Single cell RNA seq Data and Analysis

Samuele Soraggi

Bioinformatics Research Center,
Aarhus University

NGS 2020 BiRC summer course

BiRC

Content/Objectives:

time

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

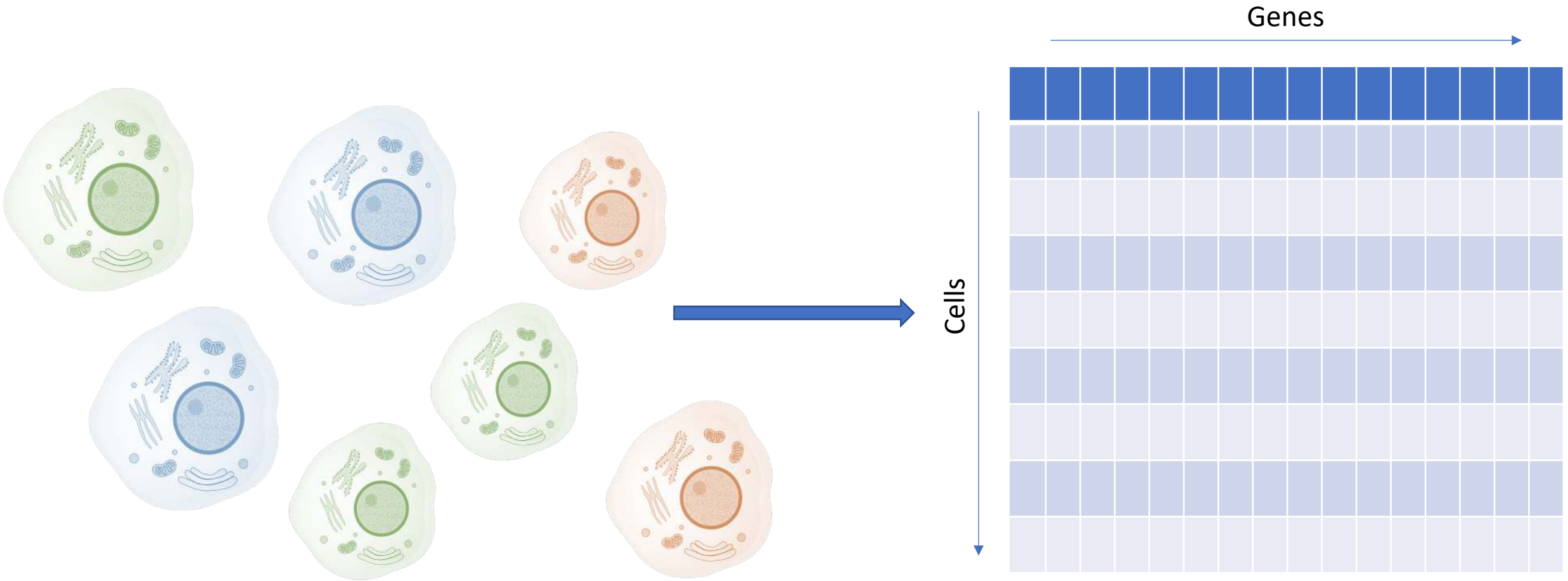
- Analysis of single cell data

- Conclusions and introduction to paper discussion

Single cell RNA seq data (scRNAseq)

Single cell data: what is it?

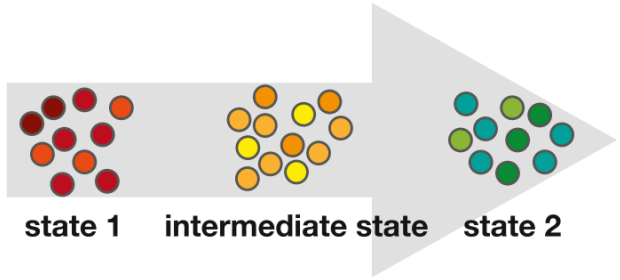
In a single cell RNA sequencing dataset (scRNAseq), we can capture the mRNA transcripts of each cell, and count how many of them are associated to a set of reference genes.

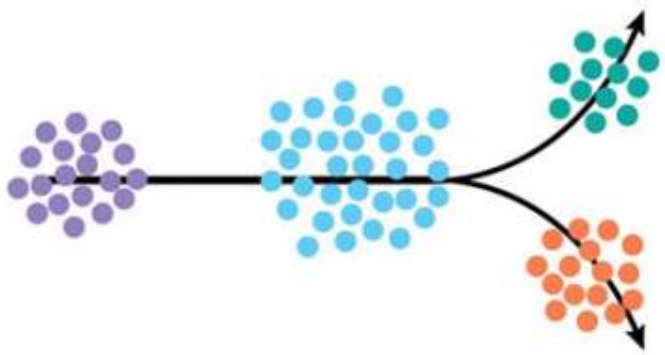


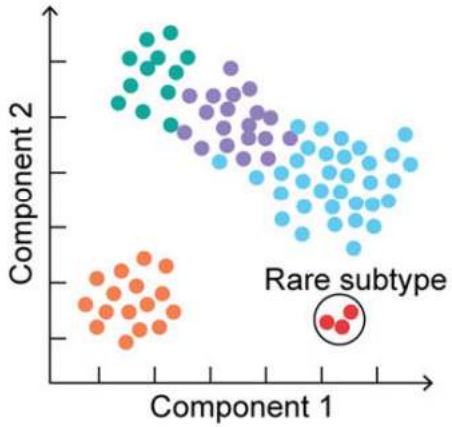
Single cell data: what is it? Why do we use it?

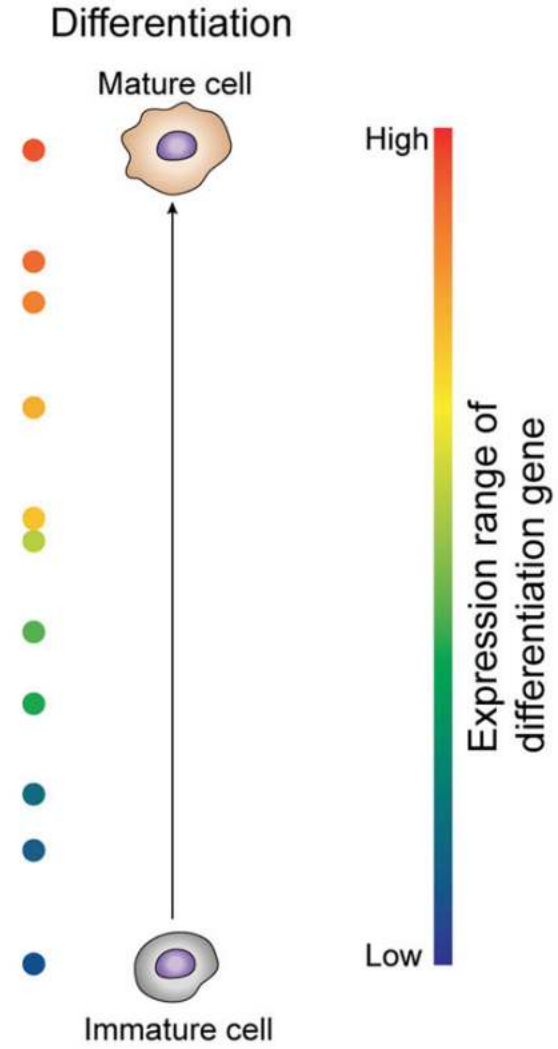
scRNAseq data gives the possibility of analyzing gene transcription at cellular level to retrieve information about

- Differentiation and fate of cells
- Cell hierarchy
- Cell response to stimuli
- Co-expression of genes
- Rare cell types
- Gradual changes in transcription rate
- ……





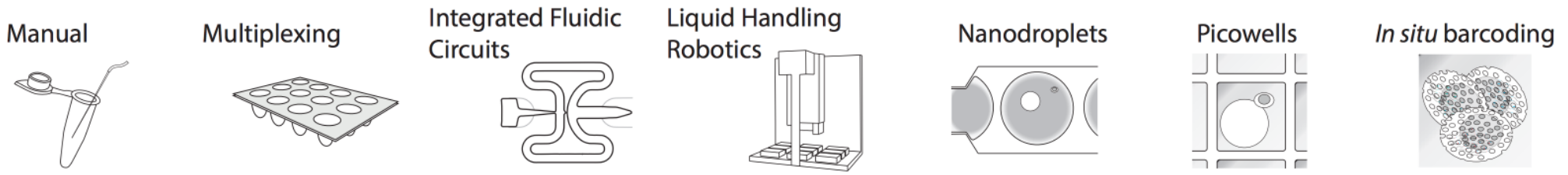


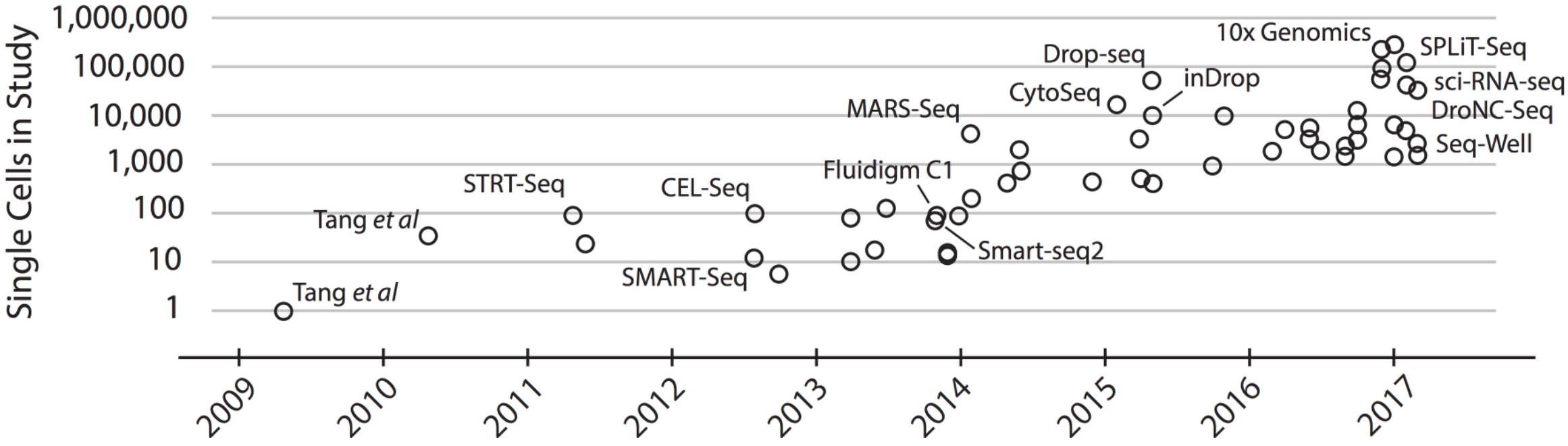


Single cell data: what is it? Why do we use it?

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

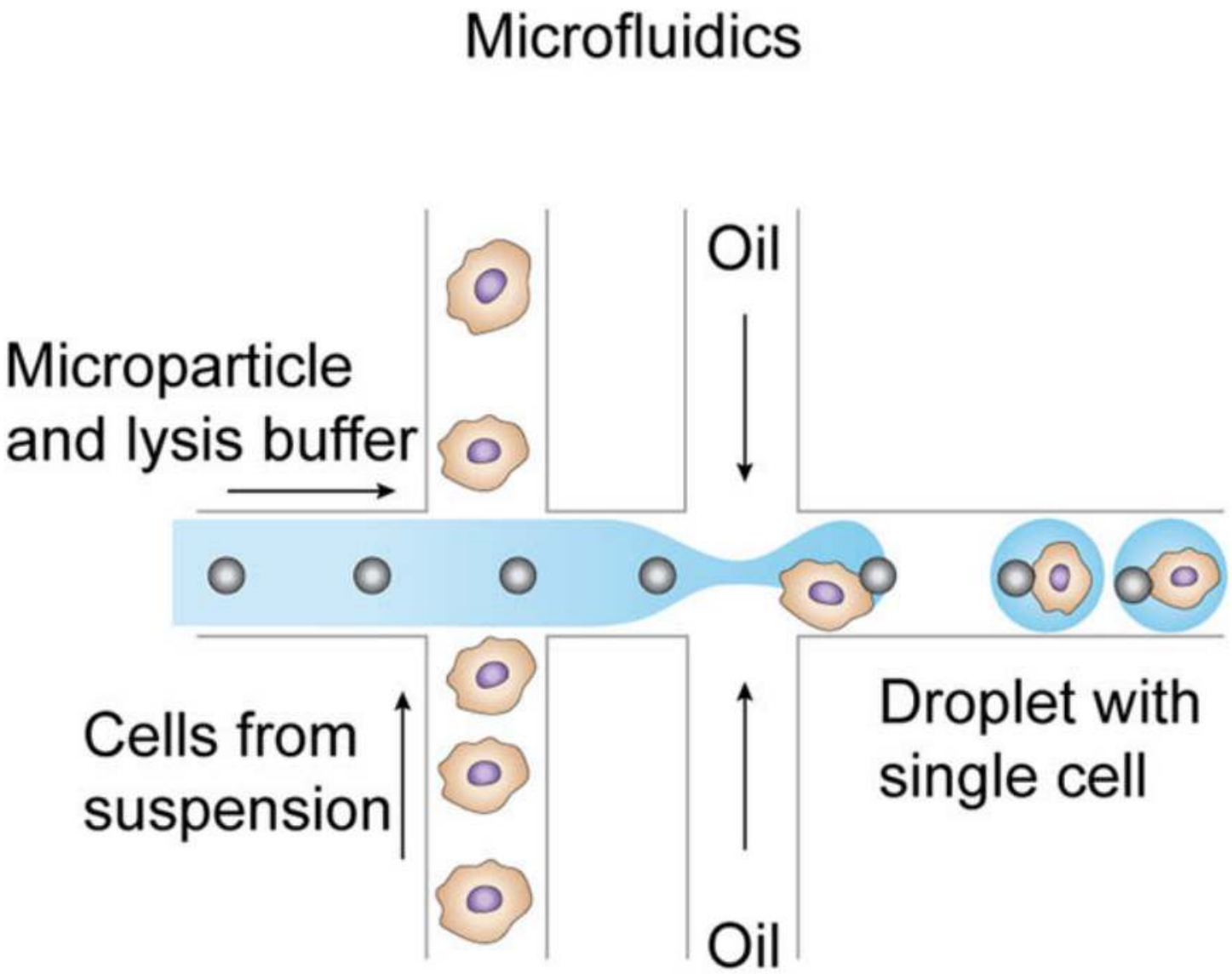
Single cell data: a recent sequencing technology

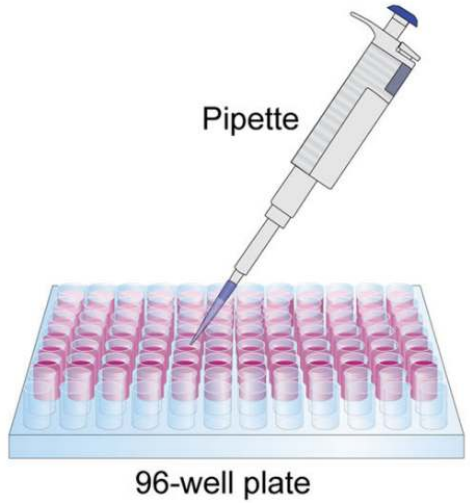


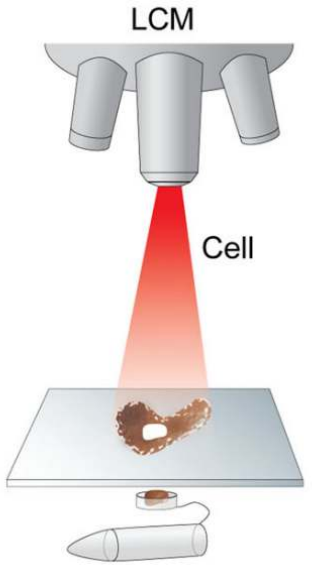


Source: Svensson et al 2017

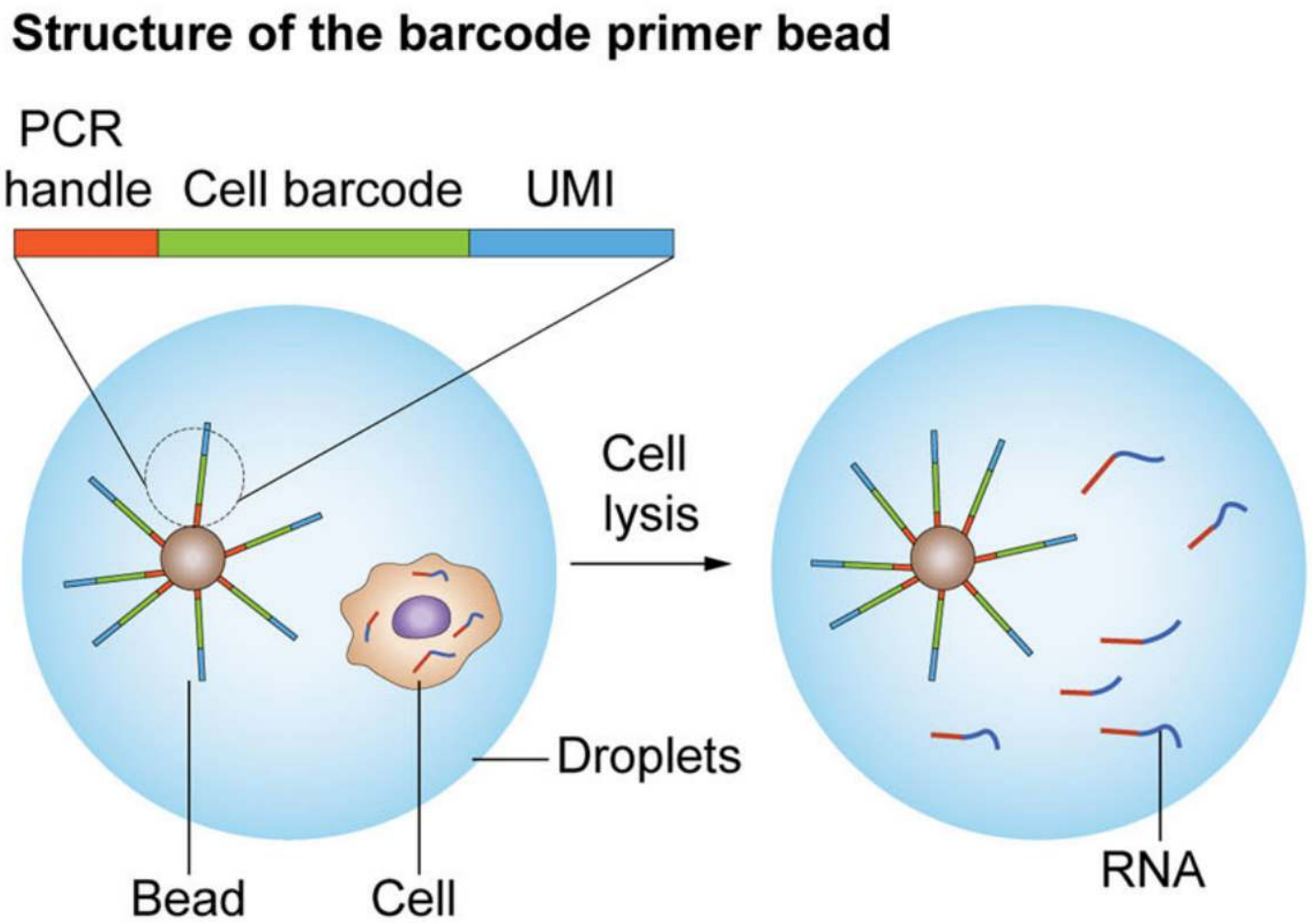
The sequencing framework: cell isolation





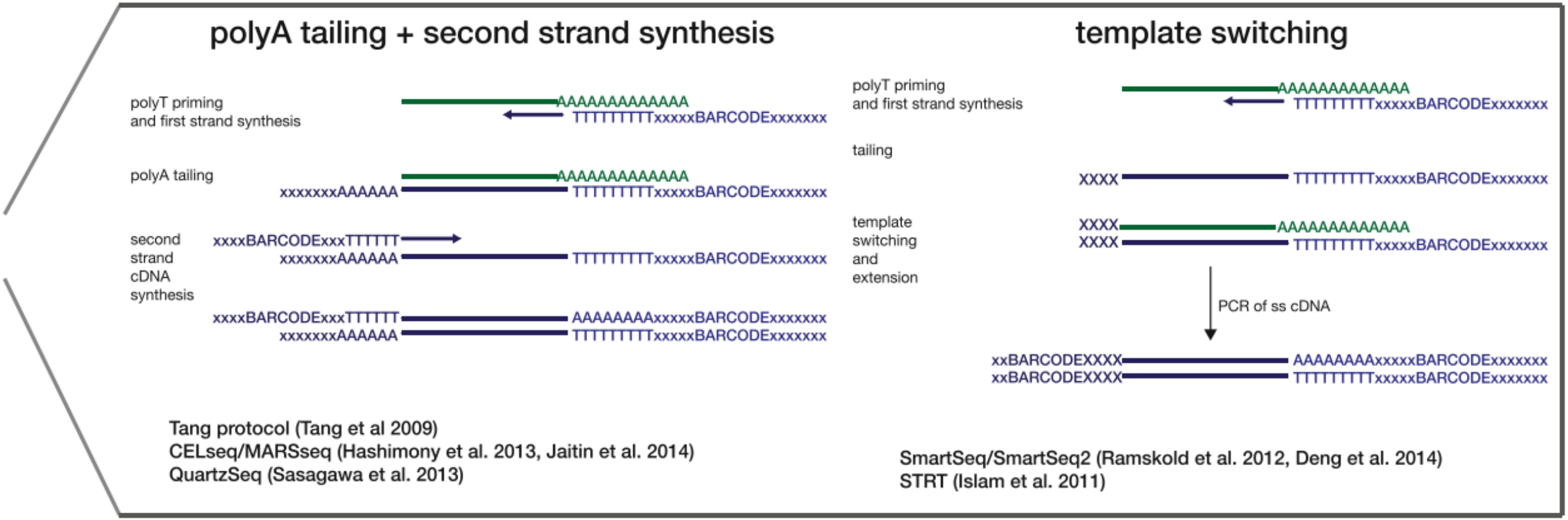


The sequencing framework: mRNA capture



The sequencing framework: cDNA RT

Reverse transcription into cDNA

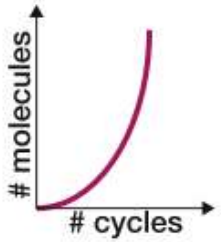


The sequencing framework: NGS sequencing

Preamplification and NGS sequencing

PCR

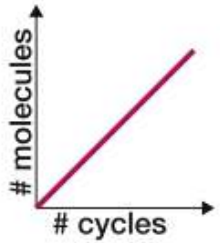
- exponential amplification
- PCR base specific biases



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)

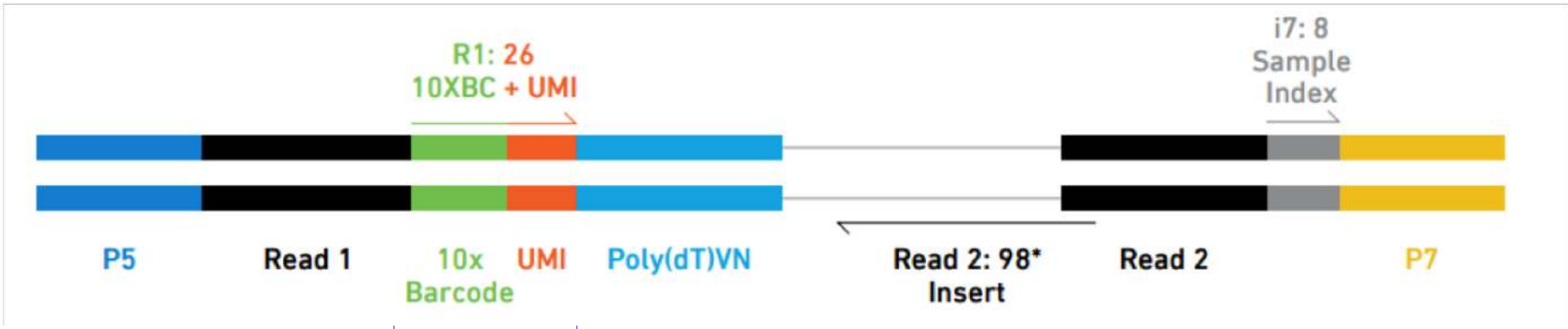
Illumina

AB SOLID

PacBio

The sequencing framework: raw reads

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

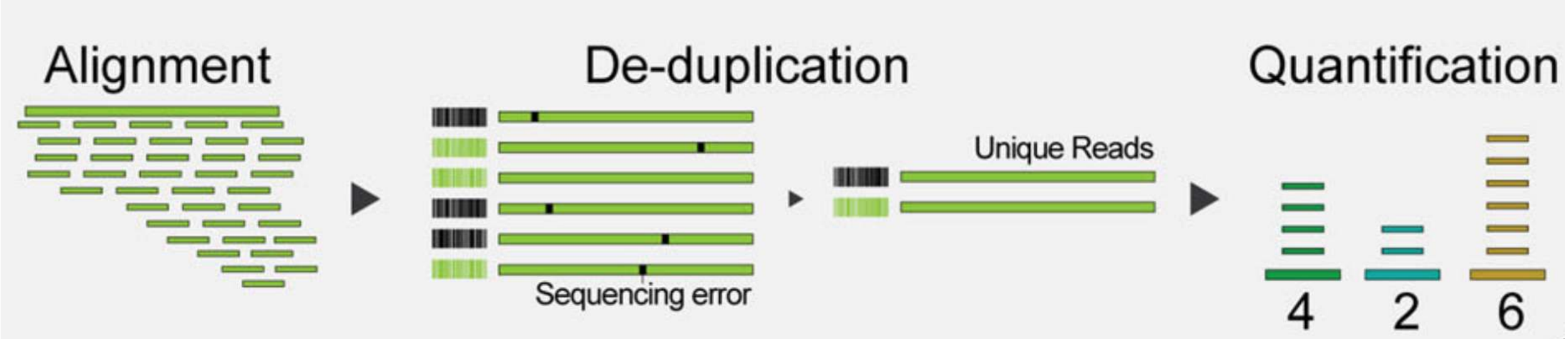



```
@SRR8363305.1 1 length=26
NTGAAGTGTTAAGACAAGCGTGAACT
+SRR8363305.1 1 length=26
# AAFFJJJJJJJJJJJJJJJJFJJJJ
```

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

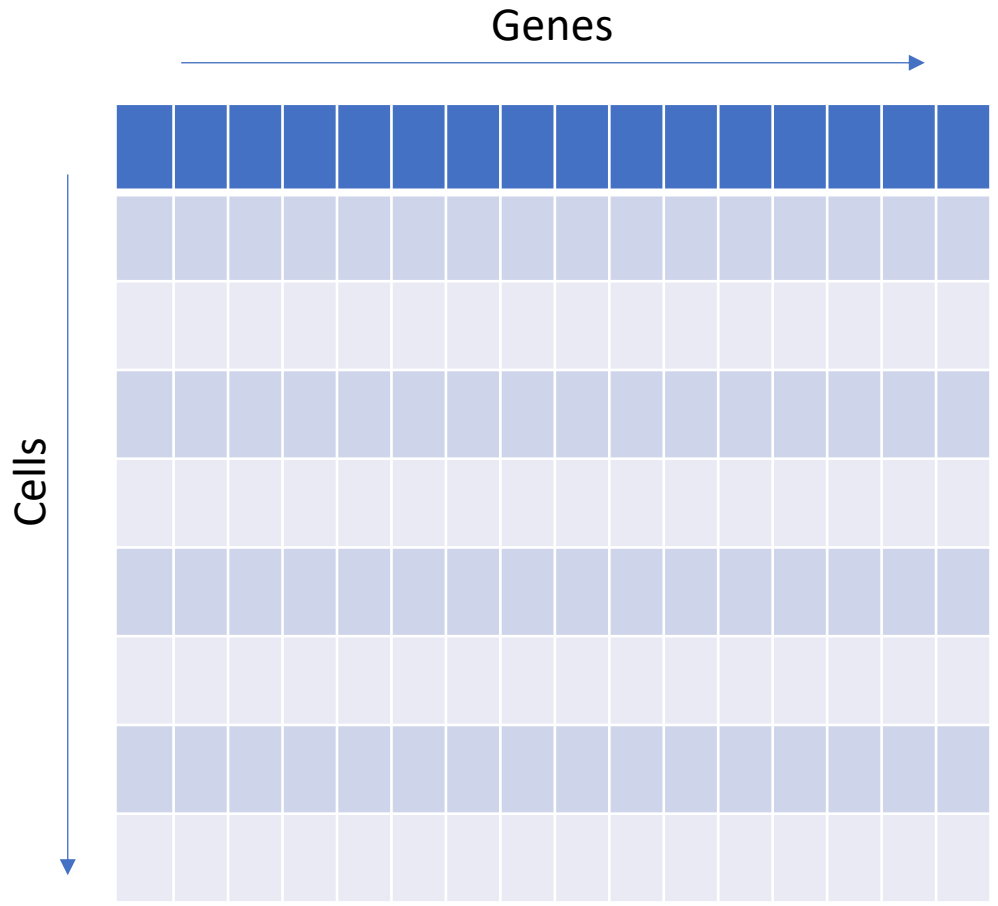
The (post)sequencing framework: align & quantify

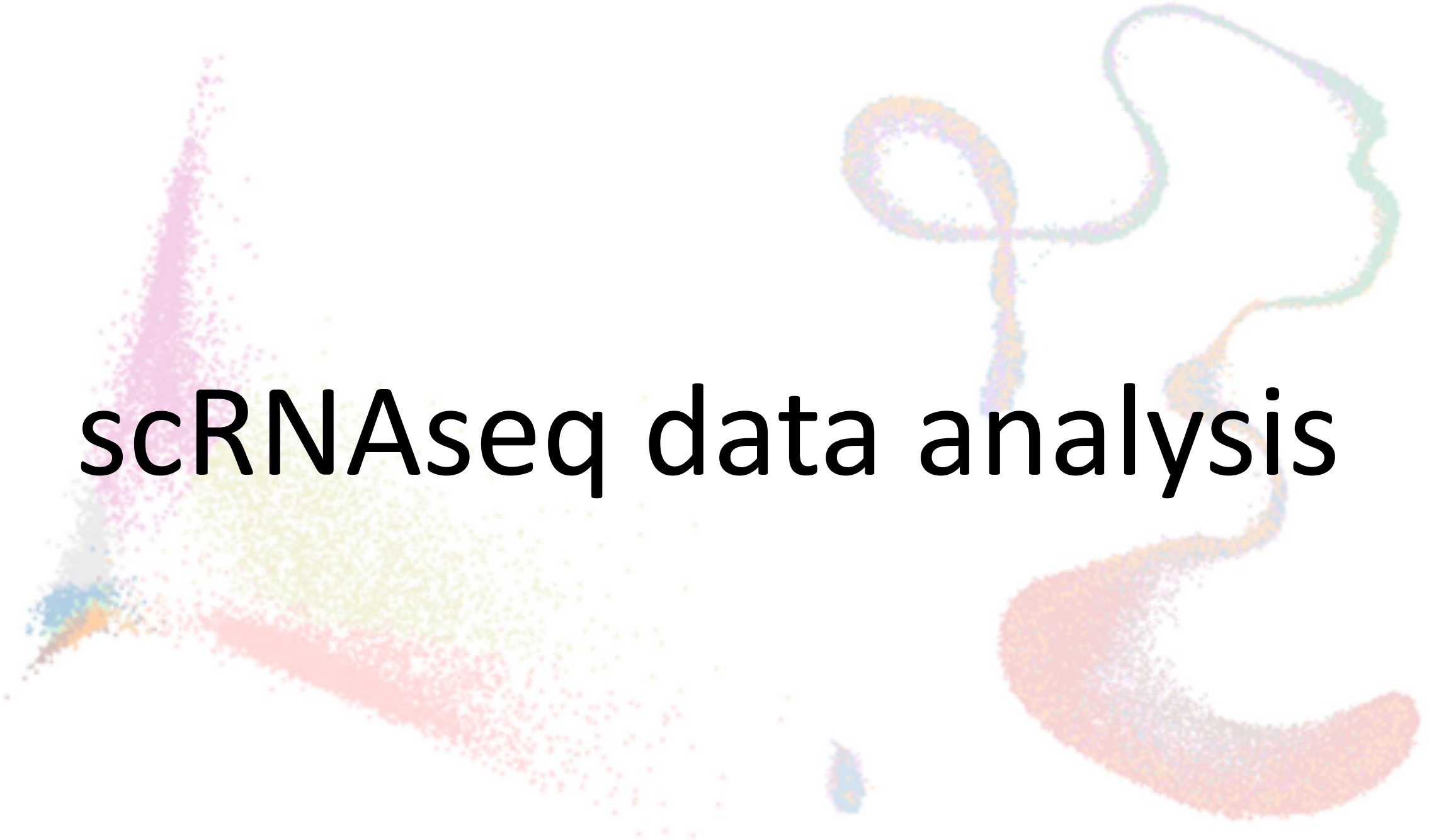
The raw fastq reads are aligned to a transcriptome to identify which reads pertain to specific genes.
The aligner has to be splice-aware and correct for sequencing errors happening in the barcodes and cDNA.
Lastly, UMIs are used to detect matching transcripts and collapse them into one if they come from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





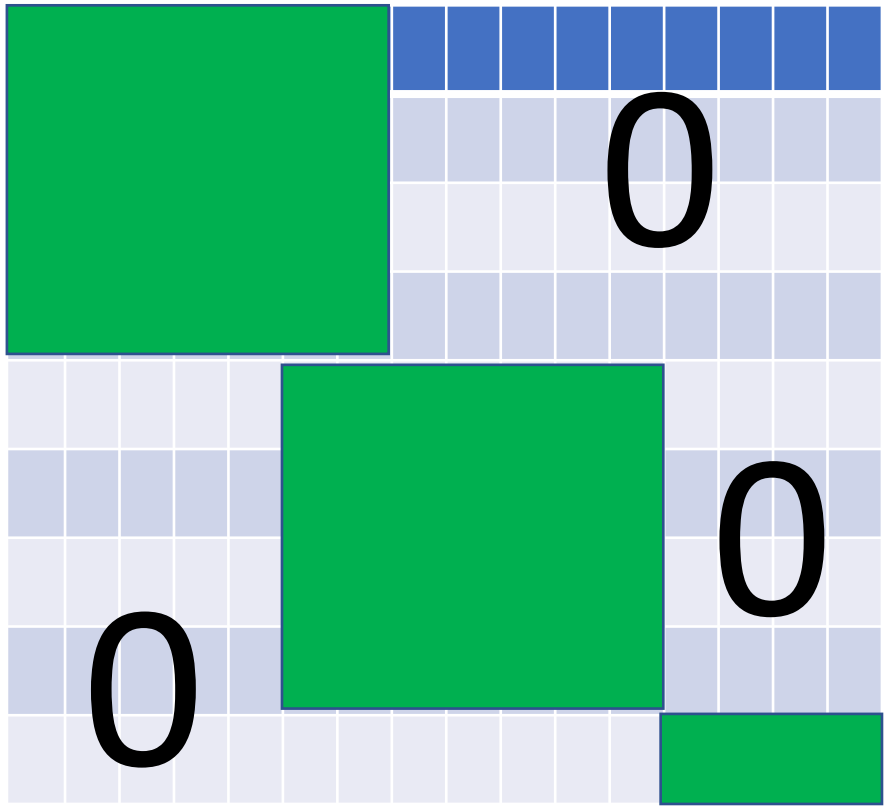
scRNAseq data analysis

Analysis: characteristics of the data

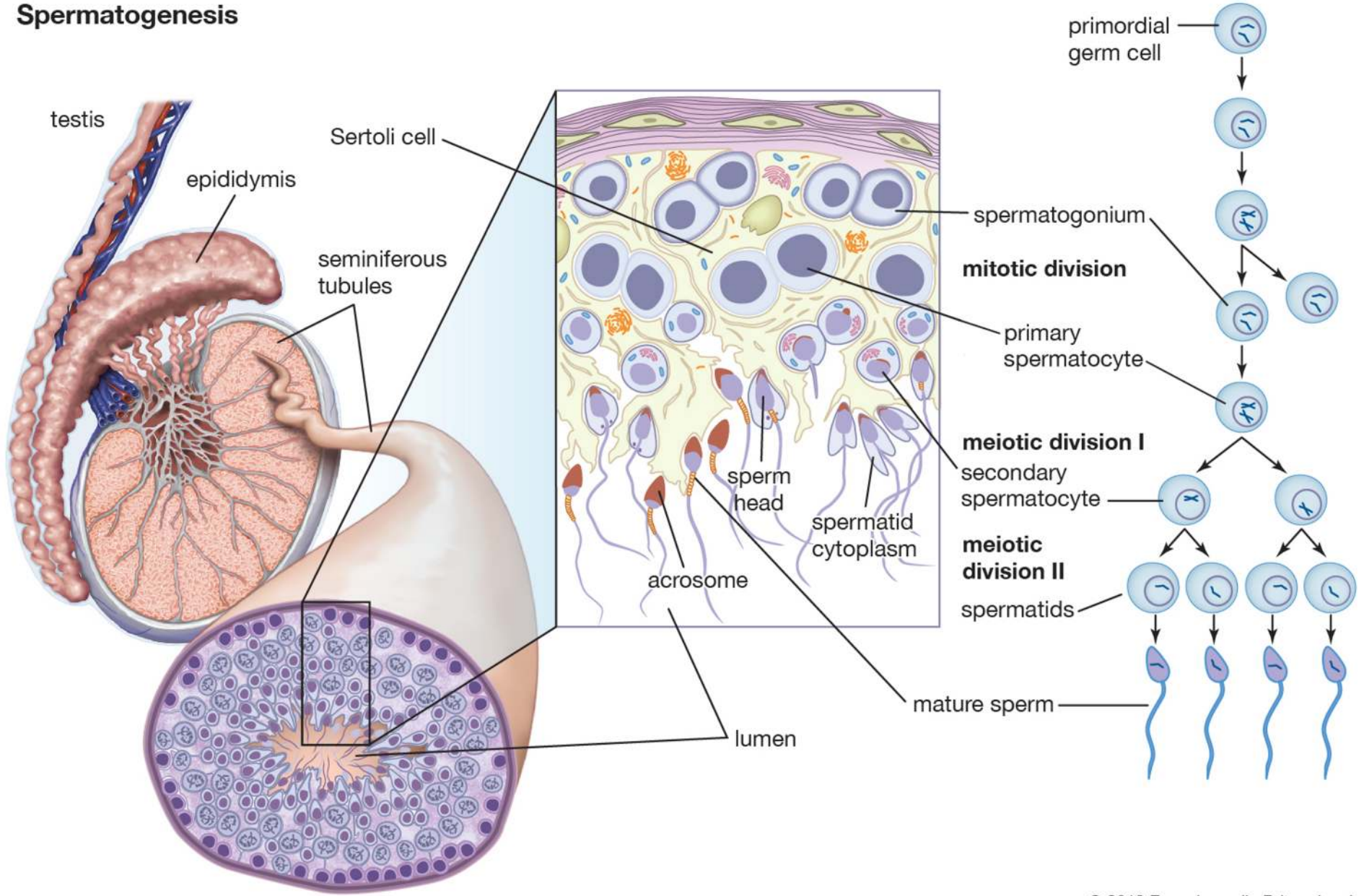
An scRNAseq dataset can be composed by thousands of cells and genes.

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**



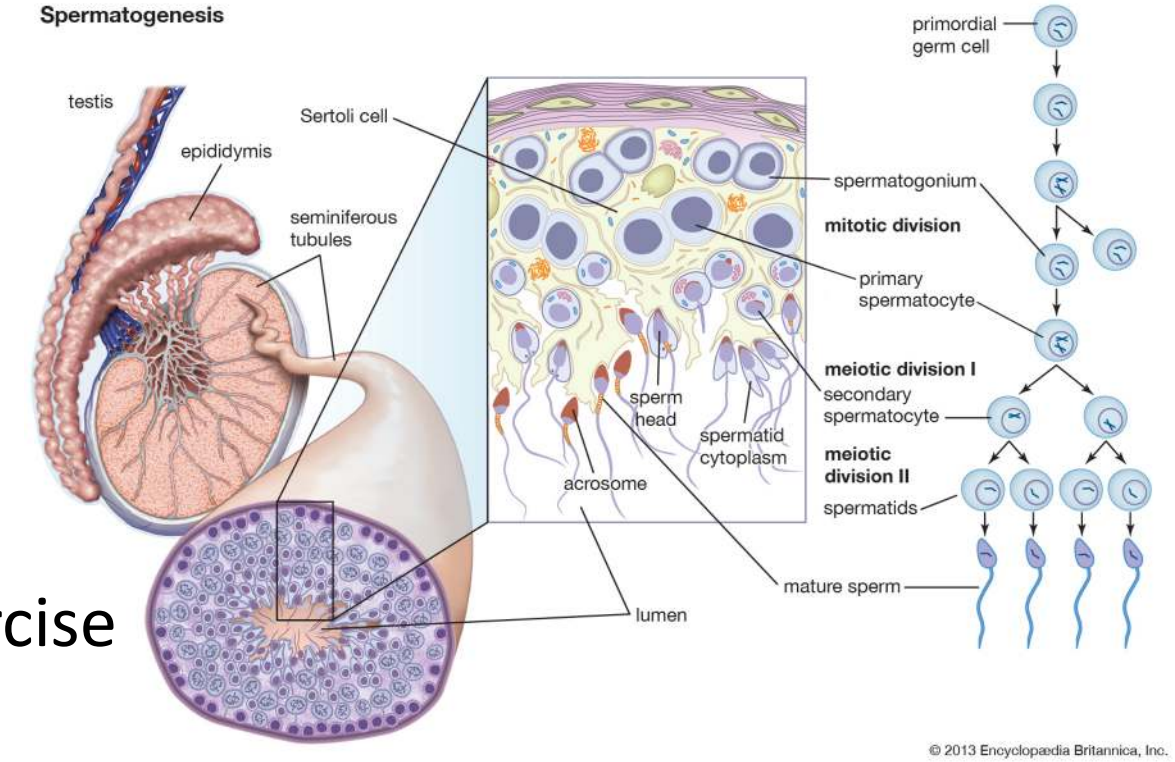
Analysis: An example dataset – human testes



Analysis: An example dataset – human testes

Human testes data:

- Composed of 65K cells
- Around 19K cells are outliers
- Data comes from 13 different datasets
- We will use a subset of this data in the exercise

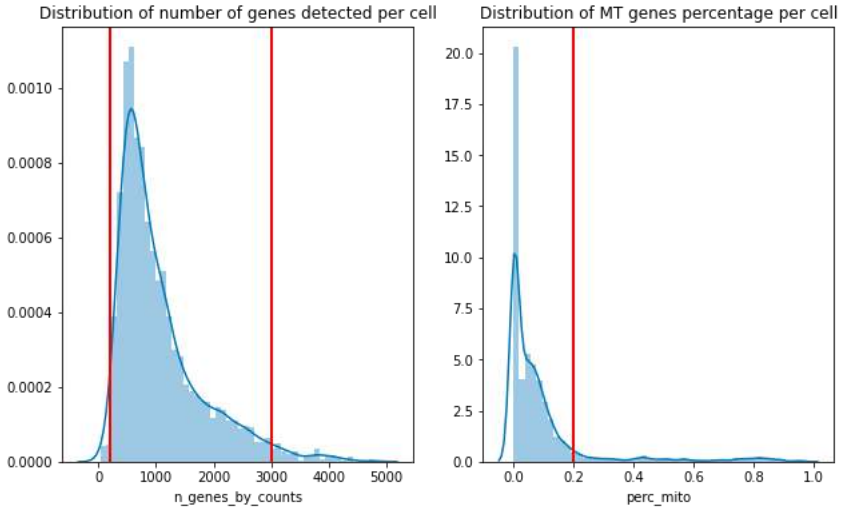


Analysis workflow: preprocessing

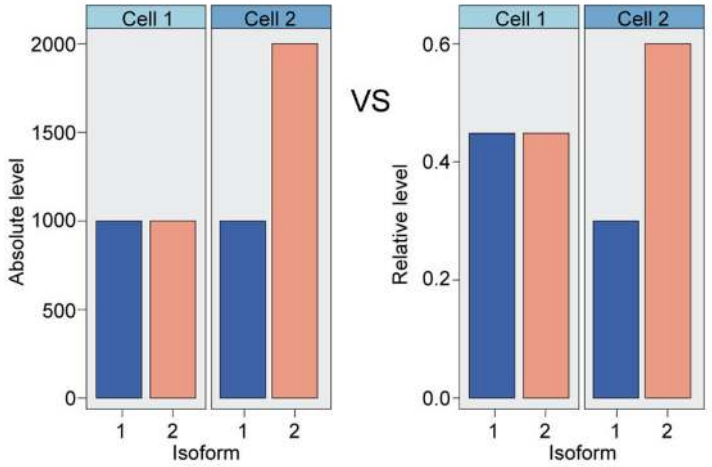
Raw matrix

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

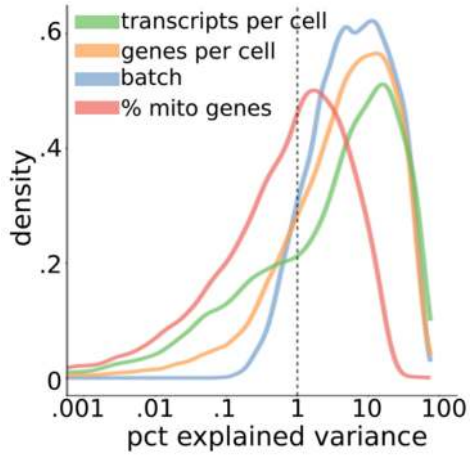
Identify cell outliers through quality measures



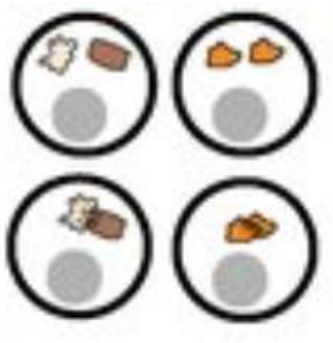
Normalize transcript counts



Identify technical/biological biases



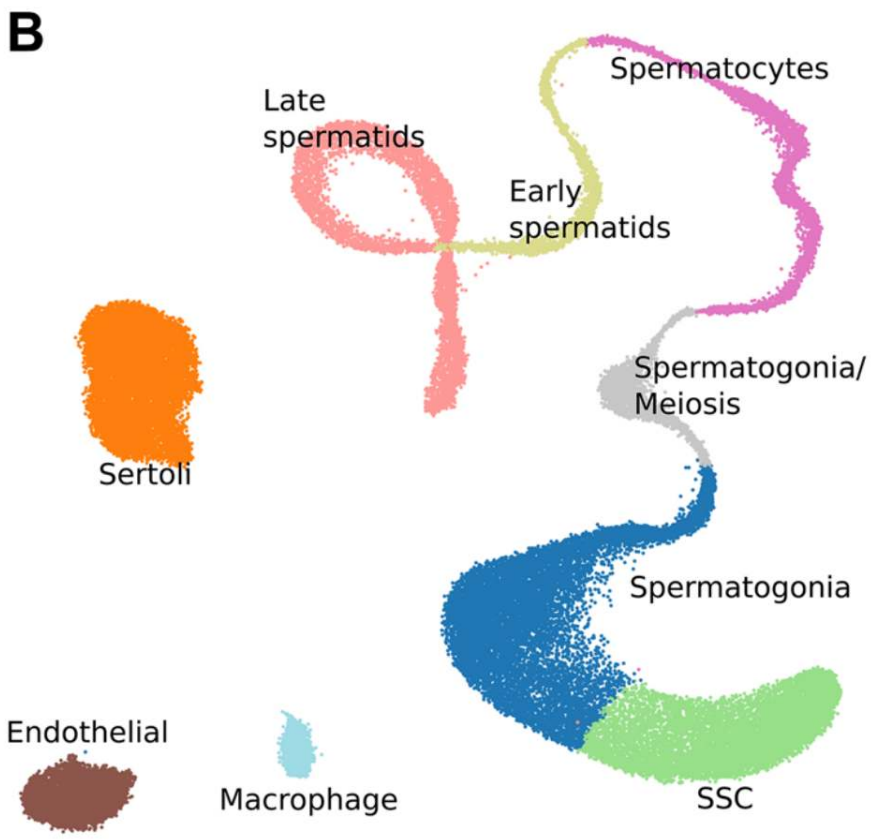
Doublet removal



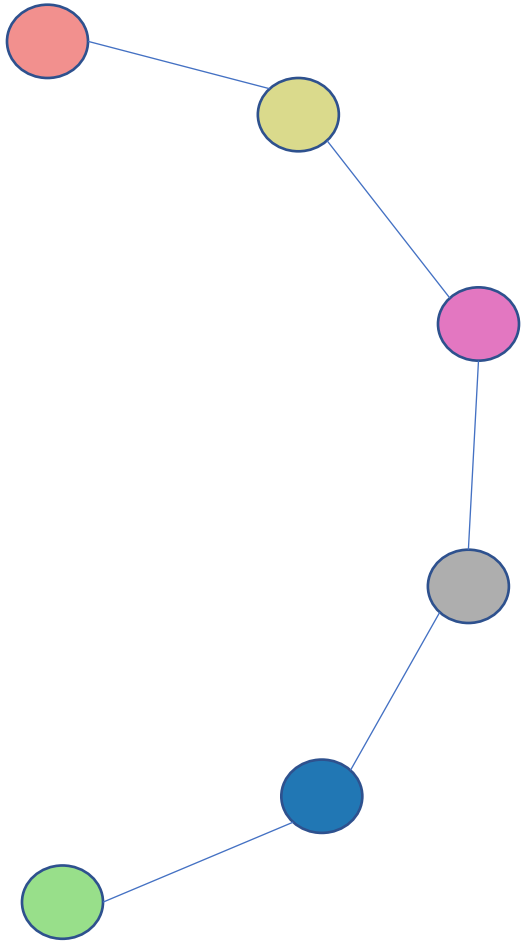
Dimensionality reduction

Analysis workflow: cell-wise analysis

Clustering and cell identification

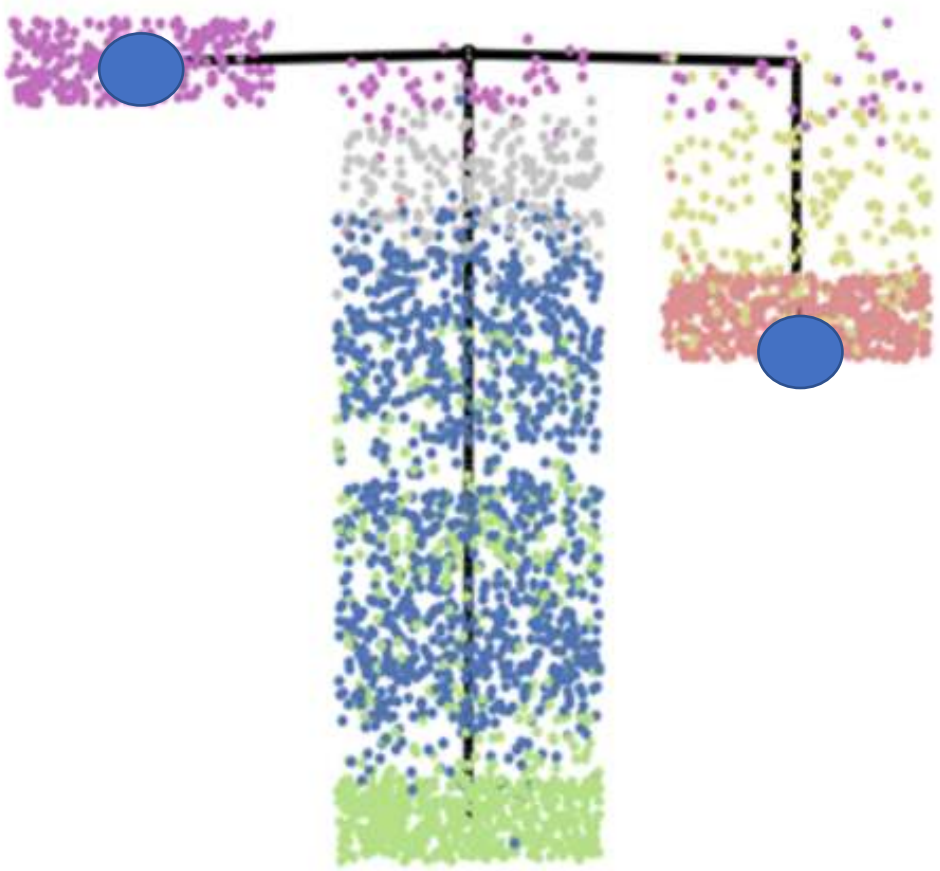


Cell hierarchy

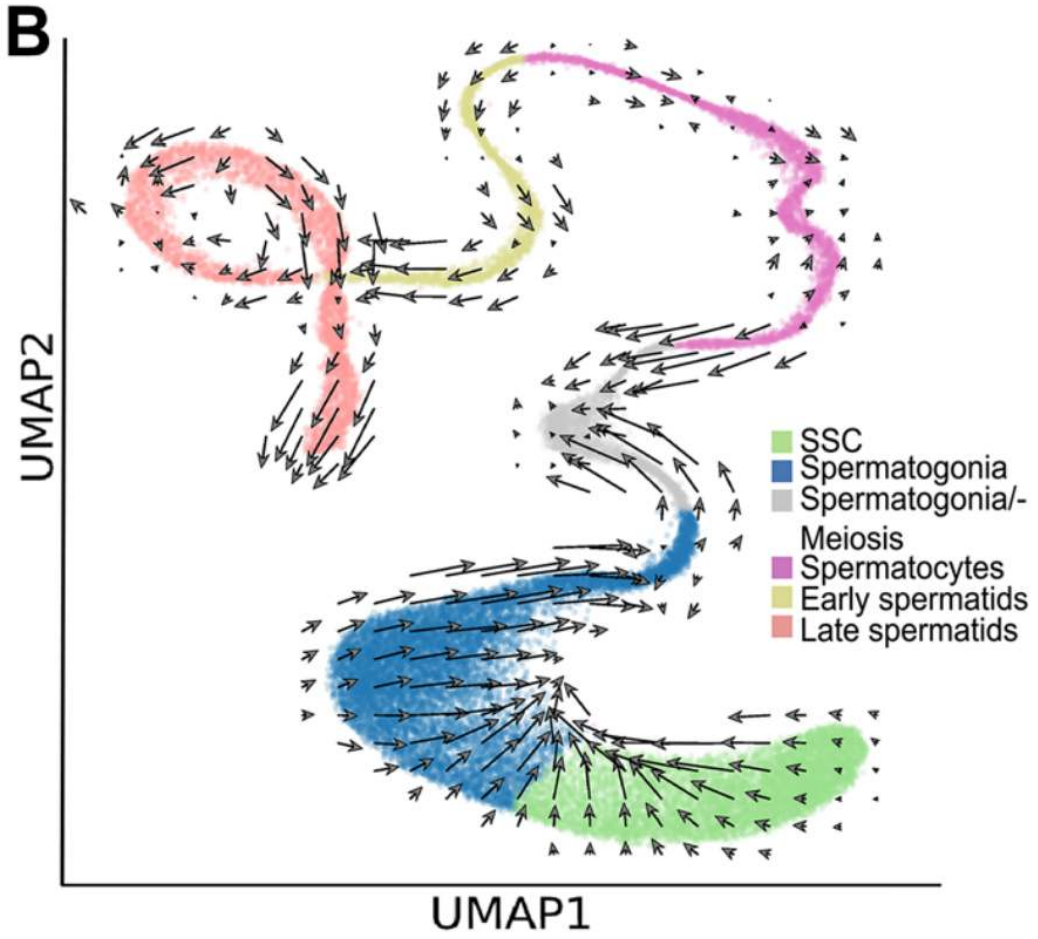


Analysis workflow: cell-wise analysis

Detection of cell fates

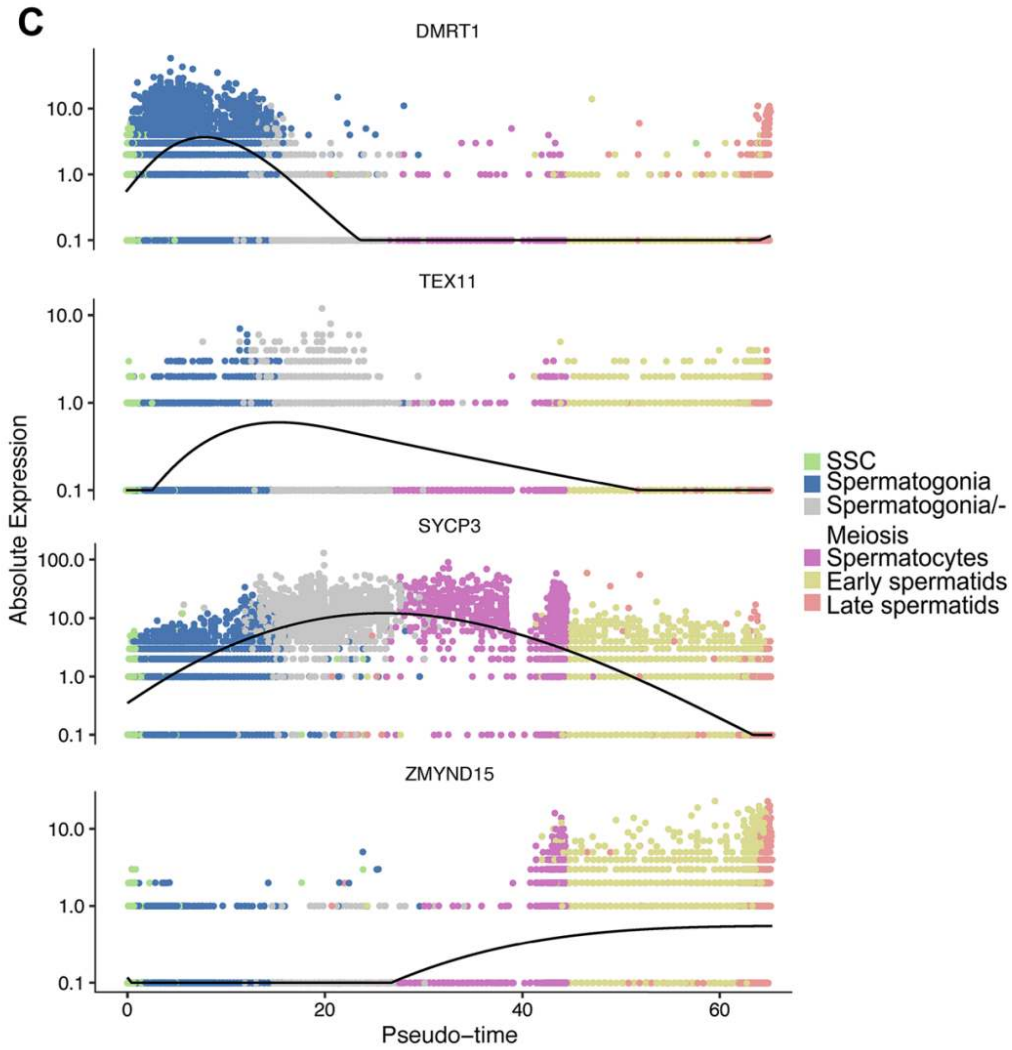


Prediction of molecular states

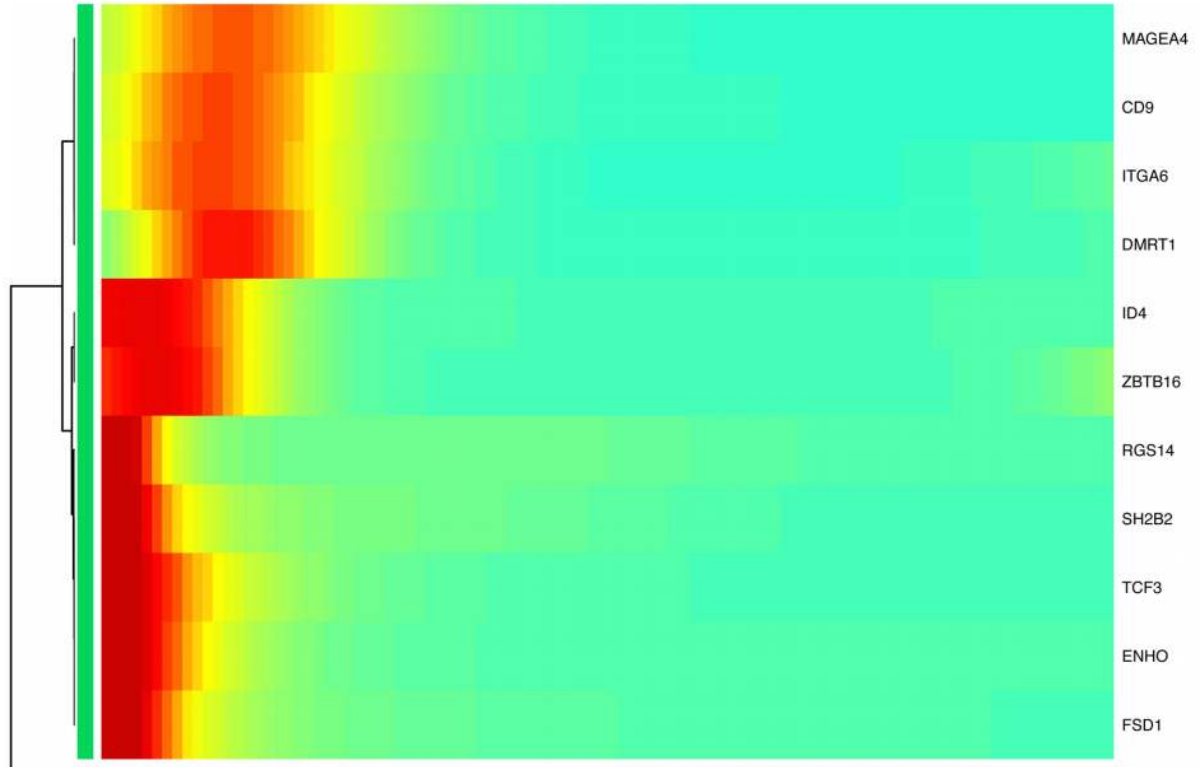


Analysis workflow: gene-wise analysis

Modeling genes trajectories in time

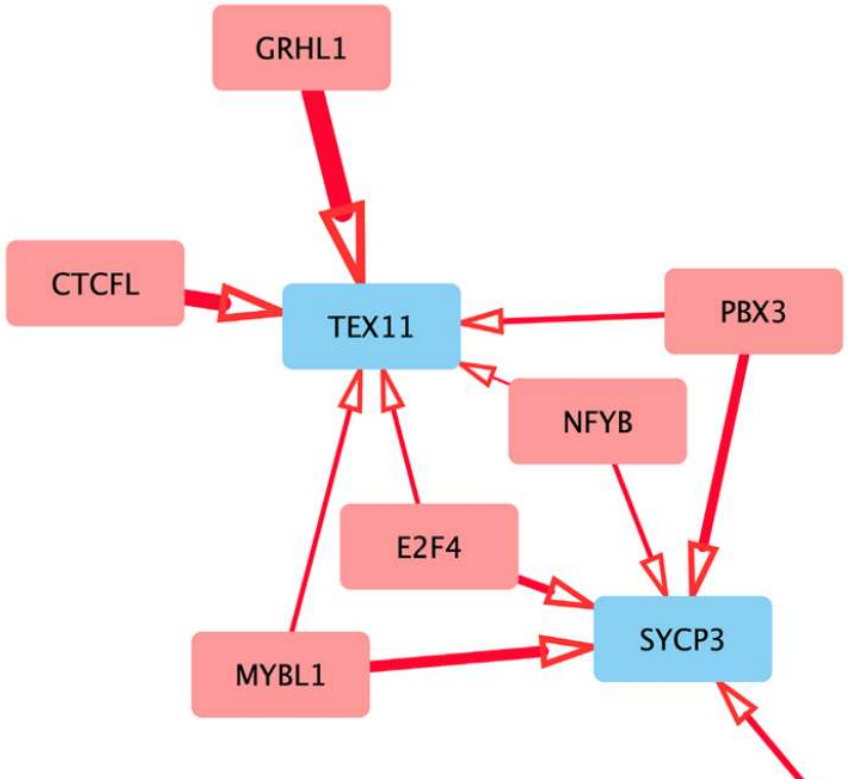


Clustering of coexpressed genes



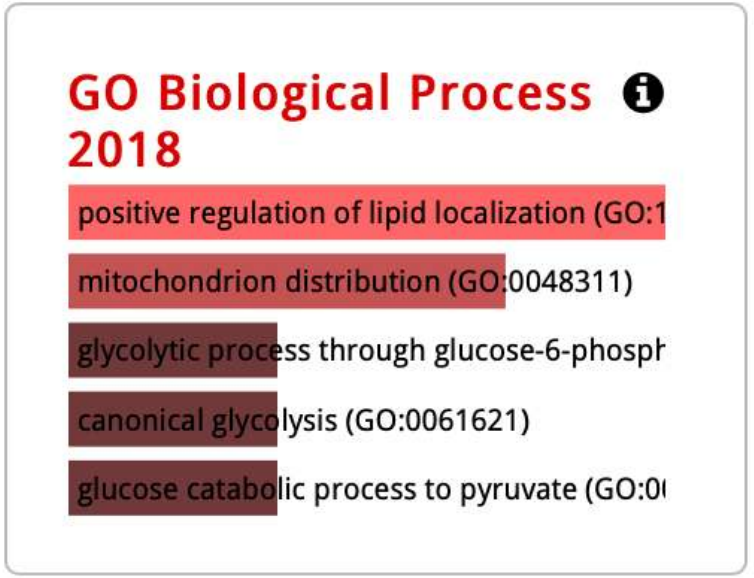
Analysis workflow: gene-wise analysis

TF-Gene Networks



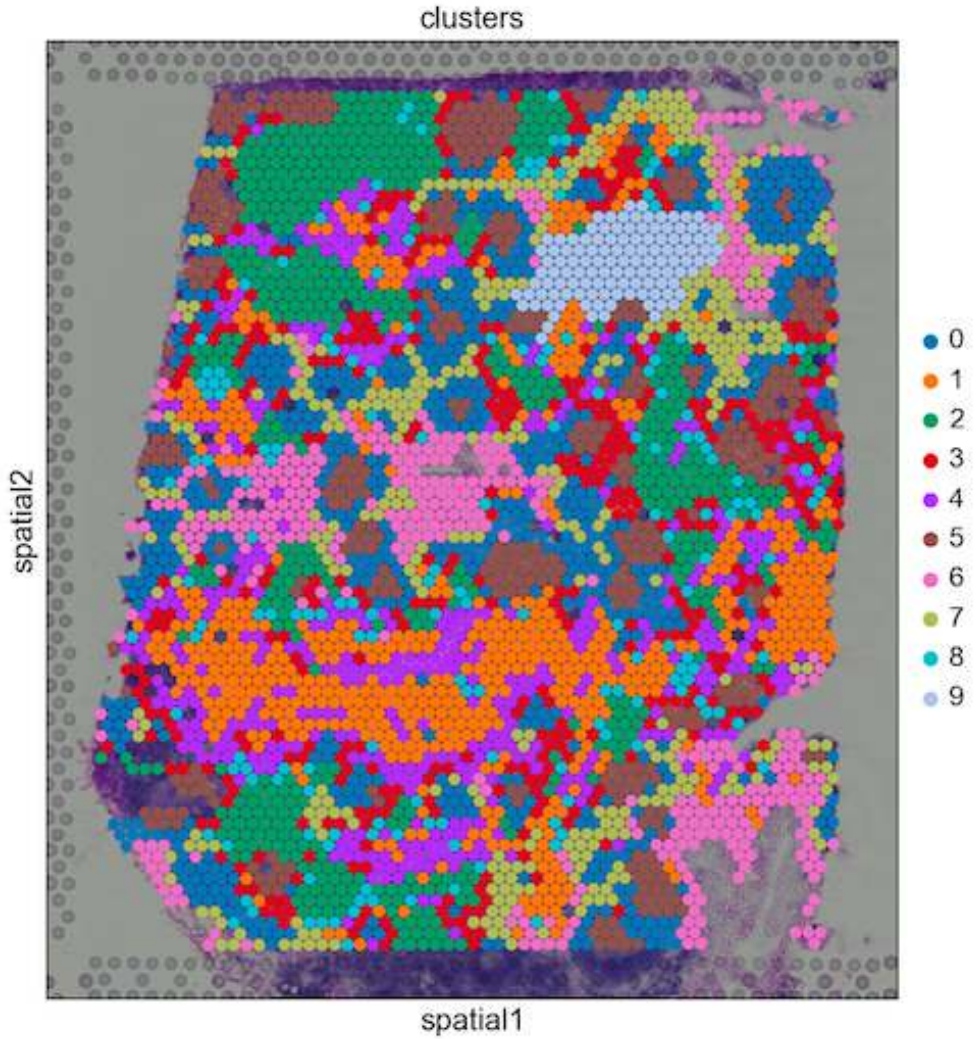
Gene Enrichment Analysis

Enrichr

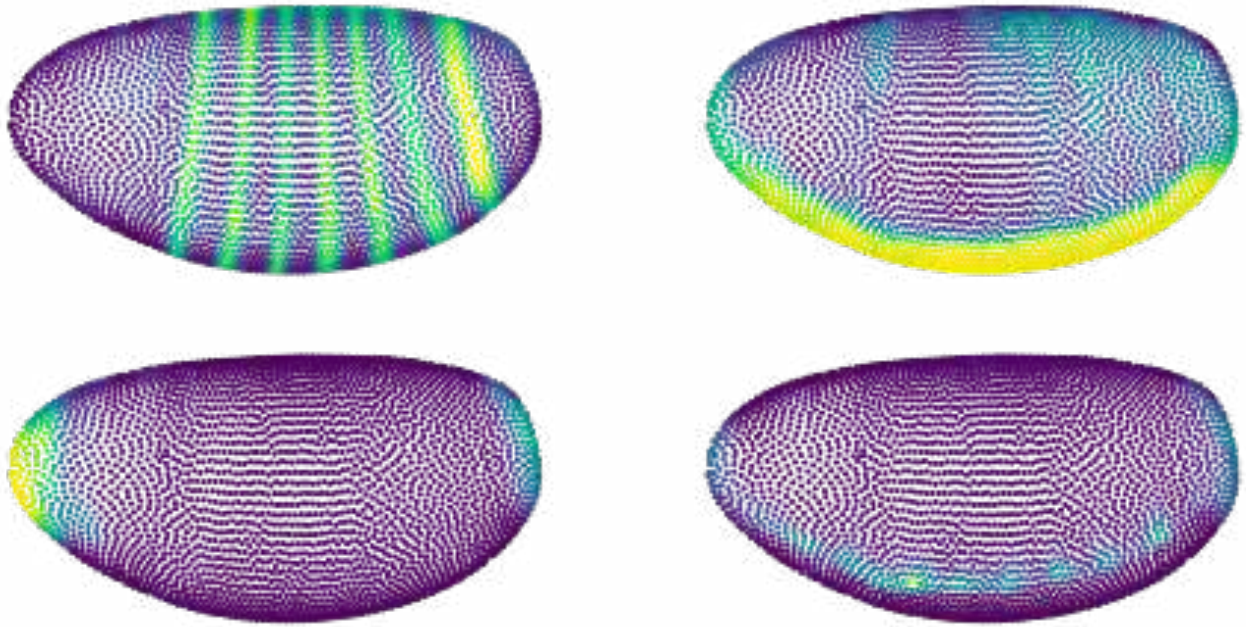


Analysis workflow: spatial analysis

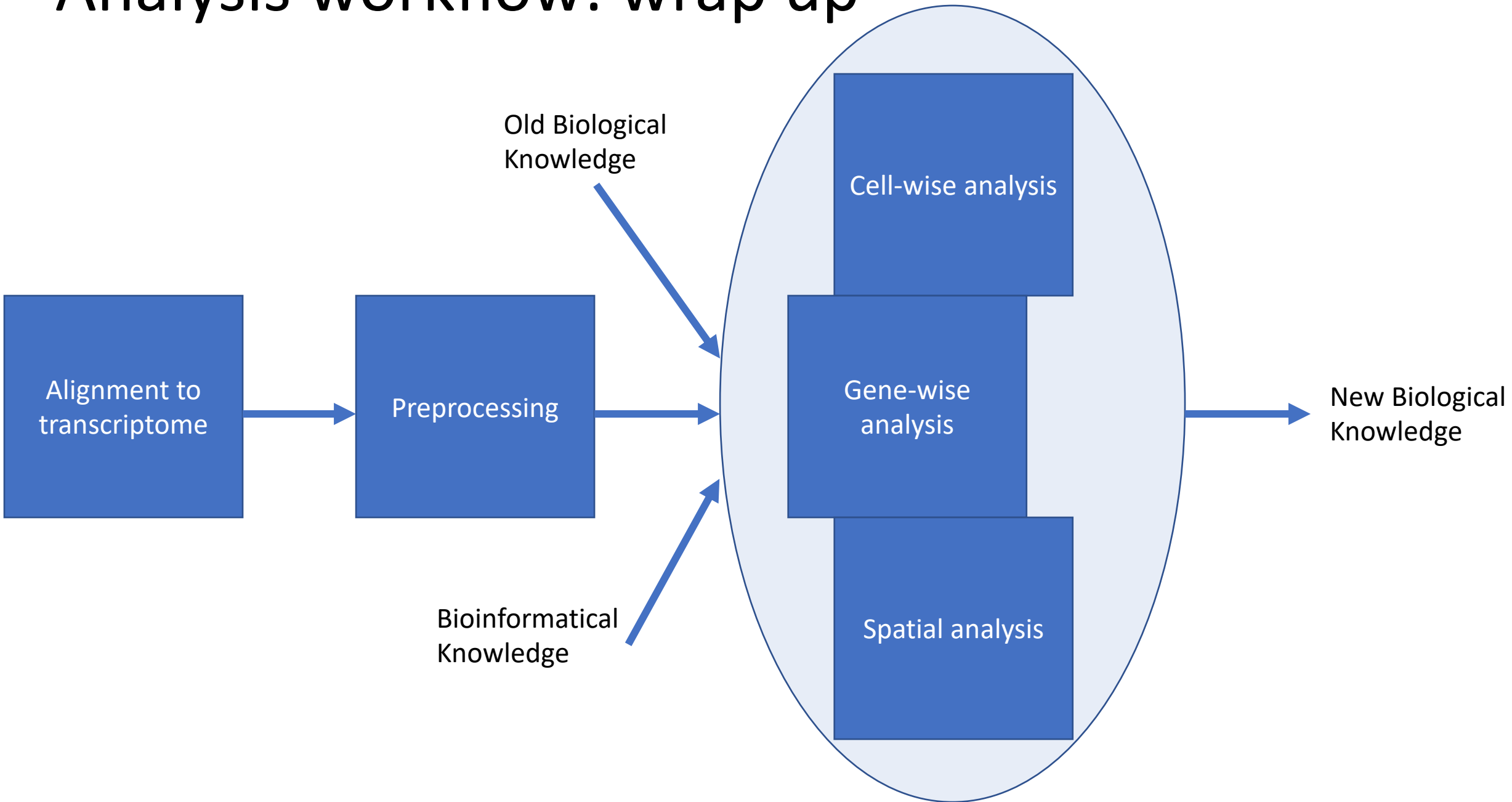
Relationship between cell location and cell types



Reconstruction of genes distribution in a tissue



Analysis workflow: wrap up



Conclusions

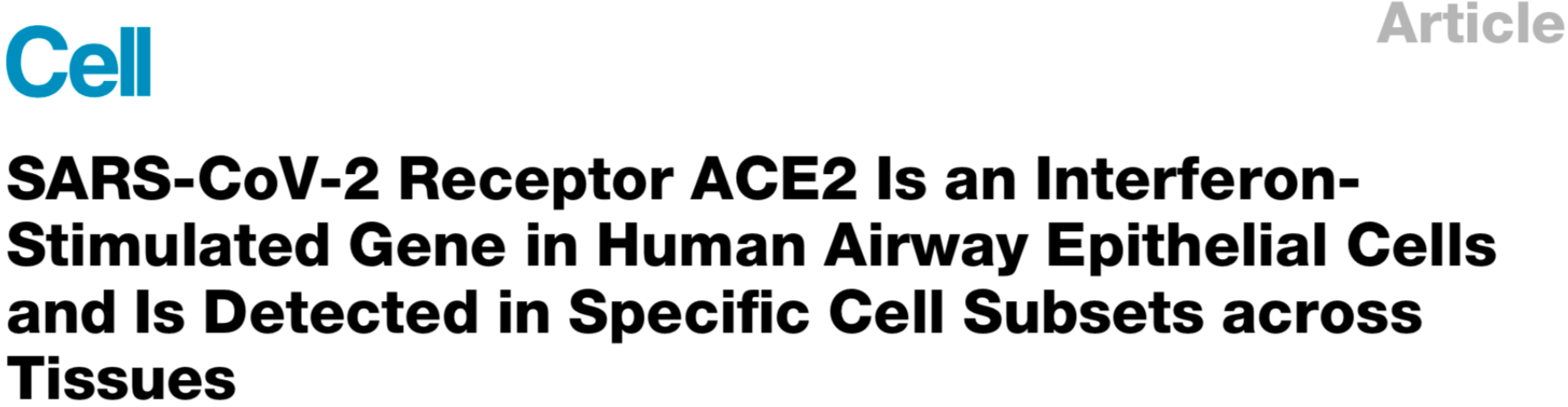
- scRNAseq data is sparse and can be noisy
- However, it allows to explore detailed biological aspects of a tissue
- There is not yet a standardised way of analyzing this data
- There are >300 scRNAseq tools
- Scanpy (python) and Seurat (R) have emerged as reliable analysis tools for many standard operations
- Yet more advanced scRNAseq analysis is still not a standard, and requires acquiring some programming expertise and 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

Paper discussion

You will discuss in group the following research paper



- Discuss the paper in groups
- Discuss answer to the questions
- Choose a person that will report the answers when we group together again