Single Cell RNA Sequencing

Ahmed Mahfouz

Assistant professor, Leiden Computational Biology Center

a.mahfouz@lumc.nl

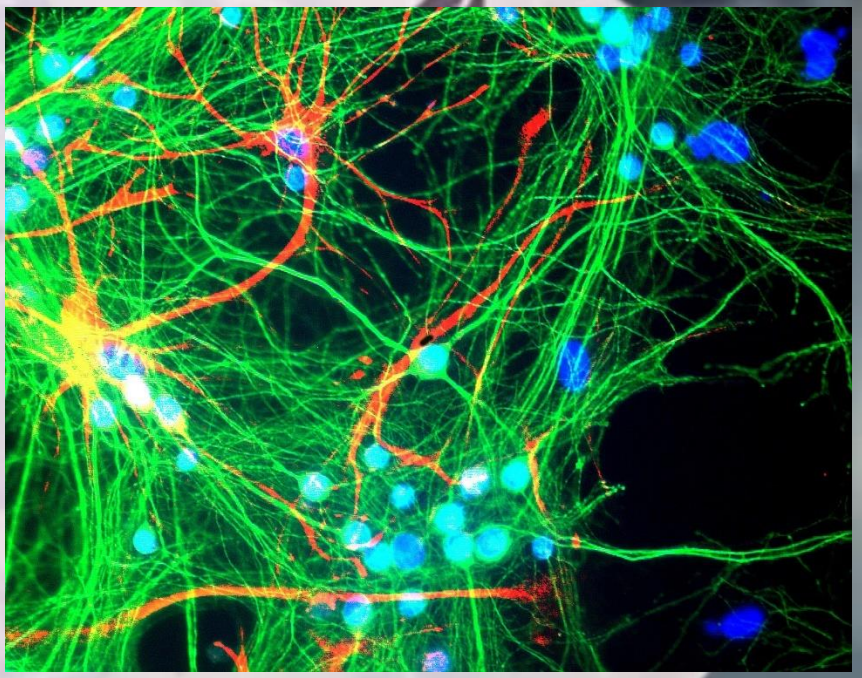
https://www.lcbc.nl/

@ahmedElkoussy

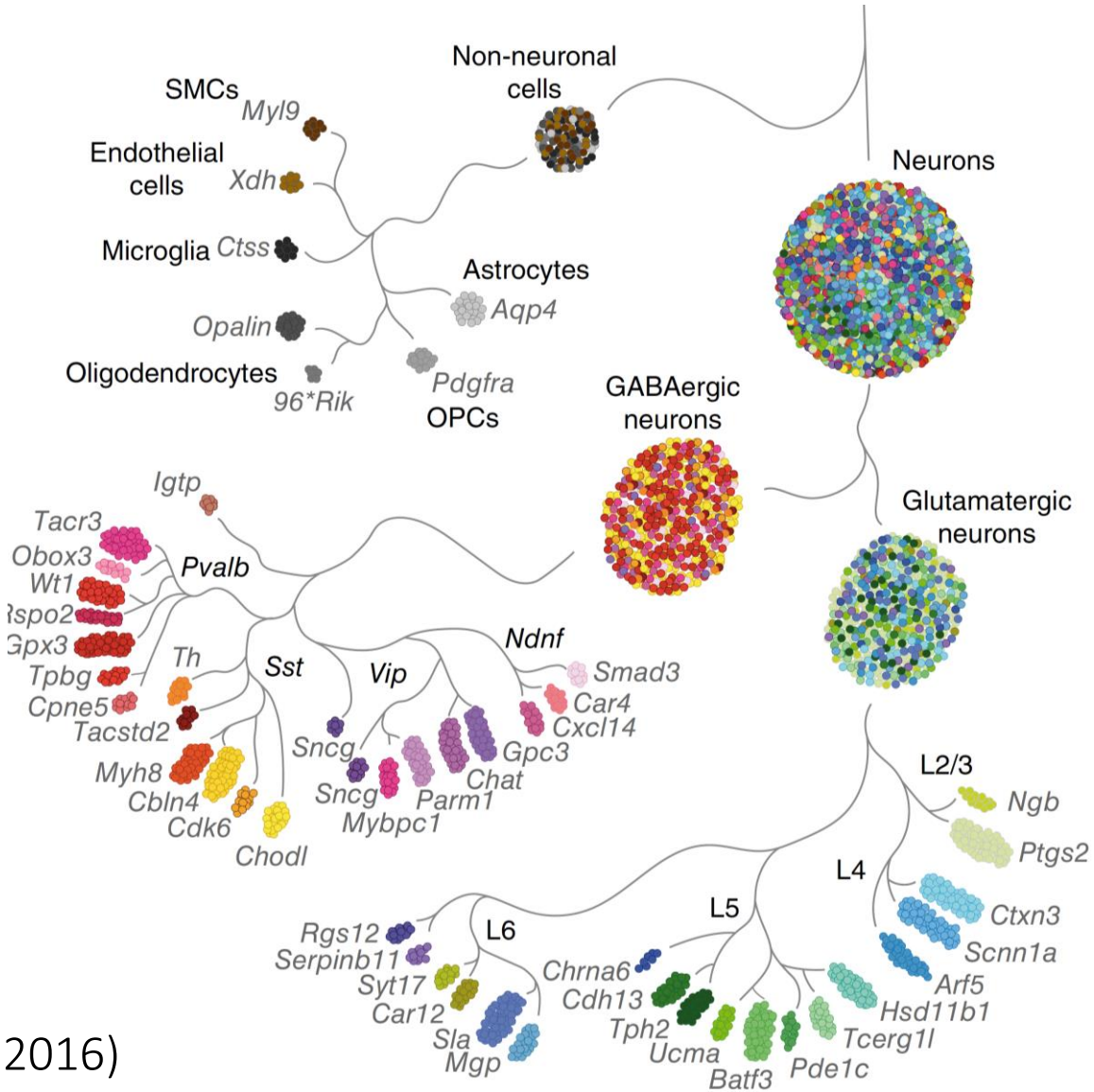
FOS Molecular Data Science – 31 October 2018

Leiden Computational Biology Center

LUMC TU Delft



Cell types in the brain

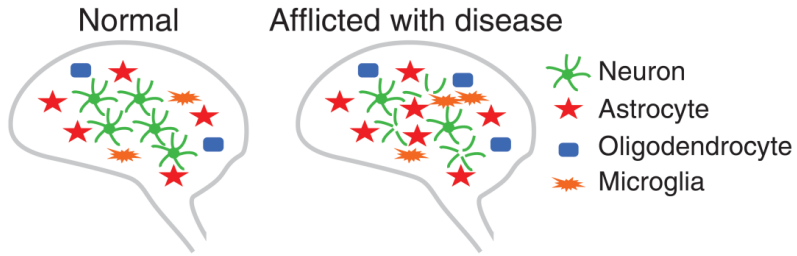


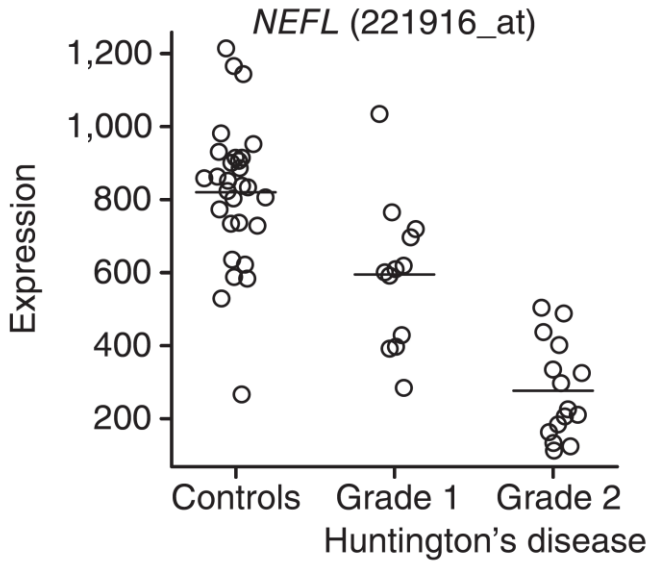
Tasic et al. (Nat. Neuroscience 2016)

Cellular heterogeneity in the brain

Human Molecular Genetics, 2006, Vol. 15, No. 6 **965–977**
doi:10.1093/hmg/ddl013
Advance Access published on February 8, 2006

Regional and cellular gene expression changes in human Huntington's disease brain





1) Decreased number of neurons?

2) Decreased expression?

3) Both?

Hodges et al. (Human Mol Genet 2006); Kuhn et al. (Nat Methods 2011)

How can we study single cells?

Technology	Measurements (P)	Cells (N)	Throughput	Pro	Con
Flow cytometry	1-15	1k-100k	big N, small P	Technically easy	Limited targets
Mass cytometry	20-50	1k-100k	big N, medium P	>P than flow	Limited targets
RNA FISH	1	~100	small N, small P	Spatial resolution	Technically hard, low throughput
Multiplex FISH	~100	100's	medium N, medium P	Spatial resolution	Technically and analytically hard
SS2 scRNA-seq	~20,000	100-1000	medium N, big P	High throughput	Sparse, low input material
Droplet scRNA-seq	~20,000	100-1M	big N, big P	High throughput	*Very* sparse, low input material

Every method has it's pros and cons.

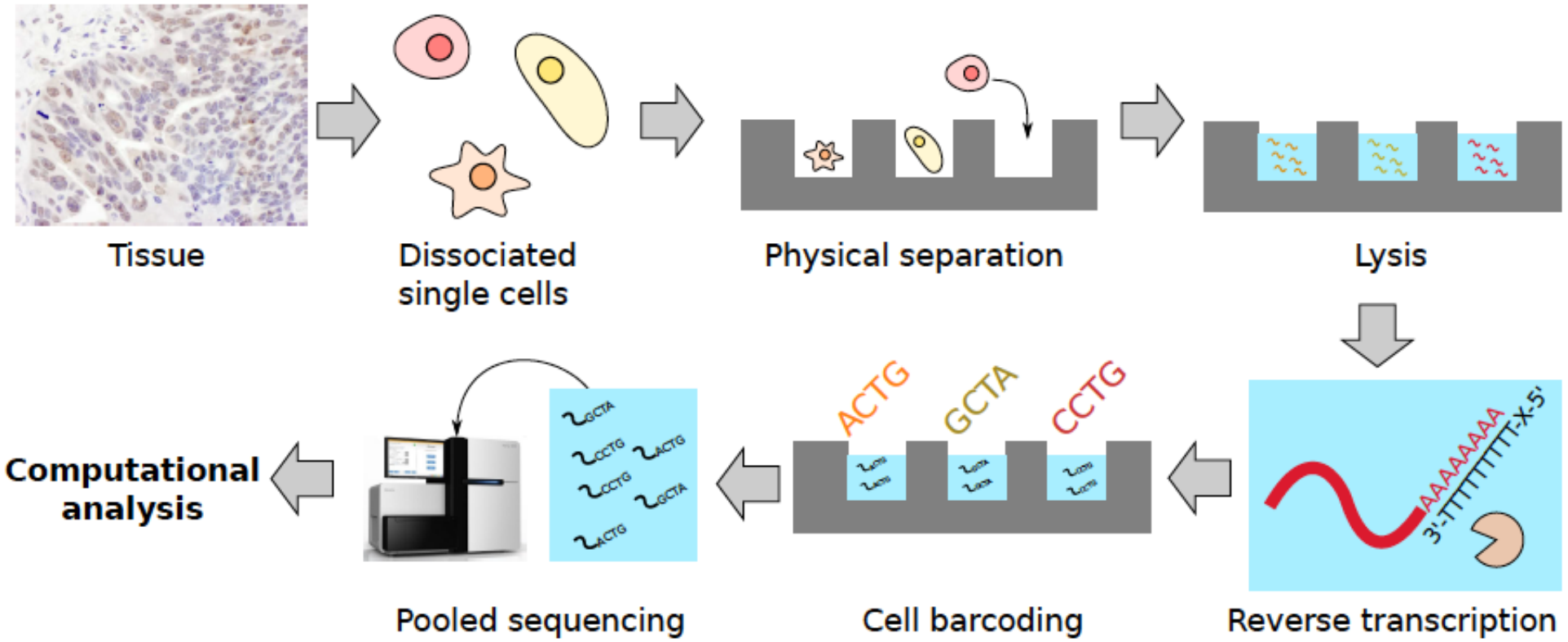
Single cell RNA-sequencing **(scRNA-seq)**

mRNA-Seq whole-transcriptome analysis of a single cell

Fuchou Tang[1,3], Catalin Barbacioru[2,3], Yangzhou Wang[2], Ellen Nordman[2], Clarence Lee[2], Nanlan Xu[2], Xiaohui Wang[2], John Bodeau[2], Brian B Tuch[2], Asim Siddiqui[2], Kaiqin Lao[2] & M Azim Surani[1]

Nature Methods 6, 377 - 382 (2009)

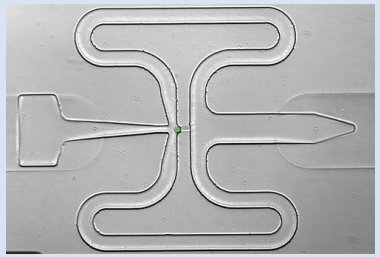
scRNA-seq Workflow



scRNA-seq Protocols

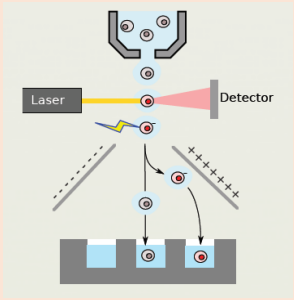
Physical separation methods

Microfluidic device



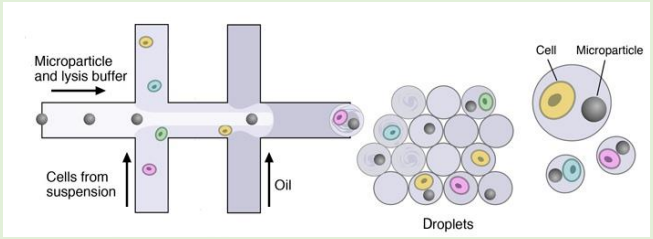
- 96 or 800 well format
- Physically check cells
- High capture efficiency
- Doublet issues
- Expensive
- Full-length cDNA (SMART-seq2)
- Spike-in control RNA
- **High gene coverage**

Plate-based



- 96 or 384 well format
- Sort specific population(s) of cells
- High capture efficiency
- Experimental design considerations
- Full-length cDNA (SMART-seq(2) or endtagging; UMIs)
- Spike-in control RNA
- **High gene coverage**

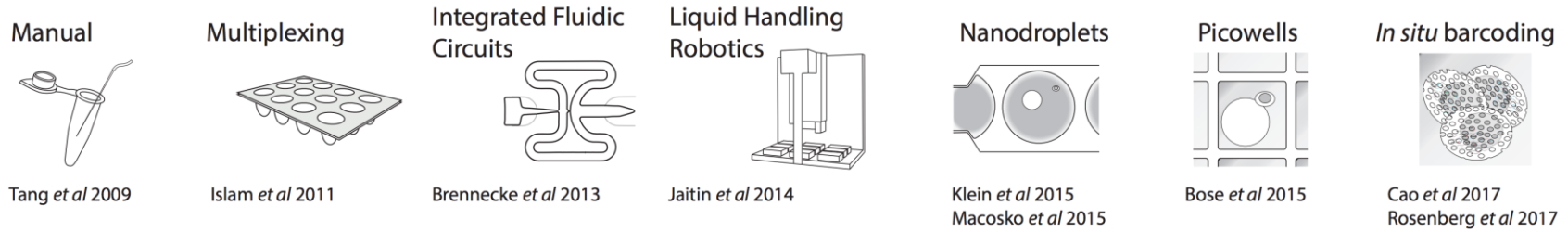
Droplet-based

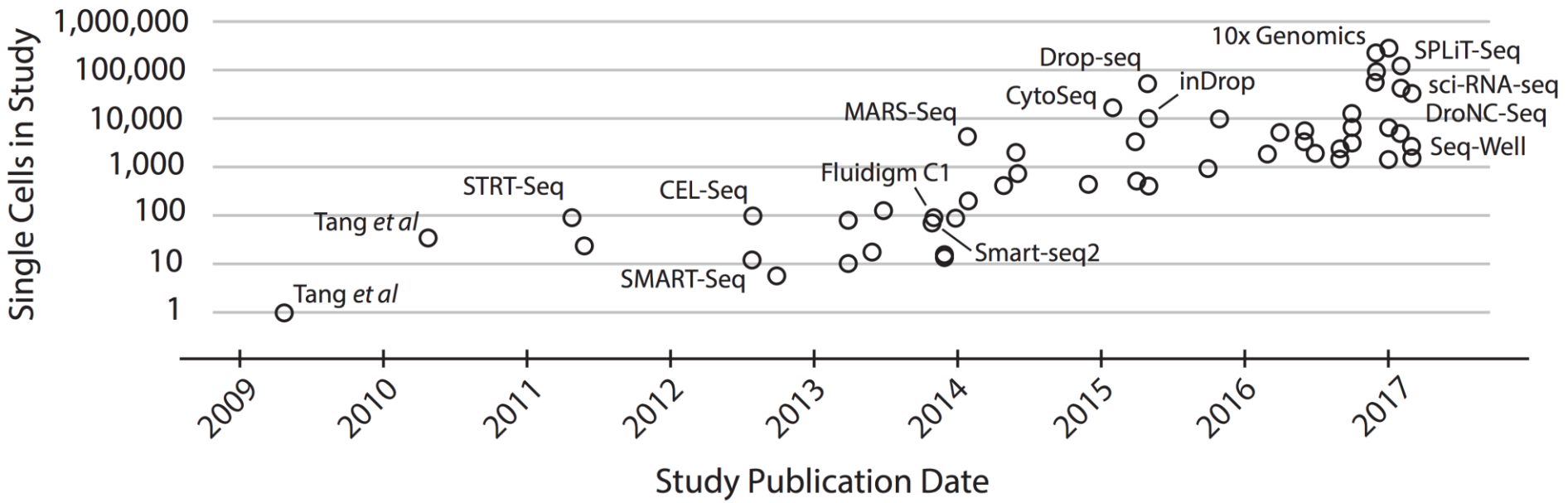


- 100-1000's of cells
- Doublet issues
- Variable capture efficiency
- Low per-cell cost
- 3' end tag; UMIs
- No spike-in control RNA
- **High cell coverage**

scRNA-seq Protocols

Number of cells

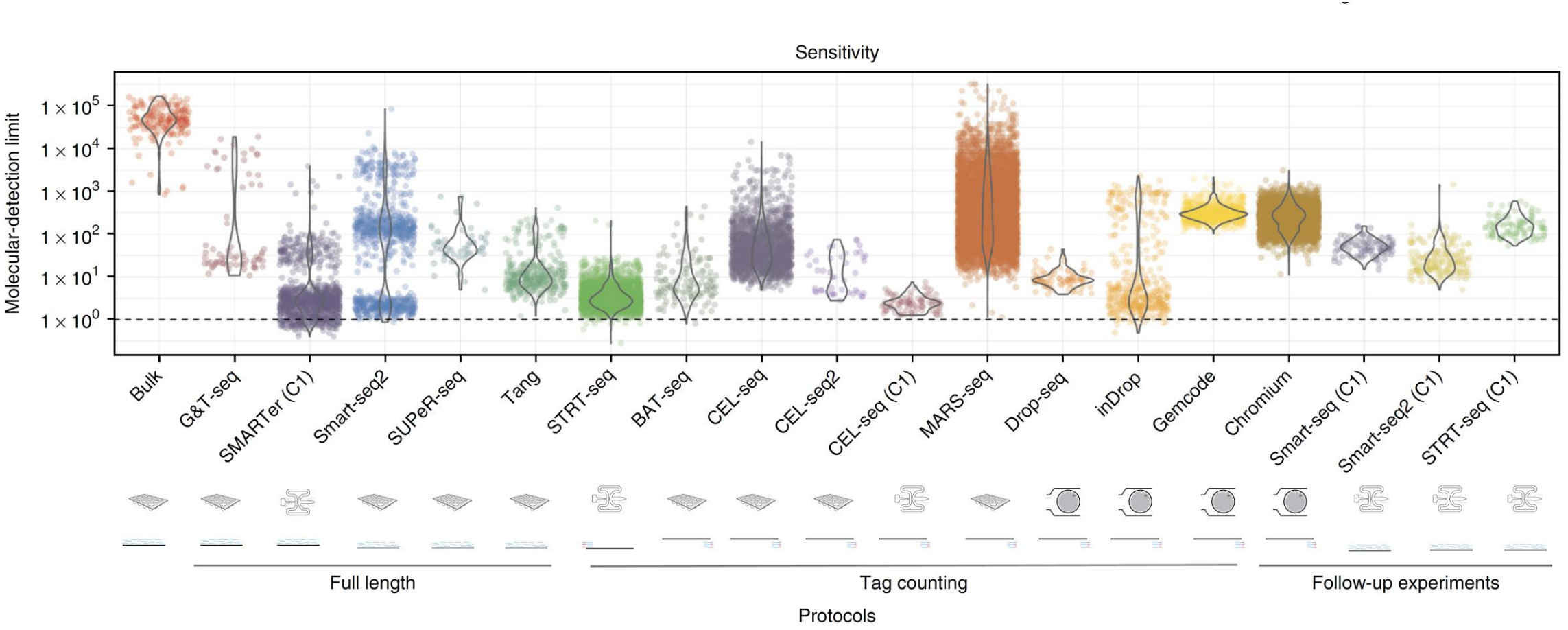




Svensson et al. Nature Methods 2017

scRNA-seq Protocols

Sensitivity



Svensson et al. Nature Methods 2017

10X: Massive Parallel Sequencing

- 1.3 million brain cells from E18 mice

10X GENOMICS™

Cell Ranger · neuron_aggregation ·

SUMMARY

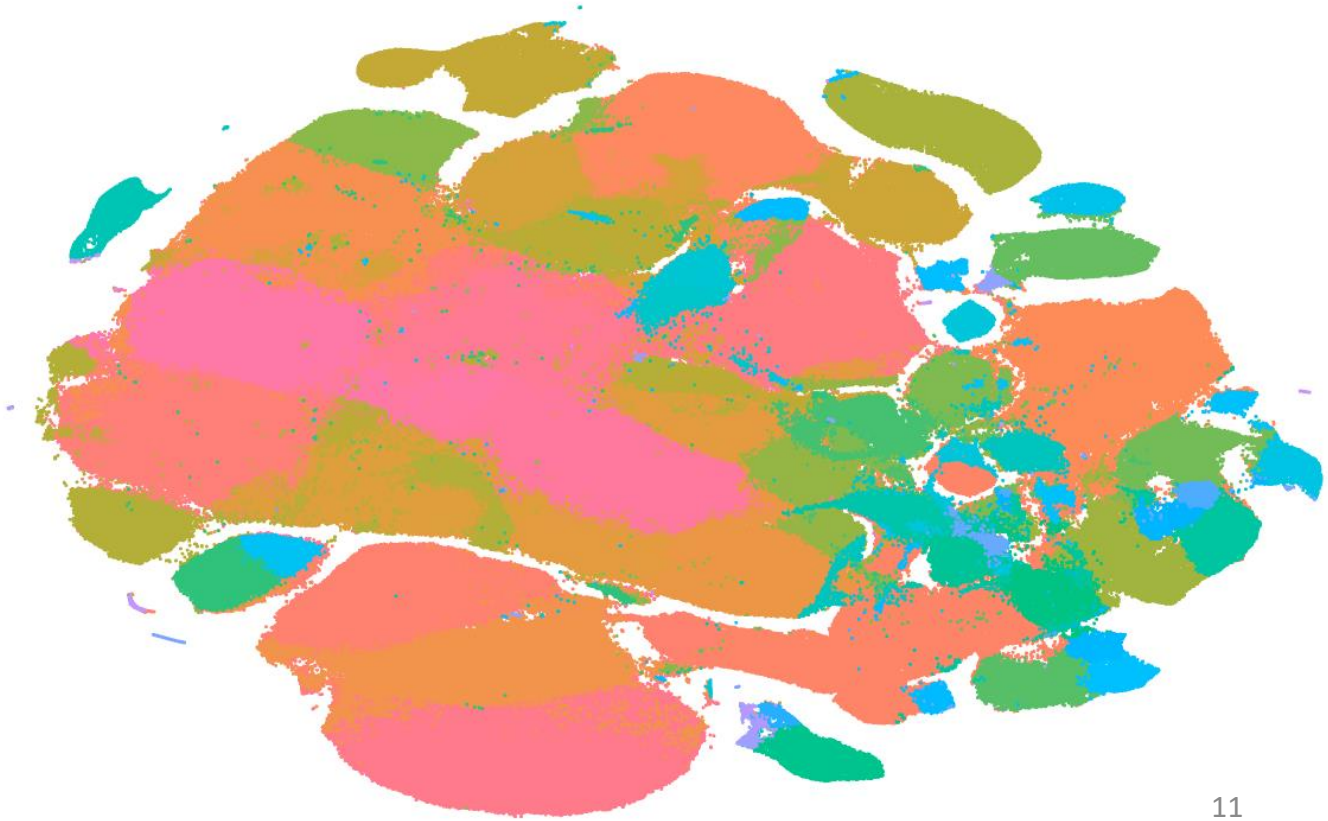
Estimated Number of Cells
1,306,127

Post-Normalization Mean Reads per Cell
18,402

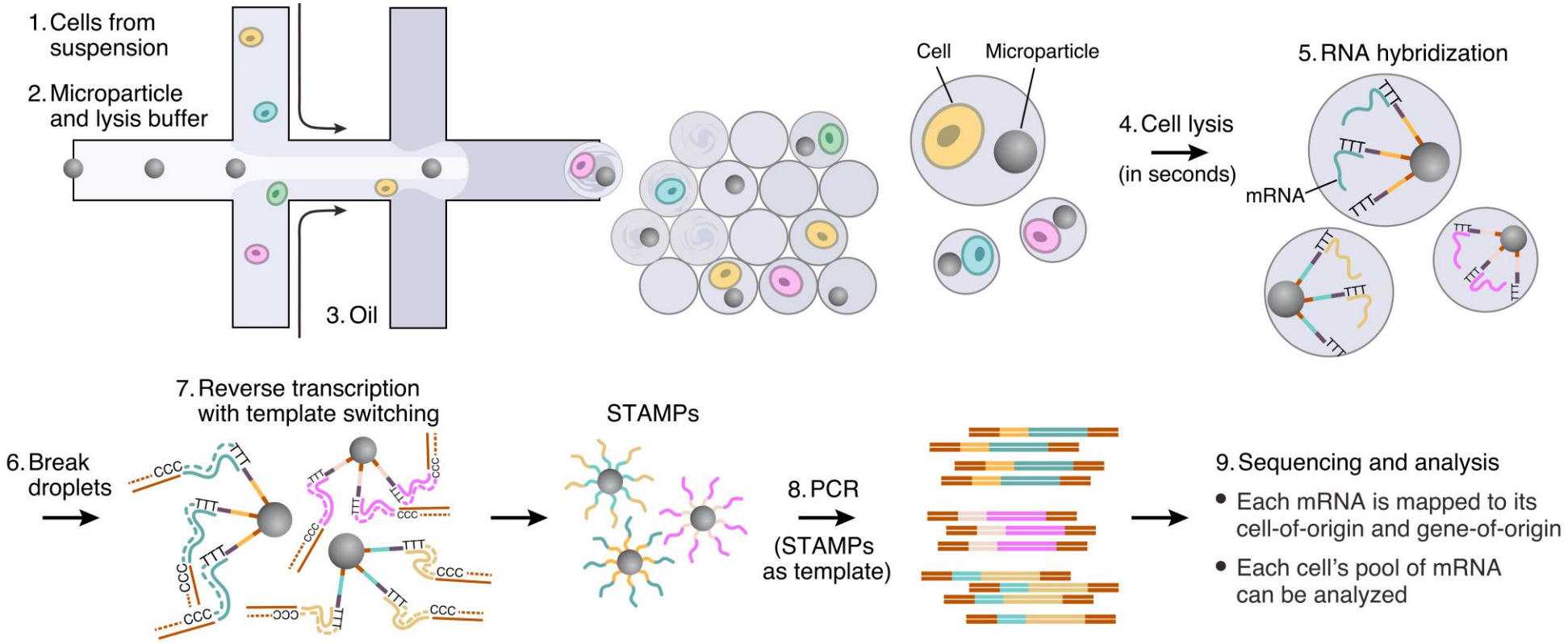
Median Genes per Cell
1,927

Sequencing

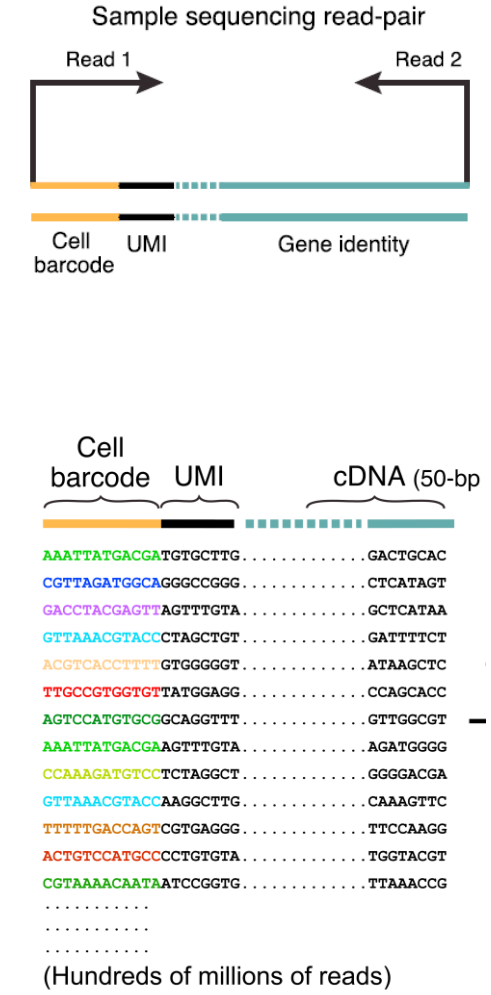
Pre-Normalization Number of Reads	30,282,984,606
Post-Normalization Number of Reads	24,036,055,221



Drop-seq



STAMPs: Single-cell transcriptomes attached to microparticles

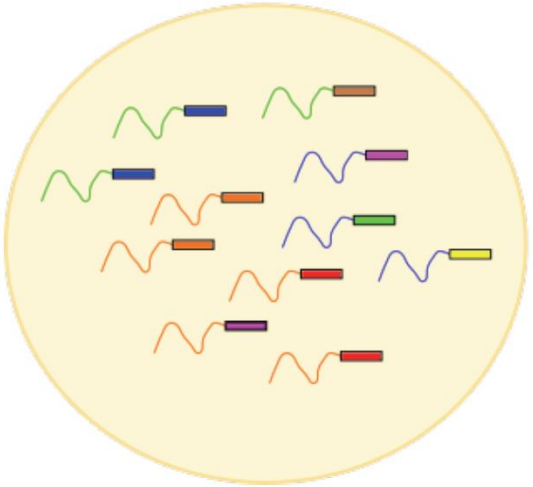


Macosko et al., Cell 2015

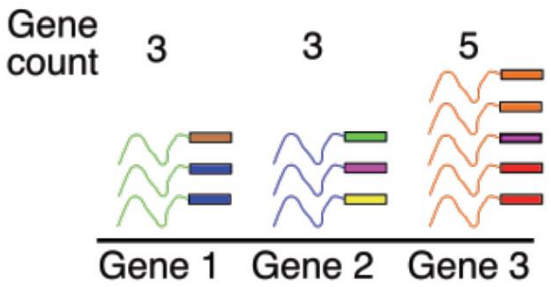
Unique Molecular Identifiers (UMIs)

- Unique molecular identifiers give (almost) exact molecule counts in sequencing experiments.
- They reduce the amplification noise by allowing (almost) complete de-duplication of sequenced fragments.

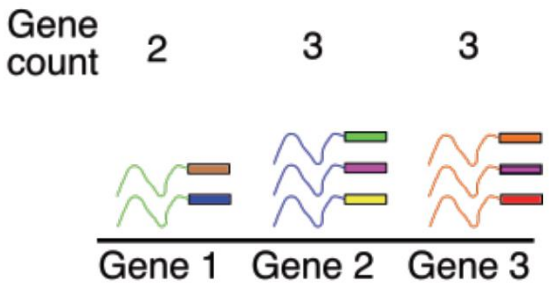
Sequenced fragments from an individual cell



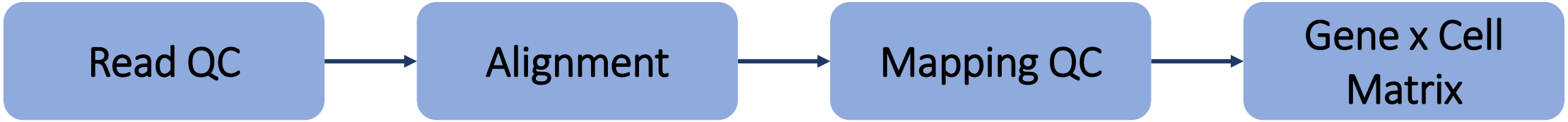
Pre de-duplication



Post de-duplication



“Typical” SMART-seq2 workflow

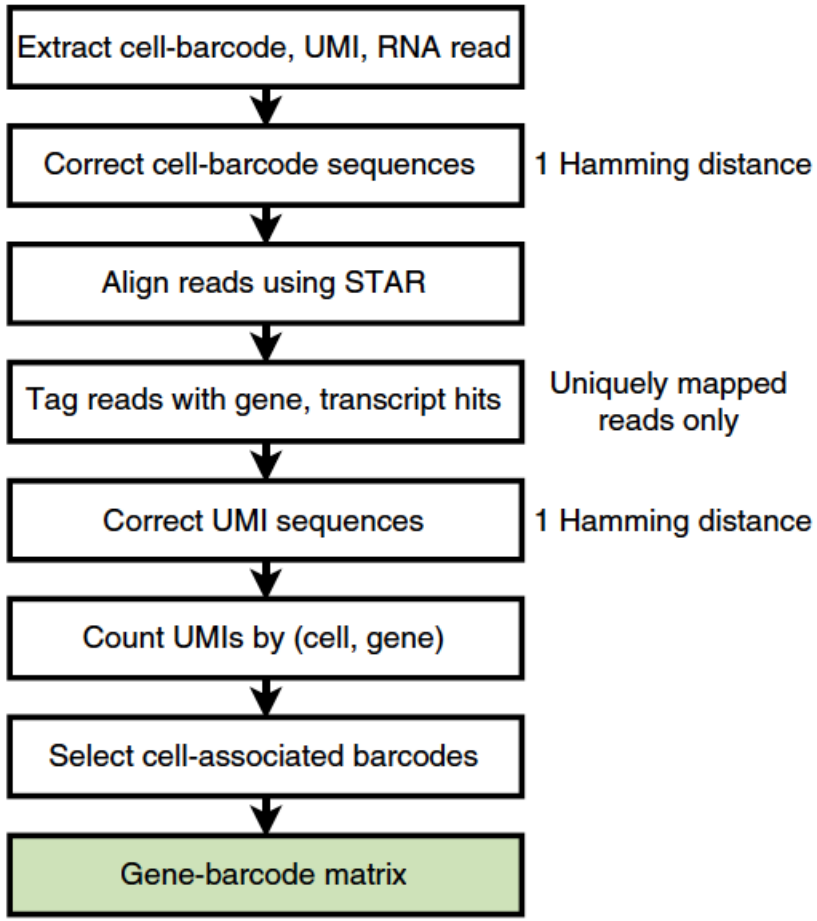


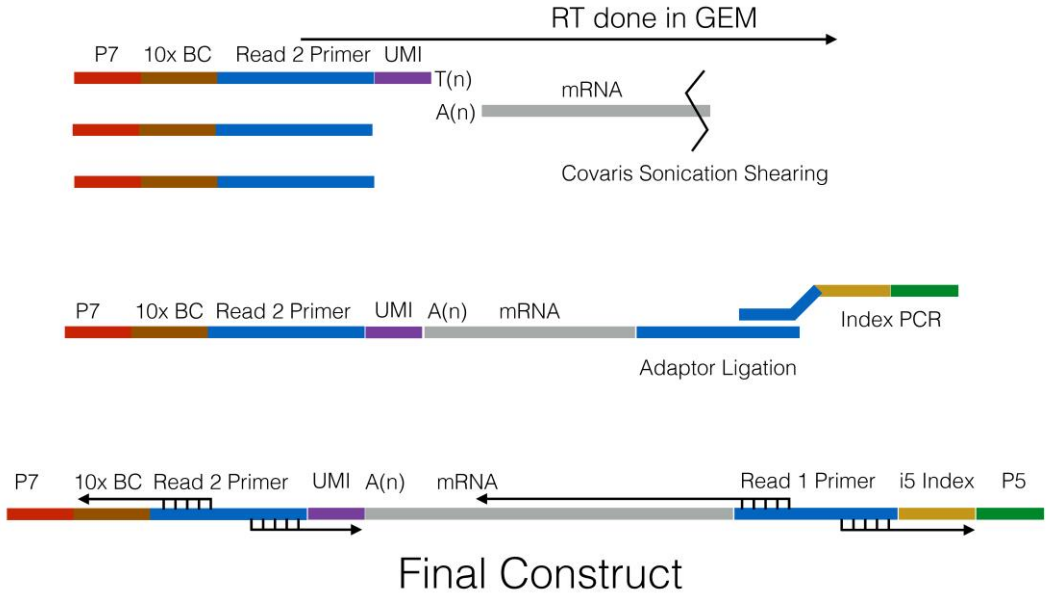
- Typically 1 library per cell, potentially many 100’s of FASTQ files
- The same tools used for bulk RNA-seq, e.g. FastQC, Star, PicardTools
- Deduplication is essential

“Typical” droplet workflow

- Prohibitively expensive to sequence 20,000 cells to 1M reads -> multiplex cells using barcodes
- 1000’s of *small* FASTQ files
- Generally run in a single pipeline, e.g. Cellranger (10X specific), DropSeq (Macosko et al.), or general (https://salmon.readthedocs.io/en/latest/alevin.html)
- Sequencing errors in cell barcodes and UMIs are a source of technical noise

Construction of the expression matrix

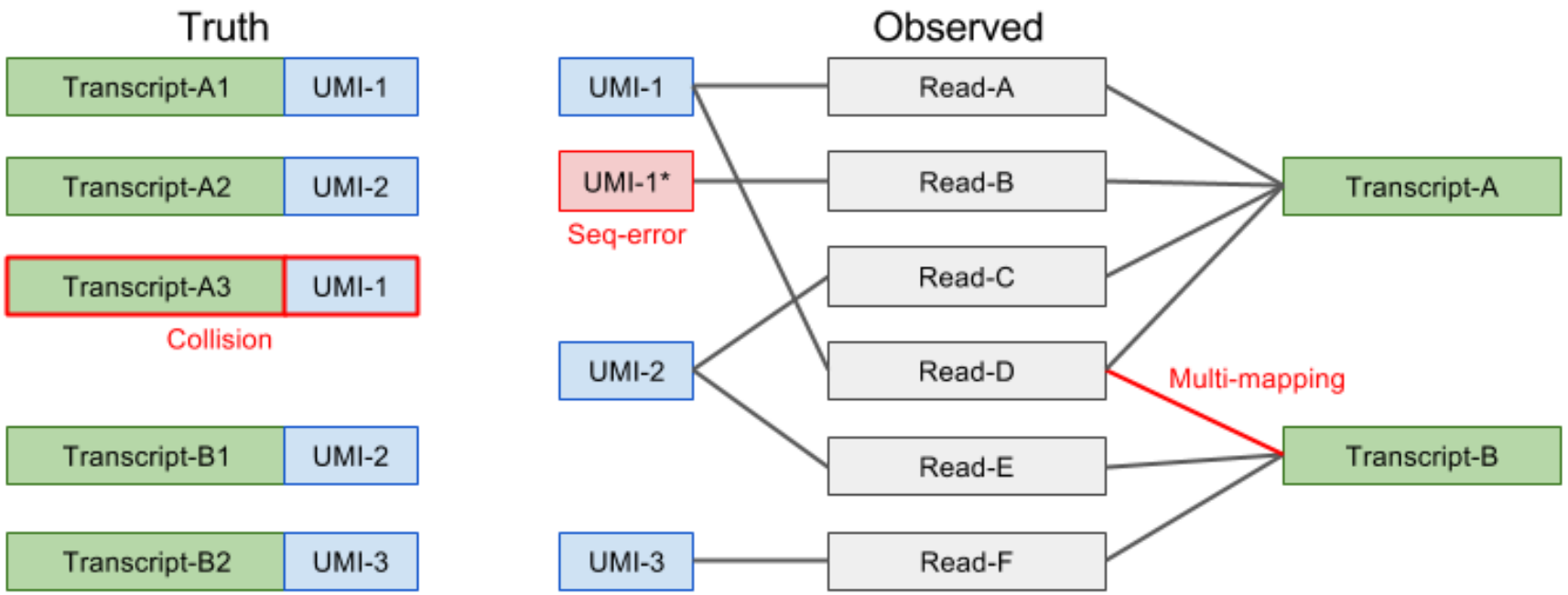




10x BC: 14bp to index GEMs
UMI: 10bp to index mRNA molecules

Zheng et al. (Nat. Comm. 2017)

De-duplication



- [UMI-tools]
- [PoissonUMIs]
- [zUMIs]
- [dropEst]

- Different UMI does not necessarily mean different molecule (PCR or sequencing errors)
- Different transcript does not necessarily mean different molecule (multiple mappings)

scRNA-seq Data Analysis

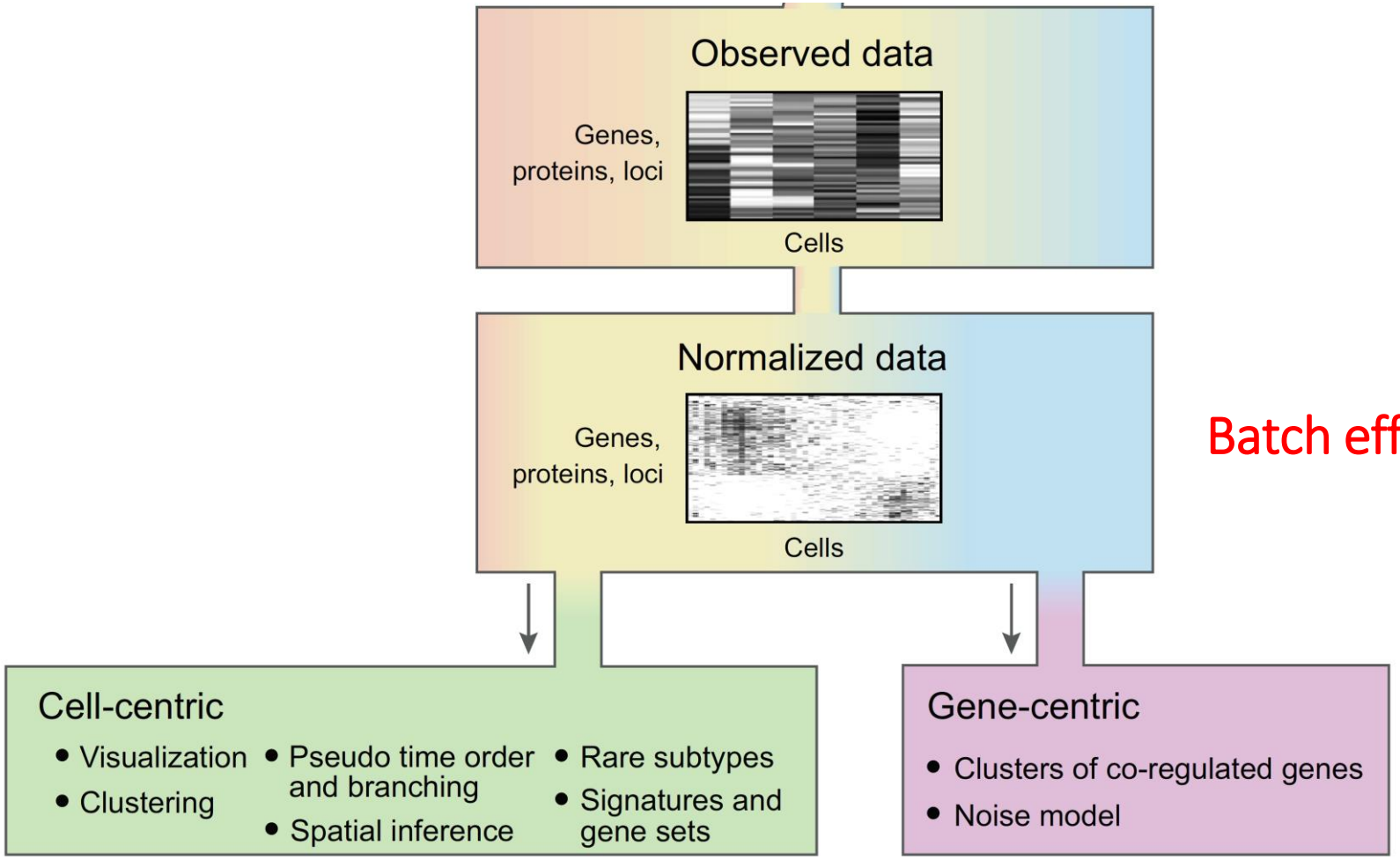
Our goal is to derive/extract real biology from technically noisy data.

What is different from bulk?

- The main sources of discrepancy between the libraries are:
  - **Amplification** (up to 1 million fold)
  - **Gene 'dropouts'**

- In both cases the discrepancies are introduced due to low starting amounts of transcripts since the RNA comes from one cell only.

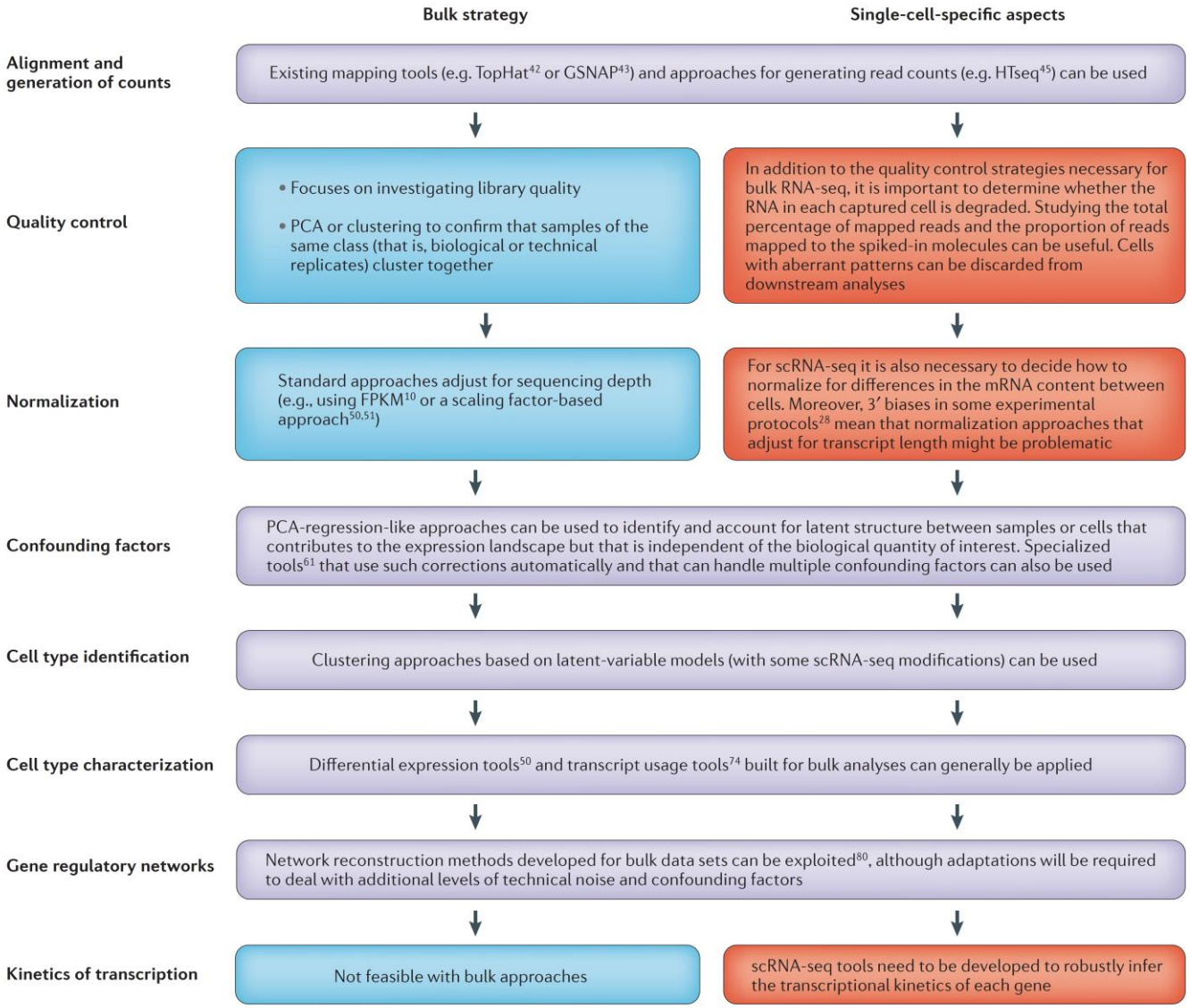
scRNA-seq Data Analysis



Batch effects?

Wagner et al. (Nat Biotech 2016)

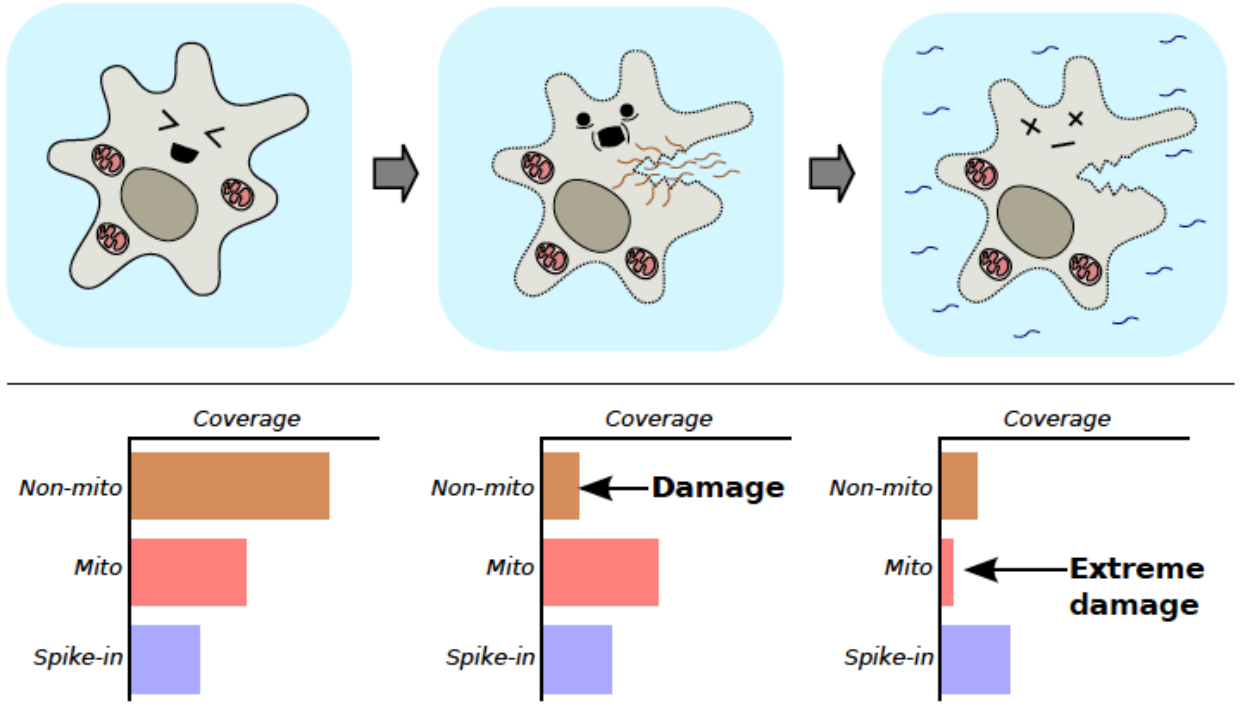
Bulk vs scRNA-seq



Stegle, Teichmann, Marioni (Nat Rev Genetics 2015)

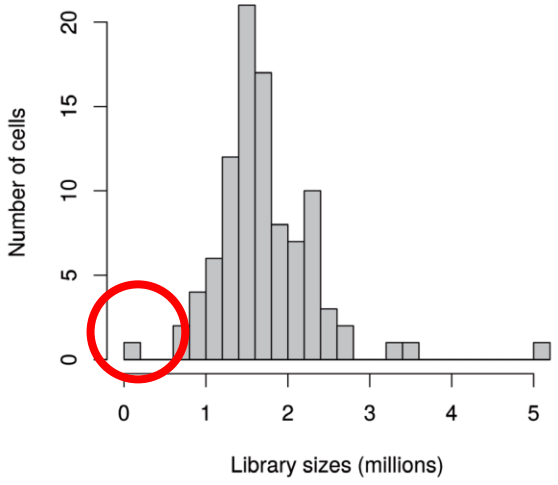
Quality control on cells (1)

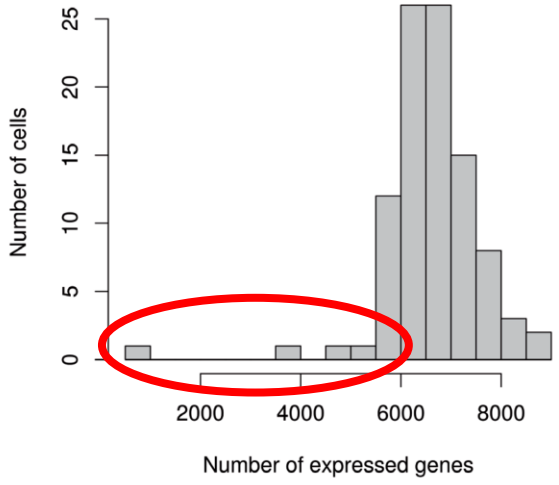
- Low sequencing depth
- Low numbers of expressed genes (i.e. any nonzero count)
- High spike-in (if present) or mitochondrial content



Quality control on cells (2)

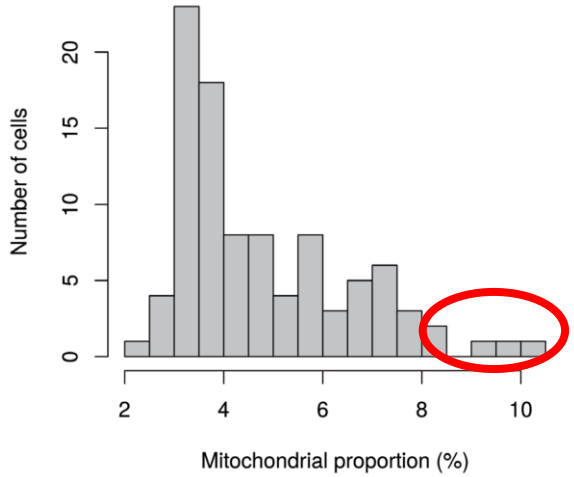
RNA has not been efficiently captured during library preparation

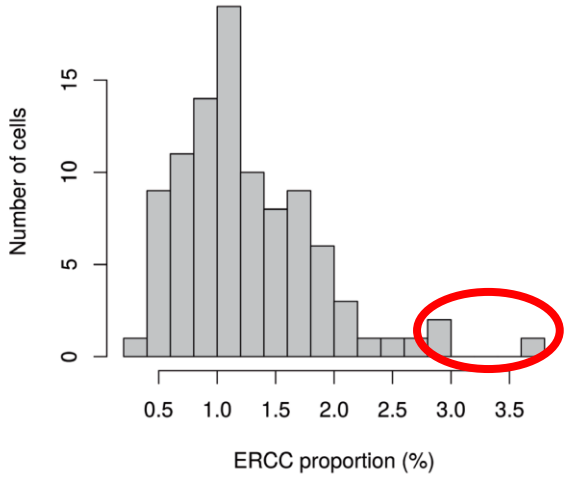




Diverse transcript population not captured

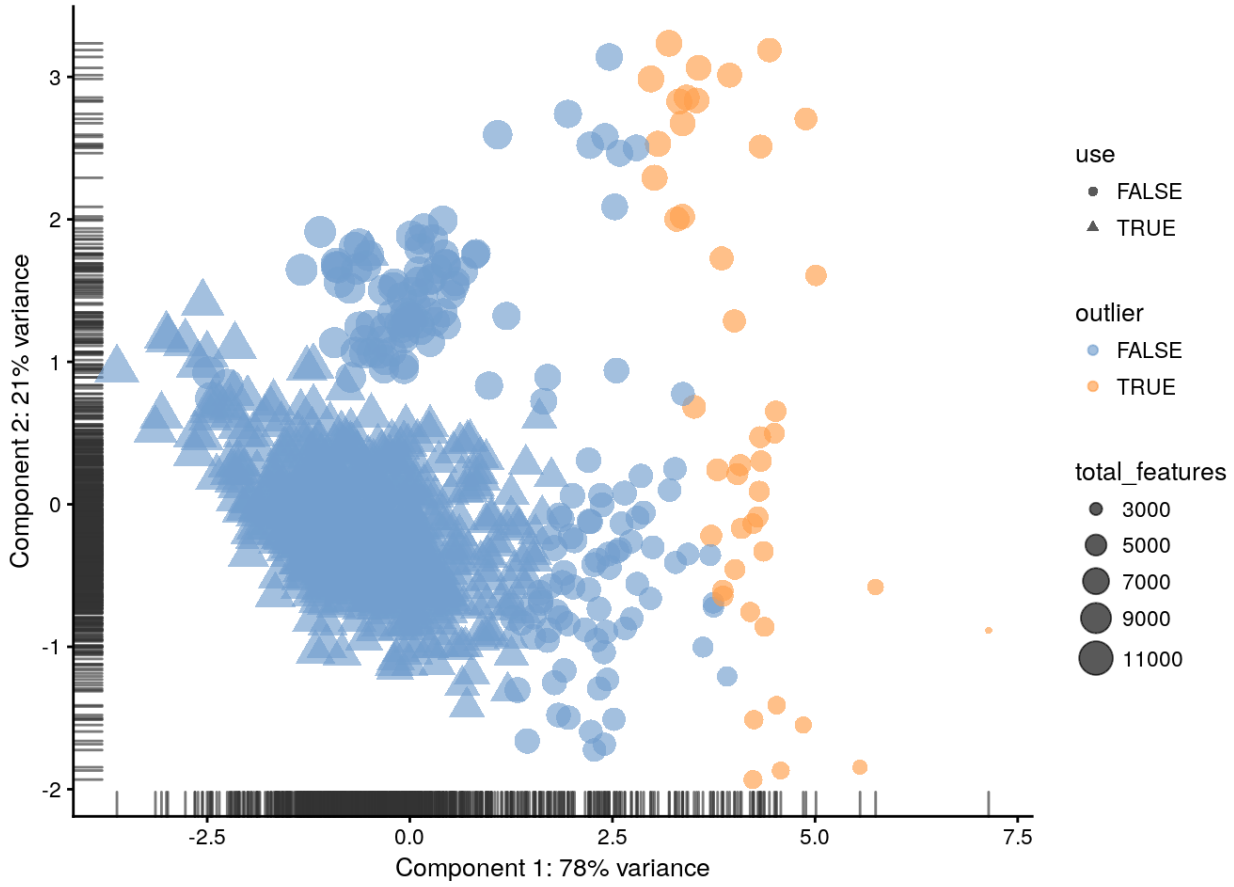
Possibly because of increased apoptosis and/or loss of cytoplasmic RNA from lysed cells





Quality control on cells (3)

PCA on a set of QC metrics

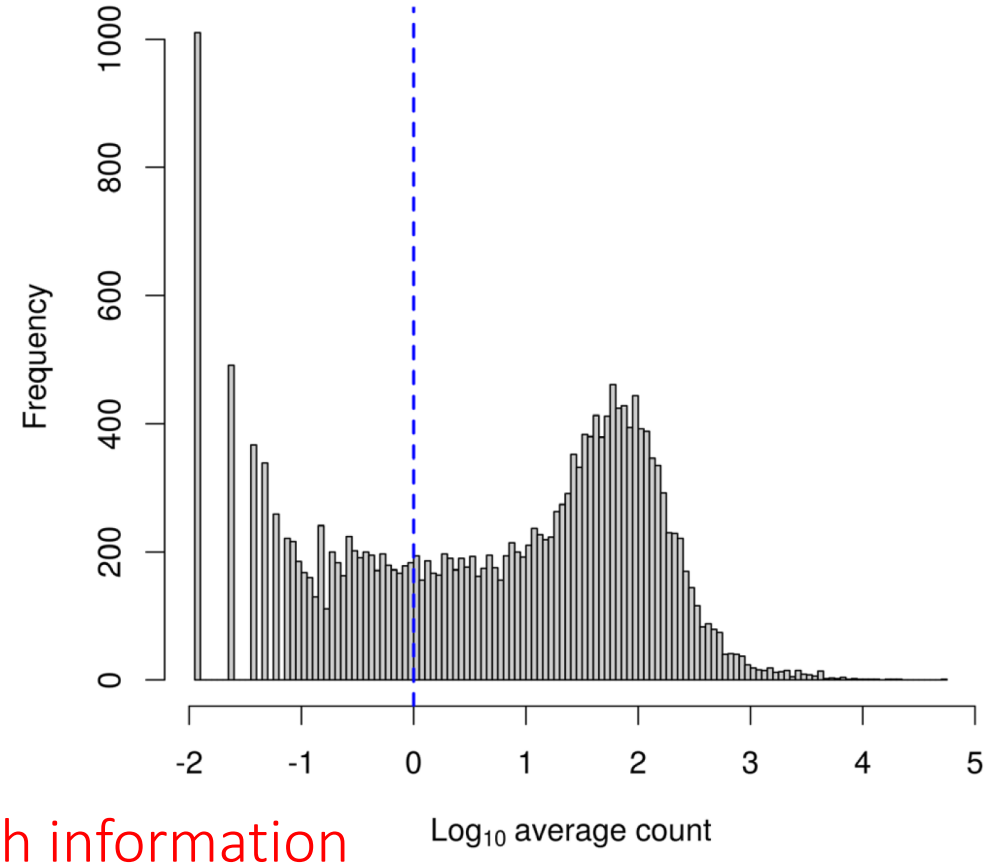


Possible features

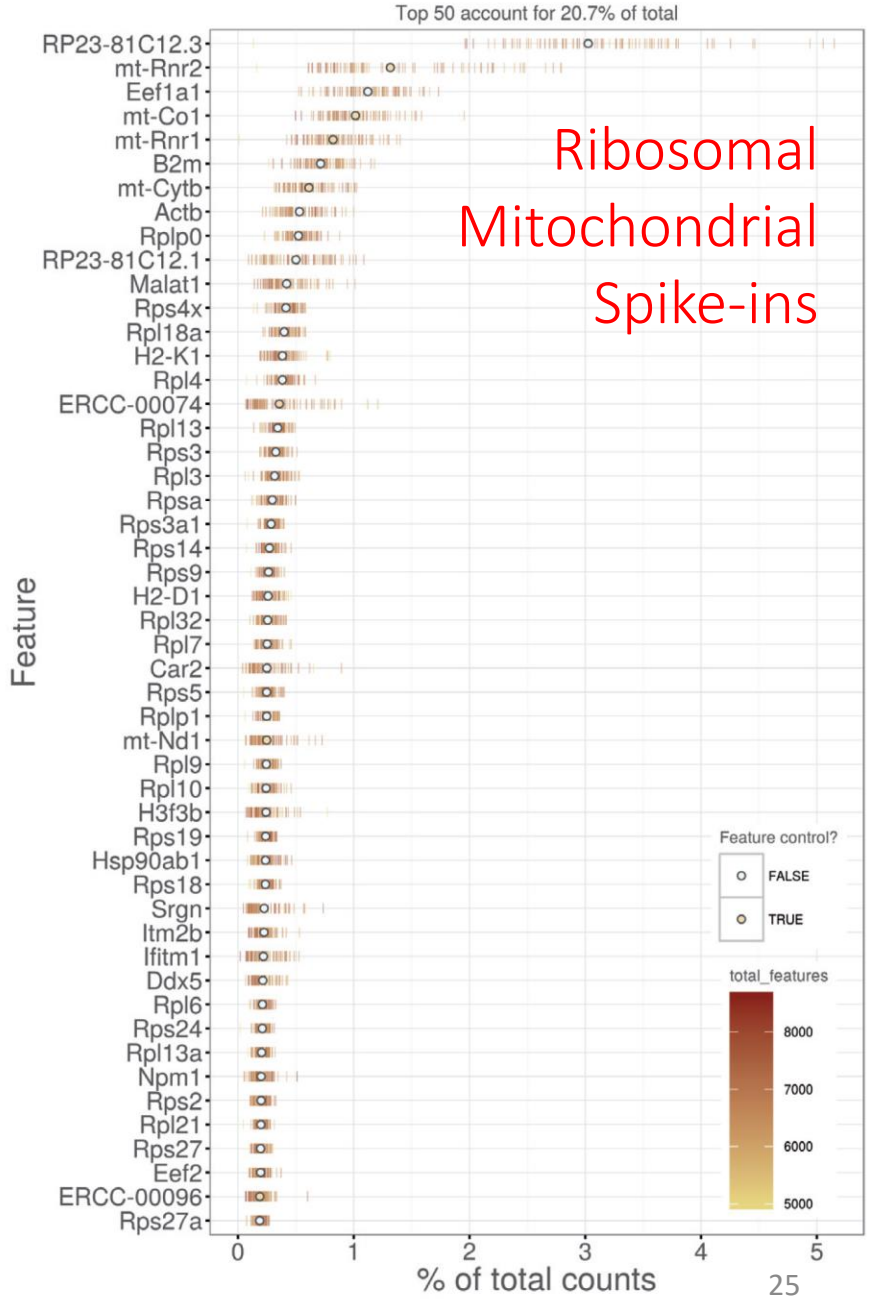
- total number of reads
- total number of features
- proportion of mitochondrial reads
- …

Interpretation!

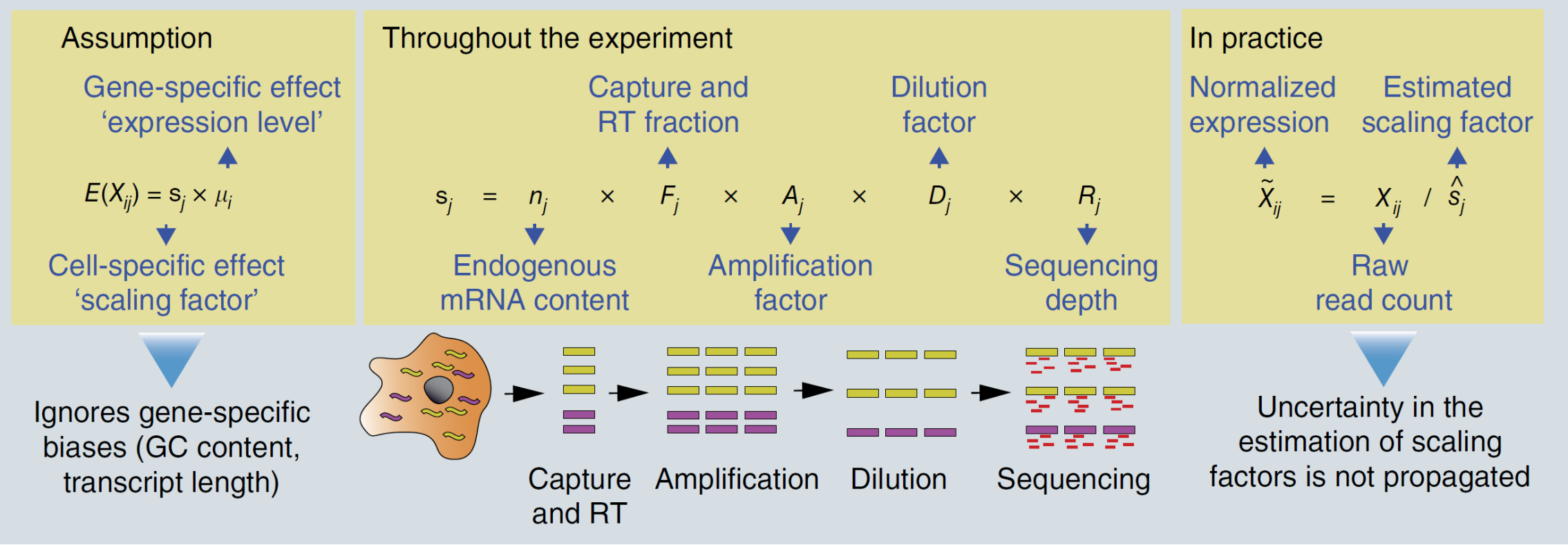
Quality control on Genes



Not enough information for reliable statistical inference



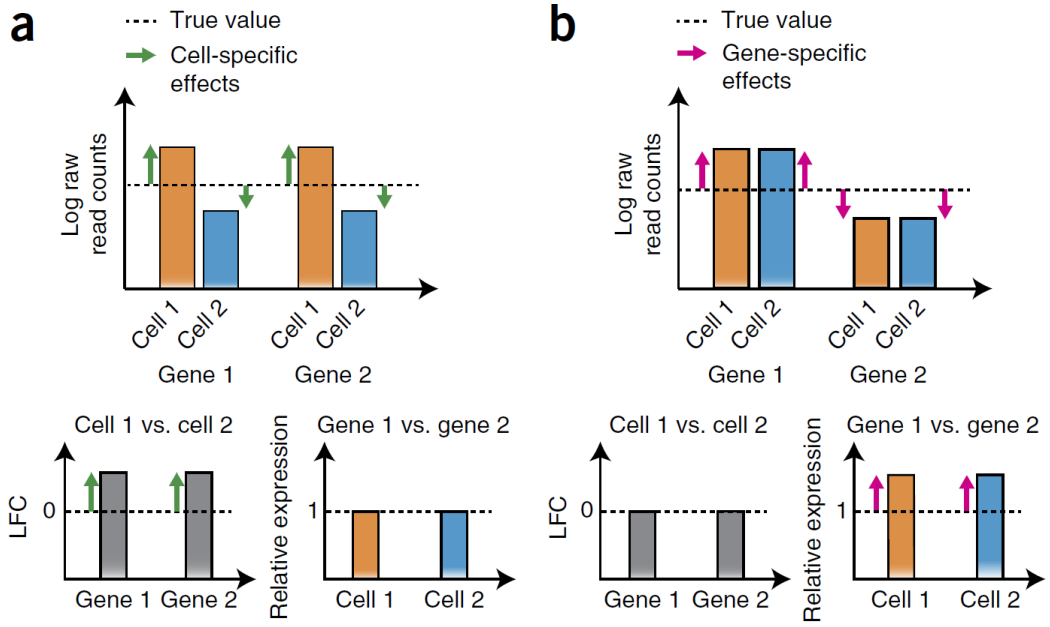
Normalization (1)



Vallejos et al. (Nat Methods 2017)

Normalization (2)

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



c

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

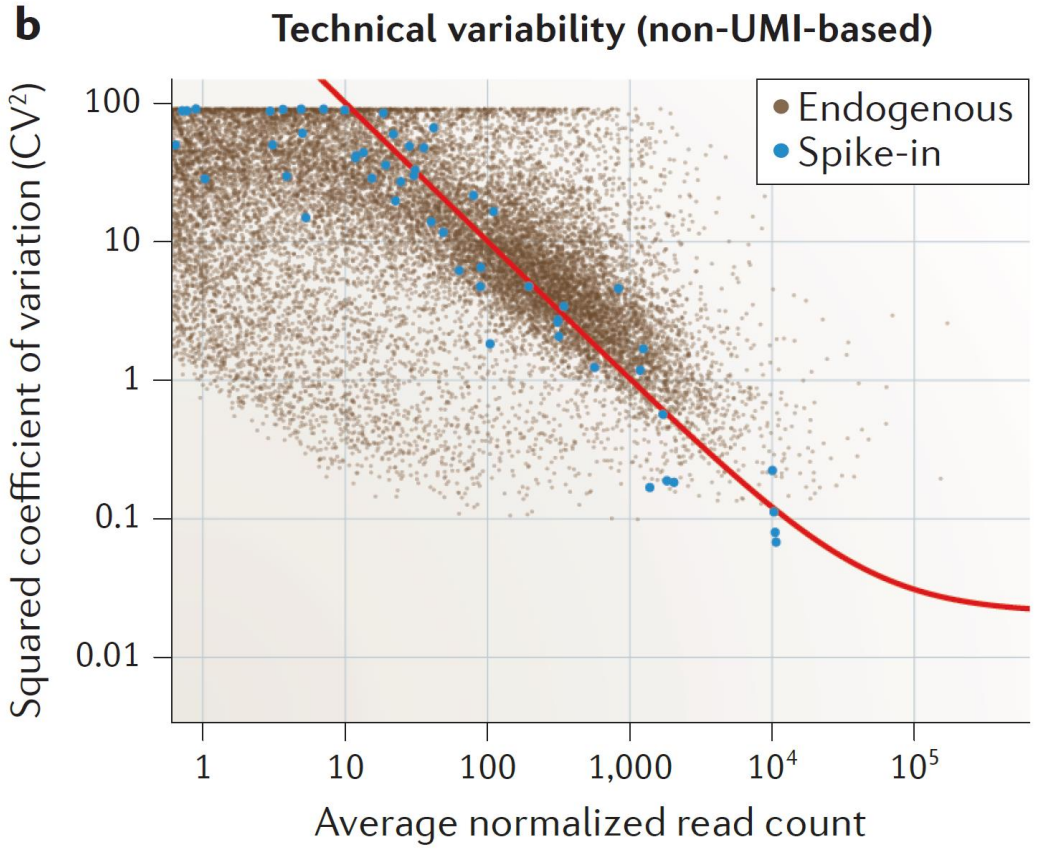
Vallejos et al. (Nat Methods 2017)

Normalization (3)

Using spike-ins

Caveats:

- The same quantity of spike-in RNA may not be consistently added to each sample
- Synthetic spike-in transcripts may not behave in the same manner as endogenous transcripts
- Not easily incorporated in all scRNA-seq protocols



Brennecke et al. (Nat Methods 2013)

Normalization (4)

To spike in or not to spike in?

Assessing the reliability of spike-in normalization for analyses of single-cell RNA sequencing data

Aaron T.L. Lun,[1] Fernando J. Calero-Nieto,[2] Liora Haim-Vilmovsky,[3,4] Berthold Göttgens,[2] and John C. Marioni[1,3,4]

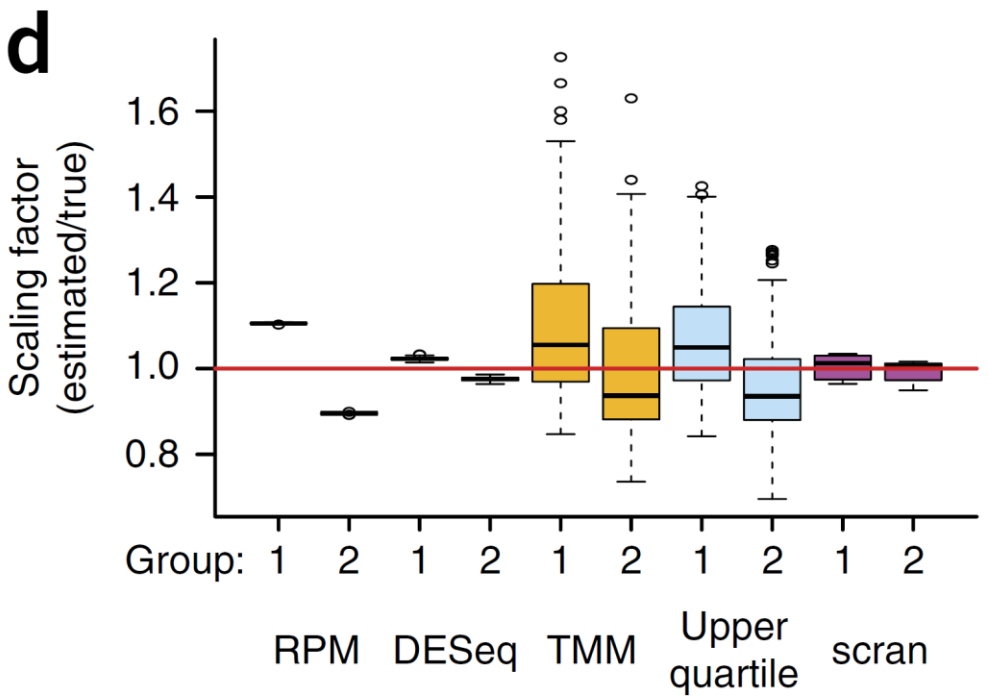
[1]*Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom;* [2]*Wellcome Trust and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0XY, United Kingdom;* [3]*EMBL European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, United Kingdom;* [4]*Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, United Kingdom*

By profiling the transcriptomes of individual cells, single-cell RNA sequencing provides unparalleled resolution to study cellular heterogeneity. However, this comes at the cost of high technical noise, including cell-specific biases in capture efficiency and library generation. One strategy for removing these biases is to add a constant amount of spike-in RNA to each cell and to scale the observed expression values so that the coverage of spike-in transcripts is constant across cells. This approach has previously been criticized as its accuracy depends on the precise addition of spike-in RNA to each sample. Here, we perform mixture experiments using two different sets of spike-in RNA to quantify the variance in the amount of spike-in RNA added to each well in a plate-based protocol. We also obtain an upper bound on the variance due to differences in behavior between the two spike-in sets. We demonstrate that both factors are small contributors to the total technical variance and have only minor effects on downstream analyses, such as detection of highly variable genes and clustering. Our results suggest that scaling normalization using spike-in transcripts is reliable enough for routine use in single-cell RNA sequencing data analyses.

Genome Research 27:1795–1806

Normalization (5)

- Bulk RNA-based methods: FPKM, CPM, TPM, upperquartile
- Log normalization (Seurat)
- Negative binomial (Monocole)
- Downsampling (RaceID)
- …

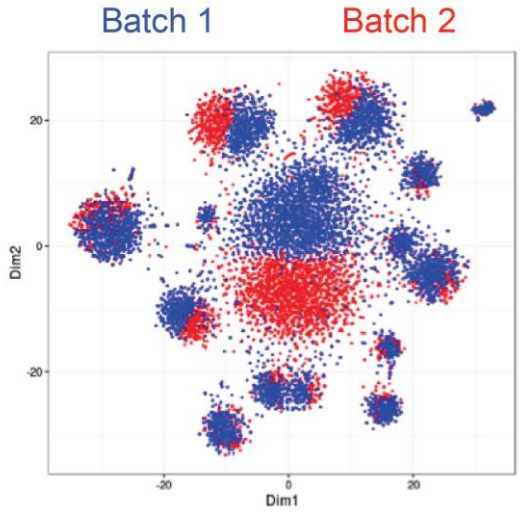


Vallejos et al. (Nat Methods 2017)

Confounders and batch effects (1)

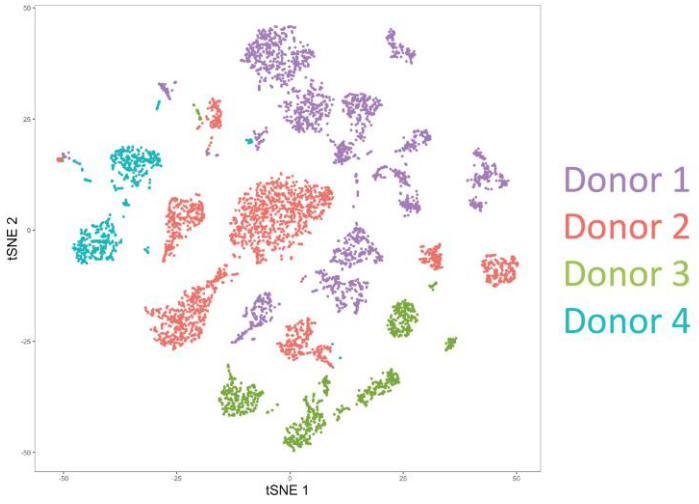
1. Technical variability
   - Changes in sample quality/processing
   - Library prep or sequencing technology
   - ‘Experimental reality’

Technical ‘batch effects’ confound downstream analysis



1. Biological variability
   - Patient differences
   - Environmental/genetic perturbation
   - Evolution! (cross-species analysis)

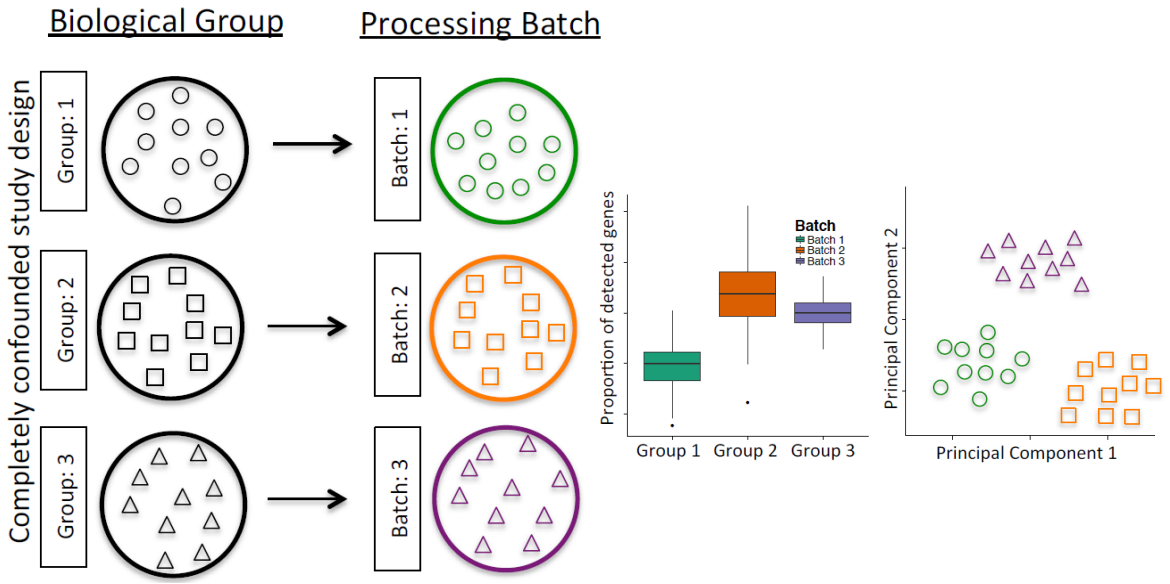
Biological ‘batch effects’ confound comparisons of scRNA-seq data



Shaham et al. (Bioinformatics 2017)

Confounders and batch effects (2)

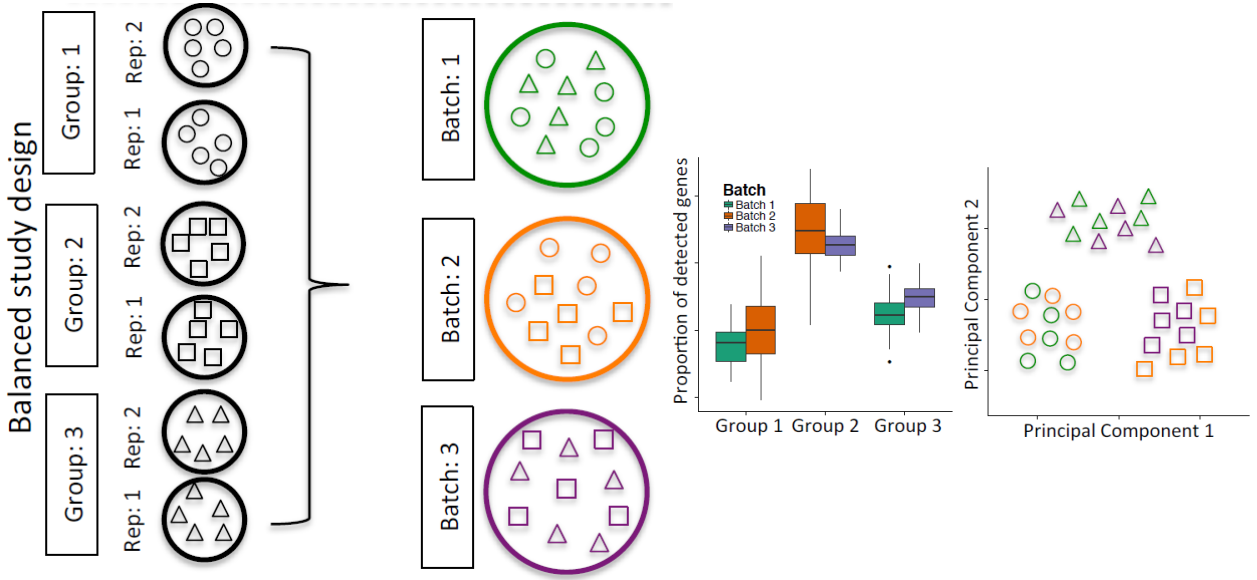
Confounded design



Don’t design your experiment like this!!!

Hicks et al. (bioRxiv 2015)

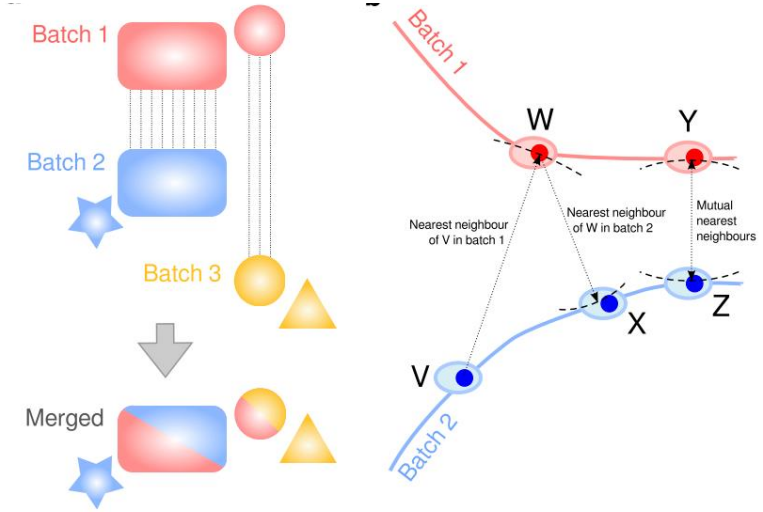
Not confounded design



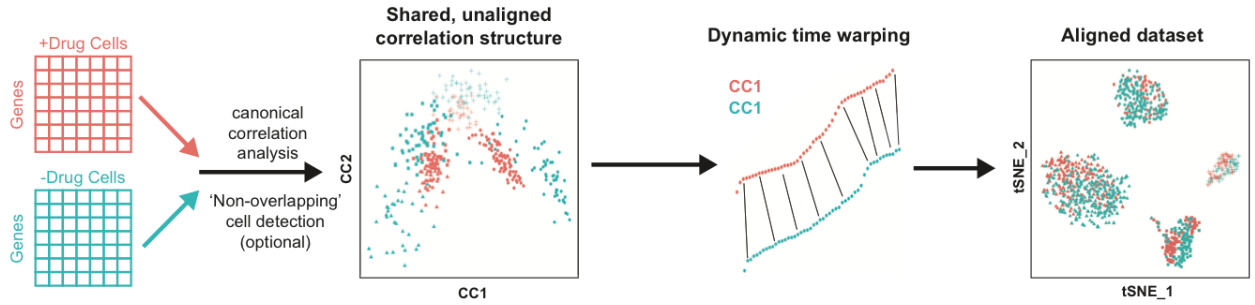
Good experimental design *does not remove batch effects*, it prevents them from biasing your results.

Confounders and batch effects (3)

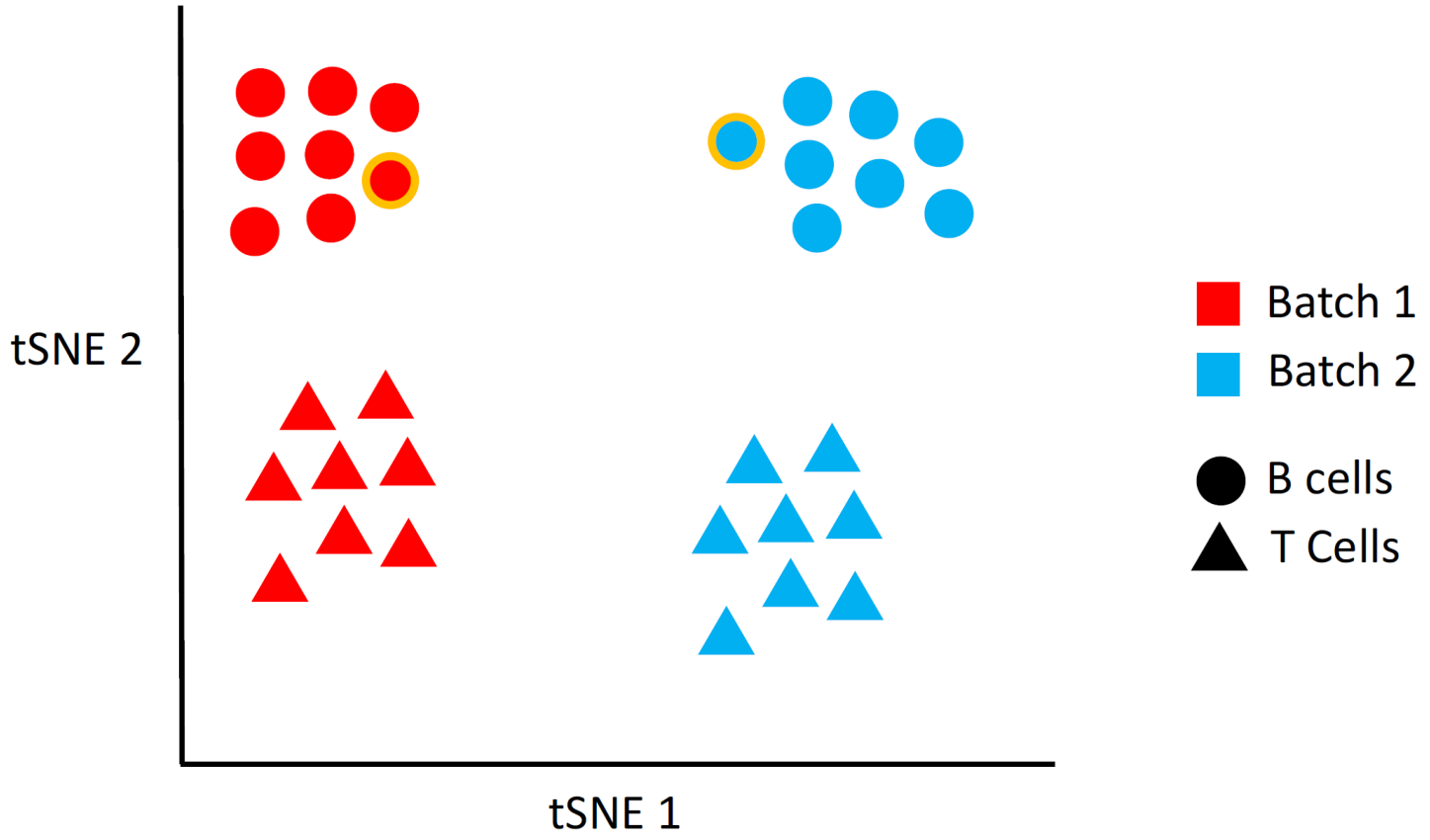
MNN-based batch correction
Haghverdi et al., (Nature Biotech 2017)



CCA-based batch correction (Seurat)
Butler et al., (Nature Biotech 2017)

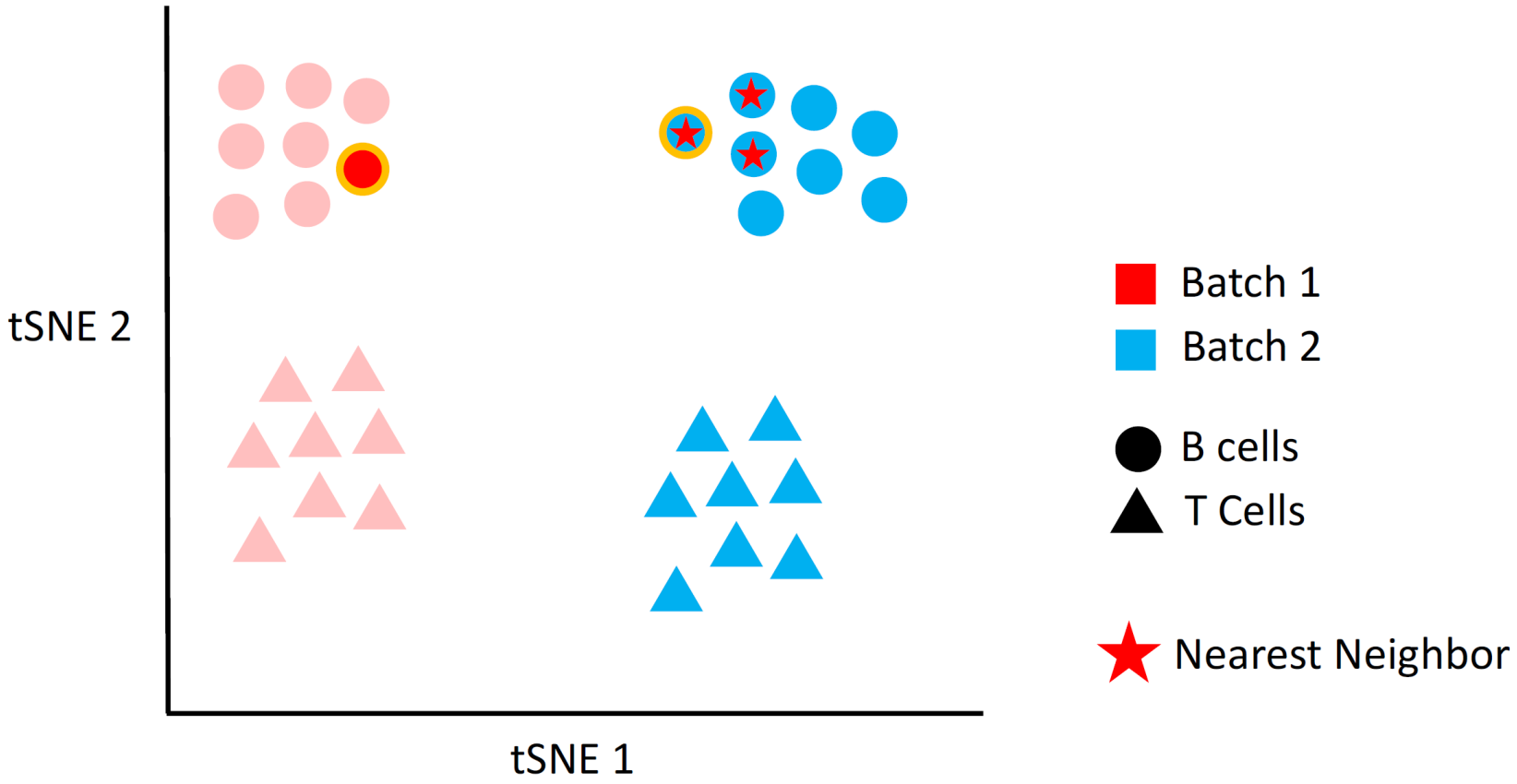


Confounders and batch effects (4)



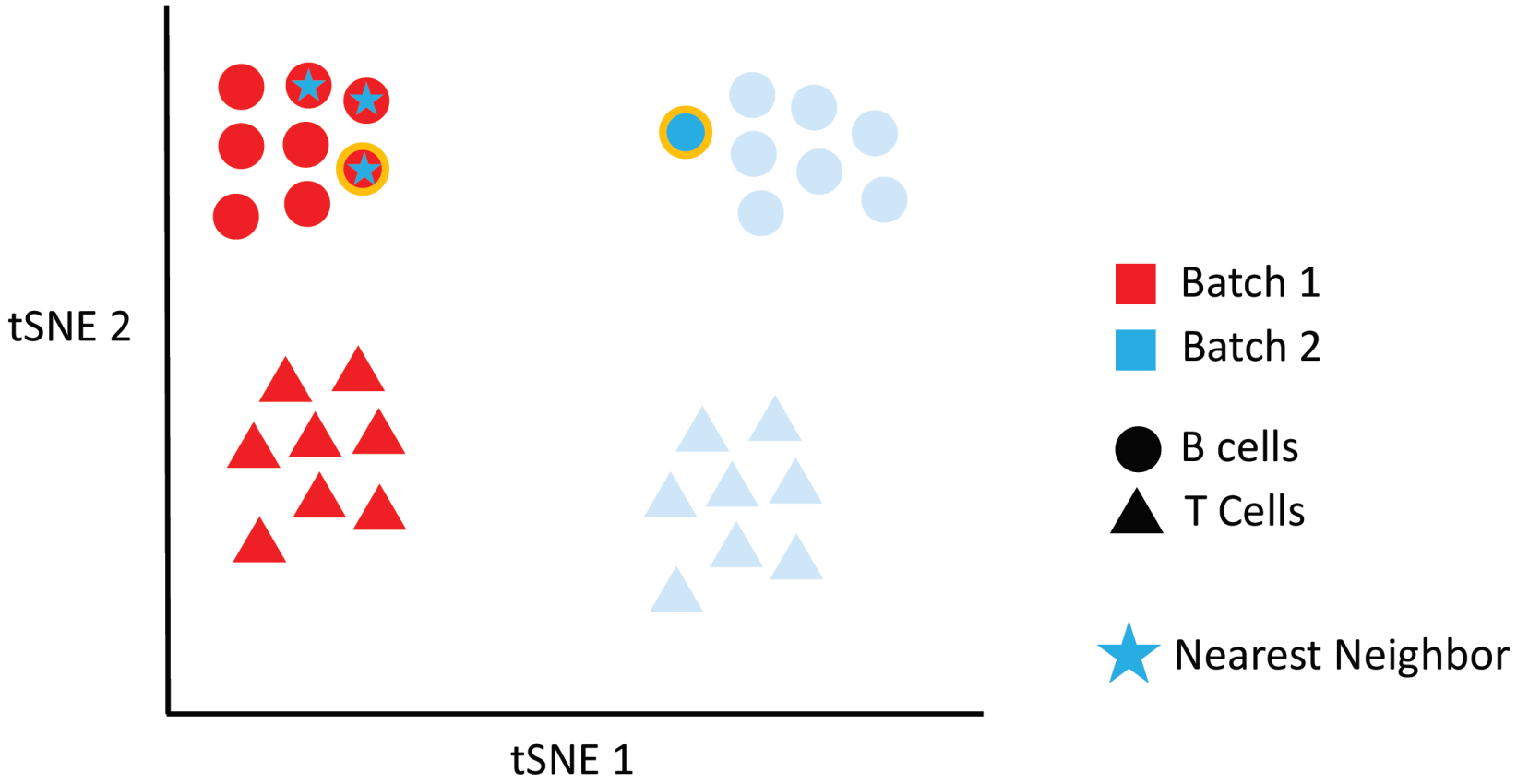
Haghverdi et al., (Nature Biotech 2017)

Confounders and batch effects (4)



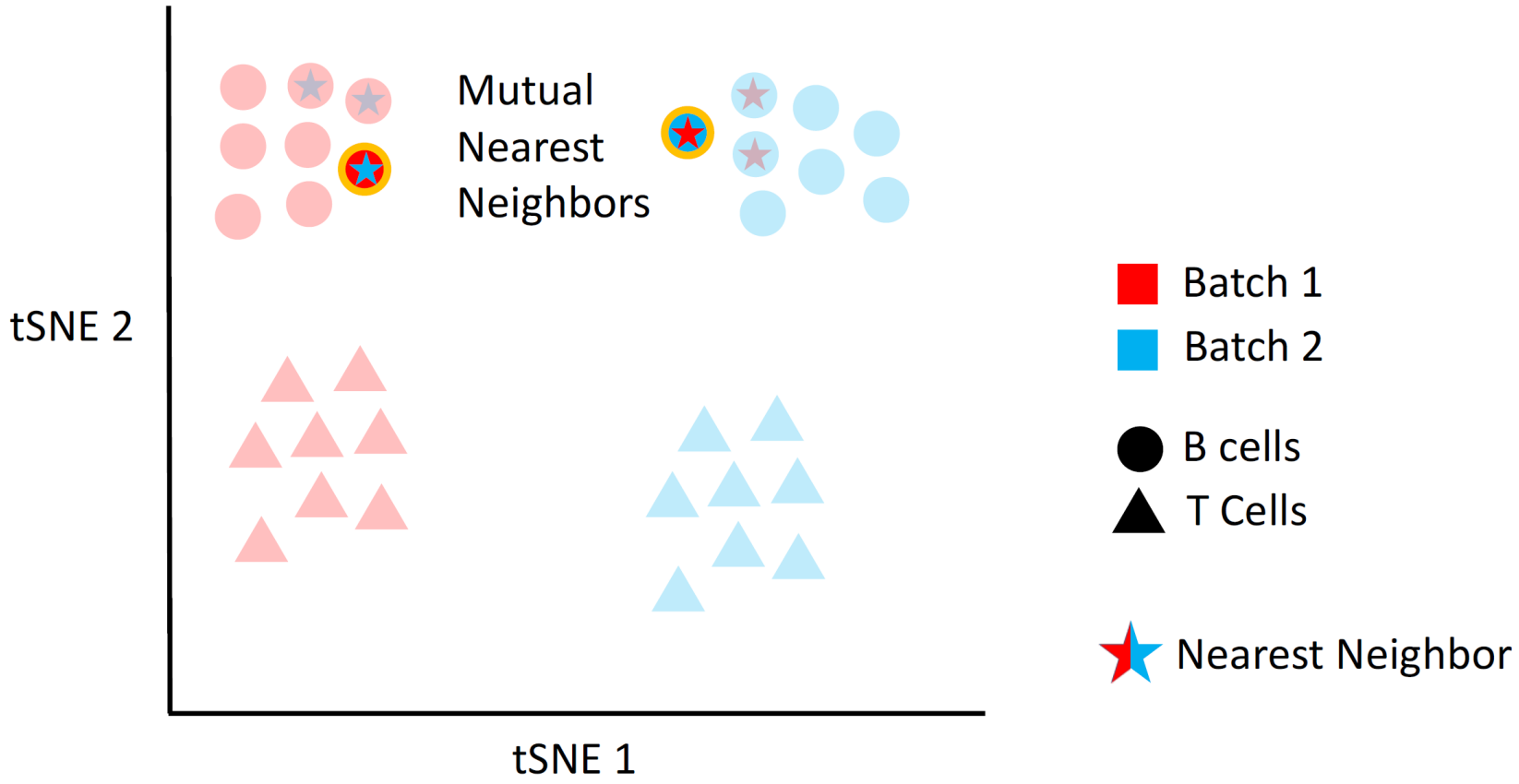
Haghverdi et al., (Nature Biotech 2017)

Confounders and batch effects (4)



Haghverdi et al., (Nature Biotech 2017)

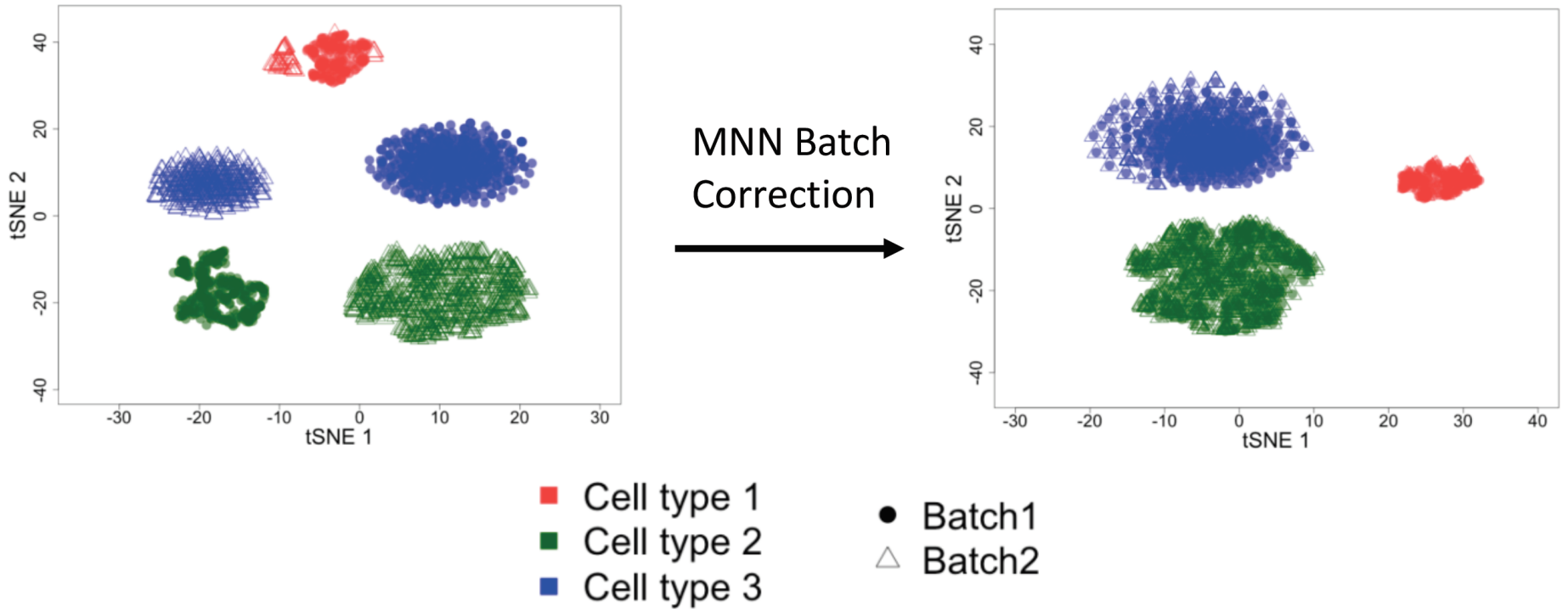
Confounders and batch effects (4)



Haghverdi et al., (Nature Biotech 2017)

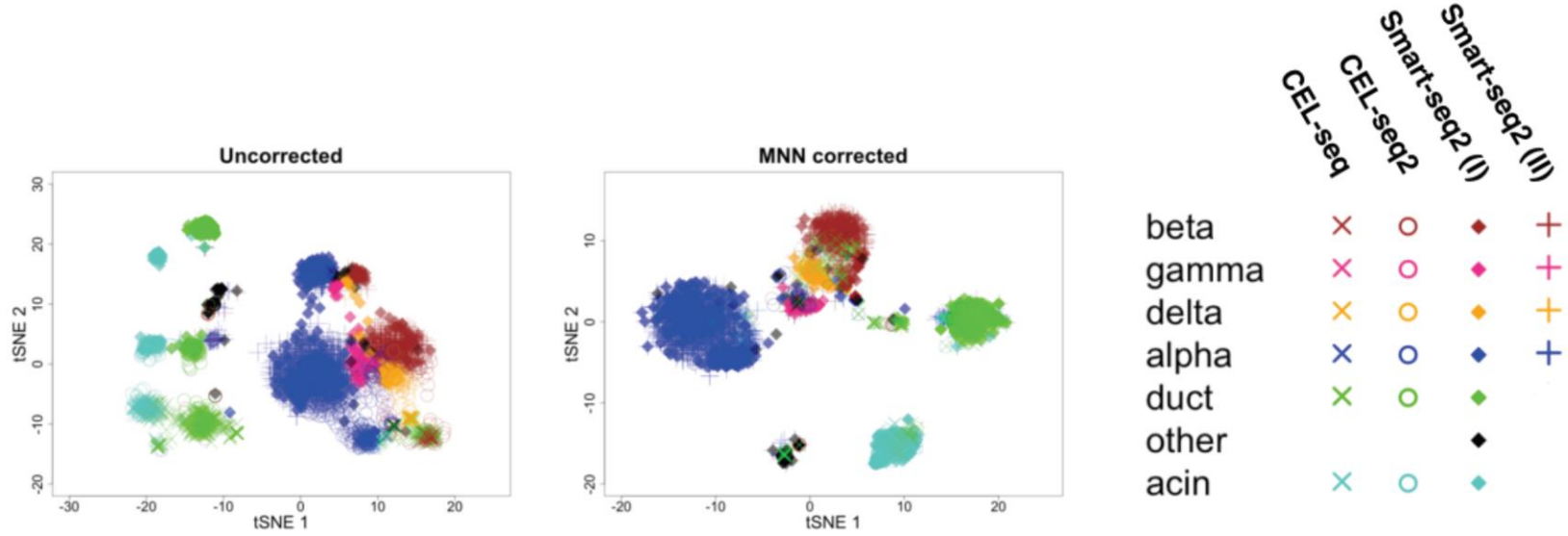
Confounders and batch effects (5)

- Enables computation of a cell-specific batch correction

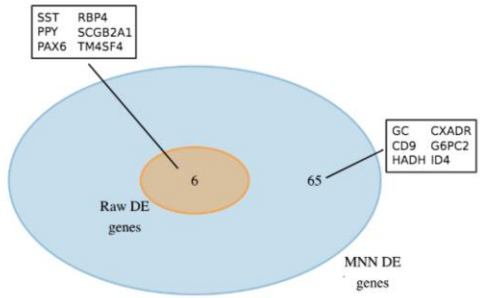


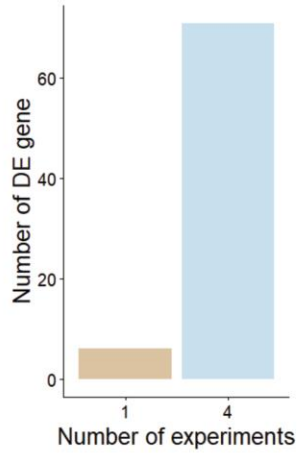
Haghverdi et al., (Nature Biotech 2017)

Confounders and batch effects (6)



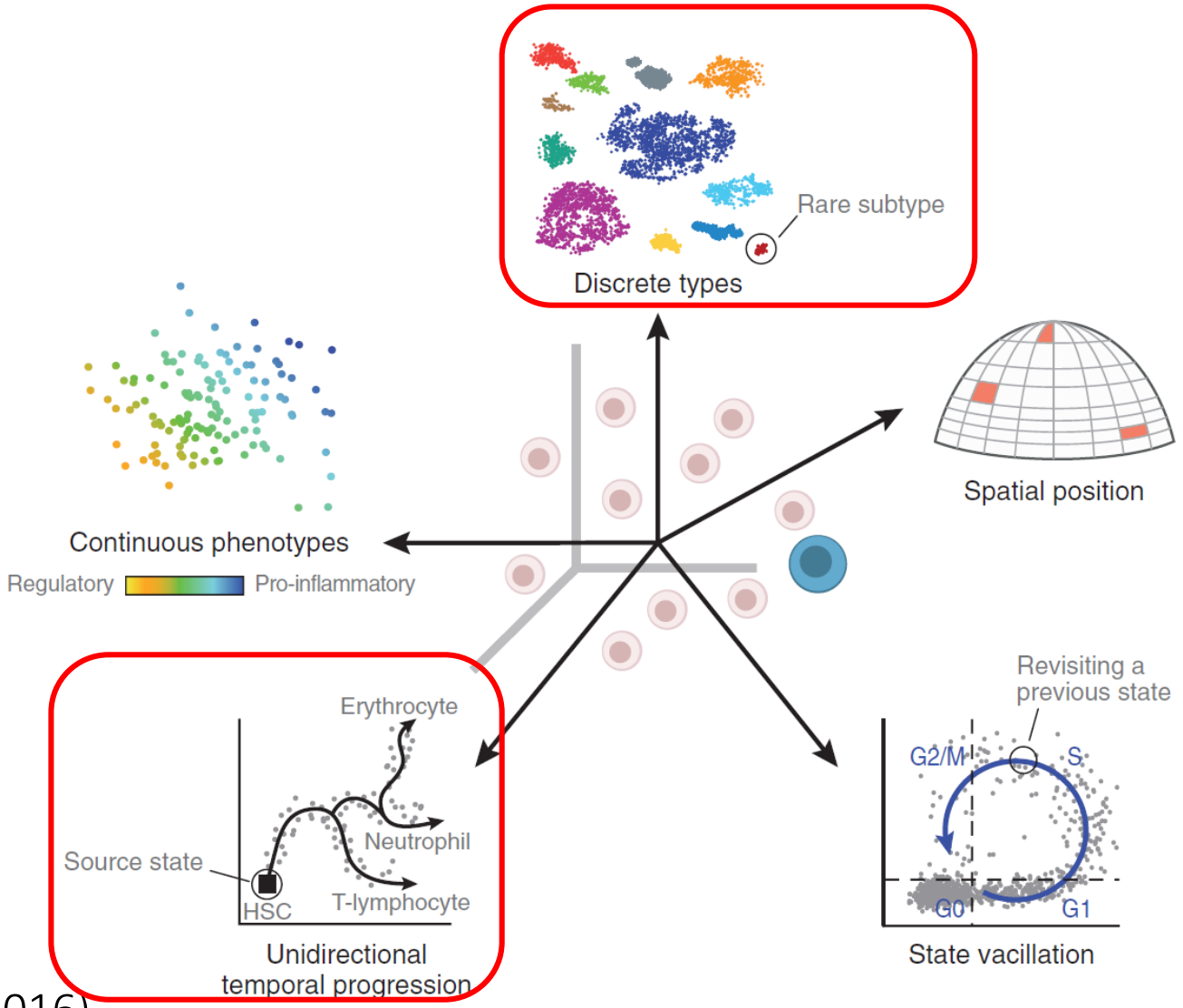
Delta vs Gamma Islet Cells





Haghverdi et al., (Nature Biotech 2017)

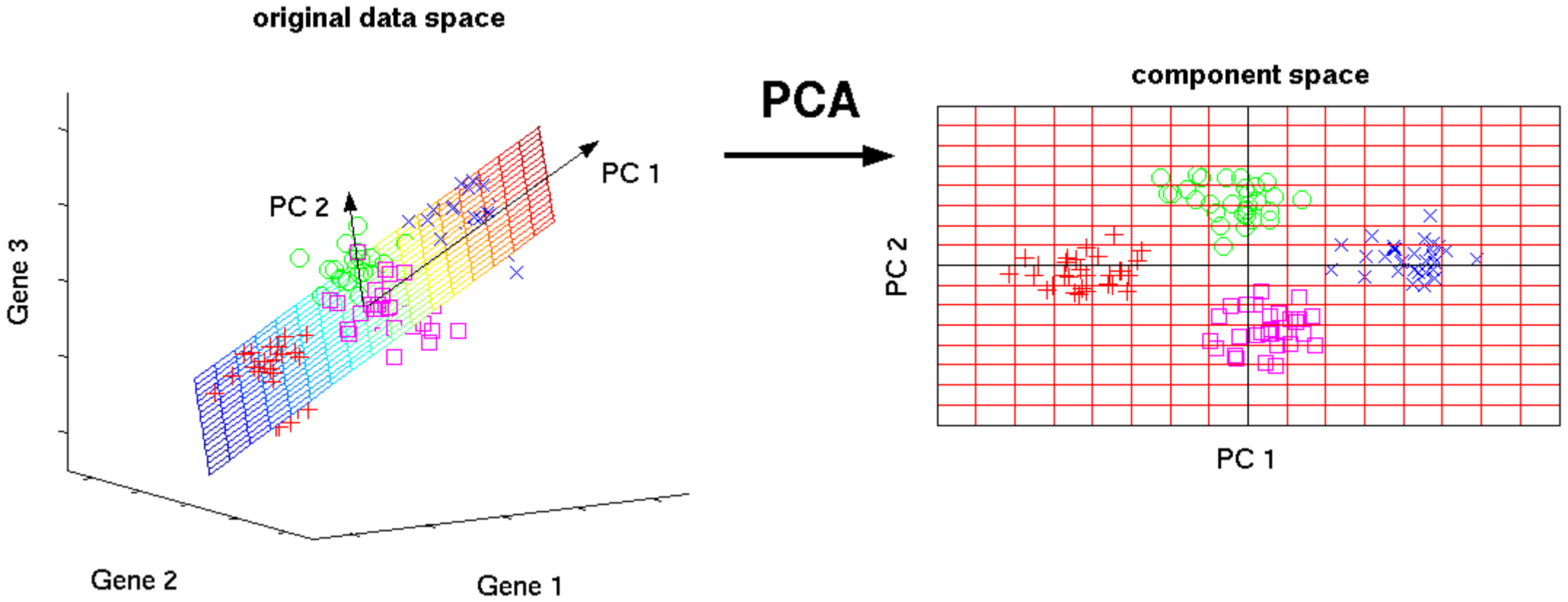
scRNA-seq Downstream Analysis



Wagner et al. (Nat Biotech 2016)

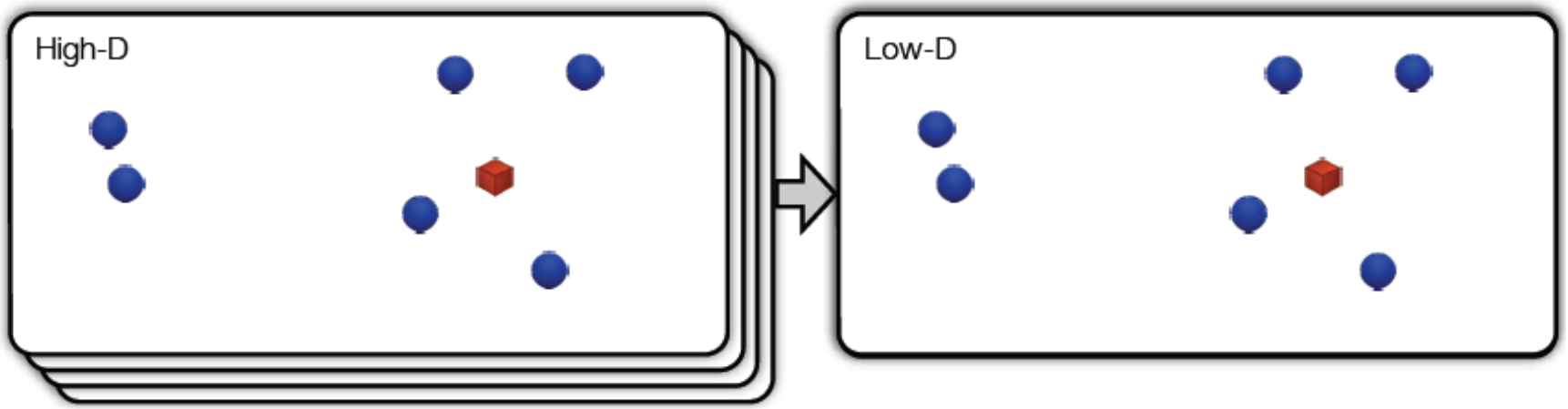
Dimensionality Reduction (1)

- PCA: Principal Component Analysis



Dimensionality Reduction (2)

- t-SNE: t-distributed stochastic neighborhood embedding



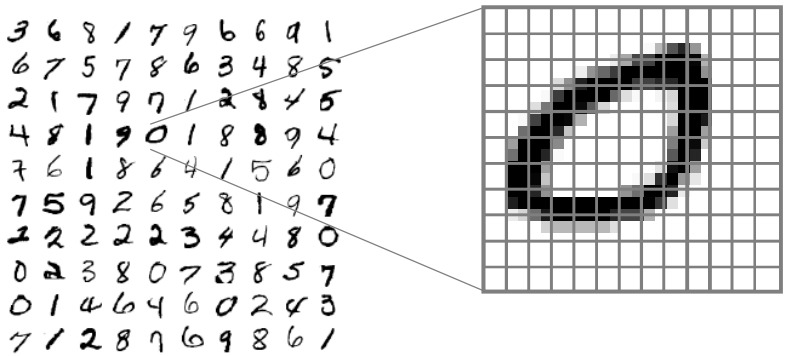
$$p_{ij} = \frac{\exp(-||\boldsymbol{x}_i - \boldsymbol{x}_j||^2/2\sigma^2)}{\sum_{k \neq l} \exp(-||\boldsymbol{x}_k - \boldsymbol{x}_l||^2/2\sigma^2)}$$

$$q_{ij} = \frac{(1 + ||\boldsymbol{y}_i - \boldsymbol{y}_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||\boldsymbol{y}_k - \boldsymbol{y}_l||^2)^{-1}}$$

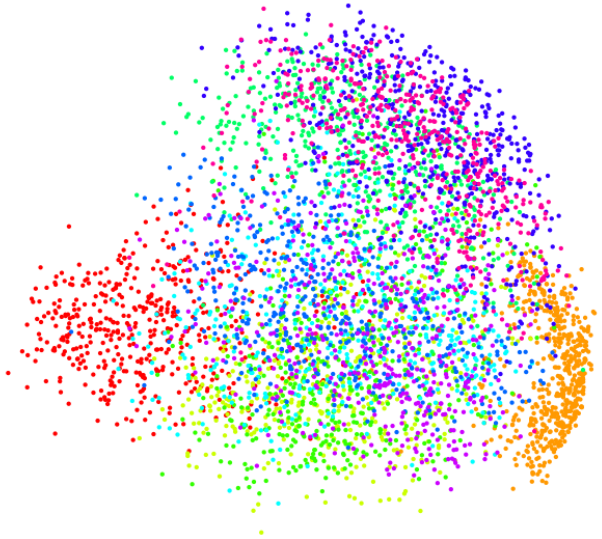
$$KL(P||Q) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

van der Maaten and Hinton, Journal of Machine Learning Research (2008)

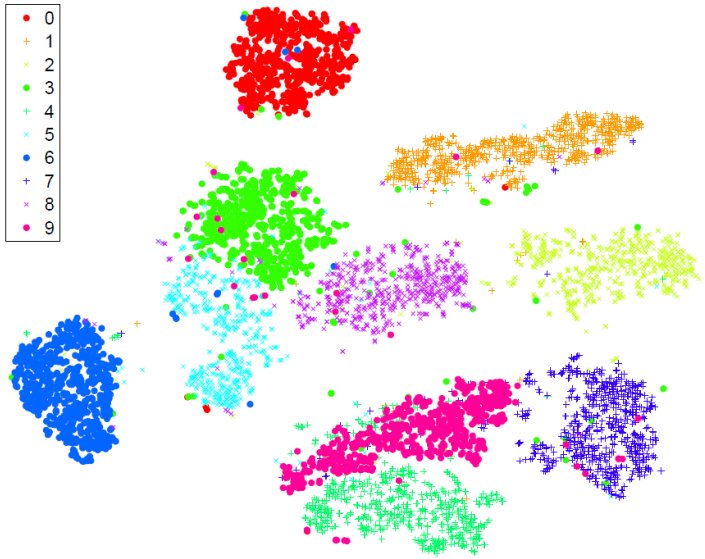
Dimensionality Reduction (3)



PCA



t-SNE

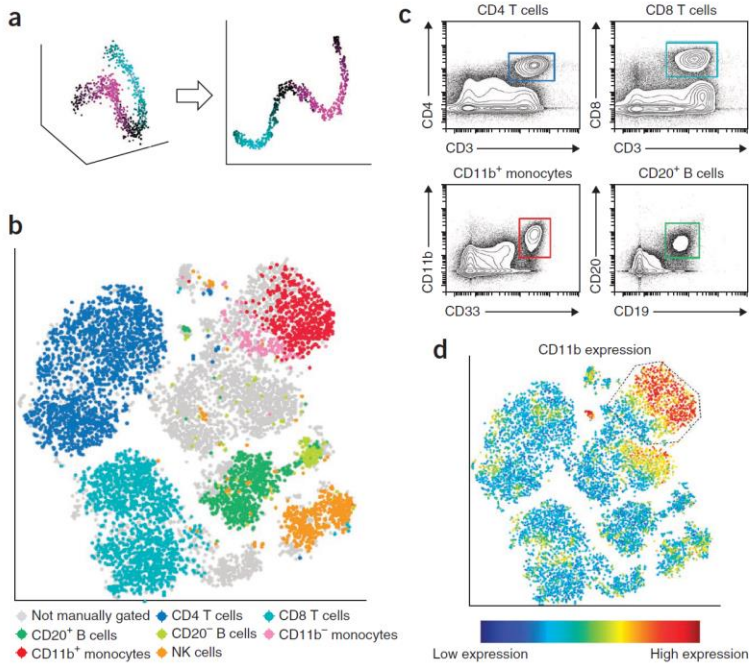


van der Maaten and Hinton, Journal of Machine Learning Research (2008)

Dimensionality Reduction (4)

viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia

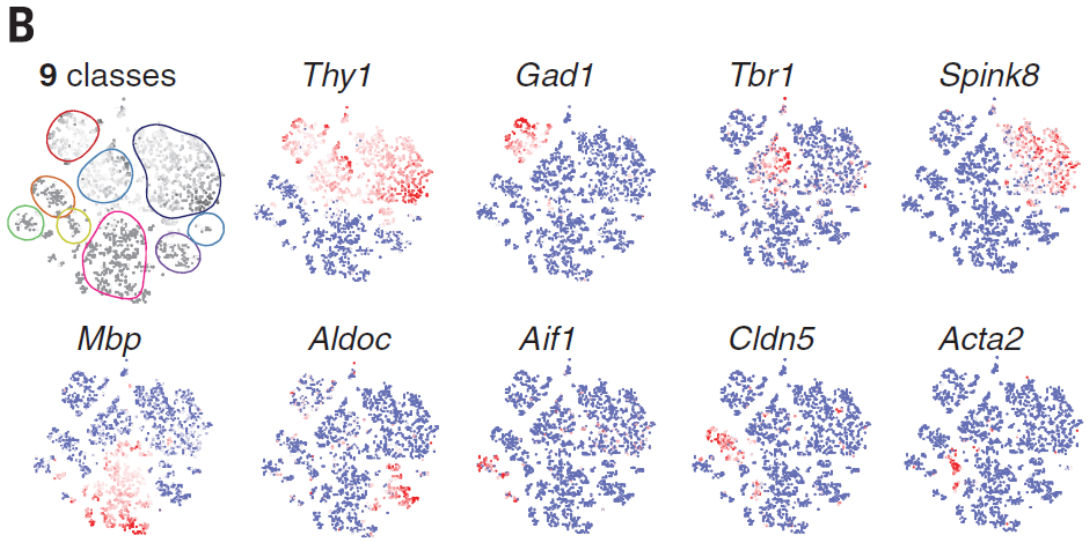
El-ad David Amir[1], Kara L Davis[2,3], Michelle D Tadmor[1,3], Erin F Simonds[2,3], Jacob H Levine[1,3], Sean C Bendall[2,3], Daniel K Shenfeld[1,3], Smita Krishnaswamy[1], Garry P Nolan[2,4] & Dana Pe'er[1,4]



Amir et al., Nature Biotechnology 2013

Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq

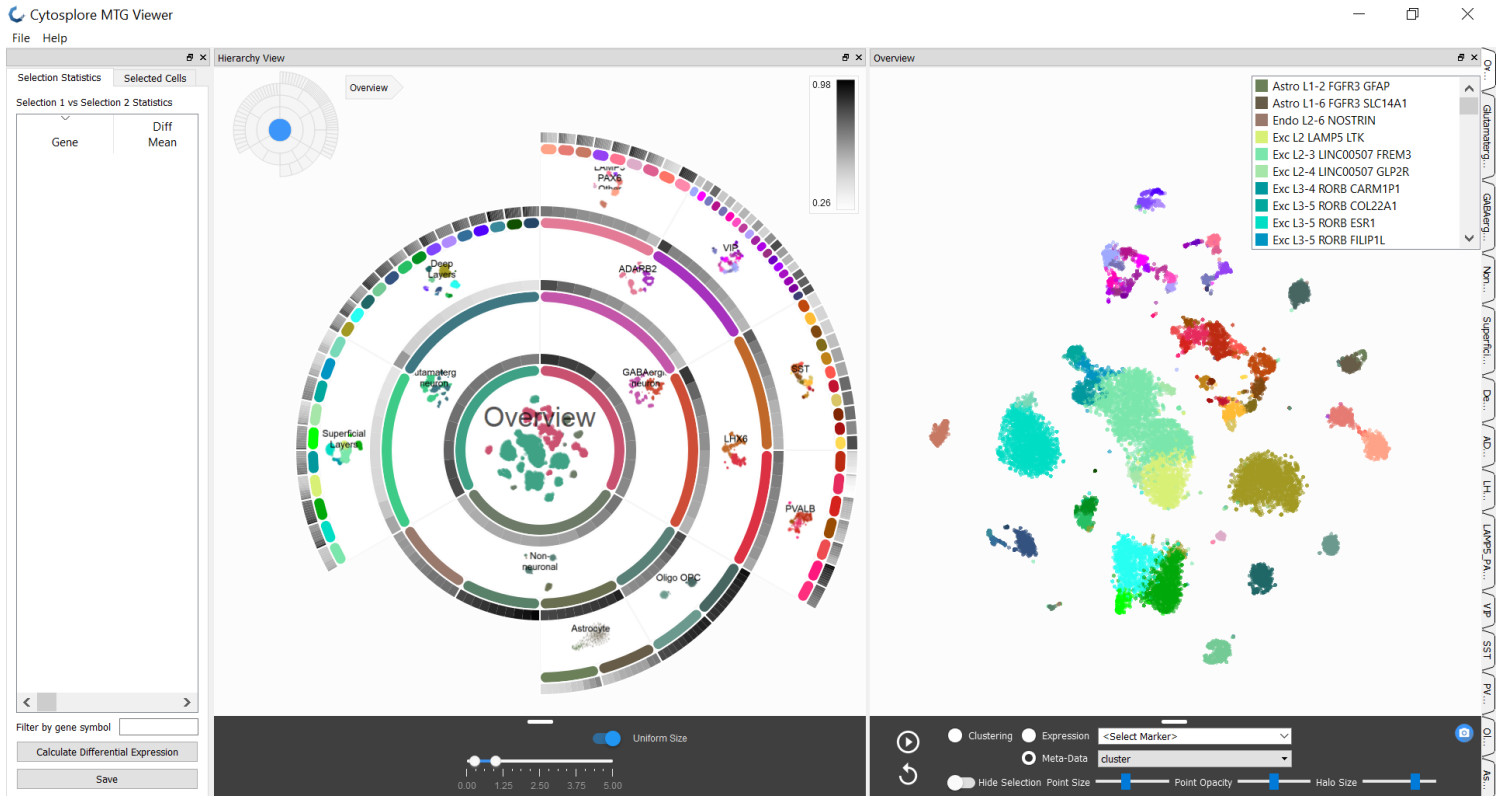
Amit Zeisel,[1*] Ana B. Muñoz-Manchado,[1*] Simone Codeluppi,[1] Peter Lönnerberg,[1] Gioele La Manno,[1] Anna Juréus,[1] Sueli Marques,[1] Hermany Munguba,[1] Liqun He,[2] Christer Betsholtz,[2,3] Charlotte Rolny,[4] Gonçalo Castelo-Branco,[1] Jens Hjerling-Leffler,[1†] Sten Linnarsson[1†]



Zeisel et al., Nature Biotechnology 2013

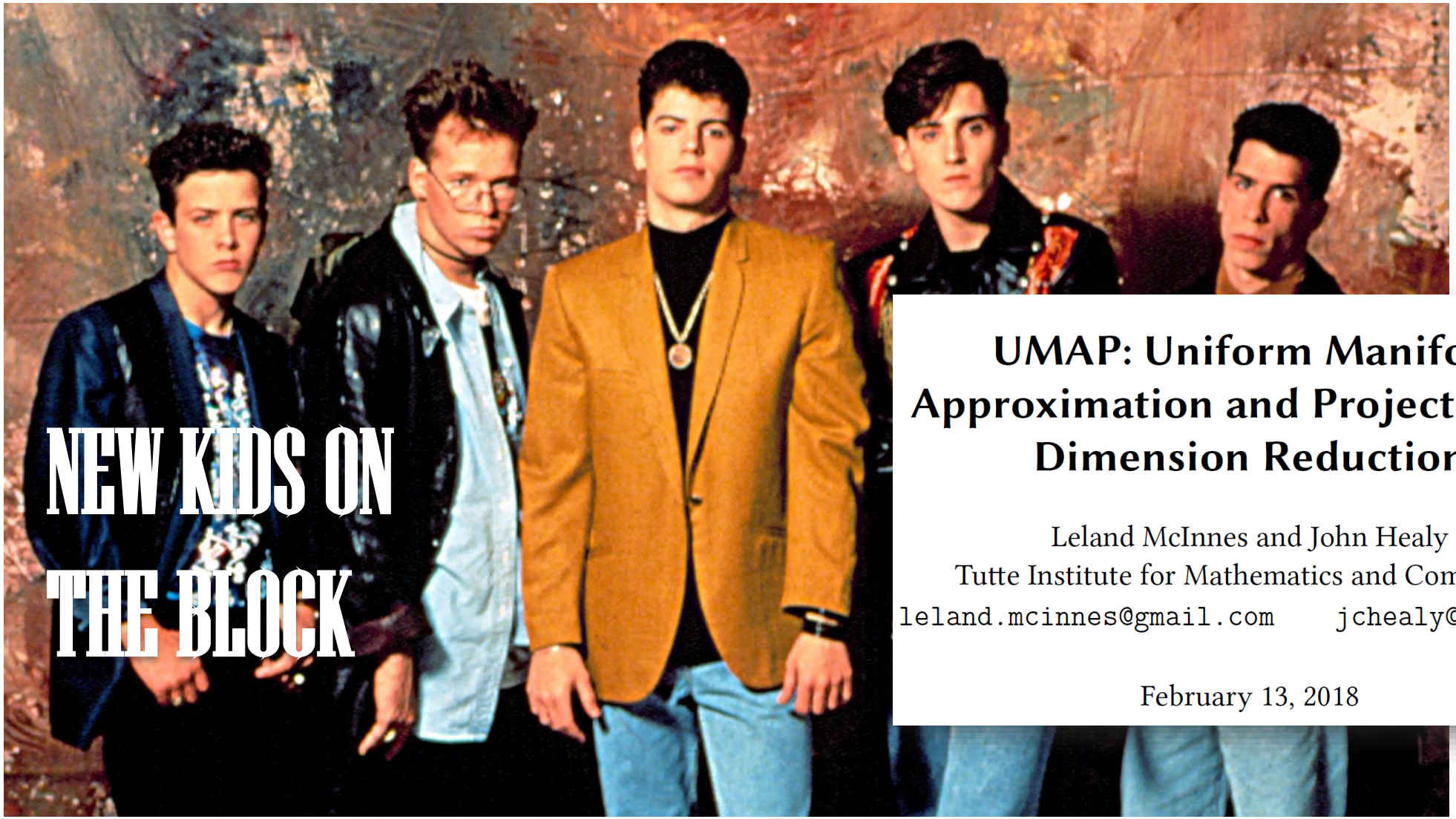
Dimensionality Reduction (5)

- CytoSplore: high performance single cell transcriptome visualizations https://viewer.cytosplore.org



Hodge et al. BioRxiv (2018)

Dimensionality Reduction (5)



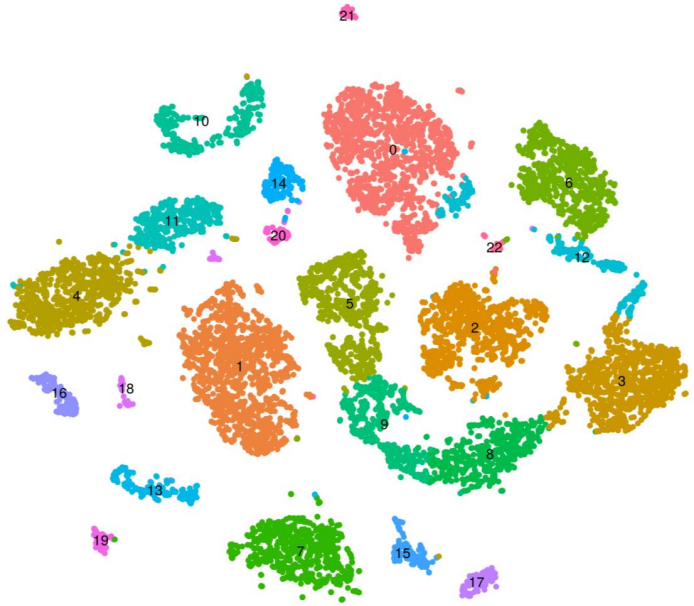
UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

Leland McInnes and John Healy
Tutte Institute for Mathematics and Computing
leland.mcinnes@gmail.com jchealy@gmail.com

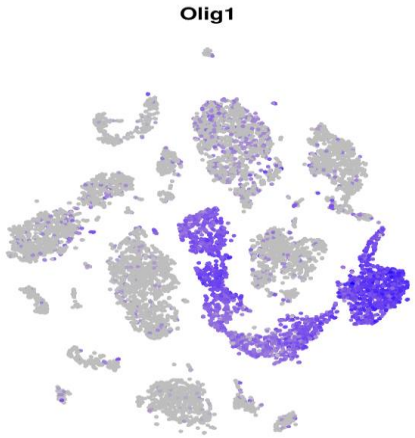
February 13, 2018

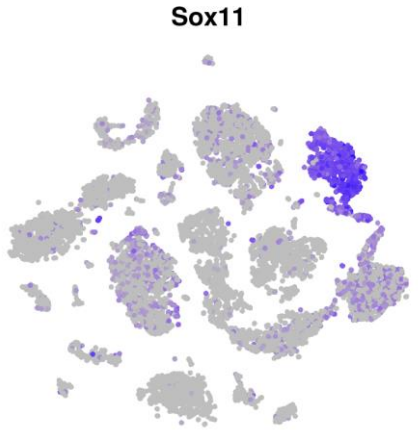
Cell type identification

Clustering



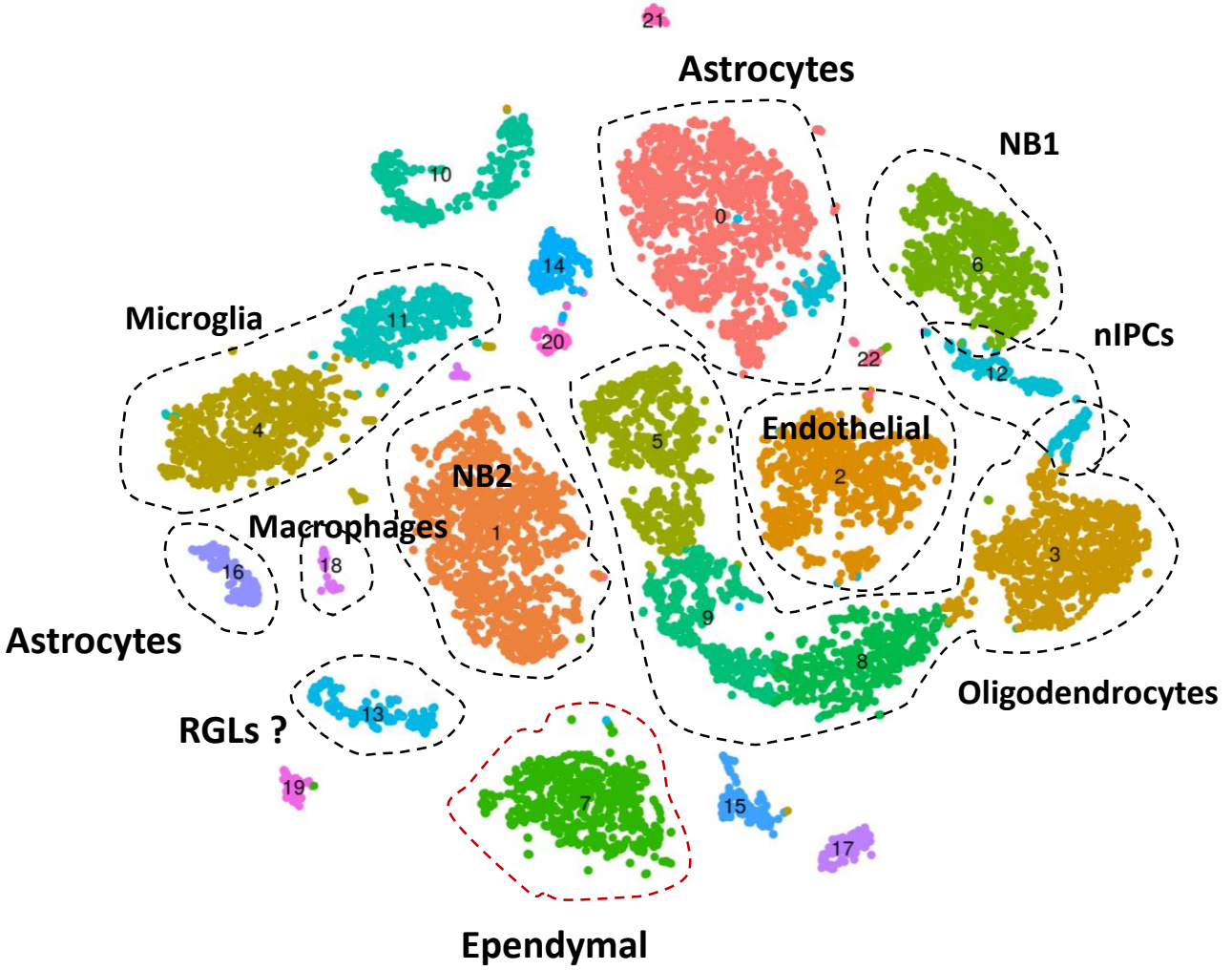
Oligodendrocytes





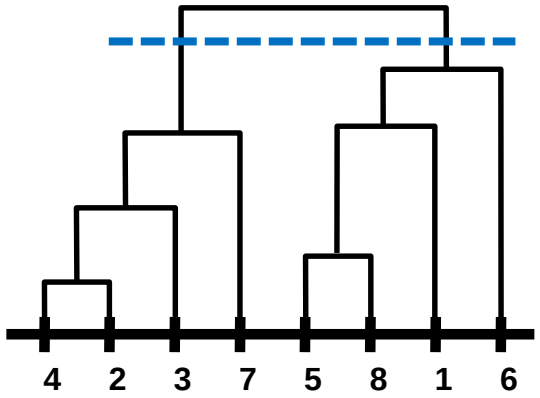
Neurons

Cell type identification

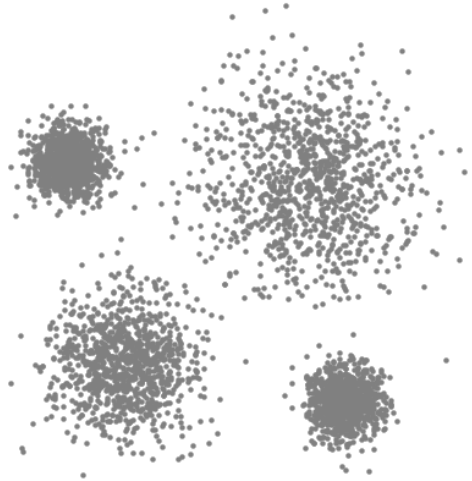


Clustering

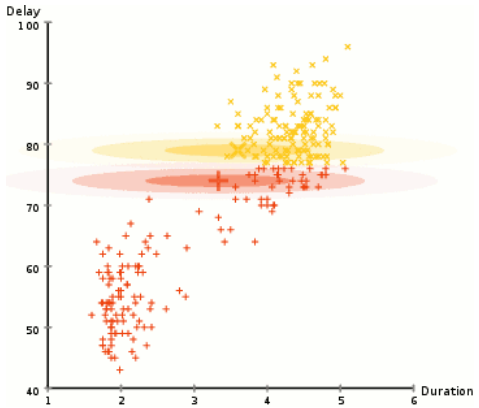
Marcel Reinders



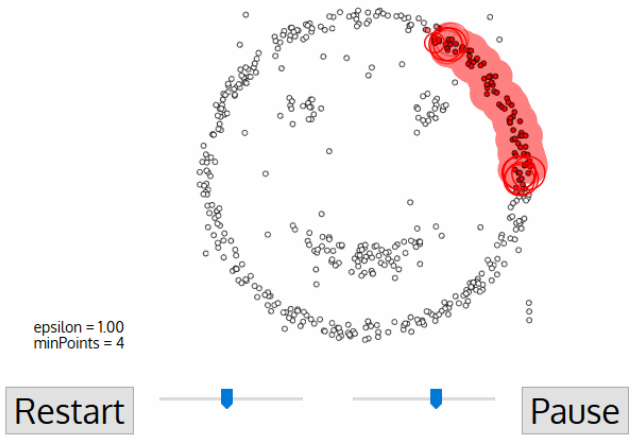
Hierarchical Clustering



Mean shift clustering



Gaussian mixture modeling

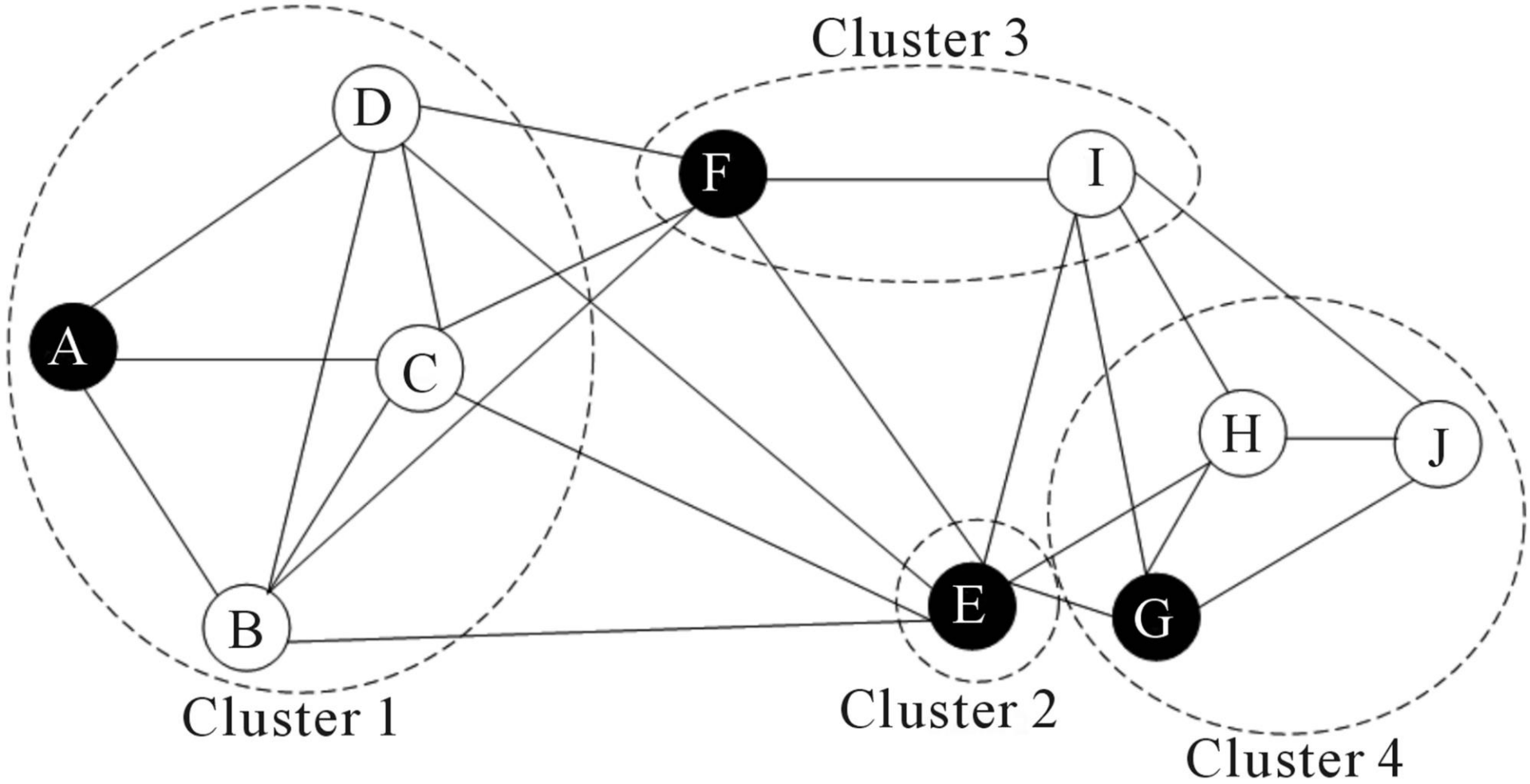


DB scan

Challenges in clustering

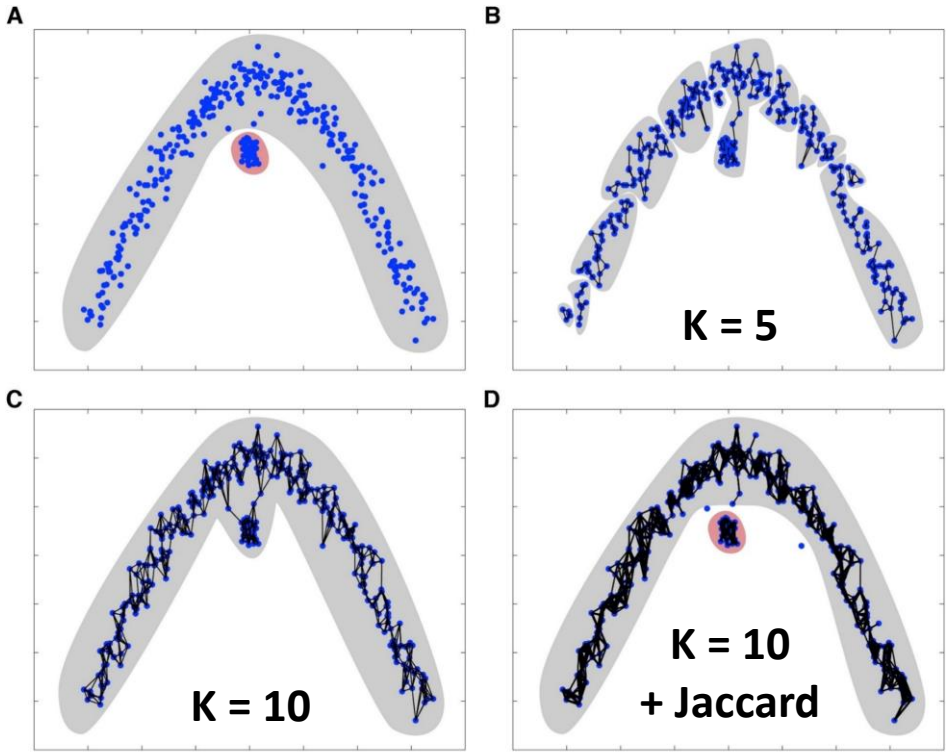
- What is the number of clusters *k*?
- What is a cell type?
- **Scalability**: in the last few years the number of cells in scRNA-seq experiments has grown by several orders of magnitude from ~10^2 to ~10^6
- Tools are not user-friendly

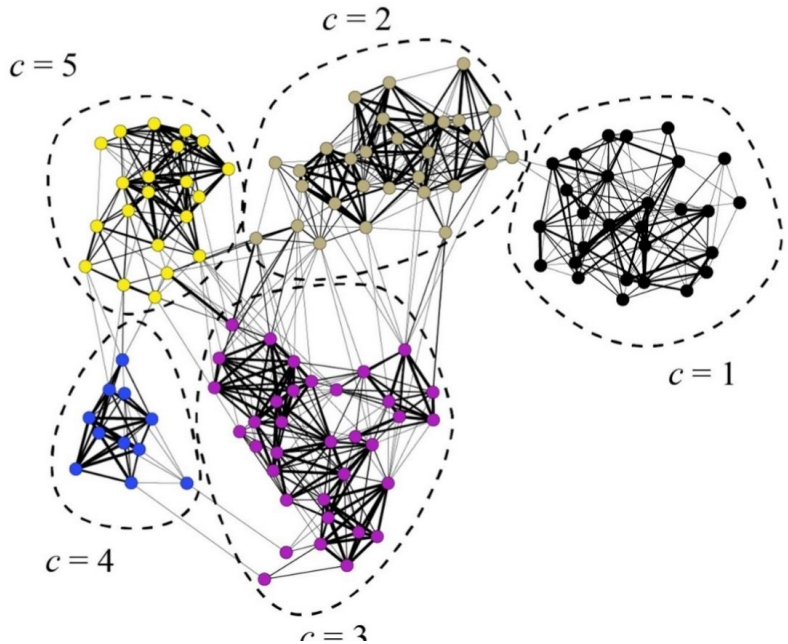
Graph-based clustering

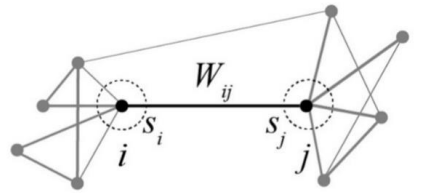


Graph-based clustering

Louvain







$$Q = \frac{1}{2m} \sum_{i,j} \left[W_{ij} - \frac{s_i s_j}{2m} \right] \delta(c_i, c_j)$$

[SNN-Cliq, Xu and Su, Bioinformatics, 2015]

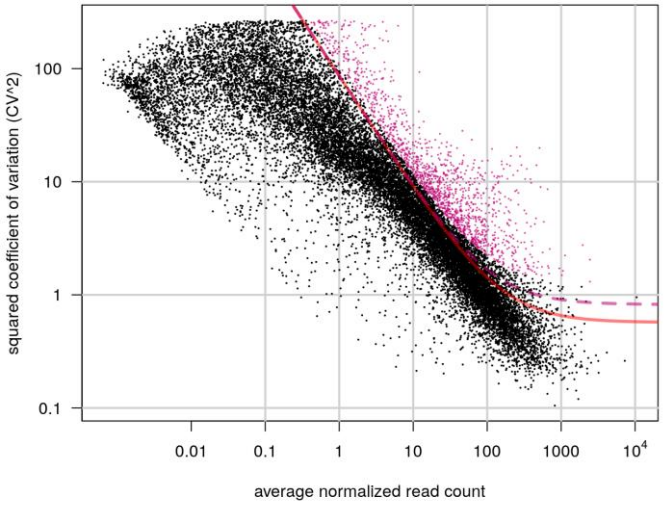
[PhenoGraph, Levine et al., Cell, 2015]

Clustering

Feature selection

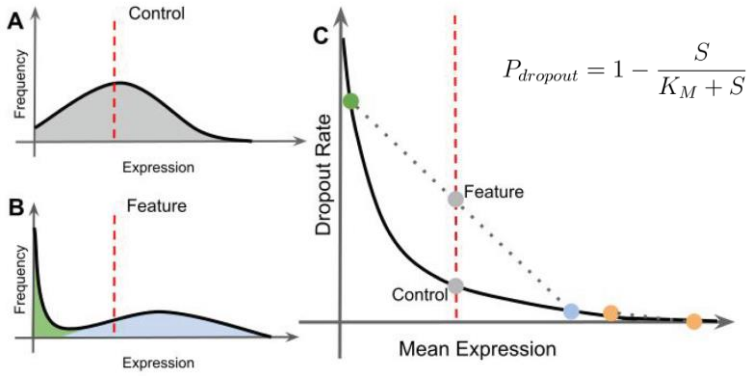
- sc-RNASeq: ~25,000 features per cell
- Only a portion of genes show biologically-relevant differences (e.g. across cell types), the majority show differences due to technical noise
- Often advantageous to perform feature selection to remove genes which only exhibit technical noise from downstream analysis
  - ➢ Increase the signal:noise ratio
  - ➢ Reduce the computational complexity of analyses

Highly Variable Genes



Brennecke et al., Nature Methods 2013

High Dropout Genes

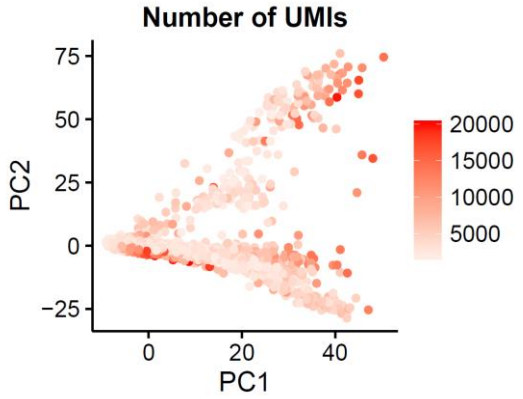


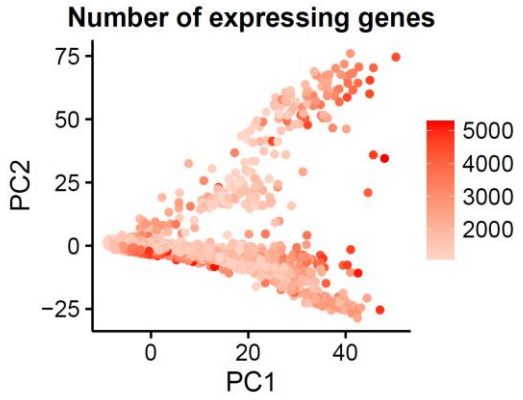
Andrews & Hemberg, bioRxiv 2016

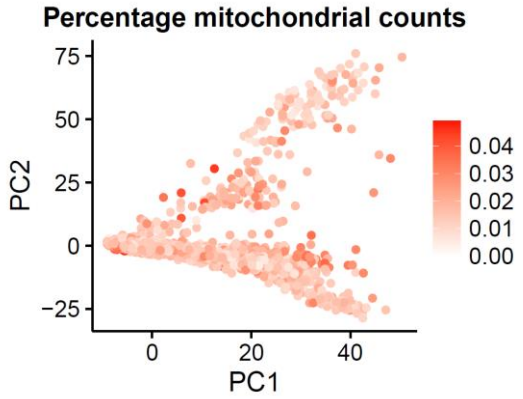
Clustering

Dealing with confounders

- Confounding factors:
  - number of detected molecules
  - number of expressing genes
  - mitochondrial gene expression
  - cell cycle
  - …
- Solution: use a linear model to regress them out

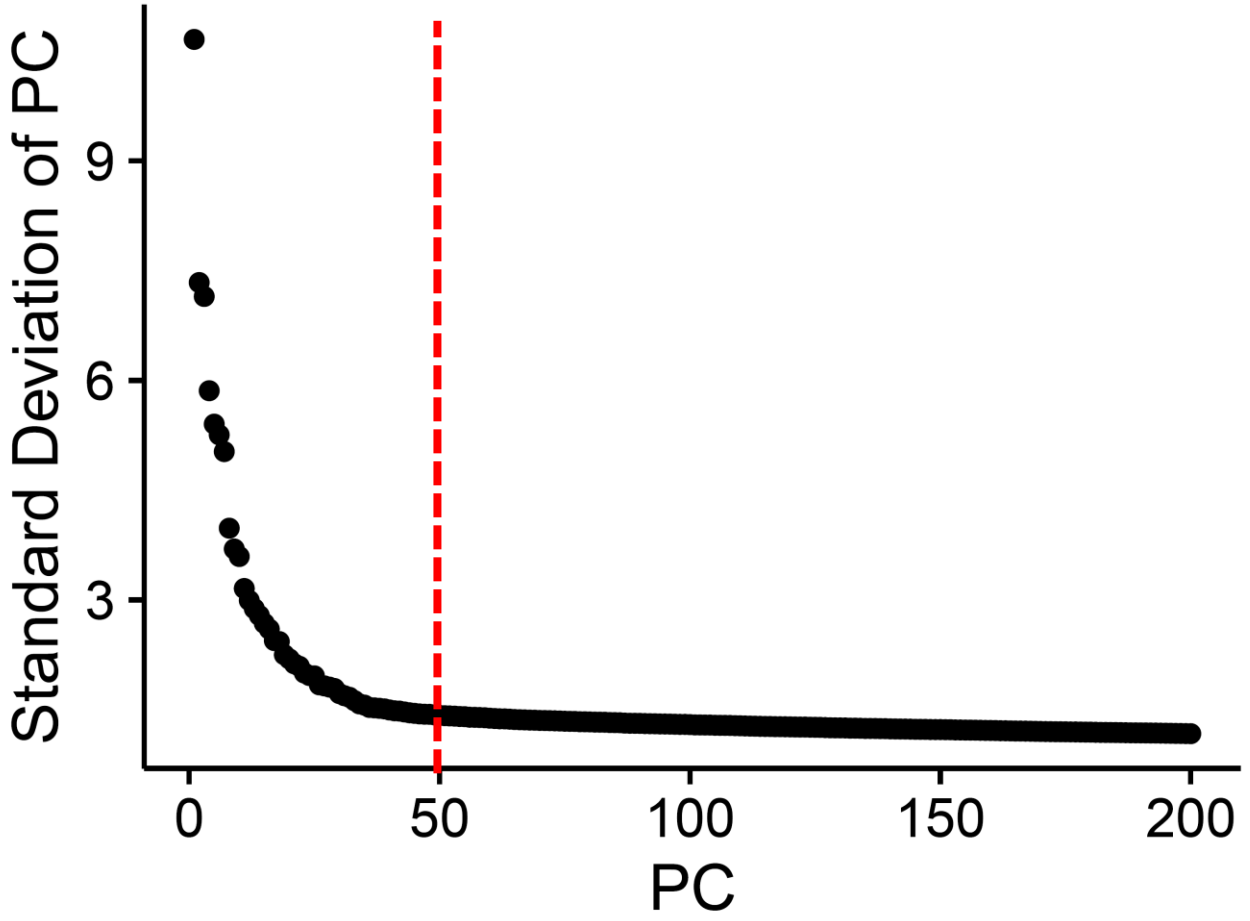






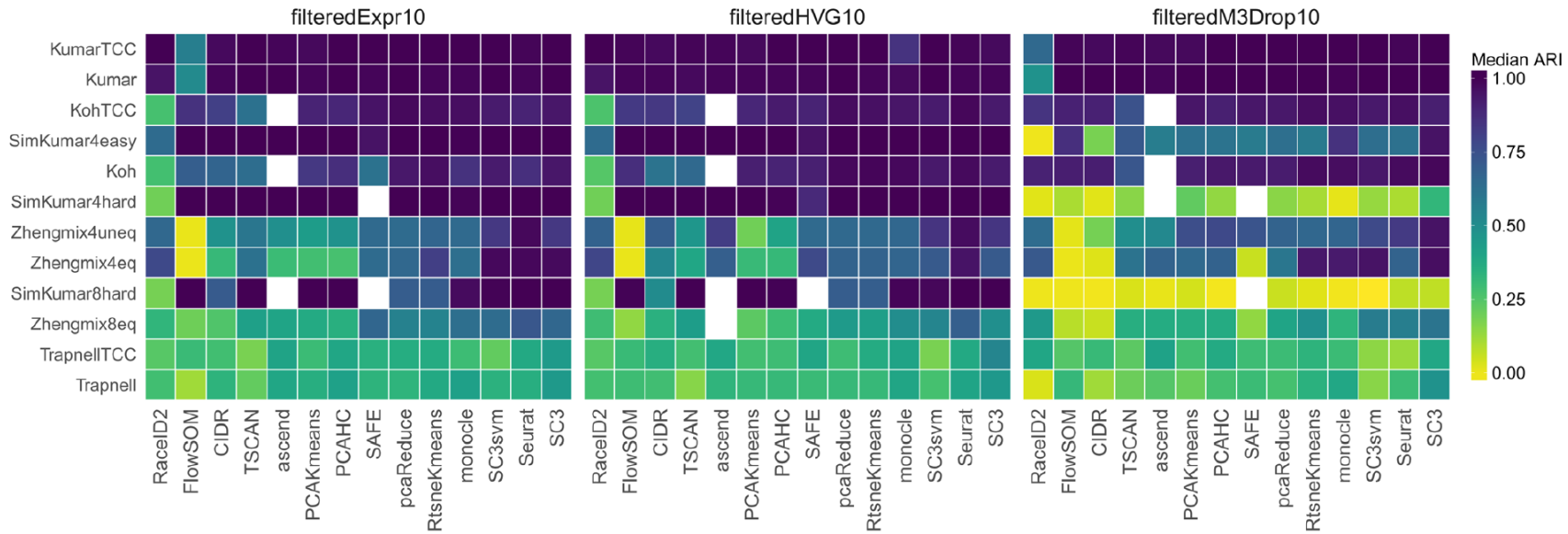
Clustering

Selecting PCs



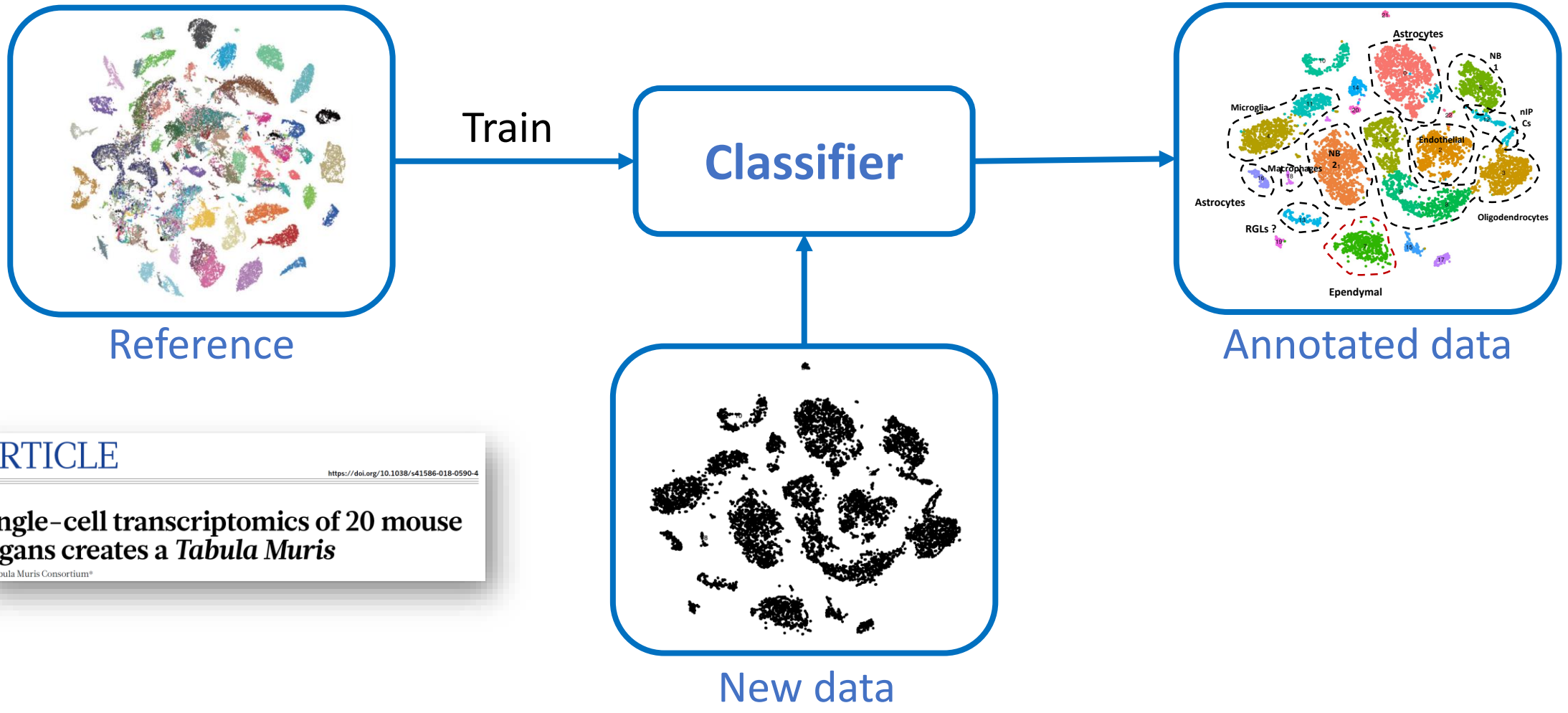
scRNA-seq Clustering

- BackSpin
- tSNE + k-means
- SC3
- SINCERA
- pcaReduce
- Seurat (a.k.a. SNN-Cliq)
- …



Duò et al. F1000Research 2018

Automatic cell type identification



ARTICLE

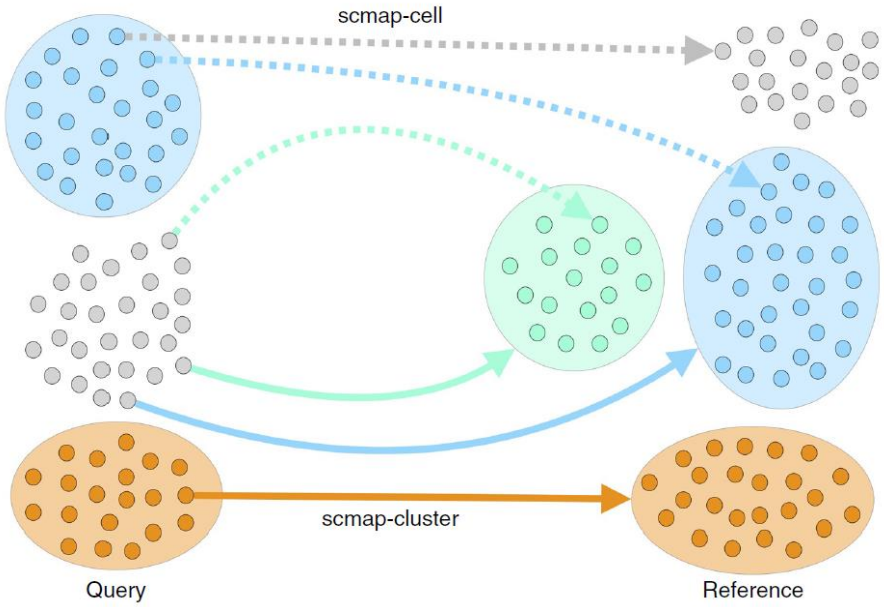
https://doi.org/10.1038/s41586-018-0590-4

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

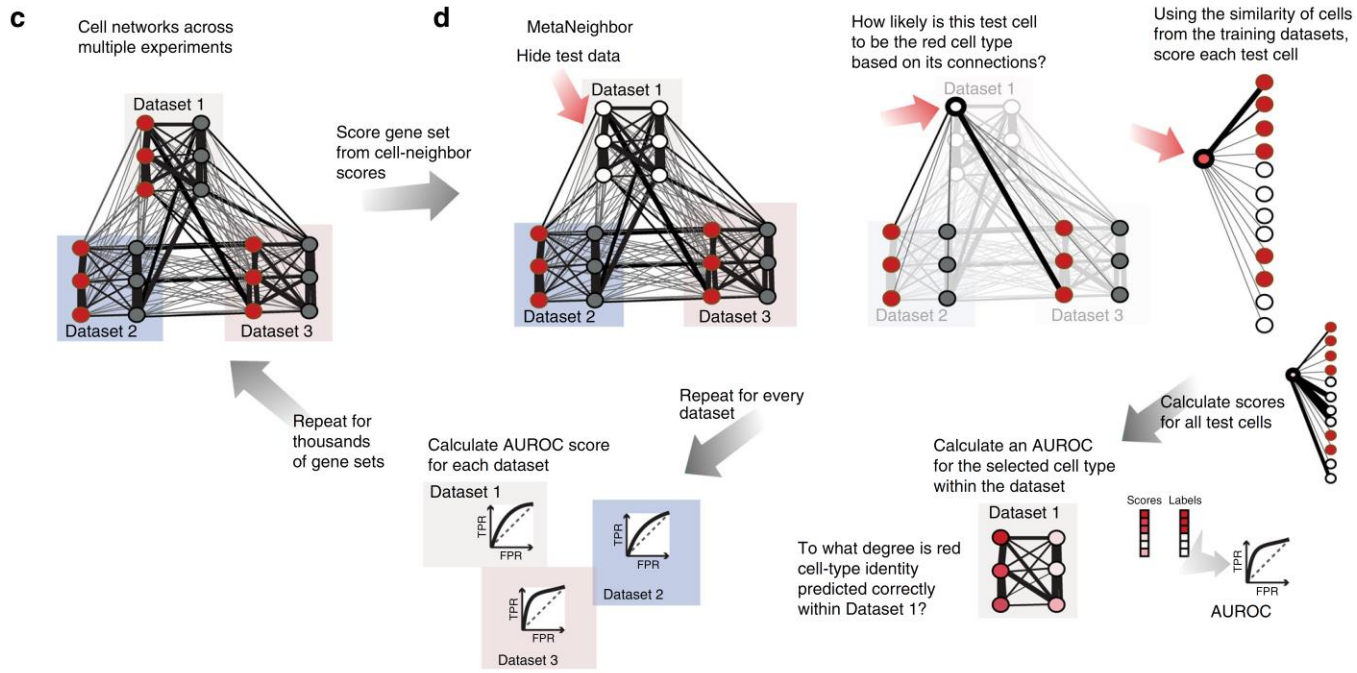
The Tabula Muris Consortium*

Single cell classification

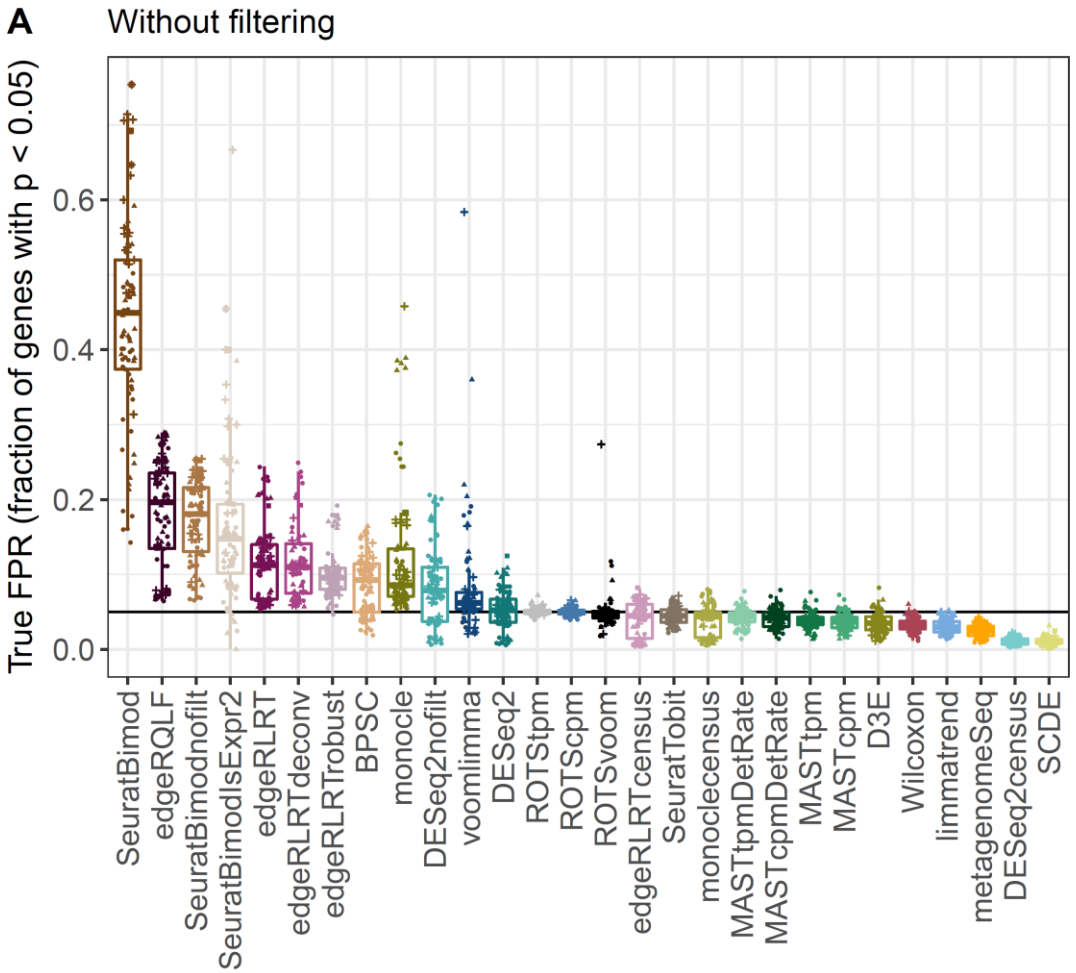
scmap
(Kiselev et al. Nature Methods 2018)

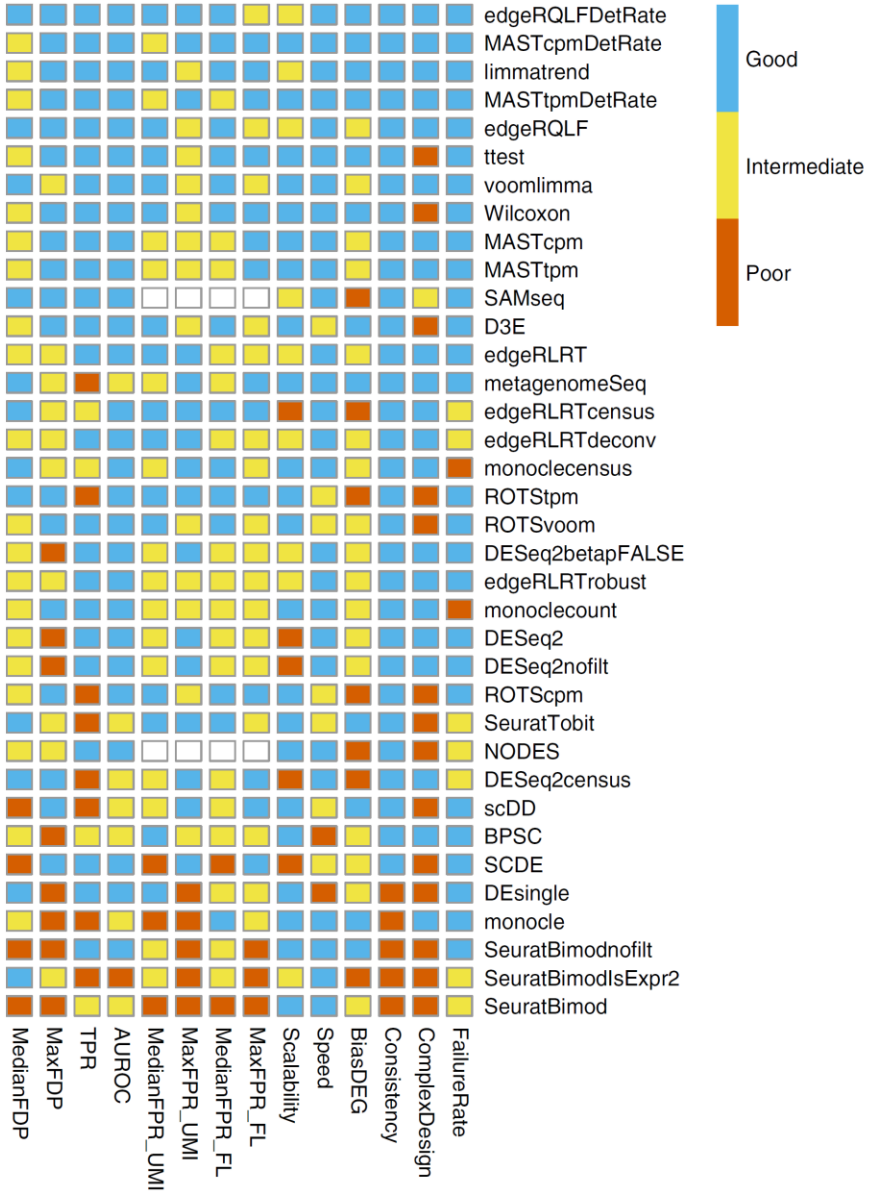


MetaNeighbor
(Crow et al. Nature Communications 2018)



Differential expression



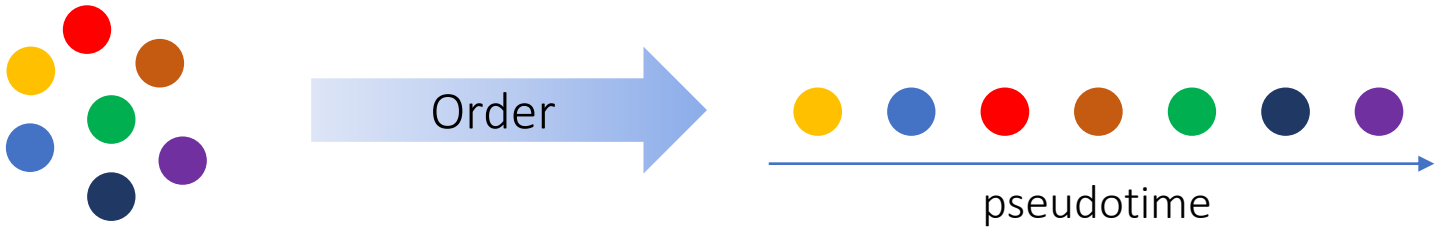


Soneson and Robinson, Nature Methods 2018

Pseudo-temporal ordering

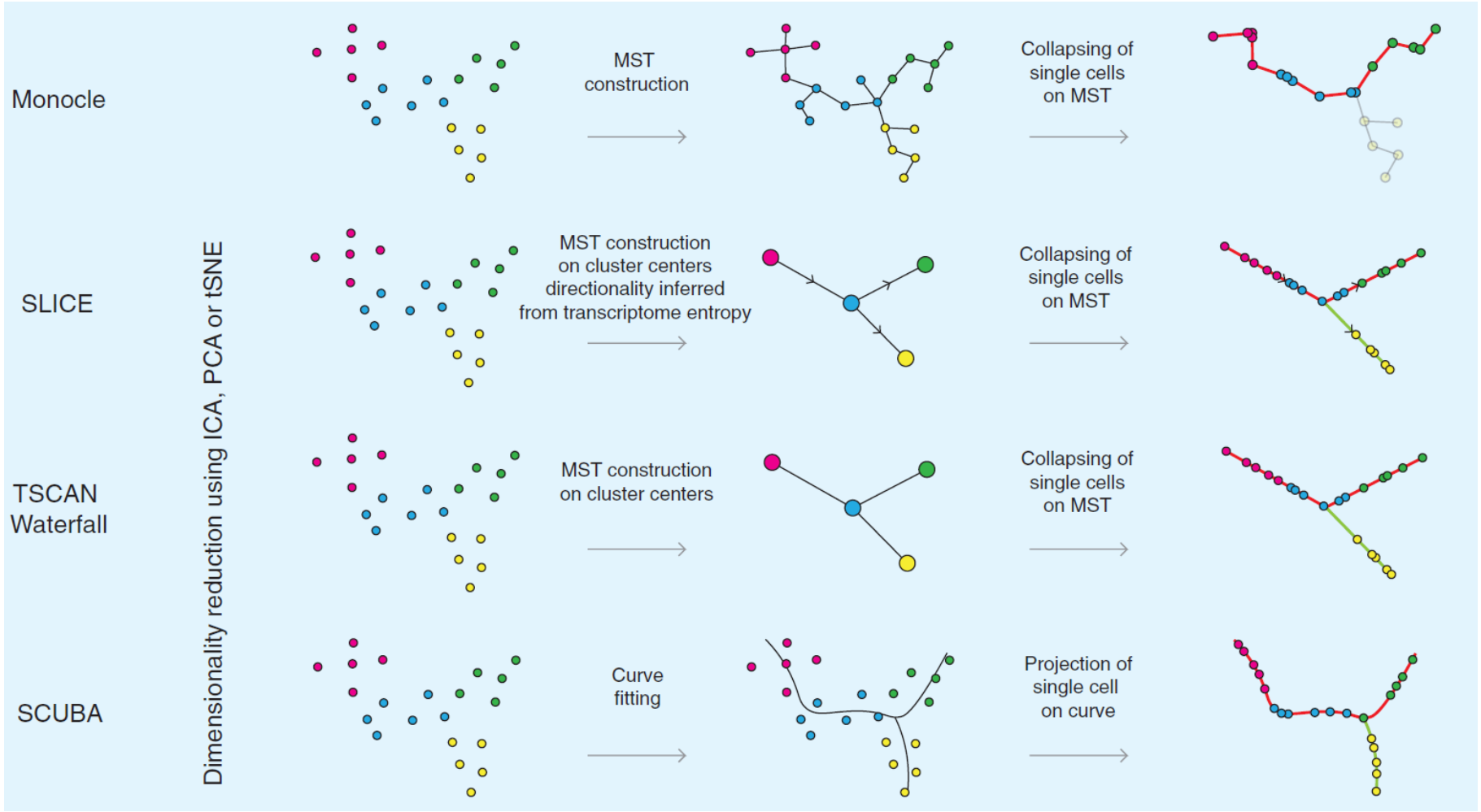
- Pseudotime: artificial measure of a cell's progression through some process (e.g. differentiation) from scRNA-seq snapshot data

- Key assumptions:
  - continuity of transcriptome changes
  - presence of all intermediate cell stages



Pseudo-temporal ordering

Dimensionality Reduction

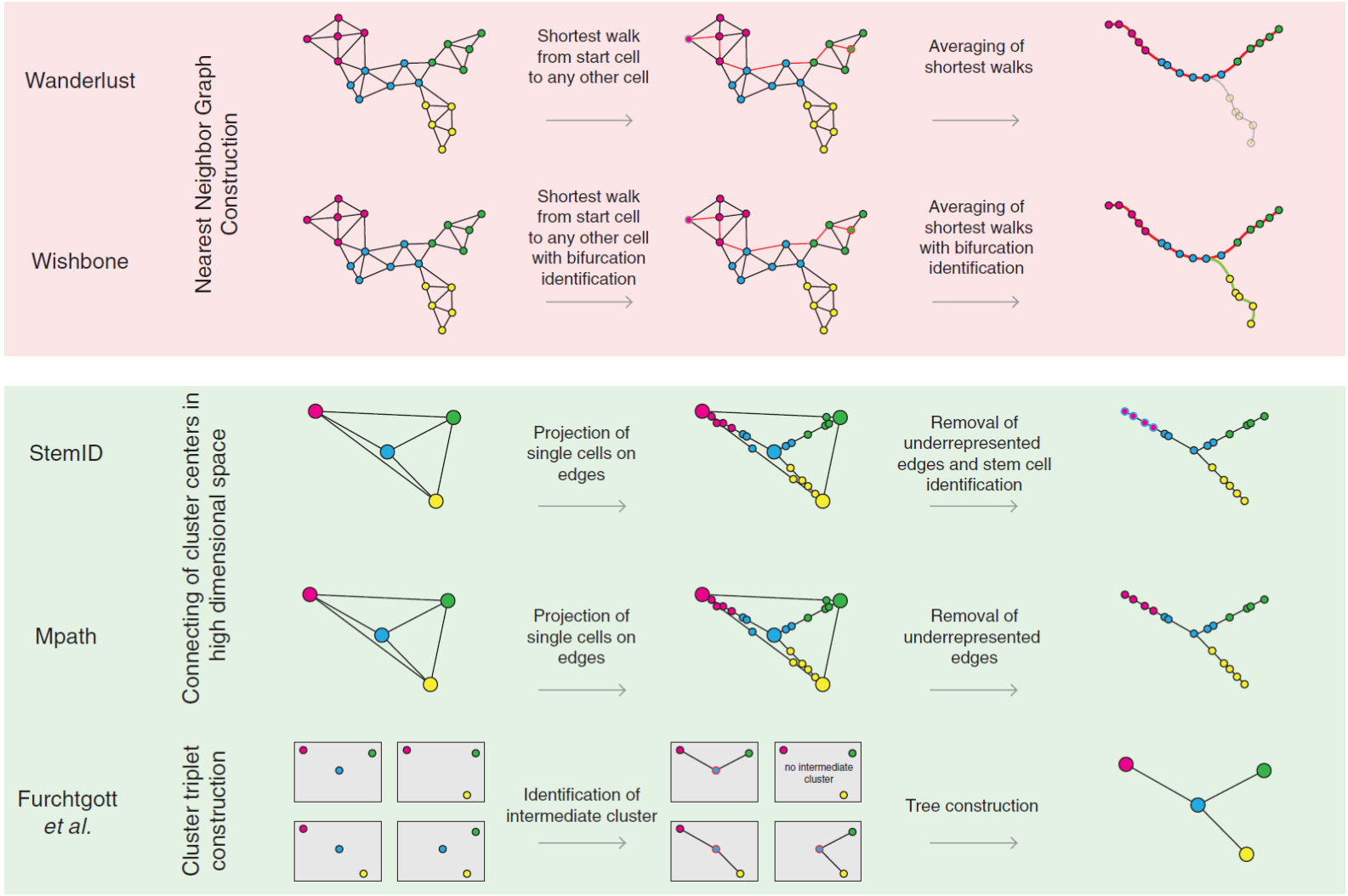


Kester and Oudenaarden, Cell Stem Cell 2018

Pseudo-temporal ordering

Nearest Neighbor Graphs

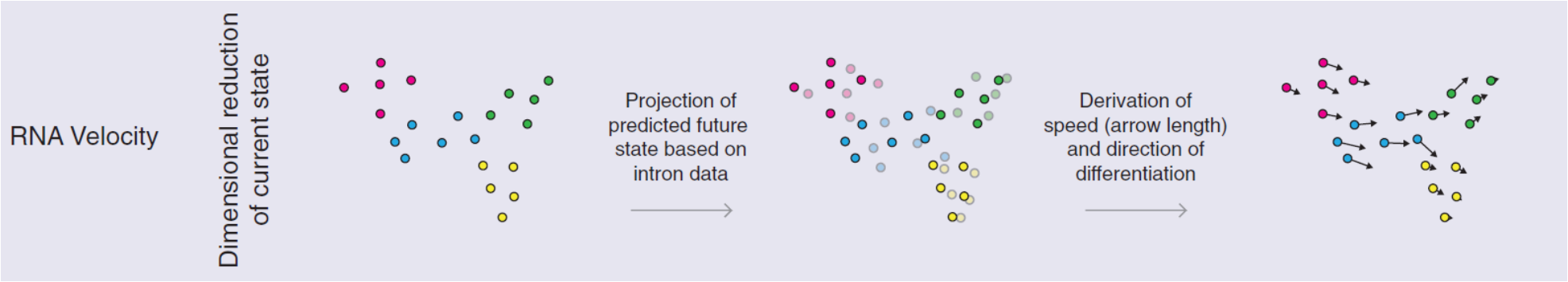
Cluster Networks



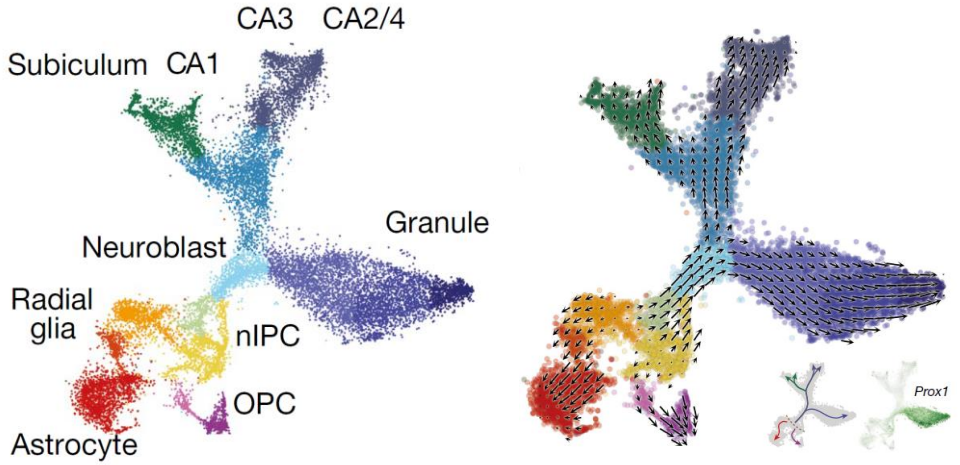
Kester and Oudenaarden, Cell Stem Cell 2018

Pseudo-temporal ordering

- RNA Velocity: time derivative of the gene expression state
  - ➢ Can be directly estimated by distinguishing between unspliced and spliced mRNAs

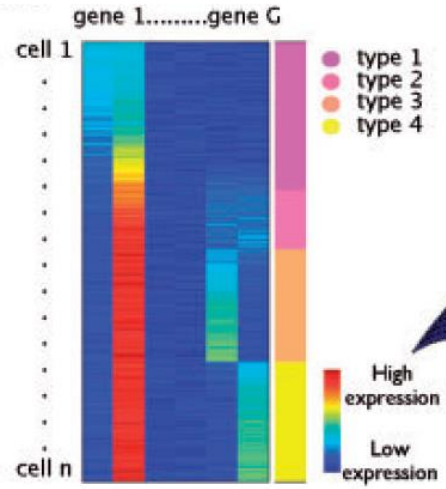


La Manno et al., Nature 2018

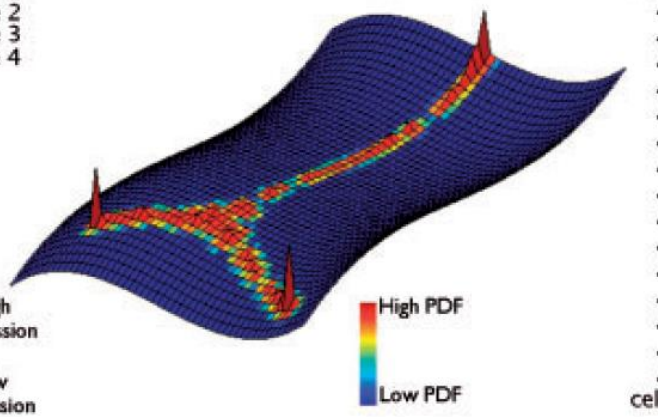


Kester and Oudenaarden, Cell Stem Cell 2018

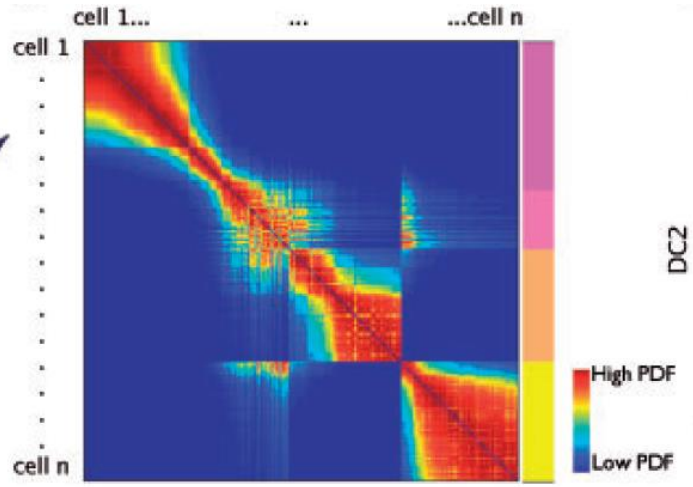
Diffusion Maps



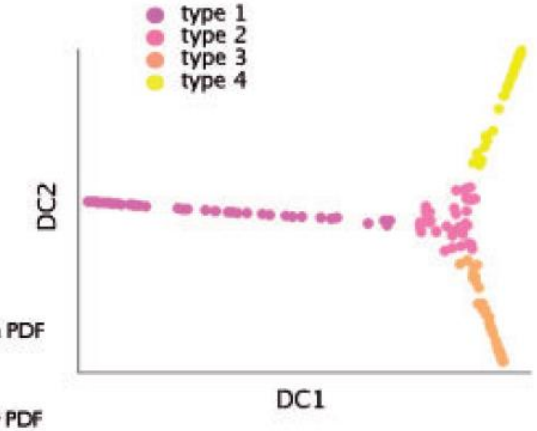
Cell X Gene matrix



Representation of each cell by a Gaussian in the G-dimensional gene space



Cell X Cell Markovian transition probability matrix

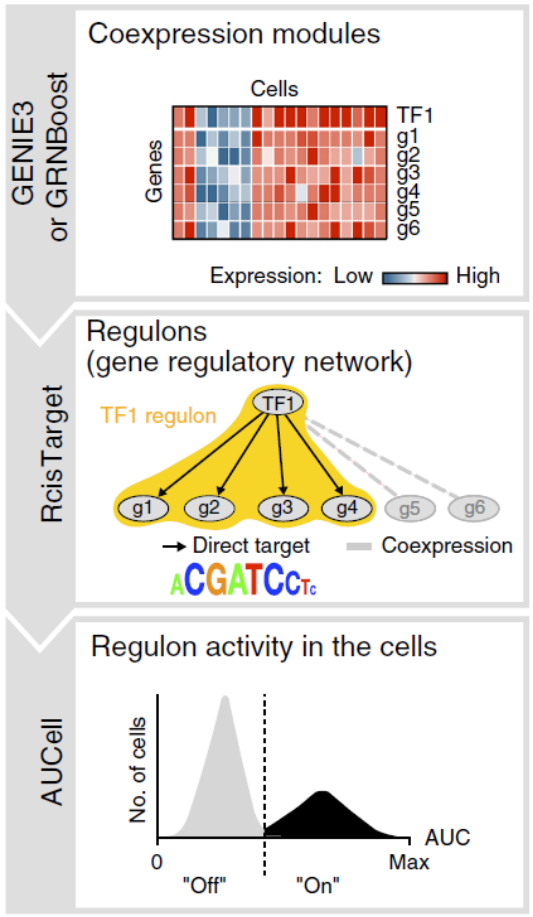


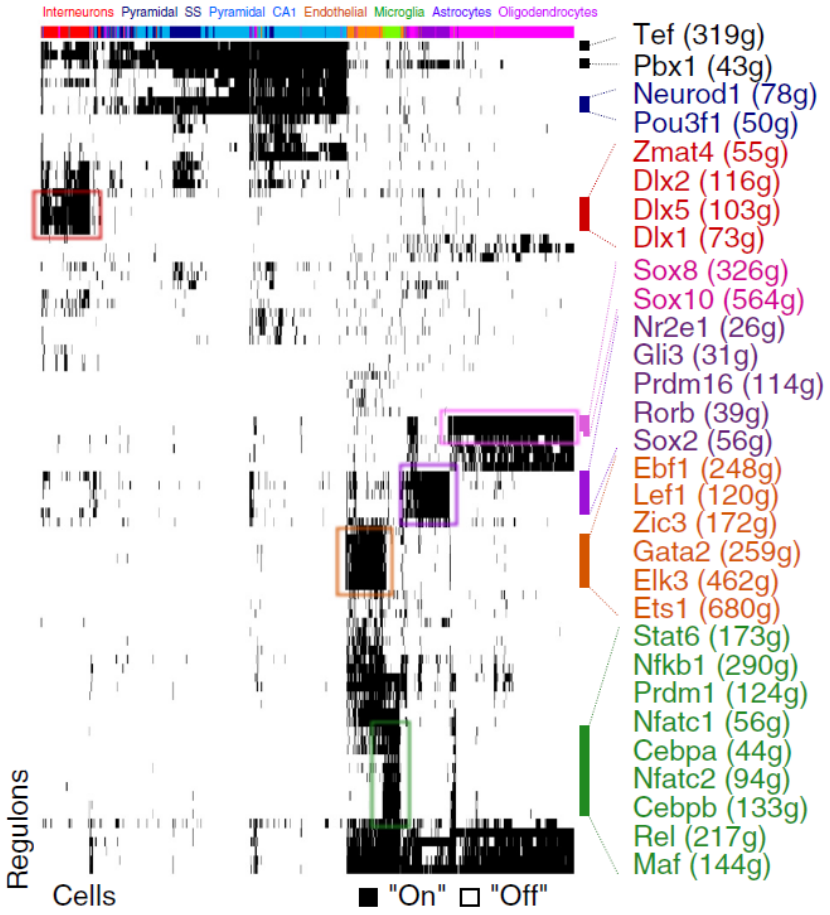
Data embedding on the first two eigenvectors of the Markovian transition matrix

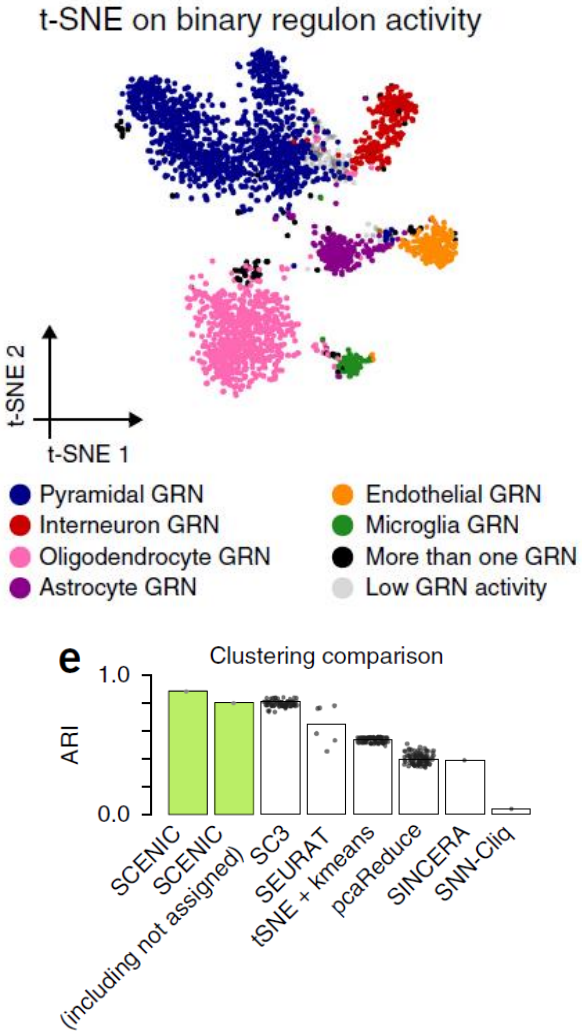
Haghverdi et al., Bioinformatics 2015

Single cell regulatory networks

SCENIC



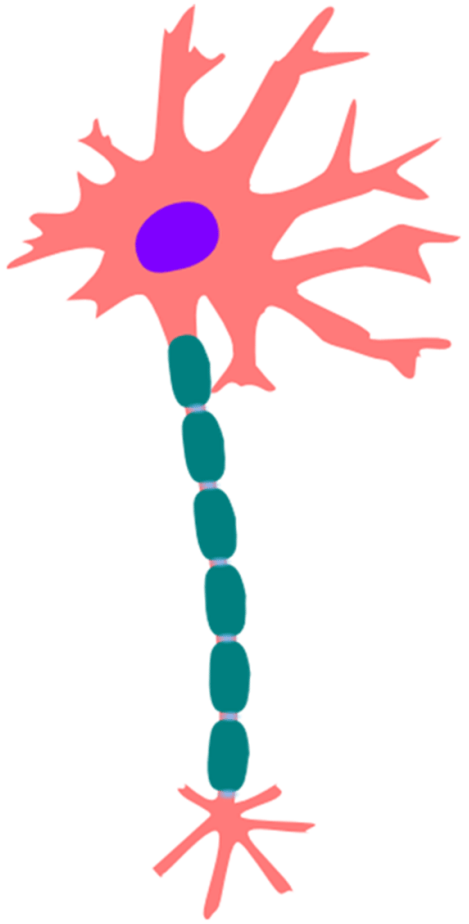




Aibar et al., Nature Methods 2018

Summary

- scRNA-seq protocols
- Constructing the cell x gene matrix
- Quality control (cells & genes)
- Normalization
- Confounding factors and batch effects
- Dimensionality reduction
- Cell type identification (clustering & classification)
- Pseudo-temporal ordering
- Single cell regulatory networks

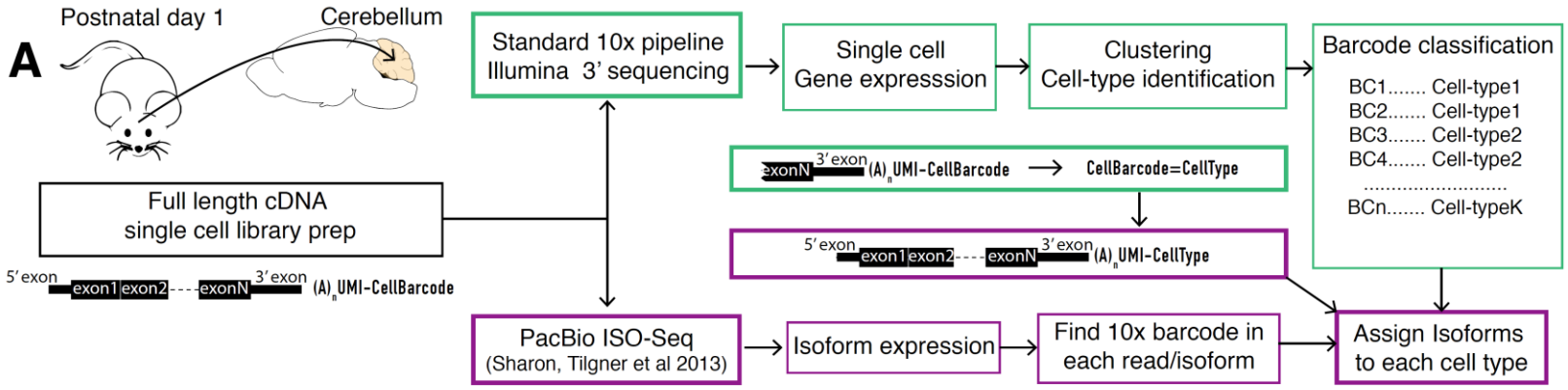


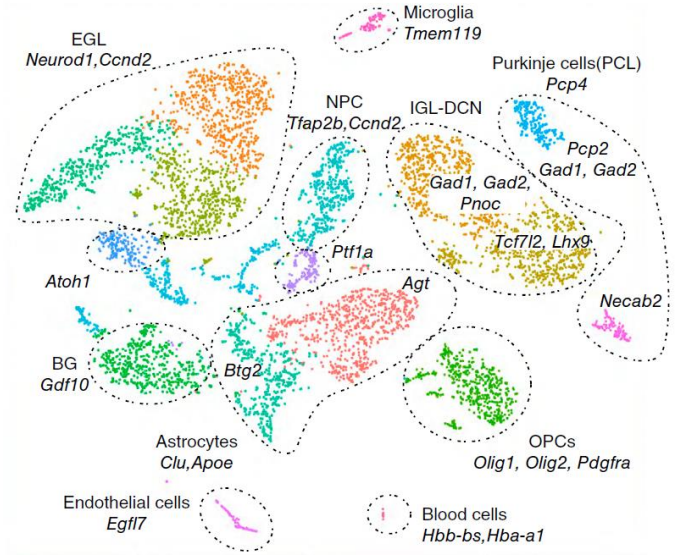
There is a lot more to it…

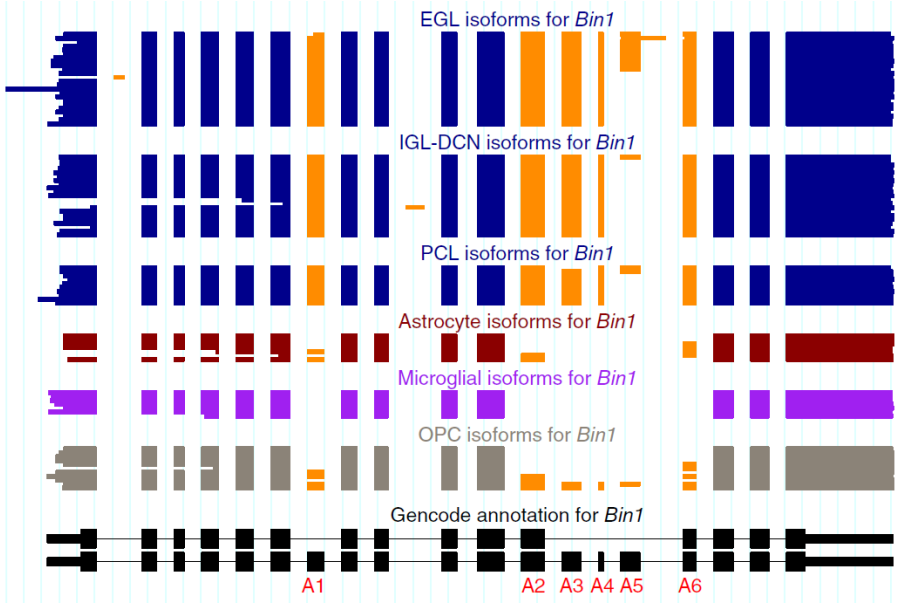
- Single nucleus sequencing
- Imputation
- scRNA-seq + protein
- Sample multiplexing
- Single cell isoform sequencing
- Cell lineage + scRNA-seq
- Spatial transcriptomics
- …

Single cell isoform RNA sequencing

ScISOr-seq







Gupta*, Collier* et al. (Nature Biotechnology 2018)

Useful Resources

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

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

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

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

- GitHub: Awesome Single Cell

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

- Recent developments in single cell genomics

https://www.dropbox.com/s/woya6ffgq8a3pkw/SingleCellGenomcisDay18_References.pdf?dl=1

Thank You

Projects Available!

1. Automatic classification of brain cells.
2. Regulatory networks underlying pancreatic beta-to-alpha cell conversions in T2D.
3. Identifying cell-type specific signaling partners of nuclear receptors.
4. …


a.mahfouz@lumc.nl


https://www.lcbc.nl/


@ahmedElkoussy