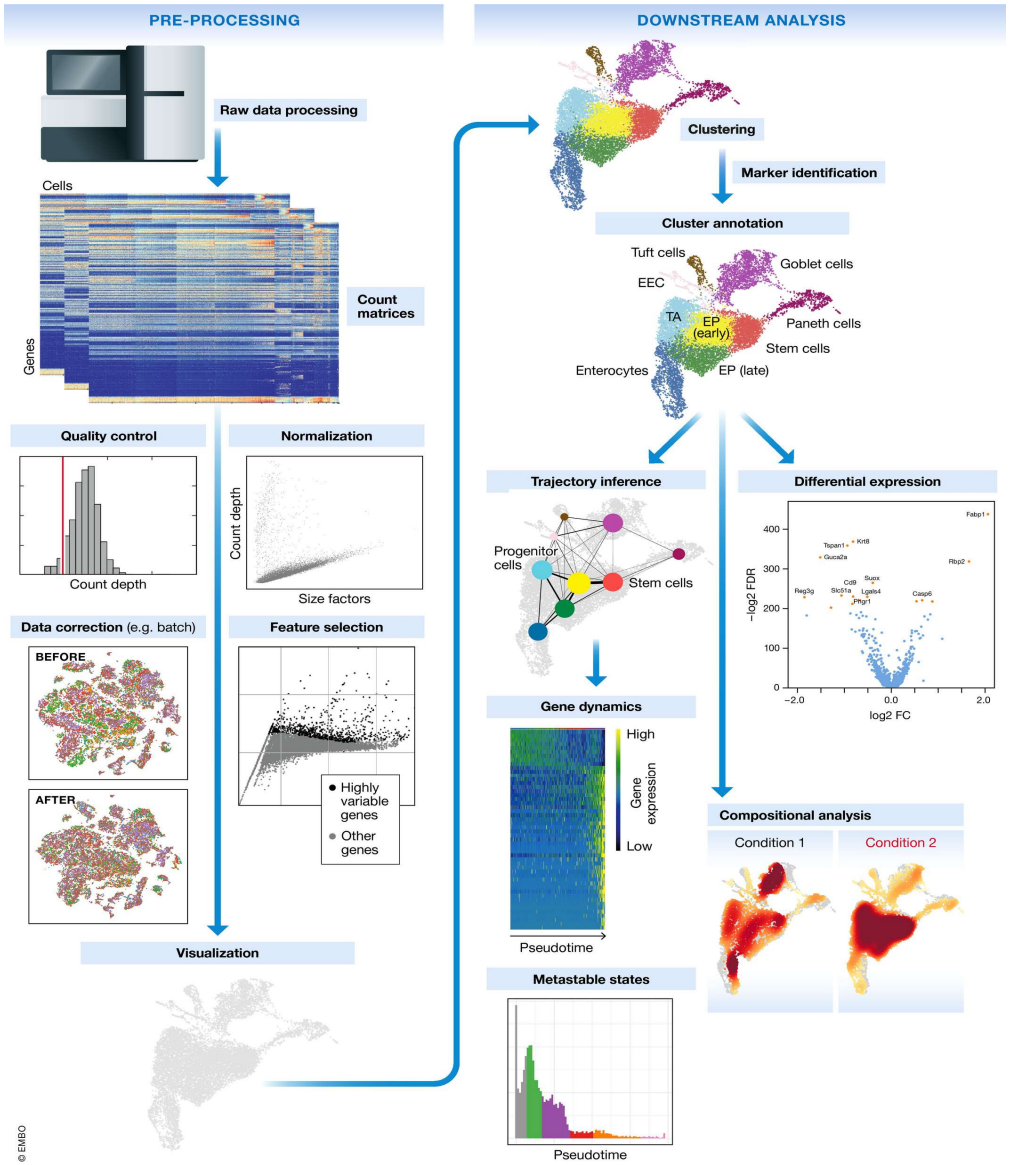
Quality Control, Normalization Experimental Design

Agnès Paquet

SincellTE 2024 - 10/21/2024

agnes.paquet@syneoshealth.com

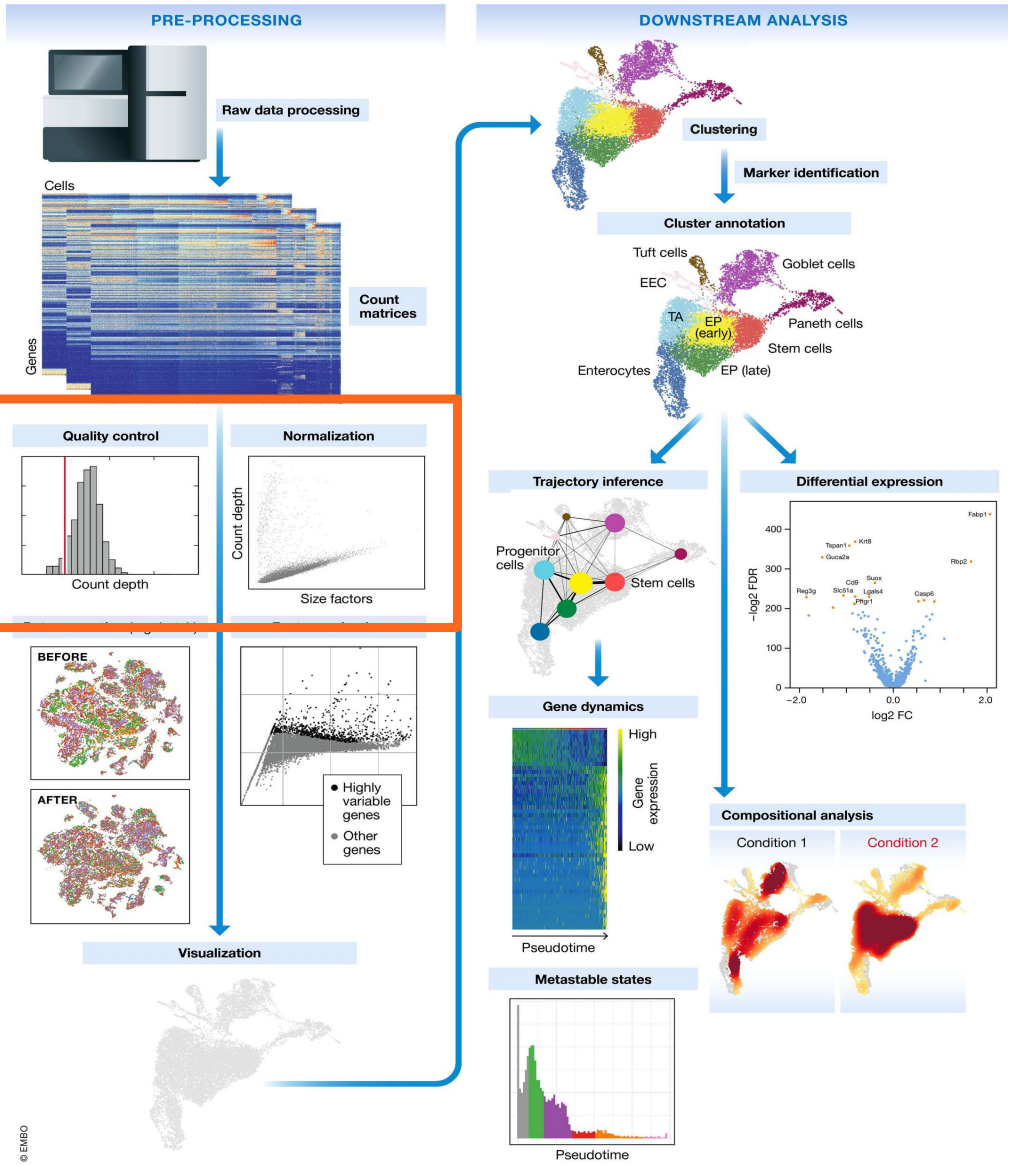
Main steps of single cell data processing



From Luecken and Theis, Mol Systems Biology 2019

Main steps of single cell data processing

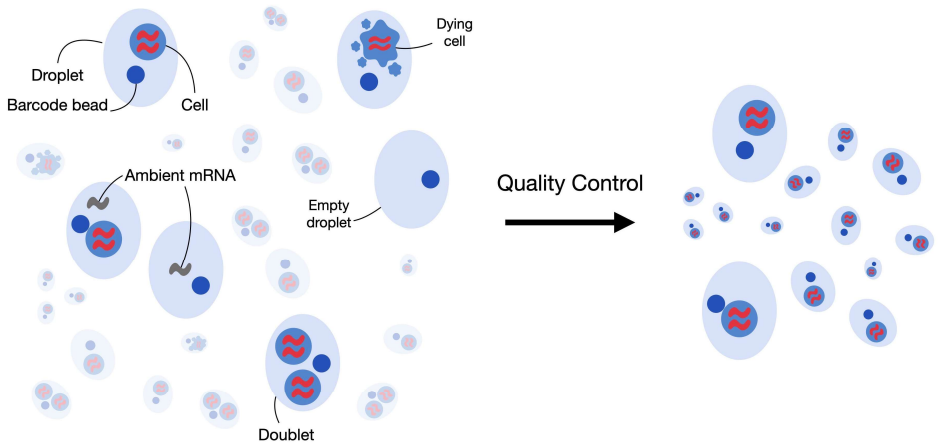
This course



From Luecken and Theis, Mol Systems Biology 2019

Aim of Quality Control

- We assumed that we have assigned 1 cell for each droplet (barcode)
- scRNAseq data quality can be impacted by technical and random noise



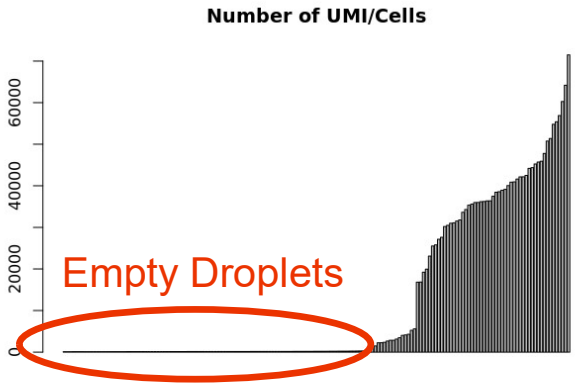
www.sc-best-practices.org

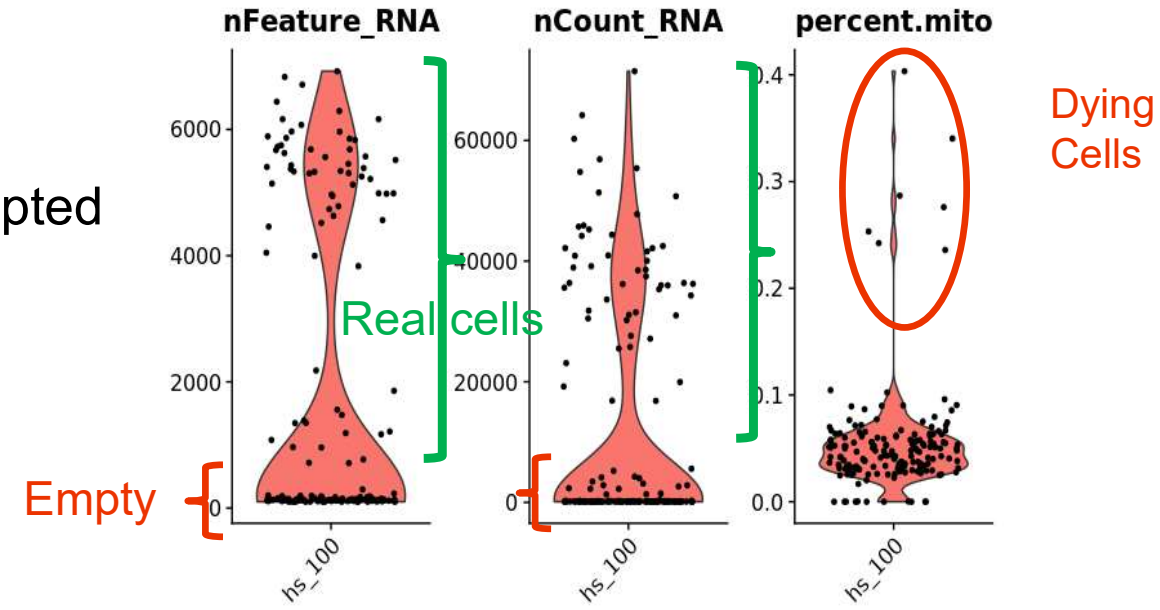
Preprocessing is required to eliminate low quality cells and clean up technical noise

1. Filter low quality cells (debris) and empty droplets
2. Remove Ambient RNA background
3. Detect and remove doublets

Detection of Poor Quality Cells

- Detection of empty droplets
  - Number of reads/UMIs per barcode
  - Number of genes detected per barcode
- Detection of dying cells
  - % of UMIs in mitochondrial genes
  - % of UMIs in Ribosomal genes
- Filtering thresholds should be adapted to your system
- Use graphs

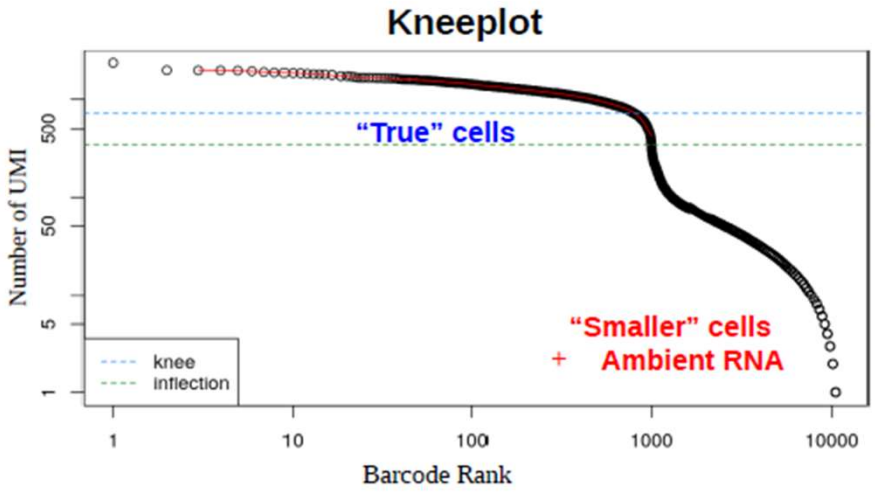




Filtering of Poor Quality Cells

- Cells with low RNA content may look like poor quality cells compared to other cells:
  - Small cells
  - Immune cells

Several rounds of analysis may be needed at this step to ensure that good quality data is not discarded

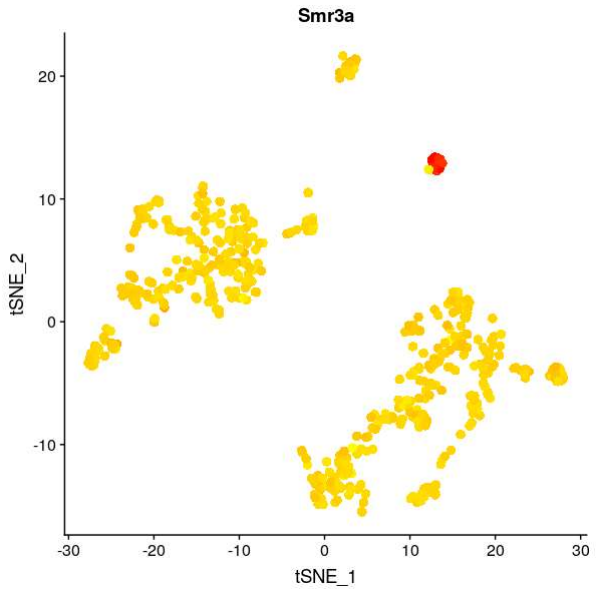


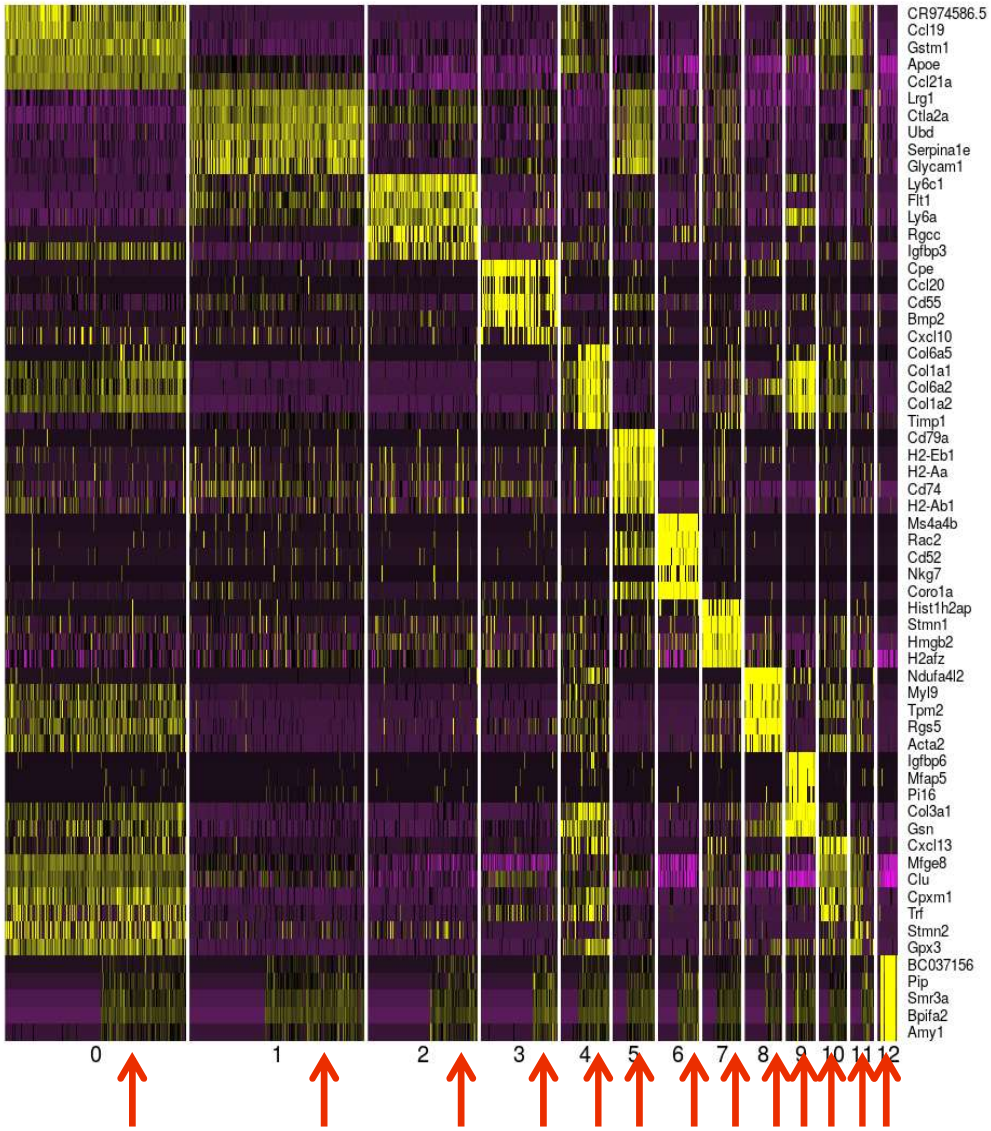
Adapted from Aglave, Montagne, Paquet

Ambient RNA correction

- RNA can leak from dying/dead cells
- Contamination of all droplets can occur

- Some tools can effectively remove this background noise

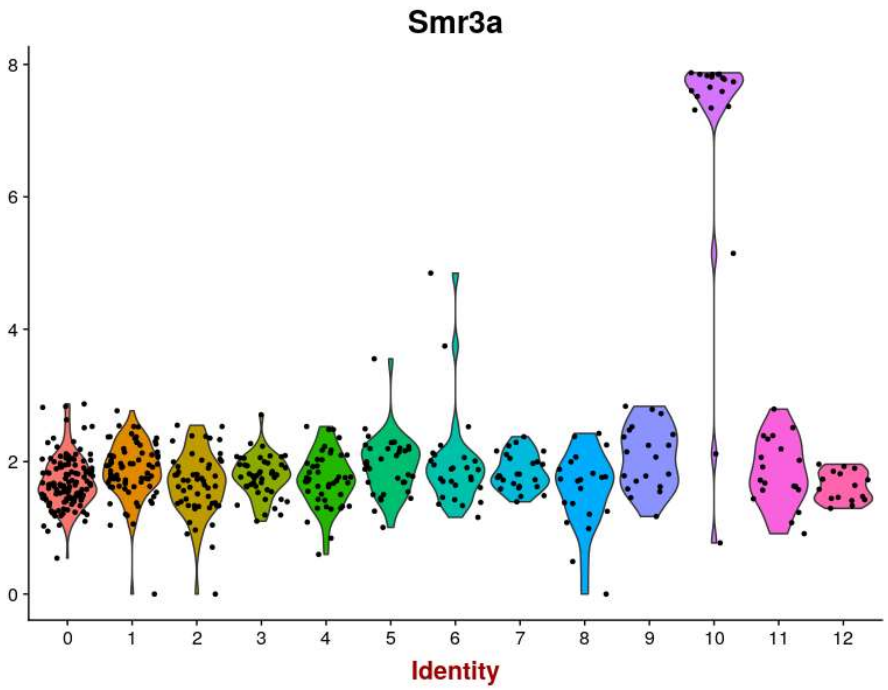




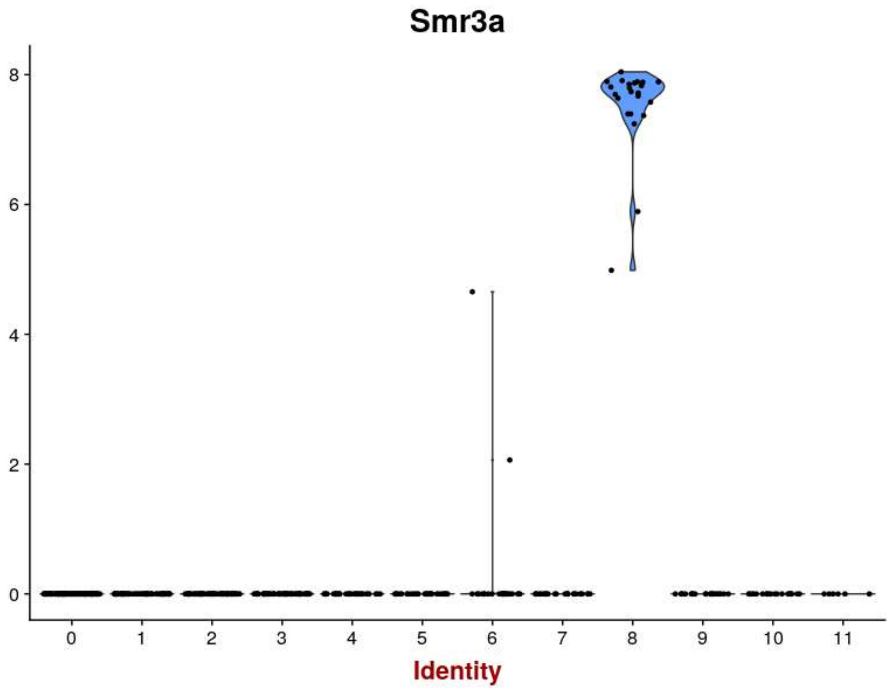
SoupX

Young MD, GigaScience 2020

BEFORE SoupX

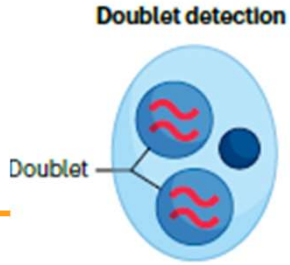


AFTER SoupX



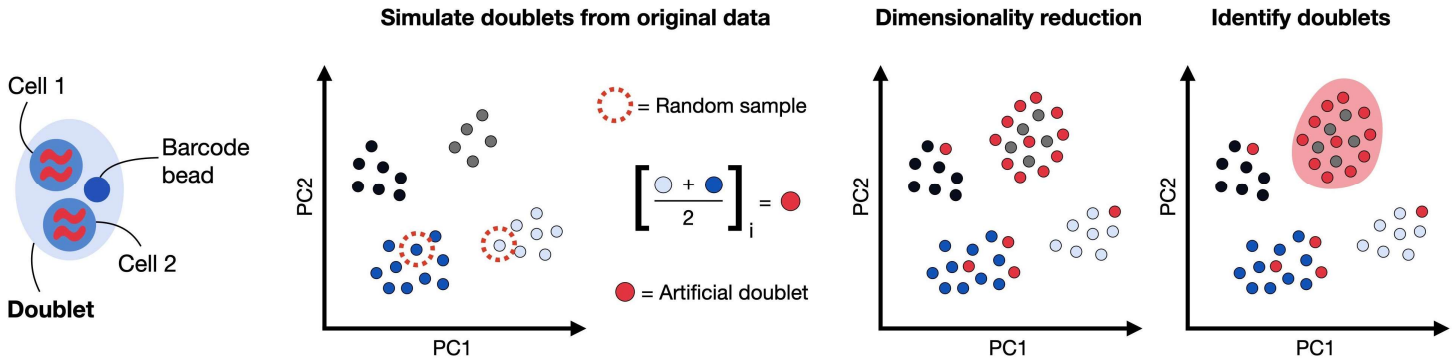
Warning: the software requires manual tuning.

Doublets/Multiplets



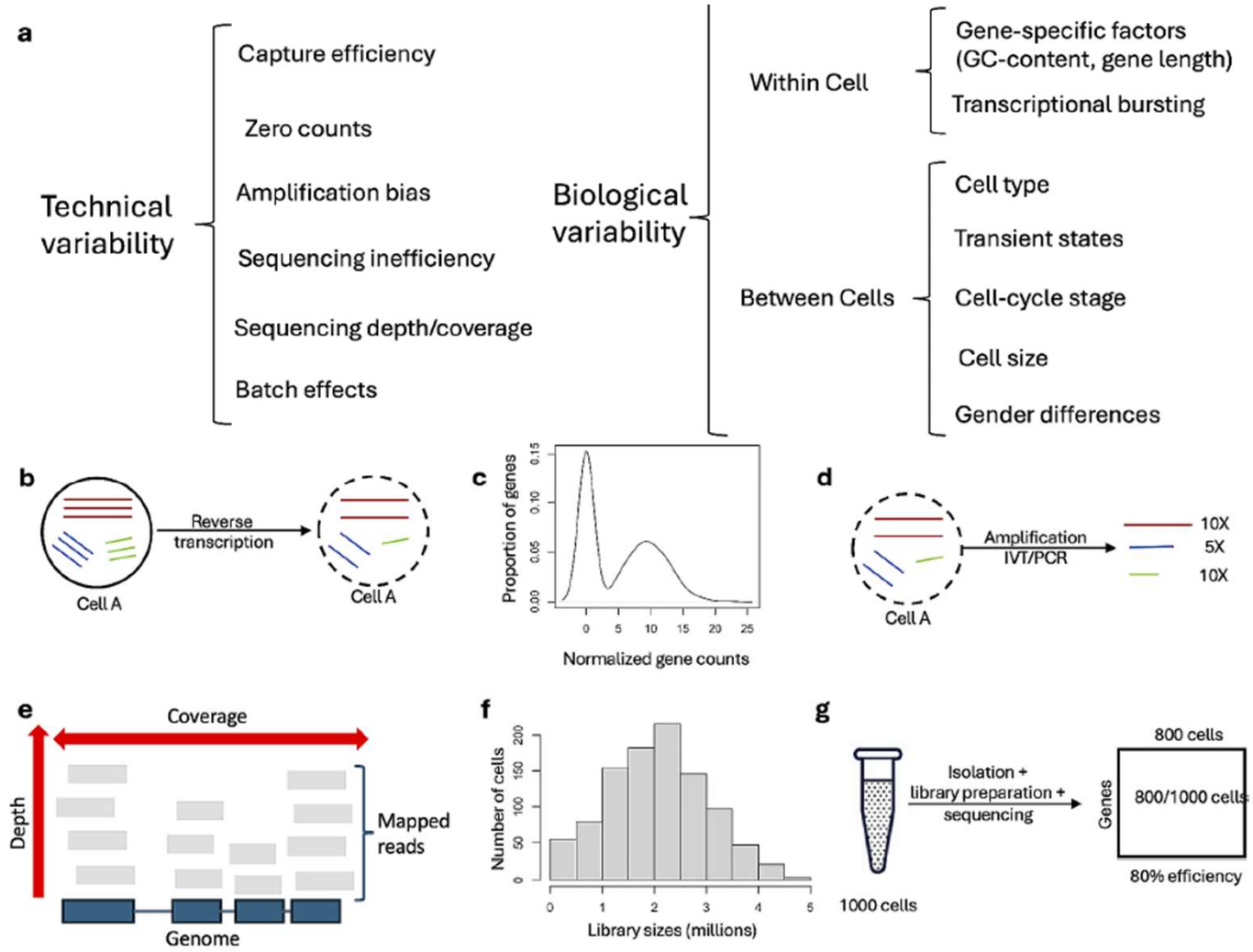
Doublets happen when two/more cells are encapsulated in the same droplet

- Number of doublets increases with cell loading density
- ***Homotypic*** : 2 or more of the same cell type
  - Harder to detect
  - Low capture efficiency -> doublet don't always have higher UMIs counts
  - Can be removed if coming from 2 individual samples (SNPs or multiplexed tags)
- ***Heterotypic*** : different cell types
  - Most problematic as they can look like an intermediate or transitioning cell type
- Several tools exist to identify doublets: scDblFinder, scds etc…



Sources of Variation

- scRNAseq show strong variability between cells and between genes.



Cuevas-Diaz Duran, 2024

scRNA-seq: 3 levels of normalization

- Normalization = Process of identifying and removing systematic variation not due to real differences between RNA treatments
  - i.e. differential gene expression.
- Goal: make gene counts comparable within and between cells.

- Gene-specific effects
  - Within cell: GC content, gene length
  - For full-length RNAseq protocols

- Cell specific effects
  - Sequencing depth
  - Aim: make count distributions comparable

C

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

Vallejos CA, 2017

- Sample/Technology-specific effects -> Data Integration
  - Batch effects (BAD)
  - Between samples variability (GOOD)

Bulk RNAseq normalization

- RPKM/FPKM/TPM/CPM (Reads/Fragments per kilobase of transcript per million reads of library)
  - Normalize for sequencing depth and transcript length at the same time

  -> ok if you have full length data

- Global scaling
  - Eg. Upper Quartile
  - If we have too many zeros, the Size Factor will be off

- Size factors calculation
  - Estimation of library sampling depth
  - DESeq2, edgeR TMM
  - Suppose that **50%** of genes are **not DE**
  - If we have too many zeros, the SF will be off

- These methods don't work well for single-cell data
  - TPM/CPM can be bias by a small number of genes carrying most of the signal
  - Quantile based methods are limited: large number of zeros -> scale factor = 0

scRNA-seq: 3 levels of normalization

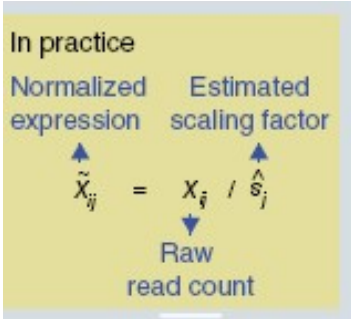
- Gene-specific effects
  - within cell: GC content, gene length
  - *Not really accounted for in droplet assays*

- Cell specific effects
  - Aim: make count distribution comparable
    1. Global scaling
    2. Variance stabilization methods
    3. Others

- Sample/Technology-specific effects -> Data Integration
  - Batch effects (BAD)
  - Between samples variability (GOOD)

Global Scaling

- Hypotheses:
  - Cell populations are homogenous
  - The RNA content is similar in all cells
  - Same scaling factor for all genes

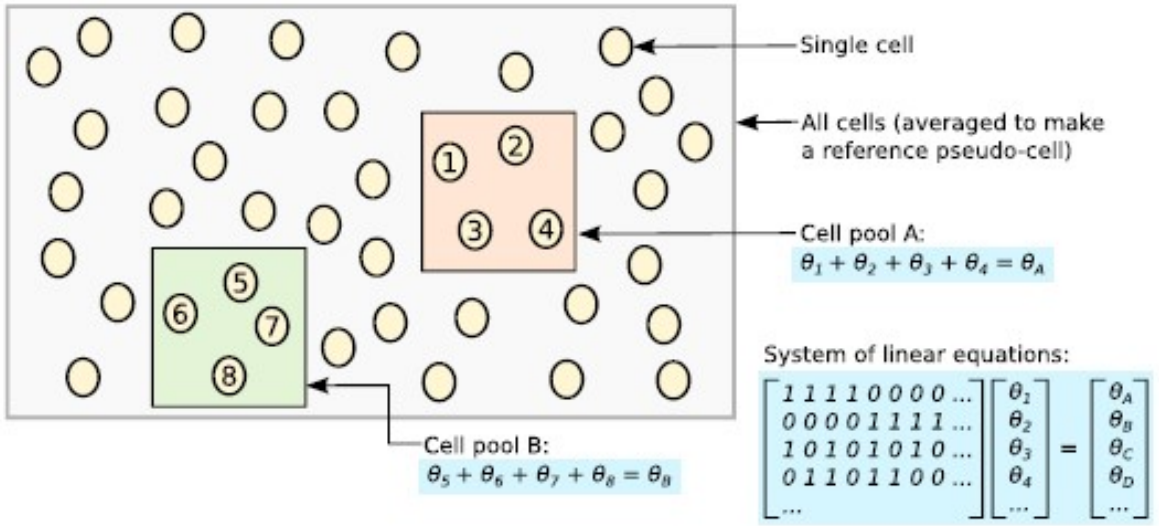


- Choice of the scaling factor
  - Median UMI counts
  - 10,000 default in Seurat / Cell Ranger

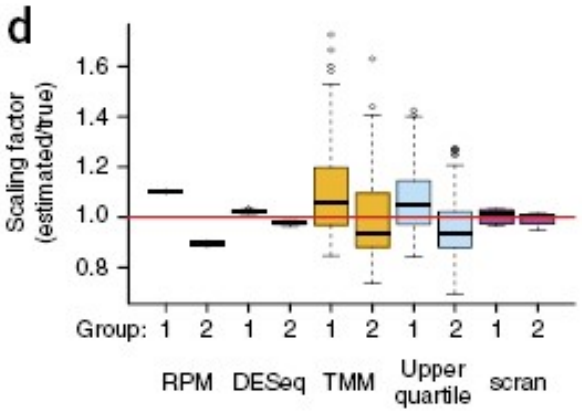
- In practice
  - Hypotheses are not always verified, but lots of people use this method anyway

Estimation of size factors using deconvolution

- Alternative method to compute the size factors
- Pool cells to reduce the number of zeros
- Estimate the size factors for the pool
- Repeat many time and use deconvolution to estimate each cell size factor
- Implemented in **scran** packages



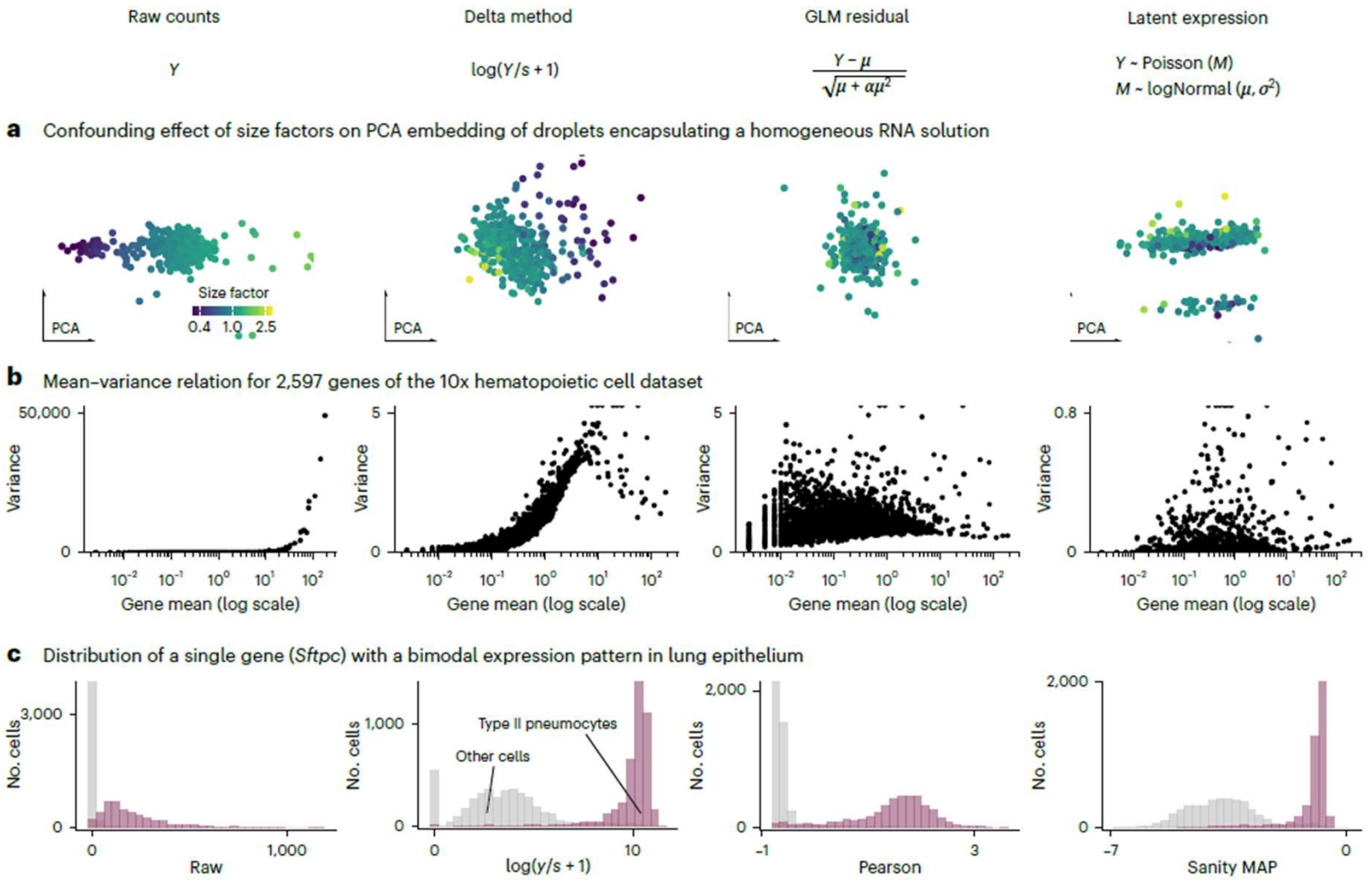
Lun, 2016



Vallejos C, 2017

Variance Stabilization

- Aim: Correct for the strong mean-variance relationship



Ahlmann-Eltze, 2023

Other methods are available…

- Normalization included in the statistical model
  - SCDE, Monocle, MAST,…
- Normalization based on spike-ins or invariant genes
  - BASICs, scNorm
- Fancy modeling
  - Modeling of single cell count data using Neg Binomial
  - ZINB-Wave, single-cell variational inference (scVI) etc
- Normalization for other biological factors
  - Known or unknown variation: Cell cycle, % mitochondrial genes…
  - Regression methods provided to account for know factors (E.g. Seurat)
  - Latent variable models to estimate and remove unknown bias (scLVM)

scRNA-seq: 3 levels of normalization

- Gene-specific effects
  - within cell: GC content, gene length

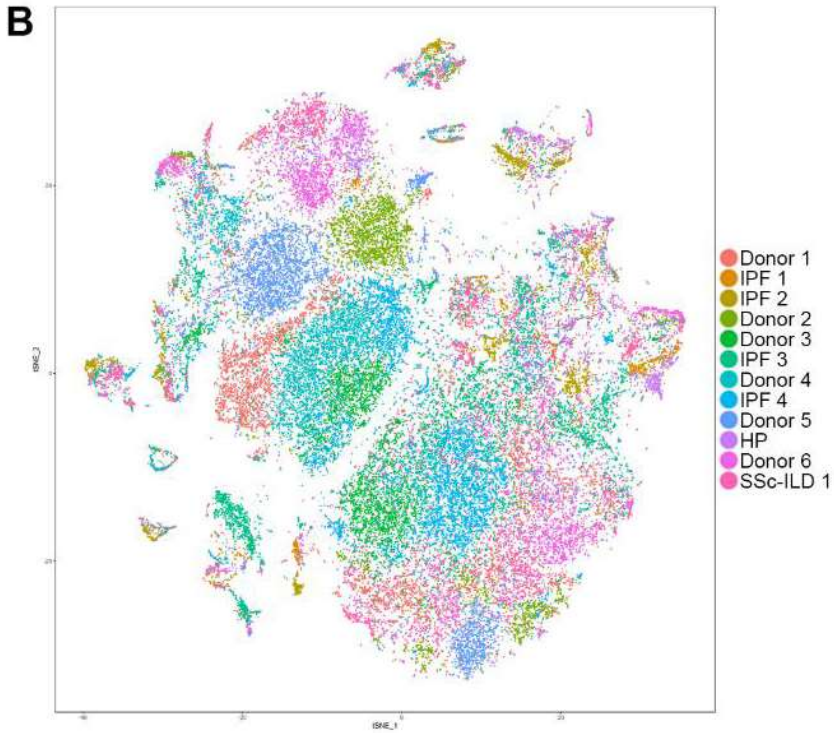
- Cell specific effects
  - Aim: make count distribution comparable
    1. Global scaling
    2. scRNA-seq specific method from scater/scran package
    3. Others

- Sample/Technology-specific effects -> Data Integration
  - Batch effects (BAD)
  - Between samples variability (GOOD)

Why do we need data integration methods?

- In practice: single cell techniques are biased
  - Variations between samples can be huge
    - donor effect +/- sampling effect
  - Samples may be processed using different technologies

- Combining datasets and applying cell-level normalization might not be enough to remove this bias



Misharin, BiorXiv 2018

Data integration

For differential analysis:

-> Choose a framework where you can add a batch term in your statistical model (e.g.: MAST, DESEq2, limma,…)

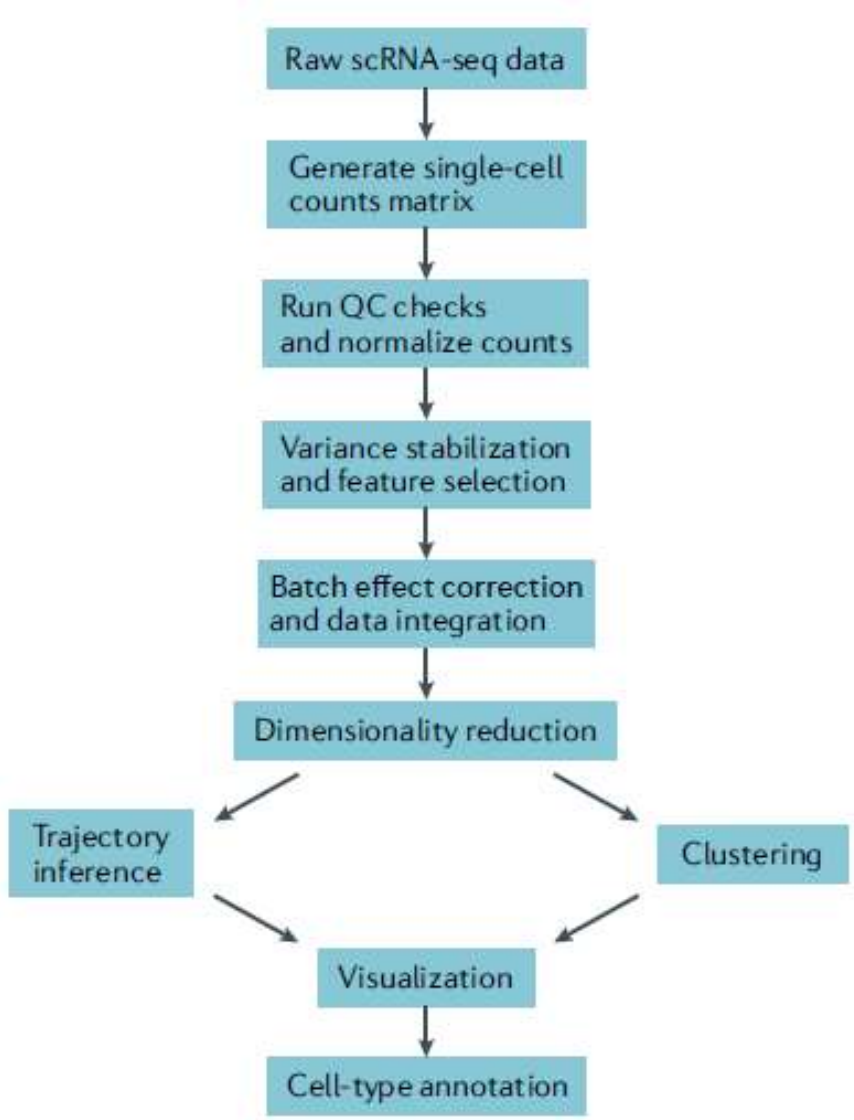
For other analyses:

- We need a method that will “merge” our datasets and remove the unwanted variation
- Non-linear transformation of cells in different proportions
- Aligns datasets from different technologies and species

Conclusion

- Parameters choice will affect the results
  - Number of features selected
  - Number of PCs used in downstream analysis
  - Clustering resolution, etc…
- Analysis will have to be repeated many times

- Normalization method should be selected based on the question of interest
- Variance Stabilization
  - Pearson Residual is best for cell type identification
  - Shifted-Log performs well for everything



Wu, Nat RevNeph 2020

References

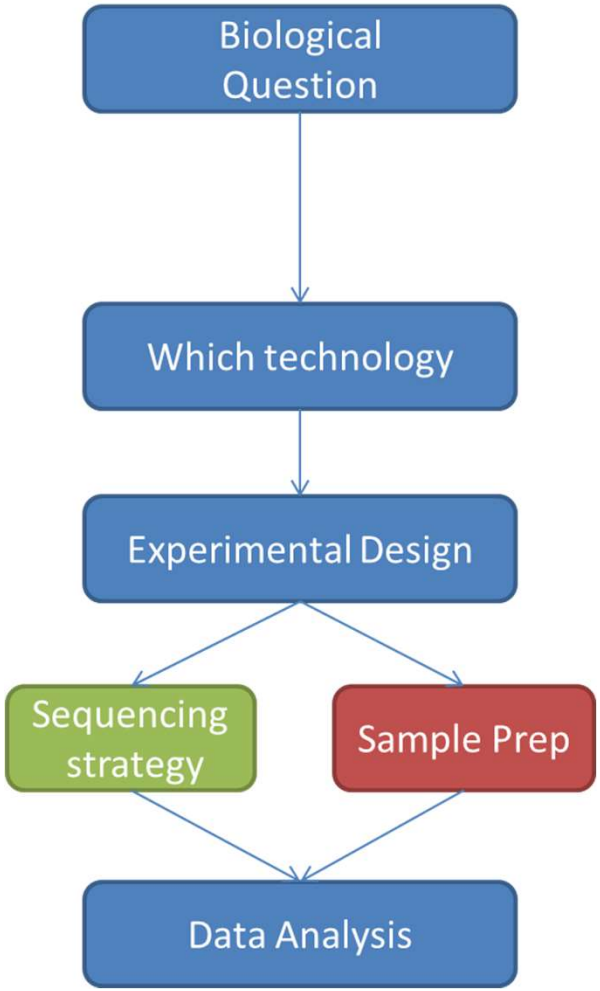
- Raquel Cuevas-Diaz Duran et al., Data normalization for addressing the challenges in the analysis of single-cell transcriptomic datasets, BMC Genomics 2024
- Heumos et al., Best practices for single-cell analysis across modalities, Nature Reviews Genetics 2023
- Ahlmann-Eltze C and Huber W, Comparison of transformations for single-cell RNA-seq data, Nature Method 2023
- https://www.sc-best-practices.org/
- Advanced Single-Cell Analysis with Bioconductor
- Germain et al, Doublet identification in single-cell sequencing data using scDblFinder, F1000Research 2022
- Dal Molin A, Di Camillo, How to design a single-cell RNA-sequencing experiment: pitfalls, challenges and perspectives, Briefings in Bioinformatics 2019
- Yan Wu and Kun Zhang, Tools for the analysis of high- dimensional single- cell RNA sequencing data, Nat Rev Nephrology
- You Yue et al, Benchmarking UMI-based single-cell RNA-seq preprocessing workflows, Genome Biol 2021
- Vallejos CA, Normalizing single-cell RNAsequencing data: challenges and opportunities, Nat Method 2017
- **Scater**: Lun A, Pooling across cells to normalize single-cell RNA sequencing data with many zero counts, Genome Biology 2016
- **Seurat:** Butler et al. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nature Biotechnology (2018).

  https://satijalab.org/

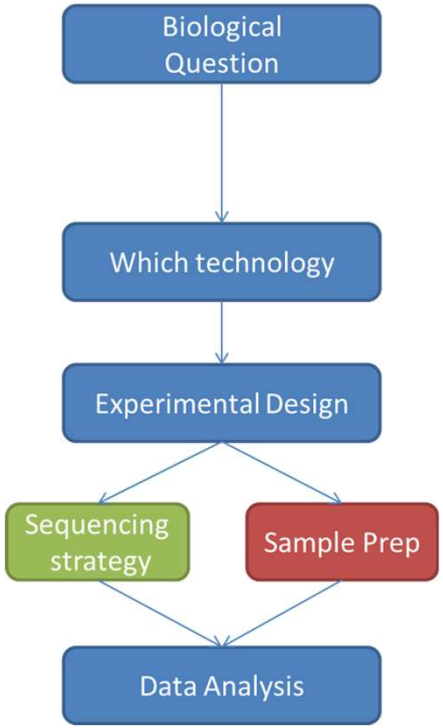
Experimental Design

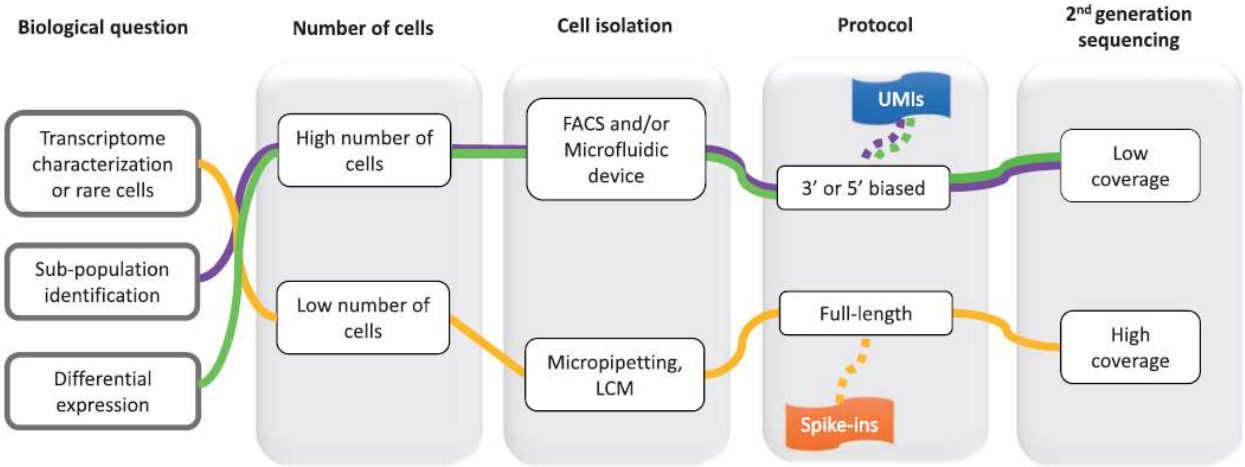
scRNAseq workflow: Bioinformatics Point of View

- What is the question ?
- What technique should we use to generate the data ?
  - Plate based / droplets
  - Full length / 3' counting with UMI
  - ➢ UNDERSTAND THE BIAS
- Experimental design
  - Sequencing strategy
    - Number of cells / number of reads
    - Spike-ins (not available for droplets)
  - Samples: Practical considerations
    - Types /number of samples
    - Cell preparation
    - Budget



Experimental design: technical considerations





Dal Molin, 2019

- Large panel of cell isolation technologies and RNAseq protocols available
- Understand protocol bias to help your collaborator select the appropriate method
- Samples: practical considerations
  - What are the major sources of variability?
  - Types / number of samples -> Biological Replicates
  - Cell preparation -> Be careful of *confounding*
  - Budget

Experimental Design

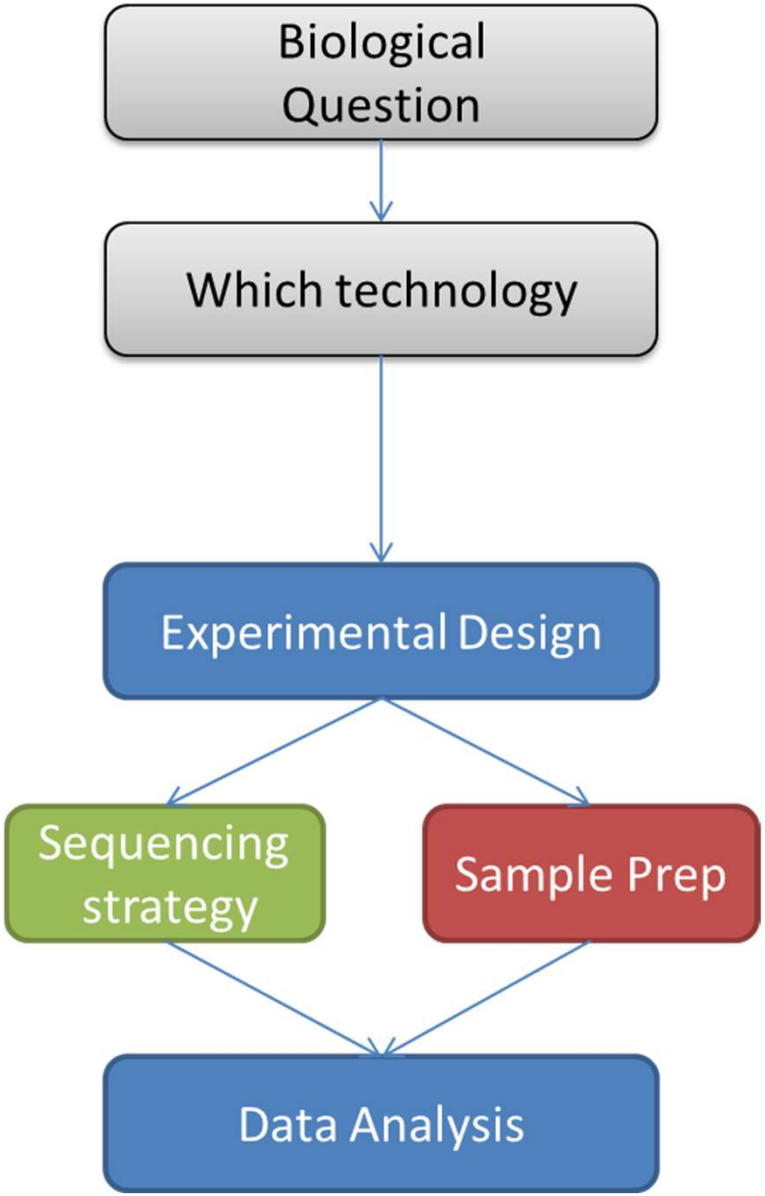
- We have a question
- We have selected a protocol

- How many samples?

- How many cells?

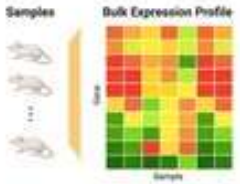
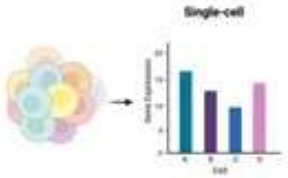
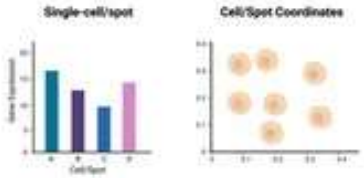
- How many reads/cell?

- How do we combine all this to minimize batch effect?



How Many Samples?

Hyeongseon Jeon, 2023

	Bulk RNA-Seq	Single-cell RNA-seq	High-throughput Spatial Transcriptomics
Level			
Data Structure	Subject x Gene Expression Count Data	Cell x Gene Expression Count Data	Cell/Spot x Gene Expression Count Data Cell/Spot 2-dimensional Coordinates
Detection Target	Differentially Expressed Genes	Differentially Expressed Genes Cell Sub-populations	Spatially Variable Genes Tissue Architecture Cell-Cell Communication

- Bulk RNAseq: each library corresponds to a biological sample
  - Biological Replicates
  - Technical replicates not recommended

- Single Cell RNAseq: 1 sample/batch = Many Cells (libraries)
  - Each cell comes from the same biological sample
  - Cells are not true replicates: there is a correlation between cells from the same sample
  - Biological replicates are needed for robustness

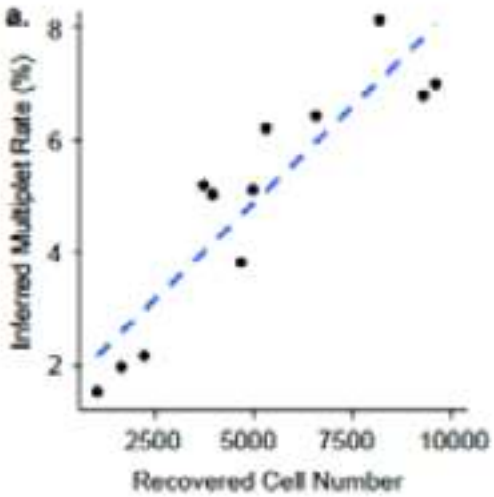
Estimating the required number of cells / sequencing depth

- Number of cells required
  - Do we have a lot of cells to begin with?
  - Are we looking for rare cells (probability estimation)?
- WARNING: doublet rate increases with higher cell numbers in droplet assays.

- Sequencing depth
  - What are the limits of my sequencer? (Novaseq or NextSeq)
  - Minimal number of reads for droplets: 50,000 reads/cells
  - Do the cells have lots of RNA ?
  - *Think about sequencing saturation*
  - *Think about dropouts generation*

Several tools are available for power calculation

Zheng 2017



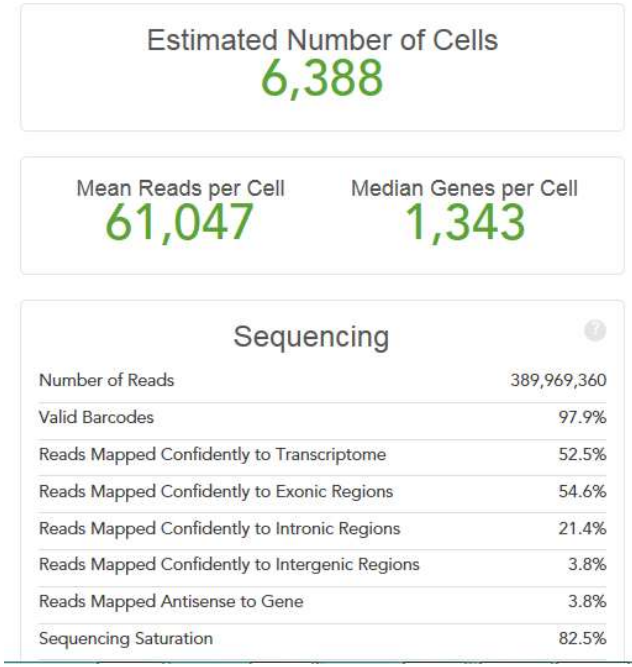
Example 1: PBMC
small cells, some don't have a lot of RNA

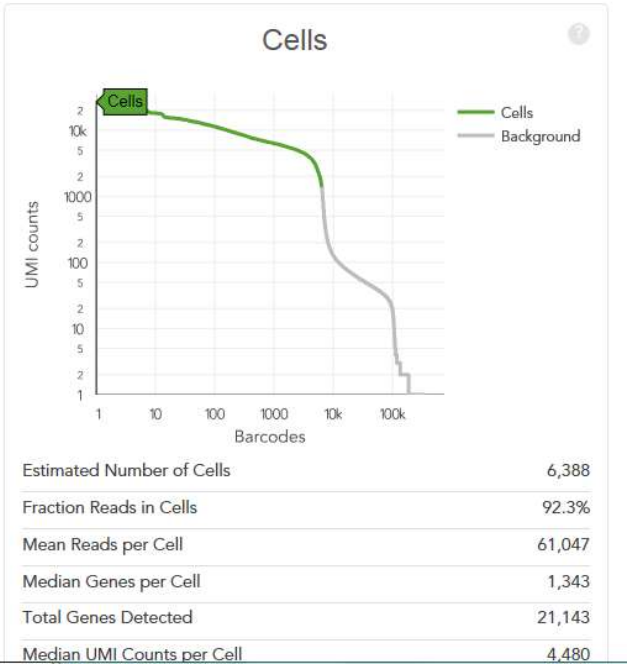
Target: 5,000 cells

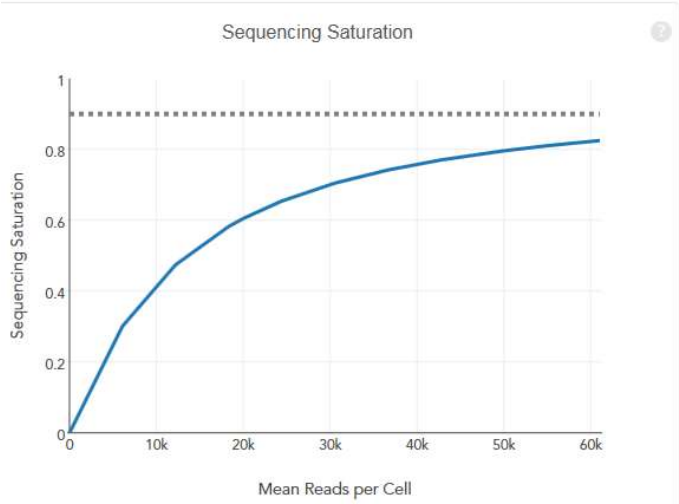
1 sample

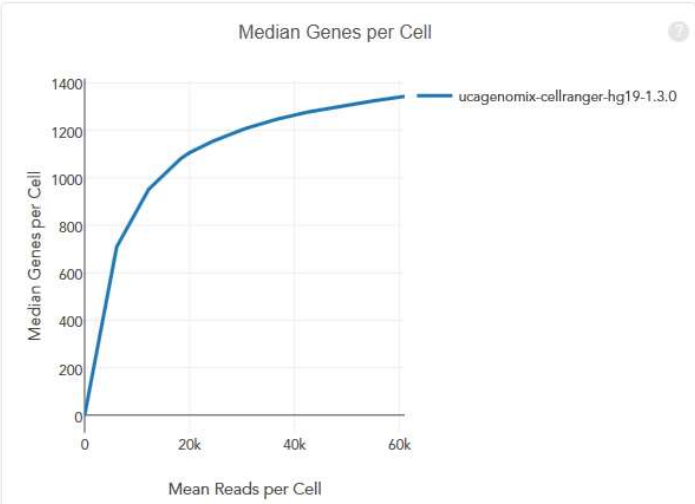
NextSeq High 75

(~400millions reads / run)









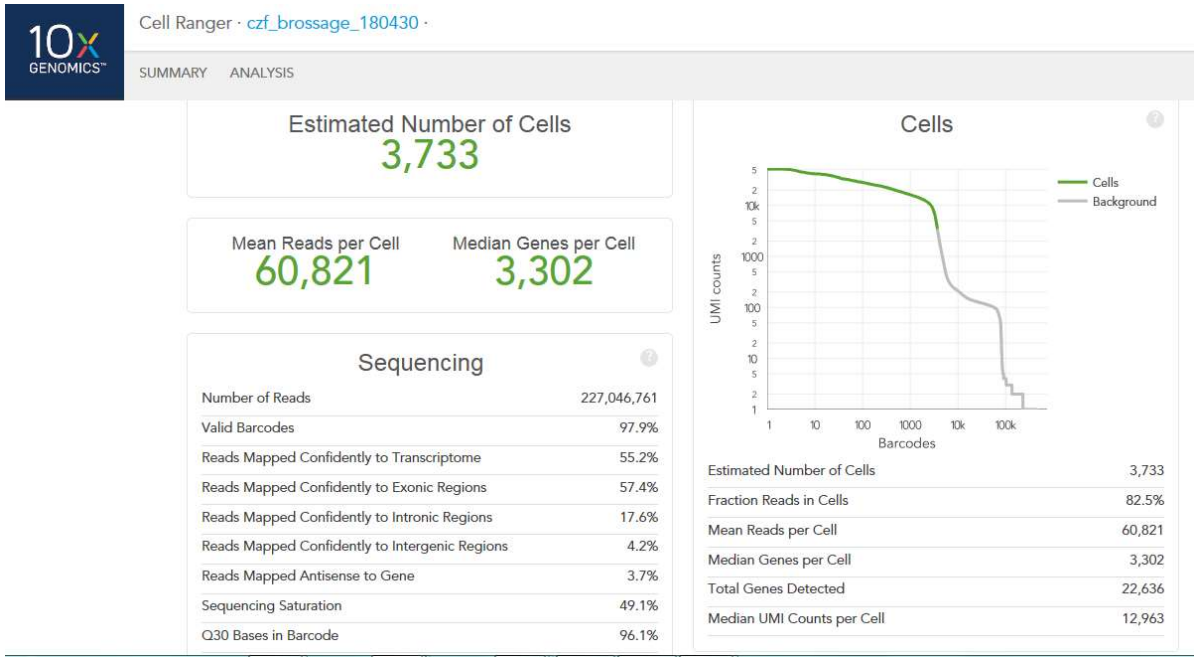
Example 2: Nasal epithelium brushing cells with lots of RNA

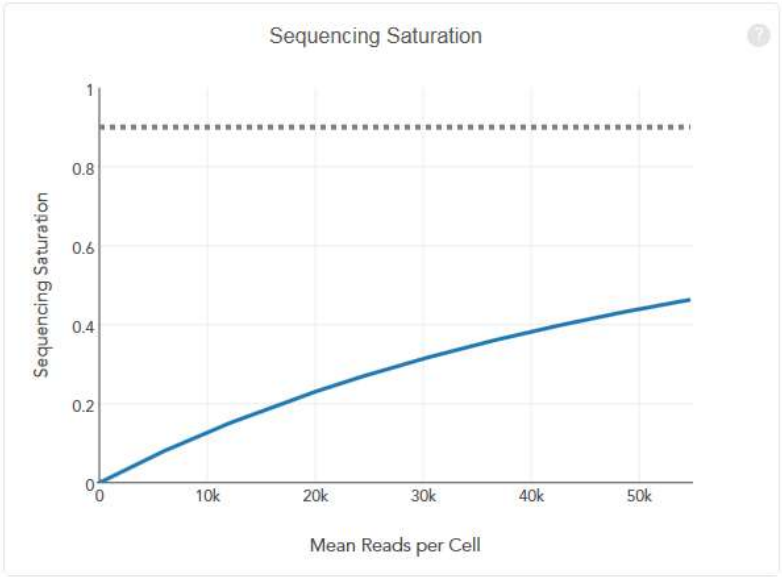
Target: 5,000 cells

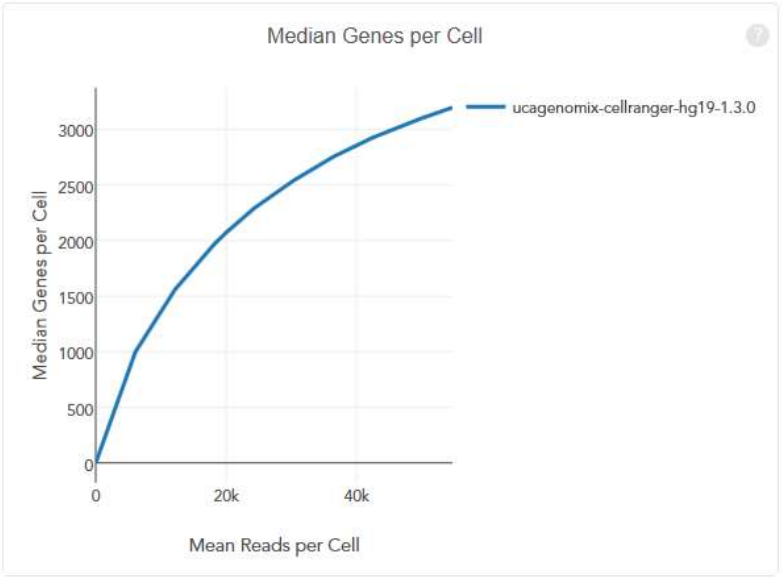
2 samples,

NextSeq High 75

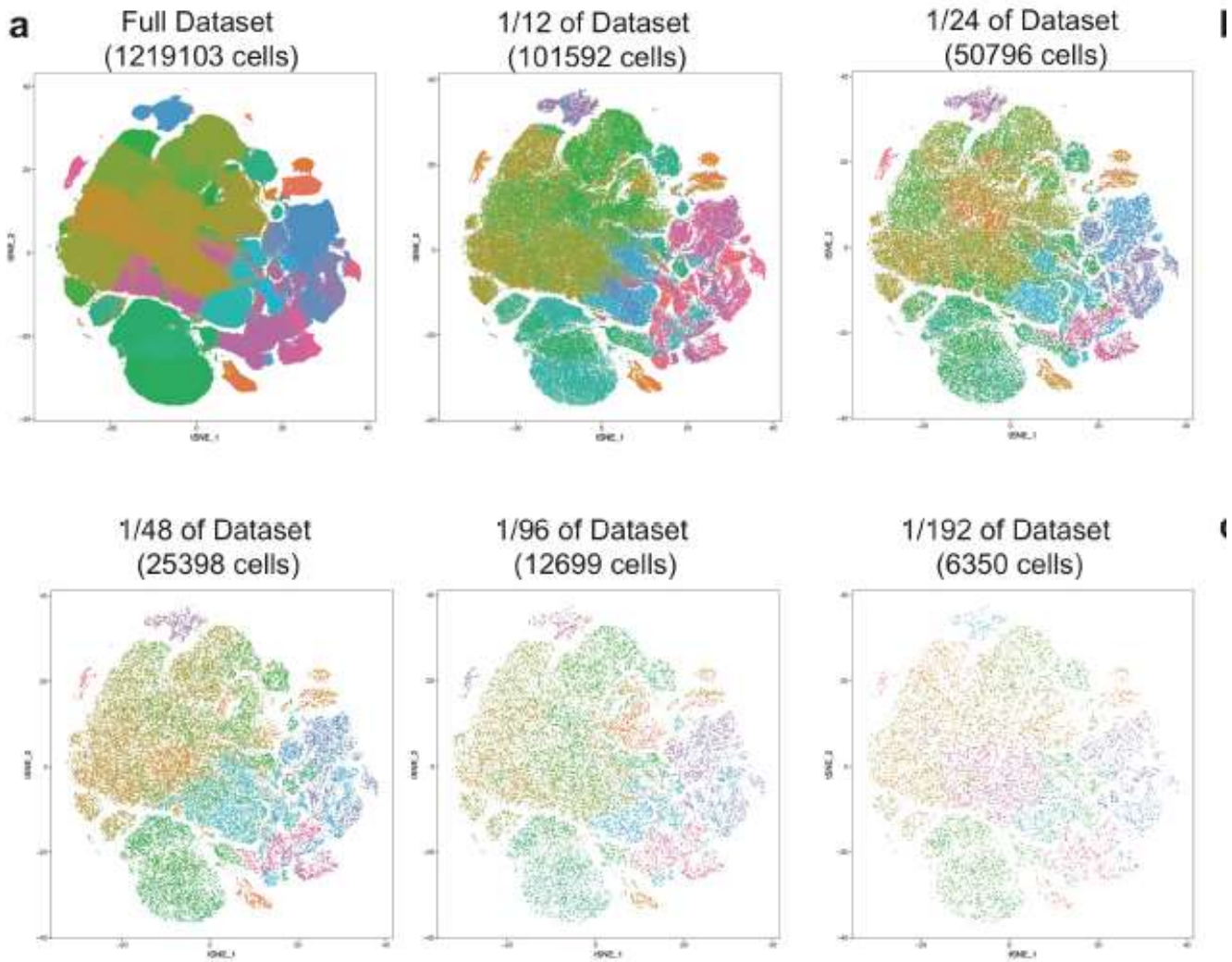
~400millions reads / run







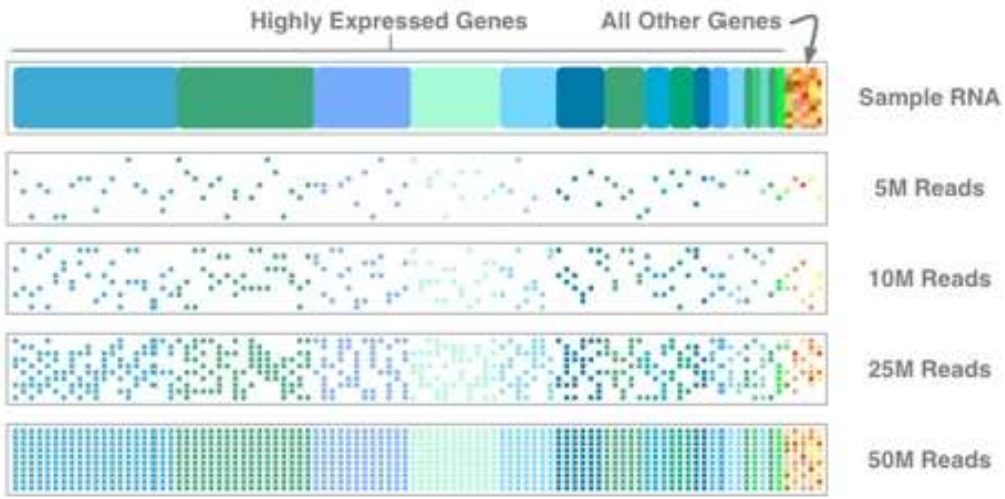
Number of cells: example of the 1.3millions cells dataset



Bhaduri A, BiorXiv 2017

Technical design: summary

- Discuss about sequencing depth with the biologist
- If the sequencing is too shallow, the statistical analysis may not be robust
  - Worst case scenario: you can't even find the biologist favorite gene
- More cells is not always better
- Sequencing depth should be the same for all samples



Sample Preparation

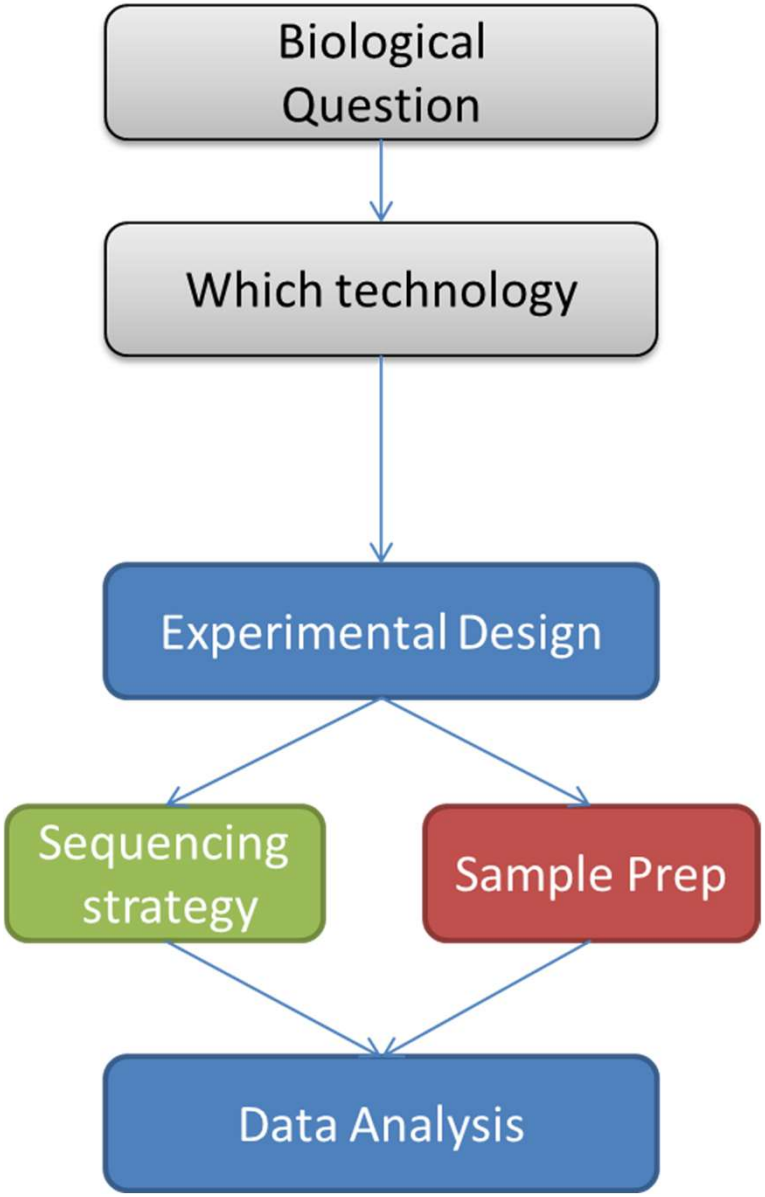
- We have a question
- We have selected a protocol

- How many samples?

- How many cells?

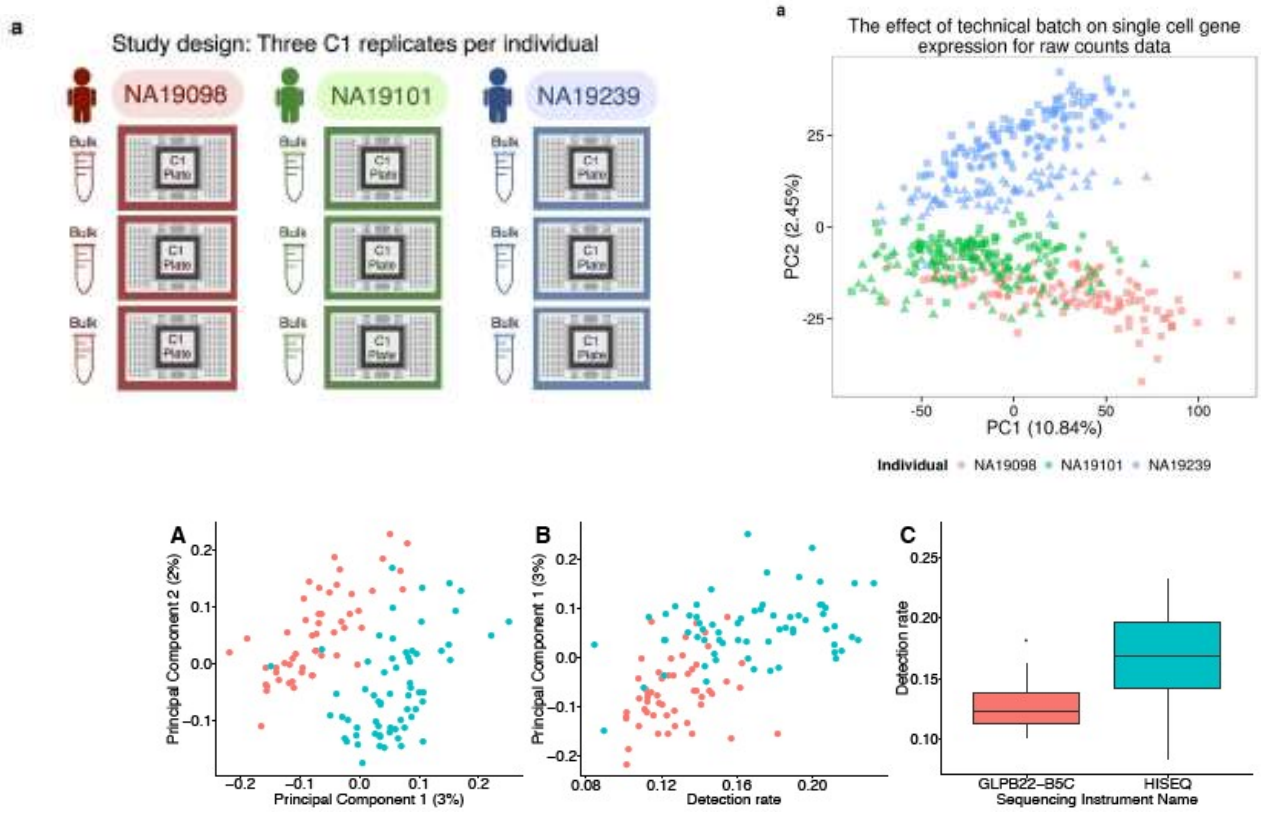
- How many reads/cell?

- How do we combine all this to minimize batch effect?



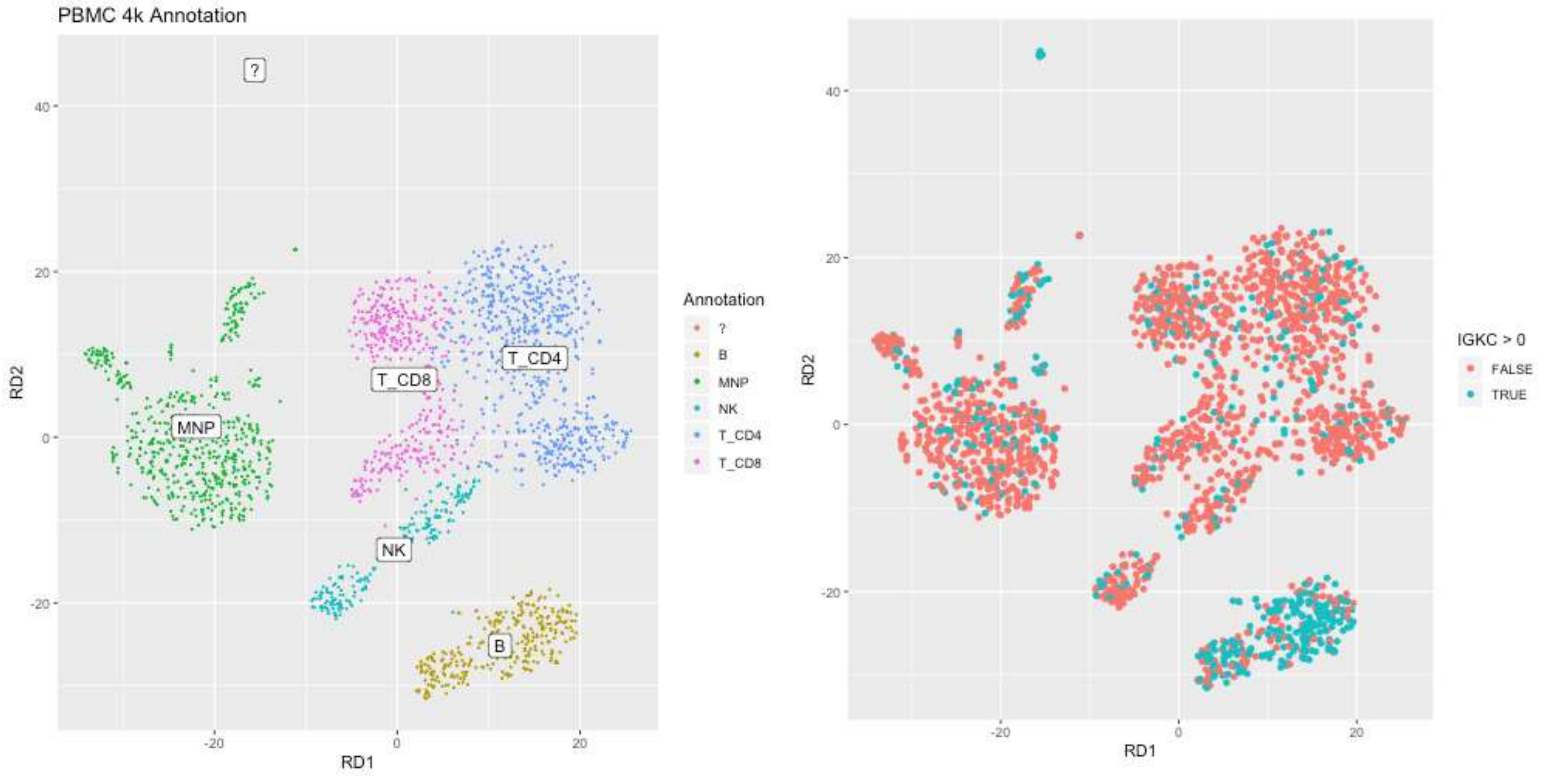
What about experimental confounding factors?

- scRNA-seq are often performed 1 sample at a time
  - Dissociation is difficult, sample are collected 1 by 1,…
  - Technological aspects vary too (seq depth, number of cells captured)
- Several studies report evidence for strong batch effects

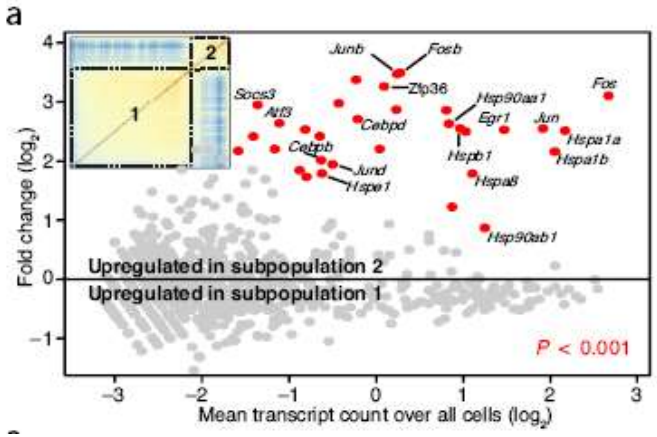


Hicks , 2017

Ambient RNA / Dissociation induced genes



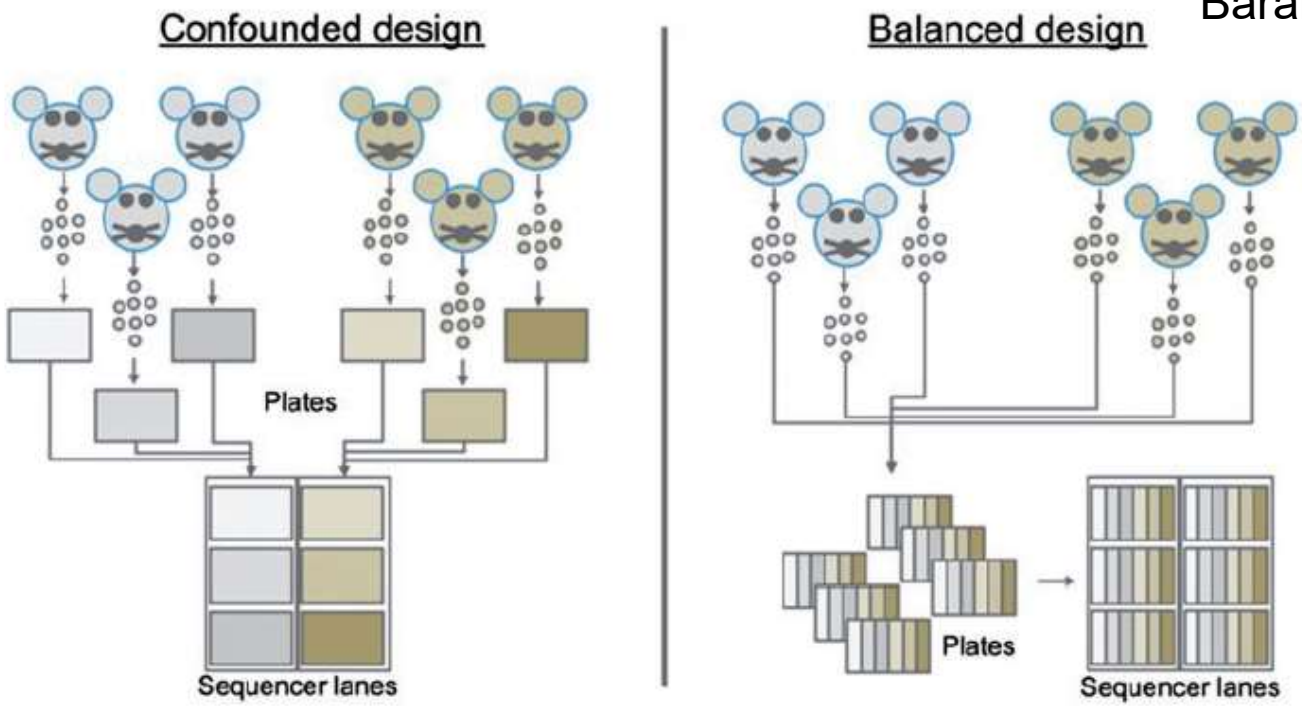
soupX tutorial
Young, 2020



Van den Brick, Nat Method 2017

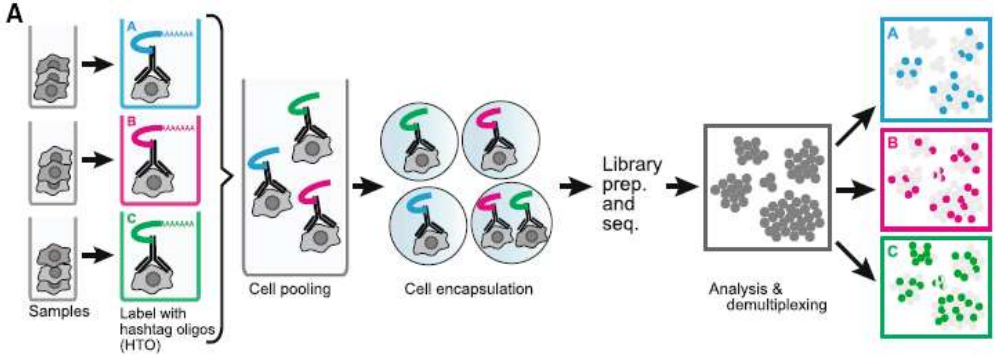
Aim for a Balanced Study Design

Baran-Gale 2018



- Balanced design may be hard to achieve for practical reasons

- Multiplexing :
  - Natural SNPs (demuxlet)
  - Expression of Xist/ChrY
  - **Cell-hashing**



Stoeckius, 2018

Example: Mouse Cell Atlases

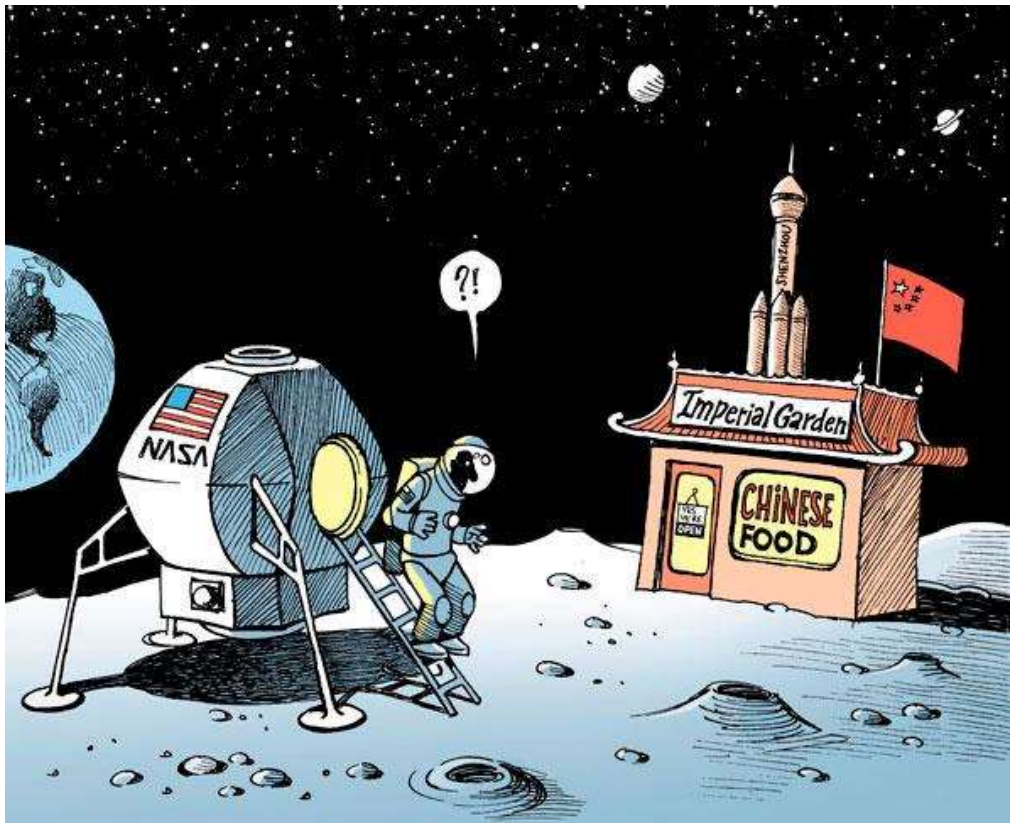
Marin Truchi, IPMC

ARTICLE

https://doi.org/10.1038/s41586-018-0590-4

Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*

The Tabula Muris Consortium*

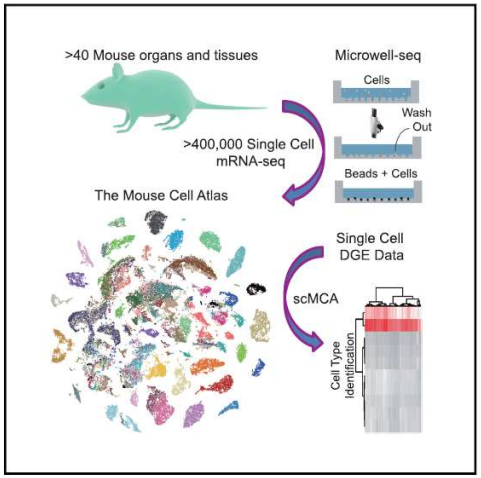


Resource

Cell

Mapping the Mouse Cell Atlas by Microwell-Seq

Graphical Abstract



Authors

Xiaoping Han, Renying Wang, Yincong Zhou, ..., Guo-Cheng Yuan, Ming Chen, Guoji Guo

Correspondence

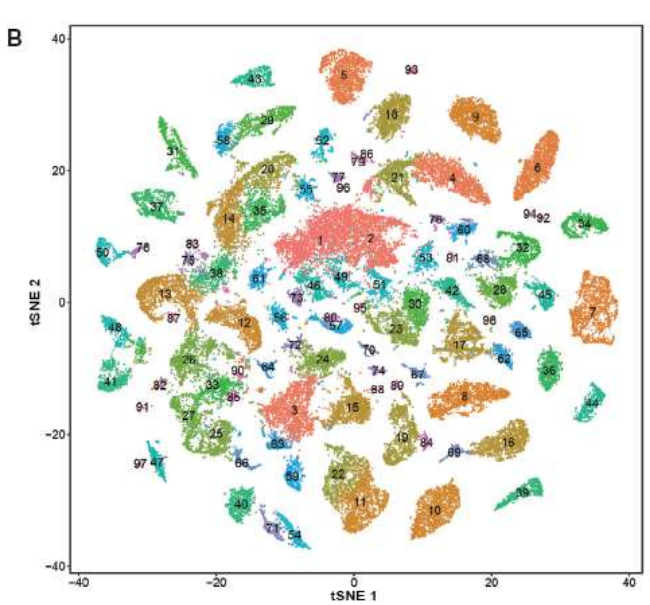
xhan@zju.edu.cn (X.H.), ggj@zju.edu.cn (G.G.)

In Brief

Development of Microwell-seq allows construction of a mouse cell atlas at the single-cell level with a high-throughput and low-cost platform.

Mouse Atlas Summary

MCA



Microwell seq

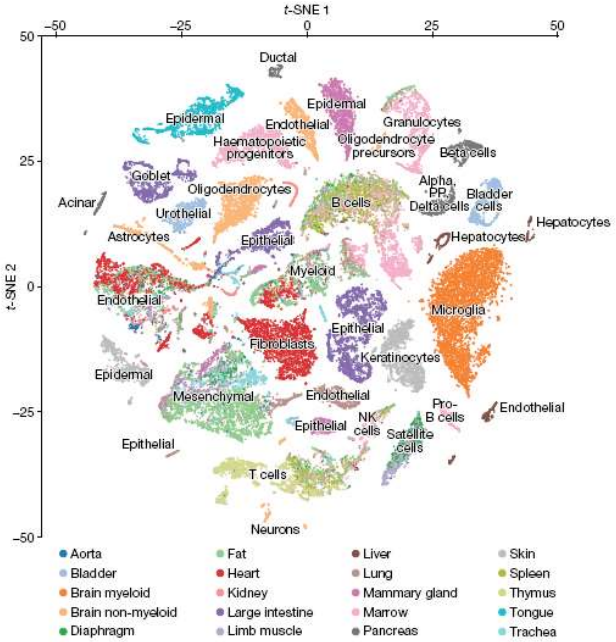
> 400,000 cells

>50 mouse tissues and cultures

> 800 cell types identified

based on 60,000 good QC cells

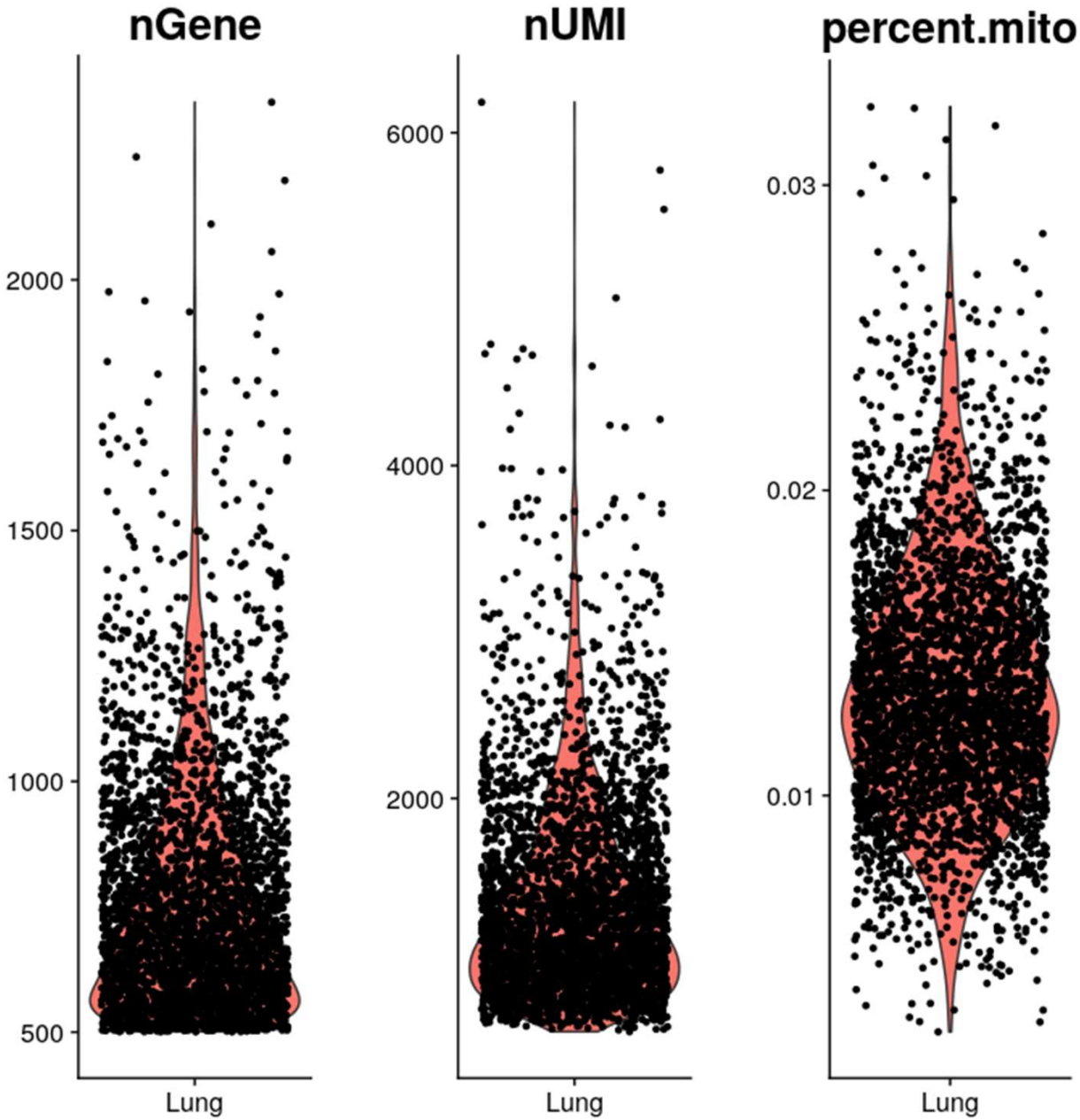
Tabula Muris



- Over 100,000 cells
- 20 organs
- *Double design*:
  - Shallow profiling using droplets
  - FACS + full length profiling

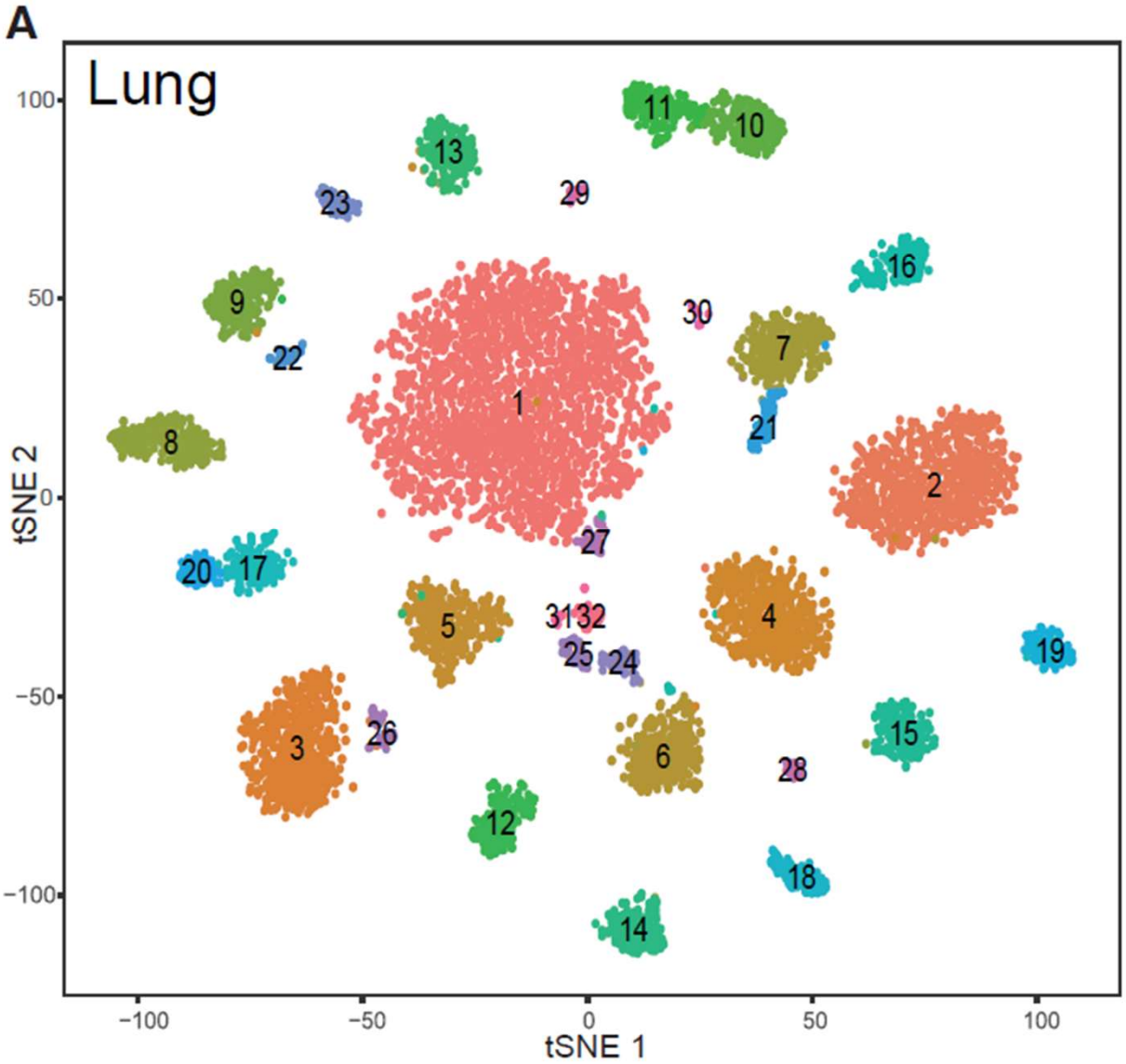
MCA Lung data (6940 cells)

Han et Al, Cell (2018)



Dropouts
96 %

MCA Lung data (6940 cells)



- **31 cell types (21 immune)**

1 AT2 Cell
2 B Cell
3 Alveolar macrophage
4 T Cell_*Cd8b1* high
5 Plasmacytoid dendritic cell
6 NK Cellhigh
7 Eosinophil granulocyte
8 Interstitial macrophage
9 Dendritic cell_*Naaa* high
10 Stromal cell_*Dcn* high
11 Stromal cell_*Inmt* high
12 Clara Cell
13 Conventional DC_*Gngt2* high
14 AT1 Cell
15 Conventional DC_*Ccl17* cell
16 Nuocyte
17 Endothelial cell
18 Ciliated cell
19 Endothelial cell_*Kdr* high
20 Endothelial cells_*Vwf* high
21 Neutrophil granulocyte
22 Dividing dendritic cells
23 Stromal cell_*Acta2* high
24 Dividing T cells
25 Dividing cells
26 Alveolar macrophage
27 Alveolar bipotent progenitor
28 Ig−producing B cell
29 Conventional DC
30 Basophil
31 Monocyte progenitor cell
32 Conventional DC

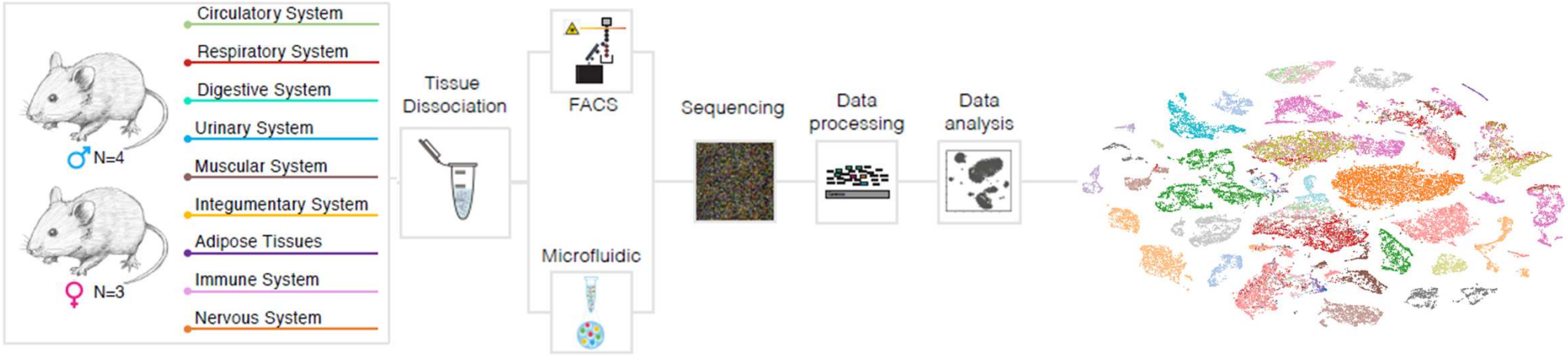
Gene expression and cell type markers available on :
http://bis.zju.edu.cn/MCA/gallery.html?tissue=Lung

ARTICLE

https://doi.org/10.1038/s41586-018-0590-4

Single-cell transcriptomics of 20 mouse organs creates a *Tabula Muris*

The Tabula Muris Consortium*



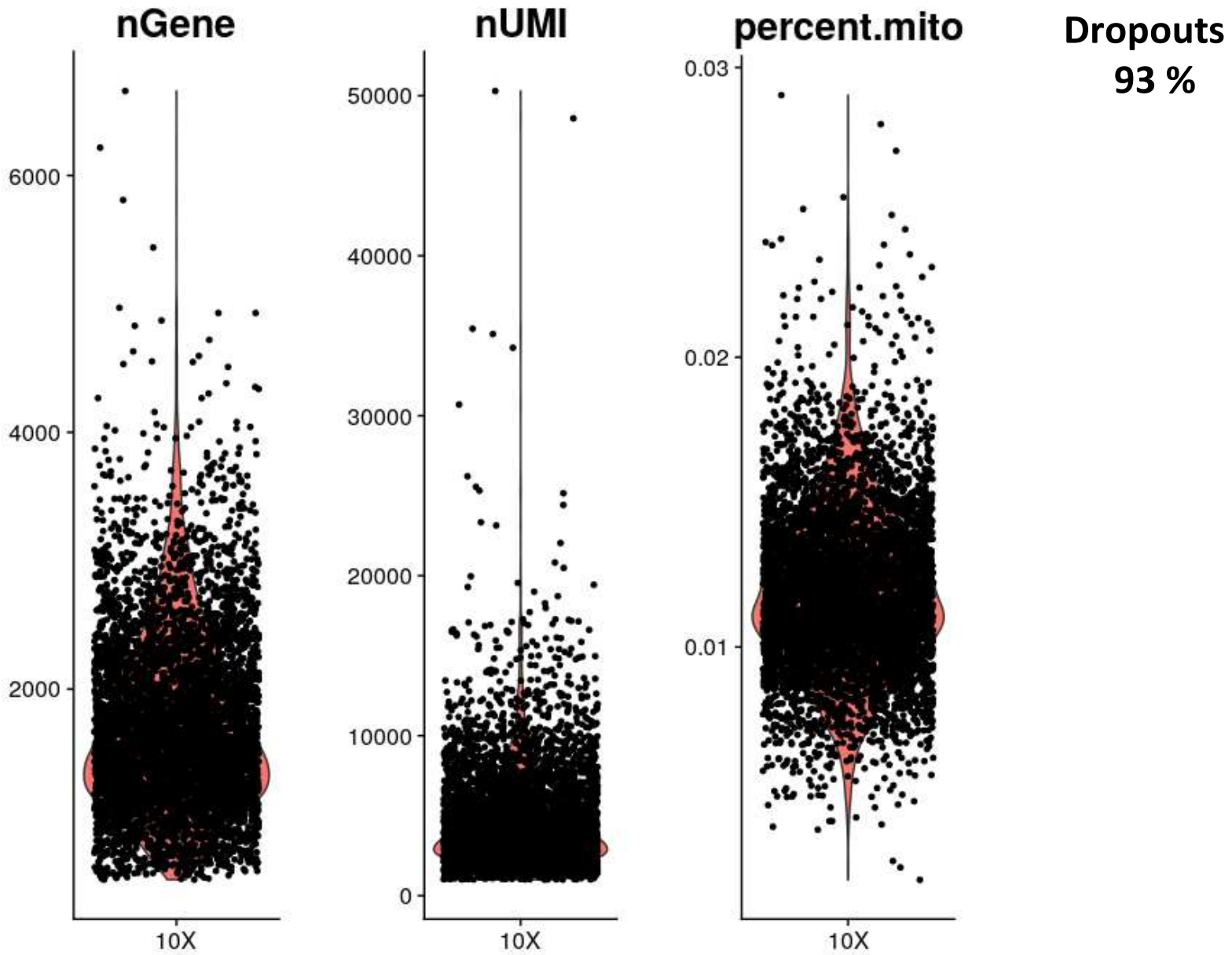
SMART-SEQ + FACS

Lung	Trachea
1620 cells	1392 cells

10X Microfluidic droplet

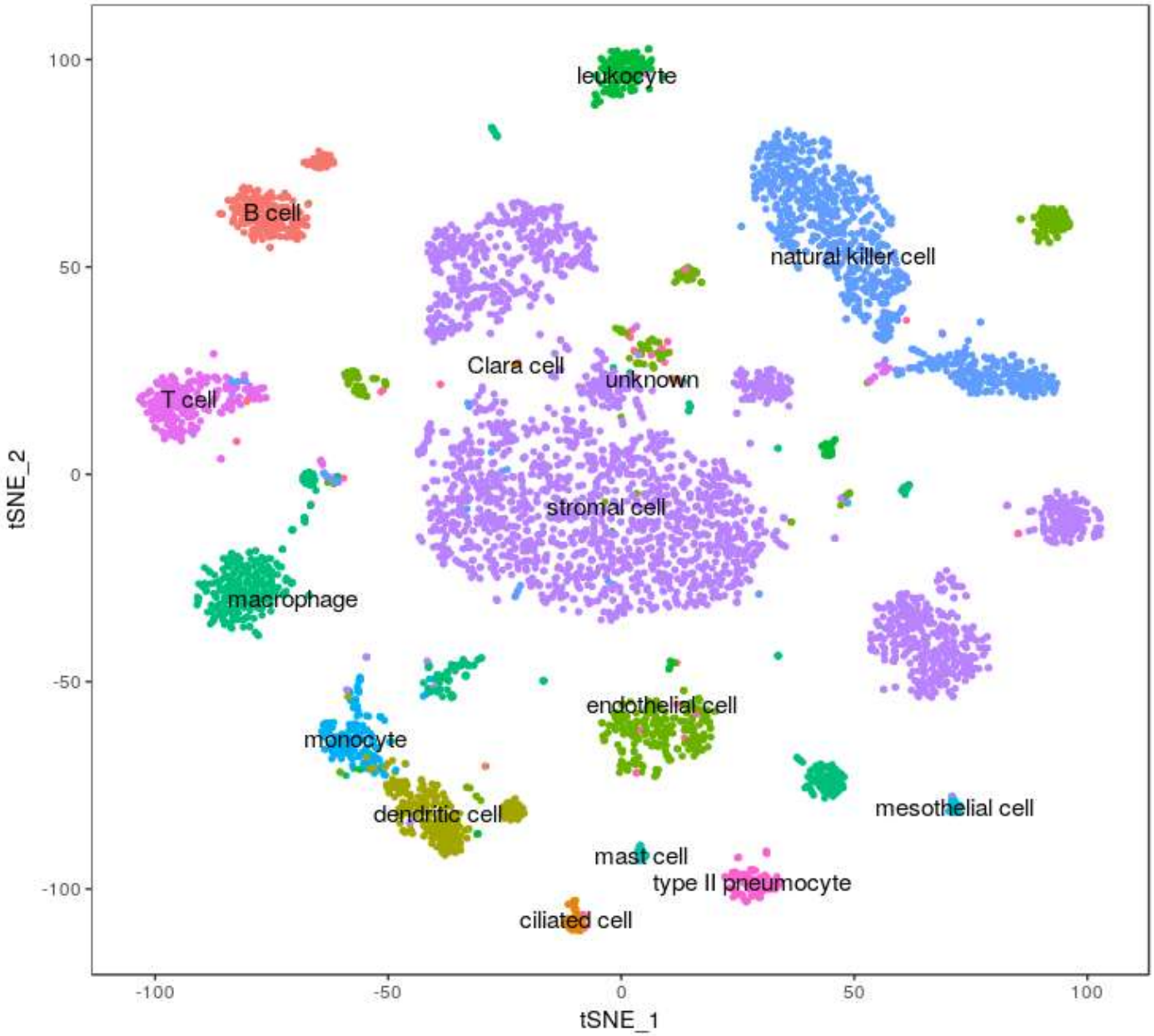
Lung	Trachea
5449 cells	11269 cells

TM Lung 10X data (**5449 cells)**



Dropouts
93 %

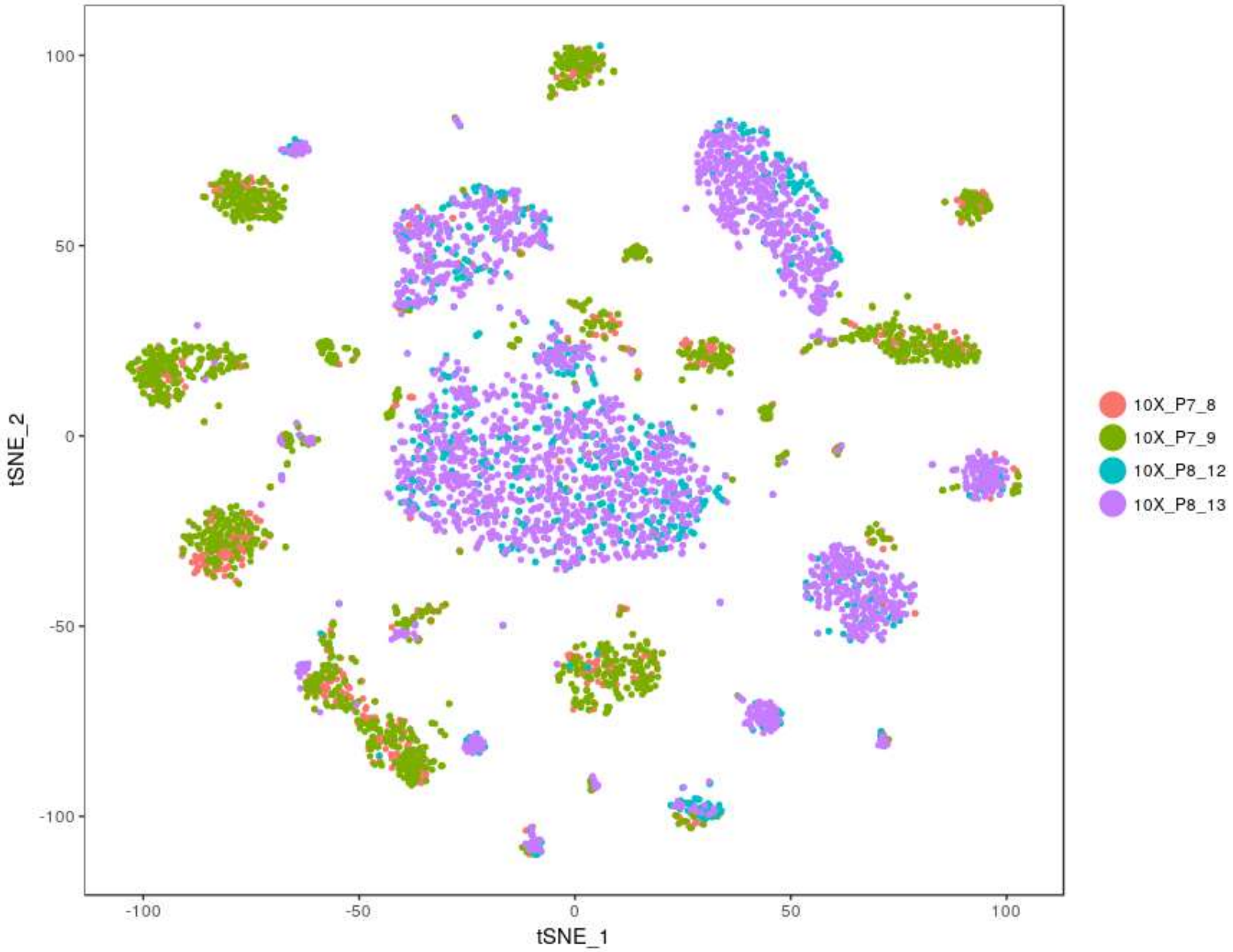
TM Lung 10X data (**5449 cells)**



- **15 cell types (8 immune)**

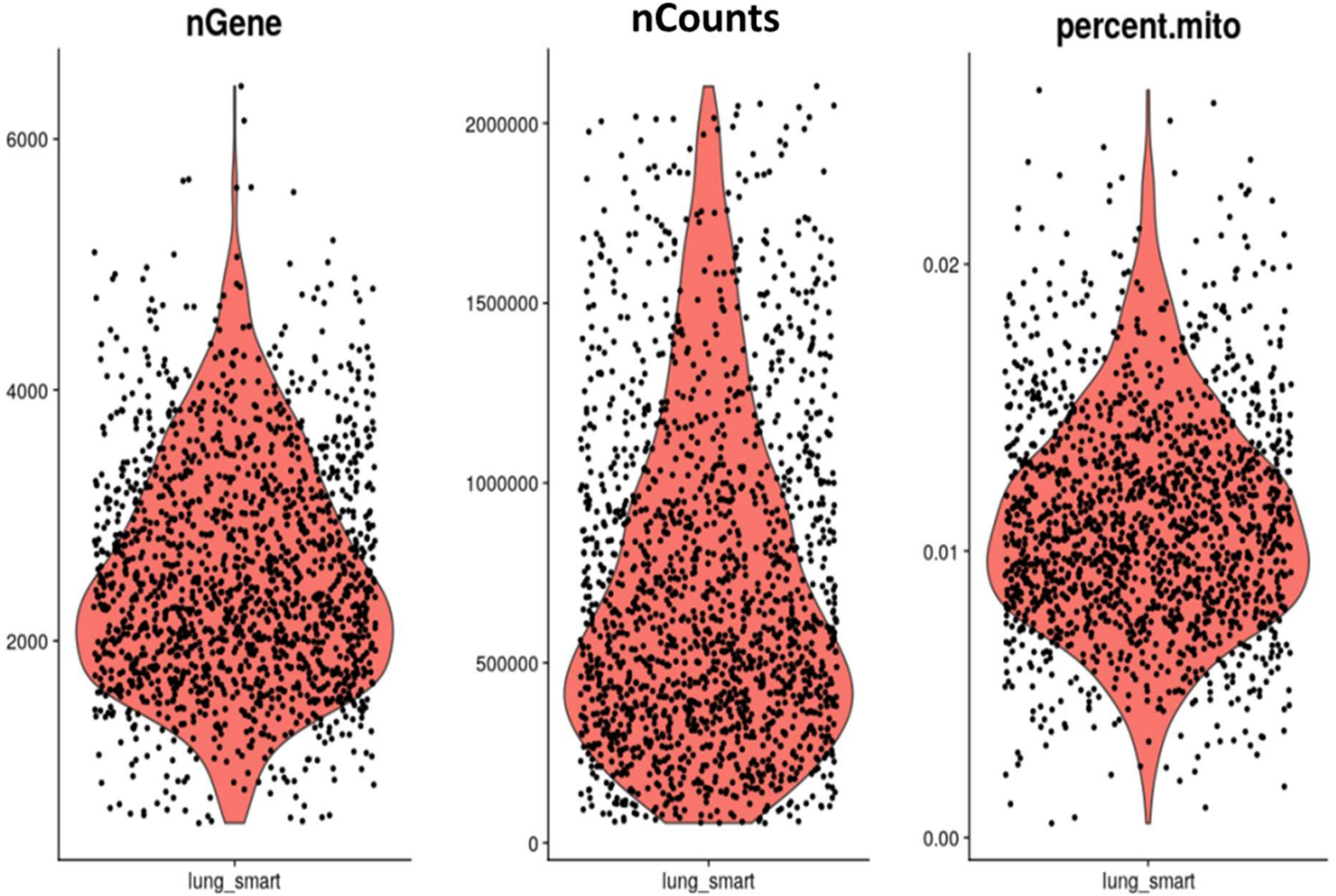
B cell	n = 205
ciliated cell	n = 41
Clara cell	n = 5
dendritic cell	n = 225
endothelial cell	n = 425
leukocyte	n = 151
macrophage	n = 456
mast cell	n = 22
mesothelial cell	n = 24
monocyte	n = 145
natural killer cell	n = 832
stromal cell	n = 2534
T cell	n = 246
type II pneumocyte	n = 89
unknown	n = 49

TM Lung 10X data (**5449 cells)**

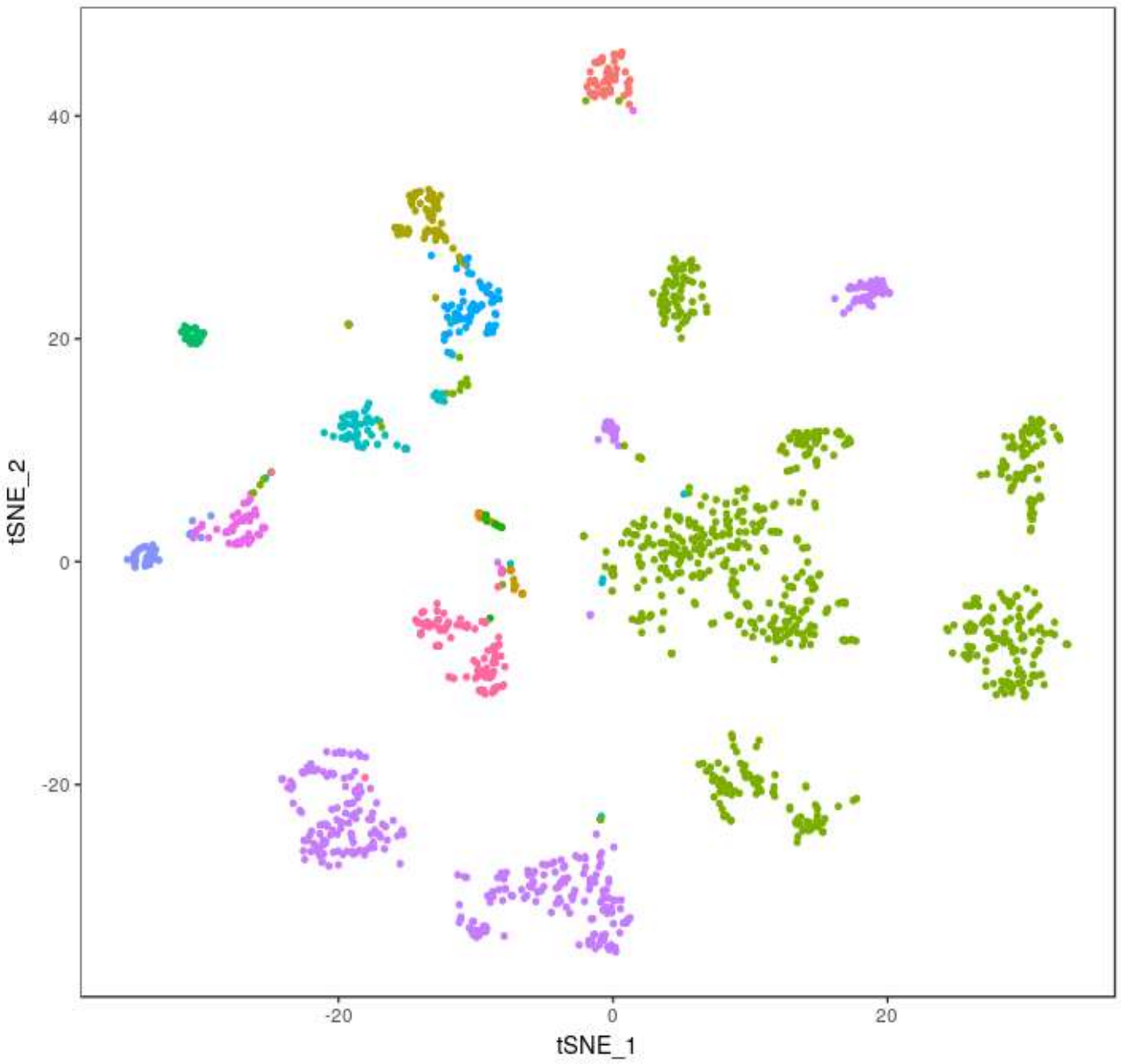


TM Lung SMART-Seq data (1620 cells)

Dropouts
89 %



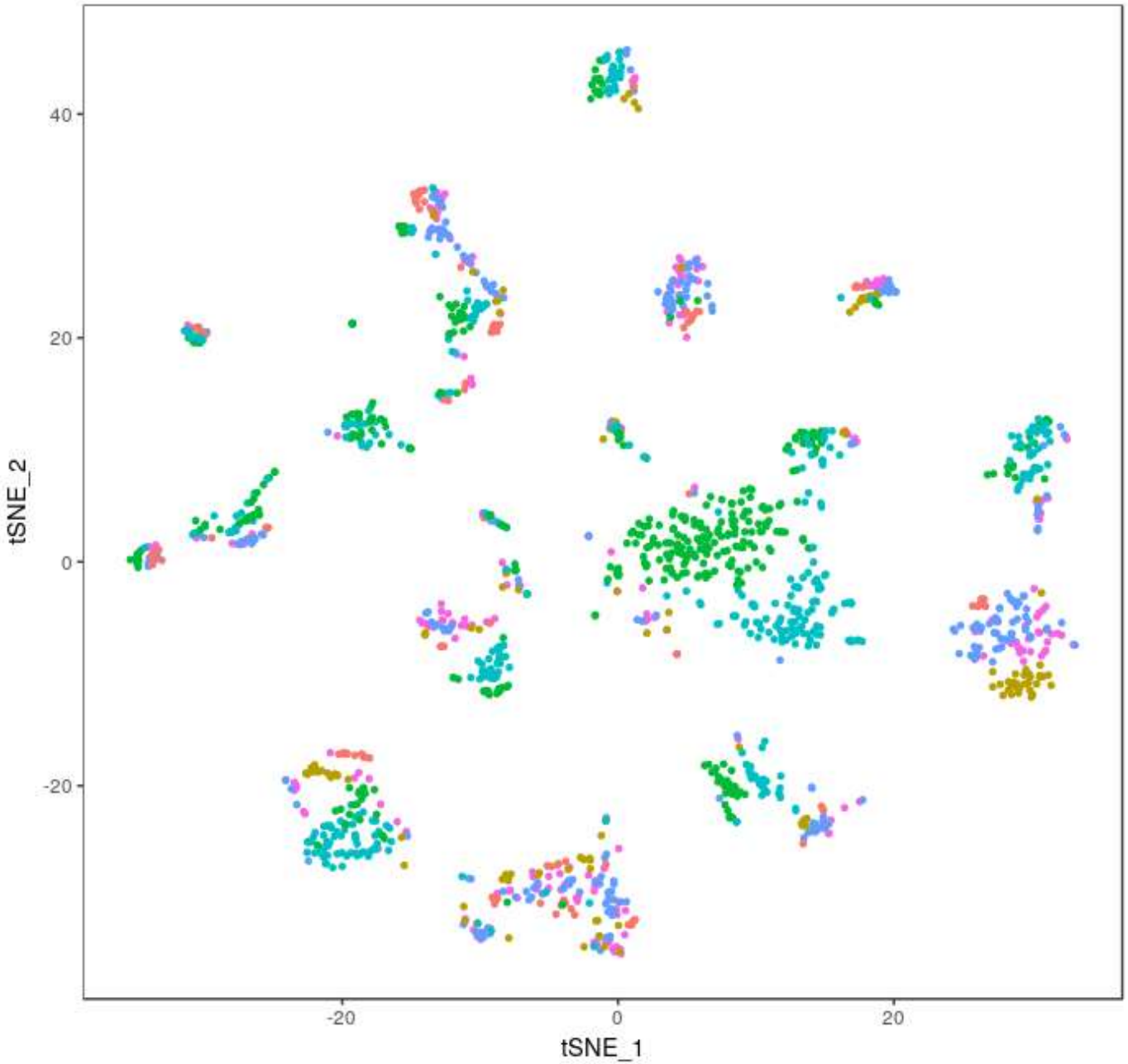
TM Lung SMART-Seq data (1620 cells)



- **16 cell types (7 immune)**

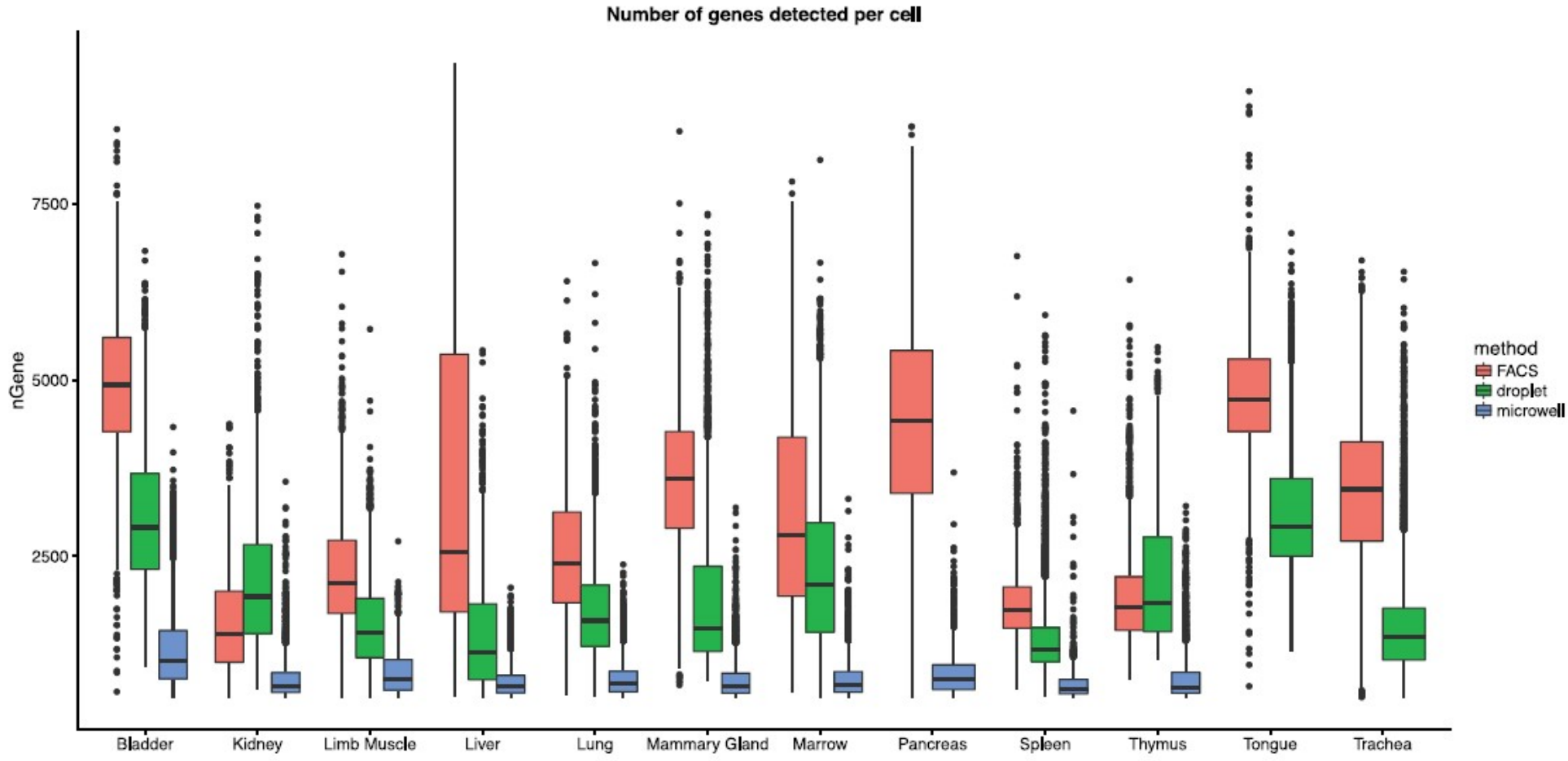
B cell
ciliated cell
Clara cell
dendritic cell
endothelial cell
epithelial cell
leukocyte
lung neuroendocrine cell
macrophage
mesothelial cell
monocyte
natural killer cell
stromal cell
T cell
type I pneumocyte
type II pneumocyte

TM Lung SMART-Seq data (1620 cells)



- **6 mice**

Mouse Atlases Sequencing depth comparison



Tabula Muris, 2018

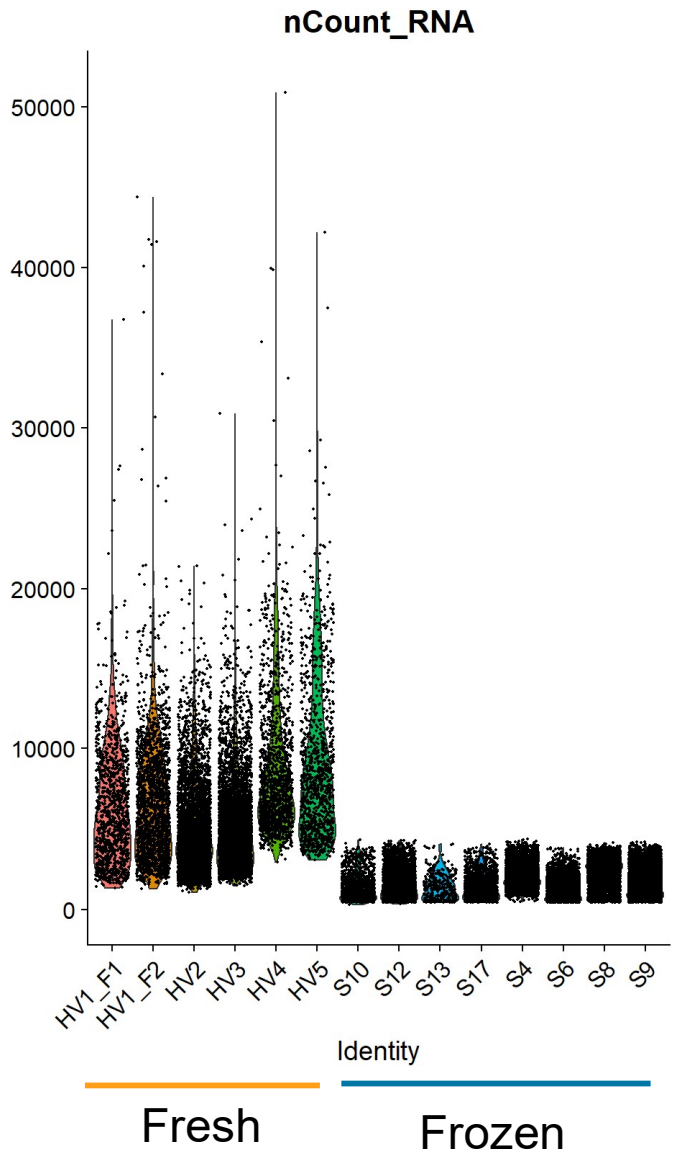
Use Existing Data to Select a Protocol

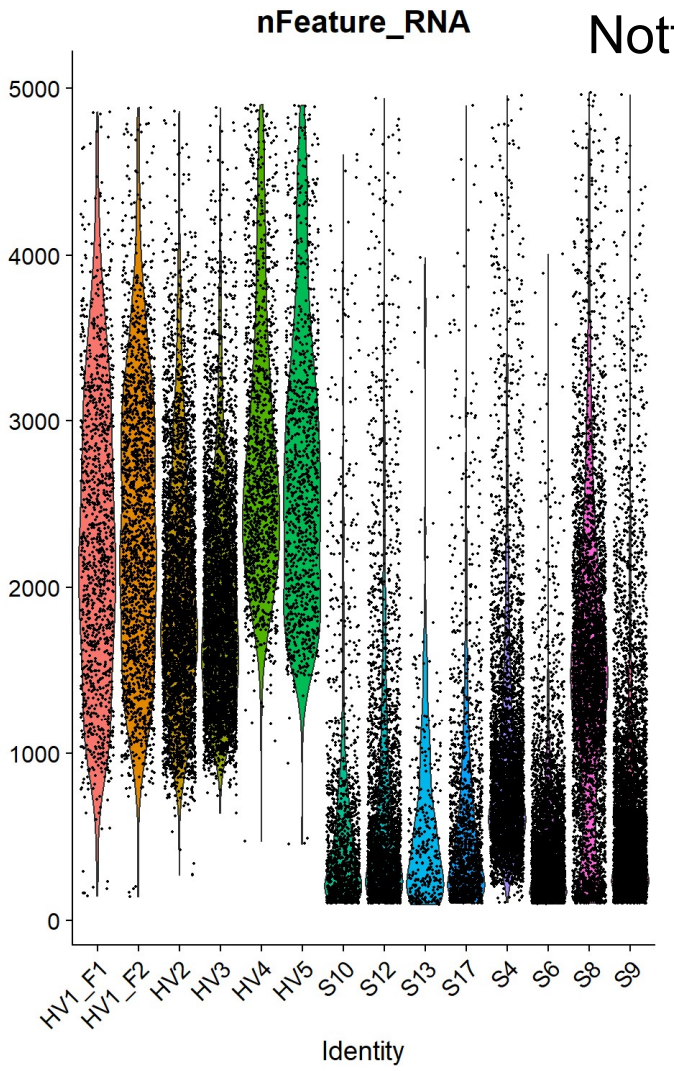
- Our collaborator is thinking about setting up a small clinical trial to study a skin disease
- She is asking for advice regarding sample collection and preparation for scRNASeq

- Clinical sample :
  - Samples collected and processed 1 by 1 if using fresh tissue
  - Some cell types are known to be degraded when frozen

- Using GEO, we reanalized 2 studies with healthy skin tissue
  - Fresh samples: GSE132802
  - Frozen samples: GSE147424

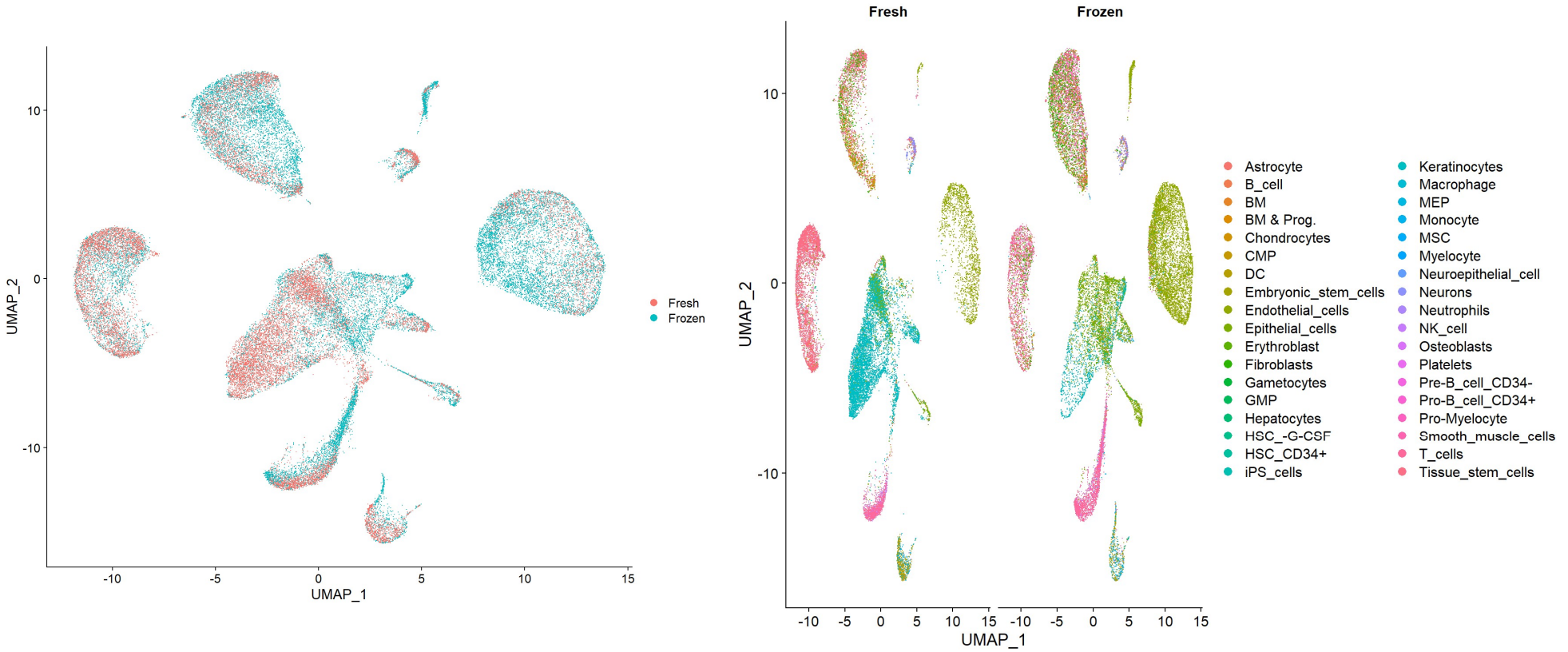
Difference in data quality is clear





Nottet, Syneos Health

Cell Type Identification



- All cell types are present in both datasets (but proportions vary)
- Differential analysis fresh vs frozen did not show a lot of DE genes

- Frozen tissue can be a solution here. A higher sequencing depth could be recommended

Conclusion

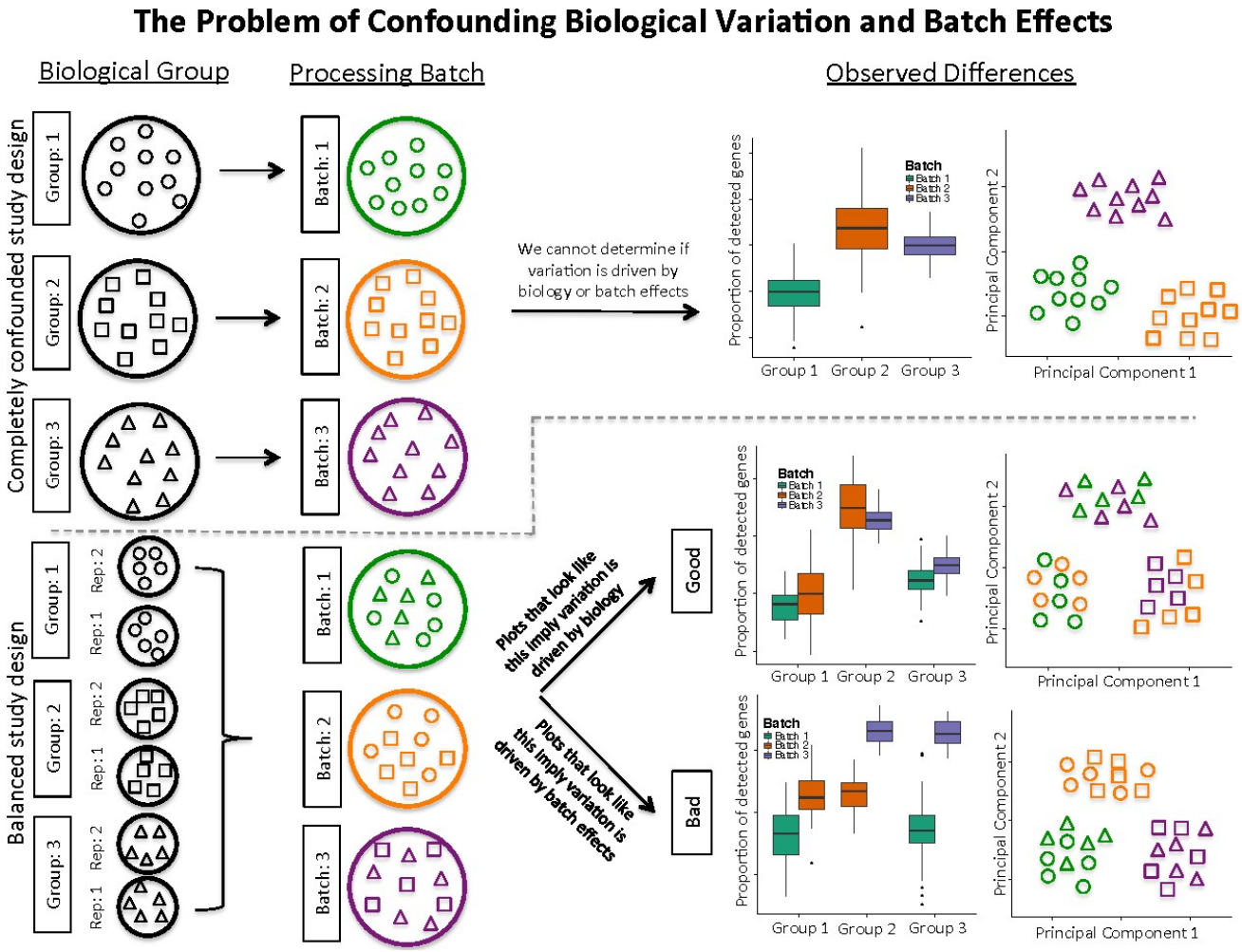
- Single Cell RNAseq data are very sensitive
- Sample/Batch effects can be very strong
- Hard problem to correct in downstream analysis
  - Batch/conditions are confounded

- New protocols based on frozen/FFPE tissue + multiplexing are available to reduce the confounding factors

References

- Hyeongseon Jeon et al., Statistical Power Analysis for Designing Bulk, Single-Cell, and Spatial Transcriptomics Experiments: Review, Tutorial, and Perspectives. Biomolecules 2023
- Svennson V et al, Power analysis of single-cell RNA-sequencing experiments, Nature Methods 2017
- Baran-Gale et al, Experimental design for single-cell RNA sequencing, Brief Functional Genomics 2017
- Tung PY et al, Batch effects and the effective design of single-cell gene expression studies, Science Reports 2017
- Arguel MJ et al, A cost effective 5 selective single cell transcriptome profiling approach with improved UMI design, Nuc Acid Res, 2017
- Chen at al, UMI-count modeling and differential expression analysis for single-cell RNA sequencing, Genome Biol 2018
- Grün D et al, Validation of noise models for single-cell transcriptomics, Nat Method 2014
- Ziegenhain C et al, Comparative Analysis of Single-Cell RNA Sequencing Methods, Molecular Cell 2017
- Hicks SC, Missing data and technical variability in single-cell RNA-sequencing experiments; Biostatistics 2017
- Kang HM et al, Multiplexed droplet single-cell RNA-sequencing using natural genetic variation, Nature Biotech 2017
- Stoeckius M, Cell 'hashing' with barcoded antibodies enables multiplexing and doublet detection for single cell genomics, BiorXiv 2017
- Van den Brick S,Single cell sequencing reveals dissociation-induced gene expression in tissue subpopulations, Nat Method 2017

Thank you



Hicks, Biostatistics 2017