3000788 Intro to Comp Molec Biol

Lecture 13: Single-cell and spatial omics

Fall 2025

Sira Sriswasdi, PhD

- Research Affairs
- Center of Excellence in Computational Molecular Biology (CMB)
- Center for Artificial Intelligence in Medicine (CU-AIM)

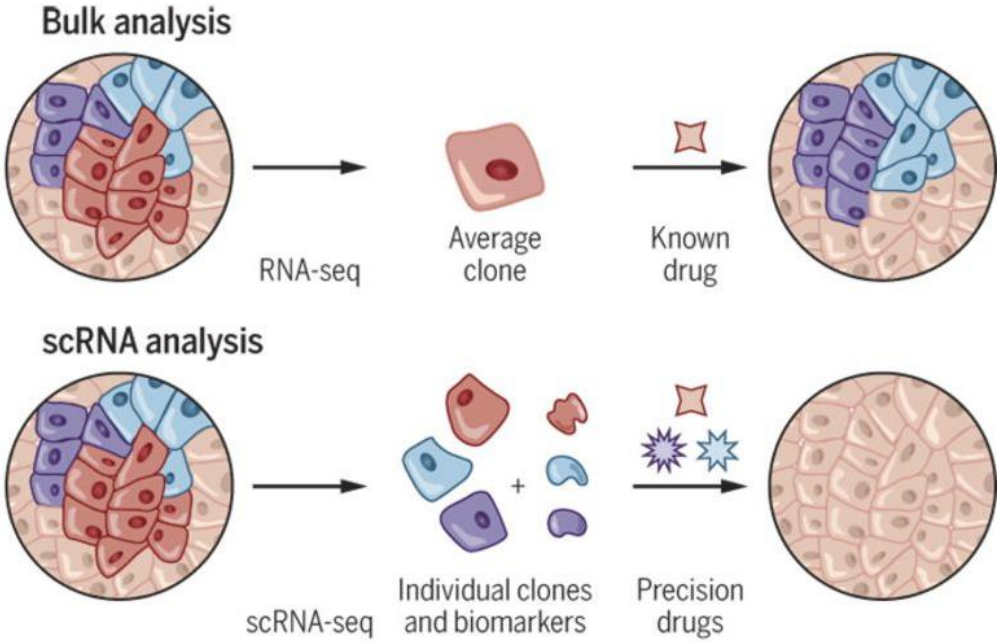
Today's agenda

- Limitation of bulk tissue omics

- Single-cell sequencing
  - Cell barcode and UMI
  - QC and batch correction

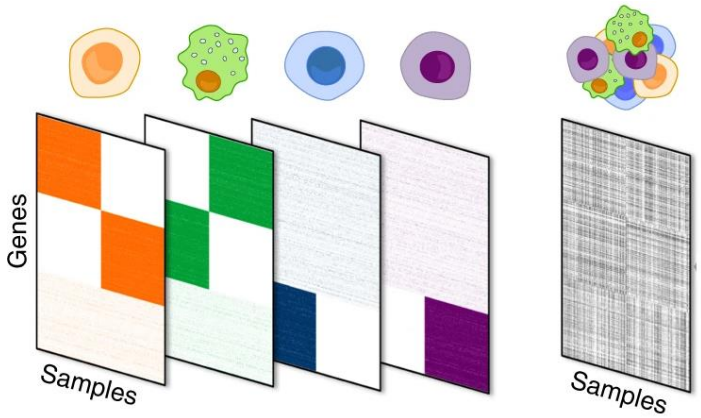
- Spatial omics

Tissue consists of multiple cell types



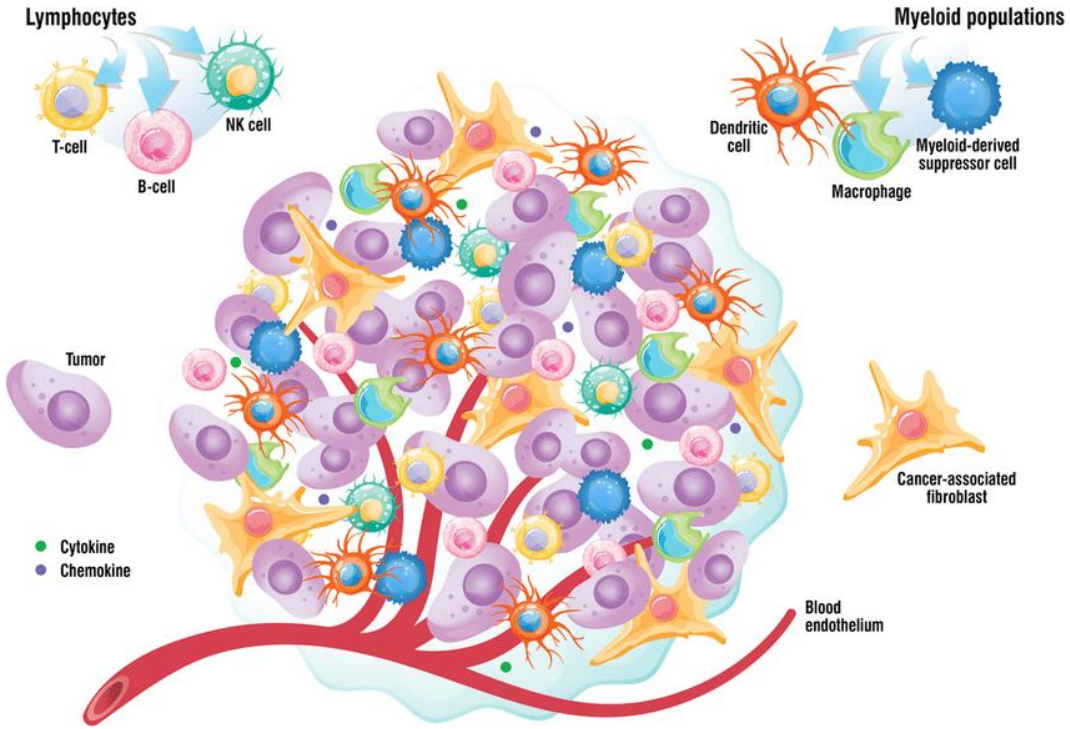
Shalek and Benson. Science Trans Med. 9:eaan4730 (2017)

Each biological sample is a mixture of many cell types



Newman *et al*. Nat Biotech 2019

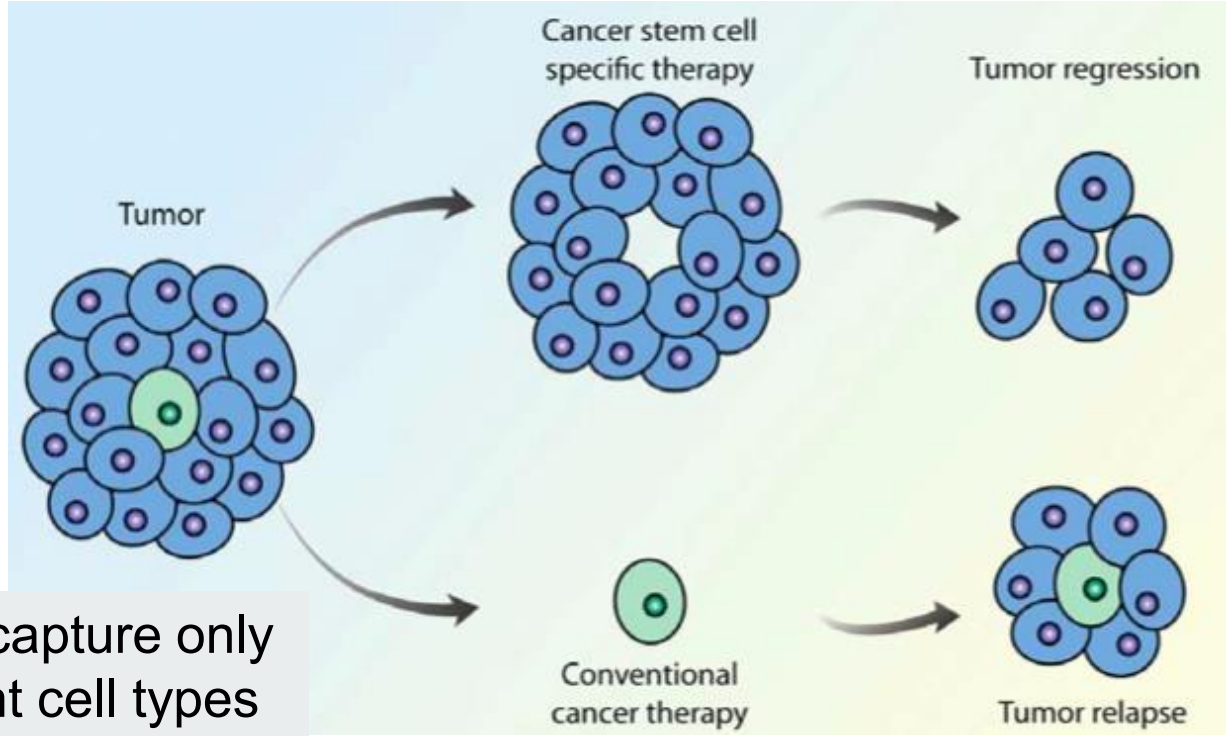
Tumor microenvironment



Zhang, J. and Veeramachaneni, N. Biomarker Research 10:5 (2022)

- **Intrinsic** = changes within the actual cancer cells
- **Tumor microenvironment** = changes among the surrounding cells recruited by cancer cells
- Bulk omics cannot disentangle the two

Cancer stem cell

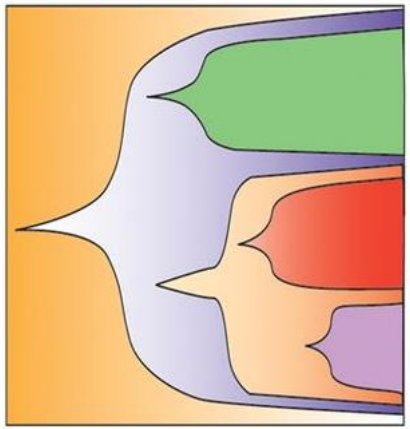


Bulk omics capture only the abundant cell types

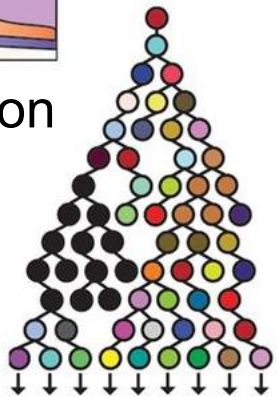
Image from https://hsci.harvard.edu/stem-cells-and-cancer

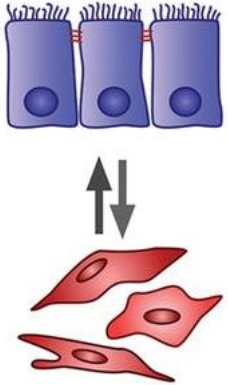
Knowledge at single-cell resolution

Heterogeneity



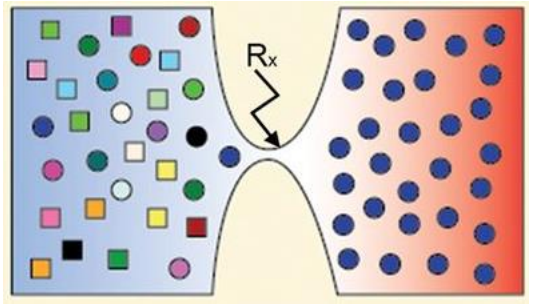
Clonal expansion



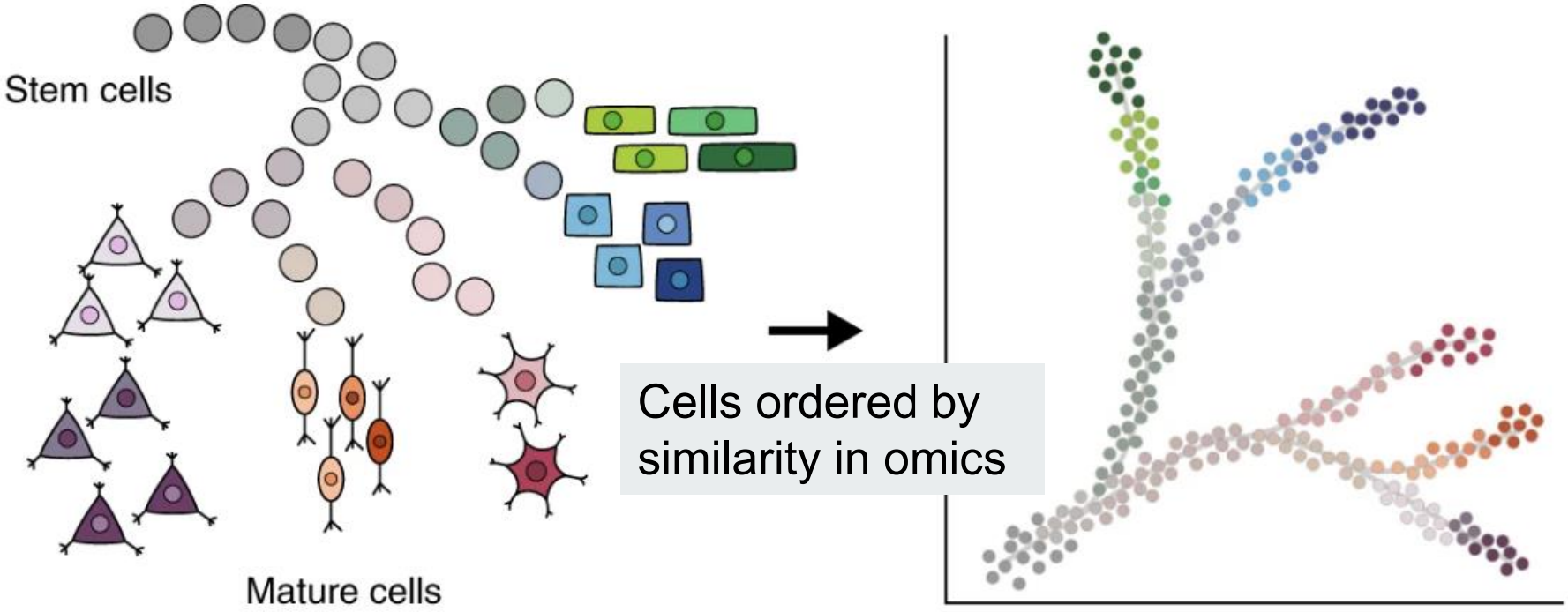


Differentiation

Treatment response



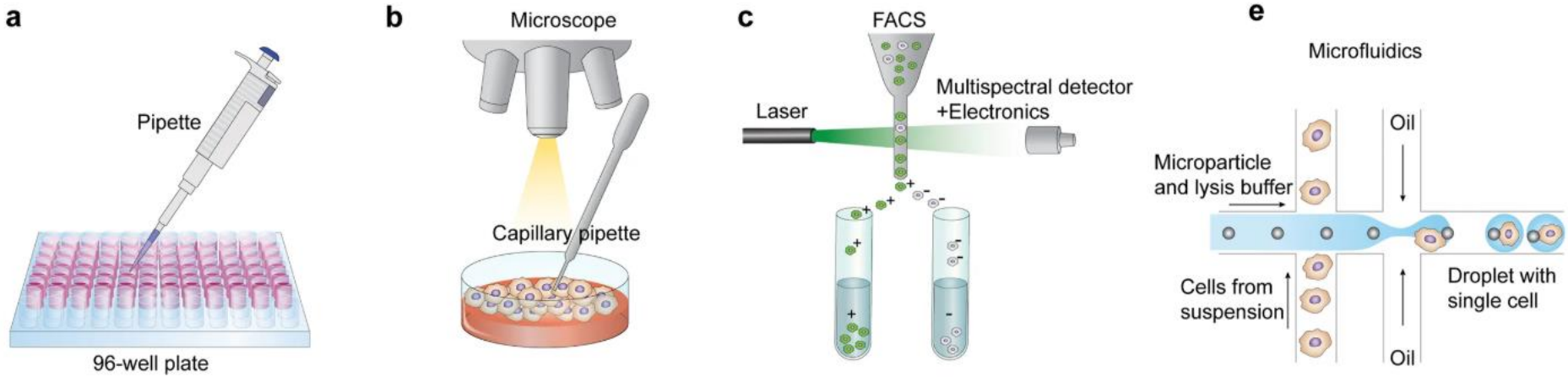
Cell development reconstruction



Source: Moon et al. Nature Biotechnology 37:1482-92 (2019)

Single-cell workflow

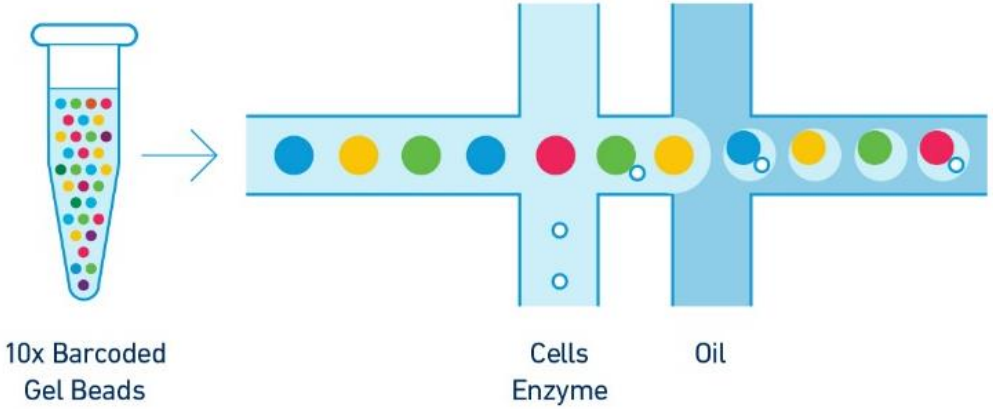
Cell isolation



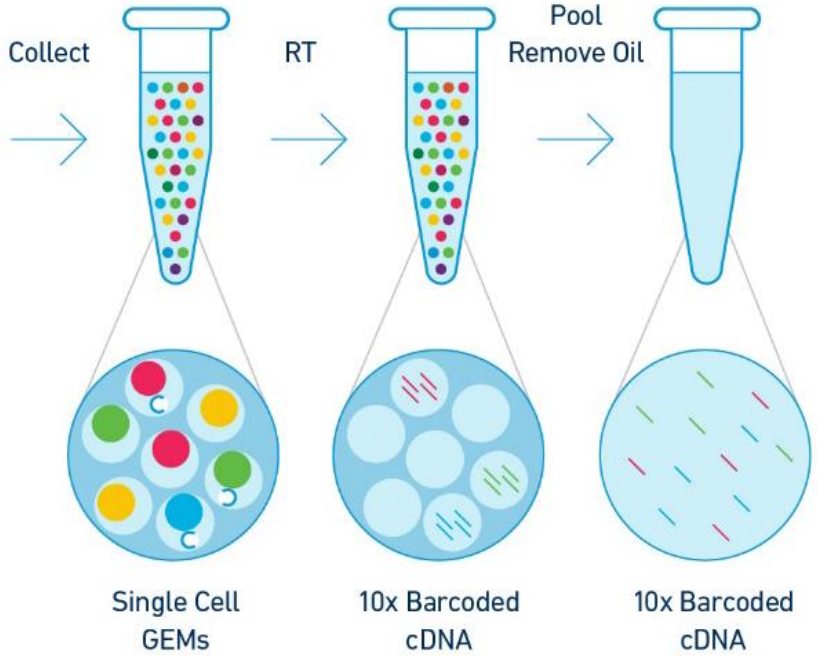
Hwang *et al.* Exp & Mol Med 50:96 (2018)

- The essence of single-cell approach is the ability to isolate individual cells, barcode them, and generate cell-specific omics data

A workflow for preparing single-cell cDNA

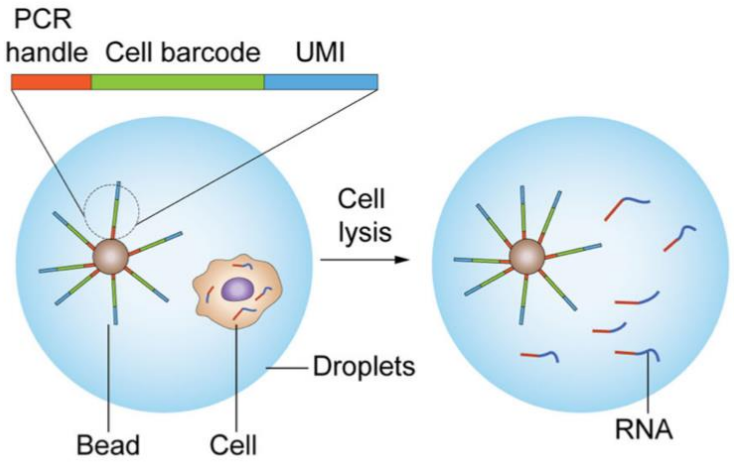


https://bauercore.fas.harvard.edu/10x-chromium-system



- Droplets of sequencing reagents
- Separate RT and PCR
- Pooled DNA sequencing

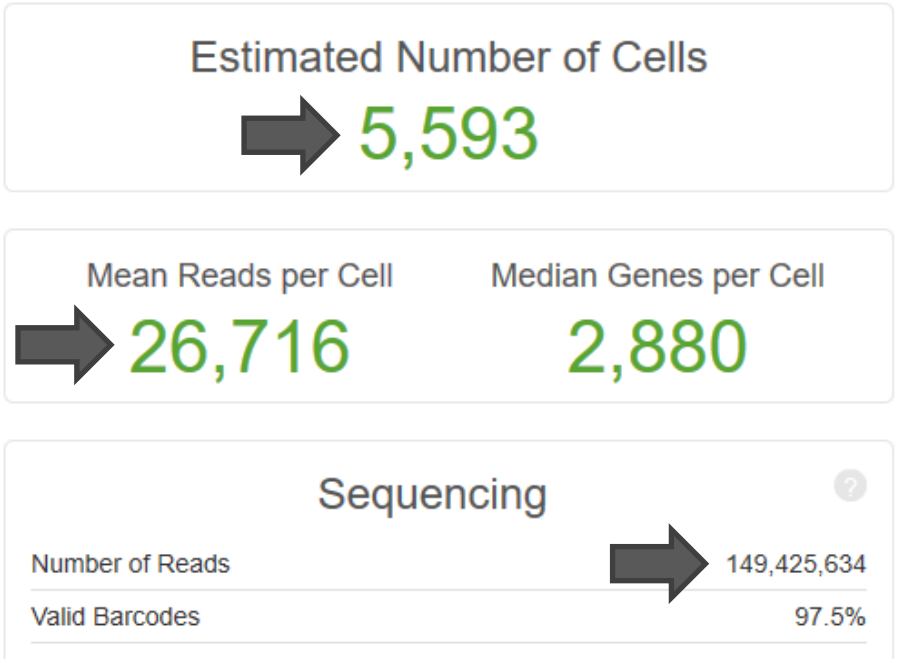
UMI and cell barcode



Hwang *et al.* Exp & Mol Med 50:96 (2018)

- All PCR adapter molecules in each droplet have the same cell barcode
  - All reads with the same barcode originate from the same cell
- Each PCR adapter contains different Unique Molecular Identifiers (UMI)
  - All reads with the same UMI originate from the same RNA molecule
- **Allow pooled sequencing**

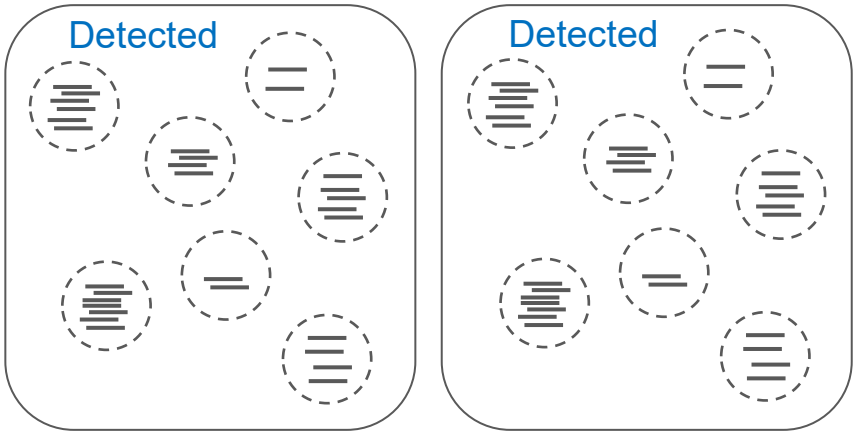
Single-cell results



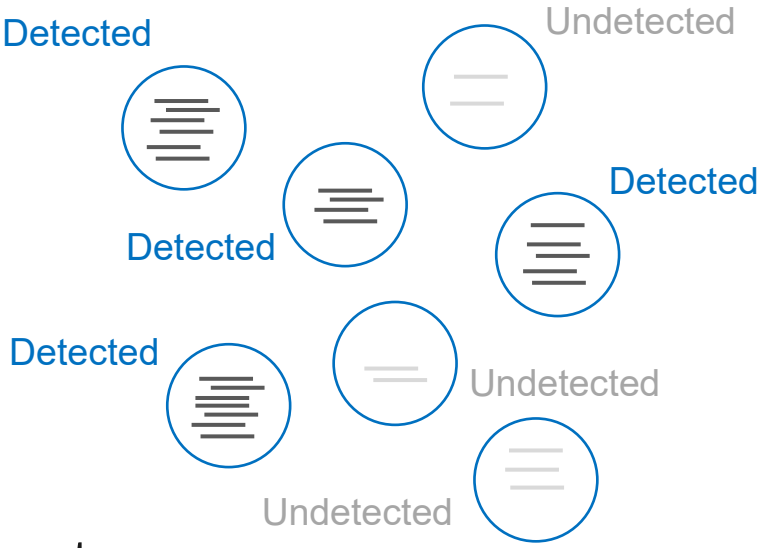
- 3,000-10,000 cells per sample
- Low read count per cell & gene
  - 20,000-50,000 reads
- Good enough for detecting cell types, but limited differential expression analysis

Detection limit issue in single-cell data

Bulk RNA-seq



Single-cell RNA-seq



- Bulk differential expression models do not fit single-cell UMI count data

Many missing values

Challenges in single-cell data analysis

- A lot of zeros in expression data
- Cells are biologically different (higher variance)
- Strong batch effect (low read count → noisy data)
- **Cells are in continuous states of development**
- Data is very large
  - Bulk: 10 samples x 10,000 genes = 0.1 million values
  - Single-cell: 5,000 cells x 3,000 gene = 15 million values

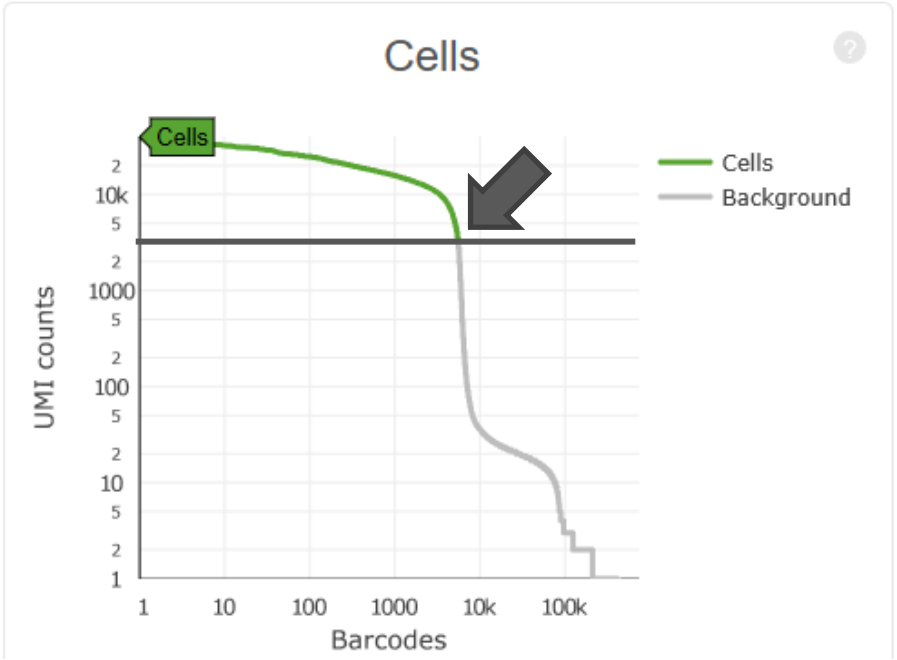
Single-cell data QC and processing

Key steps in single-cell data processing

- Quality filter
  - Low read count & gene count = non-cells
  - Very high read count & gene count = multi-cells (doublets)
  - High mitochondrial expression ~ dead cells (or special cell types)
- Normalization and impute missing values
- Multi-sample integration
  - Some cell types/genes are detected in some batches
  - Some genes are affected by conditions → affect visualization and clustering

Distinguishing cells from empty droplets

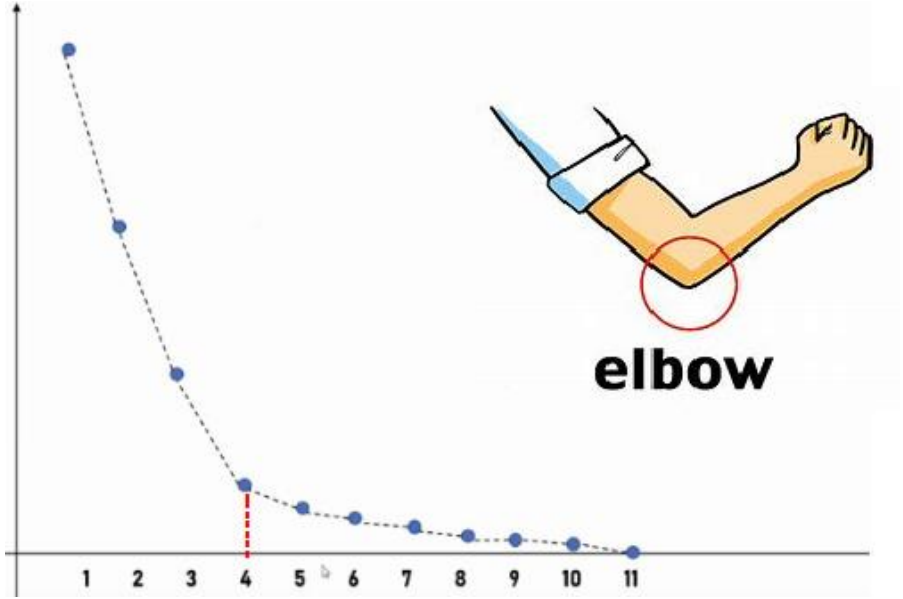
- **Assumption**: Empty droplets produce very low UMI counts
- Sort data by number of UMI per cell barcode
- Identify the “shoulder” where UMI count drops significantly



The elbow method

- Inflection point is where the graph switches from changing sharply to being flat

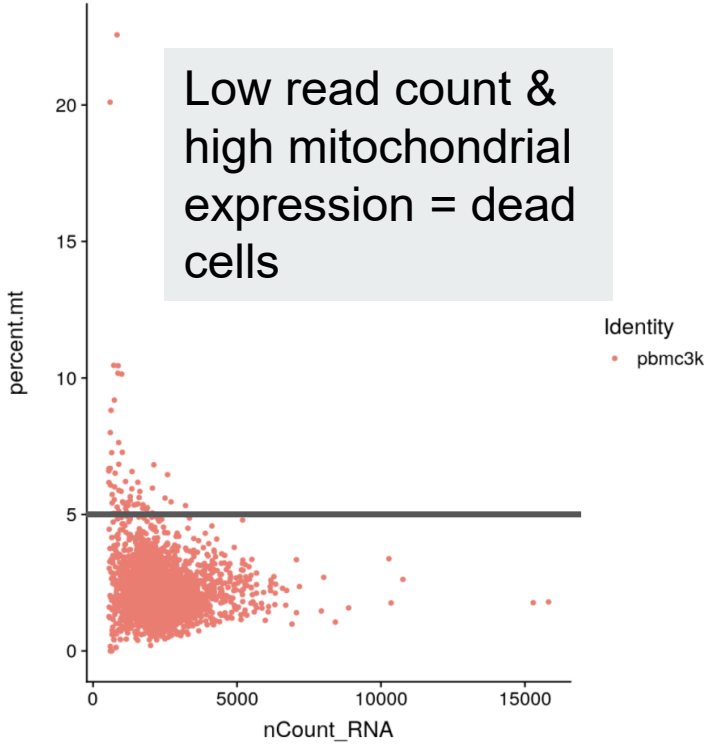
- Signify a change in certain innate characteristics
  - Population from signal to noise
  - Select model parameter



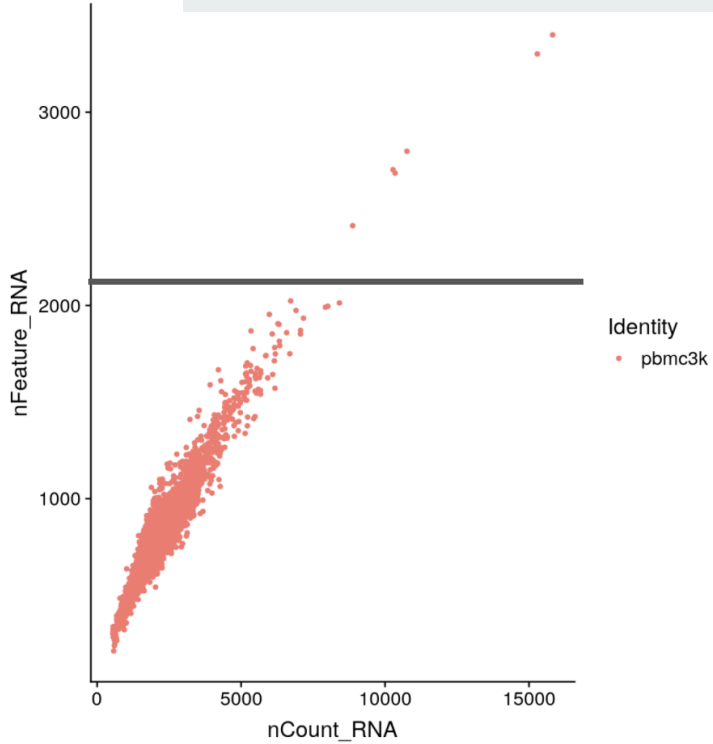
https://medium.com/@zalarushirajsinh07/the-elbow-method-finding-the-optimal-number-of-clusters-d297f5aeb189

Basic quality filters

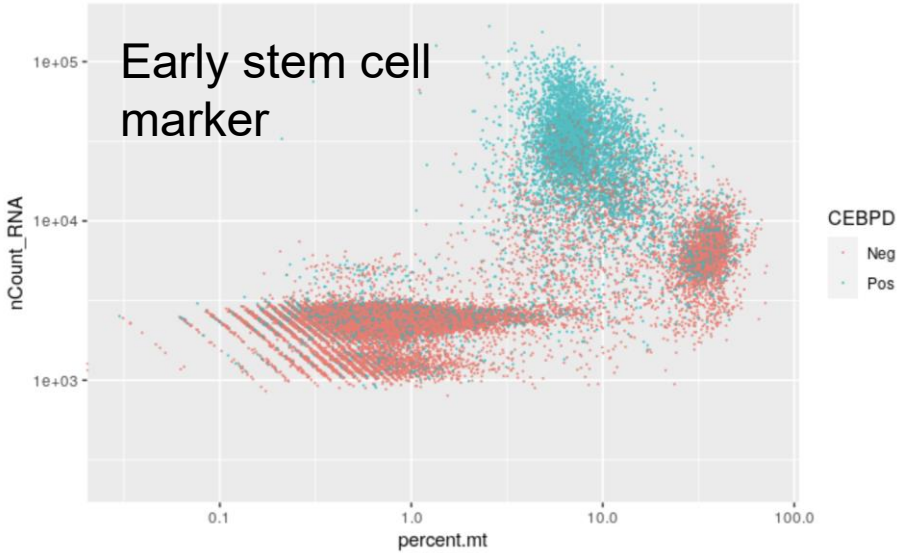
Low read count & high mitochondrial expression = dead cells

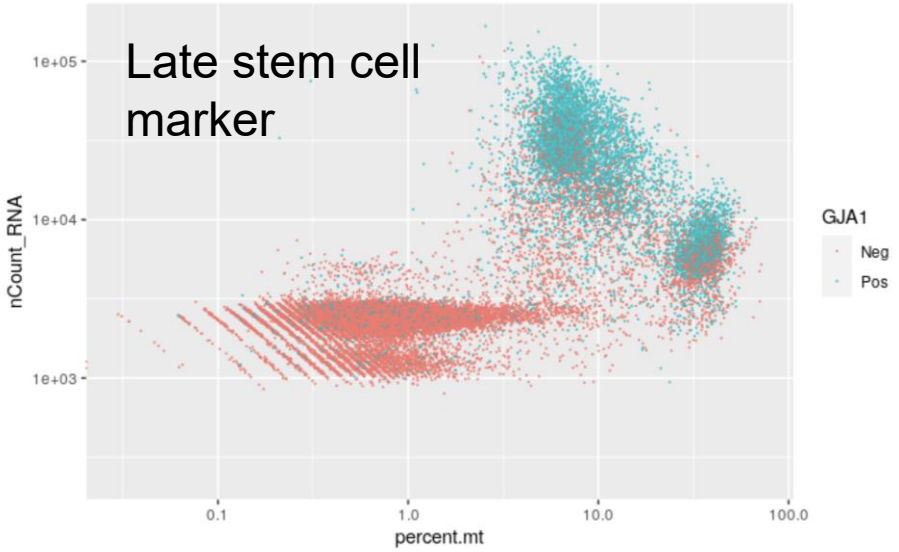


High read count & gene count = doublets



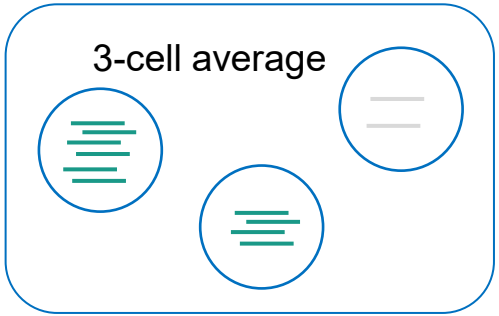
High mitochondrial expression in some cell types

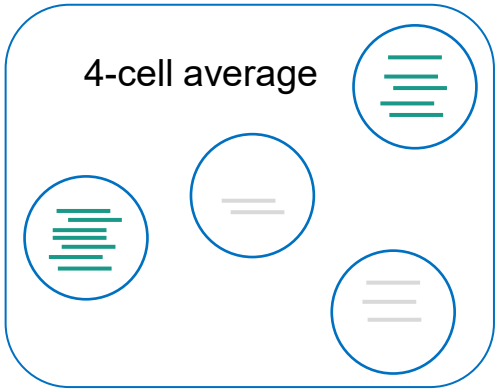




- **Example**: stem cells, muscle cells, cardiac cells
- Combine MT signature with other evidence

Normalization with pooling





Low expression

Gene A

Gene B

Gene C

Medium expression

Gene D

Gene E

Gene F

Gene G

High expression

Gene H

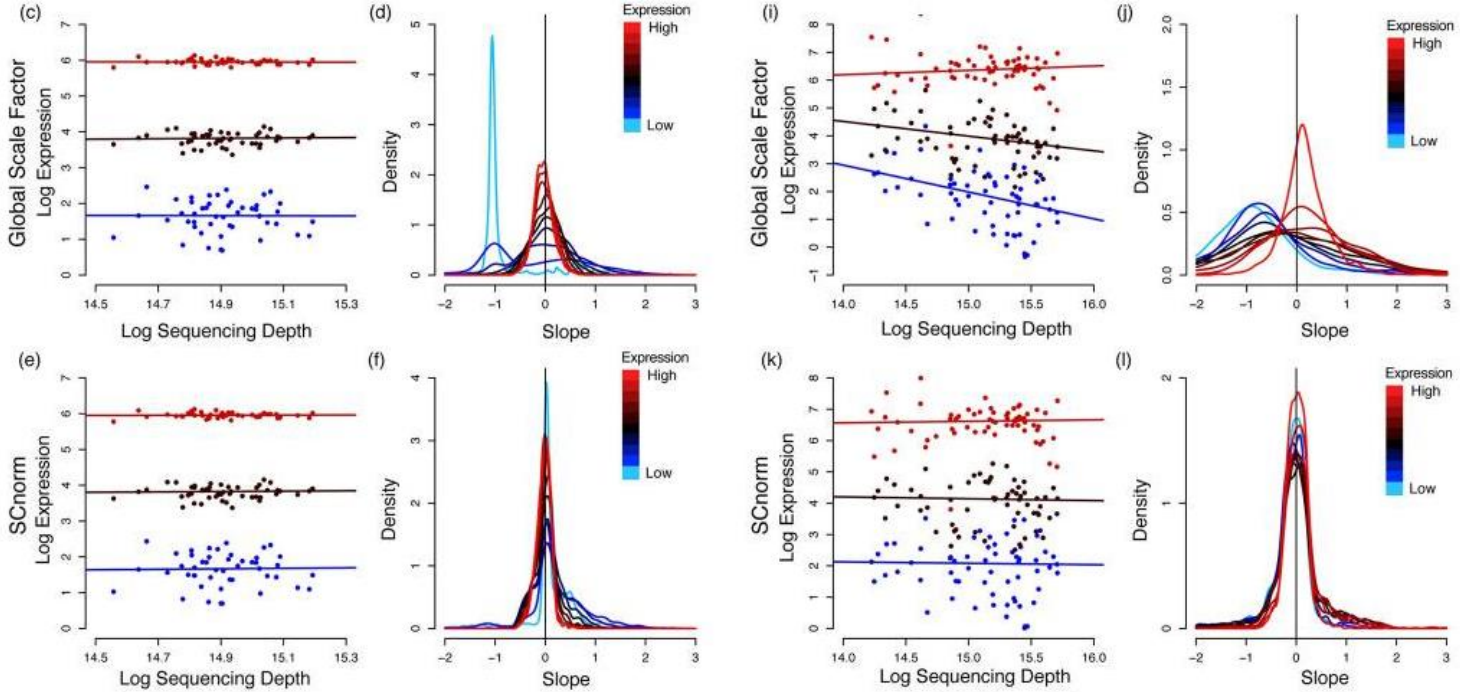
Gene I

Expression level-based correction

Expression-dependent scaling factor

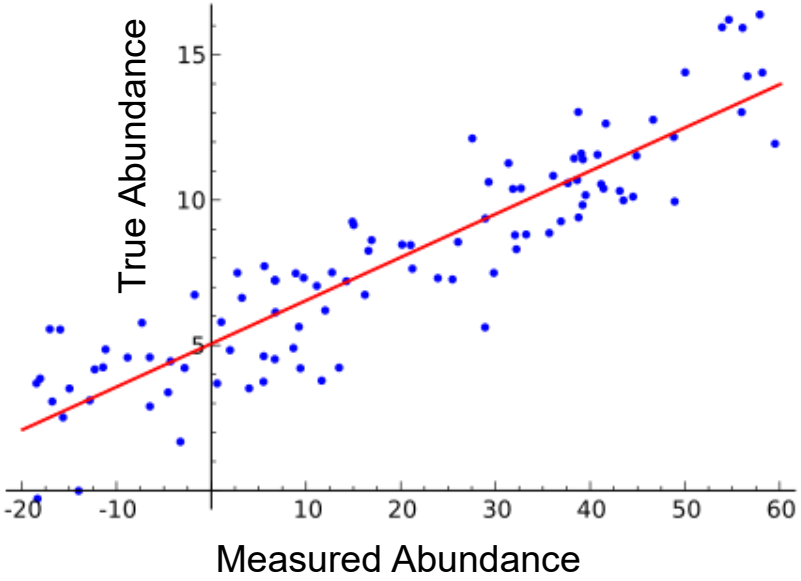
Bulk RNA-seq

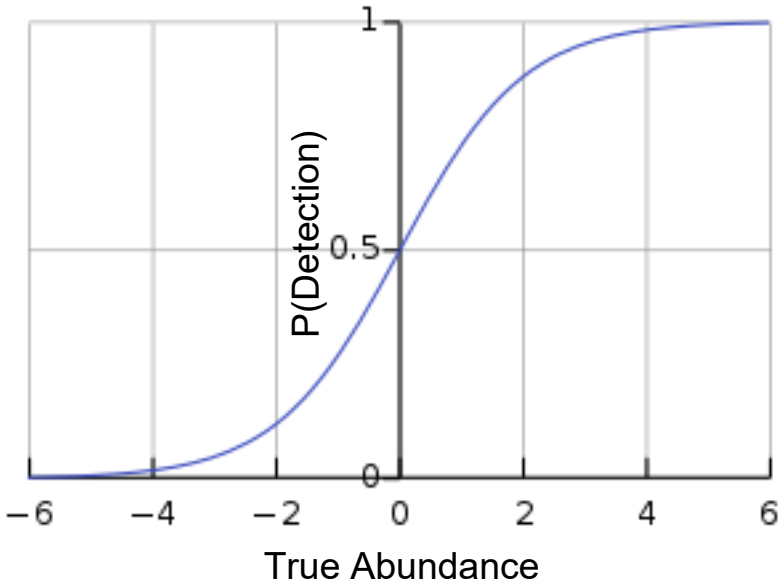
Single-cell RNA-seq



Bacher *et al.* Nat Methods 14:584-586 (2017)

A model for missing values

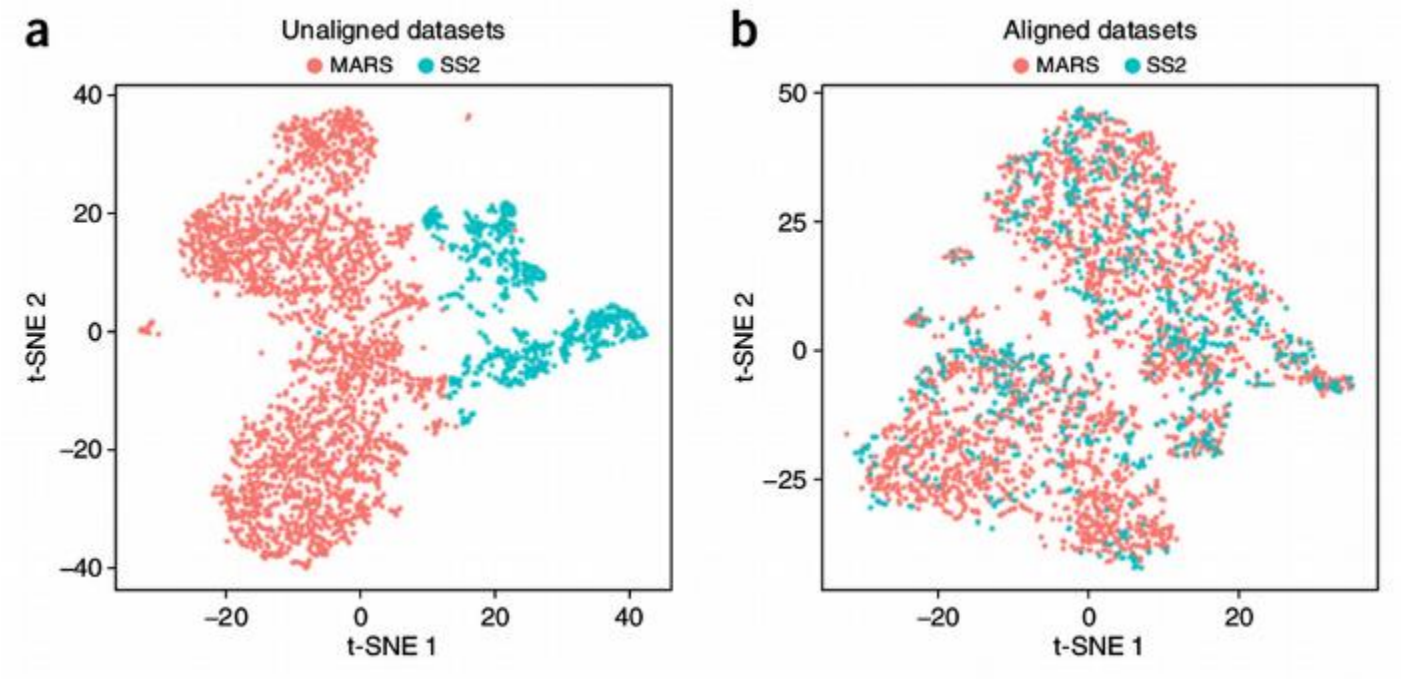




- Measured abundance = f(true abundance) x P(detection | true abundance)

Single-cell batch effect

High bias across single-cell datasets

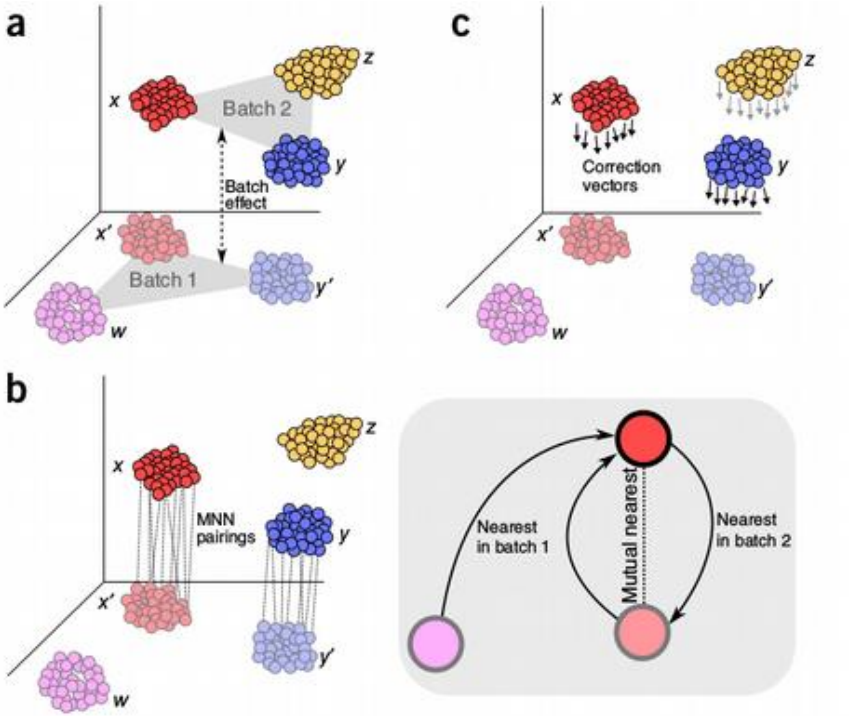


https://towardsdatascience.com/how-to-batch-correct-single-cell-7bad210c7ae1

Idea for removing batch effects

- **Assumption**: Batch effect was introduced by the low read counts (random sampling of which RNA molecules to sequence)
- Cell type A in sample 1 = Cell type A in sample 2 + noise
- Cell type A in sample 1 should be most similar to cell type A in sample 2, compared to other cell types in sample 2

Mutual nearest neighbor (MNN)

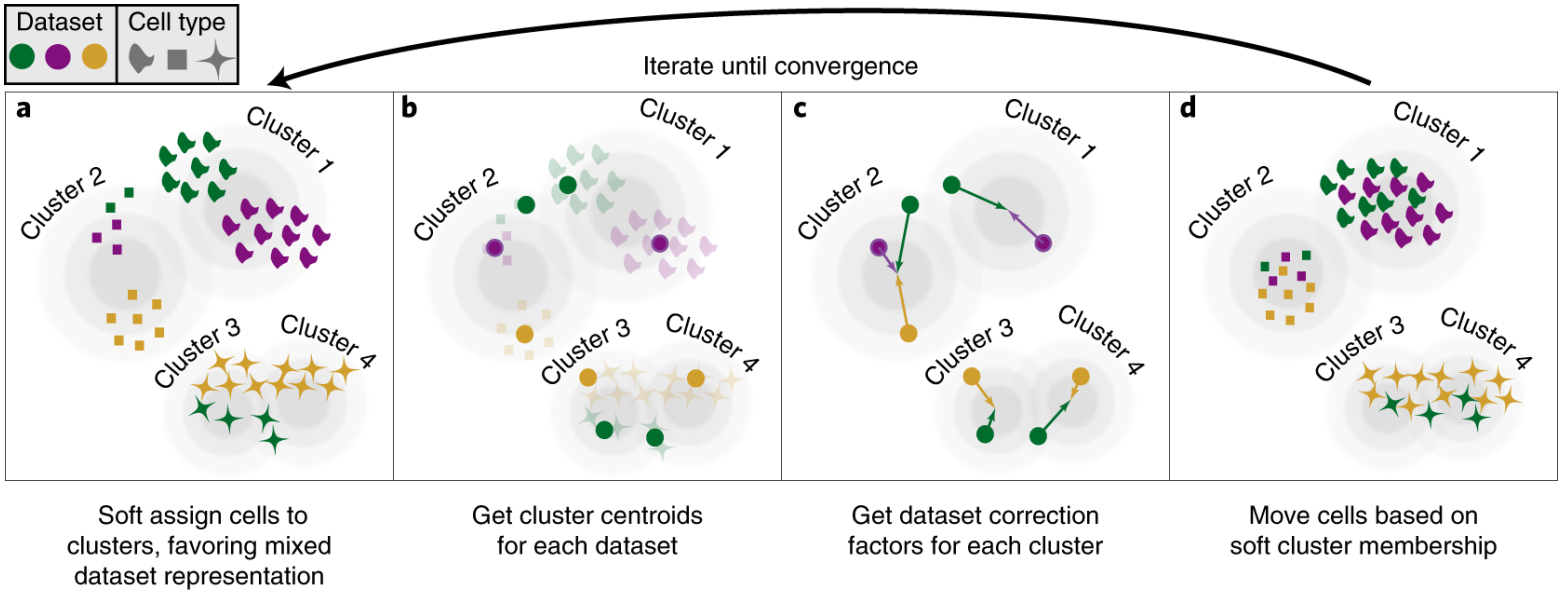


- Identify similar cells across samples
- Cell of the same type should be mapped together
- Unique cell types will show inconsistent mapping
- **Reciprocal best hit**

Haghverdi et al., Nat. Biot. 36, 2018

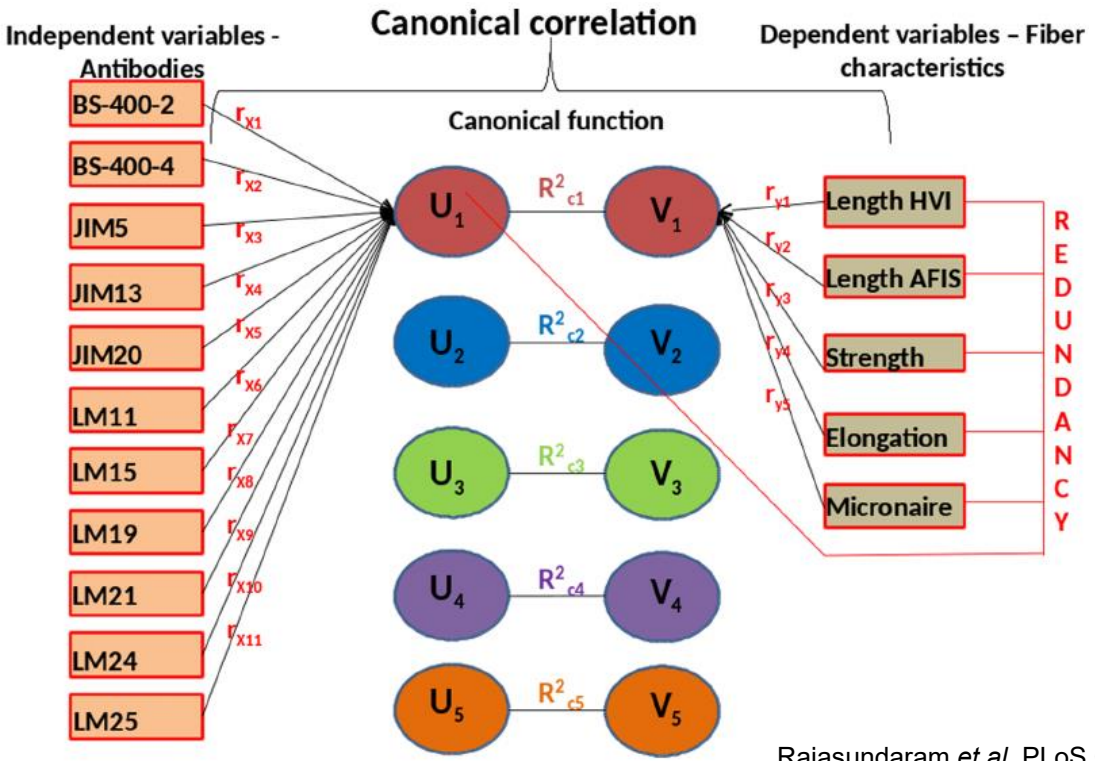
Harmony: clustering-based correction

Korsunsky, I. et al. Nature Methods 16:1289-1296 (2019)



- Iteratively cluster cells from all batches together
- Use batch-specific centroids to correct original data

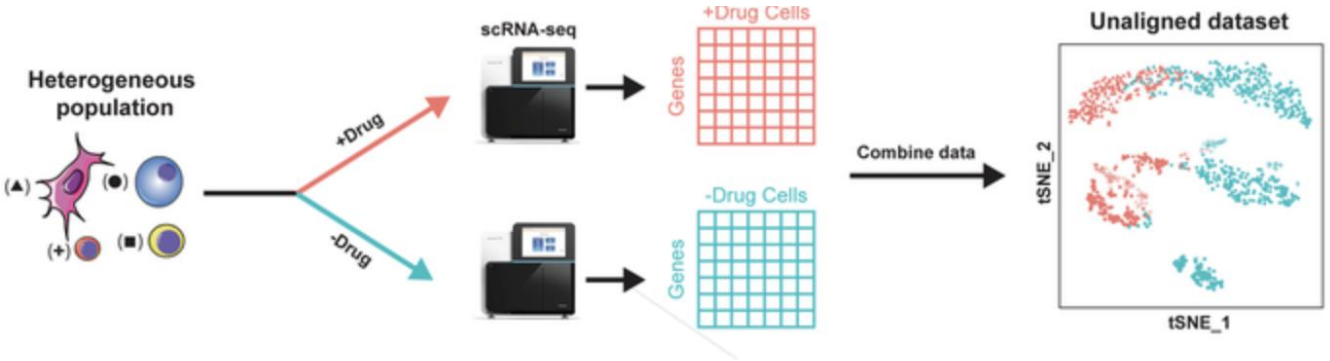
Canonical correlation analysis (CCA)



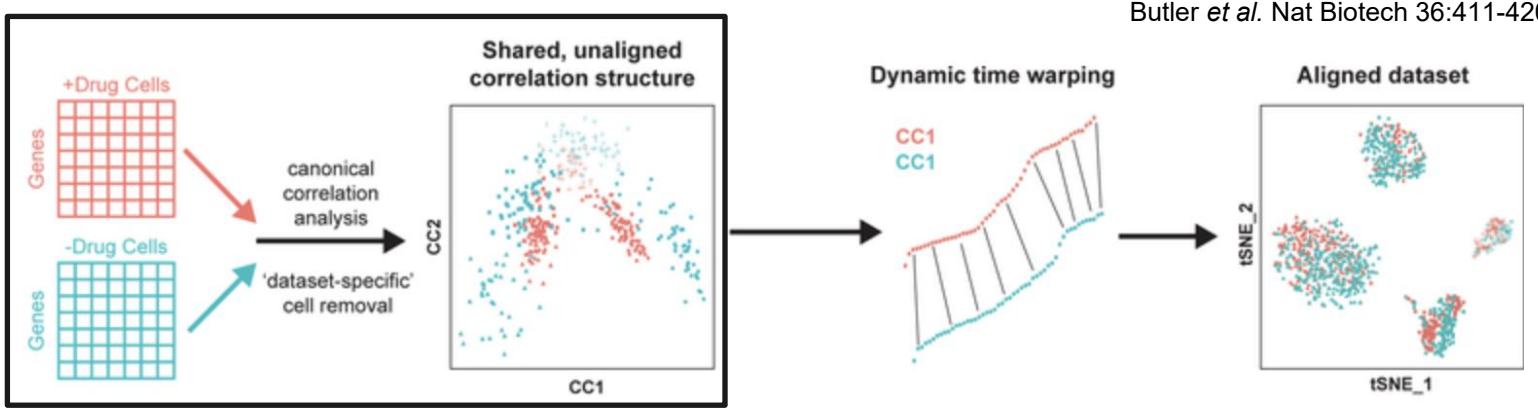
- Same samples (cells from the same tissues) with two different systems of observations (single-cell batches)
- Identify correlated structures across observations (capture tissue cell types)

Rajasundaram *et al.* PLoS ONE (2014)

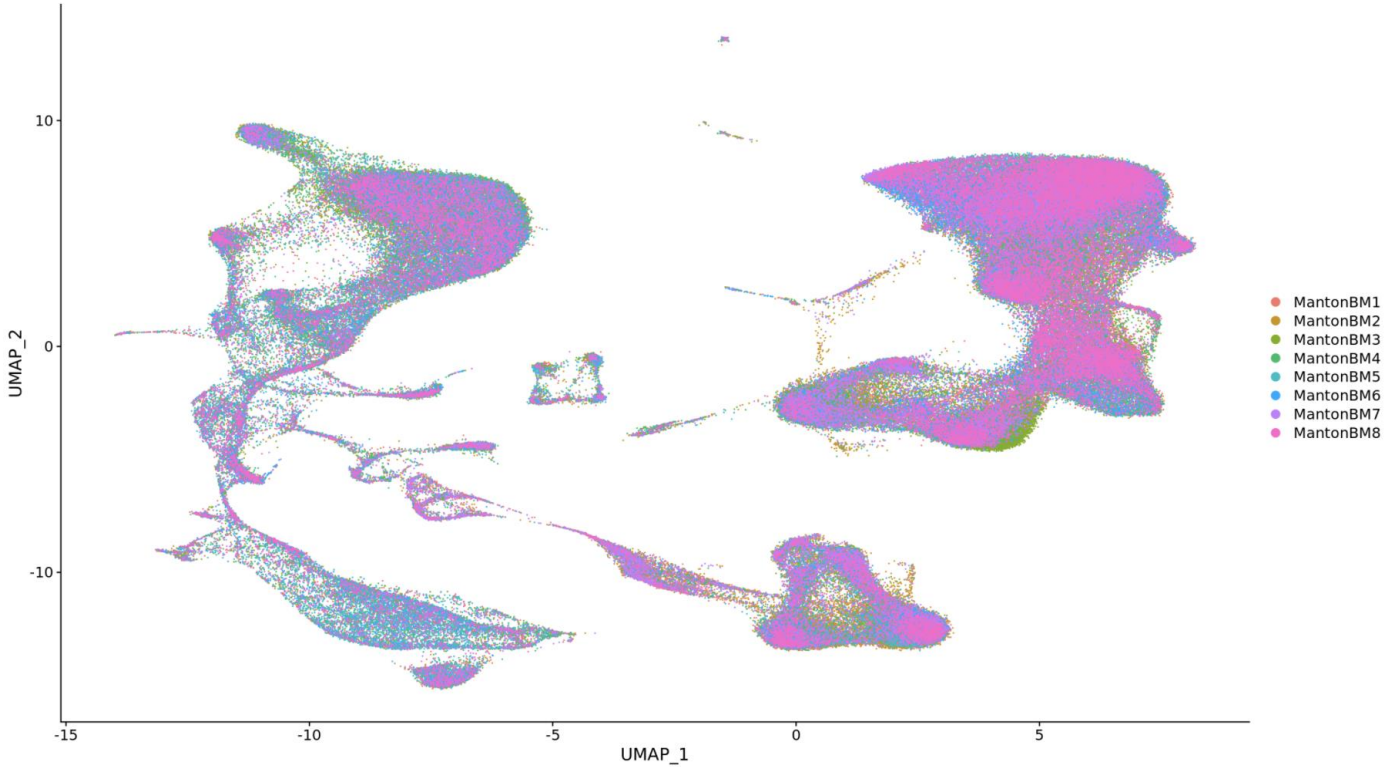
CCA applied to single-cell data



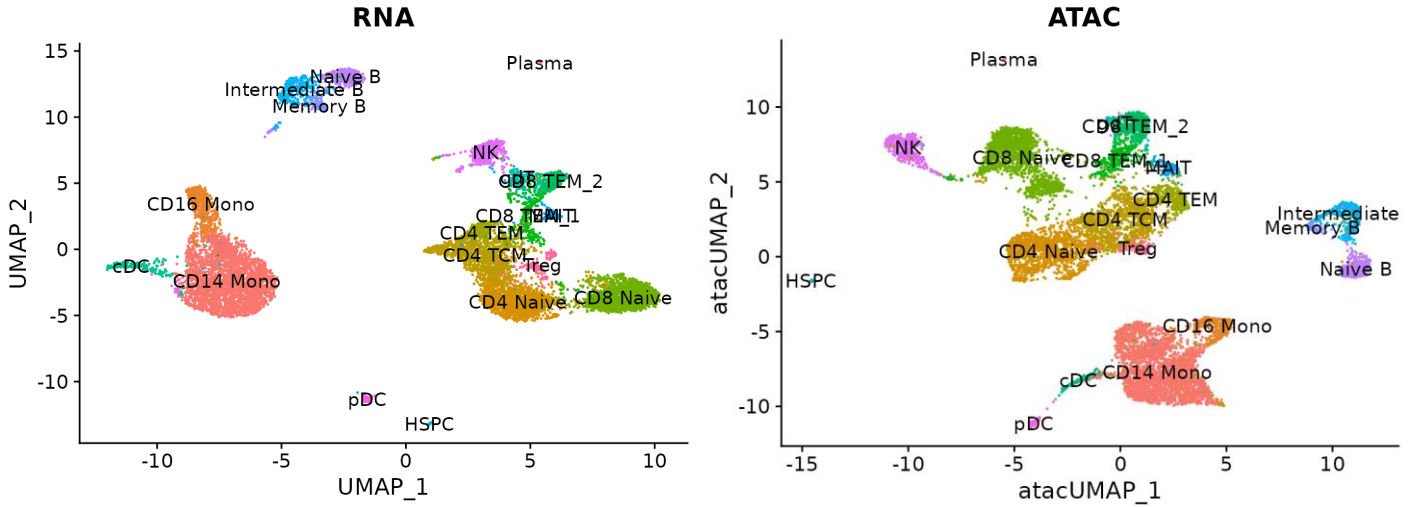
Butler *et al.* Nat Biotech 36:411-420 (2018)



Result of batch correction (sample integration)



Extension of batch effect idea to multi-omics



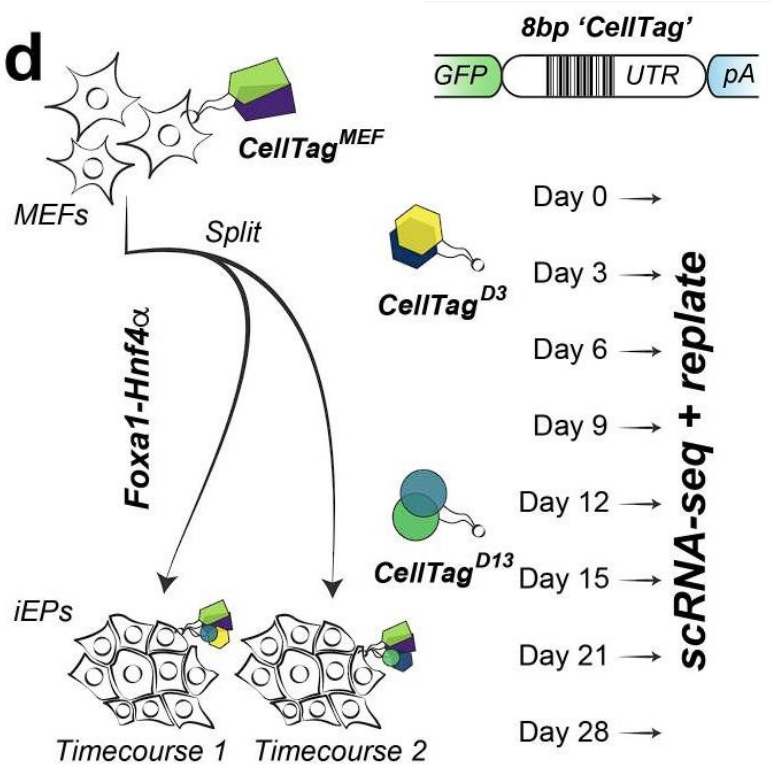
- ATAC-seq measures chromatin accessibility ~ gene expression level
- ATAC-seq and RNA-seq as two systems of observations
  - Apply CCA approach

Other single-cell related techniques

Single-nucleus sequencing

- Isolate nuclei instead of whole-cells
  - Only capture RNA expression in nucleus
  - Good enough for distinguishing cell types
- Needed for tissues that are difficult to isolate
  - Adipocyte, neuron, etc.
  - Also works well with preserved tissues

Lineage tracing

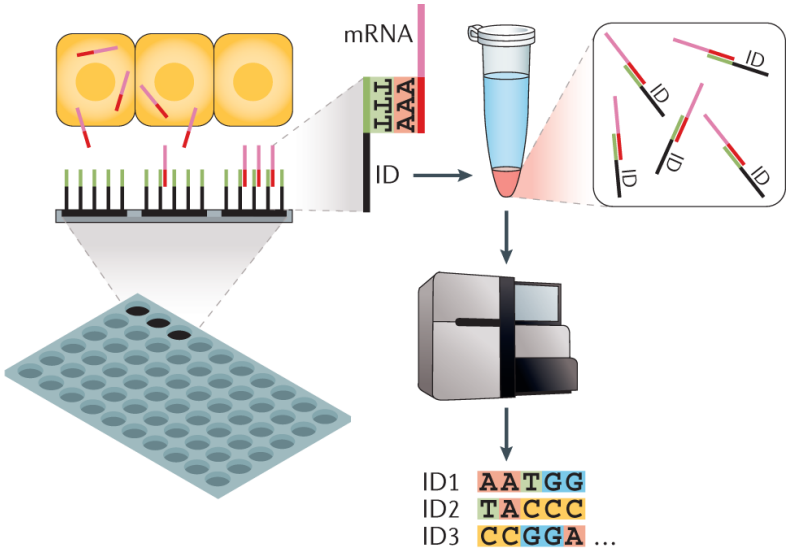


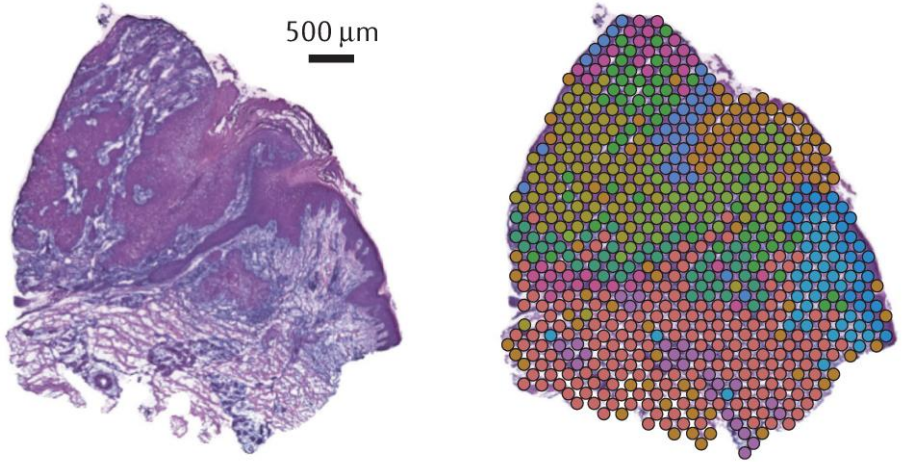
- Introduce an expressible cassette of random RNA sequences into each cell
- As cells divide, the cassettes are inherited
- Enable precise tracing of cell lineages from single-cell transcriptomics data

Biddy, B.A. et al. Nature 564:219-224 (2018)

Spatial omics

Spatial barcoding

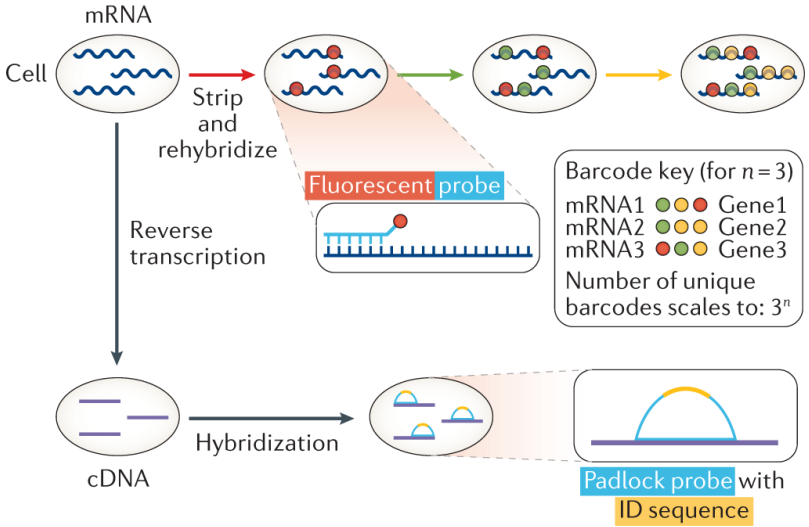


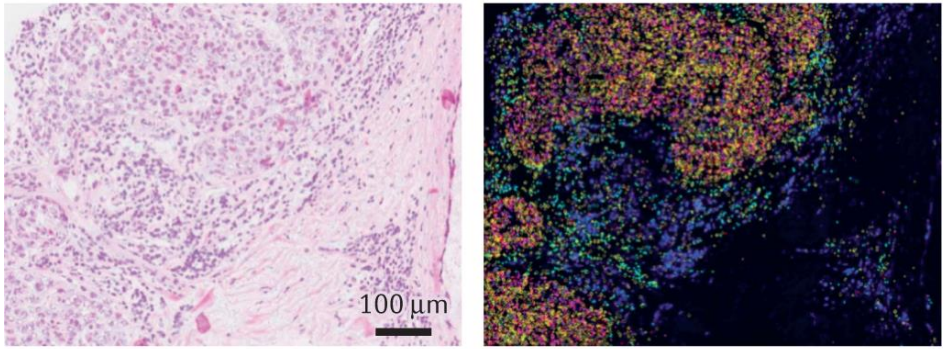


Longo et al. Nature Reviews Genetics 22:627-644 (2021)

- Single-cell data do not provide spatial information (tissue structure)
- Naïve approach: Extract location-specific samples and add barcode

Spatial imaging of markers in tissue structure



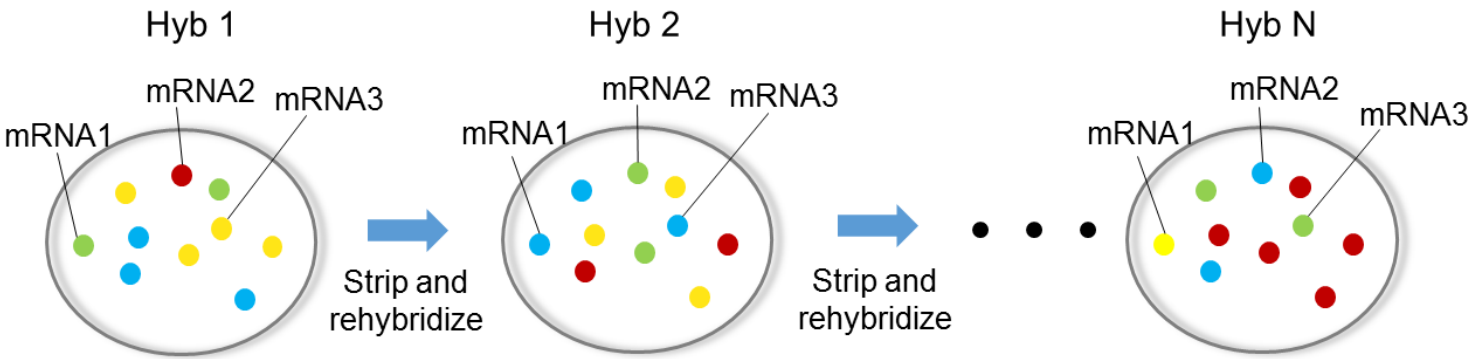


Longo et al. Nature Reviews Genetics 22:627-644 (2021)

- In-situ fluorescence labeling of selected RNA transcripts and proteins
- Latest technique can target >10,000 RNA

seqFISH: sequential hybridization readout

https://spatial.caltech.edu/seqfish/



Barcode # scales as F^N

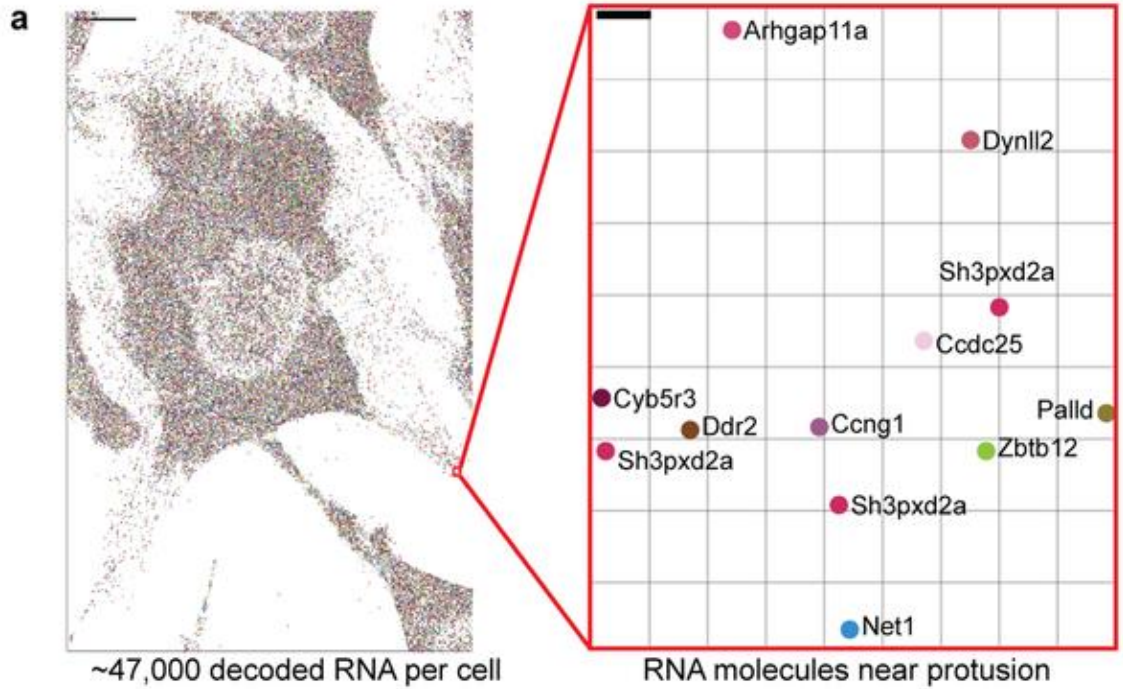
$5^3=125$

$5^6=15,625$

Barcode Key: (for N=6)

- Fluorescence probes bind to each mRNA with **distinct color sequence**

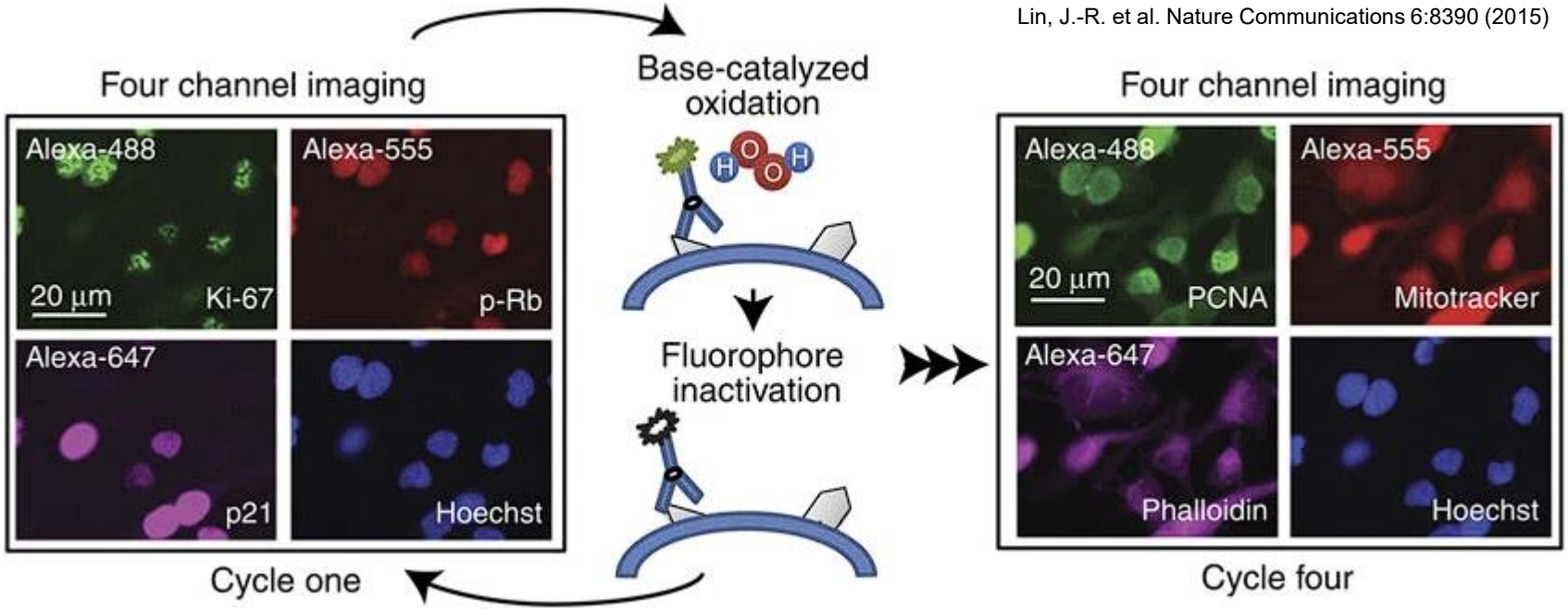
seqFISH: sequential hybridization readout



- Similar to Nanostring
- Many more possible fluorescence patterns
- Higher-resolution imaging
- Applied to fixed tissue

Eng, C.-H. L. et al. Nature 568:235-239 (2019)

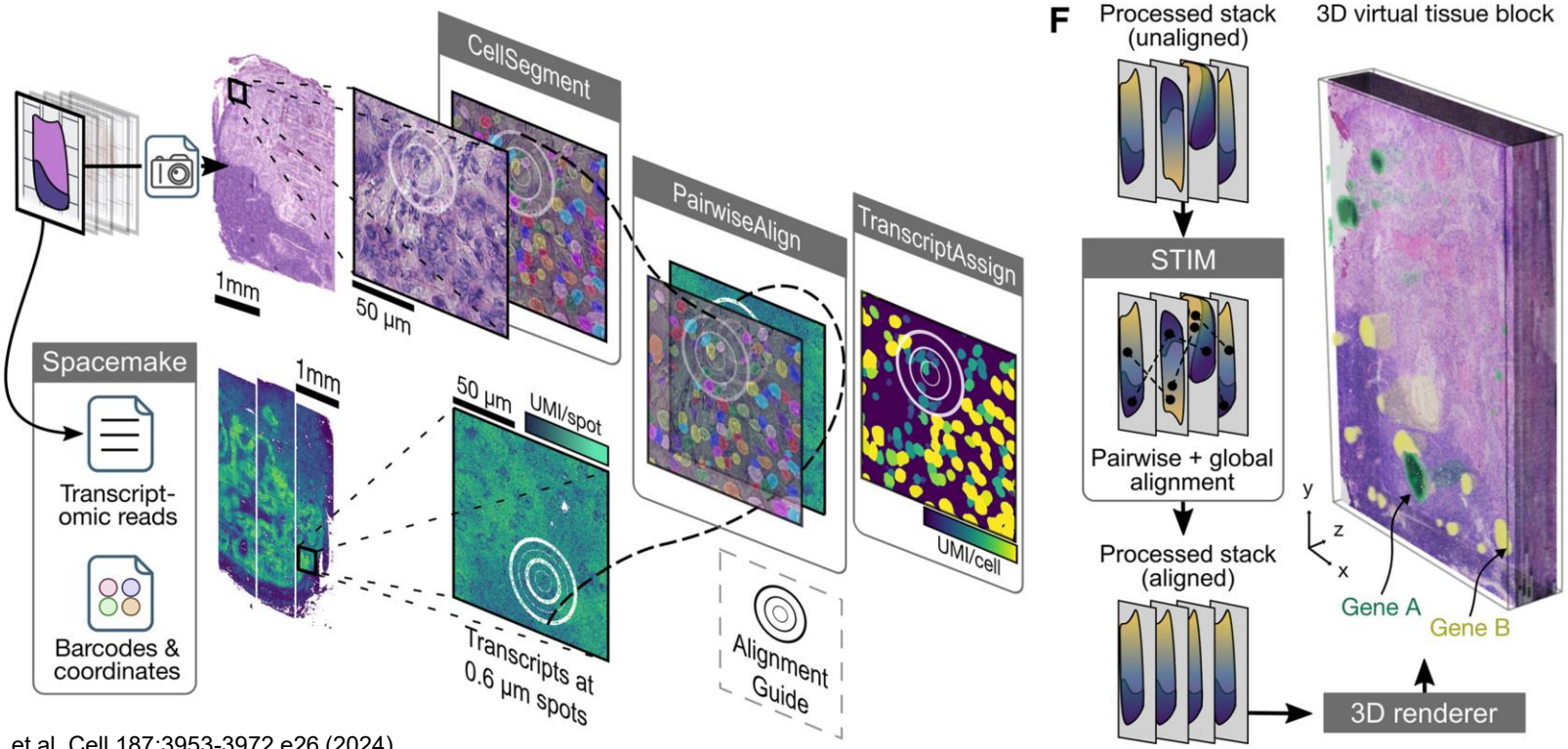
CycIF: cyclic immunofluoresence readout



Lin, J.-R. et al. Nature Communications 6:8390 (2015)

- Repeated binding and cleavage of labeled antibodies for diverse proteins

Open-ST: 3D spatial transcriptomics



Scott, M. et al. Cell 187:3953-3972.e26 (2024)

Summary

- Single-cell and spatial techniques probe heterogeneity of tissues
- Single-cell provides compositional data but no spatial information
- Spatial techniques still lack the sensitivity and resolution of single-cell
- Integration of both methods will be the future of all omics

Any question?

- See you next time