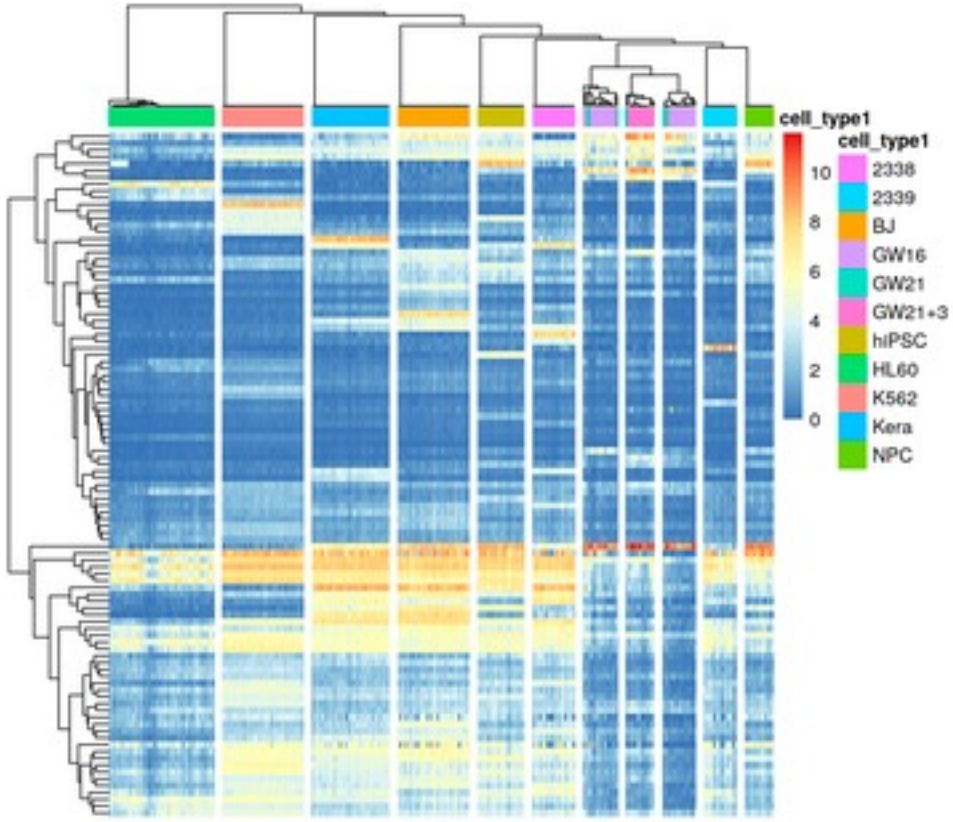
Quality control for scRNA-Seq data

Orr Ashenberg, Jacques Serizay, Fabricio Almeida-Silva

November 2025

Physalia

Courses

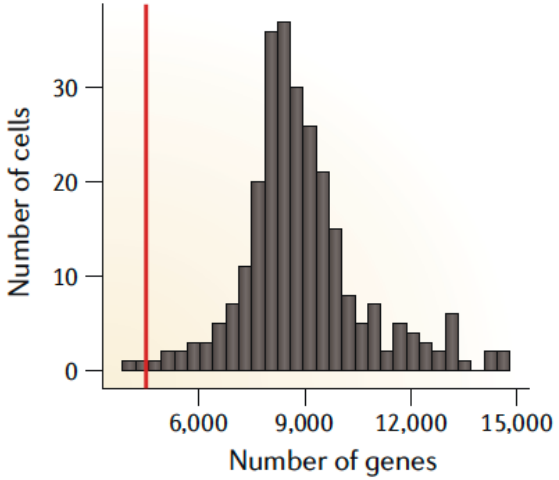


Outline: Quality control of scRNA-Seq data

- Quality control and normalization starting from gene count matrices.
- Interacting with Seurat objects.
- Next step will be dimensionality reduction, clustering, and visualization

Determining cell type, state, and function

Quality control



Normalization

Feature selection

Dimensional reduction

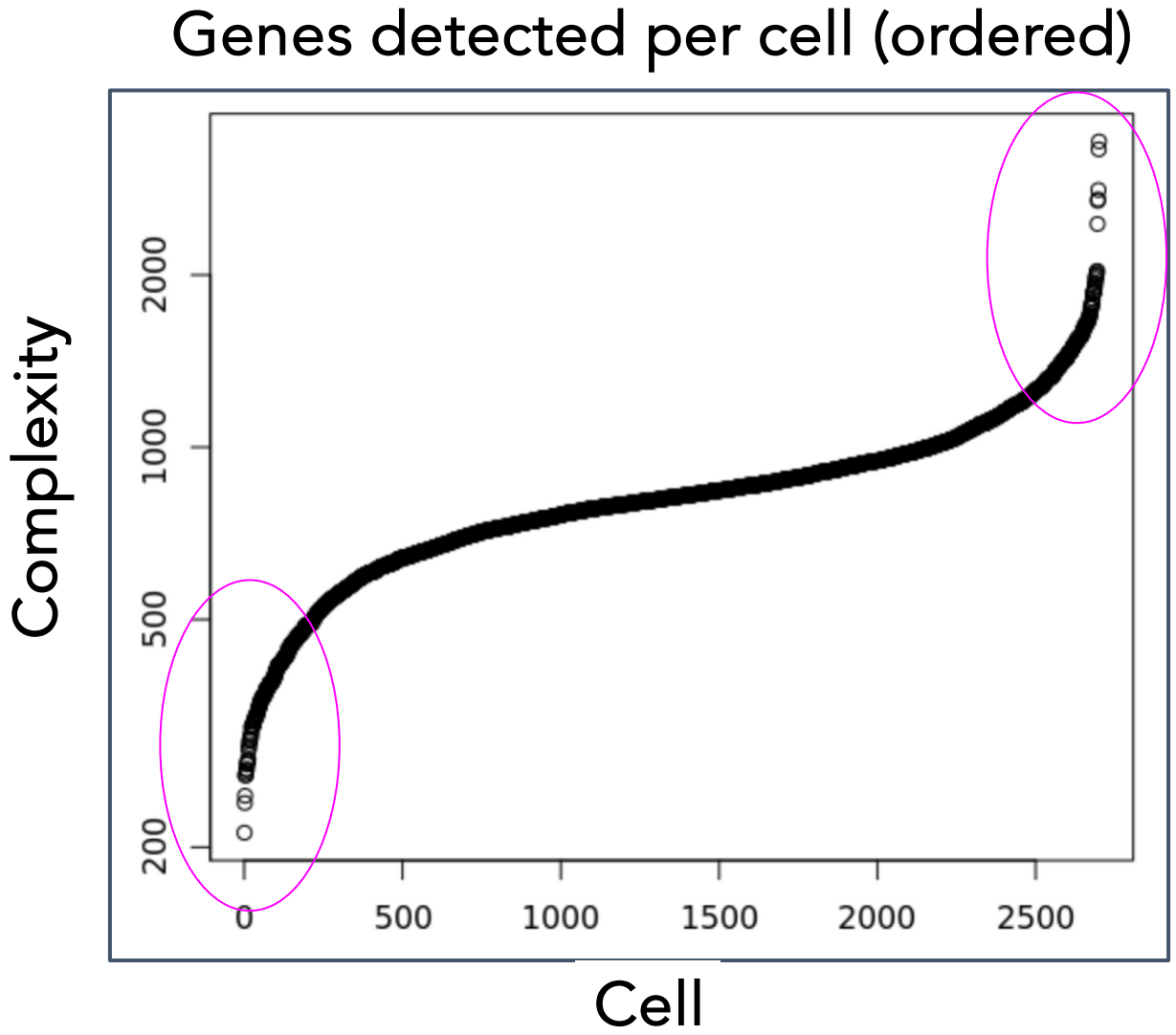
Cell-cell distances

Unsupervised clustering

Kiselev C. at *al.* (2019) *Nature Reviews Genetics*. 20: 273-282.

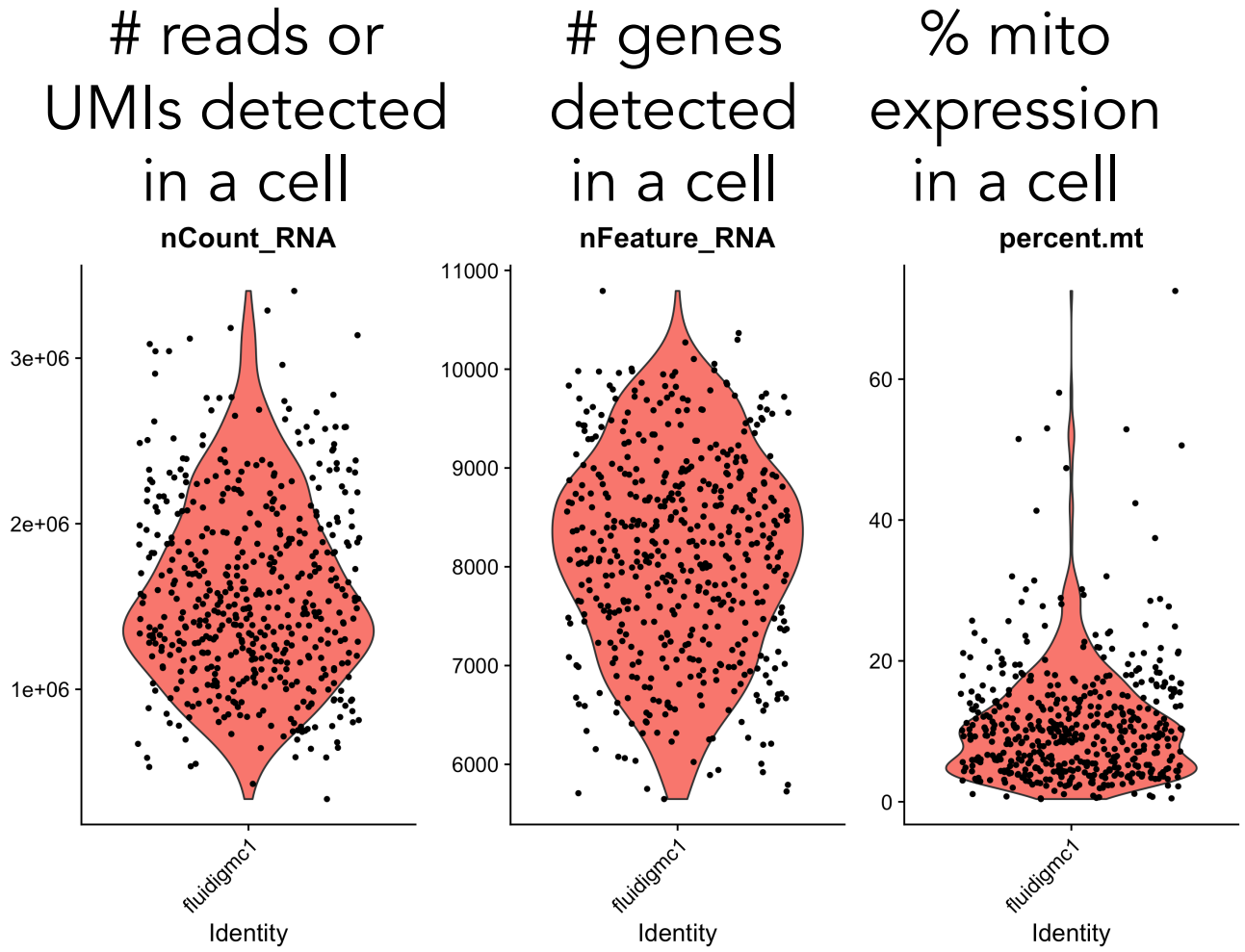
There are many quality control filters for genes and cells

Complexity =
Number of genes
detected in a cell



There are many quality control filters for genes and cells

- We filter cells based on technical or biological parameters.

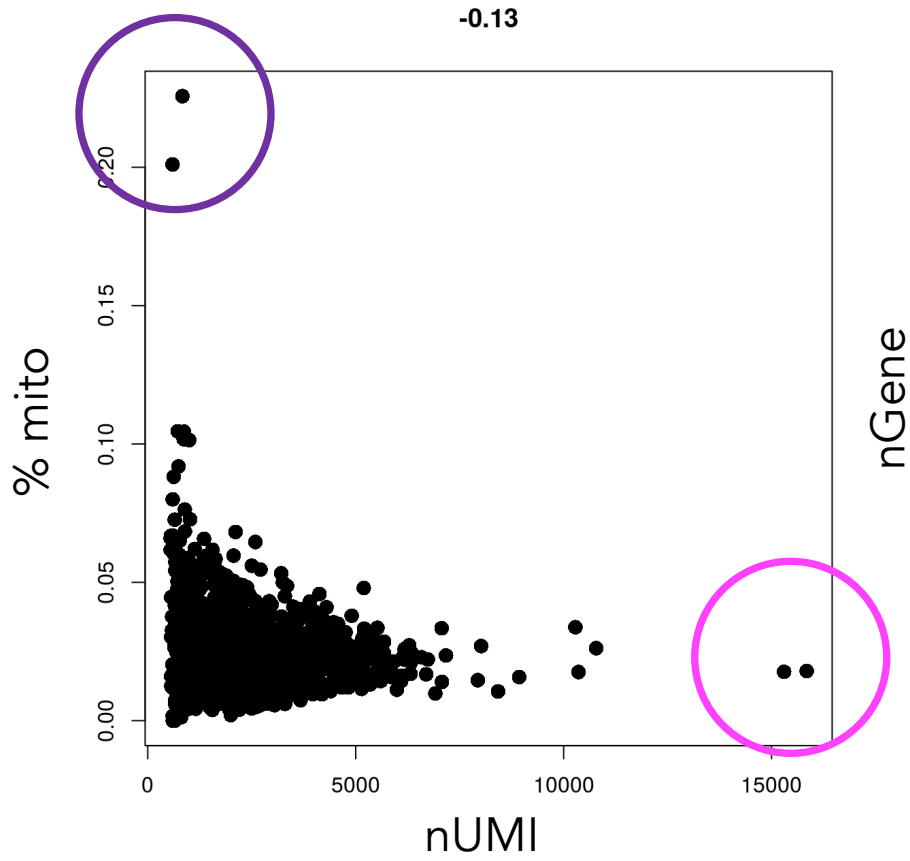


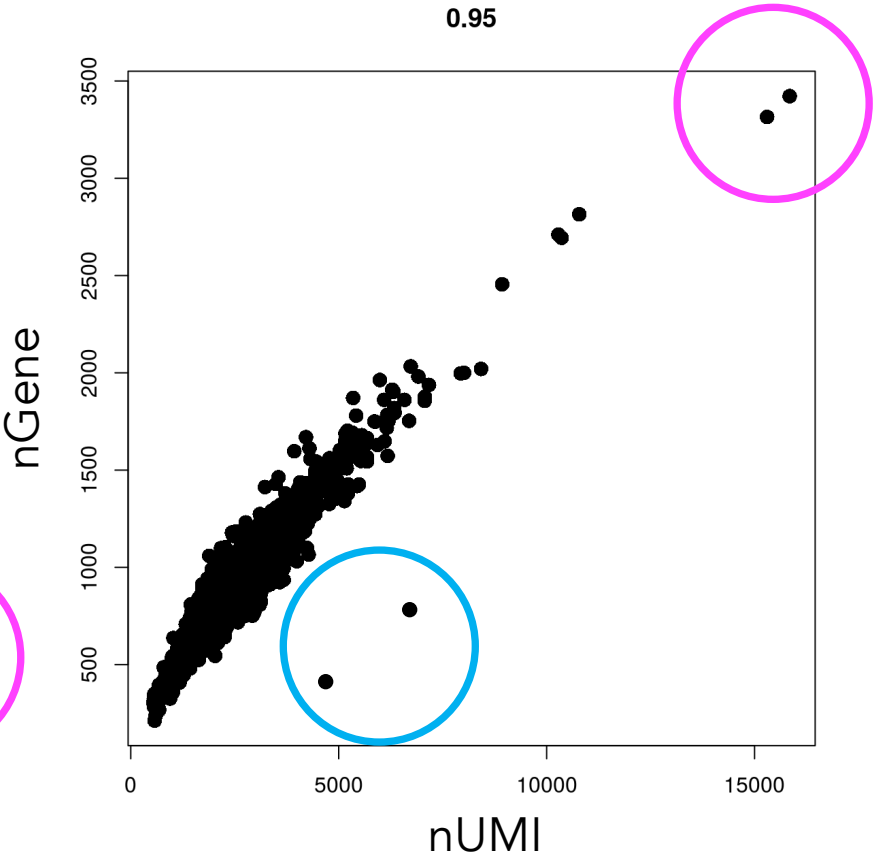
Filtering with combinations of quality control filters

Low nUMI and high % mitochondrial- Cells captured but lost a lot of the mRNA, and the mitochondrial genes were protected and retained

High nUMI & low nGene ratio – low quality library or capture rate

High nUMI & high nGene – doublets





Appropriate quality control filters vary with platform and cell types

- Different platforms set different expectations
  - e.g. Smart-Seq2 often yields more genes detected per cell than 10x Chromium.

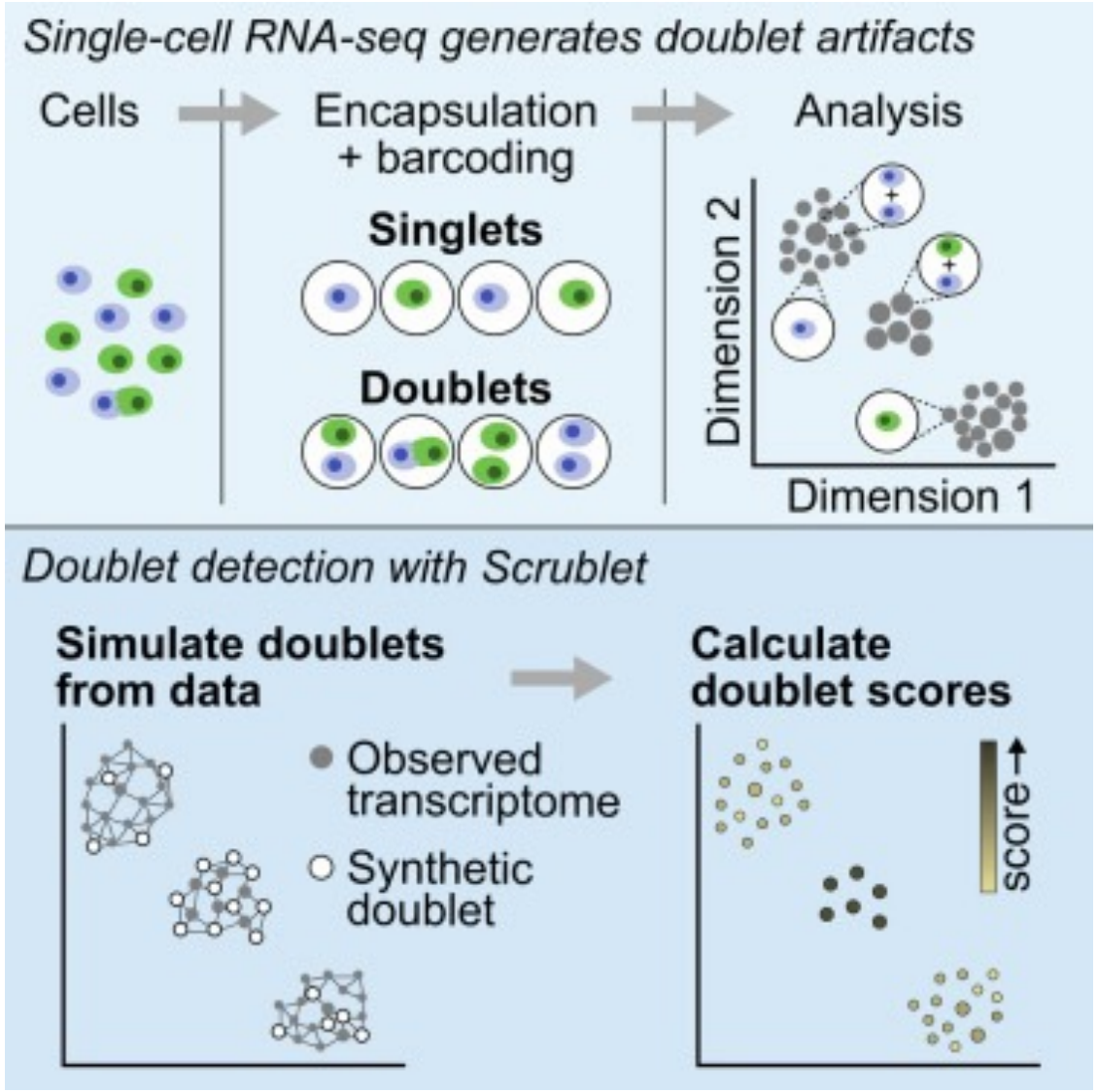
- Different cell types set different expectations
  - Immune cells normally have fewer genes detected per cell than non-immune cells
  - Malignant cells normally have more genes detected per cell than non-malignant cells

Detecting cell doublets with Scrublet

Scrublet (Single-Cell Remover of Doublets)

Wolock, Lopez, Klein. *Cell Systems* 8.4 (2019): 281-291.

Detecting cell doublets with Scrublet



Wolock, Lopez, Klein. *Cell Systems* 8.4 (2019): 281-291.

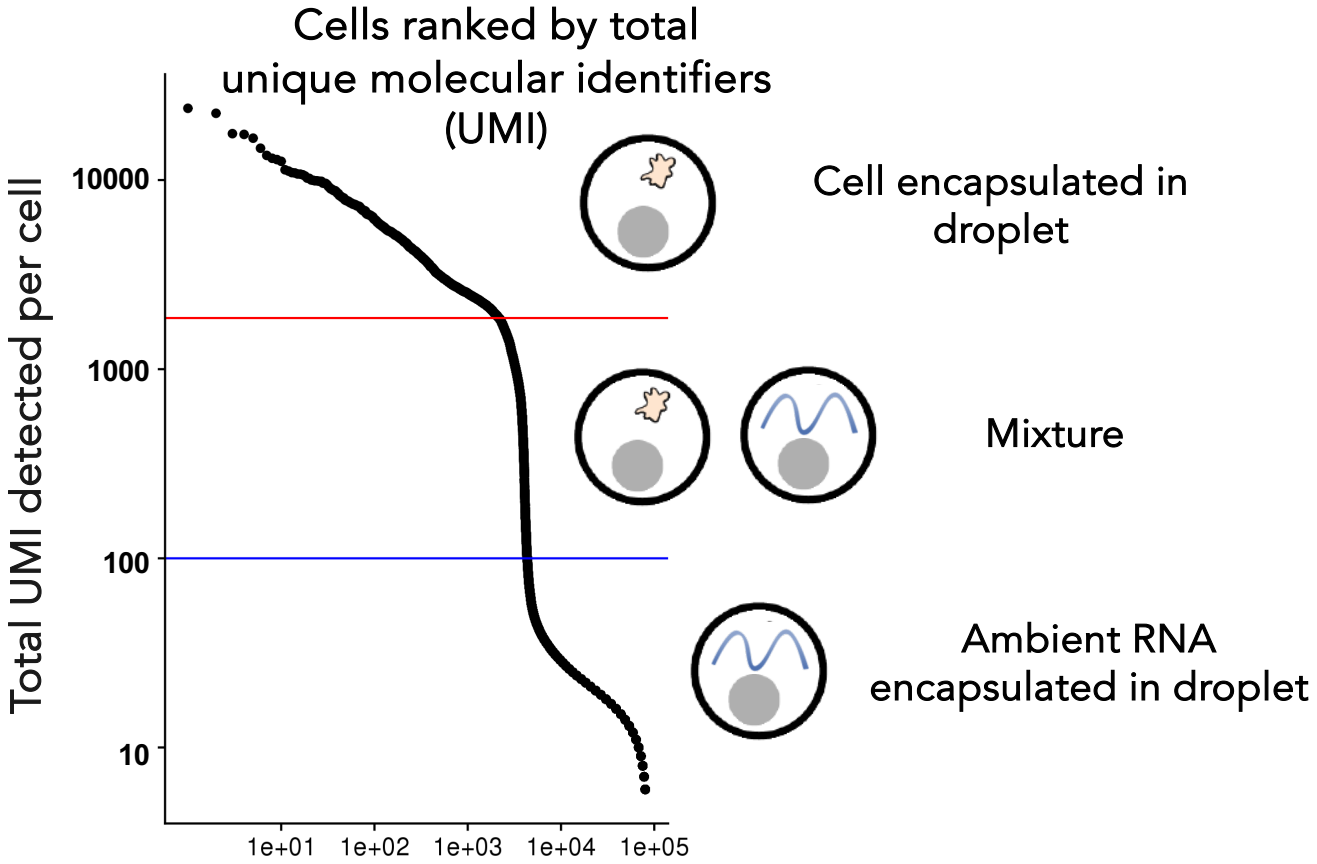
Detecting empty drops containing ambient RNA – manual

Look for transcripts expressed in unexpected cell types and remove those genes from all subsequent analysis

- e.g. hemoglobin gene expressed in a T cell

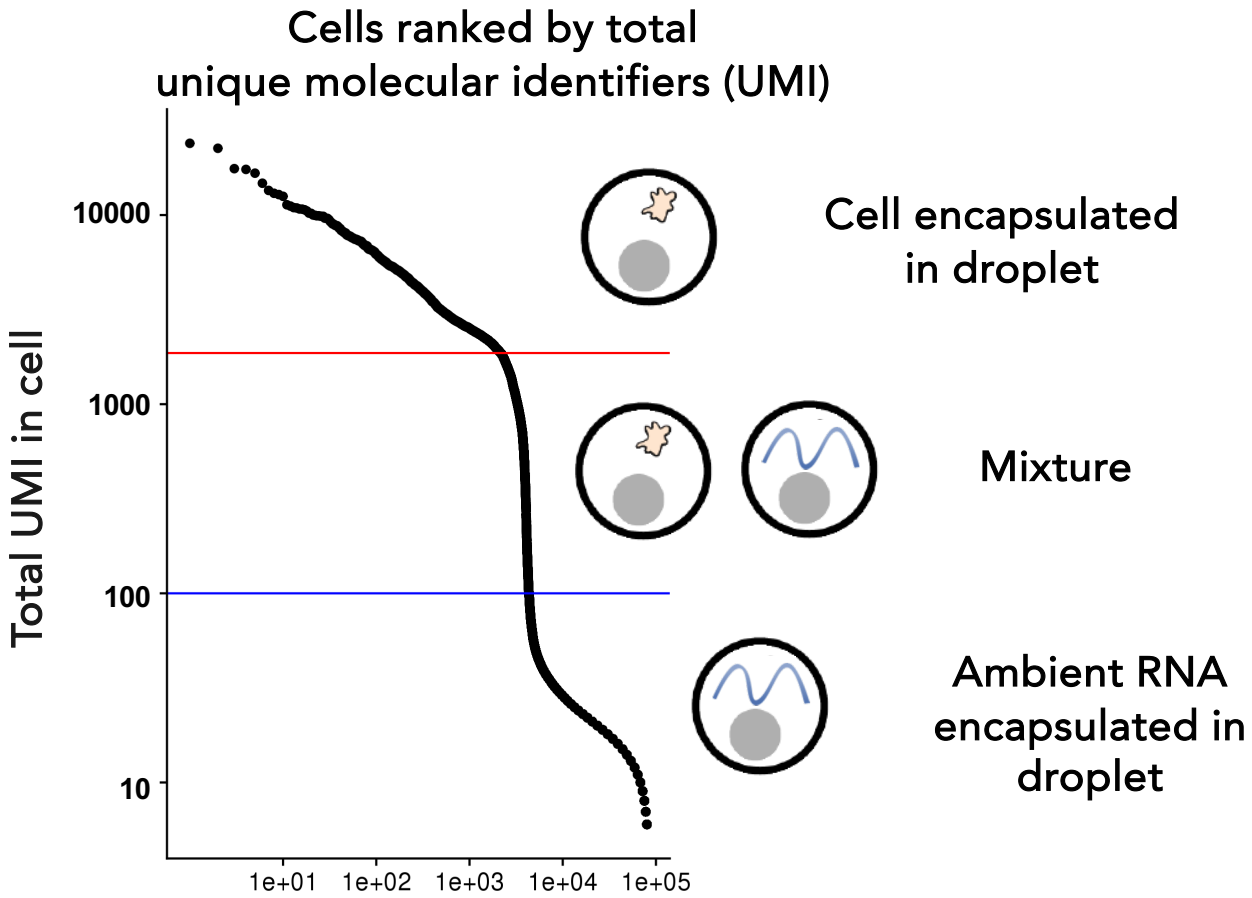
Detecting empty drops containing ambient RNA – automatic

EmptyDrops (distinguish cells from empty droplets)



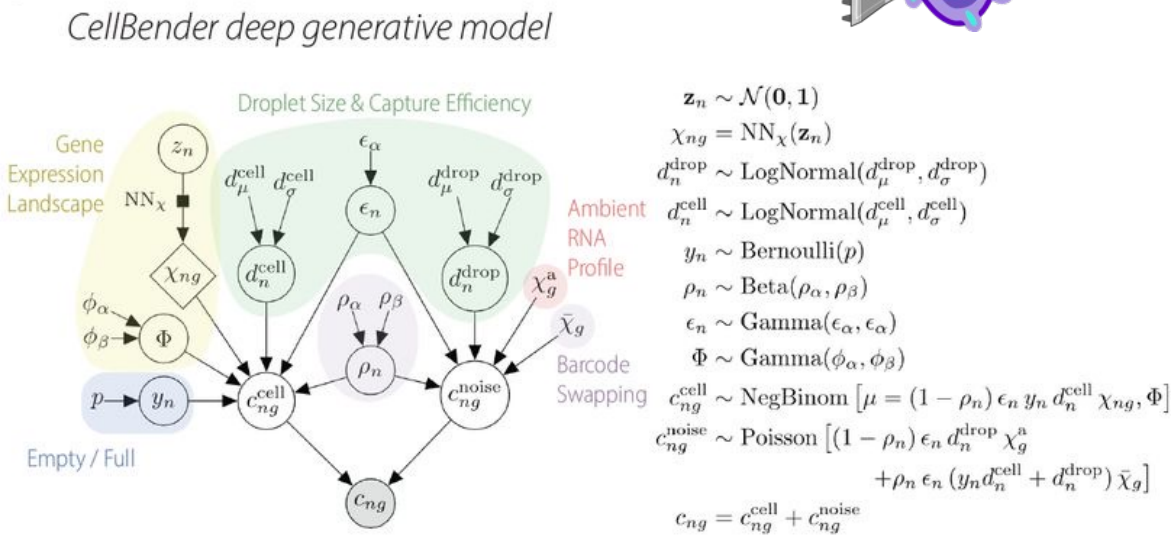
Lun, A.T.L. et al. *Genome Biol* (2019). 20, 63.

Further quality control to correct for ambient RNA



CellBender is a deep generative model of background-contaminated counts

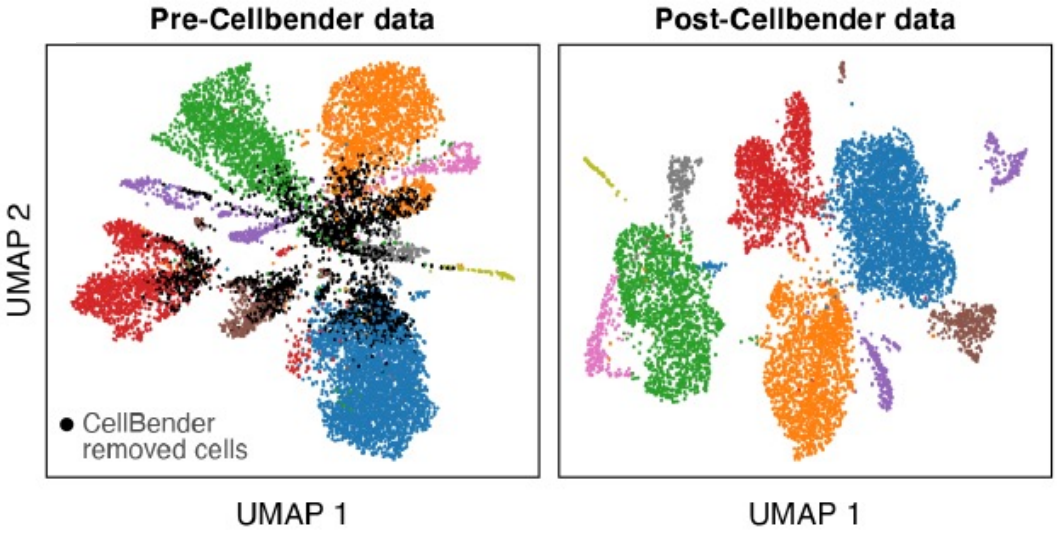
- correct ambient RNA and barcode swapping
- detect empty droplets

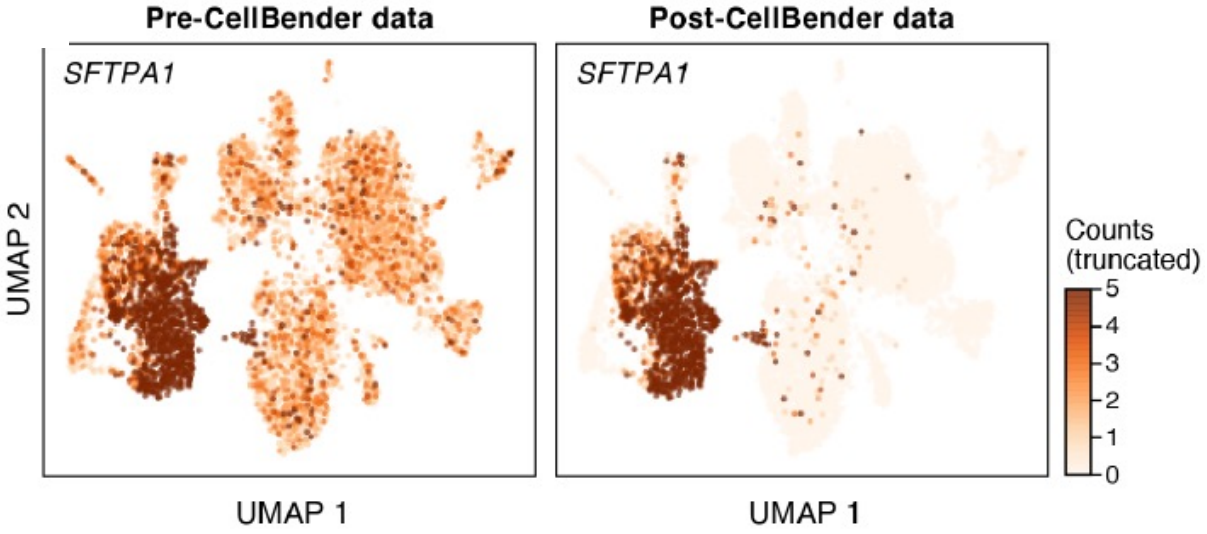


Fleming S. J. et al. *Nature Methods* (2023).

Removing ambient RNA using CellBender in COVID-19 tissue

CellBender applied to COVID-19 lung tissue





CellBender qualitative observations

- cell subsets become more distinct
- cell type marker genes become more specific
- can lower UMI and gene cutoffs, allowing for significantly more recovery of lymphocytes

Delorey T. et al. *Nature* (2021).

Quality control tips

- We often revisit quality control decisions multiple times when analyzing data.
- Start with permissive thresholds when filtering, and investigate the effects of these thresholds before applying more stringent thresholds.
- If the distributions of QC metrics differ between samples, thresholds should be determined separately for each sample to account for sample quality differences.
- Visualize QC metrics per cell subset in order to flag technical biases.

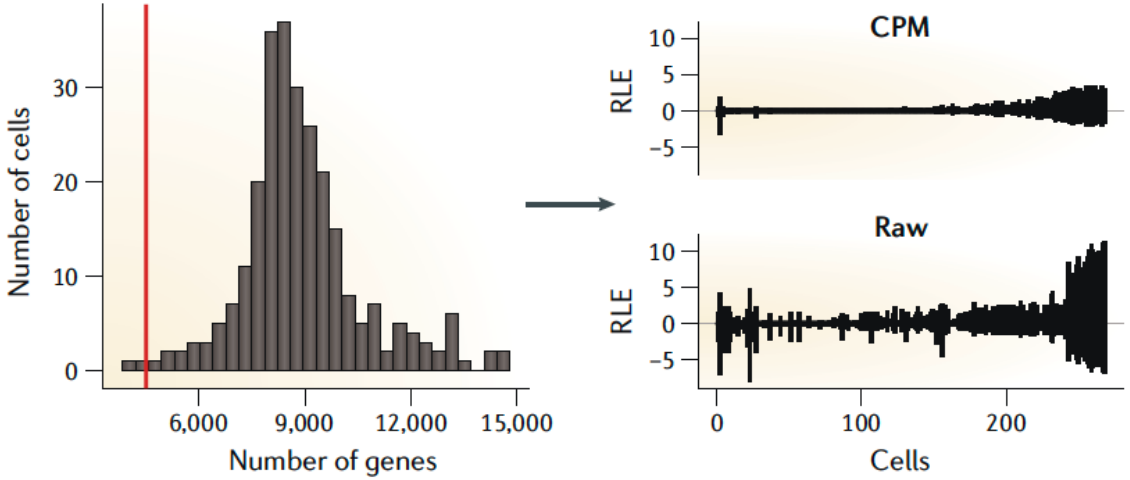
Luecken M. D. et al. *Molecular Systems Biology* (2019) 15:e8746.

Determining cell type, state, and function

Quality control

Normalization

Feature selection



Dimensional reduction

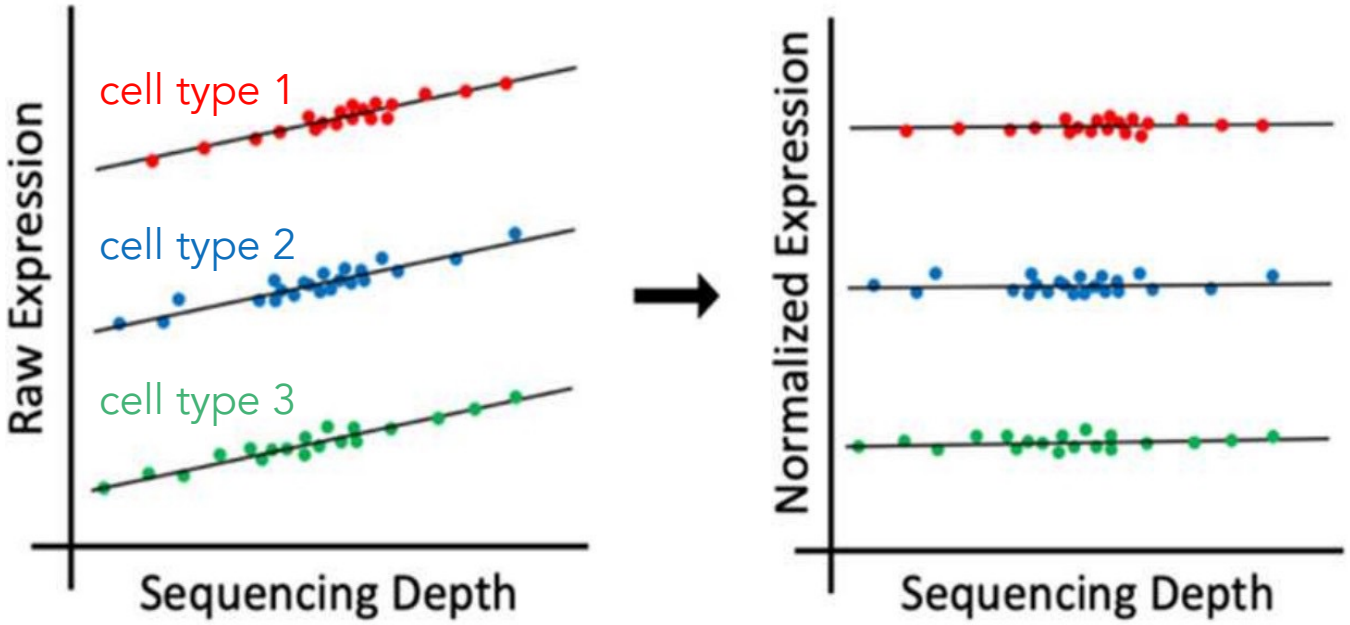
Cell-cell distances

Unsupervised clustering

Kiselev C. at *al.* (2019) *Nature Reviews Genetics*. 20: 273-282.

Single-cell RNA-Seq analysis: normalization

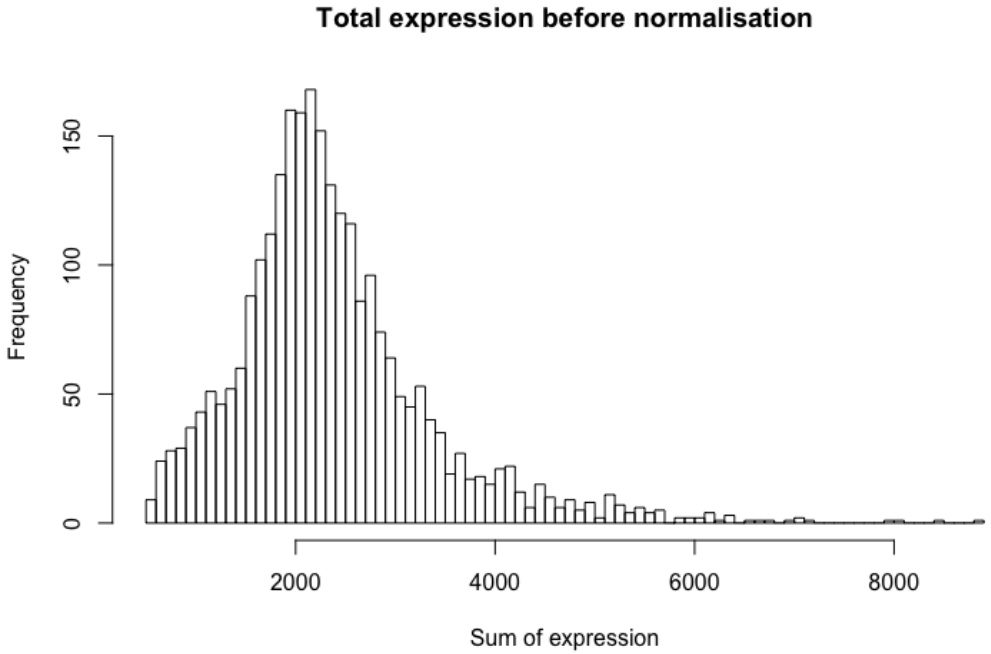
- Why normalize gene expression within a cell?



Differences in sequencing depth can lead to false differences in gene expression.

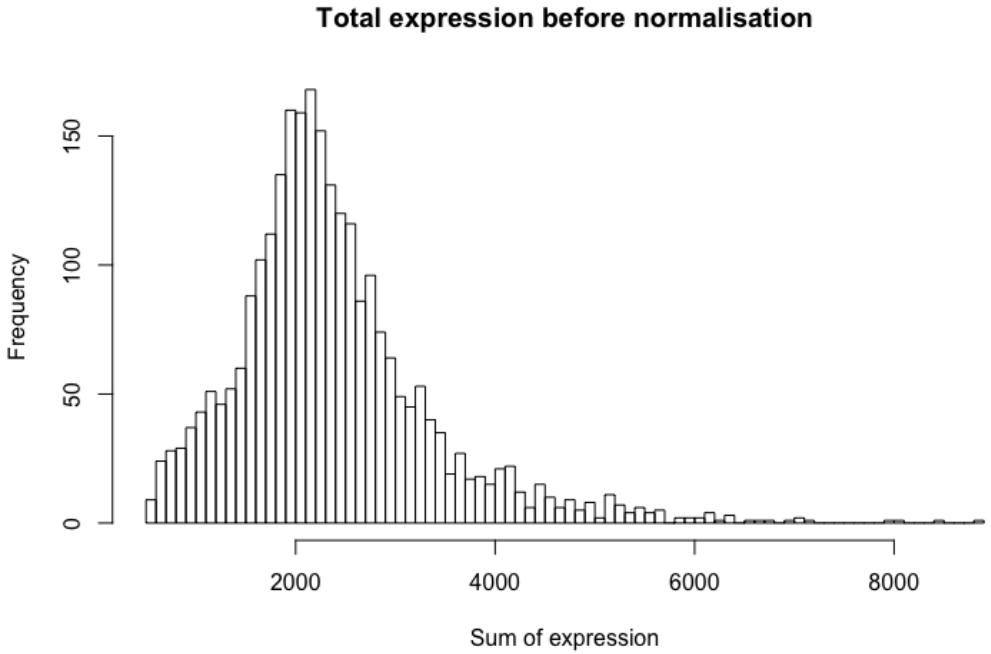
Single-cell RNA-Seq analysis: normalization

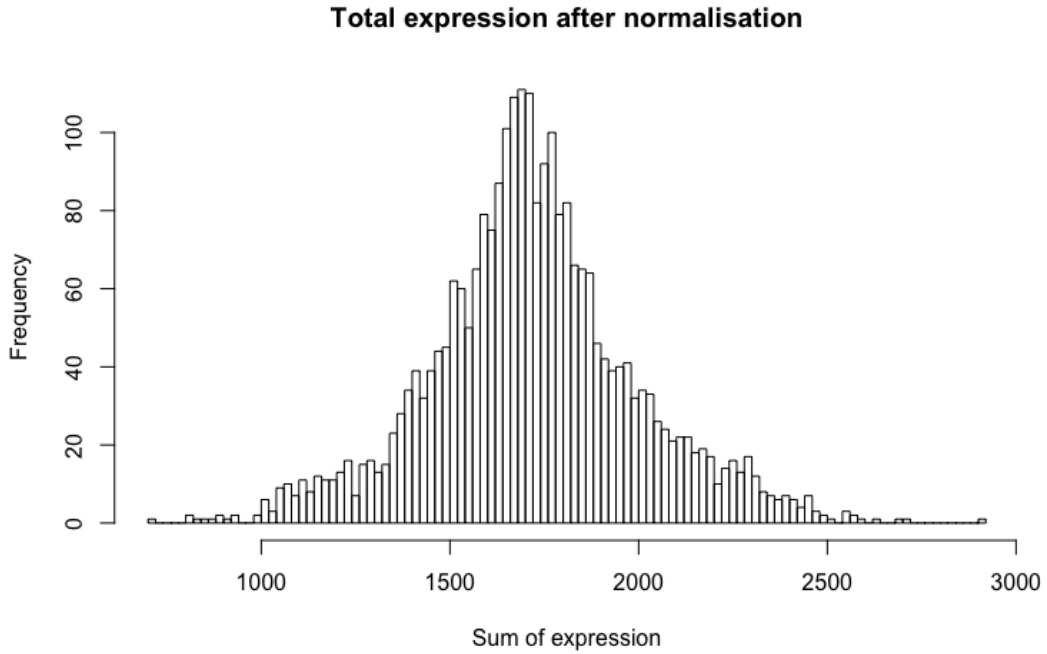
- Why normalize gene expression within a cell?
  - cells are sequenced to different depths (technical)
  - cells of different type have different amounts of mRNA (biological)
  - there are typically extreme values in distribution of gene expression
  - more highly expressed genes are more variable



Single-cell RNA-Seq analysis: normalization

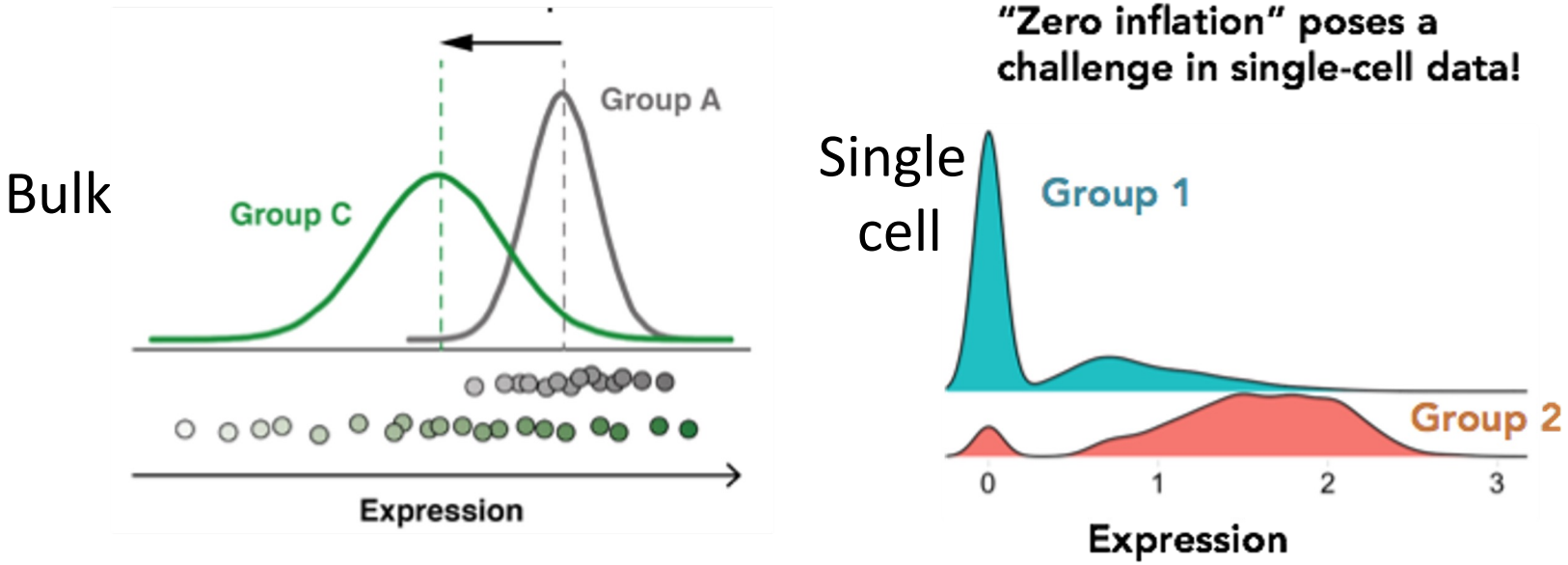
- How to normalize
  - Gene expression measurements for each cell are normalized by the total gene expression or median gene expression
  - Gene expression values then scaled to sum to 10,000 (typically), and then log(1+x)-transformed.





Is standard normalization appropriate?

Reassessing the idea that droplet scRNA-Seq is zero-inflated.

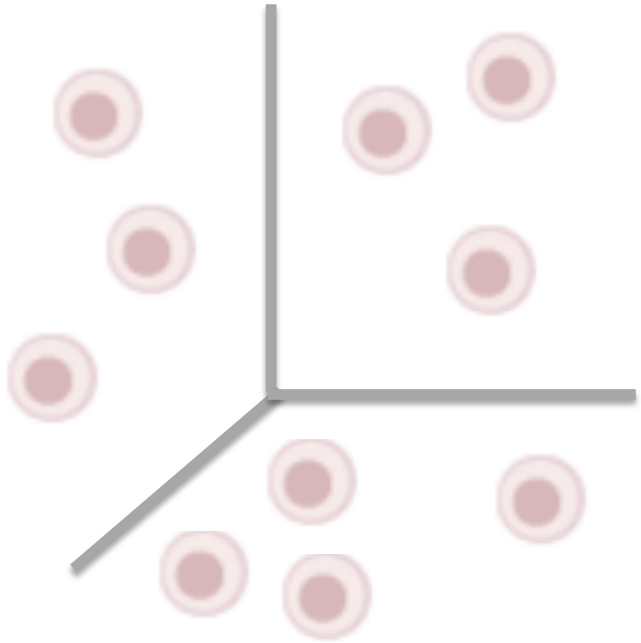


- "Droplet scRNA-seq is not zero-inflated." Svensson, *Nature Biotechnology* (2020)
- "Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model." Townes et al. *Genome Biology* (2019)
- "Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression." Hafemeister et al. *Genome Biology* (2019)
- "Statistics or biology: the zero-inflation controversy about scRNA-seq data." Jiang et al. *Genome Biology* (2022)

Identify highly variable genes

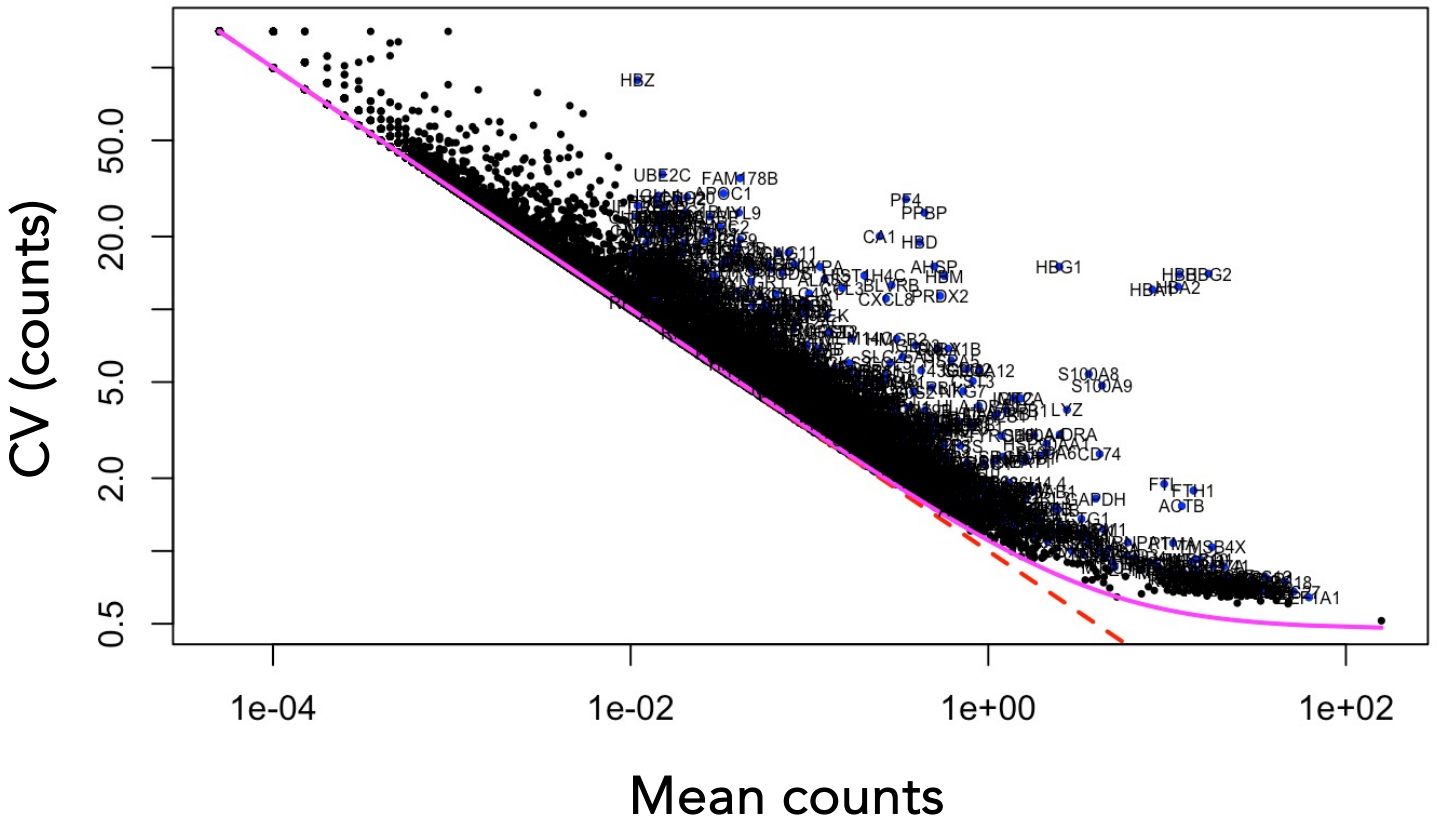
Cells are in ~20,000 dimensional space (one dimension for each gene)

- many genes are lowly detected or noisy measurements



- variable genes contain the biological signal we are interested in

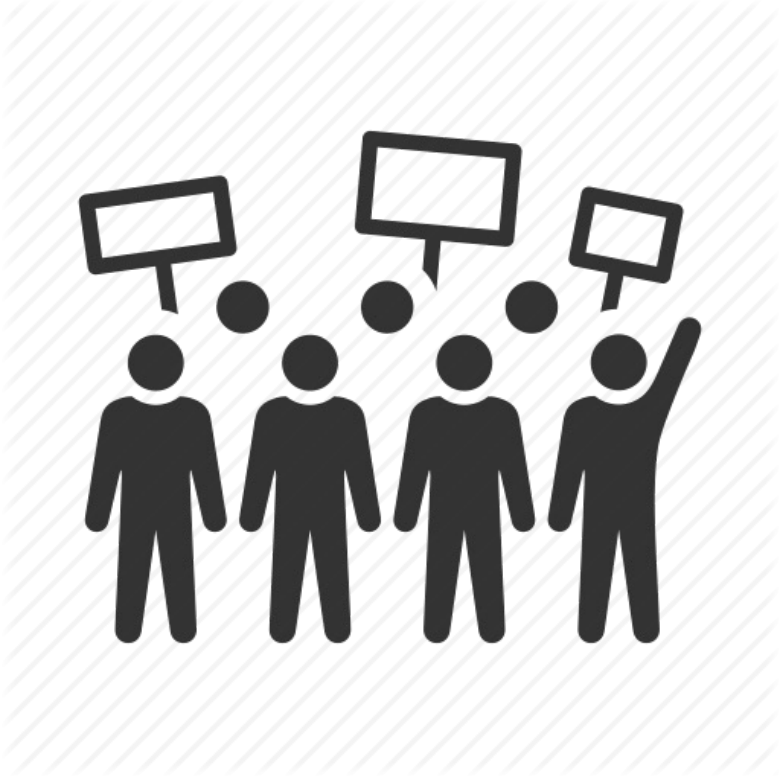
Identify highly variable genes



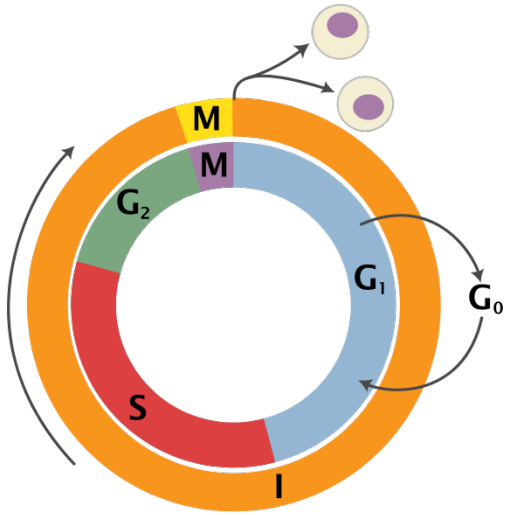
Find genes (features) that are outliers in a plot of mean of gene expression vs variance of gene expression

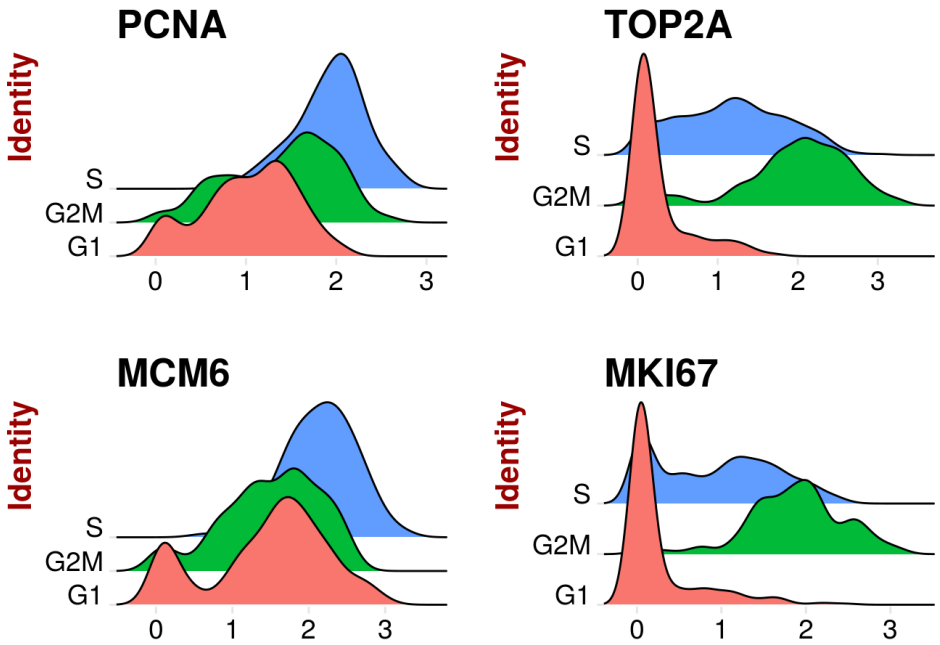
Calculating gene signatures

Relying on capturing a specific gene is not robust, but relying on a set of genes (signature) is much more stable!

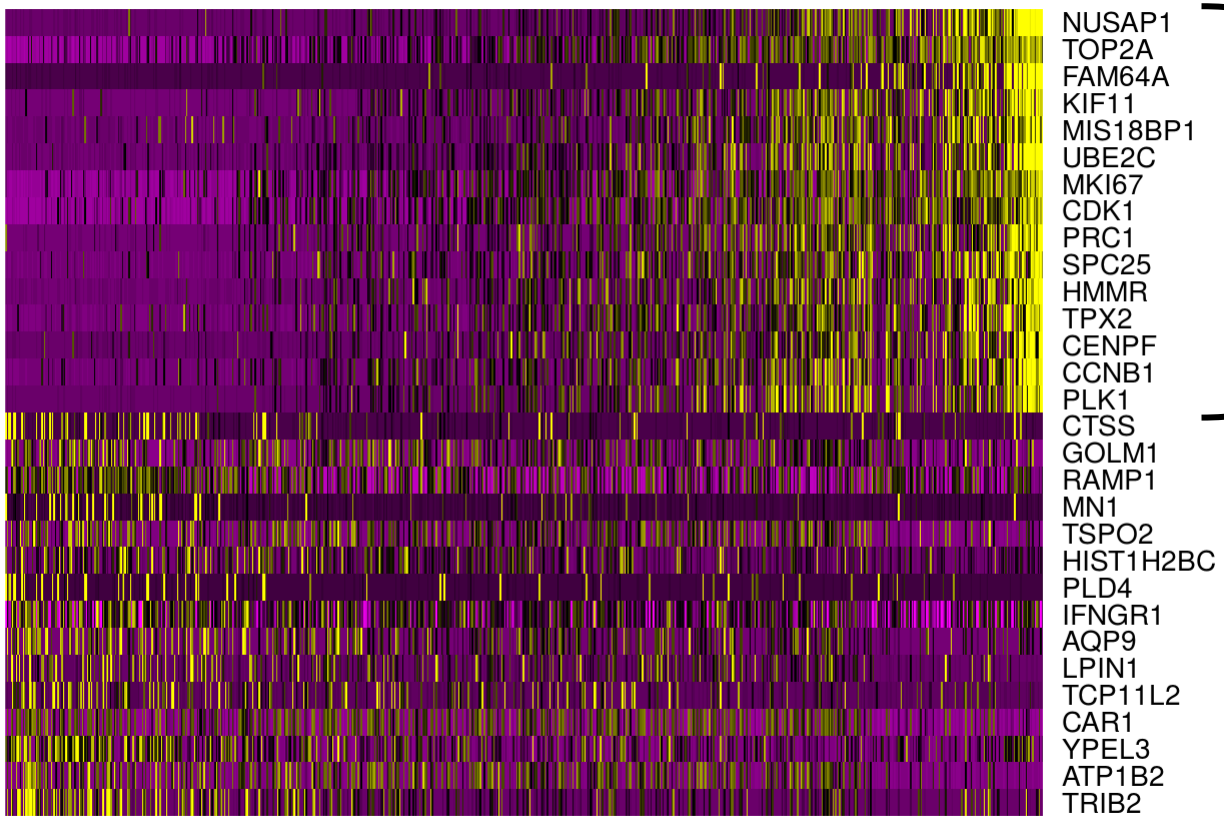


Gene signature example: cell cycle markers





variability of individual genes



Seurat and scanpy: single cell analysis toolkits

scanpy

https://satijalab.org/seurat/
https://scanpy.readthedocs.io/en/stable/

Interacting with Seurat objects

- Seurat object is used for to store 10x data and perform analysis
  - Count matrices for different assays are stored (gene expression, protein expression, chromatin accessibility, etc…)
  - Counts are stored as: counts (raw), data (normalized), scaled data (centered and scaled) in sparse matrices when possible
  - Metadata describes individual cells and genes
  - Functions for analysis (quality control, normalization, feature selection, dimensional reduction, cell-cell distances, unsupervised clustering)

https://github.com/satijalab/seurat/wiki
https://satijalab.org/seurat/essential_commands.html

Interacting with Seurat objects

```
> gcdata
An object of class Seurat
35633 features across 2000 samples within 2 assays
Active assay: RNA (33633 features)
 1 other assay present: integrated
 2 dimensional reductions calculated: pca, umap
```

Seurat object

```
> gcdata[['RNA']]@data[1:5,1:5]
5 x 5 sparse Matrix of class "dgCMatrix"
         D2ex_5 D2ex_6 D2ex_7  D2ex_11 D2ex_13
A1BG-AS1      .      .      . .               .
A1BG          .      .      . .               .
A1CF          .      .      . .               .
A2M-AS1       .      .      . .               .
A2ML1         .      .      . 1.226772        .
```

Accessing count slot from RNA assay

```
> gcdata[[]][1:5, 1:5]
        orig.ident nCount_RNA nFeature_RNA   tech integrated_snn_res.1
D2ex_5        D2ex   5745.867         2548 celseq                    4
D2ex_6        D2ex   6883.692         2619 celseq                    6
D2ex_7        D2ex   7460.202         3043 celseq                    5
D2ex_11       D2ex   8330.644         3465 celseq                    5
D2ex_13       D2ex   3891.960         1962 celseq                    6
```

Accessing cell metadata

```
> gcdata <- ScaleData(gcdata)
Centering and scaling data matrix
  |=========================================================================| 100%
```

Running analysis function

Loading data into a Seurat object

```
gcdata <- CreateSeuratObject(counts = celseq.data)
```

counts matrix

	Cell 1	Cell 2	Cell 3	Cell 4	Cell 5	Cell 6	…	Cell 70K
Tcf7	3	3	0	0	0	0	…	4
Bach2	2	4	1	0	0	0	…	2
Prf1	1	0	5	3	1	1	…	1
Gzma	0	0	3	1	0	0	…	0
Pdcd1	0	1	1	0	4	6	…	0
Eomes	0	0	1	0	3	3	…	1
…	…	…	…	…	…	…	…	…
Gene 20K	2	1	0	1	0	0	…	3

Storing counts data in dense vs sparse format

2D Arrays

cells

	1	2	3	4	5	6	7	8	
	0	0	0	0	0	0	0	0	1
	0	1	0	0	0	0	0	0	2
	0	0	0	0	0	2	0	0	3
	0	0	0	0	0	0	0	0	4
	0	0	0	0	0	0	0	0	5
	0	0	0	0	0	0	0	3	6

genes

Dense matrices

Coordinate List

2	2	1
6	3	2
8	6	3

Sparse matrices