# Single cell RNA seq Data and Analysis

Samuele Soraggi

Bioinformatics Research Center,
Aarhus University


*NGS 2022*
*summer course*

BiRC

# Content/Objectives:

time

- What is single cell data
- Why do we use it
- Sequencing framework

- Analysis of single cell data

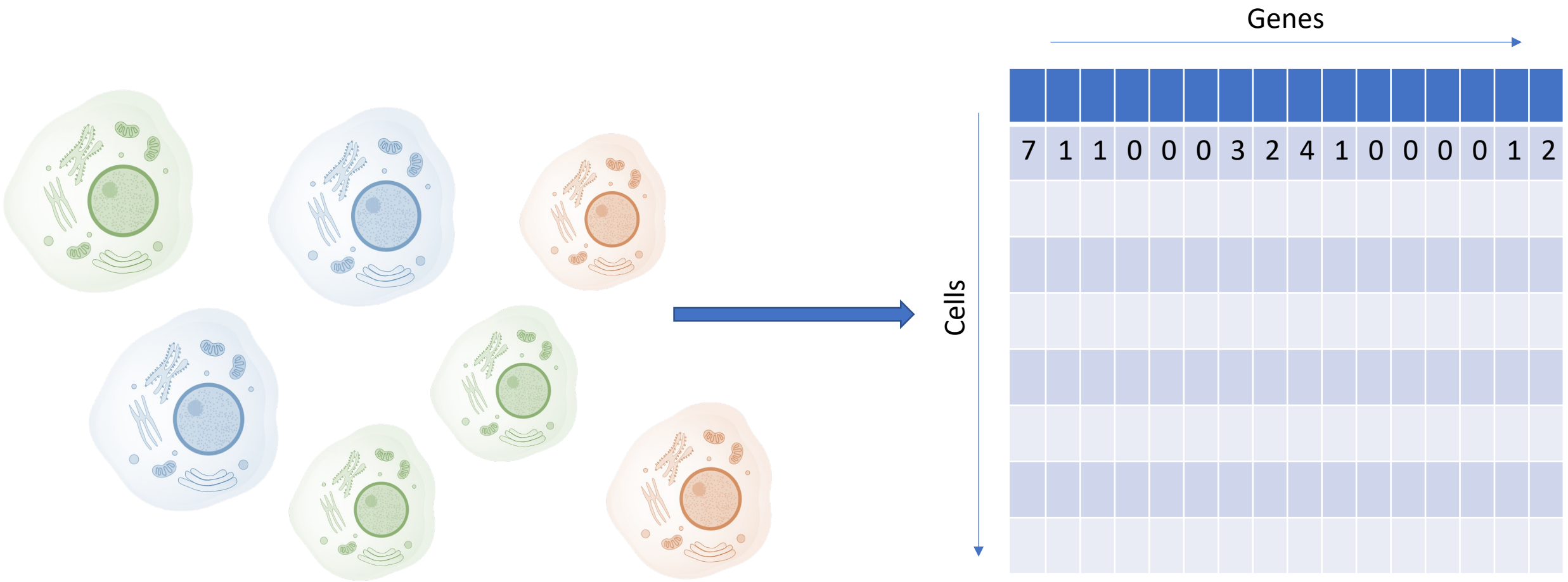
- Paper discussion

- Conclusions and introduction to exercise data

# Single cell RNA seq data (scRNAseq)

# Single cell data: what is it?

In a single cell RNA sequencing dataset (scRNAseq), we can capture mRNA transcripts of each cell, and count how many of them are associated to a set of reference genes.



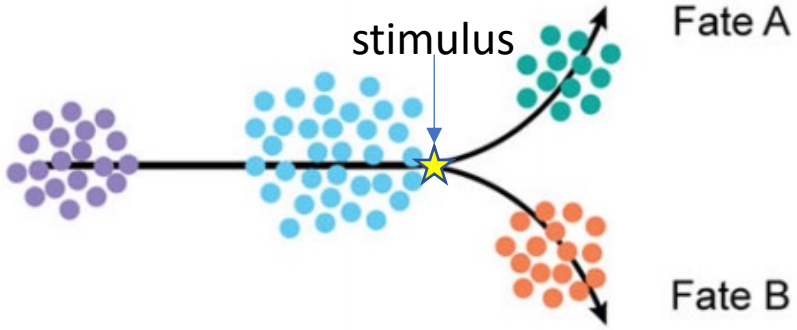
# Single cell data: what is it? Why do we use it?

scRNAseq data gives the possibility of analyzing gene transcription at cellular level. A Considerable step forward compared to bulkRNA seq data.

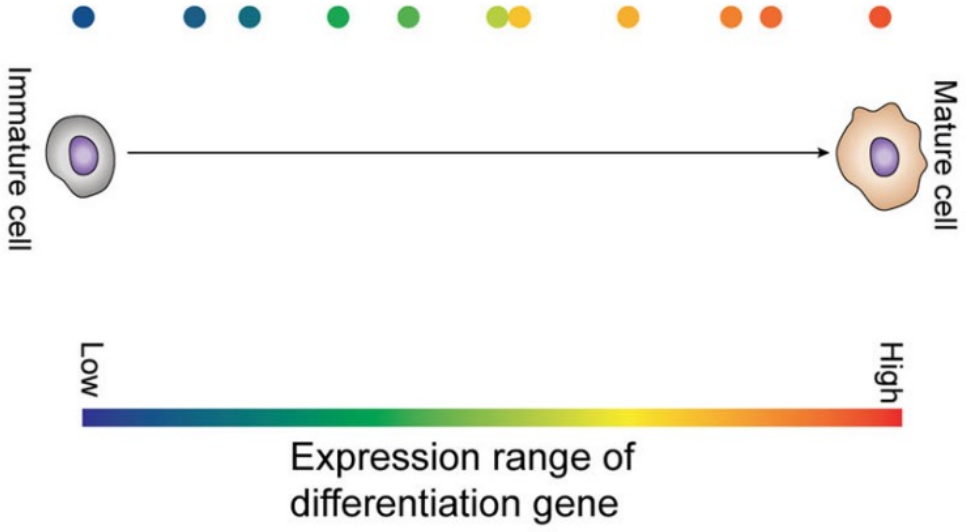
# Single cell data: what is it? Why do we use it?

scRNAseq data gives the possibility of analyzing gene transcription at cellular level to retrieve information about

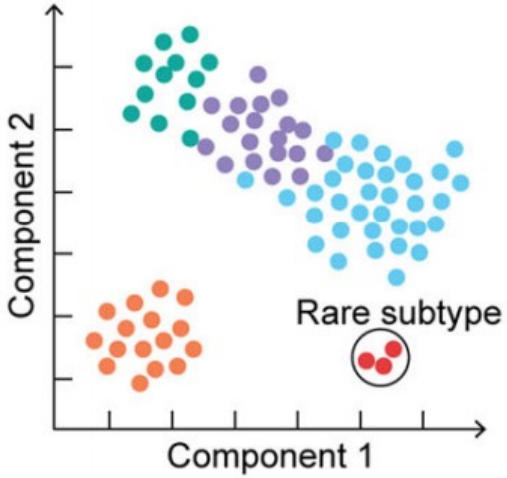
Cell fate and response to stimuli



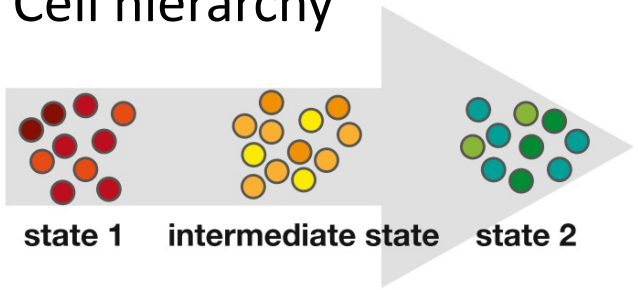

Differentiation and fate of cells




Rare cell types

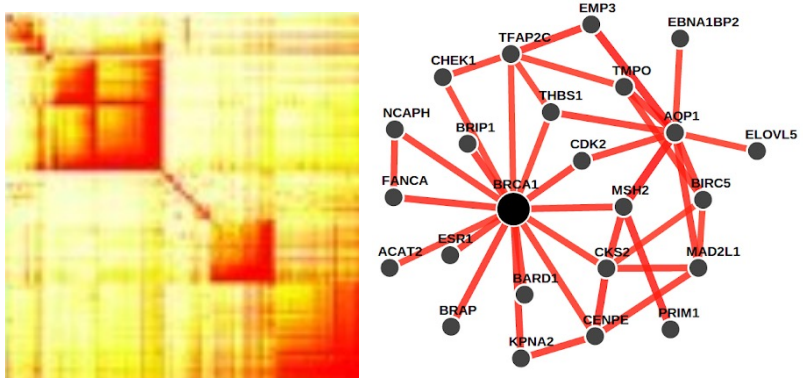



Cell hierarchy

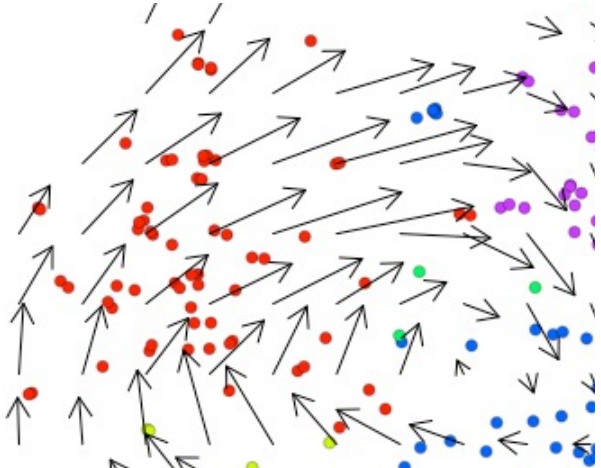



Co-expression of genes

Gene networks




Changes in transcription rate

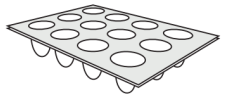


# scRNAseq usage has exploded

Manual

Tang *et al* 2009

Multiplexing

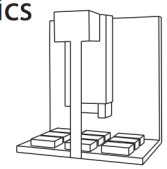


Islam *et al* 2011

Integrated Fluidic Circuits

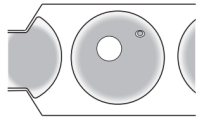
Brennecke *et al* 2013

Liquid Handling Robotics



Jaitin *et al* 2014

Nanodroplets

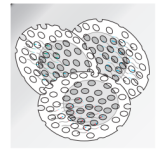


Klein *et al* 2015
Macosko *et al* 2015

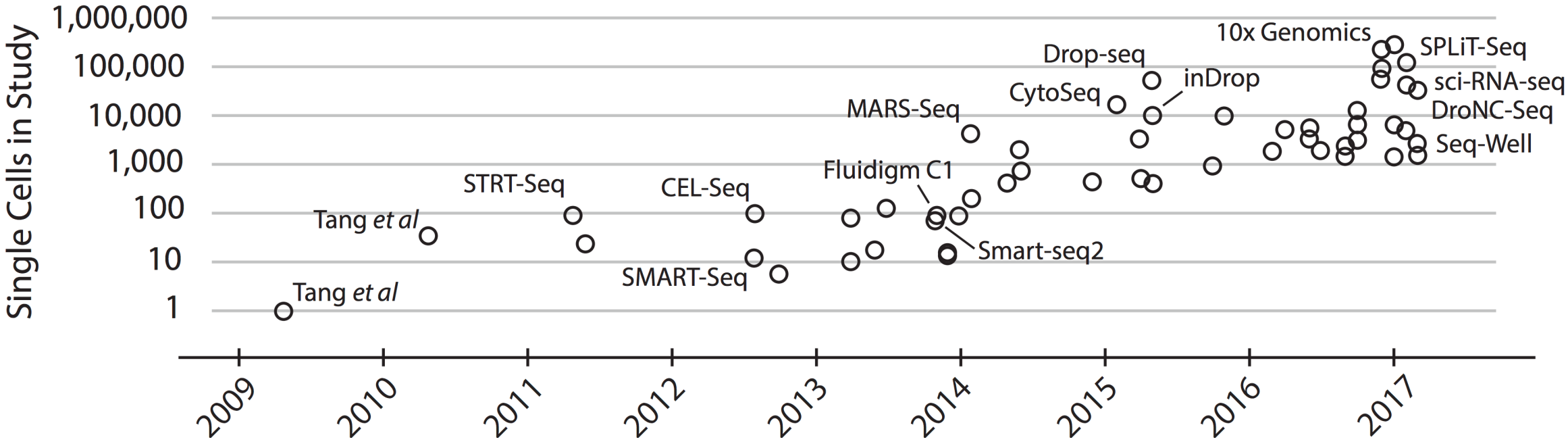
Picowells

Bose *et al* 2015

*In situ* barcoding



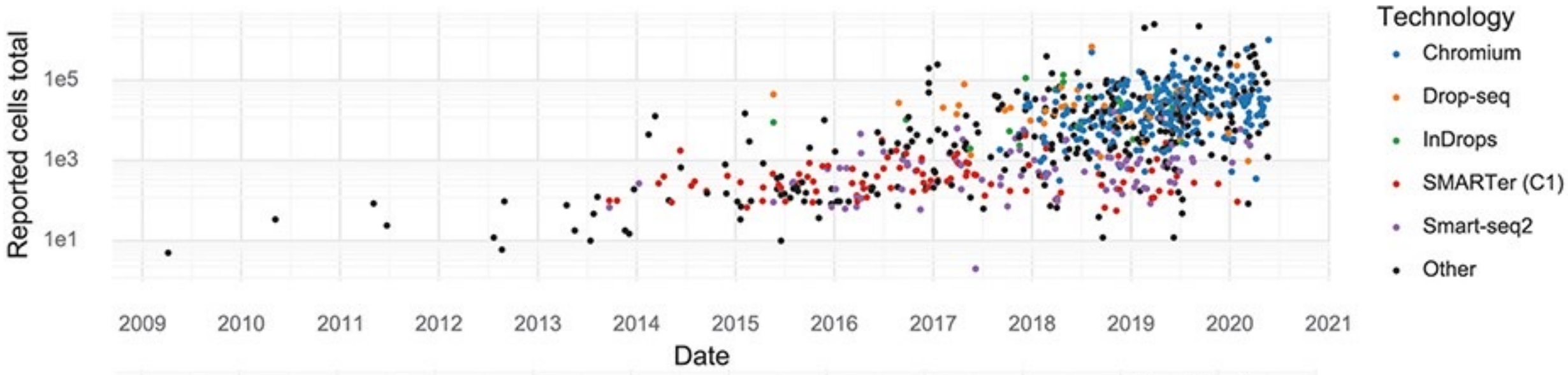
Cao *et al* 2017
Rosenberg *et al* 2017




Source: Svensson et al 2017

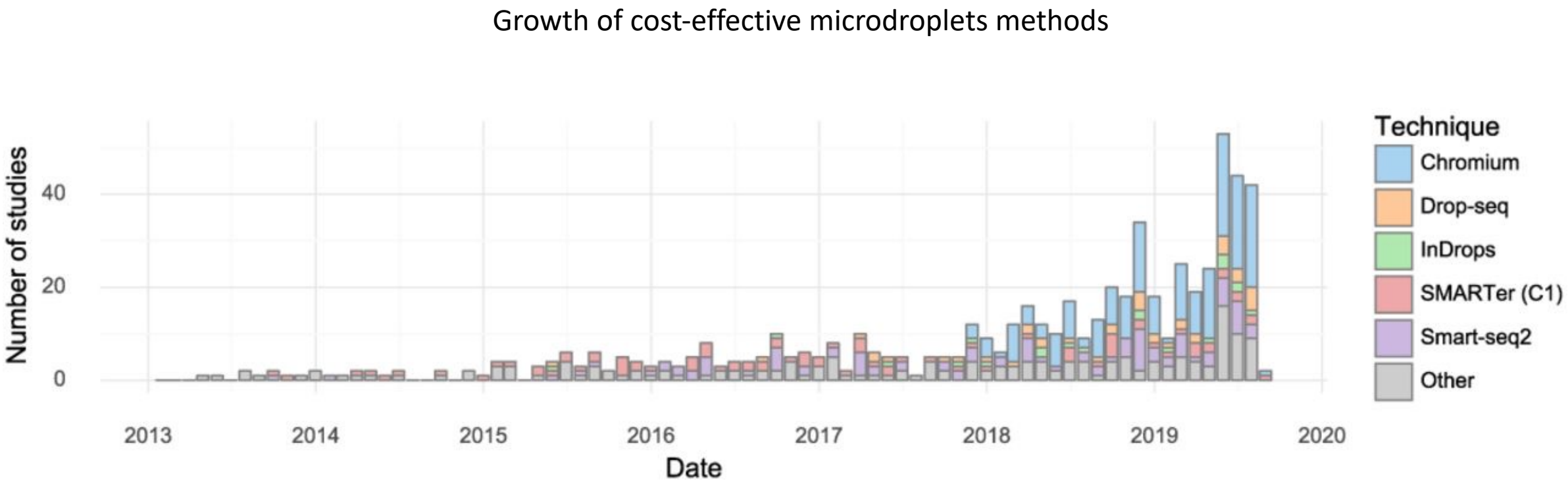
# scRNAseq usage has exploded

Increased number of cells reported per study



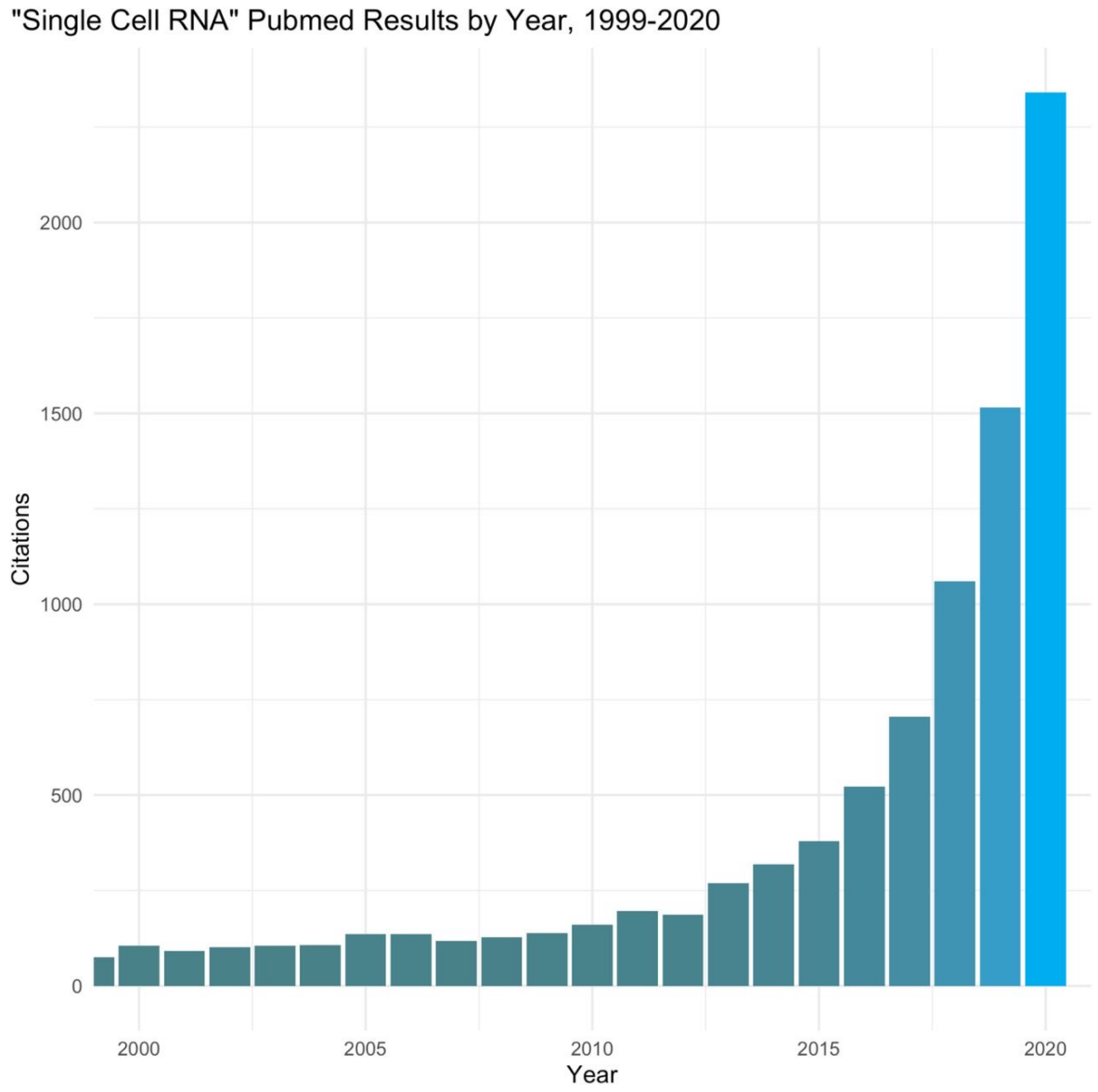

Source: Svensson and Beltrame, 2020
Interactive and Updated Graph

# scRNAseq usage has exploded




[Source: Svensson and Beltrame, 2020 Interactive and Updated Graph]

# scRNAseq usage has exploded

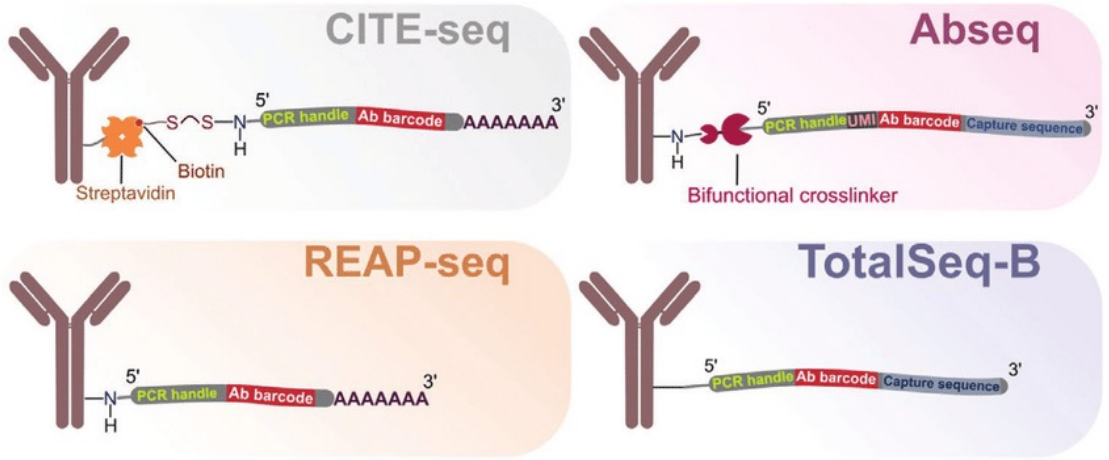



Source: Babcock et al, 2021

# Multi-OMICS frameworks are on the rise

REAPseq: proteins and mRNA simultaneously

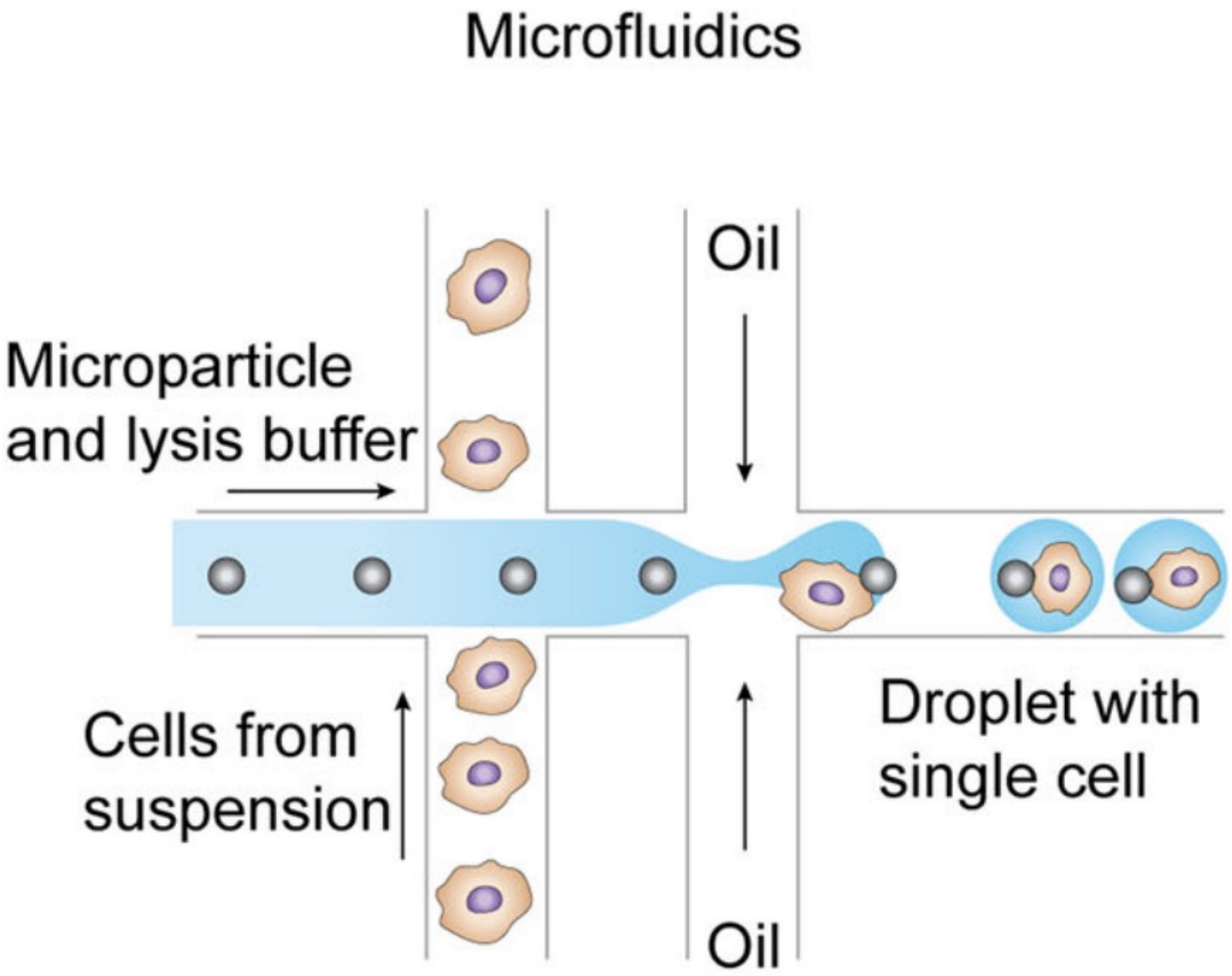
Source: Peterson et al, 2017

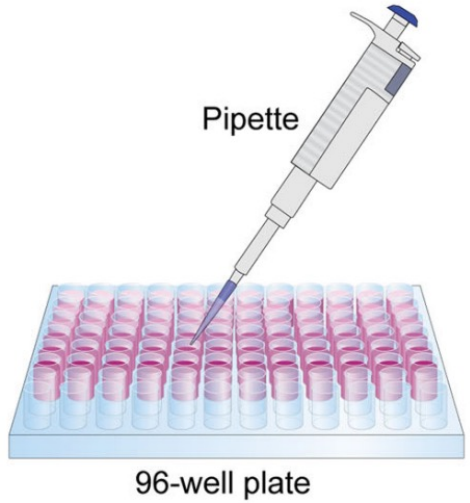


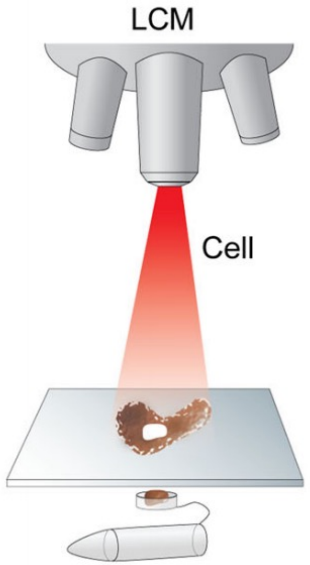

Spatial Transcriptomics: physical position and mRNA (10X visium, slideseq, …)

Source: Bergenstråhle et al, 2020

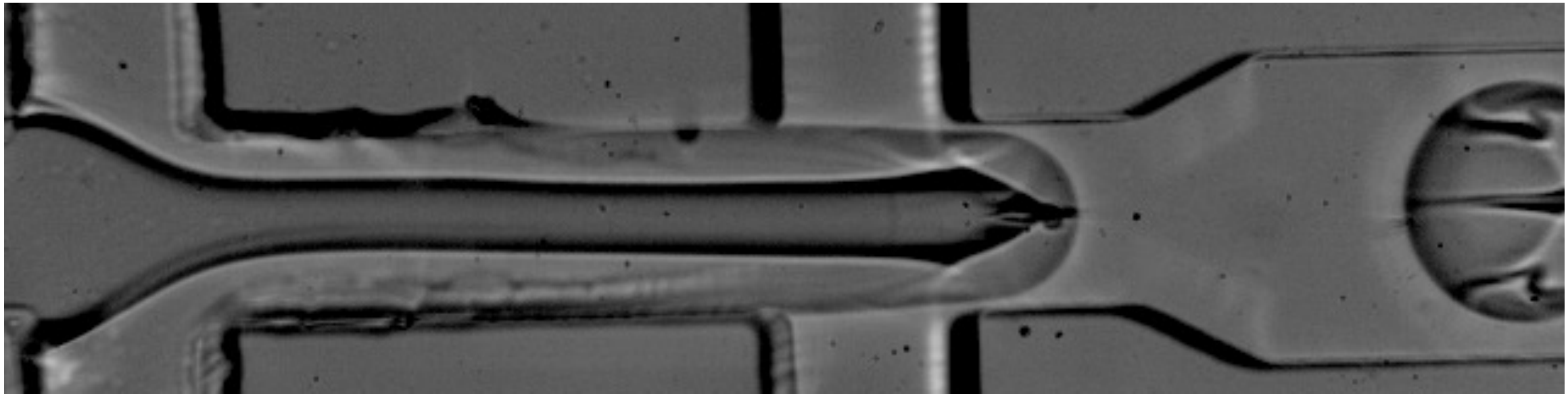
# The sequencing framework: cell isolation



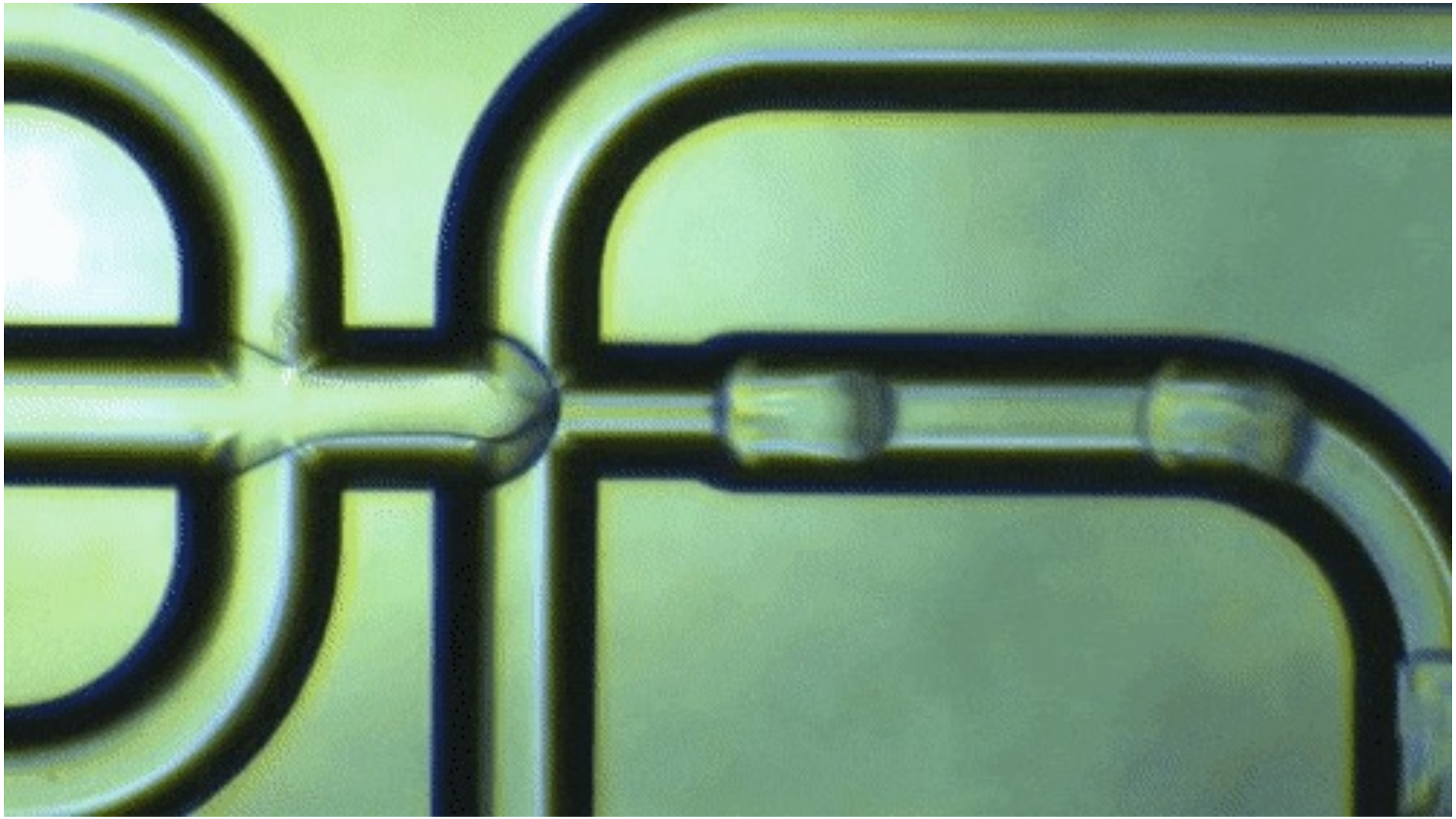



# The sequencing framework: cell isolation



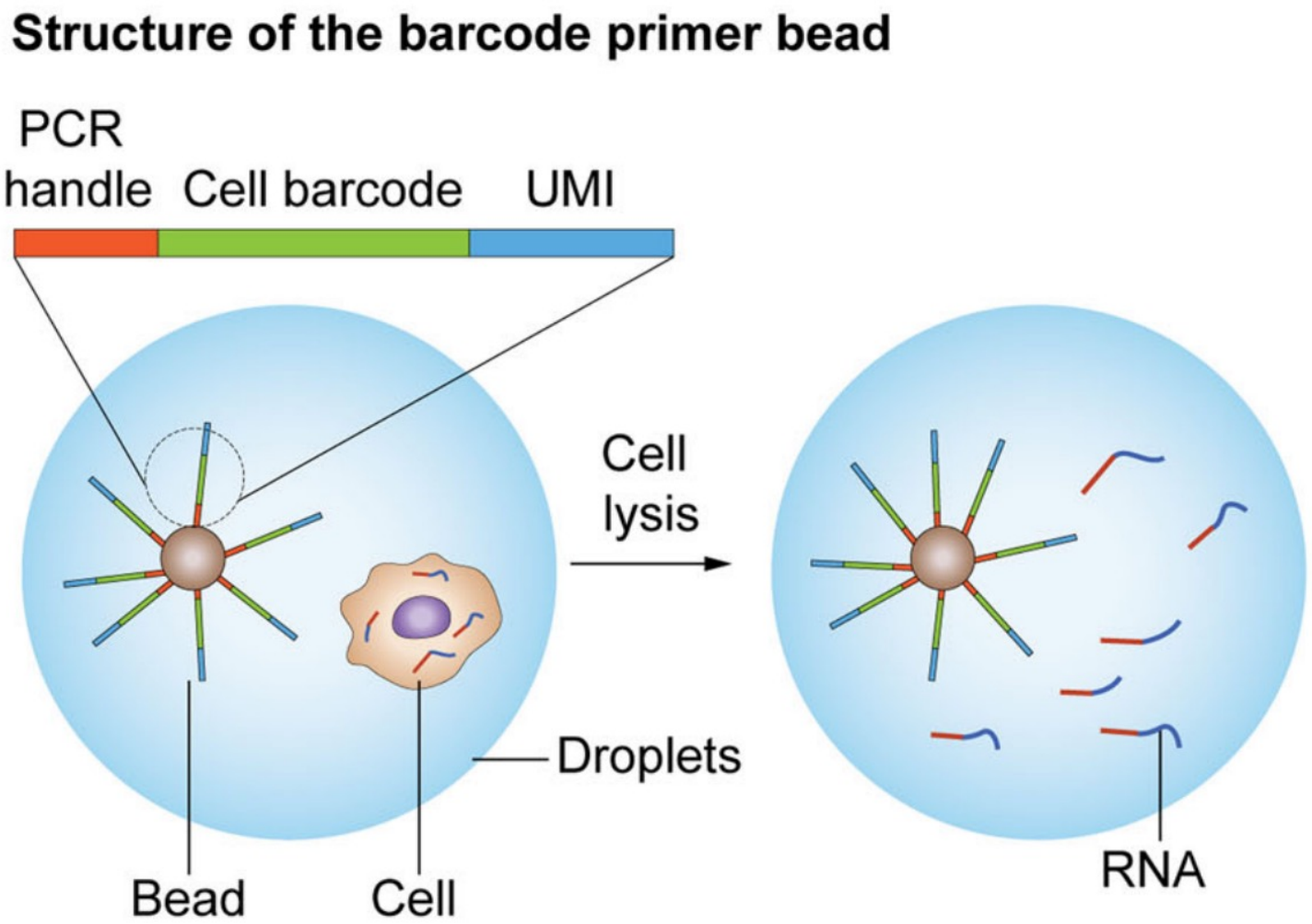
Drop-seq (Source: McCarroll Lab)

# The sequencing framework: cell isolation



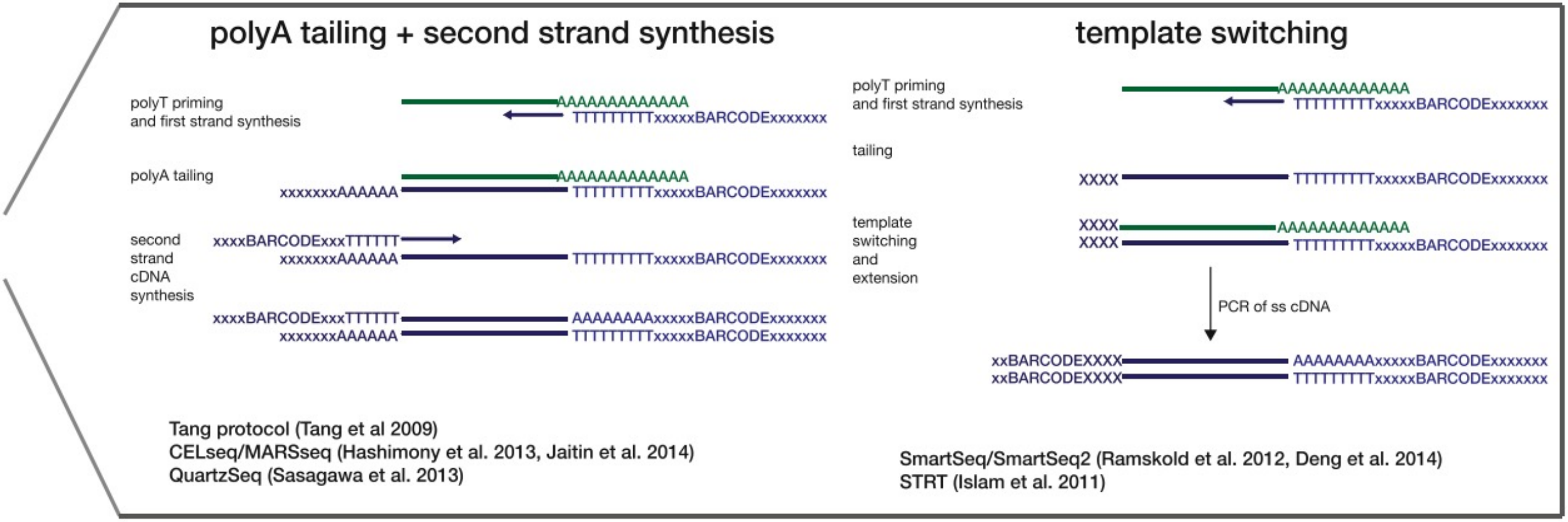
High-throughput technology (such as 10X Chromium) – faster with fewer empty droplets

# The sequencing framework: mRNA capture



# The sequencing framework: cDNA RT

Reverse transcription into cDNA

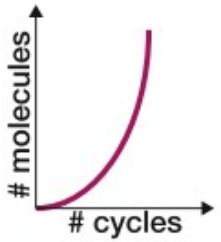


# The sequencing framework: NGS sequencing

Preamplification and NGS sequencing

## PCR

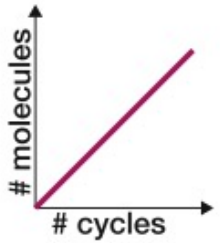
- exponential amplification
- PCR base specific biases




Tang protocol (Tang et al. 2009)
STRT (Islam et al. 2011)
SmartSeq/SmartSeq2 (Ramskold et al. 2012, Deng et al. 2014)

## IVT

- linear amplification
- 3' bias due to two rounds of reverse transcription




CELseq/MARSseq (Hashimony et al. 2013, Jaitin et al. 2014)

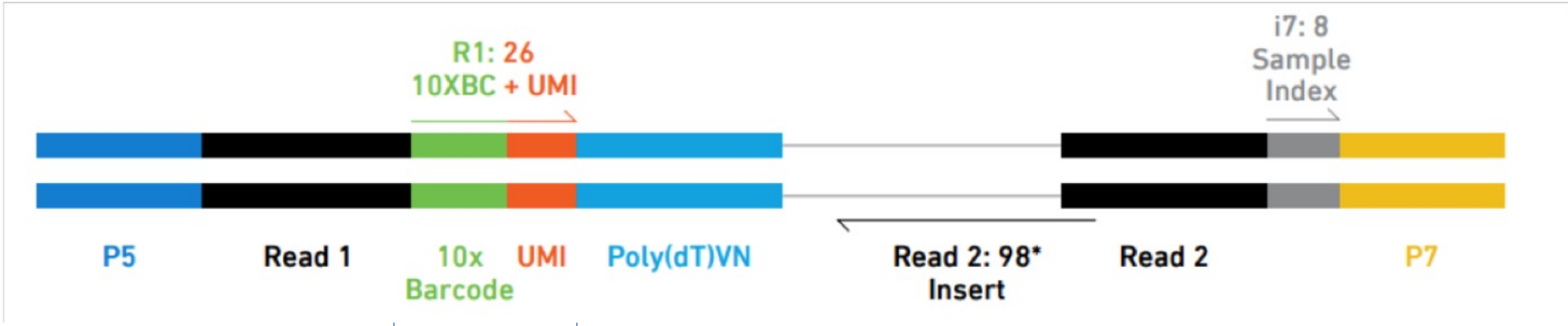
## Illumina

## AB SOLID

## PacBio

# The sequencing framework: raw reads

After NGS sequencing, a typical sequence from fastq files look like this

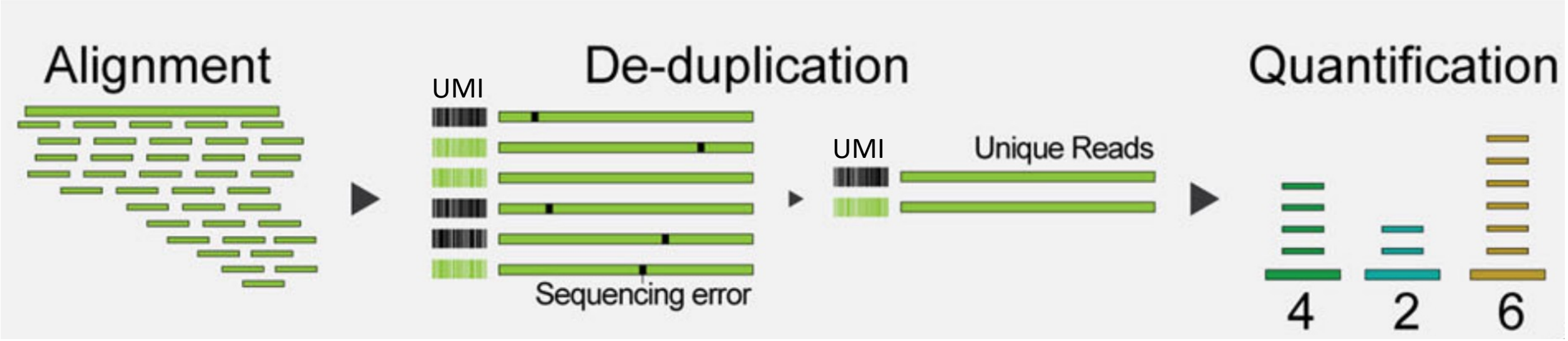



```
@SRR8363305.1 1 length=26
NTGAAGTGTTAAGACAAGCGTGAACT
+SRR8363305.1 1 length=26
# AAFFJJJJJJJJJJJJJJJJFJJJJ
```

```
@SRR8363305.1 1 length=98
NCTAAAGATCACACTAAGGCAACTCATGGAGGGGTCTT
+SRR8363305.1 1 length=98
# A<77AFJJFAAAJJJ7-7-<7FJ-7----<77--7F
```

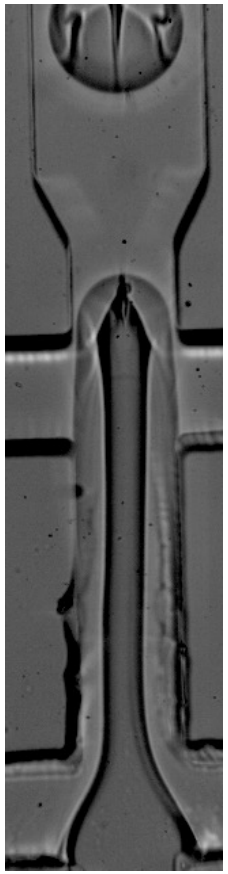
# The (post)sequencing framework: align & quantify

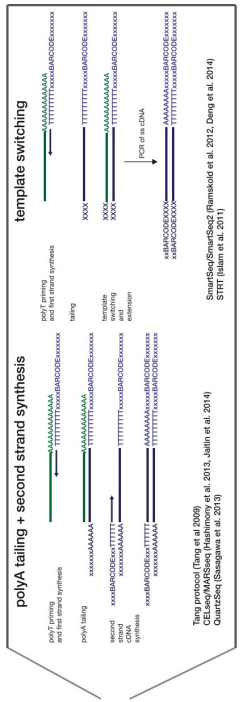
The raw fastq reads are aligned to a transcriptome to identify which reads pertain to specific genes.
The aligner has to be splice-aware and correct for sequencing errors happening in the barcodes and cDNA.
Lastly, UMIs are used to detect matching transcripts and collapse them into one if they come from the same cell.

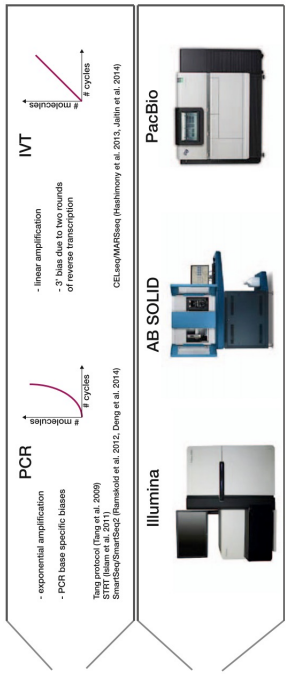


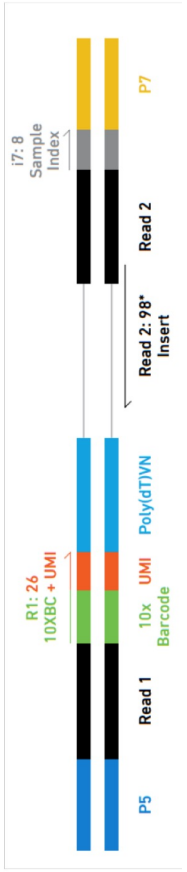
# The (post)sequencing framework: the data

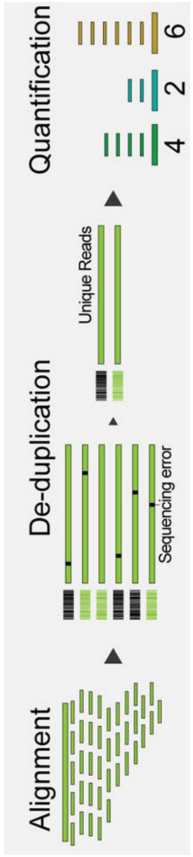
The final dataset is composed by a matrix of dimension cell x genes. For each cell, we have the quantified mRNA transcripts for each gene, detected by collapsing together matching UMI tags

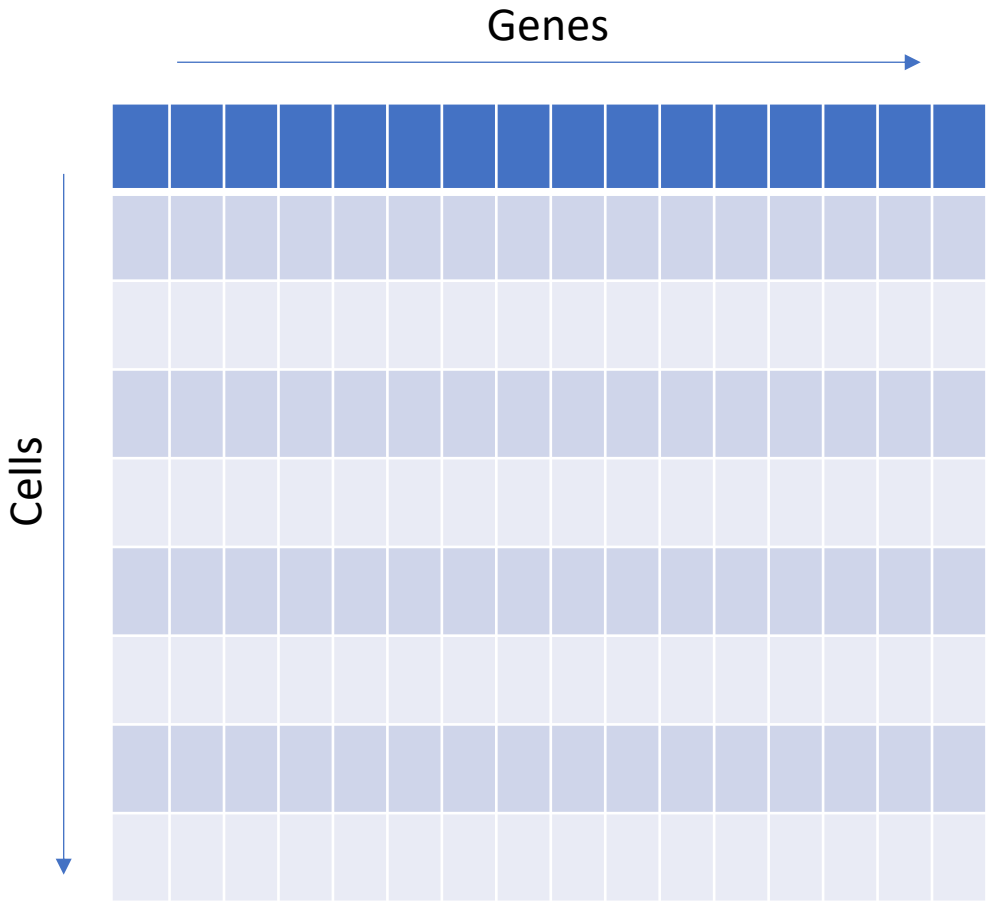




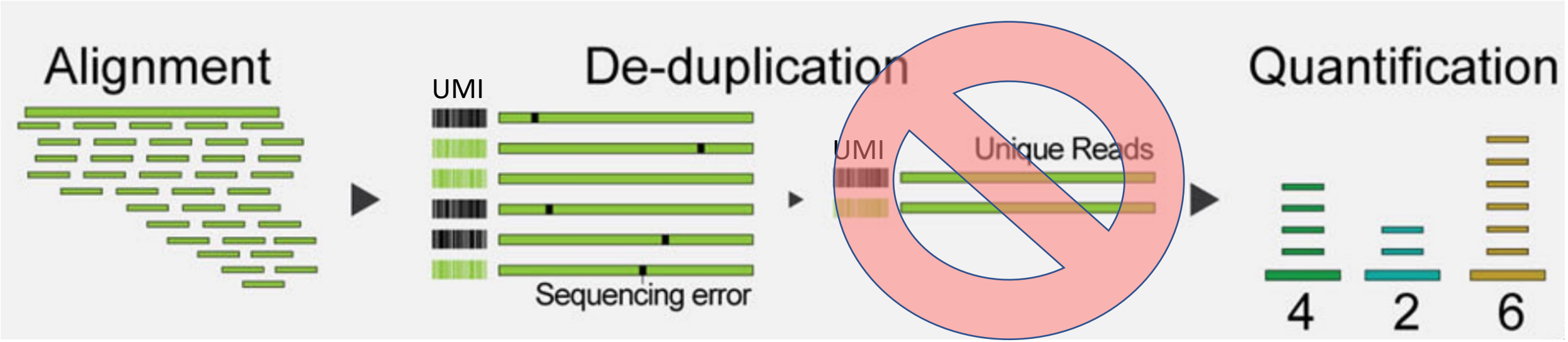



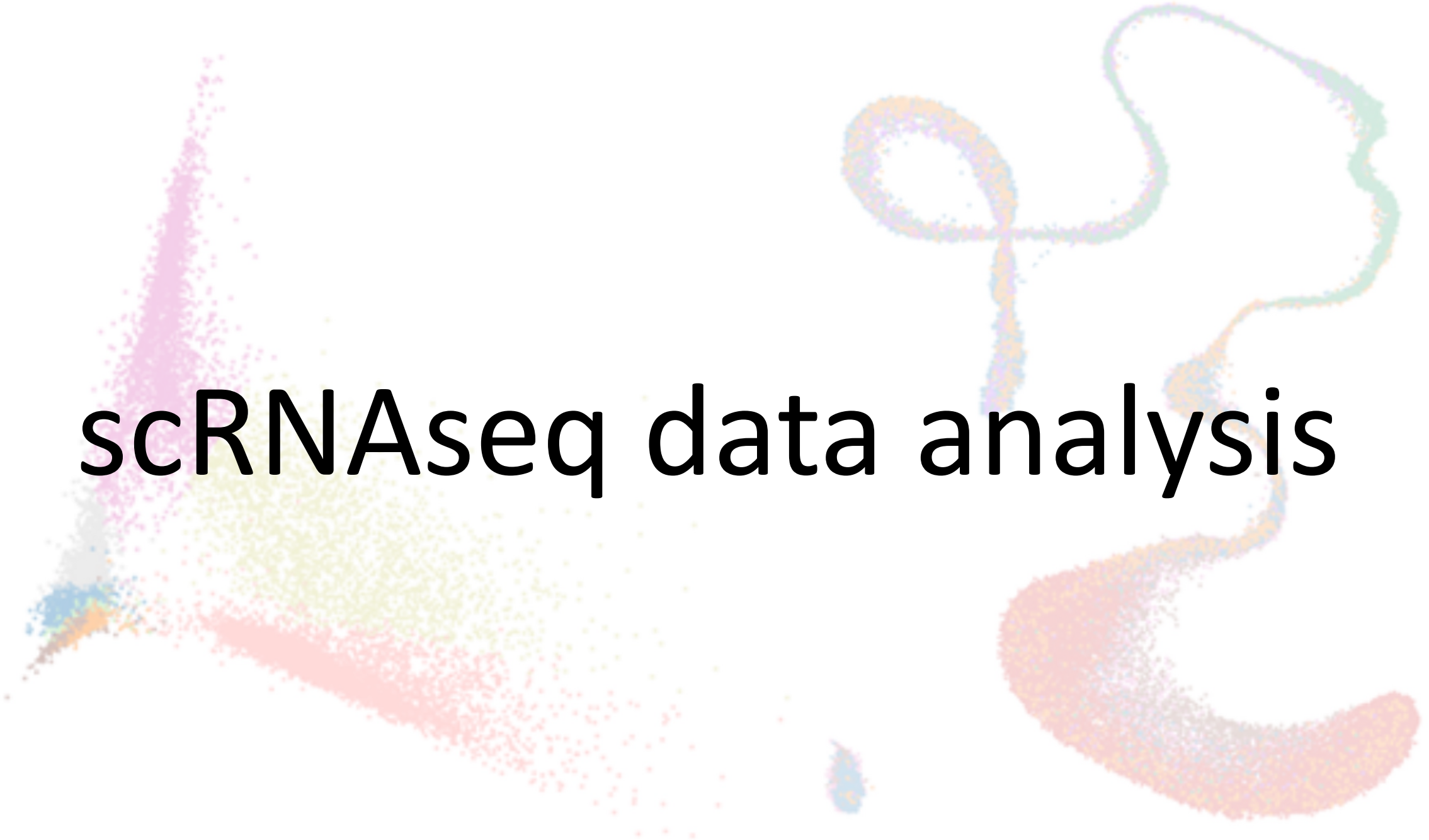



# Warning: Deduplication and UMI tags

Some technologies (most notably SMARTseq and SMARTseq 2) do not have UMI tags, but only cell-identifiers.

However, SMARTseq3 introduces the usage of UMIs.





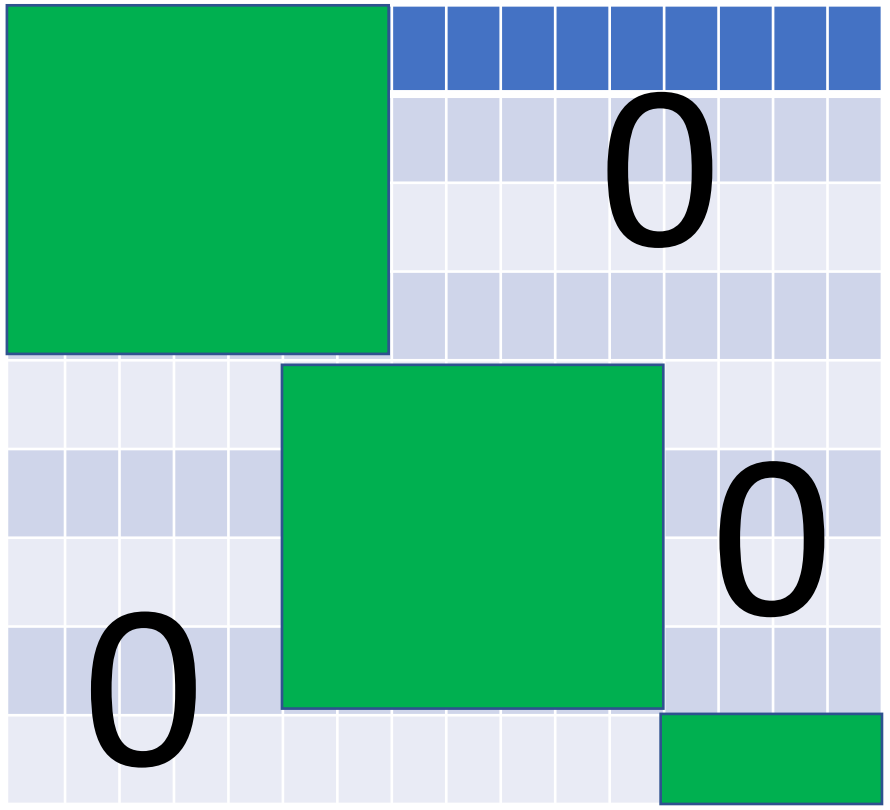
# scRNAseq data analysis

# Analysis: Data characteristics

An scRNAseq dataset can be composed by thousands of cells and genes.

Tipycally, this type of data is characterised by

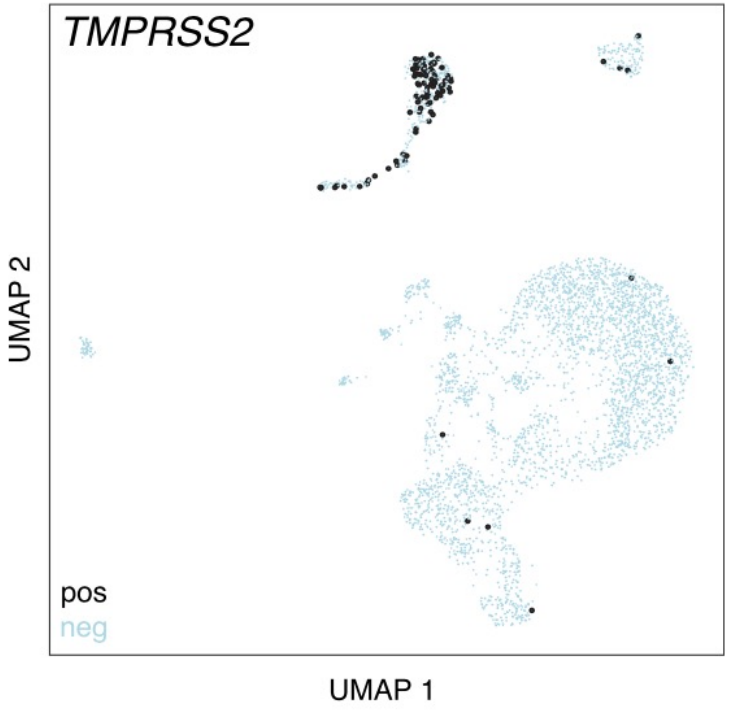
- **low capture rate** of transcripts per cell (2-10%)
- **Ambient RNA noise**
- **Empty droplets and doublets**
- **3’ RT bias (for 10X data)**
- **sparse data** (often >95% of the data matrix is zero)
- **Non-linear structure**
- **Genes expressed in modules**

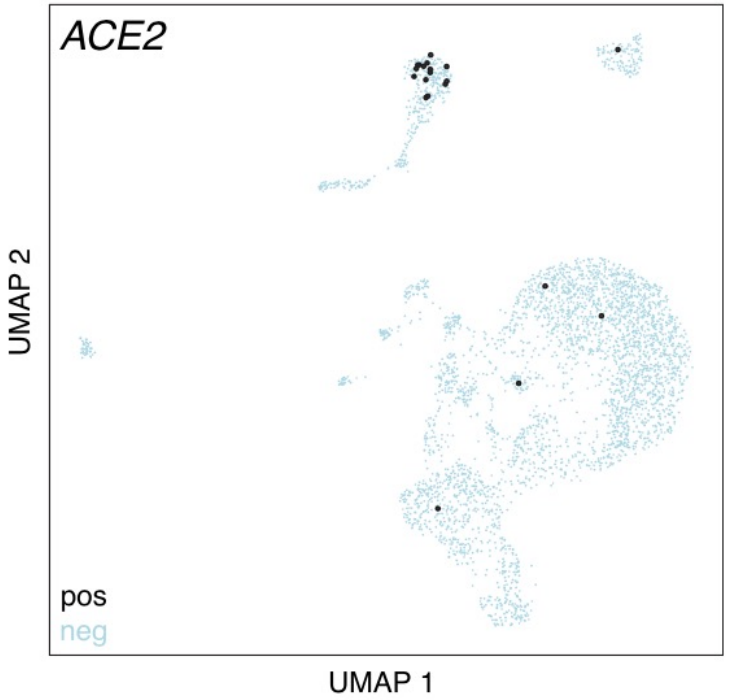


# Analysis: Data characteristics - Ziegler et al, 2020

The authors find that (page 3) non-human primate samples of lung express TMPRSS2 with ACE2 in only 3.8% of the cells.

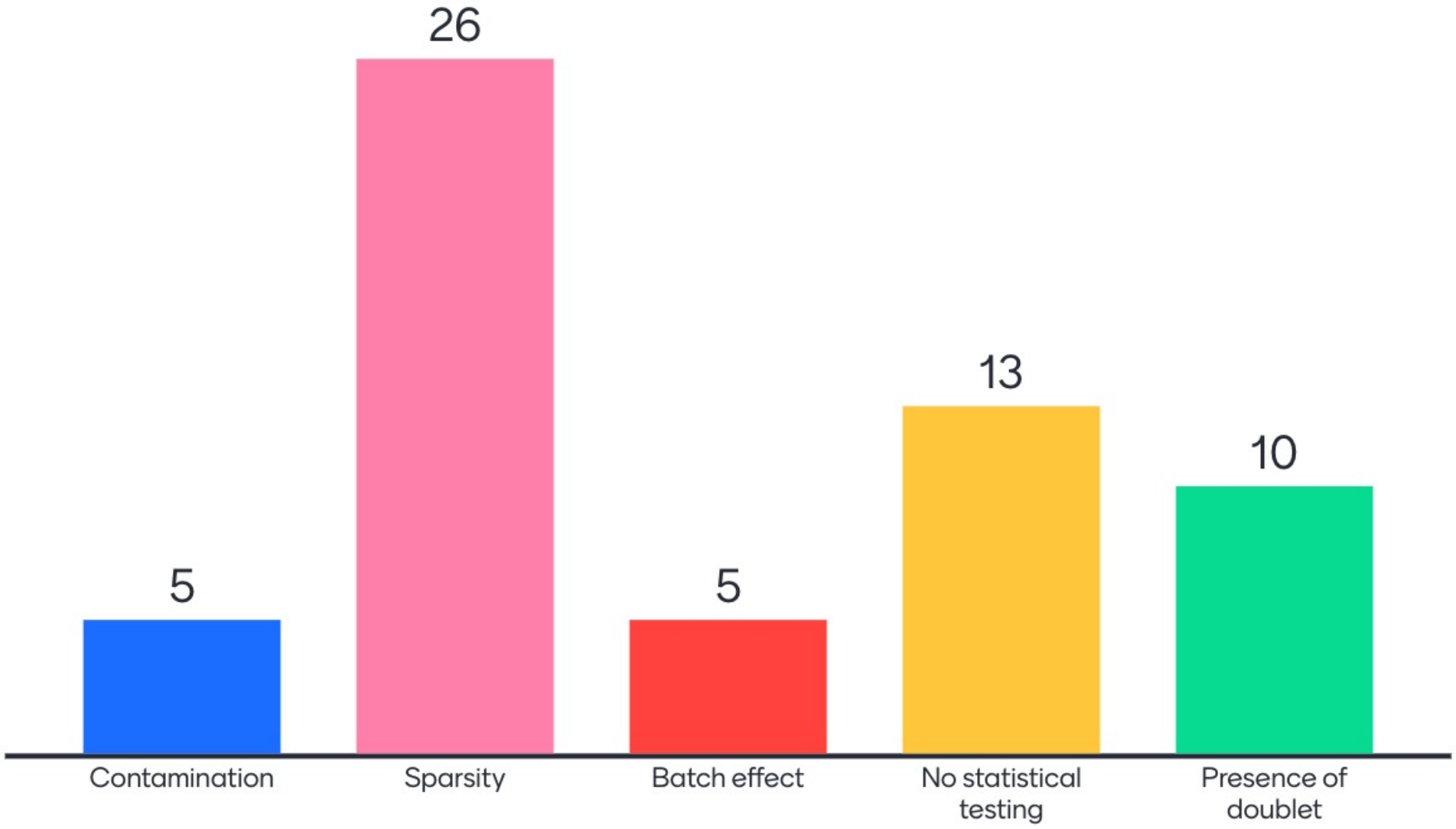
**Which characteristic of single cell data might create problems with such small cell subsets?**



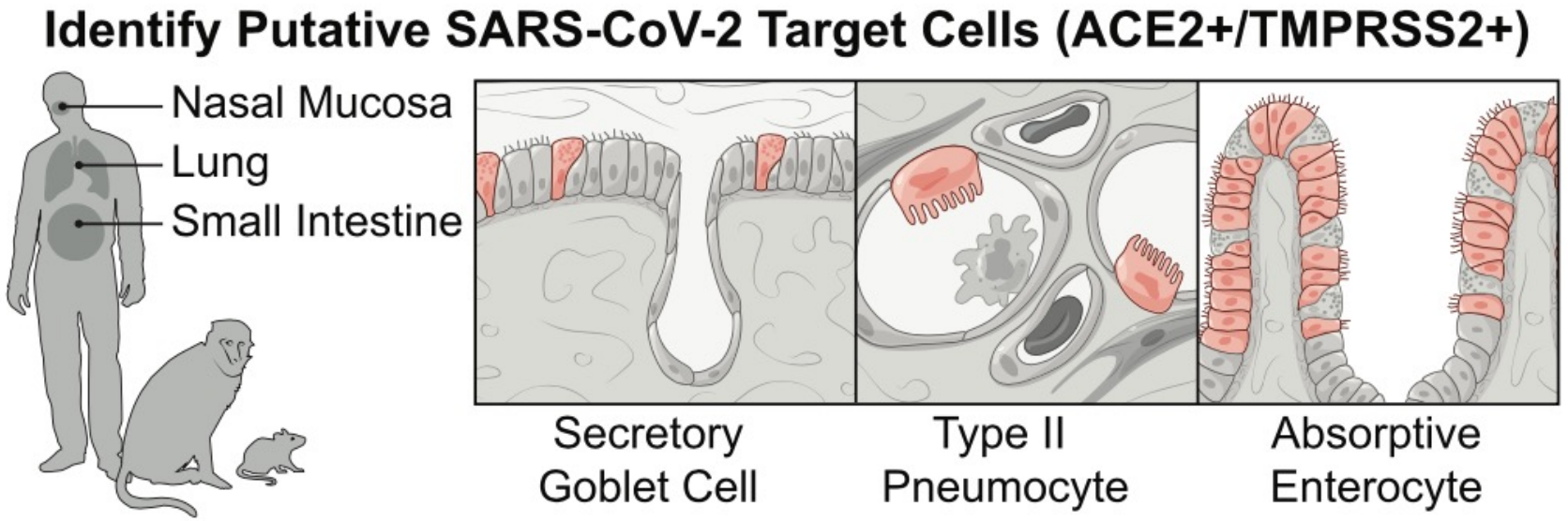




**Menti.com 68486264**

# Analysis: Data characteristics - Ziegler et al, 2020



# Analysis: Example dataset from Ziegler et al, 2020




Data from humans, N.H. primates, mice

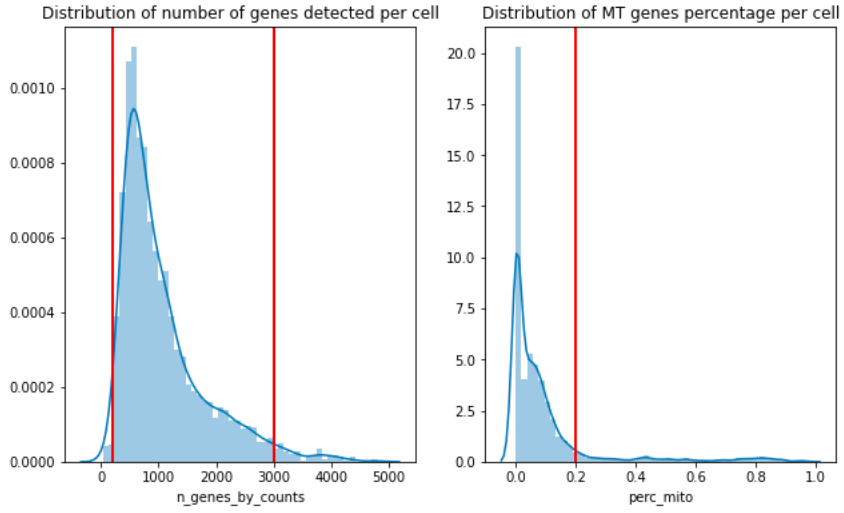
- Composed of >46K cells
- Circa 8 studies, 60 samples over 3 tissues and 2 health conditions
- Batch effect to be expected

# Analysis workflow: preprocessing

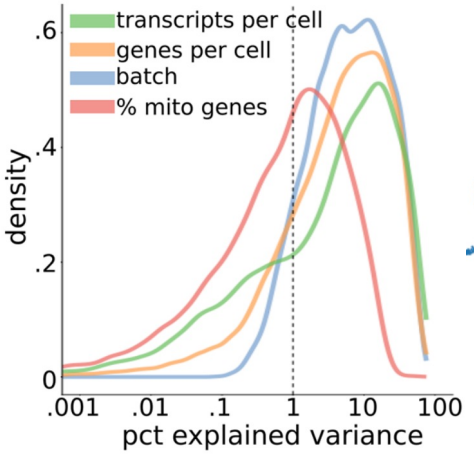
Raw matrix

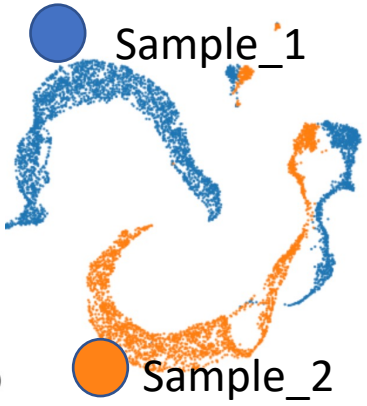
| | | |
|---|---|---|
| 20 | 24 | 60 |
| 40 | 50 | 120 |
| 10 | 16 | 30 |
| 5 | 0 | 5 |

Identify cell outliers through quality measures

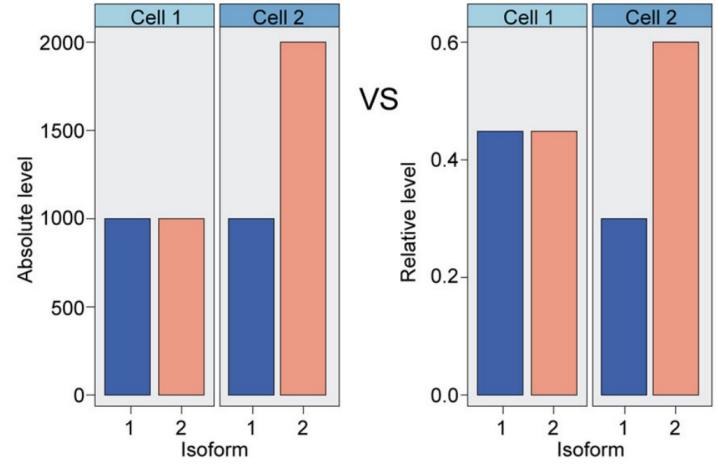



Identify technical/biological biases



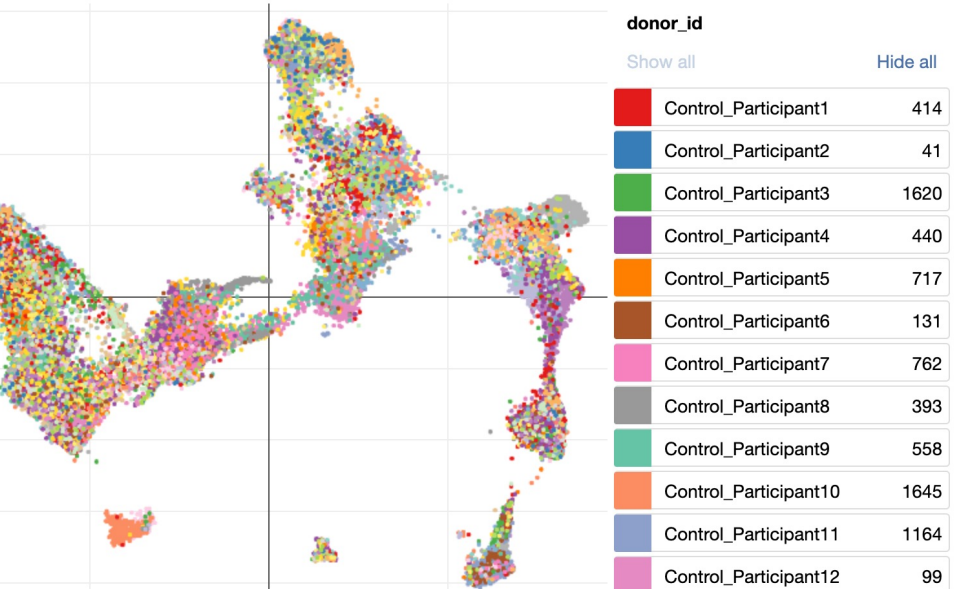




Normalize transcript counts and remove biases



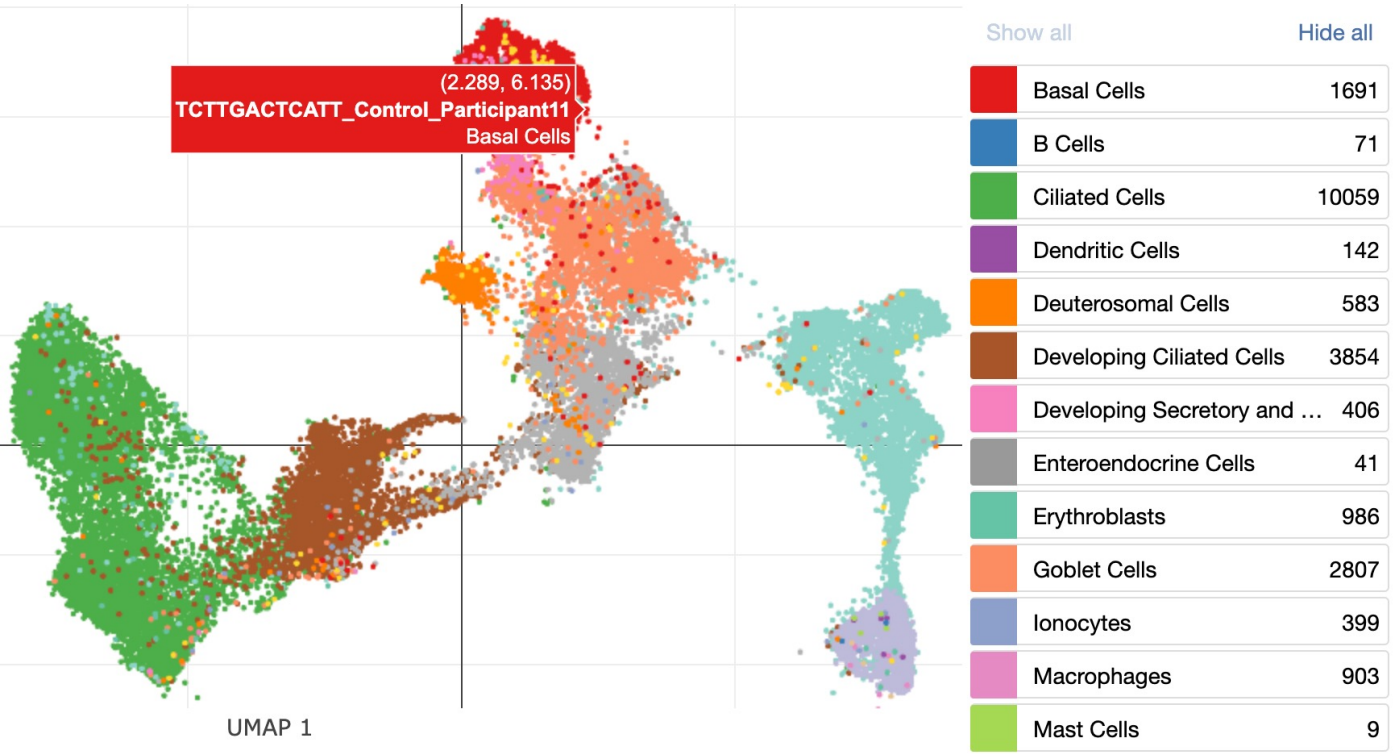

Doublet removal

[Dimensionality reduction]

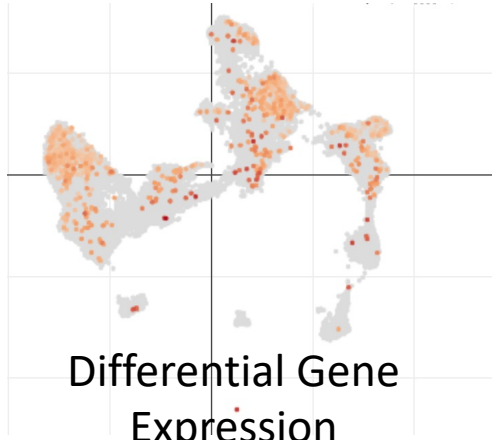


# Analysis workflow: cell-wise analysis

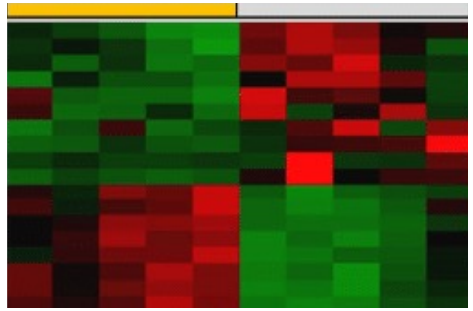
Clustering/subclustering and cell identification



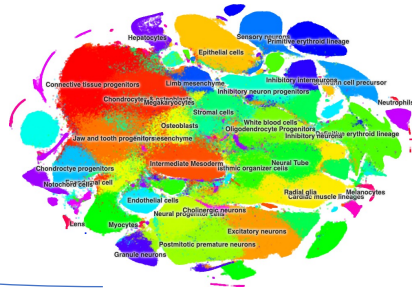

Check of Markers for cell types



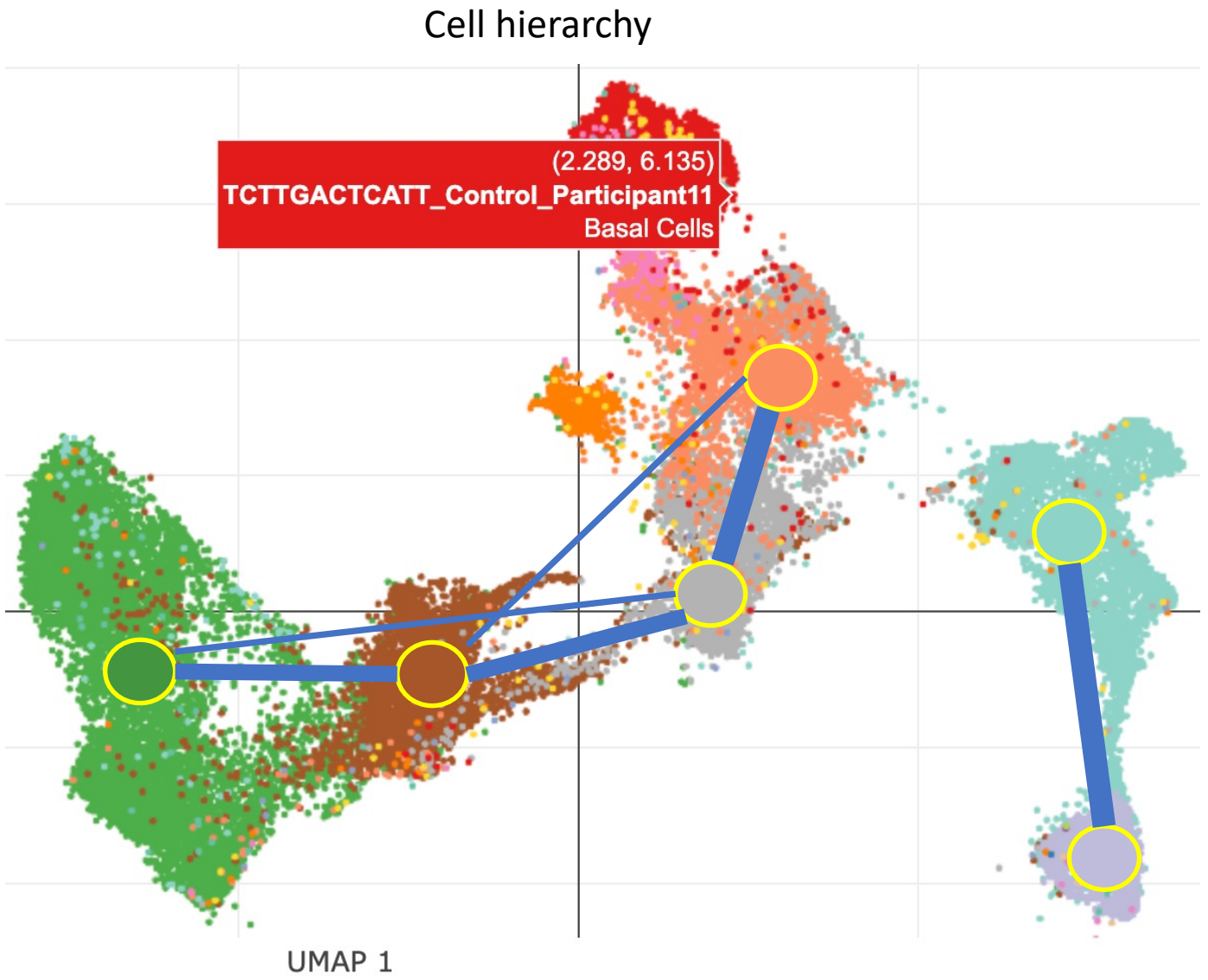
Differential Gene Expression



Annotated atlases

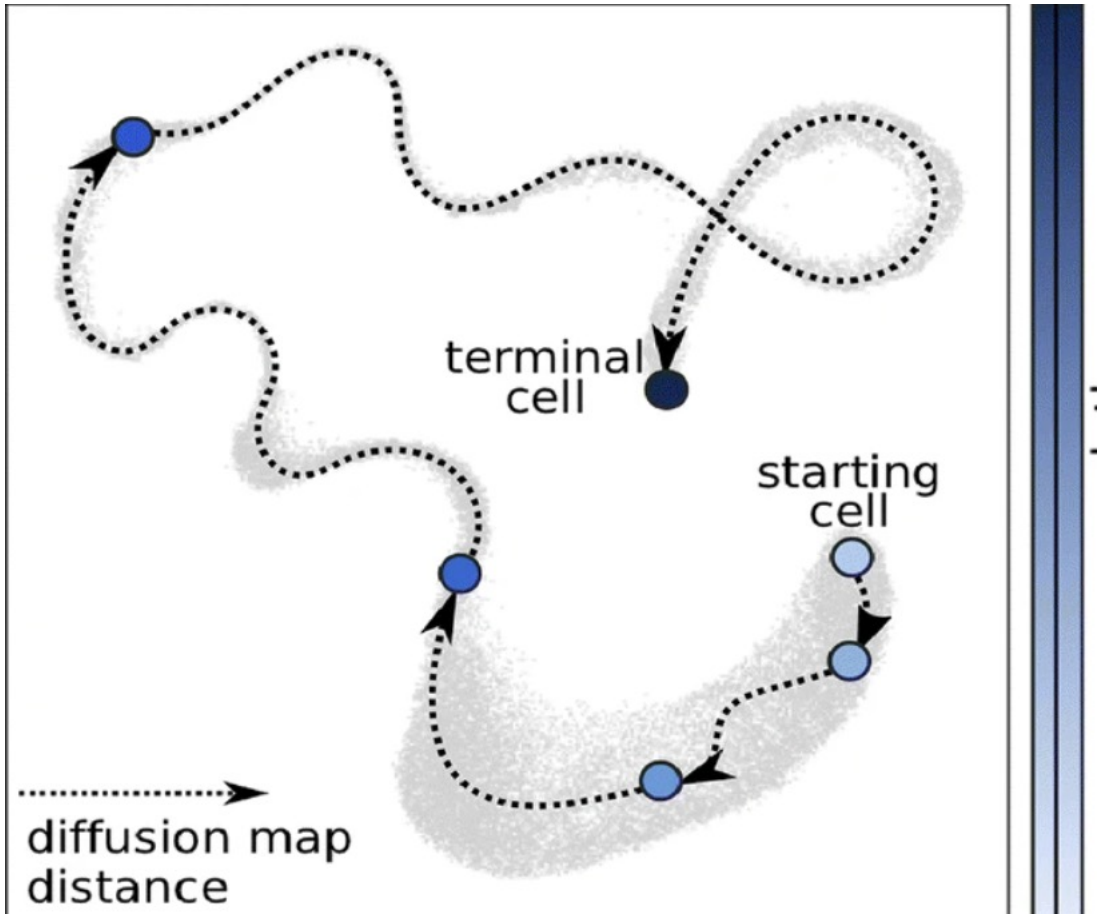


# Analysis workflow: cell-wise analysis

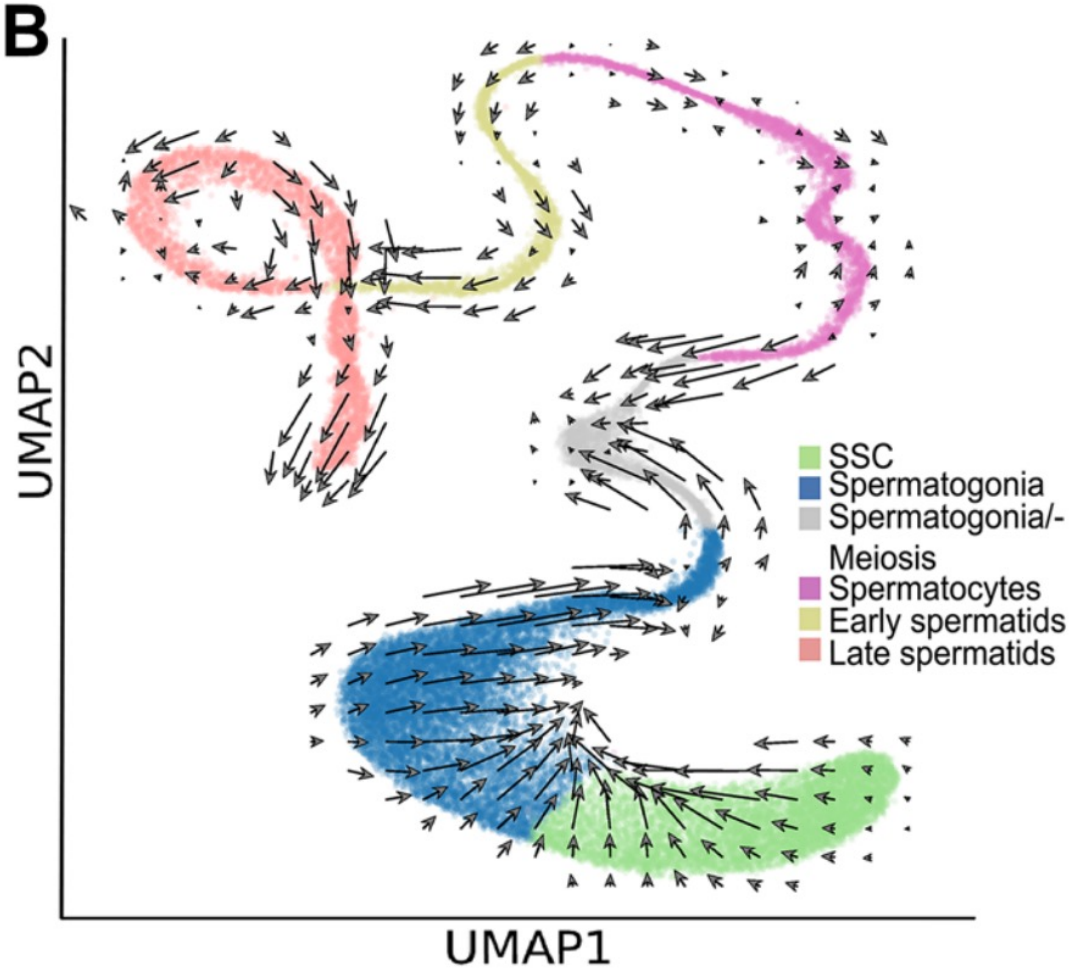


# Analysis workflow: cell-wise analysis

Pseudotimes and Detection of cell fates




Prediction of molecular states



# Analysis workflow: cell-wise analysis

**Can you identify which types of cell-wise analysis have been used in the paper?**

**Menti.com 68486264**

# Analysis workflow: cell-wise analysis

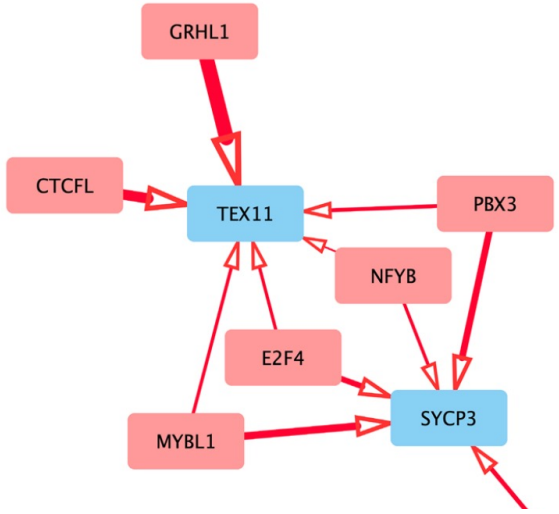
## Can you identify which types of cell-wise analysis have been used in this paper?

Mentimeter

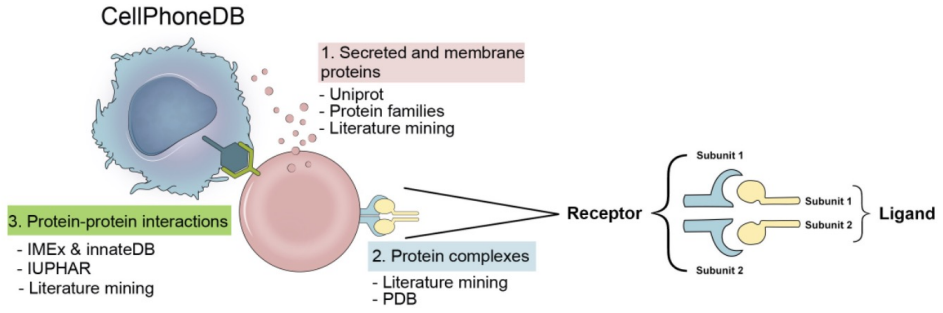
- DGE
- Clustering
- uMAP for clustering of populations.
- DGE
- Cell subset identification
- clustering + classification & gene expression

# Analysis workflow: gene-wise analysis

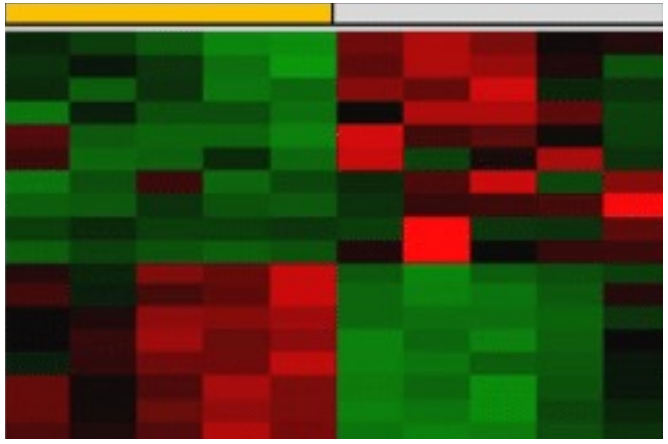
## TF-Gene Networks



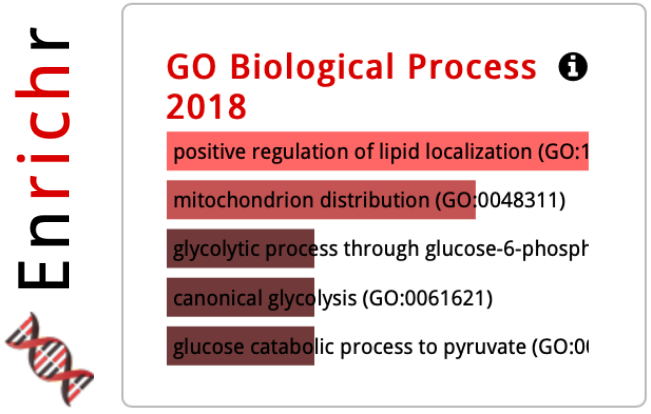

## Receptor ligand interactions



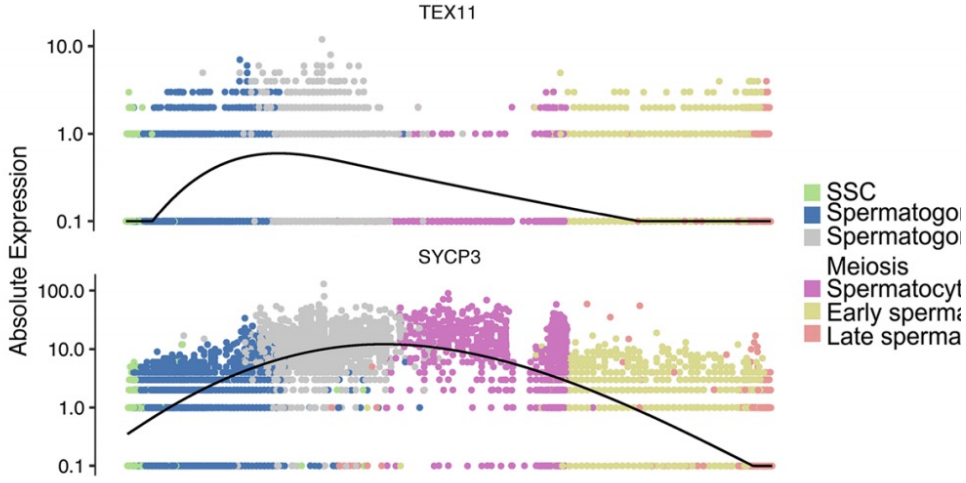

## Differential Gene expression



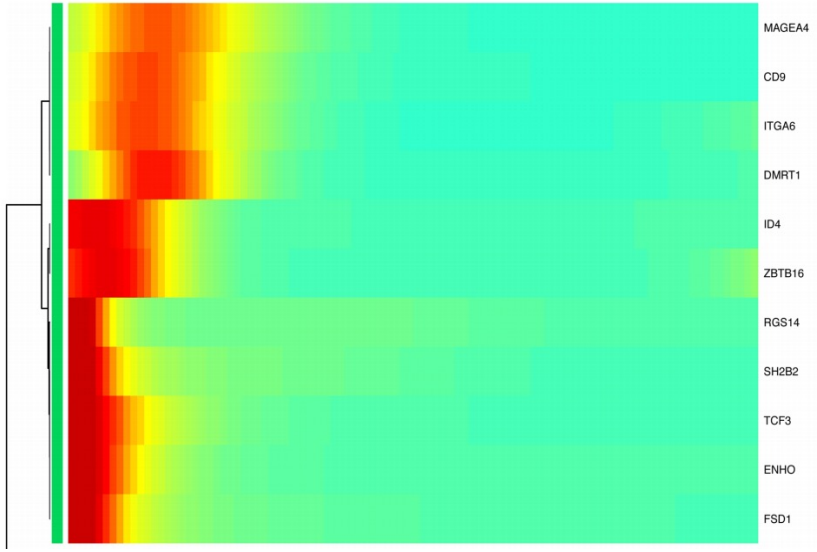
## Gene Enrichment Analysis




## Modeling genes trajectories in pseudotime




## Clustering of coexpressed genes

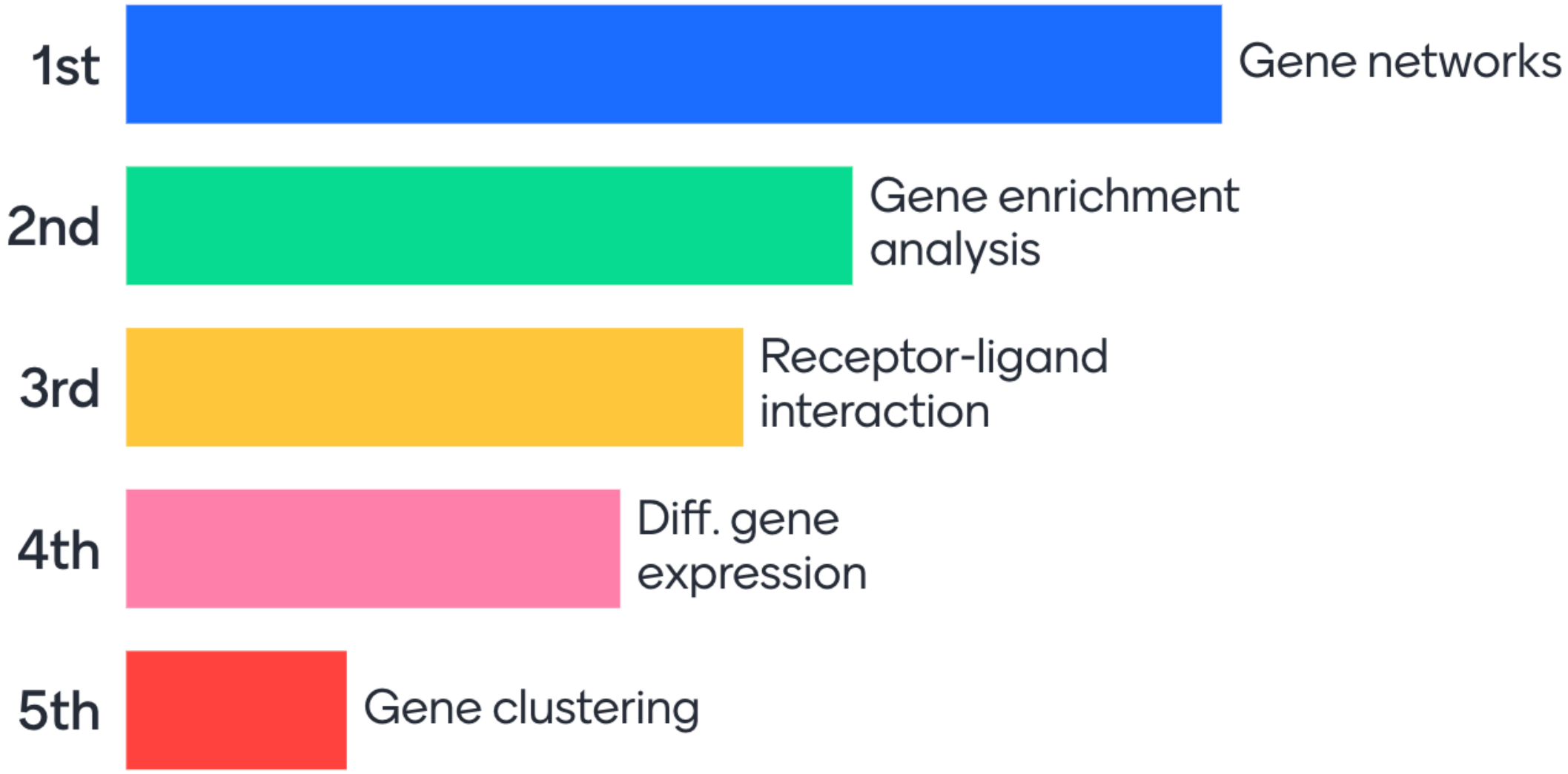


# Analysis workflow: gene-wise analysis

**The authors of the paper look only for co-transcribed genes of ACE2, but more advanced gene-wise analysis could be performed. Which would you choose?**

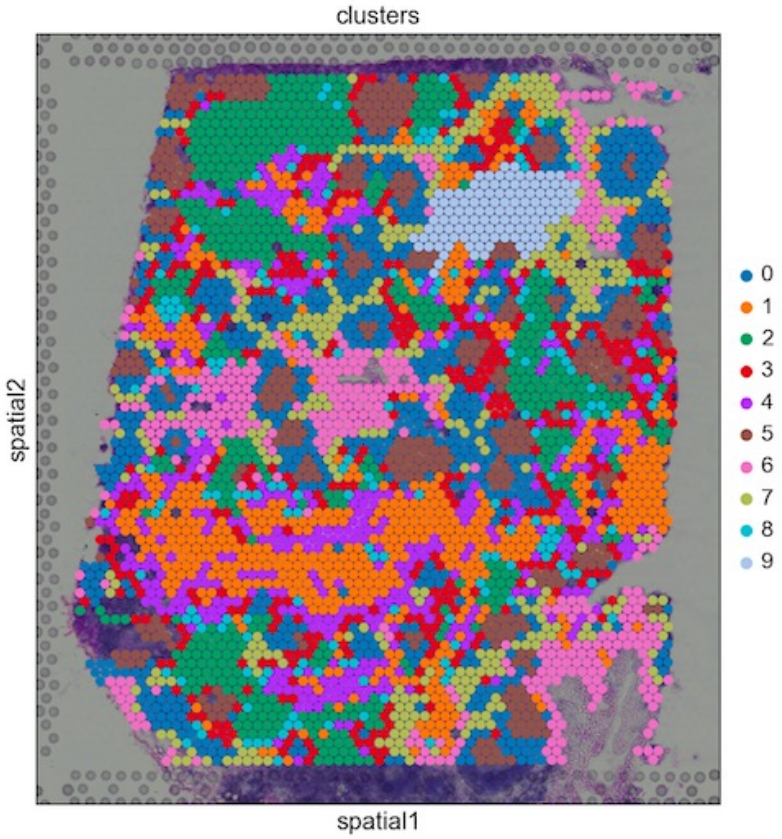
**Menti.com 43555600**

# Analysis workflow: gene-wise analysis



# Analysis workflow: spatial RNA analysis

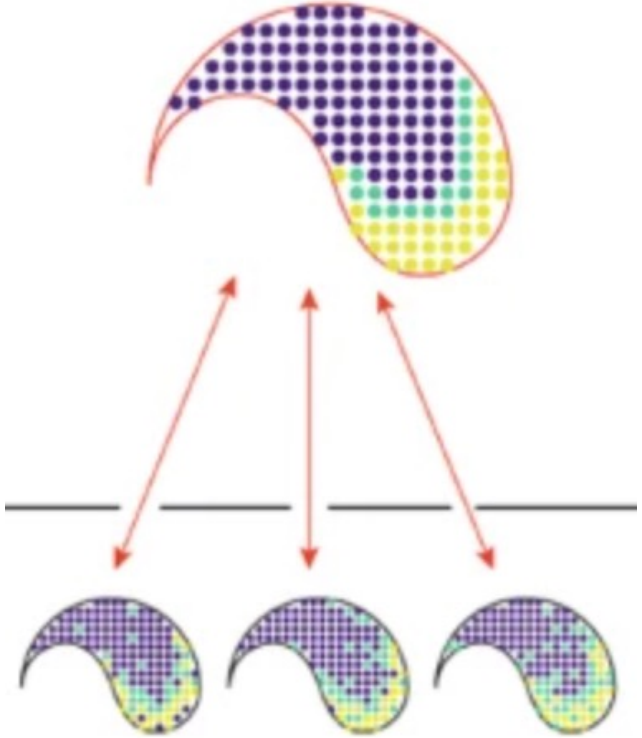
- Relationship between cell location and cell types
- Clustering based on both transcriptome and physical position




- Visualization of stacked slices
- Distances and clustering across samples

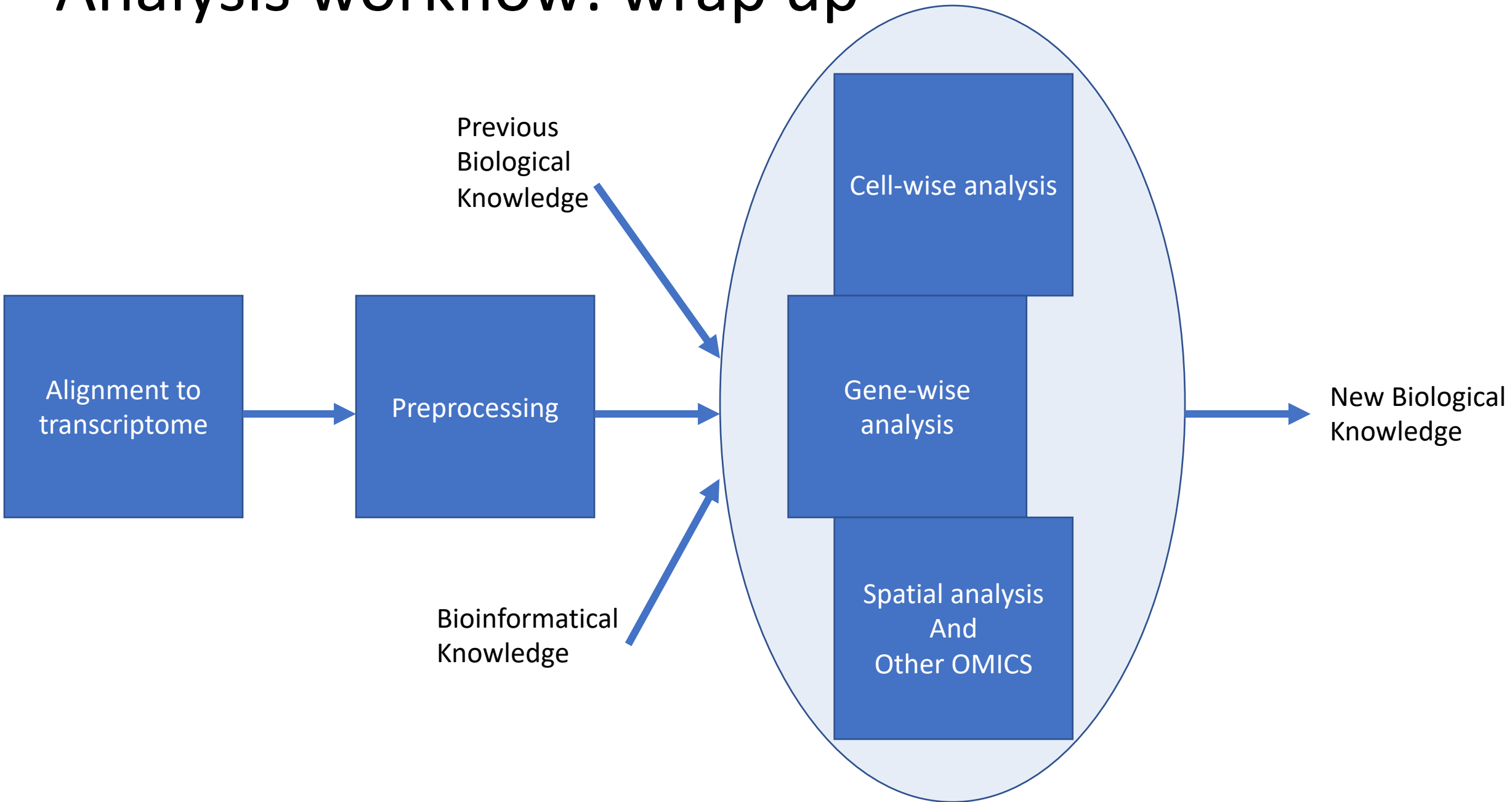
e.g.: Larsson et al 2021

- Diff. expression across physical space regions



e.g.: Svensson et al 2020

# Analysis workflow: wrap up



# Conclusions

- scRNAseq data is sparse and can be noisy
- However, it allows to explore detailed biological aspects of a tissue
- There is not yet a standardised way of analyzing this data
- There are >300 analysis tools and new methods and tech advancements
- Scanpy (python) and Seurat (R) have emerged as reliable analysis tools for many standard operations. Scanpy+Squidpy perform spatial OMICS.
- Useful acquiring some programming expertise in using both R and python packages

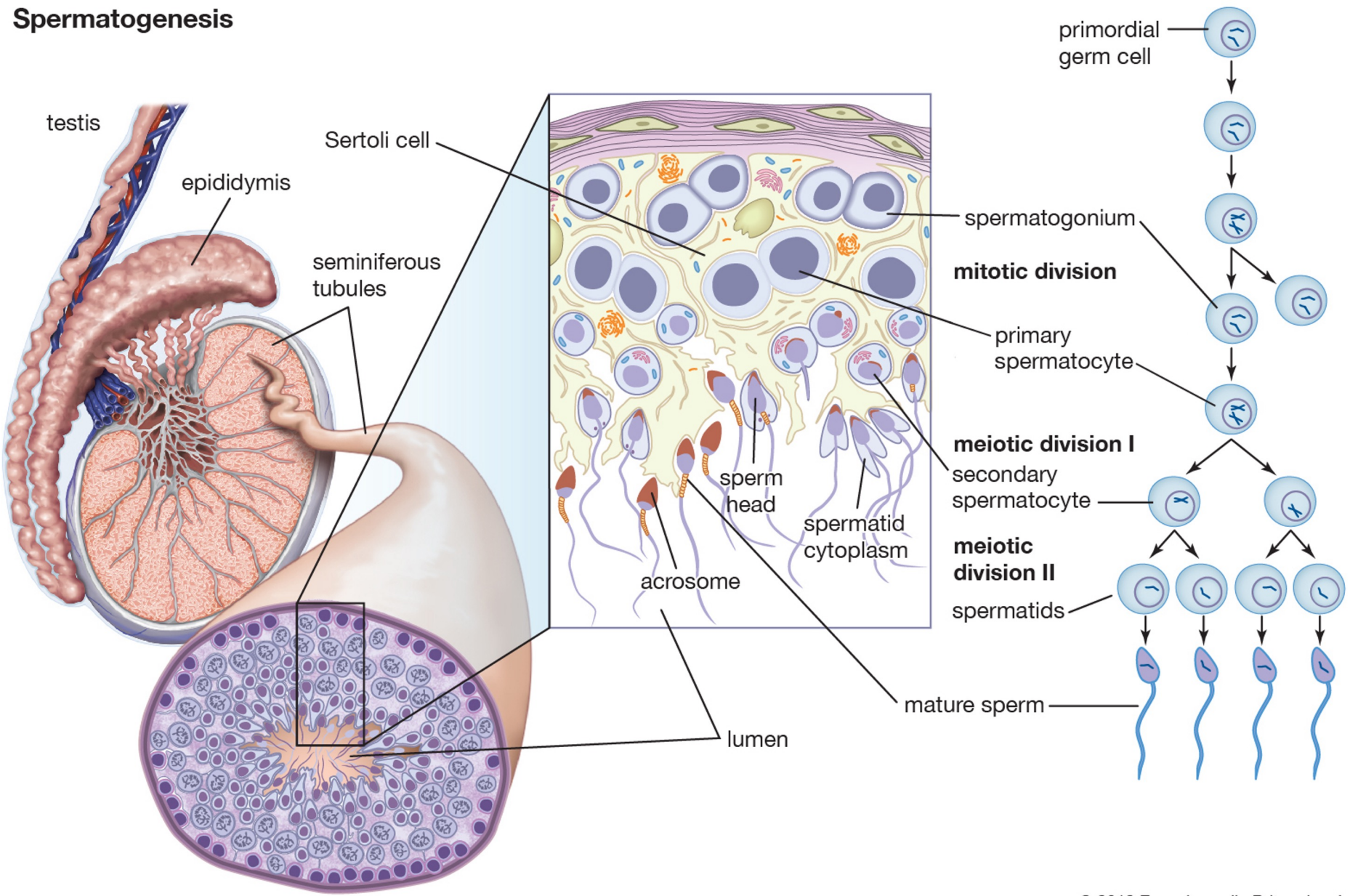
# Useful links

- The home page of Scanpy. This is a python tool. Here you have a lot of tutorials to try out some more of the single cell data analysis of this presentation. I personally suggest scanpy as your standard analysis tool.
- The home page of Seurat. This is an R tool. It is quite as good as Scanpy, but not as open and efficient, and contains less new tools than Scanpy. It can be interfaced with R Bioconductor packages quite easily.
- A list of many single cell tools and their scope

***Questions***

***Pre-exercise Break***

# Exercise dataset – human testes



# Exercise dataset – human testes

Human testes data:

- Composed of 65K cells
- Around 19K cells are outliers
- Data comes from 13 different datasets