Introduction to scRNA-seq workflow

MPG Primer Series 2/5/2025

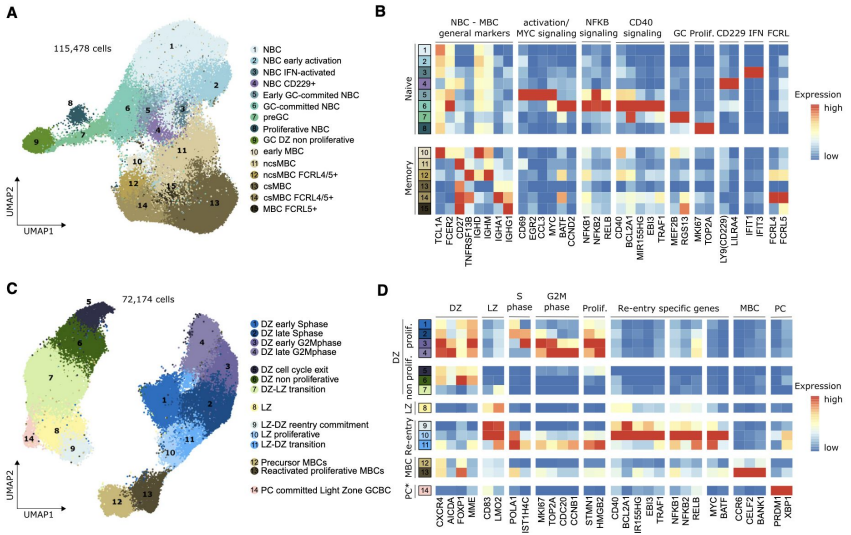
Marc Elosua Bayes

Agenda

- Basics Steps - from GEX to graph
- Quality Control
- Batch Effect & Integration Assessment
- Data Resources
- Tools Resources

Why use scRNA-seq

- Understand Cellular Heterogeneity
- Discover Novel and Rare Cell States
- Map Lineages and Developmental Trajectories
- Insights into Disease Mechanisms



Massoni-Badosa, Ramon et al. “An atlas of cells in the human tonsil.” *Immunity, 2024*

Basics Steps -

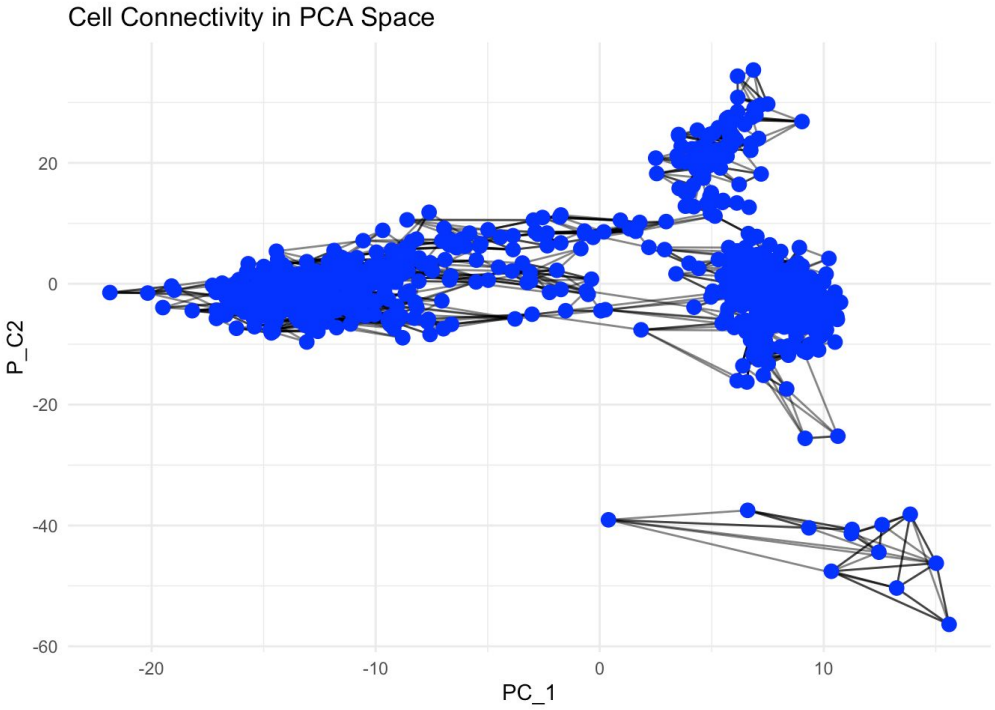
from GEX to graph

The Goal - KNN graph

Ultimately we want to build a KNN/SNN graph of our dataset.

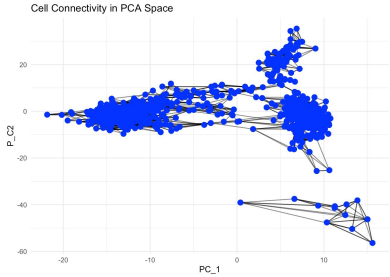
Each node is a cell and each edge is the similarity between two cells.

But… how do we measure this similarity?



Where we start - Count Matrix

		100K cells			
		Cell1	Cell2	...	CellN
30K genes	Gene1	3	2	.	13
	Gene2	2	3	.	1
	Gene3	1	14	.	18
	...	.	.	.	.
	...	.	.	.	.
	...	.	.	.	.
	GeneM	25	0	.	0



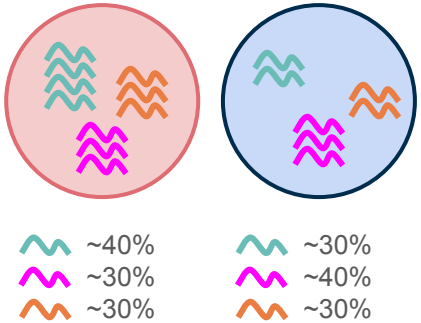
⛔ We can’t compute distances in this high dimensional space

⛔ Not memory efficient or computationally feasible

✅ We need to reduce the dimensions

Intermediate steps - Data Normalization

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0

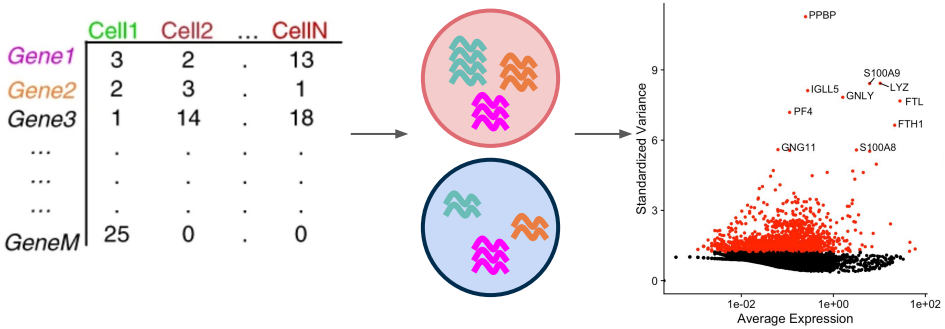


⛔ We can’t compare raw counts

⛔ scRNAseq has a lot of technical noise

✅ We need to normalize by library size and log

Intermediate steps - Feature Selection



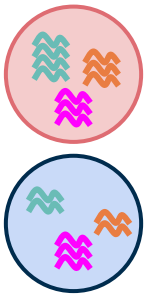
⛔ We can’t do PCA using all genes

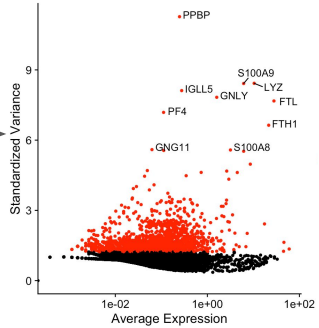
✅ We aim to to subset our genes to the most biologically relevant

✅ We select top X most informative (variable) genes

Intermediate steps - Feature Selection

	Cell1	Cell2	...	CellN
Gene1	3	2	.	13
Gene2	2	3	.	1
Gene3	1	14	.	18
...	.	.	.	.
...	.	.	.	.
...	.	.	.	.
GeneM	25	0	.	0



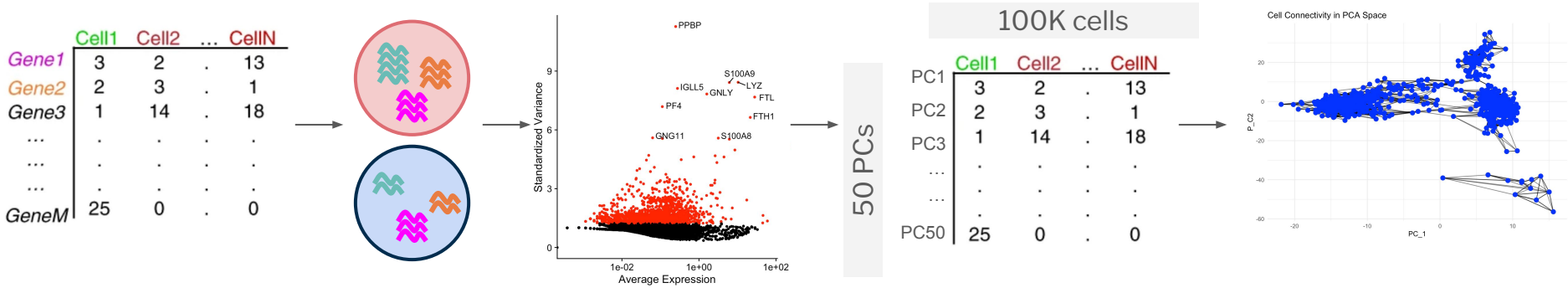


⛔ We can't do PCA using all genes

✅ We aim to to subset our genes to the most biologically relevant

✅ We select top X most informative (variable) genes

Intermediate steps - PCA

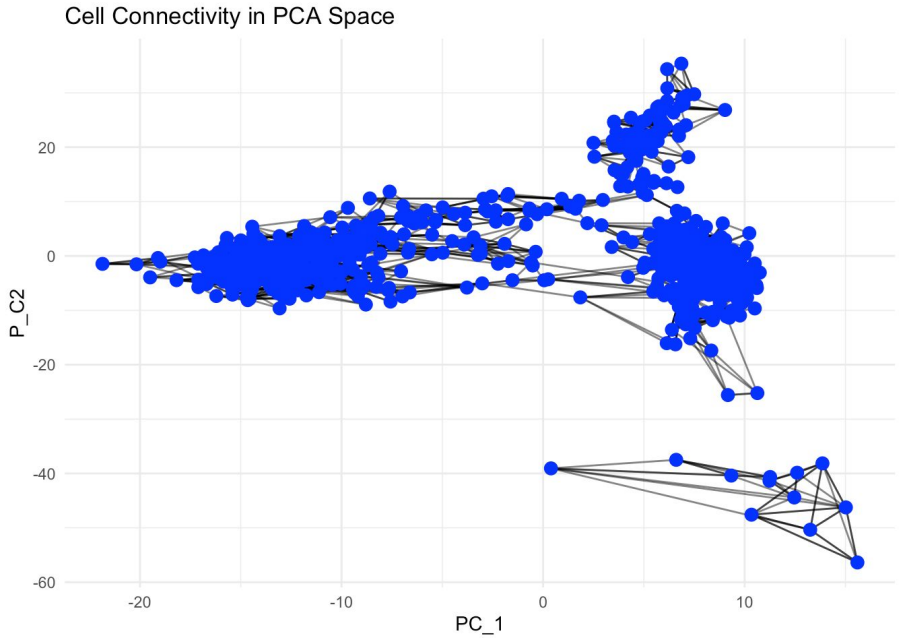


⛔ We can’t compute distances in this high dimensional space

✅ PCA reduces the dimensions to <50 learning orthogonal information

✅ We can also use NMF, scVI, SVD…

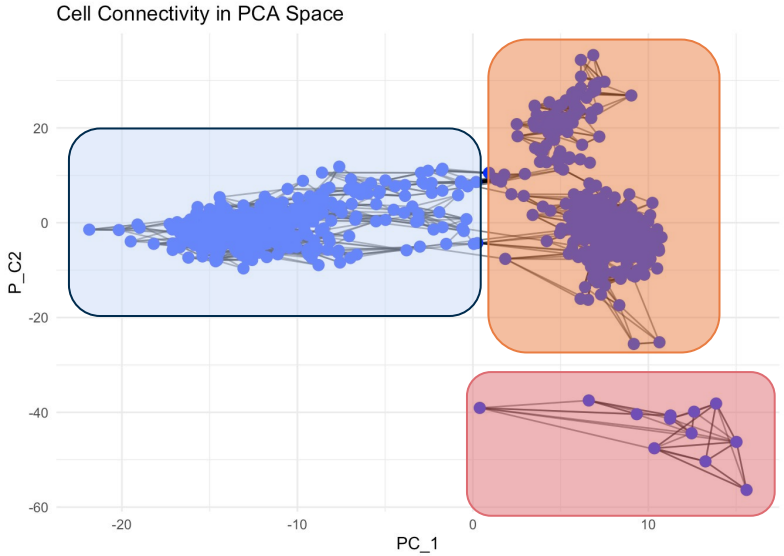
Downstream steps - Clustering



⛔ Classical methods like k-means or hierarchical clustering don’t scale

✅ We can use community detection algorithms on the SNN -

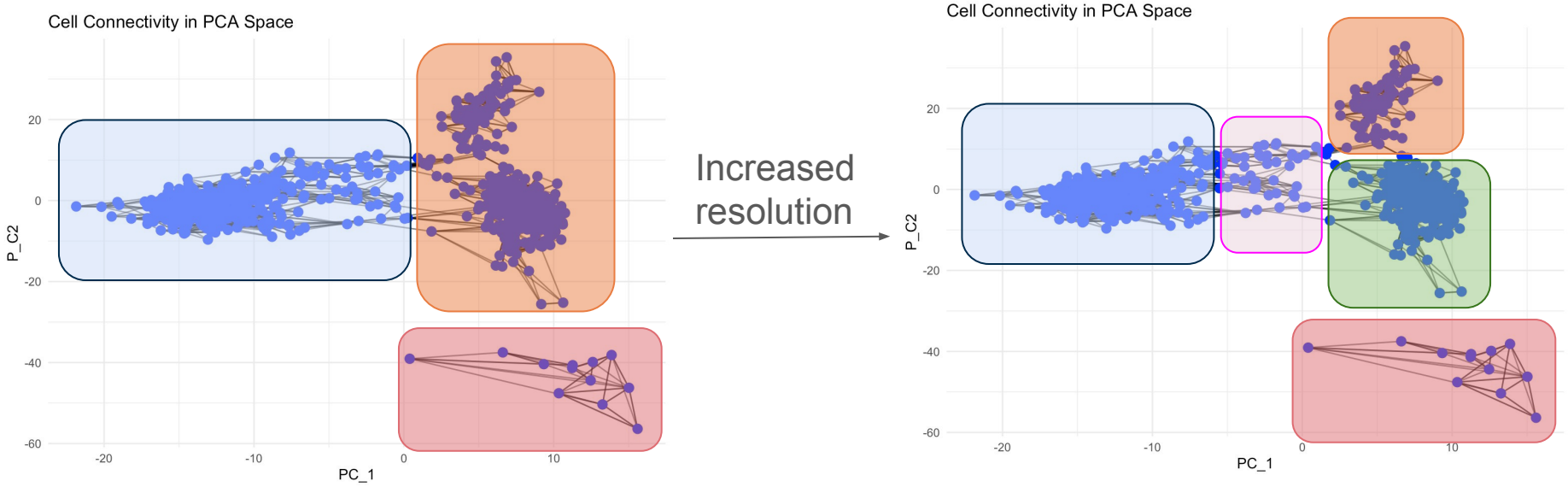
Downstream steps - Clustering



⛔ Classical methods like k-means or hierarchical clustering don’t scale

✅ We can use community detection algorithms on the SNN -

Downstream steps - Clustering

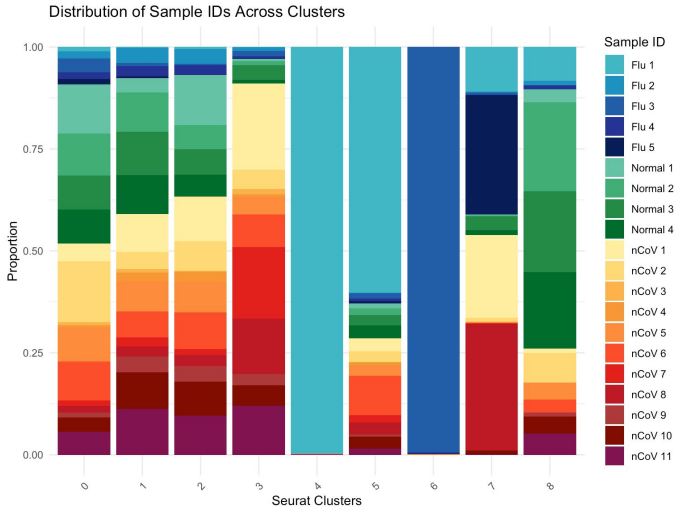


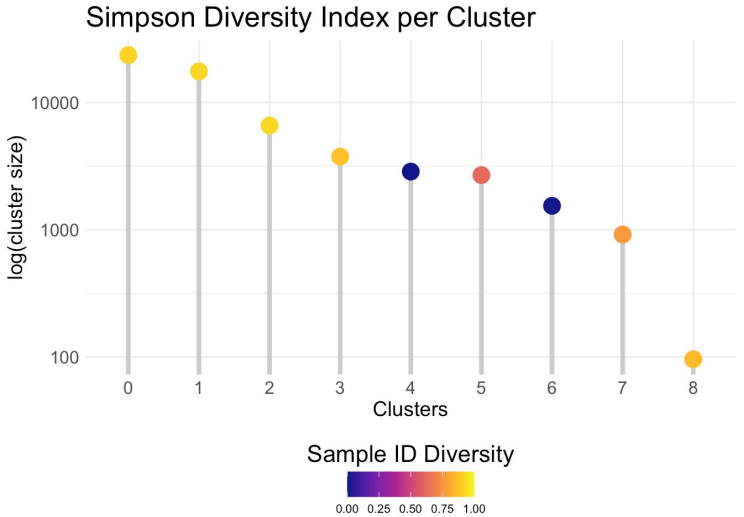
✅ Community detection algorithms have a resolution parameter

Downstream steps - Clustering Assessment

Cluster assessment

- Simpson diversity - batch specific clusters

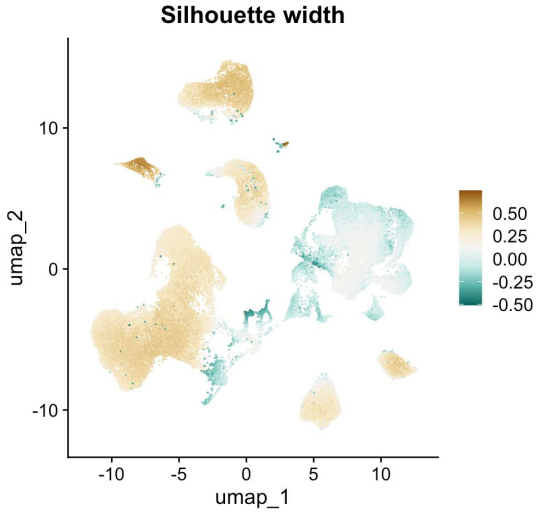


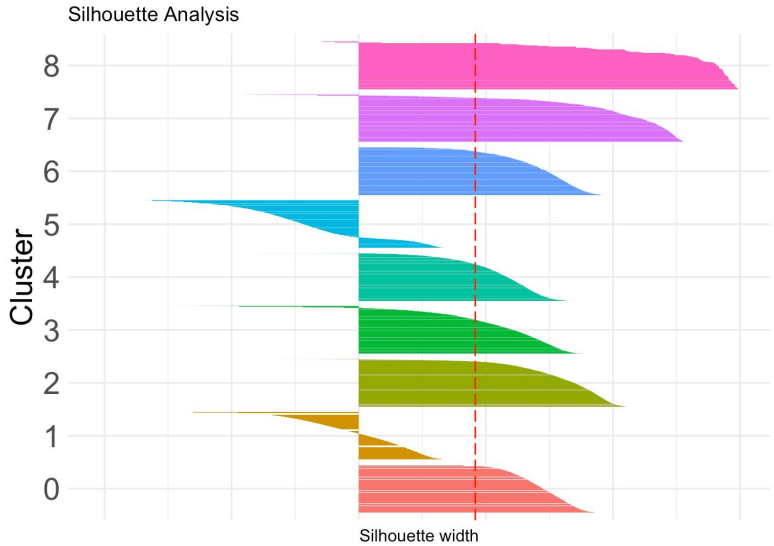


Downstream steps - Clustering Assessment

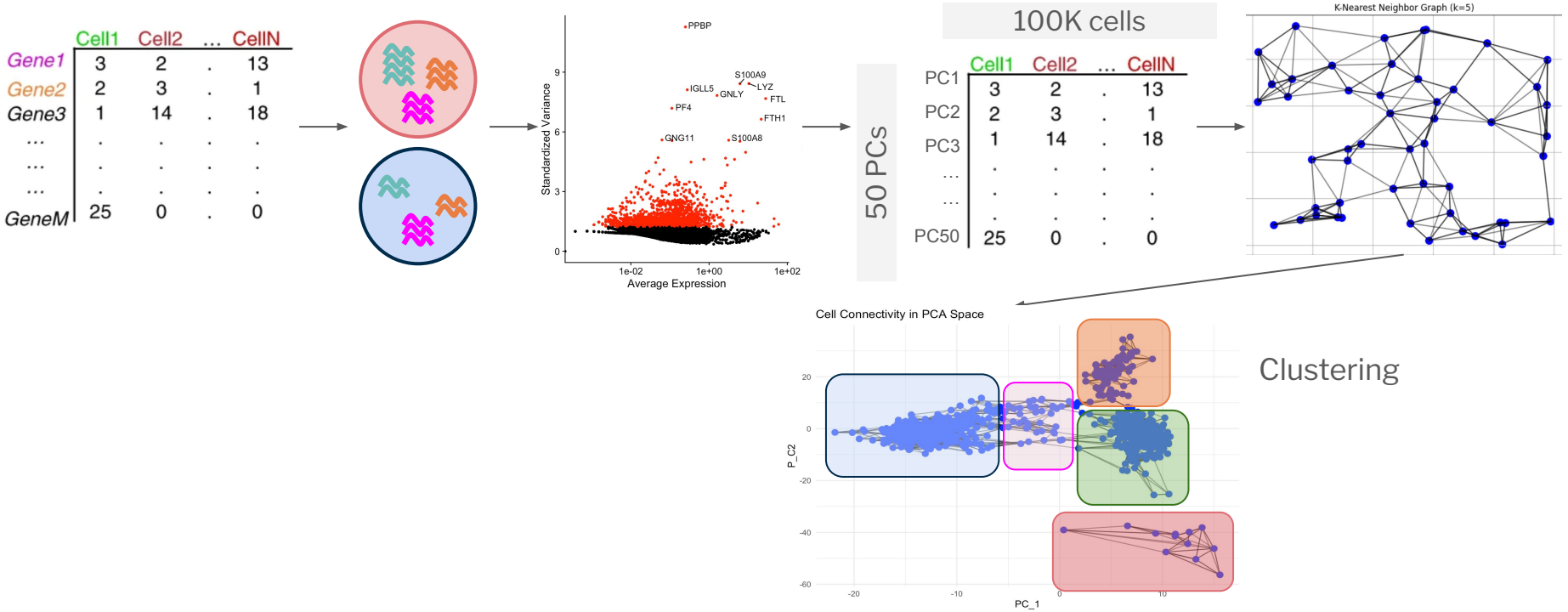
Cluster assessment

- Simpson diversity - batch specific clusters
- Silhouette score

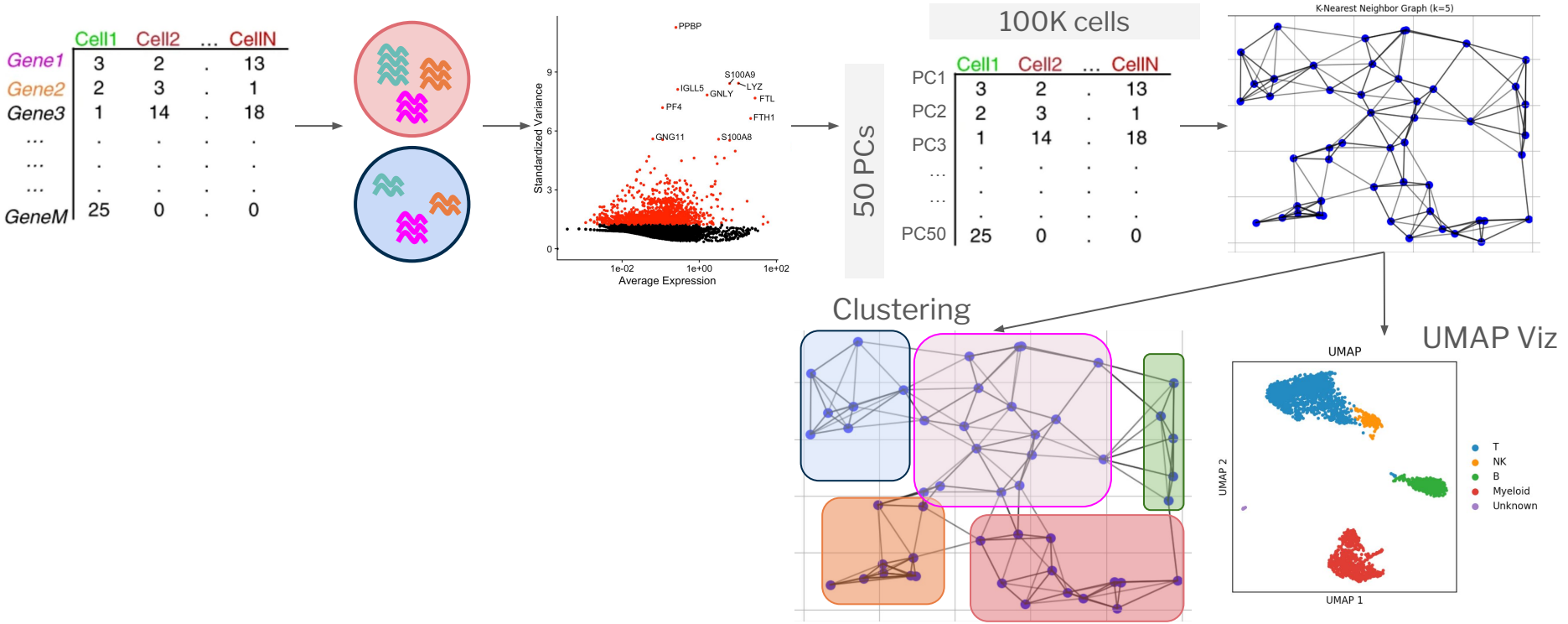




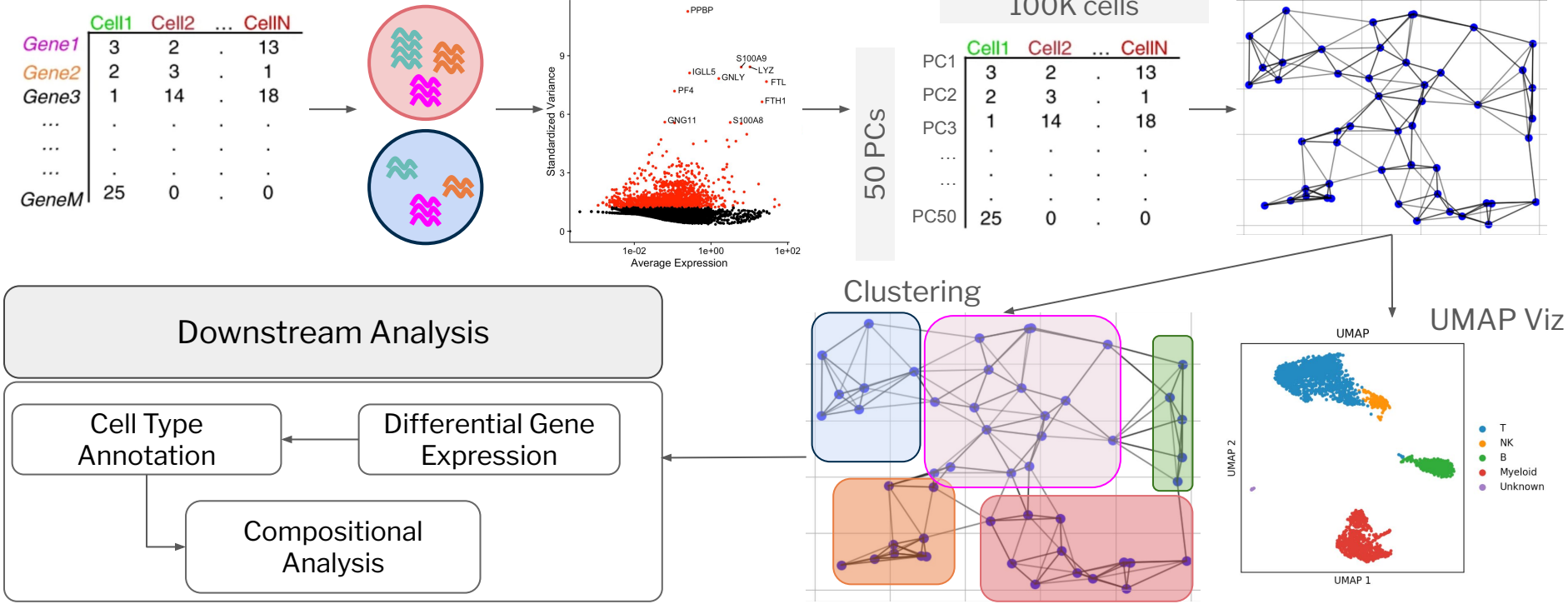
Downstream steps - Clustering



Downstream Analysis

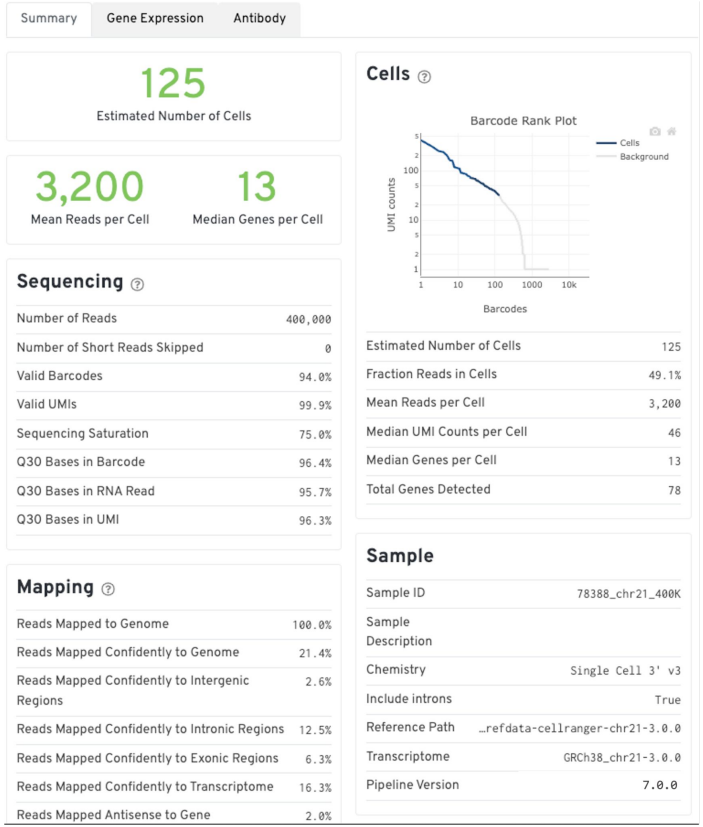


Downstream Analysis



Quality Control

Sample-Level Metrics



Goal: Identify issues at the sample level

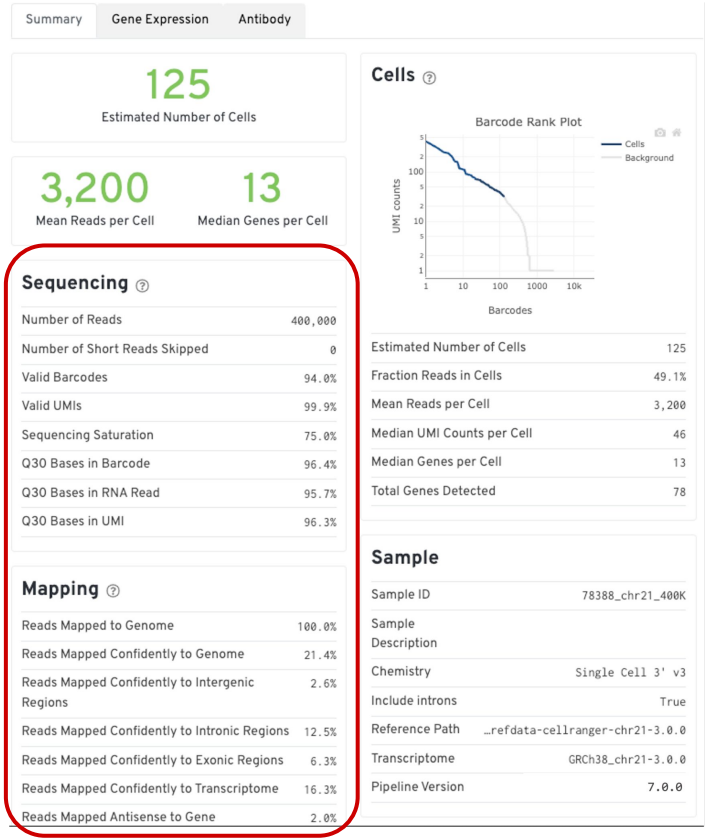
- Low number of cells
- Poor sequencing quality
- Under-sequenced samples
- Mapping quality issues

https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/outputs/cr-outputs-web-summary-count

Sample-Level Metrics

Sequencing & Alignment Quality

- Sequencing Depth
- UMI & Barcode detection
- Sequencing Confidence
- Mapping to Reference Genome

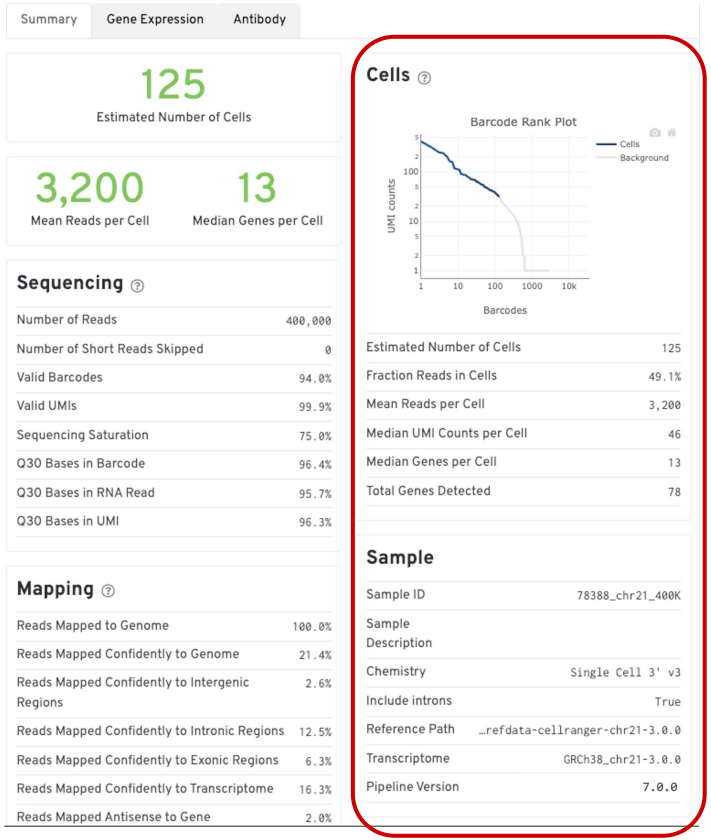


https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/outputs/cr-outputs-web-summary-count

Sample-Level Metrics

Sample & Cell Quality

- Total # of Cells & Reads
- UMIs & Genes per Cell

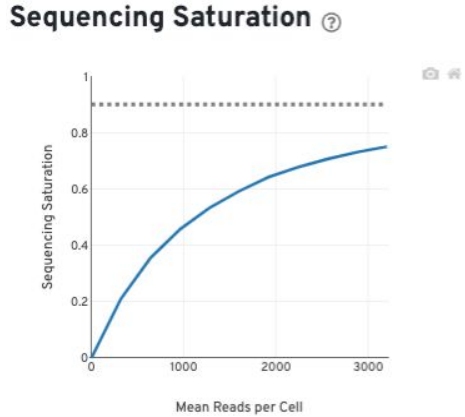


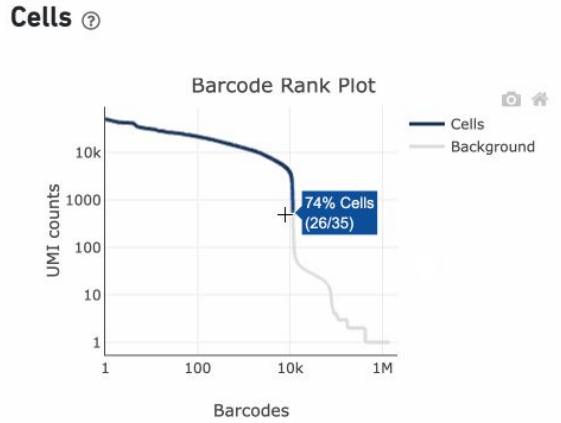
https://www.10xgenomics.com/support/software/cell-ranger/latest/analysis/outputs/cr-outputs-web-summary-count

Sample-Level Metrics

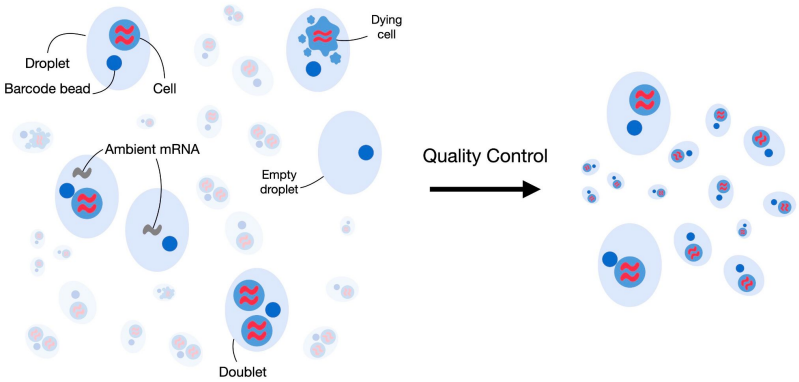
Sample & Cell Quality

- Total # of Cells & Reads
- UMIs & Genes per Cell
- Sequencing Saturation
  - Trade-offs from going deeper
- Barcode Rank Plot
  - Keep an eye out for low-complexity cells!





Cell-Level Metrics



Goal: Identify issues at the cell level

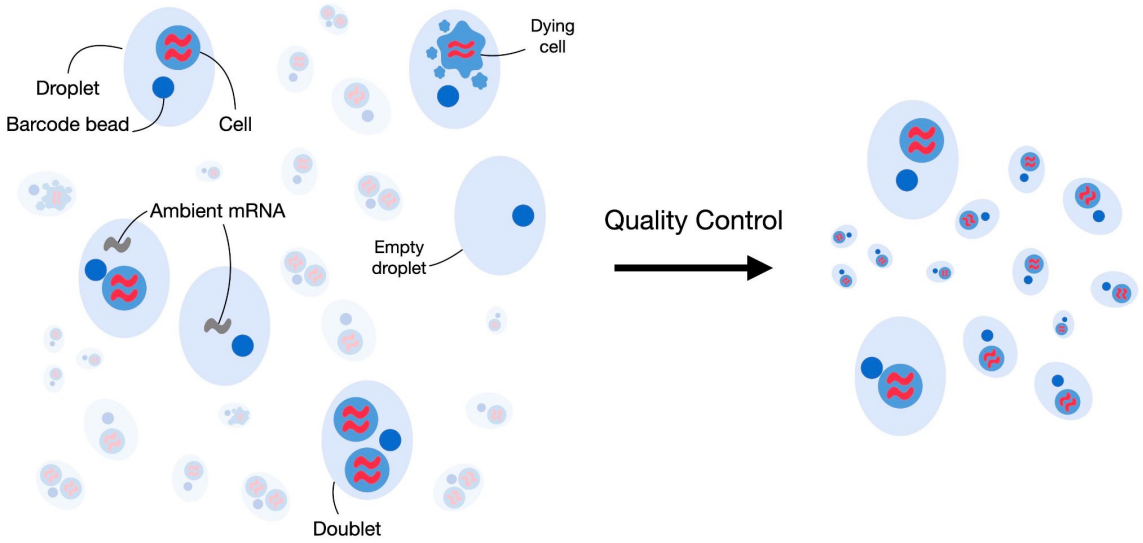
- Empty droplets
- Doublet detection
- Remove noisy artifacts

https://www.sc-best-practices.org/preprocessing_visualization/quality_control.html

Cell-Level Metrics

Key metrics to assess

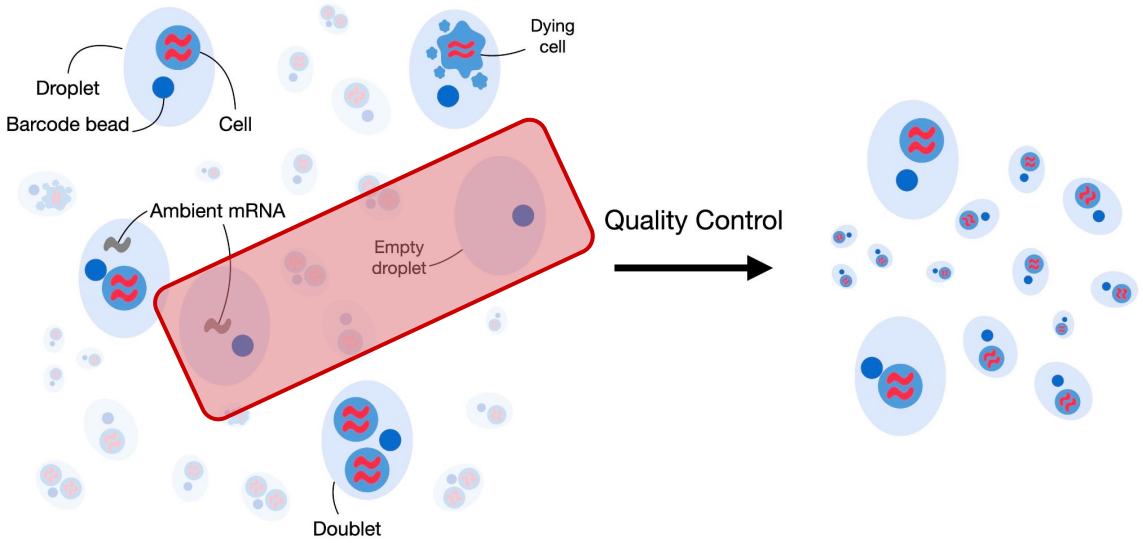
- Library Size
- Library Complexity
- Mitochondrial %
- Doublet Scores



Cell-Level Metrics

Key metrics to assess

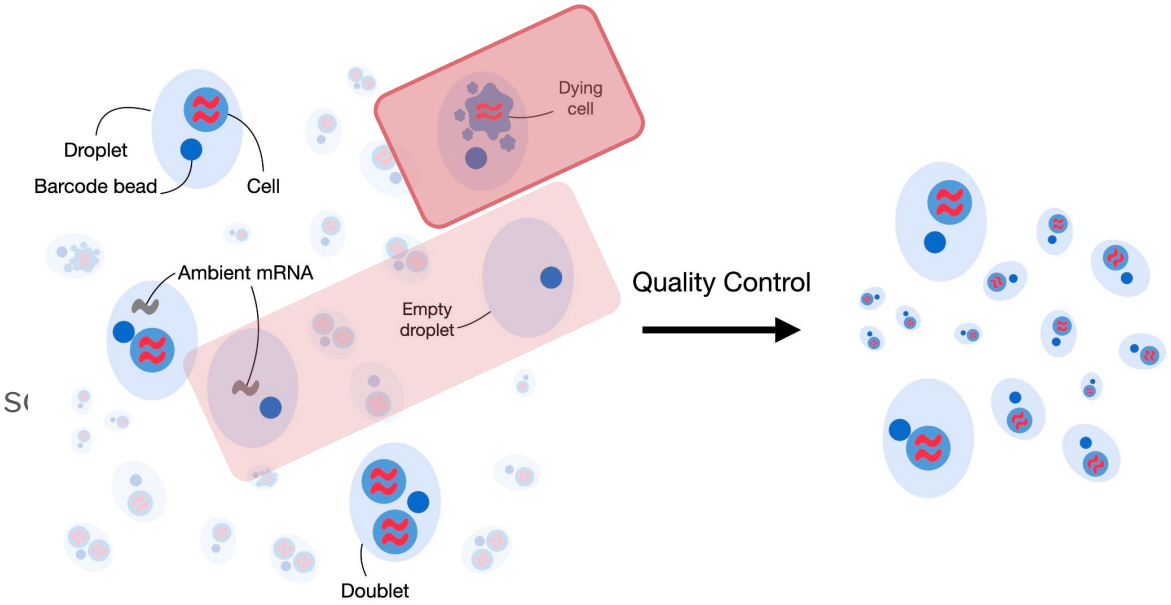
- Library Size
- Library Complexity
- Mitochondrial %
- Doublet Scores



Cell-Level Metrics

Key metrics to assess

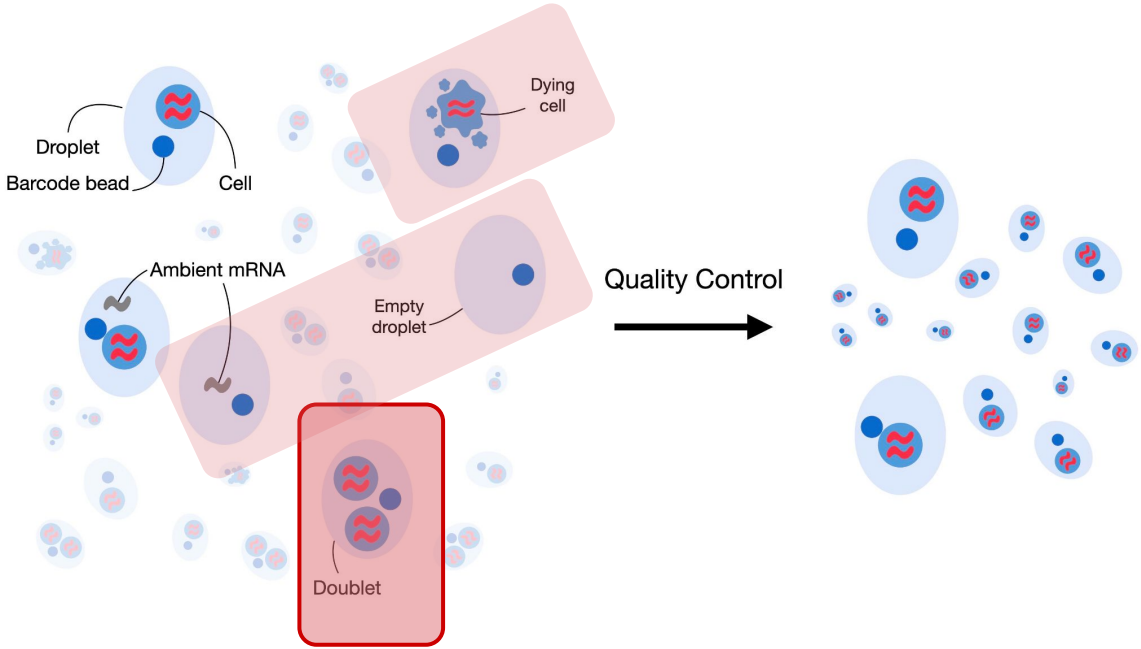
- Library Size
- Library Complexity
- Mitochondrial %
  - Depends on vs sn-RNA
- Doublet Scores



Cell-Level Metrics

Key metrics to assess

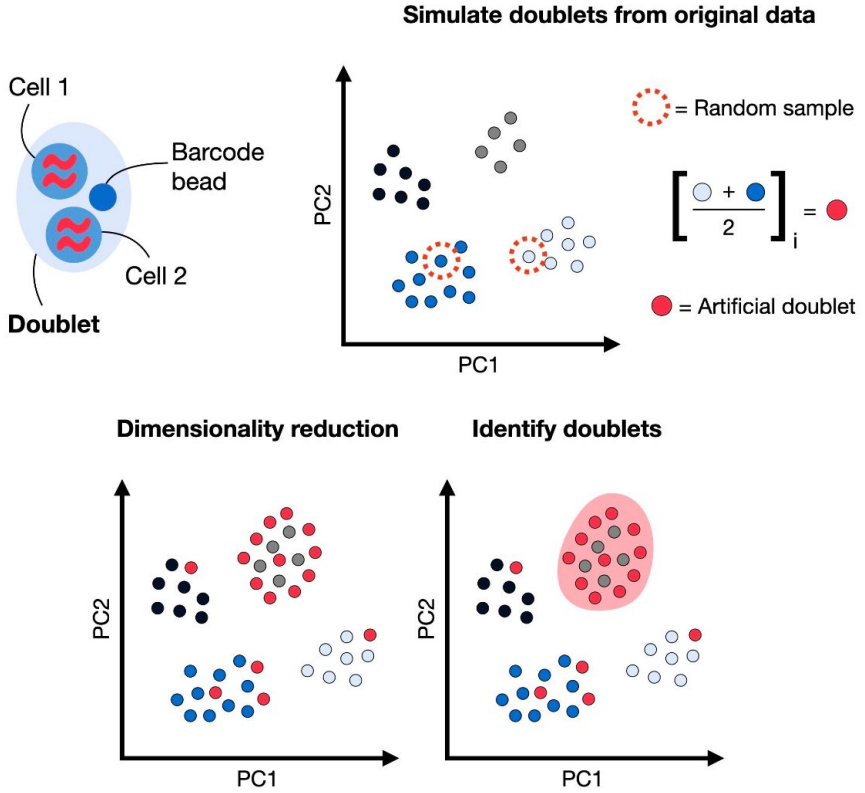
- Library Size
- Library Complexity
- Mitochondrial %
- Doublet Scores



Cell-Level Metrics - Doublet Scores

Doublet Detection Algorithms

1. Compute reduced dimension
2. Generate synthetic doublets
3. Add synthetic doublets to embedding
4. Look at neighborhood enrichment

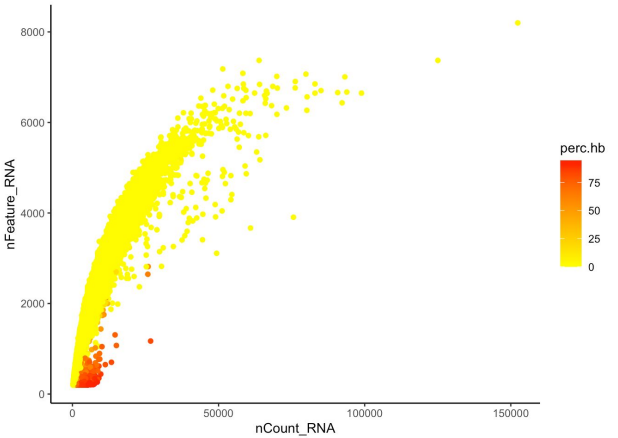


Tools

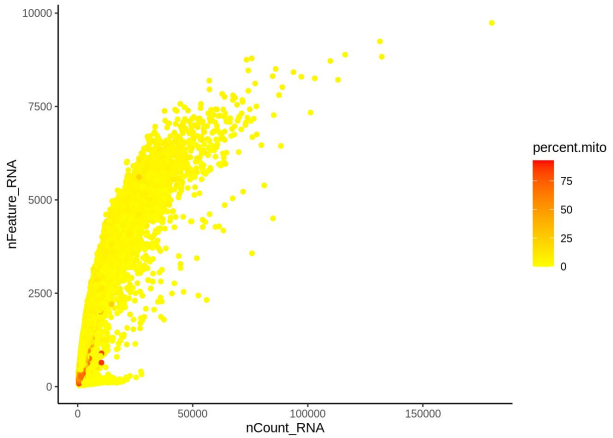
Scrublet / DoubletFinder / DoubletDetection / DoubletDecon
...

Cell-Level Metrics - Feature Covariation

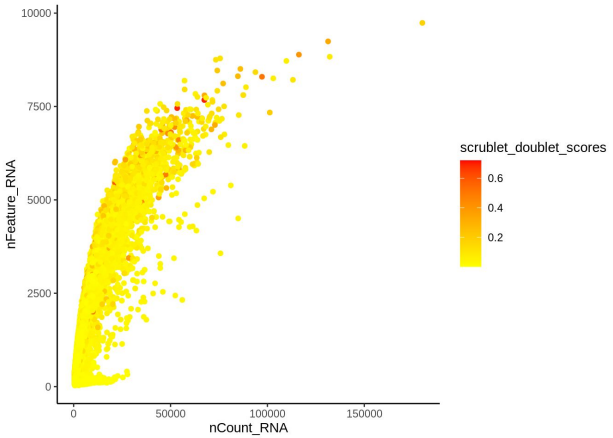
Hemoglobin %



Mitochondrial %

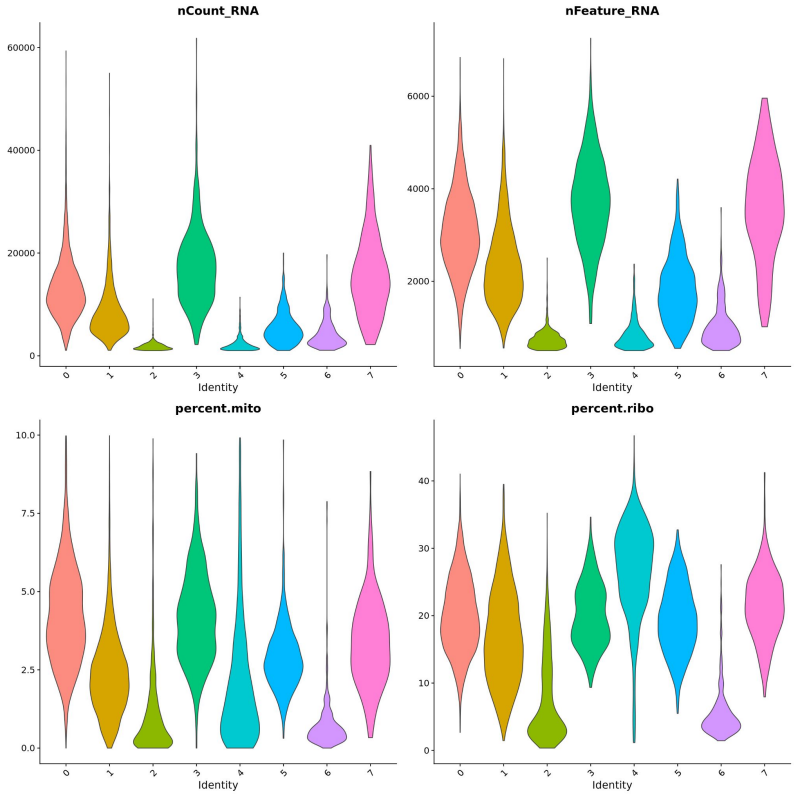


Doublet Scores %

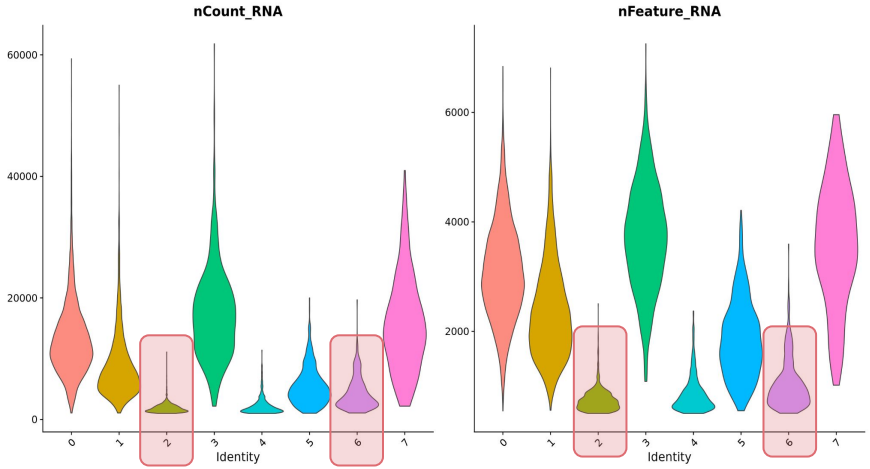


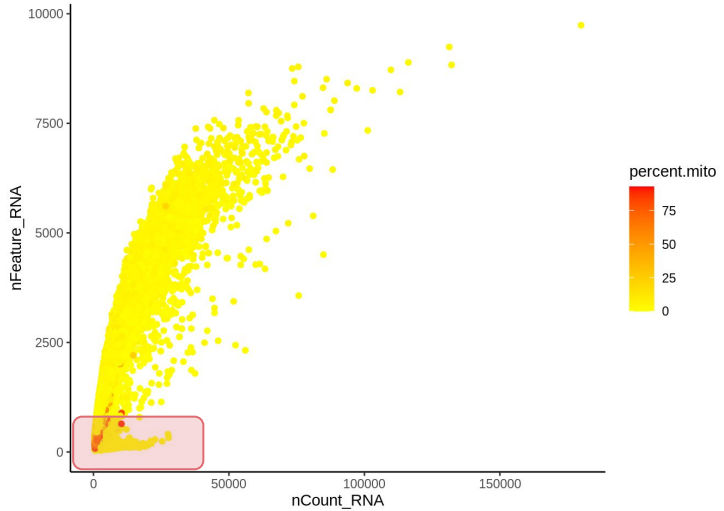
Cell-Level Metrics - Feature Covariation

Let’s look at an example…



Cell-Level Metrics - Feature Covariation



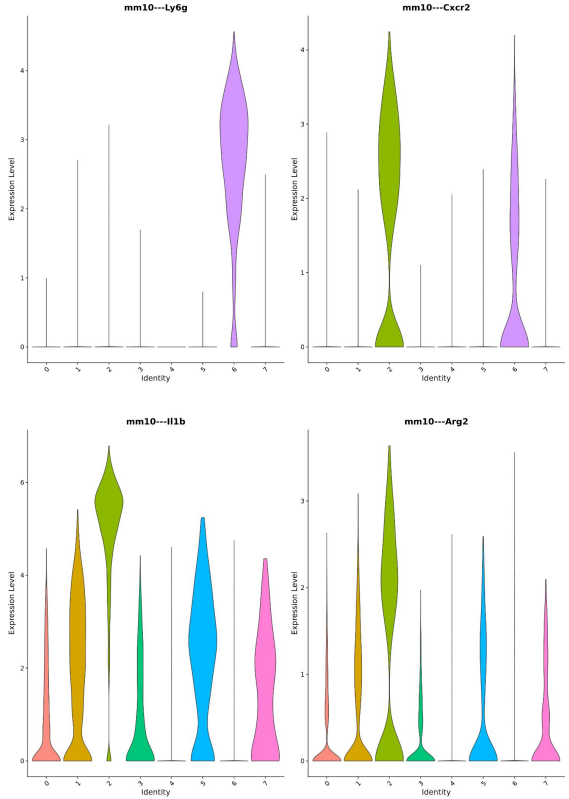


Low Quality ?

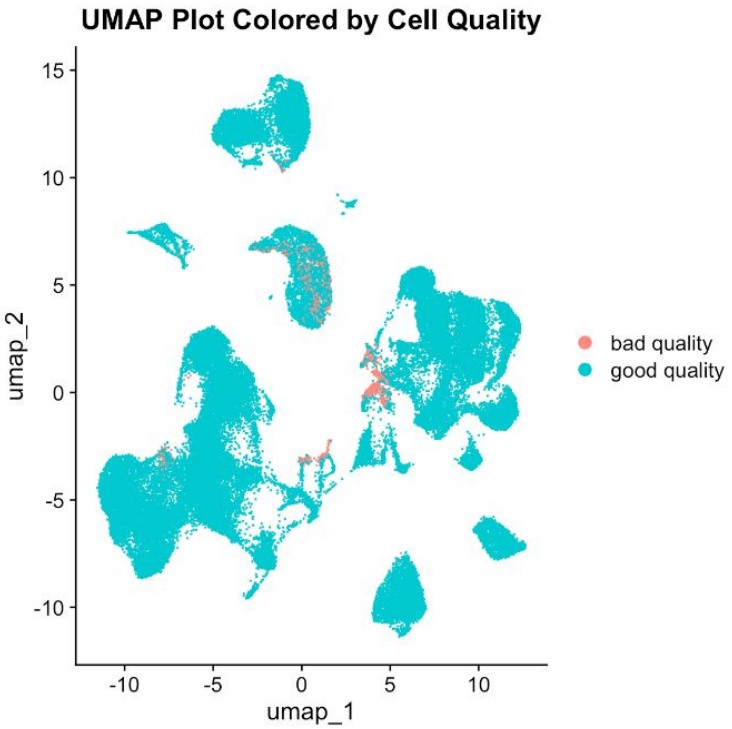
Cell-Level Metrics - Removing Low Quality Cells

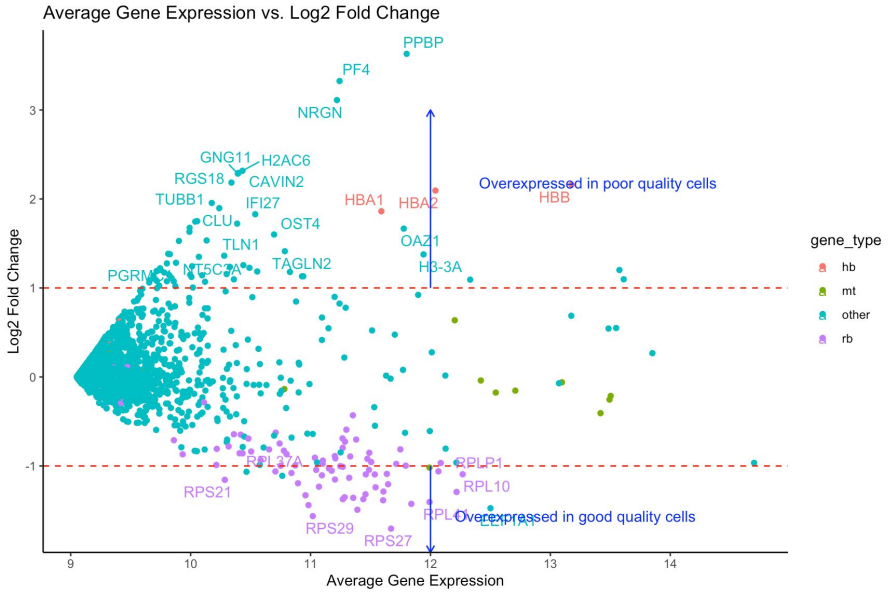
Metrics by themselves don't tell the whole picture:

- Low number of genes → Granulocytes or quiescent cells
- High mitochondrial % → Could be a muscle cell (cardiomyocyte)



Cell-Level Metrics - Diagnosing Cell Type Loss





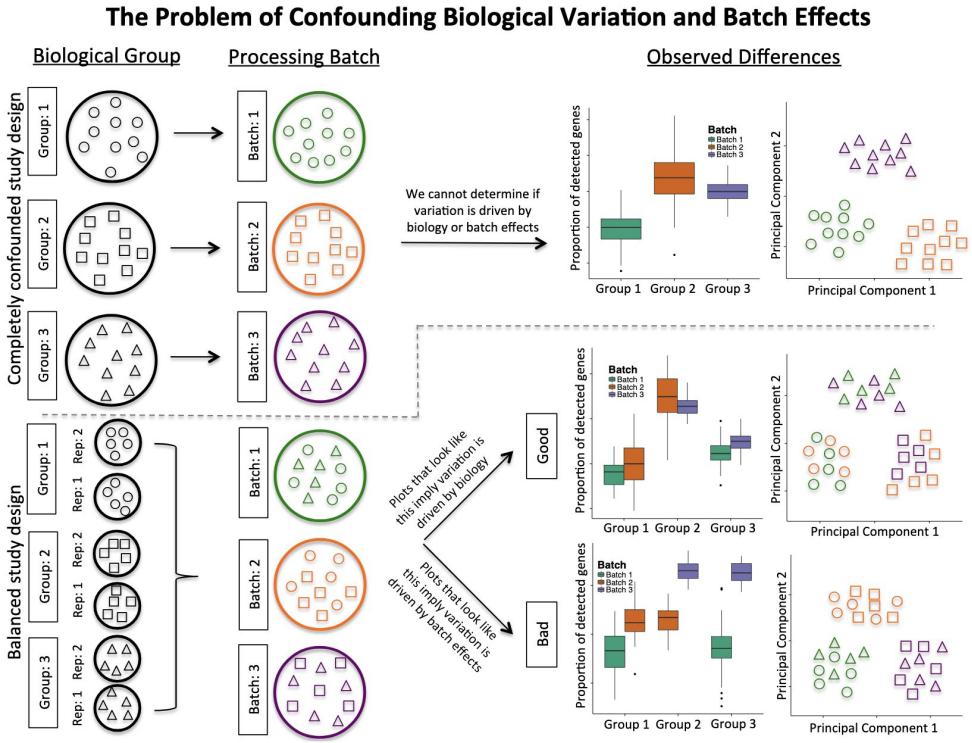
Batch Effect & Integration Assessment

What is batch?

"Batch effects in scRNA-seq experiments occur when cells from one biological group or condition are cultured, captured, and sequenced separate from cells in a second condition" - Hicks SC

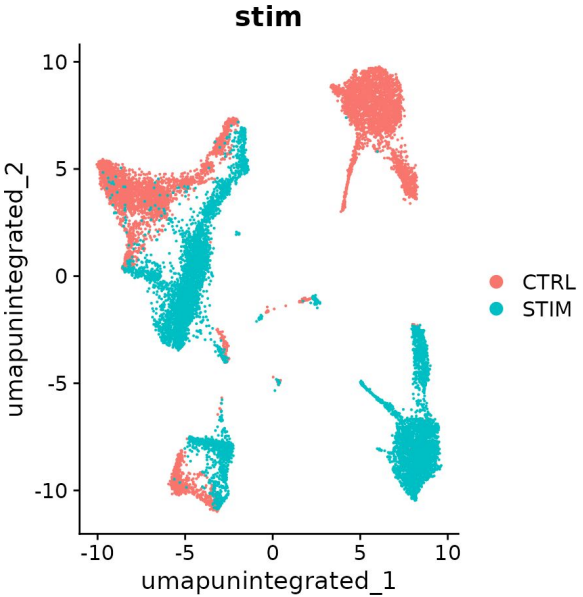
What is batch?

- Common in high throughput experiments
- Specially dangerous for single-cell data because we rely on unsupervised learning methods
- Confounds technical - biological

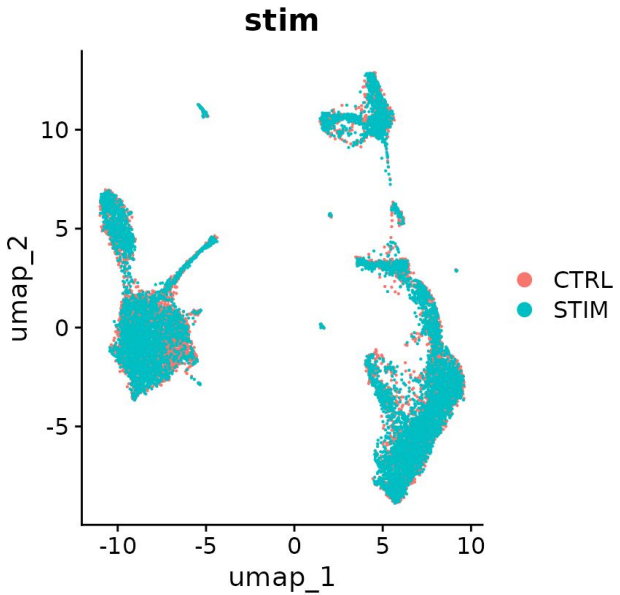


Batch Effect

Pre-integration



Post-integration



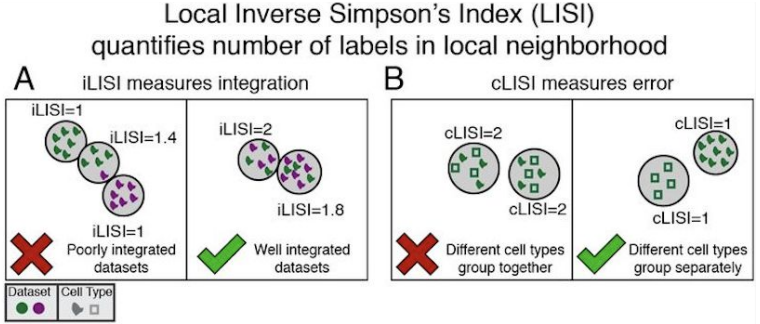
Integration Assessment

Metrics

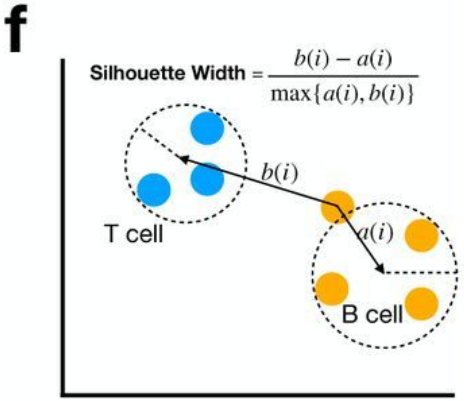
- iLISI and cLISI
- Silhouette Score
- PCA Regression

Covariate correction

- Known Batch Covariates
  - Date of collection
  - Site of collection
  - Sequencing flow cell
- Bio-conservation metrics
  - Predicted cell type label
  - Cell cycle score
  - Biological covariate of interest



Korunsky I, *et al.* Nat Methods 2019

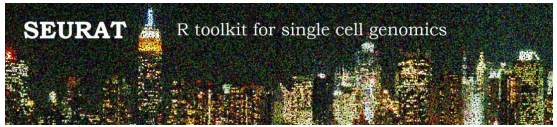


Zhao R, *et al.* bioRxiv 2024

Possible analyses

- Differential expression analyses between conditions
- Extract functional gene programs
- Map gene signatures
- Identify gene regulatory networks
- Cell-cell communication
- Compositional analyses
- Lineage tracing
- sc-eQTLs

Common Tools

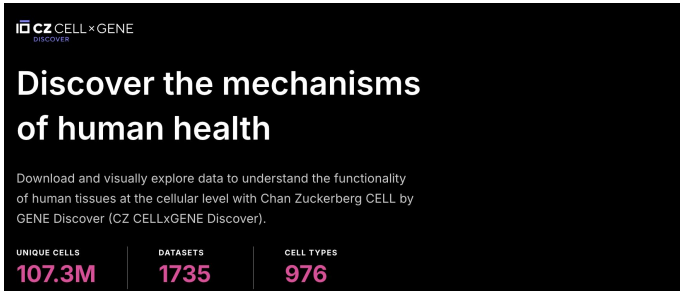


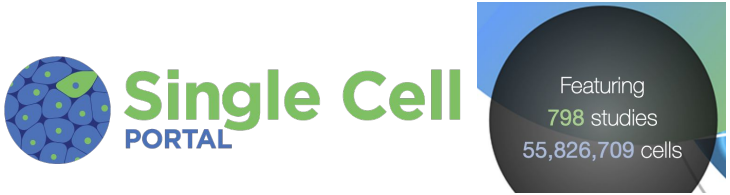
Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

scverse

scanpy

Data Repositories





Useful Resources

- Single-cell best practices
- Orchestrating Single-Cell Analysis with Bioconductor
- Current best practices in single-cell RNA-seq analysis: a tutorial
- Seurat - https://satijalab.org/seurat/
- Scanpy - https://scanpy.readthedocs.io/

Take home messages

- **Processing steps** from gene expression quantification to graph-based clustering, influences downstream biological insights.
  - **Feature selection** defines the biological space
  - **Dimensionality reduction** is a key step
- **Quality Control** is essential but should be carried out cautiously
  - QC metrics need to be checked together
  - **Diagnosing cell type loss** is essential to retain "*weird*" cell types
- **Batch effects** can be corrected with integration methods
  - Integration assessment should be done with **positive and negative controls**
  - **Overcorrecting** can remove biological signal

Acknowledgements

THE PEW CHARITABLE TRUSTS
MANTON FOUNDATION
RICHARD AND SUSAN Smith Family Foundation
aga American Gastroenterological Association
DAMON RUNYON CANCER RESEARCH FOUNDATION
Chan Zuckerberg Initiative
THE LEONA M. AND HARRY B. HELMSLEY CHARITABLE TRUST
NIH National Institute of Diabetes and Digestive and Kidney Diseases
NYSCF The New York Stem Cell Foundation
FA Food Allergy Science Initiative
abbvie
THE G. HAROLD AND LEILA Y. MATHERS FOUNDATION
NIH National Heart, Lung, and Blood Institute

ORDOVAS-MONTANES LAB
INFLAMMATION · MEMORY · IMMUNITY

CELL DISCOVERY NETWORK