Introduction to Single-Cell Sequencing Analyses

Sharon Hsu

Outline

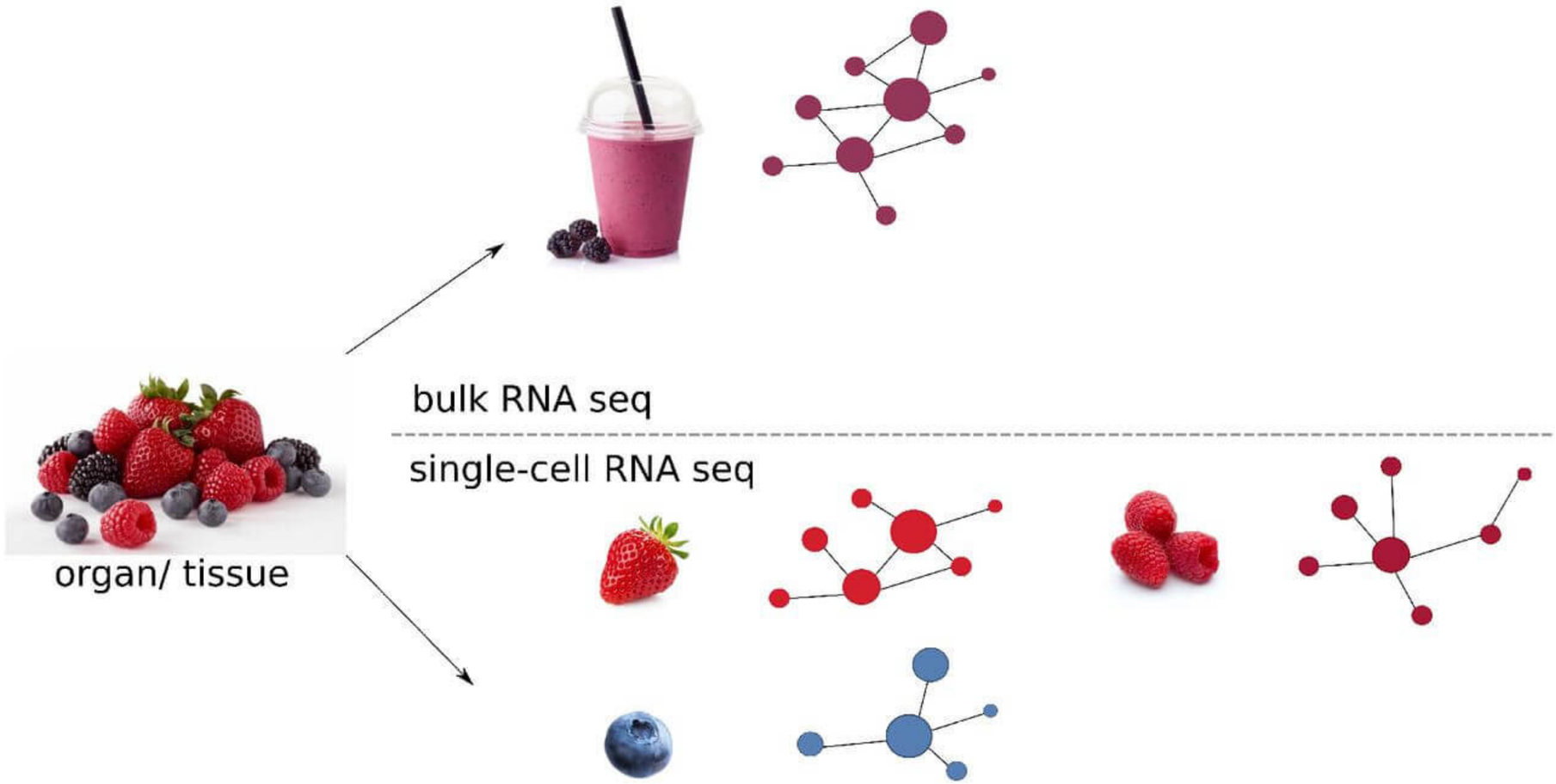
- **What** and **Why** Single-Cell Sequencing?
- **Overview** of the analysis workflow
- **Seurat** standard workflow
- **Functional** analyses
- Conclusion: challenges & future

What is Single-Cell RNA Sequencing ?

- Cells are **heterogeneous**, even within the same tissue.
- scRNA-seq allows gene expression profiling at **single-cell resolution**.
- Captures **heterogeneity** instead of averaging signals (vs bulk RNA-seq)
- Identifies **rare cell populations** and **distinct cellular states**.

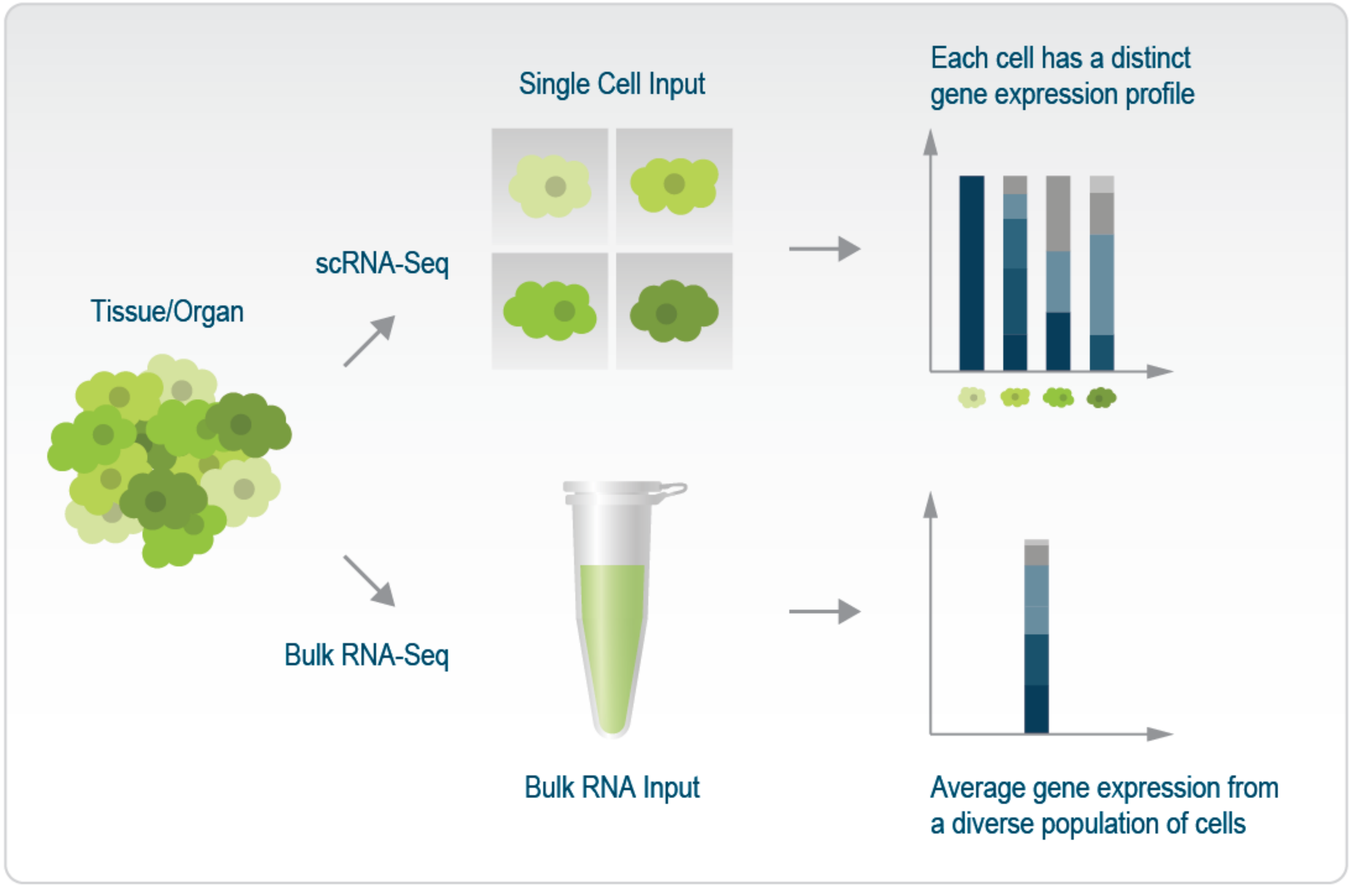
What is Single-Cell RNA Sequencing?

- Captures heterogeneity instead of averaging signals (vs bulk RNA-seq)

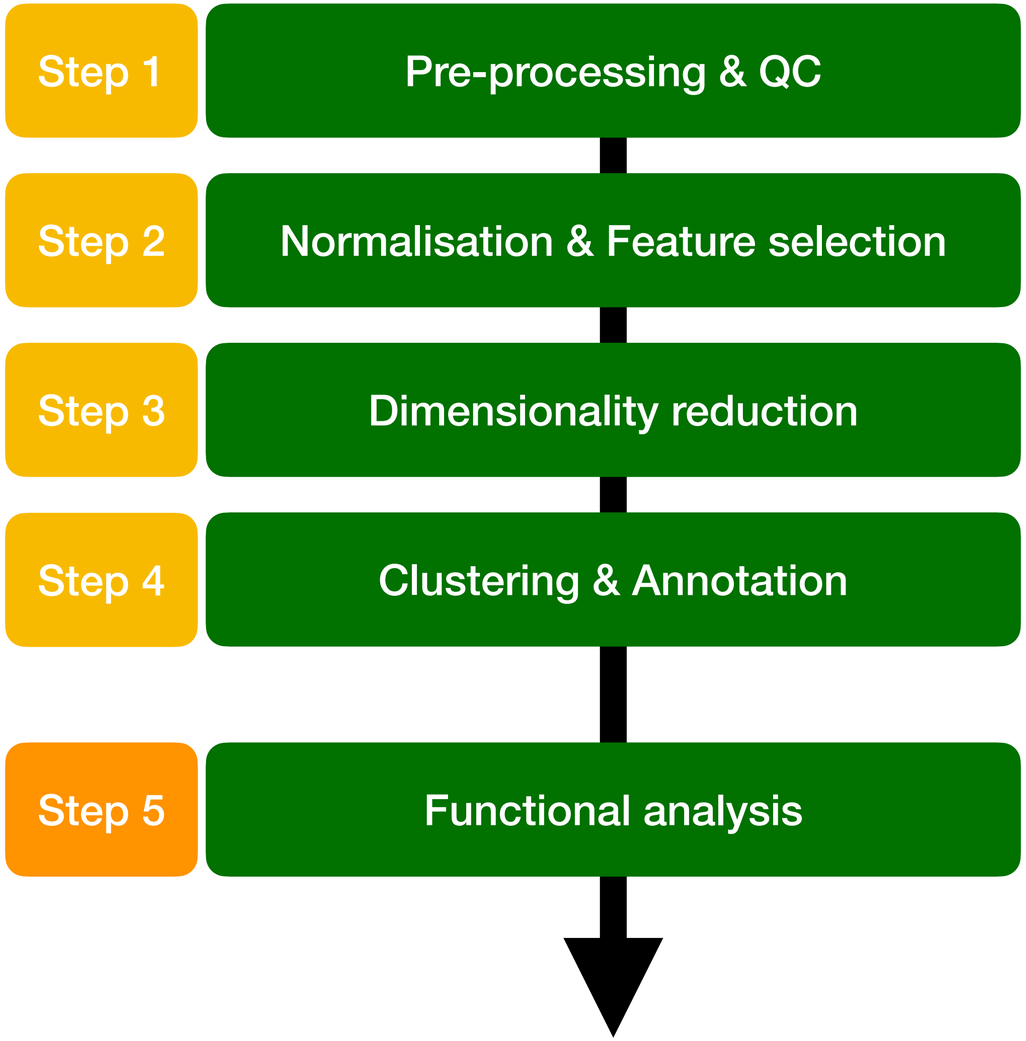


What is Single-Cell RNA Sequencing?

- Captures heterogeneity instead of averaging signals (vs bulk RNA-seq)

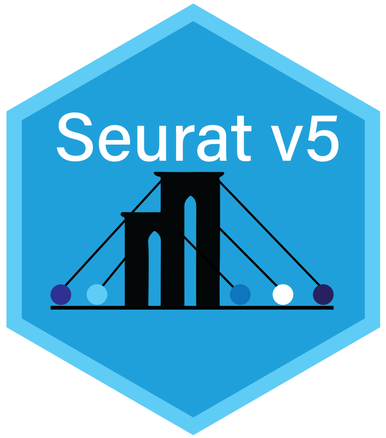


scRNA-seq Analysis Workflow



Seurat

- One of the most widely used R packages for scRNA seq analysis
- Developed by Satija Lab from New York Genome Center
- **Processing**, **analysing**, and **visualising** scRNA-seq data
- Flexible and easy to use

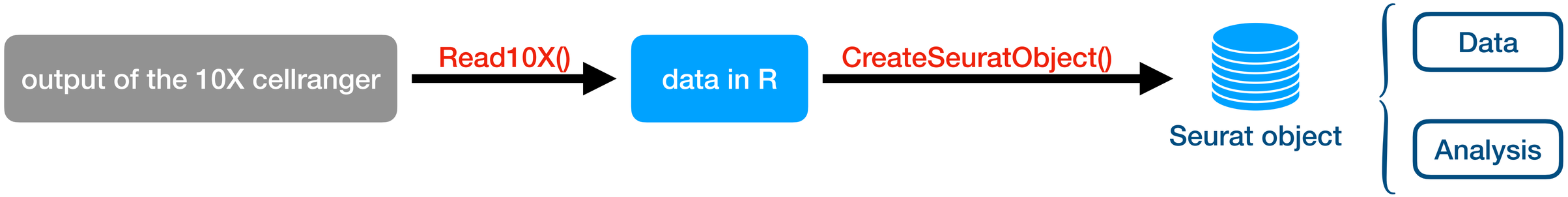


Seurat v4: Hao Y, *et al.,* Integrated analysis of multimodal single-cell data. *Cell*. 2021

Pre-processing workflow

- Raw data is often in matrix format (e.g., 10X Genomics output).
- Load data using:

```
# Load the PBMC dataset
pbmc.data <- Read10X(data.dir = "/brahms/mollag/practice/filtered_gene_bc_matrices/hg19/")
# Initialize the Seurat object with the raw (non-normalized data).
pbmc <- CreateSeuratObject(counts = pbmc.data, project = "pbmc3k", min.cells = 3, min.features =
200)
pbmc
```



Filtering Low-Quality Cells

- Remove cells based on:
  - Gene count threshold (too low = dead cell, too high = doublet).
  - Mitochondrial RNA percentage (high = damaged cell).
  - Total UMI counts per cell.
- Example code:

```
pbmc <- subset(pbmc, subset = nFeature_RNA > 200 & nFeature_RNA < 2500 & percent.mt < 5)
```

```
# Visualize QC metrics as a violin plot
VlnPlot(pbmc, features = c("nFeature_RNA", "nCount_RNA", "percent.mt"), ncol = 3)
```

Normalisation of Data

- Raw data must be normalised for proper comparison.
- Common method: **Log normalization**.
- Example code:

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

10,000 by default

Selecting informative genes

- Identifying Highly Variable Genes (HVGs) - biological meaningful!
- Directly modeling the mean-variance relationship in single cell data
- Example code:

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

# Identify the 10 most highly variable genes
top10 <- head(VariableFeatures(pbmc), 10)
```

Data scaling

- Ensures genes contribute **equally** in dimensionality reduction & clustering.
- Removes **unwanted variation** caused by differences in sequencing depth.
- Example code:

```
all.genes <- rownames(pbmc)
pbmc <- ScaleData(pbmc, features = all.genes)
```

Reducing dimensionality (PCA)

- PCA reduces the number of features while retaining variance
- Select top **principal components (PCs)** for clustering
- Visualisation: VizDimReduction(), DimPlot(), and DimHeatmap()
- Example code:

```
pbmc <- RunPCA(pbmc, features = VariableFeatures(object = pbmc))
```

Cell clustering

- **FindNeighbors()** - Graph for Clustering (map)
  - Compute a **K-nearest neighbor graph** based on the distances between cells in **PCA space**.
  - Uses **Jaccard similarity** to refine cell connections.
  - Prepares the dataset for clustering but **does not assign clusters yet**.
- **FindClusters()** – Groups Similar Cells into Clusters
  - Uses **graph-based clustering** algorithms (e.g., Louvain or SLM) to group cells.
  - Resolution affects number of clusters (higher = more clusters).

res = 0.4-1.2 recommend for 3K data

- Example code:

```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

Cell clustering

- **FindNeighbors()** - Graph for Clustering (map)
  - Compute a **K-nearest neighbor graph** based on the distances between cells in **PCA space**.
  - Uses **Jaccard similarity** to refine cell connections.
  - Prepares the dataset for clustering but **does not assign clusters yet**.
- **FindClusters()** – Groups Similar Cells into Clusters
  - Uses **graph-based clustering** algorithms (e.g., Louvain or SLM) to group cells.
  - Resolution affects number of clusters (higher = more clusters).

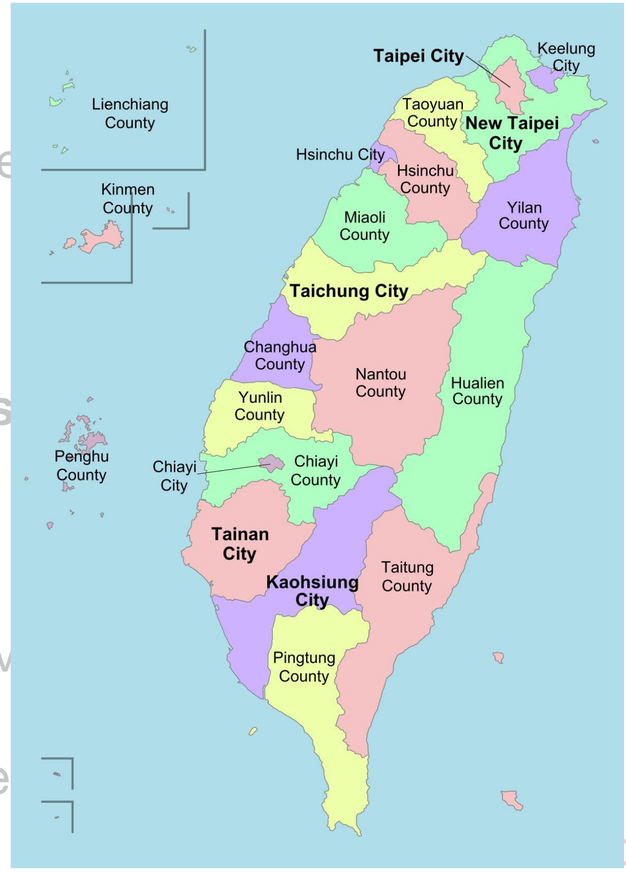
res = 0.4-1.2 recommend for 3K data

- Example code:

```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

Cell clustering

- **FindNeighbors()** - Graph for Clustering (map)
  - Comput**aph** based on the **PCA space.**
  - Uses **Ja**ell connections.
  - Prepare**but does not ass**
- **FindCluste**into Clusters
  - Uses **gr**rithms (e.g., Louv
  - Resoluti**rs** (higher = more
- Example code:



FindNeighbors()

FindClusters()

```
pbmc <- FindNeighbors(pbmc, dims = 1:10)
pbmc <- FindClusters(pbmc, resolution = 0.5)
```

Non-linear dimensionality reduction

- t-SNE or **UMAP** for visualisation
- High dimensional information → **low-dimensional space**.
- Cells in same cluster should co-localize on **UMAP**/t-SNE plots
- Preserves **local**, but **not global** relationships, over-simplified
- Use for **exploration**, not conclusions

```
pbmc <- RunUMAP(pbmc, dims = 1:10)
```

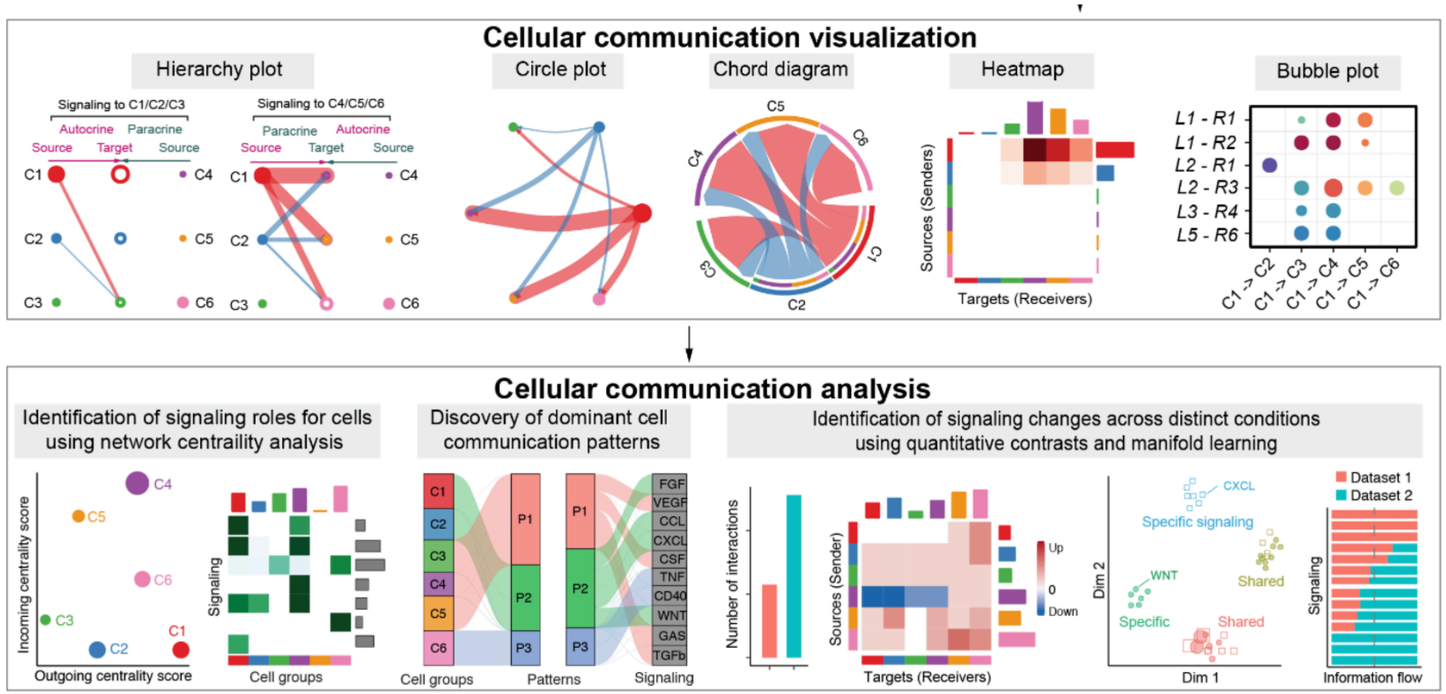
```
# note that you can set `label = TRUE` or use the LabelClusters function to help label
# individual clusters
DimPlot(pbmc, reduction = "umap")
```

Cell annotation

- Assigns **biological identity** to cell clusters
- Compares gene expression profiles to **reference datasets**.
- Common tools (combine for higher accuracy)
  - **SingleR**
  - **scCATCH**
  - **CellID + PanglaoDB**
  - **Cluster biomarkers**

Functional analysis - CellChat

- Explore cell-cell communication
- Predict **ligand-receptor interactions** and **active pathways** between cell types
- Useful visualisation for key interactions:



Functional analysis - Pseudotime analysis

- Models **cell differentiation over time**, helpful for studying **dynamic** biological processes
- Orders cells based on transcriptional similarity
- Assigns each cell a ‘pseudotime’ score (relative differentiation stage)
- Constructs a trajectory that connects cell states.
- Identifies genes driving state transitions.
- Tools for Pseudotime Analysis: **Monocle3**, Slingshot, SCORPIUS

Conclusion

- scRNA-seq captures **cellular heterogeneity** at single-cell resolution.
- Seurat provides a structured workflow for data processing & **clustering**.
- Cell annotation assigns **biological identity** to cell clusters.
- CellChat reveals **cell-cell communication** through signaling pathways.
- Pseudotime Analysis models **cell differentiation** and lineage trajectories.

Questions

l54115325@gs.ncku.edu.tw