# 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

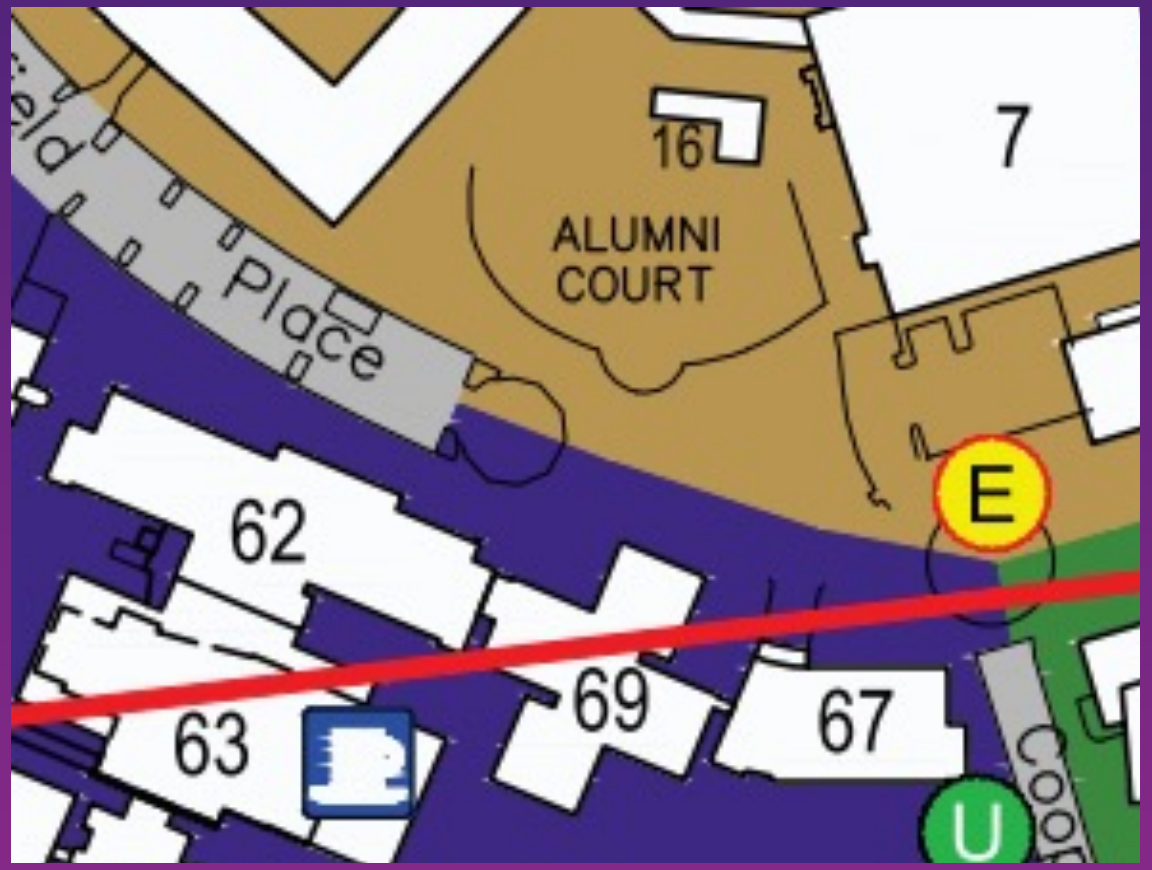
# General Information:

- We are currently located in Building 69


Emergency evacuation point

- Food court and bathrooms are located in Building 63

- 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



# Data Agreement

To maximize your learning experience, we will be working with genuine human genetic data, during this module.

Access to this data requires agreement to the following in to comply with human genetic data ethics regulations

If you haven’t done so, please email <ctr-pdg-admin@imb.uq.edu.au> with your name and the below statement to confirm that you agree with the following:

“I agree that access to data is provided for educational purposes only and that I will not make any copy of the data outside the provided computing accounts.”

# Learning materials

Instructions to access WiFi/desktop/server:

https://suave-pillow-de4.notion.site/Instruction-to-Computing-Resources-dcba658c9a584e6d80a443c5d64042d8?pvs=4

Slides and practical notes:

https://cnsgenomics.com/data/teaching/GNGWS24/module[1-6]/

# Module 2 - running the learning materials

https://github.com/GenomicsMachineLearning/qimr-teaching-2024/tree/main

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/winter_school_2024`
- `git clone https://github.com/GenomicsMachineLearning/qimr-teaching-2024`
- `~/qimr-teaching-2024/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

Room 314/315, Building 69

Aiming at interactive session, we provide the presence of a large teaching team for more one-to-one discussion.

Lecturers/Instructors: Quan Nguyen, Andrew Causer, Levi Hocki, Onkar Mulay, Prakrithi Pavithra, Andrew Newman, Xiao Tan, Feng Zhang

# Module 2 Cellular Omics – Leaning Objectives

- Technologies for generating single-cell and spatial transcriptomics data
- Technologies for other spatial omics, focusing on proteomics
- Exploratory visualisation to understand the data
- 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)
  - Analysing cell-cell interactions
- Analysing spatial proteomics data and integration with spatial transcriptomics through imaging analysis techniques
- Machine learning analysis of sequencing and imaging data

# Lecture Outline

## Day 1

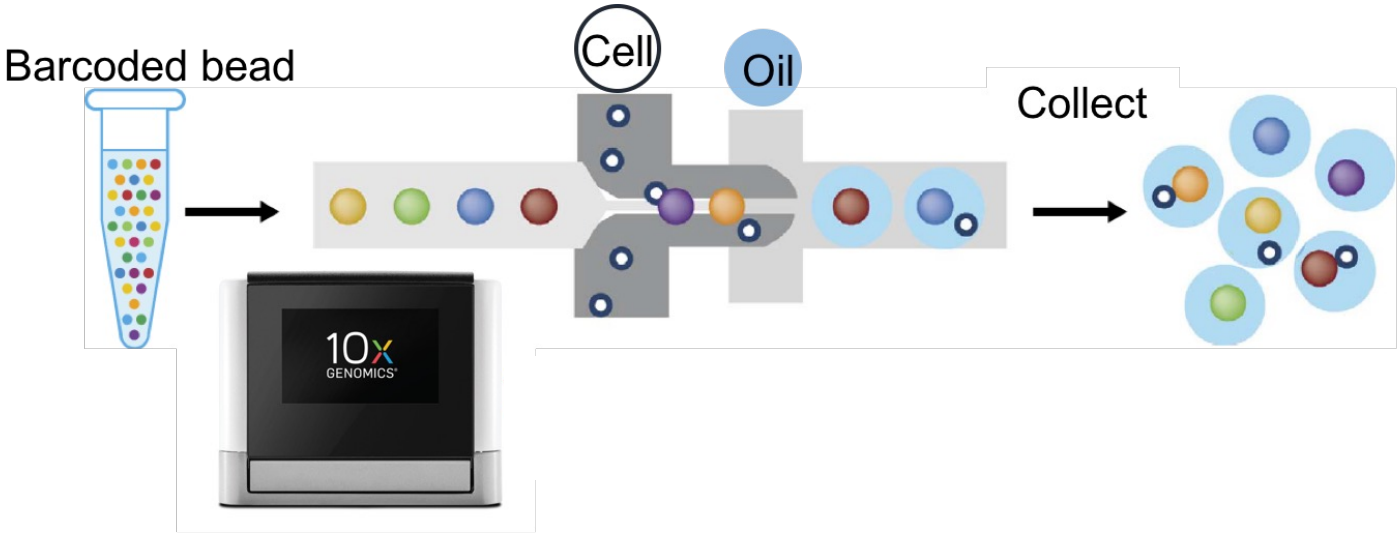
- Lecture 1: Introduction Single Cell and Spatial Transcriptomics
- Lecture 2: Defining Cell Types
- Lecture 3: Review Data Structure and Understand Spatial Concepts by Visualisation
- Lecture 4: Spatial DNA-level Analysis for Copy Number Variation
- Lecture 5: Cell Community Identification
- Lecture 6: Cell-Cell Interactions

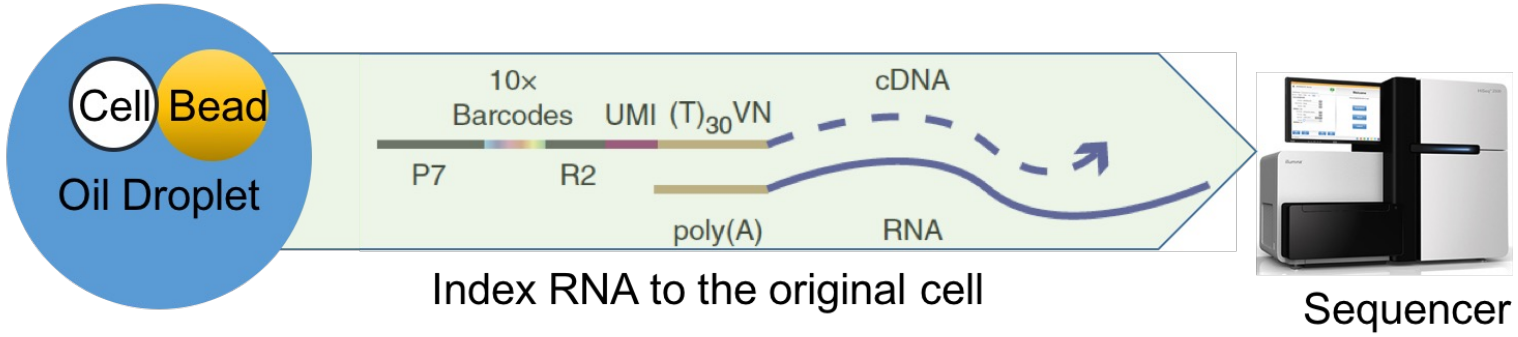
## Day 2

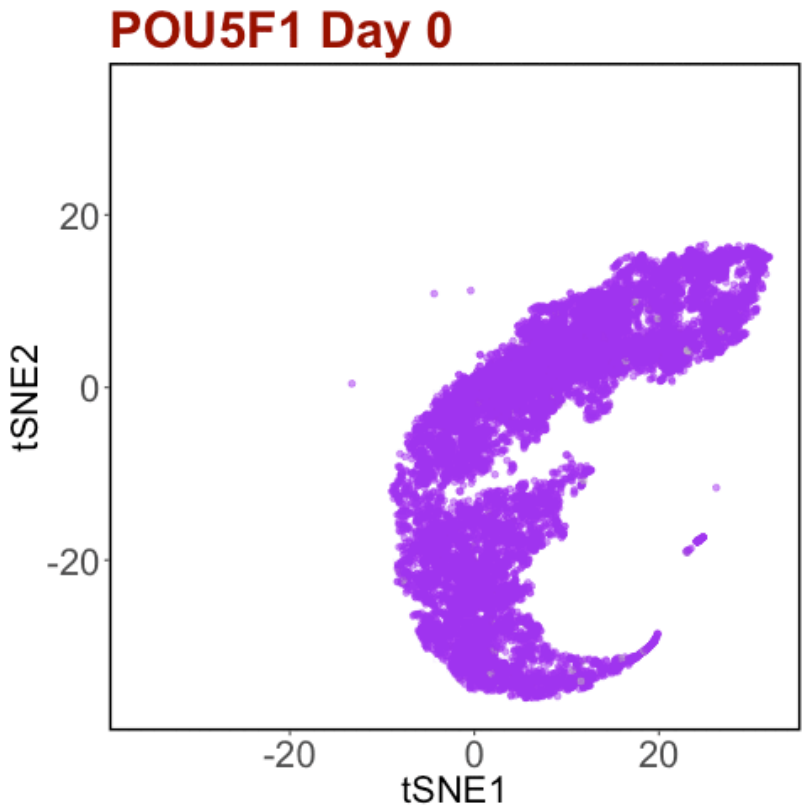
- Lecture 7: Tissue Segmentation and Spatial Statistics
- Lecture 8: Spatial Proteomics
- Lecture 9: Machine Learning

# Lecture 1: Introduction Single Cell and Spatial Transcriptomics

# Single cell RNA sequencing



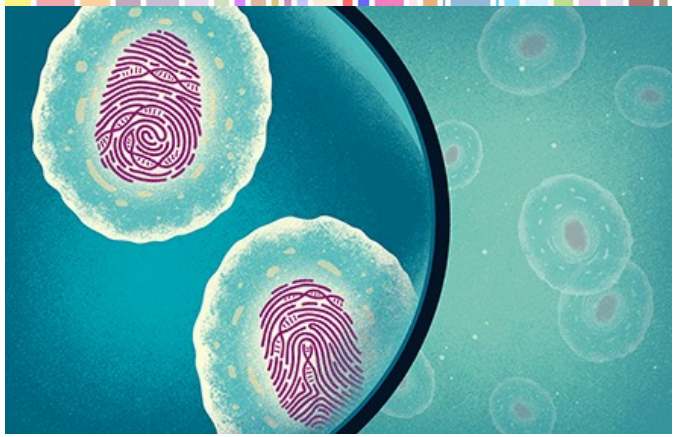




- Single-cell RNA sequencing (scRNA-seq) measures thousands of genes in a separate cell
- How: 3 barcoding steps for sample, cell and RNA molecule
- Scale: bulk RNA-seq (5 samples) vs. scRNA-seq (45 K cells), a ~900 times bigger gene count matrix

# Single cell informatics

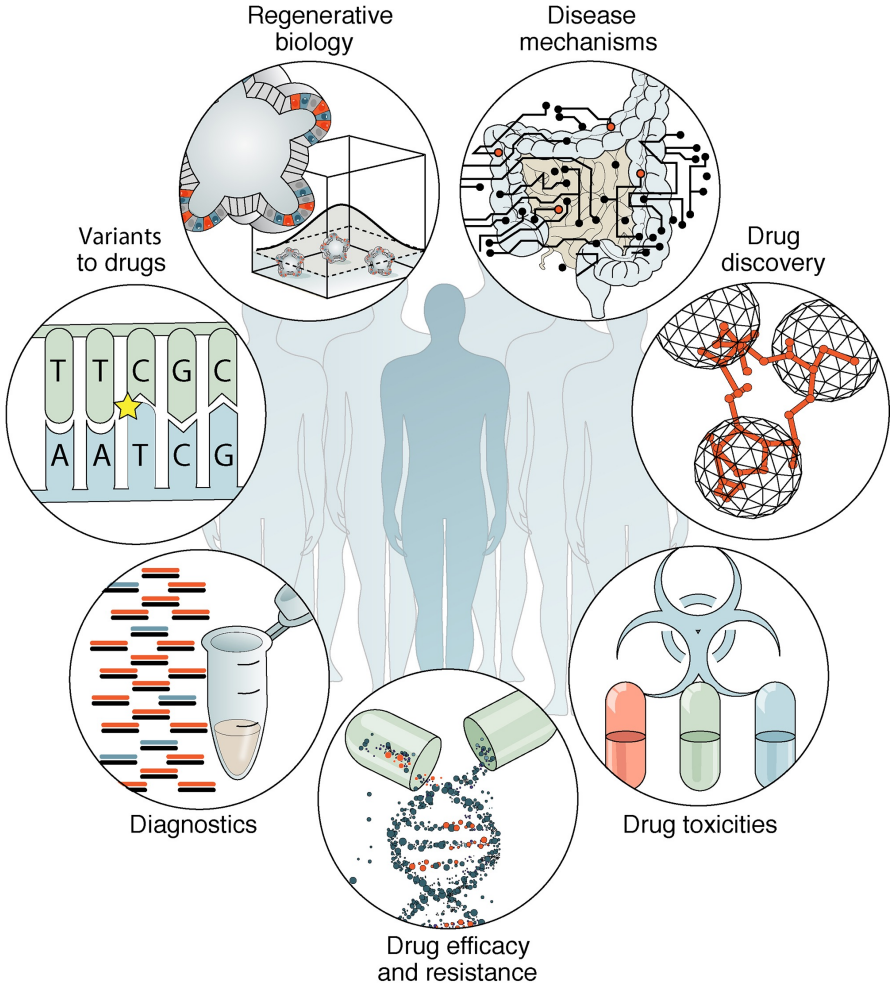
Scale



Resolution

# INFORMATICS

Precision Genomics Medicine

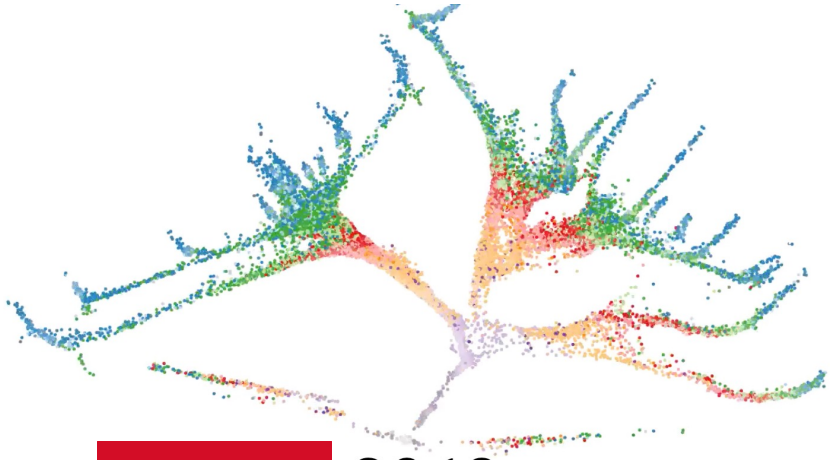



**The Genomics and Machine Learning Team**

Quan Nguyen, Andrew Causer, Levi Hocki, Onkar Mulay, Prakrithi Pavithra, Andrew Newman, Xiao Tan, Feng Zhang

# Advanced genomics technologies

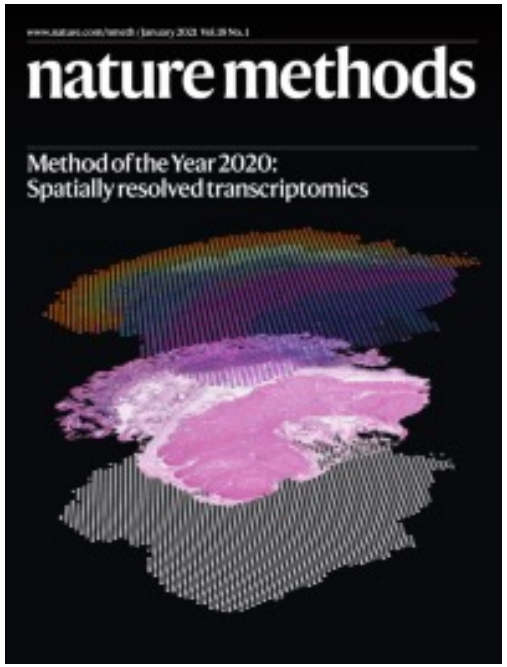
## 2018: Single Cell Transcriptomics



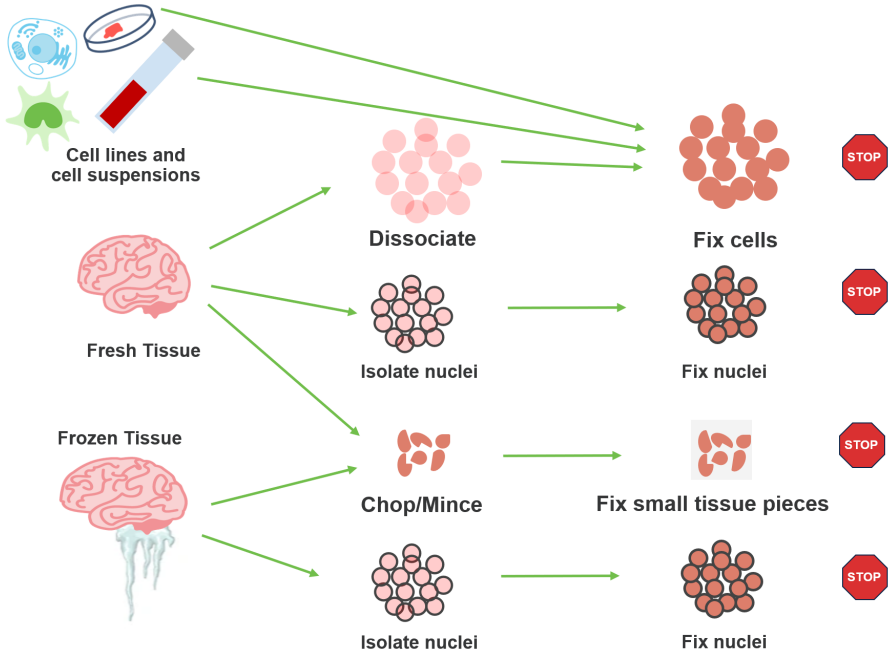
Science AAAS 2018 BREAKTHROUGH *of the* YEAR

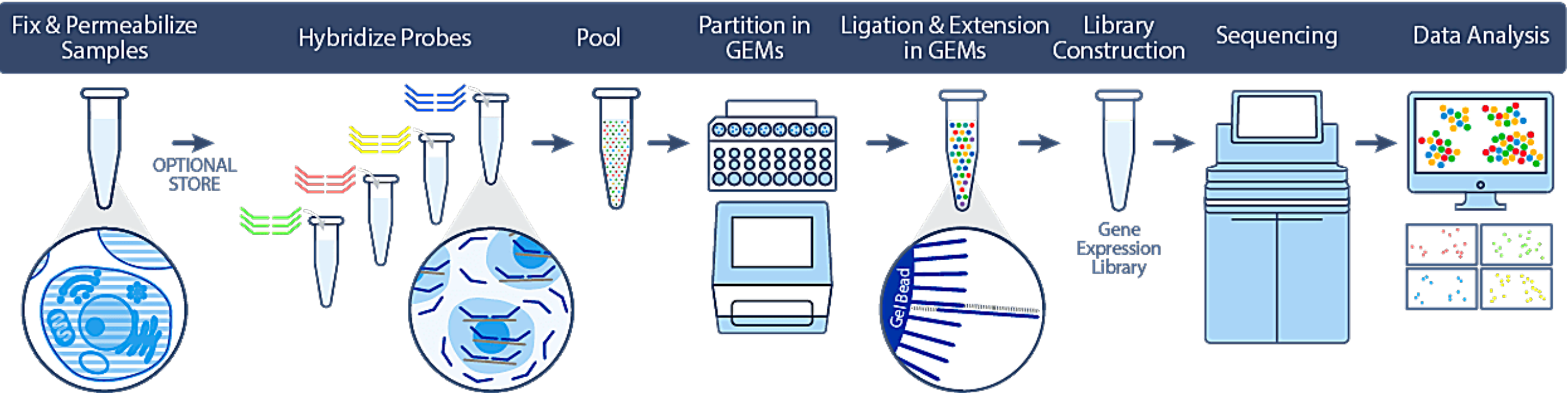
## 2019: Single Cell Multiomics


## 2020: Spatial Transcriptomics

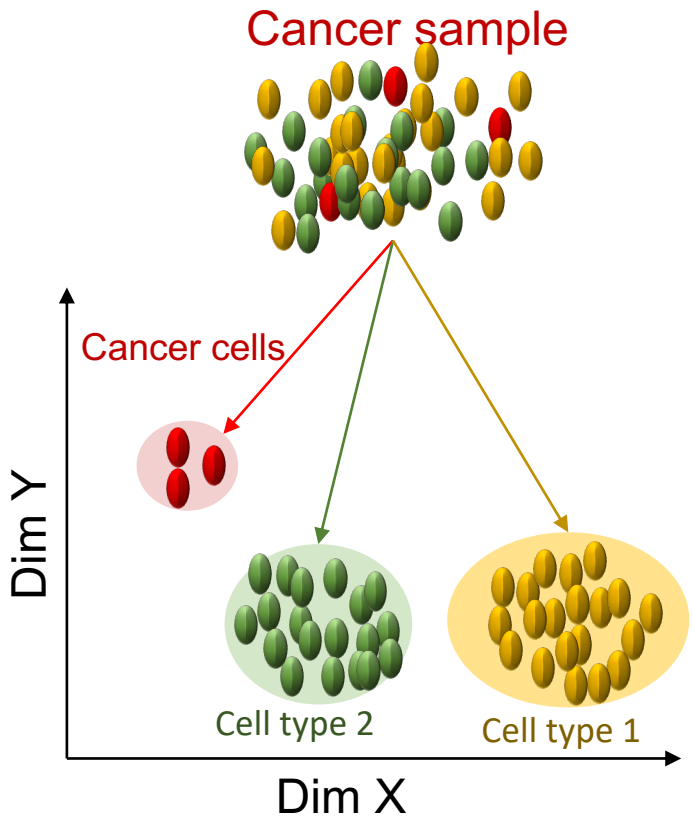


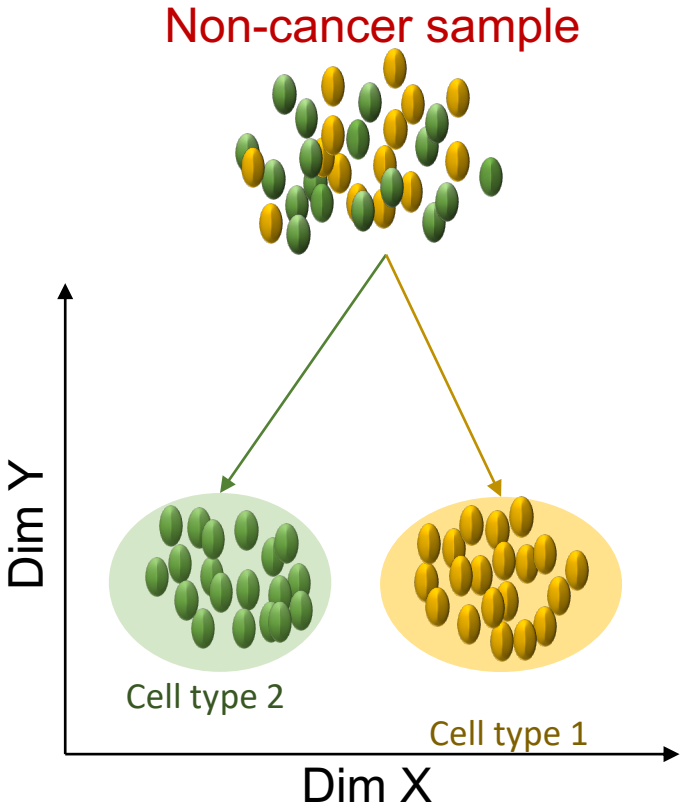
# Increase single cell experiments to millions of cells

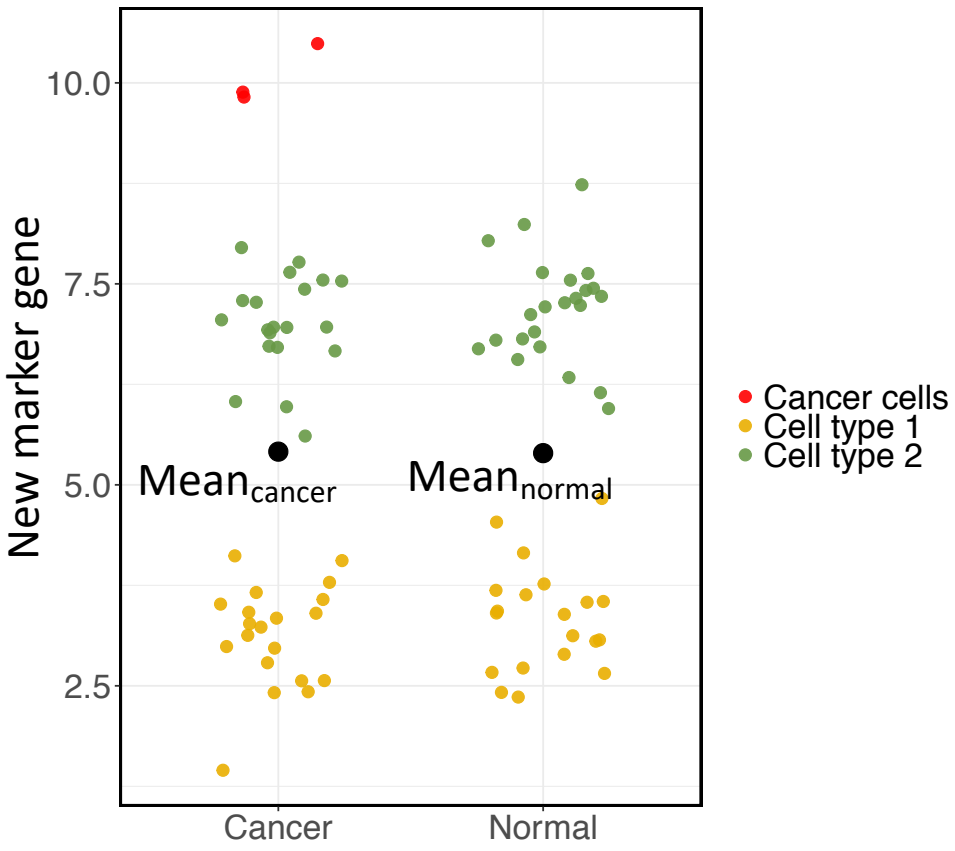


# Disease at single-cell resolution



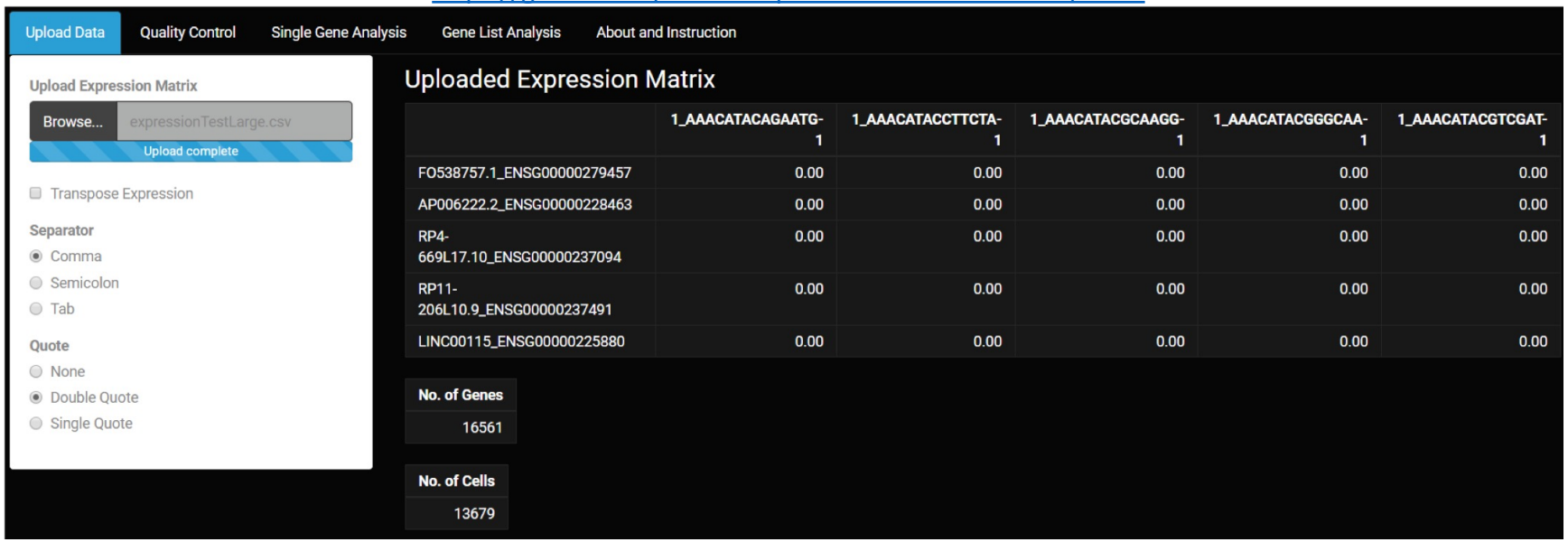



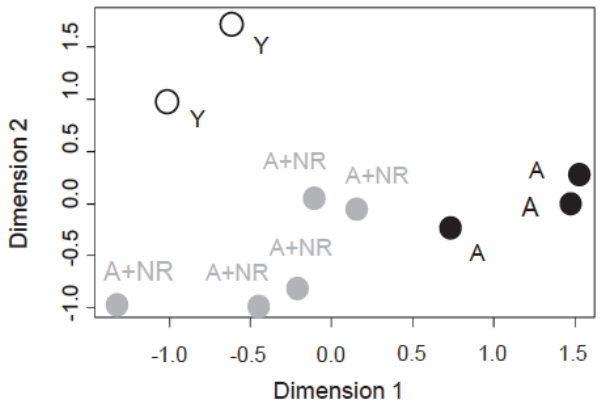

- Bulk RNA sequencing: no difference in mean expression
- Single-cell sequencing: can detect higher expression in cancer cells

# Single cell data vs. bulk data

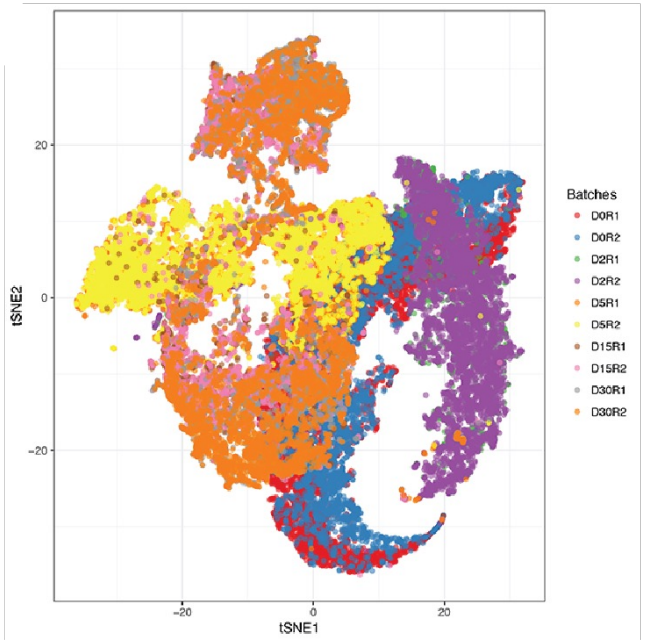
https://github.com/IMB-Computational-Genomics-Lab/scIVA




Bulk

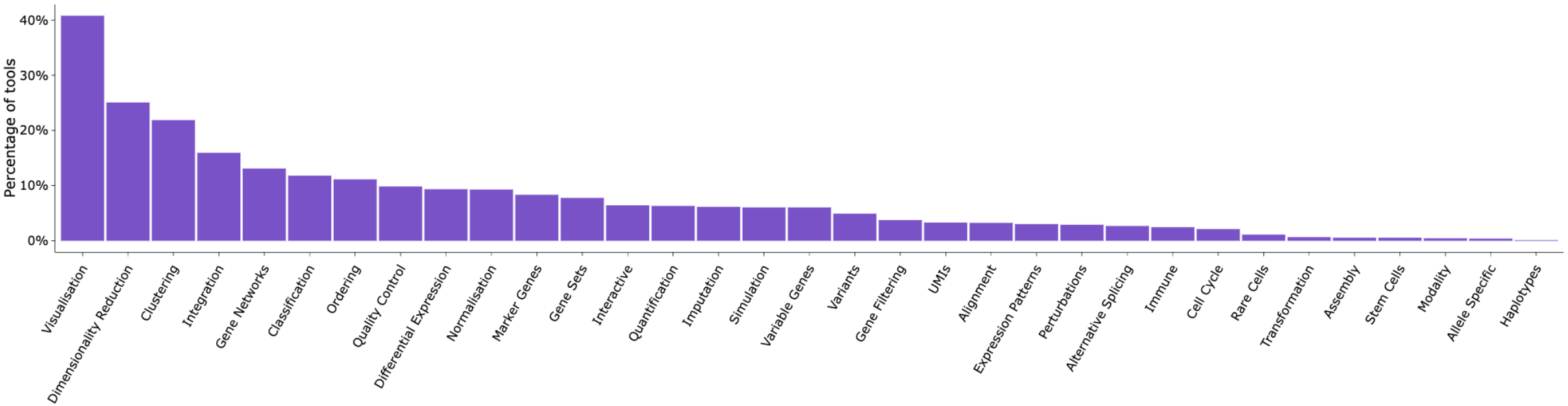


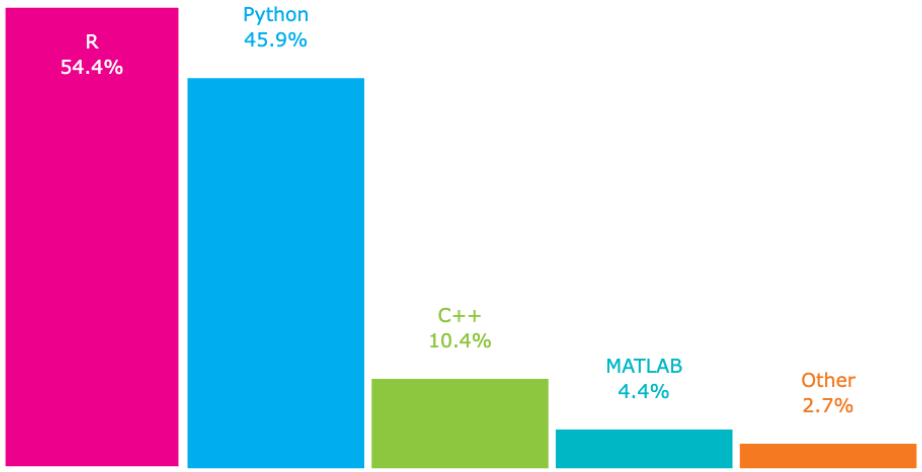

Single cell

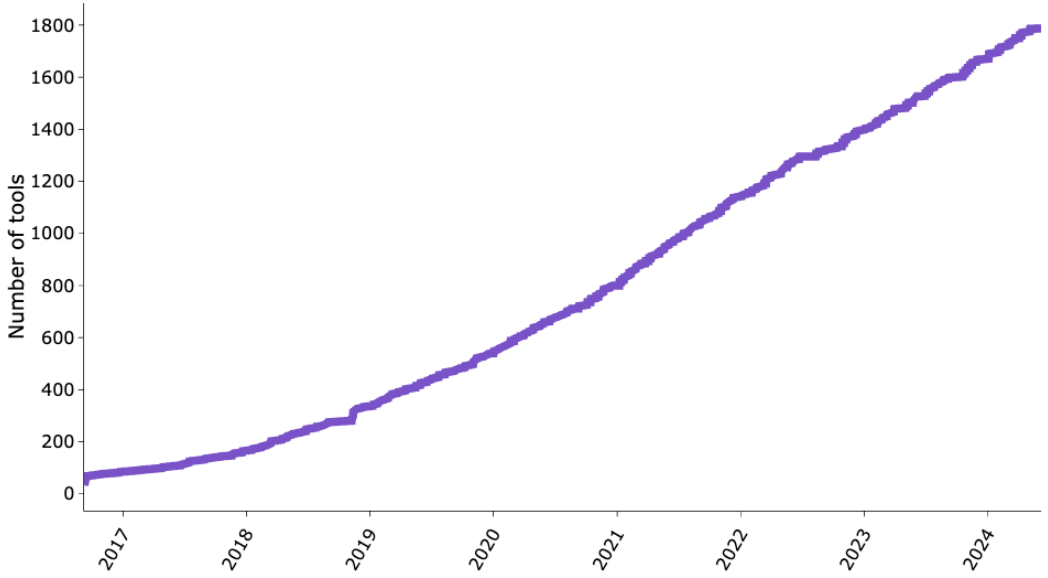


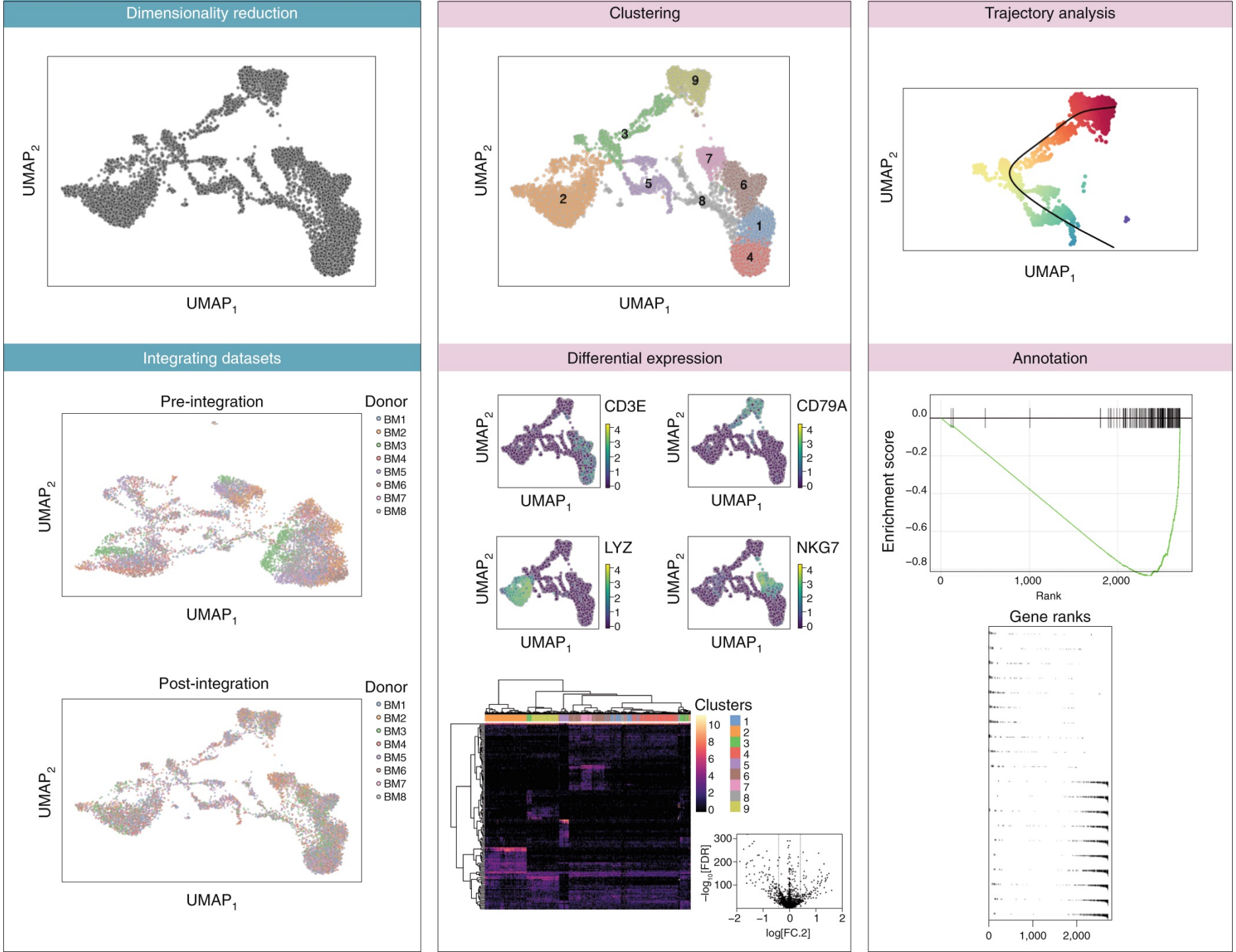

| | Single cell | Bulk |
|---|---|---|
| Noisy data | Undetected genes (zero inflation) | Deep sequencing, most genes detected |
| Cell-cell variation | Measured | Not measured |
| Data size | Thousands of cells (1 cell ~ 1 bulk sample) | 10-100 samples |

# Single cell data analysis



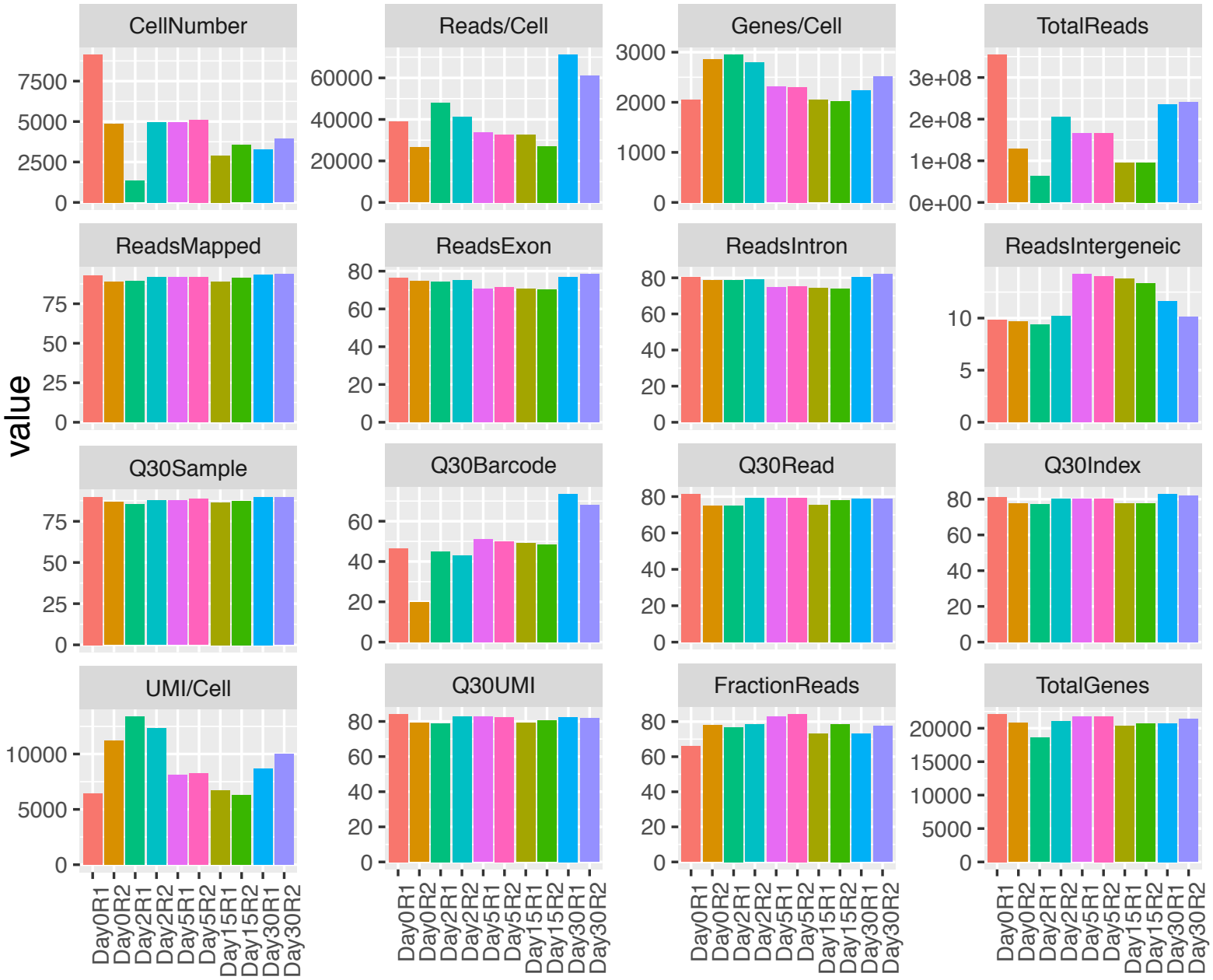




Dimensionality reduction
UMAP2
UMAP1
Clustering
9
3
7
5
6
2
8
1
4
Trajectory analysis
Integrating datasets
Pre-integration
Donor
BM1
BM2
BM3
BM4
BM5
BM6
BM7
BM8
Post-integration
Differential expression
CD3E
CD79A
LYZ
NKG7
Clusters
log[FC.2]
Annotation
Enrichment score
Rank
Gene ranks

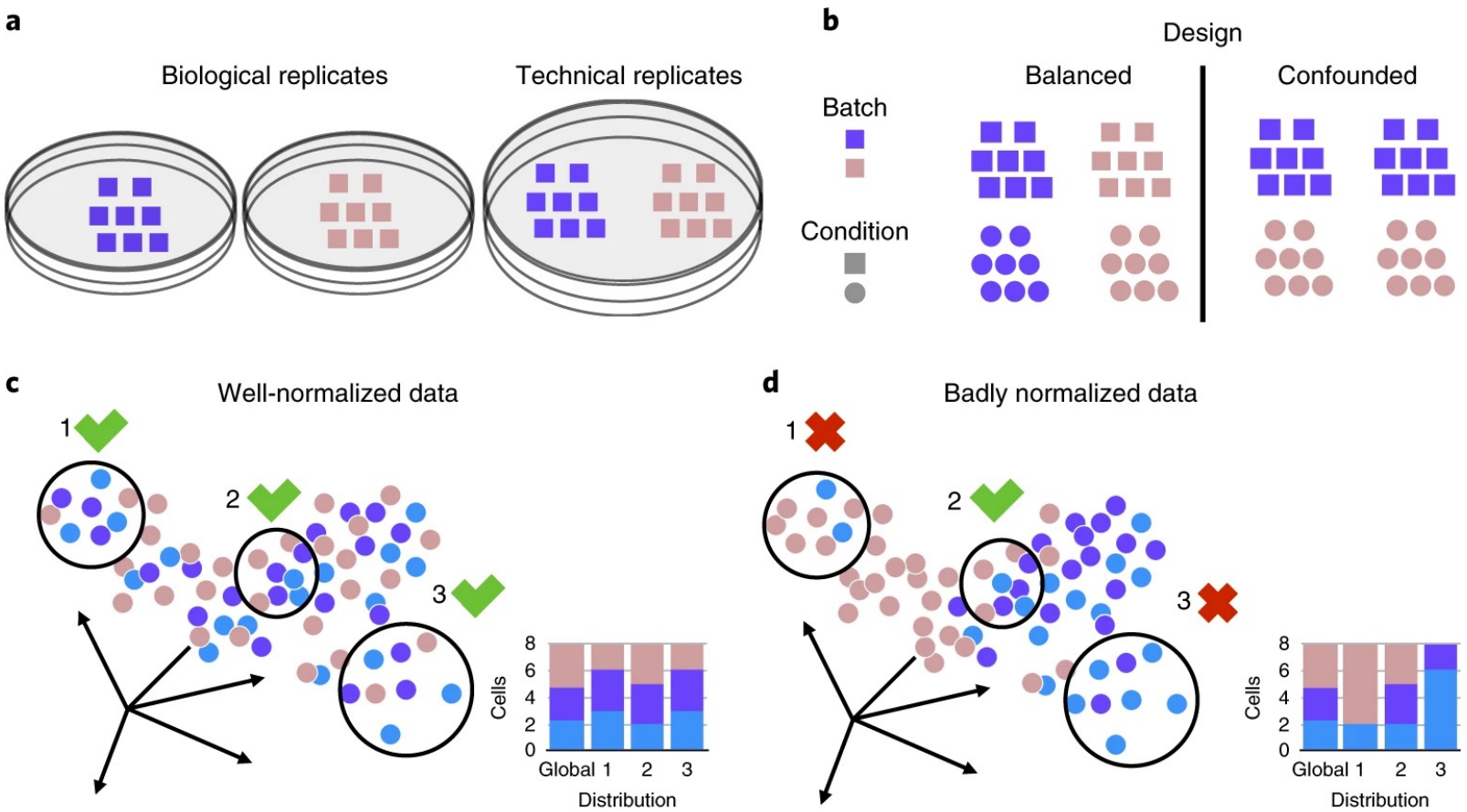
# Data quality control: a range of QC measures

- 16 QC measures
- 10 scRNA-seq libraries



# Data Normalisation - Motivation

- 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.

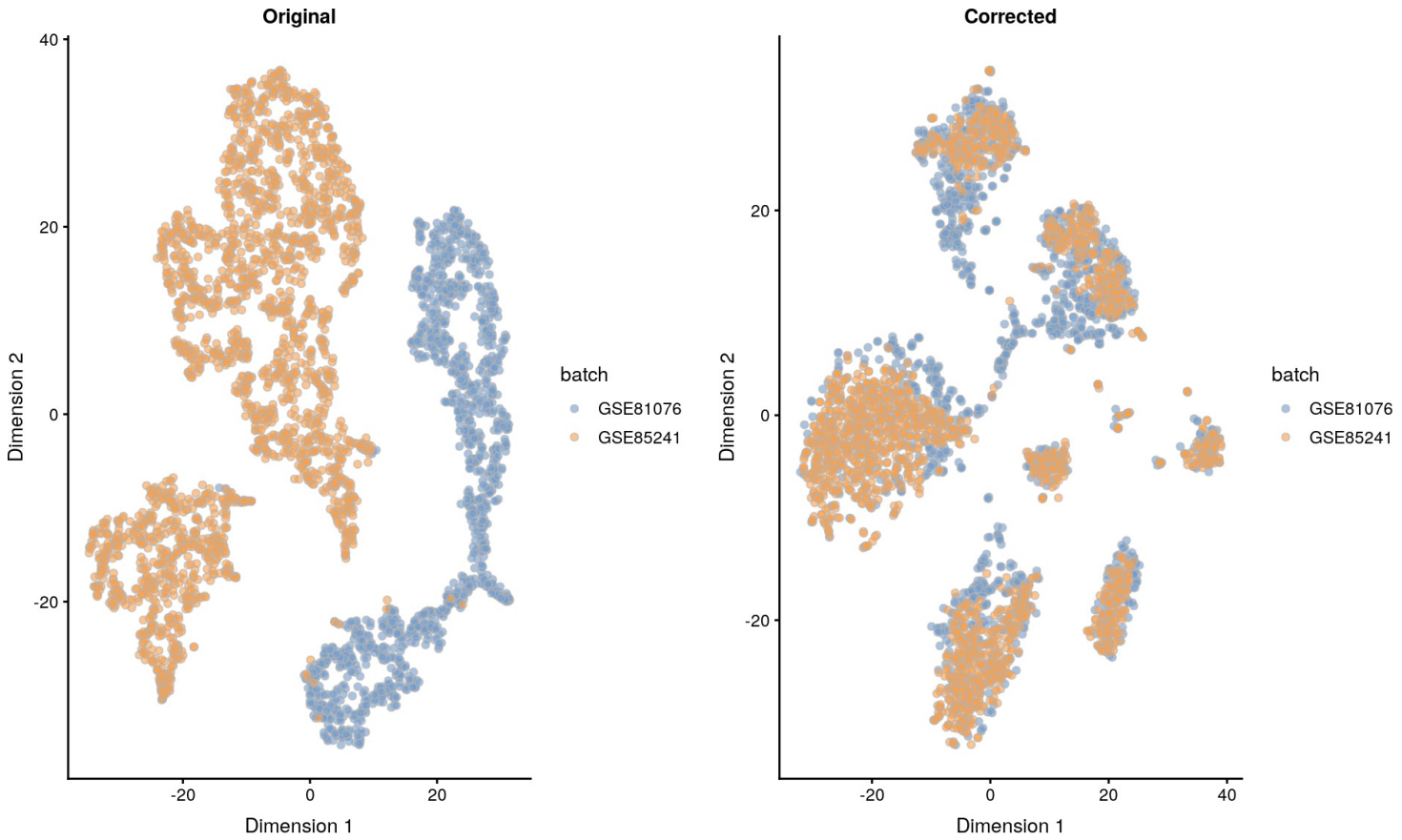



(Buttner et al, 2019)

# 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

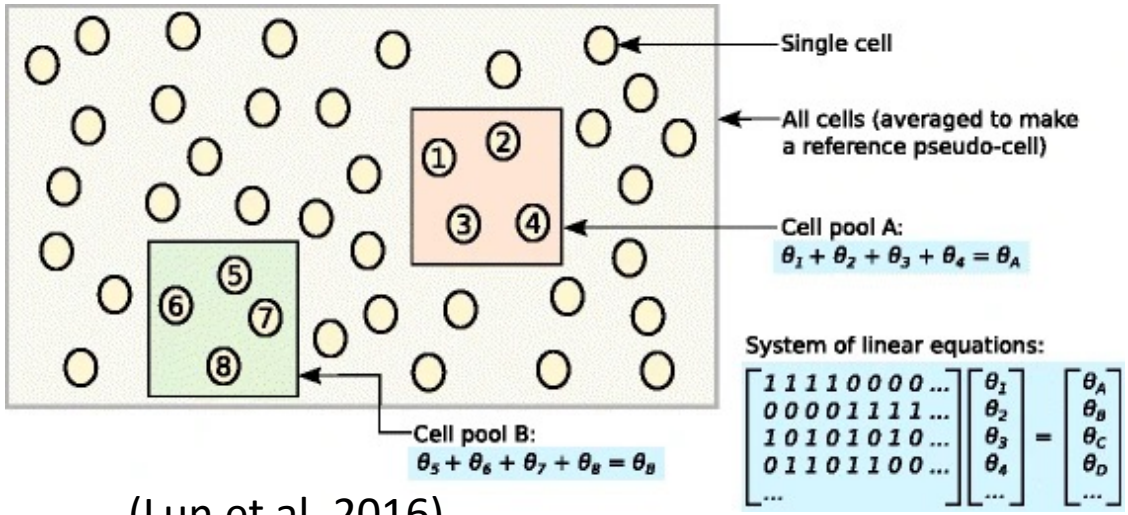
$$\text{Size factor} = \frac{\text{library size} * \text{nrCells}}{\text{Total library size}}$$

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

$\theta_A$ $\theta_B$

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

Scaling factor of cell 1

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

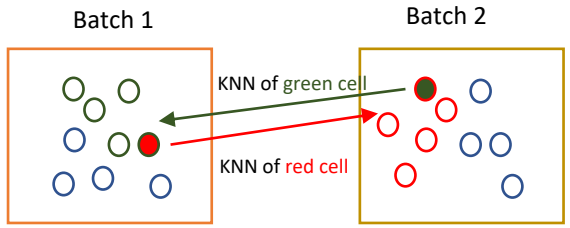
$V_{ik}$ is the sum of adjusted expression value across all cells in pool $V_k$ for gene *i*
$\lambda_{i0}$ is the expected transcript count and $\theta_j$ is the cell specific bias
$S_k$ is a pool of cell; $\theta_j \times t_j^{-1}$ is size factor for cell *j*

- 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 $\theta_j$

# Batch normalisation: Mutual nearest neighbour (MNN)

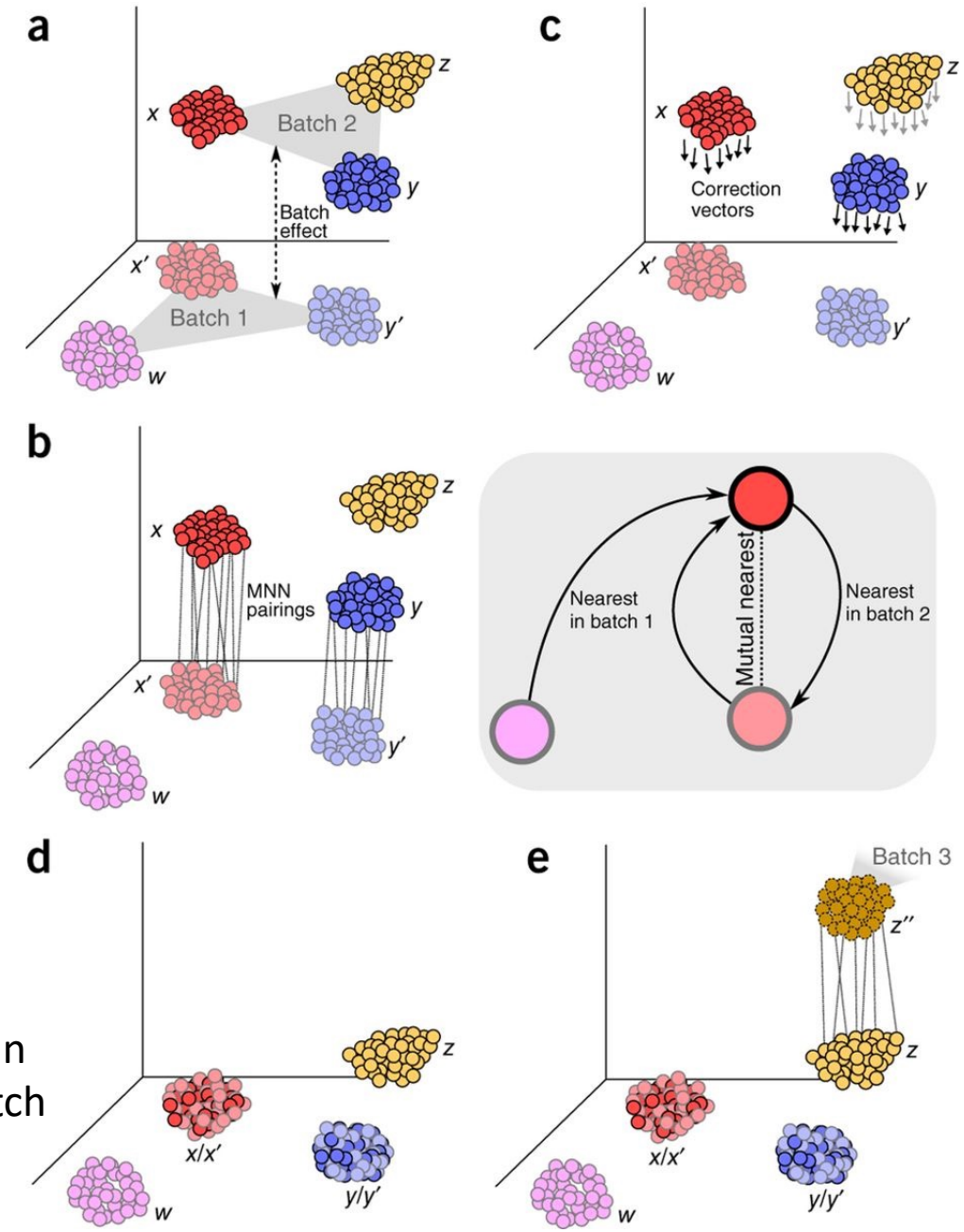
Three assumptions in MNN normalisation:

(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



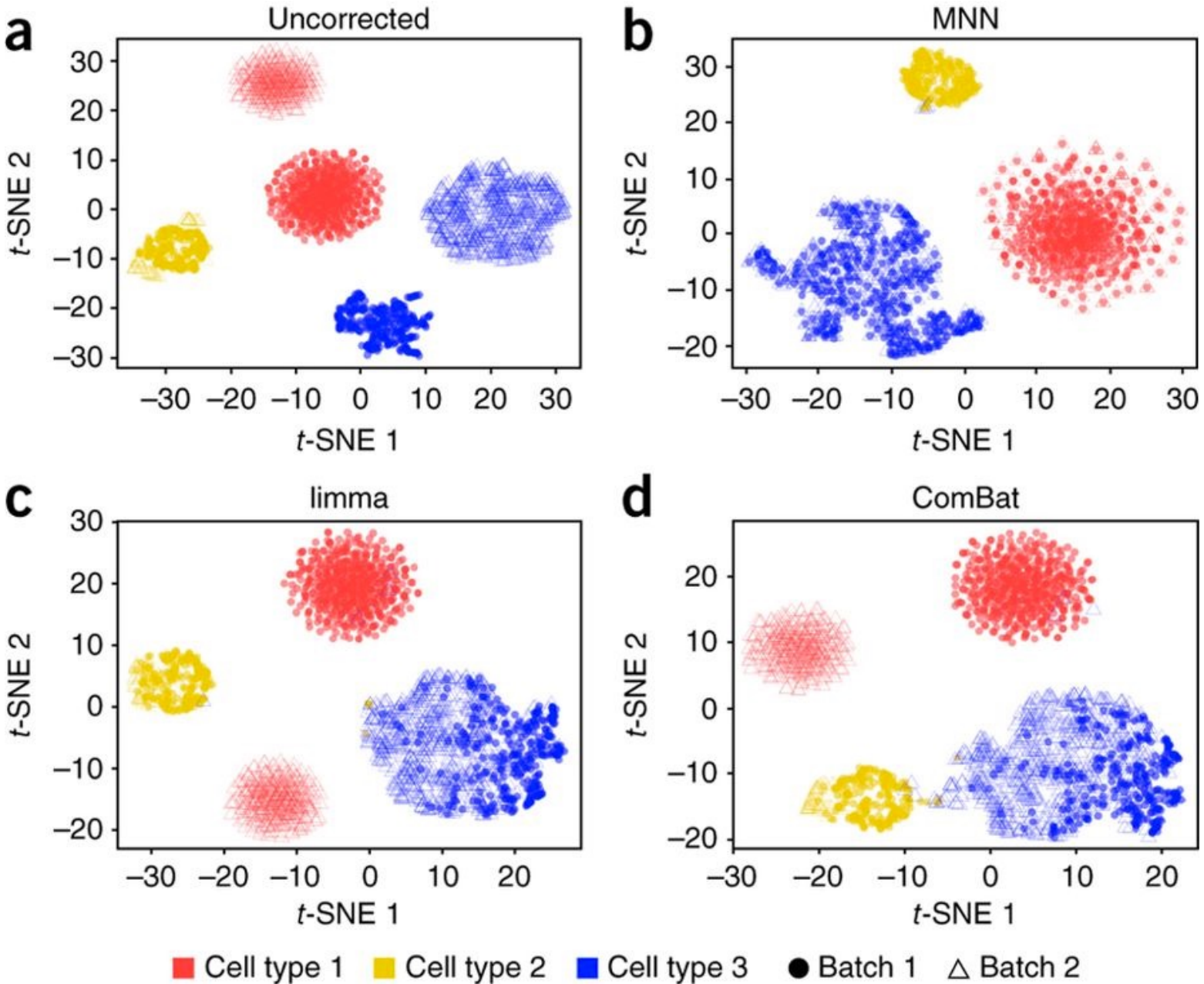

assume batch effects are mostly orthogonal to the biological manifold:
← batch effect: vertical
← biological manifold: horizontal

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:

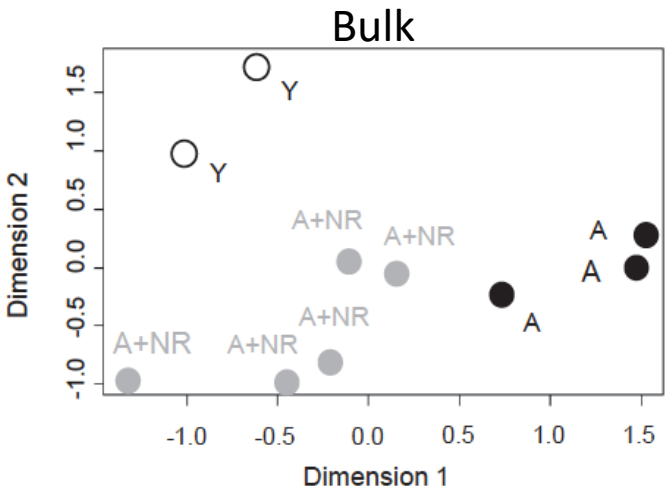
- Filters out noise
- Minimises curse of dimensionality
- Allows visualization with more separation of points
- Reduces computational load

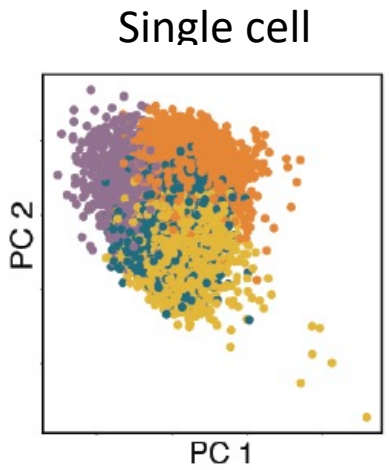
Linear approaches:

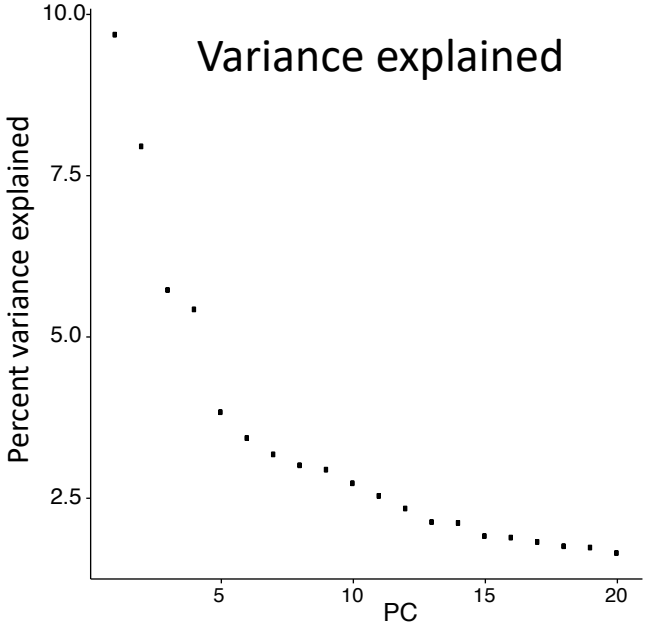
- PCA (Principal Component Analysis)
- ICA (Independent Component Analysis)
- NMF (Non-negative Matrix Factorization)

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

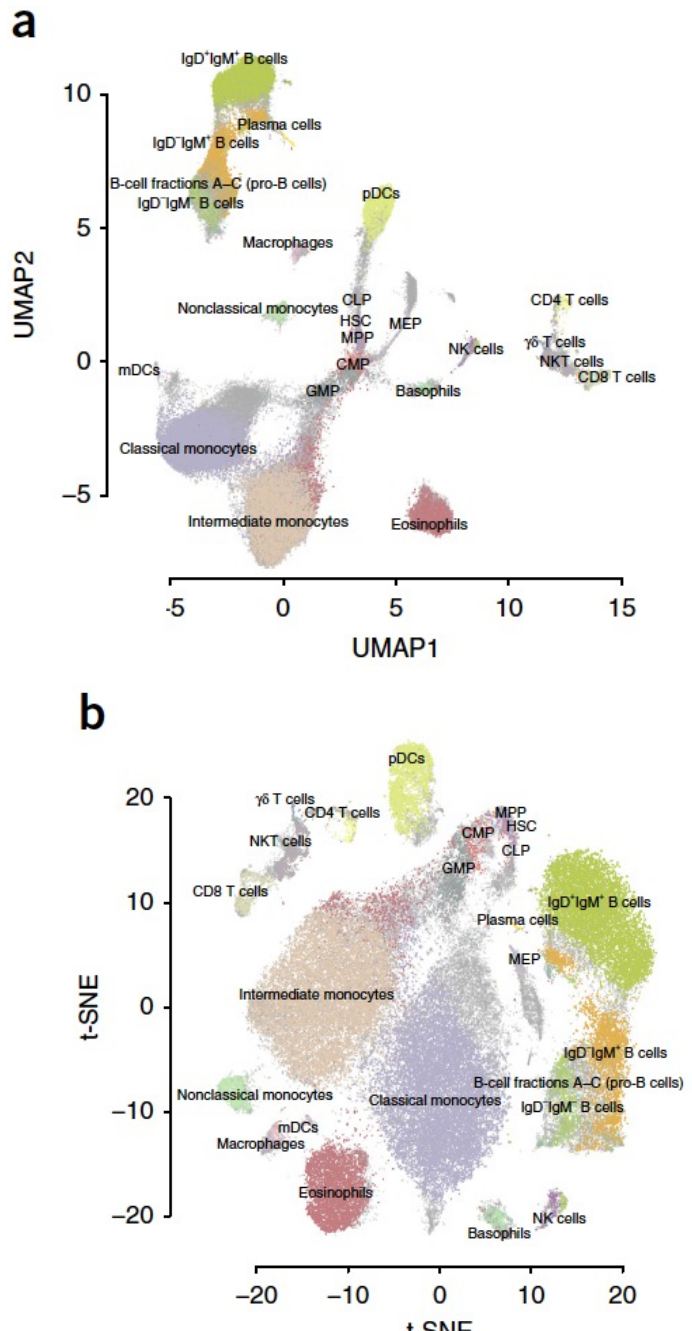


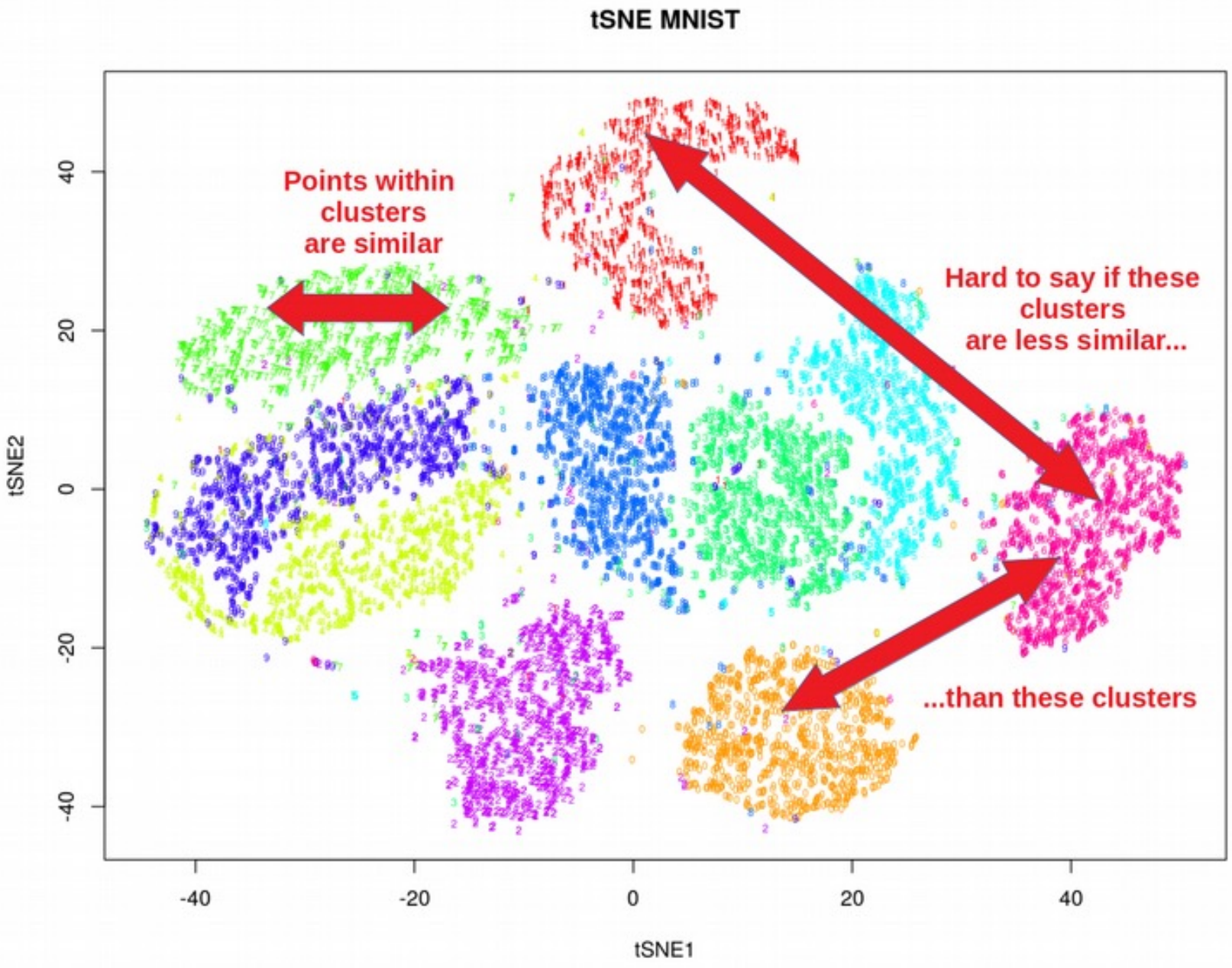



# Dimensionality reduction: nonlinear techniques

- 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 r 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



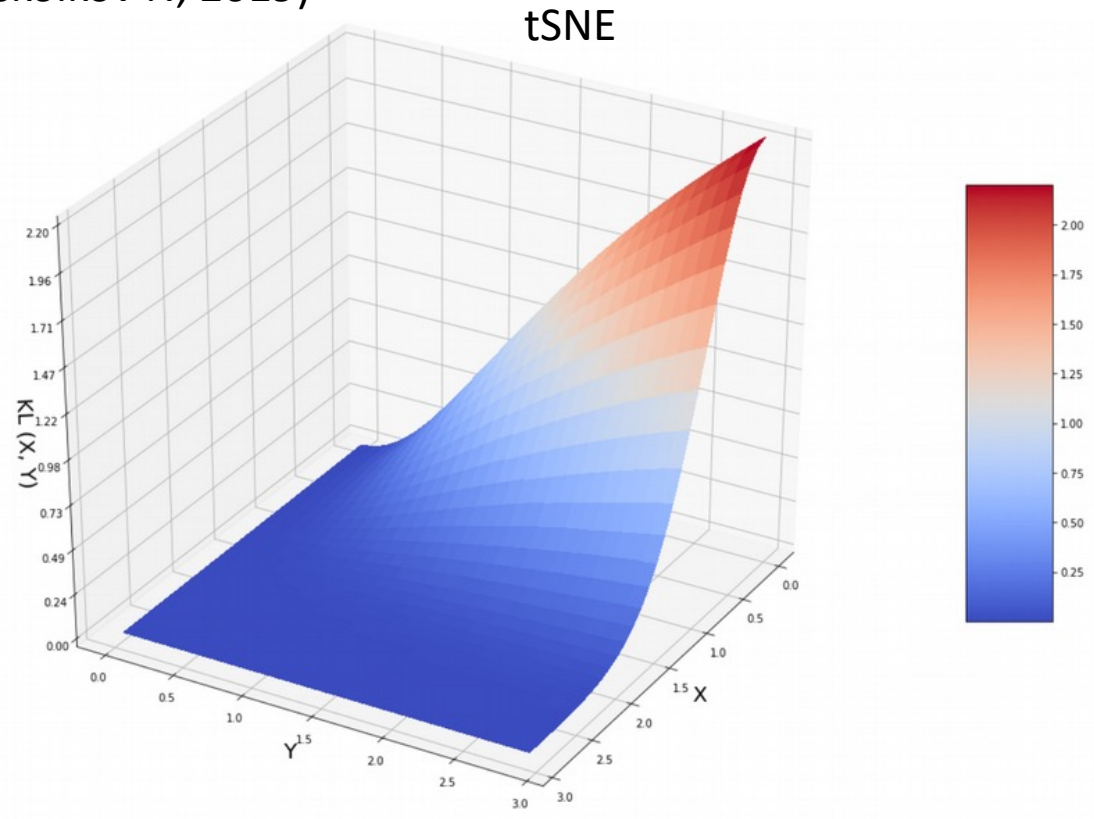
# Global vs local distance in low dimensional space




(Oskolkov N, 2019)

# tSNE does not preserve long distance - KL divergence

(Oskolkov N, 2019)




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

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

- 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:

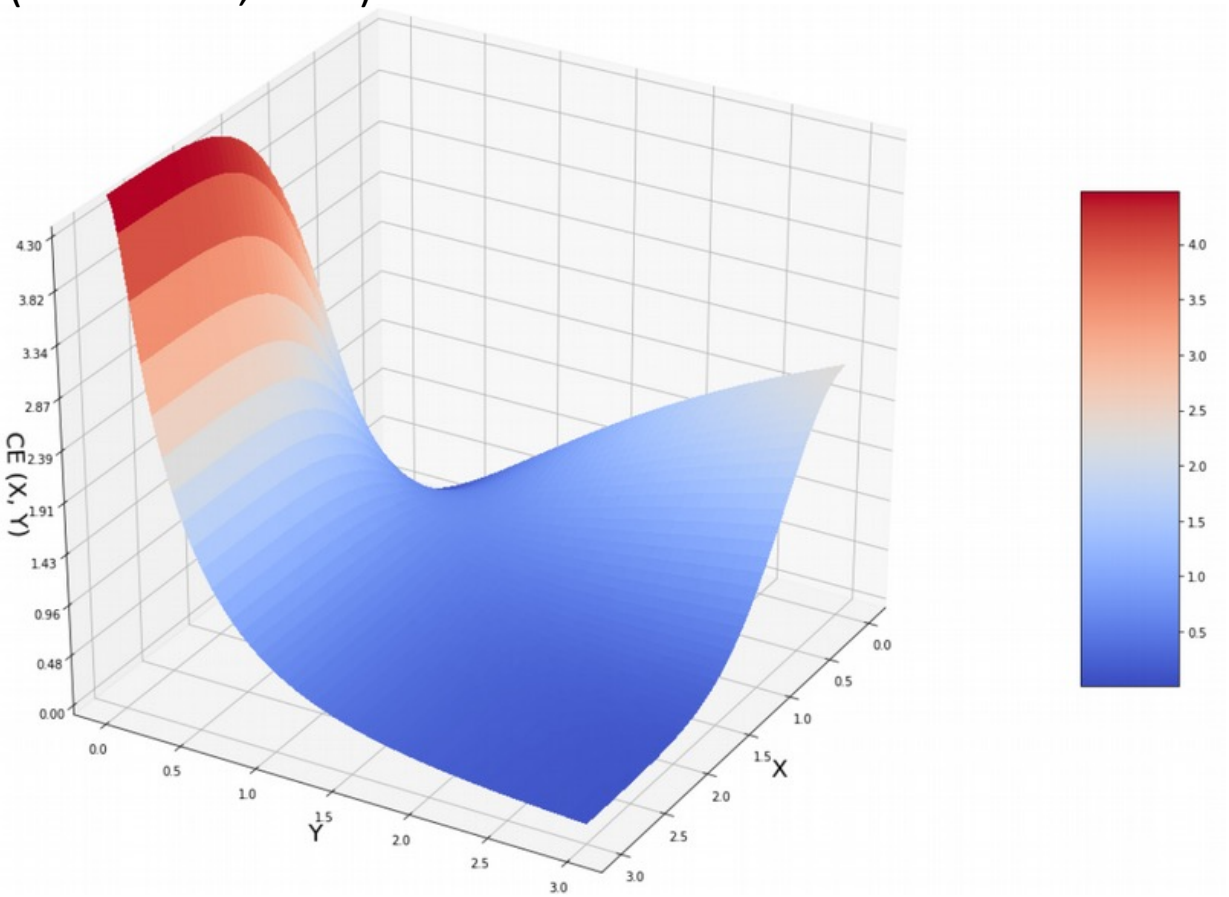
$$p_{j|i} = \frac{exp\left(\frac{-d(C_i,C_j)^2}{2\sigma_i^2}\right)}{\Sigma_{k\neq i}\ exp\left(\frac{-d(C_i,C_k)^2}{2\sigma_i^2}\right)}$$

- 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: $q_{ij} = \frac{\left(1+\|y_i-y_j\|^2\right)^{-1}}{\Sigma_{k\neq m}\ (1+\|y_k-y_m\|^2)^{-1}}$, $y_i$ 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)*

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

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

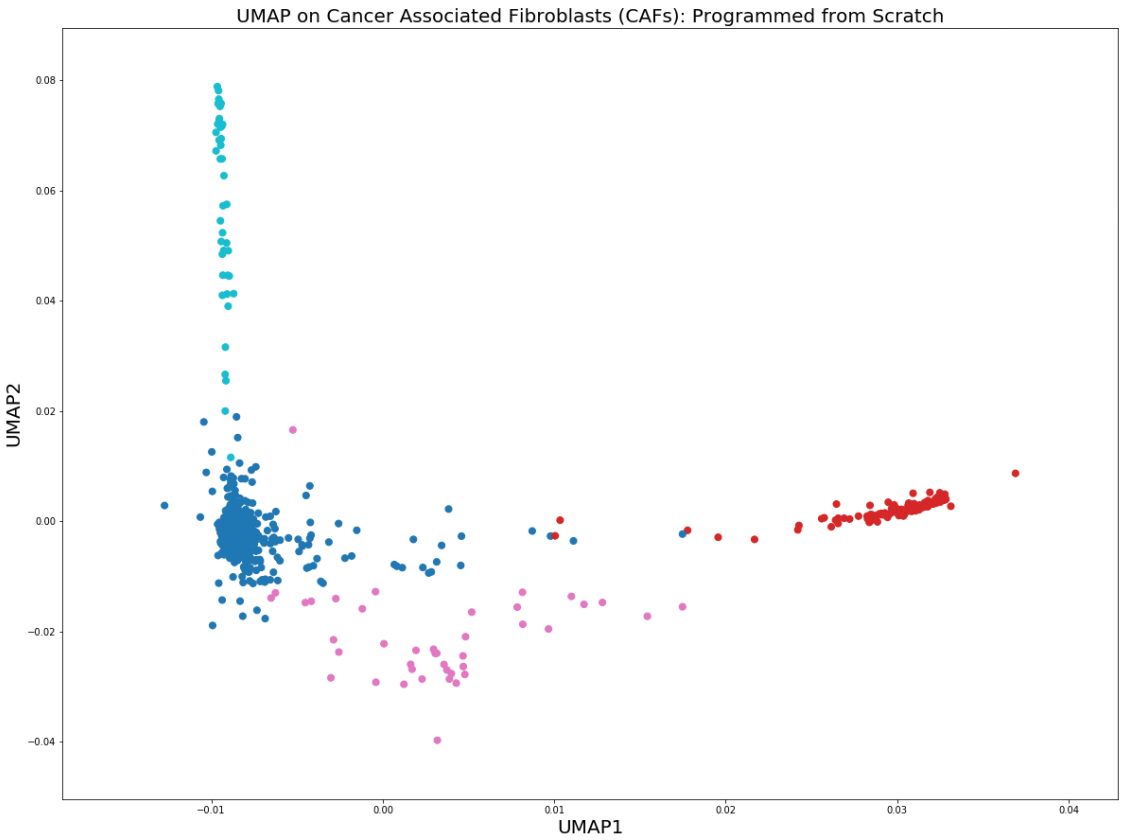
tSNE: $KL(X,Y) \approx -P(X)\log Q(Y) = e^{-X^2}\log\left(1+Y^2\right)$

# 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)

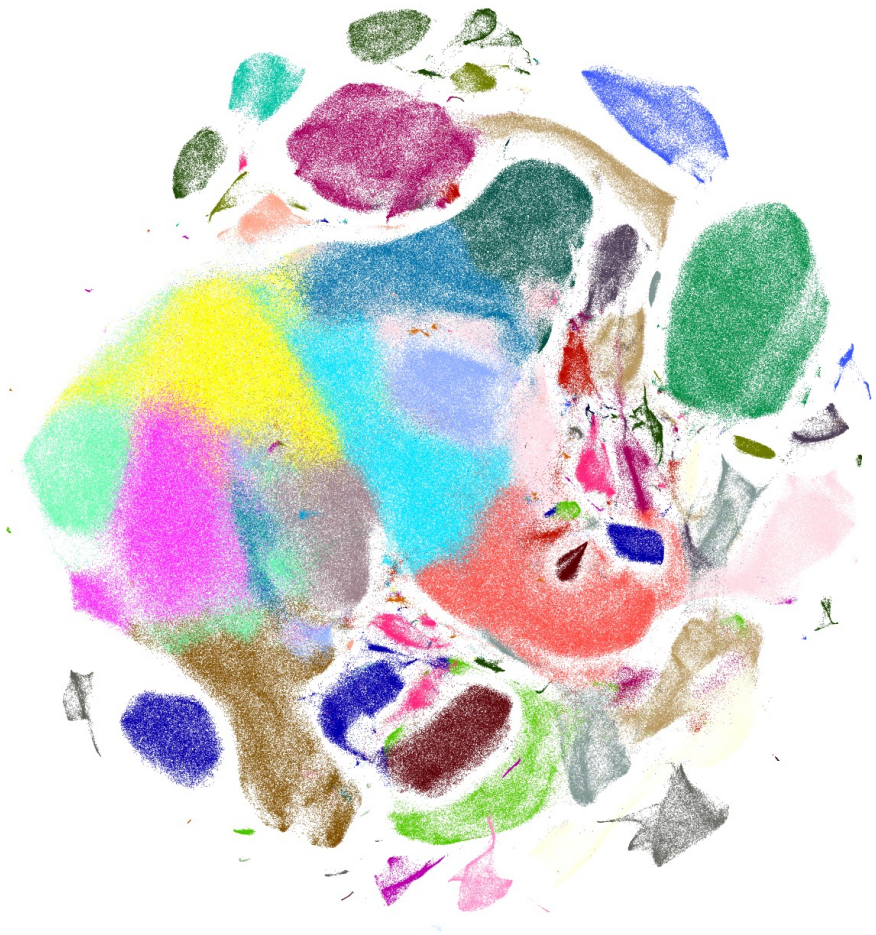
- UMAP proceeds by iteratively applying attractive (among edges) and repulsive forces (among vertices) at each edge or vertex. Convergence is guaranteed by slowly decreasing the attractive and repulsive forces of the neighbour graph.

- 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



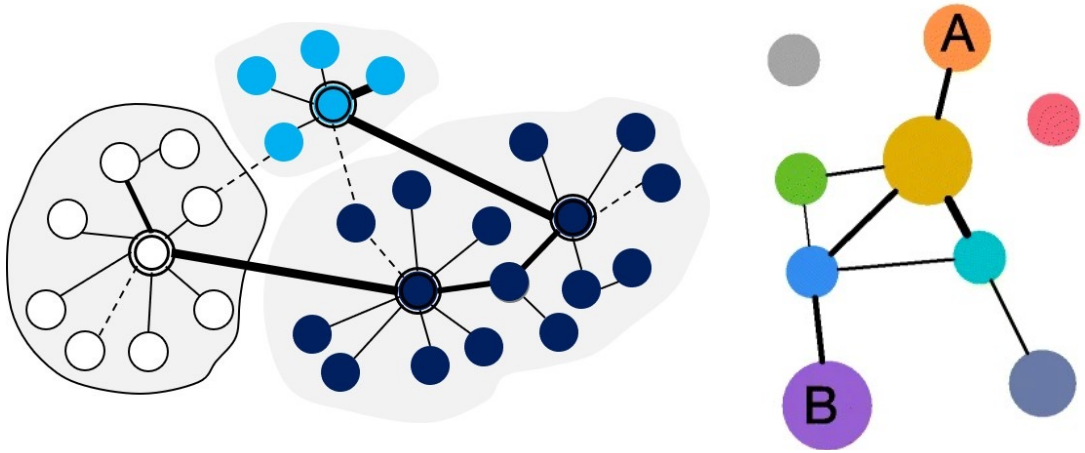
Clustering in scRNAseq is a data-driven way to find cell (sub)types at single-cell resolution

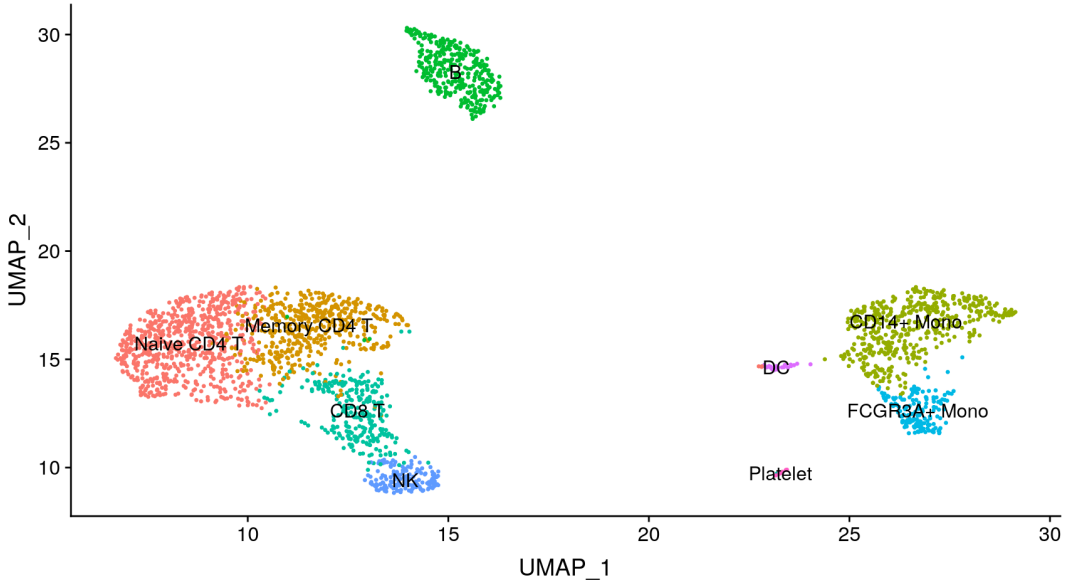
# Graph-based Clustering

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

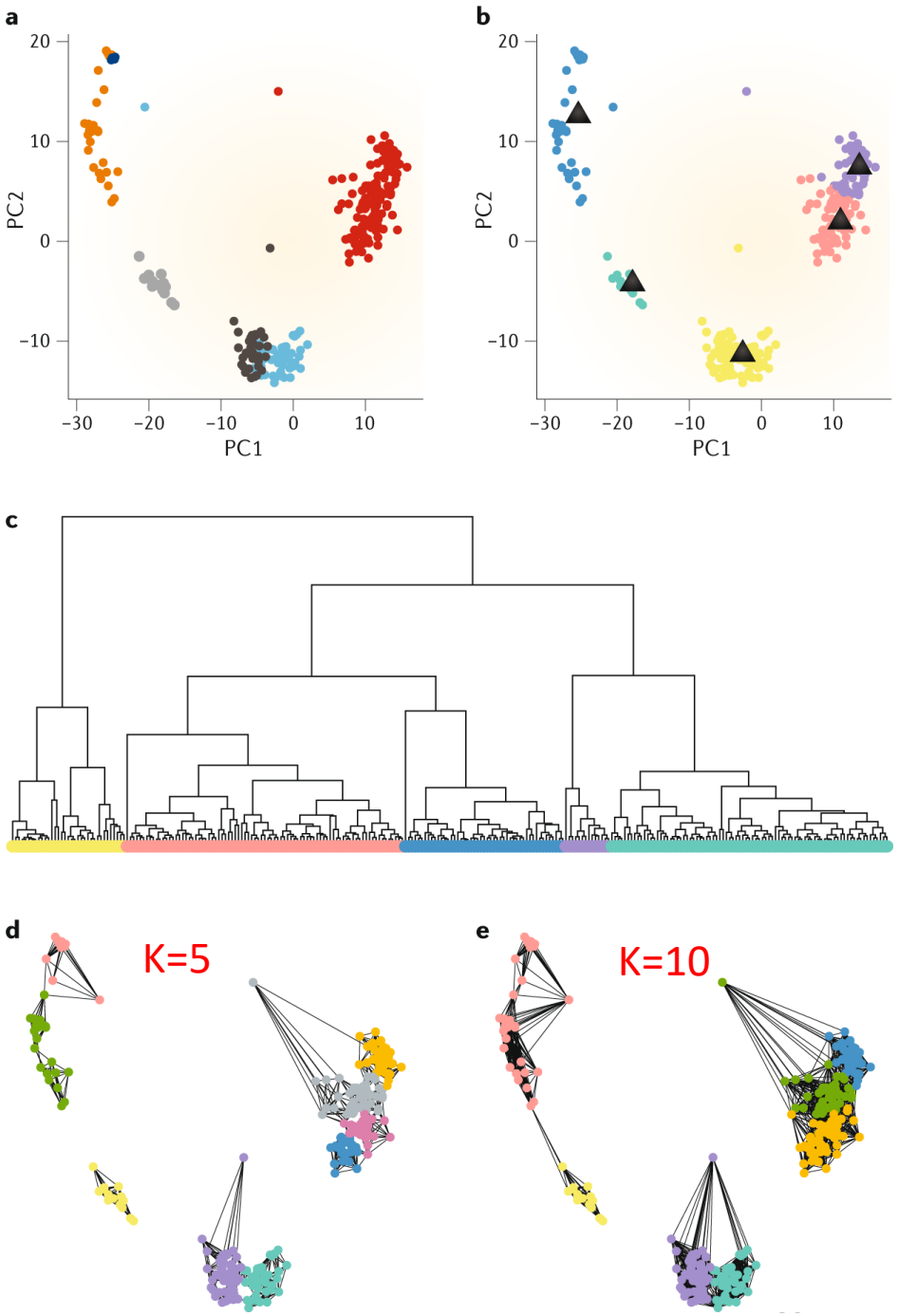


# Graph-based Clustering

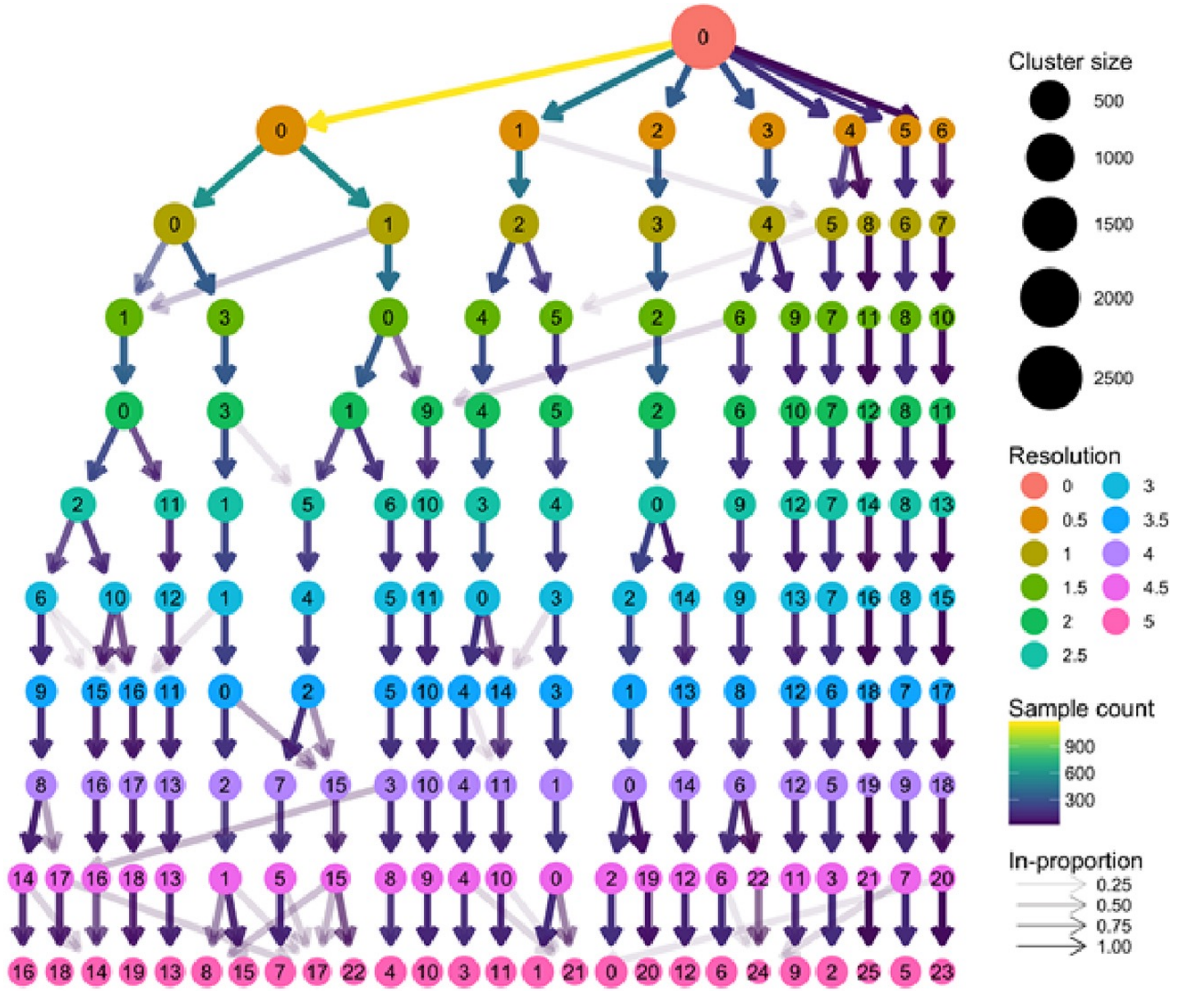
- Build shared-nearest-neighbour graph connecting the cells and finds tightly connected communities

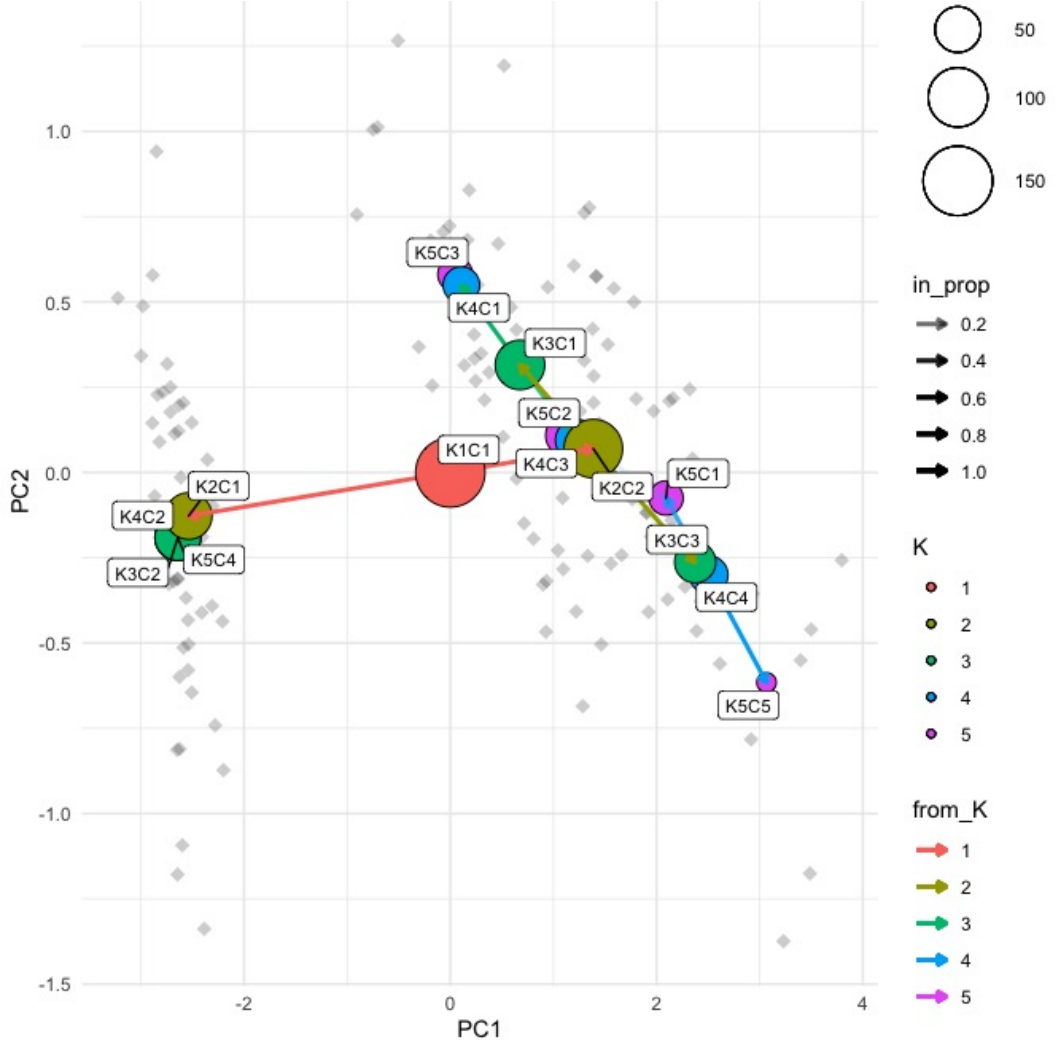
- Increasing the number of neighbours when constructing the cell–cell graph indirectly decreases the resolution of graph-based clustering



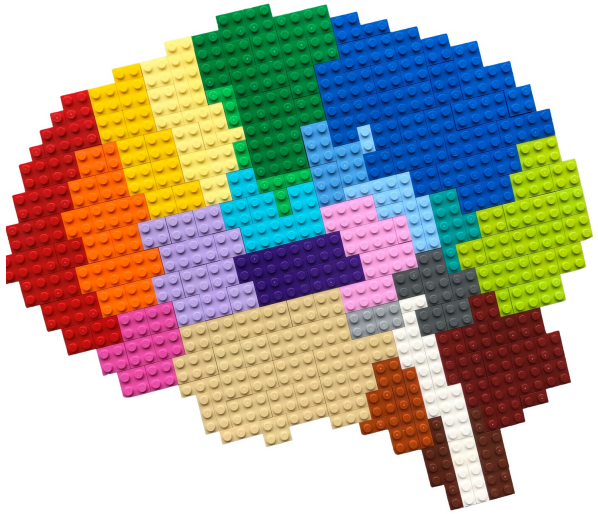

Nature Reviews Genetics, 20, (2019)

# Visualise clustering results

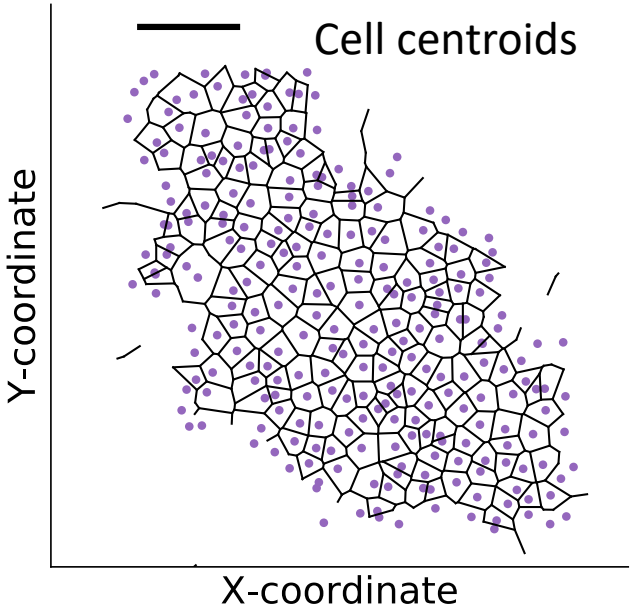


# Spatial transcriptomics approach

| | Bulk | Single cell | Spatial |
|---|---|---|---|
| Lego: (@boxia) |  |  |  |
| Fruit salad: (@LGMartelotto) |  |  |  |

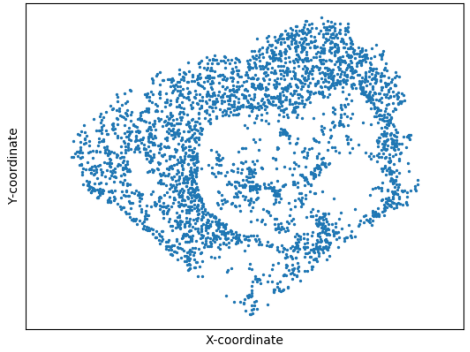
# Spatial Transcriptomics Data (seqFISH): expression + location




(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

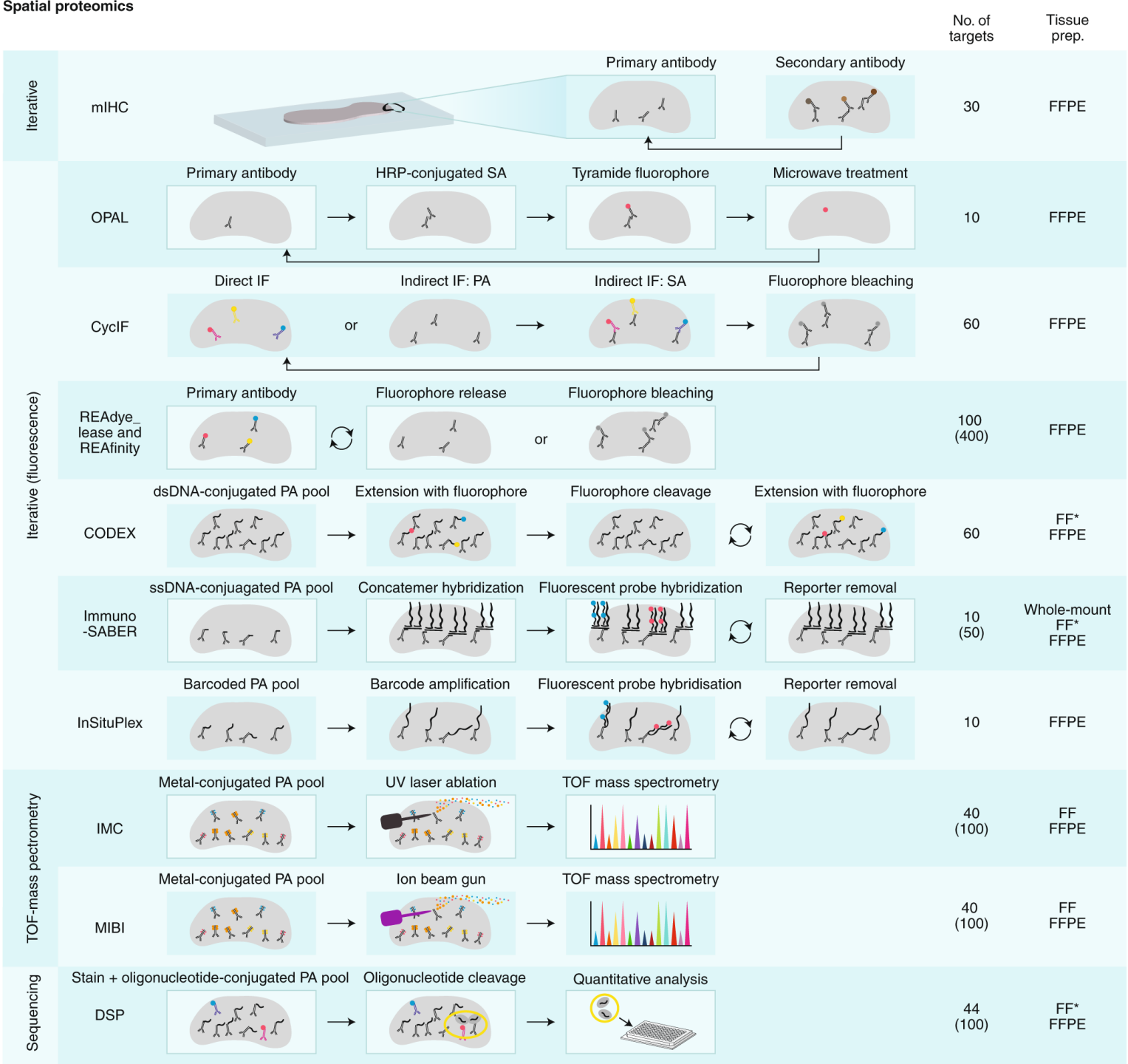



Example of seqFISH RNA in a cell: 3247 genes

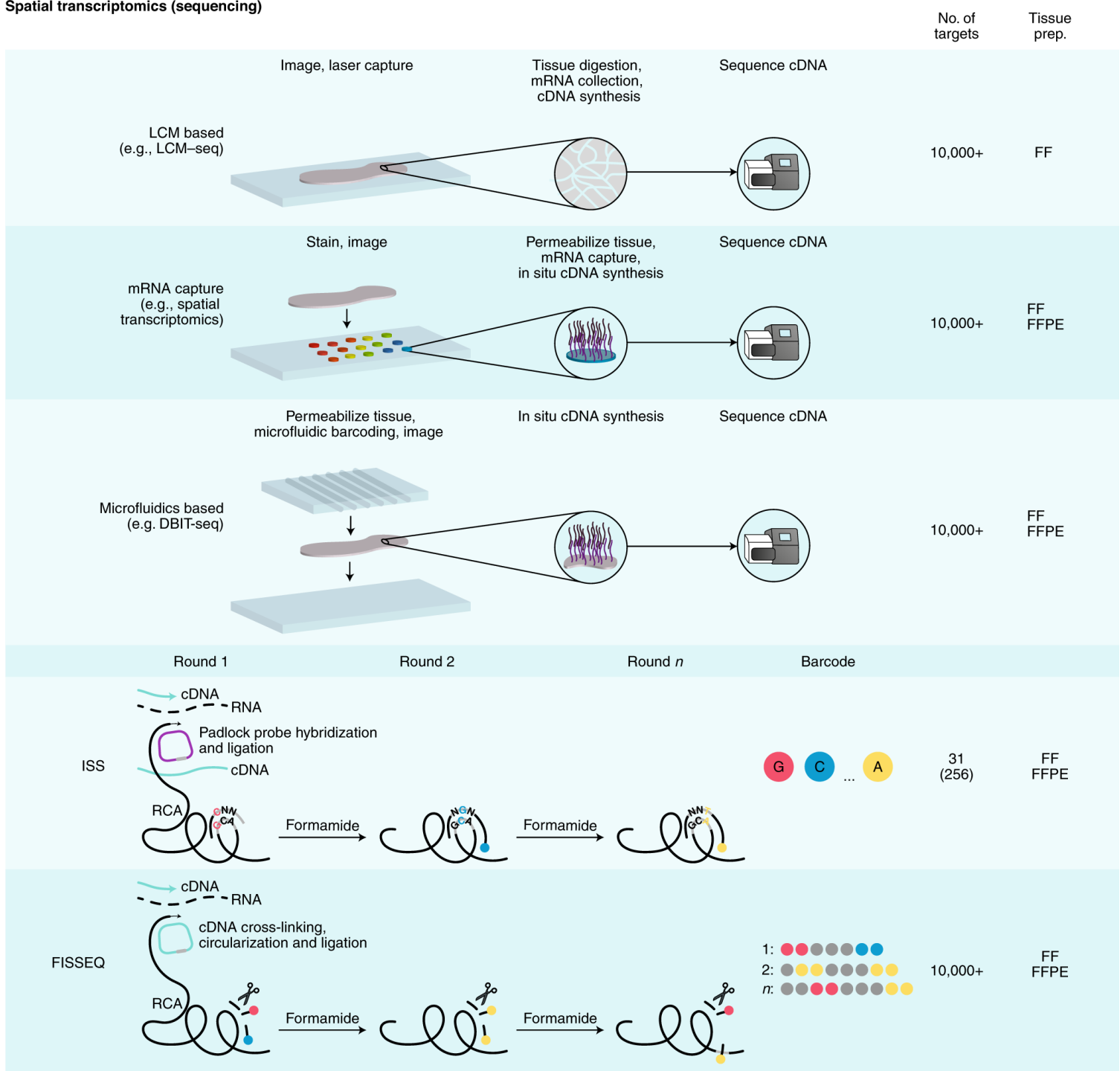
| Gene ID | 1 | 19 | 23 | 44 | 53 | 57 | 63 | 70 | 71 | 72 | ... |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 653.00 | 675.24 | 687.21 | 733.85 | 615.16 | 663.99 | 611.06 | 669.65 | 638.03 | 601.10 | ... |
| **1** | 434.34 | 428.89 | 479.06 | 472.43 | 469.95 | 464.81 | 443.74 | 417.42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

Coordinates

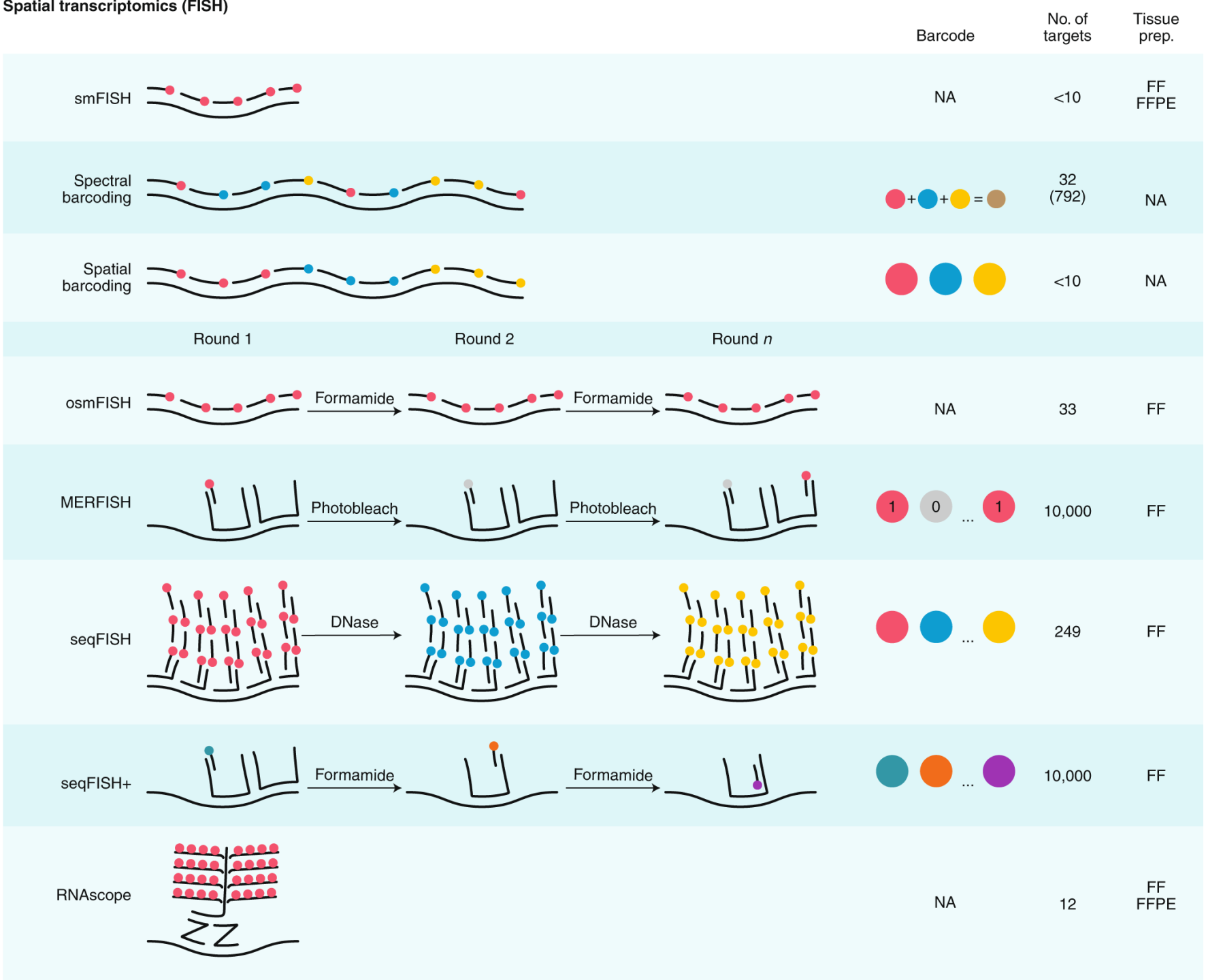
**Spatial proteomics**



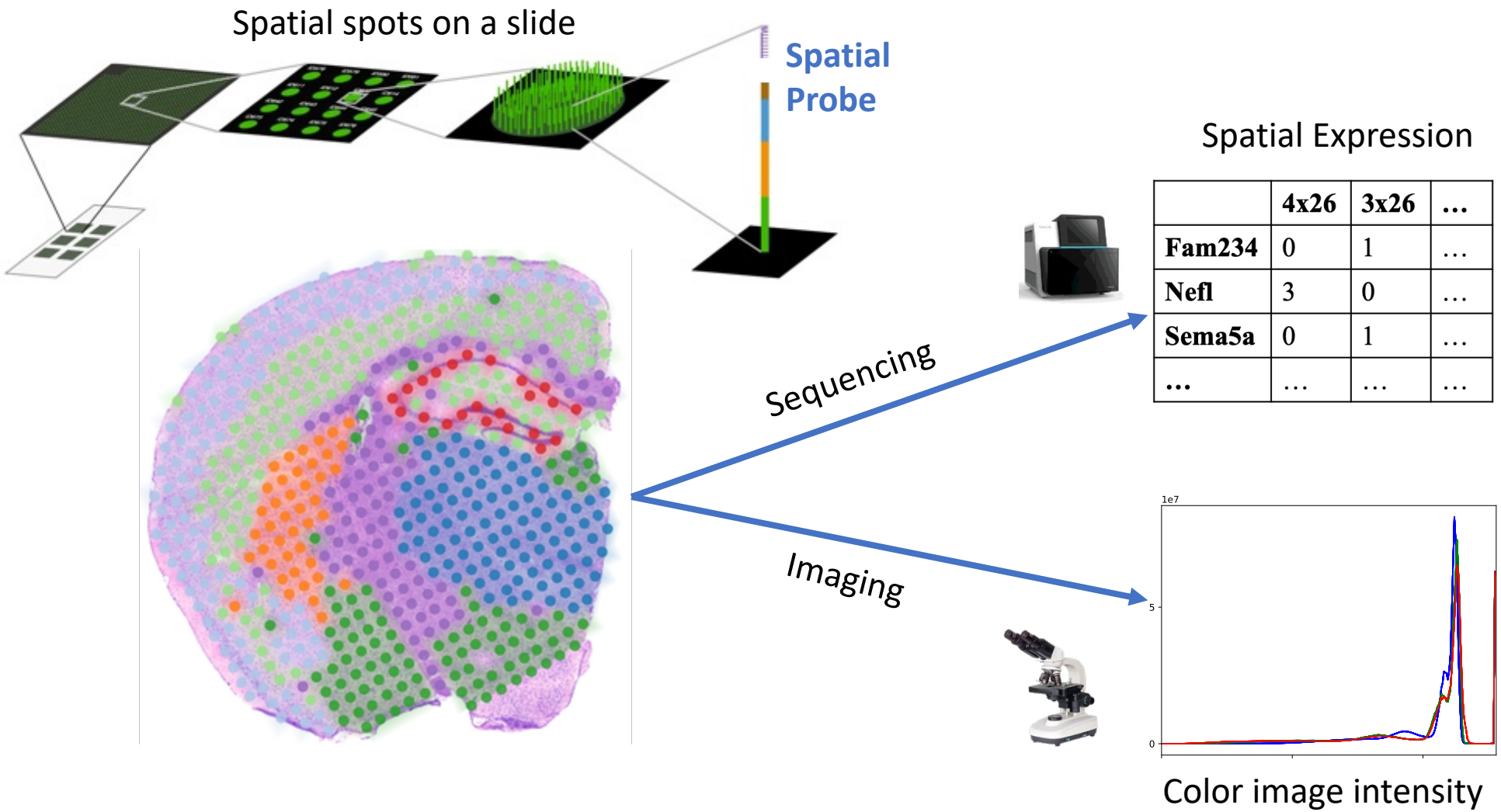
**Spatial transcriptomics (sequencing)**



**Spatial transcriptomics (FISH)**



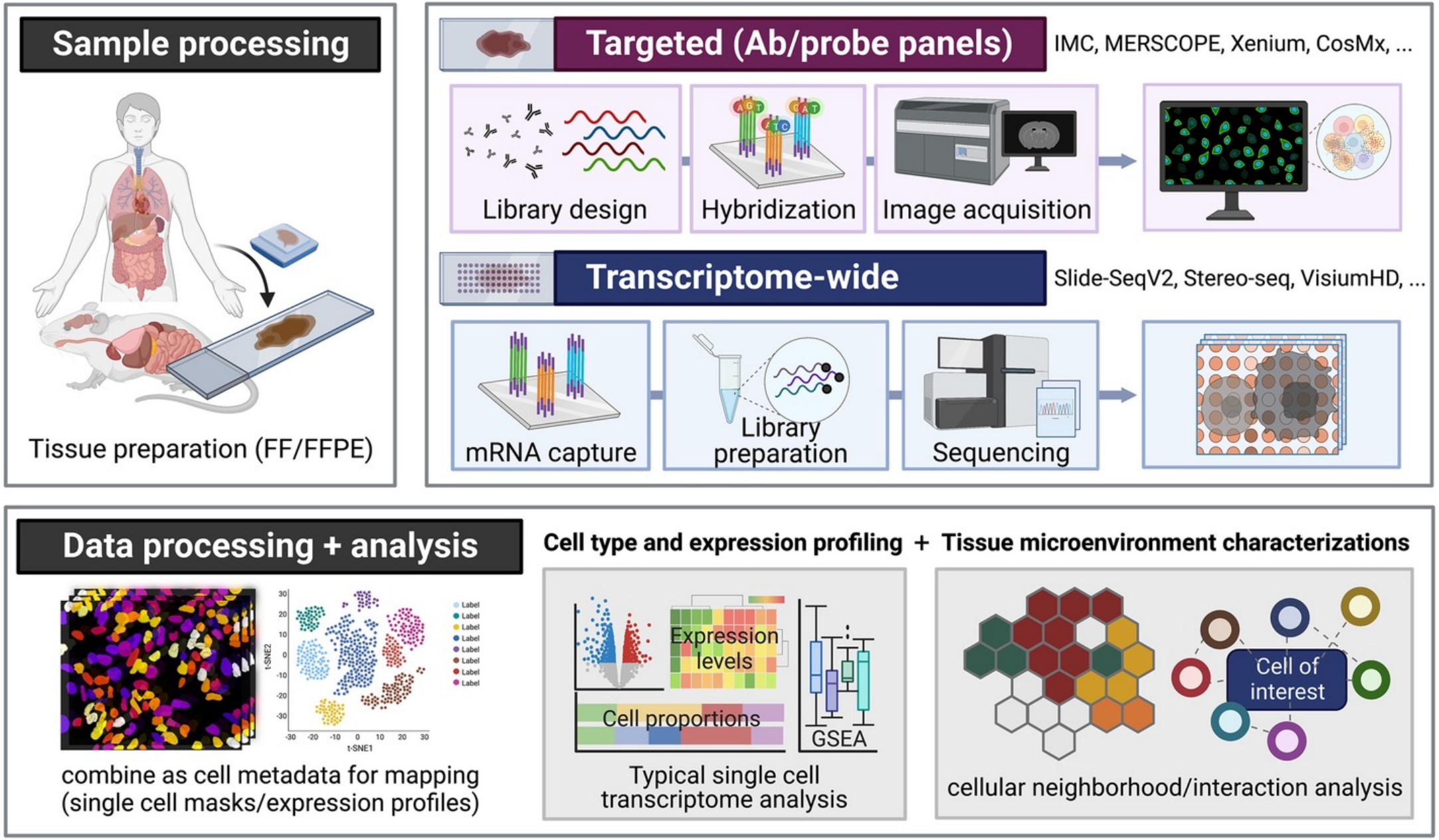
# Spatial transcriptomics adds spatial dimension and tissue morphology




| | 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.

# Analysis landscape




(Park et al, 2022)

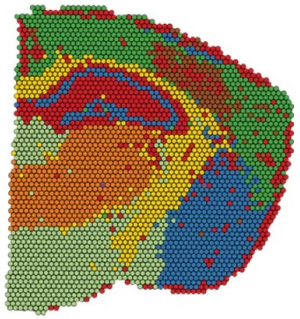
# Lecture 2: Defining Cell Types

# Module 2 – Part 2: Defining Cell Types

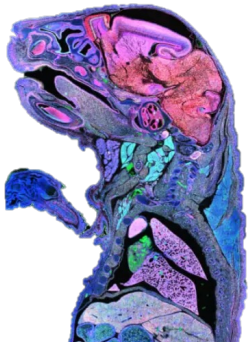
Andrew Causer

Seurat v5

R package



10X Visium



10X Xenium

# Module 2 – Part 2: Overview

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

## 1. Data Pre-Processing

- General QC – remove low quality spots/cells and genes
- Data Normalisation

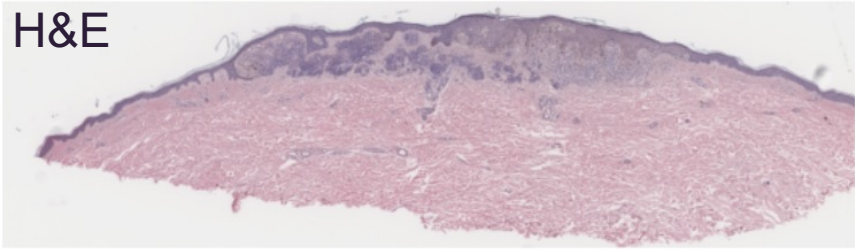
## 2. Clustering and Cell Typing

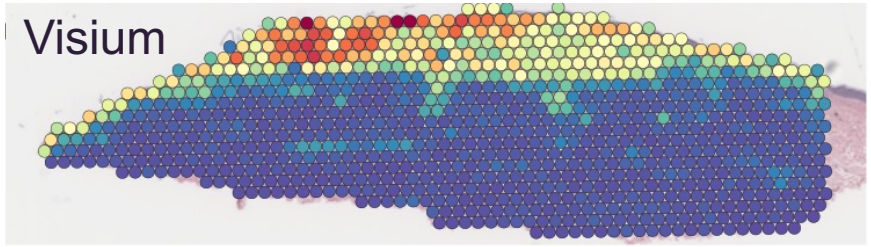
- Perform Unsupervised Clustering – group similar spots/cells together based on transcriptome
- Cluster Annotation – use marker genes to cell type clusters

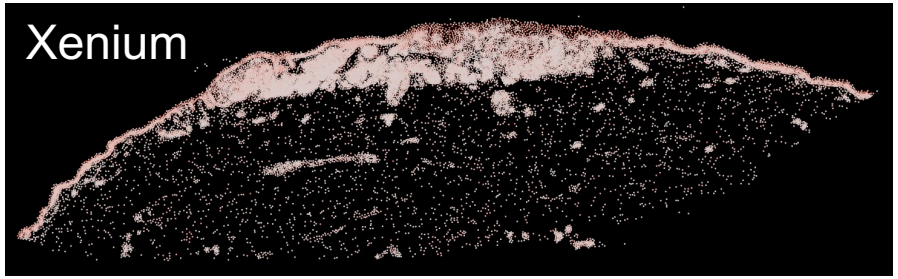
## 3. Spot Deconvolution and Single-Cell Label Transfer

- Visium Spot Deconvolution – infer the cellular composition of each spot
- Xenium Label Transfer – matches cells from a reference dataset based on genetic similarities

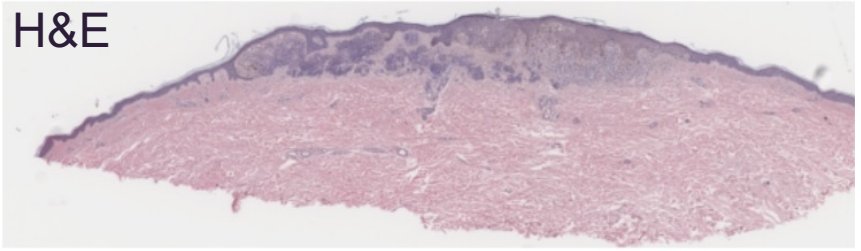
# Datasets – Melanoma (Skin)

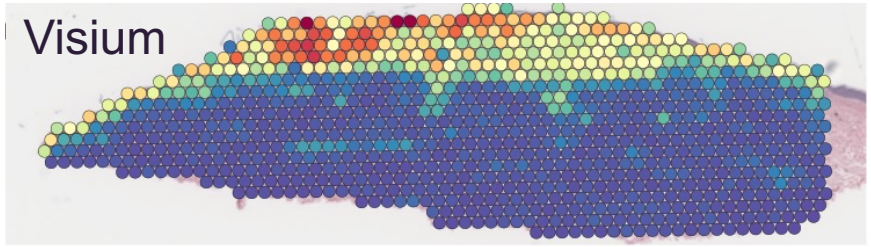


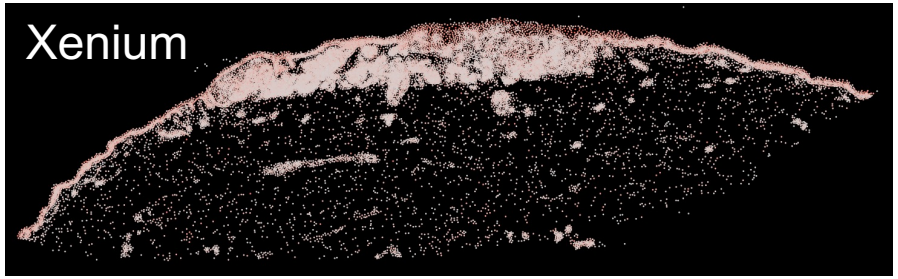



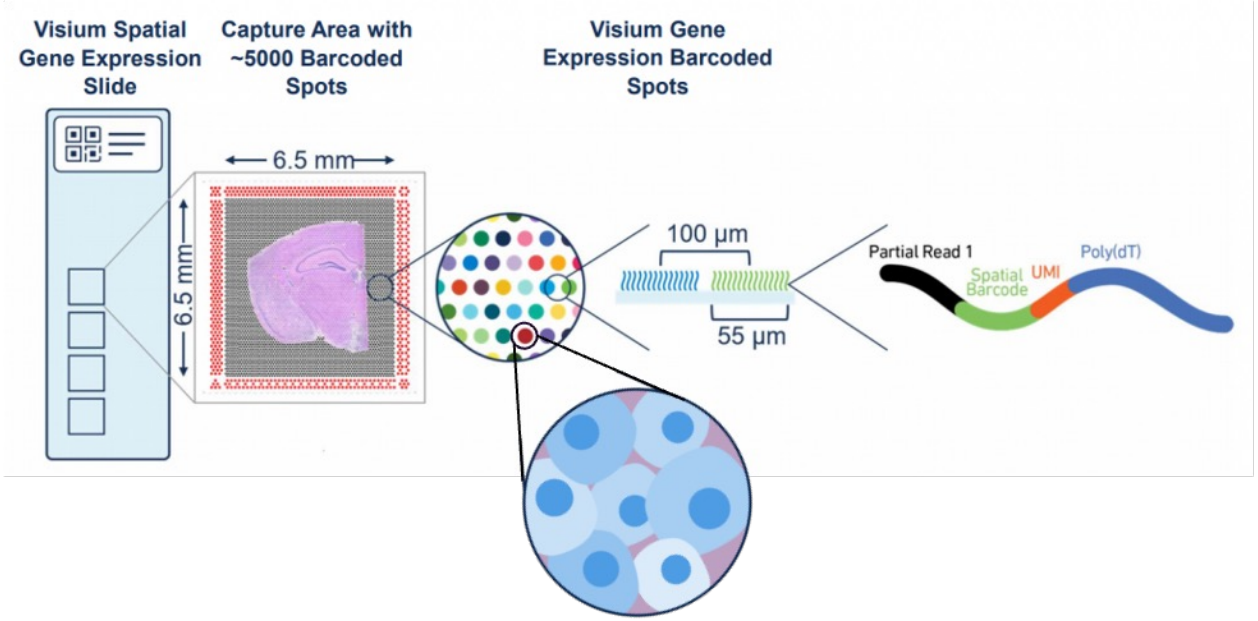
# Datasets – Melanoma (Skin)



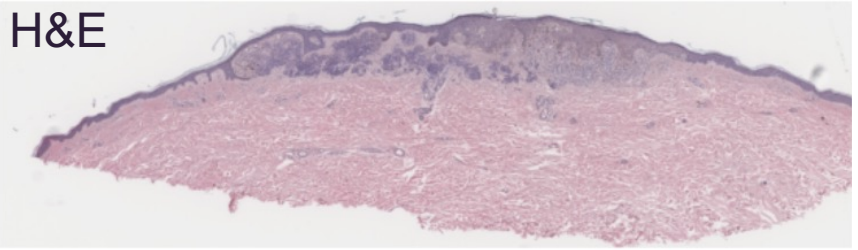



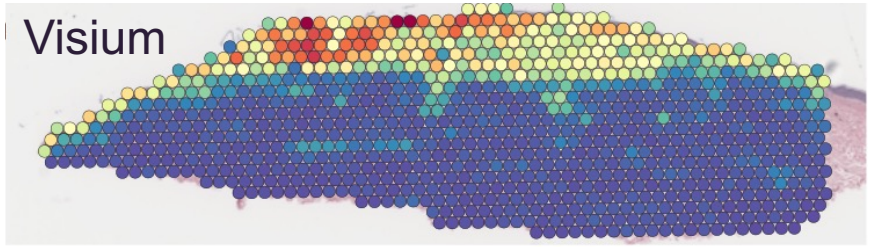

Visium

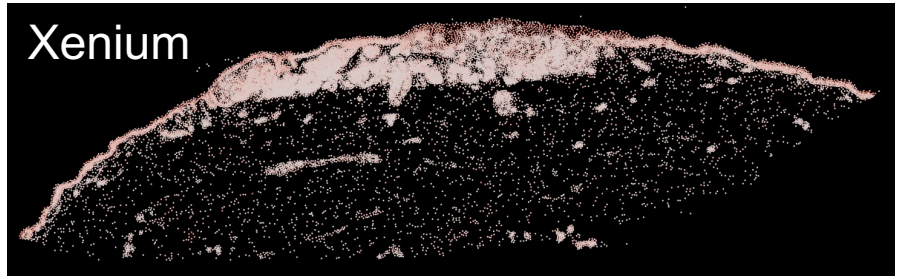


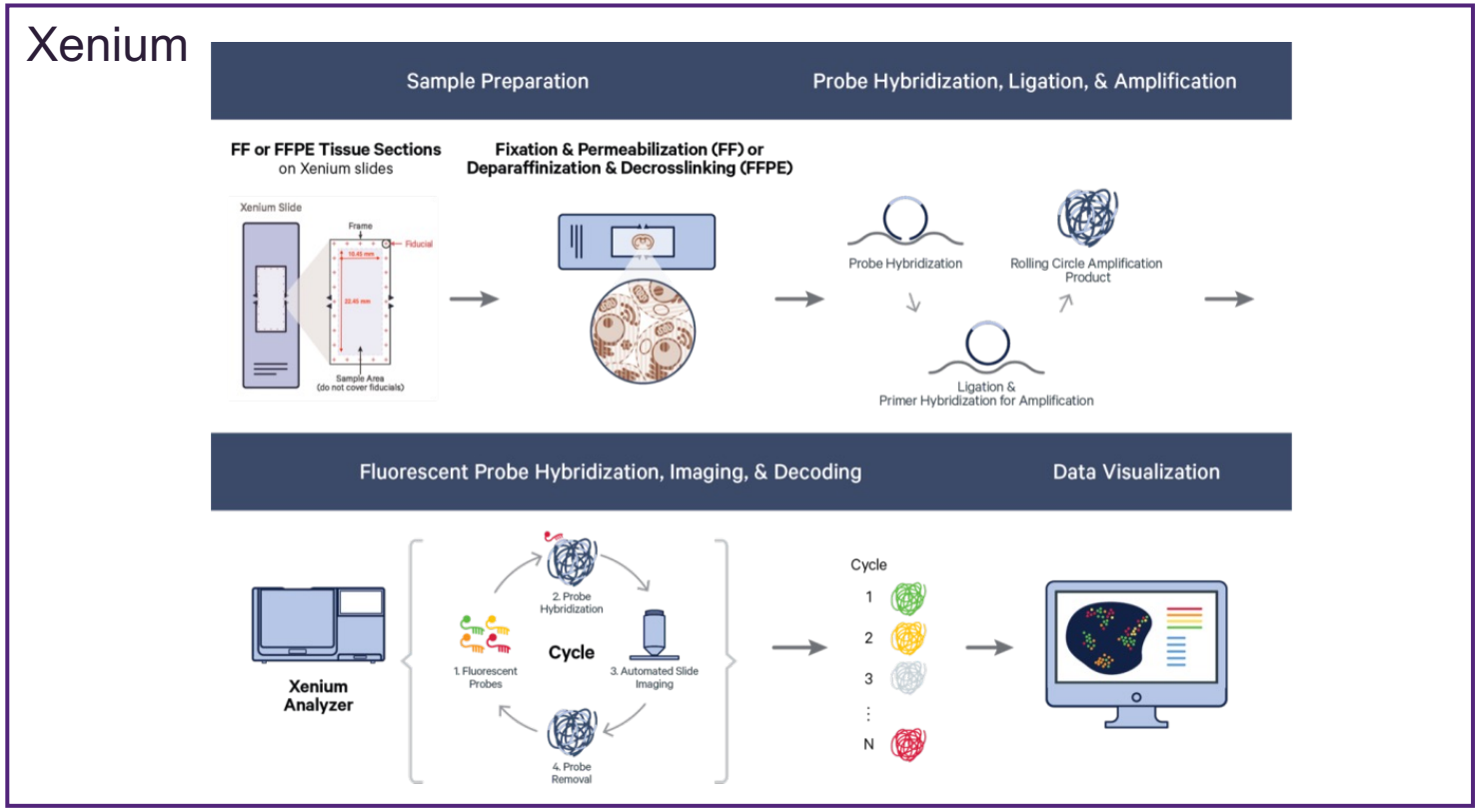

1 spot = ~2-10 cells

# Datasets – Melanoma (Skin)

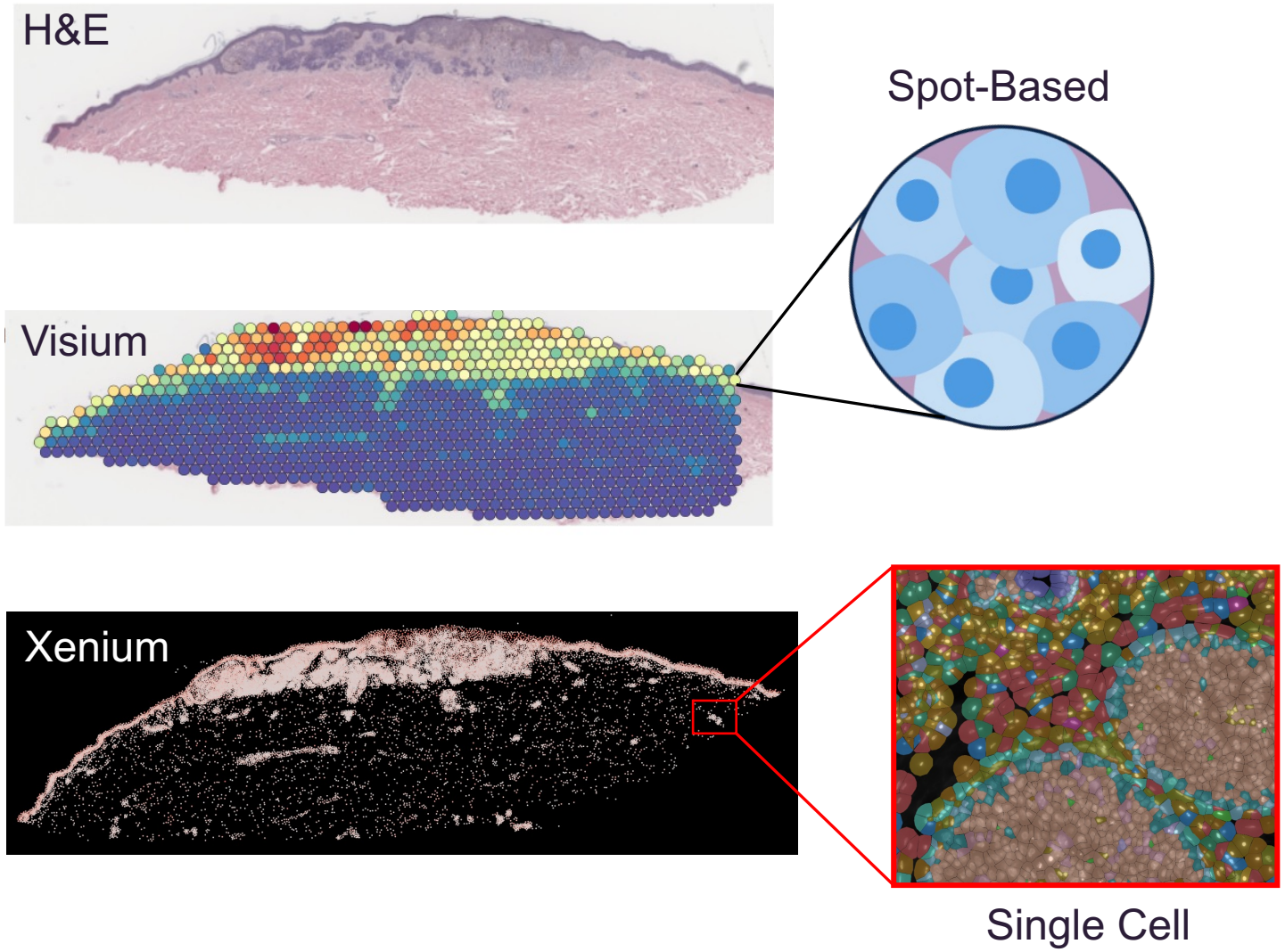


# Datasets – Melanoma (Skin)

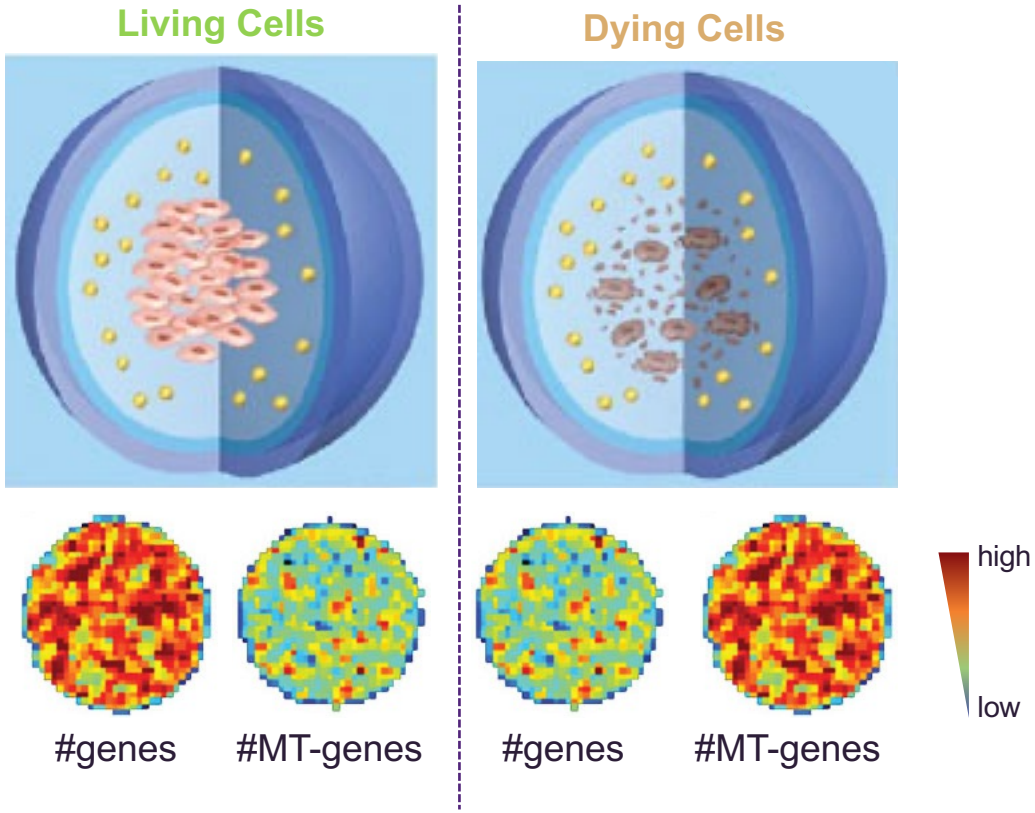



| # Data Points | # Genes |
|---|---|
| 923 spots | 18,085 |

| # Data Points | # Genes |
|---|---|
| 21,596 cells | 260 |

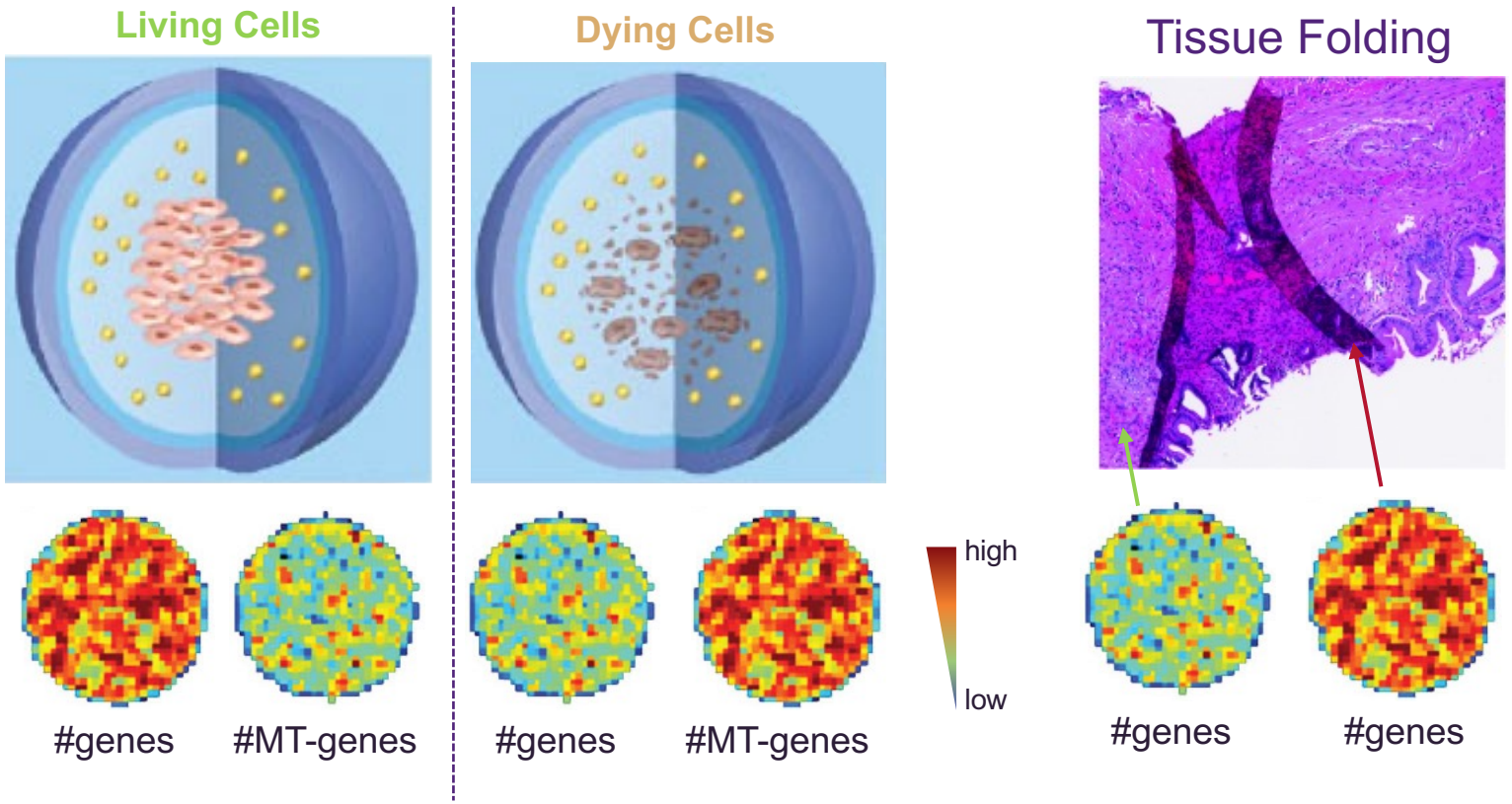
# 1. Data QC and Normalisation

*Factors of Technical Noise*



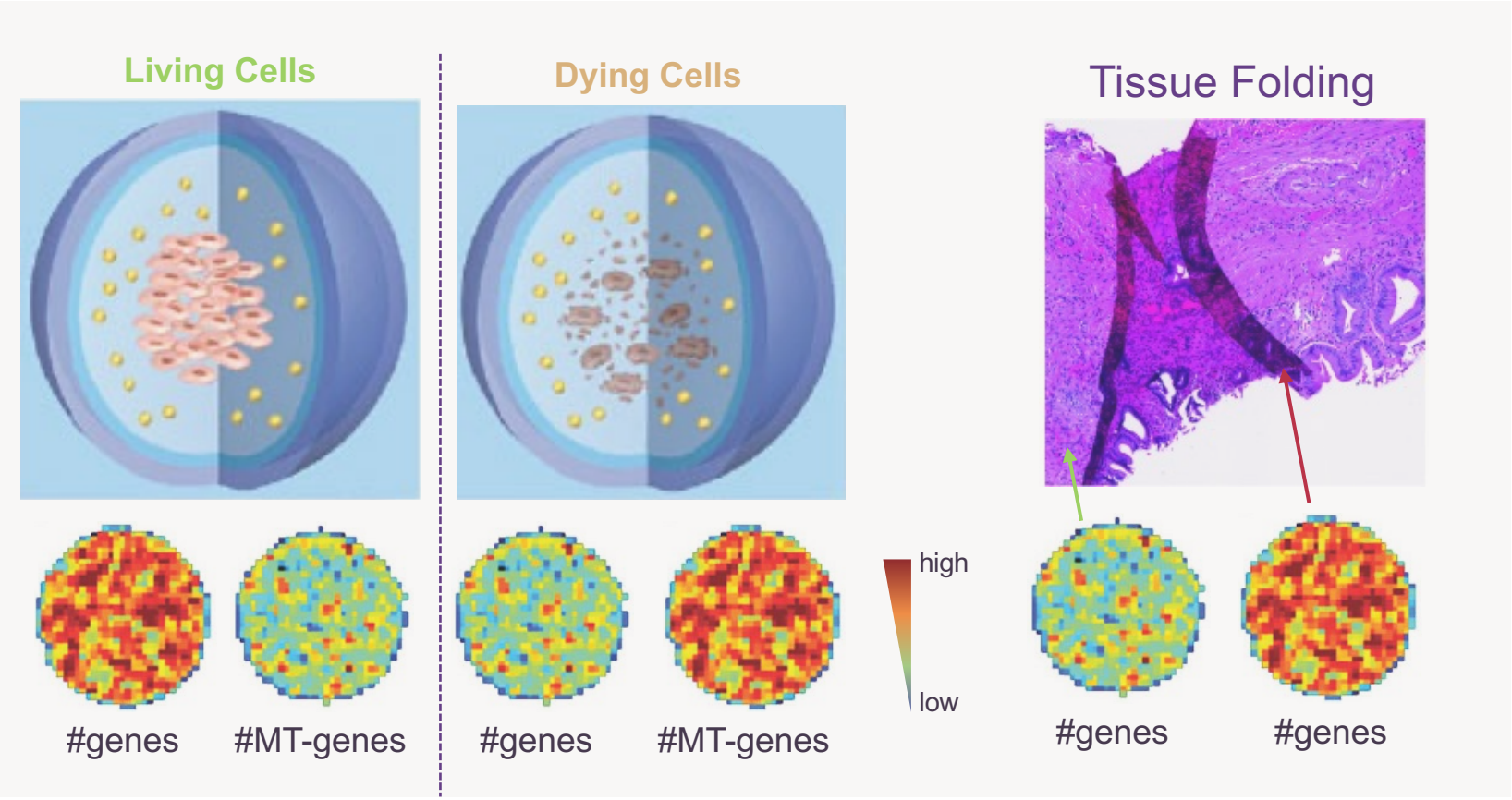
# 1. Data QC and Normalisation

*Factors of Technical Noise*

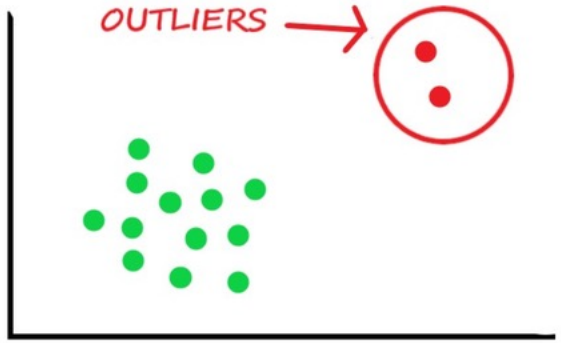


# 1. Data QC and Normalisation

*Factors of Technical Noise*



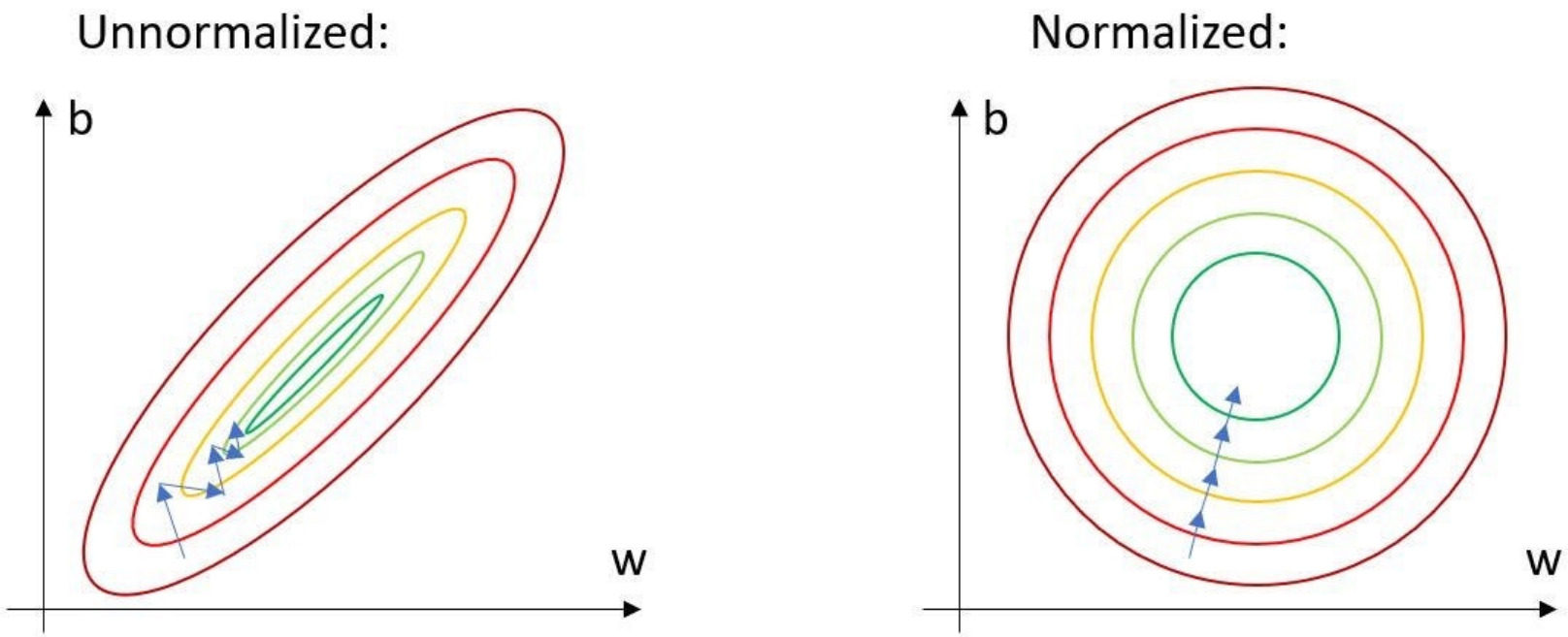

**Remove Outliers**



# 1. Data QC and Normalisation

*Data Normalisation*

**Why we normalize - Ensures comparability of gene expression between spots/cells:**

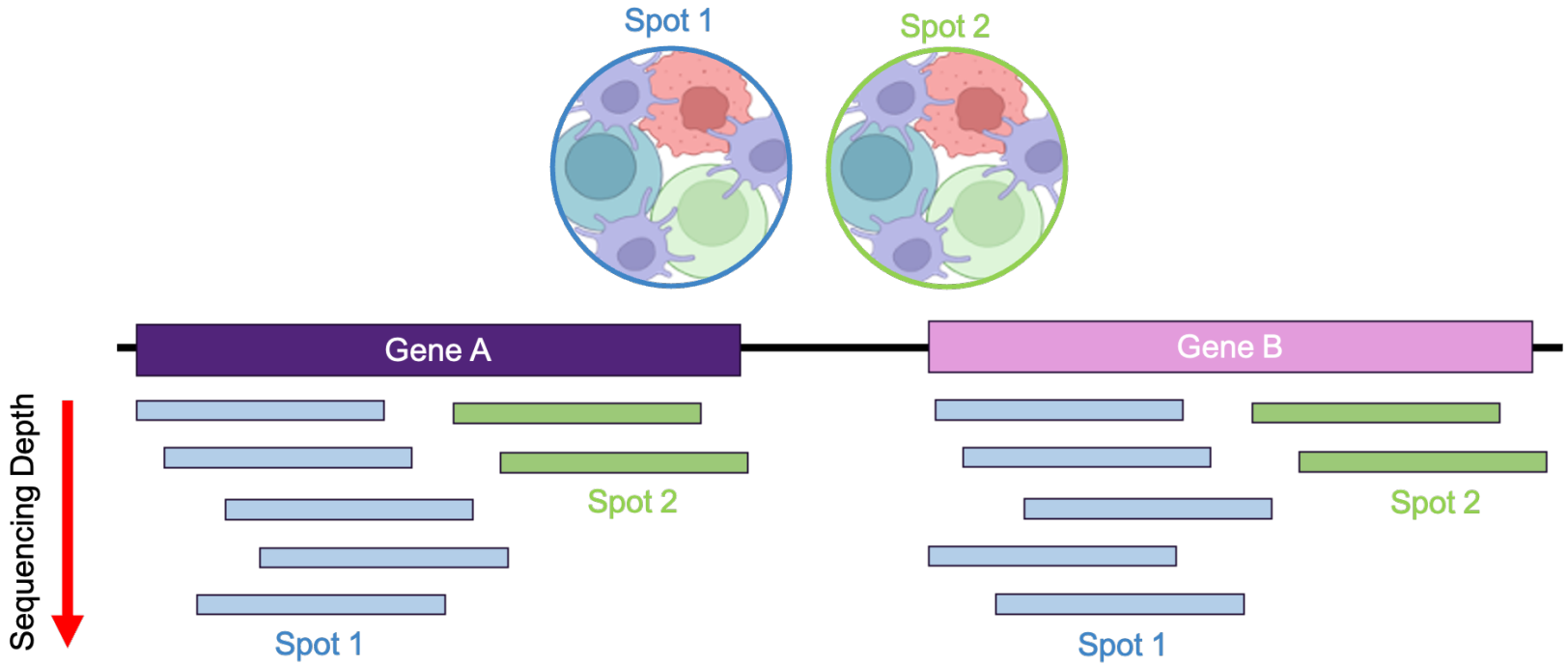


# 1. Data QC and Normalisation

*Data Normalisation*

**Why we normalize - Ensures comparability of gene expression between spots/cells:**

- *Technical noise*: capture efficiency/sequencing depth

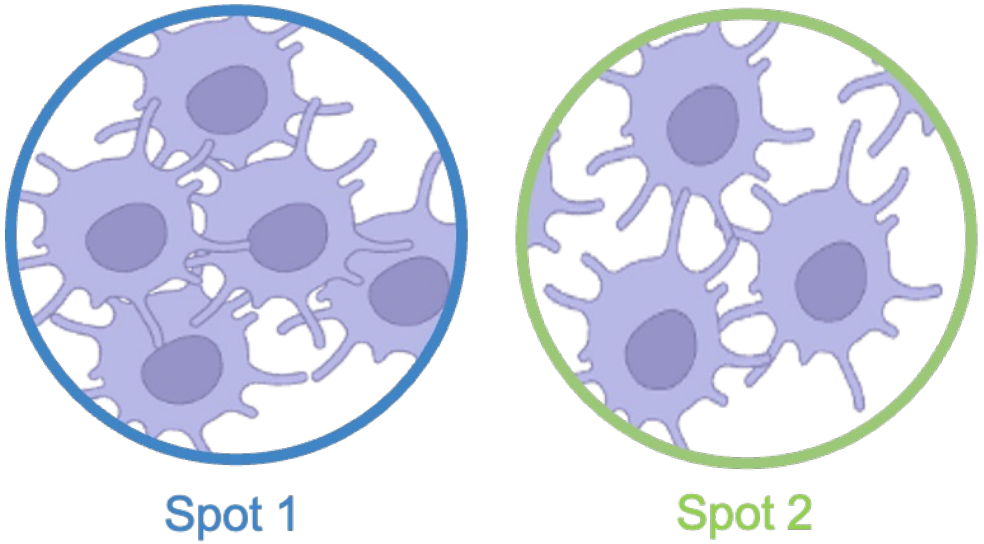


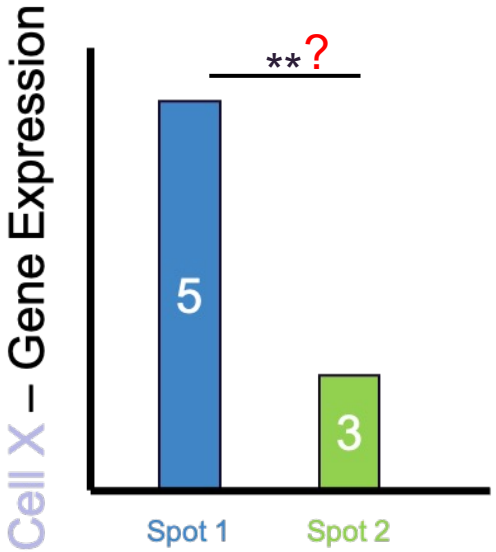
# 1. Data QC and Normalisation

*Data Normalisation*

**Why we normalize - Ensures comparability of gene expression between spots/cells:**

- *Technical noise: capture efficiency/sequencing depth*
- Biological effects: Spots may contain varying numbers of cells



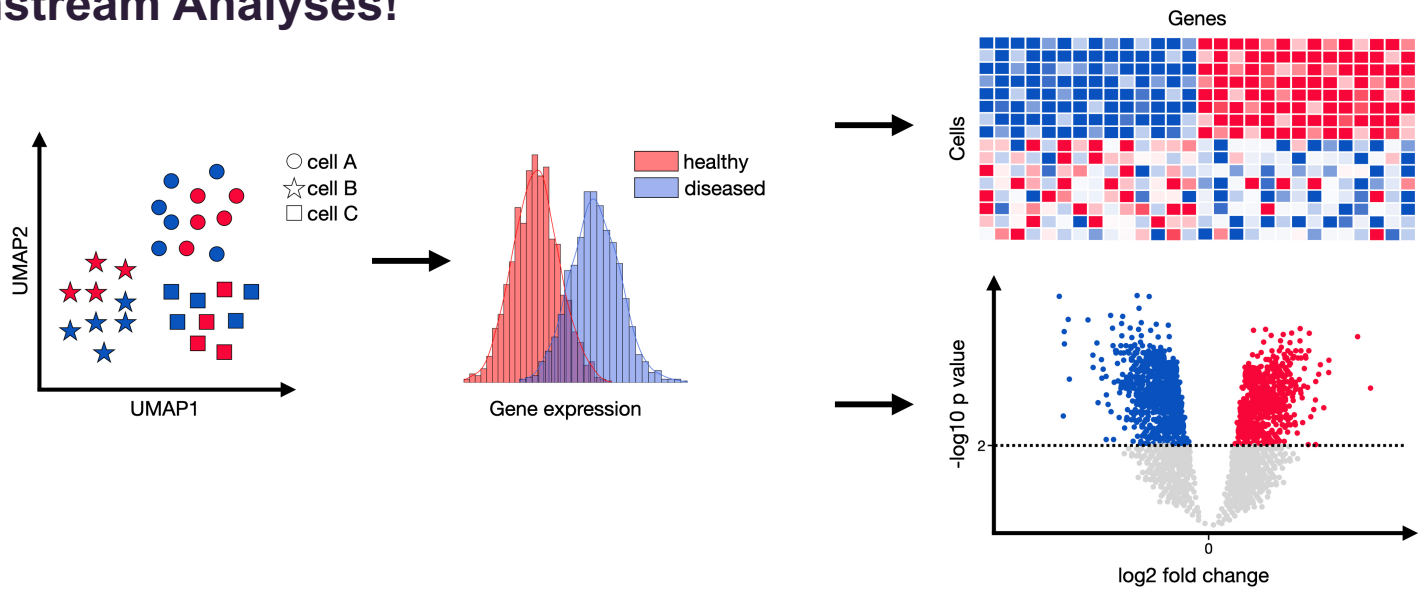
# 1. Data QC and Normalisation

*Data Normalisation*

**Why we normalize - Ensures comparability of gene expression between spots/cells:**

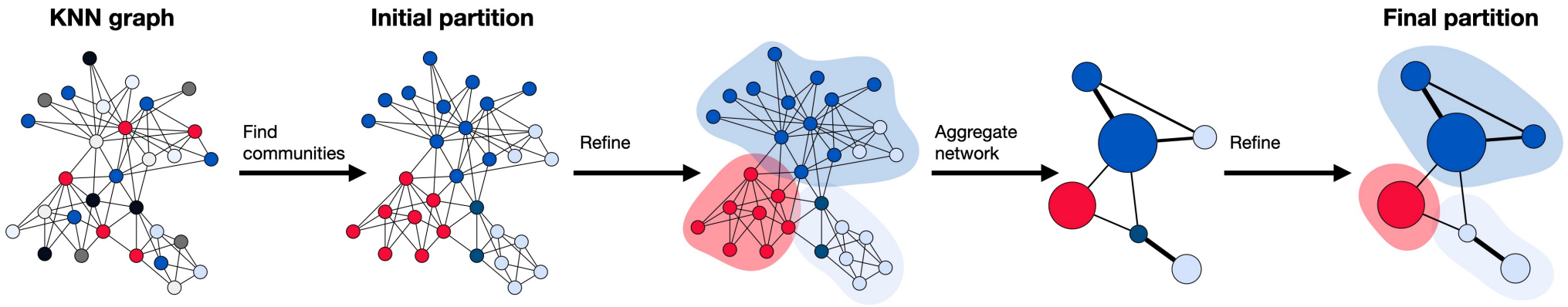
- *Technical noise*: capture efficiency/sequencing depth
- Biological effects: Spots may contain varying numbers of cells

**Need for Downstream Analyses!**



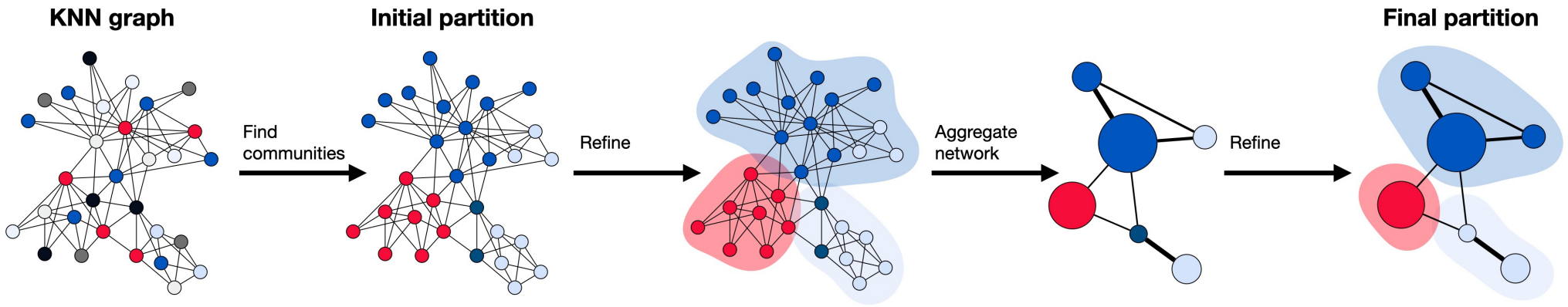
# 2. Clustering and Cell Typing

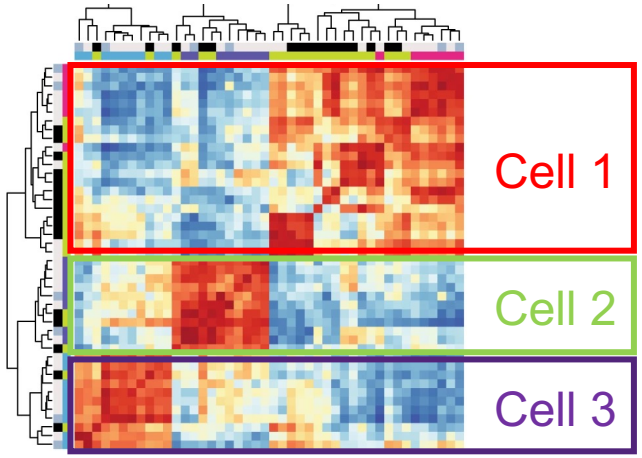
**Groups Spots/Cells together based on similar transcriptional patterns**



# 2. Clustering and Cell Typing

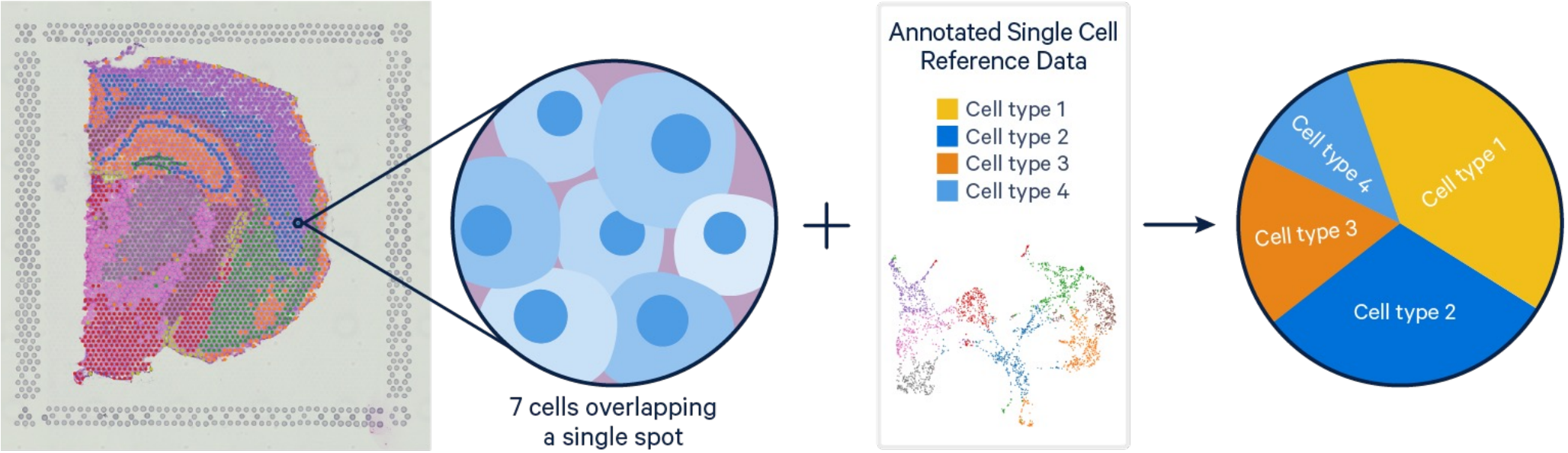
**Groups Spots/Cells together based on similar transcriptional patterns**



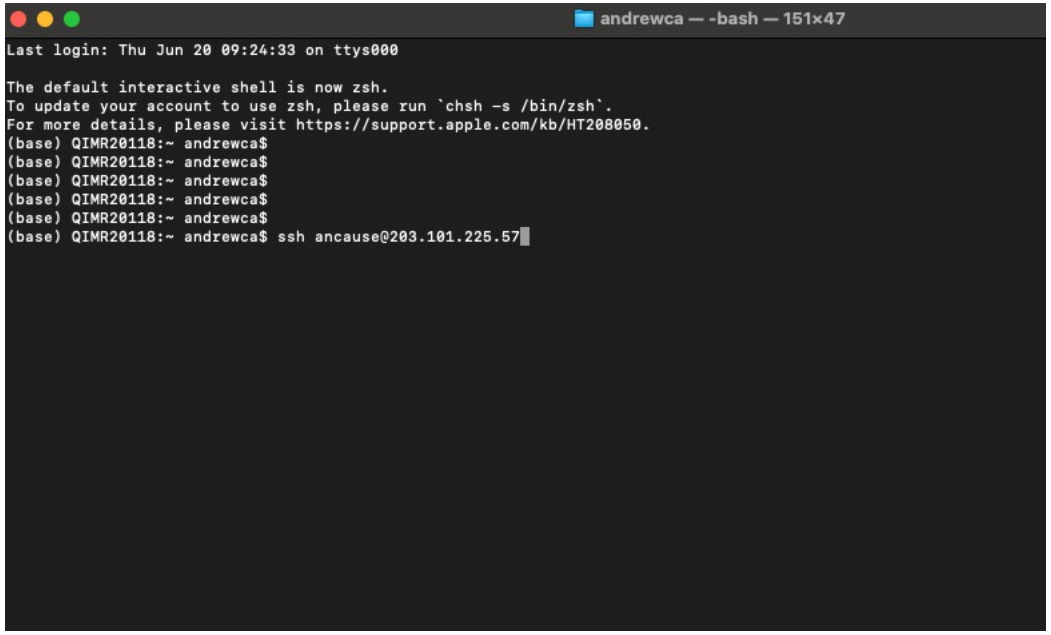
# 3. Spot Deconvolution/Label Transfer

*Spot Deconvolution*



# Running the Practical

Terminal PowerShell




1. Log into your account:

   *ssh {username}@203.101.225.57*
   **username & password from winter school email**

2. Follow these commands:

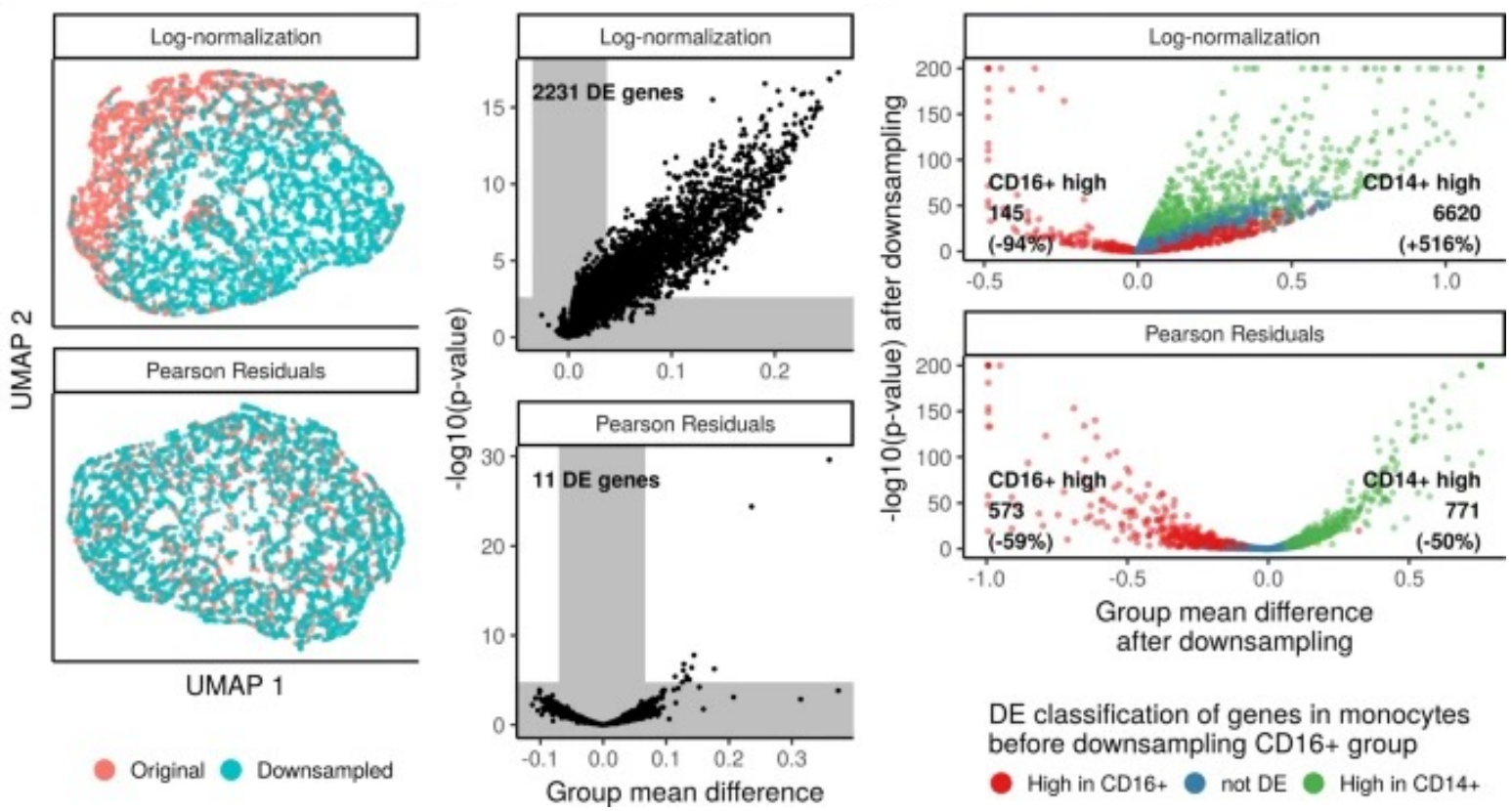
   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2024
   - git clone https://github.com/GenomicsMachineLearning/qimr-teaching-2024 /scratch/$USER/qimr-teaching-2024
   - /scratch/$USER/qimr-teaching-2024/runme.sh

3. Open JuperterNotebook:

   **/software/002-clustering-cell-typing/2.1_ST_Cell_Typeing_Tutorial.ipynb**

# 1. Data QC and Normalisation

*Data Normalisation - SCTransform*



# Lecture 3: Review Data Structure and Understand Spatial Concepts by Visualisation

Levi Hocky and Quan Nguyen

# Definition


- **Data: Collection of raw facts (numeric, categorical, etc.)**

- **Data structure: specialized format for *organizing* and *storing* data in memory that contains not only the *elements* stored but also *their relationship* to each other**

# scRNAseq or spatial transcriptomics data

- **Gene expression matrix:**
  - Row: cells/spots
  - Column: genes
- **Cells/spots metadata:**
  - Cell type
  - Batch
  - Spatial coordinates
  - …
- **Genes metadata:**
  - Reference
  - Ensembl ID
  - …
- **Image:**
  - H&E image
- **Embedding**
  - PCA
  - UMAP

| | gene_ids | feature_types | genome |
|---|---|---|---|
| **MIR1302-2HG** | ENSG00000243485 | Gene Expression | GRCh38 |

```
array([[-3.8268683e+02,  2.4569946e+02,  2.9572031e+01, ...,
        -7.4096527e+00, -1.3591890e+01, -1.5226344e+00],
       [ 8.5815186e+02,  4.6844845e+01, -5.8959357e+02, ...,
        -9.1535692e+00,  4.7668648e+01,  8.6046457e+00],
       [-5.3620459e+02, -1.2136969e+02,  8.0695274e+01, ...,
        -3.3967710e+00,  1.3312209e+00, -7.4527483e+00],
       ...,
       [ 1.8189459e+02, -4.6680363e+01, -2.7038712e+02, ...,
        -6.4620590e+00,  2.2010189e+01, -1.4795618e+01],
       [-1.9071545e+02,  3.6853920e+01, -5.3436691e+01, ...,
         3.2471569e+00, -1.2807763e+00,  6.4047074e+00],
       [-1.1925542e+02, -1.2490373e+02,  1.5722610e+02, ...,
         3.9003084e+00, -2.4630415e+00,  7.5943404e-01]], dtype=float32)
```

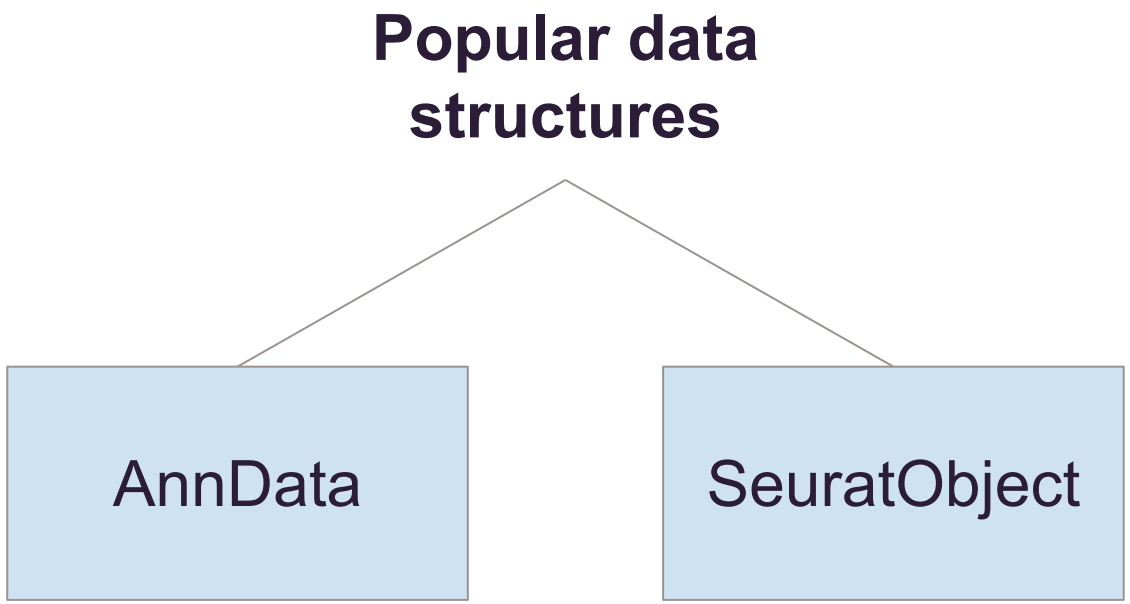
| | | | | | | |
|---|---|---|---|---|---|---|
| **FAM231C** | ENSG00000268674 | Gene Expression | GRCh38 | 8 | basal_like_1 |

3813 rows × 9 columr

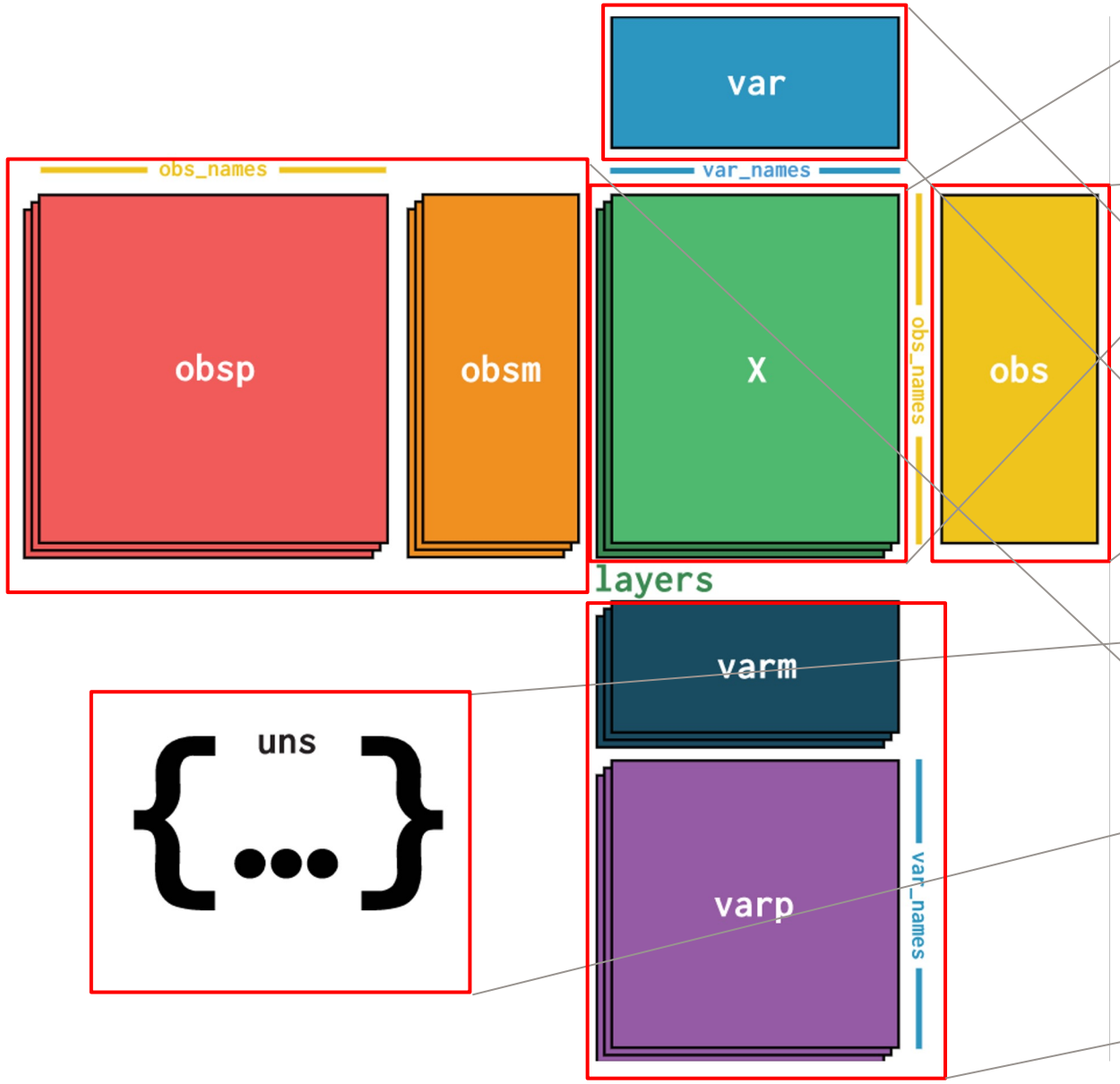
33538 rows × 3 columns

# Popular data structures





# AnnData (Annotated data) - Python




Raw counts
Normalized counts

Observations (cell/spots) metadata

Variables (genes) metadata

Image data
Unstructured data

Embedding
Features

# SeuratObject - R

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

## Seurat Object

**Assays**

Raw counts
Normalised Quantitation

**Metadata**

Experimental Conditions
QC Metrics
Clusters

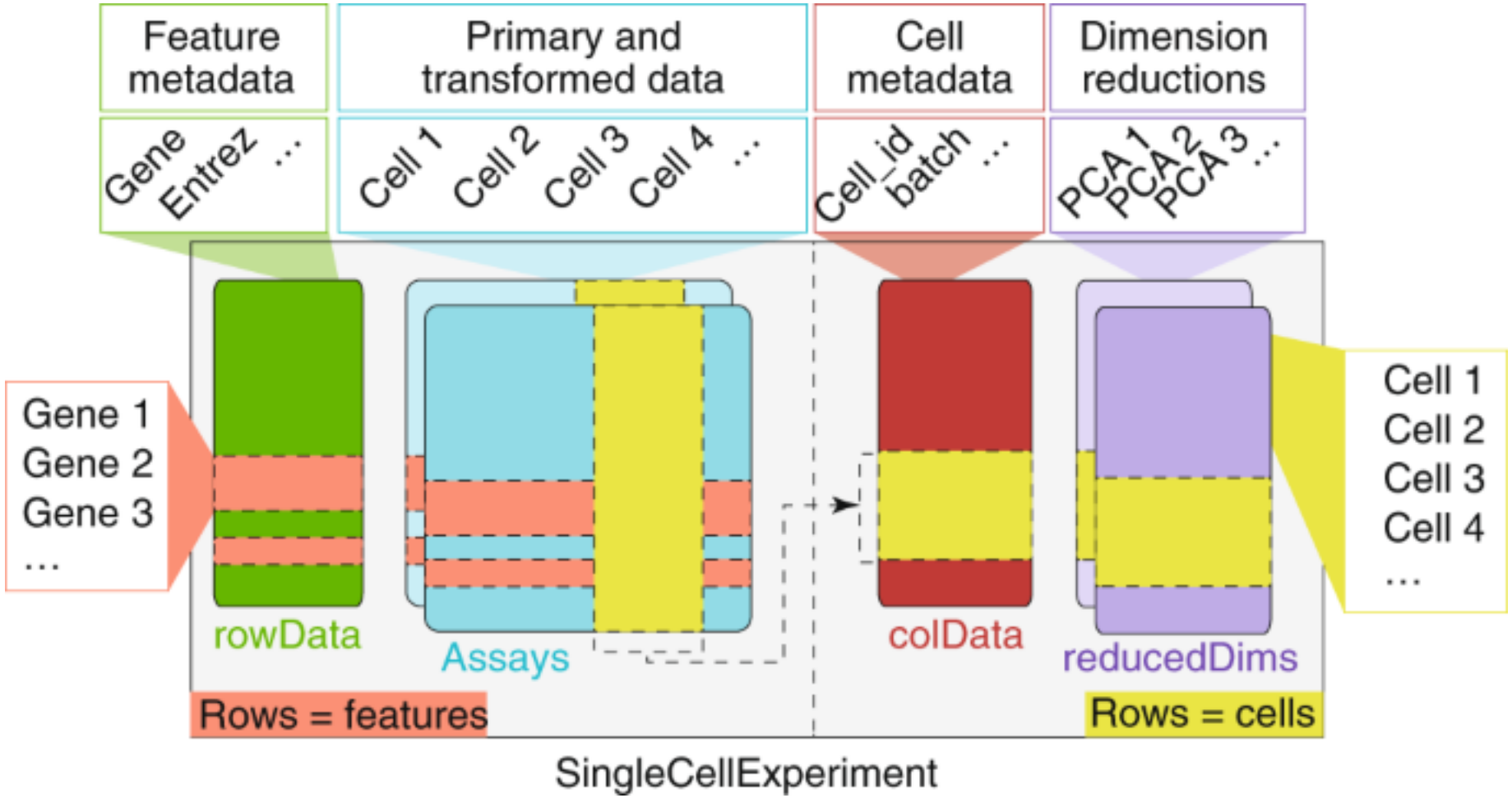
**Embeddings**

Nearest Neighbours
Dimension Reductions

**Variable Features**

Variable Gene List

# SeuratObject - R



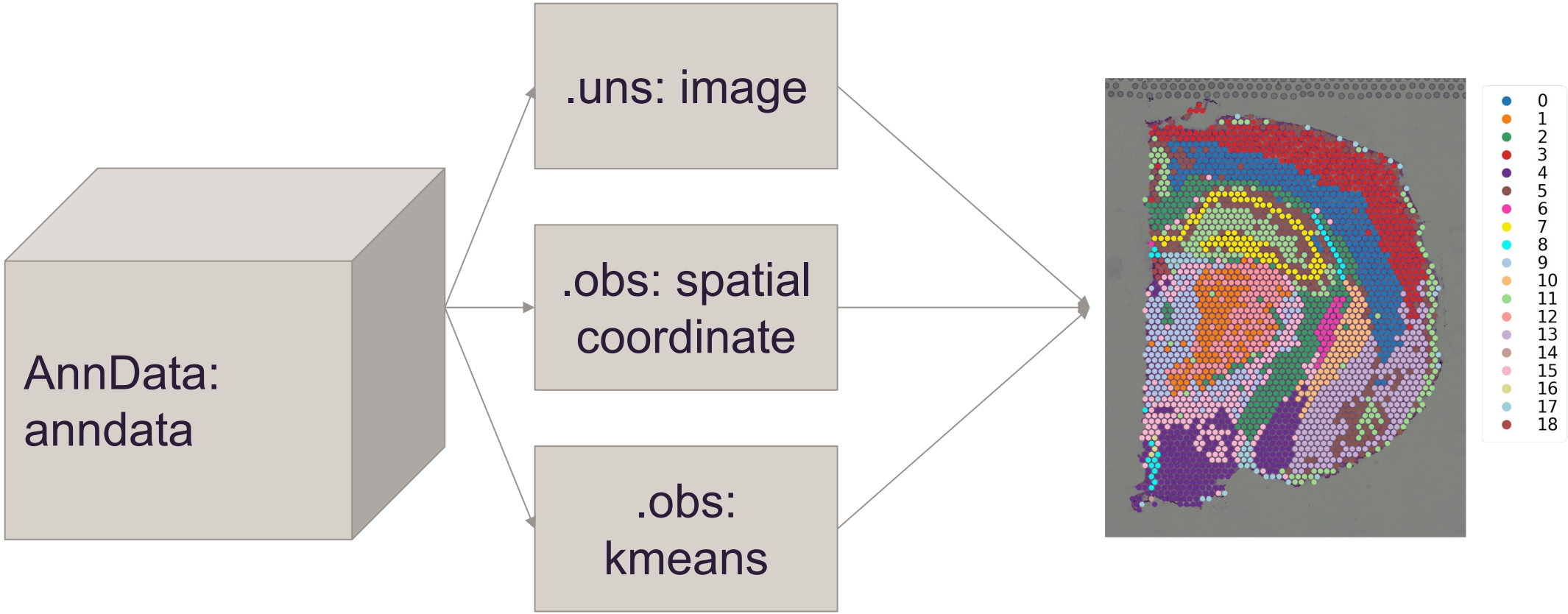
# Use case:
# Perform K-means clustering and store to AnnData

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

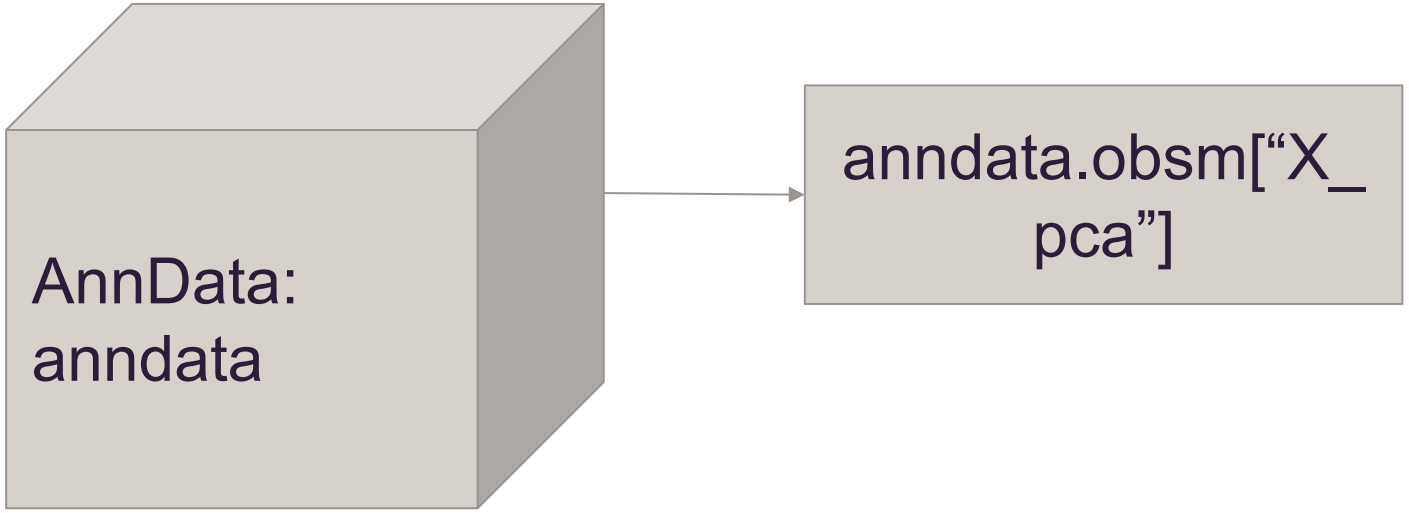
**How?**

1. **Extract the PCs components from AnnData for every cells/spots**
2. **Using external scikit-learn package for K-means clustering**
3. **Get the K-means clustering results**
4. **Add results to observation annotation of AnnData object**

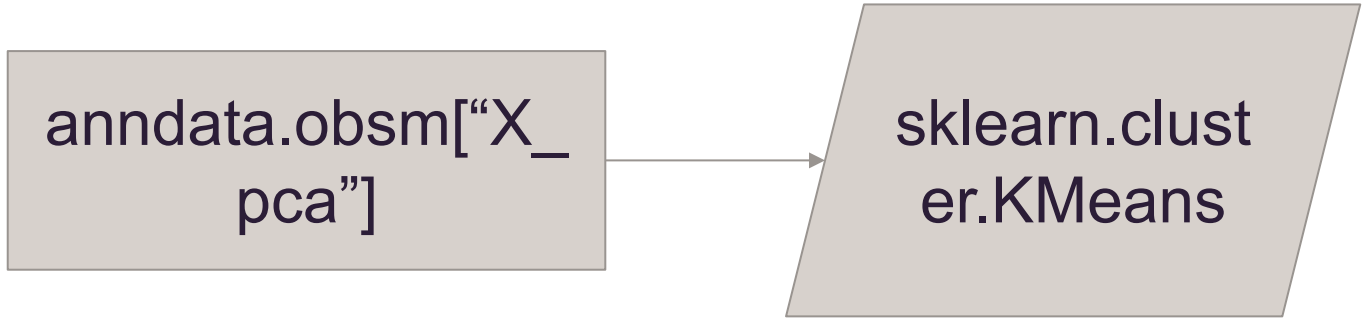
# Use case:
# Plotting Kmeans results for spatial transcriptomics



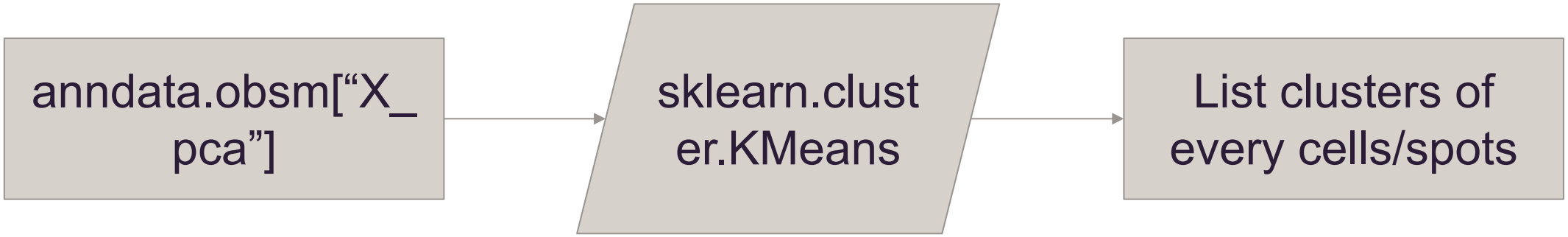
# 1. Extract the PCs components from AnnData for every cells/spots



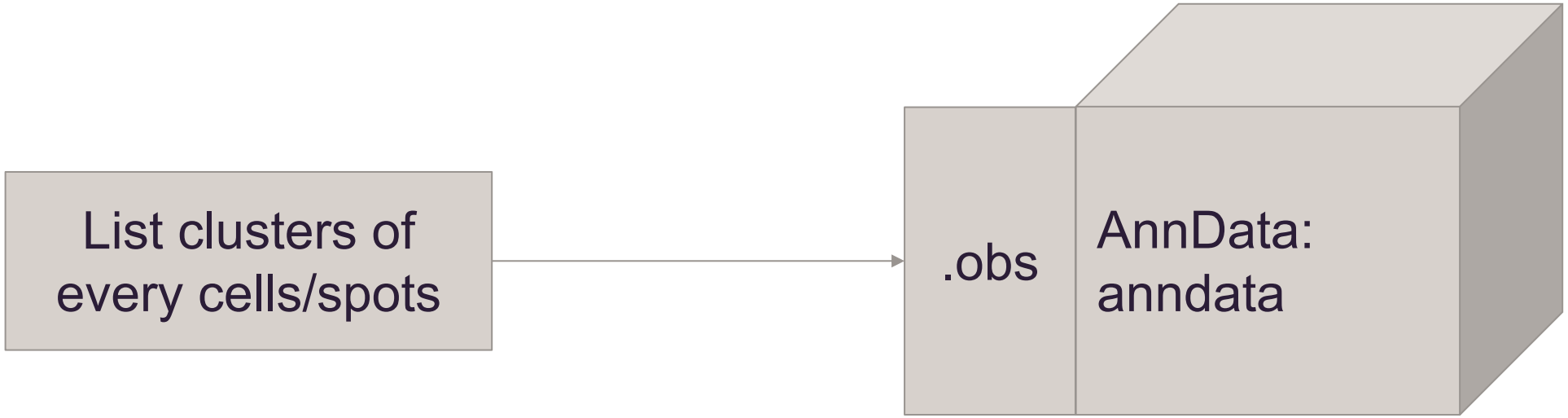
# 2. Using external scikit-learn package for K-means clustering



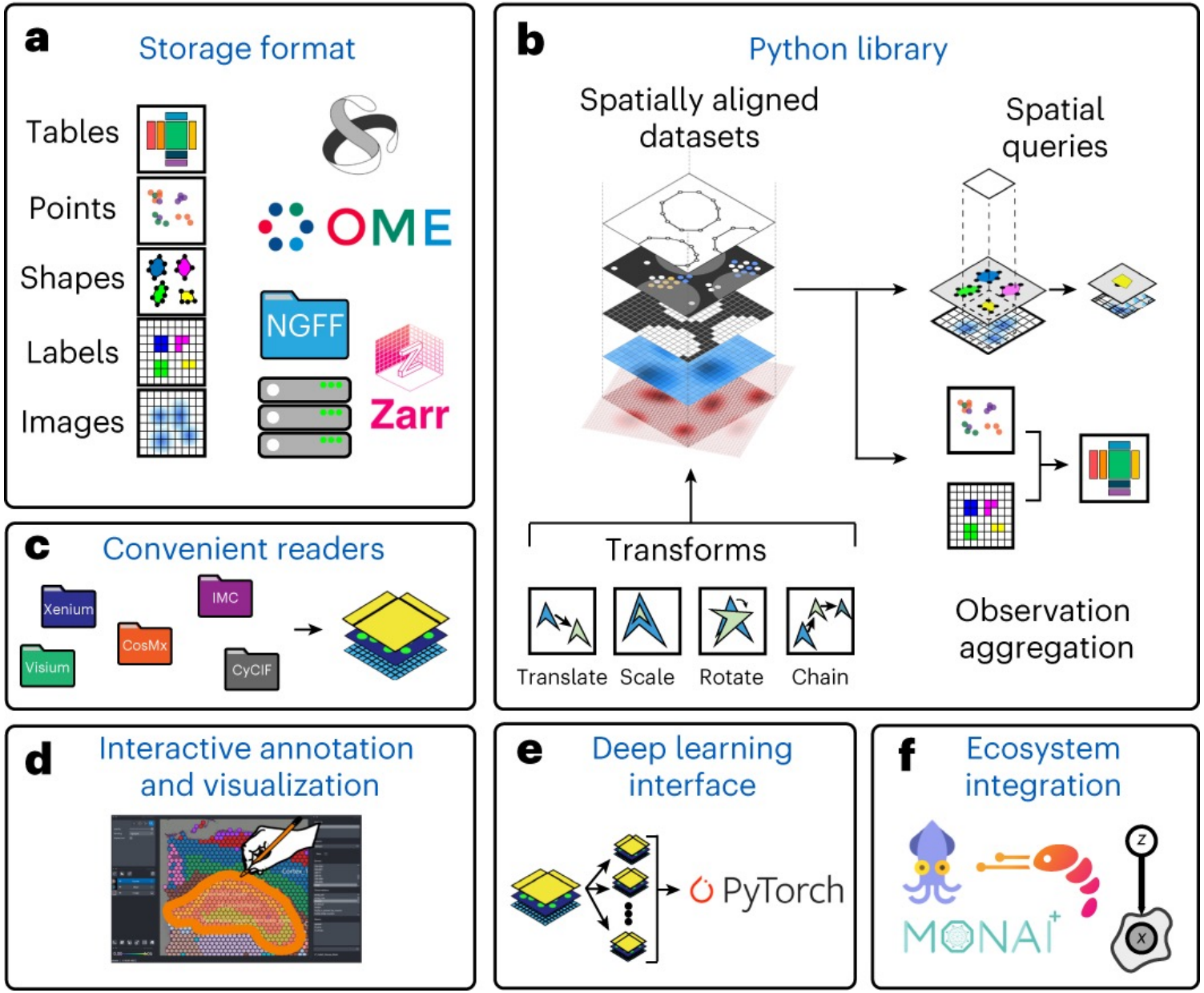
# 3. Get the K-means clustering results



# 4. Add results to observation annotation of AnnData object



# Analysis landscape

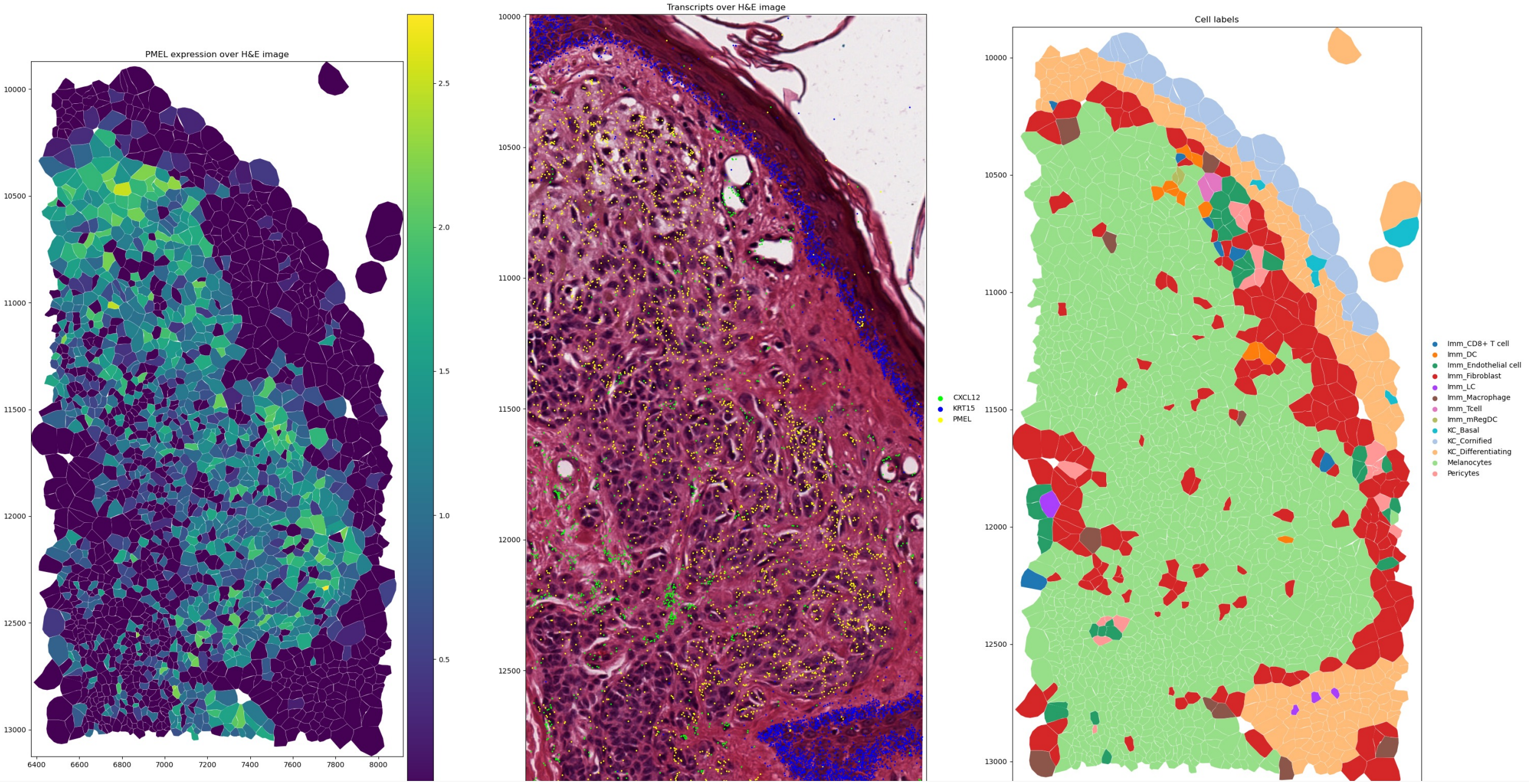



(Marcorano et al., 2024)

# Spatial Single Cell Data

```
SpatialData object with:
├── Images
│     ├── 'HE': SpatialImage[cyx] (3, 4633, 14747)
│     ├── 'morphology_focus': MultiscaleSpatialImage[cyx] (1, 37441, 11479), (1, 18720, 5739), (1, 9360, 2869),
(1, 4680, 1434), (1, 2340, 717)
│     └── 'morphology_mip': MultiscaleSpatialImage[cyx] (1, 37441, 11479), (1, 18720, 5739), (1, 9360, 2869), (1,
4680, 1434), (1, 2340, 717)
├── Labels
│     ├── 'cell_labels': MultiscaleSpatialImage[yx] (37441, 11479), (18720, 5739), (9360, 2869), (4680, 1434), (2
340, 717)
│     └── 'nucleus_labels': MultiscaleSpatialImage[yx] (37441, 11479), (18720, 5739), (9360, 2869), (4680, 1434),
(2340, 717)
├── Points
│     └── 'transcripts': DataFrame with shape: (4062390, 10) (3D points)
├── Shapes
│     ├── 'cell_boundaries': GeoDataFrame shape: (21596, 1) (2D shapes)
│     ├── 'cell_circles': GeoDataFrame shape: (21596, 2) (2D shapes)
│     └── 'nucleus_boundaries': GeoDataFrame shape: (21596, 1) (2D shapes)
└── Tables
      └── 'table': AnnData (21593, 260)
with coordinate systems:
▸ 'global', with elements:
        HE (Images), morphology_focus (Images), morphology_mip (Images), cell_labels (Labels), nucleus_labels (La
bels), transcripts (Points), cell_boundaries (Shapes), cell_circles (Shapes), nucleus_boundaries (Shapes)
```

Essentially, spatialdata is an extension of AnnData that allows for more advanced plotting and image transformations.


PMEL expression over H&E image
Transcripts over H&E image
Cell labels
CXCL12
KRT15
PMEL
Imm_CD8+ T cell
Imm_DC
Imm_Endothelial cell
Imm_Fibroblast
Imm_LC
Imm_Macrophage
Imm_Tcell
Imm_mRegDC
KC_Basal
KC_Cornified
KC_Differentiating
Melanocytes
Pericytes

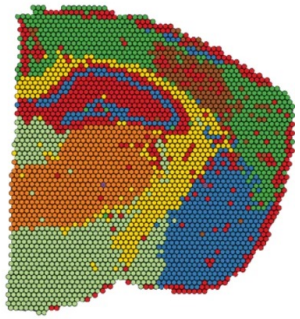
# Lecture 4: Spatial DNA-level analysis for Copy Number Variation

# **Module 2 – Part 4:** Spatial DNA-level analysis for Copy Number Variation

Prakrithi– prakrithi.pavithra@uq.edu.au

Seurat v5 package

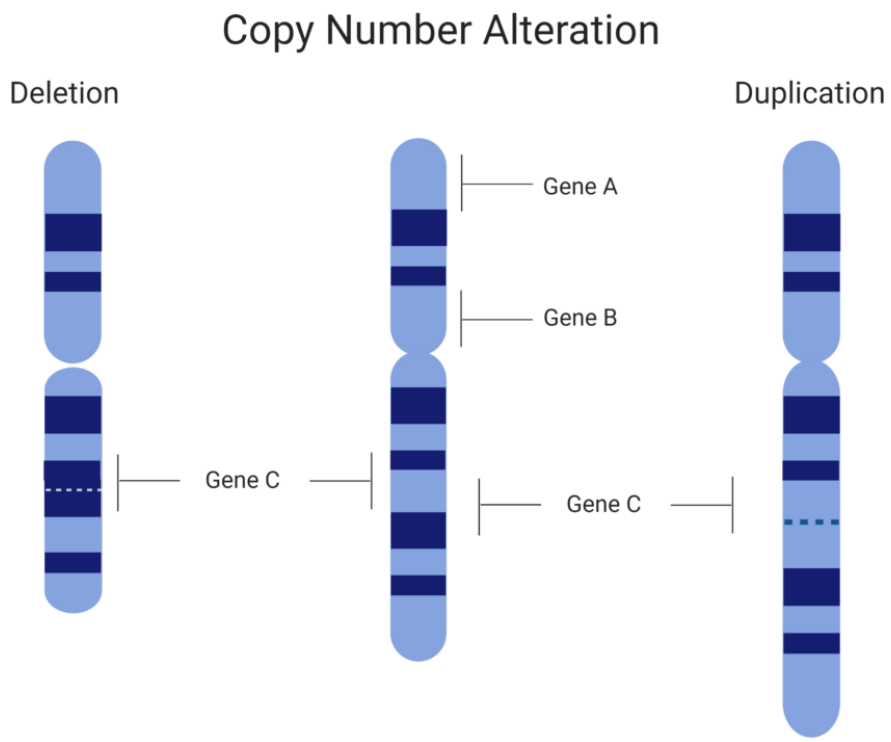
10X Chromium



10X Visium

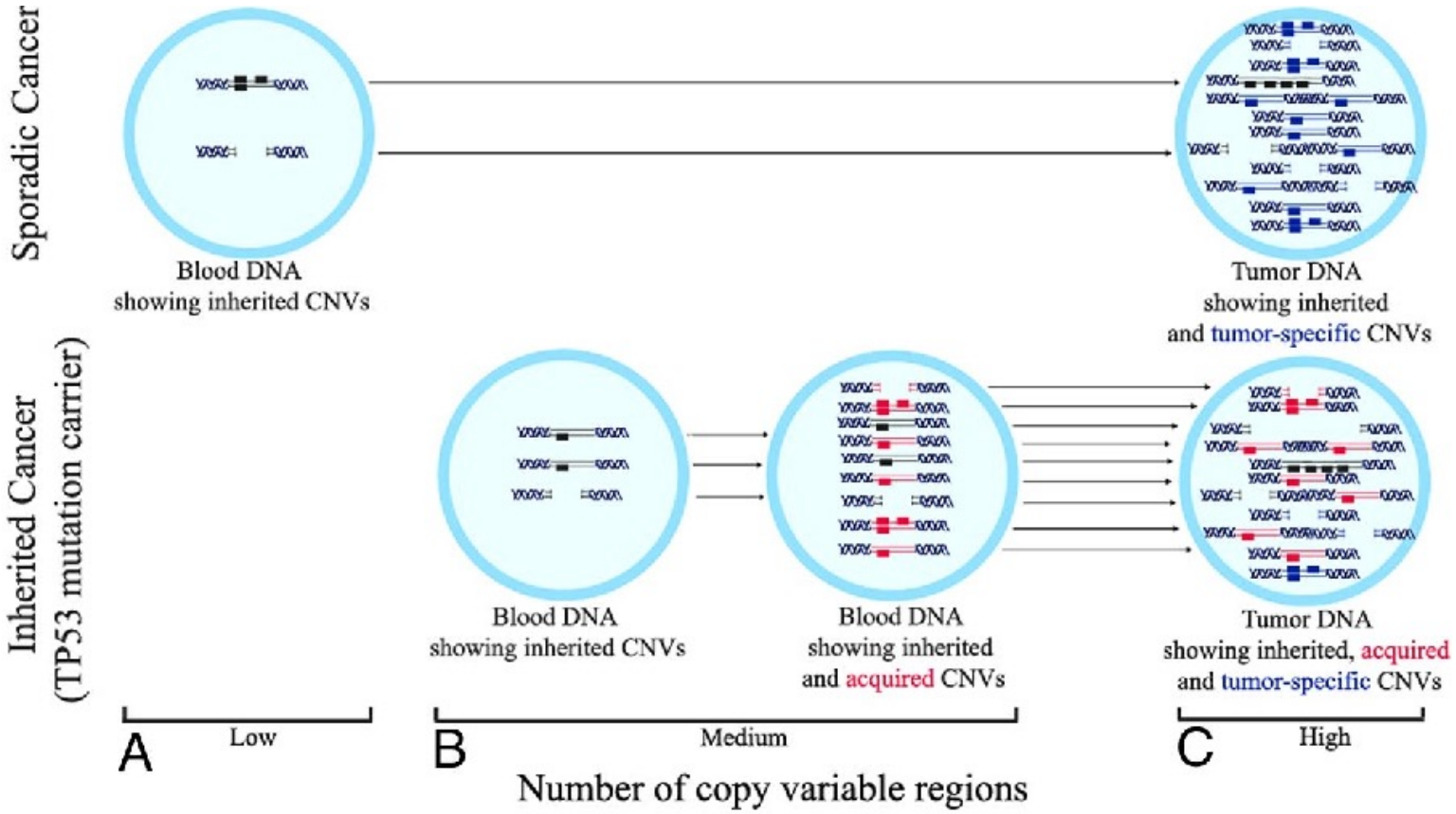
# Module 2 – Part 4: Copy Number Variations

A Copy number variation (abbreviated as CNV) refers to an instance in which the number of copies of a specific DNA segment varies among different individuals' genomes. These variations can involve deletions or duplications of segments of the genome and can range from a few kilobases to several megabases in size.



# How are CNVs related to cancer?

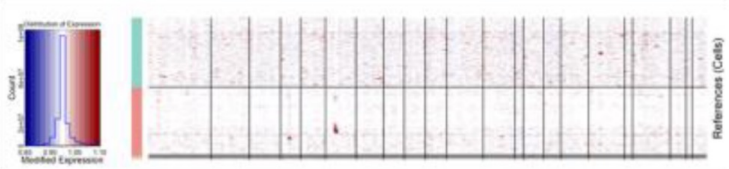
- Oncogene Amplification
- Tumor Suppressor Gene Deletion
- Genomic Instability



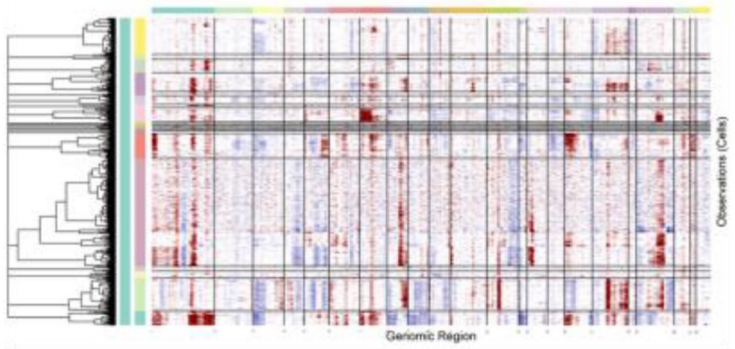

A. Shlien et al., 2008

# How can we make use of this DNA profile information for RNA-seq data?

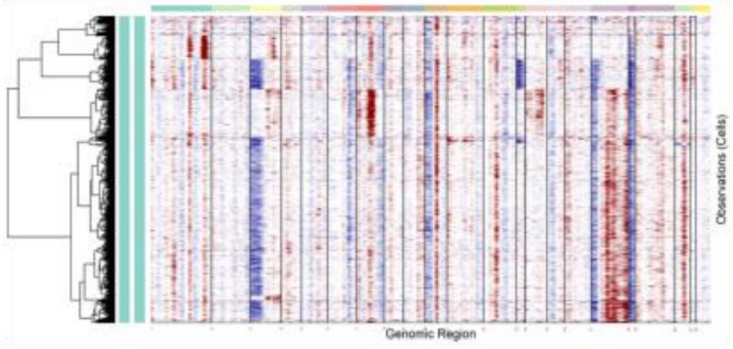
Multiple patients: Normal
No CNV pattern



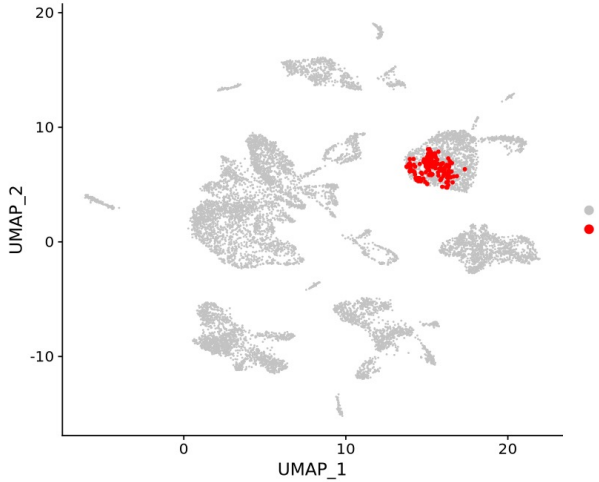
Multiple patients:Tumor
Patient-specific CNV pattern



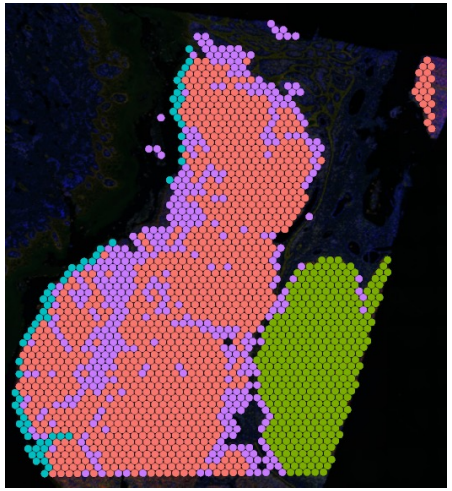

One patient: Tumor
Clonal heterogeneity




Identification of Malignant cells




Analysis of sub-clones

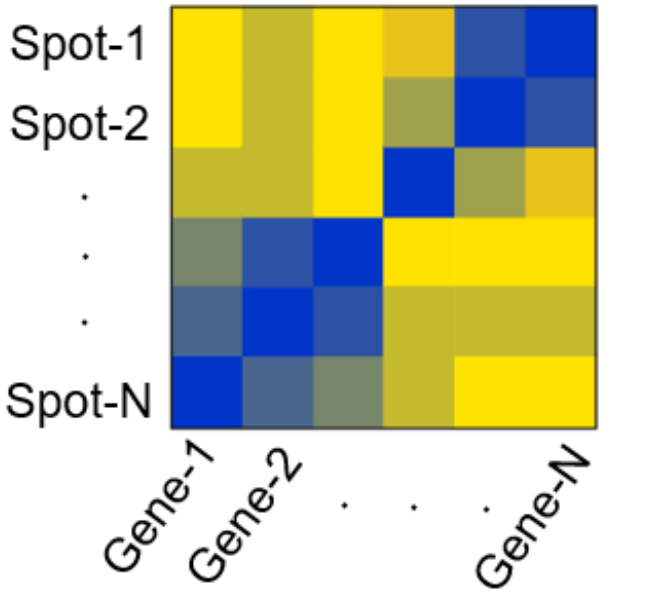


# Data Requirement

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

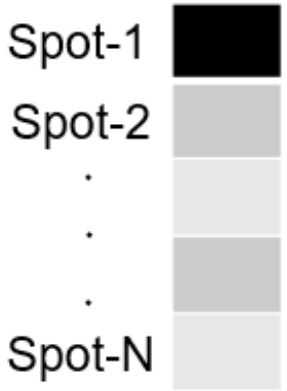
1

**Gene Expression Matrix**



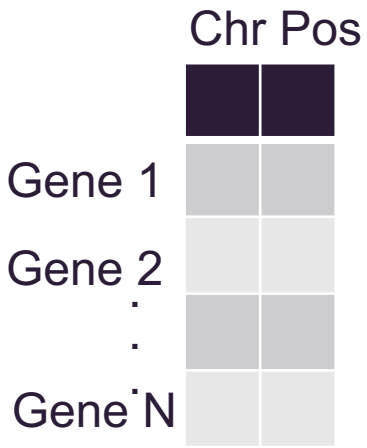

2

**Cell-type Annotation**



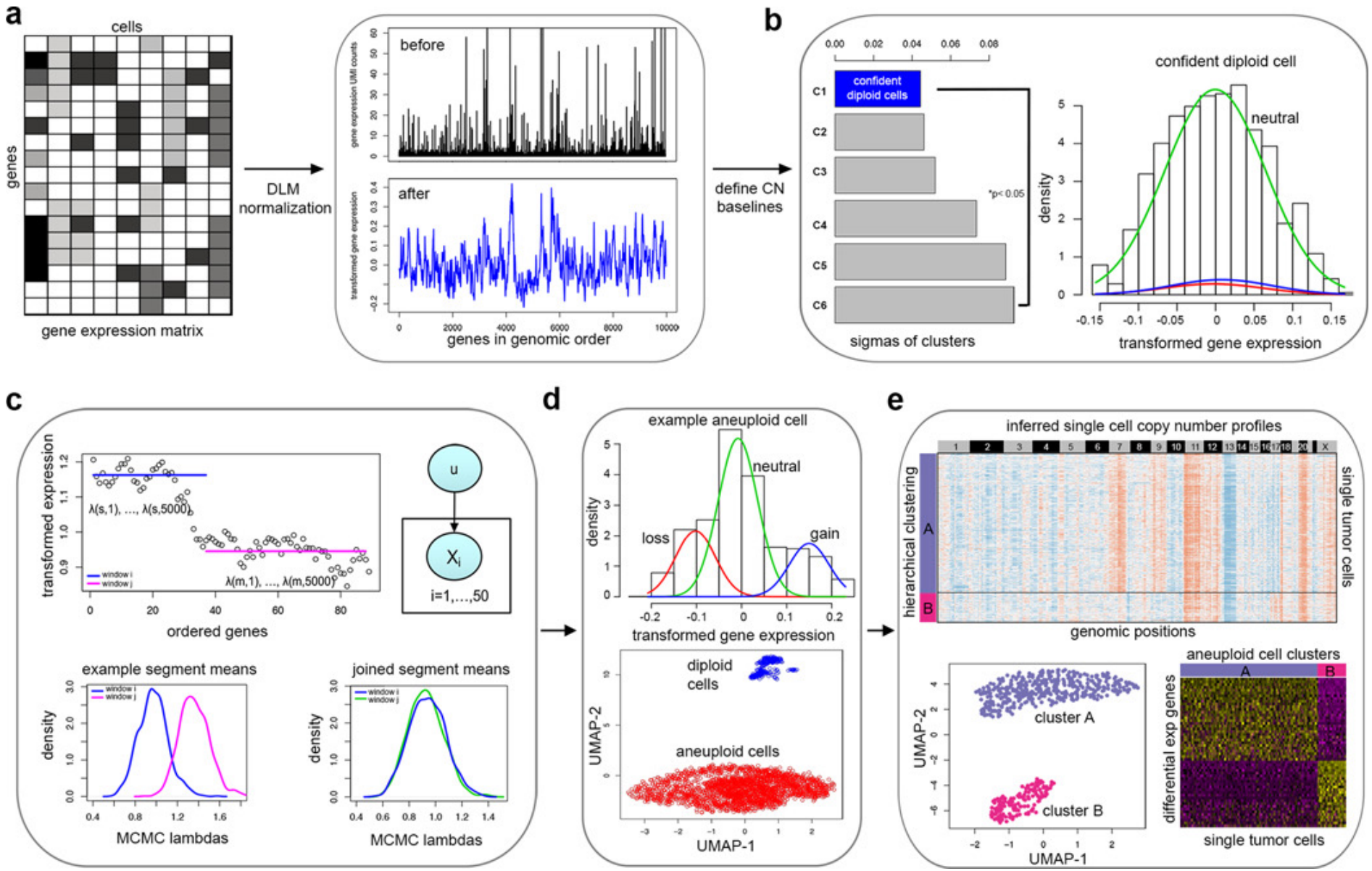

3

**Gene annotation file**



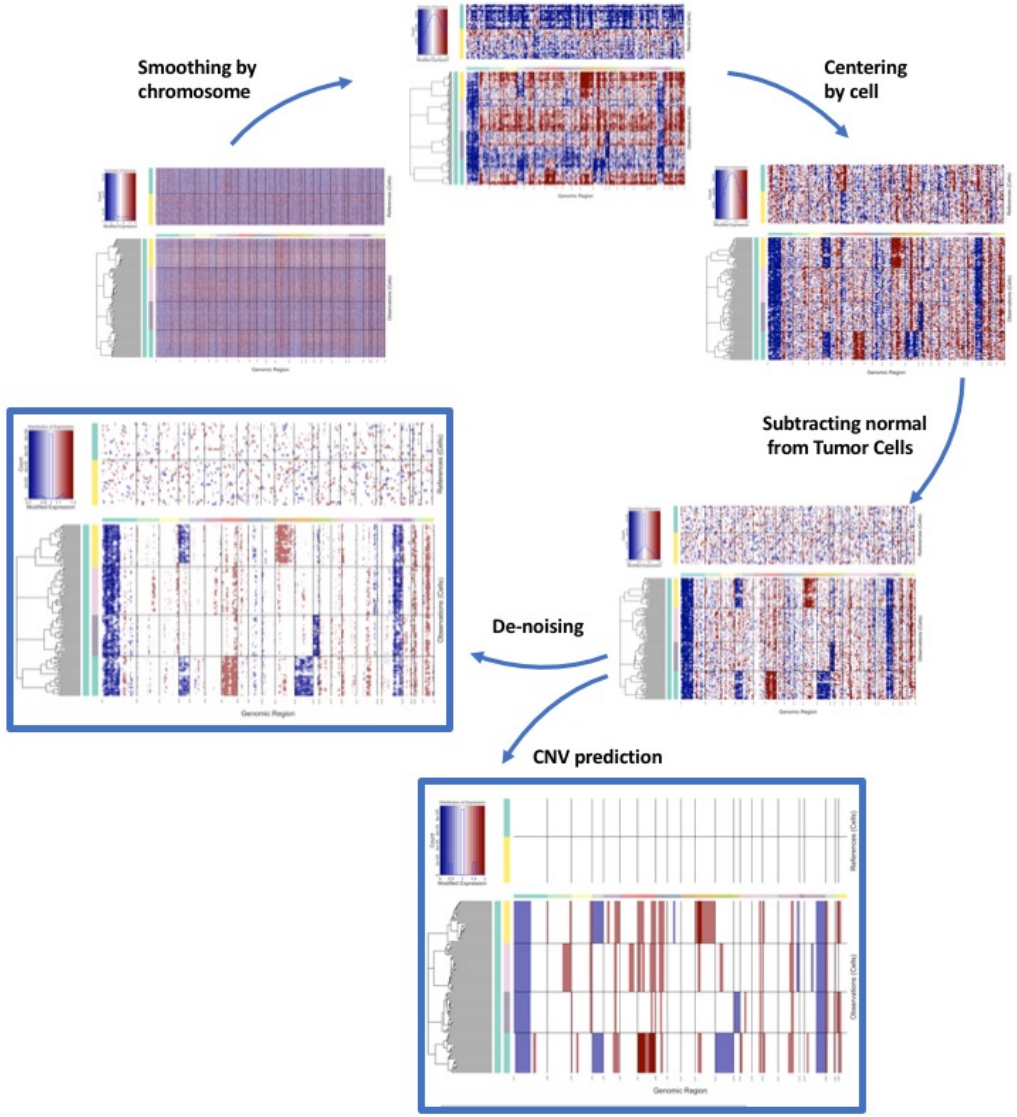
# Tools for CNV profiling : **CopyKAT**

Ruli Gao et al., 2021



# Tools for CNV profiling : **InferCNV**

- Gene ordering based on chromosomal coordinates
- The moving average is calculated by taking the mean of a fixed number of consecutive data points
- InferCNV takes in metadata of cell types and needs you to define the normal cells. If you don't know that, it uses an inbuilt normal profile reference.
- InferCNV constructs the CNV profile of a known normal sample, and then for each gene and each cell, the normal sample is subtracted from the tumor sample to determine the final tumor CNV profile of the tumor.




https://github.com/broadinstitute/infercnv/wiki/images/InferCNV_procedure.png

# Practical Session

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

**Analysis of an In-house scRNA-Seq Melanoma dataset**

- CopyKAT and InferCNV already run on this dataset – Output files are preloaded
- Visualization of results with UMAP plots

**Analysis of a publicly available Spatial Melanoma dataset**

- Dataset link https://www.10xgenomics.com/datasets/human-melanoma-if-stained-ffpe-2-standard
- Identification of tumor region and tumor sub-clones

# Lecture 5: Cell Community Identification

# Module 2 – Part 5: Cell community identification

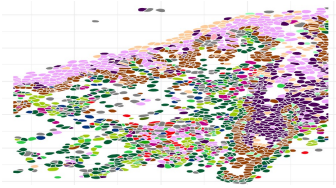
Feng Zhang and Dr Quan Nguyen

Seurat v5

R package

scikit learn

pipeline



CosMx

# Module 2 – Part 5: Overview of cell community

1. Introduction of cell community

2. HoodscanR workflow

3. NeighborhoodCoordination workflow

4. The downstream analysis of cell community identification

# Cell community identification

THE UNIVERSITY OF QUEENSLAND
AUSTRALIA

## Cell community:

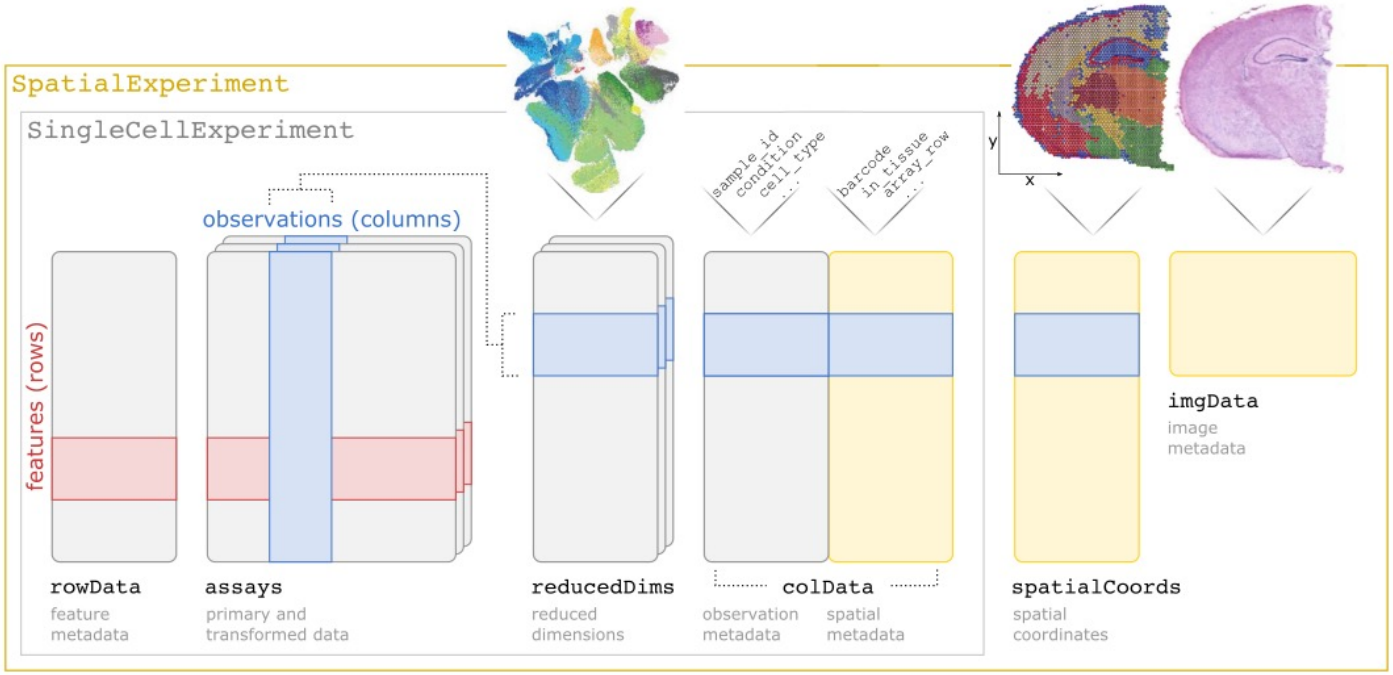
Cell community analysis characterizes the community or niche in which cells reside, which may harbor a critical tissue micro-environment that influences disease development, progression, and response to therapy.

## The biological questions to answer:

- How do the cell communities change under different conditions?
- What is the heterogeneity of cell communities?
- What is the composition of cell communities?
- How do cells within the cell community contribute to disease development, progression, and response to therapy?
- …

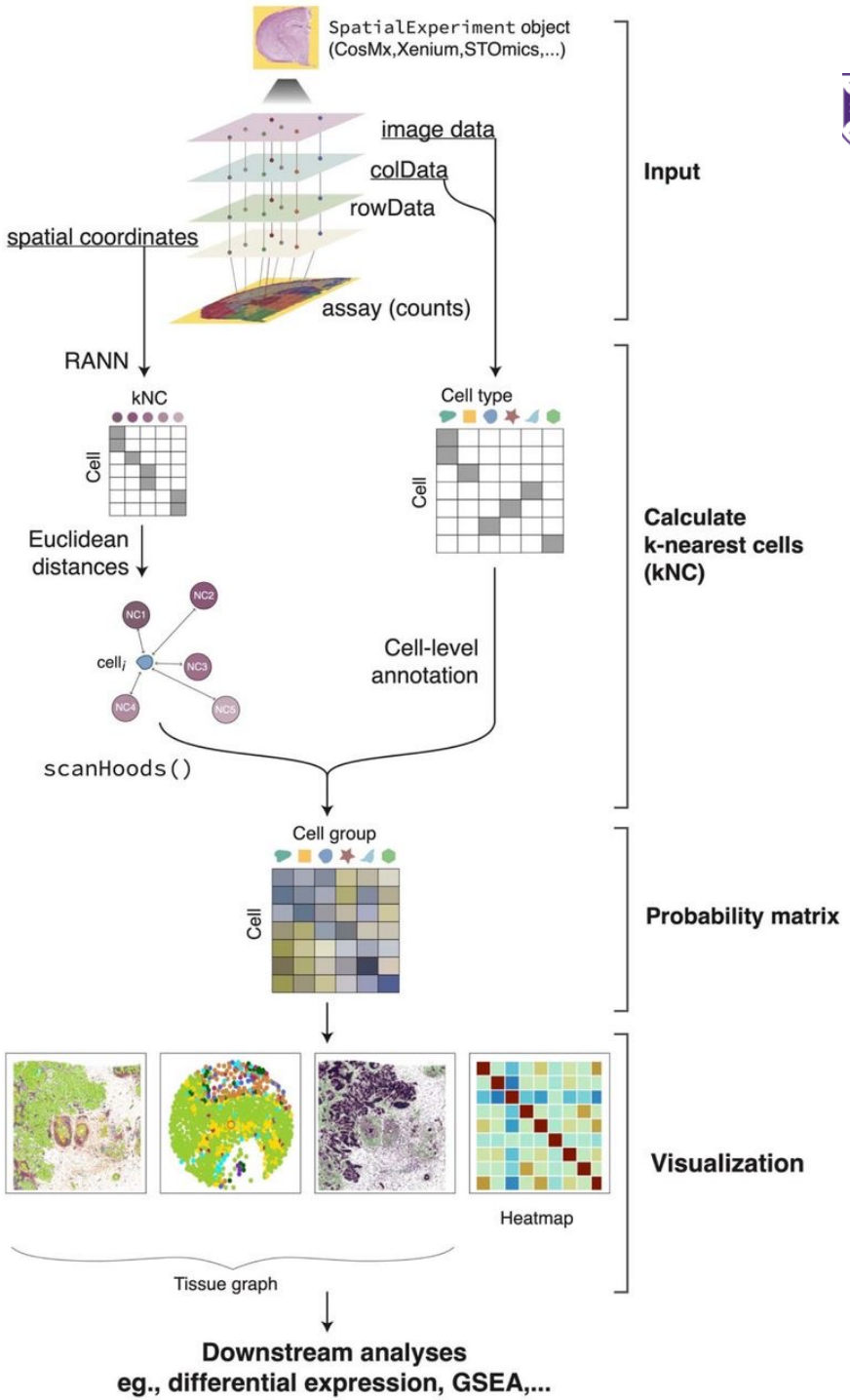
# SpatialExperiment

- assays containing expression counts
- rowData containing information on features, i.e. genes
- colData containing information on spots or cells, including nonspatial and spatial metadata
- spatialCoords containing spatial coordinates
- imgData containing image data.



# hoodscanR R package

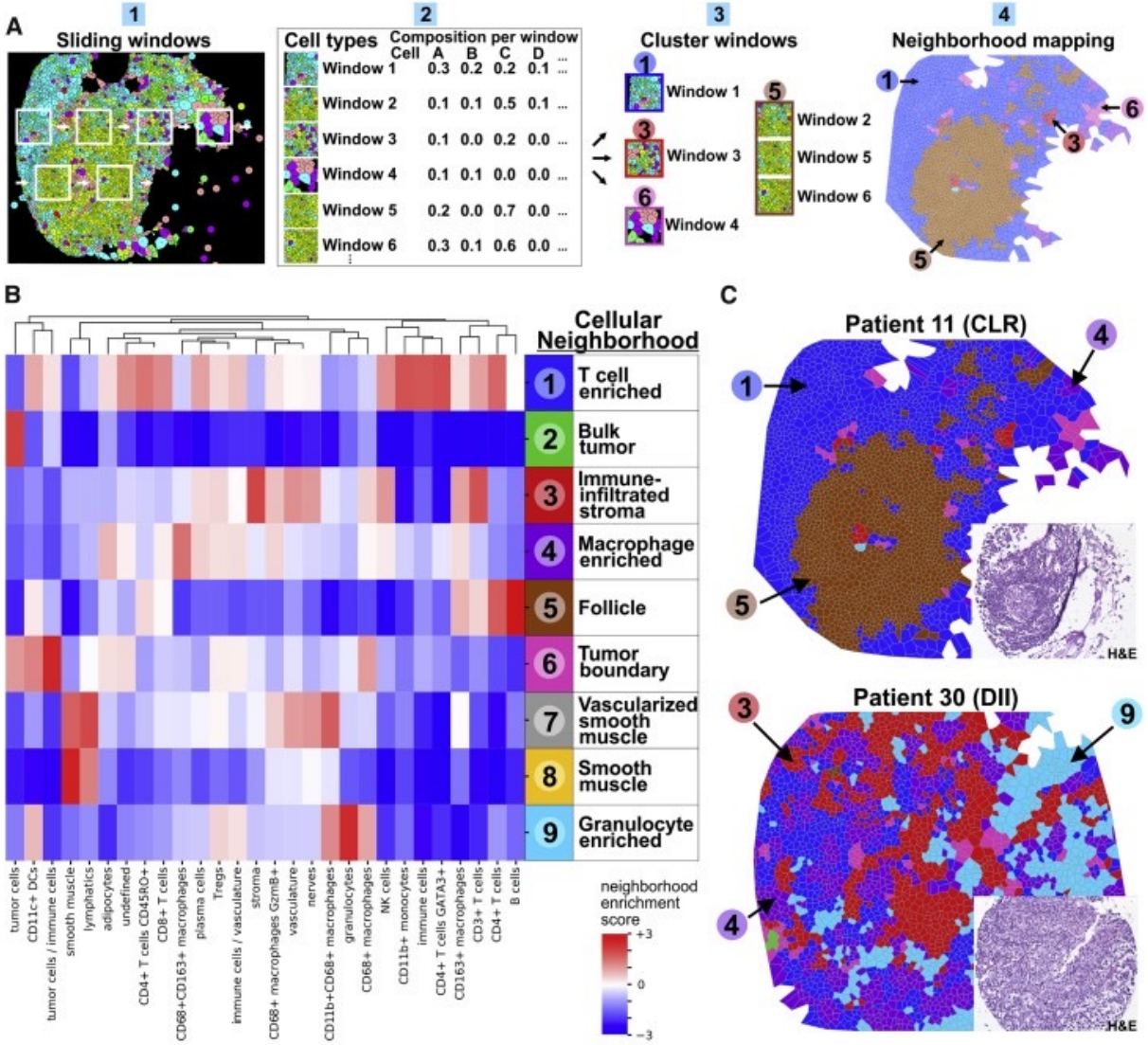
- findNearCells(): to identify K nearest cells for each cell
- scanHoods(): to generate a matrix with the probability of each cell associating with their K nearest cells
- clustByHood(): to cluster the cells by their neighborhood probability distribution




ProfileNing Liu (2024)

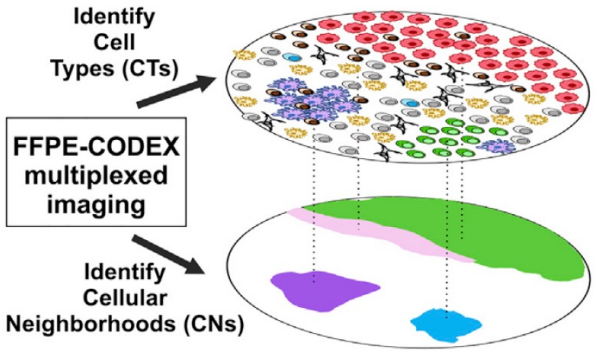
# NeighborhoodCoordination python pipeline

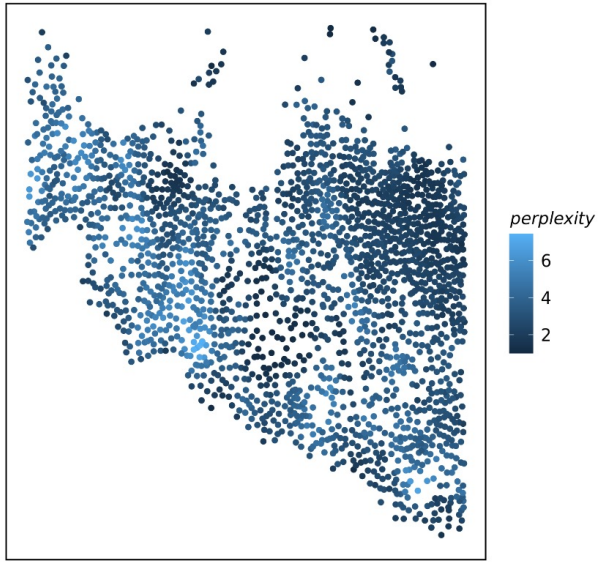
- For every cell in the tissue, its K nearest spatial neighbors, which we labeled its "window" were identified (Figure A.1).
- The cell type composition was determined per window (Figure A.2)
- All windows were clustered into different communities (Figure A.3).
- Identification of distinct cell communities based on the original cell types and their respective frequencies within each cell community (Figure B)

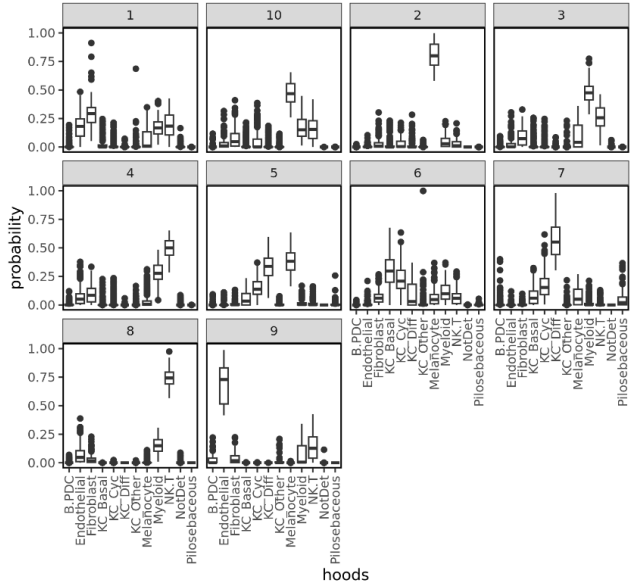


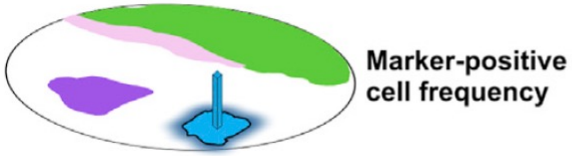

Christian M. Schürch (2020)

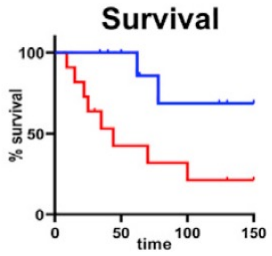
# Downstream analyses

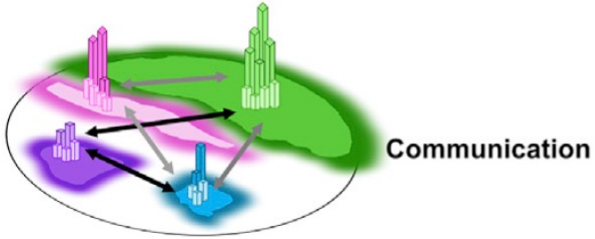


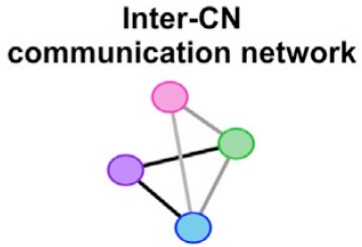




Christian M. Schürch (2020)

# Lecture 6: Cell-Cell Interactions

# Module 2 – Part 6: Cell-Cell Interactions

Onkar Mulay – o.mulay@uq.edu.au

CellChatDB

ST LEARN

MMCCI

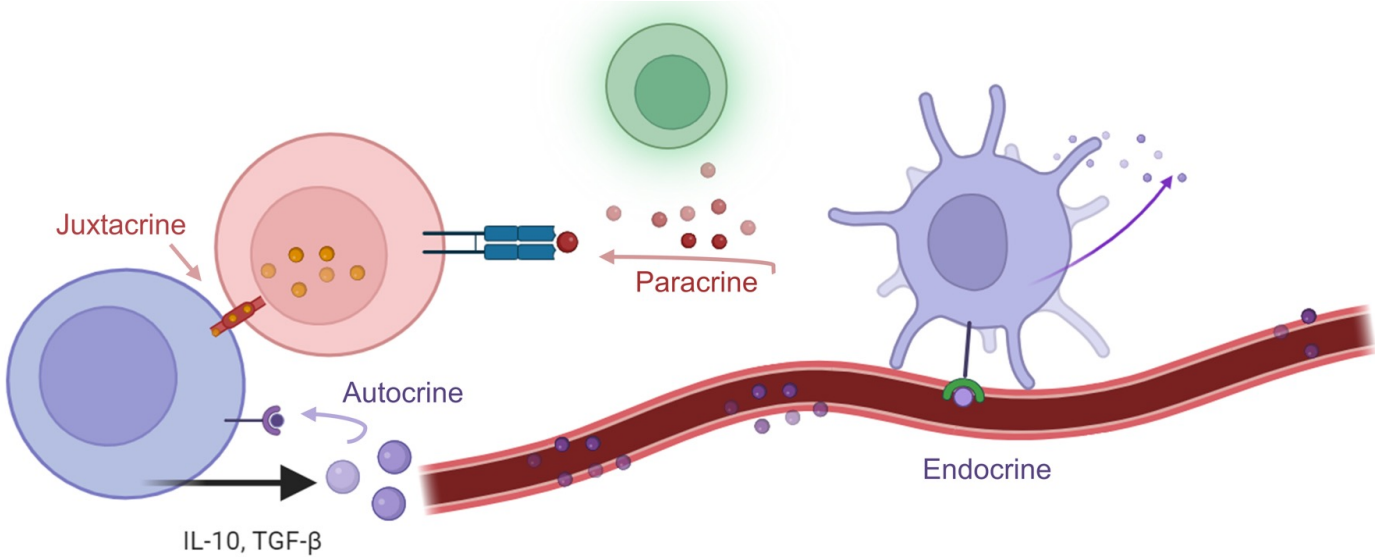
R

python

python

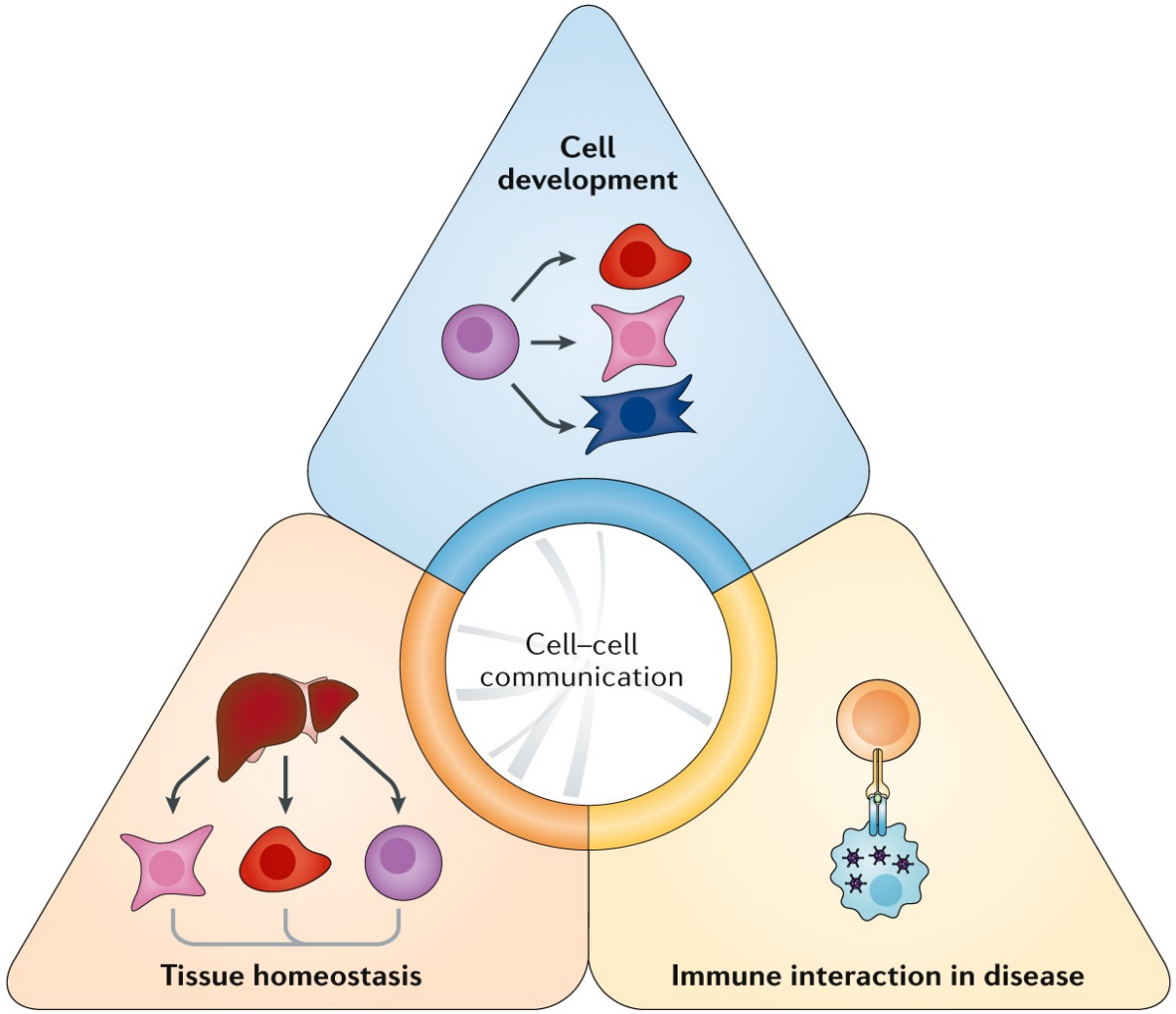
# Module 2 – Part 6: Cell-Cell Interactions

All cells depend on cell-to-cell interactions to identify and respond to stimuli in their surroundings and therefore share a microenvironment.




**Autocrine signalling** - Intracellular signalling
**Paracrine signals** - Between nearby cells
**Juxtacrine signals** - Contact-dependent or gap-junction
**Endocrine signals** - Long-distance intercellular signalling.

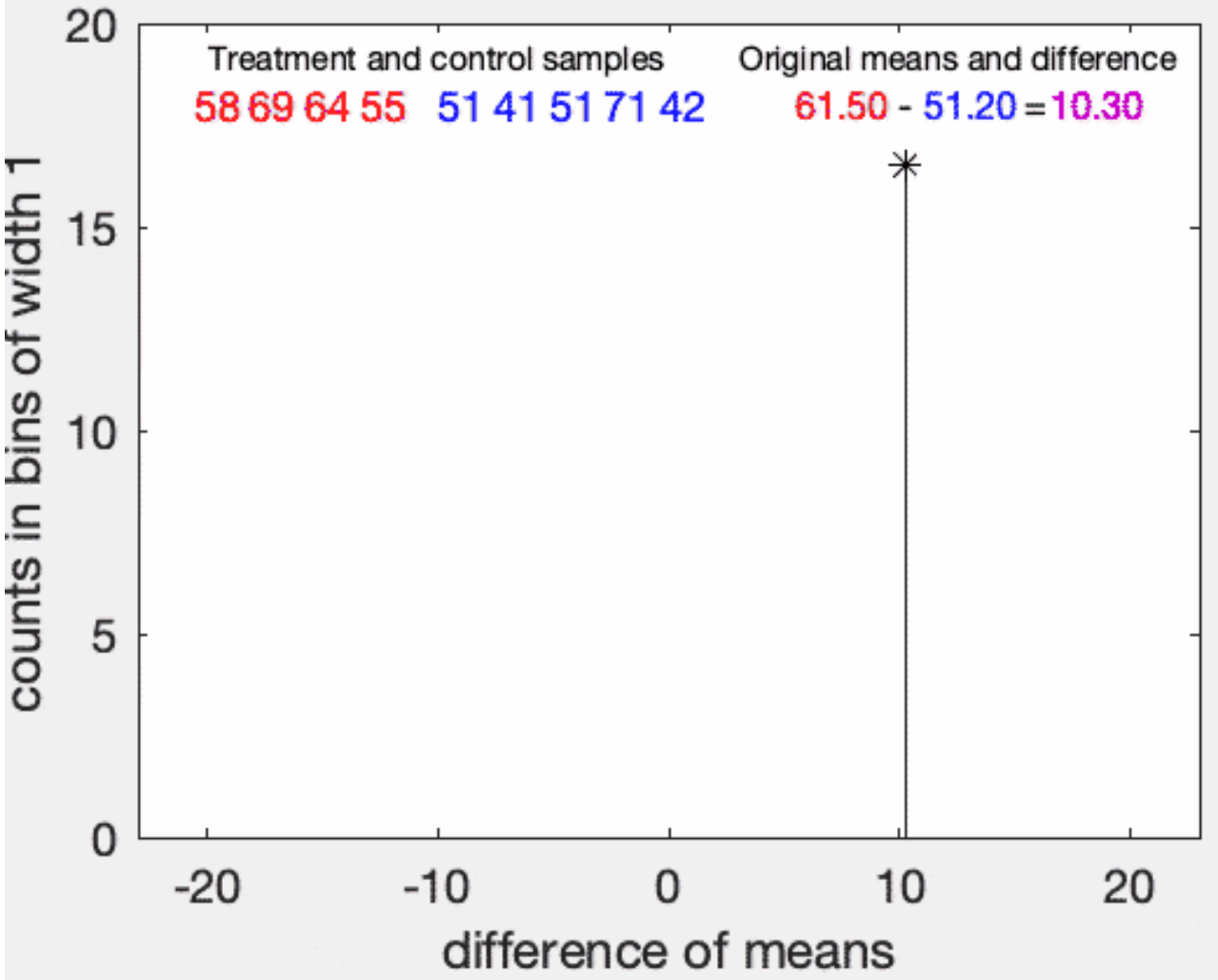
# Importance of CCI



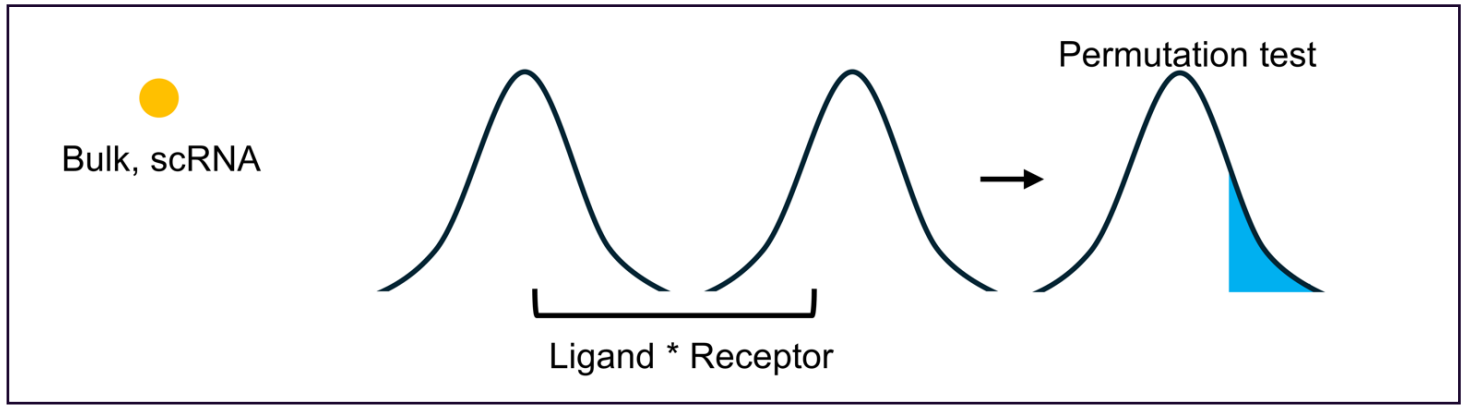

Erick Armingol et al., 2021

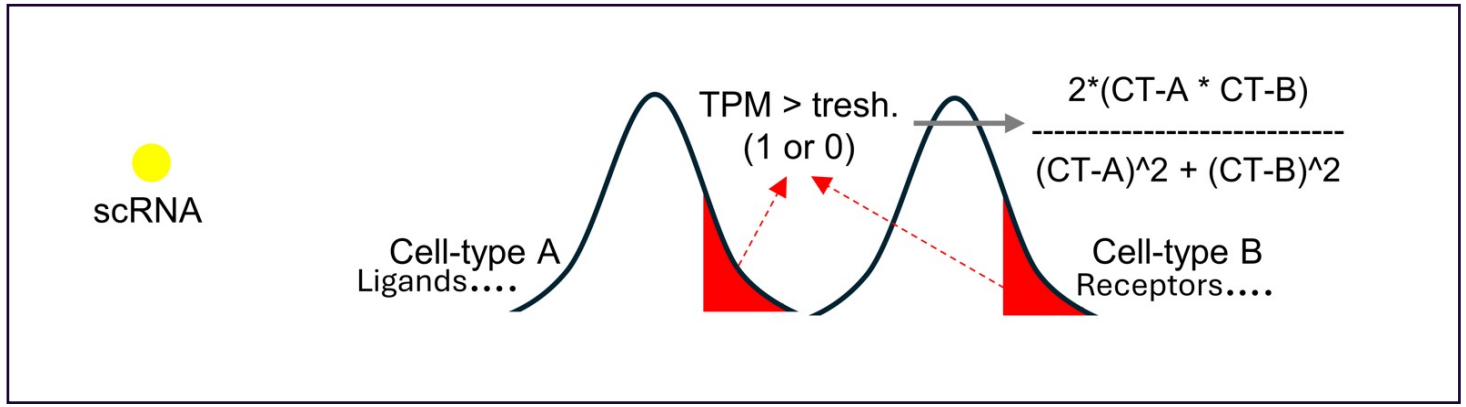
- CCI is essential for the functioning of an individual cell and allows groups of cells to communicate and coordinate to maintain homeostasis.
- When cells fail to interact correctly or misunderstand signals, it can lead to disease.

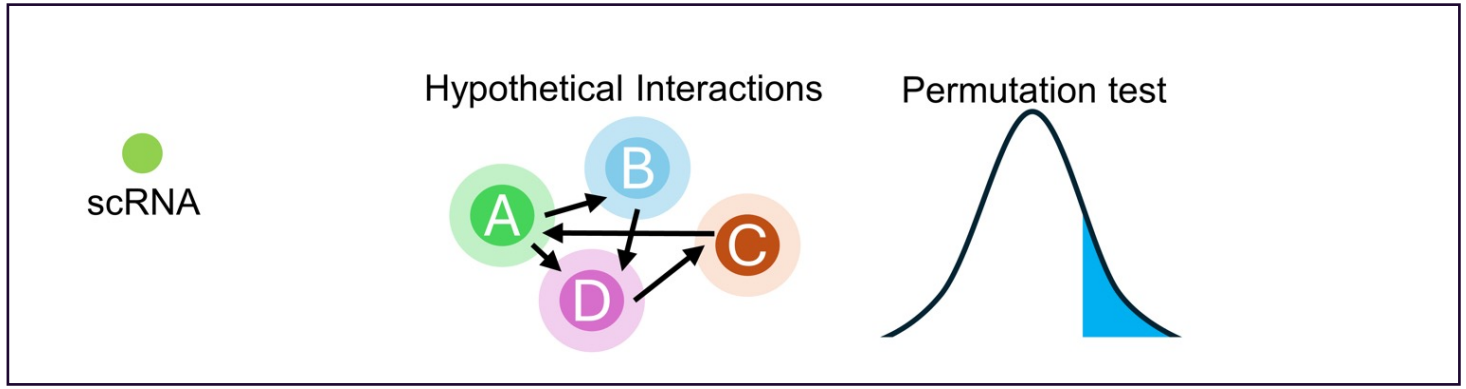
# Pre-requisite: Permutation Testing



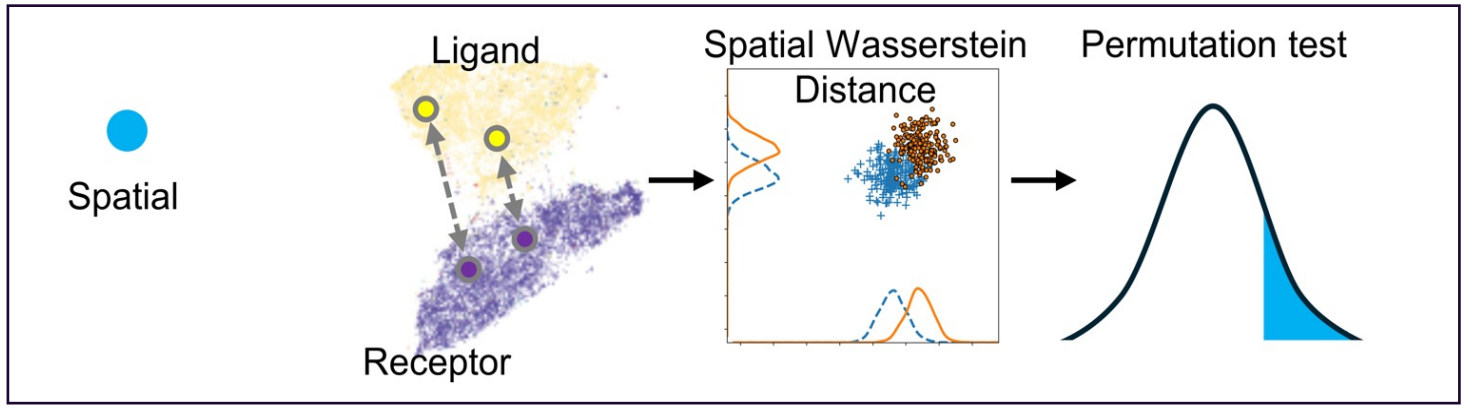
# Common Techniques for CCI

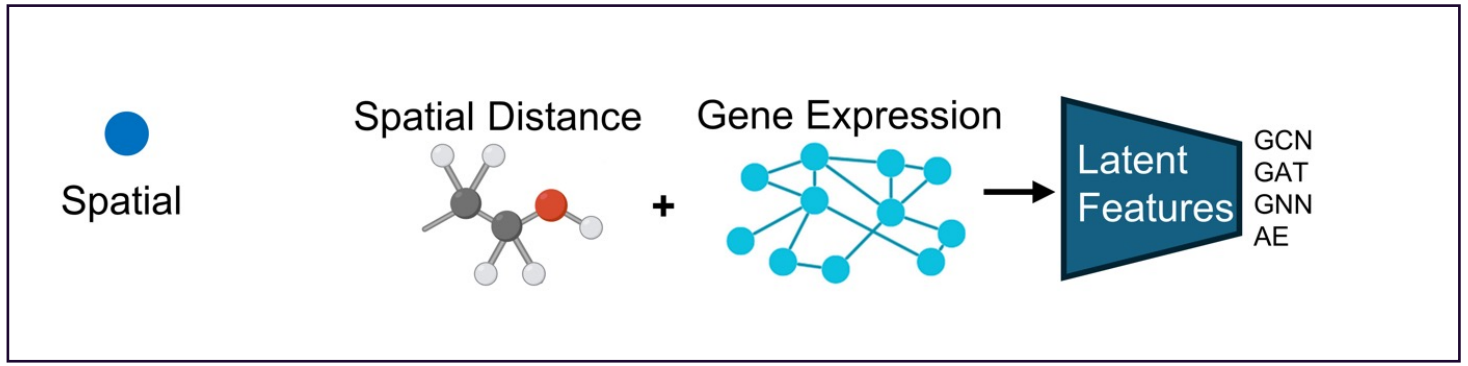


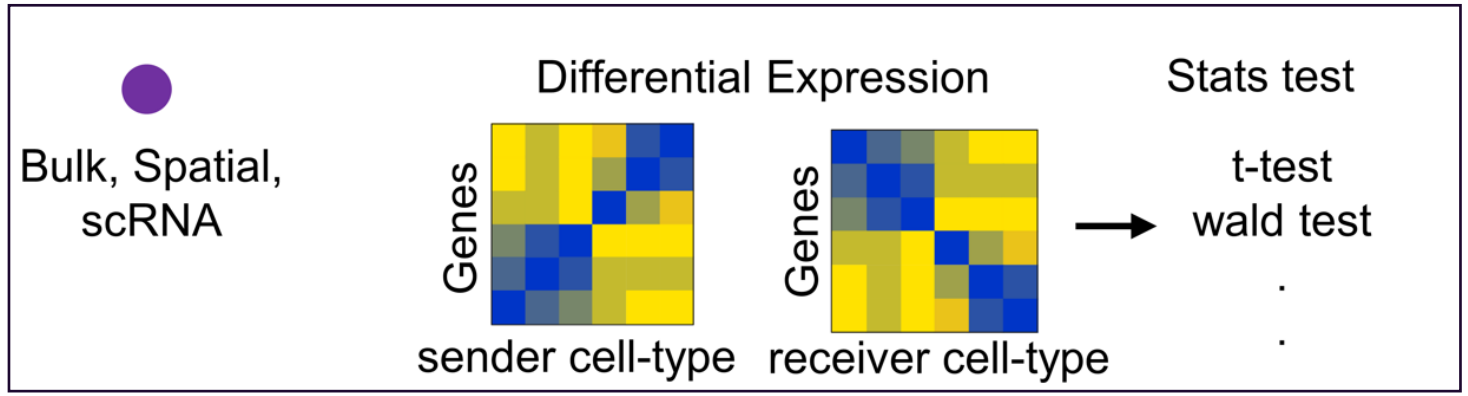



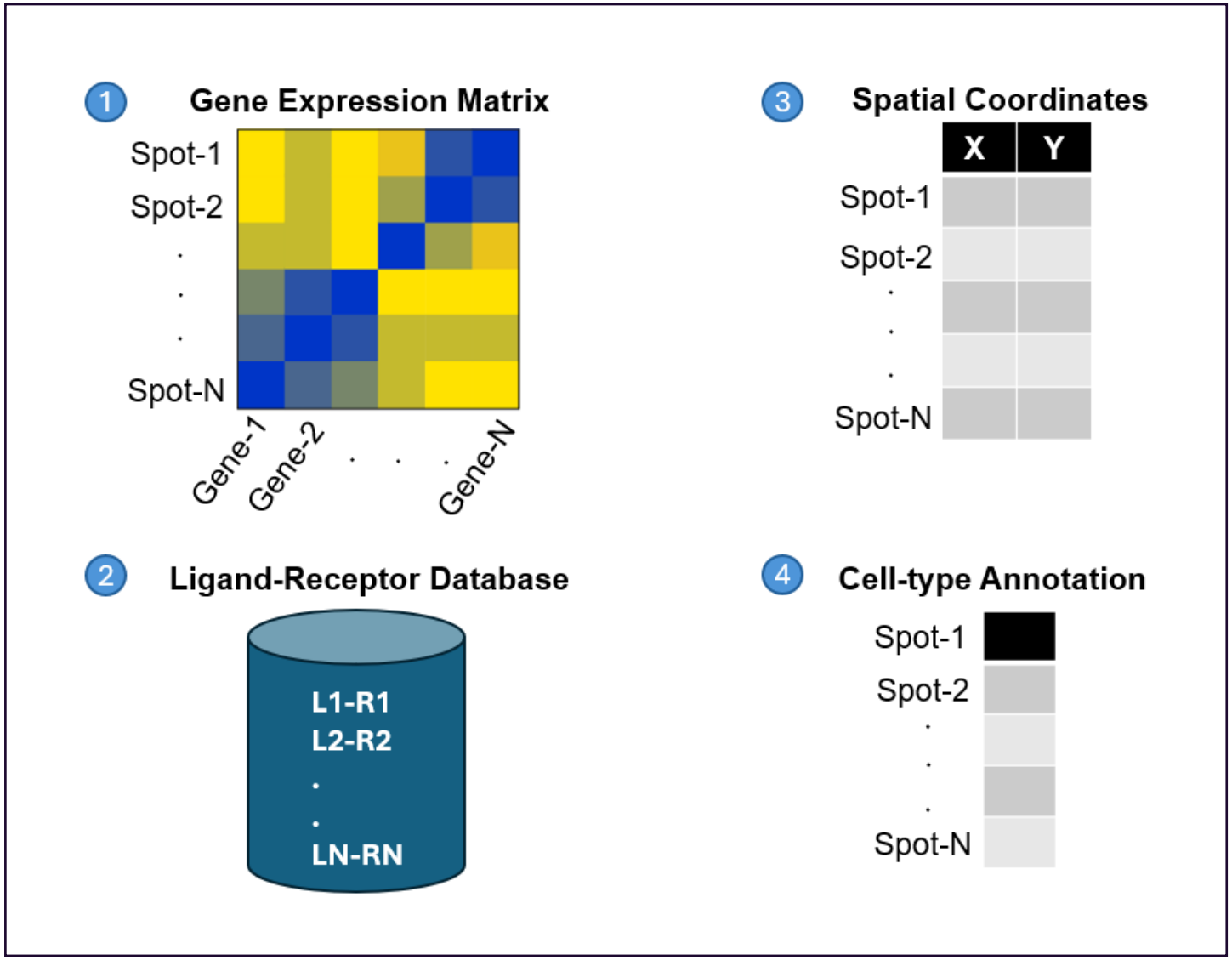
# Common Techniques for CCI



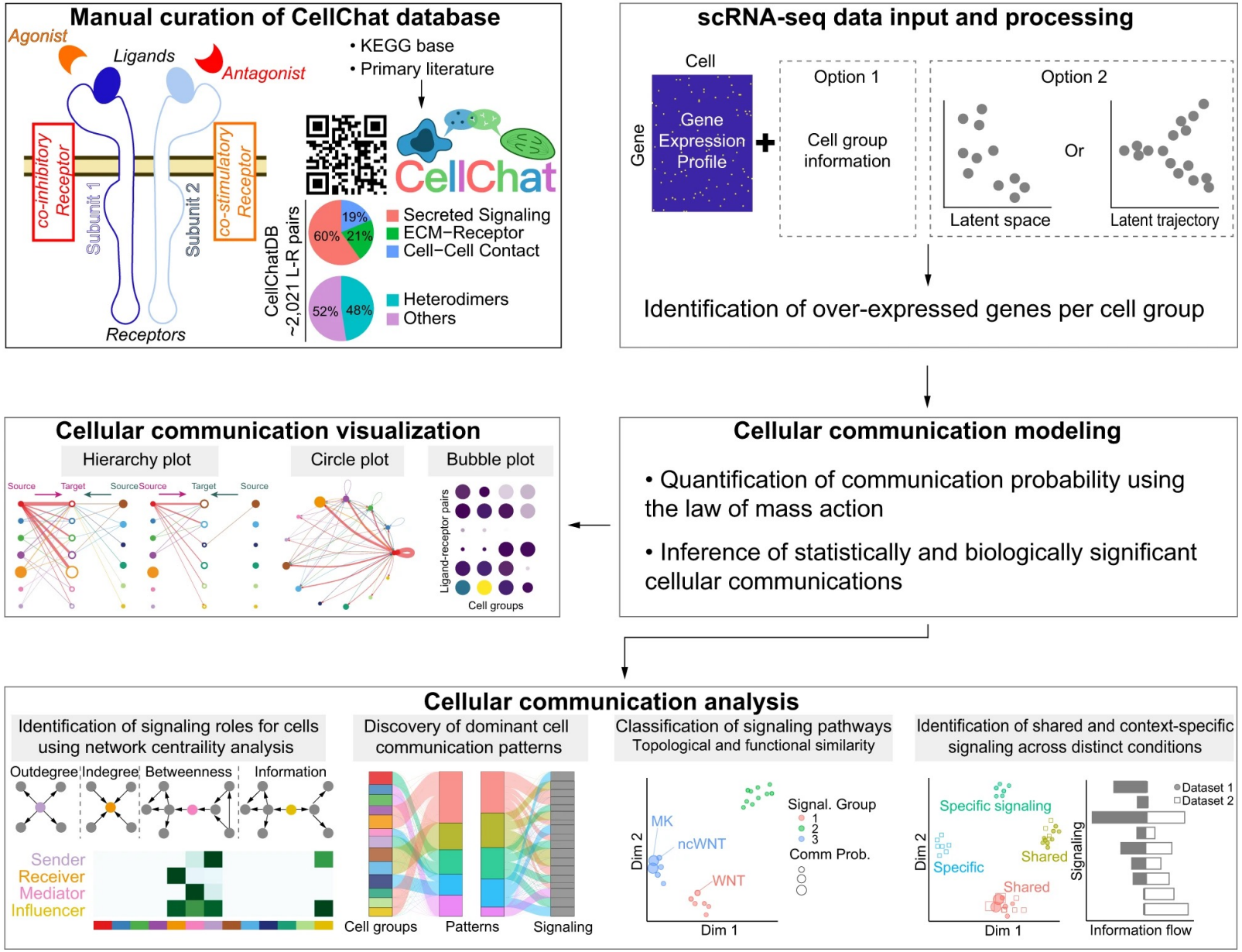



# Data Requirement for CCI

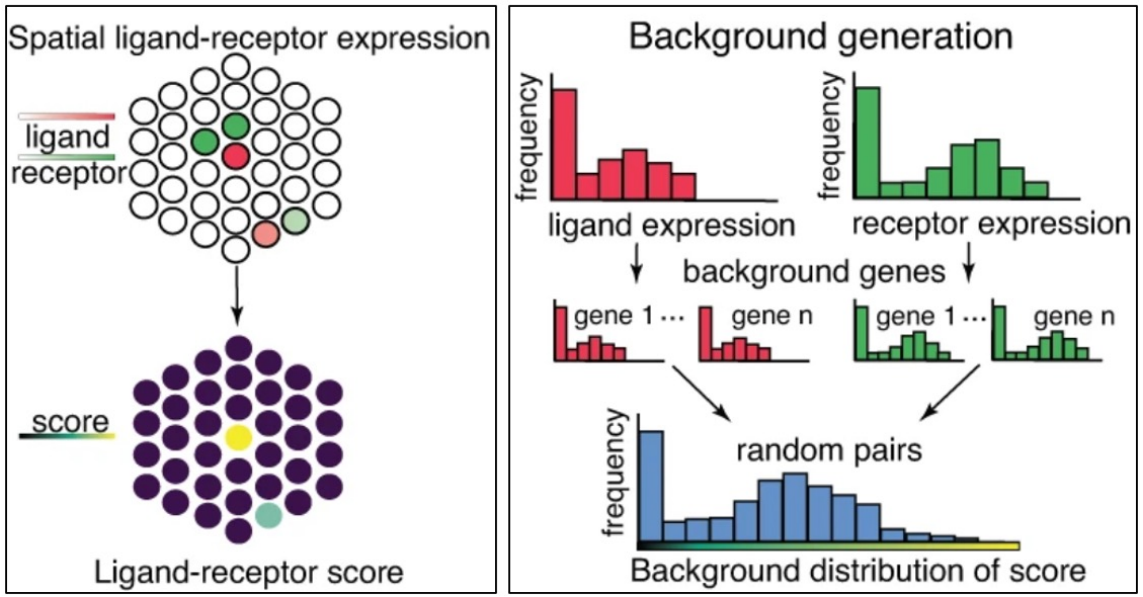


# 1. CellChat

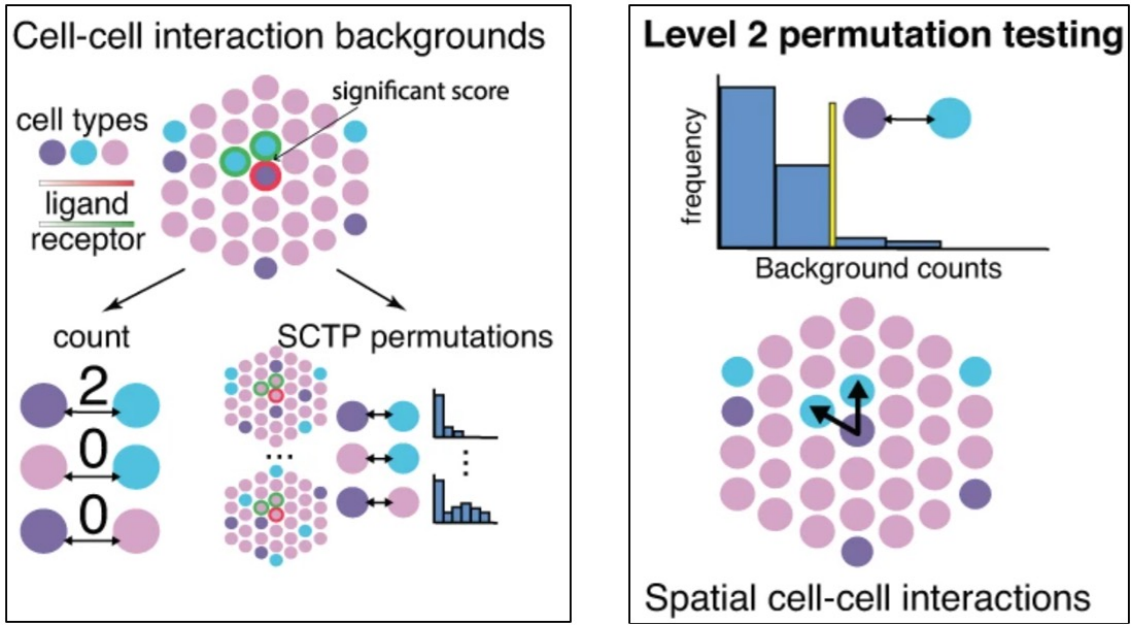
Suoqin Jin et al., 2021



# 2. stLearn



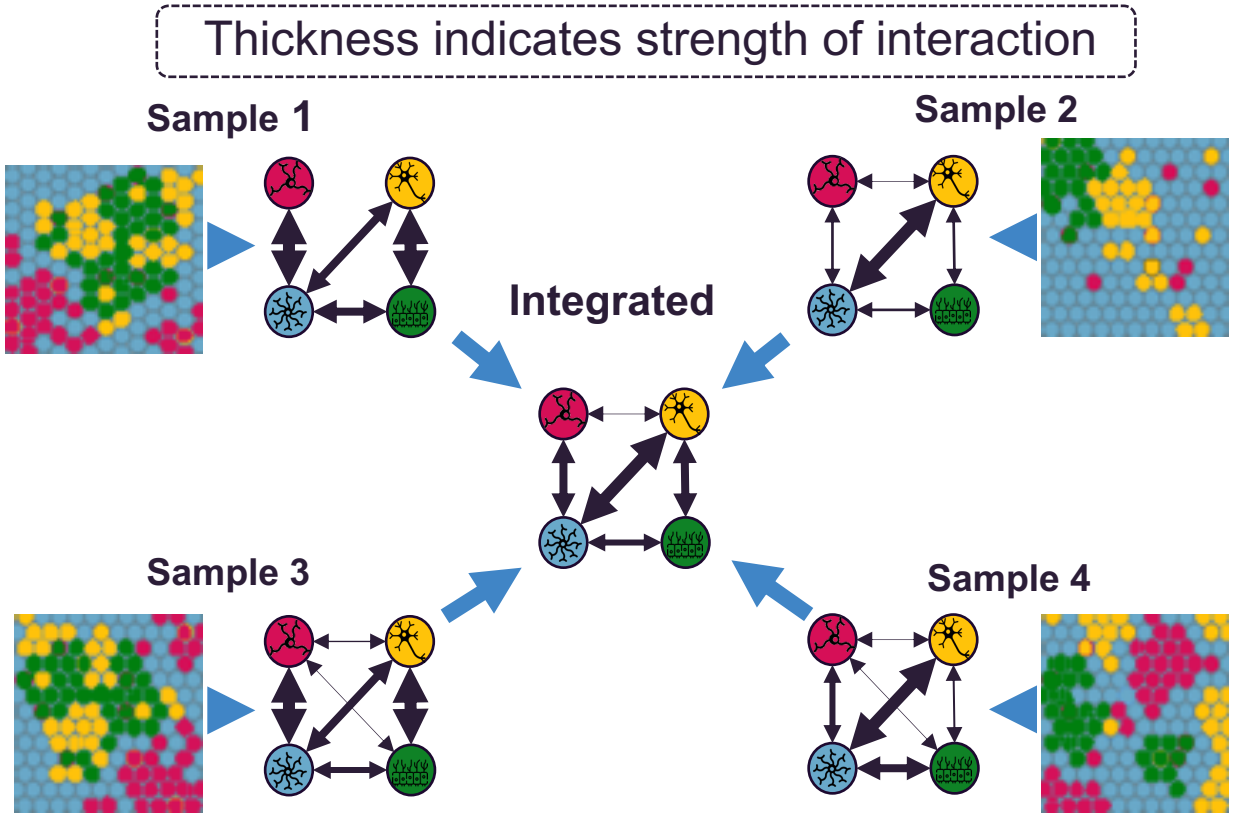

**Ligand-Receptor Interaction**




**Cell-cell Interaction**

Duy Pham et al., 2023

# 3. MMCCI

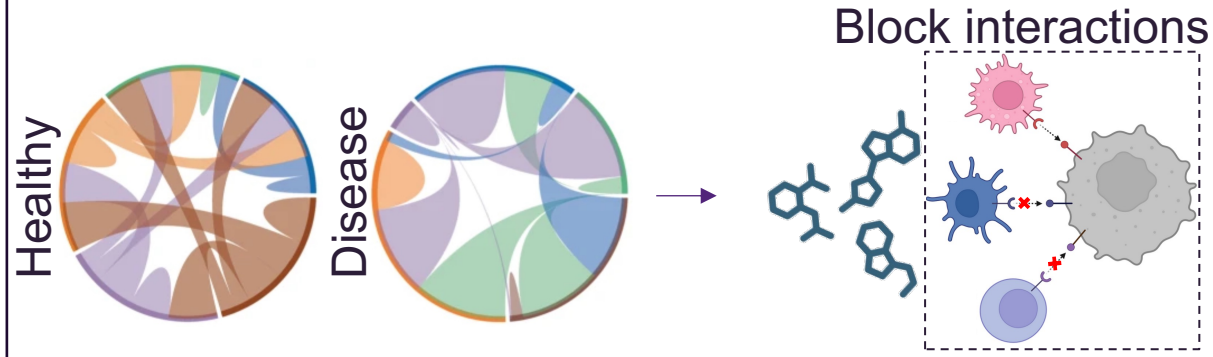



- CCI results can vary highly across individual samples, especially when using multiple modalities.
- MMCCI is a method to integrate CCI results across replicates from multiple modalities.

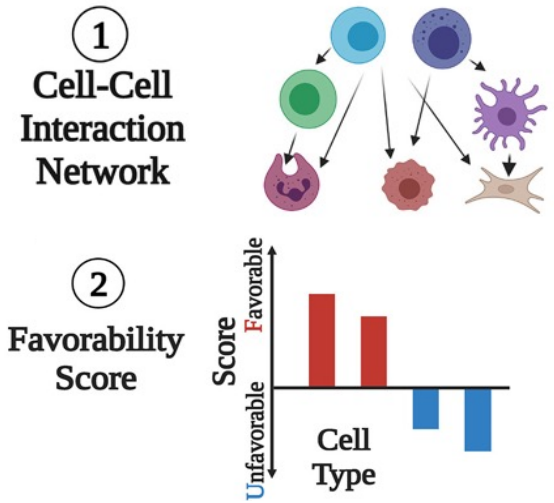
# Applications of CCI

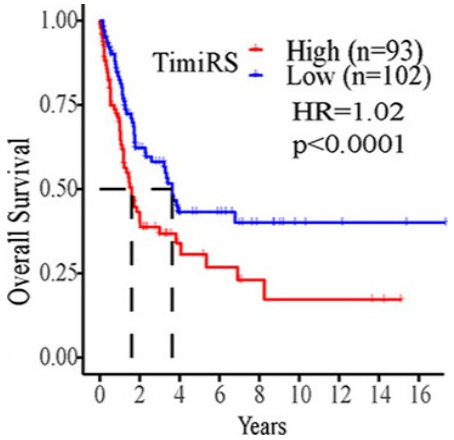


## Cell-Cell Interactions Healthy vs Disease



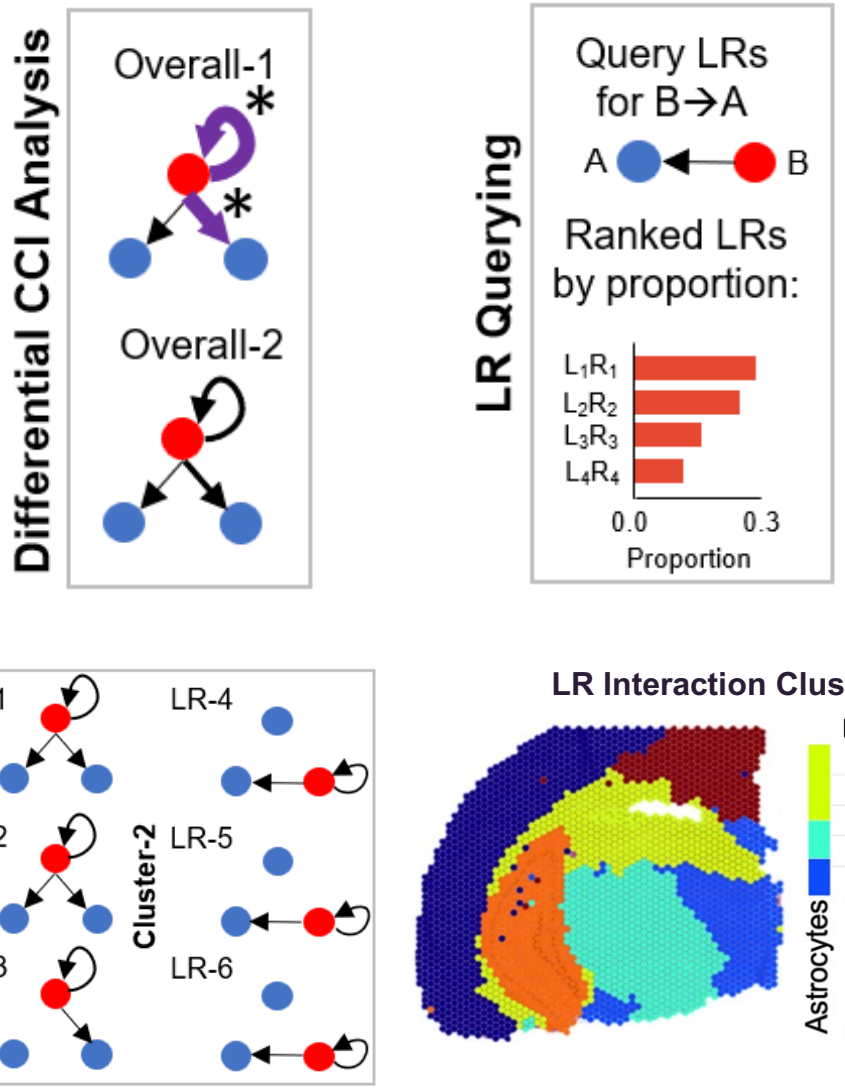

## Using CCI for prognosis




TimiGP - (Chenyang Li et al., 2023)

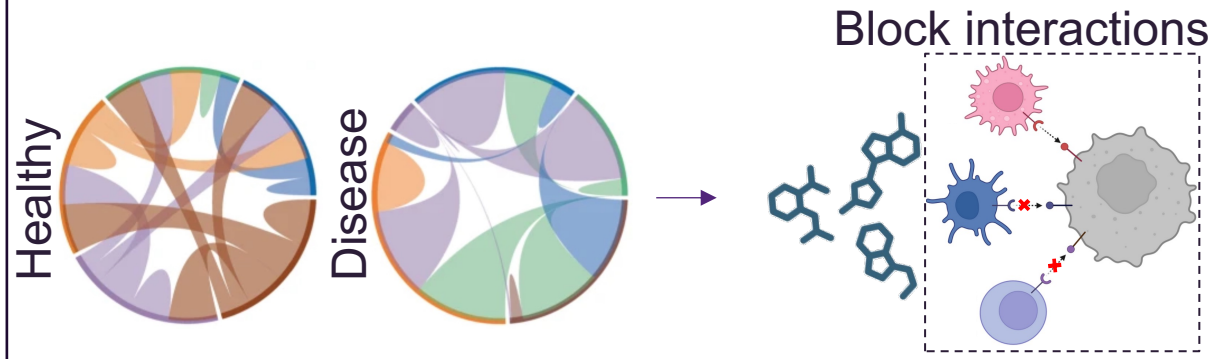
## MMCCI



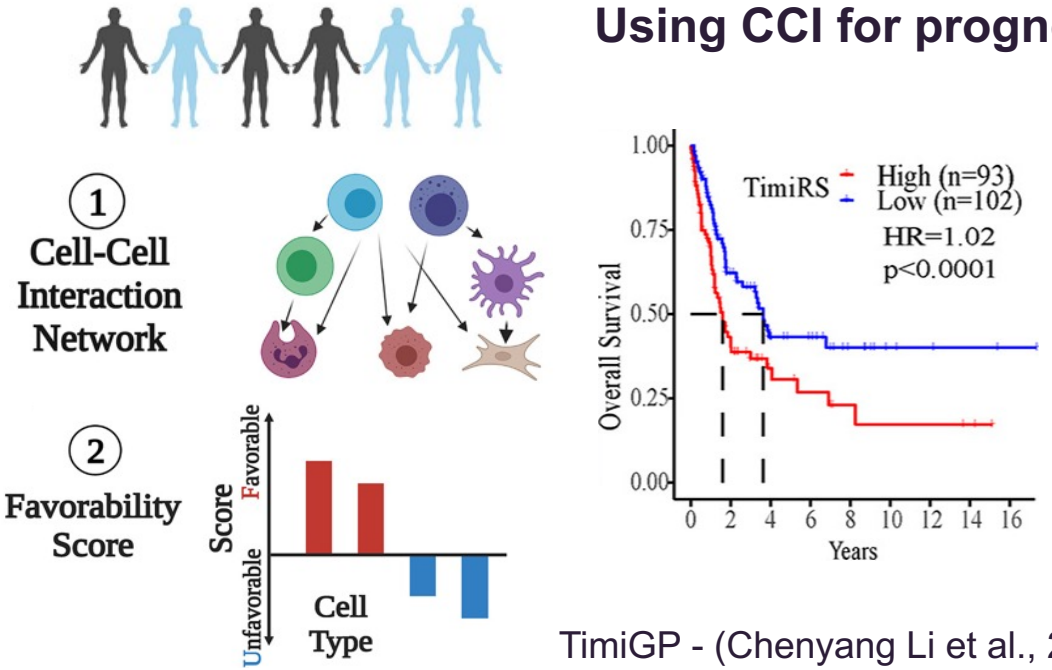
# Applications of CCI



## Cell-Cell Interactions Healthy vs Disease




## Using CCI for prognosis




TimiGP - (Chenyang Li et al., 2023)

## MMCCI