Practical Day 1: Data Structure & Understand Single Cell and Spatial Data Concepts

Prakrithi Pavithra, Xiao Tan, Quan Nguyen

Learning Objectives

- Understand data types: Single cell sequencing (scRNAseq) & spatial imaging/sequencing data
- Understand data structure: scRNAseq & spatial data
- Get to know common software: Seurat, AnnData, SpatialData
- Key concepts in data preprocessing: normalization
- Key analysis types: identify cell types
  - Kmeans, graph-based, deconvolution
  - Annotate cell type labels

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].

Practical 1: Data structure

- Prakrithi Pavithra

1. Data structure
2. Dataset
3. QC and normalisation
4. Clustering

Running the Practical

Terminal

PowerShell

1. Log into your account:

   ssh {username}@203.101.229.126
   username & password from winter school email

2. Follow these commands:

   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2026
   - git clone https://github.com/GenomicsMachineLearning/gml-teaching-2026
   - cd /scratch/$USER/gml-teaching-2026
   - /scratch/$USER/gml-teaching-2025/runme.sh

3. Open JuperterNotebook:

   000-single-cell-RNAseq/000_Single_Cell_RNAseq_Analysis.ipynb

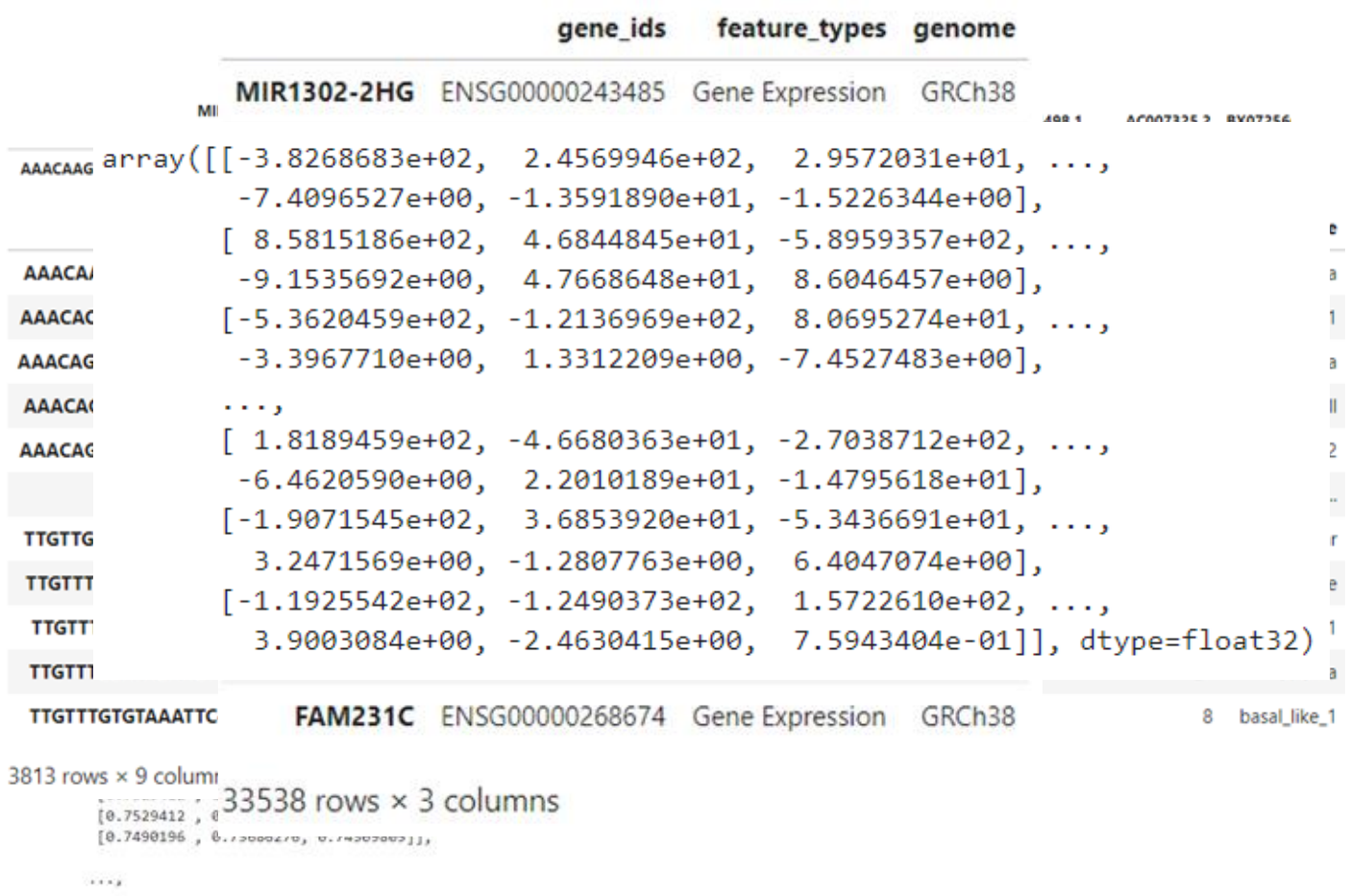
Definition

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

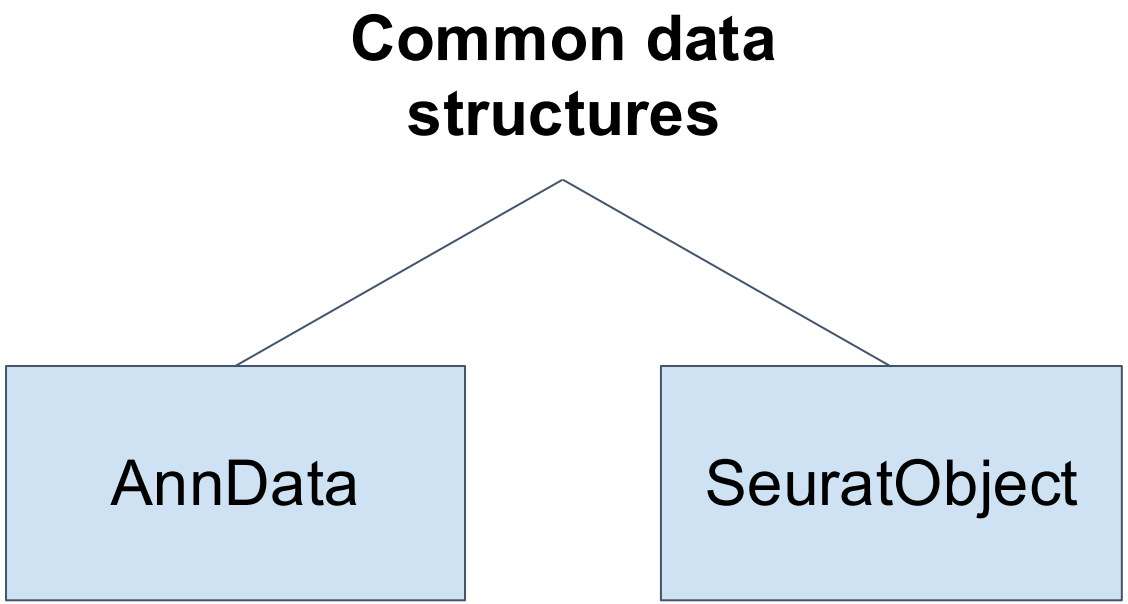
- **Data structure:** specialized format for ***organizing*** and ***storing*** data in memory that contains not only the ***elements*** stored but also ***their relationship*** to each other

scRNAseq or spatial transcriptomics data

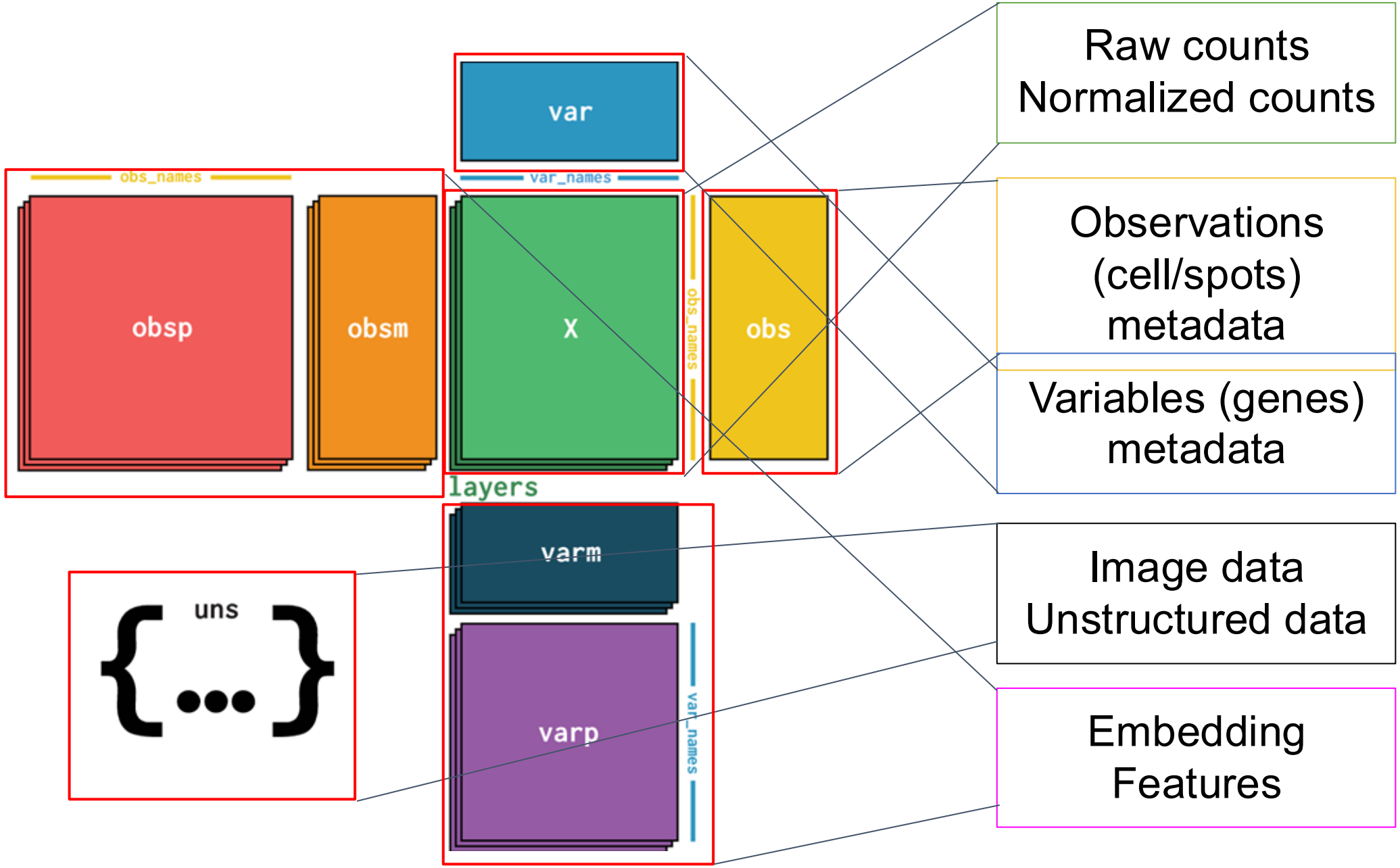
- **Gene expression matrix:**
  - Row: cells/spots
  - Column: genes
- **Cells/spots metadata:**
  - Cell type
  - Batch
  - Spatial coordinates
  - …
- **Genes metadata:**
  - Reference
  - Ensembl ID
  - …
- **Image:**
  - H&E image
- **Embedding**
  - PCA
  - UMAP



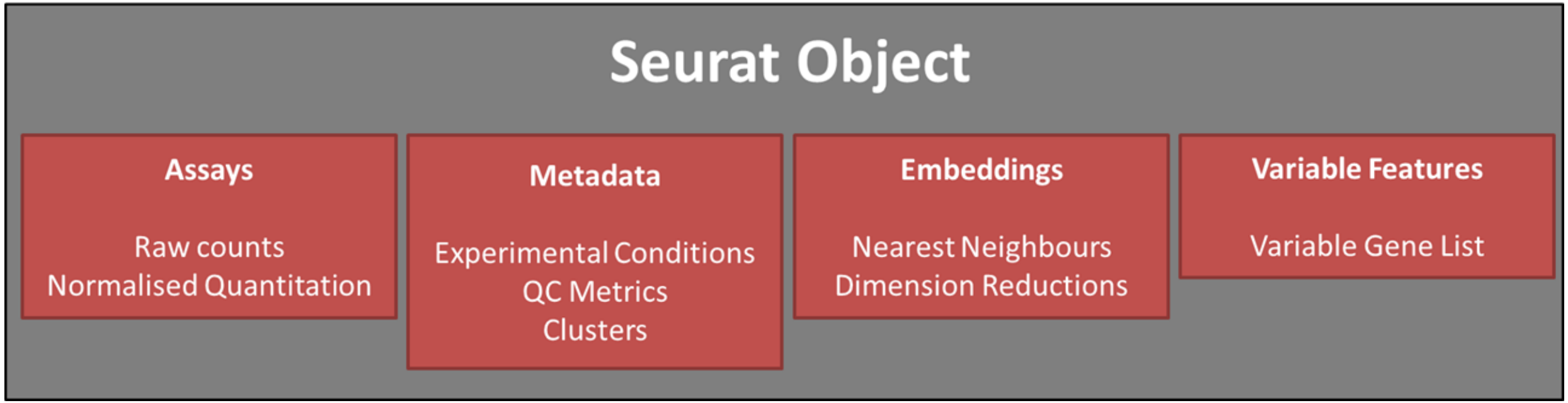
Common data structures



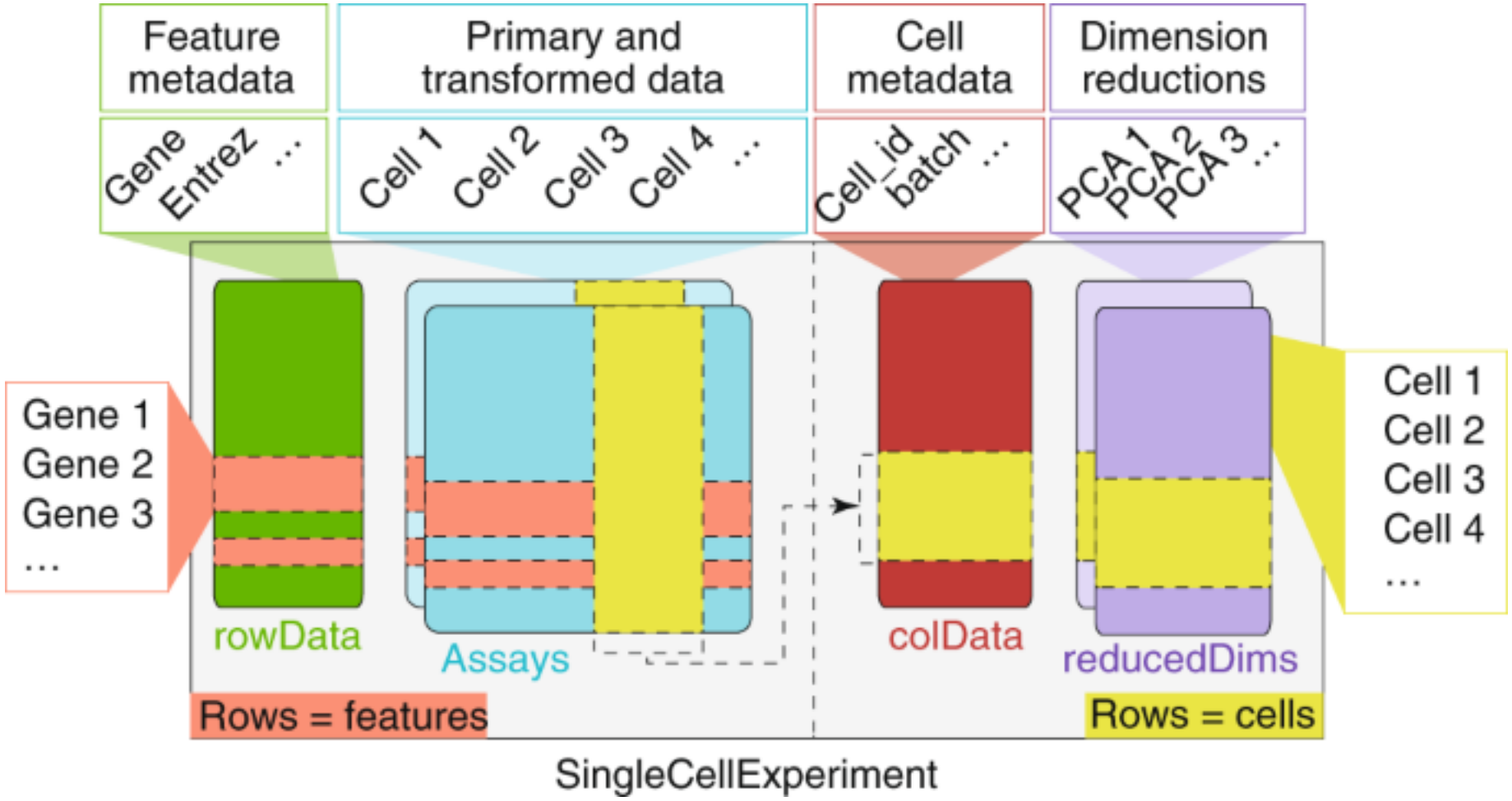
AnnData (Annotated data) - Python



SeuratObject - R



SeuratObject - R

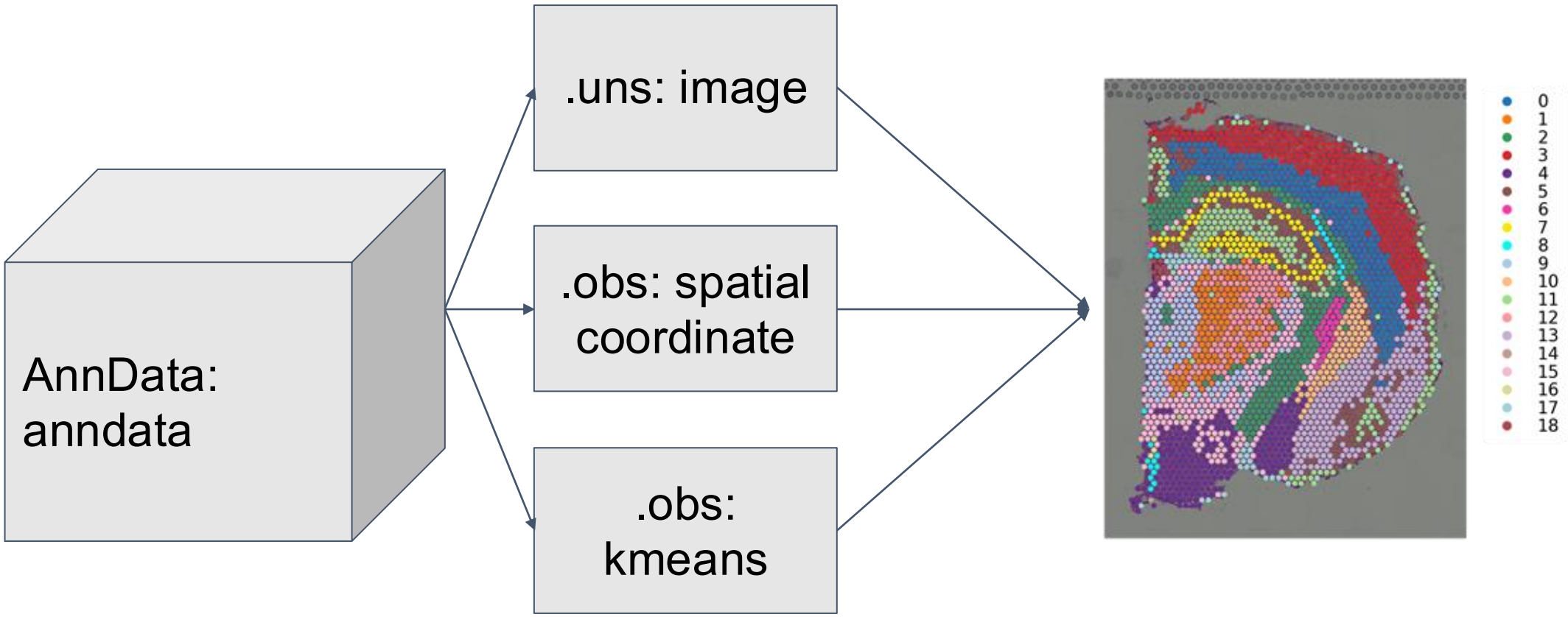


Use case:
Perform K-means clustering and store to AnnData

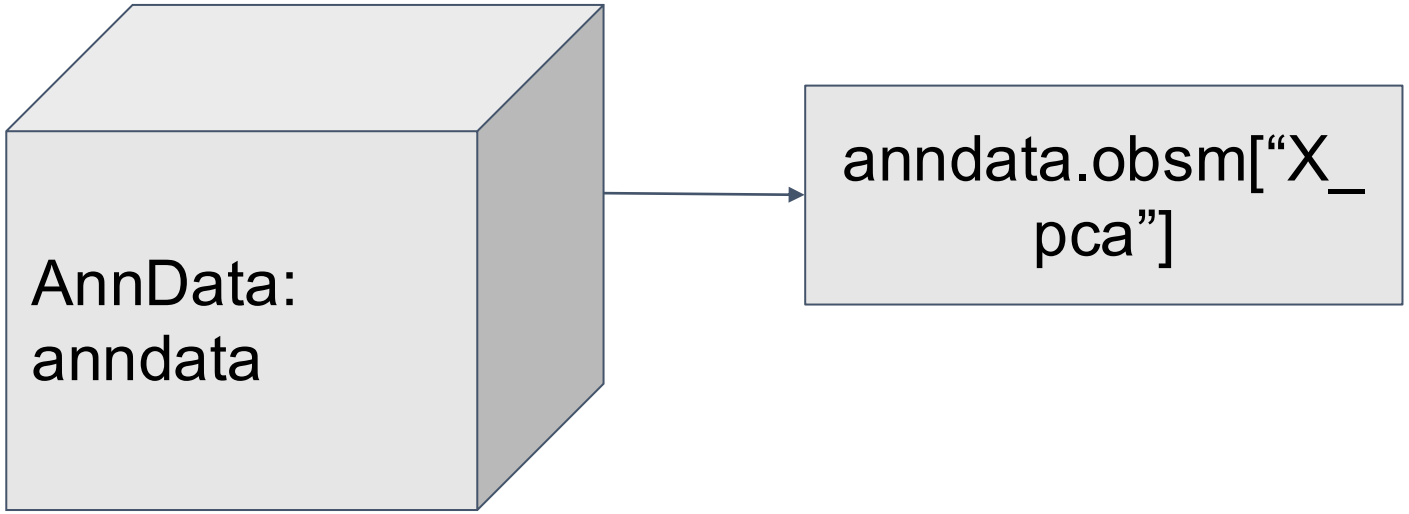
How?

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

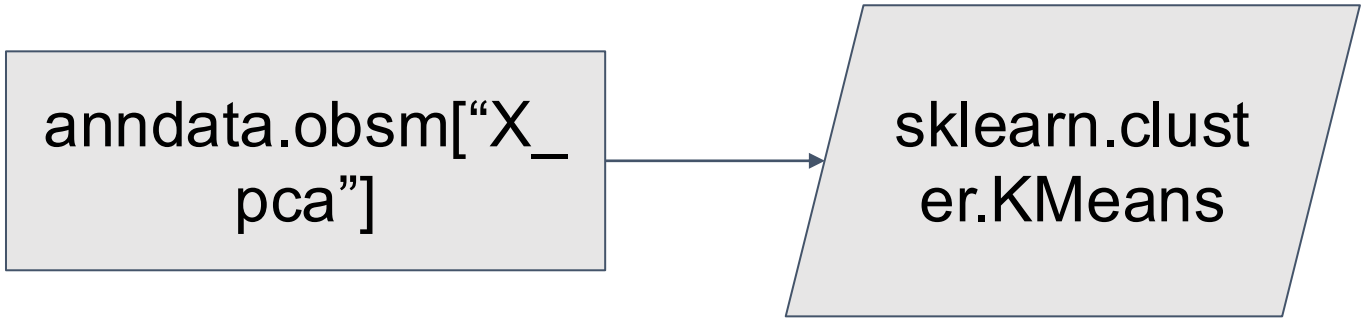
Use case:
Plotting Kmeans results for spatial transcriptomics



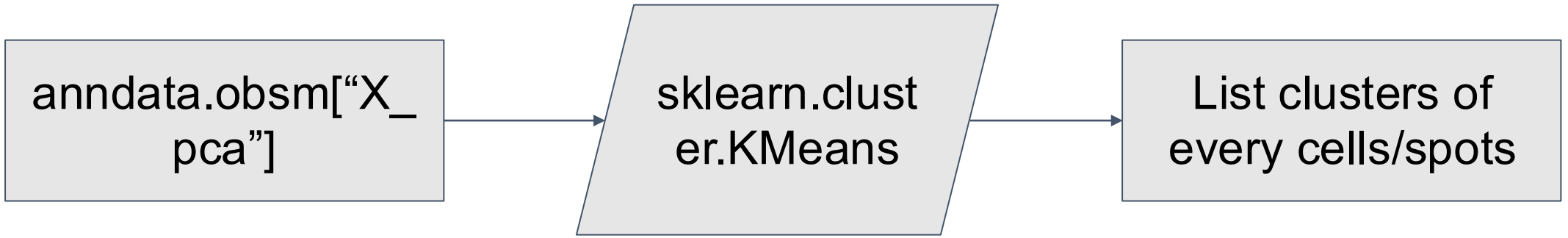
1. Extract the PCs components from AnnData for every cells/spots



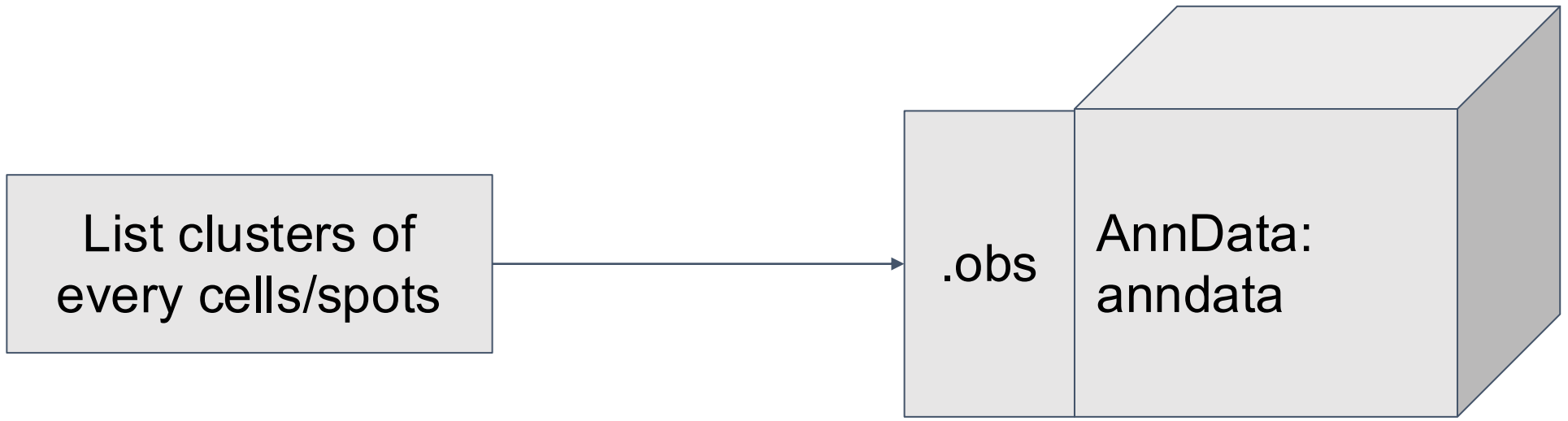
2. Using external scikit-learn package for K-means clustering



3. Get the K-means clustering results



4. Add results to observation annotation of AnnData object



Practical 1: Data structure – Single Cell Analysis

- Prakrithi Pavithra

1. Data structure
2. Dataset
3. QC and normalisation
4. Clustering

Running the Practical

Terminal

PowerShell

1. Log into your account:

   ssh {username}@203.101.229.126
   username & password from winter school email

2. Follow these commands:

   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2026
   - git clone https://github.com/GenomicsMachineLearning/gml-teaching-2026
   - cd /scratch/$USER/gml-teaching-2026
   - /scratch/$USER/gml-teaching-2026/runme.sh

3. Open JuperterNotebook:

   002-clustering-cell-typing

Datasets – PBMC (scRNA-seq)

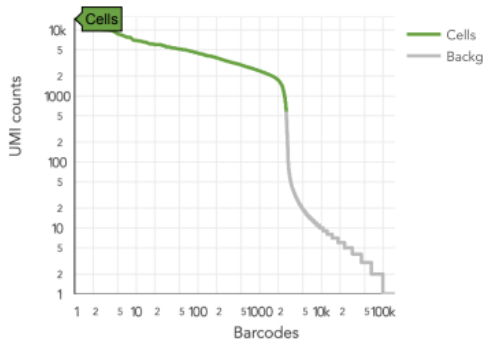
Number of Cells
2,691

Mean Reads per Cell	Median Genes per Cell
69,112	806

Sequencing

Number of Reads	185,980,783
Valid Barcodes	94.5%
Reads Mapped Confidently to Transcriptome	64.0%
Reads Mapped Confidently to Intergenic Regions	5.4%
Reads Mapped Confidently to Intronic Regions	15.8%
cDNA PCR Duplication	94.0%
Q30 Bases in Barcode	70.9%
Q30 Bases in Read 1	93.0%
Q30 Bases in Sample Index	97.2%
Q30 Bases in UMI	96.0%

Cells



Number of Cells	
Fraction Reads in Cells	
Mean Reads per Cell	
Median Genes per Cell	
Median UMI Counts per Cell	

Sample

Name	pb
Description	Peripheral blood mononuclear cells (PBMCs) healthy
Transcriptome	
Cell Ranger Version	

Peripheral blood mononuclear cells (PBMCs) from a healthy donor

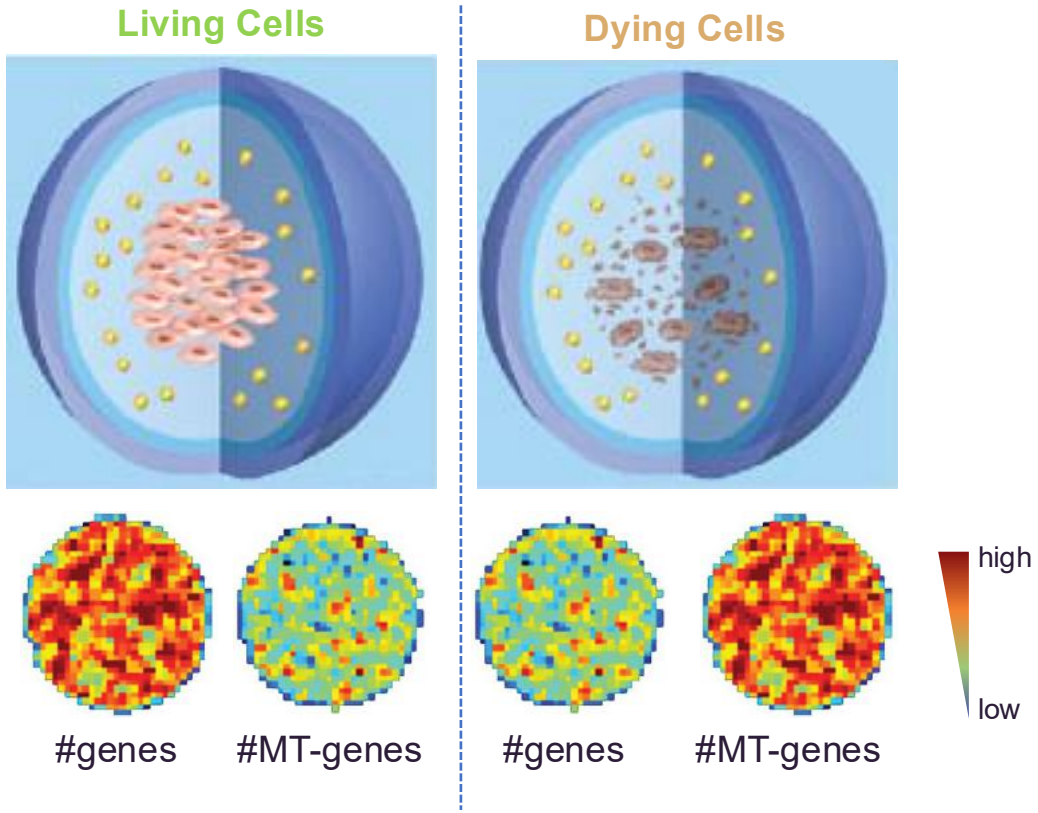
PBMCs are primary cells with relatively small amounts of RNA (~1pg RNA/cell).

- 2,700 cells detected
- Sequenced on Illumina NextSeq 500 with ~69,000 reads per cell
- 98bp read1 (transcript), 8bp I5 sample barcode, 14bp I7 GemCode barcode and 10bp read2 (UMI)
- Analysis run with --cells=3000

- https://www.10xgenomics.com/datasets/3-k-pbm-cs-from-a-healthy-donor-1-standard-1-0-0

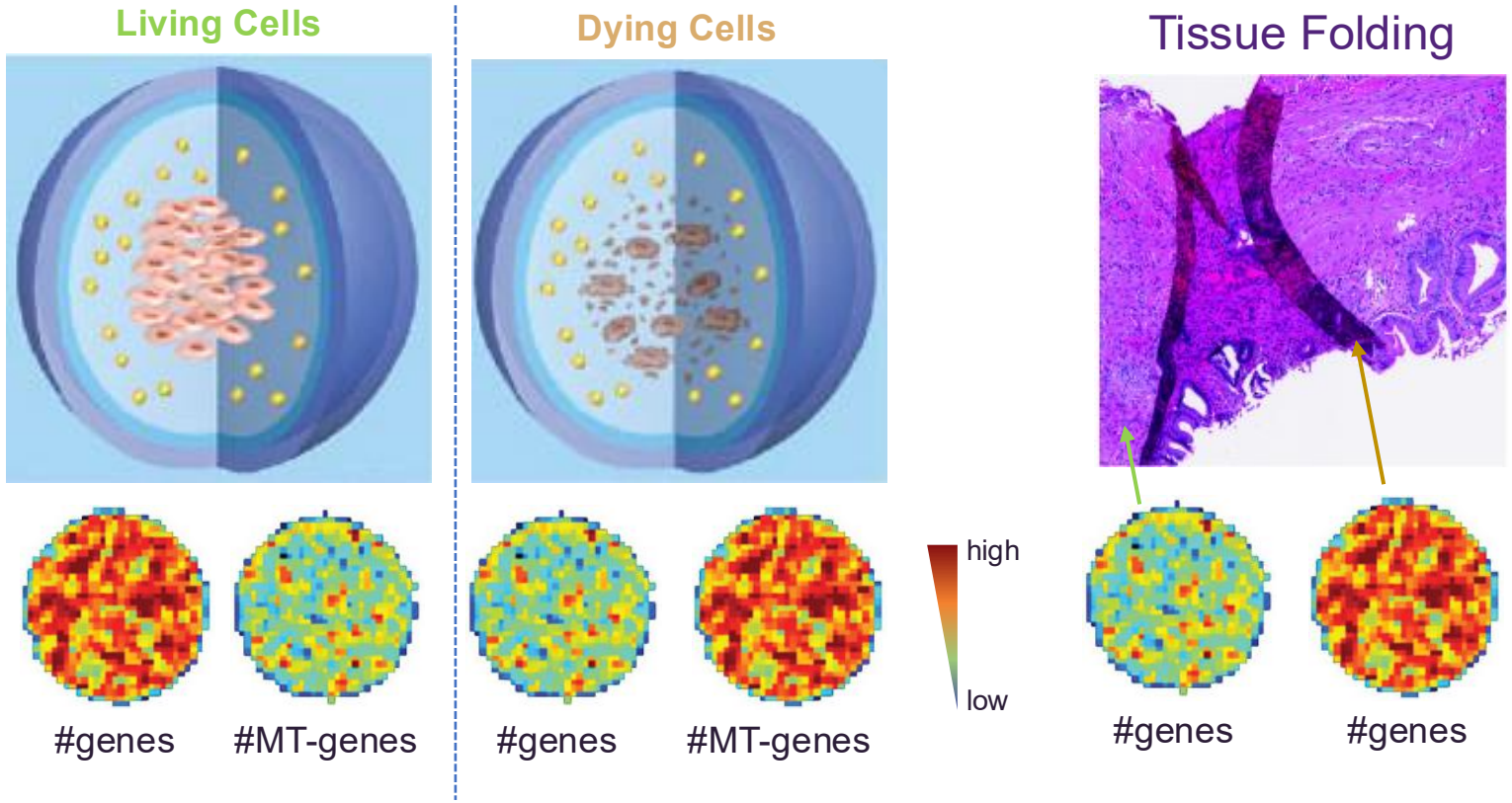
1. Data QC and Normalisation

- *Factors of Technical Noise*



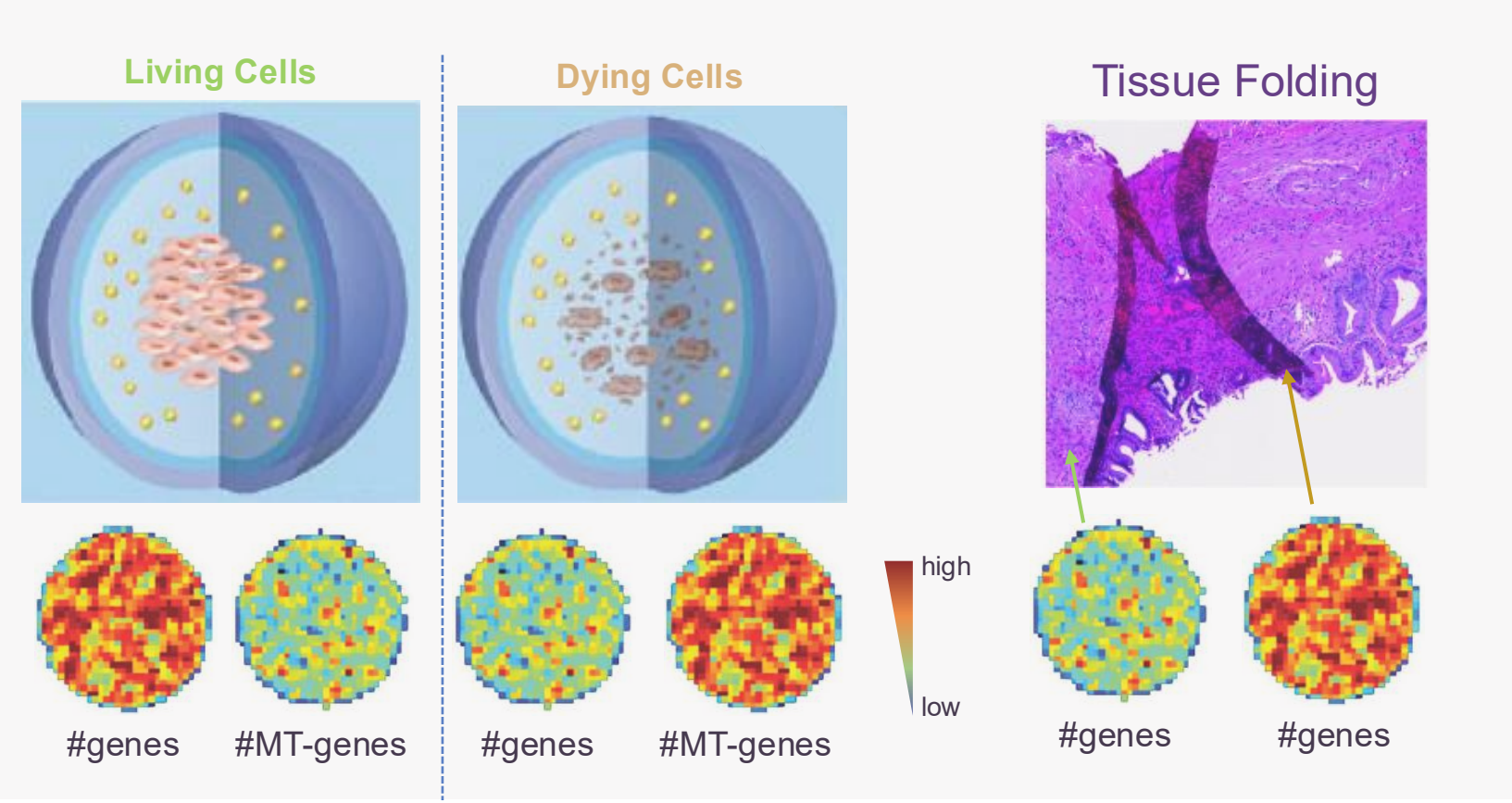
1. Data QC and Normalisation

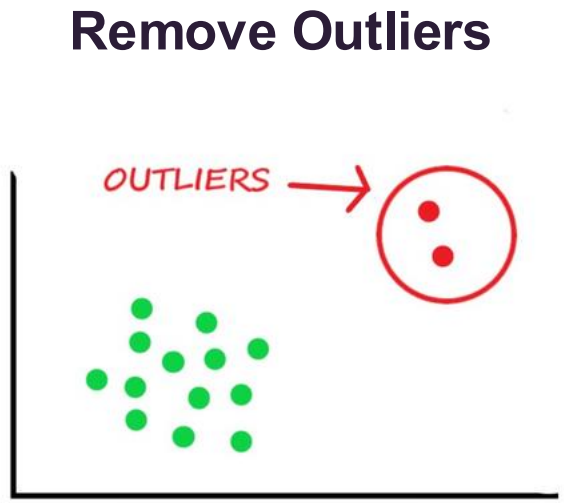
- *Factors of Technical Noise*



1. Data QC and Normalisation

- *Factors of Technical Noise*



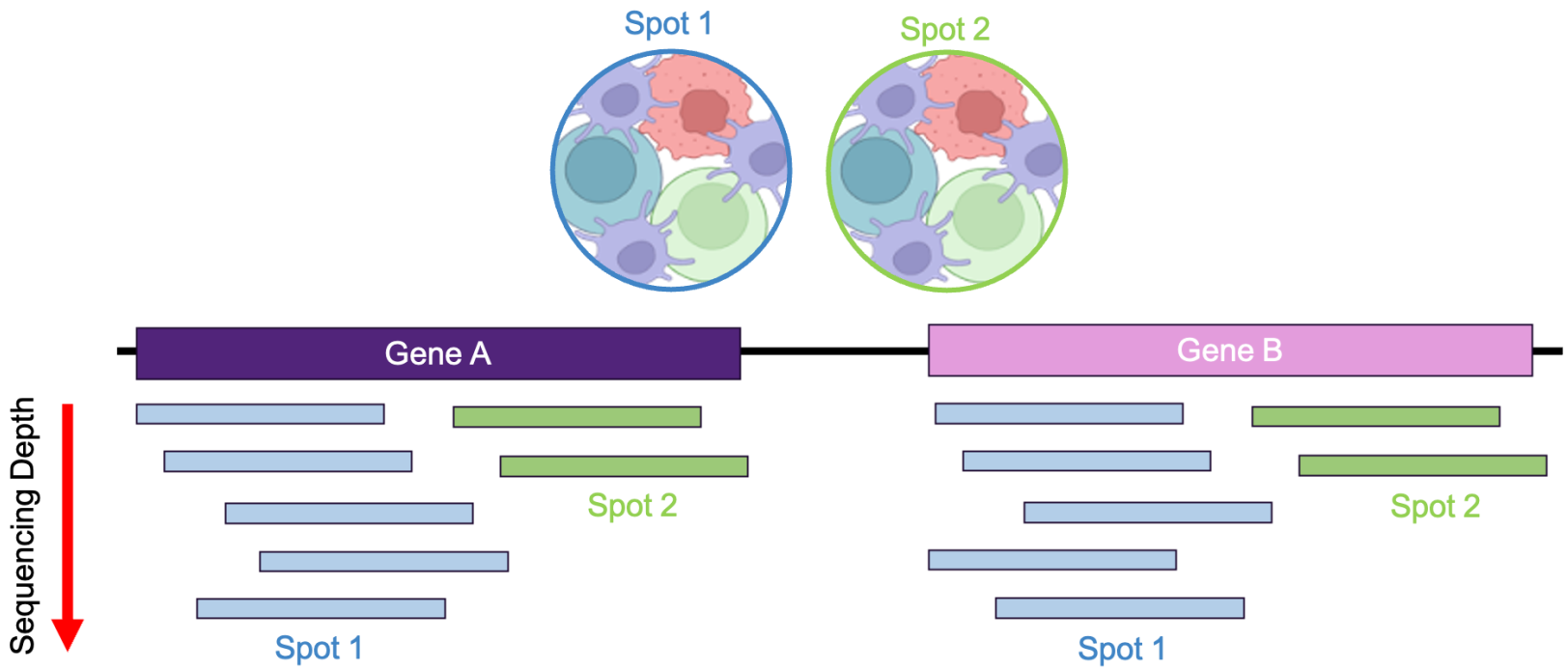


1. Data QC and Normalisation

- *Data Normalisation*

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

- *Technical noise*: capture efficiency/sequencing depth

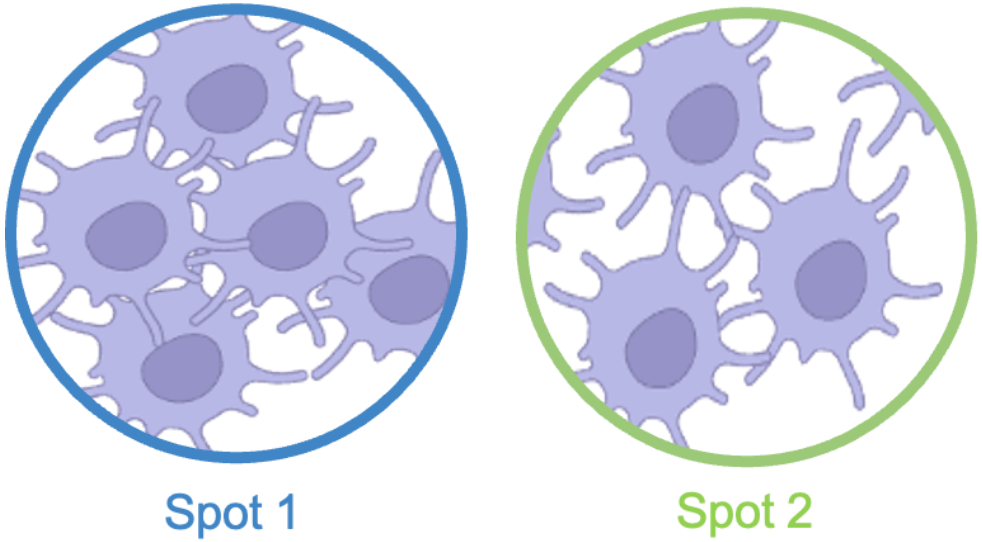


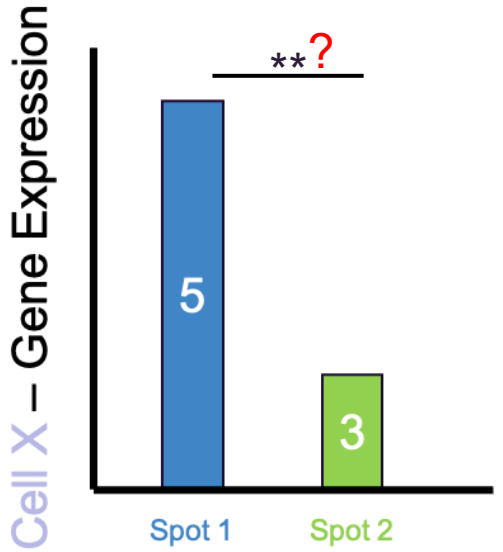
1. Data QC and Normalisation

- *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





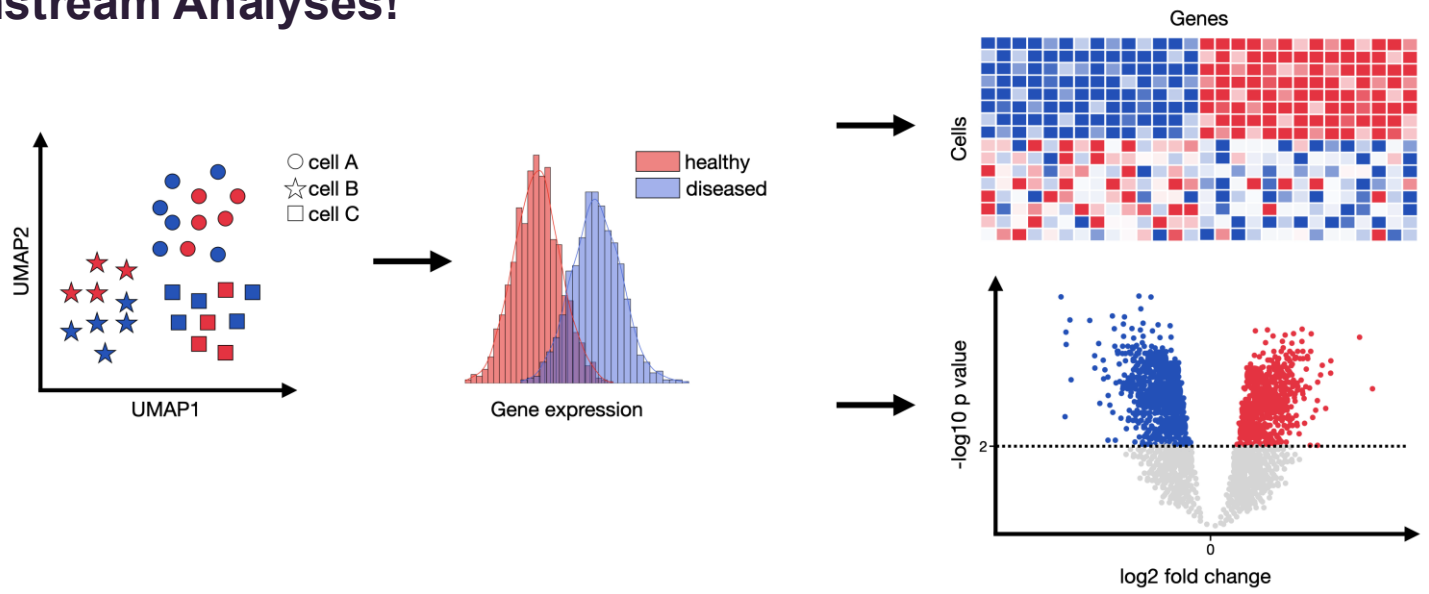
1. Data QC and Normalisation

- *Data Normalisation*

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

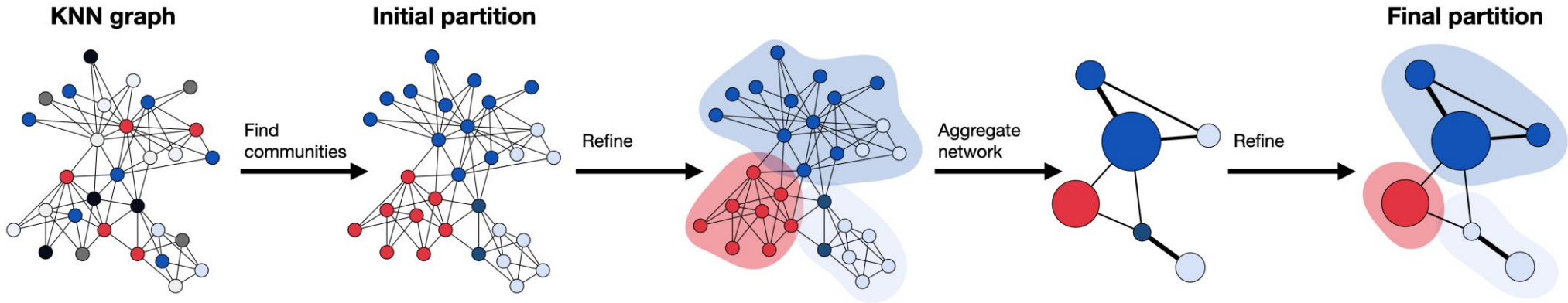
- *Technical noise*: capture efficiency/sequencing depth
- Biological effects: Spots may contain varying numbers of cells

Need for Downstream Analyses!



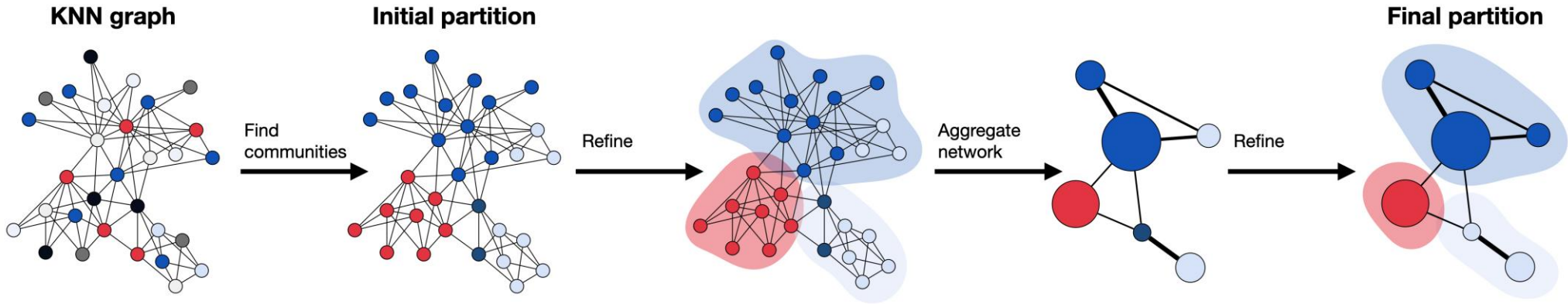
2. Clustering and Cell Typing

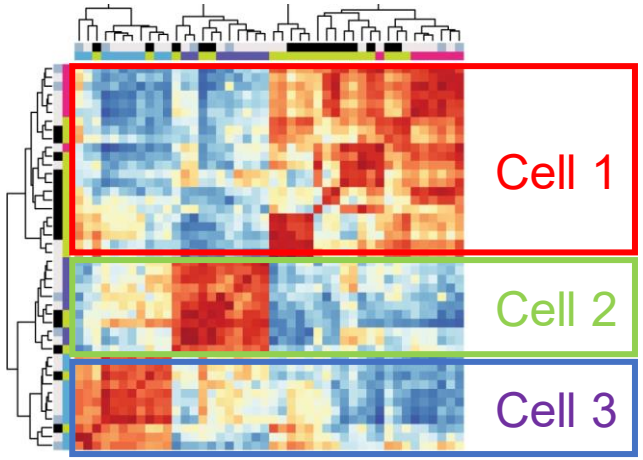
Groups Spots/Cells together based on similar transcriptional patterns



2. Clustering and Cell Typing

Groups Spots/Cells together based on similar transcriptional patterns



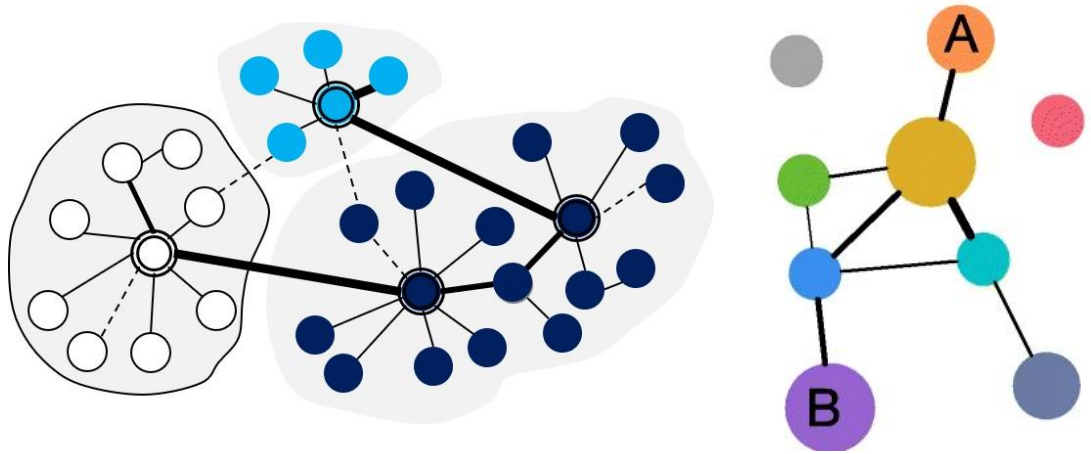


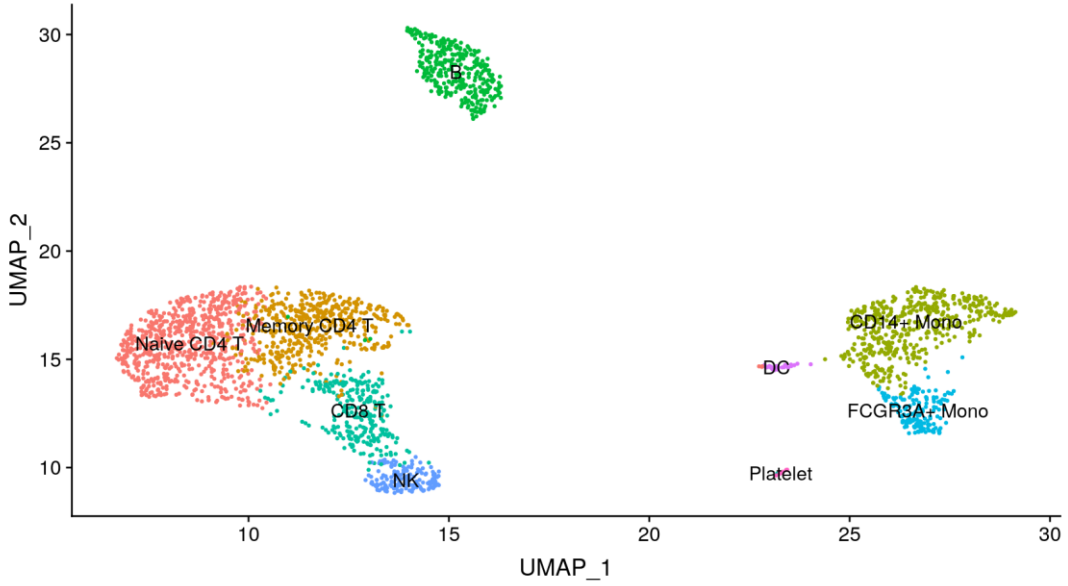
More about: Graph-based Clustering

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)

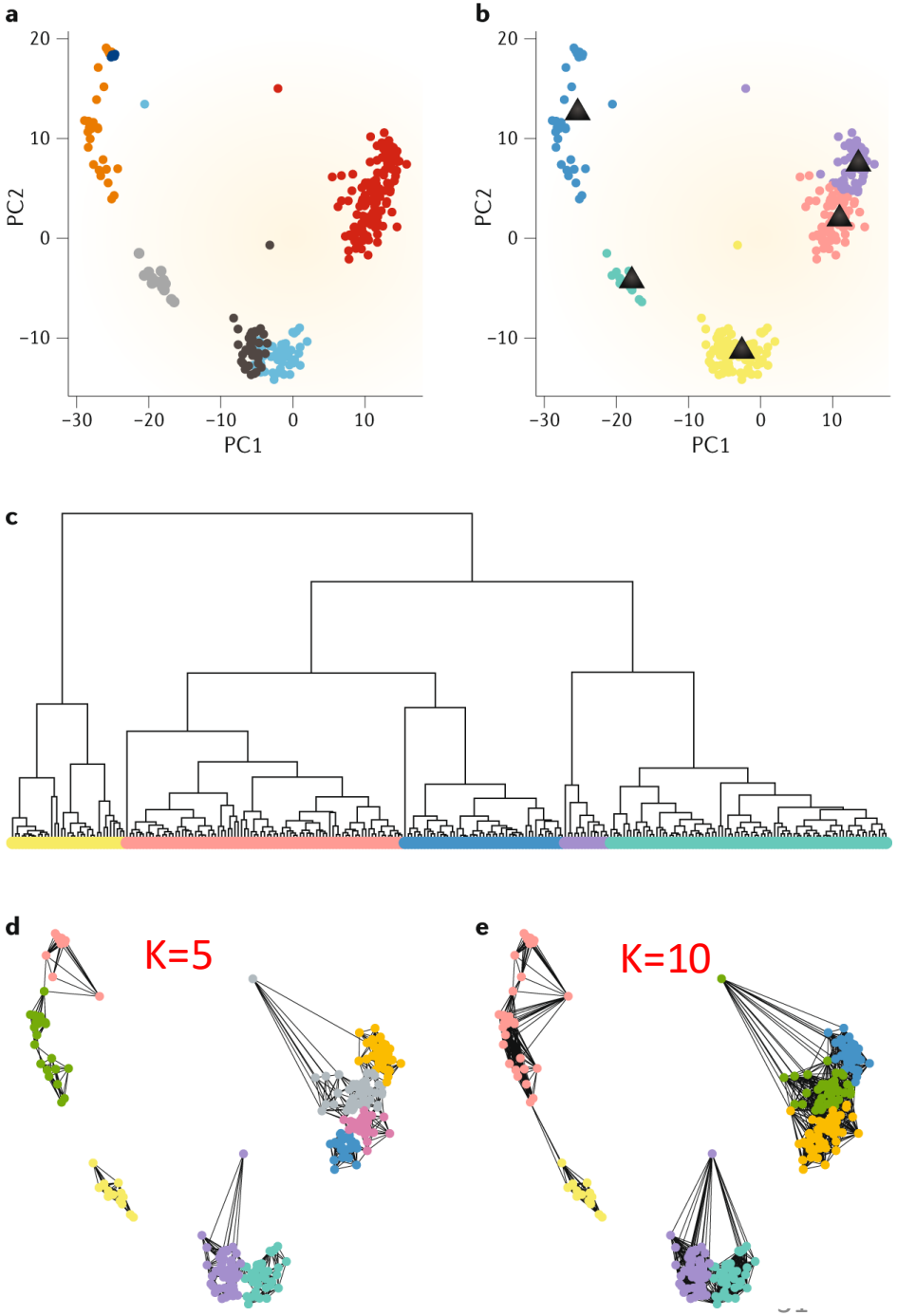
2) Community detection to partition cells in graph into groups of cells
   - Modularity optimization techniques such as the Louvain algorithm
   - Modularity: measures the density of edges inside communities to edges outside communities
   - Louvain iteratively groups cells together, with the goal of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





More about: 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)

K-Means clustering

Unsupervised machine learning algorithm.
Groups data into **K distinct clusters** based on similarity.
Each data point is assigned to the **nearest cluster centroid**.

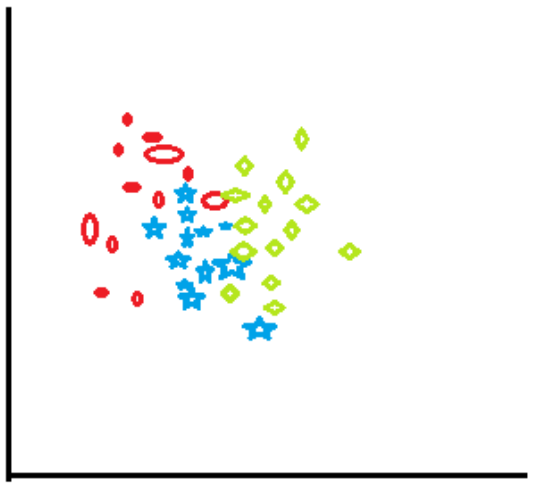


fig 1: before applying k-means clustering

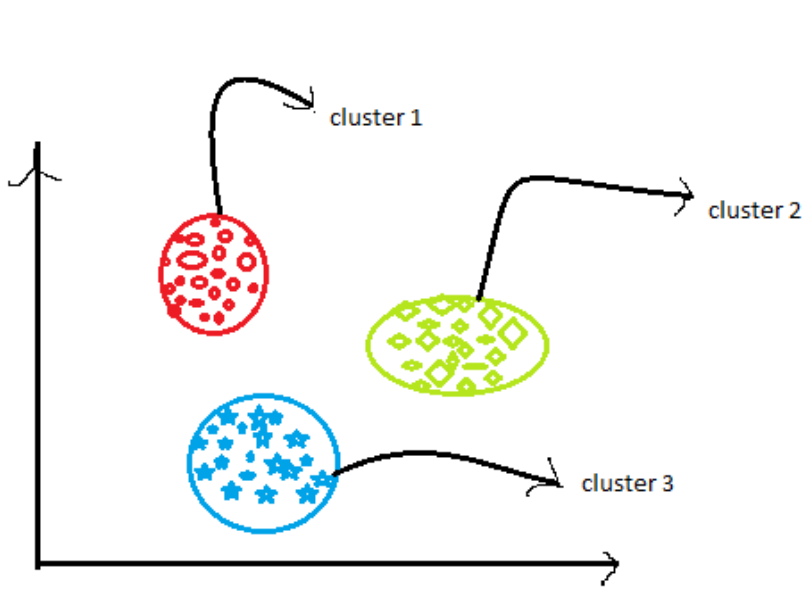
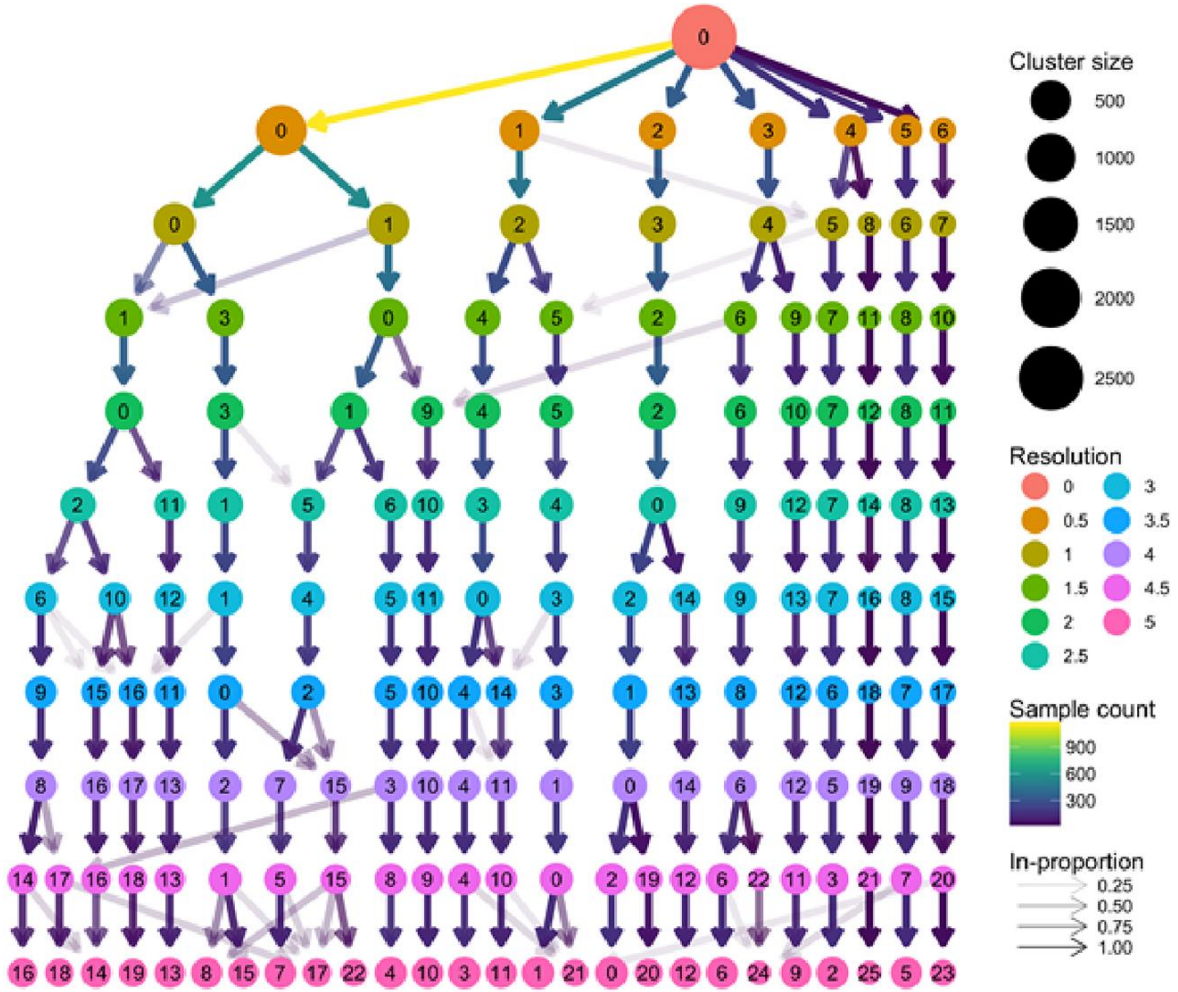


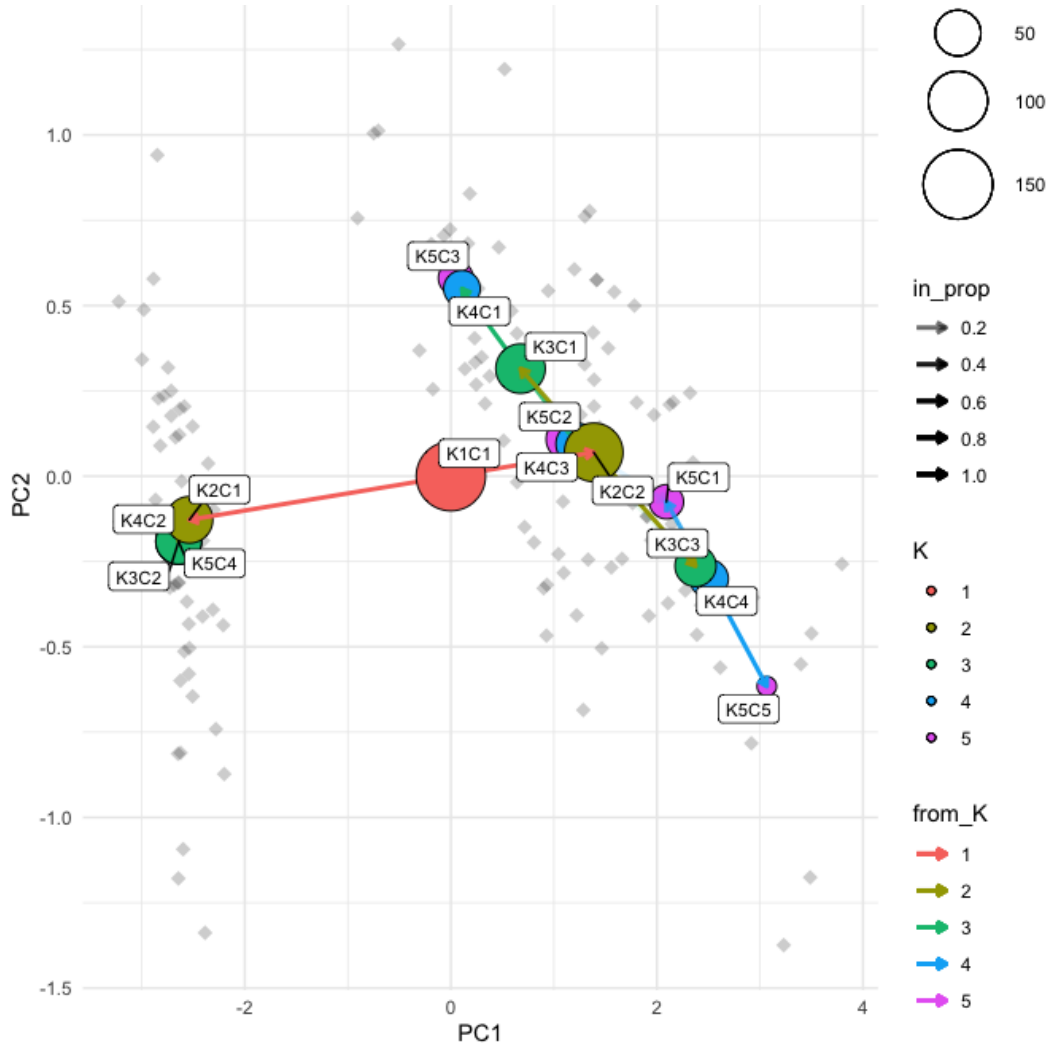
fig 2: After applying K-means clustering

How It Works

- Choose **K (number of clusters)**.
- Randomly initialize **K centroids**.
- Iteratively:
  - Assign each point to the **closest centroid**.
  - Update centroids to the **mean of assigned points**.
- Repeat until centroids **converge (no major change)**.

Visualise clustering results





Cell typing

There are 3 main ways of cell typing:

* Marker Gene Approach : Requires a curated list of genes that are specific to each cell type of interest (manually)

* Database Matching of DEGs : Matches a list of differentially expressed genes (DEGs) for each cluster to a database

* Spot Deconvolution/Cell Label Transfer : Uses a previously annotated single-cell RNA dataset to assign spots/cells labels based on similarities between transcriptomic profiles

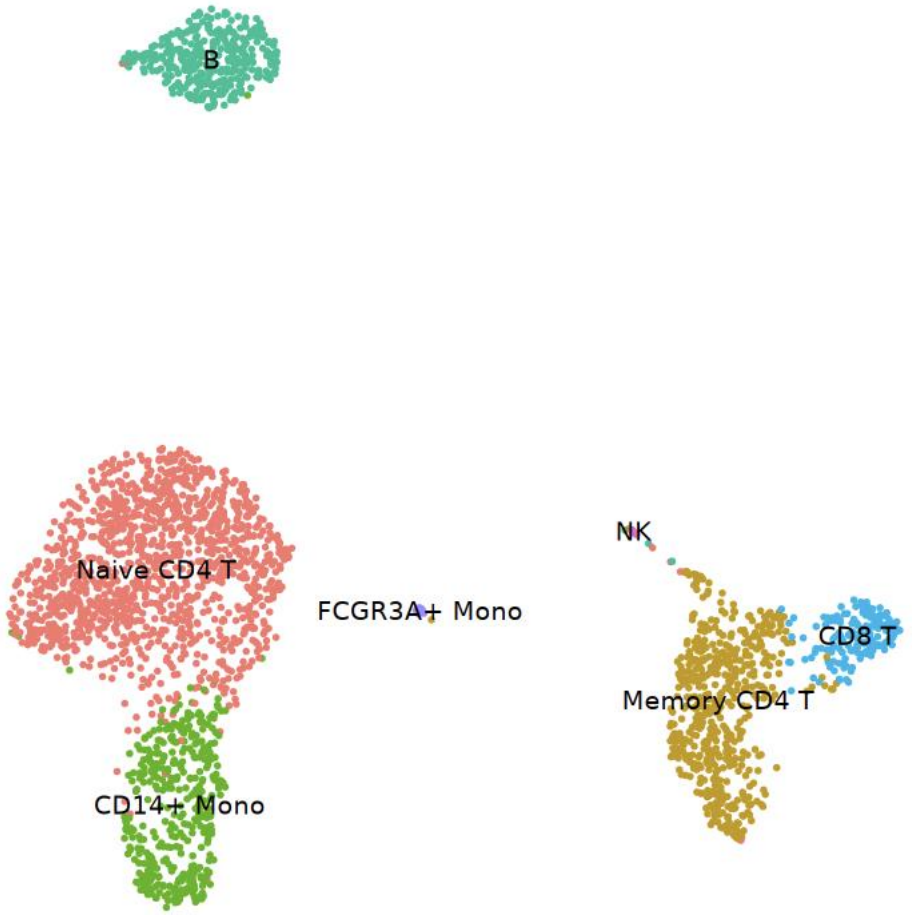
The Marker Gene Approach: Manual curation

The theory

Relies on a manually curated, pre-defined list of genes known to be highly specific to the cell types of interest. The analyst projects the expression of these canonical markers directly onto UMAP dimensional reductions or spatial tissue maps to visually and statistically confirm cluster identities.

Input: Pre-defined target gene panel + Unannotated clustered data.

Output: Visually verified, specific cell populations.



Database Matching: Unbiased DEG Profiling

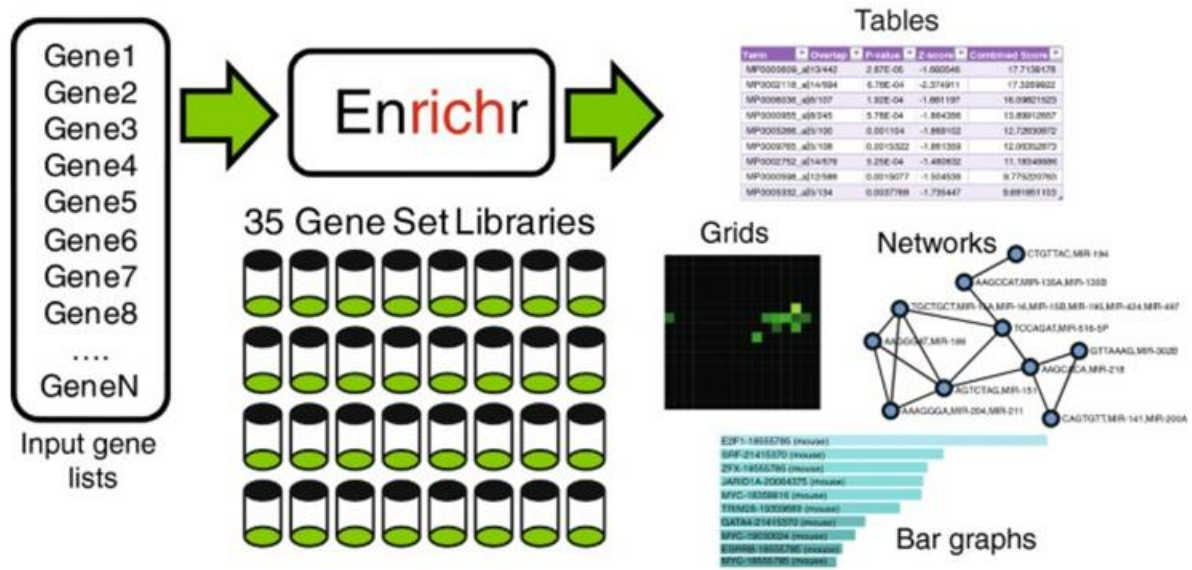
The mechanism

A completely data-driven method that evaluates the entire transcriptome to find statistical signatures.

- Step 1: Extract. Compute Differentially Expressed Genes (DEGs) to isolate the unique transcriptional fingerprint of an unidentified cluster.
- Step 2: Match. Query this empirical DEG list against comprehensive reference databases of established biological markers.

Input: Unidentified clusters + Full transcriptome matrix/ top 100-200 marker list

Output: Automated, probabilistic cell-type annotations.



Chen, E.Y., 2013. BMC Bioinformatics

Practical 2: More About Spatial Data Structure, Visualisation & Cell/Spot Typing

- Prakrithi Pavithra

1. Unique Features about Spatial Data structure
2. Dataset
3. Visualisation

Running the Practical

Terminal

PowerShell

1. Log into your account:

   ssh {username}@203.101.229.126
   username & password from winter school email

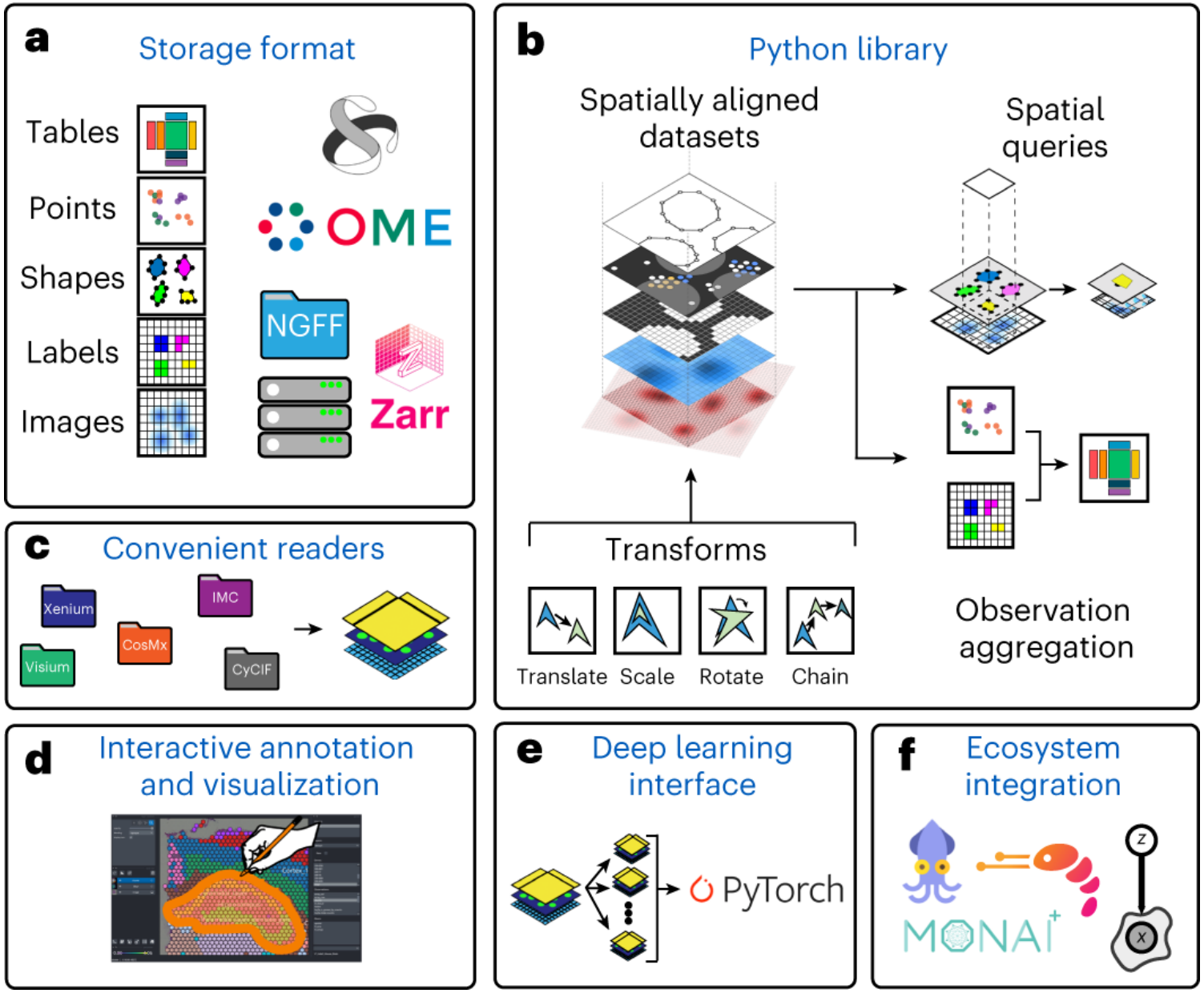
2. Follow these commands:

   - /software/bin/micromamba shell init
   - source ~/.bashrc
   - micromamba activate /software/conda-envs/winter_school_2026
   - git clone https://github.com/GenomicsMachineLearning/gml-teaching-2026
   - cd /scratch/$USER/gml-teaching-2026
   - /scratch/$USER/gml-teaching-2025/runme.sh

3. Open JuperterNotebook:

   001-spatial-single-cell/1.1_python_single_cell_visualisations.ipynb

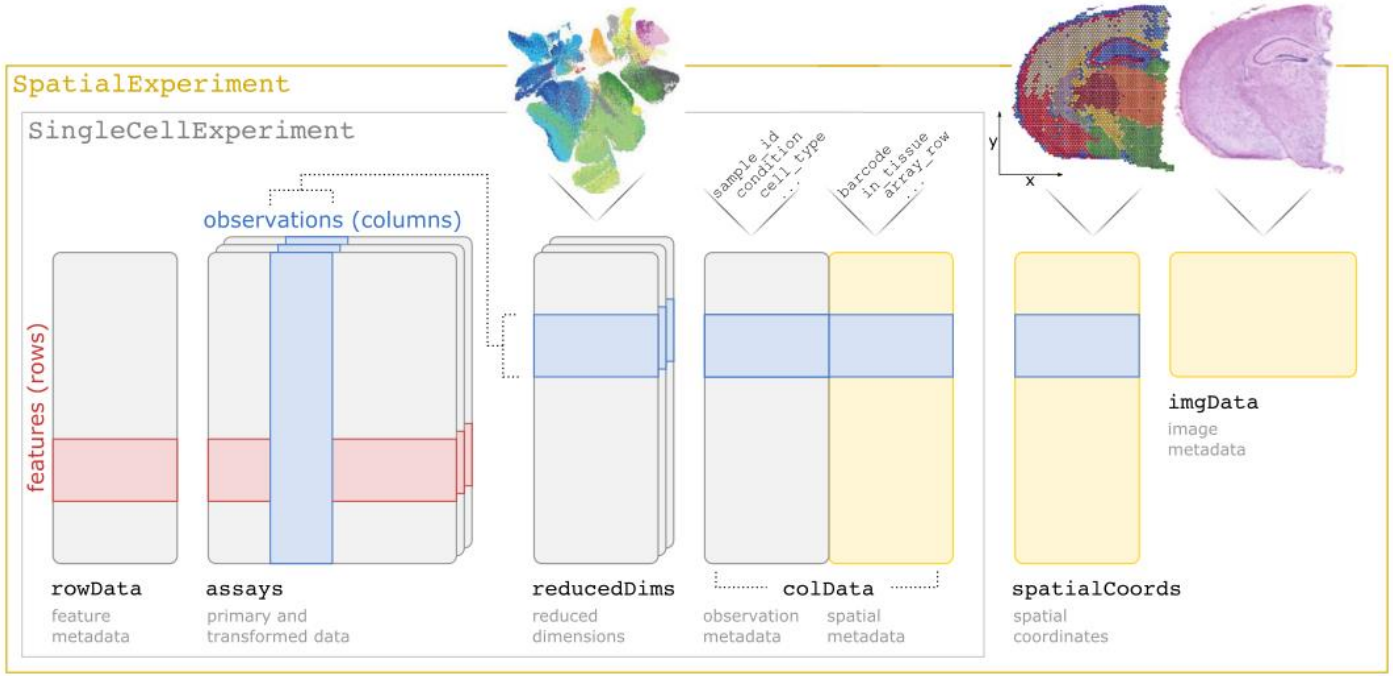
Spatial Analysis Types



(Marcorano et al., 2024)

SpatialExperiment Data Object

- 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.

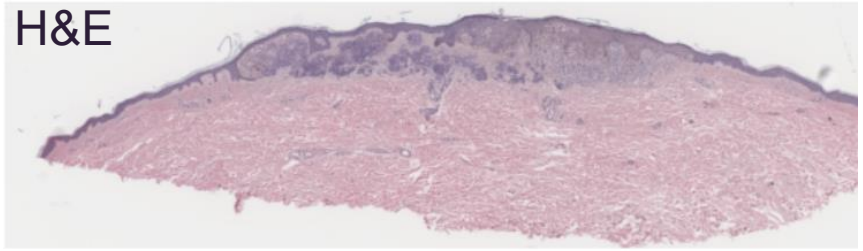


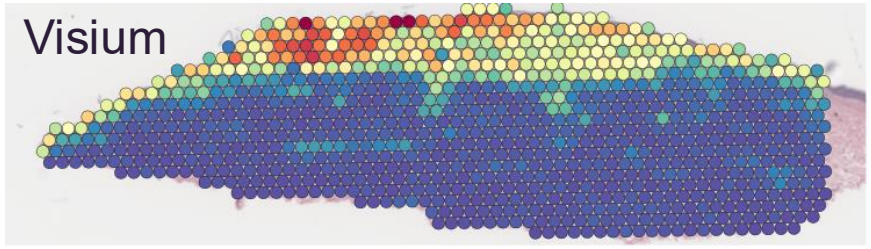
Spatial Single Cell Data

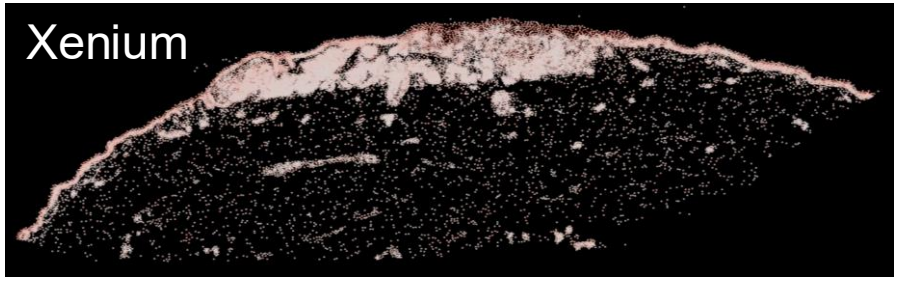
```
SpatialData object with:
├── Images
│     ├── 'HE': SpatialImage[cyx] (3, 4633, 14747)
│     ├── 'morphology_focus': MultiscaleSpatialImage[cyx] (1, 37441, 11479), (1, 18720, 5739), (1, 9360, 2869),
(1, 4680, 1434), (1, 2340, 717)
│     └── 'morphology_mip': MultiscaleSpatialImage[cyx] (1, 37441, 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.

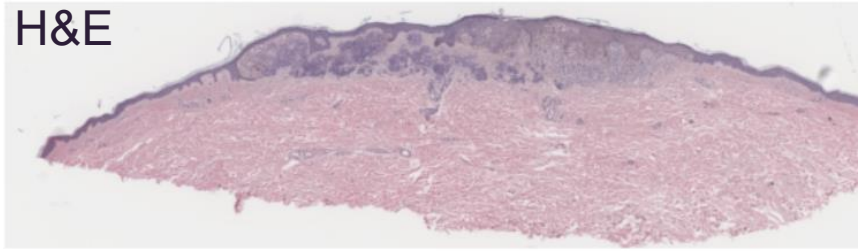
Datasets – Melanoma (Skin)

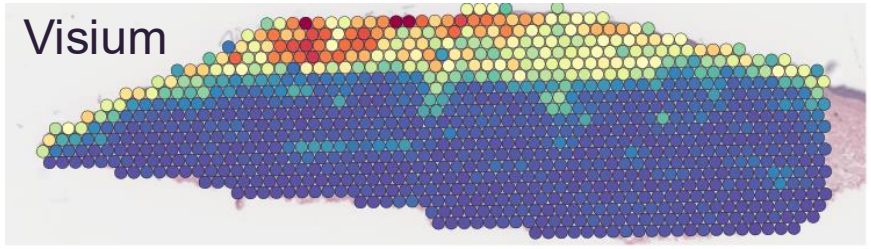


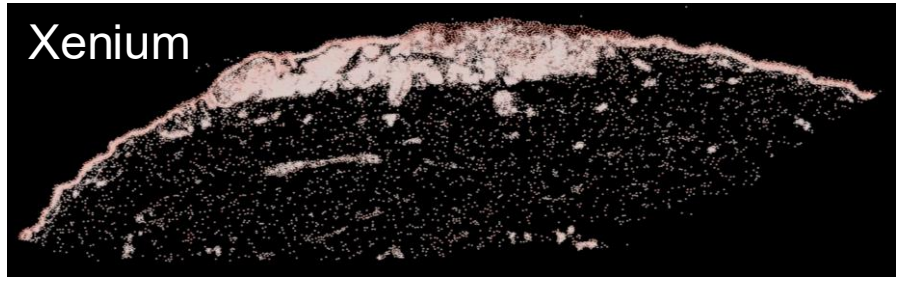




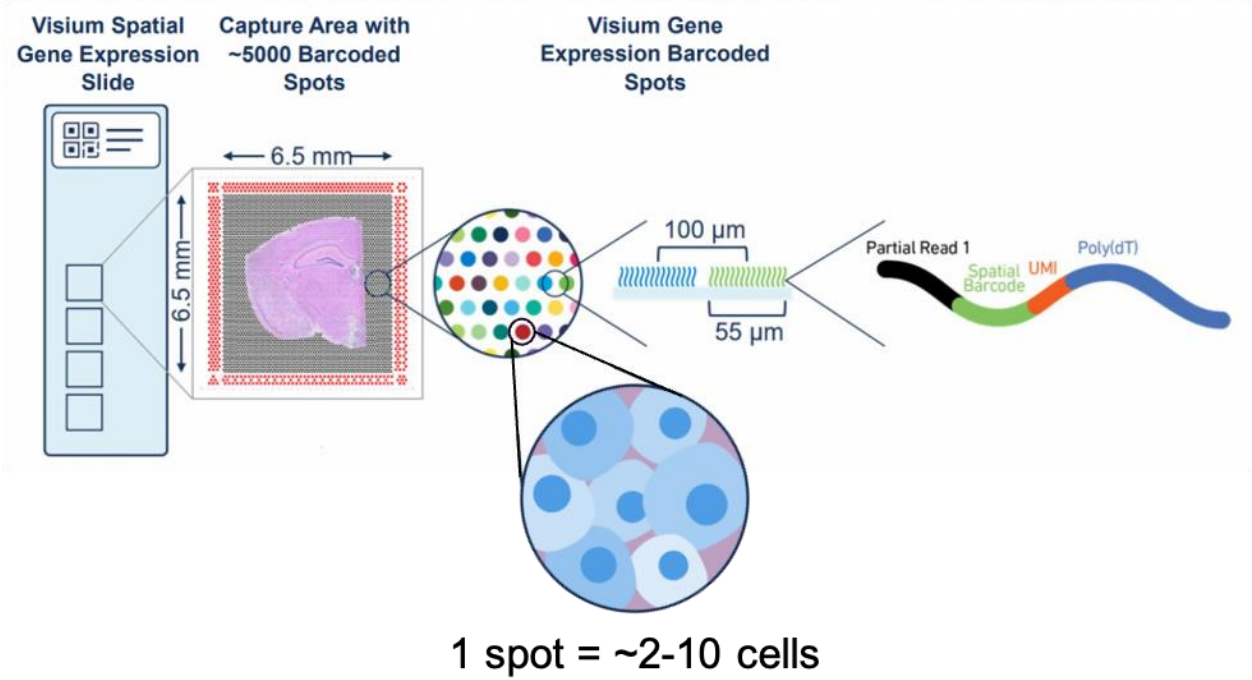
Datasets – Melanoma (Skin)



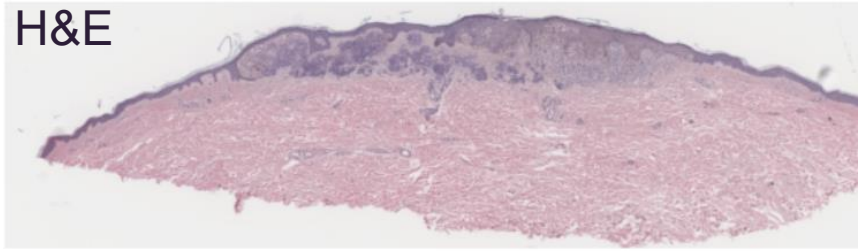


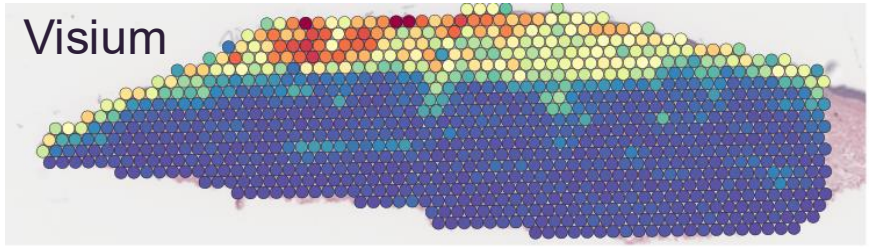


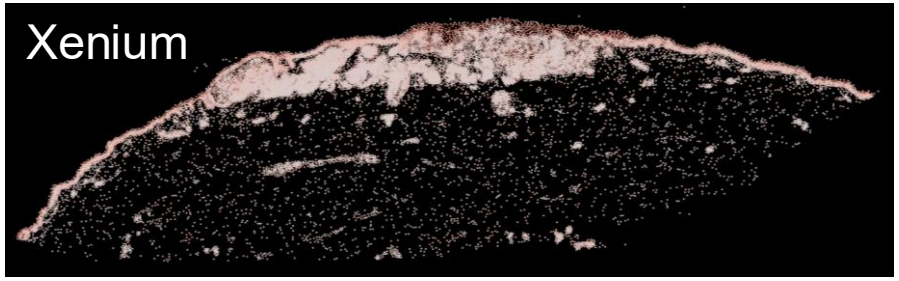
Visium



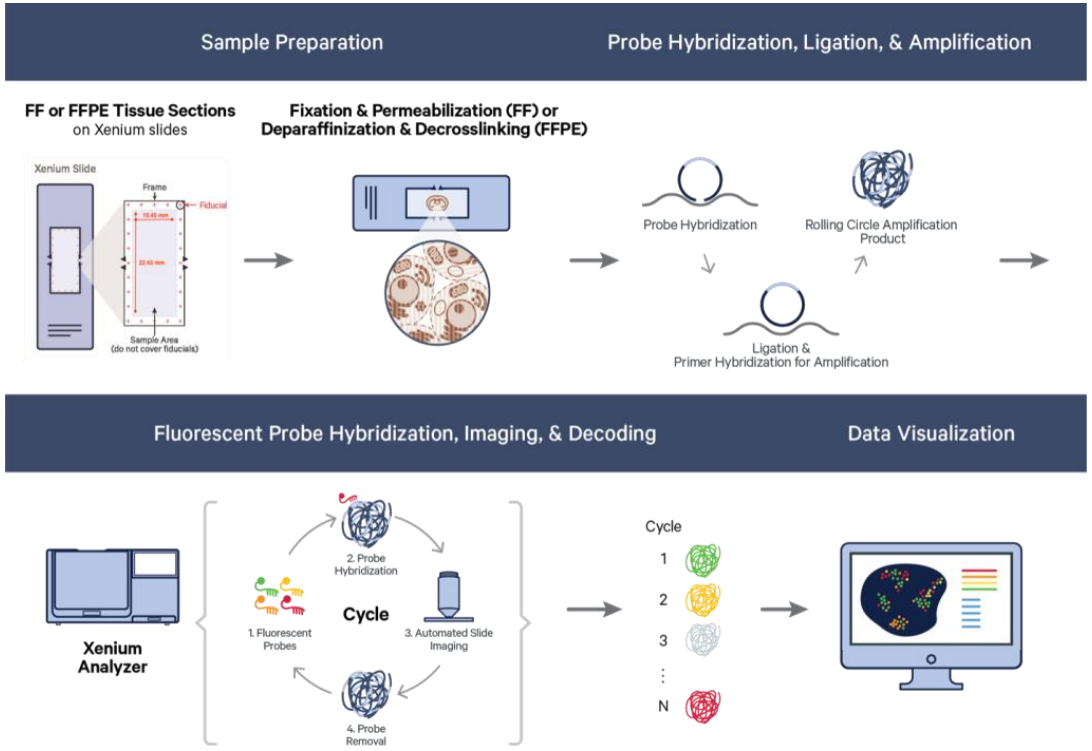
Datasets – Melanoma (Skin)



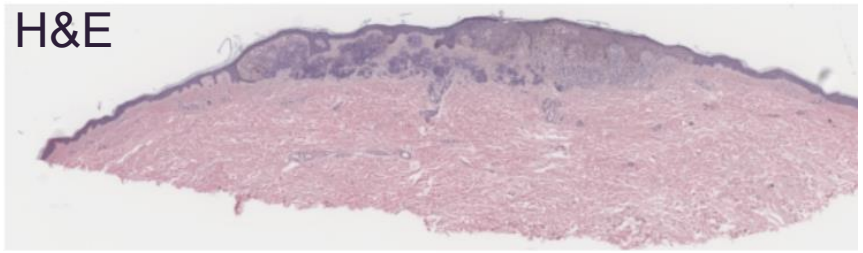


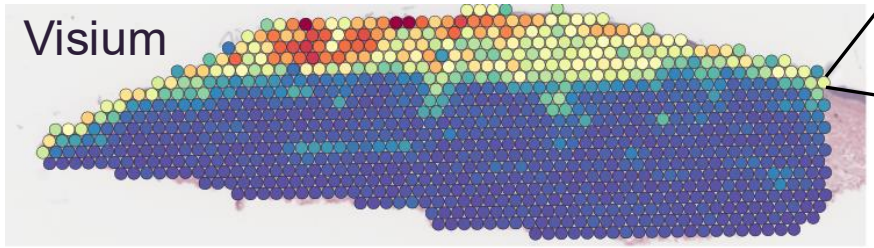


Xenium



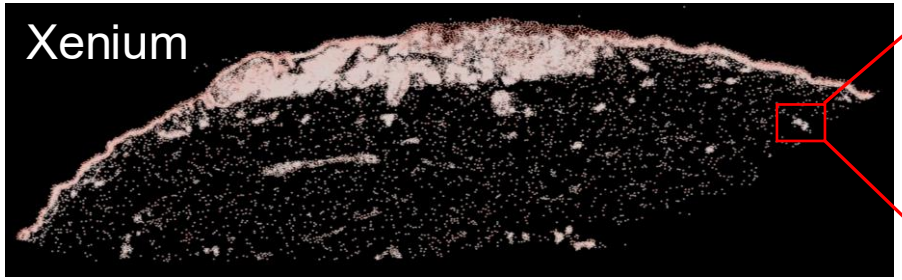
Datasets – Melanoma (Skin)

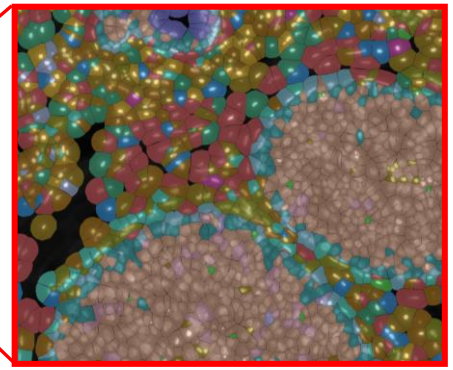




Spot-Based

# Data Points	# Genes
923 spots	18,085



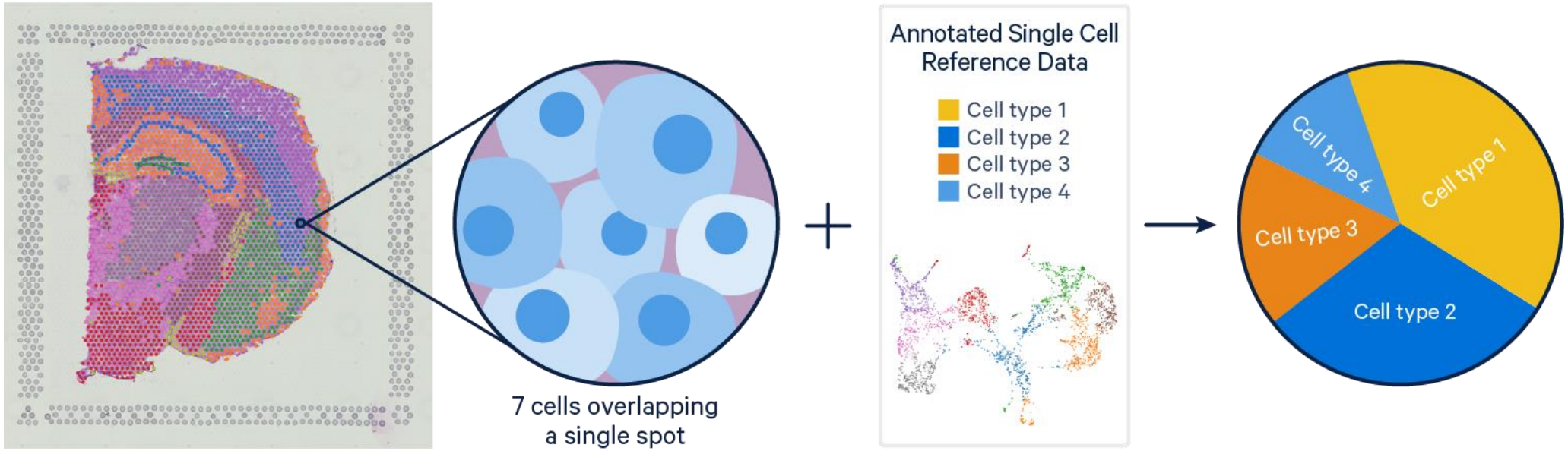


Single Cell

# Data Points	# Genes
21,596 cells	260

3. Spot Deconvolution/Label Transfer [Day 2]

- *Spot Deconvolution*



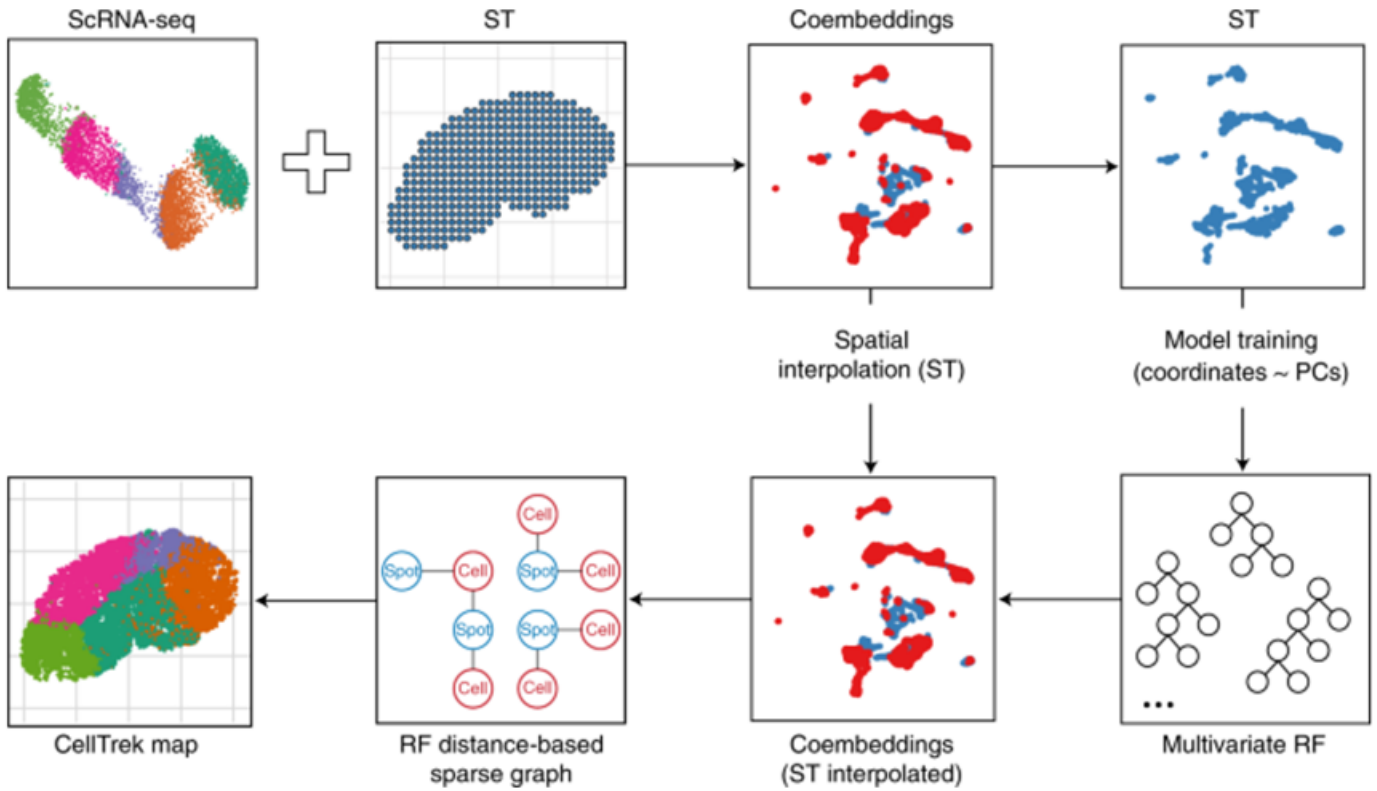
Label Transfer: Mapping Signatures to Space

The mechanism

Utilizes a previously annotated, high-resolution single-cell RNA dataset a s a "ground truth" reference. By anchoring a new, unannotated dataset (the query) to this reference based on underlying transcriptomic similarities, metadata (cell labels) is transferred over.

Input: Annotated scRNA-seq Reference + Unannotated Spatial Query.

Output: Cells/spots with mapped cell identities.



Wei, R., , 2022. *Nat. Biotechnology*

Visium spot deconvolution

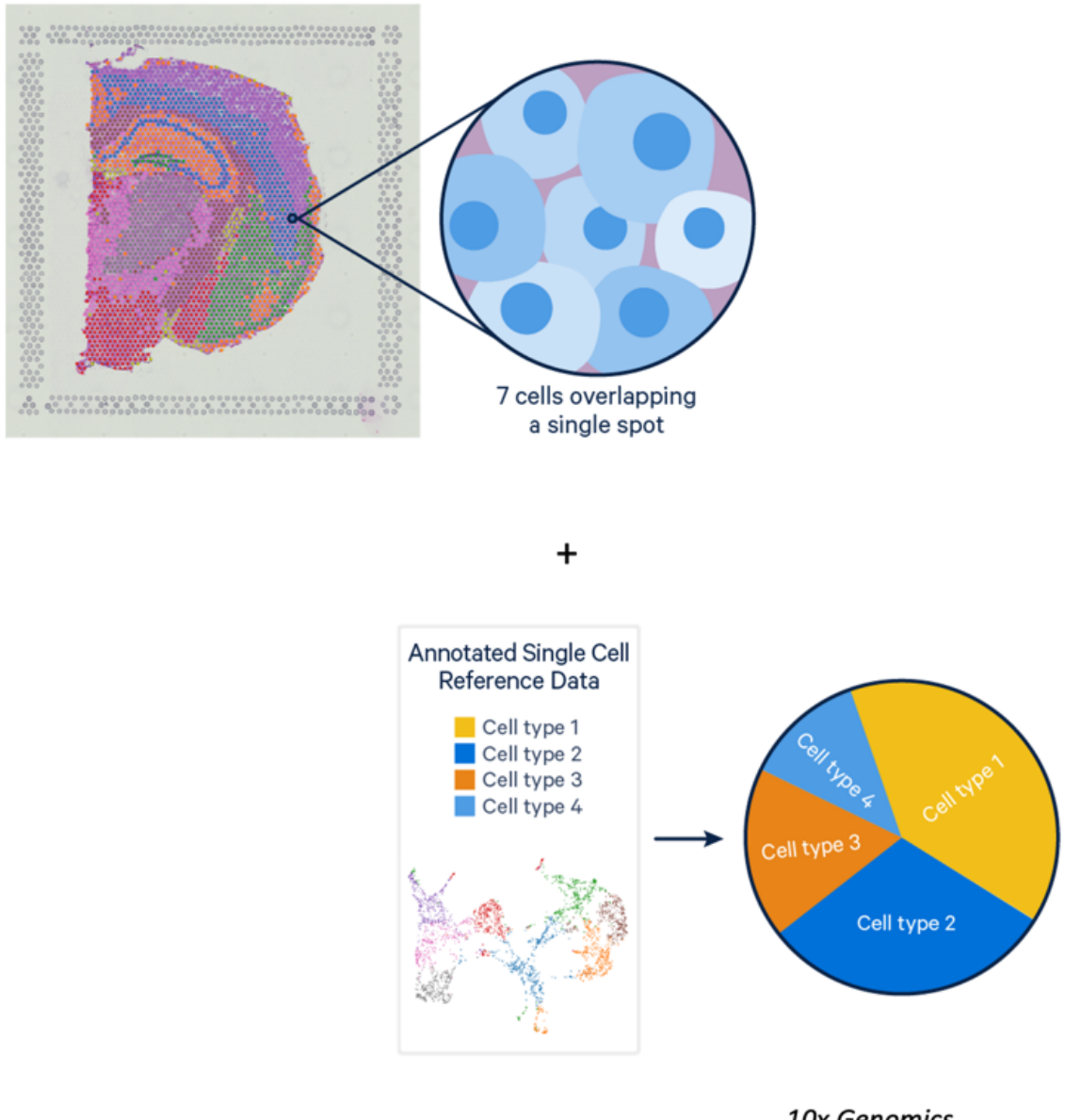
The mechanism

Visium spots: Standard 55-micron Visium spots do not capture single cells; they capture a mixture of approximately 1-9 cells.

Disentanglement: Deconvolution algorithms computationally disentangle these mixed transcriptional signals. By matching the spot's mixed gene expression against the pure signatures from the scRNA-seq reference, the algorithm estimates the exact presence and ratio of underlying cell populations within that single spot.

Input: ST data, scRNA-seq reference

Output: The final deliverable is a high-resolution spatial map detailing the estimated proportions of each cell type across the entire tissue section.



10x Genomics