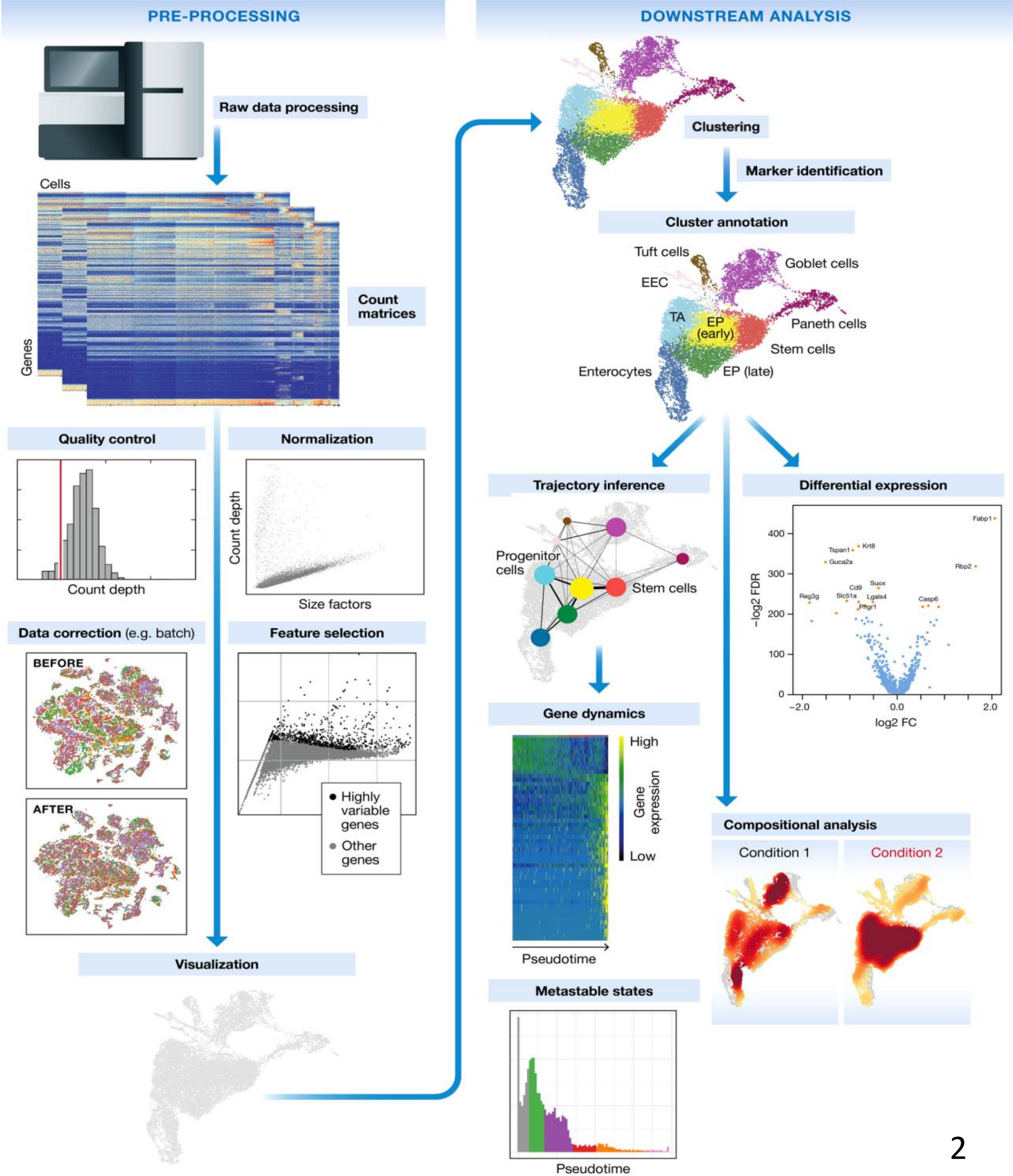
Biomedical Data Science:
Single Cell Workshop

Donglu Bai
TA Lecture for CBB 752
Biomedical Data Science: Mining and Modeling
Spring, 2023
Yale University

Single-cell RNA-seq workflow

- **Generation of the count matrix**
- **Quality control of the raw counts** – filter out poor quality cells
- **Clustering of filtered counts:** (cell types = different clusters)
- **Marker identification and cluster annotation:** identify gene markers for each cluster

Biological replicates are needed!

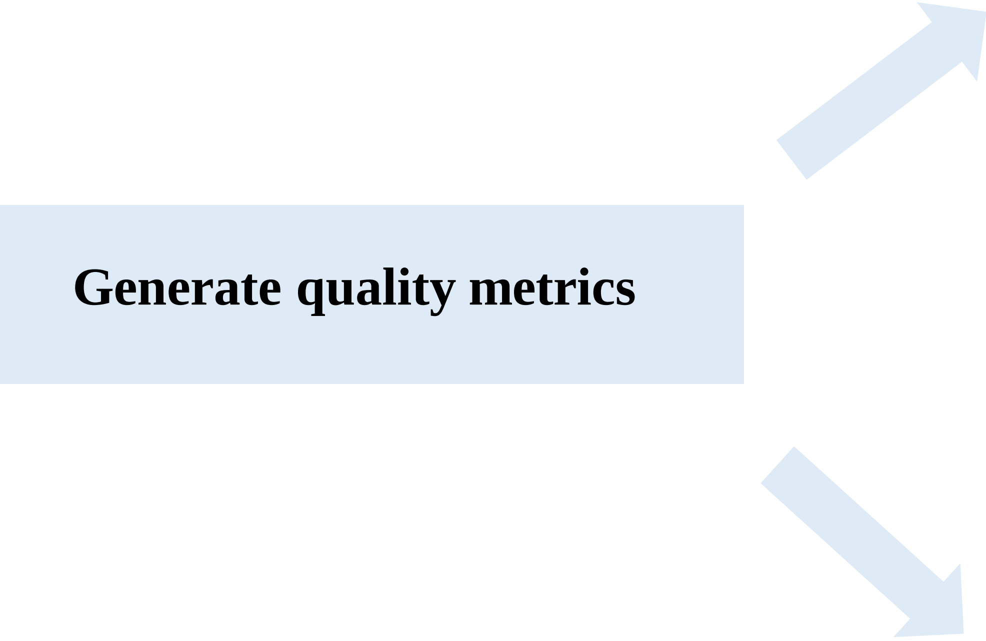


Luecken, MD and Theis, FJ. *Current best practices in single-cell RNA-seq analysis: a tutorial, Mol Syst Biol 2019 (doi: https://doi.org/10.15252/msb.20188746)*

Quality control set up

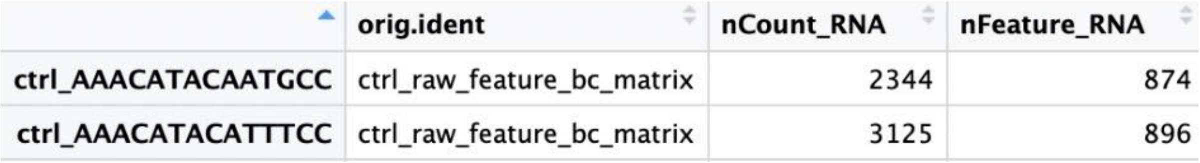
Goal:

- To filter the data to only include true cells that are of high quality
- To identify any failed samples



Generate quality metrics

Example dataset: comprised of pooled peripheral blood mononuclear cells from eight lupus patients, split into control and interferon beta-treated (stimulated) conditions



	orig.ident	nCount_RNA	nFeature_RNA
ctrl_AAACATACAATGCC	ctrl_raw_feature_bc_matrix	2344	874
ctrl_AAACATACATTTCC	ctrl_raw_feature_bc_matrix	3125	896

Sample id # of UMIs/cell # of genes/cell

Additional metrics:

1. Number of genes detected per UMI (novelty score)
2. Mitochondrial ratio

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Quality control set up

- **Novelty score**

  # of detected genes/UMI = # of genes/cell / # of UMIs per cell

- **Mitochondrial Ratio**

- PercentageFeatureSet() searches for gene identifiers that begin with MT-

- **Questions to consider:**

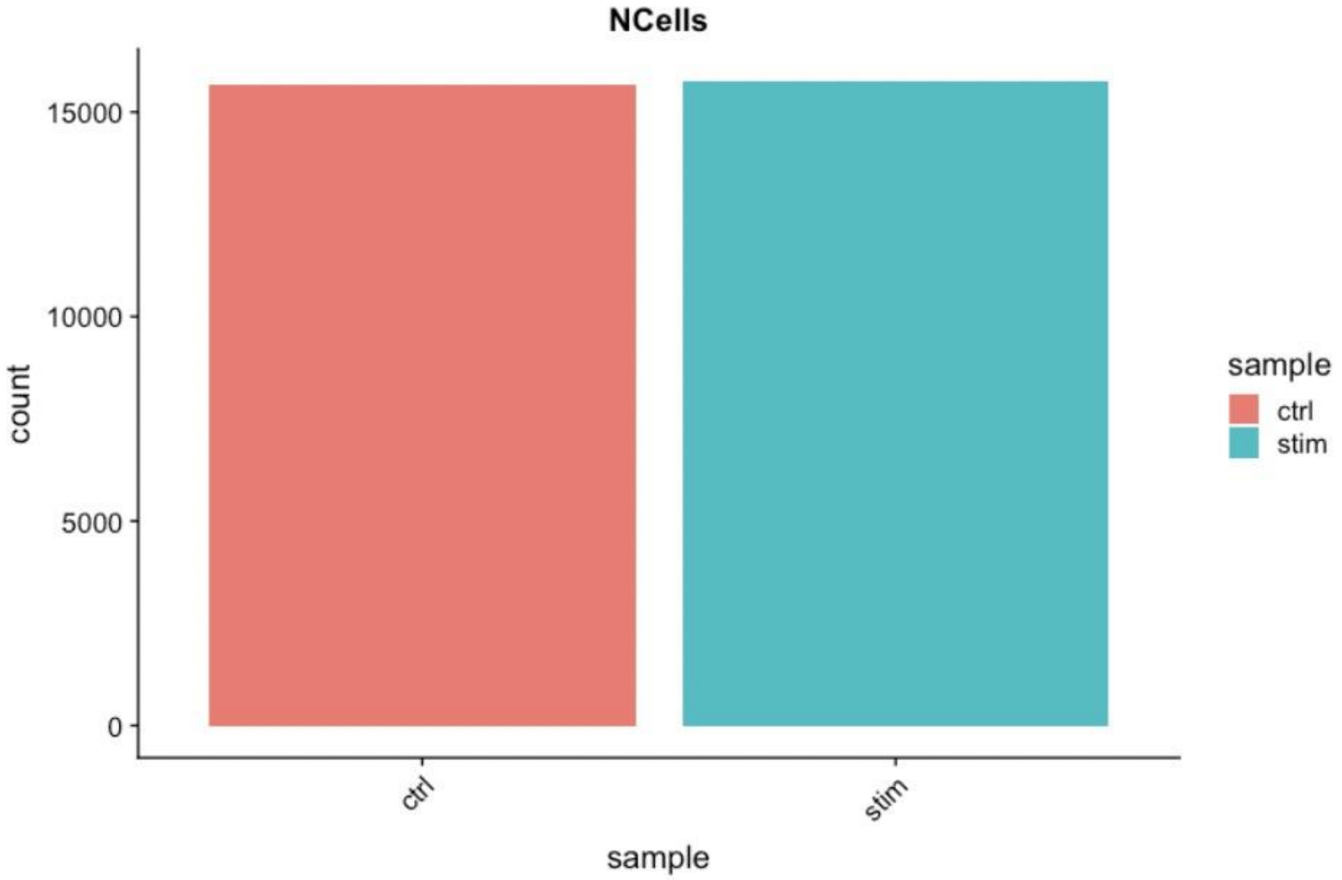
- Why aren’t we checking for doublets?
- What does high mitochondrial gene expression indicate?

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Assess the quality metrics

Cell counts
-For this experiment, between 12,000-13,000 cells are expected

-We see over 15,000 cells per sample, which is quite a bit more than the 12-13,000 expected. We have some junk 'cells' present.

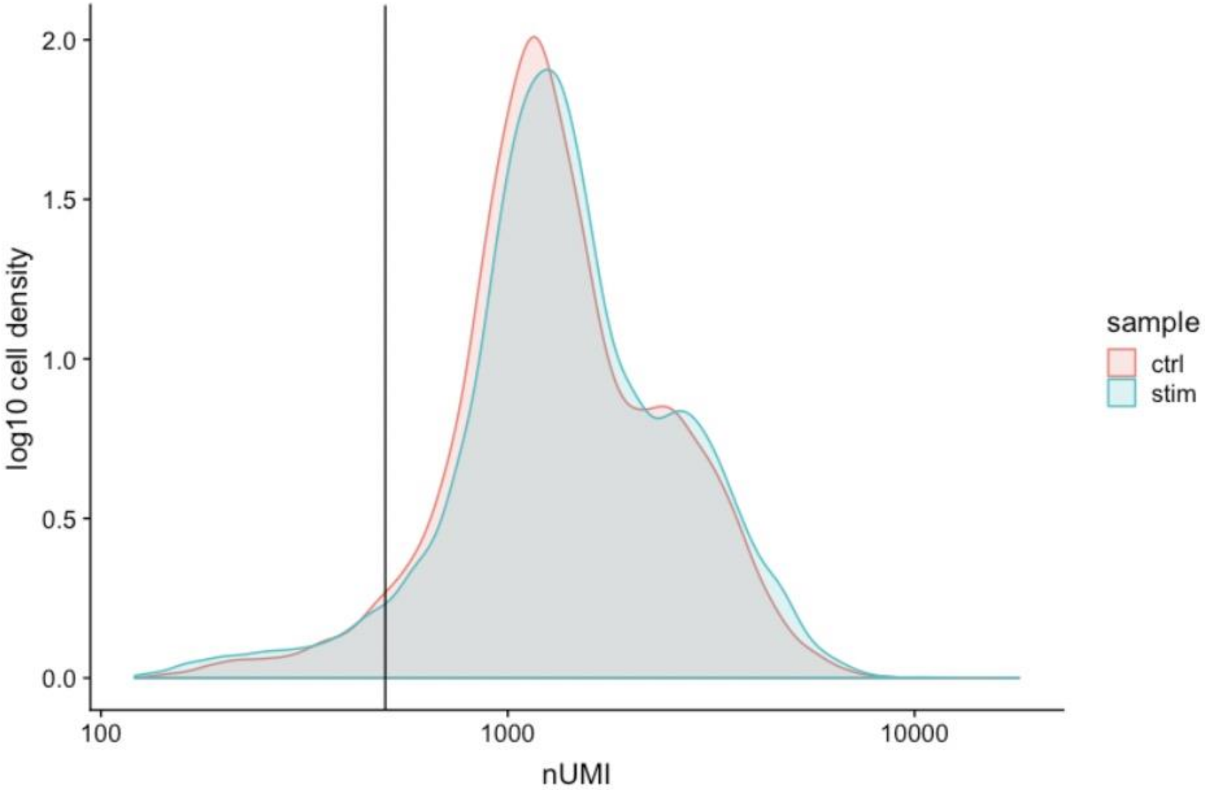


https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Assess the quality metrics

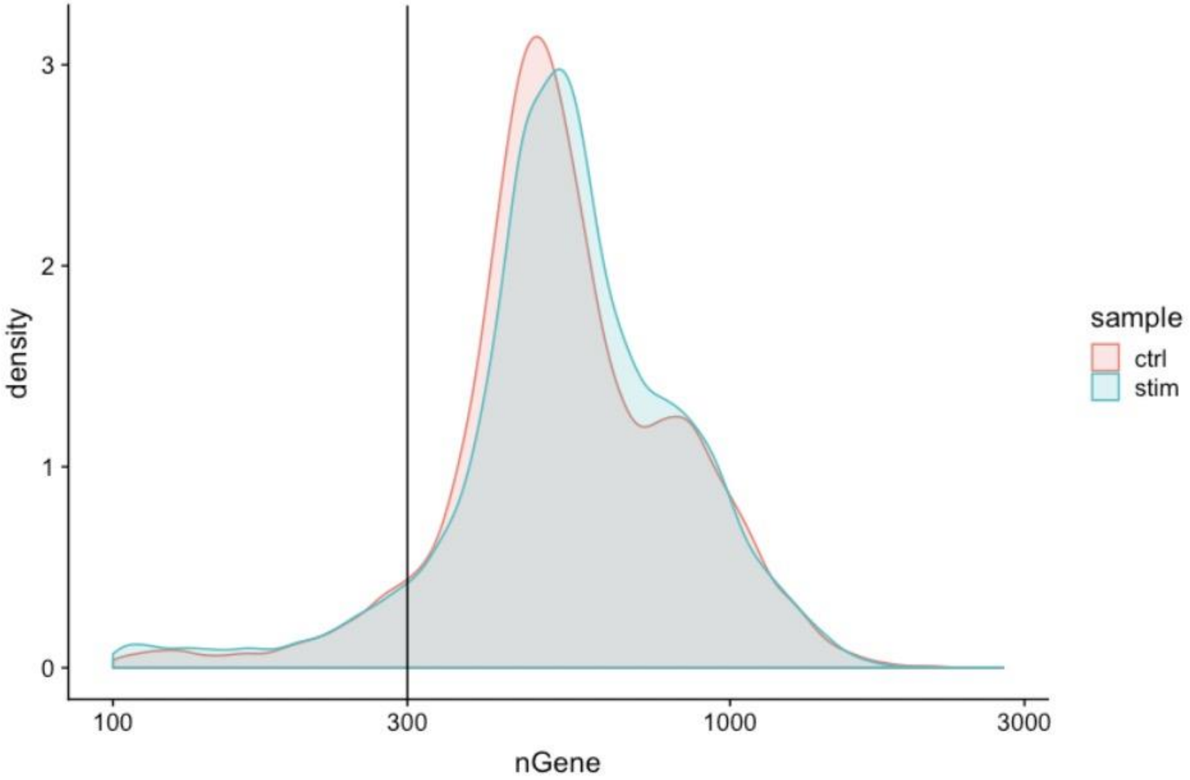
UMI counts (transcripts) per cell
-The UMI counts per cell should generally be above 500, that is the low end of what we expect.

-Most cells in both samples have 1000 UMIs or greater



Assess the quality metrics

Genes detected per cell
-Similar expectations for gene detection as for UMI detection

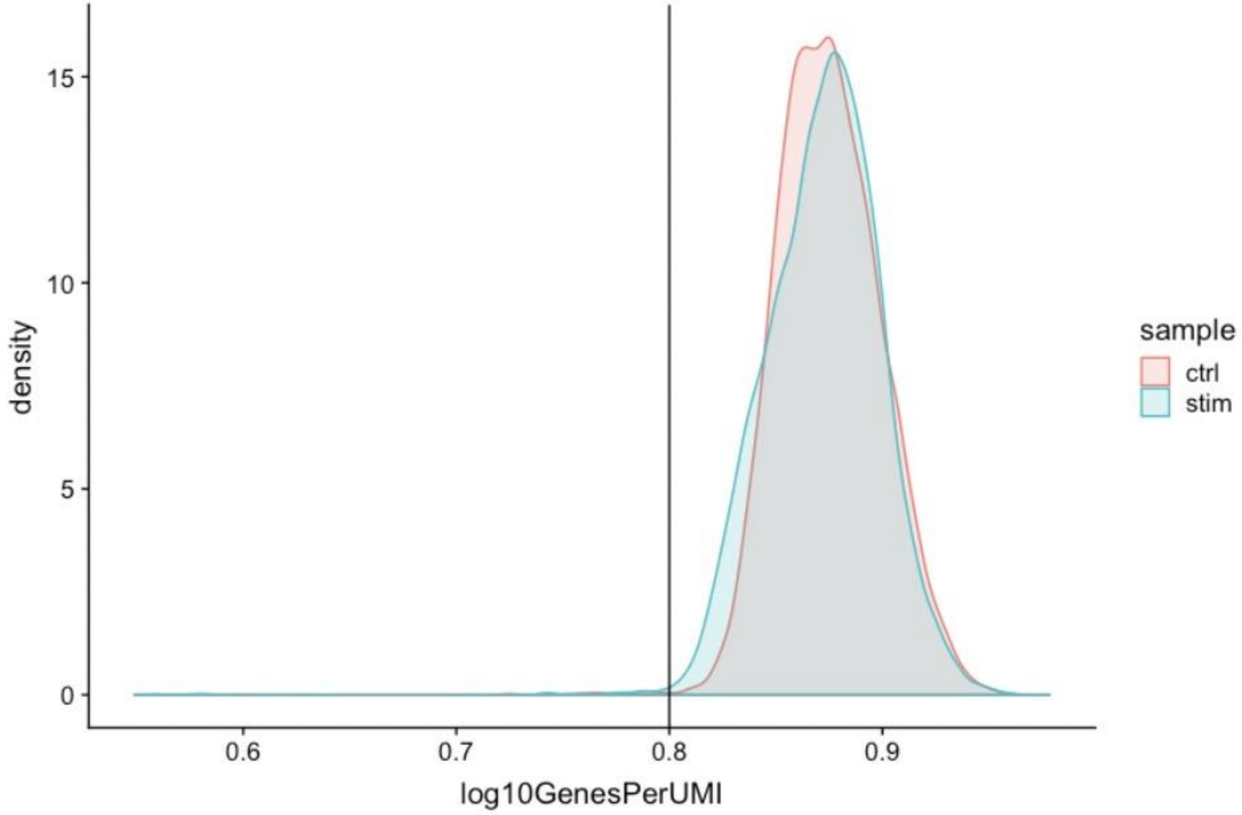


https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Assess the quality metrics

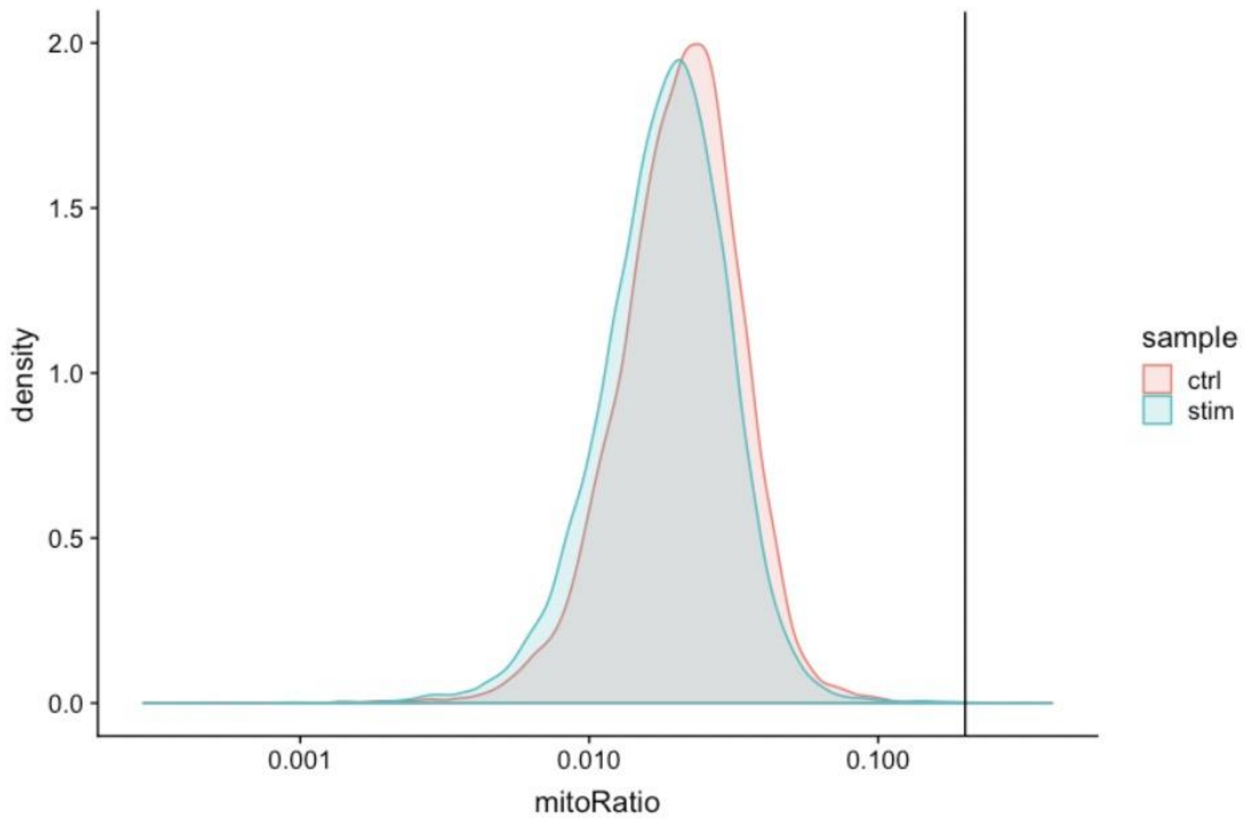
Complexity
-Novelty score for assessing how complex the RNA species are

-Generally, we expect the novelty score to be above 0.8 for good quality cells



Assess the quality metrics

Mitochondrial counts ratio
-define poor quality samples for mitochondrial counts as cells which surpass the 0.2 mitochondrial ratio mark

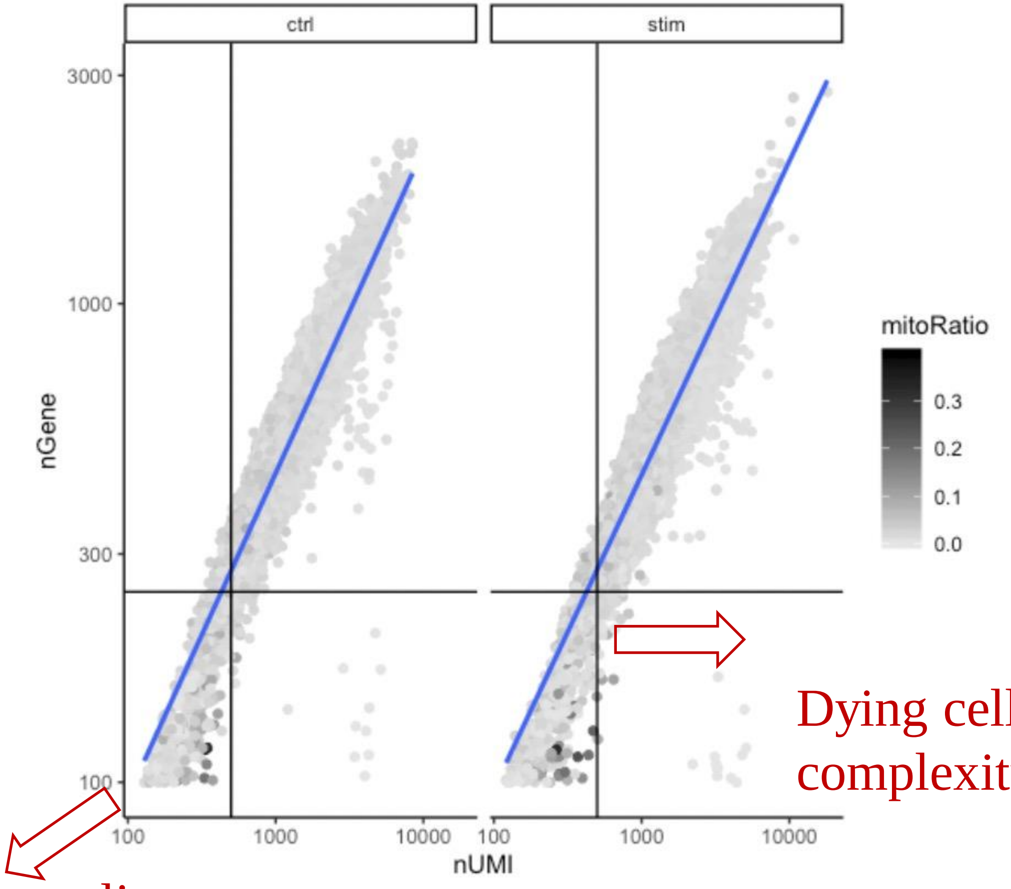


https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Assess the quality metrics

Joint filtering effects

-A general rule of thumb is to set thresholds for individual metrics as permissive as possible and consider the joint effects of these metrics



Dying cells/low complexity cell types

Cells with poor quality

High mitochondrial read fractions

Filtering

Cell-level filtering

-# of UMI >500

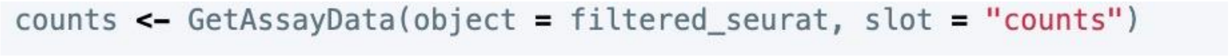
-# of Gene >250

-log10GenesPerUMI >0.8

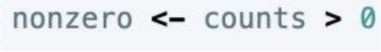
-mitoRatio <0.2

Gene-level filtering

-remove genes with zero counts from our data



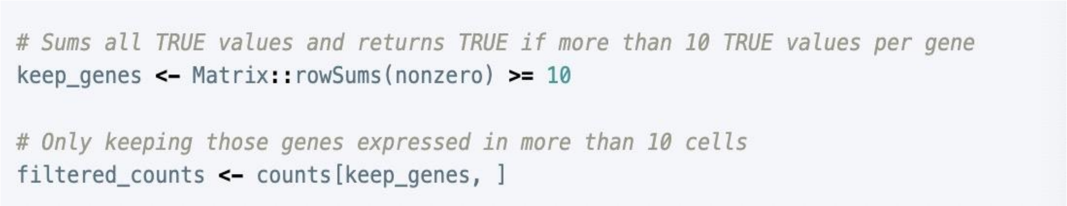
```
counts <- GetAssayData(object = filtered_seurat, slot = "counts")
```



```
nonzero <- counts > 0
```

-output a logical matrix for each gene on if there are more than zero counts per cell

-keep only genes which are expressed in 10 or more cells; remove zero count



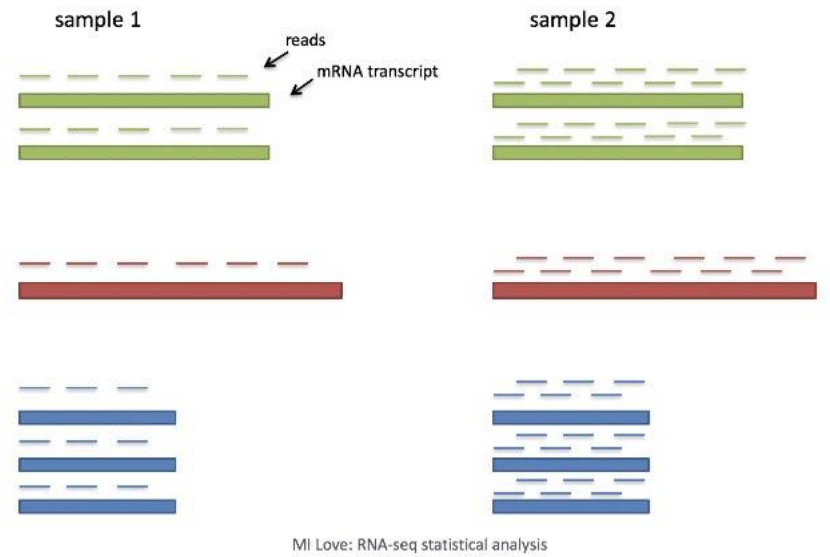
```
# Sums all TRUE values and returns TRUE if more than 10 TRUE values per gene
keep_genes <- Matrix::rowSums(nonzero) >= 10

# Only keeping those genes expressed in more than 10 cells
filtered_counts <- counts[keep_genes, ]
```

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

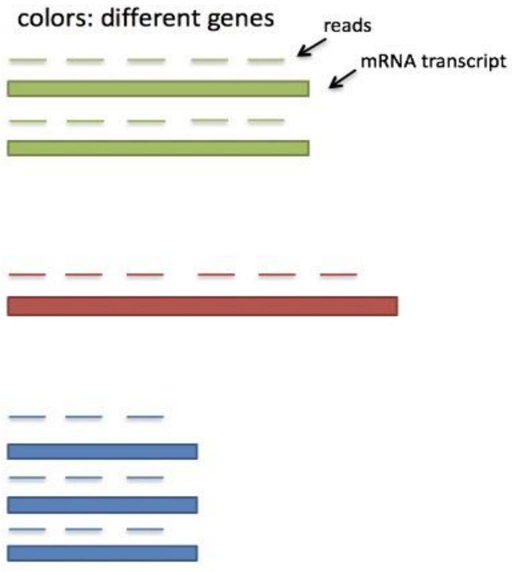
Normalization and regressing out unwanted variation

Main factors to consider

-Sequencing depth



-Gene length



Normalization and regressing out unwanted variation

Methods for scRNA-seq normalization

-Scaling

multiply each UMI count by a cell specific factor to get all cells to have the same UMI counts

(not interested in comparing absolute counts between cells)

-Transformation

Pearson residuals for transformation

(construct GLM for each gene with UMI as the response and sequencing depth as the explanatory variable to obtain residuals with normalized data values)

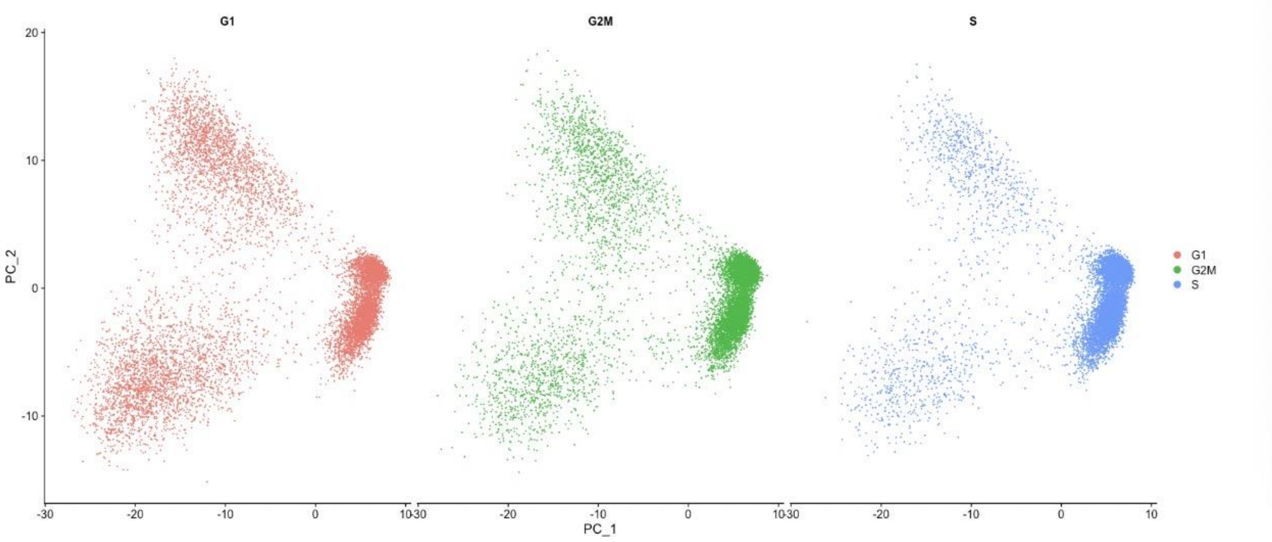
https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Normalization and regressing out unwanted variation

Explore sources of unwanted variation

-One most common biological data correction (uninteresting variation) is the effects of the cell cycle on the transcriptome
-Explore effects of cell cycle (CellCycleScoring() calculates cell phase scores)
-Determine if cell cycle is a major source of variation in our dataset using PCA

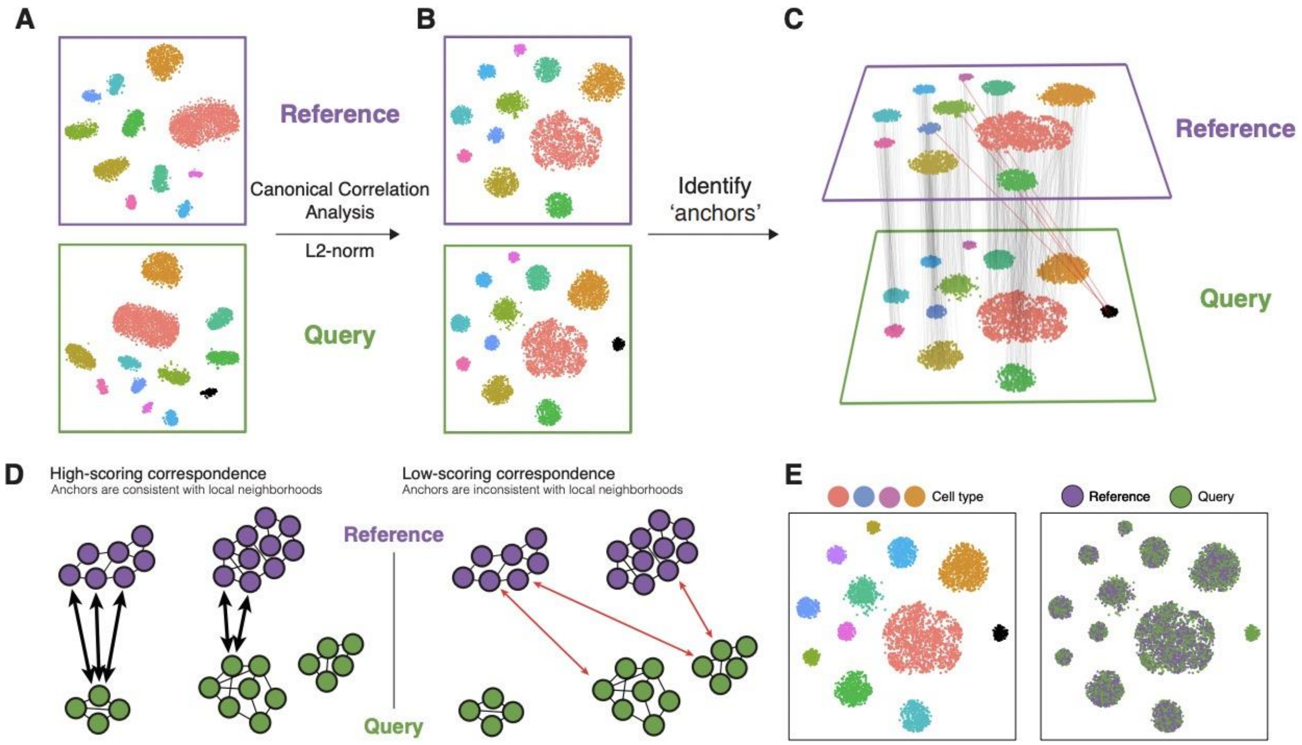
Not regress out the variation due to cell cycle



https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Integration

-Integrate the cells across conditions to ensure that cells of the same cell type cluster together
(identify cell types that are present in all conditions for interpretable downstream analysis)

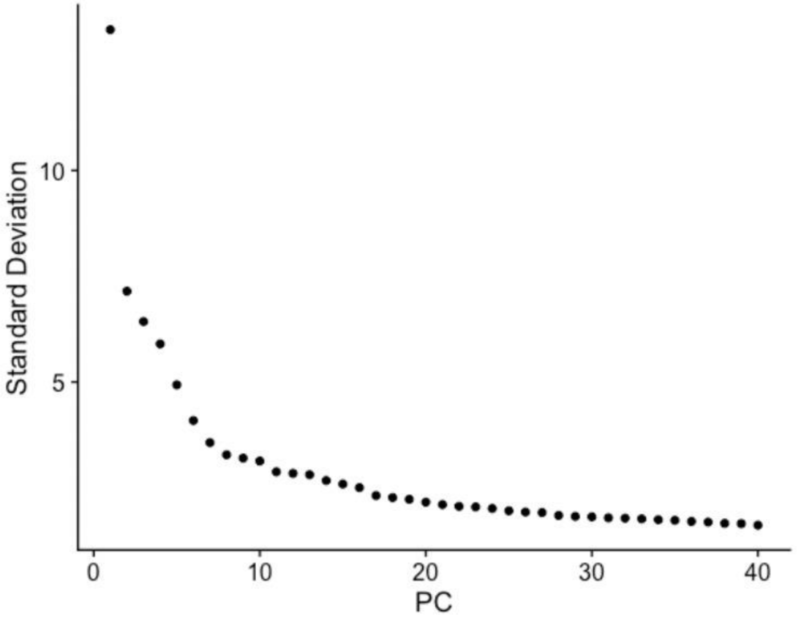


A
B
C
Reference
Canonical Correlation Analysis
L2-norm
Query
Identify 'anchors'
Reference
Query
D
High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query
E
Cell type
Reference
Query

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Clustering

Identify significant PCs



Clustering methods

-K-means clustering

Measure of similarity: Euclidean distance

Quality function: Within cluster distance

-Graph-based clustering

Memory effectiveness: Current methods aim to build sparse graphs

Curse of dimensionality: All data become sparse in high-dimensional space

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Spectral clustering

-Pre-processing

build Laplacian matrix L of the graph G

-Decomposition

Find eigenvalues and eigenvectors

Map vertices to corresponding components of second smallest eigenvectors

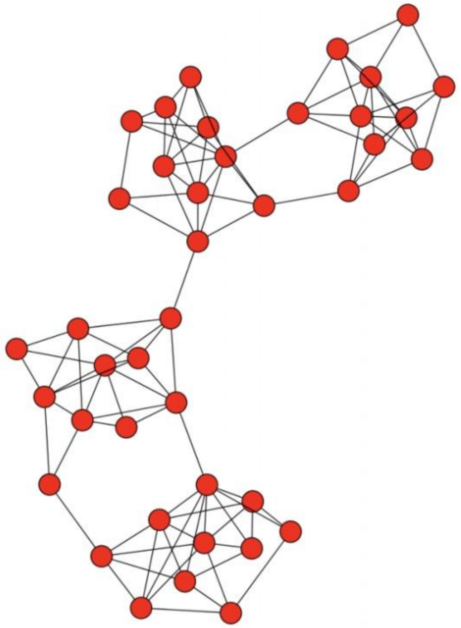
-Grouping

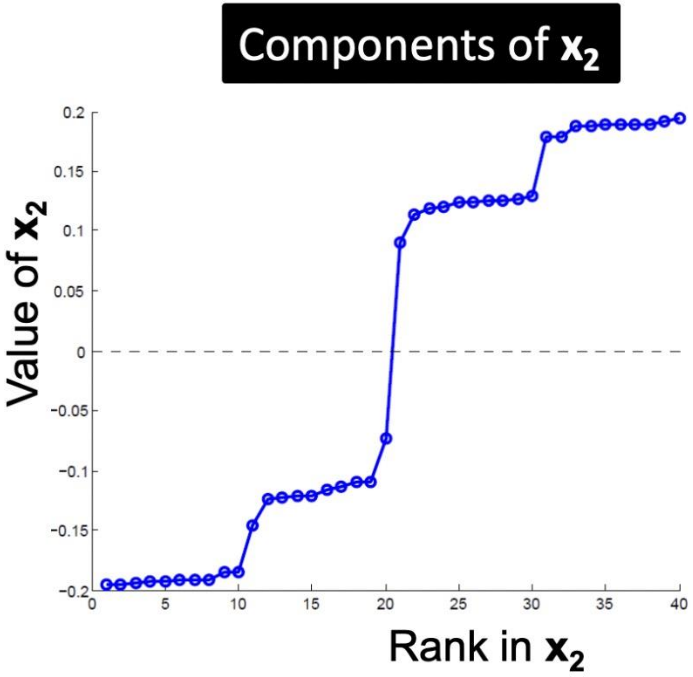
Identify clusters by splitting the sorted vector in two (split at 0 or median value)

Partition into k clusters

-Recursive bi-partitioning

-Cluster multiple eigenvectors





Avoids the curse of dimensionality by projecting data into lower-dimensional space

https://wandb.ai/syllogismos/machine-learning-with-graphs/reports/5-Spectral-Clustering--VmlldzozNzcwMDU

Louvain clustering

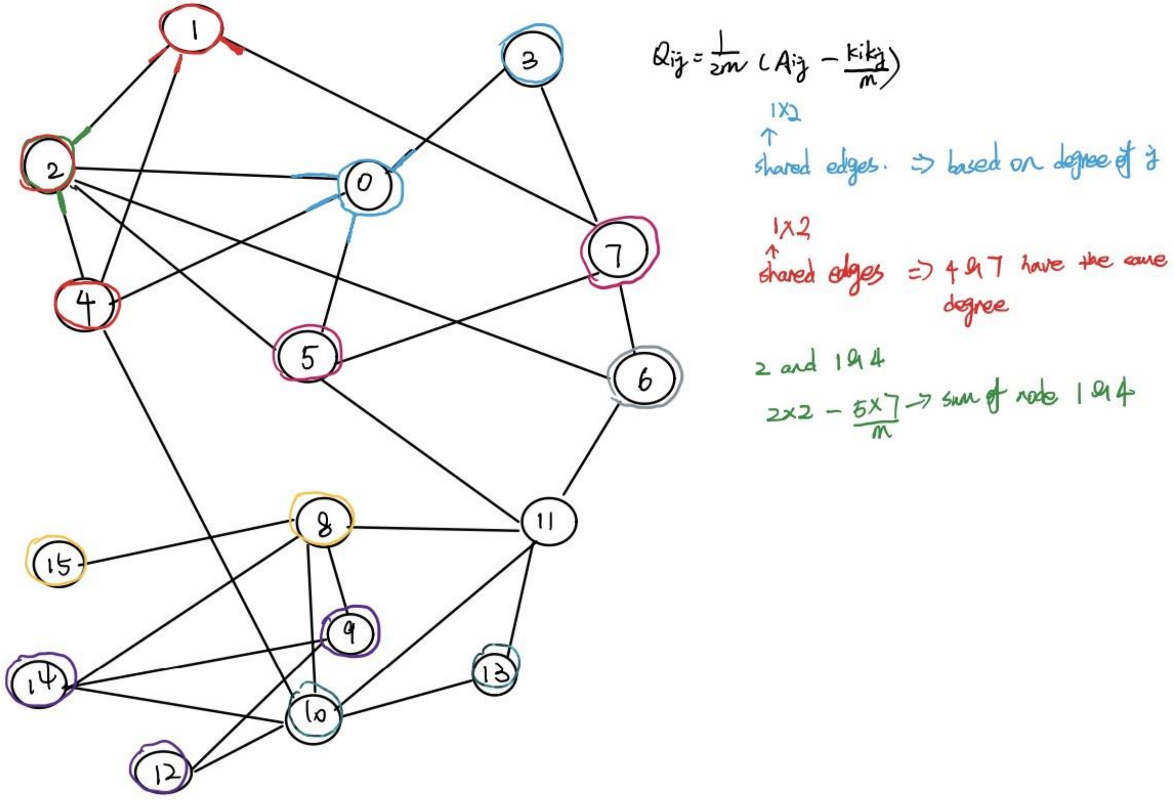
-The optimization of modularity

First small communities are found by optimizing modularity locally on all nodes; then each small community is grouped into one node and the first step is repeated

$$Q = \frac{1}{2m} \sum_{ij} \left[A_{ij} - \frac{k_i k_j}{2m} \right]$$

Fast

No need to specify k



$Q_{ij} = \frac{1}{2m}(A_{ij} - \frac{k_i k_j}{m})$

1×2

↑

shared edges. ⇒ based on degree of j

1×2

↑

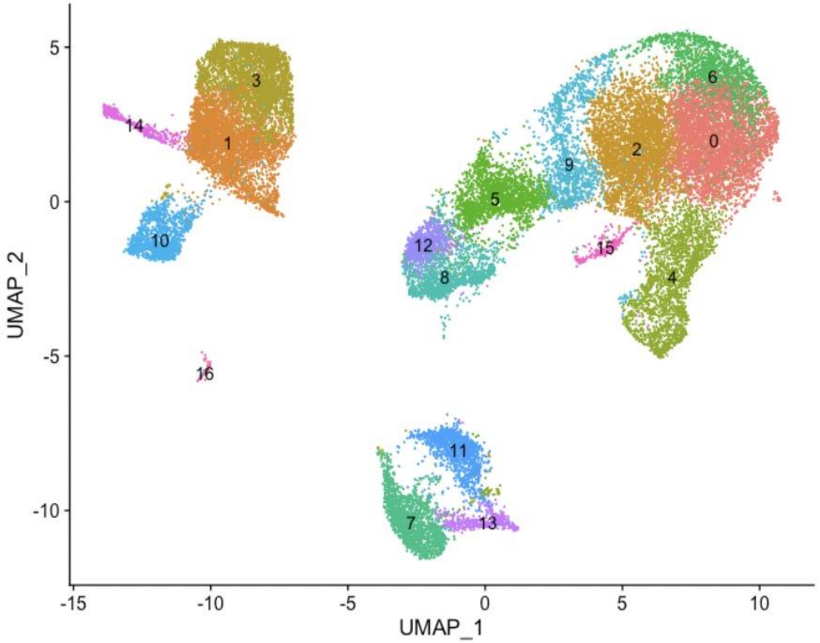
shared edges ⇒ 4 & 7 have the same degree

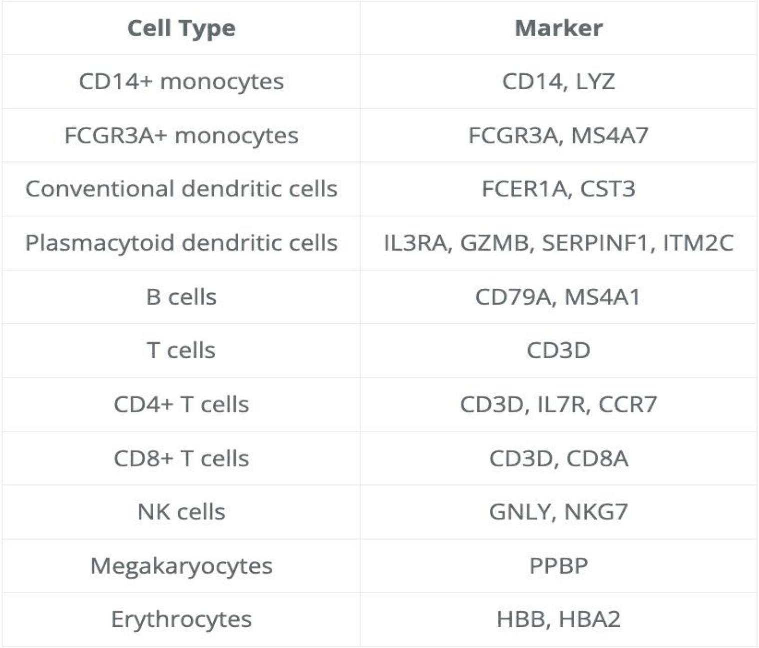
2 and 1 & 4

$2\times2 - \frac{5\times7}{m}$ → sum of node 1 & 4

Clustering quality control

Exploring known cell type markers





Cell Type	Marker
CD14+ monocytes	CD14, LYZ
FCGR3A+ monocytes	FCGR3A, MS4A7
Conventional dendritic cells	FCER1A, CST3
Plasmacytoid dendritic cells	IL3RA, GZMB, SERPINF1, ITM2C
B cells	CD79A, MS4A1
T cells	CD3D
CD4+ T cells	CD3D, IL7R, CCR7
CD8+ T cells	CD3D, CD8A
NK cells	GNLY, NKG7
Megakaryocytes	PPBP
Erythrocytes	HBB, HBA2

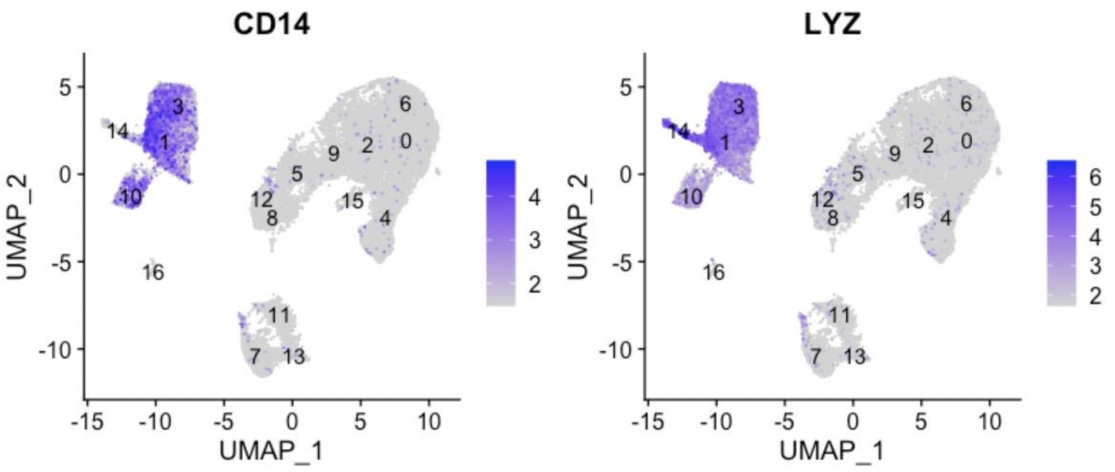
Explore cell type identities by looking for known markers

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

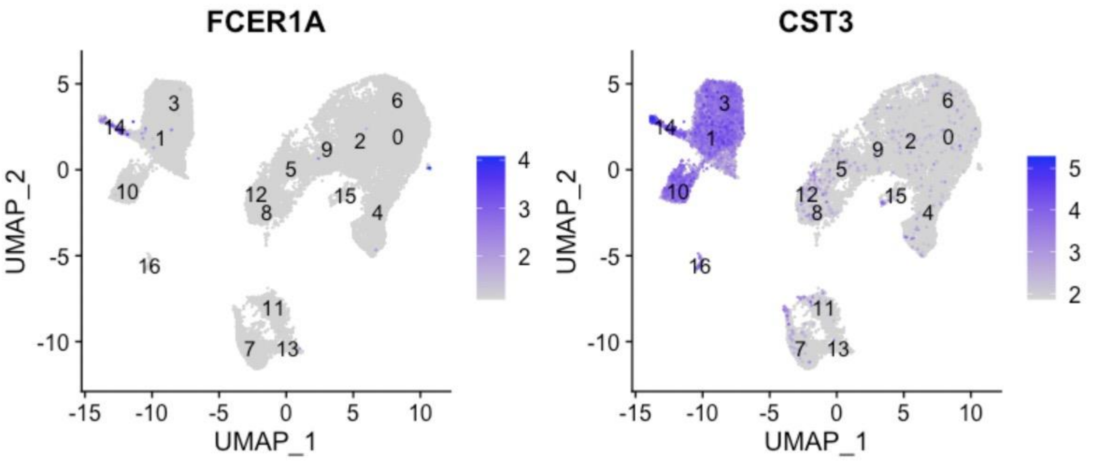
Clustering

Exploring known cell type markers

-The FeaturePlot() from Seurat makes it easy to visualize a handful of genes
-The combined expression of our chosen positive or negative markers should give us an idea of which cluster corresponds to which cell type



CD14+ monocytes appear to correspond to clusters 1 and 3

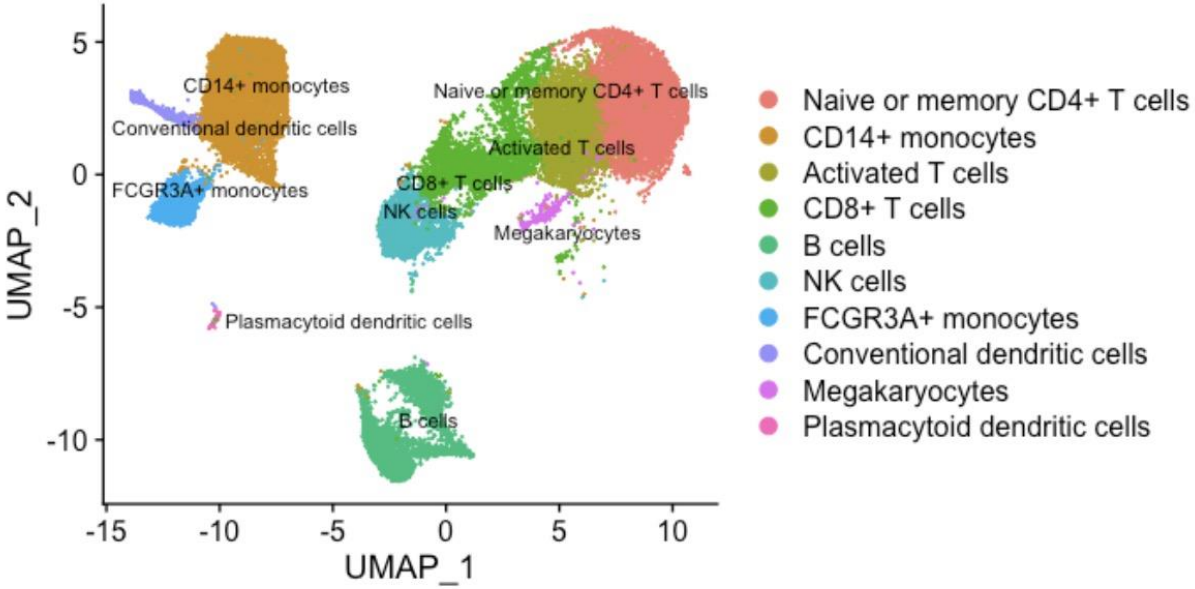


Conventional dendritic cell markers identify cluster 14

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Clustering

Exploring known cell type markers



-Experimentally validate intriguing markers

-Explore a subset of cell types to discover subclusters of cells

-Perform differential expression analysis between control and stimulations

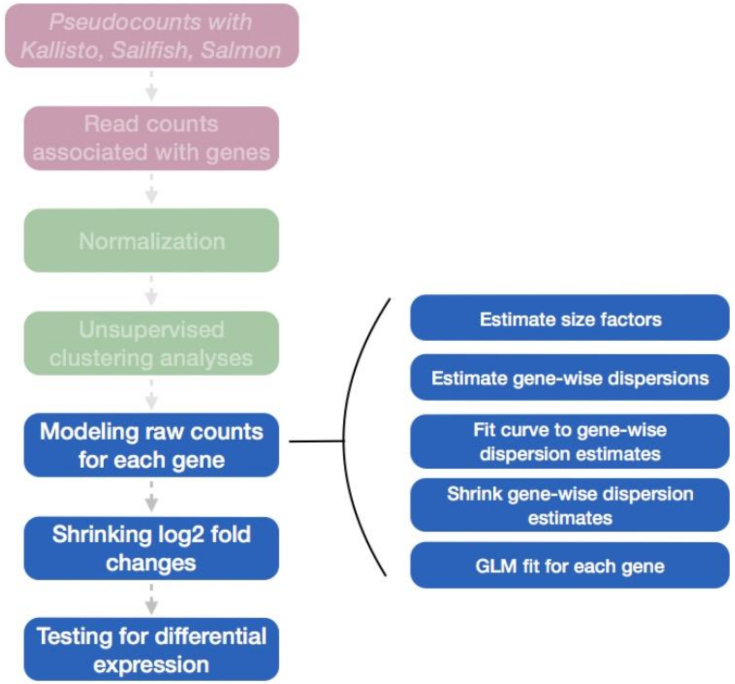
-Trajectory analysis or lineage tracing

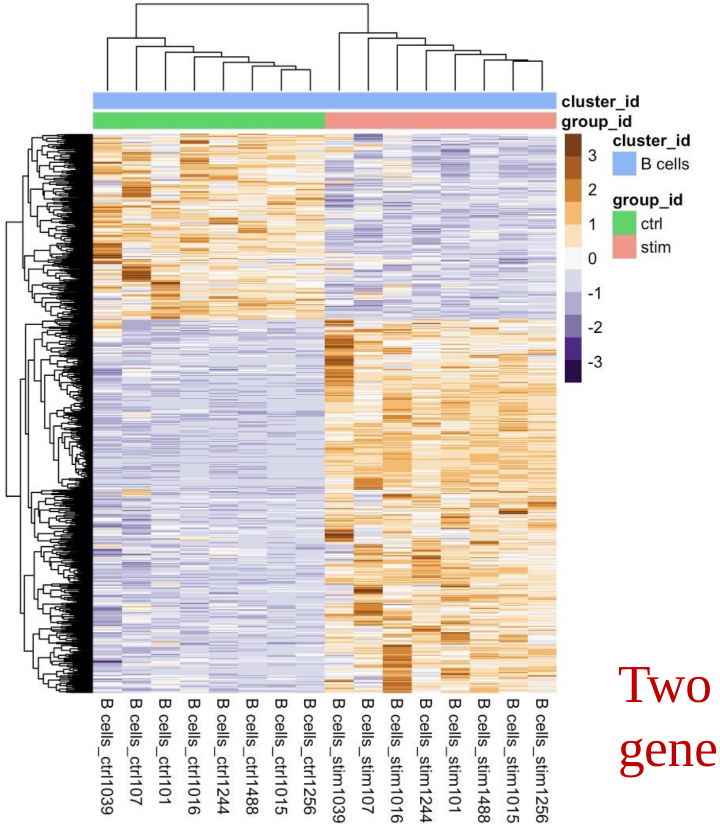
https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

DE analysis

DESeq2 for pseudobulk DE analysis

-normalizes the count data to account for differences in library sizes and RNA composition between samples


Pseudocounts with Kallisto, Sailfish, Salmon
Read counts associated with genes
Normalization
Unsupervised clustering analyses
Modeling raw counts for each gene
Shrinking log2 fold changes
Testing for differential expression
Estimate size factors
Estimate gene-wise dispersions
Fit curve to gene-wise dispersion estimates
Shrink gene-wise dispersion estimates
GLM fit for each gene



Two clear modules of genes emerge

https://hbctraining.github.io/scRNA-seq_online/lessons/04_SC_quality_control.html

Any question?

Thanks for attention!