3000788 Intro to Comp Molec Biol

Week 7: Single-cell transcriptomics

Fall 2024

Sira Sriswasdi, PhD

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

Single-cell transcriptomics key points

- Bulk transcriptomics cannot reveal heterogeneity

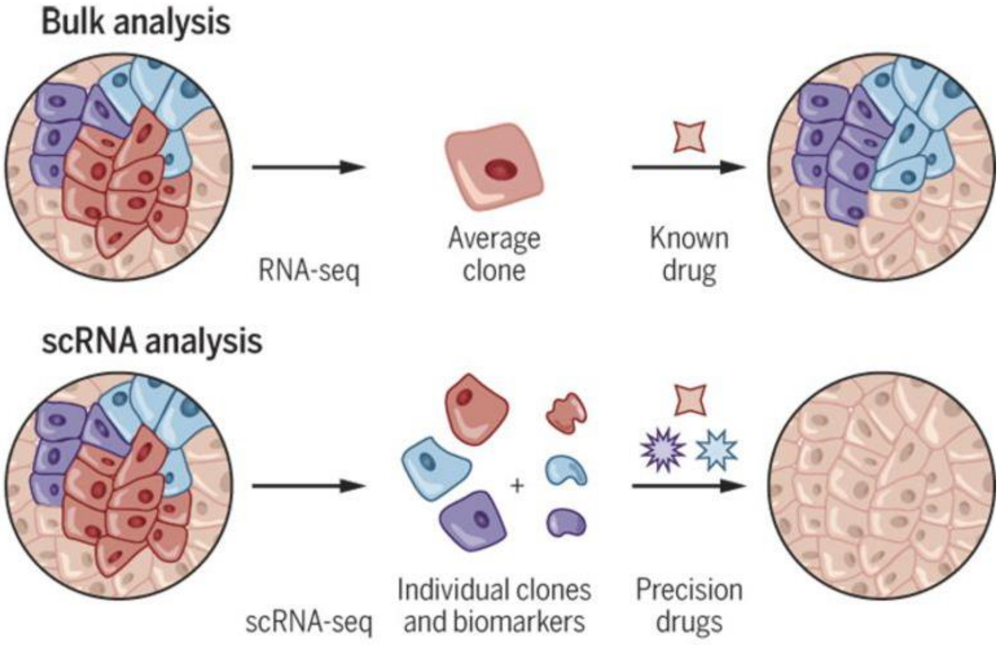
- Lower sequencing depth per cell makes data analysis difficult
- Batch effect is also very strong

- How to visualize cell-cell similarity as 3D scatter plot?

- Key analyses: cell clustering and trajectory

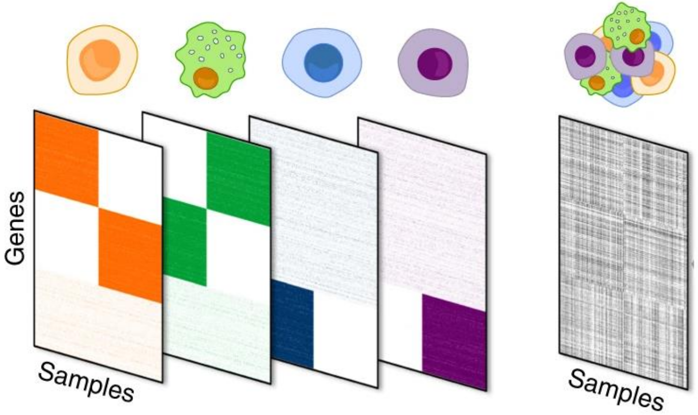
- Spatial transcriptomics

Tissue consists of multiple cell types



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

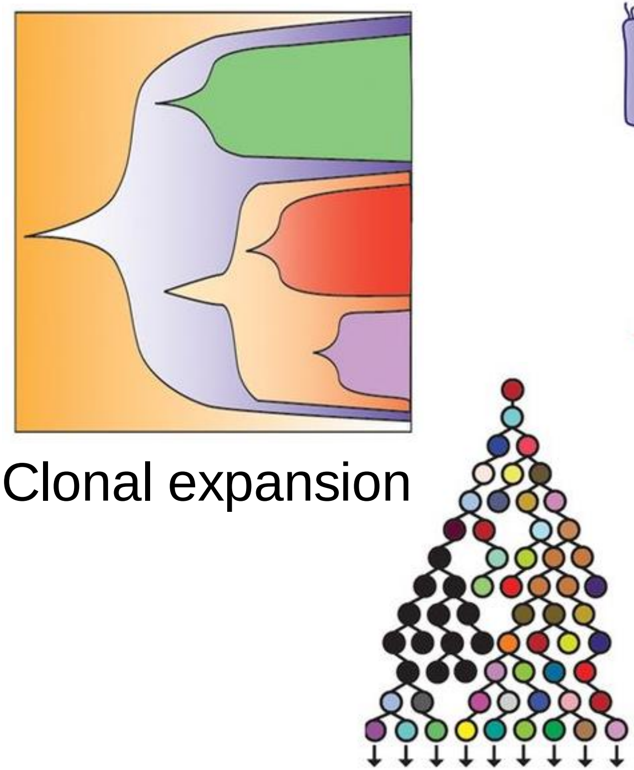
Each biological sample is a mixture of many cell types



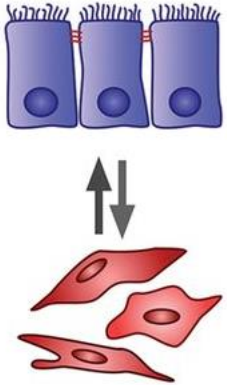
Newman *et al*. Nat Biotech 2019

Knowledge at single-cell resolution

Heterogeneity

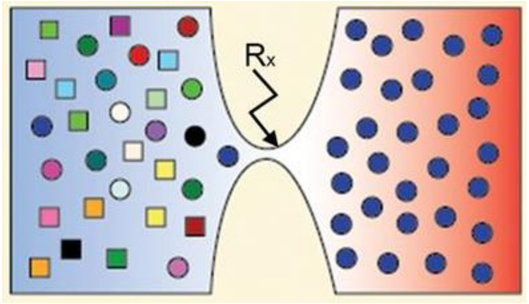


Clonal expansion

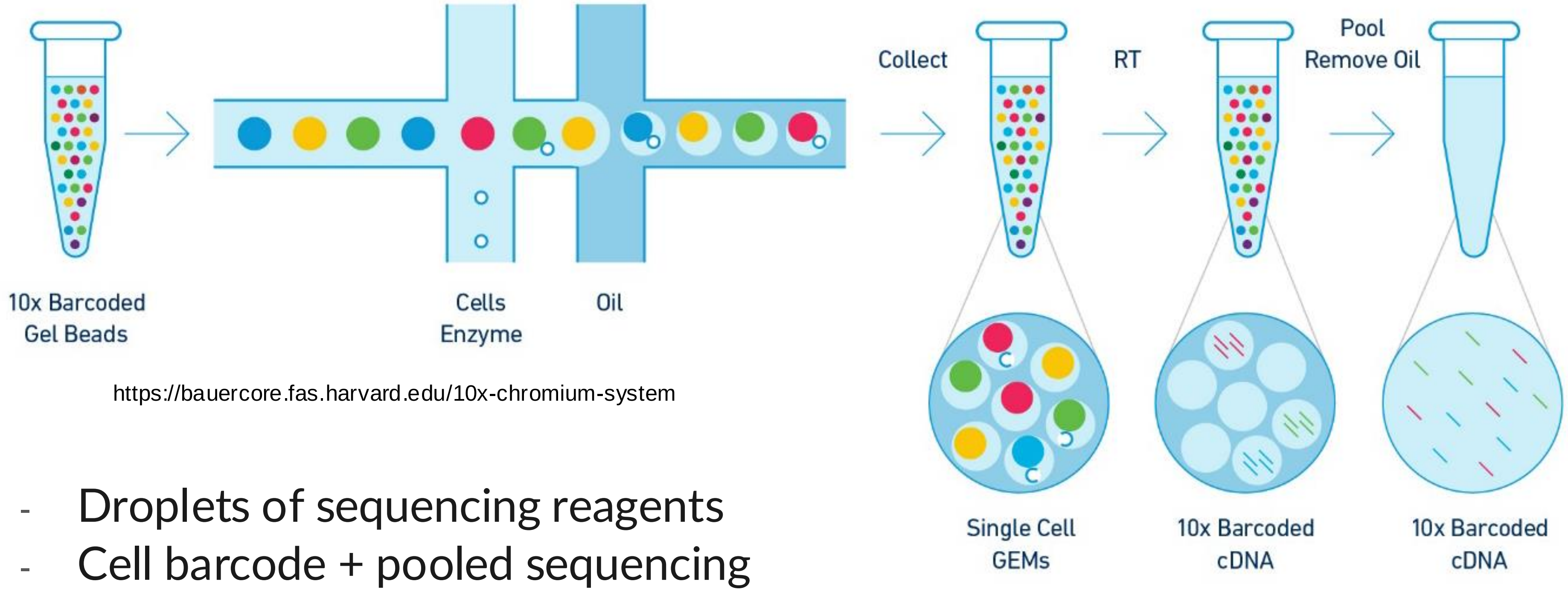


Differentiation

Treatment response



10x Genomics’ chromium technique



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

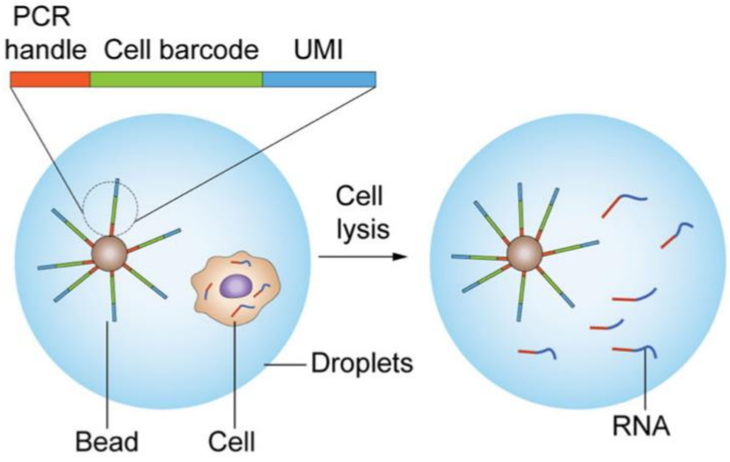
- Droplets of sequencing reagents
- Cell barcode + pooled sequencing

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

UMI and cell barcode

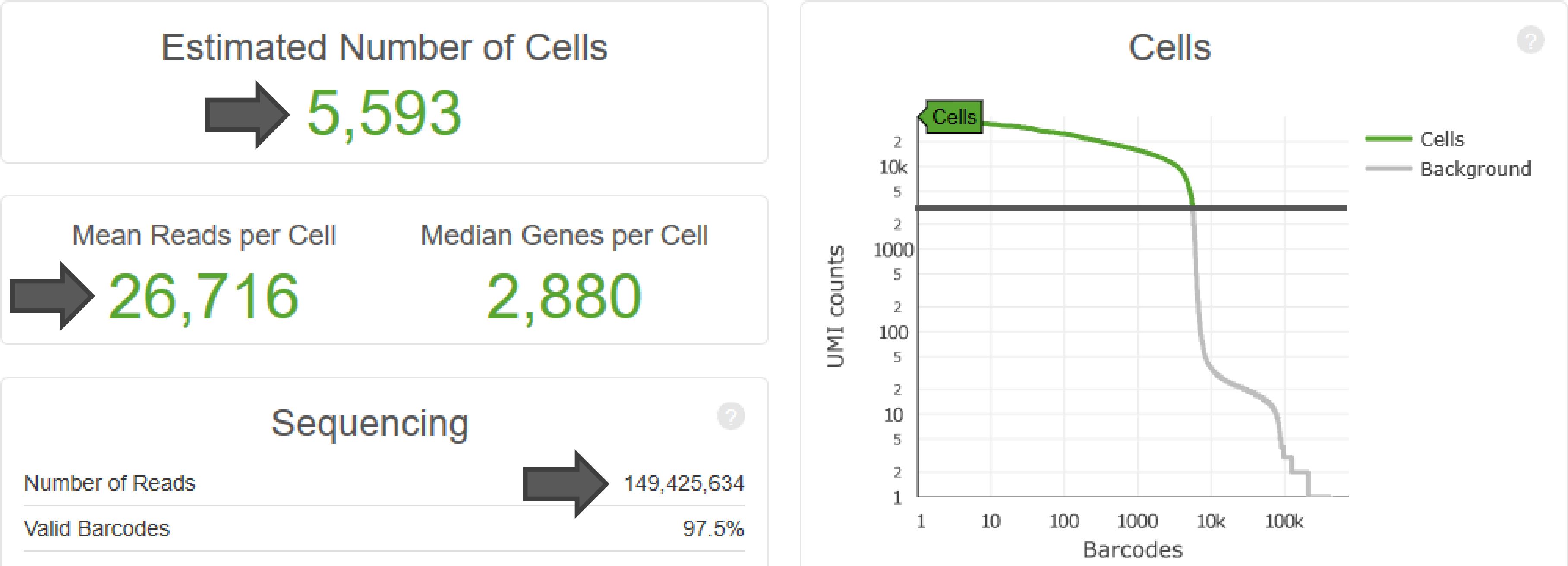


Hwang *et al.* Exp & Mol Med 50:96 (2018)

- 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%

Cells

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

Data processing and QC

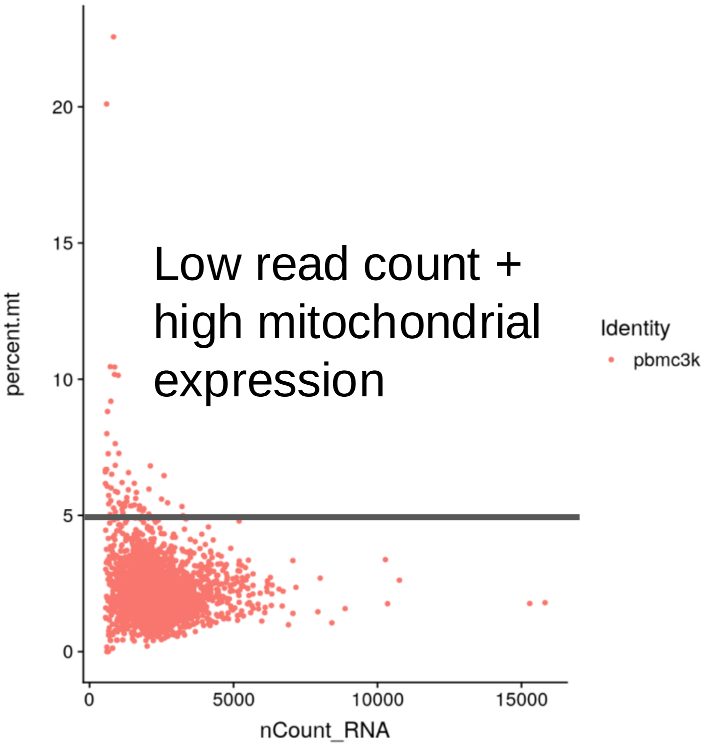
Key steps in single-cell data processing

- Quality filter
  - Low read count & gene count = non-cells
  - Very high read count & gene count = multi-cells
  - High mitochondrial expression = dead cells

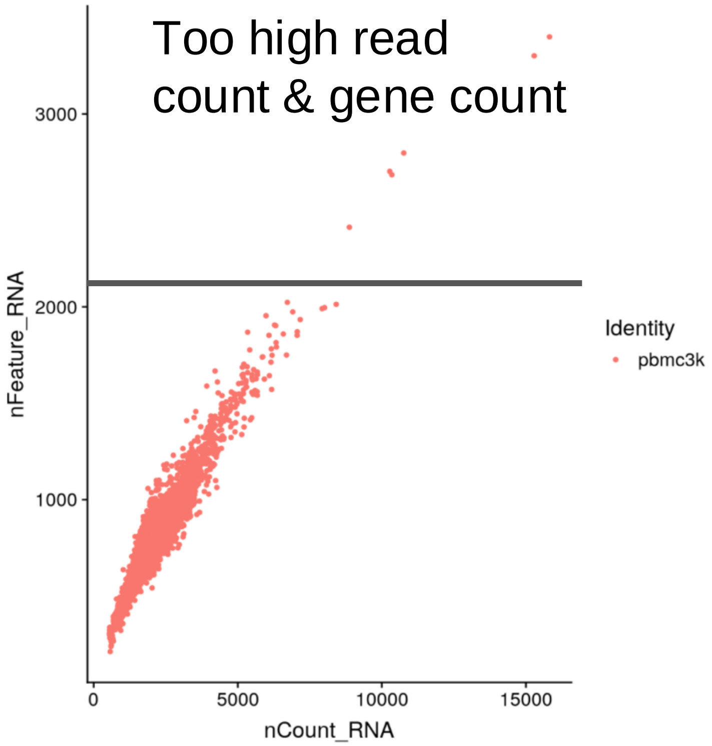
- Within-sample normalization
  - Dealing with missing expression values

- Multi-sample integration
  - Single-cell data have strong batch effects

Basic quality filters

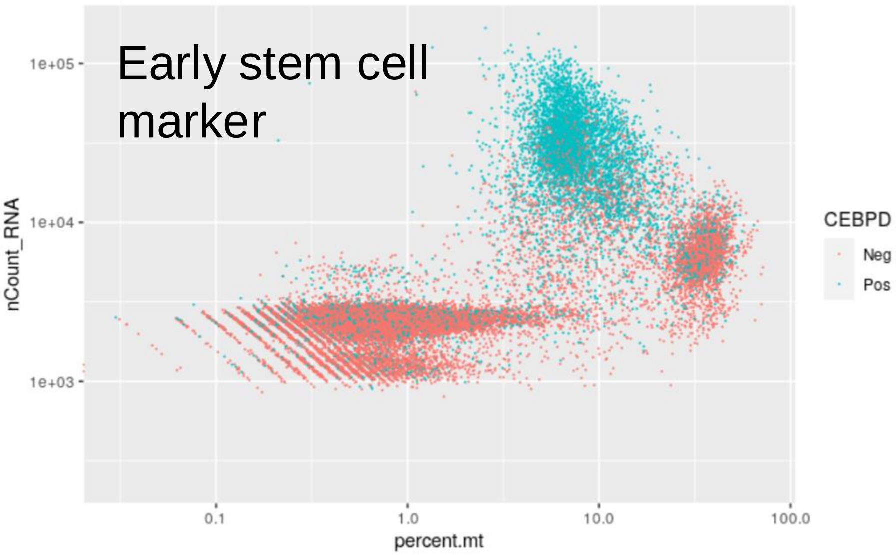


Low read count + high mitochondrial expression

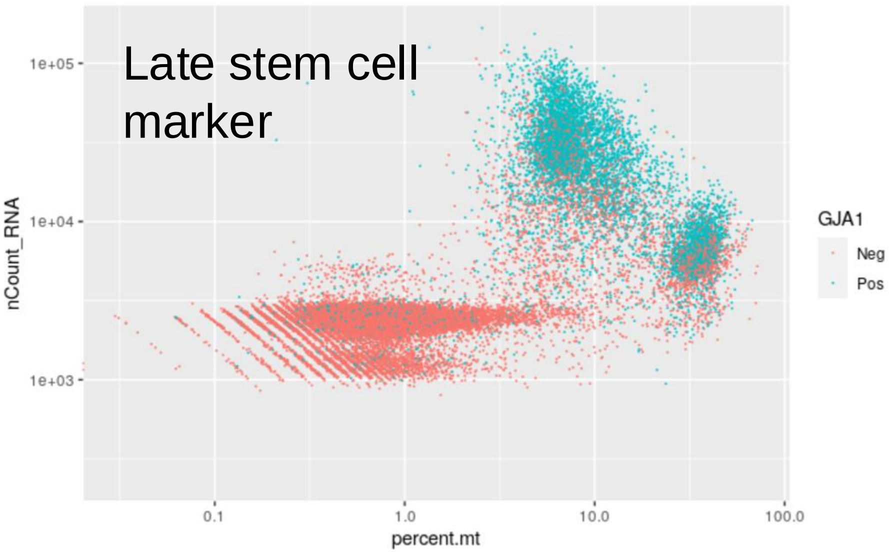


Too high read count & gene count

Stem cell markers in high-mitochondrial cells

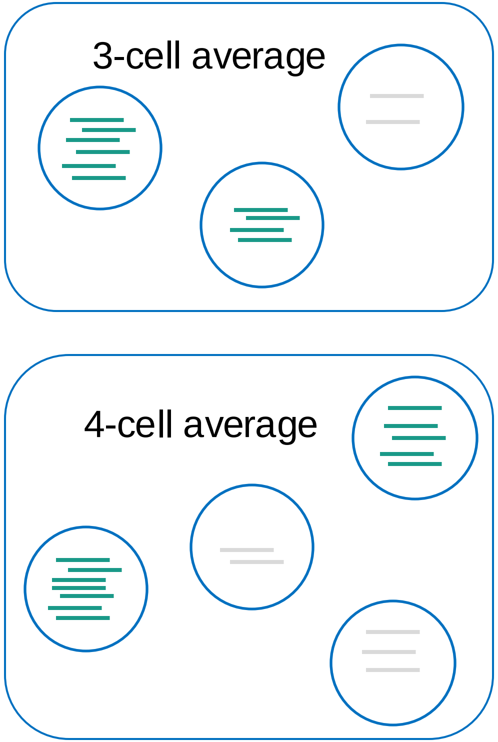


Early stem cell marker



Late stem cell marker

Normalization with pooling



Low expression

Gene A
Gene B
Gene C

Medium expression

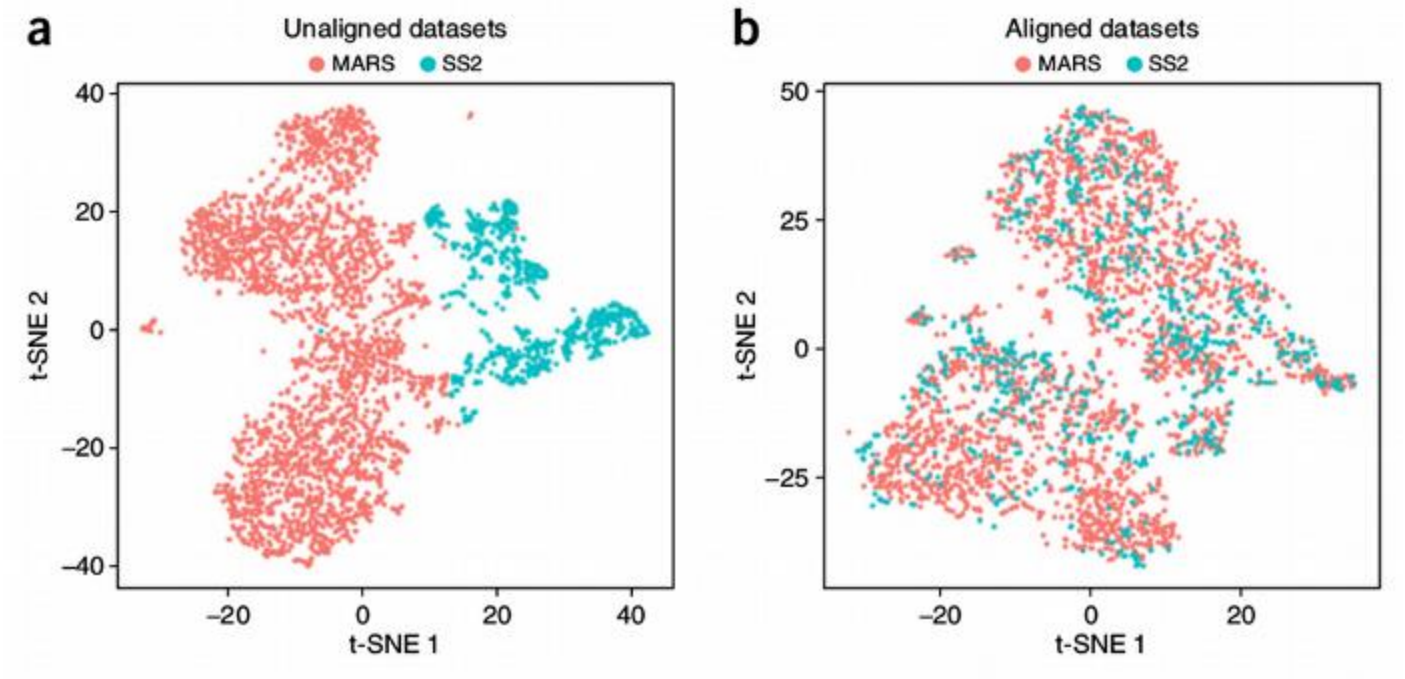
Gene D
Gene E
Gene F
Gene G

High expression

Gene H
Gene I

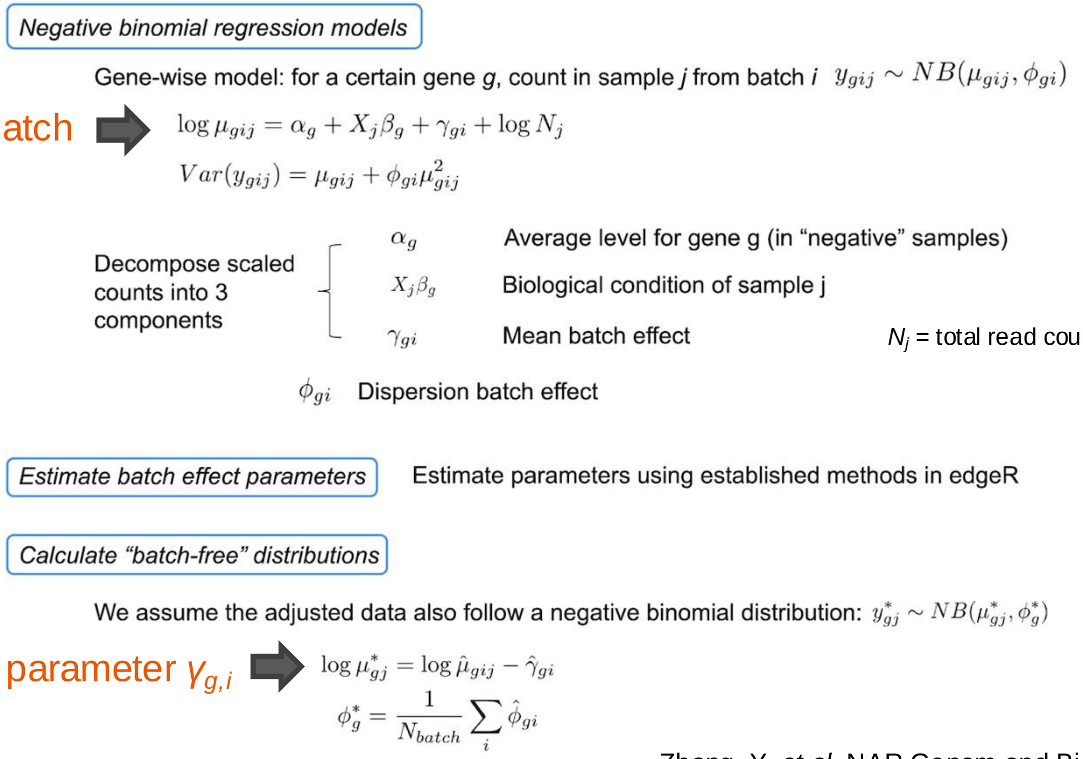
Data integration

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: $y_{gj}^* \sim NB(\mu_{gj}^*, \phi_g^*)$

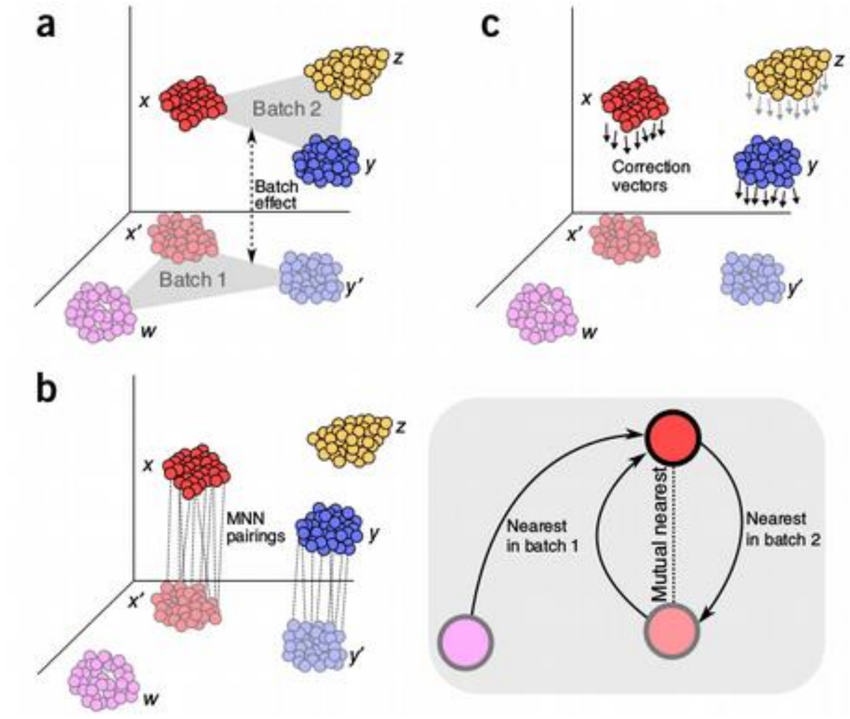
Subtract batch parameter $\gamma_{g,i}$

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

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

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

Mutual nearest neighbor (MNN)



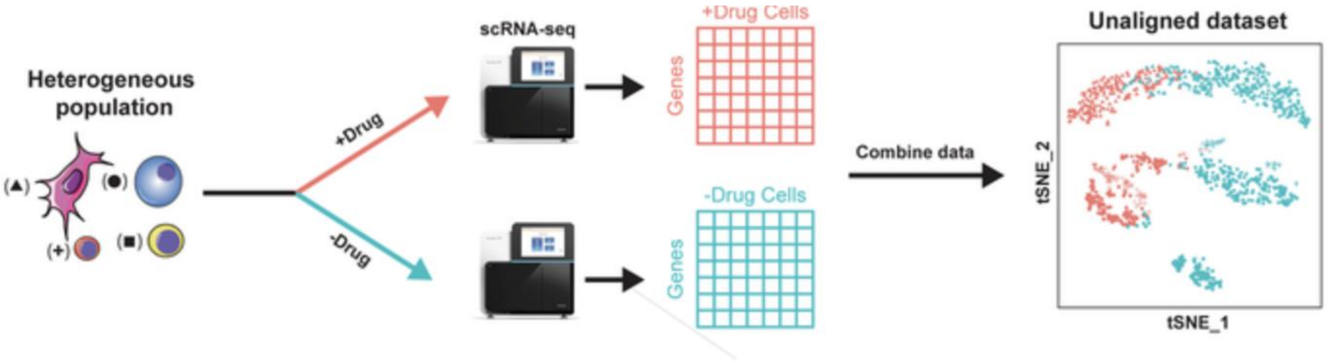
Haghverdi et al., Nat. Biot. 36, 2018

- Map clusters of cells together rather than individual cells

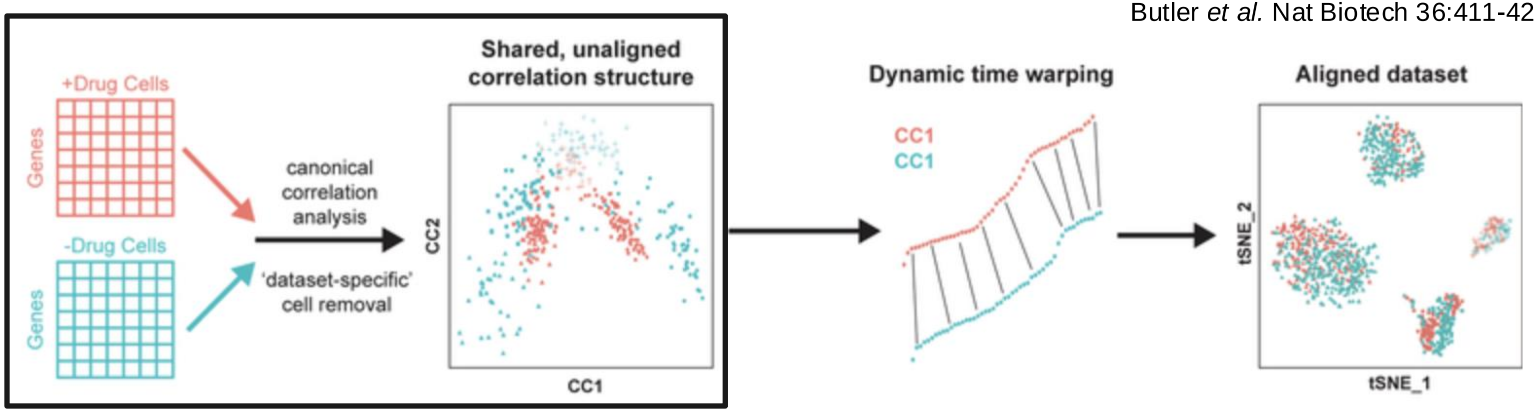
- Similar to reciprocal best hits from BLAST for identifying orthologs

- Apply the average mapping vector to unique cell types

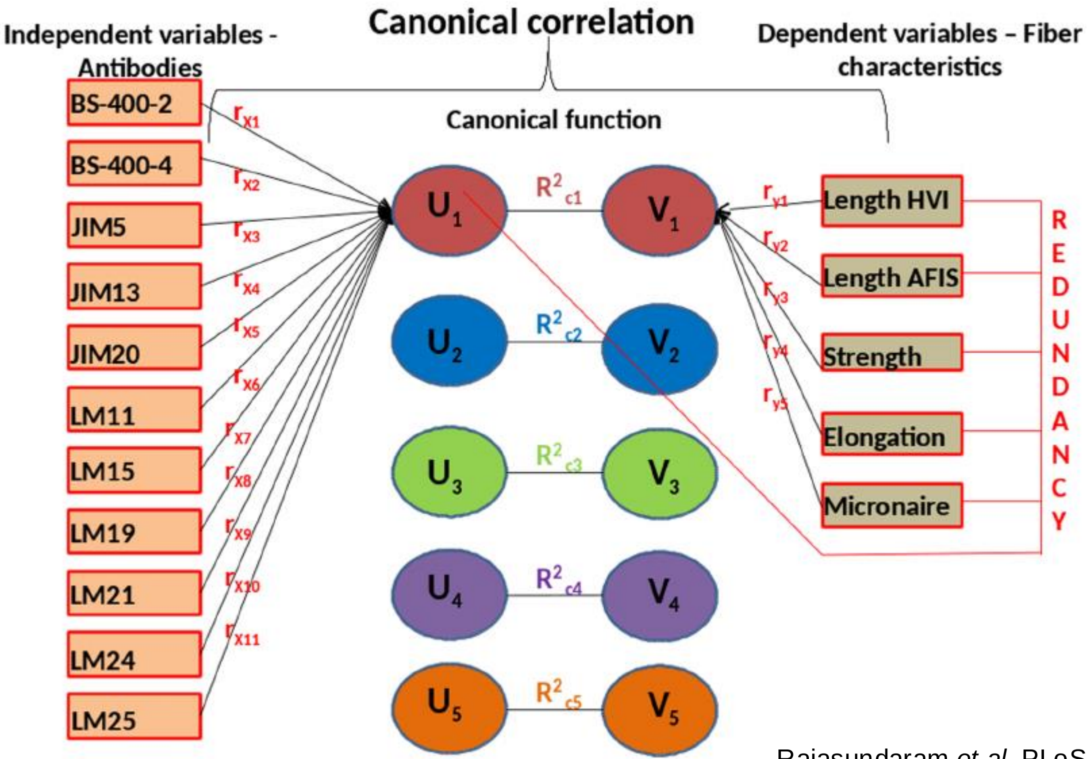
Integration via canonical correlation



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


+Drug Cells
Genes
-Drug Cells
Genes
canonical correlation analysis
'dataset-specific' cell removal
Shared, unaligned correlation structure
CC2
CC1
Dynamic time warping
CC1
CC1
Aligned dataset
tSNE_2
tSNE_1

Canonical correlation analysis (CCA)

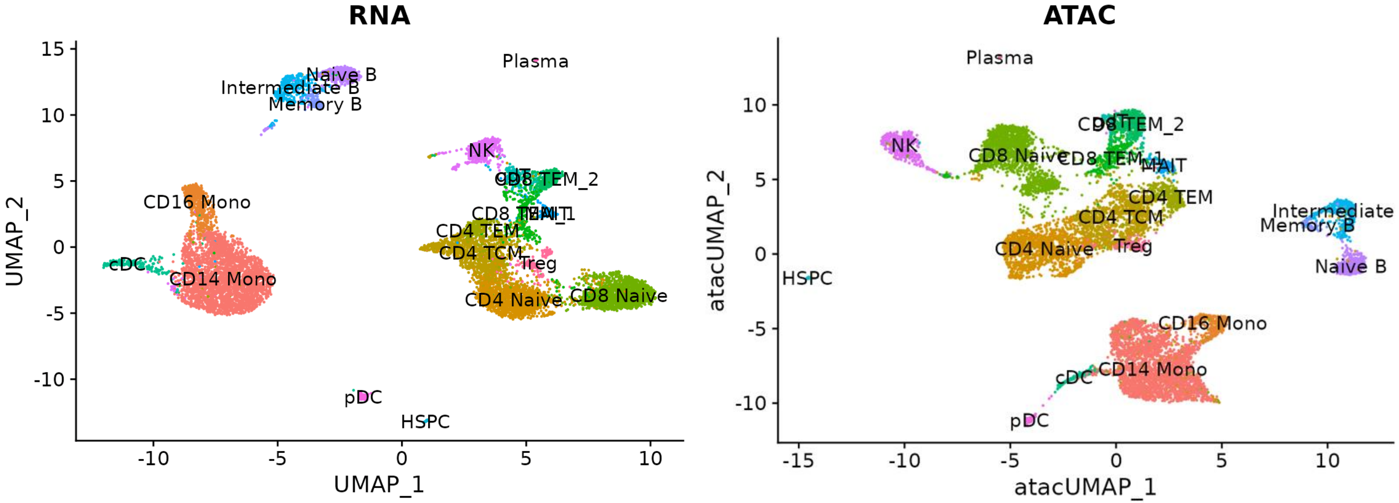


- Same samples with two different systems of observations

- Identify correlation structure observation systems (features)

Rajasundaram *et al.* PLoS ONE (2014)

Integrating RNA-seq with ATAC-seq

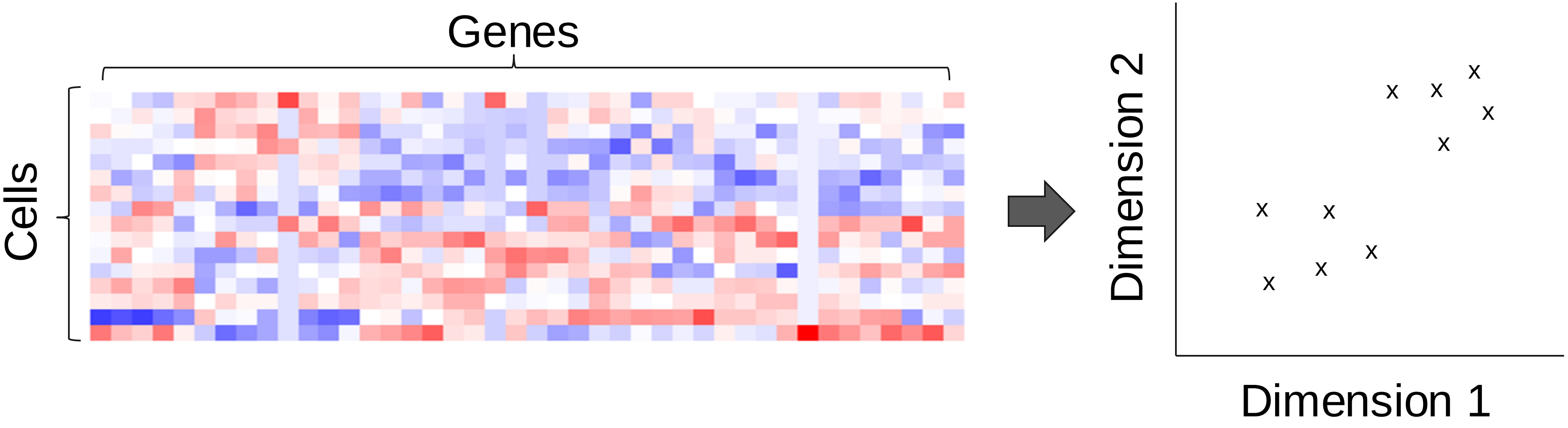


- ATAC-seq = open chromatin ~ gene expression level
- Transfer cell type label

Visualizing single-cell data

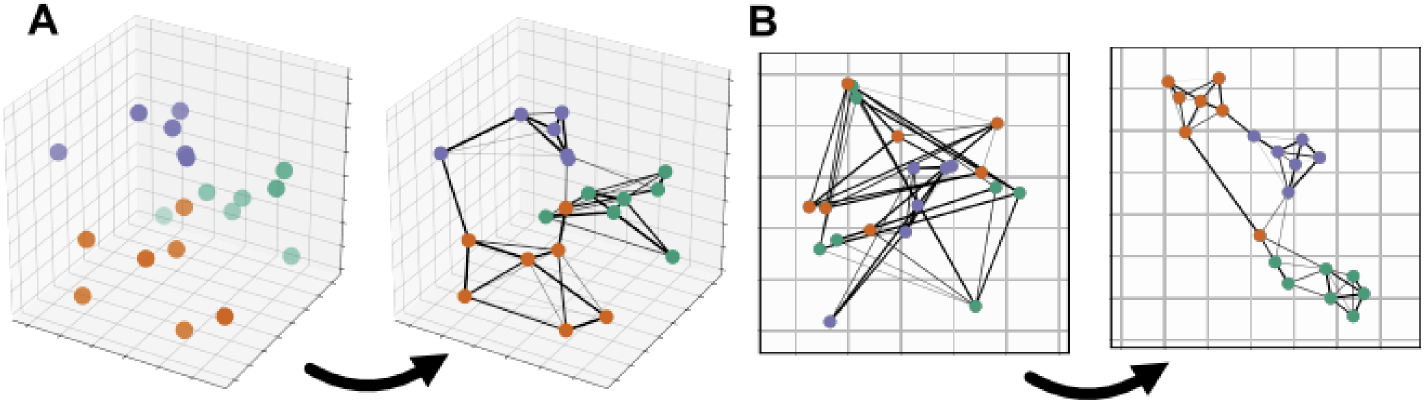
with dimensionality reduction techniques

Dimensionality reduction



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

Dimensionality reduction algorithm sketch



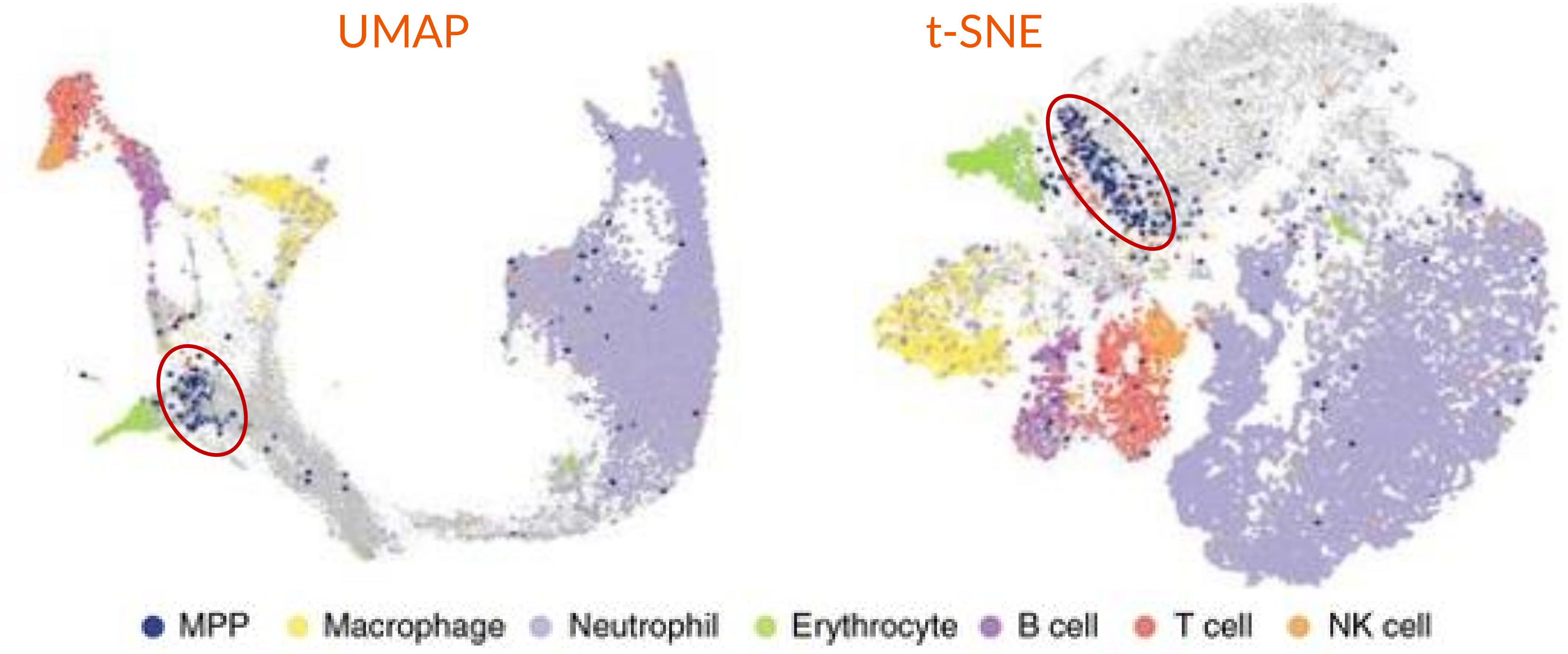
Step 1: Compute a graphical representation of the dataset

Step 2 (non-parametric): Learn an embedding that preserves the structure of the graph

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



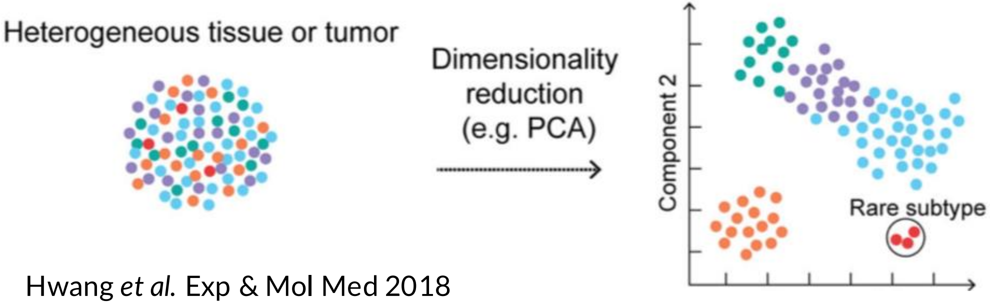
UMAP

t-SNE

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

Clustering and trajectory analyses

Cell clustering and trajectory reconstryction

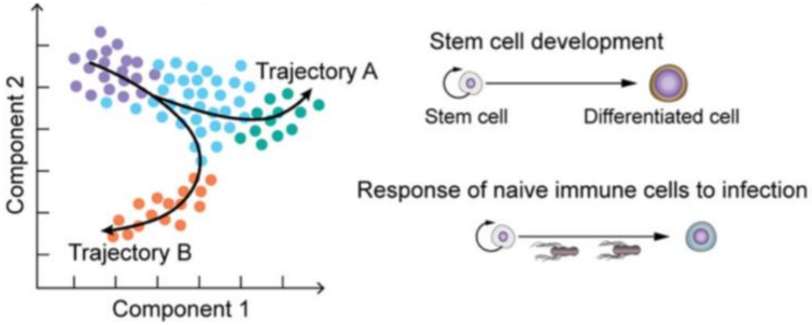


Heterogeneous tissue or tumor
Dimensionality reduction (e.g. PCA)
Component 2
Rare subtype

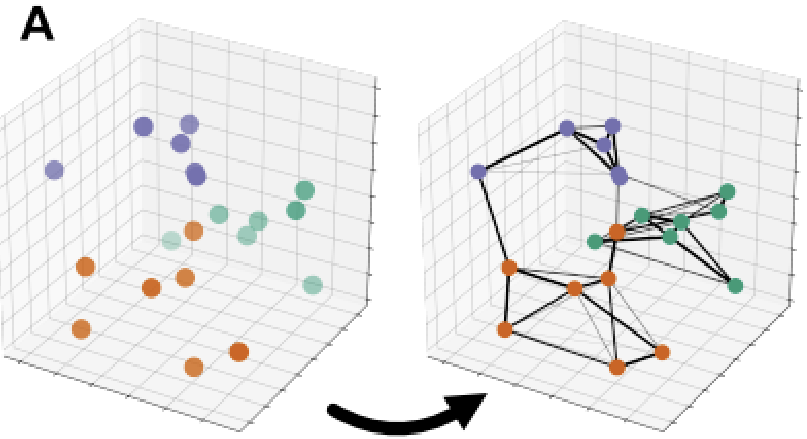
Hwang *et al.* Exp & Mol Med 2018

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

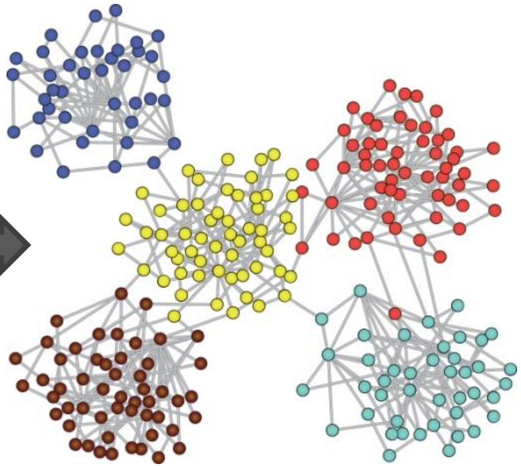
Trajectory modeling with random walk, diffusion, or Markov chain reconstruct the paths of cell development



Algorithm sketch for cell type clustering



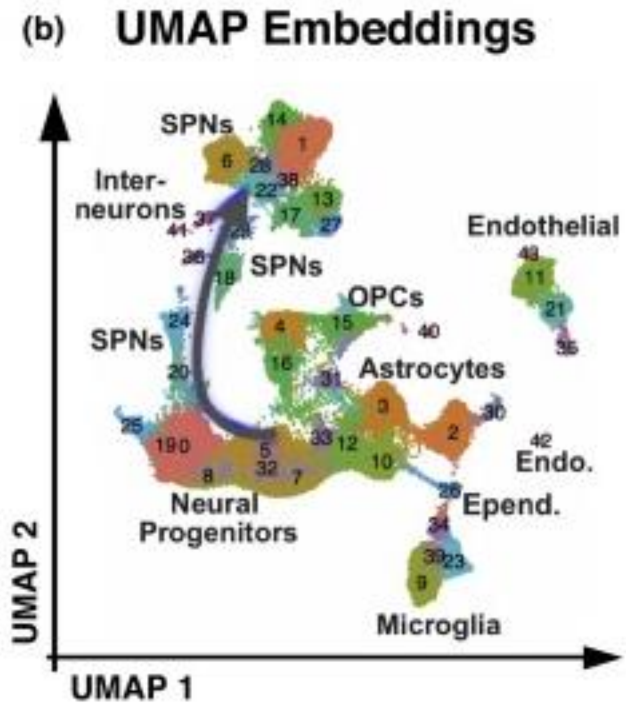
Sainburg, T. *et al.*, Neural Comput 33(11):2881-2907 (2021)



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

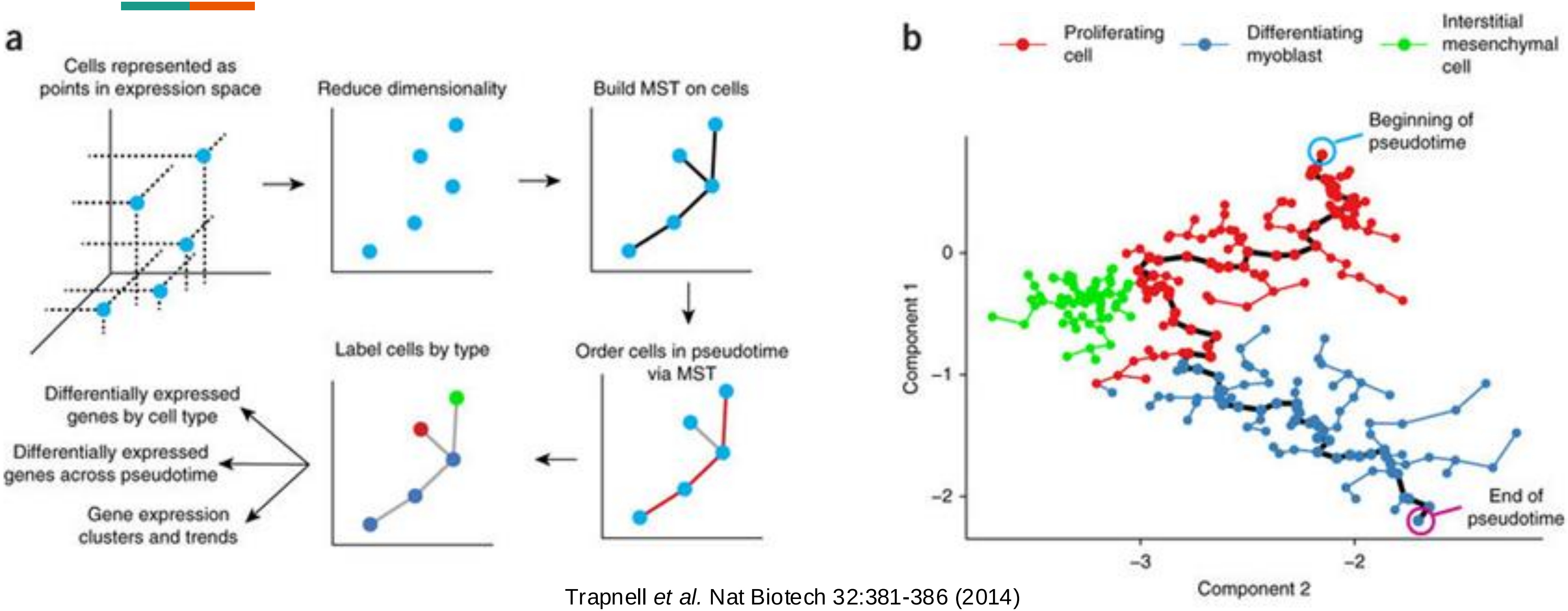
- Connect cells with similar gene expression profile
- Split network into modules with dense edges

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

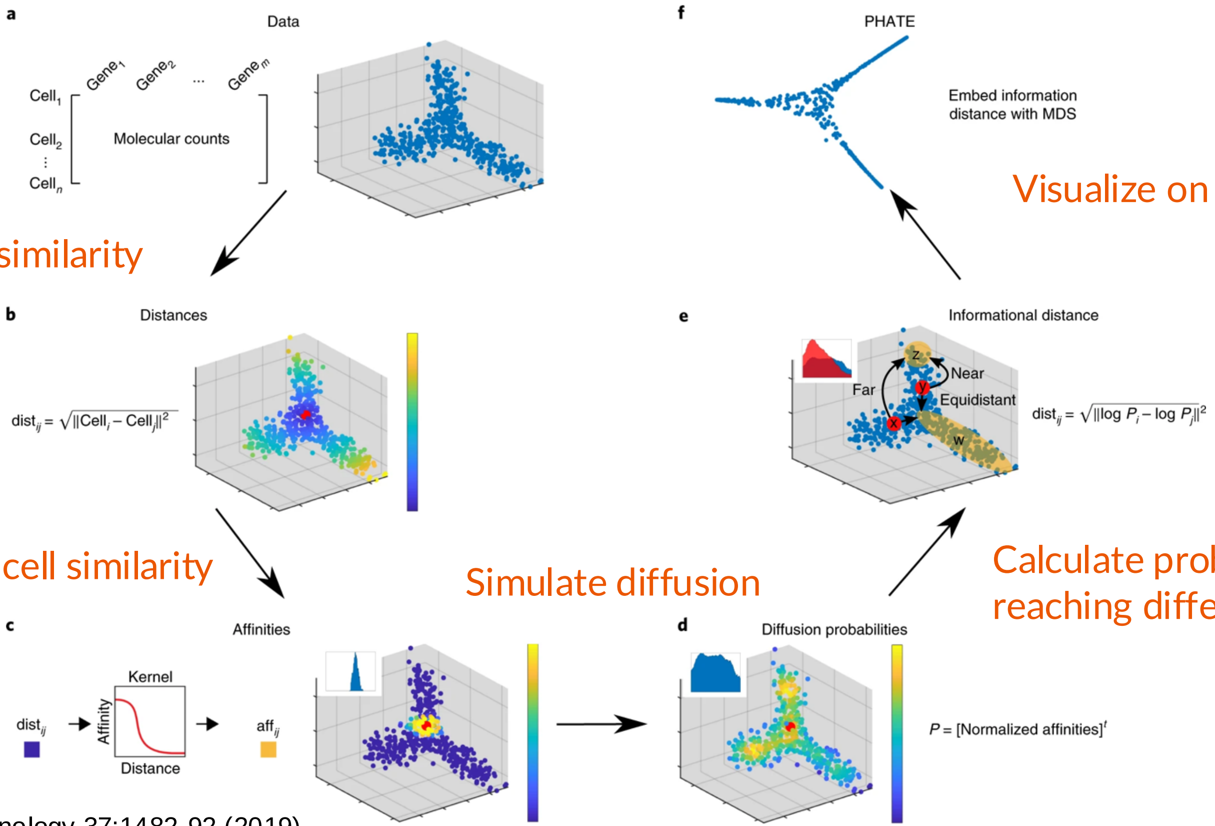


a
Cells represented as points in expression space
Reduce dimensionality
Build MST on cells
Order cells in pseudotime via MST
Label cells by type
Differentially expressed genes by cell type
Differentially expressed genes across pseudotime
Gene expression clusters and trends
b
Proliferating cell
Differentiating myoblast
Interstitial mesenchymal cell
Beginning of pseudotime
End of pseudotime
Component 1
Component 2
0
−1
−2
−3
−2

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

- Trajectory along the most similar cells

Diffusion approach for cell-cell transitions



Calculate cell-cell similarity

Visualize on 2D/3D

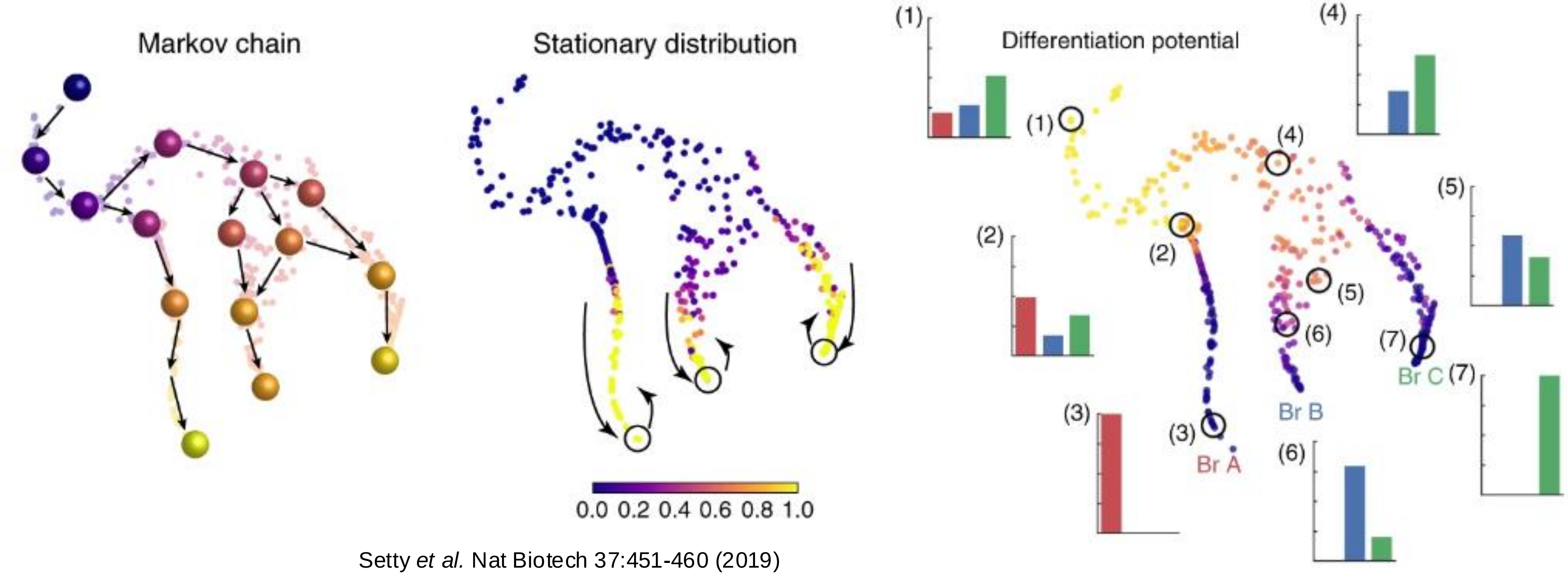
Normalize cell-cell similarity

Simulate diffusion

Calculate probability of reaching different cells

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

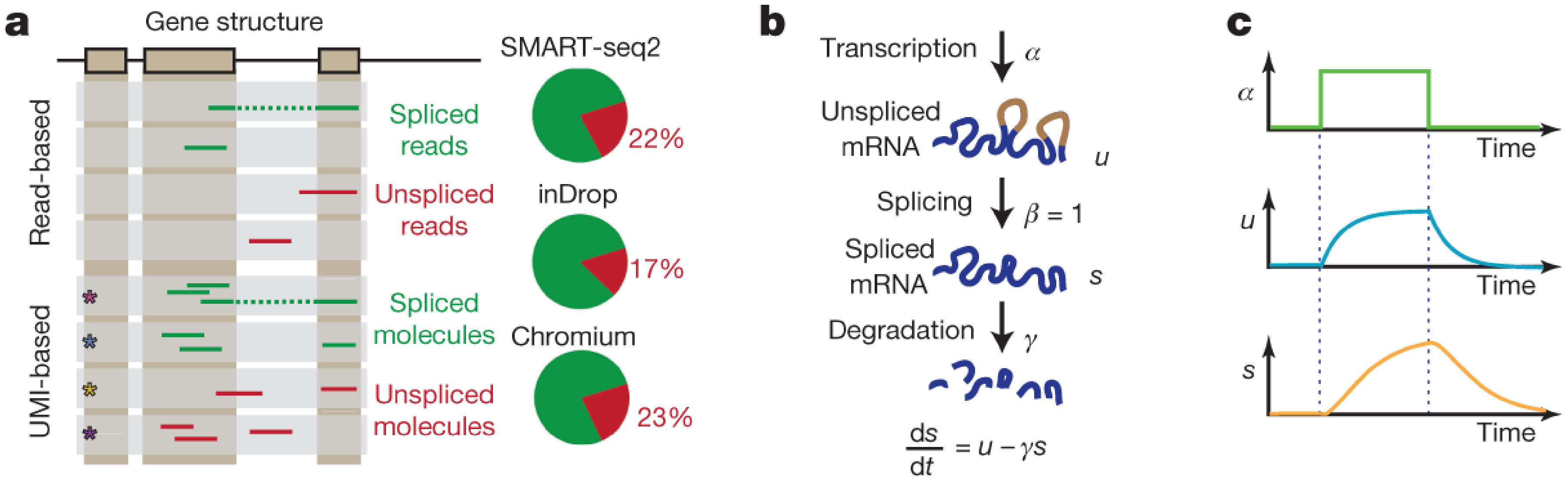
Estimating differentiation potential



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

- Differentiation potential = probability of reaching multiple final cell types

Dynamics of unspliced transcript

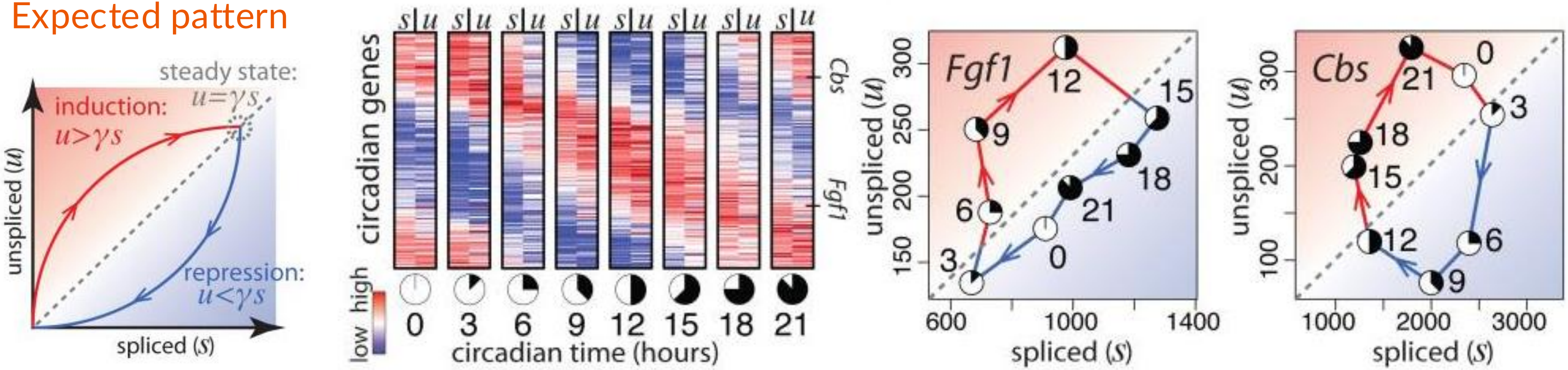


La Manno *et al.* Nature 2018

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

Proof of RNA velocity

Expected pattern

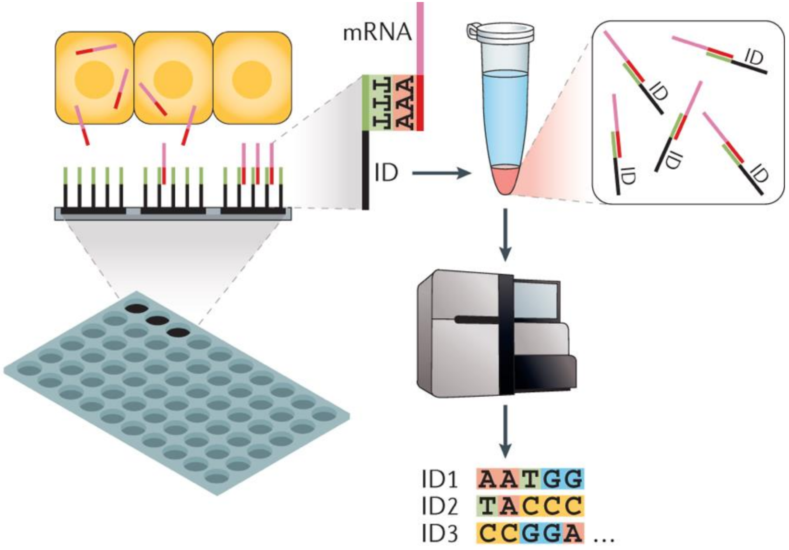


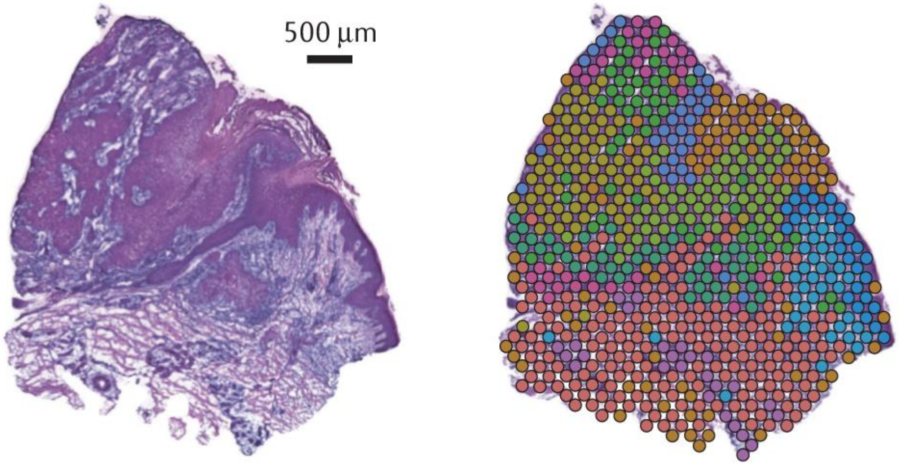
La Manno *et al.* Nature 2018

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

Spatial transcriptomics

Spatial barcoding

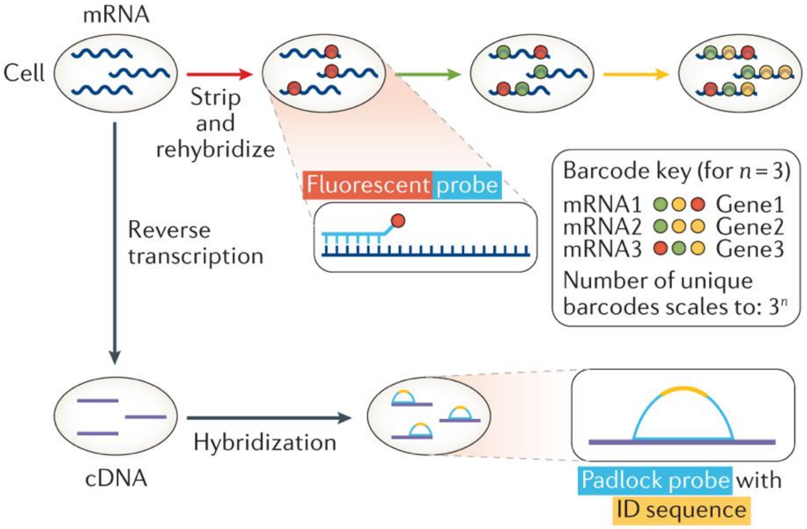


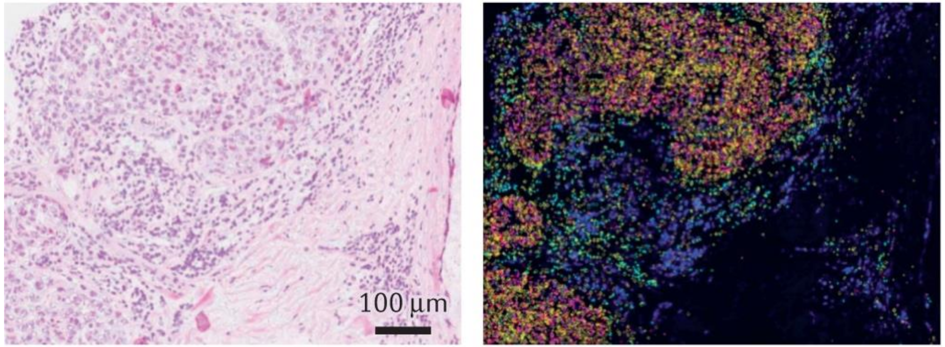


Longo et al. Nature Reviews Genetics 22:627-644 (2021)

- Spatial cell isolation + barcoding

Extended NanoString





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

- In-situ fluorescence labeling of selected RNA transcripts

Any question?