

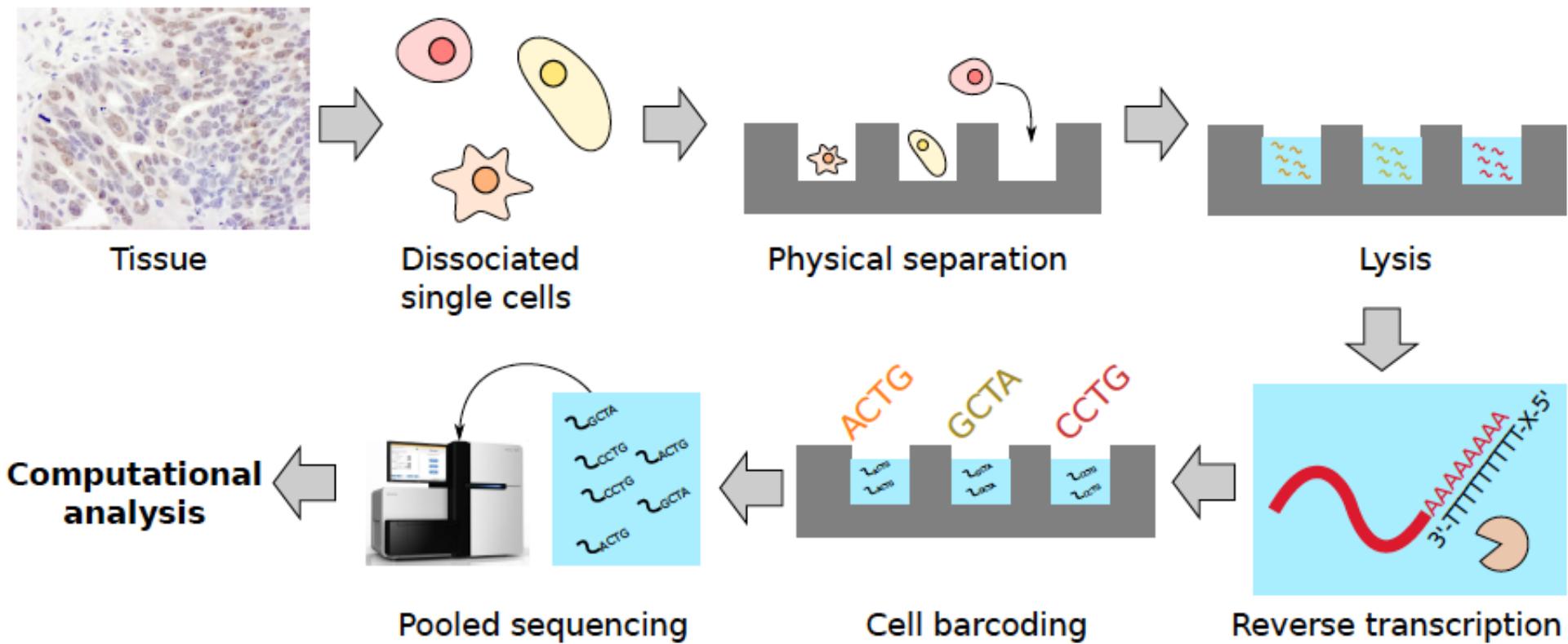
Quality Control, Normalization, and Feature Selection

Ahmed Mahfouz

Department of Human Genetics, LUMC

Pattern Recognition and Bioinformatics, TU Delft

Single cell RNA-sequencing (scRNA-seq)

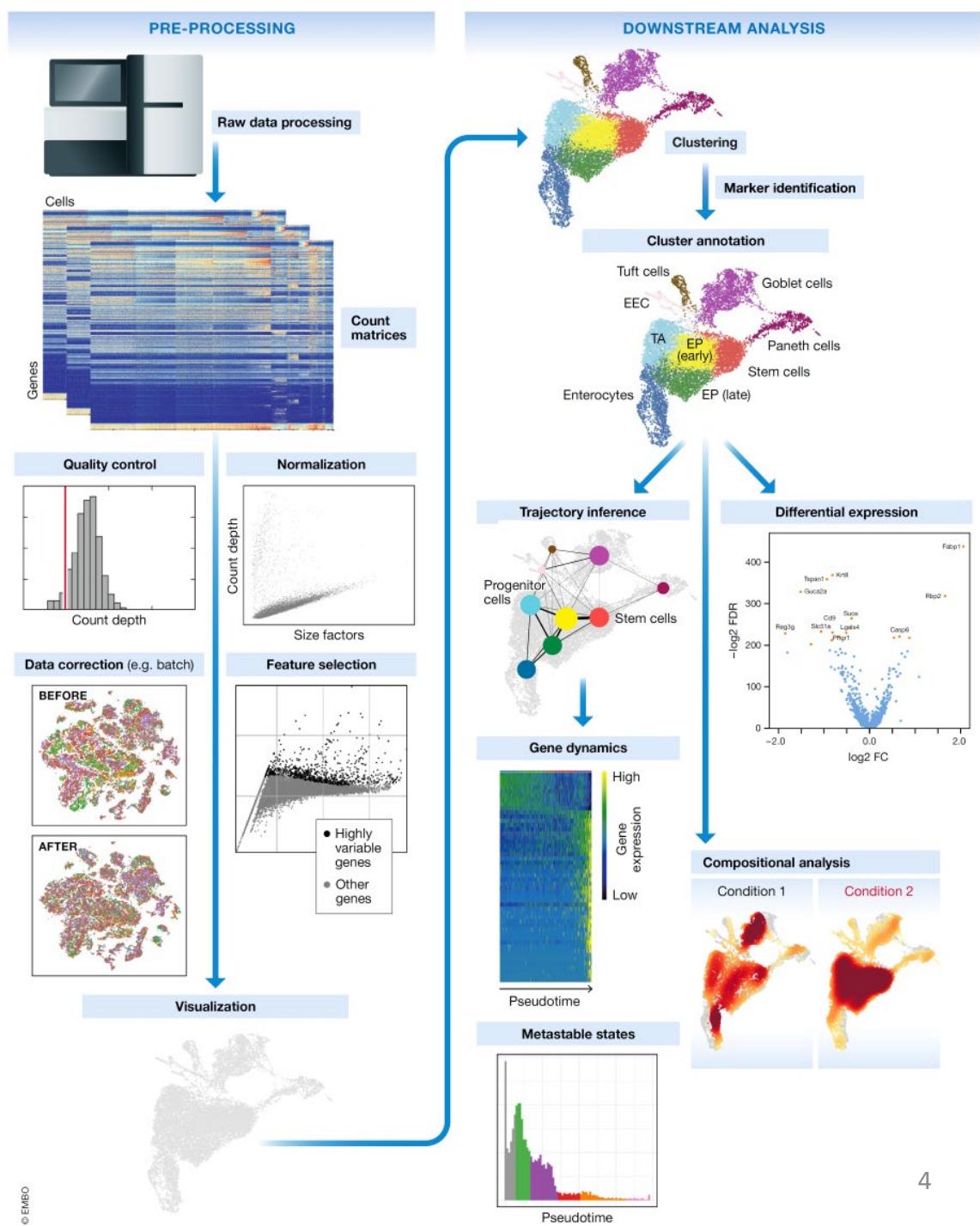


scRNA-seq Data Analysis

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

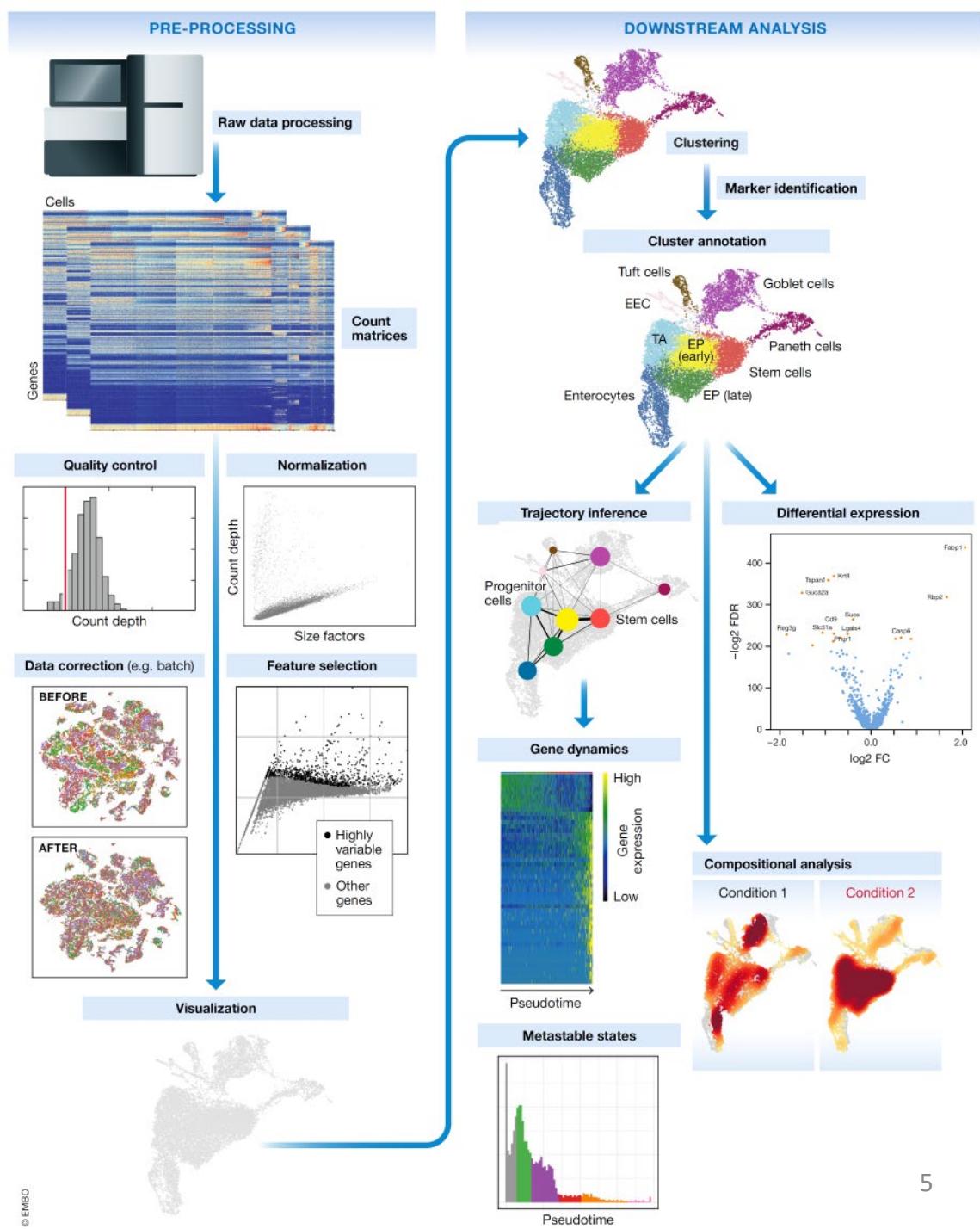
scRNA-seq Data Analysis

- 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

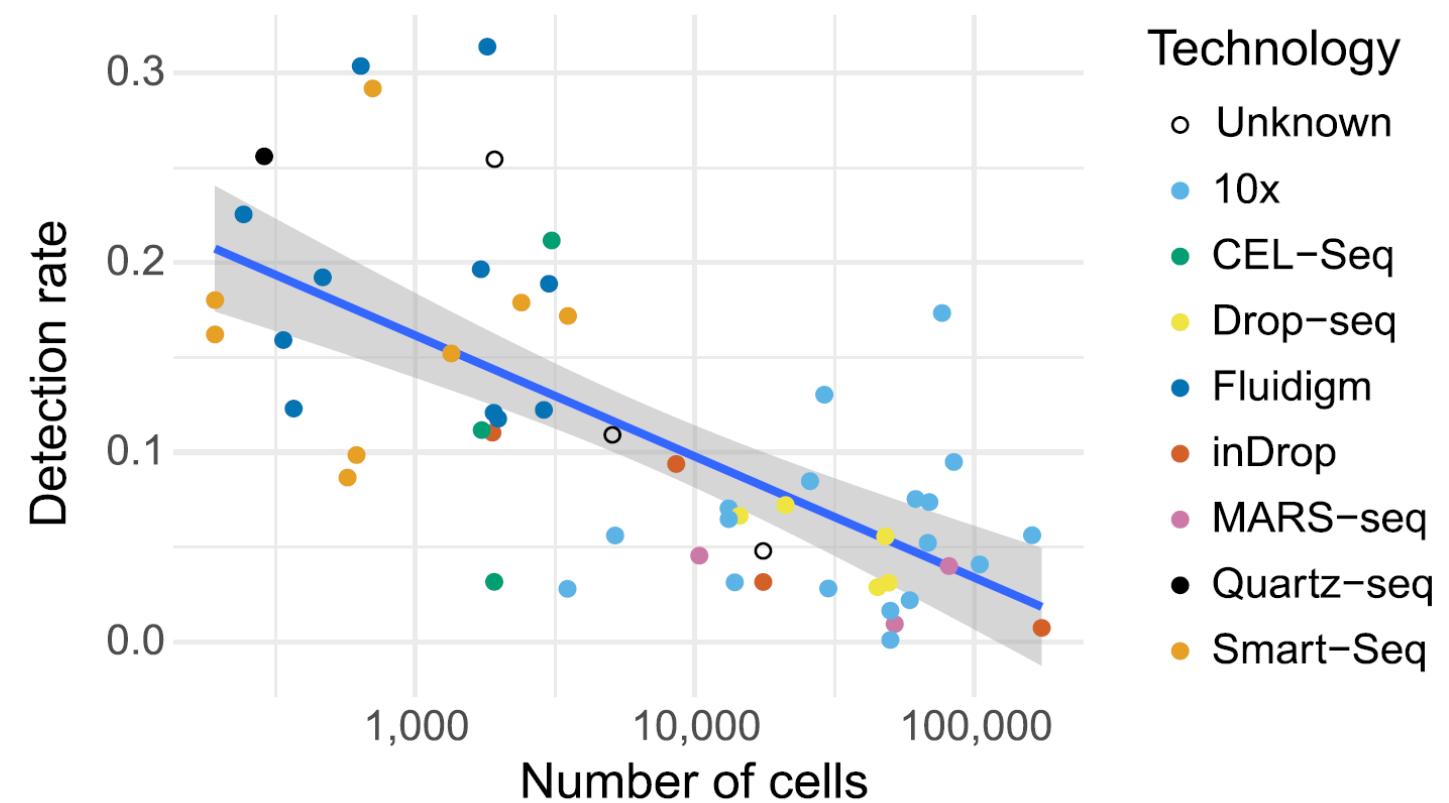
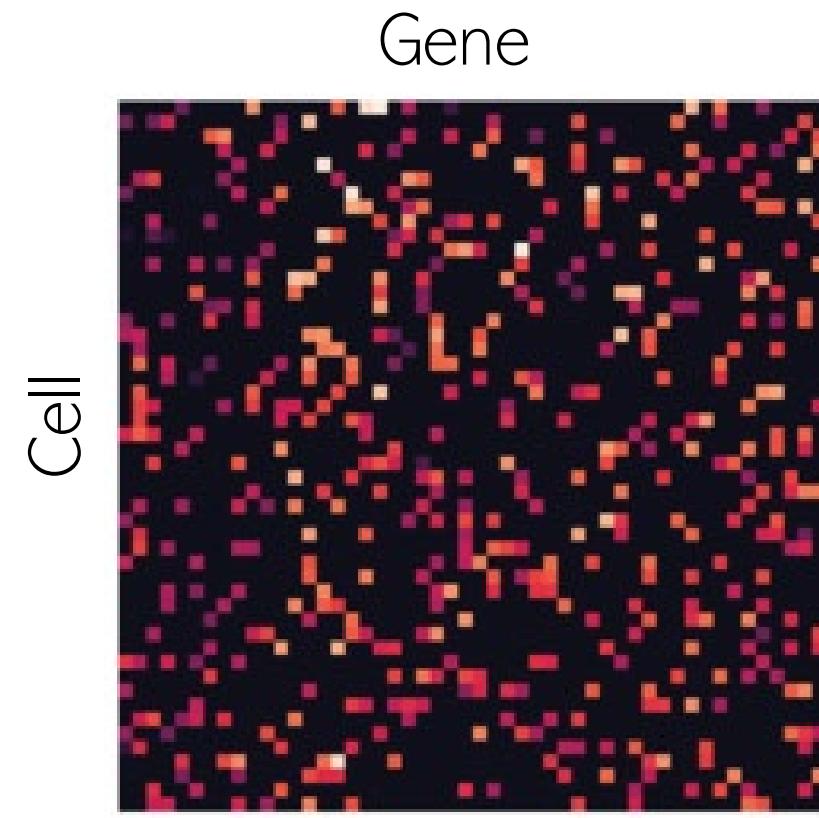
- Preprocessing:
 - Reads to count matrix 
 - Quality control (QC)
 - Normalization
 - Batch correction
 - Feature selection
- Downstream
 - Cell type identification (clustering/classification)
 - Trajectory inference
 - Differential expression
 - Compositional analysis
 - Co-expression network analysis



Our agenda

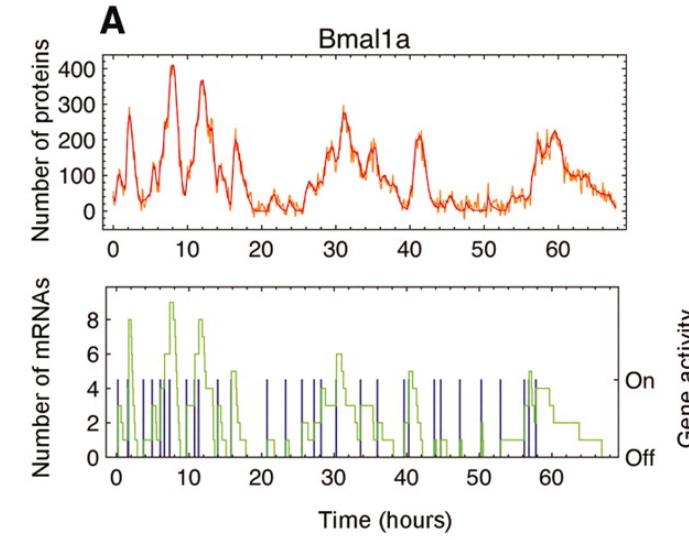
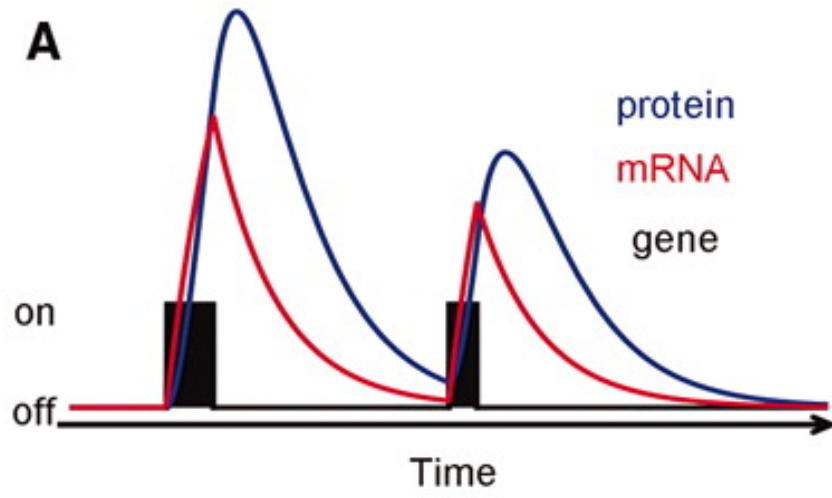
- Background on transcriptional bursting & drop-outs
- Experimental setup – what could go wrong?
- Quality control
- Normalization
- Feature selection

scRNA-seq data is mainly zeros

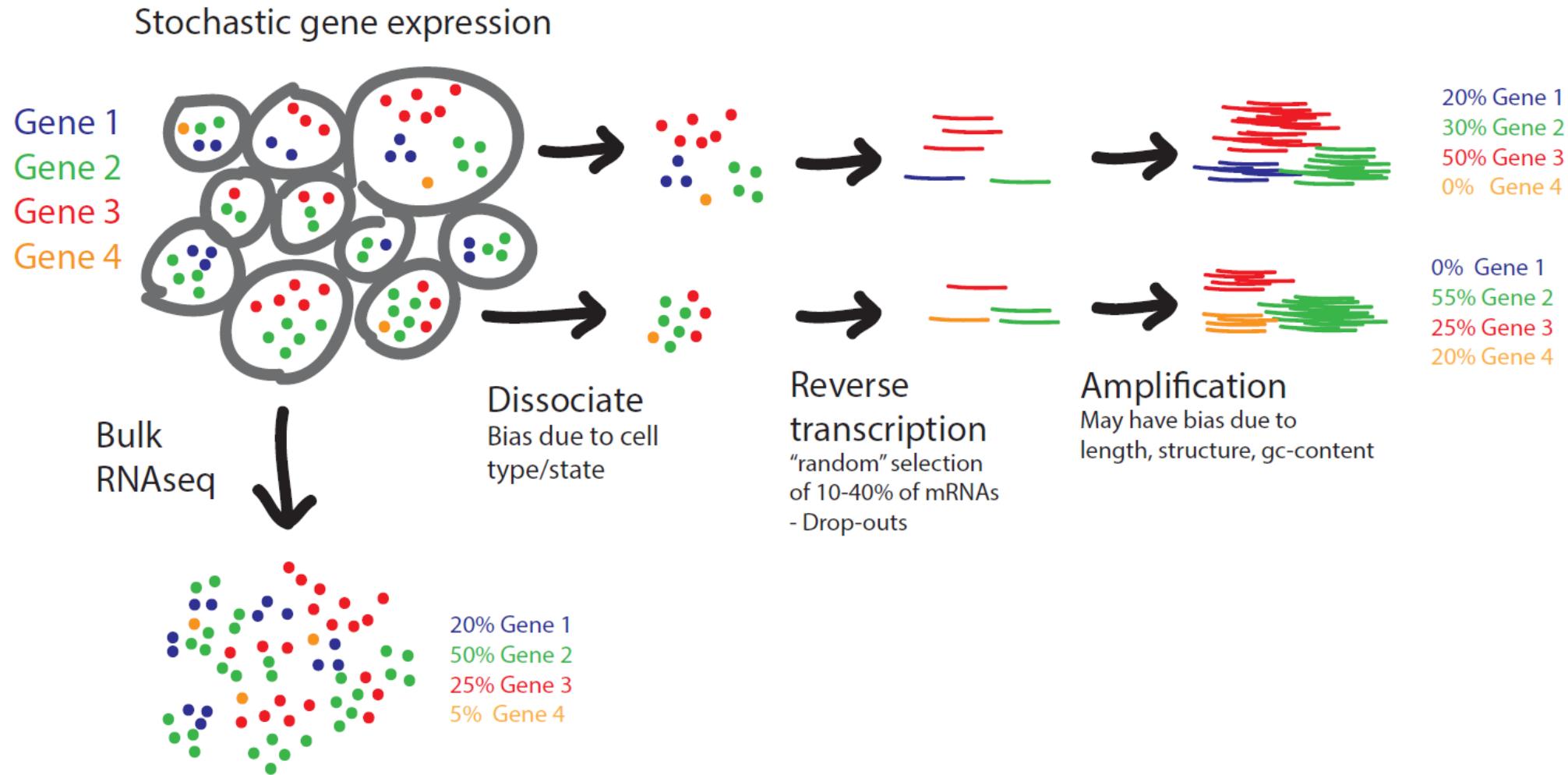


Transcriptional bursting

- 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



What could have gone wrong?

Cell dissociation

Cell capture

Cell lysis

Reverse transcription

Preamplification

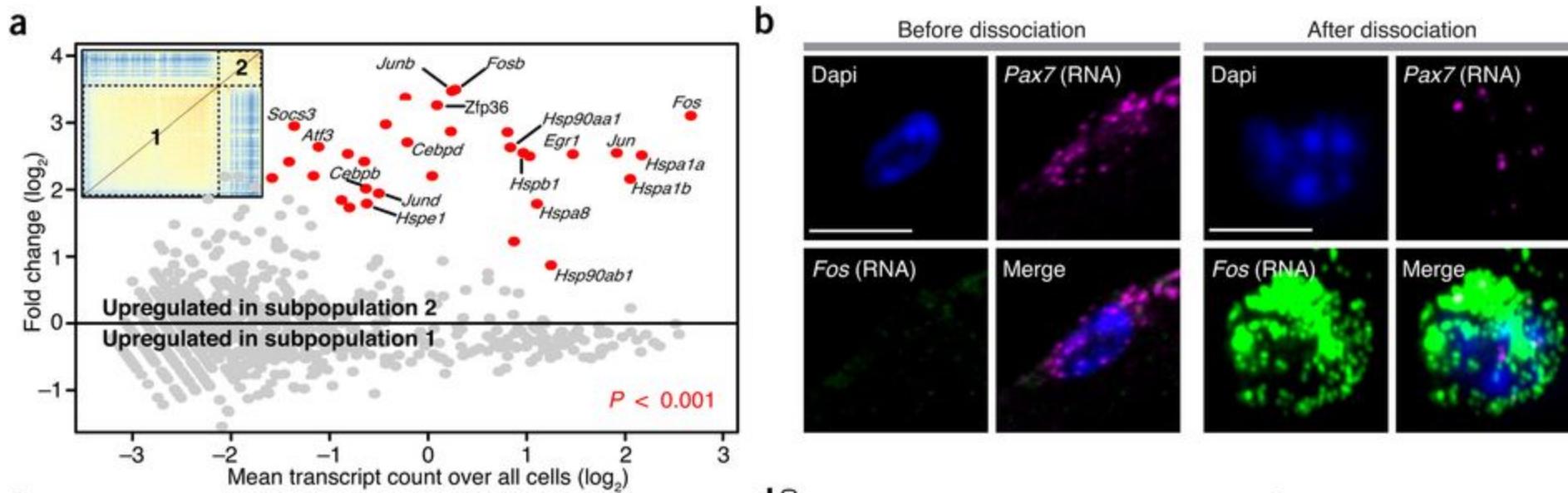
Library preparation and sequencing

Cell dissociation

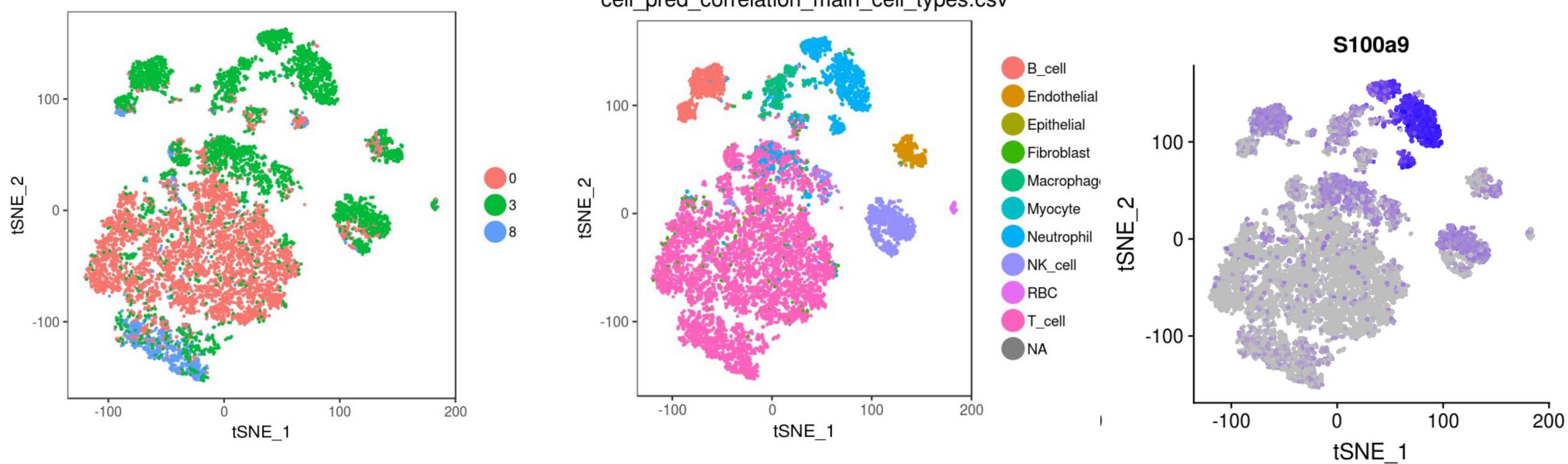
- It is critical to have healthy whole cells with no RNA leakage. Short time from dissociation to cell!
- Tissues that are hard to dissociate:
 - Laser capture microscopy (LCM)
 - Nuclei sorting
- PROBLEMS:
 - Incomplete dissociation can give multiple cells sticking together.
 - Too harsh dissociation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



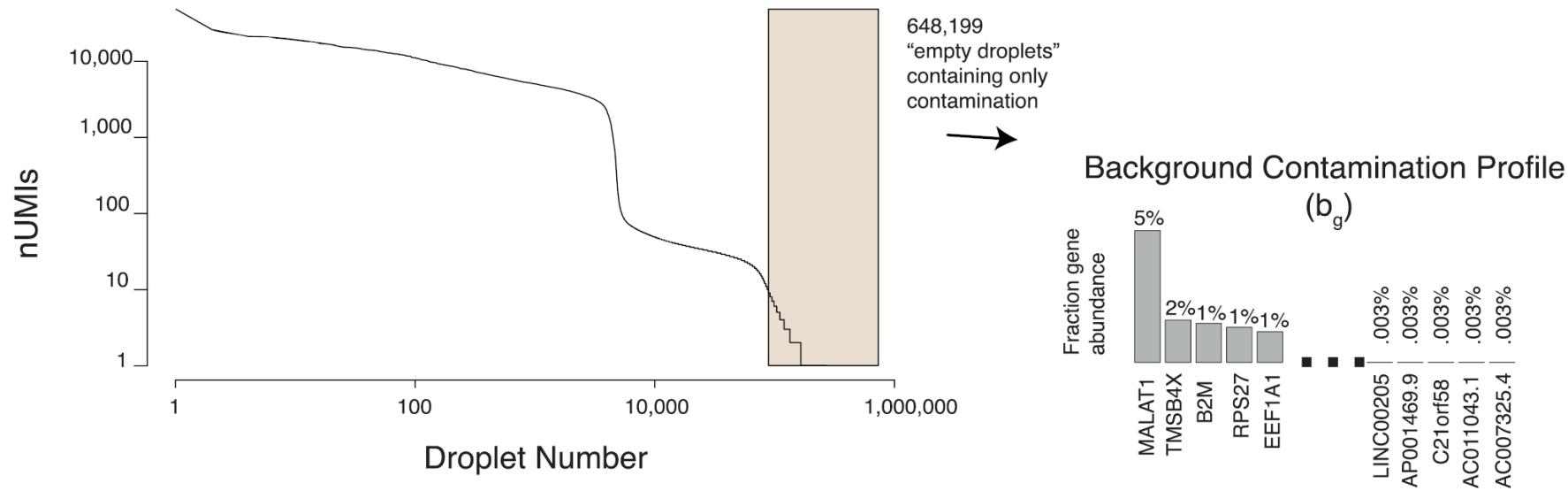
Ambient RNA



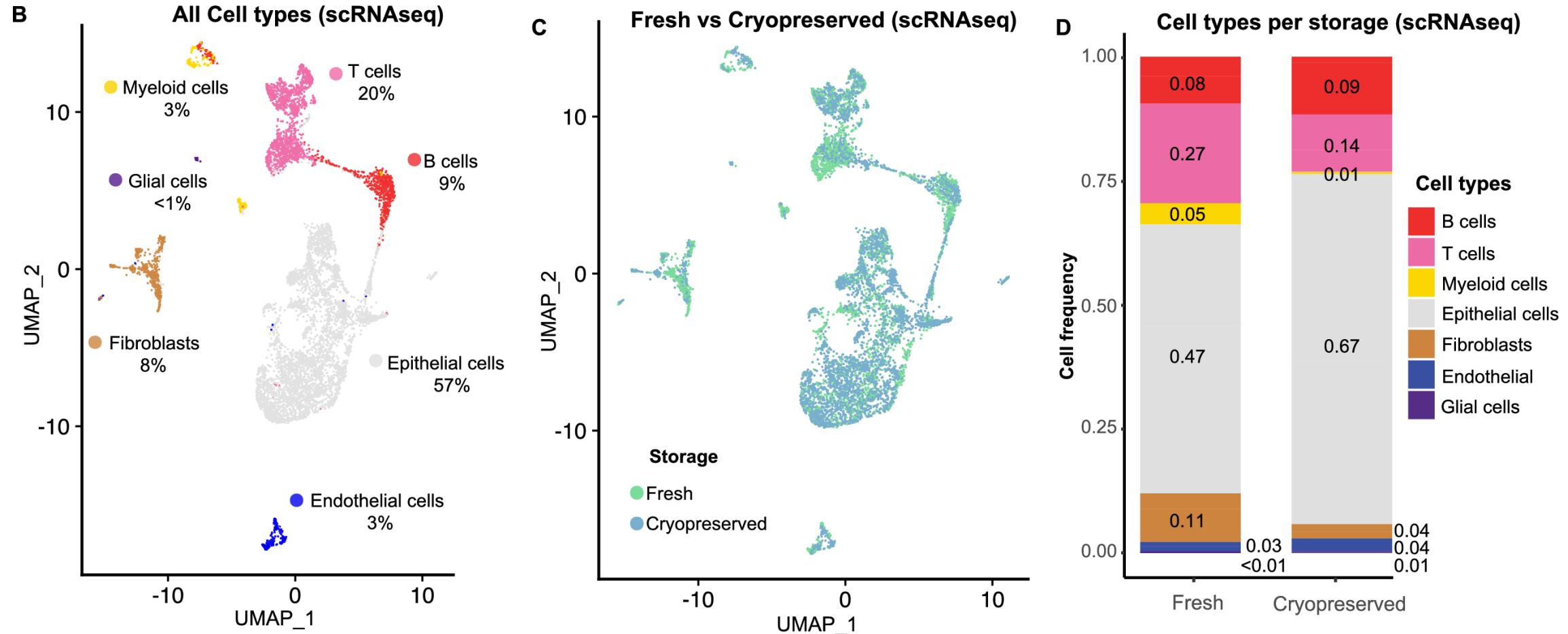
Sample from Day3 have detection of Neutrophil markers in all cells -> possibly contamination from ambient RNA.

Ambient RNA

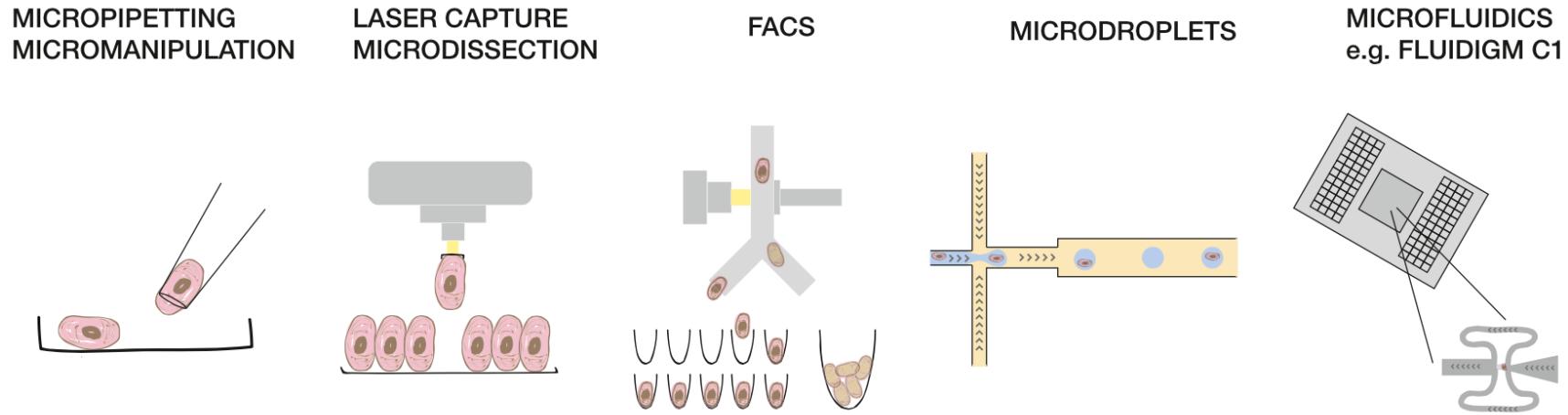
- Using empty droplets, estimate background signal
- SoupX (Young MD, GigaScience 2020)
- Cellbender (Flemming et al. BiorXiv 2022)



Biased distribution due to dissociation

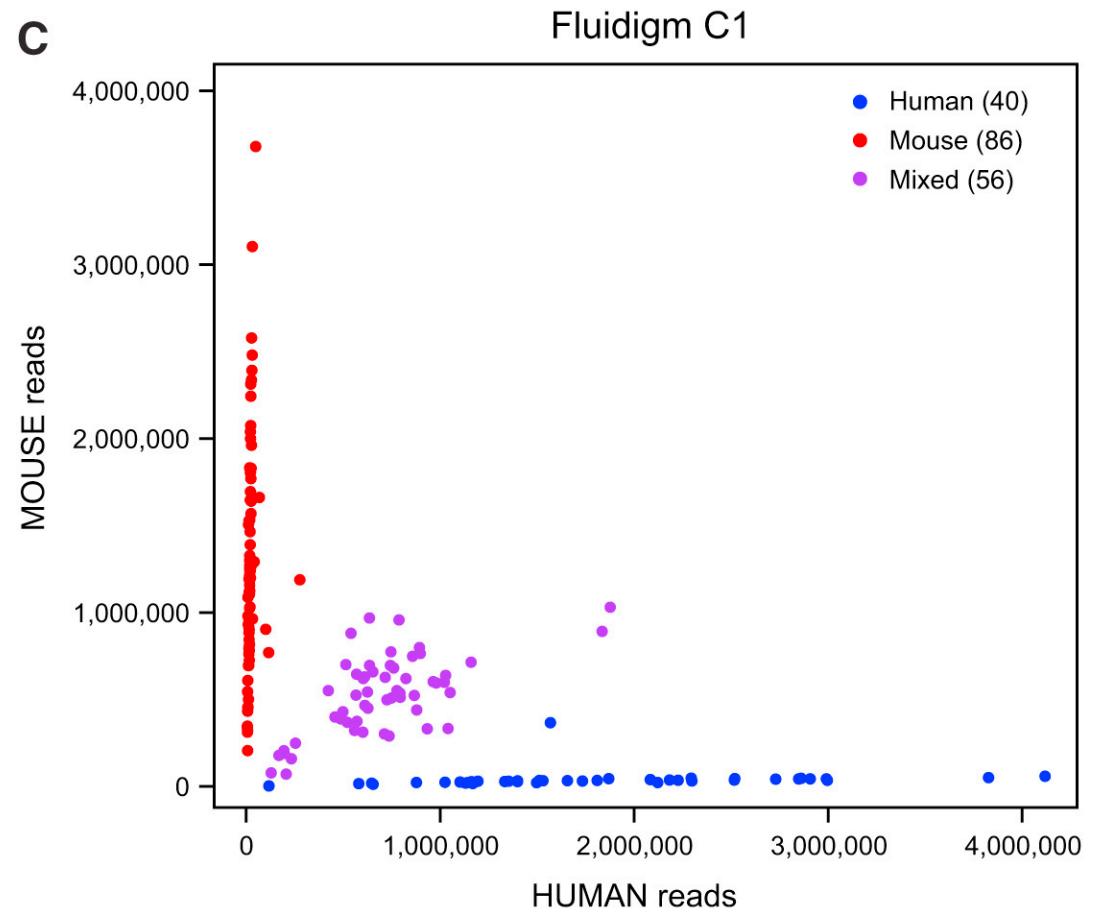
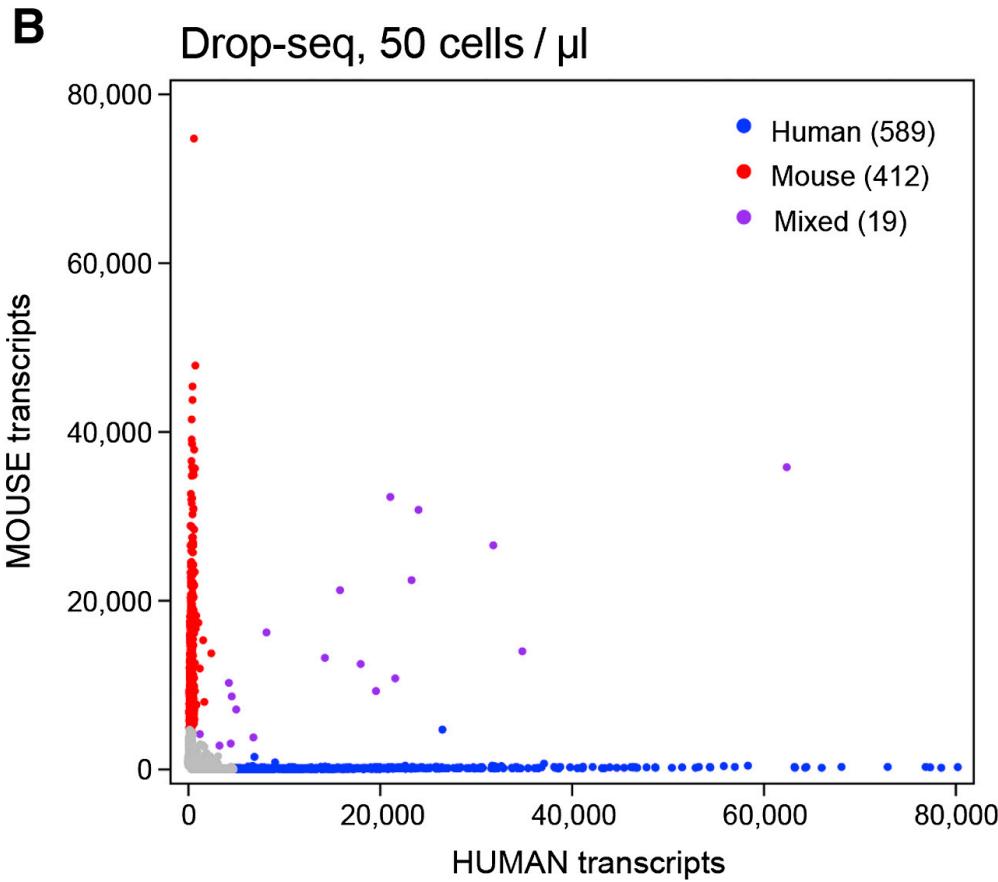


Single cell capture



- 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



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

Doubles

- 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

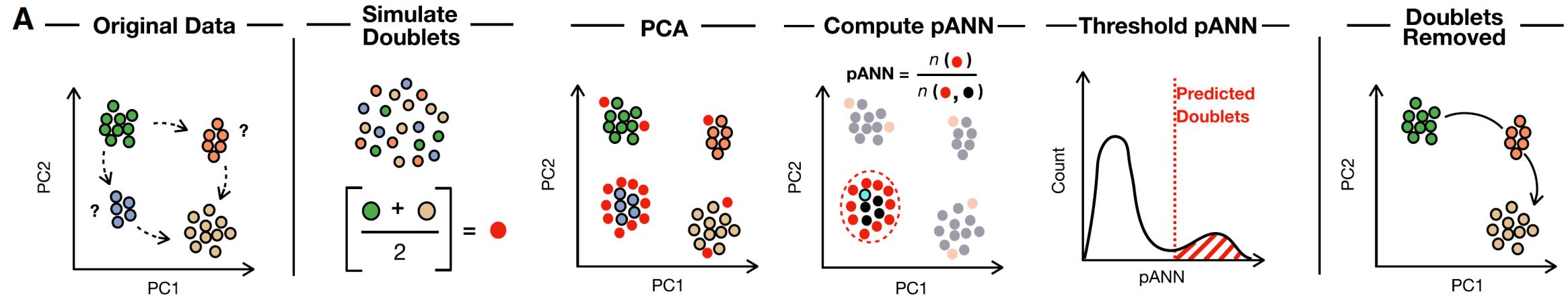
<https://github.com/AllonKleinLab/scrublet>

- DoubletDecon

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

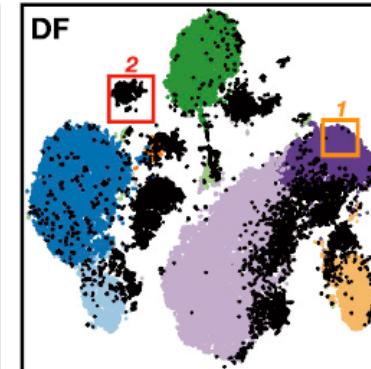
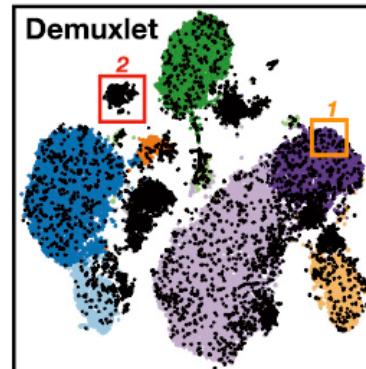
- DoubletCluster / DoubletCell in Scran

DoubletFinder

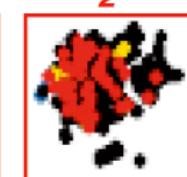
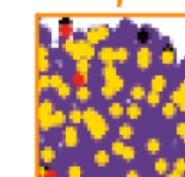


E

- Doublets
- CD14 Mono
- CD16 Mono
- CD4T
- CD8T
- NK
- DC
- MK
- B



	vs. Ground-Truth	TP	FN	FP
Ground-Truth (Demuxlet)	Dub.	Dub.	Sin.	Dub.
DoubletFinder	Dub.	Sin.	Dub.	



Other factors affecting quality

- Cell Lysis:
 - Optimal lysis conditions may vary from celltype to celltype and for nuclei vs cells.
- Reverse transcription:
 - Drop-out rate is around 90-60% depending on the method used
- Preamplification
 - Methods that uses UMIs will control for amplification bias to a large extent, but the chance of detecting a transcript that is amplified more is higher.

Cells vs nuclei

snRNAseq pros

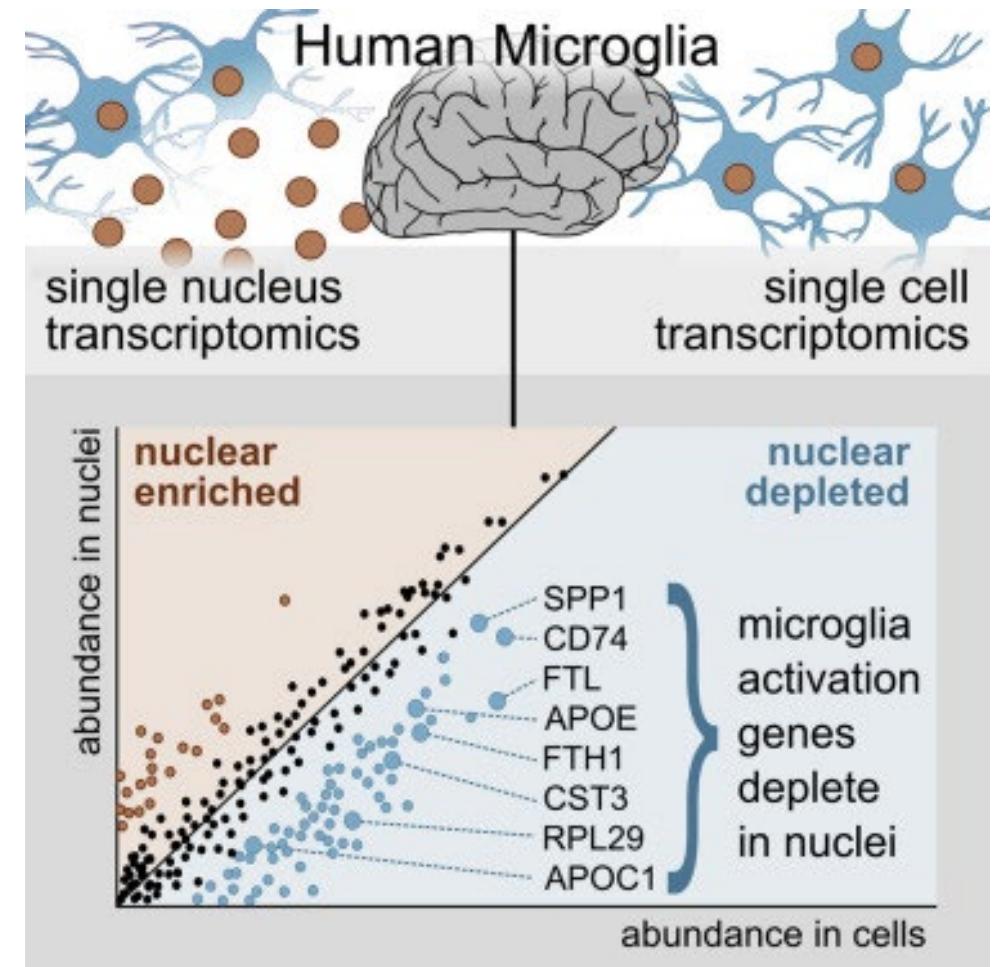
- Can avoid some biases due to dissociation.
- Hard to dissociate celltypes (e.g. neurons, muscle fibres, adipocytes)
- Frozen tissues

snRNAseq cons

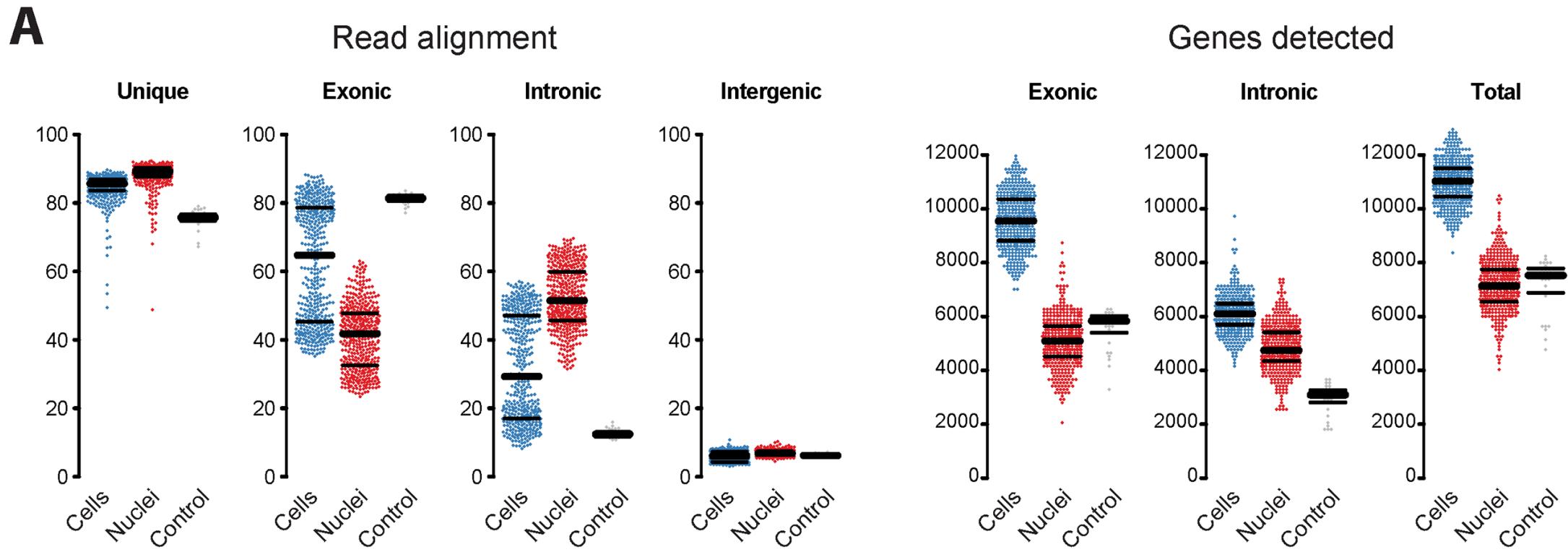
- Less mRNA per nuclei
- More dominated by nuclear lincRNAs
- Internal priming of polyA stretches in introns

Cells vs nuclei

- For some cell types there may be biased detection of genes in nuclei vs cell.



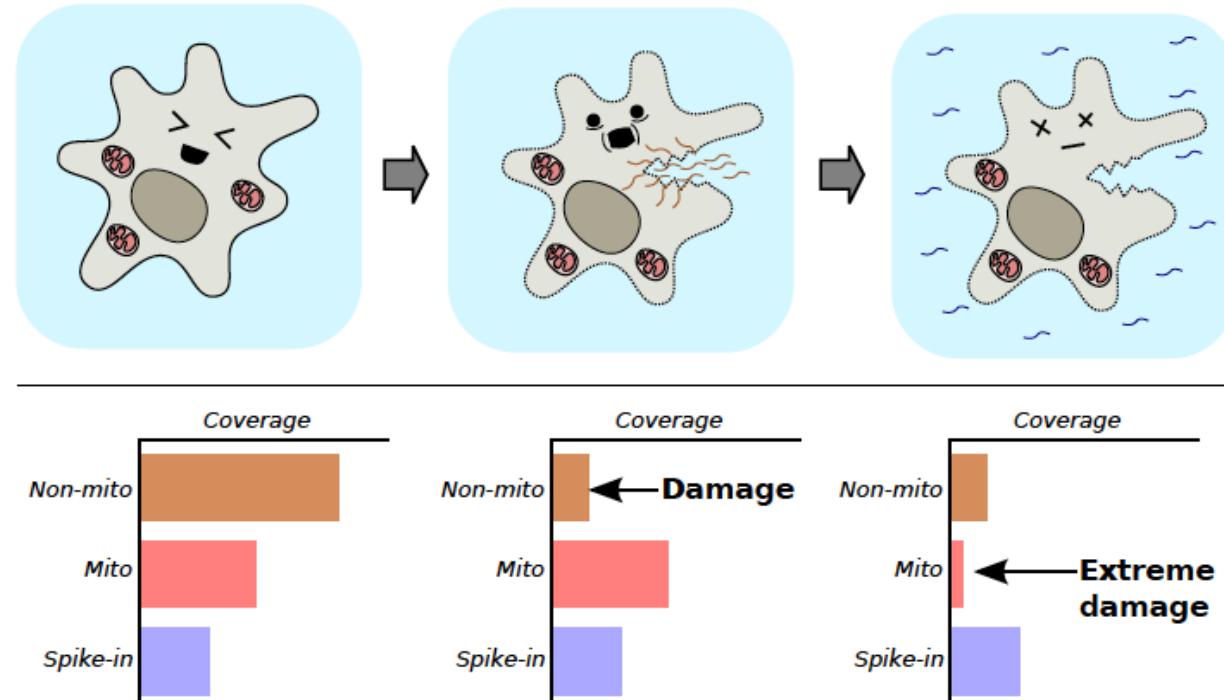
Cells vs nuclei



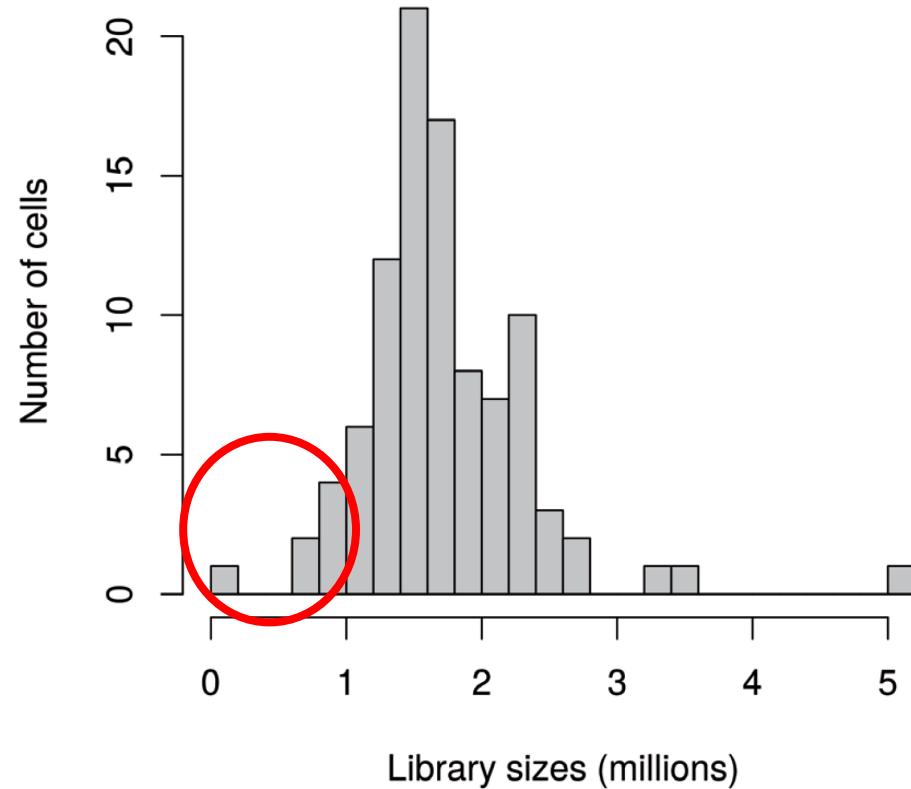
- Usually need to include intronic counts to increase transcript detection in snRNAseq
- Is now default in CellRanger also for scRNAseq

Quality control of cells (1)

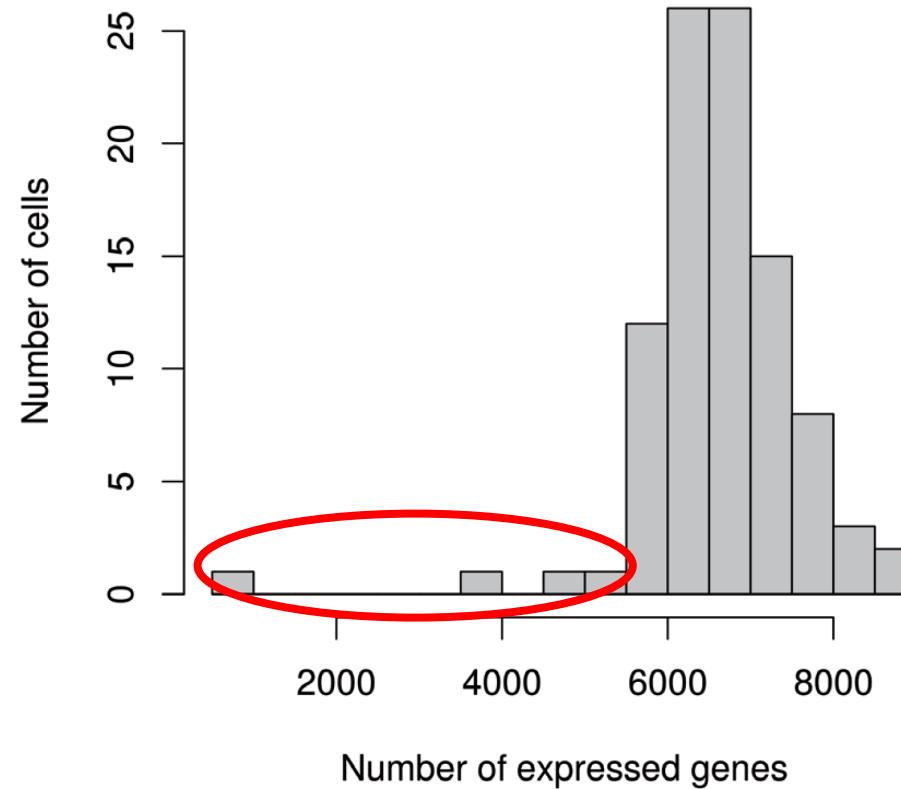
- Low sequencing depth
- Low numbers of expressed genes (i.e. any nonzero count)
- High mitochondrial content



Quality control of cells (2)

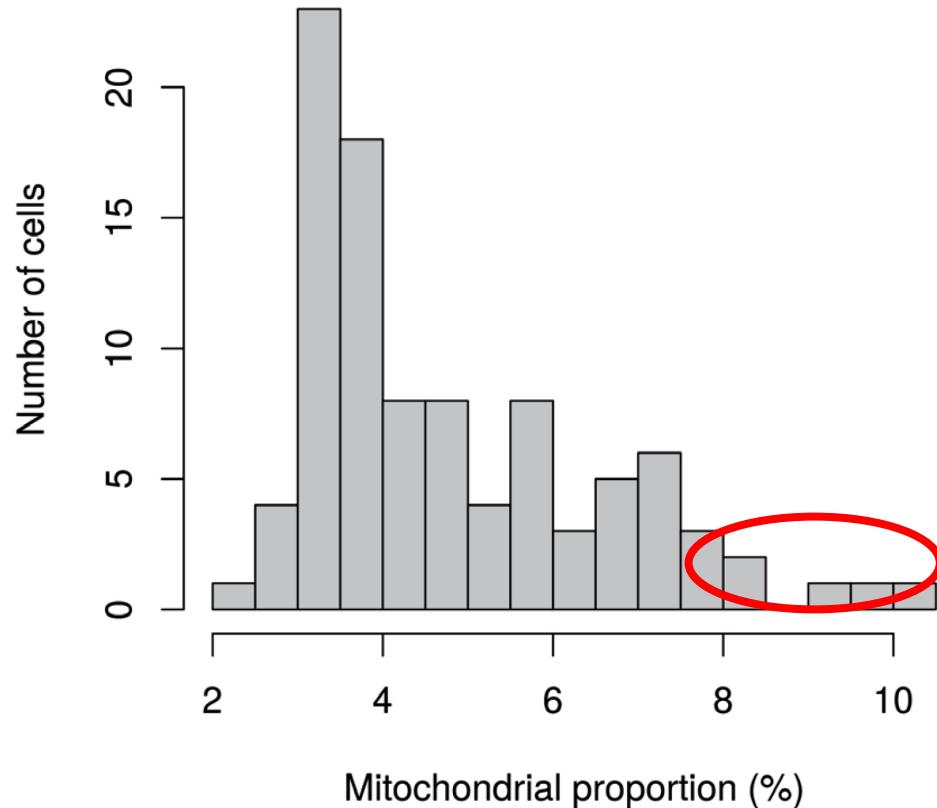


RNA has not been efficiently captured during library preparation



Diverse transcript population not captured

Quality control of cells (3)



Ribosomal RNA read fraction
Possible that degradation of
RNA leads to more templating of
rRNA-fragments.

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.

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

Sources of variation

Biological

Cell type/state

Cell cycle

Cell size

Sex, Age, ...

...

Technical

Cell quality

Library prep efficiency

Batch effects

...

Sources of variation

Biological

- Cell type/state
- Cell cycle
- Cell size
- Sex, Age, ...
- ...

Technical

- Cell quality
- Library prep efficiency
- Batch effects
- ...

To identify cell types we would like to remove all other sources of variation

UMIs do not solve everything

C

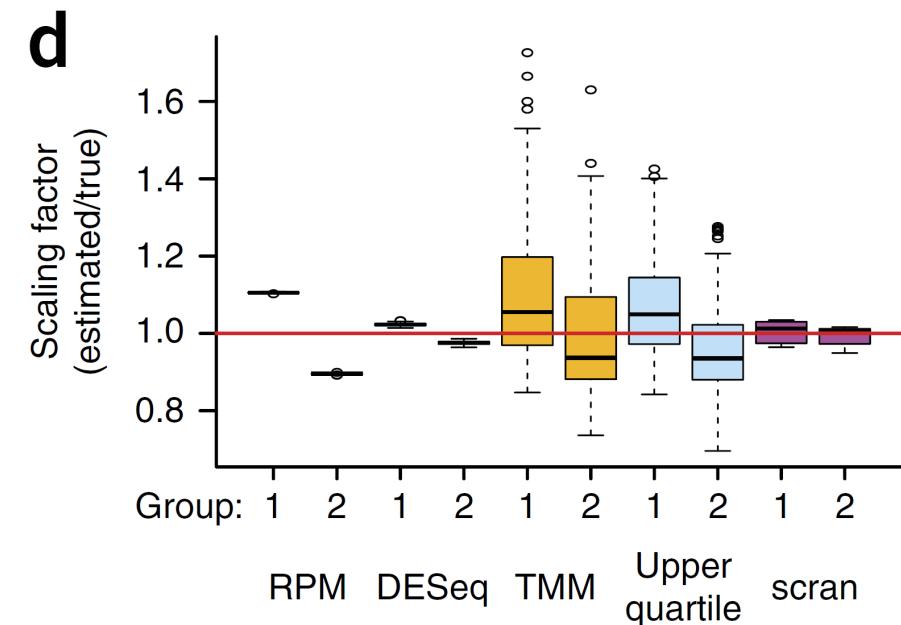
	Cell-specific effects	Gene-specific effects	Not removed by UMIs
Sequencing depth	✓		✓
Amplification	✓	✓	
Capture and RT efficiency	✓	✓	✓
Gene length		✓	
GC content	✓	✓	✓
mRNA content	✓		✓

Normalization

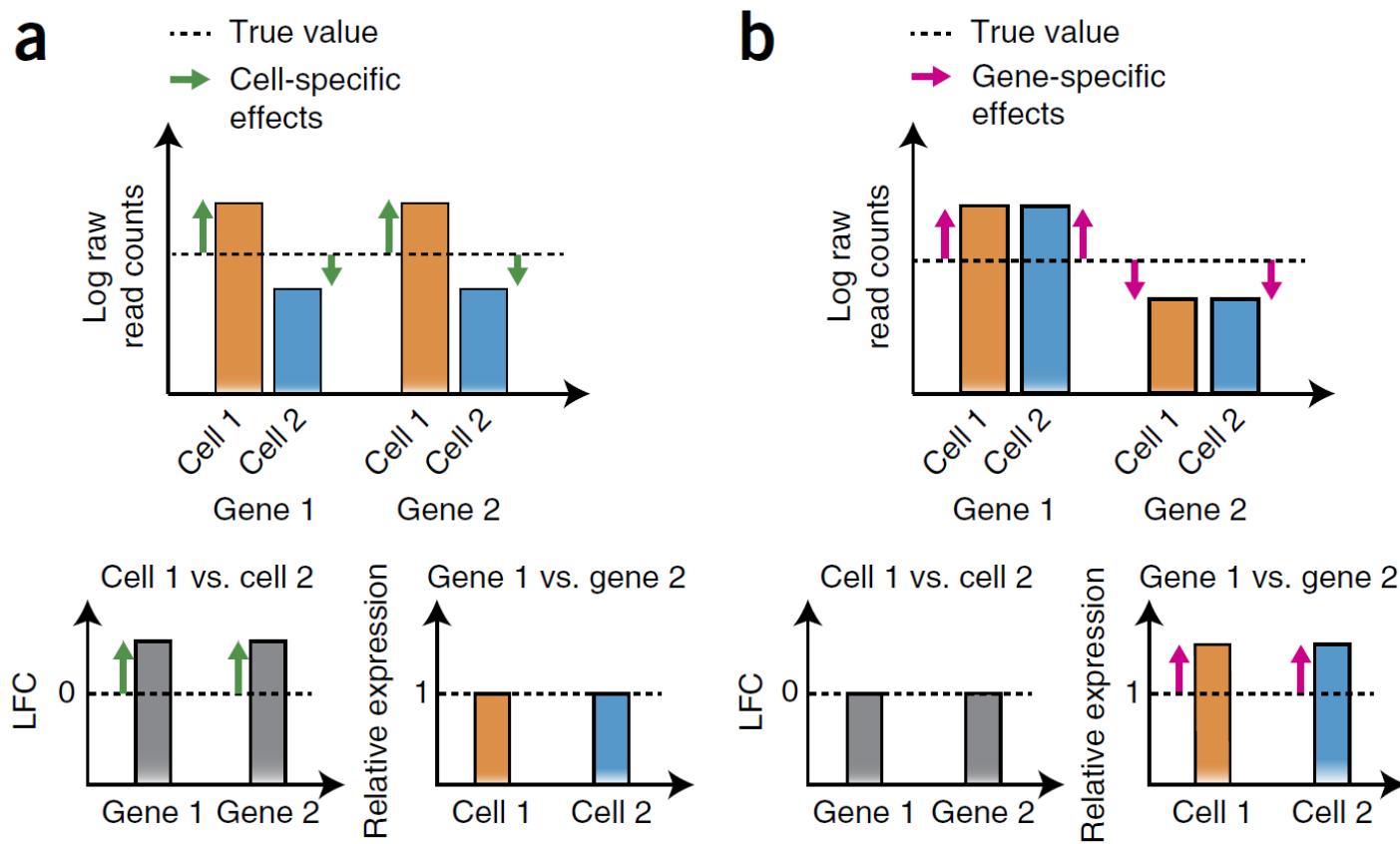
- Want to make expression comparable across samples, cells and genes.
- Involves 3 main steps:
 - Scaling
 - Transformation
 - Removal of unwanted variation

What is different from bulk RNA-seq?

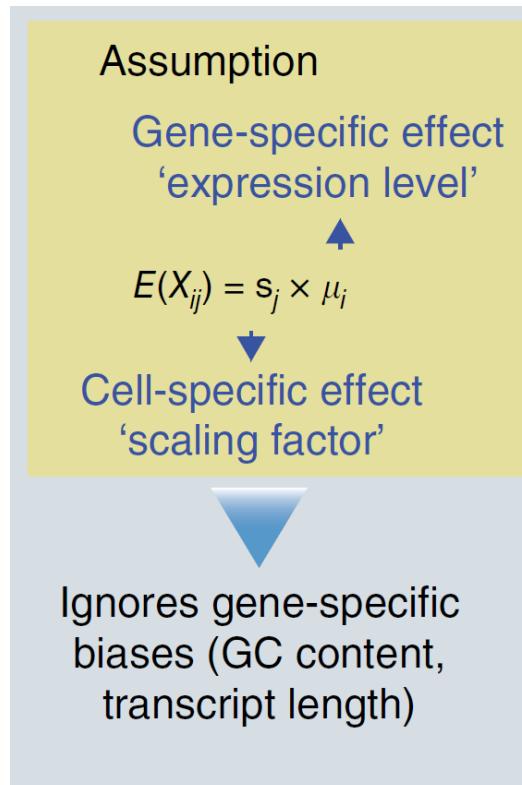
- Noise
 - Low mRNA content per cell
 - Variable mRNA capture
 - Variable sequencing depth
- Different cell types in the same sample
- Bulk RNA-seq normalization methods (FPKM, CPM, TPM, upperquartile) are not suitable



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



Normalization (1)



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)

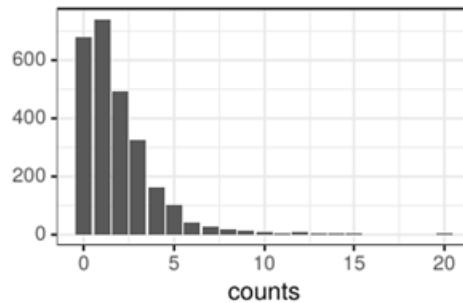
Log-normalization

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

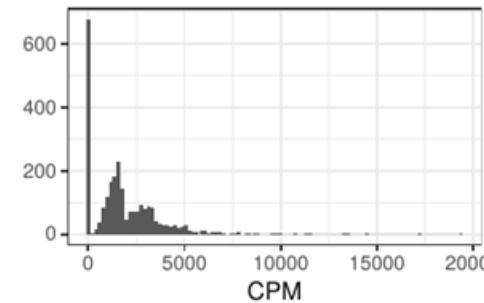
- Simplest and most commonly used normalization strategy
- Divide all counts for each cell by a cell-specific scaling factor (i.e. size factor)
- Assumes that any cell-specific bias (e.g., in capture or amplification efficiency) affects all genes equally via scaling of the expected mean count for that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

Effect of dropouts on normalization

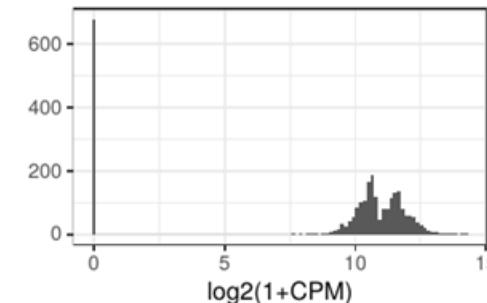
Inflation of zero counts



(a) UMI counts

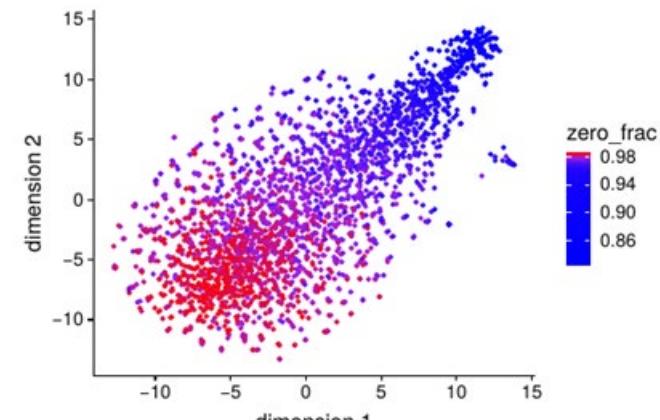
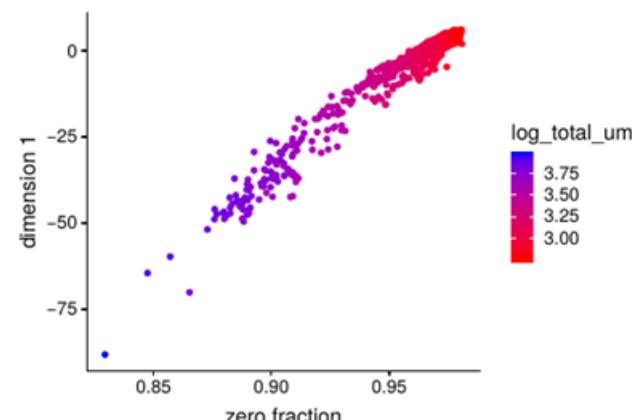


(b) counts per million (CPM)



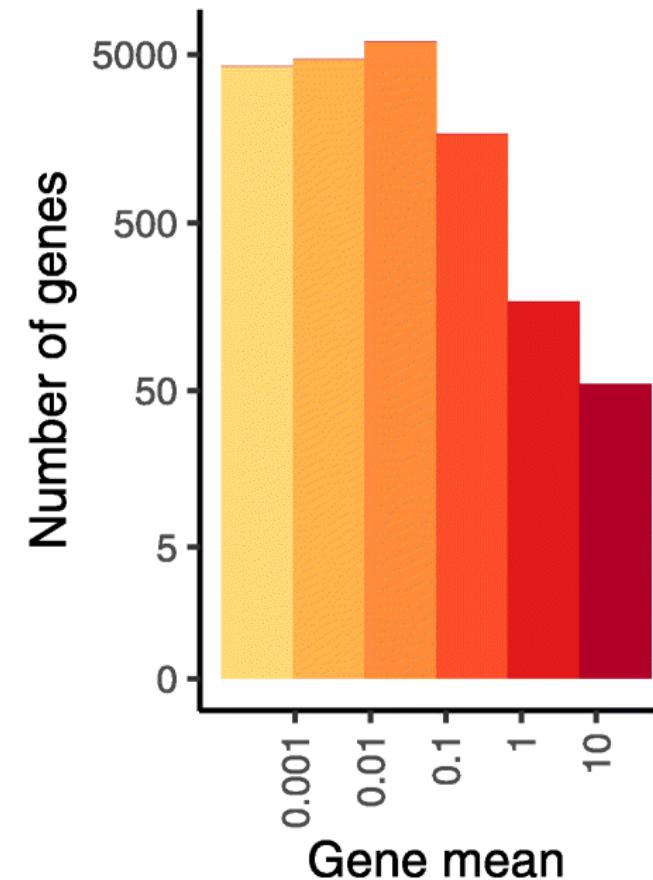
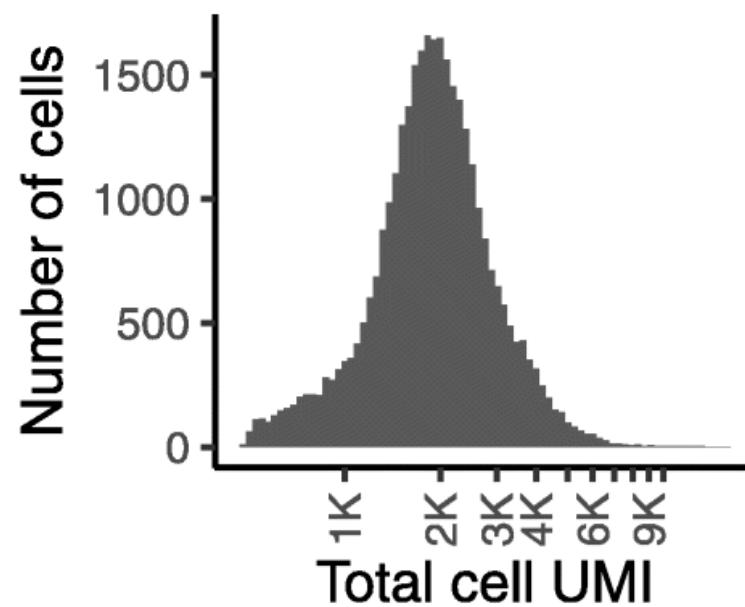
(c) log of CPM

Fraction of zeros become main source of variability



Does log-normalization (scaling) work?

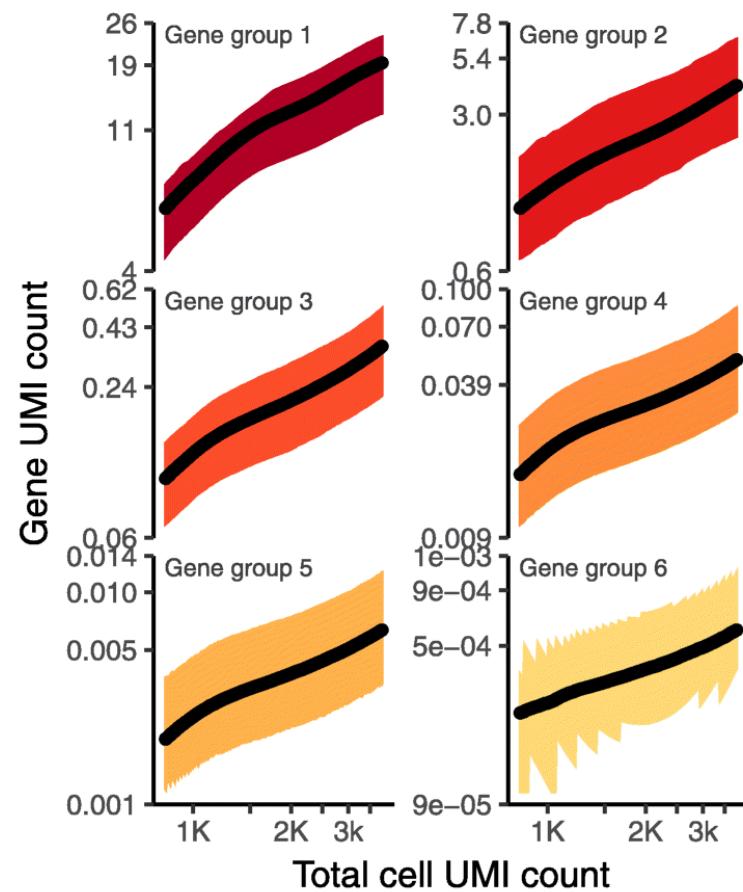
33,148 PBMCs, 10x Genomics
16,809 genes detected \geq 5 cells



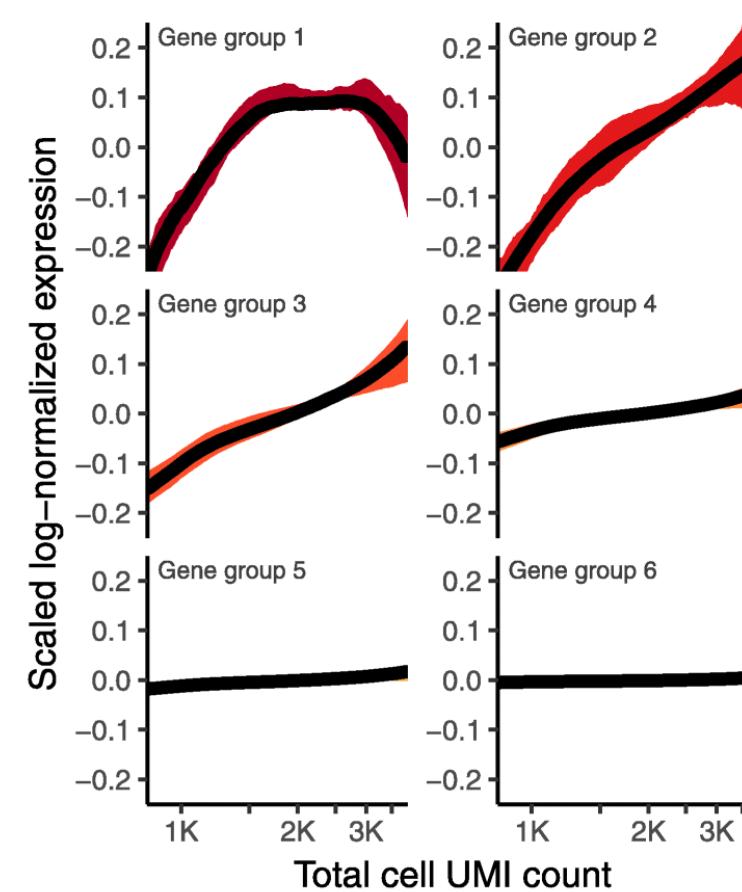
Gene group ID, size	
1,	55
2,	171
3,	1687
4,	5942
5,	4694
6,	4260

Does log-normalization (scaling) work?

Before normalization



After normalization



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 μ and variance $\mu + \frac{\mu^2}{\sigma}$

Scaling data (Z-score transformation)

- Z-score transformation: linearly transform data to a mean of zero and a standard deviation of 1
- PCA or any other type of analysis will be dominated by highly expressed genes with high variance.
- It can be wise to center and scale each gene before performing PCA

Which normalization method to use?

- Normalization has big impact on differential gene expression, but not as much on clustering
- In most cases it is enough to do sequence depth normalization and log-normalization.
- Binning by gene level (SCTransform) helps to remove the effect of different gene detection across cells.

Confounding factors

- Any source of variation that you do not expect to give separation of the cell types:
 - Cell cycle, size, sequencing depth, mitochondrial reads,...
- Regress these sources of variation out from your data
- Tools like SCTransform, ZIMB-WaVE does regression in the same step.
- **BUT**, be careful that your confounders are not related to your biological question!

A different view...

- For cell-based analysis (dimensionality reduction, clustering, cell type identification), binarized data is sufficient

Counts

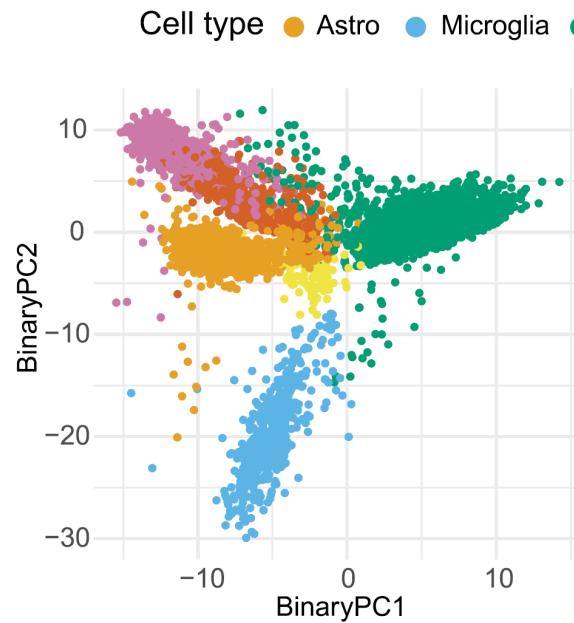
0	0	0	0	0	66	0
3	0	8	24	0	20	15
0	0	0	11	0	102	0
0	4	10	32	0	0	11
1	0	0	49	0	100	2
0	0	10	0	0	75	21
0	1	0	0	0	100	0

Detection

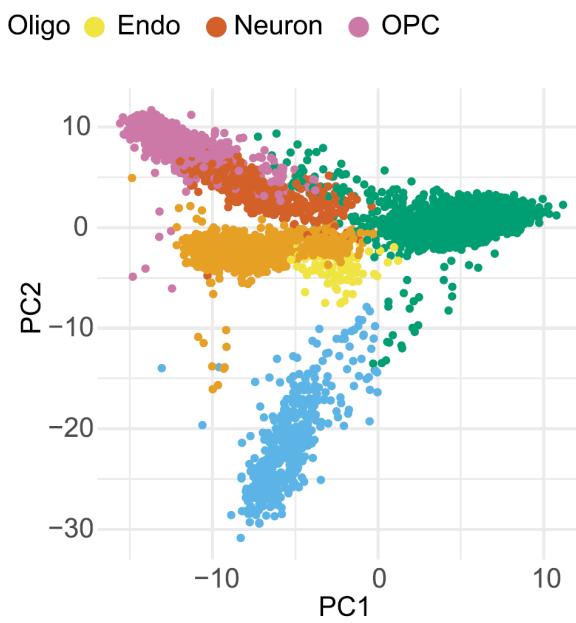
0	0	0	0	0	1	0
1	0	1	1	0	1	1
0	0	0	1	0	1	0
0	1	1	1	0	0	1
1	0	0	1	0	1	1
0	0	1	0	0	1	1
0	1	0	0	0	1	0

Binary representation of scRNA-seq data

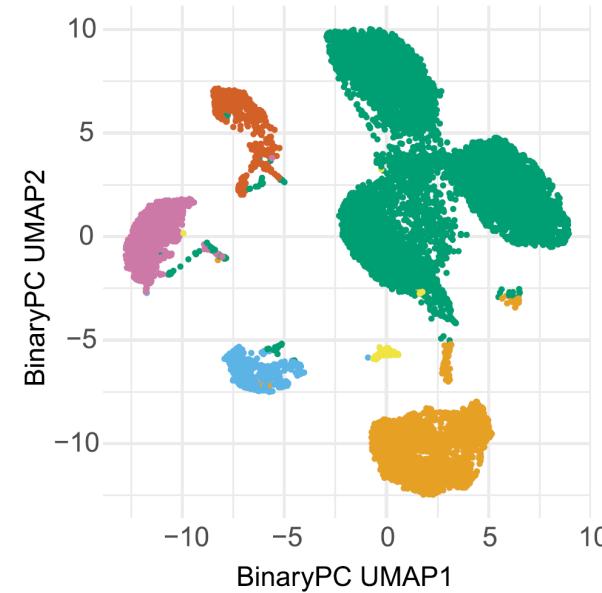
A



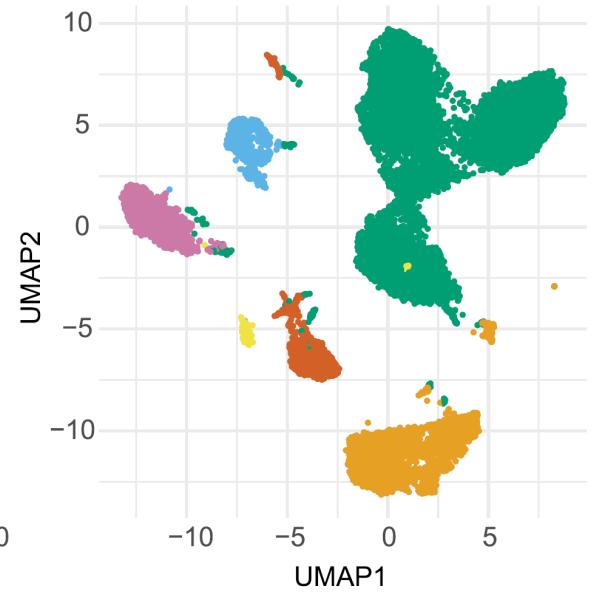
B



C



D



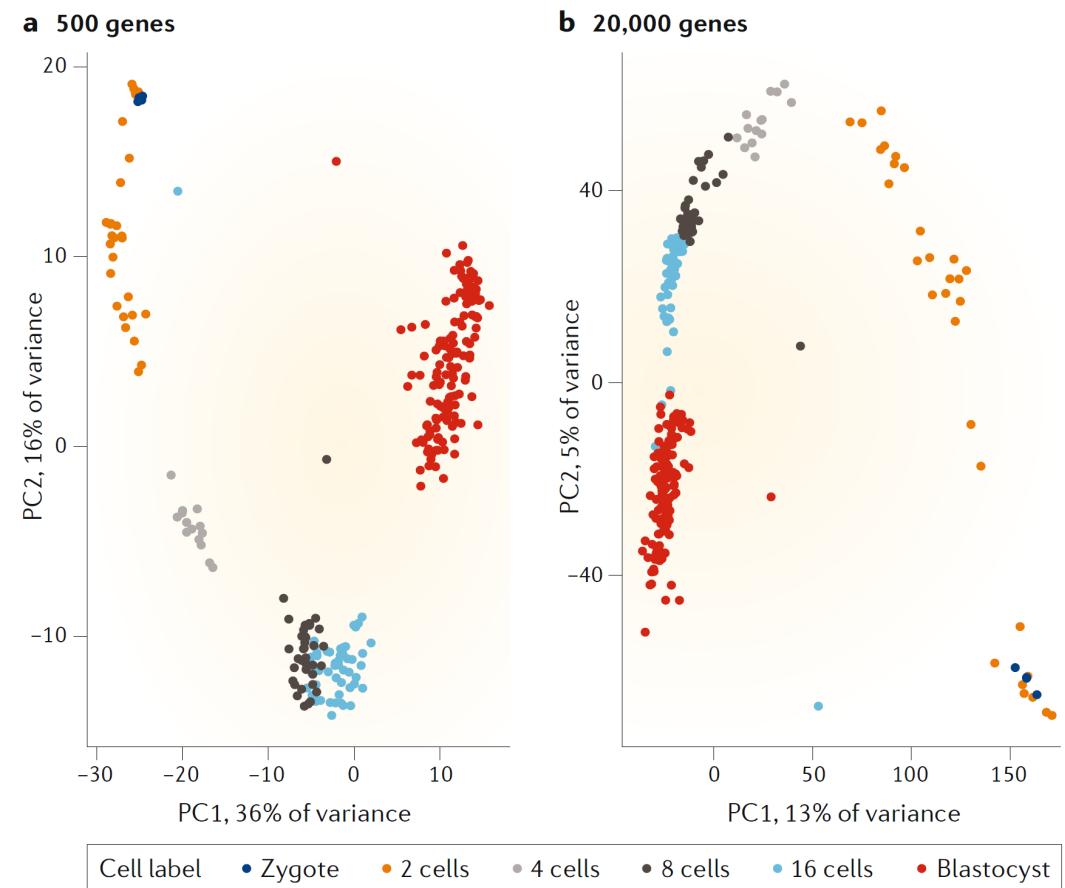
Binarized ~counts

1. Dimensionality reduction
2. Batch correction
3. Cell classification
4. Differential expression

Feature (Gene) selection

Feature selection

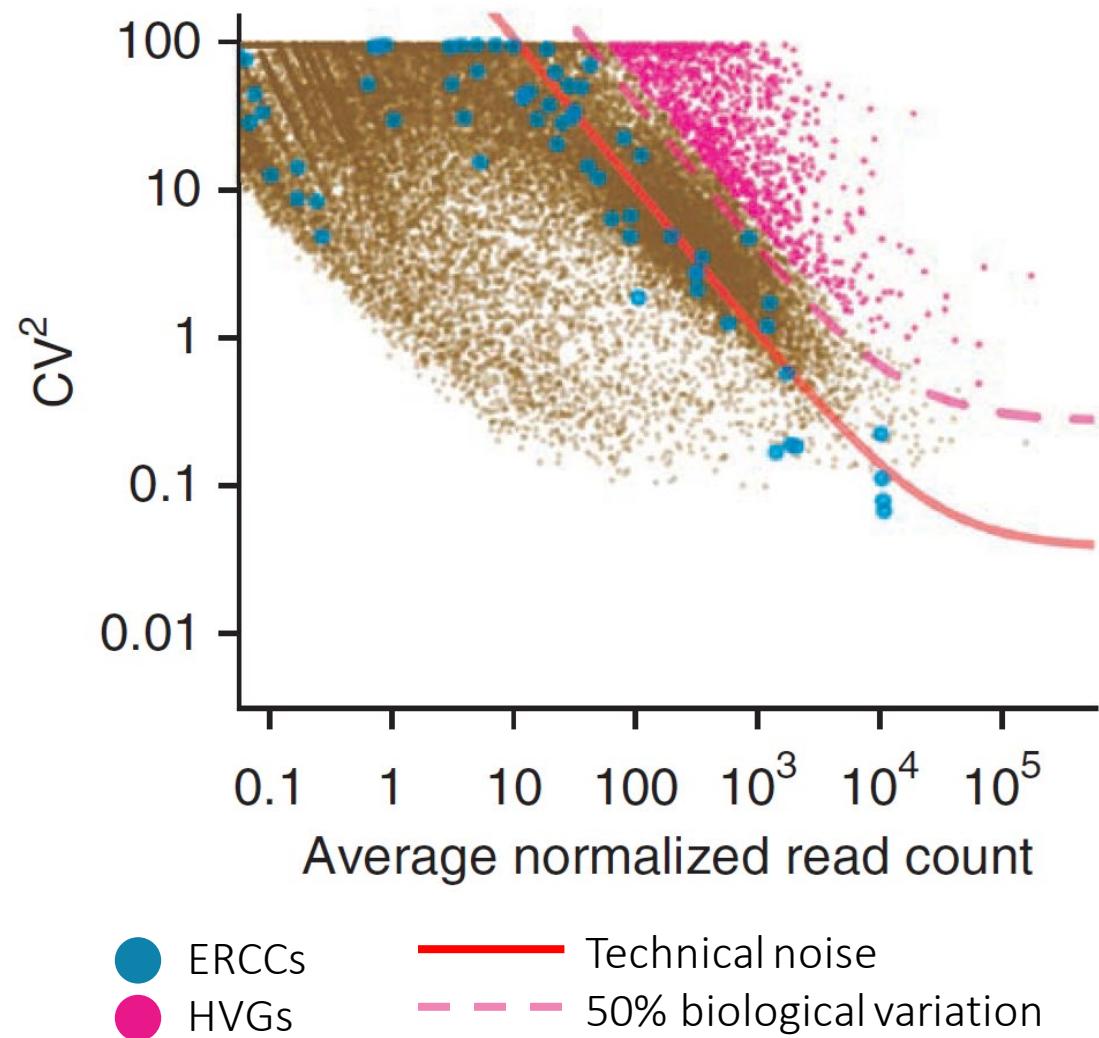
- Curse of dimensionality
 - More features (genes) -> noise dominates distances between samples (cells), effectively all cells get 'same' distance
- Remove genes which only exhibit technical noise
 - Increase the signal:noise ratio
 - Reduce the computational complexity



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

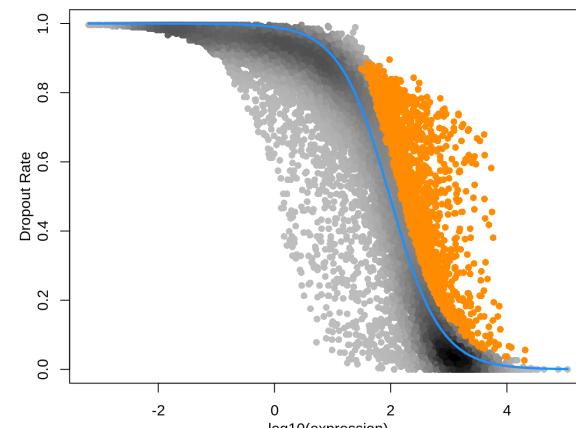
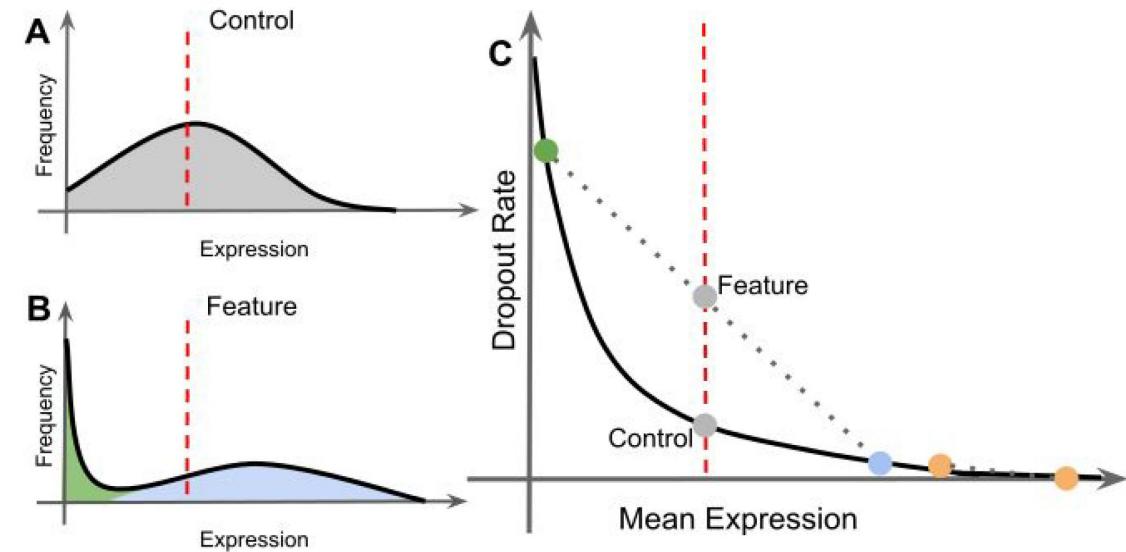
M3Drop: Dropout-based feature selection

- 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

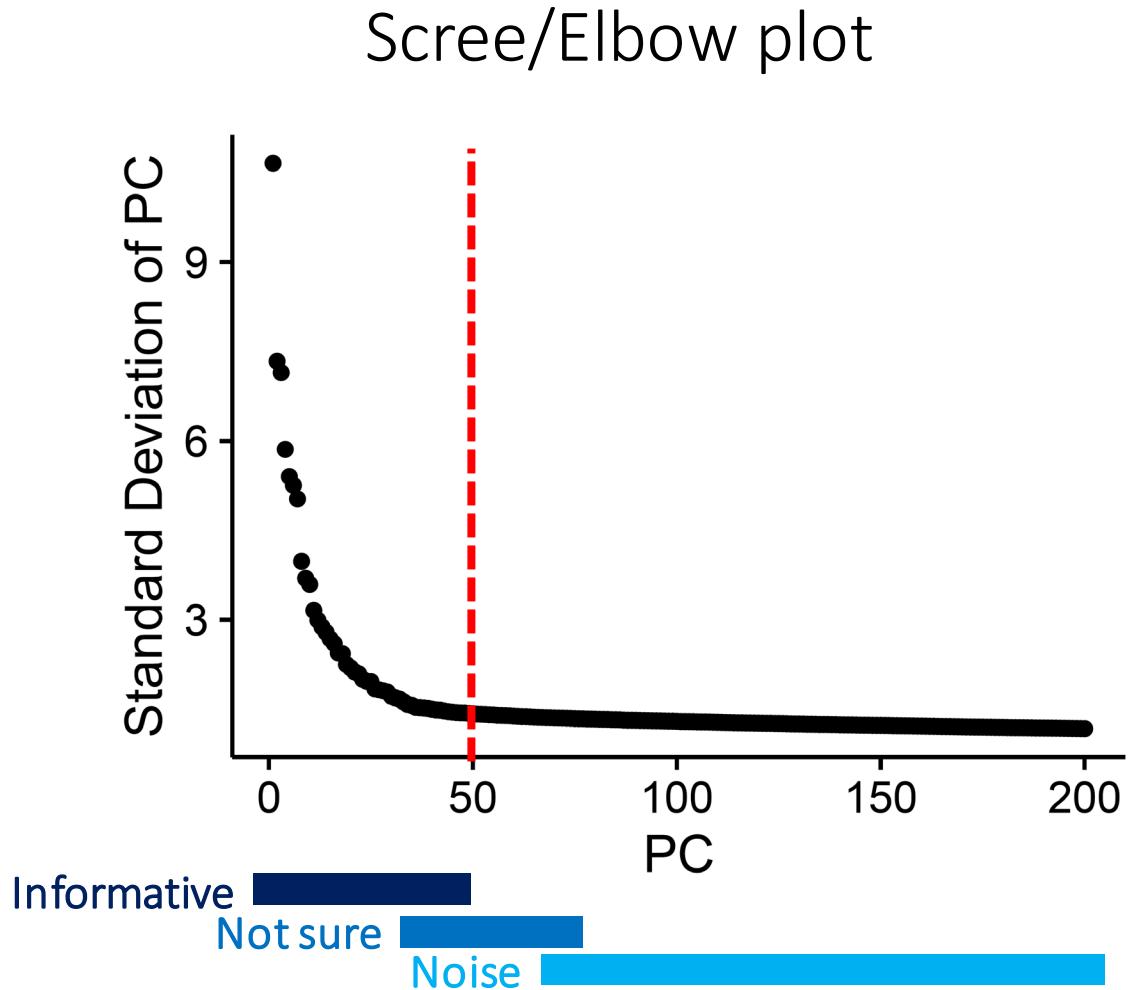
K_M : Michaelis-Menten constant



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

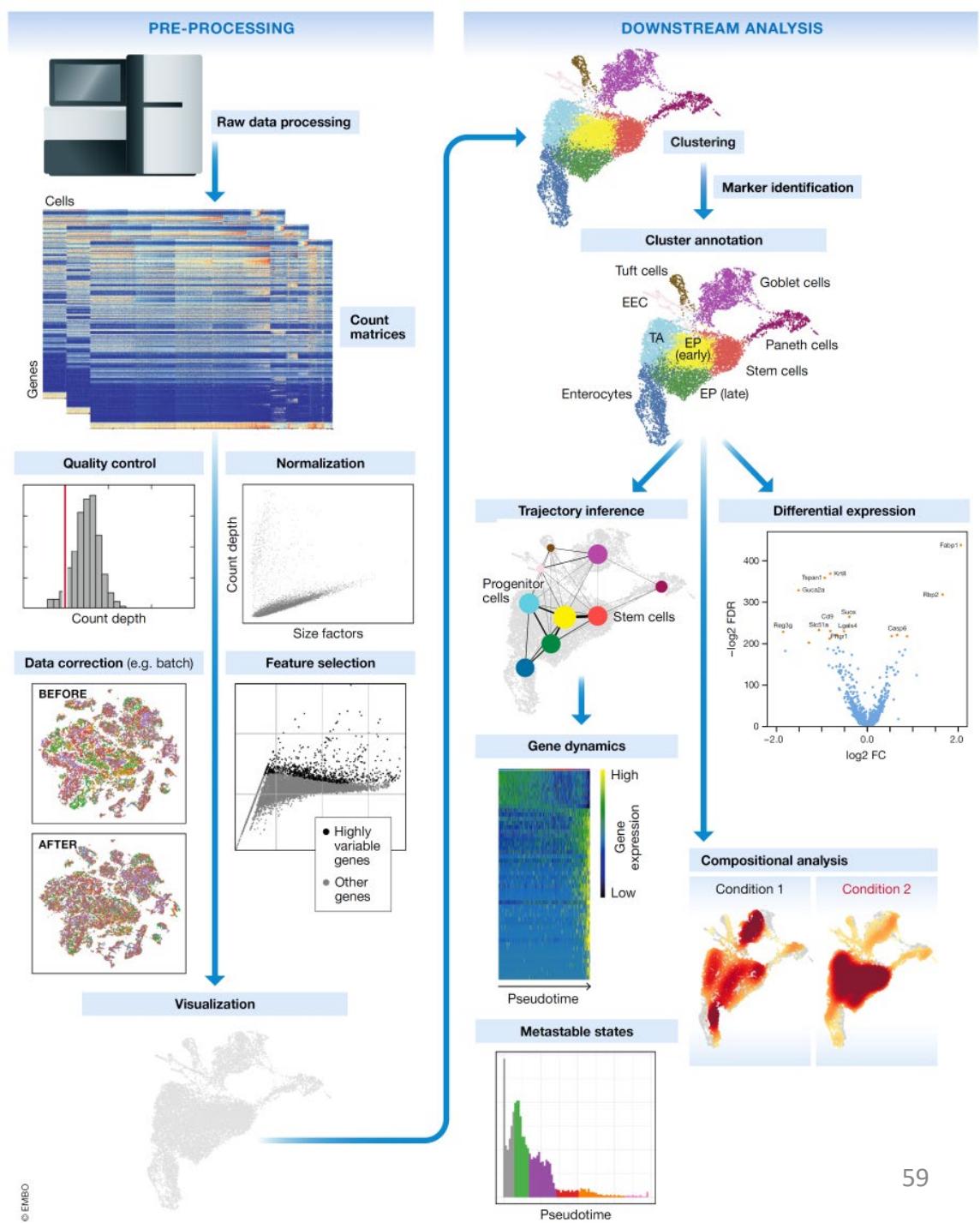


In short...

- If you have distinct celltypes – the clustering will be the same regardless of how you treat the data.
- **But**, for subclustering of similar celltypes normalization and removal of confounders may be crucial.

Summary

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



Useful Resources

- NBIS single-cell course:

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

- 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>

- GitHub: Awesome Single Cell

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

Posit Cloud available until 7 November

Thank You!

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



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