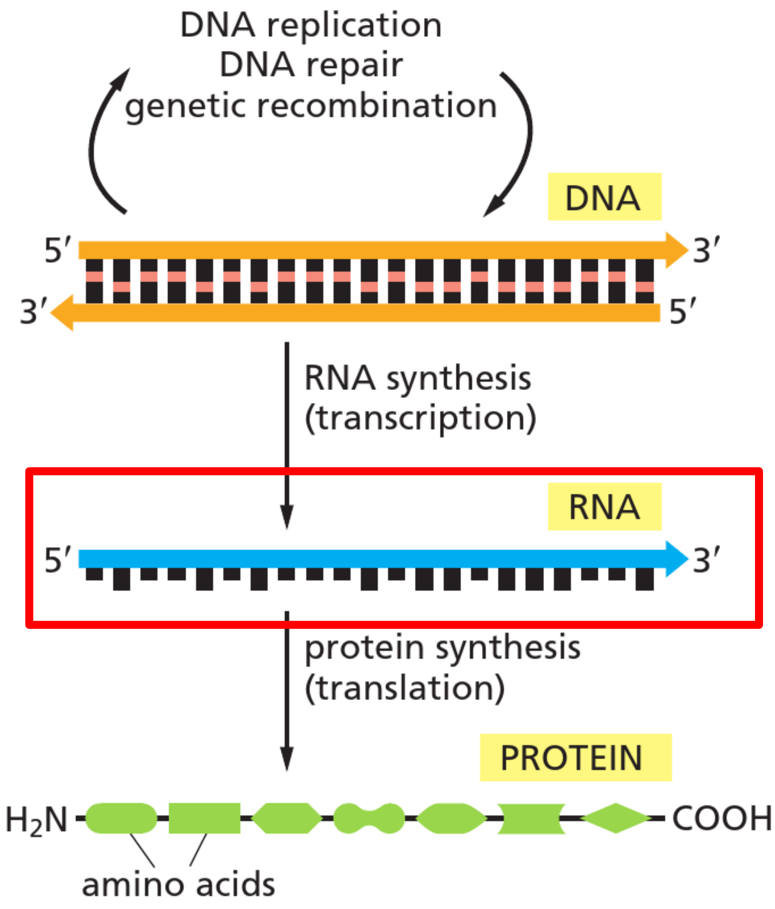
06/23/2020

Differential geometry and data science for single-cell biology

Ilia Kats and Luca Marconato
Stegle lab

Measuring gene expression

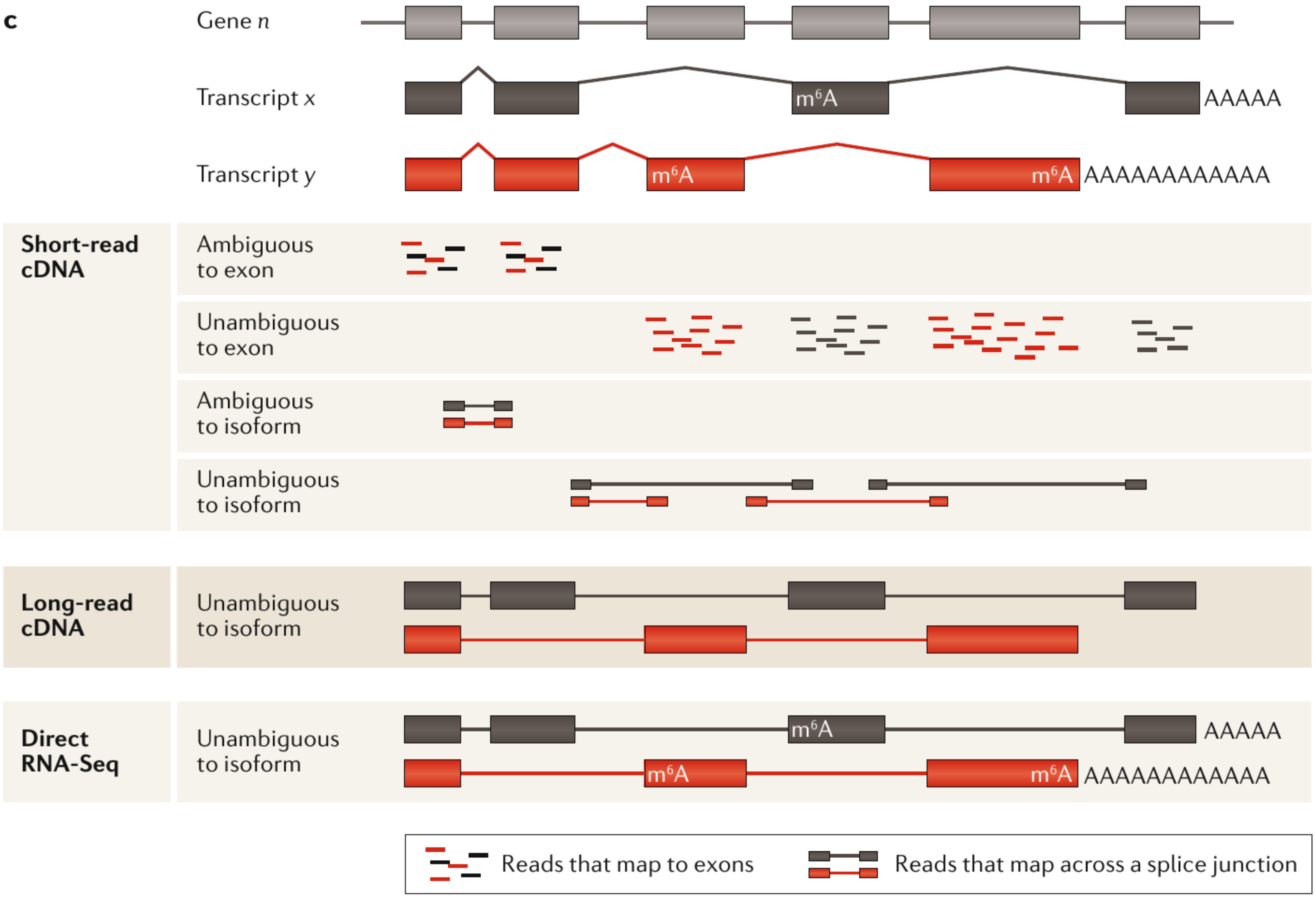


Gene expression: amount of gene product in the cell

Quantification by sequencing

Bruce Alberts, Molecular Biology of the Cell (6th edition, 2015)

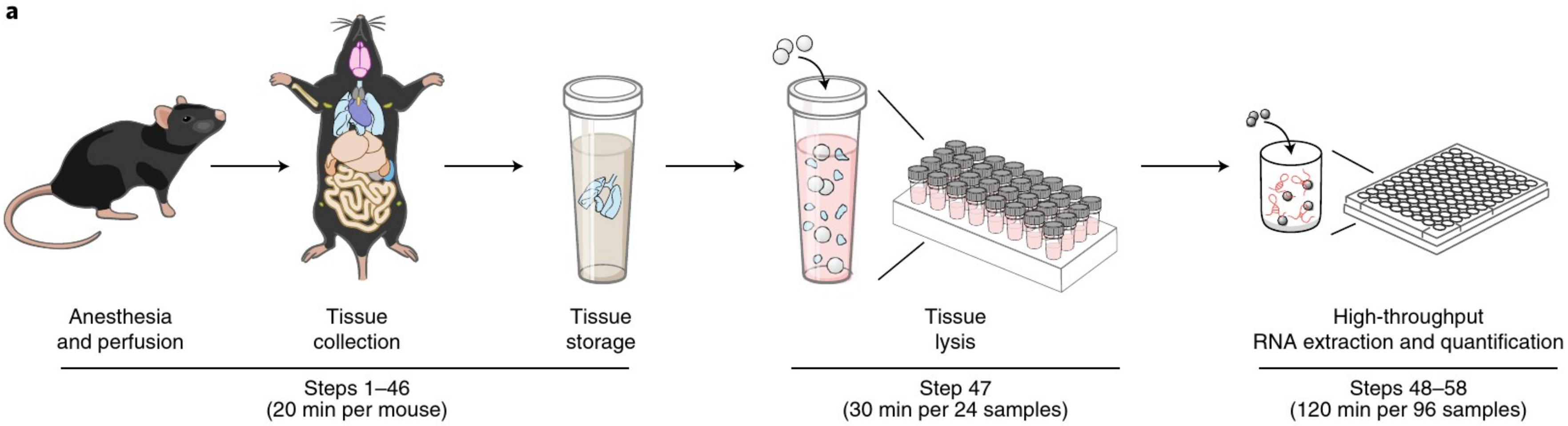
Measuring gene expression



Stark et al., 2019

RNAseq: bulk vs. single-cell

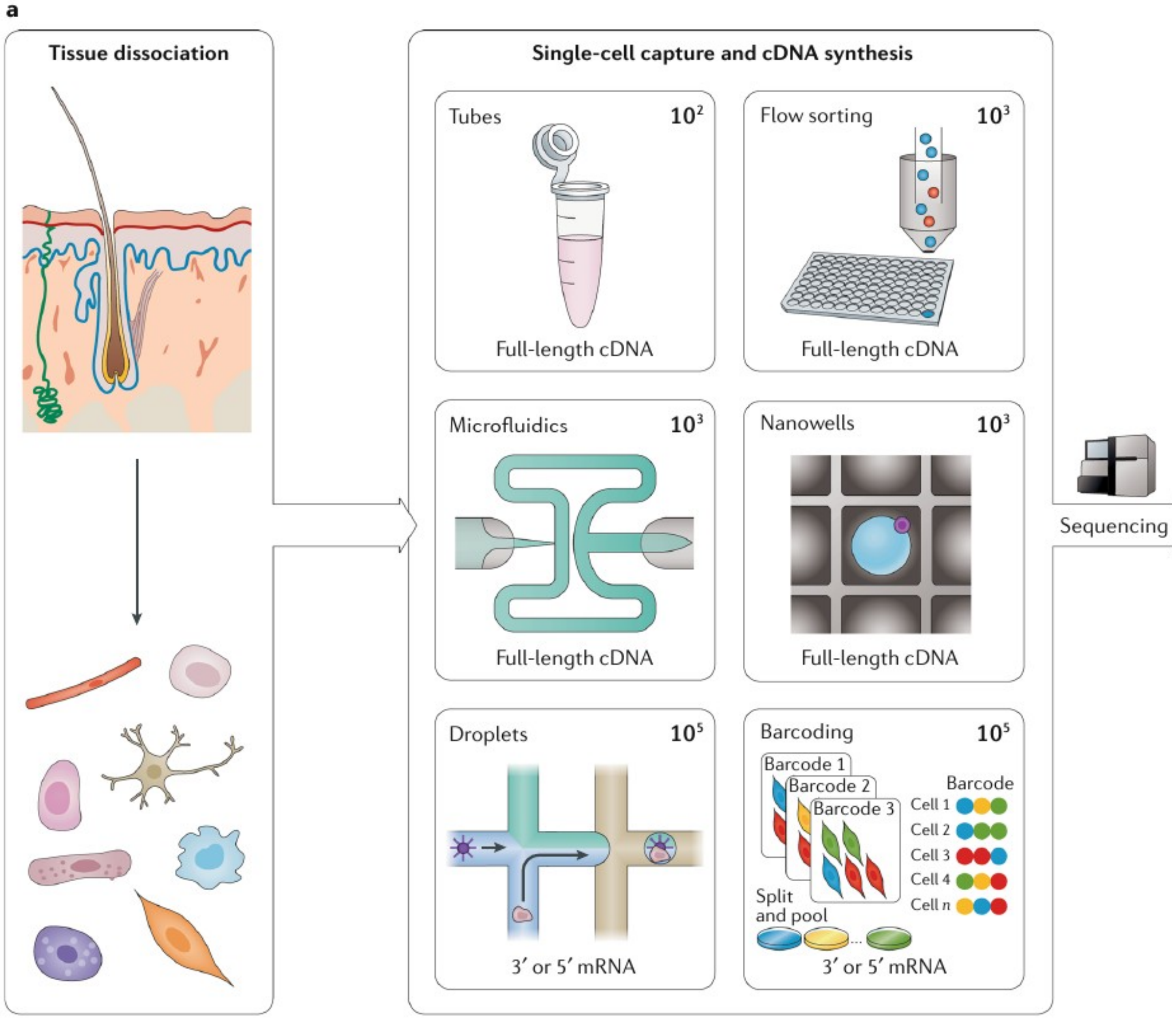
- Bulk RNAseq
  - Works with homogenized tissues/tissue culture samples
  - Measures average expression in the sample
  - High sensitivity, low throughput
  - Mostly useful for differential gene expression analysis



Pandey et al., 2020

RNAseq: bulk vs. single-cell

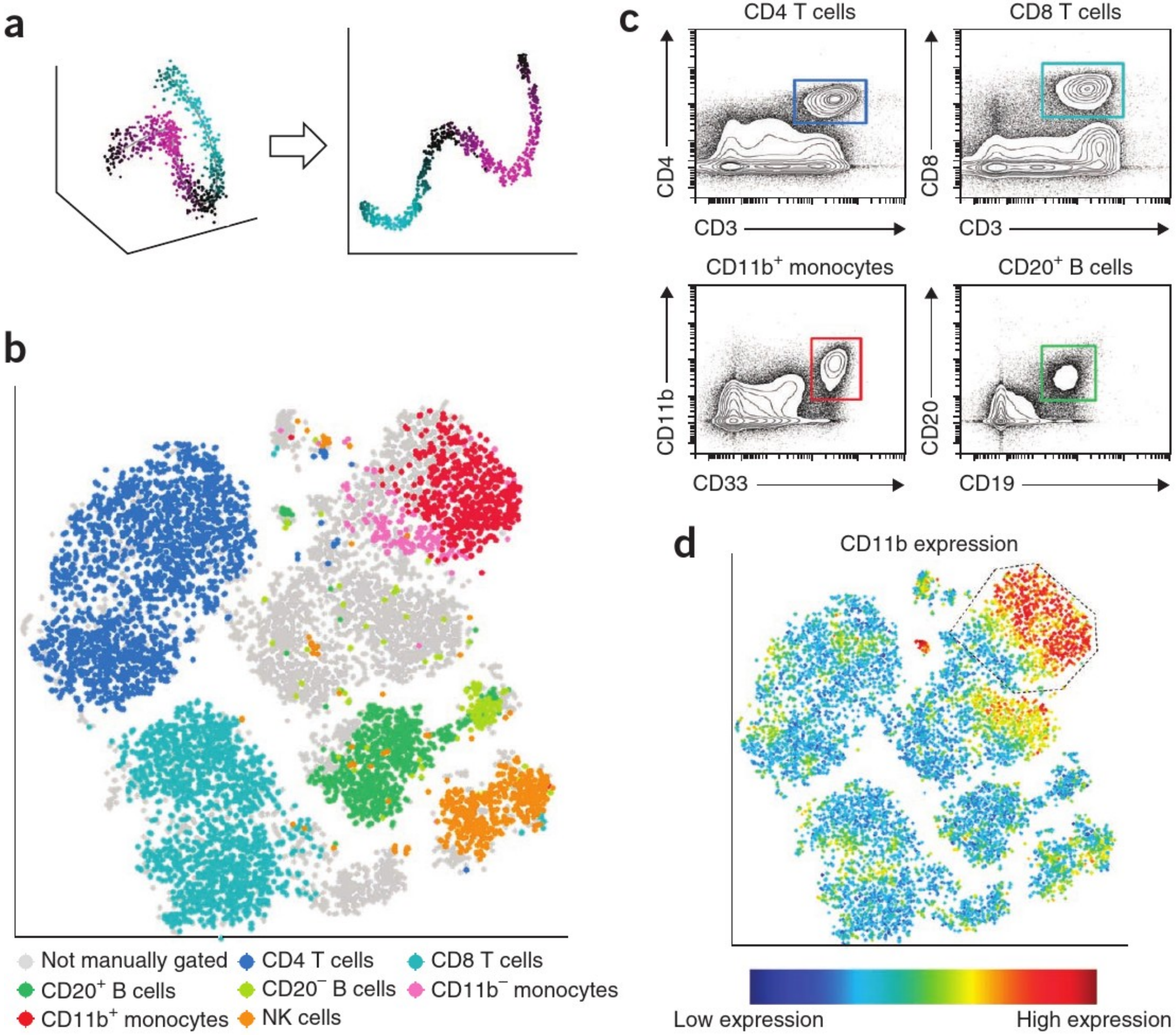
- Single-cell RNAseq
  - Measures gene expression in individual cells
  - Low sensitivity, high throughput
  - Useful for: annotation of cell types, cell differentiation analysis, differential gene expression analysis, …



Stark et al., 2019

Visualization: Dimensionality reduction

- One would like to plot the high-dimensional data in two dimensions
- Enables intuitive and easy visualization of clustering, batch correction, …
- Non-linear dimensionality reduction
  - tSNE
  - UMAP
  - PHATE
  - ...



Amir et al., 2013

tSNE (t-distributed stochastic neighbor embedding)

- Make the distribution of pairwise distances in 2 dimensions as similar as possible to the distance distribution in high-dimensional space → minimize KL-divergence
- Let x_i be the coordinates of point i in high-dimensional space and y_i the coordinates on low-dimensional visualization space

$$p_{i|j} := \frac{\exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2}\right)}{\sum_{k\neq i}\exp\left(\frac{-\|x_k - x_i\|^2}{2\sigma^2}\right)}$$

Conditional probability in feature space (not symmetric)

$$\mathbb{P}_1 \text{ with } \mathbb{P}_1((i,j)) := \frac{p_{i|j} + p_{j|i}}{2n}$$

Symmetrized joint probability

$$\mathbb{P}_2 \text{ with } \mathbb{P}_2((i,j)) := q_{ij} := \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k\neq l}(1 + \|y_k - y_l\|^2)^{-1}}$$

Joint probability in visualization space

Hinton + van der Maaten, 2008

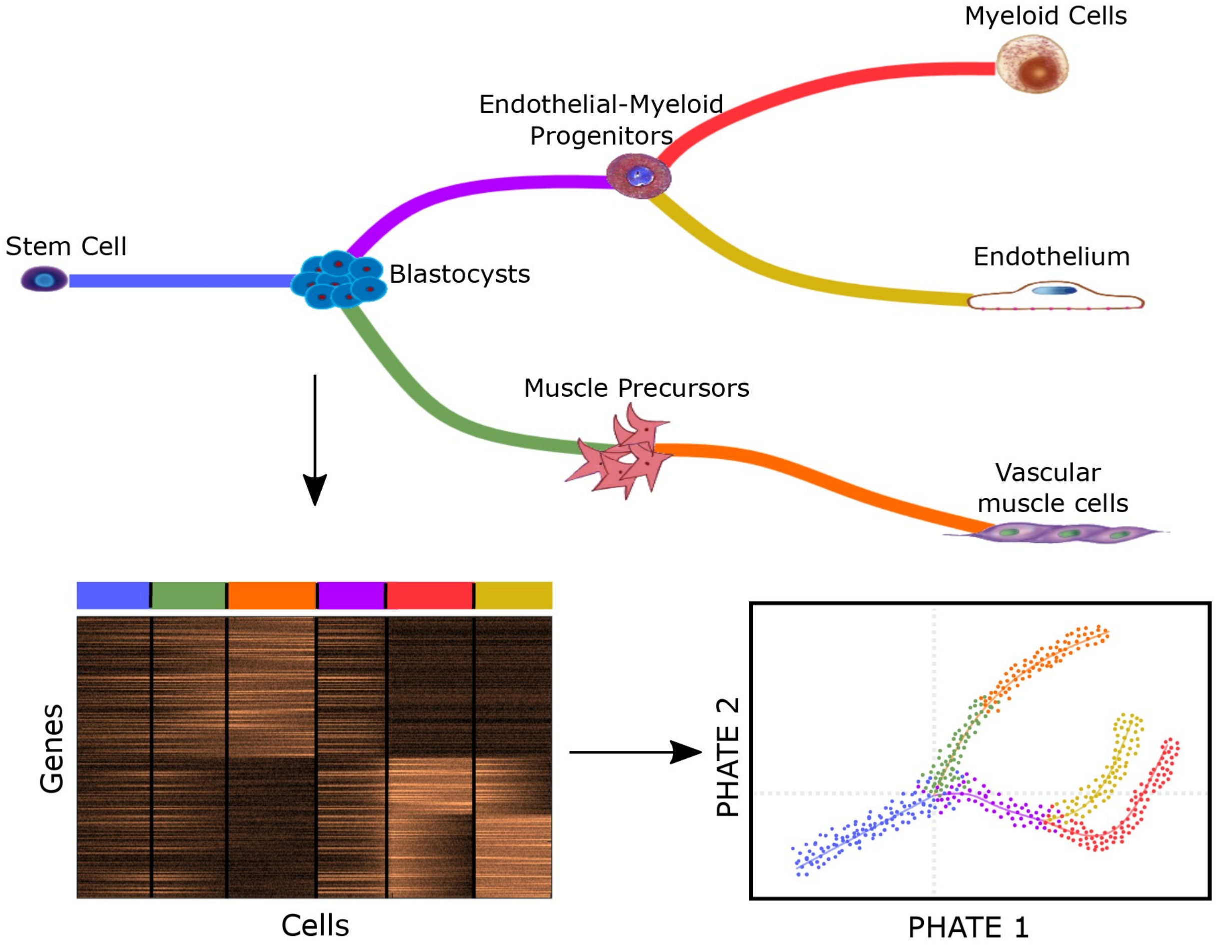
UMAP (Uniform Manifold Approximation and Projection)

- Approximate geodesic distances on data manifold by euclidean k-NN distances
- Use force-directed graph layout in visualization space
- Faster than tSNE, comparable embedding

McInnes et al., 2018

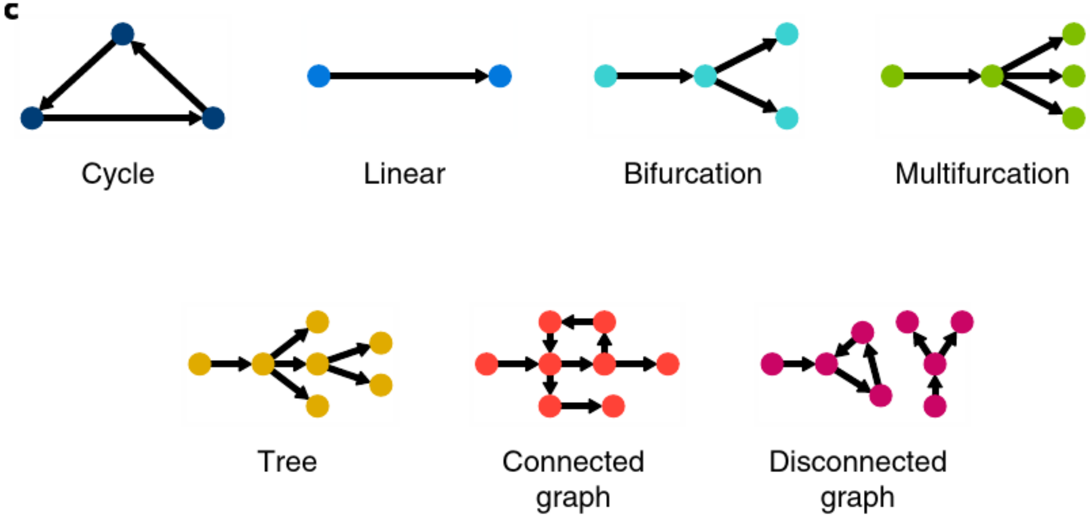
Trajectory inference

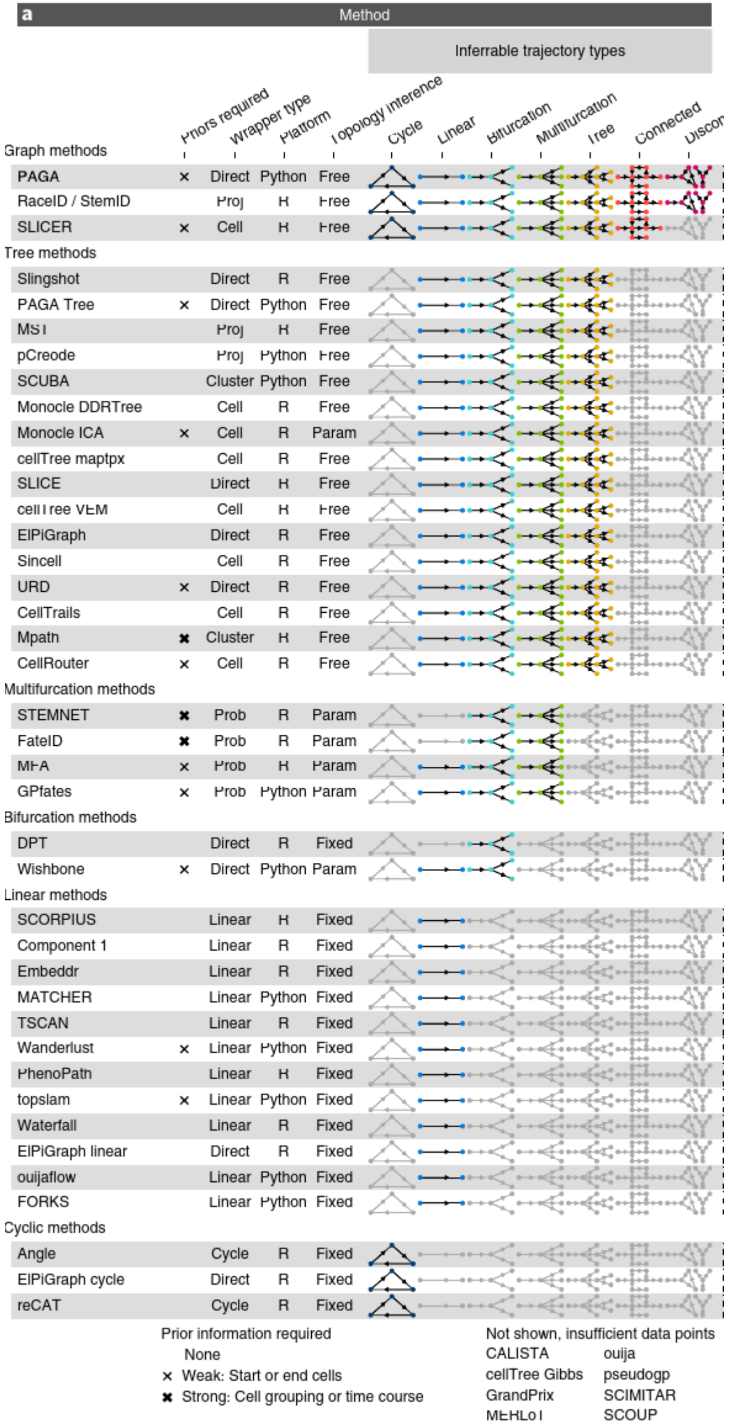
- Differentiating cells have a natural temporal ordering
- Differentiation typically not synchronized → cells from each stage are present in a sample
- Try to infer differentiation trajectory from a snapshot in time



Moon et al., 2017

Trajectory inference




a
Method
Inferrable trajectory types

	Priors required	Wrapper type	Platform	Topology inference	Cycle	Linear	Bifurcation	Multifurcation	Tree	Connected	Discor
Graph methods											
PAGA	×	Direct	Python	Free	✓	✓	✓	✓	✓	✓	✓
RaceID / StemID		Proj	R	Free	✓	✓	✓	✓	✓	✓	✓
SLICER	×	Cell	R	Free	✓	✓	✓	✓	✓	✓	
Tree methods											
Slingshot		Direct	R	Free		✓	✓	✓	✓		
PAGA Tree	×	Direct	Python	Free		✓	✓	✓	✓		
MST		Proj	R	Free		✓	✓	✓	✓		
pCreode		Proj	Python	Free		✓	✓	✓	✓		
SCUBA		Cluster	Python	Free		✓	✓	✓	✓		
Monocle DDRTree		Cell	R	Free		✓	✓	✓	✓		
Monocle ICA	×	Cell	R	Param		✓	✓	✓	✓		
cellTree maptpx		Cell	R	Free		✓	✓	✓	✓		
SLICE		Direct	R	Free		✓	✓	✓	✓		
cellTree VEM		Cell	R	Free		✓	✓	✓	✓		
ElPiGraph		Direct	R	Free		✓	✓	✓	✓		
Sincell		Cell	R	Free		✓	✓	✓	✓		
URD	×	Direct	R	Free		✓	✓	✓	✓		
CellTrails		Cell	R	Free		✓	✓	✓	✓		
Mpath	✖	Cluster	R	Free		✓	✓	✓	✓		
CellRouter	×	Cell	R	Free		✓	✓	✓	✓		
Multifurcation methods											
STEMNET	✖	Prob	R	Param			✓	✓			
FateID	✖	Prob	R	Param			✓	✓			
MFA	×	Prob	R	Param		✓	✓	✓			
GPfates	×	Prob	Python	Param		✓	✓	✓			
Bifurcation methods											
DPT		Direct	R	Fixed			✓				
Wishbone	×	Direct	Python	Param		✓	✓				
Linear methods											
SCORPIUS		Linear	R	Fixed		✓					
Component 1		Linear	R	Fixed		✓					
Embeddr		Linear	R	Fixed		✓					
MATCHER		Linear	Python	Fixed		✓					
TSCAN		Linear	R	Fixed		✓					
Wanderlust	×	Linear	Python	Fixed		✓					
PhenoPath		Linear	R	Fixed		✓					
topslam	×	Linear	Python	Fixed		✓					
Waterfall		Linear	R	Fixed		✓					
ElPiGraph linear		Direct	R	Fixed		✓					
ouijaflow		Linear	Python	Fixed		✓					
FORKS		Linear	Python	Fixed		✓					
Cyclic methods											
Angle		Cycle	R	Fixed	✓						
ElPiGraph cycle		Direct	R	Fixed	✓						
reCAT		Cycle	R	Fixed	✓						

Prior information required
None
× Weak: Start or end cells
✖ Strong: Cell grouping or time course

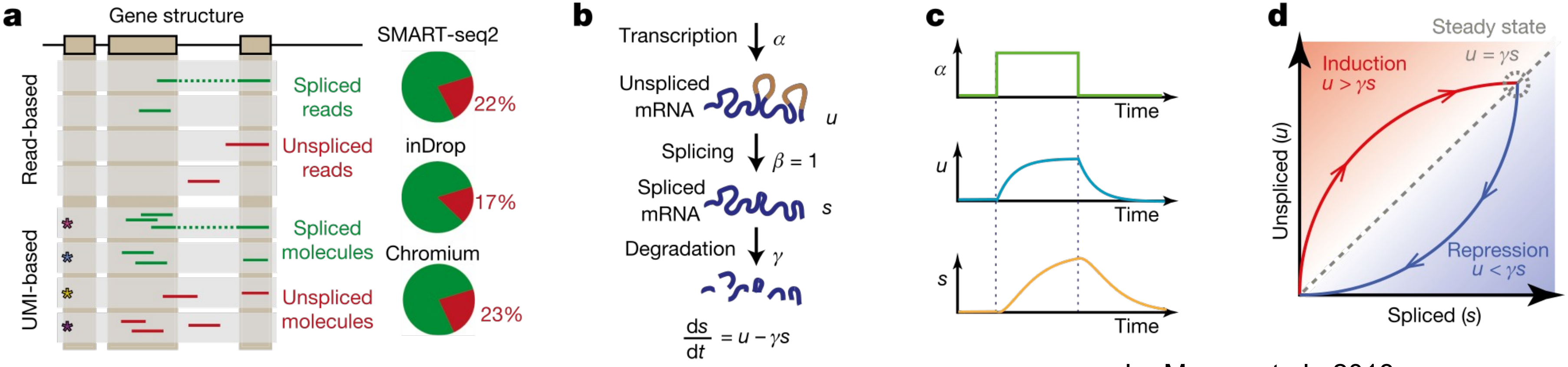
Not shown, insufficient data points

CALISTA	ouija
cellTree Gibbs	pseudogp
GrandPrix	SCIMITAR
MERLoT	SCOUP

Saelens et al., 2019

RNA velocity

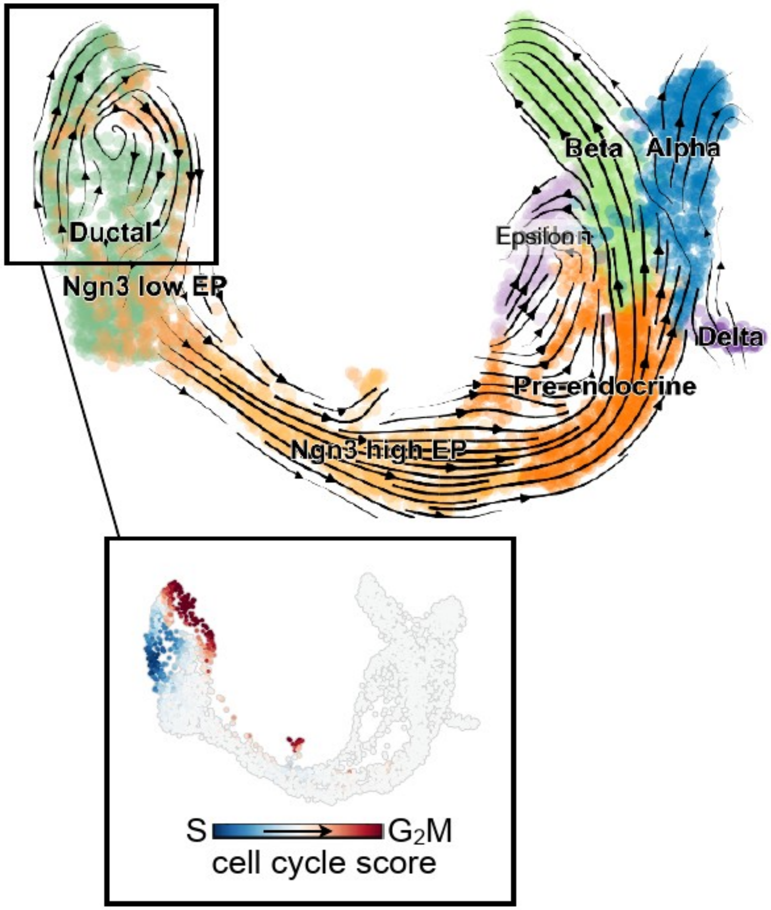
- It is possible to classify sequenced mRNAs into two classes
  - mature (old, being currently translated into protein)
  - Newly synthesized (will be translated in the future)
- This makes it possible to predict changes in gene expression for single cell (i.e. predict the future position of a cell on the manifold)
- Data extremely sparse

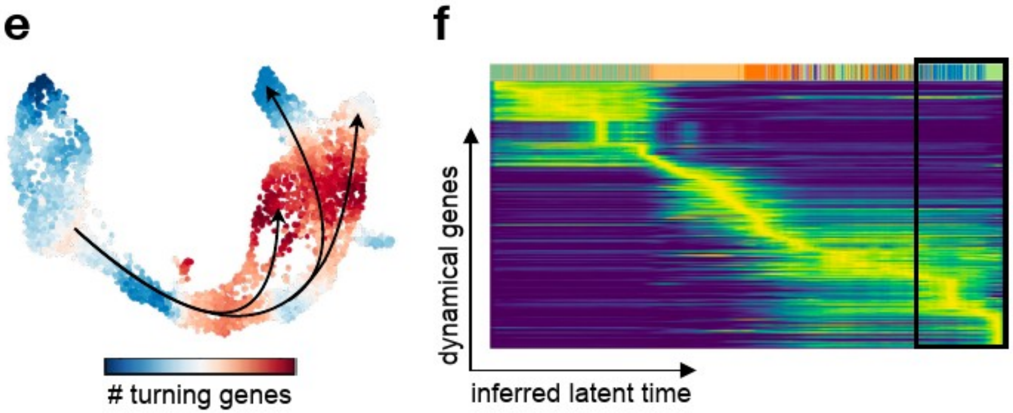


La Manno et al., 2018

RNA velocity

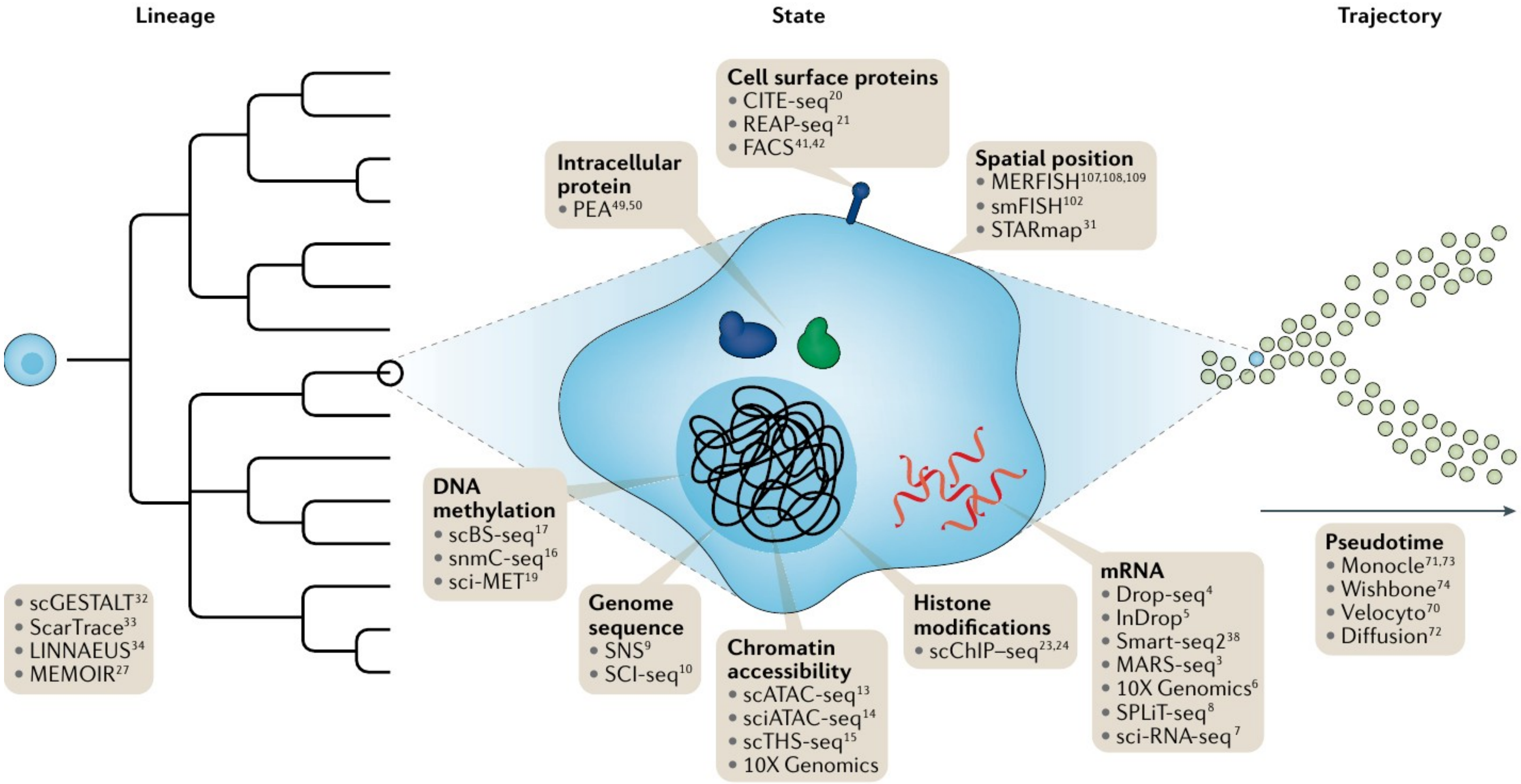
- RNA velocity can be used for semi-automatic trajectory inference





Bergen et al., 2019

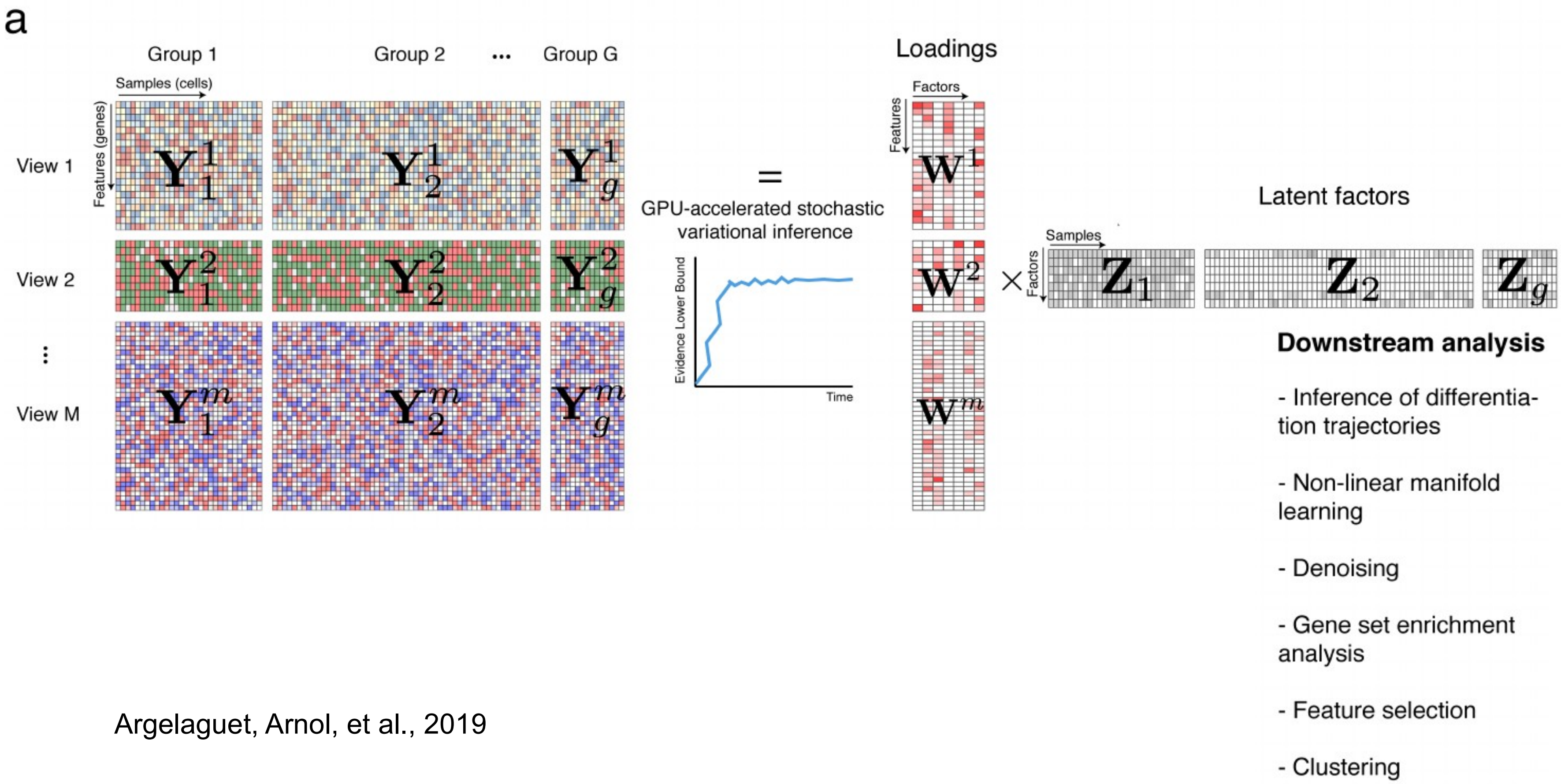
Single-cell omics



Stuart + Satija, 2019

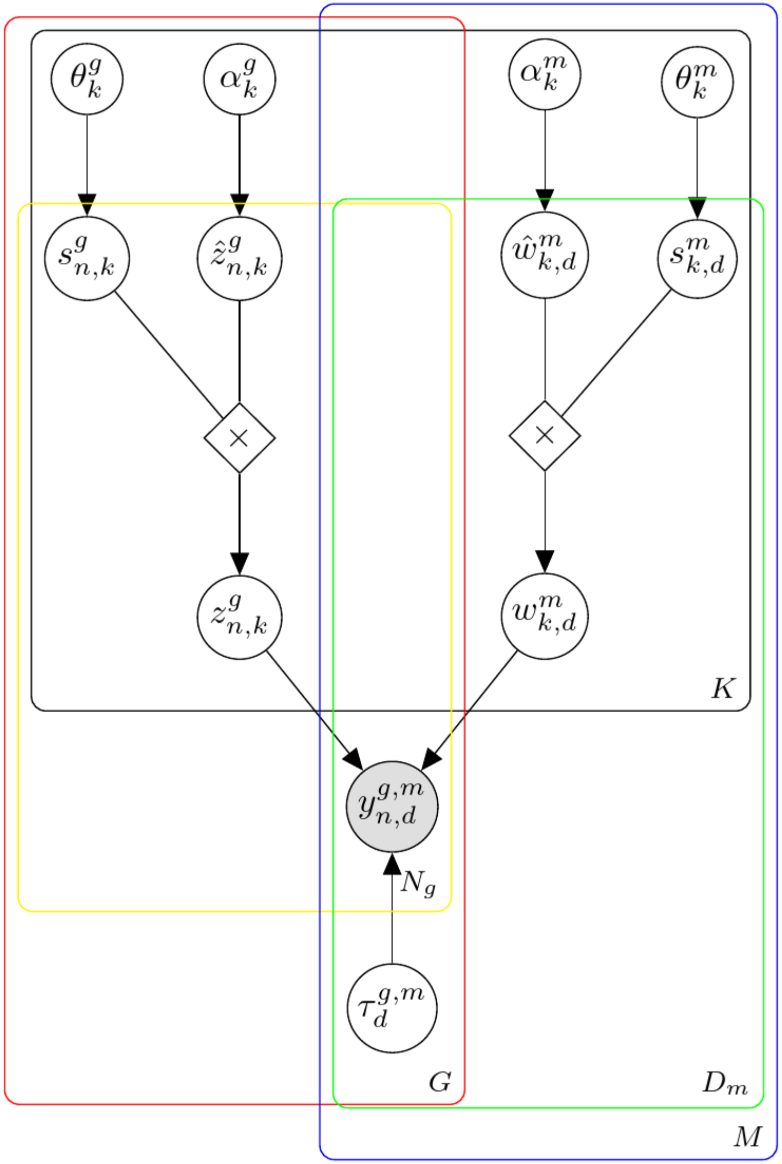
MOFA+ (Multi-Omics Factor Analysis)

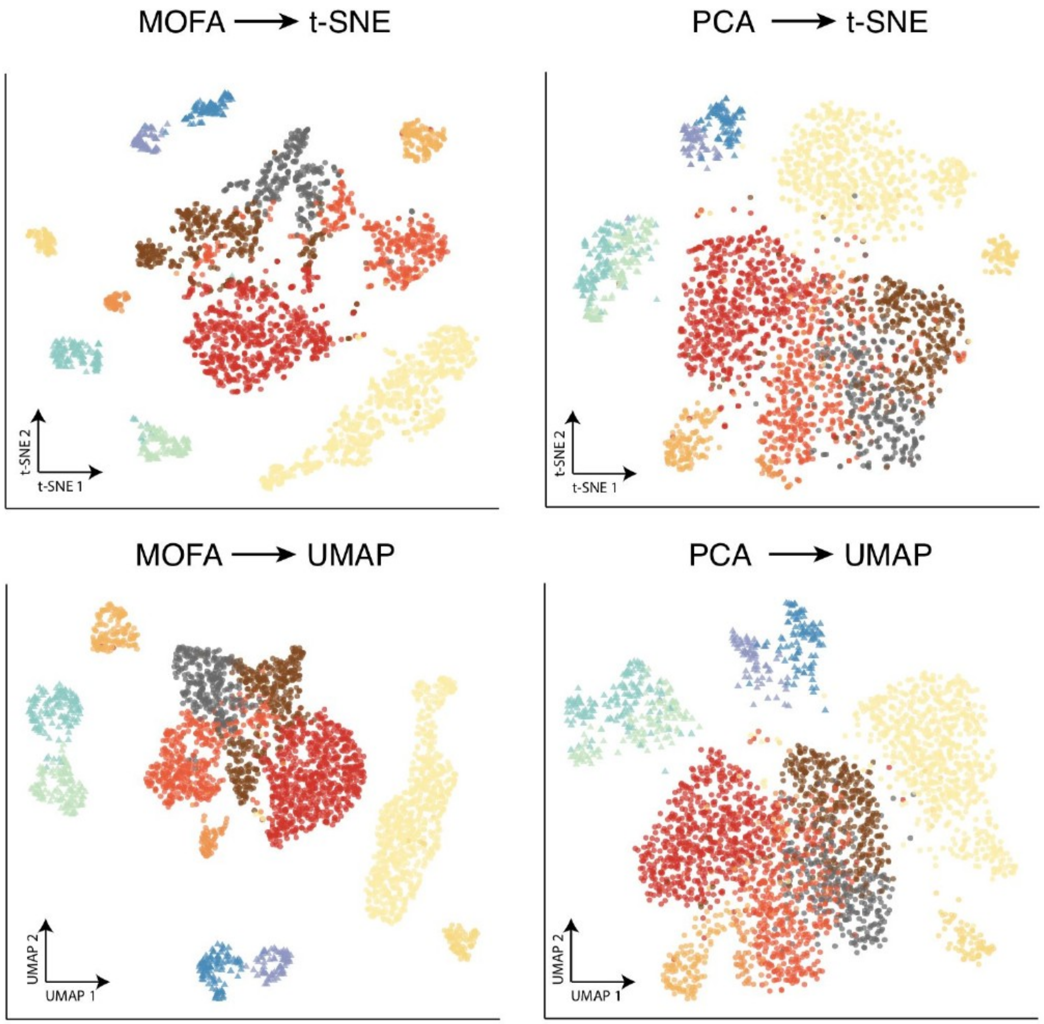
- Views: data modalities (e.g. RNAseq, protein abundance, …)
- Groups: batches/conditions



Argelaguet, Arnol, et al., 2019

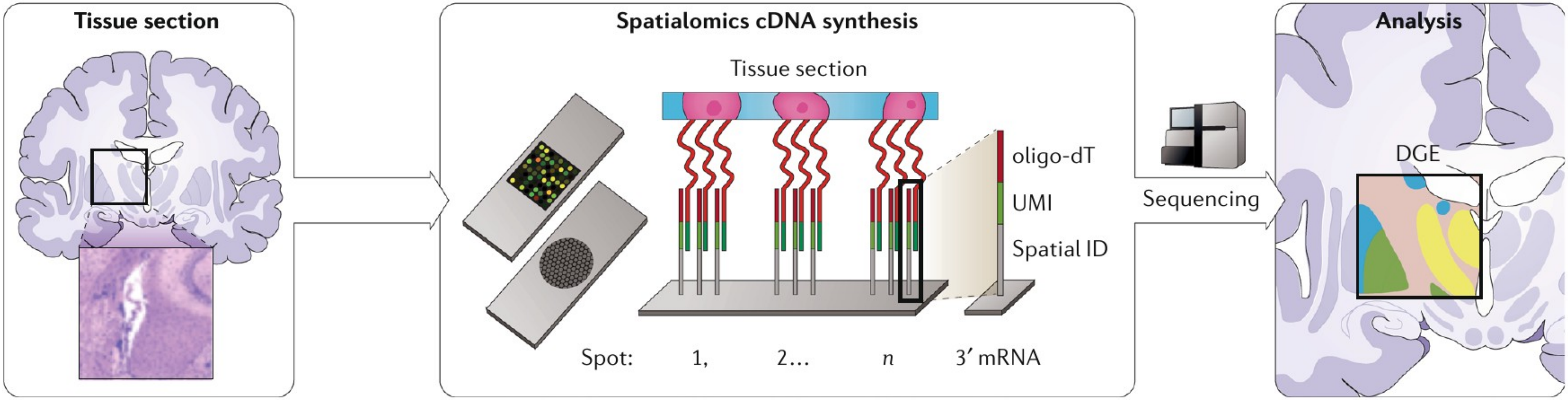
MOFA+ (Multi-Omics Factor Analysis)





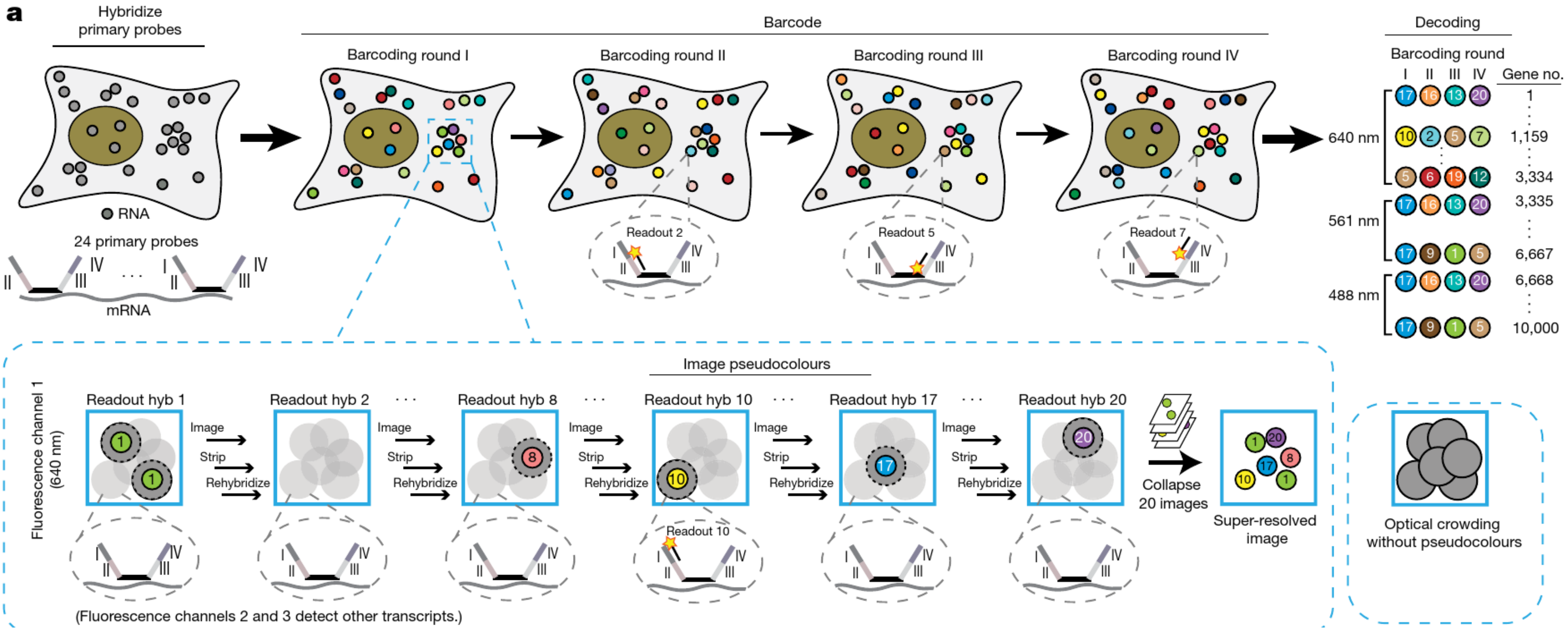
Argelaguet, Arnol, et al., 2019

Spatial omics



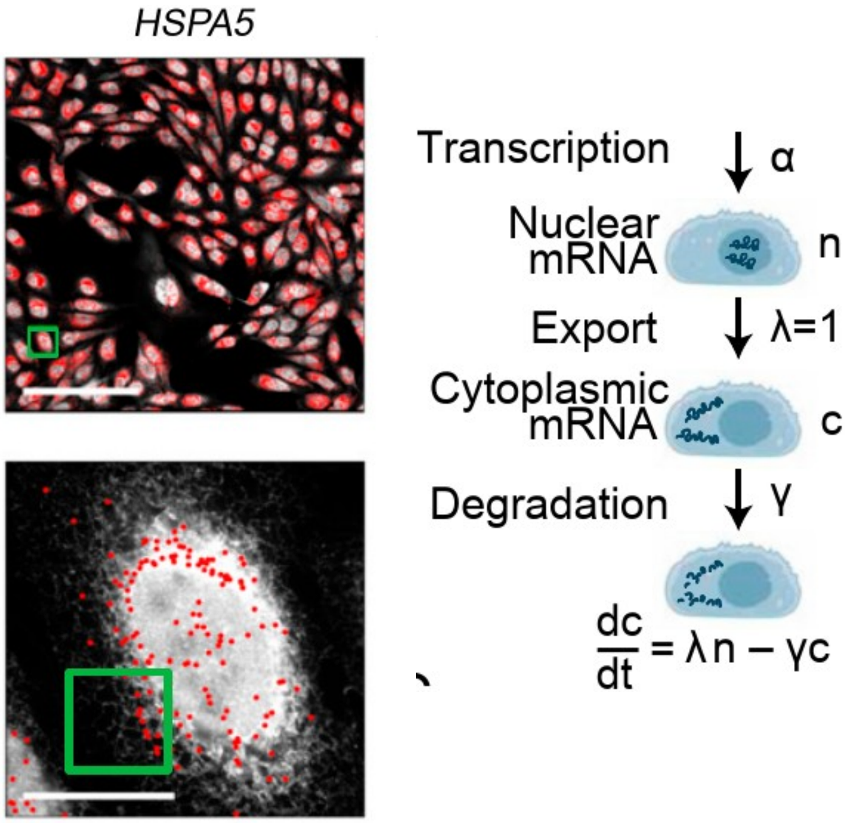
Stark et al., 2019

Spatial omics

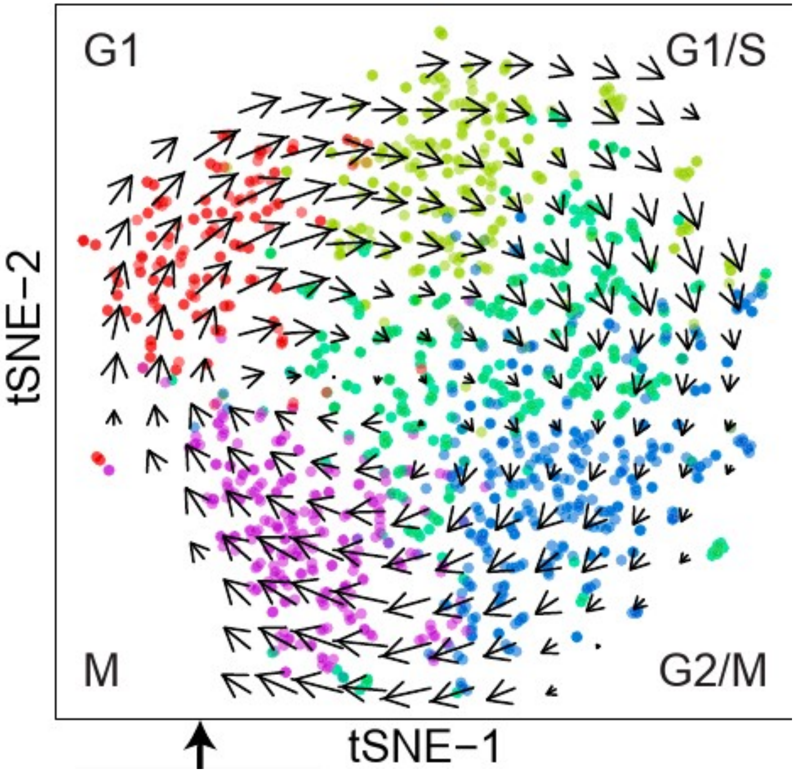


Eng et al., 2019

Spatial omics: RNA velocity



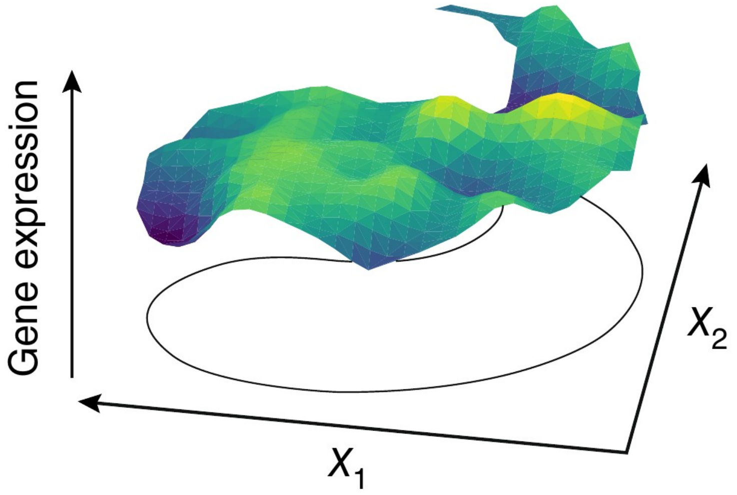
$$\frac{dc}{dt} = \lambda n - \gamma c$$

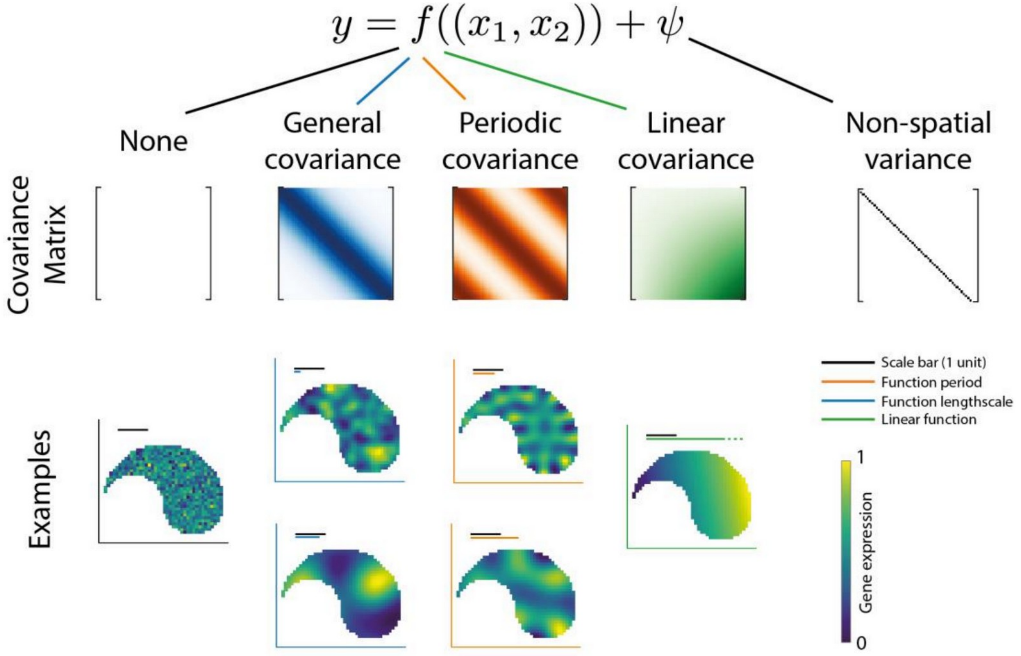


Xia et al, 2019

SpatialDE

- Models spatial gene expression as Gaussian Process
- Detection of spatially variable genes, classification into types of spatial variability





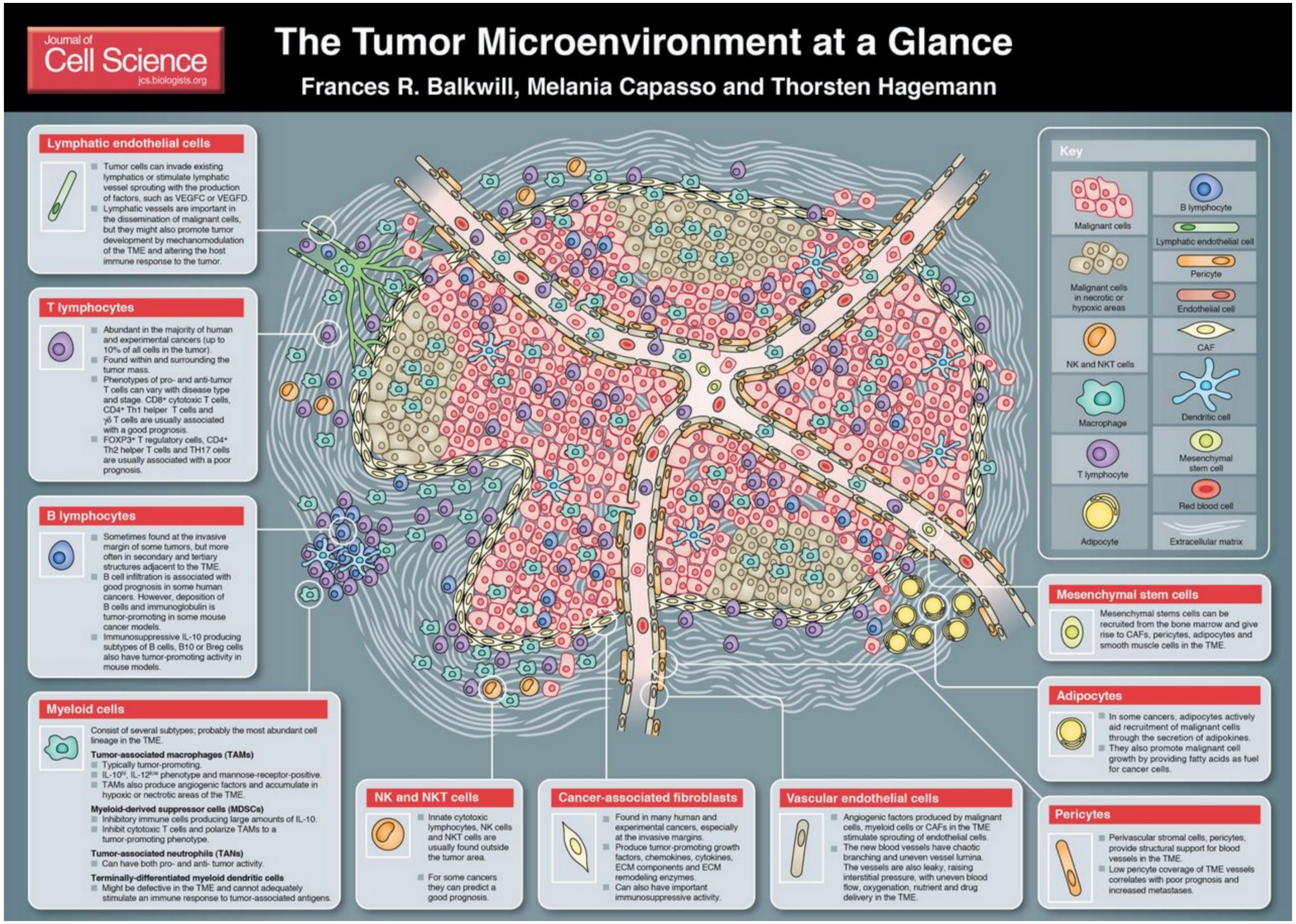
Svensson et al., 2018

General issues in single cell omics

- Data sets are becoming larger very fast
  - Currently: largest available scRNAseq data set has approx. 10^6 cells
  - Commercial platform for spatial RNAseq: 5000 spots
- Methods need to be computationally efficient and scale well
  - Increasing use of GPUs
- Very sparse data, large proportion of missing values
- No ground truth
  - Difficult to know if a method is doing The Right Thing™
- Data typically non-Gaussian
  - Makes exact calculations difficult, inefficient, or impossible
  - Sometimes transformation to approximate Gaussianity possible

Microenvironments

- Core idea: interactions/communications between cells



The Tumor Microenvironment at a Glance

Frances R. Balkwill, Melania Capasso and Thorsten Hagemann

Lymphatic endothelial cells

- Tumor cells can invade existing lymphatics or stimulate lymphatic vessel sprouting with the production of factors, such as VEGFC or VEGFD.
- Lymphatic vessels are important in the dissemination of malignant cells, but they might also promote tumor development by mechanomodulation of the TME and altering the host immune response to the tumor.

T lymphocytes

- Abundant in the majority of human and experimental cancers (up to 10% of all cells in the tumor).
- Found within and surrounding the tumor mass.
- Phenotypes of pro- and anti-tumor T cells can vary with disease type and stage. $CD8^+$ cytotoxic T cells, $CD4^+$ Th1 helper T cells and γδ T cells are usually associated with a good prognosis.
- $FOXP3^+$ T regulatory cells, $CD4^+$ Th2 helper T cells and TH17 cells are usually associated with a poor prognosis.

B lymphocytes

- Sometimes found at the invasive margin of some tumors, but more often in secondary and tertiary structures adjacent to the TME.
- B cell infiltration is associated with good prognosis in some human cancers. However, deposition of B cells and immunoglobulin is tumor-promoting in some mouse cancer models.
- Immunosuppressive IL-10 producing subtypes of B cells, B10 or Breg cells also have tumor-promoting activity in mouse models.

Myeloid cells

Consist of several subtypes; probably the most abundant cell lineage in the TME.

Tumor-associated macrophages (TAMs)

- Typically tumor-promoting.
- IL-10^{hi}, IL-12^{low} phenotype and mannose-receptor-positive.
- TAMs also produce angiogenic factors and accumulate in hypoxic or nectrotic areas of the TME.

Myeloid-derived suppressor cells (MDSCs)

- Inhibitory immune cells producing large amounts of IL-10.
- Inhibit cytotoxic T cells and polarize TAMs to a tumor-promoting phenotype.

Tumor-associated neutrophils (TANs)

- Can have both pro- and anti- tumor activity.

Terminally-differentiated myeloid dendritic cells

- Might be defective in the TME and cannot adequately stimulate an immune response to tumor-associated antigens.

NK and NKT cells

- Innate cytotoxic lymphocytes, NK cells and NKT cells are usually found outside the tumor area.
- For some cancers they can predict a good prognosis.

Cancer-associated fibroblasts

- Found in many human and experimental cancers, especially at the invasive margins.
- Produce tumor-promoting growth factors, chemokines, cytokines, ECM components and ECM remodeling enzymes.
- Can also have important immunosuppressive activity.

Vascular endothelial cells

- Angiogenic factors produced by malignant cells, myeloid cells or CAFs in the TME stimulate sprouting of endothelial cells.
- The new blood vessels have chaotic branching and uneven vessel lumina. The vessels are also leaky, raising interstitial pressure, with uneven blood flow, oxygenation, nutrient and drug delivery in the TME.

Mesenchymal stem cells

Mesenchymal stems cells can be recruited from the bone marrow and give rise to CAFs, pericytes, adipocytes and smooth muscle cells in the TME.

Adipocytes

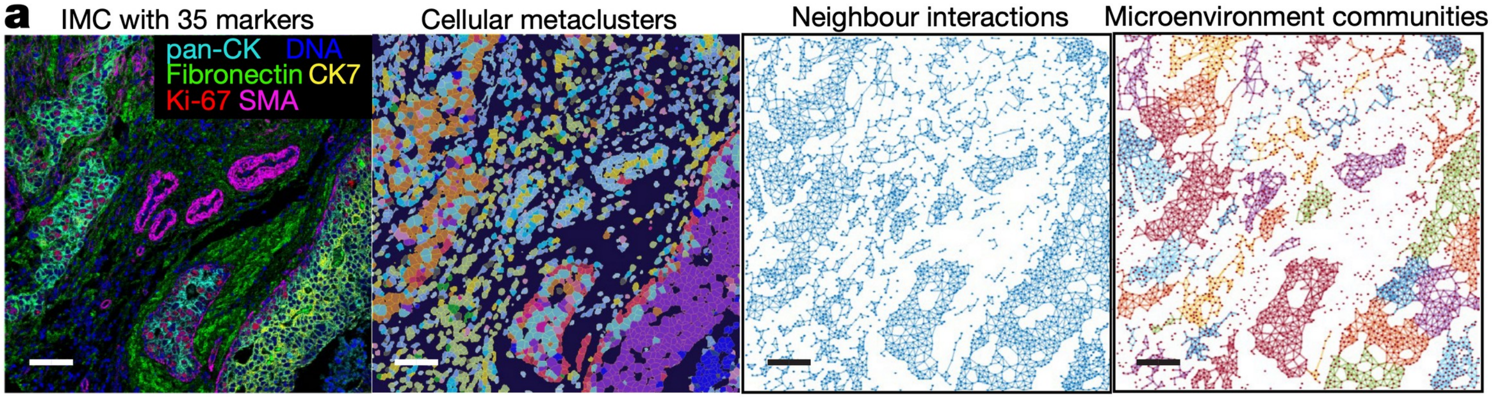
- In some cancers, adipocytes actively aid recruitment of malignant cells through the secretion of adipokines.
- They also promote malignant cell growth by providing fatty acids as fuel for cancer cells.

Pericytes

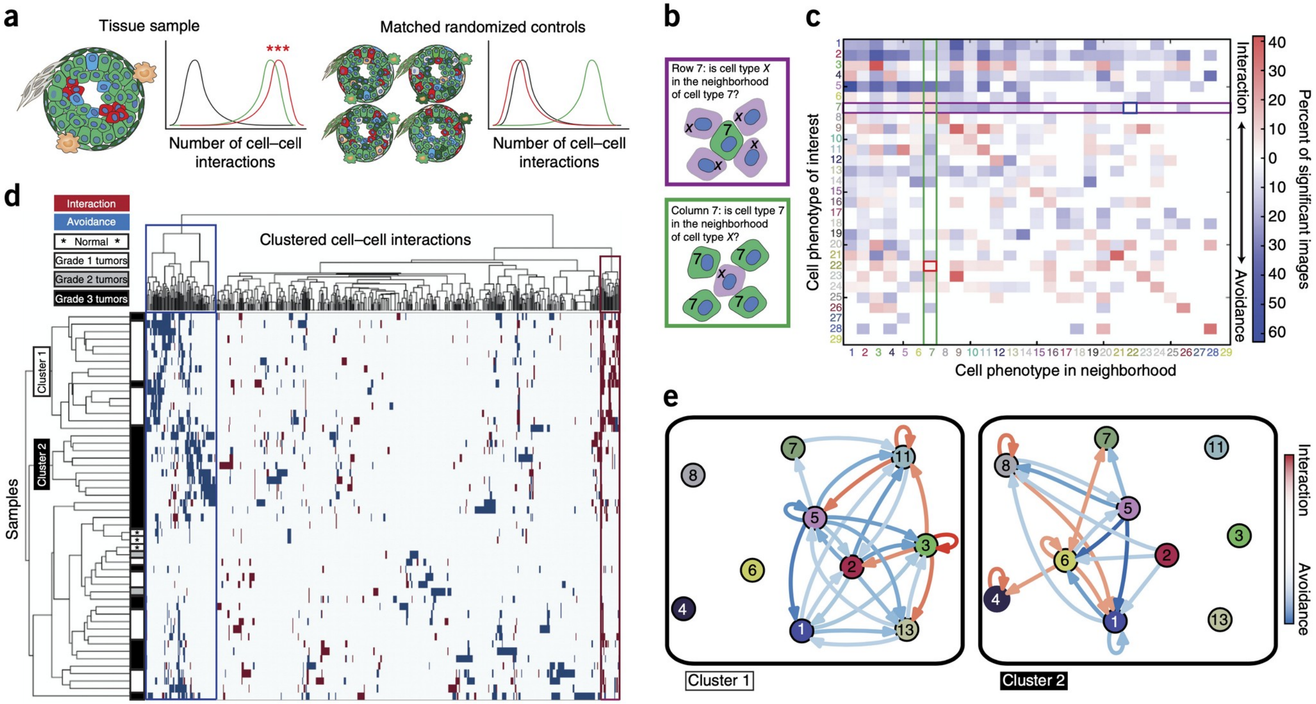
- Perivascular stromal cells, pericytes, provide structural support for blood vessels in the TME.
- Low pericyte coverage of TME vessels correlates with poor prognosis and increased metastases.

Balkwill et al. The tumor microenvironment at a glance (J. Cell Sci 2012)

Microenvironments



Adapted from Jackson et al. The single-cell pathology landscape of breast cancer (Nature 2020)



Approach:

1) Instance segmentation

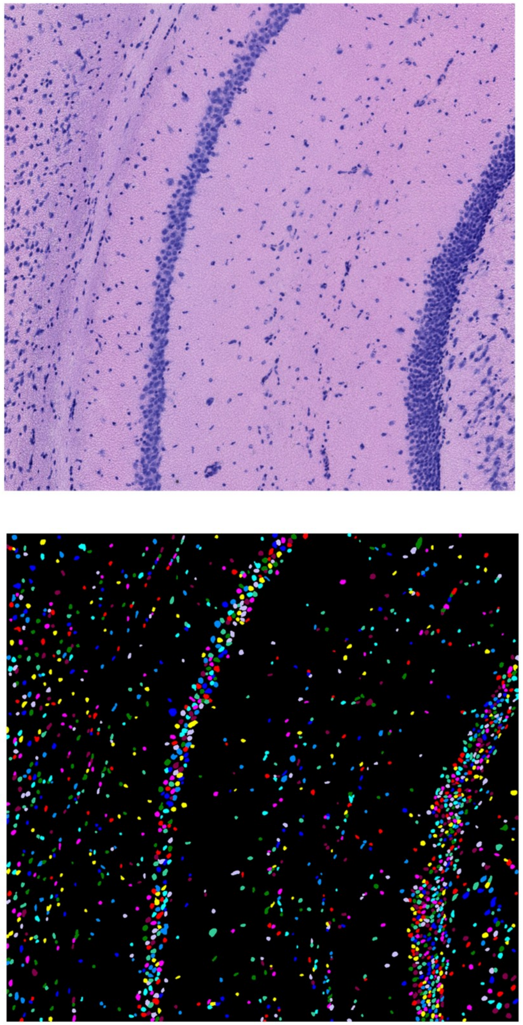
2) Graph construction, features averaging pixel signal

3) Permutation test

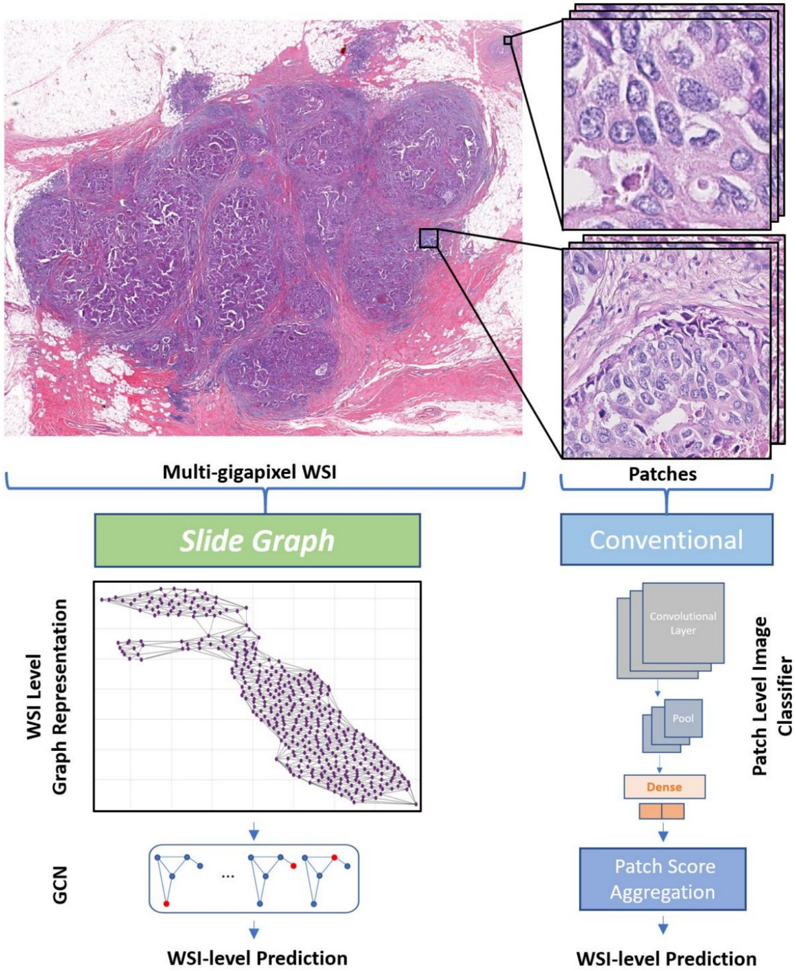
4) Clustering

Schapiro et al. histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data (Nat. methods 2017)

Microenvironments



Applying the method presented in Carpenter et al. (Nat. methods, 2018)



WSI = Whole slide image

Lu et al. Capturing Cellular Topology in Multi-Gigapixel Pathology Images (CVPR 2020)

Approach:

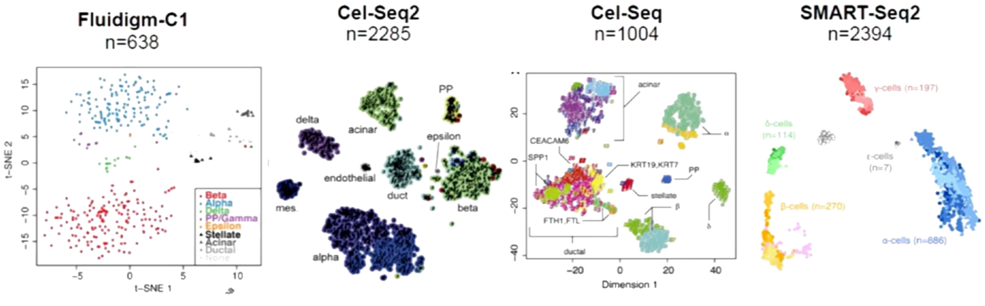
1) Instance segmentation from H&E stained images

2) Graph construction, features by clustering image features

3) Predictions with GCN

Alignment

- Same biology, different experiments → need to match the data

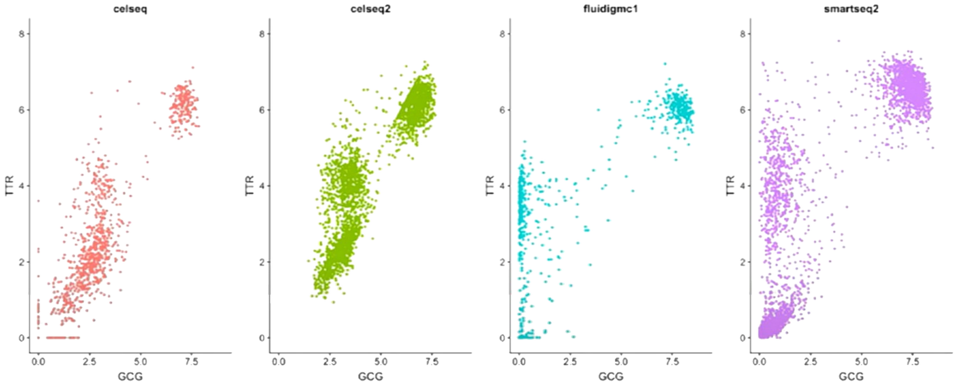


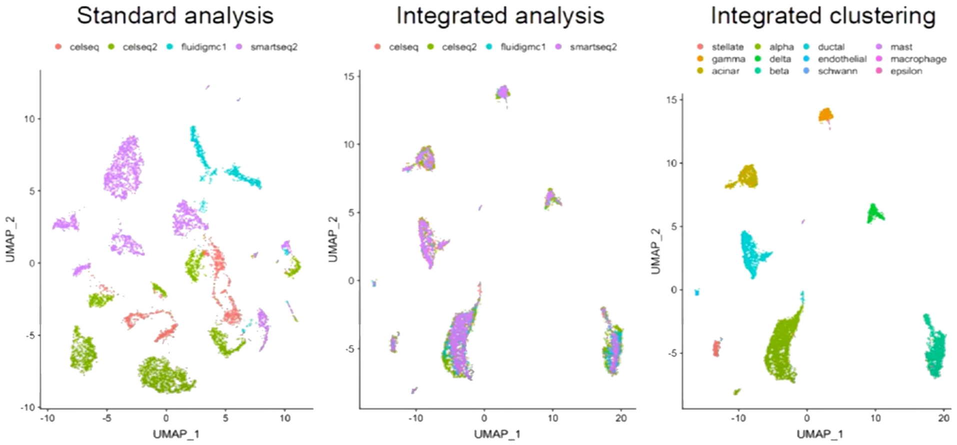
Lawlor et al
Genome Research, 2016

Muraro et al
Cell Systems, 2016

Grun et al
Cell Stem Cell, 2016

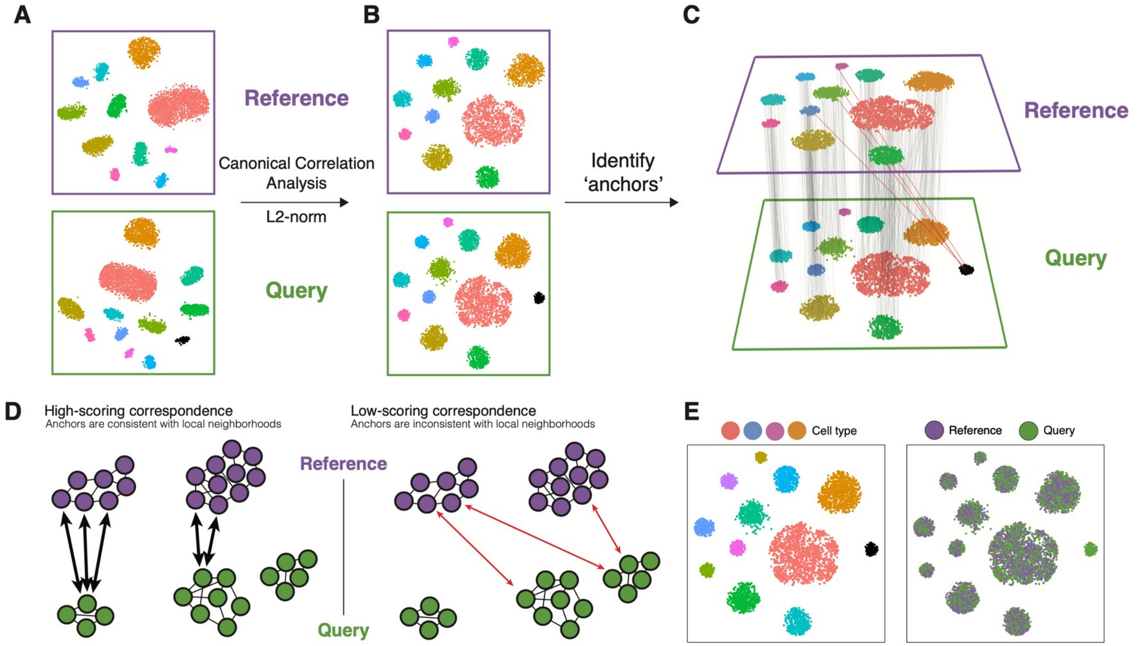
Segerstolpe et al
Cell Metabolism, 2016





Adapted from Comprehensive Integration of Single Cell Data (presentation of Rahul Satija)

Alignment


A
B
C
Reference
Query
Canonical Correlation Analysis
L2-norm
Identify 'anchors'
Reference
Query
D
High-scoring correspondence
Anchors are consistent with local neighborhoods
Low-scoring correspondence
Anchors are inconsistent with local neighborhoods
Reference
Query
E
Cell type
Reference
Query

Satija et al. Comprehensive Integration of Single Cell Data (bioRxiv 2018)

Approach:

1) Joint dimensionality reduction

$$X_{f,c}, Y_{f,c} \quad f \text{ features}, c \text{ cells}, \text{same features}$$

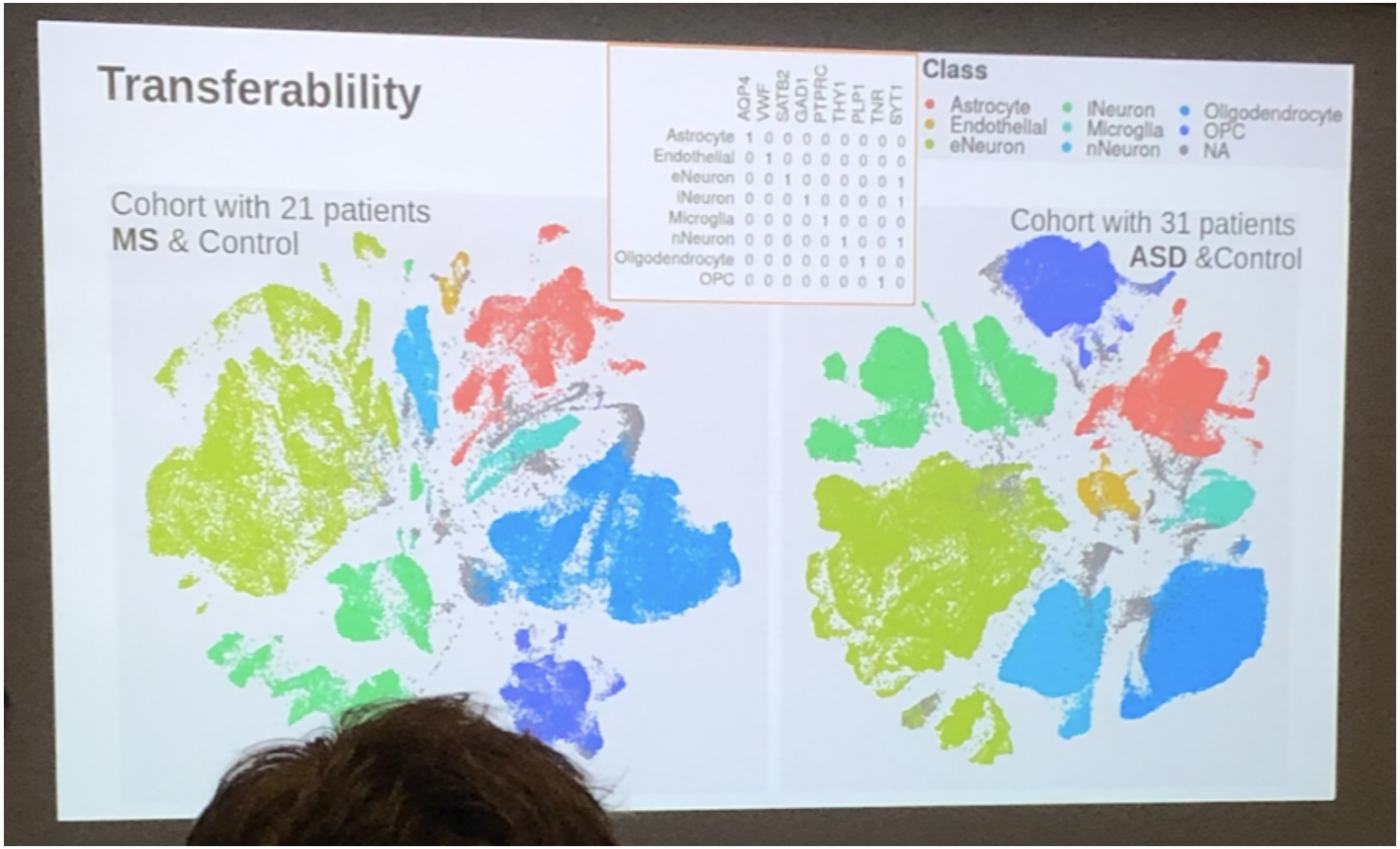
$$\max_{u,v} u^\top X^\top Y v, \quad \|u\|^2 \leq 1, \|v\|^2 \leq 1$$

2) Finding anchors by mutual k nearest neighbors

3) Anchor weighting (by a Gaussian kernel) and alignment

Alignment

- Idea: considering the topology for improving the alignment



Transferablility

	AQP4	VWF	SATB2	GAD1	PTPRC	THY1	PLP1	TNR	SYT1
Astrocyte	1	0	0	0	0	0	0	0	0
Endothelial	0	1	0	0	0	0	0	0	0
eNeuron	0	0	1	0	0	0	0	0	1
iNeuron	0	0	0	1	0	0	0	0	1
Microglia	0	0	0	0	1	0	0	0	0
nNeuron	0	0	0	0	0	1	0	0	1
Oligodendrocyte	0	0	0	0	0	0	1	0	0
OPC	0	0	0	0	0	0	0	1	0

Picture from a presentation of F. Frauhammer, DKFZ

Some directions that could be worth exploring

- Topology-aware manifold aligner
- Differential expression in manifolds
- Geometric deep learning

Datasets:

- Many modalities
  - Paired (many modalities for the same entity)
  - Unpaired (from the same or from different samples)
- High dimensionality (>20000 genes in humans)
- Multi-channel, high-resolution images
- Data with tree-like latent structure (suitable for hyperbolic embeddings)

Thank you
for your attention!

dkfz. GERMAN
CANCER RESEARCH CENTER
IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

SpatialDE

- Boundary effects?
- Euclidean distance smaller than geodesic distance
- Difference in detected patterns?