Acknowledgement of Country

The University of Queensland (UQ) acknowledges the Traditional Owners and their custodianship of the lands on which we meet.

We pay our respects to their Ancestors and their descendants, who continue cultural and spiritual connections to Country.

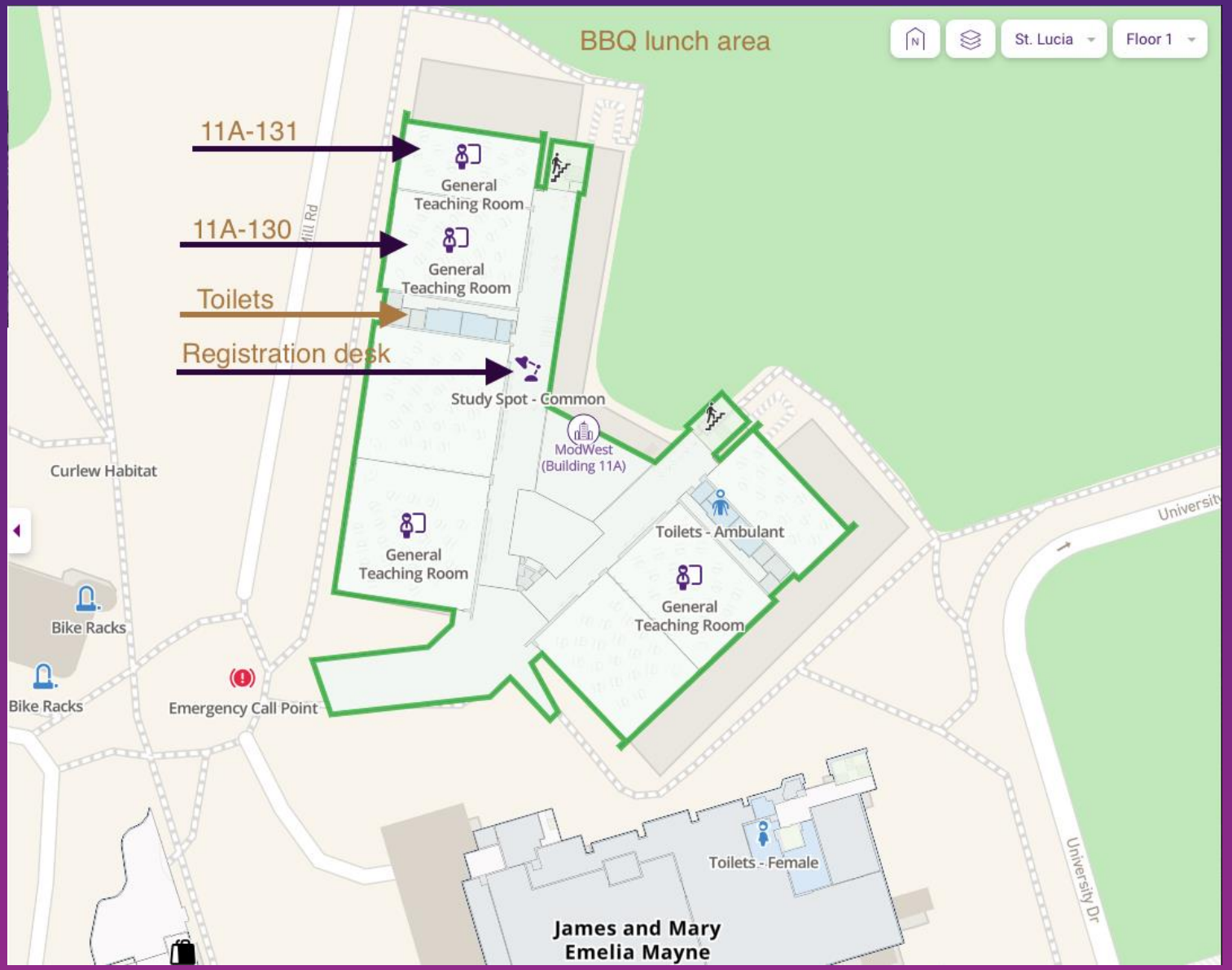
We recognise their valuable contributions to Australian and global society.

Image: Digital reproduction of *A guidance through time* by Casey Coolwell and Kyra Mancktelow

General Information:

- We are currently located in Building 11A MODWEST
  - Bathrooms
  - Vending machines
- Food court and other bathrooms are located in Building 63 or Building 21B
- If you are experiencing cold/flu symptoms or have had COVID in the last 7 days please ensure you are wearing a mask for the duration of the module

EMERGENCY CONTACT/UQ SECURITY: 3365 3333



Learning materials

Instructions to access WiFi/desktop/server:

https://cnsgenomics.com/data/teaching/GNGWS26/module0/

The winter school server is available until **24th July 2026** (2 weeks after the course)

Slides and practical notes for this module:

https://cnsgenomics.com/data/teaching/GNGWS26/module2/

Module 2 Cellular Omics

Room 315/316, Building 69

Lecturers/Instructors: Quan Nguyen, Xiao Tan, Prakrithi Pavithra

Module 2 - running the learning materials

https://github.com/GenomicsMachineLearning/gml-teaching-2026

Copy and paste each of the following lines into your terminal once you have logged into the workshop server:

- `/software/bin/micromamba shell init`
- `source ~/.bashrc`
- `micromamba activate /software/conda-envs/gml-teaching`
- `git clone https://github.com/GenomicsMachineLearning/gml-teaching-2026`
- `~/qimr-teaching-2026/runme.sh`

The output will look something like:

```
Port 3502 is available

Command to create ssh tunnel:
ssh -N -L 3502:10.10.10.10:3502 foo@10.10.10.10
Use a Browser on your local machine to go to:
localhost:3502  (prefix w/ https:// if using password)

[I 2024-06-20 05:57:41.633 ServerApp] Extension package jupyter_lsp took 0.1372s to import
[I 2024-06-20 05:57:44.647 ServerApp]     http://127.0.0.1:3502/tree?token=abc123
```

- Copy the line beginning with "ssh" into a new terminal, on your local computer, and hit [Enter].
- Copy the text beginning with "http://127.0.0.1" into a new tab in your browser, and hit [Enter].

Module 2 Cellular Omics – Leaning Objectives

- [Day 1] How the data get generated experimentally – latest sequencing and imaging technologies:
  - Technologies for generating single-cell and spatial transcriptomics data
  - Technologies for other spatial omics, focusing on proteomics

- [Day 1] Exploratory visualisation to understand the data

- [Days 1, 2] Statistical analyses to discover new biological processes and biomarkers associated with disease, including cells, genes and groups of cells within the tissue. This includes:
  - Identifying cell types
  - Finding gene markers
  - Mapping cell neighbourhoods (cell communities)
  - Multiomics integration
    - Spatial transcriptomics + proteomics
    - gsMAP: incorporating GWAS with Spatial transcriptomics

- [Day 2] Machine learning analysis of sequencing and imaging data
  - Key concepts
  - Practical hands-on analyses
  - Transformer models for gene expression
  - Vision transformer for H&E images

Module 2 Cellular Omics – Schedule

- Afternoon 6th July: 1 pm – 4 pm
  - ✓ Session 1: Single cell and spatial technologies (1:00-2:00)
  - ✓ Session 2: Practical - (2:00-2:45)
    - ✓ Single cell analysis (2:00-2:30)

  Break (2:45-3:00)
  - ✓ Session 3: spatial analysis (3:00-3:30)
  - ✓ Session 4: Visualise spatial single cell data (3:30-3:50)
  - ✓ Closing: 3:50-4

- Morning 7th July (9 am – 12 pm) Practical for single cell and spatial analysis
  - ✓ Session 1: Cell typing (9-10)
  - ✓ Session 2: Community analysis (10-10:15)

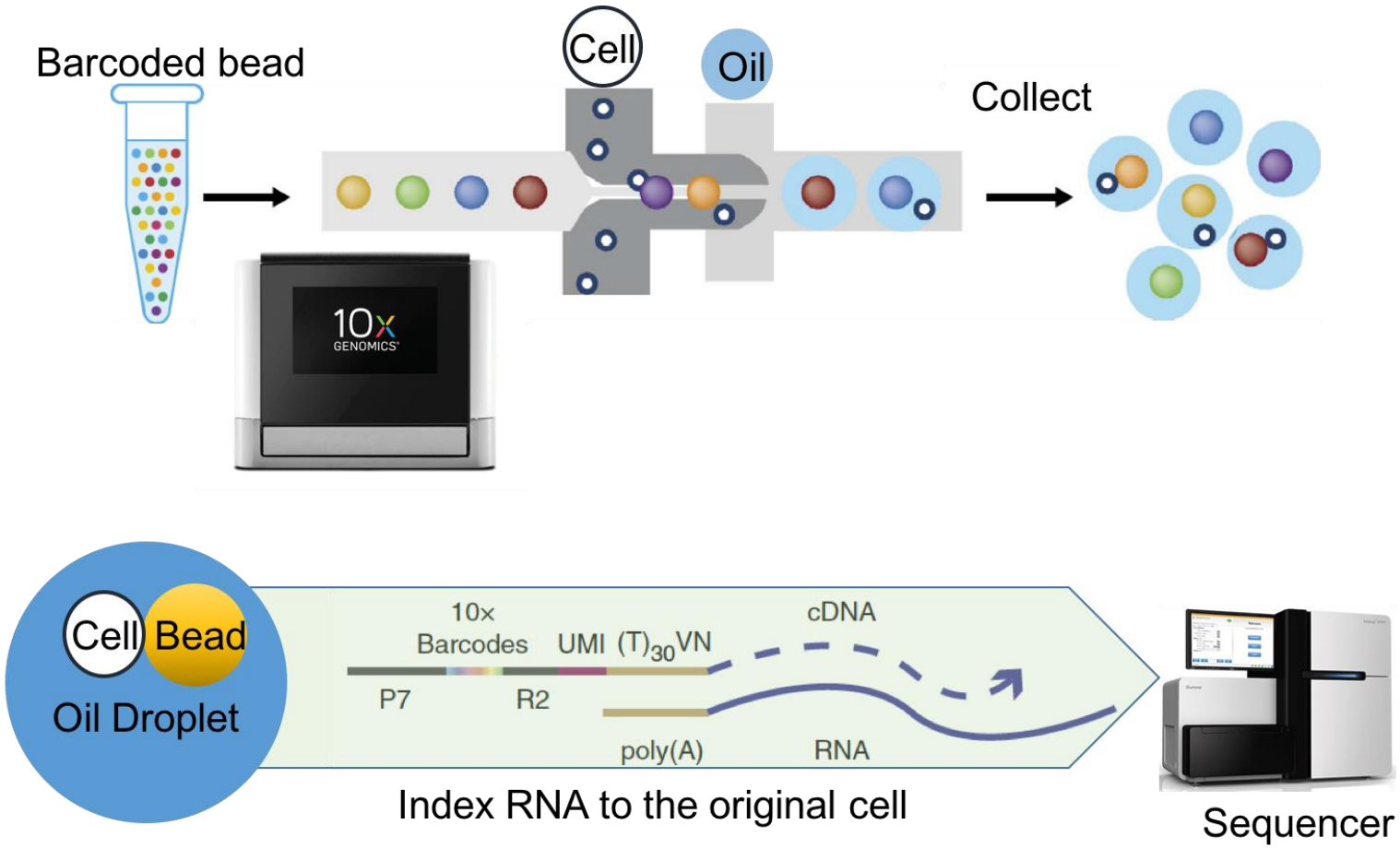
  Break: 10:15-10:30 am
  - ✓ Session 3: Multiomics Integration lecture (10:30-11:30am)
  - ✓ Practical: gsMap (11.30-12 am)

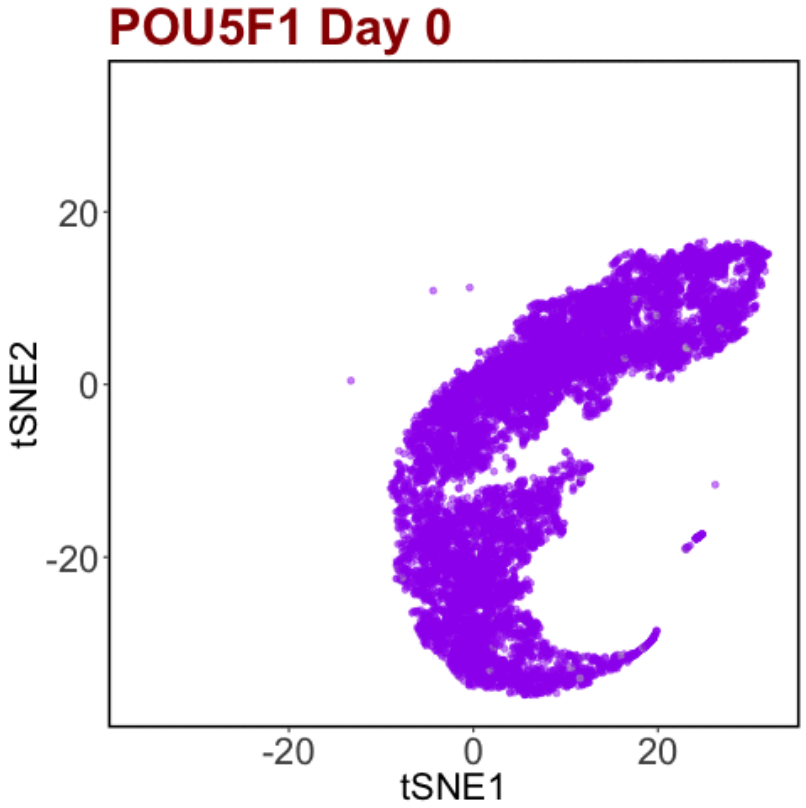
- Morning 7th July (1 pm – 4 pm) Practical for machine learning
  - ✓ Session 1: Deep learning for spatial data (1:00-1:30)
  - ✓ Session 2: Autoencoder for gene expression (1:30 -2:45)

  Break (2:45-3:00)
  - ✓ Session 3: transformer for genes and images(3:00-3:50)
  - ✓ Closing: 3:50-4:00

Lecture 1: Single Cell Transcriptomics

Single cell RNA sequencing





- Single-cell RNA sequencing (scRNA-seq) measures thousands of genes in a separate cell
- How: 3 barcoding steps for sample, cell and RNA molecule
- Scale: bulk RNA-seq (5 samples) vs. scRNA-seq (45 K cells), a ~900 times bigger gene count matrix

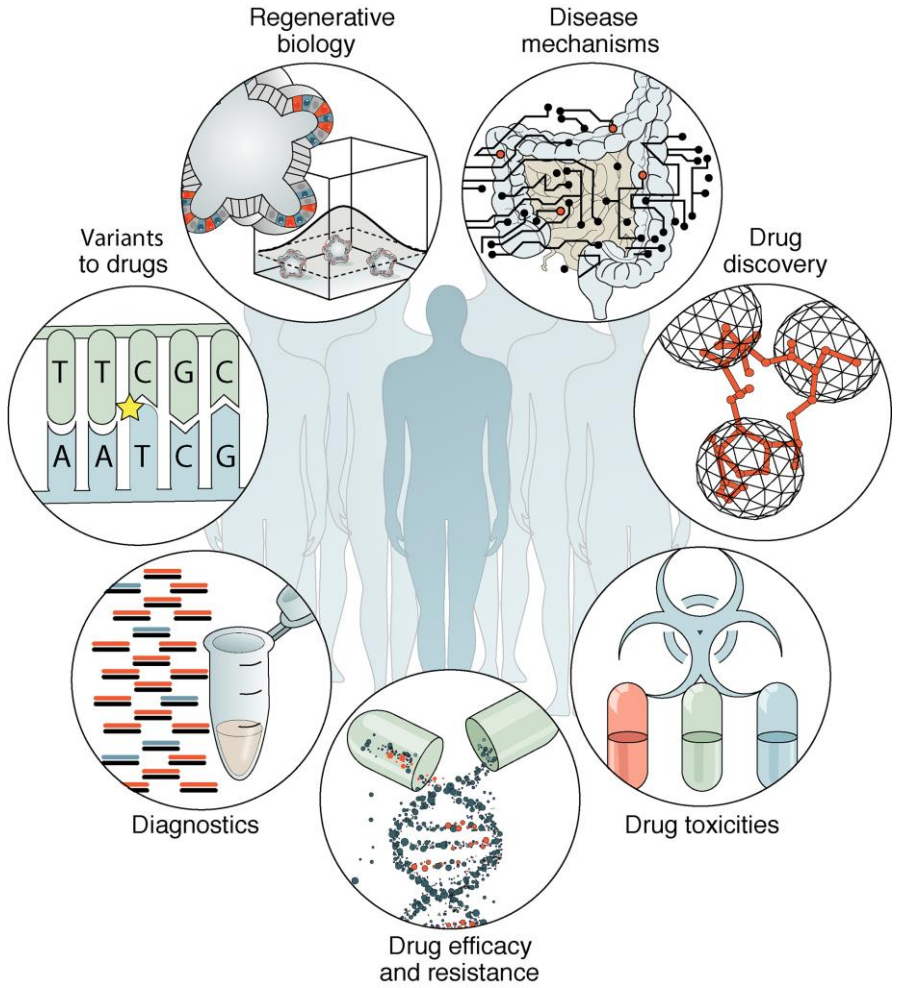
Single cell informatics

Scale

Resolution

INFORMATICS

Precision Genomics Medicine

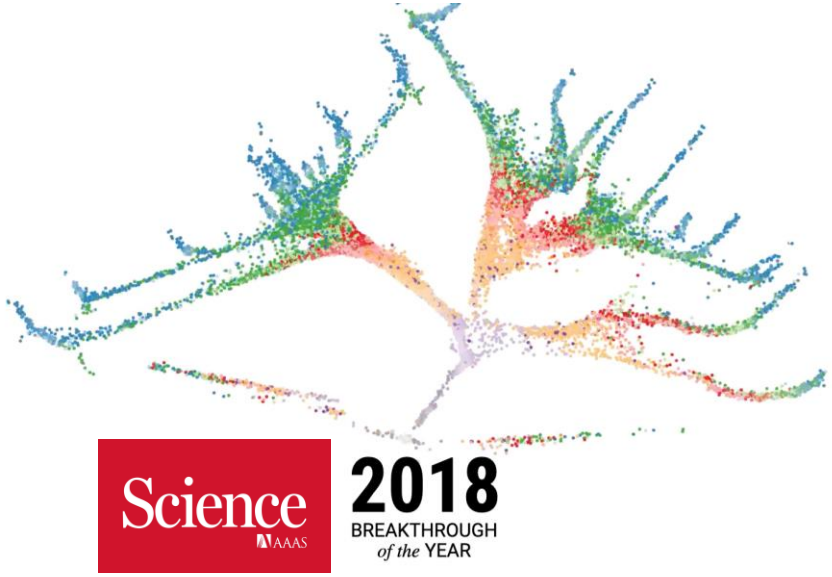


The Genomics and Machine Learning Team

Quan Nguyen, Onkar Mulay, Prakrithi Pavithra, Xiao Tan

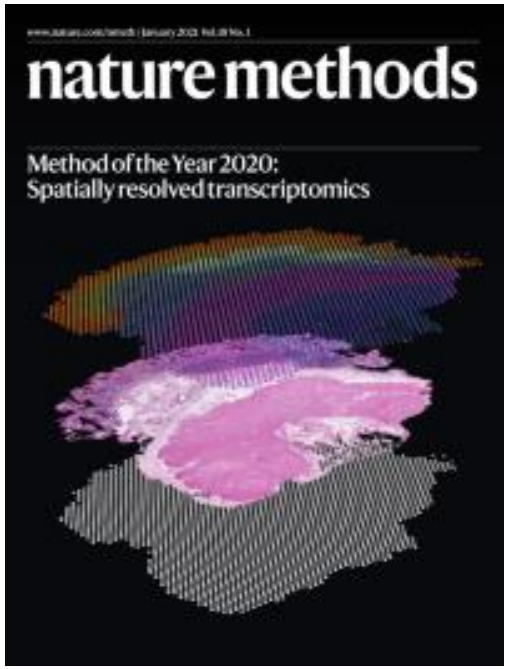
Advanced genomics technologies

2018: Single Cell Transcriptomics

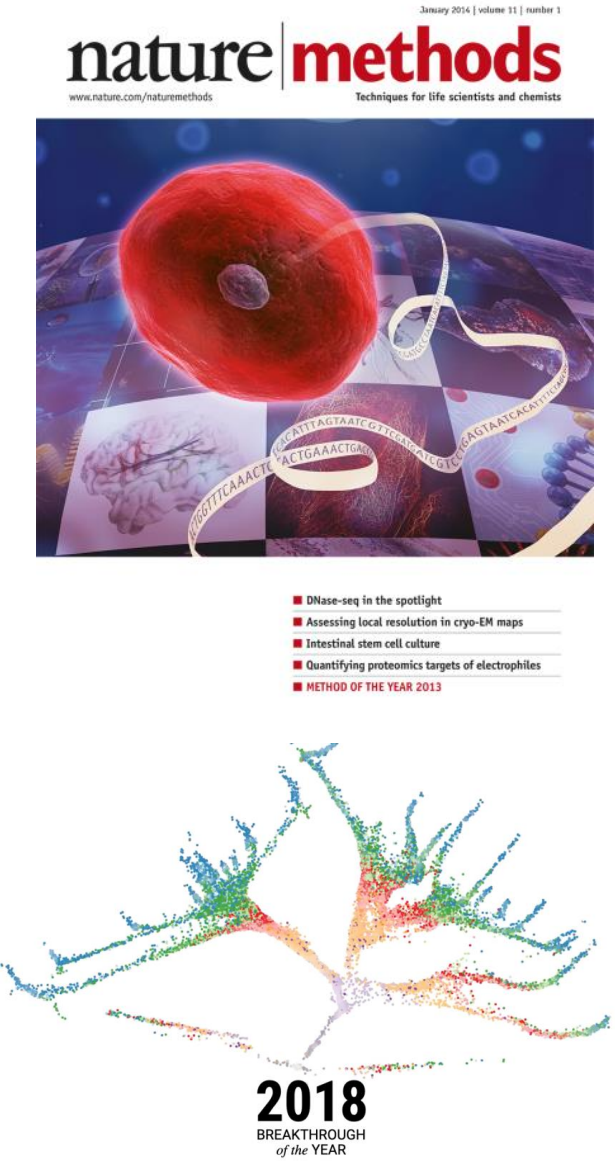


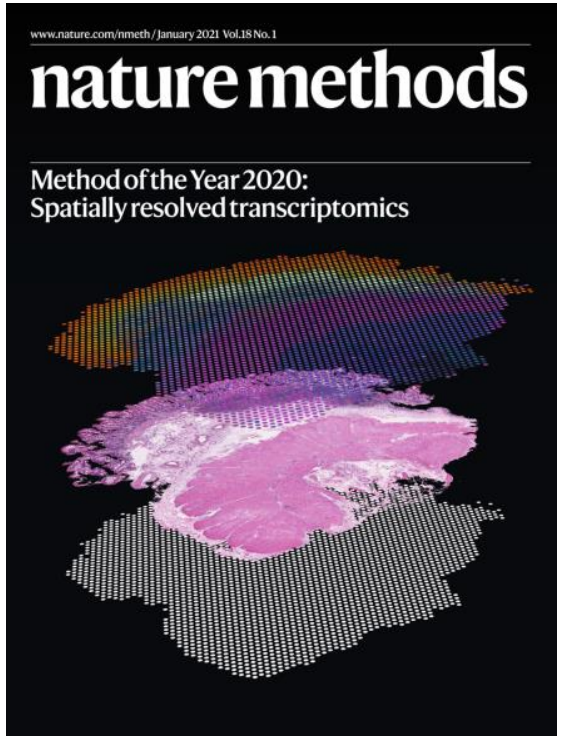
2019: Single Cell Multiomics

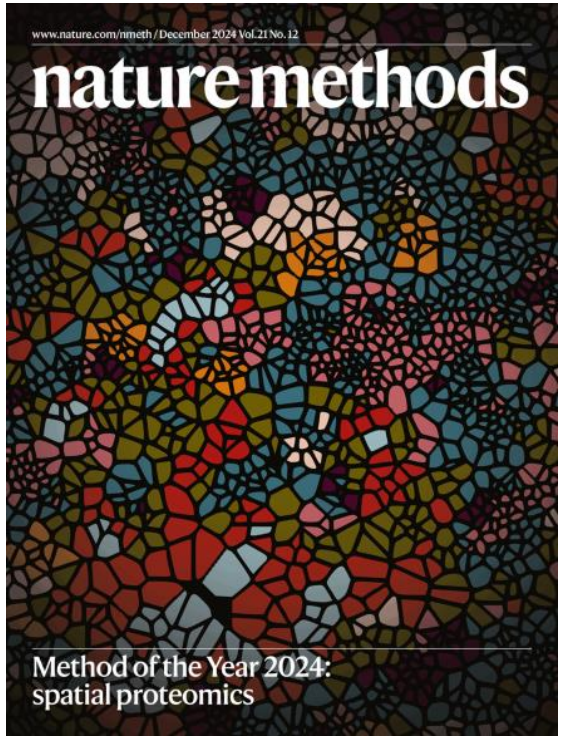
2020: Spatial Transcriptomics



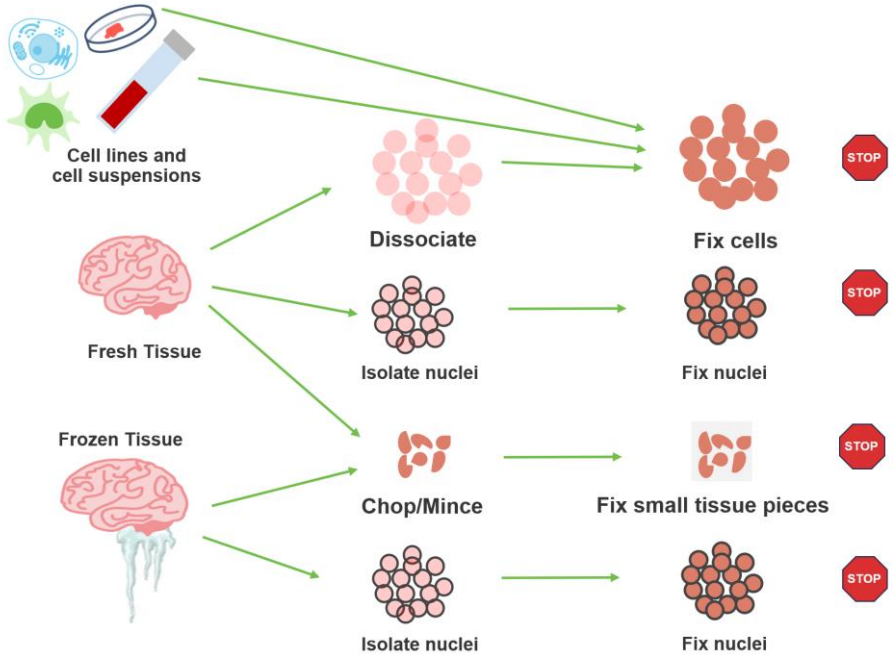
Single cell resolution and spatial context

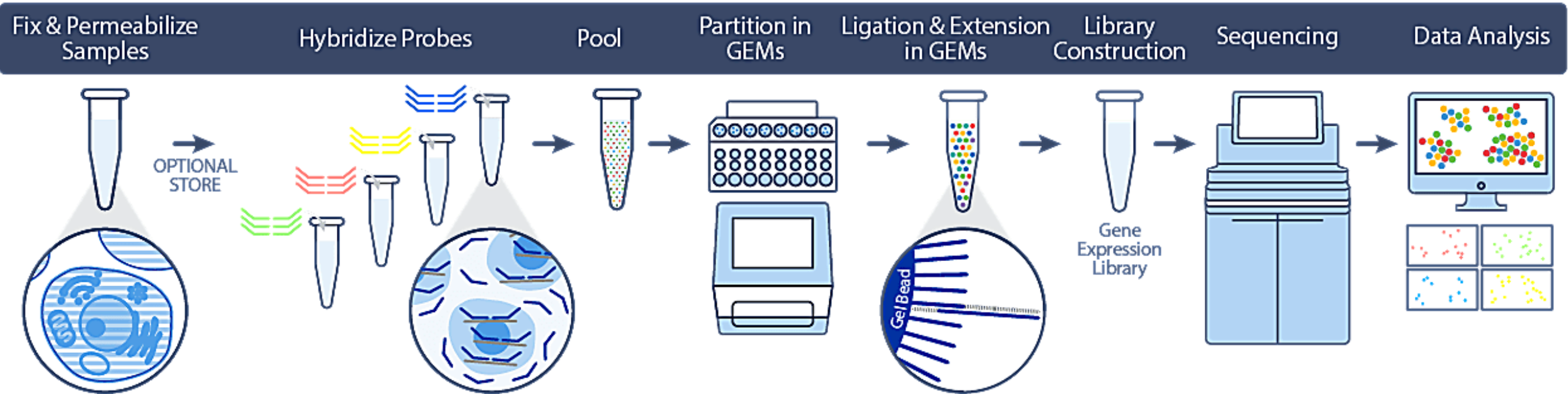




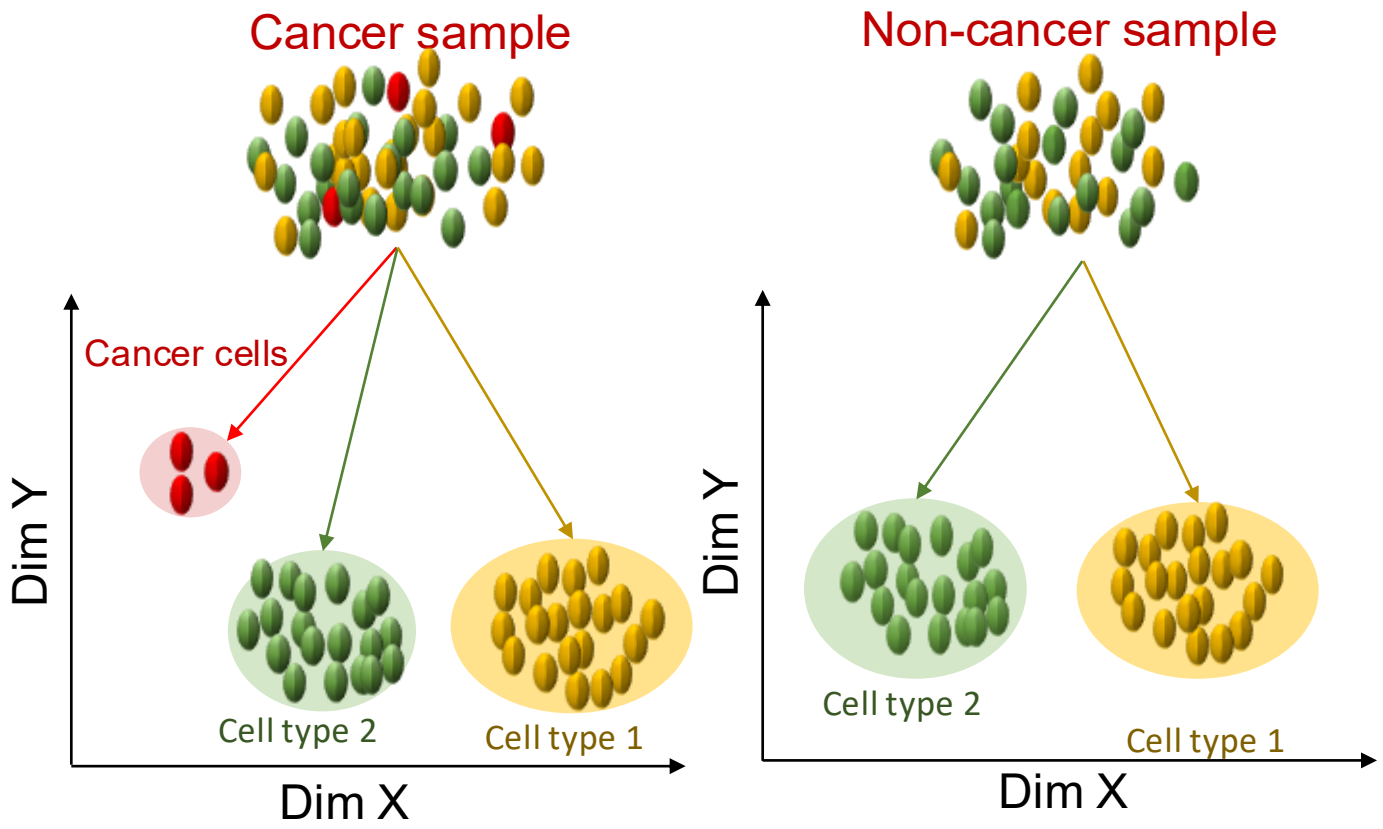


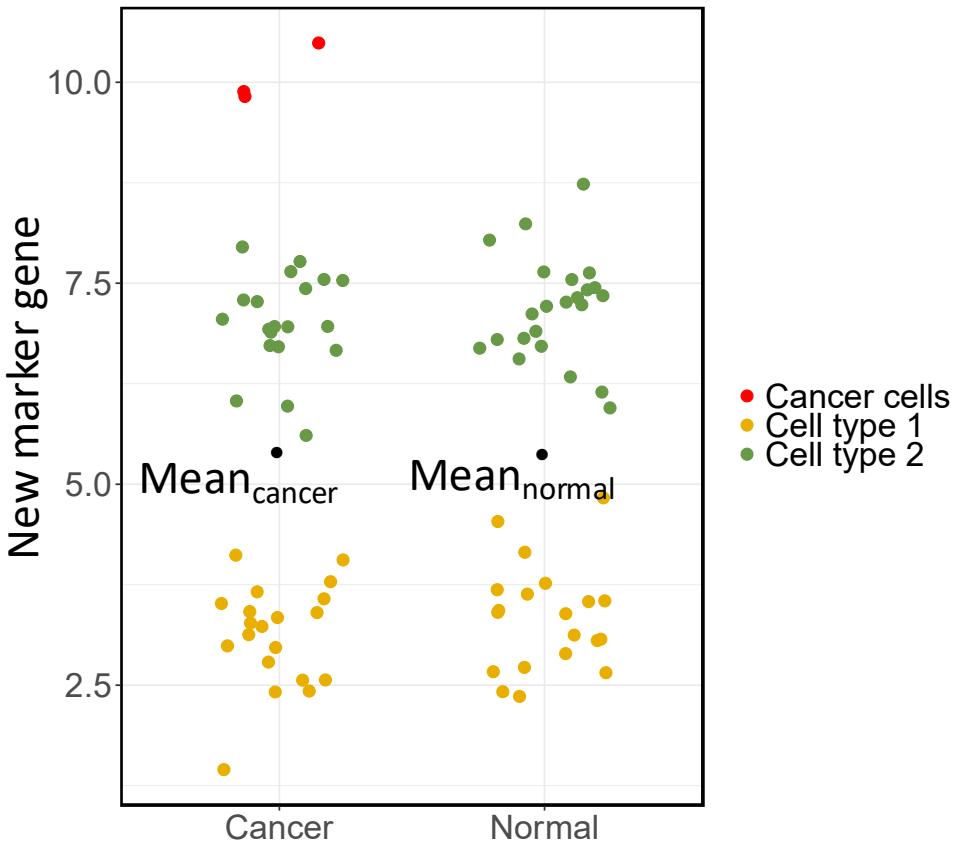
Increase single cell experiments to millions of cells





Disease at single-cell resolution

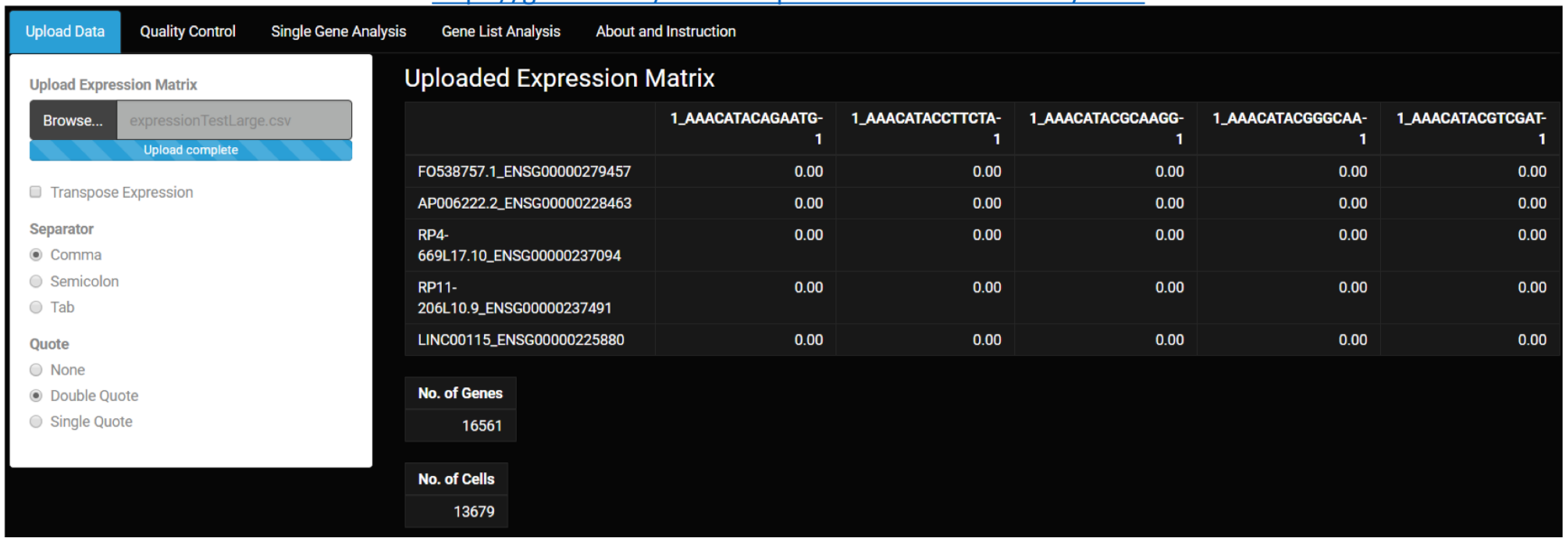




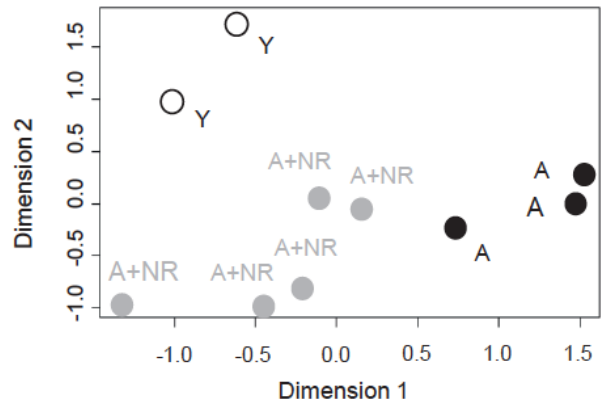
- Bulk RNA sequencing: no difference in mean expression
- Single-cell sequencing: can detect higher expression in cancer cells

Single cell data vs. bulk data

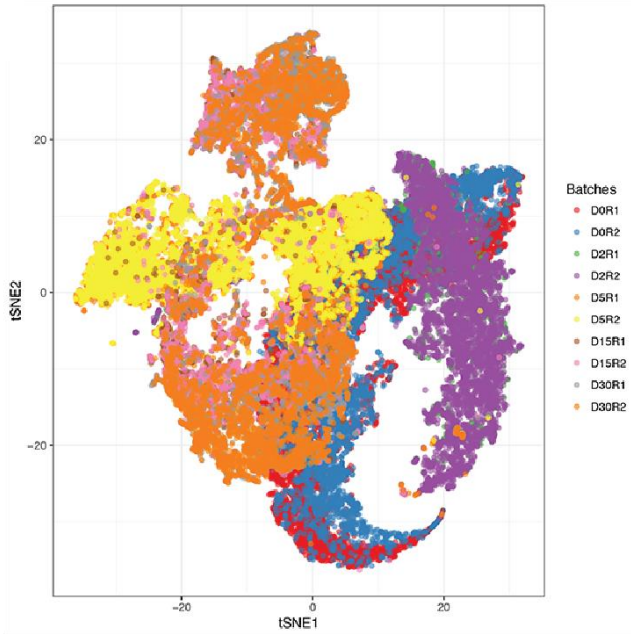
https://github.com/IMB-Computational-Genomics-Lab/scIVA



Bulk

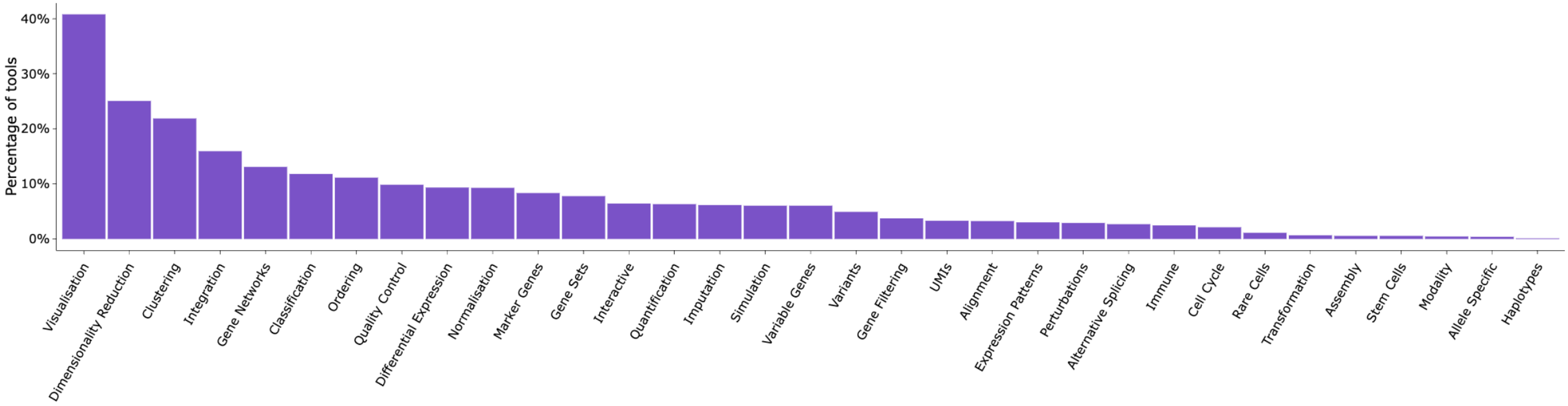


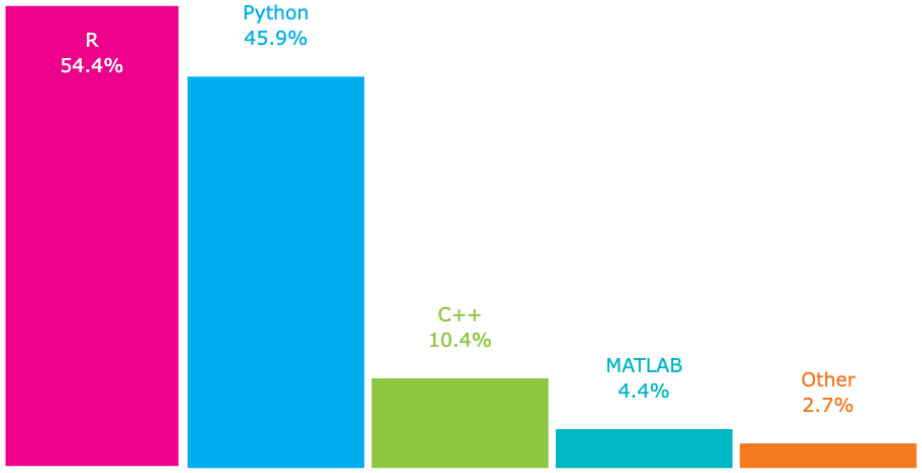
Single cell

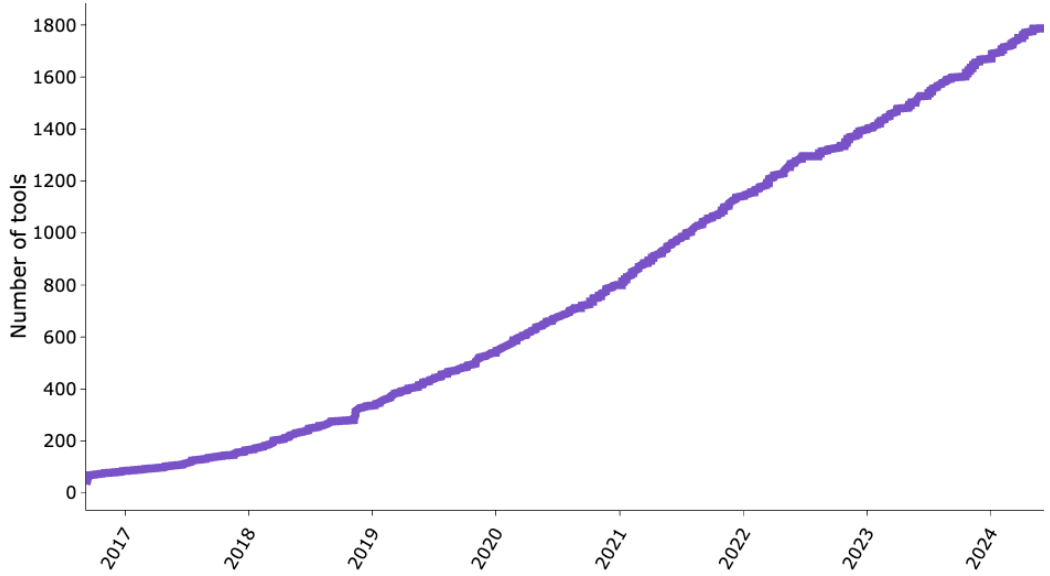


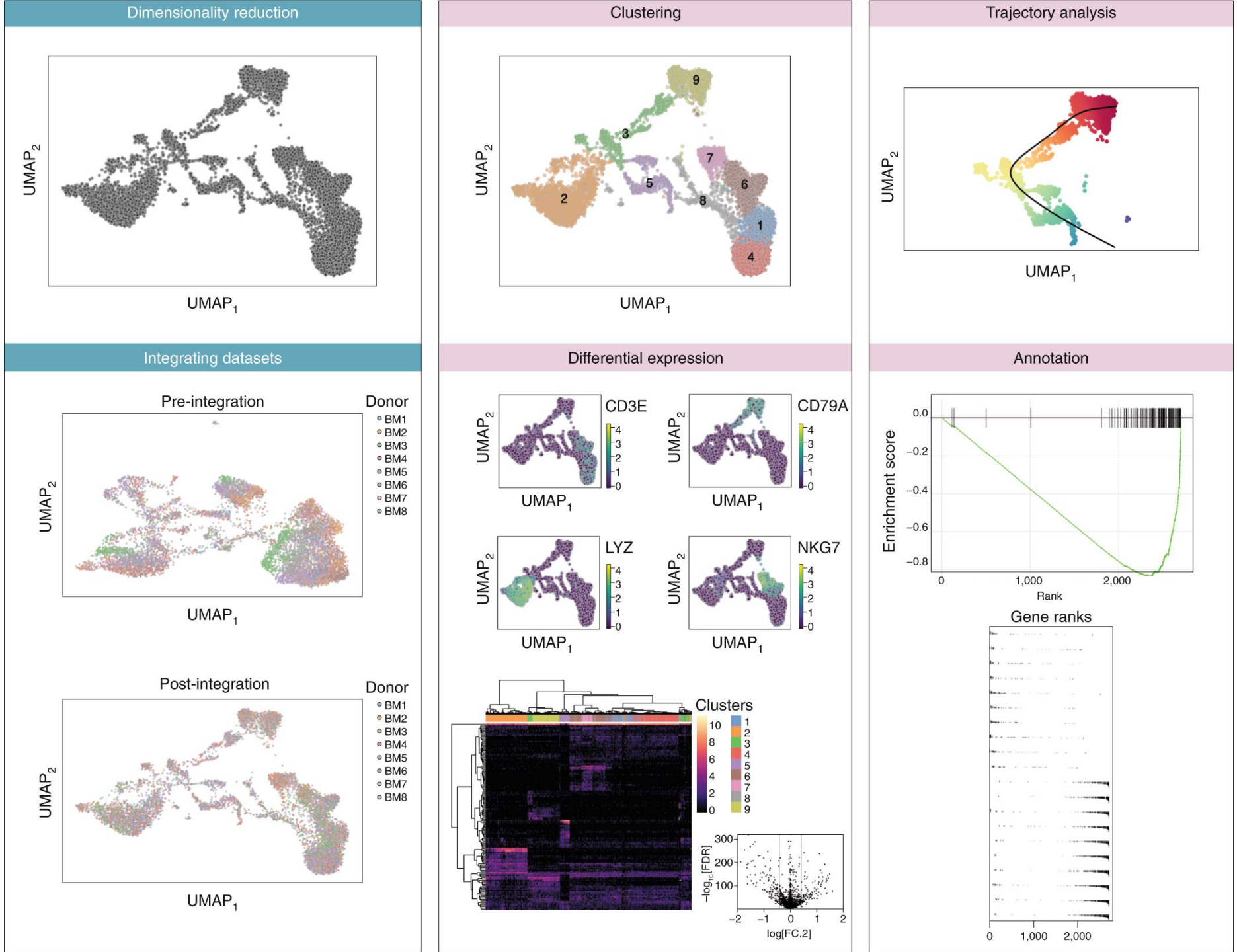
	Single cell	Bulk
Noisy data	Undetected genes (zero inflation)	Deep sequencing, most genes detected
Cell-cell variation	Measured	Not measured
Data size	Thousands of cells (1 cell ~ 1 bulk sample)	10-100 samples

Single cell data analysis types



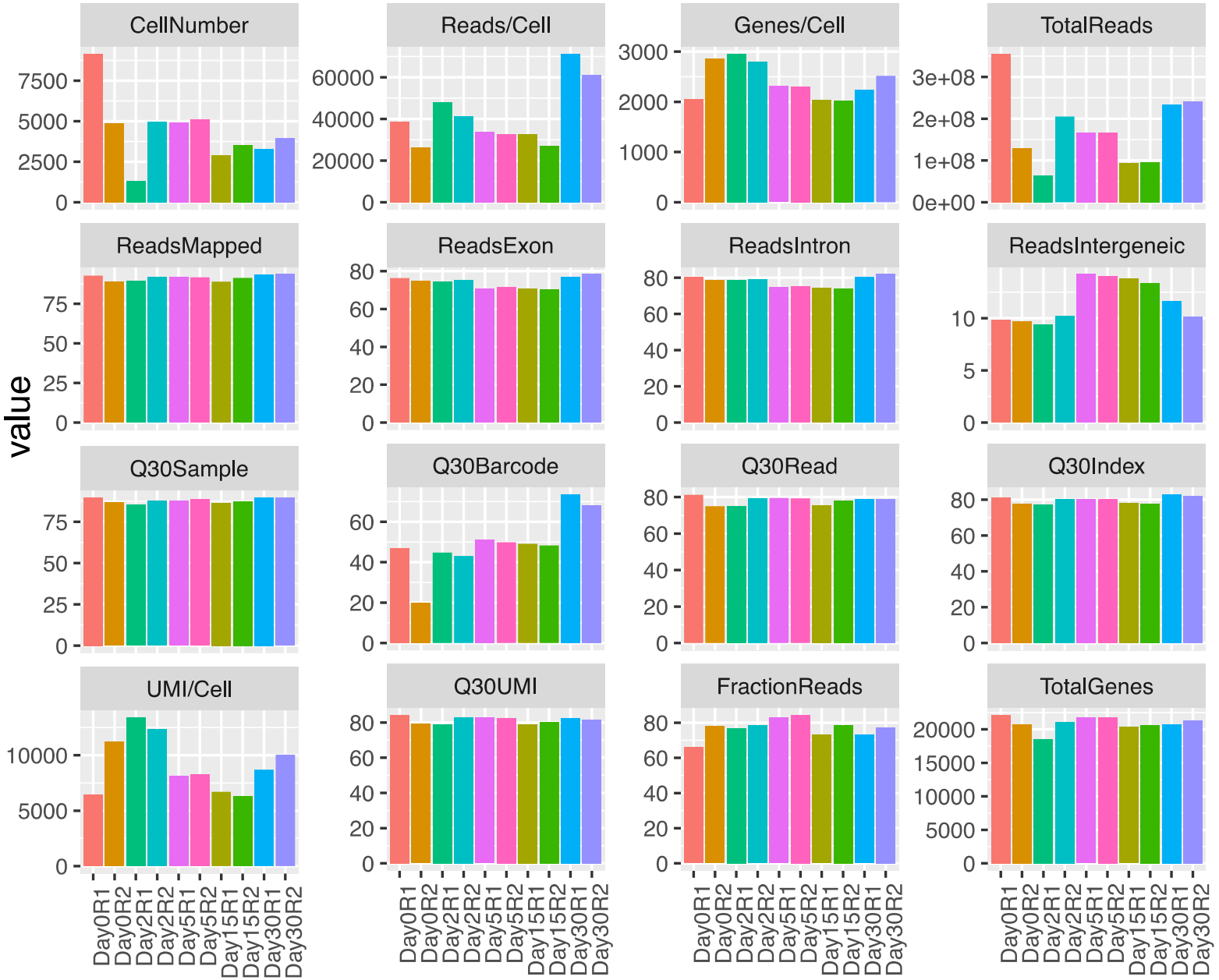





Dimensionality reduction
UMAP2
UMAP1
Clustering
9
3
7
5
6
2
8
1
4
Trajectory analysis
Integrating datasets
Pre-integration
Donor
BM1
BM2
BM3
BM4
BM5
BM6
BM7
BM8
Post-integration
Differential expression
CD3E
CD79A
LYZ
NKG7
Clusters
Annotation
Enrichment score
Rank
Gene ranks
−log10[FDR]
log[FC.2]

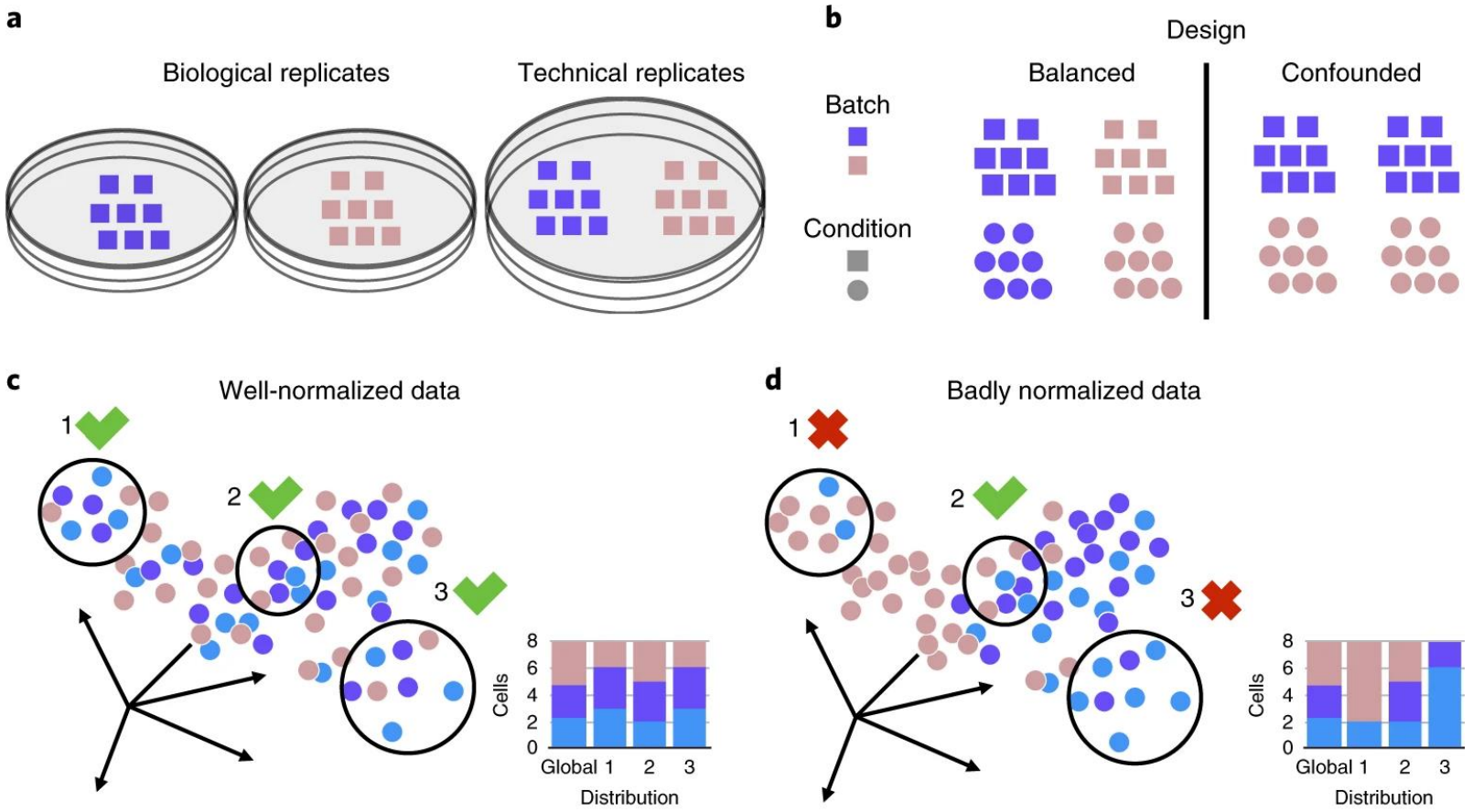
Data quality control: a range of QC measures

- 16 QC measures
- 10 scRNA-seq libraries



Why Data Normalisation

- Batch effects: technical differences induced by the operator or other experimental artifacts
- Often observe systematic differences in sequencing coverage between libraries (or cells)
- Normalization aims to remove these differences
- Such that they do not interfere with comparisons of the expression profiles between cells
- Ensure heterogeneity or differential expression within the cell population are driven by biology and not technical biases.

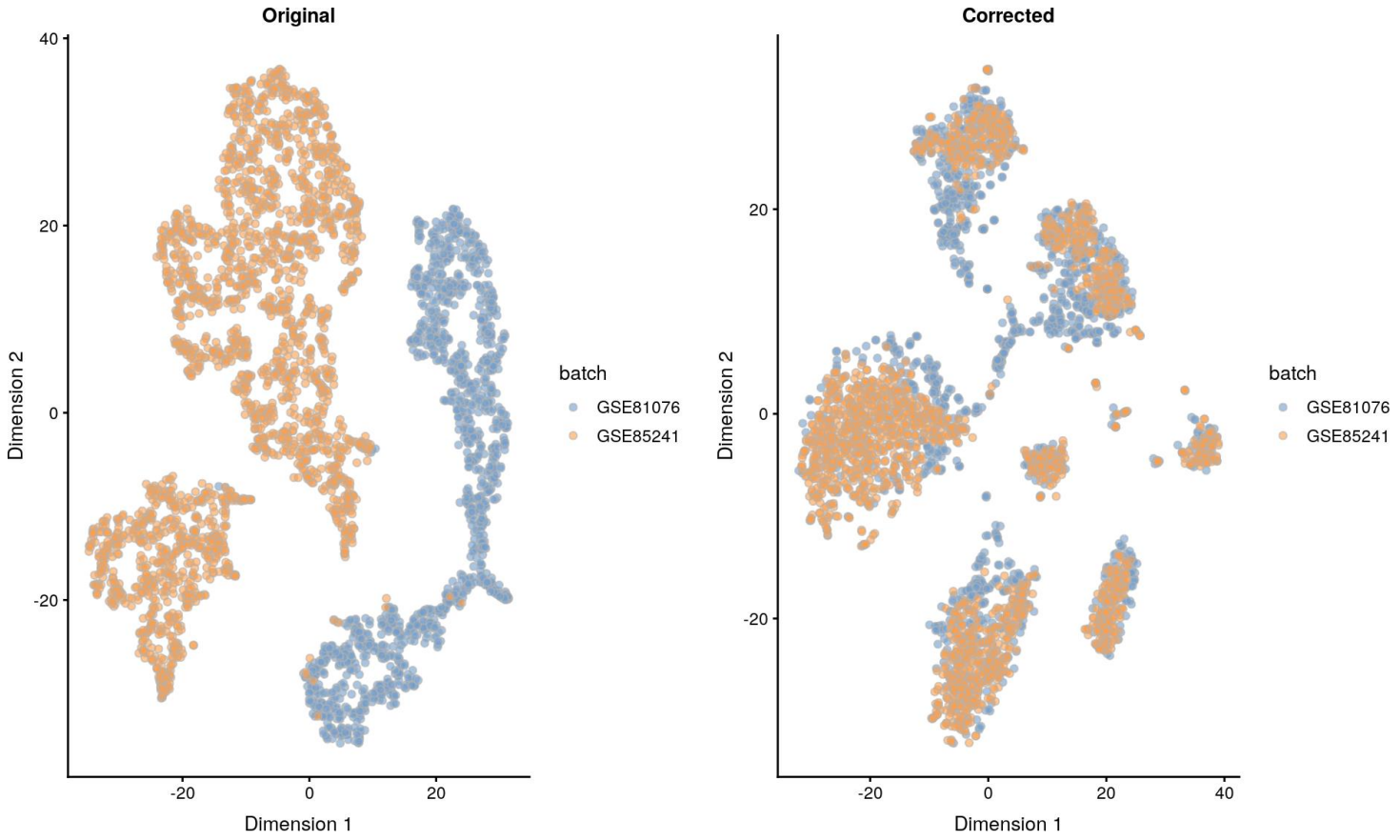


(Buttner et al, 2019)

Three levels of single cell data normalization

Three levels of technical variation in scRNA-seq data:

- Gene-specific effects within a cell: GC content, gene length
- Cell specific effects within a sample: each cell is amplified separately, causing amplification bias among cells
- Batch effects within a study: sample preparation or technology-specific effects



Cell to cell normalization: Library size normalization

	Cell1	Cell2	Cell3	Cell4	Cell5
gene1	0	0	0	0	0
gene2	0	0	0	0	0
gene3	3	0	1	0	1
gene4	0	1	3	3	0
gene5	1	4	2	1	2
colsum / library size	4	5	6	4	3
factor	0.91	1.14	1.36	0.91	0.68
Normalized library size	4.40	4.39	4.41	4.40	4.41

Total library size = 22

Nrcells = 5

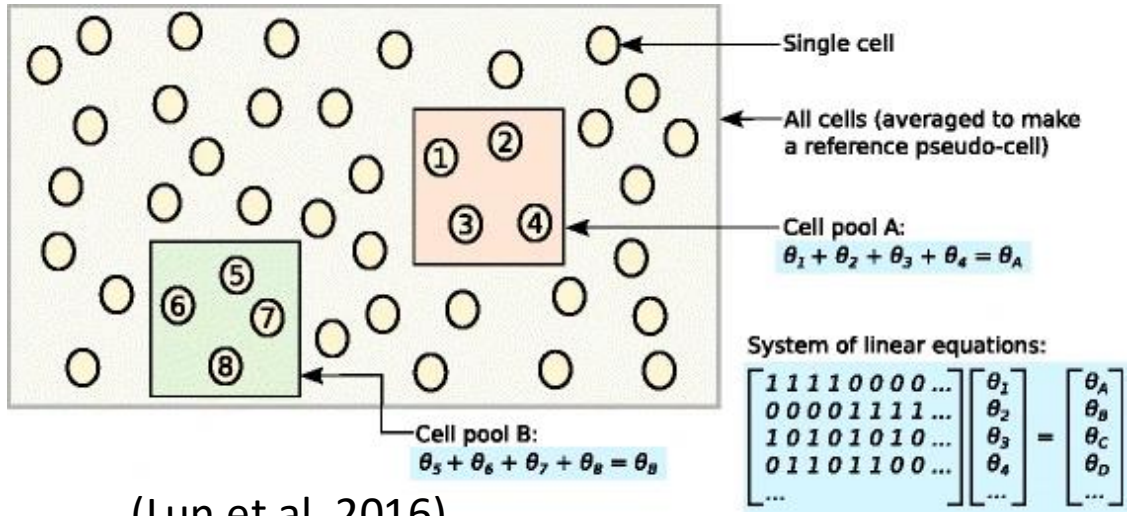
$$\text{Size factor} = \frac{\text{library size} * \text{nrCells}}{\text{Total library size}}$$

The mean size factor across all cells is equal to 1

Normalized expression values are on the same scale as the original counts,

Useful for interpretation especially when dealing with transformed data

Cell to cell normalisation: a pooling strategy to solve zero inflation



(Lun et al, 2016)

	Pool A		Pool B		Sum(poolA)	Sum(PoolB)	average	Sum(poolA)/ average	Sum(poolB)/ average
	Cell 1	Cell 2	Cell 3	Cell 4					
g1	0	0	0	0	0	0	0	0	0
g2	0	0	0	0	0	0	0	0	0
g3	3	0	1	0	**3**	1	4/4	3/4/4	1/4/4
g4	0	1	3	3	1	6	7/4	1/7/4	6/7/4
g5	1	4	2	1	5	3	8/4	5/8/4	3/8/4

→ A demo

θ_A θ_B

$= \theta_{cell1} + \theta_{cell2}$

Scaling factor of cell 1

$$E(V_{ik}) = \lambda_{i0} \sum_{\mathrm{j} \epsilon \mathrm{S_k}} \theta_j \times t_j^{-1}$$

V_{ik} is the sum of adjusted expression value across all cells in pool V_k for gene *i*

λ_{i0} is the expected transcript count and θ_j is the cell specific bias

S_k is a pool of cell; $\theta_j \times t_j^{-1}$ is size factor for cell *j*

- Each cell is considered as a sequencing library, so the total reads per cell need to be normalised
- Pool cells to reduce the number of zeros
- Estimate the size factors for the pool
- Repeat many time and use deconvolution to estimate each cell size factor θ_j

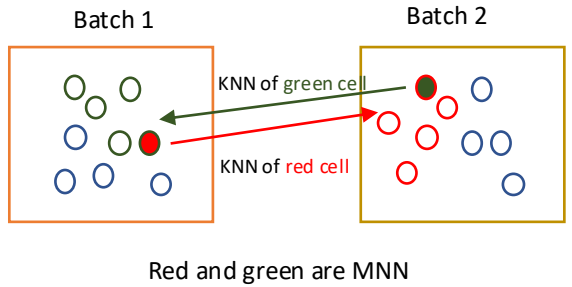
Batch normalisation: Mutual nearest neighbour (MNN)

Three assumptions in MNN normalisation:

(i) there is at least one cell population that is present in both batches,

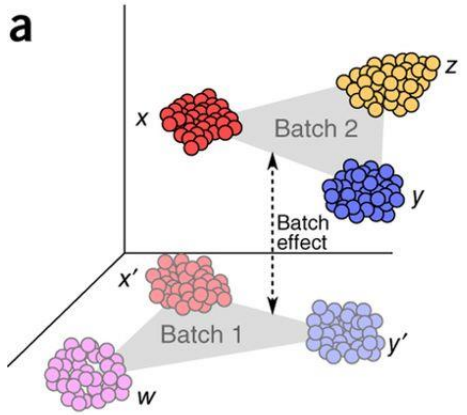
(ii) the batch effect is almost orthogonal to the biological subspace, and

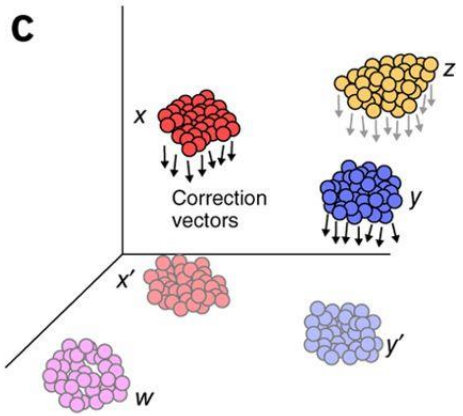
(iii) the batch-effect variation is much smaller than the biological-effect variation between different cell types



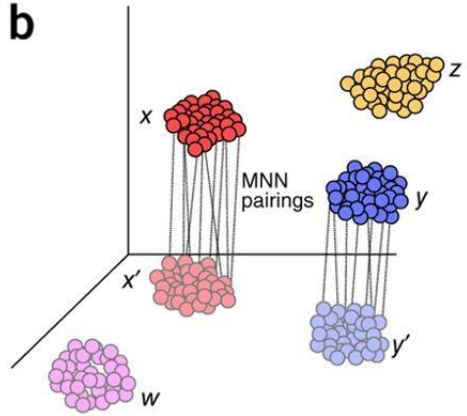
Red and green are MNN

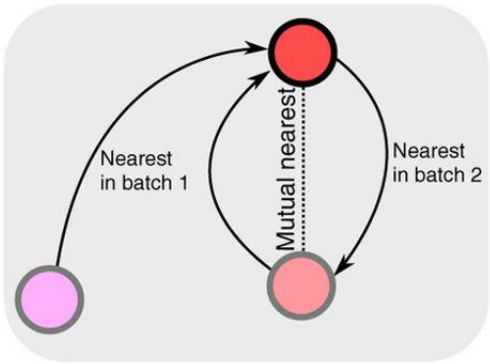
Find KNN in anther Batch





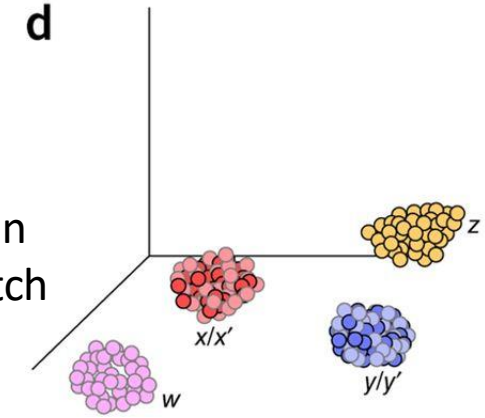
assume batch effects are mostly orthogonal to the biological manifold:
← batch effect: vertical
← biological manifold: horizontal

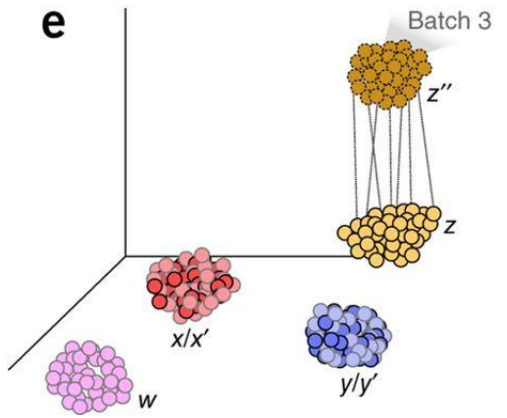




the cosine normalization

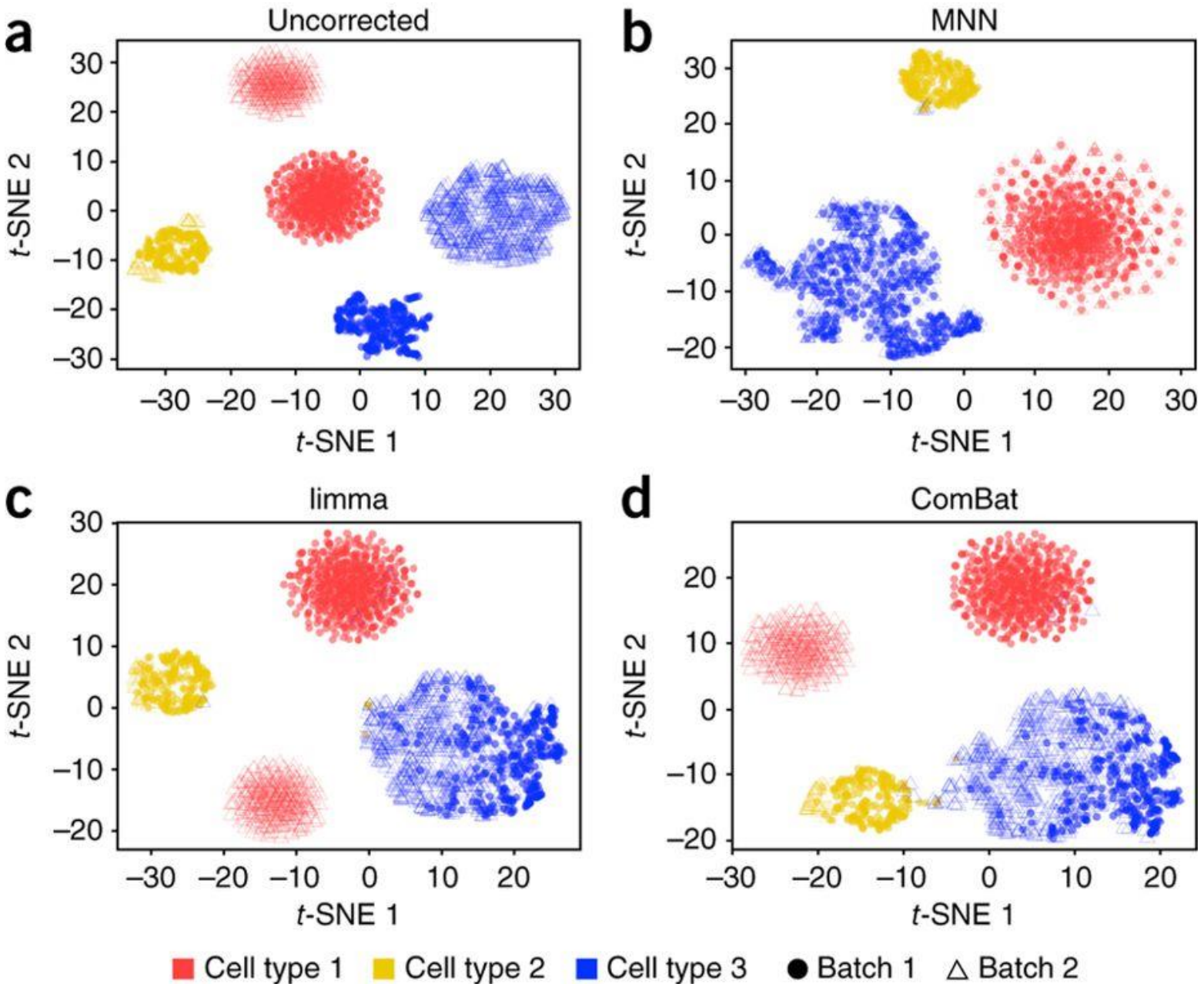
$$Y_x \leftarrow \frac{Y_x}{\|Y_x\|}$$





(Haghverdi et al, 2018)

Batch normalisation: Mutual Nearest Neighbour (MNN)



Dimensionality reduction: linear techniques

Why dimensionality reduction:

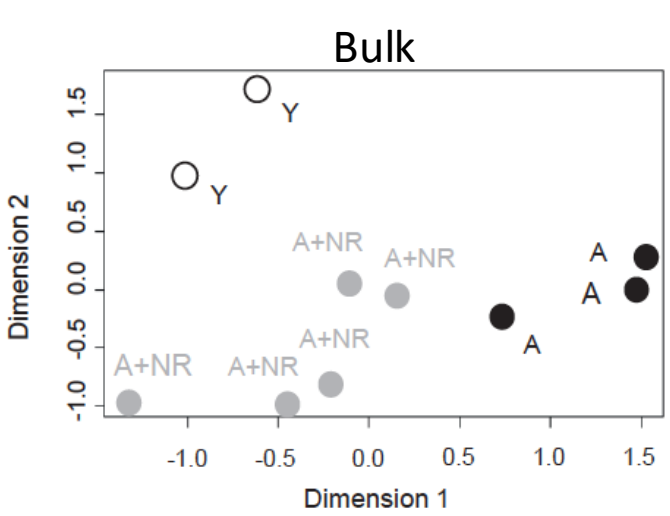
- Filters out noise
- Minimises curse of dimensionality
- Allows visualization with more separation of points
- Reduces computational load

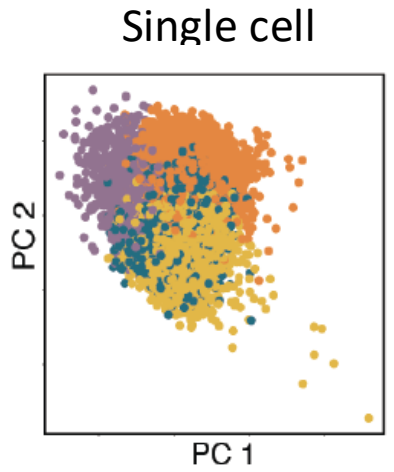
Linear approaches:

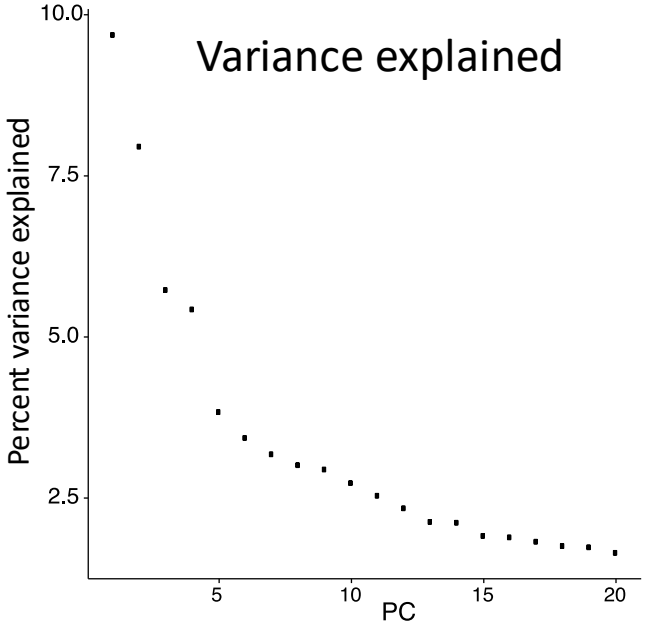
- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- NMF (Non-negative Matrix Factorization)

Linear approaches:

- Capture the dimensions with higher variance
- Quantitative way to assess the amount of retained dimensions
- Preserve both long-range and short-range distance (i.e. cells that are very different or very similar)
- Different to bulk RNAseq data, the first few dimensions are not enough to capture scRNAseq data structure well

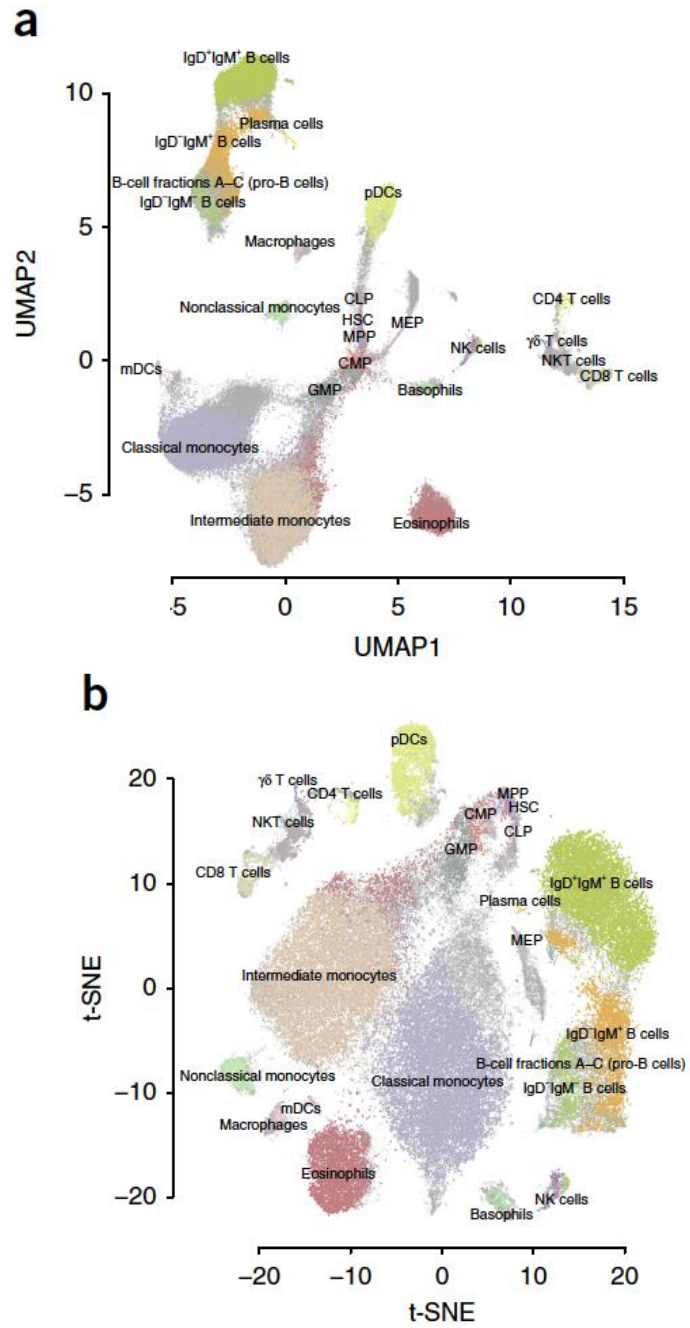




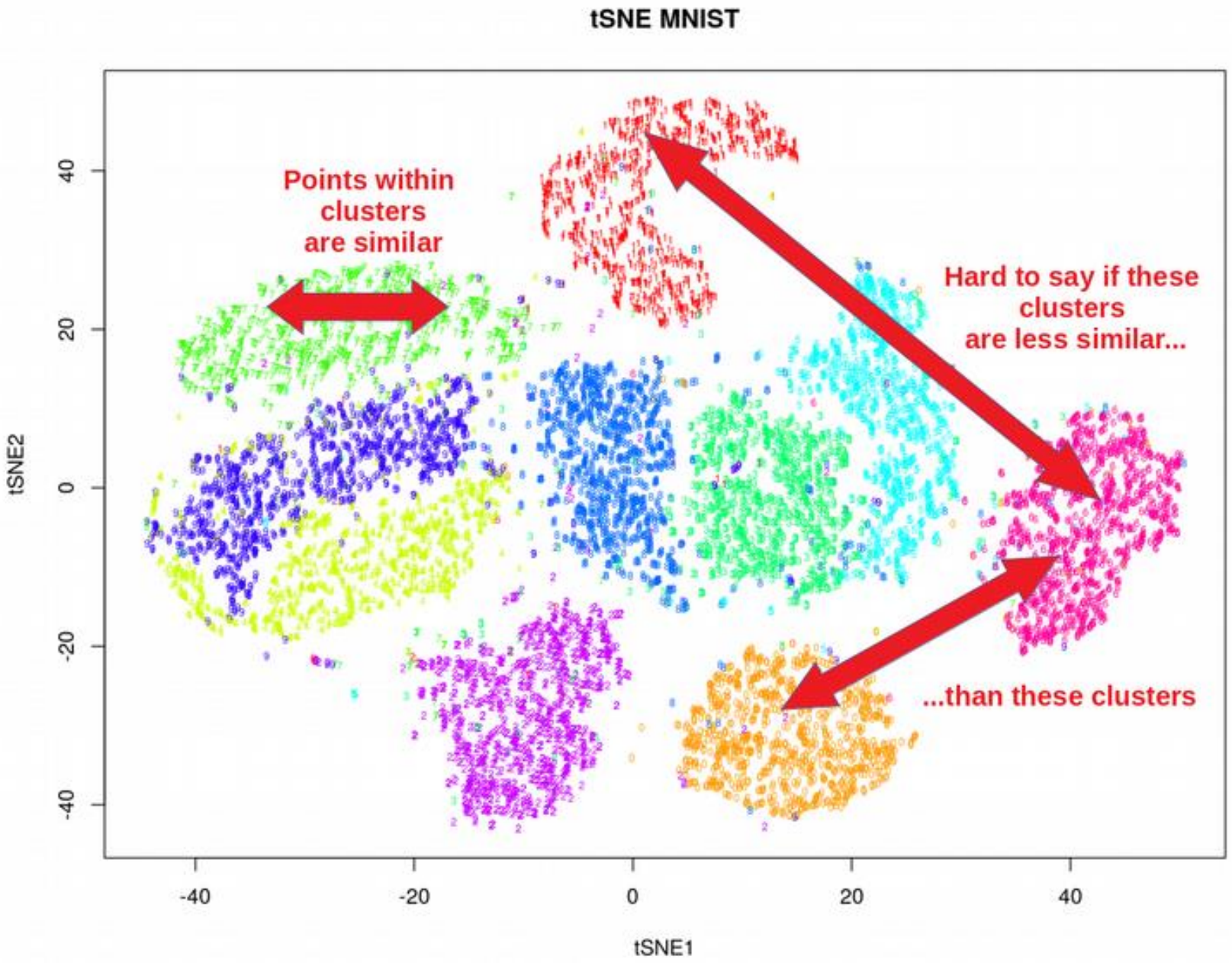


Dimensionality reduction: nonlinear techniques

- MDS (Multidimensional Scaling)
- Uniform manifold approximation and projection (UMAP)
- t-distributed Stochastic Neighbour Embedding (t-SNE)
- UMAP and tSNE: nonlinear embedding (mapping) of data points from high dimensional space to low dimensional space, so that the probability distance between these two space (KL diverge · cross entropy) is minimised
- Both methods: class of k-neighbour based graph learning algorithms, strong influence of hyperparameters, non-deterministic (stochastic)
- Nonlinear techniques solve the overcrowding representation, which is often seen in linear approaches for large scRNA-seq data
- UMAP preserves local & more of the global data structure than t-SNE



Global vs local distance in low dimensional space

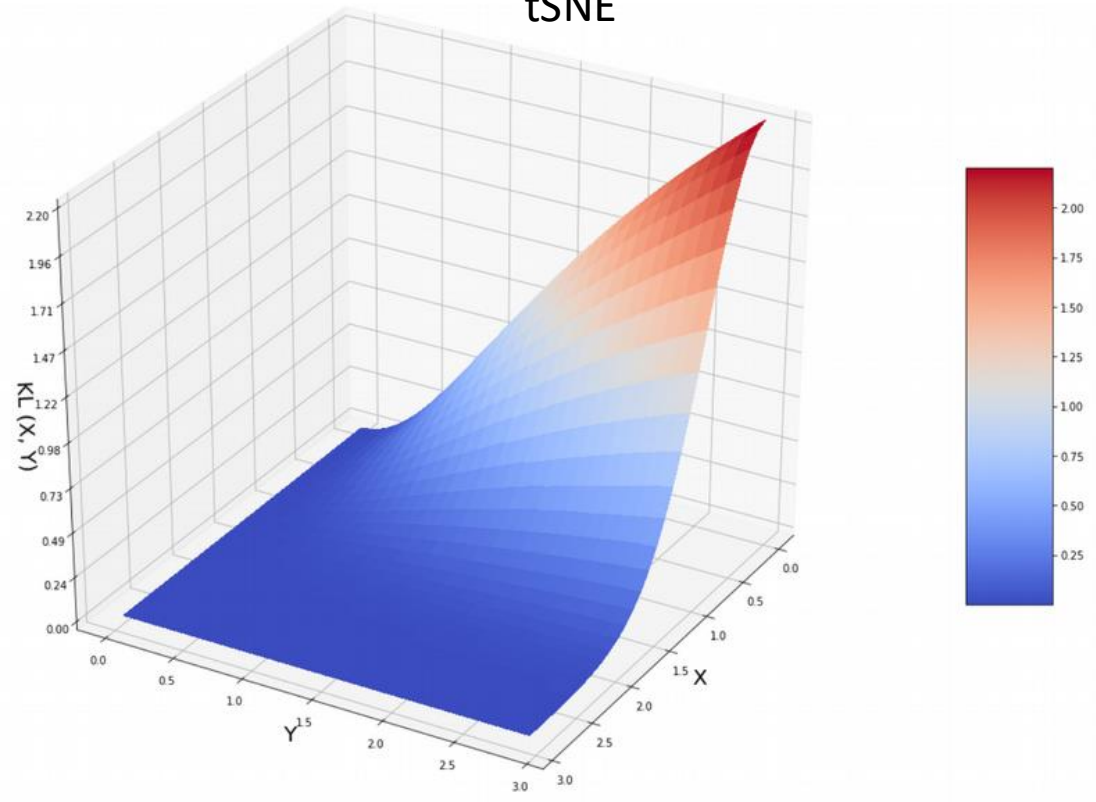


(Oskolkov N, 2019)

tSNE does not preserve long distance - KL divergence

(Oskolkov N, 2019)

tSNE



tSNE minimises Kullback-Leiber divergence *KL(X,Y)*

$$KL(X,Y) \approx -P(X)\log Q(Y) = e^{-X^2}\log\left(1+Y^2\right)$$

- The embedding minimizes the Kullback-Leiber divergence of the distribution from Q to P calculated as: $KL(X,Y) = \sum_{i\neq j} p_{ij}\log\frac{p_{ij}}{q_{ij}} \approx e^{-X^2}\log(1+Y^2)$
- The probability distance between two neighbouring cells is the joint probabilities $p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}$
- Conditional probability of cell C_j given cell C_i is calculated as:

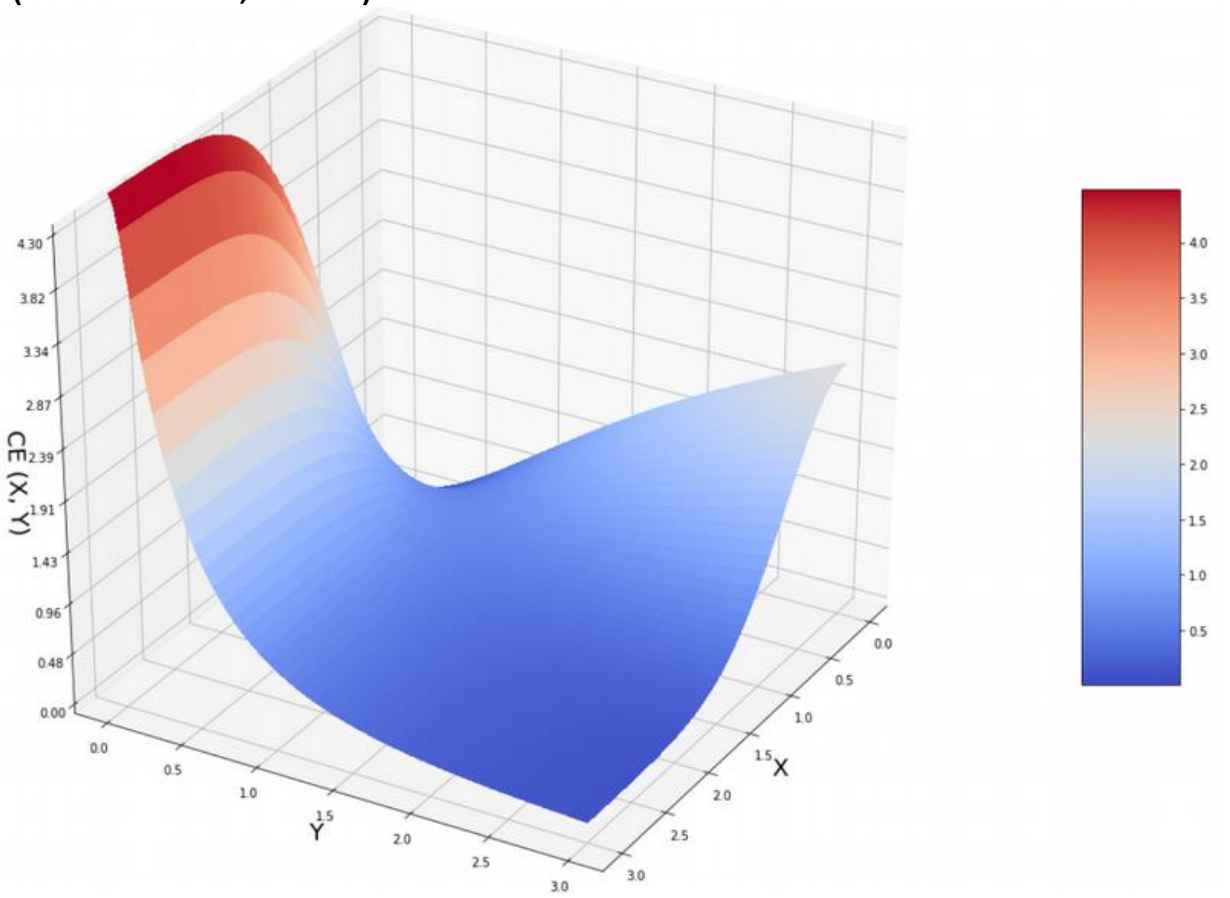
$$p_{j|i} = \frac{exp\left(\frac{-d\left(C_i,C_j\right)^2}{2\sigma_i^2}\right)}{\Sigma_{k\neq i}\ exp\left(\frac{-d(C_i,C_k)^2}{2\sigma_i^2}\right)}$$

- For large distances X in high dimensions, the exponential term approaching 0, **so Y can be basically any value from 0 to ∞ and KL remains small**
- For small X, to minimise KL (cost/penalty), Y is small
- Pairwise similarity in t-SNE space: $q_{ij} = \frac{\left(1+\|y_i - y_j\|^2\right)^{-1}}{\sum_{k\neq m}(1+\|y_k - y_m\|^2)^{-1}}$, y_i and y_j are corresponding mapped points of cells C_i and C_j to t-SNE space, and **q_{ij} follows *t* distribution to avoid crowding**

UMAP preserves long distance - cross entropy

(Oskolkov N, 2019)

UMAP



$$X \to 0 : CE(X,Y) \approx \log\left(1+Y^2\right)$$

When X small, Y is also approaching 0 to minimize CE

$$X \to \infty : CE(X,Y) \approx \log\left(\frac{1+Y^2}{Y^2}\right)$$

When X large, Y is also large to minimize CE

UMAP minimises cross entropy *CE(X,Y)*

$$CE(X,Y) = P(X)\log\left(\frac{P(X)}{Q(Y)}\right) + (1-P(X))\log\left(\frac{1-P(X)}{1-Q(Y)}\right)$$

$$\approx e^{-X^2}\log\left(1+Y^2\right) + \left(1-e^{-X^2}\right)\log\left(\frac{1+Y^2}{Y^2}\right)$$

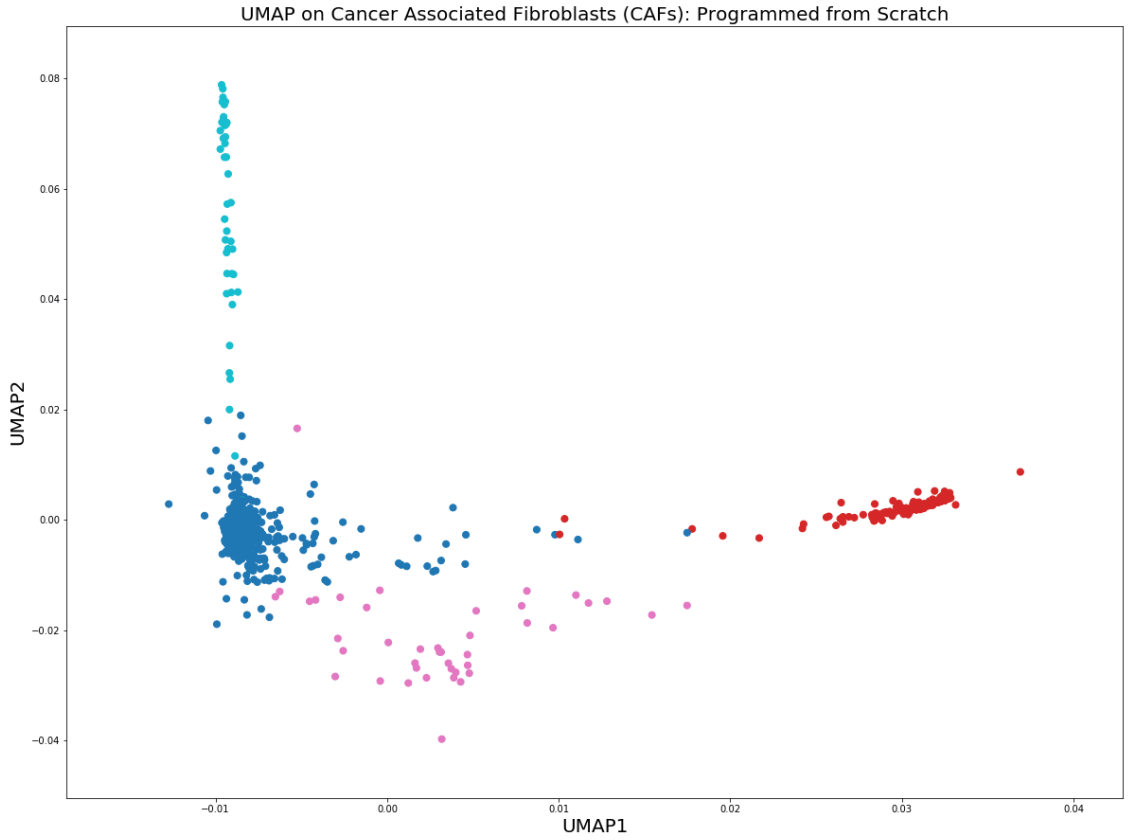
tSNE: $KL(X,Y) \approx -P(X)\log Q(Y) = e^{-X^2}\log\left(1+Y^2\right)$

More about UMAP vs tSNE

- To learn low-dimensional embeddings, UMAP assigns initial low-dimensional coordinates using **Graph Laplacian** (force directed graph layout algorithm) in contrast to **random normal initialization** used by tSNE. Therefore, UMAP is less dependent on random state (not changing from run to run)

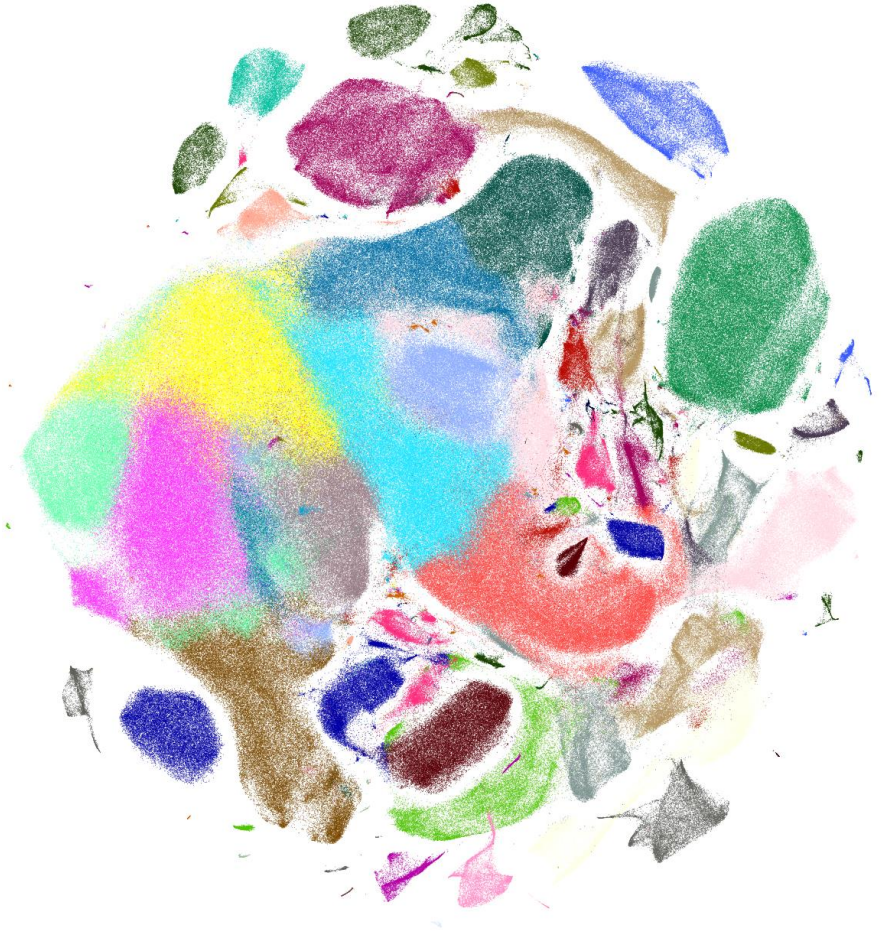
- UMAP proceeds by iteratively applying attractive (among edges) and repulsive forces (among vertices) at each edge or vertex. Convergence is guaranteed by slowly decreasing the attractive and repulsive forces of the neighbour graph.

- UMAP has no computational restrictions on embedding dimension, making it viable as a general-purpose dimension reduction technique for machine learning (tSNE can only embed to 2-3 dimensions)



(Oskolkov N, 2019)

Single Cell Clustering Analysis



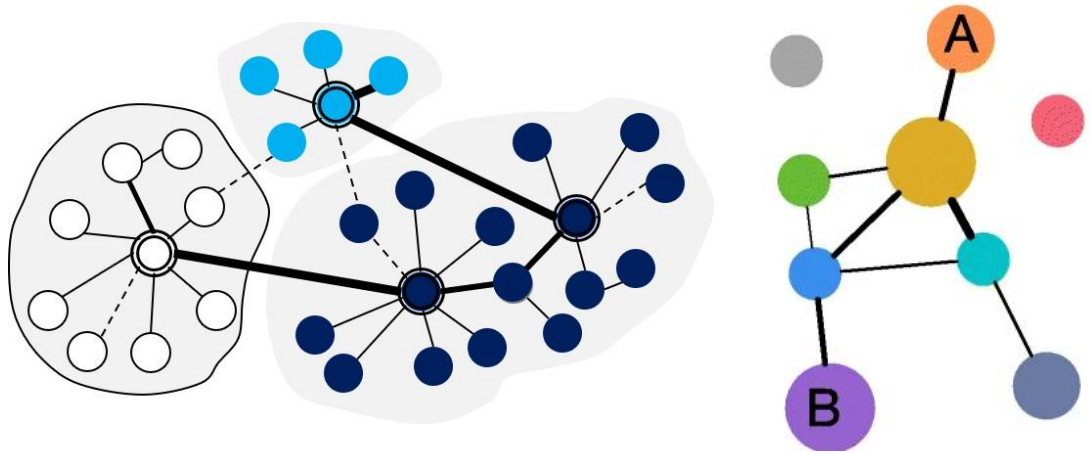
Clustering in scRNAseq is a data-driven way to find cell (sub)types at single-cell resolution

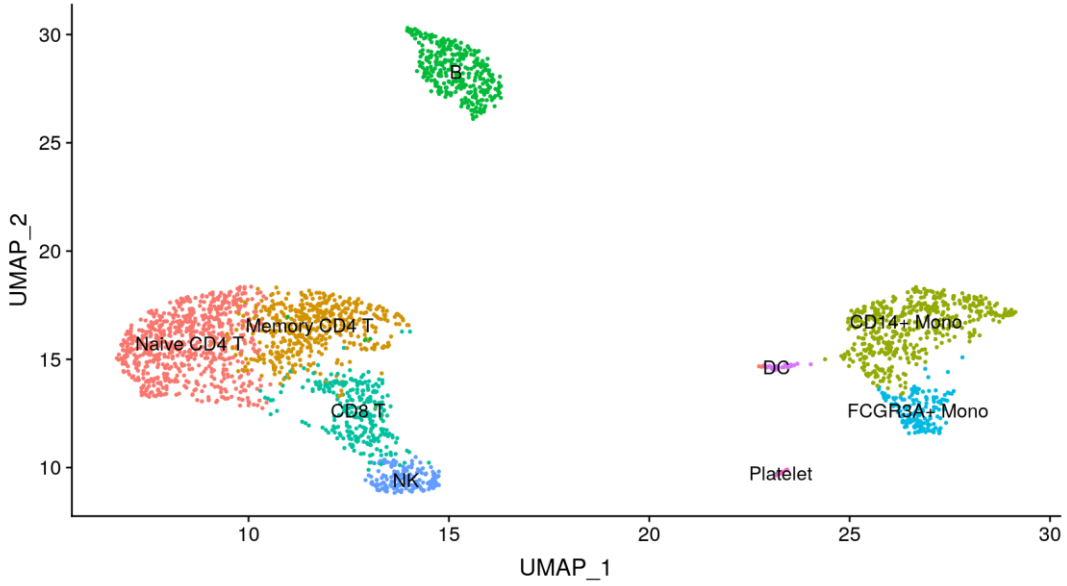
Graph-based Clustering

Two main steps:

1) Embed cells in a graph structure:
   - K-nearest neighbour (KNN) graph (cells with similar expression patterns identified by Euclidean distance in PCA space)
   - Edge weights between any two cells based on the shared overlap in their local neighbourhoods (Jaccard similarity)

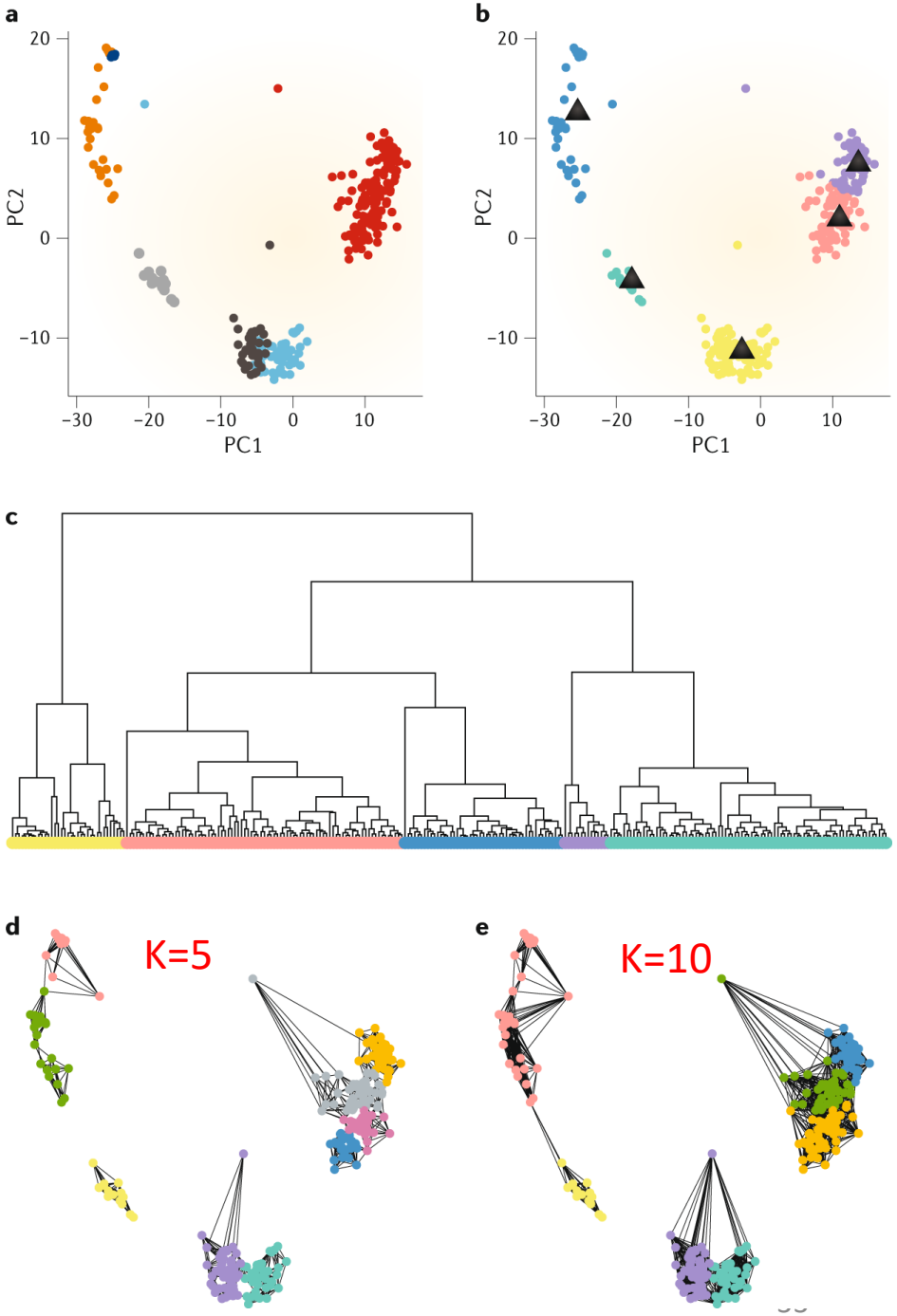
2) Community detection to partition cells in graph into groups of cells
   - Modularity optimization techniques such as the Louvain algorithm
   - Modularity: measures the density of edges inside communities to edges outside communities
   - Louvain iteratively groups cells together, with the goal of optimizing the standard modularity function





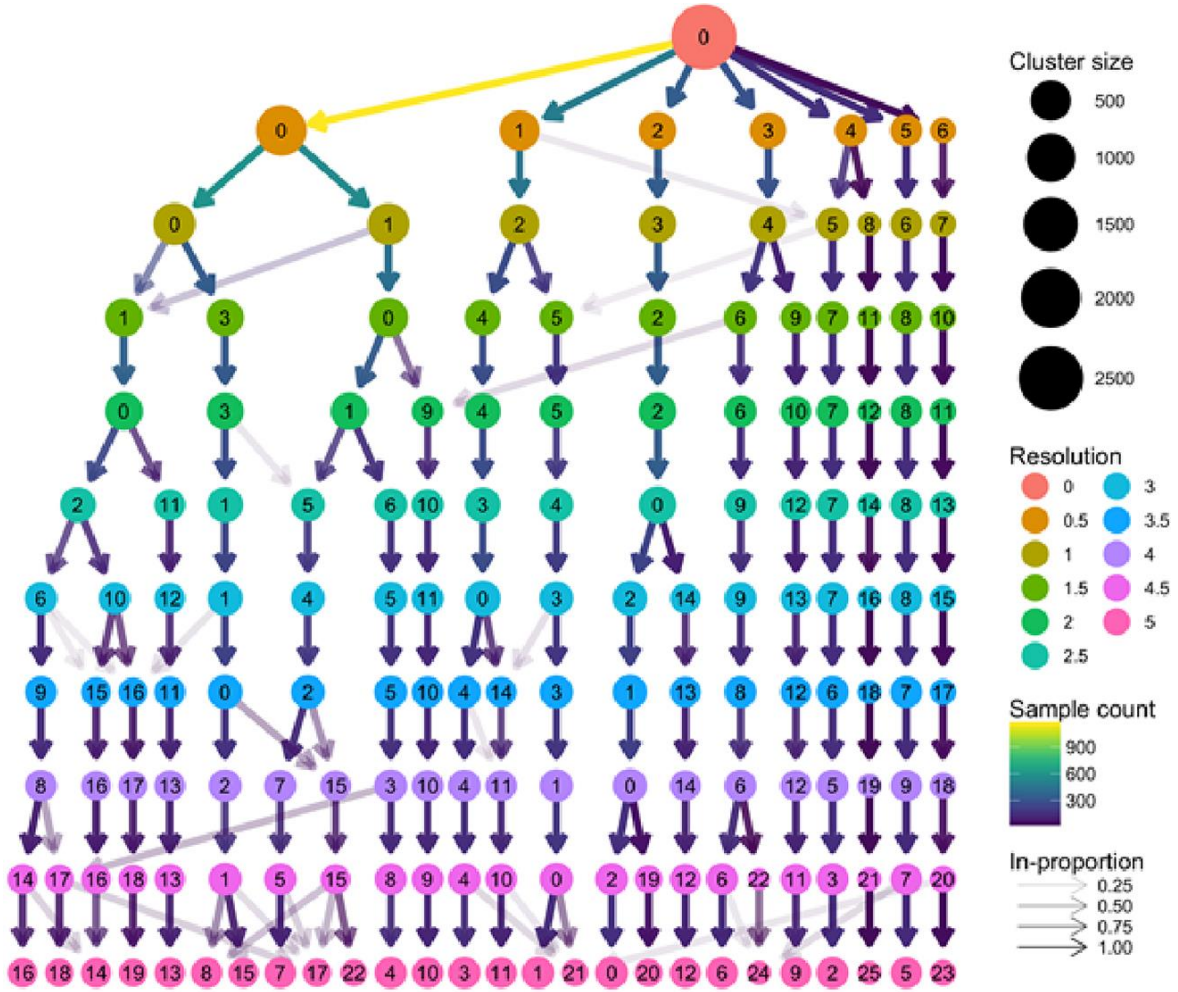
Graph-based Clustering

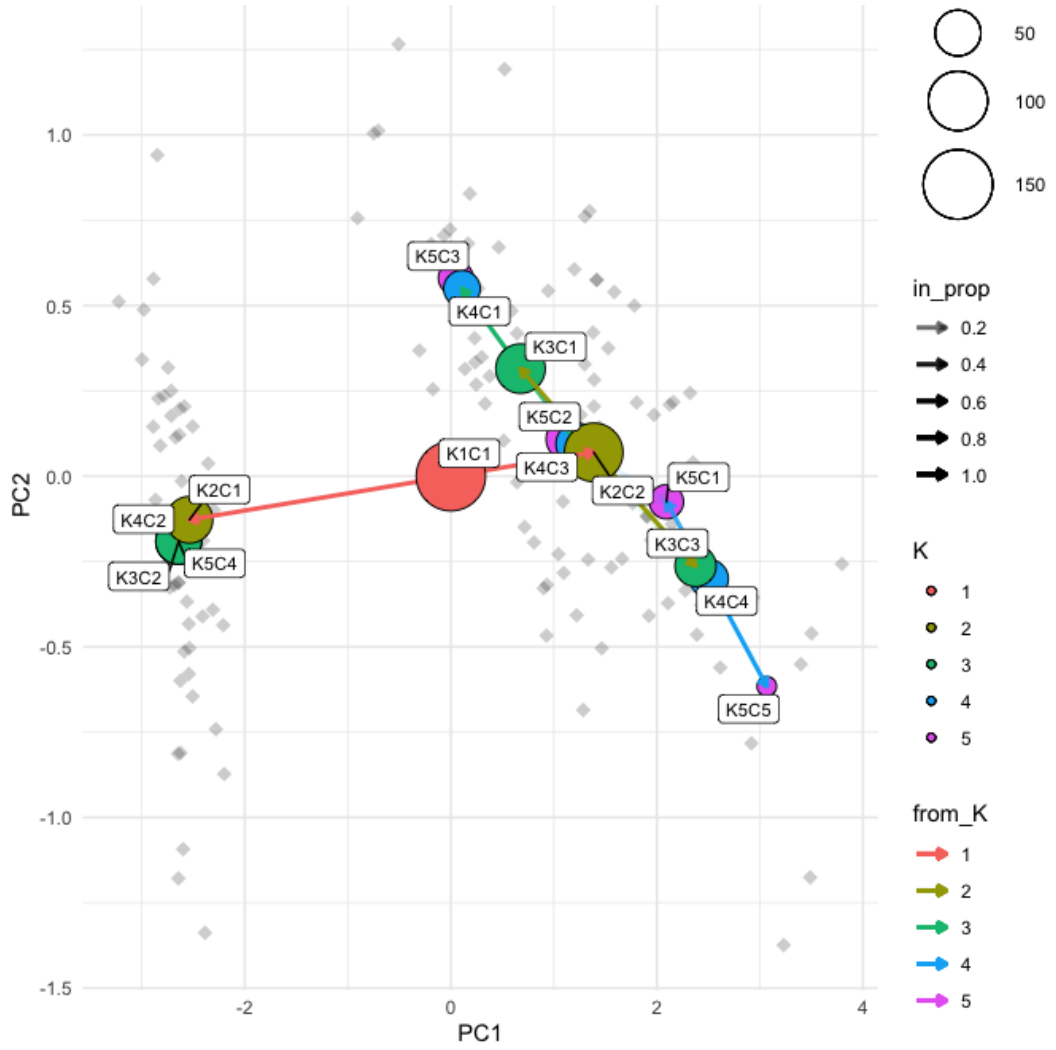
- Build shared-nearest-neighbour graph connecting the cells and finds tightly connected communities
- Increasing the number of neighbours when constructing the cell–cell graph indirectly decreases the resolution of graph-based clustering



Nature Reviews Genetics, 20, (2019)

Visualise clustering results





Lecture 2: Spatial Transcriptomics

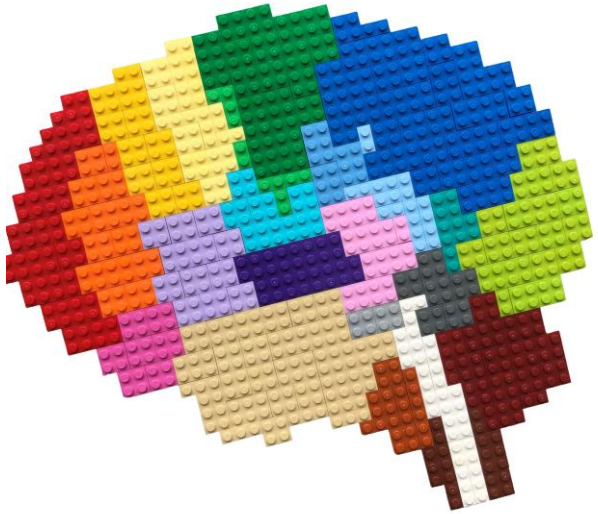
Spatial transcriptomics approach

Bulk

Single cell

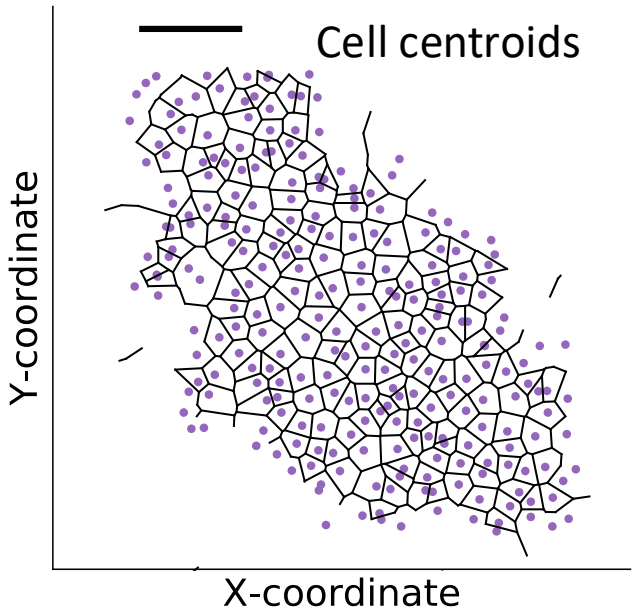
Spatial

Lego:
(@boxia)



Fruit salad:
(@LGMartelotto)

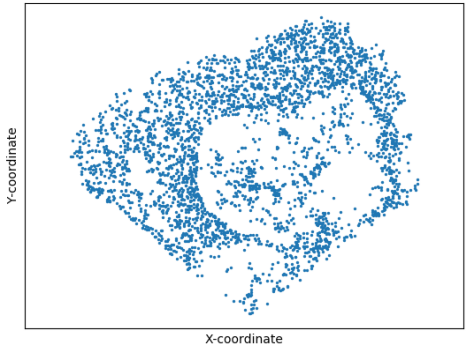
Spatial Transcriptomics Data (seqFISH): expression + location



(2050 cells and ~10,000 genes)

	Field of View	Cell ID	X	Y	Aanat	Aasdh	Aatf	Abat	Abca16	Abca17	...
0	0	1	1766.40	283.42	0	0	2	0	0	0	...
1	0	2	1891.40	348.38	0	0	0	0	2	0	...
2	0	3	1548.70	351.11	0	0	0	0	0	0	...
3	0	4	1657.60	357.37	0	0	0	2	0	0	...
4	0	5	1767.40	392.22	0	0	0	0	0	0	...

Fluorescence single molecule counts

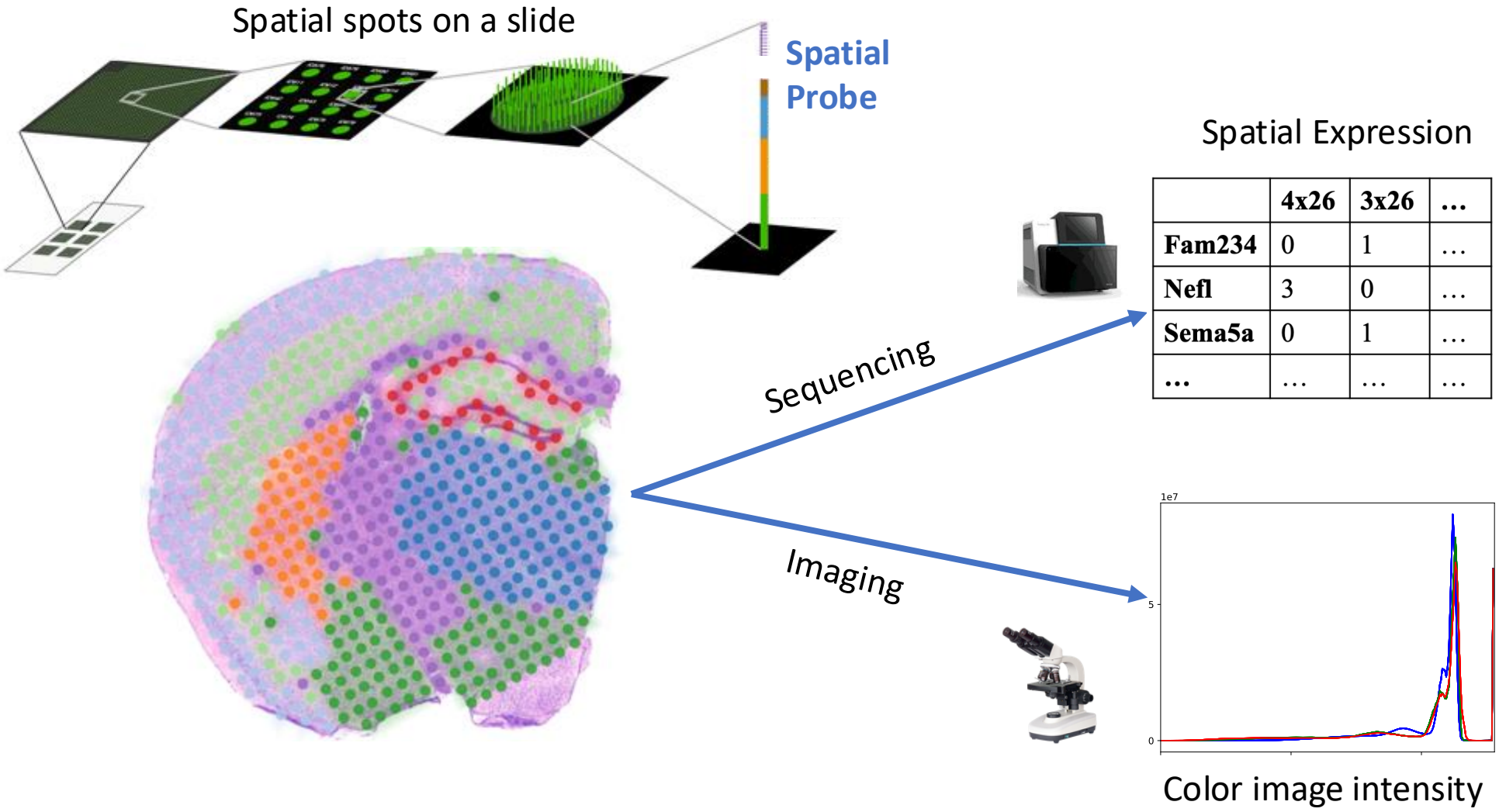


Example of seqFISH RNA in a cell: 3247 genes

Gene ID	**1**	**19**	**23**	**44**	**53**	**57**	**63**	**70**	**71**	**72**	**...**
0	653.00	675.24	687.21	733.85	615.16	663.99	611.06	669.65	638.03	601.10	...
1	434.34	428.89	479.06	472.43	469.95	464.81	443.74	417.42	430.46	472.07	...

Coordinates

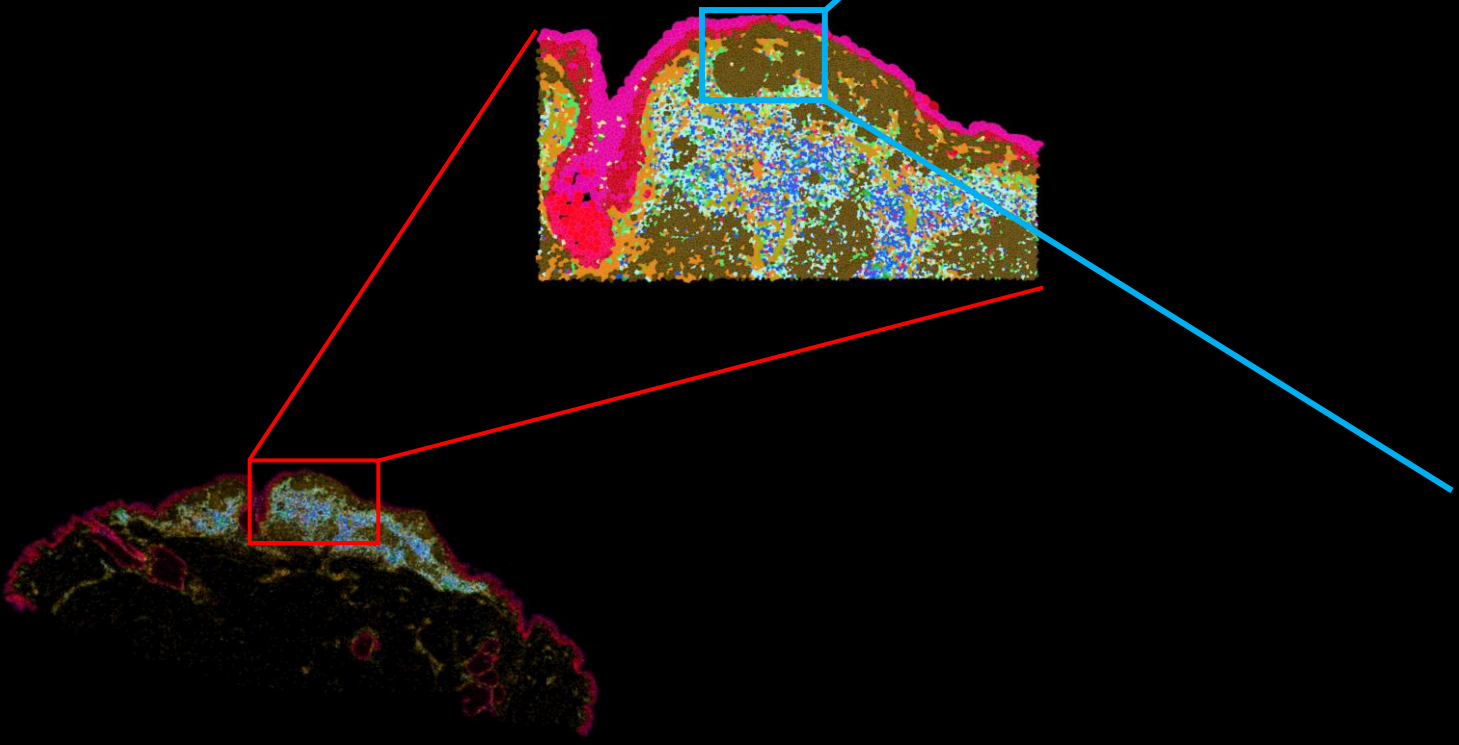
Spatial transcriptomics adds spatial dimension and tissue morphology

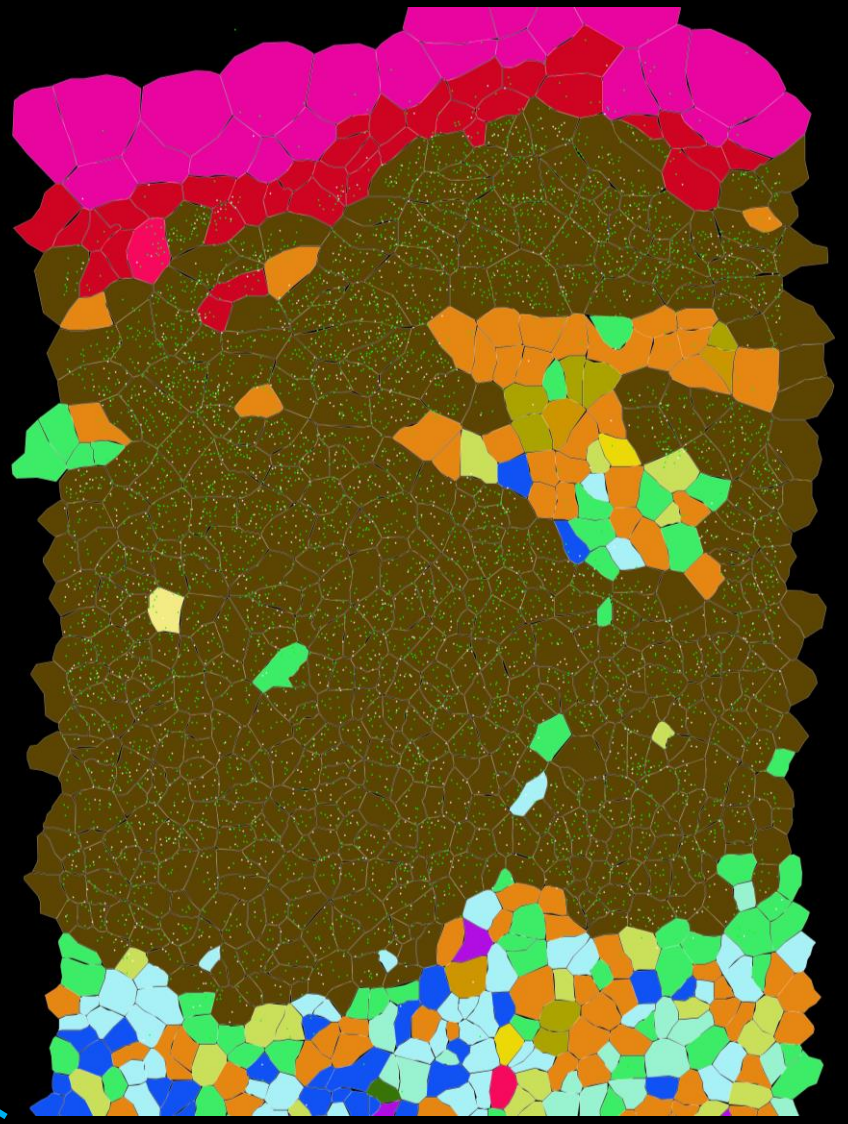


	4x26	3x26	…
Fam234	0	1	…
Nefl	3	0	…
Sema5a	0	1	…
…	…	…	…

- On-tissue expression profiling (>20,000 genes); each spot contains ~1-9 cells; tissue < 6.5 mm x 6.5 mm
- Other spatial technologies are different (complementary) in resolution, throughput, scale, sensitivity ect.

Xenium - Spatial transcriptomics at single cell resolution

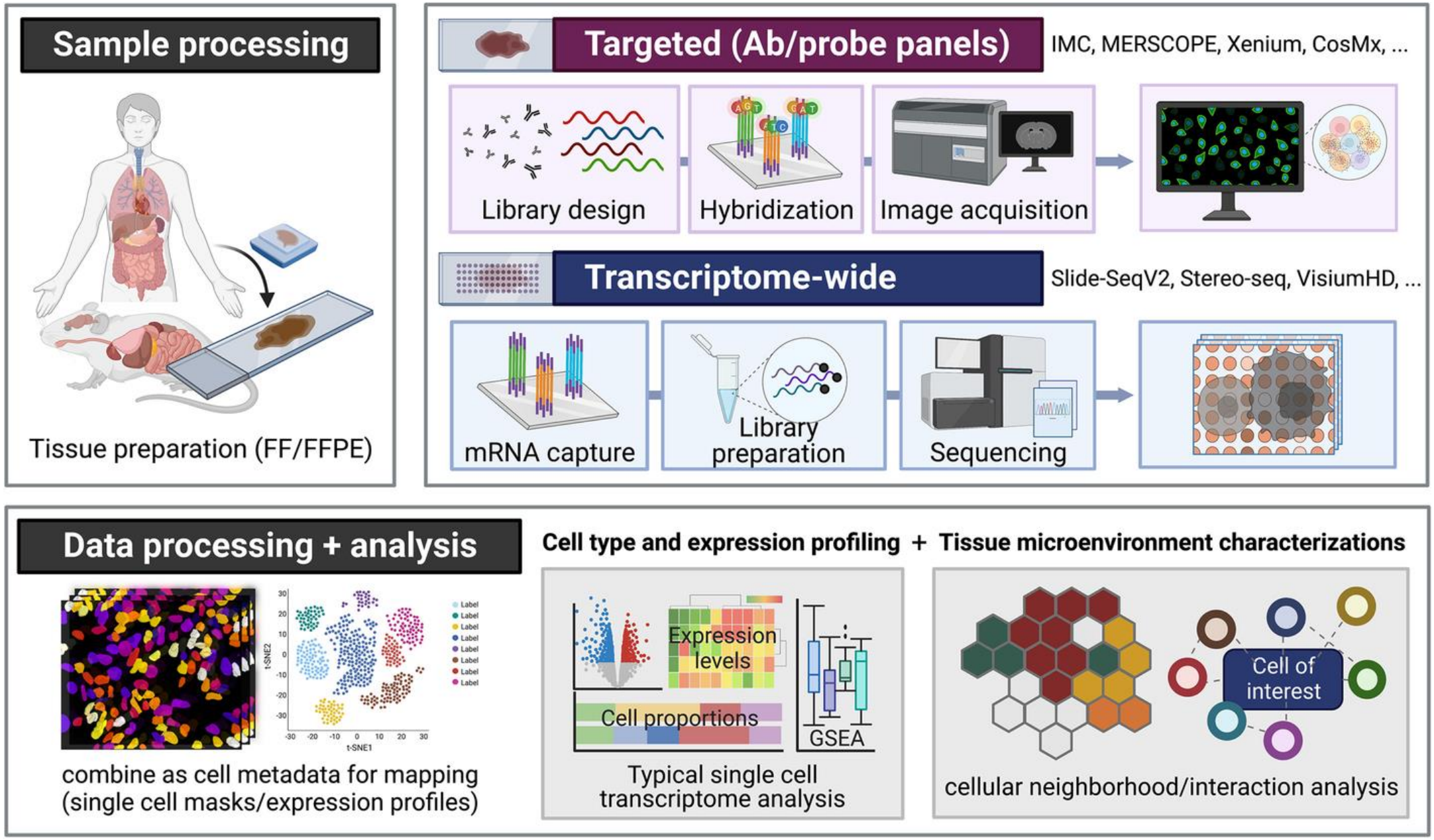




cell_types

- Melanocytes
- Imm_CD8+ T cell
- Imm_DC
- Imm_Fibroblast
- Imm_Macrophage
- Imm_Tcell
- KC_Differentiating
- KC_Basal
- KC_Cornified
- Imm_LC
- Imm_Mast Cells
- Imm_Endothelial c
- Imm_pDC
- Imm_Bcells
- Imm_NK

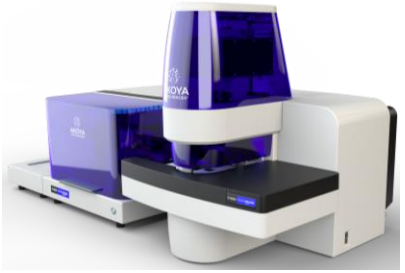
Spatial Analysis Types



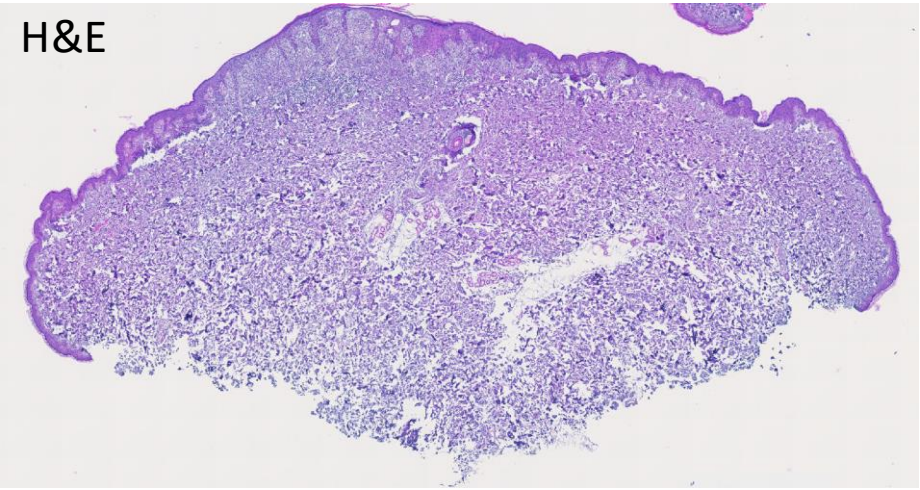
(Park et al, 2022)

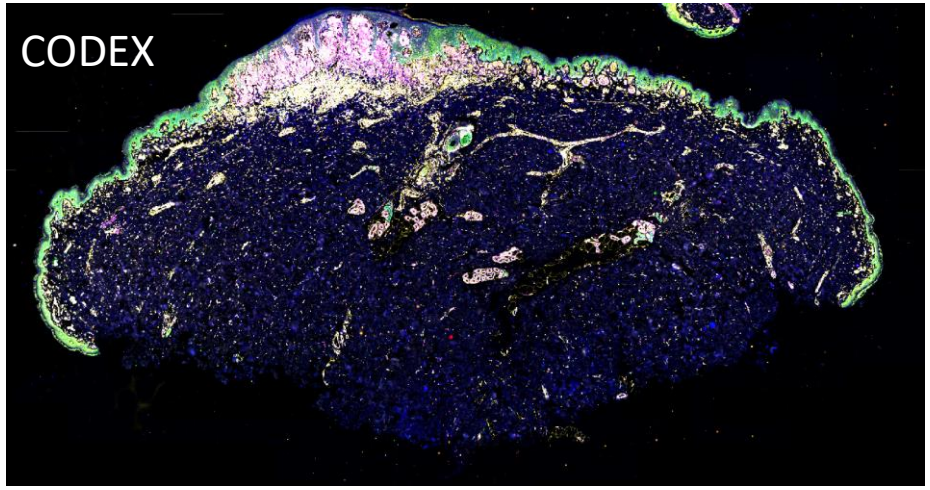
Lecture 3: Spatial Proteomics

Datasets – Melanoma (Skin)

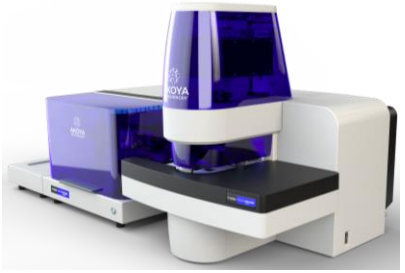


PhenoCycler-Fusion
(CODEX)

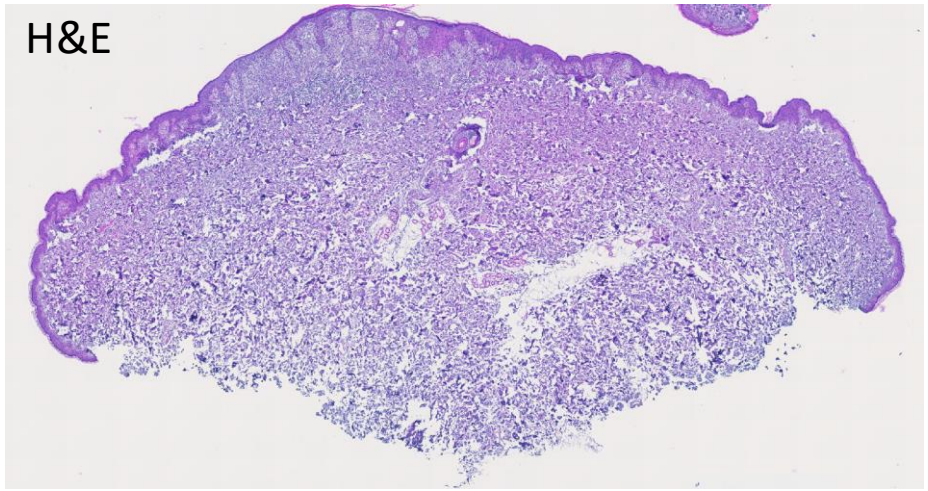


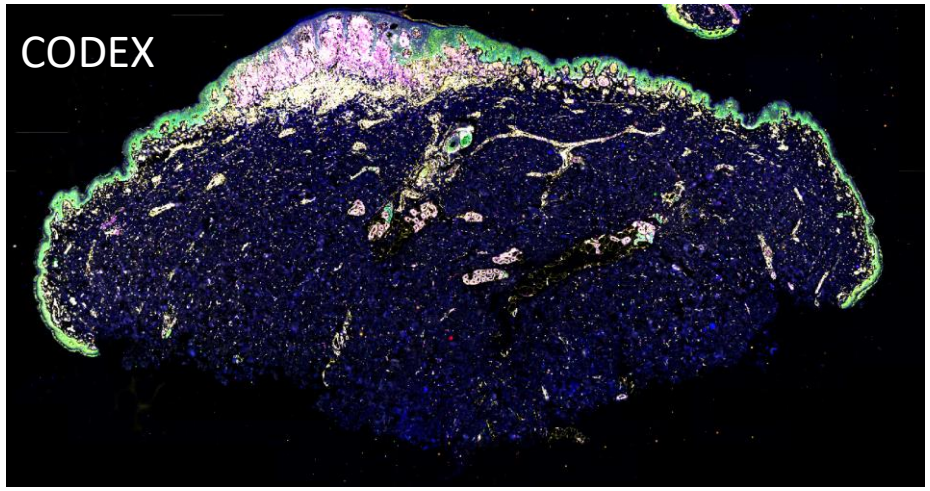


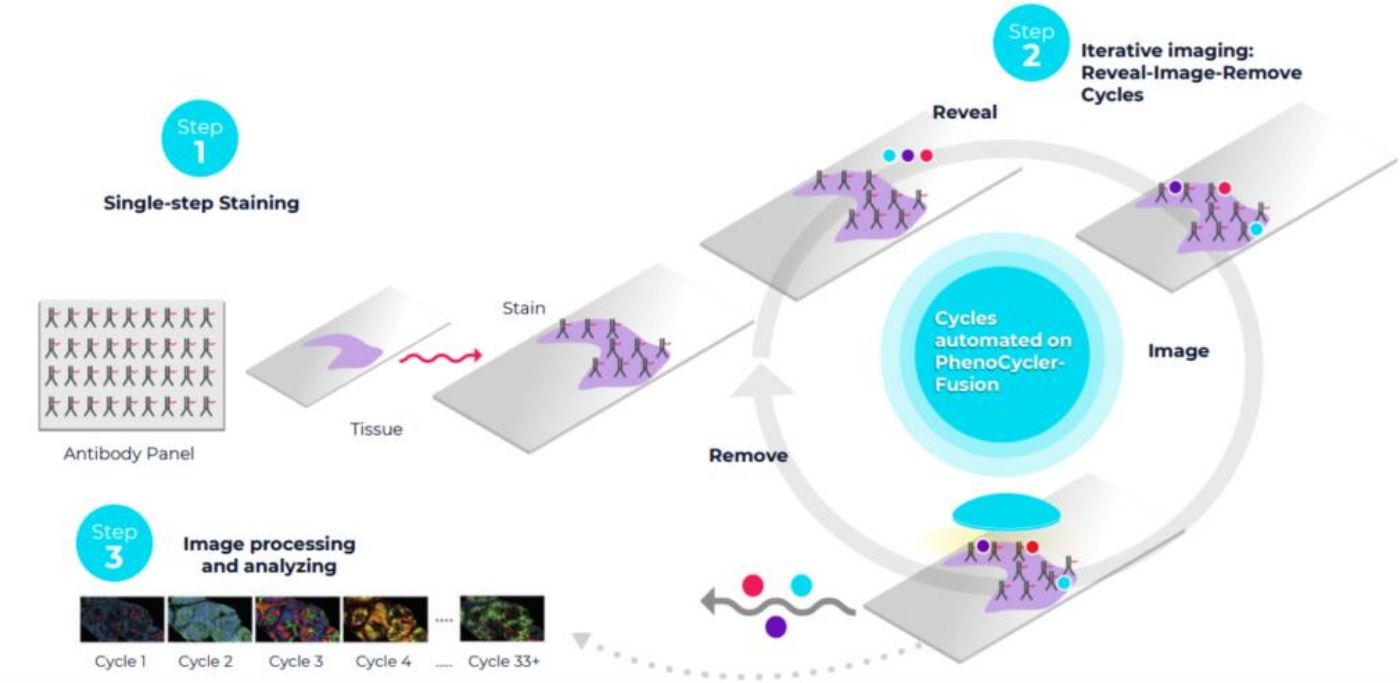
Datasets – Melanoma (Skin)



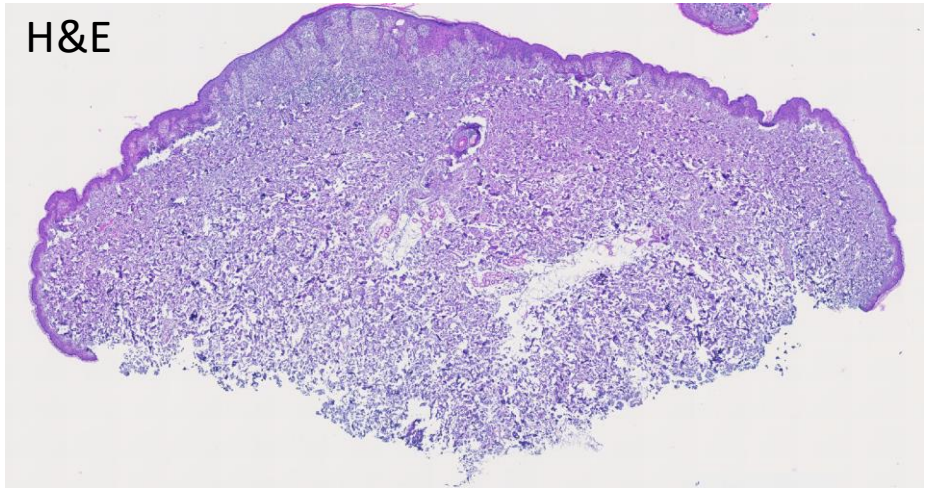
PhenoCycler-Fusion (CODEX)



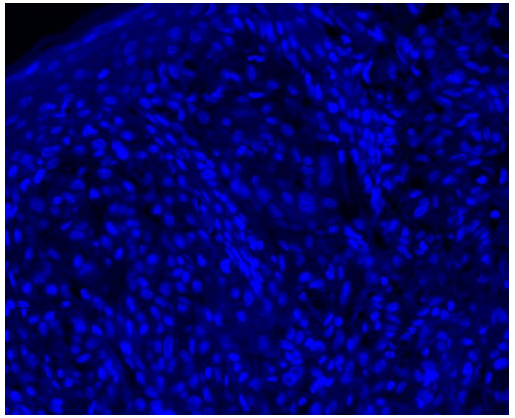




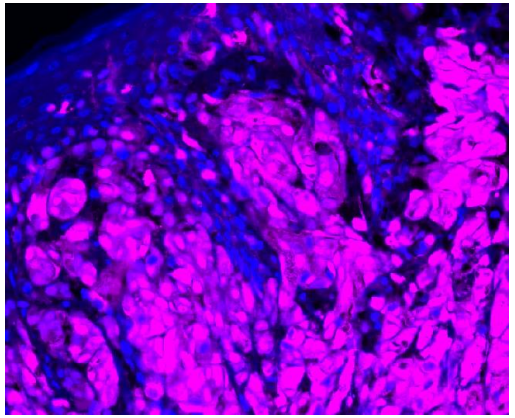
Datasets – Melanoma (Skin)

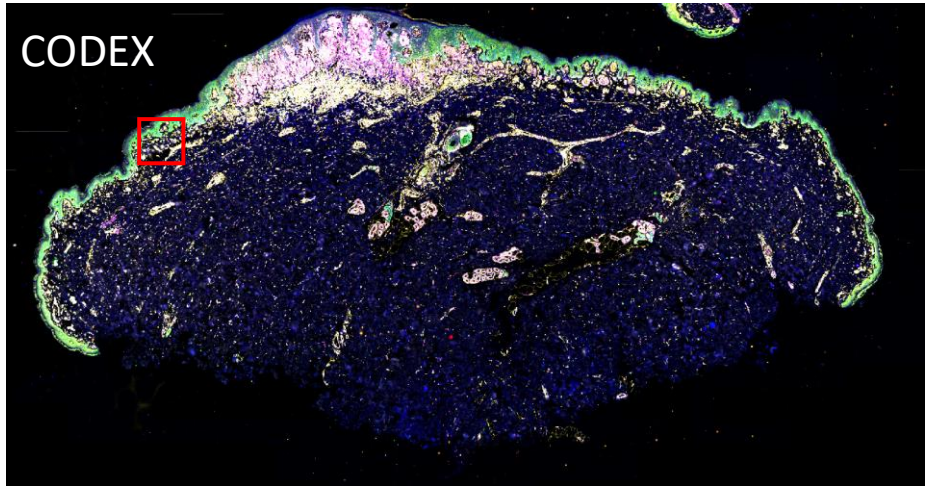


DAPI

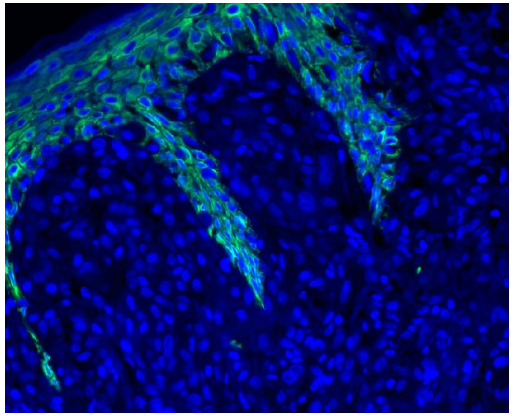


S100B

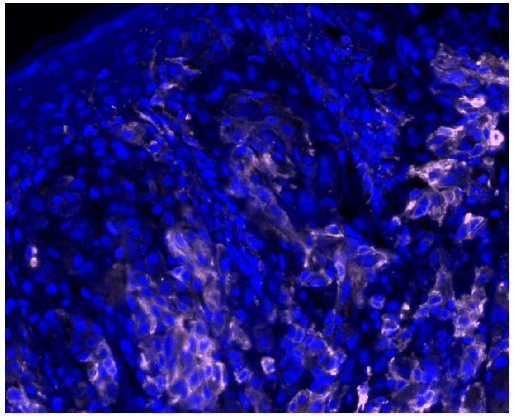




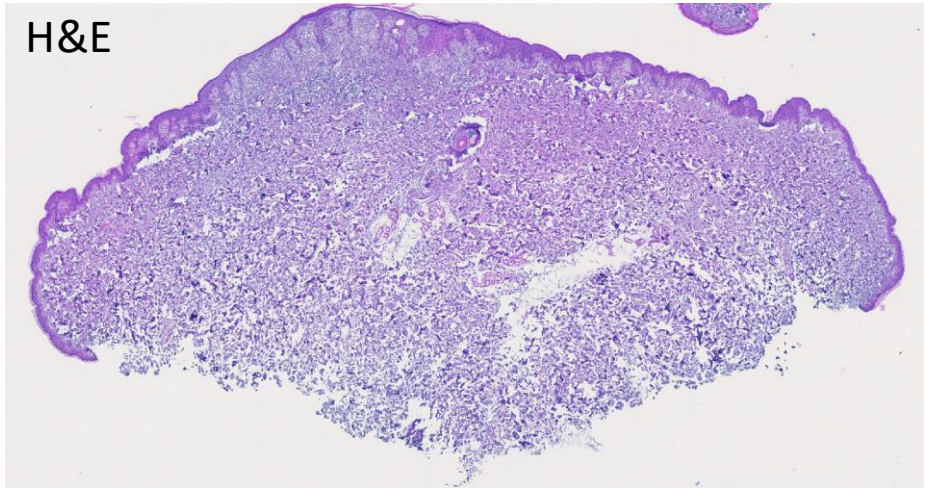
Keratin 14

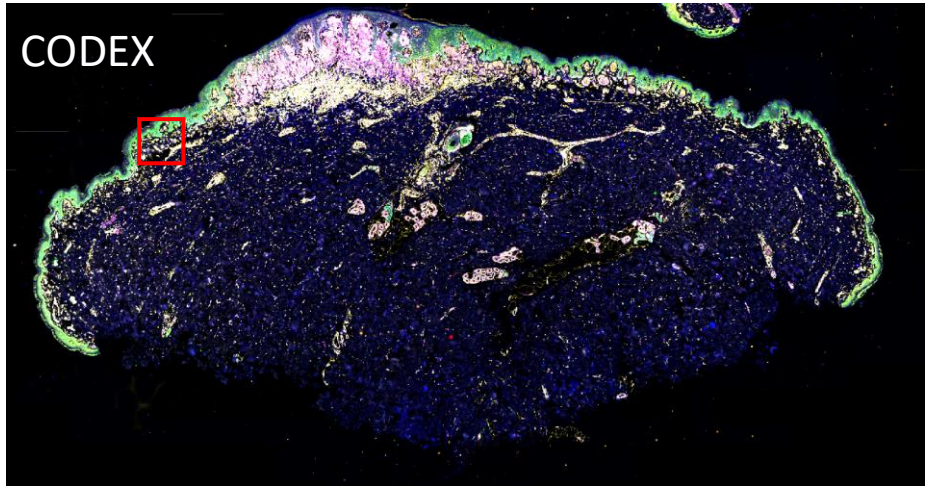


PMEL

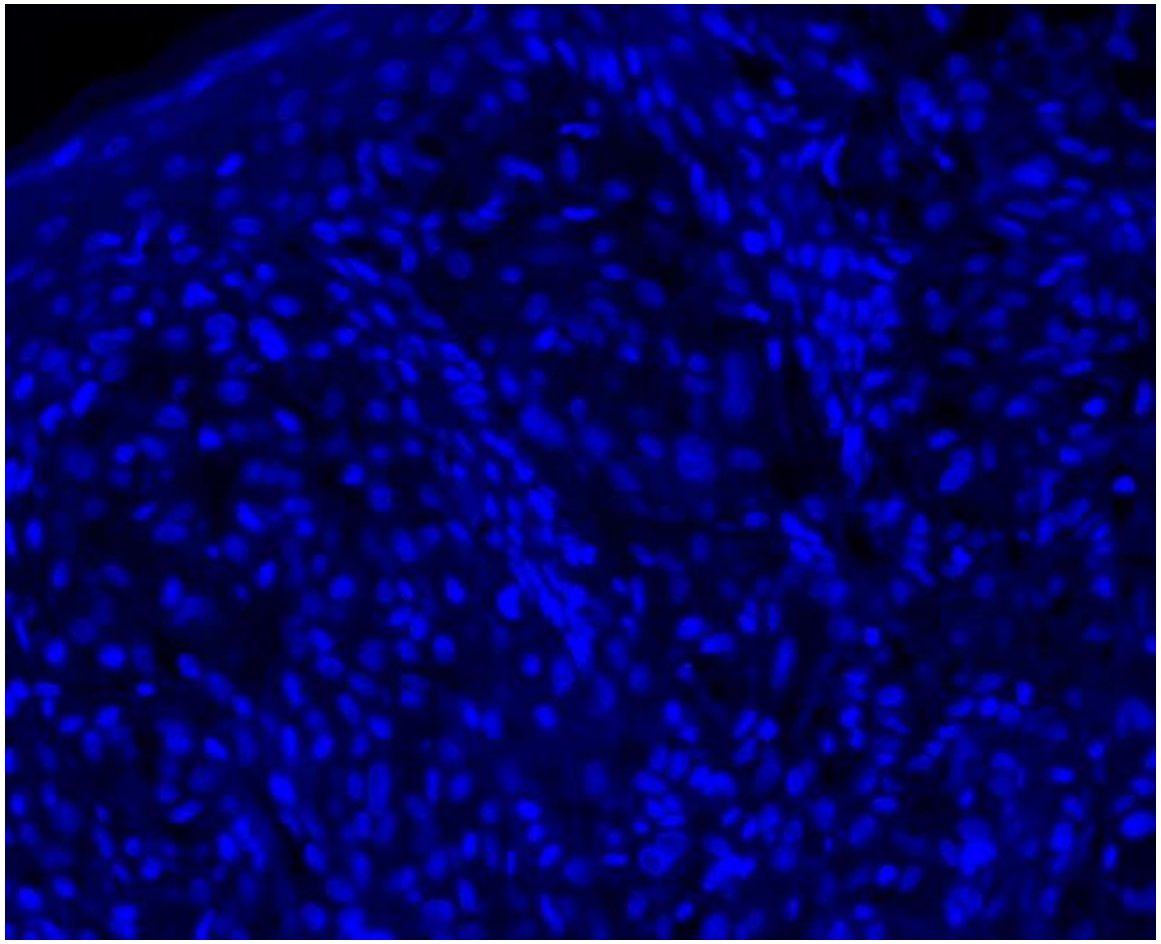


Datasets – Melanoma (Skin)

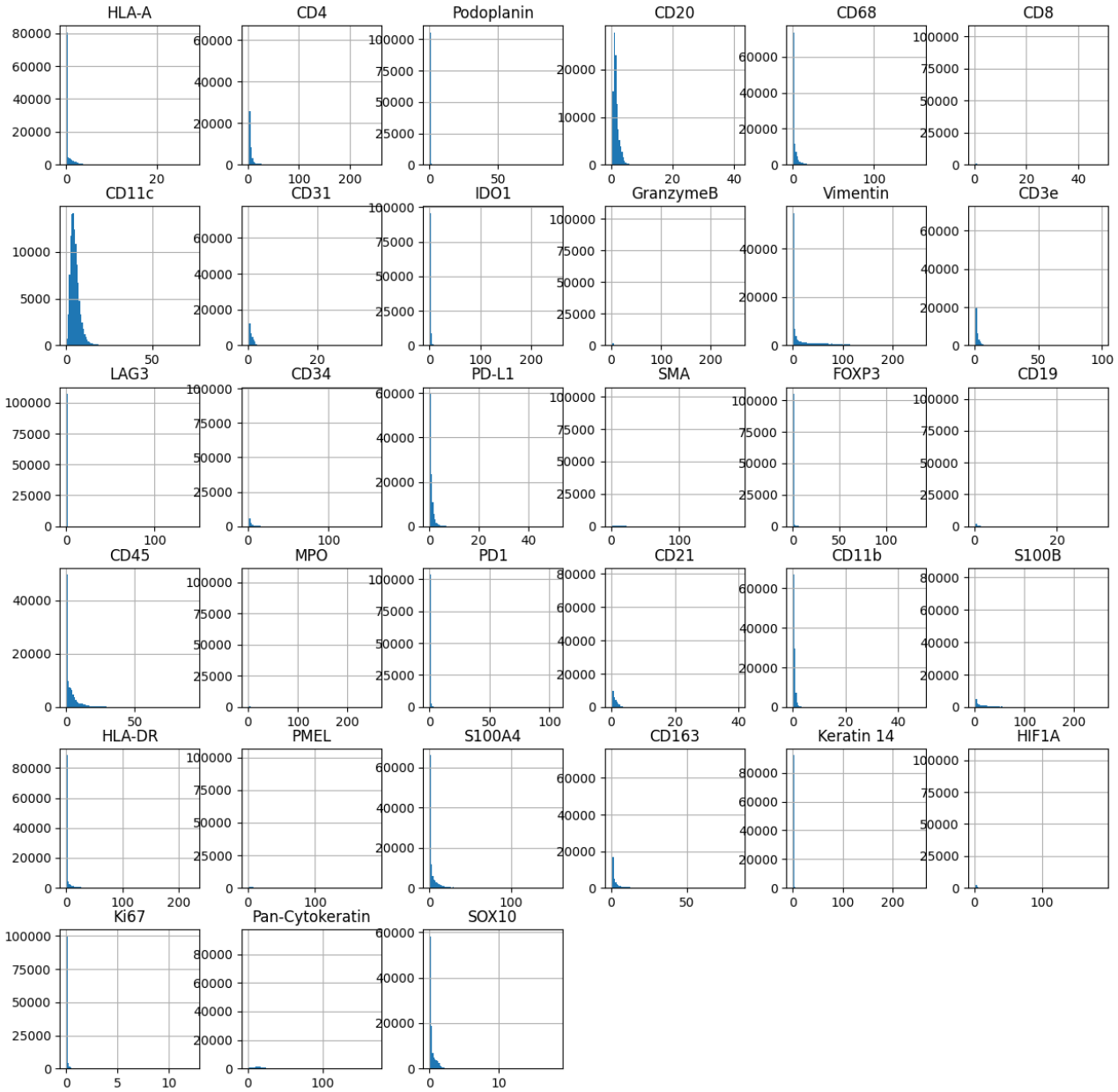




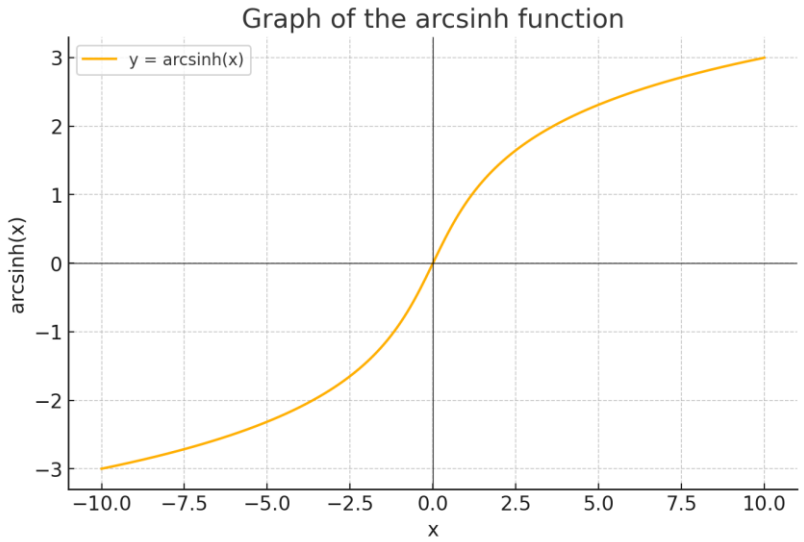
DAPI Cell segmentation



1. Protein Measurements



Normalisation: $$\operatorname{asinh}\left(\frac{X}{5Q(0.2, X)}\right)$$



Standardisation: $$z = \frac{(X - \mu)}{\sigma}$$

1. Protein Measurements