

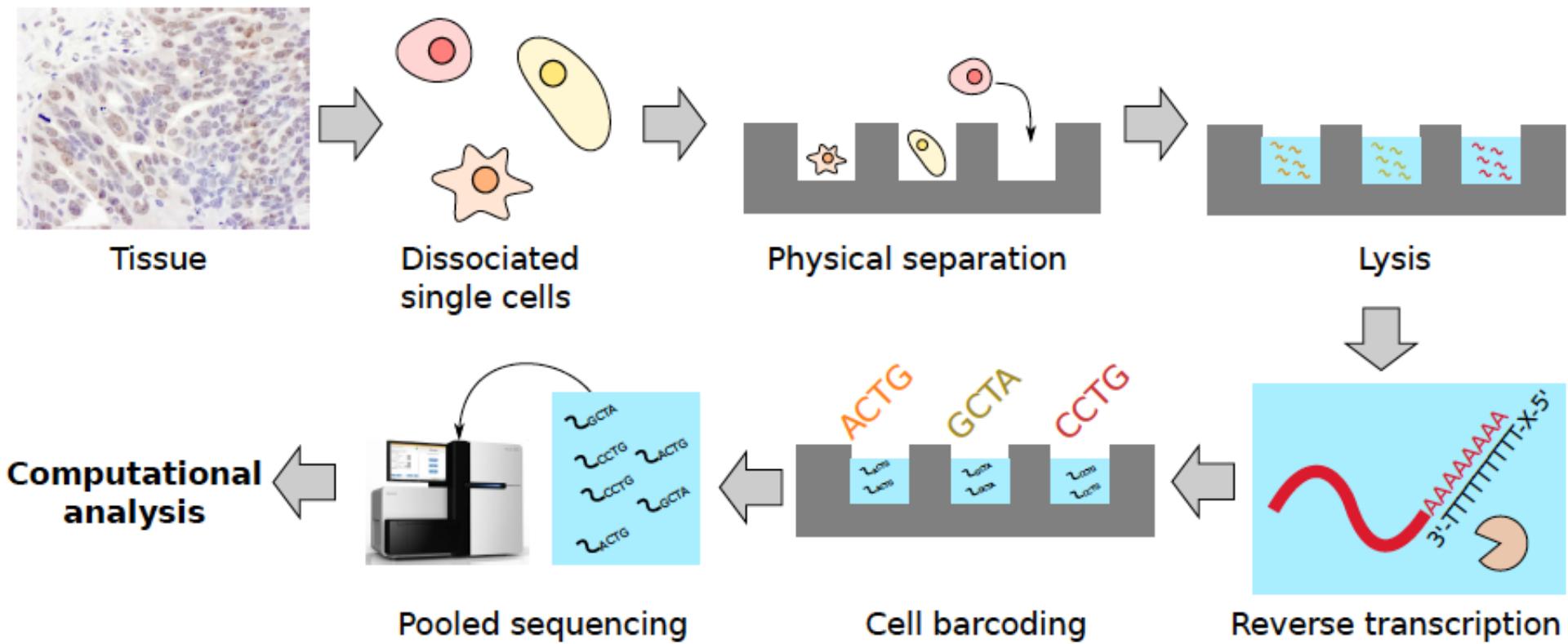
# Quality Control and Normalization of Single Cell RNA-seq Data

Ahmed Mahfouz

Department of Human Genetics, Leiden University Medical Center  
Pattern Recognition and Bioinformatics, TU Delft

[mahfouzlab.org](http://mahfouzlab.org)  
@ahmedElkoussy

# Single cell RNA-sequencing (scRNA-seq)

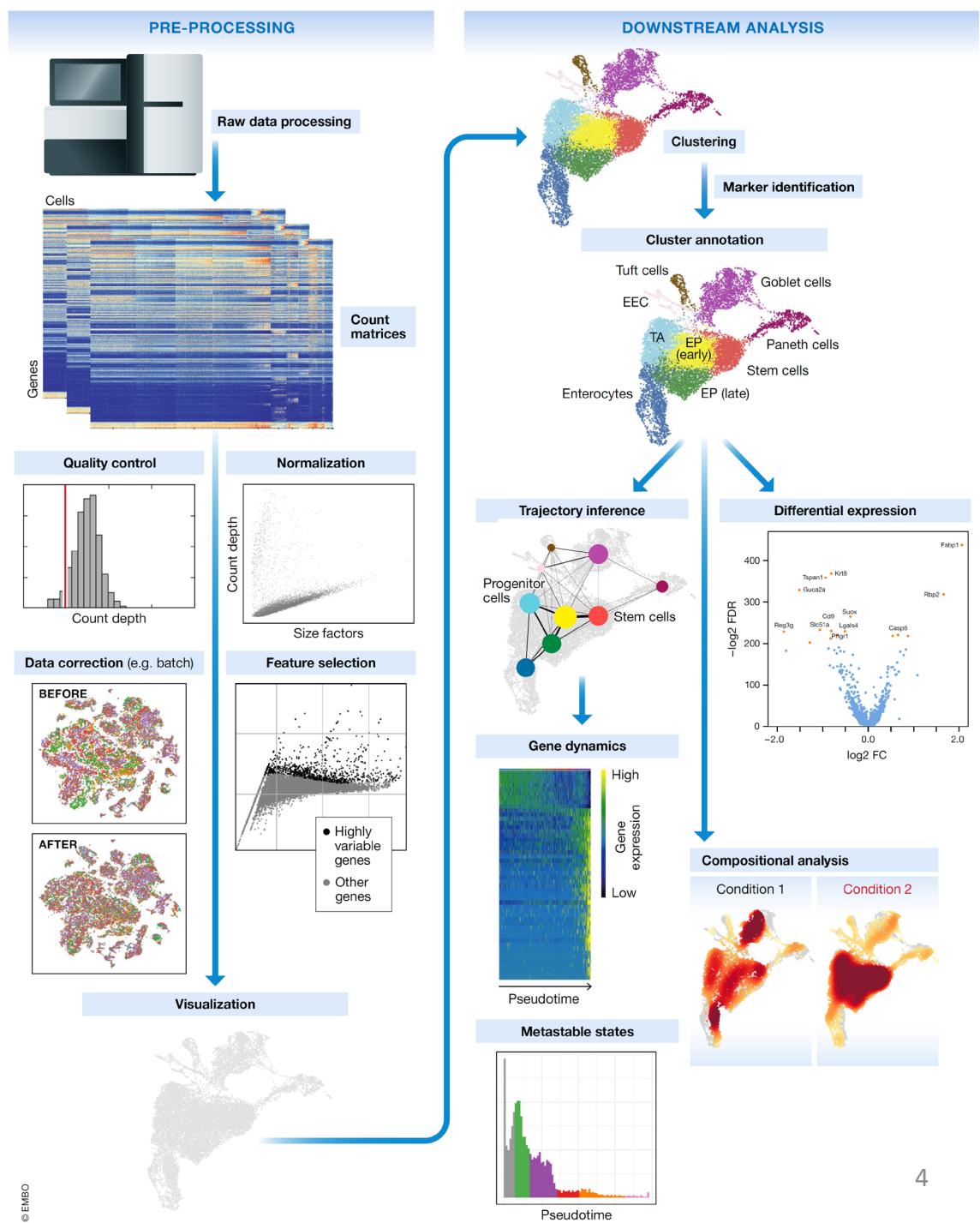


# scRNA-seq Data Analysis

Our goal is to derive/extract real biology from  
technically noisy data

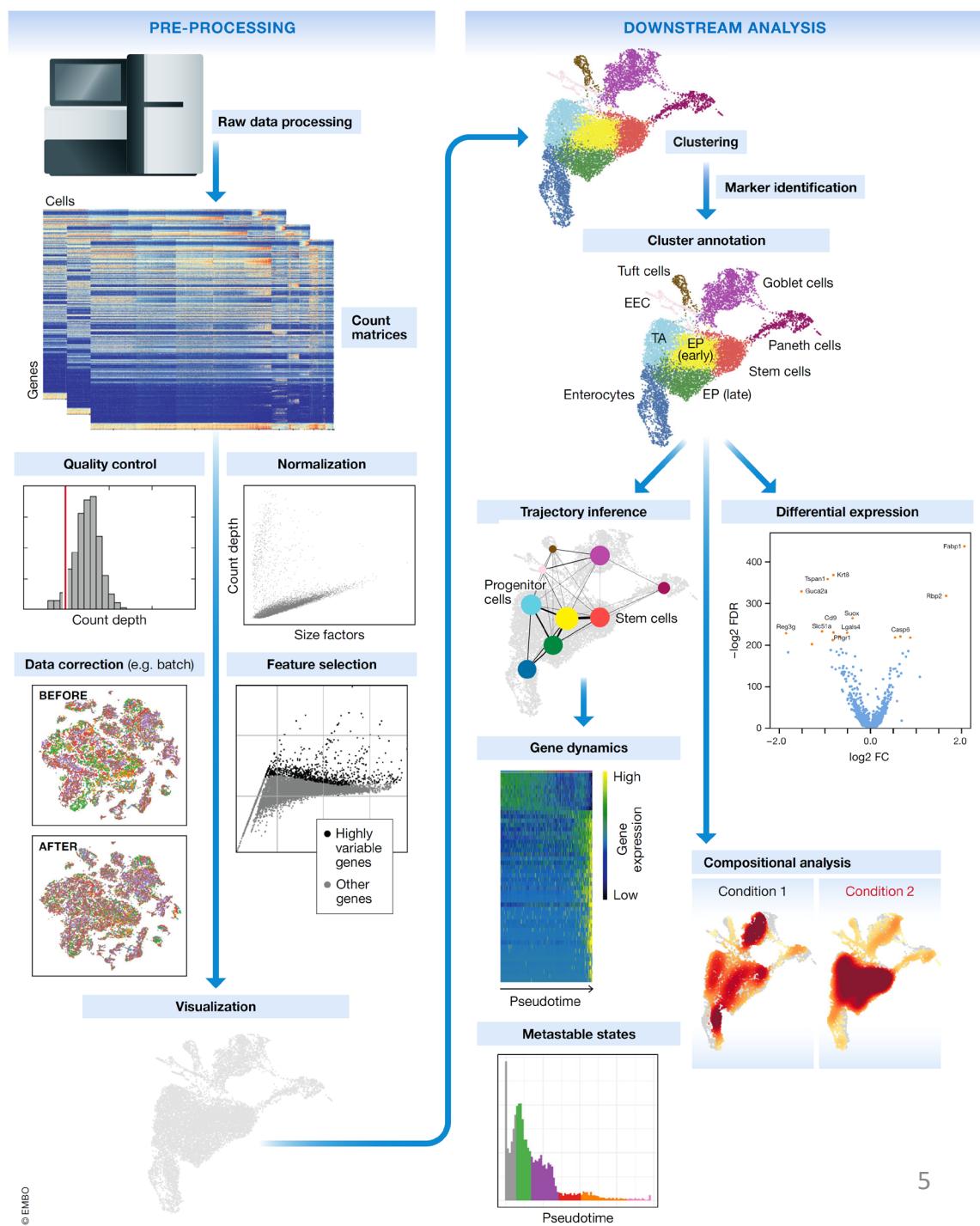
# scRNA-seq Data Analysis

- Preprocessing:
  - Reads to count matrix
  - Quality control (QC)
  - Normalization
  - Batch correction
  - Feature selection
- Downstream
  - Cell type identification (clustering/classification)
  - Trajectory inference
  - Differential expression
  - Compositional analysis
  - Co-expression network analysis



# scRNA-seq Data Analysis

- Preprocessing:
  - Reads to count matrix 
  - Quality control (QC)
  - Normalization
  - Batch correction
  - Feature selection
- Downstream
  - Cell type identification (clustering/classification)
  - Trajectory inference
  - Differential expression
  - Compositional analysis
  - Co-expression network analysis



# Normalization vs Batch correction



- **Normalization:** occurs regardless of the batch structure and only considers technical biases.
- **Batch correction:** only occurs across batches and must consider both technical biases and biological differences.
- *Technical biases:* tend to affect genes in a similar manner, or at least in a manner related to their biophysical properties (e.g., length, GC content).
- *Biological differences:* highly unpredictable.

# Course materials

<https://github.com/LeidenCBC/MGC-BioSB-SingleCellAnalysis2020>

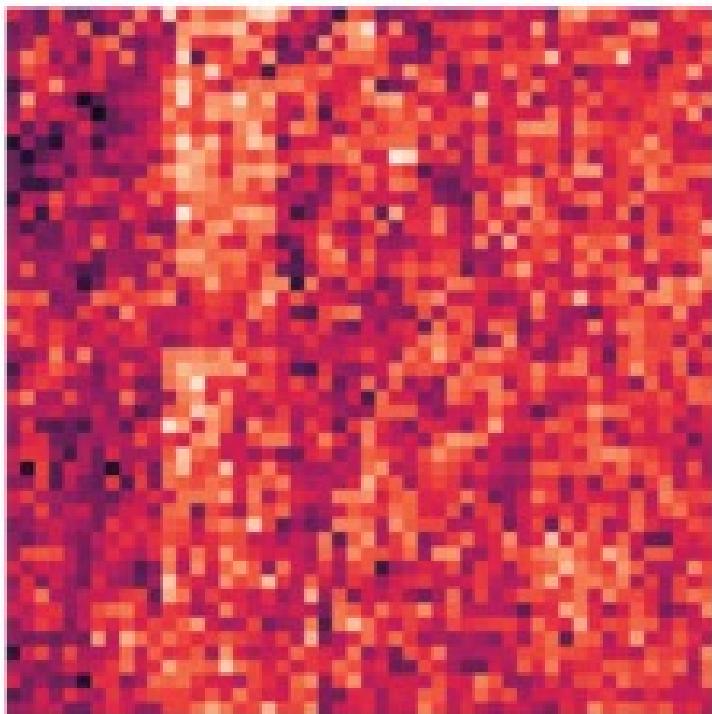
Credits: Åsa Björklund (NBIS, SciLifeLab)

# Our agenda

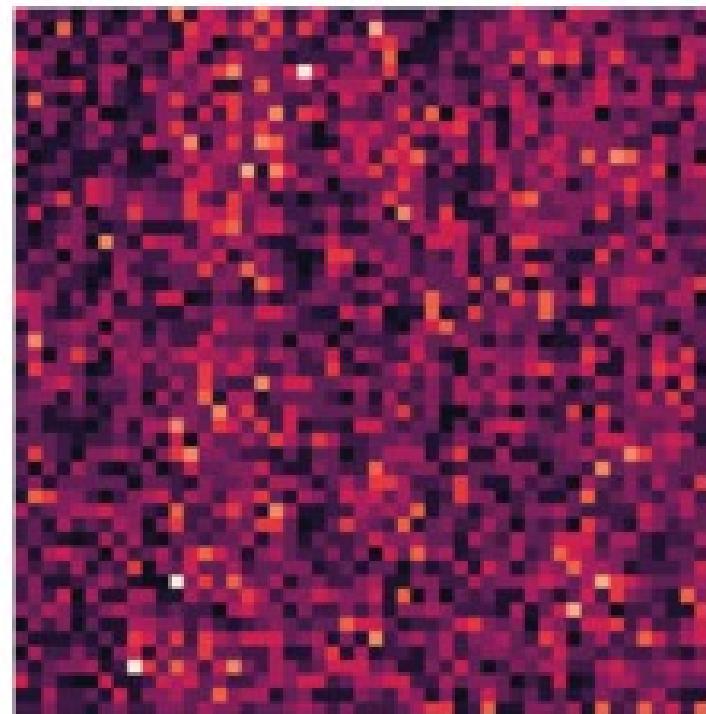
- Background on transcriptional bursting & drop-outs
- Experimental setup – what could go wrong?
- Quality control
- Normalization
- Feature selection

# Which matrix best resembles scRNA-seq data?

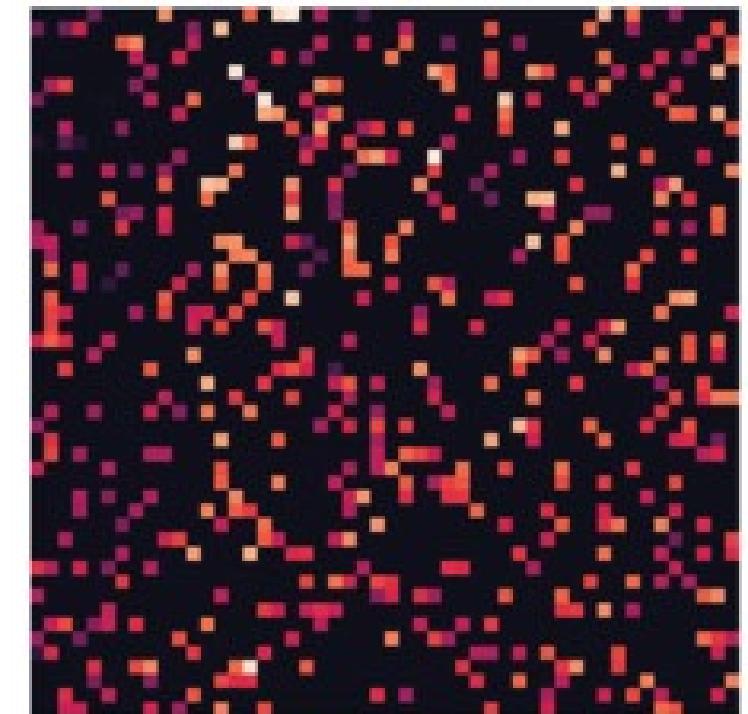
Original  
matrix



80% of the molecules  
are not counted

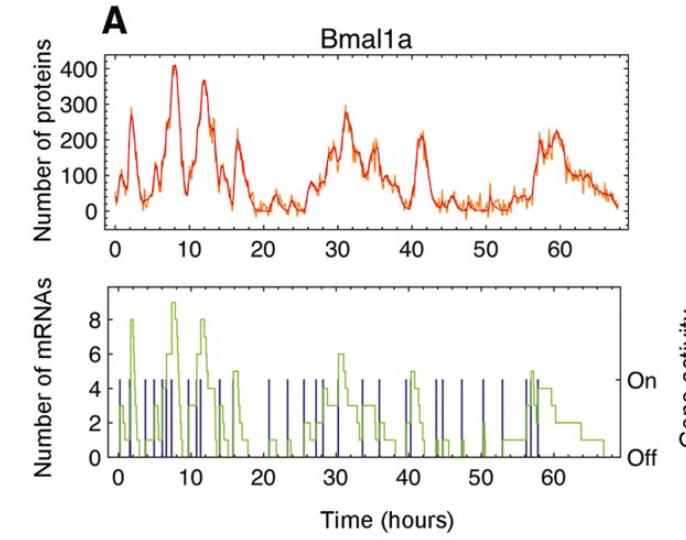
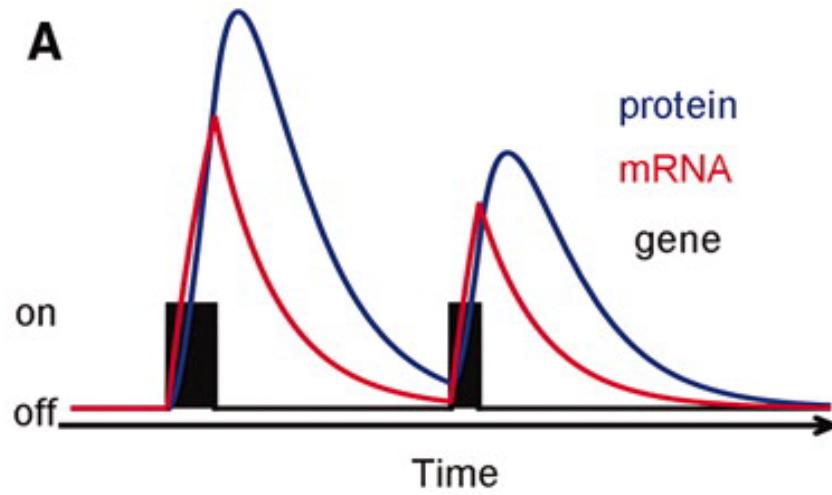


80% zeros

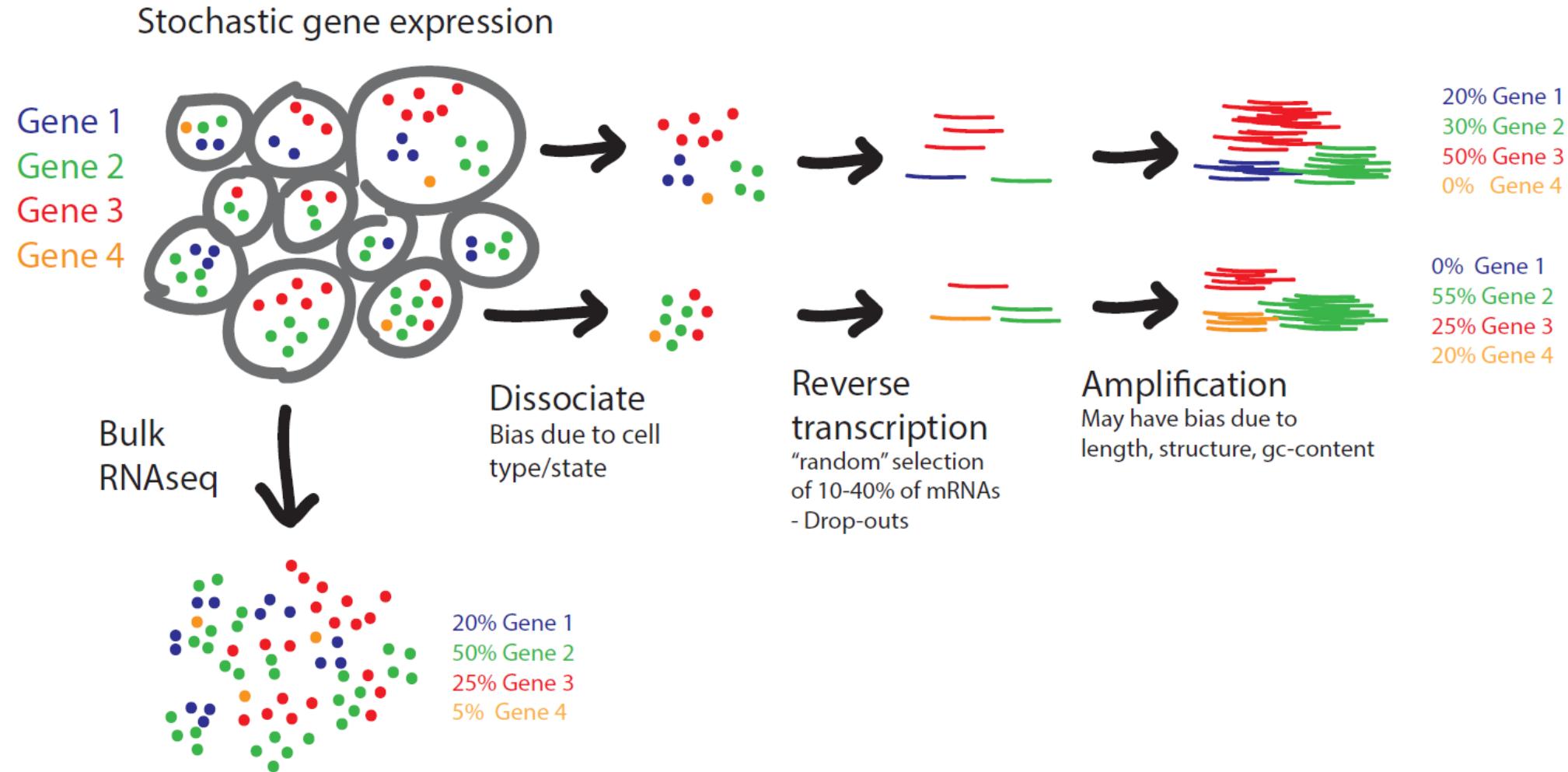


# Transcriptional bursting

- Burst frequency and size is correlated with mRNA abundance
- Many TFs have low mean expression (and low burst frequency) and will only be detected in a fraction of the cells

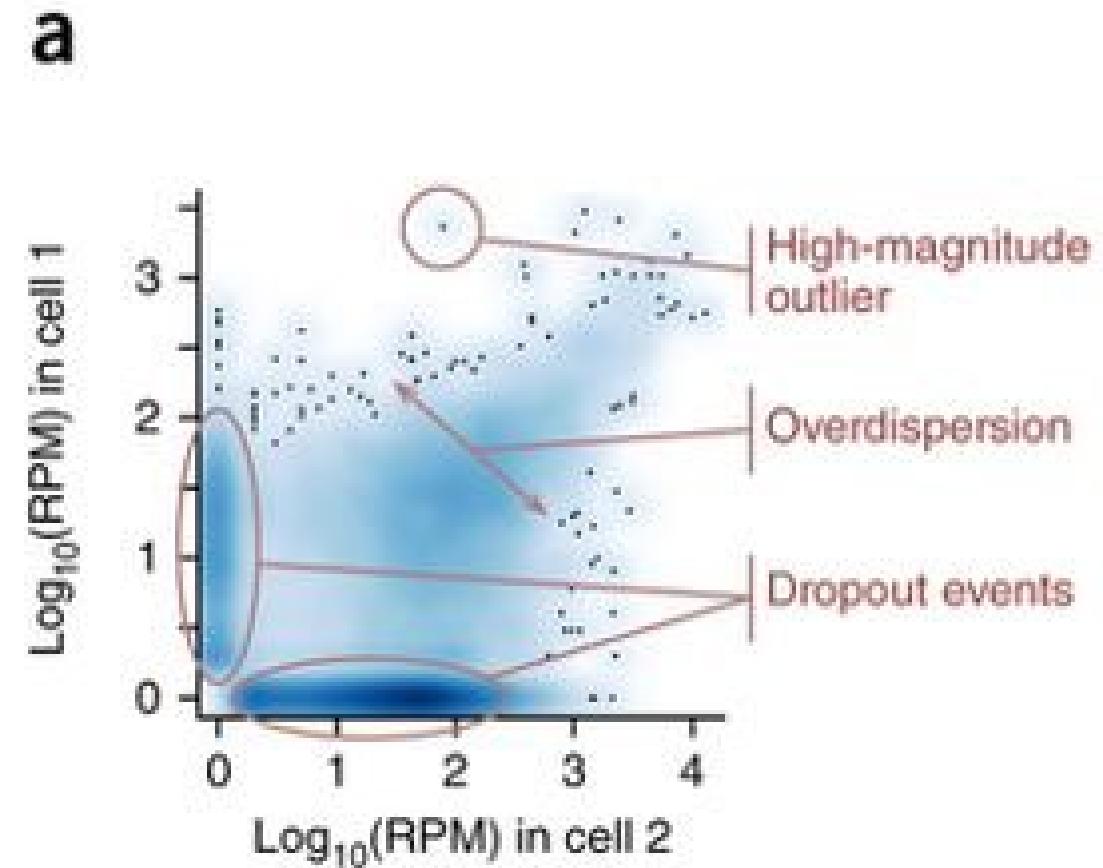


# Bursting, drop-outs and amplification bias



# Problems compared to bulk RNA-seq

- Amplification bias
- Drop-out rates
- Transcriptional bursting
- Background noise
- Bias due to cell-cycle and cell size
- Often clear batch effects



# What could have gone wrong?

Cell dissociation

Cell capture

Cell lysis

Reverse transcription

Preamplification

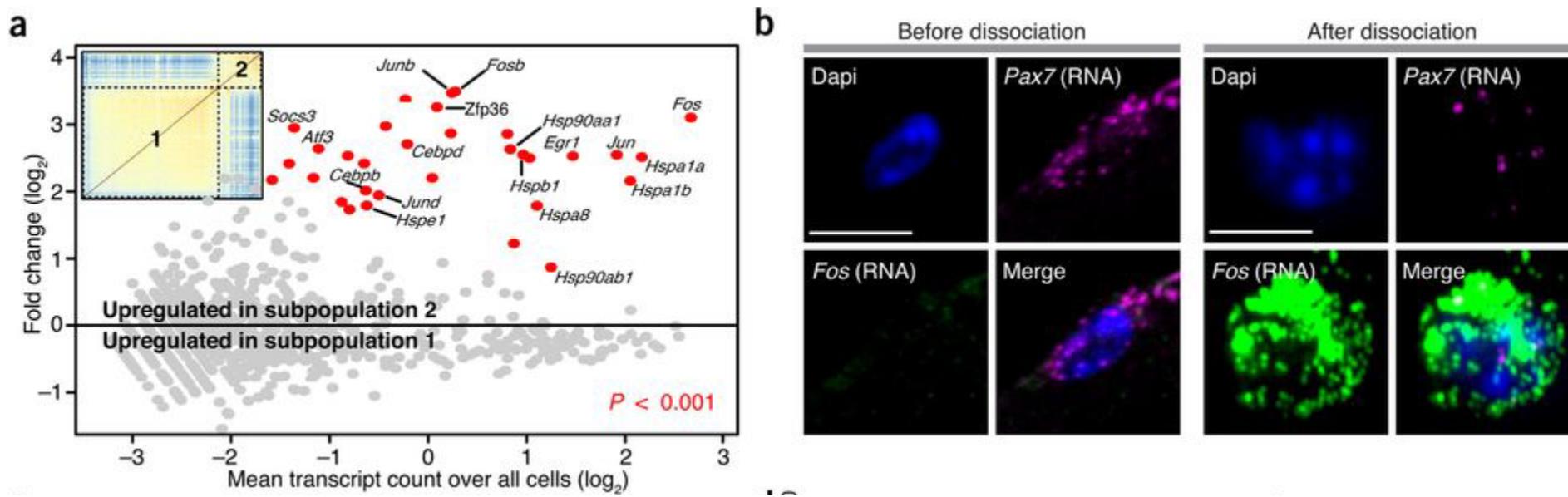
Library preparation and sequencing

# Cell dissociation

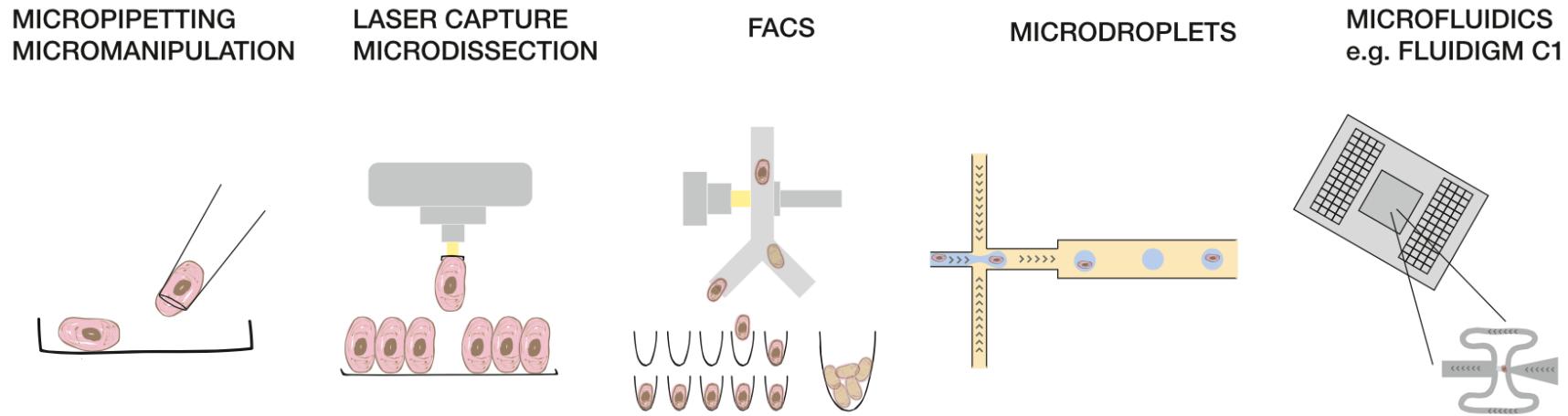
- It is critical to have healthy whole cells with no RNA leakage. Short time from dissociation to cell!
- Tissues that are hard to dissociate:
  - Laser capture microscopy (LCM)
  - Nuclei sorting
- PROBLEMS:
  - Incomplete dissociation can give multiple cells sticking together.
  - Too harsh dissociation may damage cells -> RNA degradation and RNA leakage.
  - Leakage of RNA – background signal.

# Dissociation artifacts

- Dissociation may bias your cell populations
- Dissociation protocols may introduce transcriptional changes.



# Single cell capture

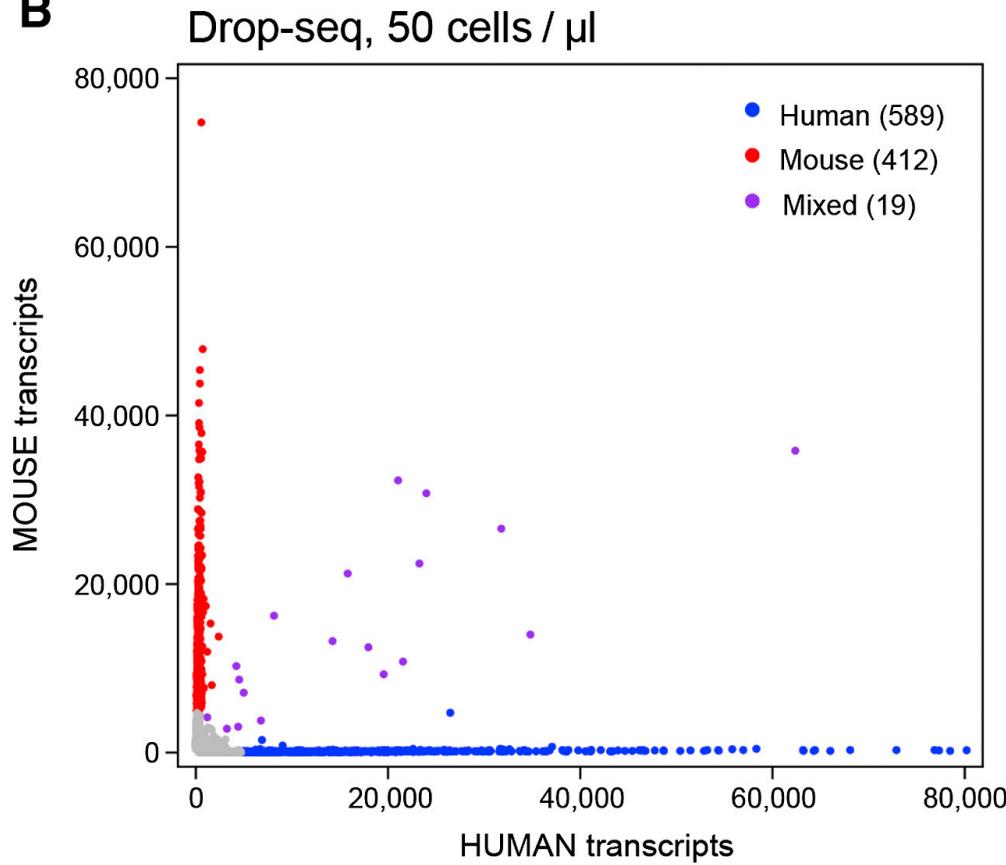


- PROBLEMS:

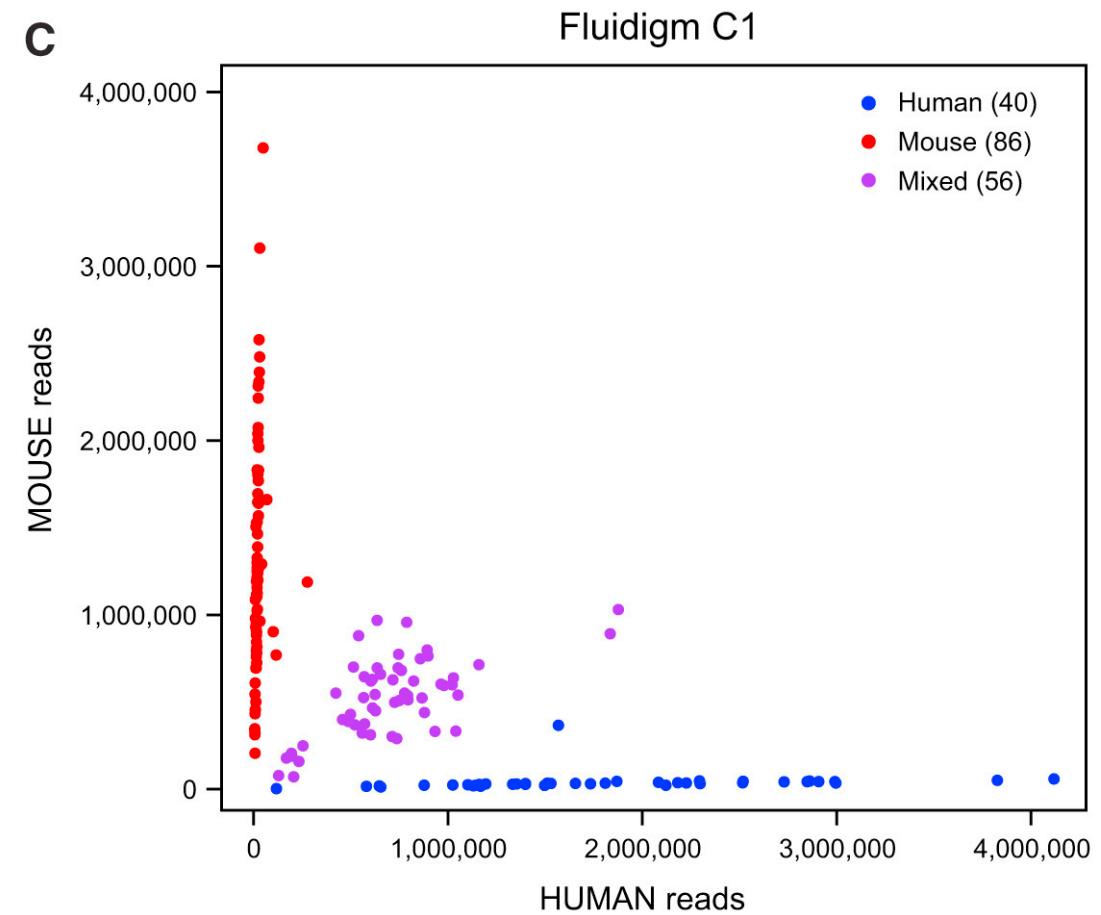
- All these methods may give rise to empty wells/droplets, and also duplicates or multiples of cells.
- Size selection bias for many of the methods – dropseq has upper limit for cell size.
- Biased selection of certain cell type(s)
- Long time for sorting may damage the cells

# scRNA-seq is not always single-cell

B



C



# 10x doublet rate

Multiplet Rate (%)	# of Cells Loaded	# of Cells Recovered
~0.4%	~870	~500
~0.8%	~1700	~1000
~1.6%	~3500	~2000
~2.3%	~5300	~3000
~3.1%	~7000	~4000
~3.9%	~8700	~5000
~4.6%	~10500	~6000
~5.4%	~12200	~7000
~6.1%	~14000	~8000
~6.9%	~15700	~9000
~7.6%	~17400	~10000

# Doubllets

- High number of detected genes or UMIs – can be a sign of multiples
  - But, beware so that you do not remove all cells from a larger cell type.
- After clustering – check if you have cells with signatures from multiple clusters.
- A combination of those 2 features would indicate duplicates.
- With 10X you should have a feeling for your doublet rate based on how many cells were loaded

# Doublet detection

- DoubletFinder

<https://github.com/chris-mcginnis-ucsf/DoubletFinder>

- Scrublet

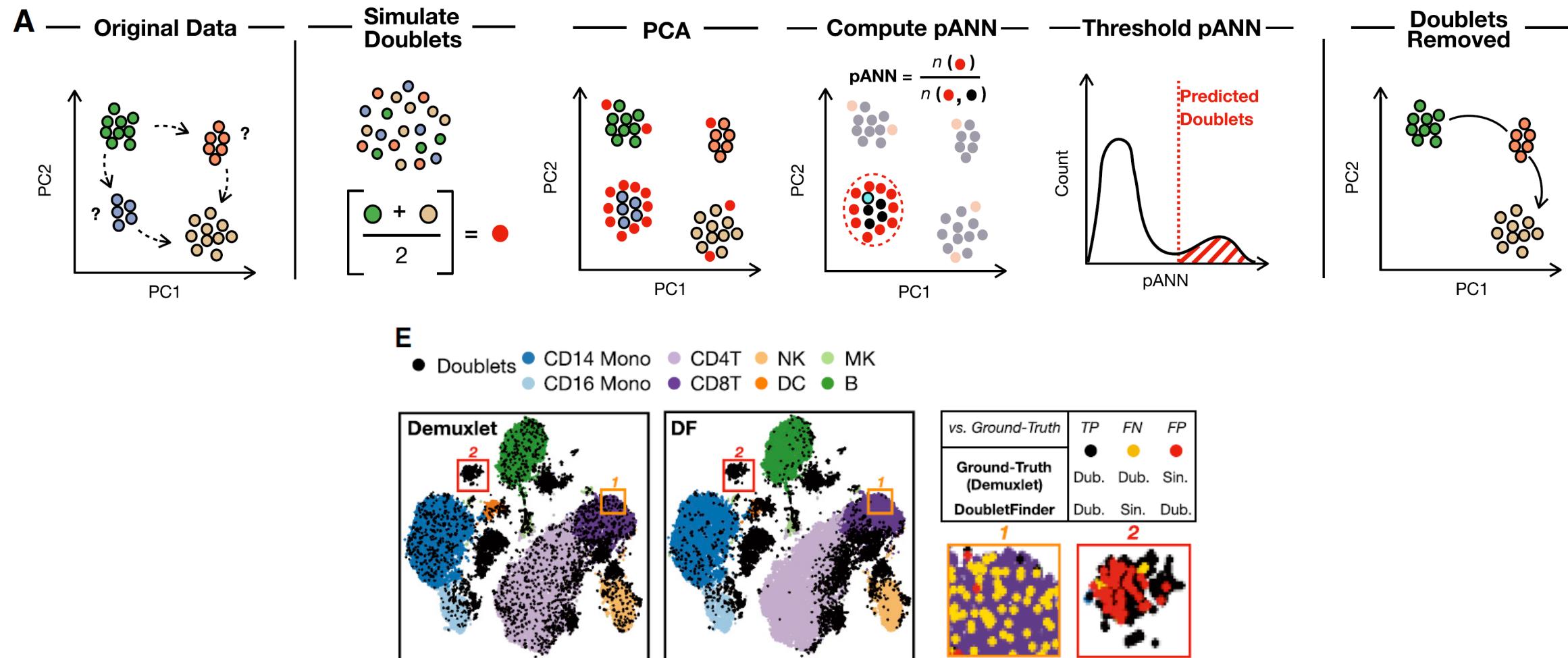
<https://github.com/AllonKleinLab/scrublet>

- DoubletDecon

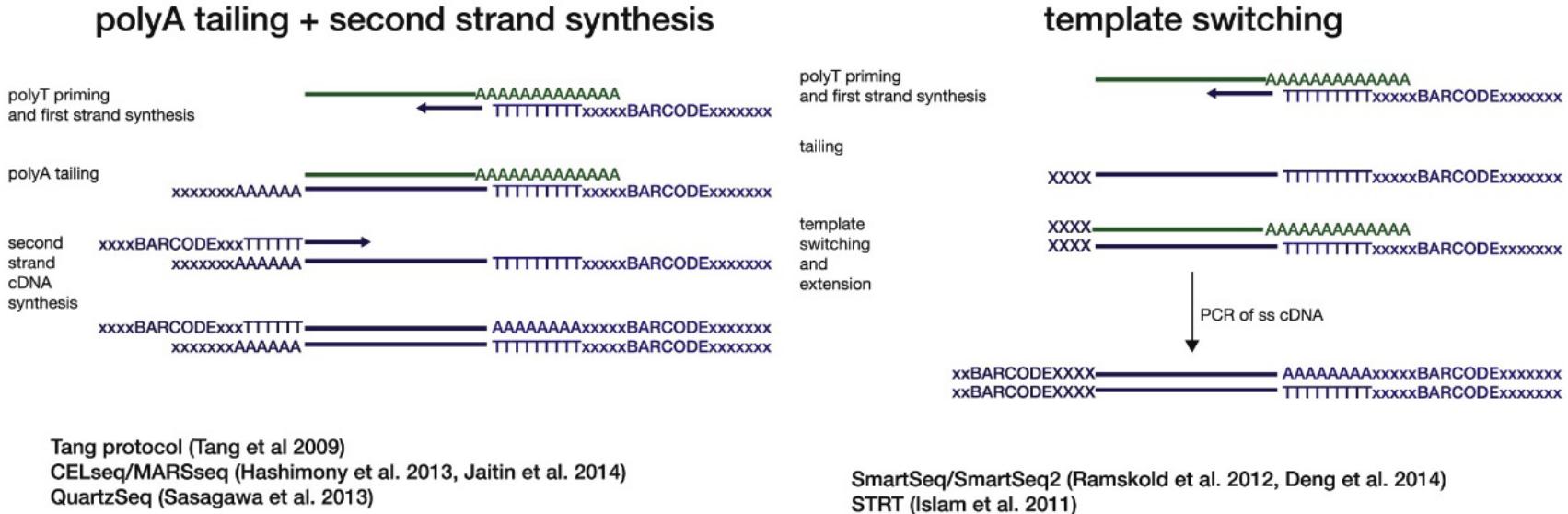
<https://github.com/EDePasquale/DoubletDecon>

- DoubletCluster / DoubletCell in Scran

# DoubletFinder



# Reverse transcription



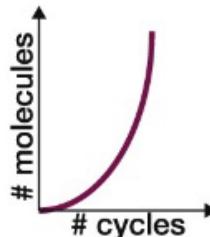
- Efficiency of reverse transcription is the key to high sensitivity.
- Drop-out rate is around 90-60% depending on the method used.
- Two libraries with the same method using the same cell type may have very different drop-out rates.

# Preamplification

## PCR

- exponential amplification
- PCR base specific biases

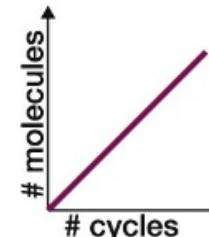
Tang protocol (Tang et al. 2009)  
STRT (Islam et al. 2011)  
SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)



## IVT

- linear amplification
- 3' bias due to two rounds of reverse transcription

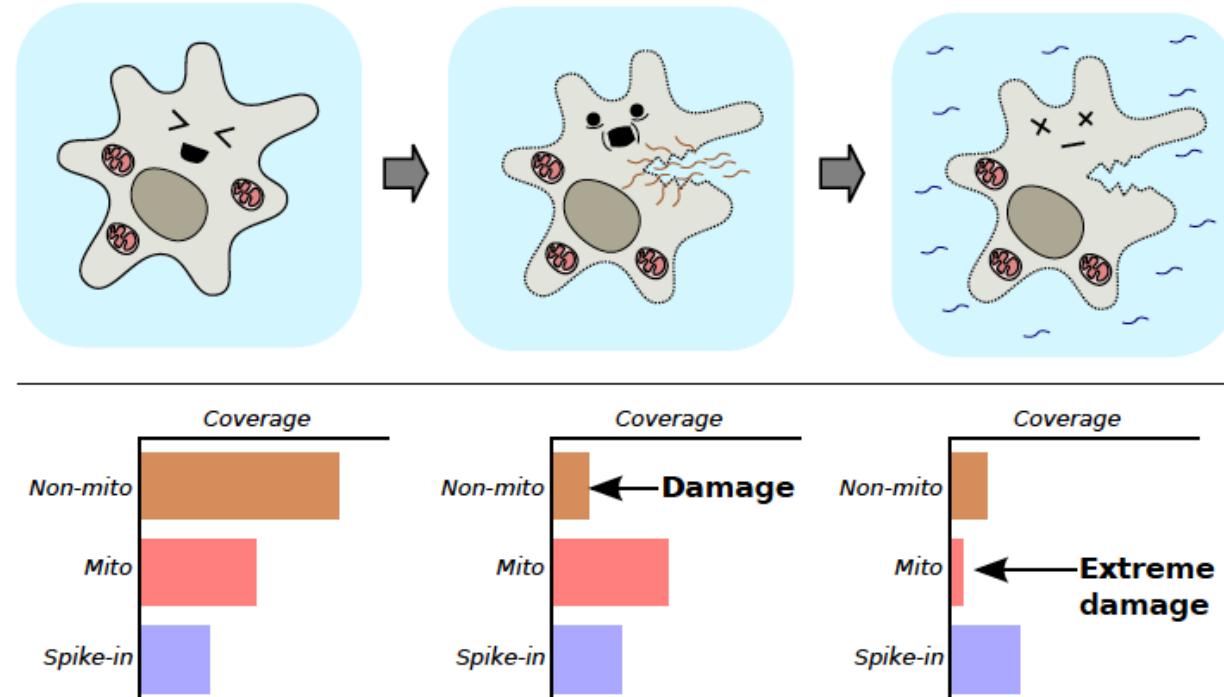
CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)



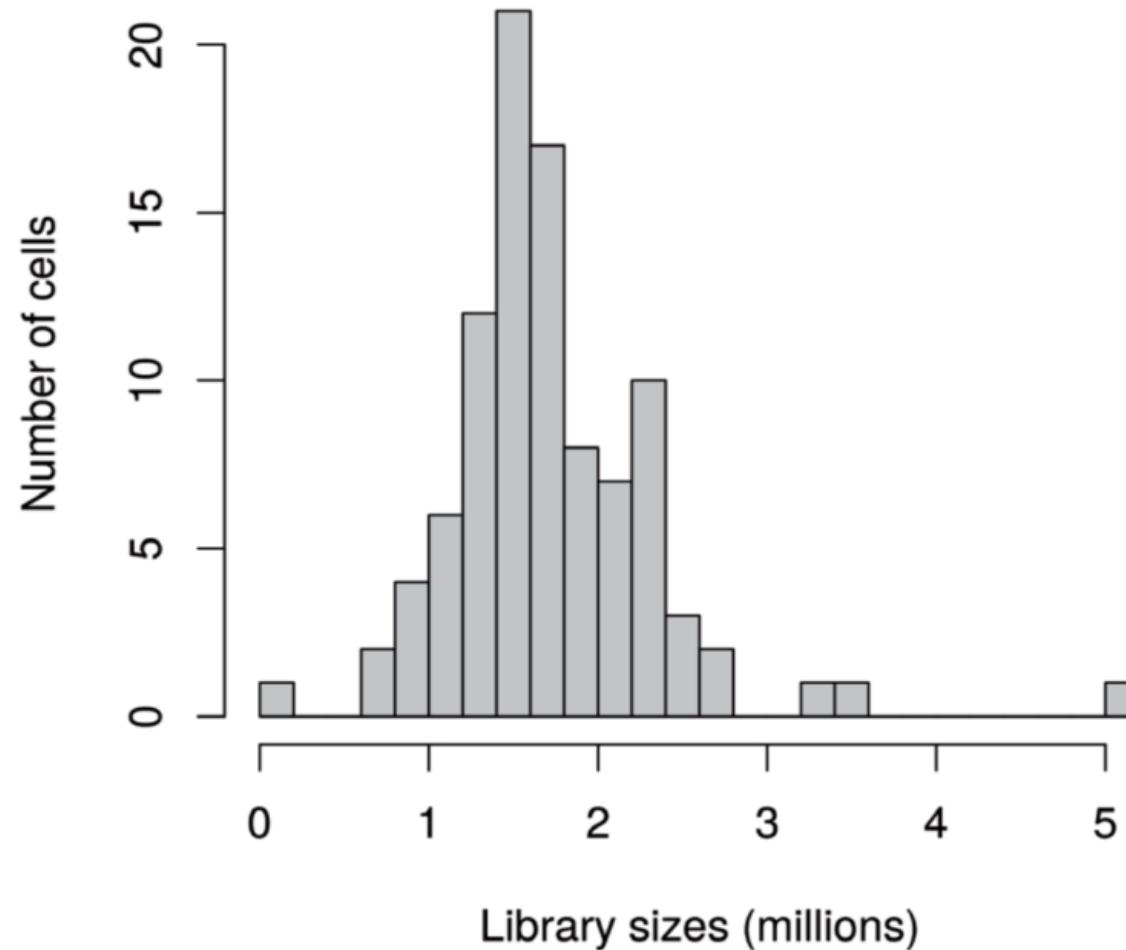
- Any amplification step will introduce a bias in the data.
- Methods that use UMIs will control for this to a large extent, but the chance of detecting a transcript that is amplified more is higher.
- Full length methods like SmartSeq2 have no UMIs, so we cannot control for amplification bias.

# Quality control of cells (1)

- Low sequencing depth
- Low numbers of expressed genes (i.e. any nonzero count)
- High spike-in (if present) or mitochondrial content



# Select low quality cells



Powered by  **Poll Everywhere**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at [pollev.com/app](https://pollev.com/app)

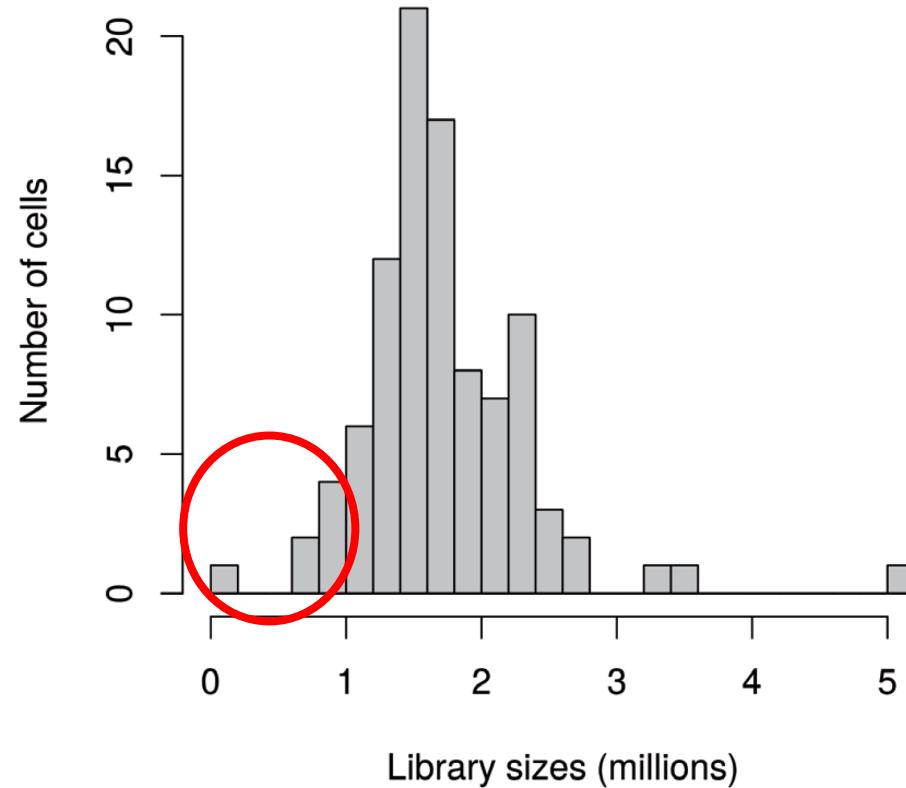
# Why do you think these cells are of low quality?

Top

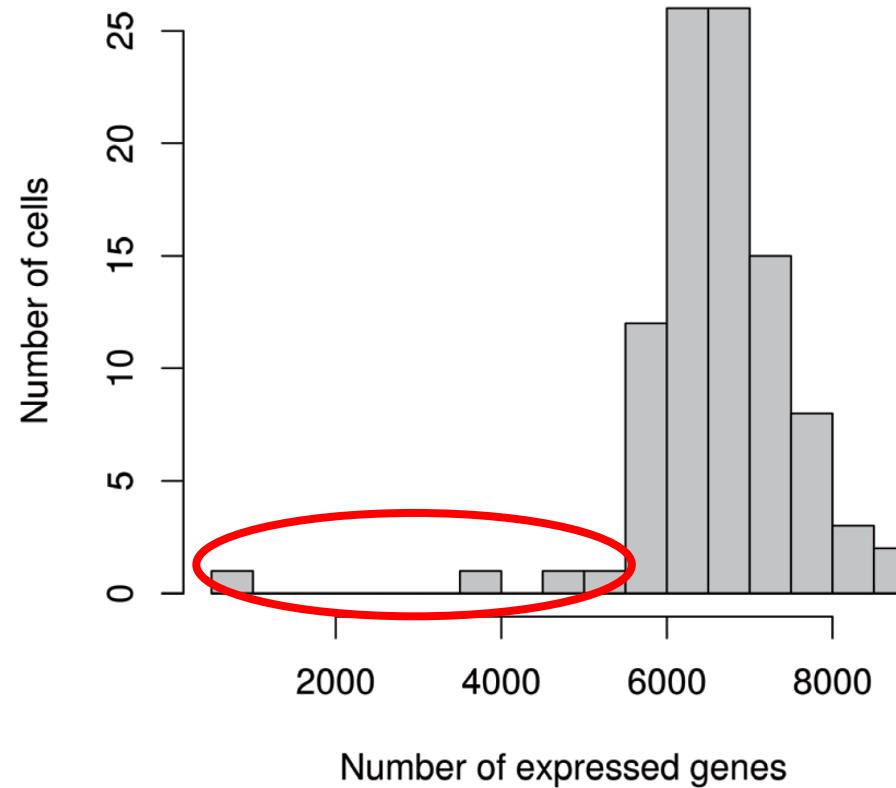
Powered by  Poll Everywhere

Start the presentation to see live content. For screen share software, share the entire screen. Get help at [pollev.com/app](http://pollev.com/app)

# Quality control of cells (2)

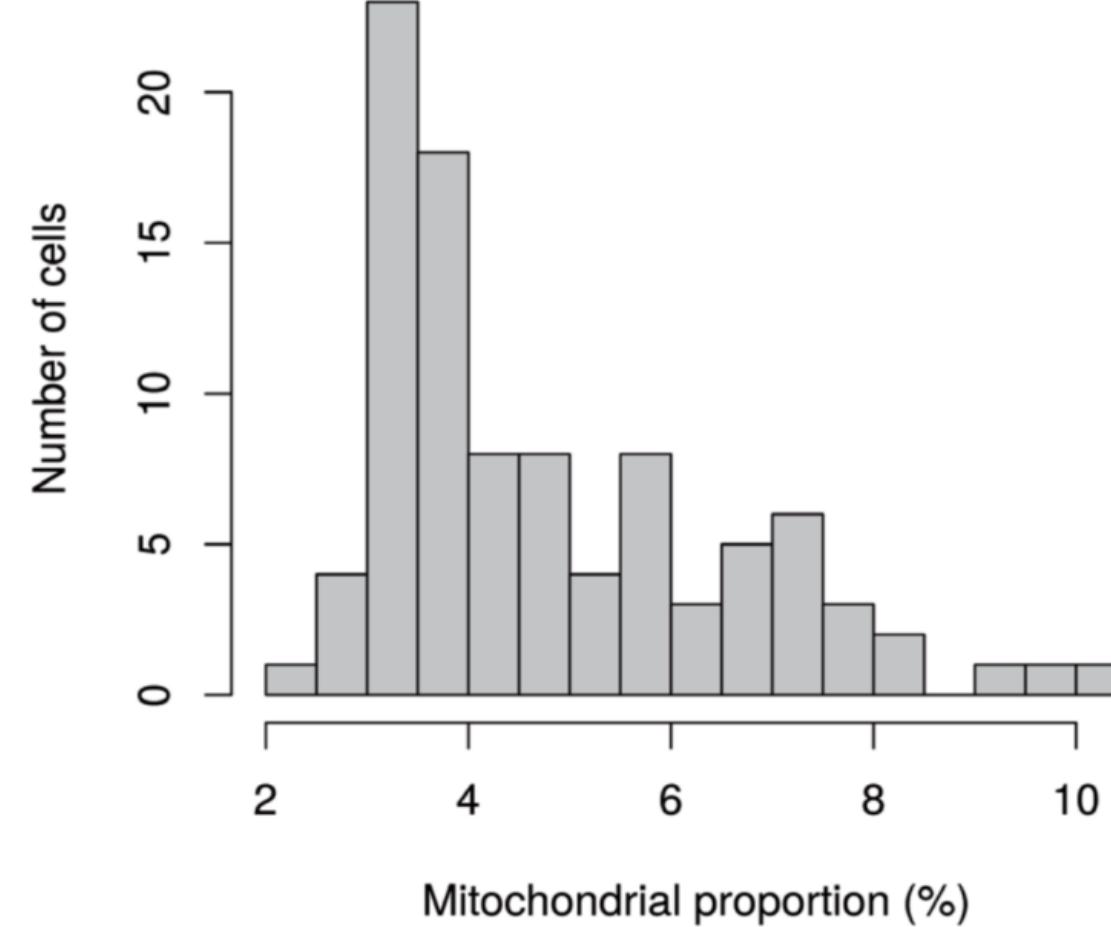


RNA has not been efficiently captured during library preparation



Diverse transcript population not captured

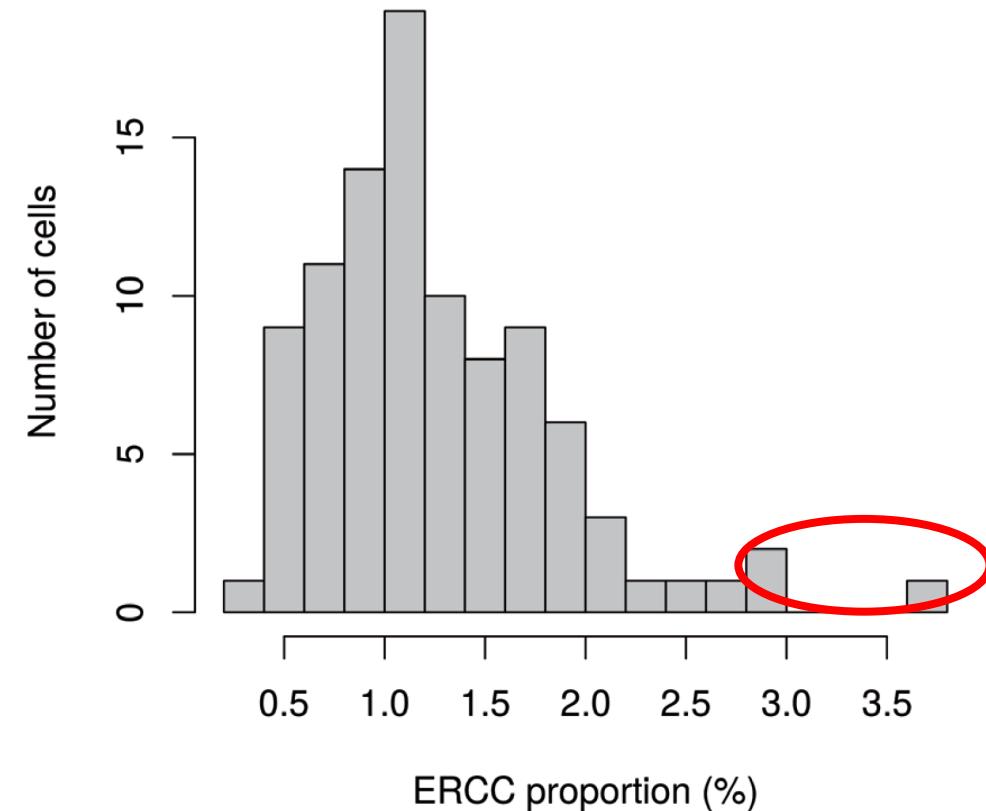
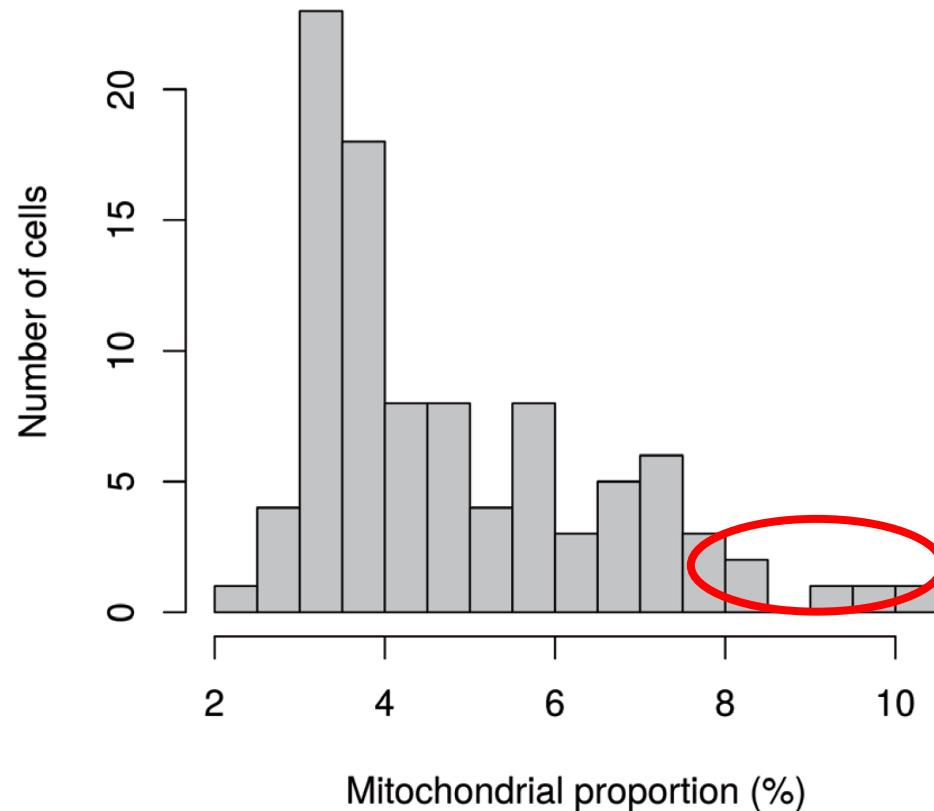
# Select low quality cells



Powered by  **Poll Everywhere**

Start the presentation to see live content. For screen share software, share the entire screen. Get help at [pollev.com/app](http://pollev.com/app)

# Quality control of cells (2)

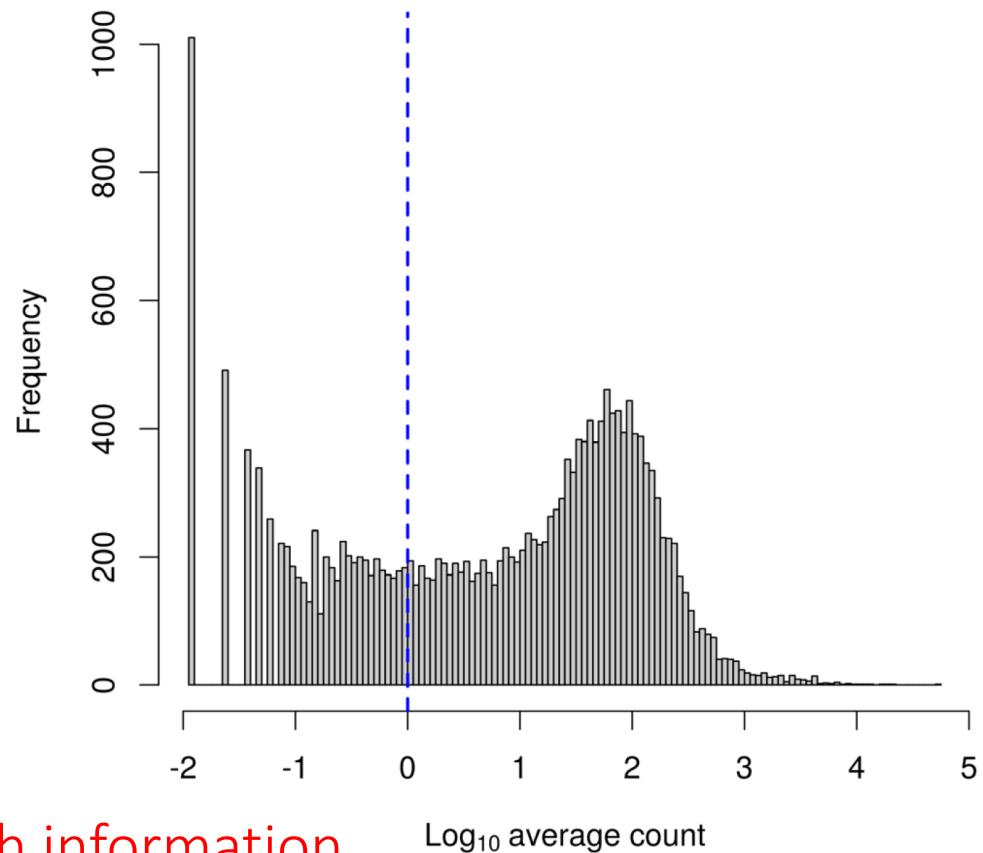


Possibly because of increased apoptosis  
and/or loss of cytoplasmic RNA from lysed cells

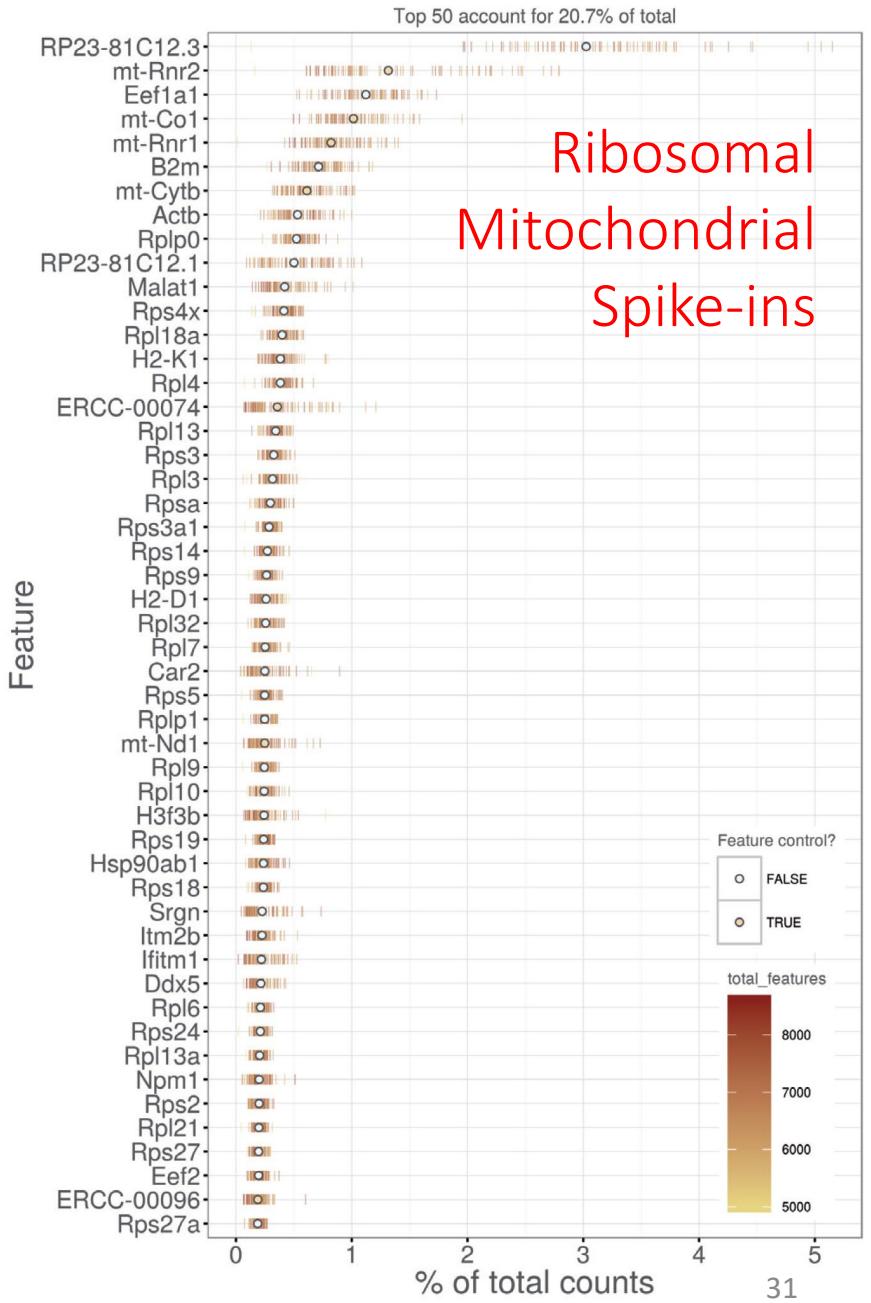
# Deciding on cutoffs for filtering

- Do you have a homogeneous population of cells with similar sizes?
- Is it possible that you will remove cells from a smaller cell type?
- Examine PCA/tSNE/UMAP before and after filtering and make a judgment on whether to remove more or less cells.

# Quality control of genes



Not enough information  
for reliable statistical  
inference



# QC (pitfalls and recommendations)

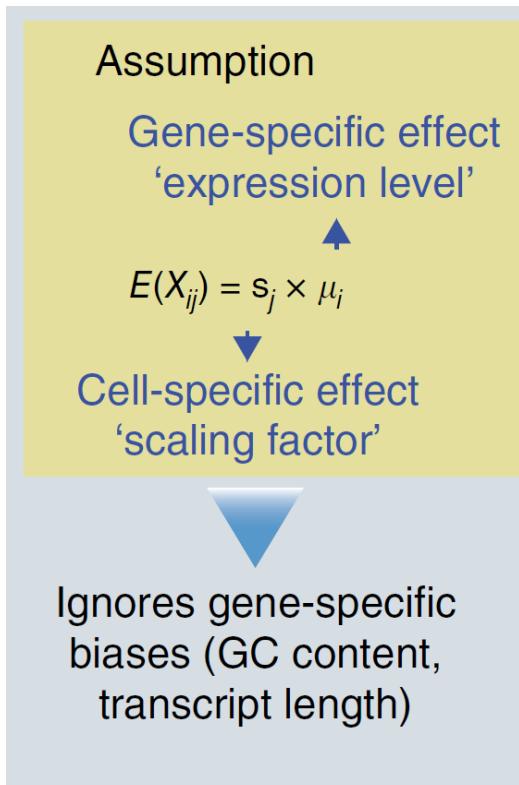
- Perform QC by finding outlier peaks in the number of genes, the count depth and the fraction of mitochondrial reads. Consider these covariates jointly instead of separately.
- Be as permissive of QC thresholding as possible, and revisit QC if downstream clustering cannot be interpreted.
- If the distribution of QC covariates differ between samples, QC thresholds should be determined separately for each sample to account for sample quality differences as in Plasschaert et al (2018).

# Check!

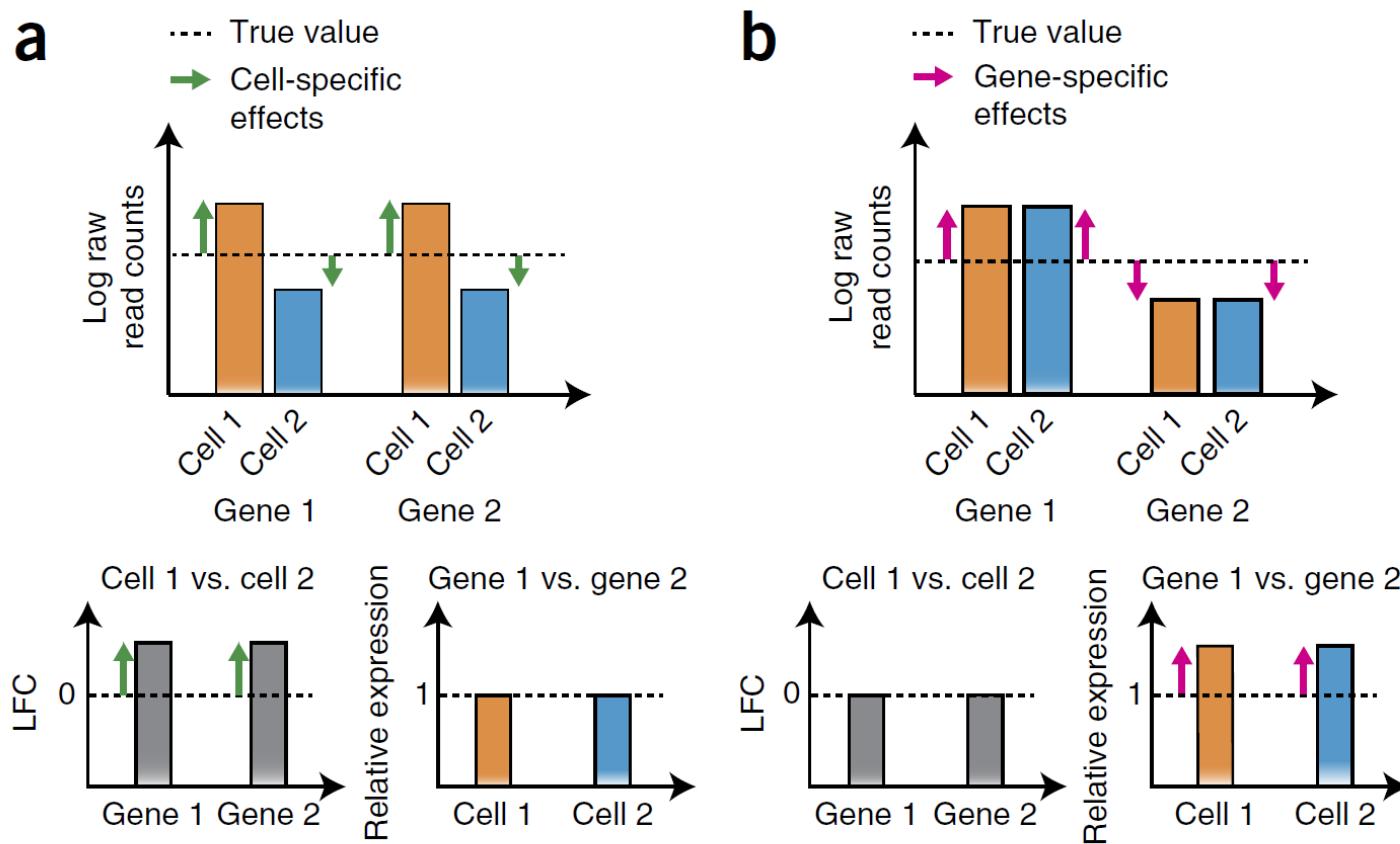
- Always go back to QC-stats after doing downstream analysis (clustering/lineage analysis etc.)
- Are your findings correlated with technical factors?

# Normalization

# Normalization (1)



# Cell- and gene-specific effects in RNA-seq experiments



When poll is active, respond at **PollEv.com/ahmedmahfouz577**

## Which effects are NOT removed by UMIs?

Sequencing depth

Amplification

Capture RT and efficiency

Gene length

GC content

mRNA content

Total Results: 0

# Which effects are NOT removed by UMIs?

Sequencing depth

Amplification

Capture RT and efficiency

Gene length

GC content

mRNA content

## Which effects are NOT removed by UMIs?

Sequencing depth

Amplification

Capture RT and efficiency

Gene length

GC content

mRNA content

# Which effects are NOT removed by UMIs?

C

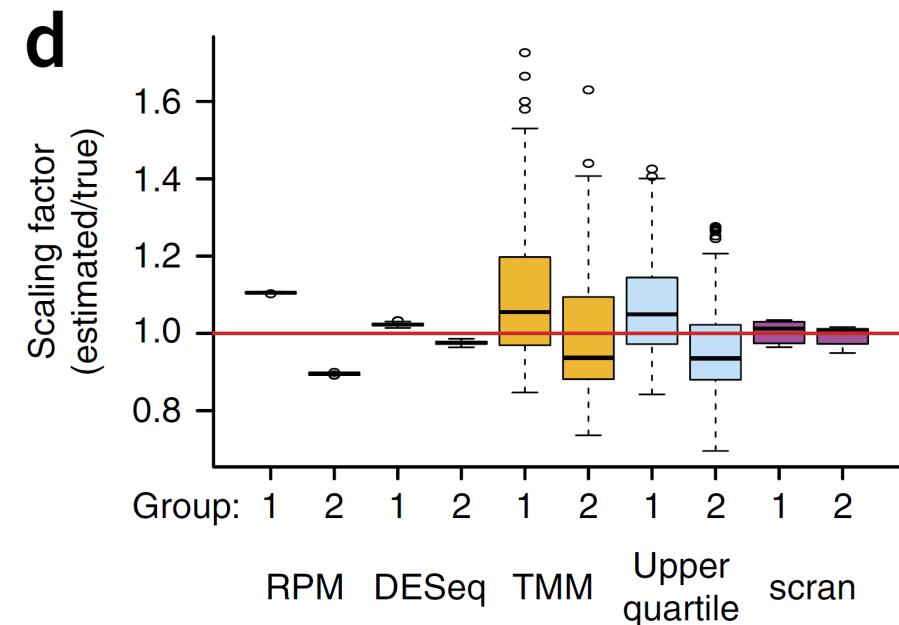
	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

# Normalization (2)

- The aim is bring all cells onto the same distribution to remove biases
- We want to preserve biological variability, not introduce new technical variation
- Primary source of bias is sequencing depth – scale down counts accordingly
- Need a method that is robust to sparsity and composition bias

# What is different from bulk RNA-seq?

- Noise
  - Low mRNA content per cell
  - Variable mRNA capture
  - Variable sequencing depth
- Different cell types in the same sample
- Bulk RNA-seq normalization methods (FPKM, CPM, TPM, upperquartile) are based on per-gene statistics → not suitable for zero-inflated data



# Normalization methods

1. Size factor scaling methods
  - Log-normalization
2. Probabilistic methods
  - scTransform (Hafemeister & Satija Genome Biol 2019)
  - ZINB-WaVE (Risso et al. Nature Comm 2018)

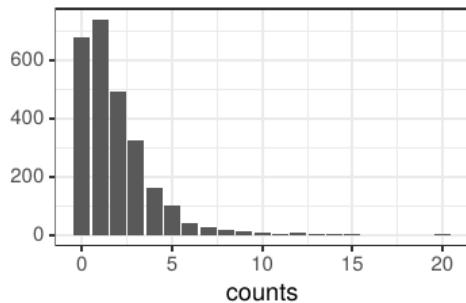
# Log-normalization

$$Y_{ij} = \log_e\left(\frac{X_{ij}}{\sum_i X_{ij}} \times 10,000\right) + 1$$

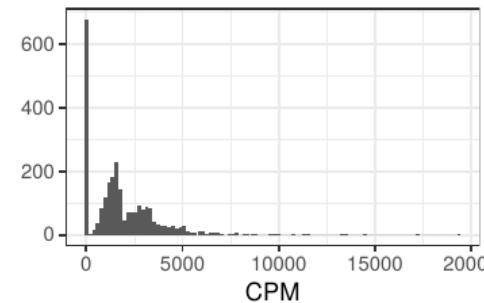
- Simplest and most commonly used normalization strategy
- Divide all counts for each cell by a cell-specific scaling factor (i.e. size factor)
- Assumes that any cell-specific bias (e.g., in capture or amplification efficiency) affects all genes equally via scaling of the expected mean count for that cell
- Modified CPM normalization
- Seurat, scanpy, 10X Cell Ranger: log-normalization

# Effect of dropouts on normalization

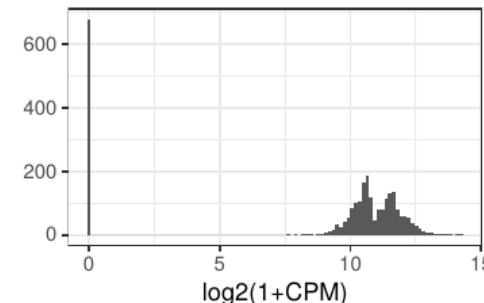
**Inflation of zero counts**



(a) UMI counts

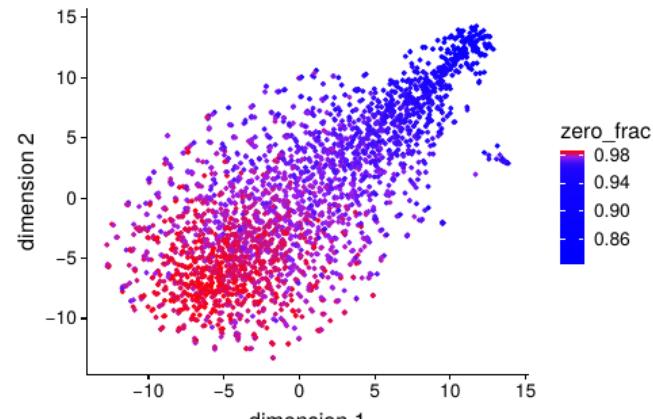
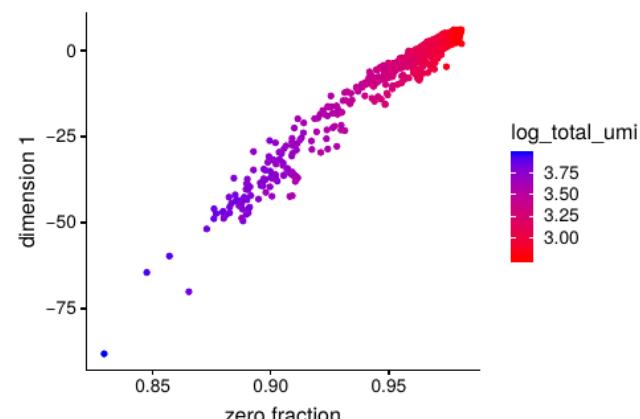


(b) counts per million (CPM)



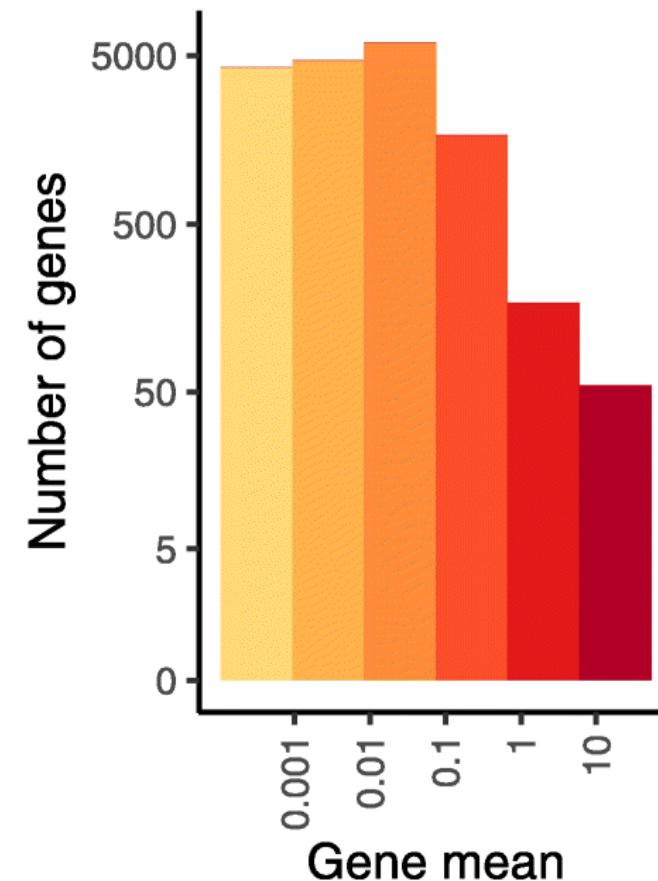
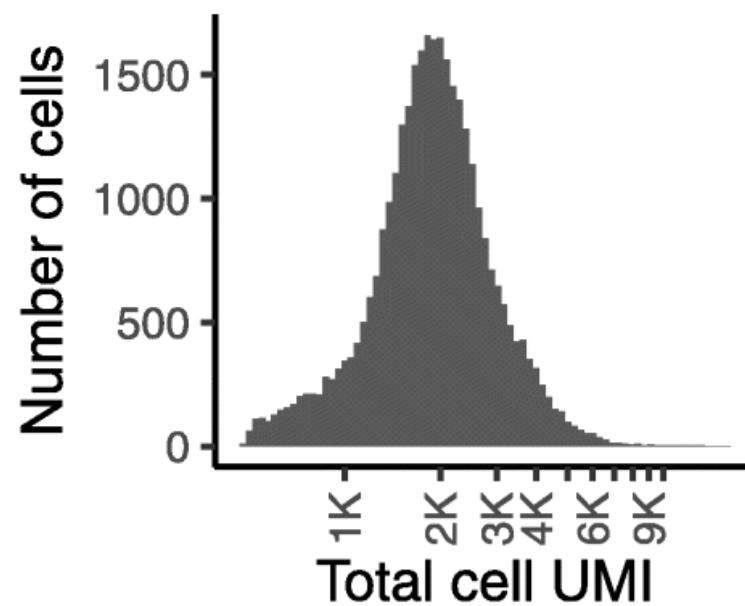
(c) log of CPM

**Fraction of zeros become main source of variability**



# Does log-normalization (scaling) work?

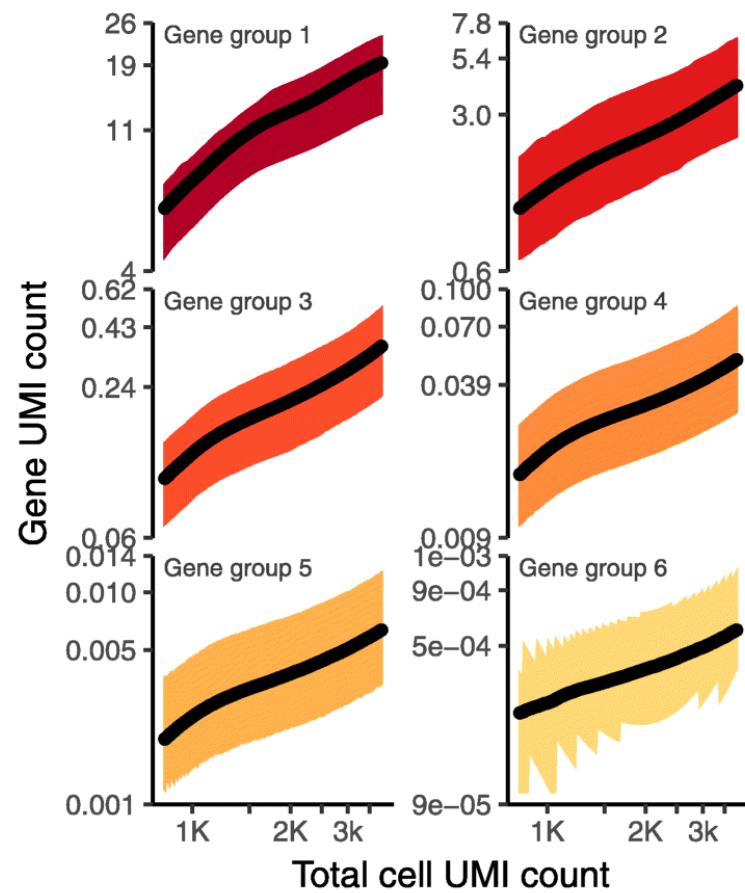
33,148 PBMCs, 10x Genomics  
16,809 genes detected  $\geq 5$  cells



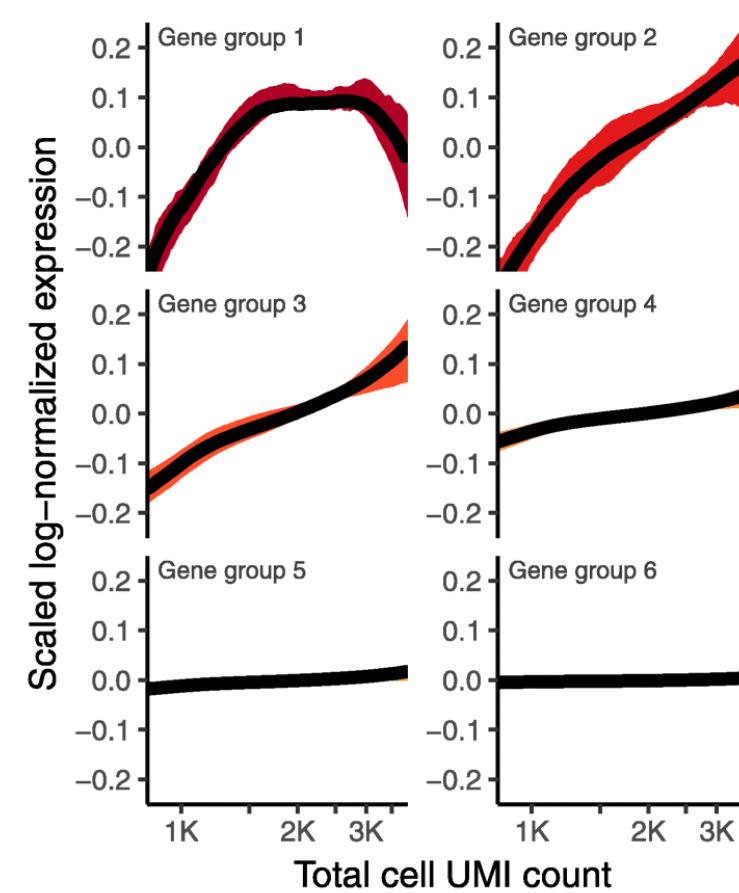
Gene group ID, size	
1,	55
2,	171
3,	1687
4,	5942
5,	4694
6,	4260

# Does log-normalization (scaling) work?

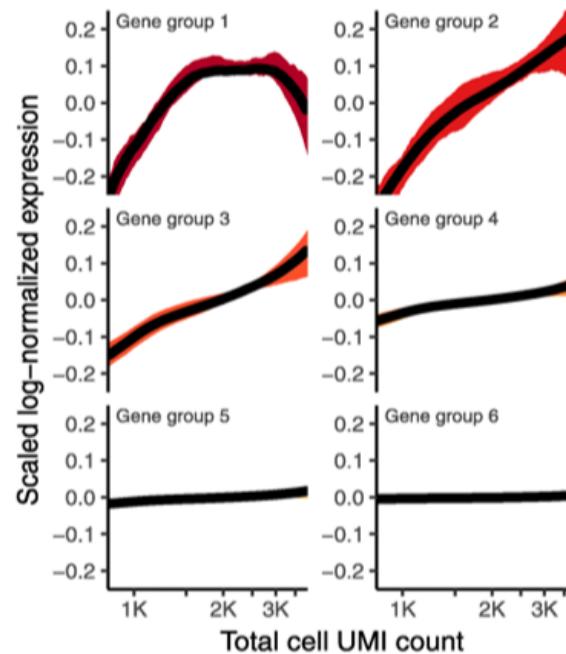
Before normalization



After normalization



# Which groups of genes are effectively normalized?



Gene group 1

Gene group 2

Gene group 3

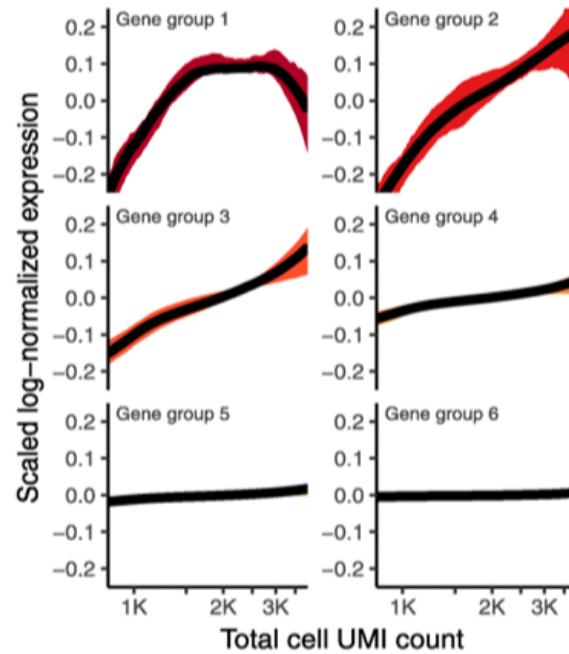
Gene group 4

Gene group 5

Gene group 6

Total Results: 0

# Which groups of genes are effectively normalized?



- Gene group 1
- Gene group 2
- Gene group 3
- Gene group 4
- Gene group 5
- Gene group 6

# Modeling scRNAseq data

- Model the UMI counts for a given gene using a generalized linear model

$$\log(\mathbb{E}(x_i)) = \beta_0 + \beta_1 \log_{10} m + e_i$$

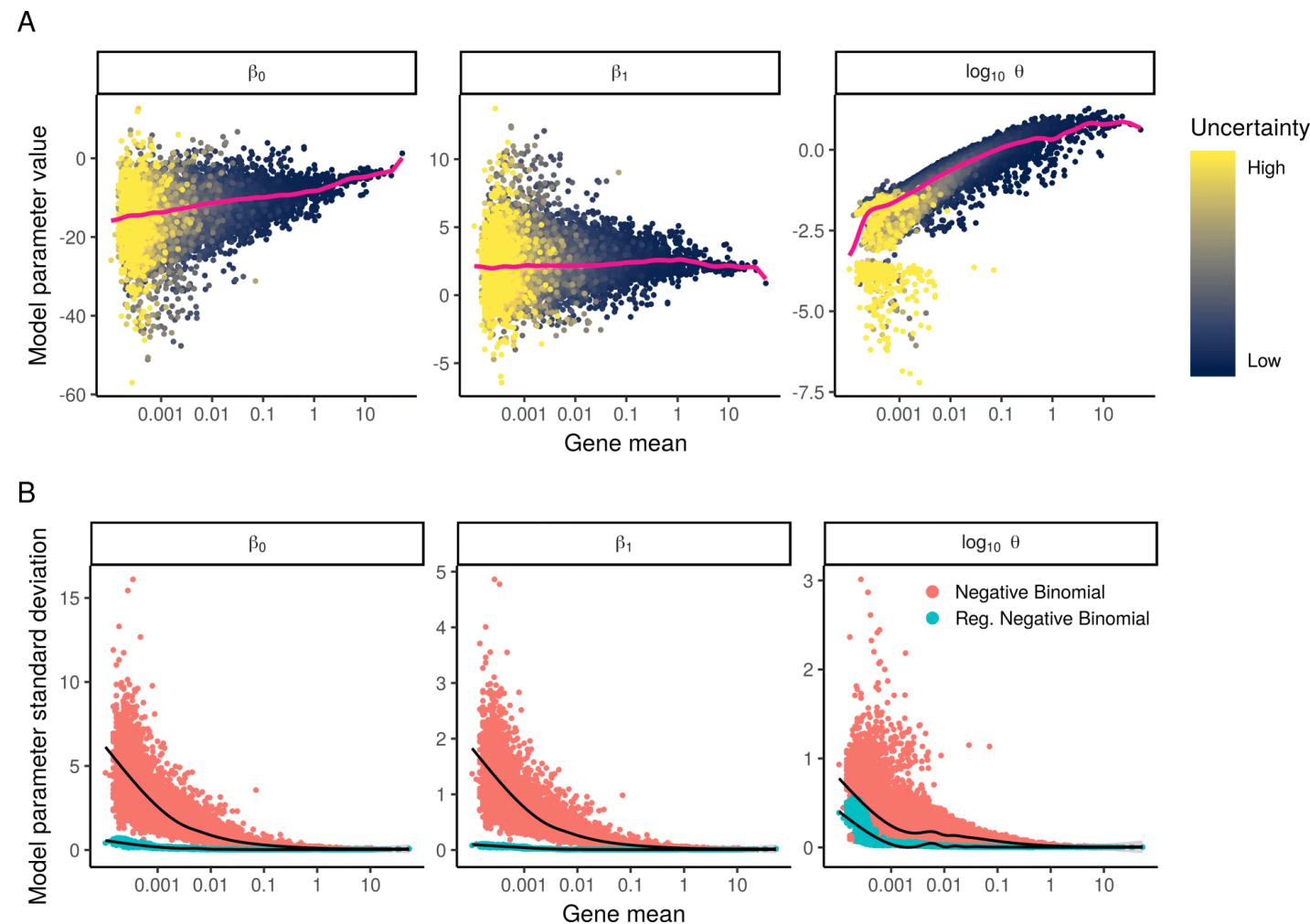
$x_i$ : vector of UMI counts assigned to gene  $i$

$m$ : vector of molecules assigned to the cells, i.e.,  $m_j = \sum_i x_{ij}$

$e_i$ : negative binomial (NB) error distribution, parameterized with mean  $\mu$  and variance  $\mu + \frac{\mu^2}{\sigma}$

# Modeling scRNAseq data

- BUT, modeling each gene separately results in overfitting
- Solution: regularize all model parameters, including the NB dispersion parameter  $\theta$ , by sharing information across genes



# Modeling scRNAseq data

## scTransform: Regularized negative binomial regression

**Step1:** fit independent regression models per gene

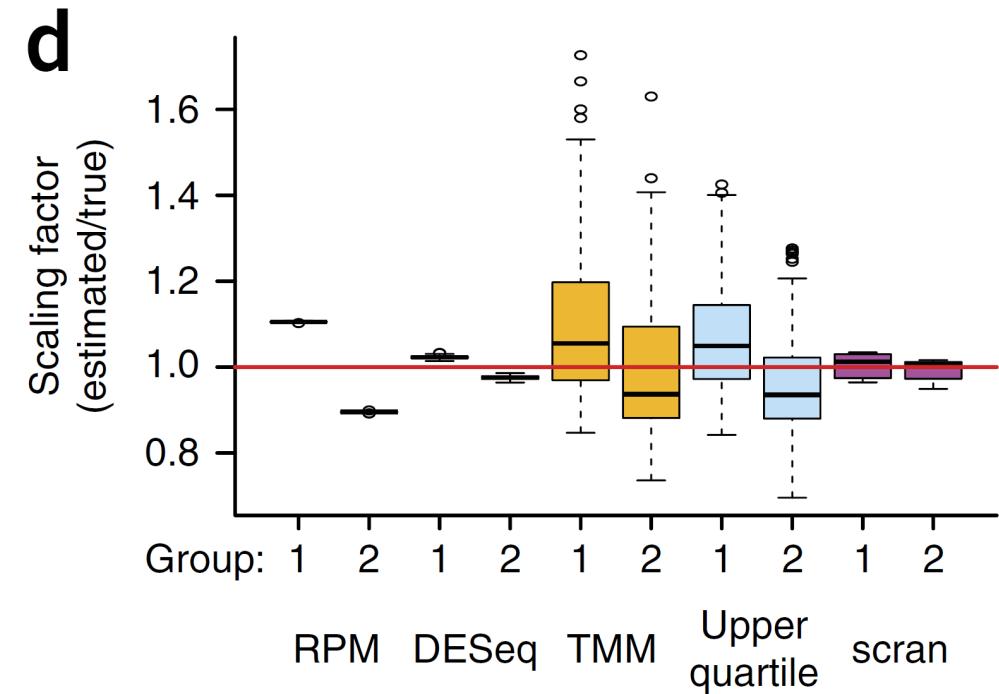
**Step2:** exploit the relationship of model parameter values and gene mean to learn global trends in the data (kernel regression with a normal kernel)

**Step3:** use the regularized regression parameters to transforms UMI counts into Pearson residuals:

$$z_{ij} = \frac{x_{ij} - \mu_{ij}}{\sigma_{ij}},$$
$$\mu_{ij} = \exp(\beta_{0i} + \beta_{1i} \log_{10} m_j),$$
$$\sigma_{ij} = \sqrt{\mu_{ij} + \frac{\mu_{ij}^2}{\theta_i}},$$

# Normalization (5)

- Bulk RNA-based methods: FPKM, CPM, TPM, upperquartile (*NOT APPROPRIATE*)
- Log normalization (Seurat)
- Negative binomial (Monocle)
- Zero-inflated negative binomial (ZINB) models
- scTransform (regularized NB regression)
- ...



Performance Assessment and Selection of  
Normalization Procedures for Single-Cell RNA-Seq  
Cole et al, Cell Systems 2019

# Normalization (pitfalls and recommendations)

- We recommend scran for normalization of non-full-length datasets. An alternative is to evaluate normalization approaches via scone especially for plate-based datasets. Full-length scRNA-seq protocols can be corrected for gene length using bulk methods.
- There is no consensus on scaling genes to 0 mean and unit variance. We prefer not to scale gene expression.
- Normalized data should be  $\log(x+1)$ -transformed for use with downstream analysis methods that assume data are normally distributed.

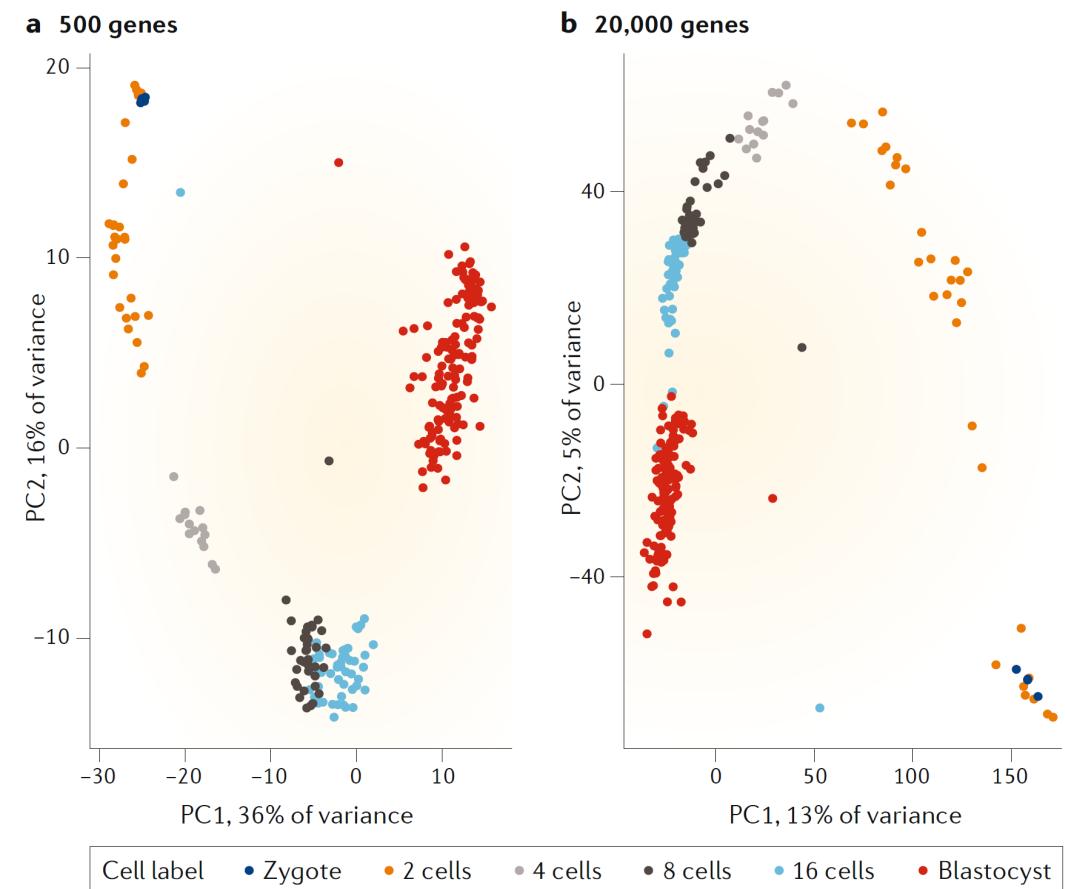
# More on Normalization...

- Svensson, V. Droplet scRNA-seq is not zero-inflated. *Nat. Biotechnol.* **38**, 147–150 (2020).  
<https://doi.org/10.1038/s41587-019-0379-5>
- Sarkar, A., Stephens, M. Separating measurement and expression models clarifies confusion in single-cell RNA sequencing analysis. *Nat Genet* **53**, 770–777 (2021).  
<https://doi.org/10.1038/s41588-021-00873-4>
- Lause, J., Berens, P. & Kobak, D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol* **22**, 258 (2021). <https://doi.org/10.1186/s13059-021-02451-7>

# Feature selection

# Feature selection

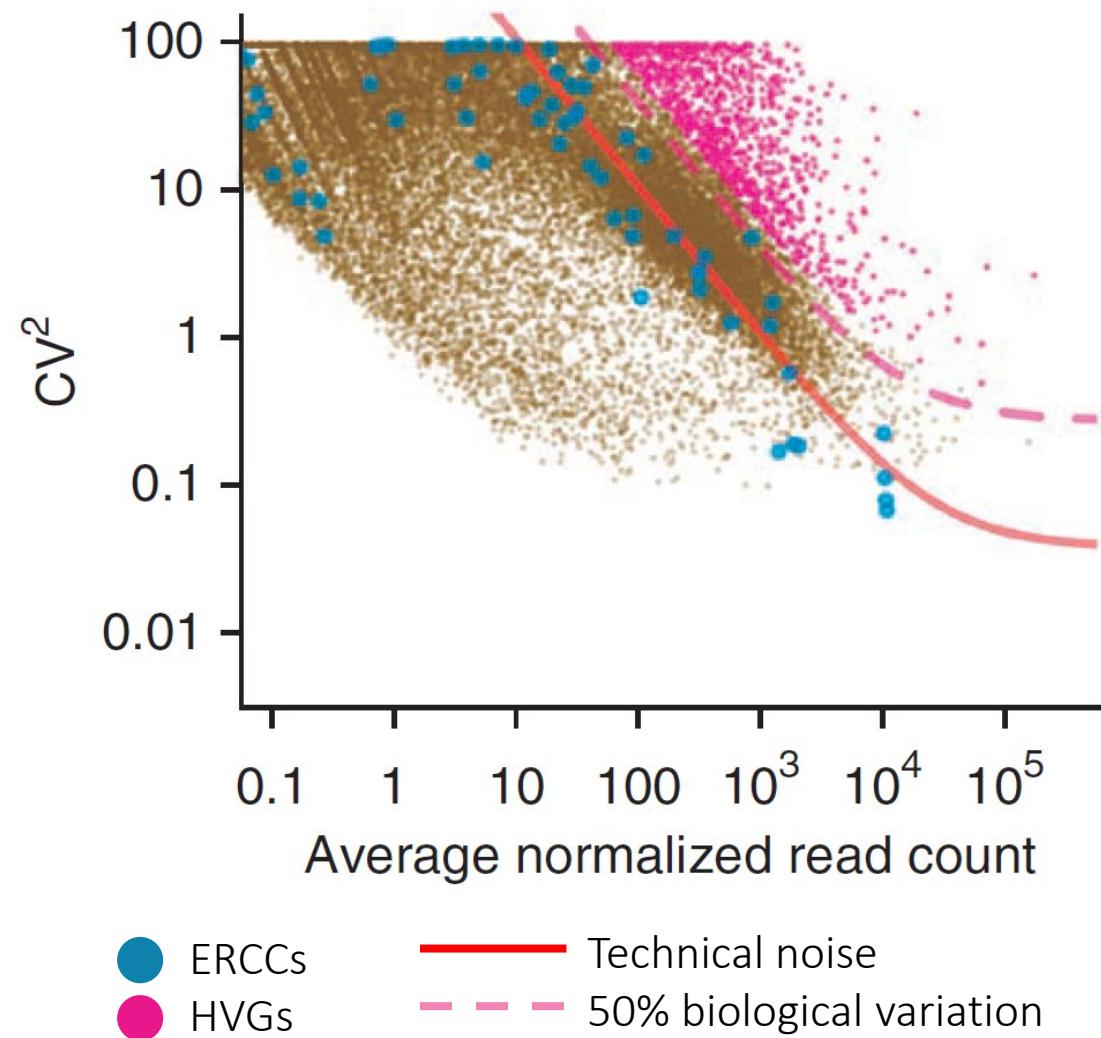
- Curse of dimensionality
  - More features (genes) -> noise dominates distances between samples (cells), effectively all cells get 'same' distance
- Remove genes which only exhibit technical noise
  - Increase the signal:noise ratio
  - Reduce the computational complexity



# Feature selection

## *Highly Variable Genes (HVG)*

- $CV = \frac{var}{mean} = \frac{\sigma}{\mu}$
- Fit a gamma generalized linear model to spike ins (ERCCs)
- No ERCCs?  
Estimate technical noise based on all genes



# Feature selection

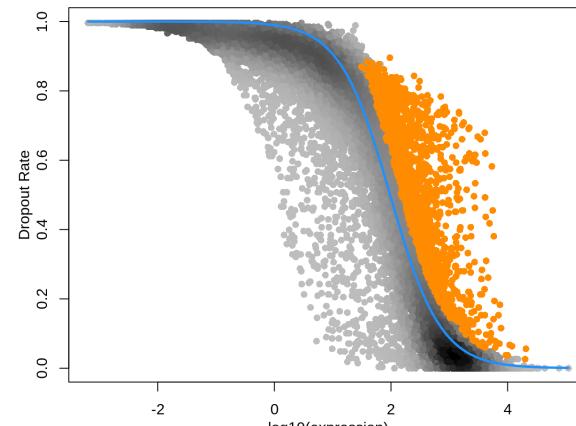
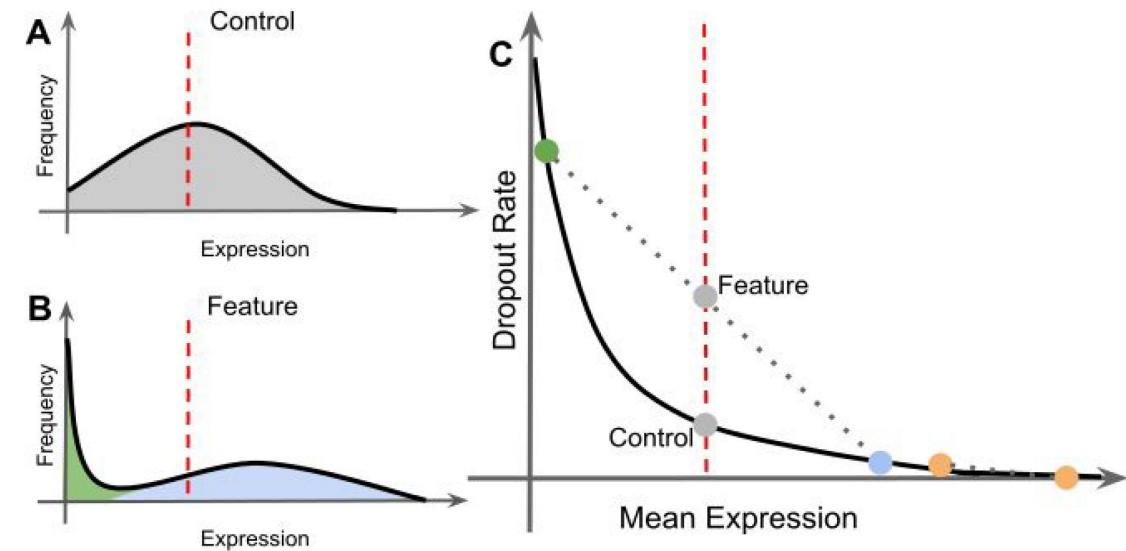
## *M3Drop: Dropout-based feature selection*

- Reverse transcription is an enzyme reaction thus can be modelled using the Michaelis-Menten equation:

$$P_{dropout} = 1 - \frac{S}{K_M + S}$$

$S$ : average expression

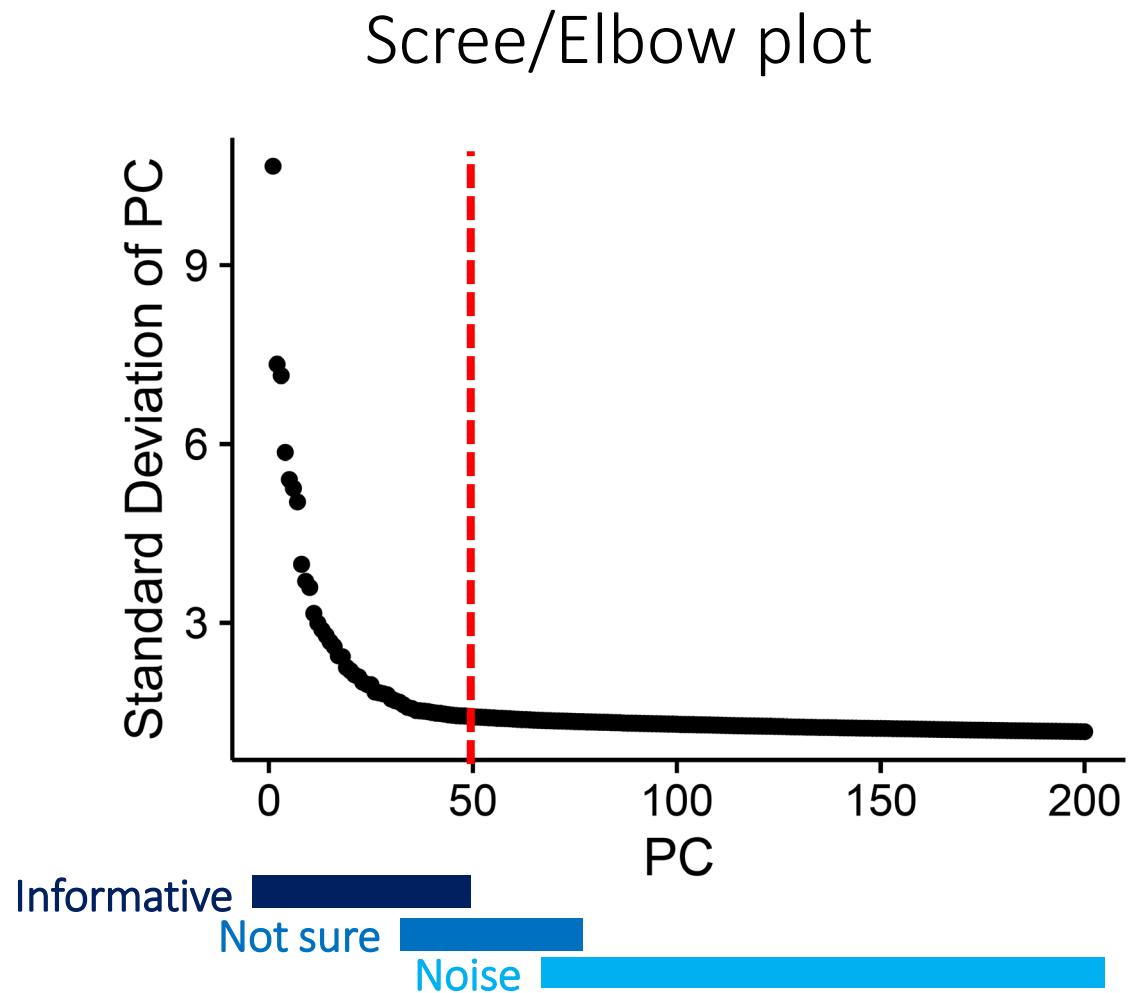
$K_M$ : Michaelis-Menten constant



# Feature selection

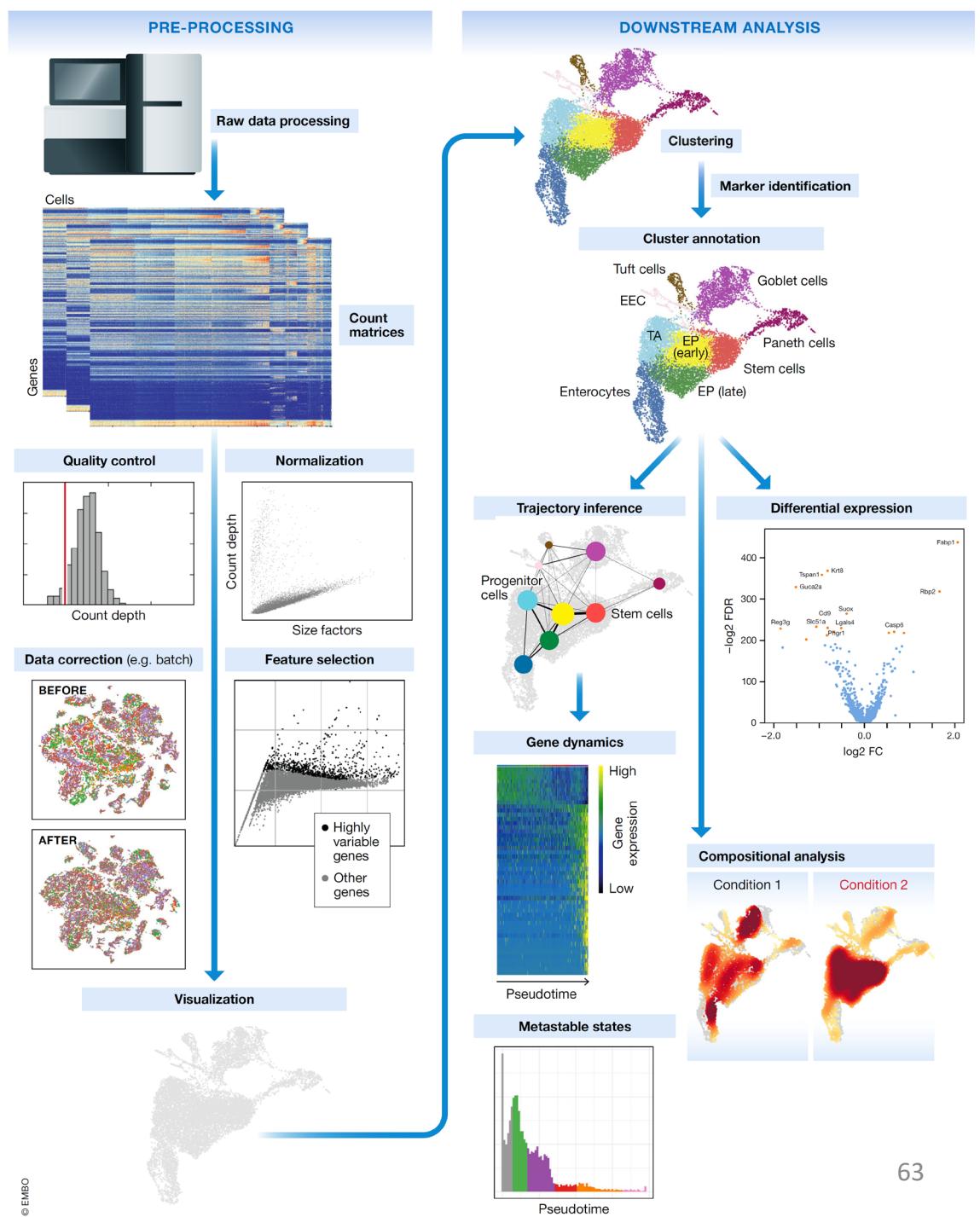
## *Selecting principal components*

- To overcome the extensive technical noise in scRNA-seq data, it is common to cluster cells based on their PCA scores
- Each PC represents a ‘metagene’ that (linearly) combines information across a correlated gene set



# Summary

- Preprocessing:
  - Reads to count matrix
  - Quality control (QC) ✓
  - Normalization ✓
  - Batch correction
  - Feature selection ✓



# Useful Resources

- Best practices in single cell RNA-seq analysis (Luecken & Theis, MSB 2019)

<https://www.embopress.org/doi/pdf/10.15252/msb.20188746>

- Orchestrating Single-Cell Analysis with Bioconductor

<https://osca.bioconductor.org/>

- Single Cell Course (Martin Hemberg Lab, Wellcome Trust Sanger):

<http://hemberg-lab.github.io/scRNA.seq.course>

- Aaron Lun's single cell workflow (very detailed):

<https://www.bioconductor.org/packages/release/workflows/html/simpleSingleCell.html>

- GitHub: Awesome Single Cell

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

- Recent developments in single cell genomics

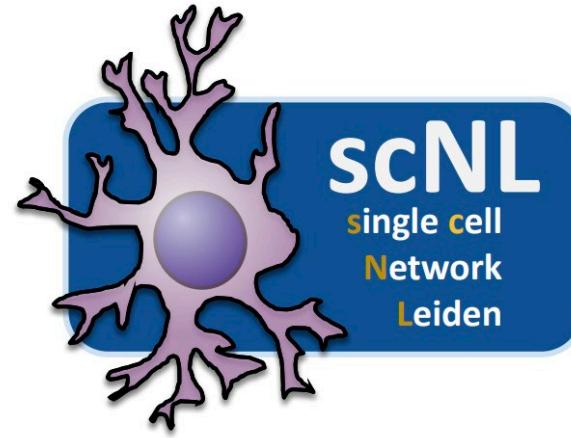
[https://www.dropbox.com/s/woya6ffgq8a3pkw/SingleCellGenomicsDay18\\_References.pdf?dl=1](https://www.dropbox.com/s/woya6ffgq8a3pkw/SingleCellGenomicsDay18_References.pdf?dl=1)

# Practicals

- You received an invite for RStudio Cloud
- Click on the assignment to start
- Work in pairs, find your partner in the General channel under the Files tab
- Ask your questions in the right Questions\_Lab(DAY) channel
- Check in the schedule which TAs are available for each session
- RStudio Cloud will be available until 18 Nov 2021

# Thank You!

 a.mahfouz@lumc.nl  
 mahfouzlab.org  
 @ahmedElkoussy



<https://www.singlecell.nl/>