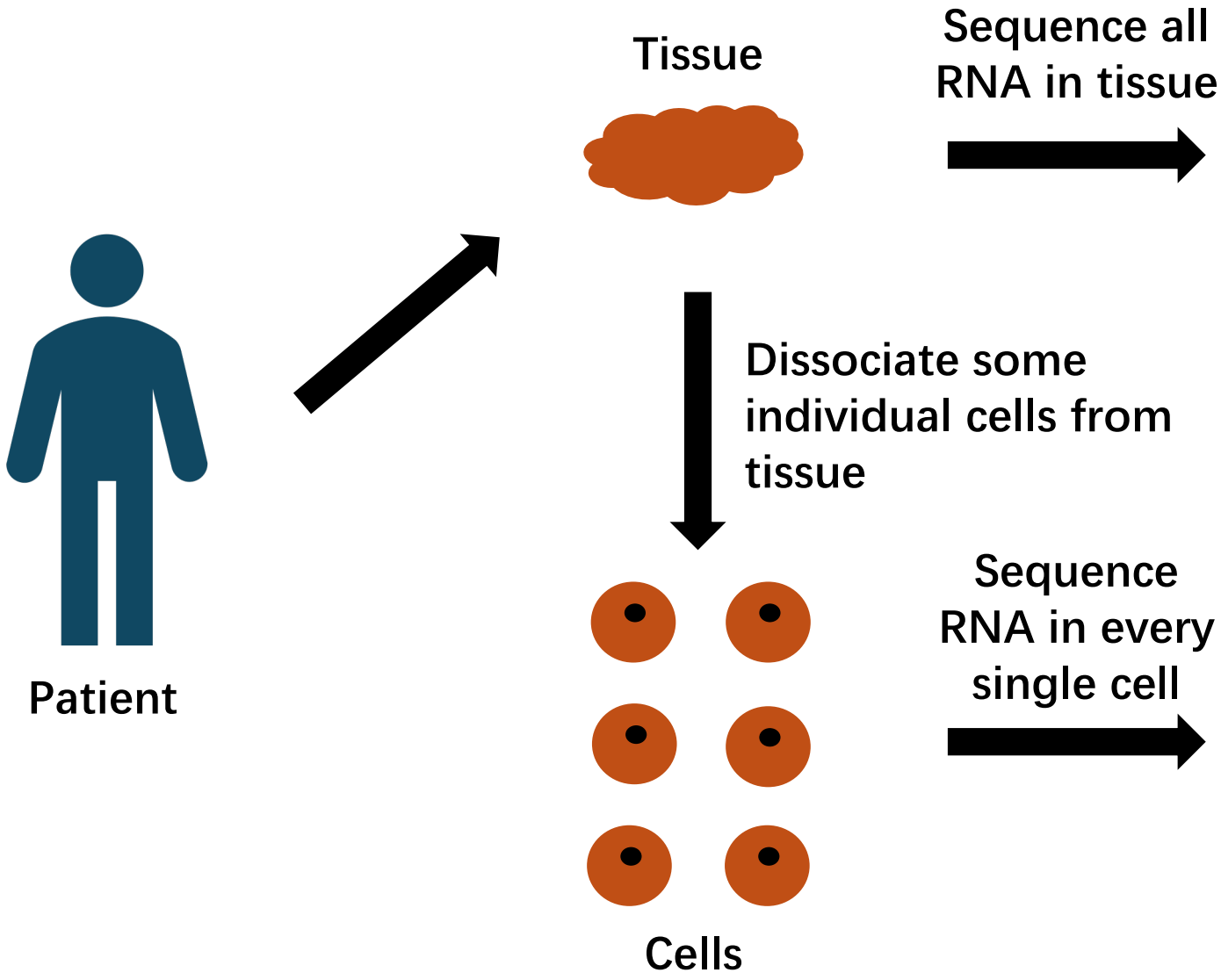
Current best practices in single-cell RNA-seq analysis: a tutorial

Melbourne Integrative Genomics

Xiaochen, Zhang

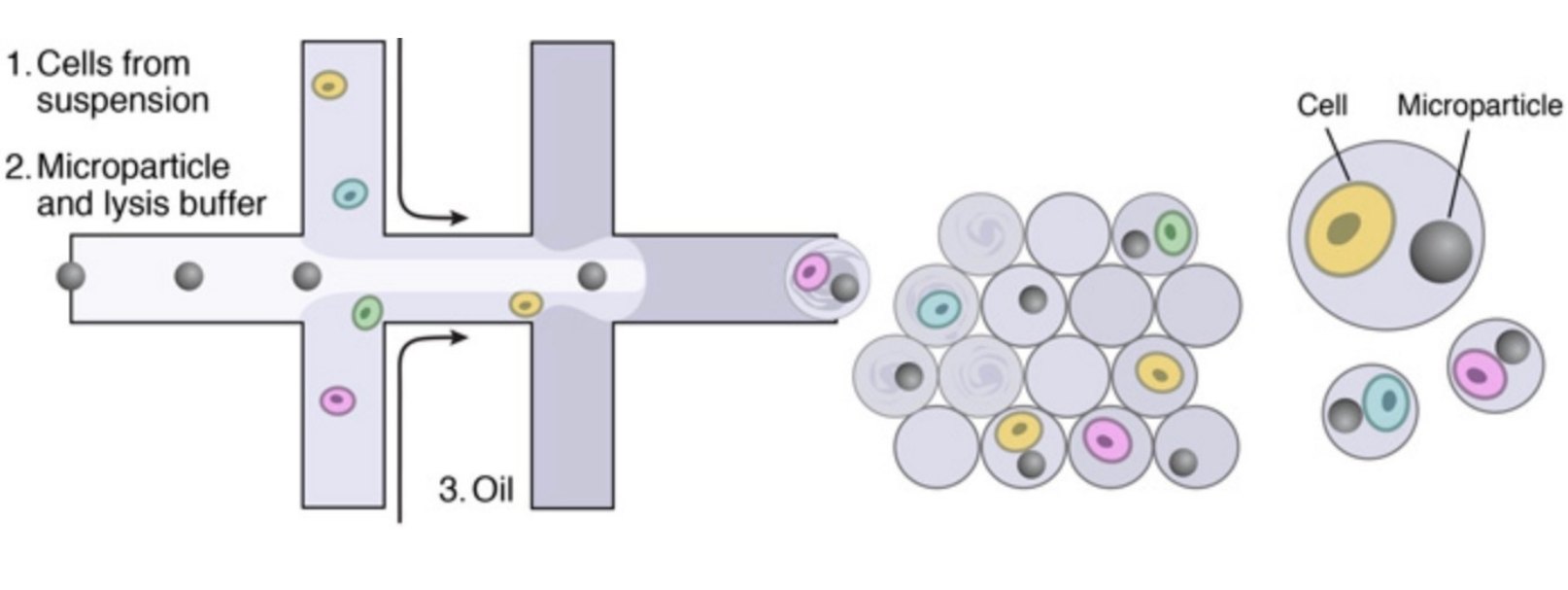
Luecken MD, Theis FJ. (2019).

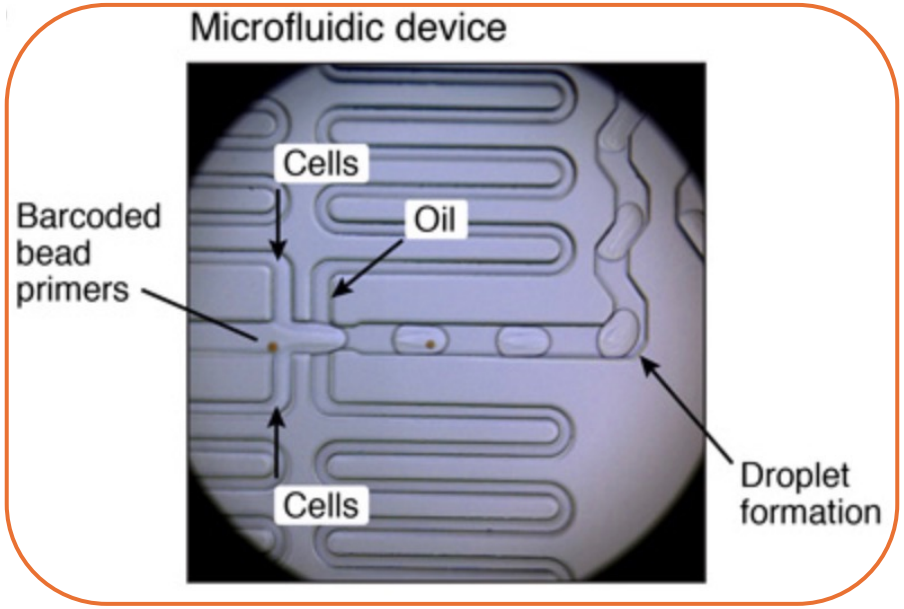
scRNA-seq: A part of transcriptome



Dissociation

Most scRNA-seq experiments use droplet-based platforms for dissociation.



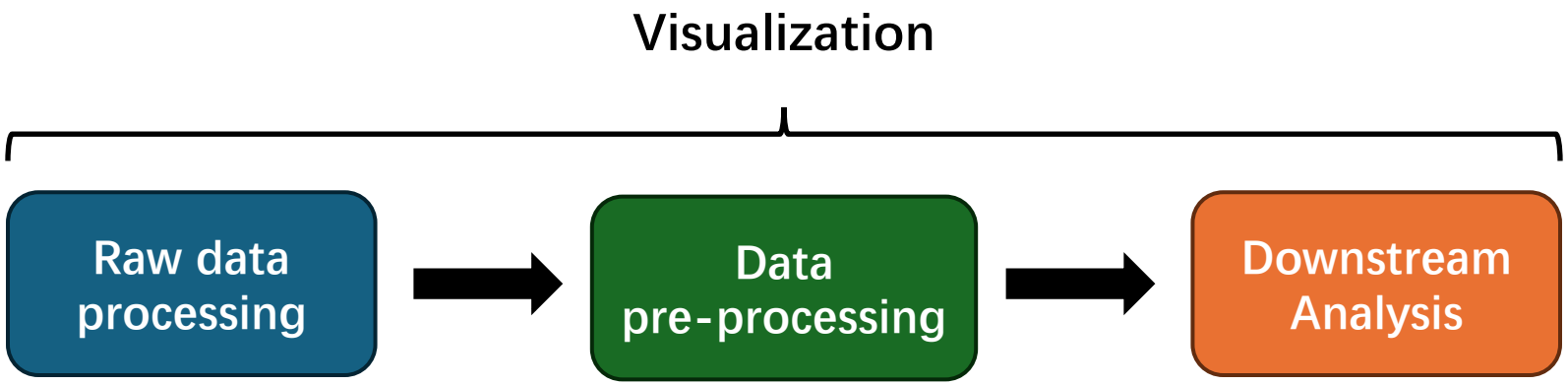


Some scRNA-seq experiments use plate-based platforms for dissociation.

We assign each dissociated cell a barcode consisting of 4-20 bases to determine its identity.

Macosko EZ et al. (2015);
More details: Baran et al. (2018).

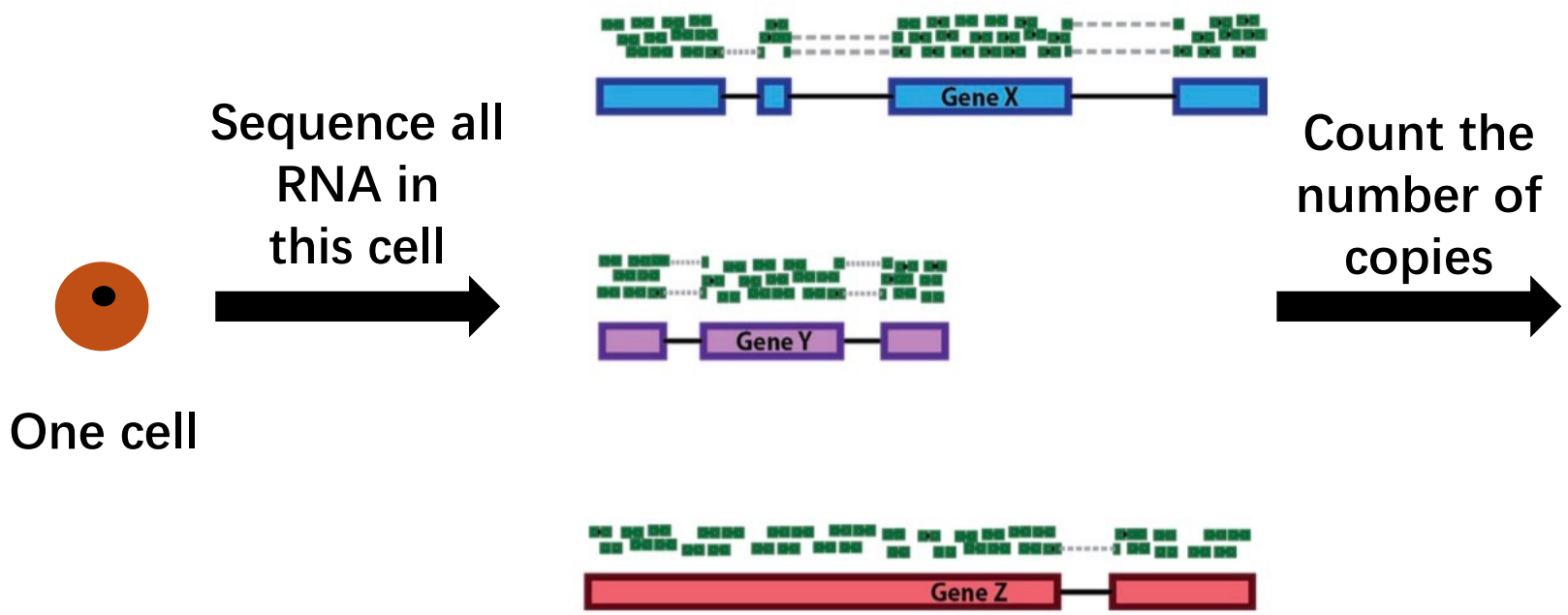
High-level pipeline for analyzing scRNA-seq data



Luecken MD, Theis FJ. (2019).

Raw data processing: from sample to counts

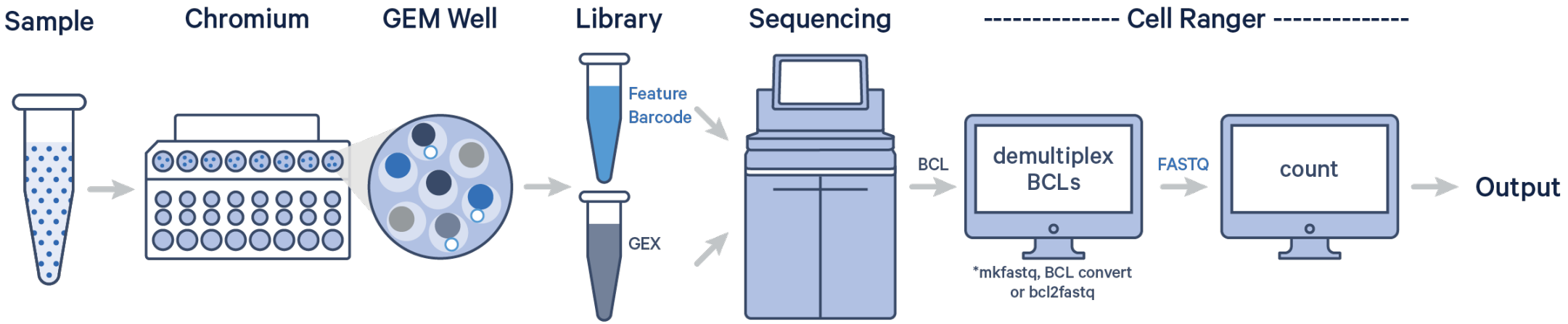
Raw data processing



Gene	Count
Gene X	3
Gene Y	3
Gene Z	2

Raw data processing: from sample to counts

Raw data processing



e.g Software Cell Ranger from 10x Genomics

More information about Cell Ranger:
https://www.10xgenomics.com/support/software/cell-ranger/latest/getting-started/cr-what-is-cell-ranger

A real scRNA-seq matrix

Raw data processing

	pbmc2_10X_V2_AAACCTGAGATGGGTC	pbmc2_10X_V2_AAACCTGAGCGTAATA	pbmc2_10X_V2_AAACCTGAGCTAGGCA	pbmc2_10X_V2_AAACCTGAGGGTCTCC
TSPAN6	0	0	0	0
TNMD	0	0	0	0
DPM1	0	0	0	0
SCYL3	0	0	0	0
C1orf112	0	0	0	0
FGR	0	0	2	0
CFH	0	0	0	0
FUCA2	0	0	0	0
GCLC	0	0	0	0
NFYA	0	0	0	0
STPG1	0	0	0	0
NIPAL3	0	1	0	0
LAS1L	0	0	0	0
ENPP4	0	0	0	0

High dimension (typically > 10,000 cells * 20,000 genes)

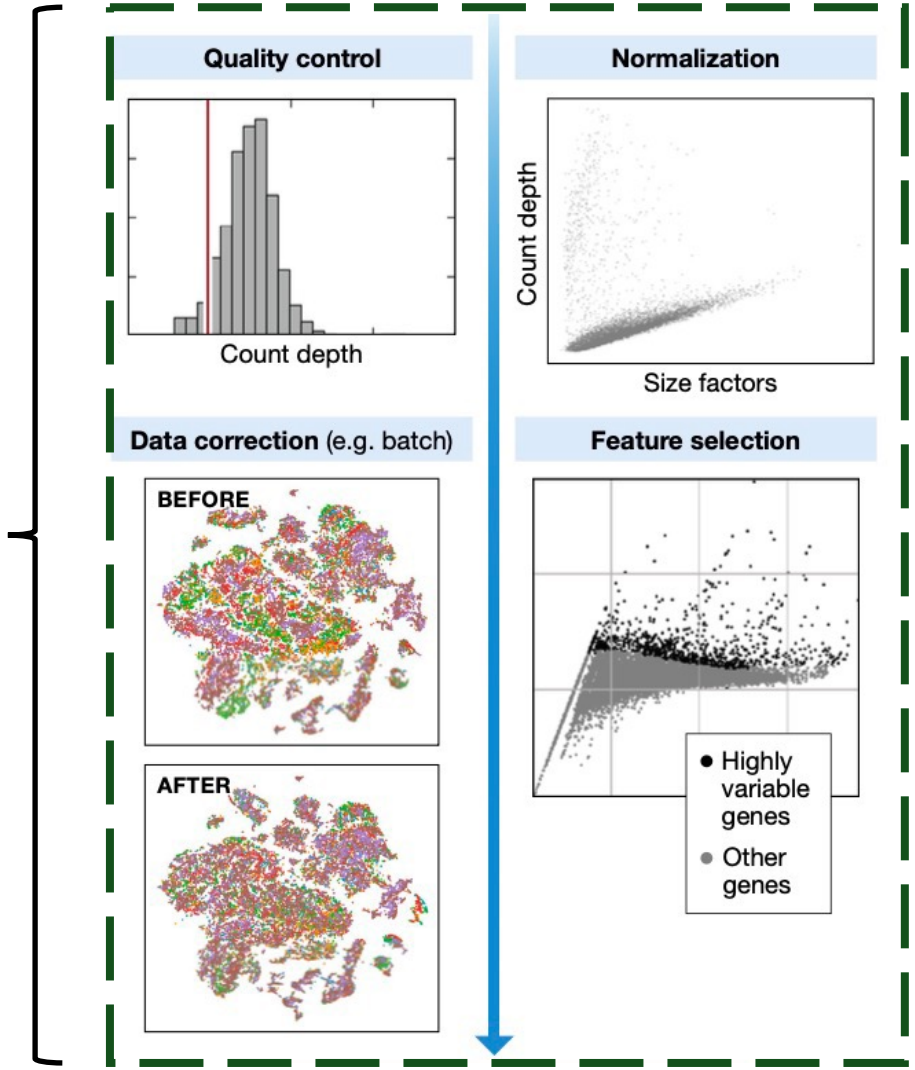
Dimension reduction tools

Very Sparse (more than 90% of the matrix includes 0)

Feature selection, new computational methods

Data pre-processing: Overview

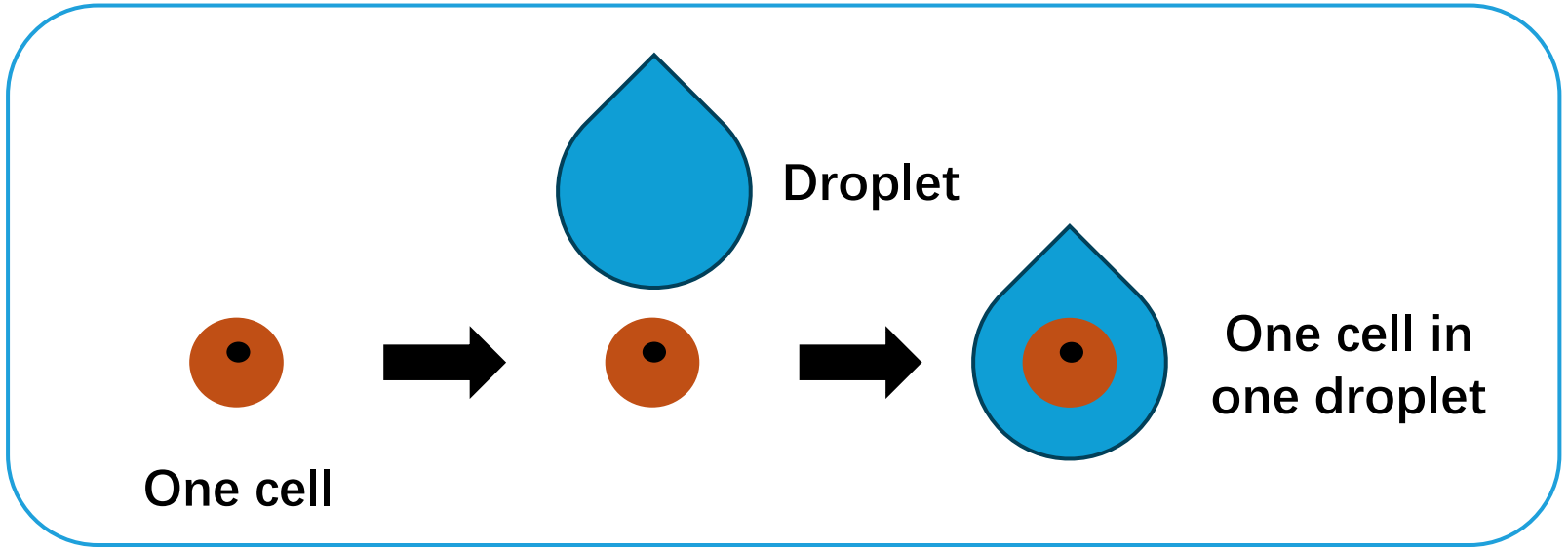
Data pre-processing



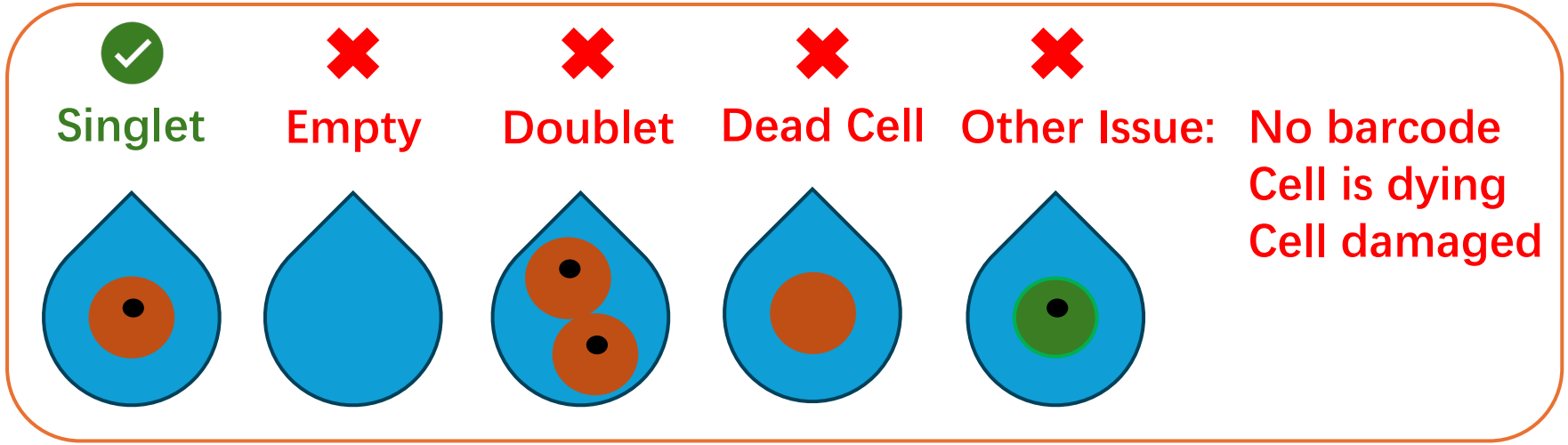
Luecken MD, Theis FJ. (2019).

Data pre-processing: Quality Control

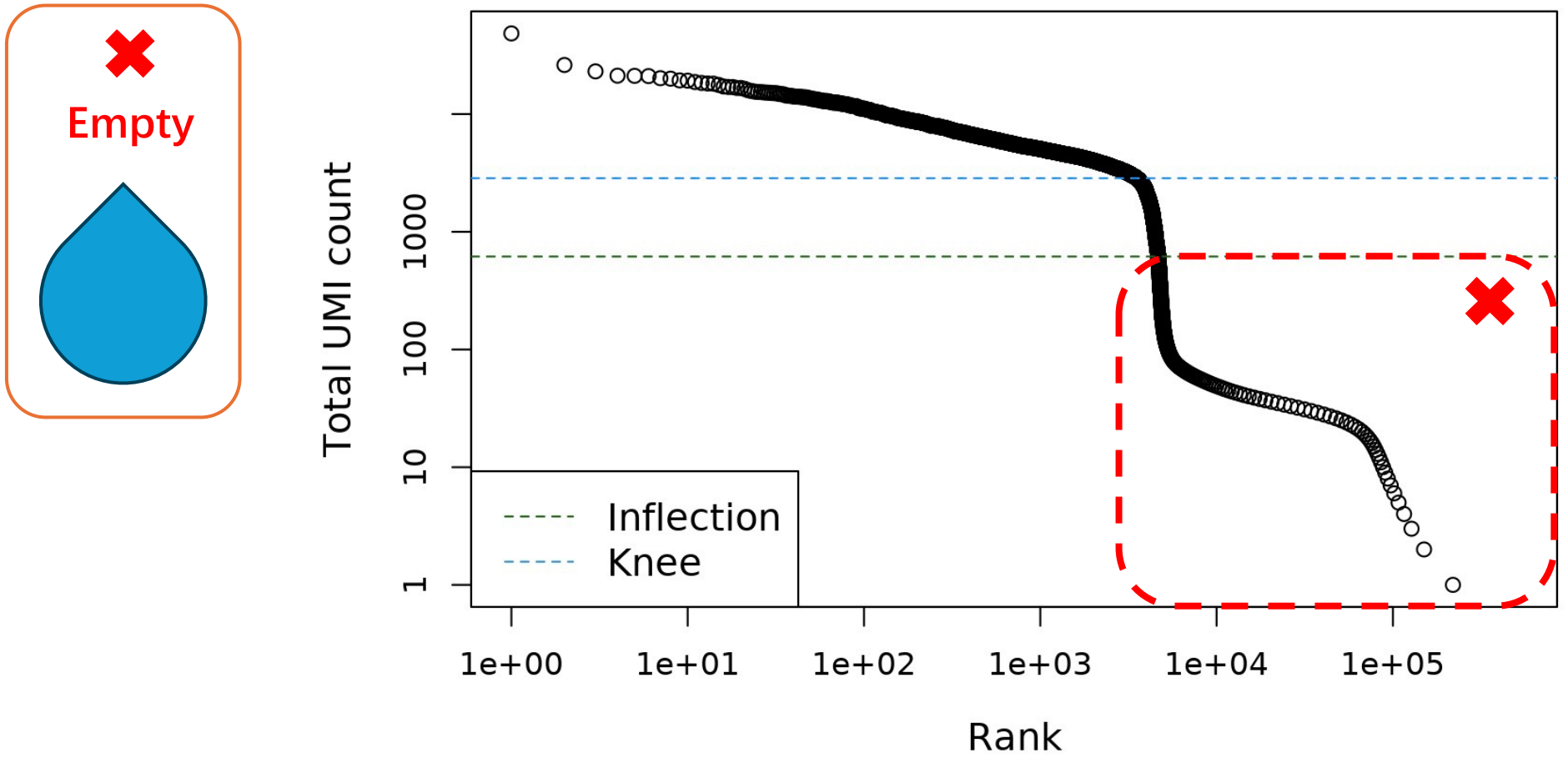
Ideal Situation:



Actual Situation:



Data pre-processing: Quality Control

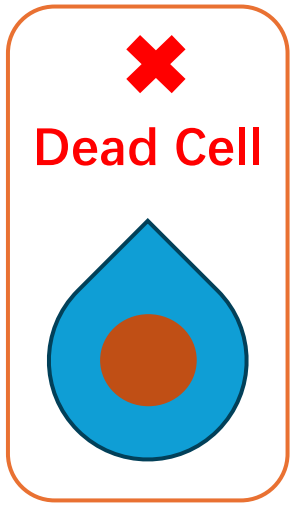


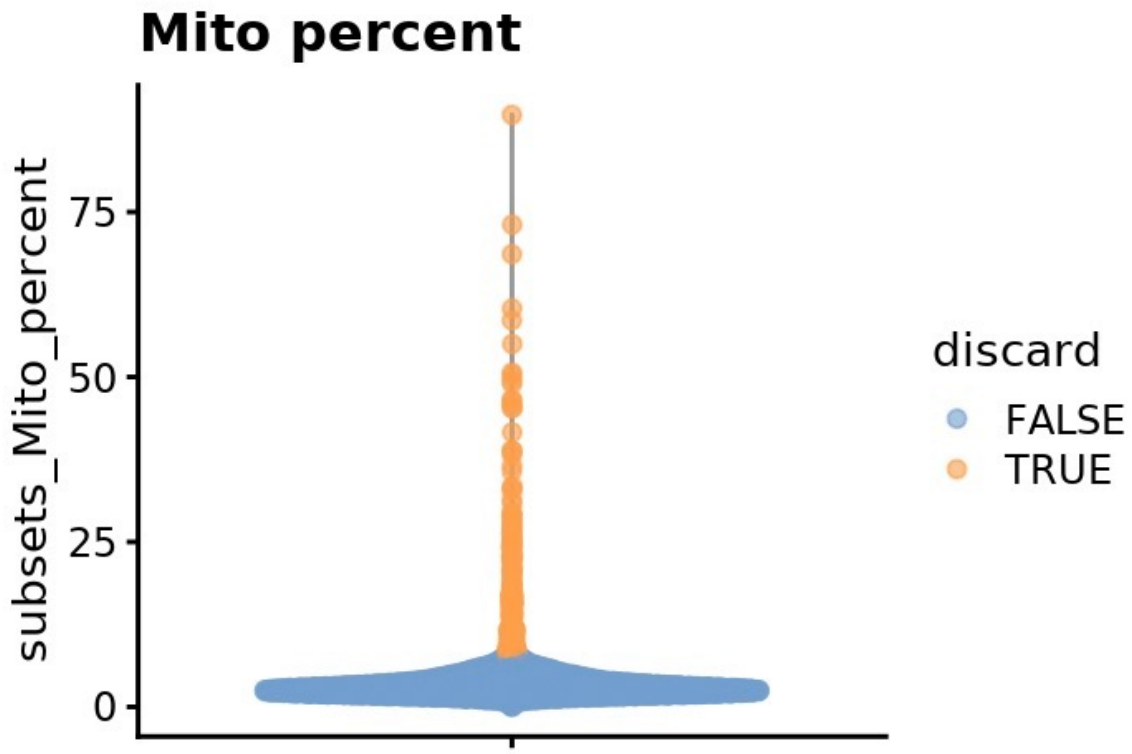
Empty droplets have a low library size (= total RNA counts).

Figure from Davis McCarthy's lab (MIG)

Quality Control

Data pre-processing



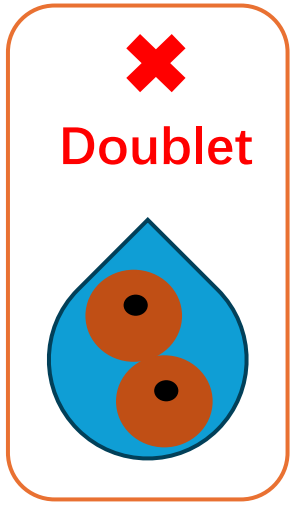


Dead (dying) cells have a high percentage of mitochondrial gene counts.

Figure from Davis McCarthy's lab (MIG)

Quality Control

Data pre-processing



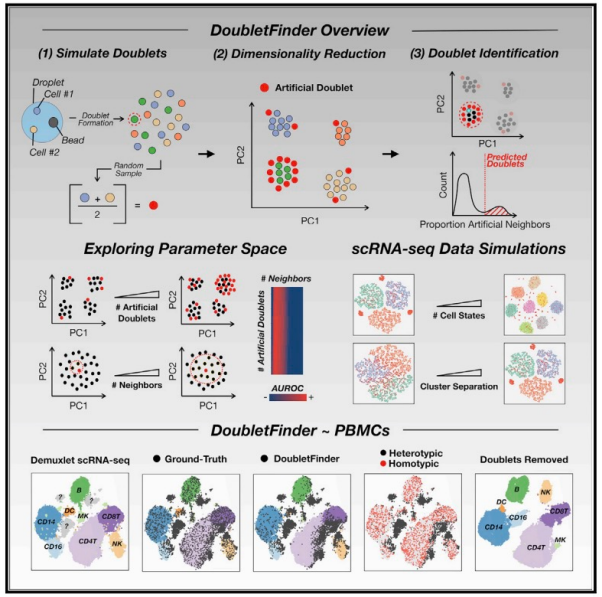
Many software are designed specifically to remove doublets.
For example:

Brief Report

Cell Systems

DoubletFinder: Doublet Detection in Single-Cell RNA Sequencing Data Using Artificial Nearest Neighbors

Graphical Abstract



Authors

Christopher S. McGinnis,
Lyndsay M. Murrow, Zev J. Gartner

Correspondence

zev.gartner@ucsf.edu

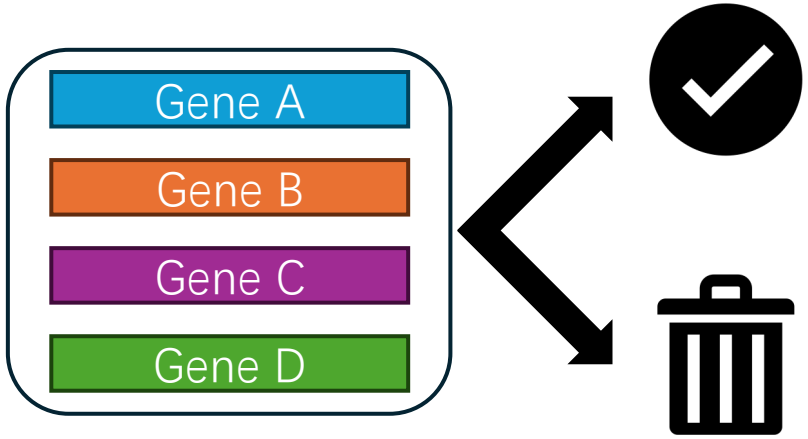
In Brief

scRNA-seq data interpretation is confounded by technical artifacts known as doublets—single-cell transcriptome data representing more than one cell. Moreover, scRNA-seq cellular throughput is purposefully limited to minimize doublet formation rates. By identifying cells sharing expression features with simulated doublets, DoubletFinder detects many real doublets and mitigates these two limitations.

Data pre-processing: Feature Selection

Most of genes are not useful

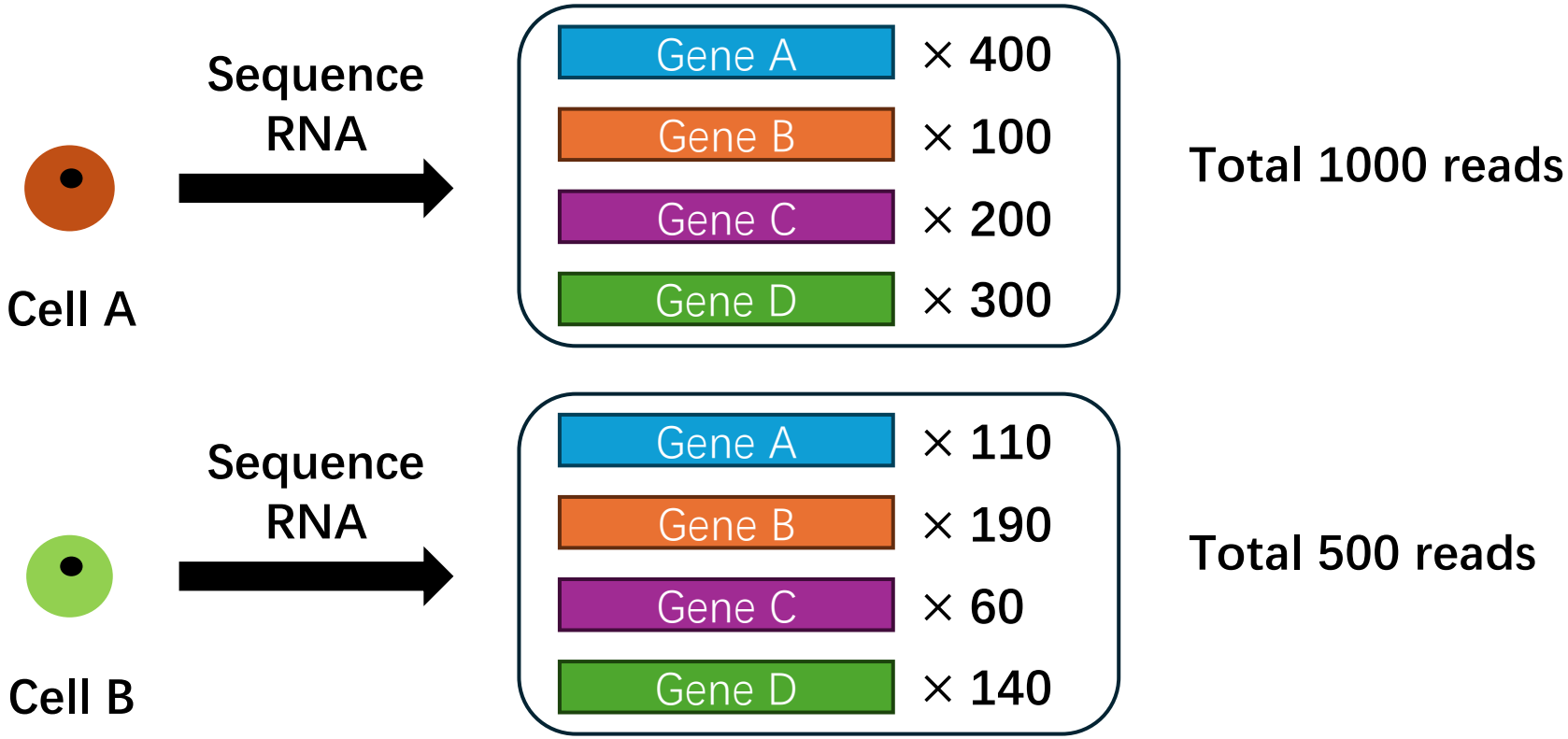
- **Genes with low or no expression levels**
- **Housekeeping genes**



We often select the top 3000 ~ 5000 most variable genes.

Data pre-processing: Normalization

Purpose: Remove technical variation (Sampling part of RNA)



Estimate a size factor for each cell that can be used to adjust for library size

Data pre-processing: Data Correction

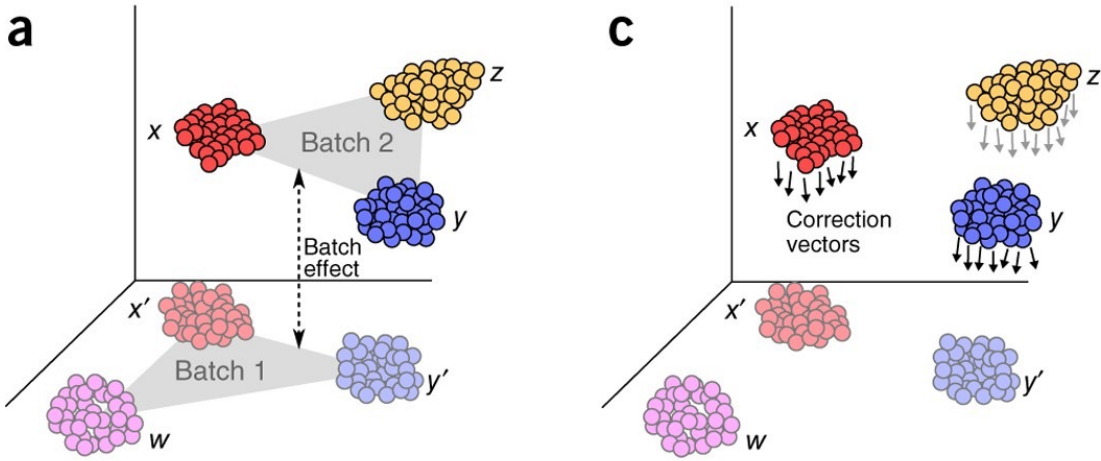
Remove batch effect and data integration

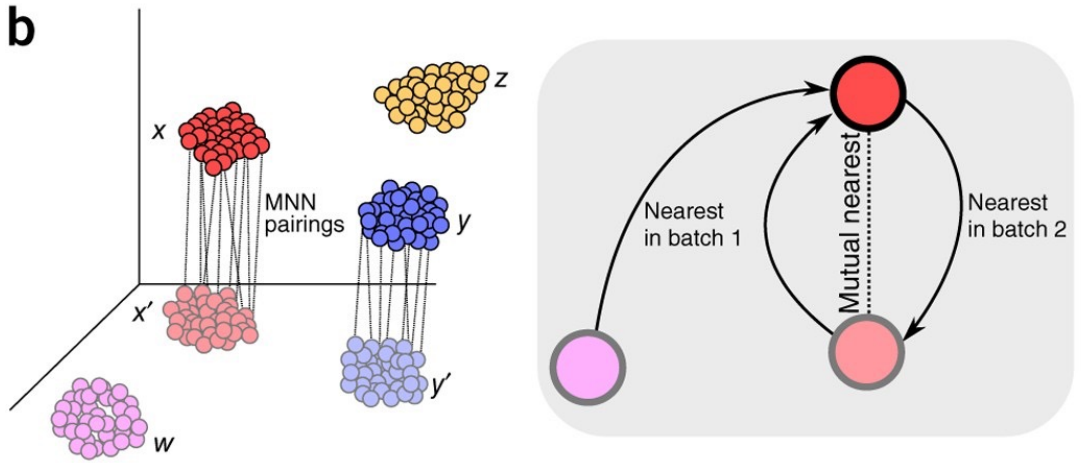
Tools:

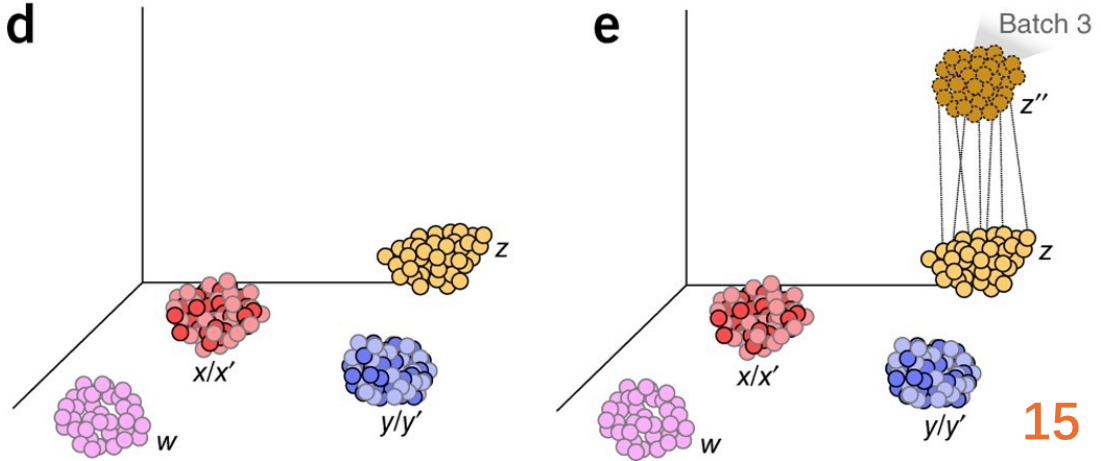
Matching mutual nearest neighbours (MNN)
Haghverdi et al. (2018)

Canonical correlation analysis (CCA)
Satija et al.(2015)

Harmony (R package)
Korsunsky et al. (2019)







Data pre-processing

Data pre-processing: Dimension Reduction

Data pre-processing

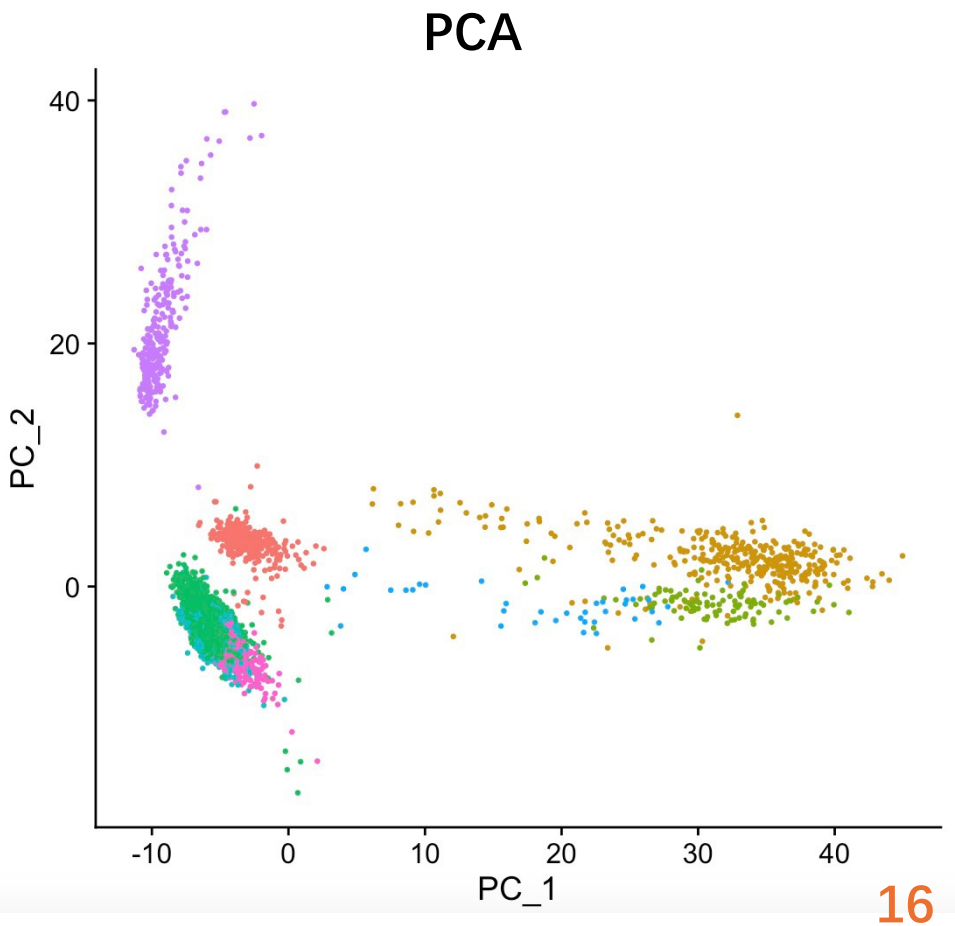
Purpose: Address high dimensionality and gene correlation

Method: Principal Component Analysis (PCA)

Use PCA to compress data into linear combinations of genes that capture the largest amount of variation in the data

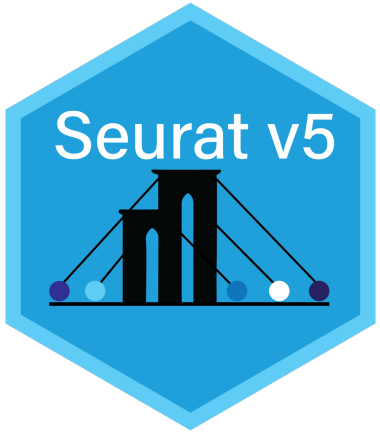
PCs can be used as input to downstream steps to save computing time

Not for visualization purposes
(2D PCA is not enough to summarize most of the information)



Summary of data pre-processing: tools

Data pre-processing



Seurat is an R package that provides full-process automatic data pre-processing.

If there are no personalized requirements, each step of data pre-processing only requires one line of code.

Seurat can also do some downstream analysis.

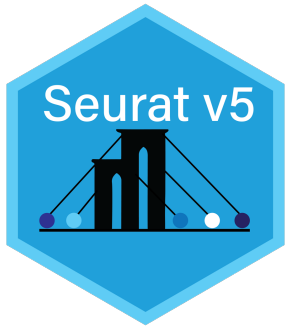
We will use it in the practical session.

Hao et al. (2023).

Summary of data pre-processing

Data pre-processing

Step 1	Quality Control	Filter out poor-quality cells
Step 2	Normalization	Remove technical variation
Step 3	Feature Selection	Select the top ~5,000 most variable gene
Step 4	Dimension Reduction	PCA to simplify downstream operations
Step 5	Data Correction	CCA to remove the batch effect between two patients.



Visualization

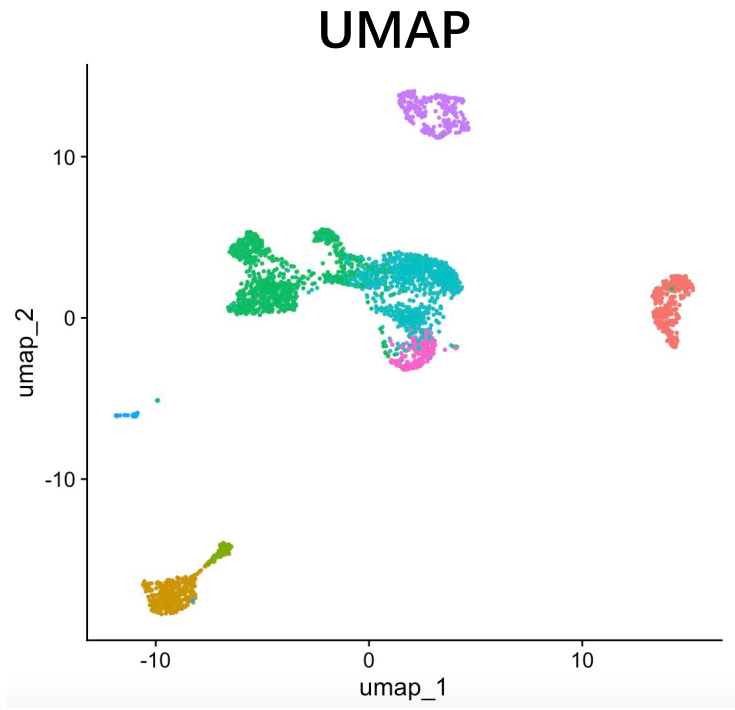
Visualize high-dimensional data in 2D

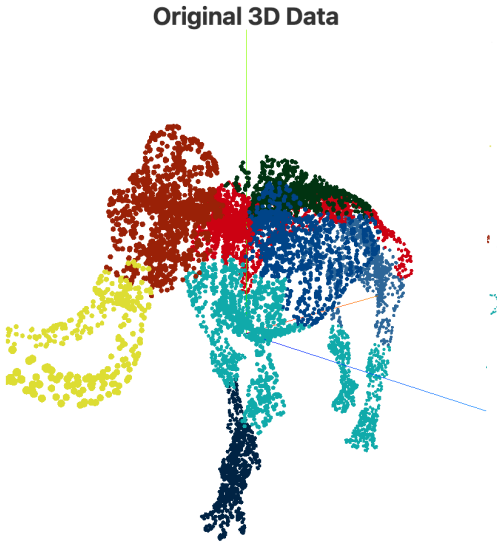
Popular technique:

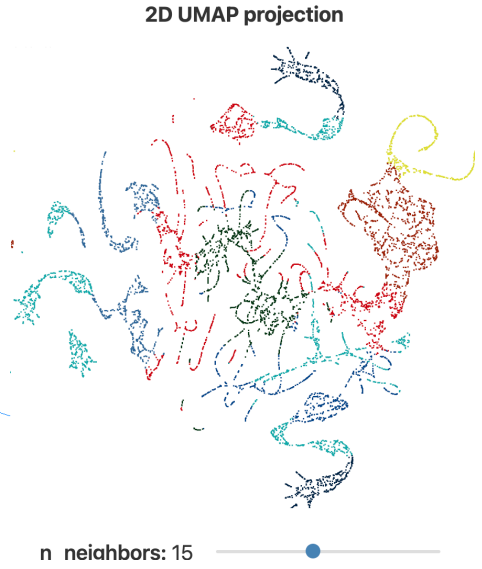
- **Uniform manifold approximation and project (UMAP)**

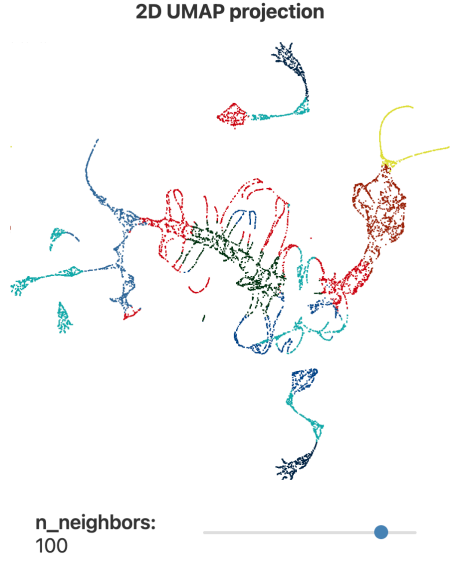
Features:

- **Non-linear dimension reduction**
- **Good at preserving local structures: e.g. clusters**
- **Fast (by using PCs from PCA)**









Visualization

UMAP is good enough,

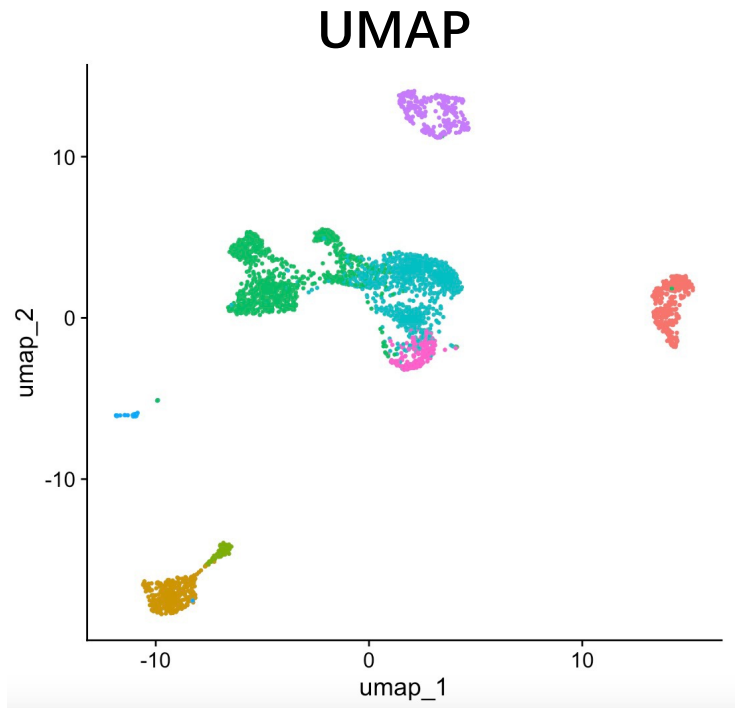
But:

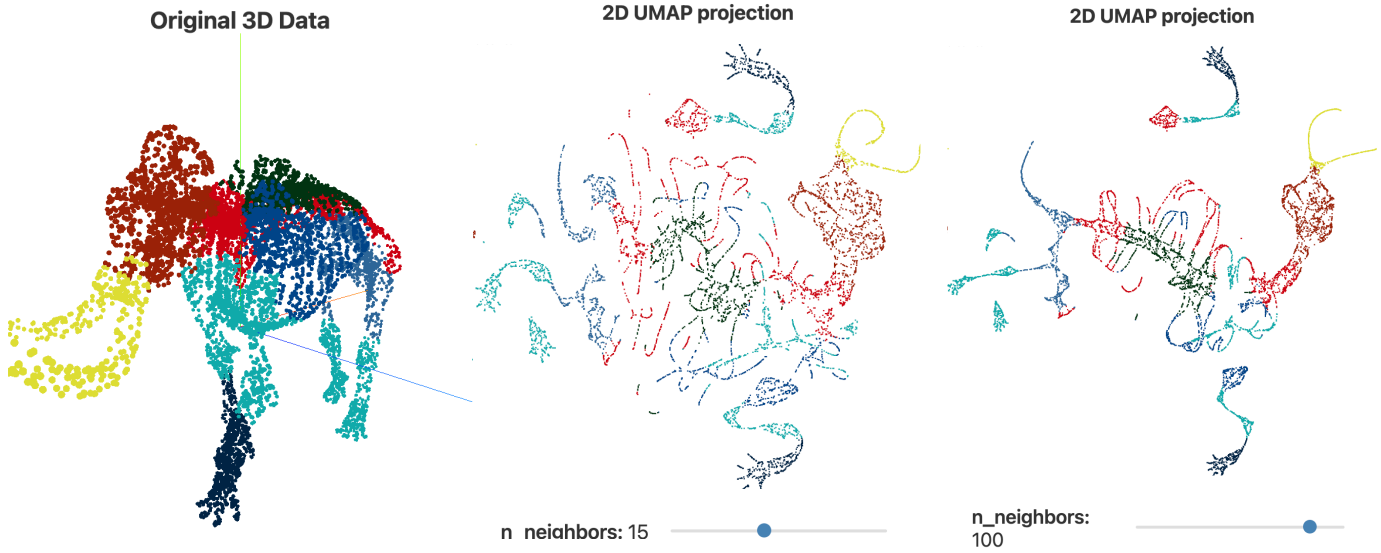
Only for viz, not suitable for downstream analysis

- Clusters might be spurious
- Distance between clusters may not mean anything

Careful tuning based on biology

Without PCA, UMAP is slow





Interactive web app for understanding UMAP:
https://pair-code.github.io/understanding-umap/#:~:text=The%20biggest%20difference%20between%20the,meaningful%20than%20in%20t%2DSNE.

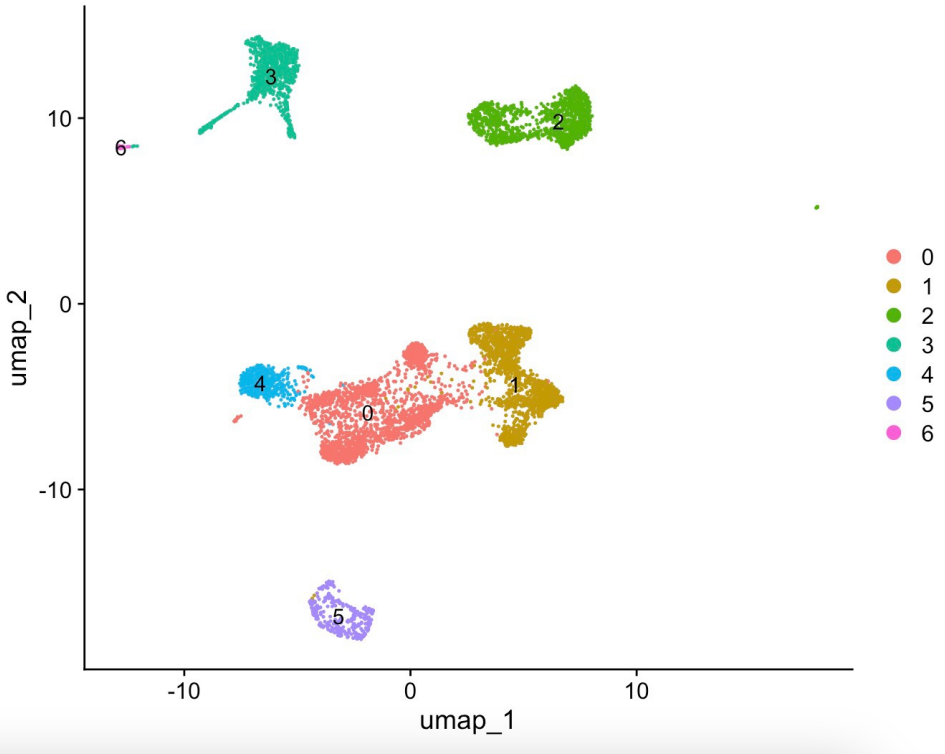
Downstream Analysis: Clustering

Cells that are statistically similar tend to have similar biological functions.

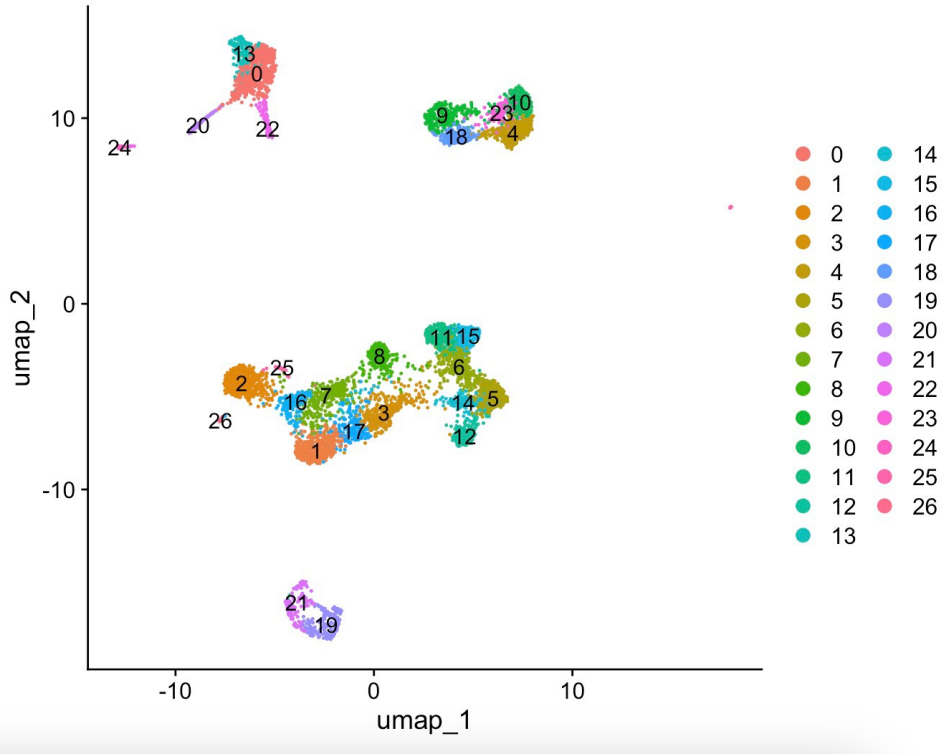
We often use Leiden algorithm or Louvain algorithm to do clustering.

Based on biological problems, resolution can be different.

Resolution=0.1:



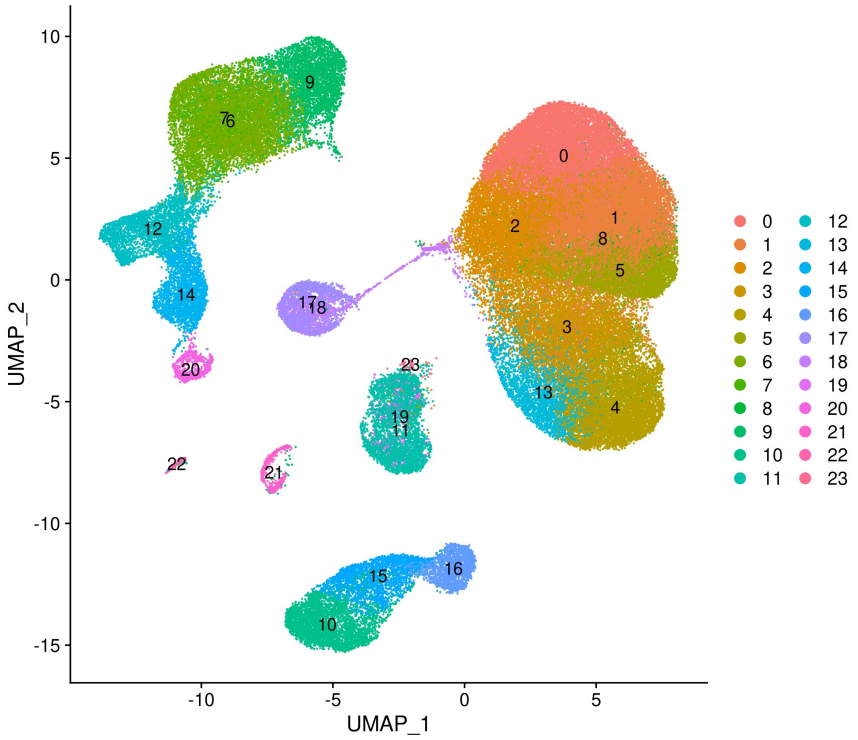
Resolution=3:



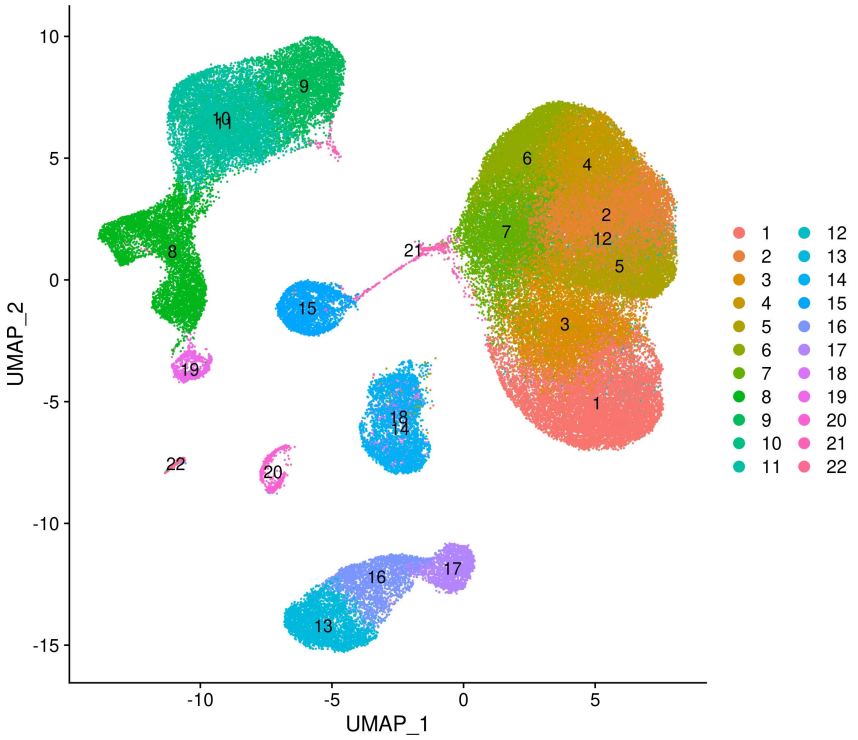
Downstream Analysis: Clustering

Downstream Analysis

Louvain algorithm



Leiden algorithm



Louvain algorithm: Classic, default by Seurat, often good-enough result

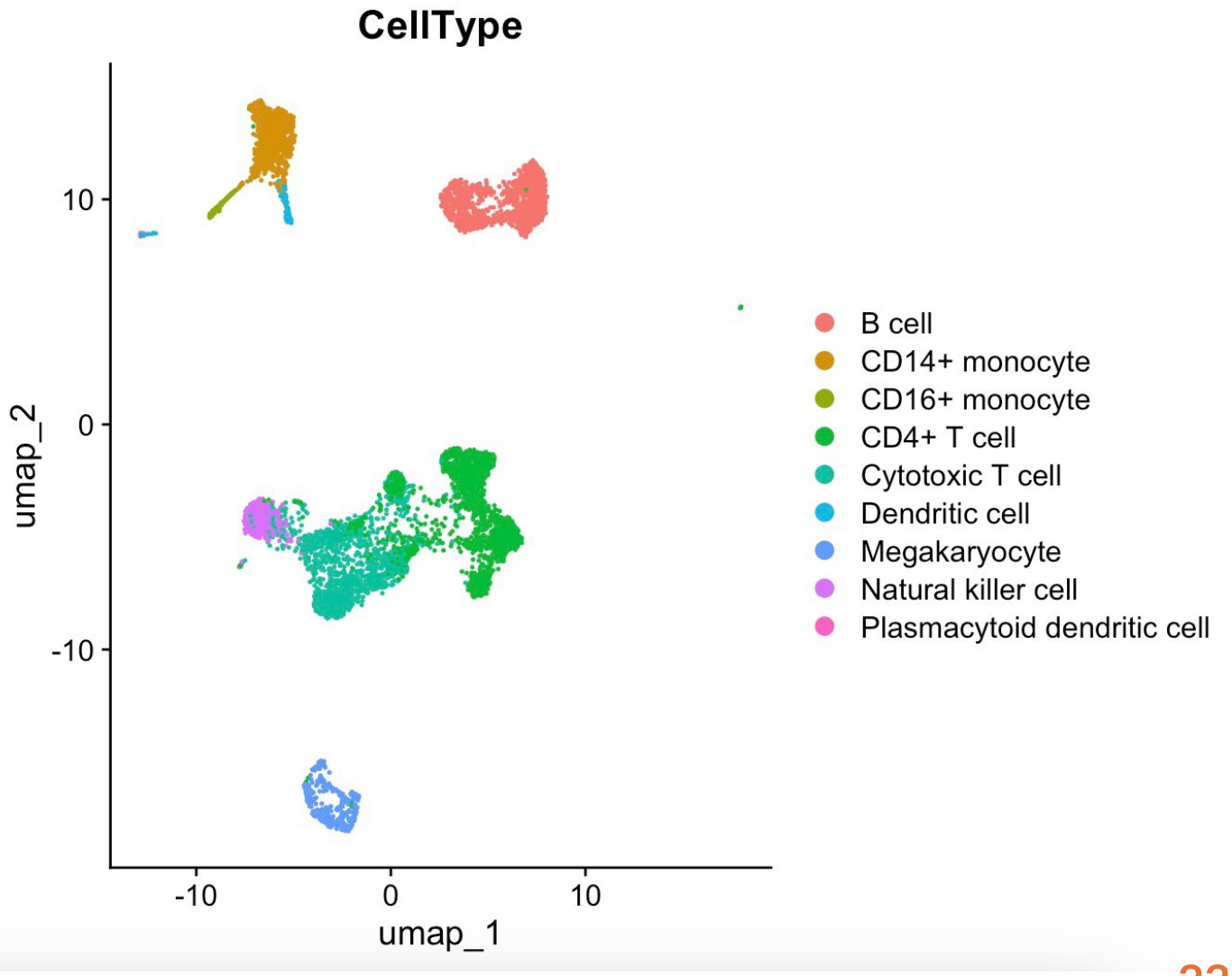
Leiden algorithm: similar to Louvain but faster and leads to better-connected clusters

Downstream Analysis: Cell type annotation

Assign biological annotation to clusters

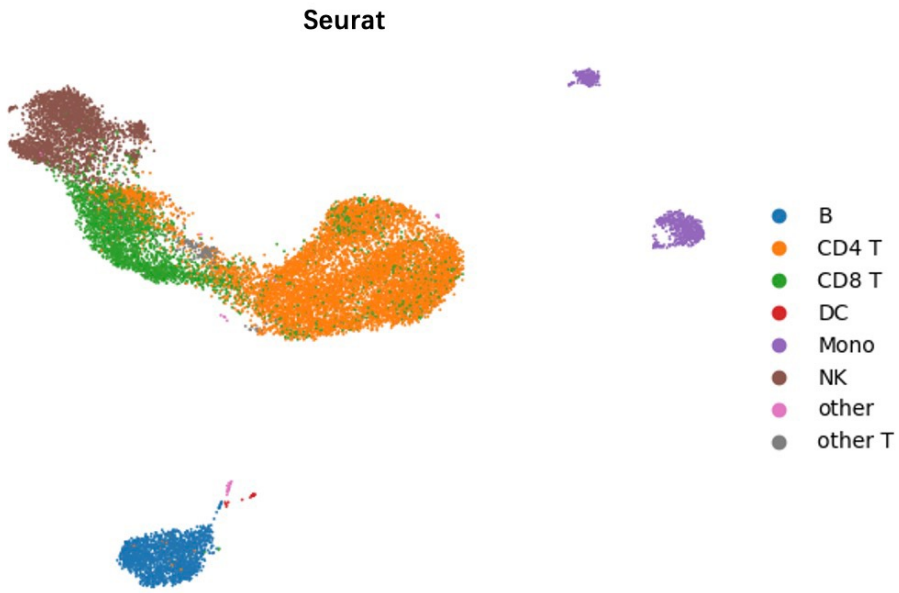
Manually or Automatically

Different cell types have different biological functions

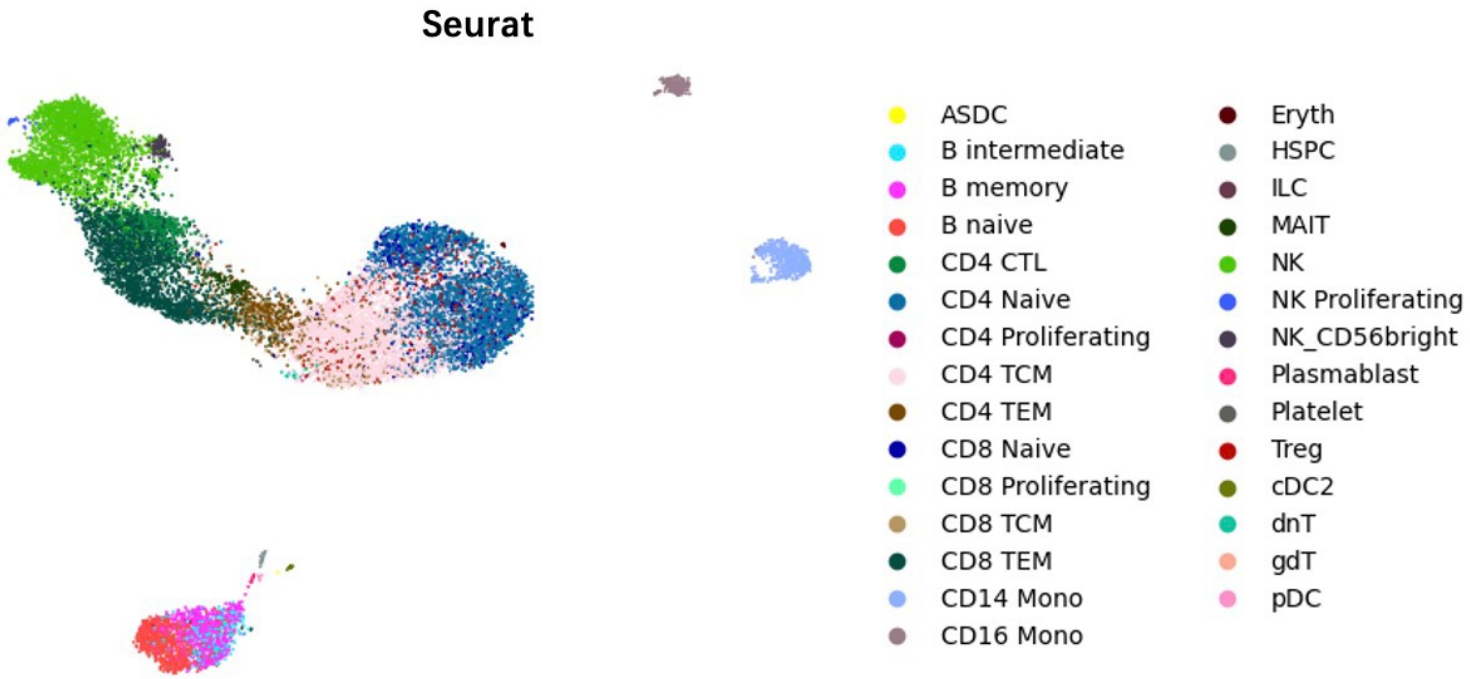


Downstream Analysis: Cell type annotation

Annotation is hierarchical with different resolution
(main label and fine label)



Main label



Fine label

Downstream Analysis: Cell type annotation

Some tools can give every cell an annotation (cell resolution)

Cell Resolution tool:

Higher resolution

Lower accuracy

Cluster Resolution tool:

Lower resolution

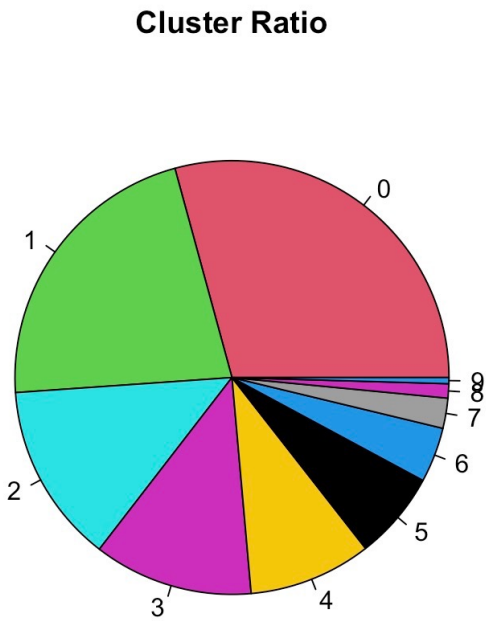
Higher accuracy

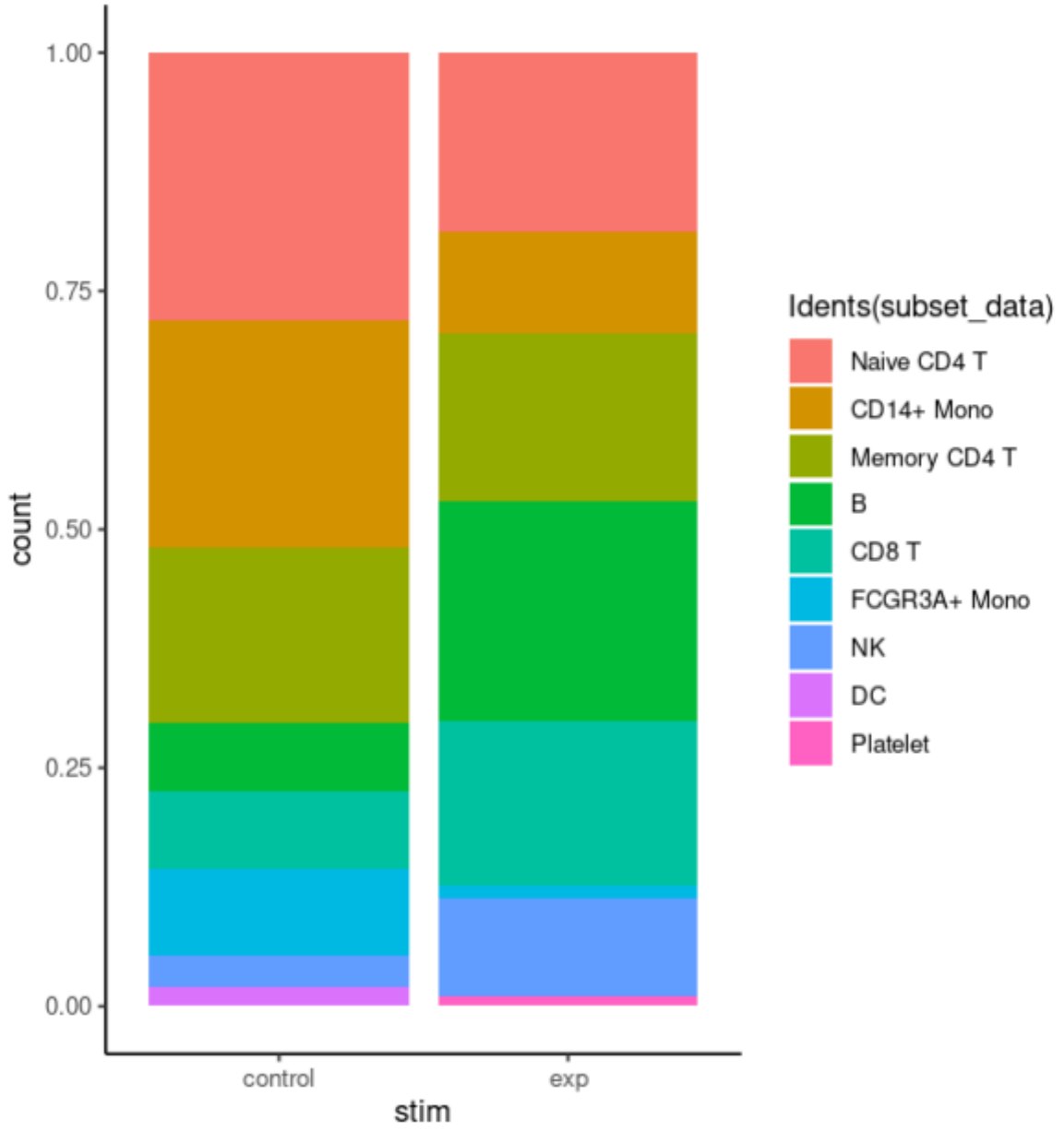
SingleR: Aran et al. (2019).
Seurat V5: Hao al. (2023).

Downstream Analysis: Proportion test

What is the composition of cell types in a tissue?

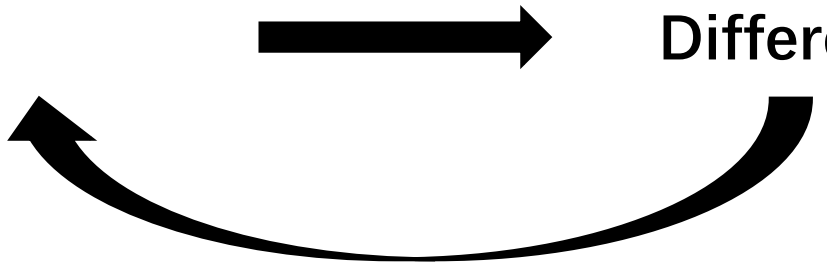
Compare different tissues: Hypothesis testing





Downstream Analysis: Differential expression gene

Downstream Analysis



Challenge: Too many cells, time consuming

Solution: Easy test (Wilcoxon signed-rank test)

If you are interested in a small number of cells, you can choose more sophisticated statistical methods to reduce errors.

Downstream Analysis: Differential expression gene

Downstream Analysis

Output: A gene list with p-value

	p_val	avg_log2FC	pct.1	pct.2	p_val_adj	cluster	gene
CD8A	3.422184e-244	2.8485180	0.781	0.198	1.153071e-239	0	CD8A
GZMH	1.579843e-230	2.2058942	0.841	0.203	5.323123e-226	0	GZMH
CD8B	1.083422e-225	2.8787964	0.704	0.138	3.650481e-221	0	CD8B
NKG7	2.727302e-220	1.7742412	0.993	0.396	9.189372e-216	0	NKG7
CST7	3.763309e-213	1.6251535	0.968	0.304	1.268009e-208	0	CST7
FGFBP2	7.654762e-191	2.0610470	0.759	0.188	2.579195e-186	0	FGFBP2
CCL5	1.375759e-173	1.0483501	0.995	0.427	4.635482e-169	0	CCL5
HLA-B	2.945992e-165	0.7062837	1.000	0.952	9.926224e-161	0	HLA-B
HLA-C	2.977181e-158	0.8237553	0.998	0.925	1.003132e-153	0	HLA-C
CD3D	5.072259e-157	1.6565923	0.882	0.388	1.709047e-152	0	CD3D
TRAC	6.865046e-148	1.6137421	0.853	0.370	2.313109e-143	0	TRAC
GZMM	8.282338e-148	1.7895641	0.781	0.311	2.790651e-143	0	GZMM
MALAT1	8.620873e-140	0.5624246	1.000	0.905	2.904717e-135	0	MALAT1
B2M	2.399743e-138	0.5020925	1.000	0.998	8.085694e-134	0	B2M
GZMA	2.367523e-130	1.3805791	0.773	0.266	7.977132e-126	0	GZMA
CD3E	3.399315e-127	1.2128783	0.958	0.499	1.145365e-122	0	CD3E

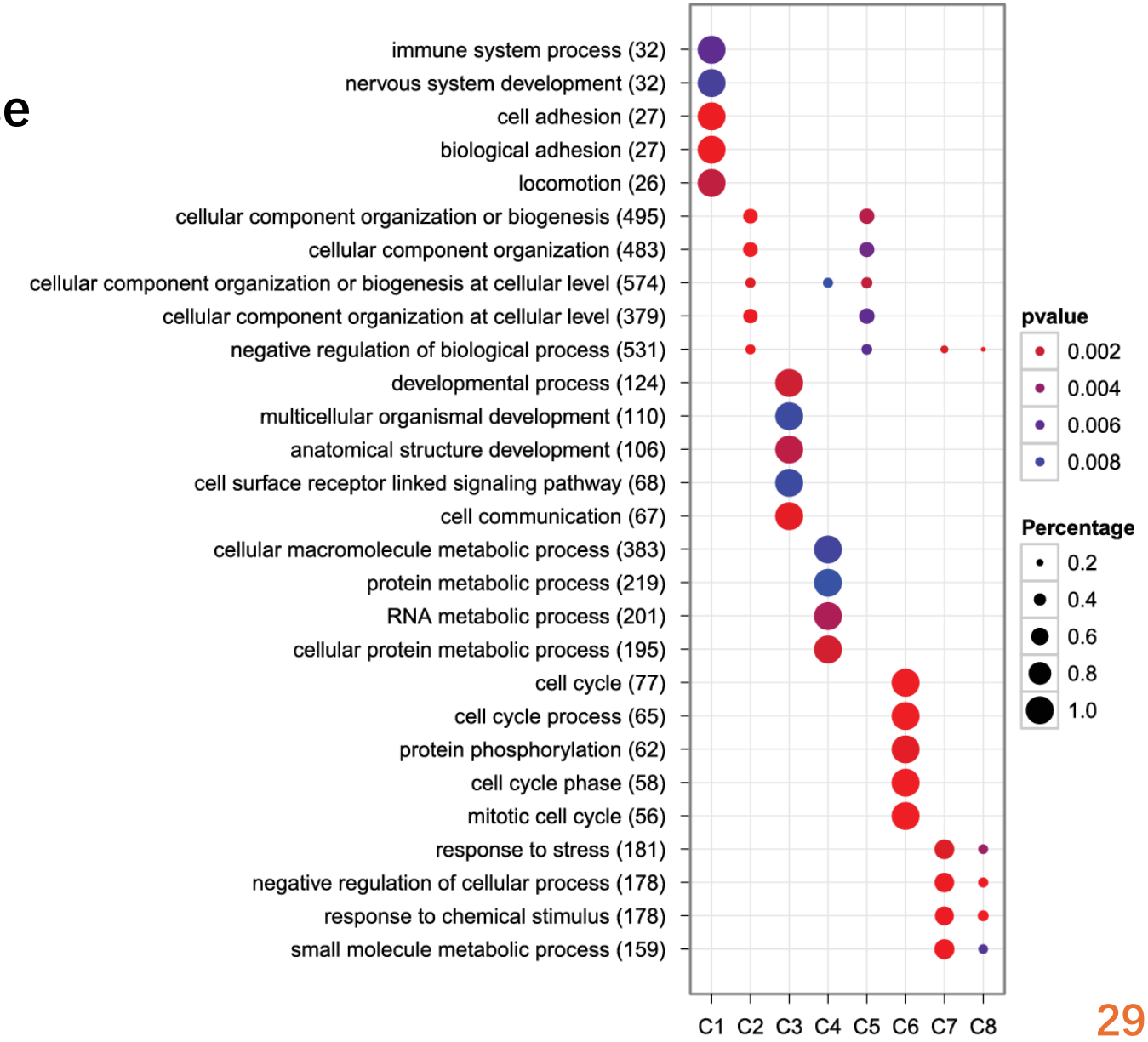
Downstream Analysis: Gene Ontology

Downstream Analysis

Input: Gene list and a biological database

Output: Biological terms and p-value

R package: clusterProfiler



Yu et al. (2012).

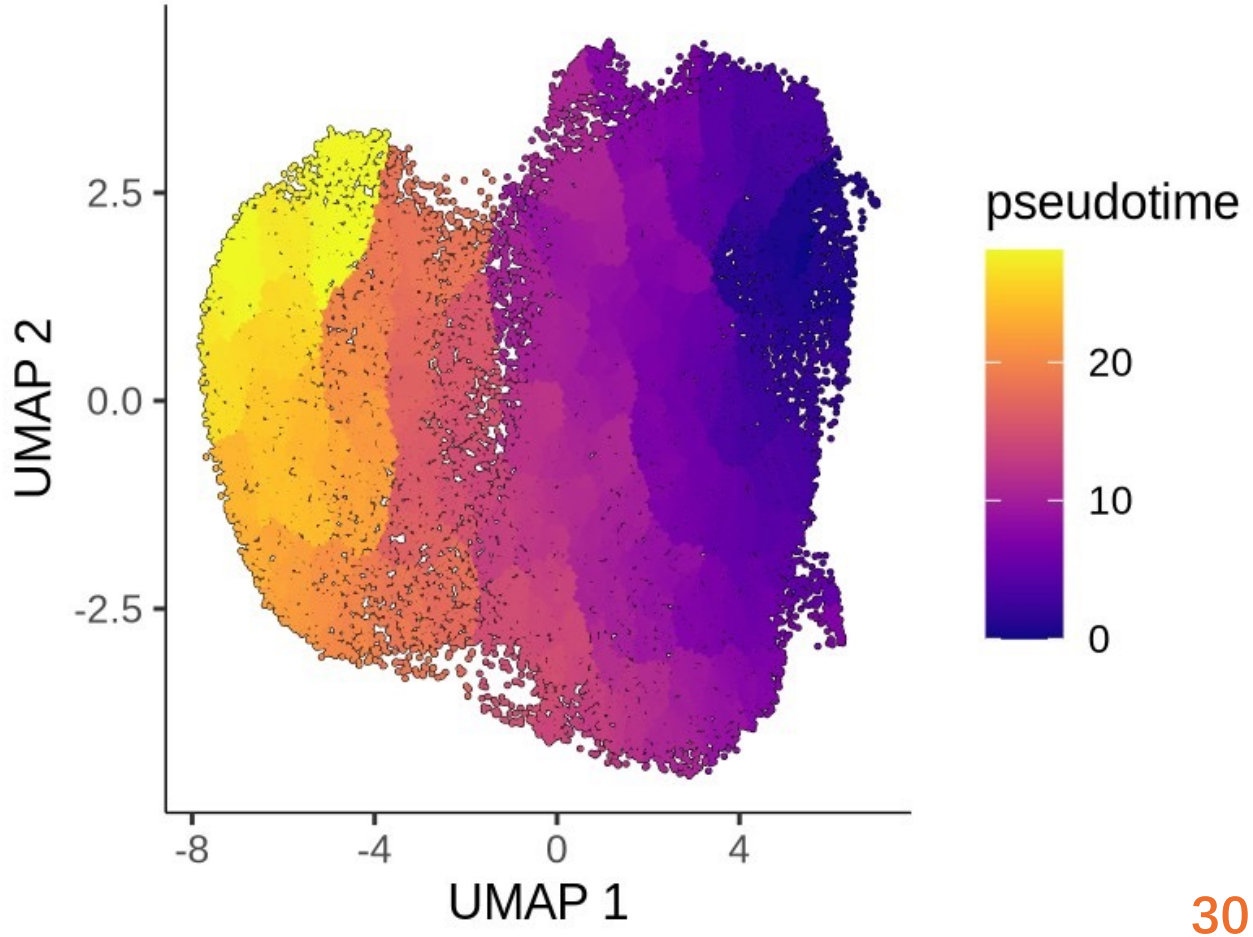
Downstream Analysis: Trajectory inference

Also known as Pseudo-time analysis.

We want to know how cells develop.

Monocle can infer how cells develop (grow, differentiate) from some cells to some other cells.

Also outputs a list of genes associated with development.



Trapnell et al. (2014).

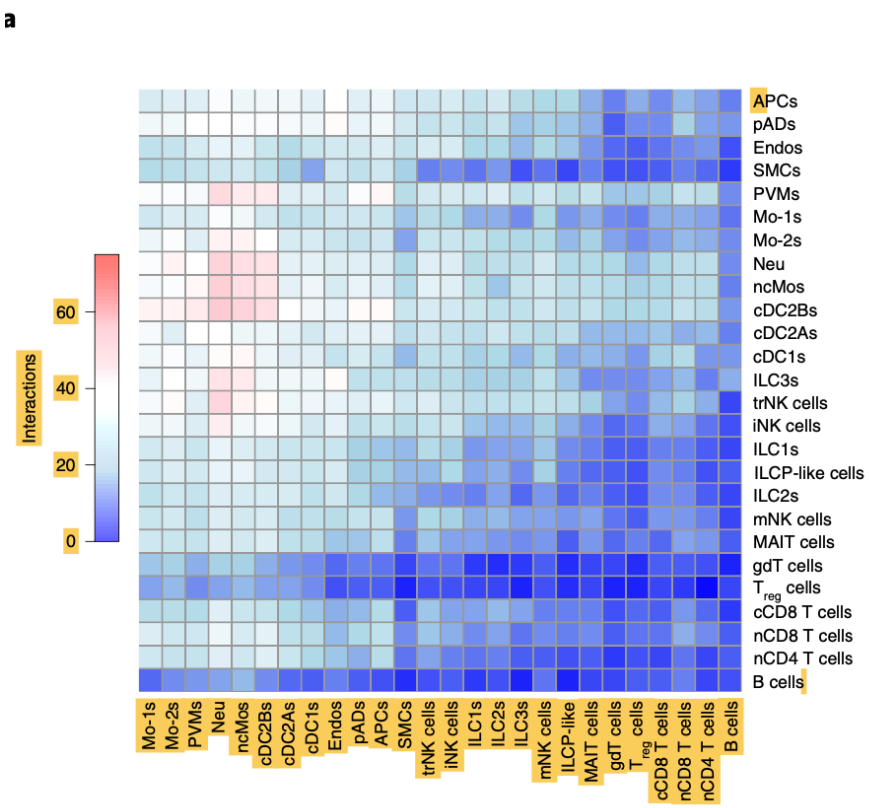
Downstream Analysis: Cell-cell interaction

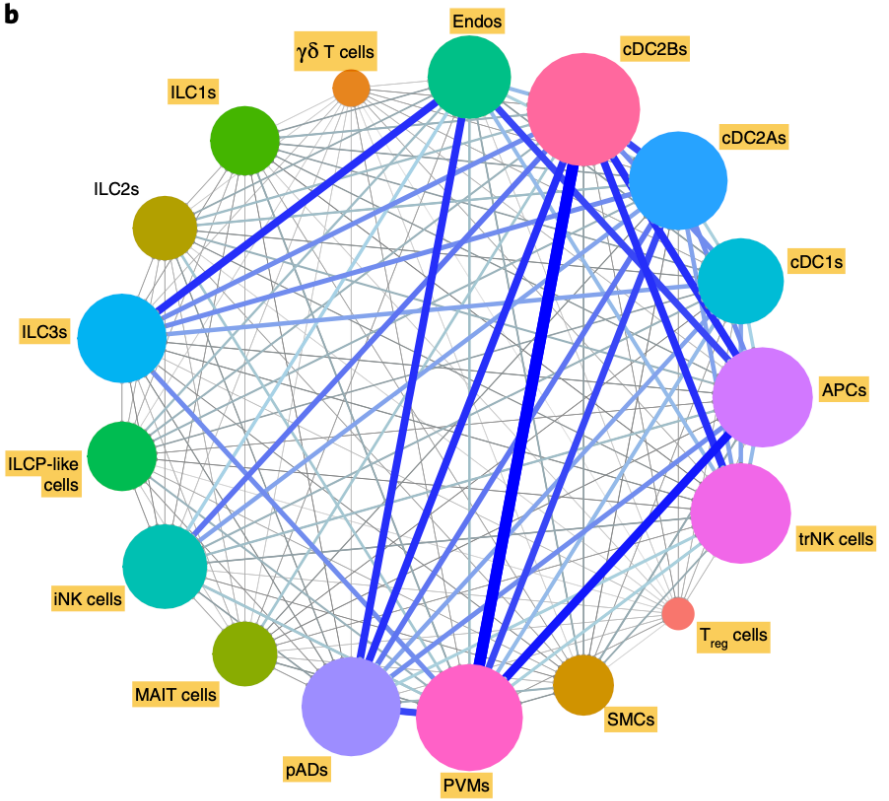
Downstream Analysis

Understand interaction between cell types

Make sure which genes are important in the cell-cell interaction

Software: CellPhoneDB

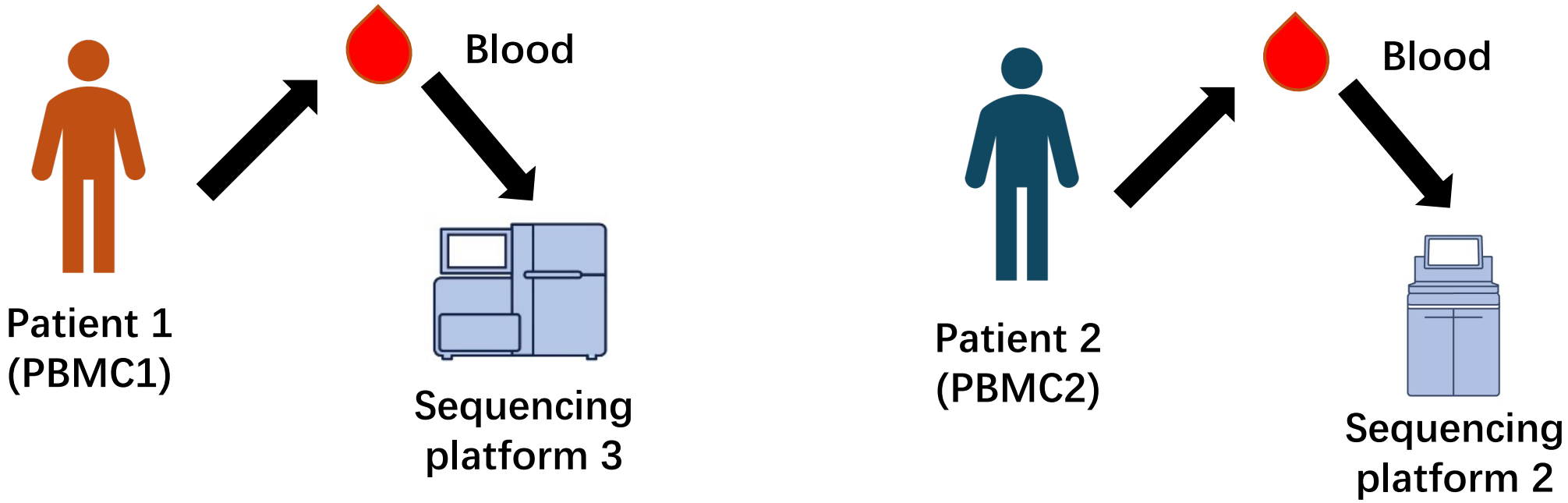




Efremova et al. (2020).

Case study (practical session)

Single-cell sequencing of peripheral blood mononuclear cells (PBMC) from two patients.



Challenge: Remove batch effect between two experiments*

* Batch effects can arise from patients, protocols, etc.

Thank you!

My Email:
xiaoczhang3@student.unimelb.edu.au