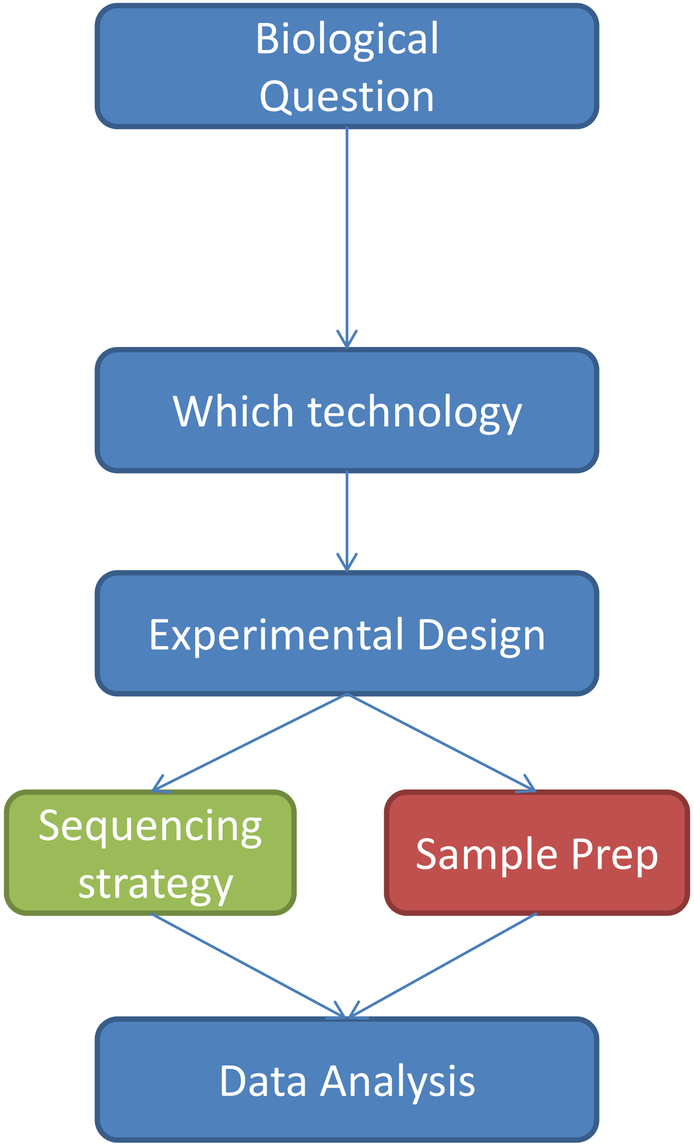
Experimental Design
Quality Control, Normalization

Agnes Paquet

SincellTE, Roscoff 05/02/2019

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



Technology Overview

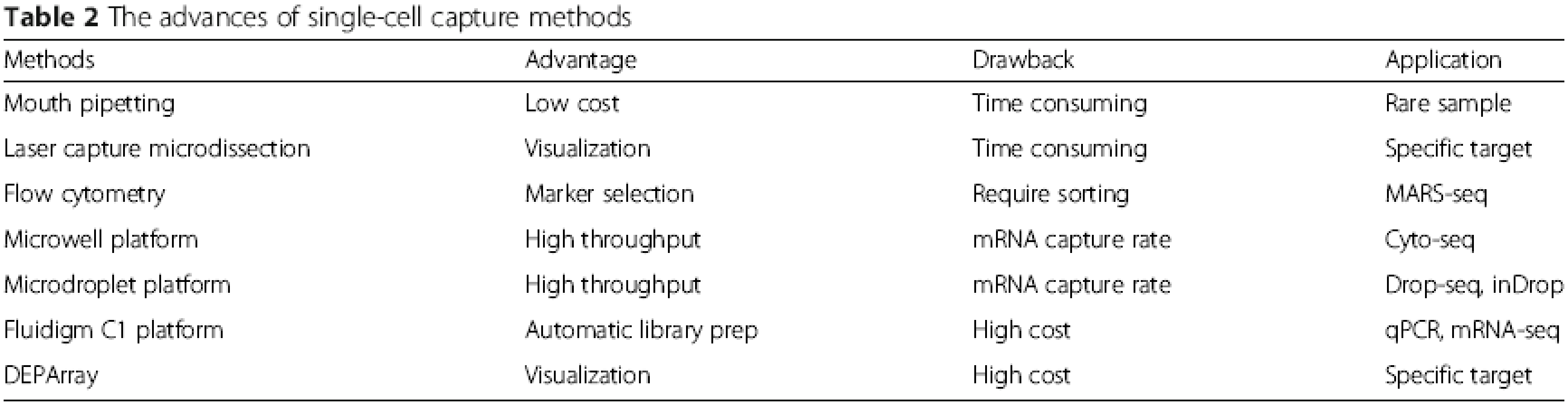


Table 2 The advances of single-cell capture methods

Methods	Advantage	Drawback	Application
Mouth pipetting	Low cost	Time consuming	Rare sample
Laser capture microdissection	Visualization	Time consuming	Specific target
Flow cytometry	Marker selection	Require sorting	MARS-seq
Microwell platform	High throughput	mRNA capture rate	Cyto-seq
Microdroplet platform	High throughput	mRNA capture rate	Drop-seq, inDrop
Fluidigm C1 platform	Automatic library prep	High cost	qPCR, mRNA-seq
DEPArray	Visualization	High cost	Specific target

Ye et al, Journal of Hema and Onco 2017

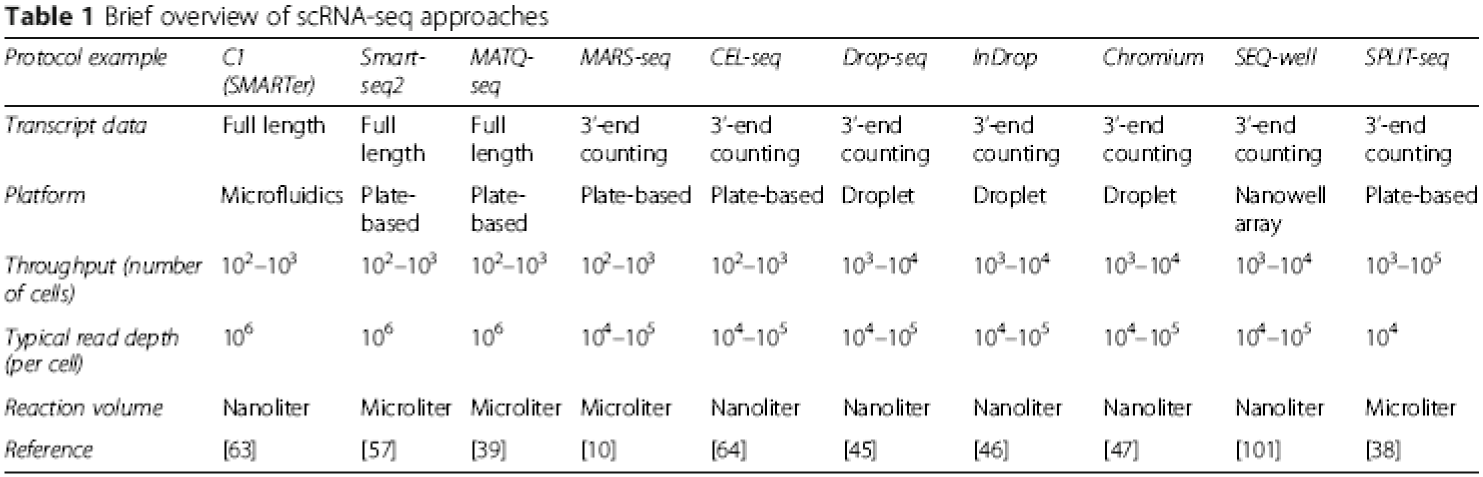


Table 1 Brief overview of scRNA-seq approaches

Protocol example	*C1 (SMARTer)*	*Smart-seq2*	*MATQ-seq*	*MARS-seq*	*CEL-seq*	*Drop-seq*	*InDrop*	*Chromium*	*SEQ-well*	*SPLIT-seq*
Transcript data	Full length	Full length	Full length	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting	3′-end counting
Platform	Microfluidics	Plate-based	Plate-based	Plate-based	Plate-based	Droplet	Droplet	Droplet	Nanowell array	Plate-based
Throughput (number of cells)	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^2–10^3	10^3–10^4	10^3–10^4	10^3–10^4	10^3–10^4	10^3–10^5
Typical read depth (per cell)	10^6	10^6	10^6	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4–10^5	10^4
Reaction volume	Nanoliter	Microliter	Microliter	Microliter	Nanoliter	Nanoliter	Nanoliter	Nanoliter	Nanoliter	Microliter
Reference	[63]	[57]	[39]	[10]	[64]	[45]	[46]	[47]	[101]	[38]

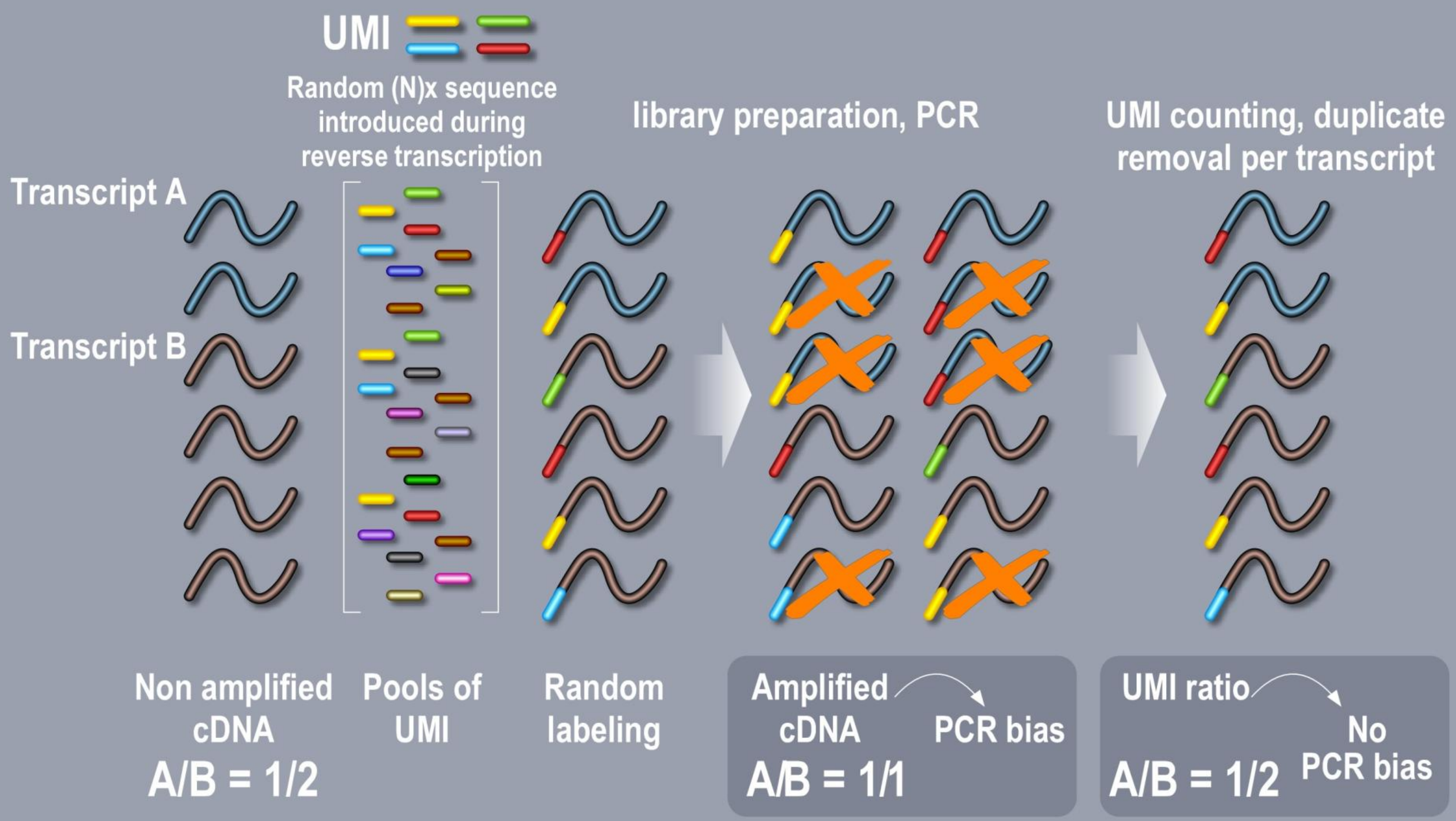
Haque et al, Genome Medicine 2017

- ➢Throughput: 10s vs. 1000s of cells?
- ➢ Full length protocols: required for splicing, inferring CNV, TCR/BCR profiling

Experimental design: technical point of view

- UMI design
- Use of Spike-ins
- Discuss about sequencing design
  - Number of cells
  - Sequencing depth

Unique Molecular Identifier (*Islam et al., Nature Methods, 2014*)

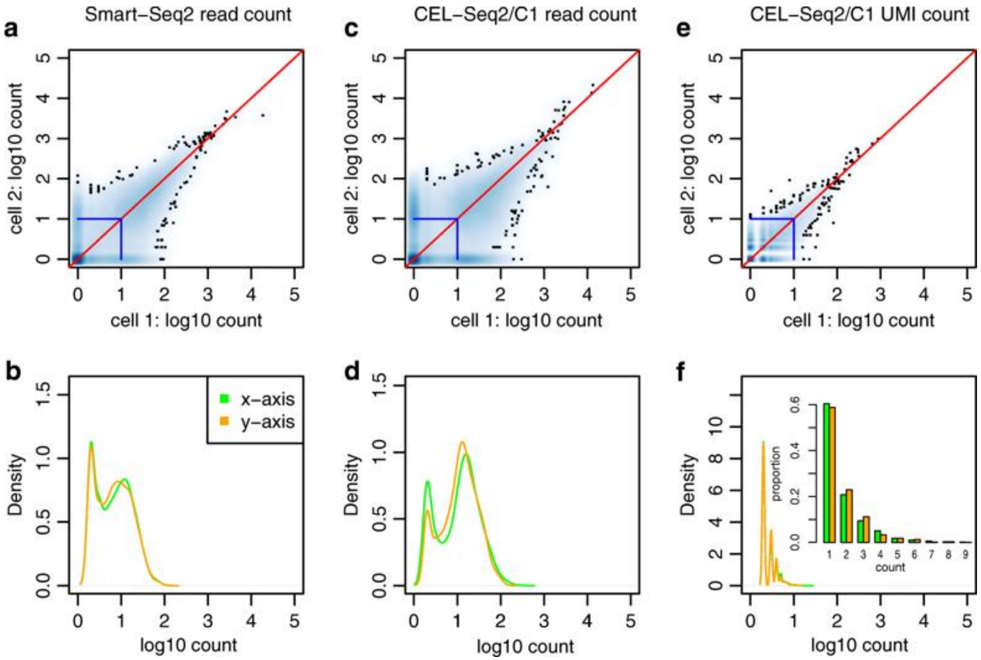


UMI allow a more precise profiling

UMIs : *Kivioja, T. et al. Counting absolute numbers of molecules using unique molecular identifiers. Nat Meth 9, 72-74 (2012)*
UMIs for single cell transcriptome: *Islam, S. et al. Quantitative single-cell RNA-seq with UMI . Nat Methods 11, (2014).*

UMI design and bias correction

- UMI-based protocols allow for PCR bias correction
- Improved accuracy of gene expression measures



Chen, Genome Bio 2018

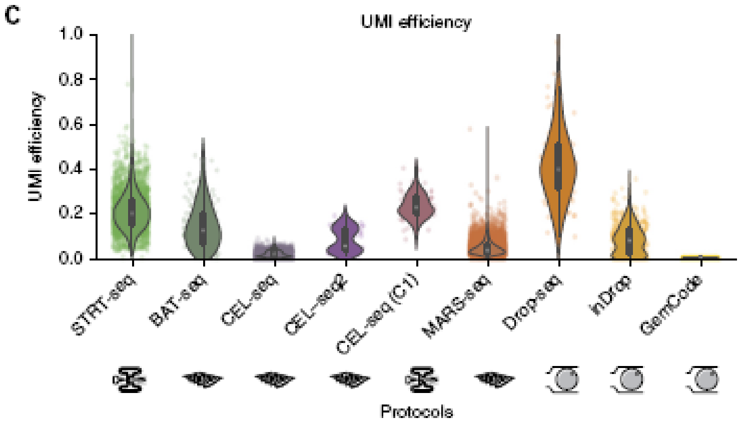
UMI Design Bias

- Design limits
  - N=4-10bp barcodes -> 4^N possible UMIs

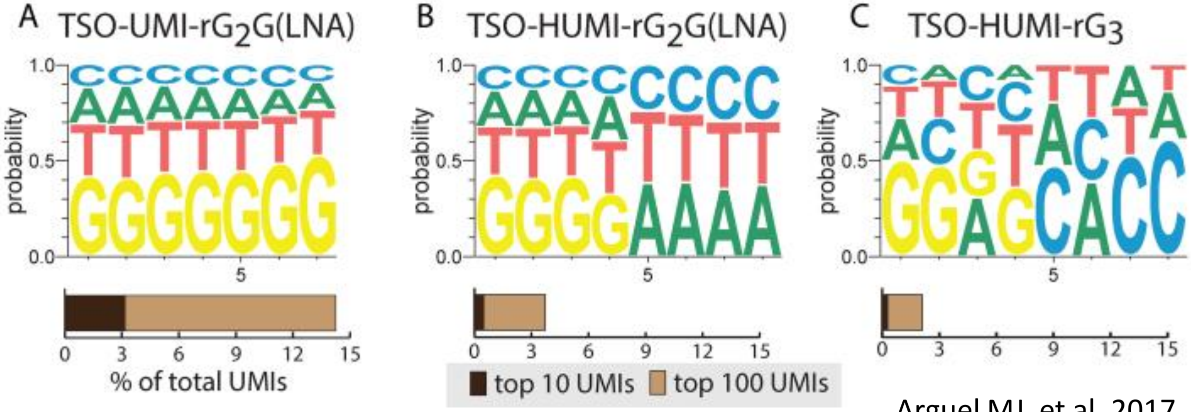
    N=5 -> 1024 UMIs available

    N=10 -> 1,048,576 UMIs available

  - GC content bias

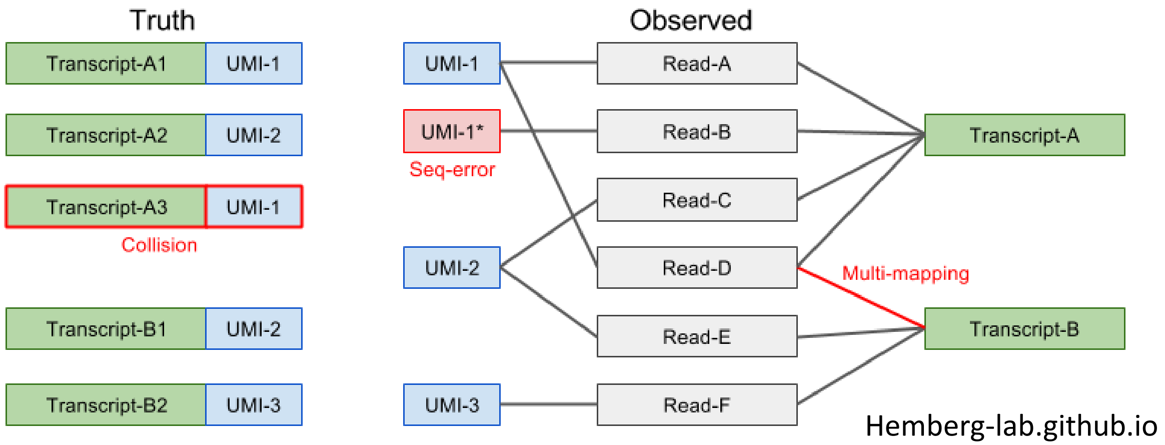


Svensson V, et al 2017



Arguel MJ et al, 2017

Handling of sequencing errors



- Same UMI does not necessarily mean same molecule
  - Biases in UMI frequency and short UMIs
- Correction for UMI saturation:
  - e.g. Grün, 2014

- Different UMI does not necessarily mean different molecule
  - Sequencing errors
- Different transcript does not necessarily mean different molecule
  - Mapping errors/multi-mapping

➢ Error correction using edit distance (ed=1 standard for 8-10bp UMI)

Ref: UMI-tools, Smith T, Genome Res 2017

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

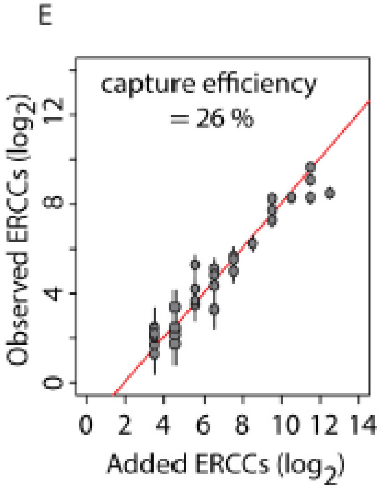
Spike-ins use in scRNA-seq

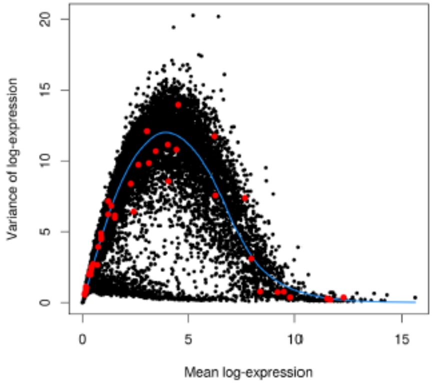
- Estimate protocol capture efficiency
  - How many of the spiked molecules did we detect?
- Comparison of protocols performance
  - Level of detection in low expressed genes
  - See Svensson V. et al, 2017

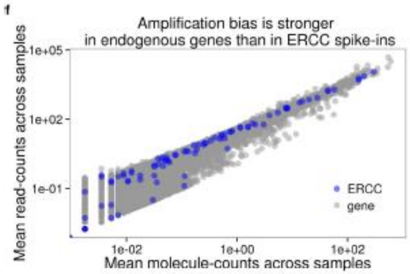
- Estimate technological noise
  - Help for detection of highly variable genes

- 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

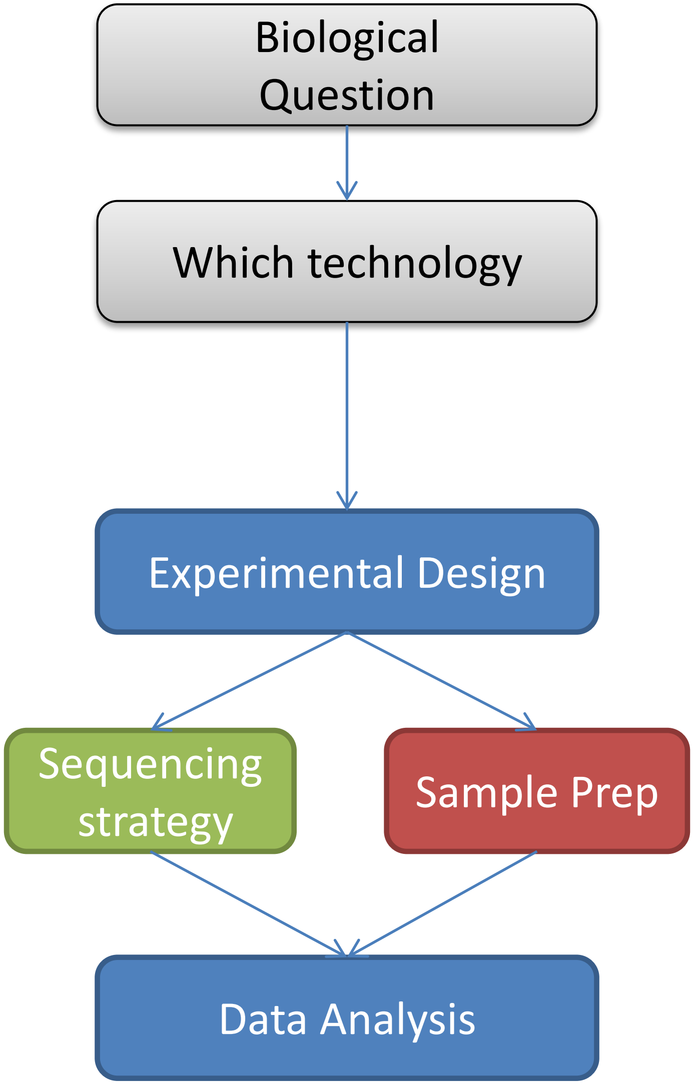
- We have a question
- We have selected a protocol

- How many samples?

- How many cells?

- How many reads/cell?

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

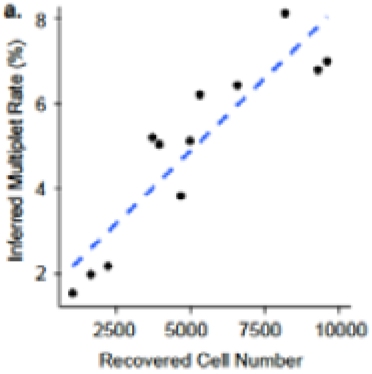


Estimating the required number of cells / sequencing depth

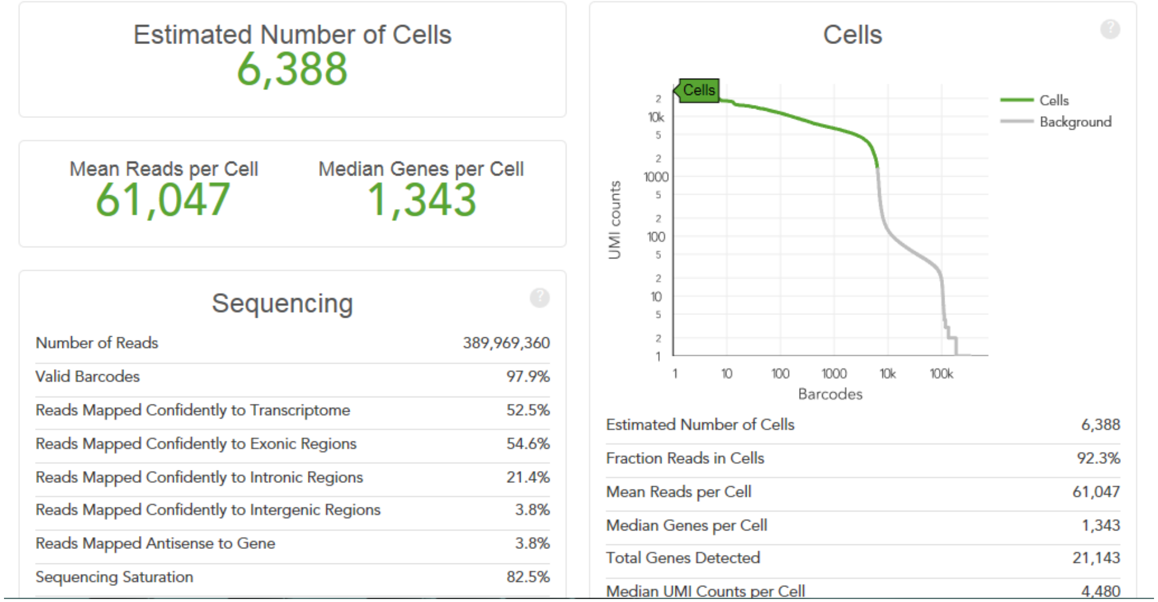
- Number of cells required
  - Do we have a lot of cells to begin with?
  - Are we looking for rare cells (probability estimation)?
- WARNING: doublet rate increases with higher cell numbers in droplet assays.

- 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

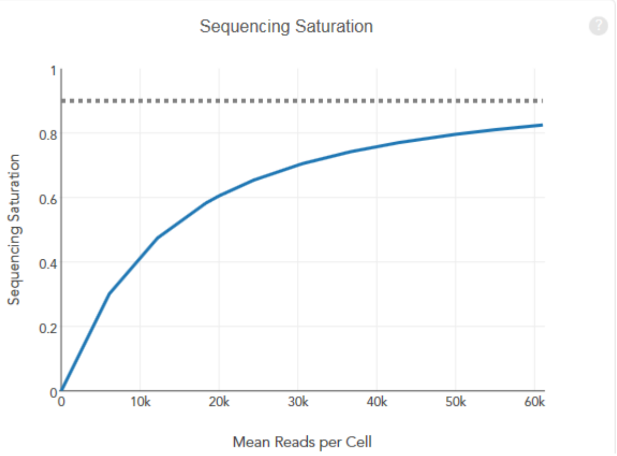
Median Genes per Cell
1,343

Sequencing

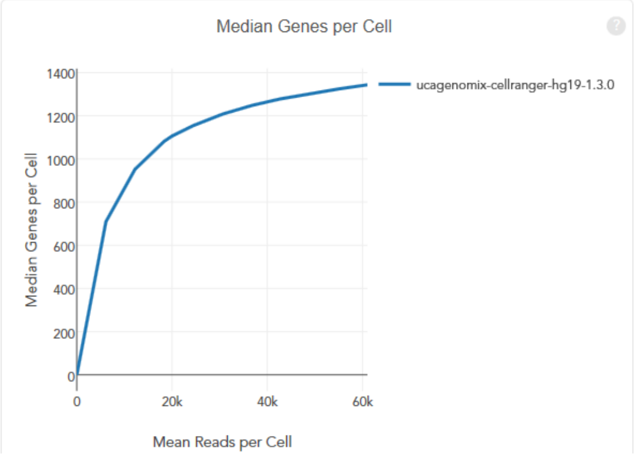
Number of Reads	389,969,360
Valid Barcodes	97.9%
Reads Mapped Confidently to Transcriptome	52.5%
Reads Mapped Confidently to Exonic Regions	54.6%
Reads Mapped Confidently to Intronic Regions	21.4%
Reads Mapped Confidently to Intergenic Regions	3.8%
Reads Mapped Antisense to Gene	3.8%
Sequencing Saturation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



Sequencing Saturation

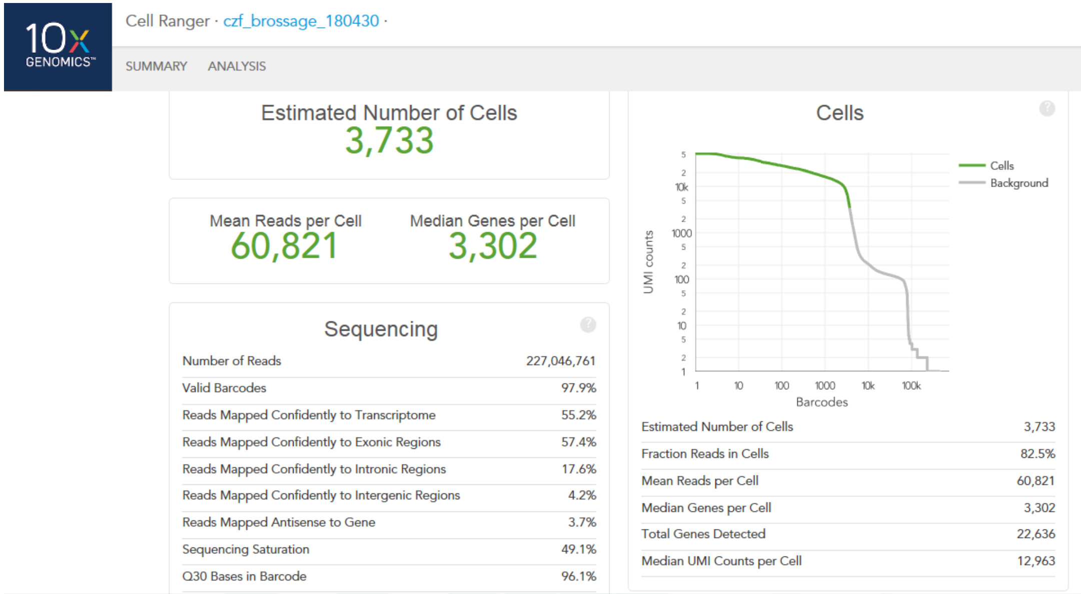


Median Genes per Cell

Example 2: Nasal epithelium brushing (cells with lots of RNA)

-Target: 5,000 cells

- 2 samples, NextSeq High 75 (~400millions reads / run)



Estimated Number of Cells
3,733

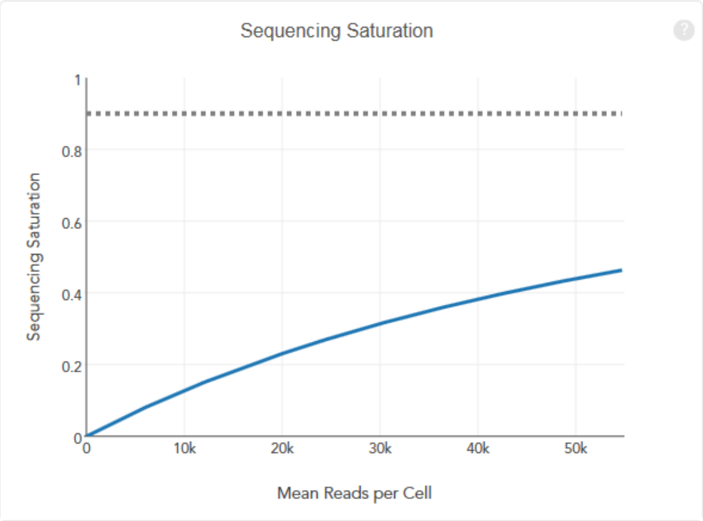
Mean Reads per Cell	Median Genes per Cell
60,821	3,302

Sequencing

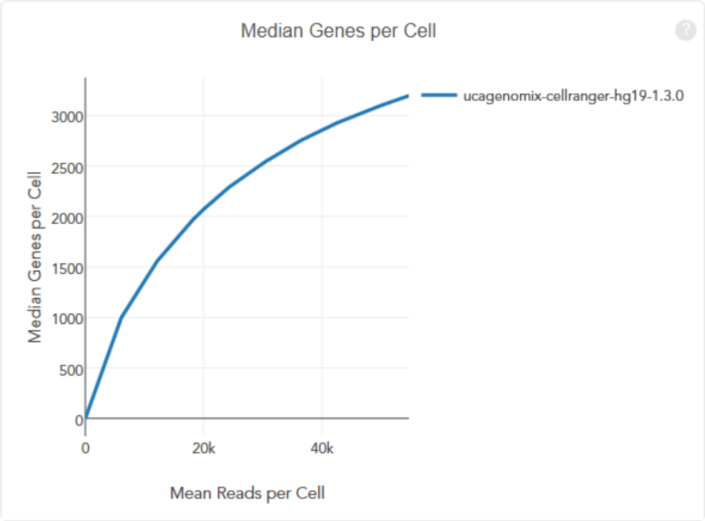
Number of Reads	227,046,761
Valid Barcodes	97.9%
Reads Mapped Confidently to Transcriptome	55.2%
Reads Mapped Confidently to Exonic Regions	57.4%
Reads Mapped Confidently to Intronic Regions	17.6%
Reads Mapped Confidently to Intergenic Regions	4.2%
Reads Mapped Antisense to Gene	3.7%
Sequencing Saturation	49.1%
Q30 Bases in Barcode	96.1%

Estimated Number of Cells	3,733
Fraction Reads in Cells	82.5%
Mean Reads per Cell	60,821
Median Genes per Cell	3,302
Total Genes Detected	22,636
Median UMI Counts per Cell	12,963

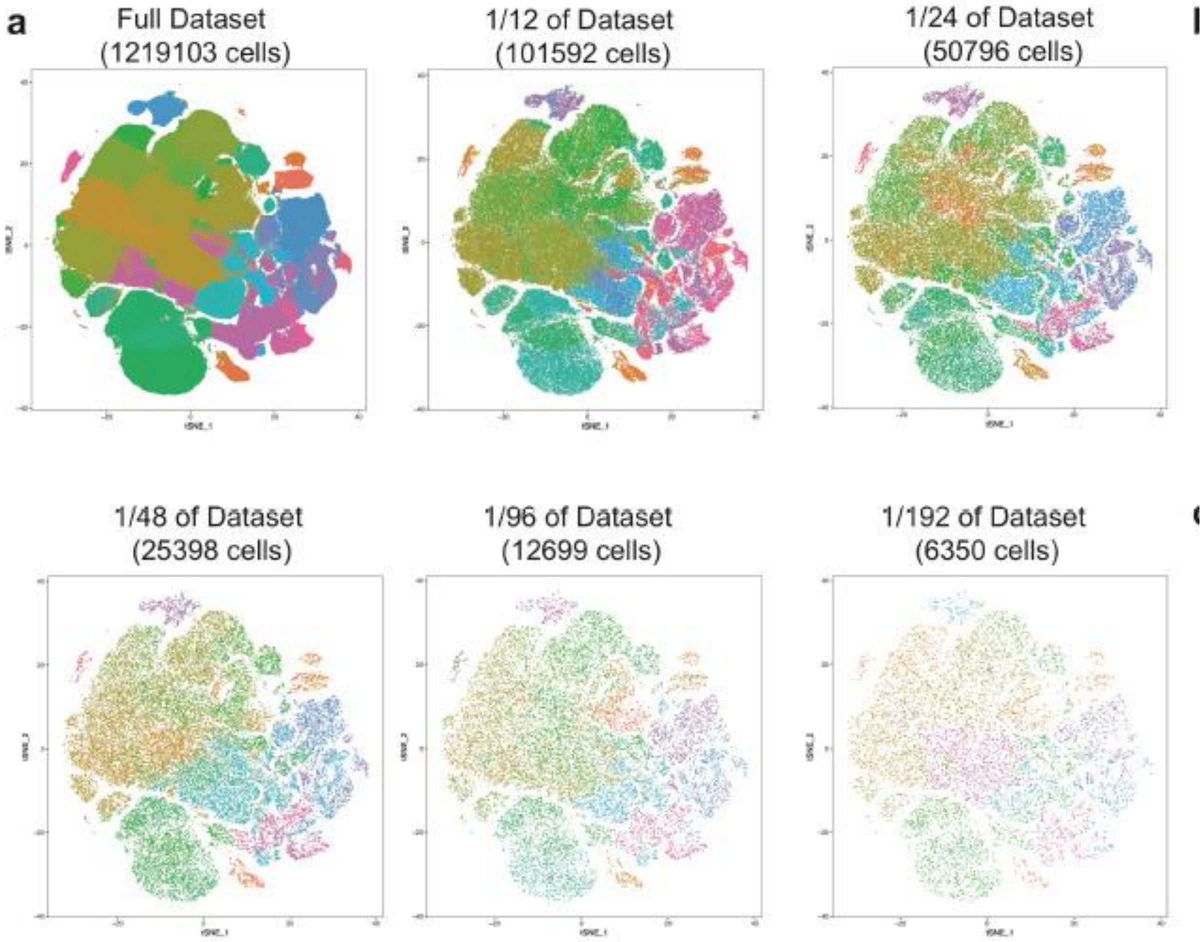
Sequencing Saturation



Median Genes per Cell



Number of cells: example of the 1.3millions cells dataset



Full Dataset (1219103 cells)

1/12 of Dataset (101592 cells)

1/24 of Dataset (50796 cells)

1/48 of Dataset (25398 cells)

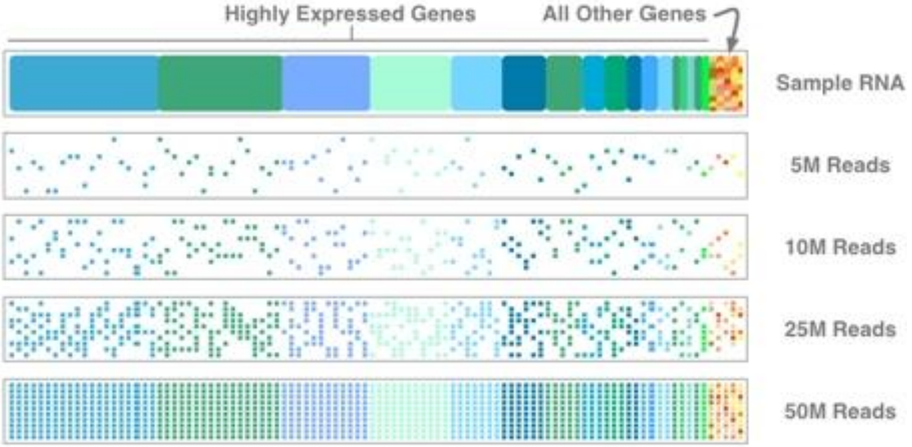
1/96 of Dataset (12699 cells)

1/192 of Dataset (6350 cells)

Bhaduri A, BiorXiv 2017

Summary of technical design

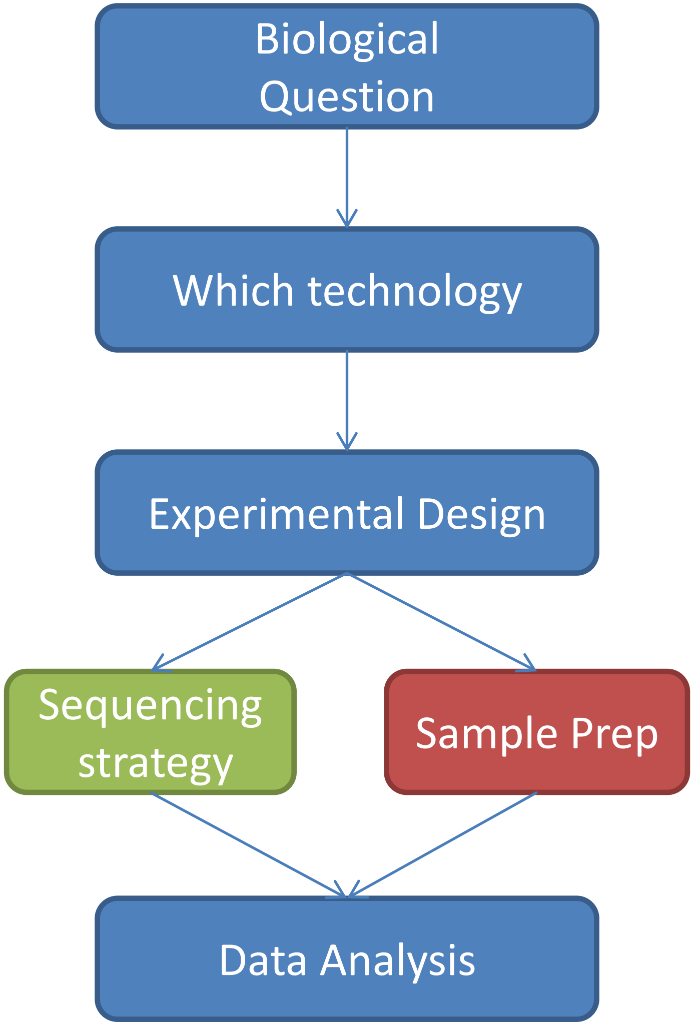
- Discuss about sequencing depth with the biologist
- If the sequencing is too shallow, the statistical analysis may not be robust
  - Worst case scenario: you can't even find the biologist favorite gene
- More cells is not always better
- Sequencing depth should be the same for all samples



Single RNAseq workflow: bioinformatics point of view

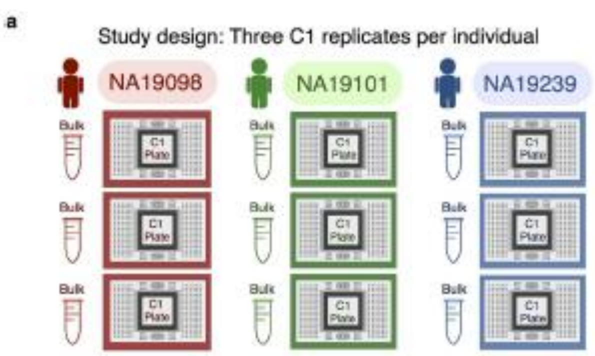
- What technique should we use to generate the data ?
  - Plate based / droplets
  - Full length / 3’ counting with UMI
  - ➢ UNDERSTAND THE BIAS

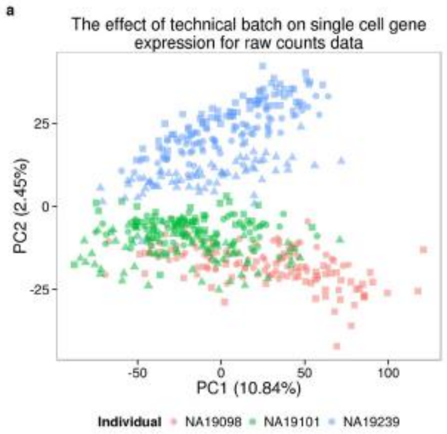
- Experimental design
  - Sequencing strategy
    - UMI design
    - Spike-ins
    - How to sequence
  - **Samples: Practical considerations**
    - Types /number of samples
    - Cell preparation -> *confounding*
    - Budget



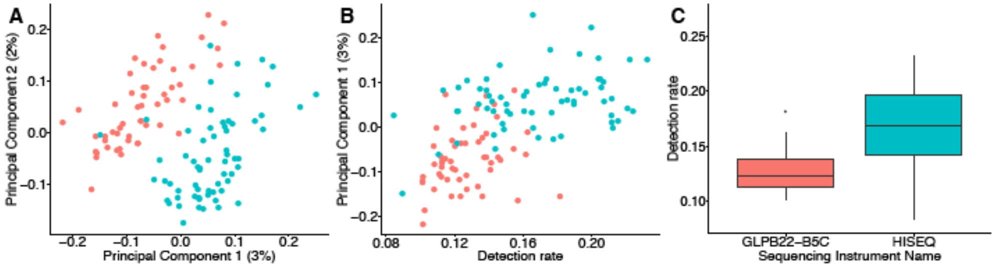
What about experimental confounding factors ?

- Most scRNA-seq are 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





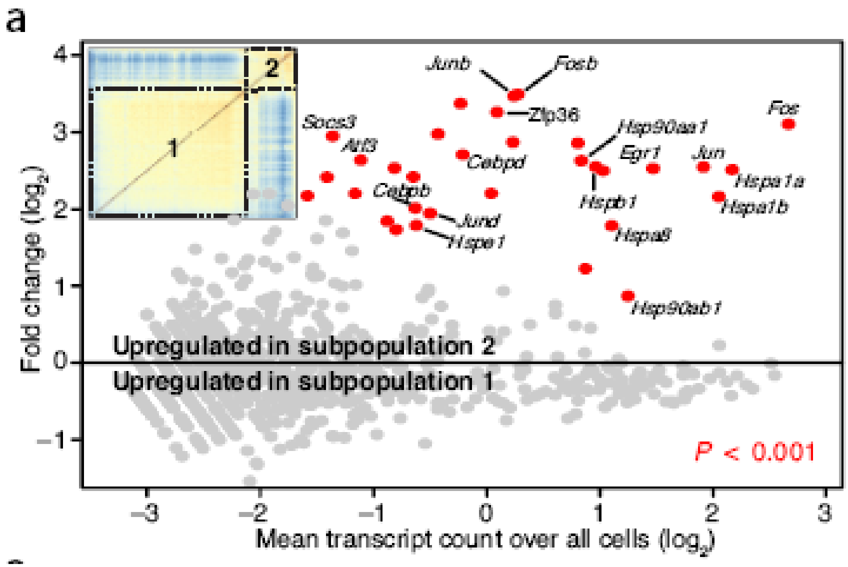
Tung, 2017



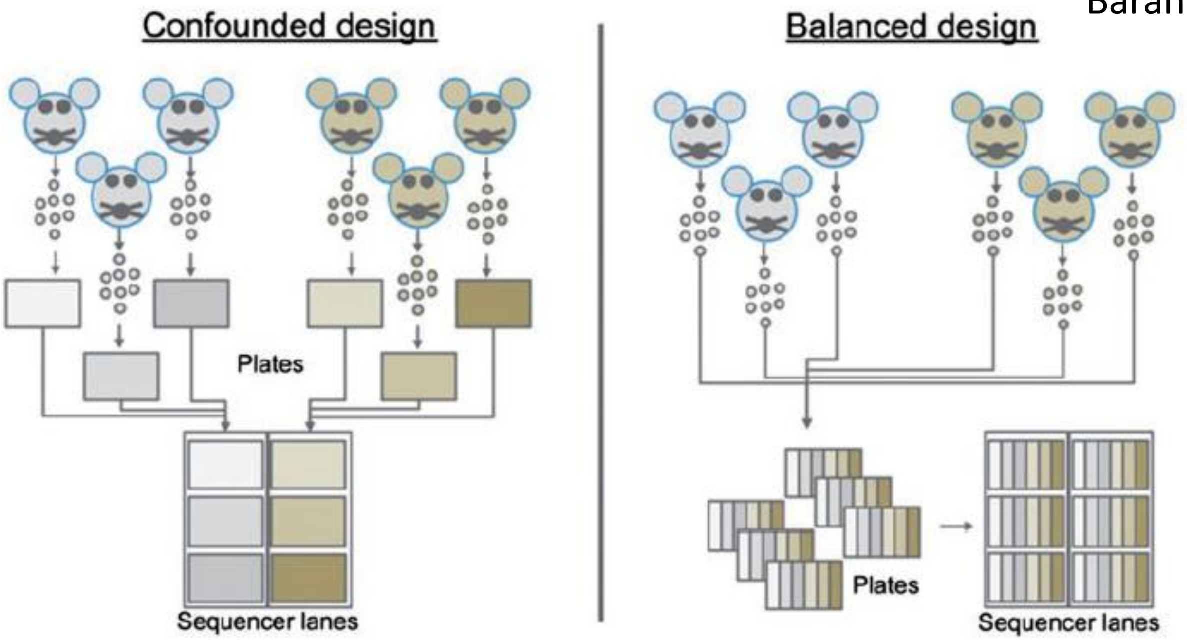
Hicks , 2017

Dissociation induced genes

- Van den Brick S, Nat Method 2017



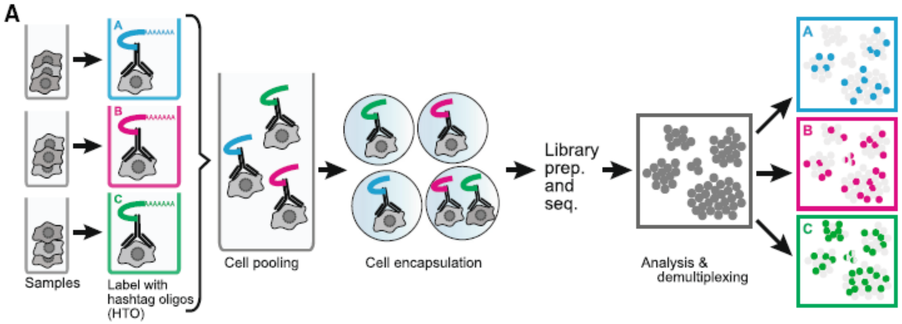
Perfect study design



Baran-Gale 2018

- Balanced design will be hard to achieve for practical reasons

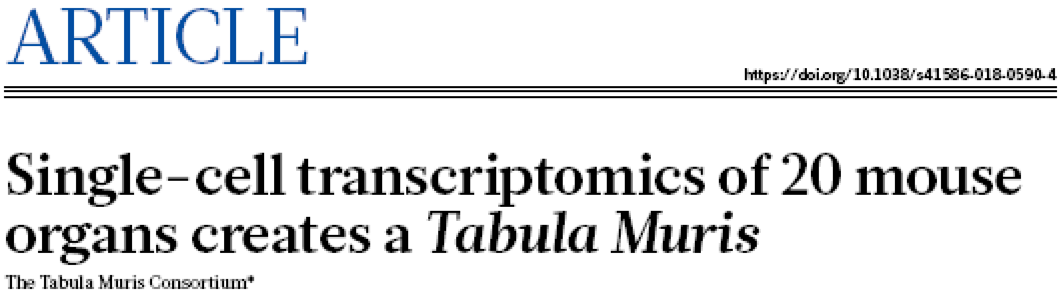
- Multiplexing :
  - Natural SNPs (demuxlet)
  - Expression of Xist/ChrY
  - **Cell -hashing**



Stoeckius, 2018

Example: Mouse Cell Atlases

Marin Truchi, IPMC

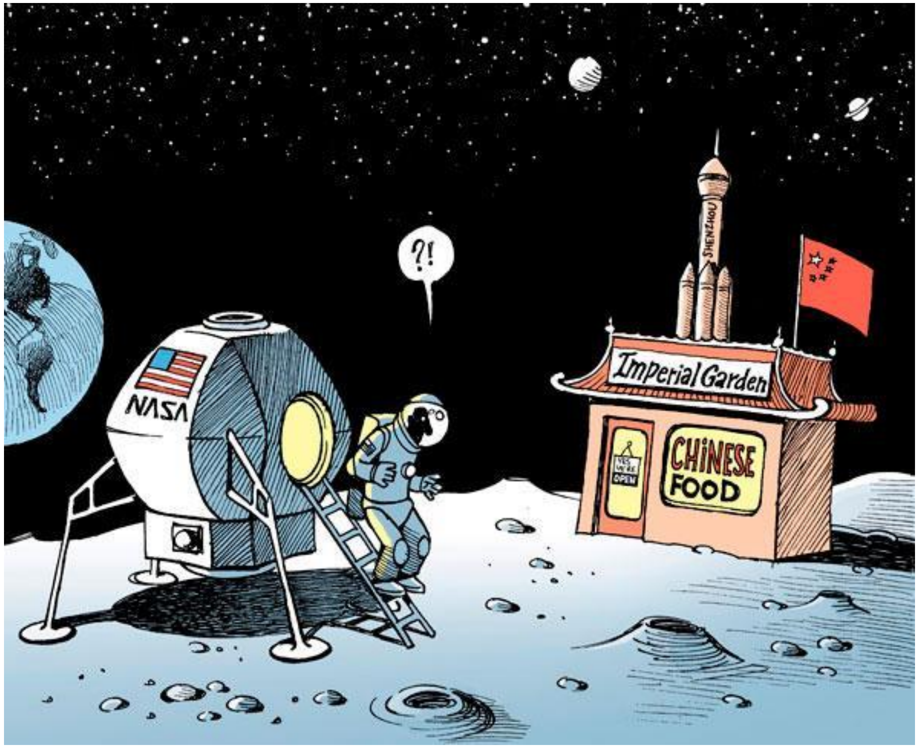


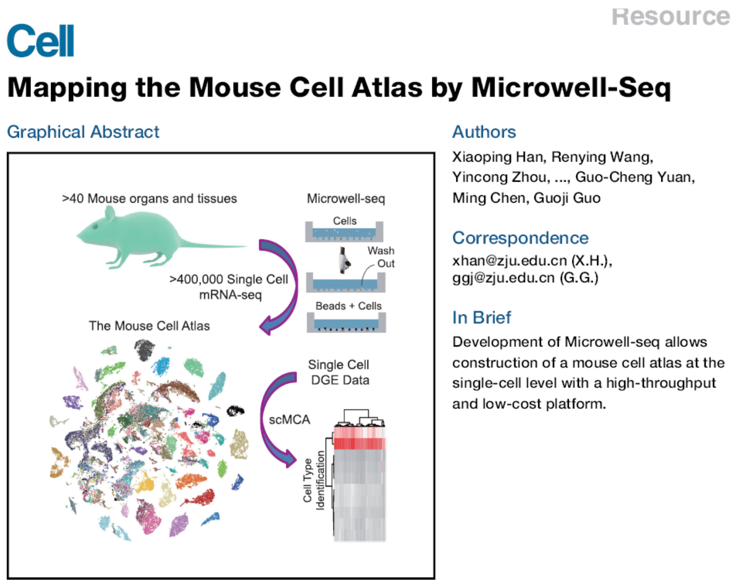
ARTICLE

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

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

The Tabula Muris Consortium*





Cell

Resource

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.)

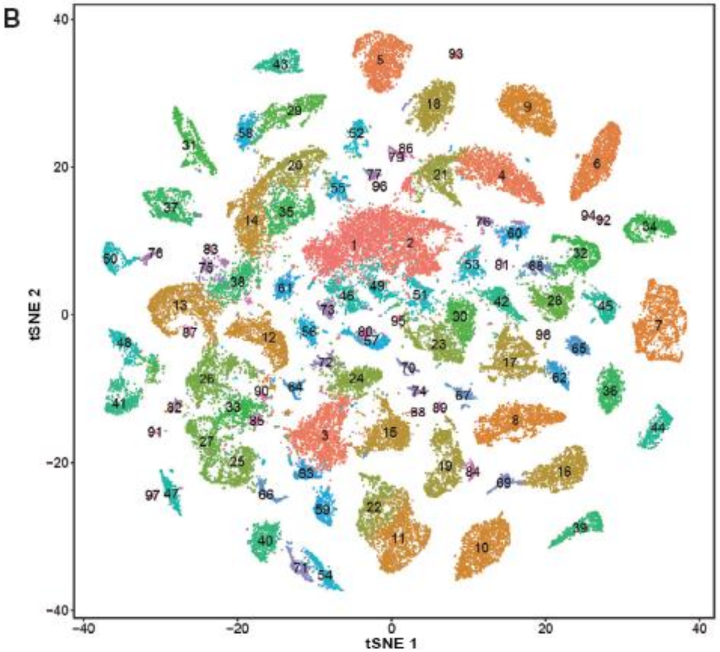
In Brief

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

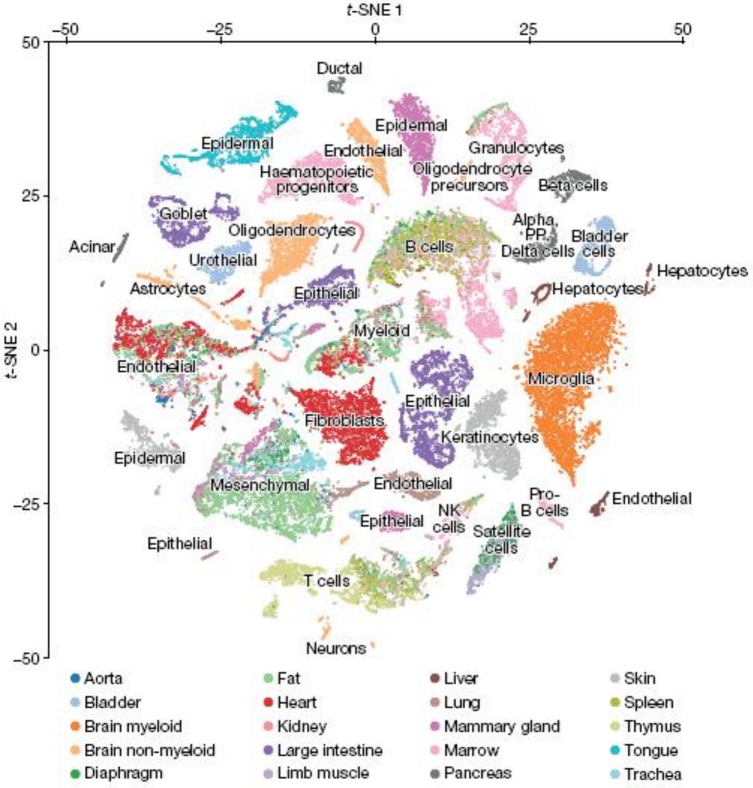
  based on 60,000 good QC cells



ISSUE: how do you deal with >100,000 cells?

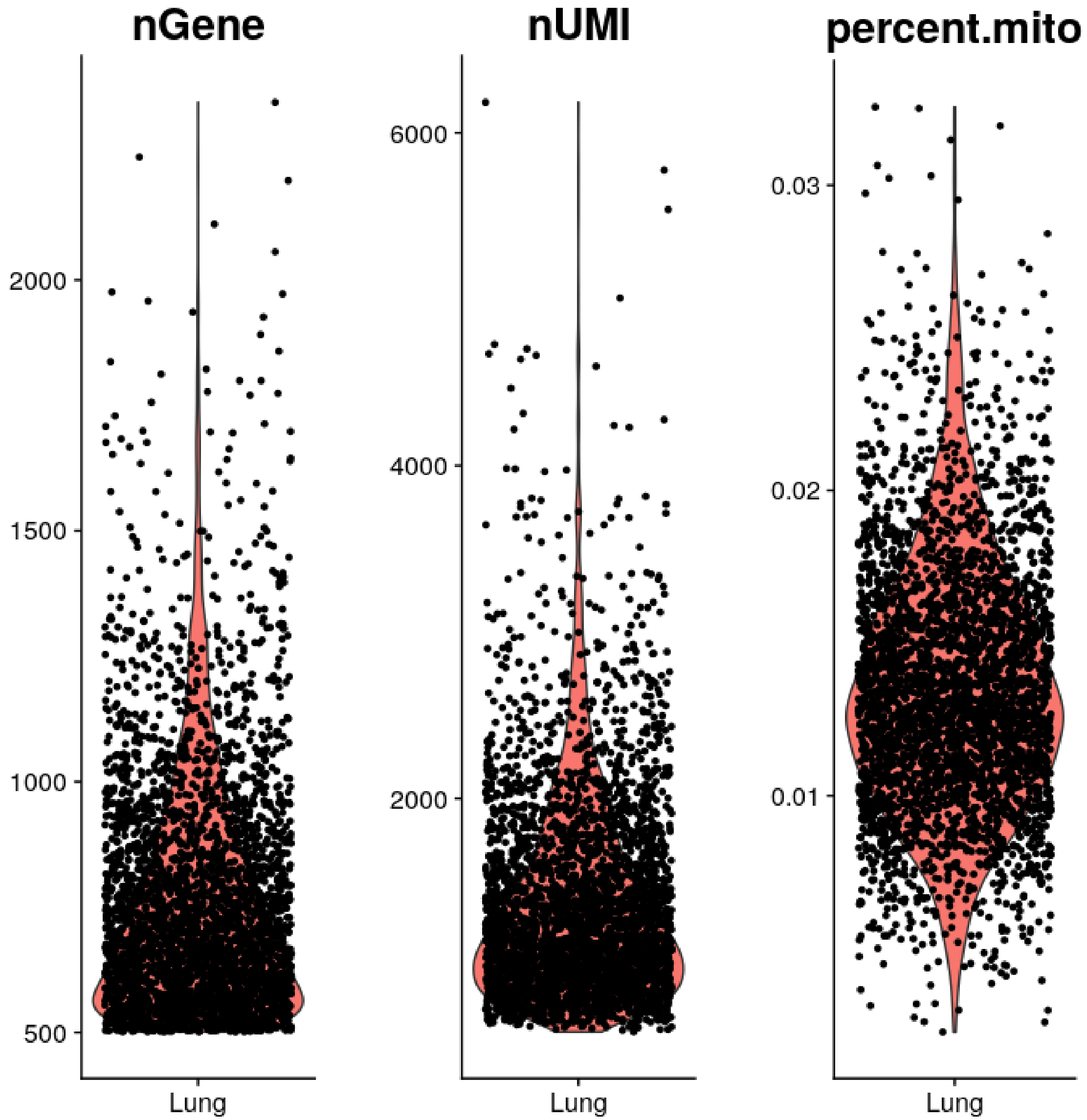
Tabula Muris Summary

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



MCA Lung data (6940 cells)

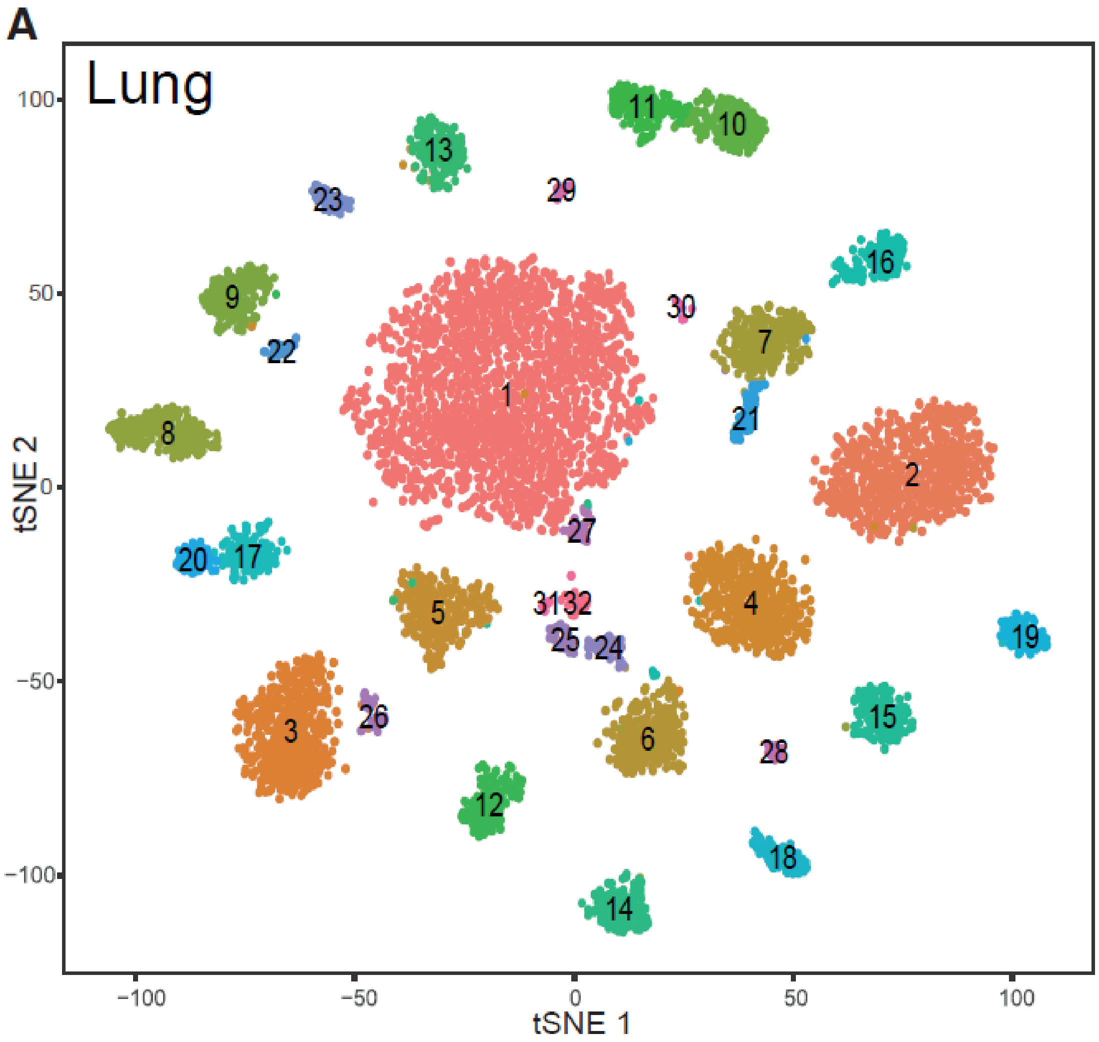
Han et Al, Cell (2018)



Dropouts 96 %

MCA Lung data (6940 cells)

Han et Al, Cell (2018)



- **30 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

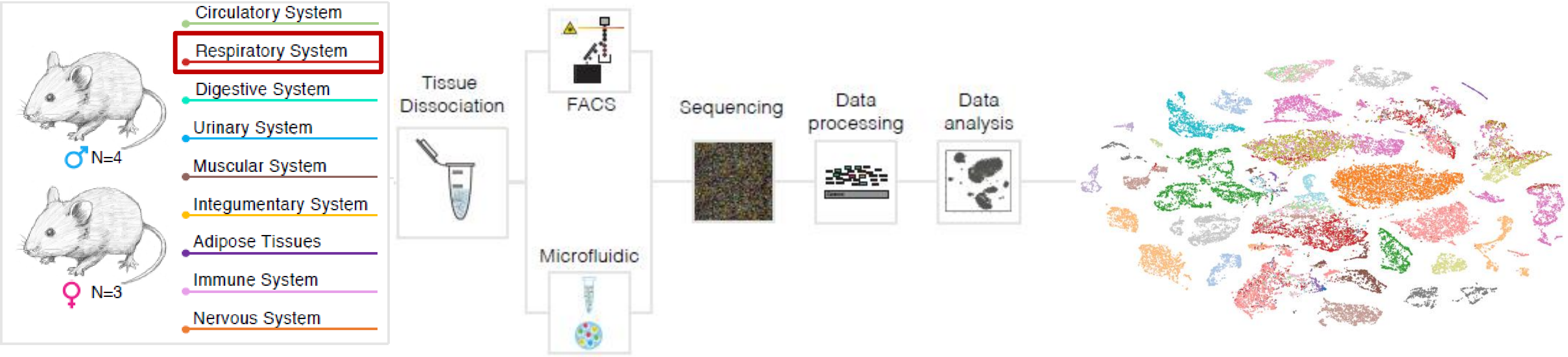
Gene expression and cell type markers available on : http://bis.zju.edu.cn/MCA/gallery.html?tissue=Lung

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

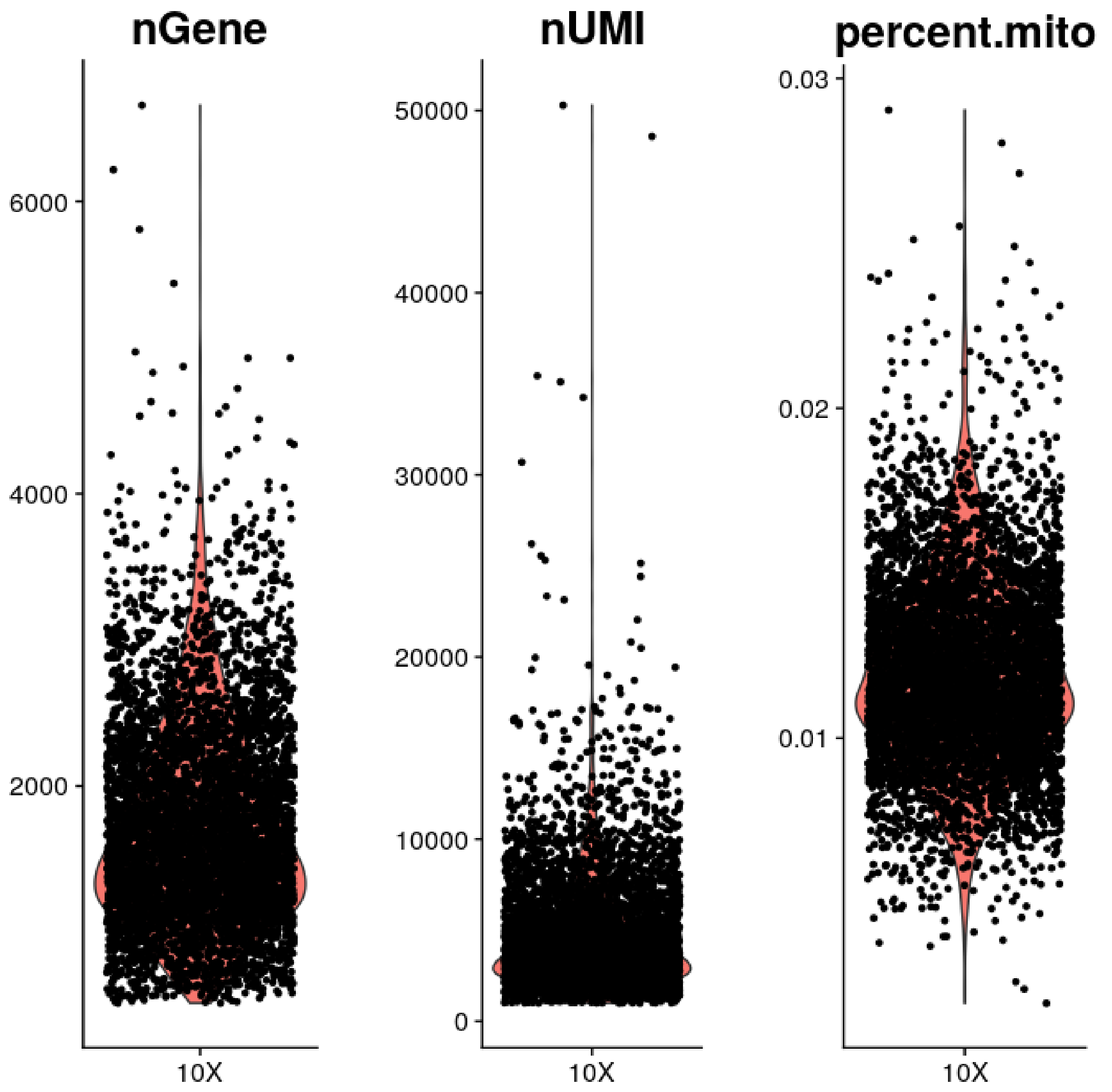
Lung	Trachea
1620 cells	1392 cells

10X Microfluidic droplet

Lung	Trachea
5449 cells	11269 cells

TM Lung 10X data (5449 cells)

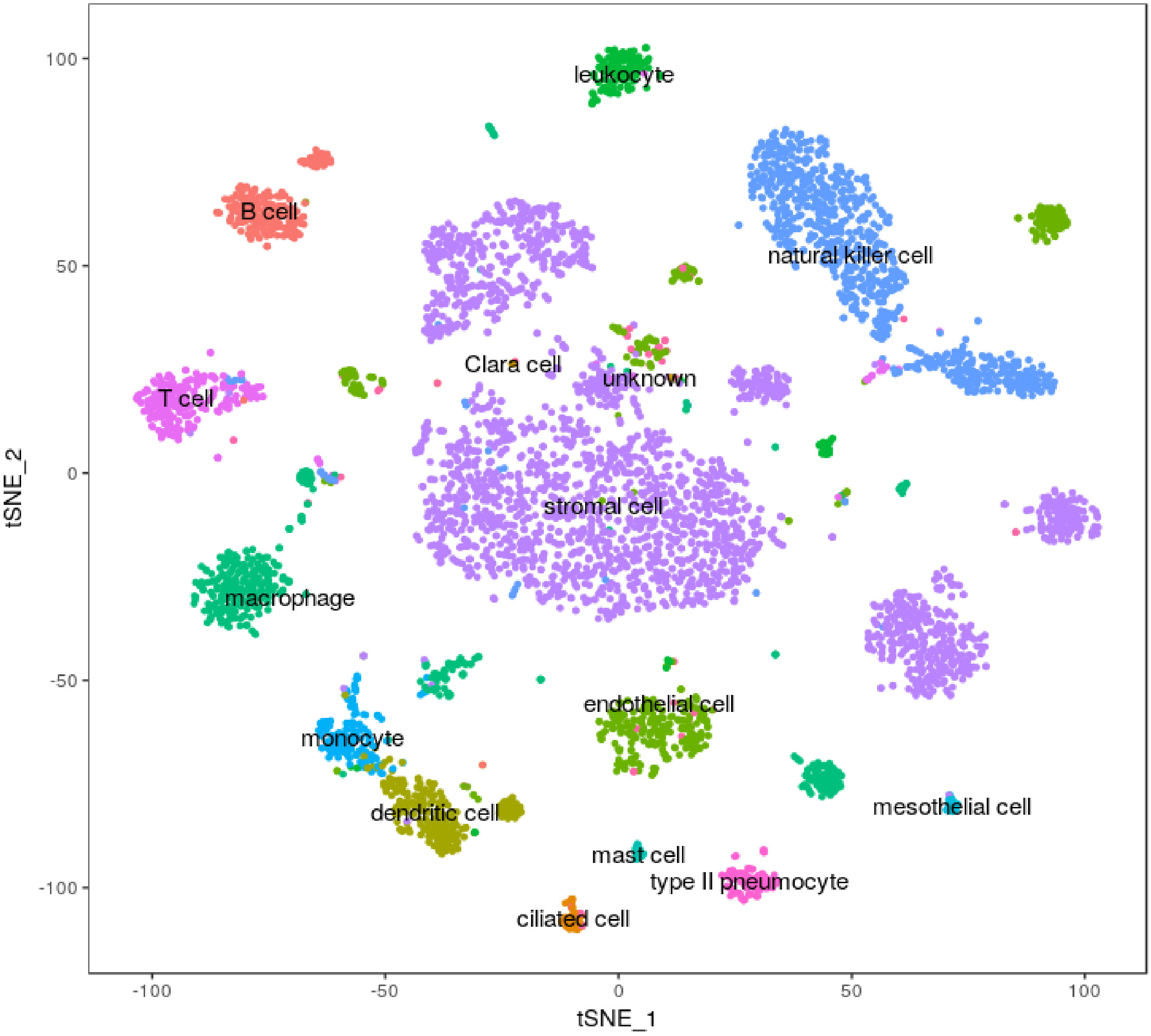
QC metrics



Dropouts 93 %

TM Lung 10X data (5449 cells)

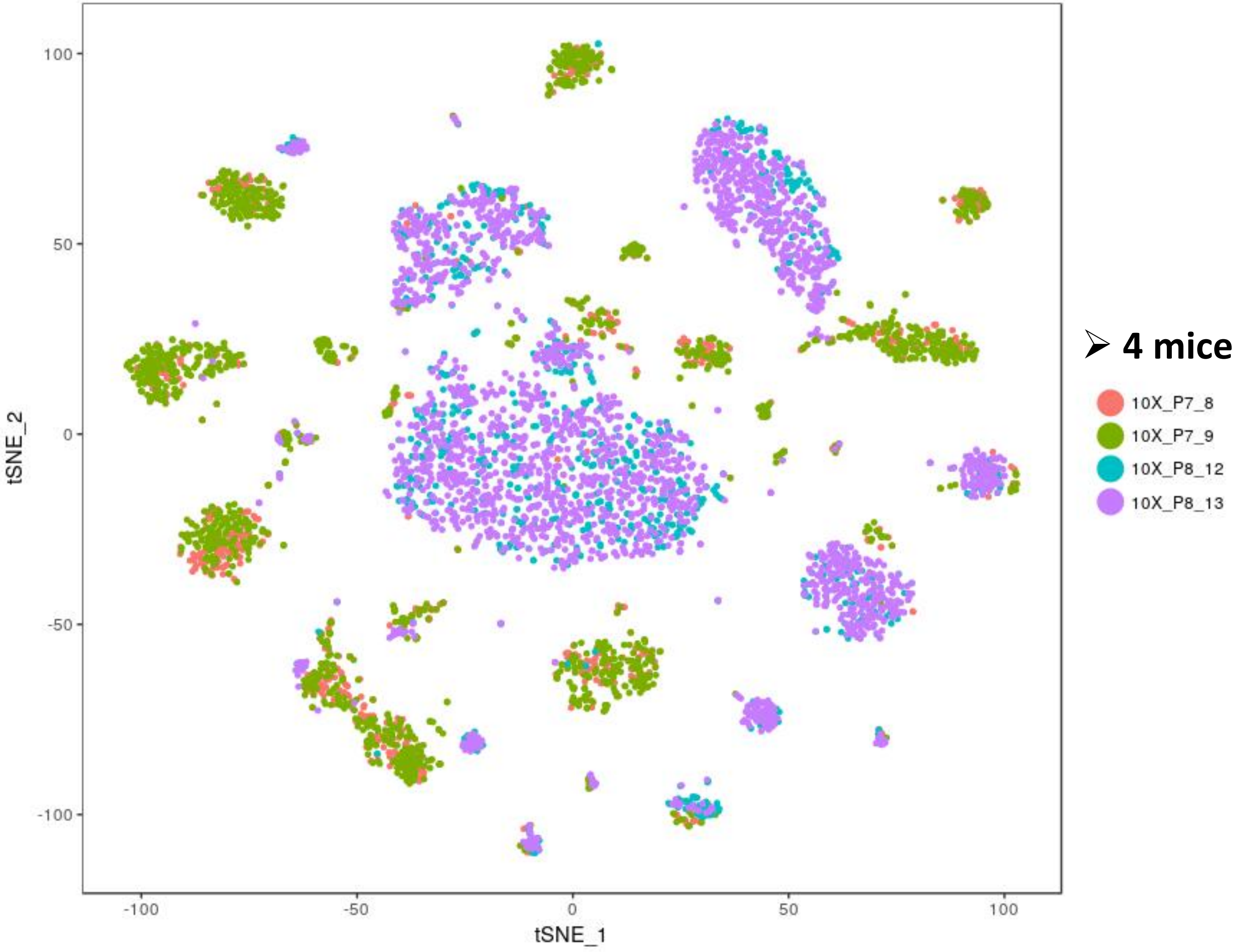
T-SNE of cell types



- 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

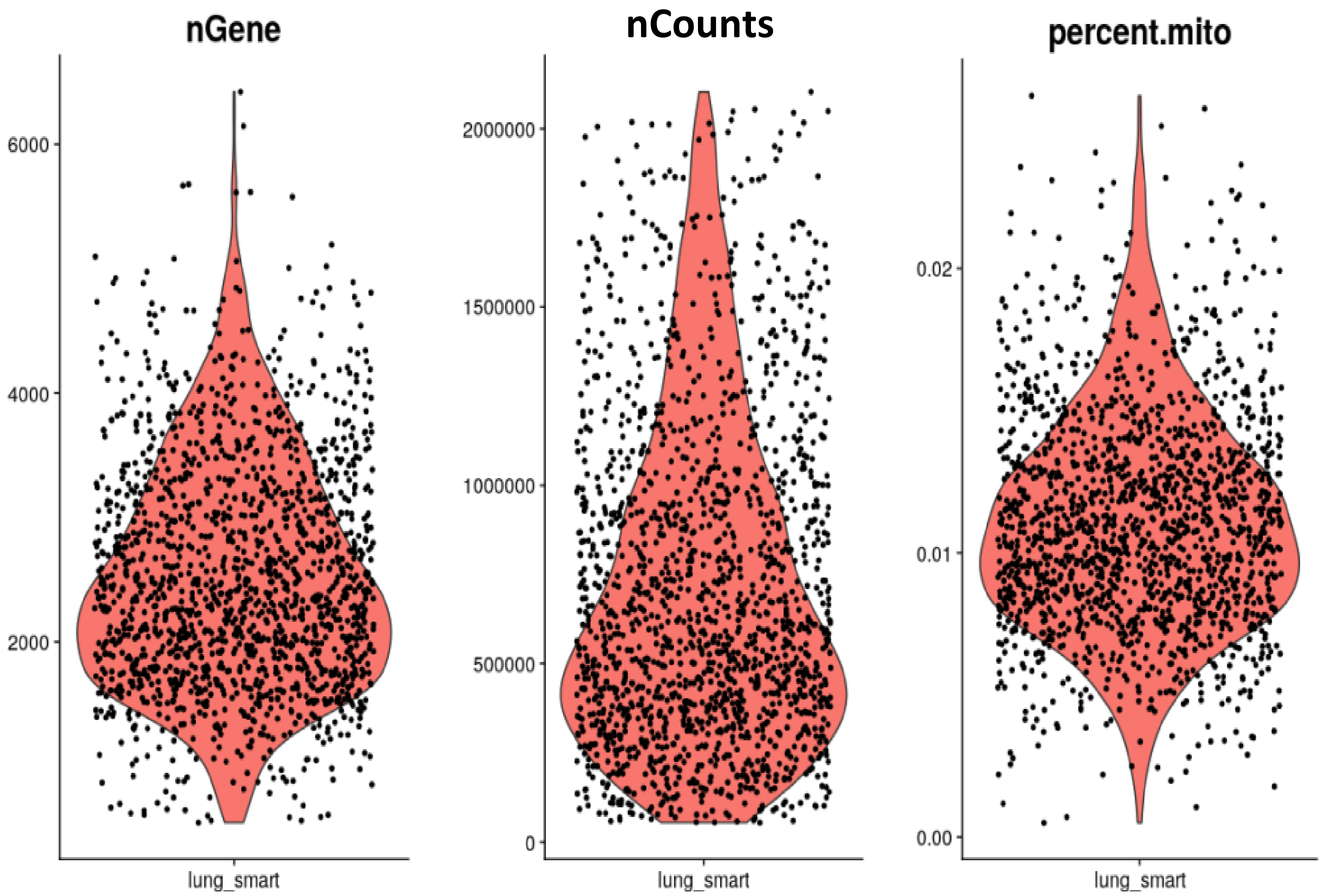
TM Lung 10X data (5449 cells)



- 4 mice

TM Lung SMART-Seq data (1620 cells)

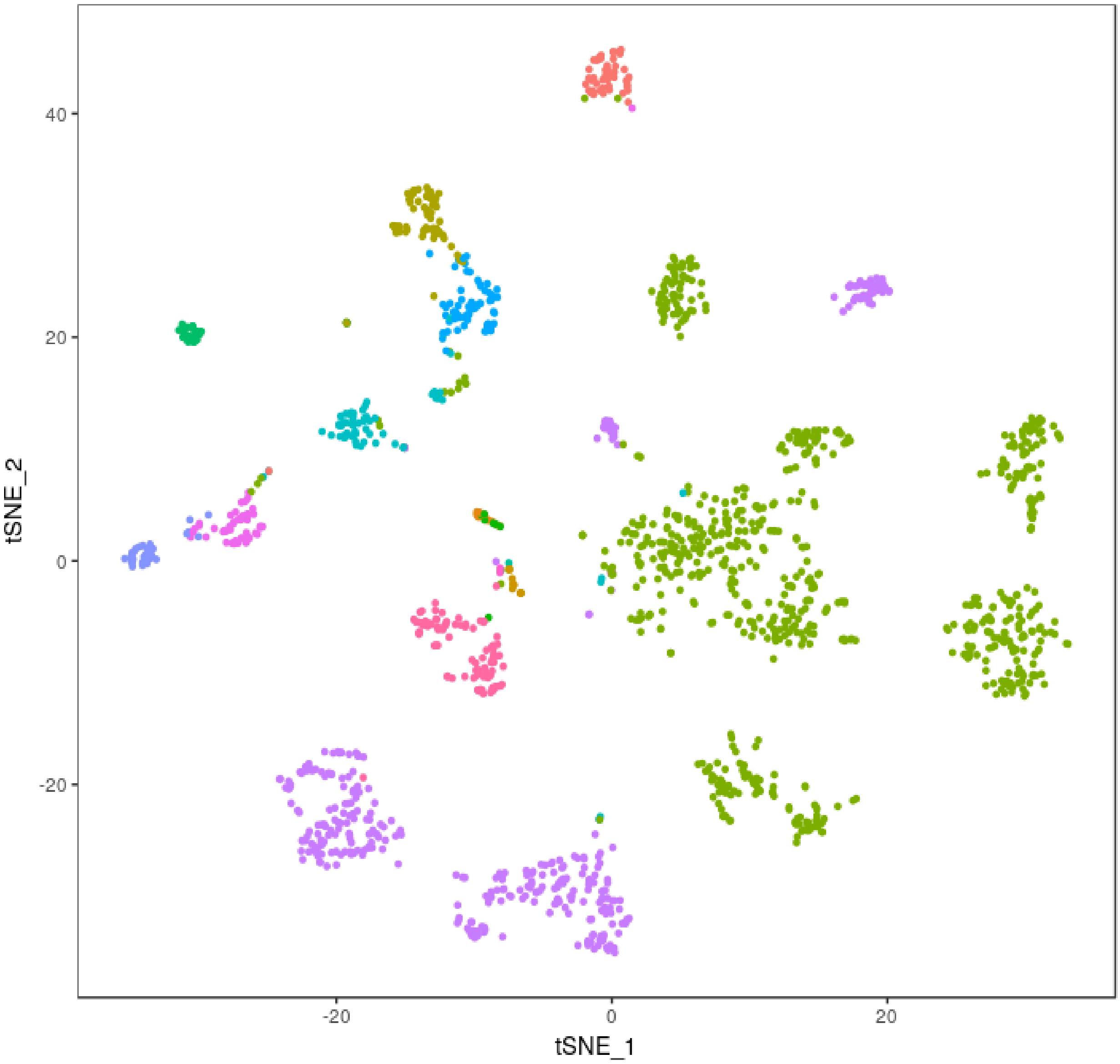
QC metrics



Dropouts 89 %

TM Lung SMART-Seq data (1620 cells)

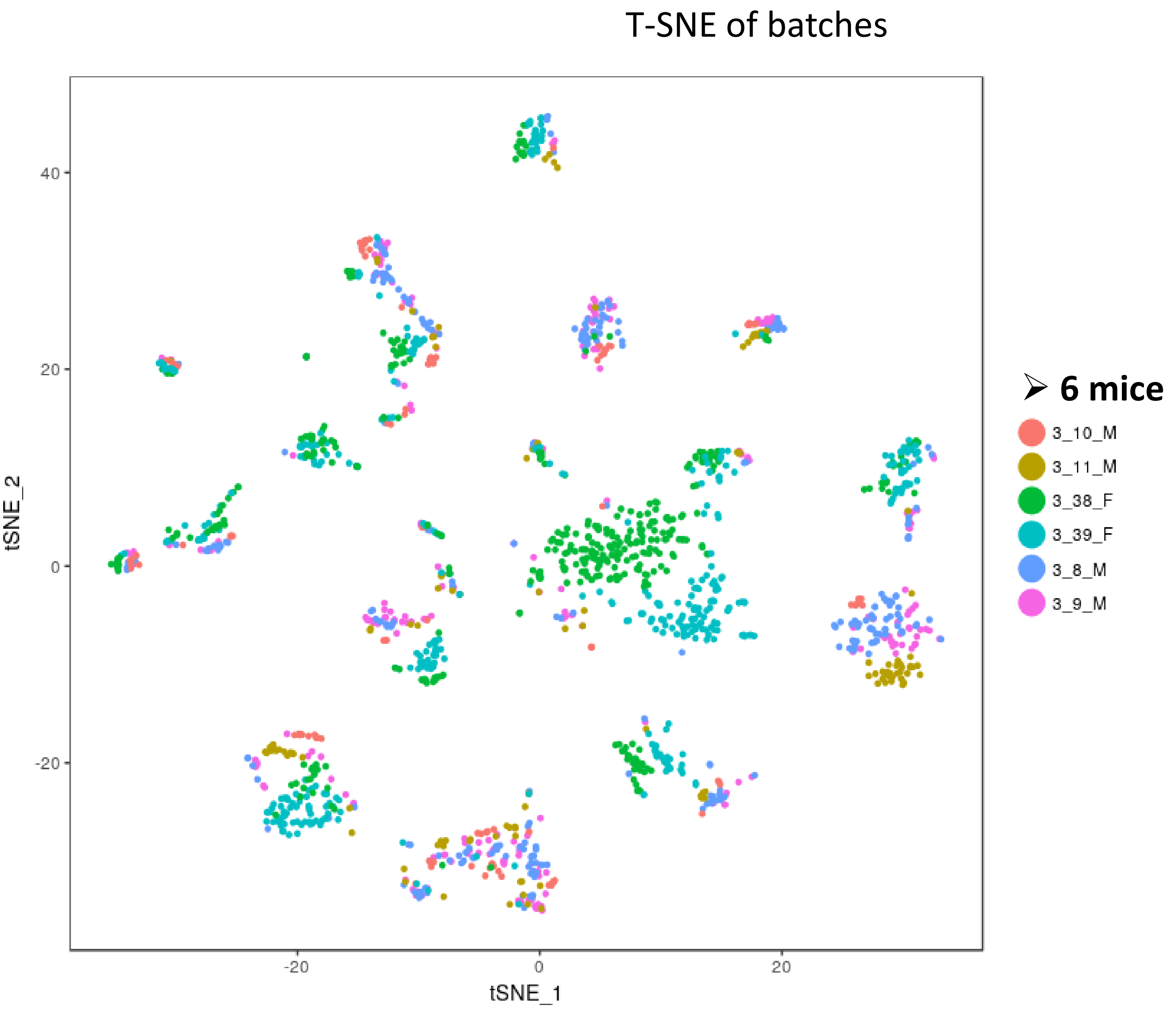
T-SNE of cell types



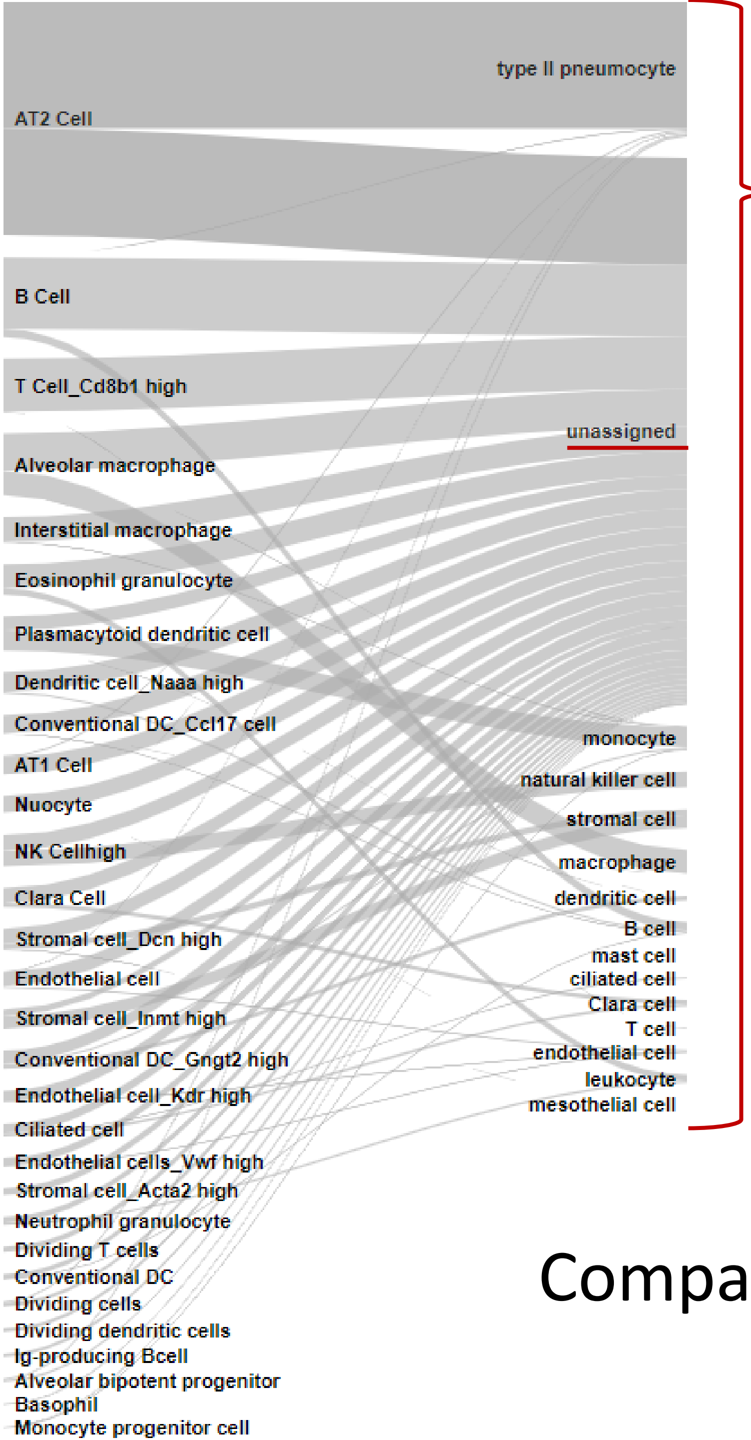
- 16 cell types (7 immune)

●	B cell	n = 55
●	ciliated cell	n = 14
●	Clara cell	n = 13
●	dendritic cell	n = 69
●	endothelial cell	n = 738
●	epithelial cell	n = 9
●	leukocyte	n = 31
●	lung neuroendocrine cell	n = 2
●	macrophage	n = 69
●	mesothelial cell	n = 2
●	monocyte	n = 65
●	natural killer cell	n = 36
●	stromal cell	n = 366
●	T cell	n = 55
●	type I pneumocyte	n = 2
●	type II pneumocyte	n = 94

TM Lung SMART-Seq data (1620 cells)


T-SNE of batches

- **6 mice**

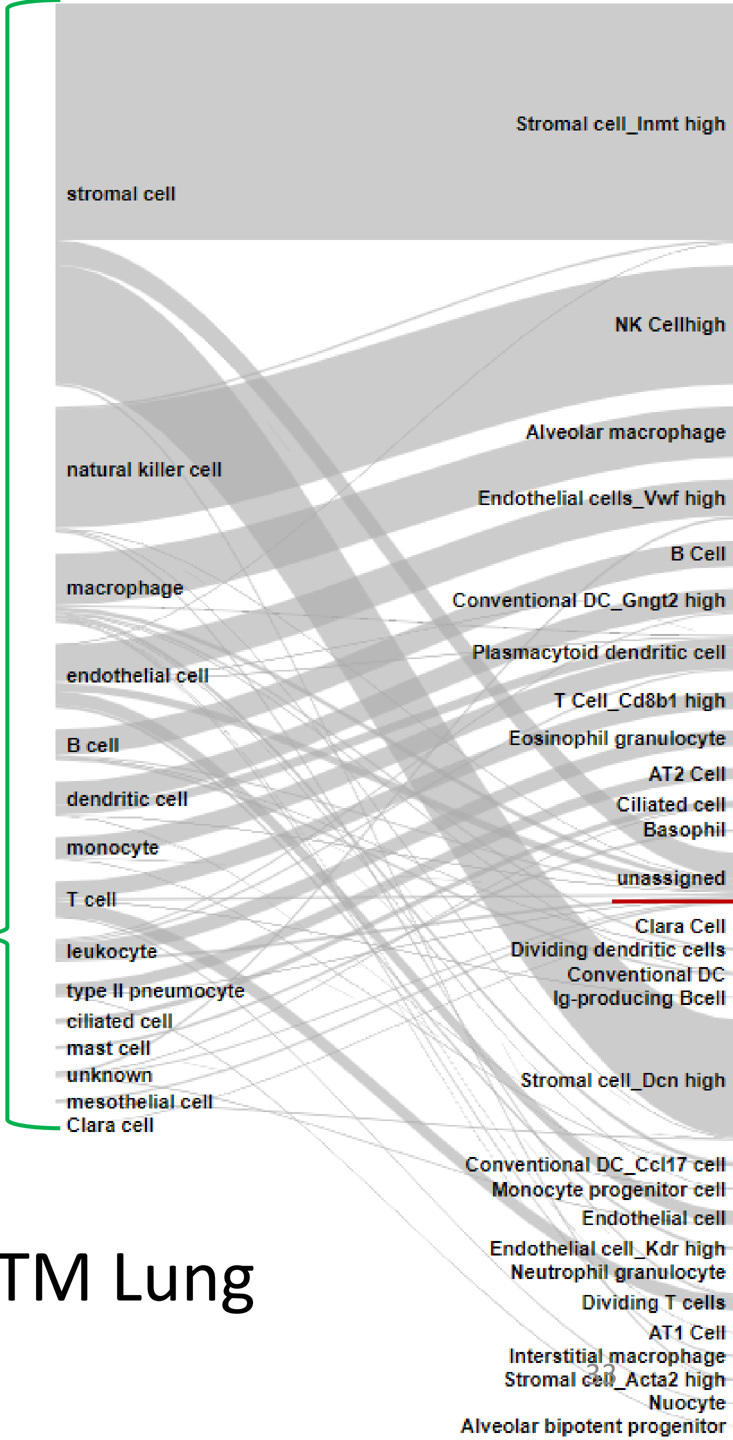


MCA (query)
on
TM (ref)

35%

90%

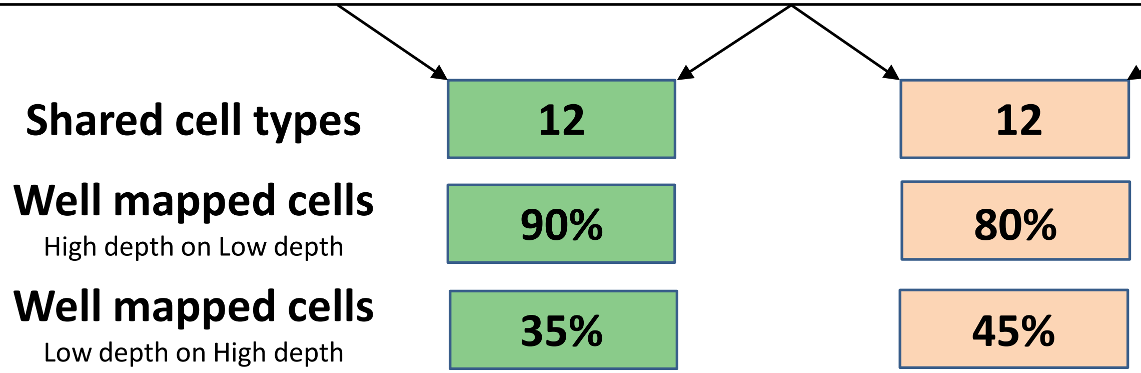
TM (query)
on
MCA (ref)



Comparing MCA and TM Lung 10X datasets

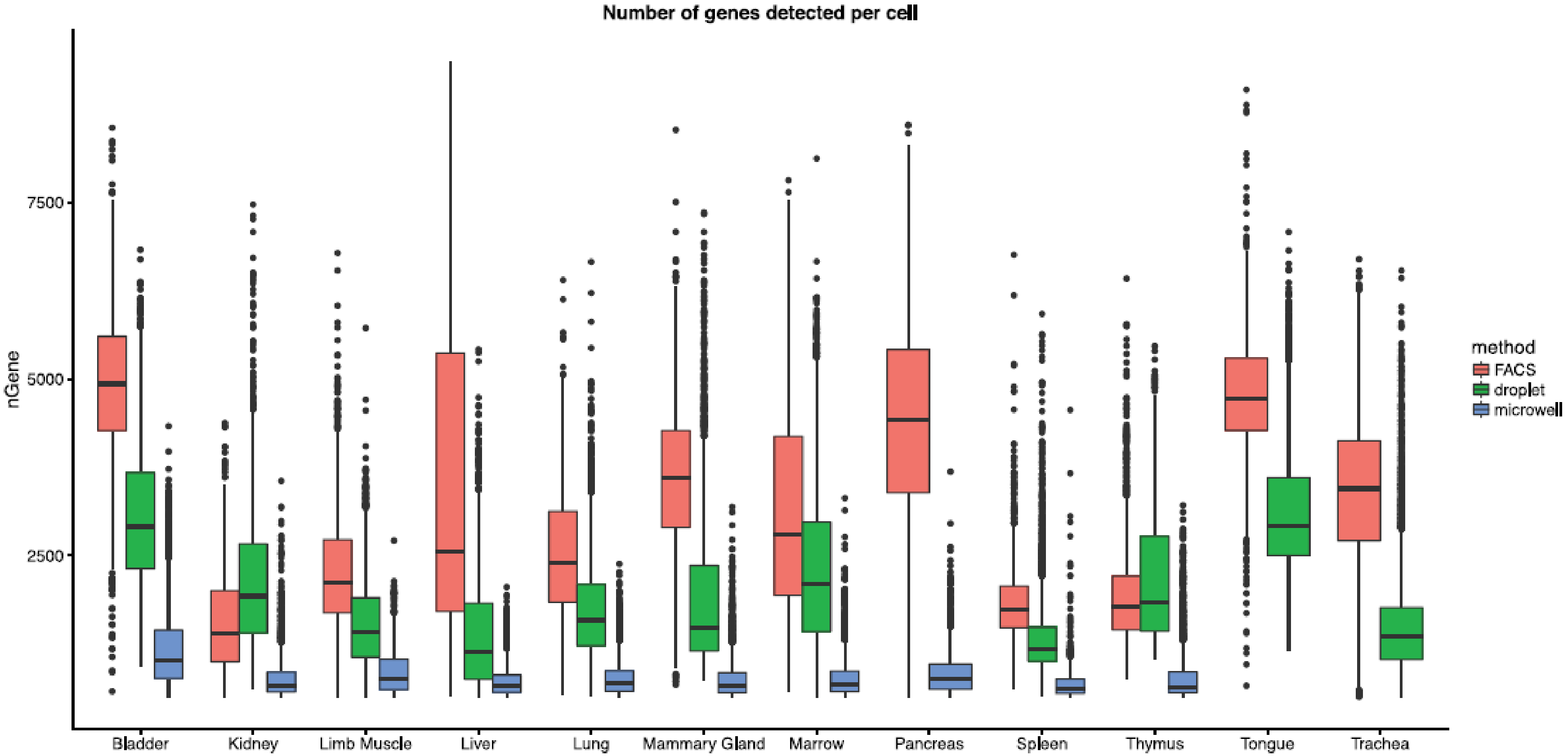
Conclusion

	MCA lung 10X	TM lung 10X	TM lung SMART-Seq
Nb of Cells	6940	5449	1620
Nb of cell types	30	15	16
Sequencing depth **(mean of detected genes)**	764	1200	2000
% of Dropouts	96%	93%	89%



Shared cell types	**12**	**12**
Well mapped cells High depth on Low depth	**90%**	**80%**
Well mapped cells Low depth on High depth	**35%**	**45%**

Comparison of the Mouse Atlases

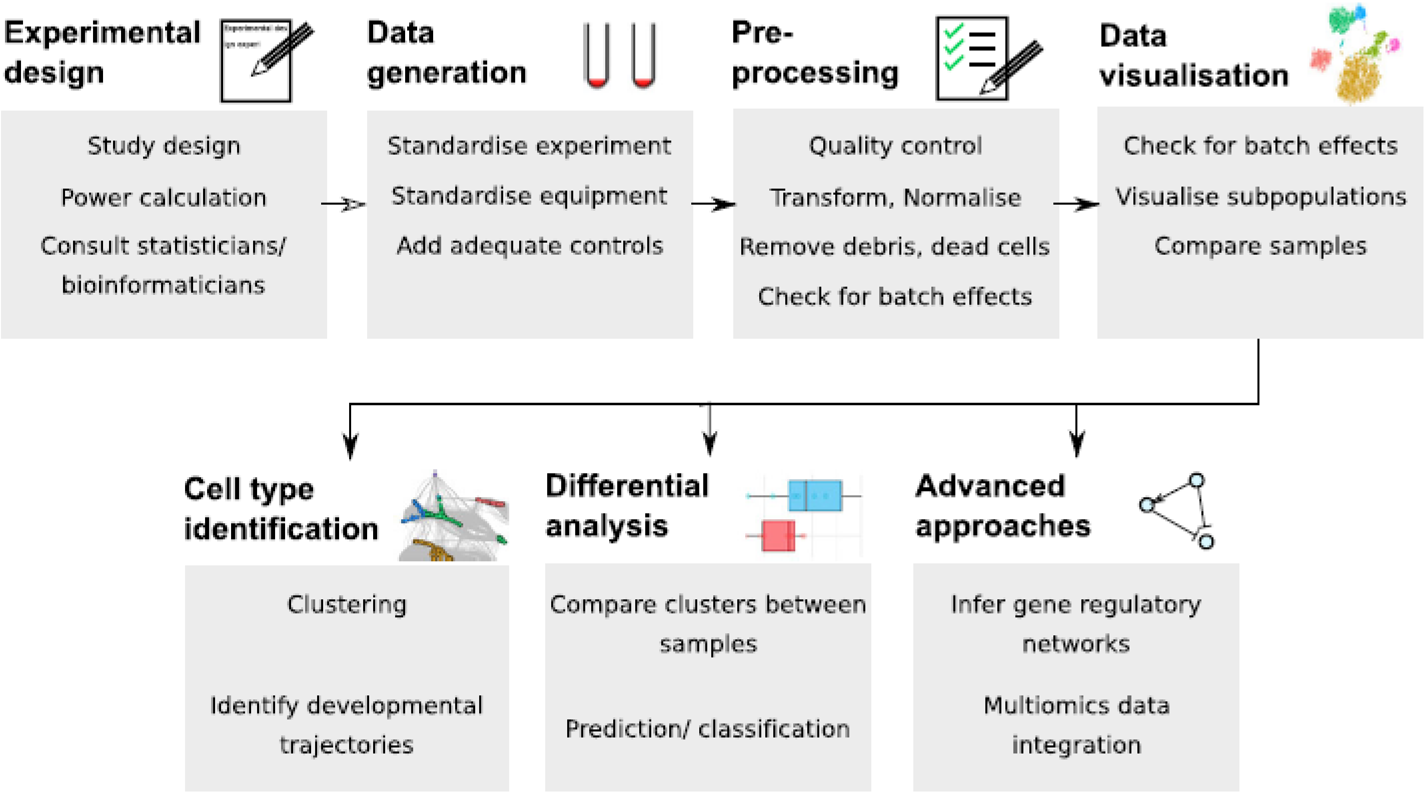


Tabula Muris, 2018

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

Standard Data Analysis Pipeline



Experimental design

Study design

Power calculation

Consult statisticians/ bioinformaticians

Data generation

Standardise experiment

Standardise equipment

Add adequate controls

Pre-processing

Quality control

Transform, Normalise

Remove debris, dead cells

Check for batch effects

Data visualisation

Check for batch effects

Visualise subpopulations

Compare samples

Cell type identification

Clustering

Identify developmental trajectories

Differential analysis

Compare clusters between samples

Prediction/ classification

Advanced approaches

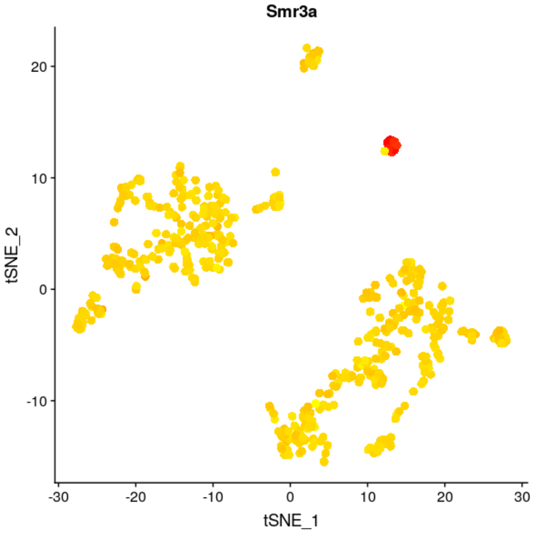
Infer gene regulatory networks

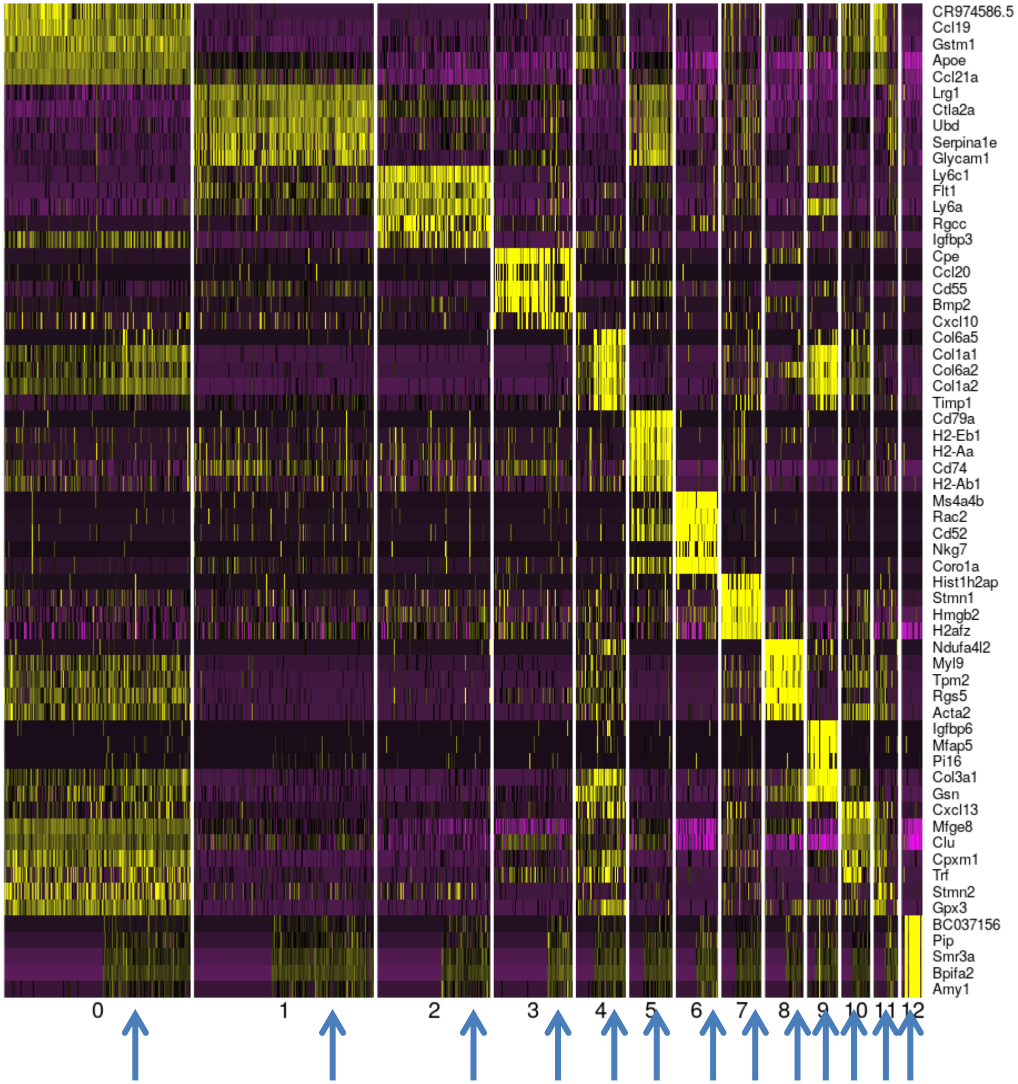
Multiomics data integration

Todorov, 2018

Preprocessing Steps

- Filtering of poor quality cells
  - Number of genes/UMI detected
  - % mitochondrial genes…
- Remove doublets
  - doubletFinder, scrublet
- Check for background issues
  - SoupX

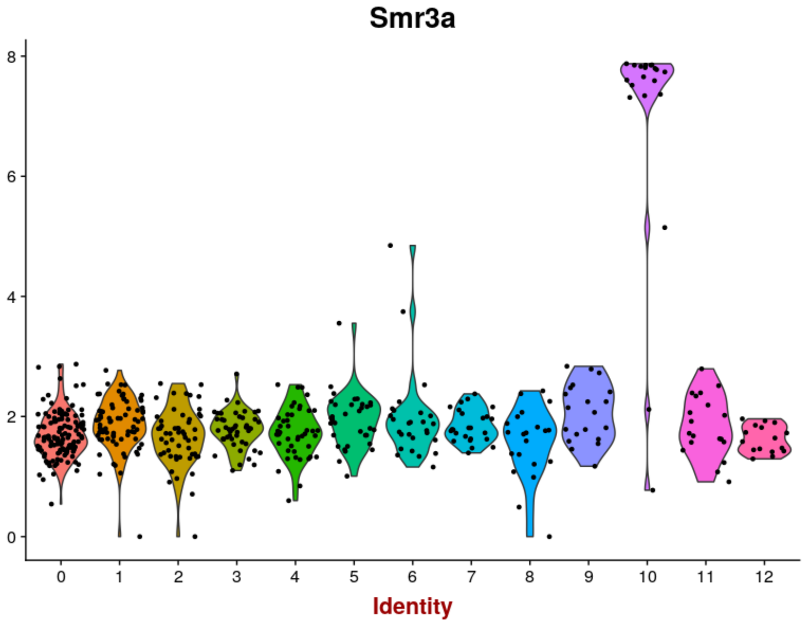




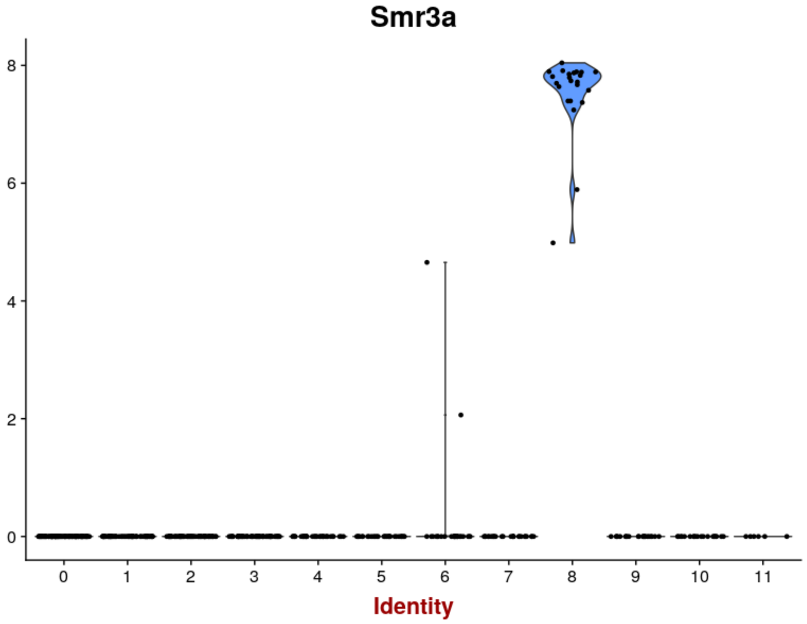
SoupX

Young MD, BiorXiv 2018

BEFORE SoupX



AFTER SoupX



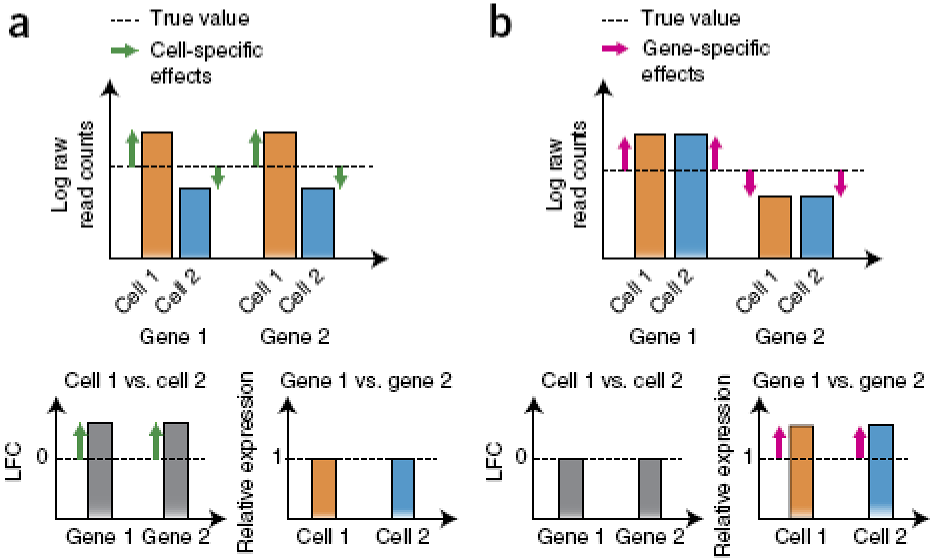
Warning: the software requires manual tuning.

NORMALIZATION

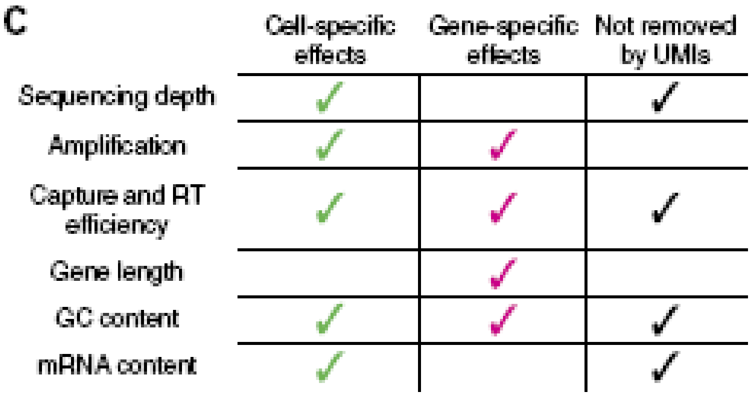
- Process of identifying and removing systematic variation not due to real differences between RNA treatments i.e. differential gene expression.

- 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 **<u>not DE</u>**
  - 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

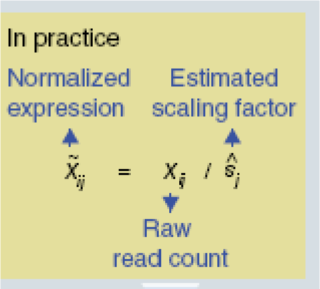
  1. Global scaling
  2. scRNA-seq specific method from scater/scran package
  3. Others

- 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

Normalized expression

Estimated scaling factor

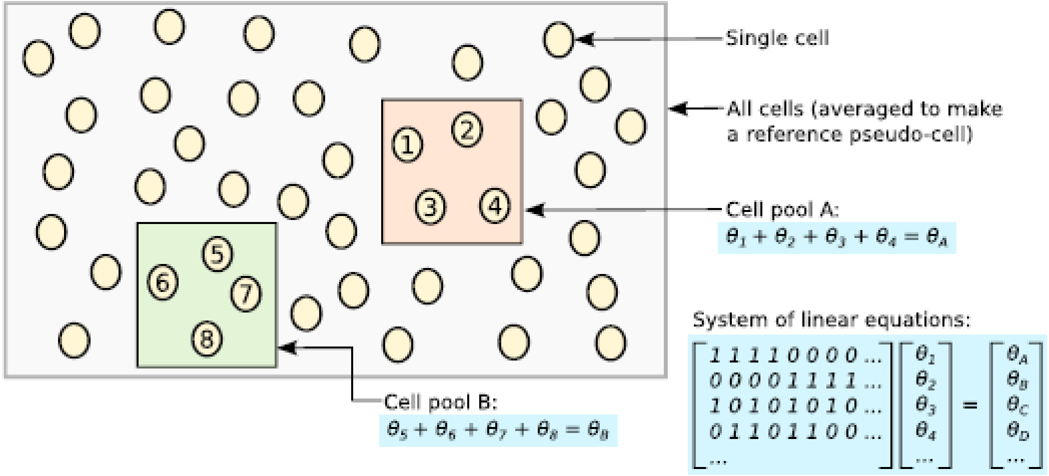
$\tilde{x}_{ij} = x_{ij} / \hat{s}_j$

Raw read count

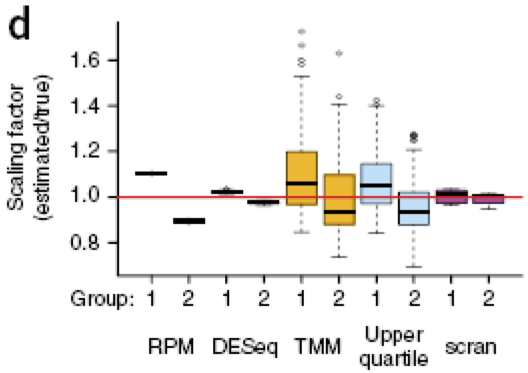
- In practice
  - Hypotheses are not always verified, but lots of people use this method anyway

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


Single cell
All cells (averaged to make a reference pseudo-cell)
Cell pool A:
$\theta_1 + \theta_2 + \theta_3 + \theta_4 = \theta_A$
Cell pool B:
$\theta_5 + \theta_6 + \theta_7 + \theta_8 = \theta_B$
1 2 3 4 5 6 7 8
System of linear equations:
$\begin{bmatrix} 1\,1\,1\,1\,0\,0\,0\,0 \ldots \\ 0\,0\,0\,0\,1\,1\,1\,1 \ldots \\ 1\,0\,1\,0\,1\,0\,1\,0 \ldots \\ 0\,1\,1\,0\,1\,1\,0\,0 \ldots \\ \ldots \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \\ \theta_4 \\ \ldots \end{bmatrix} = \begin{bmatrix} \theta_A \\ \theta_B \\ \theta_C \\ \theta_D \\ \ldots \end{bmatrix}$

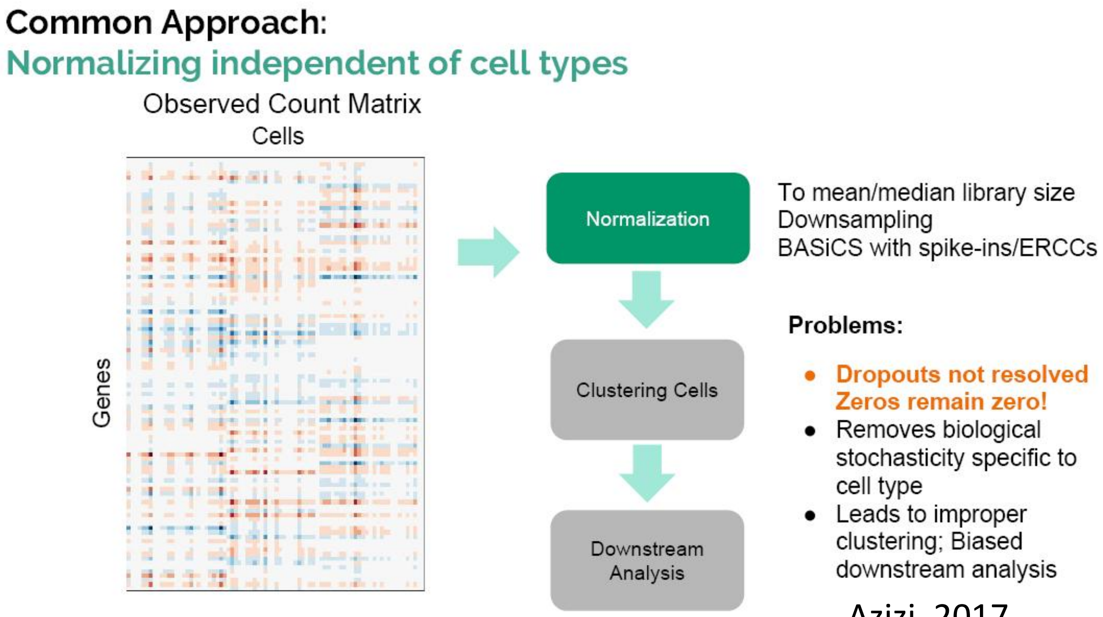
Lun, 2016



Vallejos C, 2017

Other methods are available

- Normalization included in the statistical model
  - SCDE, Monocle, MAST,…
- Normalization based on spike-ins or invariant genes
  - BASICs
- Can we be more creative?



Common Approach:

Normalizing independent of cell types

To mean/median library size
Downsampling
BASiCS with spike-ins/ERCCs

Problems:

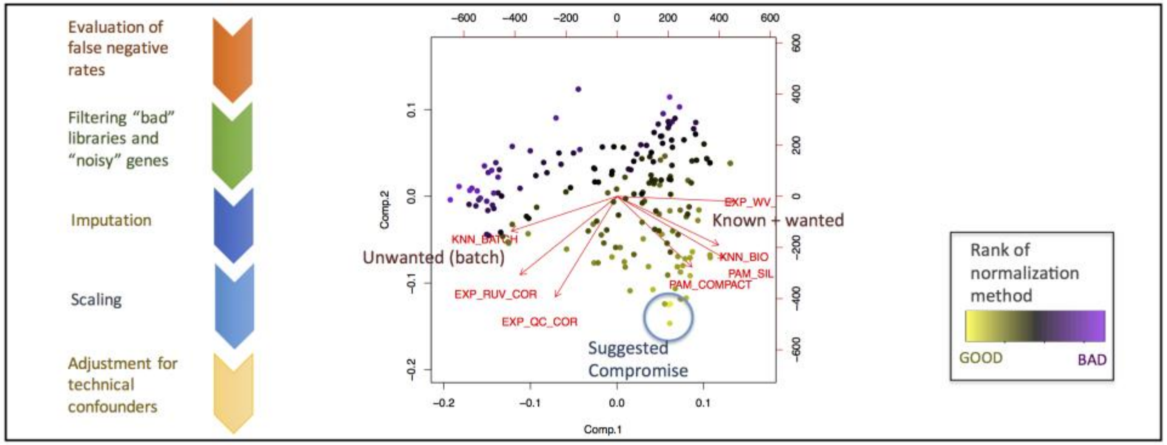
- **Dropouts not resolved Zeros remain zero!**
- Removes biological stochasticity specific to cell type
- Leads to improper clustering; Biased downstream analysis

Azizi, 2017

Integration of imputation, pre-processing and clustering

- SCONE (R package)

Cole M, Risso D (2018). scone: Single Cell Overview of Normalized Expression data. R package version 1.4.0.



Nir Yosef

- BISCUIT (R package)

Bayesian Inference for Single-cell Clustering and Imputing

Elham Azizi, 2017

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

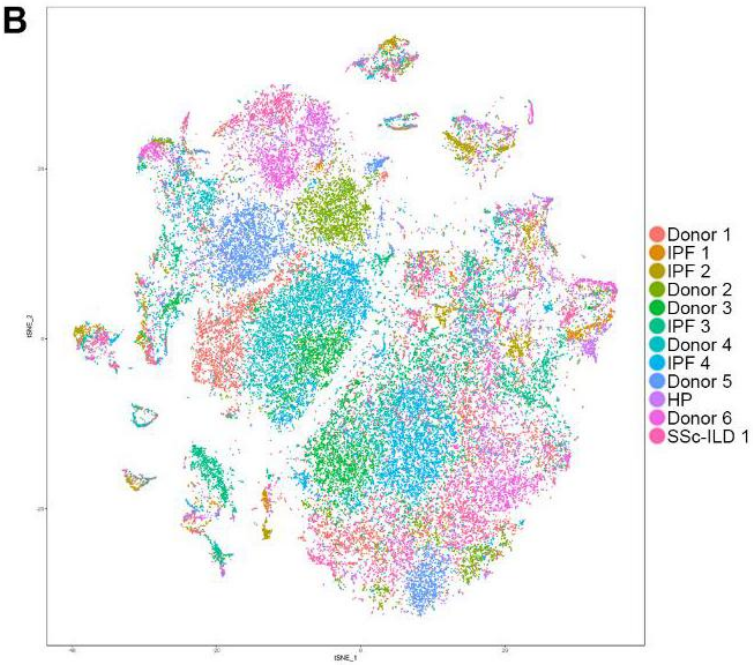
- Cell specific effects
  - Aim: make count distribution comparable
  1. Global scaling
  2. scRNA-seq specific method from scater/scran package
  3. Others

- 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



Misharin, BiorXiv 2018

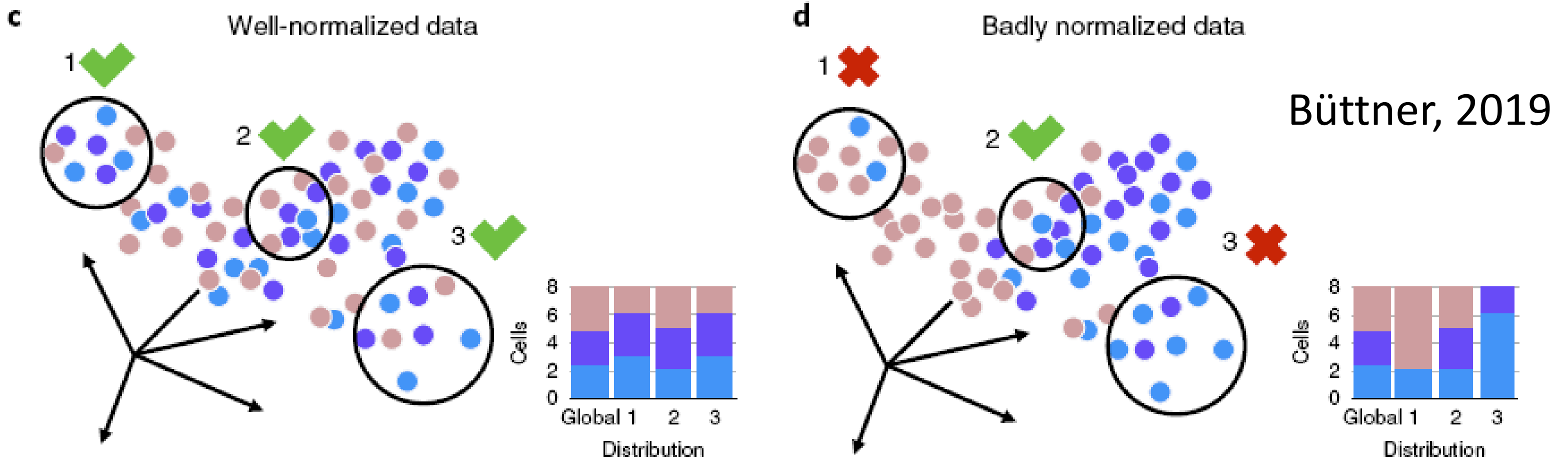
Data integration

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



Büttner, 2019

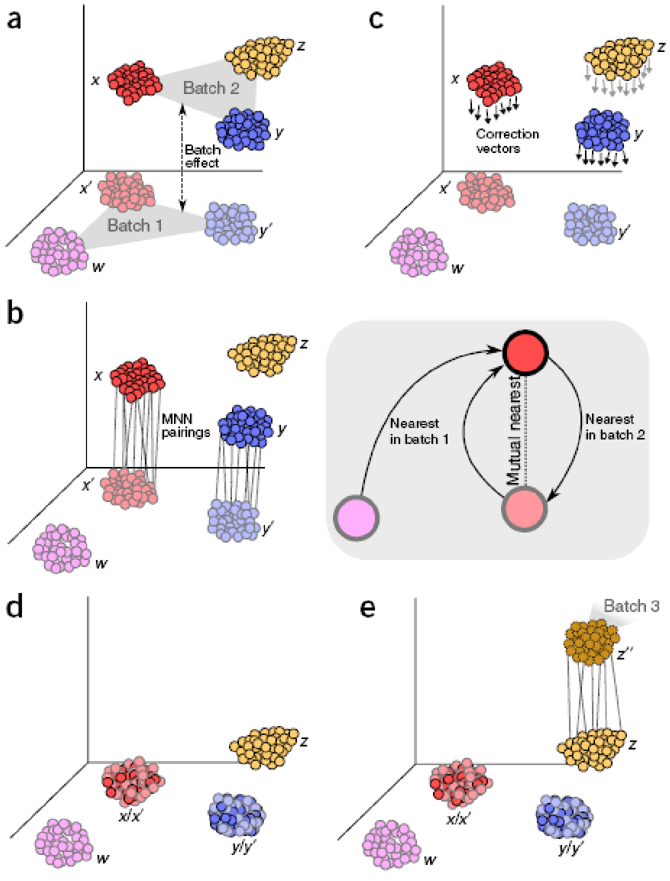
Current Data Integration Method

- MNN: Haghverdi, 2018
- Harmony, Korsunsky BiorXiv 2018
- Seurat V3, Stuart BiorXiv 2018

- ComBat (sva)

- No gold-standard yet

- Performance assessment?
  - Visual inspection
  - kBET (Büttner, 2019)
  - Other metrics?



Haghverdi, 2018

References

- 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

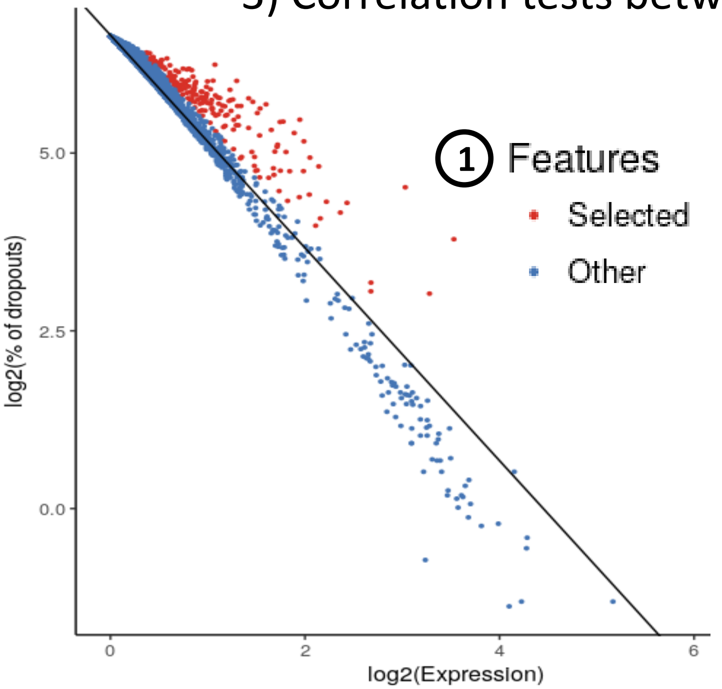
Comparing datasets with SC-Map

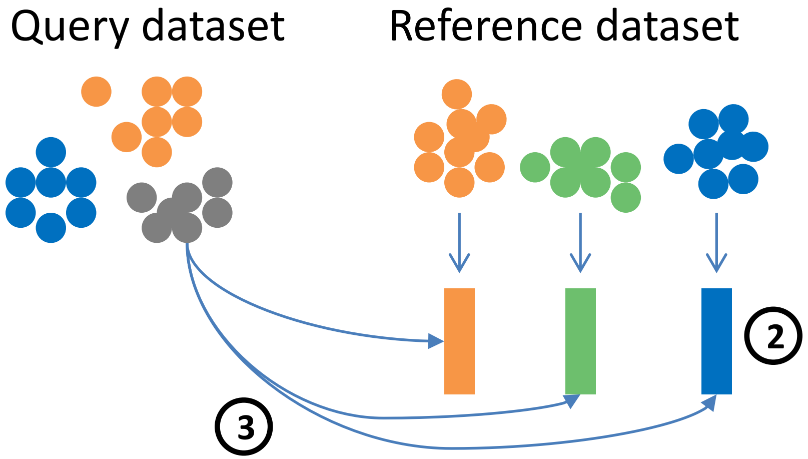
Kiselev et al., Nature Methods (2018)

SC-Map = R package with a label-centric approach, focused on trying to identify equivalent cell-types across datasets by comparing individual cells or groups of cells.

Method used = sc-map cluster

3 steps :   1) Selection of most informative genes (cell types markers)
2) Compute expression median for selected genes in all cells of each cluster
3) Correlation tests between each query cells and reference expression profiles


Features
● Selected
● Other
log2(% of dropouts)
log2(Expression)



Query dataset
Reference dataset
2
3