

Single-cell RNA sequencing (scRNA-seq) data analysis tutorial



Ashok Kumar Sharma, Ph.D.
Research Bioinformatician II
Digestive and Liver Diseases
Cedars Sinai Medical Center



@ashoks773



@ashoks773



@sharma-ak

cedars-sinai.org

Original Study: Single-cell RNA-seq analysis of stromal vascular cells (SVCs) isolated from primary adipose tissue.

Study Workflow:

This study investigates the translocation of viable gut microbiota to mesenteric adipose tissue and its role in the formation of creeping fat (CrF) in Crohn's disease (CD). By performing single-cell RNA sequencing (scRNA-seq) on stromal vascular cells isolated from primary adipose tissue, the study compares healthy and inflamed tissue from Crohn's disease patients to explore microbial impacts on tissue remodeling.

Key Insights

The study identifies **Clostridium innocuum** as a key microbe translocating to mesenteric adipose tissue in CD patients, triggering immune activation and adipose tissue remodeling. Single-cell analysis reveals CrF as pro-fibrotic and pro-adipogenic, with a rich immune response to microbial signals. This was further validated in gnotobiotic mice colonized with **C. innocuum**.

Impact:

This research provides a novel mechanistic link between microbial translocation and the development of creeping fat in CD, revealing potential therapeutic targets for managing this extra-intestinal manifestation of the disease. The study also opens new avenues for understanding the microbial influence on chronic inflammation and tissue remodeling in Crohn's disease.

Outline

- Part1: Download dataset including metadata and create Seurat Object
- **Part2: Perform QC and filtering**
- **Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering**
- **Part4: Cell Type Annotations using Known Markers and singleR**
- **Part5: Marker identification – Cluster specific and/or disease specific (if any)**
- **Part6: Pathway enrichment analysis**
- **Part7: Cell-Cell Interaction/Communication Analysis**

Recourses: Datasets and tutorial links to be followed during this course

Dataset

BioProject: PRJNA659007

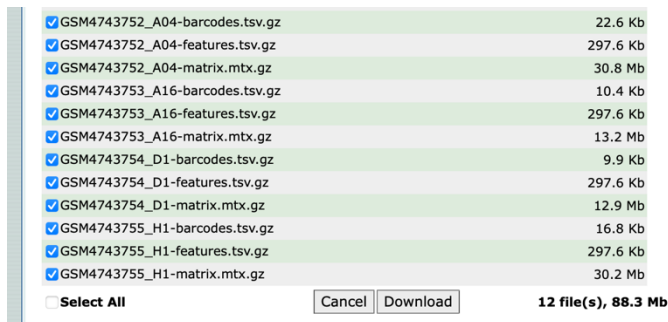
- Raw data: "GSE156776"
- <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156776>
- **Chromium samples:** 2 Crohn's disease patients, 2 non-IBD controls.
- **Final GitHub:**

Detailed public tutorials

- **scRNA-seq analysis course:**
<https://www.singlecellcourse.org/introduction-to-single-cell-rna-seq.html>
- **For cell type assignments:**
<https://bioconductor.org/books/release/SingleRBook/>
- **For cell cell Communication analysis**
<https://rpubs.com/HHJ/921311>

Download scRNA seq Data and create Seurat Object

<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156776>



<input checked="" type="checkbox"/> GSM4743752_A04-barcodes.tsv.gz	22.6 Kb
<input checked="" type="checkbox"/> GSM4743752_A04-features.tsv.gz	297.6 Kb
<input checked="" type="checkbox"/> GSM4743752_A04-matrix.mtx.gz	30.8 Mb
<input checked="" type="checkbox"/> GSM4743753_A16-barcodes.tsv.gz	10.4 Kb
<input checked="" type="checkbox"/> GSM4743753_A16-features.tsv.gz	297.6 Kb
<input checked="" type="checkbox"/> GSM4743753_A16-matrix.mtx.gz	13.2 Mb
<input checked="" type="checkbox"/> GSM4743754_D1-barcodes.tsv.gz	9.9 Kb
<input checked="" type="checkbox"/> GSM4743754_D1-features.tsv.gz	297.6 Kb
<input checked="" type="checkbox"/> GSM4743754_D1-matrix.mtx.gz	12.9 Mb
<input checked="" type="checkbox"/> GSM4743755_H1-barcodes.tsv.gz	16.8 Kb
<input checked="" type="checkbox"/> GSM4743755_H1-features.tsv.gz	297.6 Kb
<input checked="" type="checkbox"/> GSM4743755_H1-matrix.mtx.gz	30.2 Mb
<input type="checkbox"/> Select All	
Cancel Download	12 file(s), 88.3 Mb

Download **files** for each sample:

- barcodes.tsv.gz
- features.tsv.gz
- matrix.mtx.gz

Download **SraRunTable.csv**

Create Seurat Object:

Read matrix, features and barcode files using **ReadMtx**

Check Seurat data structure:

`print(dim(CrF_merged_seurat))` : # of Genes * # of Cells

Check Associated metadata:

`unique(CrF_merged_seurat@meta.data$Patient_ID)` # Total 4 patient IDs

GEO_Accession	Patient_ID	condition	tissue_status
GSM4743752	A04	Crohn's disease	inflamed
GSM4743753	A16	Crohn's disease	inflamed
GSM4743754	D1	Non-IBD control	healthy
GSM4743755	H1	Non-IBD control	healthy

Outline

- Part1: Select study, download dataset including metadata and create Seurat Object
- **Part2: Perform QC and filtering**
- Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- Part6: Pathway enrichment analysis
- Part7: Cell-Cell Interaction/Communication Analysis

Metrics can be used for Quality Filtering

Total counts

Reads or UMIs) detected in given **cell**/spot across all genes

Overall RNA content captured for a **cell**/spot



May indicate low RNA content or poor capture efficiency



May also suggest doublets or sequencing artifacts

N genes by counts

of Features with non-zero expression in given **cell**/spot

How many genes are actively expression in a **cell**/spot



May indicate dead or dying cells or empty spots in spatial data



Could indicate doublets (two **cells**/spots captured together)

Mitochondrial Genes (MT Genes):

% of genes are mitochondrial in a given **cell**/spot

MT % often correlates negatively with the UMI count

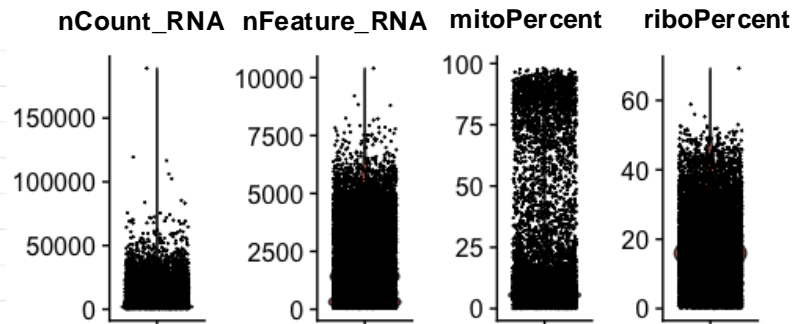
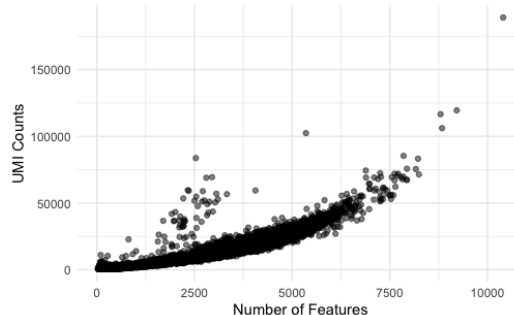
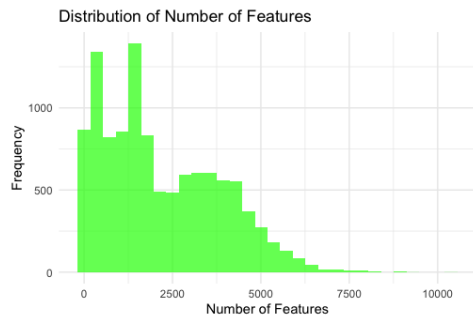
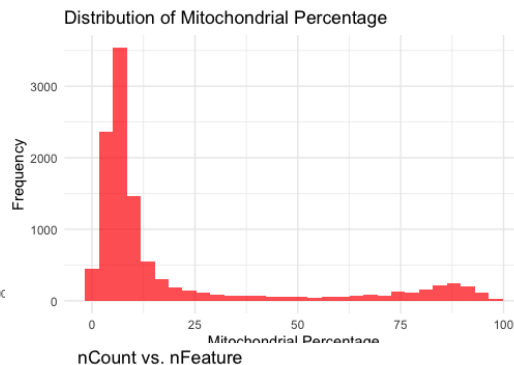
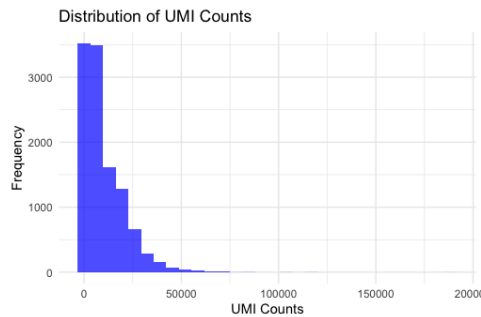


Lower percentage indicate high quality cells



Higher percentage usually indicates a lower quality cell

Explore QC metrics: by plotting them



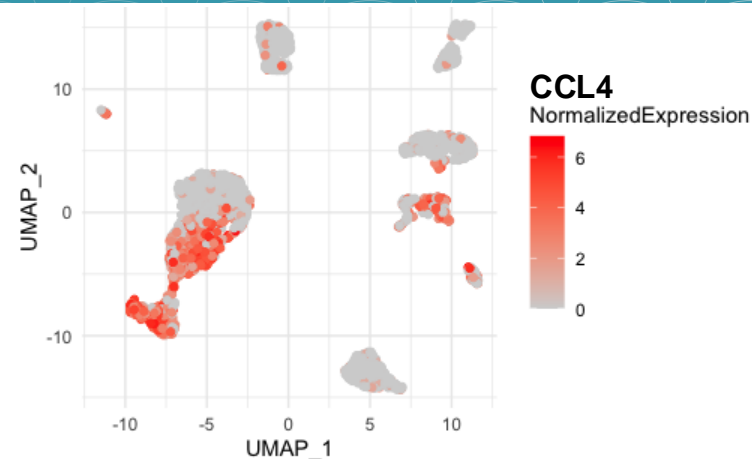
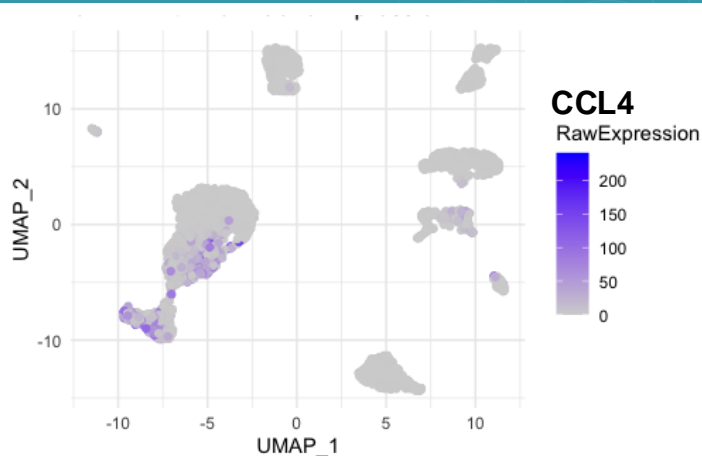
```
## Counts :: UMIs  
min_counts <- 200  
max_counts <- 5000  
## Features :: genes  
min_features <- 200  
max_features <- 7500  
## Mitochondrial %  
max_mito_ratio <- 20
```

Check Seurat data structure again after filtering:
`print(dim(merged_seurat_filtered))` : # of Genes * # of Cells

Outline

- Part1: Select study, download dataset including metadata and create Seurat Object
- Part2: Perform QC and filtering
- **Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering**
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- Part6: Pathway enrichment analysis
- Part7: Cell-Cell Interaction/Communication Analysis

Exploration of scRNA-seq Datasets: Check Effect of Normalization



Normalization: `NormalizeData()`

Default Normalization method: `LogNormalize`

$$\text{Normalized value} = \log_2 \left(\frac{\text{Feature Count}}{\text{Total Counts per Cell}} \times \text{Scaling Factor} + 1 \right)$$

CPM is similar to Seurat's `LogNormalize`, but without the logarithmic transformation. Other methods: CLR and RC

Scaling: `ScaleData()`

Function standardizes (or z-scores) the gene expression values for each gene across all cells

Exploration of scRNA-seq Datasets: PCA vs. UMAP: Key Concepts and Comparison

Principal Component Analysis (PCA):

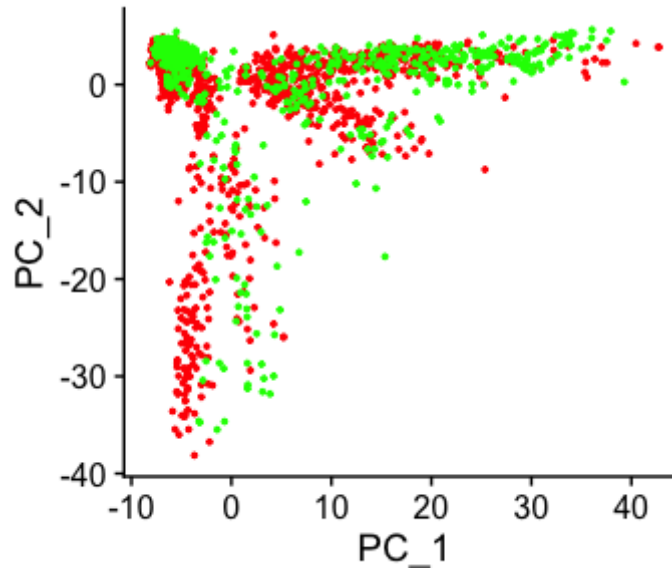
- Linear dimensionality reduction method.
- Preserves **global structure** in the data.
- Reduces data variance by projecting onto orthogonal axes (principal components).
- Useful for visualizing and interpreting data with **linear relationships**.
- **Limitations:** May fail to capture **nonlinear patterns** in complex datasets.

Uniform Manifold Approximation and Projection (UMAP):

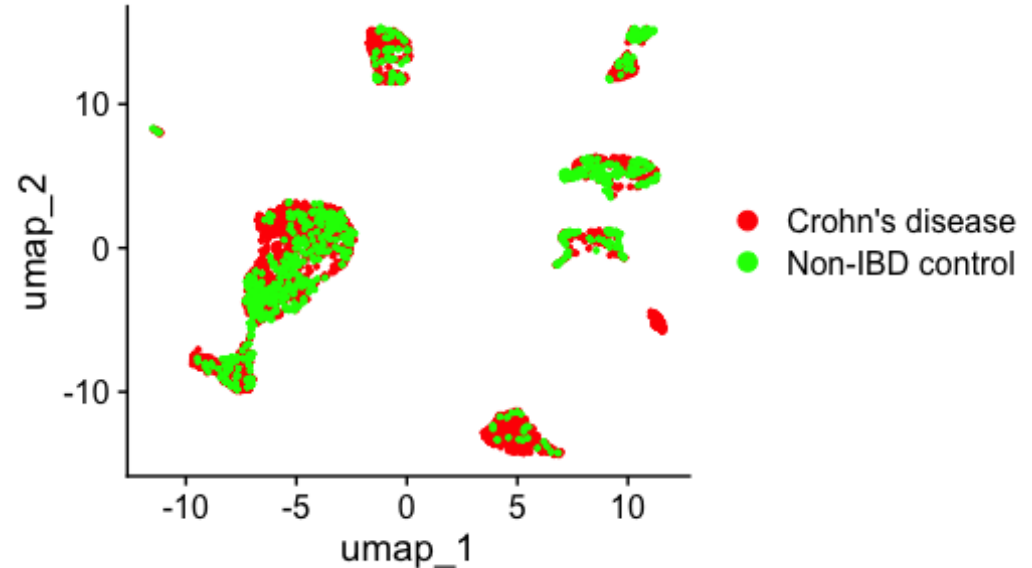
- Nonlinear dimensionality reduction method.
- Focuses on preserving **local structure** while maintaining some global structure.
- Constructs a graph based on distances and optimizes a lower-dimensional representation.
- Ideal for **complex, high-dimensional data**, such as single-cell RNA-seq.
- **Strength:** Can better capture relationships in noisy or non-linear data.

Exploration of scRNA-seq Dataset: PCA vs. UMAP: Key Concepts and Comparison

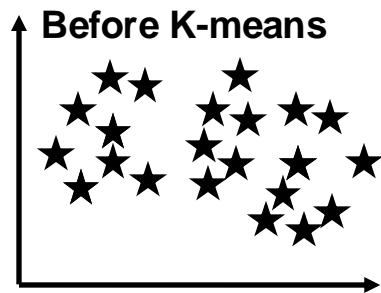
Principal Component Analysis (PCA):



Uniform Manifold Approximation and Projection (UMAP):



Biological Analysis – Clustering Basics



Methods:

Hierarchical clustering

k-means

Graph-based methods

Tools:

SINCERA

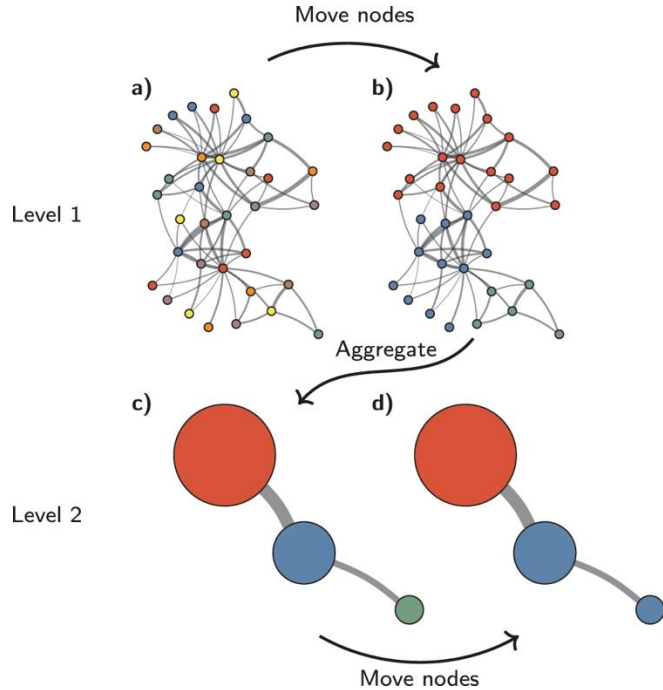
SC3

tSNE + k-means

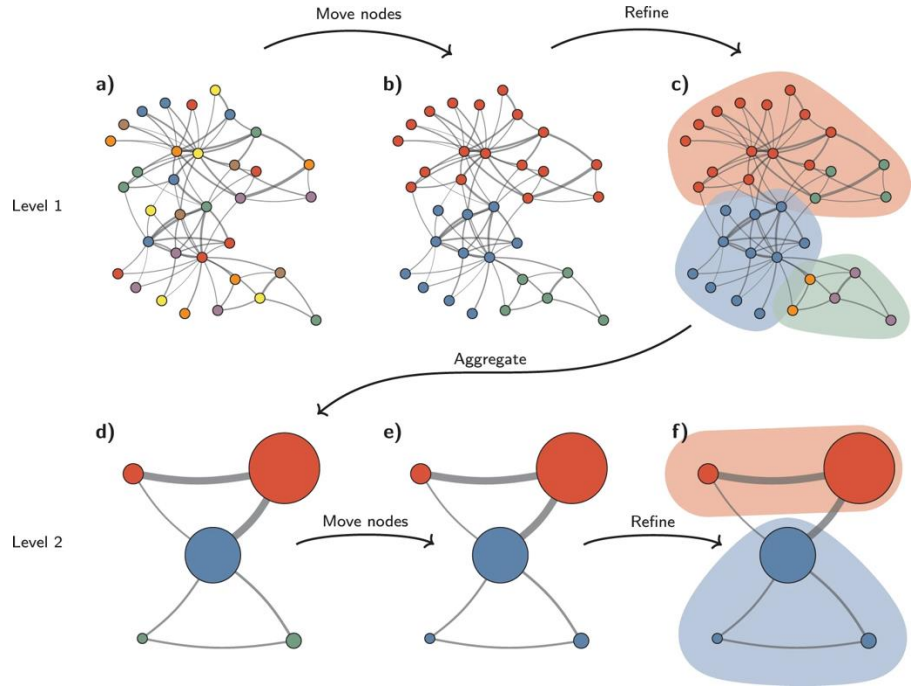
Seurat clustering

Graph based clustering methods

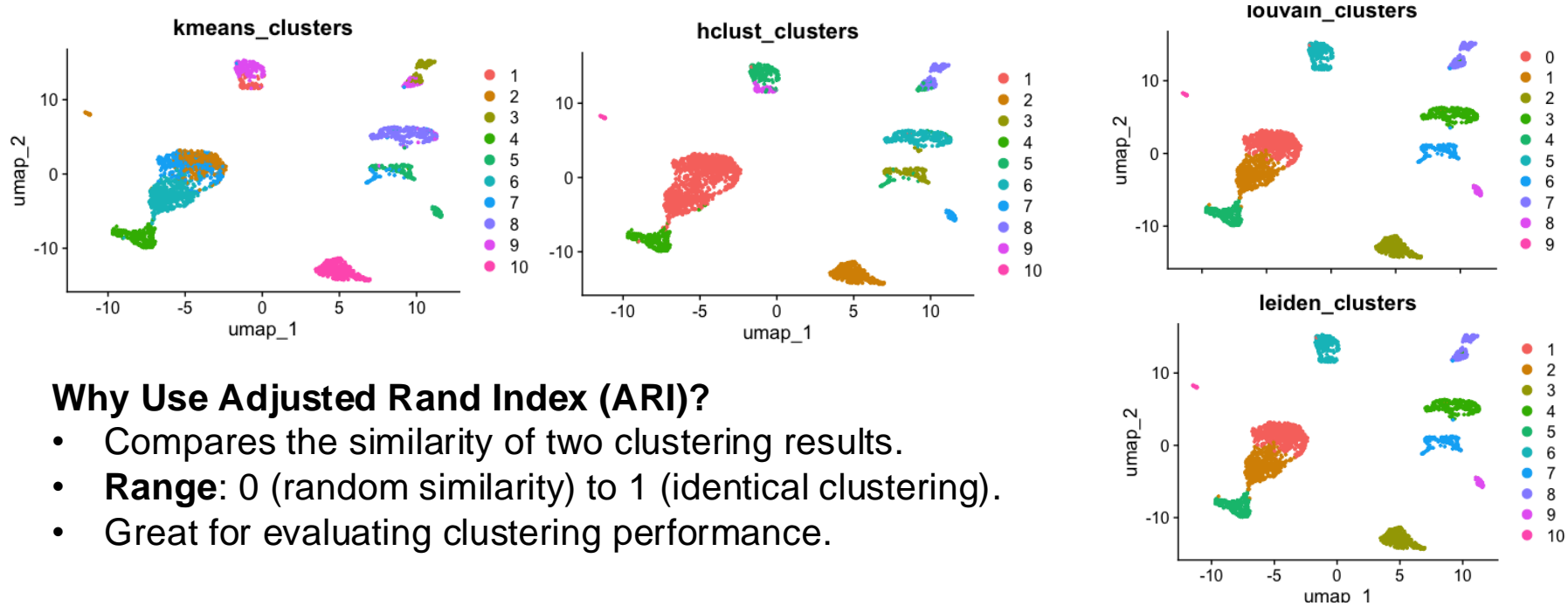
Louvain algorithm



Leiden algorithm



Exploration of scRNA-seq Dataset: Compare Different Clustering methods



Why Use Adjusted Rand Index (ARI)?

- Compares the similarity of two clustering results.
- **Range:** 0 (random similarity) to 1 (identical clustering).
- Great for evaluating clustering performance.

ARI (Louvain vs K-means): **0.72**

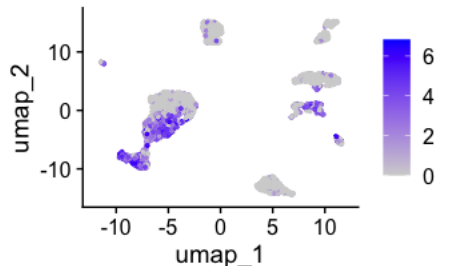
ARI (Louvain vs hclust): **0.66**

ARI (Louvain vs Leiden): **0.99**

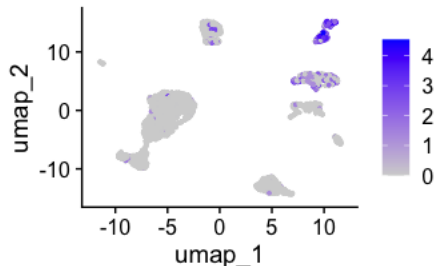
Feature plots & UMAP Visualization: Exploring Data Groupings

FeaturePlot

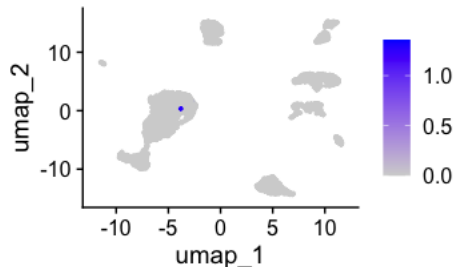
CCL4



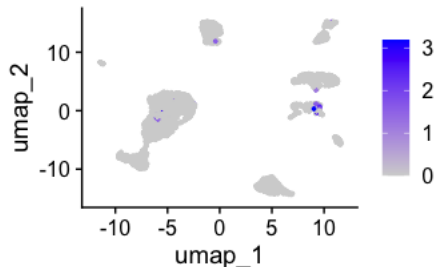
TPM2



IL21

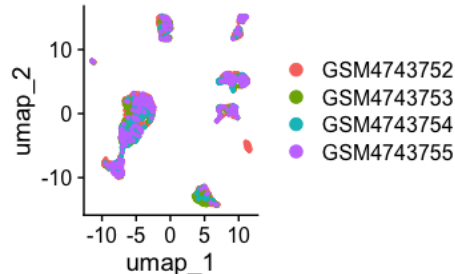


IL10

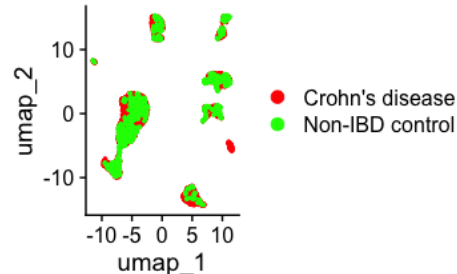


UMAP

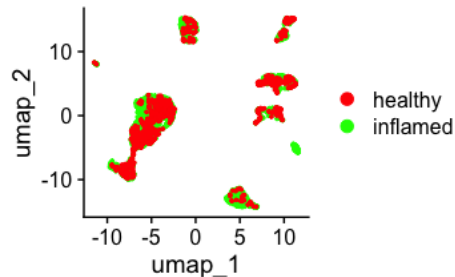
GEO_Accession



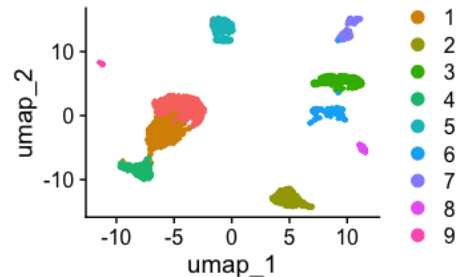
condition



tissue_status



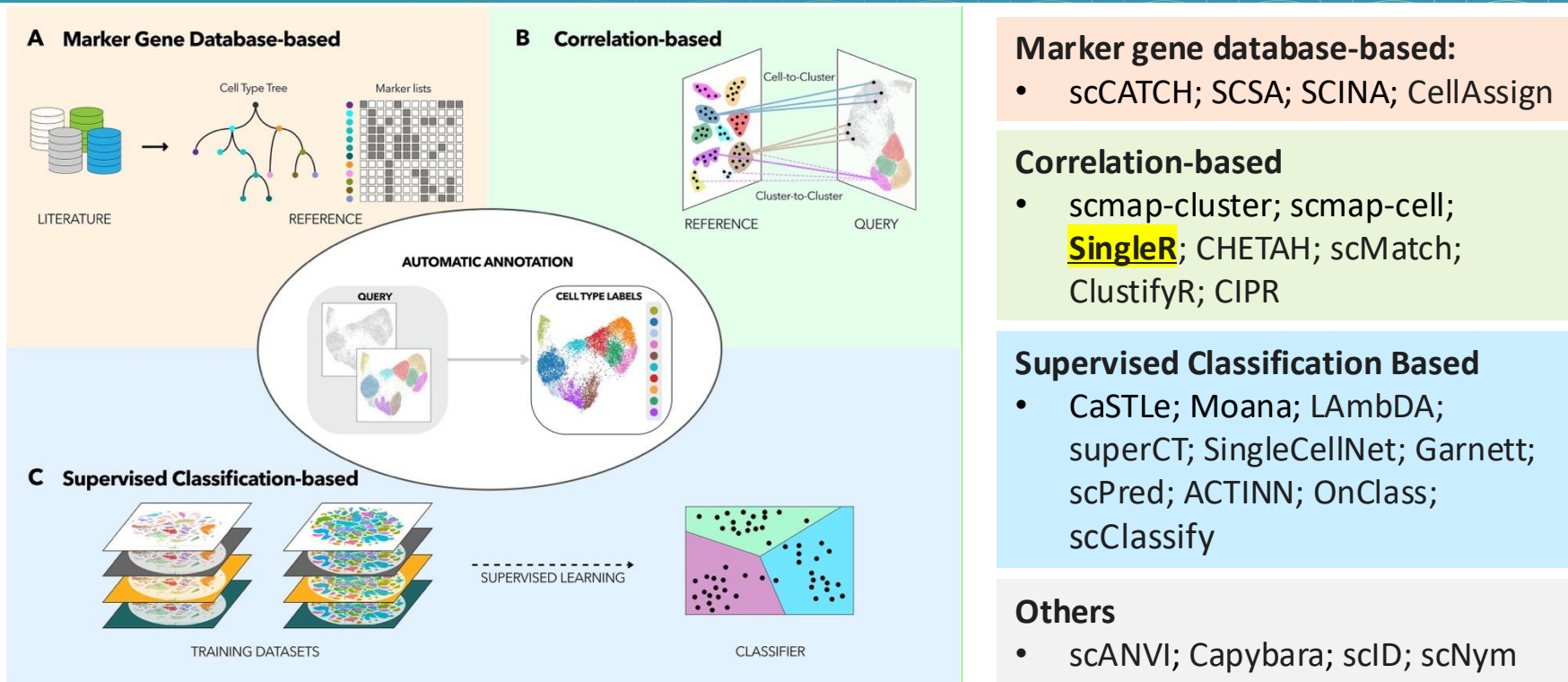
seurat_clusters



Outline

- Part1: Select study, download dataset including metadata and create Seurat Object
- Part2: Perform QC and filtering
- Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- Part6: Pathway enrichment analysis
- Part7: Cell-Cell Interaction/Communication Analysis

Approaches for cell type annotation of scRNA-seq datasets



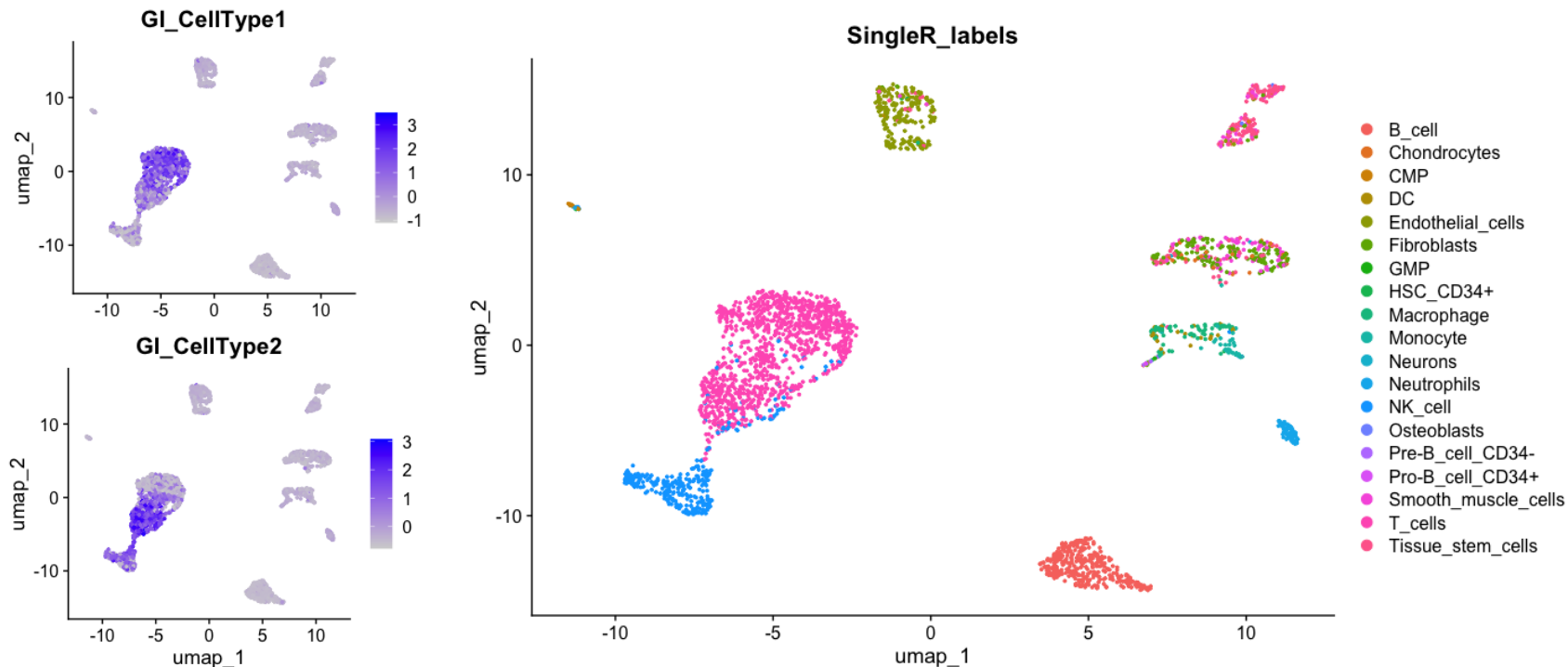
Part4: Detailed cell types assignment with SingleR

Use reference dataset with known labels to assign new cells from test dataset – based on similarities in expression profiles

Common application is to predict **cell type** in a new dataset

1. **Robust Nearest-Neighbors Classification:** SingleR uses a variant of nearest-neighbors with adjustments for better label resolution
2. **Spearman Correlation:** Calculates the correlation between the test cell and each reference sample, using Spearman correlation to reduce batch effect issues.
3. **Marker Genes:** Only marker genes from pairwise label comparisons are used, increasing label separation accuracy.
4. **Per-Label Scoring:** A score for each label is set by the 0.8 quantile of correlations across its samples, accounting for varying reference sample sizes.
5. **Label Prediction:** The label with the highest score becomes SingleR's predicted label for the cell.
6. **Fine-Tuning (Optional):** Refines predictions by iteratively narrowing down to the closest labels based on updated scores.

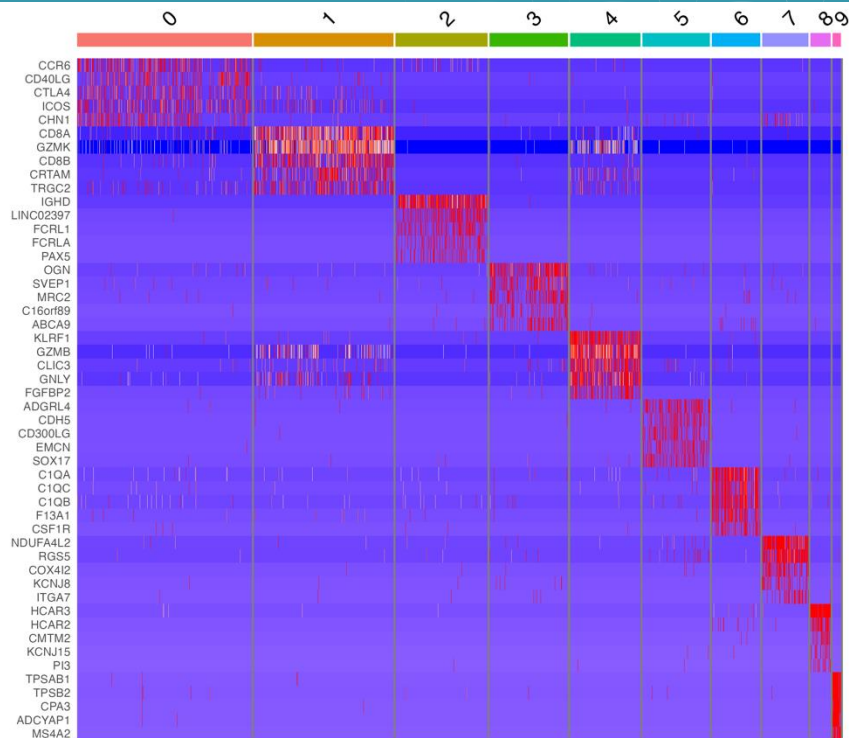
Exploration of scRNA-seq Dataset: Cell Type Annotations using Known Markers & singleR



Outline

- Part1: Select study, download dataset including metadata and create Seurat Object
- Part2: Perform QC and filtering
- Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- Part6: Pathway enrichment analysis
- Part7: Cell-Cell Interaction/Communication Analysis

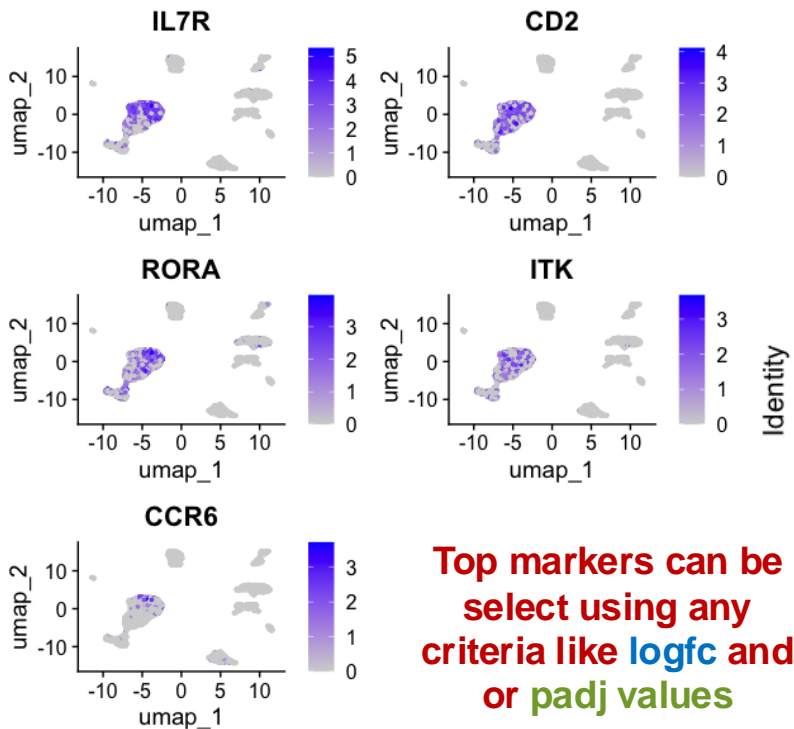
Exploration of scRNA-seq Dataset: Cluster specific Markers identified using Wilcoxon rank-sum test



```
markers <- FindAllMarkers(
  object = merged_seurat_filtered,
  assay = "RNA",
  logfc.threshold = 1, # Minimum log-fold
  change (default is 0.25)
  test.use = "wilcox", # Use Wilcoxon rank-
  sum test
  min.pct = 0.25,      # Minimum
  percentage of cells expressing the feature
  only.pos = FALSE     # Only return
  positive markers
)
```

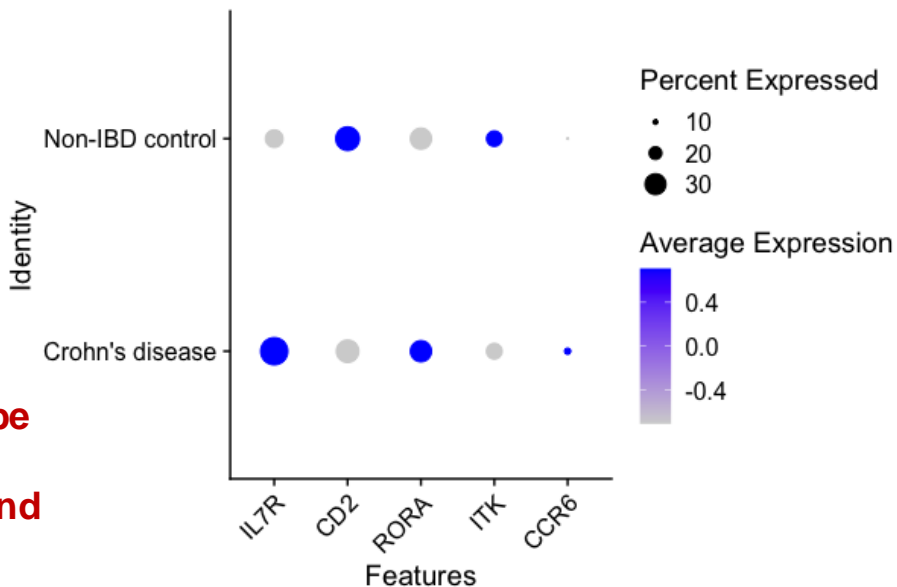
Top 5 markers of each cluster:
use any criteria like **logfc** and or **padj** values

Exploration of scRNA-seq Dataset: Disease specific Markers identified using Wilcoxon rank-sum test



Top markers can be select using any criteria like **logfc** and **padj** values

FindMarkers() – Function to get Disease or Cluster Specific Marker Genes



Outline

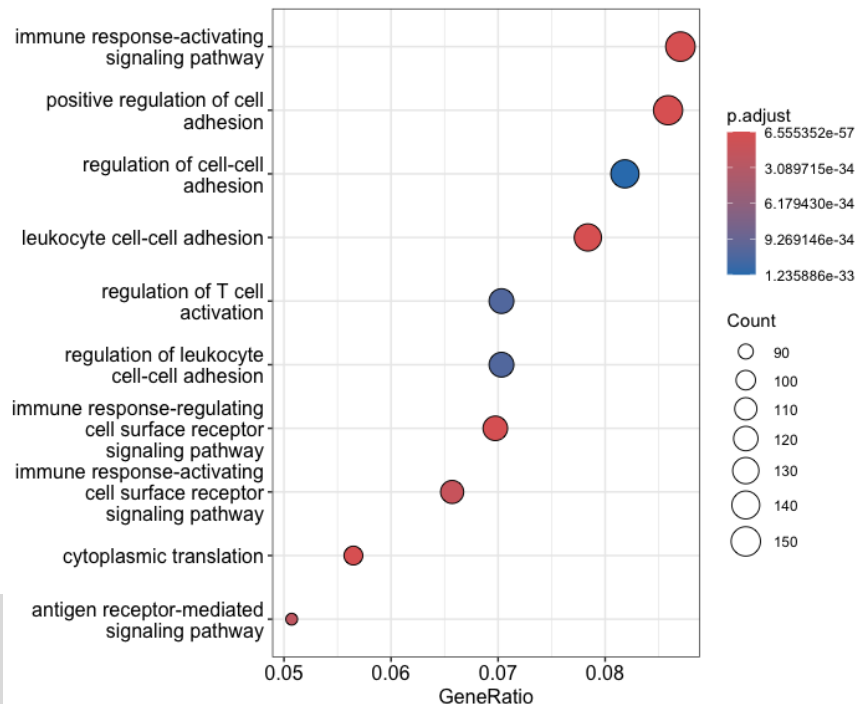
- Part1: Select study, download dataset including metadata and create Seurat Object
- Part2: Perform QC and filtering
- Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- **Part6: Pathway enrichment analysis**
- Part7: Cell-Cell Interaction/Communication Analysis

Exploration of scRNA-seq Dataset: Pathway Enrichment Analysis Using GO Terms

Steps in the Analysis:

- **Gene Selection:** Selected significant genes based on adjusted p-value < 0.05 .
- **ID Conversion:** Converted gene symbols to Entrez IDs for compatibility with enrichment tools.
- **GO Enrichment:**
 - Conducted Gene Ontology (GO) analysis for the Biological Process (BP) category.
 - Adjusted p-values using the Benjamini-Hochberg (BH) method to control for false discoveries.

enrichGO() – Function to get the significantly enriched pathways using selected marker genes



Outline

- Part1: Select study, download dataset including metadata and create Seurat Object
- Part2: Perform QC and filtering
- Part3: Normalization, Scaling and downstream analysis including Ordination and Clustering
- Part4: Cell Type Annotations using Known Markers and singleR
- Part5: Marker identification – Cluster specific and/or disease specific (if any)
- Part6: Pathway enrichment analysis
- **Part7: Cell-Cell Interaction/Communication Analysis**

Exploration of scRNA-seq Dataset: Cell-Cell Communication Analysis using CellChat

•Database Selection:

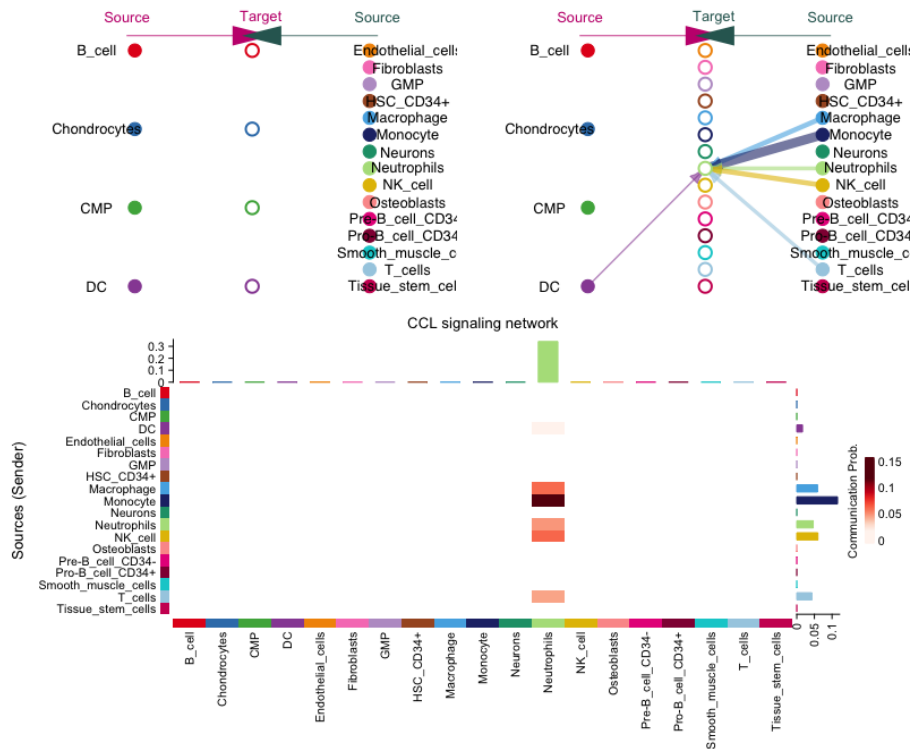
- Utilized the **CellChatDB.human** data base, which contains curated knowledge of cell signaling pathways.

•Data Processing:

- Subsetted relevant signaling data for computational efficiency.
- Computed **communication probability** based on gene expression and pathway activation.

•Pathway Aggregation:

- Integrated signaling networks across multiple pathways to identify key interaction hubs.



Network Plot:

A hierarchical layout to show directionality and strength of communication.

Heatmap:

Reveals pairwise signaling activity, making patterns between cell types visually accessible.

Follow this tutorial:
https://github.com/ashoks773/SingleCell_CrohnsDataAnalysis

Thank you

Ashok Kumar Sharma, Ph.D.
Research Bioinformatician II
Digestive and Liver Diseases
Cedars Sinai Medical Center



@ashoks773

@ashoks773

@sharma-ak

