Experimental Design Normalization

Agnès Paquet

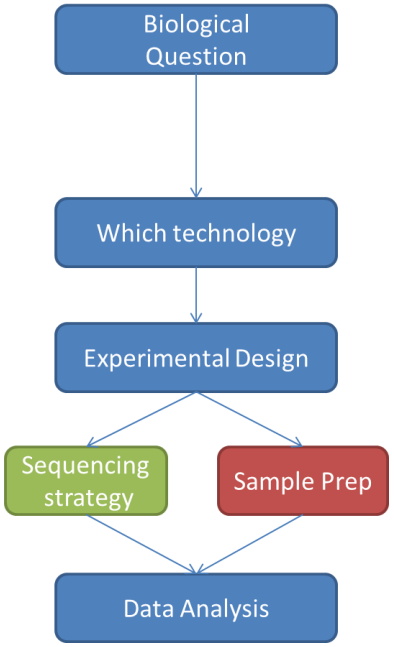
SincellTE 2022 - 01/10/2022

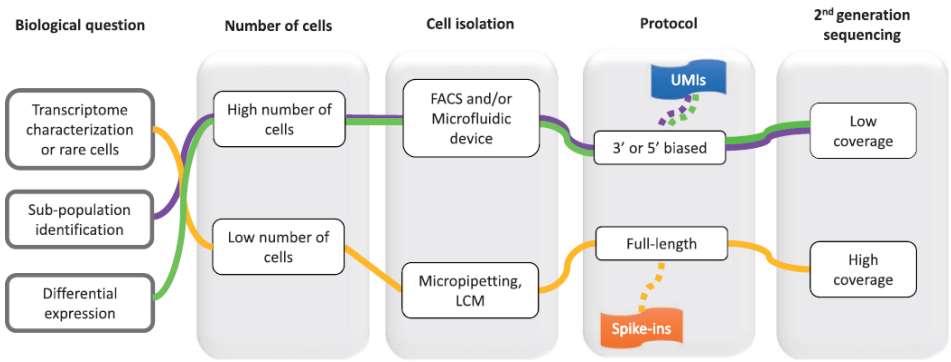
agnes.paquet@syneoshealth.com

Single RNAseq workflow: bioinformatics point of view

- 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
    - UMI design
    - Spike-ins
    - Sequencing strategy?
    - Number of cells
  - Samples: Practical considerations
    - Types /number of samples
    - Cell preparation -> *confounding*
    - Budget





Dal Molin, 2019

Technical point of view

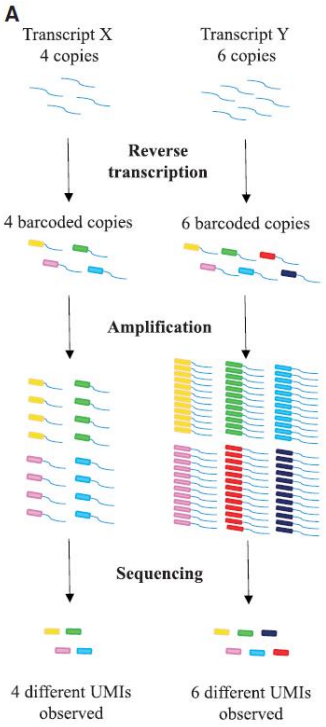
1. UMI design

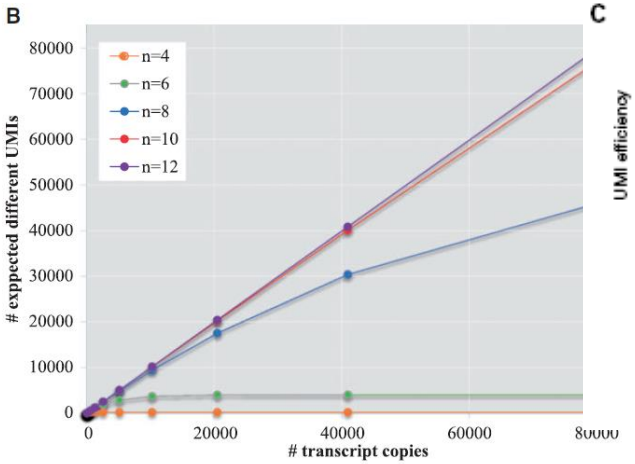
2. Use of Spike-ins

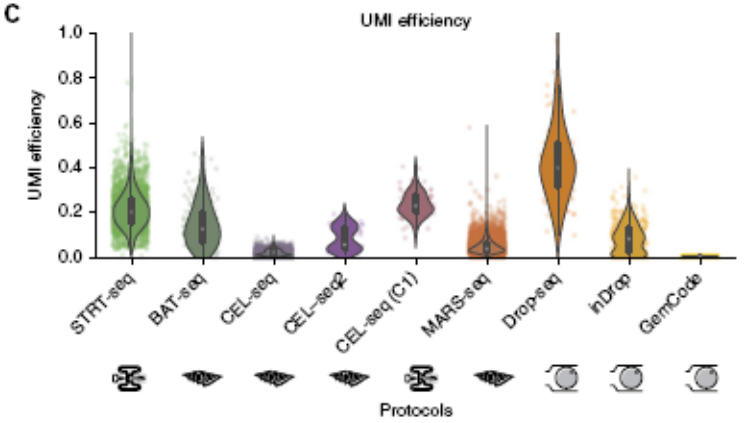
3. Discuss about sequencing design
   - Number of cells
   - Sequencing depth

UMI design

Dal Molin, 2019







Svensson V, et al 2017

- Unique Molecular Identifier (Islam et al, 2014)
- UMI-based protocols allow for PCR bias correction
- Improved accuracy of gene expression measures (E.g.: Chen, Genome Bio 2018)
- Design limits: be careful of saturation
  - N=4-10bp barcodes -> 4^N possible UMIs
  - N=5 -> 1024 UMIs available
  - N=10 -> 1,048,576 UMIs available

Spike-ins

- Spike-ins are molecules that are added in known concentration to the library

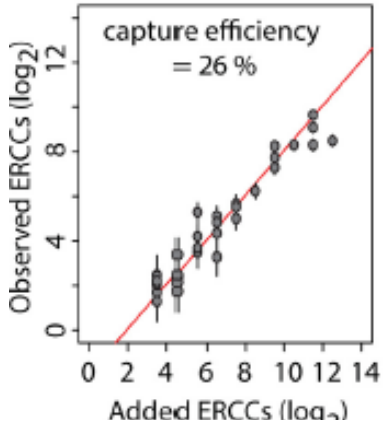
- Used to assess protocol accuracy and reproducibility

- ERCC
  - 92 bacterial RNA species, different lengths, GC contents
  - 22 abundance levels, 2 mixes for fold-change accuracy assessment

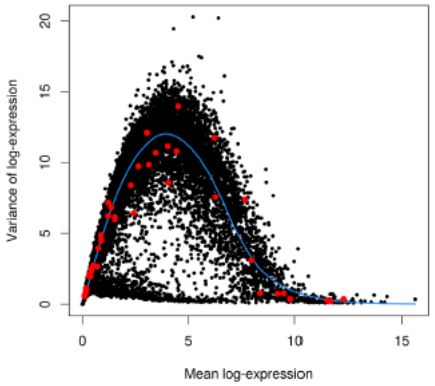
- SIRV
  - 69 artificial transcripts
  - Mimic human genes
  - Main difference: Used for isoforms detection

Spike-ins use in scRNA-seq

- Estimate protocol capture efficiency
  - How many of the spiked molecules did we detect?
- Comparison of protocols performance
  - Level of detection in low expressed genes
  - See Svensson V. et al, 2017

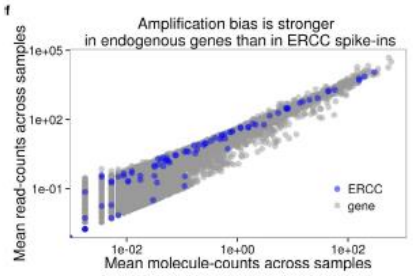


- Estimate technological noise
  - Help for detection of highly variable genes



- Issue 1: spike-ins behave differently than endogenous genes
  - May introduce more bias

- Issue 2: Spike-ins can't be used in droplet assays
  - Even incorporation in all droplets
  - Reads will be used to sequence only spike-ins



Experimental Design

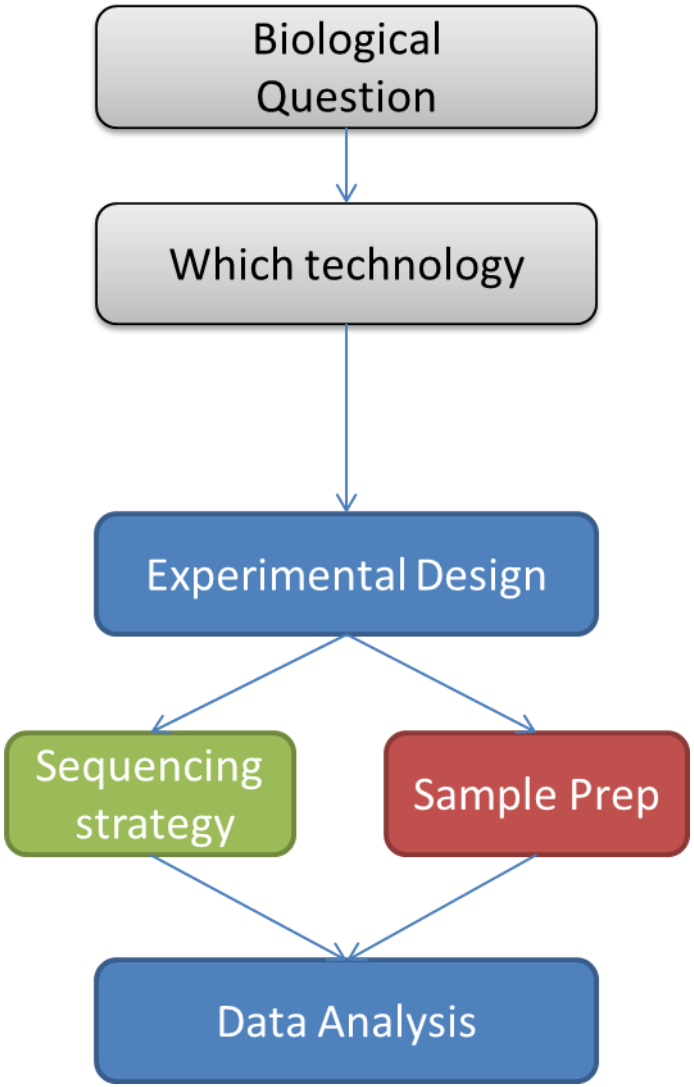
- We have a question
- We have selected a protocol

- How many samples?

- How many cells?

- How many reads/cell?

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

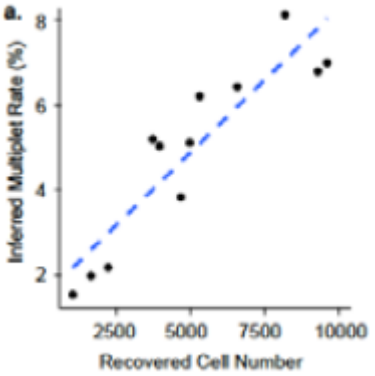


Estimating the required number of cells / sequencing depth

- 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*

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)

Estimated Number of Cells
6,388

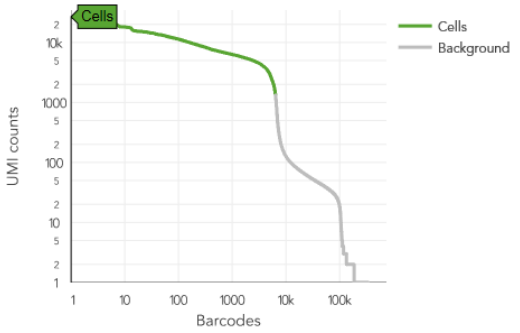
Mean Reads per Cell
61,047

Median Genes per Cell
1,343

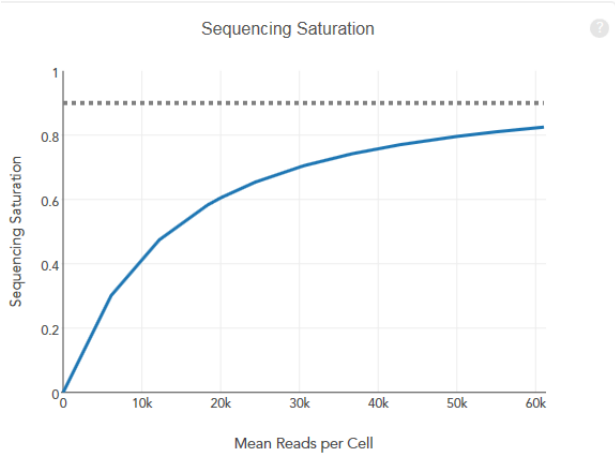
Sequencing

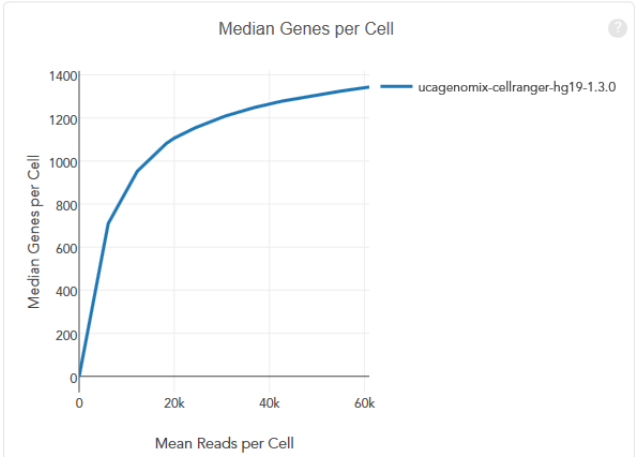
Number of Reads	389,969,360
Valid Barcodes	97.9%
Reads Mapped Confidently to Transcriptome	52.5%
Reads Mapped Confidently to Exonic Regions	54.6%
Reads Mapped Confidently to Intronic Regions	21.4%
Reads Mapped Confidently to Intergenic Regions	3.8%
Reads Mapped Antisense to Gene	3.8%
Sequencing Saturation	82.5%

Cells



Estimated Number of Cells	6,388
Fraction Reads in Cells	92.3%
Mean Reads per Cell	61,047
Median Genes per Cell	1,343
Total Genes Detected	21,143
Median UMI Counts per Cell	4,480





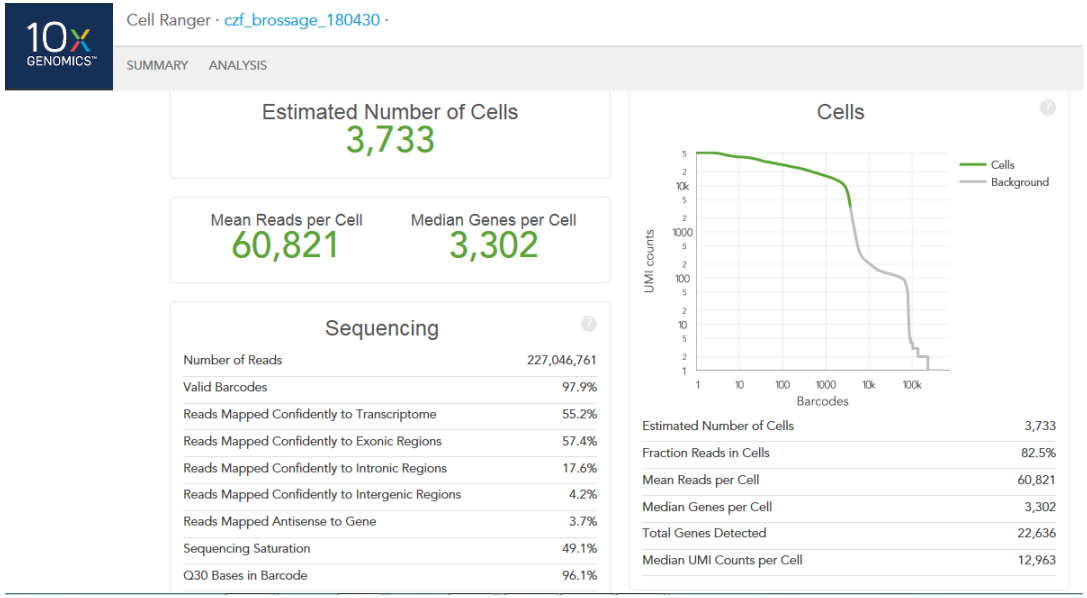
Example 2: Nasal epithelium brushing cells with lots of RNA

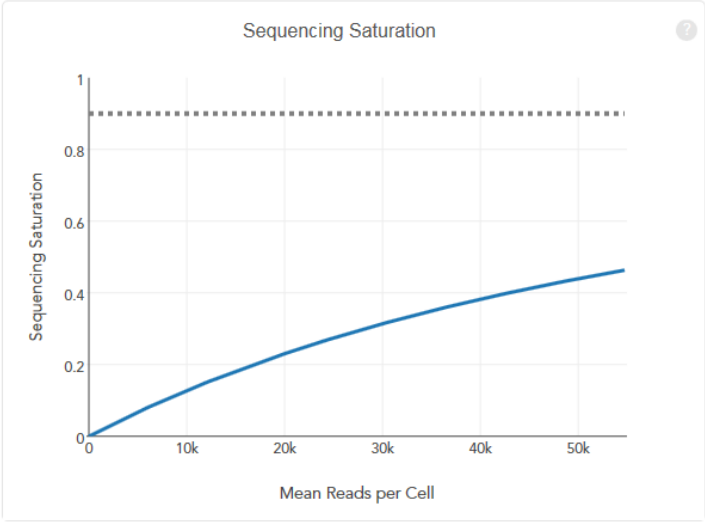
Target: 5,000 cells

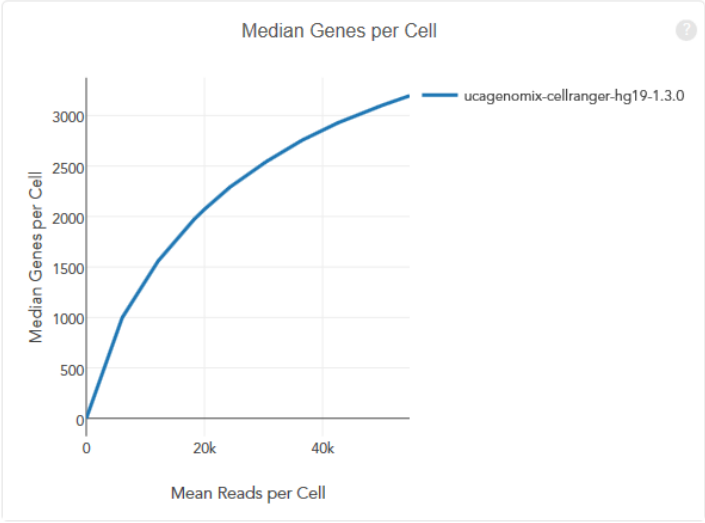
2 samples,

NextSeq High 75

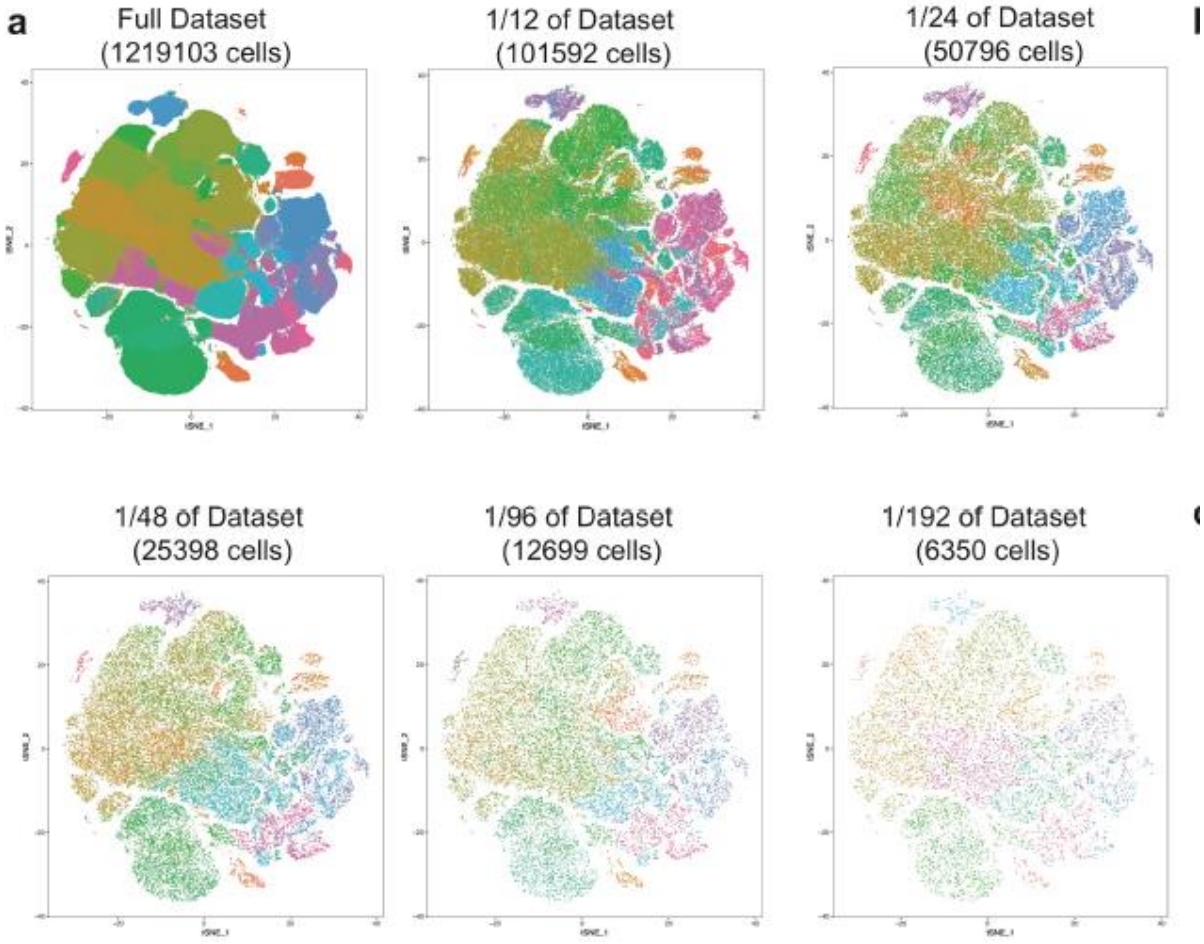
~400millions reads / run







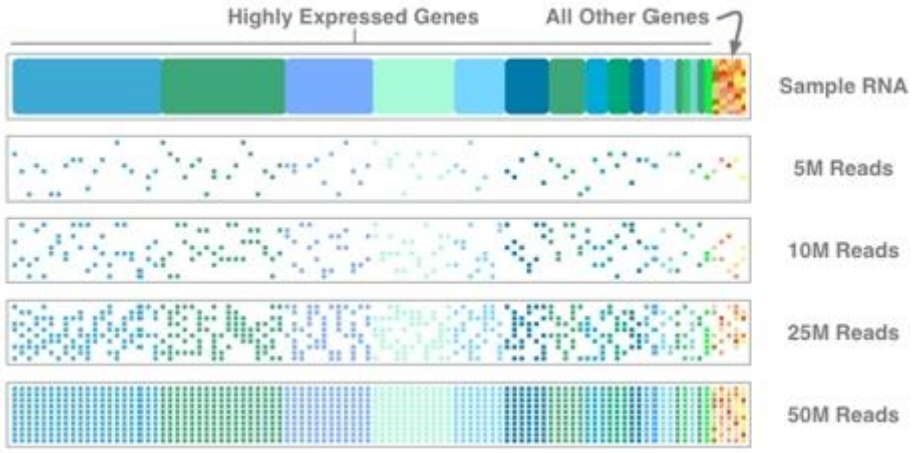
Number of cells: example of the 1.3millions cells dataset



Bhaduri A, BiorXiv 2017

Technical design: summary

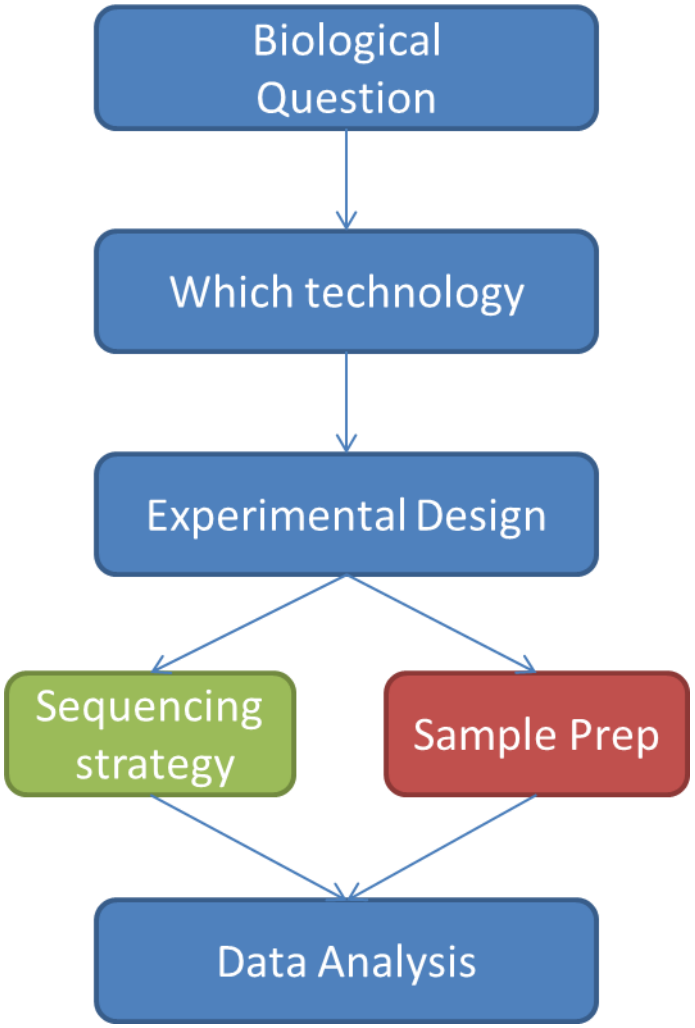
- Discuss about sequencing depth with the biologist
- If the sequencing is too shallow, the statistical analysis may not be robust
  - Worst case scenario: you can't even find the biologist favorite gene
- More cells is not always better
- Sequencing depth should be the same for all samples



Sample preparation

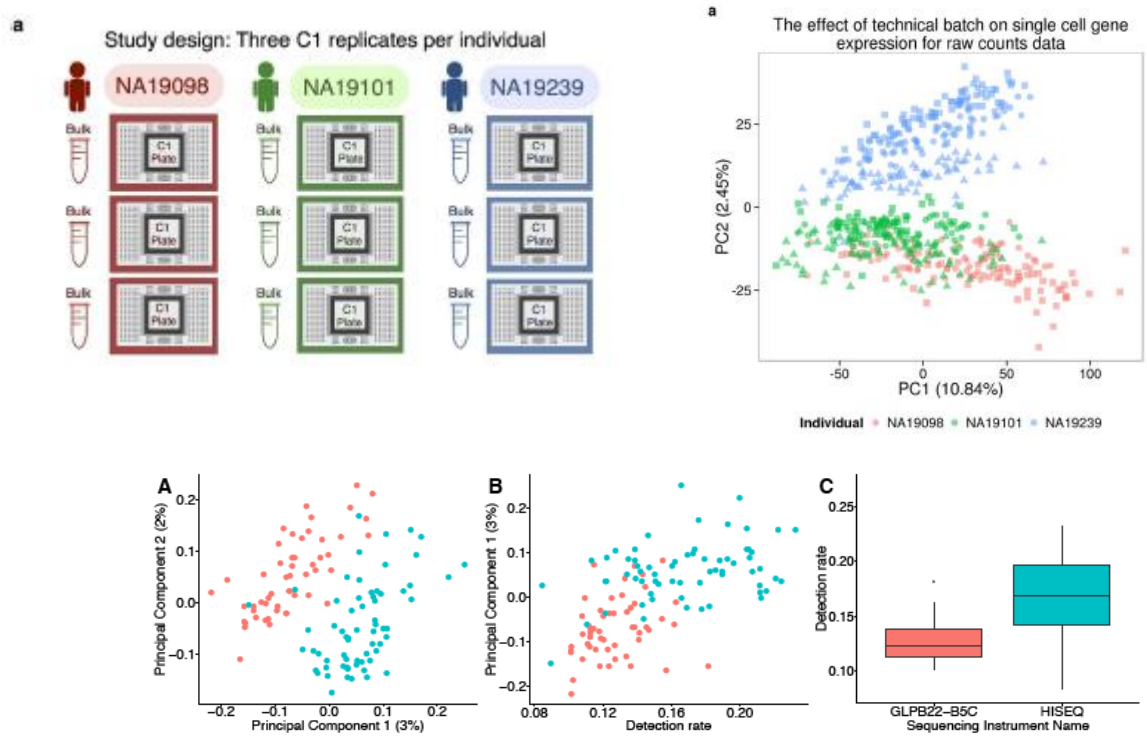
- What technique should we use to generate the data ?
  - Plate based / droplets
  - Full length / 3’ counting with UMI
  - ➢ UNDERSTAND THE BIAS

- Experimental design
  - Sequencing strategy
    - UMI design
    - Spike-ins
    - How to sequence
  - **Samples: Practical considerations**
    - Types /number of samples
    - Cell preparation -> *confounding*
    - Budget



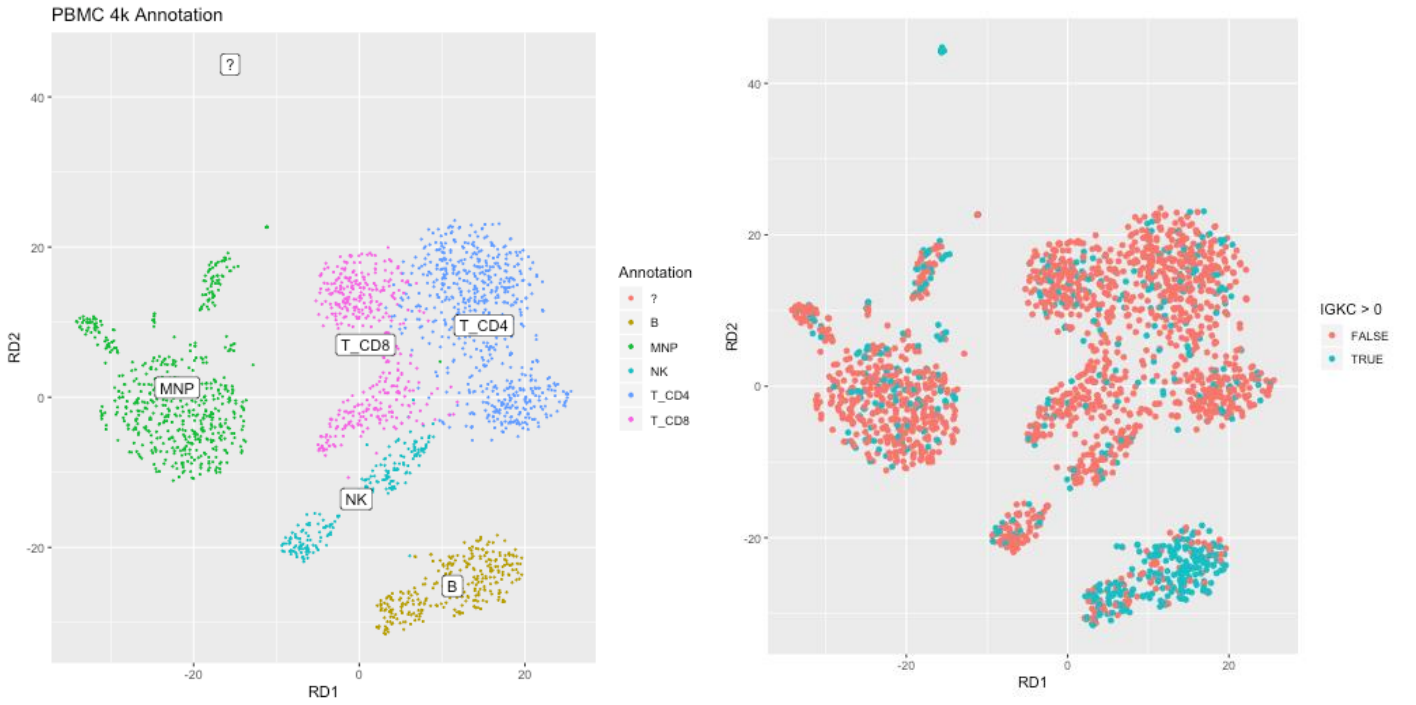
What about experimental confounding factors ?

- 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

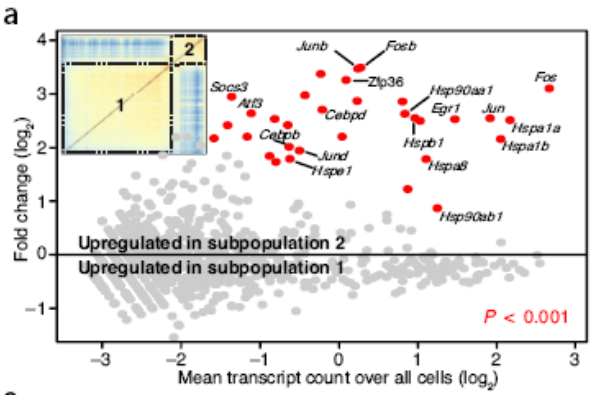


Hicks , 2017

Ambient RNA / Dissociation induced genes



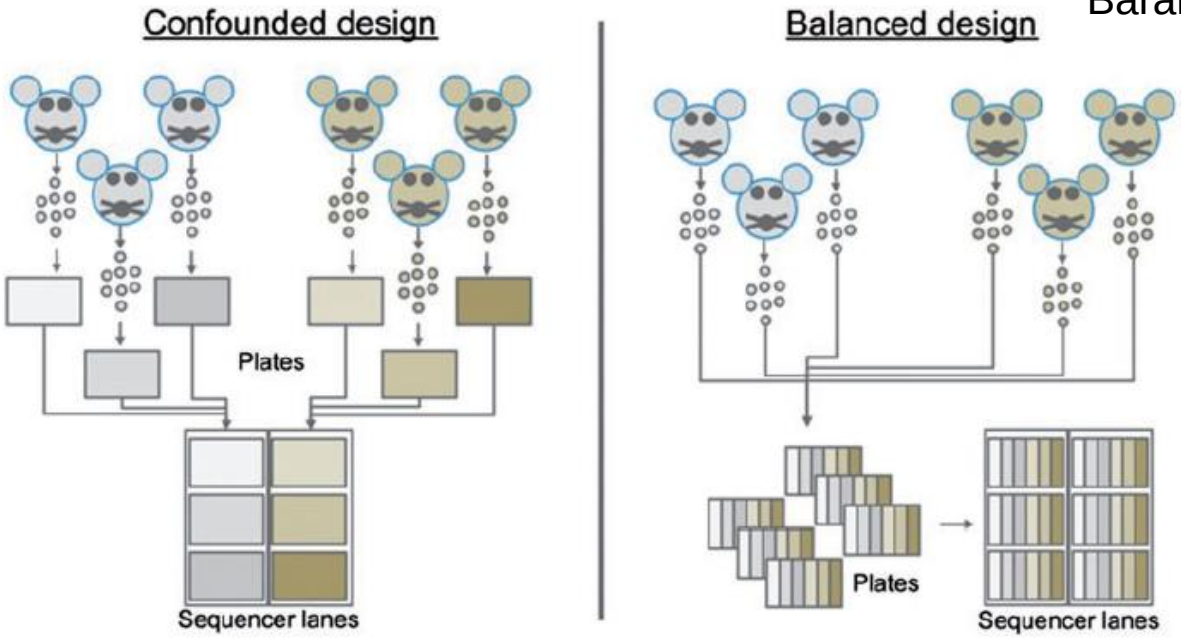
soupX tutorial
Young, 2020



Van den Brick, Nat Method 2017

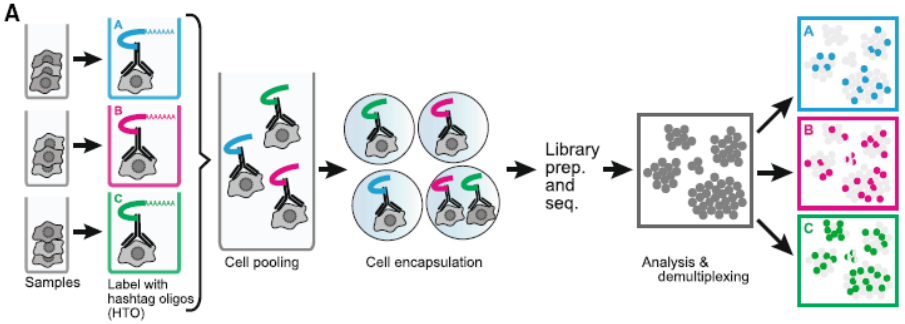
Perfect study design

Baran-Gale 2018



- Balanced design will be hard to achieve for practical reasons

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



Stoeckius, 2018

Example 1: Mouse Cell Atlases

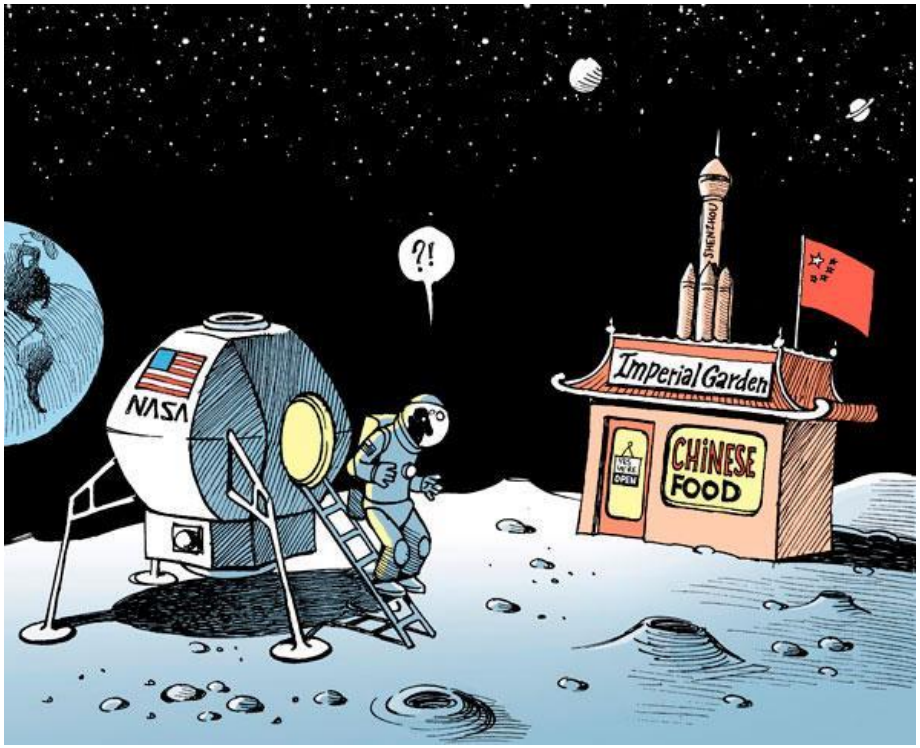
Marin Truchi, IPMC

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*

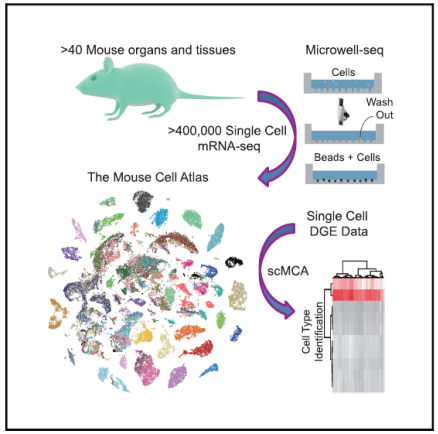


Resource

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

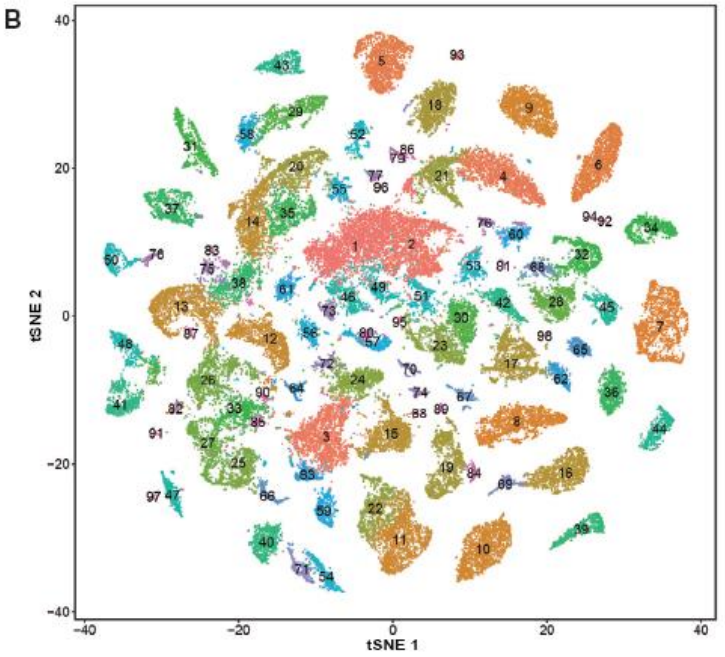
xhan@zju.edu.cn (X.H.), ggj@zju.edu.cn (G.G.)

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.

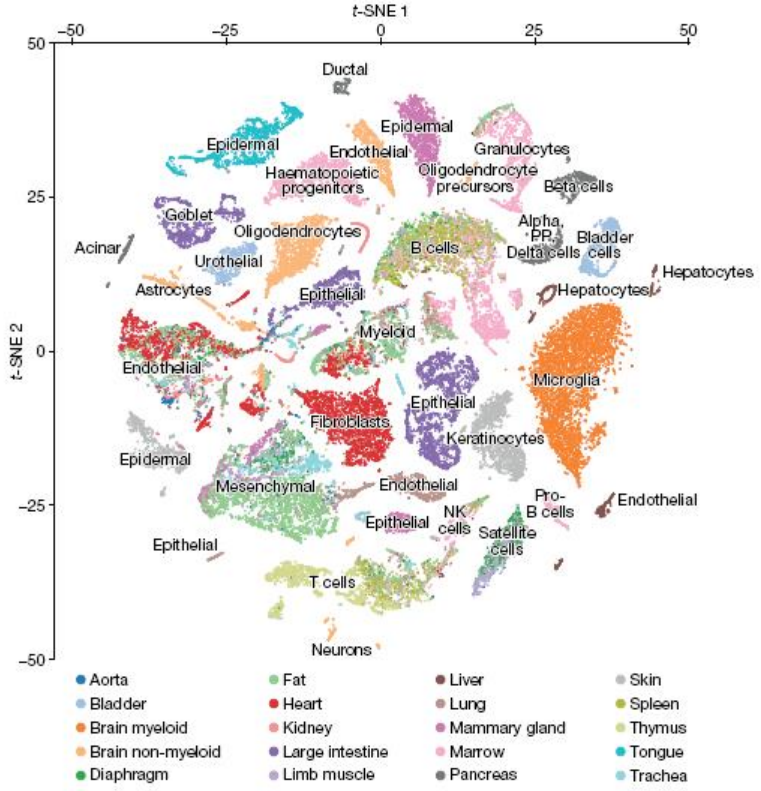
Mouse Cell Atlas Summary

- > 400,000 cells
- >50 mouse tissues and cultures
- > 800 cell types identified
  based on 60,000 good QC cells



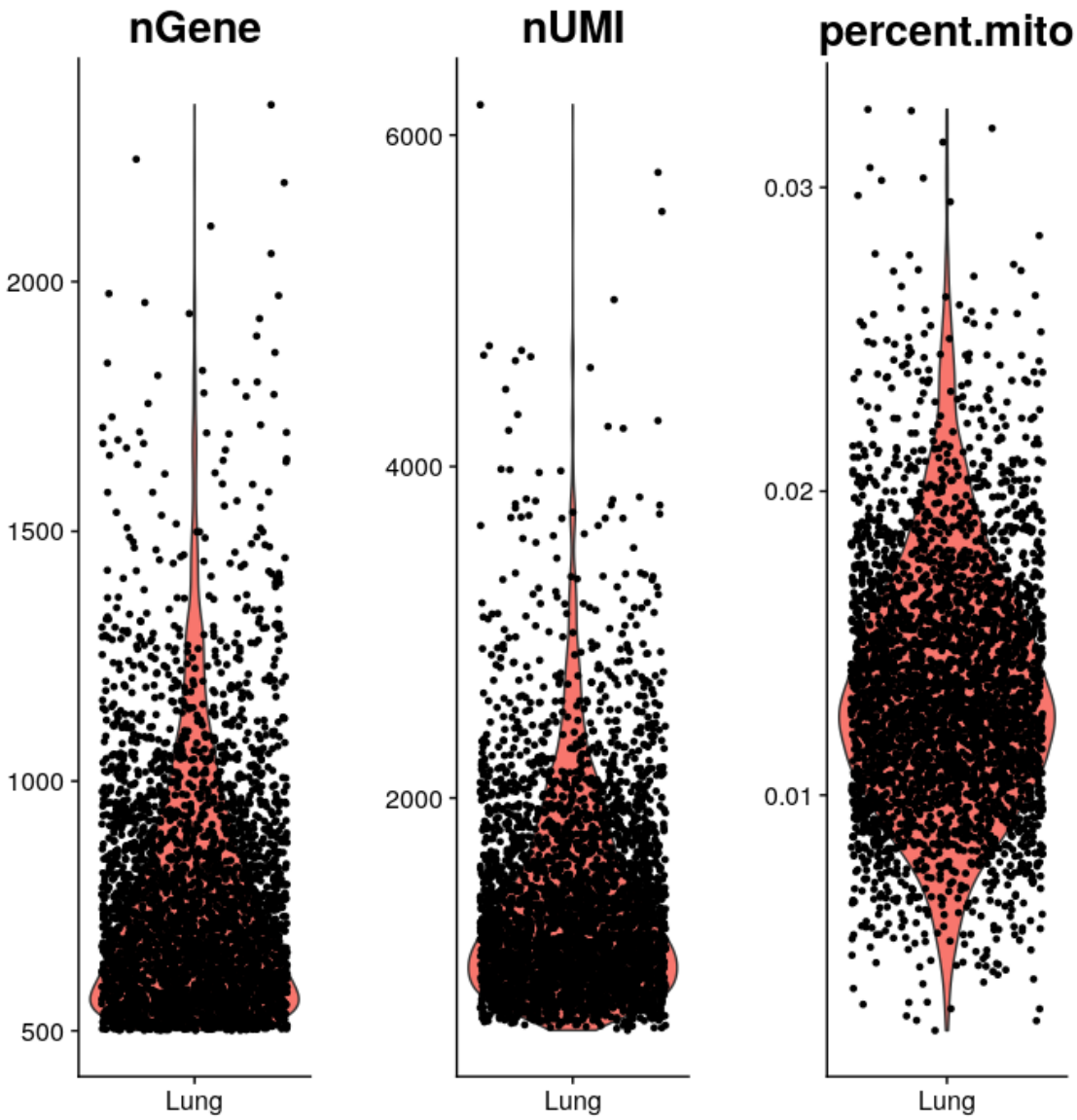
Tabula Muris Summary

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



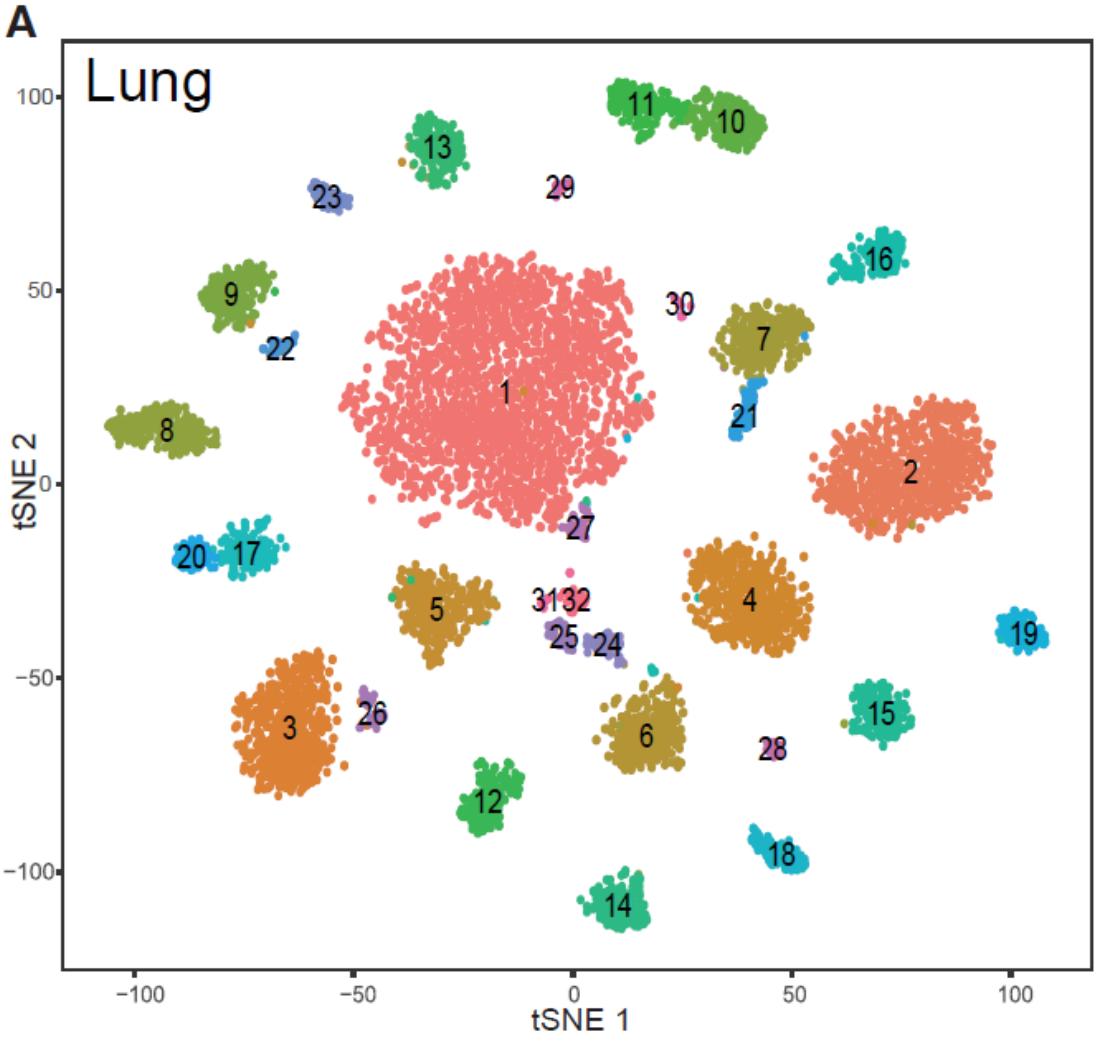
MCA Lung data (6940 cells)

Han et Al, Cell (2018)



Dropouts
96 %

MCA Lung data (6940 cells)



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

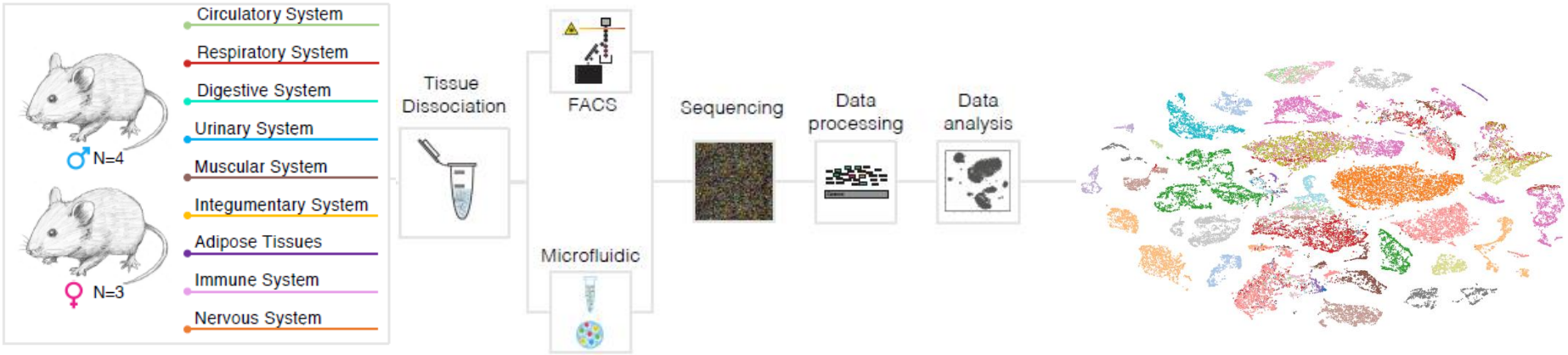
Gene expression and cell type markers available on :
http://bis.zju.edu.cn/MCA/gallery.html?tissue=Lung

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*



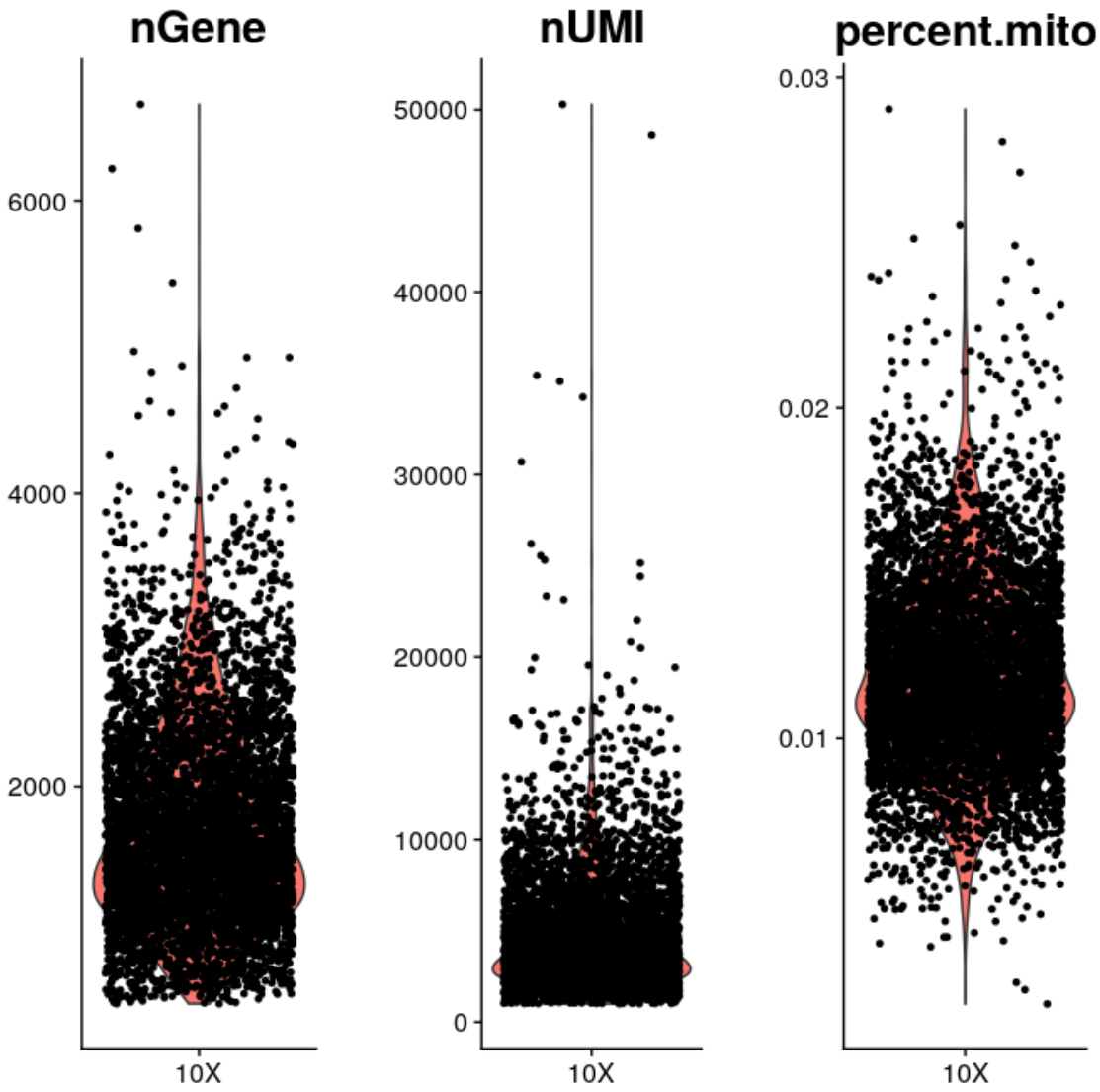
SMART-SEQ + FACS

Lung	Trachea
1620 cells	1392 cells

10X Microfluidic droplet

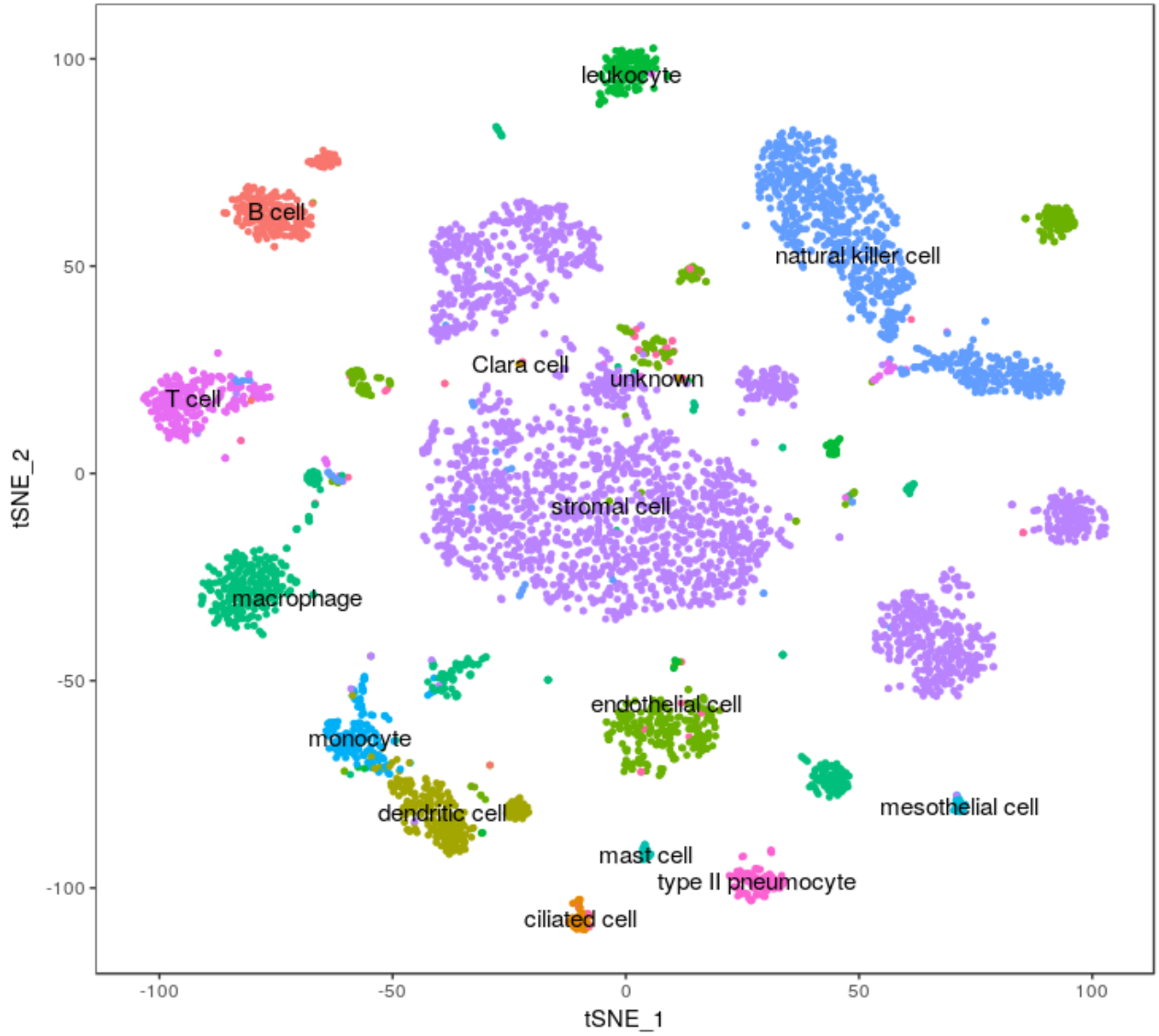
Lung	Trachea
5449 cells	11269 cells

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



Dropouts
93 %

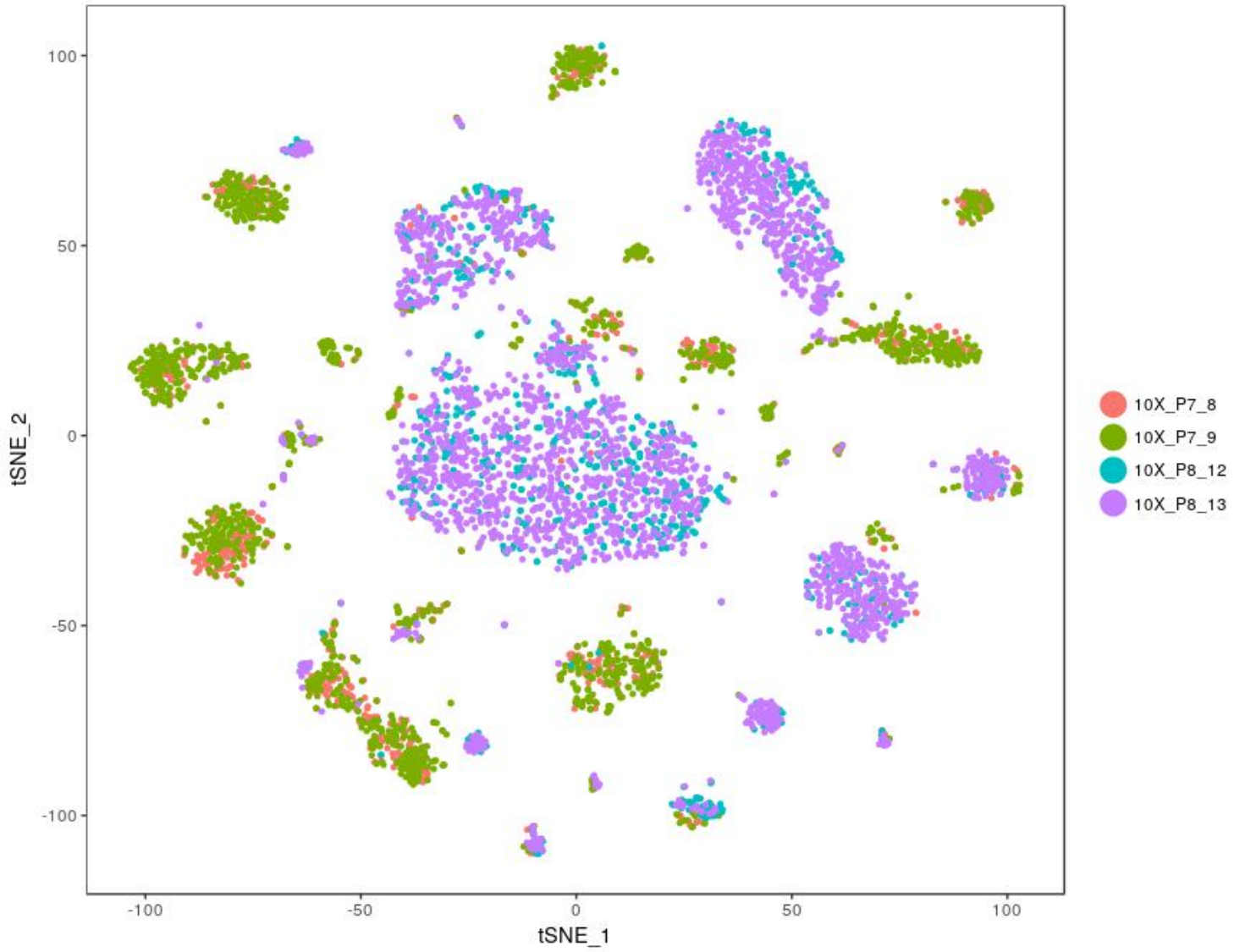
TM Lung 10X data (**5449 cells)**



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

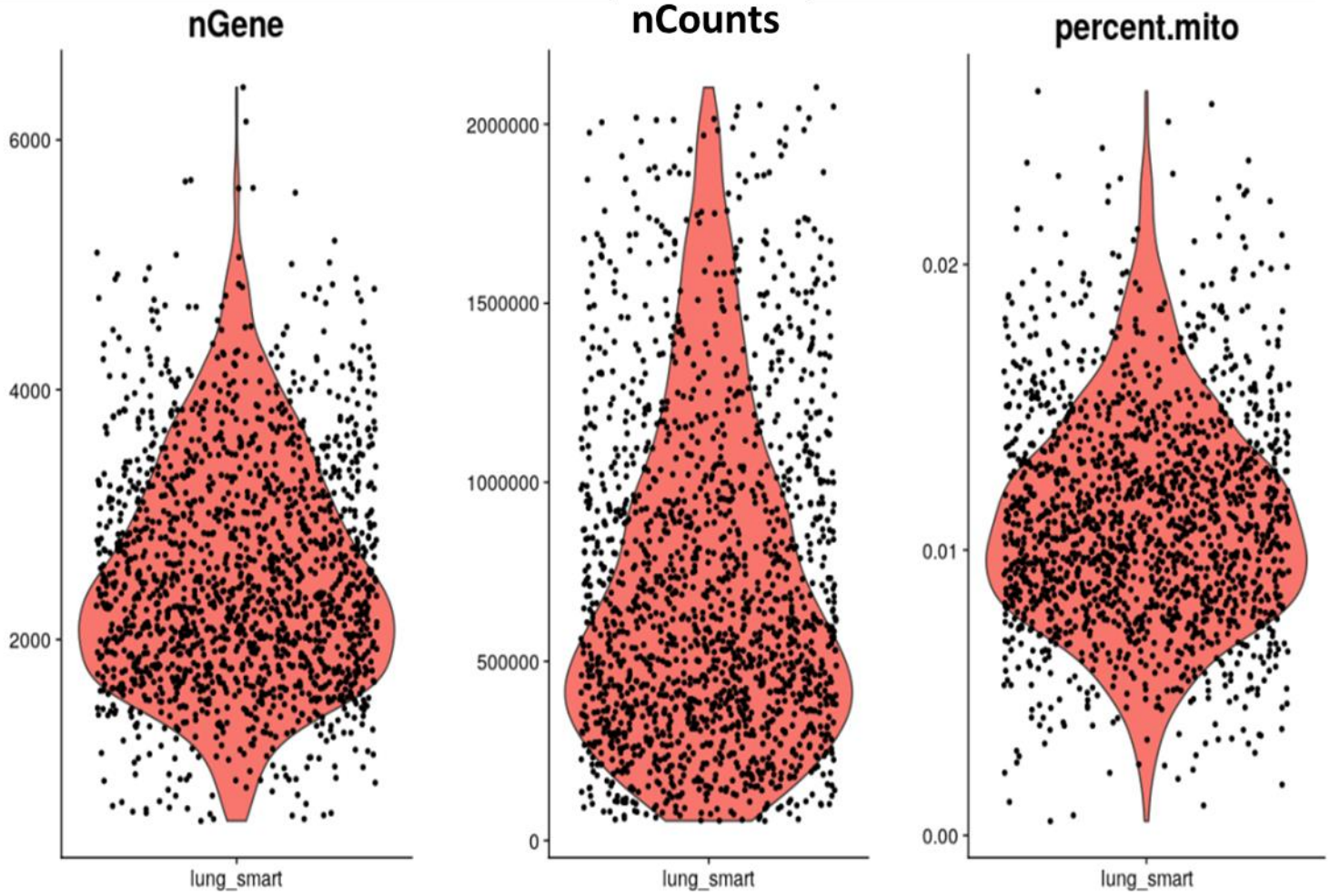
B cell	n = 205
ciliated cell	n = 41
Clara cell	n = 5
dendritic cell	n = 225
endothelial cell	n = 425
leukocyte	n = 151
macrophage	n = 456
mast cell	n = 22
mesothelial cell	n = 24
monocyte	n = 145
natural killer cell	n = 832
stromal cell	n = 2534
T cell	n = 246
type II pneumocyte	n = 89
unknown	n = 49

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

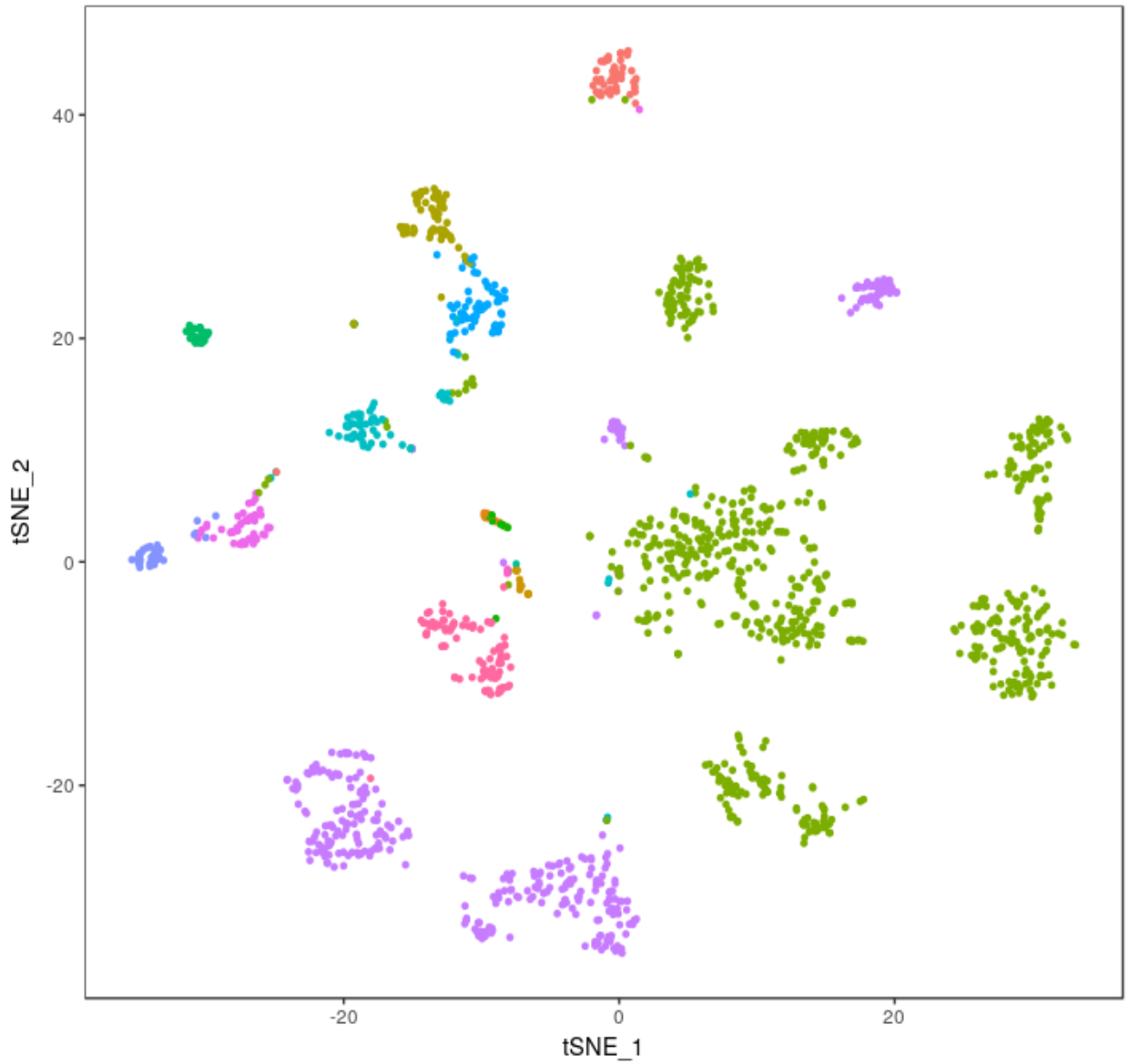


TM Lung SMART-Seq data (1620 cells)

Dropouts
89 %



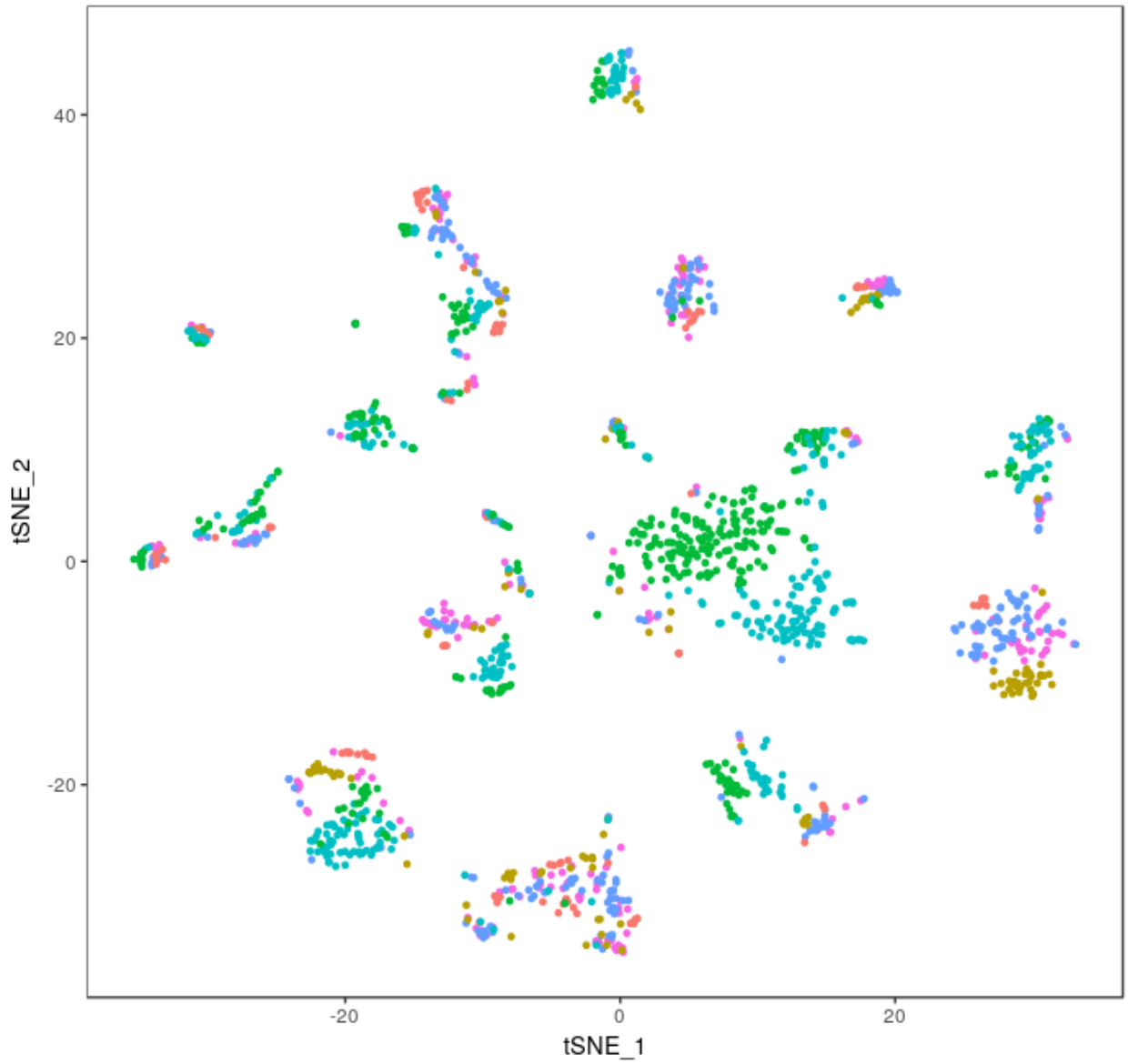
TM Lung SMART-Seq data (1620 cells)



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

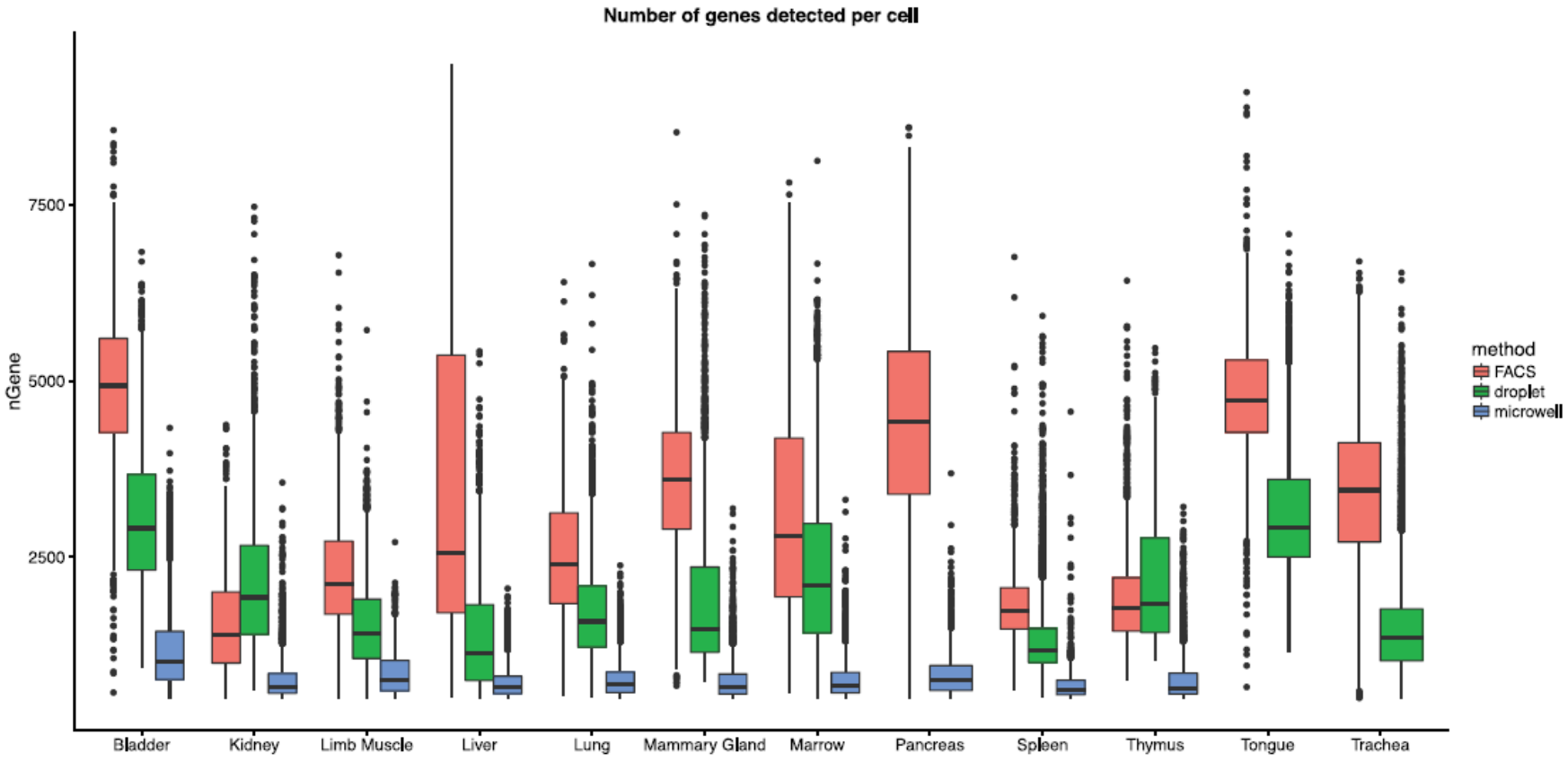
B cell
ciliated cell
Clara cell
dendritic cell
endothelial cell
epithelial cell
leukocyte
lung neuroendocrine cell
macrophage
mesothelial cell
monocyte
natural killer cell
stromal cell
T cell
type I pneumocyte
type II pneumocyte

TM Lung SMART-Seq data (1620 cells)



- **6 mice**

Mouse Atlases Sequencing depth comparison



Tabula Muris, 2018

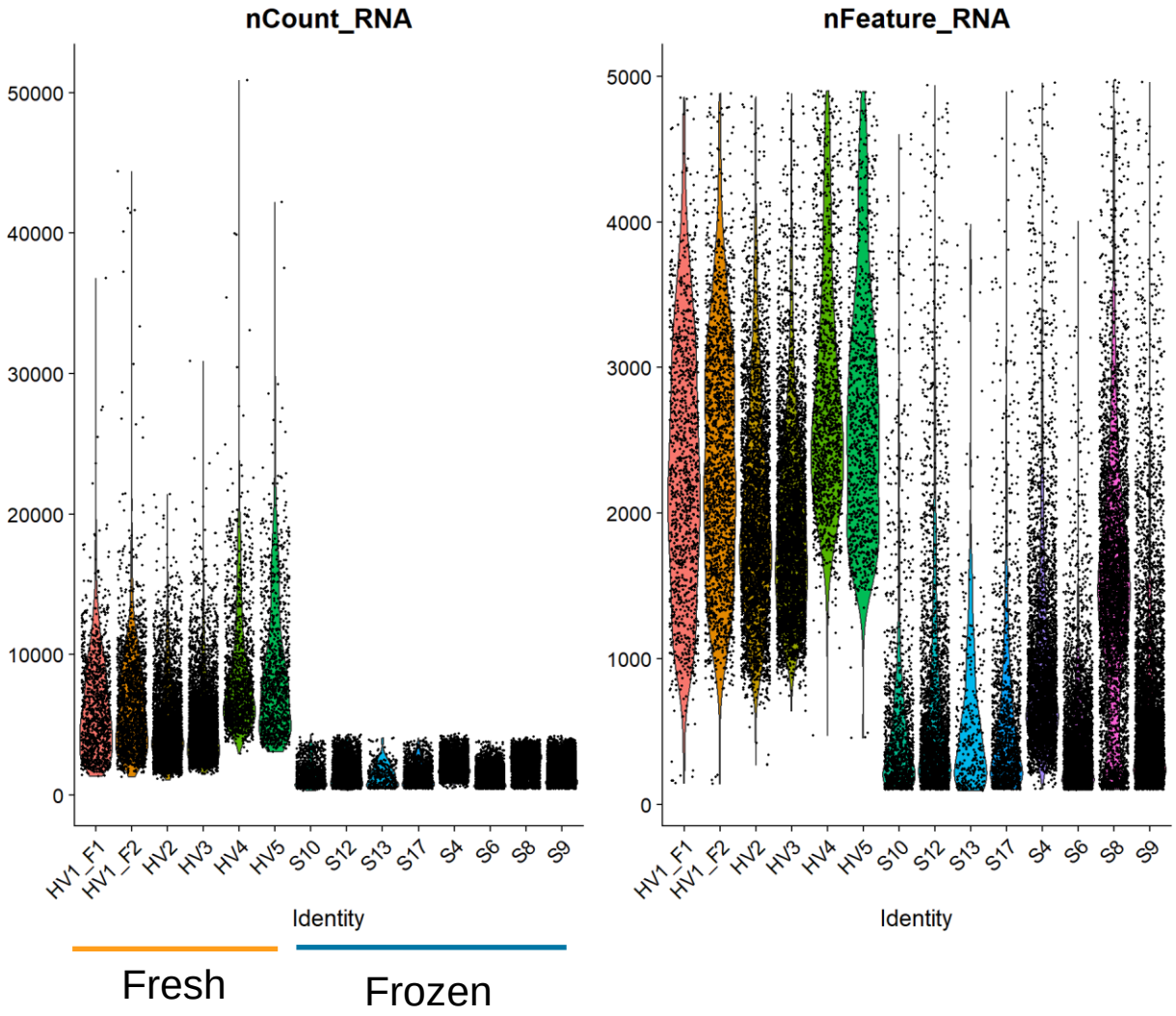
Example 2: Skin biopsies

- 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

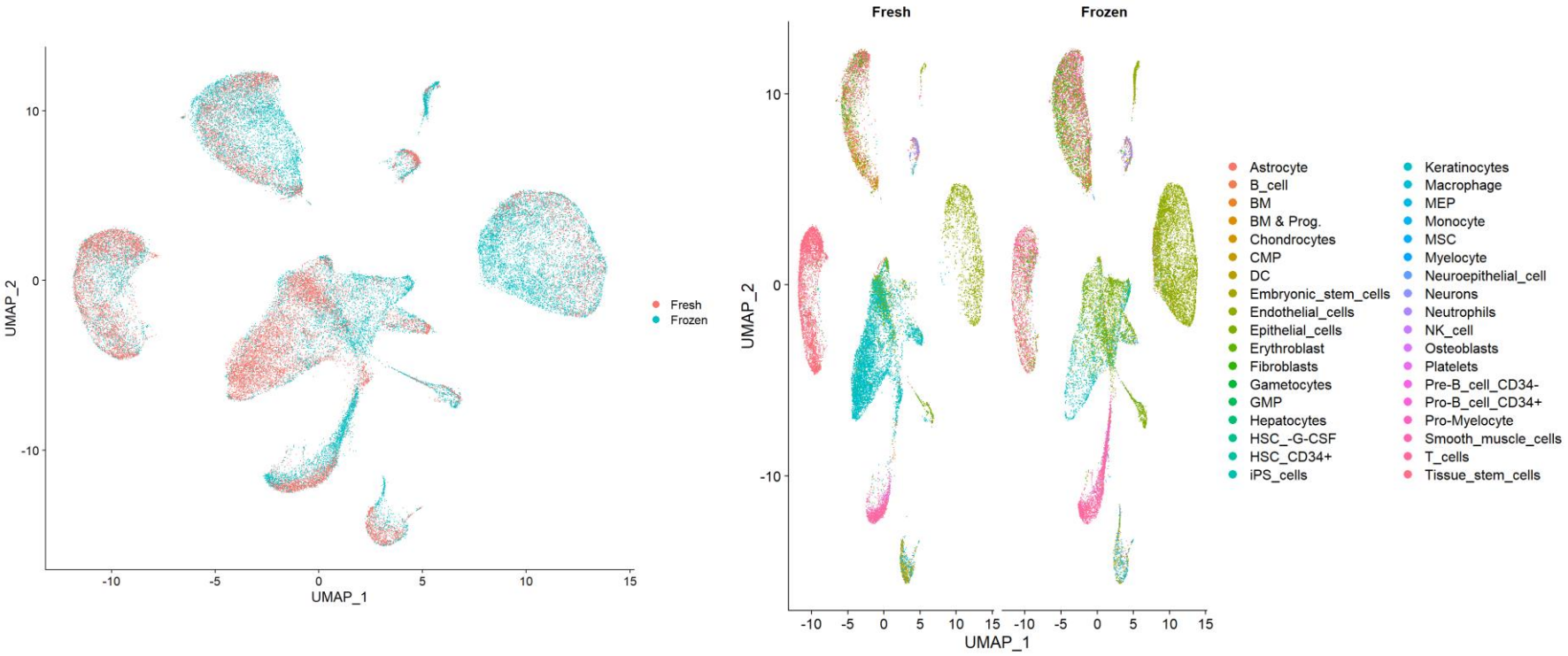
- Using GEO, we reanalized 2 studies with healthy skin tissue
  - Fresh samples: GSE132802
  - Frozen samples: GSE147424

Difference in data quality is clear

Nicolas Nottet, Syneos Health



Cell Type identification



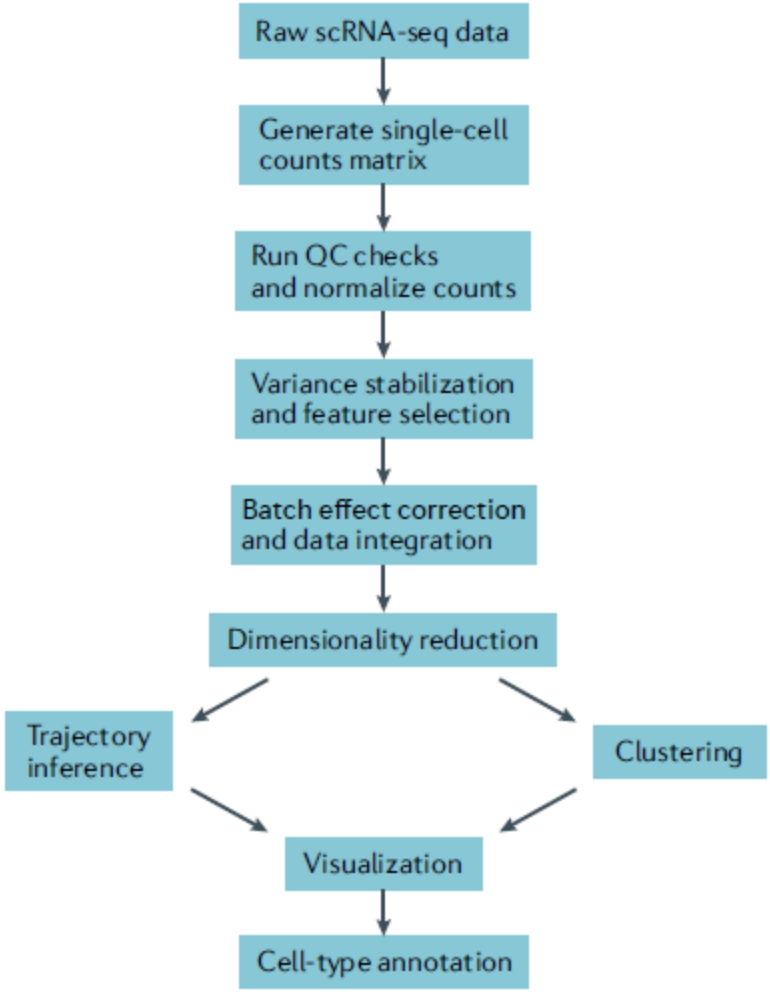
- All cell types are present in both datasets (but proportions vary)
- Differential analysis fresh vs frozen did not show a lot of DE genes

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

References

- 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

Single Cell RNAseq data analysis workflow



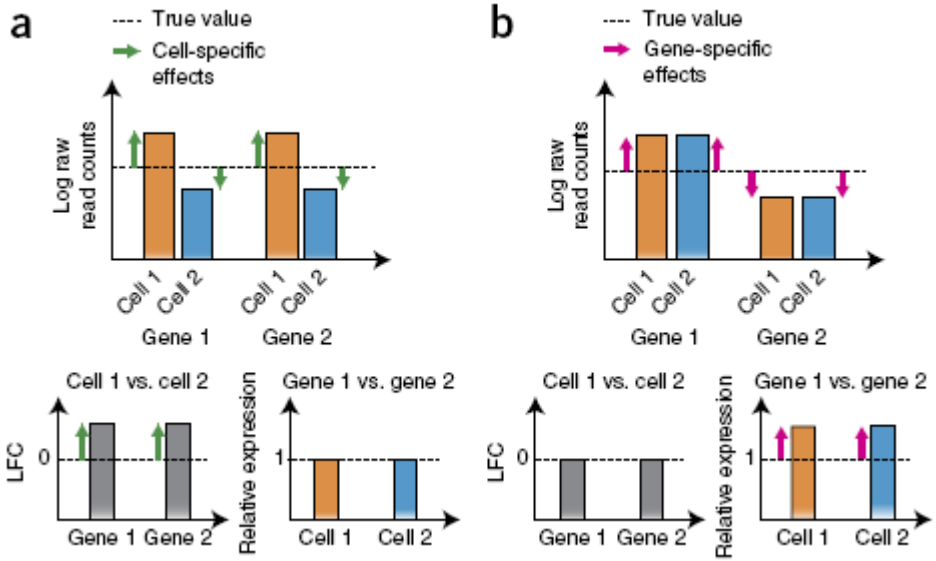
Wu, Nat RevNeph 2020

Normalization

- Process of identifying and removing systematic variation not due to real differences between RNA treatments i.e. differential gene expression.

- Cell-specific effects

- Gene-specific effects



Vallejos CA, 2017

c

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

scRNA-seq: 3 levels of normalization

- Gene-specific effects
  - within cell: GC content, gene length
- Cell specific effects
  - Aim: make count distributions comparable
- 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 SF 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

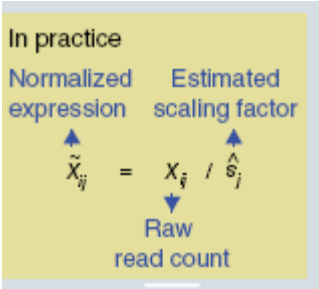
- Gene-specific effects
  - within cell: GC content, gene length
  - ***Not really accounted for in droplet assays***

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

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

Global Scaling

- Hypotheses:
  - Cell populations are homogenous
  - The RNA level is similar in all cells

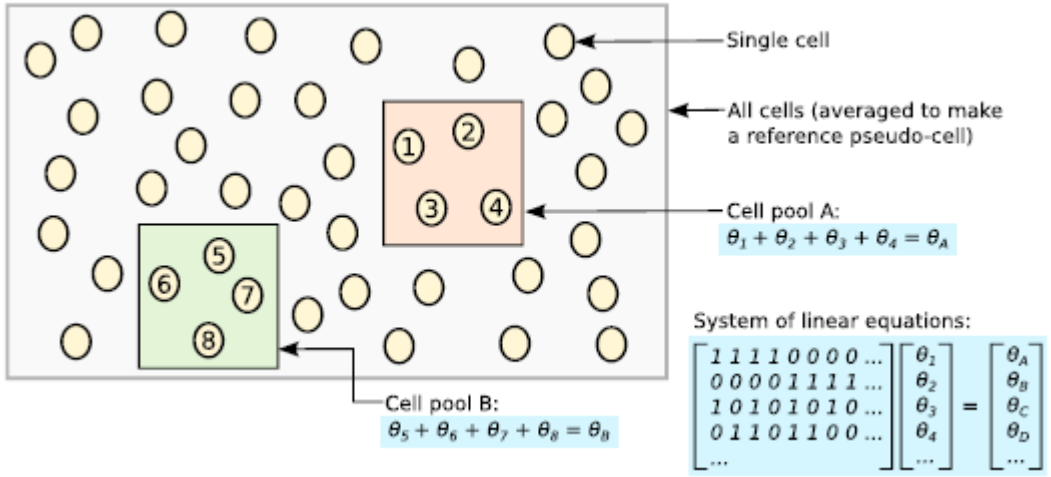


- Choice of the scaling factors
  - Median UMI counts
  - 10,000 default in Seurat / Cell Ranger

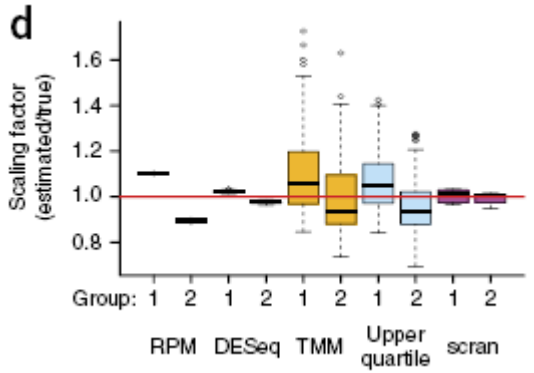
- In practice
  - Hypotheses are not always verified, but lots of people use this method anyway

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 **scater/scran** packages



Lun, 2016



Vallejos C, 2017

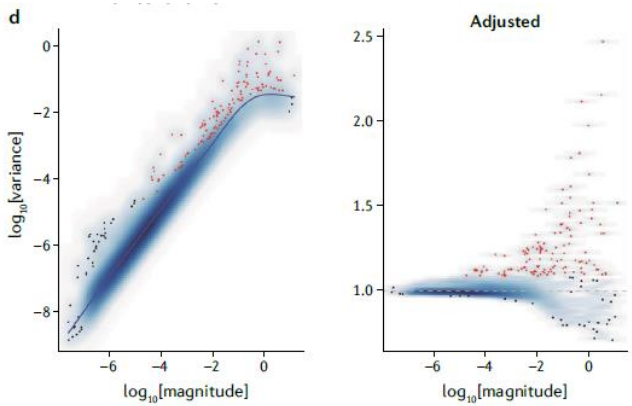
More advanced methods are available

- Normalization included in the statistical model
  - SCDE, Monocle, MAST,…

- Normalization based on spike-ins or invariant genes
  - BASICs, scNorm

- Variance stabilization
  - Correct for strong mean-variance relationship
  - Included in Seurat, Pagoda2, SCANPY

- Fancy modeling
  - Modeling of single cell count data using Neg Binomial
  - ZINB-Wave, single-cell variational infernece (scVI) etc



Wu 2020

Normalization for other biological factors

- Known or unknown variation
  - Cell cycle, number of genes detected, % mitochondrial genes…
- Regression methods provided to account for know factors
  - 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

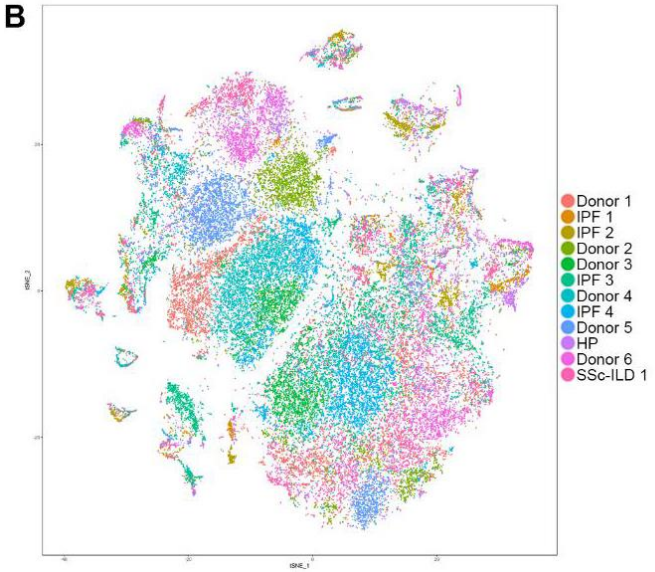
- Sample/Technology-specific effects -> Data Integration
  - Batch effects (BAD)
  - Between samples variability (GOOD)

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

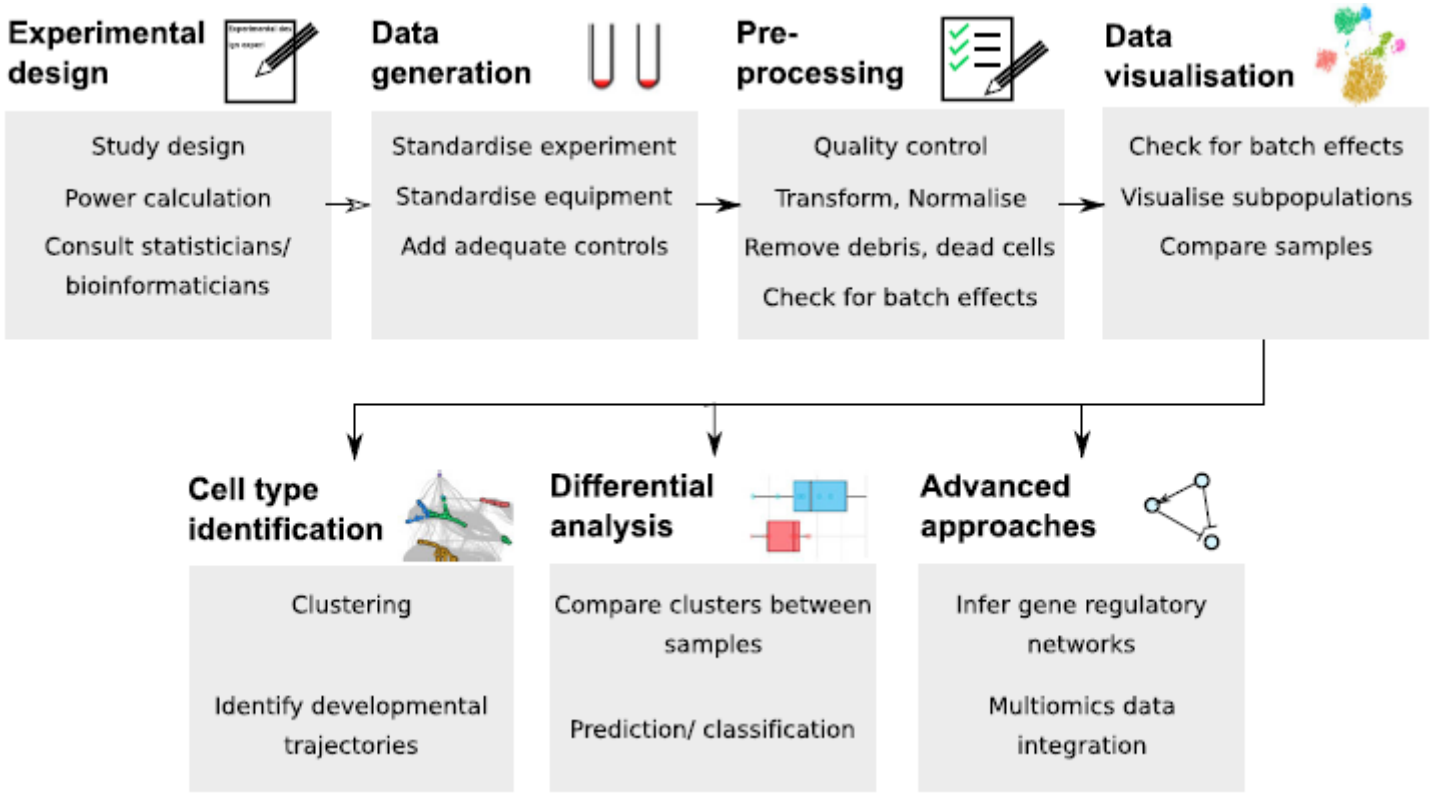
- Combining datasets and applying cell-level normalization might not be enough to remove this bias

- More details in next session



Misharin, BiorXiv 2018

Conclusion



Todorov, 2018

References

- 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/
- https://bioconductor.org/help/course-materials/2017/BioC2017/Day2/InvitedSpeakers/Biscuit_Azizi.pdf
- Haghverdi, L., Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat. Biotechnol. 2018.
- kBET: Maren Büttner et al, A test metric for assessing single-cell RNA-seq batch correction, Nat Methods 2019
- Yan Wu and Kun Zhang, Tools for the analysis of high- dimensional single- cell RNA sequencing data, Nat Rev Neph 2020
- Cakir B**,** Comparison of visualization tools for single-cell RNAseq data, NAR Genomics and Bioinformatics, 2020

Thank you