
# Introduction to Spatial Transcriptomic (ST) Data Analysis

**Jenny Wu**

Director of Bioinformatics

Genomics Research and Technology Hub

Chao Family Comprehensive Cancer Center

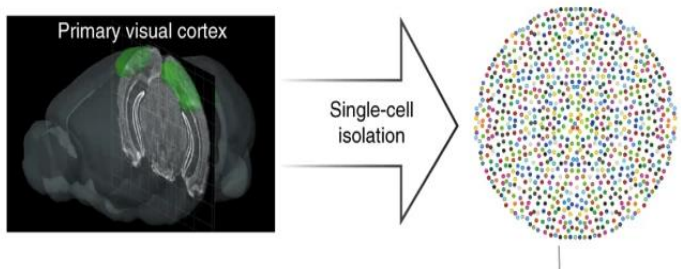
UC Irvine

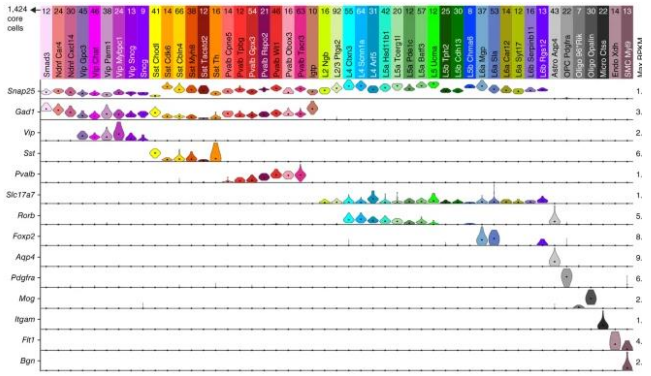
# Outline

- Why spatial omics
- Spatial transcriptomic technologies: Sequencing vs Imaging based
- Data analysis pipeline and applications
  - ✓ Pre-processing: platform dependent
  - ✓ Downstream analysis and visualization
- Pre-processing with Space Ranger for Visium HD and Xenium 5K onboard analysis
- Downstream analysis pipeline - Seurat Workflow
  - ✓ Data import
  - ✓ QC, filtering and feature selection
  - ✓ Dimension reduction and clustering
  - ✓ Data visualization and integration
- Advanced topics: cell segmentation, cell type deconvolution, integration with scRNA data, and cell-cell communication

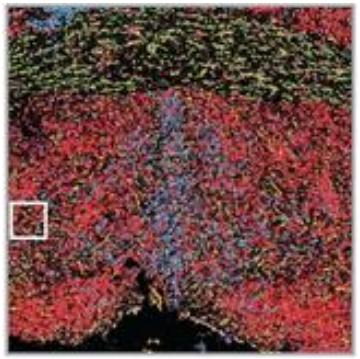
# Why Spatial Biology

- Single cell sequencing reveals cellular heterogeneity



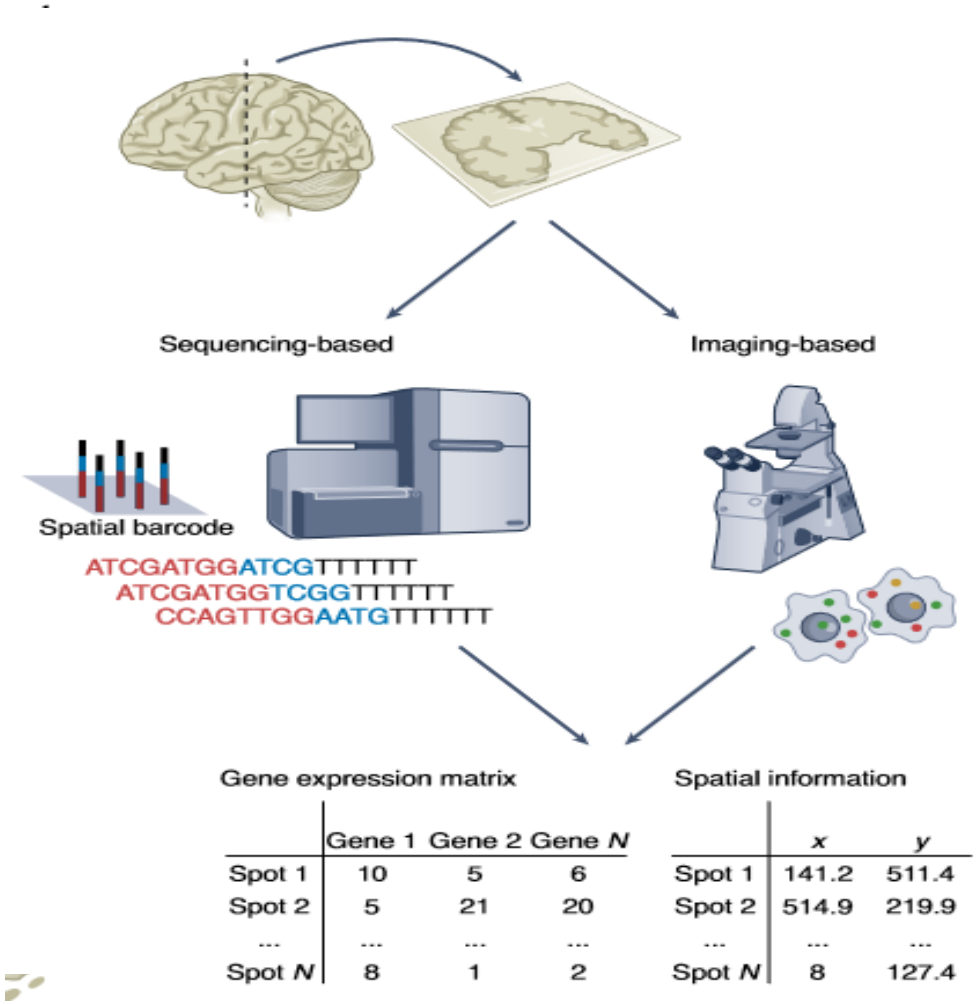



- Spatial information is important



# Two Classes of ST Technology

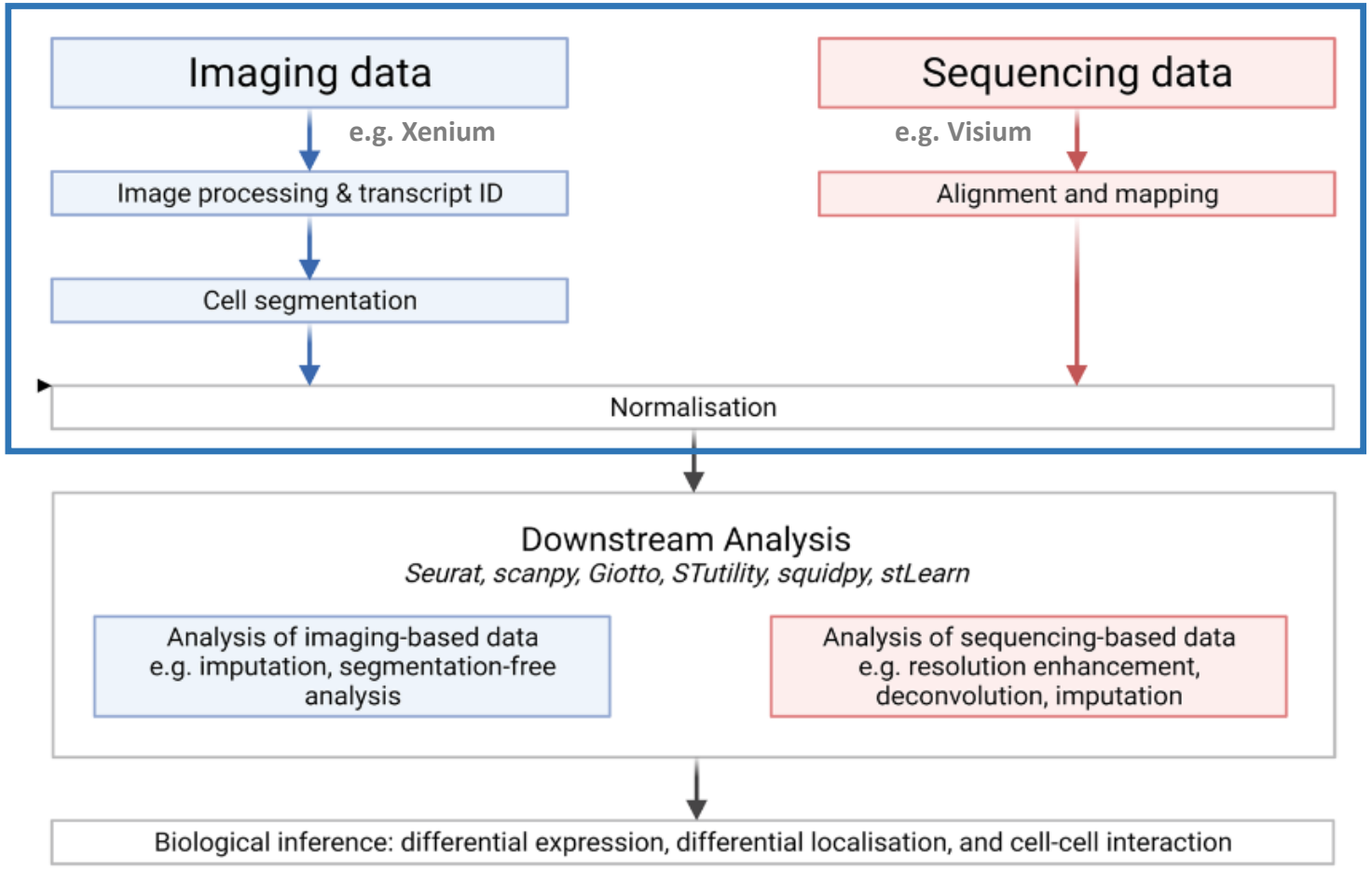
- Sequencing Based (SST) gene detection
  - *10x Visium, Slide-seq2, Stereo-seq etc*
  - Whole transcriptome but not true single cell resolution
- Imaging based (IST) gene detection
  - *10x Xenium, MERFISH, Resolve etc*
  - Subcellular resolution but limited gene throughput




Tian *et al*, 2023

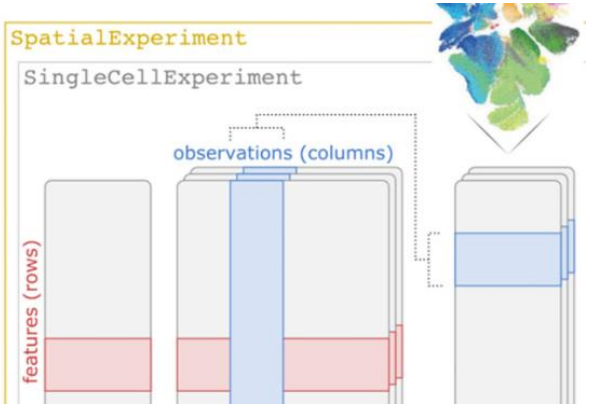
# General Data Analysis Workflow

- Pre-processing is handled by proprietary software such as *Space Ranger or* on Xenium Analyzer instrument.
- Downstream analysis may be performed with a wide range of transcriptomics analysis packages: *Seurat, Scanpy, Squidpy, Giotto* etc.




Williams *et al*, 2022

# Analytical tools for Downstream Analysis


Giotto

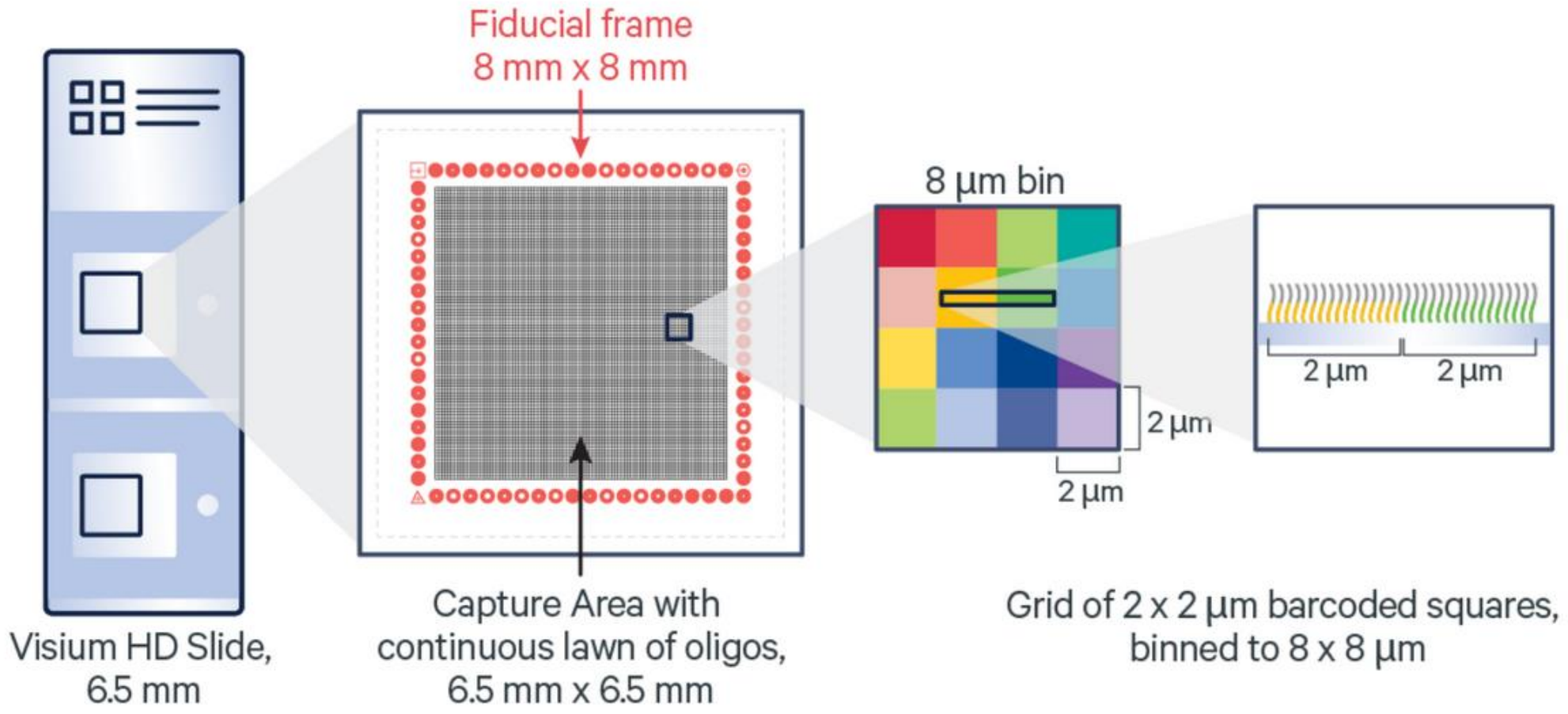
squidpy

QuPath

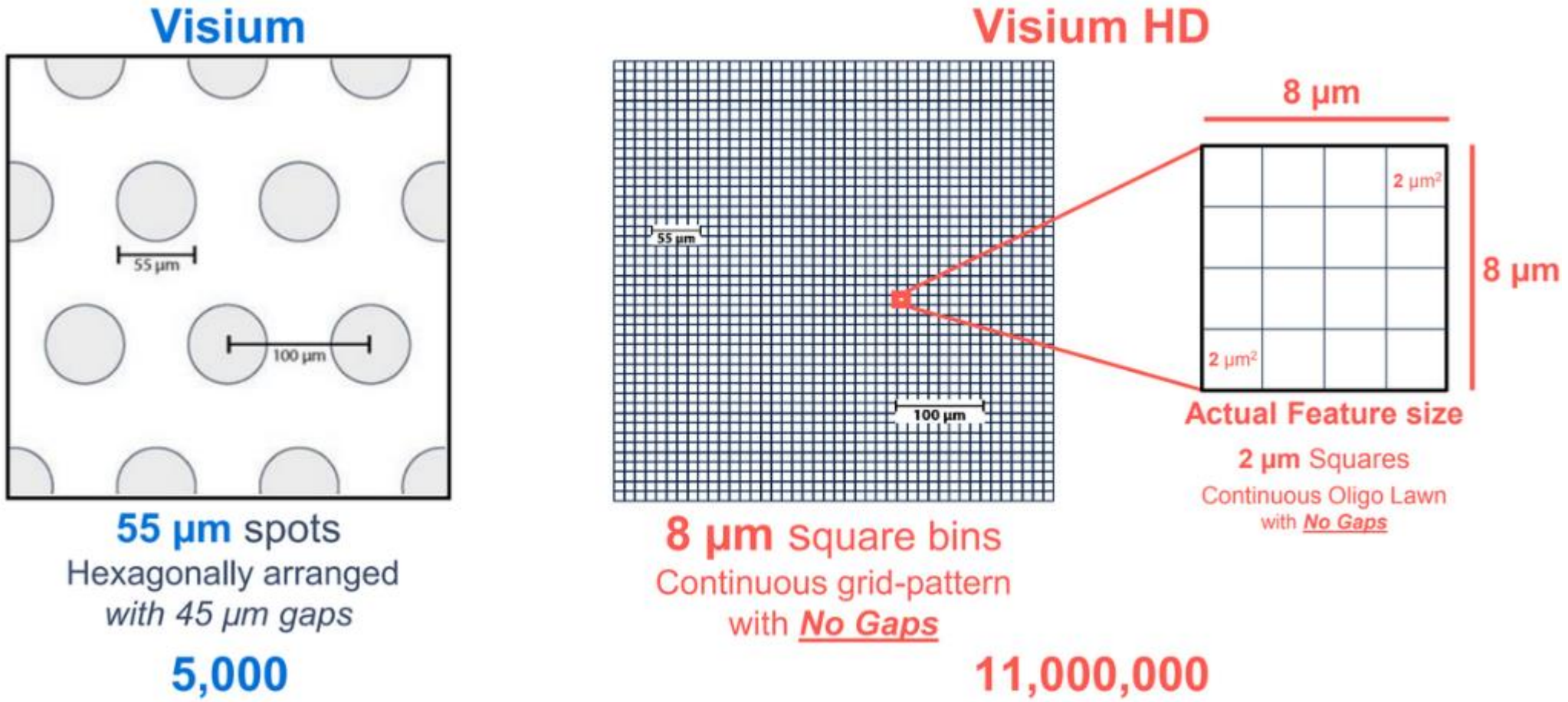
sopa

Spatial omics pipeline and analysis

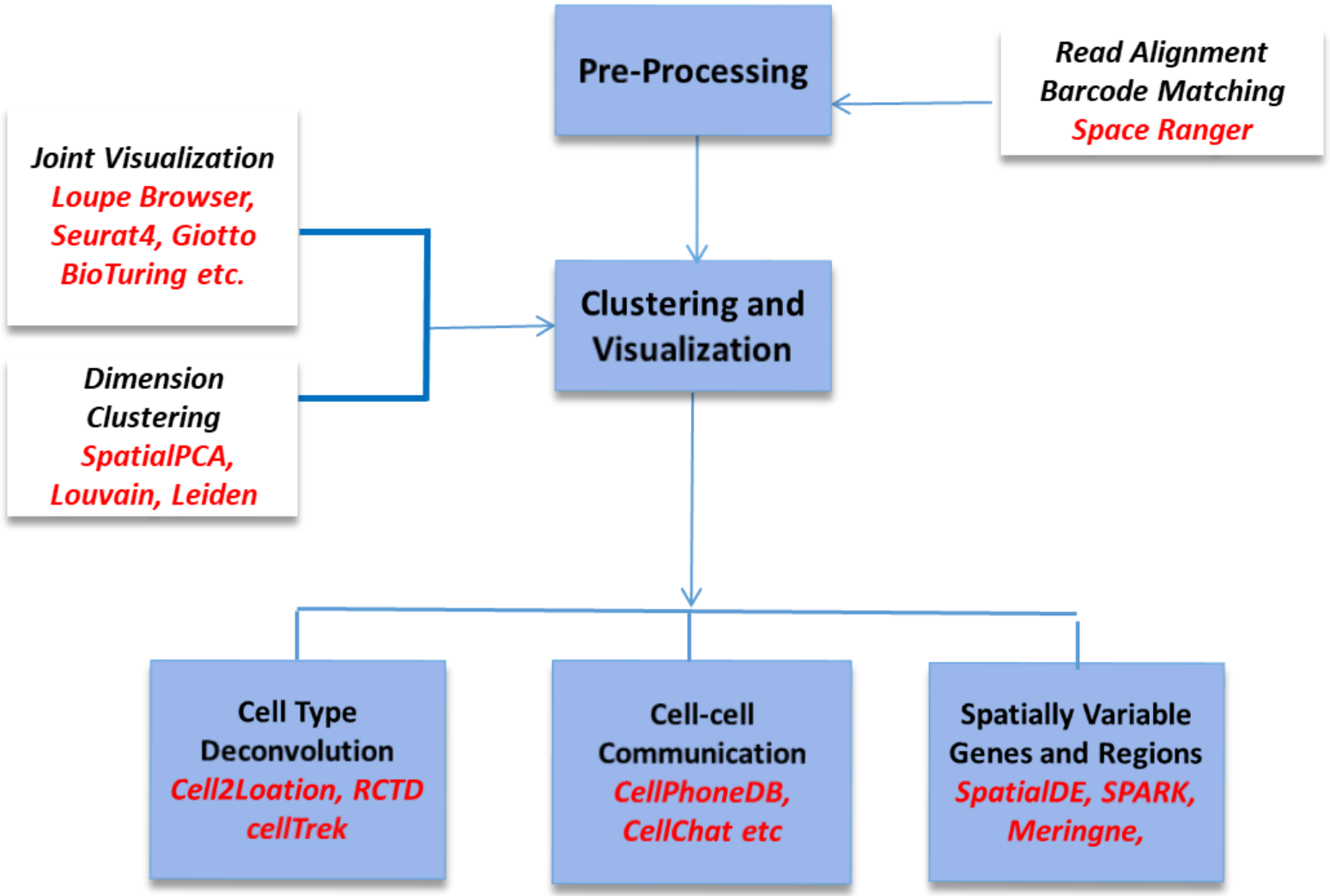
# Visium HD


Fiducial frame
8 mm x 8 mm
8 µm bin
2 µm
2 µm
2 µm
2 µm
Visium HD Slide,
6.5 mm
Capture Area with
continuous lawn of oligos,
6.5 mm x 6.5 mm
Grid of 2 x 2 µm barcoded squares,
binned to 8 x 8 µm

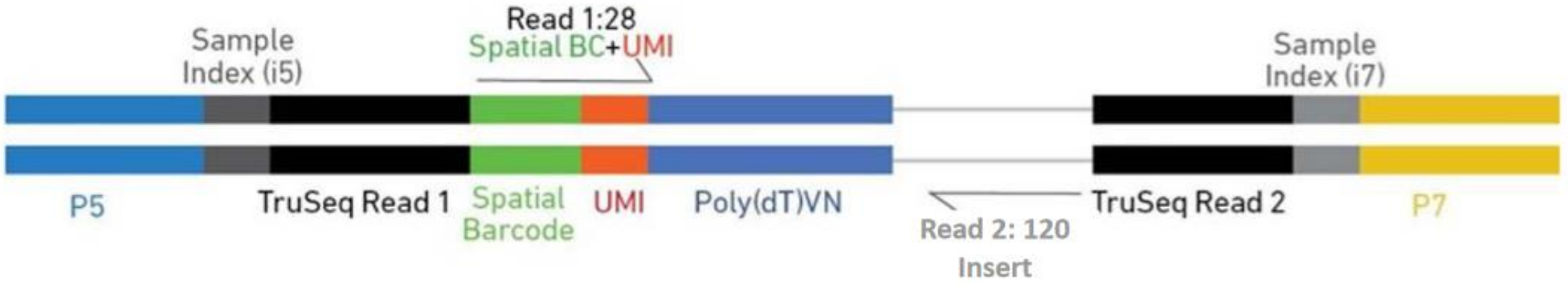
# Visium HD vs Visium



# Visium (HD) Data Analysis Work Flow



# 10x Visium HD Data and Space Ranger Count




```
$ cd /home/jdoe/runs
$ spaceranger count --id=sample345 \ #Output directory
                    --transcriptome=/home/jdoe/refdata/GRCh38-2020-A \ #Path to Reference
                    --fastqs=/home/jdoe/runs/HAWT7ADXX/outs/fastq_path \ #Path to FASTQs
                    --sample=mysample \ #Sample name from FASTQ filename
                    --image=/home/jdoe/runs/images/sample345.tiff \ #Path to brightfield image
                    --slide=V19J01-123 \ #Slide ID
                    --area=A1 \ #Capture area
                    --localcores=8 \ #Allowed cores in localmode
                    --localmem=64 #Allowed memory (GB) in localmode
```

https://www.10xgenomics.com/support/software/space-ranger/latest

# Space Ranger Output: Web Summary

10x GENOMICS Space Ranger • count

## Visium_HD_Human_Breast_Cancer_Fresh_Frozen - Gene expression library of Fresh Frozen Human Breast Cancer (Visium HD) using the Human Whole Transcriptome Probe Set

Summary | Image Alignment | Bin-Level Metrics

### Key Metrics

| | |
|---|---|
| 472,859 | 1581.5 |
| Number of 8 µm binned squares under tissue | Mean reads per 8 µm bin |
| 772.8 | 17,527 |
| Mean UMIs per 8 µm bin | Total genes detected |

### Mapping

| | |
|---|---|
| Reads Mapped to Probe Set | 98.5% |
| Reads Mapped Confidently to Probe Set | 98.1% |
| Reads Mapped Confidently to the Filtered Probe Set | 94.8% |
| Reads Half-Mapped to Probe Set | 0.2% |
| Reads Split-Mapped to Probe Set | 0.2% |

### Total UMI Count to Image Alignment

The total UMI count in each 8 µm bin is overlaid onto the tissue image below to assess bin alignment and tissue detection. The highest value on the color scale corresponds to the 98th percentile of UMIs per 8 µm bin under tissue, excluding bins with no UMIs.

Check that the overlay matches tissue morphology and covers all of the tissue of interest within the capture area.

If the overlay is inaccurate, use the manual fiducial alignment and tissue detection workflow in Loupe Browser.

Additional QC images are in the Image Alignment tab for review.

All bins | Bins under tissue  UMI Color scale: Jet

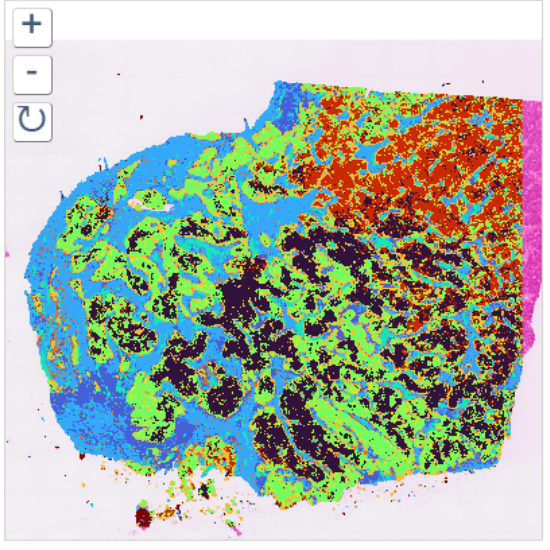
UMI counts

0 200 400 600 800 1000 1200 1400 1600

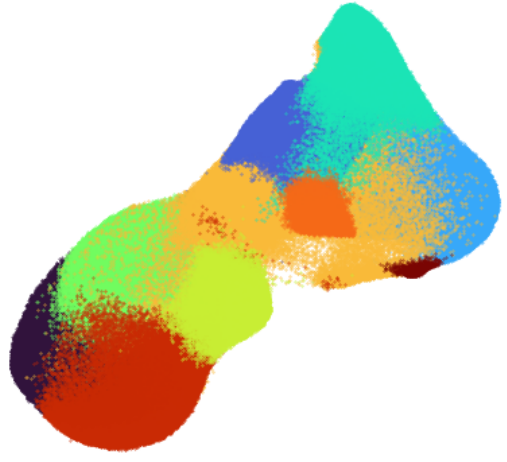
### Clustering

8 µm bin | 16 µm bin

Tissue plot with 8µm bins colored by Graph-based clustering



UMAP Projection of 8µm bins colored by Graph-based clustering



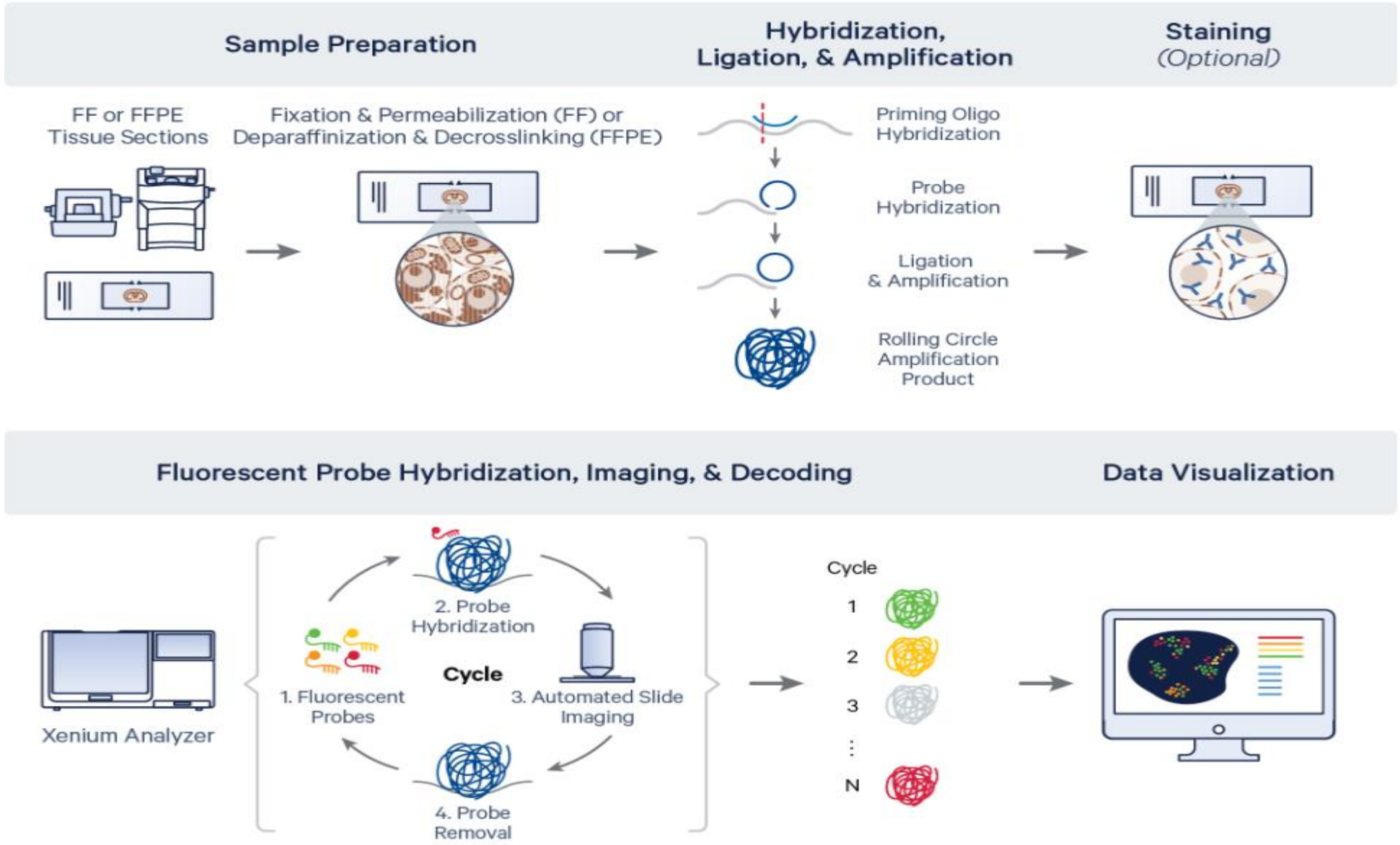
Select All
Cluster 1
Cluster 2
Cluster 3
Cluster 4
Cluster 5
Cluster 6
Cluster 7
Cluster 8
Cluster 9
Cluster 10

Bin opacity:  Display Tissue

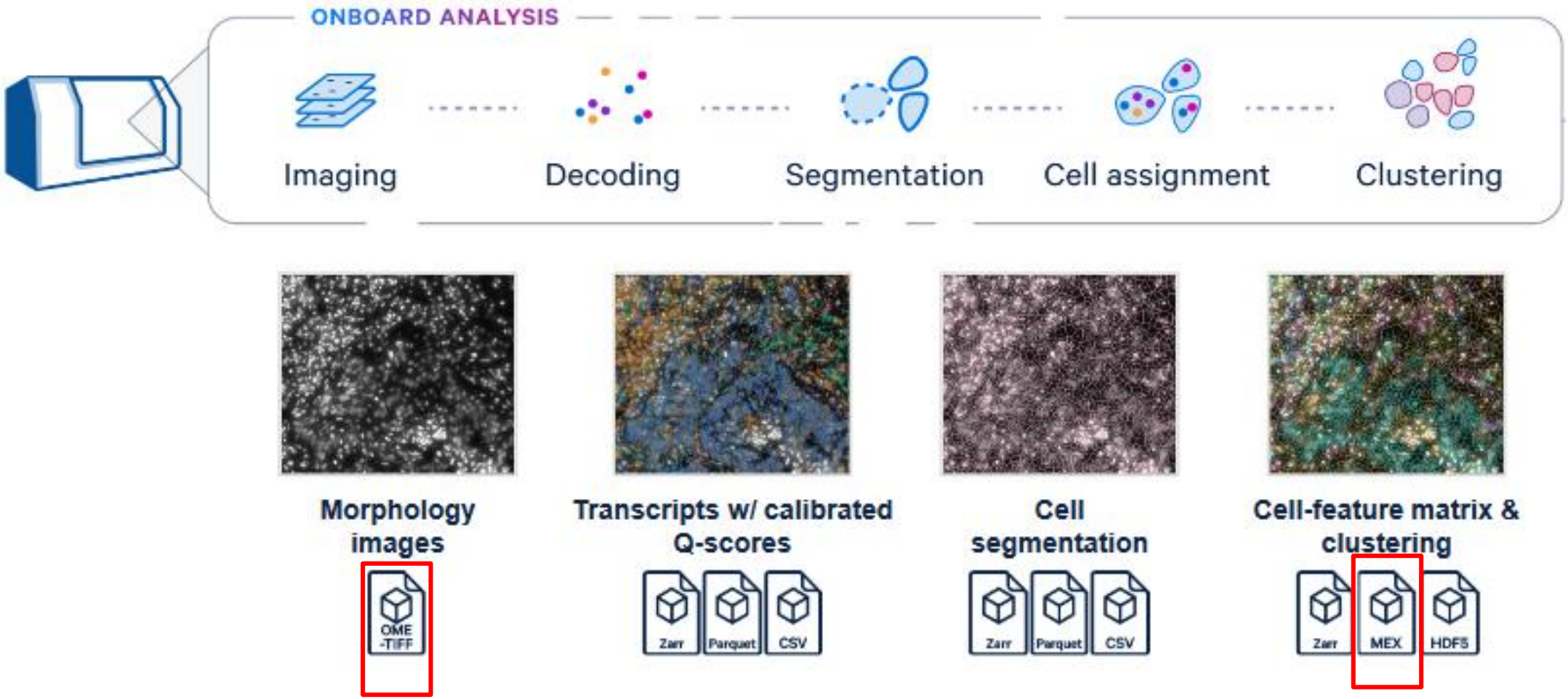
### Top Features by Cluster (Log2 fold-change, p-value)

| Feature | | Cluster 1 | | Cluster 2 | | Cluster 3 | | Cluster 4 | | Cluster 5 | | Cluster 6 | | C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | Name | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2FC | p-value | L2F |
| ENSG00000182704 | TSKU | 1.95 | 4e-26 | -4.05 | 2e-27 | -0.28 | 3e-1 | -0.76 | 8e-2 | -0.04 | 1e+0 | -2.98 | 9e-10 | -2 |
| ENSG00000196228 | SULT1C3 | 1.90 | 8e-25 | -4.18 | 9e-29 | -1.07 | 2e-6 | -0.69 | 1e-1 | 0.97 | 1e-4 | -3.19 | 9e-11 | -2 |
| ENSG00000189058 | APOD | 1.55 | 2e-16 | -4.26 | 2e-29 | -0.67 | 5e-3 | -0.68 | 2e-1 | 1.12 | 5e-6 | -3.19 | 9e-11 | -2 |
| ENSG00000198650 | TAT | 1.35 | 2e-12 | -4.05 | 2e-27 | -0.75 | 1e-3 | -0.58 | 3e-1 | 1.34 | 1e-8 | -3.07 | 3e-10 | -2 |
| ENSG00000164434 | FABP7 | 1.28 | 3e-11 | -3.67 | 5e-24 | -0.25 | 4e-1 | -0.26 | 1e+0 | 0.73 | 8e-3 | -2.76 | 9e-9 | -2 |

# Xenium Prime 5K



# Xenium Onboard Analysis Output Formats




https://www.10xgenomics.com/support/software/xenium-onboard-analysis/latest/analysis/

# Xenium 5K Web Summary

10X GENOMICS **Xenium Analysis Summary**

Run name
Mouse FF Brain Hemisphere (Coronal) 5K

Region name
Mouse FF Brain Hemisphere (Coronal)

Run start time
Jun 28, 2024, 20:28 GMT

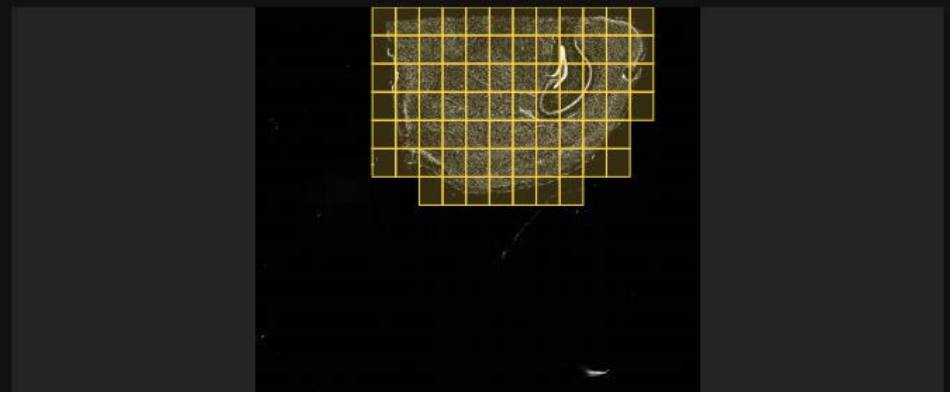
**Summary** | Decoding | Cell Segmentation | Analysis | Image QC

## Key Metrics

| 63,173 | 1,088 | 1,692.3 | 149,297,699 |
|---|---|---|---|
| Number of cells detected | Median transcripts per cell | Nuclear transcripts per 100 μm² | Total high quality decoded transcripts |

## Sample Region Summary

| | |
|---|---|
| Region name | **Mouse FF Brain Hemisphere (Coronal)** |
| Slide ID | **N/A** |
| Cassette name | **N/A** |
| Preparation method | **FF** |



# Xenium 5K Web Summary

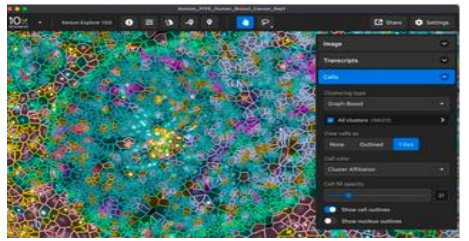
# Various Tools for Xenium Analysis

**Xenium Onboard Analysis**

- Assigns decoded transcripts to segmented cells.
- Includes clustering and differential expression.
- On-instrument Analysis Summary file

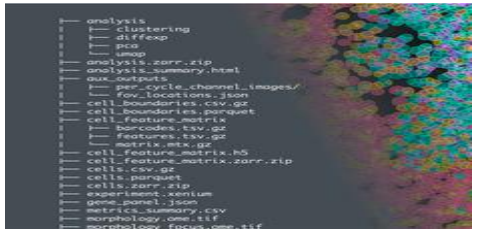
**Xenium Explorer**

- Interactive data exploration and visualization tool
- Pinpoints specific transcripts, check cell segmentation, and inform downstream analysis.
- Runs on Windows and macOS.



**Xenium Ranger**

- Enables reanalysis and custom segmentation.
- Runs on a range of Linux distributions.



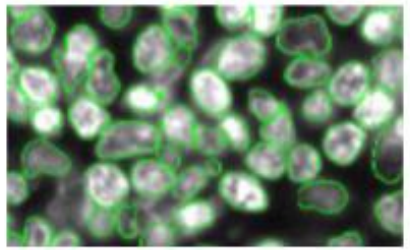

**Community Tools**

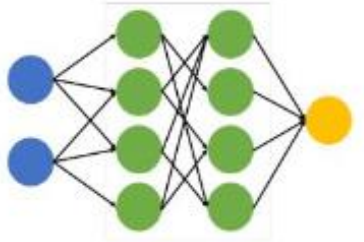
- Primarily programming libraries with some stand-alone tools (*e.g.*, Seurat)
- Developed by the broader research community.
- Not officially supported by 10x Genomics

squidpy

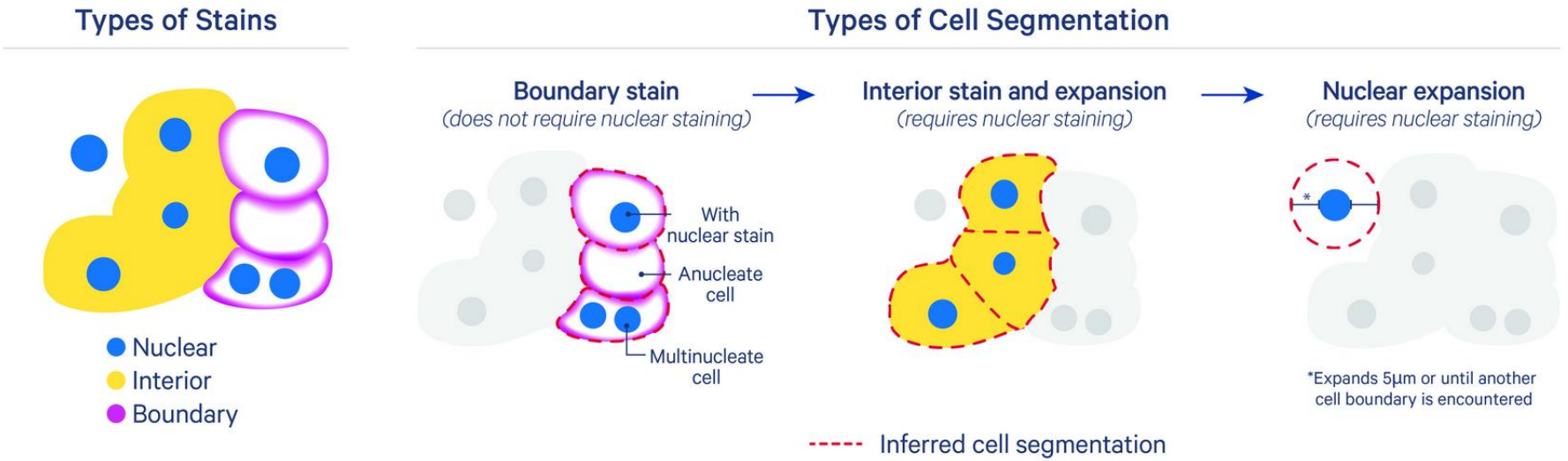
# Imaging Based ST: Cell Segmentation

- Many of the downstream analyses and interpretations of the spatially resolved data depend on the ability to resolve individual cells
  - *Cellpose*, *Baysor (Xenium)*
  - *Stardist (Visium HD)*, *deepcell*
  - Spot-based Spatial cell-type Analysis by Multidimensional mRNA density estimation (*SSAM*)
  - *DL* can be computational intensive



# 10x Xenium Multimodal Segmentation



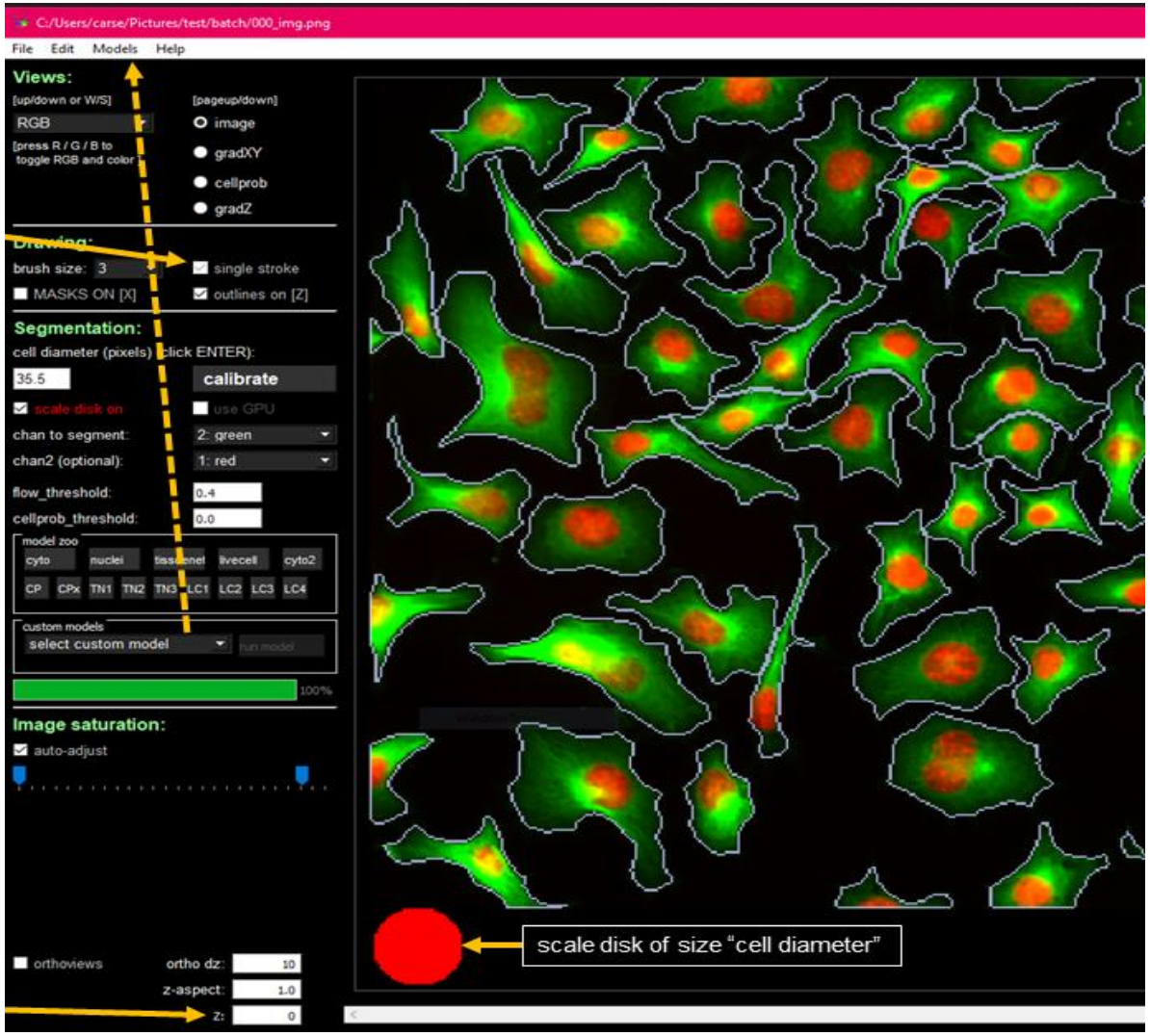

The segmentation results are prioritized in this order for each cell:

- Segment cells based on their cell boundary stain
- Segment cells based on expansion from the nucleus to the cell interior stain edge
- Nuclear expansion

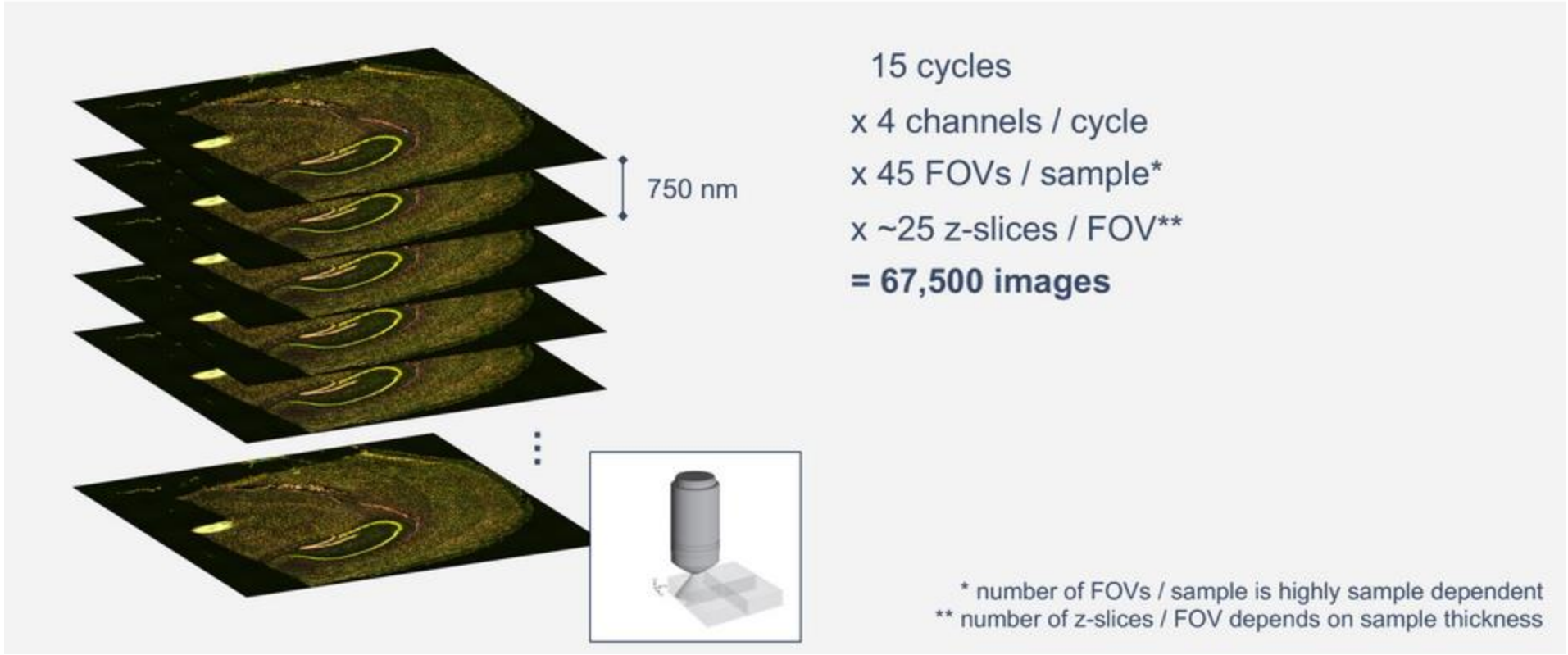
# Optimizing Xenium Segmentation

## Cellpose + Baysor

- Xenium's current segmentation: Cookie cutter style, 2D, nuclei segmentation based on DAPI, followed by an expansion of the segmentation masks with 15 micron
- Nuclei segmentation using Cellpose and the assignment of reads to individual cells using Baysor



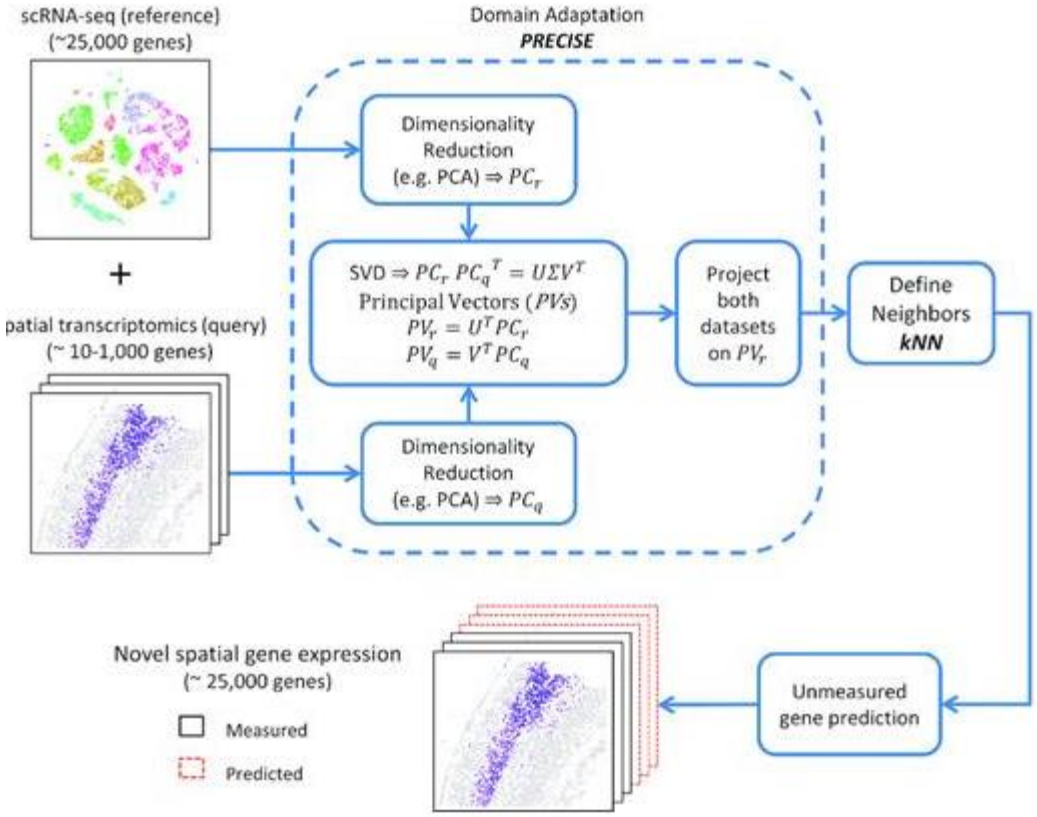
# Increasing Volume of Data



# Xenium: Gene Imputation

- Predicting gene expression from a reference scRNA-seq onto the cellular-resolution ST dataset

- *gimVI, SpaGE, Tangram, SpaOTsc; RCTD, Seurat integration, Liger …*



# Seurat: Data Import: Visium HD

## Load a 10x Genomics Visium Spatial Experiment into a `Seurat` object

Source: `R/preprocessing.R`

Load a 10x Genomics Visium Spatial Experiment into a `Seurat` object

```
Load10X_Spatial(
  data.dir,
  filename = "filtered_feature_bc_matrix.h5",
  assay = "Spatial",
  slice = "slice1",
  bin.size = NULL,
  filter.matrix = TRUE,
  to.upper = FALSE,
  image = NULL,
  ...
)
```

# Seurat: Data Import: Xenium

## Read and Load 10x Genomics Xenium in-situ data

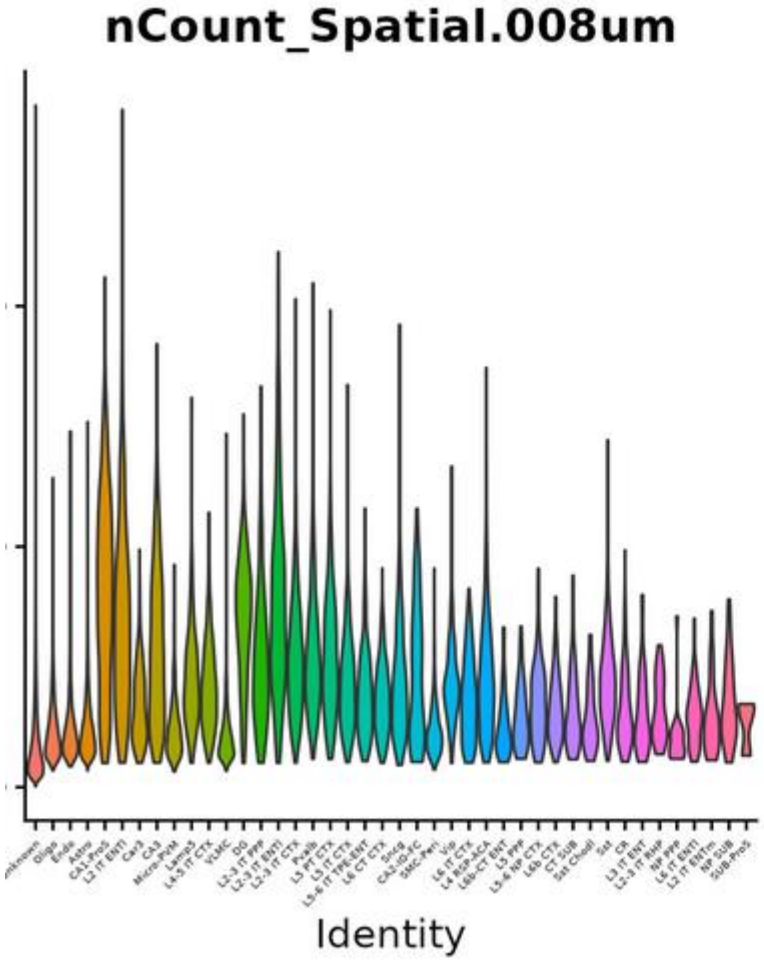
Source: `R/convenience.R`, `R/preprocessing.R`

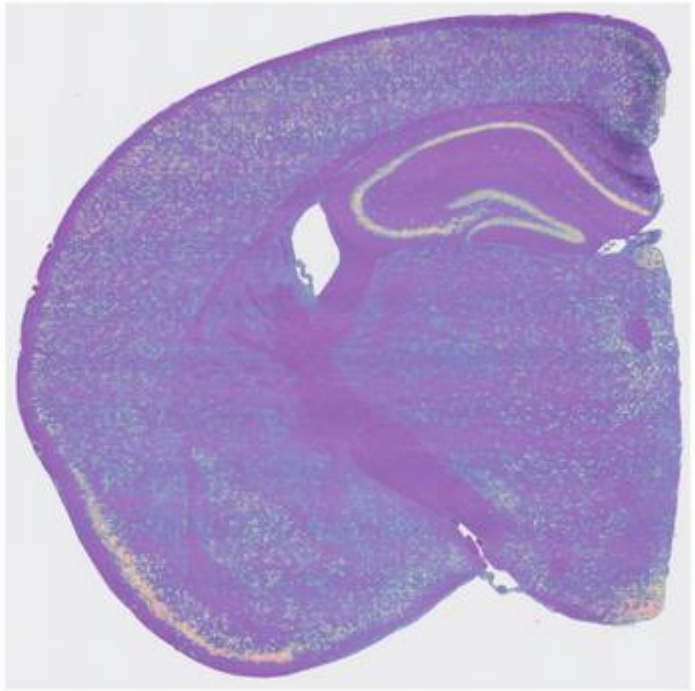
Read and Load 10x Genomics Xenium in-situ data

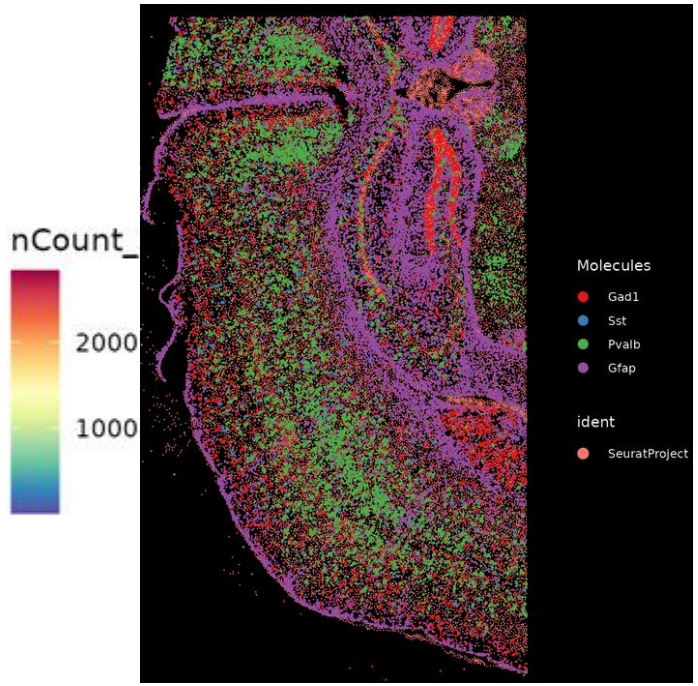
```
LoadXenium(data.dir, fov = "fov", assay = "Xenium")

ReadXenium(
  data.dir,
  outs = c("matrix", "microns"),
  type = "centroids",
  mols.qv.threshold = 20
)
```

# Seurat: QC and Filtering



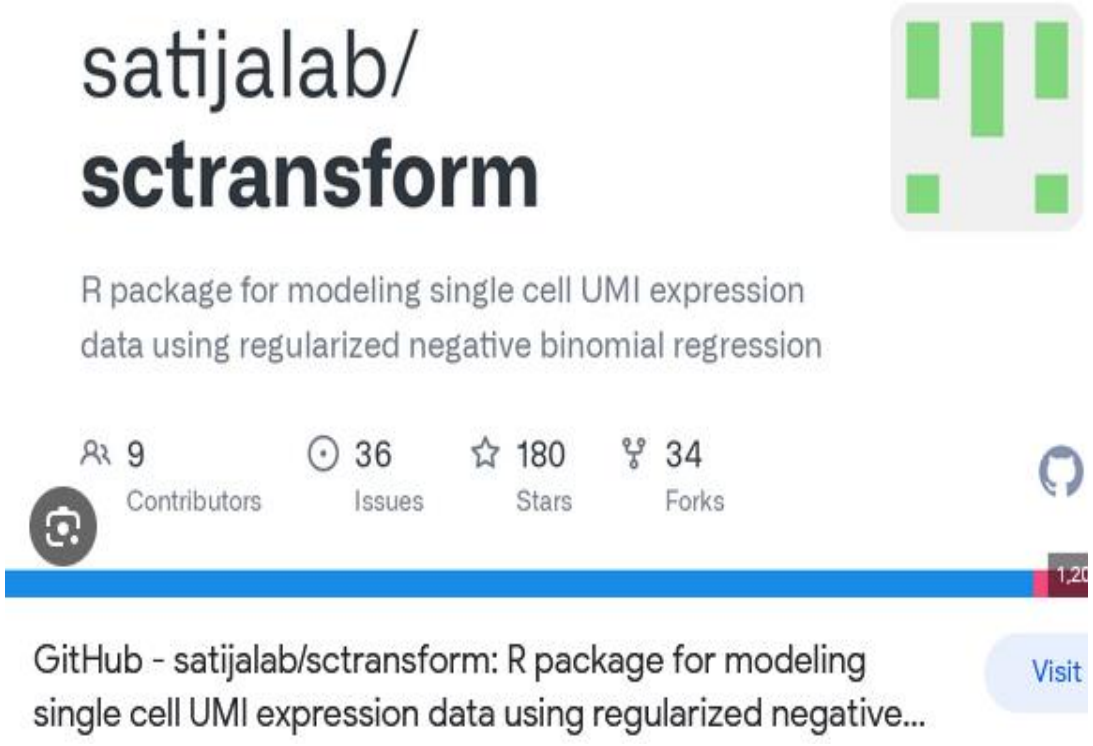






vln.plot <- VlnPlot(object, features = "nCount_Spatial.008um", pt.size = 0)

count.plot <- SpatialFeaturePlot(object, features = "nCount_Spatial.008um")

# Normalization and Feature Selection



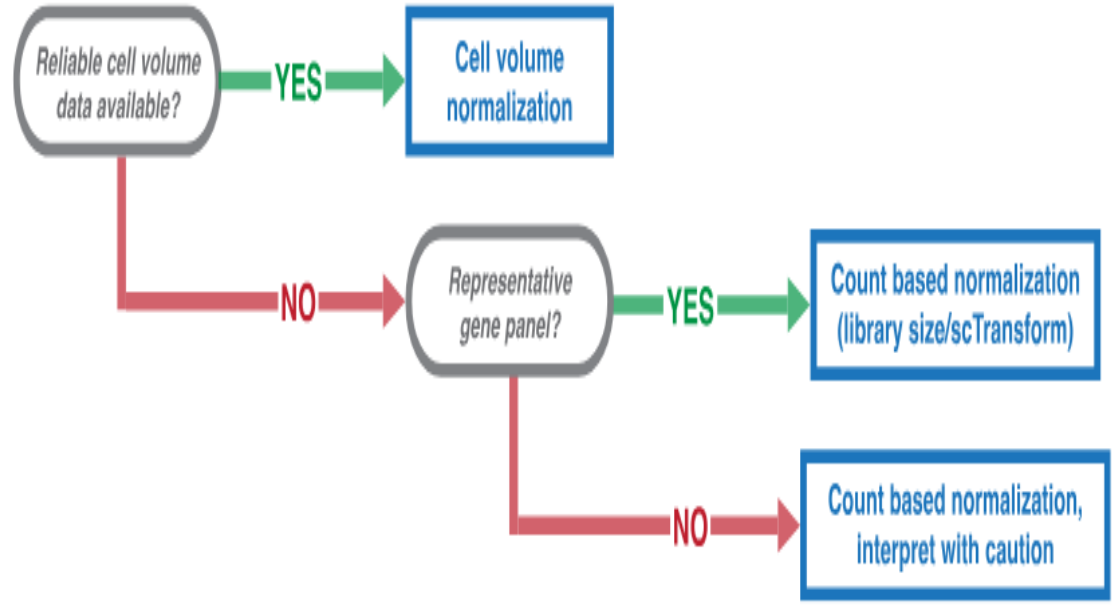




**Fig. 7** Decision tree summarizing recommendations for gene count normalization method selection for im-SRT data

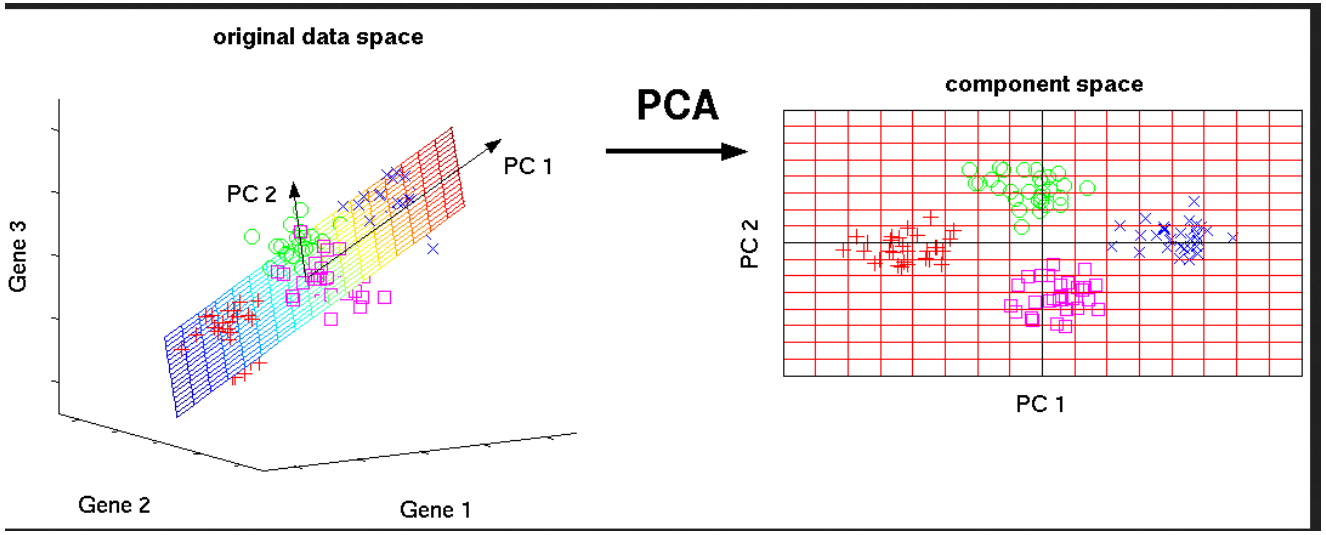
SpaNorm

(https://www.biorxiv.org/content/10.1101/2024.05.31.596908v1)

brain <- SCTransform(brain, assay = "Spatial", verbose = FALSE")

# Linear Dimension Reduction

- Principle Component Analysis (PCA) is a standard technique for visualizing high dimensional data and for data pre-processing.



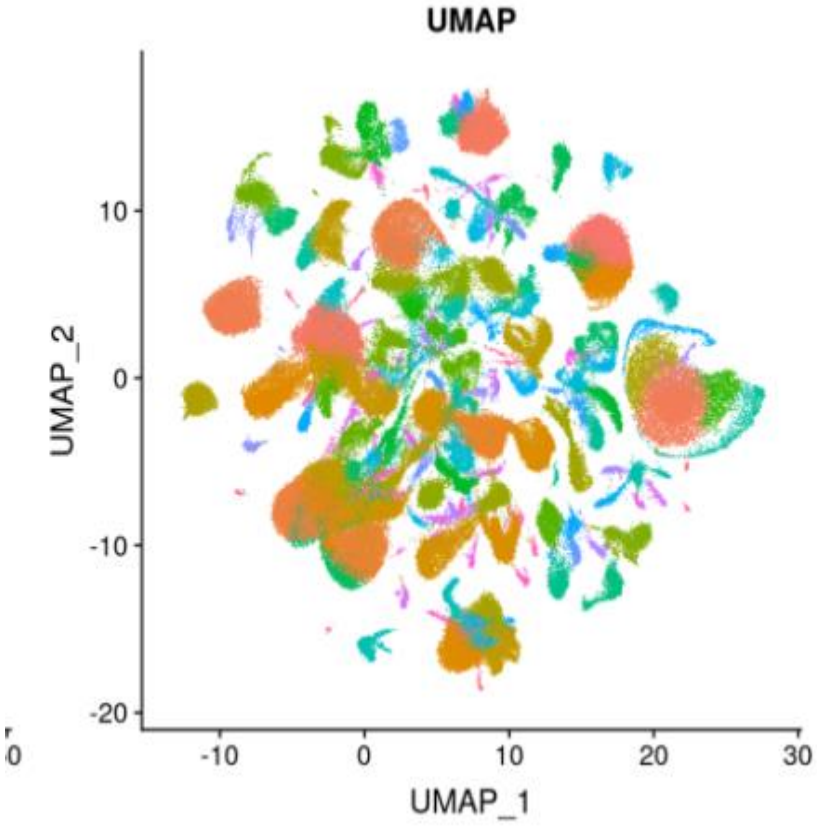

Matthias Scholz, 2015

```
brain <- RunPCA(brain, assay = "SCT", verbose = FALSE)
```

HVGs from SCTransform go into PCA

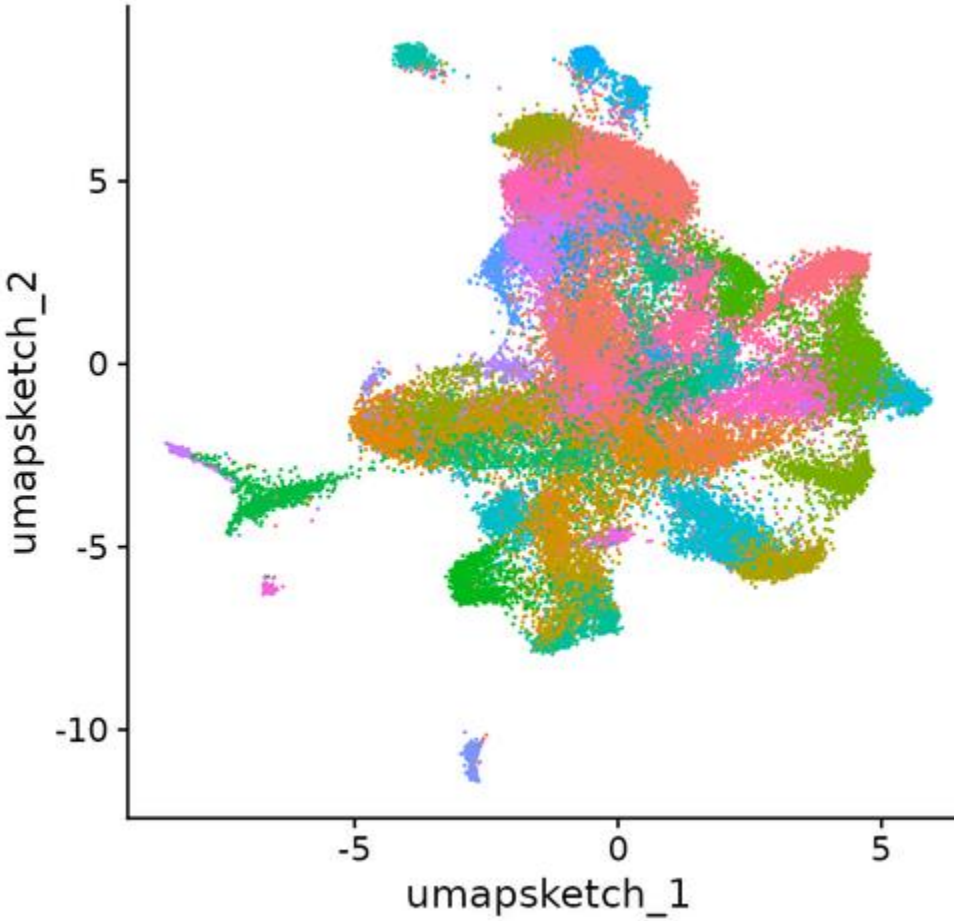
# Clustering

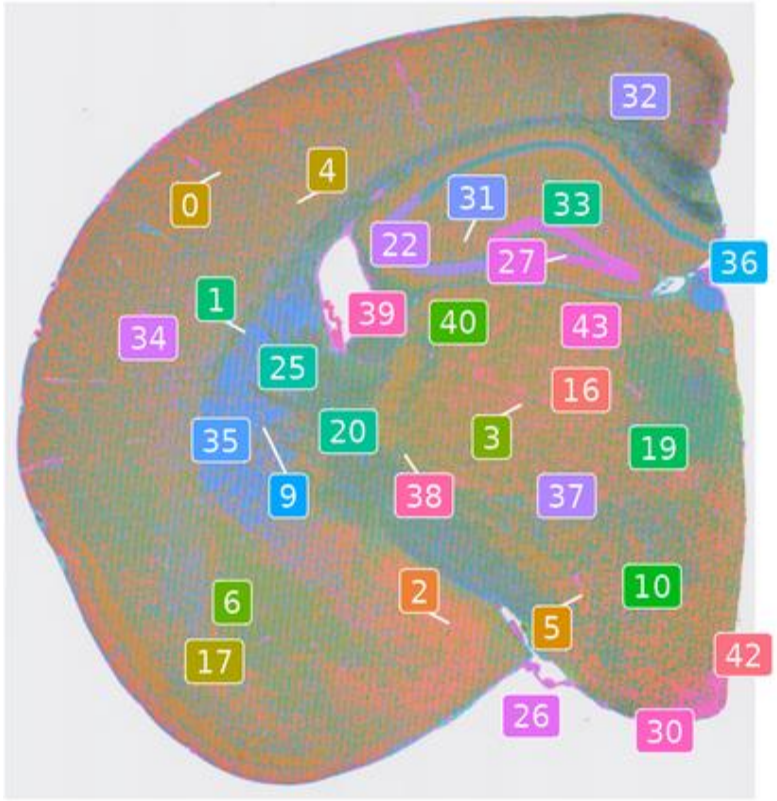
- Distance/similarity metrics
- Algorithm choices
  - K means
  - Hierarchical
  - Density based
  - Graph partition
  - …

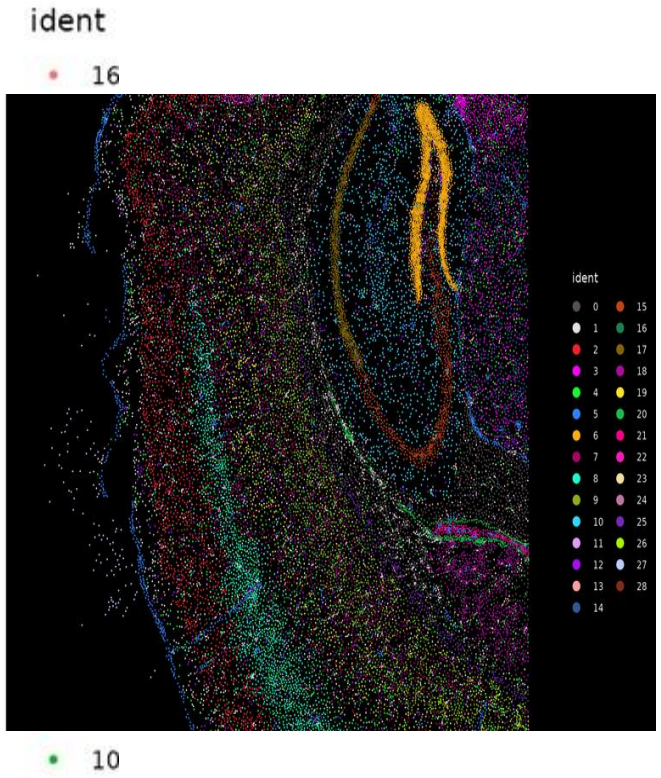



```
brain <- FindNeighbors(brain, reduction = "pca", dims = 1:30)
brain <- FindClusters(brain, verbose = FALSE)
```

# Non Linear Dimension Reduction : UMAP



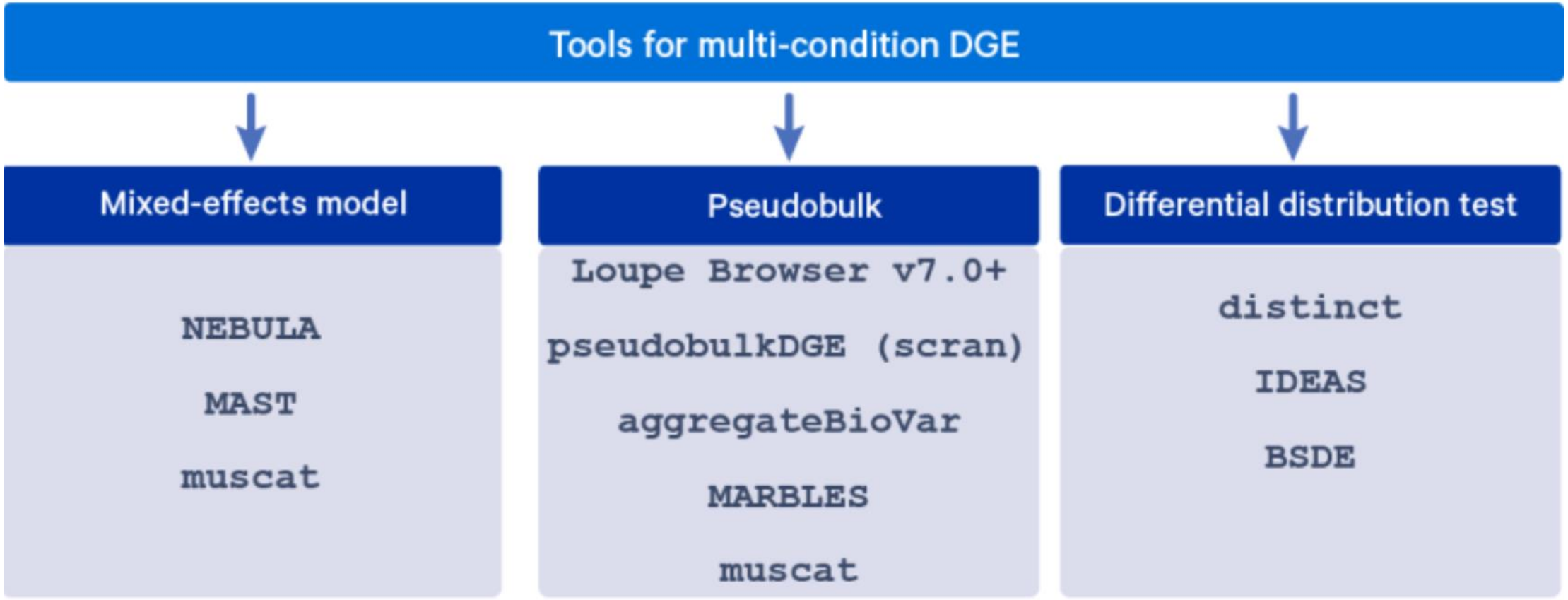




brain <- RunUMAP(brain, reduction = "pca", dims = 1:30)

UMAP is not clustering!

# Differential Gene Expression

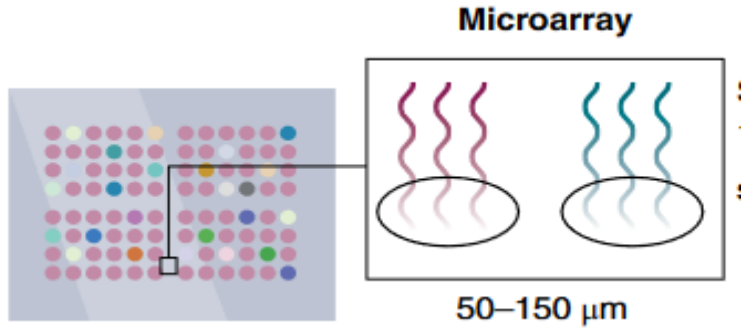


# Sequencing Based ST Data Analysis: Visium

- Cell Type Deconvolution
  - *RCTD*
  - *SpatialDWLS (Giotto)*

- Alignment of scRNA data with SST Data
  - *CytoSpace*
  - *Tangram*
  - *Cell2location*

  - *BayesSpace*

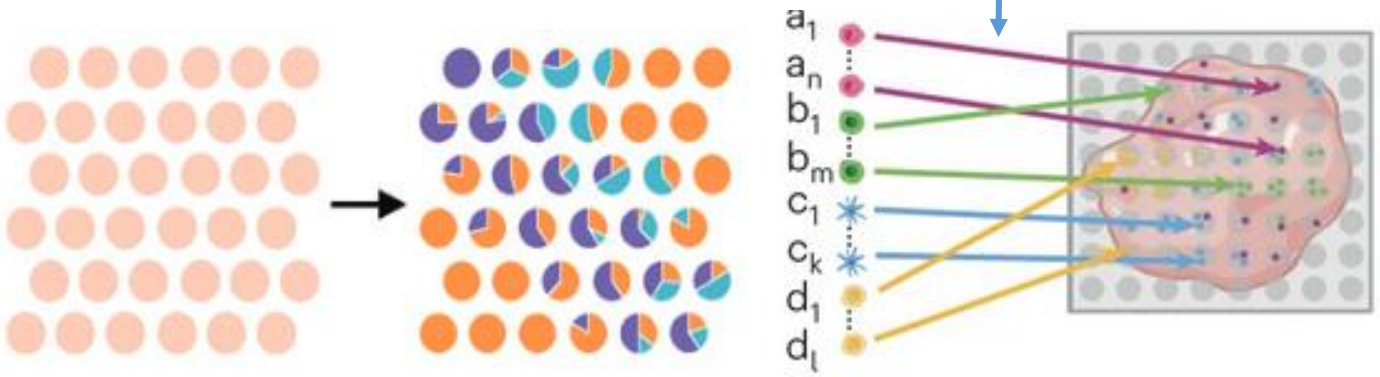



Gene expression matrix

| | Gene 1 | Gene 2 | Gene *N* |
|---|---|---|---|
| Spot 1 | 10 | 5 | 6 |
| Spot 2 | 5 | 21 | 20 |
| ... | ... | ... | ... |
| Spot *N* | 8 | 1 | 2 |

Spatial information

| | *x* | *y* |
|---|---|---|
| Spot 1 | 141.2 | 511.4 |
| Spot 2 | 514.9 | 219.9 |
| ... | ... | ... |
| Spot *N* | 8 | 127.4 |

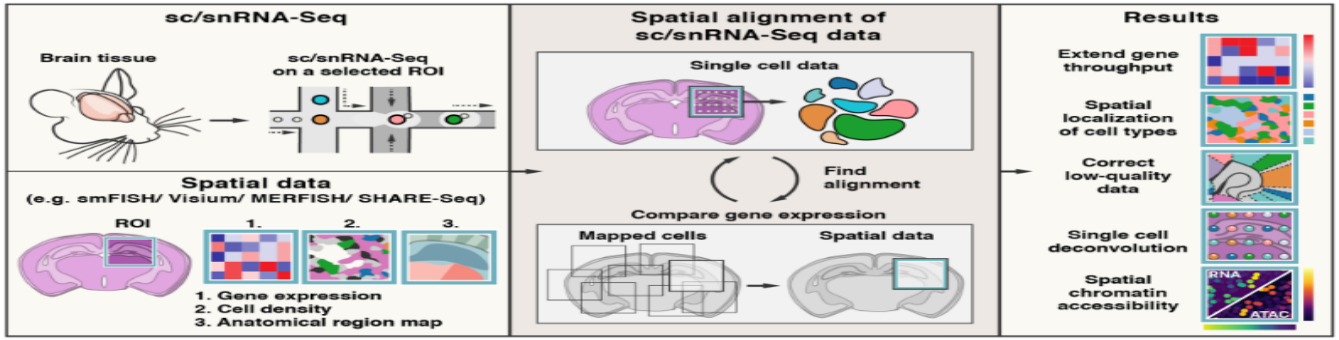


# Integrating scRNA Data with ST Data

## TANGRAM

pypi package 1.0.3

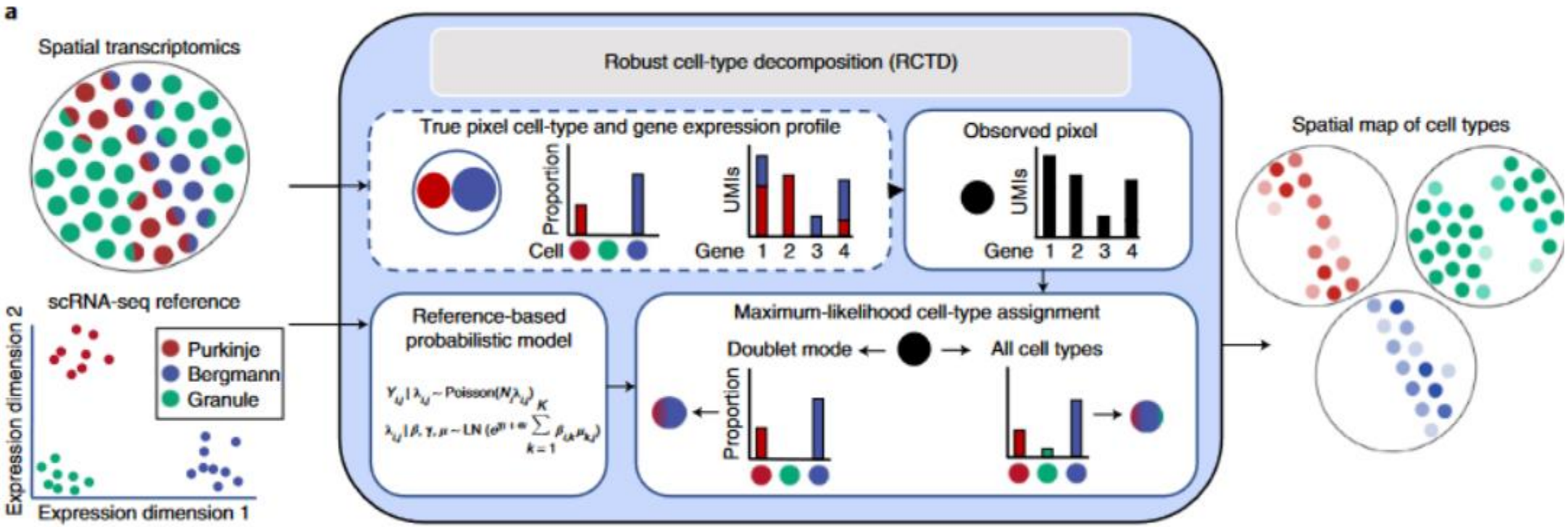
Tangram is a Python package, written in PyTorch and based on scanpy, for mapping single-cell (or single-nucleus) gene expression data onto spatial gene expression data. The single-cell dataset and the spatial dataset should be collected from the same anatomical region/tissue type, ideally from a biological replicate, and need to share a set of genes. Tangram aligns the single-cell data in space by fitting gene expression on the shared genes. The best way to familiarize yourself with Tangram is to check out our tutorial and our documentation. Open in Colab
If you don't use squidpy yet, check out our previous tutorial.




CytoSPACE

scvi-tools

# RCTD : Robust Cell Type Decomposition

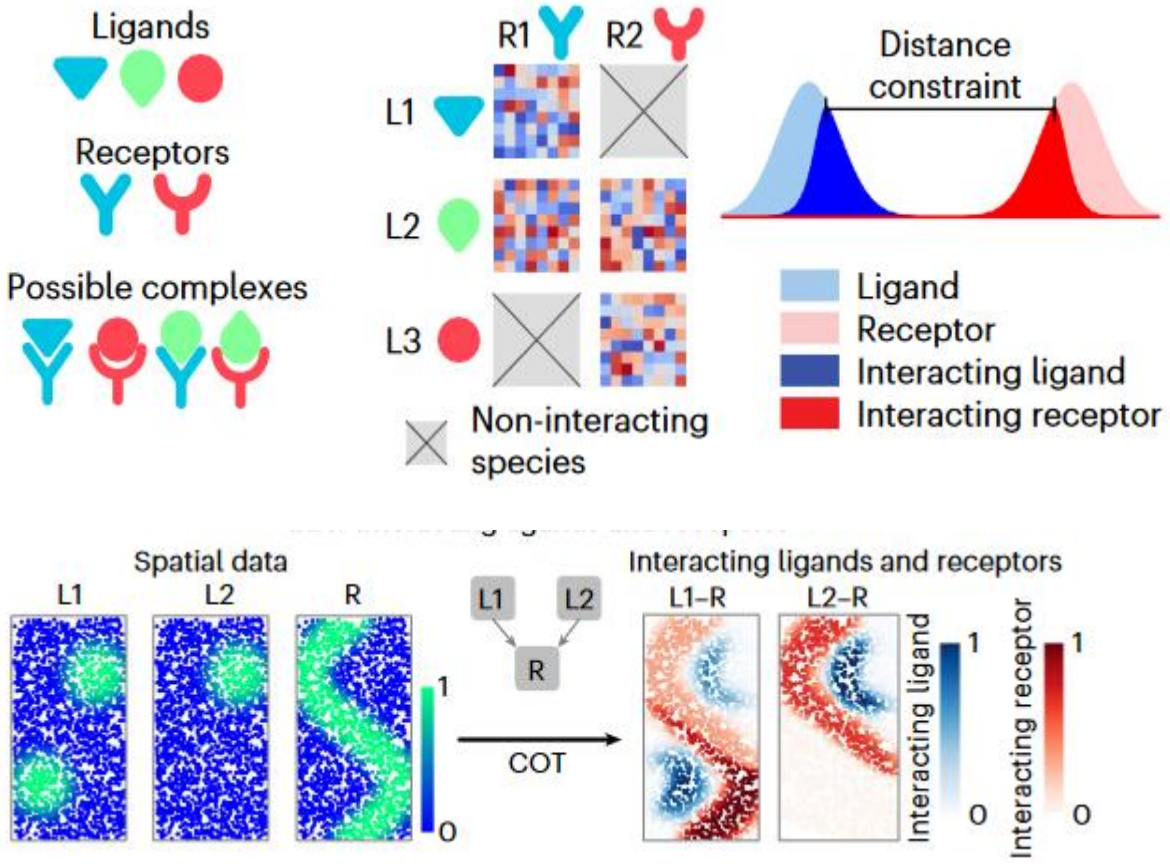



```
# run RCTD with many cores
RCTD <- create.RCTD(query, reference, max_cores = 8)
RCTD <- run.RCTD(RCTD, doublet_mode = "doublet")
```

Cable *et al,* 2022

# Downstream Analysis: Cell-Cell Communication

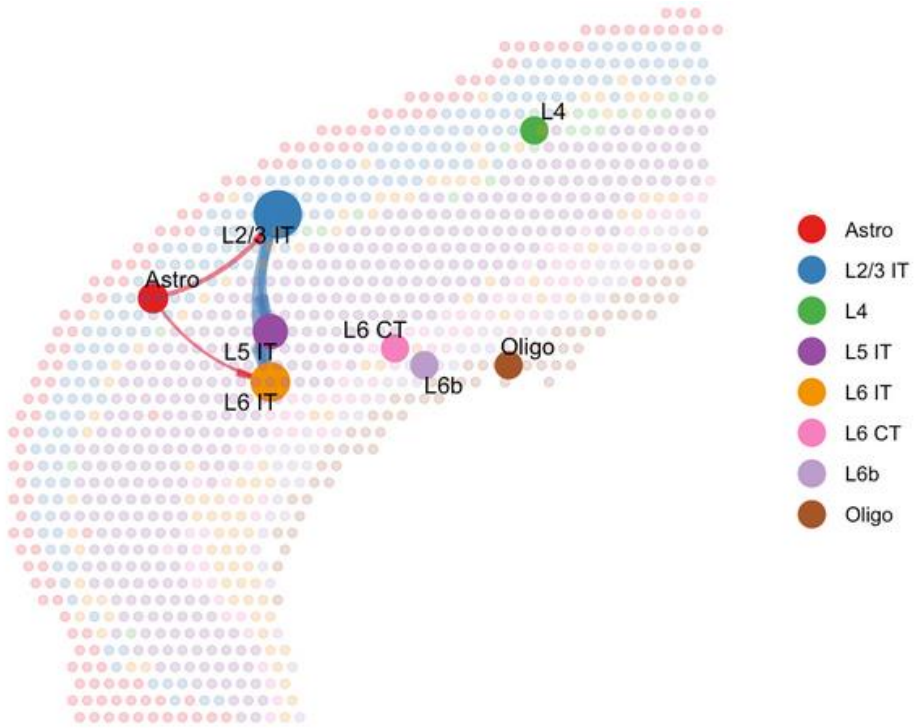
- Spatial CellChat (v2)
- COMMOT: collective optimal transport based
- SpaTalk : knowledge-graph-based cell-cell communication inference




Cang *et al*, 2023

# CellChat v2

- Currently mostly applicable to sequencing based data
- Interaction range set to 250 micron
- Can combine replicates and compare different conditions



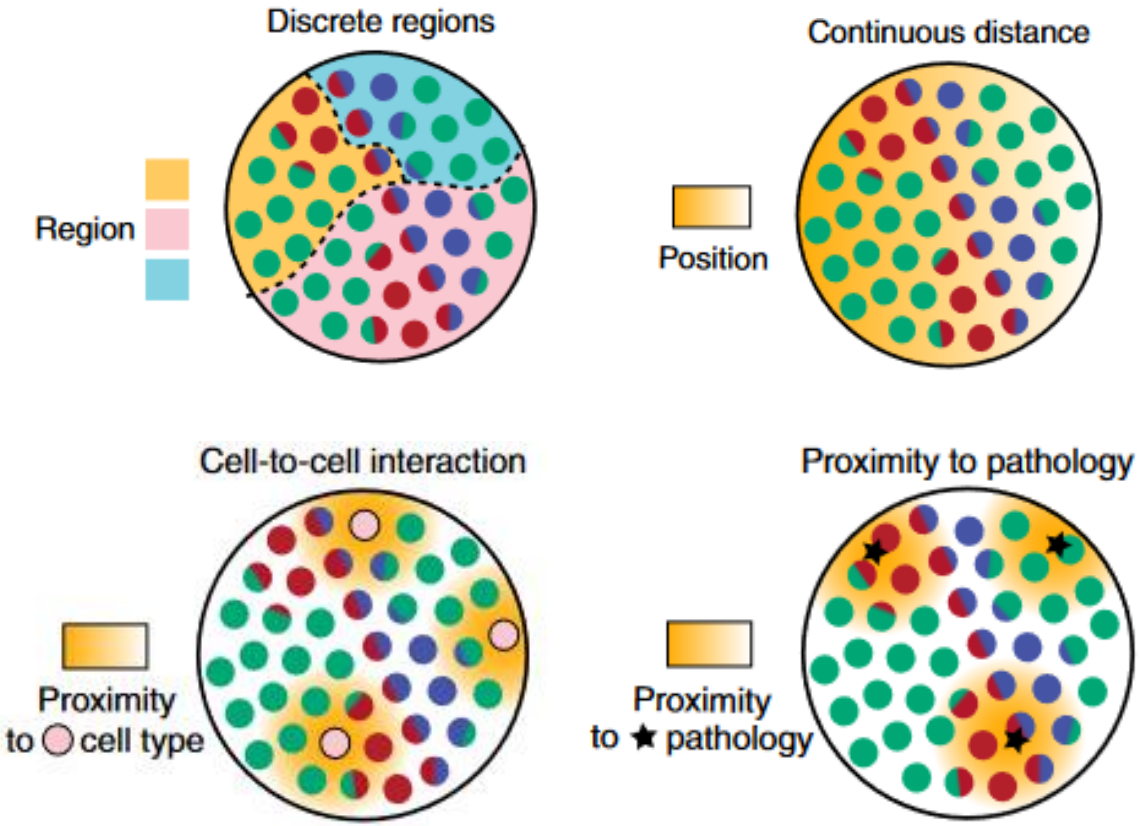

```
>cellchat <- createCellChat(object = data.input, meta = meta, group.by = "labels",  datatype =
"spatial", coordinates = spatial.locs, spatial.factors = spatial.factors)
>cellchat <- identifyOverExpressedGenes(cellchat)
>cellchat <- identifyOverExpressedInteractions(cellchat)
>cellchat <- computeCommunProb(cellchat, type = "truncatedMean", trim = 0.1,   distance.use =
TRUE, interaction.range = 250, contact.dependent = TRUE, contact.range = 100)
```

# Challenges in ST Data Analysis

- Wide range of protocols and data processing pipelines
- A larger variety of file formats and data structures due to heterogeneity of methodologies
- No standardized tissue and quality control measurements or benchmarks
- Increased data volume, run time and memory usage demands more hardware and hands on time
- Computation tools are evolving rapidly

# Downstream Analysis: Spatial DE

- Cell type specific differential gene expression (*C-SIDE*)
  - Takes into account cell types
  - Accounting for localization of other cell types
  - Can incorporate covariates and replicates
  - Can detect DE due to pathology, anatomical regions, cell-cell interactions etc




*Implemented in spacexr package*

Cable *et al*, 2022

# Primary Analysis with 10x Space Ranger