Single Cell RNA-seq Analysis

Ahmed Mahfouz

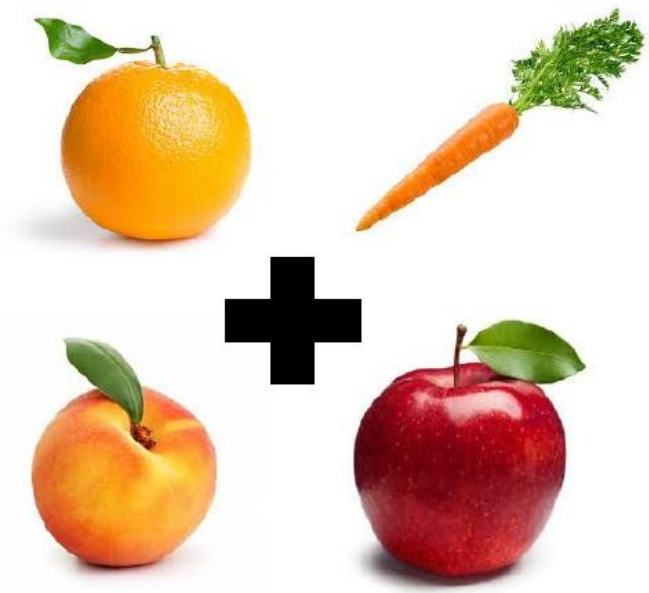
Human Genetics Department, LUMC
Leiden Computational Biology Center, LUMC
Delft Bioinformtaics Lab, TU Delft

Leiden
Computational Biology Center

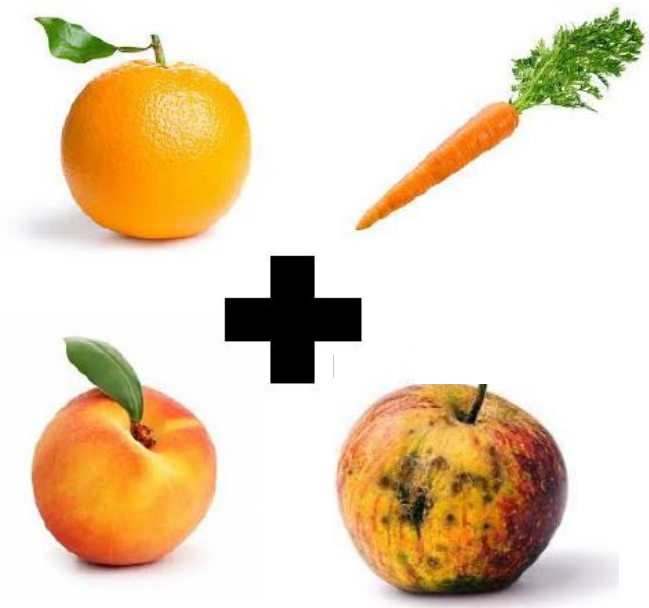
LUMC

TU Delft

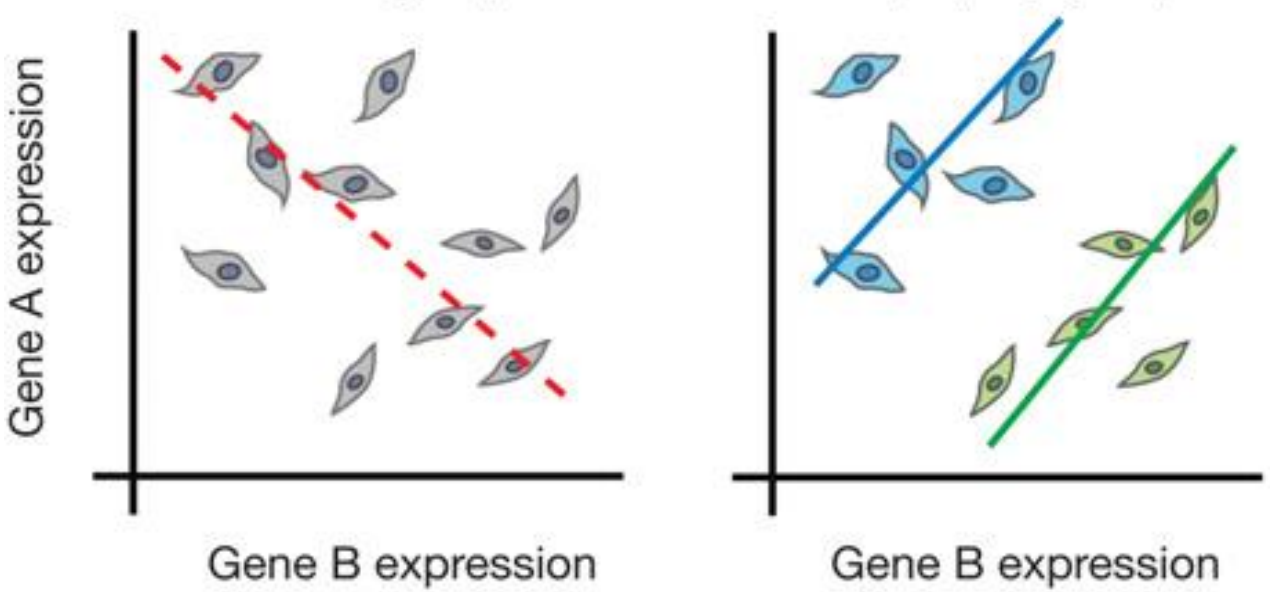
Why single cells?



Why single cells?



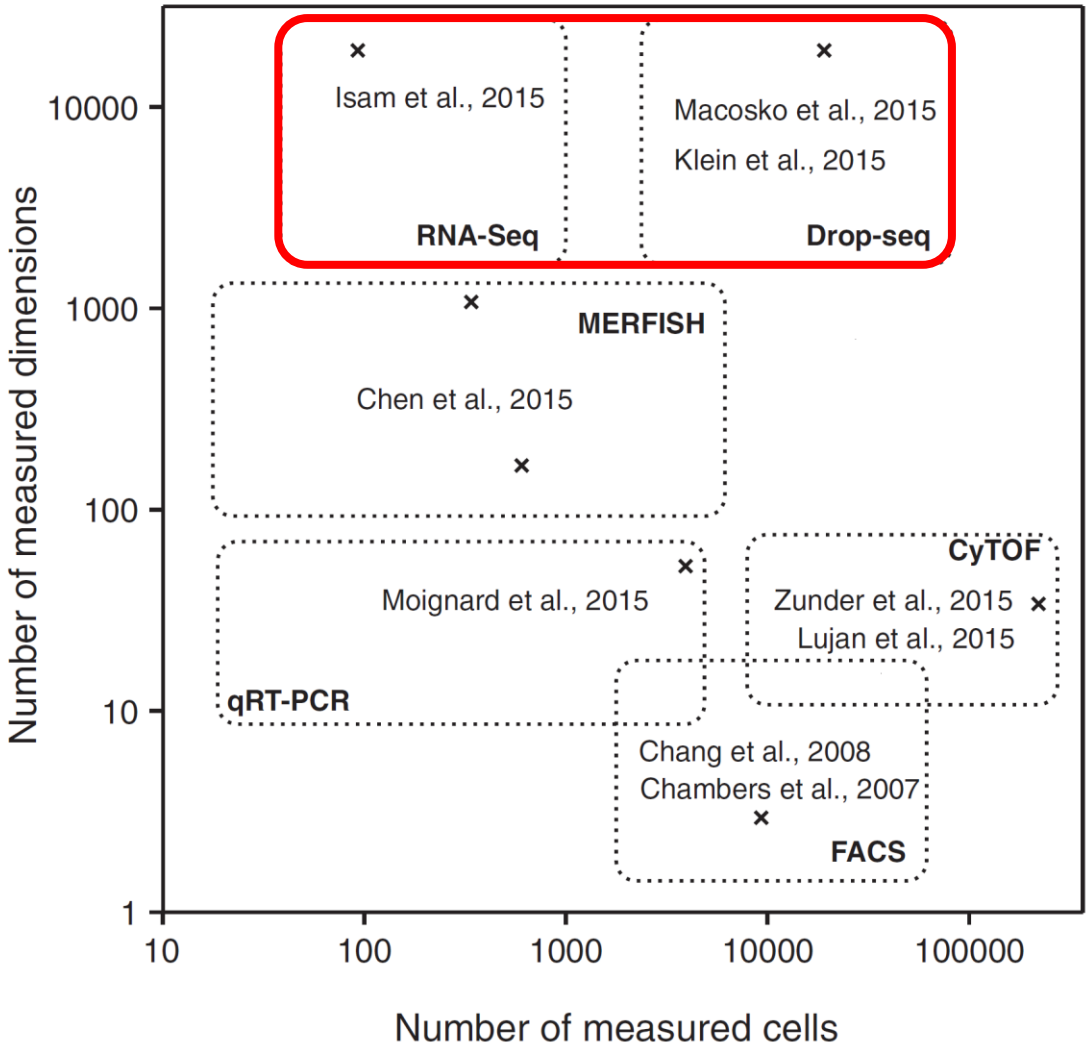
Why single cells?



Simpson's Paradox describes the misleading effects that arise when averaging signals from multiple individuals.

Trapnell (Genome Research 2015)

How can we study single cells?



Every method has it's pros and cons.

Marr et al. *Curr Op Biotech* 2016

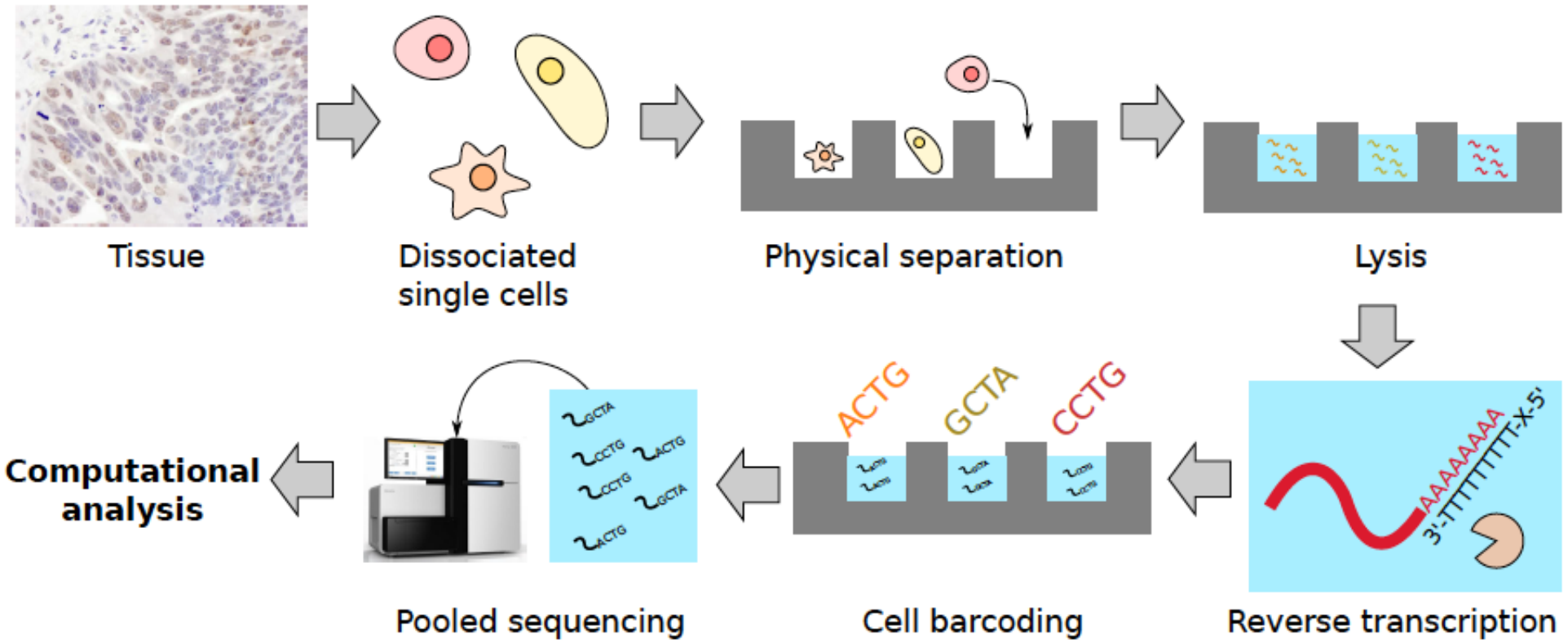
Goals of today

- Introduce single cell RNA-sequencing (scRNA-seq)
- Outline analysis workflow
- Go through (some) steps and discuss best practices

Agenda

09:00 – 10:00	**Lecture part 1:** quality control, normalization, batch correction, feature selection
10:00 – 10:15	---*Break*---
10:15 – 12:30	**Practical part 1:** quality control, normalization, batch correction, feature selection
12:30 – 13:30	---*Lunch*---
13:30 – 14:15	**Lecture part 2:** dimensionality reduction, cell type identification
14:15 – 15:00	**Practical part 2:** dimensionality reduction, cell type identification
15:00 – 15:15	---*Break*---
15:15 – 16:00	**Lecture part 3:** trajectory inference
16:00 – 17:00	**Practical part 3:** trajectory inference

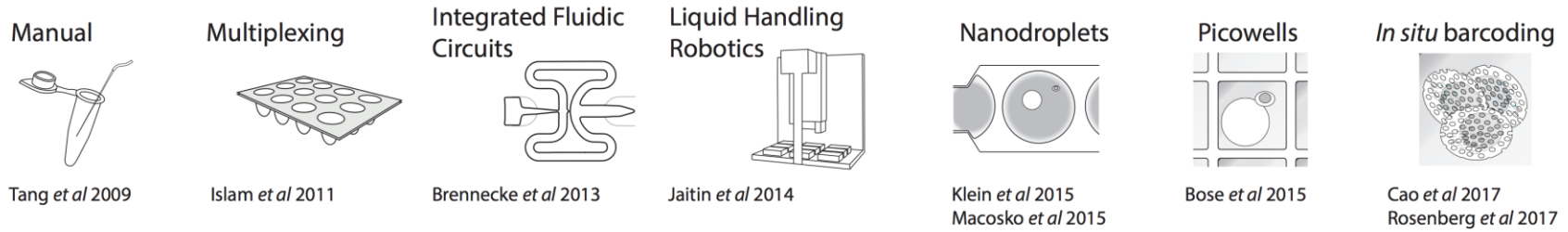
Single cell RNA-sequencing (scRNA-seq)

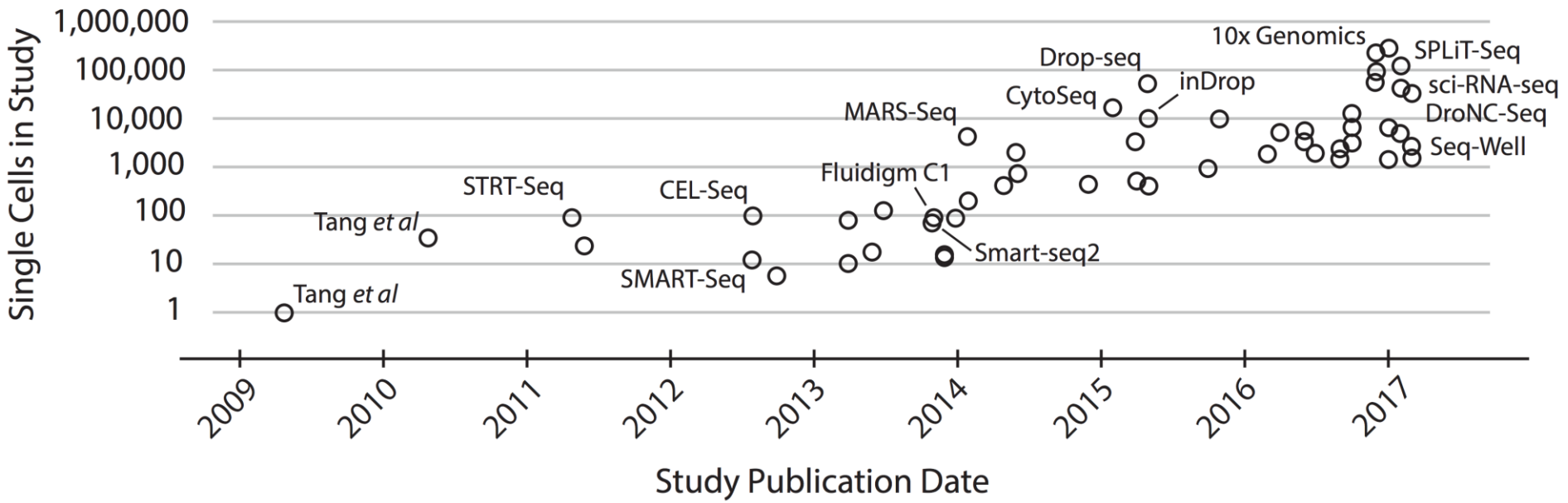


Aaron Lun (Marioni Group, EMBL-EBI)

scRNA-seq Protocols

Number of cells

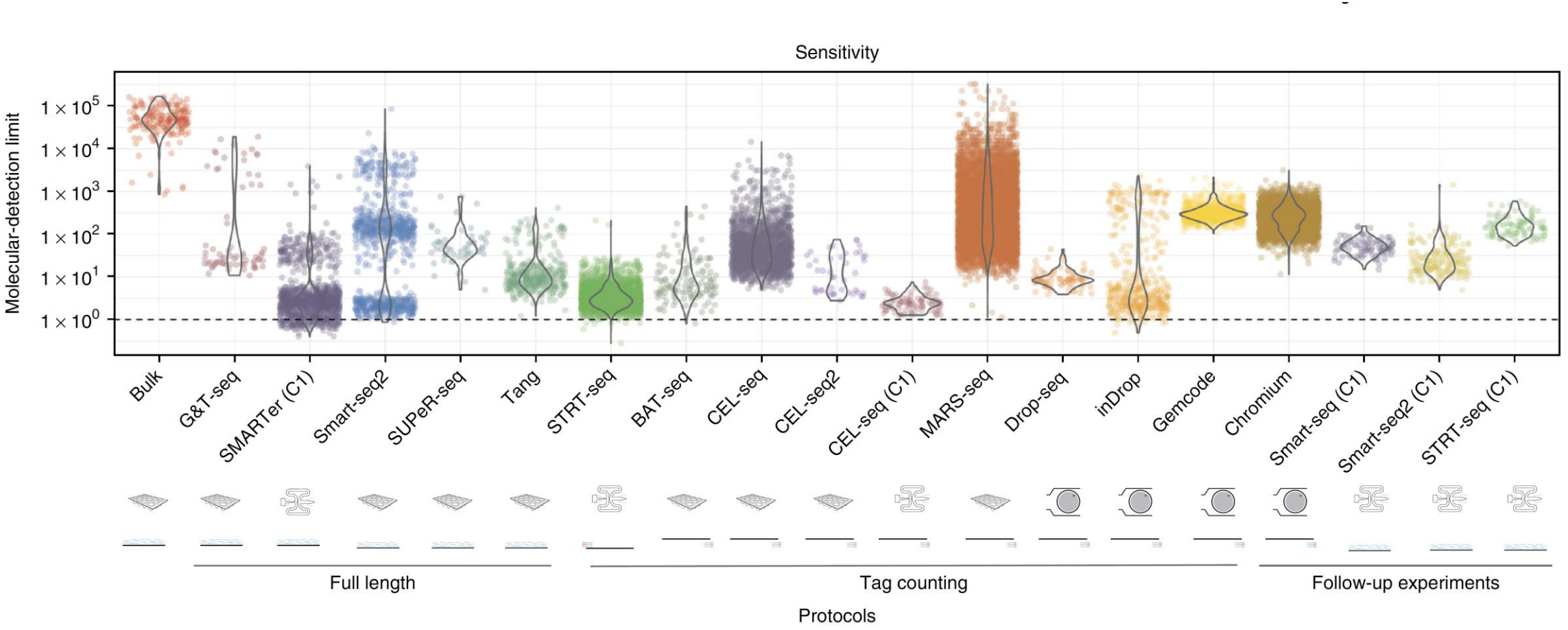




Svensson et al. Nature Methods 2017

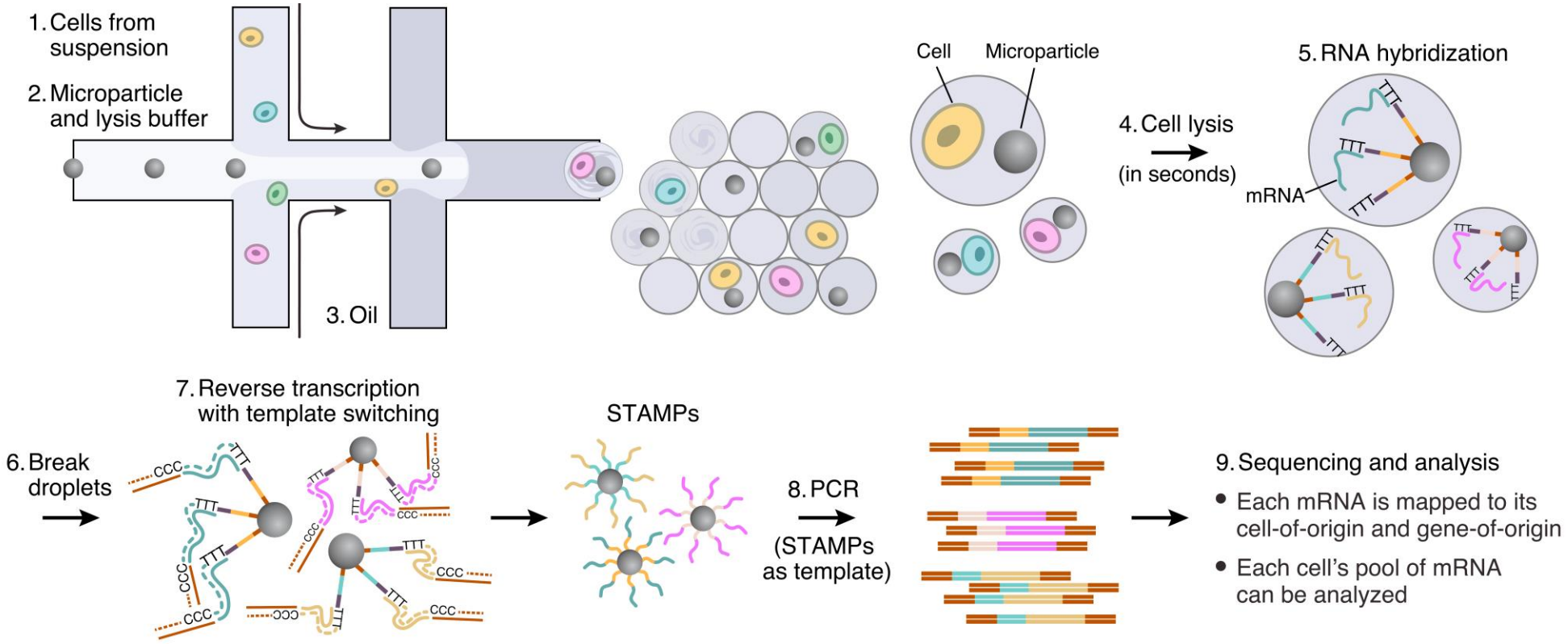
scRNA-seq Protocols

Sensitivity

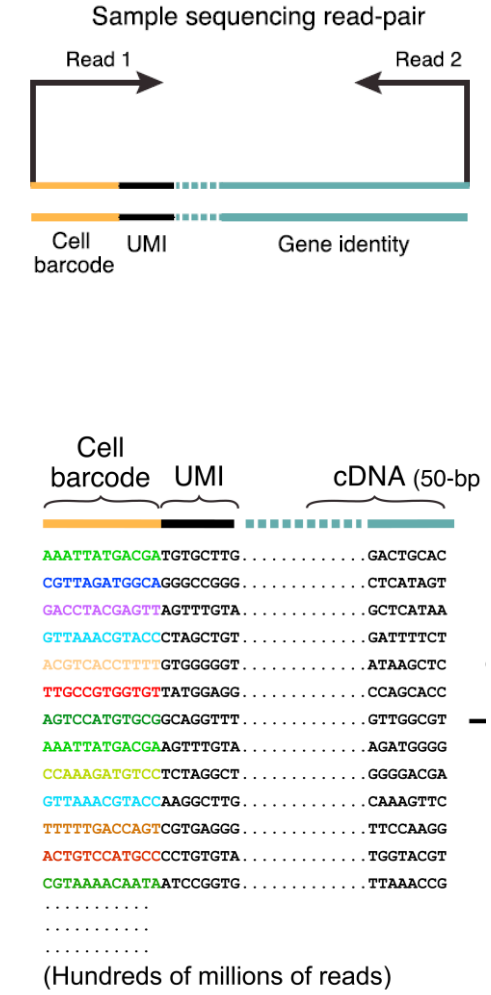


Svensson et al. Nature Methods 2017

Drop-seq



STAMPs: Single-cell transcriptomes attached to microparticles

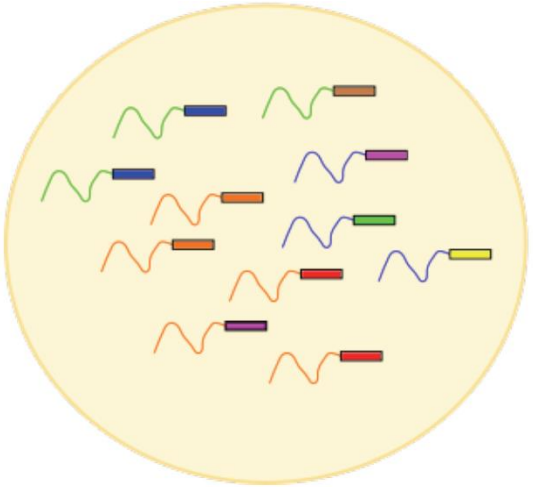


Macosko et al., Cell 2015

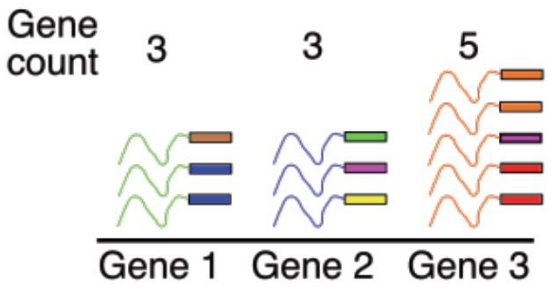
Unique Molecular Identifiers (UMIs)

- 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.

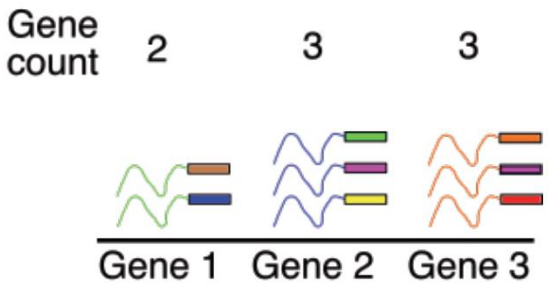
Sequenced fragments from an individual cell



Pre de-duplication



Post de-duplication

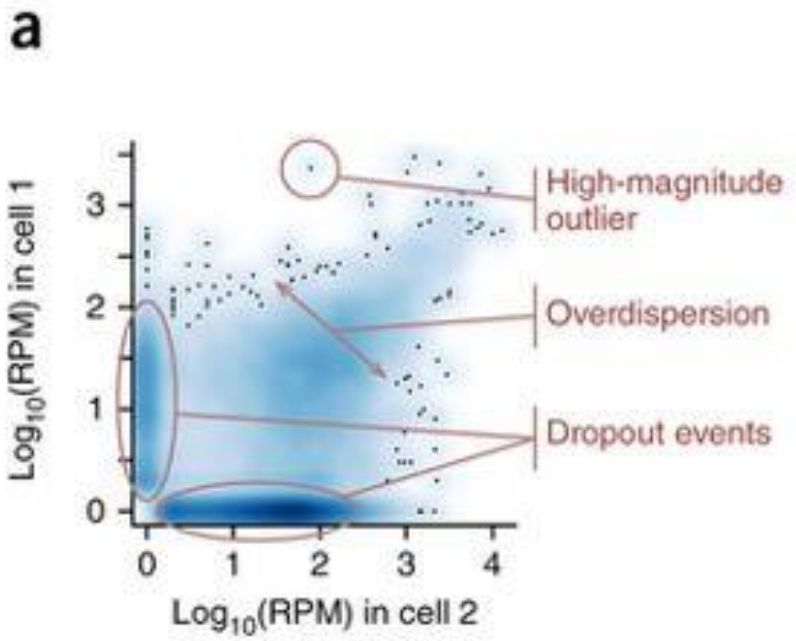


scRNA-seq Data Analysis

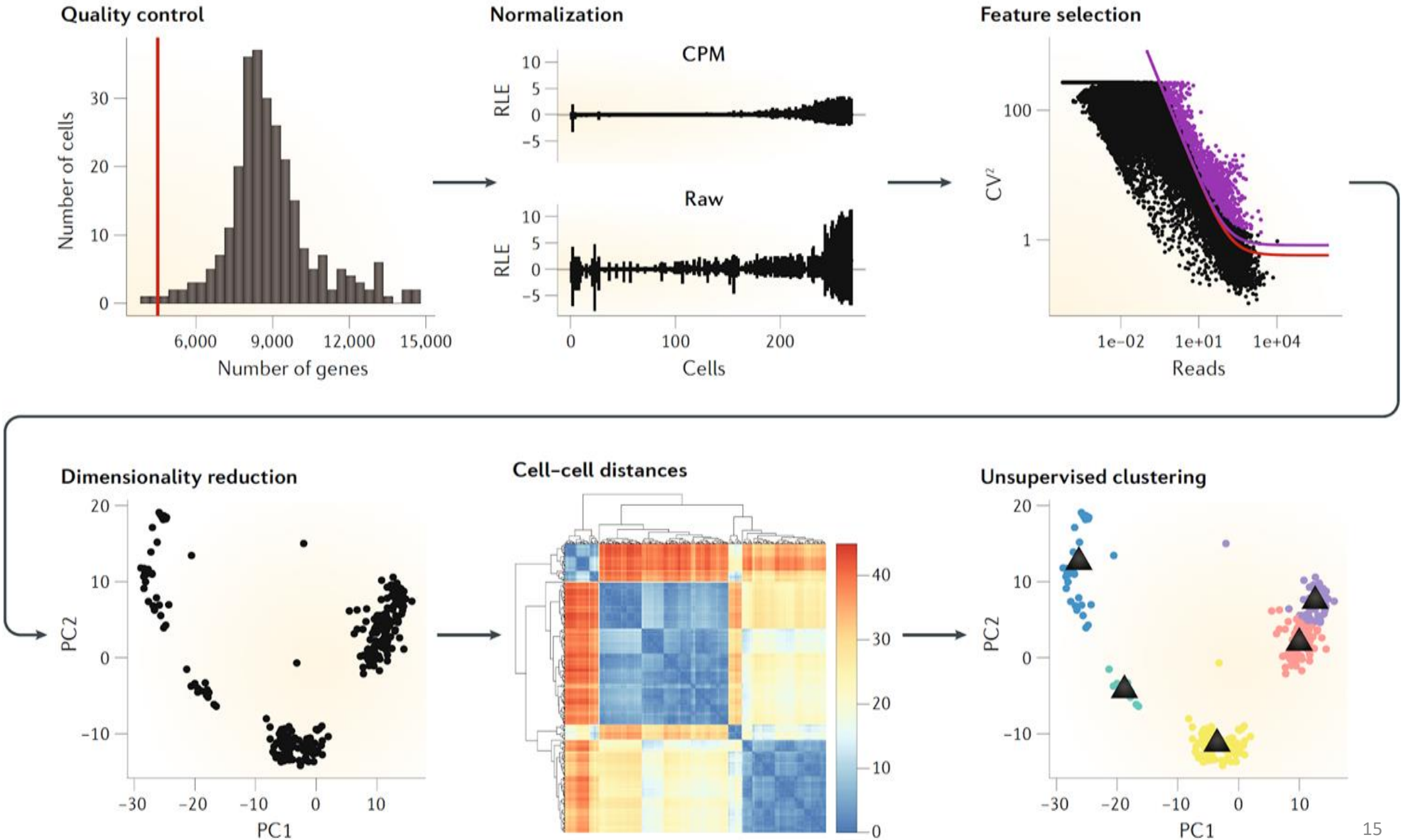
Our goal is to derive/extract real biology from technically noisy data

Problems compared to bulk RNA-seq

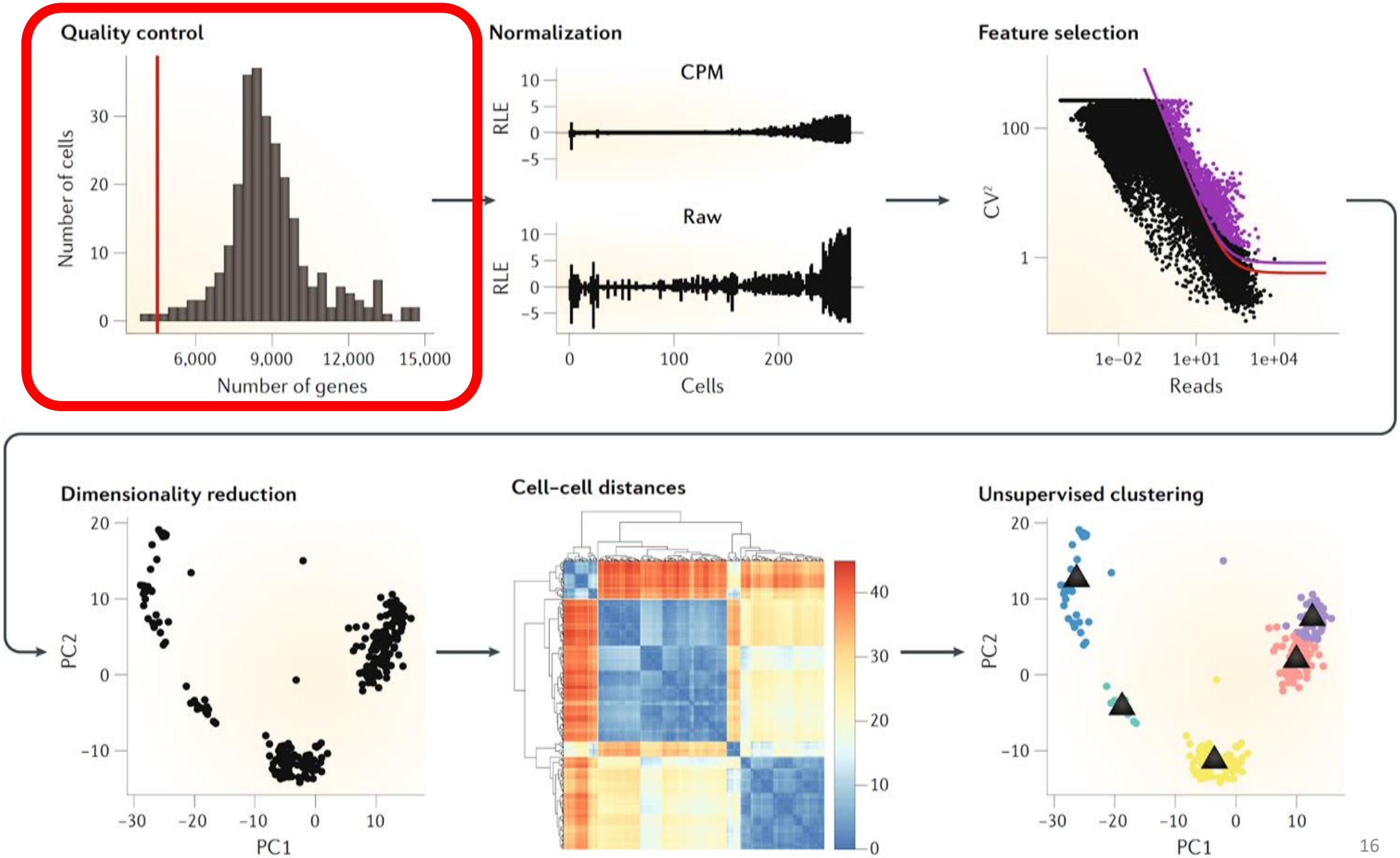
- Amplification bias
- Drop-out rates
- Transcriptional bursting
- Background noise
- Bias due to cell-cycle, cell size and other factors
- Often clear batch effects



Karchenko et al. (Nature Methods 2014)



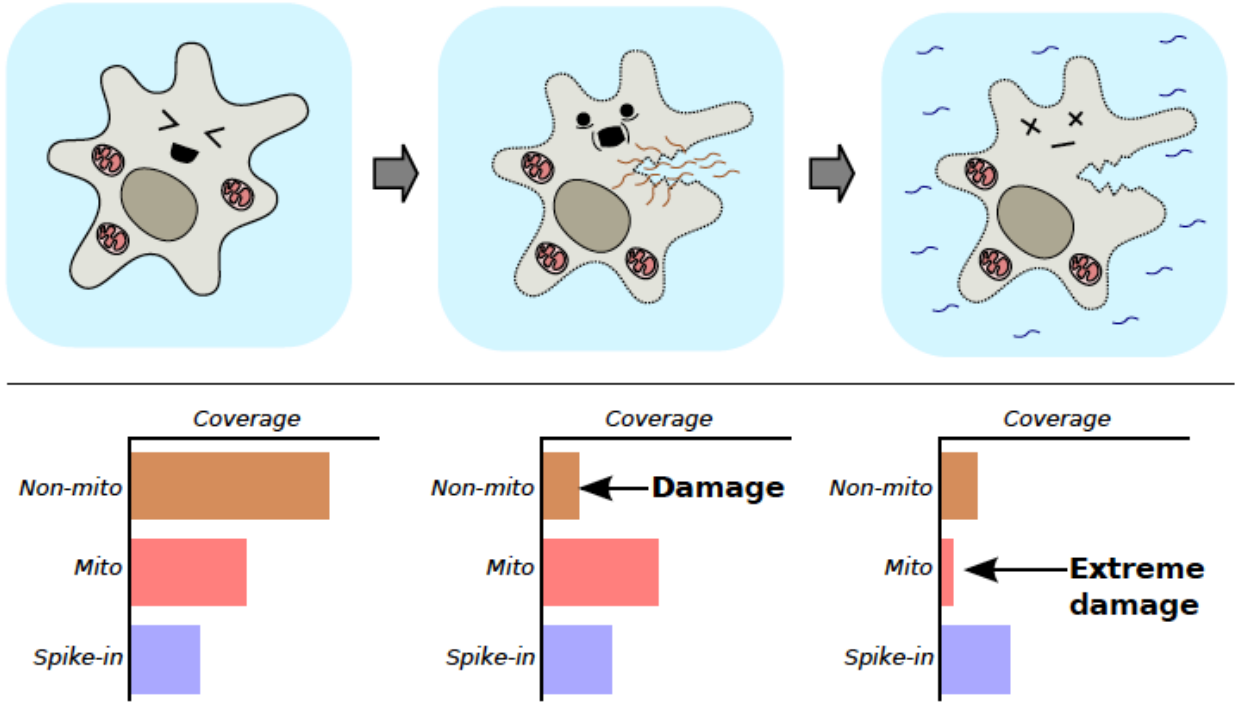
Kiselev et al. (Nature review Genetics 2019)



Kiselev et al. (Nature review Genetics 2019)

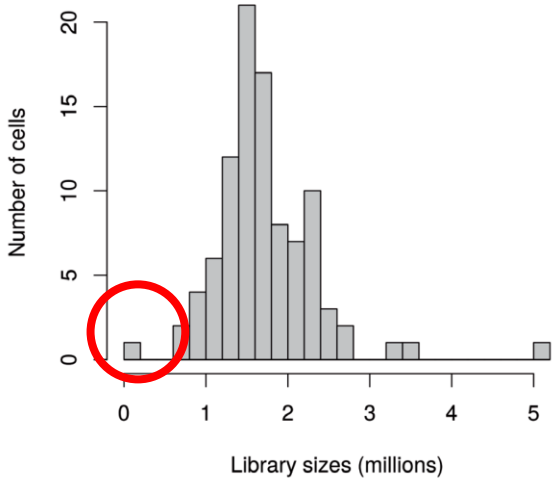
Quality control of cells (1)

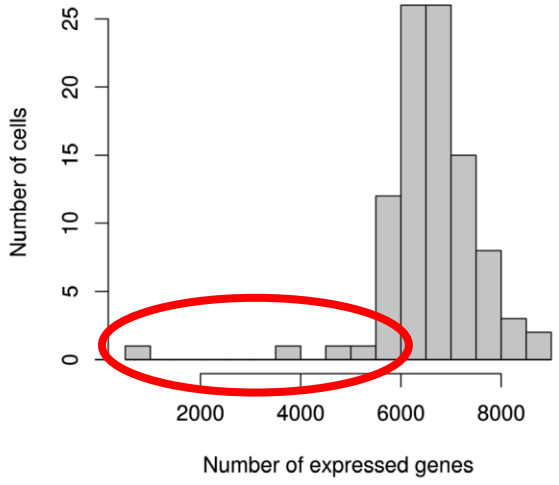
- Low sequencing depth
- Low numbers of expressed genes (i.e. any nonzero count)
- High spike-in (if present) or mitochondrial content



Quality control of cells (2)

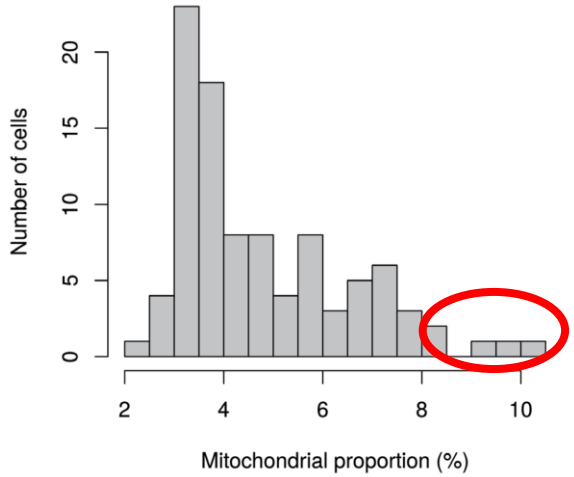
RNA has not been efficiently captured during library preparation

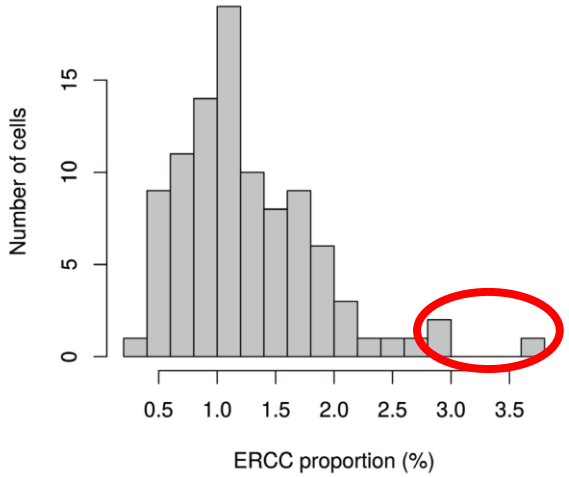




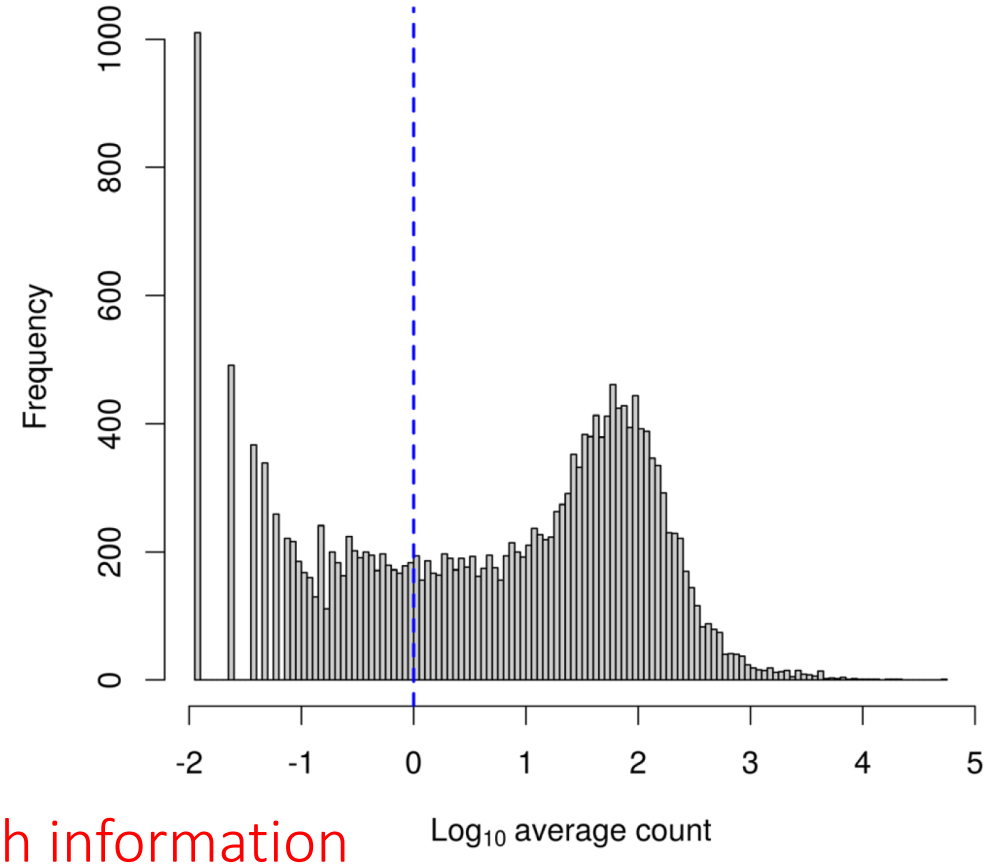
Diverse transcript population not captured

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

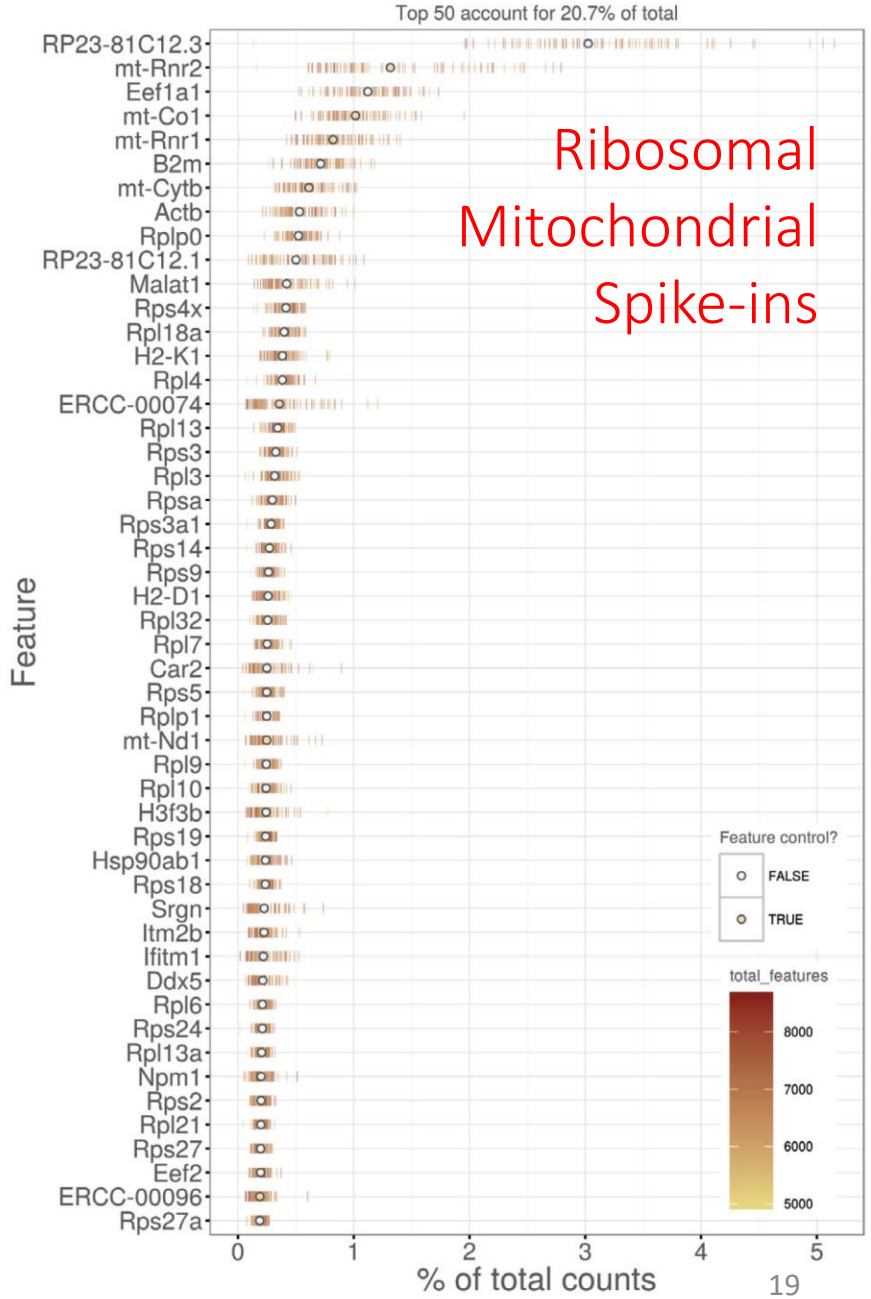




Quality control of genes



Not enough information for reliable statistical inference



Ribosomal
Mitochondrial
Spike-ins

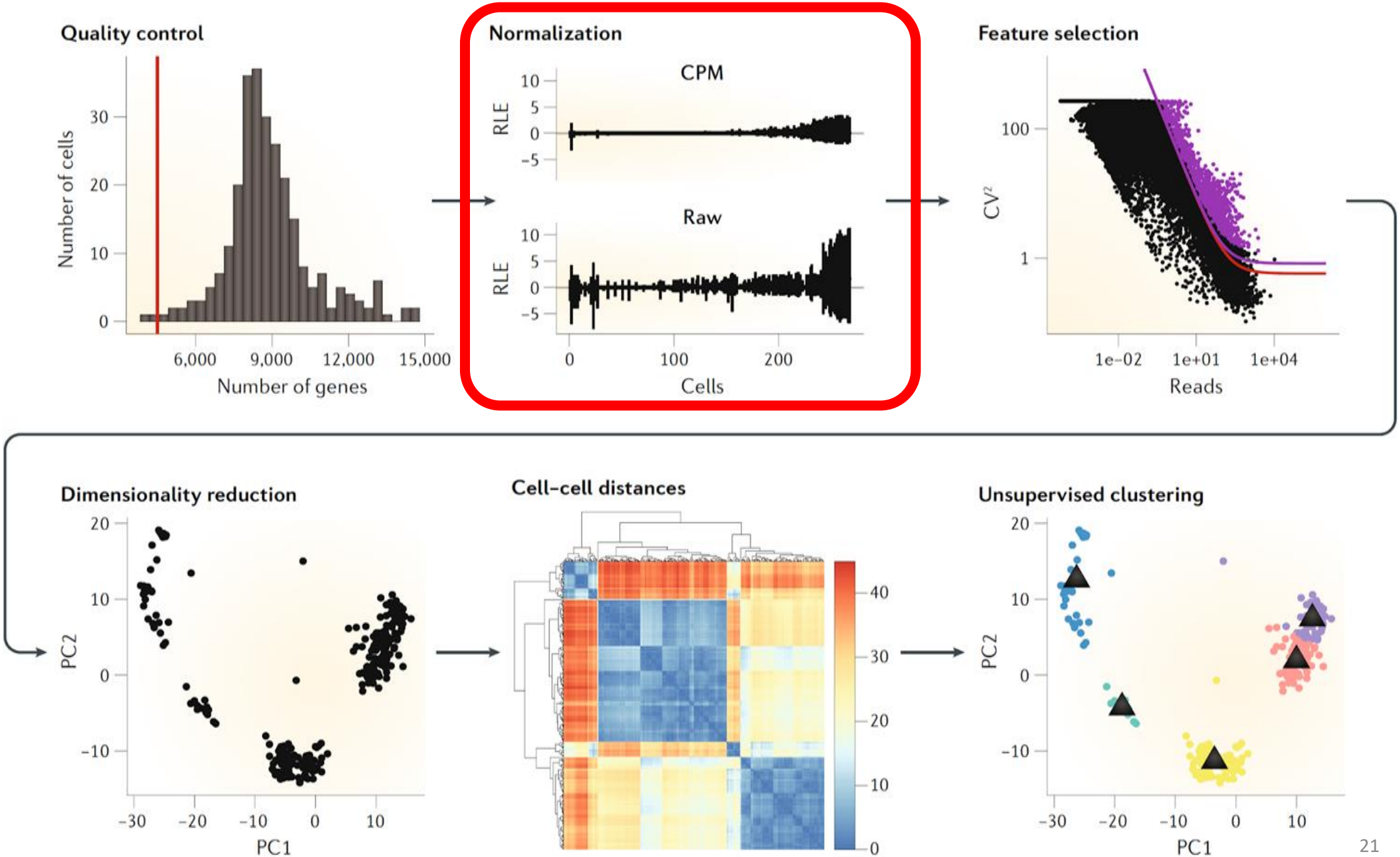
QC (pitfalls and recommendations)

- Perform QC by finding outlier peaks in the number of genes, the count depth and the fraction of mitochondrial reads. Consider these covariates jointly instead of separately.

- Be as permissive of QC thresholding as possible, and revisit QC if downstream clustering cannot be interpreted.

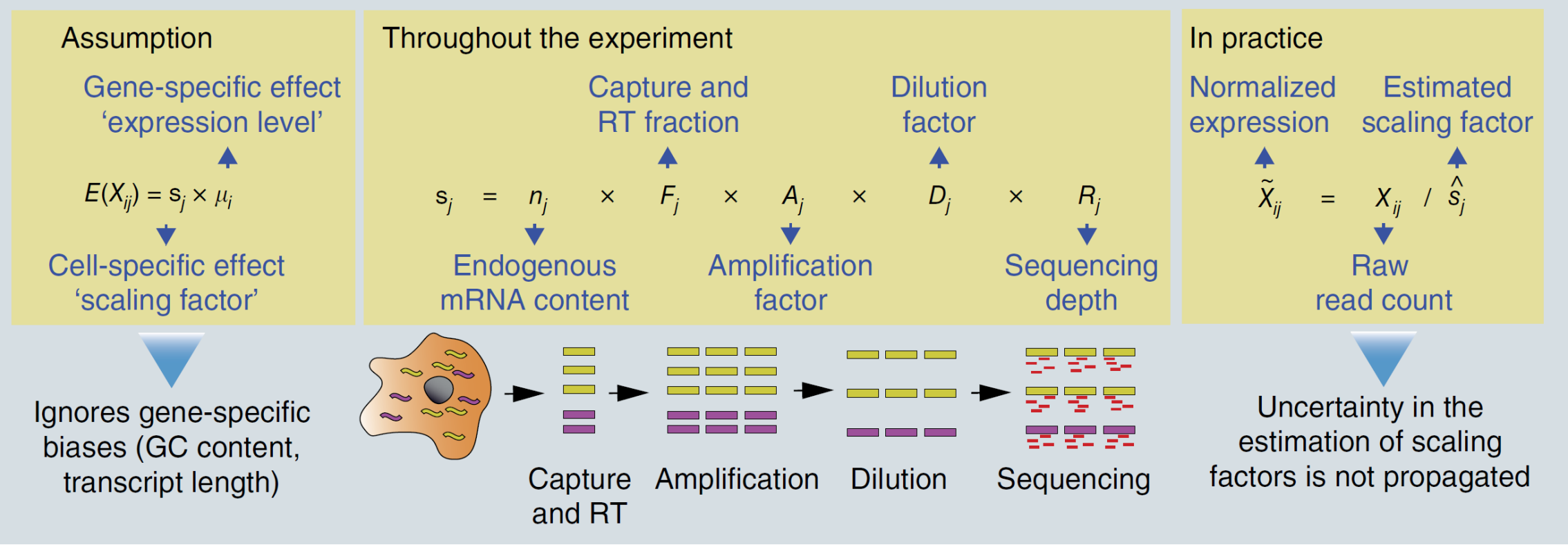
- If the distribution of QC covariates differ between samples, QC thresholds should be determined separately for each sample to account for sample quality differences as in Plasschaert et al (2018).

Luecken and Theis (MSB 2019)



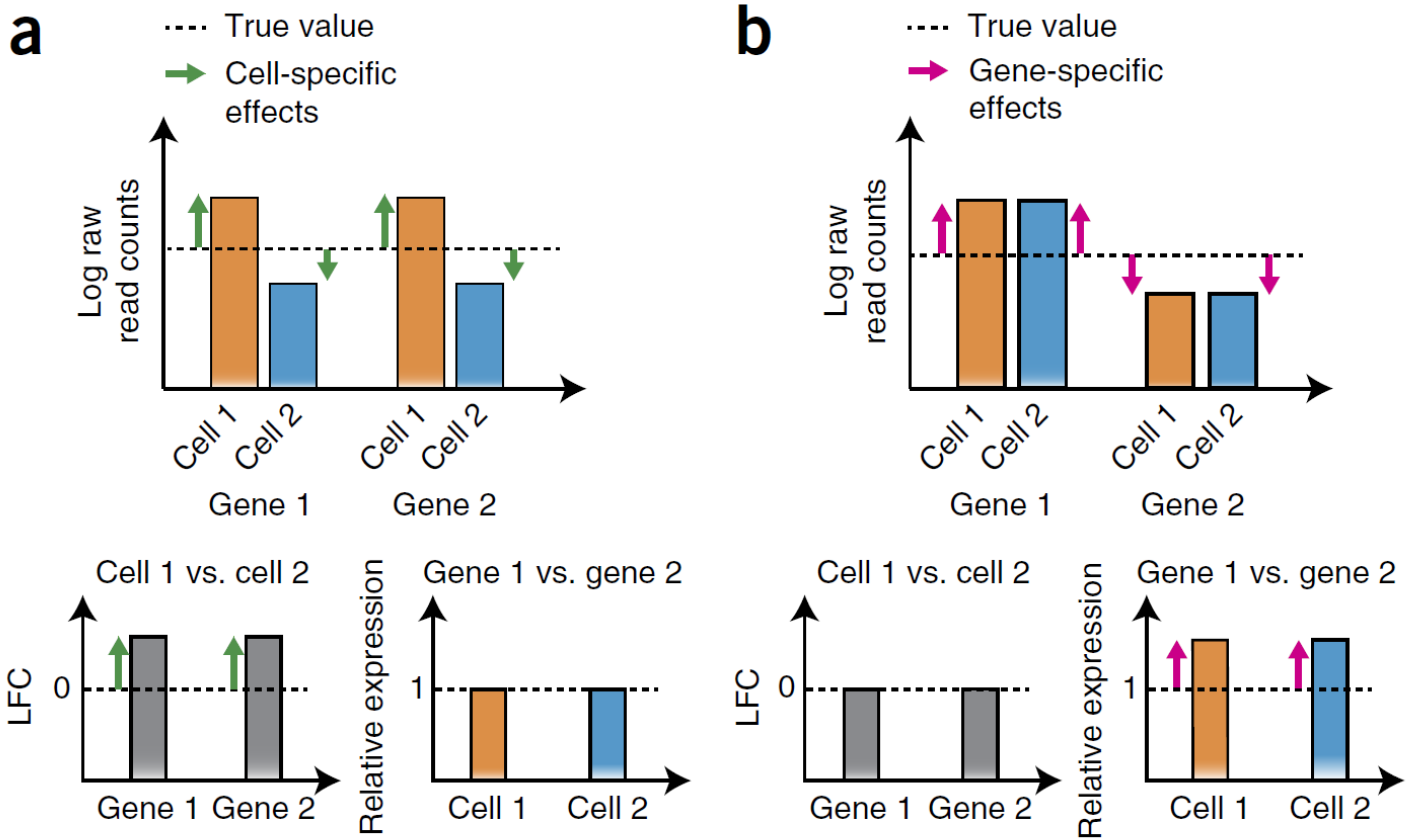
Kiselev et al. (Nature review Genetics 2019)

Normalization (1)



Vallejos et al. (Nat Methods 2017)

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



Vallejos et al. (Nat Methods 2017)

Which effects are removed by UMIs?

c

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

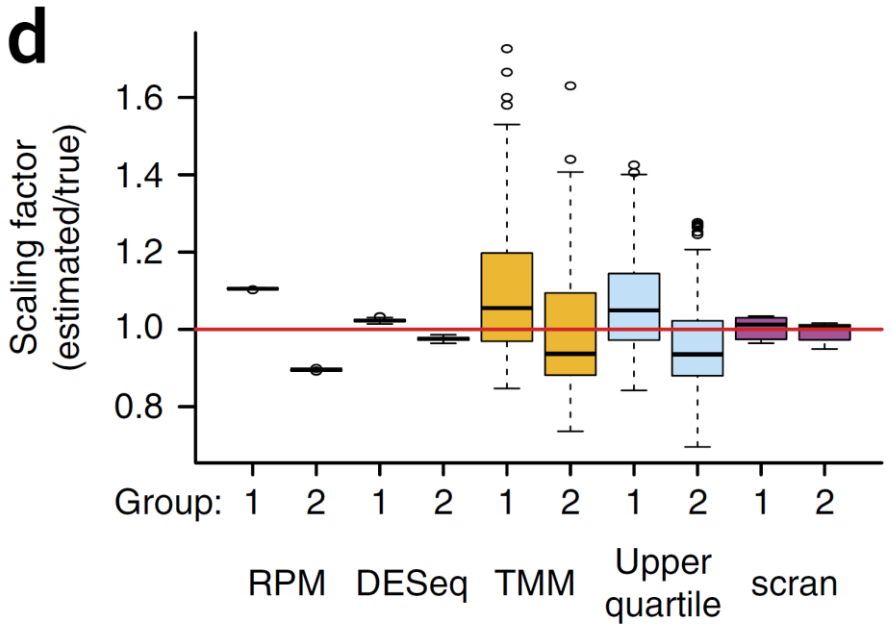
Normalization (2)

- 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

Vallejos et al. (Nat Methods 2017)

What is different from bulk RNA-seq?

- Noise
  - Low mRNA content per cell
  - Variable mRNA capture
  - Variable sequencing depth
- Different cell types in the same sample
- Bulk RNA-seq normalization methods (FPKM, CPM, TPM, upperquartile) are based on per-gene statistics –> not suitable for zero-inflated data



Vallejos et al. (Nat Methods 2017)

Normalization methods

1. Size factor scaling methods

2. Probabilistic methods (Zero-inflated negative binomial (ZINB) models).

   E.g. ZINB-WaVE, Risso et al. (Nature Comm 2018).

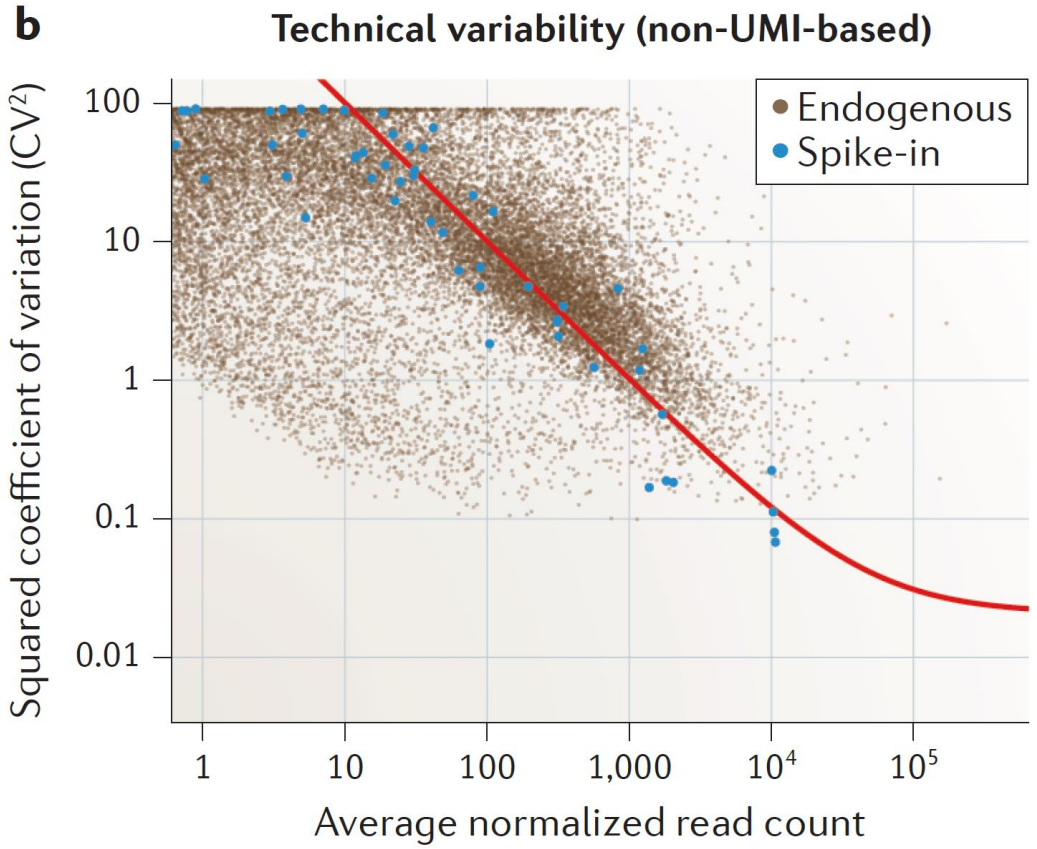
Size factor scaling methods

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

Using Spike-In RNA

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 (not in droplet-based)



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]

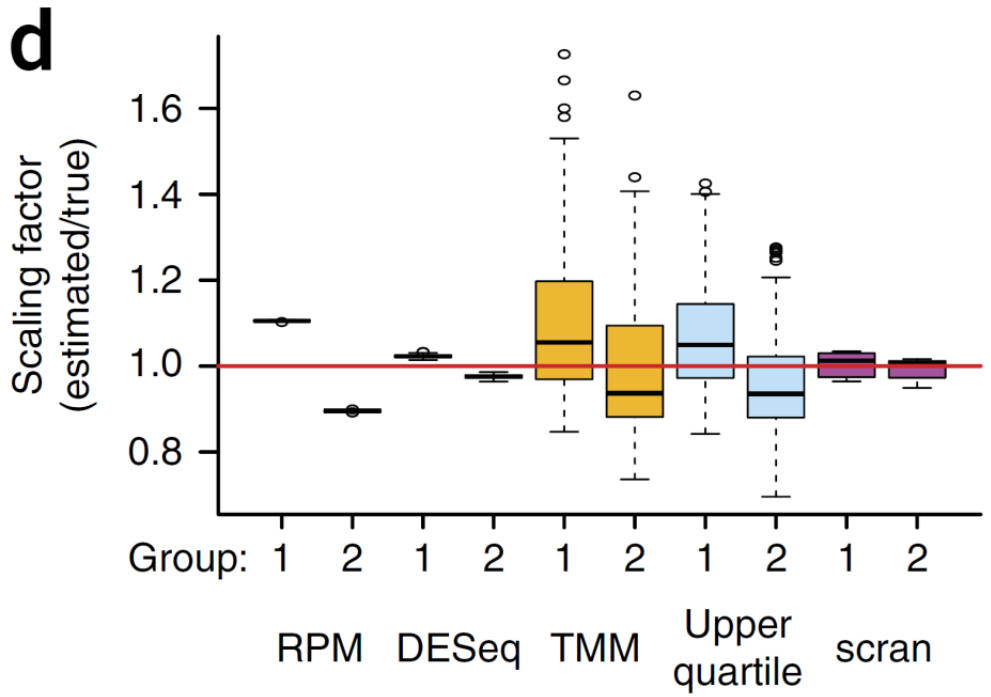
[1]*Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom;* [2]*Wellcome Trust and MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge CB2 0XY, United Kingdom;* [3]*EMBL European Bioinformatics Institute, Wellcome Genome 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 *(NOT APPROPRIATE)*
- Log normalization (Seurat)
- Negative binomial (Monocole)
- Zero-inflated negative binomial (ZINB) models
- …



Performance Assessment and Selection of Normalization Procedures for Single-Cell RNA-Seq
Cole et al, Cell Systems 2019

Vallejos et al. (Nat Methods 2017)

Normalization (pitfalls and recommendations)

- We recommend scran for normalization of non-full-length datasets. An alternative is to evaluate normalization approaches via scone especially for plate-based datasets. Full-length scRNA-seq protocols can be corrected for gene length using bulk methods.

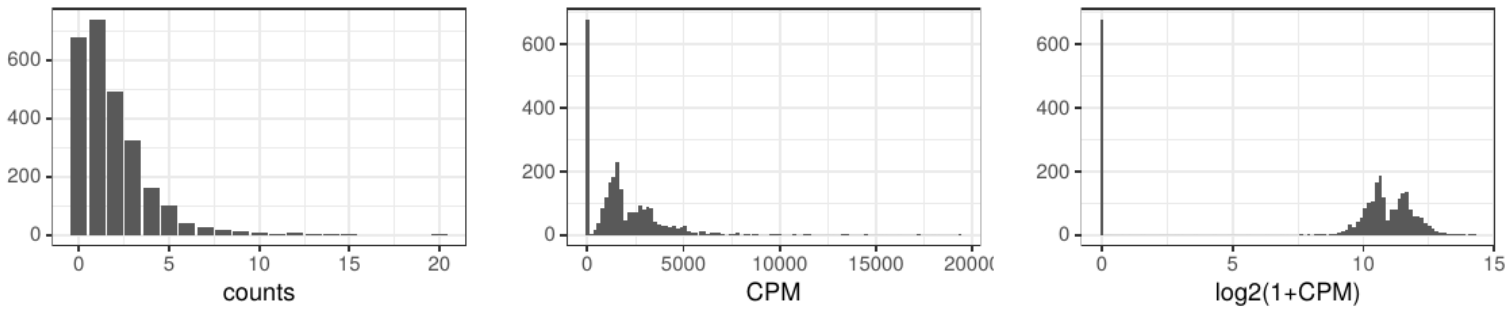
- There is no consensus on scaling genes to 0 mean and unit variance. We prefer not to scale gene expression.

- Normalized data should be log(x+1)-transformed for use with downstream analysis methods that assume data are normally distributed.

Luecken and Theis (MSB 2019)

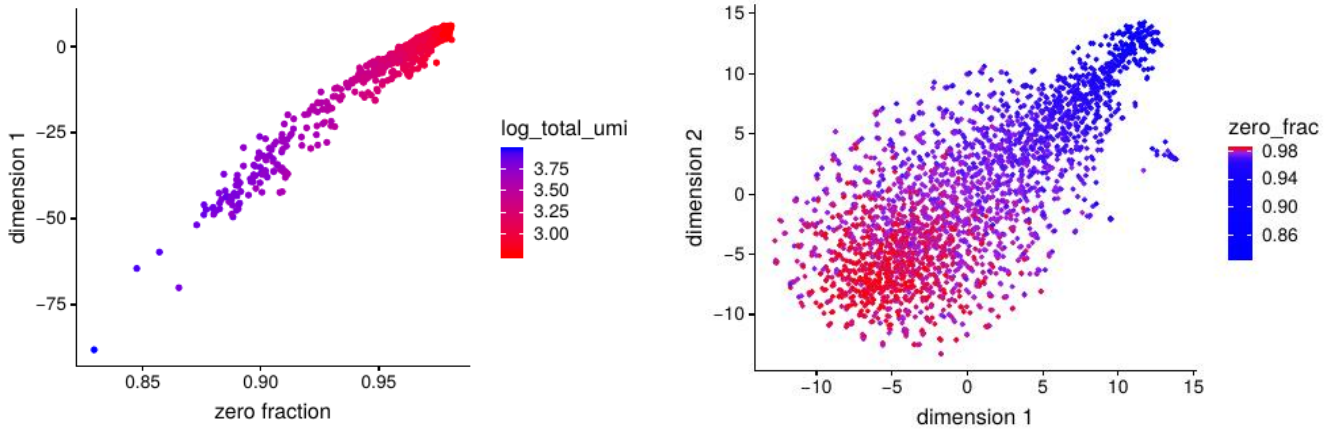
Effect of dropouts on normalization

Inflation of zero counts



(a) UMI counts (b) counts per million (CPM) (c) log of CPM

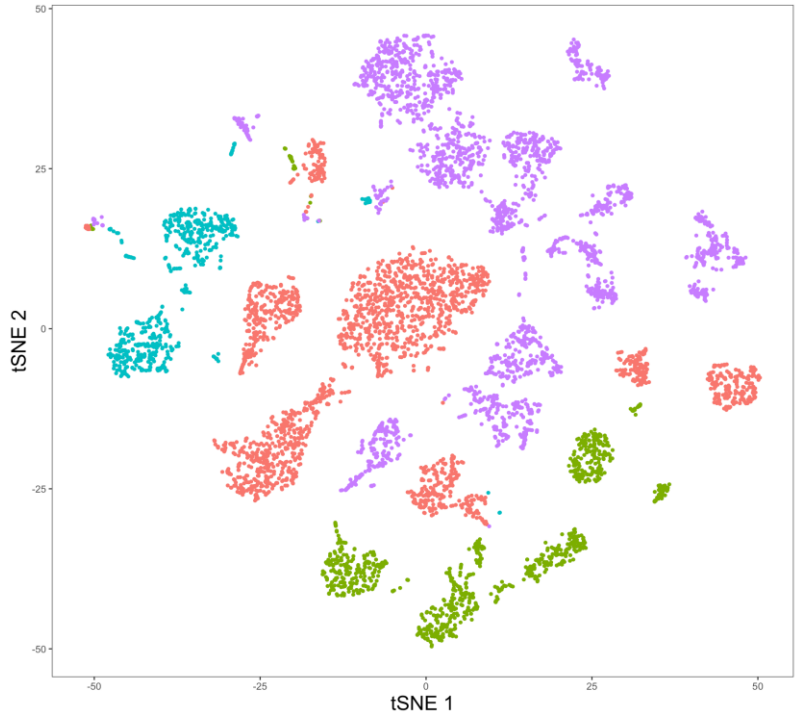
Fraction of zeros become main source of variability



Towns et al. (bioRxiv 2019); Svensson (bioRxiv 2019)

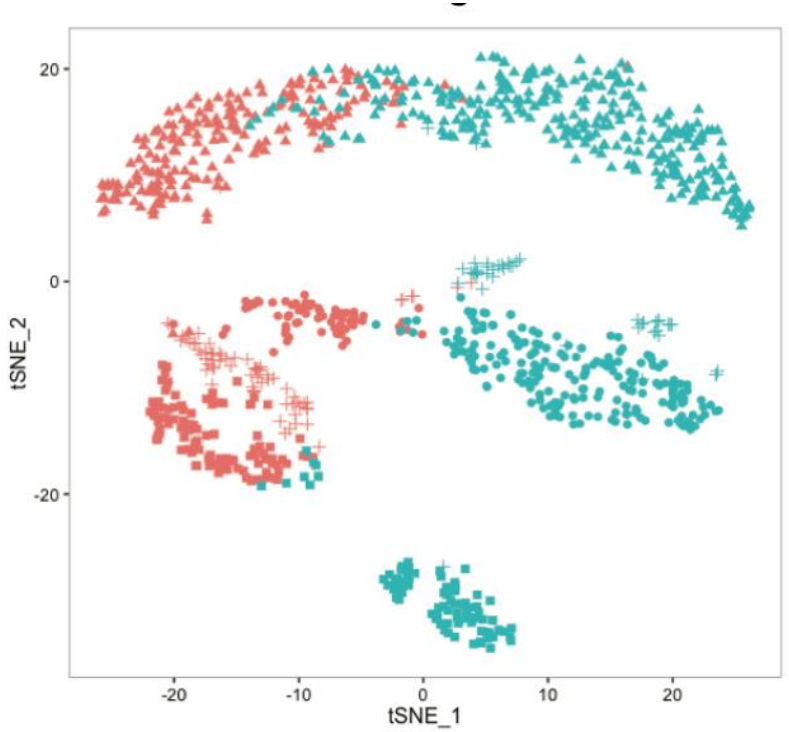
Batch correction

Why integrate?



Donor 1
Donor 2
Donor 3
Donor 4

Same tissue from different donors



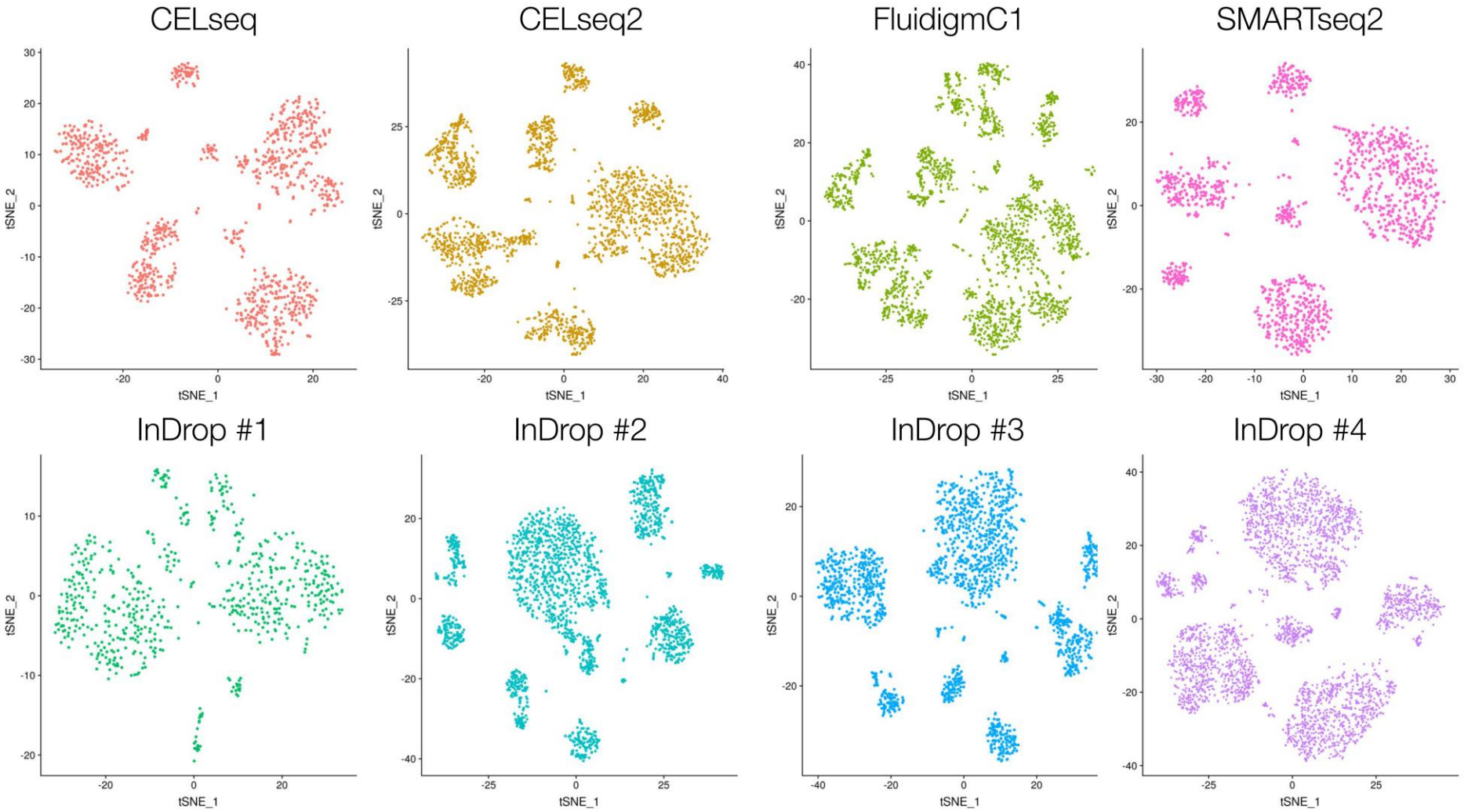
Sick
Healthy

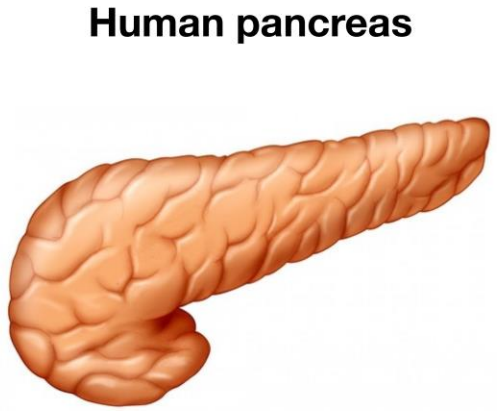
▲ Cell Type 1
■ Cell Type 2
● Cell Type 3

Cross condition comparisons

Building a cell atlas

8 maps of the human pancreas





Baron et al. 2016, *Cell Syst.*

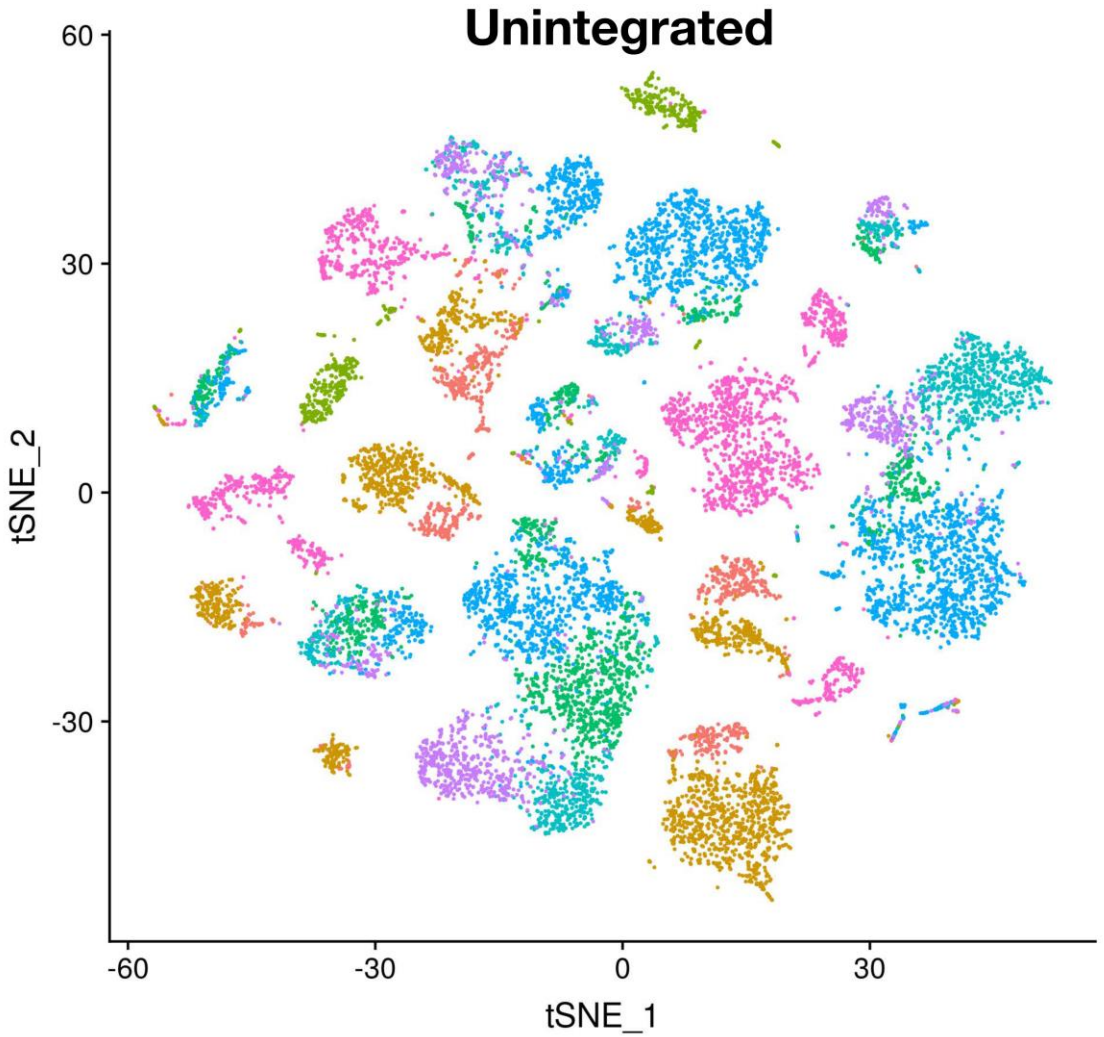
Lawlor et al. 2017, *Genome Res.*

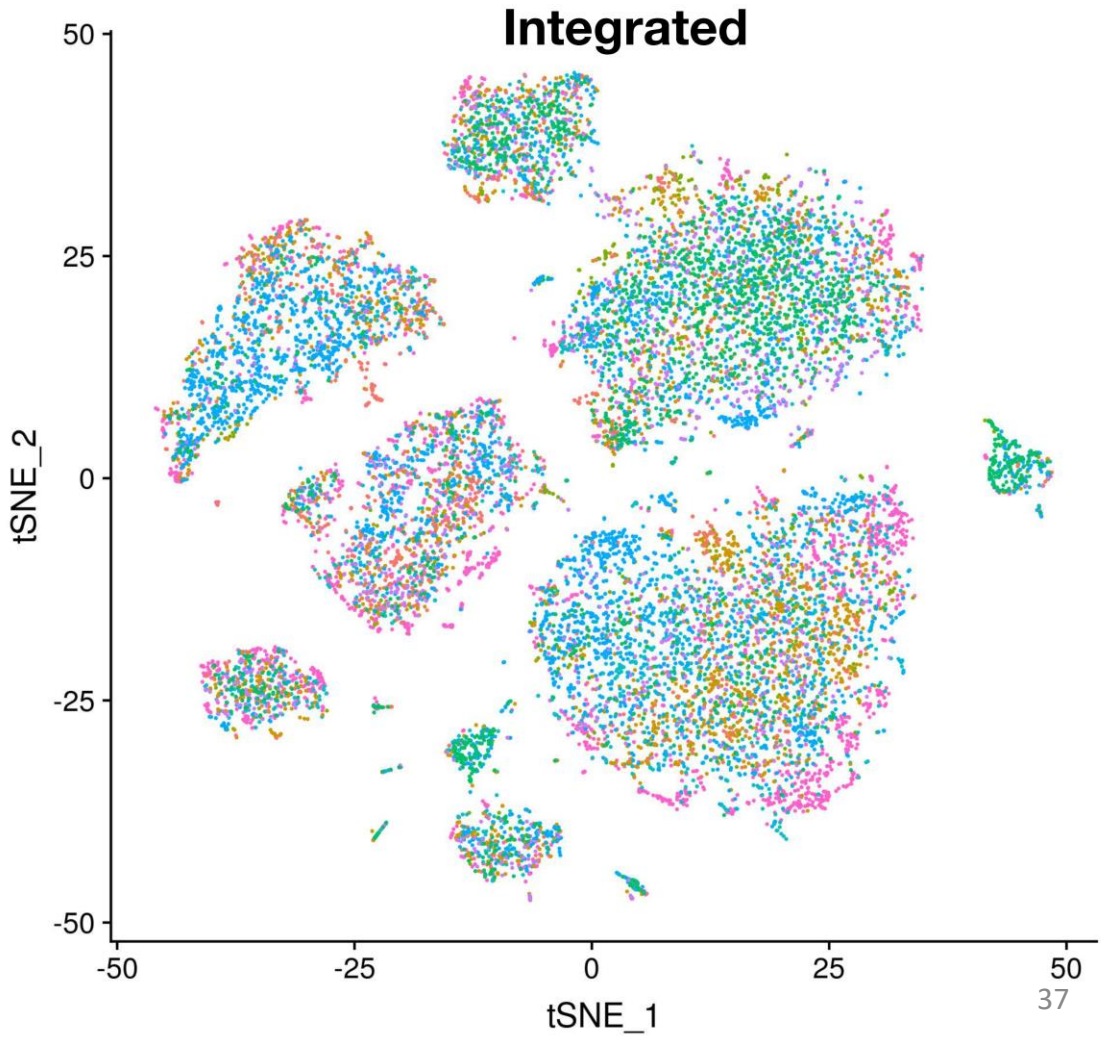
Grun et al. 2016, *Cell Stem Cell*

Muraro et al. 2016, *Cell Syst.*

Building a cell atlas

8 maps of the human pancreas

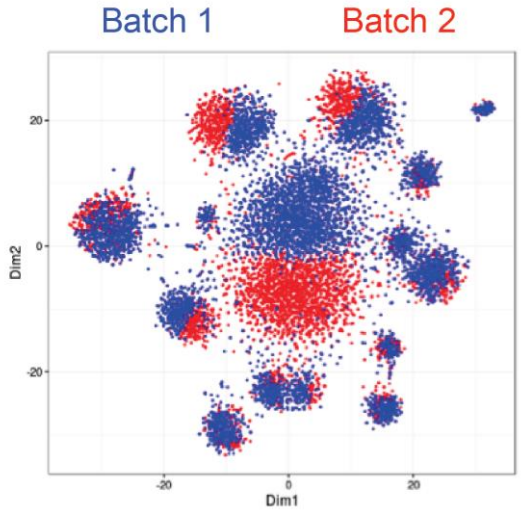




Confounders and batch effects (1)

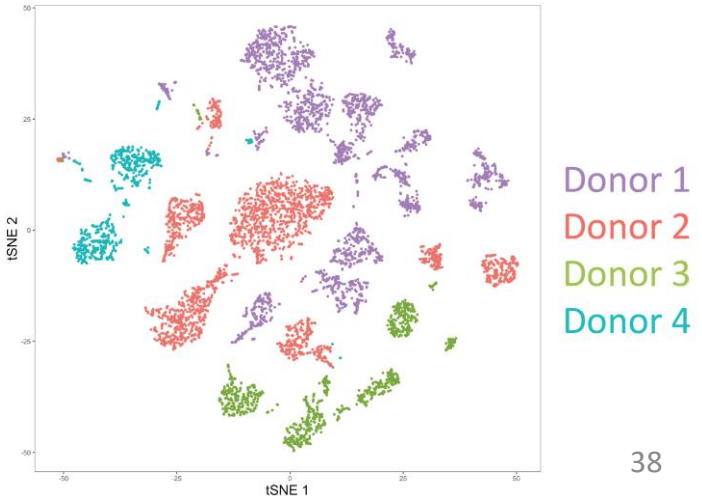
1. Technical variability
   - Changes in sample quality/processing
   - Library prep or sequencing technology
   - ‘Experimental reality’

Technical ‘batch effects’ confound downstream analysis



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

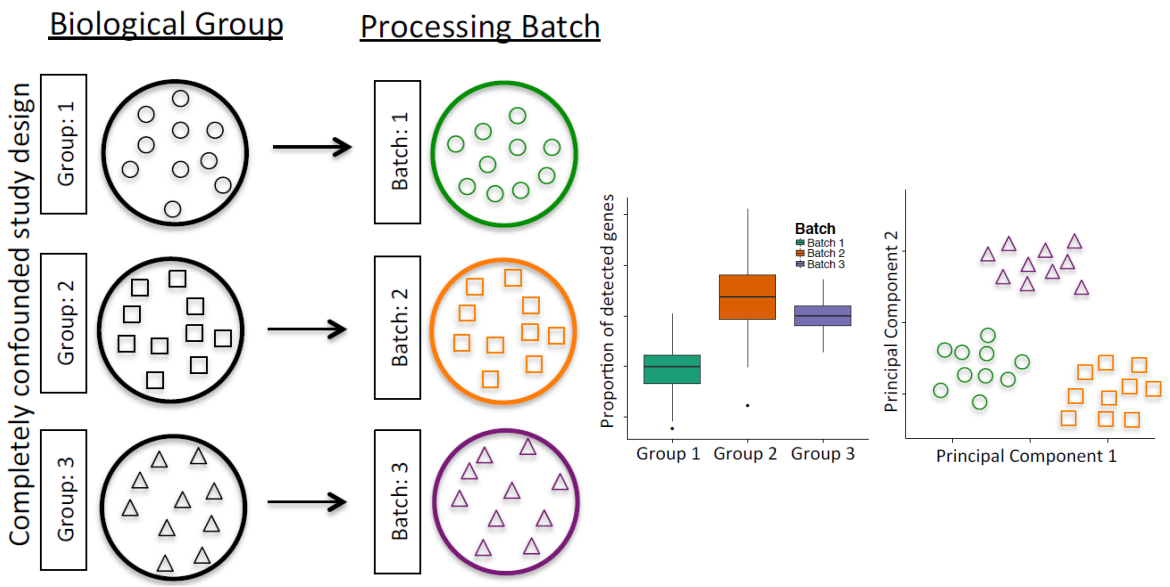
Biological ‘batch effects’ confound comparisons of scRNA-seq data



Shaham et al. (Bioinformatics 2017)

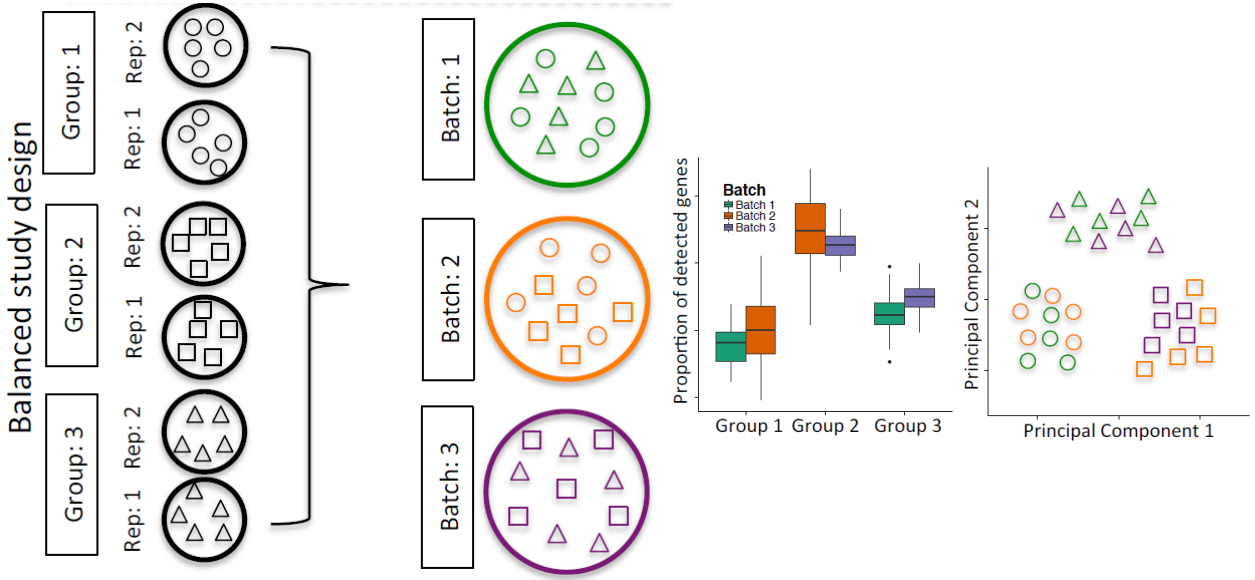
Confounders and batch effects (2)

Confounded design



Don’t design your experiment like this!!!

Not confounded design



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

Hicks et al. (bioRxiv 2015)

Batch correction methods

- Many good options have been developed for bulk RNA-seq data:
  - RUVseq() or svaseq()
  - Linear models with e.g. removeBatchEffect() in limma or scater
  - ComBat() in sva
  - …

- But bulk RNA-seq methods make modelling assumptions that are likely to be violated in scRNAseq data
  - The composition of cell populations are either known or the same across batches
  - Batch effect is additive: batch-induced fold-change in expression is the same across different cell subpopulations for any given gene

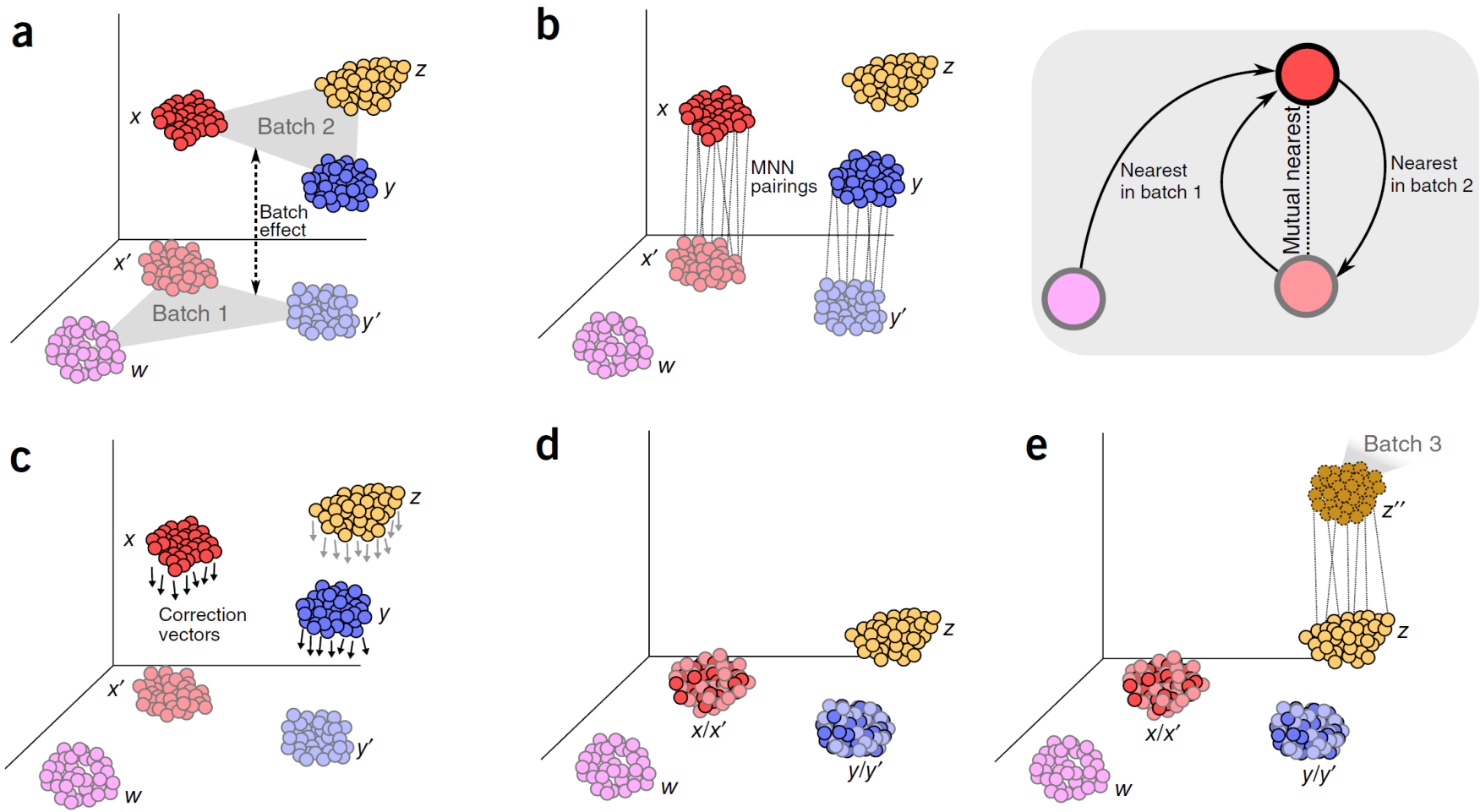
Batch correction methods

- MNNcorrect (https://doi.org/10.1038/nbt.4091)
- CCA + anchors (Seurat v3) (https://doi.org/10.1101/460147)
- CCA + dynamic time warping (Seurat v2) (https://doi.org/10.1038/nbt.4096)
- LIGER (https://doi.org/10.1101/459891)
- Harmony (https://doi.org/10.1101/461954)
- Conos (https://doi.org/10.1101/460246)
- Scanorama (https://doi.org/10.1101/371179)
- scMerge (https://doi.org/10.1073/pnas.1820006116)
- …

Two broad strategies:

- Joint dimension reduction
- Graph-based joint clustering

Mutual Nearest Neighbors (MNN)

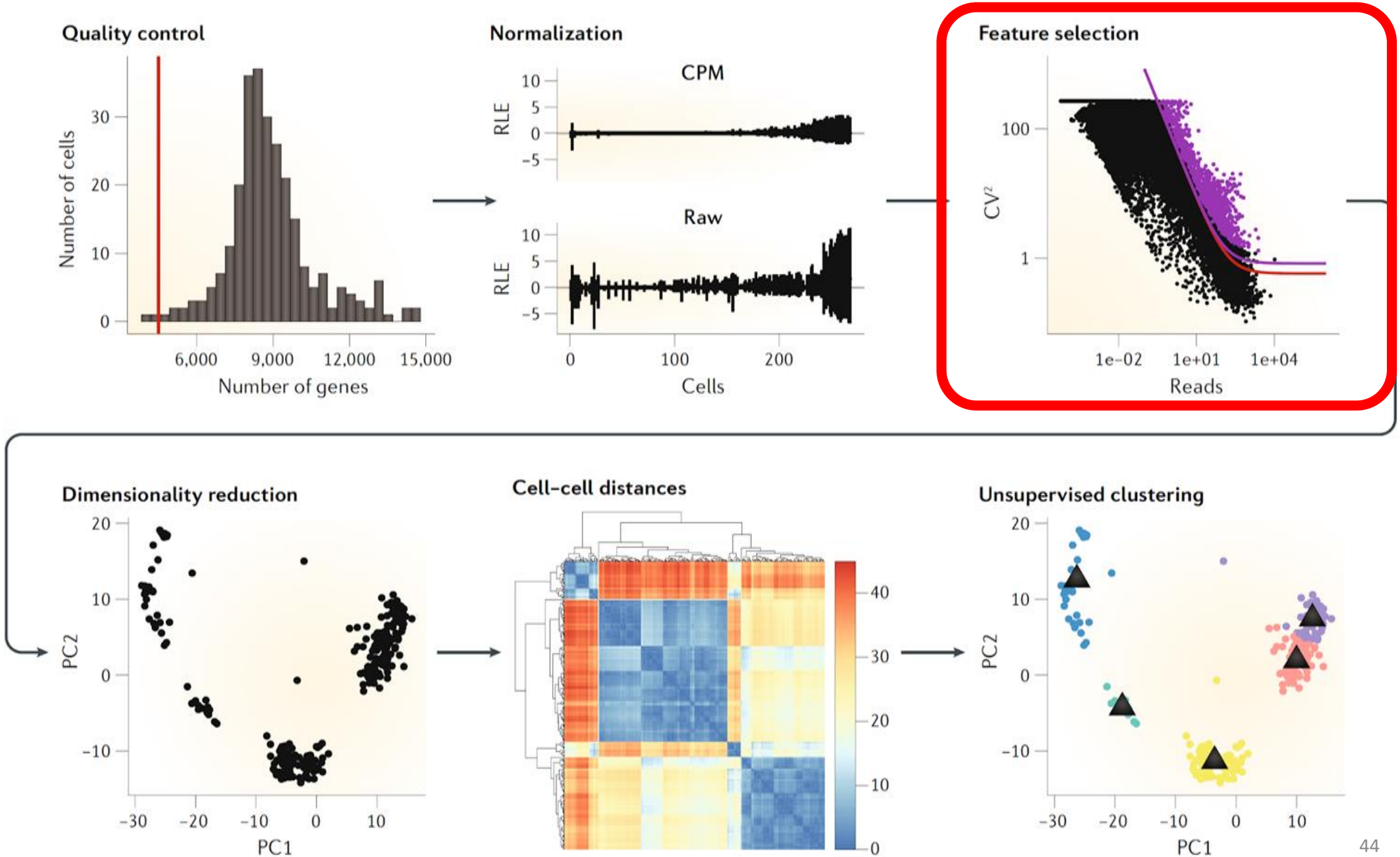


Haghverdi et al., (Nature Biotech 2017)

Using the corrected values

- Batch correction facilitates cell-based analysis of population heterogeneity in a consistent manner across batches.
  - No need to identify mappings between separate clusterings
  - Increased number of cells allows for greater resolution of population structure

- BUT…
- It is not recommended to use the corrected expression values for gene-based analyses (e.g. differential expression)
- Arbitrary correction algorithms are not obliged to preserve the magnitude (or even direction) of differences in per-gene expression when attempting to align multiple batches



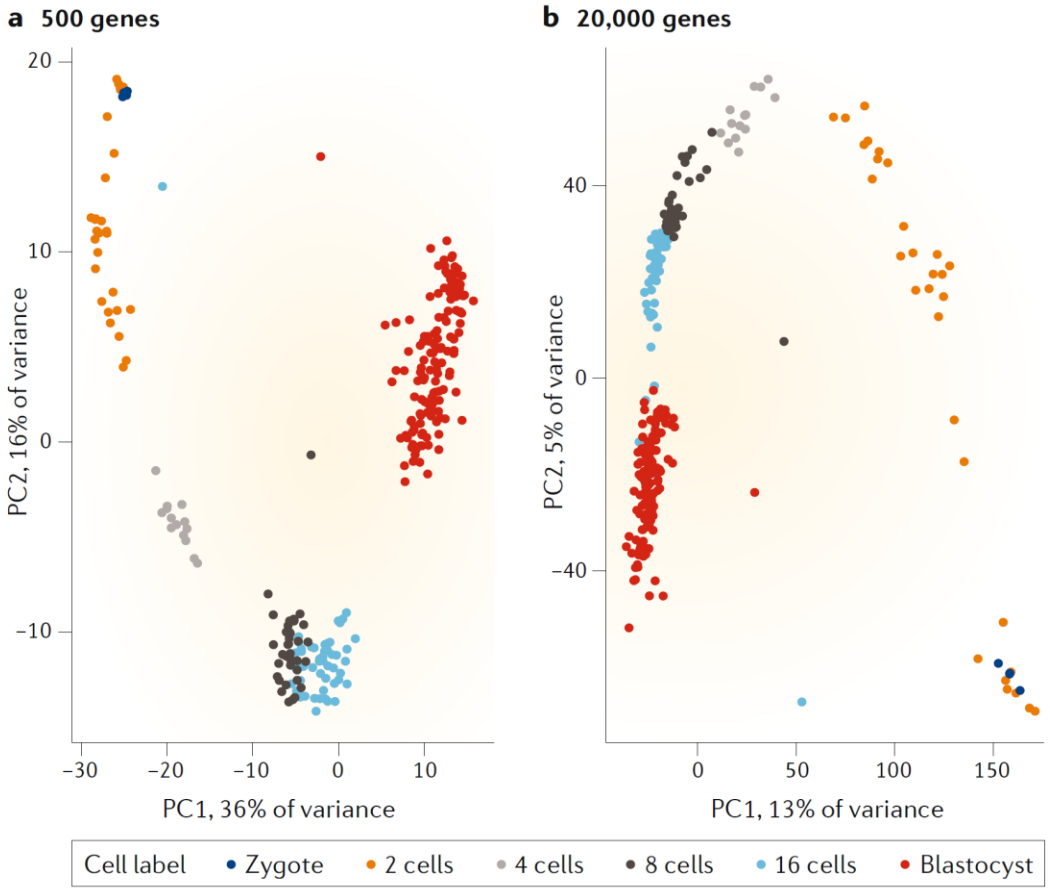
Kiselev et al. (Nature review Genetics 2019)

Feature selection

- Curse of dimensionality:

More features (genes) -> smaller distances between samples (cells)

- Remove genes which only exhibit technical noise
  - Increase the signal:noise ratio
  - Reduce the computational complexity



Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)

Feature selection

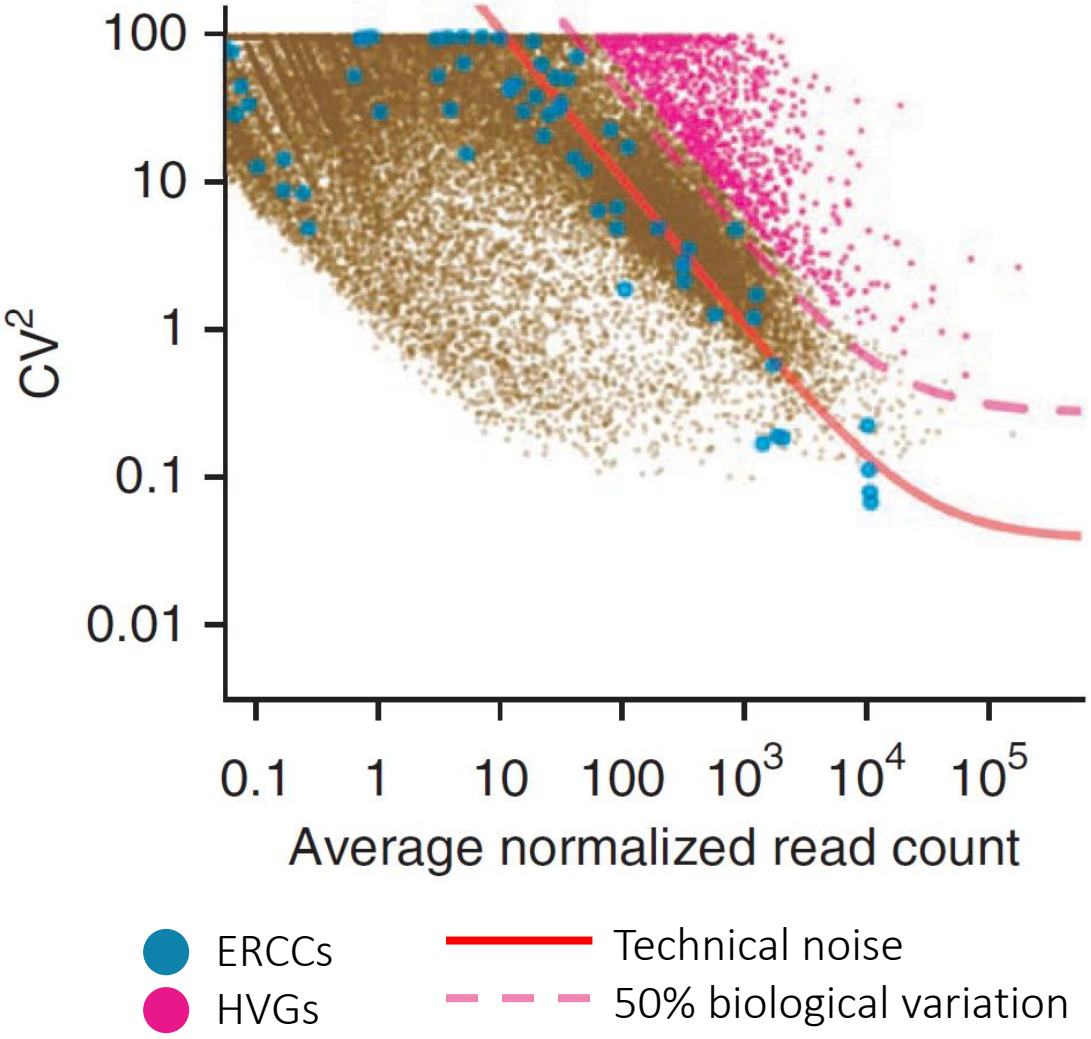
Highly Variable Genes (HVG)

- $CV = \frac{var}{mean} = \frac{\sigma}{\mu}$

- Fit a gamma generalized linear model

- No ERCCs?

  -> estimate technical noise based on all genes



Brennecke et al. (https://doi.org/10.1038/nmeth.2645)

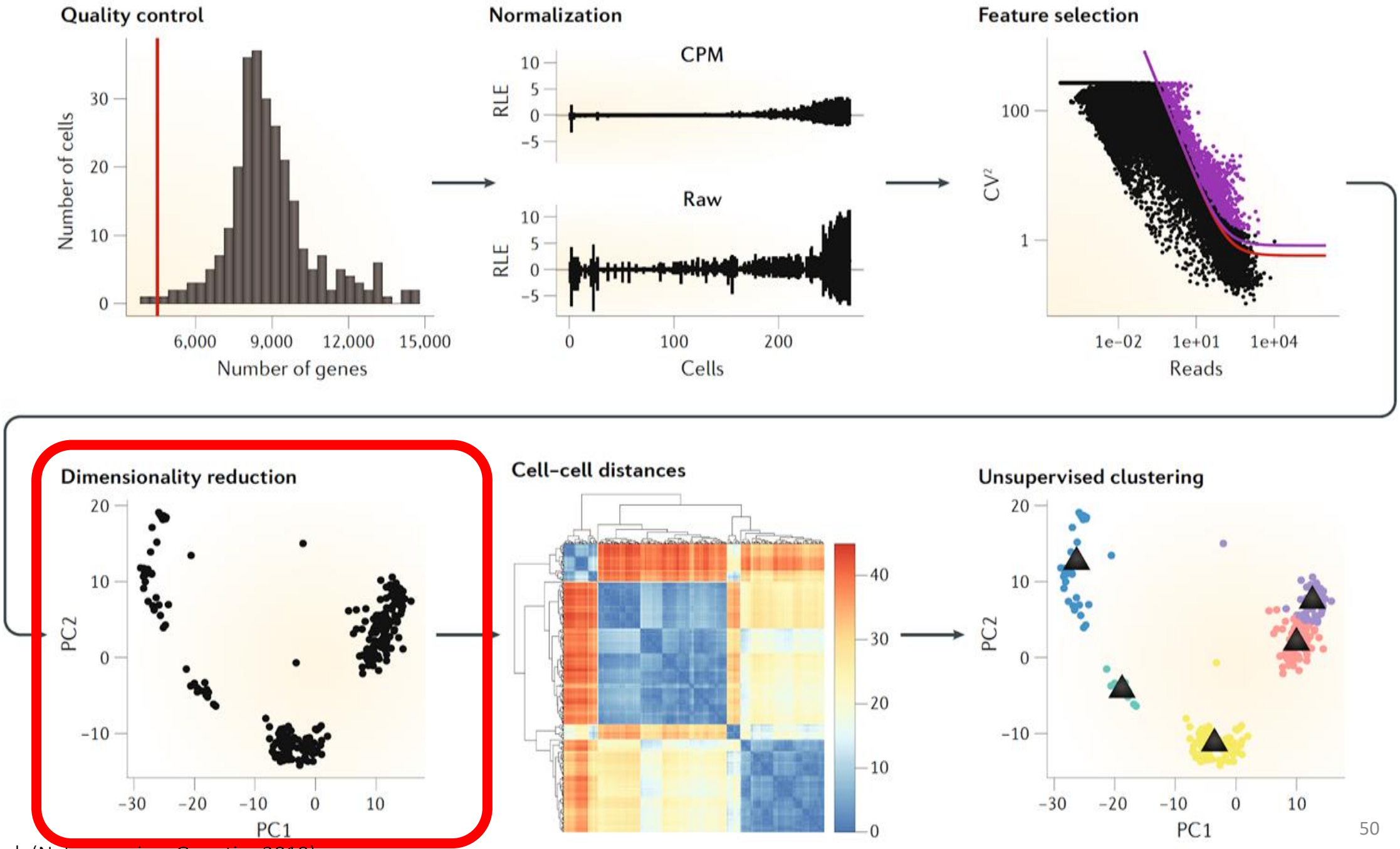
Feature Selection (pitfalls and recommendations)

- We recommend selecting between 1,000 and 5,000 highly variable genes depending on dataset complexity.

- Feature selection methods that use gene expression means and variances cannot be used when gene expression values have been normalized to zero mean and unit variance, or when residuals from model fitting are used as normalized expression values. Thus, one must consider what pre-processing to perform before selecting HVGs.

Luecken and Theis (MSB 2019)

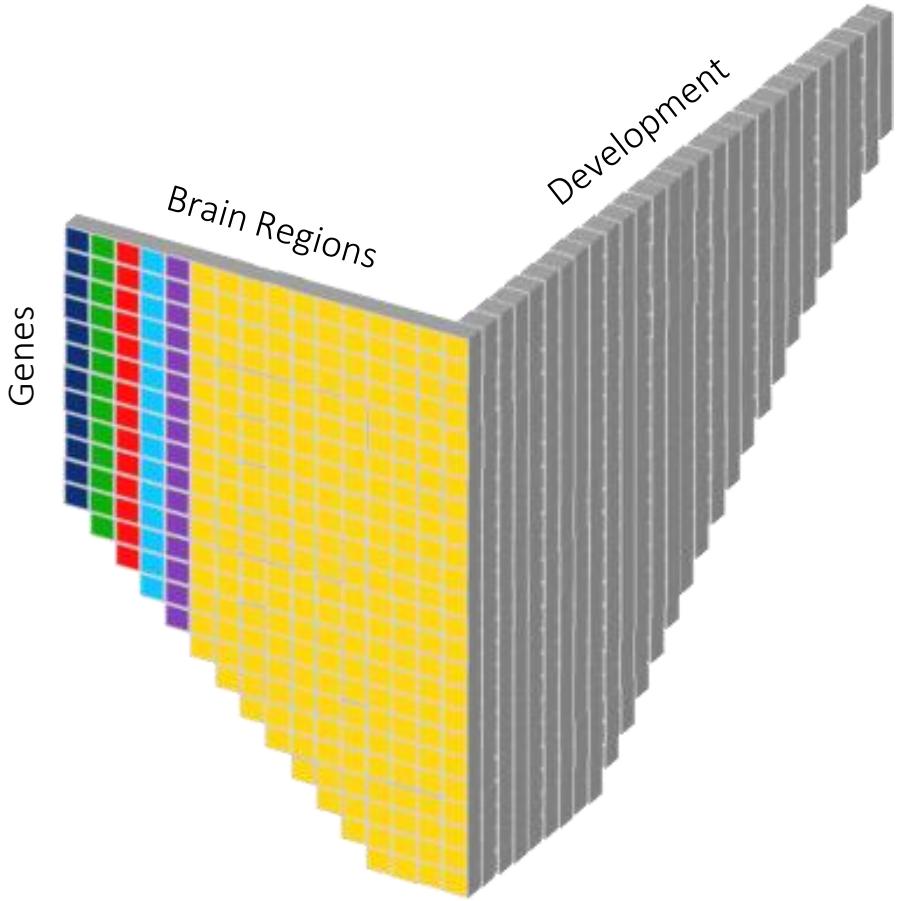
End part 1



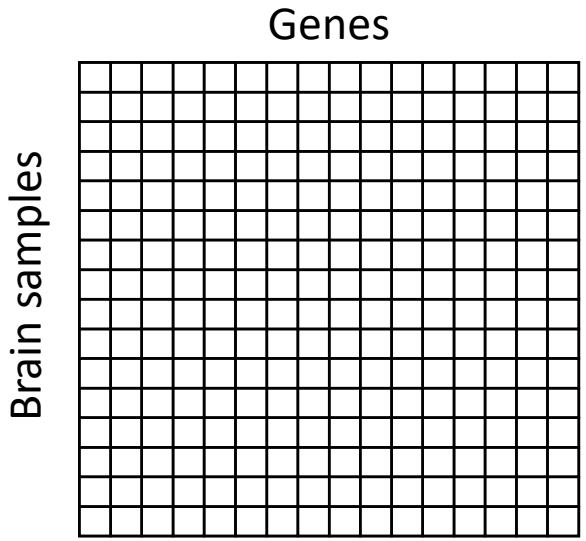
Kiselev et al. (Nature review Genetics 2019)

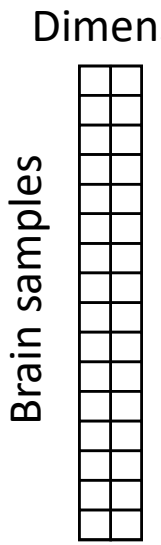
High dimensional data

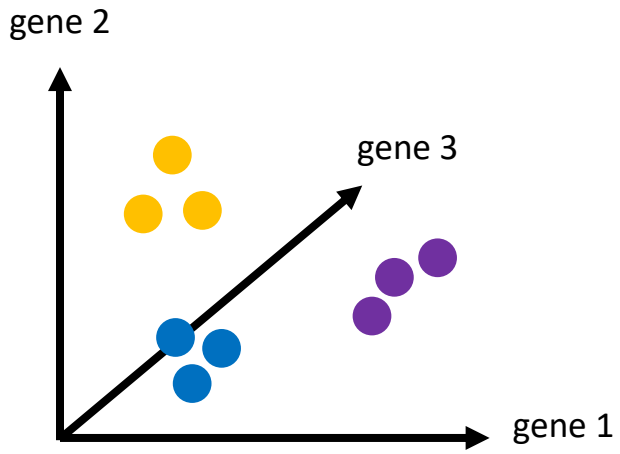
- We have huge amounts of complex data (many samples x many genes)
- We want to reduce complexity for analysis

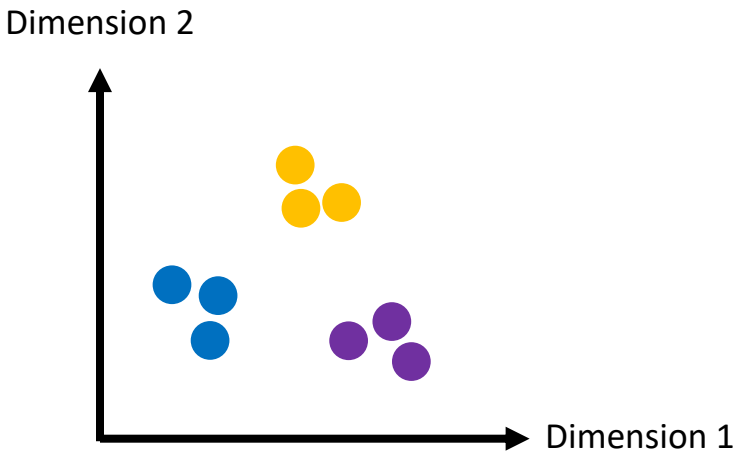


Dimensionality reduction





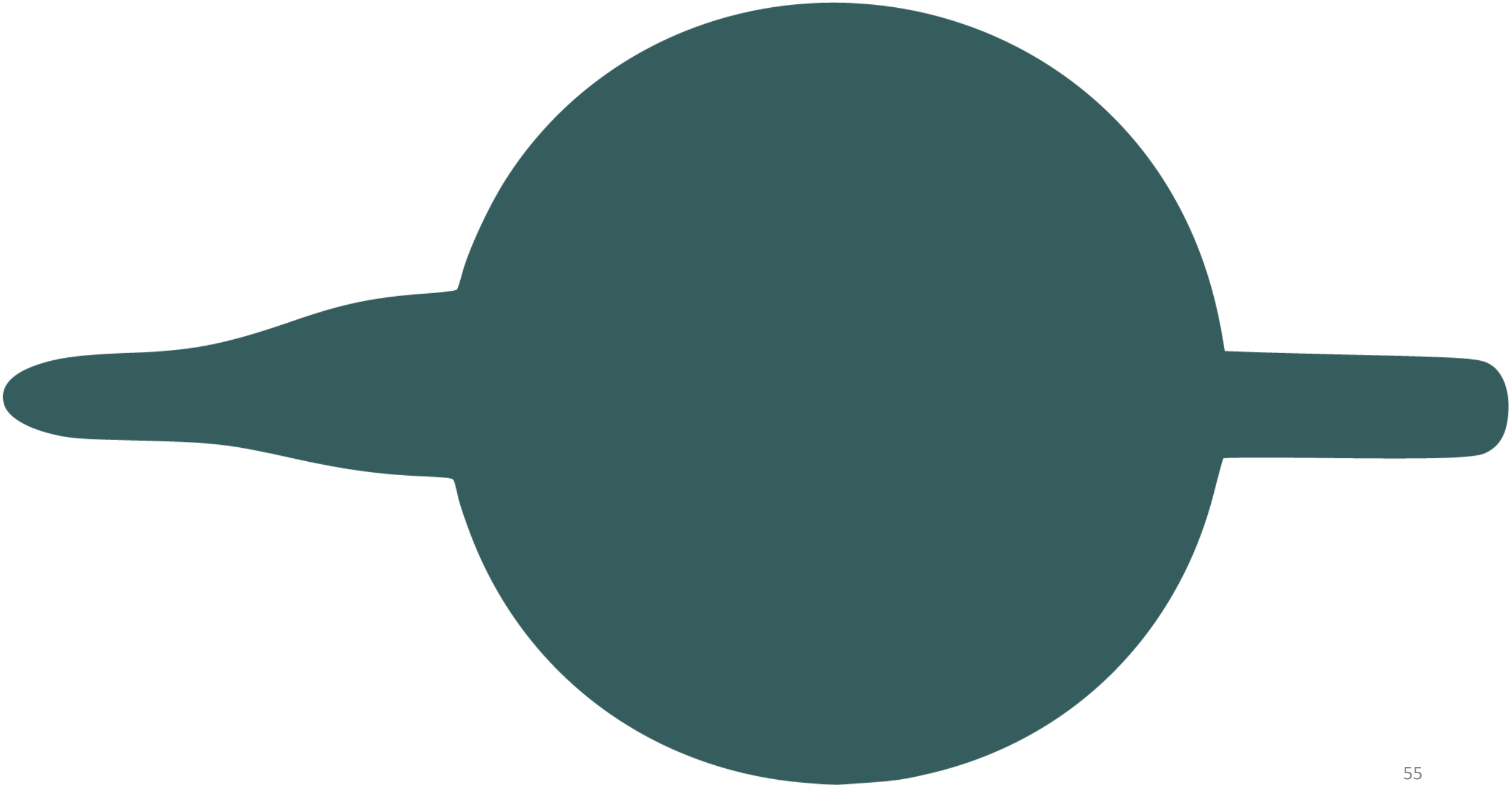


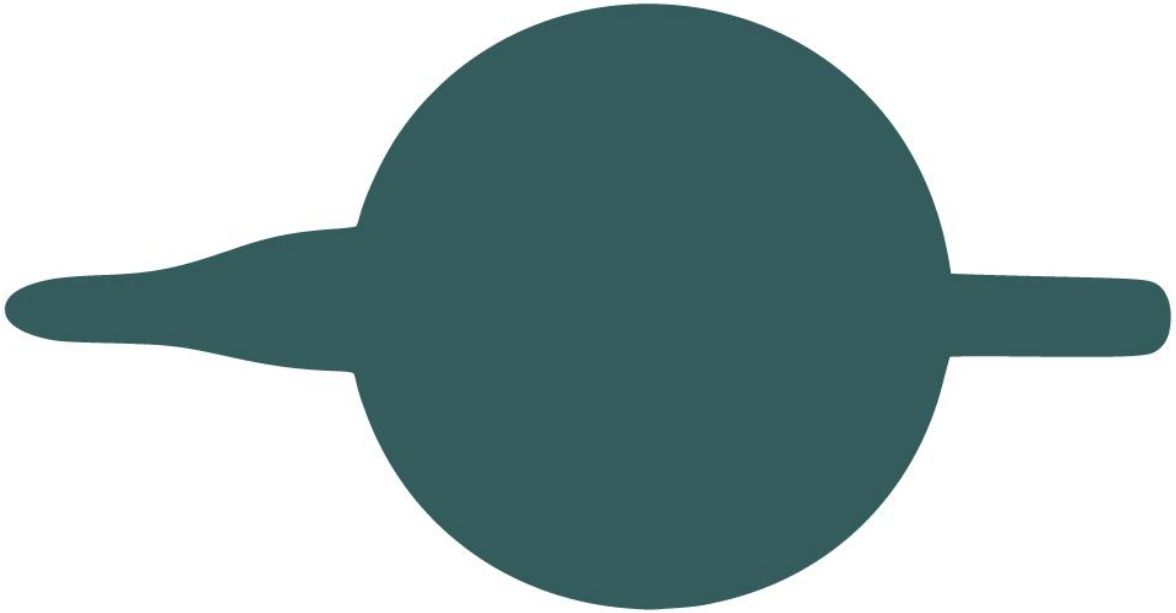


Why Dimensionality reduction?

- Simplify complexity, so it becomes easier to work with
  - “Remove” redundancies in the data
  - Identify the most relevant information (find and filter noise)
  - Reduce computational time for downstream procedures e.g. clustering

- Visualization





Dimensionality reduction (1)

Matrix factorization

Graph-based

Auto-encoders

PCA	linear		
ICA	linear		
MDS	non-linear		
Sparce NNMF	non-linear	2010	https://pdfs.semanticscholar.org/664d/40258f12ad28ed0b7d4 c272935ad72a150db.pdf
cPCA	non-linear	2018	https://doi.org/10.1038/s41467-018-04608-8
ZIFA	non-linear	2015	https://doi.org/10.1186/s13059-015-0805-z
ZINB-WaVE	non-linear	2018	https://doi.org/10.1038/s41467-017-02554-5
Diffusion maps	non-linear	2005	https://doi.org/10.1073/pnas.0500334102
Isomap	non-linear	2000	10.1126/science.290.5500.2319
t-SNE	non-linear	2008	https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- BH t-SNE	non-linear	2014	https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
- Flt-SNE	non-linear	2017	arXiv:1712.09005
LargeVis	non-linear	2018	arXiv:1602.00370
UMAP	non-linear	2018	arXiv:1802.03426
PHATE	non-linear	2017	https://www.biorxiv.org/content/biorxiv/early/2018/06/28/120378.full.pdf
scvis	non-linear	2018	https://doi.org/10.1038/s41467-018-04368-5
VASC	non-linear	2018	https://doi.org/10.1016/j.gpb.2018.08.003

Dimensionality reduction (1)

Matrix factorization

Graph-based

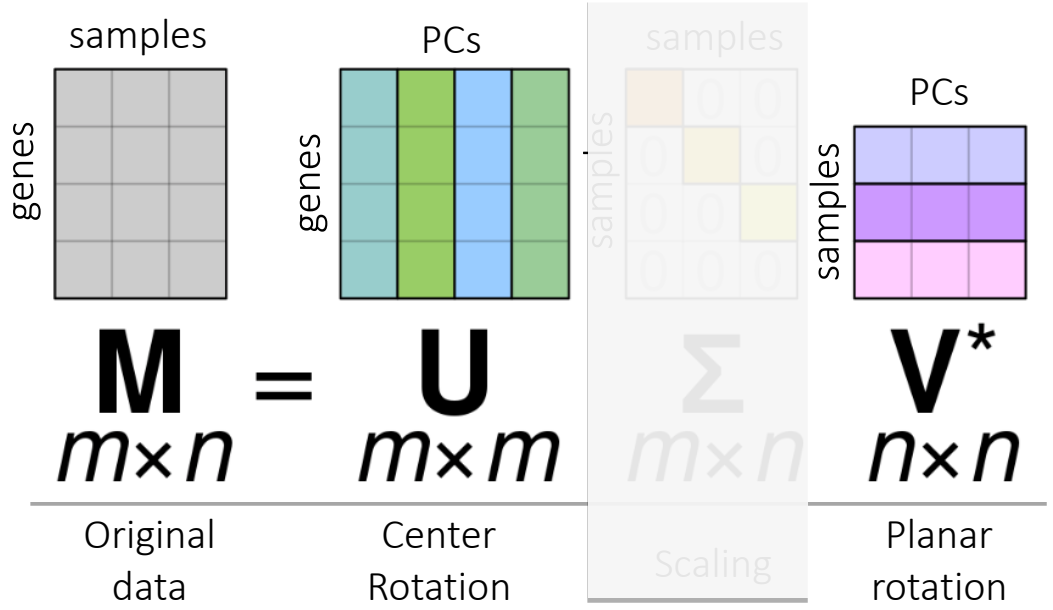
Auto-encoders

PCA	linear		
ICA	linear		
MDS	non-linear		
Sparce NNMF	non-linear	2010	https://pdfs.semanticscholar.org/664d/40258f12ad28ed0b7d4c272935ad72a150db.pdf
cPCA	non-linear	2018	https://doi.org/10.1038/s41467-018-04608-8
ZIFA	non-linear	2015	https://doi.org/10.1186/s13059-015-0805-z
ZINB-WaVE	non-linear	2018	https://doi.org/10.1038/s41467-017-02554-5
Diffusion maps	non-linear	2005	https://doi.org/10.1073/pnas.0500334102
Isomap	non-linear	2000	10.1126/science.290.5500.2319
t-SNE	non-linear	2008	https://lvdmaaten.github.io/publications/papers/JMLR_2008.pdf
- BH t-SNE	non-linear	2014	https://lvdmaaten.github.io/publications/papers/JMLR_2014.pdf
- Flt-SNE	non-linear	2017	arXiv:1712.09005
LargeVis	non-linear	2018	arXiv:1602.00370
UMAP	non-linear	2018	arXiv:1802.03426
PHATE	non-linear	2017	https://www.biorxiv.org/content/biorxiv/early/2018/06/28/120378.full.pdf
scvis	non-linear	2018	https://doi.org/10.1038/s41467-018-04368-5
VASC	non-linear	2018	https://doi.org/10.1016/j.gpb.2018.08.003

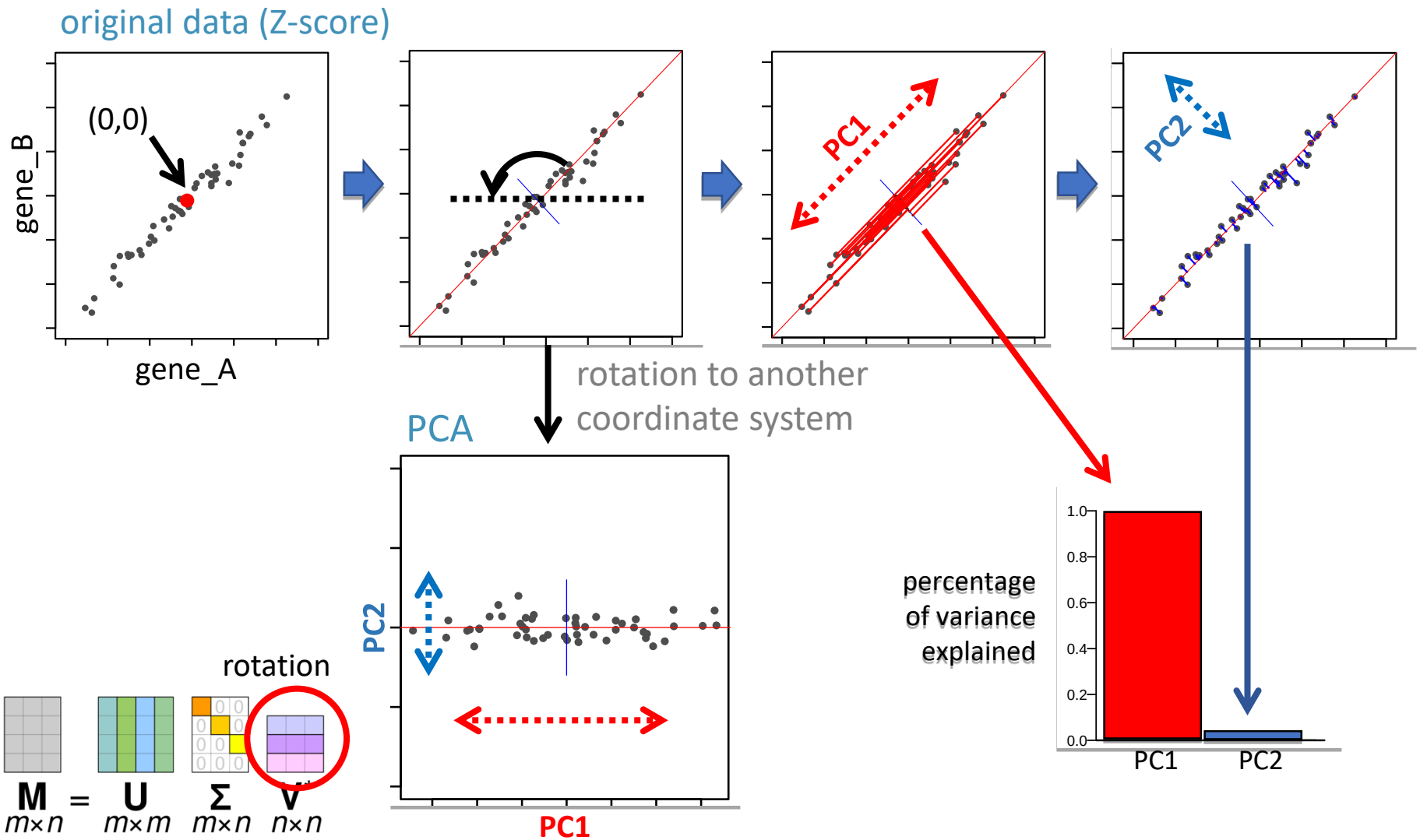
Principle Component Analysis (PCA)

- It is a LINEAR algebraic method of dimensionality reduction.

- It is a case inside Singular Value Decomposition (SVD) method (data compression)
  - Any matrix can be decomposed as a multiplication of other matrices (Matrix Factorization).

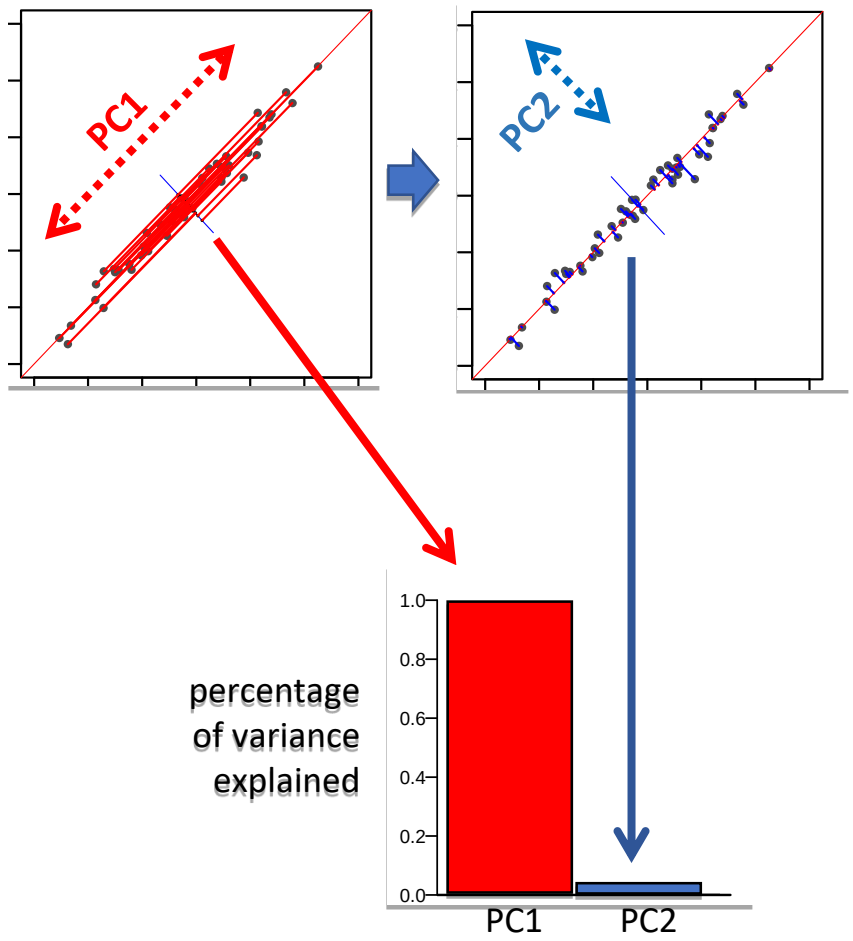


Principle Component Analysis (PCA)



Principle Component Analysis (PCA)

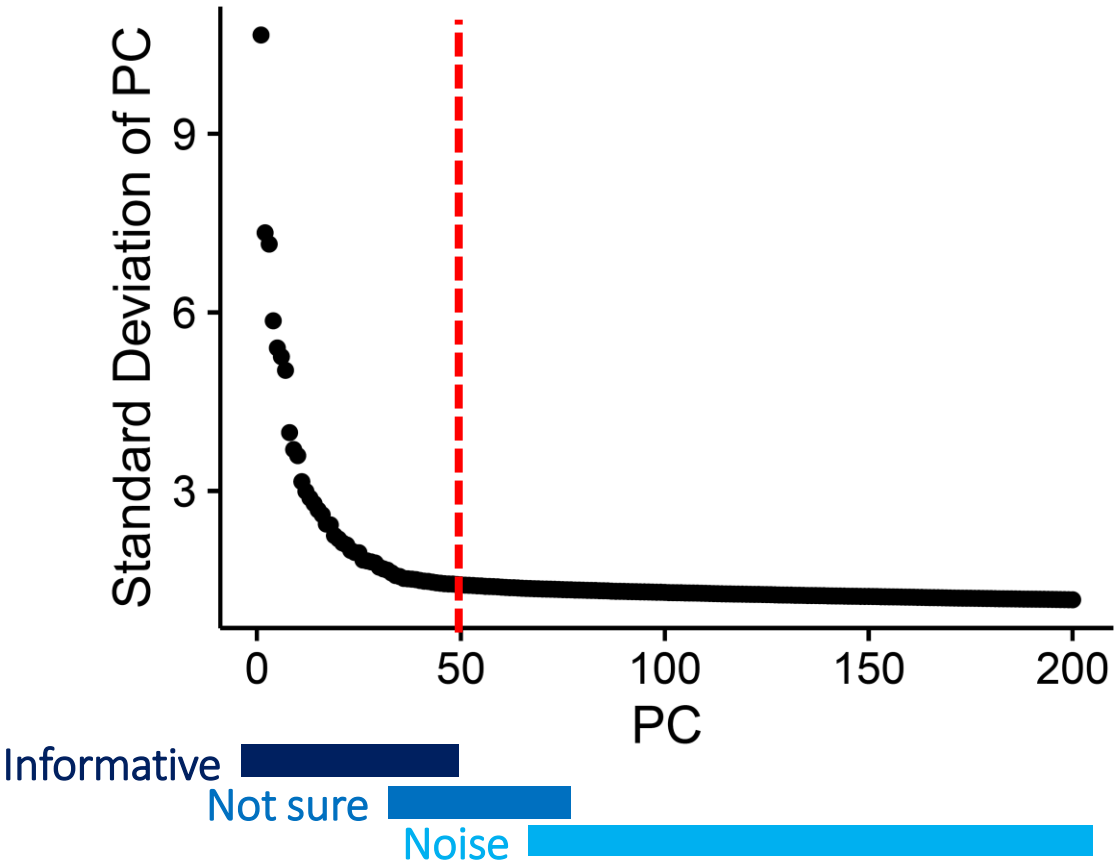
- PC1 explains >98% of the variance
- 1 PC thus represents 2 genes very well
  - “Removing” redundancy
- PC2 is nearly insignificant in this example
  - Could be disregarded



Principle Component Analysis (PCA)

- In reality…

Scree/elbow plot

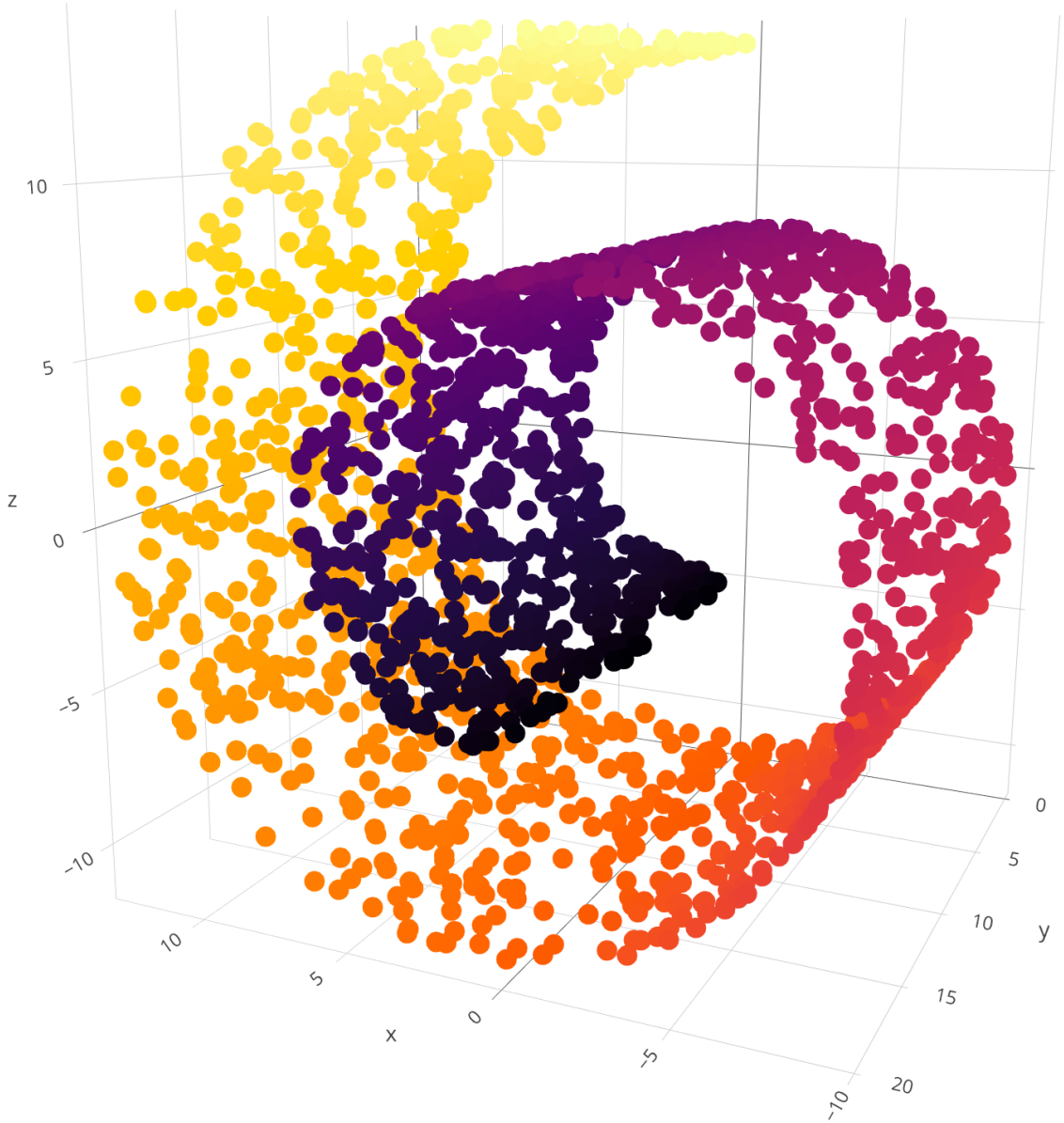


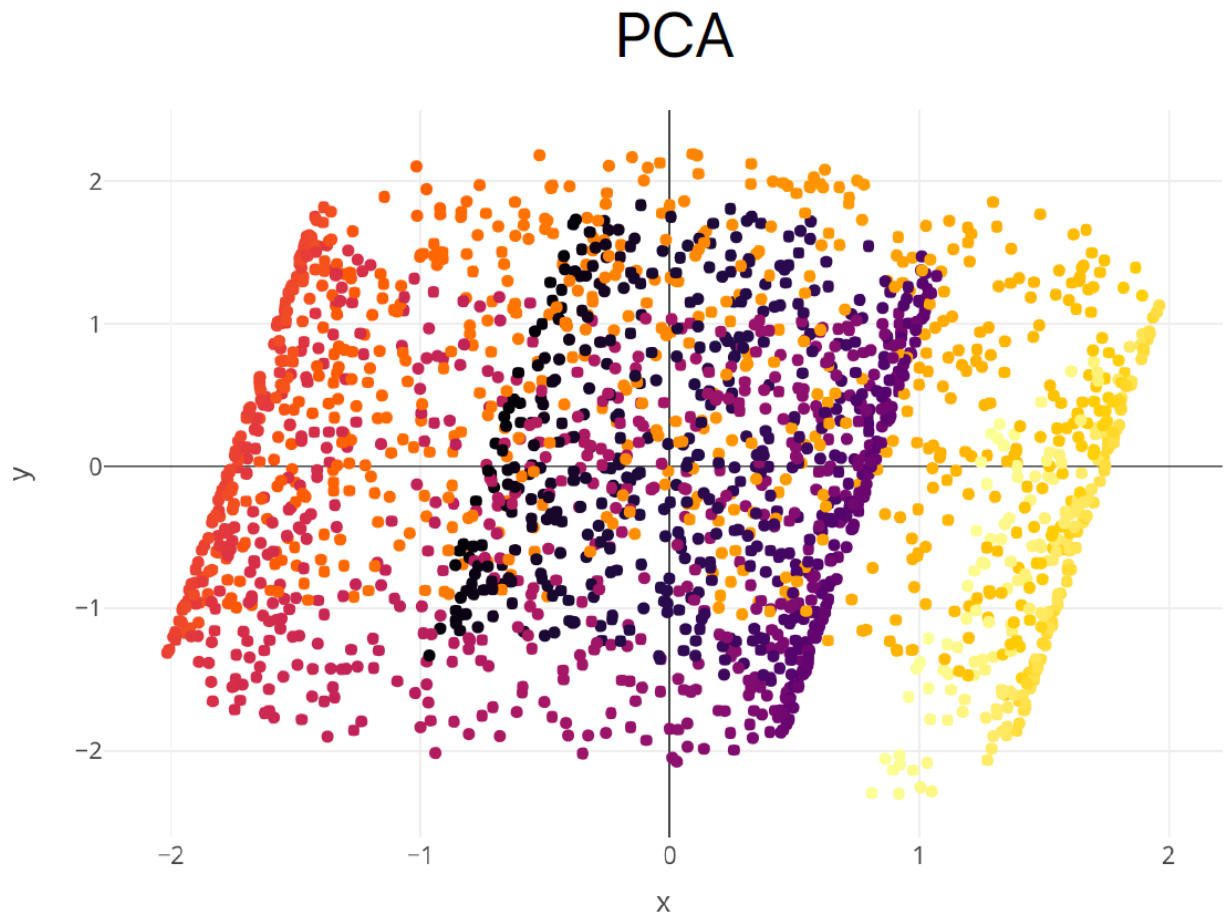
Principle Component Analysis (PCA)

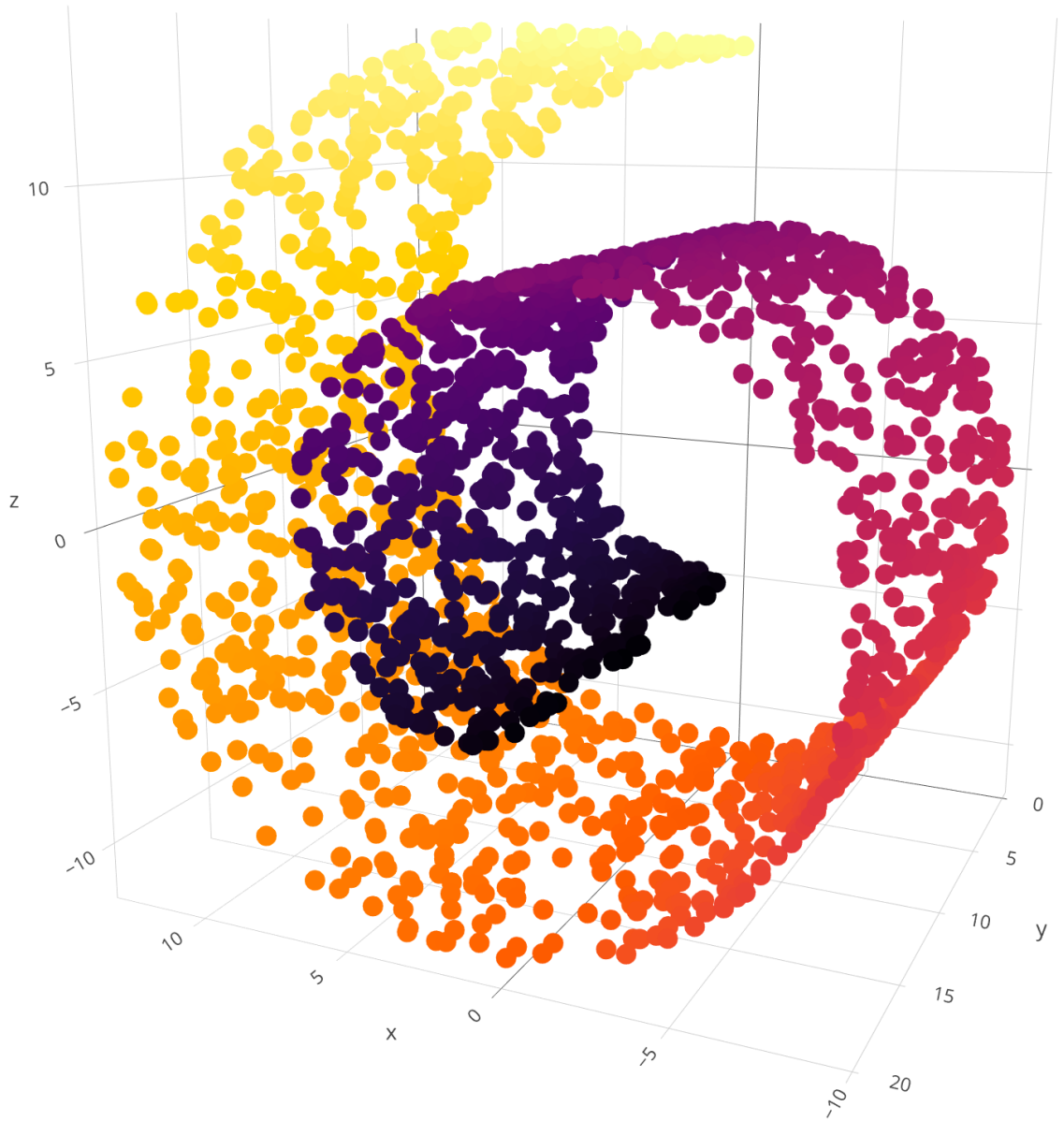
- LINEAR method of dimensionality reduction
- The TOP principal components contain higher variance from the data
- Can be used as FILTERING, by selecting only the top significant PCs
- *It is an interpretable/parametric dimensionality reduction*

Problems

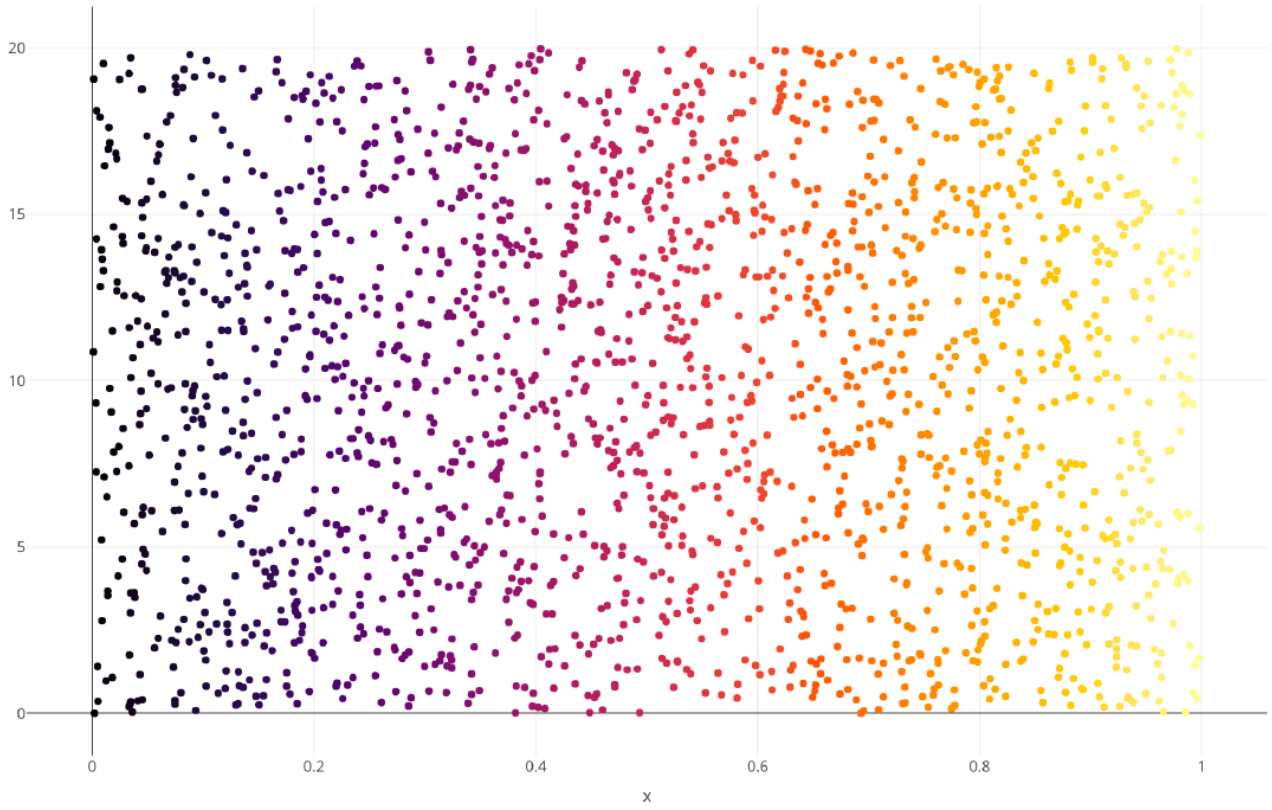
- It performs poorly to separate cells in 0-inflated data types
- Cell sizes and sequencing depth are usually captured in the top PCs


10
5
z
0
−5
−10
10
5
x
0
−5
−10
0
5
10
15
20
y


PCA
2
1
y
0
−1
−2
−2
−1
0
1
2
x



Better

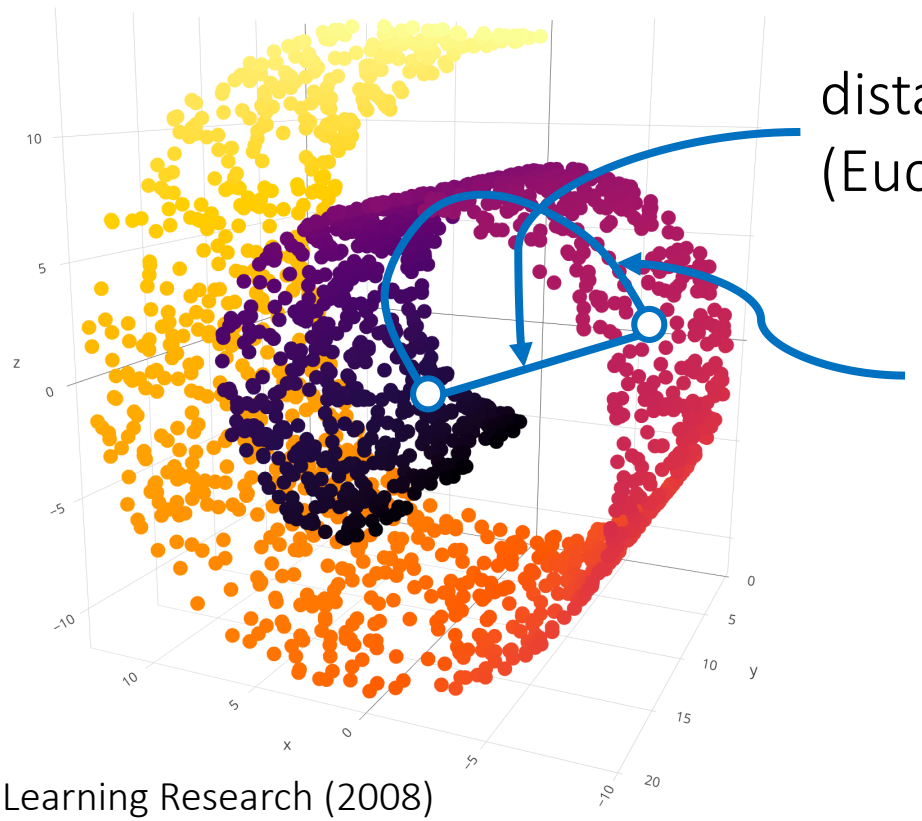


t-SNE

t-distributed stochastic neighborhood embedding

- It is a graph-based NON-LINEAR dimensionality reduction

Manifold



distance in PCA space
(Euclidean distance)

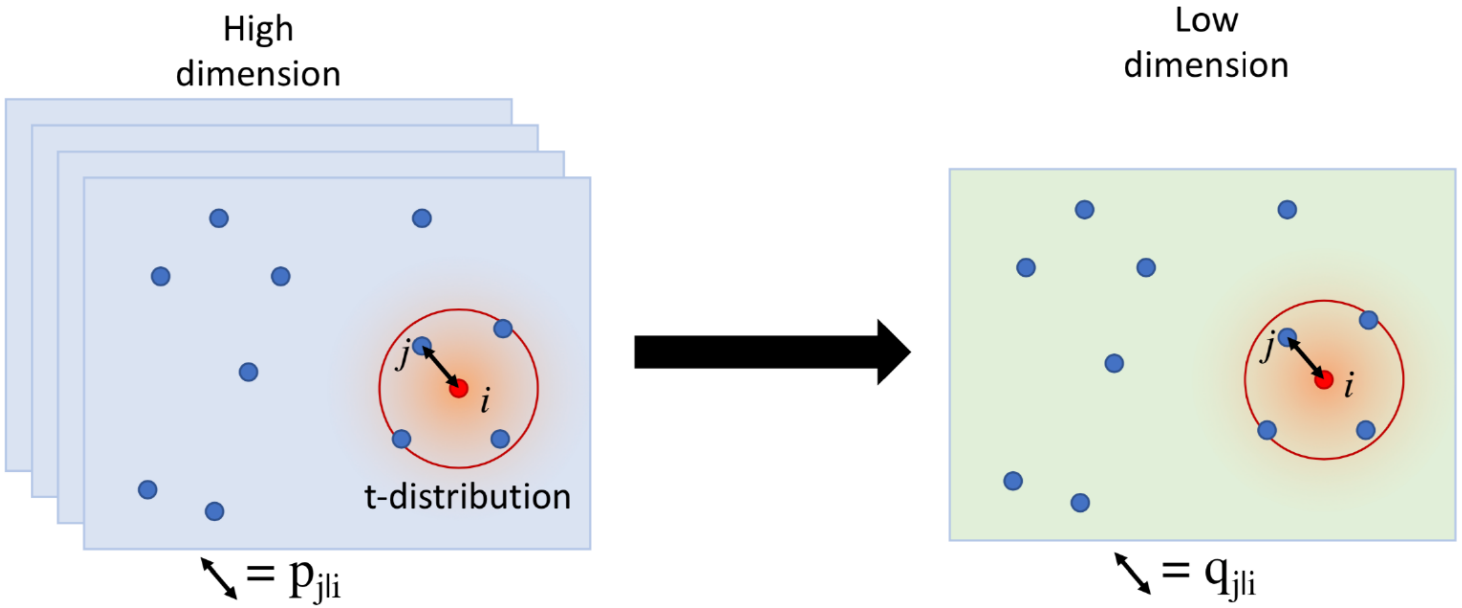
distance in tSNE space

t-SNE calculates the distances based on the distance to the neighbor cell

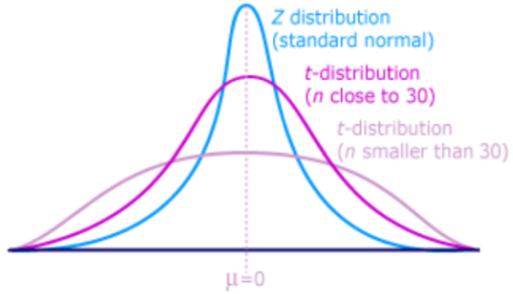
van der Maaten and Hinton, Journal of Machine Learning Research (2008)

t-SNE

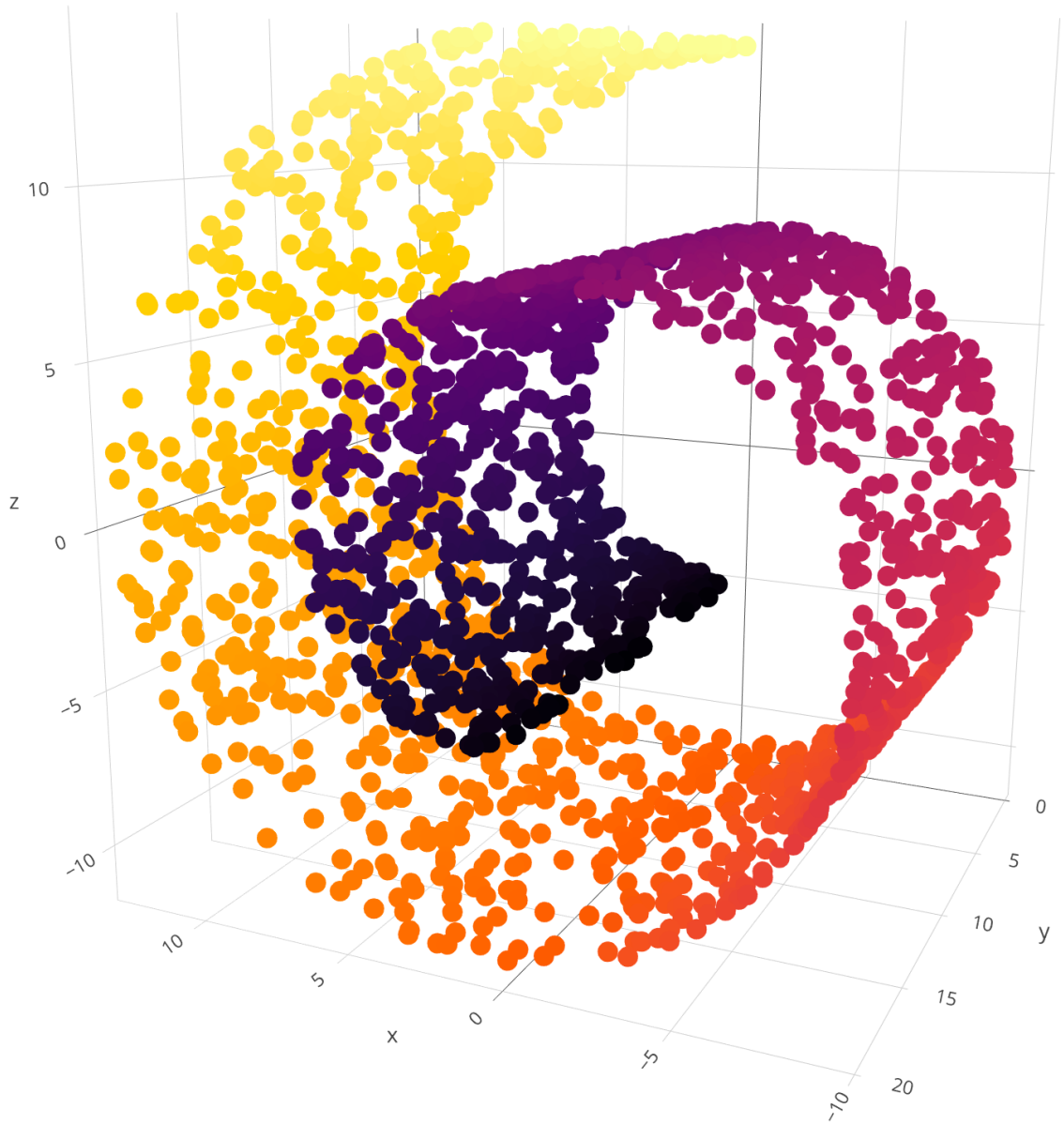
t-distributed stochastic neighborhood embedding

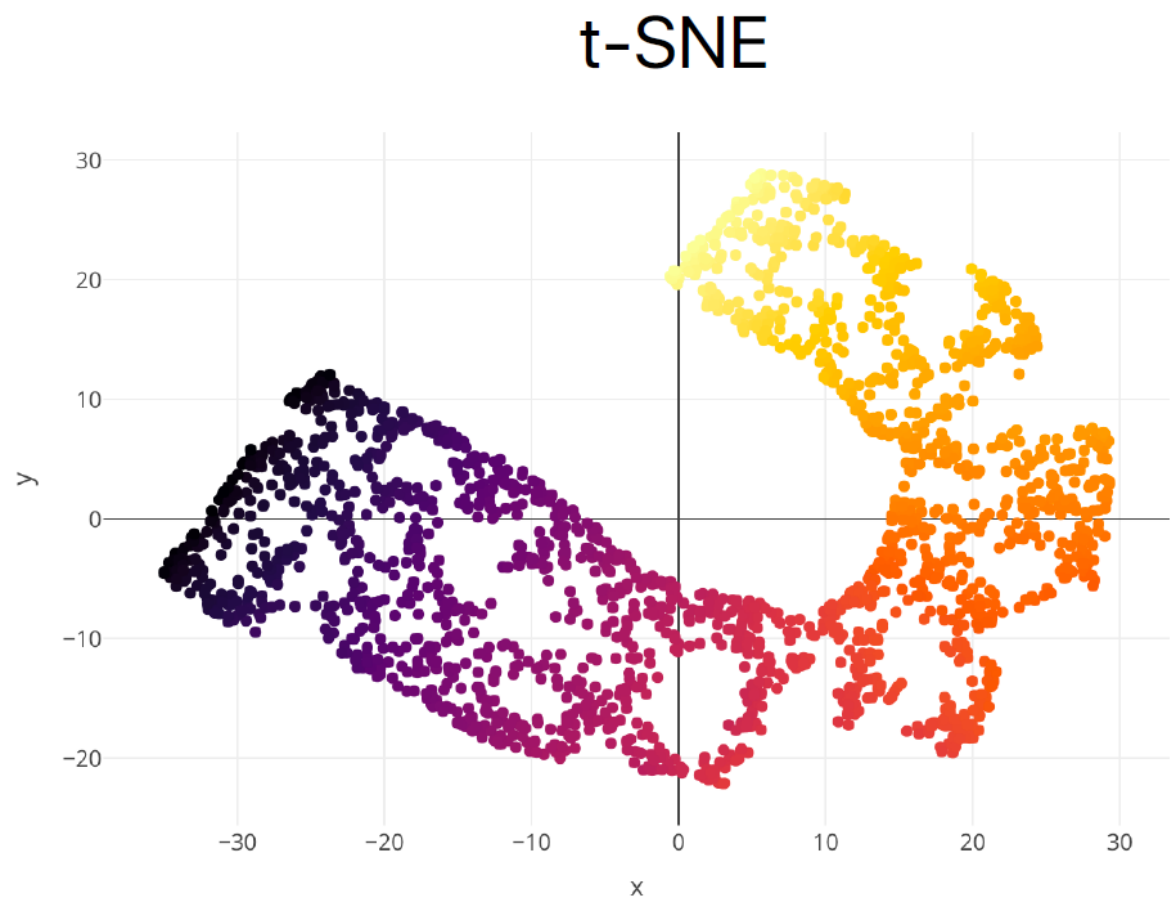


p_{jli} and q_{jli} measure the conditional probability that a point i would pick point j as it's nearest neighbor, in high (p) and low (q) dimensional space respectively.



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


10
5
z
0
−5
−10
10
5
x
0
−5
−10
0
5
10
15
20
y


t-SNE
30
20
10
y
0
−10
−20
−30
−20
−10
0
10
20
30
x

t-SNE

t-distributed stochastic neighborhood embedding

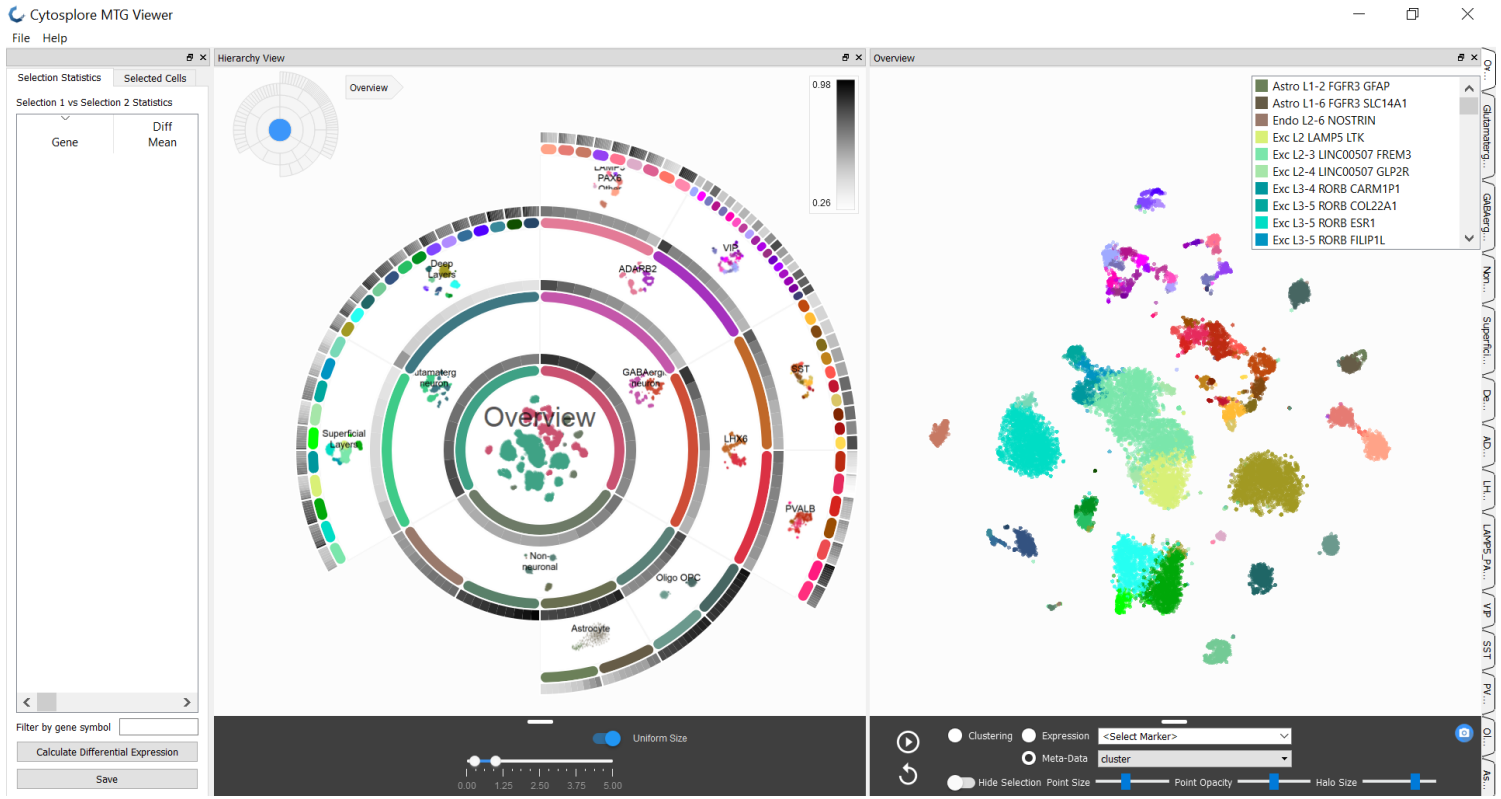
- NON-LINEAR method of dimensionality reduction
- It is the current GOLD-STANDARD method in single cell data (including scRNA-seq)
- Can be run from the top PCs (e.g.: PC1 to PC10)

Problems

- It does not learn an explicit function to map new points
- It's cost function is not convex – This means that the optimal t-SNE cannot be computed
- Many hyper-parameters need to be defined empirically (dataset-specific)

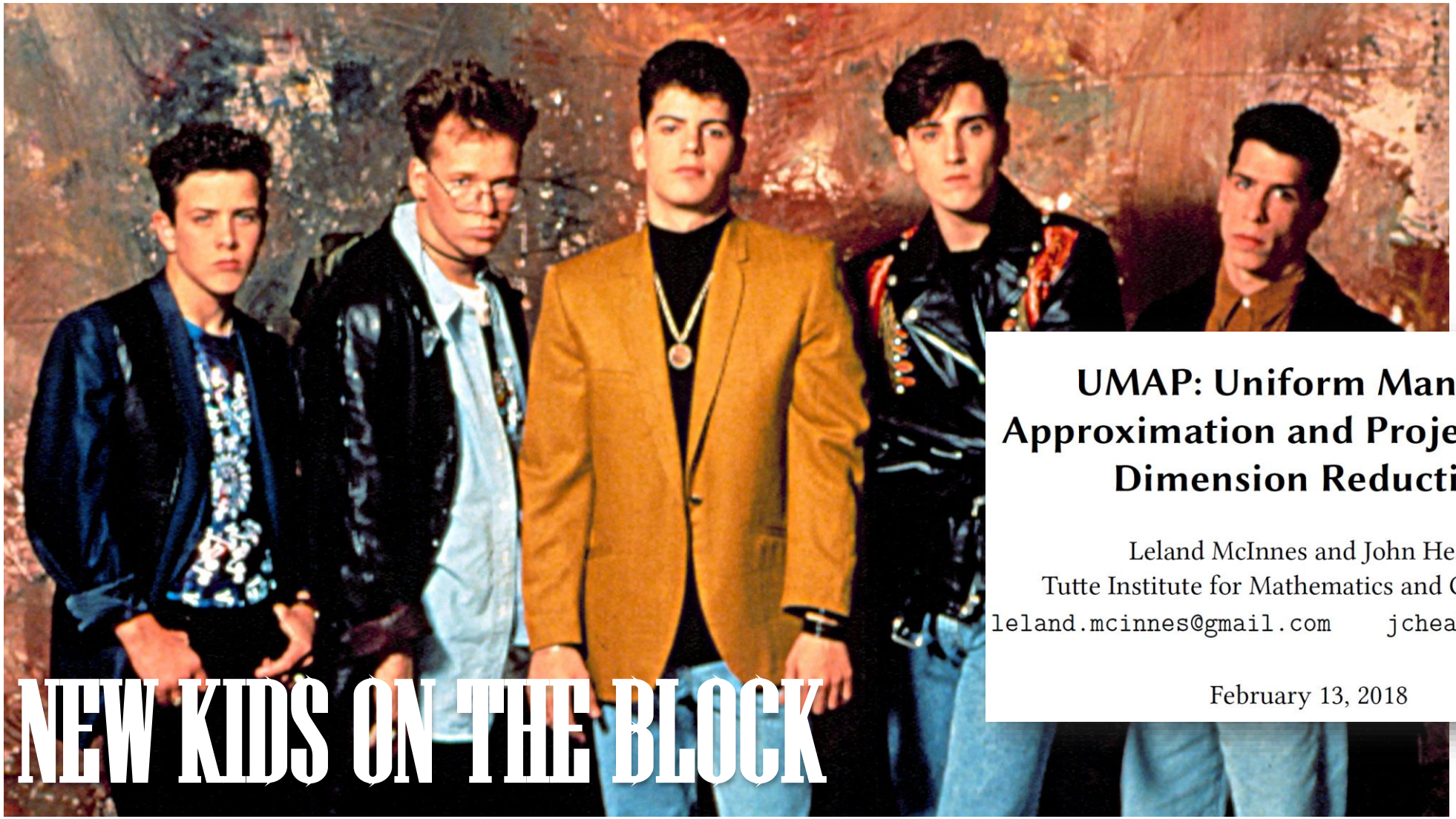
Dimensionality Reduction (5)

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



Hodge et al. BioRxiv (2018)

Dimensionality Reduction (5)

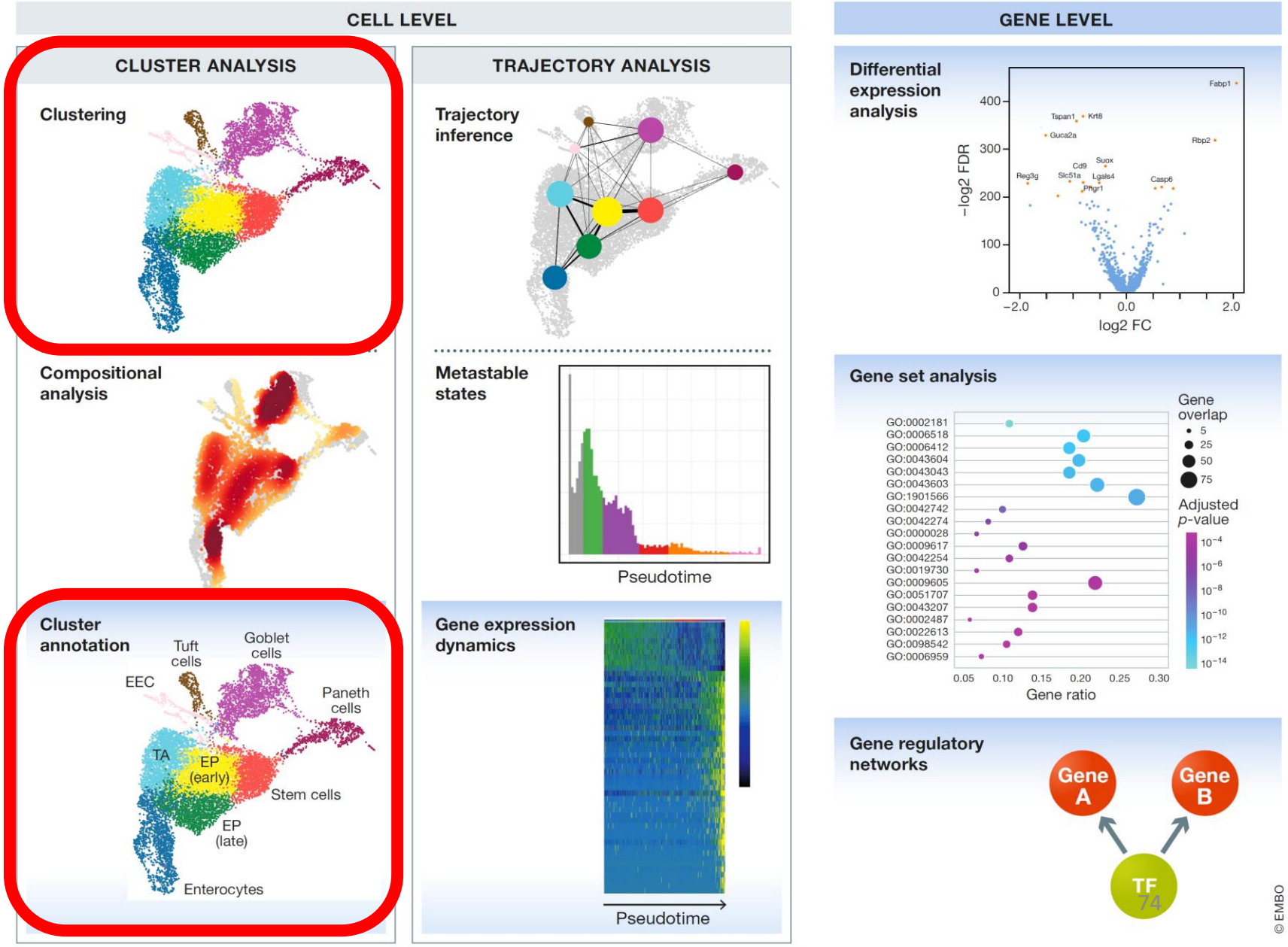


UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction

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

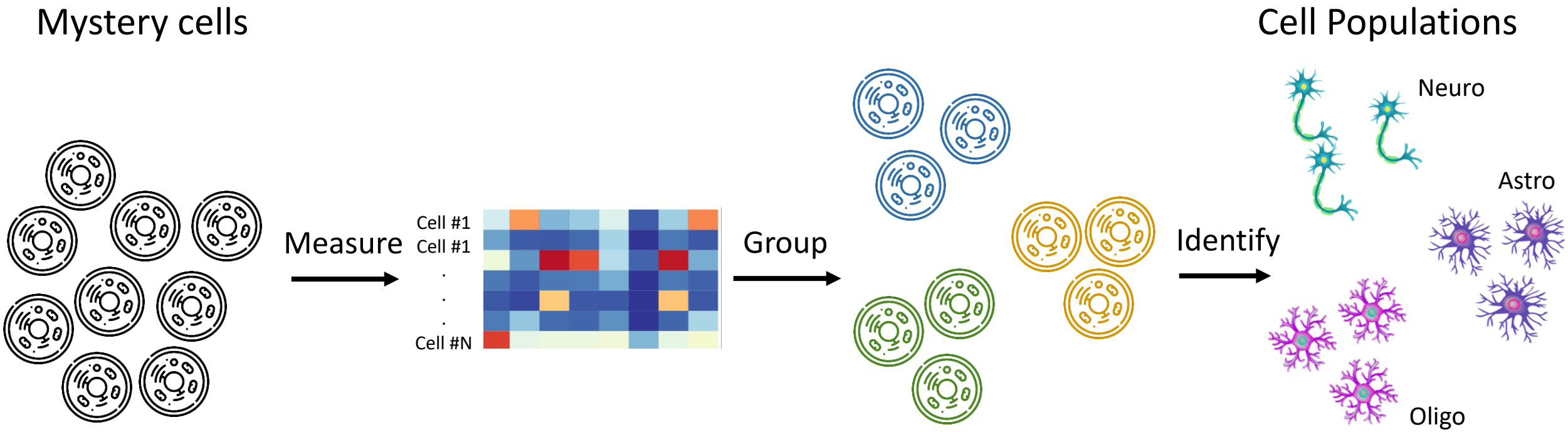
February 13, 2018

scRNA-seq Downstream Analysis

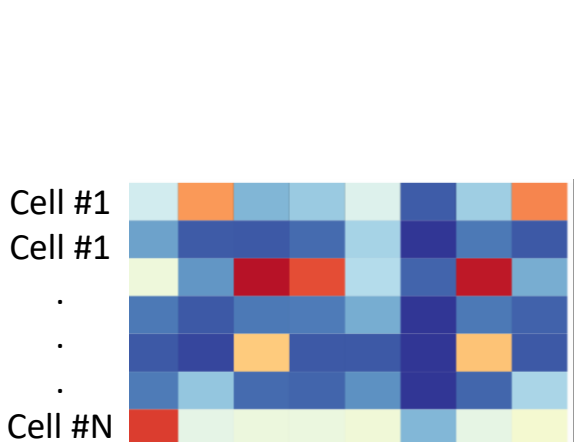


Luecken and Theis (MSB 2019)

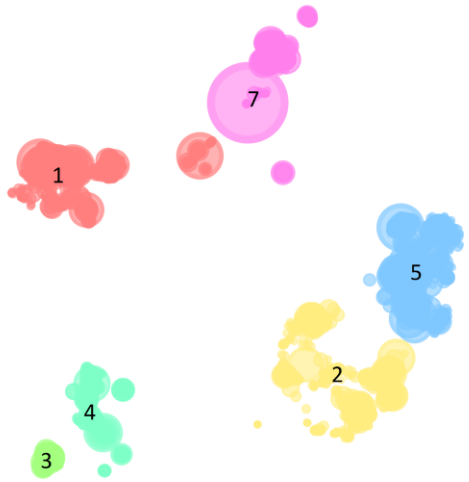
How can we identify cell populations?

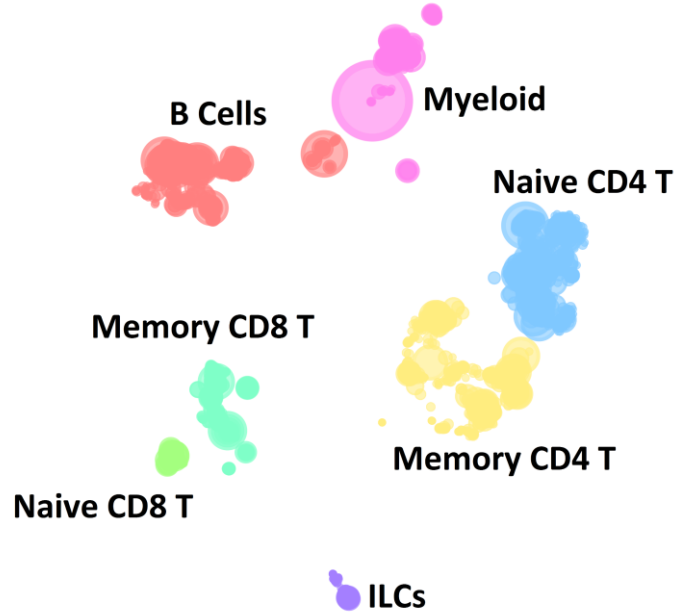


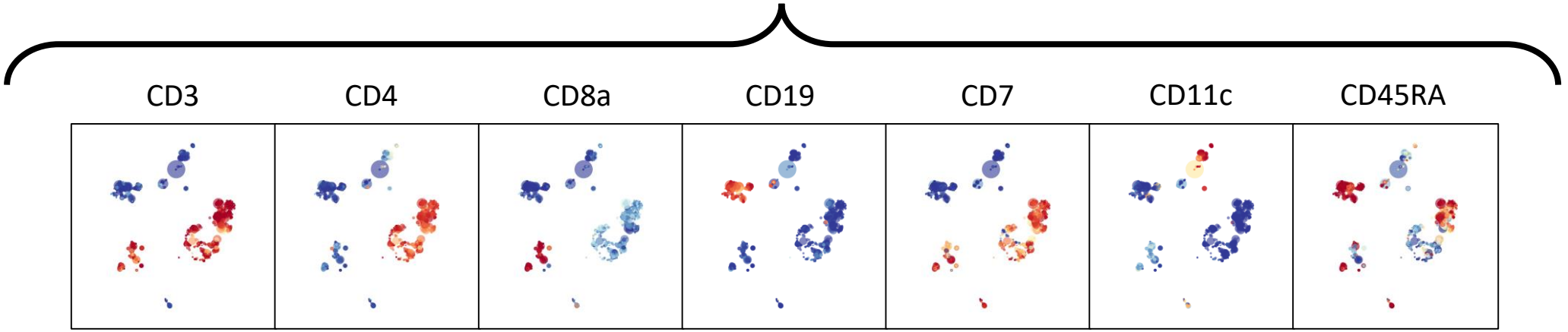
Unsupervised approach



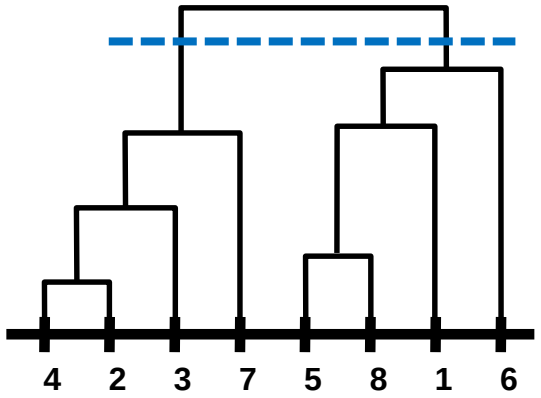
PCA
tSNE
UMAP
...



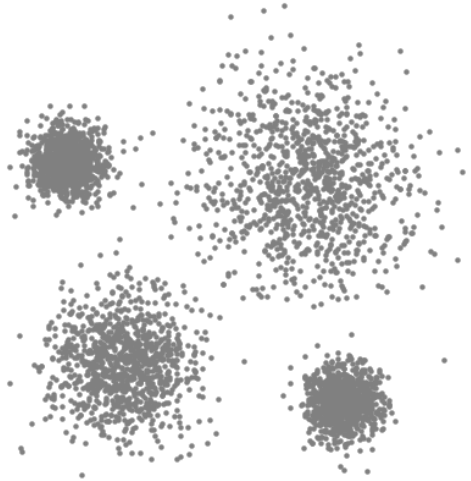




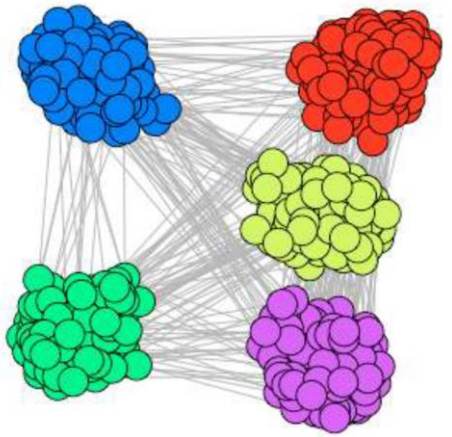
Many clustering approaches



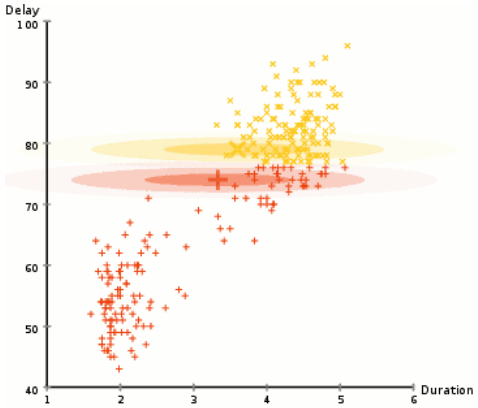
Hierarchical Clustering



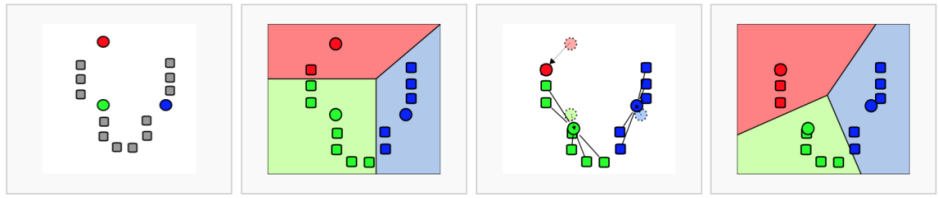
Mean shift clustering



Graph-based clustering



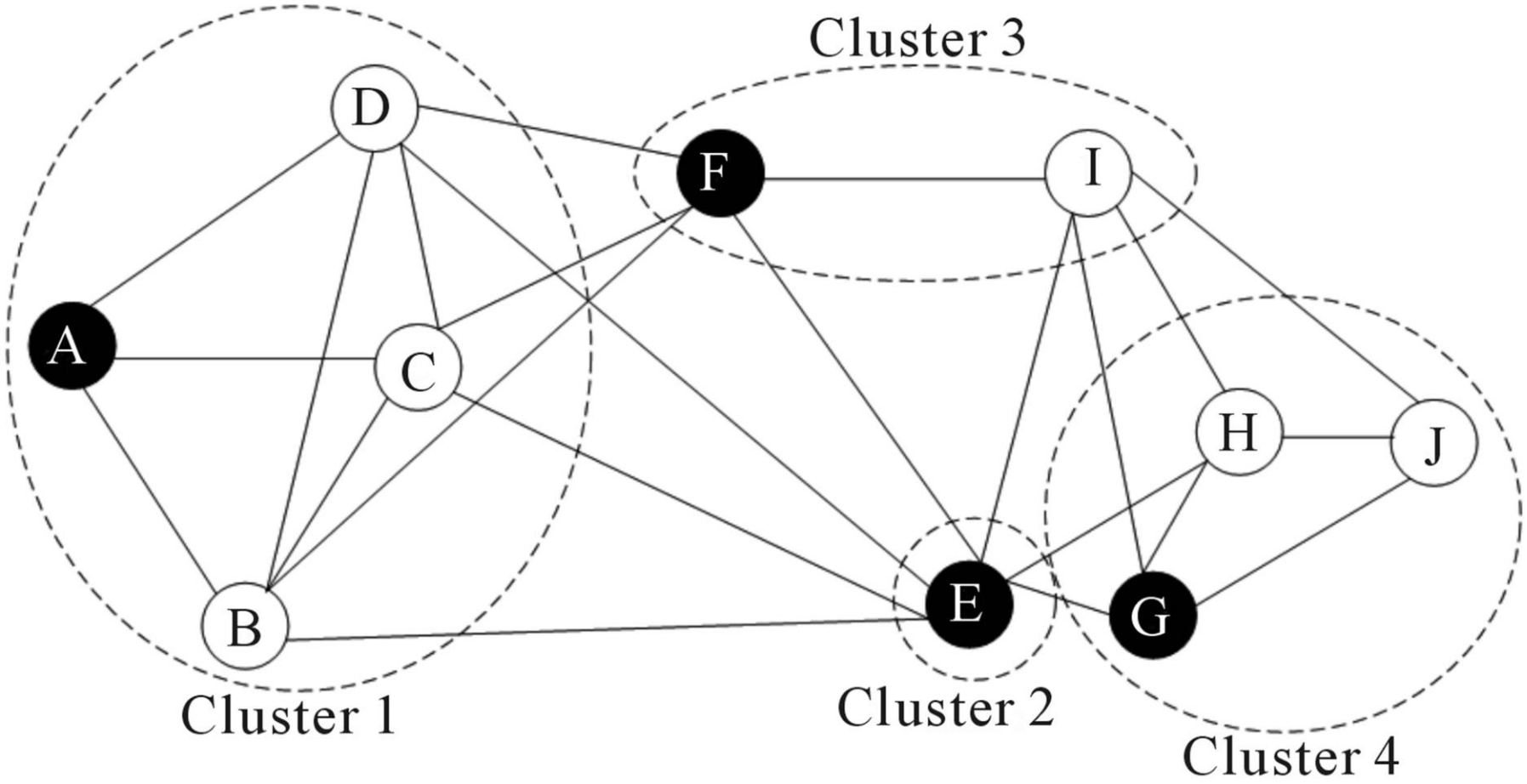
Gaussian mixture modeling



k-means clustering

Graph-based clustering

Nodes -> cells
Edges -> similarity



Graph Types

- **k-Nearest Neighbor (kNN) graph**

A graph in which two vertices *p* and *q* are connected by an edge, if the distance between *p* and *q* is among the *k*-th smallest distances from *p* to other objects from *P.*

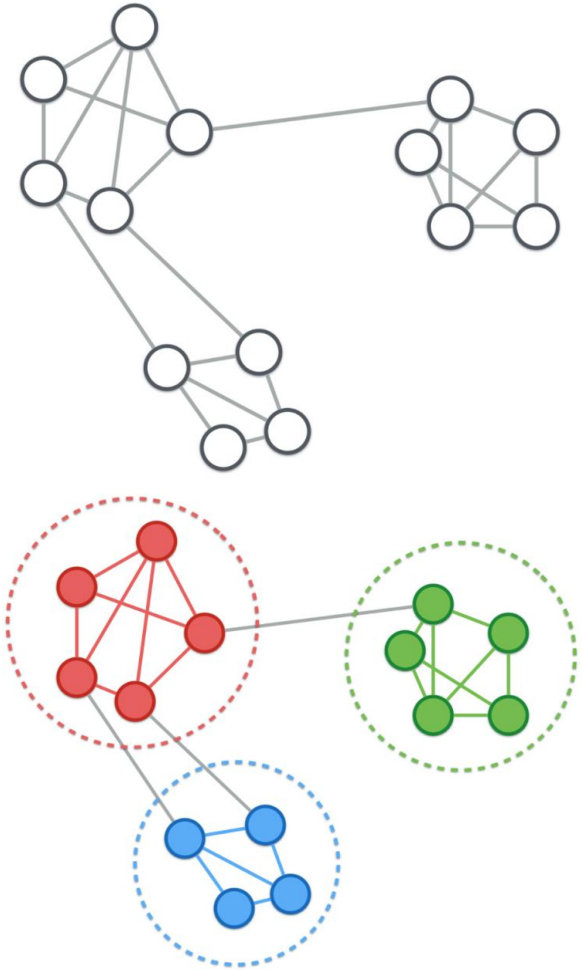
- **Shared Nearest Neighbor (SNN) graph**

A graph in which weights define proximity, or similarity between two nodes in terms of the number of neighbors (i.e., directly connected nodes) they have in common.

Graph clustering (Community detection)

- **Community detection:** find a group (community) of nodes with more edges inside the group than edges linking nodes of the group with the rest of the graph.

- Algorithms for community detection:
  - Spectral clustering
  - Louvain
  - Markov clustering
  - …



scRNA-seq clustering methods

Name	Year	Method type	Strengths	Limitations
scanpy[4]	2018	PCA + graph-based	Very scalable	May not be accurate for small data sets
Seurat (latest)[3]	2016			
PhenoGraph[32]	2015			
SC3 (REF.[22])	2017	PCA + *k*-means	High accuracy through consensus, provides estimation of *k*	High complexity, not scalable
SIMLR[24]	2017	Data-driven dimensionality reduction + *k*-means	Concurrent training of the distance metric improves sensitivity in noisy data sets	Adjusting the distance metric to make cells fit the clusters may artificially inflate quality measures
CIDR[25]	2017	PCA + hierarchical	Implicitly imputes dropouts when calculating distances	
GiniClust[75]	2016	DBSCAN	Sensitive to rare cell types	Not effective for the detection of large clusters
pcaReduce[27]	2016	PCA + *k*-means + hierarchical	Provides hierarchy of solutions	Very stochastic, does not provide a stable result
Tasic et al.[28]	2016	PCA + hierarchical	Cross validation used to perform fuzzy clustering	High complexity, no software package available
TSCAN[41]	2016	PCA + Gaussian mixture model	Combines clustering and pseudotime analysis	Assumes clusters follow multivariate normal distribution
mpath[45]	2016	Hierarchical	Combines clustering and pseudotime analysis	Uses empirically defined thresholds and a priori knowledge
BackSPIN[26]	2015	Biclustering (hierarchical)	Multiple rounds of feature selection improve clustering resolution	Tends to over-partition the data
RaceID[23], RaceID2 (REF.[115]), RaceID3	2015	*k*-Means	Detects rare cell types, provides estimation of *k*	Performs poorly when there are no rare cell types
SINCERA[5]	2015	Hierarchical	Method is intuitively easy to understand	Simple hierarchical clustering is used, may not be appropriate for very noisy data
SNN-Cliq[80]	2015	Graph-based	Provides estimation of *k*	High complexity, not scalable

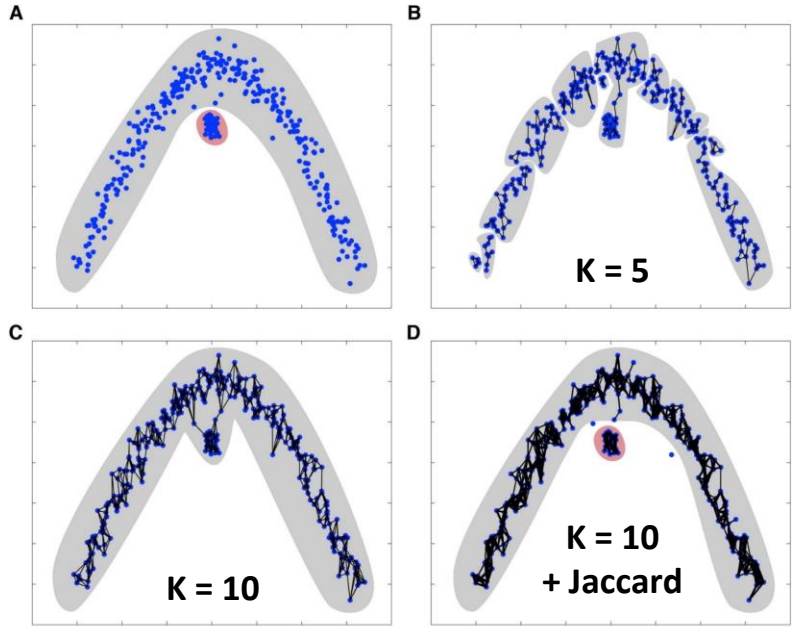
Kiselev et al. (https://doi.org/10.1038/s41576-018-0088-9)

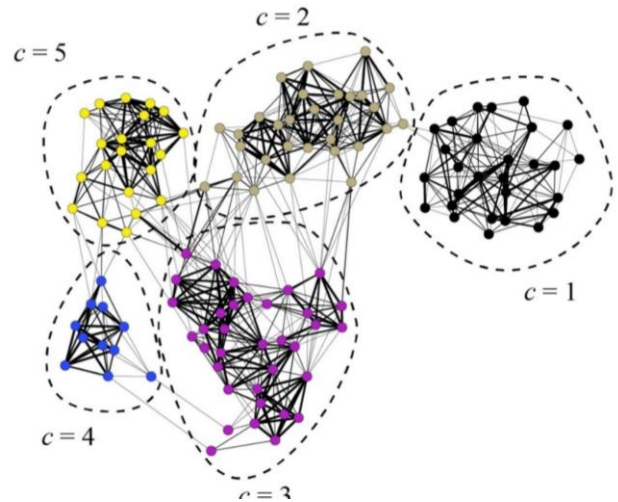
Seurat

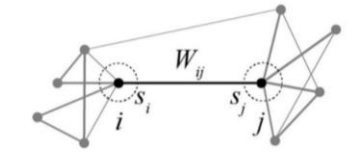
1) Construct KNN (k-nearest neighbor) graph based on the Euclidean distance in PCA space.

2) Refine the edge weights between any two cells based on the shared overlap in their local neighborhoods (Jaccard distance).

3) Cluster cells by optimizing for modularity (Louvain algorithm)



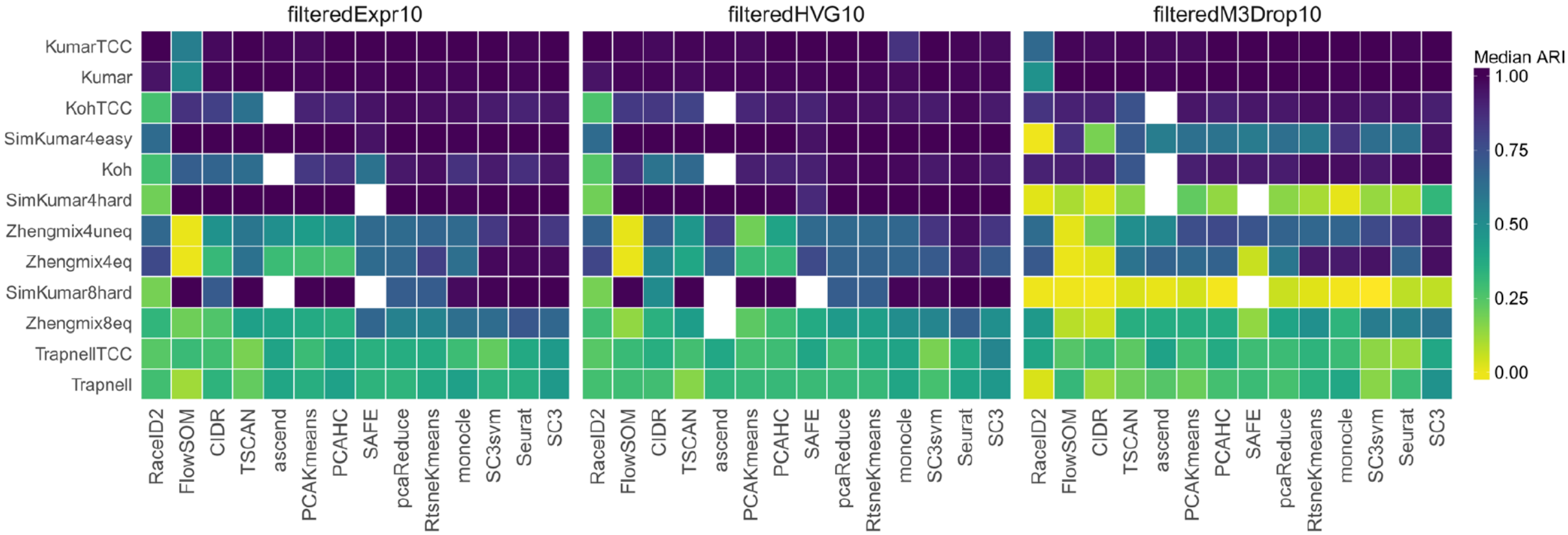




$$Q = \frac{1}{2m}\sum_{i,j}\left[W_{ij} - \frac{s_i s_j}{2m}\right]\delta(c_i, c_j)$$

Xu and Su (https://doi.org/10.1093/bioinformatics/btv088)
Levine et al. (https://doi.org/10.1016/j.cell.2015.05.047)

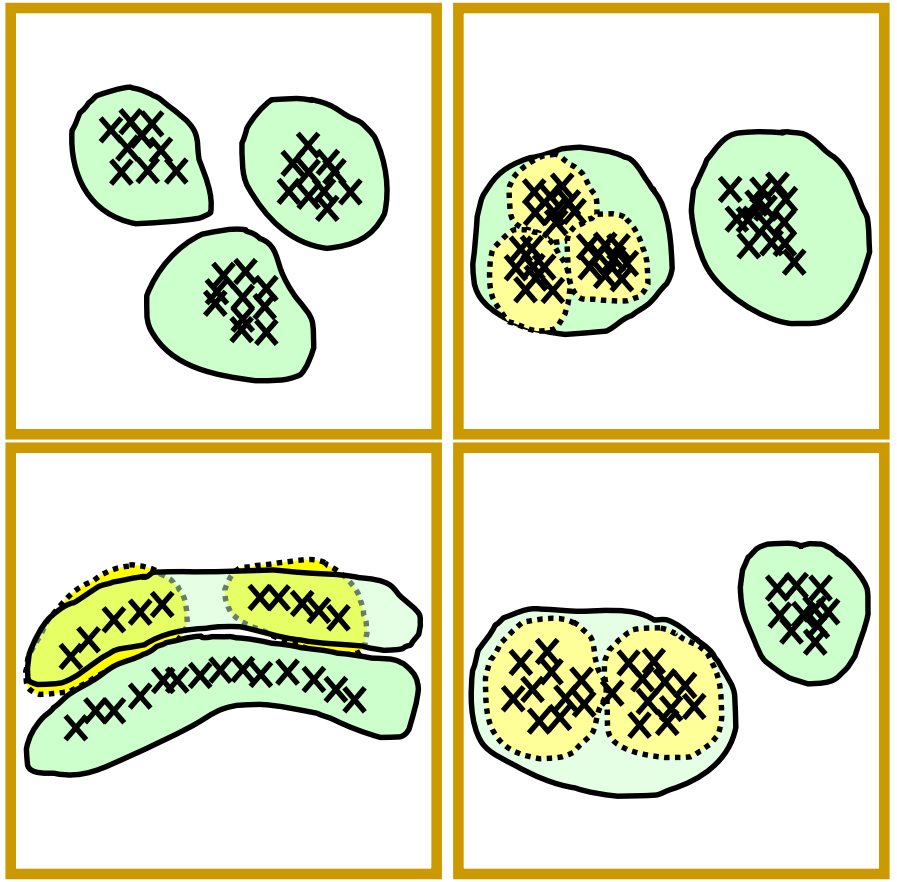
Benchmarking scRNA-seq clustering methods



Duò et al. (https://doi.org/10.12688/f1000research.15666.2)

Clustering is subjective!

- Principle choices
  - Similarity measure
  - Algorithm
- Different choice leads to different results
  - Subjectivity becomes reality
- Cluster process
  - Validate, interpret (generate hypothesis), repeat steps



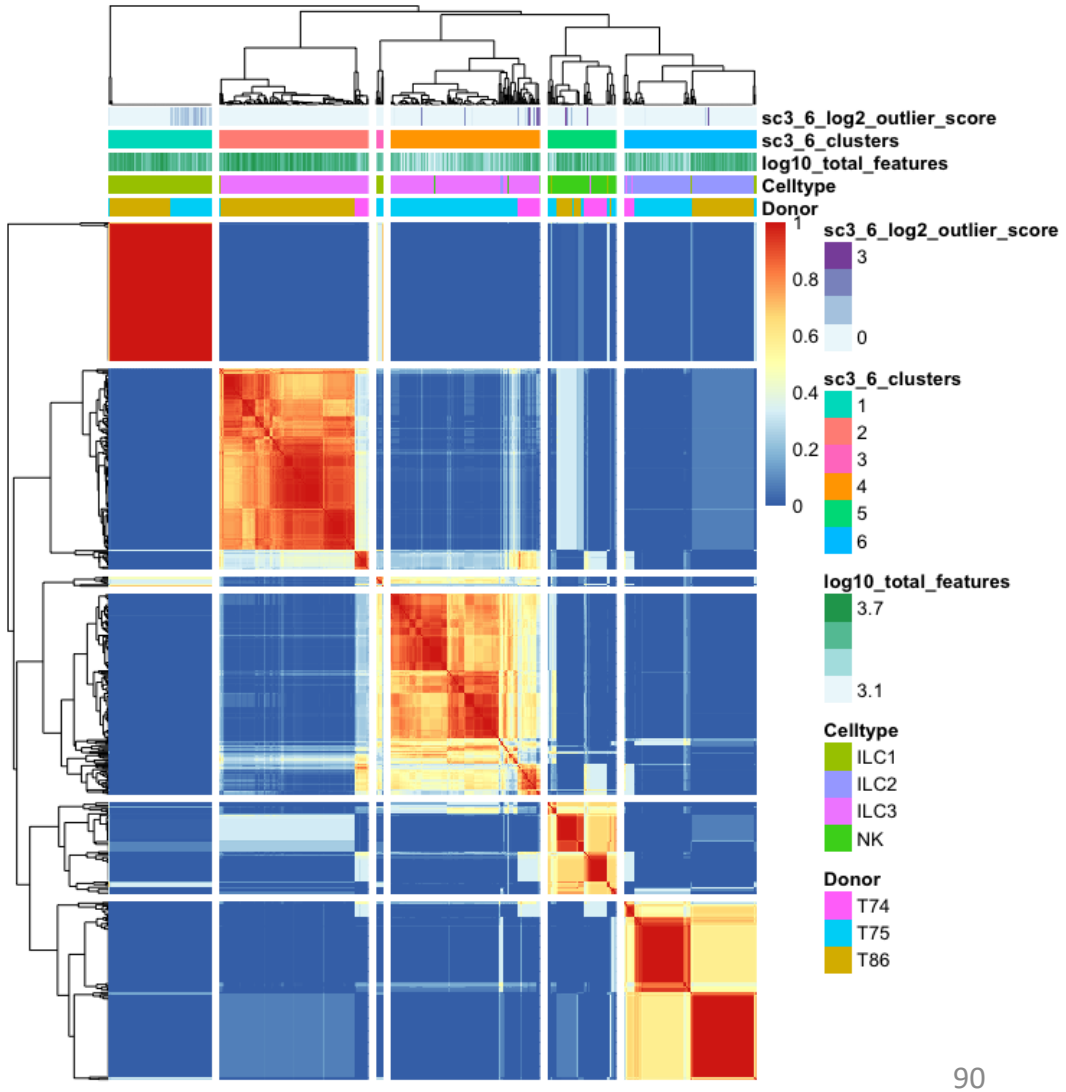
How many clusters do you really have?

- It is hard to know when to stop clustering – you can always split the cells more times.

- Can use:
  - Do you get any/many significant DE genes from the next split?
  - Some tools have automated predictions for number of clusters – may not always be biologically relevant

Always check QC data

- Is what your splitting mainly related to batches, qc-measures (especially detected genes)?



From clusters to cell identities

- Using lists of DE genes and prior knowledge of the biology

- Using lists of DE genes and comparing to other scRNAseq data or sorted cell populations

Databases with celltype gene signatures

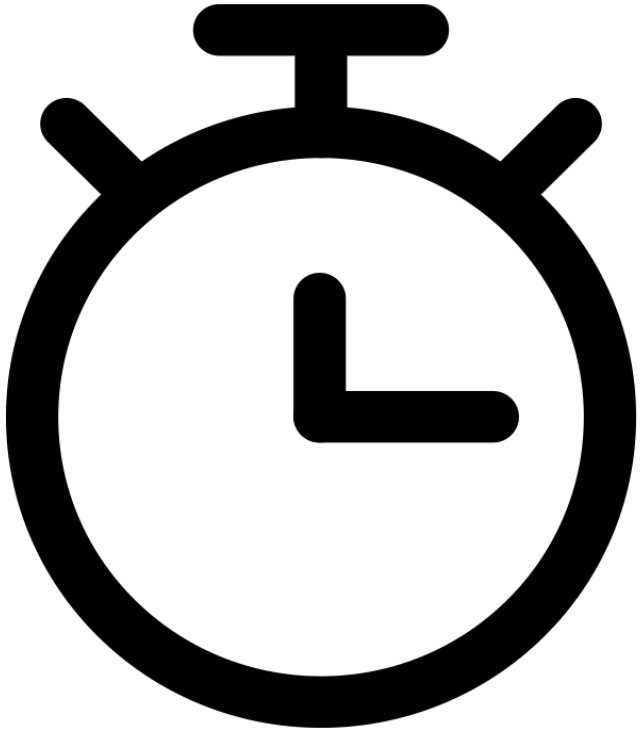
- PanglaoDB (https://panglaodb.se/)
  - Human: 295 samples, 72 tissues, 1.1 M cells
  - Mouse: 976 samples, 173 tissues, 4 M cells
  - Franzén et al (https://doi.org/10.1093/database/baz046)

- CellMarker (http://biocc.hrbmu.edu.cn/CellMarker/)
  - Human: 13,605 cell markers of 467cell types in 158 tissues
  - Mouse: 9,148 cell makers of 389 cell types in 81 tissues
  - Zhang et al. (https://doi.org/10.1093/nar/gky900)

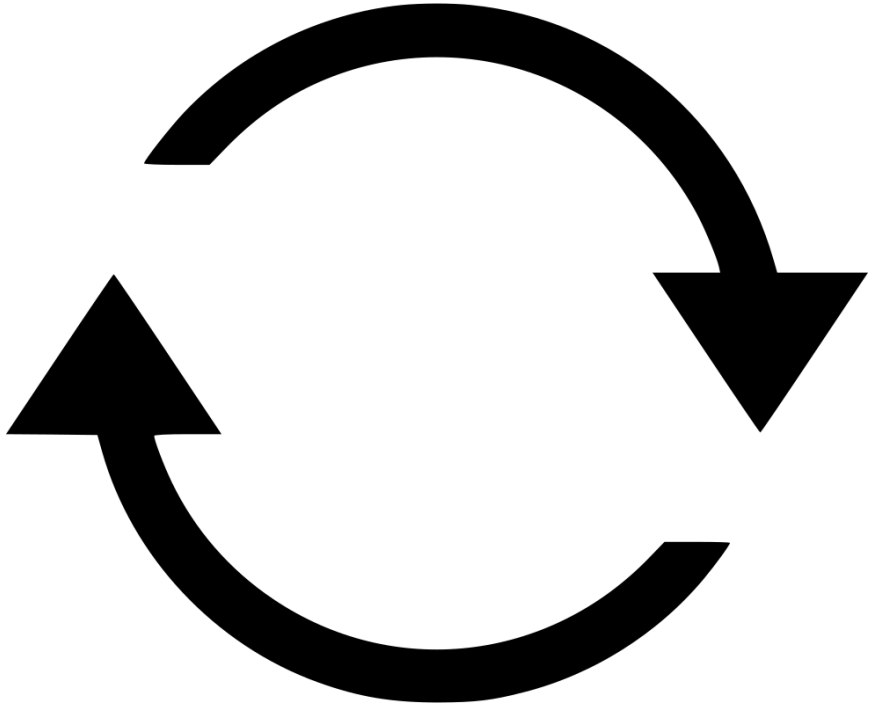
Challenges in clustering

- What is a cell type?
- What is the number of clusters *k*?
- **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

Limitations of unsupervised methods



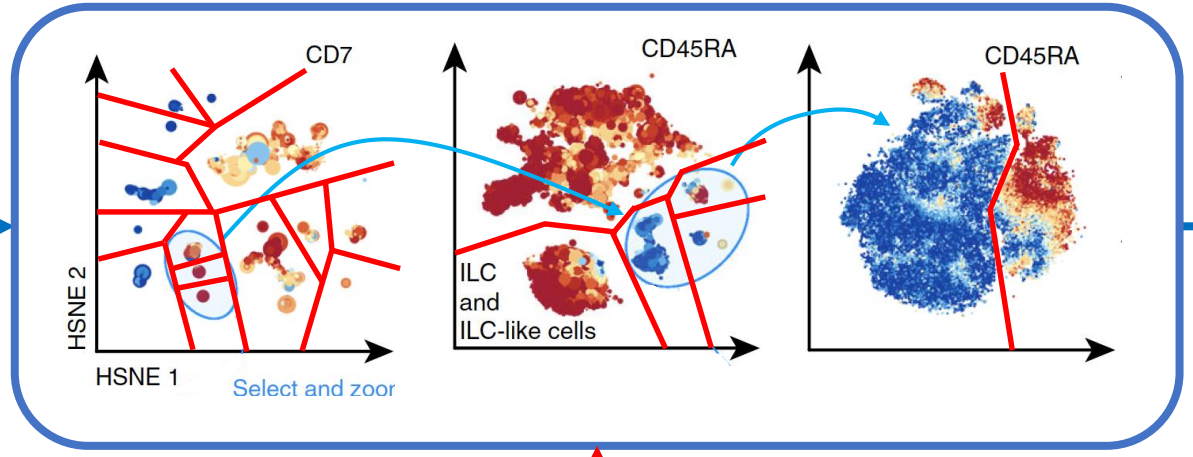
Time



Reproducibility

Supervised cell type identification

New cells



New cells
Annotated

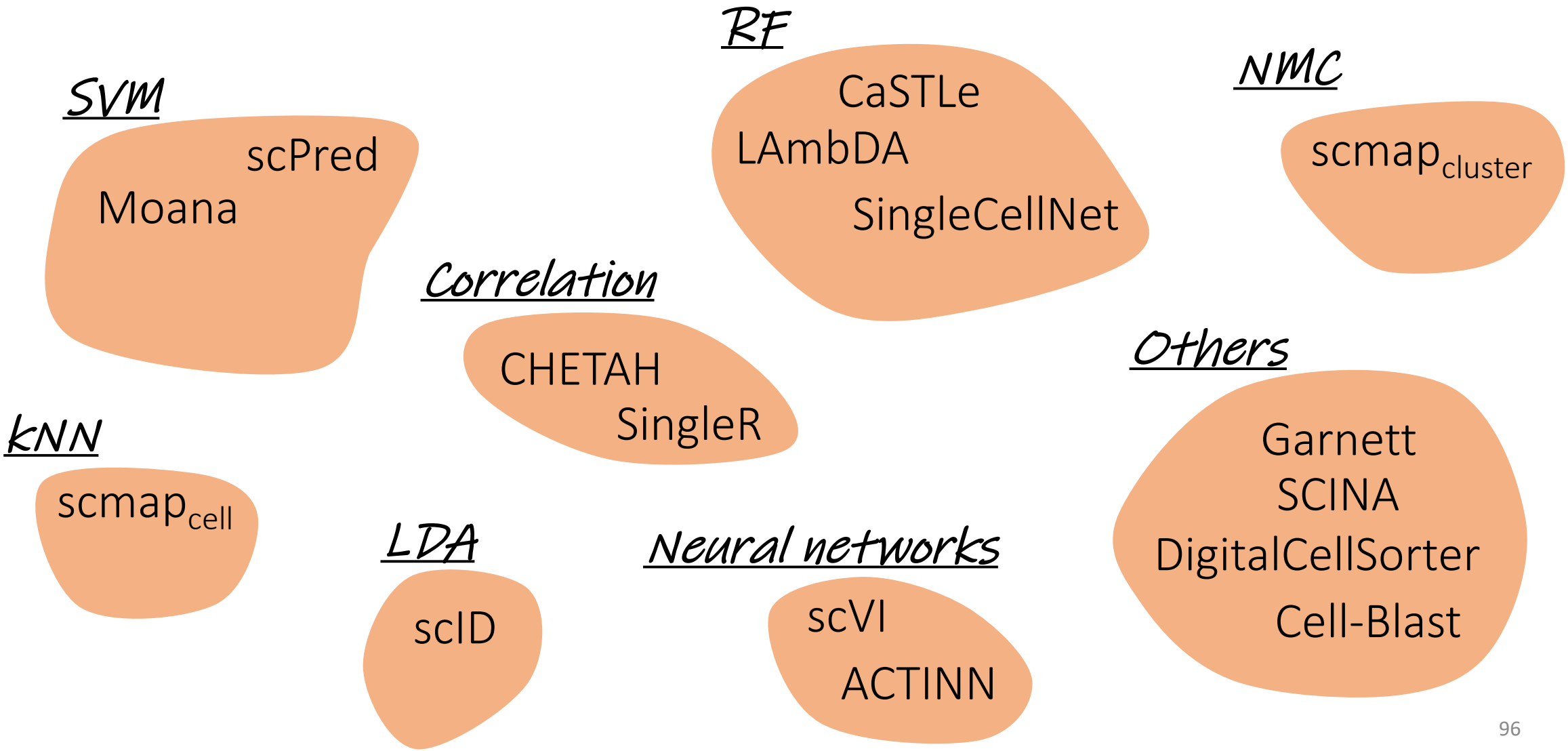
New cells
unassigned

Marker genes
file

Annotated
Cells

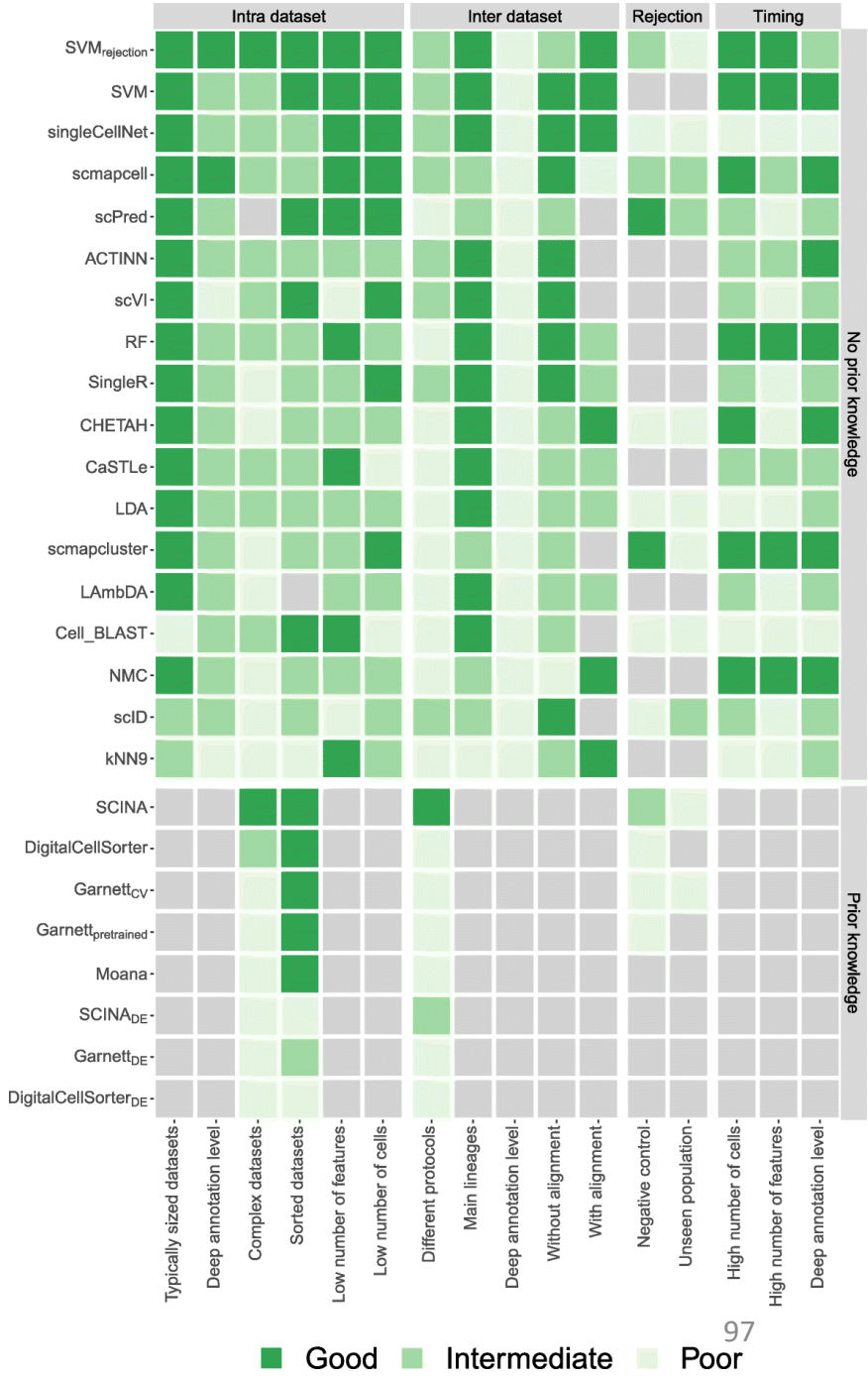
Atlas
Pilot data
First replicate
…

16 existing classifiers (April 2019)



Benchmarking

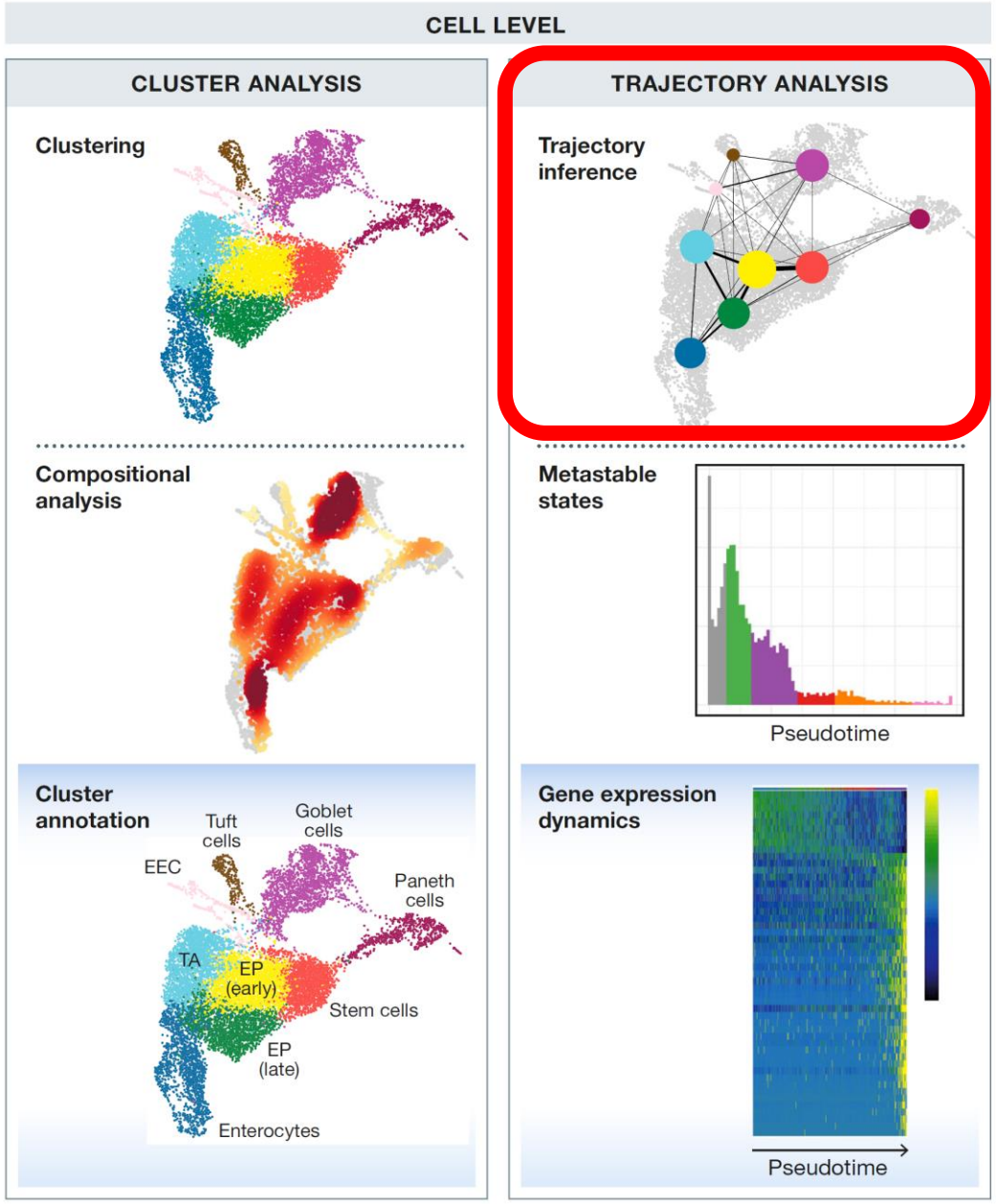
- Simple, off-the-shelf outperform special scRNA-seq classifiers
- Incorporating prior knowledge is not beneficial

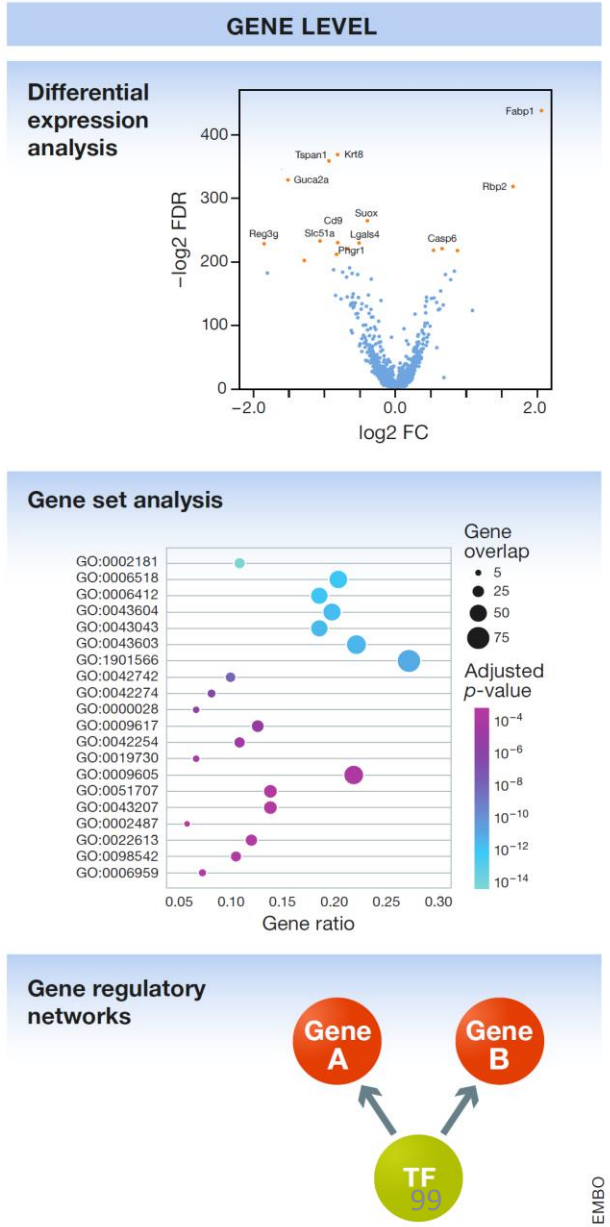


Abdelaal*, Michielsen* et al. (Genome Biology 2019)

End Part 2

scRNA-seq Downstream Analysis

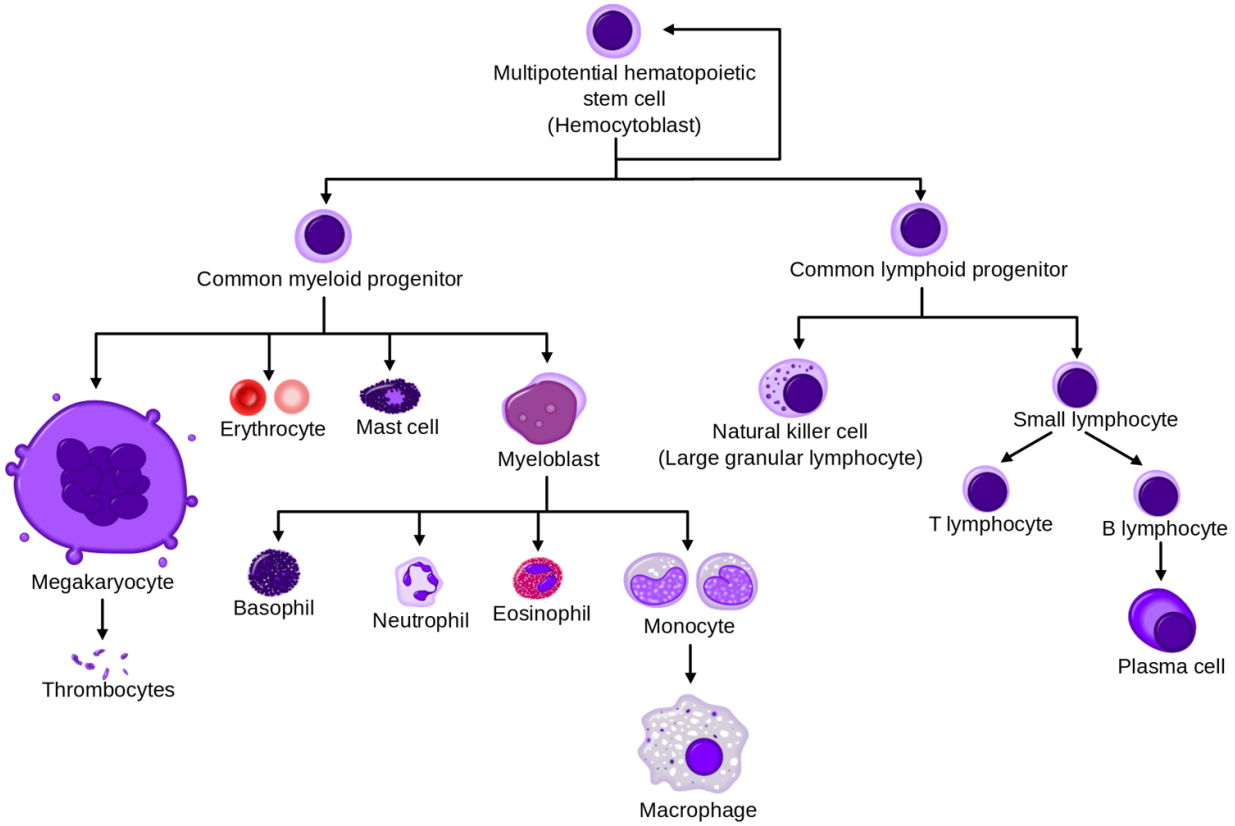




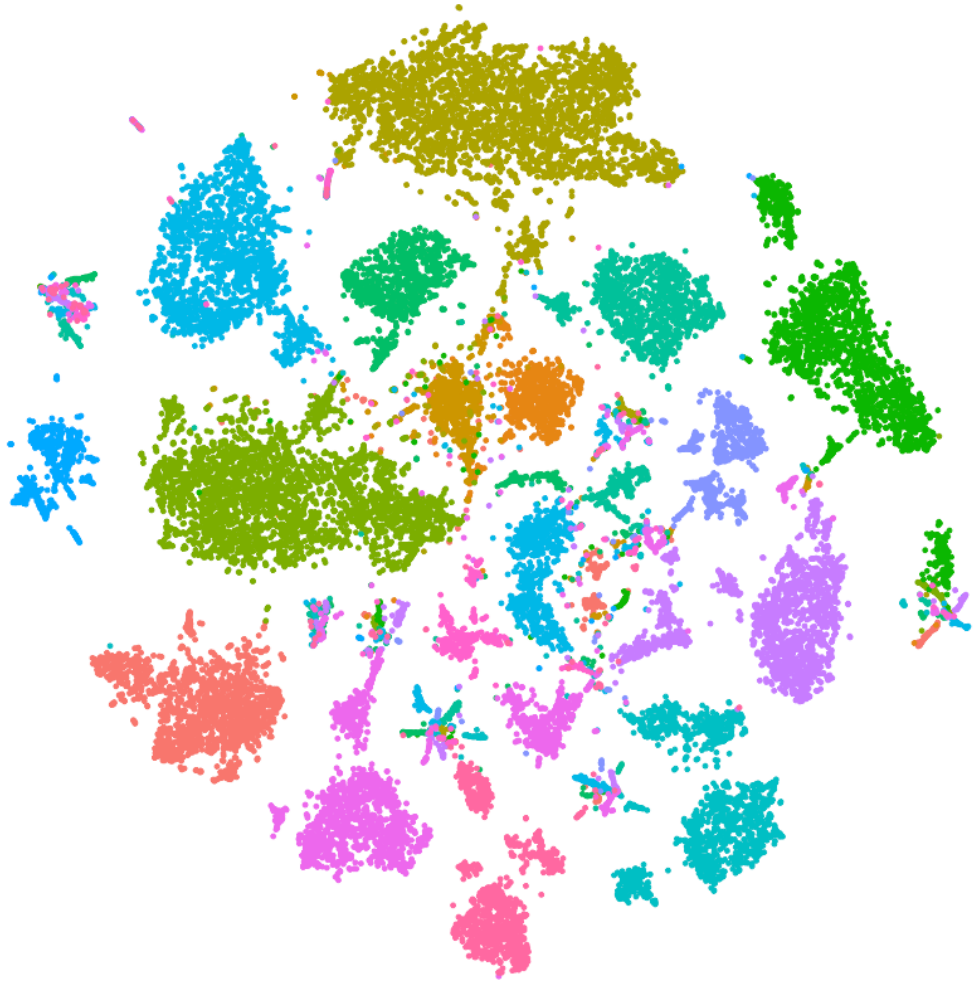
Luecken and Theis (MSB 2019)

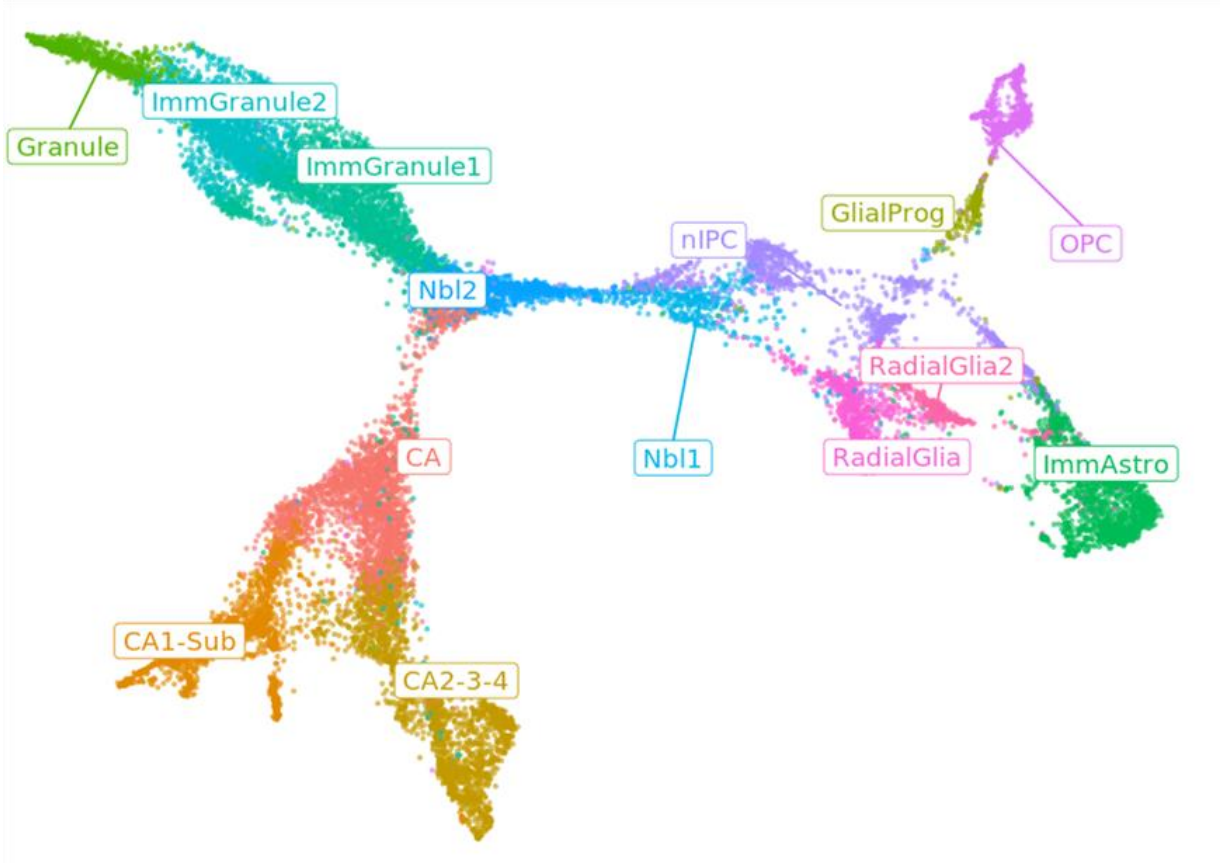
Differentiating cells

- Cells that differentiate display a continuous spectrum of states
- Individual cells will differentiate in an unsynchronized manner
  -> Each cell is a snapshot of differentiation time



Clustering of differentiating cells

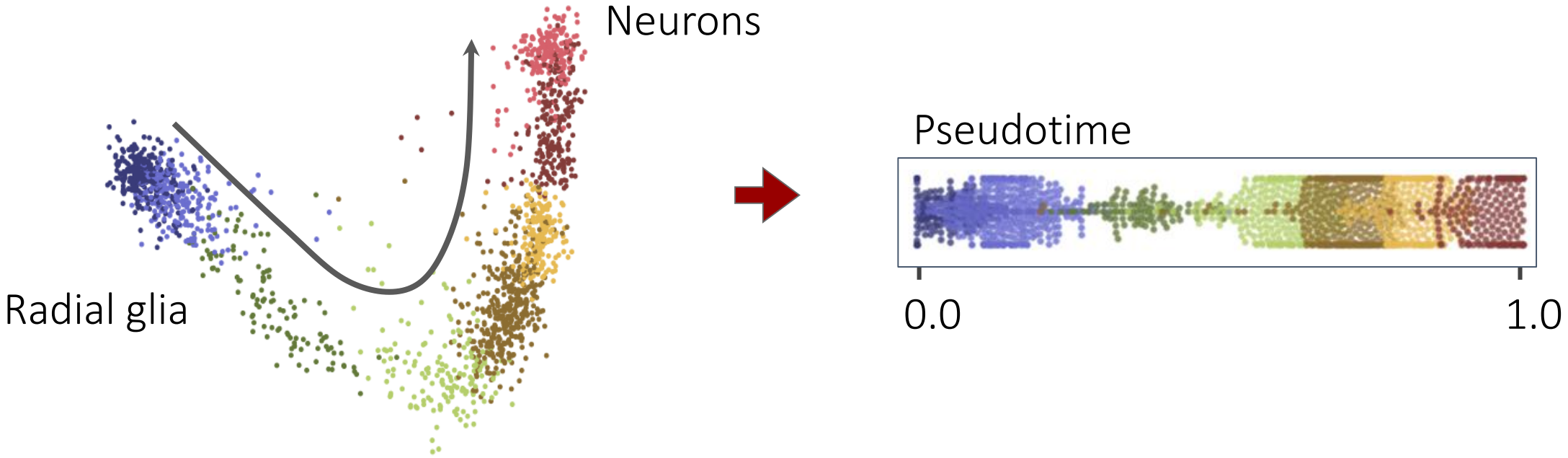




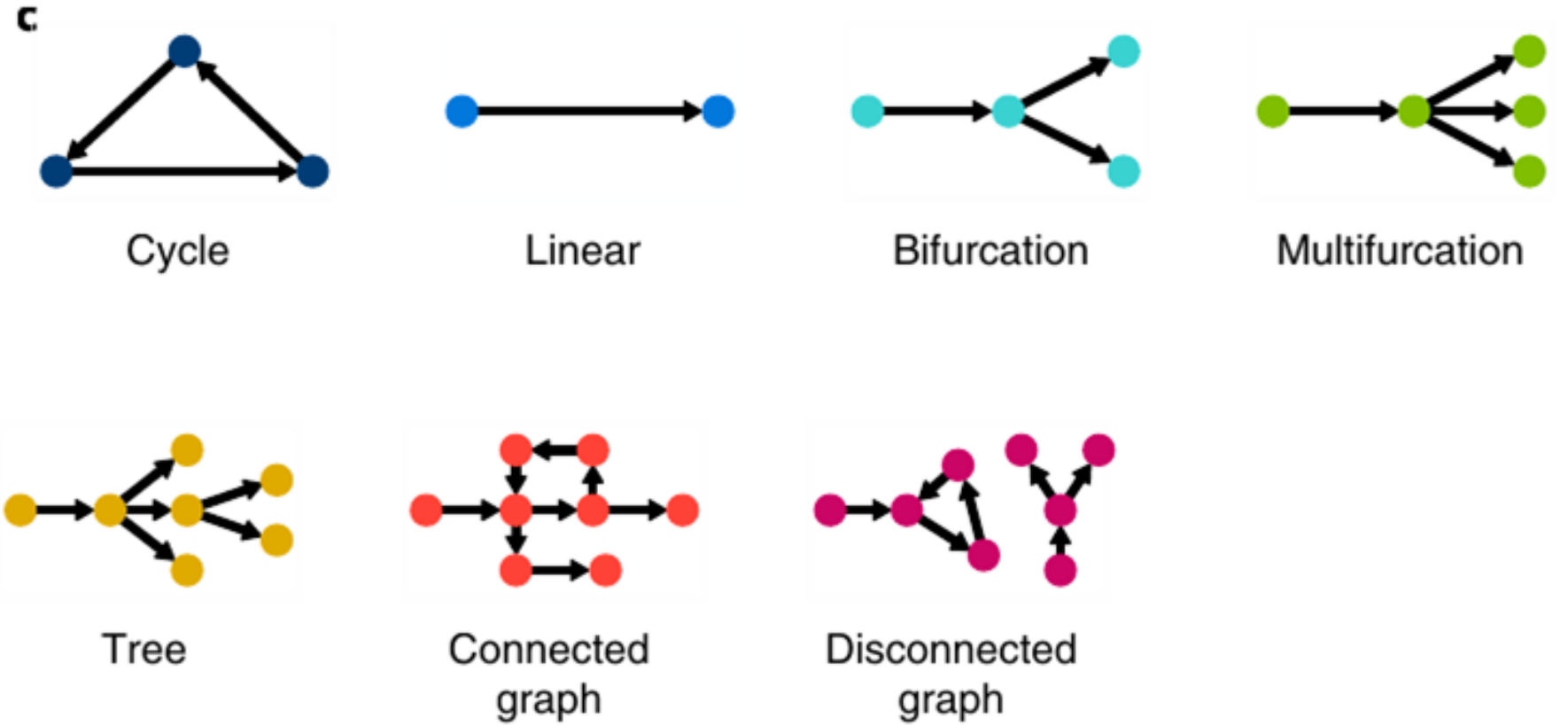
https://dimension-reduction.github.io/

Trajectory / pseudotime inference

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



Trajectory structure

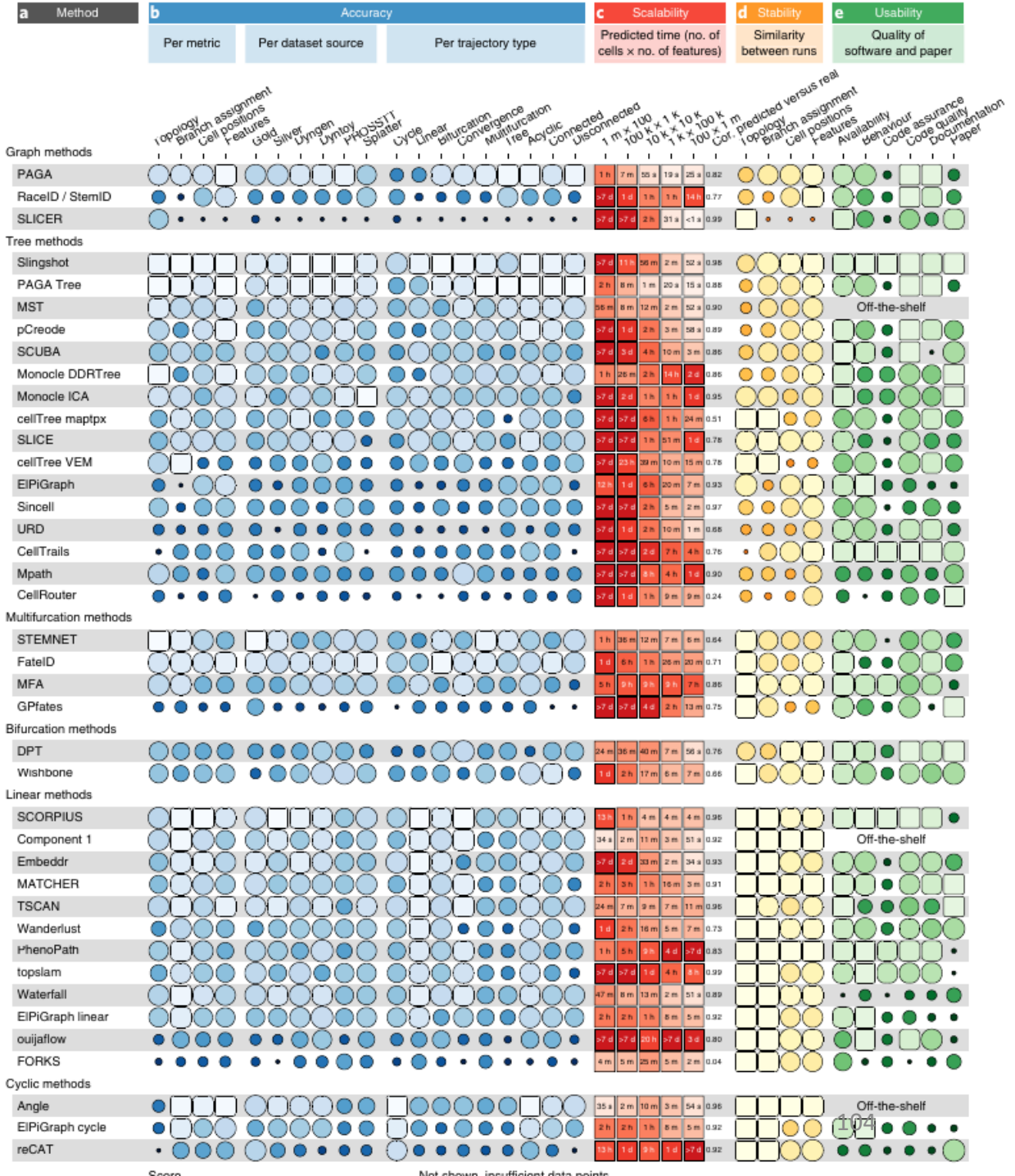


Saelens et al. (Nature Biotechnology 2018)

Many many trajectory inference methods

- Comprehensive evaluation by Saelens et al.

- Interactive website to select best tool for your data
  - http://guidelines.dynverse.org/



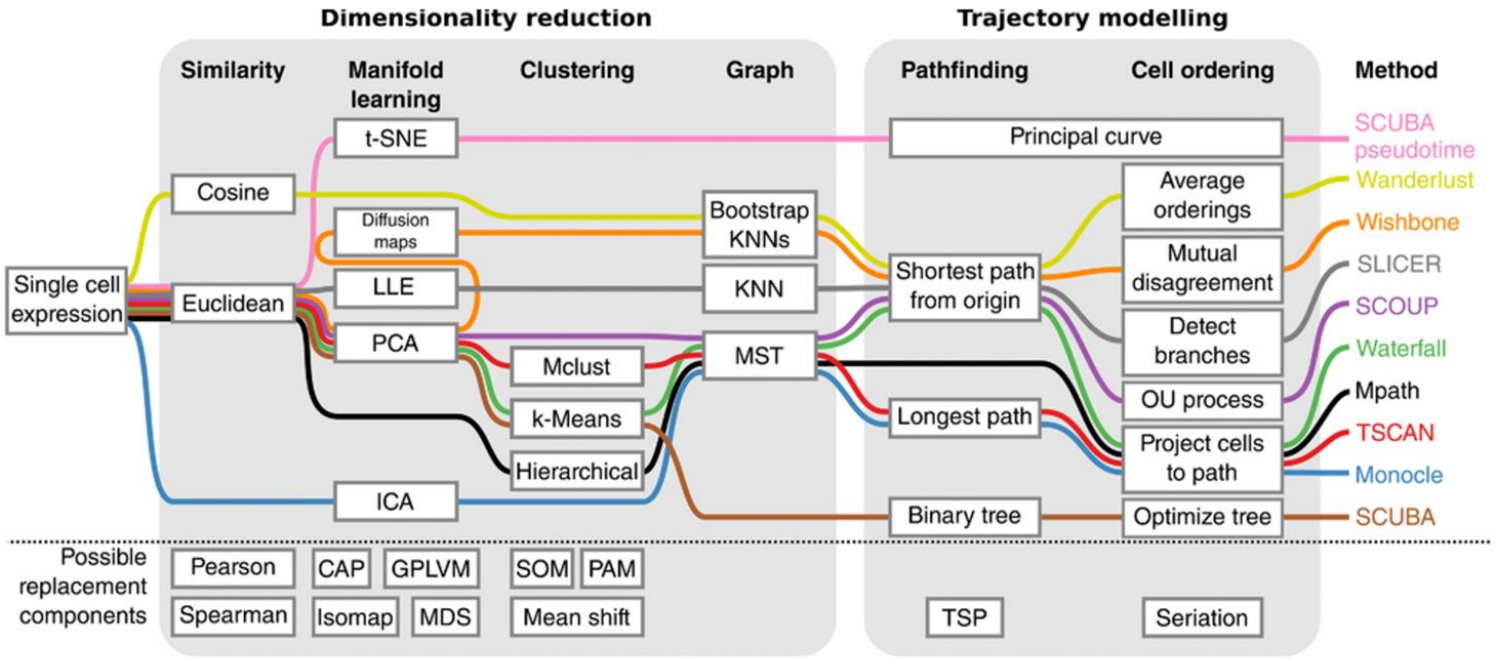
Saelens et al (2019) Nat Biotechnology

Trajectory inference methods

- To be discussed:
  - Monocle 1 (Trapnell et al., Nature Biotech 2014)
  - Slingshot (Street et al., BMC Genomics 2018)
  - Monocle 2 (Qiu et al., Nature Methods 2017)
  - Ouija (Campbell & Yau Bioinformatics 2019)

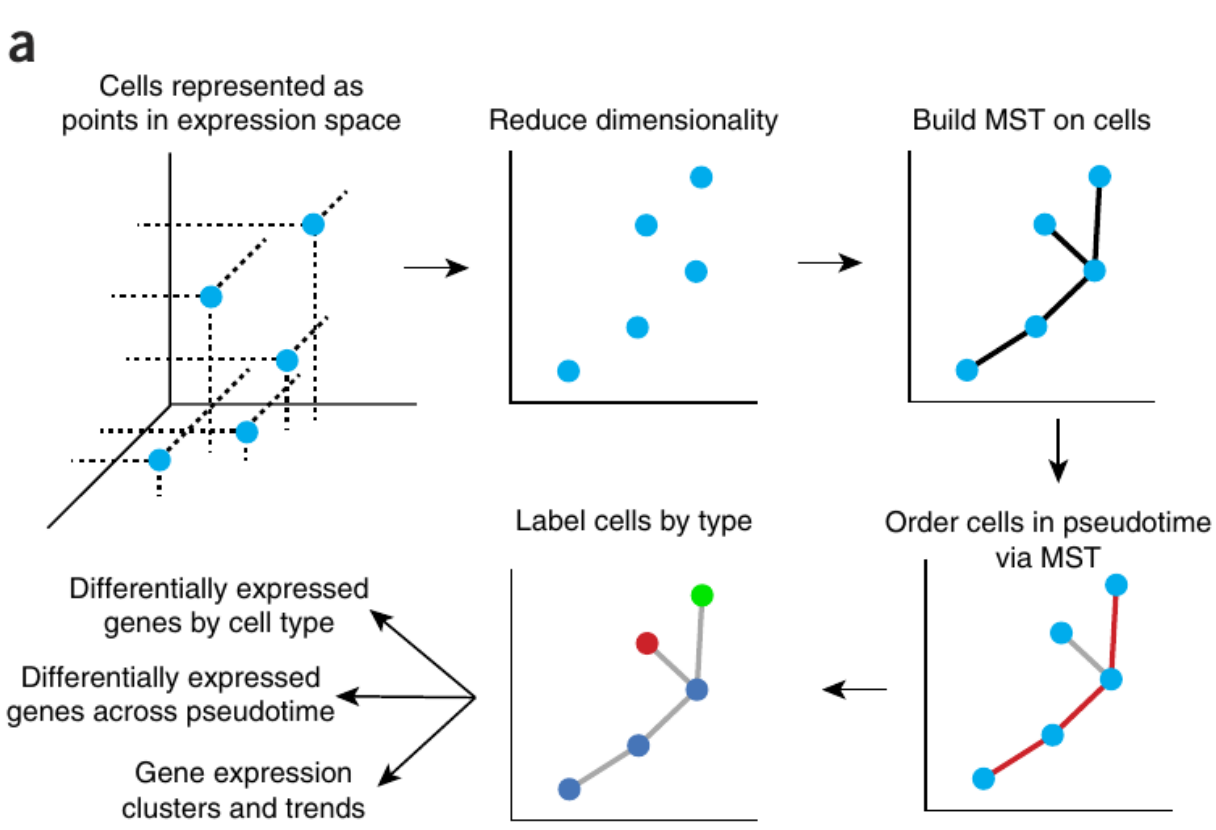
General trajectory inference pipeline

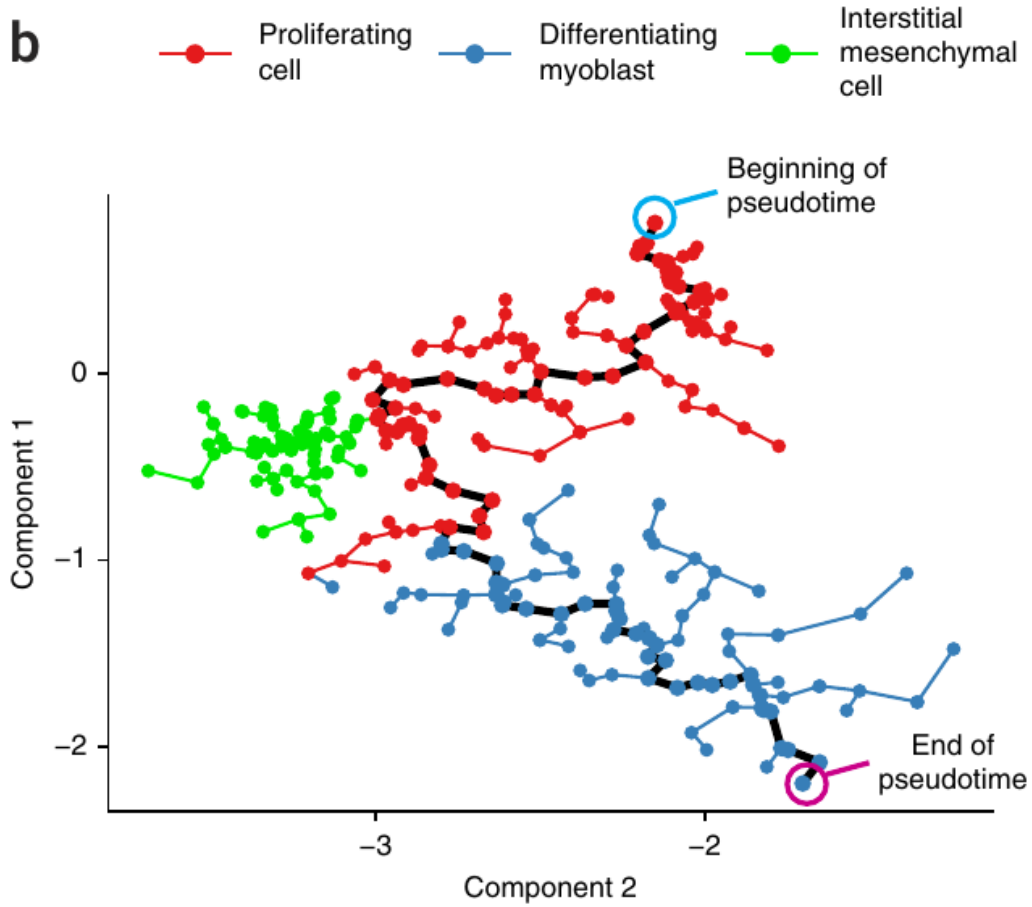
1. Dimensionality reduction
2. Trajectory fitting
3. Pseudotime assignment



Cannoodt et al, 2016

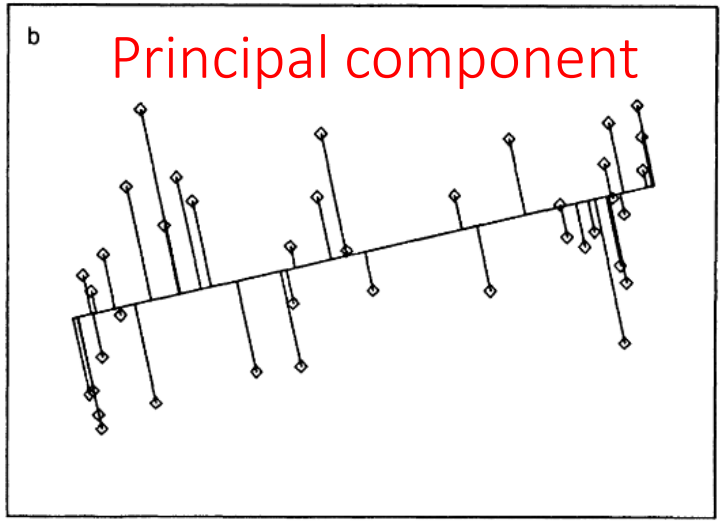
Minimum spanning trees
(Monocle 1)

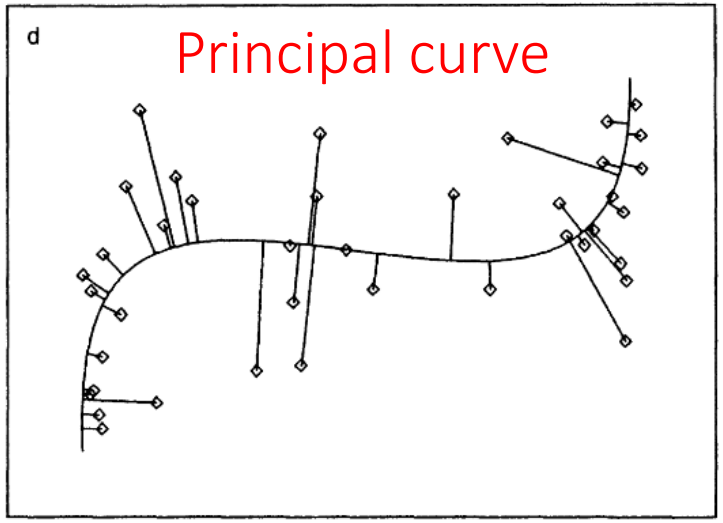




Trapnell et al., (Nature Biotech 2014)

Principle curves





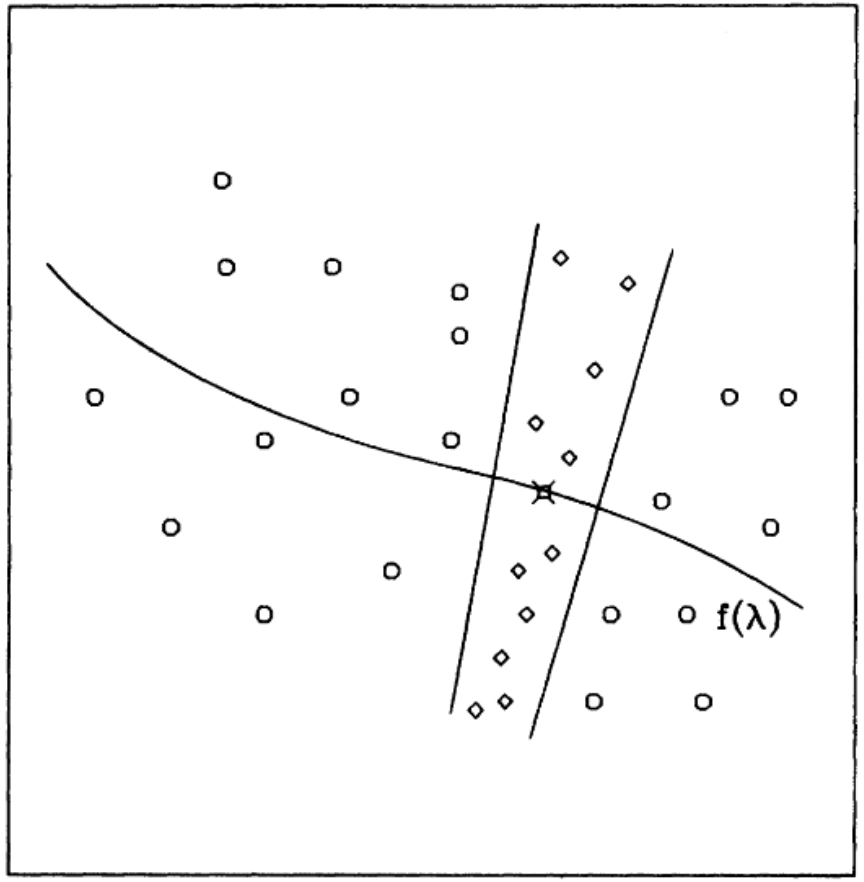
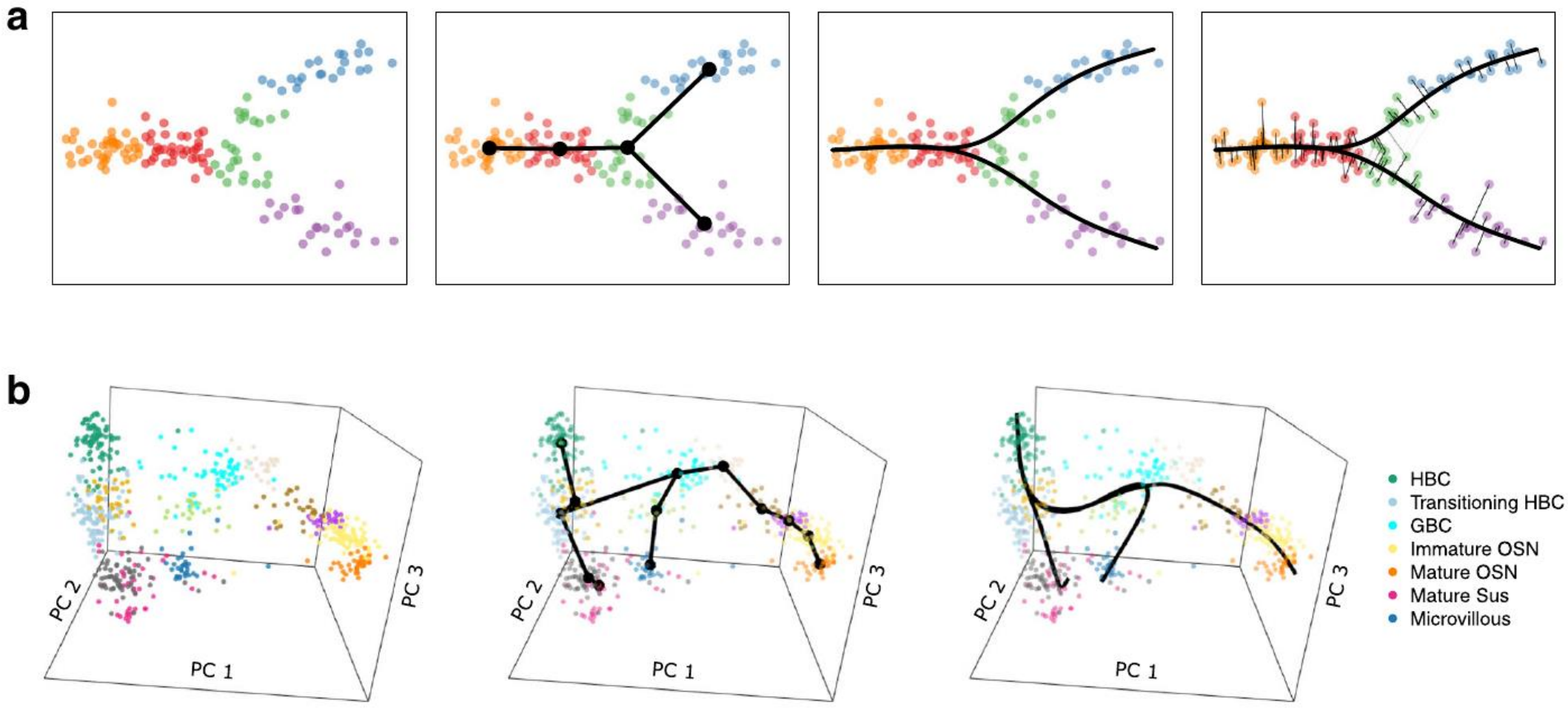


Figure 3. Each point on a principal curve is the average of the points that project there.

Hastie, Stuetzle, Principal curves, 1989

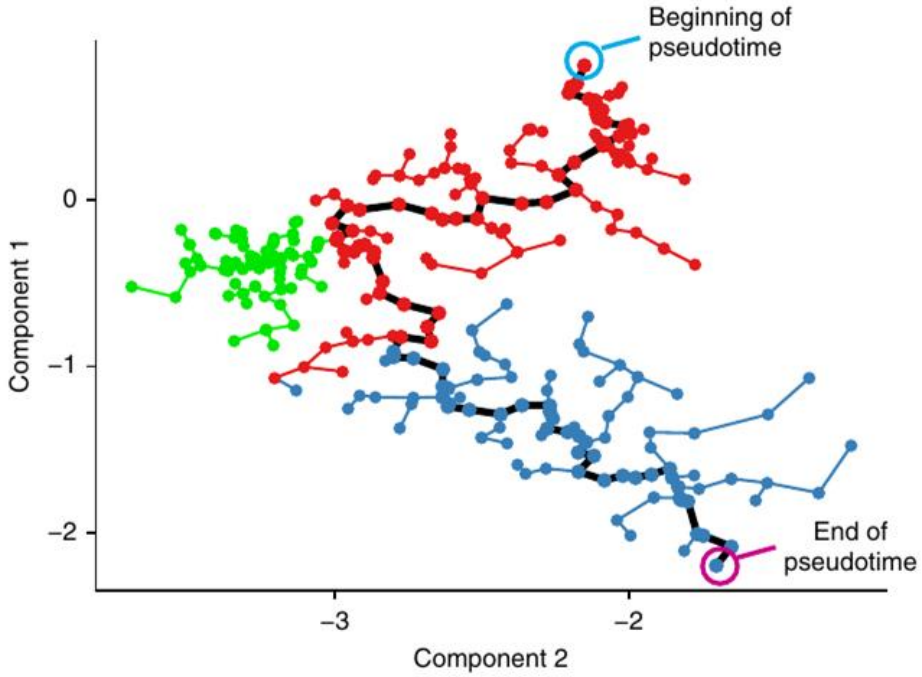
Slingshot

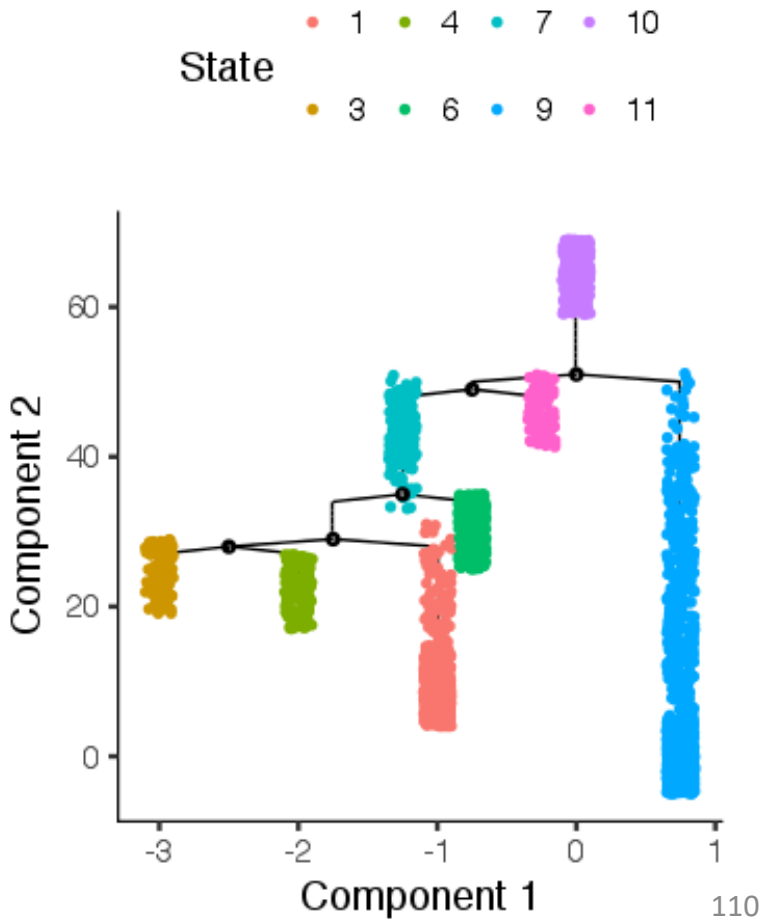


Street et al. (BMC Genomics 2018)

Monocle 2

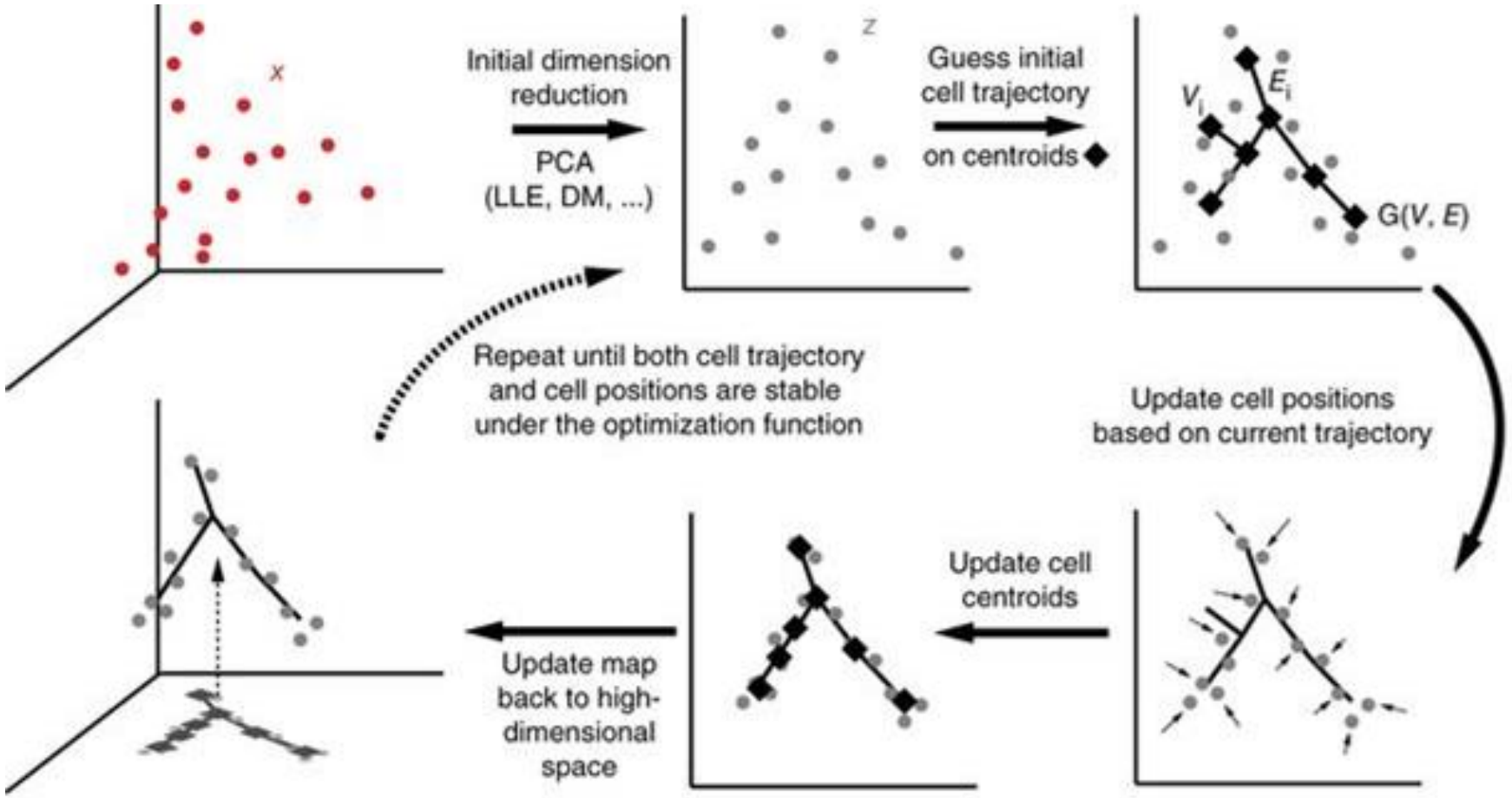
- End goal: Fit any arbitrary graph on the data





Qiu et al., Nature Methods 2017

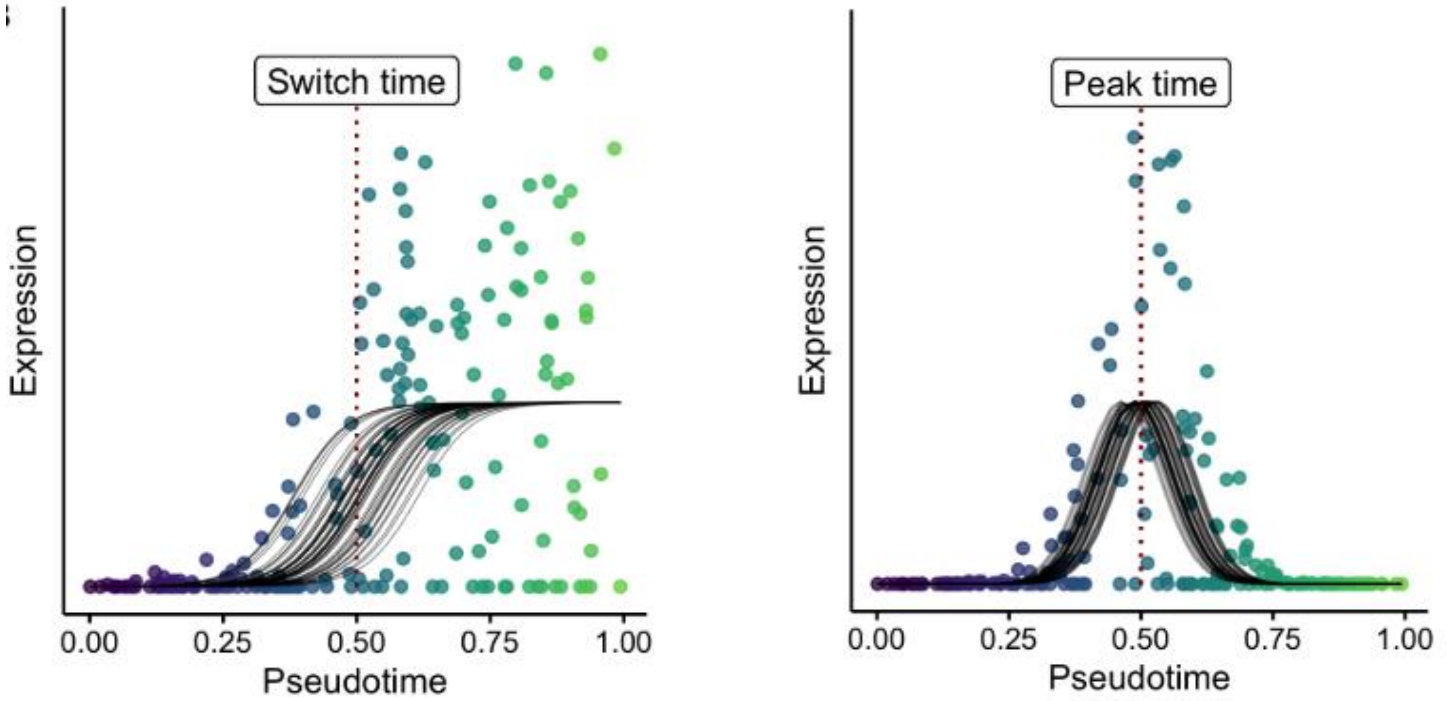
Monocle 2



Qiu et al., Nature Methods 2017

Ouija

- Model a small set of marker genes instead of fitting trajectory on complete transcriptome
- Switch focus to interpretability



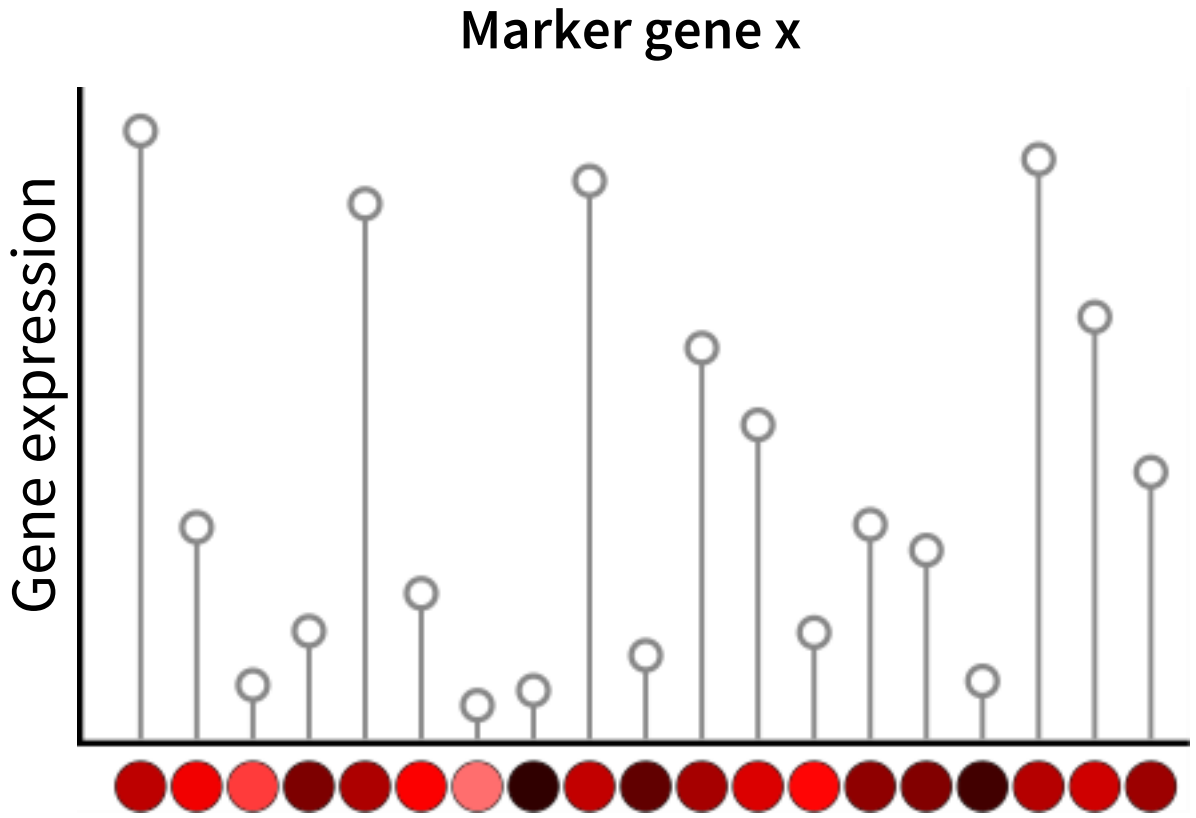
Campbell & Yau (Bioinformatics 2019)

Ouija intuition

True ordering:

Random cell ordering:

Goodness-of-fit: low

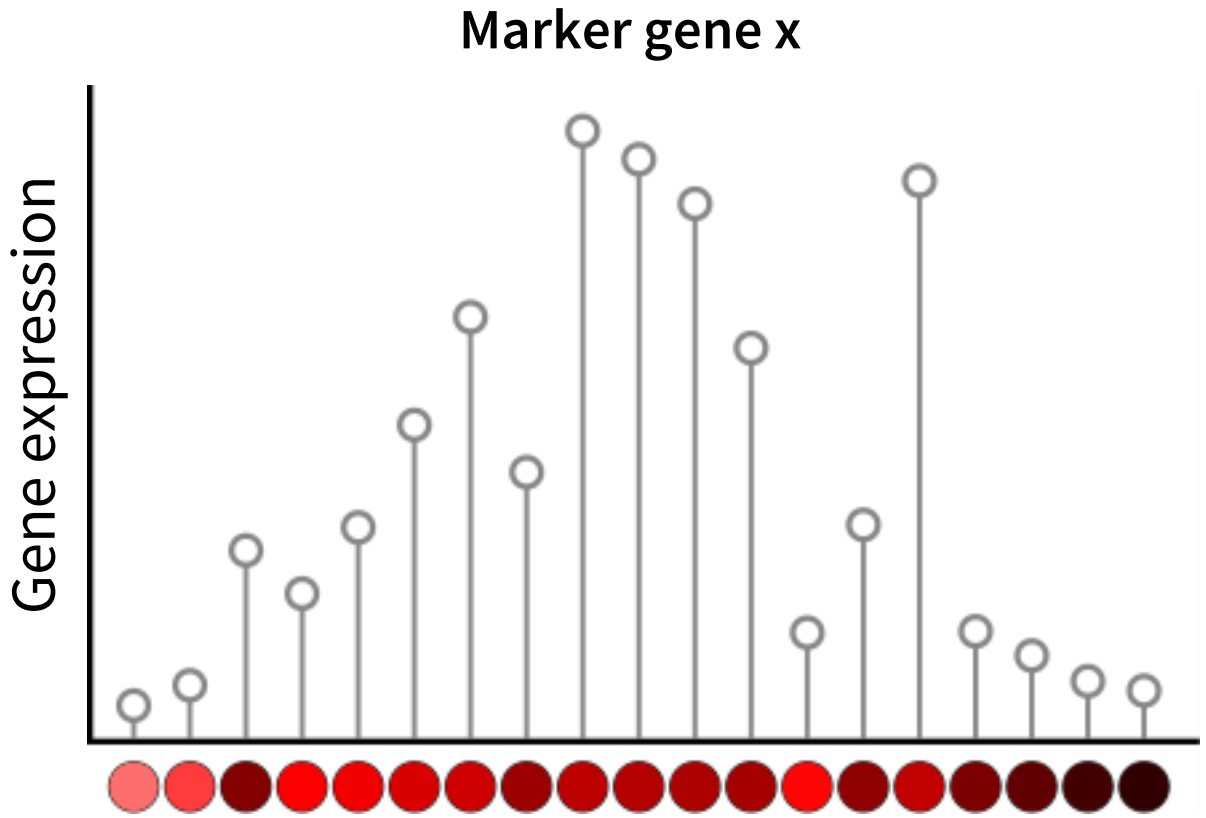


Ouija intuition

True ordering:

Optimize iteration: 100

Goodness-of-fit: mid

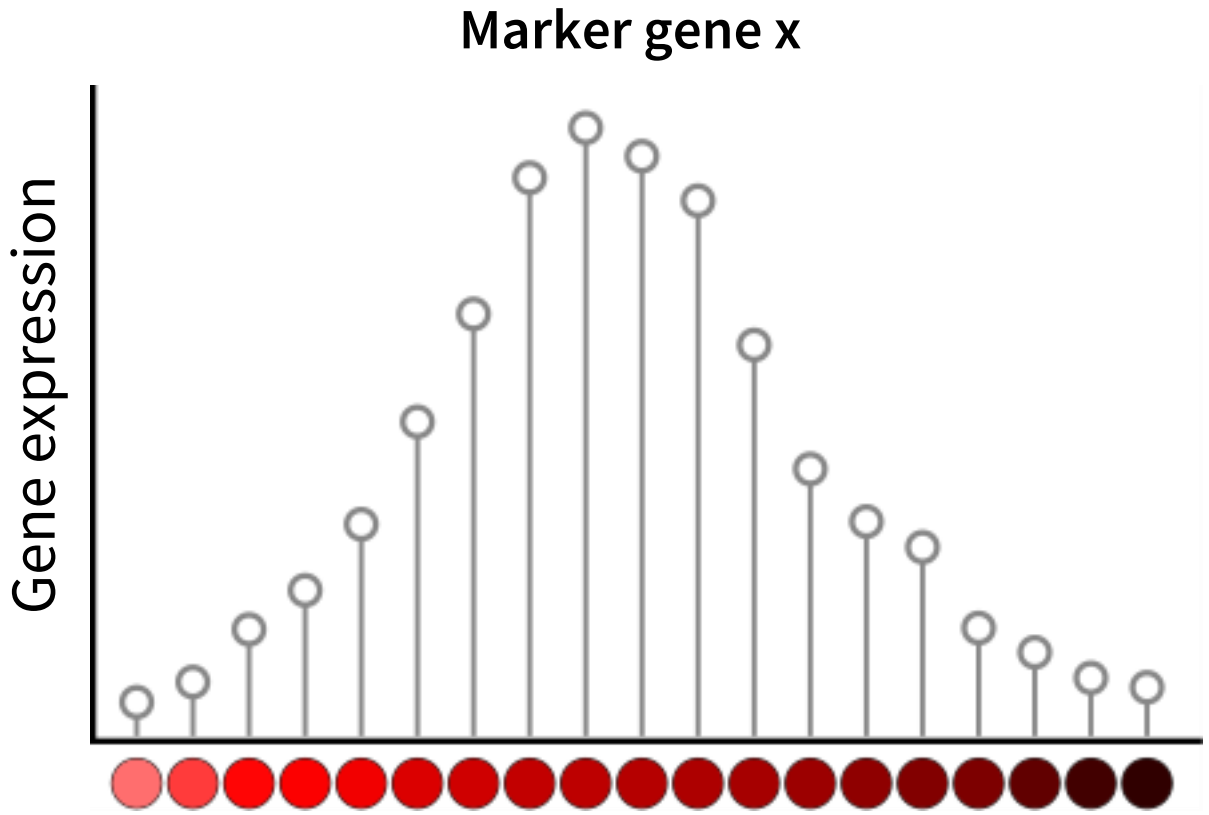


Ouija intuition

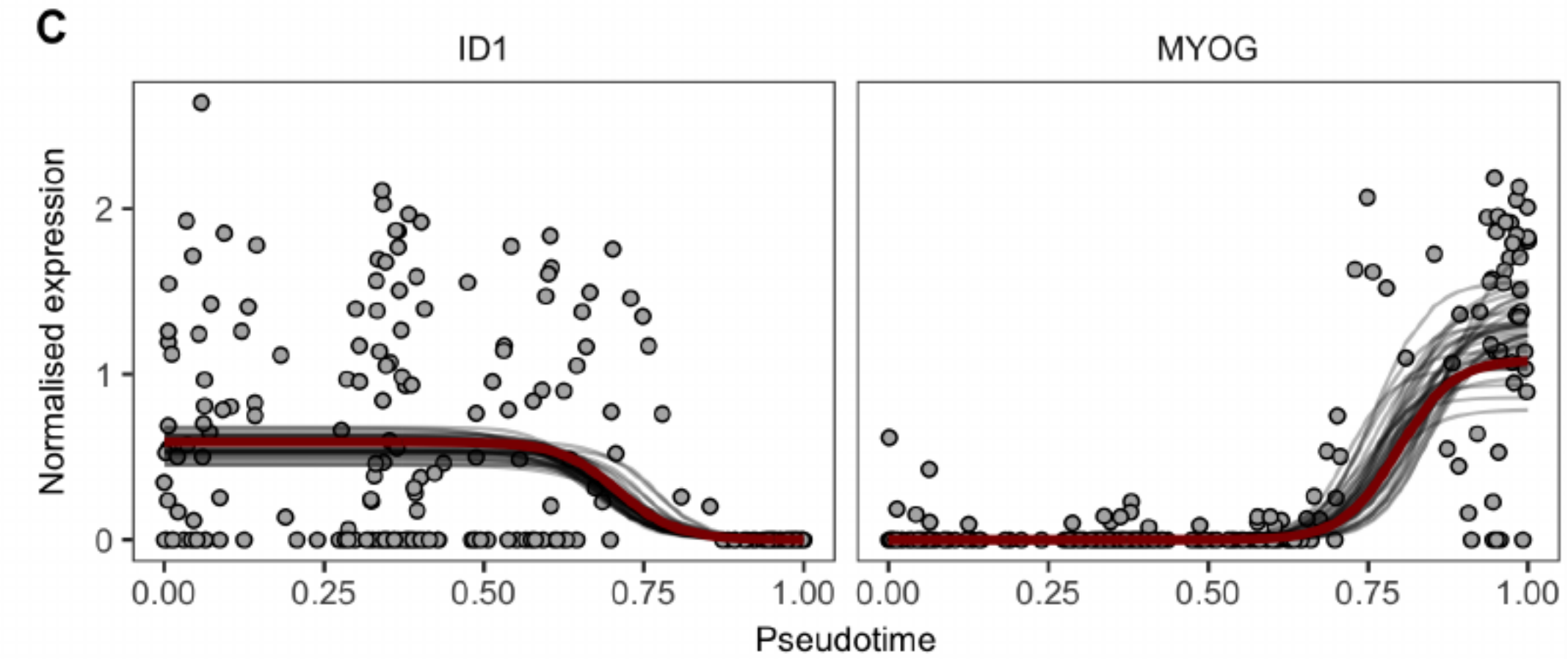
True ordering:

Optimize iteration: 500

Goodness-of-fit: high

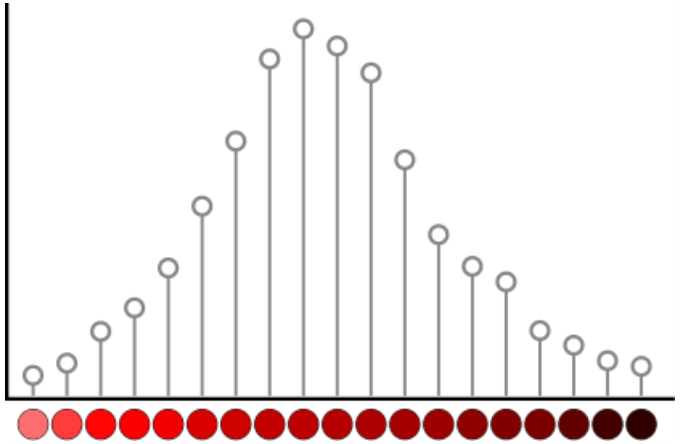


Ouija probabilistic modelling

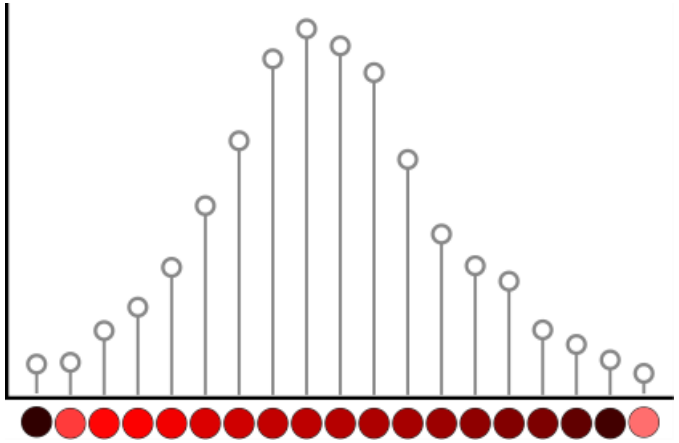


Campbell & Yau (Bioinformatics 2019)

Marker gene x

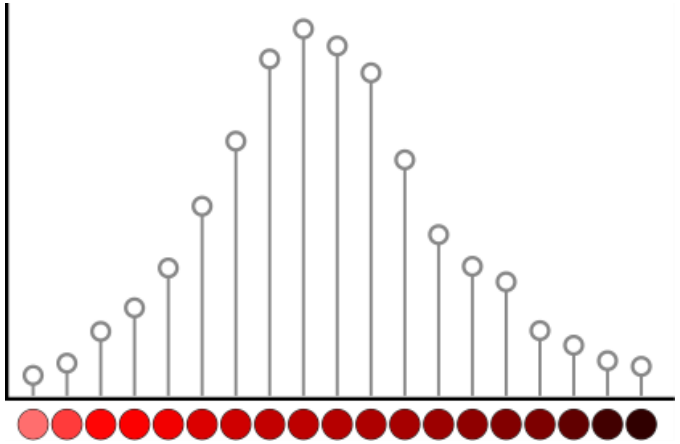


Goodness-of-fit: high



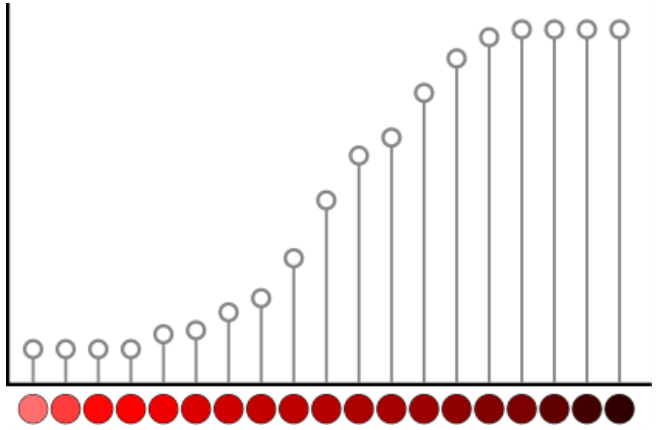
Goodness-of-fit: high

Marker gene x

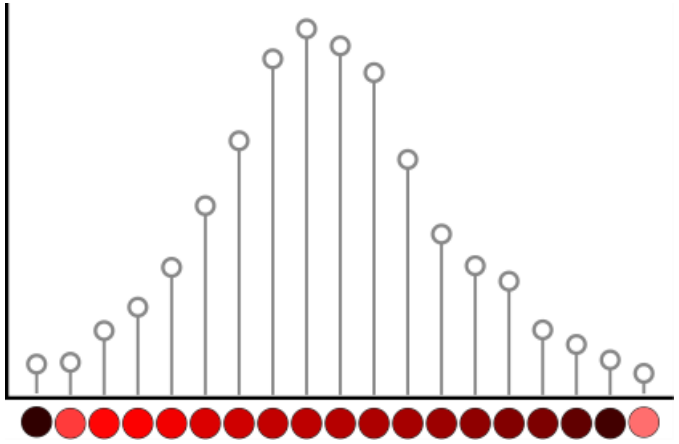


Goodness-of-fit: high

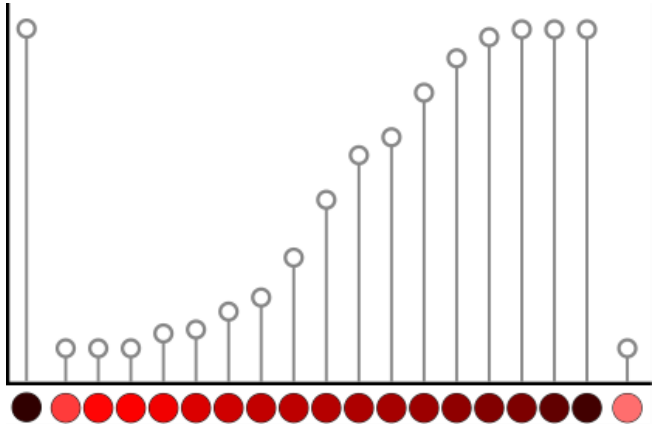
Marker gene y



Goodness-of-fit: high



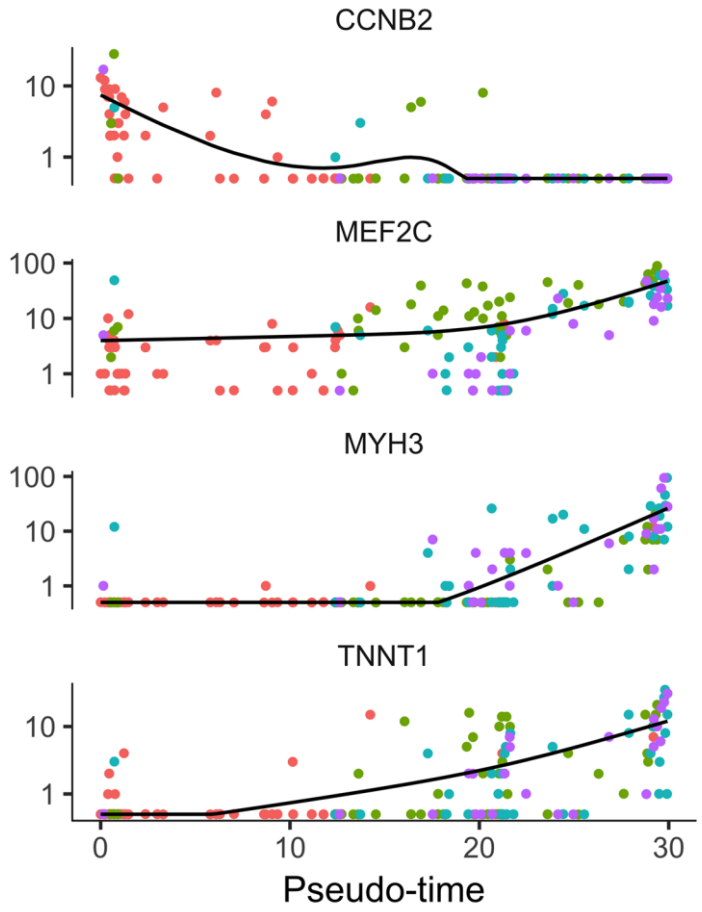
Goodness-of-fit: high



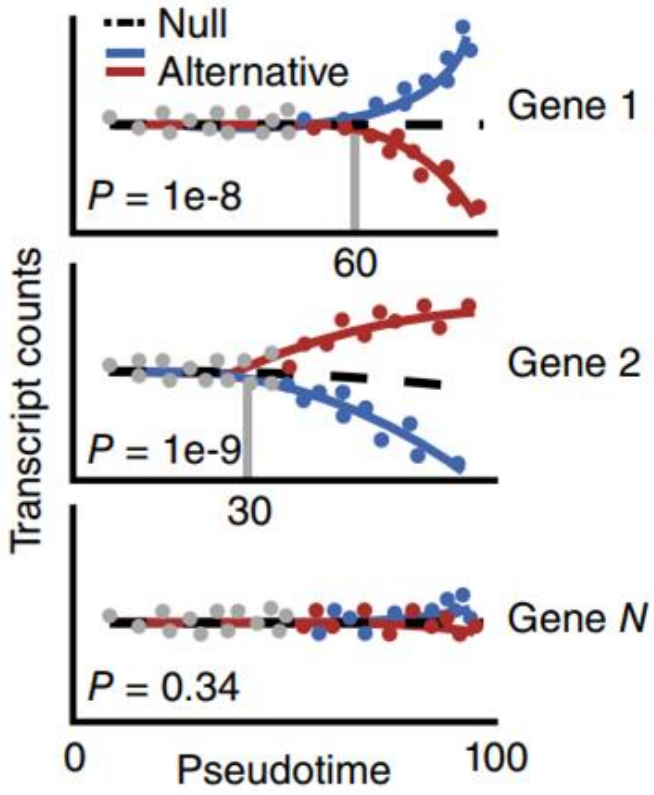
Goodness-of-fit: mid

Pseudotime analysis

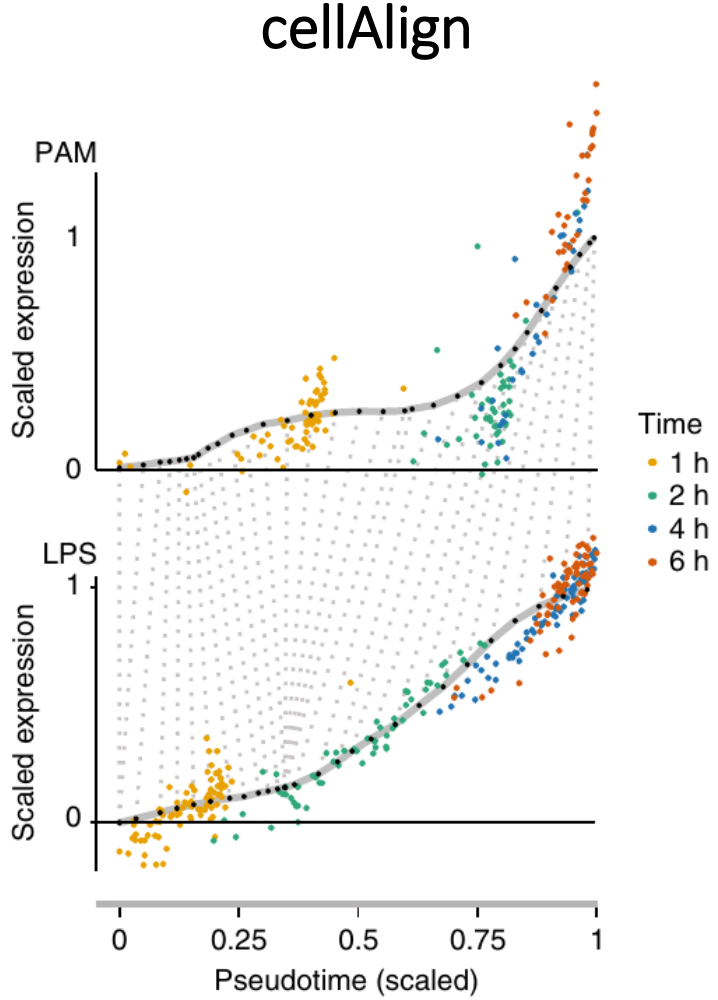
Pseudotime-gene relation



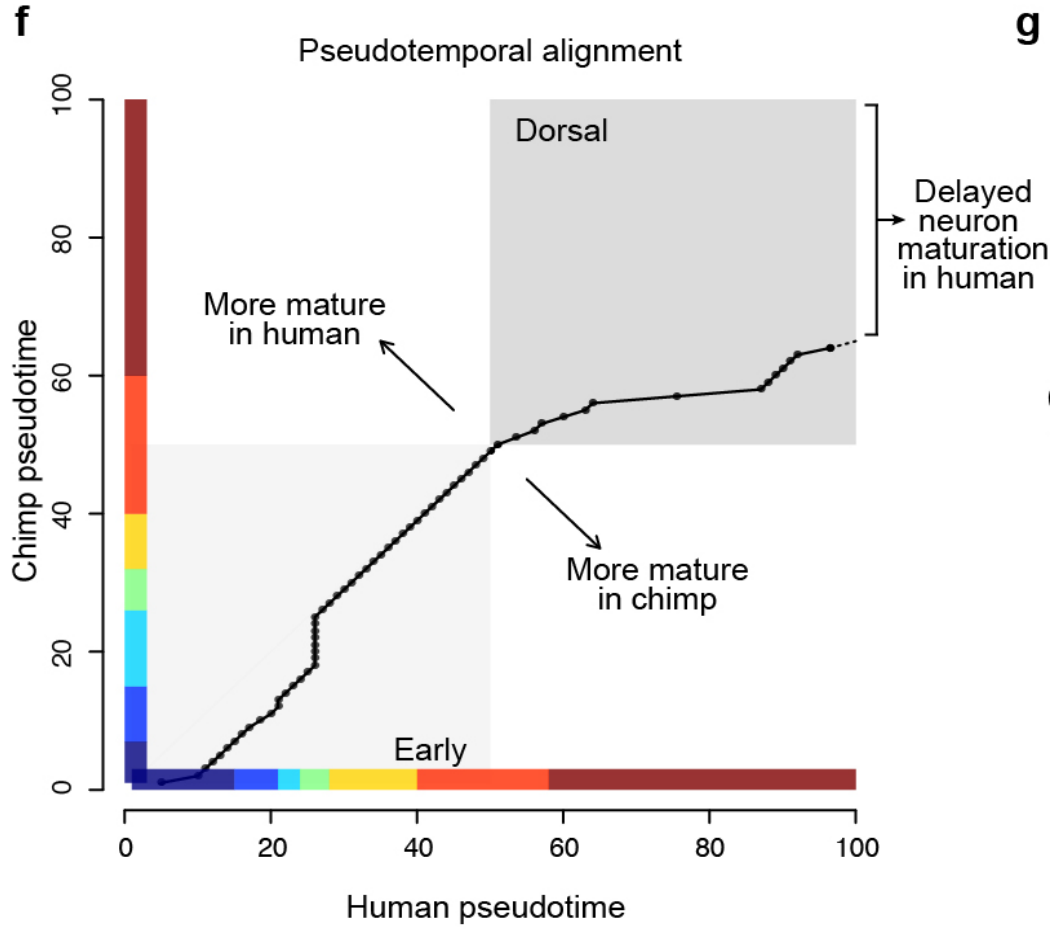
Detect branching genes



Comparing trajectories



Alpert et al. (Nature Methods 2018)



Kanton, Boyle, He et al. (Nature 2019)

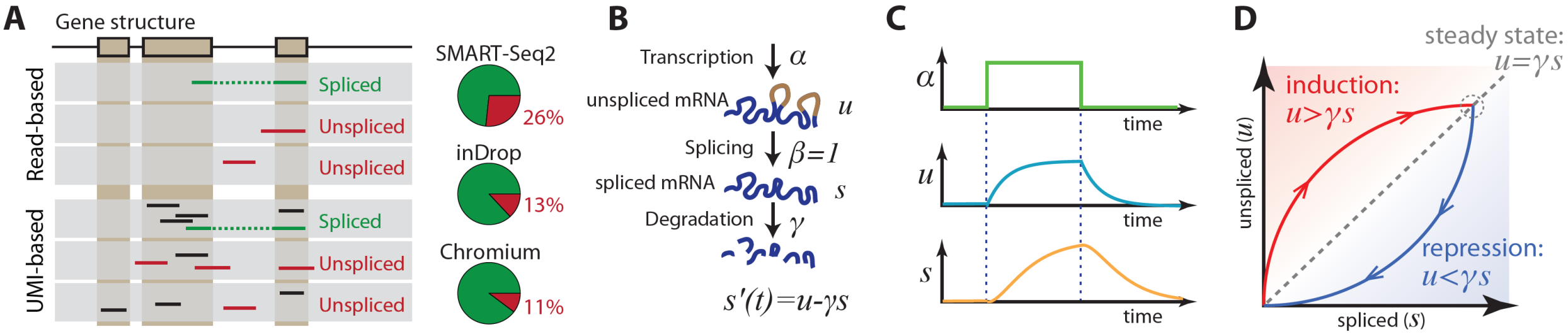
Is pseudotime inference a solved problem?

Fundamental limits on dynamic inference from single-cell snapshots

Caleb Weinreb[a], Samuel Wolock[a], Betsabeh K. Tusi[b], Merav Socolovsky[b], and Allon M. Klein[a,1]

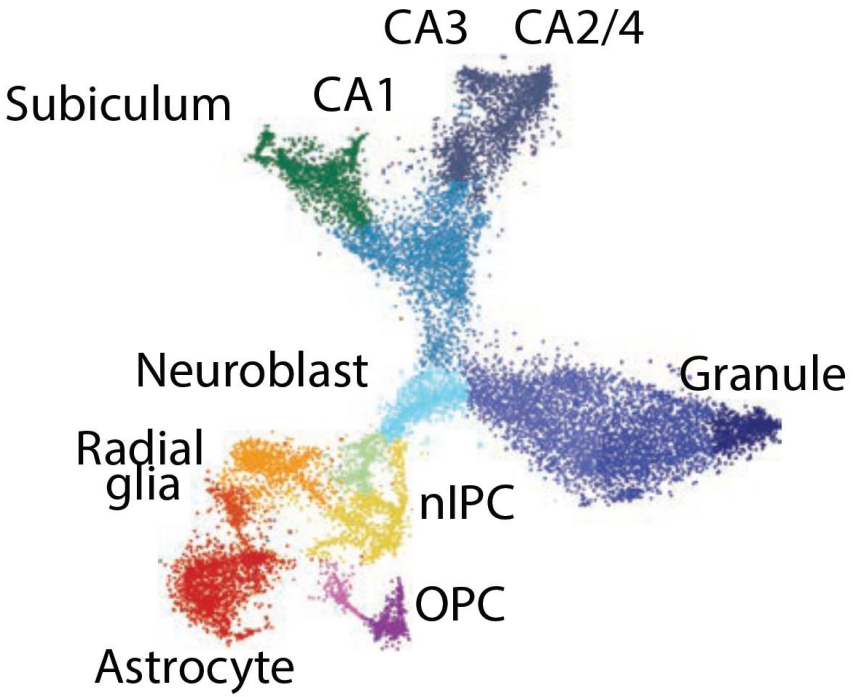
"The general challenge, even with perfect data, is that many regulatory mechanisms can generate the same dynamic process, and many dynamic processes can give rise to the same distribution."

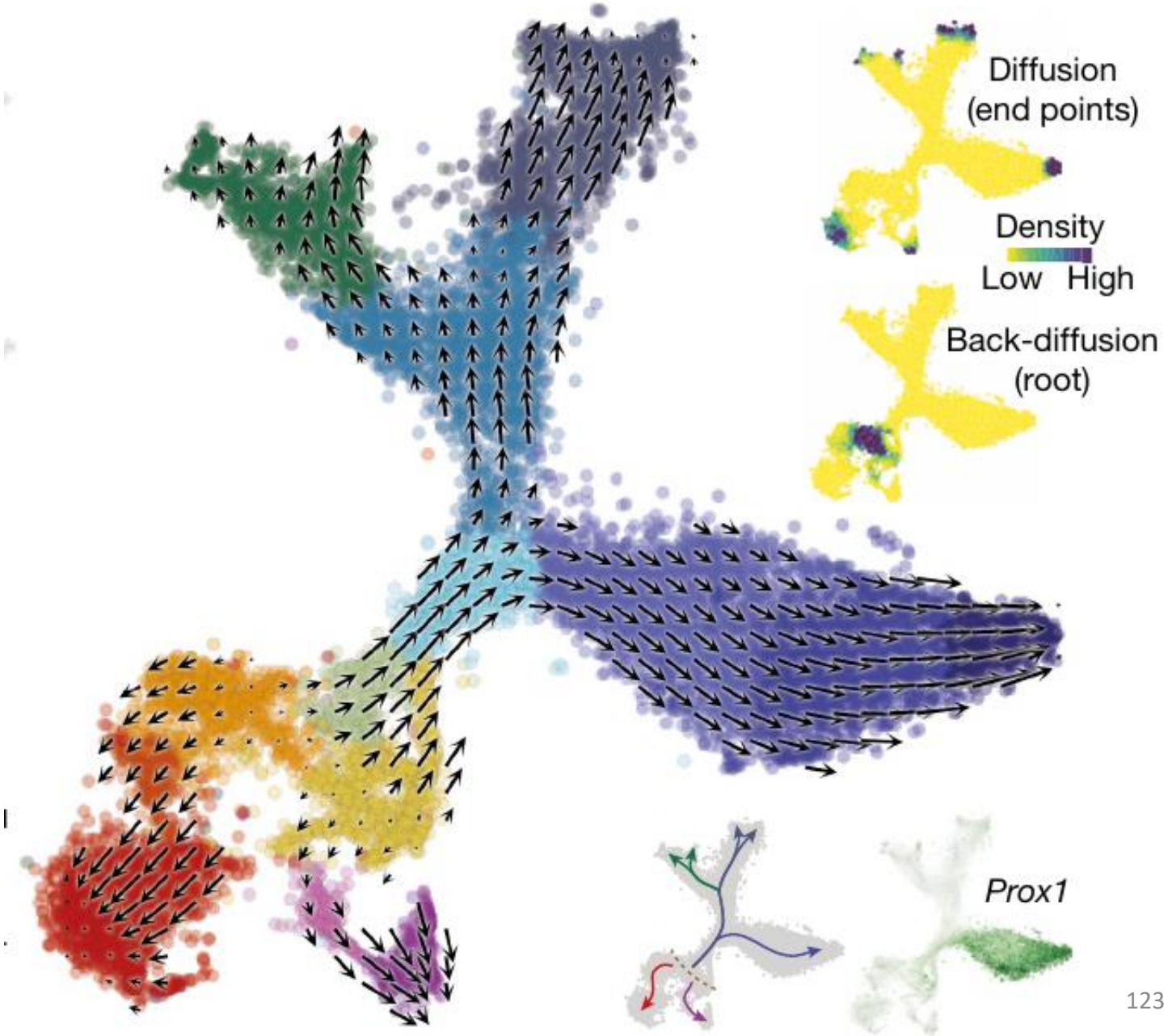
RNA Velocity



Le Manno et al. (Nature 2018)

RNA velocity





Le Manno et al. (Nature 2018)

Summary of today

- Overview of single-cell assays/platforms/protocols
- Quality control
- Normalization
- Data integration
- Feature selection
- Dimensionality reduction
- Cell type identification
- Trajectory inference

There is more to it…

- Constructing the cell x gene matrix
- Single cell regulatory networks
- Imputation
- Single cell multi-omics
- Sample multiplexing
- Single cell isoform sequencing
- Cell lineage + scRNA-seq
- Spatial transcriptomics
- …

Useful Resources

- Best practices in single cell RNA-seq analysis (Luecken & Theis, MSB 2019)

https://www.embopress.org/doi/pdf/10.15252/msb.20188746

- Orchestrating Single-Cell Analysis with Bioconductor

https://osca.bioconductor.org/

- Single Cell Course (Martin Hemberg Lab, 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

a.mahfouz@lumc.nl
https://www.lcbc.nl/
@ahmedElkoussy