Spatial Transcriptomics Data Analysis

GitHub: https://github.com/ashoks773/SpatialTranscriptomicsWorkflow

Ashok Kumar Sharma, Ph.D.
Research Bioinformatician II
Digestive and Liver Diseases
Cedars Sinai Medical Center

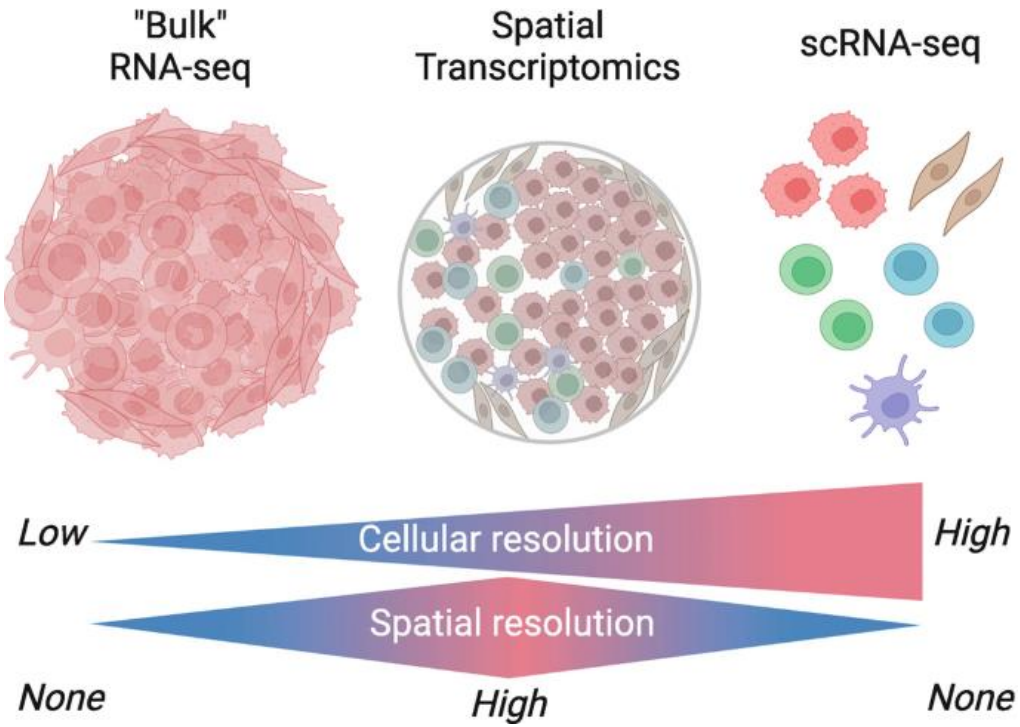
@ashoks773
@ashoks773
@sharma-ak

cedars-sinai.org

Spatial Transcriptomics Analysis Workflow – Outline

- **Introduction to Spatial Transcriptomics**
- Select and Download the Dataset
- Load and Create Single-Cell Object
- Normalization, PCA, UMAP, Clustering, and Visualization
- Marker Gene Identification
- Cell Type Annotations Using Different Methods

Transcriptomic Approaches: Bulk RNA-seq, scRNA-seq, and Spatial Transcriptomics

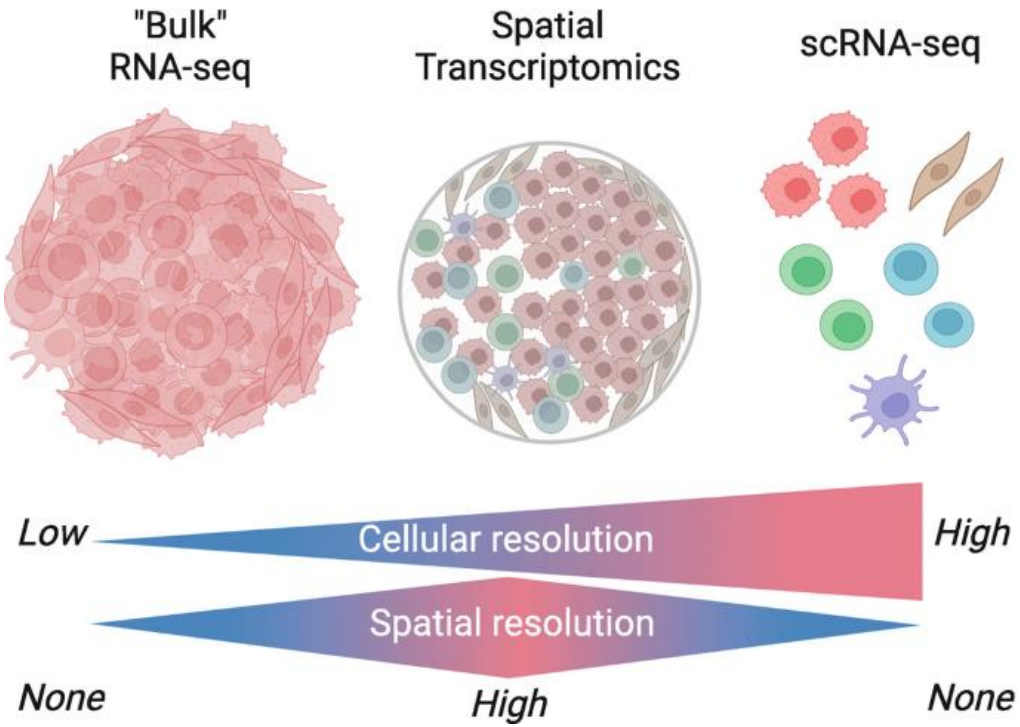


Overview of Techniques

- Measure av. Gene expression across cells
- Gene expression at Cell level
- Capture spatial context – helpful to link molecular data to tissue architecture

Ospina et. Al; Methods in Molecular Biology, 2023

Transcriptomic Approaches: Bulk RNA-seq, scRNA-seq, and Spatial Transcriptomics

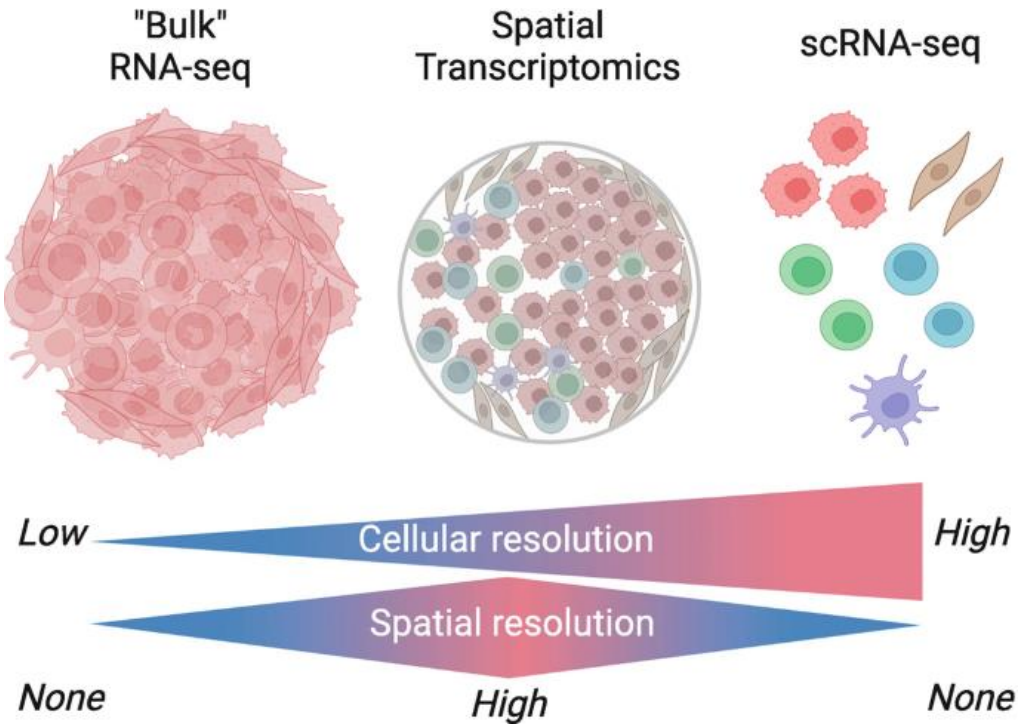


Cellular Resolution

- Aggregate signals from diverse cell types
- High resolution- identify distinct cell types and states
- Moderate resolution – depending upon the technique – preserve tissue organization

Ospina et. Al; Methods in Molecular Biology, 2023

Transcriptomic Approaches: Bulk RNA-seq, scRNA-seq, and Spatial Transcriptomics

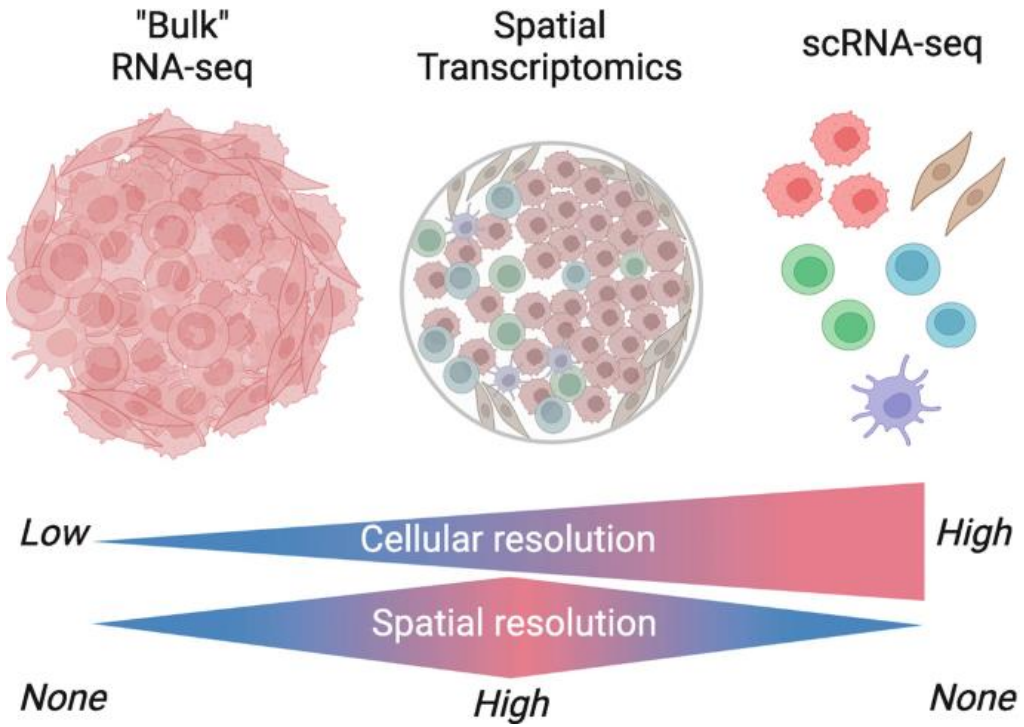


Spatial Resolution

- No spatial resolution
- Limited spatial resolution
- Directly links gene expression to spatial locations

Ospina et. Al; Methods in Molecular Biology, 2023

Transcriptomic Approaches: Bulk RNA-seq, scRNA-seq, and Spatial Transcriptomics

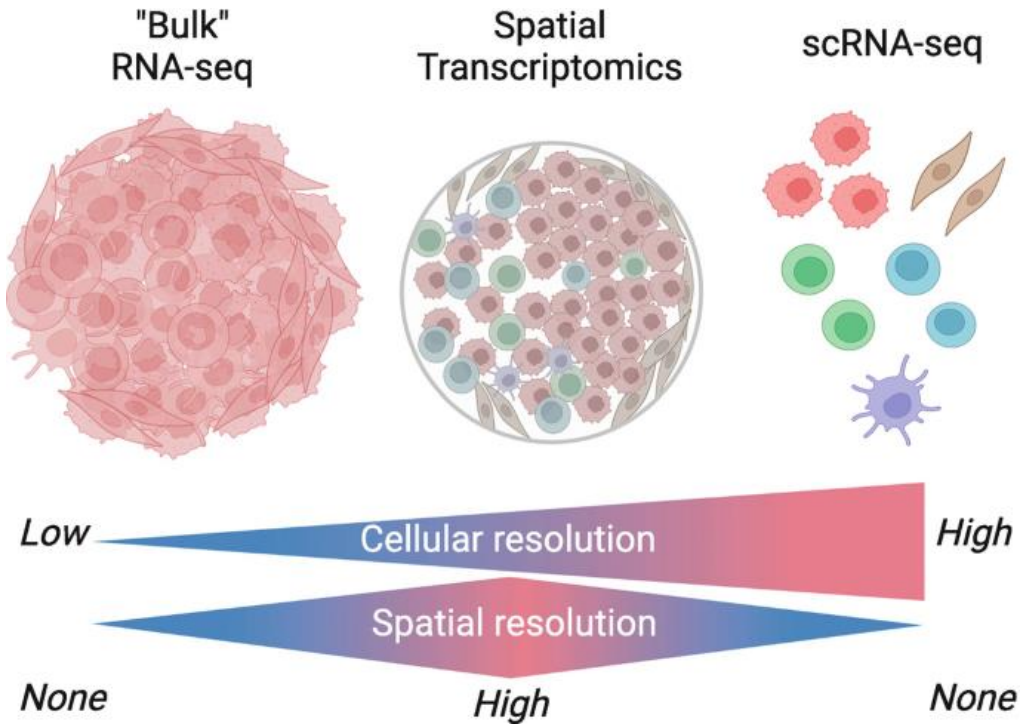


Applications

- Suitable for large-scale exploratory studies like disease profiling
- Useful for exploring the cellular diversity in the tissue (e.g., tumor, immune responses)
- Ideal for studying tissue architecture and cellular interactions in context (e.g., development, pathology)

Ospina et. Al; Methods in Molecular Biology, 2023

Transcriptomic Approaches: Bulk RNA-seq, scRNA-seq, and Spatial Transcriptomics



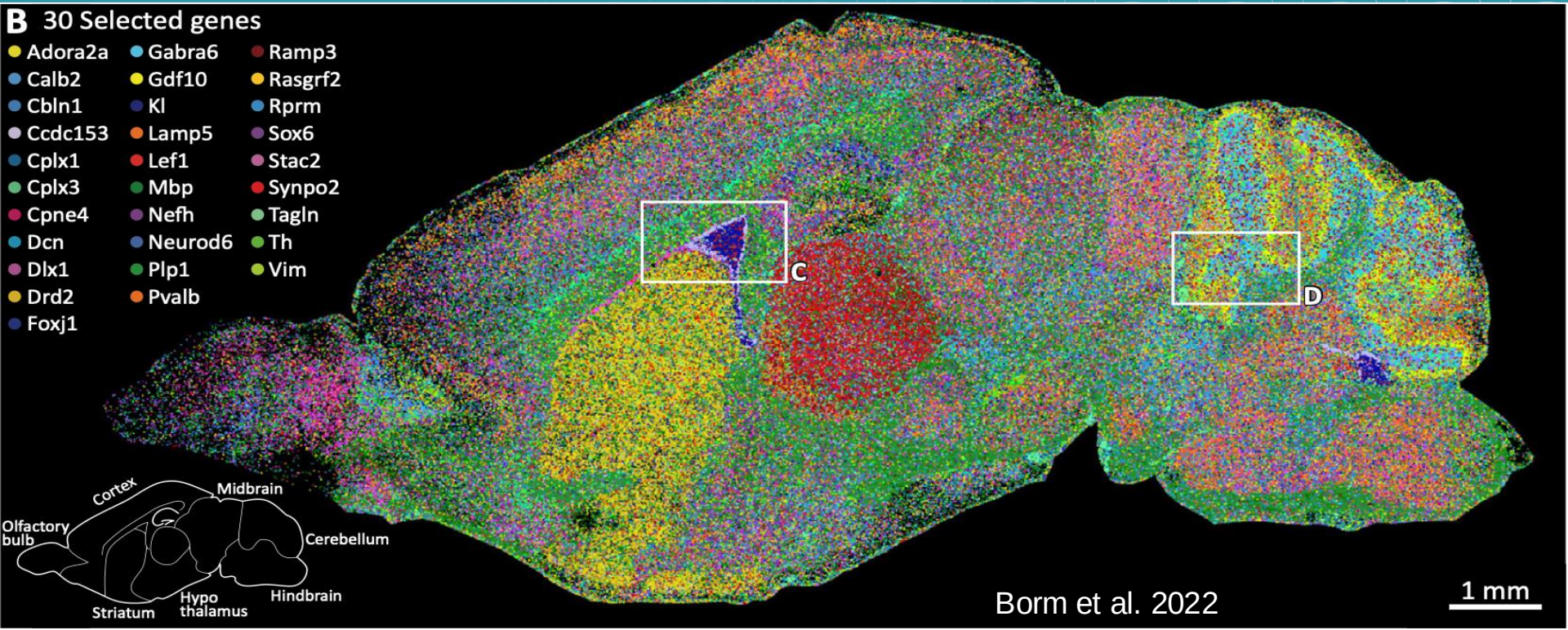
Limitations

- Cannot distinguish cell-type specific effects
- May miss spatial interactions without further context
- Potentially lower sensitivity for lowly expressed genes compared to Bulk methods

Ospina et. Al; Methods in Molecular Biology, 2023

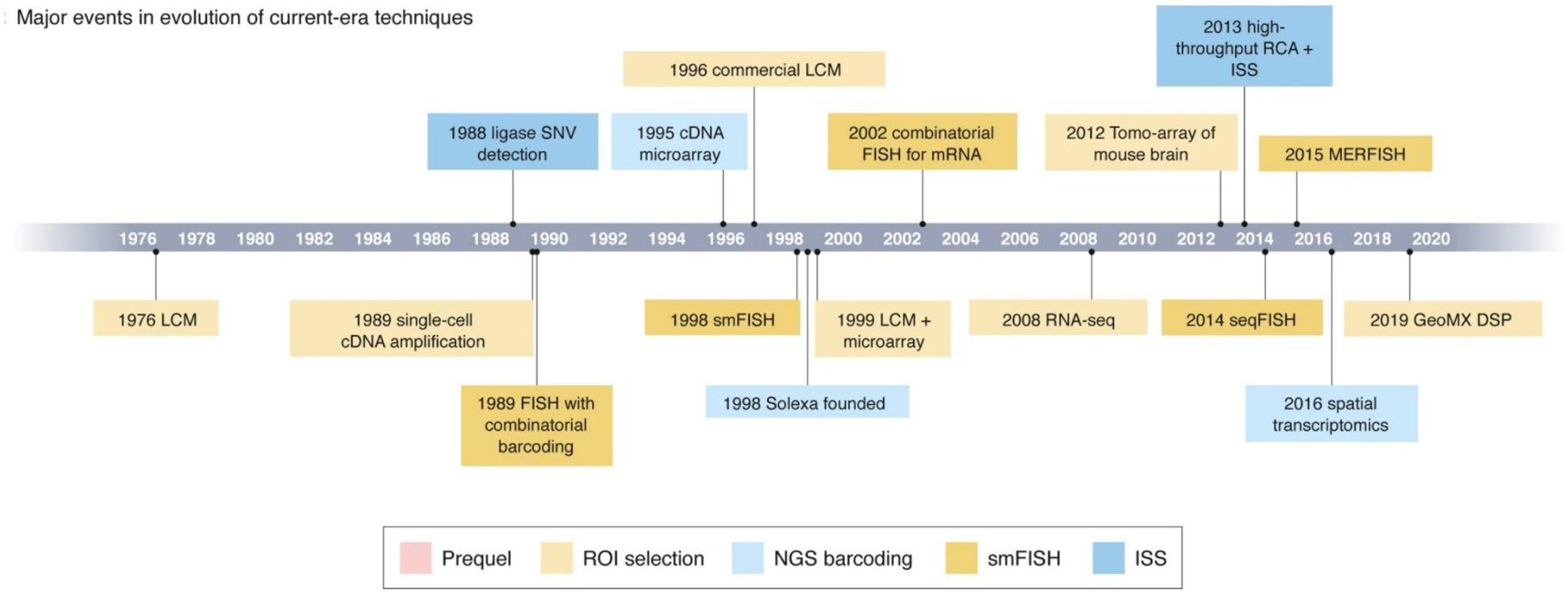
What is spatial Transcriptomics

Technologies that make *transcripts (RNA)* seen, while preserving their spatial information in the cell or tissue, with high-throughput (many cells and many genes)



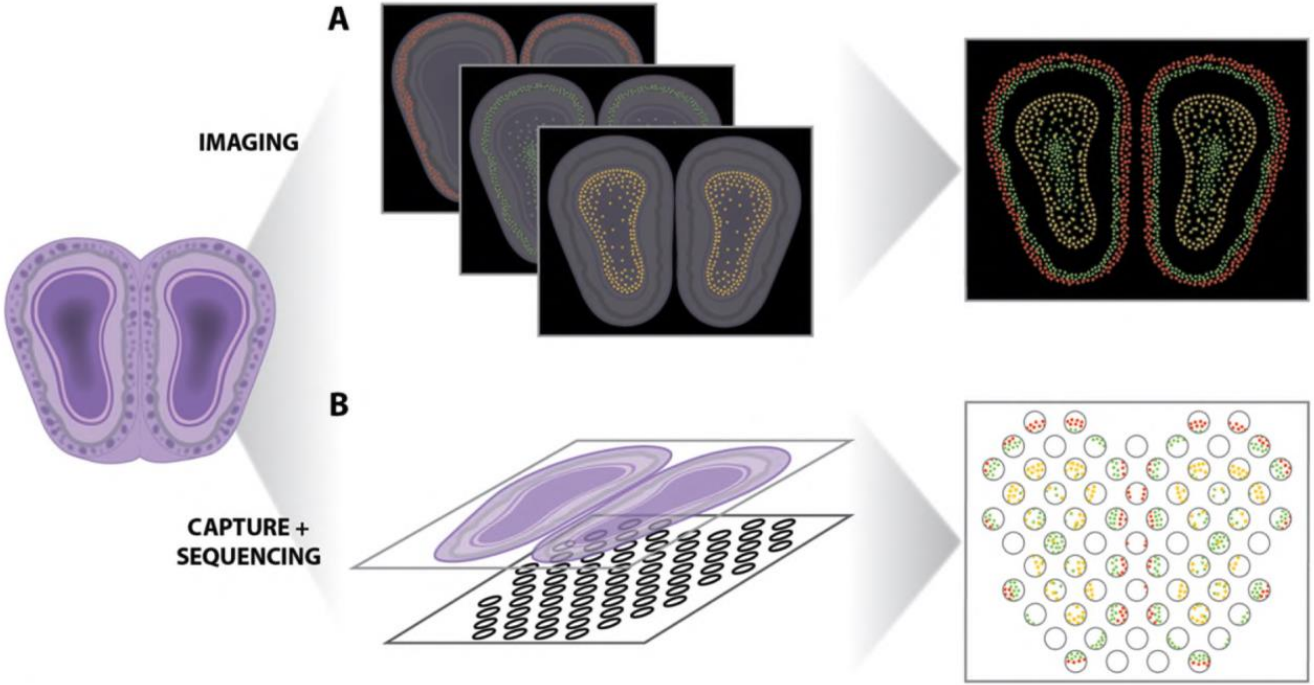
Borm et al. 2022

Major events in evolution of current-era techniques



Moses et. Al; Nature Methods, 2022

Two major Technologies – for Spatial transcriptomics



Imaging-based

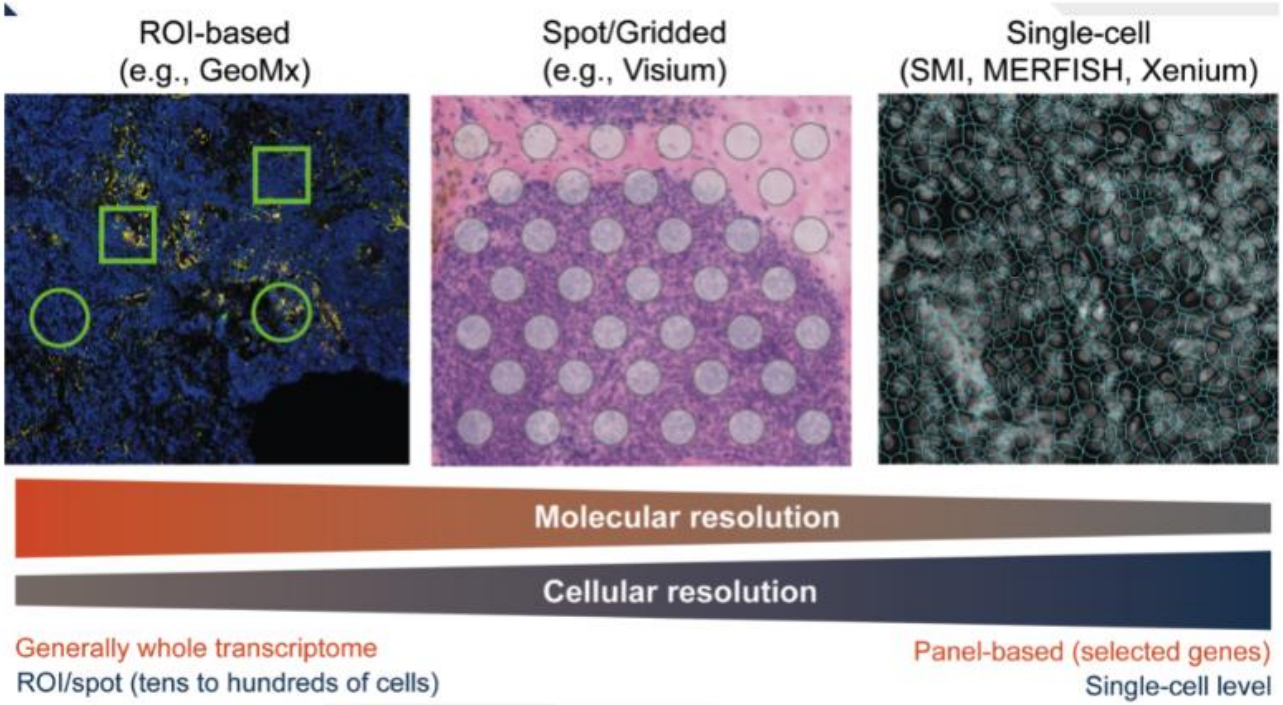
- MERFISH
- seqFISH
- STARmap

Sequencing-based

- Slide-seq
- 10X Visium

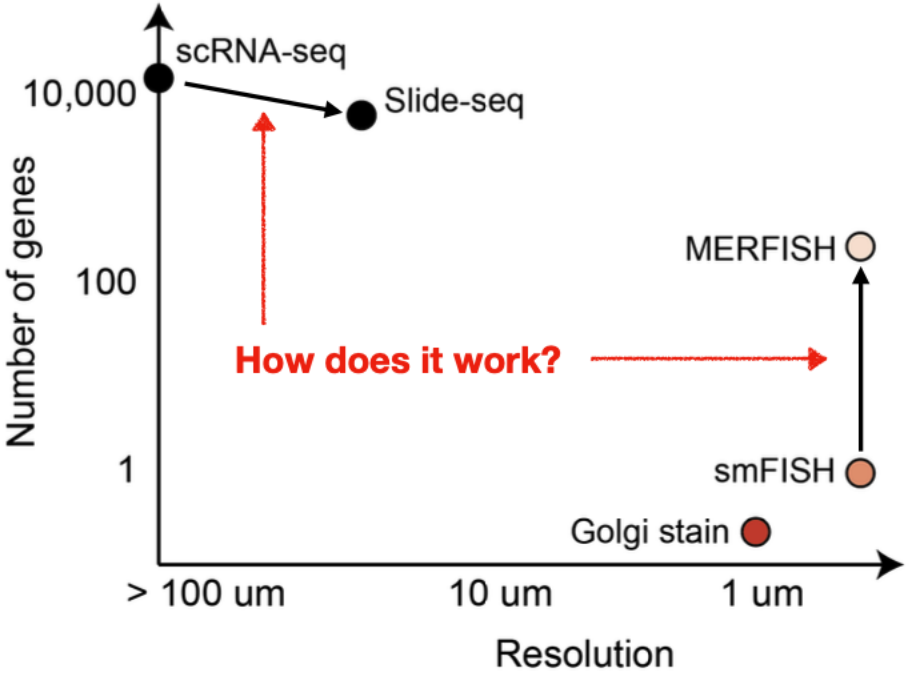
Atta and Fan, 2021, Nat Comm

Overview of a spatial transcriptomics workflow



https://hutchdatascience.org/Choosing_Genomics_Tools/spatial-transcriptomics-1.html

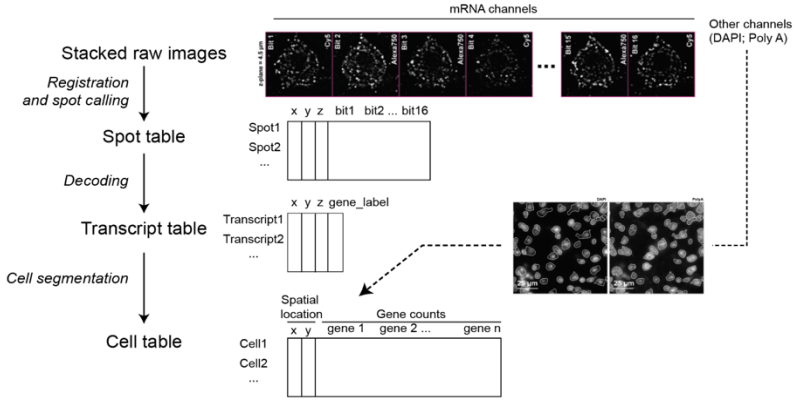
Sequencing-based vs Imaging-based assays



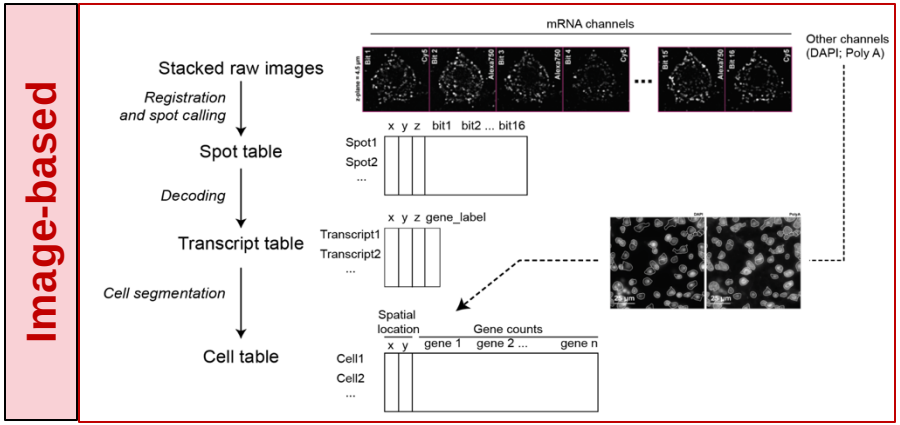
- **Multiplexing**
- **Resolution**
- **Throughput**
- **Sensitivity**

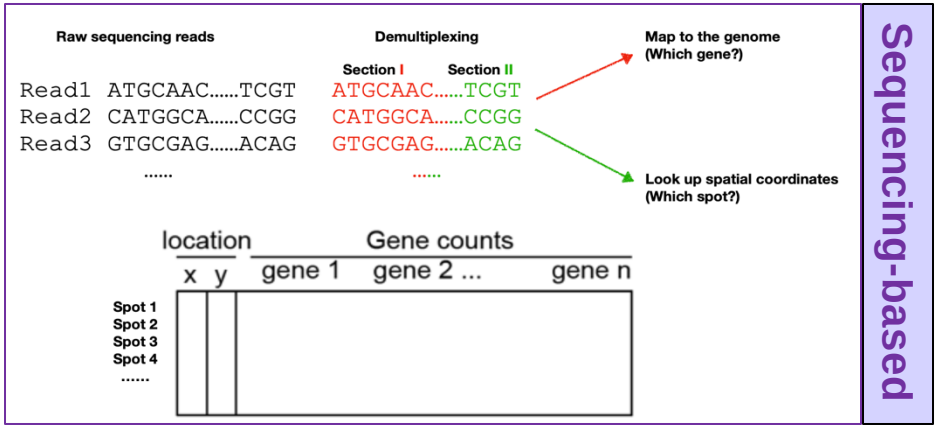
Data processing workflow
Image-based & Sequencing-based

Image-based



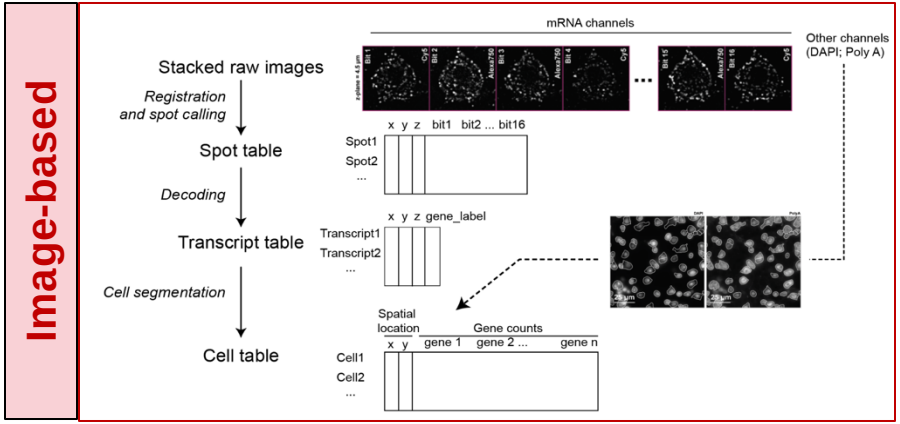
Data processing workflow
Image-based & Sequencing-based

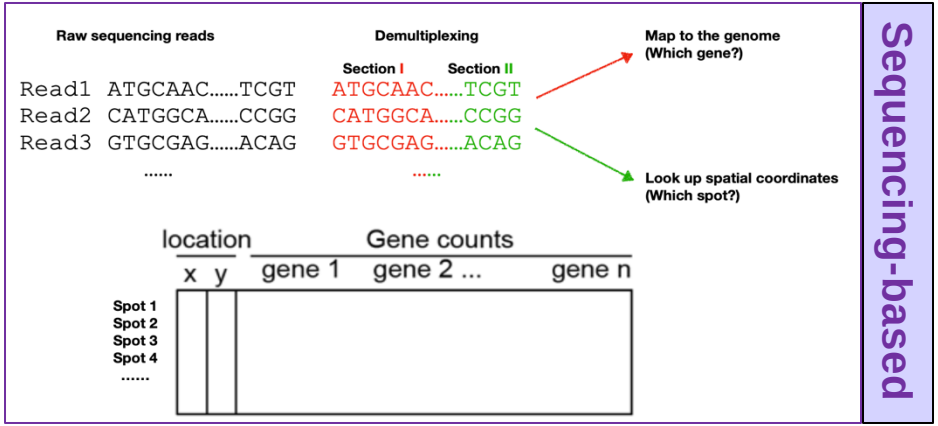




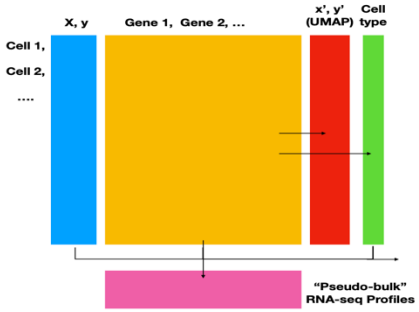
Ref: UCLA Collaboratory workshop (W31)

Data processing workflow
Image-based & Sequencing-based





Analysis overview



- [x, y, one gene at a time] — spatial distribution of gene expression
- [All genes] — dimensionality reduction (PCA, UMAP)
- [All genes] — clustering/cell typing (k-means, hierarchical, Leiden)
- [x, y, cell type] — spatial proximity, interaction, and spatial enrichment of cell types.

Ref: UCLA Collaboratory workshop (W31)

Spatial Transcriptomics Analysis Workflow – Outline

- Introduction to Spatial Transcriptomics
- **Select and Download the Dataset**
- Load and Create Single-Cell Object
- Normalization, PCA, UMAP, Clustering, and Visualization
- Marker Gene Identification
- Cell Type Annotations Using Different Methods

Recourses: Datasets and tutorial links to be followed during this course

Spatial transcriptomics analysis of neoadjuvant cabozantinib and nivolumab in advanced hepatocellular carcinoma identifies independent mechanisms of resistance and recurrence

Dataset

- Spatial transcriptomics with Visium (10x Genomics)
- **Samples: 7** HCC-patients (4 Responder and 3 NonResponders)
- **Raw data: "GSE238264"**
  - https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=**GSE238264**

Learnings

- **Step1:** Load and Create single Cell Object
- **Step2:** Normalization PCA; UMAP; Clustering; Visualization
- **Step3:** Marker gene identification for each cluster; sample or response group
- **Step4:** Cell Type Annotations using CellTypist

Requirements – Python libraries

Data processing; normalization

numpy
pandas
matplotlib
seaborn
sklearn
umap-learn
umap
scanpy

For Cell Type Annotation

Seurat
celldex
SingleCellExperiment
celltypist
azimuth
scType
singleR

To use R based annotations methods in Python

rpy2 (to use R consol directly in Python)

Spatial Transcriptomics Analysis Workflow – Outline

- Introduction to Spatial Transcriptomics
- Select and Download the Dataset
- **Load and Create Single-Cell Object**
- Normalization, PCA, UMAP, Clustering, and Visualization
- Marker Gene Identification
- Cell Type Annotations Using Different Methods

Step1: Overview of the AnnData Object - hepatocellular carcinoma (HCC) resection specimens (n=7)

AnnData object with
n_obs × n_vars =
17292 × 36601

- **obs**: 'in_tissue', 'array_row', 'array_col', 'sample', 'center_x', 'center_y'
- **var**: 'gene_ids', 'feature_types'
- **obsm**: 'spatial'

Total observations in 7 samples

obs: Observation Metadata (Rows/Cells or Spots)

- **obs** stores information related to each observation (cell or spatial location). This can include sample-specific details, spatial coordinates, and metadata labels for grouping or comparisons

var: Variable Metadata (Columns/Genes or Features)

- **var** contains metadata for each variable (usually genes). This includes details about each gene or feature being measured.

Obsm: Multi-dimensional Observation Metadata (Spatial Coordinates)

- **obsm** holds multi-dimensional data, often matrices, related to observations (cells or spots). It's especially useful for embeddings or spatial coordinates.

uns: Unstructured Metadata (Annotations or Color Maps)

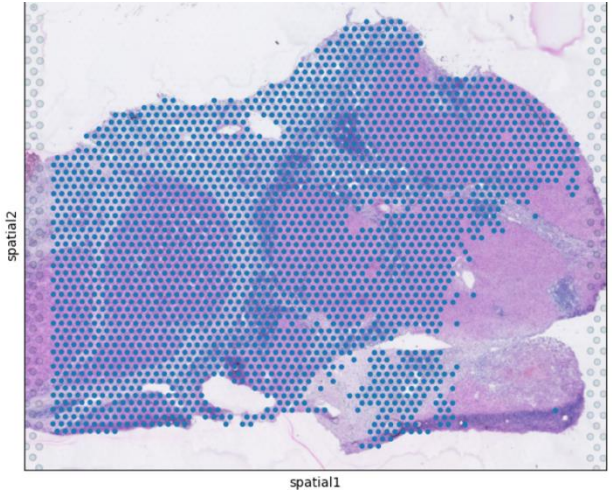
- **uns** is used for unstructured data, like color maps, analysis parameters, or general annotations that don't fit neatly into rows or columns.

QC metrics

First check – weather the **gene is mitochondrial** often identified by the names starting with MT- or mt-

- Proportion of RNA counts derived from mitochondrial genes in each cell
  - **Low %:** Indicates healthy cells with minimal stress or apoptosis, as mitochondria typically contribute only a small proportion of the total RNA in viable cells
  - **High %:** May indicate cellular stress, apoptosis, or poor-quality data, as dying cells often show increased mitochondrial RNA relative to the total RNA content

HCC1R



Other Checks can be:

- Total counts
- # of expressed genes
- % contribution by top genes

QC metric calculation: sc.pp.calculate_qc_metrics

Total counts	Reads or UMIs) detected in given cell/**spot** across all genes	Overall RNA content captured for a cell/**spot**	↓ May indicate low RNA content or poor capture efficiency ↑ May also suggest doublets or sequencing artifacts
N genes by counts	# of Genes with non-zero expression in given cell/**spot**	How many genes are actively expression in a cell/**spot**	↓ May indicate dead or dying cells or empty **spots** in spatial data ↑ Could indicate doublets (two cells/**spots** captured together)
PCT counts in top 100 genes	How much of top **100 genes** contribute to the total counts	Weather a few genes dominate the transcriptome of a cell/**spot**	↓ Suggest a balanced gene expression profiles across cells/**spots** ↑ May suggest biases or the presence of dominant genes; potential artifacts

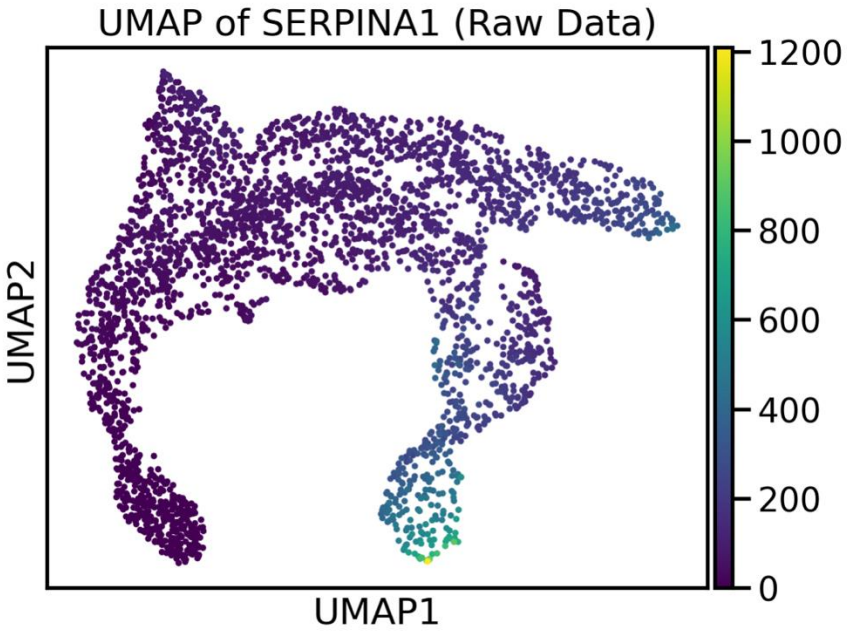
$$\text{pct\_counts\_in\_top\_100\_genes} = \left(\frac{\sum_{i=1}^{100} \text{counts}_i}{\text{total\_counts}} \right) \times 100$$

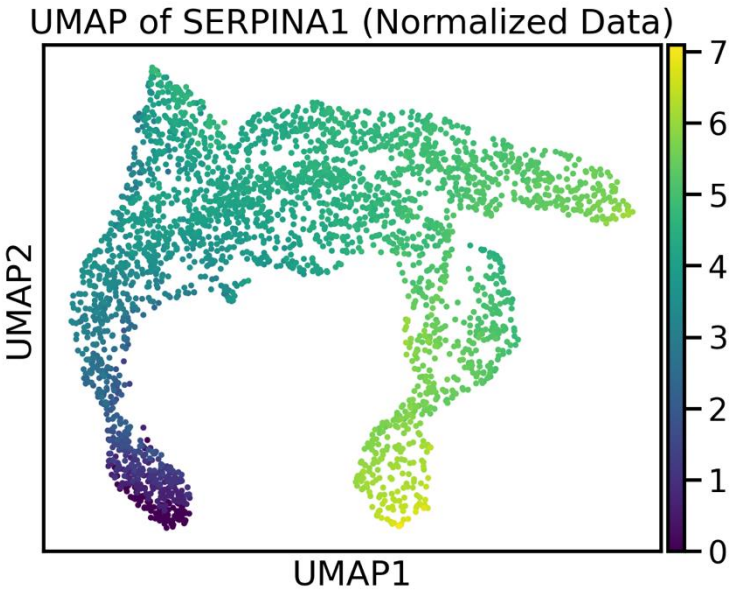
Spatial Transcriptomics Analysis Workflow – Outline

- Introduction to Spatial Transcriptomics
- Select and Download the Dataset
- Load and Create Single-Cell Object
- **Normalization, PCA, UMAP, Clustering, and Visualization**
- Marker Gene Identification
- Cell Type Annotations Using Different Methods

Why normalization is Important – hepatocellular carcinoma (HCC) resection specimen HCC1R – UMAP

Apply log1p normalization: sc.pp.log1p(ad_viz)

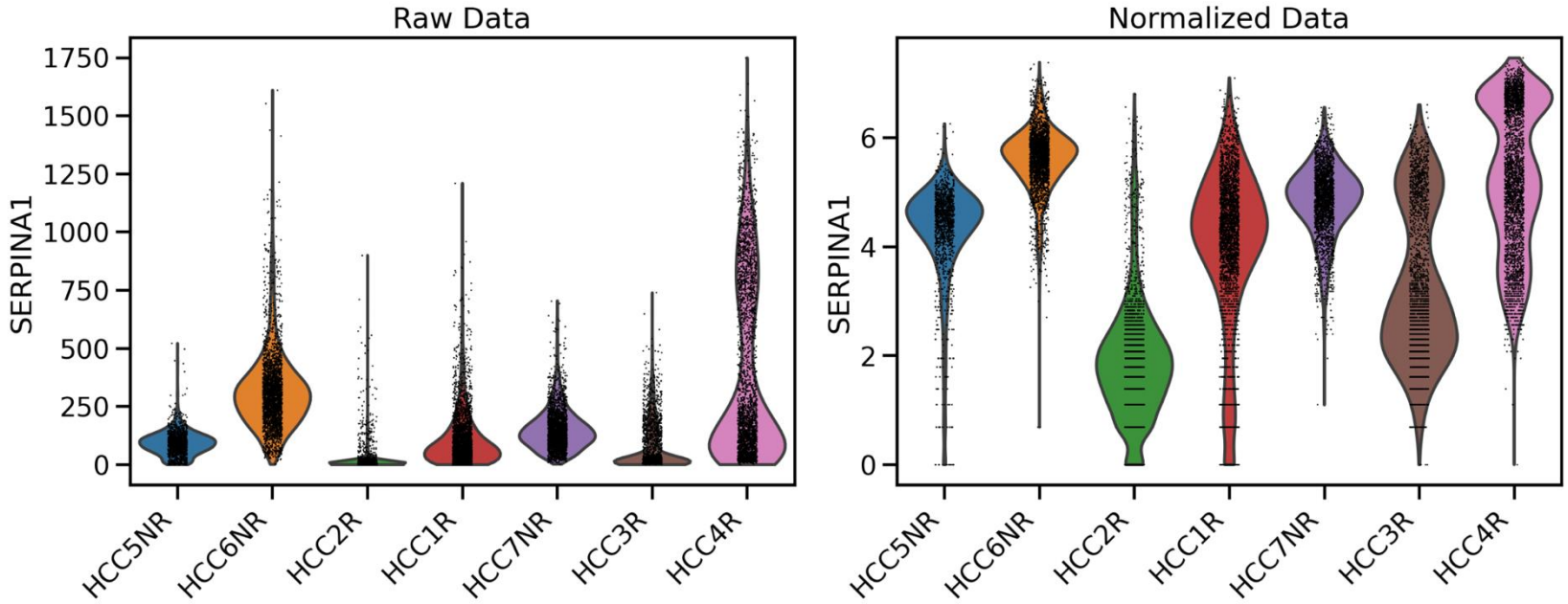




https://github.com/ashoks773/SpatialTranscriptomicsWorkflow/blob/main/notebooks/Exercise_Normalization_AKS.ipynb

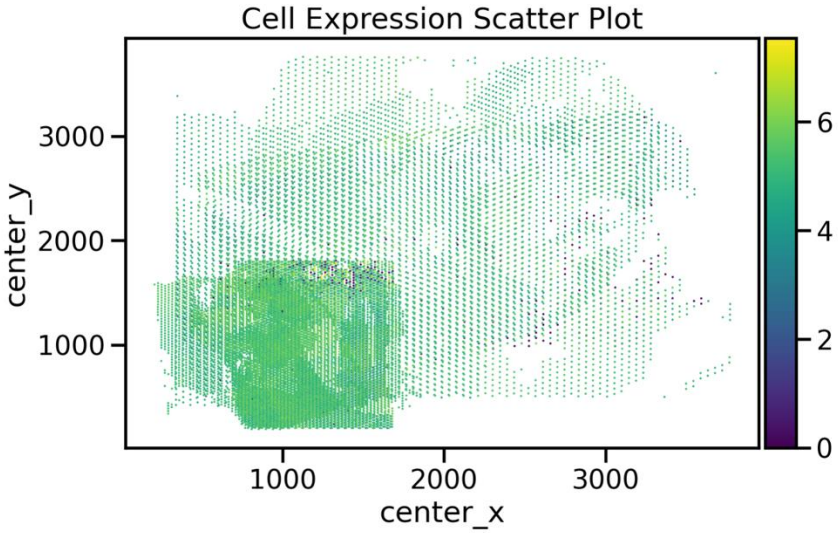
Why normalization is Important – hepatocellular carcinoma (HCC) resection specimen HCC1R – Selected Gene Expression

Apply log1p normalization: sc.pp.log1p(ad_viz)



Why Dimensionality reduction is important

Gene of interest: SERPINA1



PCA: principal component analysis

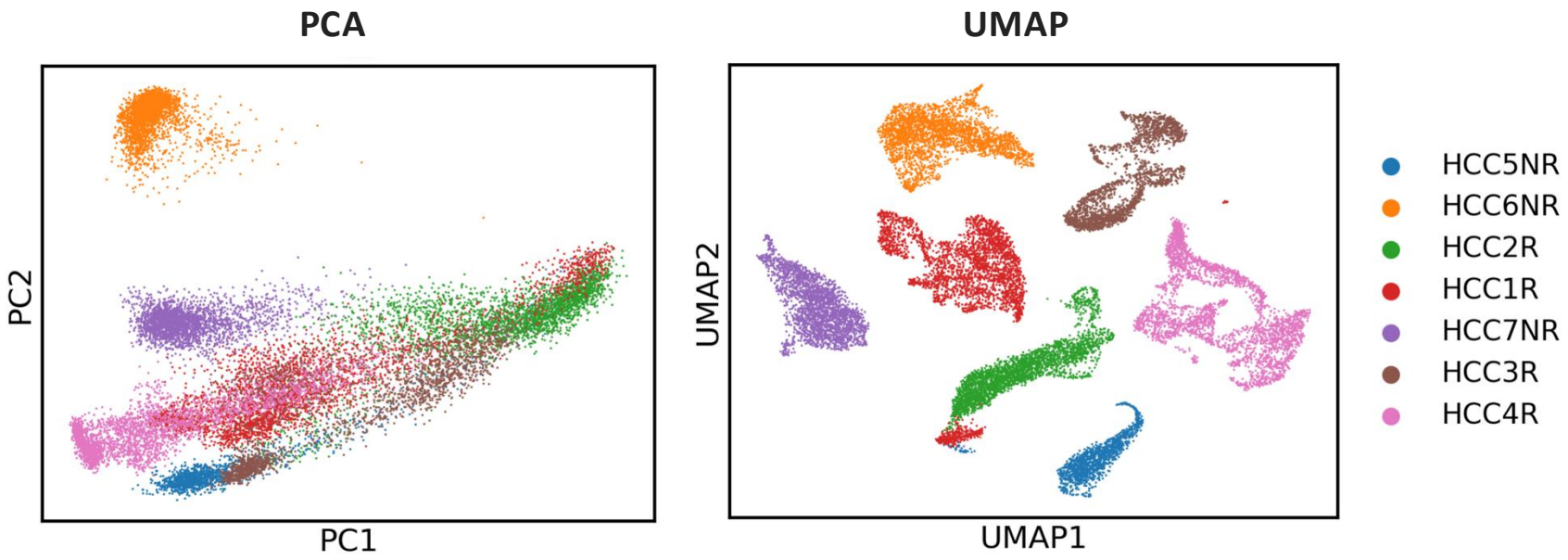
- Simple, linear, clear, robust, fast
- Preserves global structure (everything) as much as possible

UMAP: uniform manifold approximation and projection for dimension reduction

- Complex, non-linear, flexible, slower but reasonably fast
- Preserves local structure (neighborhood) as much as possible

https://www.singlecellcourse.org/basic-quality-control-qc-and-exploration-of-scrna-seq-datasets.html

PCA and UMAP

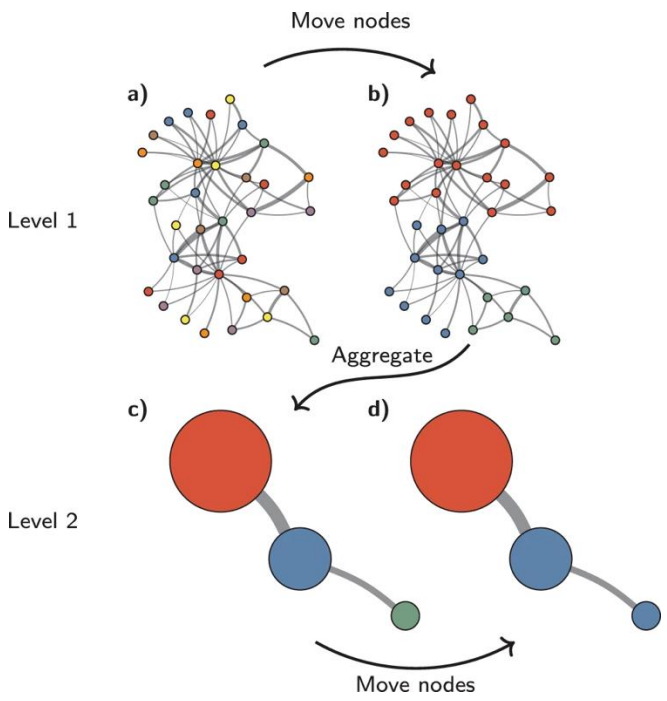

PCA
PC2
PC1
UMAP
UMAP2
UMAP1
HCC5NR
HCC6NR
HCC2R
HCC1R
HCC7NR
HCC3R
HCC4R

Clustering

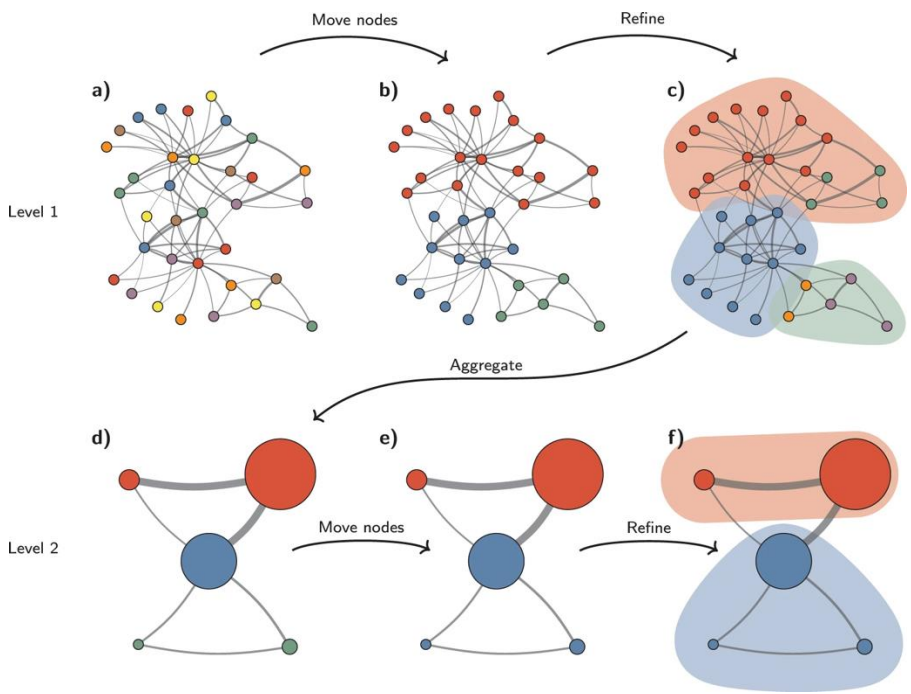
Methods:	Tools:
Hierarchical clustering	SINCERA
k-means	SC3
Graph-based methods	tSNE + k-means
	Seurat clustering

Graph based clustering methods

Louvain algorithm



Leiden algorithm

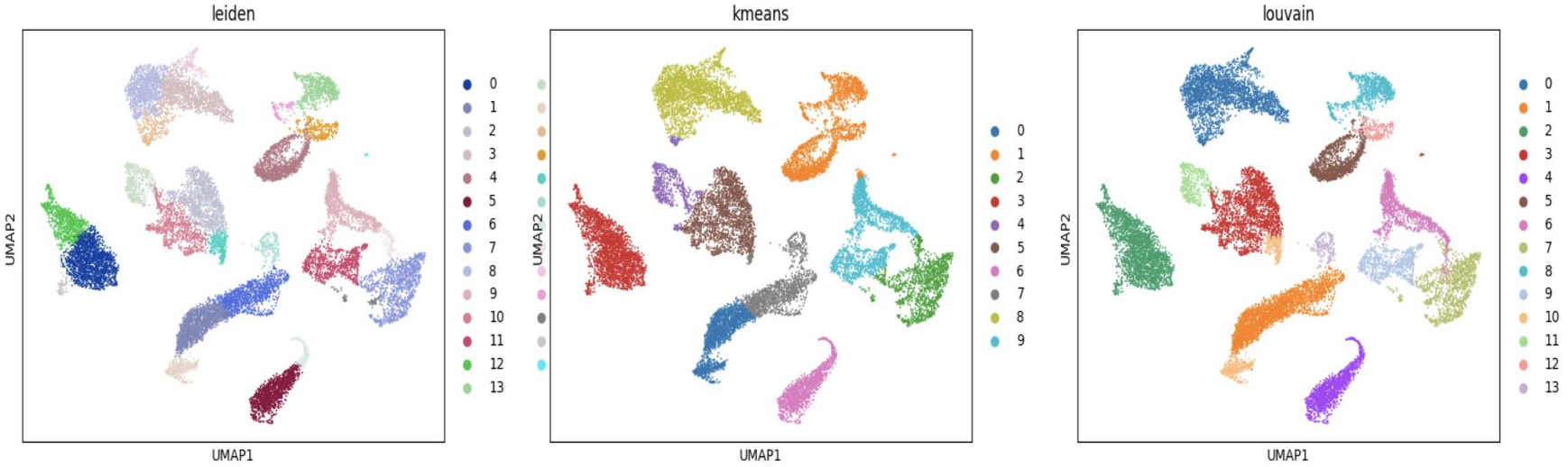


https://www.nature.com/articles/s41598-019-41695-z

Graph based clustering methods

Features	Louvain Clustering	Leiden Clustering
Type of Method	Graph-based (community detection)	Graph-based (community detection)
Algorithm	Greedy optimization algorithm	Improved refinement algorithm based on Louvain
Modularity Optimization	Louvain optimizes for modularity (a measure of the density of edges within clusters vs. between clusters)	Leiden optimizes for modularity as well but includes a refinement step to ensure all nodes within a cluster are well-connected
Partition Stability	Can lead to disconnected or poorly connected clusters (clusters with weak intra-cluster connections)	Ensures that clusters are well-connected internally, which addresses Louvain's tendency to form disconnected clusters
Resolution	Both support resolution parameters, though Louvain may lead to less fine-grained control over community sizes	Leiden typically offers better control and fine-tuning of resolution parameters
Performance	Faster but may lead to less accurate clustering and is more prone to converging on suboptimal solutions	Slightly slower but more robust, reaching more optimal solutions and typically providing higher-quality clusters
Hierarchical Structure	Can be used in a hierarchical setting, but not inherently hierarchical like classic hierarchical clustering	Can also be used in hierarchical clustering; has better handling of hierarchical or multi-level structures in complex graphs
Scalability	Generally faster in large-scale networks but less accurate on some data types	More computationally intensive but more reliable for high-resolution or complex community detection

Comparison of different clustering approach



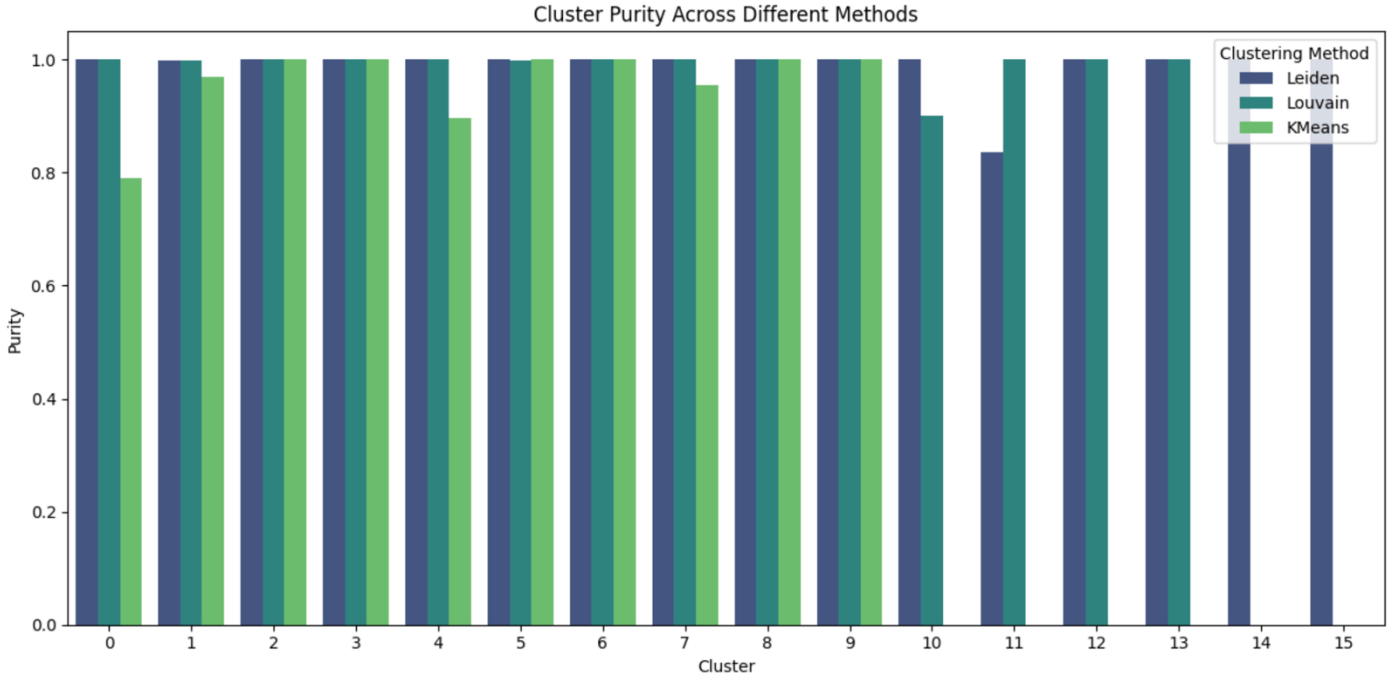
ARI and NMI between

leiden and kmeans: ARI: 0.79, NMI: 0.87

louvain and kmeans ARI: 0.78, NMI: 0.86

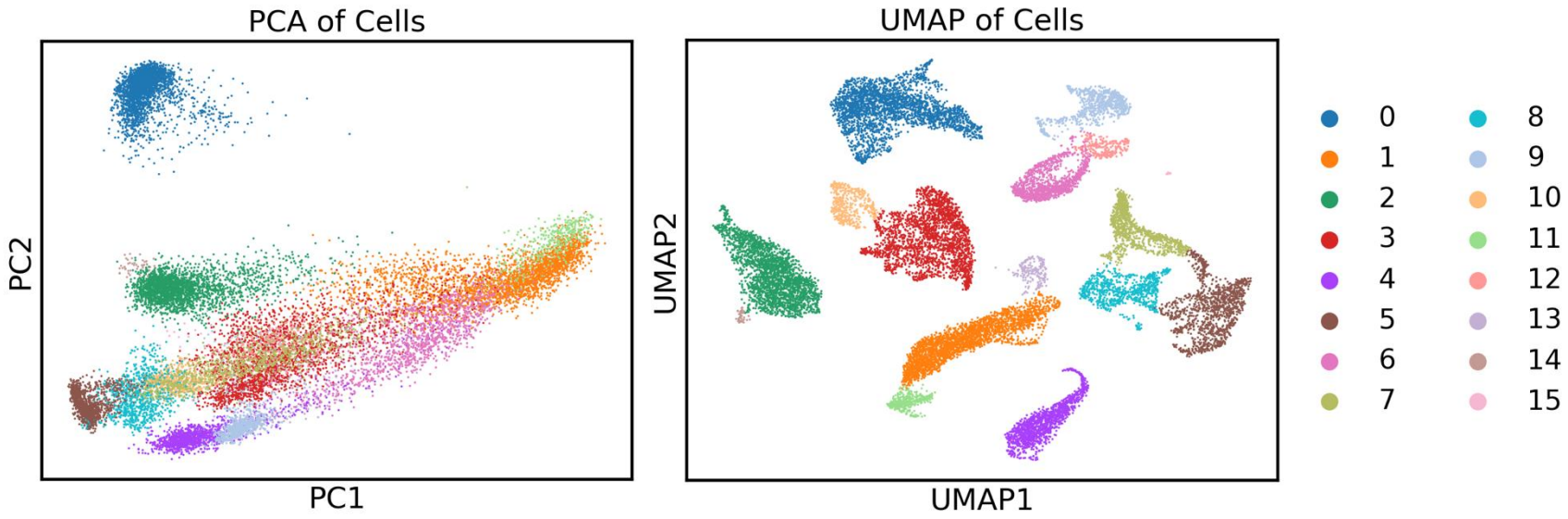
louvain and leiden ARI: 0.95, NMI: 0.96

Comparison of different clustering approach – Cluster Purity

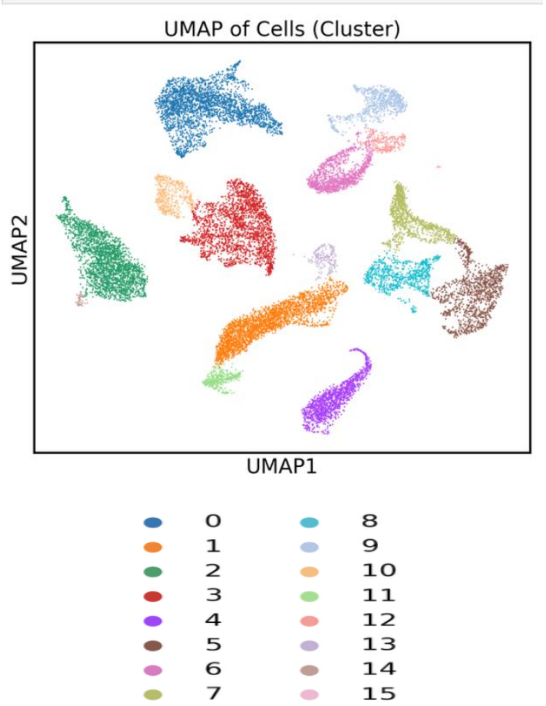


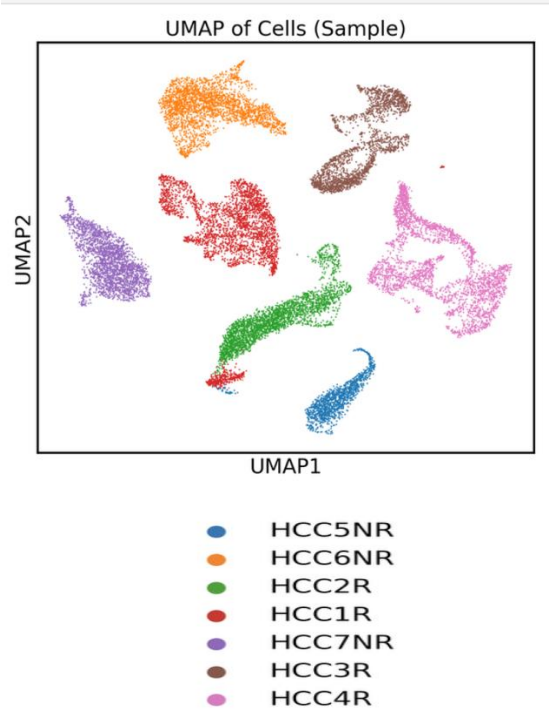
Cluster Purity Summary:
Leiden: 0.99
Louvain: 0.99
KMeans: 0.96

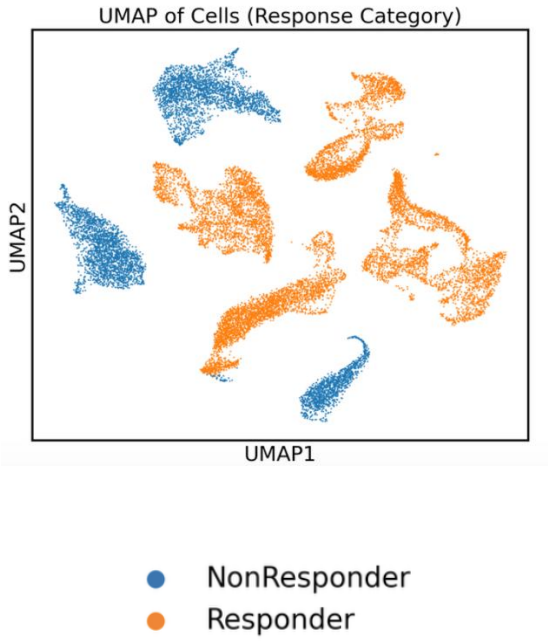
PCA and UMAP – After clustering (Louvain)



UMAP Clusters by Sample and Groups







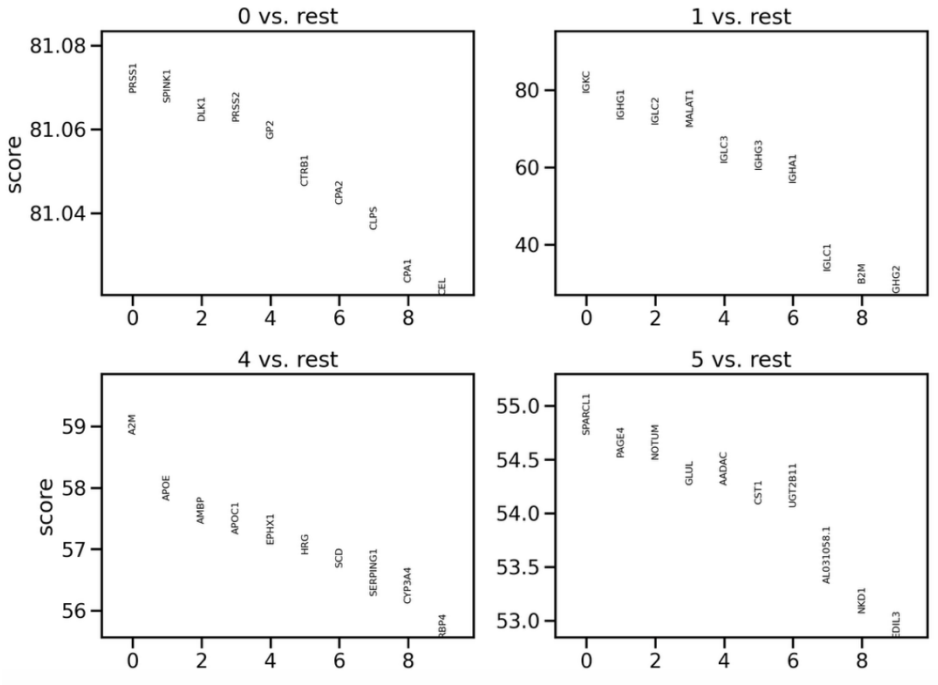
Spatial Transcriptomics Analysis Workflow – Outline

- Introduction to Spatial Transcriptomics
- Select and Download the Dataset
- Load and Create Single-Cell Object
- Normalization, PCA, UMAP, Clustering, and Visualization
- **Marker Gene Identification**
- Cell Type Annotations Using Different Methods

Marker gene identification

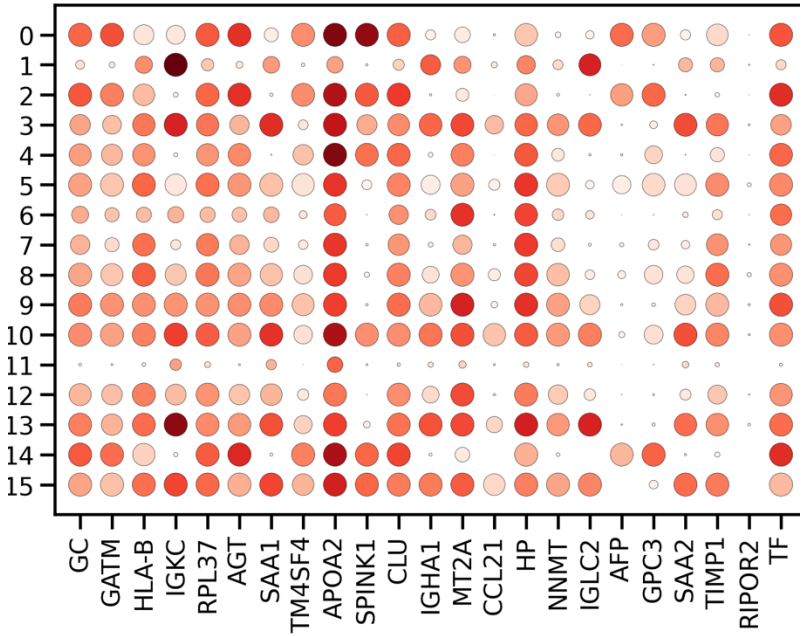
Key Steps:

- Data Preparation and Clustering (**already done**)
- **Ranking Marker Genes (Cluster specific markers)**
- **Statistical Testing**
- Marker Gene Plotting

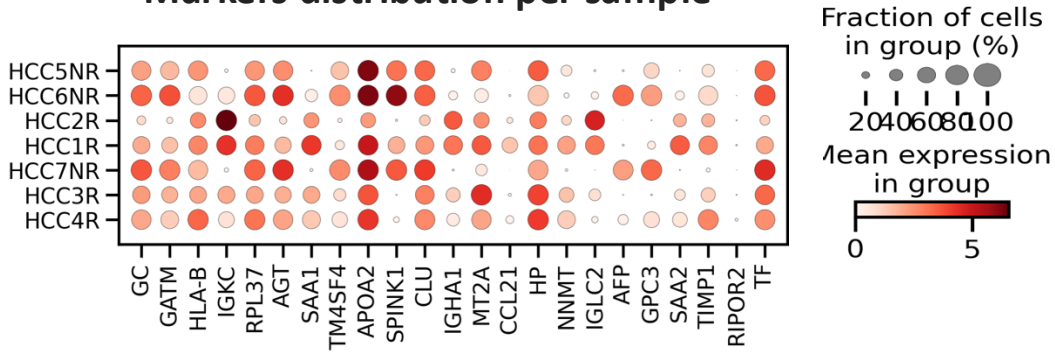


Marker gene identification

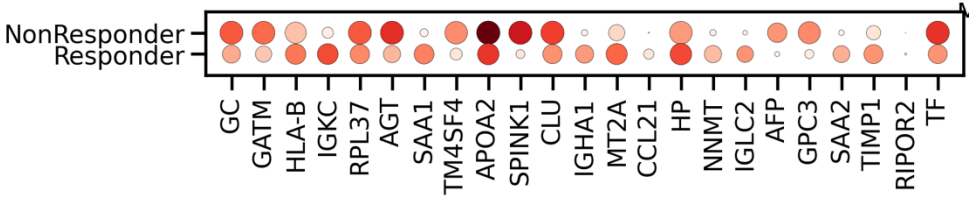
Markers distribution per cluster



Markers distribution per sample



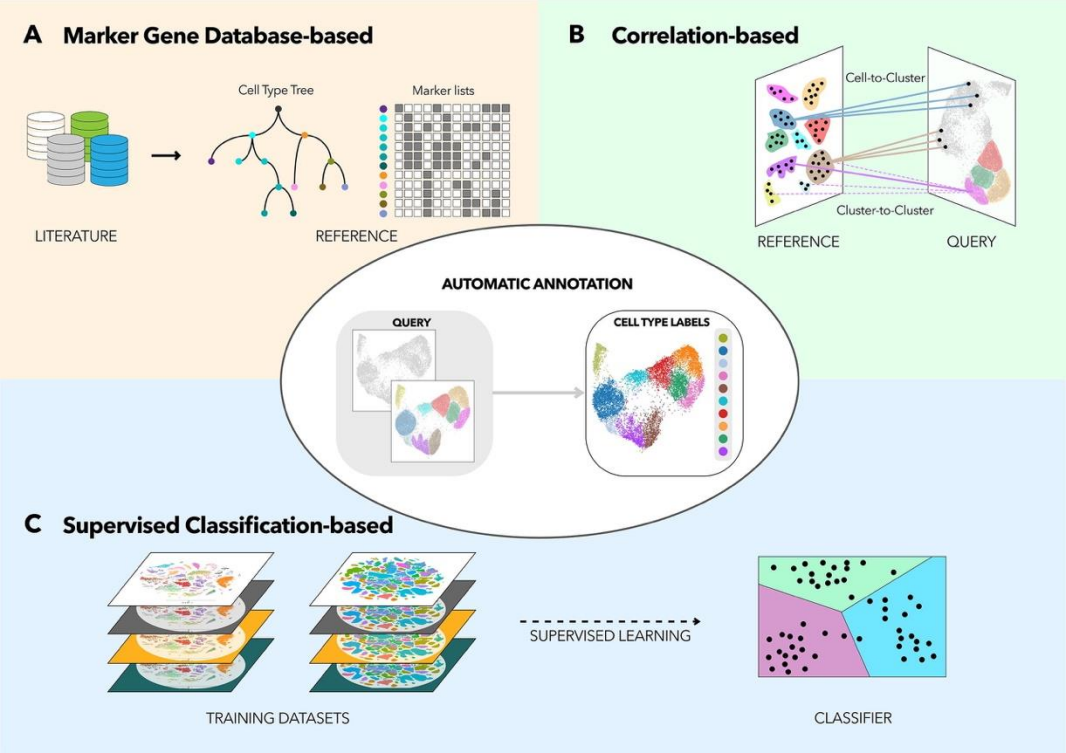
Markers distribution per Condition



Spatial Transcriptomics Analysis Workflow – Outline

- Introduction to Spatial Transcriptomics
- Select and Download the Dataset
- Load and Create Single-Cell Object
- Normalization, PCA, UMAP, Clustering, and Visualization
- Marker Gene Identification
- **Cell Type Annotations Using Different Methods**

Overview of Cell Type Annotation Methods



Marker gene database-based:

- scType; MAESTRO; SCINA; CellAssign

Correlation-based

- scmap-cluster; scmap-cell; SingleR; CHETAH; scMatch; ClustifyR; CIPR

Supervised Classification Based

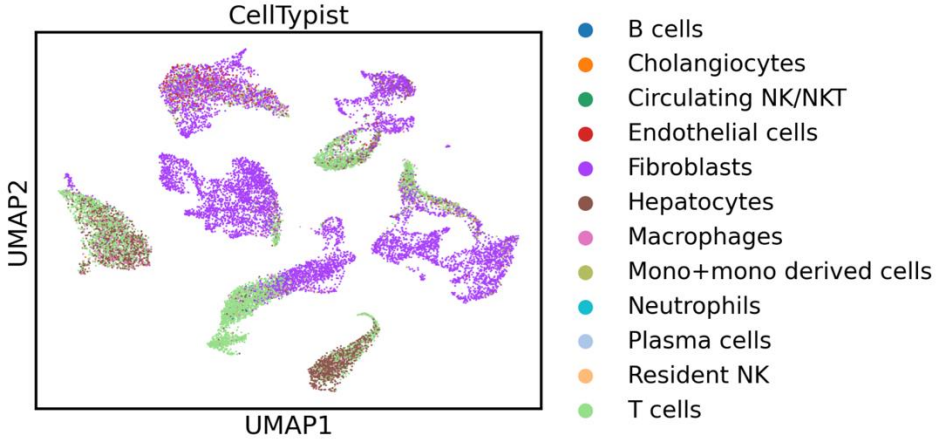
- CellTypist; CaSTLe; Moana; LAmbDA; superCT; SingleCellNet; Garnett; scPred; ACTINN; OnClass; scClassify

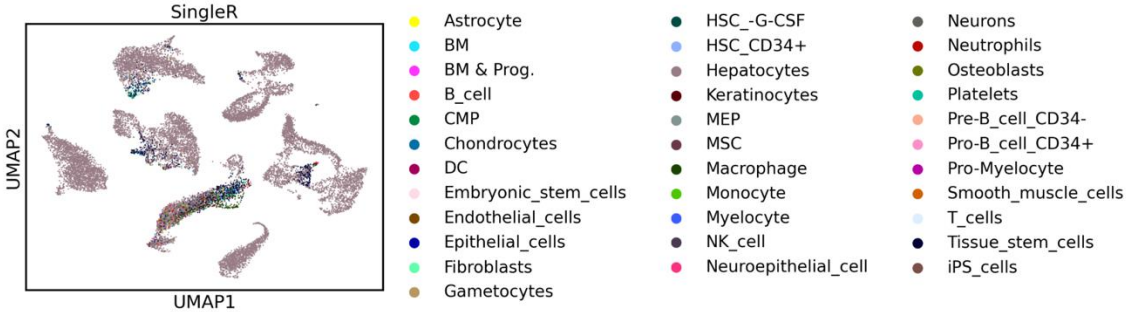
Knowledge Based

- Azimuth; SingleCellNet

https://www.sciencedirect.com/science/article/pii/S2001037021000192

Cell Typist and singleR on smaller subset (considering 10000 genes)





https://github.com/ashoks773/SpatialTranscriptomicsWorkflow/blob/main/notebooks/Exercise_CellTypeAnnotation_AKS.ipynb

Thank you

GitHub: https://github.com/ashoks773/SpatialTranscriptomicsWorkflow

Cedars Sinai