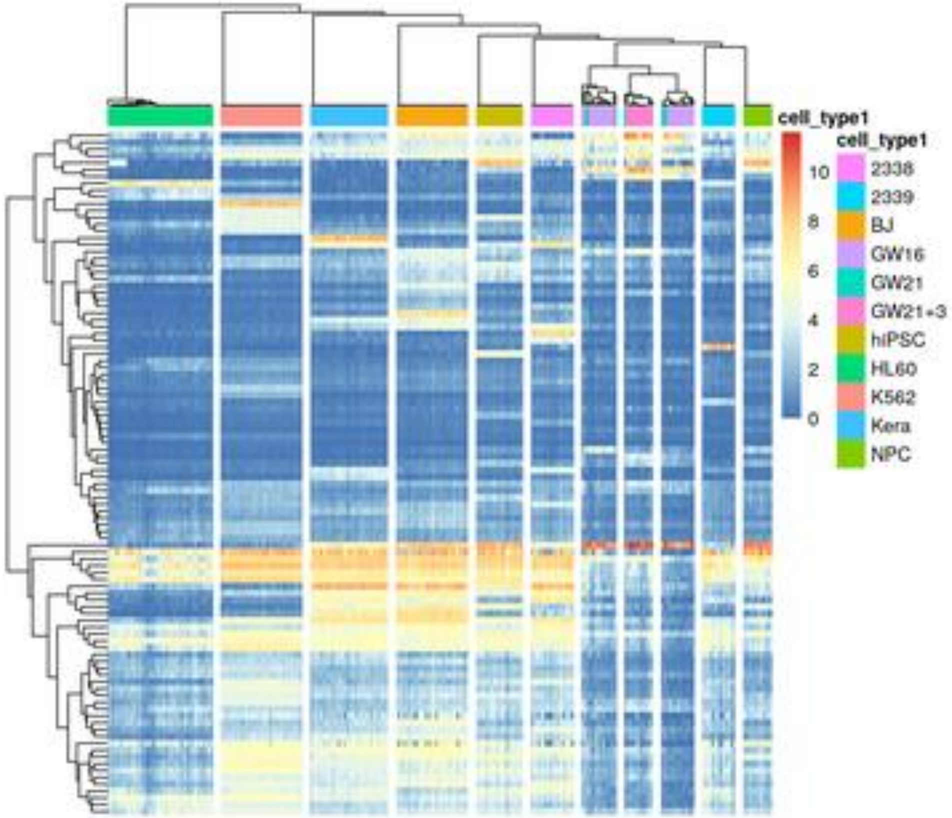
Quality control for scRNAseq data

Orr Ashenberg, Jacques Serizay, Fabrício Almeida-Silva

November, 2024

Physalia
Courses



Lecture topics

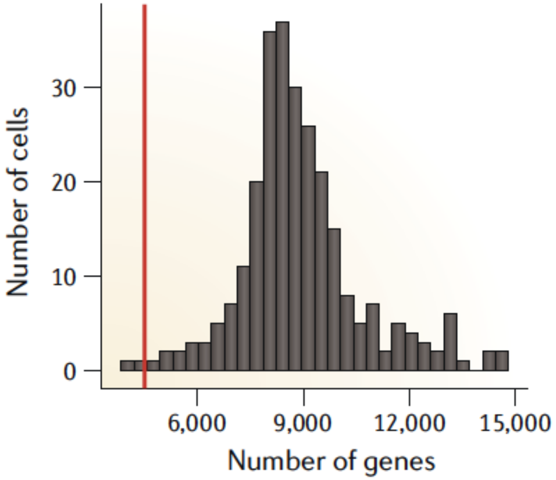
- Interacting with Seurat objects

- Quality control, normalization, and feature selection starting from raw count or expression matrices

- 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

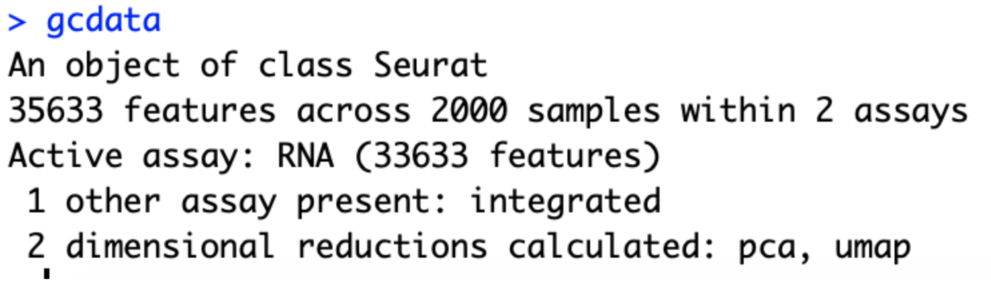
Kiselev C. at *al.* (2019) *Nature Reviews Genetics*. 20: 273-282.

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)

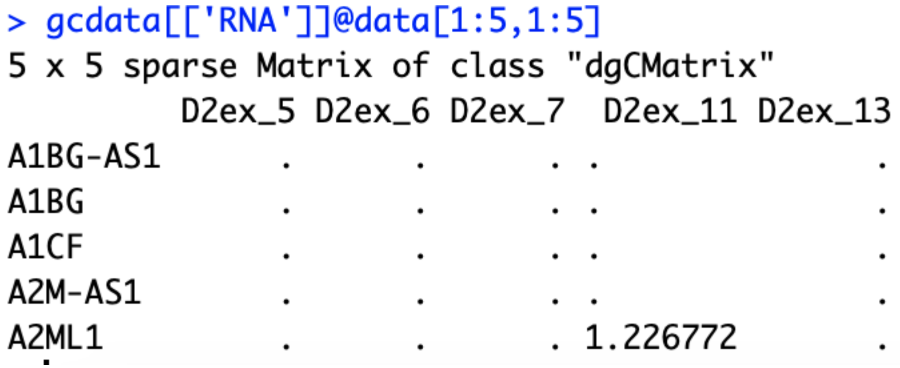
https://github.com/satijalab/seurat/wiki
https://satijalab.org/seurat/essential_commands.html

Interacting with Seurat objects



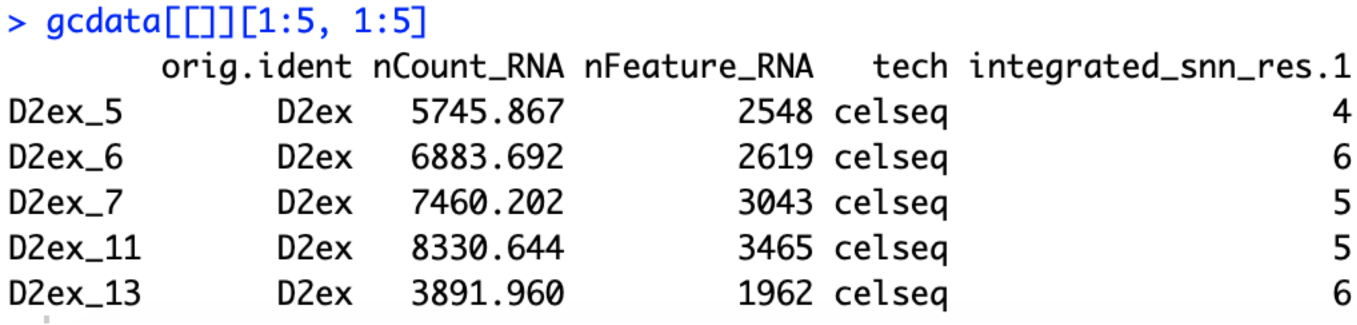
```
> gcdata
An object of class Seurat
35633 features across 2000 samples within 2 assays
Active assay: RNA (33633 features)
 1 other assay present: integrated
 2 dimensional reductions calculated: pca, umap
```

Seurat object



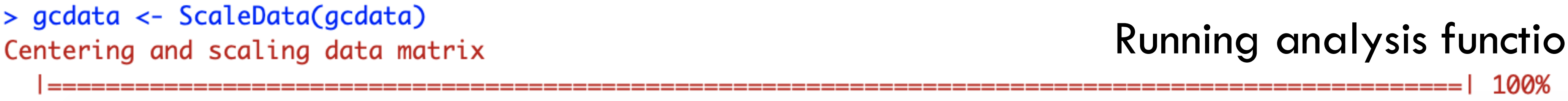
```
> gcdata[['RNA']]@data[1:5,1:5]
5 x 5 sparse Matrix of class "dgCMatrix"
         D2ex_5 D2ex_6 D2ex_7  D2ex_11 D2ex_13
A1BG-AS1      .      .      . .              .
A1BG          .      .      . .              .
A1CF          .      .      . .              .
A2M-AS1       .      .      . .    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



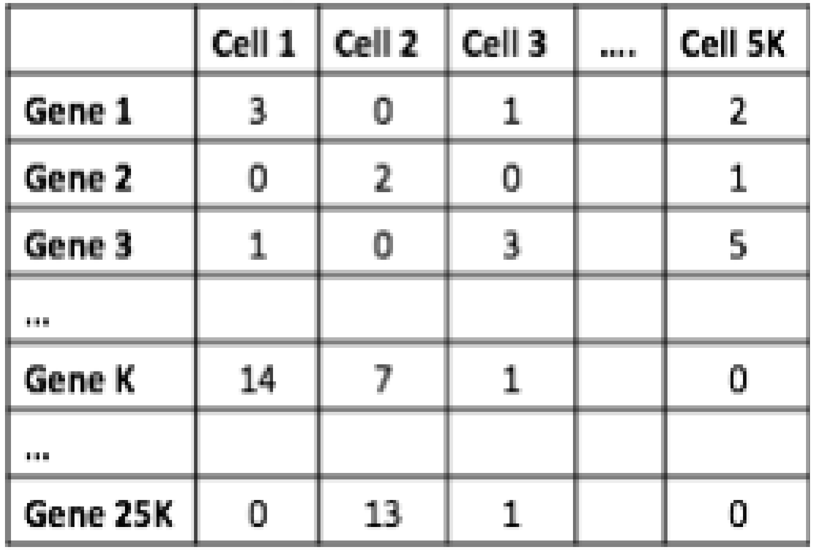
```
> gcdata <- ScaleData(gcdata)
Centering and scaling data matrix
  |=====================================================================================| 100%
```

Running analysis function

Loading data into a Seurat object

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

counts matrix



	Cell 1	Cell 2	Cell 3	....	Cell 5K
Gene 1	3	0	1		2
Gene 2	0	2	0		1
Gene 3	1	0	3		5
...					
Gene K	14	7	1		0
...					
Gene 25K	0	13	1		0

Storing counts data in dense vs sparse format

2D Arrays

cells

1	2	3	4	5	6	7	8	genes
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

Dense matrices

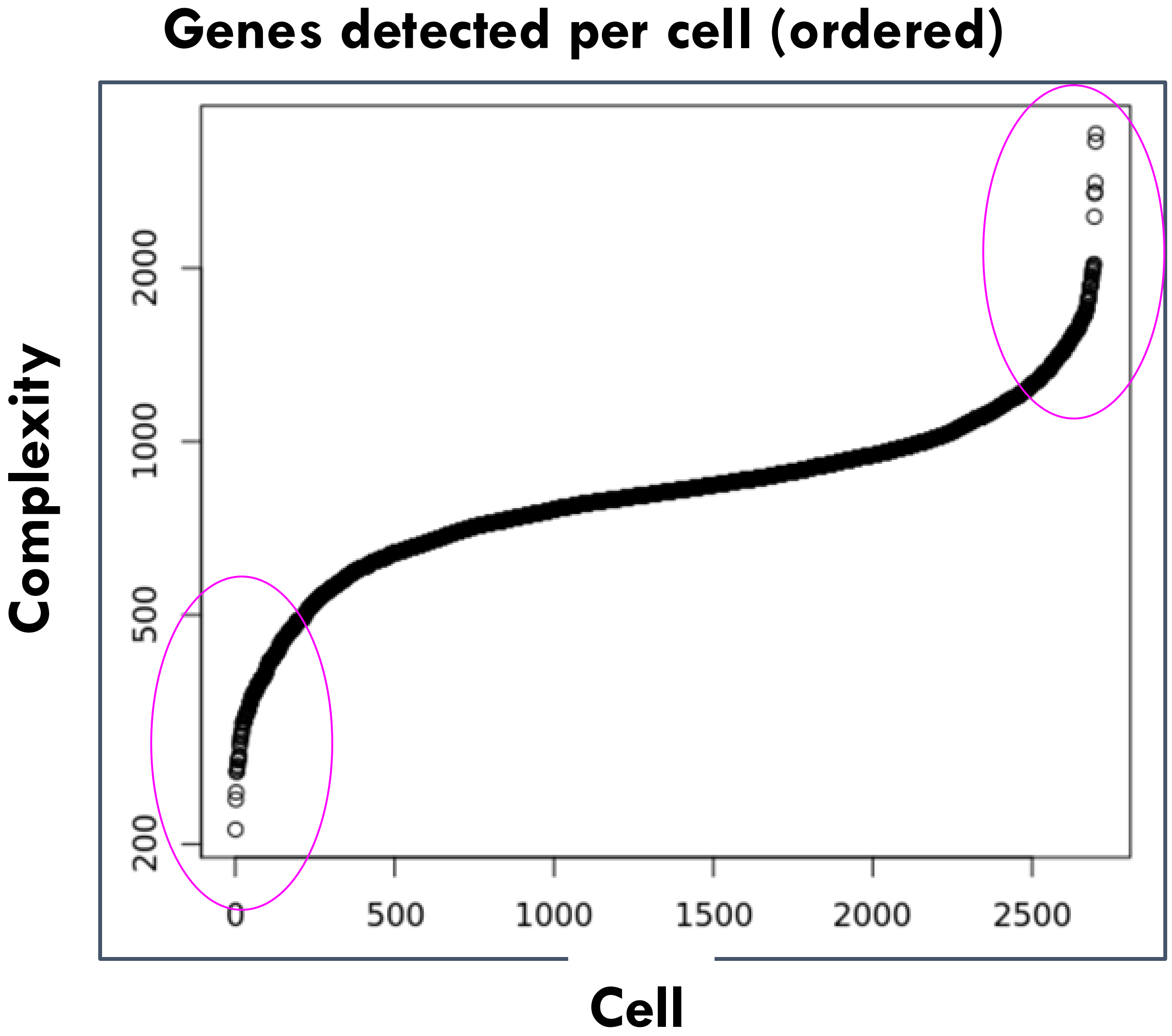
Coordinate List

2	2	1
6	3	2
8	6	3

Sparse matrices

There are many quality control filters for genes and cells

Complexity =
Number of genes
detected in a cell



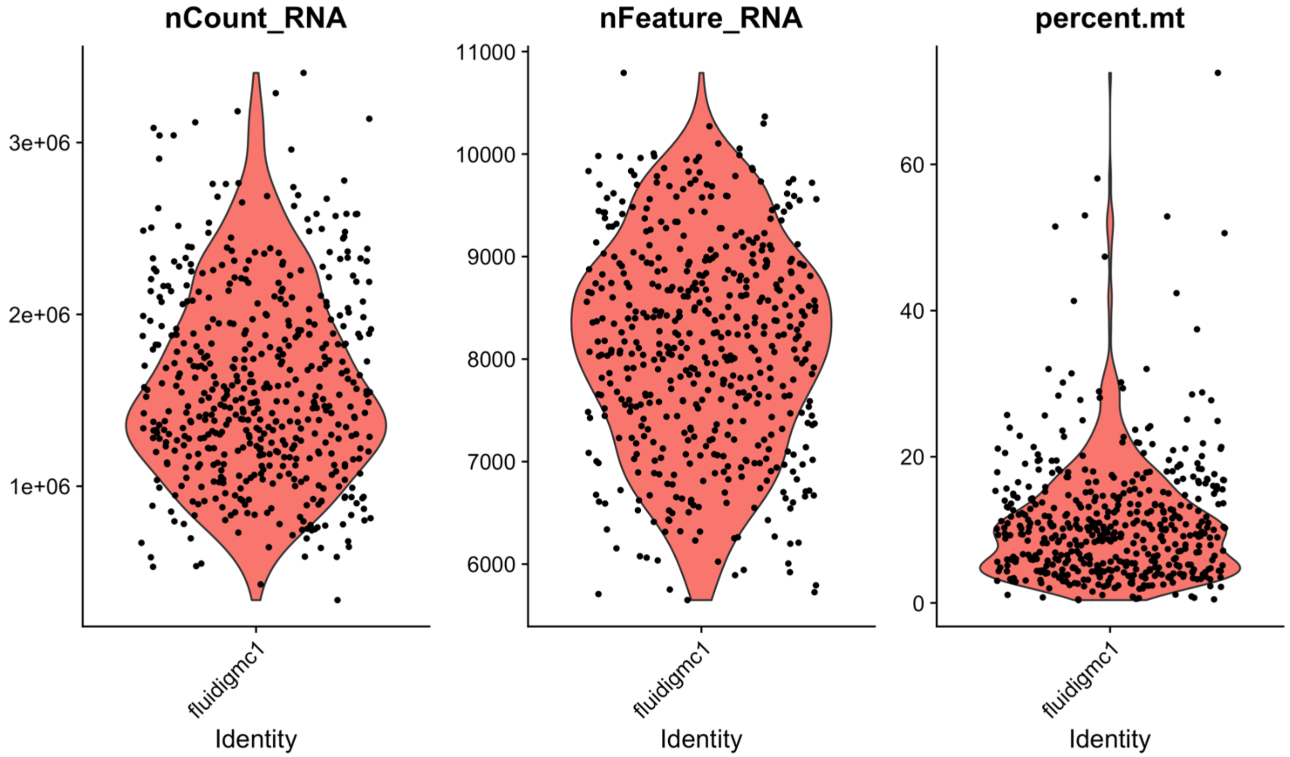
There are many quality control filters for genes and cells

- We filter cells based on technical or biological parameters.

# reads or UMIs detected in a cell

# genes detected in a cell

% mito expression in a cell

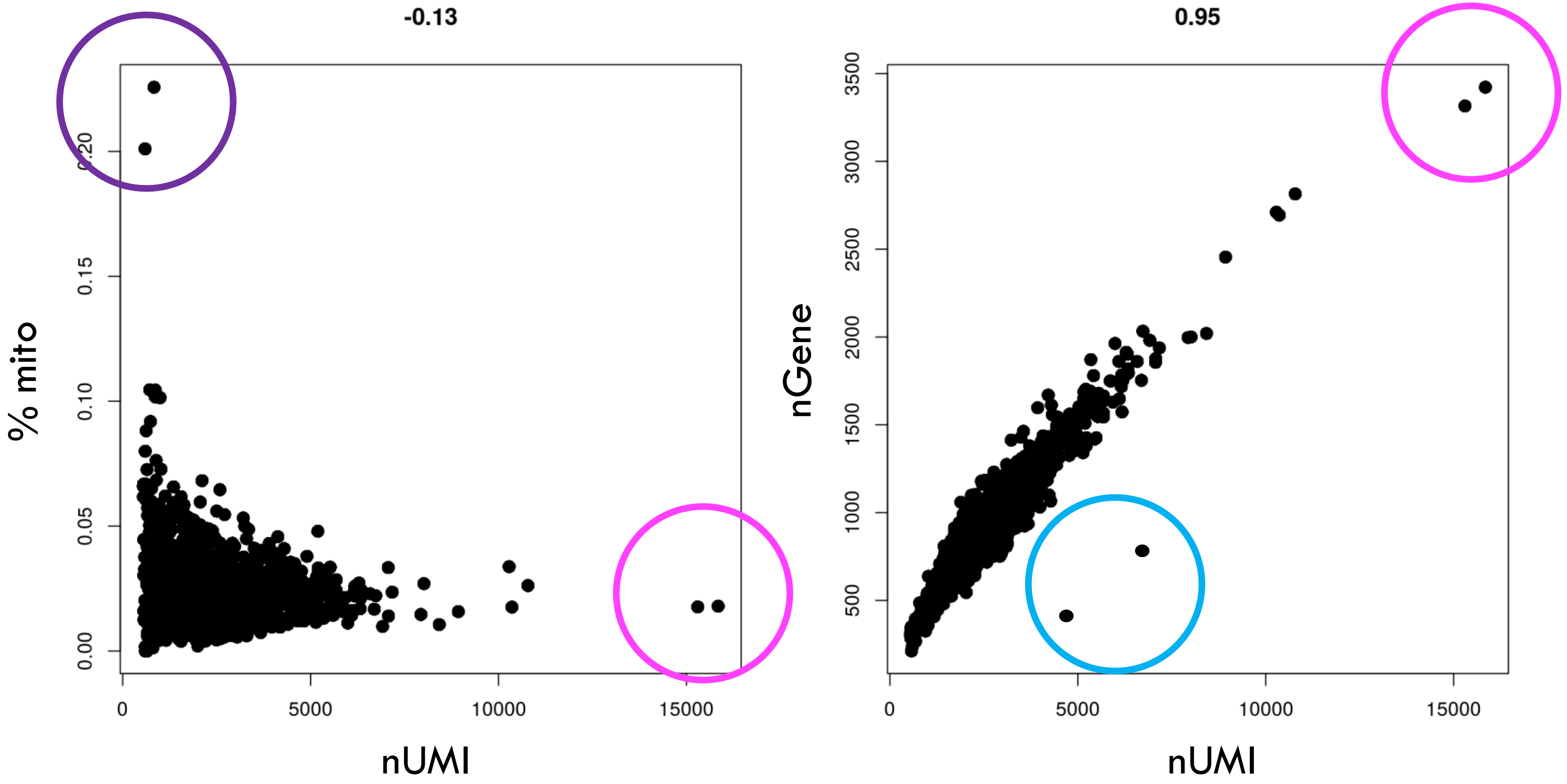


Filtering with combinations of quality control filters

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

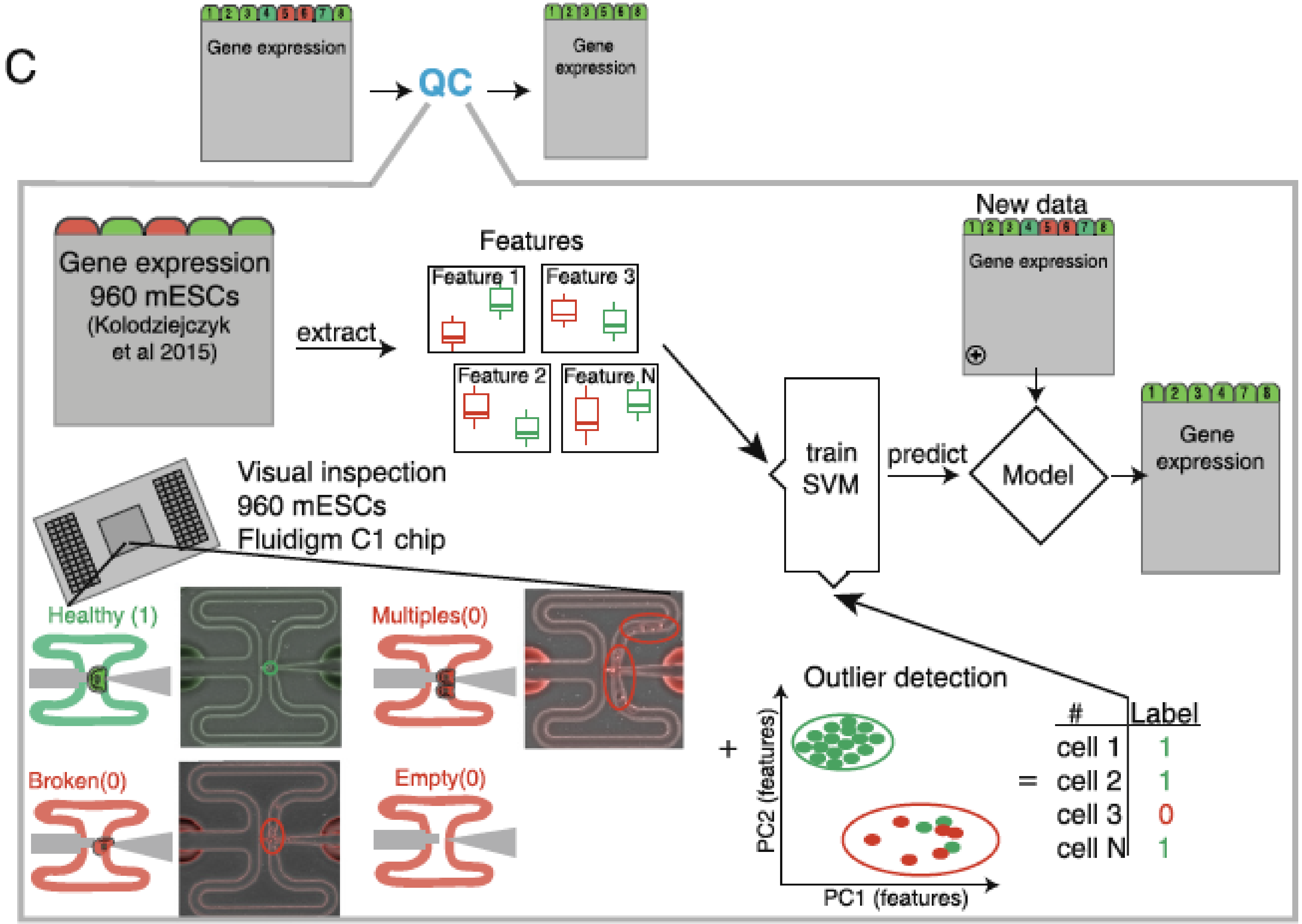


Appropriate quality control filters vary with platform and cell types

- 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

A classifier for low-quality cells

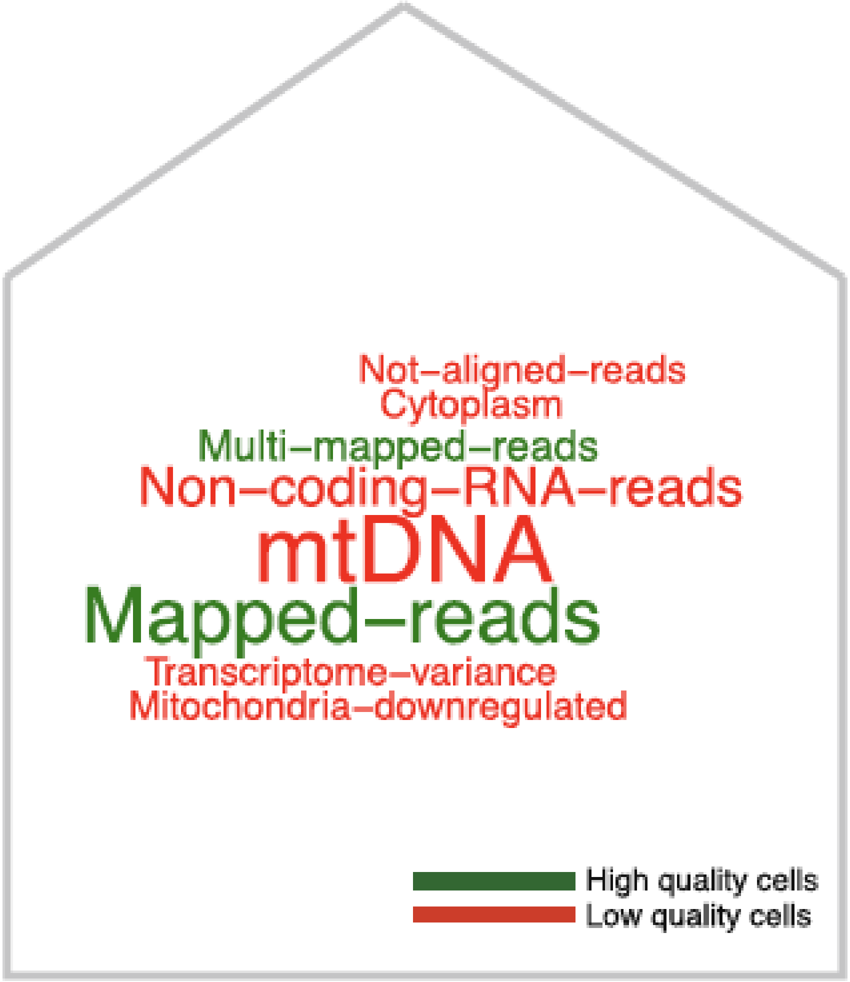


"The pipeline takes advantage of a highly-curated set of generic features that are incorporated into a machine learning algorithm to identify low quality cells."

Ilicic T. et al. (2016) *Genome Biology* 17: 29.

A classifier for low-quality cells

SVM feature weights



What are caveats to this approach?

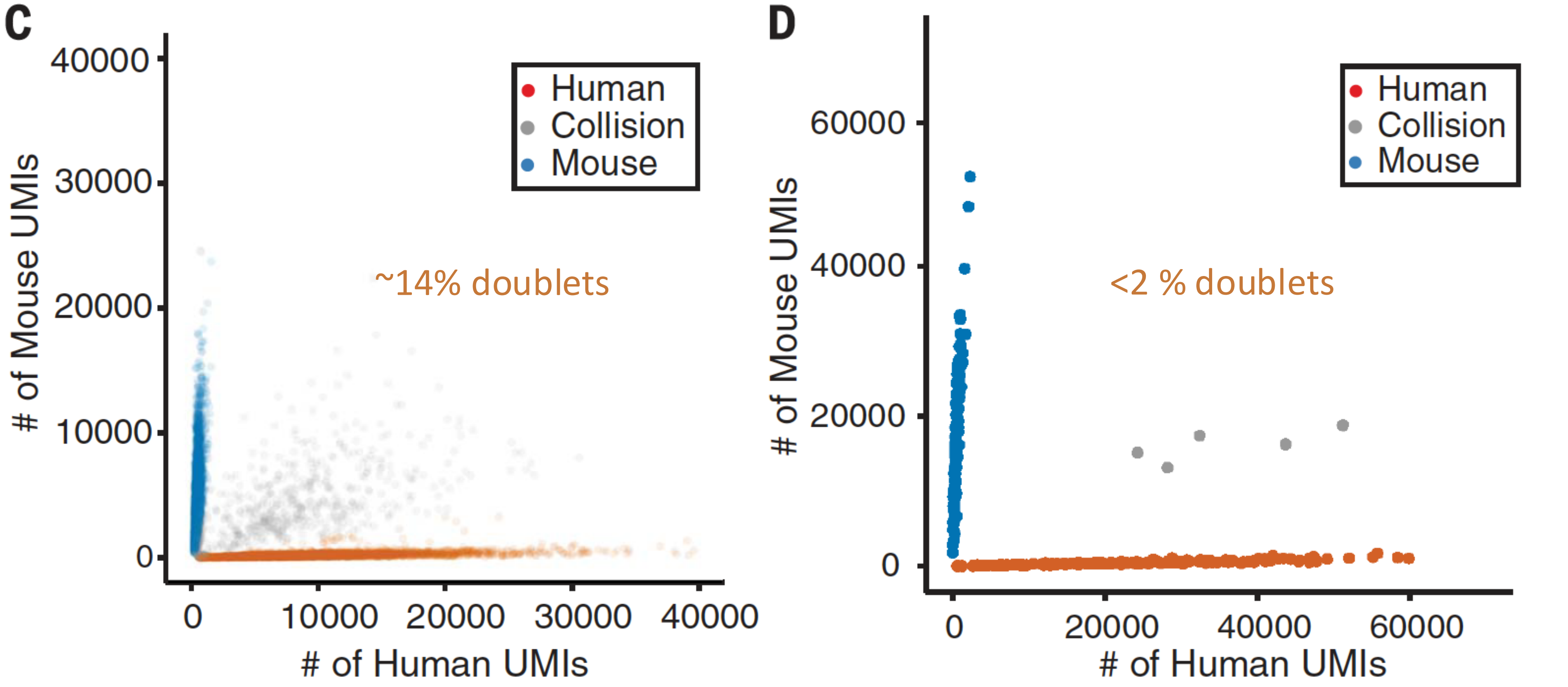
Ilicic T. et al. (2016) *Genome Biology* 17: 29.

Other quality control filters for genes and cells

- Doublets
  - number of genes ⬆
  - number of UMIs ⬆
  - percentage of mitochondrial gene expression ≈
- Ambient RNA and empty droplets
  - number of genes ⬇
  - number of UMIs ⬇
  - percentage of mitochondrial gene expression ⬆
- Barcode swapping

Cell doublets can be misleading

Because of the setup, it is possible that two or more cells can enter the same droplet. Studies estimate doublet frequency through a “mixed-species” experiment



~14% doublets

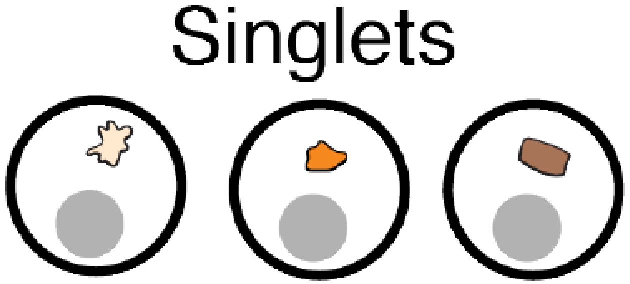
<2 % doublets

The doublet frequency is positively correlated with throughput

Slide courtesy of Karthik Shekhar

Detecting cell doublets with Scrublet

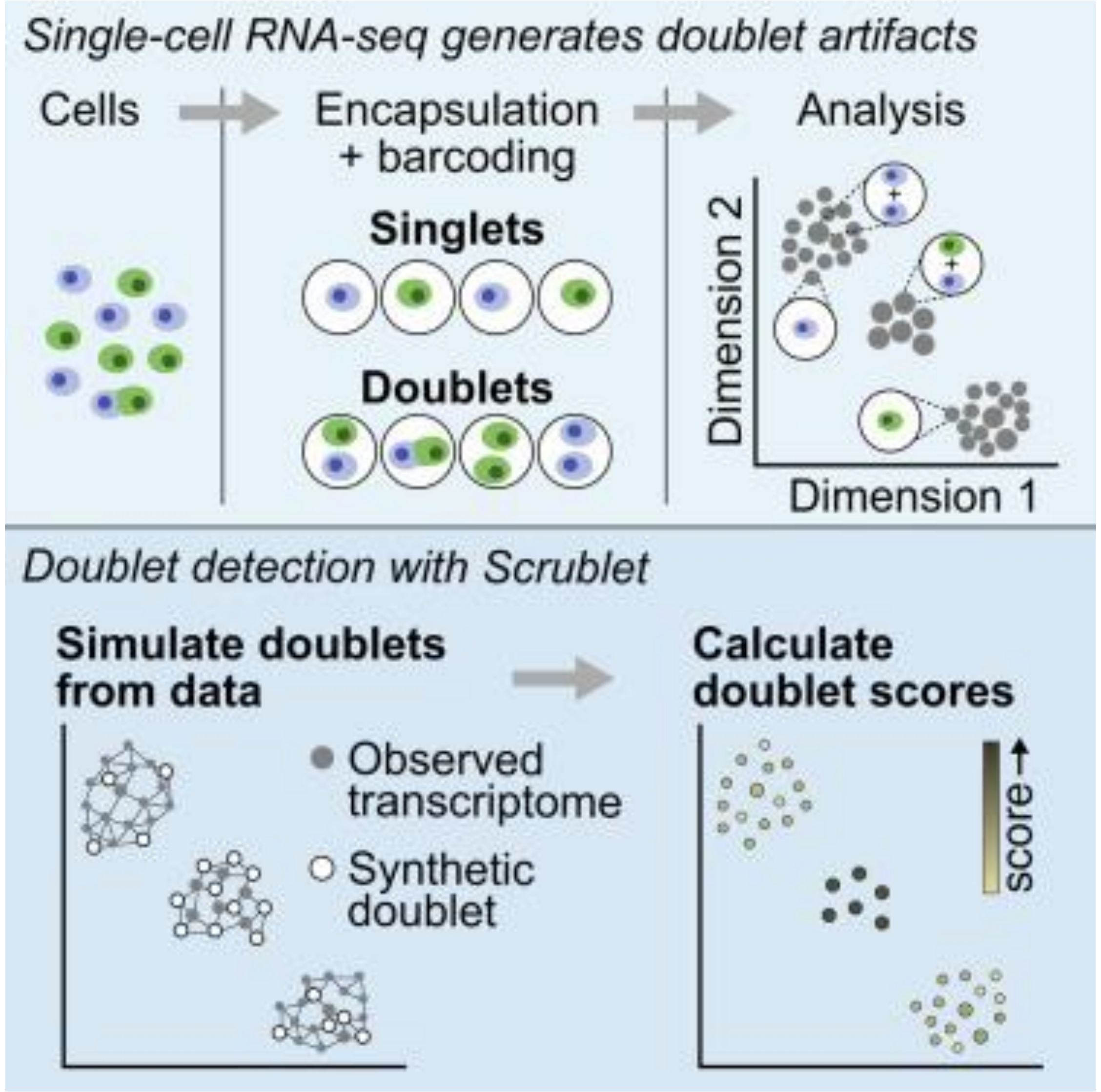
Scrublet (Single-Cell Remover of Doublets)



🤗

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

Detecting cell doublets with Scrublet


Single-cell RNA-seq generates doublet artifacts
Cells
Encapsulation + barcoding
Analysis
Singlets
Doublets
Dimension 2
Dimension 1
Doublet detection with Scrublet
Simulate doublets from data
● Observed transcriptome
○ Synthetic doublet
Calculate doublet scores
score

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

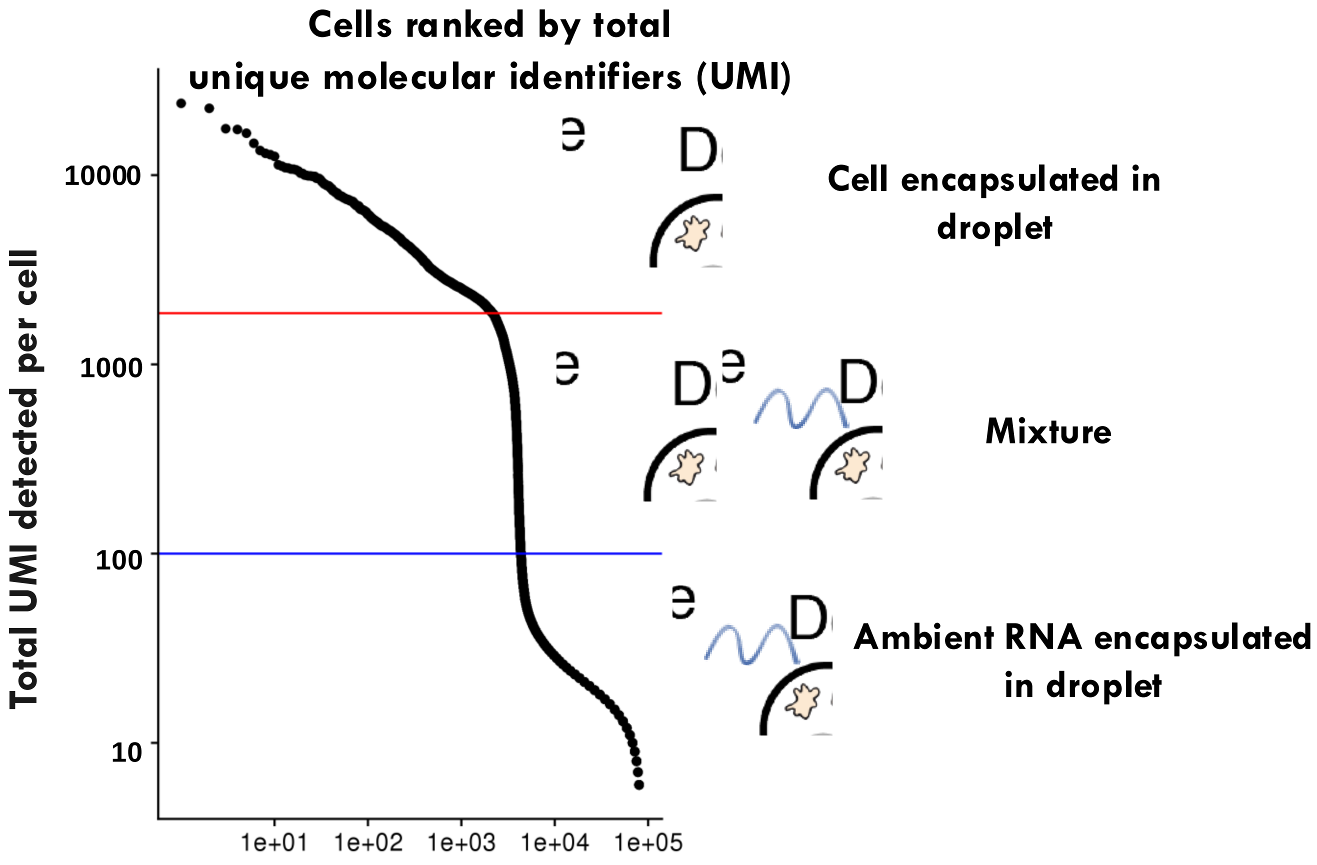
Detecting empty drops containing ambient RNA – manual

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

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

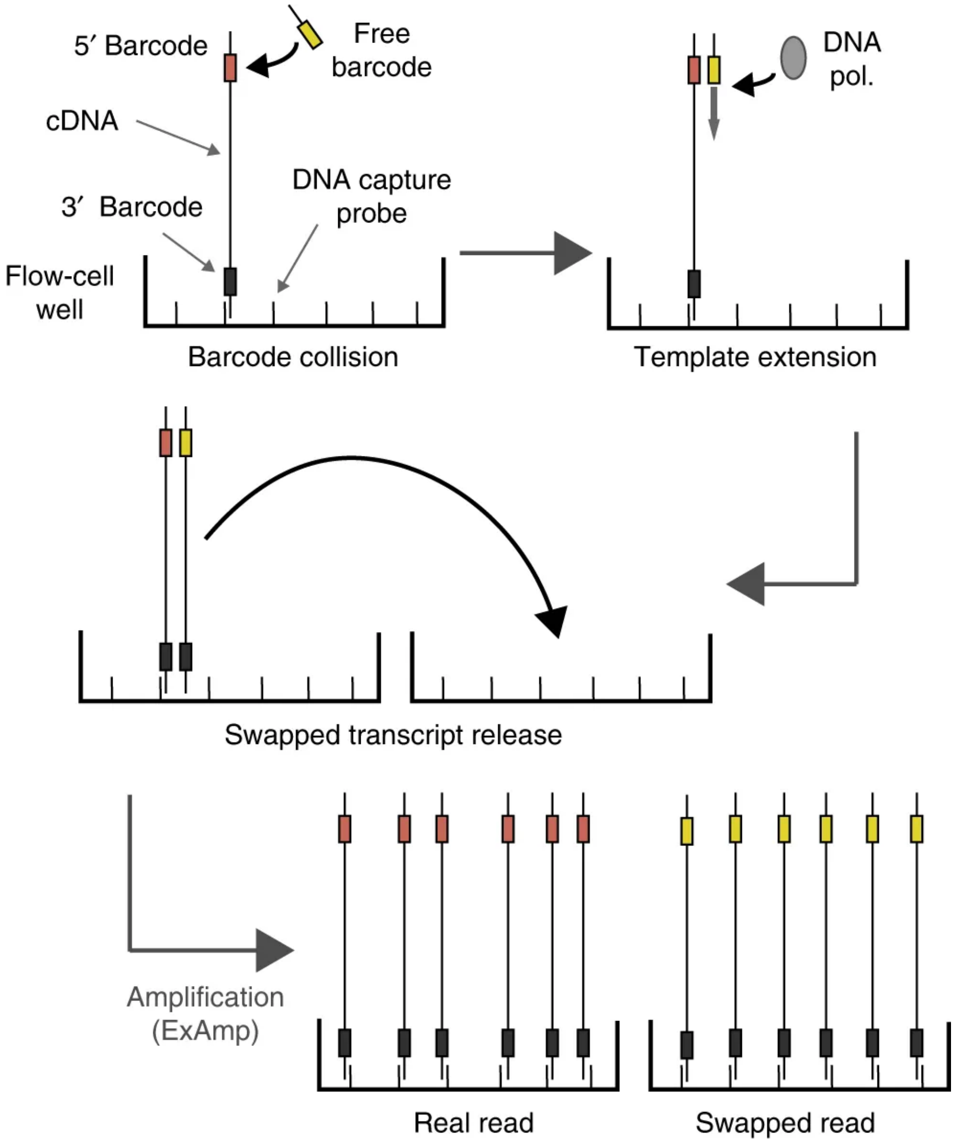
Detecting empty drops containing ambient RNA – automatic

EmptyDrops (distinguish cells from empty droplets)



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

Detecting barcode swapping in multiplexed samples



"Barcode swapping is a phenomenon that occurs upon multiplexing samples on the Illumina 4000 sequencer. Molecules from one sample are incorrectly labelled with *sample* barcodes from another sample, resulting in their misassignment upon demultiplexing."

"Specifically, we considered molecules across **multiplexed samples that contain the same combination of unique molecular identifier, cell barcode, and aligned gene**."

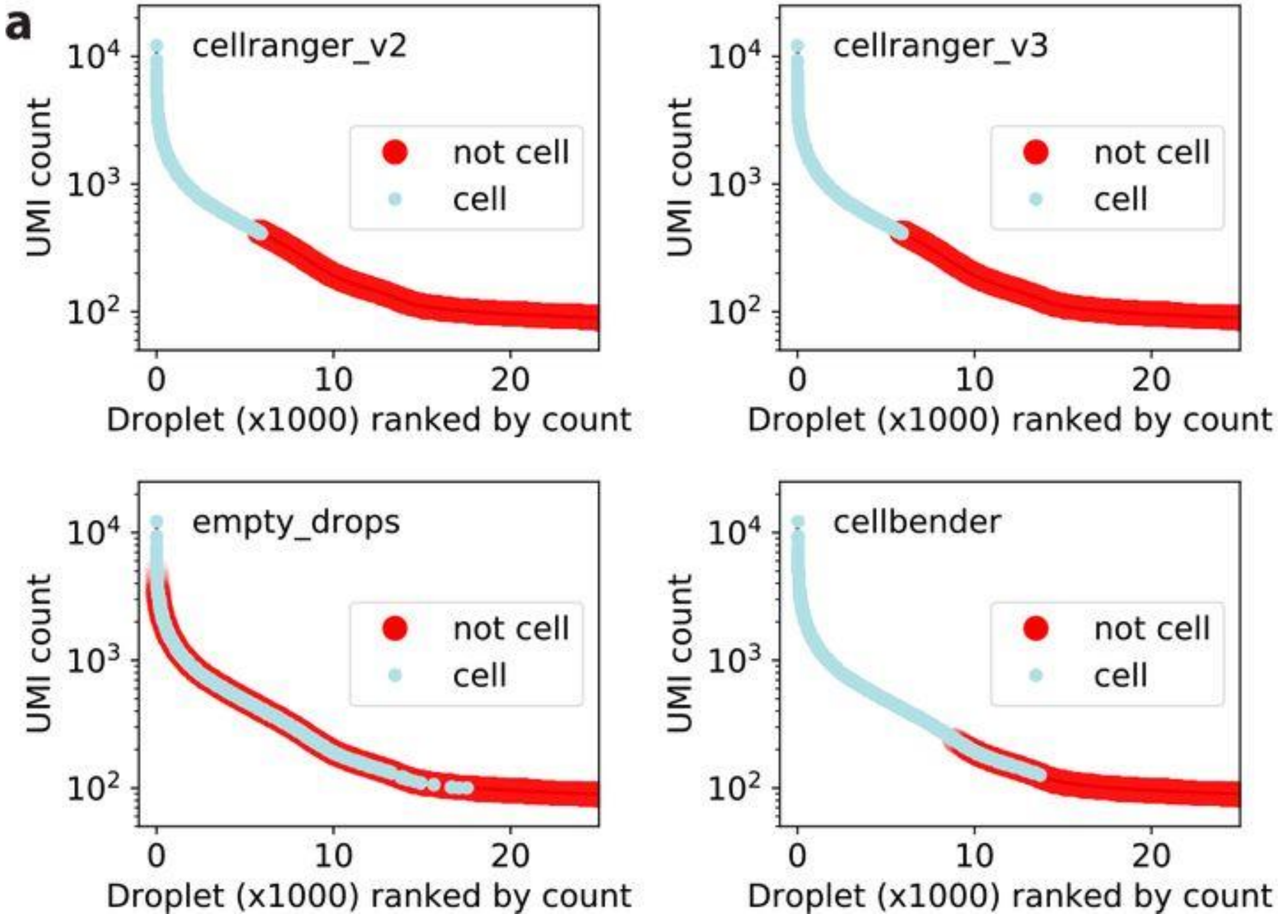
Griffiths, J. A. et al. *Nature* Communications (2018).9: 2667.

https://bioconductor.org/packages/devel/bioc/vignettes/DropletUtils/inst/doc/DropletUtils.html#removing-the-effect-of-barcode-swapping

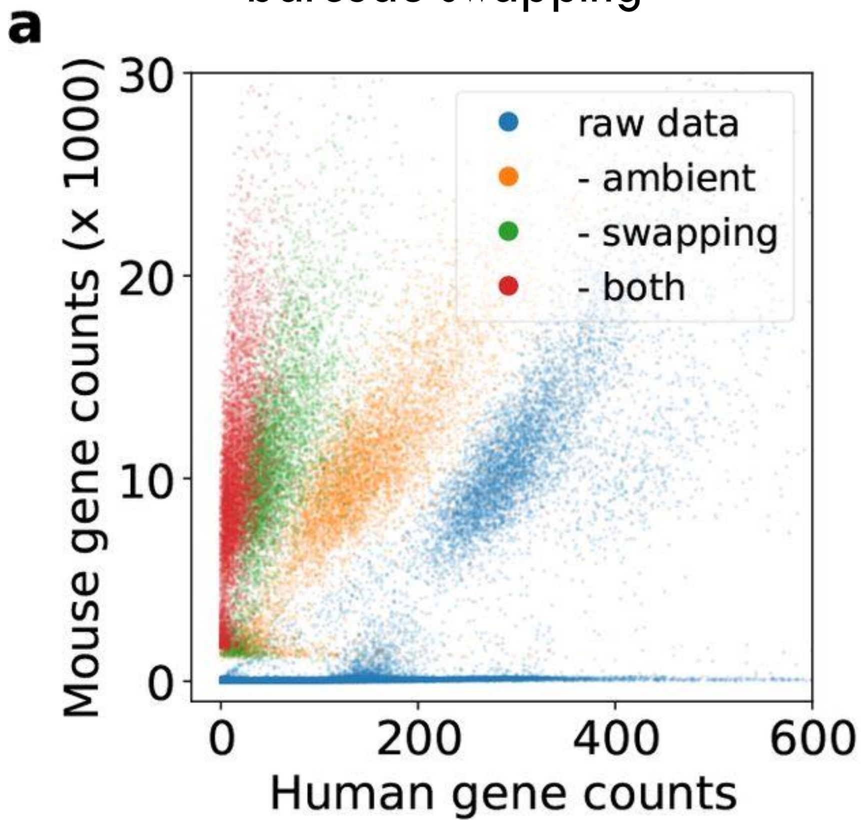
Tool to detect empty drops, correct ambient RNA and barcode swapping

CellBender remove ambient background and barcode swapping via deep learning

Detecting empty droplets



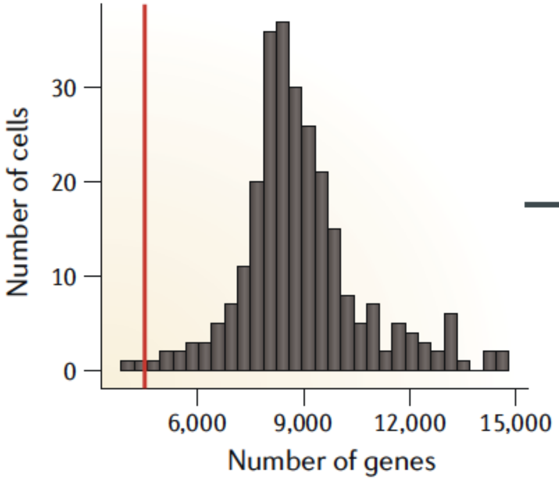
Correcting ambient RNA and barcode swapping



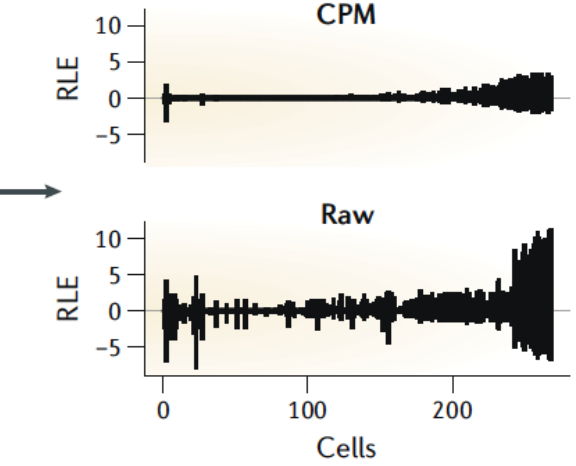
Fleming et al. *bioRxiv* (2019)

Determining cell type, state, and function

Quality control



Normalization



Feature selection

Dimensional reduction

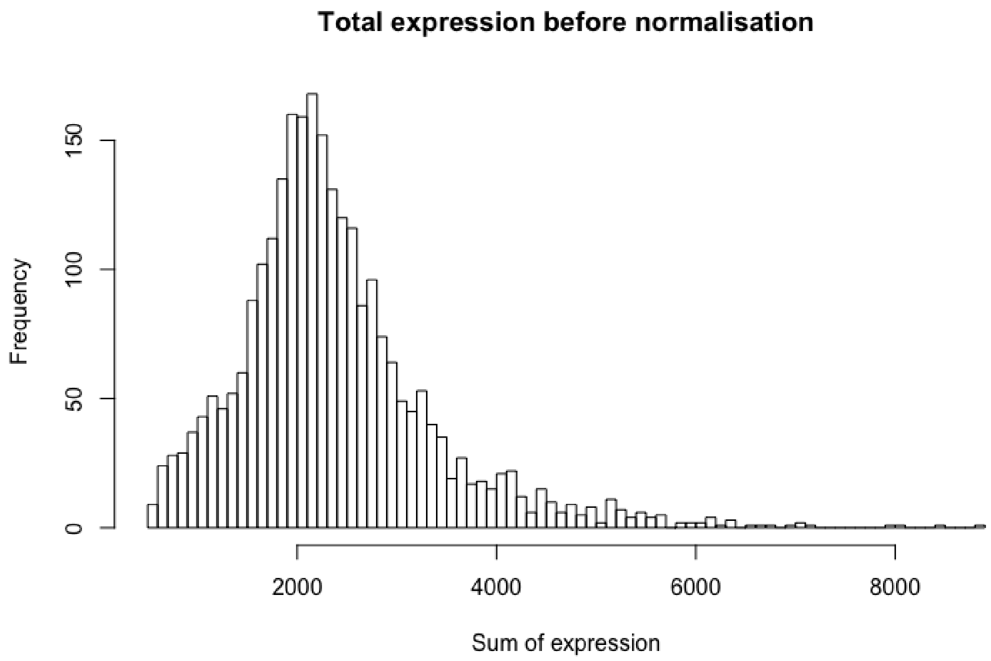
Cell-cell distances

Unsupervised clustering

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

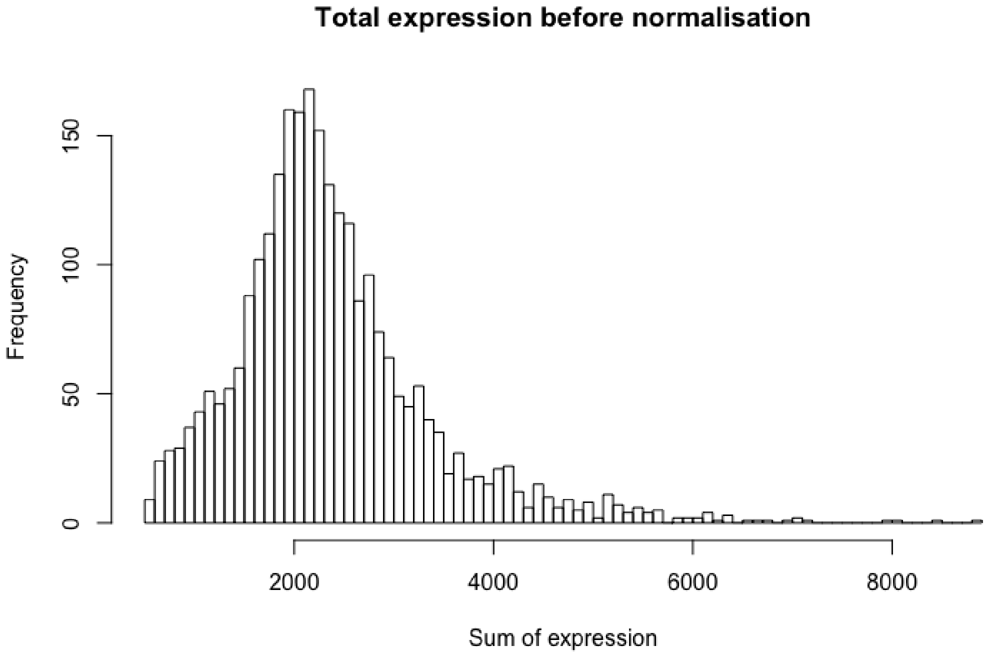
Normalizing gene expression in each cell

- 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

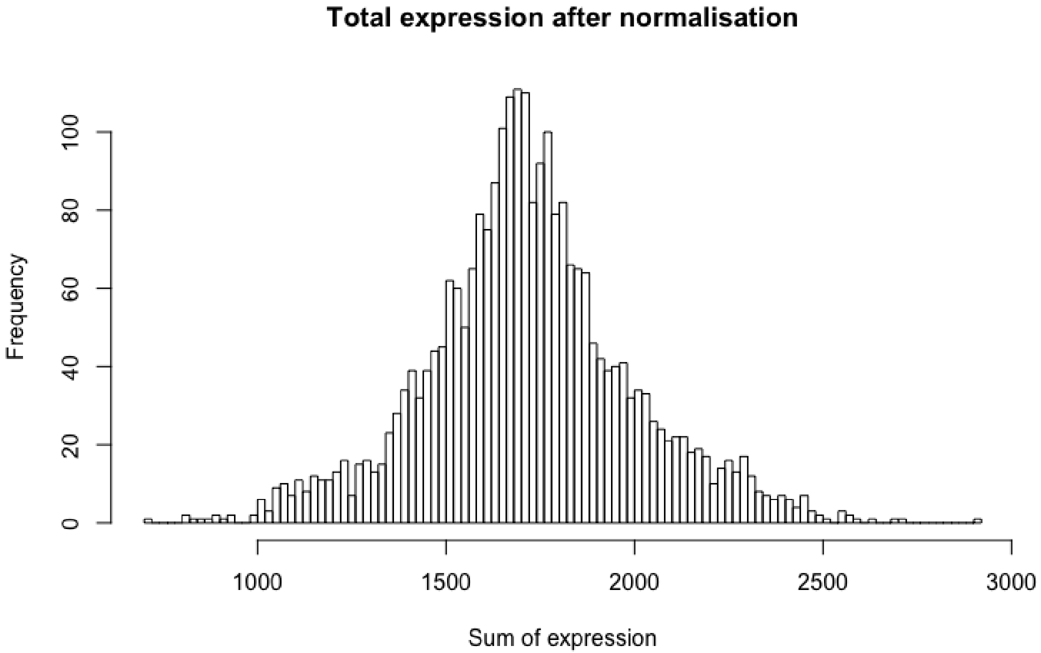


Normalizing gene expression in each cell

- 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-transformed.



Total expression before normalisation

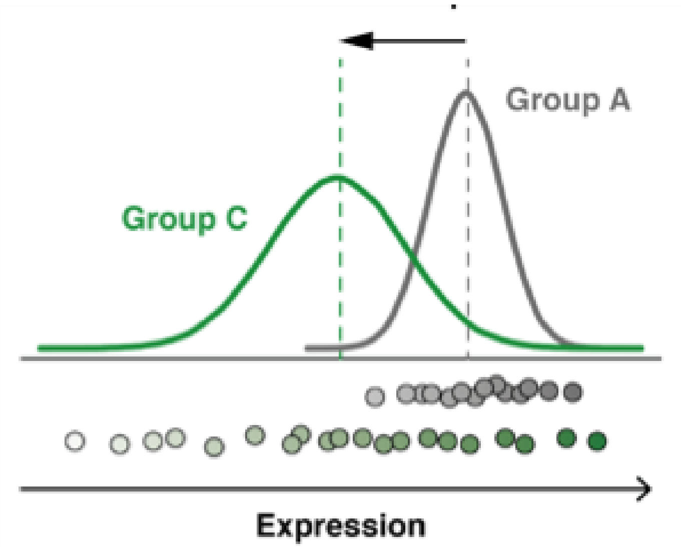


Total expression after normalisation

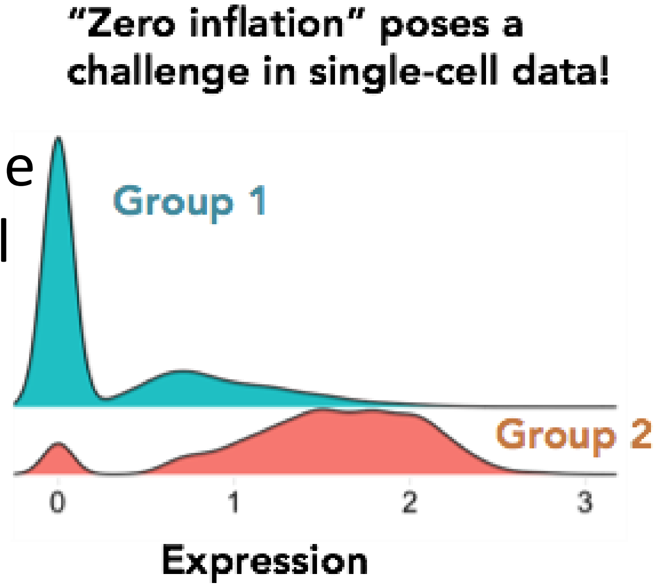
Is standard normalization appropriate?

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

Bulk

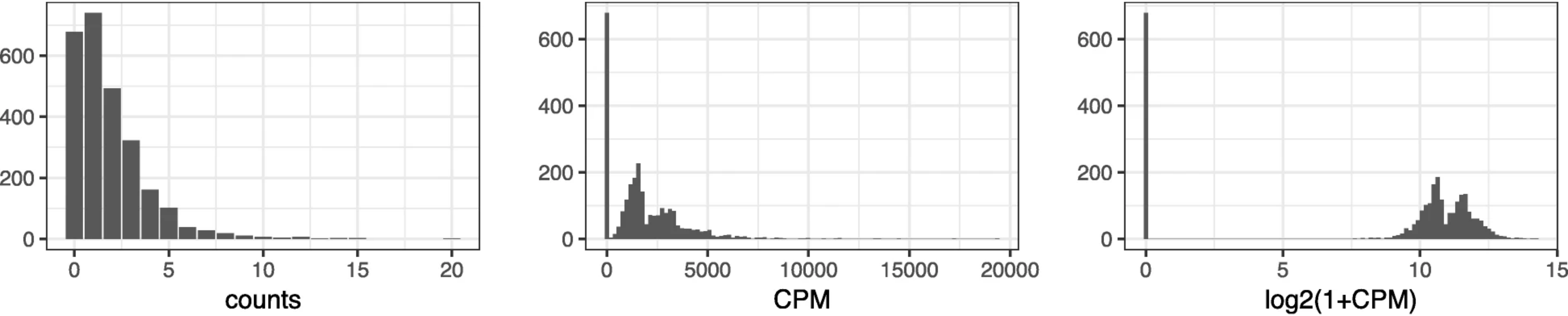


Single cell



- "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)

Is standard normalization appropriate?



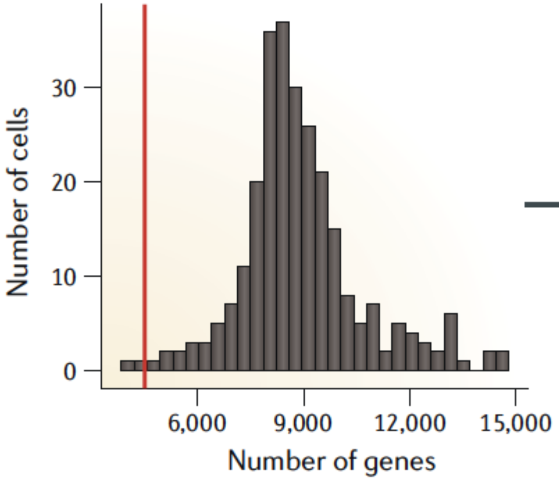
(a) UMI counts (b) counts per million (CPM) (c) log of CPM

Example of how current approaches to normalization and transformation artificially distort differences between zero and nonzero counts. **a** UMI count distribution for gene ENSG00000114391 in the monocytes biological replicates negative control dataset. **b** Counts per million (CPM) distribution for the exact same count data. **c** Distribution of $log_2(1+CPM)$ values for the exact same count data

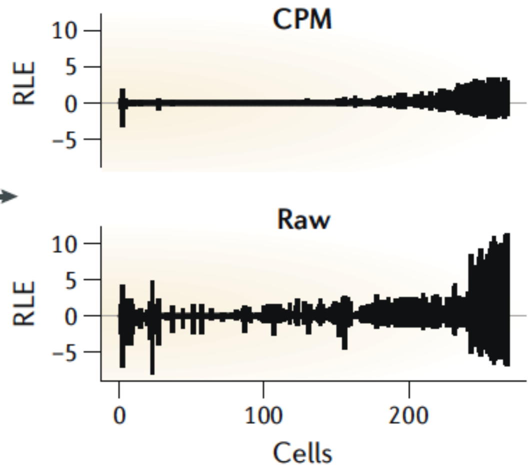
William Townes F. et al. Genome Biology (2019) 20: 295.

Determining cell type, state, and function

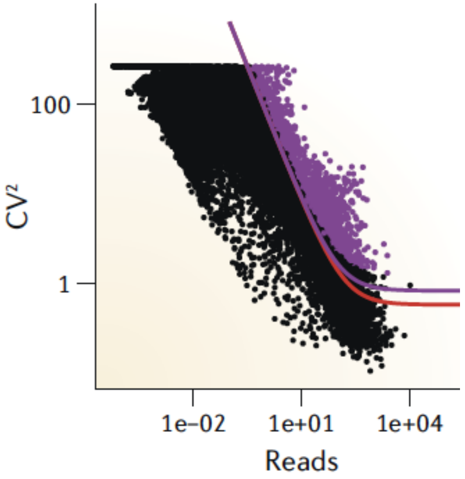
Quality control



Normalization



Feature selection



Dimensional reduction

Cell-cell distances

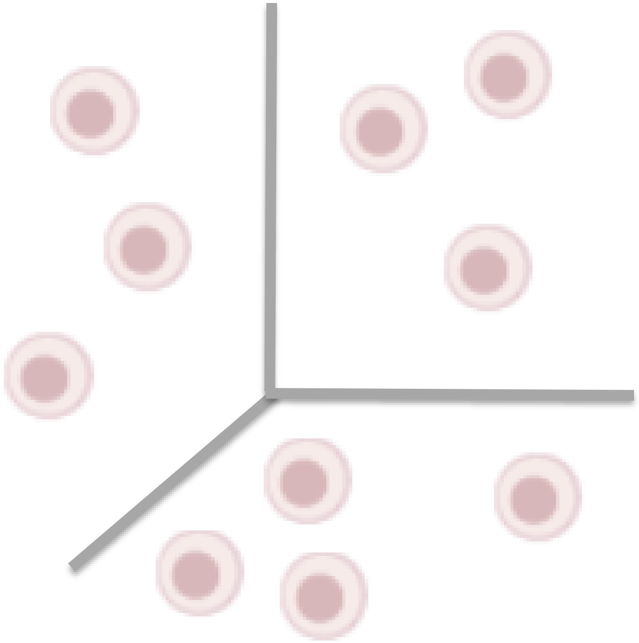
Unsupervised clustering

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

Identify highly variable genes

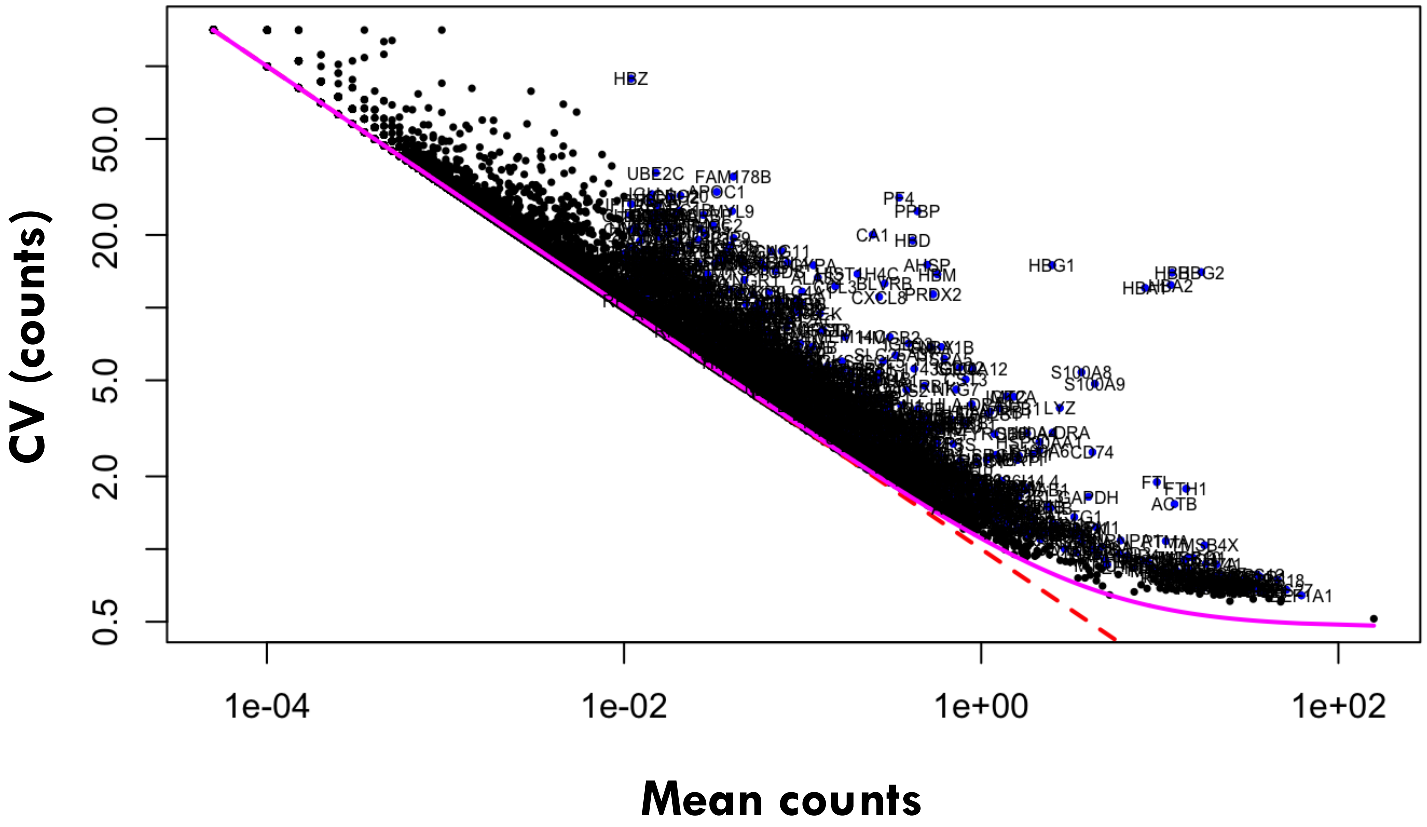
Cells are in ~20,000 dimensional space (one dimension for each gene)

- many genes are lowly detected or noisy measurements



- variable genes contain the biological signal we are interested in

Identify highly variable genes

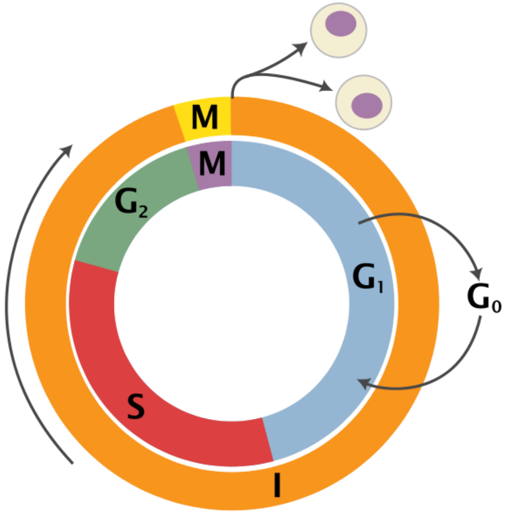


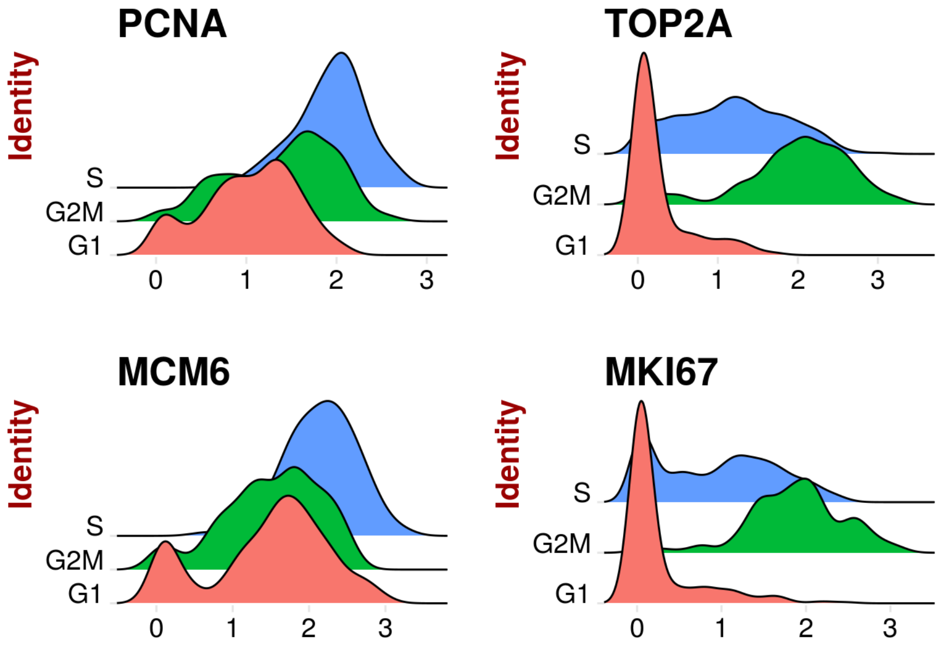
Find genes (features) that are outliers in a plot of mean of gene expression vs variance of gene expression

Calculating gene signatures

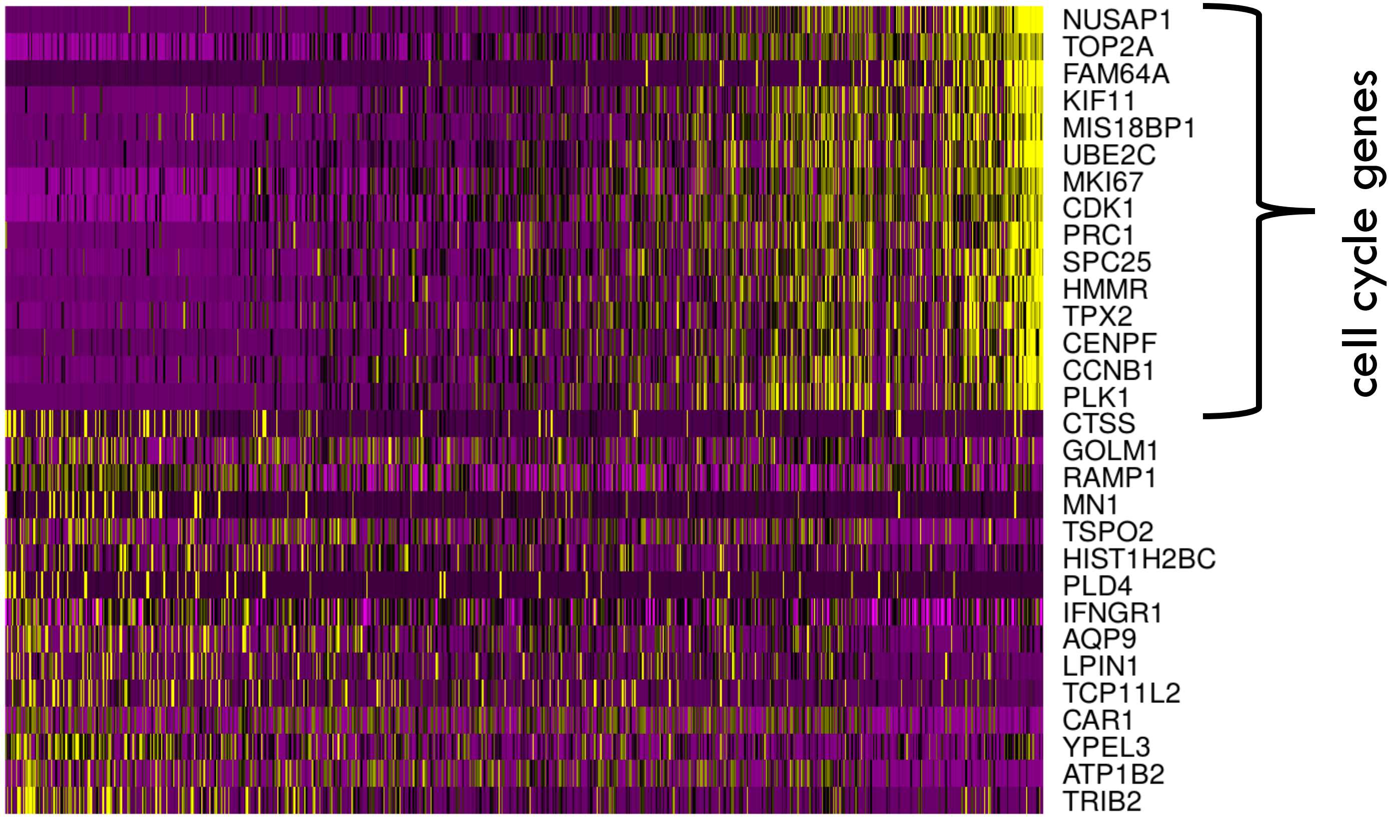
Relying on capturing a specific gene is not robust, but relying on a set of genes (signature) is much more stable!

Gene signature example: cell cycle markers





variability of individual genes


NUSAP1
TOP2A
FAM64A
KIF11
MIS18BP1
UBE2C
MKI67
CDK1
PRC1
SPC25
HMMR
TPX2
CENPF
CCNB1
PLK1
CTSS
GOLM1
RAMP1
MN1
TSPO2
HIST1H2BC
PLD4
IFNGR1
AQP9
LPIN1
TCP11L2
CAR1
YPEL3
ATP1B2
TRIB2
cell cycle genes