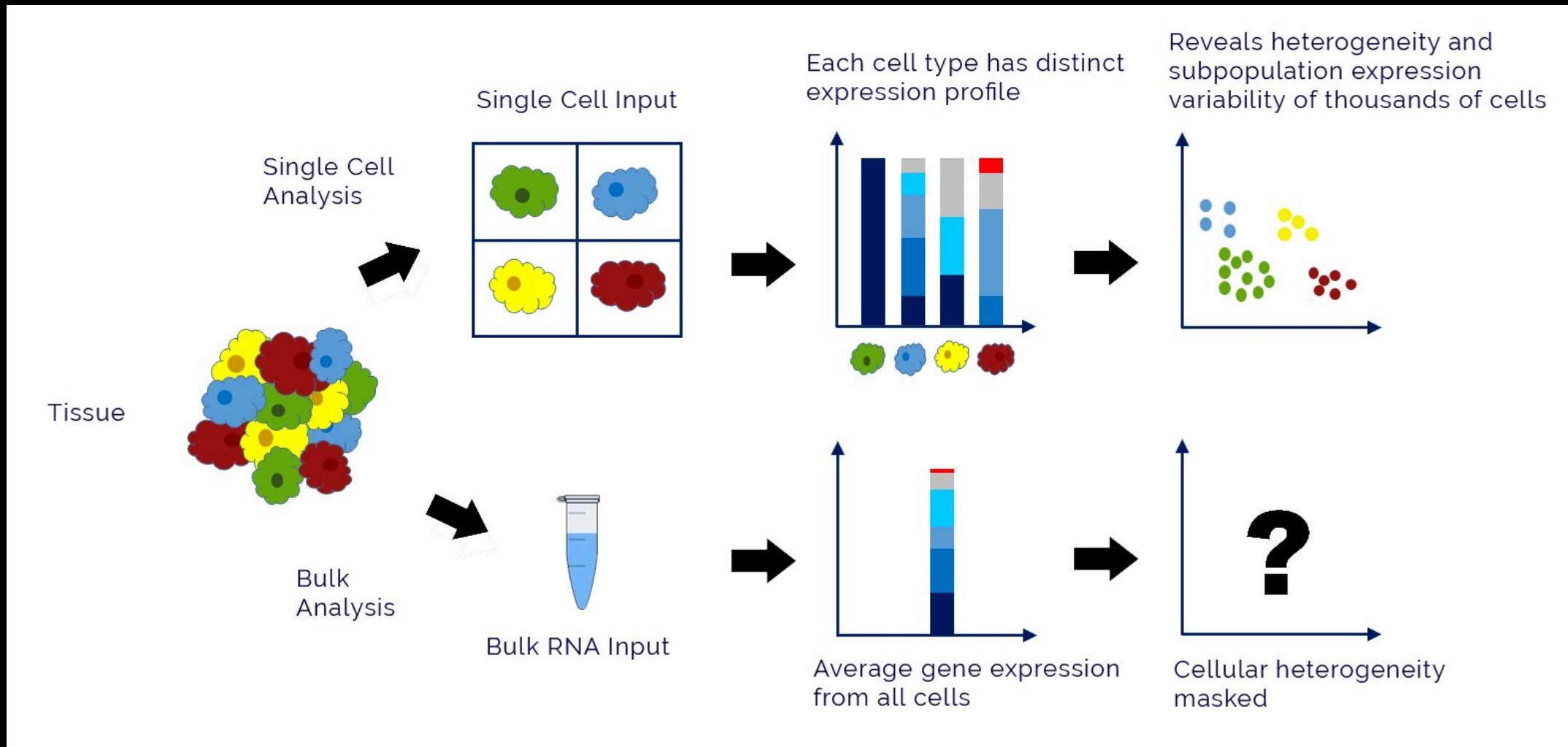
Misconceptions

- *I need a supercomputer to do any real analysis*
- *It will take days to analyze a single dataset*
- *I need to have extensive coding experience*
- *There's not a dataset relevant to the question that I am asking*

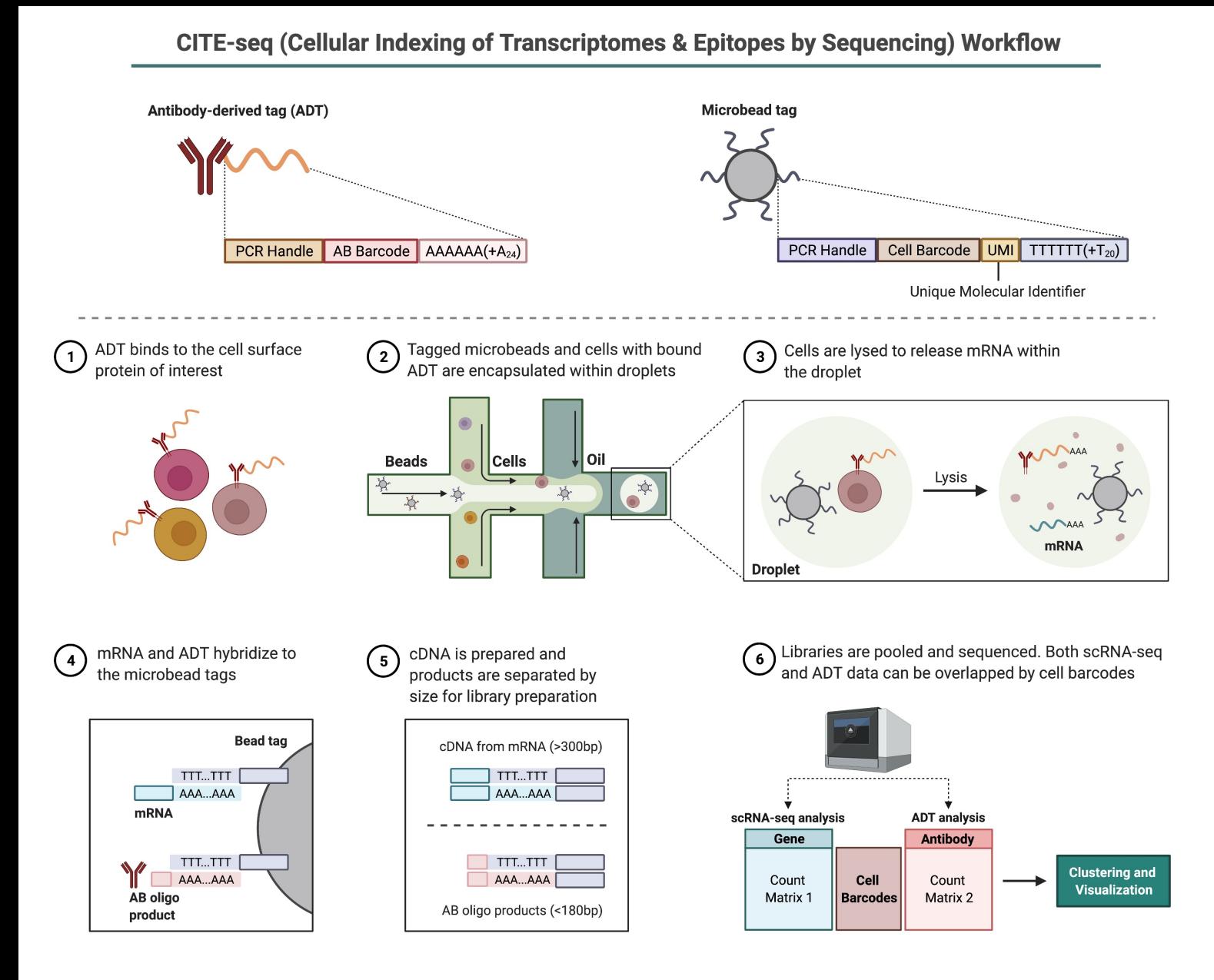
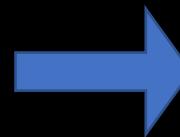
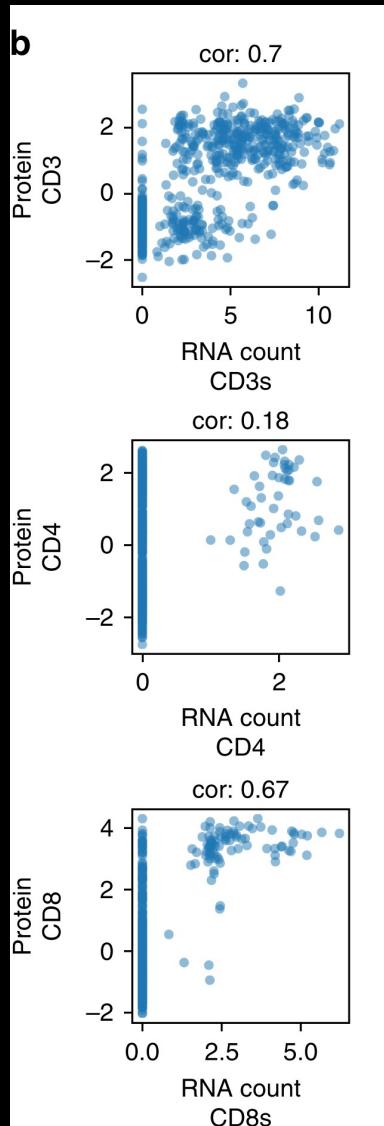
Brief Review of scRNASeq



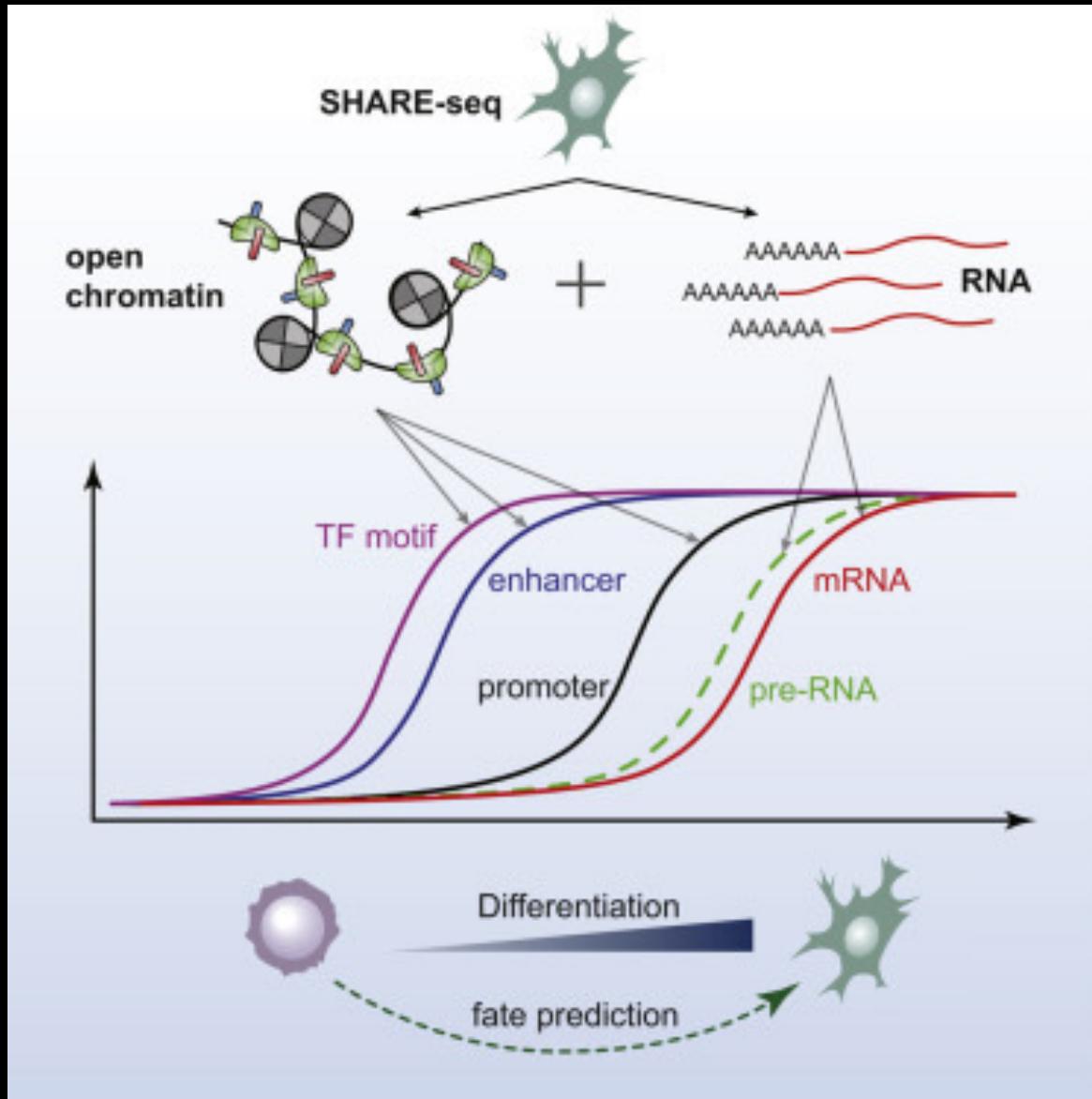
Why is scRNAseq important?



CITE/REAP-seq



Epigenetic scRNAseq/SHARE-Seq



ScRNAseq Drawbacks

- Cost
- Tissue Procurement/Isolation
- Appropriate QC is often complex

Public scRNAseq datasets are
widely available and easily
accessible

Human Cell Atlas

12.2M Cells

ALL CELLS



Kidney

Liver

Lung

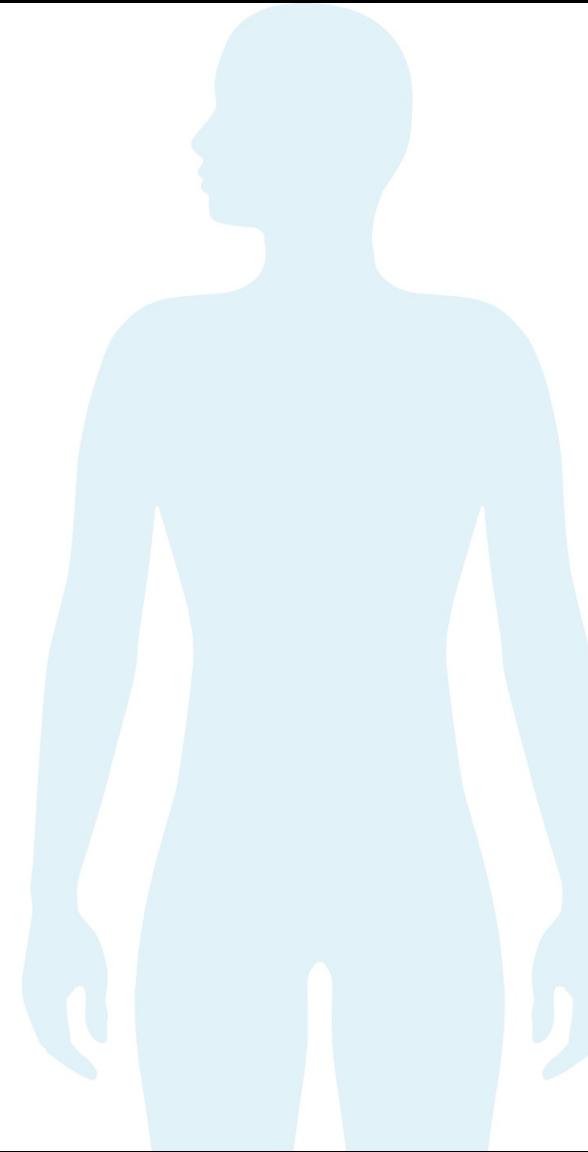
Heart

Immune System

Skin



View More Organs



247
LABS

Using NIH GEO

- Access: <https://www.ncbi.nlm.nih.gov/geo>

The screenshot shows the NIH Gene Expression Omnibus (GEO) homepage. At the top left is the "Gene Expression Omnibus" logo. Below it is a brief description: "GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles." To the right is the GEO logo and a search bar labeled "Keyword or GEO Accession" with a "Search" button. A red circle with a white slash icon is positioned above the search bar, pointing towards it with a yellow arrow.



From Manuscripts

Data availability

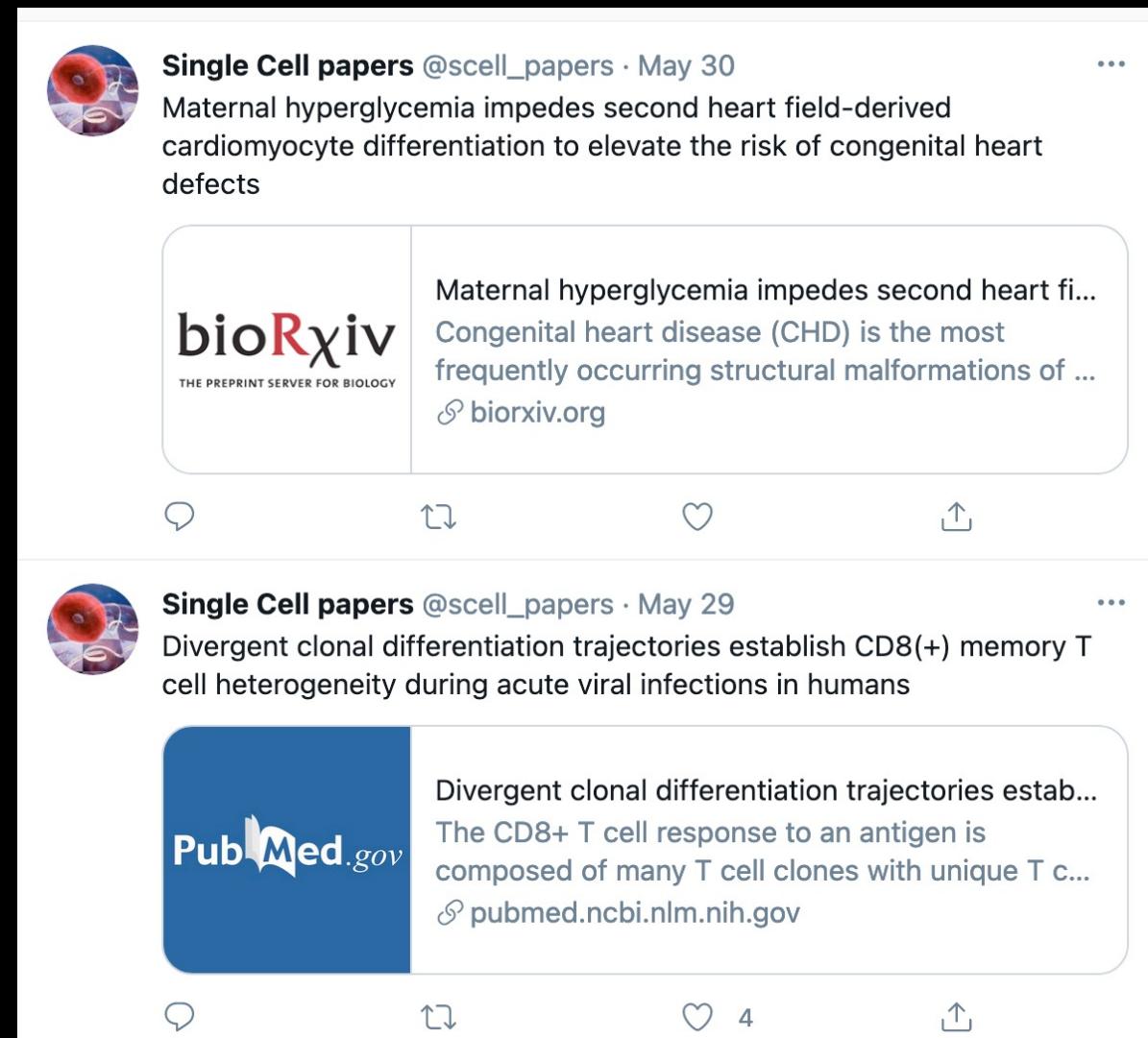
The scRNA-seq data generated for this study have been deposited in ArrayExpress under [E-MTAB-8656](#). The reference genome sequence was downloaded from Ensembl (http://www.ensembl.org/Mus_musculus) and used for alignment of the scRNA-seq data. To evaluate stem cell priming, scRNA-seq data were obtained from the Single Cell Portal (https://portals.broadinstitute.org/single_cell/study/small-intestinal-epithelium) and used to define gene sets for differentiated sub-lineages of epithelial cells. The lists of marker genes used to annotate types of epithelial, mesenchymal and immune cells in Fig. [3b, f](#) and Extended Data Figs. [6d, 8a–c](#) are given in Supplementary Tables [1, 3](#). Gene sets used in Fig. [3d, e](#) and Extended Data Fig. [7a, b, k–m](#) are provided in Supplementary Table [2](#). Source data are provided with this paper.

Other Useful Sources

- https://twitter.com/scell_papers



A screenshot of a Twitter profile page for the account @scell_papers. The profile picture shows a red blood cell. The bio reads: "A twitter bot account for papers employing sequencing of single cells". The location is listed as Vienna, Austria, and the account was joined in March 2015. It has 27 following and 8,527 followers. A note indicates it is followed by IdoAmitLab, CZI Science, and 5 others. Navigation tabs at the bottom include Tweets (underlined), Tweets & replies, Media, and Likes.



The image shows two tweets from the account @scell_papers. The first tweet, dated May 30, discusses maternal hyperglycemia and its impact on heart development. It includes a logo for bioRxiv and a link to biorxiv.org. The second tweet, dated May 29, discusses divergent clonal differentiation trajectories in CD8+ memory T cells during viral infections. It includes a logo for PubMed.gov and a link to pubmed.ncbi.nlm.nih.gov. Both tweets have engagement icons below them.

Single Cell papers @scell_papers · May 30
Maternal hyperglycemia impedes second heart field-derived cardiomyocyte differentiation to elevate the risk of congenital heart defects

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Maternal hyperglycemia impedes second heart fi...
Congenital heart disease (CHD) is the most frequently occurring structural malformations of ...
biorxiv.org

Single Cell papers @scell_papers · May 29
Divergent clonal differentiation trajectories establish CD8(+) memory T cell heterogeneity during acute viral infections in humans

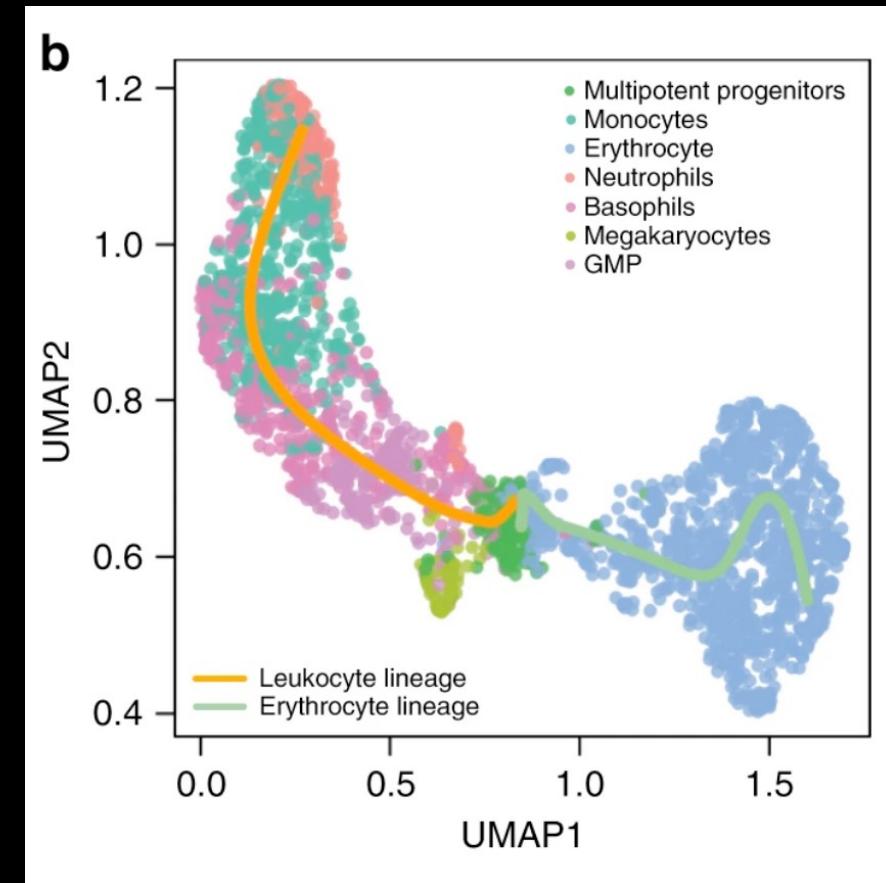
PubMed.gov

Divergent clonal differentiation trajectories estab...
The CD8+ T cell response to an antigen is composed of many T cell clones with unique T c...
pubmed.ncbi.nlm.nih.gov

How to analyze

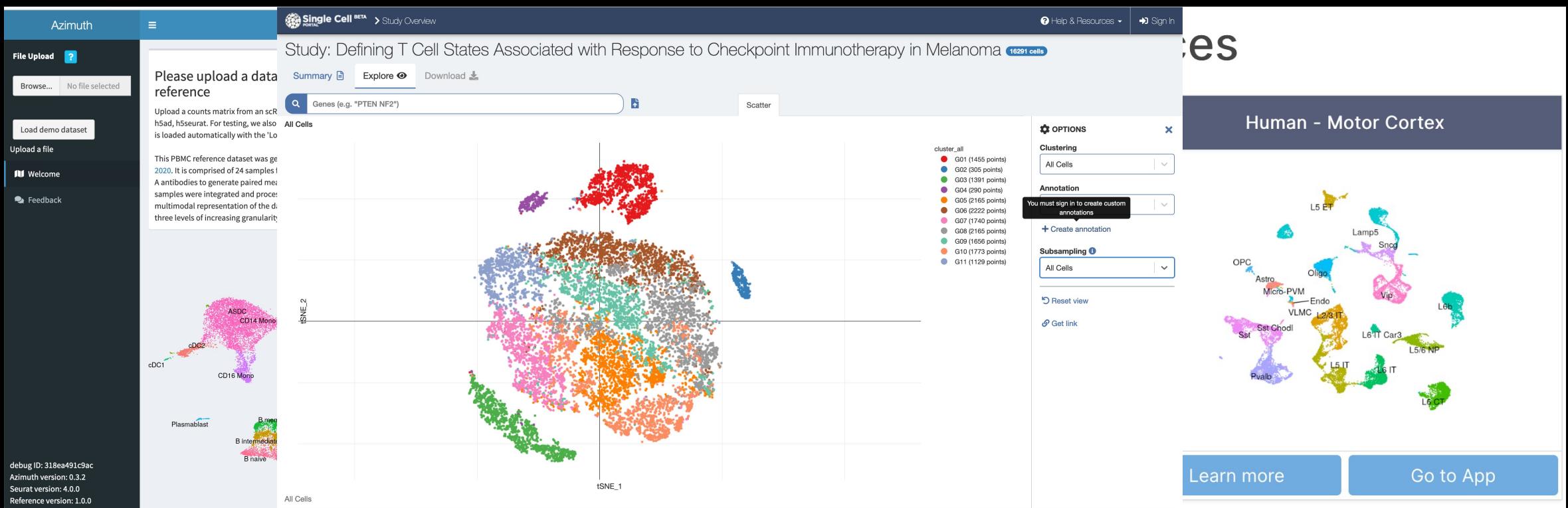
- Seurat
- Main alternative: Monocle

The screenshot shows the Seurat 4.0.2 website. The top navigation bar includes links for "Seurat 4.0.2", "Install", "Get started", "Vignettes", "Extensions", "FAQ", "News", "Reference", and "Archive". The main content area features a large, colorful UMAP plot of single-cell genomics data. Overlaid on the plot is the text "SEURAT" in large, bold, white letters, accompanied by a small yellow Eiffel Tower icon. Below the plot, the text "R toolkit for single cell genomics" is visible. At the bottom of the page, the text "Official release of Seurat 4.0" is displayed.



How to analyze

- Web-based platforms
 - Azimuth: <https://azimuth.hubmapconsortium.org/>
 - Single Cell Portal : https://singlecell.broadinstitute.org/single_cell



Step-by-step walkthrough

- Installing Software
- Downloading datafiles
 - Human Cell Atlas
 - NIH GEO
- Setting up workspace in RStudio
- Inputting downloaded datafiles
- Analyzing scRNAseq data with Seurat
- Generating Figures

Installing Software

- Minimum Hardware Requirements

- 16GB RAM
- i5 Processor

1. [Install R](#)
2. [Install Rstudio](#)
3. [Install Rtools](#)
 - Windows only



- R Learning Resources

- [Coursera](#)
- [EdX](#)
- [Melbourne](#)
- [Data Science for Immunologists \(book\)](#)

Downloading Files - HCA

- Access the HCA Data Portal
- Tabula Muris
 - Large Intestine

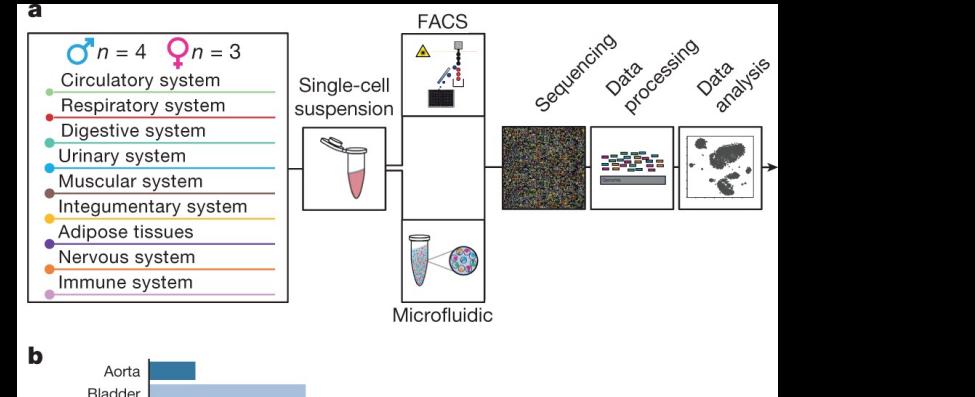
Article | Published: 03 October 2018

FACS_matrices.zip

Mus musculus

adip
aort
blac
bon
brai
diap
heal
kidr
larg
liver
lung
mar
mus
pan
skin
sple
thyro
tonc
trac

Project Information
Project Metadata
Project Matrices
External Resources



Tabula Muris: Transcriptomic characterization of 20 organs and tissues from *Mus musculus* at single cell resolution

Select Project

Project Information

Description

We have created a compendium of single cell transcriptome data from the model organism *Mus musculus* comprising more than 100,000 cells from 20 organs and tissues. These data represent a new resource for cell biology, revealing gene expression in poorly characterized cell populations and allowing for direct and controlled comparison of gene expression in cell types shared between tissues, such as T-lymphocytes and endothelial cells from distinct anatomical locations. Two distinct technical approaches were used for most tissues: one approach, microfluidic droplet-based 3'-end counting, enabled the survey of thousands of cells at relatively low coverage, while the other, FACS-based full length transcript analysis, enabled characterization of cell types with high sensitivity and coverage. The cumulative data provide the foundation for an atlas of transcriptomic cell biology.

Downloading Files - GEO

 NCBI  Gene Expression Omnibus

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO
NCBI > GEO > Accession Display [?](#) Not logged in | Login [?](#)

Scope: Format: Amount: GEO accession: GSE109774

Series GSE109774 Query DataSets for GSE109774

Status	Public on Mar 19, 2018
Title	Tabula Muris: Transcriptomic characterization of 20 organs and tissues from Mus musculus at single cell resolution
Organism	Mus musculus
Experiment type	Expression profiling by high throughput sequencing
Summary	We have created a resource of single cell transcriptome data from the model organism <i>Mus musculus</i> . Contributor: The Tabula Muris Consortium The full list of contributors to this dataset can be found in the corresponding publication.
Overall design	Single cell RNA sequencing of single cells across 20 tissues of 3 month aged mice
Citation(s)	Tabula Muris Consortium., Overall coordination., Logistical coordination., Organ collection and processing. et al. Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris. <i>Nature</i> 2018 Oct;562(7727):367-372. PMID: 30283141

Supplementary file	Size	Download	File type/resource
GSE109774_Bladder.tar.gz	114.9 Mb	(ftp)(http)	TAR
GSE109774_Brain_Microglia.tar.gz	315.3 Mb	(ftp)(http)	TAR
GSE109774_Brain_Neurons.tar.gz	382.0 Mb	(ftp)(http)	TAR
GSE109774_Colon.tar.gz	288.8 Mb	(ftp)(http)	TAR
 GSE109774_Fat.tar.gz	395.2 Mb	(ftp)(http)	TAR
GSE109774_Heart.tar.gz	468.4 Mb	(ftp)(http)	TAR
GSE109774_Kidney.tar.gz	55.2 Mb	(ftp)(http)	TAR
GSE109774_Liver.tar.gz	65.7 Mb	(ftp)(http)	TAR
GSE109774_Lung.tar.gz	128.3 Mb	(ftp)(http)	TAR
GSE109774_Mammary.tar.gz	183.4 Mb	(ftp)(http)	TAR
GSE109774_Marrow.tar.gz	364.7 Mb	(ftp)(http)	TAR
GSE109774_Muscle.tar.gz	139.6 Mb	(ftp)(http)	TAR
GSE109774_Pancreas.tar.gz	135.1 Mb	(ftp)(http)	TAR
GSE109774_RAW.tar	374.5 Mb	(http)(custom)	TAR (of TAR)
GSE109774_Skin.tar.gz	169.2 Mb	(ftp)(http)	TAR
GSE109774_Spleen.tar.gz	112.8 Mb	(ftp)(http)	TAR
GSE109774_Thymus.tar.gz	103.5 Mb	(ftp)(http)	TAR
GSE109774_Tongue.tar.gz	102.4 Mb	(ftp)(http)	TAR
GSE109774_Trachea.tar.gz	95.9 Mb	(ftp)(http)	TAR
GSE109774_list_of_SRR_accessions_and_raw_filenames.txt.gz	774.0 Kb	(ftp)(http)	TXT

Installing Software Packages in R

- General script for installing given package
 - `install.packages('NAME OF PACKAGE')`
- Some packages require downloading and compilation from Github
 - Github: online code repository

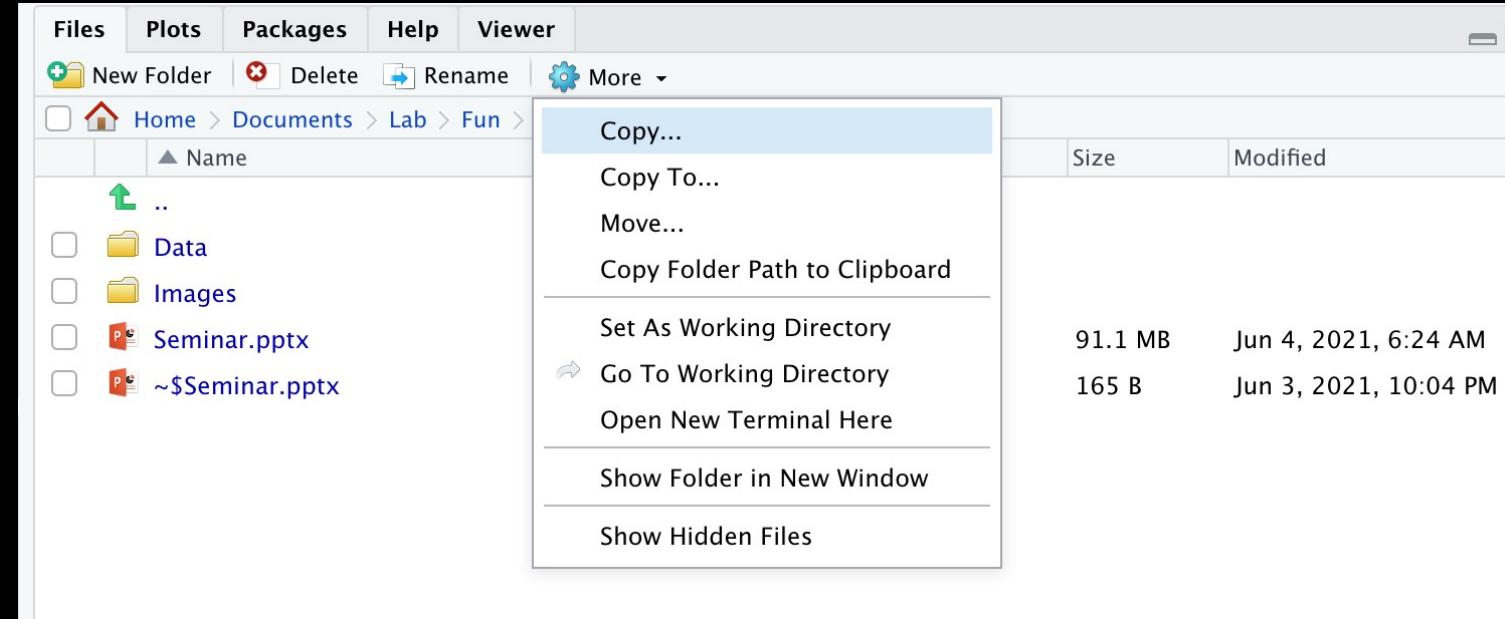
```
#install packages available from CRAN
install.packages('Seurat', 'tidyverse', 'msigdbr', 'ggpubr')

#install CellID from bioconductor
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
##The following initializes usage of Bioc devel
BiocManager::install(version='devel')
BiocManager::install("CellID")
```

Set up environment

- Load Packages
 - `library(PACKAGE NAME)`
- Set Workspace
 - Simplest way is to navigate using RStudio

```
#load packages in workspace
library(Seurat)
library(tidyverse)
library(CellID)
library(ggpubr)
library(msigdbr)
```

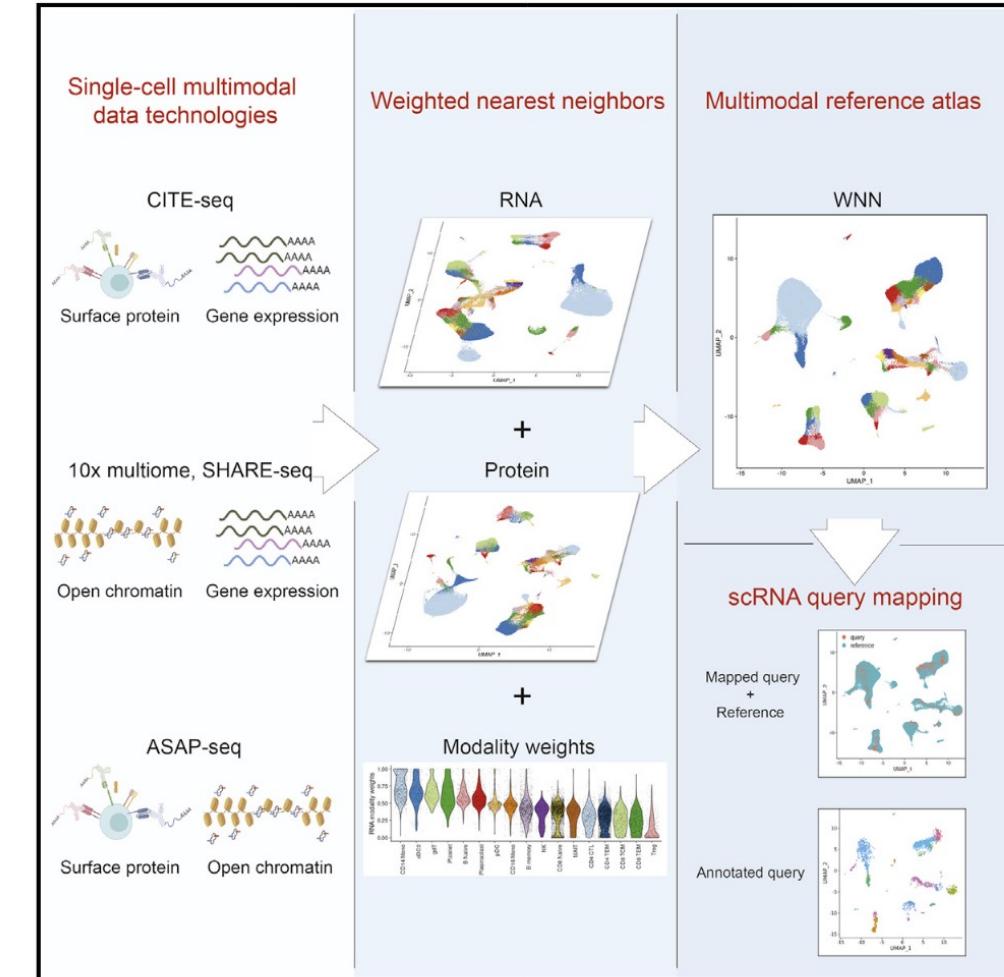


Seurat

- [Website](#)
- [Installation](#)
- [Tutorial](#)

Integrated analysis of multimodal single-cell data

Graphical abstract



Authors

Yuhan Hao, Stephanie Hao,
Erica Andersen-Nissen, ...,
Raphael Gottardo, Peter Smibert,
Rahul Satija

Correspondence

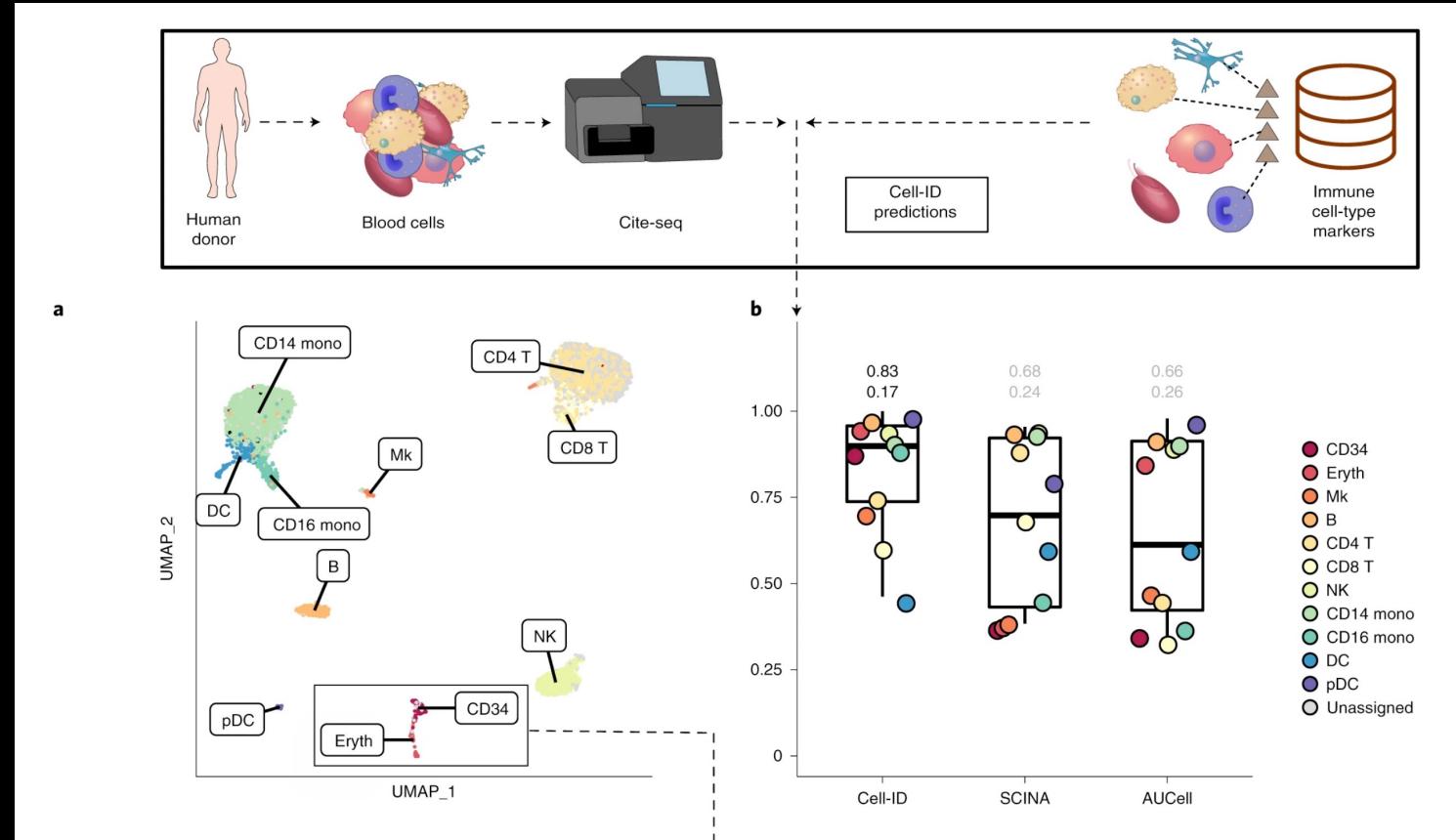
rsatija@nygenome.org (R.S.),
smibertp@gmail.com (P.S.)

In brief

A framework that allows for the integration of multiple data types using single cells is applied to understand distinct immune cell states, previously unidentified immune populations, and to interpret immune responses to vaccinations.

Cell-ID

- Repository
 - Contains installation instructions, documentation, etc.
- Allows for:
 - Identification of cell-type at single-cell, rather than cluster, level
 - Easy method to score cells for gene signature expression



Preparing data files for input

- Creation of Seurat object requires two datasets

- Gene expression matrix

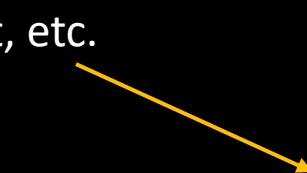
- Rows: Cells
 - Columns: Genes

- Cell Metadata

- Rows: Cells
 - Columns: Information re: identity of cell
 - E.g. tissue, identity, treatment, etc.



	<i>Irf8</i>	<i>Myc</i>	<i>Hyal1</i>
Cell.1	5	2	1
Cell.2	0	0	2
Cell.3	3	4	6

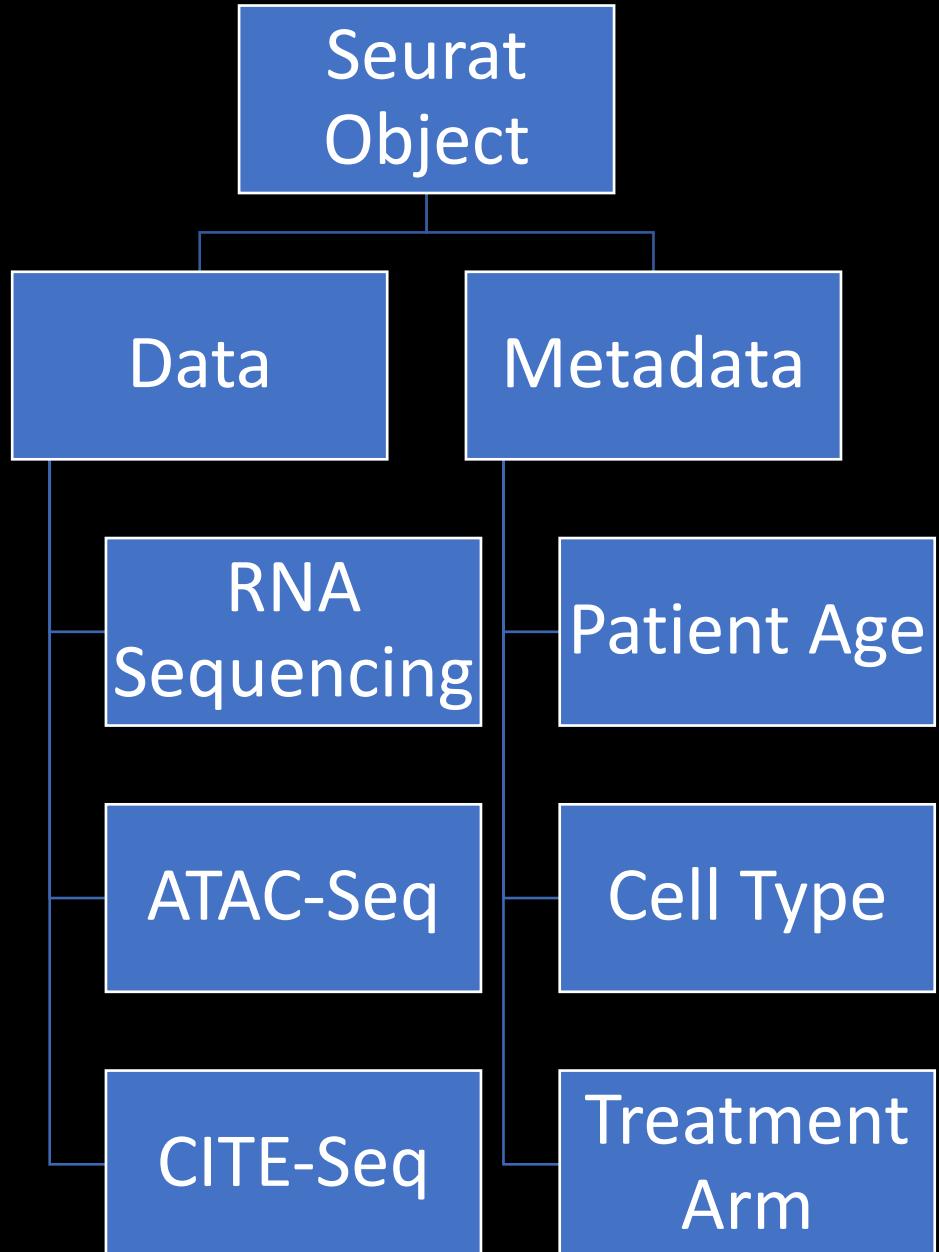


	<i>Age</i>	<i>Gender</i>	<i>Tissue</i>
Cell.1	54	M	C
Cell.2	44	F	B
Cell.3	36	F	C

Seurat Objects

- Contain both assay and metadata
 - Assay: RNA, protein, DNA, etc.
 - Metadata: age, cell source, etc.
 - Accessible with \$ operator: Colon\$age

```
#Create Seurat object
##read in Colon expression dataset; note read_csv used as file csv
colon <- read_csv("FACS/Large_Intestine-counts.csv")
##set cell to rowname
colon <- tibble::column_to_rownames(colon, var = "X1")
##filter out rownames that do not encode for protein in mice
colon <- colon[rownames(colon) %in% MgProteinCodingGenes,]
##read in metadata
colon.anno <- read_csv("annotations_facs.csv")
##set cell to rowname
colon.anno <- tibble::column_to_rownames(colon.anno, var = "cell")
##Create seurat object, filtering out genes that are not expressed in >=5 cells,
##or cells that do not express at least 200 genes
colon <- CreateSeuratObject(counts = colon, meta.data = colon.anno,
                           min.cells = 5, min.features = 200)
```

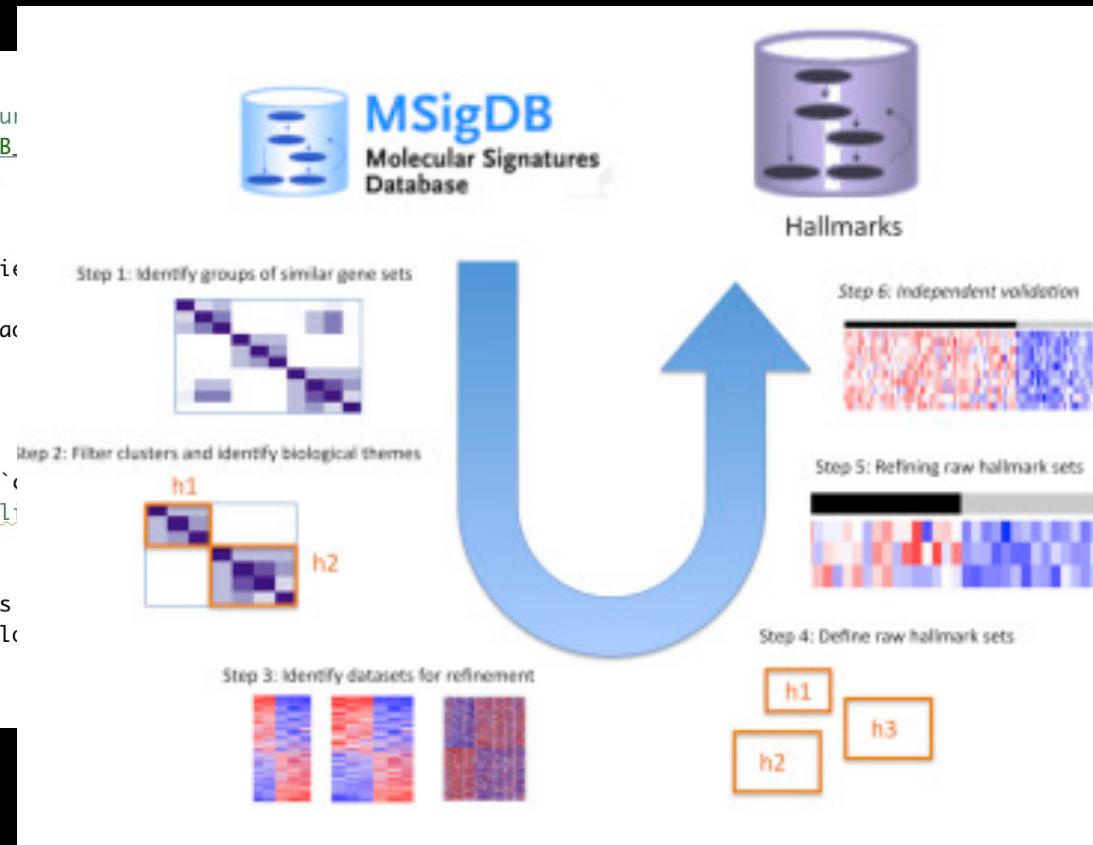


Analyze Seurat Object

```
#Process files according to standard Seurat pipeline, scaling out mitochondrial reads
#so as not to be used in defining variable features for downstream analysis
colon <- colon %>%
  Seurat::NormalizeData() %>%
  Seurat::FindVariableFeatures() %>%
  Seurat::ScaleData(vars.to.regress = c("percent.mt"))
colon <- Seurat::RunPCA(colon, ncs = 100, features = VariableFeatures(colon))
##Use Elbowplot to find 'inflection' point in PCs to use as cutoff
ElbowPlot(colon, ndims = 100)
##Number of dims = number of PCs used; if you wanted to use first 50 PCs, would imput 1:50 instead of 1:30
colon <- FindNeighbors(colon, dims = 1:30)
##Increasing resolution will increase the number of clusters identified by Seurat
colon <- FindClusters(colon, resolution = 0.5)
colon <- RunUMAP(colon, dims = 1:30)
```

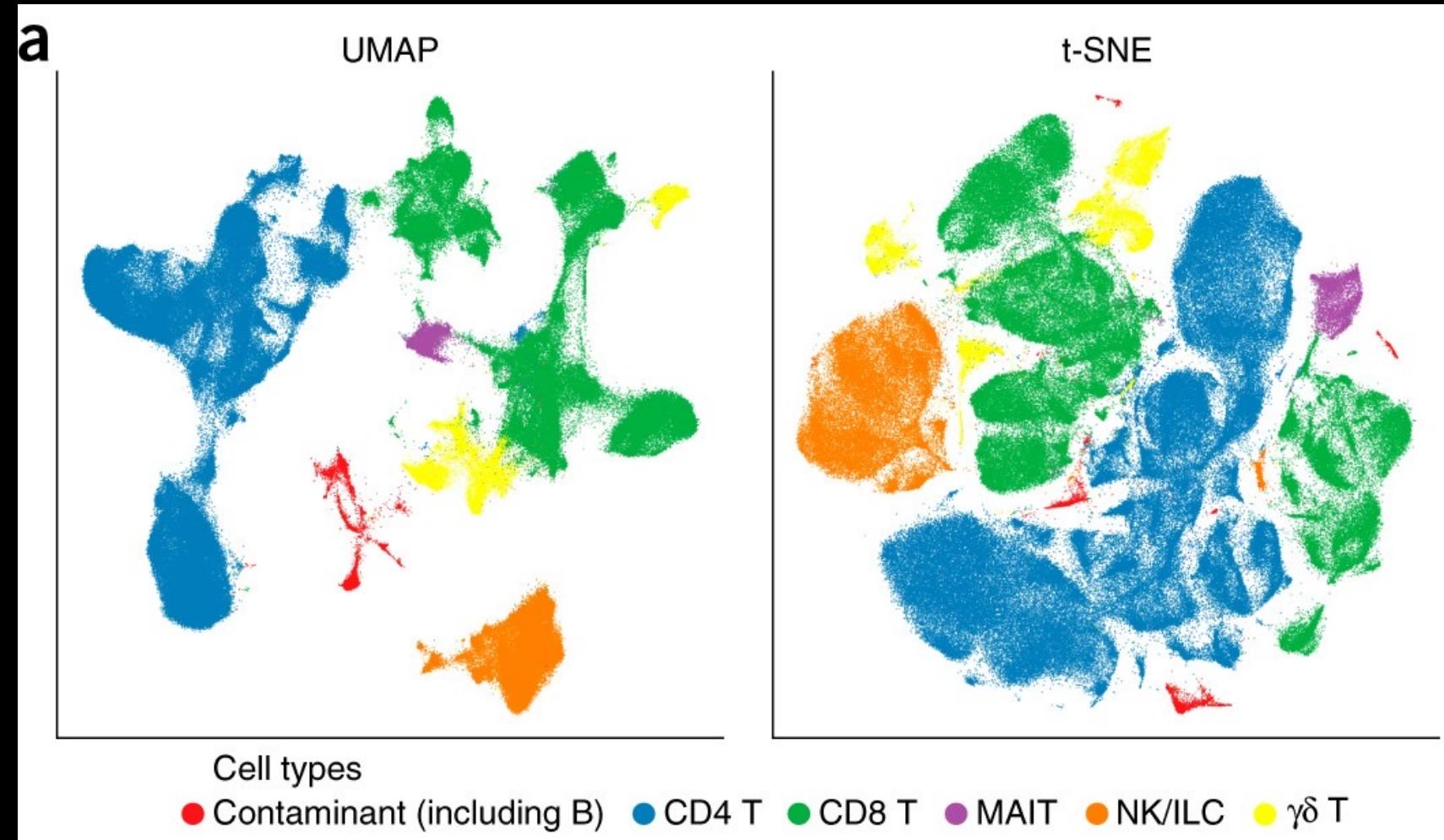
Annotate Cell Identity

```
#Score cells for HALLMARK signatures
##downloads hallmark signatures
msigdf <- msigdbr::msigdbr(species = "mouse")
##creates hallmark signatures in Seurat object
HallmarkDF <- split(x = msigdf$genes, 
                      HallmarkDF <- as.list(HallmarkDF)
##scores cells against hallmark
HGT_Hallmark <- RunCellHGT(color = "Hallmark")
##loads scores back into Seurat
colon@assays[["Hallmark"]] <- CreateSeuratObject(
  DefaultAssay(colon) <- "Hallmark"
  #label cells with CellID
  ##reads in Panglao dataset containing cell specific signatures
  panglao <- read_tsv("https://panglaodb.se/markers/PanglaoDB")
  ##filtering dataset for signatures associated for GI tract
  panglao_colon <- panglao %>% filter(organ == "GI tract")
  ##restricting to mouse specific genes
  panglao_colon <- panglao_colon %>% filter(str_detect(species, "mouse"))
  ##ensuring signatures are calculated same way as our genes
  panglao_colon$`official gene symbol` <- str_to_title(panglao_colon$`official gene symbol`)
  ##converts dataframe to list
  panglao_colon <- panglao_colon %>%
    group_by(`cell type`) %>%
    summarise(geneset = list(`official gene symbol`))
  colon_gs <- setNames(panglao_colon$geneset, panglao_colon$`cell type`)
  ##creates MCA (multiple correspondence analysis dimensional reduction)
  colon <- RunMCA(colon)
  ##Runs prediction algorithm
  HGT_colon_gs <- RunCellHGT(colon, pathways = colon_gs, dims = 2)
  colon_gs_prediction <- rownames(HGT_colon_gs)[apply(HGT_colon_gs, 1, sum)]
  ##adds annotations to Seurat object
  colon$colon_gs_prediction <- colon_gs_prediction
```

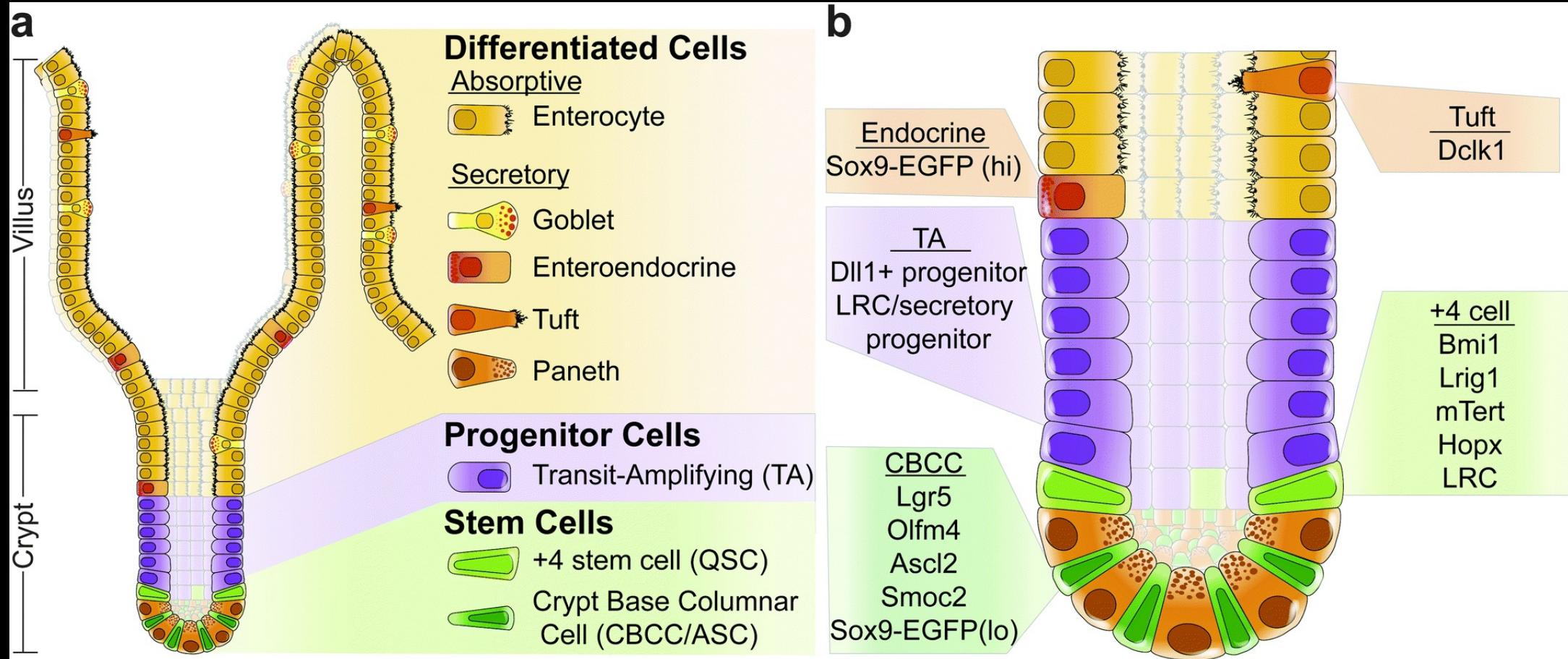


What is UMAP?

- **UMAP:** Uniform Manifold Approximation and Projection
 - X/Y axis: Dimensional Representations
 - Dot: Cell
 - Clusters: Similar cells

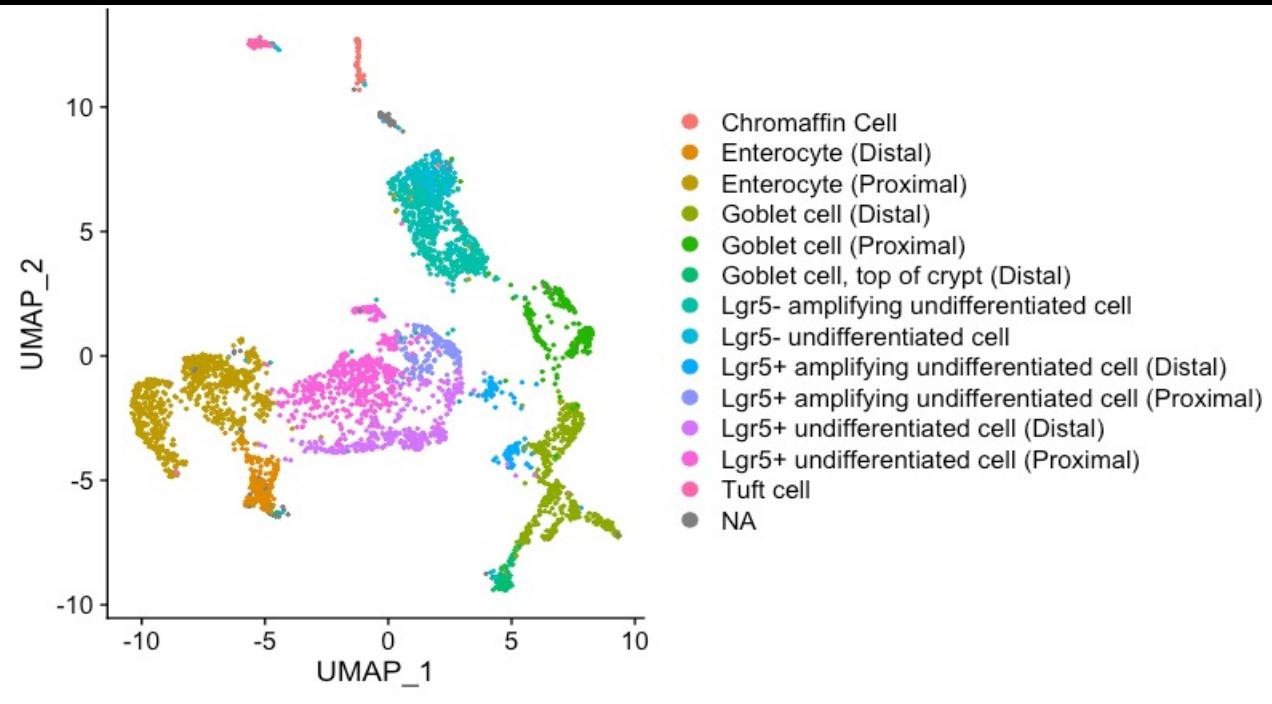


Colon Crypt Organization

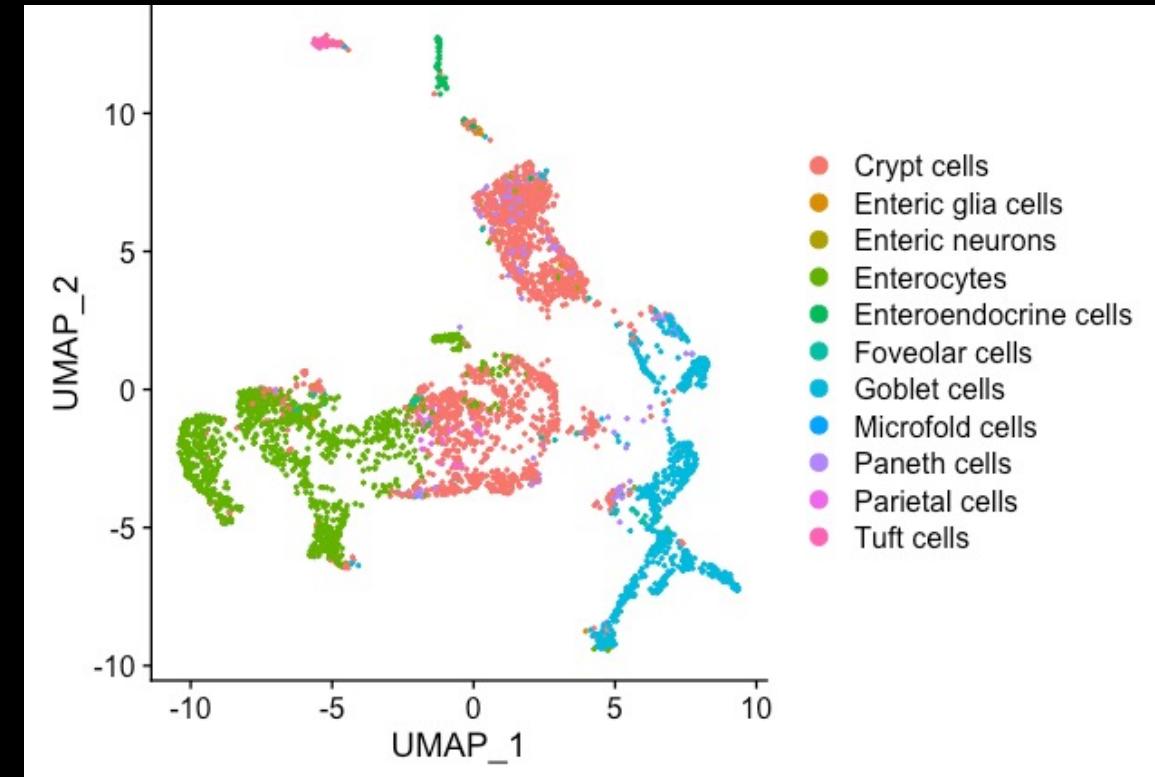


Analyze Cell Identity

Tabula Muris Annotations



Cell-ID Annotations



Analyze Gene Expression

Article

Cell Reports

Myeloid-Derived Suppressor Cells Produce IL-10 to Elicit DNMT3b-Dependent IRF8 Silencing to Promote Colitis-Associated Colon Tumorigenesis

Graphical Abstract

The graphical abstract illustrates the study's findings. It starts with 'DSS Chronic Inflammation' leading to 'MDSCs'. MDSCs produce 'IL-10', which acts on 'STAT3'. STAT3 activation leads to the recruitment of 'DNMT3b + DNMT1' to the DNA promoter region between -1000 and +1000 relative to the transcription start site ('+'). This results in the silencing of the 'IRF8' gene, represented by a red 'X' over the gene name. The silencing of IRF8 leads to the development of 'CAC' (Colon Adenocarcinoma). A bar chart at the bottom shows the expression levels of DNMT3b and DNMT1 across the promoter region.

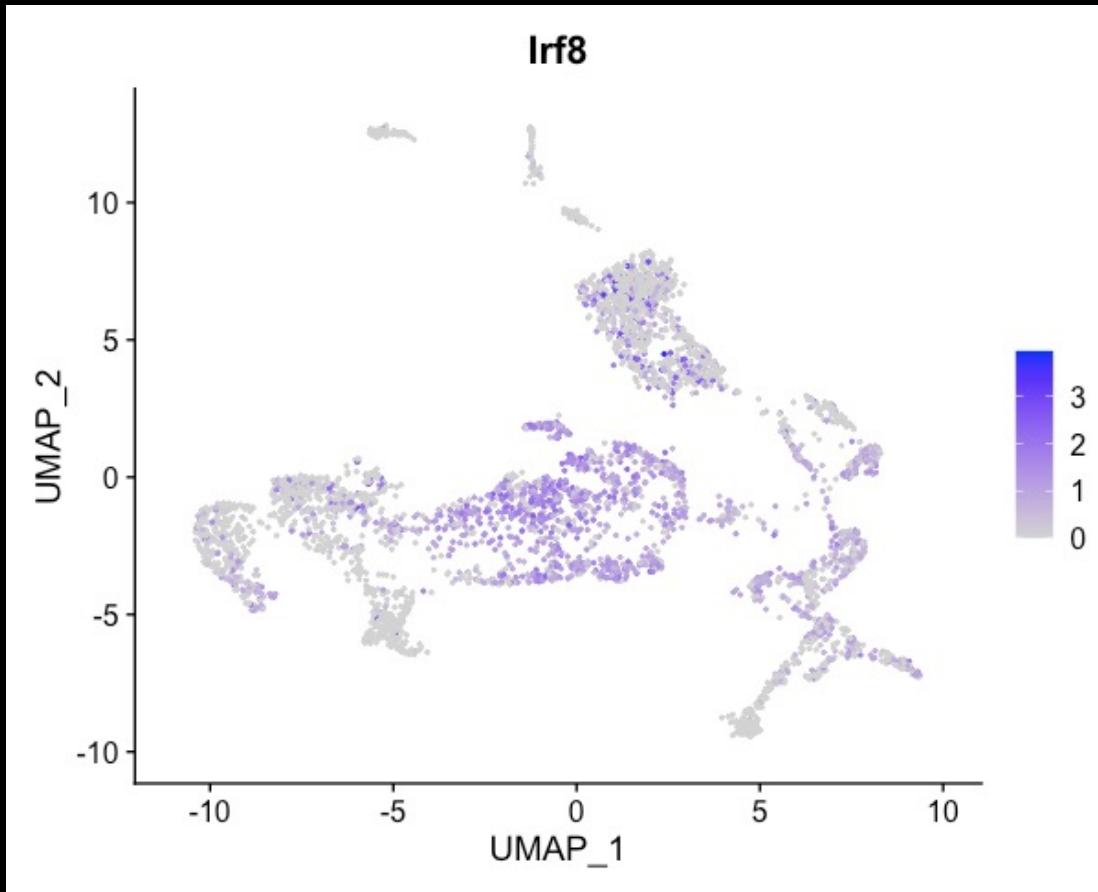
Authors
Mohammed L. Ibrahim, John D. Klement, Chunwan Lu, ..., Phillip J. Buckhaults, Herbert C. Morse III, Kebin Liu

Correspondence
kliu@augusta.edu

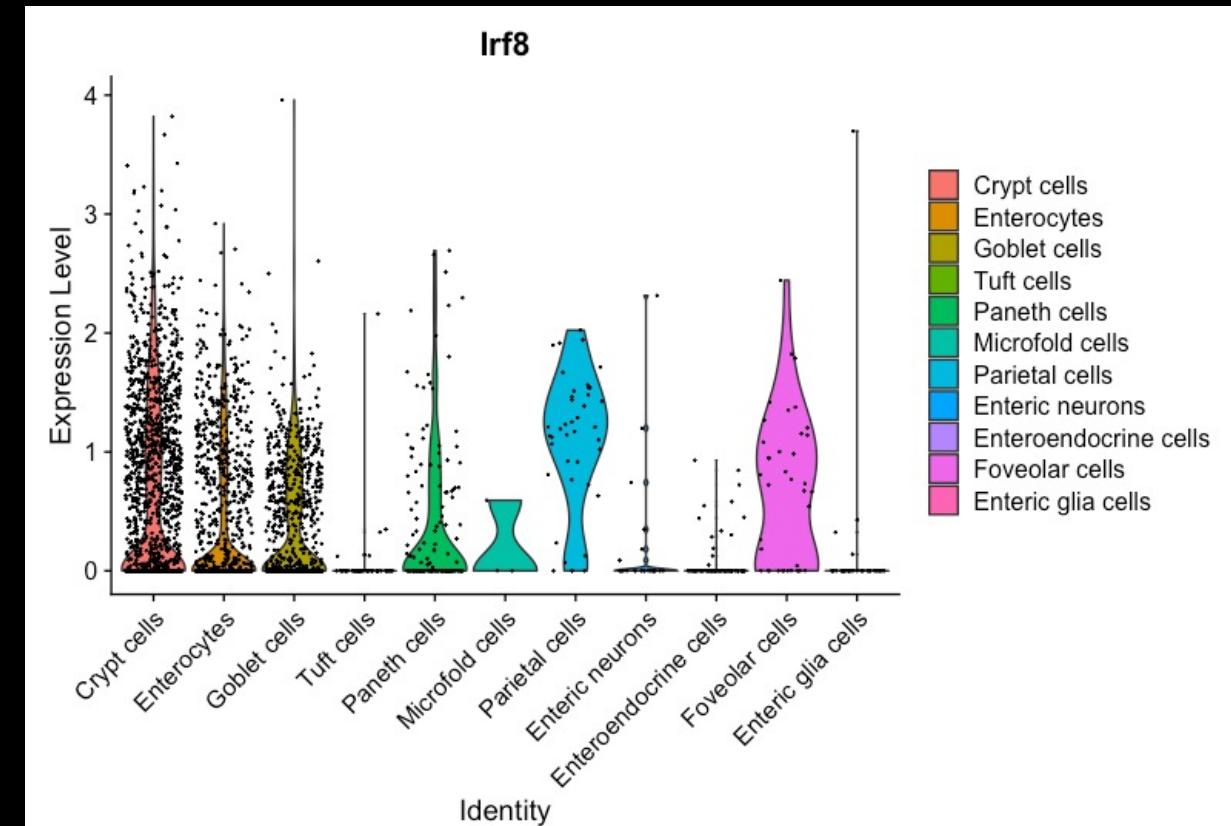
In Brief
Ibrahim et al. report that chronic inflammation induces colonic accumulation of myeloid-derived suppressor cells (MDSCs) that upregulates IL-10. IL-10 directly regulates STAT3 activation to upregulate DNMT3b to silence tumor suppressor IRF8 in colonic epithelial cells. The MDSC-IL-10-STAT3-DNMT3b-IRF8 pathway links chronic inflammation to colon cancer initiation.

How does IRF8 mediate its function as a tumor suppressor?

Where is IRF8 expressed?



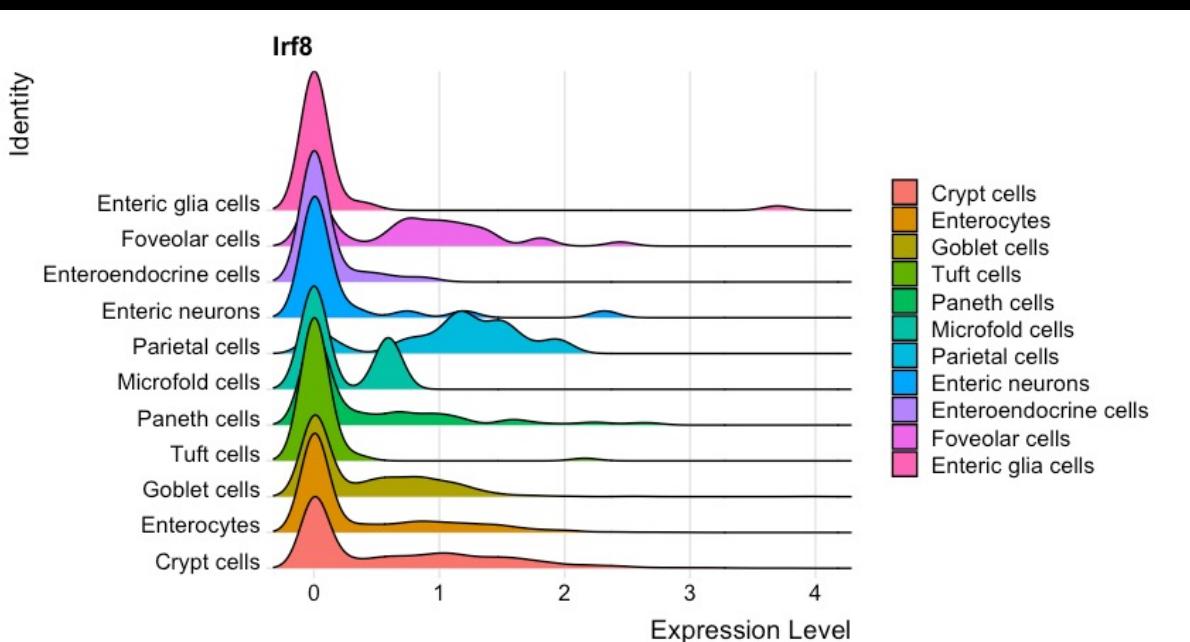
```
FeaturePlot(colon, "Irf8")
```



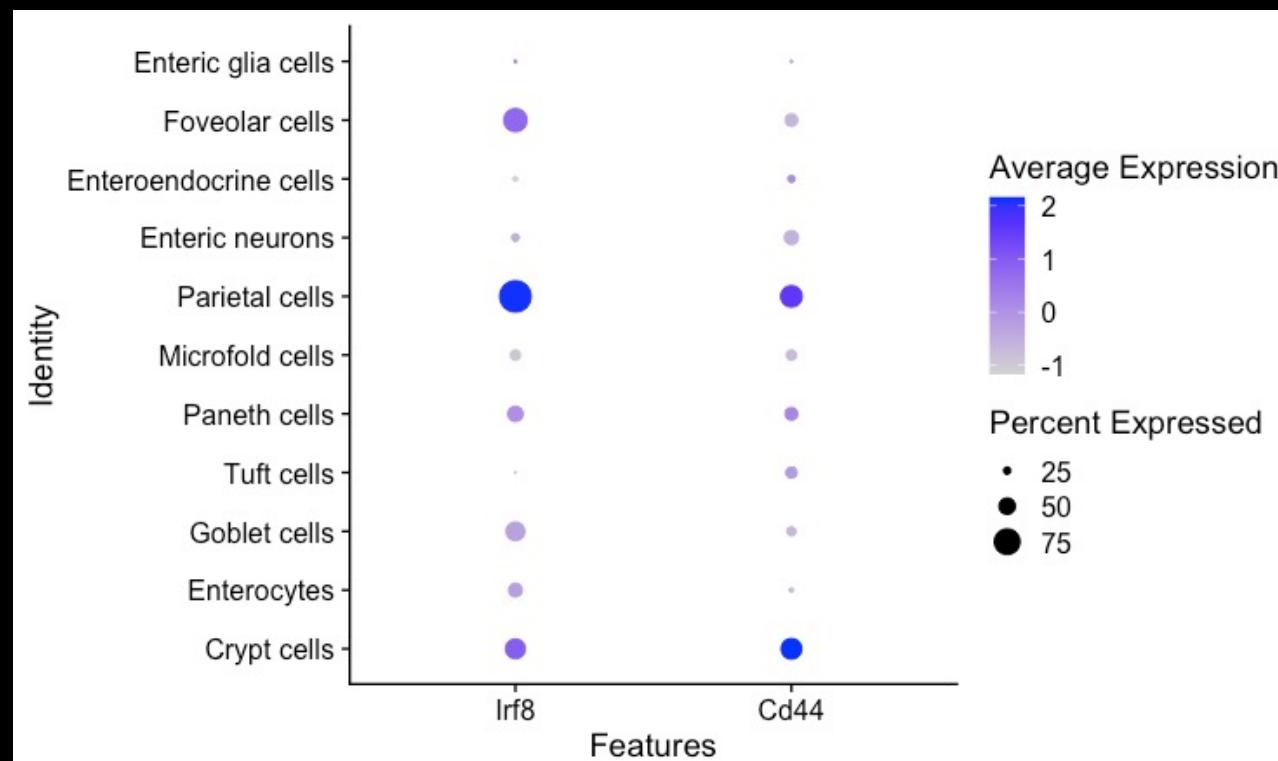
```
VlnPlot(colon, "Irf8")
```

- Crypt cells
- Enterocytes
- Goblet cells
- Tuft cells
- Paneth cells
- Microfold cells
- Parietal cells
- Enteric neurons
- Enteroendocrine cells
- Foveolar cells
- Enteric glia cells

Where is IRF8 expressed?

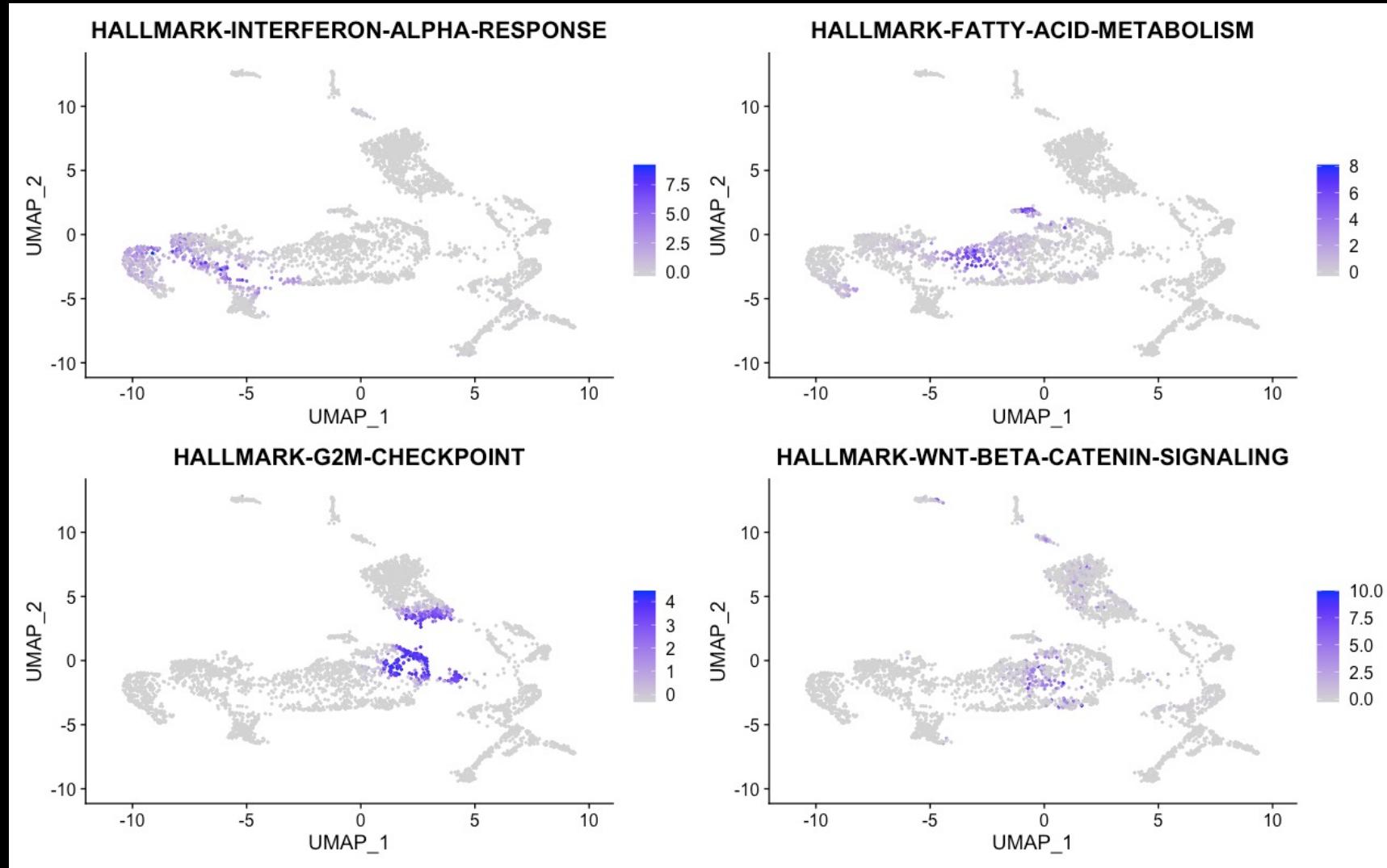


```
RidgePlot(colon, "Irf8")
```



```
DotPlot(colon, features = c("Irf8", "Cd44"))
```

Analyze Gene Signature Expression

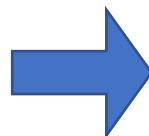


Is there is a difference between IRF8+ and
IRF8- crypt cells?



10 lines of code...

```
#Vln Plot differences  
Idents(colon) <- "colon_gs_prediction"  
DefaultAssay(colon) <- "RNA"  
FeaturePlot(colon, "Irf8")  
entero <- subset(colon, idents = "Crypt Cells")  
irf8.mean <- mean(entero[["RNA"]])@data[["Irf8", ]]  
entero$irf8 <- entero[["RNA"]])@data[["Irf8", ]] > irf8.mean  
DefaultAssay(entero) <- "Hallmark"  
Idents(entero) <- "irf8"  
irf8.markers <- Seurat::FindMarkers(entero, ident.1 = T, ident.2 = F)  
DoHeatmap(entero, features = entero@assays[["Hallmark"]])@counts@Dimnames[[1]],  
  slot = "scale.data", label = F) + viridis::scale_fill_viridis()
```

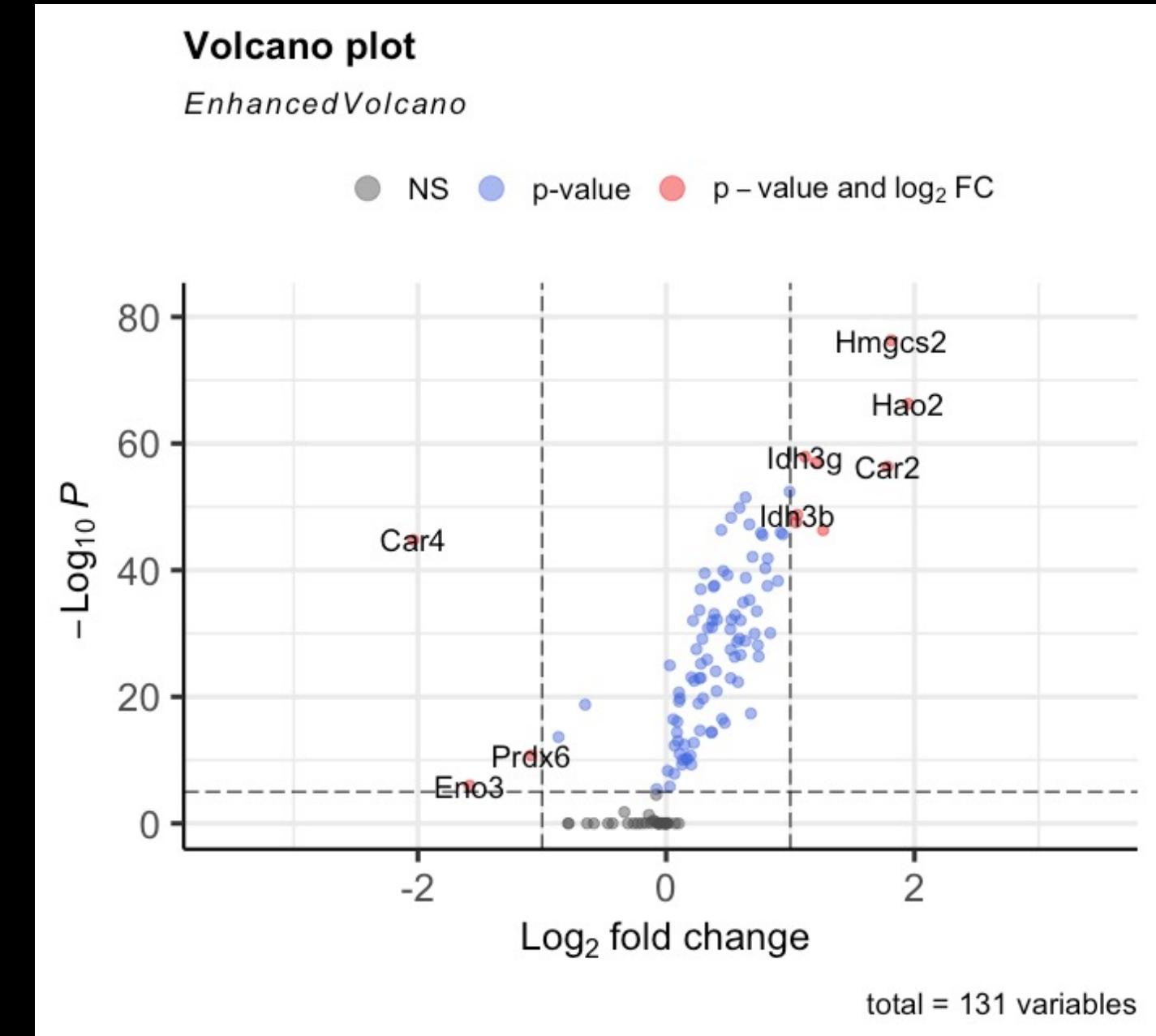


IRF8-LOW

IRF8-HIGH



What genes drive the difference in FA Oxidation?



Misconceptions

- *I need a supercomputer to do any real analysis*
 - Most analyses can be run on desktop computers that we use in lab
- *It will take days to analyze a single dataset*
 - Generally can be accomplished in minutes to hours
- *I need to have extensive coding experience*
 - Use of pre-made packages in R make data analysis and visualization easy
- *There's not a dataset relevant to the question that I am asking*
 - With hundreds of human and mouse datasets, it's likely that there is an easily accessible dataset suitable for exploratory analysis

PowerPoint and all script used
to generate figures

<https://github.com/jdklement/Seminar-210604>

Email: jklement@augusta.edu

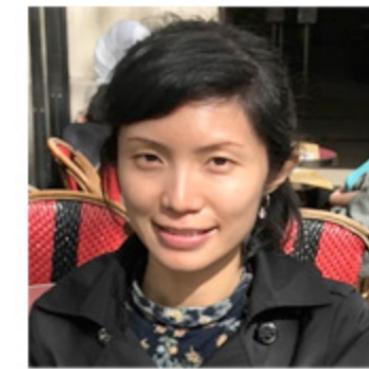
Thank you!



Dr. Kebin Liu, PI



Dafeng Yang
Lab Manager



Chunwan Lu
Assistant Res Scientist



John D. Klement
MD-Ph.D Student



Alyssa D. Smith
Ph.D Student



Dakota S. Booth
MD-Ph.D Student