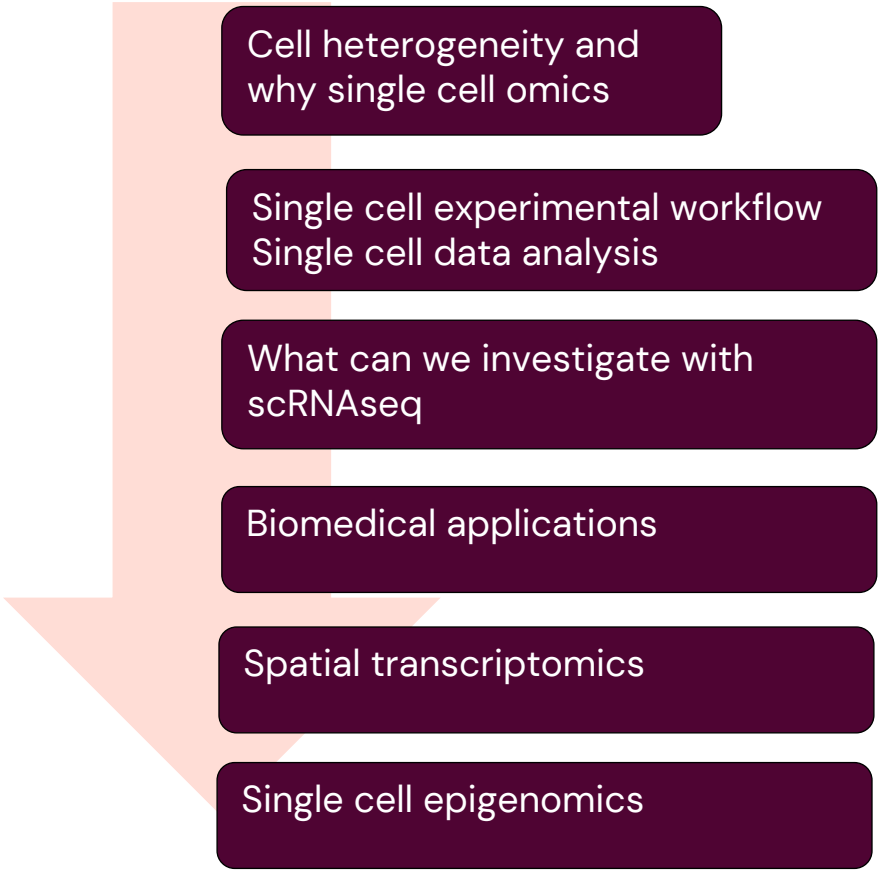
Single cell omics

Master's Programme in Translational Physiology and Pharmacology

Omics in science – bioinformatic analysis and visualization of gene regulation

Tina Gorsek Sparovec, PhD
tina.gorsek@ki.se

Overview


Cell heterogeneity and why single cell omics
Single cell experimental workflow
Single cell data analysis
What can we investigate with scRNAseq
Biomedical applications
Spatial transcriptomics
Single cell epigenomics

Cellular heterogeneity

Nearly all cellular systems are heterogeneous.

Multicellular organisms undergo specialization. What does it mean? What does it result in?
Cellular complexity, specific cellular functions, despite having nearly identical genomic architecture.

Cellular heterogeneity plays a critical role in development, homeostasis and disease.

Biochemical processes (gene, protein expression), cell cycle status and tissue micro-environment drive cellular heterogeneity even within the cell population, which raise a need to investigate these processes on a cellular level.

How to tackle the issue of cellular heterogeneity?

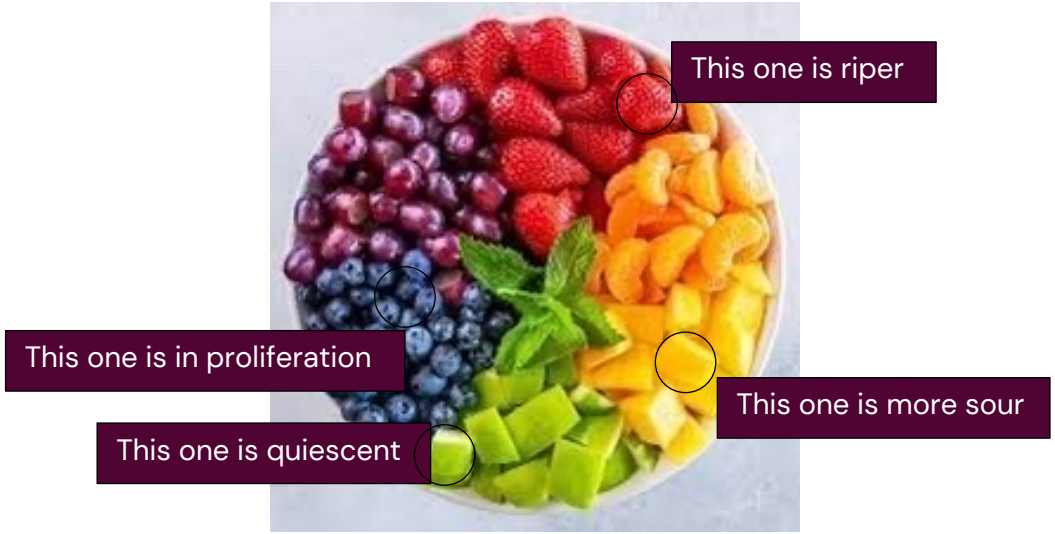
E.g. Bulk RNAsequencing

E.g. single cell RNAsequencing

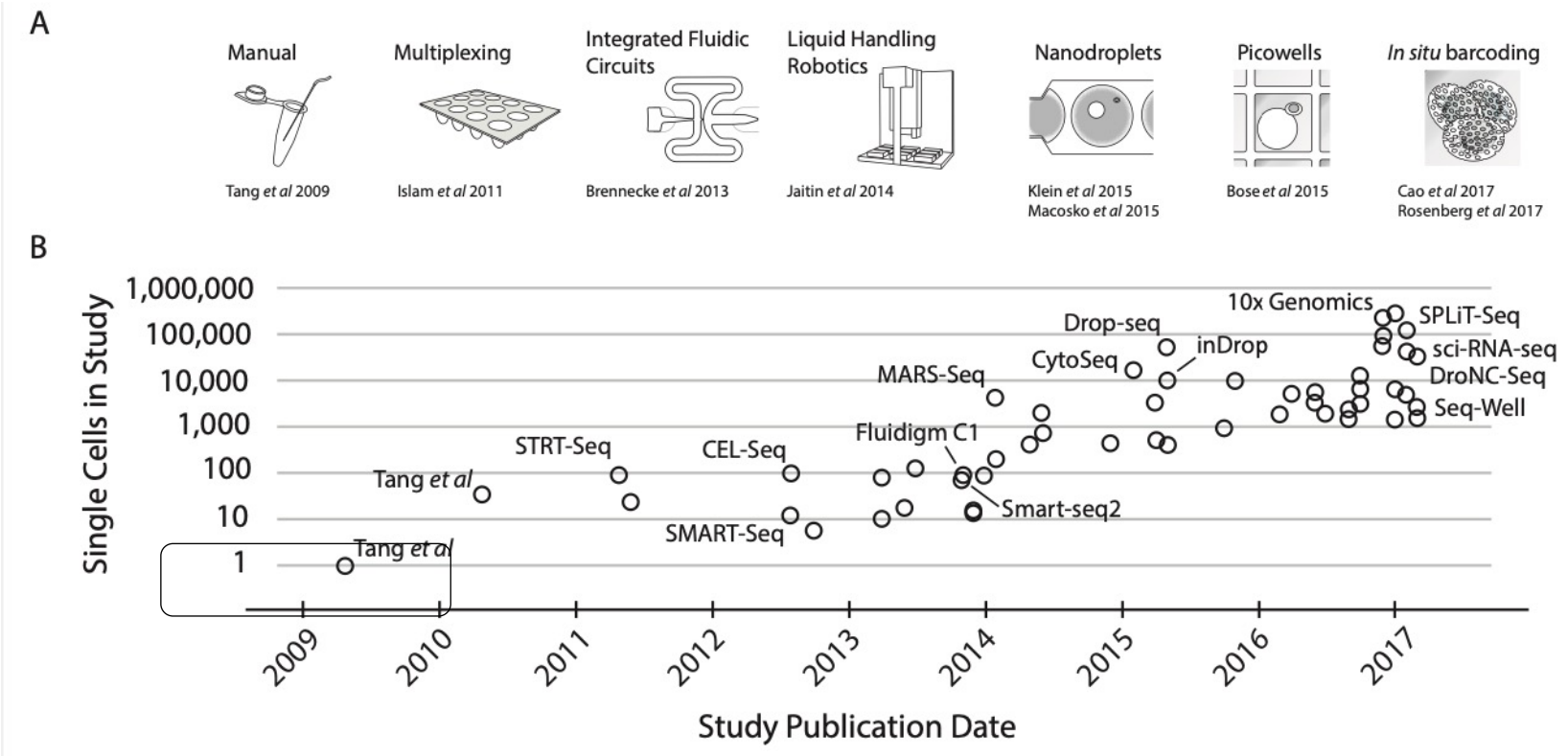
Yields information for a cell population,
but unable to decipher cellular heterogeneity,
cell stage, cell-cell interactions etc.

Single cell omics

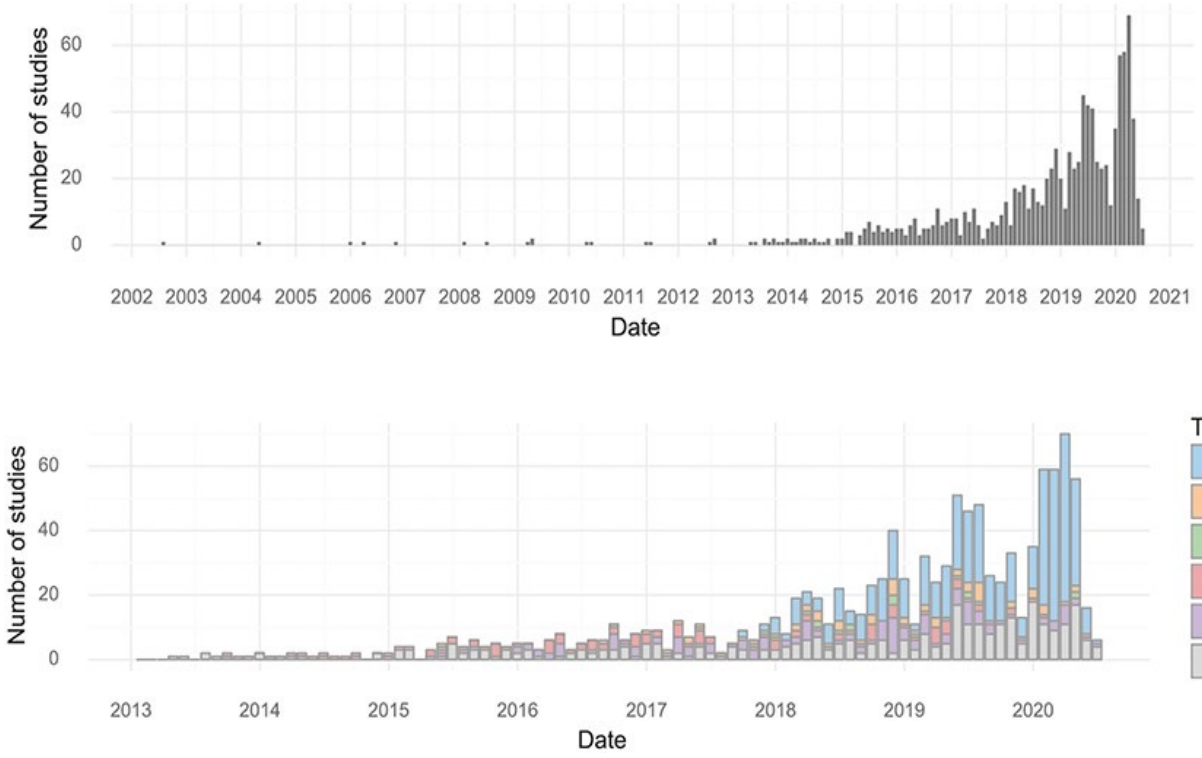
- Single cell omics offer an insignt into genome, transcriptome, epigenome, proteome and other omics modalities at the level of individual cell.
- Capturing a full spectrum of cellular subpopulations, cell states and dynamics, identify lineage trajectories, cell fate decisions and cellular responses to external stimuli.



Exponential increase in throughput – it all started with one cell



Where is it going

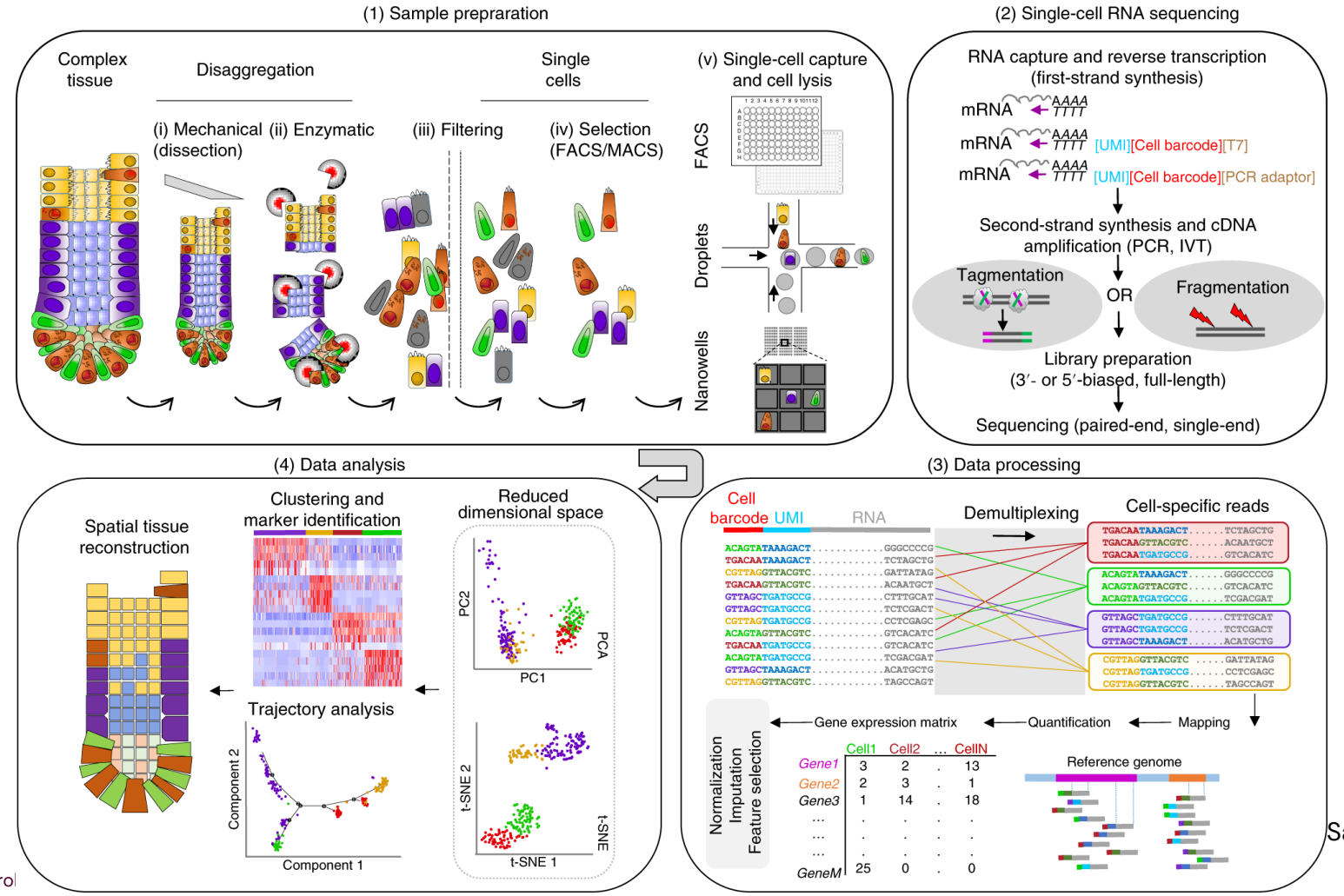


Svensson et al. 2020

Single cell omics

transcriptomics	epigenomics	proteomics	metabolome
Understanding gene expression patterns	Understanding gene regulation and cell identity	Diversity of protein expression within individual cell	Diversity of metabolite expression within individual cell
Abundance of all transcribed RNA in a cell, cell-cell interaction, cell states etc.	DNA methylation, histone modification, chromatine accessibility, chromosome conformation	Posttranscriptional modifications, signaling mechanisms, protein binding	intracellular, cell membrane–bound, and secreted metabolites.

Workflow



Sanger Institute

1. Sample preparation

Aim
to obtain single cell suspension

Input material
frozen or fresh tissue, liquid tissue (blood)

Sample preparation protocol is **optimized to specific tissue**/organism and it is one of the most cruical steps!

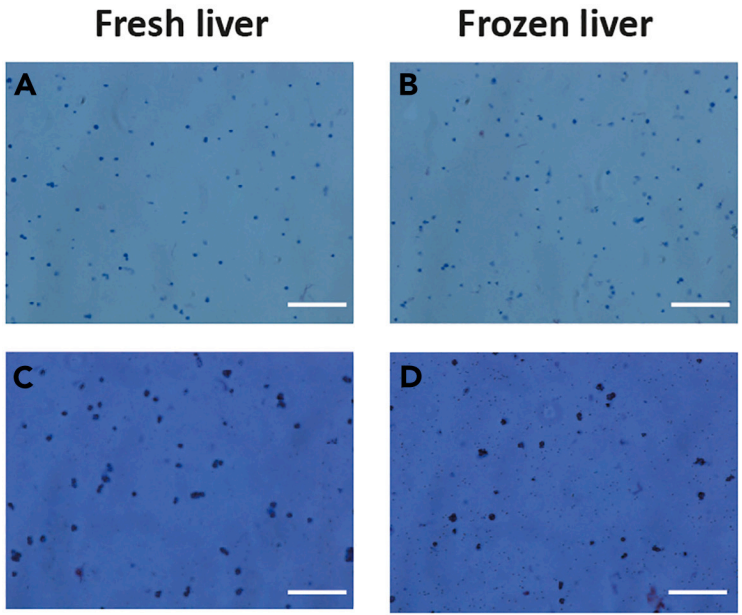
E.g.: plant material require additional processing steps to efficiently remove cell wall;

Only high-quality single cell suspension can lead to successful single-cell studies.

Challenges:
- fragility of the starting sample,
- physical stress (potentially damages the cells)
- choice of buffers
- duration of the cell dissociation
- fibrotic or scarred tissue

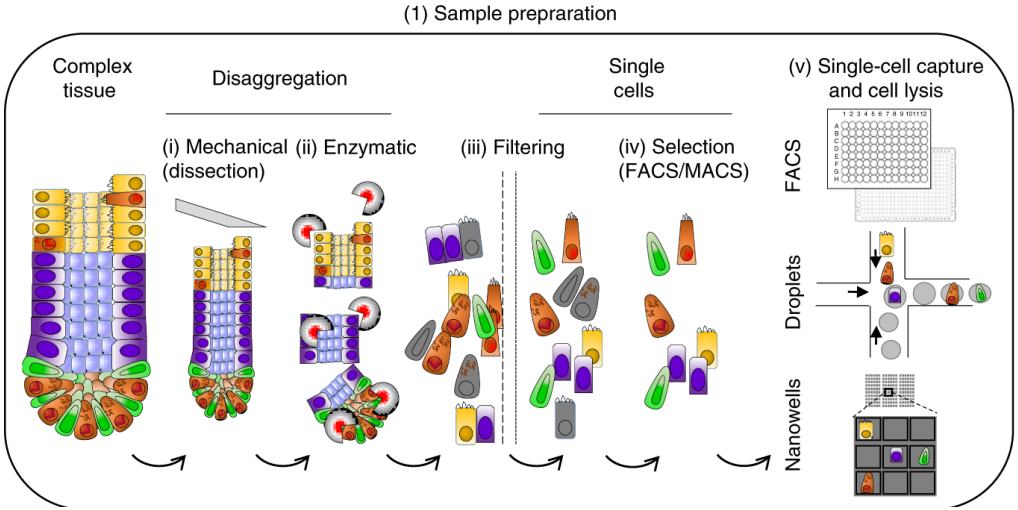
Single nuclei RNA sequencing – Modification of scRNAseq

- Single cell isolation from **frozen samples** yields low number of cells – single nuclei sequencing
- Challenges with **tissue dissociation** and **cell size (multinucleated cells)** can be overcome.
- Nuclear RNA usually contains a higher proportion of unprocessed RNA, with more of the sequenced transcripts containing introns.



Loft et al 2021 Protocol

1. Sample preparation



Quality control check of the isolated cells/nuclei!

- **Tissue dissociation**:
  mechanical (mincing, dissecting)
  enzymatic
- **Filtering**:
  to avoid cell aggregations, and remove debris
- **Cell selection for cell enrichment**:
  if we want a certain population of cells (e.g immune cells) we sort them with FACS/MACS
- **Cell capture**

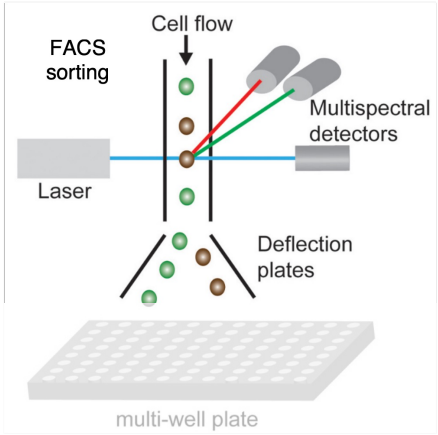
1. Sample preparation – cell capture

Isolating cells into individual wells of the plate using pipetting, microdissection or FACS.

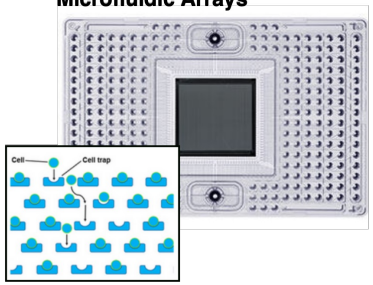
<u>Advantage</u>:
assessment of cell quality and discarding of damage cells is possible!

<u>Cons:</u>
often low-throughput, higher amount of work.

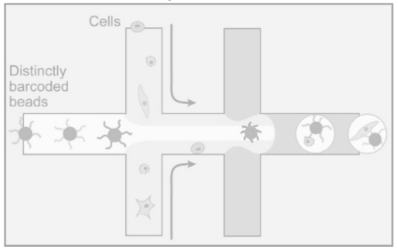
Microtitre Plates



Microfluidic Arrays



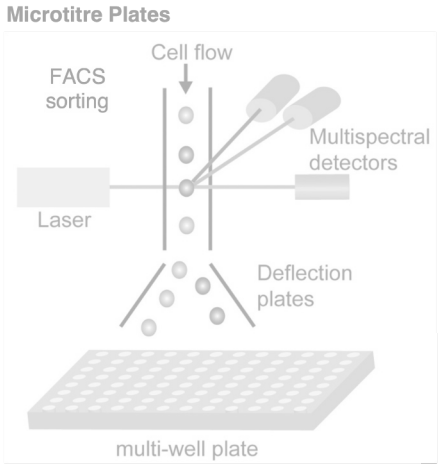
Microfluidic Droplets

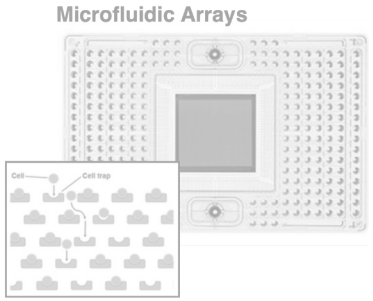


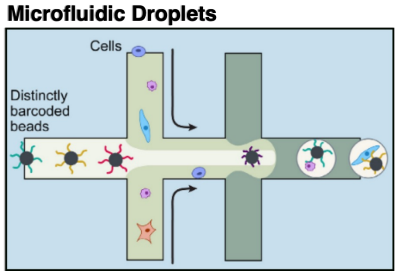
Integrated system of capturing cells and performing reactions necessary for the library preparation.

<u>Cons</u>:
low number of cells captured – issue if dealing with rare cell types

1. Sample preparation – cell capture







Offers the highest throughput and is the most popular method nowadays.

Principle:
it encapsulates individual cell inside a **nanoliter-sized oil droplet** together with the **bead**.

Each bead is loaded with enzymes and barcodes. What is a barcode?

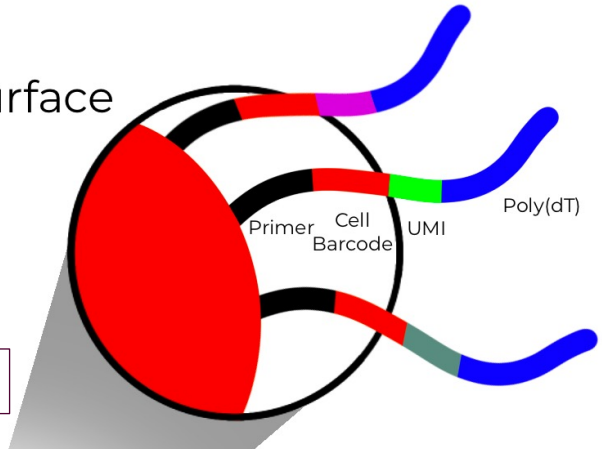
Unique Molecular Identifiers

- Short nucleotide sequences used to distinguish molecules rather then associate transcripts with cells – gene expression quantification.

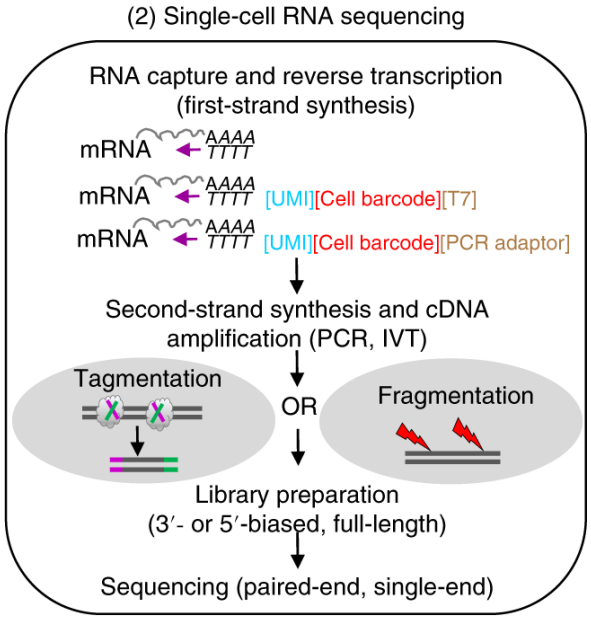
- Nucleotide sequences are random with a very low likelihood of a duplication within a single bead – each mRNA read is assigned a unique UMI

Cell barcode – label the cell source

UMI – labels the mRNA read



1. Sample preparation – single cell RNA sequencing



Post cell capture, individual cells/droplets are lysed

Conversion into cDNA via reverse transcription

cDNA amplification

Library construction and library QC

Sequencing

Sequencing

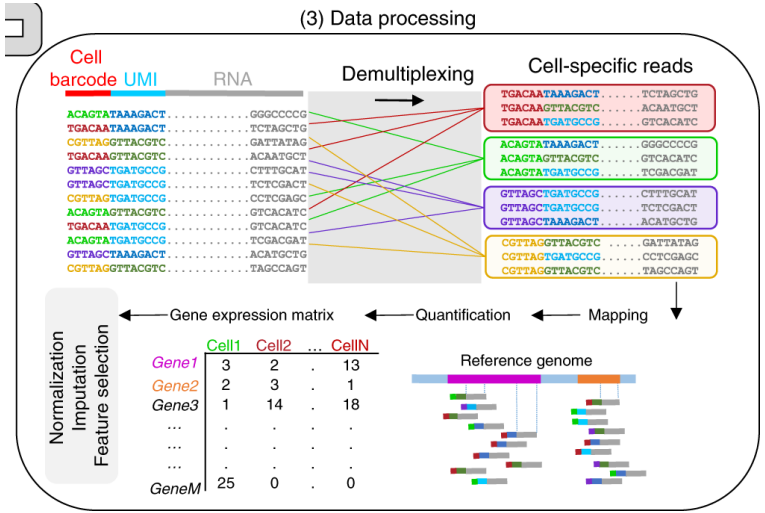
- Sequencing of cDNA library

- Currently used technologies:
  - Single-end reads (either 3’ or 5’)
  - Paired-end reads
  - Long reads (eg SMART-seq2)

Tag-based transcript quantification

- Combining tag-based protocol with UMI improve the accuracy of transcript quantification (PCR amplification step creates several duplicate copies, because the amplification is exponential- this could lead to unfair representation. Therefore, UMI is part of sequencing read and is computationally taken into account)

3. Processing raw scRNAseq data in droplet technology



1) Formatting reads and filtering noisy cellular barcodes

2) Demultiplexing the samples
assigning reads to the cell

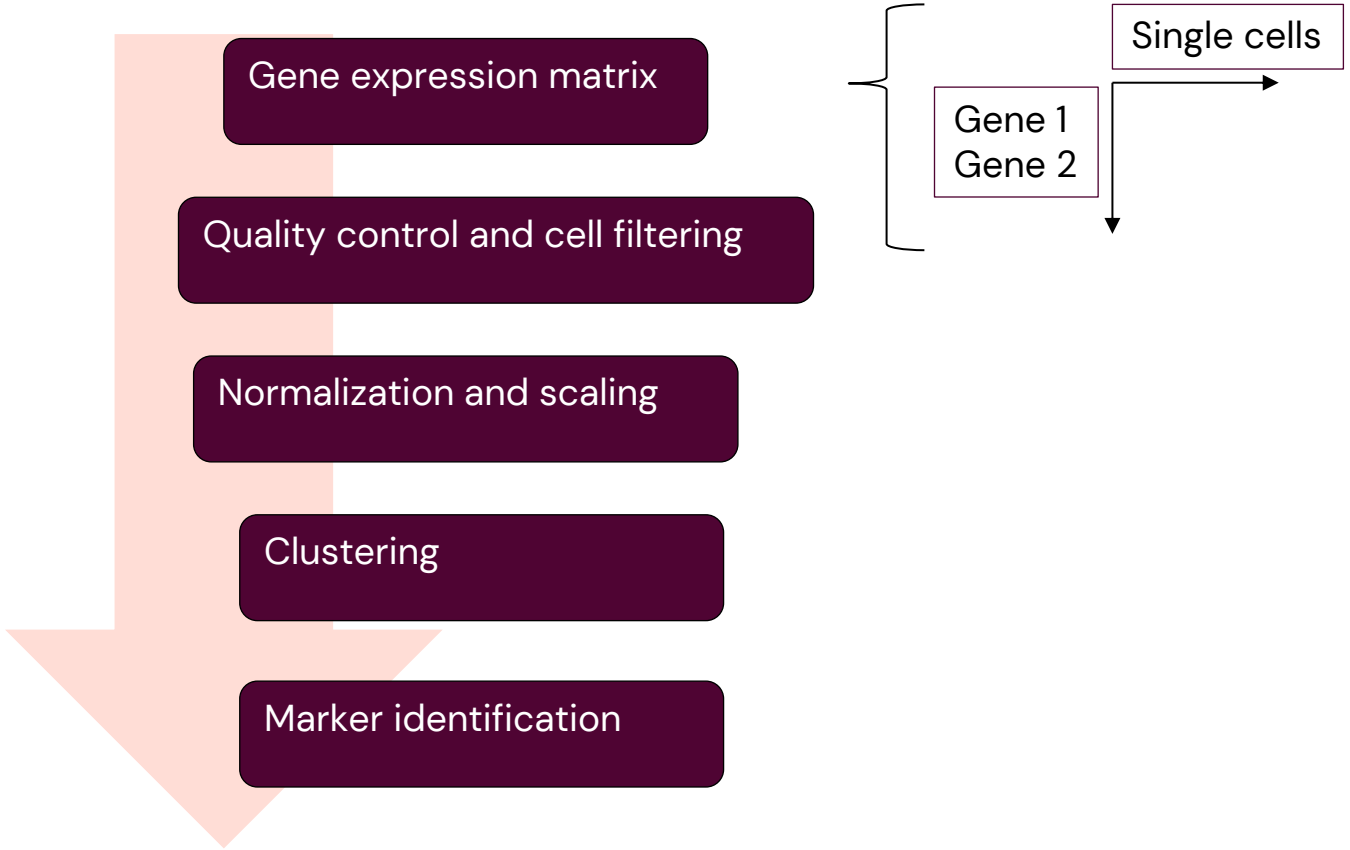
3) Mapping to the transcriptome
alligning the reads to the reference genome,
(what about non-model organisms?)

4) Quantification of reads

Several software suits or pipelines to perform preprocessing (Cell Ranger, kallisto, STARsolo)

Gene expression matrix –
result of gene expression quantification

HBC


Gene expression matrix
Quality control and cell filtering
Normalization and scaling
Clustering
Marker identification
Single cells
Gene 1
Gene 2

4. Data analysis - Quality control and filtering

Quality control:

process of improving data by removing identifiable errors from the data set. Typically the first step performed upon data acquisition.

Be aware: QC process alters the data, however it cannot turn bad data into useful one (for example – bad single cell suspension)

Extreme caution not to introduce new features into data.

Aim

We want the same information only more accurate (without noise).

4. Quality control and filtering

Removal of poor quality cells (if not successful → technical noise, which can obscure biological signals of interest)

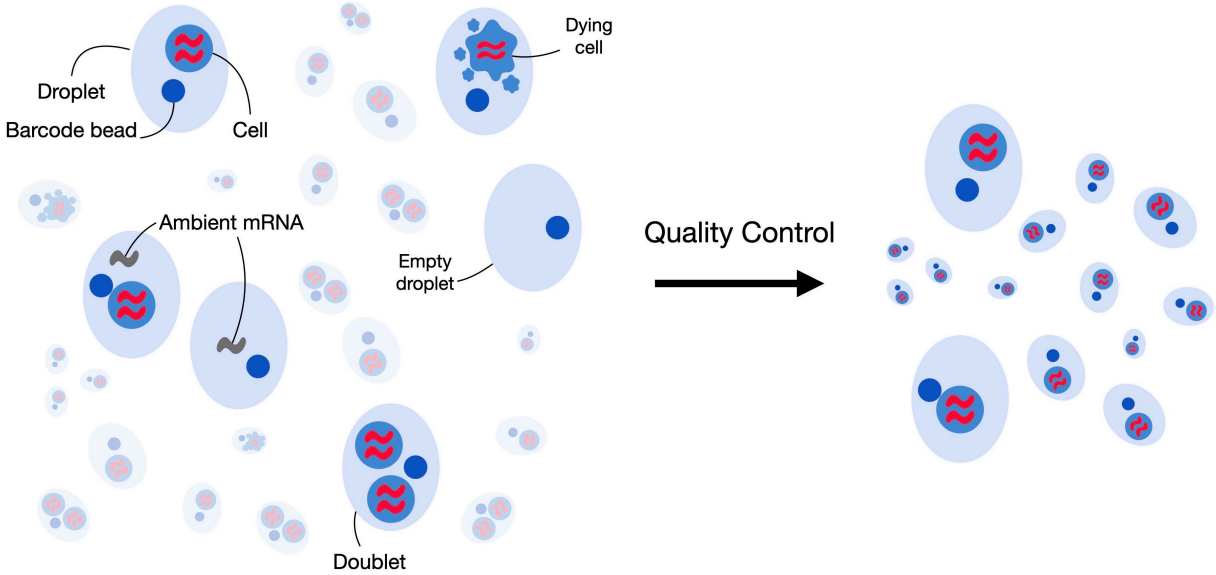
QC measures vary from experiment to experiment

<u>In scRNAseq it includes</u>:

- removal of the mitochondrial (low quality cells , ribosomal genes;
- duplets, filtering unique features;
- removal of low quality cells and empty droplets

4. Quality control and filtering

Removal of poor quality cells (if not successful → technical noise, which can obscure biological signals of interest)



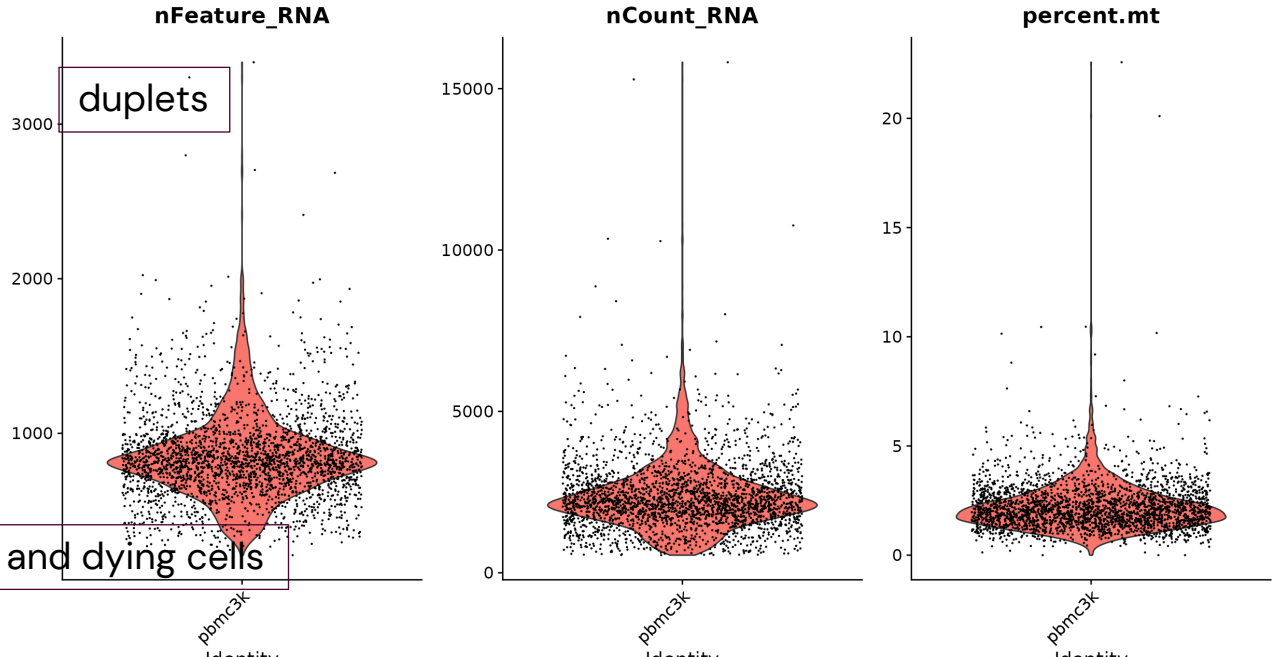
Heumos et al 2023 Nat Rev gen

4. How do we perform quality control

Visualization is cruical after each step of quality control! Have we eliminated enough or too much noise?

Unique featurs

Mitochondrial count



Low quality and dying cells

4. Keep in mind!

Every QC process introduces error. However, the impact of the new error should be smaller than the impact of errors that QC corrects.

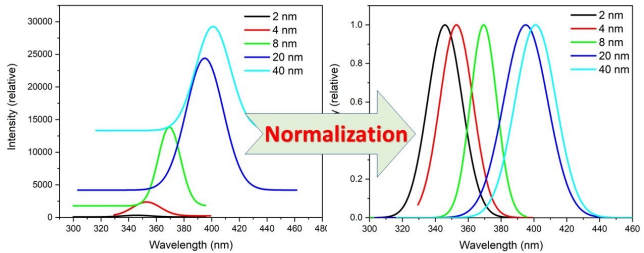
Example: some cells will have higher fraction of mitochondrial counts and should not be filtered out.

Extensive tweeking of QC with aim to ”improve” data can lead to overfitting. What is overfitting?

4. Normalization and dimensionality reduction

To be able to compare expression values across the samples we need to perform **normalization**.
Counts for each gene is proportional to the expression of RNA (of interest) and other factors (noise).
Normalization – adjusting raw counts to account for the ”other factors – noise”. D

1) Scaling
   due to differences in the amount of mRNA between cells
2) Transformation
   applying function/transform to to each measurement onto a common scale

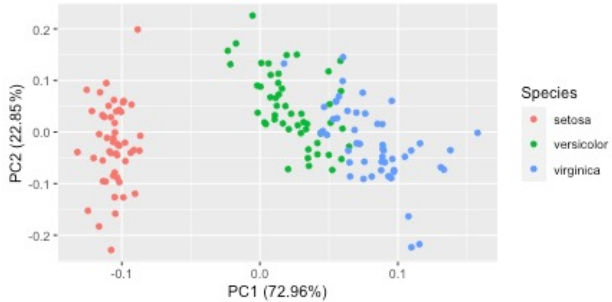


Example: Log normalization – using logarithmic transformation to standardize data with high variance. It transforms the data on the scale that approximates normality.

4. Linear dimensional reduction

Transformation of data from high-dimensional space into a low-dimensional space by applying linear transformation.

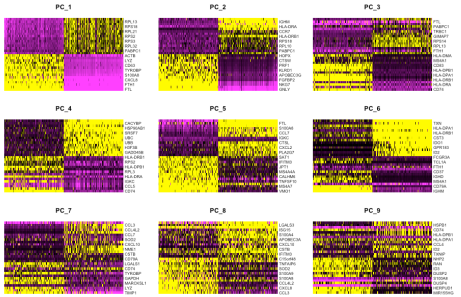
<u>Principal component analysis (PCA)</u>
dimensionality reduction in continuous data. PCA defines new axes through the data to capture the highest amount of variation possible.
Aim: to reduce the number of variables while preserving as much information as possible



4. Clustering analysis

Aim: assigning cells to groups of similar cells and determination of the cell identity

1) Find neighbors (based on constructed K-nearest neighbour graph it draws edges between cells with similar expression patterns)
2) Find clusters (grouping of the cells together)



Visualization of clusters is performed through dimensionality reduction techniques (t-SNE, UMAP)

t-distributed stochastic neighbor embedding (t-SNE) and Uniform Manifold Approximation and Projection (UMAP)

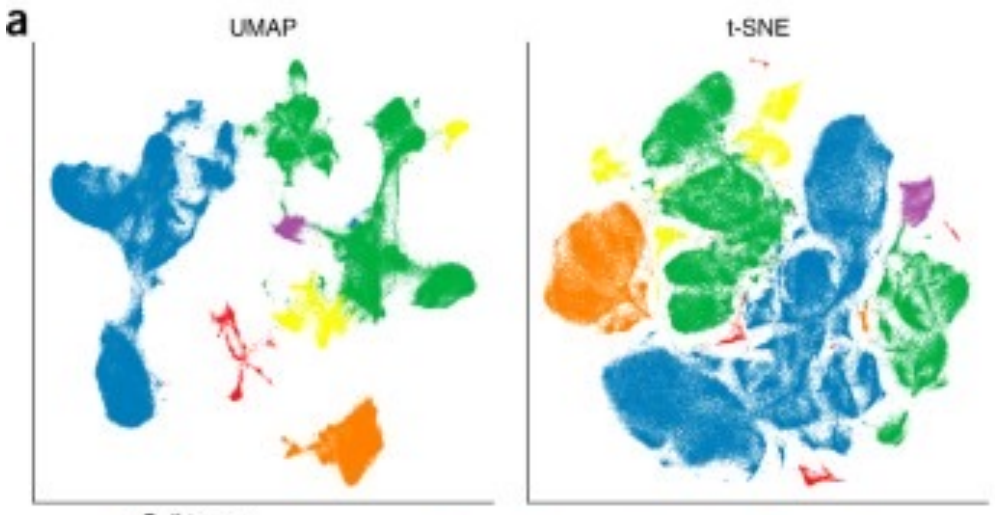
probabilistic method (algorithms) for visualizing cells with similar local neighborhoods in low dimensional space.

t-SNE
if two points are close, they are most likely close in the higher dimensional space. Same if points are farther away.

UMAP
similar principle, with better preservation of global structure in the final projection and with clearer separation of the groups of similar categories from each others.

https://pair-code.github.io/understanding-umap/

Dimensionality reduction for visualizing single-cell data using UMAP



UMAP provides fastest run times, higher reproducibility and most meaningful organization of cell clusters

T-SNE problems:
-cluster size not interpretable
-cluster distances not interpretable
-random noise can look non-random

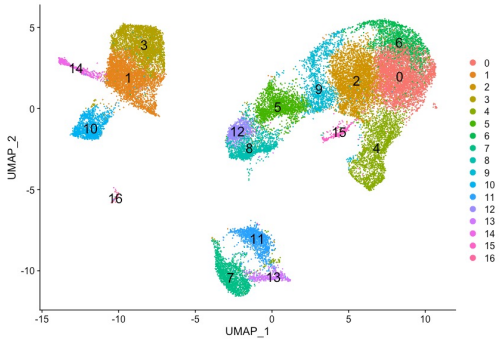
Becht 2019 Nat Biotech

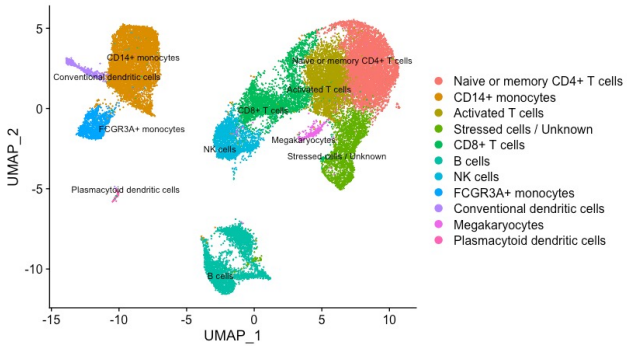
4. Marker identification

Aim: determine gene markers of each cluster and to identify cell types in each clusters

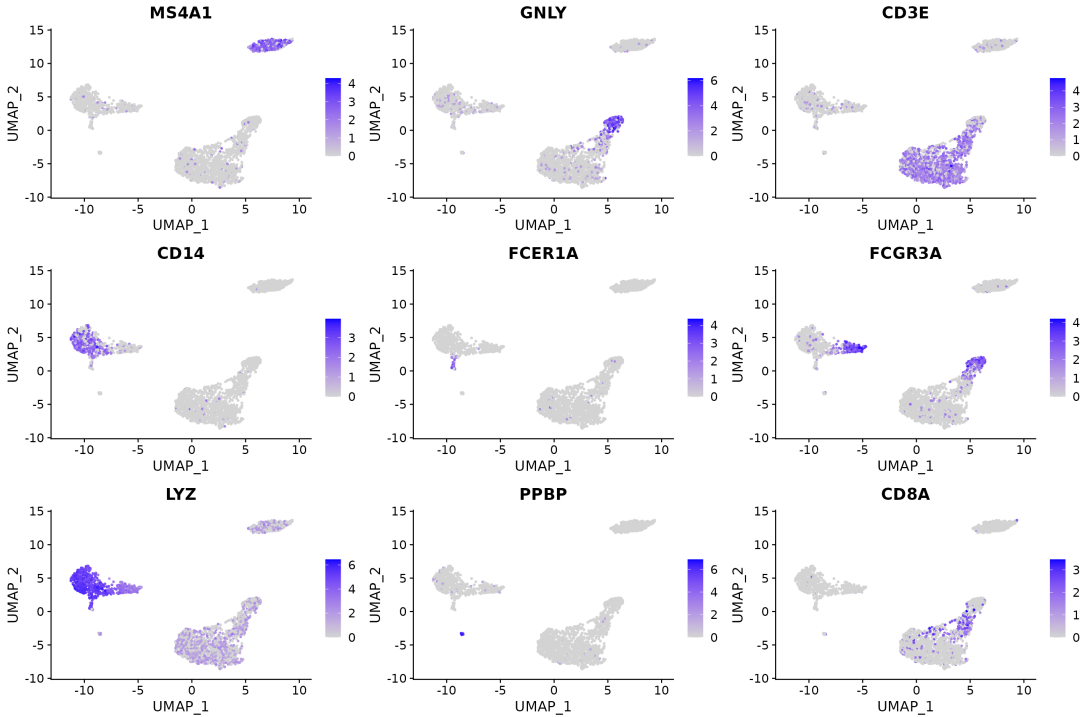
Compares each clusters against all other clusters to identify potential marker genes (differential expression analysis)

Assign cell name to the cluster



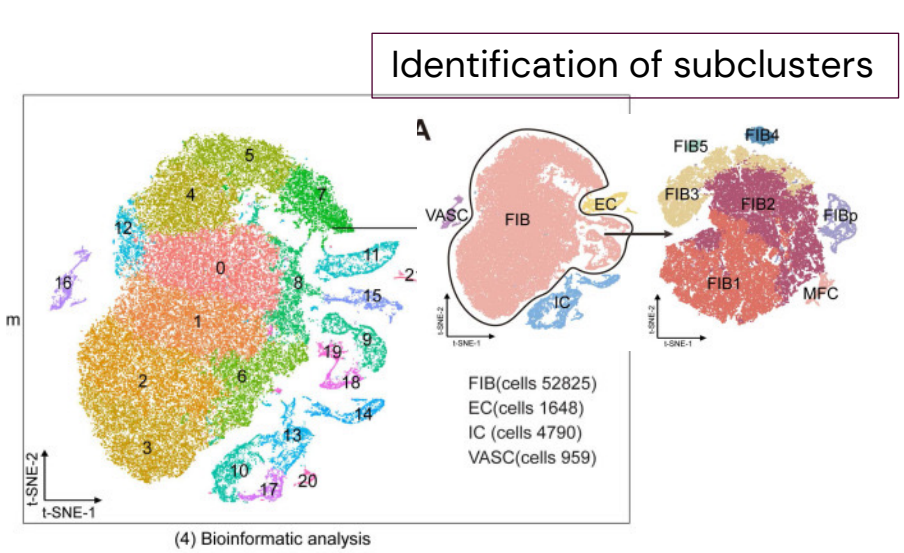


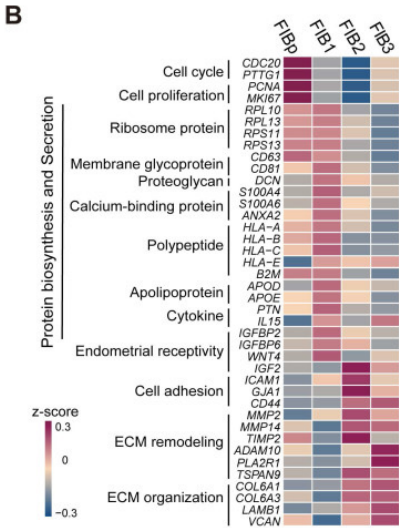
Marker gene expression



What can we investigate with sc-RNAseq – subpopulations of the cells

Comparing gene expression (differentially expressed genes -DEGs) between subclusters and identification of cell subpopulations.

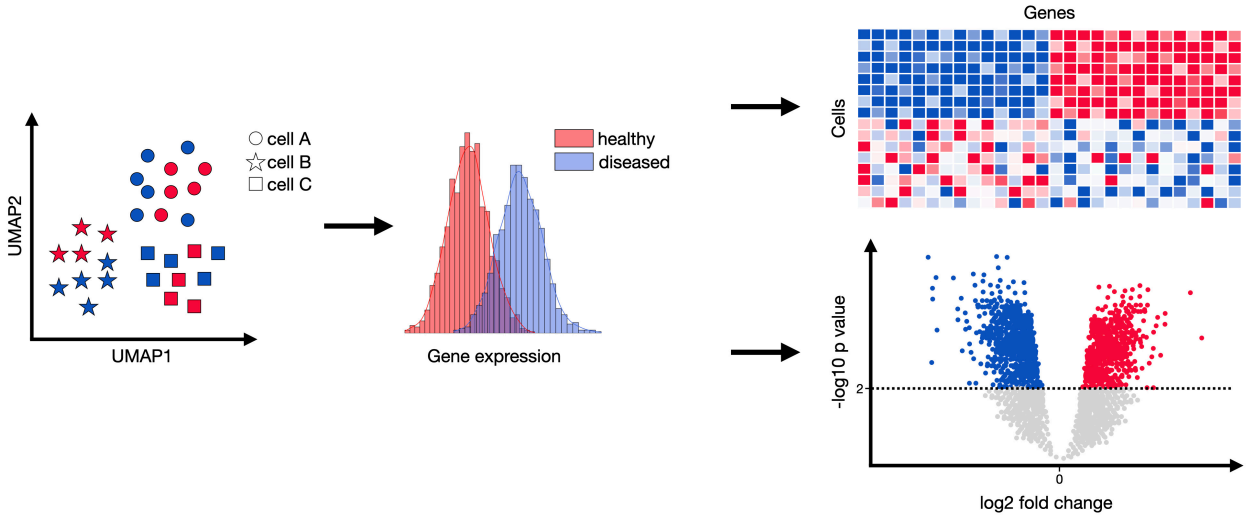




Lai et al 2022 Theranostics

What can we investigate with scRNA-seq – differential gene expression (DEG) analysis

Investigation of molecular variability deriving from different experimental (ctrl:stimulation) or biological conditions (healthy:disease).
- Principle of bulkRNAseq data analysis.



Heumos et al
2023 Nat Rev gen

What can we investigate with scRNA-seq – DEG analysis –> Gene Set Enrichment Analysis (GSEA)

Aim: identify gene programs (biological processes, gene ontologies and molecular pathways) that are overrepresented in the experimental/biological condition compared to control, based on DEGs.

Sourcing information from currated databases

GSEA
https://www.gsea-msigdb.org/gsea/msigdb

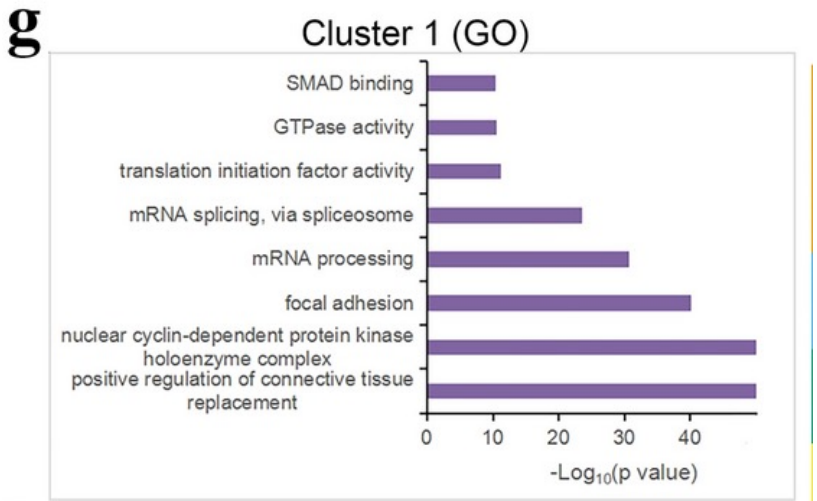
KEGG pathway database
https://www.genome.jp/kegg/pathway.html

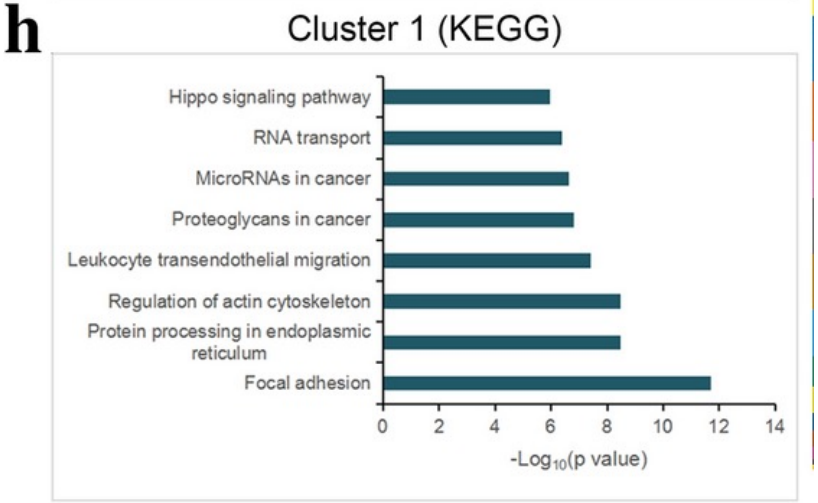
Reactome
https://reactome.org

Gene Ontology (GO)
https://geneontology.org

WikiPathways
https://www.wikipathways.org

Example:

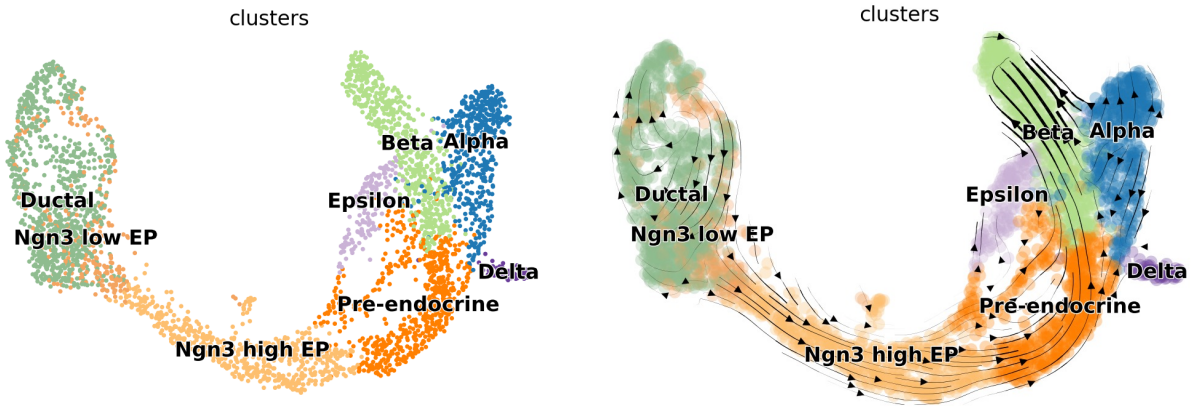

g
Cluster 1 (GO)
SMAD binding
GTPase activity
translation initiation factor activity
mRNA splicing, via spliceosome
mRNA processing
focal adhesion
nuclear cyclin-dependent protein kinase holoenzyme complex
positive regulation of connective tissue replacement
0
10
20
30
40
-Log10(p value)


h
Cluster 1 (KEGG)
Hippo signaling pathway
RNA transport
MicroRNAs in cancer
Proteoglycans in cancer
Leukocyte transendothelial migration
Regulation of actin cytoskeleton
Protein processing in endoplasmic reticulum
Focal adhesion
0
2
4
6
8
10
12
14
-Log10(p value)

What can we investigate with scRNA-seq – RNA velocity

Single cell sequencing is a snapshot – information in the cell is captured at one particular time.

Based on different transcription states of the cells we predict transcriptional dynamics



La Manno et al 2018 Nature

Challenges of scRNAseq

- Large volume of data
  Data output is large, requiring higher amounts of memory to analyse, larger storage requirements and longer run times of analysis

- Low depth of sequencing per cell
  Often detecting only 10-50% of the transcriptome per cell.
  Genes with zero counts – not being expresed or transcripts not detected

- Biological variability across cells/samples
  Biological variation can also be a result of:
  transcriptional bursting (gene transcription not turned on all the time for all the genes)
  continuous or discrete cell identities (e.g. changes in the pro/anti inflammatory status)
  environmental stimuli (local environment of the cell influence gene expression)
  temporal changes (cellular processes – cell cycle can affect gene expression profiles of individual cells)

- Technical variability across cells/samples (batch effects)

Insights into complex diseases

Discovery of **rare cell popularions** in the context of evolution of diseases (tumors)

Cell-type atlases
integration of multi-omics data from data from genomic, epigenomic, transcriptomic and proteomic modalities. (The cancer genome atlas – 33 cancer types over 20,000 primary cancer samples and controls)

Biomedical applications

Tissue and tumor heterogeneity

Investigation of the **differentiation processes**, **tracing cell fate**, identification of new transitional cell states

Novel pathways and network identification

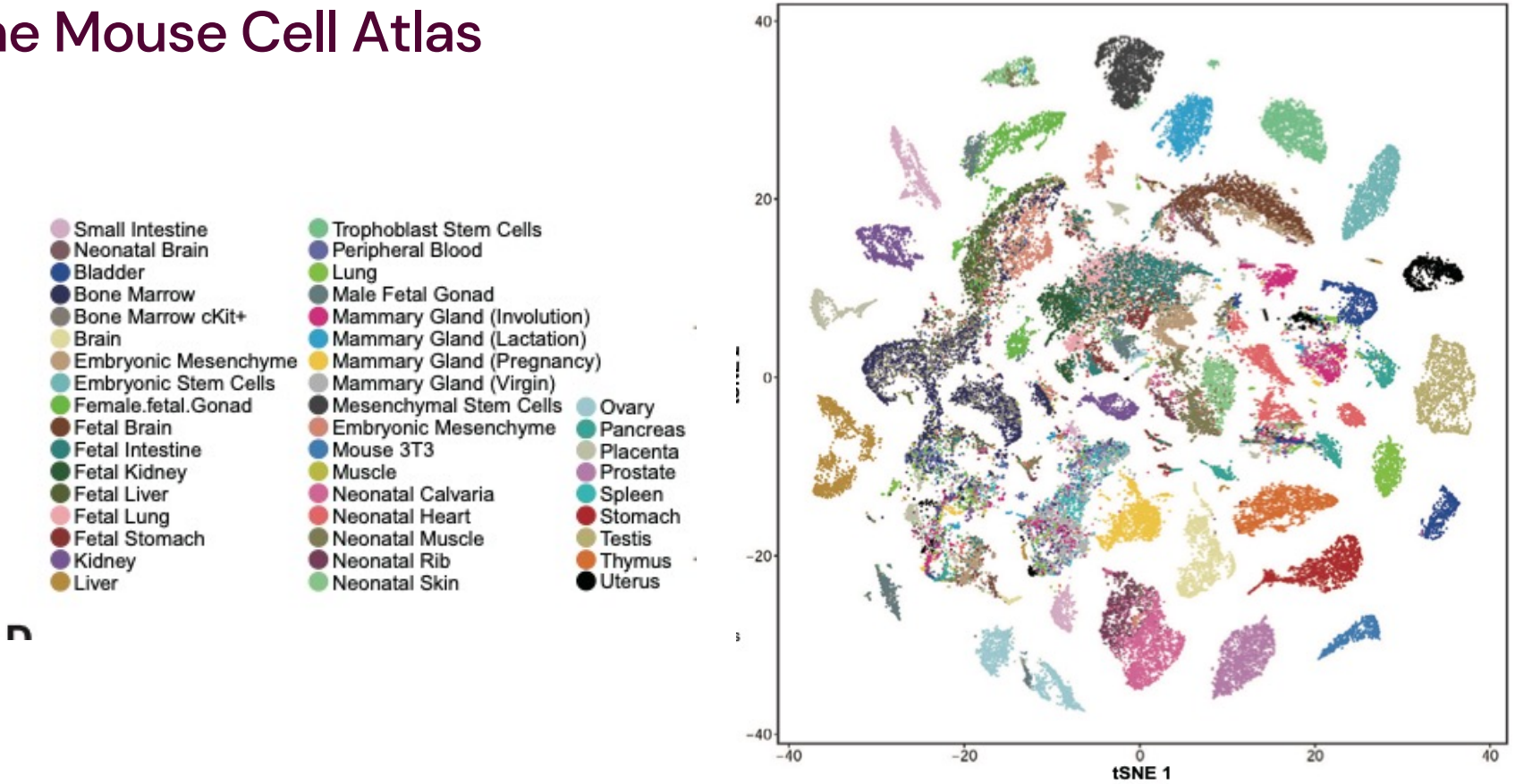
Biomarker discovery

Disease monitoring and progression assessment

Facilitating drug screening for personalized treatment

Baysoy et al 2023 Nat Rev Mol Cell Biol

The Mouse Cell Atlas



Han et al 2018 Cell

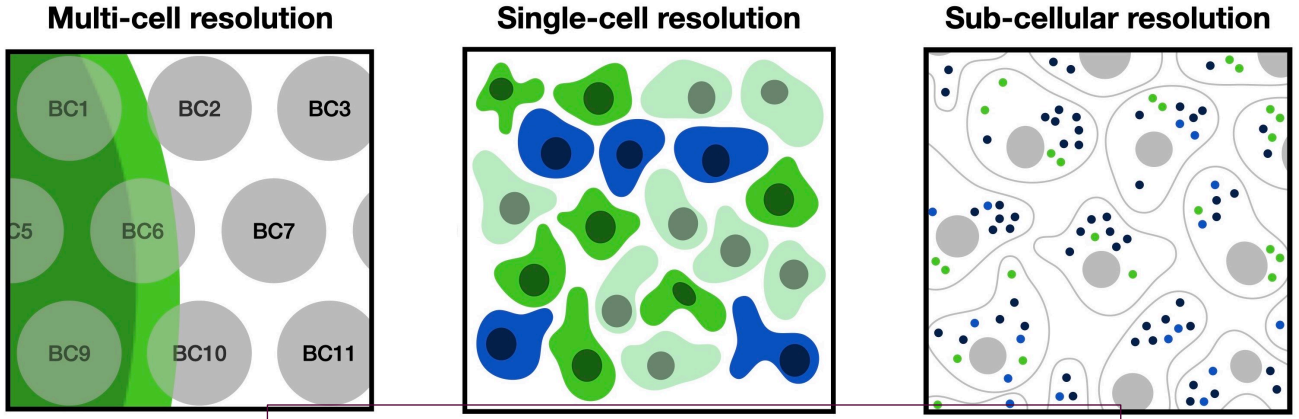
Single-cell data resolved in space

Single-cell genomics technologies enable us to charecterize and identify cellular identities and their dependencies on genomic scale.

Loss of information on spatial organization

Preserving spatial information

Liangchen et al 2023
Computational and Structural Biotechnology Journal



The spot either directly catches single cells or spot on the scale of single-cell.

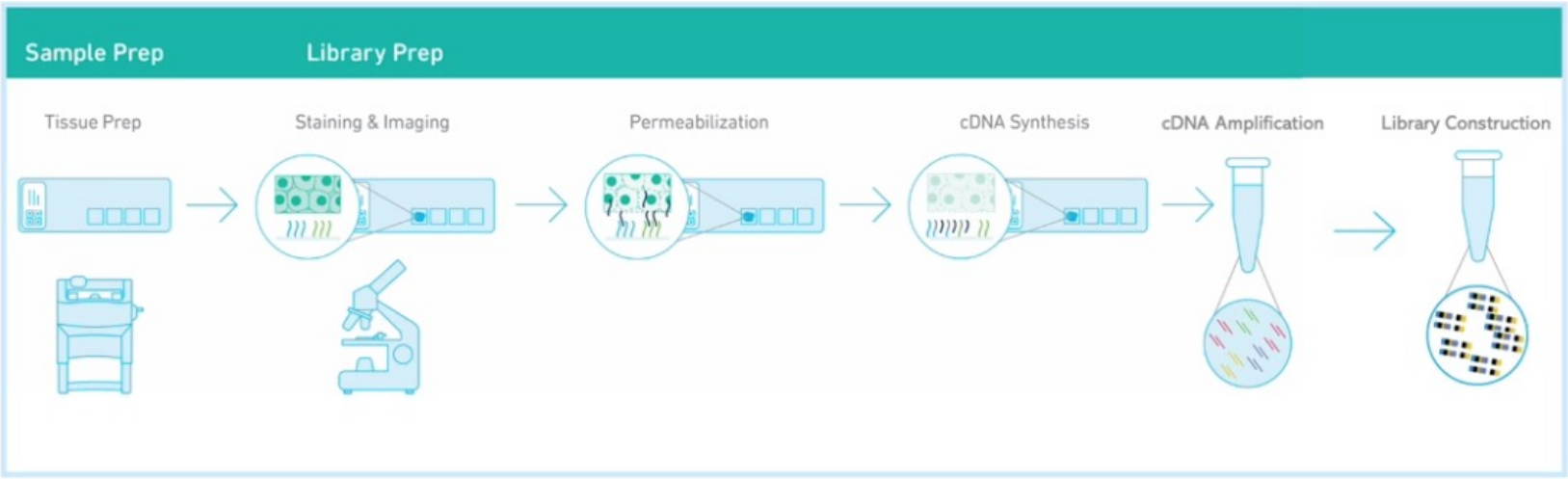
Captures omics measurement among several cells. Each spot captures several cells and can be decomposed with deconvolution methods.

Captures the position of individual RNA Molecules through single-molecule imaging or spatial Barcoding with spots smaller than single-cells.

Deconvolution – computational method that estimates proportion of different cell types in a sample.

Heumos et al 2023 Nat Rev gen

Spatial transcriptomics principle



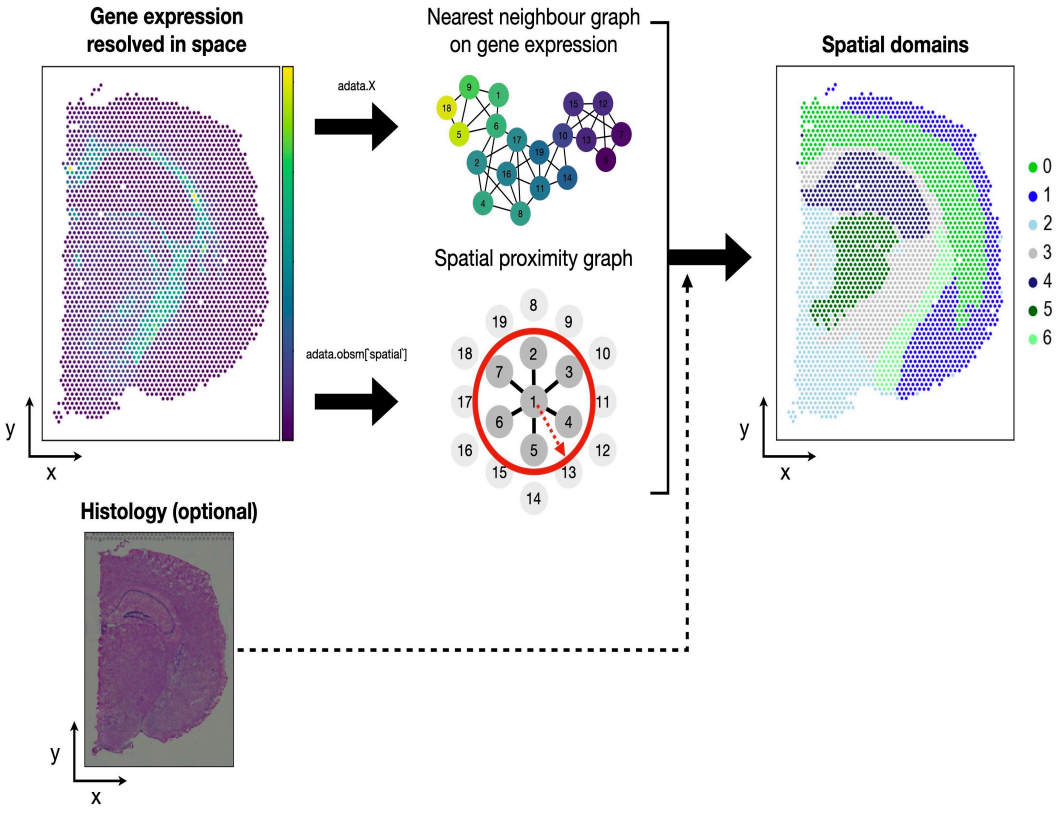
Sample preparation
Loading the tissue section onto the capture area (cryo or FFPE)

Library construction
Permabilization of the tissue (release of mRNA from the cells/spots),
Reverse transcription (barcodes and UMI),

Sequencing and data analysis, integration with other omics

10X Genomics

Spatial transcriptomics enables:



- cellular neighborhoods analysis to understand cellular composition, cell-cell communication.

- Prediction of **spatial domains,** due to orthogonal information aside from gene expression matrix (clusters that account for similarities in gene expression and spatial proximity).

Star Protocols,
Heumos et al 2023 Nat Rev gen

Single cell epigenomics

Transcript levels alone do not reveal the nuances of gene regulation within individual cells.

Epigenomics
DNA accessability, transcription factor binding, DNA and histone modifications, enhancer-promoter contact maps and 3D genome organization provide insight into promoter regulation of gene expression.

Epigenomics give deep insight into biological function, reveals disease associated pertubations/ transcription factors and aging processes.

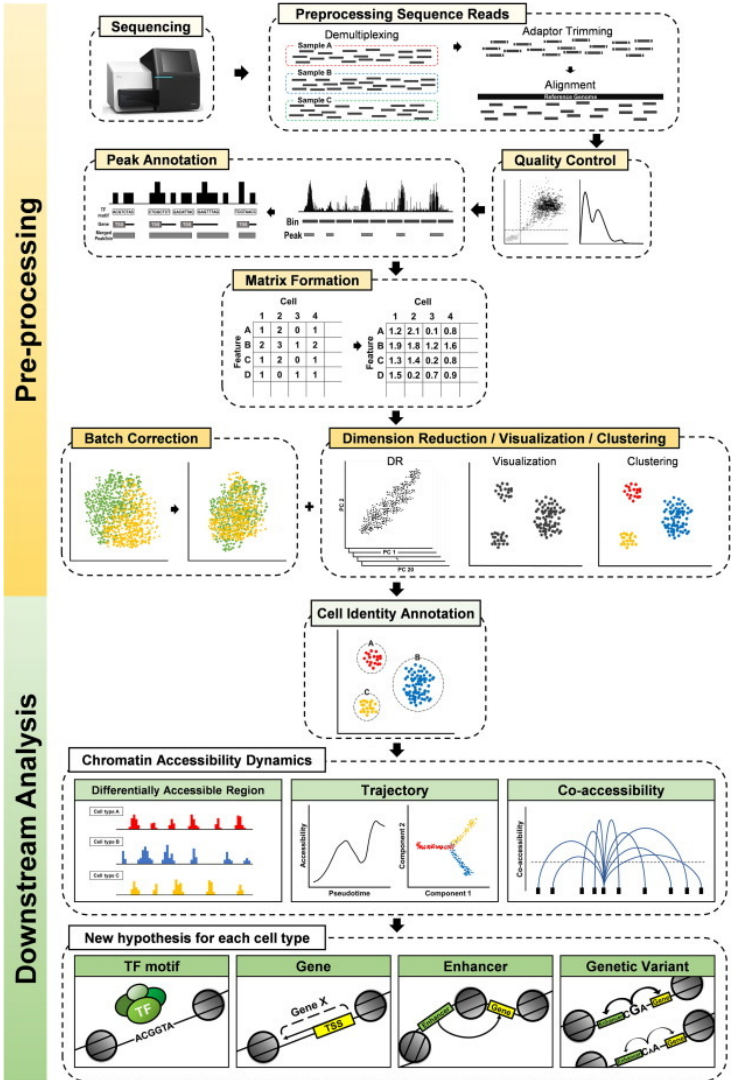
Most widely used epigenomic investigation in single cell = Assay for Transposase Accessible Chromatin sequencing (**ATAC-seq**)

scATAC-seq

Principle: transposase-mediated insertion of sequencing primers into open regions in the chromatin (promoters, enhancers and TF binding sites).

scATAC-seq provides information on chromatin accessible regions that indicate active regulatory regions at single cell resolution.

Seungbyn et al 2020



Application

Identification of specific **TFs** for different cell types within heterogeneous cell population and their role in cellular differentiation.

Identification of **enhancers** (cis regulatory elements) and their proximal and distal interacrions with other regulatory elements.

Relating epigenetic features to GWAS enables identification of **SNP** and **eQTL** (as genetic variants) is possible due to their association with cis-regulatory elements.

Take home message

Further reading

- Single cell best practices
  https://www.sc-best-practices.org/preamble.html
- Biostar Handbook of Bioinformatics
  https://www.biostarhandbook.com
- Computational biology lecture
  https://github.com/pachterlab/Bi-BE-CS-183-2023/#readme
- HBC training github
  https://github.com/hbctraining/scRNA-seq_online/blob/master/lessons/02_SC_generation_of_count_matrix.md
- UMAP
  https://umap-learn.readthedocs.io/en/latest/how_umap_works.html
  https://pair-code.github.io/understanding-umap/
- PCA explanation
  https://www.youtube.com/watch?v=_UVHneBUBW0

- Questions? -> tina.gorsek@ki.se


KAROLINSKA INSTITUTET
ANNO 1810
Karolinska
Institutet