@koenvdberge_Be

Single-cell RNA-seq analysis

Koen Van den Berge

Ghent University

koen.vandenberge@ugent.be

GHENT UNIVERSITY

This module

- In the single-cell RNA-sequencing module, the aim is to explain the applicability of these datasets and the general concepts behind their analysis

- Understanding the concept will be more beneficial in the long term than knowing which tool you should use right now

- I will therefore not spend time on providing a recommendation of current best tools for such an analysis, since the field is evolving very fast

- If you are interested in this, please check one of the last slides on workflows and current best practices

- **Please ask questions if something is unclear**

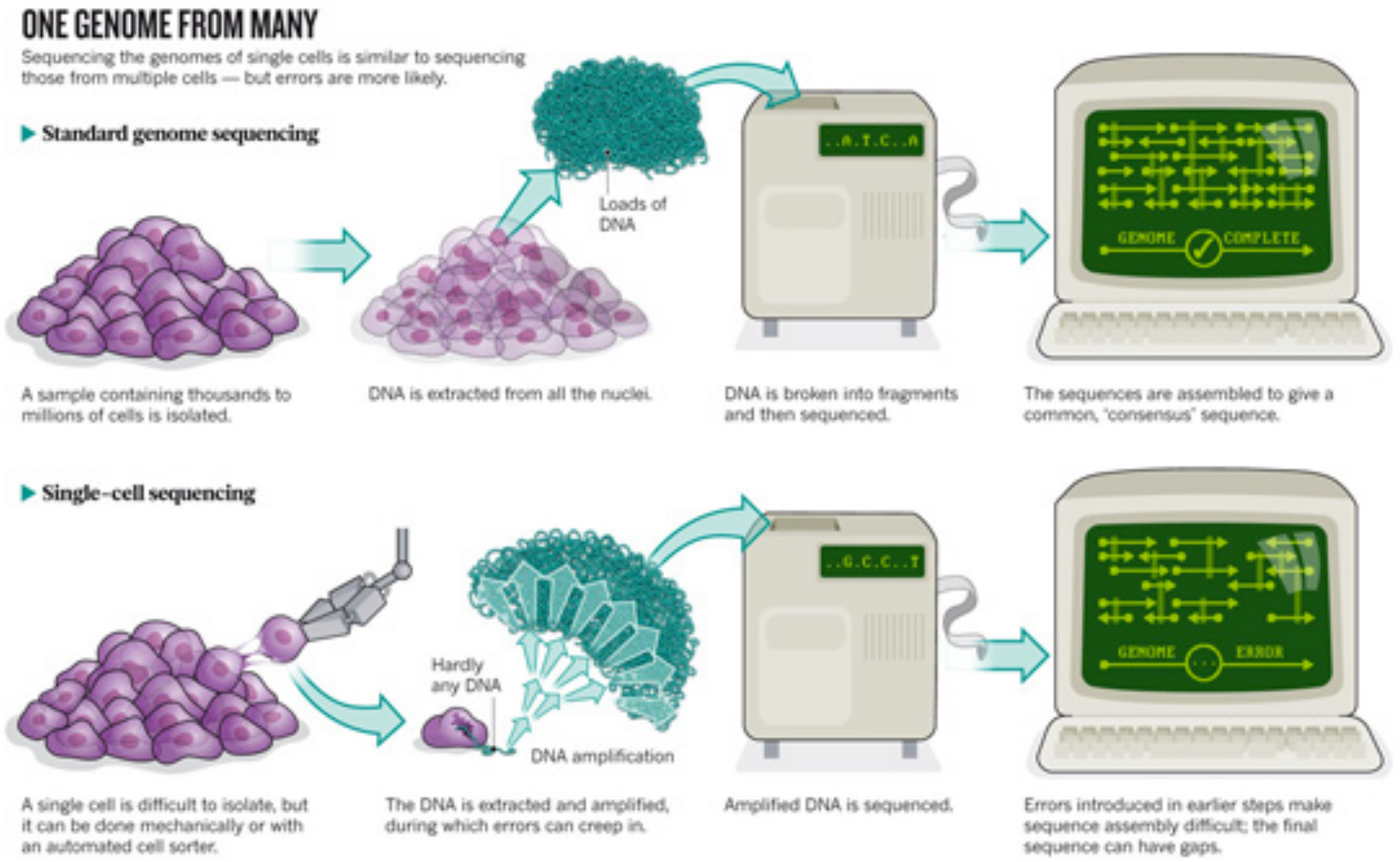
Table of contents

Table of contents

In single-cell RNA-sequencing (scRNA-seq), the RNA of a single cell is sequenced

Bulk:

Single-cell:



https://scitechdaily.com/images/one-genome-from-many.jpg

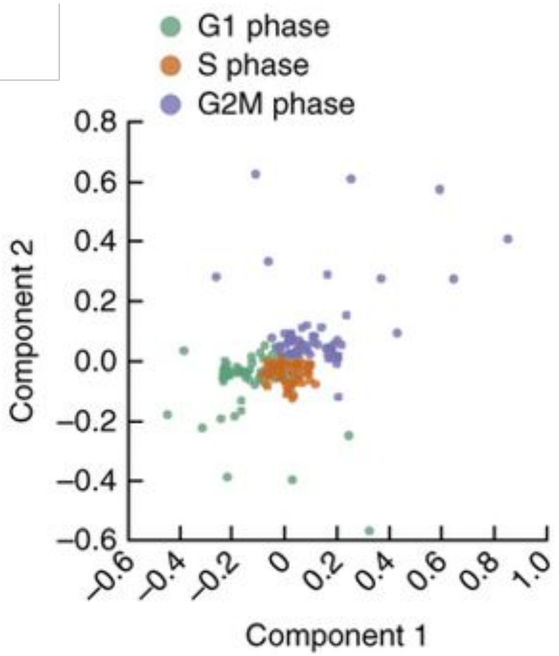
scRNA-seq allows disentanglement of complex biological systems

- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells

scRNA-seq allows disentanglement of complex biological systems

- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells

Dissecting cell states

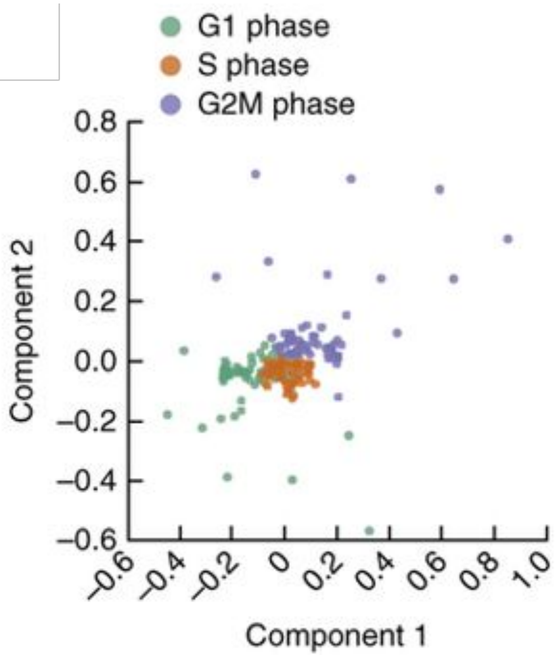


Buettner *et al.* (2015)

scRNA-seq allows disentanglement of complex biological systems

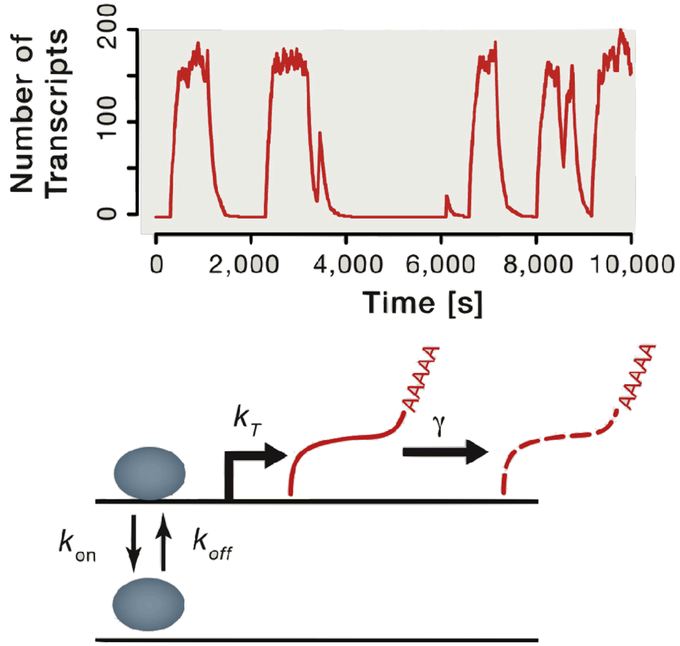
- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells

Dissecting cell states



Buettner *et al.* (2015)

Transcriptional regulation

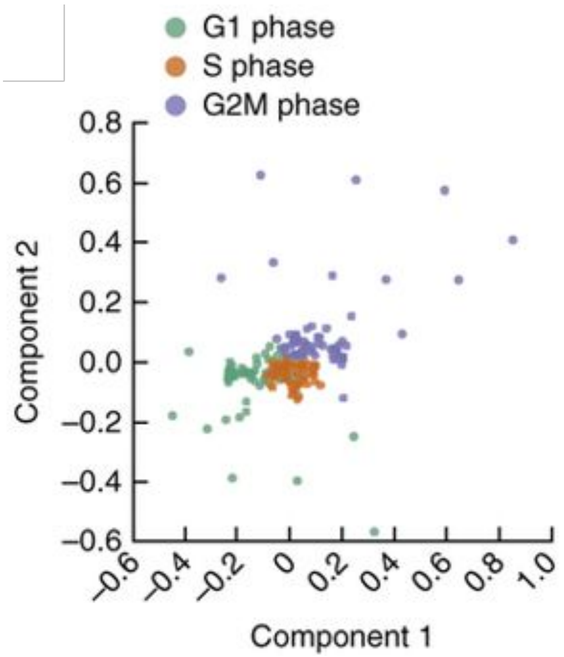


Grün *et al.* (2013)

scRNA-seq allows disentanglement of complex biological systems

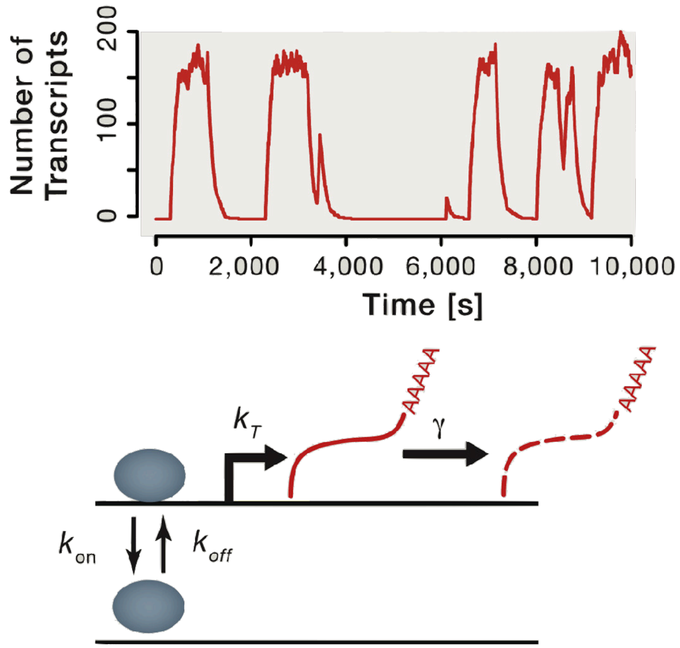
- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells

Dissecting cell states



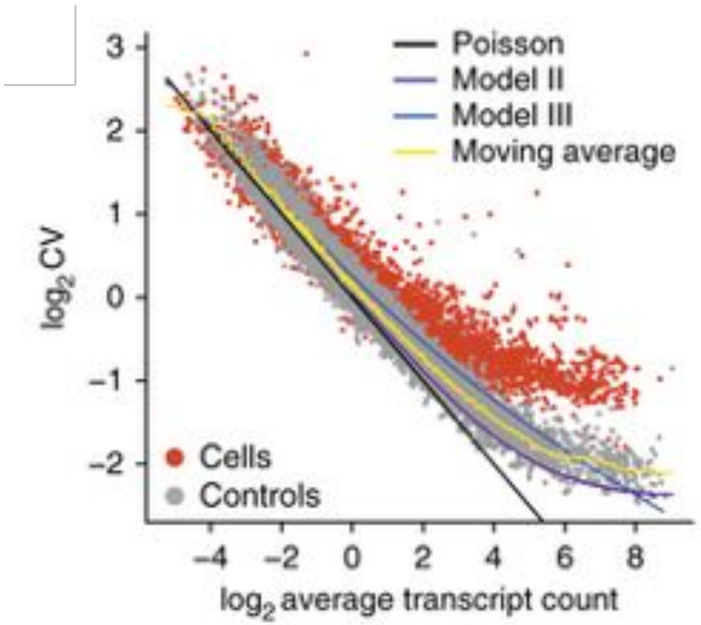
Buettner *et al.* (2015)

Transcriptional regulation



Grün *et al.* (2013)

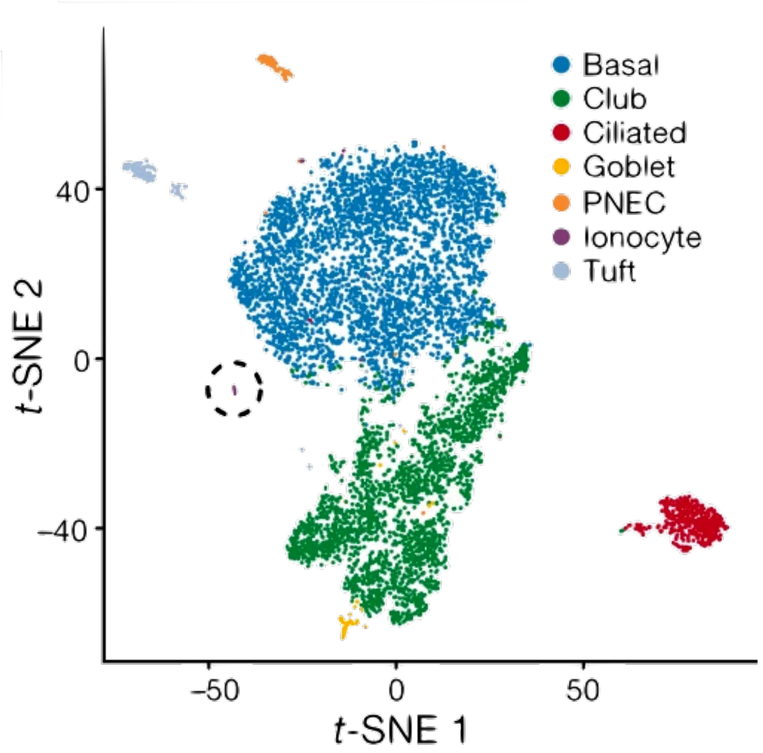
Assess technical variability



Grün *et al.* (2014)

scRNA-seq allows disentanglement of complex biological systems

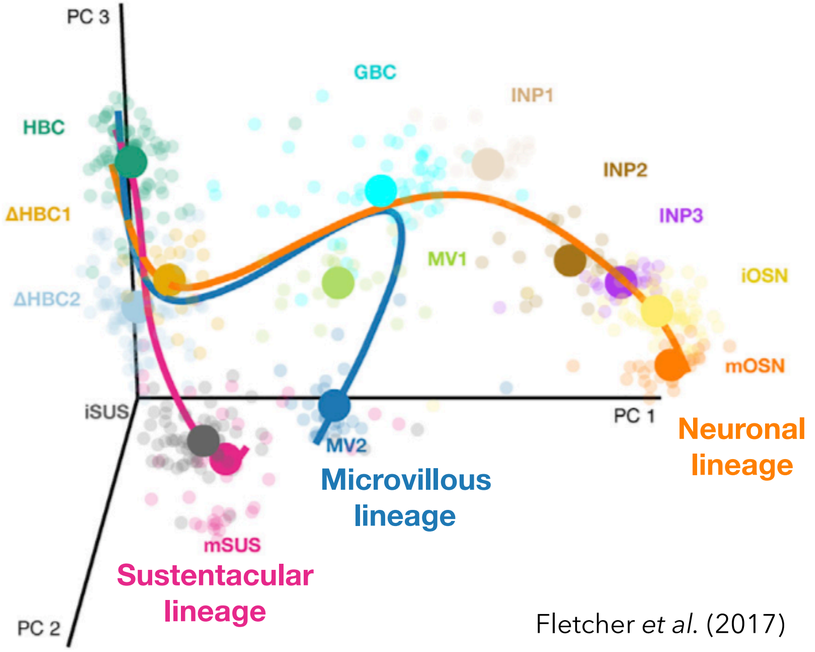
- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells
  - Identification of novel and rare cell types



Montoro *et al.* (2018)

scRNA-seq allows disentanglement of complex biological systems

- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells
  - Identification of novel and rare cell types
  - Reconstructing single-cell developmental/activational trajectories (e.g. development of stem cell to a mature cell type, activation of cells following treatment)



Fletcher *et al.* (2017)

scRNA-seq allows disentanglement of complex biological systems

- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells
  - Identification of novel and rare cell types
  - Reconstructing single-cell developmental/activational trajectories (e.g. development of stem cell to a mature cell type, activation of cells following treatment)
  - Studying sparsely occurring cell populations (e.g., stem cells, embryogenesis)

scRNA-seq allows disentanglement of complex biological systems

- Gene expression data on a single-cell level allows us to answer hypotheses of interest that were previously unavailable with bulk RNA-seq
  - Heterogeneity of gene expression between single cells
  - Identification of novel and rare cell types
  - Reconstructing single-cell developmental/activational trajectories (e.g. development of stem cell to a mature cell type, activation of cells following treatment)
  - Studying sparsely occurring cell populations (e.g., stem cells, embryogenesis)

- A quick note on terminology: cell identity represents the combined effect of cell type (permanent feature, e.g. neuron) and cell state (transient feature, e.g. cell cycle stage)

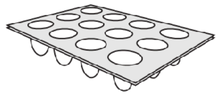
Table of contents

scRNA-seq remains a fast-paced field with continuous active developments

Manual

Tang *et al* 2009

Multiplexing

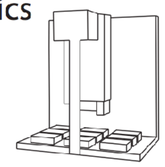


Islam *et al* 2011

Integrated Fluidic Circuits

Brennecke *et al* 2013

Liquid Handling Robotics



Jaitin *et al* 2014

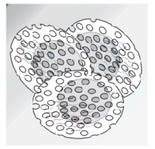
Nanodroplets

Klein *et al* 2015
Macosko *et al* 2015

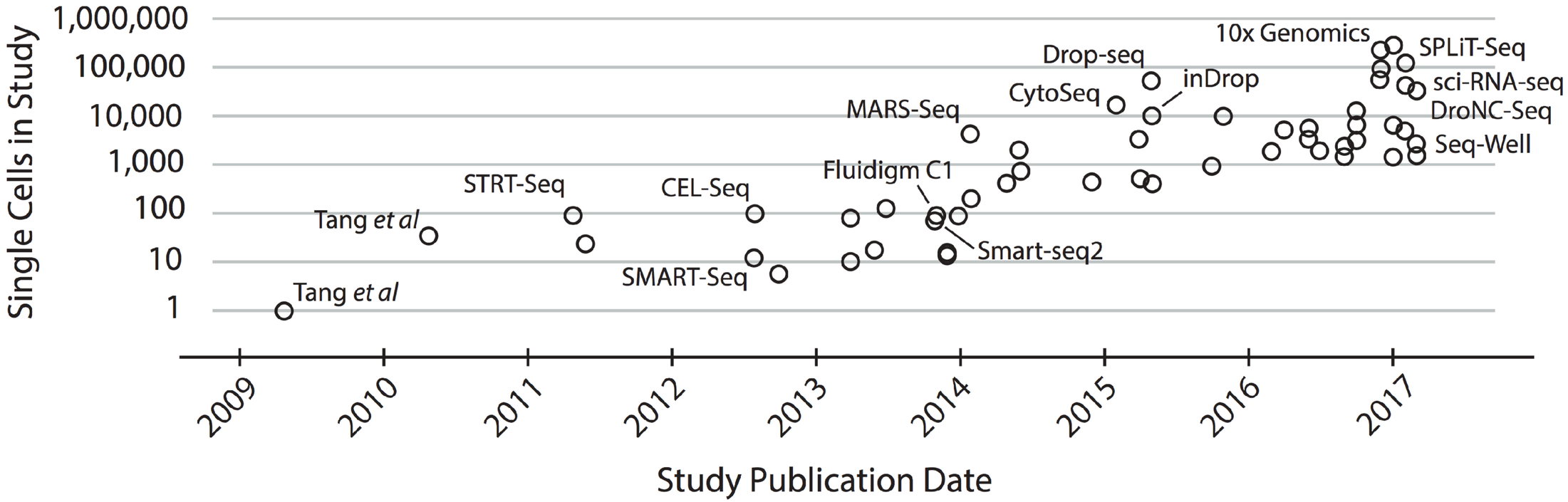
Picowells

Bose *et al* 2015

In situ barcoding

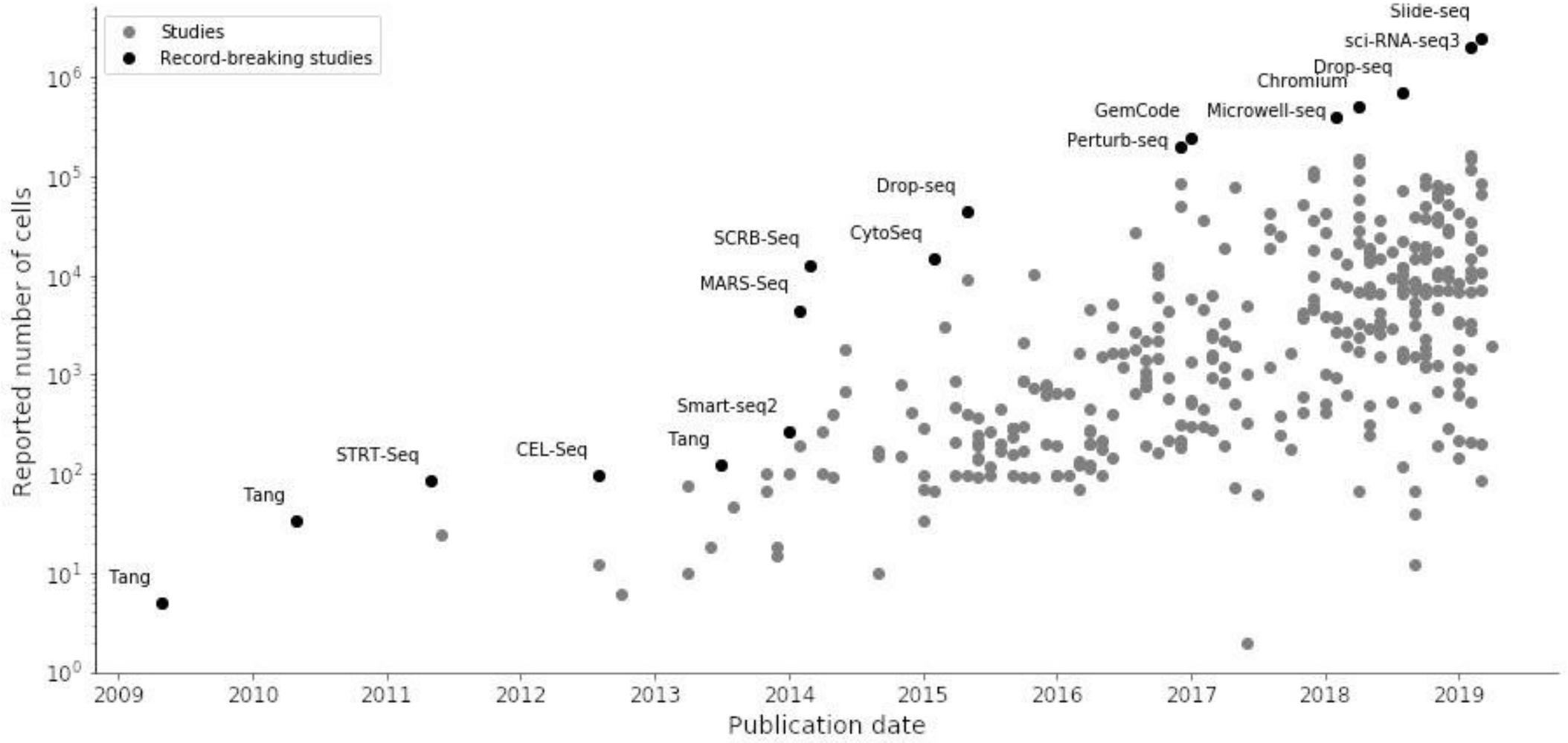


Cao *et al* 2017
Rosenberg *et al* 2017



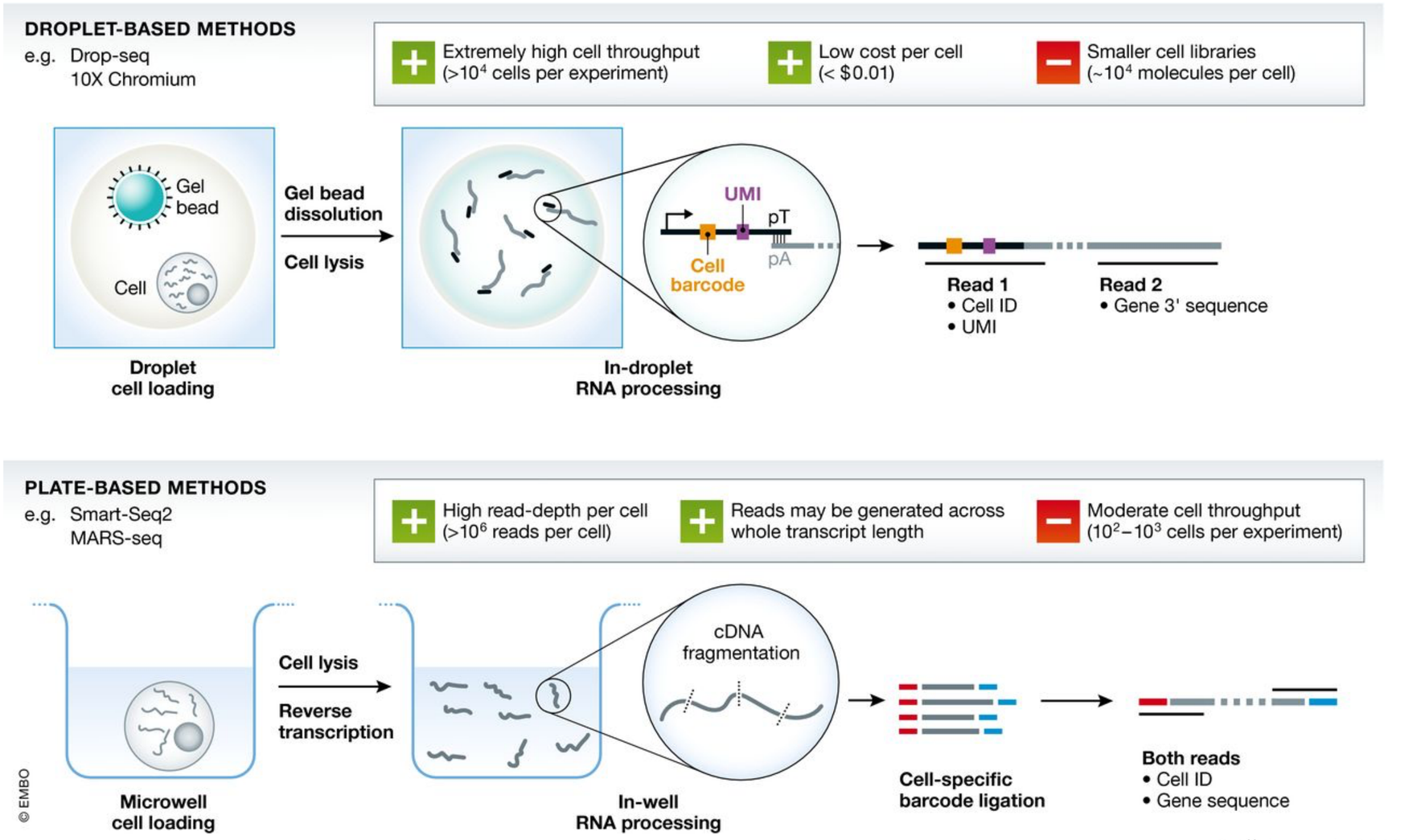
Svensson *et al.* (2018)

scRNA-seq remains a fast-paced field with continuous active developments



https://twitter.com/vallens/status/1113982015517282304

scRNA-seq protocols may (roughly) be classified in plate-based and droplet-based



Griffiths *et al.* (2018)

Cool video demonstrating the Drop-seq protocol: https://vimeo.com/128484564

scRNA-seq protocols may (roughly) be classified in plate-based and droplet-based

- Deciding which protocol you should use essentially depends on the question you wish to answer

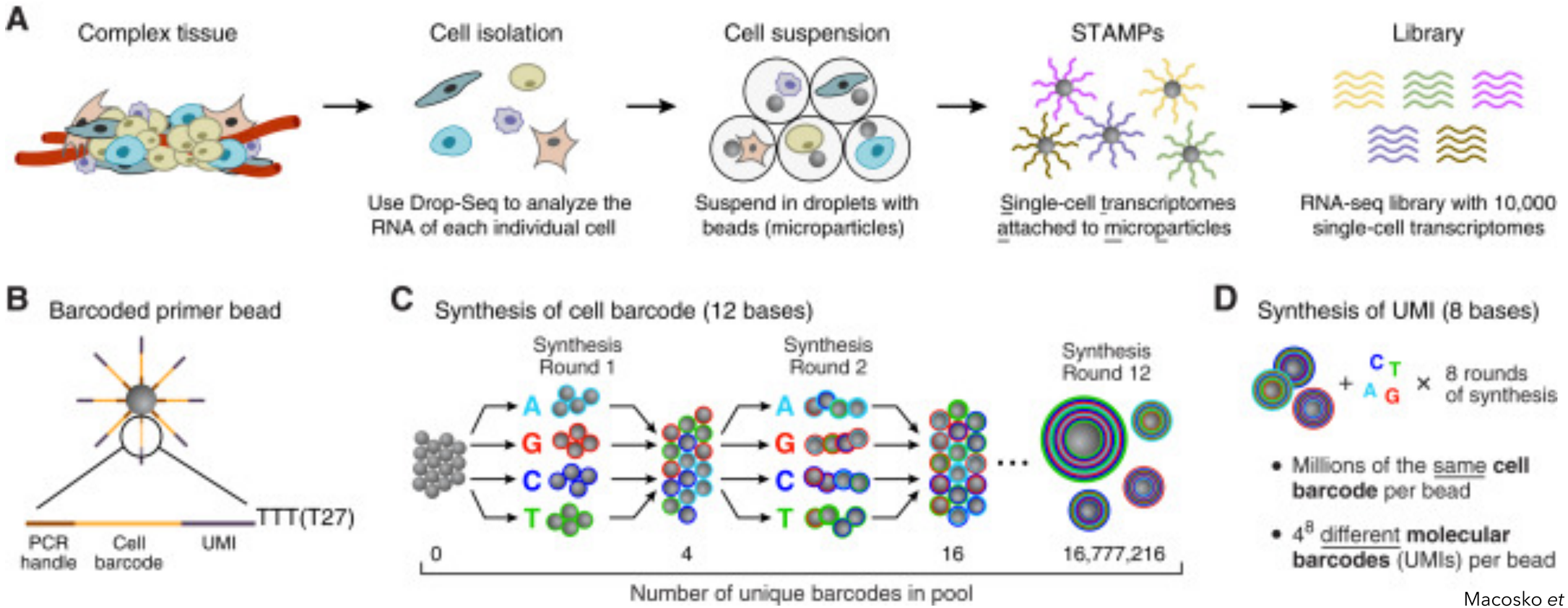
- Droplet-based protocols are more suited for
  - Examining the composition of a tissue
  - Identifying novel / rare cell types

- Plate-based protocols are more suited for
  - Studying a rare cell population with known surface markers (through FACS sorting)
  - Isoform-level analysis (full-length transcript information)
  - Marker gene discovery?

- In general, while droplet-based protocols allow for a higher throughput, plate-based protocols seem to have a higher signal-to-noise ratio per cell

Expression quantification differs between protocols

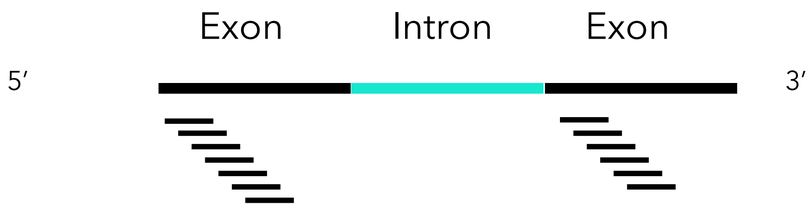
- Plate-based protocols adopt read counting (like in bulk RNA-seq), while droplet-based protocols typically adopt unique molecular identifiers (UMIs) to quantify gene expression



Macosko *et al.* (2015)

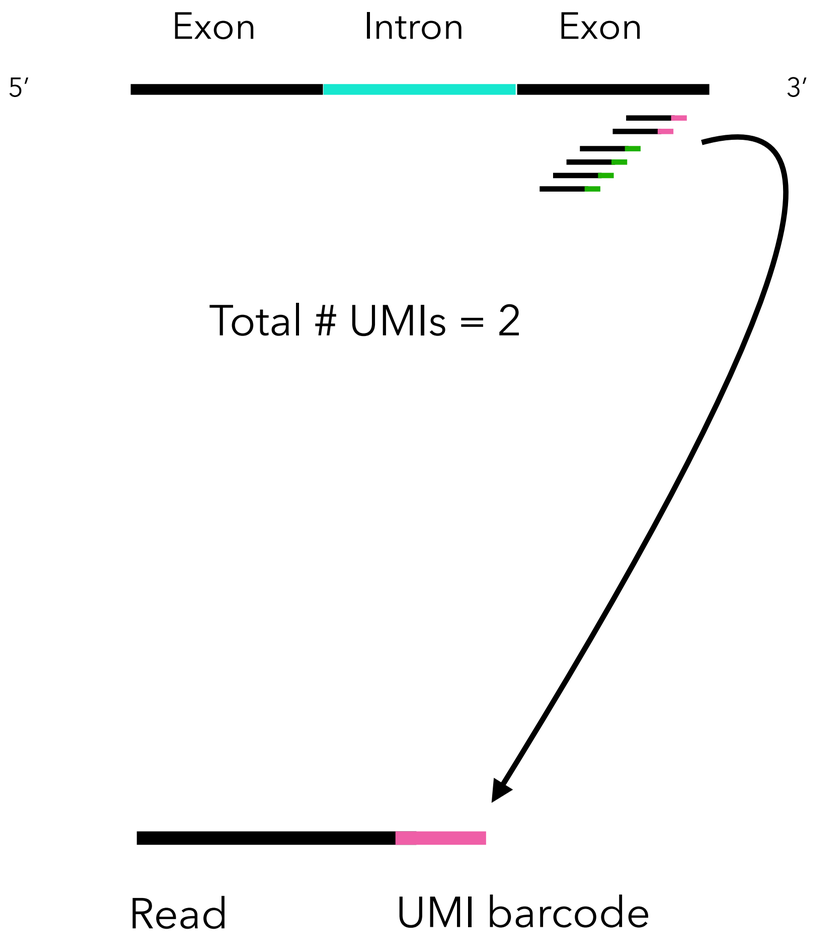
Expression quantification differs between protocols

READ COUNTING



Total # reads = 13

UMI COUNTING



Total # UMIs = 2

Expression quantification differs between protocols

- Read counts are affected by e.g. gene length, sequencing depth and PCR amplification bias
- UMIs were introduced to avoid this, however this is only true if every cell is sequenced to saturation, see Vallejos *et al*. (2017)
- Even if UMIs are used, between-cell normalization is still crucial to obtain reliable results
- Due to the counting strategy, UMI counts can be interpreted as a proxy for the number of transcripts originally present in the cell (<-> read counting)

Table of contents

Expression quantification differs between protocols

- Due to different counting strategies, count matrices are very different between protocols, and thus the data analysis strategies may also vary
- Example: plate-based SMART-Seq dataset

```
> counts
                cell1 cell2 cell3 cell4 cell5
Atp6v1h           432    94     0     0     0
Oprk1               0     0     0     0     0
4732440D04Rik       2     2     1    13     0
Rb1cc1            235     0     0    41    14
St18                0     0     0     0     0
```

- High positive counts mixed with zeros
- Zero abundance roughly ~50-75%
- Typically lower number of cells

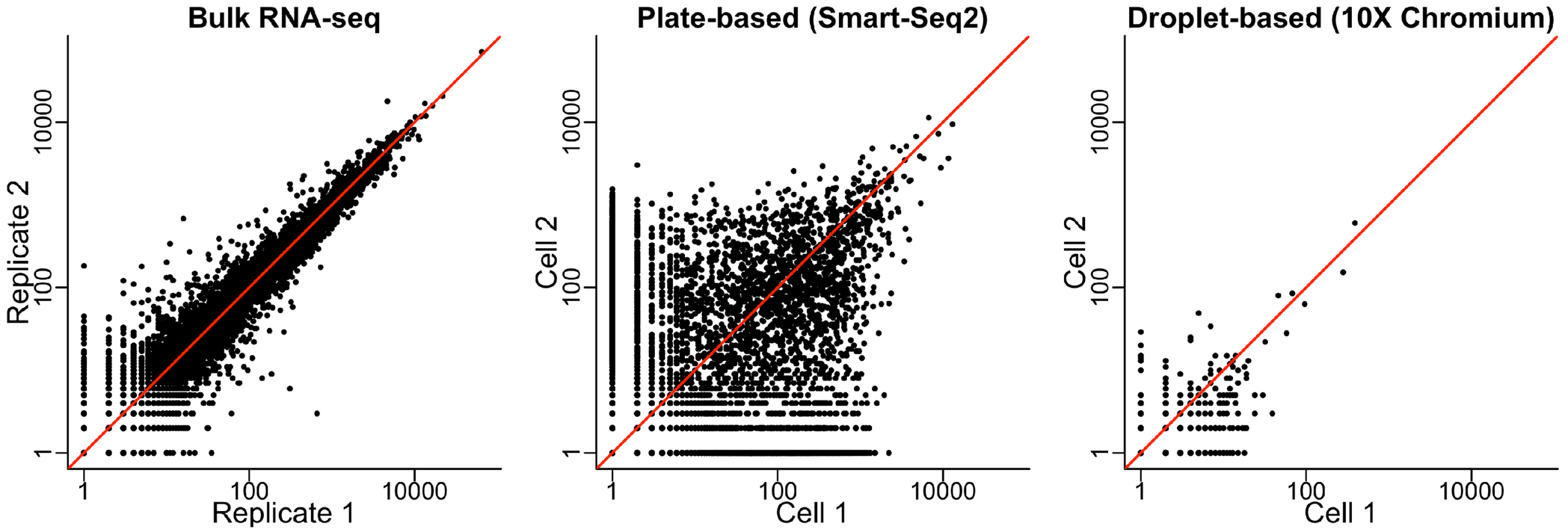
- Example: droplet-based UMI dataset

```
> as.matrix(exprs(cds)[1:5,1:5])
                W31105 W31106 W31107 W31108 W31109
0610007L01Rik       0      0      1      1      0
0610009O20Rik       0      0      0      0      3
0910001L09Rik       0      1      1      1      1
1100001G20Rik       0      0      0      0      0
1110004E09Rik       0      1      2      1      2
```

- Low positive counts mixed with even more zeros
- Zero abundance typically >90%
- Typically higher number of cells

scRNA-seq data is sparse, more variable than bulk RNA-seq data

- As in bulk RNA-seq, a scRNA-seq dataset is often summarized in a count matrix, where rows represent genes and columns represent cells
- In scRNA-seq, the number of samples (cells) are generally much higher as compared to bulk RNA-seq; plate-based protocols easily gather hundreds of cells, while droplet-based protocols produce datasets of several thousands of cells

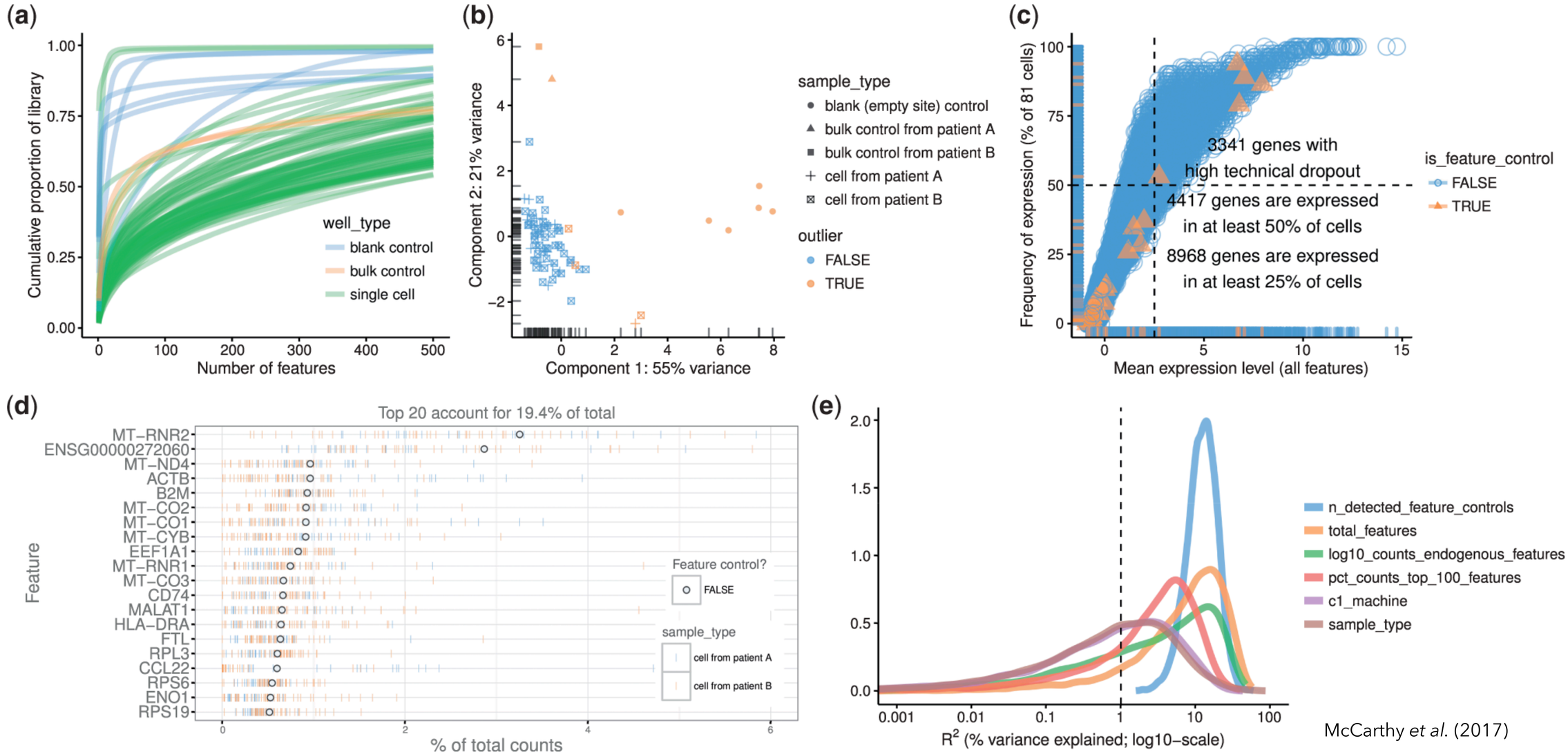


scRNA-seq requires thorough quality control (QC)

- Identification and removal of
  - low-quality (e.g. dead/damaged) cells
  - doublets (droplets/wells containing 2+ cells)
  - empty droplets (droplets/wells without any cells)
- Identification of these cells typically occurs in a data-driven way
- In plate-based protocols, checking for these may also occur through microscopic observation (but, time-intensive!)
- Example 1: doublet detection by combining single cells and searching for nearest neighbours (DoubletFinder method, McGinnnis *et al*. (2019))
- Example 2: empty droplet detection by testing for a significant deviation from an ambient solution (EmptyDrops method, Lun *et al*. (2019))

scRNA-seq requires thorough quality control (QC)

Diagnostic plots from R/Bioconductor package scater: https://bioconductor.org/packages/release/bioc/html/scater.html. For more examples, check the conquer website: http://imlspenticton.uzh.ch:3838/conquer/



McCarthy *et al.* (2017)

Table of contents

The data analysis workflow in scRNA-seq is different from bulk RNA-seq

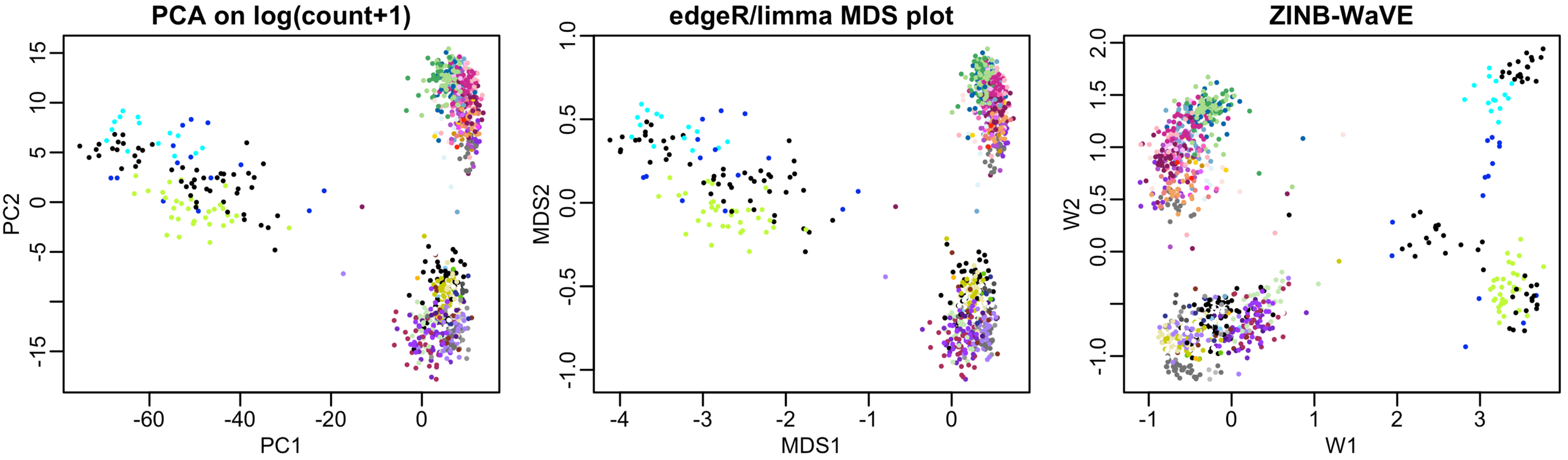
- In bulk RNA-seq, we typically know which groups we want to compare (e.g., treatment vs. control)
- In single-cell RNA-seq, we are often interested in comparing gene expression between different cell types
- However, we first must identify the cell types in order to be able to compare them
- Therefore, in scRNA-seq, the differential expression analysis is usually preceded by identification of cell identity, typically through clustering in reduced dimensionality
- Note, that the definition of cell identity can be vague, and may include both cell type (e.g., leukocyte vs. erythrocyte) and cell state (e.g., cell cycle phase)

Dimensionality reduction methods for scRNA-seq data

- The goal of dimensionality reduction (DR) is to reduce our G x C matrix to a Q x C matrix, where Q << G, while retaining as much signal in the data as possible
- This may serve multiple purposes, such as visualization, identification of batch effects and clustering in reduced dimensionality
- Traditional DR methods are insufficient, e.g. PCA is inappropriate for count data (Townes *et al.* 2019)
- **Many** dimensionality reduction methods are being used in scRNA-seq
- Most popular ones are non-linear DR methods, e.g. t-SNE and UMAP

Dimensionality reduction methods for scRNA-seq data

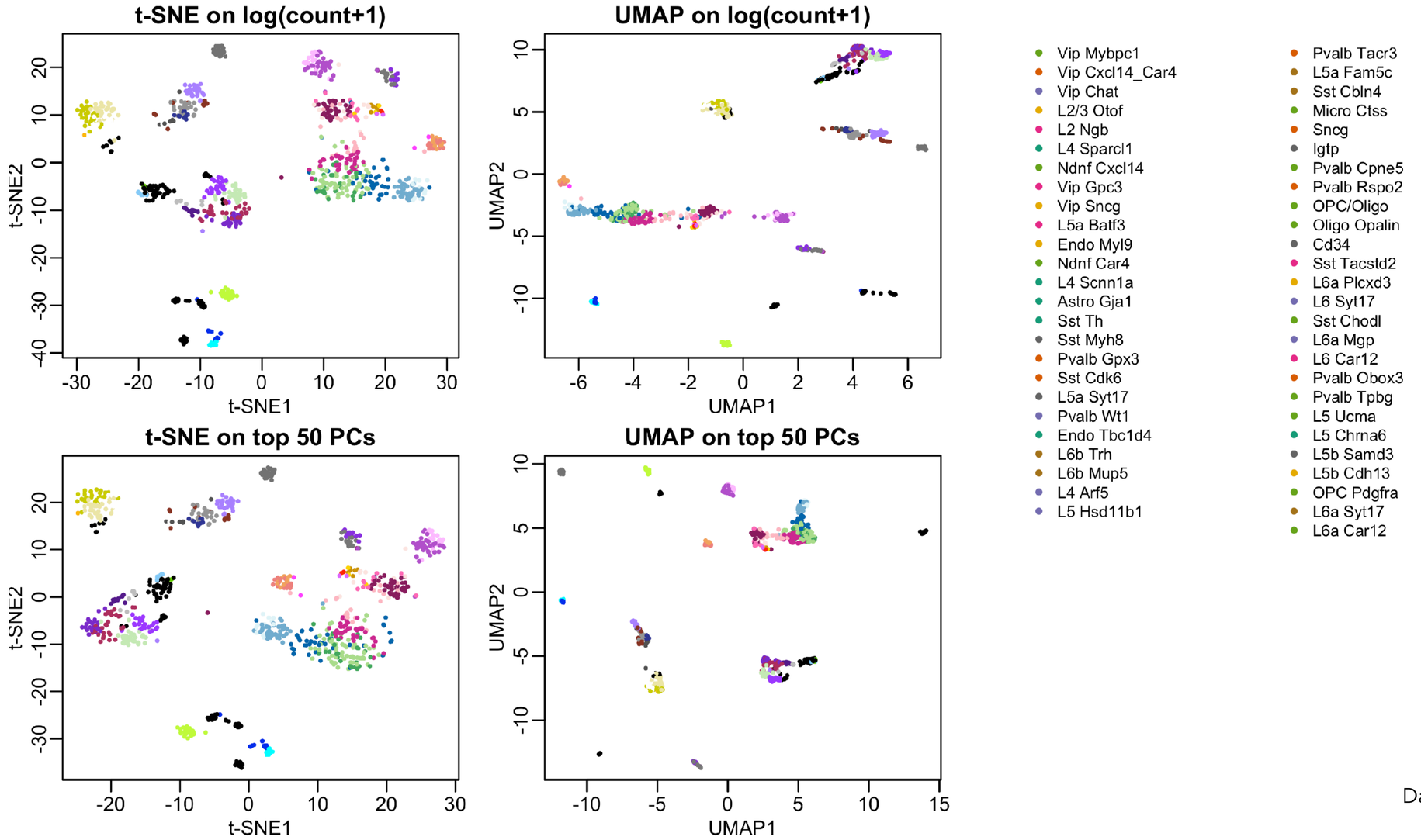
Demonstration of linear dimensionality reduction methods, cells are colored by cell-type.



Data from Tasic *et al.* (2016)

Dimensionality reduction methods for scRNA-seq data

Demonstration of non-linear dimensionality reduction methods, cells are colored by cell-type.



Data from Tasic *et al.* (2016)

Clustering methods for scRNA-seq data

- If we assume that a different cell identity is reflected by a different gene expression profile, we can cluster cells to identify cell types
- In clustering, the goal is to group cells together that have similar expression profiles
- Clustering typically occurs in reduced dimension or based on a subset of interesting genes (curse of dimensionality)
- Example: k-means clustering:
  - START: set the number of clusters k, randomly choose k cells to be cluster centroids
  - 1. Find closest centroid for each cell
  - 2. Group cells together that share the closest centroid
  - 3. Update centroid based on current group of cells
  - Repeat 1-3 until convergence
  - Visualization: http://shabal.in/visuals/kmeans/1.html

Clustering methods for scRNA-seq data

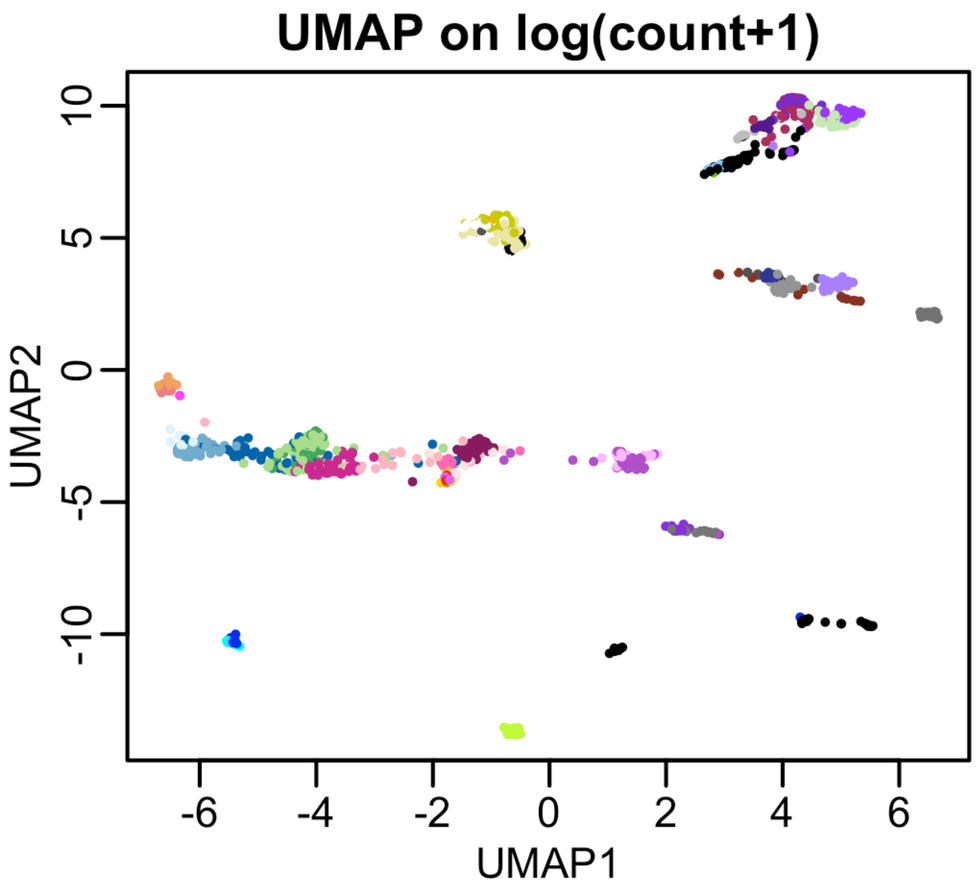
- k-means has some drawbacks (requires choice of k, results may vary over several iterations due to randomness of starting point)
- Therefore, alternative methods have been introduced, which include
  - Graph-based methods: cluster cells that are connected together (e.g., using nearest neighbours), e.g. Seurat
  - Consensus clustering: cluster cells that are often clustered together over several clustering algorithms (some cells will be unclustered), e.g. RSEC (Risso *et al.* (2018))
  - Iterative clustering: Recluster initially derived clusters, e.g. Tasic *et al.* (2016)

Why don’t we cluster the full gene expression matrix directly?

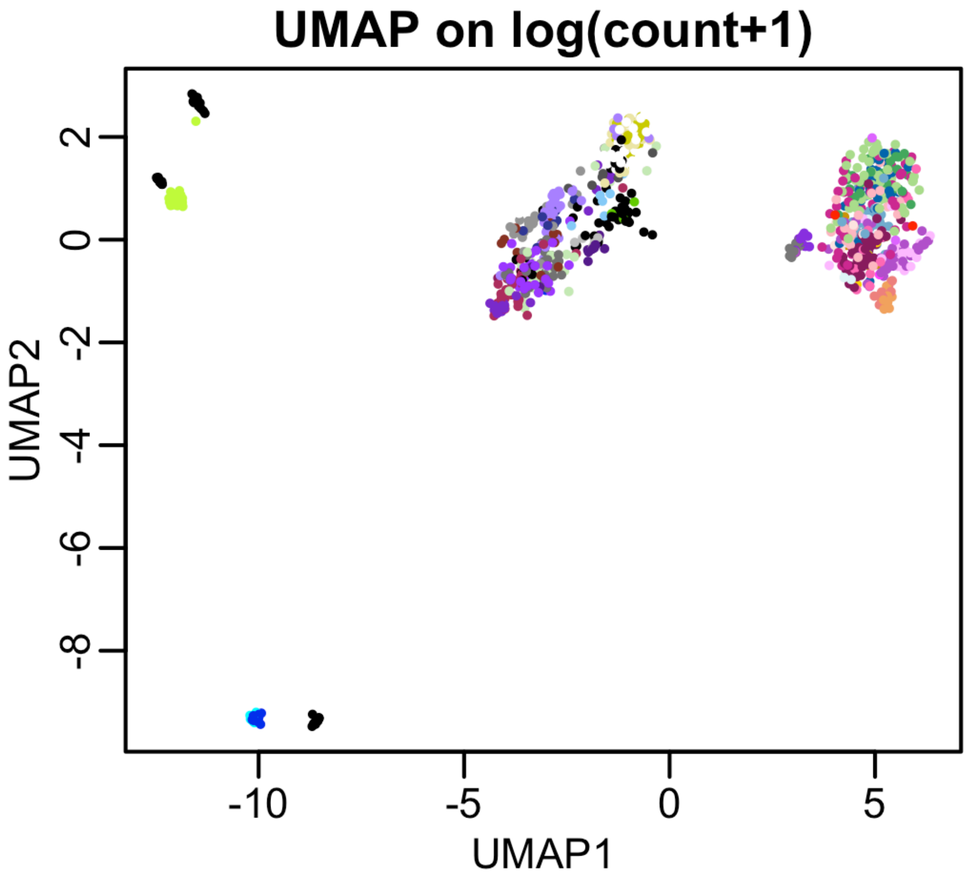
- Why do we need dimensionality reduction or feature selection of interesting genes before clustering?
- Curse of dimensionality: hard to extract the underlying signal if the number of variables (genes) is much larger than the number of samples (cells)

Why don’t we cluster the full gene expression matrix directly?

Based on 500 most variable genes



Based on 10K most variable genes



Data from Tasic *et al.* (2016)

Table of contents

Differential expression analysis in scRNA-seq data

- In differential expression analysis, the aim is to discover marker genes that differentiate cell types or biological groups
- The statistical models used in scRNA-seq typically build on the GLM framework (see 'crash course on GLMs' in bulk RNA-seq slides)
- Count-based bulk RNA-seq DE methods (e.g., edgeR, DESeq2) can be directly leveraged to scRNA-seq data from droplet-based protocols!
- However, scRNA-seq data from plate-based protocols may suffer from zero inflation, and accounting for this can improve performances (Van den Berge, Perraudeau *et al.* (2018))

Count distributions and zero inflation

- Reminder: bulk RNA-seq DE methods assume a negative binomial (NB) model on gene expression Y_{gi} for gene g in sample i

$$\begin{cases} Y_{gi} & \sim & NB(\mu_{gi}, \phi_g) \\ \log(\mu_{gi}) & = & \eta_{gi} \\ \eta_{gi} & = & \mathbf{X}_i \beta_g + log(N_i), \end{cases}$$

where $E(Y_{gi}) = \mu_{gi}$, and $\text{Var}(Y_{gi}) = \mu_{gi} + \phi_g \mu_{gi}^2$.

Model definition:

1. We assume gene expression follows a NB distribution
2. We model the mean of the NB using a log-link
3. The linear predictor is modeled according to covariates X_i (e.g. control/treatment) and offset N_i

Love *et al. Genome Biology* (2014) 15:550
DOI 10.1186/s13059-014-0550-8

Genome Biology

BIOINFORMATICS APPLICATIONS NOTE Vol. 26 no. 1 2010, pages 139–140
doi:10.1093/bioinformatics/btp616

Gene expression

edgeR: a Bioconductor package for differential expression analysis of digital gene expression data

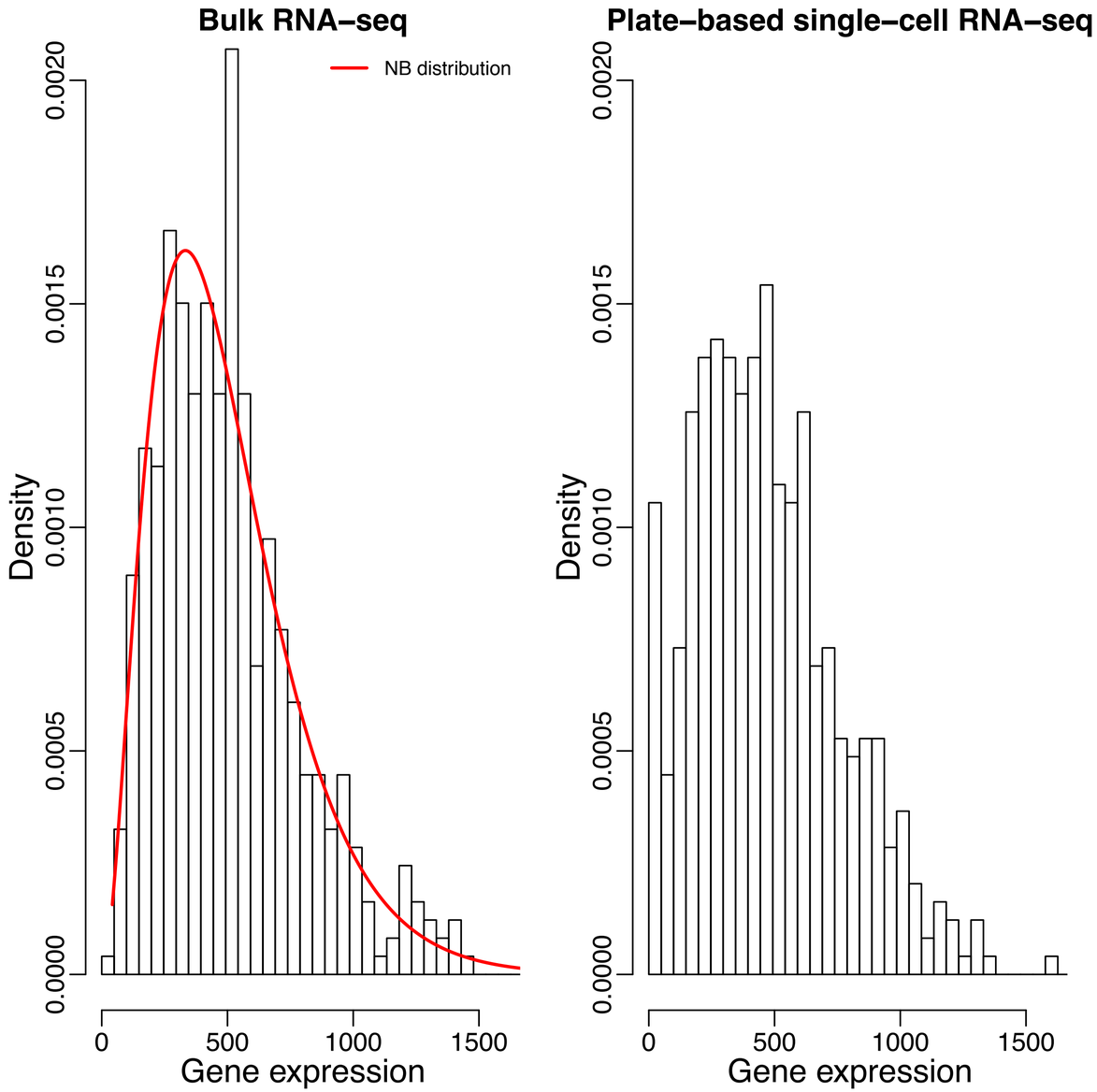
Mark D. Robinson[1,2,*,†], Davis J. McCarthy[2,†] and Gordon K. Smyth[2]

METHOD Open Access

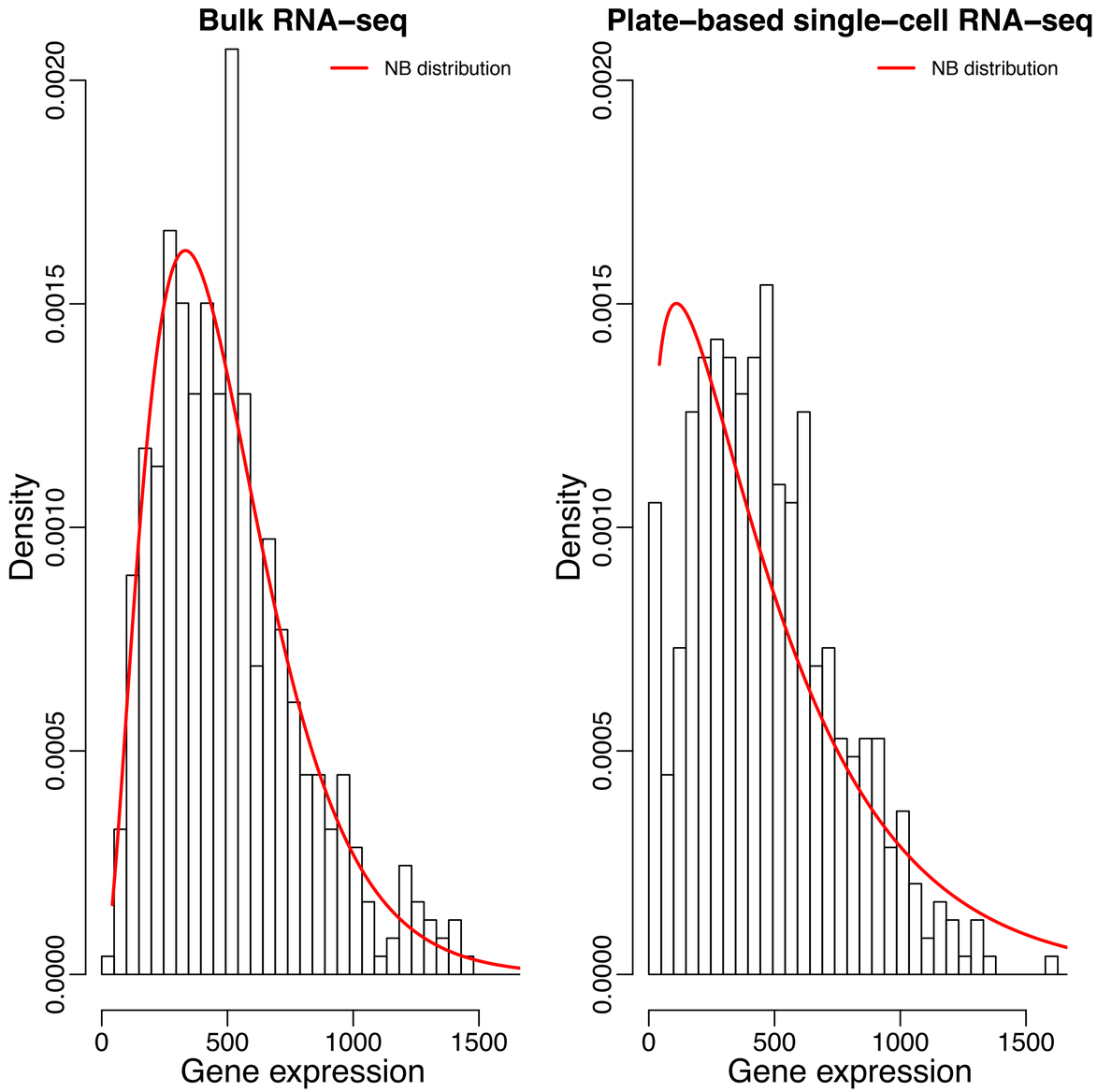
Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

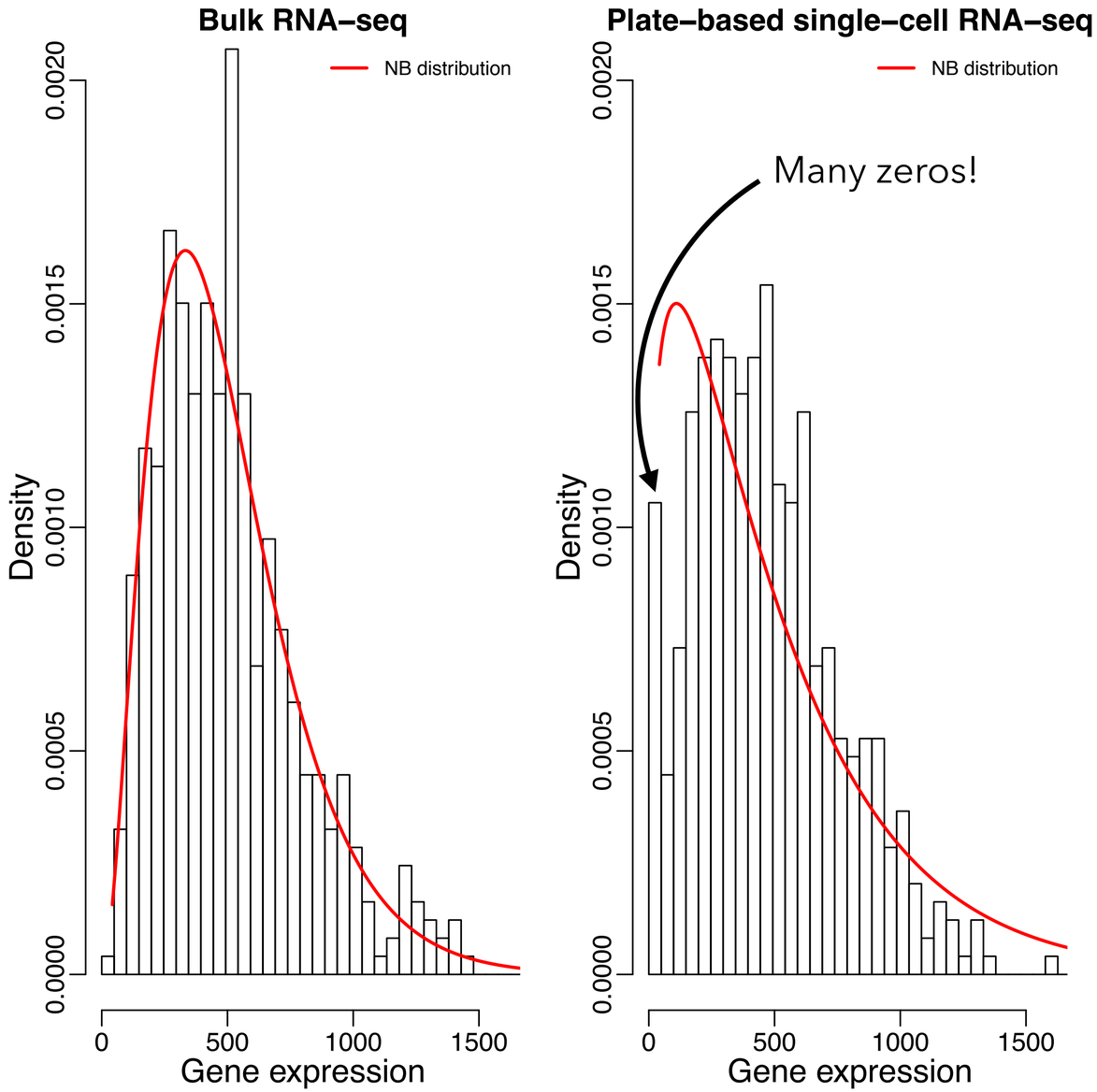
Count distributions and zero inflation



Count distributions and zero inflation



Count distributions and zero inflation



Count distributions and zero inflation

The zero-inflated negative binomial distribution (ZINB) is a two-component mixture distribution

$$Y_{gi} \sim \pi_{gi}\delta + (1 - \pi_{gi})f_{NB}(\mu_{gi}, \phi_g),$$

consisting of

- a point mass at zero to account for zero inflation, $\pi_{gi}\delta$
- A count component to model gene expression counts, $(1 - \pi_{gi})f_{NB}(\mu_{gi}, \phi_g)$

Count distributions and zero inflation

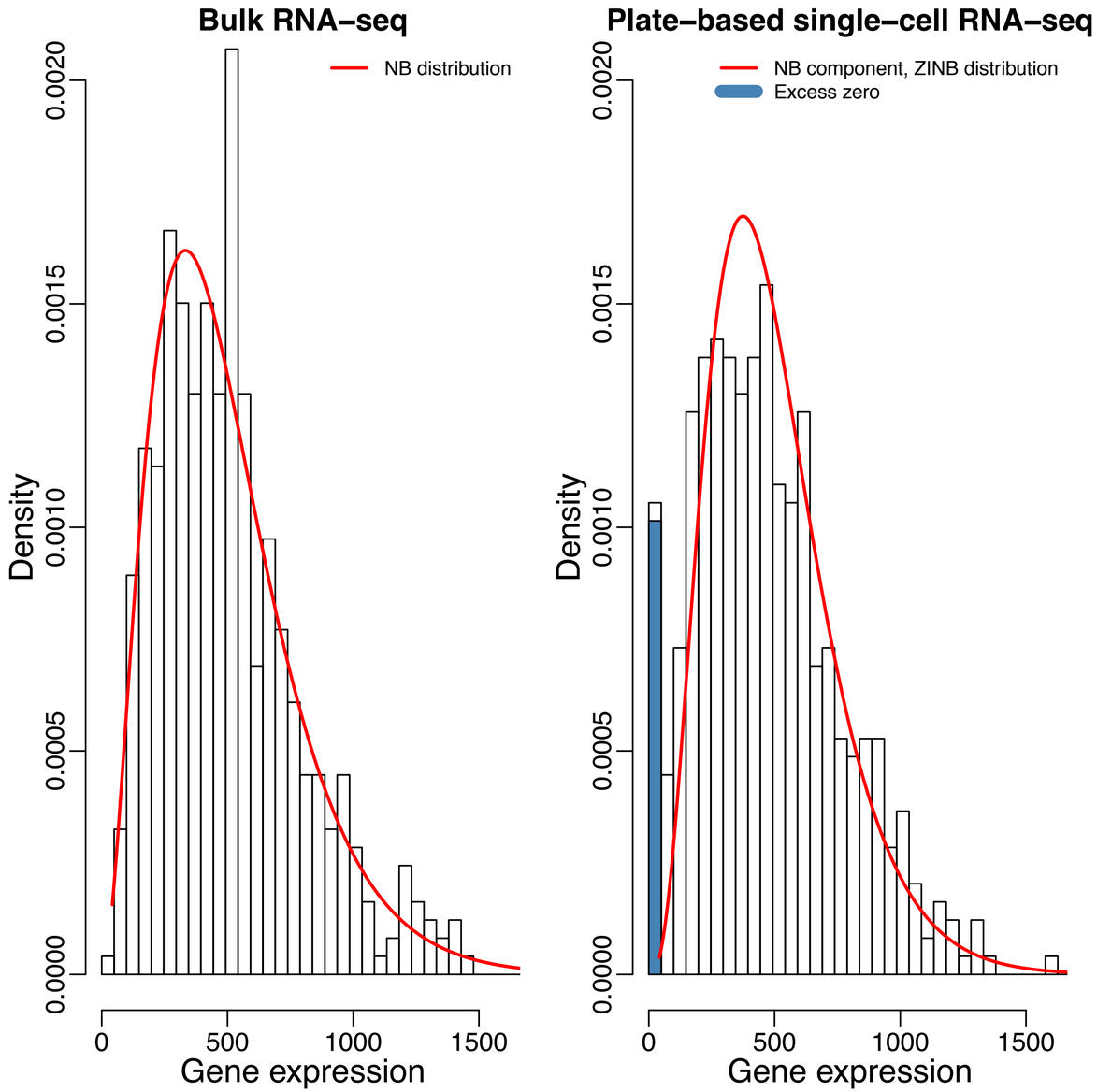
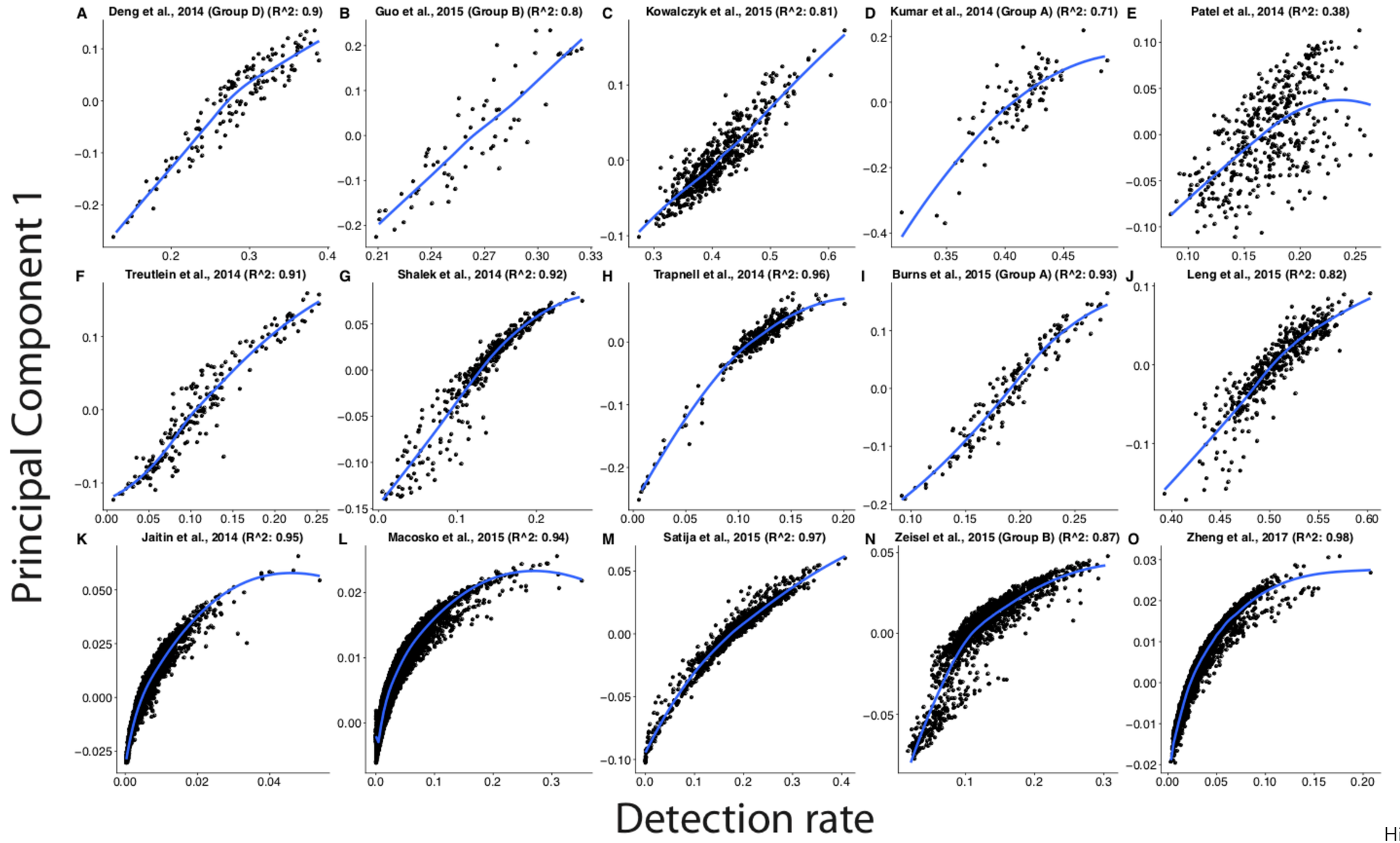


Table of contents

Batch effects

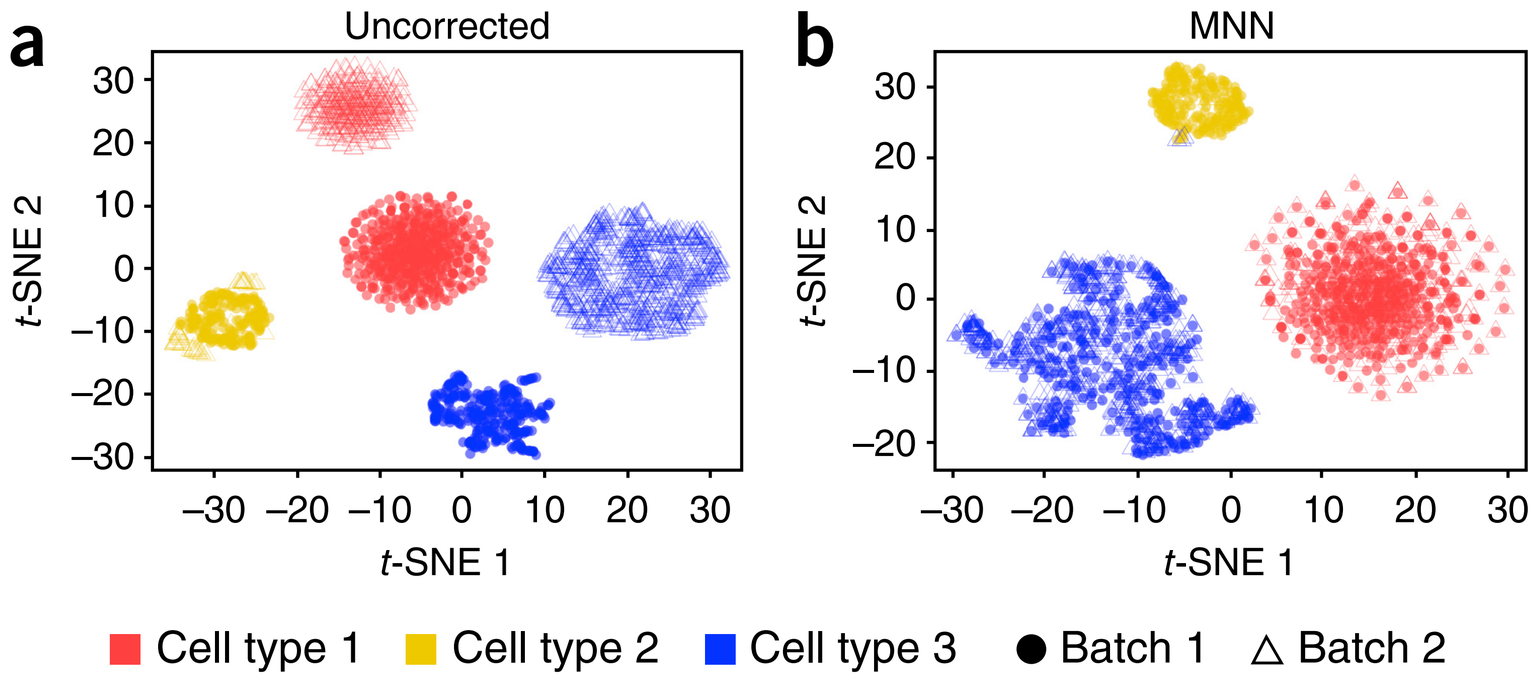
- Batch effects are systematic technical variation in the dataset that are not of interest
- This may represent known sources of variation, e.g. plate effects, different sequencing runs
- Sometimes, batch effects may be unobserved, and hence they must be estimated from the data, see ZINB-WaVE (Risso *et al.* (2018b))
- In the Experimental Design session, we looked into incorporating the batch effect as a covariate in the mean model
- Also in scRNA-seq, care **must** be taken to avoid confounding, e.g. do not separate control and treatment cells on two different plates for plate-based scRNA-seq

Nuisance effects may influence dimensionality reduction



Hicks *et al.* (2018)

Removing batch effects aids identification of biological cell types



Haghverdi *et al.* (2018)

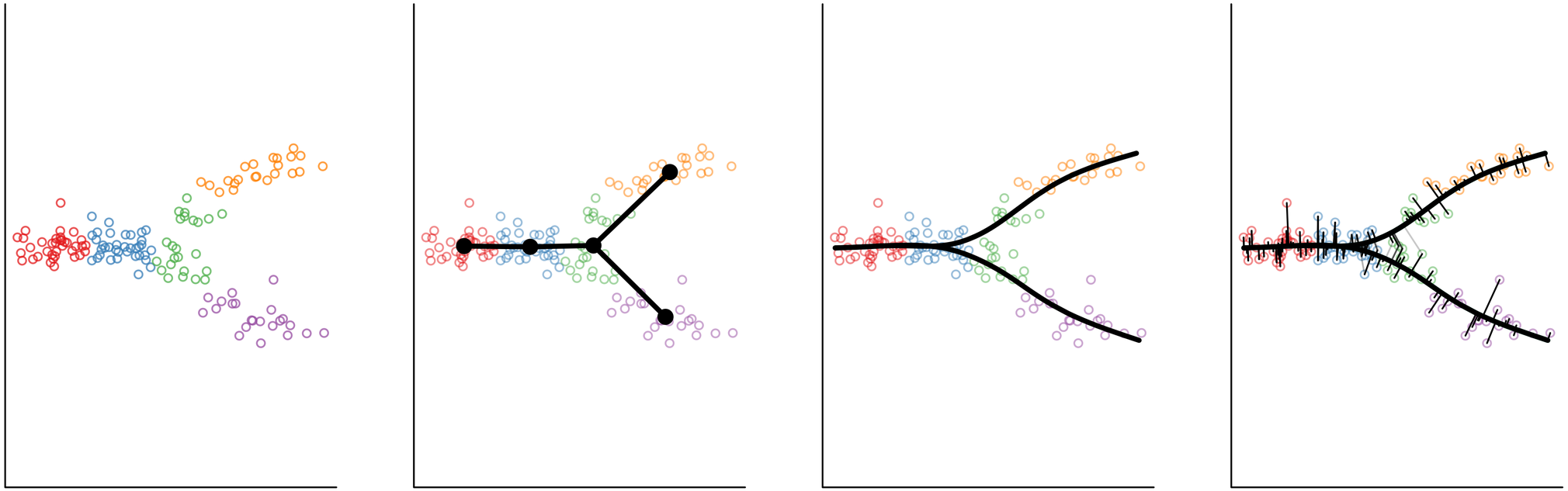
Post-selection inference

- There is a caveat in first identifying cell types and then performing differential expression analysis on the same data: we expect an increased false positive rate, if there is no true biological signal between the compared groups
- Intuitively: **we are using the same data twice**
- A quick taste of what happens: https://gist.github.com/koenvandenberge/c07d56c7c69e1c927291027329c7f34e

Table of contents

scRNA-seq allows the study of dynamic biological systems

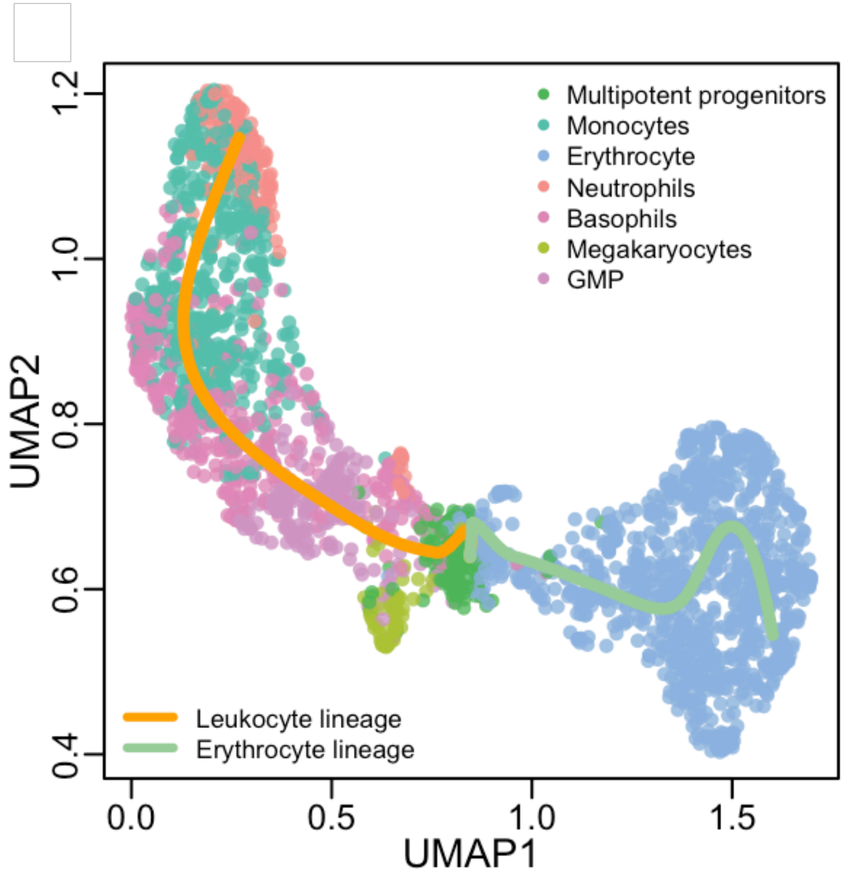
- Dynamical systems are often best represented by a continuous transition
- This continuity is represented with a trajectory
- Based on the trajectory, one can estimate pseudotime for each cell
- Pseudotime corresponds to the length of the trajectory, and can be considered as a proxy for true developmental time



Street *et al.* (2018)

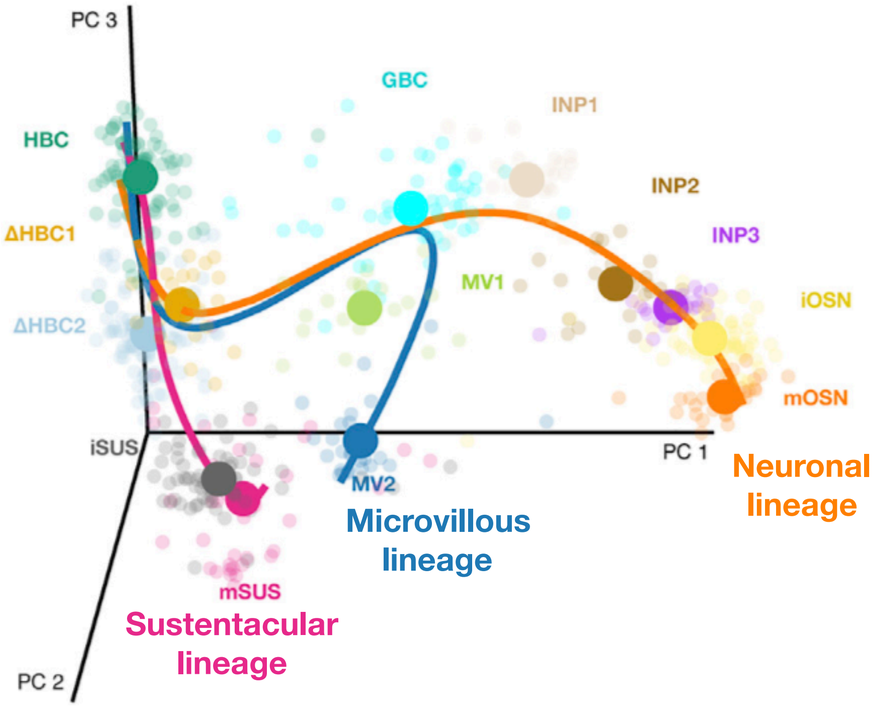
scRNA-seq allows the study of dynamic biological systems

Haematopoiesis



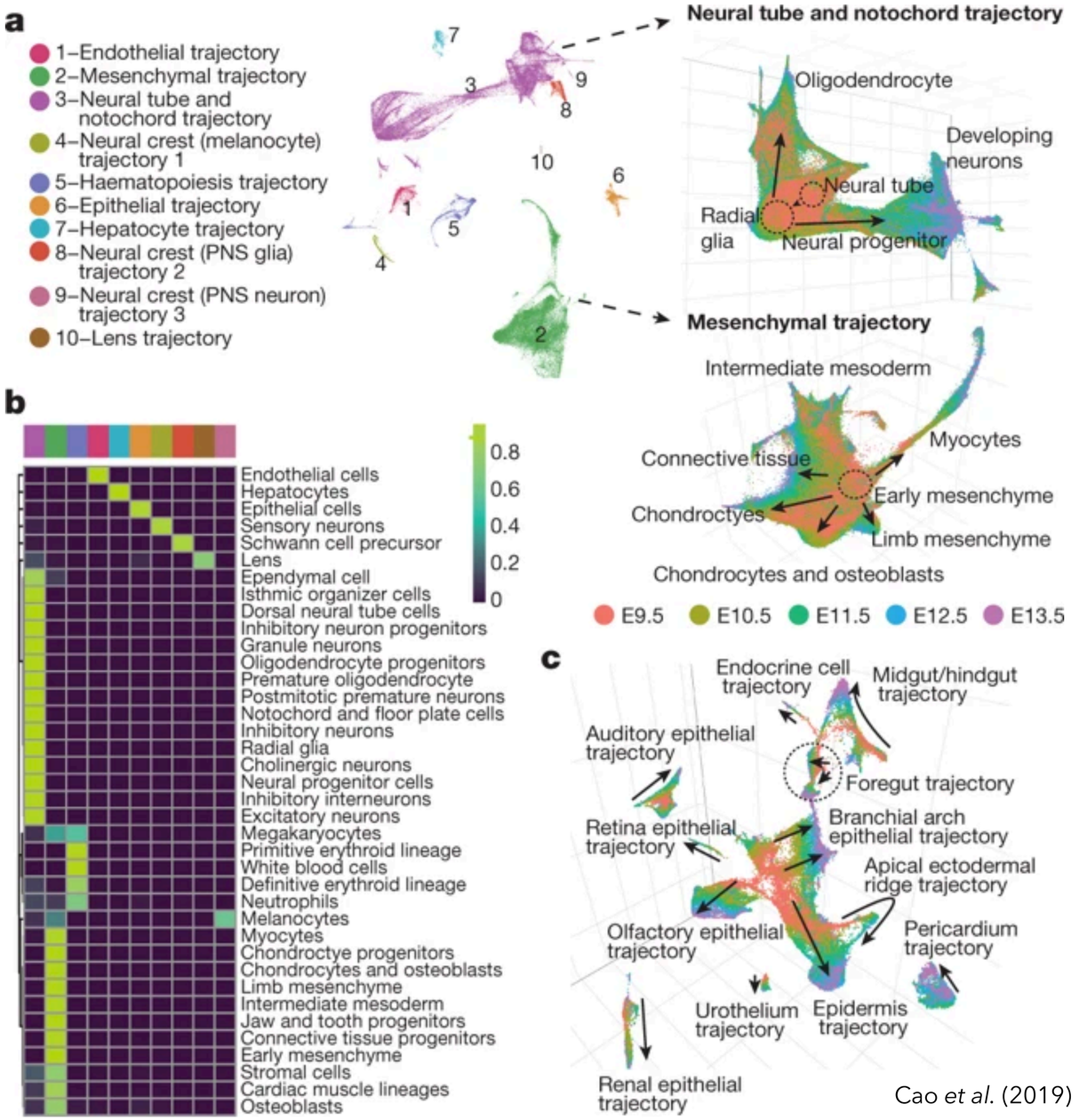
Paul *et al.* (2015)

Olfactory epithelium development



Fletcher *et al.* (2017)

scRNA-seq allows the study of dynamic biological systems



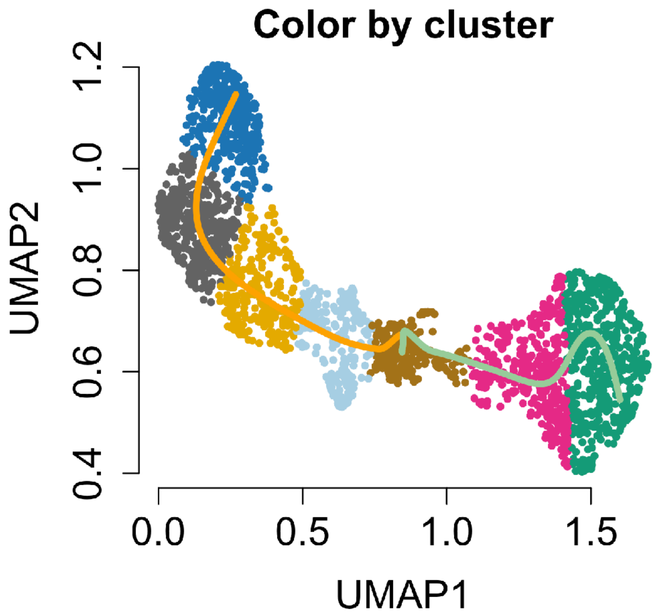
Cao *et al.* (2019)

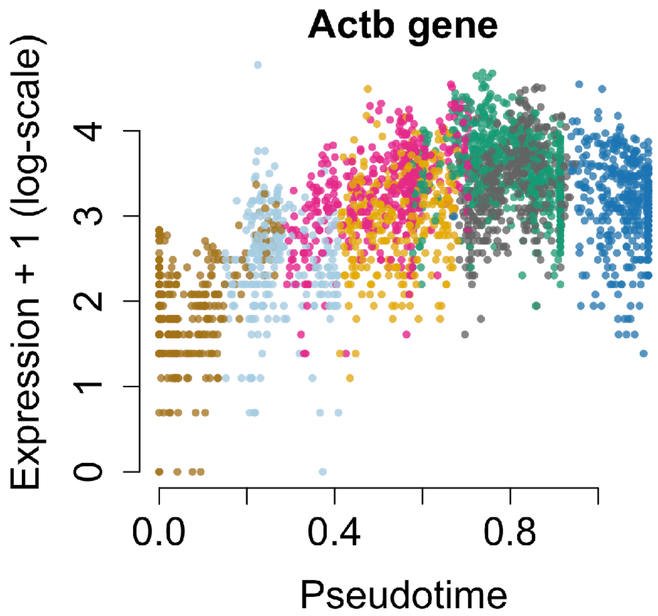
Table of contents

In dynamic systems, groups for differential expression analysis cannot be easily derived

Previous work has performed cluster-based comparisons. This is suboptimal, because:

- Heterogeneous clusters
- Clusters (often) have no fixed biological meaning
- Which clusters to compare?
- Many comparisons per gene inflates gene-level FDR
- How to derive shortlist of interesting genes?

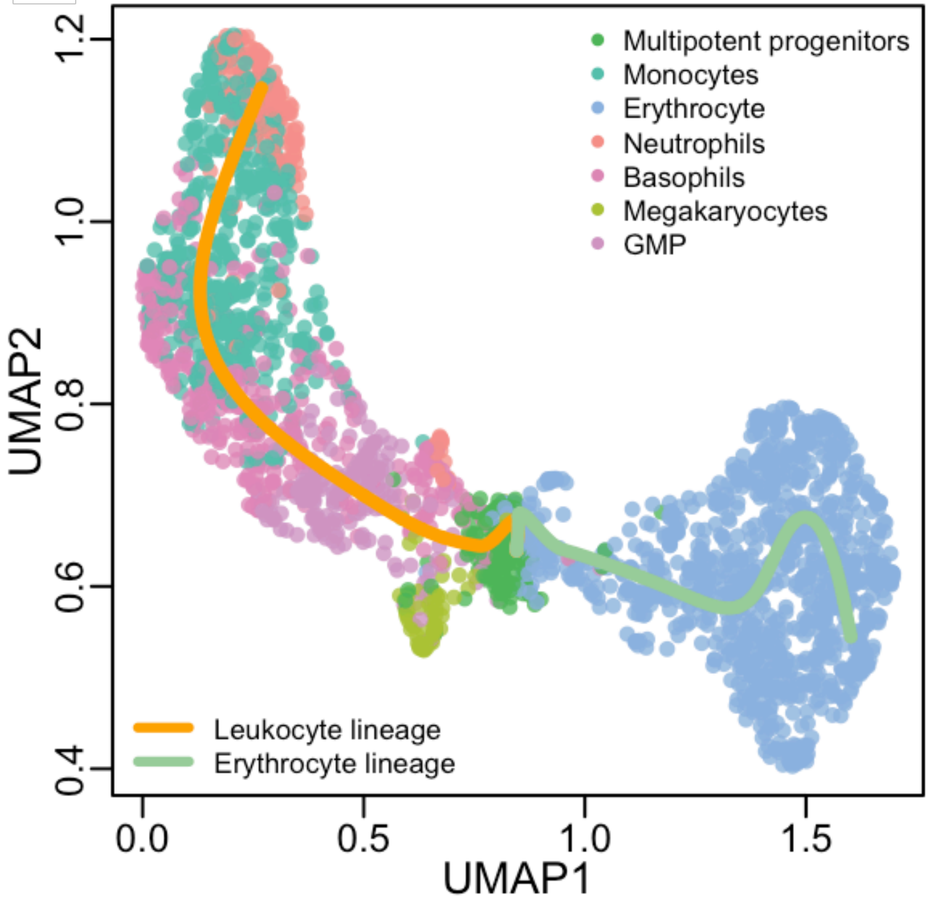


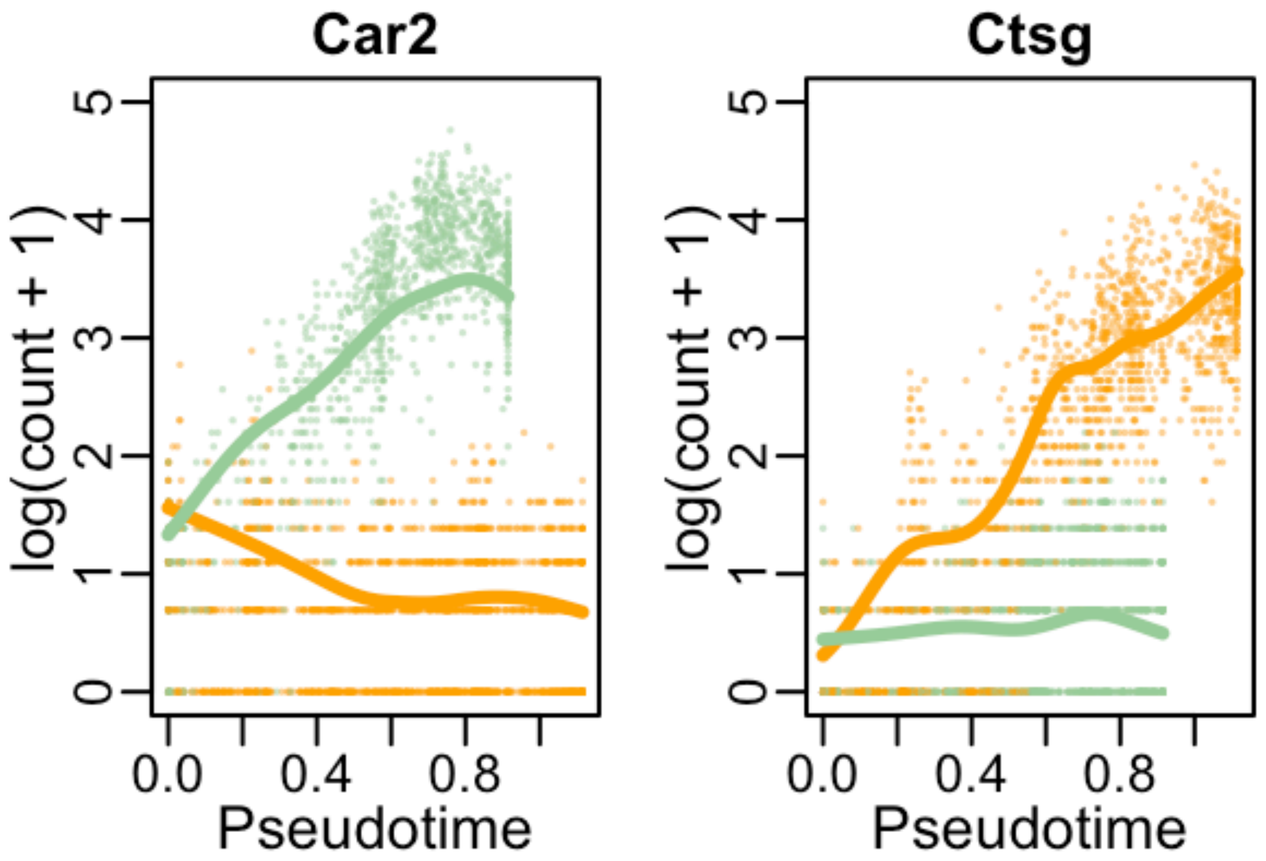


Data from Paul *et al.* (2015)

In dynamic systems, groups for differential expression analysis cannot be easily derived

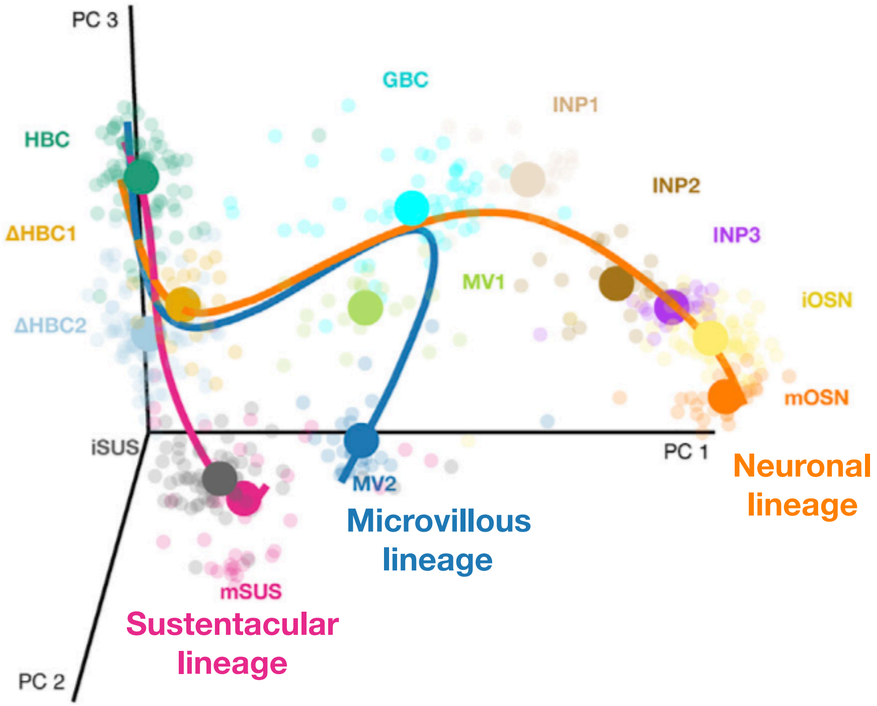
Instead, several methods have proposed smoothing gene expression along pseudotime

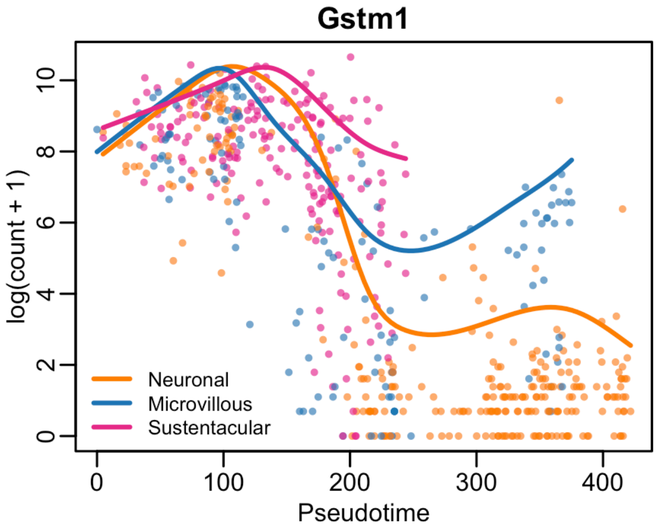


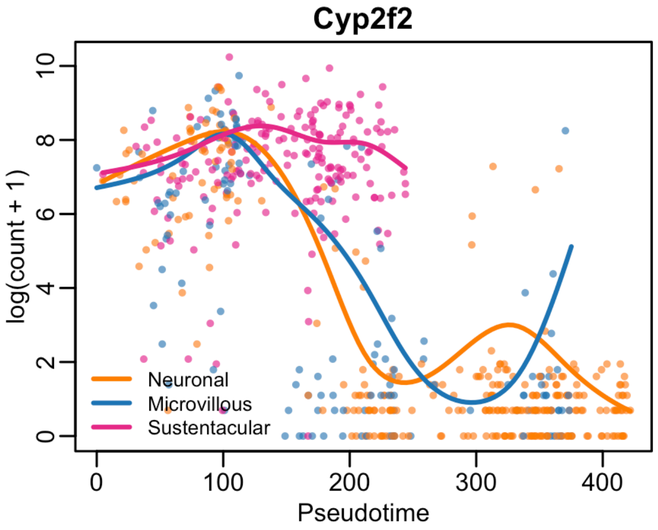


In dynamic systems, groups for differential expression analysis cannot be easily derived

Instead, several methods have proposed smoothing gene expression along pseudotime



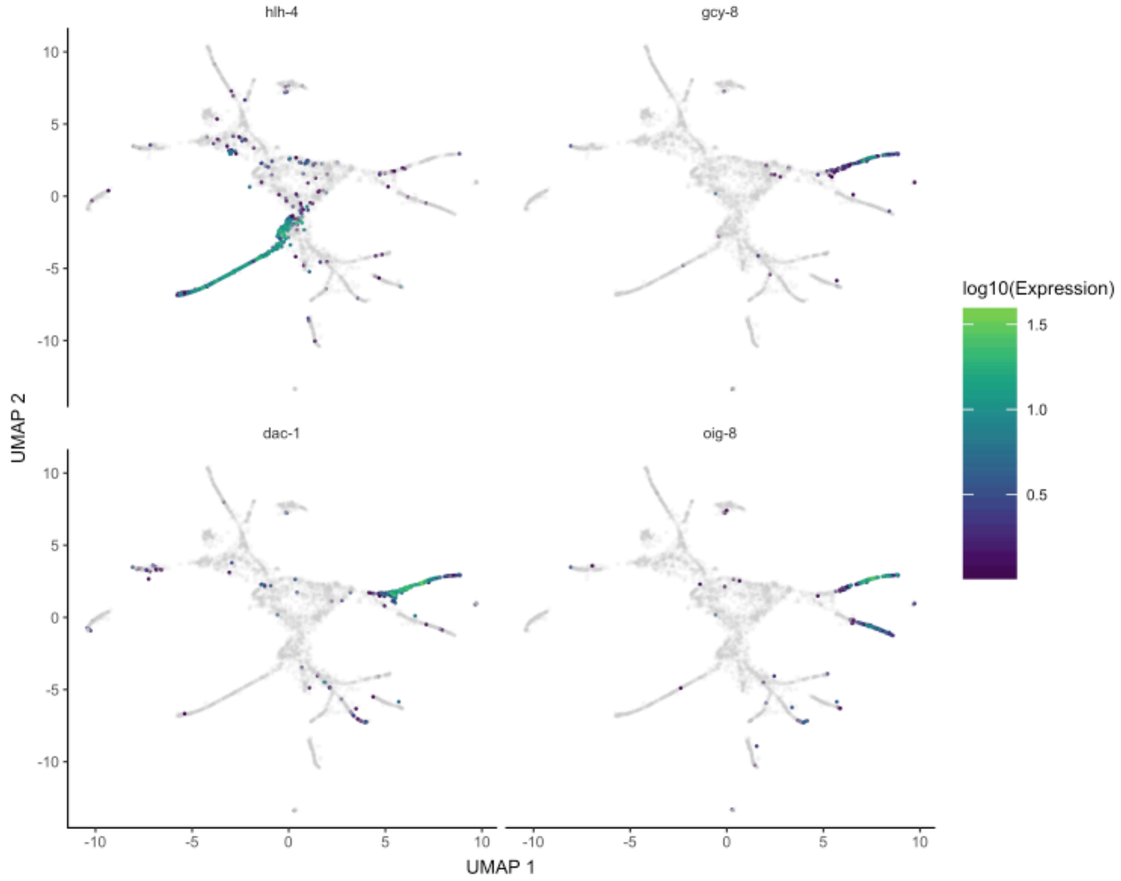




In dynamic systems, groups for differential expression analysis cannot be easily derived

Several methods have proposed smoothing gene expression along pseudotime

- Monocle 3 allows a range of functionalities, see https://cole-trapnell-lab.github.io/monocle3/monocle3_docs/#differential-expression-analysis

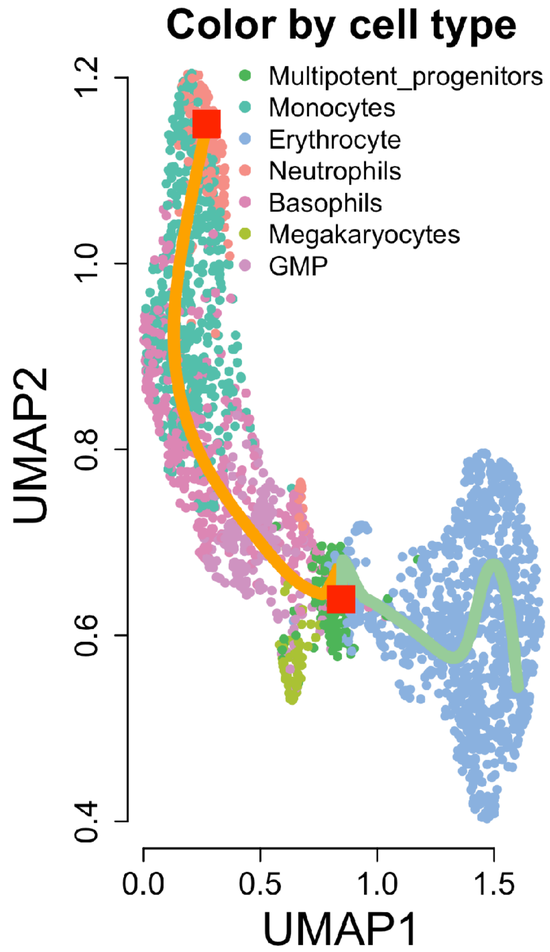


In dynamic systems, groups for differential expression analysis cannot be easily derived

- tradeSeq allows assessment of interpretable hypotheses
  - Within-lineage differential expression
    - Association of gene expression with pseudotime
    - Comparing progenitor vs. differentiated cell population
  - Between-lineage differential expression
    - Global differential expression pattern
    - Compare end points of lineages
    - Different expression pattern in a confined region selected by the user

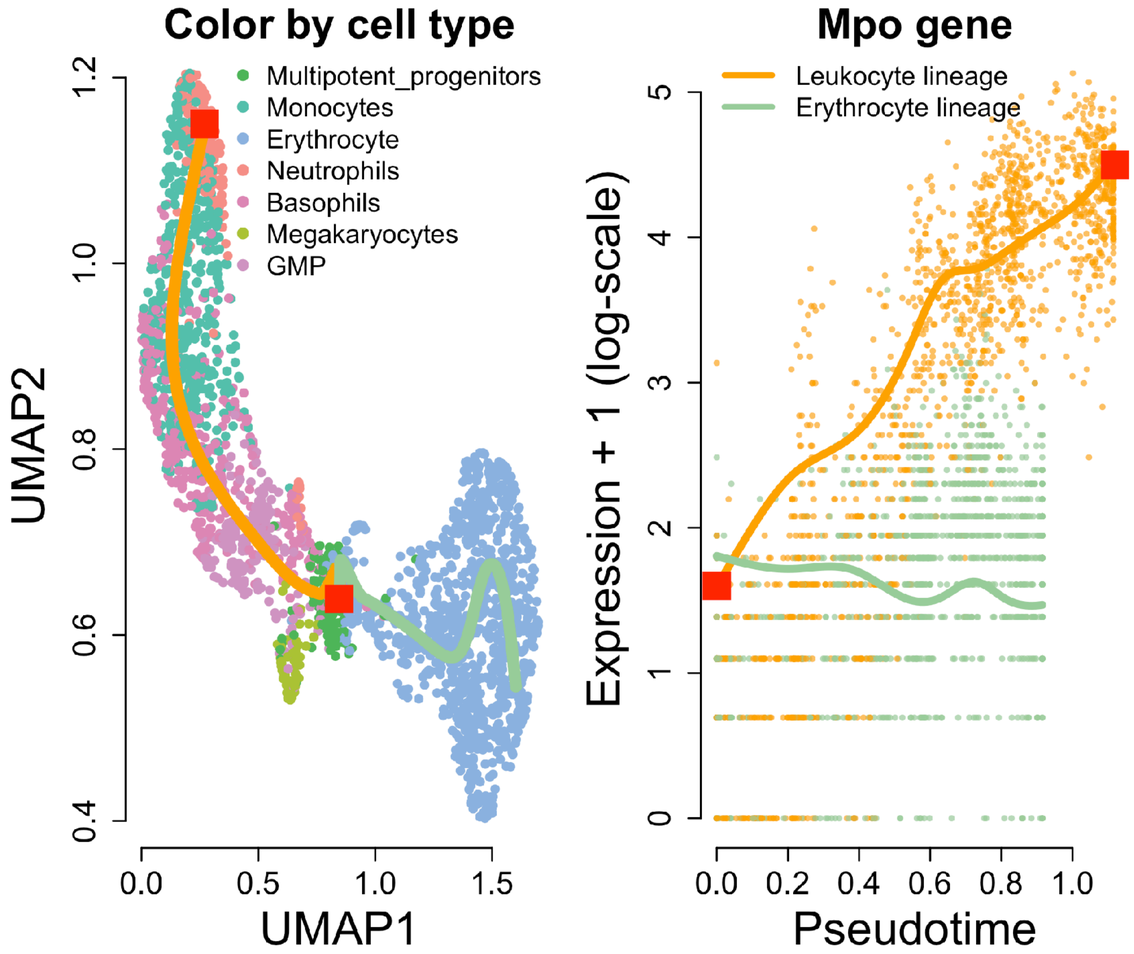
Example: Within-lineage differential expression with tradeSeq

- Global association of gene expression with pseudotime for a lineage
- Comparing start versus end points of a lineage (shown)



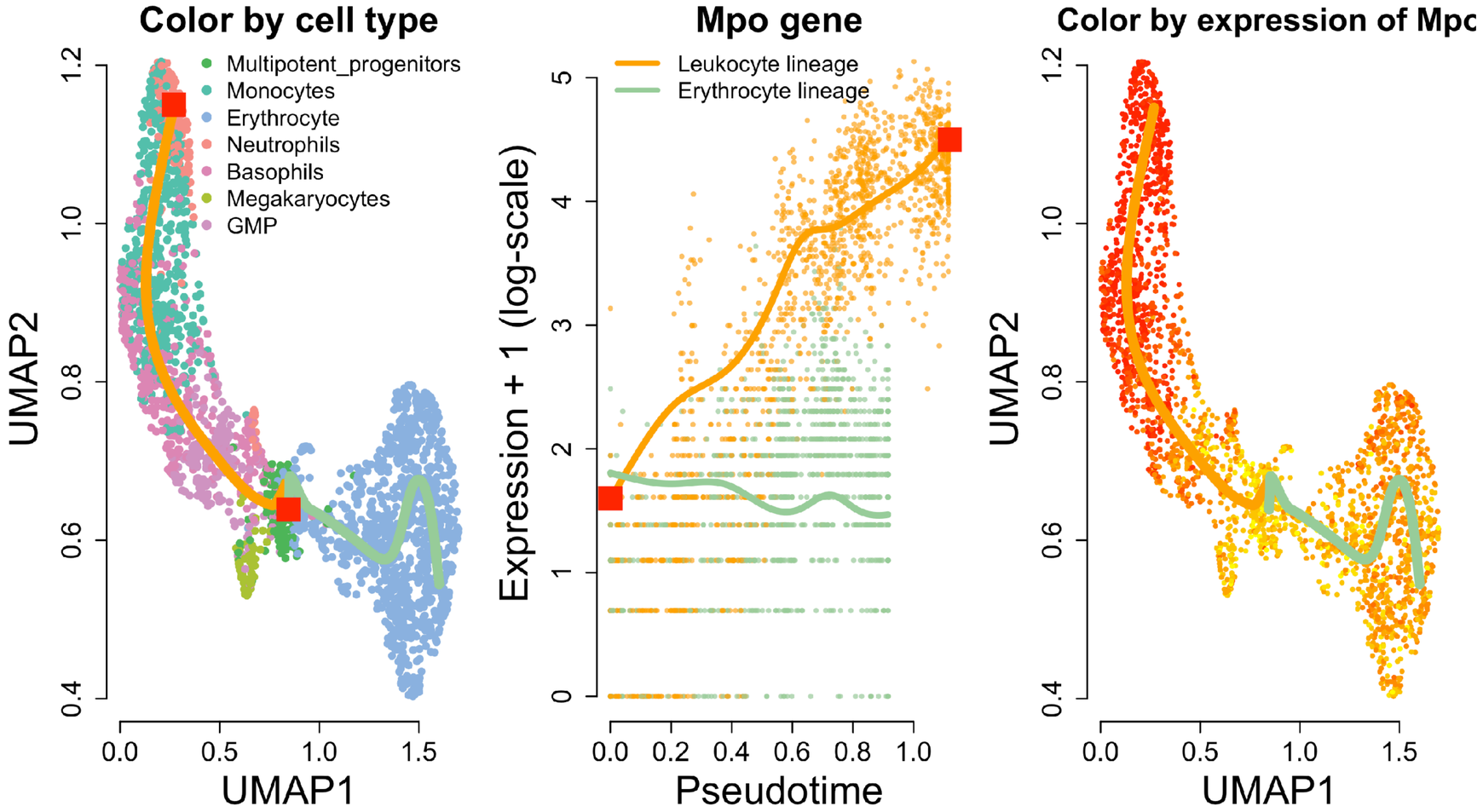
Example: Within-lineage differential expression with tradeSeq

- Global association of gene expression with pseudotime for a lineage
- Comparing start versus end points of a lineage (shown)

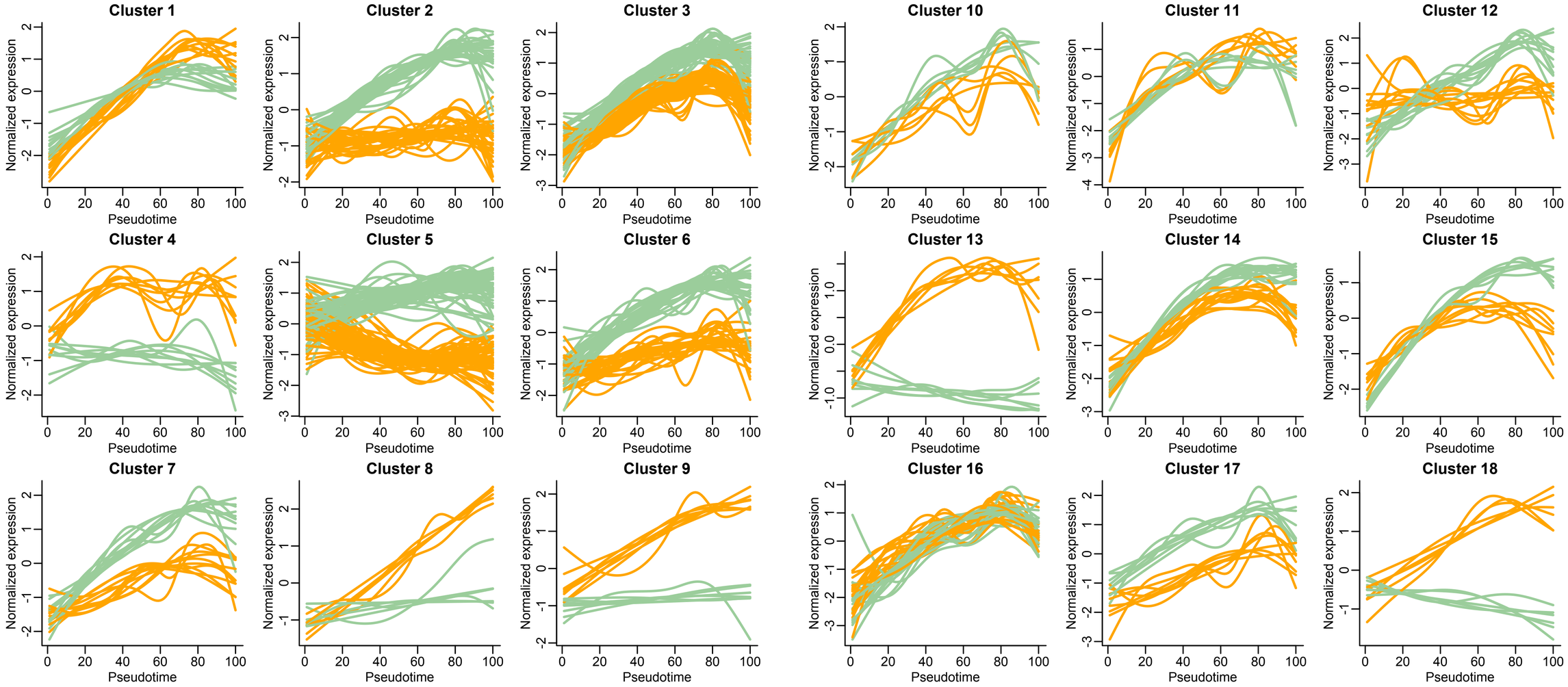


Example: Within-lineage differential expression with tradeSeq

- Global association of gene expression with pseudotime for a lineage
- Comparing start versus end points of a lineage (shown)



Genes can be clustered according to their expression pattern

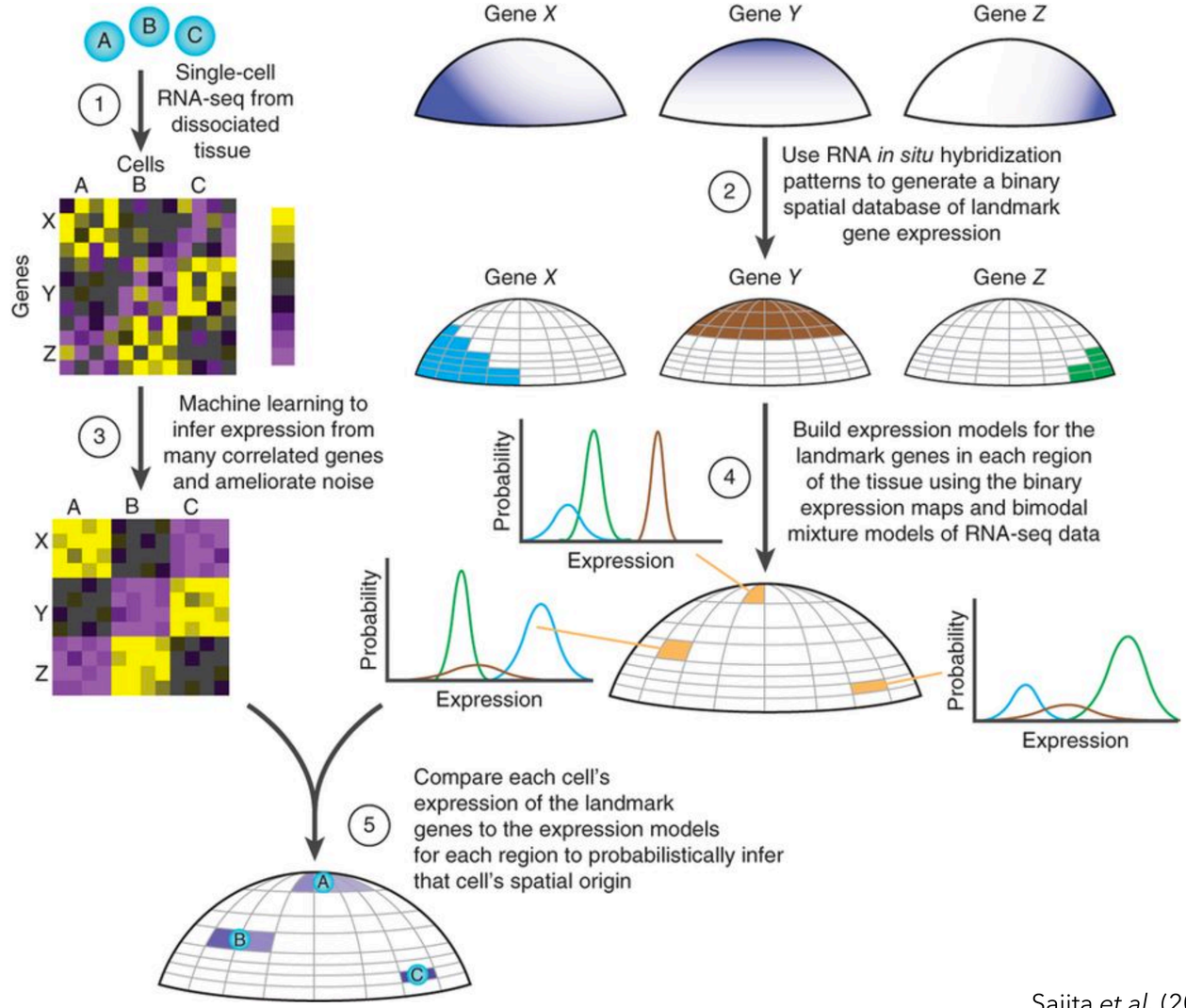


Van den Berge *et al.* (2019)

Table of contents

Novel technologies are allowing for spatial scRNA-seq

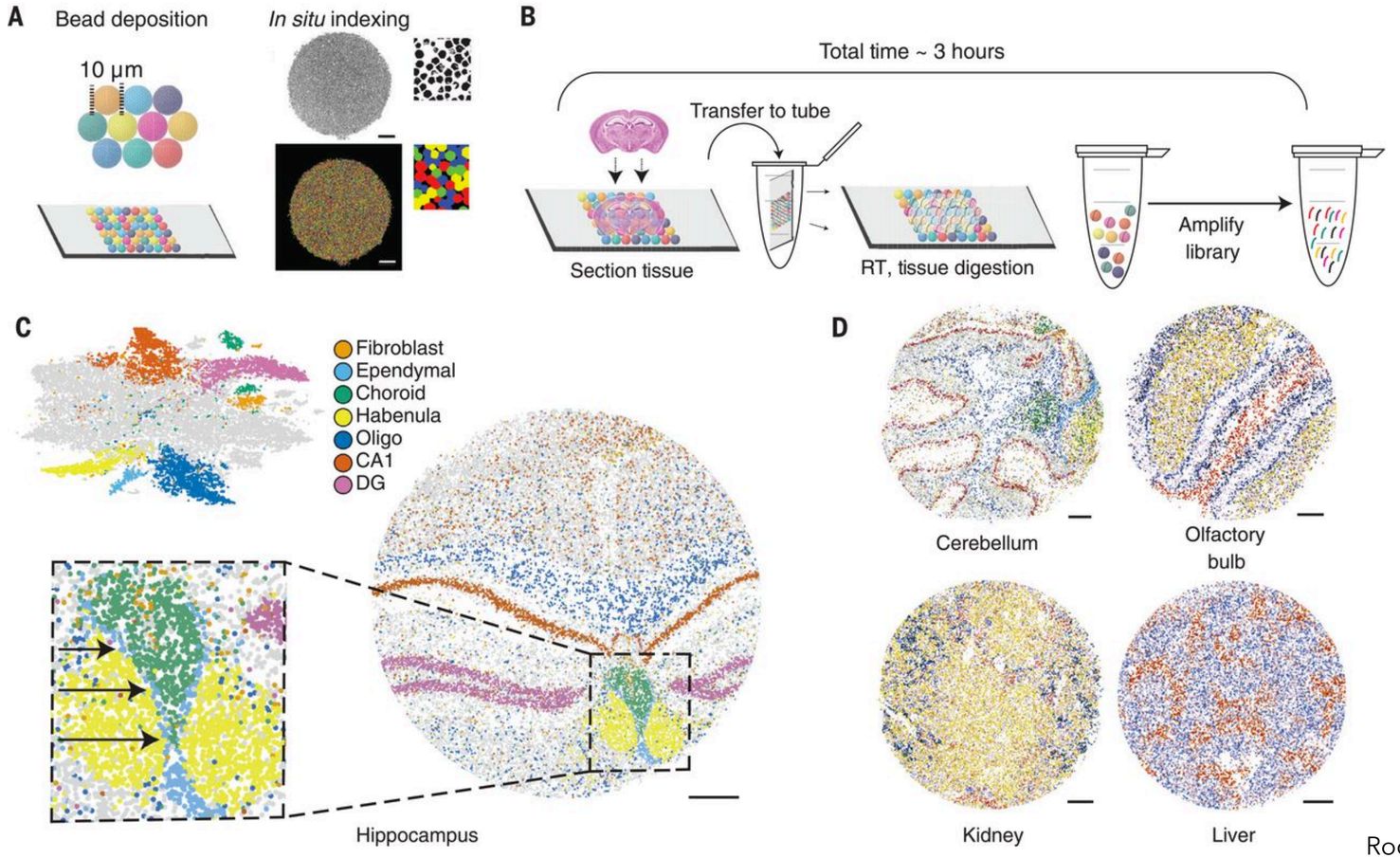
- Map dissociated cells using landmark genes from which spatial expression is known (Satija *et al*. (2015))



Sajita *et al*. (2015)

Novel technologies are allowing for spatial scRNA-seq

- Using fluorescence in situ hybridization over multiple rounds, e.g., seqFISH+ (Eng *et al.* (2019))
- Transferring tissue section on a surface covered with barcoded beads (Rodriques *et al.* (2019))



Rodriques *et al.* (2019)

Recent years: development of many single-cell multi-omics protocols

- RNA is only one of the many modalities one can study within a single cell
- Many novel developments have focussed on assessing RNA simultaneously with other modalities within the same single cell, e.g.
  - REAP-seq, CITE-seq: RNA and protein abundance
  - sci-CAR: RNA abundance and chromatin conformation (i.e., ATAC-seq)
  - G&T-seq: DNA-seq and RNA-seq
  - sc-GEM: RNA-seq, with genotype and methylation information
  - …

Want to get your hands dirty? Here's where to start

- A step-by-step workflow for low-level analysis of single-cell RNA-seq data with Bioconductor: https://f1000research.com/articles/5-2122

- Bioconductor workflow for single-cell RNA sequencing: Normalization, dimensionality reduction, clustering and lineage inference https://f1000research.com/articles/6-1158

- Current best practices in single-cell RNA-seq analysis: https://www.embopress.org/doi/pdf/10.15252/msb.20188746

- Orchestrating single-cell analysis with Bioconductor: https://www.biorxiv.org/content/10.1101/590562v1

- Hemberg Lab single-cell RNA-seq course website: https://scrnaseq-course.cog.sanger.ac.uk/website/index.html

References

- Buettner et al. 2015 Nature Biotechnology https://www.nature.com/articles/nbt.3102
- Larsson et al. 2019 Nature https://www.nature.com/articles/s41586-018-0836-1
- Grün et al. 2013 Cell https://www.sciencedirect.com/science/article/pii/S0092867415013537
- Grün et al. 2014 Nature methods https://www.nature.com/articles/nmeth.2930
- Montoro et al. 2018 Nature https://www.nature.com/articles/s41586-018-0393-7
- Fletcher et al. 2017 Cell https://www.sciencedirect.com/science/article/pii/S1934590917301273
- Svensson et al. 2018 Nature Protocols https://www.nature.com/articles/nprot.2017.149
- Griffiths et al. 2018 Molecular Systems Biology http://msb.embopress.org/content/14/4/e8046
- Vallejos et al. 2017 Nature Methods https://www.nature.com/articles/nmeth.4292
- McCarthy et al. 2017 Bioinformatics https://academic.oup.com/bioinformatics/article/33/8/1179/2907823
- McGinnis et al. 2019 Cell Systems https://www.sciencedirect.com/science/article/abs/pii/S2405471219300730

References

- Lun et al. 2019 Genome Biology https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1662-y
- Macosko et al. 2015 Cell https://doi.org/10.1016/j.cell.2015.05.002
- Townes et al. 2019 bioRxiv https://www.biorxiv.org/content/10.1101/574574v1
- Seurat: https://satijalab.org/seurat/
- Risso et al. 2018 PLoS Comp Bio https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006378
- Van den Berge, Perraudeau et al. 2018 Genome Biology https://genomebiology.biomedcentral.com/articles/10.1186/s13059-018-1406-4
- Hicks et al. 2018 Biostatistics https://academic.oup.com/biostatistics/article/19/4/562/4599254
- Risso et al. 2018b Nature Communications https://www.nature.com/articles/s41467-017-02554-5
- Paul et al. 2015 Cell https://doi.org/10.1016/j.cell.2015.11.013
- Fletcher et al. 2017 Cell Stem Cell https://doi.org/10.1016/j.stem.2017.04.003
- Cao et al. 2019 Nature https://www.nature.com/articles/s41586-019-0969-x

References

- Street et al. 2018 BMC Genomics https://bmcgenomics.biomedcentral.com/articles/10.1186/s12864-018-4772-0
- Van den Berge et al. 2019 bioRxiv https://www.biorxiv.org/content/10.1101/623397v1
- Satija et al. 2015 Nature Biotechnology https://www.nature.com/articles/nbt.3192
- Eng et al. 2019 Nature https://www.nature.com/articles/s41586-019-1049-y
- Rodriques et al. 2019 Science https://science.sciencemag.org/content/363/6434/1463
- Haghverdi et al. 2018 Nature Biotechnology https://www.nature.com/articles/nbt.4091