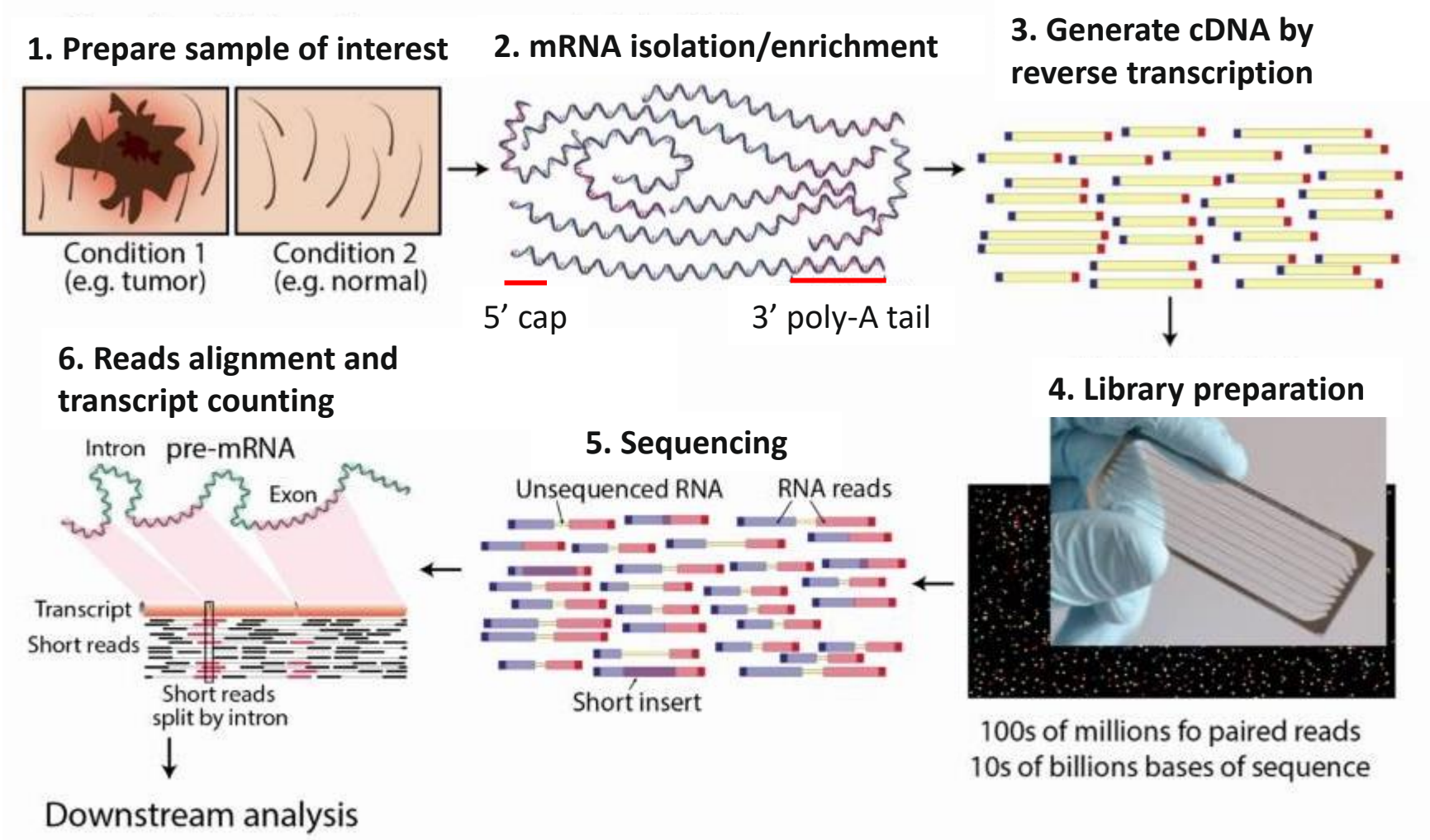
An Introduction of bioinformatic approach to single-cell RNA sequencing

Presented by Chen-Yang Huang
Steve Rozen Lab, Center of Computational Biology, Duke-NUS Medical School

2022-01-24

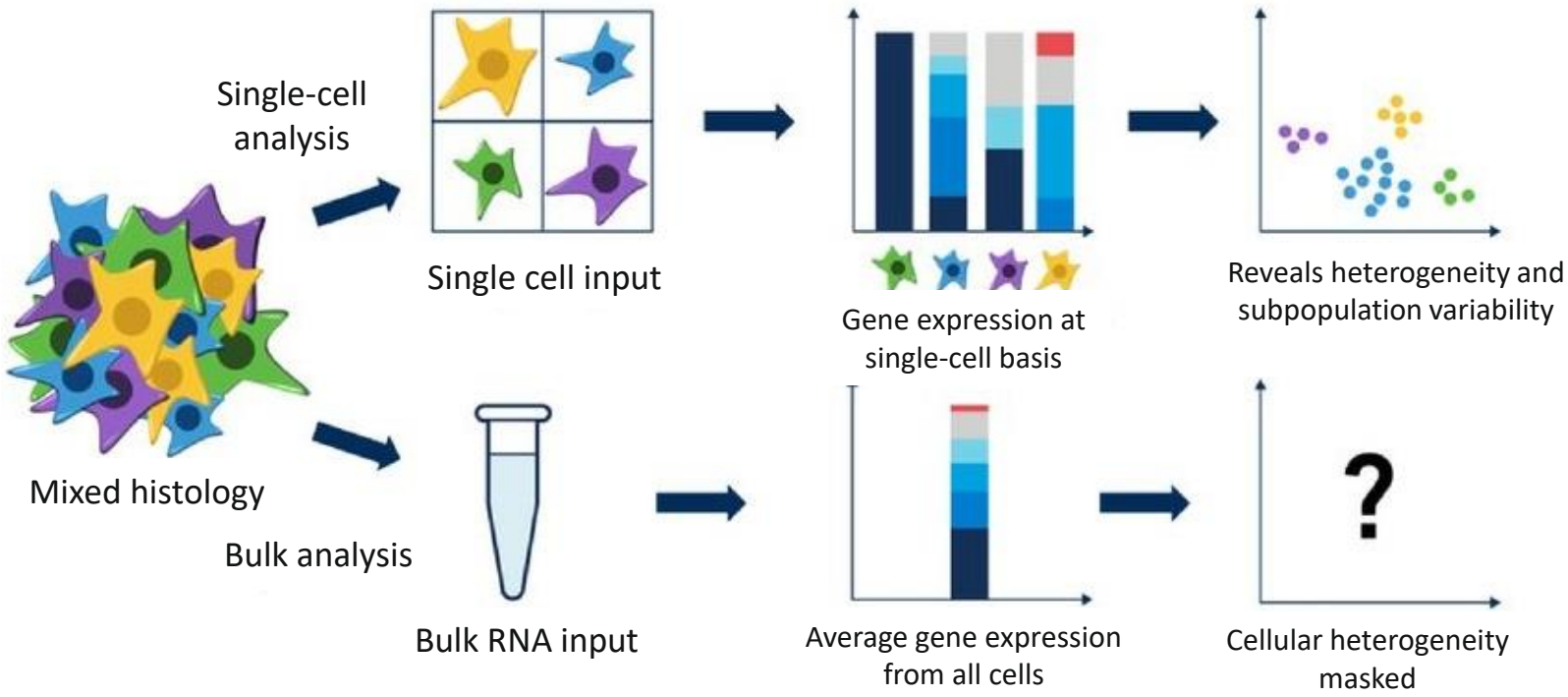
RNA sequencing: a major breakthrough to study transcriptome



From bulk RNA to single-cell RNA sequencing (scRNAseq): filling the research gap

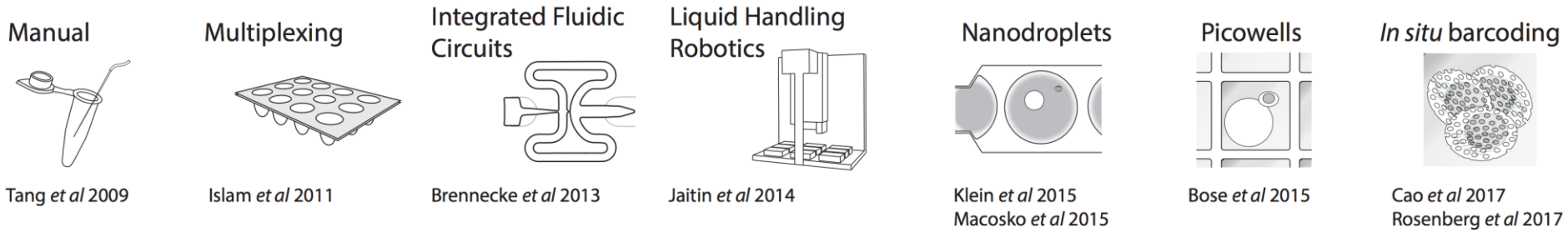
Bulk RNA seq:

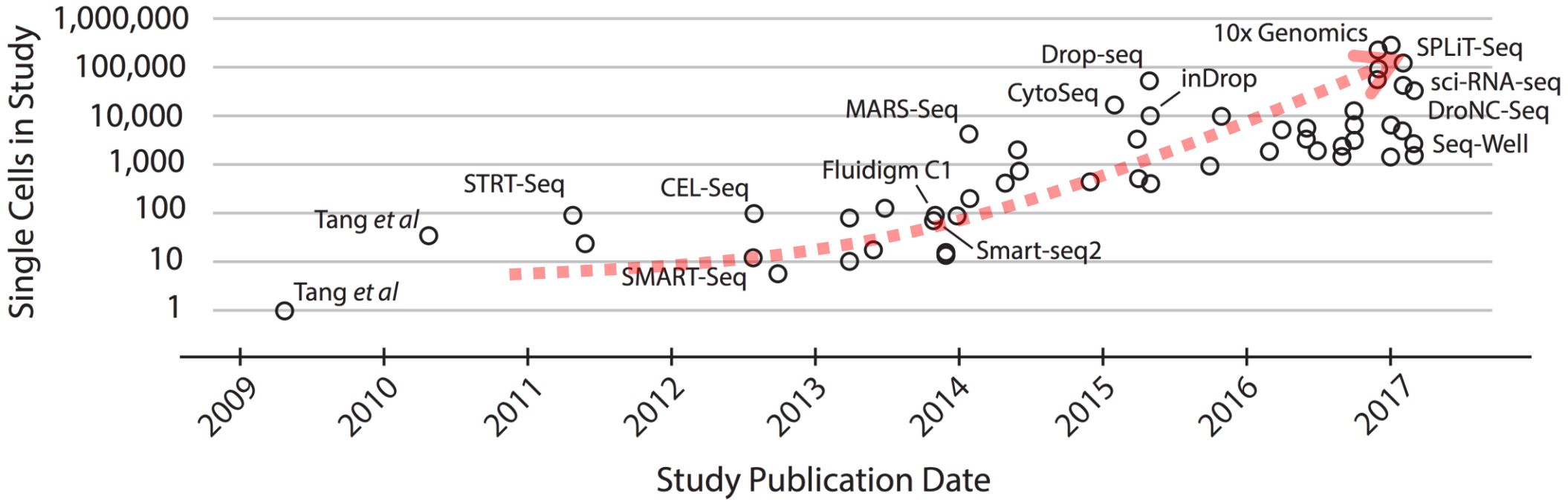
- Measures the average expression level for each gene across a large population of input cells
- Insufficient for studying heterogeneous systems, e.g. early development studies, complex tissues (brain, immune cells)
- Does not provide insights into the stochastic nature of gene expression



https://www.10xgenomics.com/

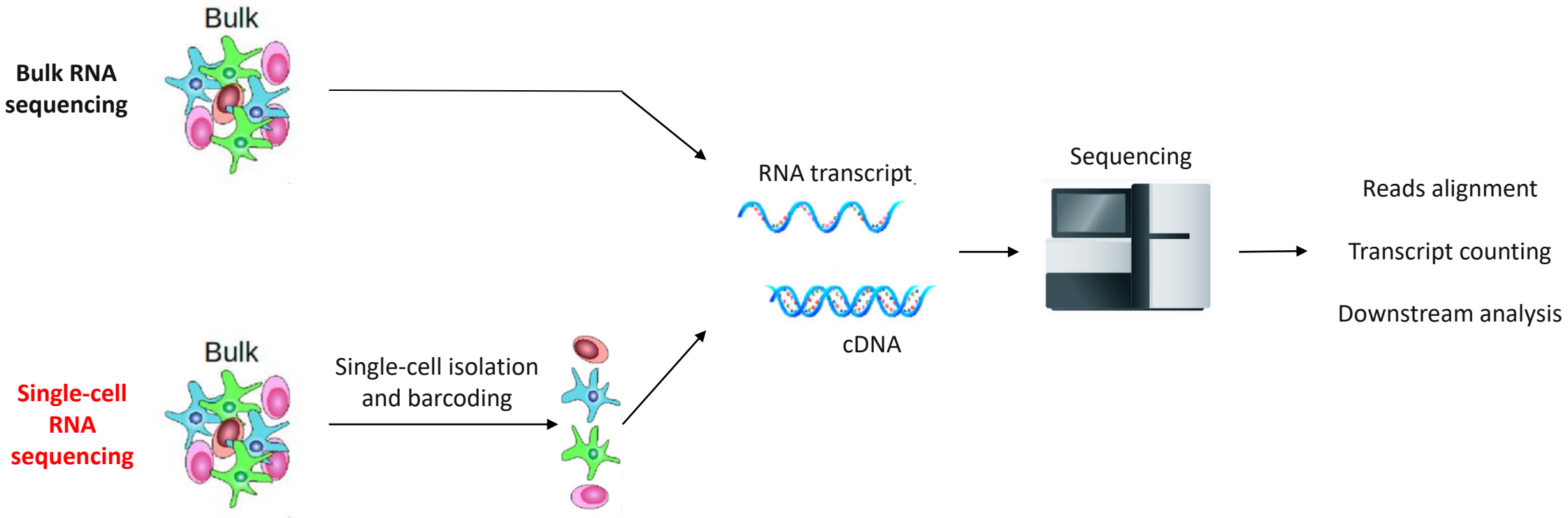
Moore's law in scRNAseq: exponential growth of sequencing capacity in the last decade





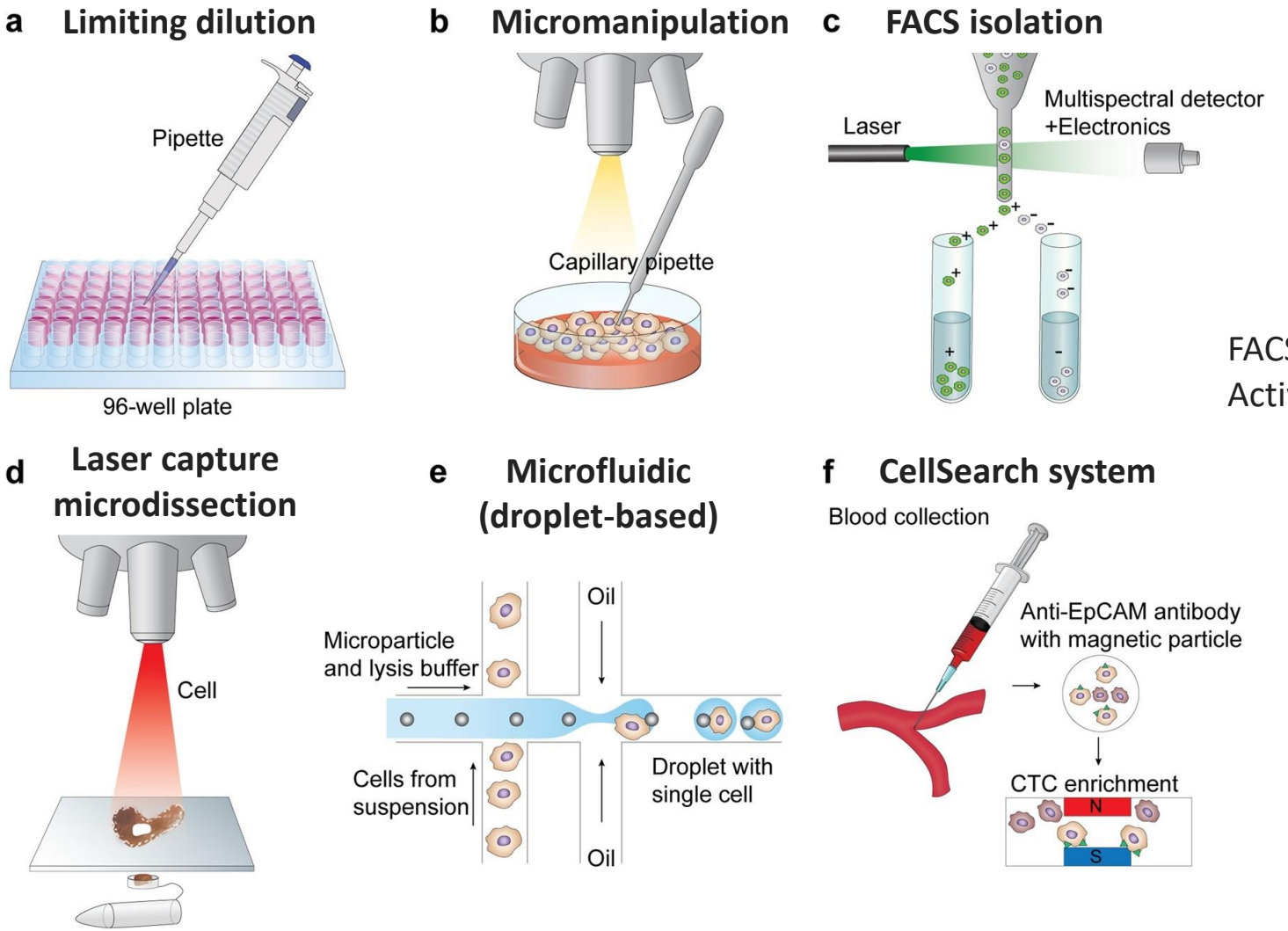
Svensson, 2017

Integration of cell-specific barcode into bulk RNA sequencing technology



Hwang, 2018

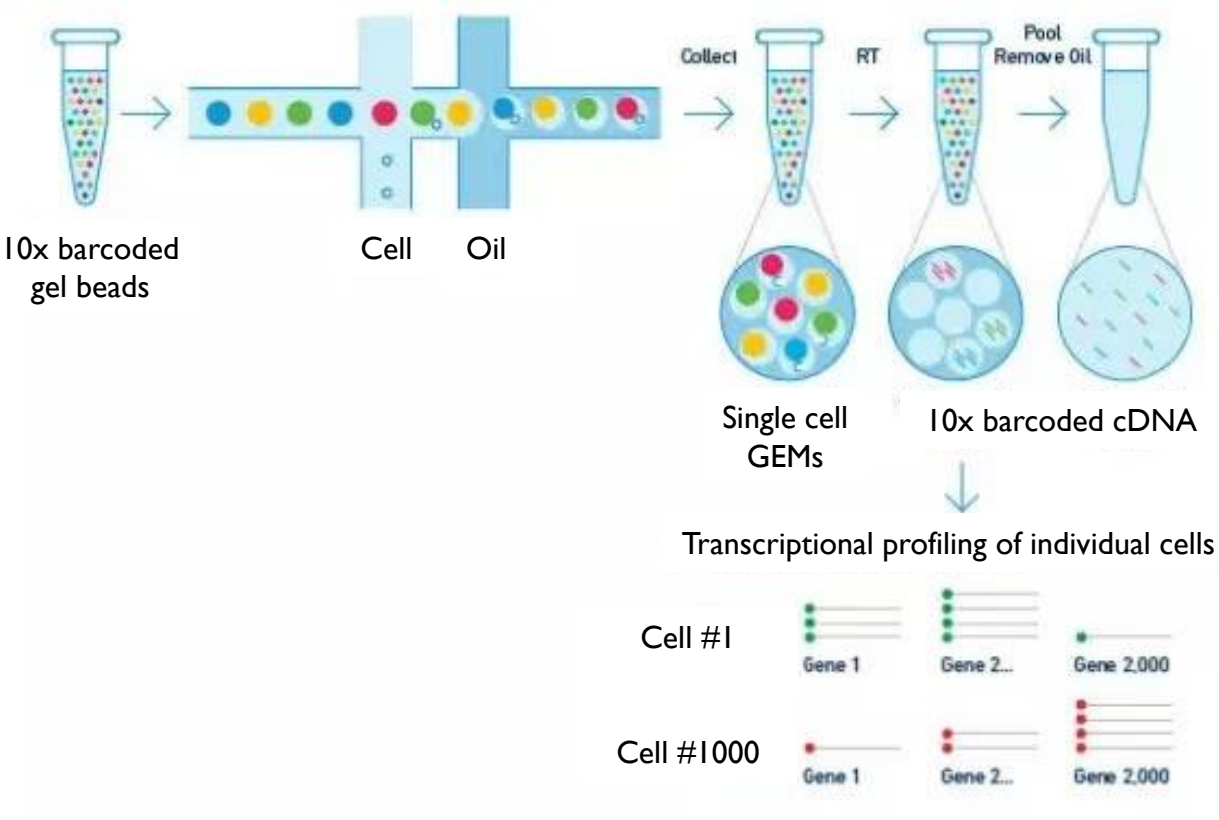
Single cell isolation methods

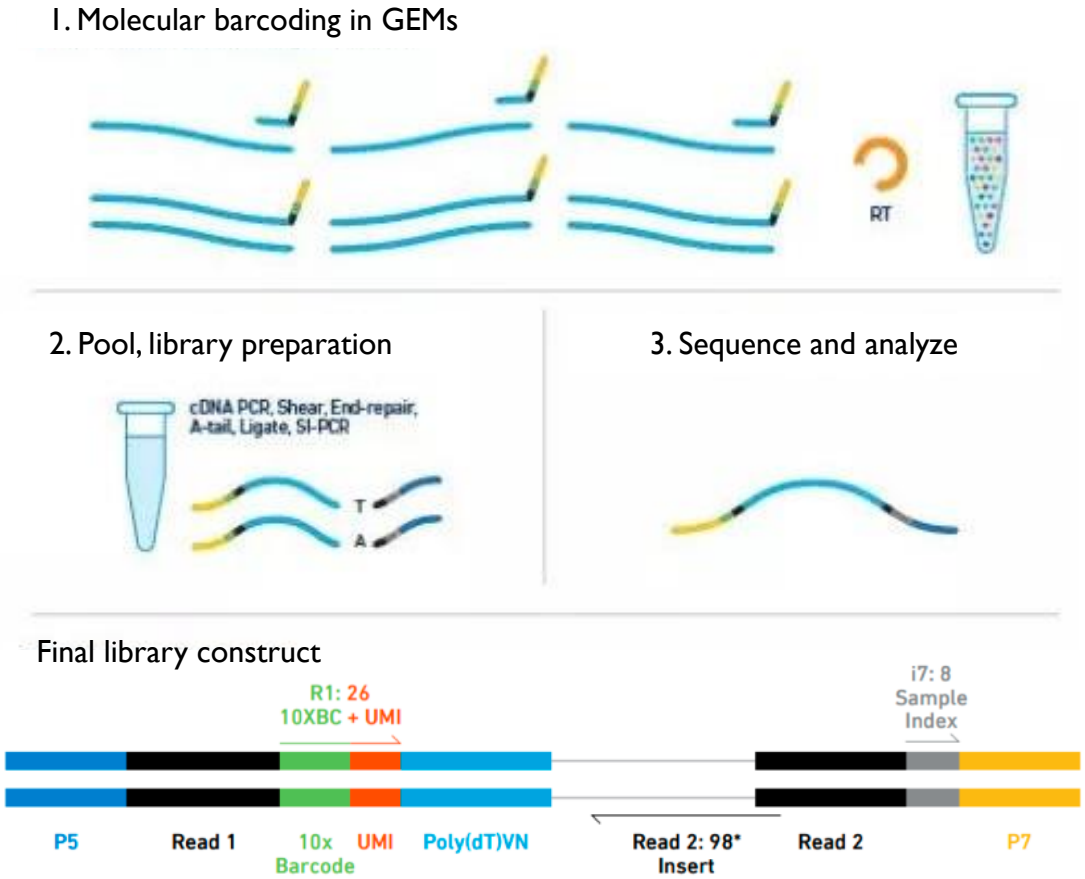


FACS: Fluorescence-Activated Cell Sorting

Hwang, 2018

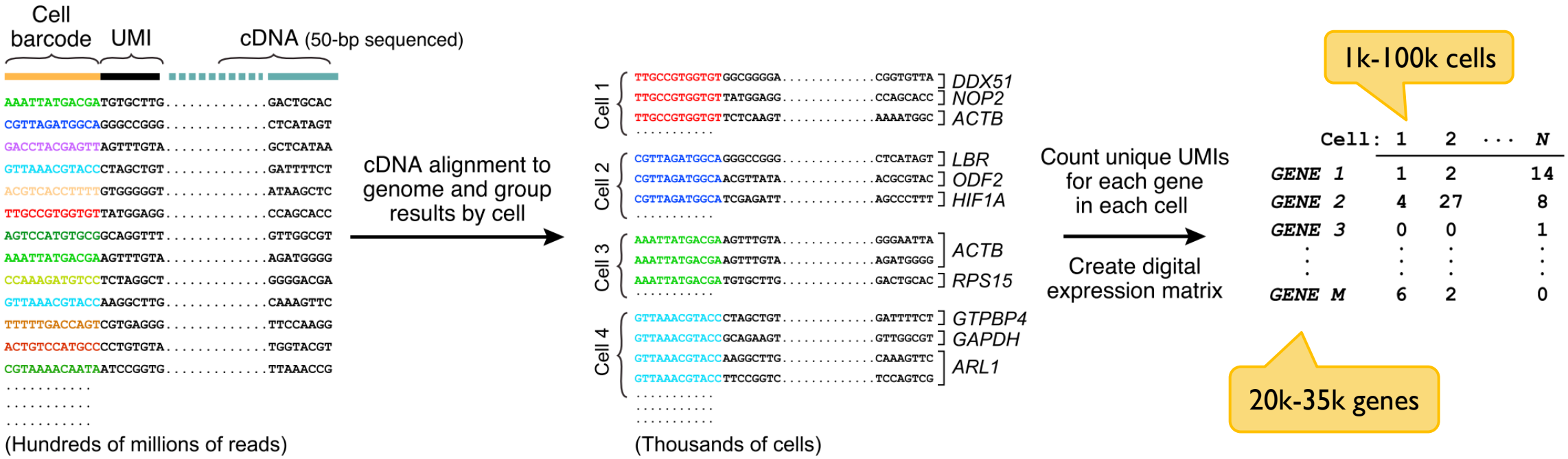
Droplet-based single cell transcriptomic profiling and cell barcoding





https://www.10xgenomics.com/

scRNAseq: reads alignment and transcript counting



UMI: Unique molecular identifier

Macosko EZ, Cell 2015

Comparison of common scRNAseq protocols

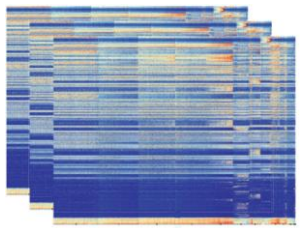
	SMART-seq2	CEL-seq2	STRT-seq	Quartz-seq2	MARS-seq	Drop-seq	inDrop	Chromium	Seq-Well	sci-RNA-seq	SPLiT-seq
Single-cell isolation	FACS, microfluidics	FACS, microfluidics	FACS, microfluidics, nanowells	FACS	FACS	Droplet	Droplet	Droplet	Nanowells	Not needed	Not needed
Second strand synthesis	TSO	RNase H and DNA pol I	TSO	PolyA tailing and primer ligation	RNase H and DNA pol I	TSO	RNase H and DNA pol I	TSO	TSO	RNase H and DNA pol I	TSO
Full-length cDNA synthesis?	Yes	No	Yes	Yes	No	Yes	No	Yes	Yes	No	Yes
Barcode addition	Library PCR with barcoded primers	Barcoded RT primers	Barcoded TSOs	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers	Barcoded RT primers and library PCR with barcoded primers	Ligation of barcoded RT primers
Pooling before library?	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Library amplification	PCR	In vitro transcription	PCR	PCR	In vitro transcription	PCR	In vitro transcription	PCR	PCR	PCR	PCR
Gene coverage	Full-length	3'	5'	3'	3'	3'	3'	3'	3'	3'	3'
Number of cells per assay											

Chen X, 2018

General workflow of computational approach to scRNAseq data

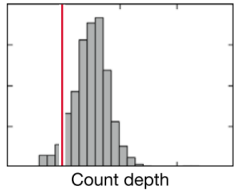
Raw data processing

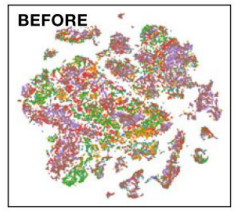
Count matrices

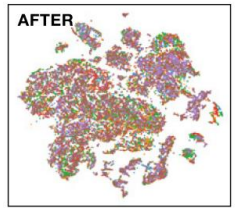


Quality control

Data correction (batch effect)



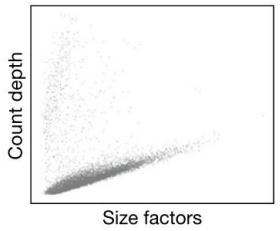


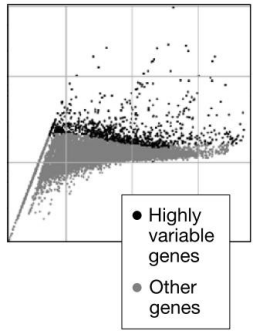


Normalization

Feature selection

Dimensionality Reduction

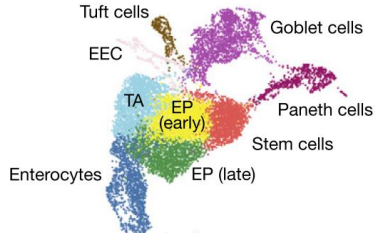


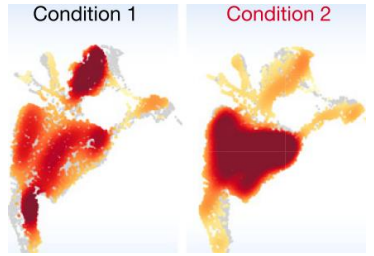


Clustering

Marker identification

Cluster annotation

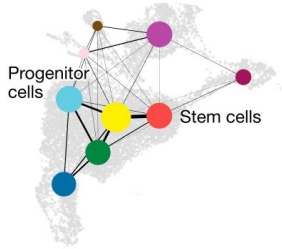


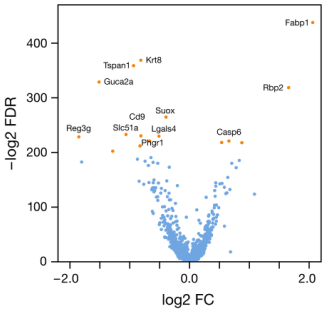


Downstream analysis

Trajectory inference

Differential expression





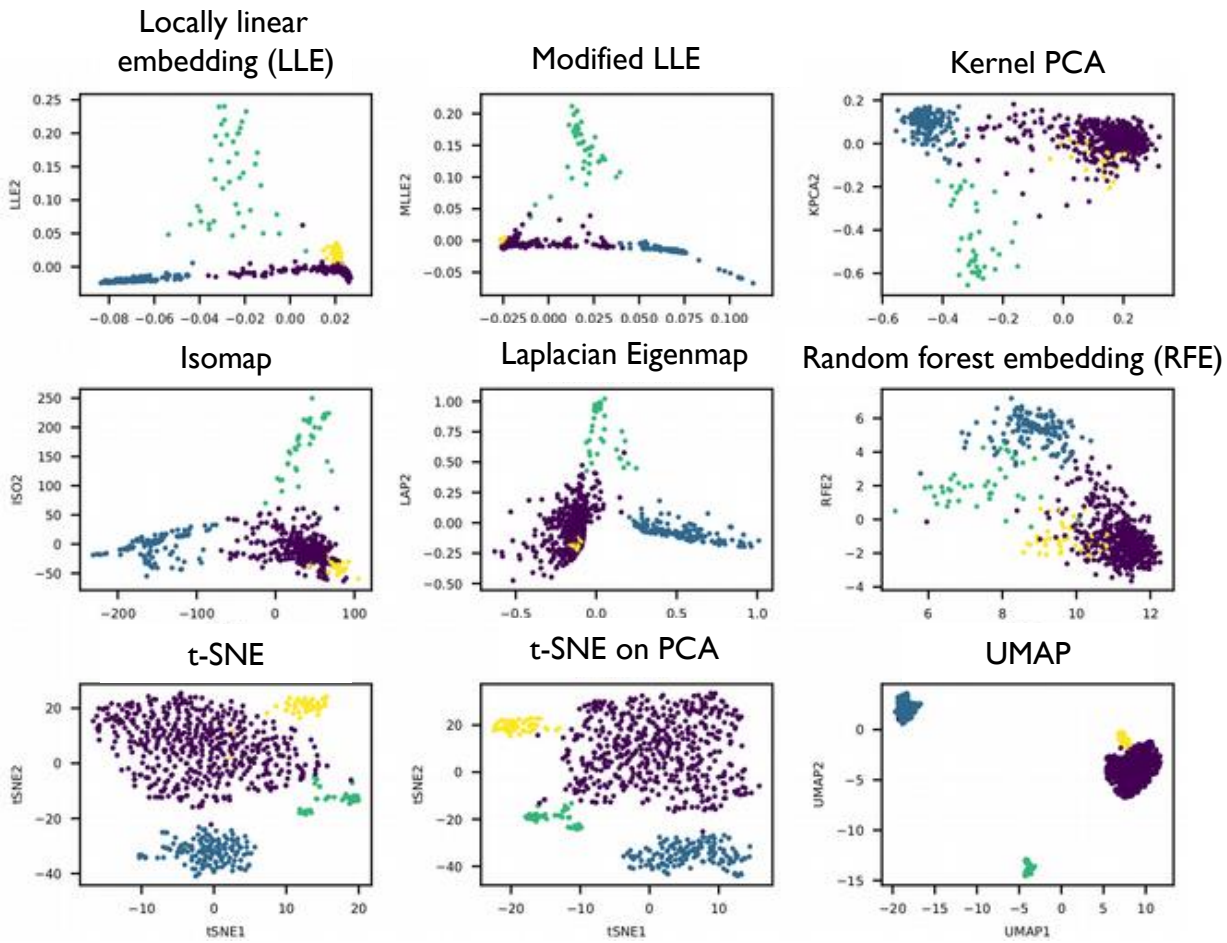
Kiselev VY, Nat Rev Genetics 2019.

Dimensionality reduction in scRNAseq data

Goal of dimensionality reduction:

- Transform the dataset into a more compact and interpretable representation
- Improve the performance of downstream analysis

- Linear methods:
  - Principal component analysis (PCA)
  - Zero inflated factor analysis (ZIFA)
  - Non-negative matrix factorization (NMF)
- Non-linear methods:
  - t-distributed stochastic neighborhood embedding (t-SNE)
  - Uniform manifold approximation and projection (UMAP)
  - Diffusion maps



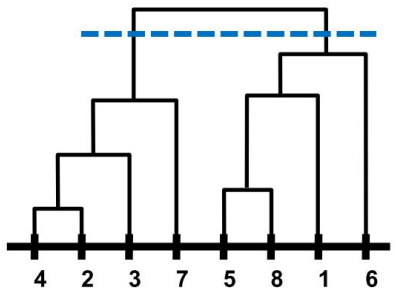
https://towardsdatascience.com/

Unsupervised clustering in scRNAseq

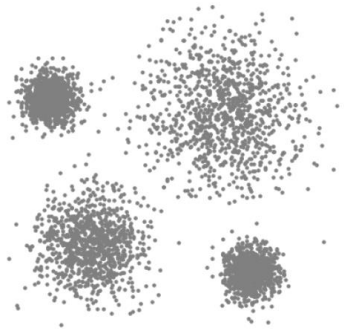
Goal of unsupervised clustering:

- using coherent and unbiased approach to discover the natural groupings of a set of dataset

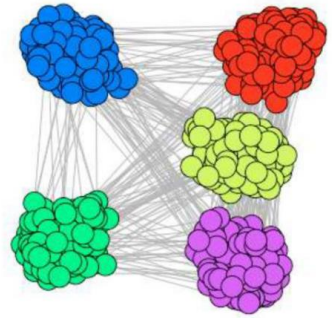
Types of method:



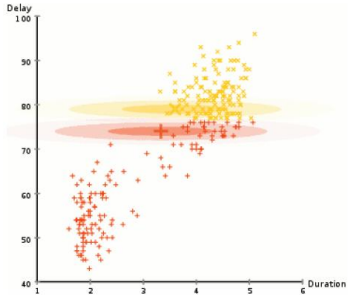
Hierarchical Clustering



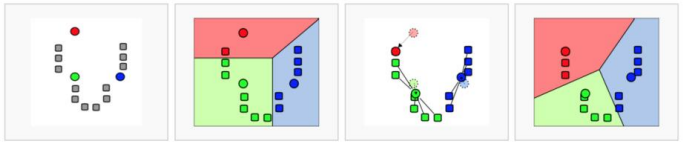
Mean shift clustering



Graph-based clustering



Gaussian mixture modeling



k-means clustering

UMAP produces faster and more meaningful clustering results than t-SNE

t-distributed stochastic neighborhood embedding (t-SNE):

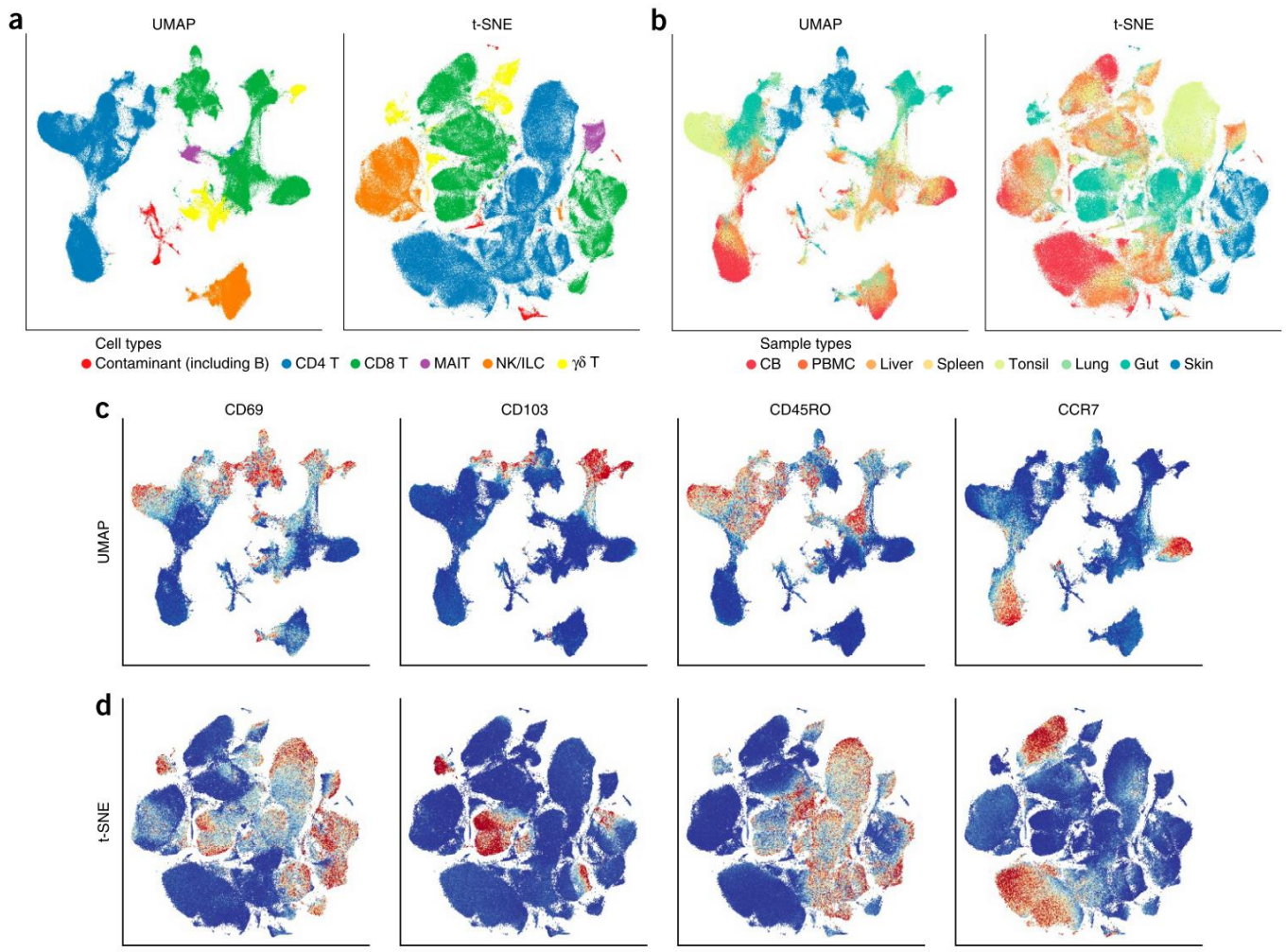
- prioritizes the local structure of the data, weaken the global properties of the datasets

Uniform manifold approximation and projection (UMAP):

- generates a nearest neighbours graph of the cells, weighting each cell–cell connection by the strength of similarity (global structure)

UMAP vs. t-SNE

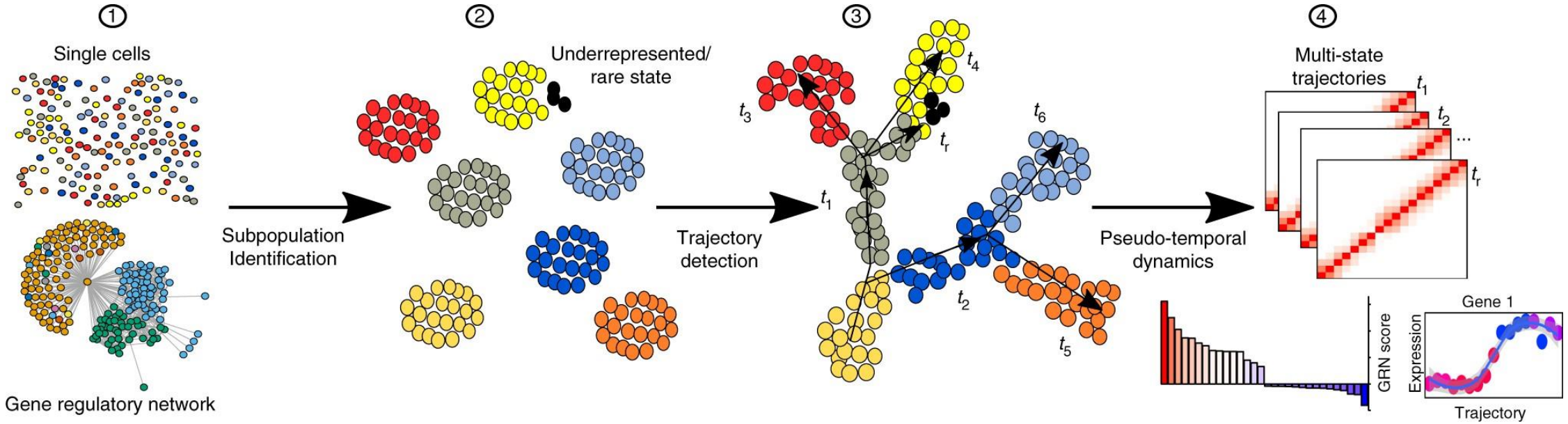
- faster run times with higher reproducibility
- more meaningful organization of cell clusters



Becht E, Nat Biotech 2018.

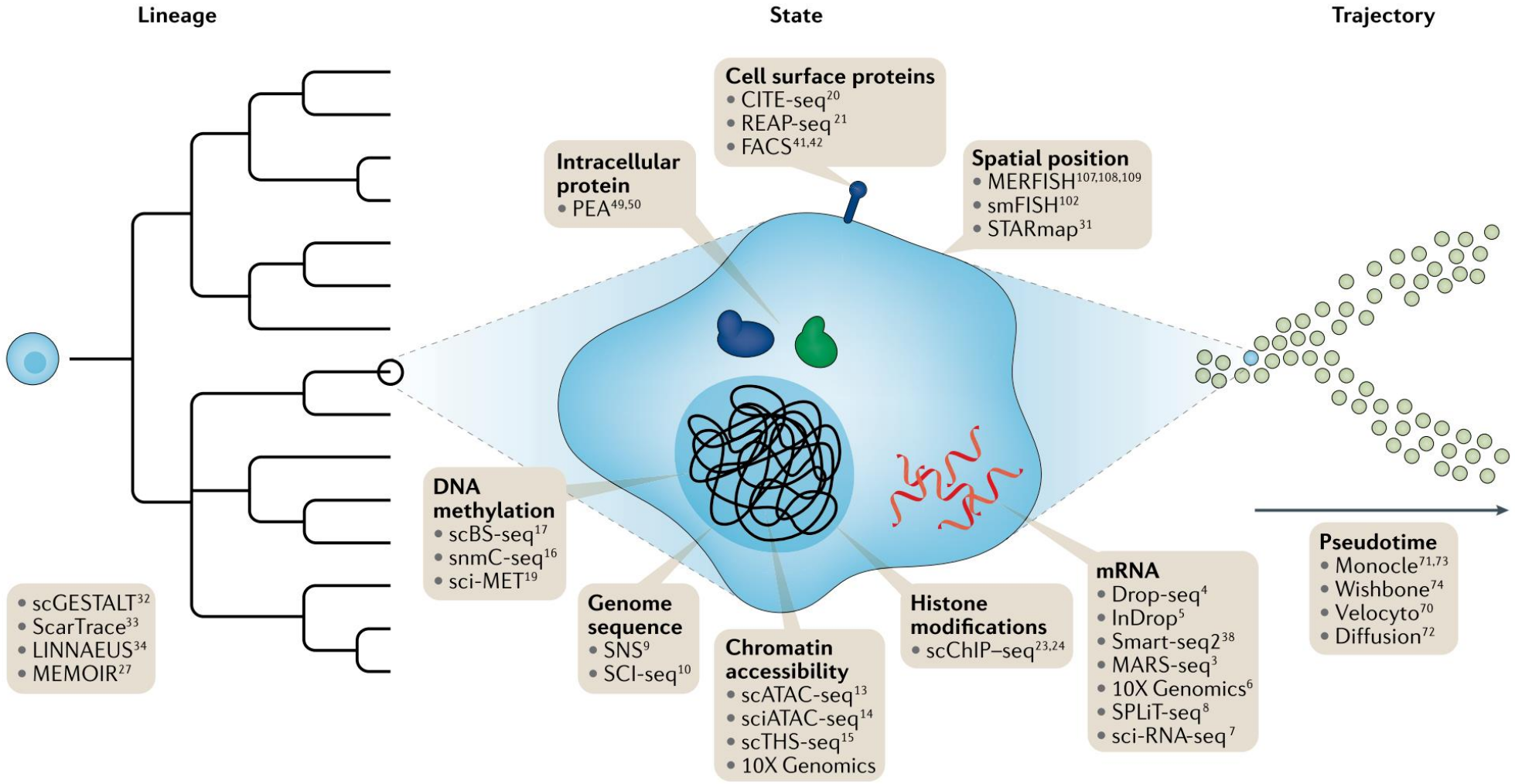
Single-cell trajectory inference and pseudotime analysis: finding a developmental track on transcriptomic data

- Trajectory inference:
  - order cells along a trajectory based on similarities in their expression patterns
  - can be linear, bifurcating, or more complex topologies
  - offer an transcriptome-wide understanding of a dynamic process, allowing the objective identification of new (primed) subsets of cells



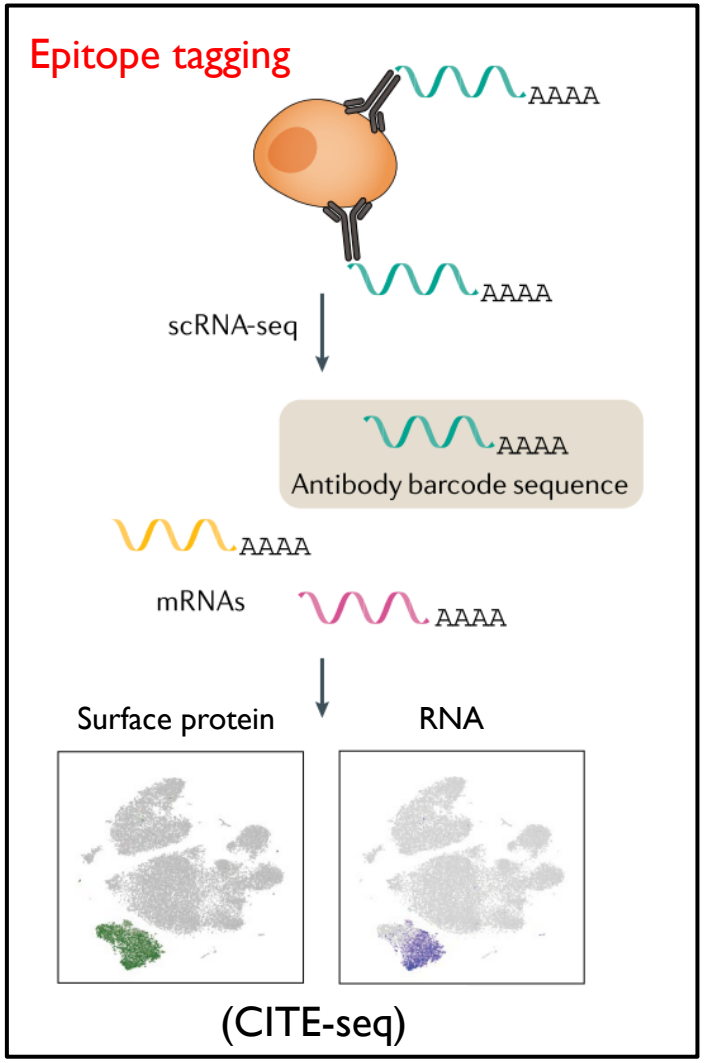
Rocha EL, Nat Biotech 2018

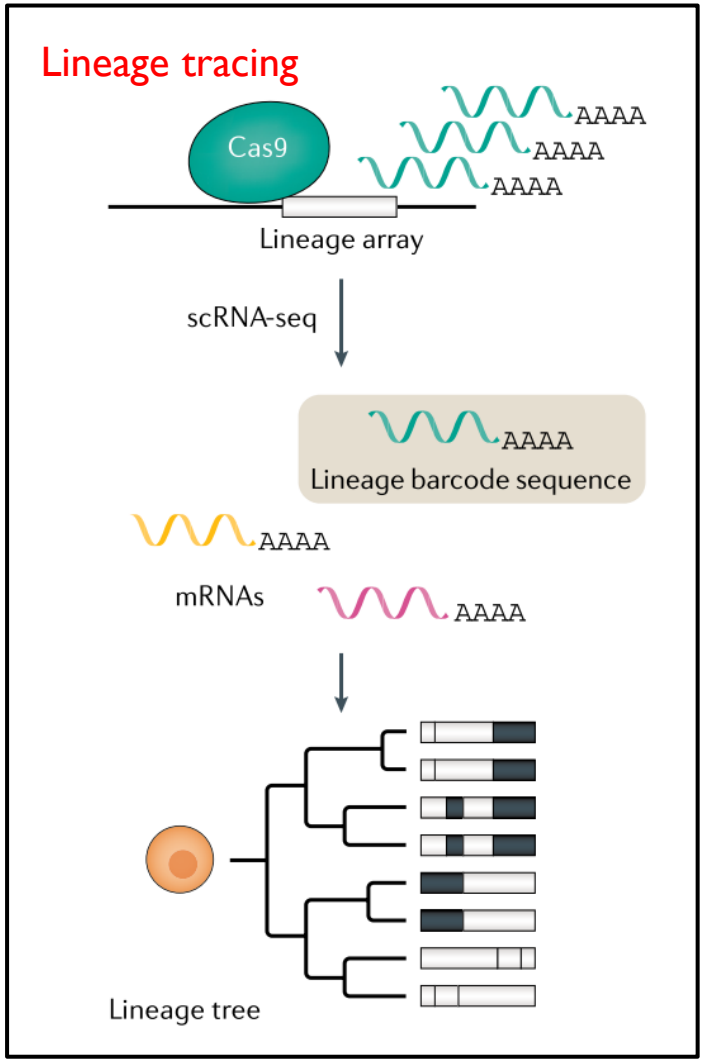
Beyond the boundary of transcriptome: multimodal and integrative methods for single-cell analysis

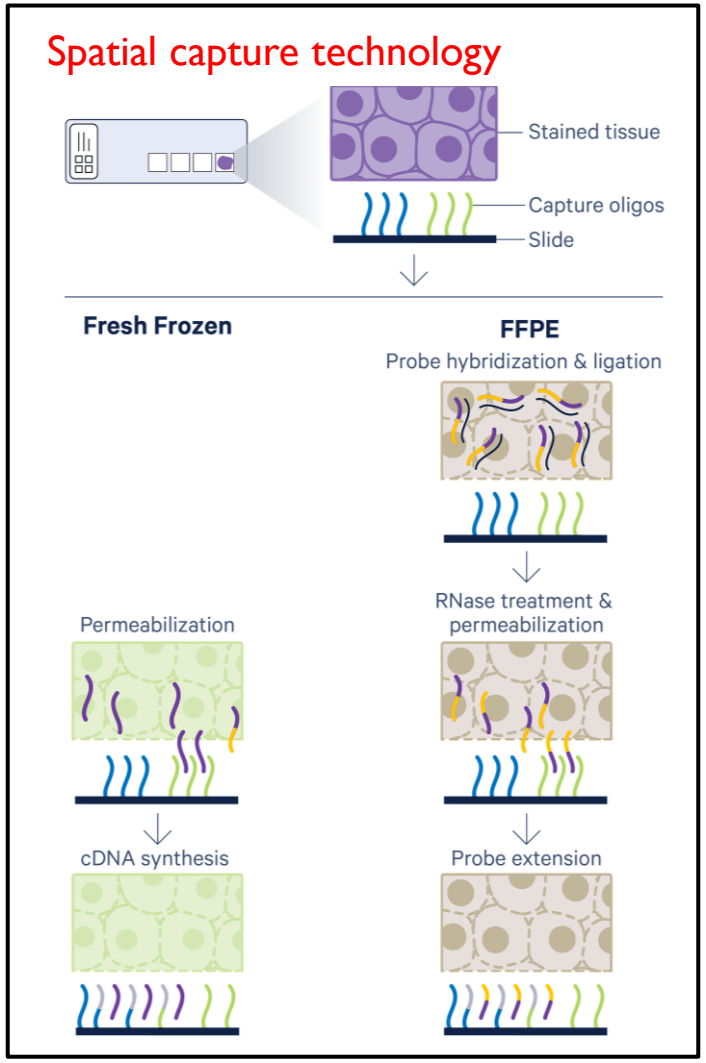


Stuart T, 2019

Examples of innovative barcoding strategies used in multimodal single-cell sequencing technology







Stuart T, 2019

Frequently encountered problems during scRNAseq data analysis

- Differentiating empty droplet, singlet and doublet (two cells captured in one droplet)
  - Setting up a threshold using UMI counts, number of distinct gene per cell…
  - Other computational approach: 'CellBender', 'DoubletFinder'…
- Choosing parameters to filter out low-quality cells
  - Mitochondrial ratio, ribosomal ratio…
  - Other computational approach: 'miQC'…
- Correcting batch effect before integrating multiple scRNAseq datasets into one
  - Mutual nearest neighbors (MNN), canonical correlation analysis (CCA), variational autoencoder (VAE)
- Correcting cell cycle variation from scRNAseq data

Bioinformatic tools for analysis of scRNAseq data

Toolkit	Main function	Platform	Weblink
Cellranger	Fastq processing	Linux	https://support.10xgenomics.com/single-cell-gene-expression/software/pipelines/latest/installation
DropEst	Fastq processing	Linux	https://github.com/hms-dbmi/dropEst
Seurat	Dimension reduction, clustering, differential expression	R	https://satijalab.org/seurat/
Monocle	Dimension reduction, clustering, trajectory inference	R	http://cole-trapnell-lab.github.io/monocle-release/
Scanpy	Dimension reduction, clustering, differential expression	Python	https://scanpy.readthedocs.io/
DEsingle	Differential expression	R	https://miaozhun.github.io/DEsingle/
scCATCH	Cell type annotation	R	https://github.com/ZJUFanLab/scCATCH
Slingshot	Trajectory inference	R	https://bioconductor.org/packages/slingshot/
GeneSwitches	Differential expression on pseudotime	R	https://geneswitches.ddnetbio.com/

* More about public scRNAseq dataset: http://support.10xgenomics.com/single-cell/datasets

- Thank you & stay safe!

- My email address:
  e0470761@u.duke.nus.edu

- Steve Rozen Lab:
- Prof. Steve Rozen
- Faculties:
  - Yu Shun Shing, Willie
  - Wu Yang
  - Jiang Nanhai
  - Huang Chen Yang
  - Zhang Yinuo
  - Zhang Shenli
  - Ho Szu-Chi
  - Elizabeth Tan
  - Chia Huan Chern