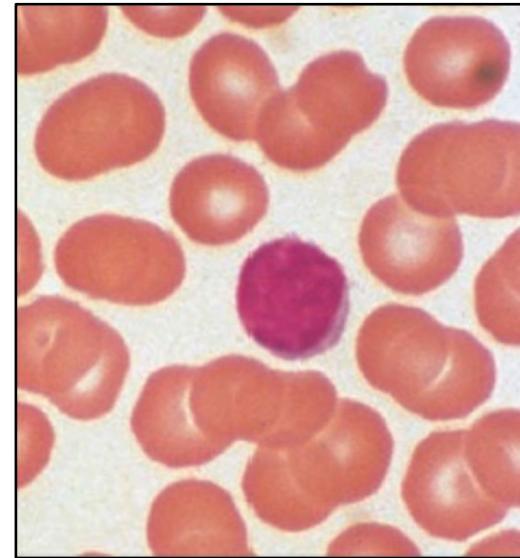
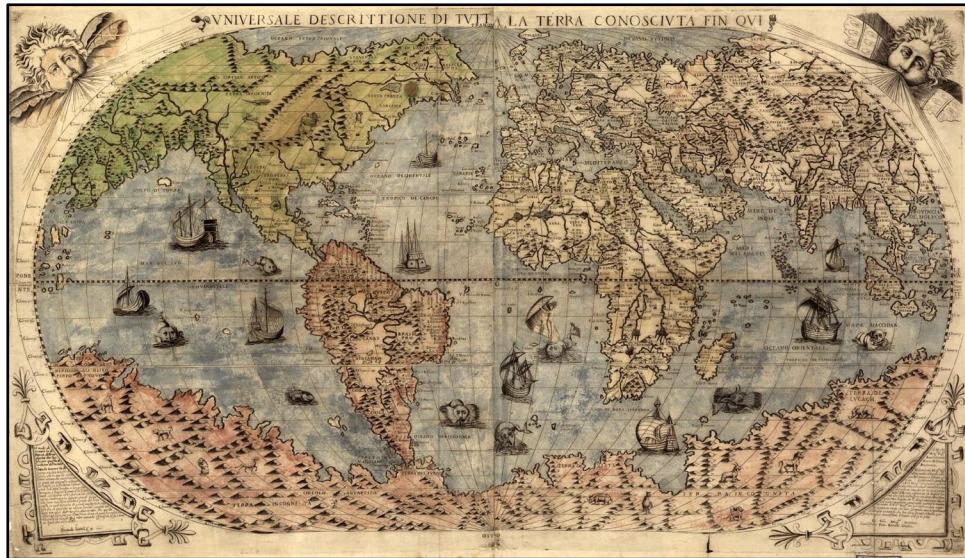


Quality control of scRNA-Seq data

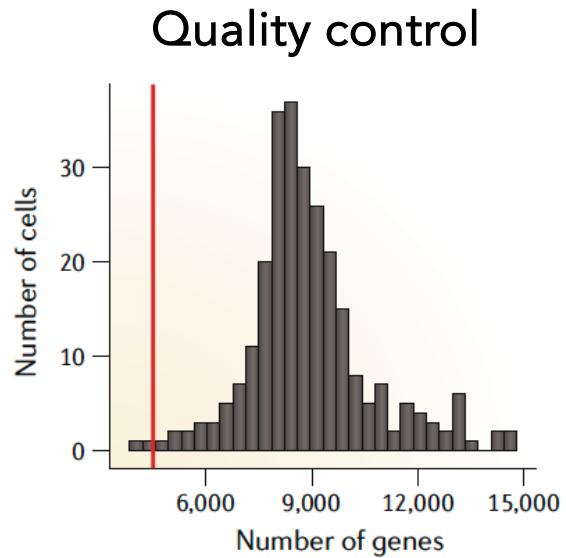
HCA Latin America
Single-Cell RNA-Seq Computational Workshop
October 23, 2023



Outline: Quality control of scRNA-Seq data

- Quality control and normalization starting from gene count matrices.
- Interacting with Seurat objects.

Determining cell type, state, and function



Normalization

Feature selection

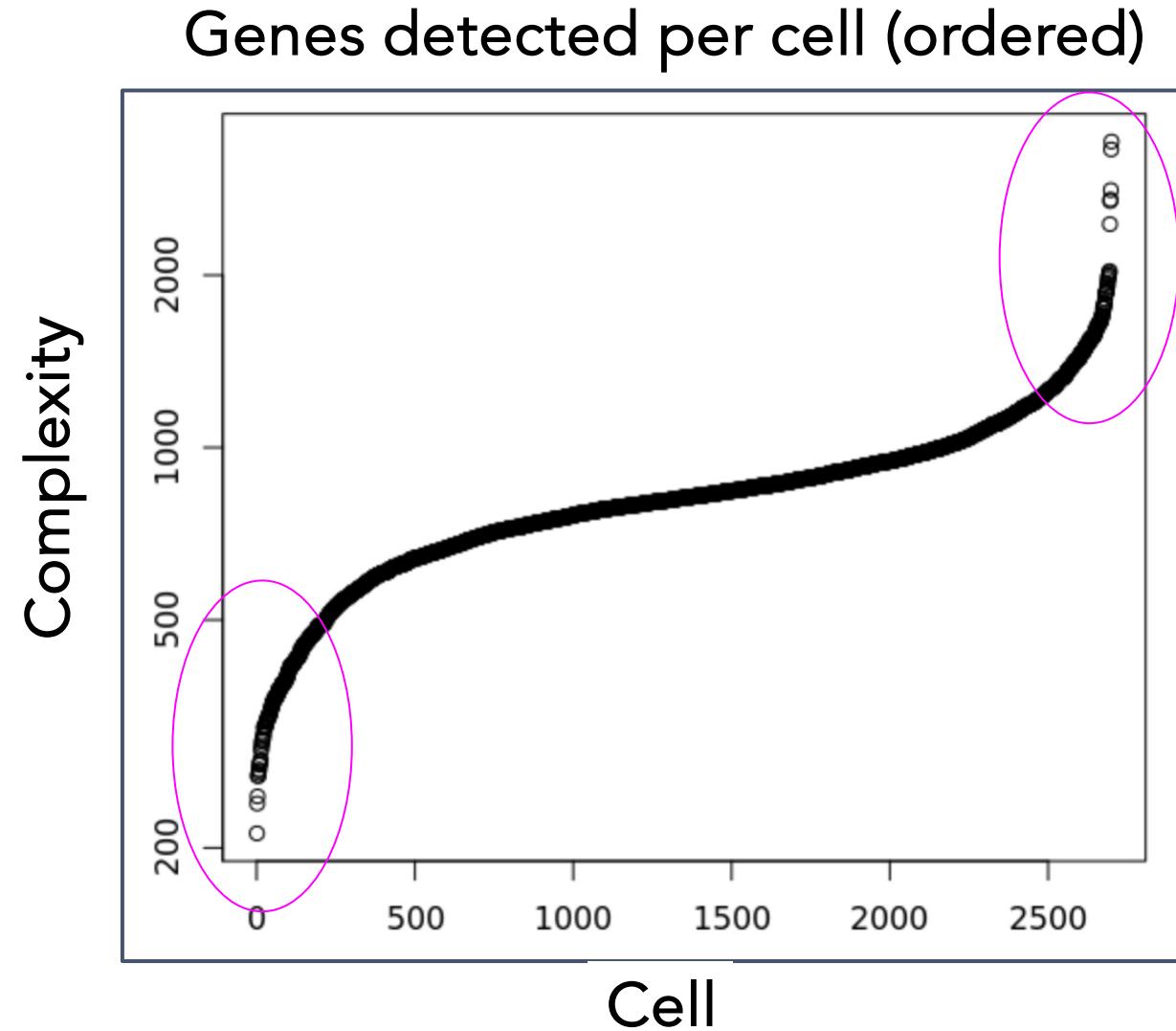
Dimensional reduction

Cell-cell distances

Unsupervised clustering

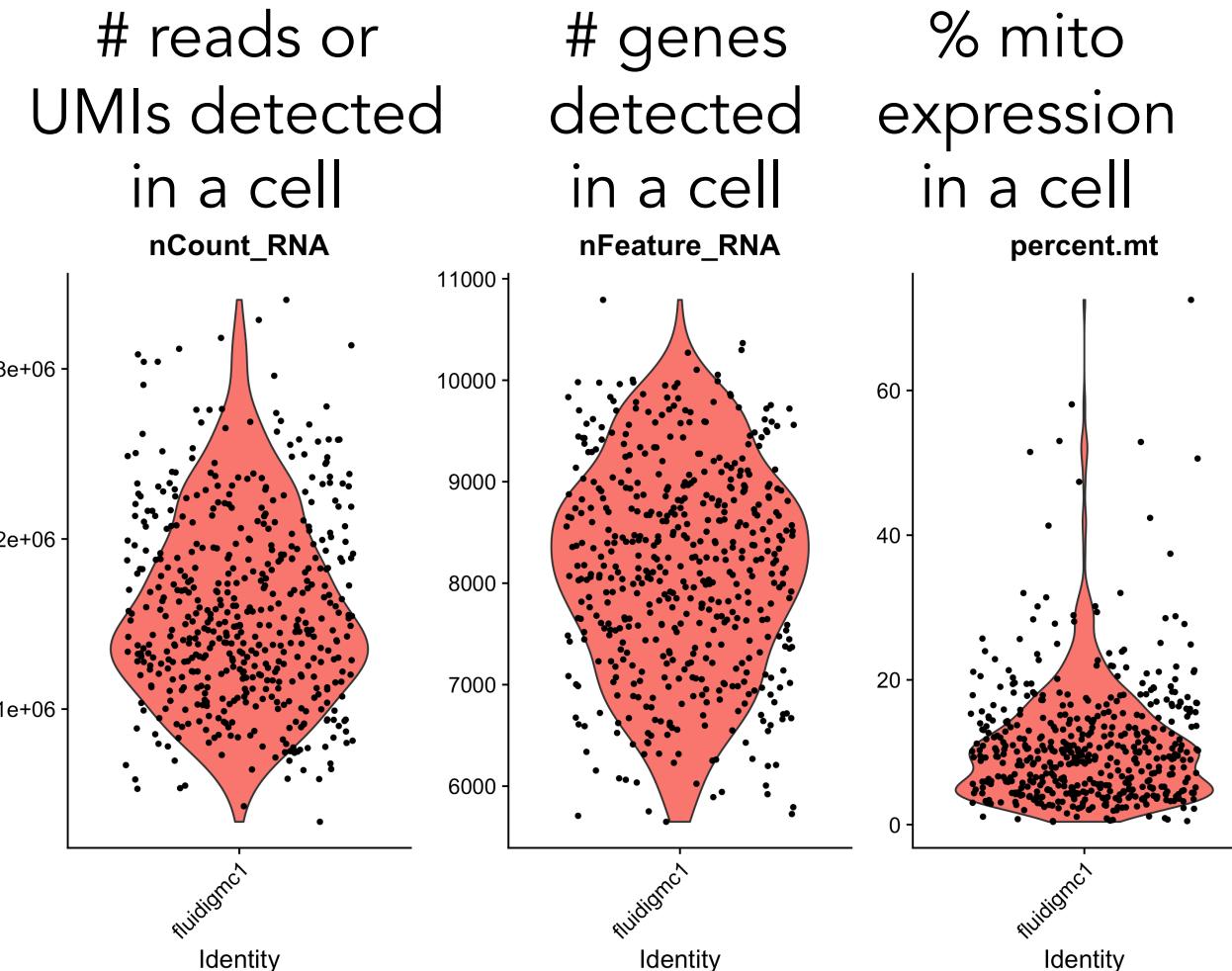
There are many quality control filters for genes and cells

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.

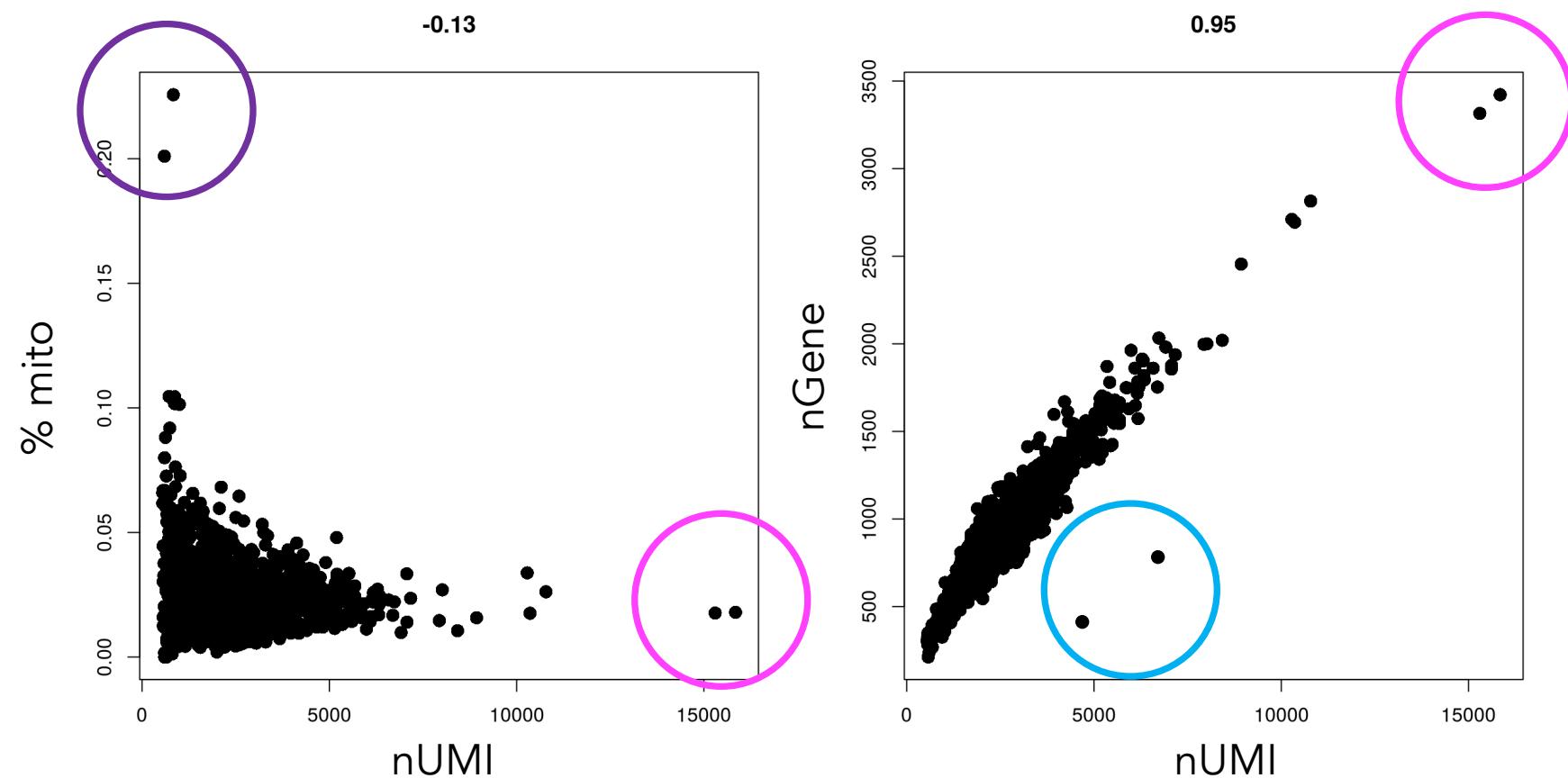


Filtering with combinations of quality control filters

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 – doubles



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

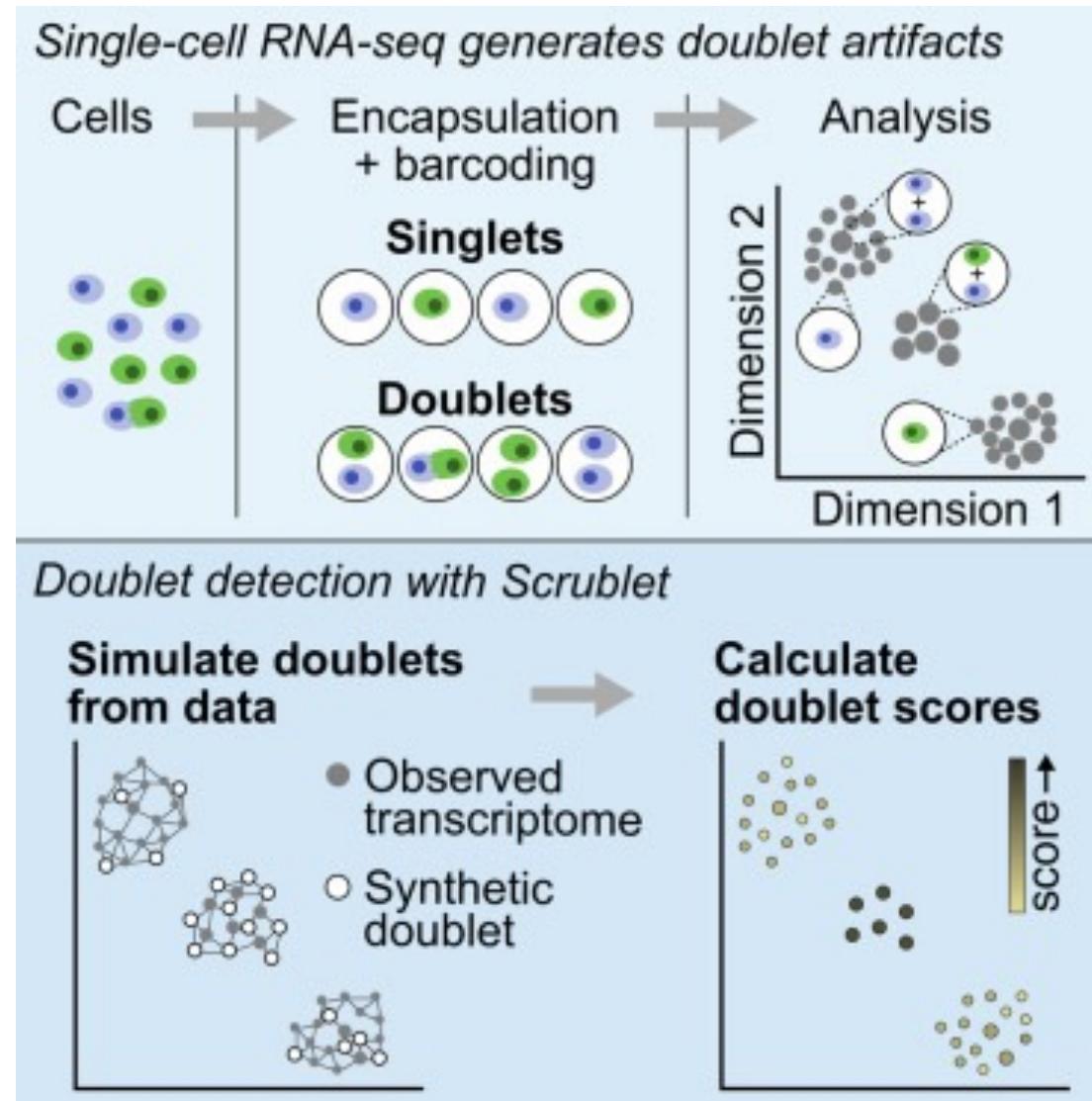
Detecting cell doublets with Scrublet

Scrublet (Single-Cell Remover of Doublets)

Singlets



Detecting cell doublets with Scrublet



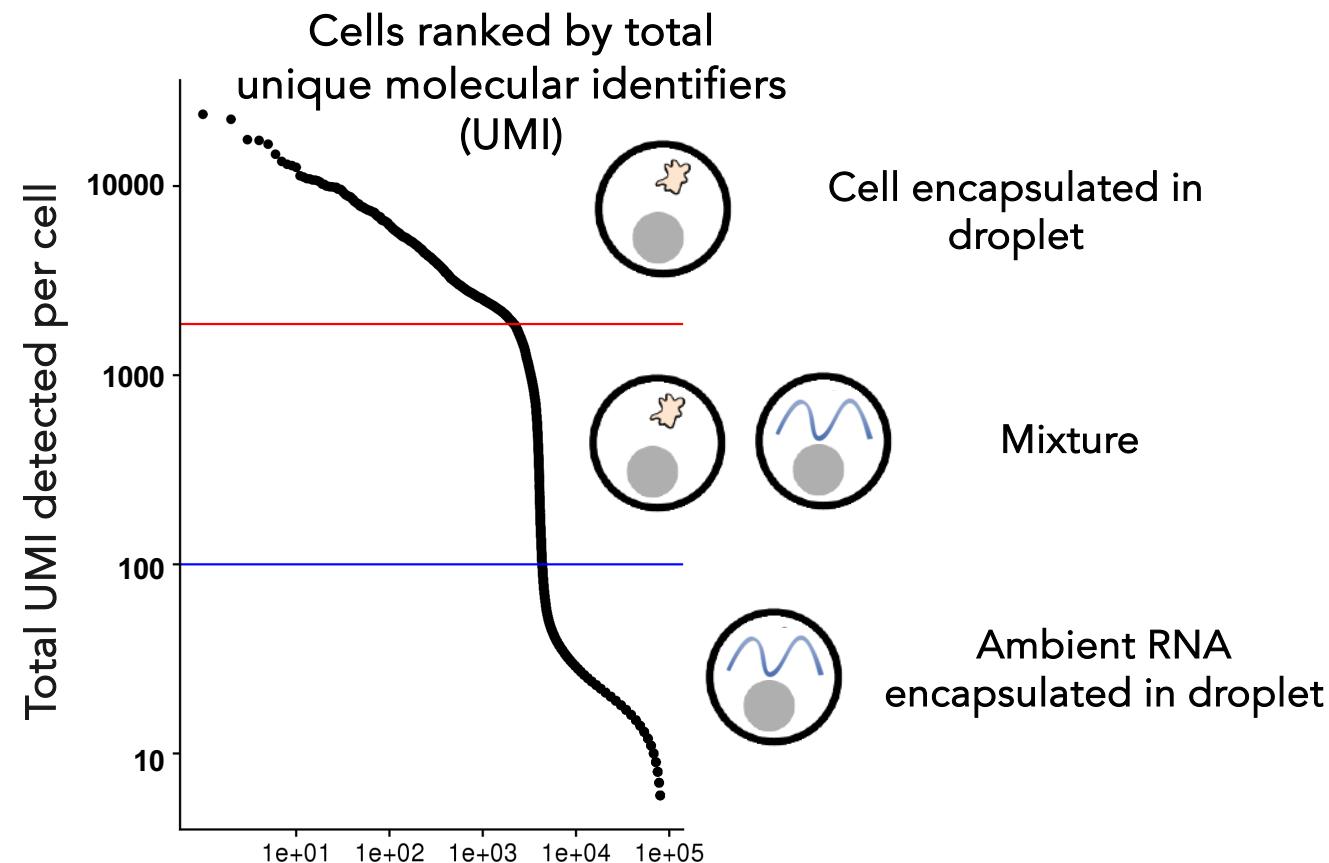
Detecting empty drops containing ambient RNA – manual

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

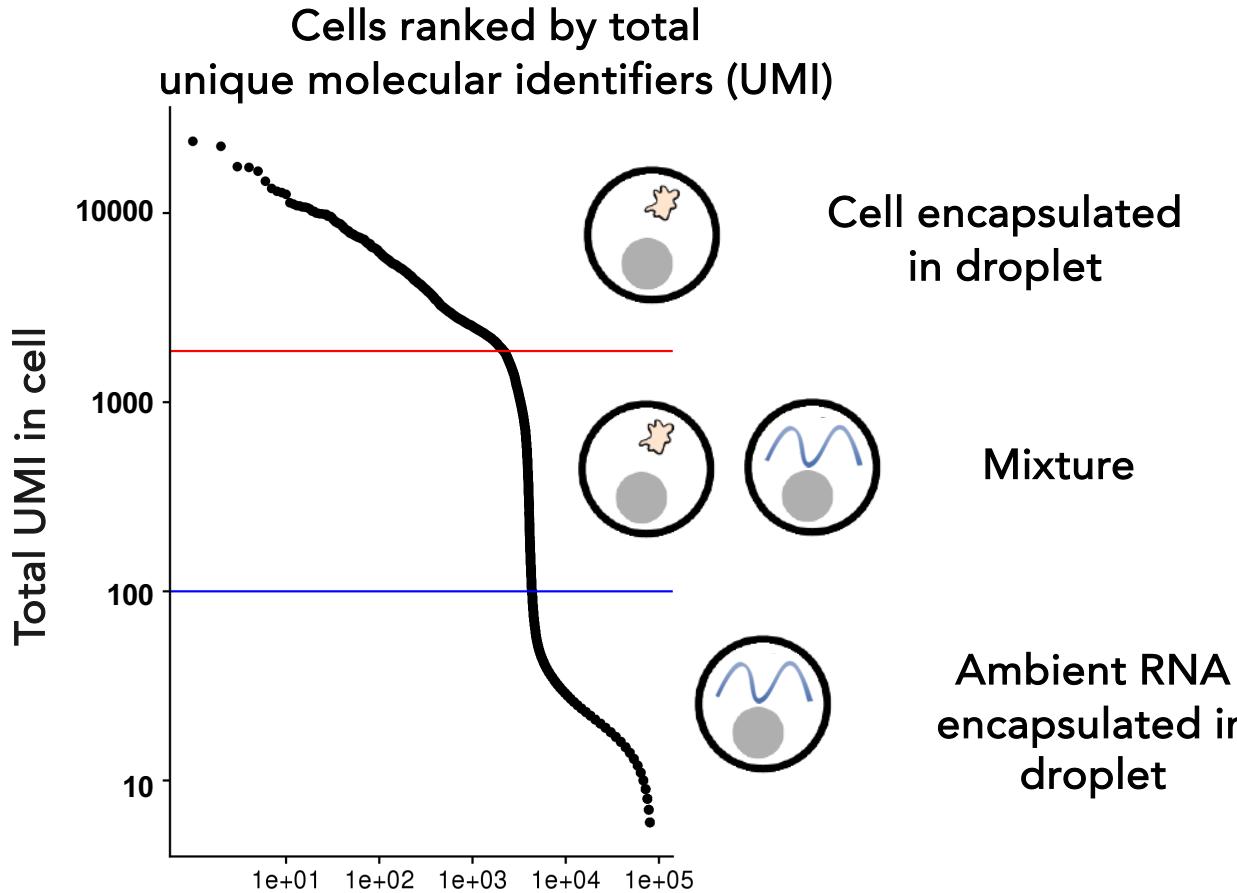
- e.g. hemoglobin gene expressed in a T cell

Detecting empty drops containing ambient RNA – automatic

EmptyDrops (distinguish cells from empty droplets)

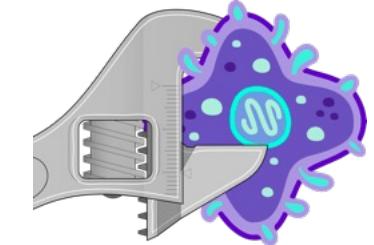


Further quality control to correct for ambient RNA

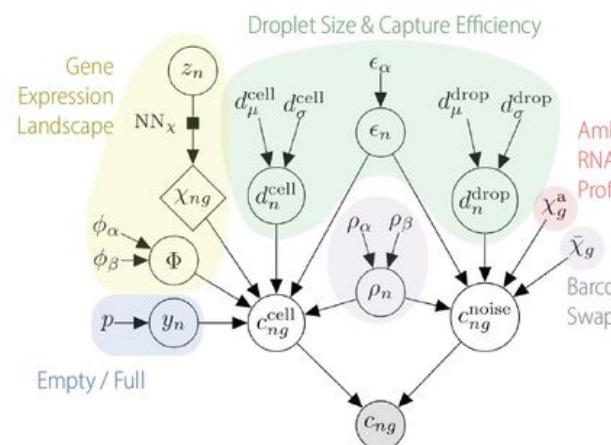


CellBender is a deep generative model of background-contaminated counts

- correct ambient RNA and barcode swapping
- detect empty droplets



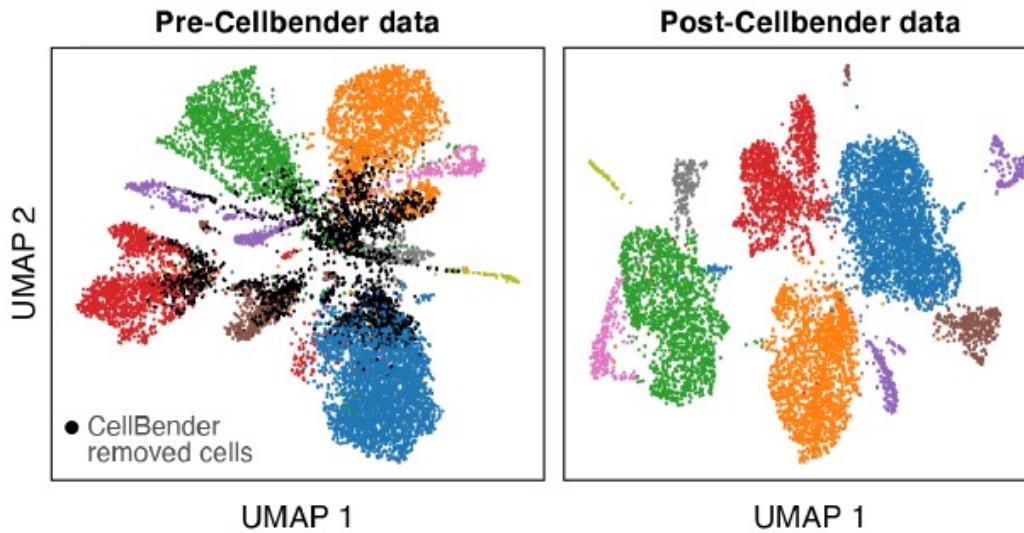
CellBender deep generative model



$$\begin{aligned}
 z_n &\sim \mathcal{N}(\mathbf{0}, \mathbf{1}) \\
 \chi_{ng} &= \text{NN}_x(z_n) \\
 d_n^{\text{drop}} &\sim \text{LogNormal}(d_{\mu}^{\text{drop}}, d_{\sigma}^{\text{drop}}) \\
 d_n^{\text{cell}} &\sim \text{LogNormal}(d_{\mu}^{\text{cell}}, d_{\sigma}^{\text{cell}}) \\
 y_n &\sim \text{Bernoulli}(p) \\
 \rho_n &\sim \text{Beta}(\rho_{\alpha}, \rho_{\beta}) \\
 \epsilon_n &\sim \text{Gamma}(\epsilon_{\alpha}, \epsilon_{\alpha}) \\
 \Phi &\sim \text{Gamma}(\phi_{\alpha}, \phi_{\beta}) \\
 c_{ng}^{\text{cell}} &\sim \text{NegBinom} [\mu = (1 - \rho_n) \epsilon_n y_n d_n^{\text{cell}} \chi_{ng}, \Phi] \\
 c_{ng}^{\text{noise}} &\sim \text{Poisson} [(1 - \rho_n) \epsilon_n d_n^{\text{drop}} \chi_g^a + \rho_n \epsilon_n (y_n d_n^{\text{cell}} + d_n^{\text{drop}}) \bar{\chi}_g] \\
 c_{ng} &= c_{ng}^{\text{cell}} + c_{ng}^{\text{noise}}
 \end{aligned}$$

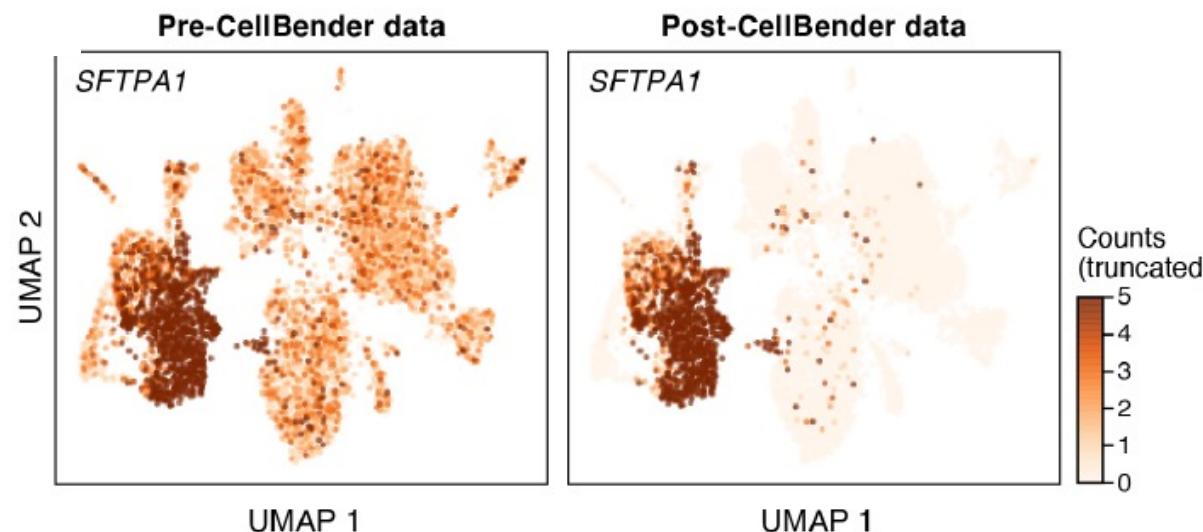
Removing ambient RNA using CellBender in COVID-19 tissue

CellBender applied to COVID-19 lung tissue



CellBender qualitative observations

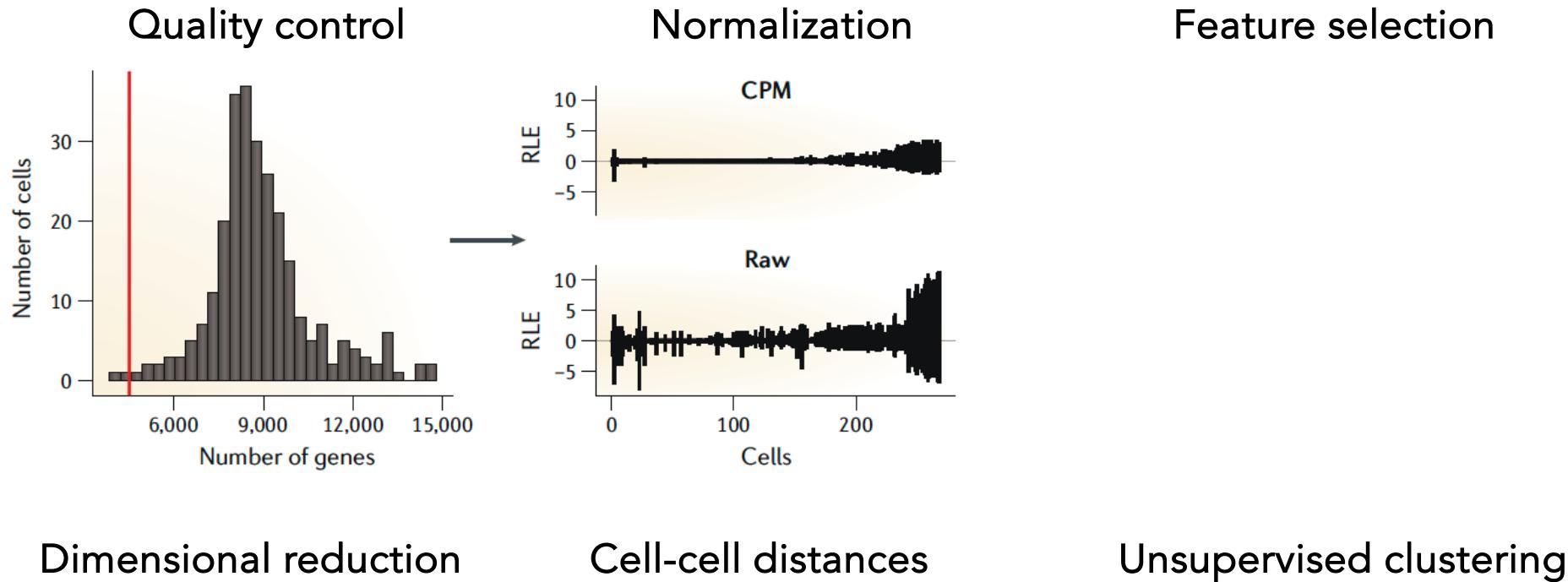
- cell subsets become more distinct
- cell type marker genes become more specific
- can lower UMI and gene cutoffs, allowing for significantly more recovery of lymphocytes



Quality control tips

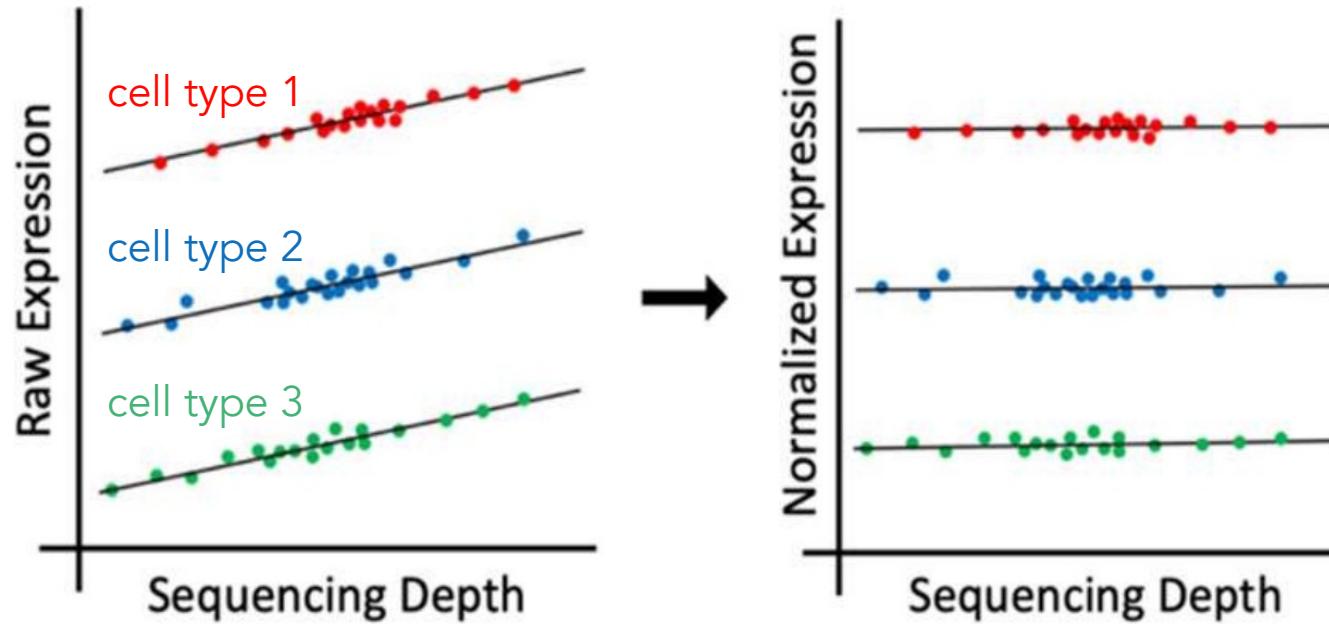
- We often revisit quality control decisions multiple times when analyzing data.
- Start with permissive thresholds when filtering, and investigate the effects of these thresholds before applying more stringent thresholds.
- If the distributions of QC metrics differ between samples, thresholds should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

Determining cell type, state, and function



Single-cell RNA-Seq analysis: normalization

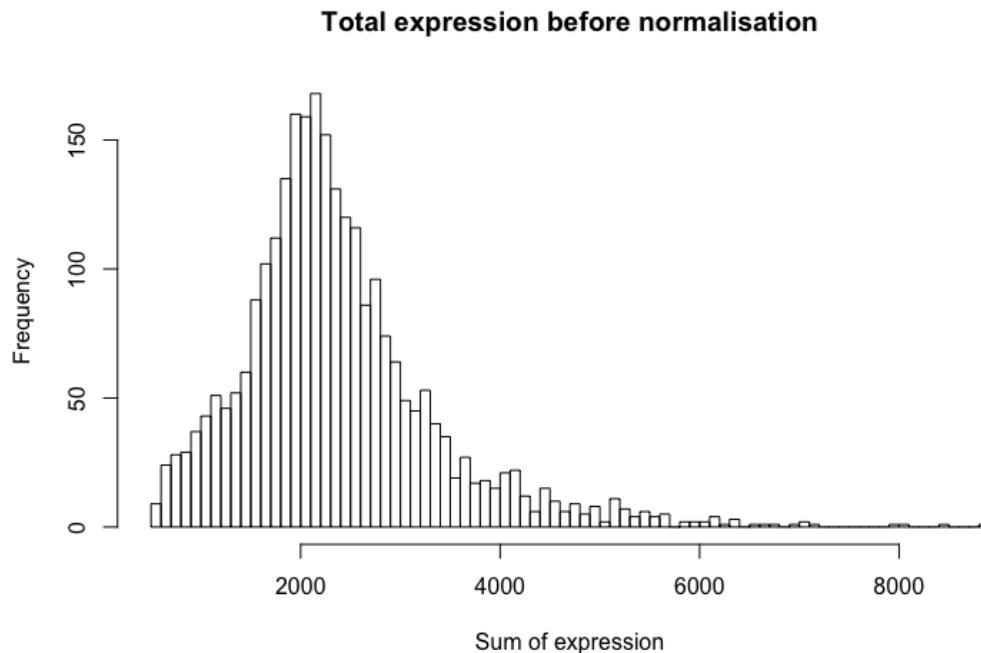
- Why normalize gene expression within a cell?



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

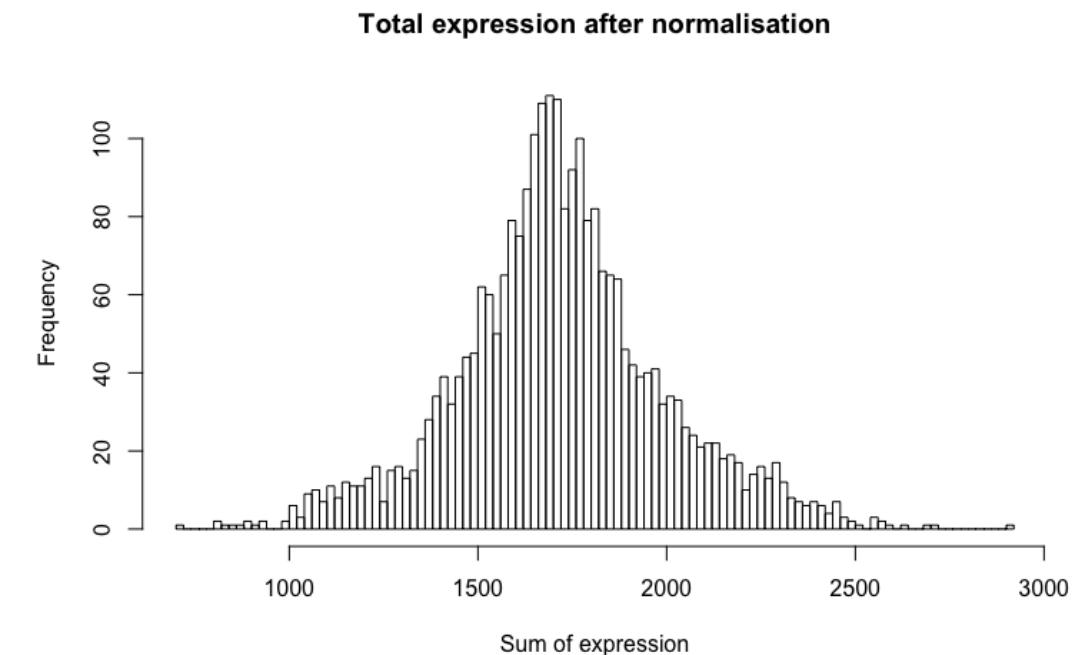
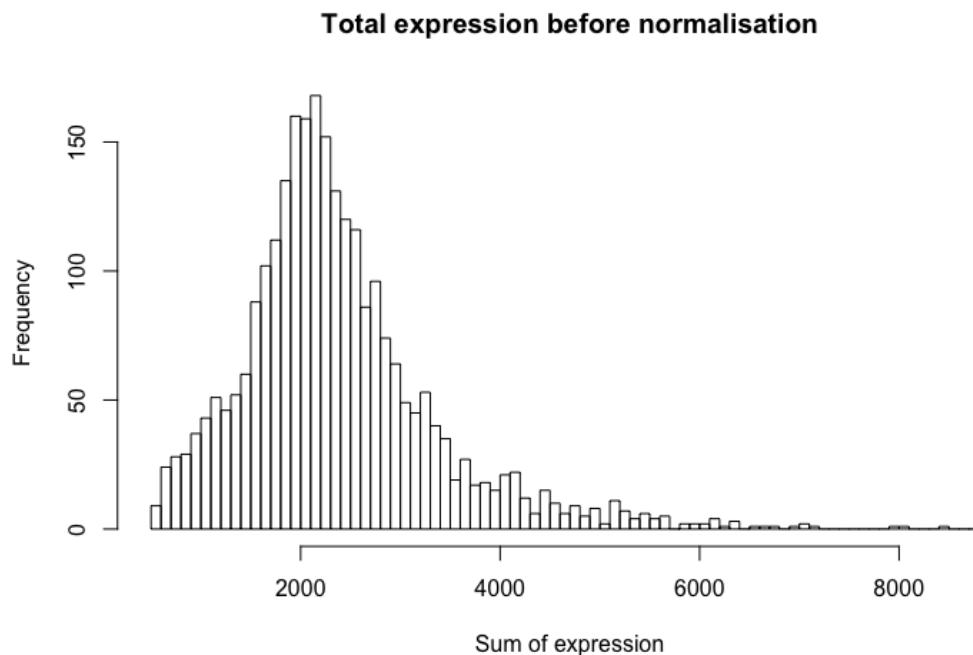
Single-cell RNA-Seq analysis: normalization

- 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



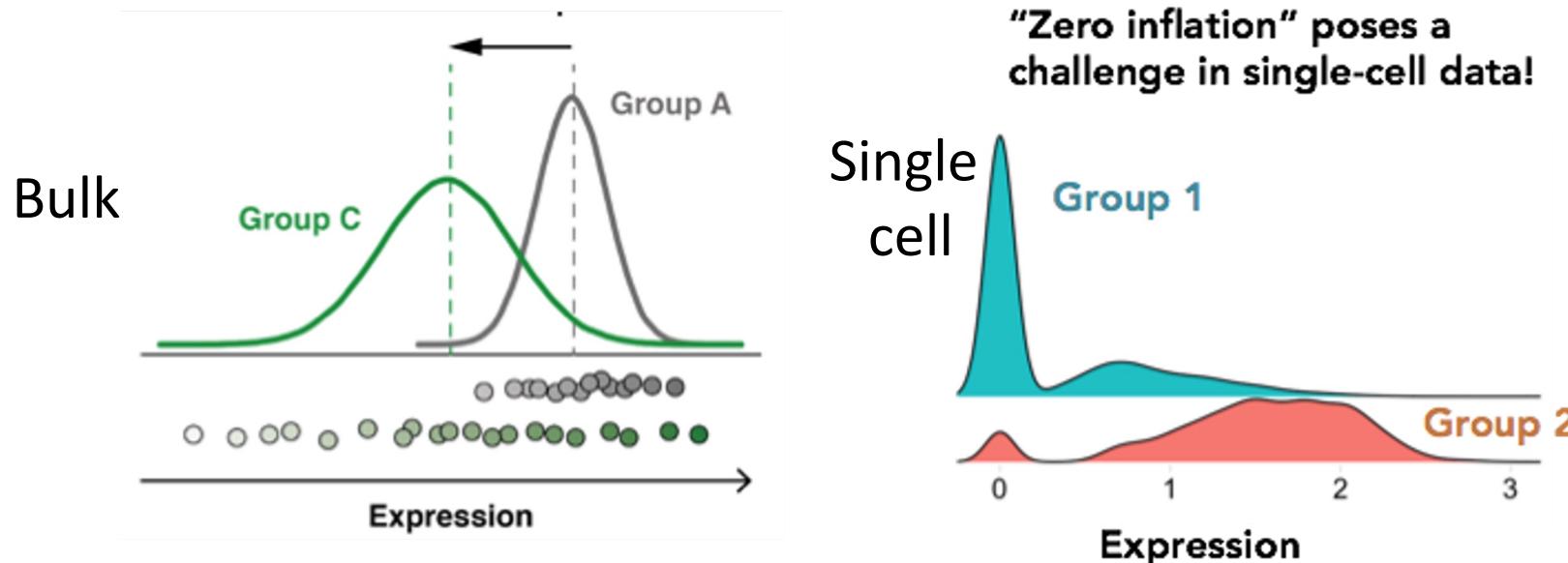
Single-cell RNA-Seq analysis: normalization

- 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)

Seurat and scanpy: single cell analysis toolkits



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)



SEURAT

R toolkit for single cell genomics

<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

Running analysis function

|=====| 100%

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
<i>Tcf7</i>	3	3	0	0	0	0	...	4
<i>Bach2</i>	2	4	1	0	0	0	...	2
<i>Prf1</i>	1	0	5	3	1	1	...	1
<i>Gzma</i>	0	0	3	1	0	0	...	0
<i>Pdcd1</i>	0	1	1	0	4	6	...	0
<i>Eomes</i>	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
0	1	0	0	0	0	0	0
0	0	0	0	0	2	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	3

Dense matrices

Coordinate List

genes

1	2	1
2	3	2
3	6	3

Sparse matrices

A few single-cell resources

Comprehensive list of single-cell resources

<https://github.com/seandavi/awesome-single-cell>

<https://www.scrna-tools.org/>

Computational packages for single-cell analysis

<https://satijalab.org/seurat/>

<https://scanpy.readthedocs.io/>

<http://bioconductor.org/packages-devel/workflows/html/simpleSingleCell.html>

<https://scverse.org/>

eLife Commentary on the Human Cell Atlas (HCA) <https://elifesciences.org/articles/27041>

Nature Commentary on the HCA <https://www.nature.com/news/the-human-cell-atlas-from-vision-to-reality-1.22854>

Online courses

<https://www.singlecellcourse.org/>

<https://lmweber.org/OSTA-book/>

<https://www.sc-best-practices.org/>

<https://bioconductor.org/books/release/OSCA/>

<https://uppsala.instructure.com/courses/52011>

Single cell data repositories

<http://jinglebells.bgu.ac.il/>

<https://www.nxn.se/single-cell-studies/gui>

<https://data.humancellatlas.org/>

<https://tabula-sapiens-portal.ds.czbiophub.org/>

<https://tabula-muris.ds.czbiophub.org/>