3000788 Intro to Comp Molec Biol

Lecture 14: Single-cell transcriptomics

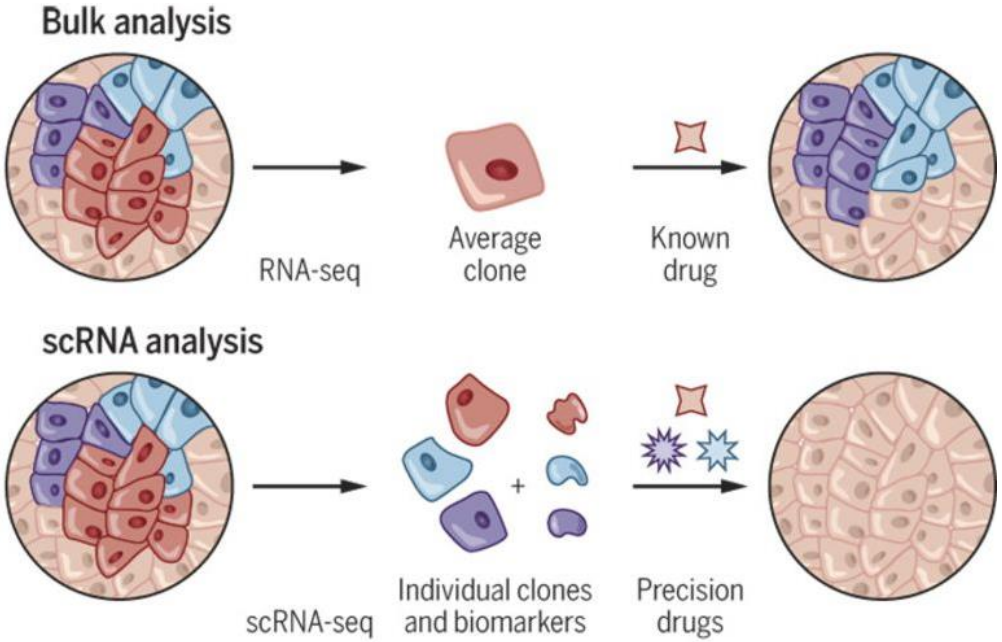
October 2, 2023

Sira Sriswasdi, PhD

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

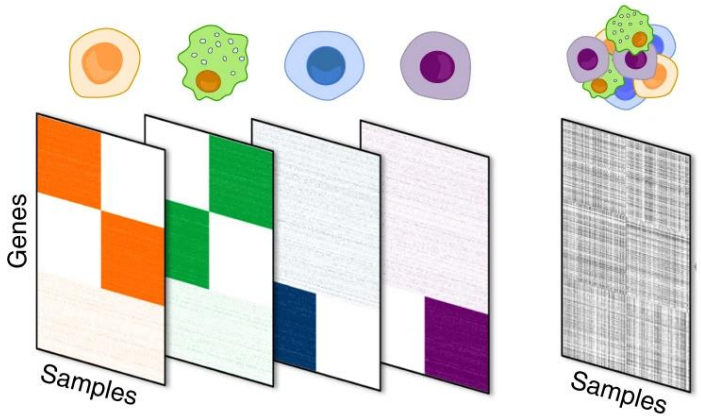
Why single-cell?

Tissue consists of multiple cell types



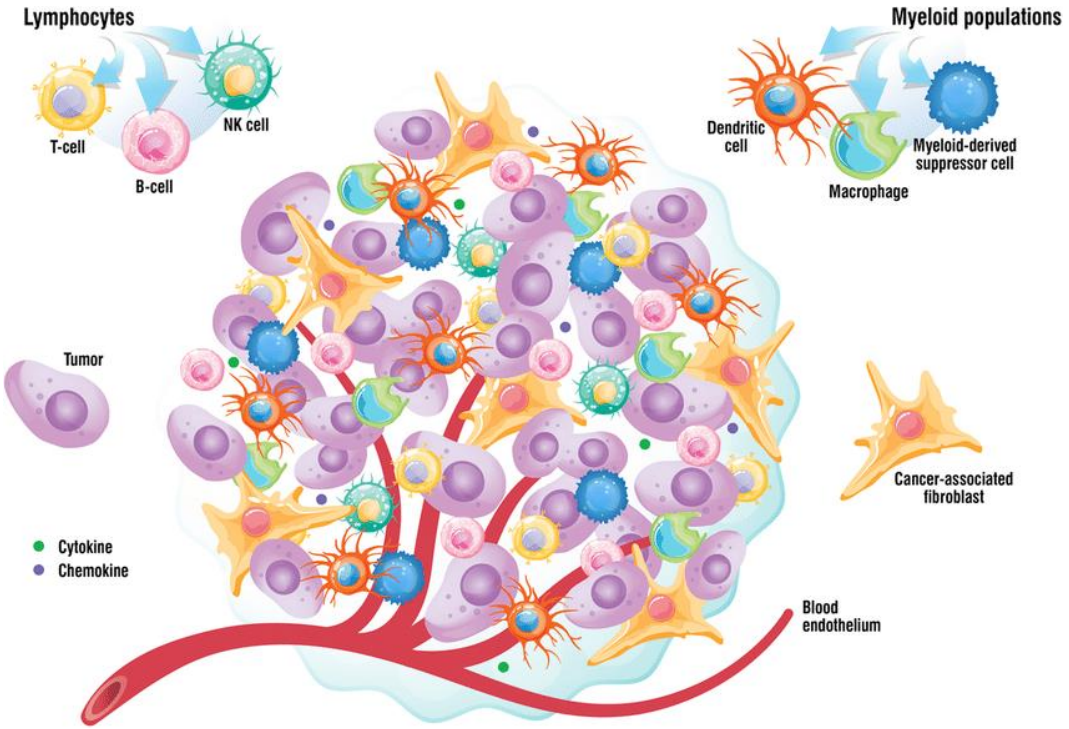
Shalek and Benson. Science Trans Med. 9:eaan4730 (2017)

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)

Cancer stem cell

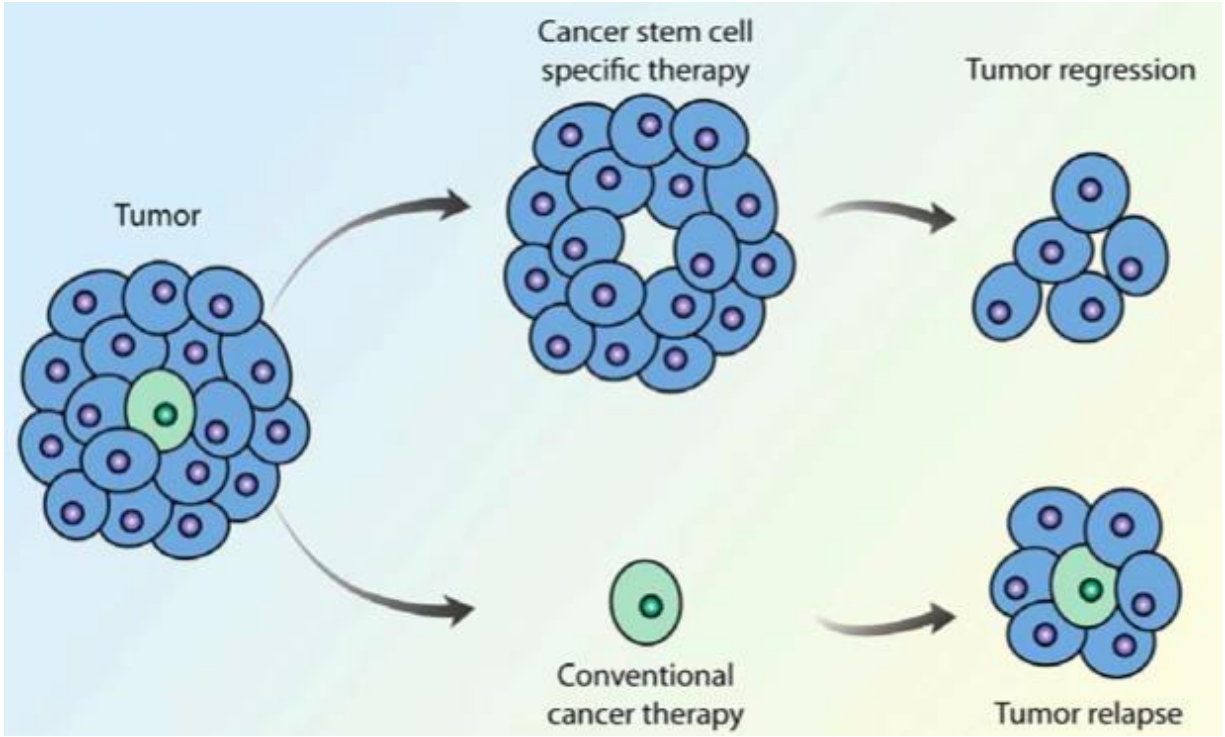
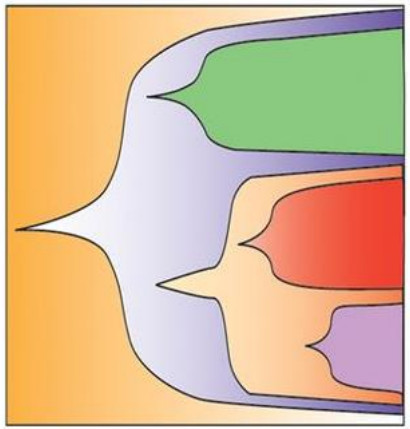


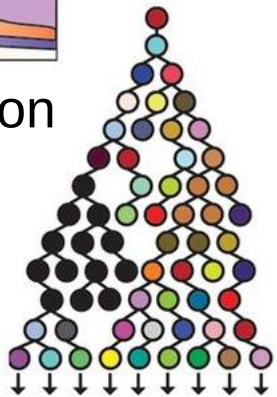
Image from https://hsci.harvard.edu/stem-cells-and-cancer

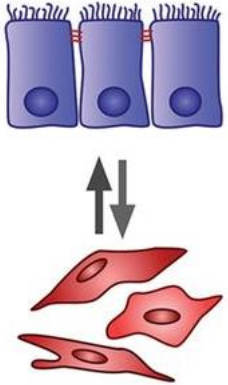
Knowledge at single-cell resolution

Heterogeneity



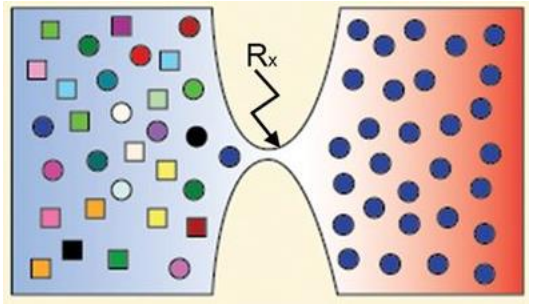
Clonal expansion



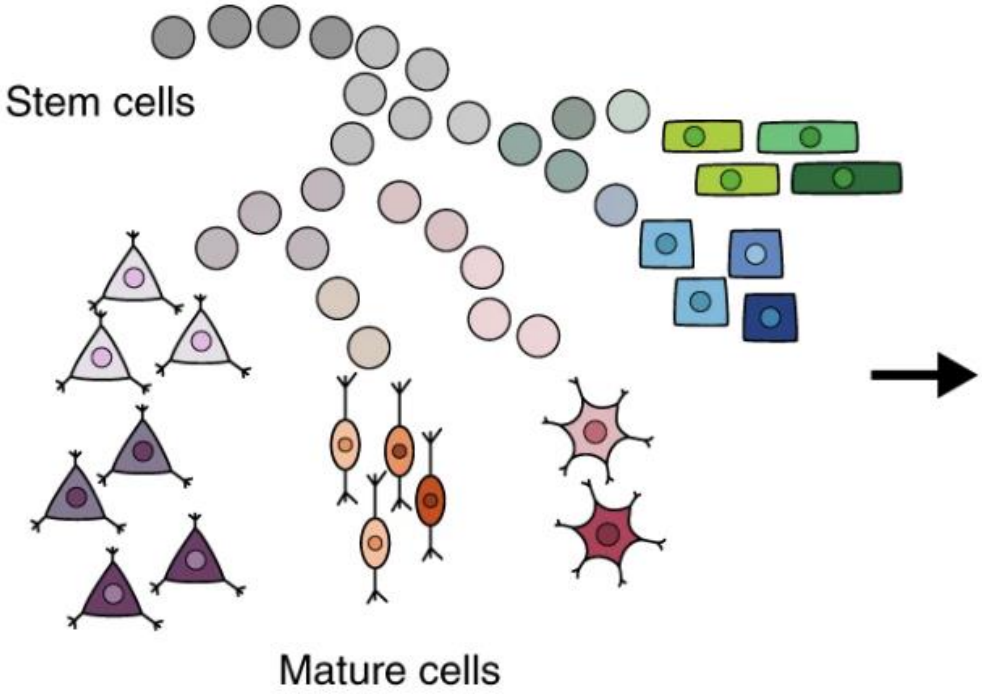


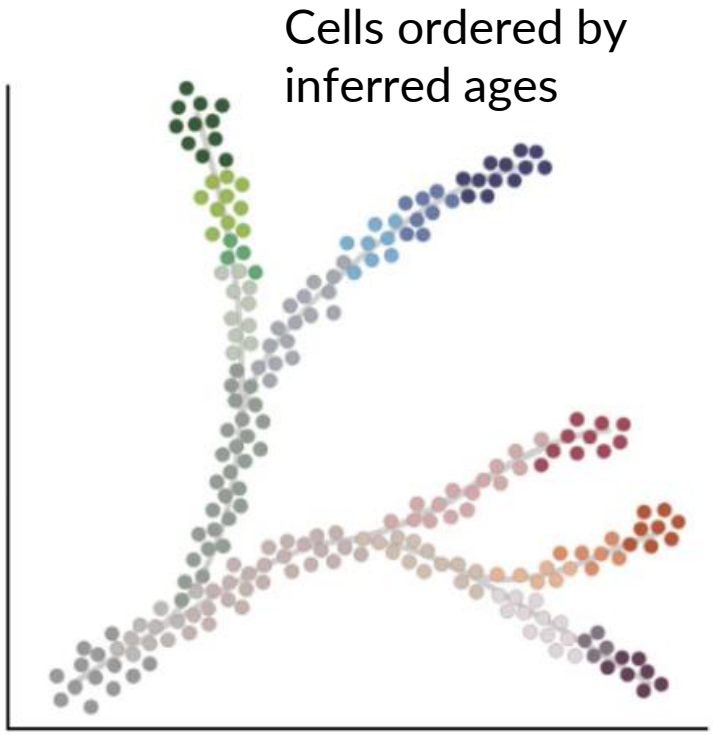
Differentiation

Treatment response



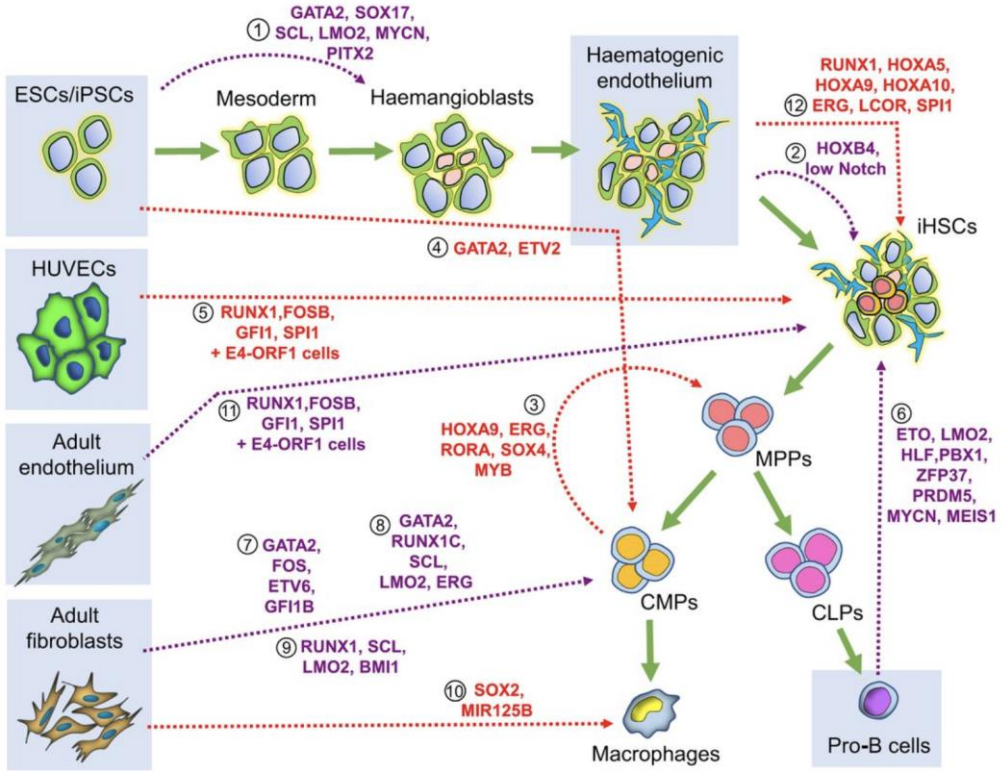
Cell development through single-cell data





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

Detecting gene switches

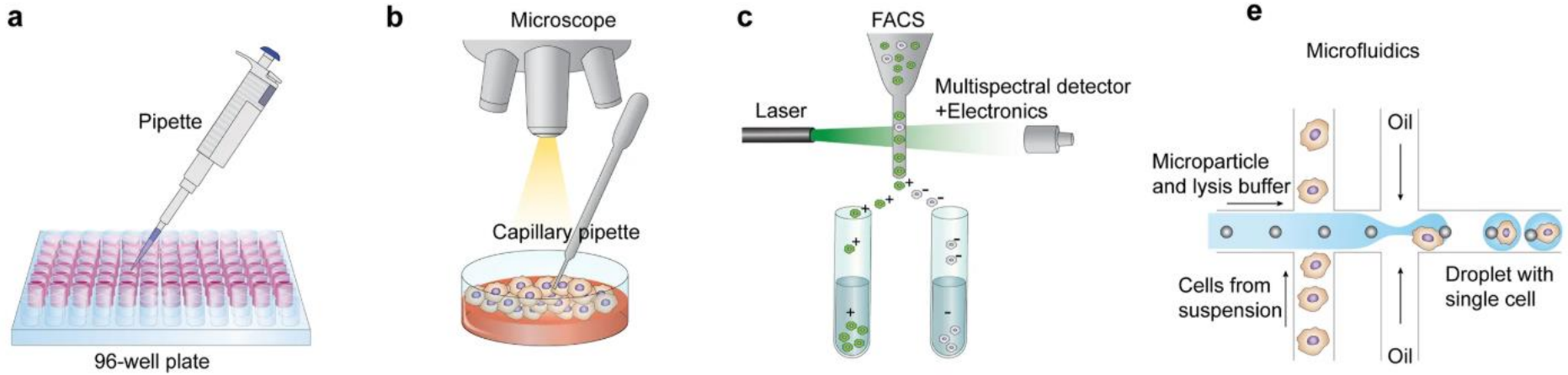


- Genes whose expressions:
- Change over inferred developmental time
- Diverge between two developmental branches
- Turn on/off across cell type

Source: Ivanovs et al. Development 144:2323-37 (2017)

Single-cell vs bulk transcriptomics

Cell isolation techniques



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

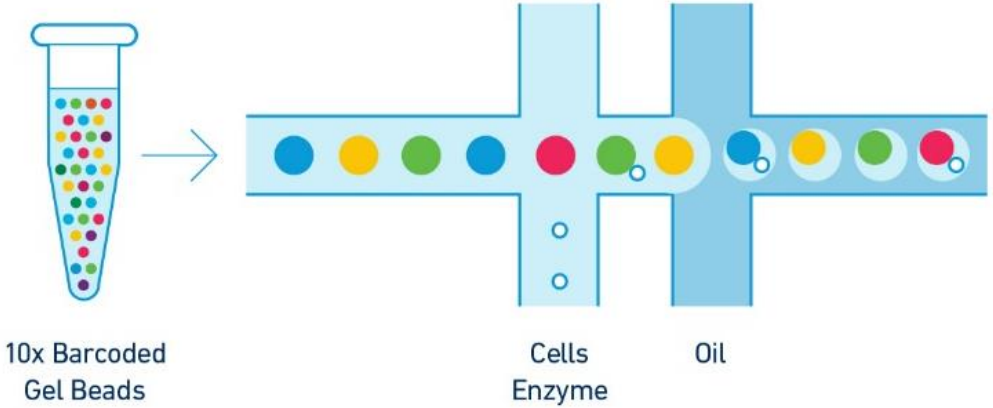
Diverse protocol choices

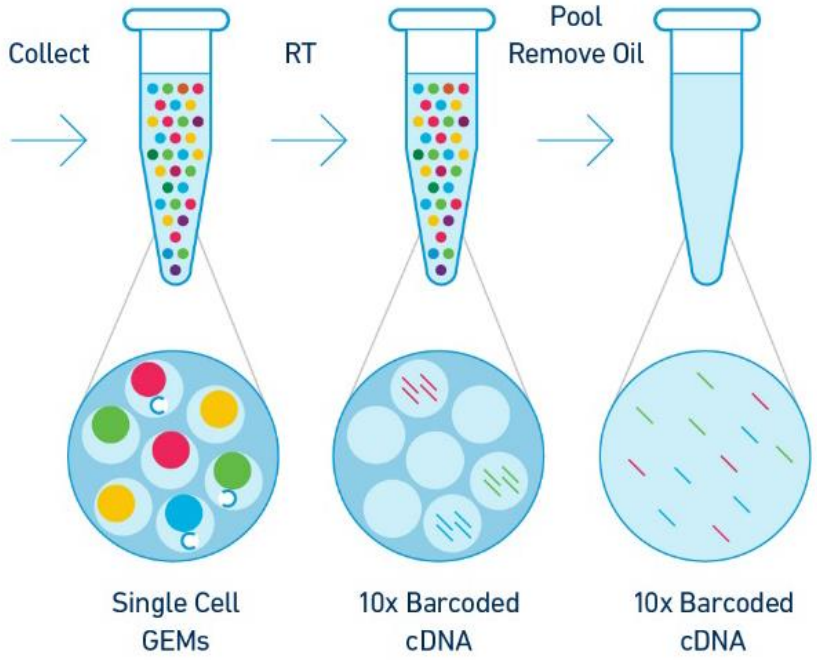
Methods	Transcript coverage	UMI possibility	Strand specific	References
Tang method	Nearly full-length	No	No	Tang et al., 2009
Quartz-Seq	Full-length	No	No	Sasagawa et al., 2013
SUPeR-seq	Full-length	No	No	Fan X. et al., 2015
Smart-seq	Full-length	No	No	Ramskold et al., 2012
Smart-seq2	Full-length	No	No	Picelli et al., 2013
MATQ-seq	Full-length	Yes	Yes	Sheng et al., 2017
STRT-seq and STRT/C1	5′-only	Yes	Yes	Islam et al., 2011, 2012
CEL-seq	3′-only	Yes	Yes	Hashimshony et al., 2012
CEL-seq2	3′-only	Yes	Yes	Hashimshony et al., 2016
MARS-seq	3′-only	Yes	Yes	Jaitin et al., 2014
CytoSeq	3′-only	Yes	Yes	Fan H.C. et al., 2015
Drop-seq	3′-only	Yes	Yes	Macosko et al., 2015
InDrop	3′-only	Yes	Yes	Klein et al., 2015
Chromium	3′-only	Yes	Yes	Zheng et al., 2017
SPLiT-seq	3′-only	Yes	Yes	Rosenberg et al., 2018
sci-RNA-seq	3′-only	Yes	Yes	Cao et al., 2017
Seq-Well	3′-only	Yes	Yes	Gierahn et al., 2017
DroNC-seq	3′-only	Yes	Yes	Habib et al., 2017
Quartz-Seq2	3′-only	Yes	Yes	Sasagawa et al., 2018

Chen *et al.* Front Genet. 10:317 (2019)

- Mostly sequence 3’ ends of transcript
- Unique Molecular Identifier (UMI) = PCR barcode

10x Genomics' chromium technique

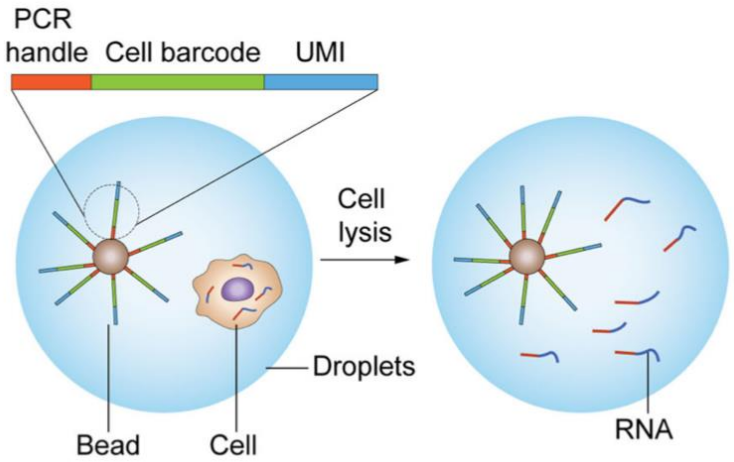




https://bauercore.fas.harvard.edu/10x-chromium-system

- Droplets of sequencing reagents
- Cell barcode + pooled sequencing

UMI and cell barcode



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

- All beads in each droplet have the same cell barcode
  - Reads with the same barcode came from the same cell
- Each PCR adapter contains different Unique Molecular Identifiers (UMI)
  - Reads with the same UMI came from the same original RNA molecule

Single-cell vs bulk data

Estimated Number of Cells

5,593

Mean Reads per Cell

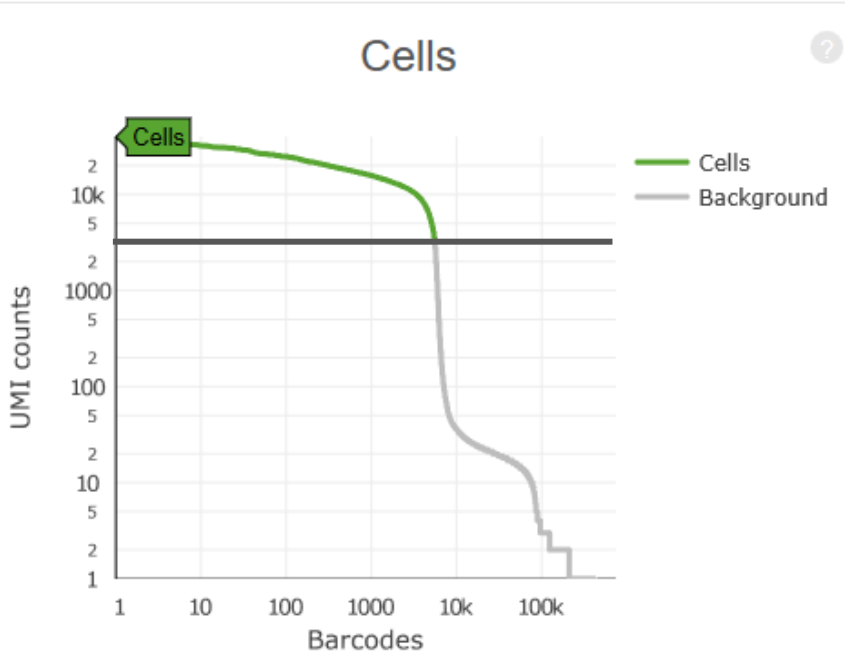
26,716

Median Genes per Cell

2,880

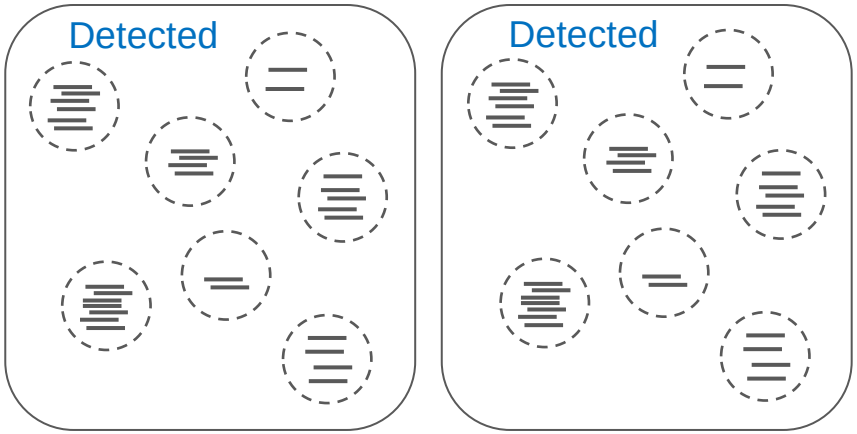
Sequencing

Number of Reads	149,425,634
Valid Barcodes	97.5%

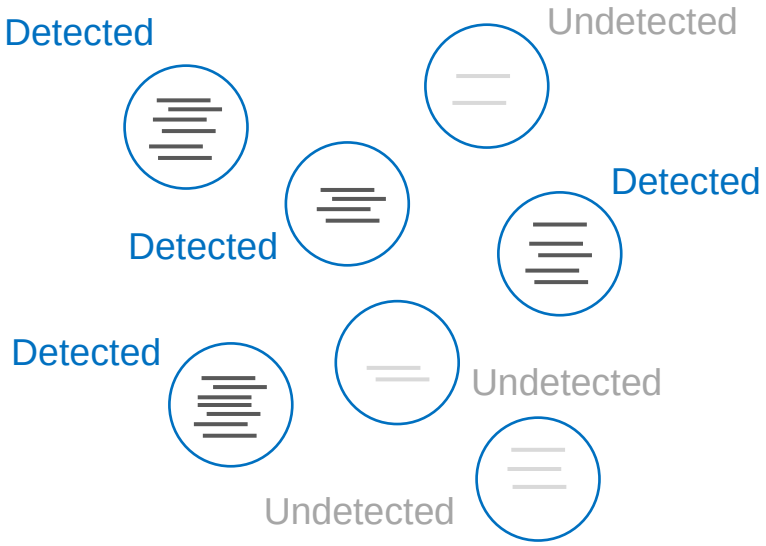


Detection limit of single-cell data

Bulk RNA-seq



Single-cell RNA-seq



Challenges in single-cell data analysis

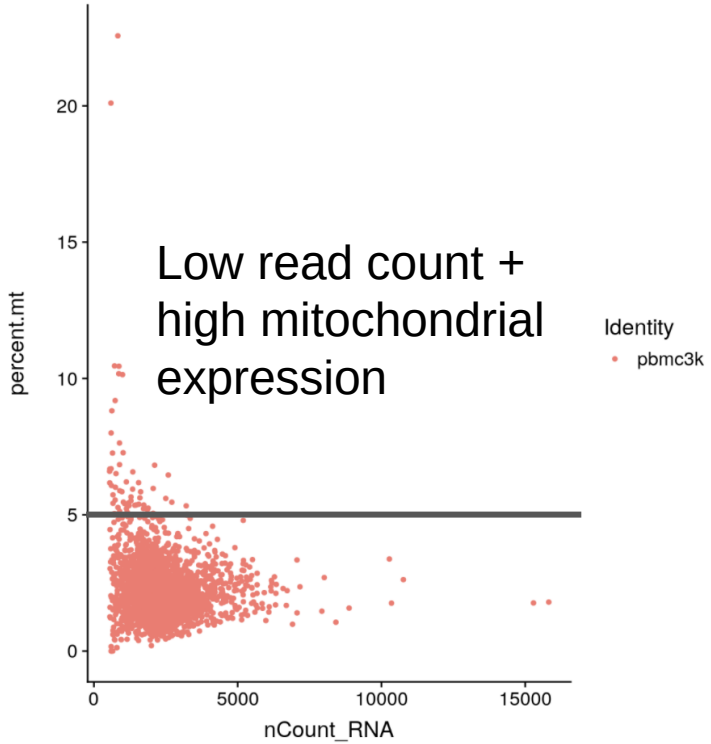
- Low read count per cell and gene
  - A lot of zeros in expression data
- Cells are biologically different
  - High variance across cells
- Cells are in continuous states of development
  - Not just control vs treatment
- Data is very large (256 GB of RAM for medium project)
  - 10,000 cells x 5,000 genes

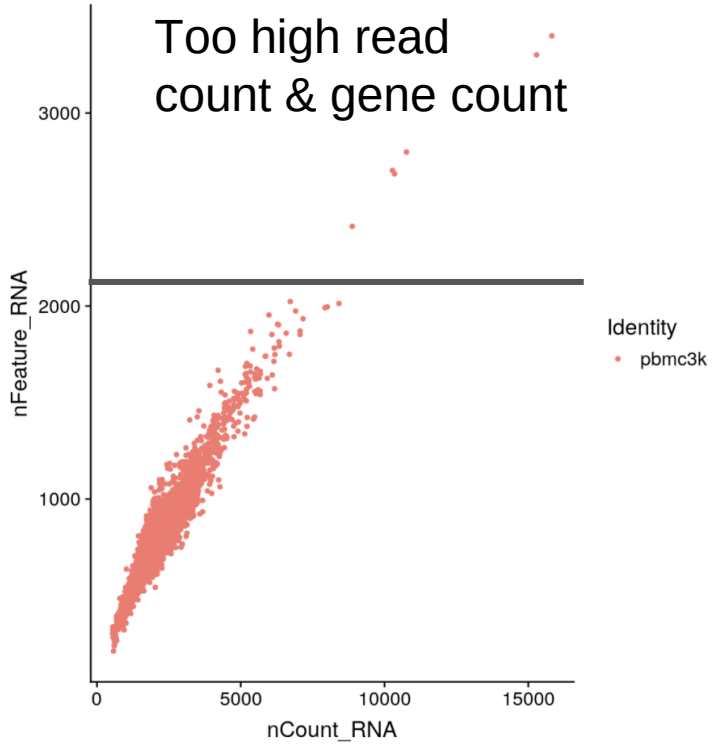
Single-cell data preprocessing

Key steps in single-cell data processing

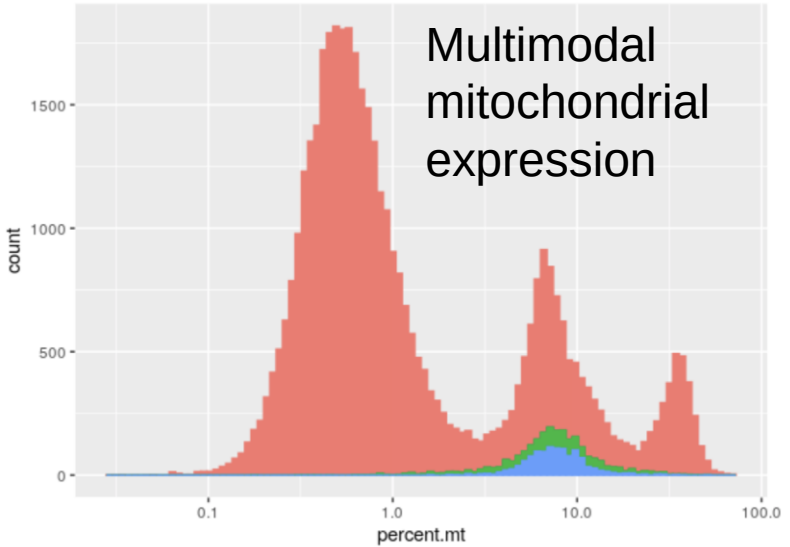
- Quality filter
  - Low read count & gene count = non-cells
  - Very high read count & gene count = multi-cells
  - High mitochondrial expression = dead cells
- Within-sample normalization
  - Dealing with missing expression values
- Multi-sample integration
  - Single-cell data have strong batch effects

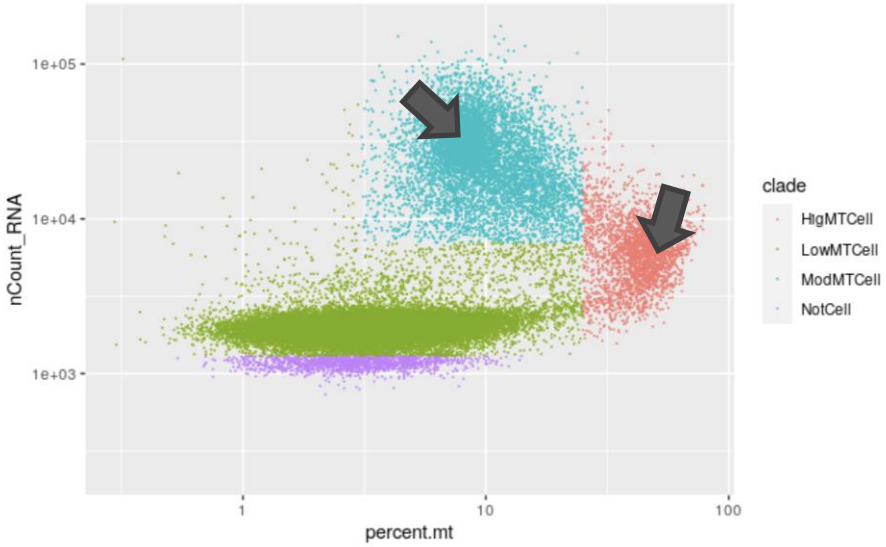
Basic quality filters





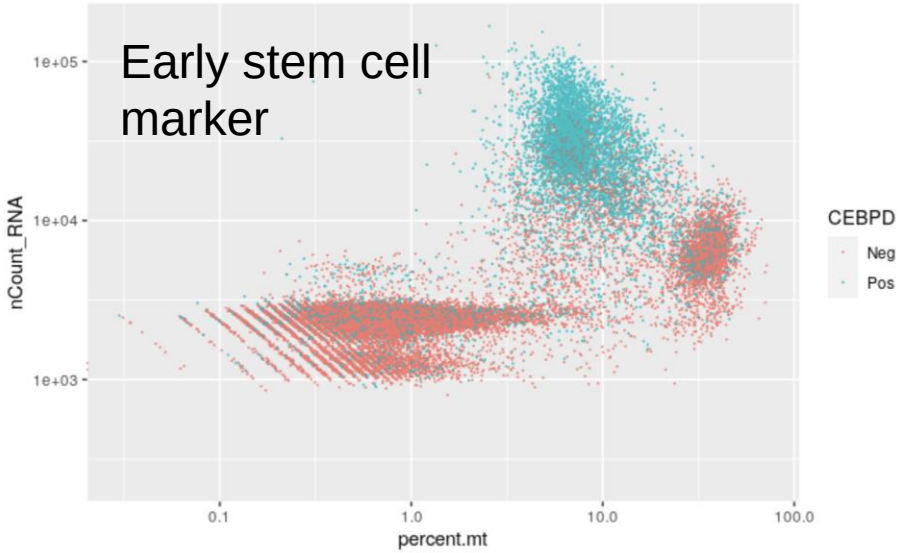
An exception

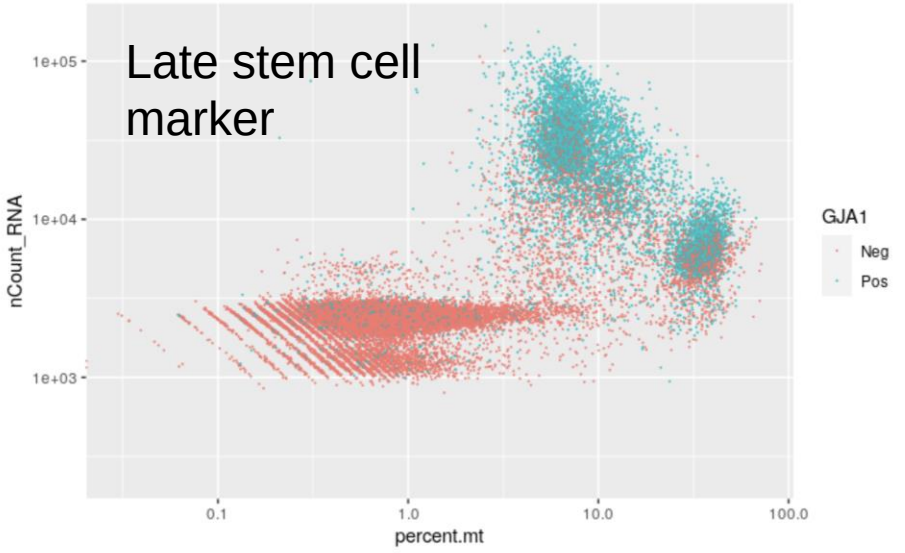




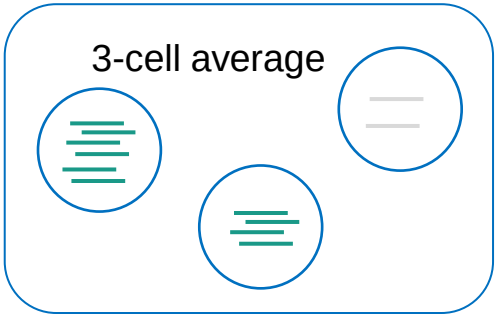
- High mitochondrial activity in stem cells and some cell types

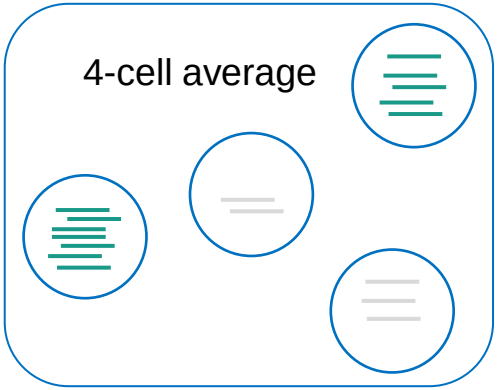
Stem cell markers in high-mitochondrial cells





Normalization with pooling





Low expression

Gene A

Gene B

Gene C

Medium expression

Gene D

Gene E

Gene F

Gene G

High expression

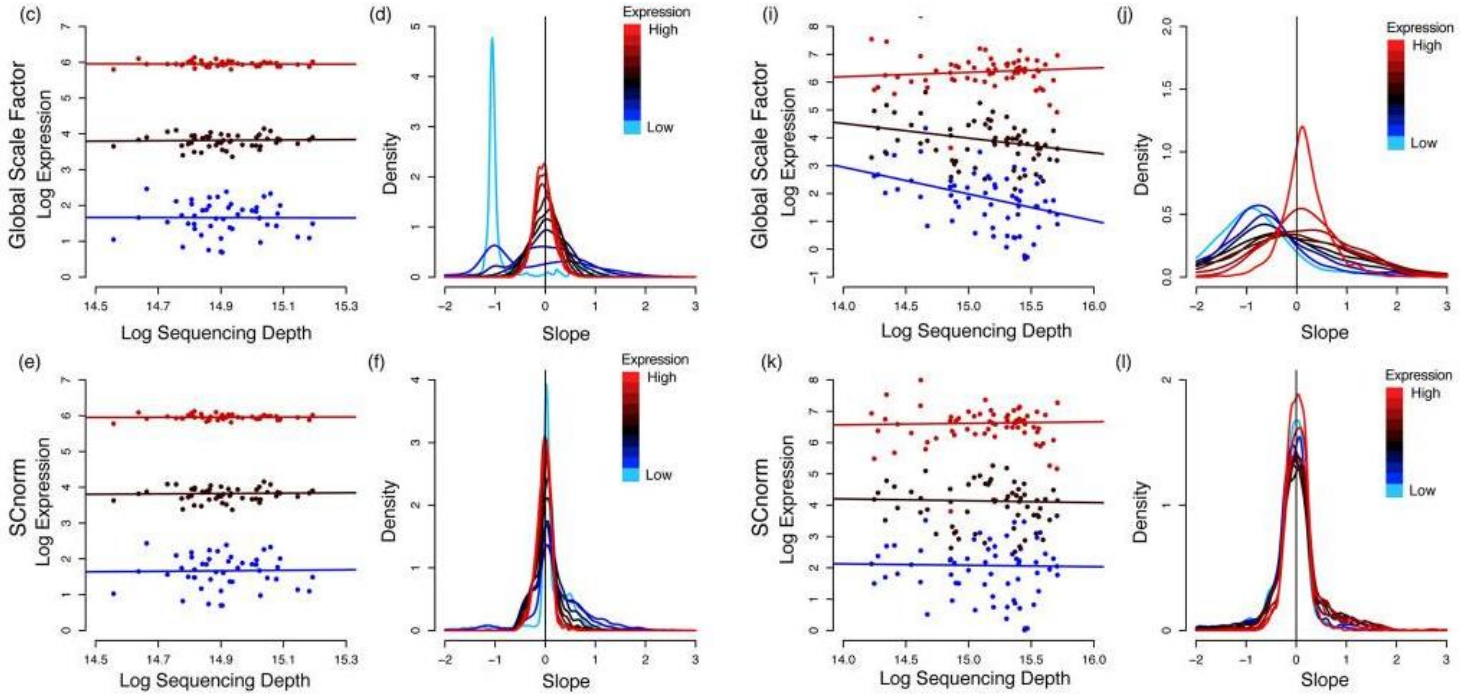
Gene H

Gene I

Expression-dependent scaling factor

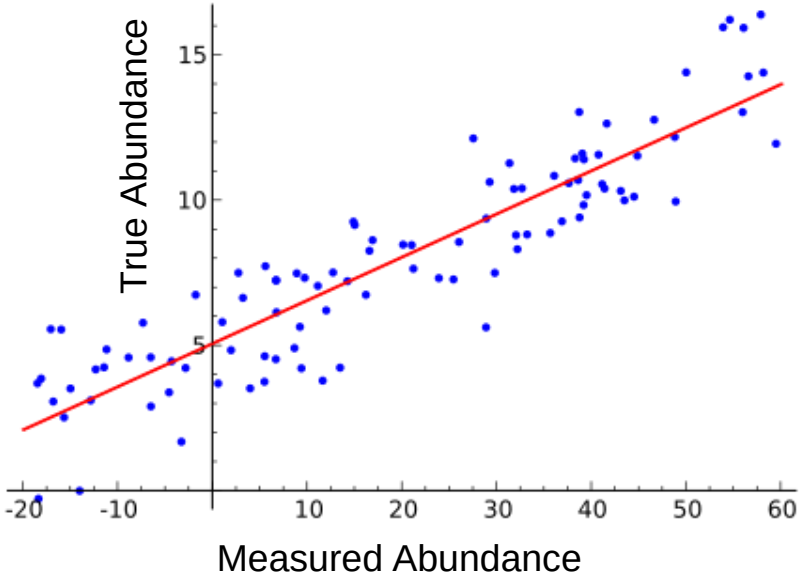
Bulk RNA-seq

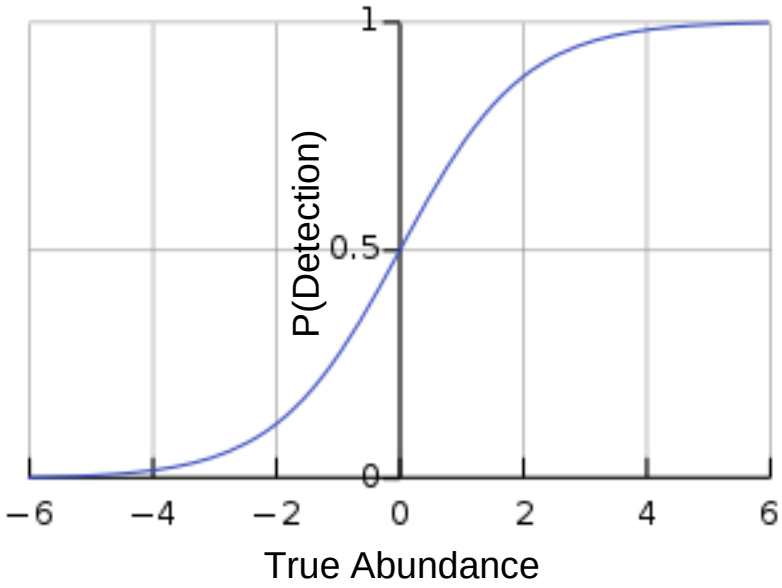
Single-cell RNA-seq



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

Modeling of detection probability

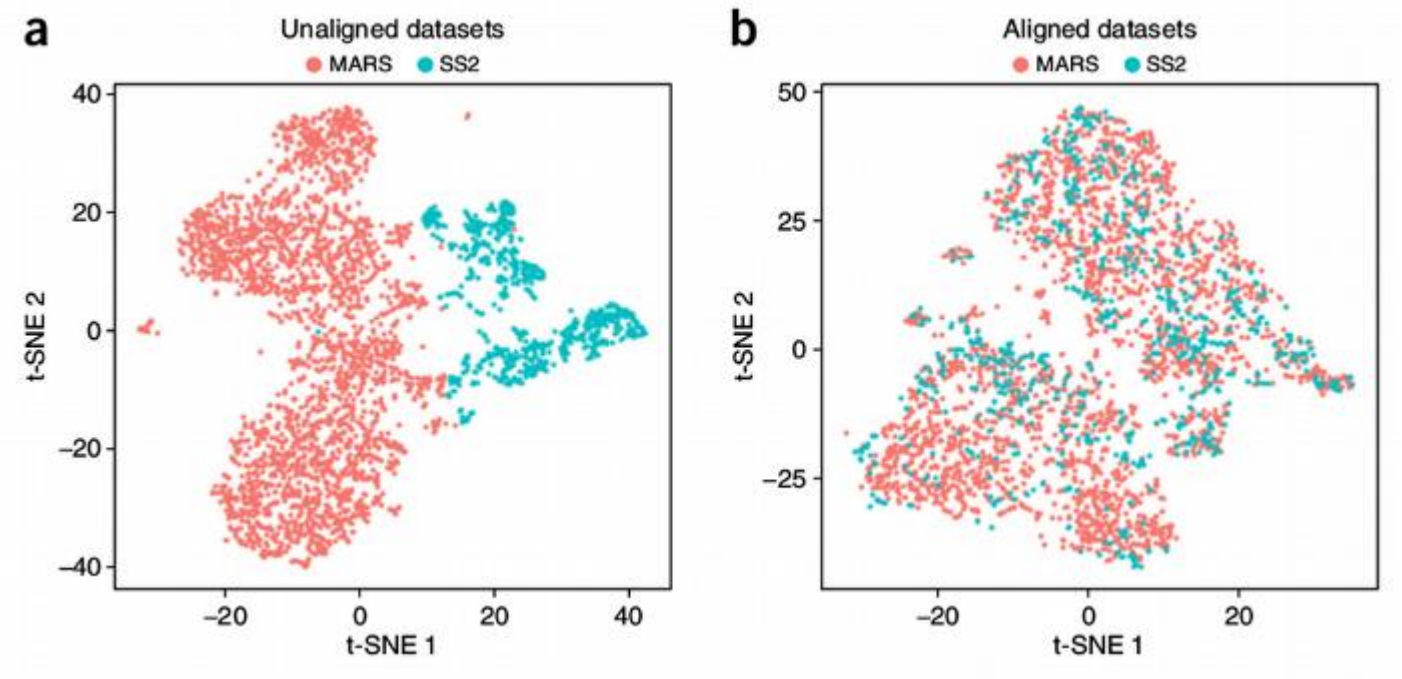




- Measured abundance = f(true abundance) x P(detection | true abundance)

Single-cell data integration

High bias across datasets



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

Linear effect removal (Combat-Seq)

Negative binomial regression models

Gene-wise model: for a certain gene *g*, count in sample *j* from batch *i* $y_{gij} \sim NB(\mu_{gij}, \phi_{gi})$

Explicit addition of batch parameter $\gamma_{g,i}$

$$\log \mu_{gij} = \alpha_g + X_j \beta_g + \gamma_{gi} + \log N_j$$

$$Var(y_{gij}) = \mu_{gij} + \phi_{gi} \mu_{gij}^2$$

Decompose scaled counts into 3 components

- α_g — Average level for gene g (in "negative" samples)
- $X_j \beta_g$ — Biological condition of sample j
- γ_{gi} — Mean batch effect

N_j = total read count for sample j

ϕ_{gi} Dispersion batch effect

Estimate batch effect parameters — Estimate parameters using established methods in edgeR

Calculate "batch-free" distributions

We assume the adjusted data also follow a negative binomial distribution: $y^*_{gj} \sim NB(\mu^*_{gj}, \phi^*_g)$

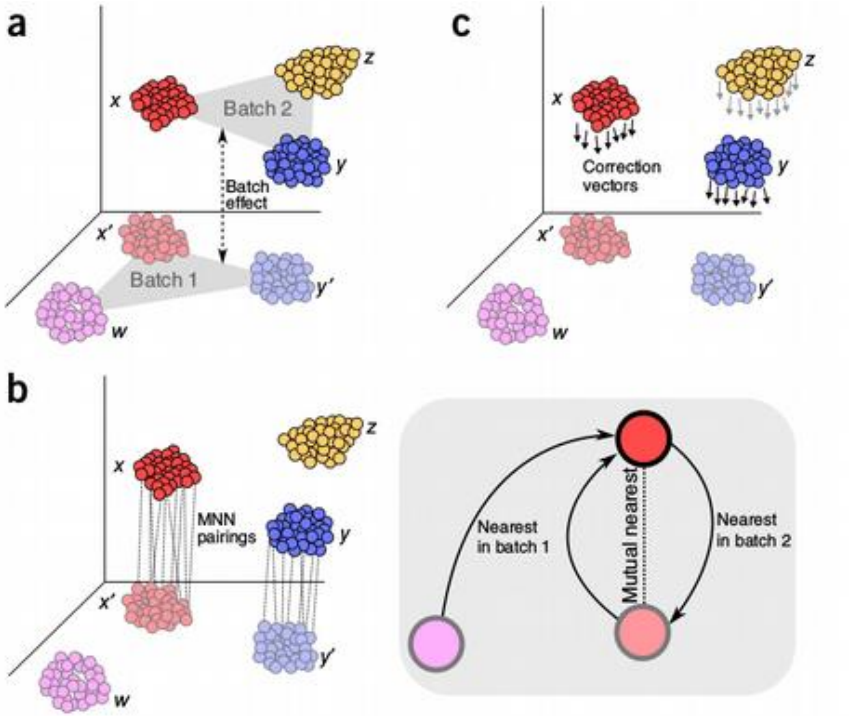
Subtract batch parameter $\gamma_{g,i}$

$$\log \mu^*_{gj} = \log \hat{\mu}_{gij} - \hat{\gamma}_{gi}$$

$$\phi^*_g = \frac{1}{N_{batch}} \sum_i \hat{\phi}_{gi}$$

Zhang, Y. *et al.* NAR Genom and Bioinfo 2:lqaa078 (2020)

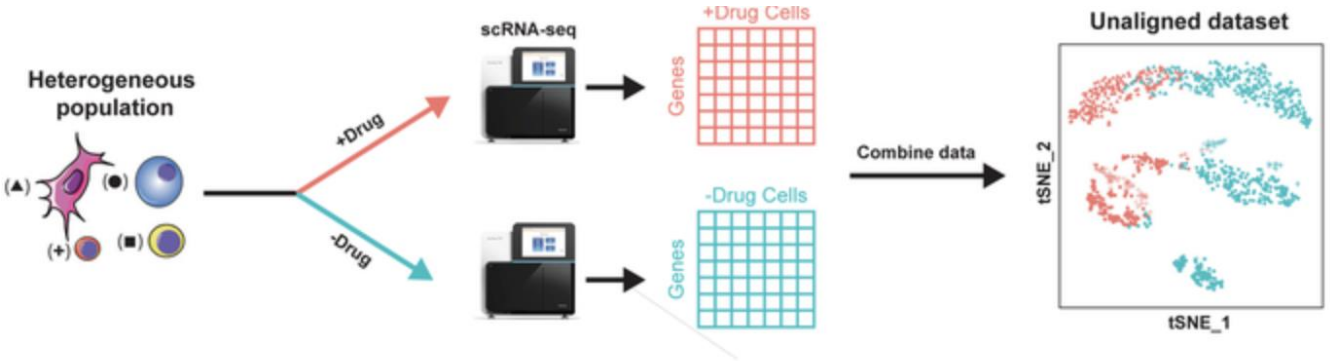
Mutual nearest neighbor (MNN)



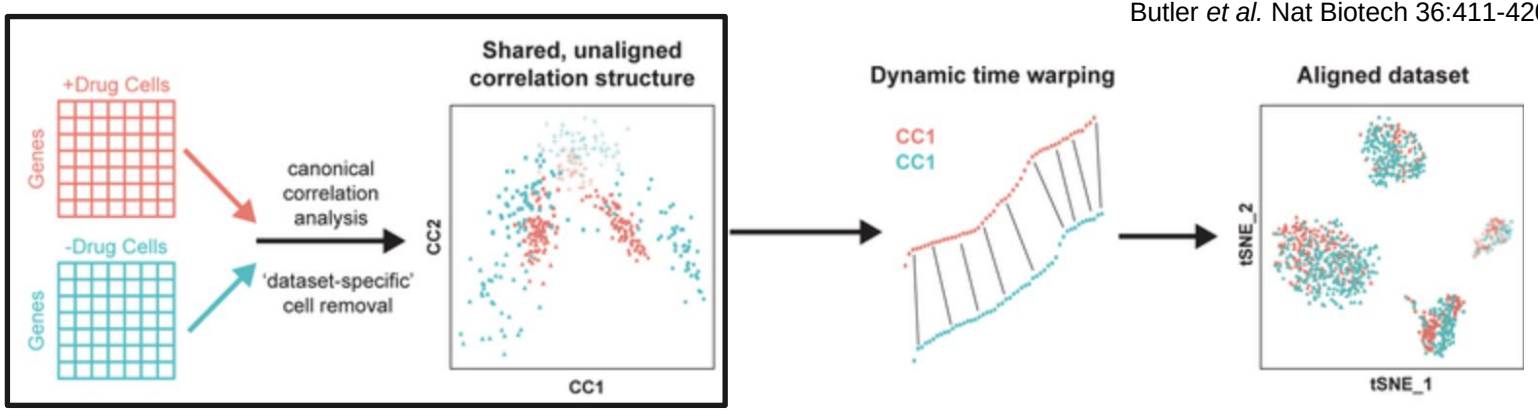
- Map clusters of cells together rather than individual cells
- Similar to reciprocal best hits from BLAST for identifying orthologs
- Apply the average mapping vector to unique cell types

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

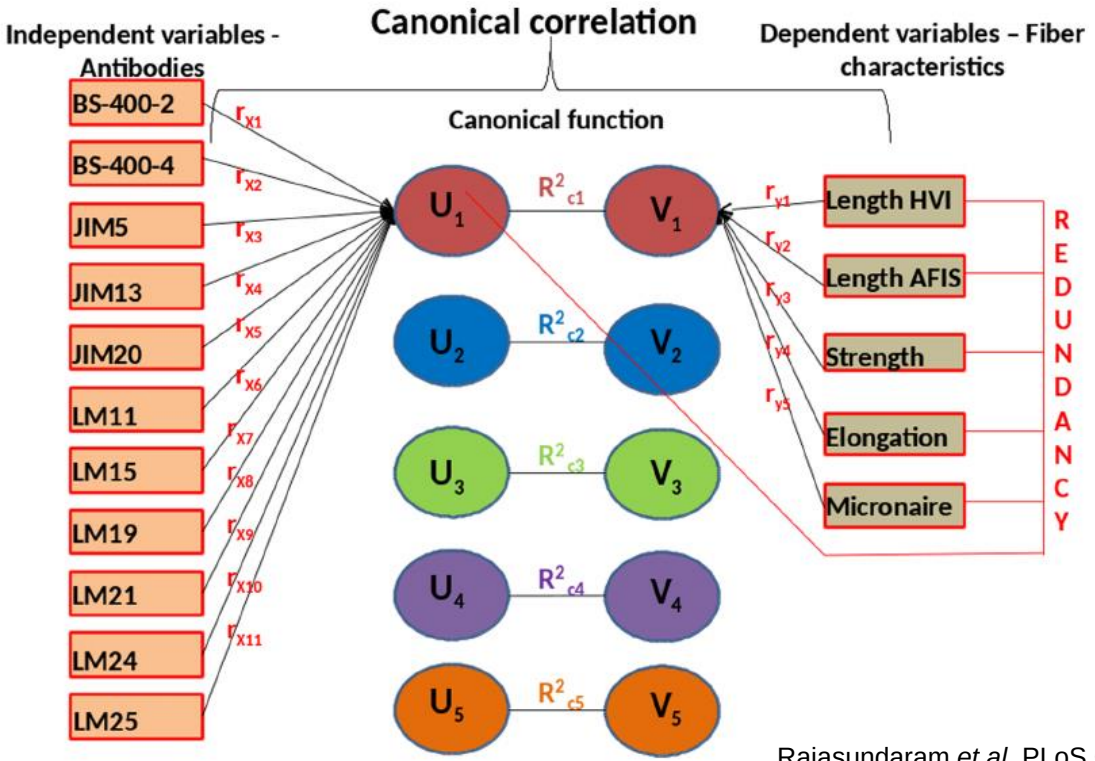
Integration via canonical correlation



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



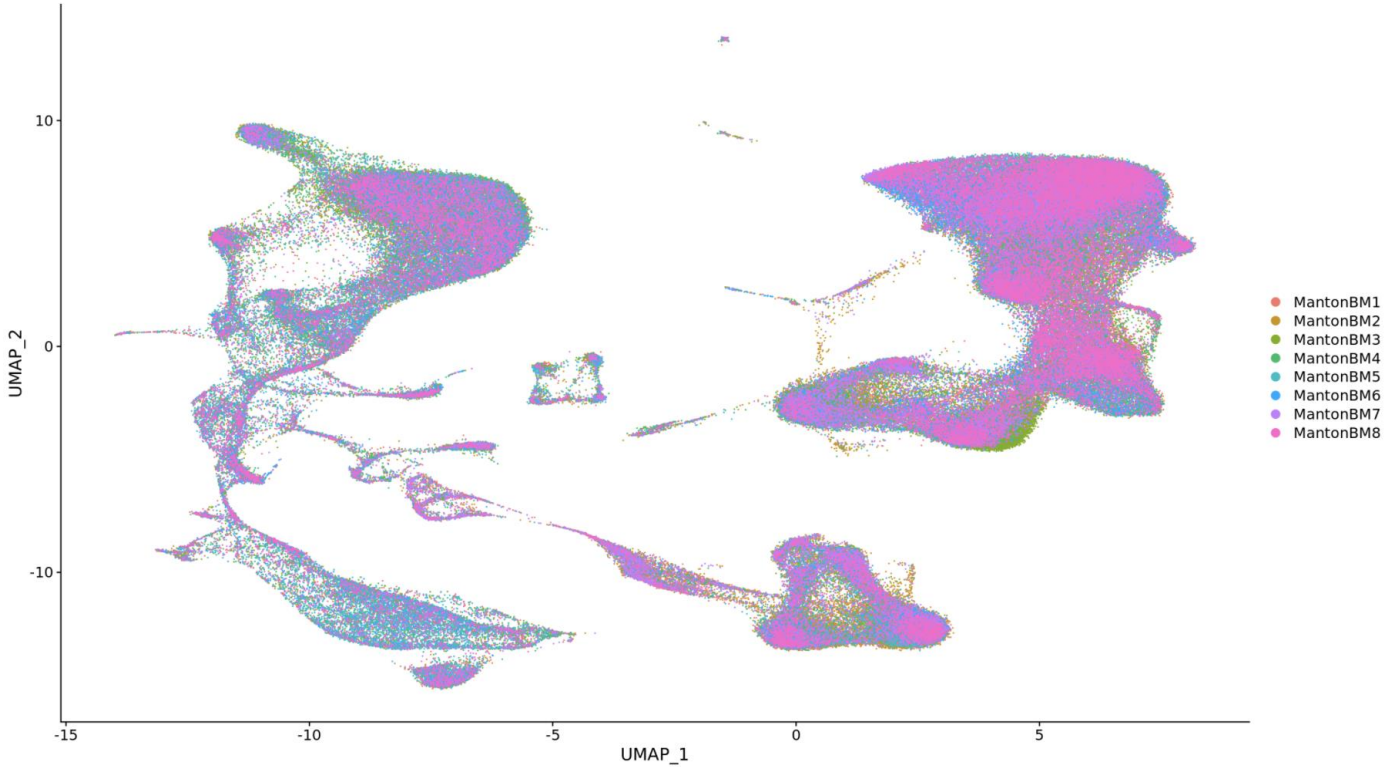
Canonical correlation analysis (CCA)



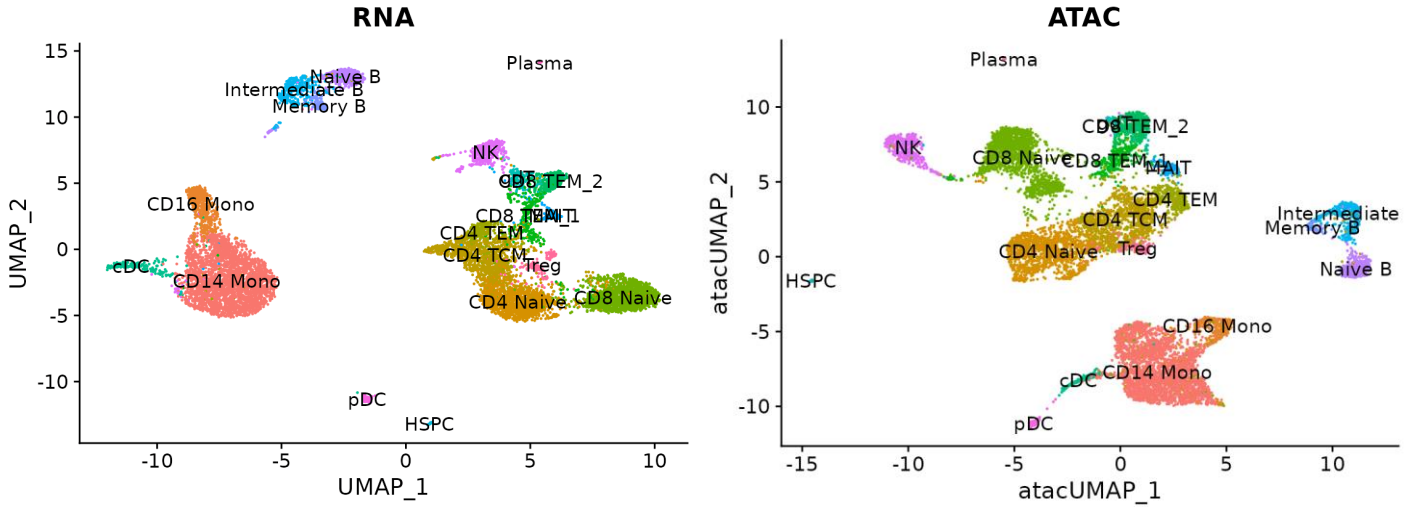
- Same samples with two different systems of observations
- Identify correlation structure observation systems (features)

Rajasundaram *et al.* PLoS ONE (2014)

Good data integration



Integrating RNA-seq with ATAC-seq

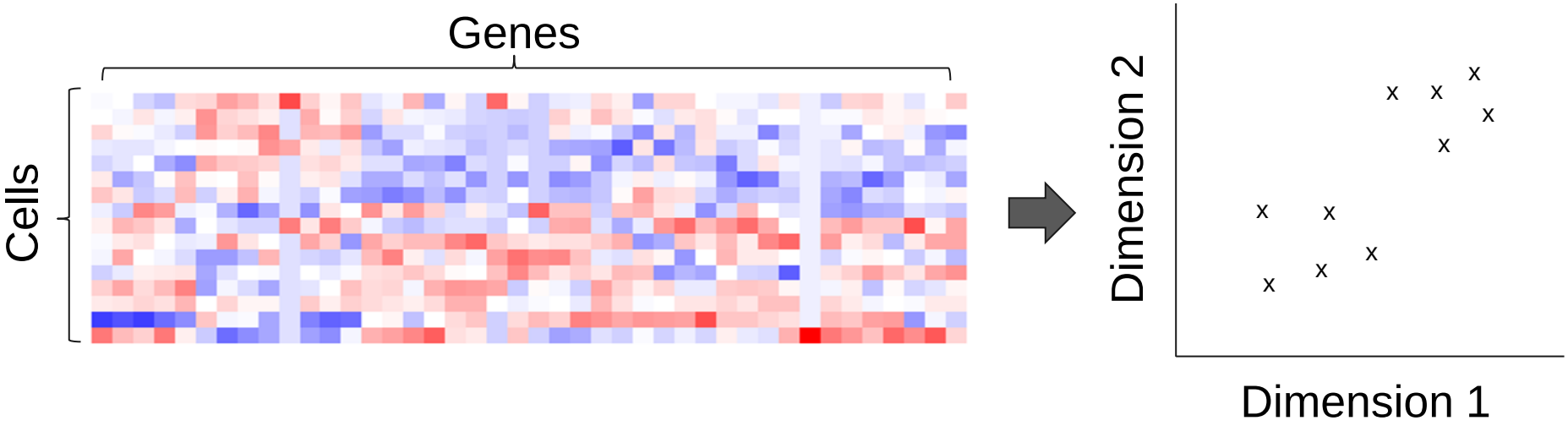


- ATAC-seq = open chromatin ~ gene expression level
- Transfer cell type label

Visualizing single-cell data

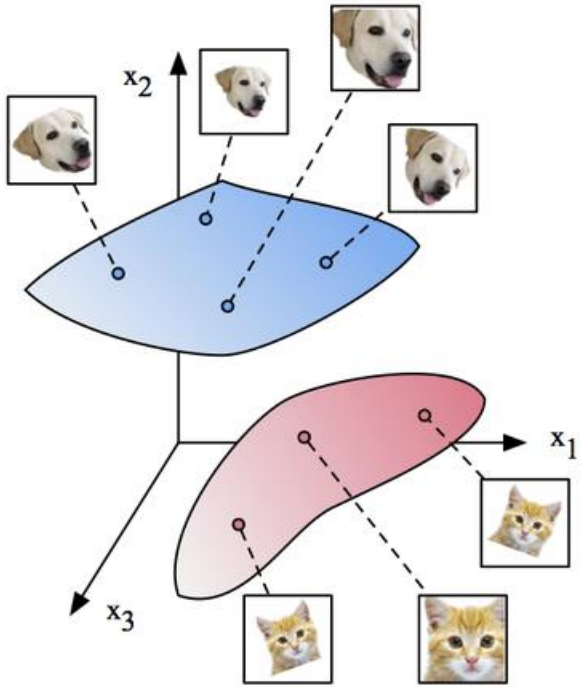
with dimensionality reduction techniques

Dimensionality reduction

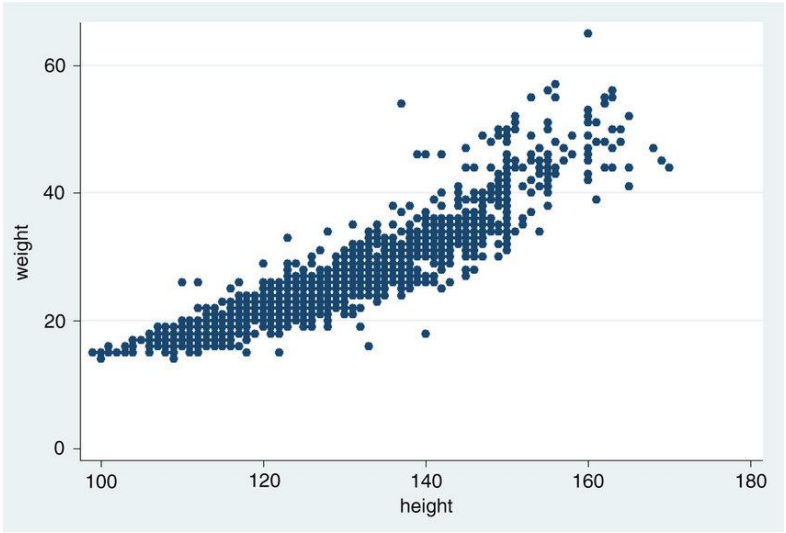


- Collapse high-dimensional data on to 2D or 3D scatter plot that preserve some information in original dimension

Manifold hypothesis



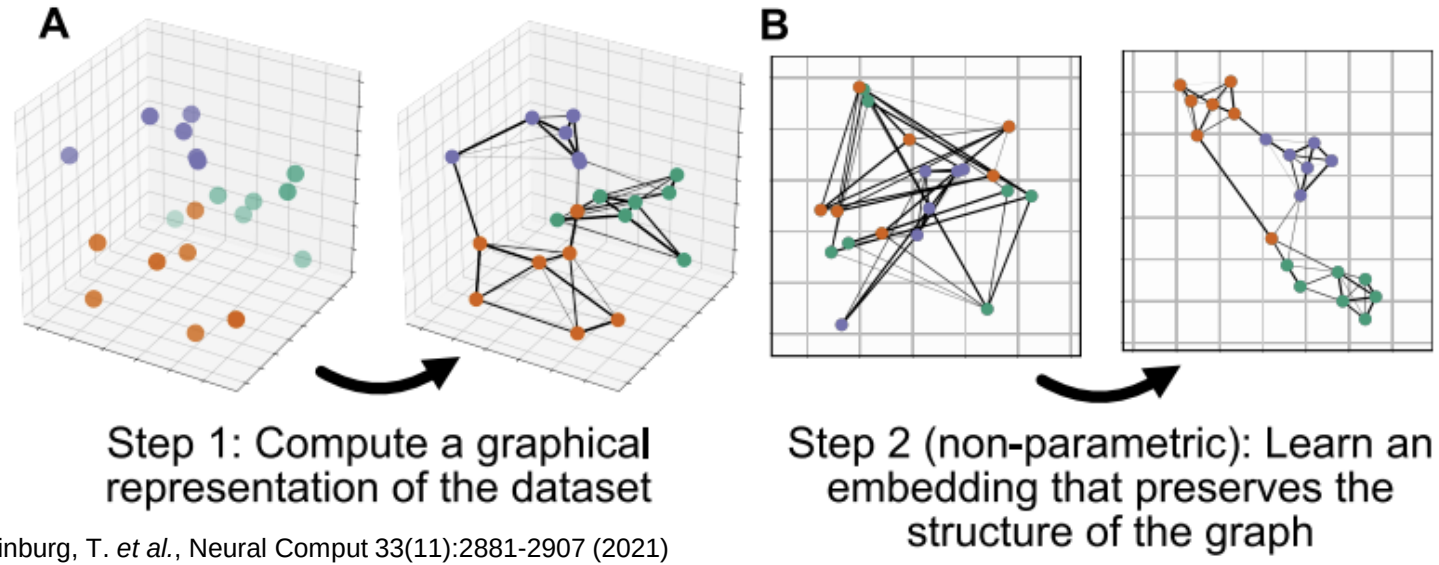
"Real-world, high-dimensional data lie on some low-dimensional manifolds"



Chung, S. et al. "Classification and Geometry of General Perceptual Manifolds"

Nordin, P. et al. Global Health Action 7:25351 (2014)

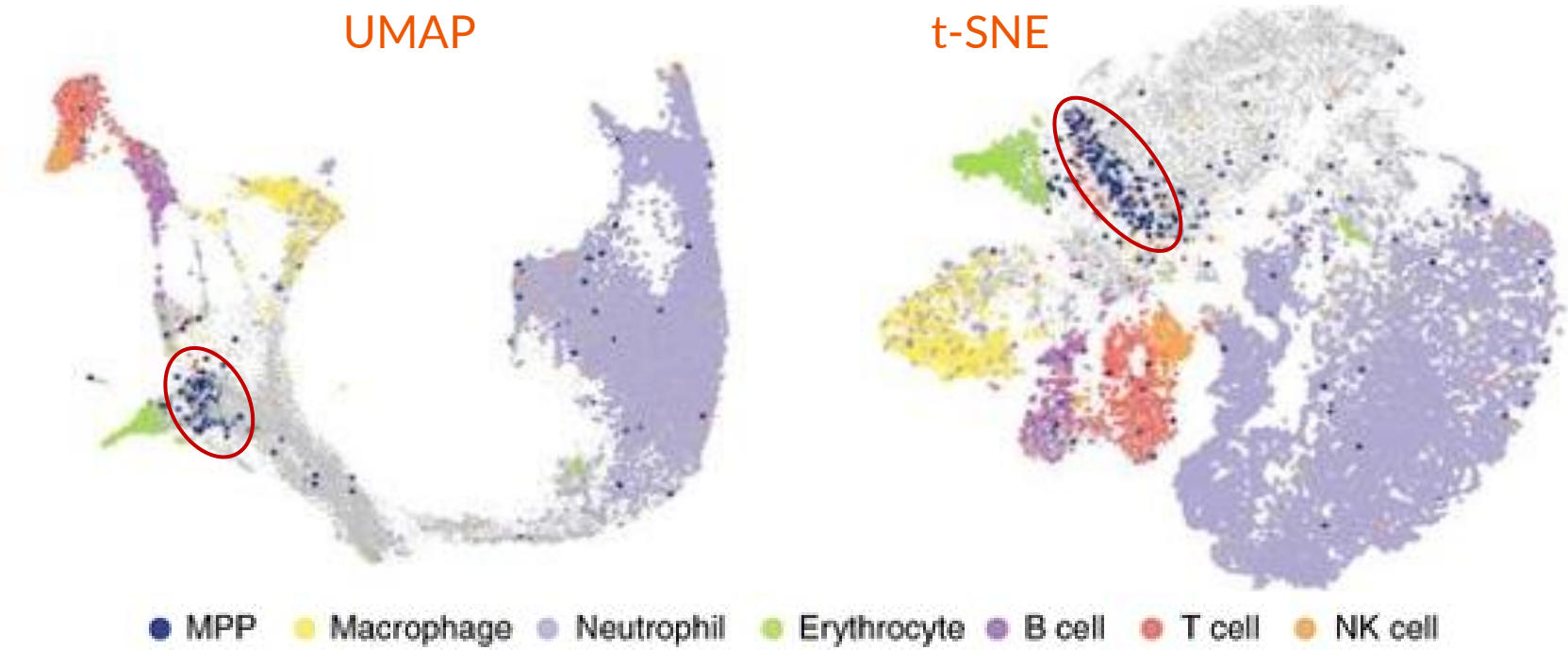
Dimensionality reduction algorithm sketch



Sainburg, T. *et al.*, Neural Comput 33(11):2881-2907 (2021)

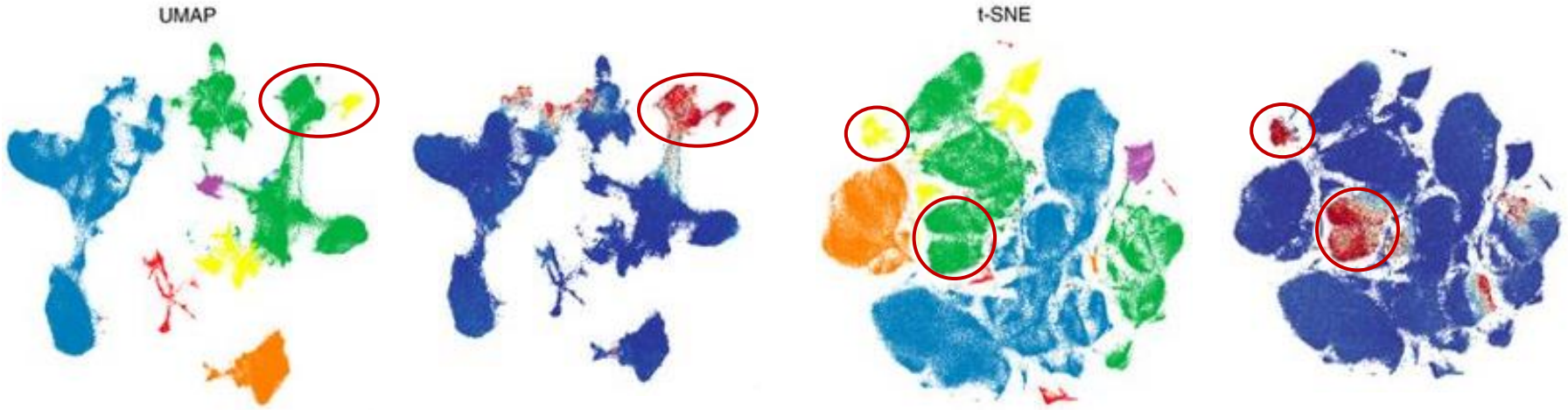
- Similarity = correlation in gene expression across cells

t-SNE vs UMAP on single-cell data



Becht, E. et al. Nature Biotechnology 37:38-44 (2019)

t-SNE vs UMAP on single-cell data

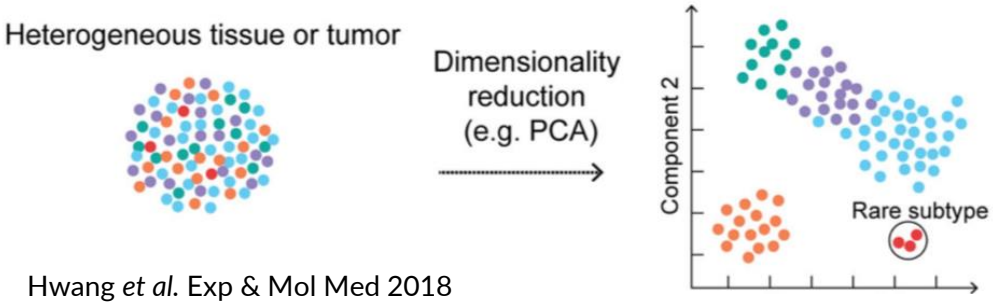


Becht, E. et al. Nature Biotechnology 37:38-44 (2019)

- Both are equally good at detecting individual cell types
- But UMAP is better at capturing transitions across cell types

Single-cell analysis overview

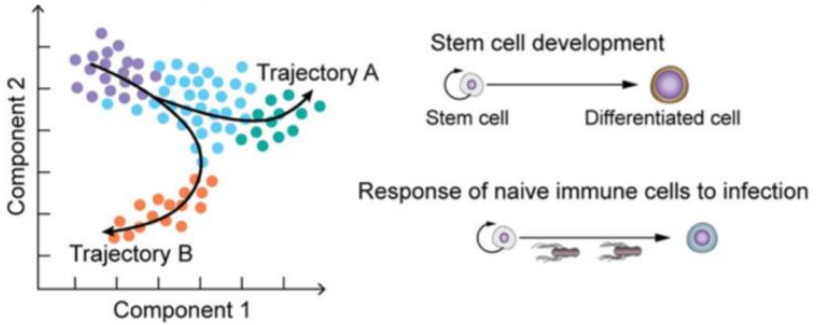
Cell clustering and trajectory reconstryction



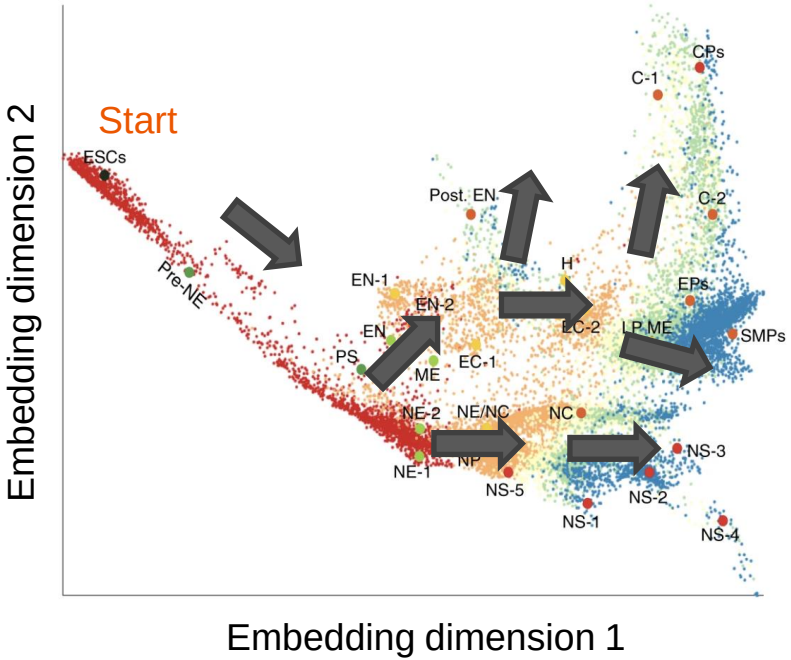
Hwang *et al.* Exp & Mol Med 2018

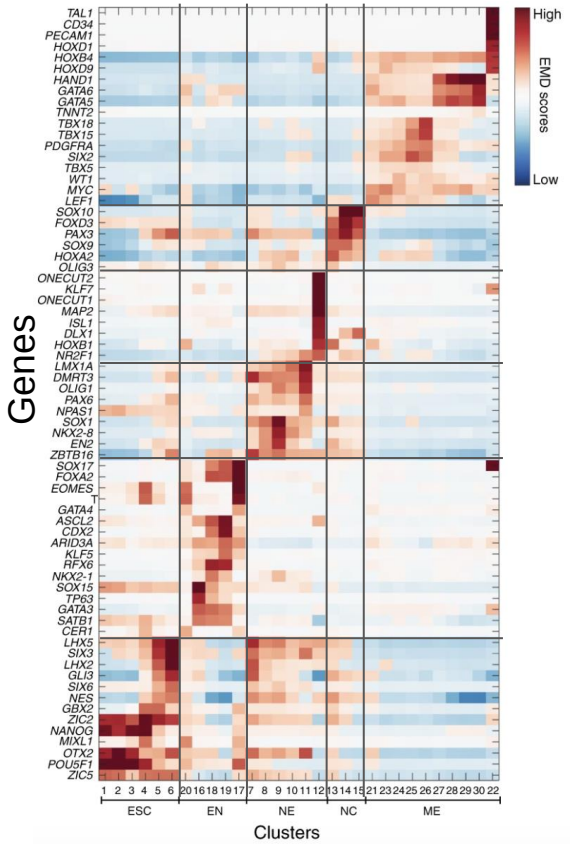
Clustering of cells with similar omics signatures reveal groups of different cell types and developmental stages

Trajectory modeling with random walk, diffusion, or Markov chain reconstruct the paths of cell development



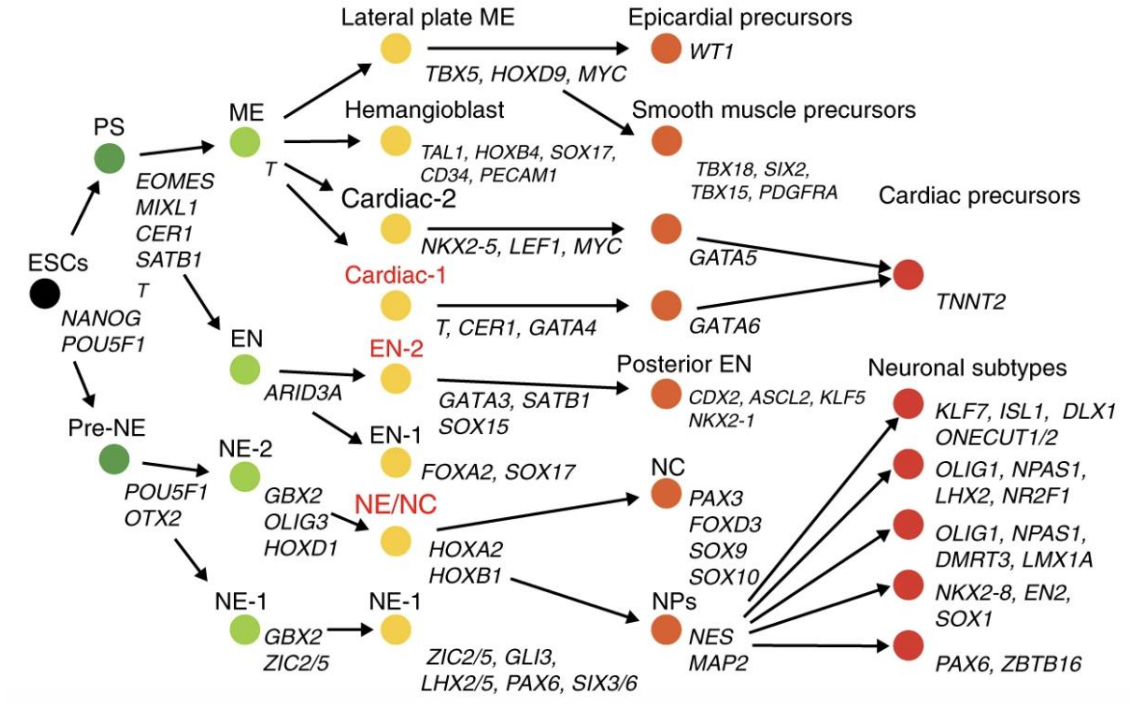
An end-to-end example





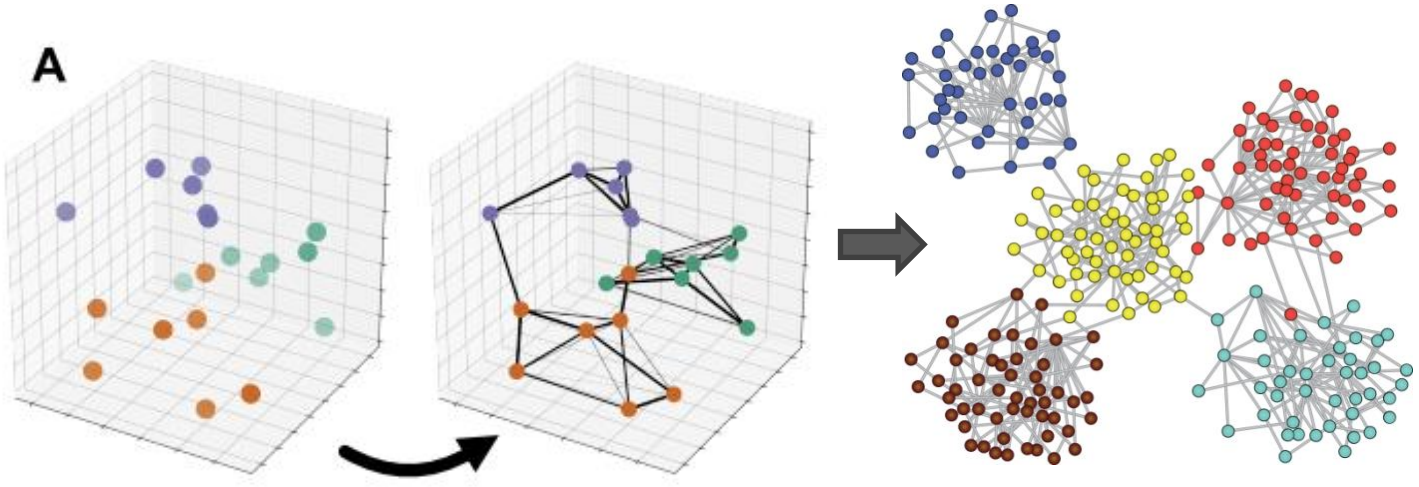
Source: Moon et al. Nature Biotechnology 37:1482-92 (2019)

An end-to-end example



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

Algorithm sketch for cell type clustering



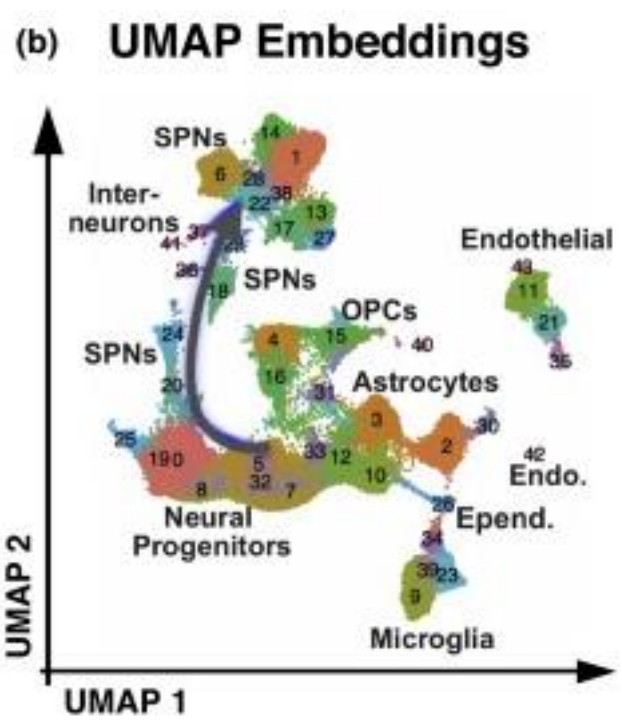
Sainburg, T. *et al.*, Neural Comput 33(11):2881-2907 (2021)

https://github.com/topics/graph-clustering

- Connect cells with similar gene expression profile
- Split network into modules with dense edges

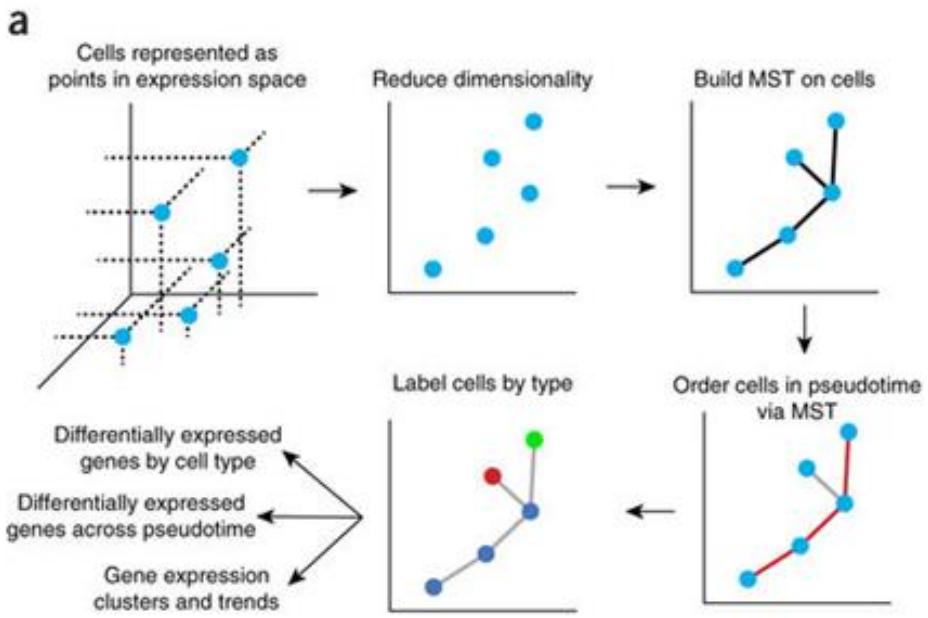
Cell developmental trajectory

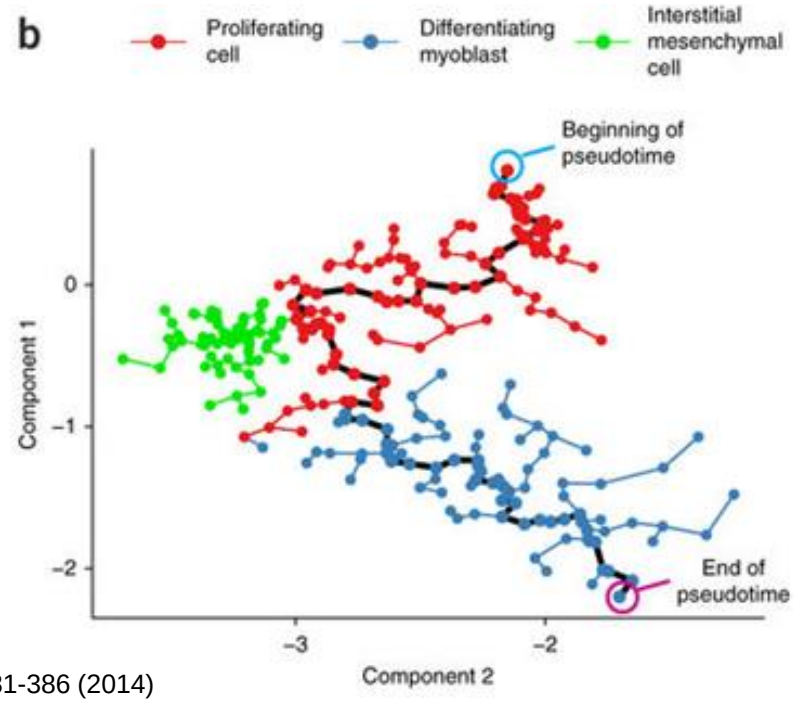
Cells can be arrange along developmental path



- Cells are in continuous developmental states
- Similar gene expression implies similar state
- Reconstruct pseudotime
- Identify important genes for development
  - Expression change along the trajectory
  - Expression switch at a particular time point

Minimum spanning tree

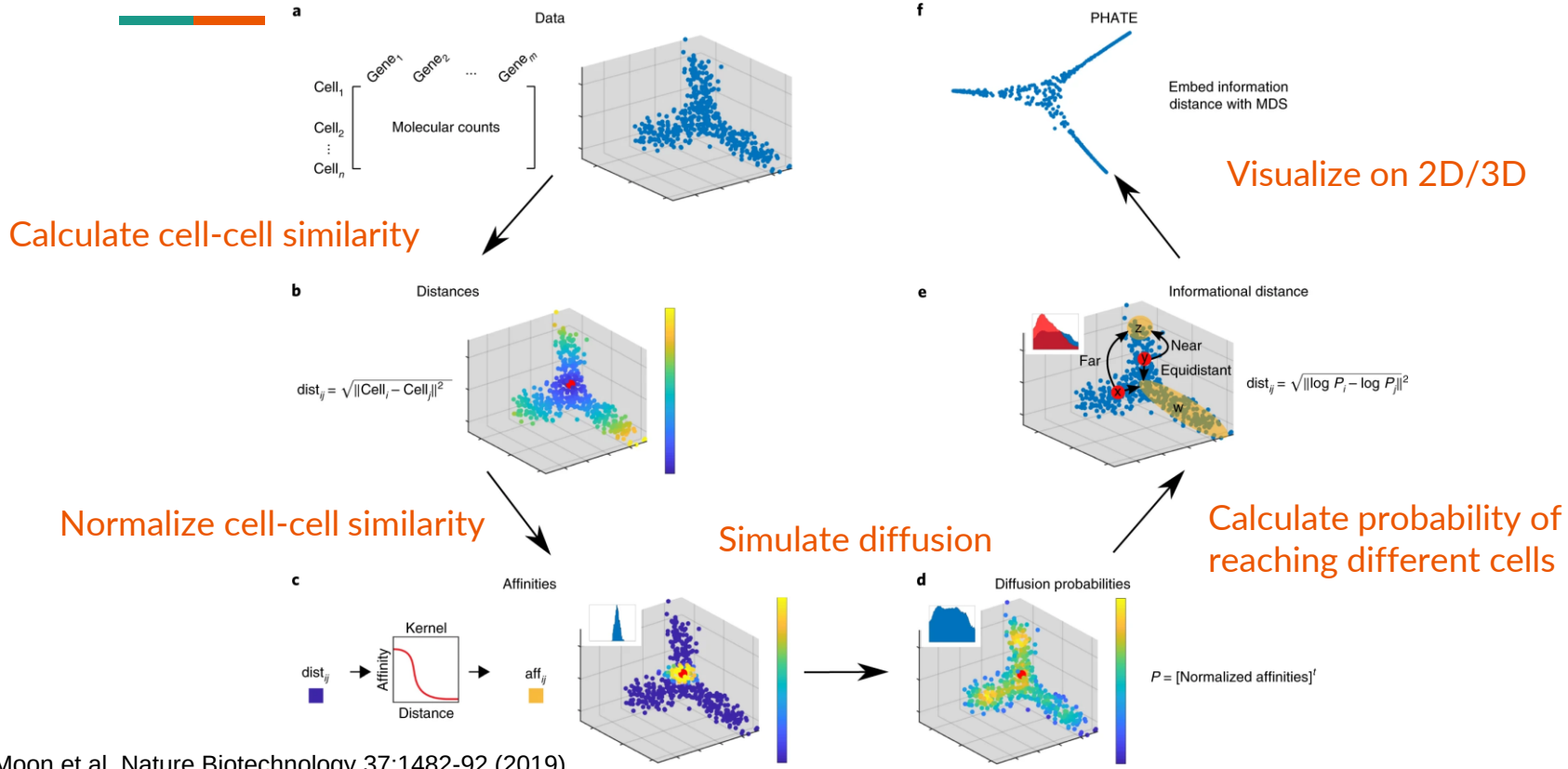




Trapnell *et al.* Nat Biotech 32:381-386 (2014)

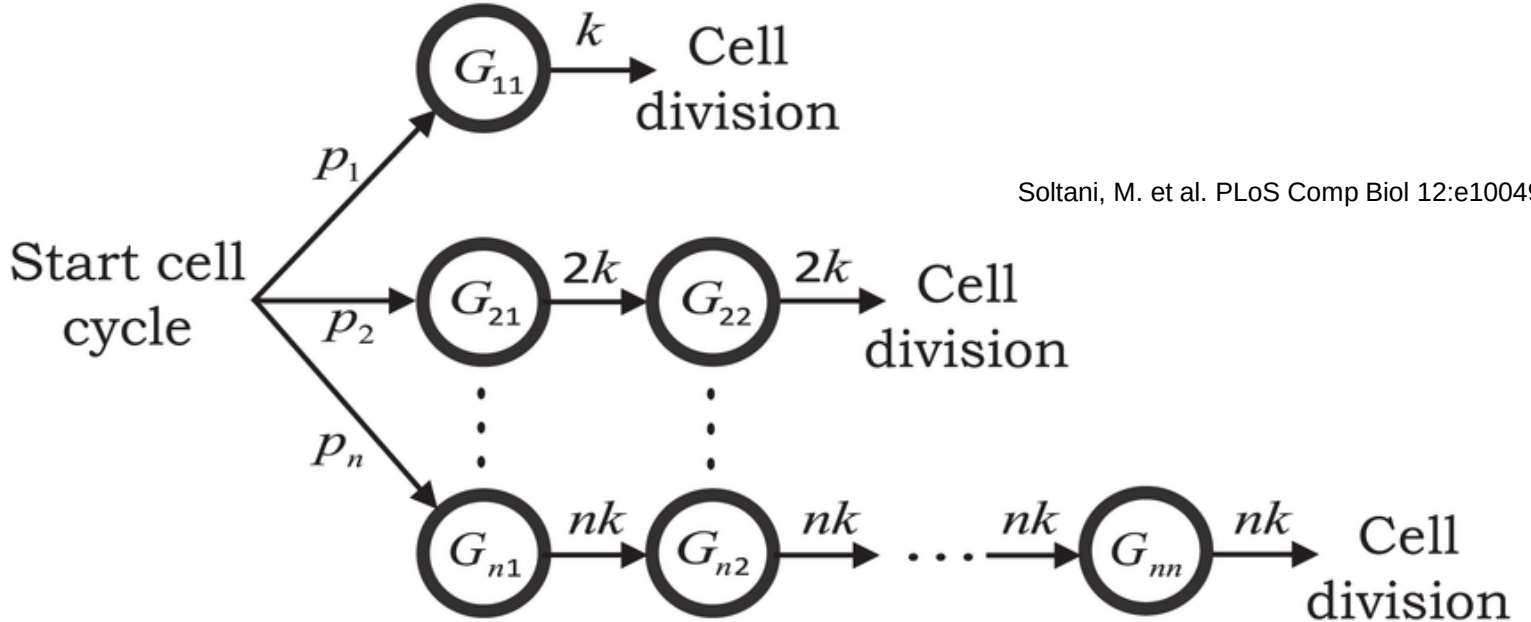
- Trajectory along the most similar cells

Diffusion approach for cell-cell transitions



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

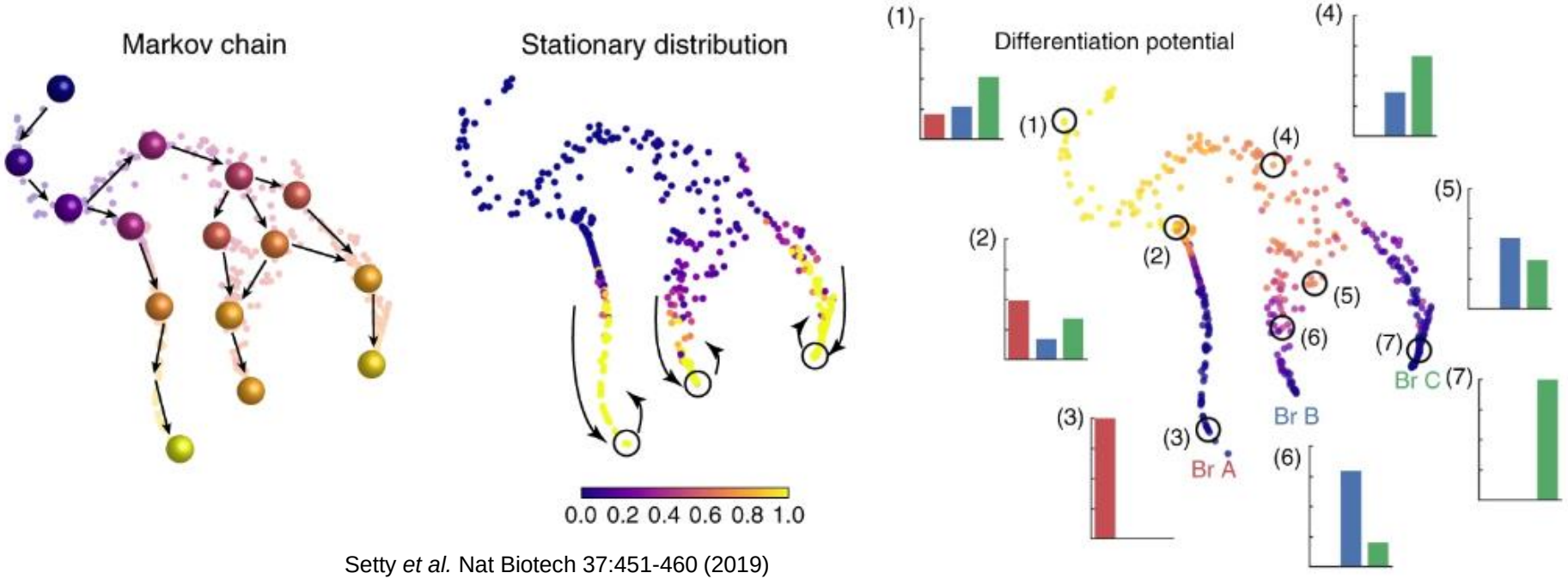
Markov chain model for cell development



Soltani, M. et al. PLoS Comp Biol 12:e1004972

- Cell state + state transition probability + dependency on previous state

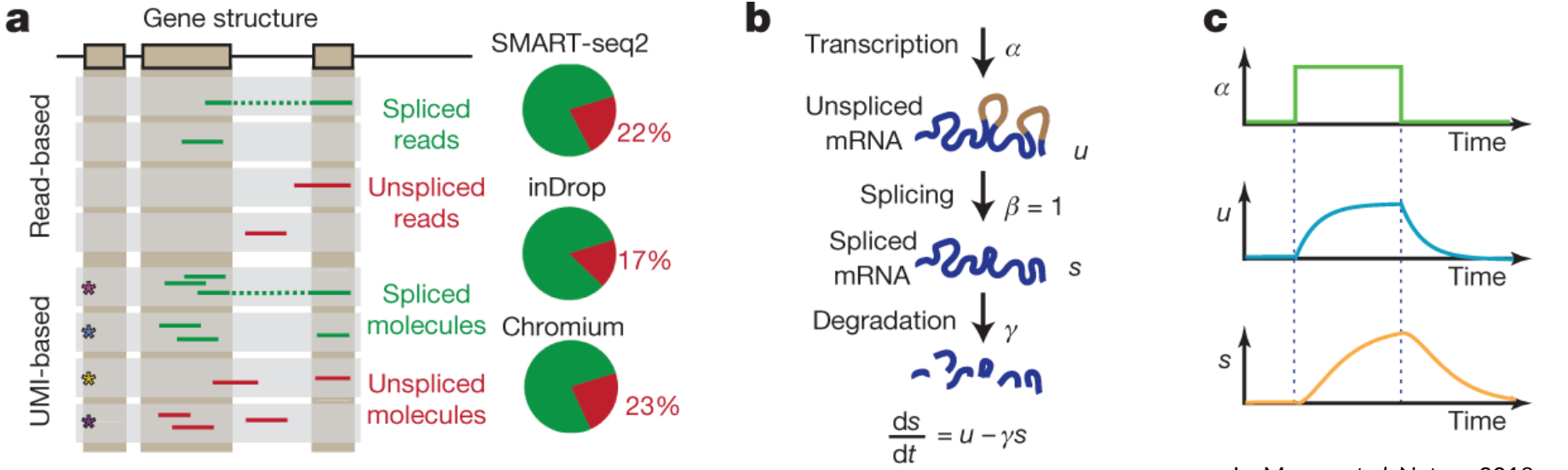
Estimating differentiation potential



Setty *et al.* Nat Biotech 37:451-460 (2019)

- Differentiation potential = probability of reaching multiple final cell types

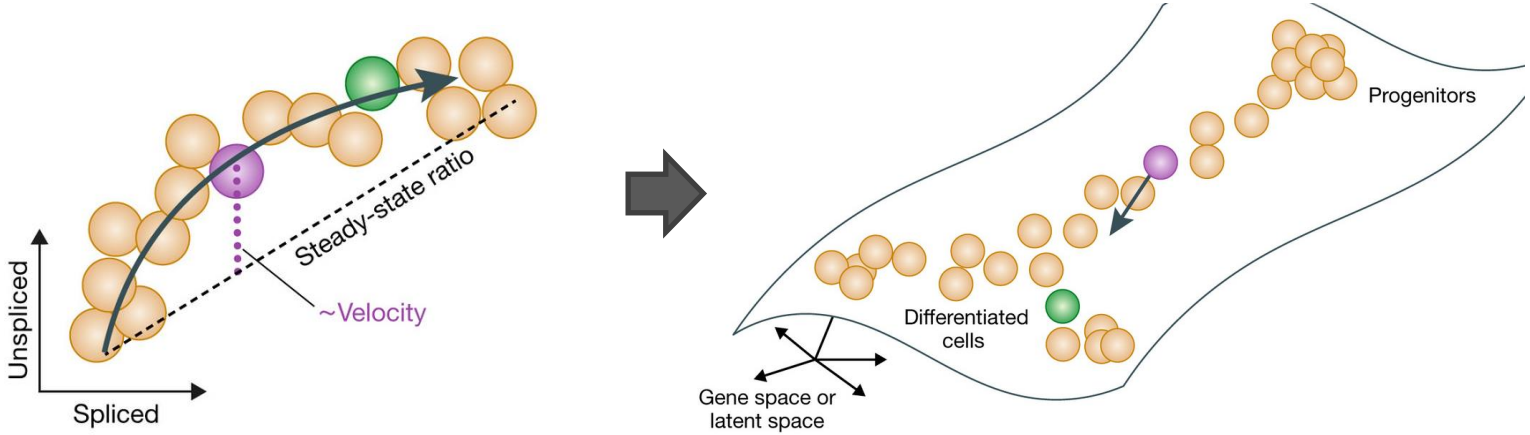
Dynamics of unspliced transcript



La Manno *et al.* Nature 2018

- When a gene is activated, level of unspliced transcripts rises first
- When a gene is repressed, level of unspliced transcripts drops first

RNA velocity model

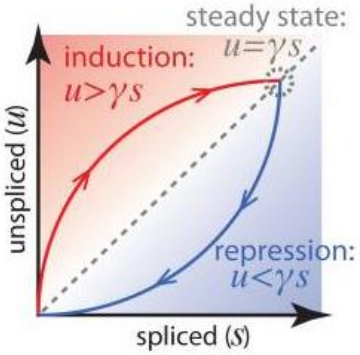


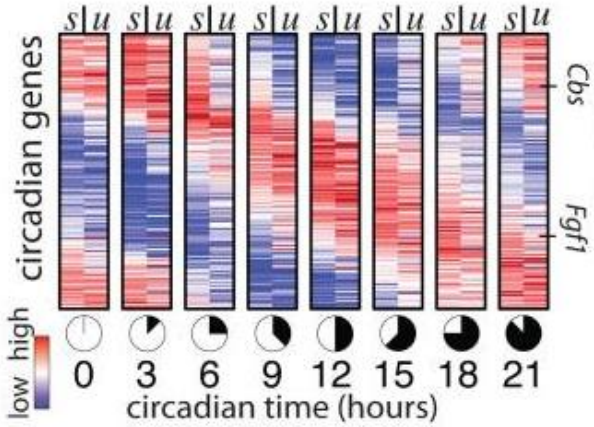
Bergen, V. *et al.* Mol Sys Biol 17:e10282 (2021)

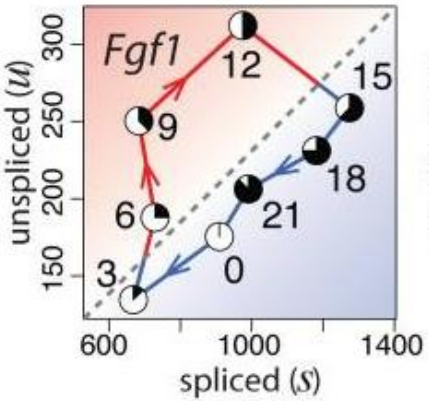
- Ratio of spliced and unspliced isoforms tells gene activation state
- Compare to nearby cells to identify direction of activation or repression

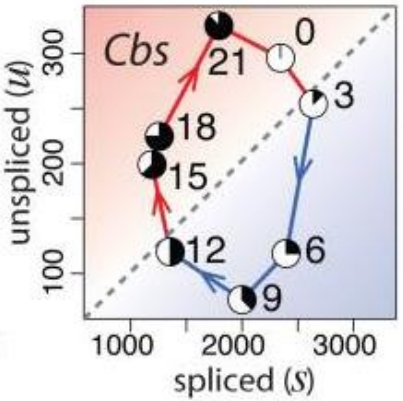
Proof of RNA velocity

Expected pattern







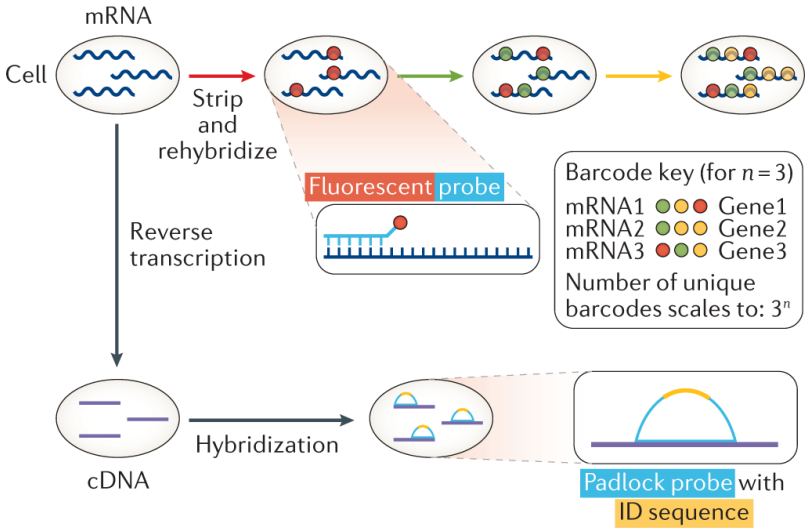


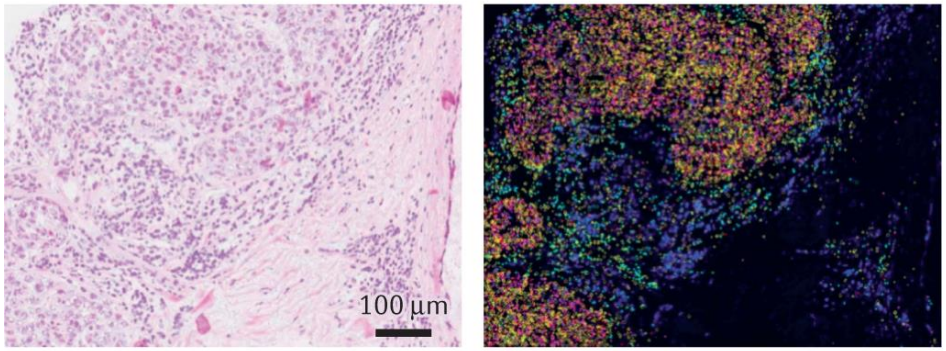
La Manno *et al.* Nature 2018

- Circadian genes are genes whose expression cycle with the time of day

Spatial transcriptomics

Extended NanoString

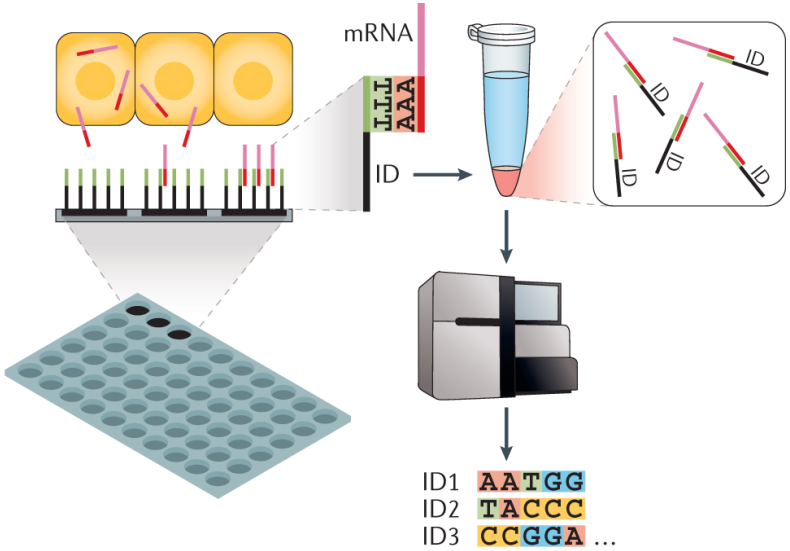


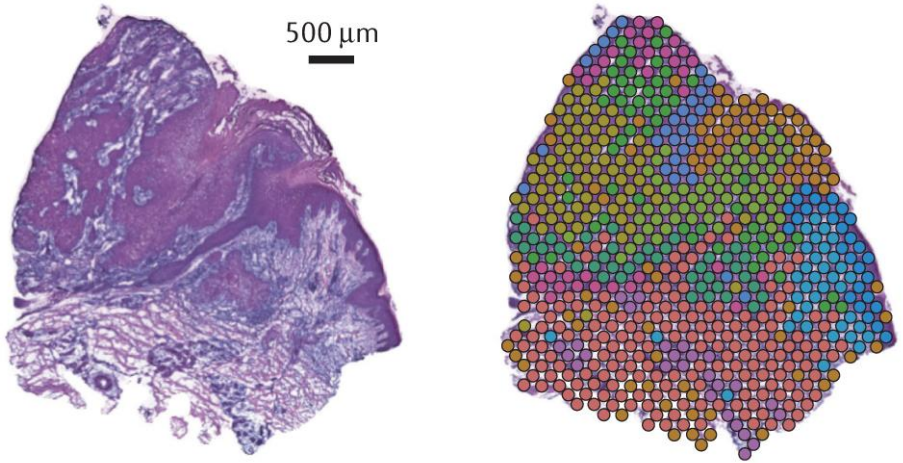


Longo et al. Nature Reviews Genetics 22:627-644 (2021)

- In-situ fluorescence labeling of selected RNA transcripts

Spatial barcoding





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

- Spatial cell isolation + barcoding

Single-nucleus sequencing

- Isolate nuclei instead of whole-cells
  - Only capture RNA expression in nucleus

- Good for cells that are difficult to isolate: adipocyte, neuron, etc.
  - Also works well with preserved tissues

Summary

- Benefits of single-cell technology
- Difference between single-cell and bulk transcriptomics
- New analysis ideas
  - Visualization of high-dimensional data (preview)
  - Cell type clustering
  - Developmental trajectory reconstruction
- Spatial transcriptomics

Any question?

- See you on October 9