CRI Bioinformatics Bootcamp 2025

Day 6: Introduction to Spatial Transcriptomics

Nataly Naser Al Deen, PhD | Memorial Sloan Kettering Cancer Center
Nick Ceglia, PhD | Memorial Sloan Kettering Cancer Center

THURSDAY, MAY 22, 2025

9 AM–5 PM *SPATIAL TRANSCRIPTOMICS*

Nataly Naser Al Deen, PhD | Memorial Sloan Kettering Cancer Center
Nick Ceglia, PhD | Memorial Sloan Kettering Cancer Center

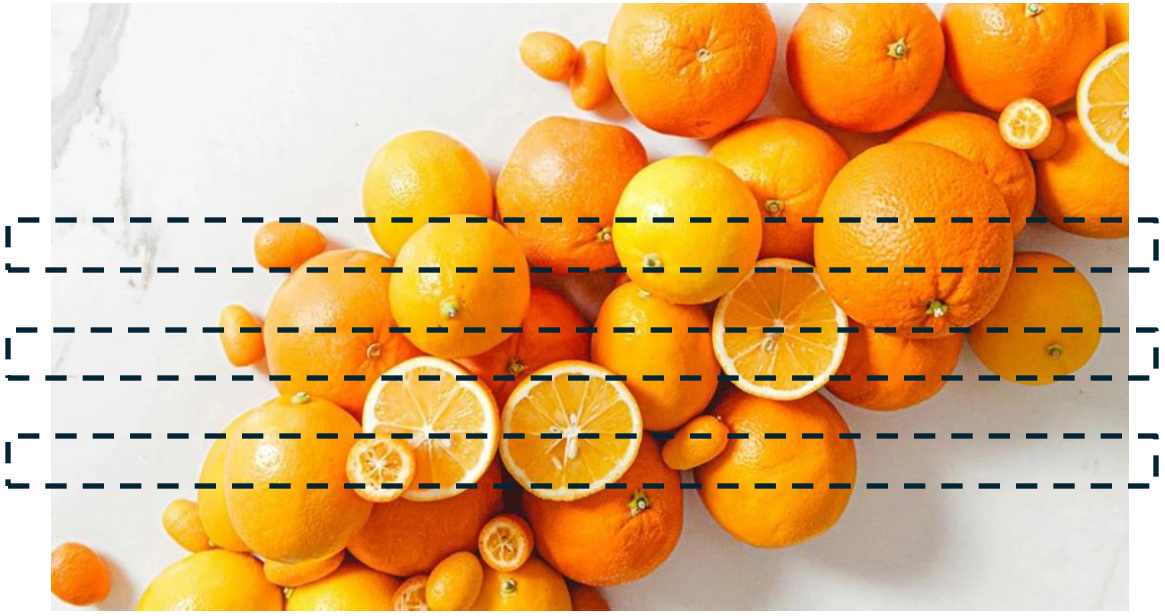
8:00 AM	Breakfast - Bonnet Creek Ballroom Foyer
9:00 AM	Intro to spatial: Visium, Xenium, CosMx
10:00 AM	Sample pre-processing
10:30 AM	Creating a Seurat object, introducing motifs and neighborhood clustering (Banksy/Graffiti)
11:30 AM	Custom binning (provide python code, stardist)
12:00 PM	Office Hour "Working Lunch" - Lunch at Deep Blu Seafood Grille *(in hotel lobby)*
1:30 PM	Cell typing, CNV and trajectory analysis
2:30 PM	Hands-on: unsupervised clustering, cell typing, loupe browser
4:30 PM	Day 6 Summary/Questions
4:45 PM	Closing and Feedback
5:00 PM	Bootcamp Ends

Introduction to spatial transcriptomics: Visium, Xenium, CosMx

Bulk RNA

ScRNA-seq

Spatial Transcriptomics

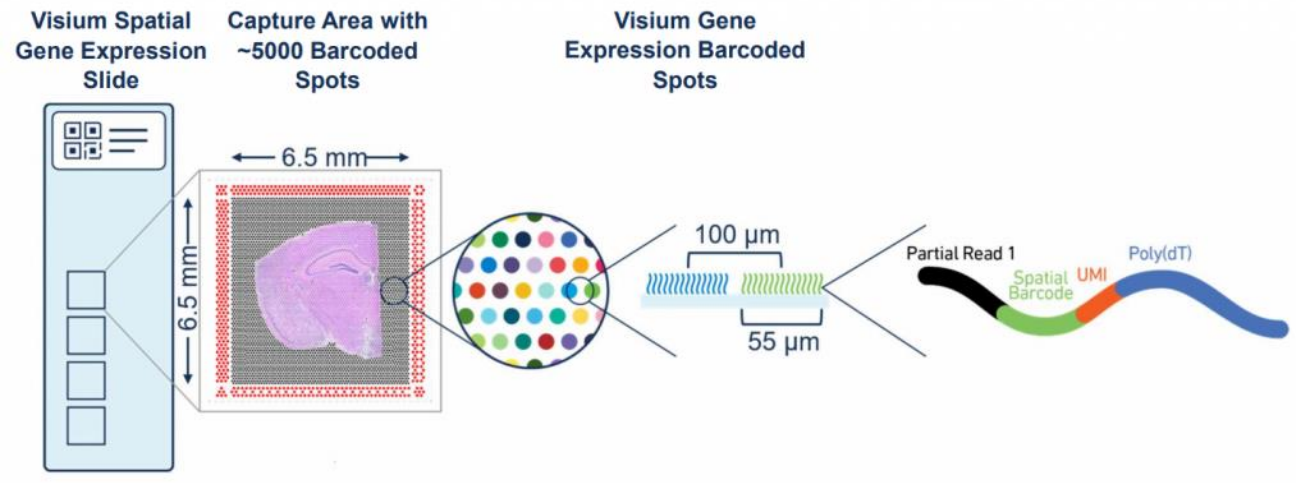


Serial sections will have different cells/parts of cells based on the snapshot in time of the bulk tissue

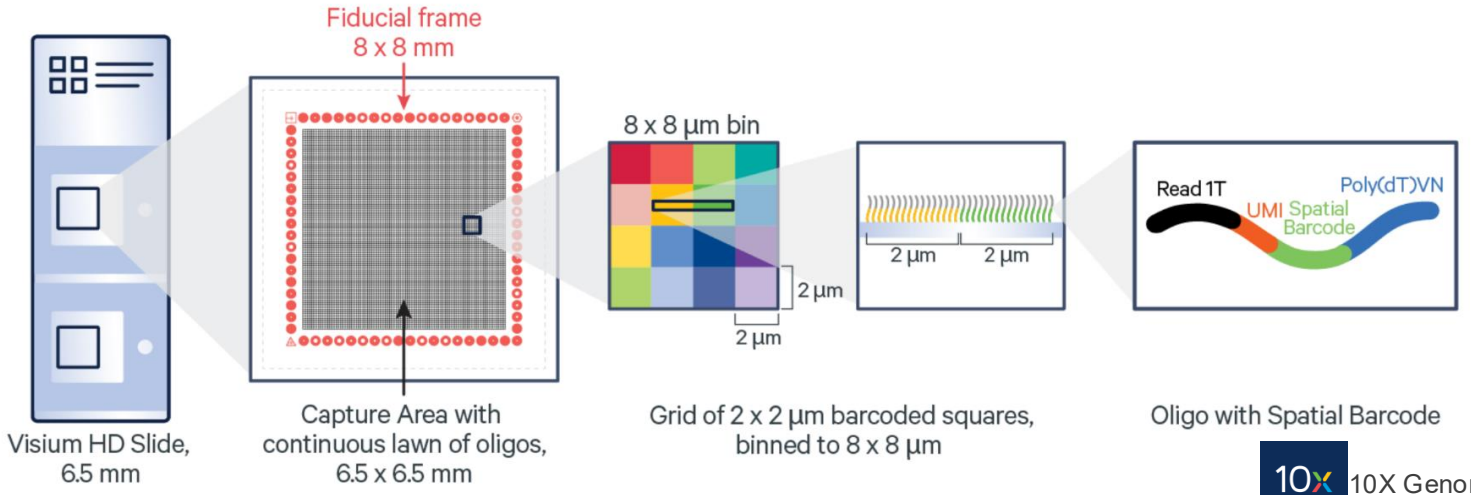
Visium HD

- Material (FFPE 5µm/OCT 10 µm)- Archived H&E/IF, USS, Block)
- Fiducial frame size/chemistry (6.5x6.5mm, 2x2µm bin)
- Sample QC
- Recommended slides for anti-detachment
- Experimental procedure (CytAssist)
- Library Chemistry (FFPE probe based /OCT cDNA synthesis)
- Data delivery (FATSQ and high res H&E/IF image) - needs SpaceRanger
- Loupe Browser
- Discovery technique
- No protein panel, but allows up to ≈3 mpIF on the same section prior to CytAssist run

Classical Visium enabled 18,000 gene probe-based sequencing at a 55µm diameter (1-10 cell resolution/ spot) with spots missing tissue every 55µm, covering a maximum of 5000 spots per 6.5x6.5 mm fiducial frame (also available at 11x11 mm)

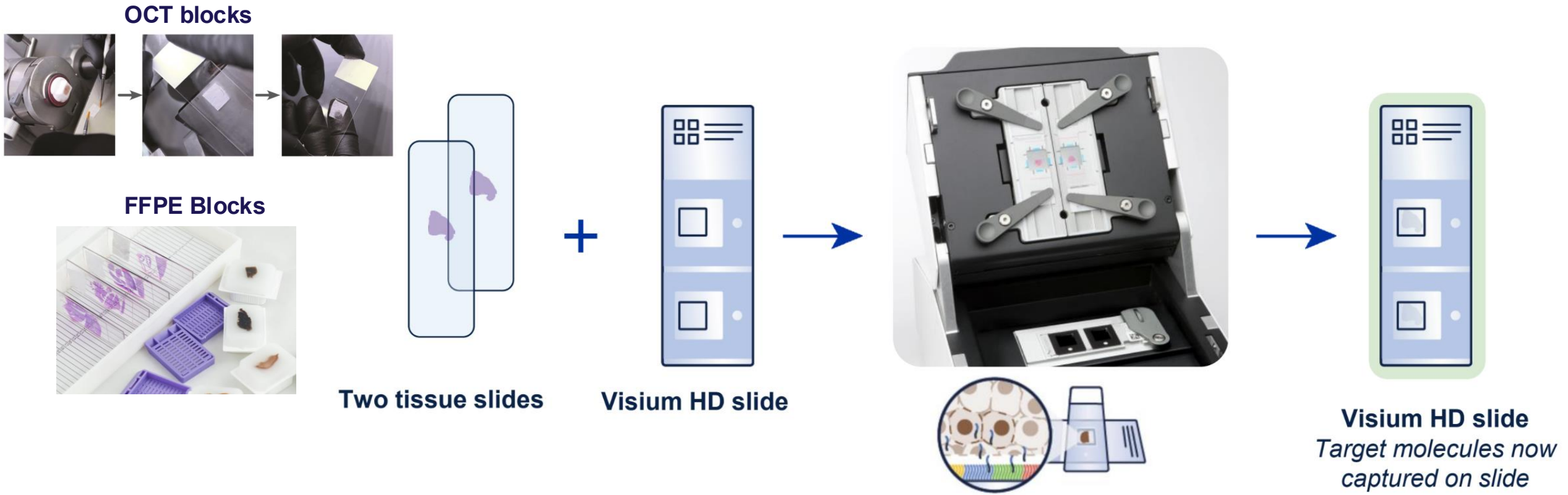


Visium HD slides contain two 6.5 x 6.5 mm Capture Areas with a continuous lawn of oligonucleotides arrayed in millions of 2x2µm barcoded squares without gaps, achieving single cell–scale spatial resolution. The data is output at 2 µm, as well as multiple bin sizes. The 8x8 µm bin is the recommended starting point for visualization and analysis



10X Genomics blog, Olivia Habern, April 2024

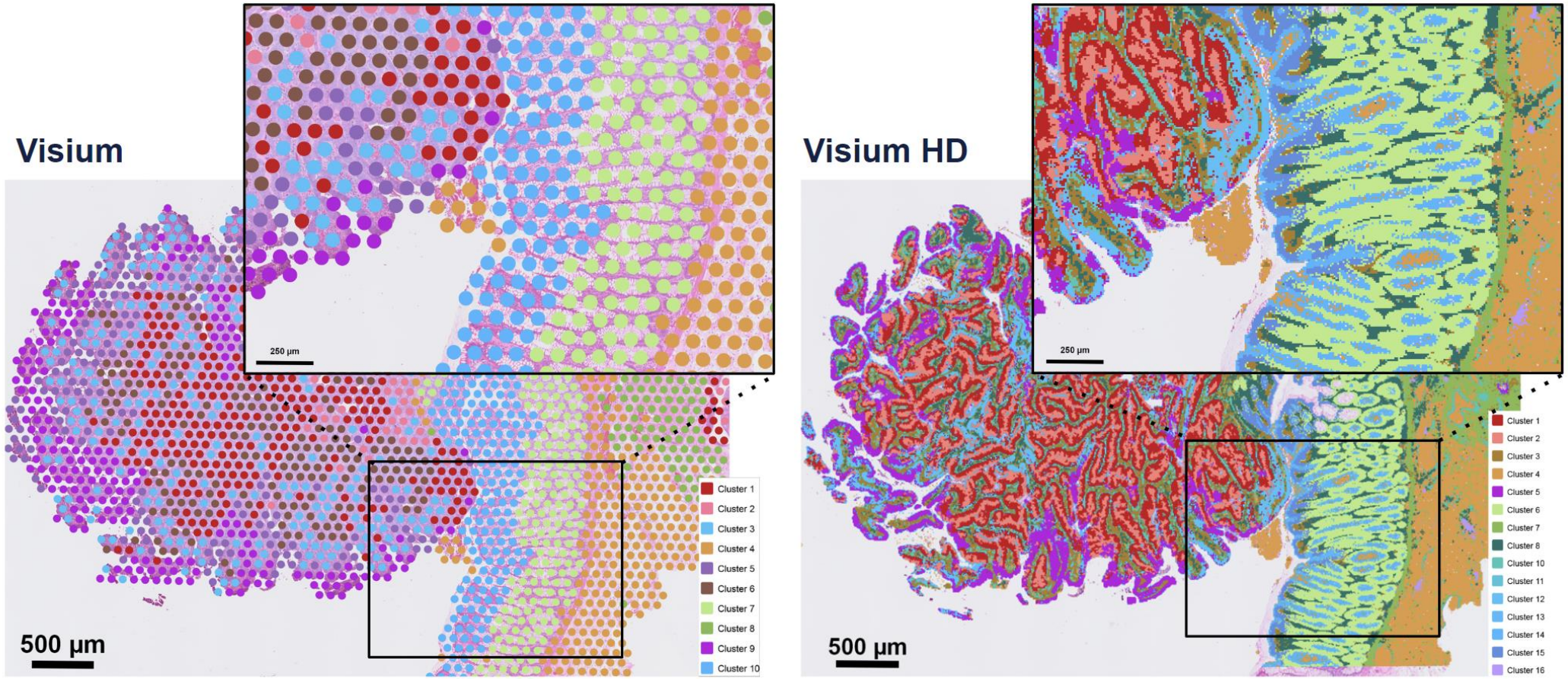
Researchers can start the workflow from **fresh frozen**, **fixed frozen**, or **FFPE** tissue blocks and freshly section onto glass slides, or from **pre-sectioned** tissue slides. After the H&E or IF staining steps, tissues are treated with whole transcriptome probe panels, enabling hybridization and ligation of probe pairs to their targets.



Tissue slides and Visium HD slides are loaded into the Visium CytAssist instrument, where they are brought into proximity with one another. Gene expression probes are released from the tissue, enabling capture by the spatially barcoded oligonucleotides present in a hydrogel on the Visium slide surface.

10X Genomics blog, Olivia Habern, April 2024

A comparison of Visium v2 data (left) and Visium HD data (right) in FFPE human colorectal cancer, demonstrating the enhanced discovery power of whole transcriptome spatial gene expression at single cell–scale resolution.



10X Genomics blog, Olivia Habern, April 2024

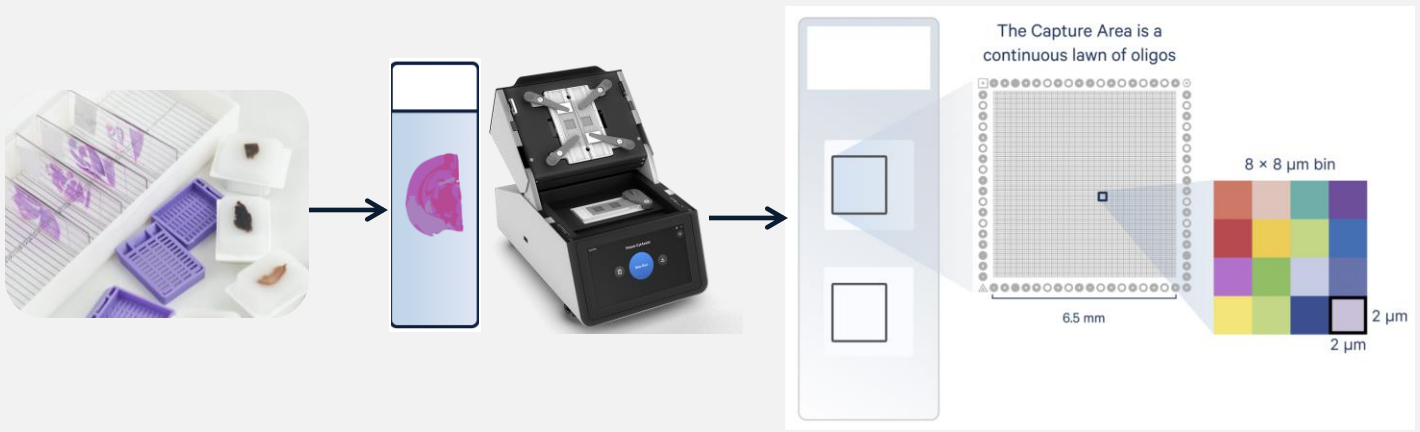
Visium Spatial Transcriptomics HD enables **18,000 whole transcriptome** on 5µm sections of **FFPE blocks** or **archived (un)stained slides** with DV200>30% at a (2µmx2µm) resolution

CRI Cancer Research Institute™

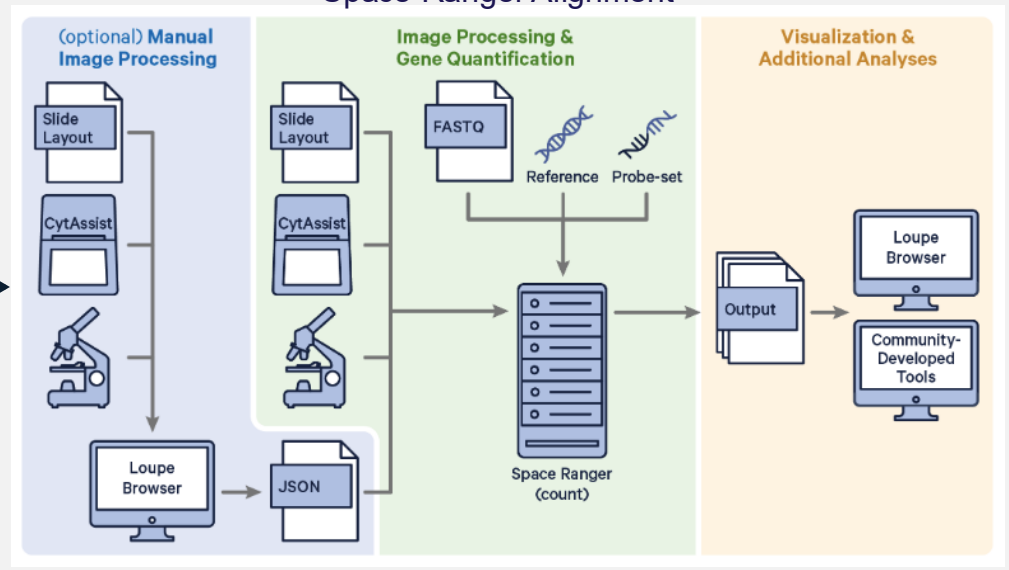
Input: 5µm FFPE/OCT tissue (6.5x6.5mm)

Output: high resolution H&E (brightfield) or IF (Fluorescent) image (≈1GB) and FASTQ files

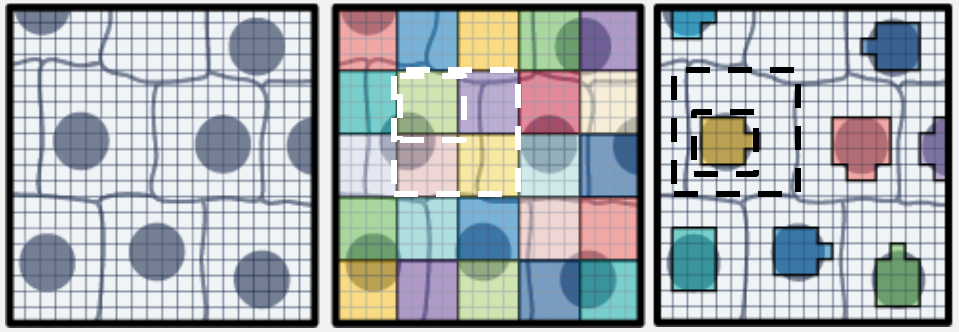
A Visium HD experimental planning



B Space Ranger Alignment

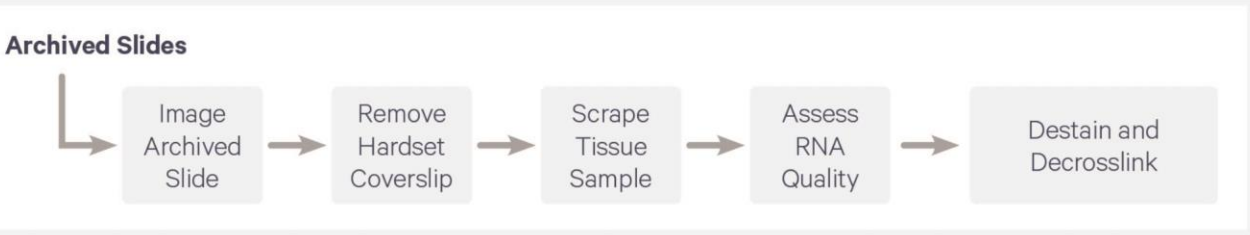


C Nuclei segmentation and custom binning



D Third party tool analysis

Using an archived H&E/IF stained slide for the Visium HD QC and CytAssist run



a. Cool a metal block on dry ice for **5–10 min**.

b. Gently immerse archived slide in Xylene Jar 1. Secure the jar cap to prevent xylene loss.

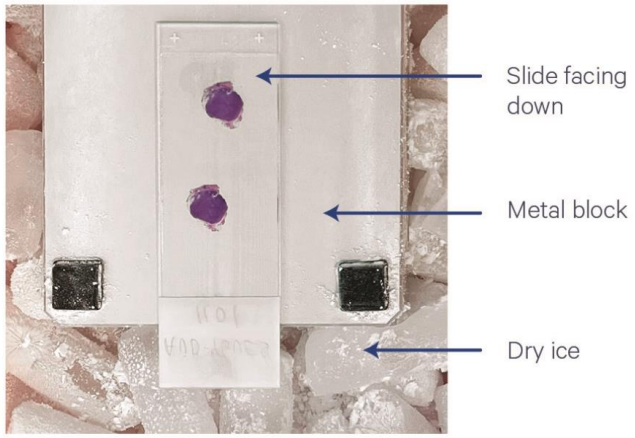
c. Incubate for **5 min**.

d. Remove excess xylene from archived slide with a lint-free laboratory wipe.

e. Place on pre-cooled metal block with the coverslipped tissue sections facing down.

f. Wait **1 min**.

Archived Slide on Metal Block

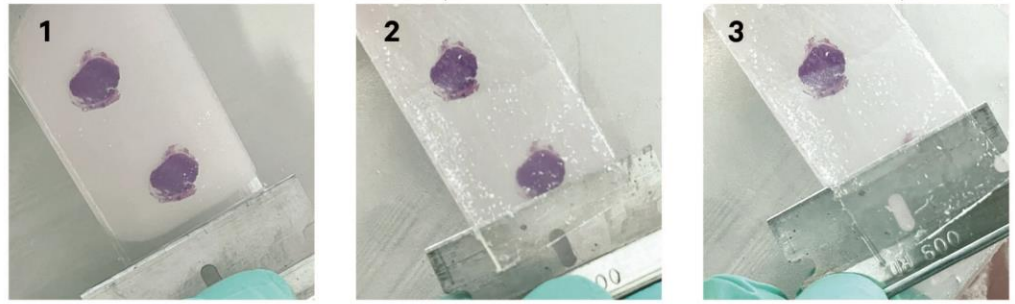


Removing Coverslip with Razor Blade

Align edge of razor blade to right most edge of coverslip

Slowly wedge razor blade between slide and coverslip until coverslip releases from slide

Holding razor blade towards coverslip, slowly work blade to the left to lift coverslip off slide

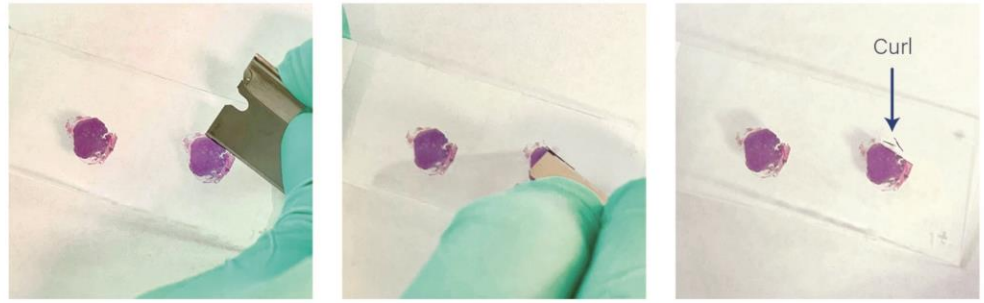


Scrape Tissue for RNA Quality Assessment

Scrape small portion of tissue at edge in one motion *(tissue will curl on itself)*

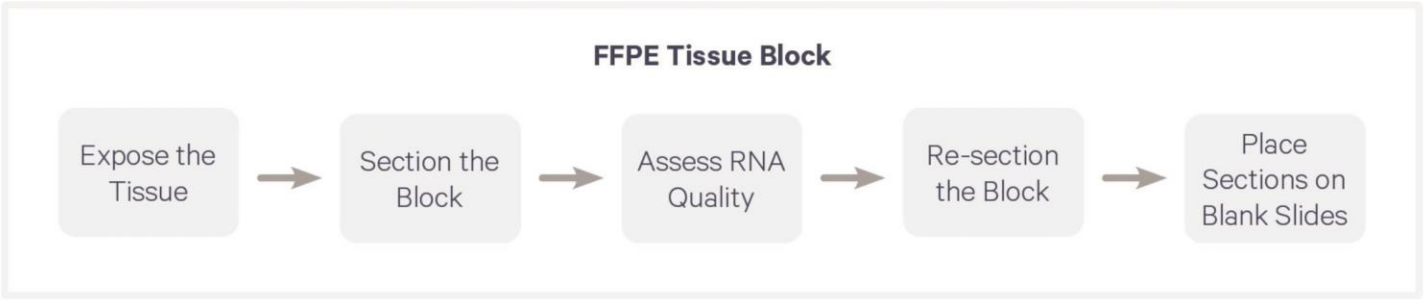
Cut off small portion of tissue from opposite side

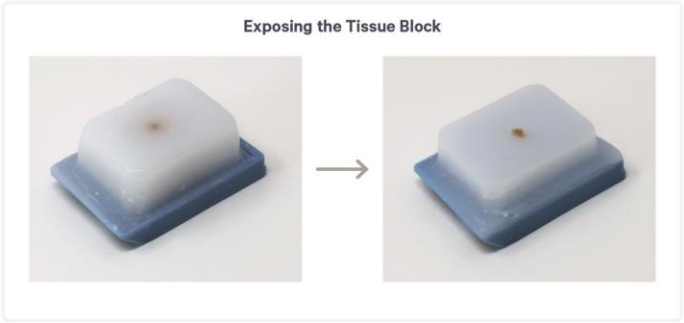
Use small curl of tissue for RNA quality assessment

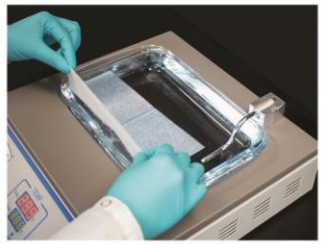


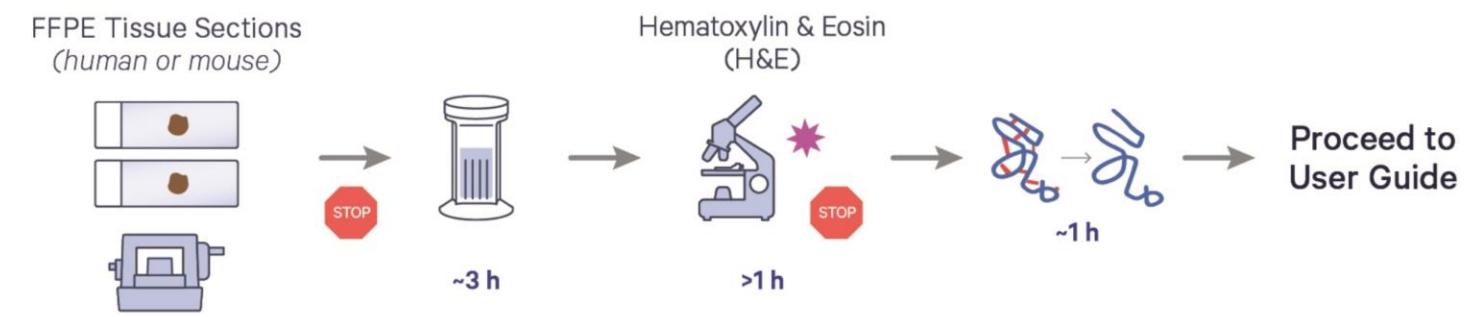
The remaining section on the archived slide may be used for the Visium HD workflow. If necessary, store the slide in a sealed slide mailer in a desiccator kept in the dark at 4°C for up to two hours to allow time for RNA evaluation.

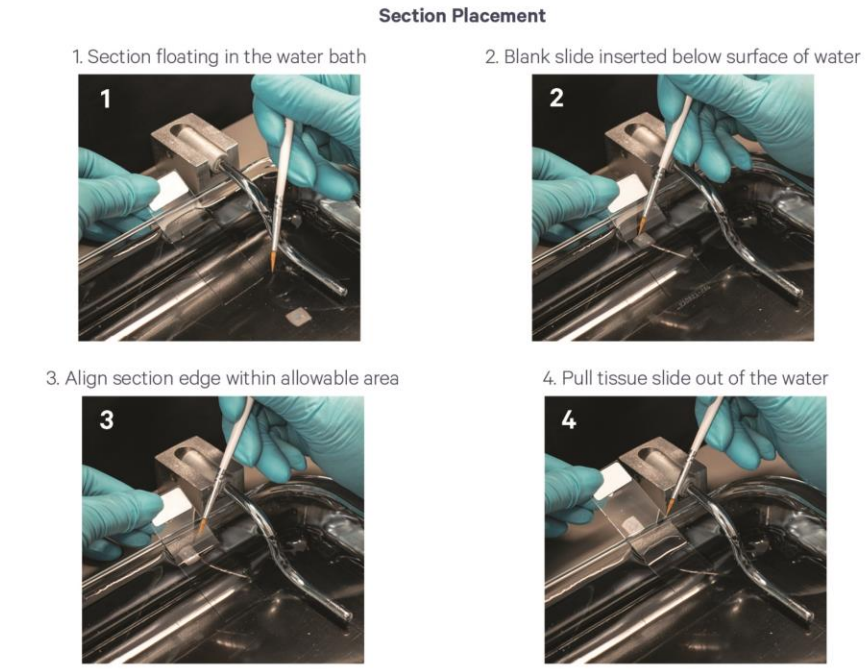
Using an FFPE tissue block for the Visium HD QC and CytAssist run



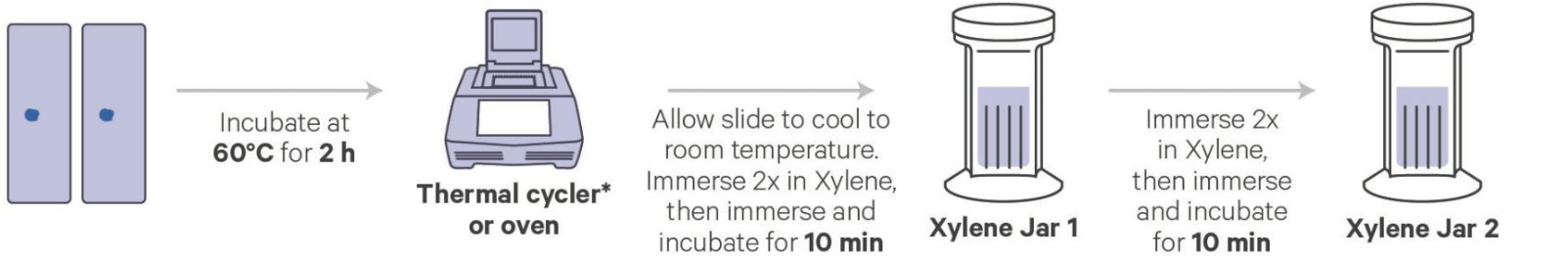


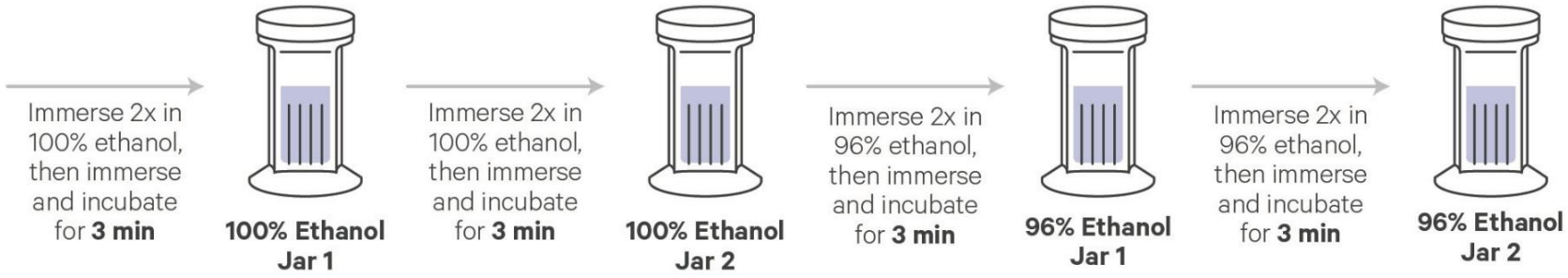


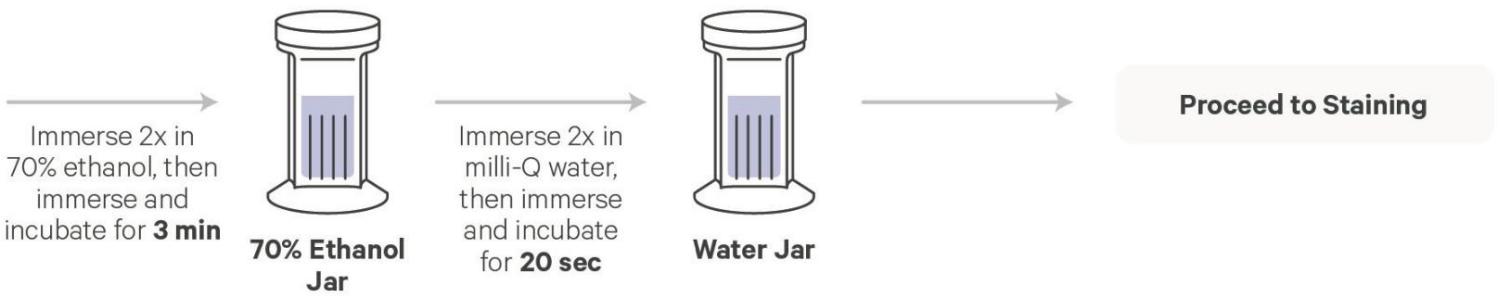




Deparaffinization ~3 h



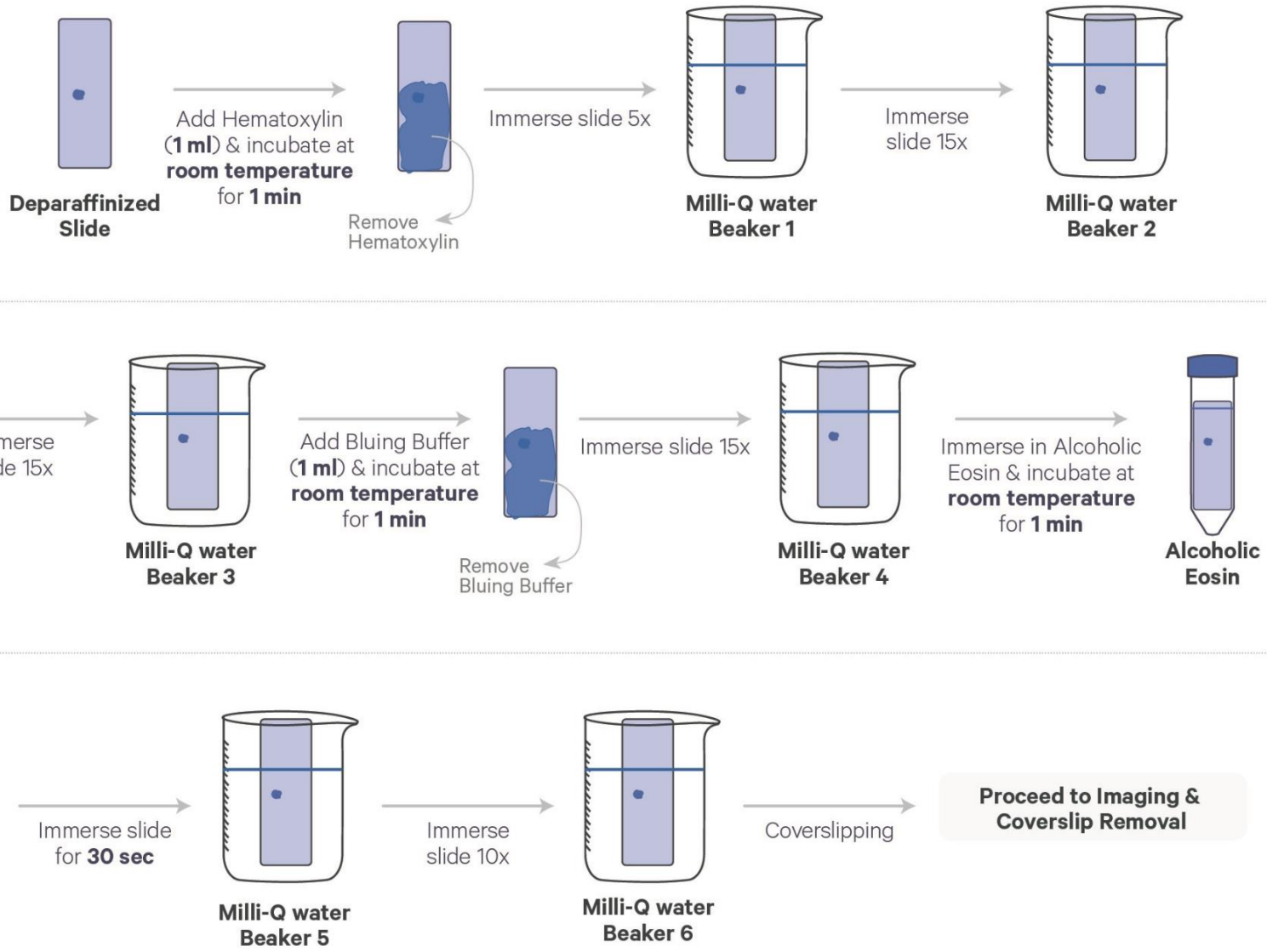




Slide in image is representative

**Keep the thermal cycler lid open during incubation*

H&E Staining & Coverslipping ~15 min



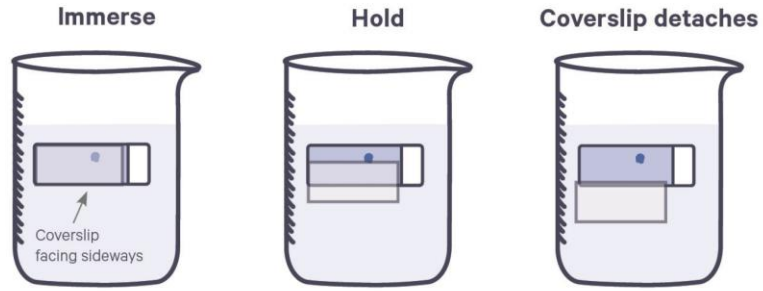
Wipe excess liquid from the back of the slide after each immersion series

CRI Cancer Research Institute™

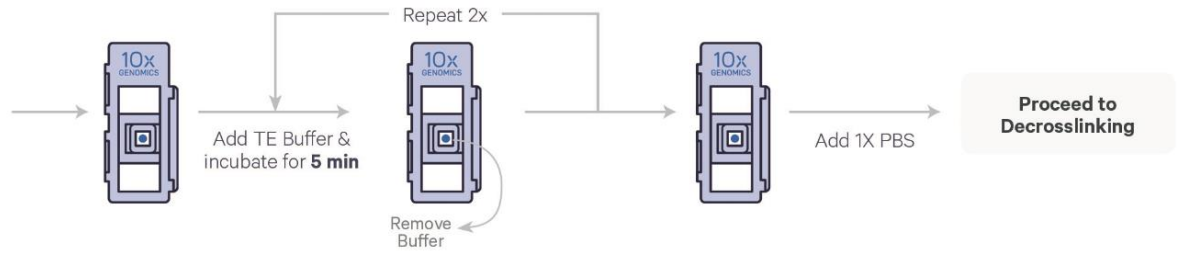
Cover slip removal:

After high resolution microscopy image is acquires

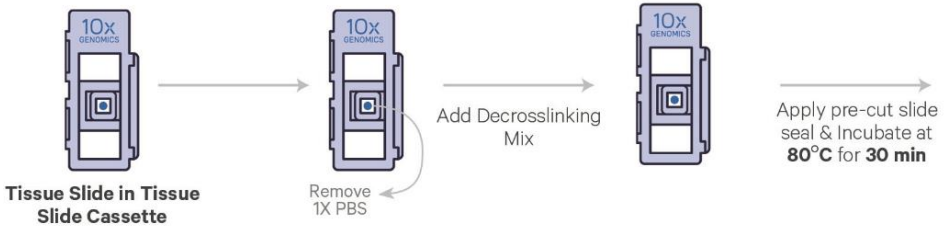
Prior to de-staining and decrosslinking

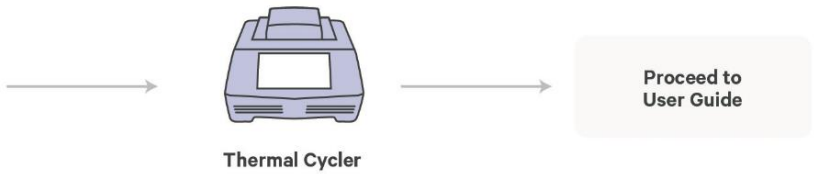


Destaining ~30 min



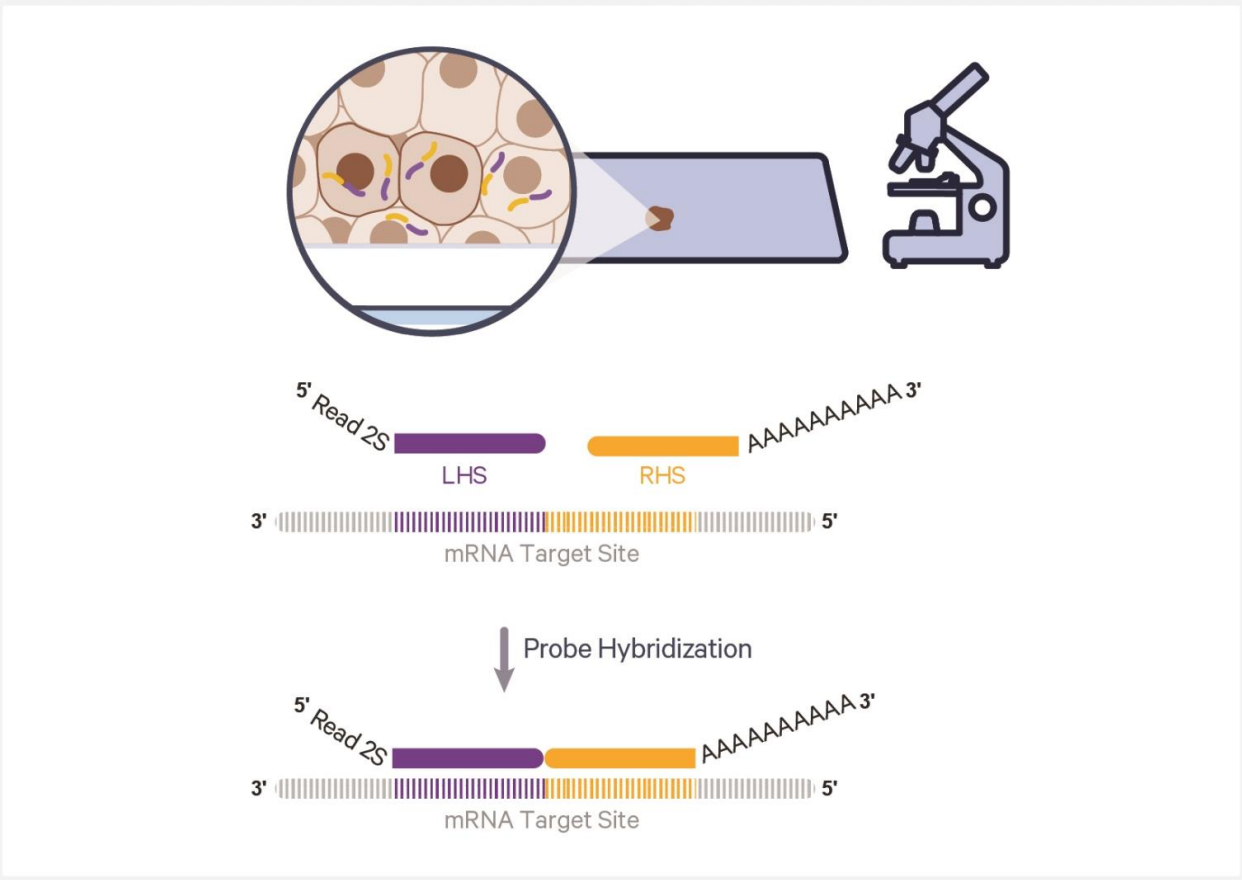
Decrosslinking ~30 min





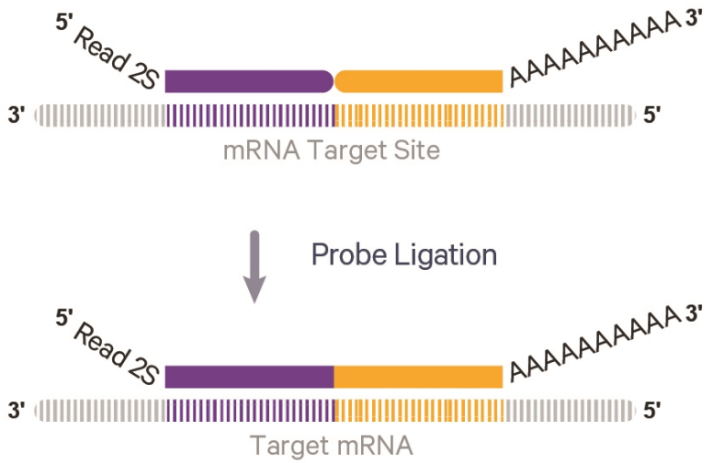
Step 1: Probe Hybridization

The human or mouse whole transcriptome probe panel, consisting of ~3 specific probes for each targeted gene, is added to the deparaffinized, destained, and decrosslinked tissues. Together, probe pairs hybridize to their complimentary target RNA.



Probe Ligation

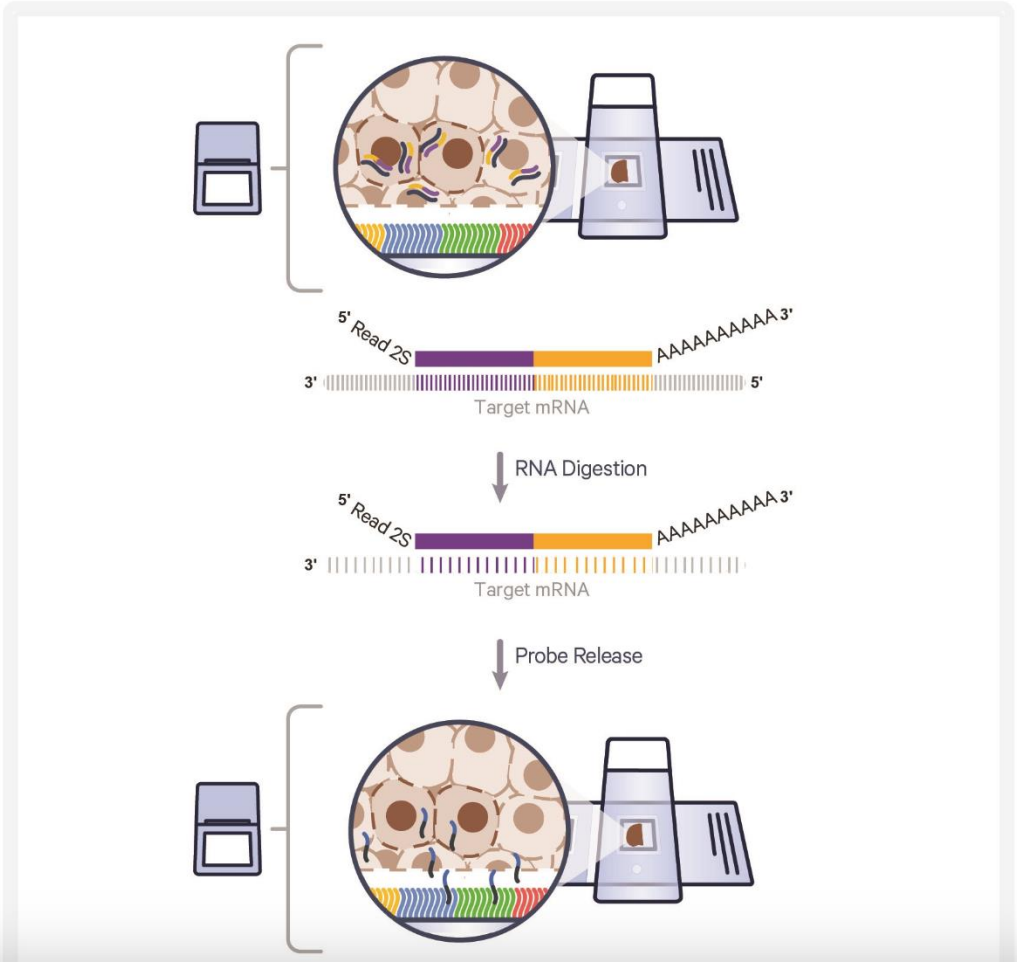
After hybridization, a ligase is added to bridge the junction between the probe pairs that have hybridized to RNA, forming a ligation product.

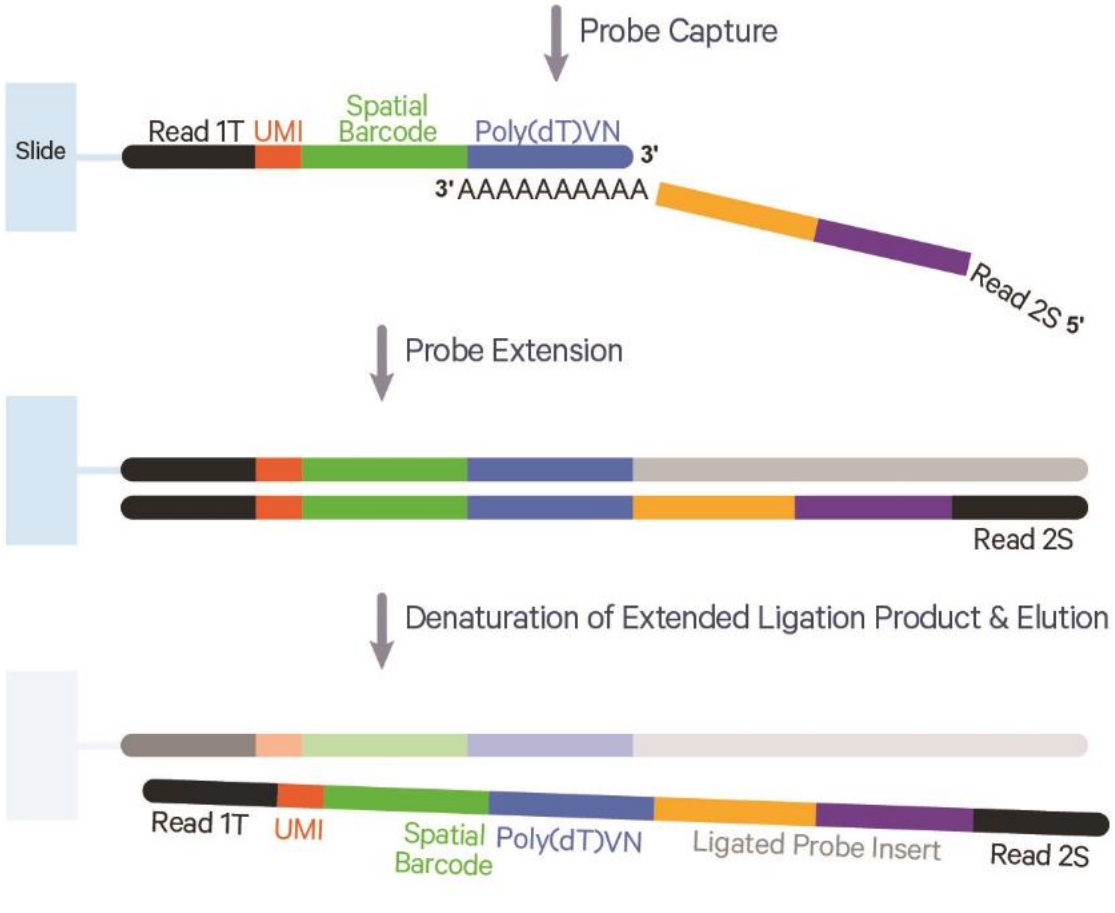


Probe Release & Extension

Probe release and capture occurs in the Visium CytAssist instrument. The single stranded ligation products are released from the tissue upon RNase treatment and captured on the Visium slide. Once ligation products are captured, the slides can be removed from the instrument.

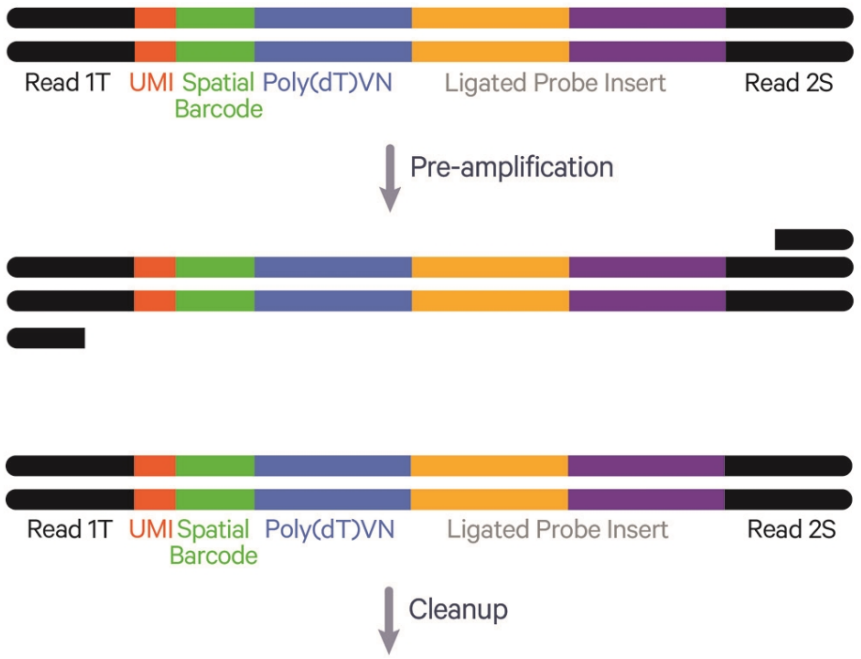
Ligation products are extended by the addition of the Spatial Barcode, UMI, and partial Read 1 primer. This generates spatially-barcoded ligation products, which can then be carried forward for library preparation.





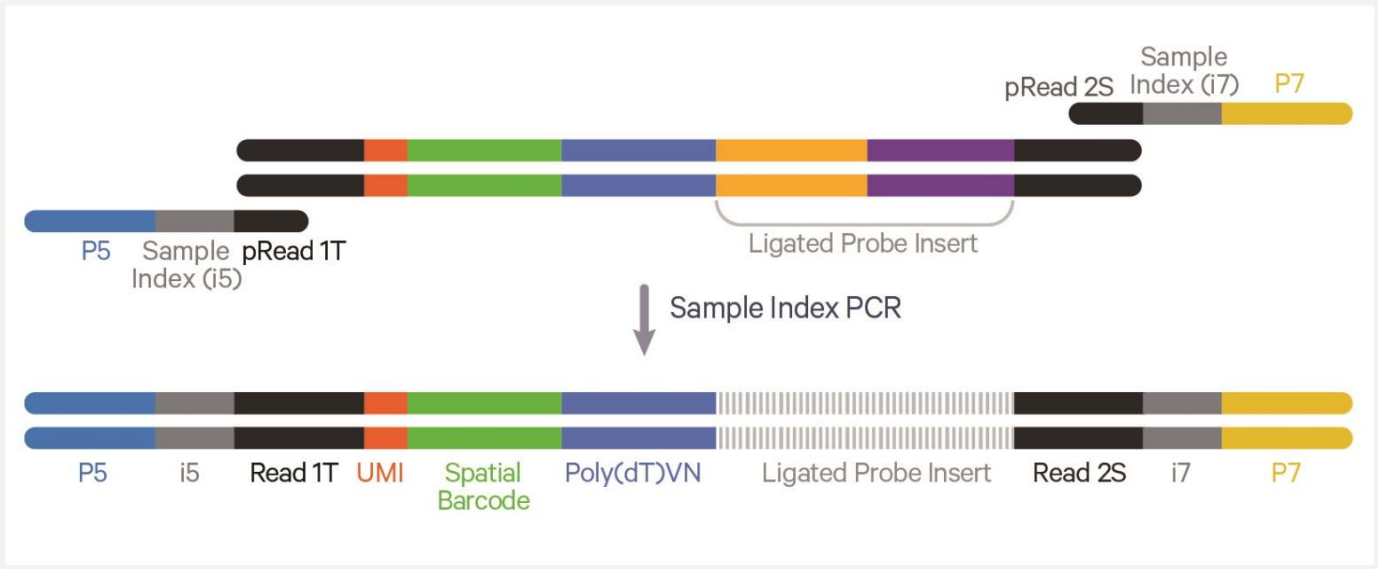
Pre-amplification and SPRIselect

To generate ample material for library construction, barcoded ligation products are amplified. This pre-amplification is followed by SPRIselect cleanup.



Step 6: Visium HD Spatial Gene Expression - Probe-based Library Construction

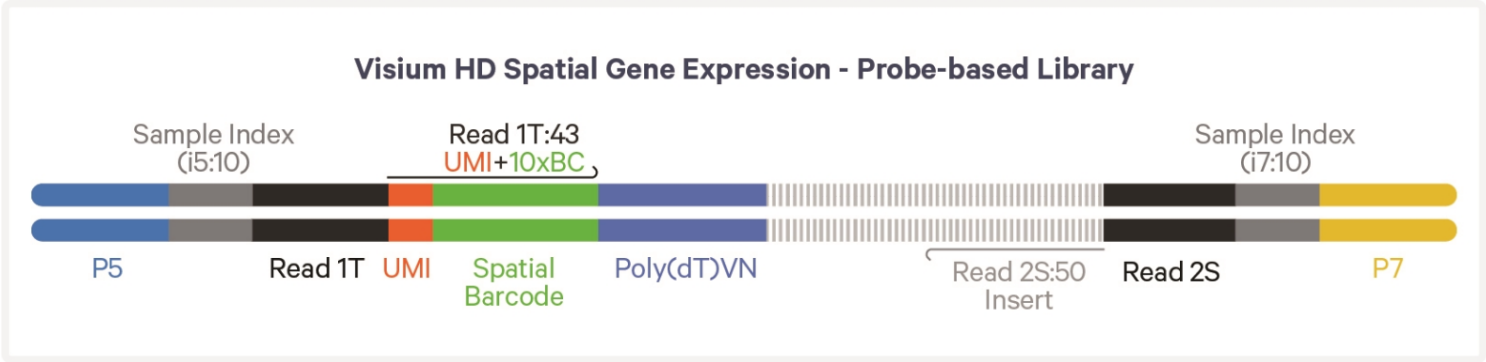
Pre-amplification material is collected for qPCR to determine sample index PCR cycle number for gene expression libraries. The amplified material then undergoes indexing via sample index PCR to generate final library molecules. The final libraries are cleaned up by SPRIselect, assessed on a Bioanalyzer or a similar instrument, quantified by qPCR, and then sequenced.



Sequencing

A Visium HD Spatial Gene Expression - Probe-based library comprises standard Illumina paired-end constructs, which begin and end with P5 and P7 adaptors. The 43 bp UMI and Spatial Barcode are encoded in TruSeq Read 1, while Small RNA Read 2 (Read 2S) is used to sequence the ligated probe product.

Illumina sequencer compatibility, sample indices, and library loading information are summarized in the Sequencing section.

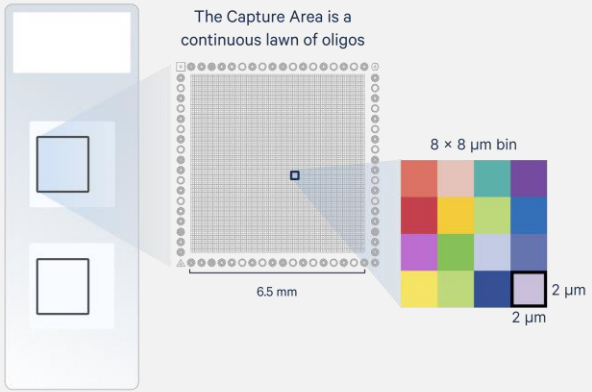


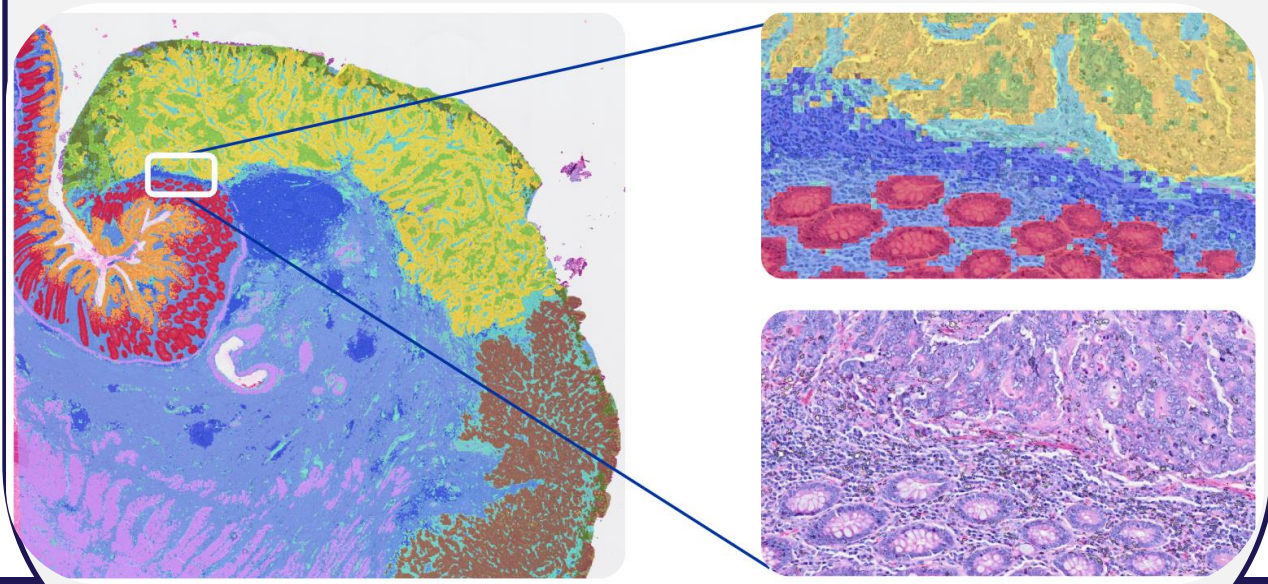
Xenium

- Material (freshly cut block on the Xenium Slide)
- Can add several samples per slide
- Non-destructive protocol, can perform H&E or IF or even Visium HD after the protocol is done
- Slide dimensions (22.5x10.5mm)
- Machine cycles, image-based sequencing
- Xenium Explorer, performs cell segmentation, and delivers pre-processed data
- Single cell/subcellular resolution (50nm)
- No protein panel
- Fully customizable panel (very expresive) and off the shelf panels for specific tissue/cancer types and a pan-cancer panel (500 more sensitive, 5000 available but less recommended)
- Validation technique

Visium HD

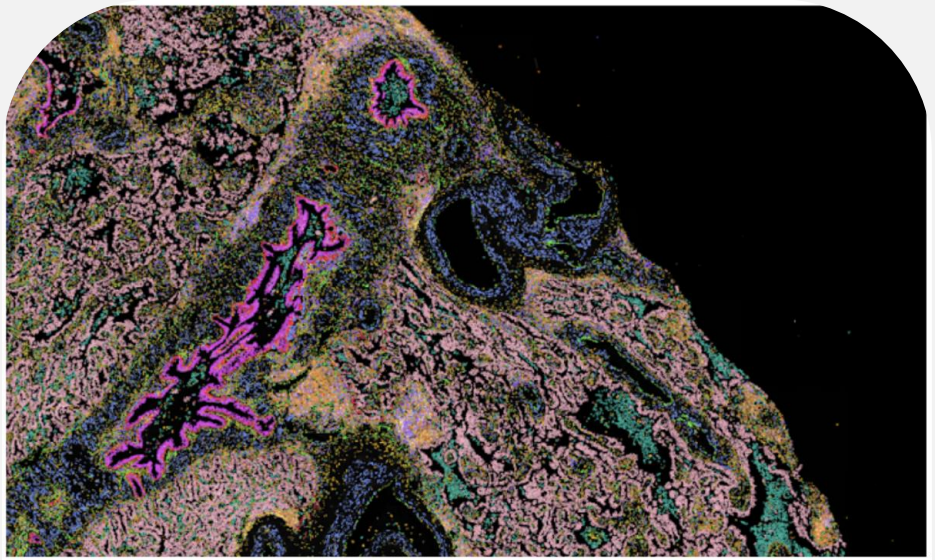
Visium HD on human colon cancer with unsupervised gene expression clustering and zoomed-in samples of clustered gene expression in 8 x 8 µm bins

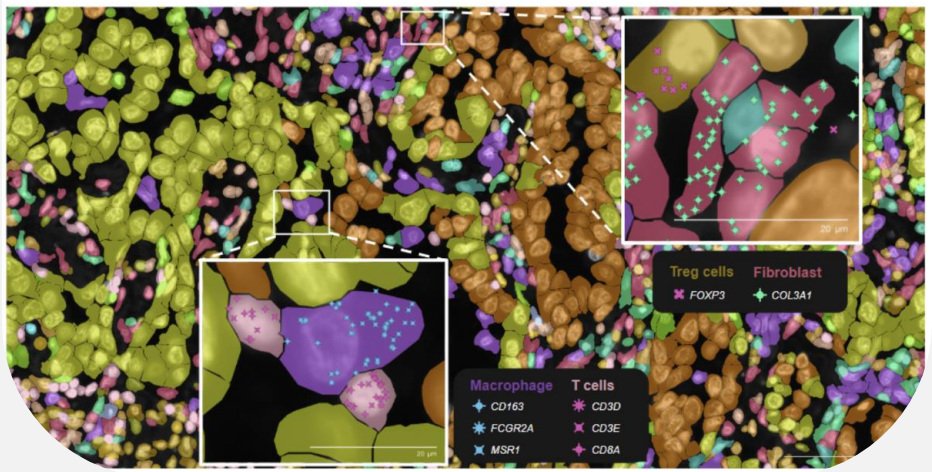




Xenium

Xenium Explorer analysis of human lung cancer with segmented cells and selected transcripts associated with immune cell types localized within individual cells.





Visium HD vs Xenium

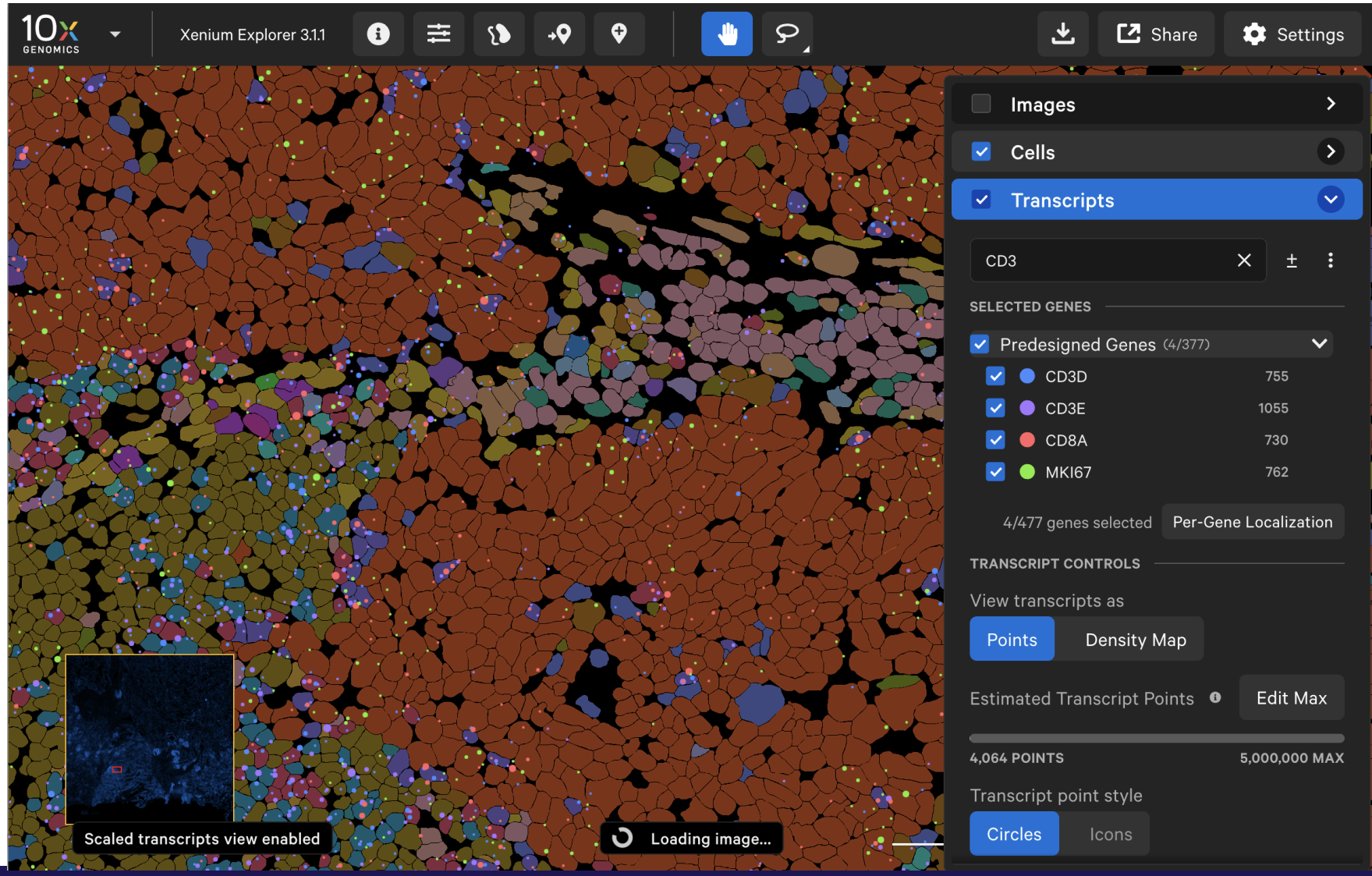
		Visium HD	Xenium Prime 5K
Technical Specs	**Methodology**	Sequencing-based spatial transcriptomics	Imaging-based spatial transcriptomics
	Data acquisition	NGS	High-resolution imaging
	Analyzable area	6.5 x 6.5 mm per tissue section (2 per Visium slide)	22.5 x 10.5 mm per slide (2 slides per run)
Plexity	**Number of RNA targets**	Whole transcriptome (~20,000 transcripts)	Up to 5,000 genes
	Customization options	Custom probe spike-in for custom and exogenous genes	Up to 100 custom genes added to pre-designed panels Advanced custom probes for exogenous genes and gene isoforms *Note: Standalone custom 480-plex panels are also available on Xenium*
Sensitivity	**Transcript abundance (total transcripts per analyzed area)**	~1.5-fold greater than Xenium Prime 5K*	Tissue dependent

Visium HD vs Xenium

	Visium HD	Xenium Prime 5K
Resolution	Single cell scale (data output at 2 x 2 µm and multiple bin sizes; 8 x 8 µm bin is a recommended starting point for analysis)	Subcellular (transcripts localized within single cells with < 50 nm (XY-localization) and < 100 nm (Z-localization) precision)
Cell segmentation	Requires third-party tools	Multimodal cell segmentation performed during instrument run
Multiomics: Additional assays on same section	Pre-run: Morphology (H&E) or IF	Post-run: Morphology (H&E), IF, and Visium/Visium HD
Tissue Fixation: Sample compatibility	FF, fixed frozen, and FFPE	FF and FFPE
Organisms	Human and mouse, with non-human primate and rat possible depending on customer support guidance	Human and mouse *Note: Other species are available on standalone panels with up to 480 genes through Advanced Custom Design*
Tissue format	Pre-sectioned/archived slides Tissue block (sectioned onto standard glass slides)	Tissue block (must be sectioned onto Xenium slide)

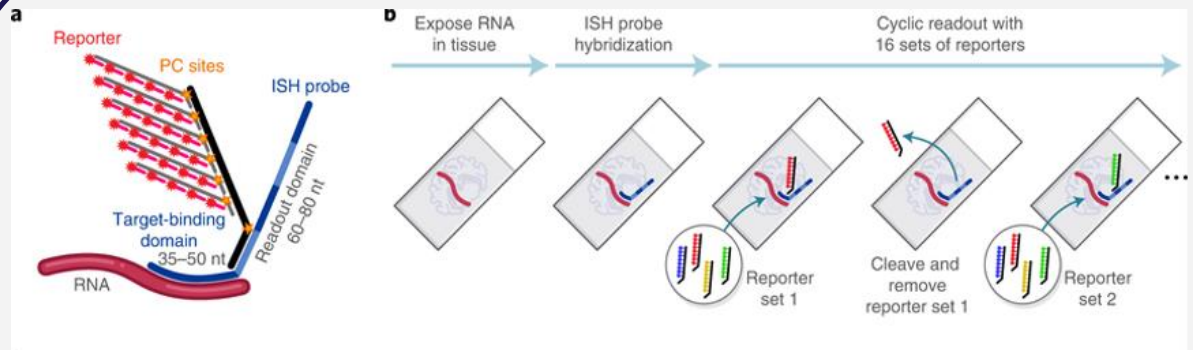
Live example of Xenium Explorer

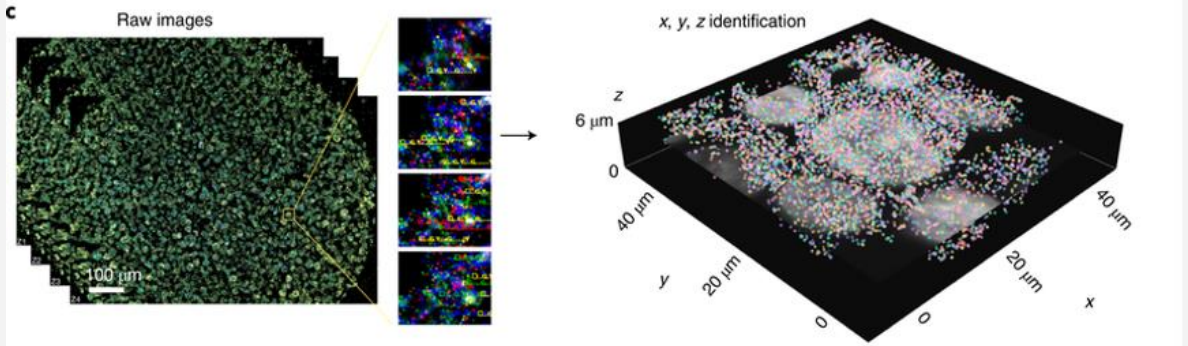
- analysis
- analysis_summary.html
- analysis.zarr.zip
- aux_outputs
- cell_boundaries.csv.gz
- cell_boundaries.parquet
- cell_feature_matrix
- cell_feature_matrix.h5
- cell_feature_matrix.zarr.zip
- cells.csv.gz
- cells.parquet
- cells.zarr.zip
- experiment.xenium
- gene_panel.json
- metrics_summary.csv
- morphology_focus
- morphology.ome.tif
- nucleus_boundaries.csv.gz
- nucleus_boundaries.parquet
- transcripts.parquet
- transcripts.zarr.zip

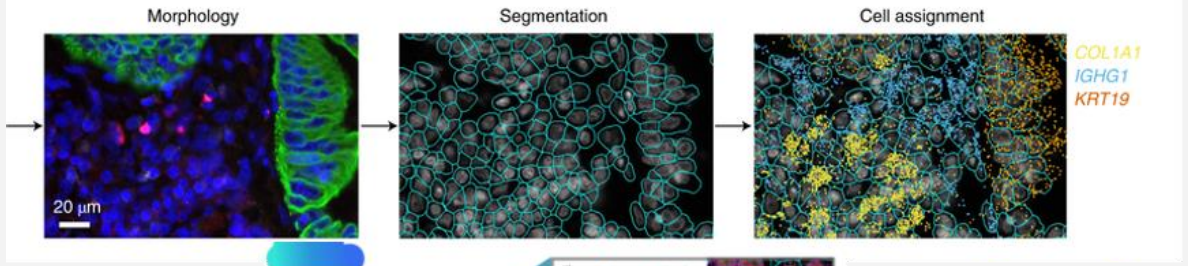


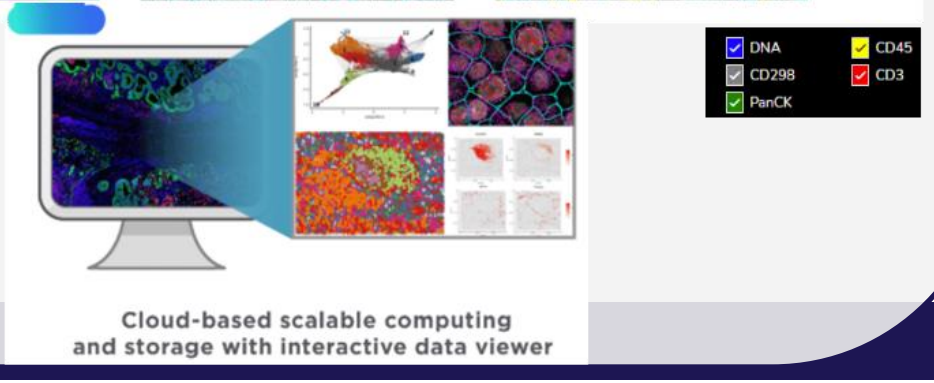
CosMx

- Material (freshly cut block on a few recommended slides)
- Can add several samples per slide
- Destructive protocol,
- Slide dimensions (20x15mm)
- Machine cycles, image-based sequencing
- AtoMx platform, performs cell segmentation, and delivers pre-processed data
- Single cell/subcellular resolution (50nm)
- No protein panel on the same slide, but available for adjacent tissue 64-plex vs 120 plex
- Can add a few customizable panel (very expresive) and off the shelf panels pan-cancer (1000 and 6000? plex)
- Validation technique

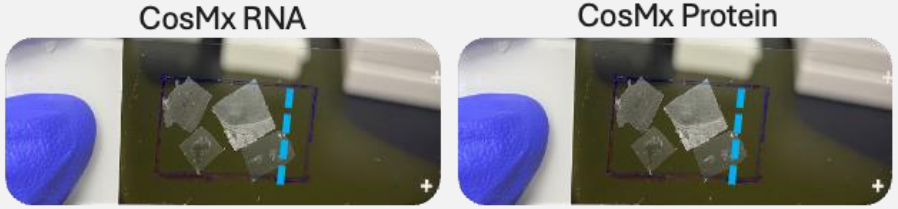


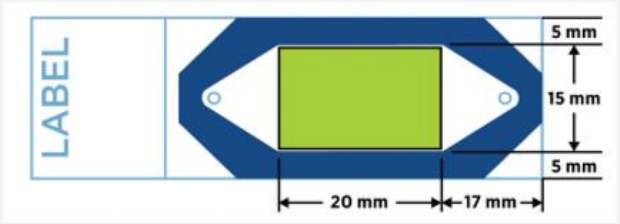






Storage at 4C after sectioning for RNA <2 weeks, and for protein <8 weeks
working area/slide = 15x20mm (≈4 samples)





Comparatively longer run time than other ISH techniques

CosMx™ SMI Turnaround Time Estimation

Analyte	Plex Format	Slide Format	Turnaround Time (Days)				
RNA	100	2-Slide	2	3	4	8	11
		4-Slide	2	4	8	-	-
	1000	2-Slide	3	4	7	13	-
		4-Slide	4	7	13	-	-
Protein	64	2-Slide	1	3	5	10	-
		4-Slide	2	5	10	-	-
Tissue Area Coverage			16 mm²	50 mm²	100 mm²	200 mm²	300 mm²
Fields of View			62	193	386	772	1158

Turnaround estimates are designed as experimental guidance and may change slightly from run to run.

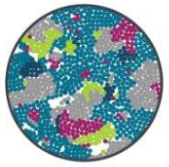
FOR RESEARCH USE ONLY. Not for use in diagnostic procedures.

nano

CRI Cancer Research Institute™

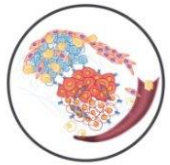
nanoString
A BRUKER COMPANY

Key Applications



Cell Atlasing/Cell Typing

Discover and map cell types using expression profiles of known RNA and protein targets



Disease State

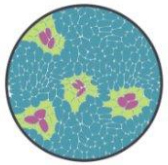
Visualize and quantify molecular (RNA / protein) and cellular organizational changes in a tissue

Ligand-Receptor Interaction

Analyze expression and interactions of up to 100 classic ligand-receptor pairs between interacting cells

Biomarker Discovery

Reveal differential gene expression and pathways in the same cell types depending on their spatial location



Tissue Microenvironment

Understand cellular neighborhoods by examining individual cells and their interacting neighboring cells

CosMx AtoMx Platform

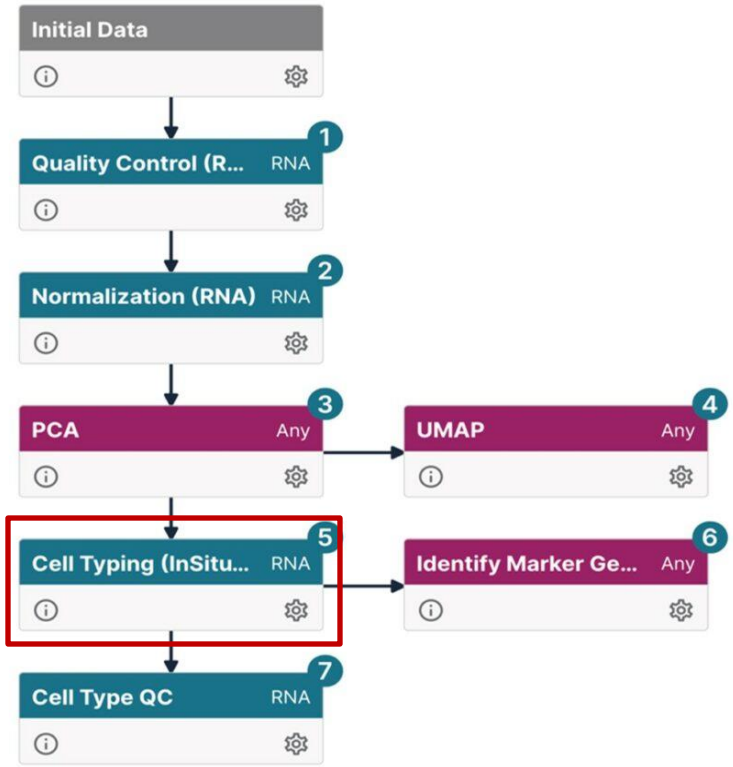
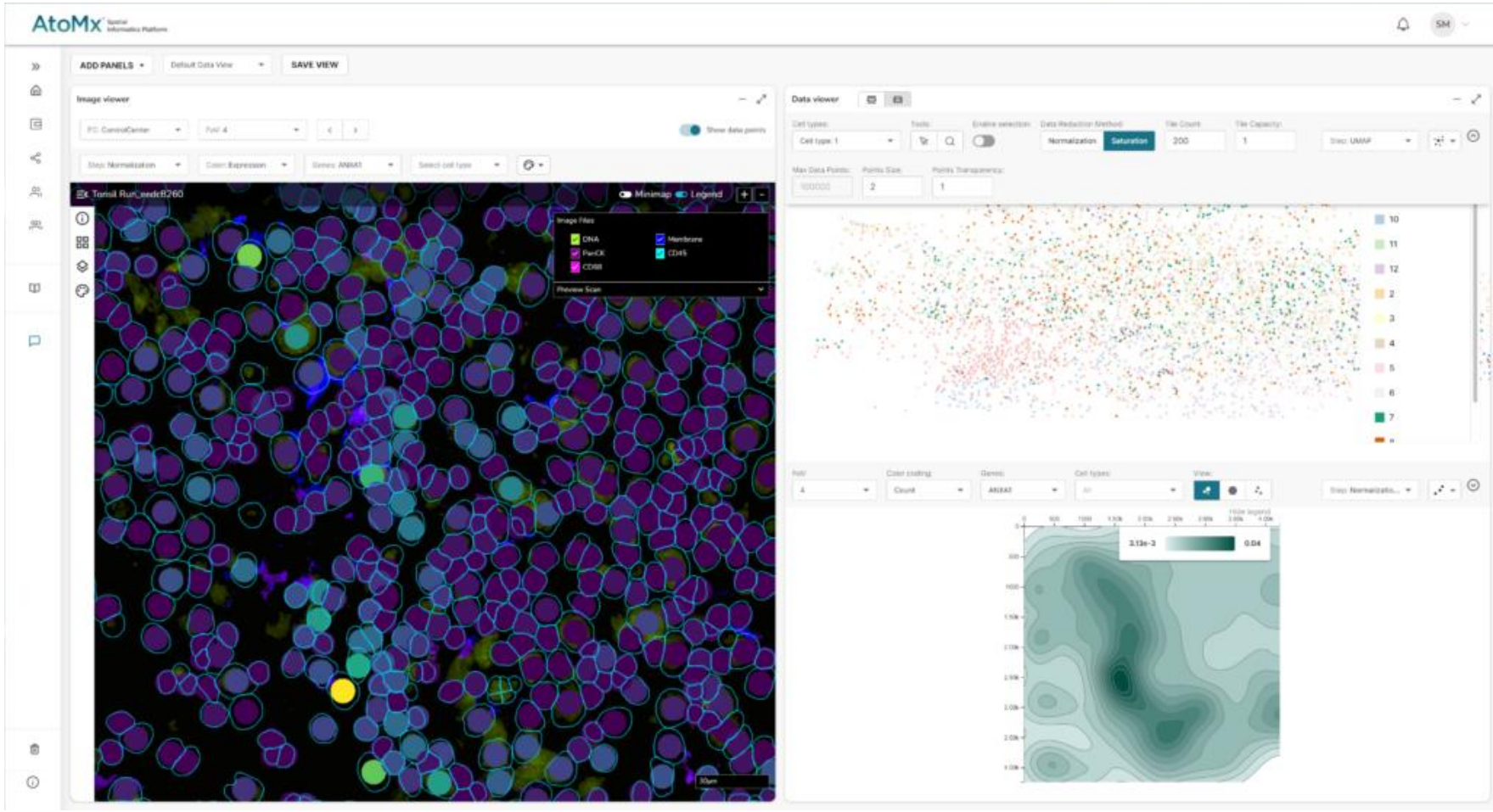


Figure 1: the standard AtoMx cell typing pipeline for CosMx RNA assays.



Can add publicly available gene/cell type matrix from scRNA data to transfer cell type labels

Sample Pre-processing: Visium HD

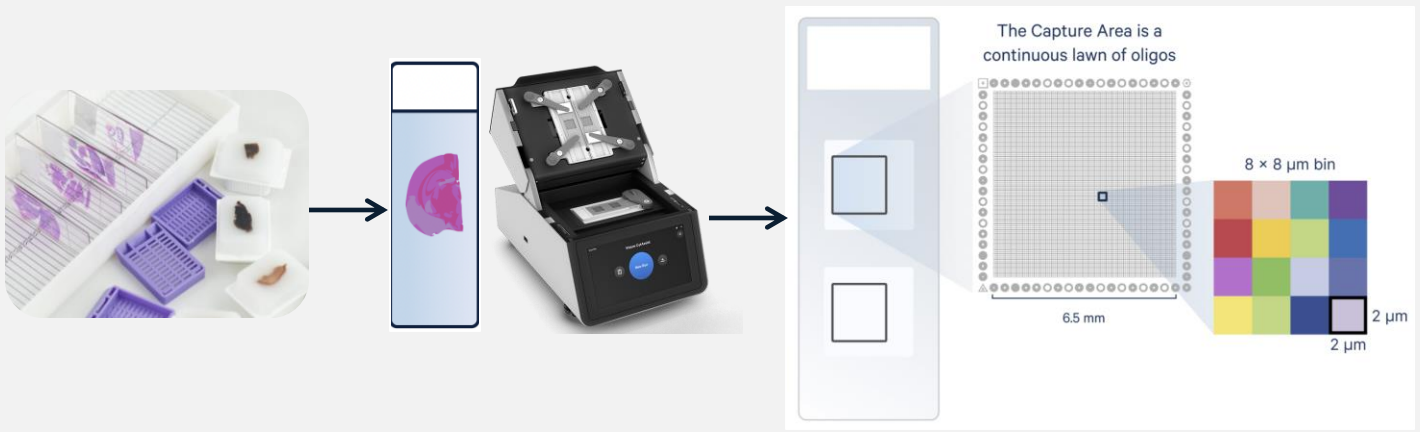
Visium Spatial Transcriptomics HD enables **18,000 whole transcriptome** on 5µm sections of **FFPE blocks** or **archived (un)stained slides** with DV200>30% at a (2µmx2µm) resolution

CRI Cancer Research Institute™

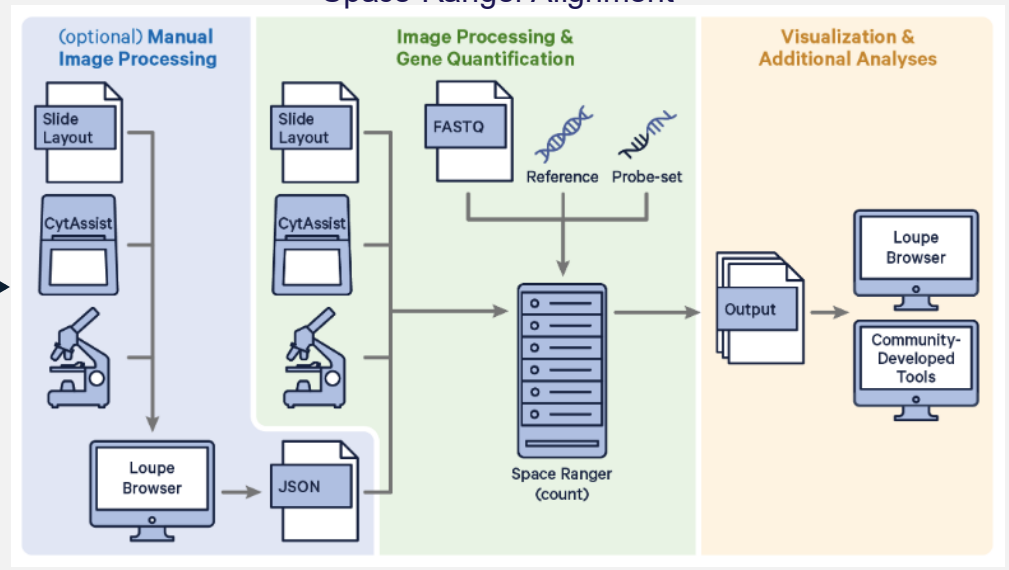
Input: 5µm FFPE/OCT tissue (6.5x6.5mm)

Output: high resolution H&E (brightfield) or IF (Fluorescent) image (≈1GB) and FASTQ files

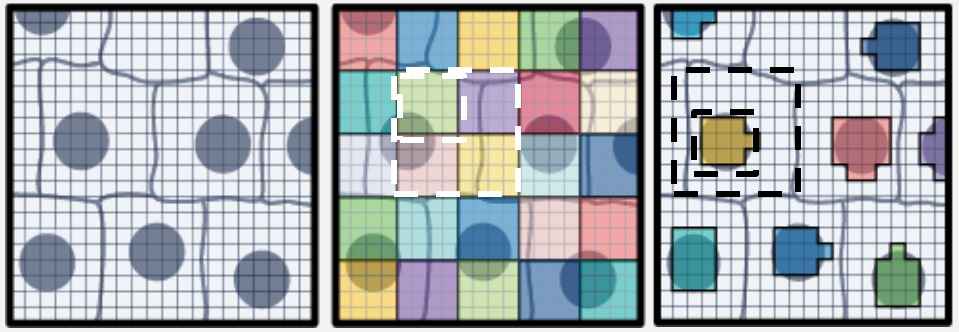
A Visium HD experimental planning



B Space Ranger Alignment



C Nuclei segmentation and custom binning



D Third party tool analysis

Input files

Before running the pipeline, check that you have the following inputs prepared:

- The corresponding CytAssist image in `TIFF` format (`--cytaimage`)
- Microscope image (optional) in either `TIFF`, `QPTIFF`, `BTF`, or `JPEG` format:
  - `--image` for a brightfield microscope image
  - `--darkimage` for a dark background fluorescence microscope image
  - `--colorizedimage` for a composite colored fluorescence microscope image
- Slide parameters specified by:
  - `--slide` & `--area` if `spaceranger` has access to internet (optional with CytAssist metadata)
  - `--slidefile`, `--slide` & `--area` if `spaceranger` has no access to the internet (the slide layout file must be directly downloaded
  - `--unknown-slide` if Visium slide details are unknown
- The reference transcriptome (`--transcriptome`). Human and mouse references are available for download, or build a custom reference
- The probe set CSV (`--probe-set`). These files can be found in the `probe_sets` directory in the Space Ranger software package or downloaded here
- In most cases Space Ranger will perform automatic image to fiducial alignment and tissue detection. When automatic alignment fails, you must also run manual alignment in Loupe Browser and specify the alignment JSON file with the `--loupe-alignment` option.

SpaceRanger Input: high resolution H&E (brightfield) or IF (Fluorescent) image (≈1GB) and FASTQ files (sometimes json files)

SpaceRanger Outputs: can be in customized bins, multiple of 2, up to 100x100µm bin

Show Owner/Mode Show Dotfiles Filter:

Showing 3 rows - 0 rows selected

Type	Name	Size	Modified at
	square_016um	-	12/13/2024 2:02:00 AM
	square_008um	-	12/13/2024 2:01:51 AM
	square_002um	-	12/13/2024 2:01:26 AM

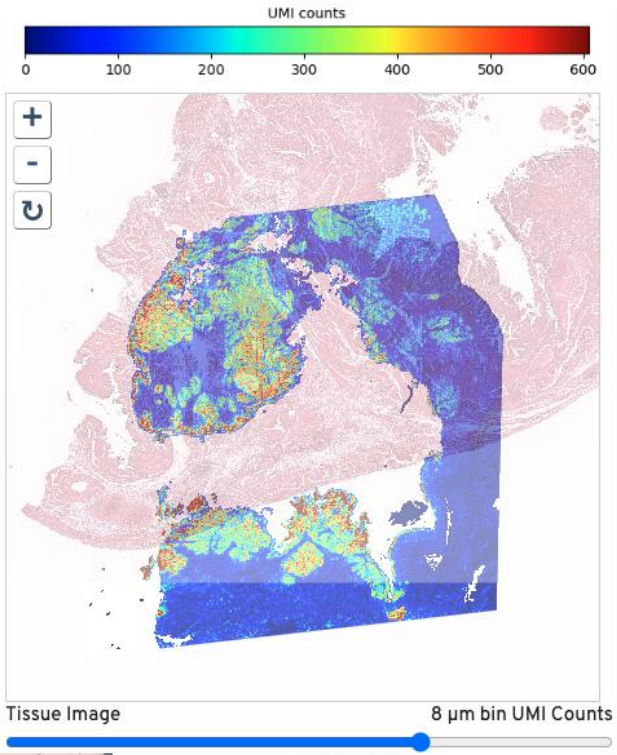
We choose the 8µmx8µm bin to achieve closer to a single cell and since it is the smallest bin that can produce a cLoupe Browser interactive file

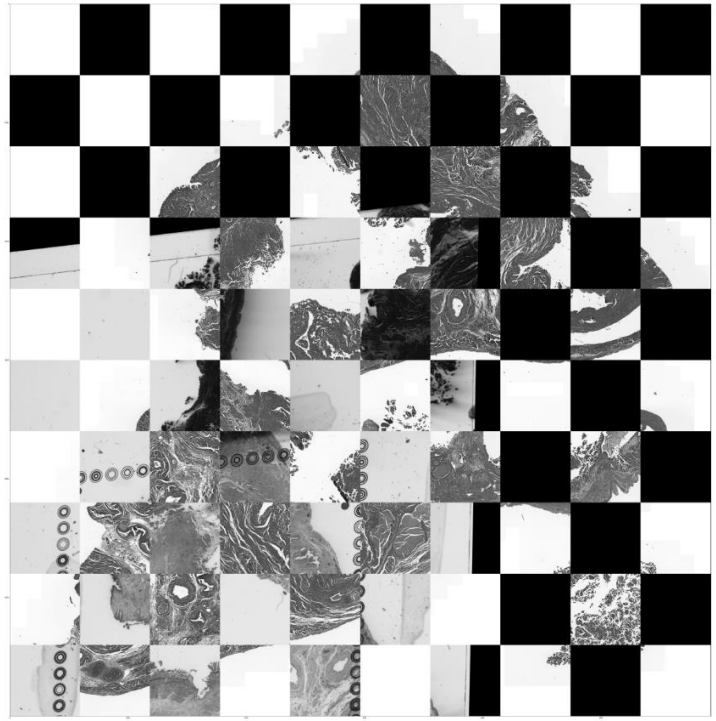
Tissue Troubleshooting: First step is to always examine the outputs **web summary** for any errors, warnings, or tissue mis-alignment

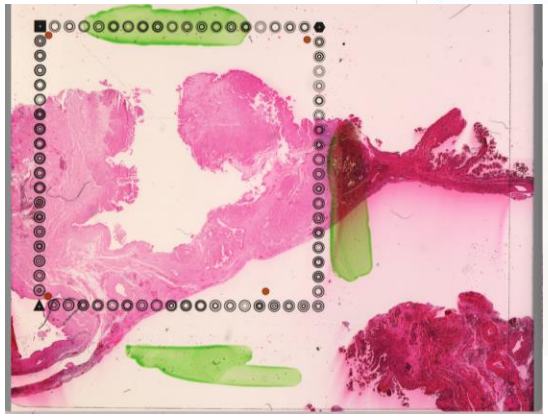
Reads Half-Mapped to Probe Set	0.2%
Reads Split-Mapped to Probe Set	0.1%

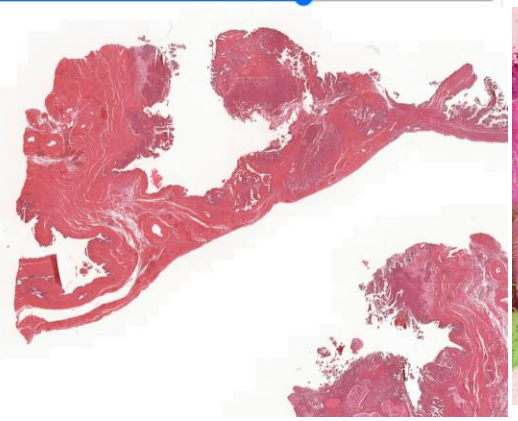
Sequencing

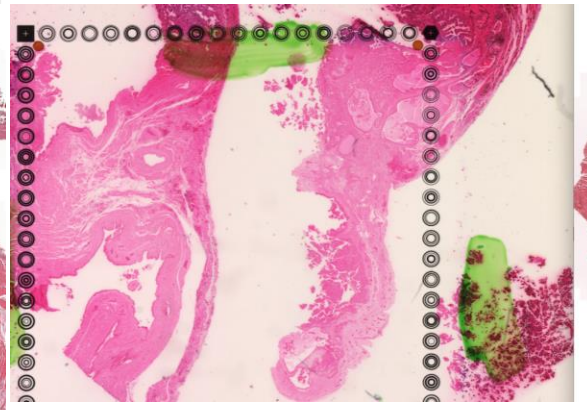
Number of Reads	781,188,369
Valid Barcodes	95.6%
Valid UMIs	99.7%
Sequencing Saturation	90.4%
Q30 Bases in Barcode	98.1%
Q30 Bases in Probe Read	97.9%
Q30 Bases in UMI	97.0%
Fraction of Bins Under Tissue 8 µm	59.1%
Fraction Reads in Squares Under Tissue	94.1%

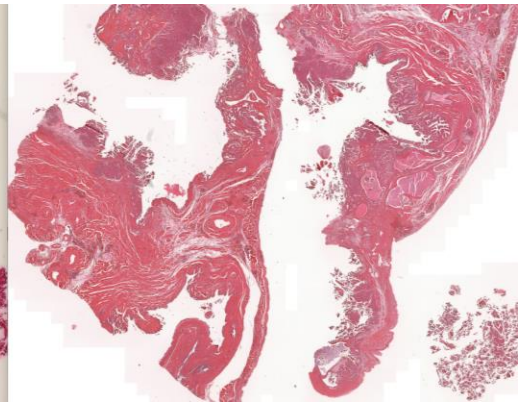








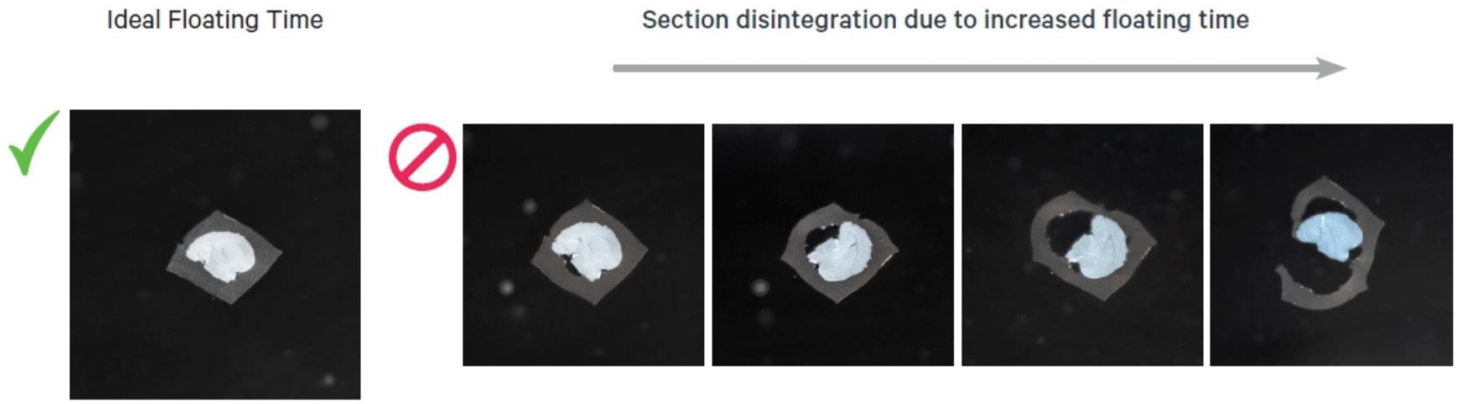




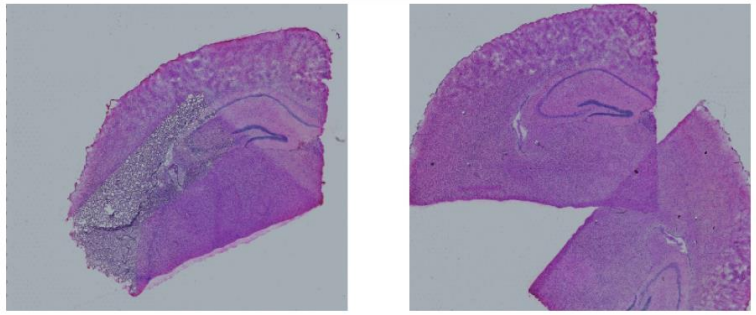
Tissue Troubleshooting: Prior to the Visium HD CytAssist alignment by examining the H&E QC section

Troubleshooting

Ideal Floating Time Determination



Incorrect Placement of Tissue Sections



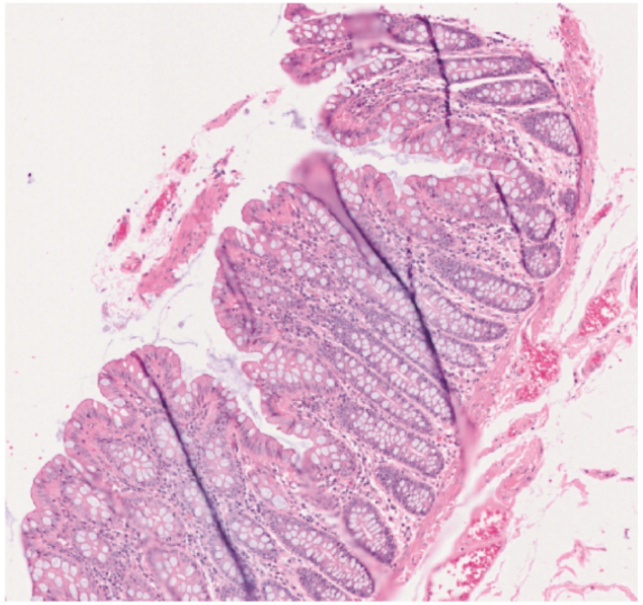
Folded tissue section

Overlapping sections

Tissue Troubleshooting: Prior to the Visium HD CytAssist alignment by examining the H&E QC section

Common Artifacts that cause Detachment

Wrinkles



Causes

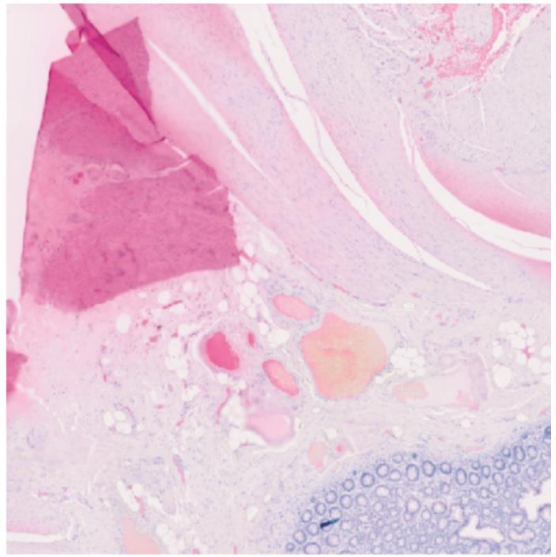
- Section compression (due to warm block or dull blade) during sectioning leads to wrinkle formation. These wrinkles become permanent when placed in the water bath.
- Accumulated wax or static electricity on microtome parts also contribute to section compression.

Troubleshooting

- Ensure that the block is well hydrated.
- Adjust temperature down and increase float time.
- Gently and gradually lay FFPE sections onto water bath surface, lengthwise.
- Utilize a new blade.
- Ensure microtome is cleaned with 100% ethanol to minimize static and section compression (bunching on blade).

Tissue Troubleshooting: Prior to the Visium HD CytAssist alignment by examining the H&E QC section

Folds



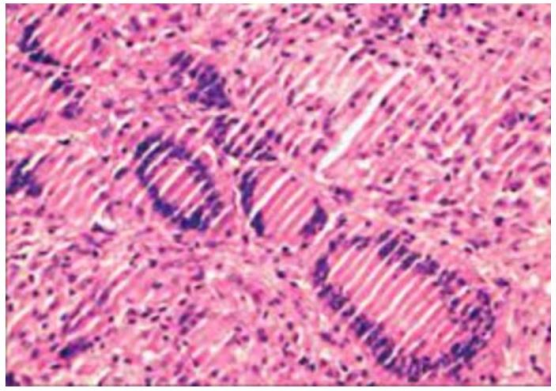
Causes

- Mostly happens when placing the section on the water bath especially when the section is uneven.
- If the fold is at the edge this most likely can happen during sectioning or mounting on the slide.

Troubleshooting

- Gently and gradually lay FFPE ribbons or sections onto water bath surface, lengthwise.
- If sections curl during sectioning, gently flatten them with a brush before floating.

Venetian Blinds or Shatter



Causes

- Parallel lines in the section mostly appear due to dry tissue because of under-hydration of the block in the ice bath.
- Less likely due to dull blade or loose parts of the microtome.

Troubleshooting

- Increase incubation time of the block in an ice bath.
- Tighten down components of microtome and make sure the blade is at a correct angle.

While samples should be carefully QC-ed before proceeding to the Visium HD protocol for any detachment, folding, and DV200, in some instances where no tissue block is available (depleted block), and tissues have already been sectioned and archived, correcting for these mis-alignment/detachment can be done after sequencing and before pre-processing using 10X Loupe Browser

Tissue Troubleshooting: After the Visium HD Run

Few instances for manual fiducial alignment:

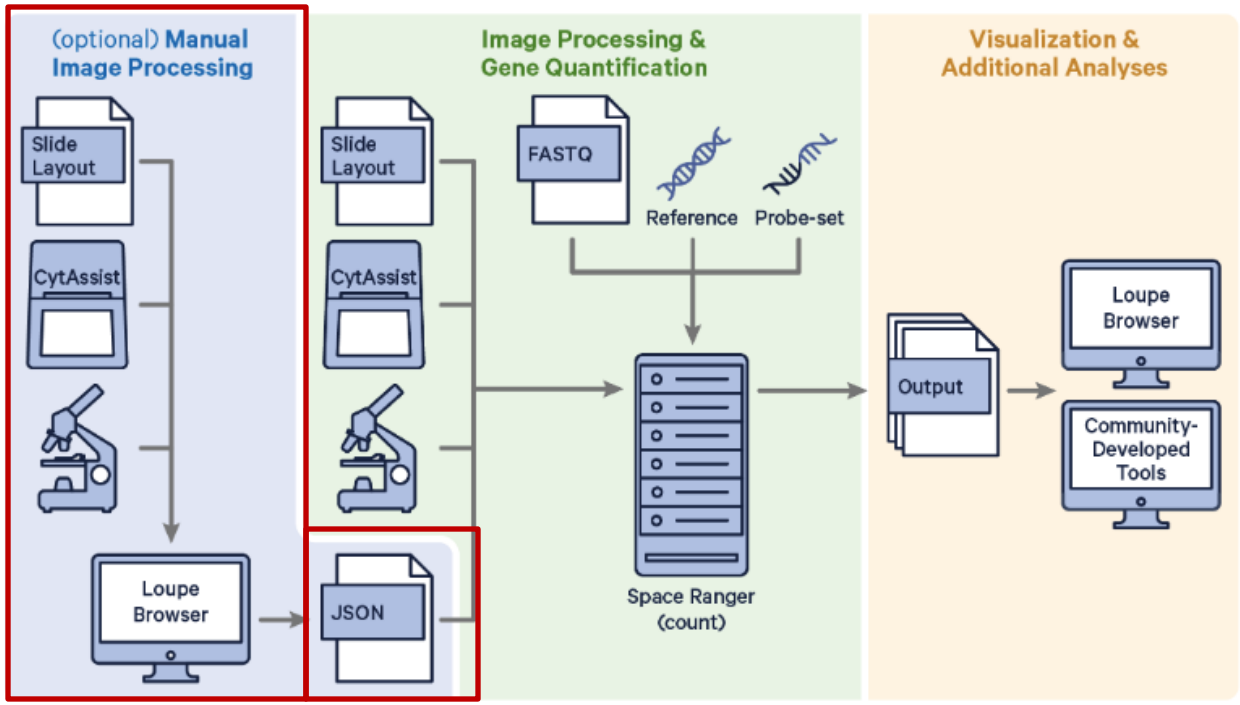
a) when the tissue covers the fiducial frame (old Visium only, no issue with Visium HD)

b) when there is any tissue detachment

c) Sometimes with deeper sequencing, some empty bins might have bleeding over of signal from neighboring bins with tissue present

Visium HD Analysis with spaceranger count

Space Ranger v3.0 and later can analyze Visium HD Spatial Gene Expression datasets generated from formalin fixed paraffin embedded (FFPE) human or mouse samples and the Visium CytAssist instrument. The `spaceranger count` pipeline is necessary to analyze these data.

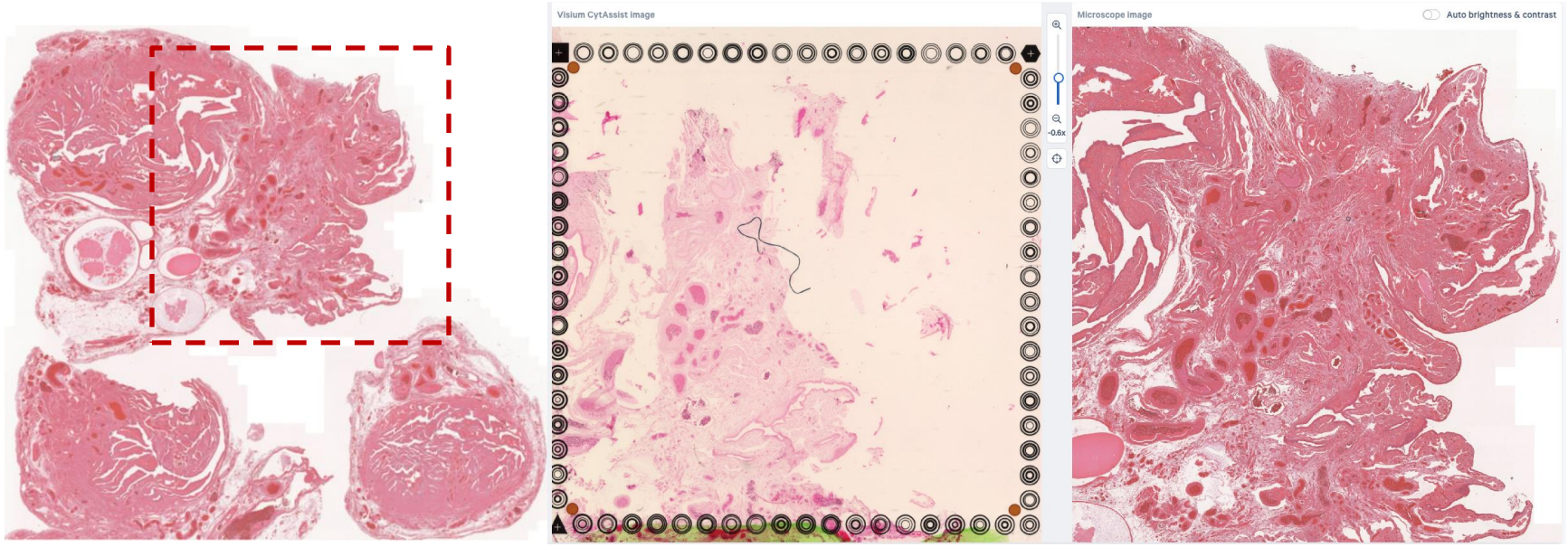


Input files

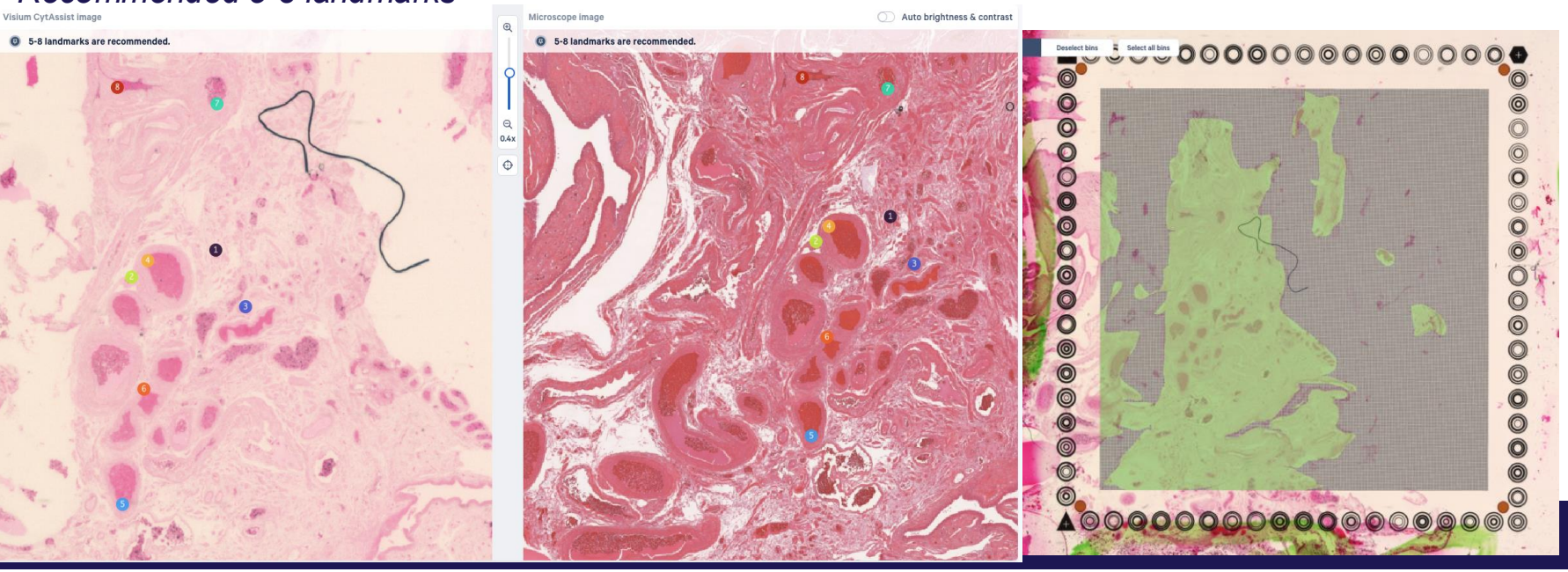
Before running the pipeline, check that you have the following inputs prepared:

- The corresponding CytAssist image in `TIFF` format (`--cytaimage`)
- Microscope image (optional) in either `TIFF`, `QPTIFF`, `BTF`, or `JPEG` format:
  - `--image` for a brightfield microscope image
  - `--darkimage` for a dark background fluorescence microscope image
  - `--colorizedimage` for a composite colored fluorescence microscope image

Tissue Troubleshooting After the Visium HD Run: Manual fiducial alignment when there is any **tissue detachment**

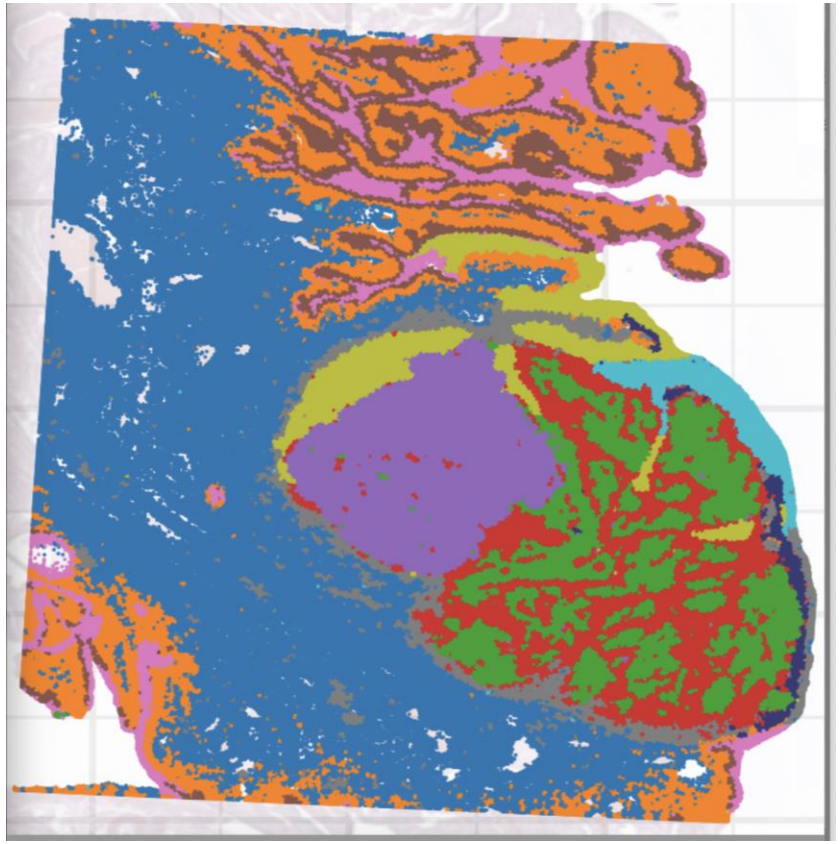


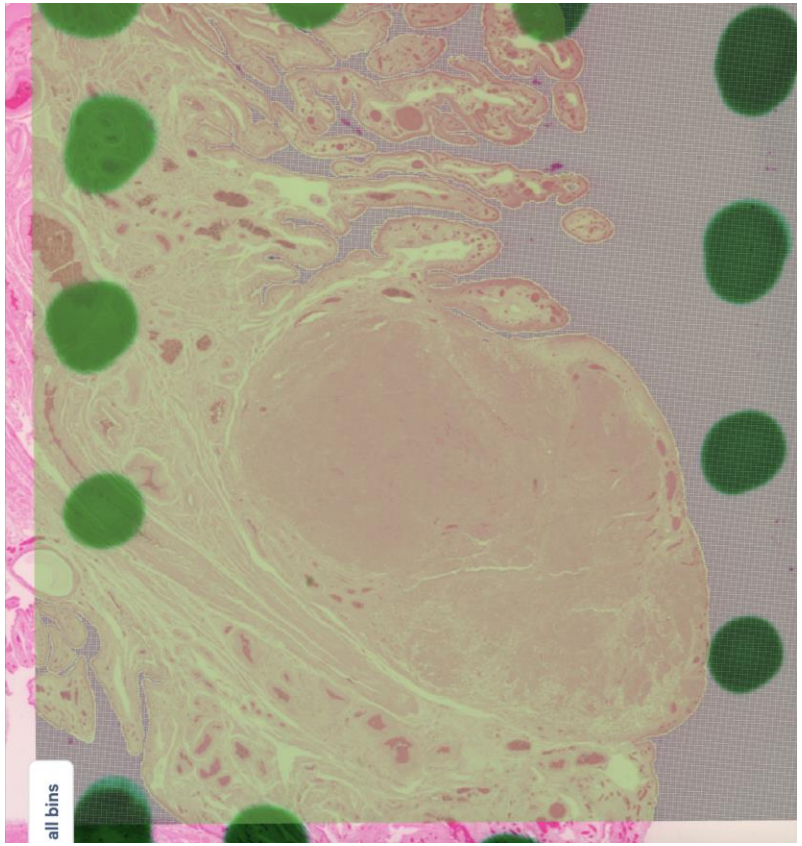
Recommended 5-8 landmarks

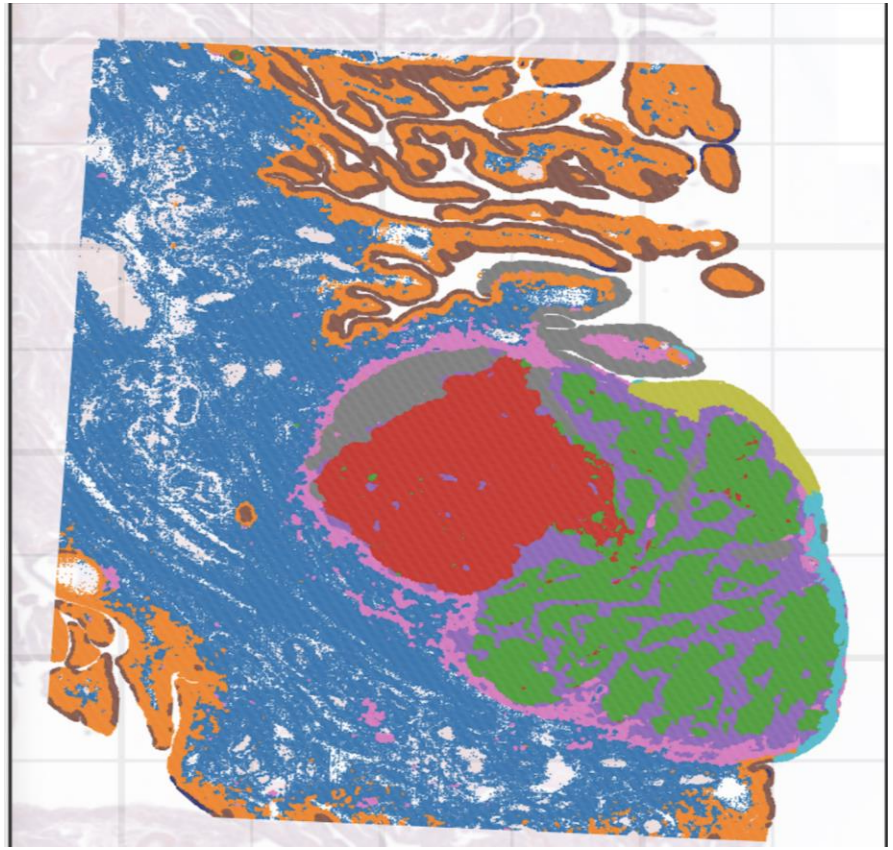


Recommended slides to avoid detachment: Schott Nexterion Slide (Schott, PN-1800434)

Tissue Troubleshooting After the Visium HD Run: With opting for deeper sequencing, some empty bins might have bleeding over of signal from neighboring bins with tissue present



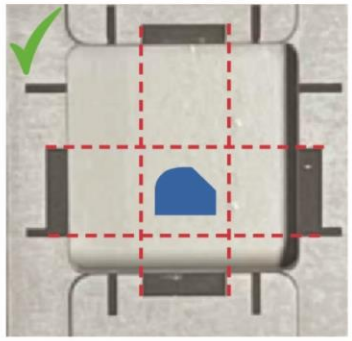




The actual working area is 6x6mm and not 6.5x6.5mm. Avoid placing your important ROI closer to the edges while aligning the H&E/IF slide to the CytAssist gasket

Correct

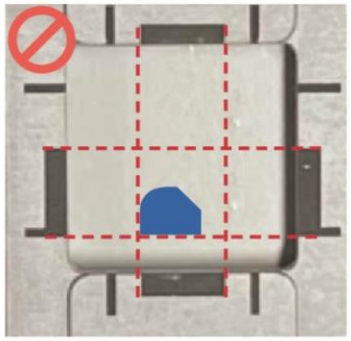
Extend lines from the alignment guides to create an imaginary square. Place the tissue on the center of that square.

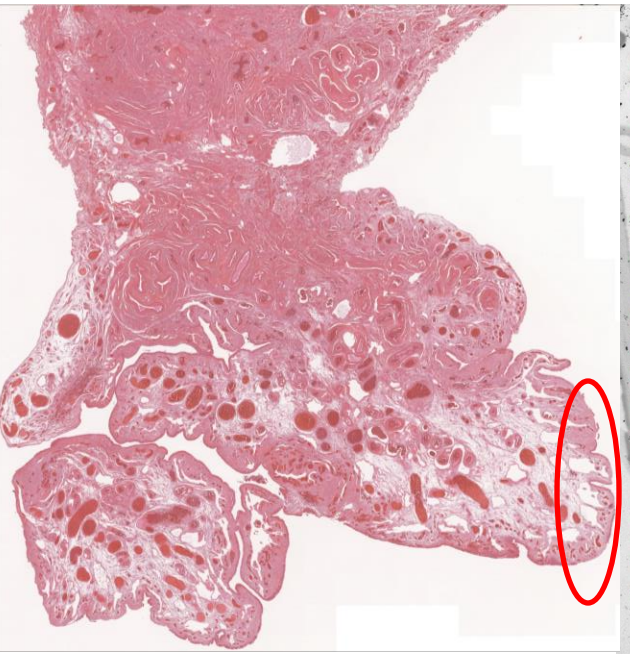


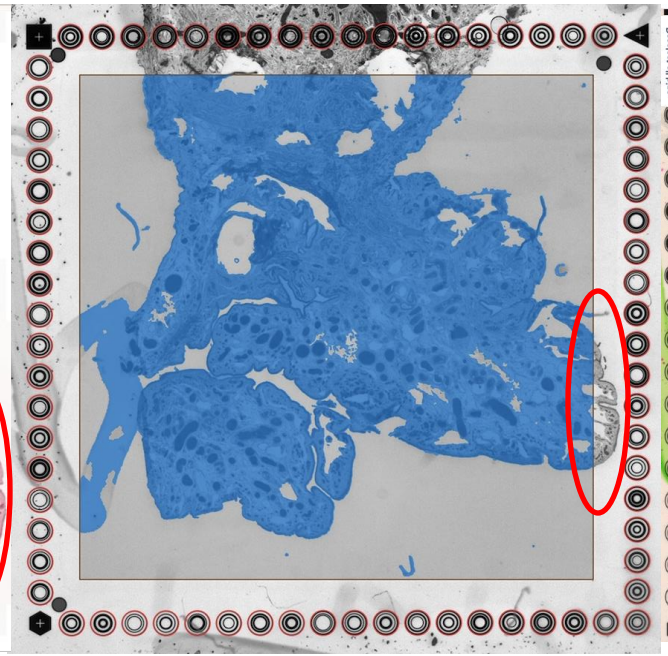
6.5 mm capture area used as an example

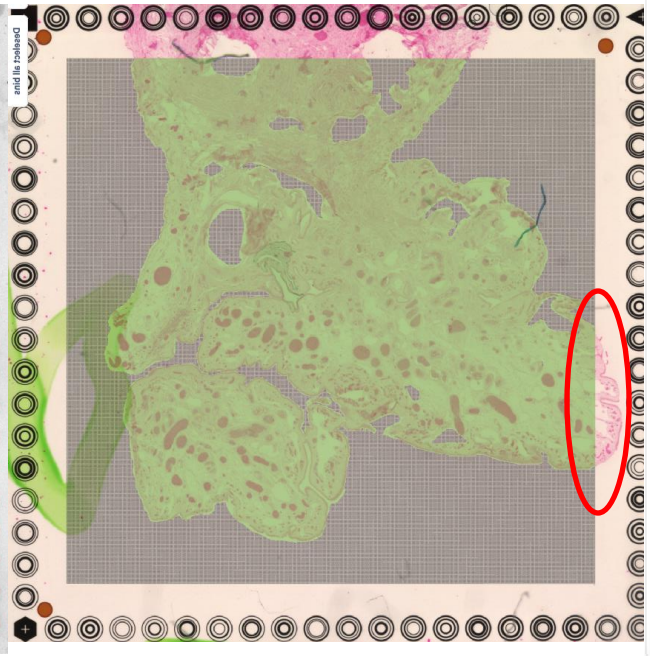
Incorrect

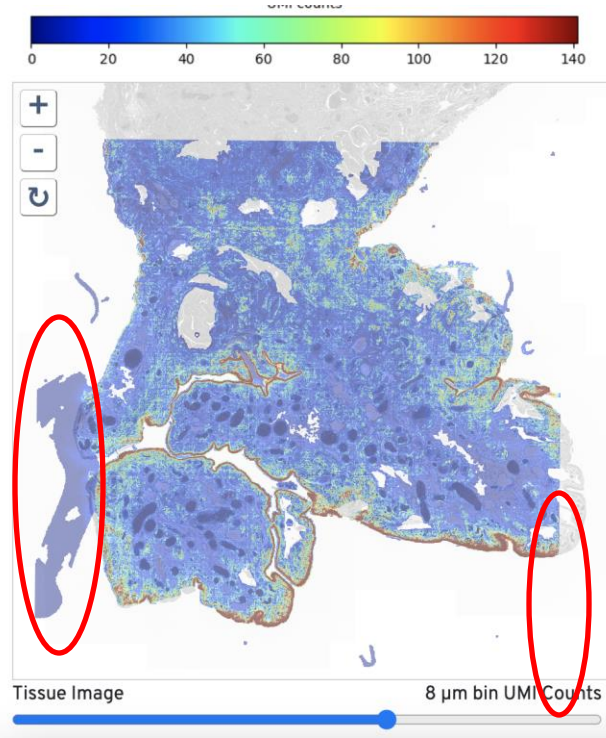
Tissue is within the alignment guides, but should be centered within the imaginary square, not aligned with the top or bottom line.











Live Demo Loupe Browser 8 Manual alignment and json file extraction

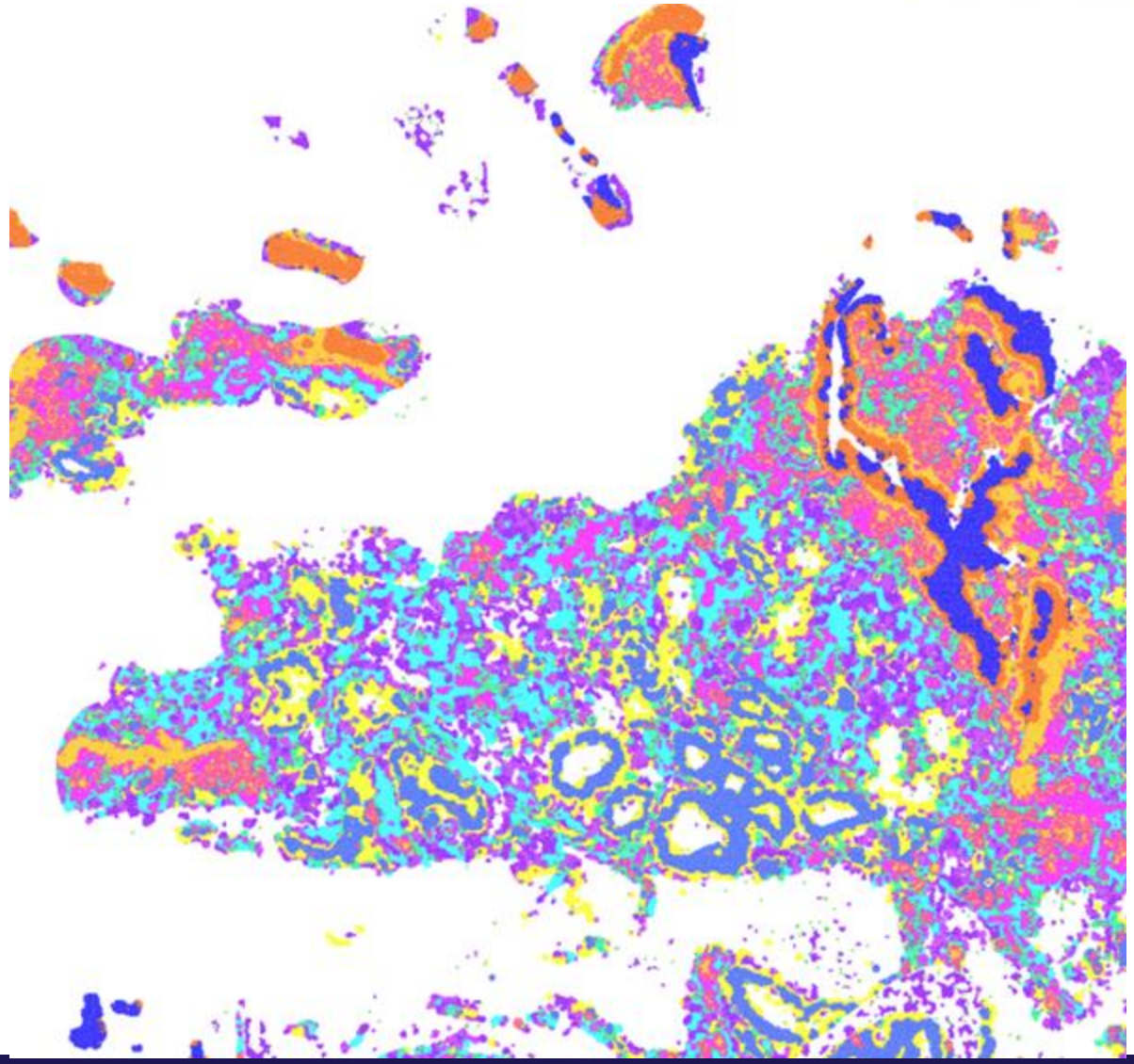
Motifs and neighborhood clustering (Banksy, Grafiti)

Creating a Seurat Object

Graph Representation and Spatial Motifs

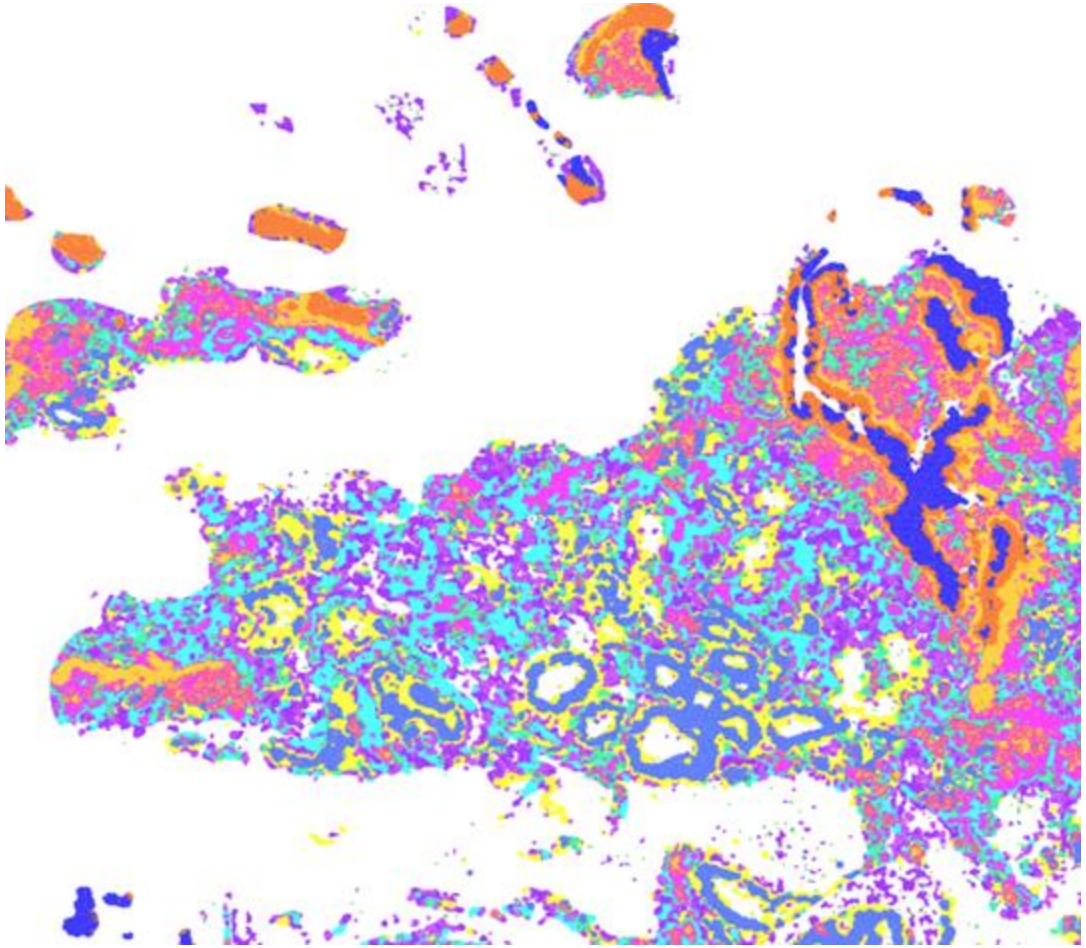
Spatial motif identification in ST is the partitioning of cells or bins into distinct groups that are both spatially and transcriptionally coherent.

- Fundamental step for spatial transcriptomic/IF analyses... similar to clustering/cell typing scrna-seq.
- Not a one to one mapping with cell type.



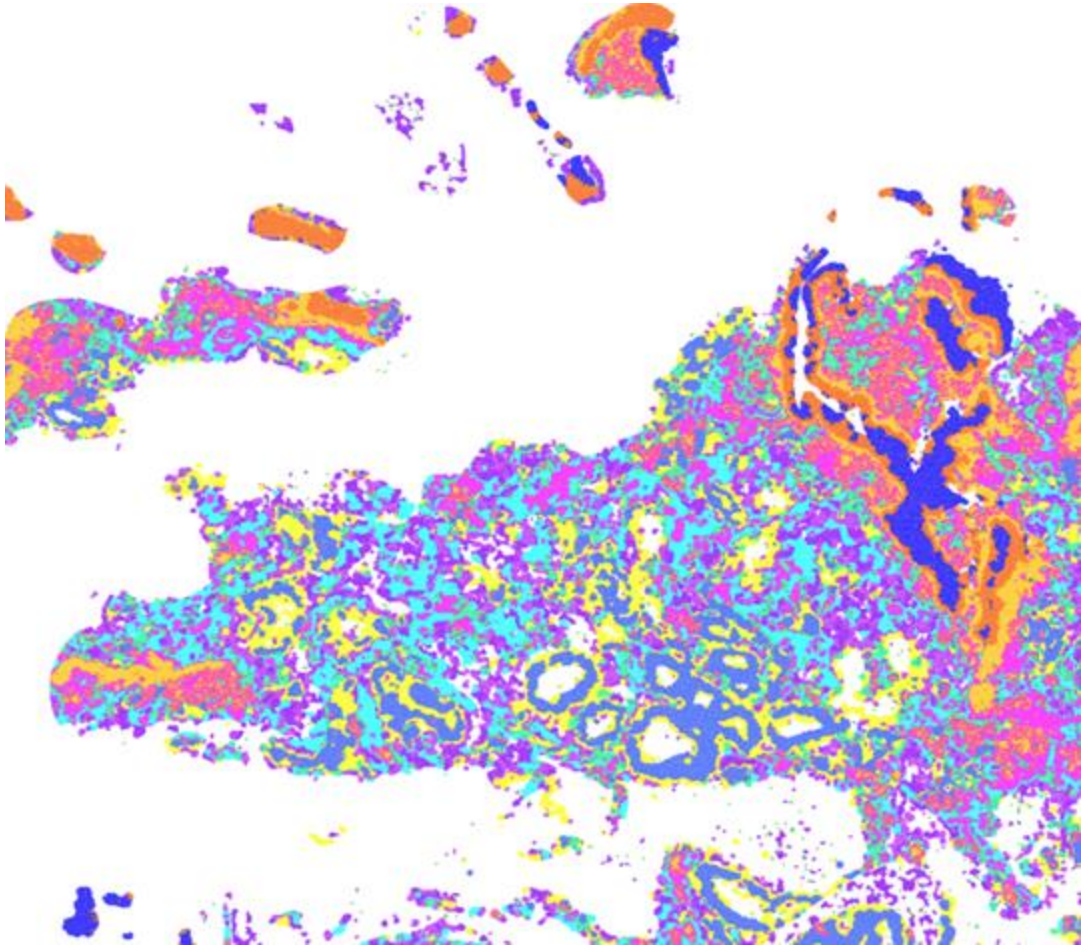
Spatial Motif

•A set of genes whose expression levels covary in a **localized, contiguous tissue region**, forming a coherent transcriptional signature.
•Manifests as a **continuous spatial pattern** across neighboring spots or cells, not scattered randomly.
•Marks a **functional niche or microenvironment** (e.g. immune infiltration front, stromal–tumor boundary) that underlies specific biological processes.



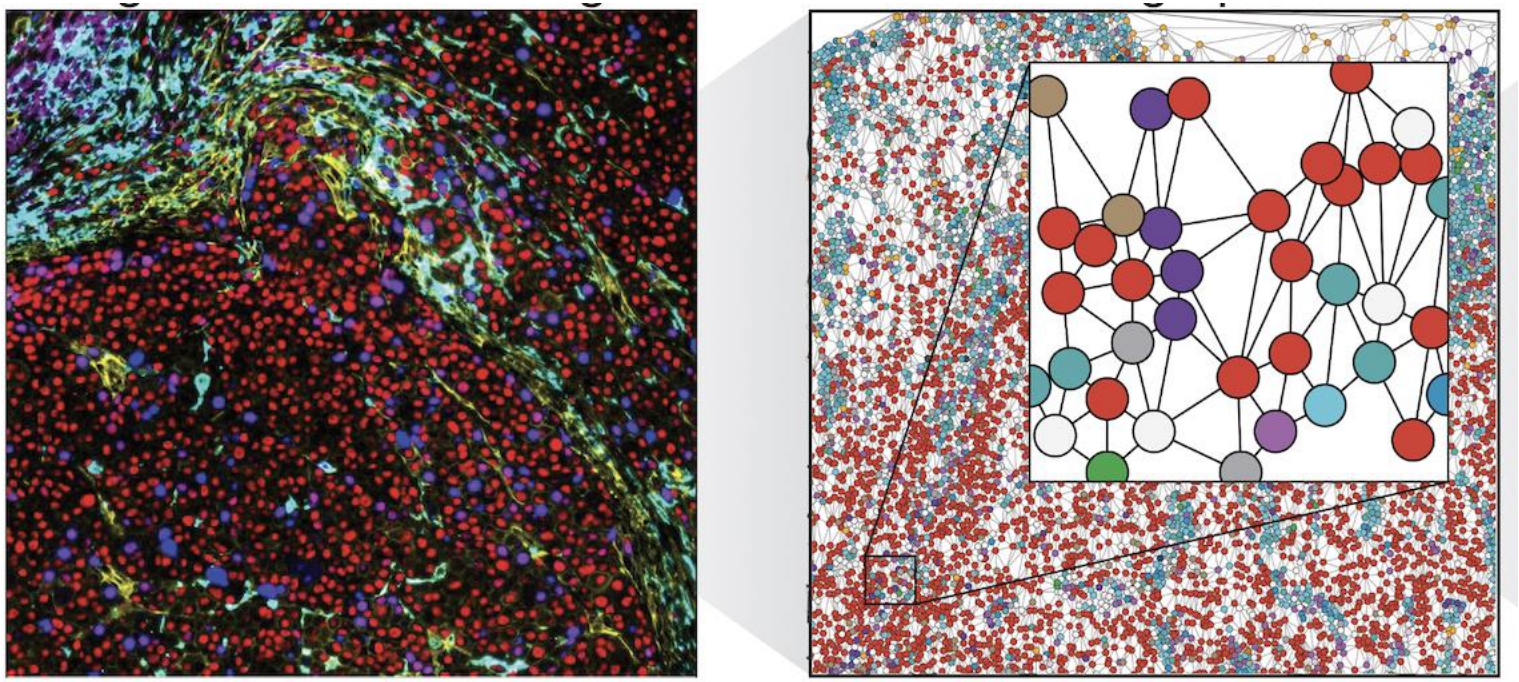
Spatial Motif

•A set of genes whose expression levels covary in a **localized, contiguous tissue region**, forming a coherent transcriptional signature.
•Manifests as a **continuous spatial pattern** across neighboring spots or cells, not scattered randomly.
•Marks a **functional niche or microenvironment** (e.g. immune infiltration front, stromal–tumor boundary) that underlies specific biological processes.



Graphs

- Cells (or spots) ↔ nodes
- Spatial relationships ↔ edges
- Captures local context & interactions
- Enables powerful graph algorithms
- Flexible multi-modal integration
- Scalable & interpretable



Graph Representation

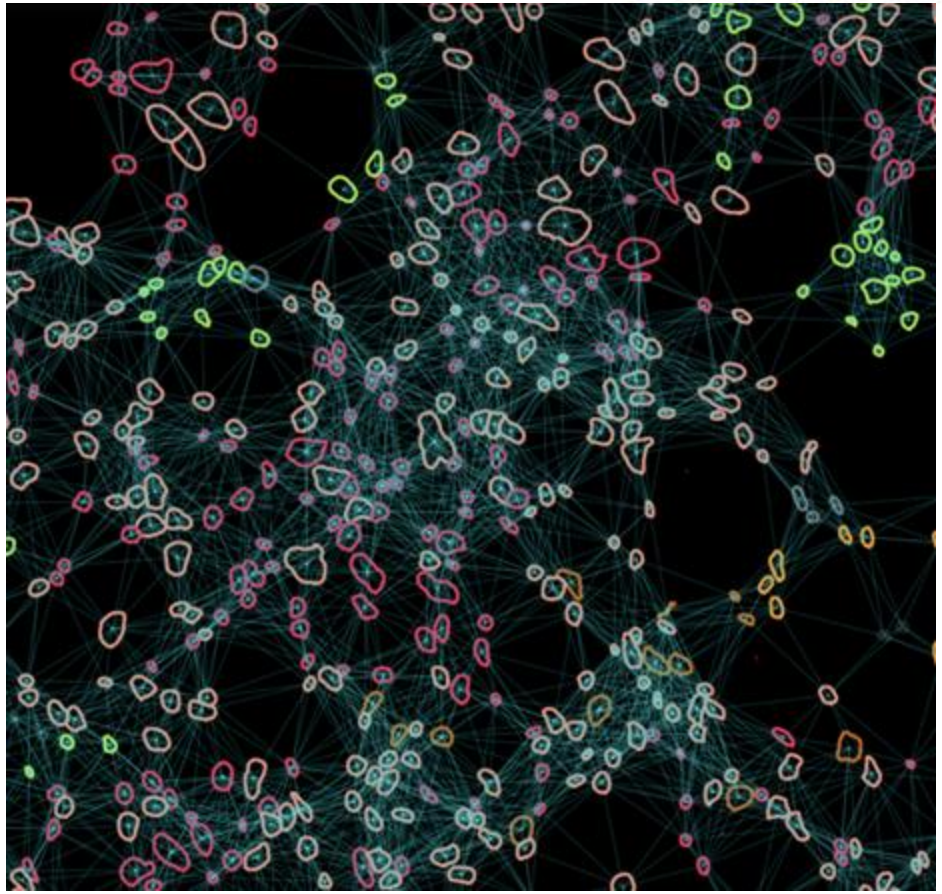
Graph Definition

Let $G = (V, E)$ be an undirected, weighted graph where:

- V is the set of nodes (cells) in the spatial domain.
- $E < V \times V$ is the set of edges, representing spatial reletionships between nodes.
- Each edge $(i, j) \in E$ has an associated weight w_{ij}, which encodes the spatial distance between node i and node j. We assume $w_{jj} > 0$ and the weights are symmetric (i.e., $w_j = w_j$).

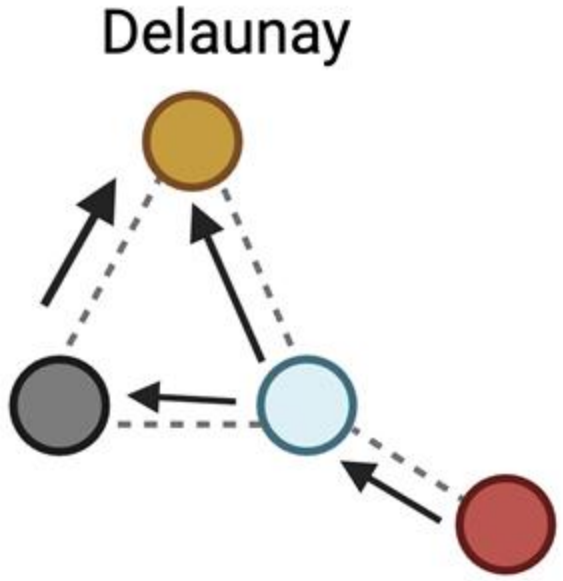
Node Attributes

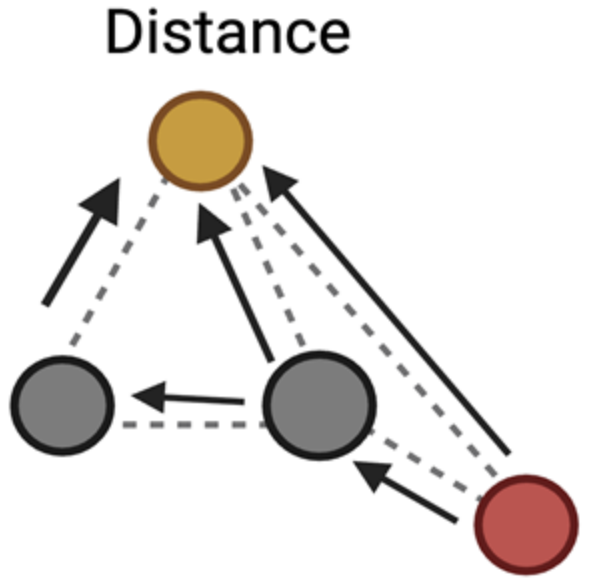
Each node $v \in V$ has an associated feature vector $x_v \in R^d$ that represents the gene expression or phenotypic profile of the cell.



Edges: Cell-Cell Connectivity

- There are different methods for defining a connection.
- **Distance based** graph construction allows any cell with in a 30 um region to directly affects any other cell. (*Rings in spot-based lawns*)
- While **Delaunay triangulation** describes a situation where only the closest cells directly interact, with an indirect interaction “passing through” through intermediate neighbors.





Neighborhood Aggregation

Start with node features

Each node begins with its own feature list (e.g., gene expression levels, protein markers).

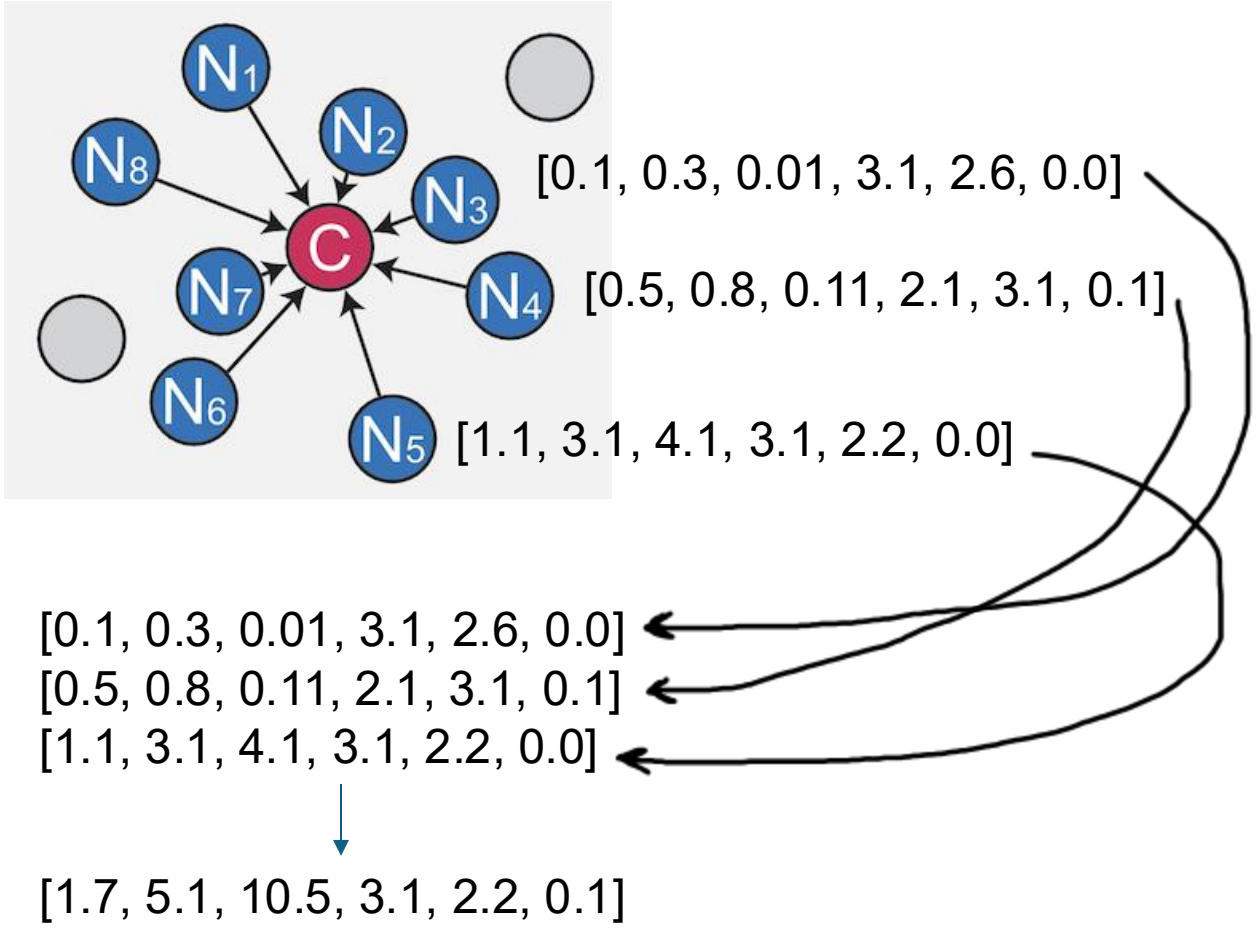
Gather neighbor features

For a given node, look up all directly connected neighbors and pull together each neighbor's feature list into a combined table.

Aggregate the neighbor table

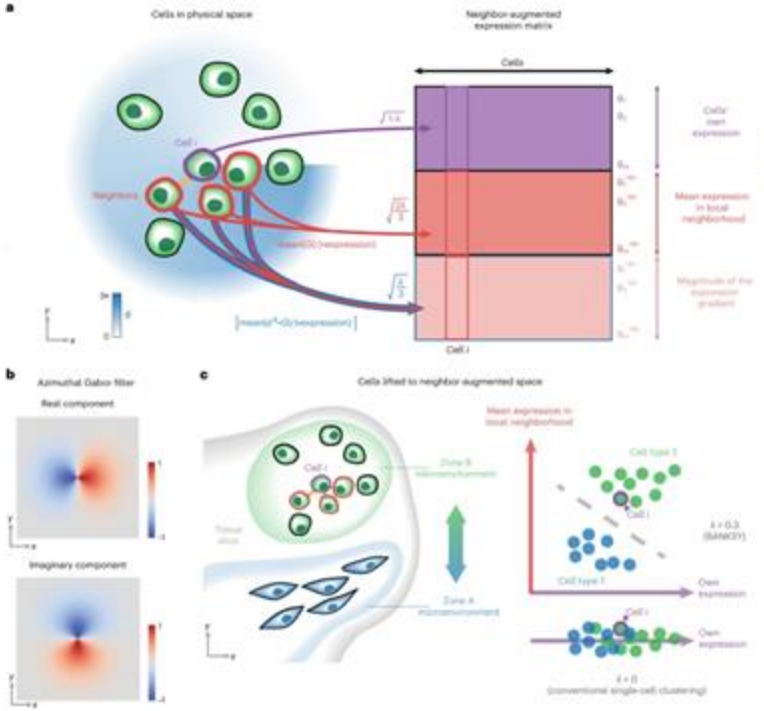
Collapse that table back down to one summary feature list by applying a simple operation across each column—common choices are:

- **Mean** of each feature
- **Sum** of each feature
- **Maximum** of each feature



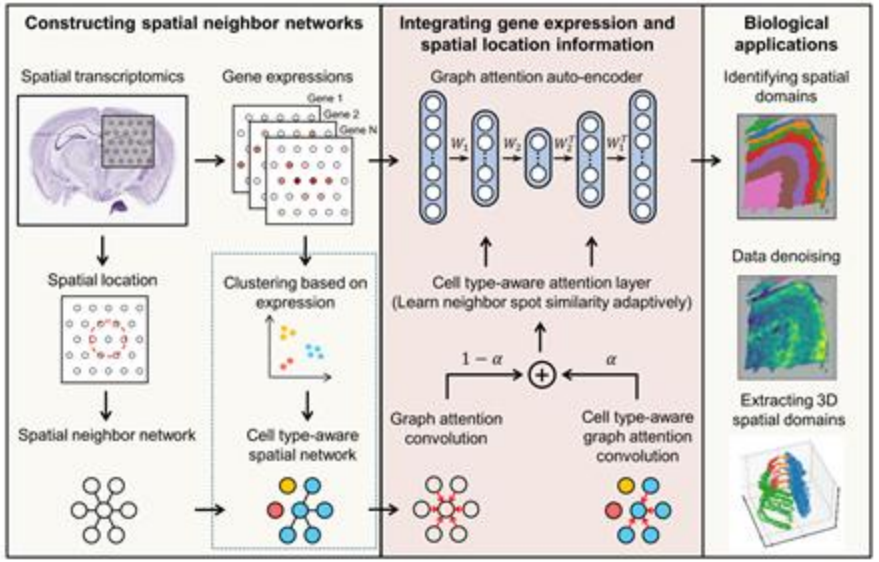
Graph-based Methods (Banksy)

- Dominant paradigm is graph-based methods.
- Graph-based methods
  - Algorithmic:
    - UTAG
    - Cellcharter
    - **Banksy -** Azimuthal Gabor filter (AGF)
  - Deep learning:
    - GraphST
    - **Stagate** - Attention-based



BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis

Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, Hwee Kuan Lee, Kok Hao Chen ✉ & Shyam Prabhakar ✉



Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder

Kangning Dong & Shihua Zhang ✉

Attention-based GNN autoencoder that generates joint embeddings.
Uses gene expression pre-clustering as input and smoothing of cluster assignment output by neighbor.

Method Focus - Bansky

- Unifies these two spatial clustering problems by embedding cells in a product space of their own and the local neighborhood transcriptome.
- BANKSY revealed unexpected niche-dependent cell states in the mouse brain.
- Outperformed competing methods on domain segmentation and cell typing benchmarks.
- Faster and more scalable than existing methods.

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis

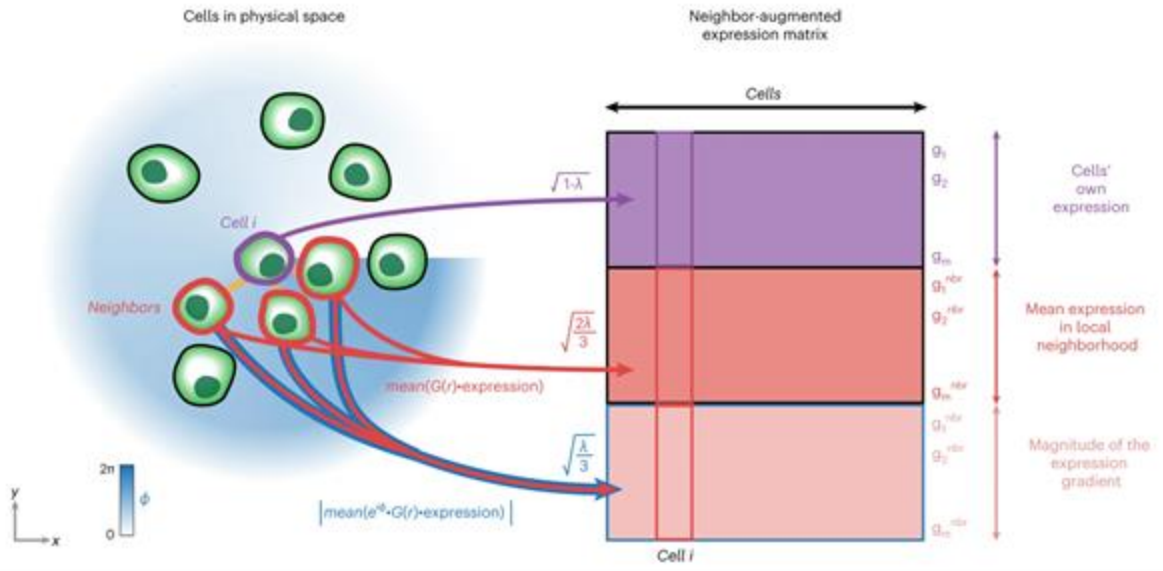
Vipul Singhal, Nigel Chou, Joseph Lee, Yifei Yue, Jinyue Liu, Wan Kee Chock, Li Lin, Yun-Ching Chang, Erica Mei Ling Teo, Jonathan Aow, Hwee Kuan Lee, Kok Hao Chen ✉ & Shyam Prabhakar ✉

Why BANKSY

- Spatially aware cell typing methods
  - scRNA methods are common, but do not use any spatial information.
  - Spatially informed methods make some invalid assumptions (MERINGUE)
    - distant cells are less similar
  - Spatially informed methods don't scale.
- Tissue domain identification methods
  - Problem because each tissue domain could include multiple cell types.
  - Domain segmentation methods encouraged physically proximal cells to have the same label.
    - This assumes that a cell's transcriptome resembles the average transcriptome of cells in its tissue domain, which is not always valid because diverse cell types are commonly intermingled within a single domain.
    - Im not sure this makes sense…..
  - NNs overfit.
  - All methods don't scale or have not been tested at scale.

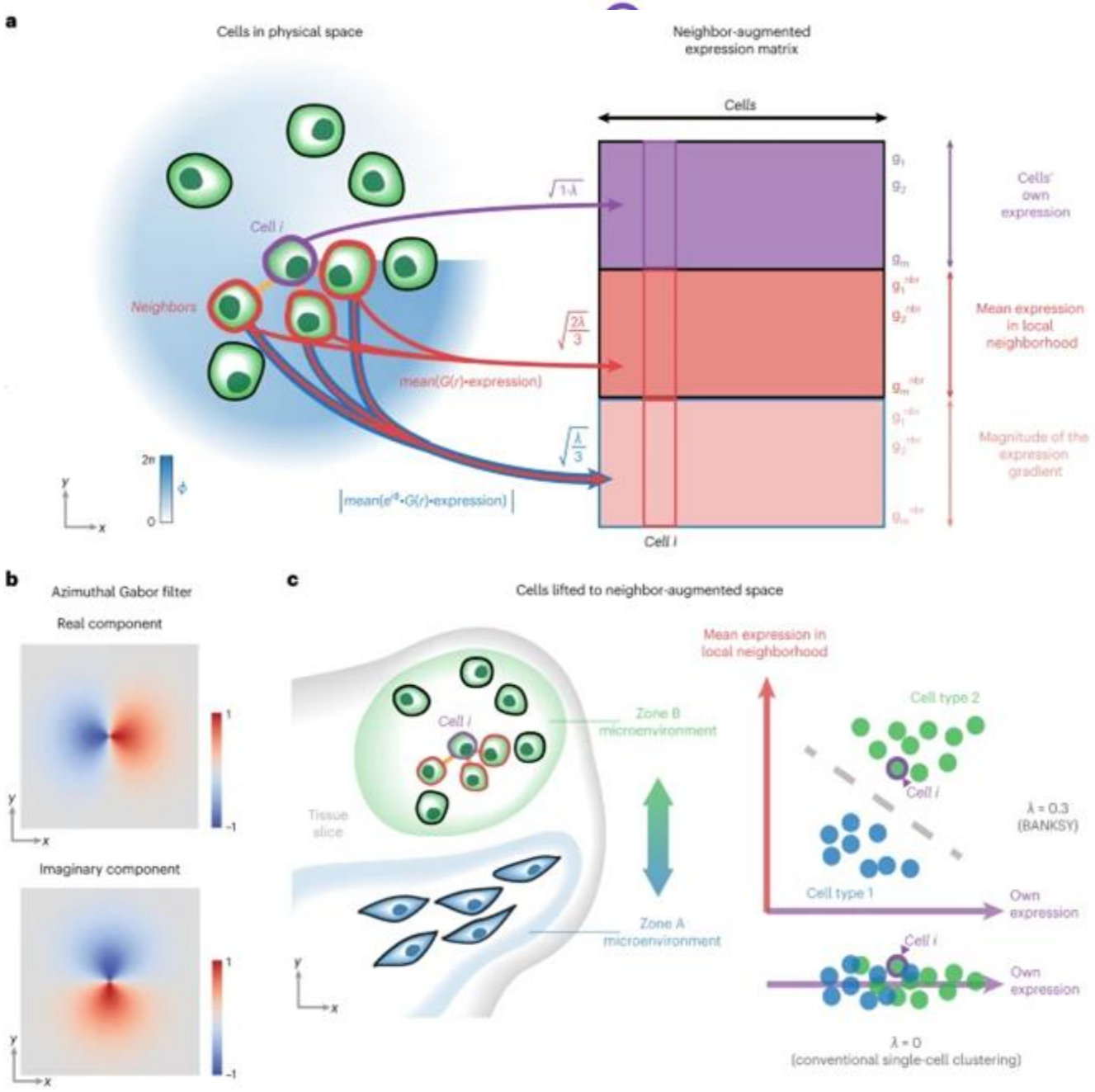
How does Banksy work?

- **B**uilding **A**ggregates with a **N**eighborhood **K**ernel and **S**patial **Y**ardstick
- Leverages the fact that a cell's state can be more fully represented by considering both its own transcriptome and that of its local microenvironment.
- Pair of spatial kernels to encode the transcriptomic texture of the microenvironment
  - Weighted mean of gene expression in each cell's neighborhood
  - Azimuthal Gabor filter (AGF)
- Hyperparameter can tune BANKSY to accurately detect tissue domains rather than cell types



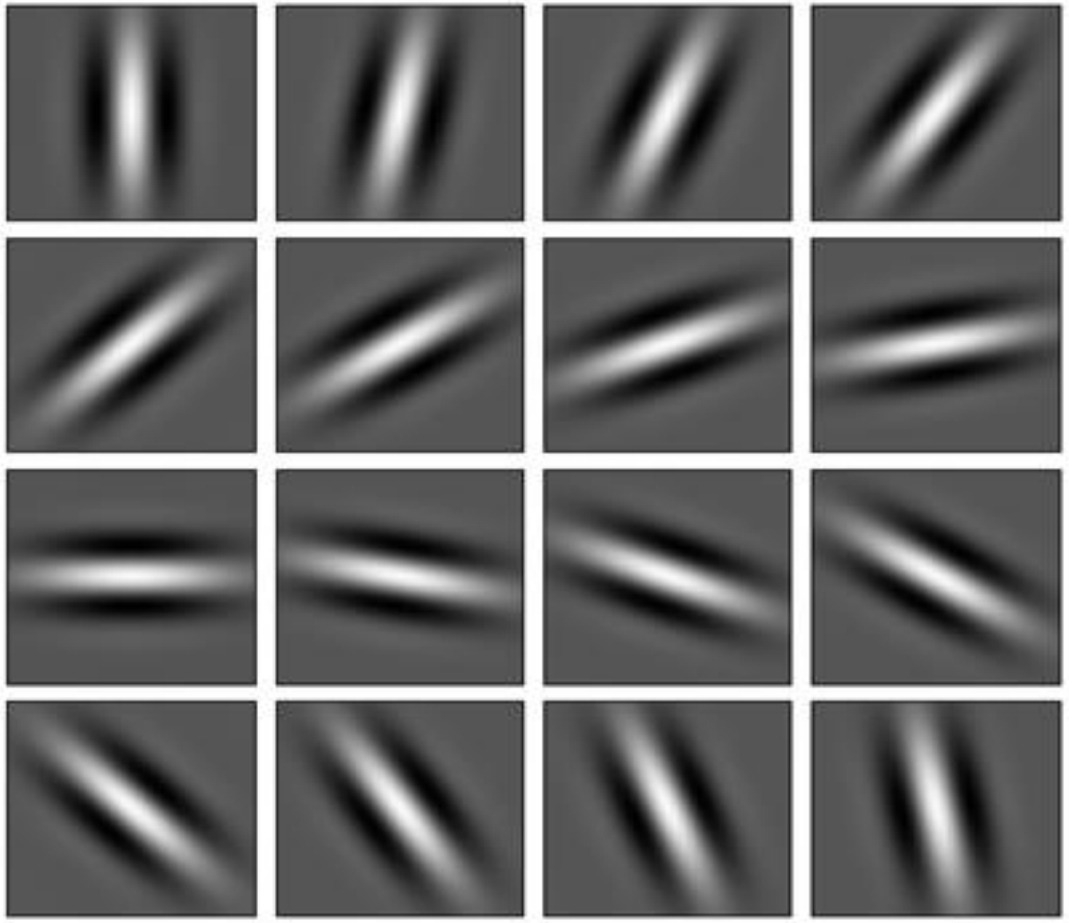
Neighborhood Aggregation

- Mean neighborhood expression and the AGF (Fig. 1a,b) to represent the transcriptomic microenvironment around each cell.
- The AGF (Fig. 1b), which can be thought of as measuring the gradient of gene expression in each cell's neighborhood is invariant to rotation.
- These additional features are used to embed cells in a neighbor-augmented product space (pink blocks under purple).
- Dimensionality reduction applied to this augmented matrix.
- Graph clustering can be performed using any graph partitioning algorithm (leiden/louvain).



Gabor Filters

- Used in computer vision, image processing, and pattern recognition.
- Lens that highlights certain features like edges.
- These are seen at different rotations of the lens.
- Spatial transcriptomics data can be thought of as "images" where the "pixels" are cells, and their "intensities" correspond to gene expression values.
- Spatial patterns and detect regions with specific features correspond to identifying tissue boundaries, spatial domains, or niche-specific cell clusters.



Gabor Filter Continued

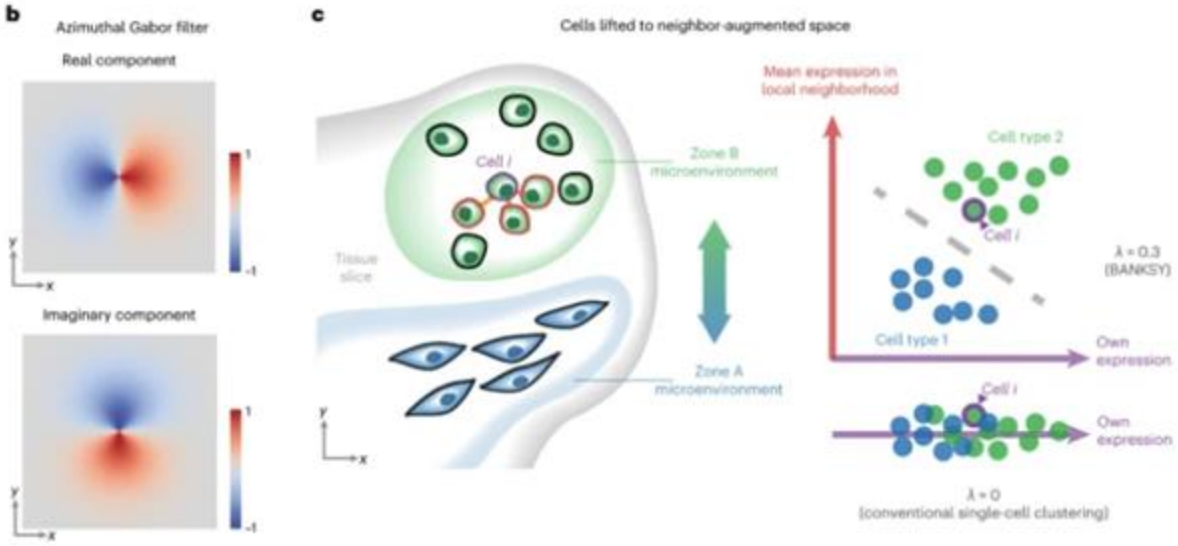
$$\mathrm{AGF}(u,q) = \sum_{v \in N(u)} g_q(v) \cdot \exp\left(i\phi_{uv}\right) \cdot \exp\left(-\frac{r_{uv}^2}{2\sigma^2}\right)$$

Where:

- $N(u)$ = neighborhood of cells around u (often includes 2 × k_{geom} spatial neighbors to have sufficient data for estimating gradients)
- $g_q(v)$ = gene expression of gene q in neighboring cell v
- $\exp(i\phi_{uv})$ = **complex sinusoid** that captures the angular information of the cell v relative to u (just like in a Gabor filter, this term "picks out" features in specific directions)
- $\exp(-r_{uv}^2/2\sigma^2)$ = **Gaussian-like wave modulation** that downweights distant cells (so that closer cells are weighted more heavily)
- σ = controls the scale of the distance-based decay, similar to the bandwidth of a Gaussian kernel

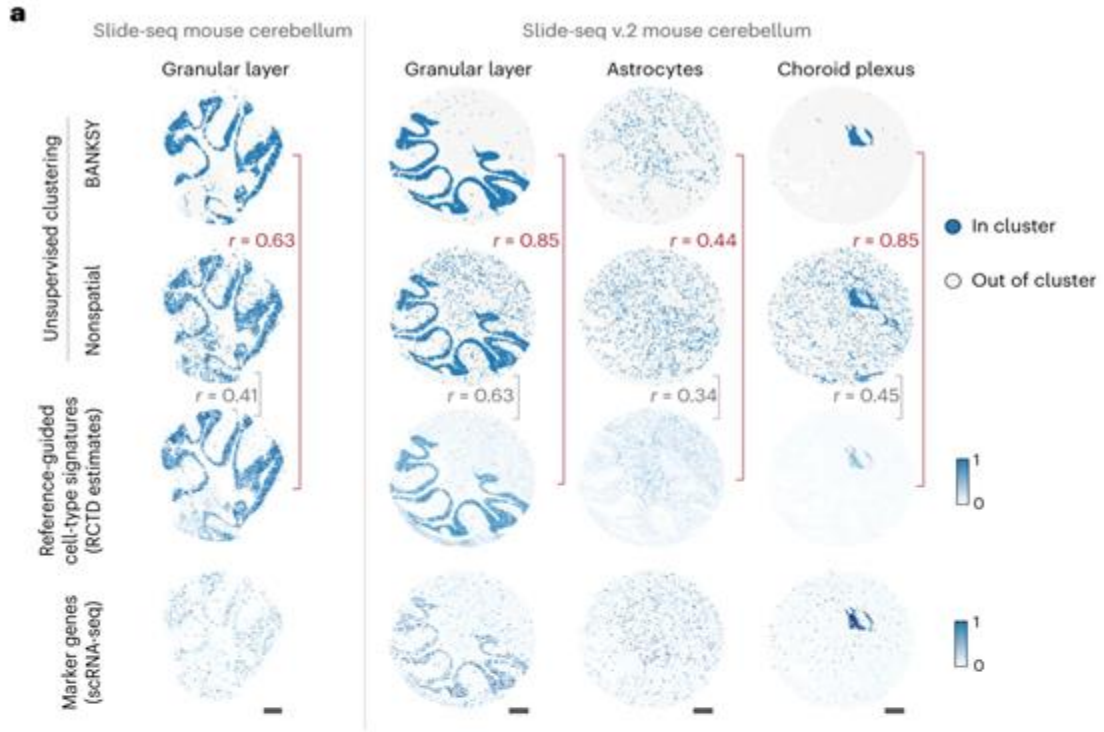
Gene Expression Patterns

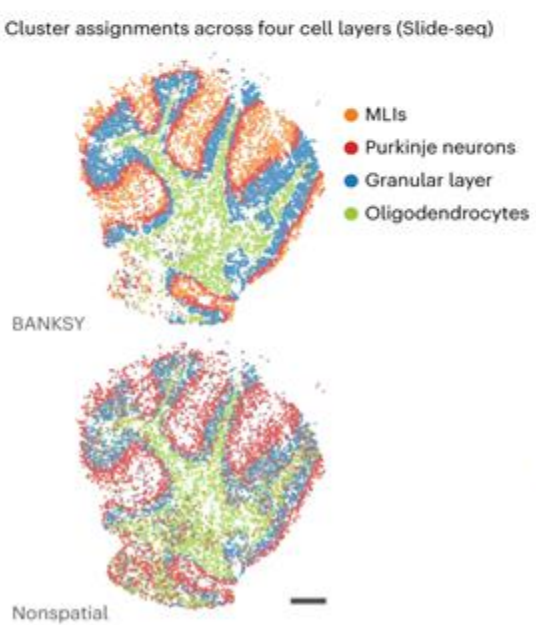
- Textures
  - Diffuse all around center cell.
  - Gradient direction (an edge or polarization of the cells in a block).
  - Concentration of high or low expression in neighboring cells.
  - Waves of expression in different layers of cells.
- Lambda parameter weights the linear combination of all three matrices.

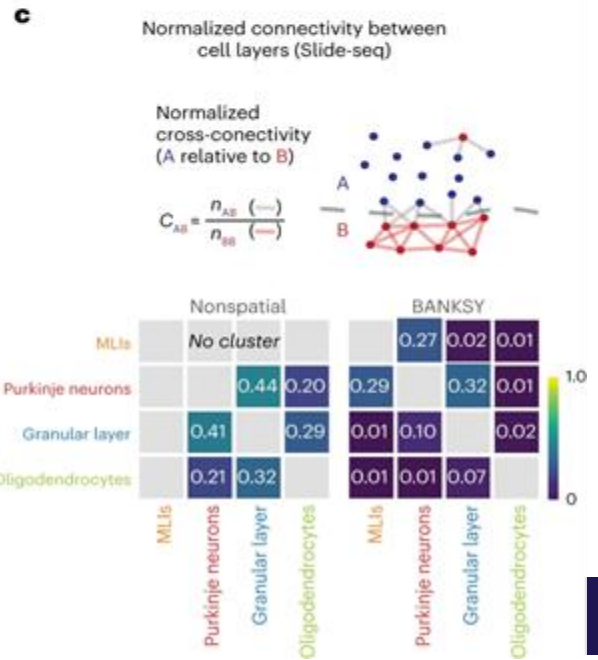


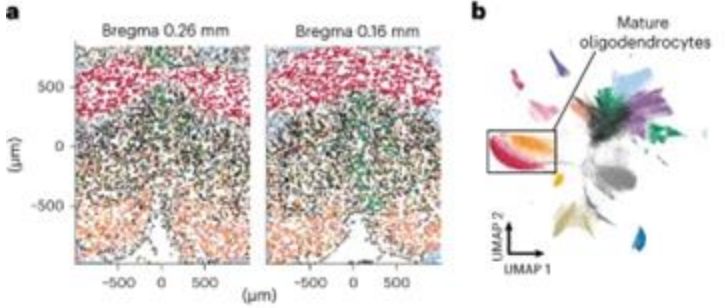
$$\mathcal{B} = \begin{bmatrix} \sqrt{1-\lambda}\,\mathcal{C} \\ \sqrt{\lambda/\mu}\,\mathcal{M} \\ \sqrt{\lambda/(2\mu)}\,\mathcal{G} \end{bmatrix} \in \mathbb{R}^{3p\times N} \quad (2)$$

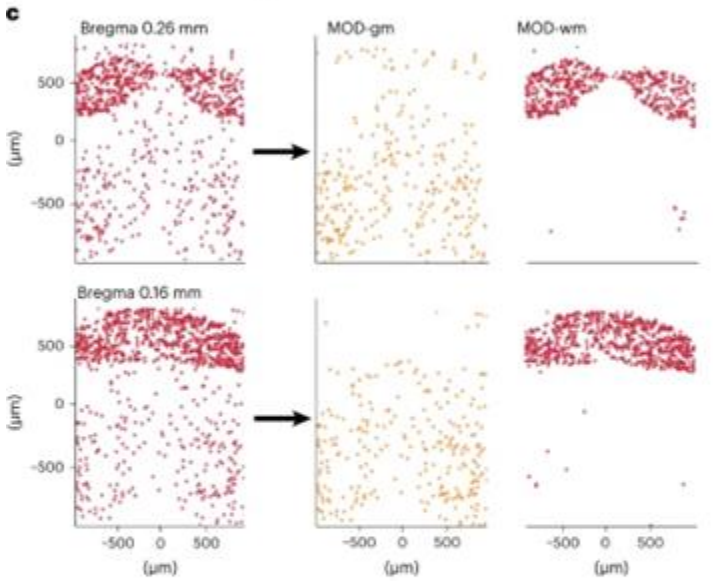
where $\mu = 1.5$ is a normalization factor to ensure the convexity of the linear combination of distance matrices (Supplementary Section 2) and $\lambda \in [0, 1] \subset \mathbb{R}$ is a mixing parameter that controls the relative weights of the three component matrices. For cell typing, we used a default value of $\lambda = 0.2$ to construct the neighbor-augmented matrix $\mathcal{B}$; for domain segmentation, we used $\lambda = 0.8$ for the creation of the corresponding $\mathcal{B}$ matrix. In either case, the subsequent processing of this matrix was similar: PCA (20 principal components as a default) for dimensionality reduction, followed by Leiden clustering for community detection. Finally, we note that at $\lambda = 0$, the algorithm only takes the cells' own expression into account and reduces to nonspatial (conventional) clustering.

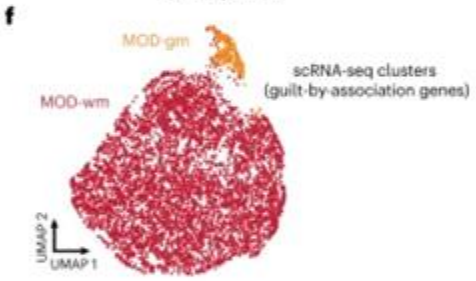

a
Slide-seq mouse cerebellum
Slide-seq v.2 mouse cerebellum
Granular layer
Granular layer
Astrocytes
Choroid plexus
Unsupervised clustering
BANKSY
Nonspatial
Reference-guided cell-type signatures (RCTD estimates)
Marker genes (scRNA-seq)
r = 0.63
r = 0.41
r = 0.85
r = 0.63
r = 0.44
r = 0.34
r = 0.85
r = 0.45
In cluster
Out of cluster
1
0
1
0

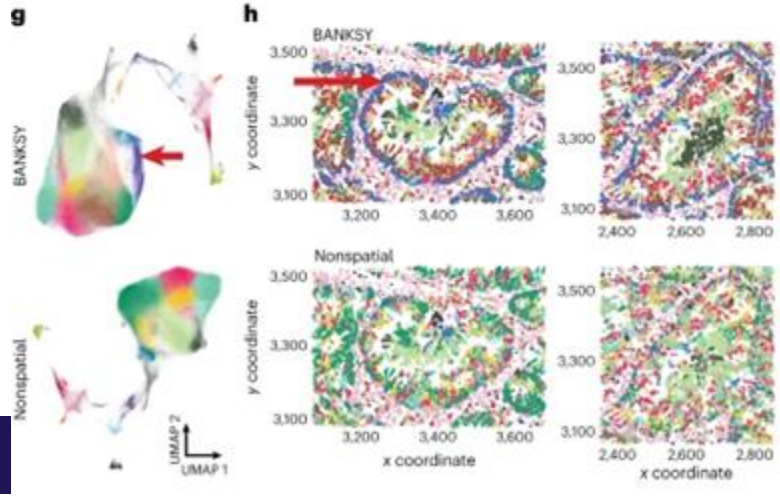

b
Cluster assignments across four cell layers (Slide-seq)
MLIs
Purkinje neurons
Granular layer
Oligodendrocytes
BANKSY
Nonspatial


c
Normalized connectivity between cell layers (Slide-seq)
Normalized cross-conectivity (A relative to B)
A
B
Nonspatial
BANKSY
MLIs
Purkinje neurons
Granular layer
Oligodendrocytes
No cluster
0.44
0.20
0.41
0.29
0.21
0.32
0.27
0.02
0.01
0.29
0.32
0.01
0.01
0.10
0.02
0.01
0.01
0.07
1.0
0


a
Bregma 0.26 mm
Bregma 0.16 mm
(μm)
b
Mature oligodendrocytes
UMAP 2
UMAP 1
d
MOD-wm
MOD-gm
Expression
Mlc1
Gjak
Cldn2
Syt4
Gad1
Plin3
Gnht
Sln
Olig3
Mbp
Lpar1
Tih
Ucn3
Cck
Cells


c
Bregma 0.26 mm
MOD-gm
MOD-wm
Bregma 0.16 mm
(μm)
e
BANKSY's neighbor-augmented space metagene plot
MOD-gm
MOD-wm
Neighbor expression
Own expression


f
MOD-gm
MOD-wm
scRNA-seq clusters
(guilt-by-association genes)
UMAP 2
UMAP 1

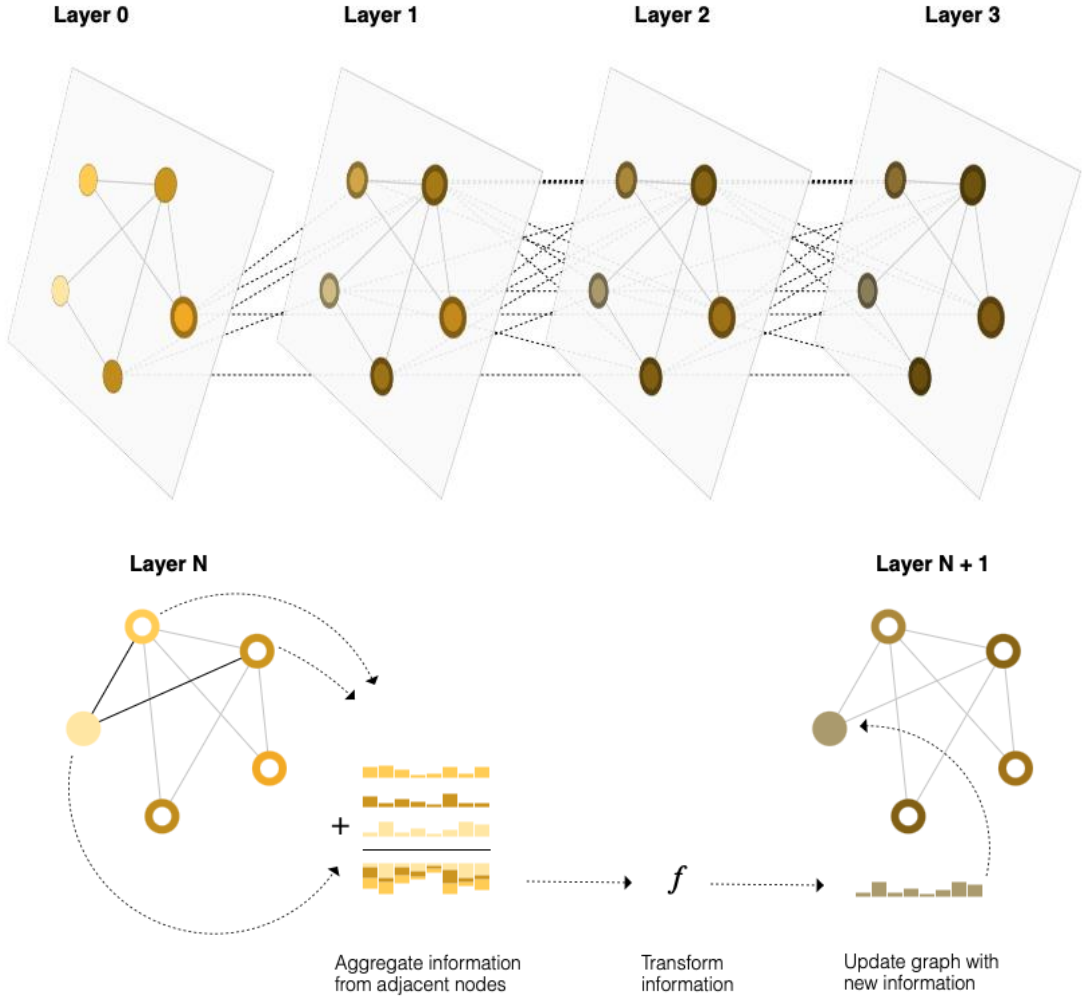

g
BANKSY
Nonspatial
UMAP 2
UMAP 1
h
BANKSY
Nonspatial
y coordinate
x coordinate
i
Cell cycle metagene
UMAP 2
UMAP 1
Cell state
cE01 (TA-like proliferation)
cE02 (stem-like)
cE03 (stem-like)
cE04 (TA-like proliferation)
cE05 (TA-like proliferation)
cE06 (stem/TA-like)
cE07 (enterocytes)
cE08 (cycling)
cE10 (immature goblet)
cE11 (enterocytes)
cE12 (cycling)
cE13 (stem/TA-like)
cI01 (monocytes)
cI02 (granulocytes)
cI03 (macrophage-like)
cI04 (CD4+ T_reg)
cS01 (MMP3+ CAF)
cS02 (GREM1+ CAF)
cS03 (pericytes)
cS04 (endothelial)
cS05 (myofibroblast)
cS06 (CAF stem niche)
j
BANKSY
Nonspatial

Graph Neural Networks

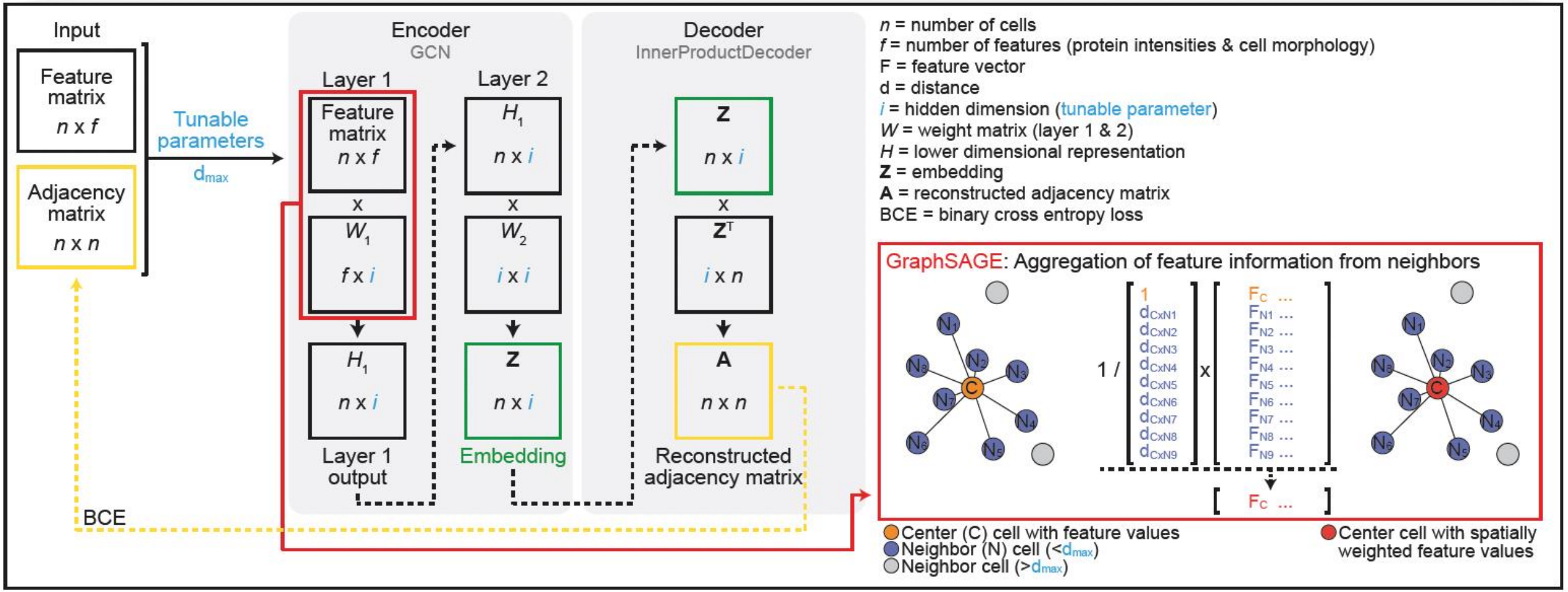
- **Neural nets for graphs**
- Unlike CNNs/RNNs on grids or sequences, GNNs ingest **nodes + edges**, learning representations that respect irregular topology.
- **Message-passing paradigm**
- In each layer, a node gathers (“messages”) features from its neighbors, aggregates them (mean/sum/attention), then updates its own embedding via a small neural network.
- **Layer stacking = multi-hop context**
- Stacking L layers lets each node incorporate information from up to L-steps away, capturing broader structural patterns.
- The learning algorithm optimizes both how you **weigh** neighbor messages (e.g. attention coefficients) and how you **combine** self vs. neighbor info, tailoring the model to your task.

Graph Neural Networks

- **Layer 1 gathers direct neighbors**: each node aggregates features only from its immediate neighbors (1-hop).
- **Layer 2 reaches 2-hop neighbors**: the updated embeddings from Layer 1 include neighbor context, so when Layer 2 aggregates, it effectively pulls in information from neighbors of neighbors.
- **Stacking layers expands receptive field**: an L-layer GNN can incorporate data from up to L-hop neighborhoods around each node.
- **Deeper context at the cost of smoothing**: more hops bring broader information but risk over-smoothing, where node embeddings become too similar across the graph.
- **Design trade-off**: choose layer depth based on how far relational signals (e.g. signaling cascades or spatial niches) truly extend in your tissue graph.



Graph Neural Networks



Need to optimize single nucleus/single cell custom binning using **StarDist** (eosin leaching issue with the high-res image)

10x GENOMICS Products Resources Support Hub Company

Analysis Guides /

Nuclei Segmentation and Custom Binning of Visium HD Gene Expression Data

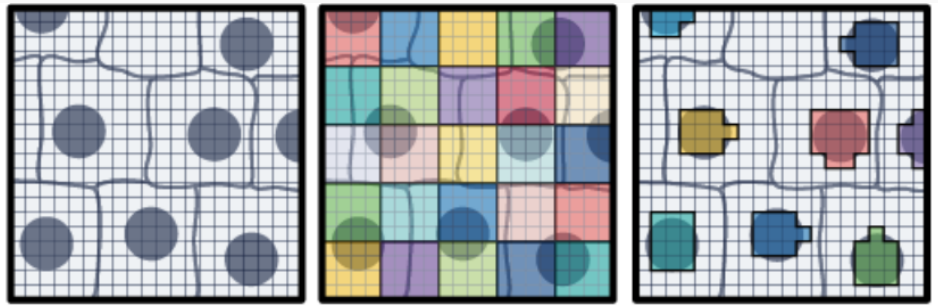


Figure 1: Two approaches for binning the 2x2 μm barcode squares in Visium HD data. The capture area of a Visium HD slide is a continuous grid of barcodes (left panel). The Space Ranger pipeline by default will create square bins (shown here as 8 x 8 μm squares of a single color) that tile the entire tissue containing capture area (middle panel). An alternative is to use the nuclei stain from a high-resolution H&E microscope image to group together barcodes that underlie the same cell nuclei (right panel).

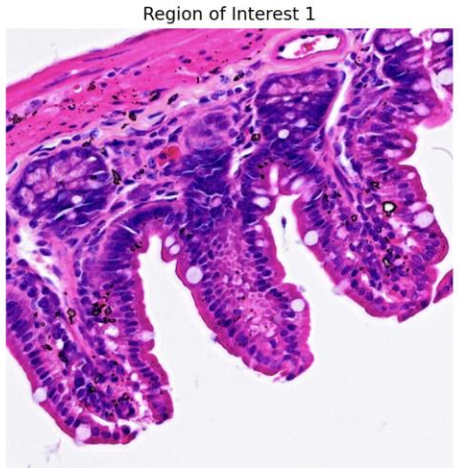
Requirements

- High-resolution H&E image of the sample used in the Visium HD Gene Expression assay
- Output from a completed Space Ranger count analysis
- Familiarity with Python programming

The image taken by the CytAssist is not suitable for this analysis, becuase it is only a cytoplasmic eosin stain and lacks the hematoxylin stain for cell nuclei.

Custom Binning

•**Segment nuclei** on your high-resolution histology image (e.g. with CellProfiler, ilastik, or deep-learning tools like StarDist).
•**Extract centroids** of each segmented nucleus.
•**Build custom bins** around each centroid:
•Define a circular or polygonal region (radius ~5–10 µm) around each nucleus.
•Or generate Voronoi tiles from centroids to partition the tissue.
•**Map reads to bins**:
•Convert Visium barcoded spots (or raw read coordinates) into spatial points.
•For each read/spot, assign it to the nearest custom bin region.
•Aggregate gene counts within each bin.
•**Create a pseudo-single-cell "expression matrix"** where each bin is treated like a cell: rows = genes, columns = bins.
•**Tools & tips**



Creating a Seurat Object: QC and unsupervised clustering (Banksy and Sketching/Projecting)

Visium HD Spatial Gene Expression Library, Human Breast Cancer, IF (FFPE) - 10X genomics

https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-human-breast-cancer-ffpe-if

Visium HD Spatial Gene Expression Library, Human Kidney (FFPE) - 10X genomics

https://www.10xgenomics.com/datasets/visium-hd-cytassist-gene-expression-libraries-human-kidney-ffpe

SpaceRanger Input: high resolution H&E (brightfield) or IF (Fluorescent) image (≈1GB) and FASTQ files (sometimes json files)

SpaceRanger Outputs: can be in customized bins, multiple of 2, up to 100x100µm bin

Show Owner/Mode Show Dotfiles Filter:

Showing 3 rows - 0 rows selected

Type	Name	Size	Modified at
	square_016um	-	12/13/2024 2:02:00 AM
	square_008um	-	12/13/2024 2:01:51 AM
	square_002um	-	12/13/2024 2:01:26 AM

We choose the 8µmx8µm bin to achieve closer to a single cell and since it is the smallest bin that can produce a cLoupe Browser interactive file

Creating and QC-ing a Visium HD Seurat Object

Visium_HD_Human_Breast_Cancer_FFPE - Gene expression library of FFPE Human Breast Cancer (Visium HD) using the Human Whole Transcriptome Probe Set

Summary | Image Alignment | Bin-Level Metrics

Key Metrics

663,857	557.3
Number of 8 µm binned squares under tissue	Mean reads per 8 µm bin
96.9	18,010
Mean UMIs per 8 µm bin	Total genes detected

Mapping

Reads Mapped to Probe Set	97.3%
Reads Mapped Confidently to Probe Set	87.0%
Reads Mapped Confidently to the Filtered Probe Set	84.9%
Reads Half-Mapped to Probe Set	1.0%
Reads Split-Mapped to Probe Set	9.3%

Sequencing

Number of Reads	369,943,706
Valid Barcodes	90.5%
Valid UMIs	99.9%
Sequencing Saturation	77.3%
Q30 Bases in Barcode	91.2%
Q30 Bases in Probe Read	89.7%
Q30 Bases in UMI	90.5%
Fraction of Bins Under Tissue 8 µm	94.5%
UMIs per sq mm of Tissue	1,514,525
Reads per sq mm of Tissue	7,837,798
Fraction Reads in Squares Under Tissue	99.3%

Total UMI Count to Image Alignment

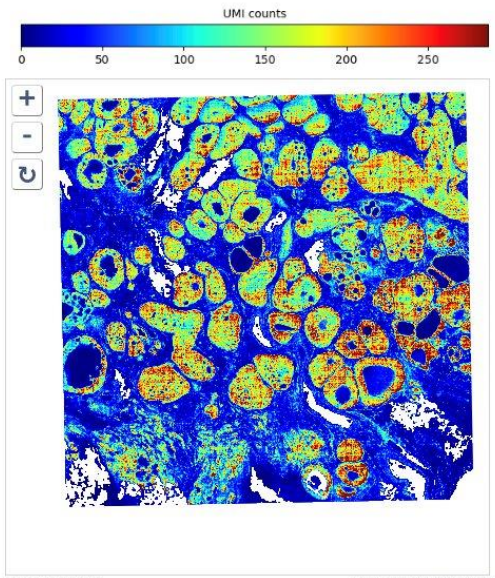
The total UMI count in each 8 µm bin is overlaid onto the tissue image below to assess bin alignment and tissue detection. The highest value on the color scale corresponds to the 98th percentile of UMIs per 8 µm bin under tissue, excluding bins with no UMIs.

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



10x GENOMICS Space Ranger • count

Visium_HD_Human_Breast_Cancer_FFPE - Gene expression library of FFPE Human Breast Cancer (Visium HD) using the Human Whole Transcriptome Probe Set

Summary | Image Alignment | Bin-Level Metrics

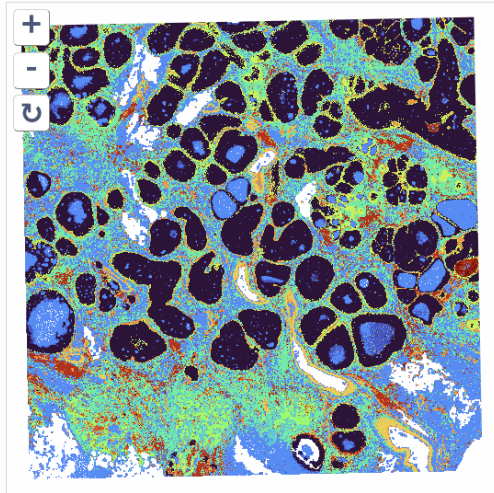
Bin Metrics Overview

Bin Size (µm)	8 µm	16 µm
Number of Bins Under Tissue	663,857	167,545
Mean UMI Counts per Bin	96.9	384.6
Mean Genes per Bin	89.9	327.0

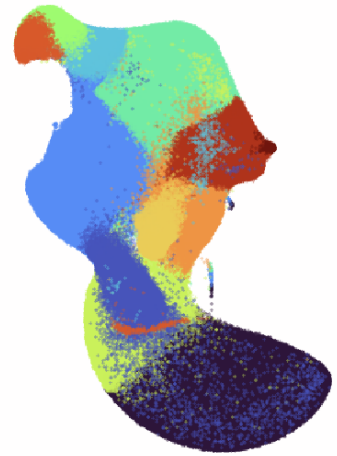
Clustering

8 µm bin | 16 µm bin

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

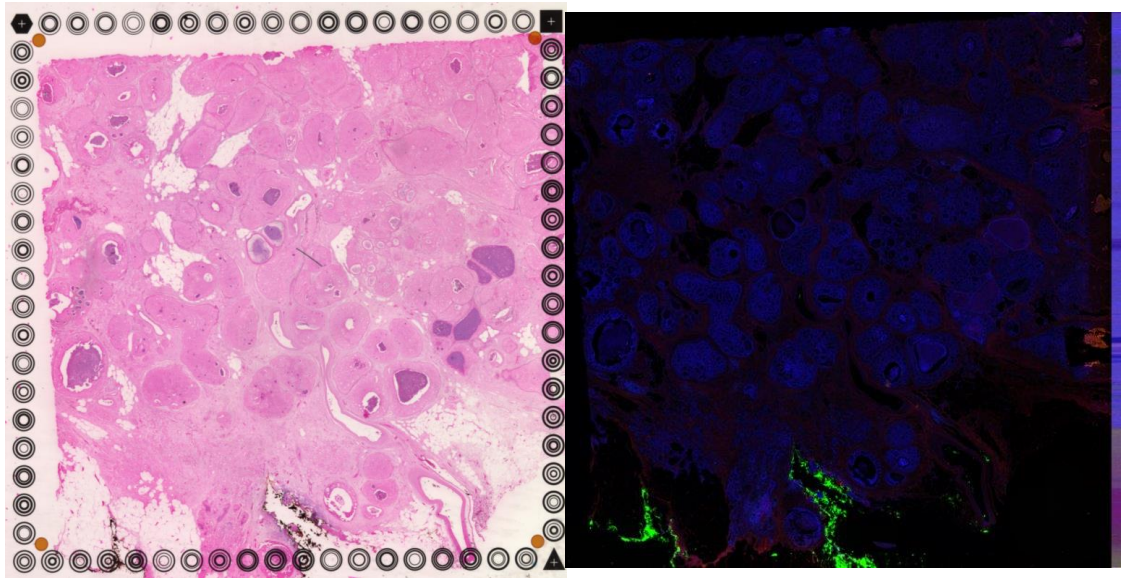


UMAP Projection of 8µm bins colored by Graph-based clustering

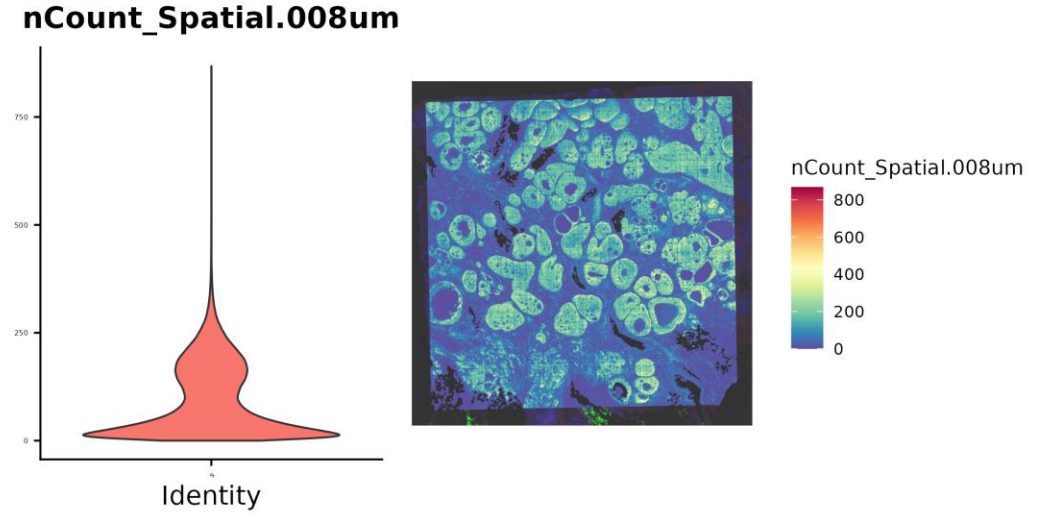


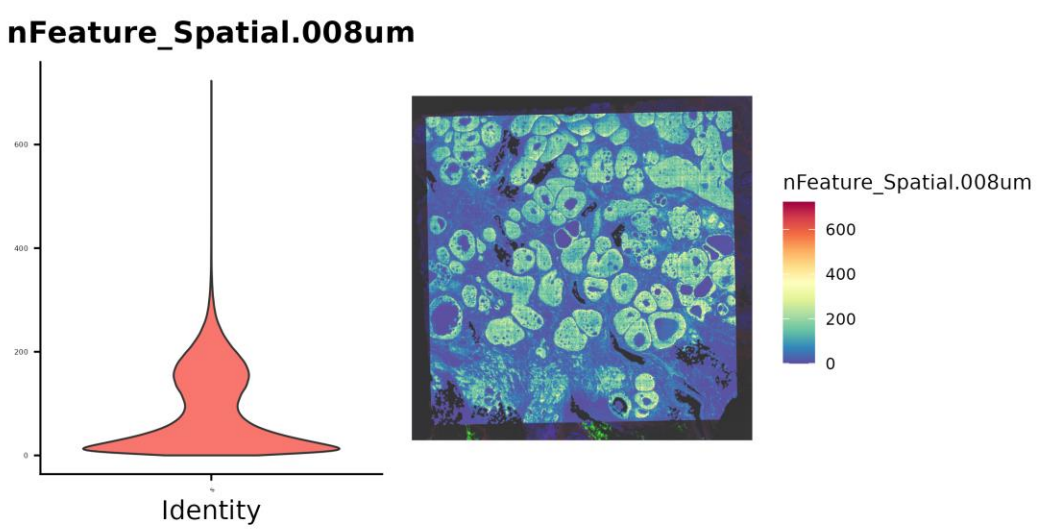
Bin opacity: Display Tissue

```
##############################################
##########    Visium HD Seurat      ###########
##############################################
library(Seurat)

dirs <- c("/data1/shahs3/users/naseran1/work/projects/CRI_Spatial/data/Visium_HD_Breast_dataset_CRI/Output_Files/")

# Load individual sample
object <- Load10X_Spatial(data.dir = dirs, bin.size = c(8))
```



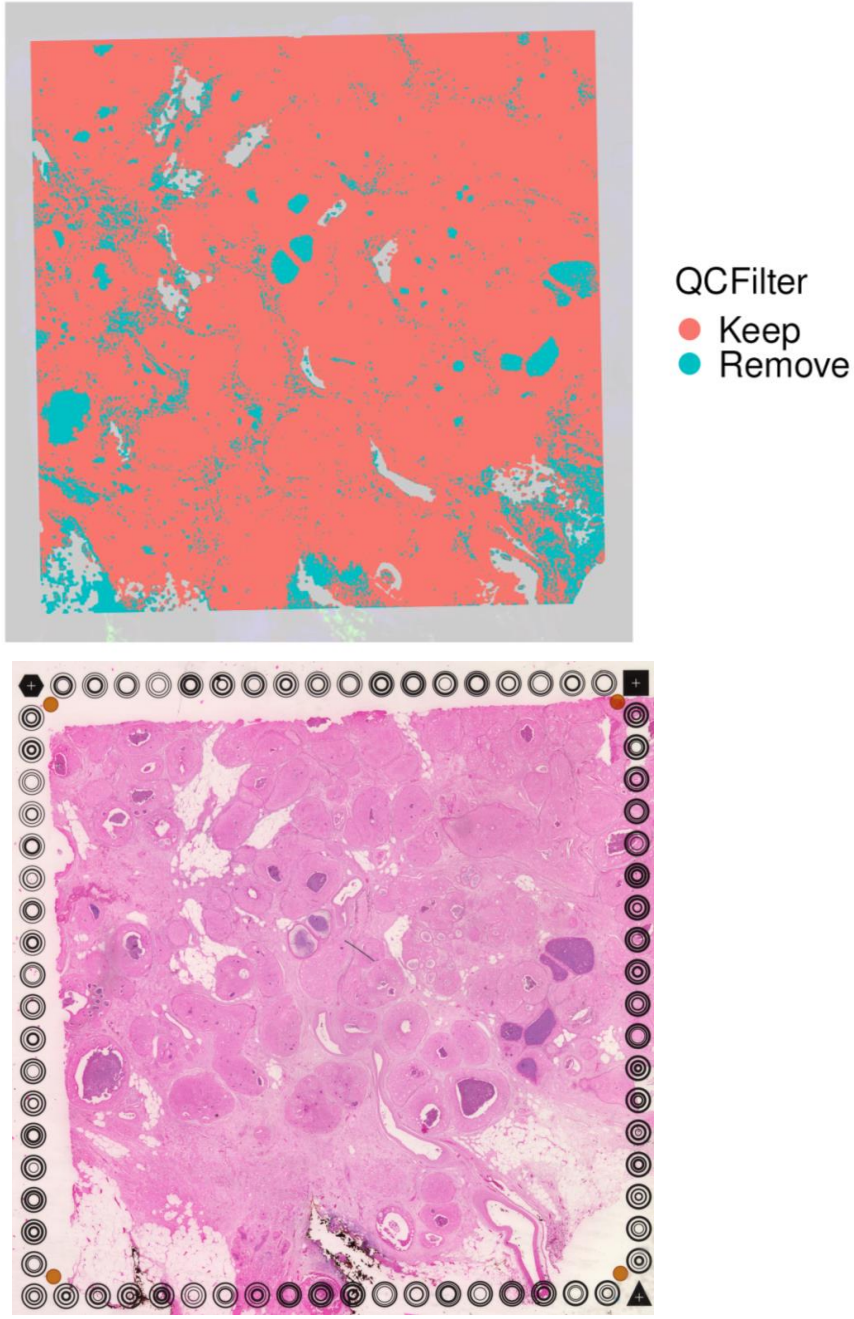
```
############################ View the raw counts and features/genes

vln.plot1 <- VlnPlot(object, features = "nCount_Spatial.008um", pt.size = 0, raster=FALSE)
count.plot <- SpatialFeaturePlot(object, features = "nCount_Spatial.008um", pt.size.factor = 4,
                                 image.alpha = 0.8, alpha=c(1,1))
vln.plot2 <- VlnPlot(object, features = "nFeature_Spatial.008um", pt.size = 0, raster=FALSE)
feature.plot <- SpatialFeaturePlot(object, features = "nFeature_Spatial.008um", pt.size.factor = 4,
                                   image.alpha = 0.8, alpha=c(1,1))

(vln.plot1 | count.plot) / (vln.plot2 | feature.plot)
```





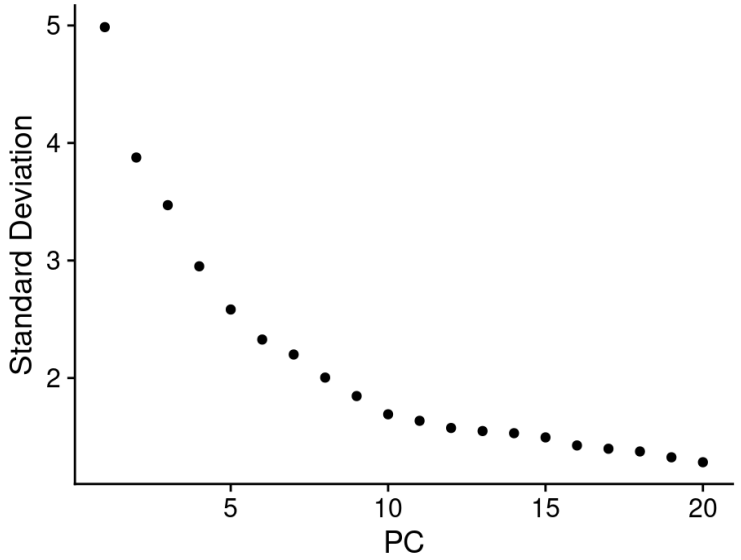
Seurat Object QC

```
##############################################
##########       Object QC          ###########
##############################################

# Add % MT
object[["MT.percent"]] <- PercentageFeatureSet(object, pattern = "^MT-")

# UMI and Gene Threshold
object$QCFilter<-ifelse(object$MT.percent < 25 &
                          object$nCount_Spatial.008um > 10  &
                          object$nFeature_Spatial.008um > 10 ,"Keep","Remove")

cells_plot <- object[[]] %>% filter(QCFilter == "Keep") %>% rownames(.)

p1 <- SpatialDimPlot(object, group.by = "QCFilter", pt.size.factor = 4, image.alpha = 0.2, alpha=c(1,0.2)) +
  theme(
    legend.position = "right",
    legend.key.size = unit(0.3, "cm"),
    plot.title = element_text(size = 16, hjust = 0.5, family = "Arial"),
    legend.text = element_text(size = 14, family = "Arial"),
    axis.text = element_text(size = 14, family = "Arial"),
    legend.title = element_text(size = 14, family = "Arial"),
    axis.title = element_text(size = 14, family = "Arial")) +
  guides(fill = guide_legend(override.aes = list(size = 4)))

ggsave(filename = file.path(sample_dir, "QC.png"), plot = p1, width = 5, height = 5, dpi = 300)

object<-subset(object,cells=colnames(object)[object$QCFilter=="Keep"])
object <- UpdateSeuratObject(object)
```

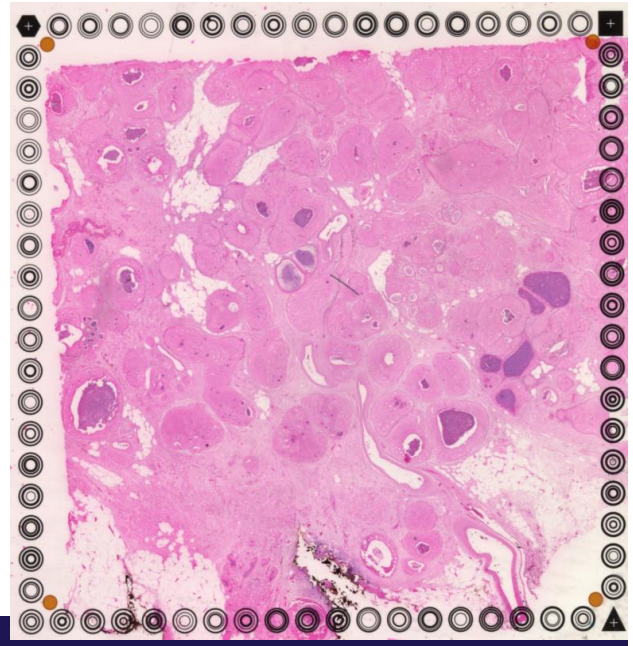


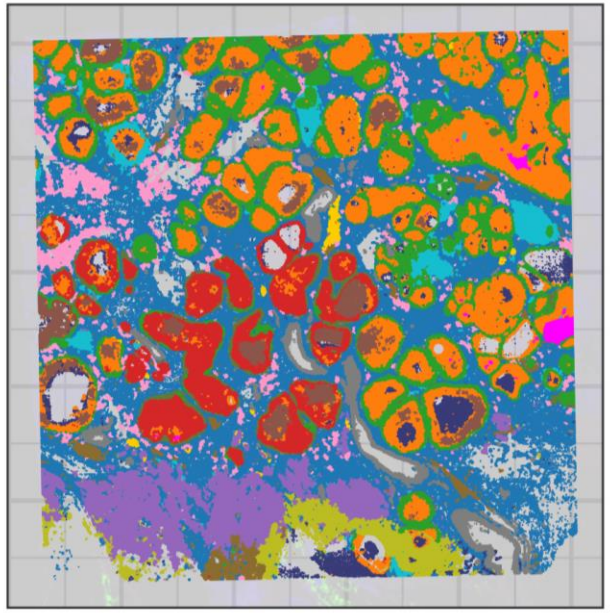
Banksy unsupervised neighborhood clustering (single object)

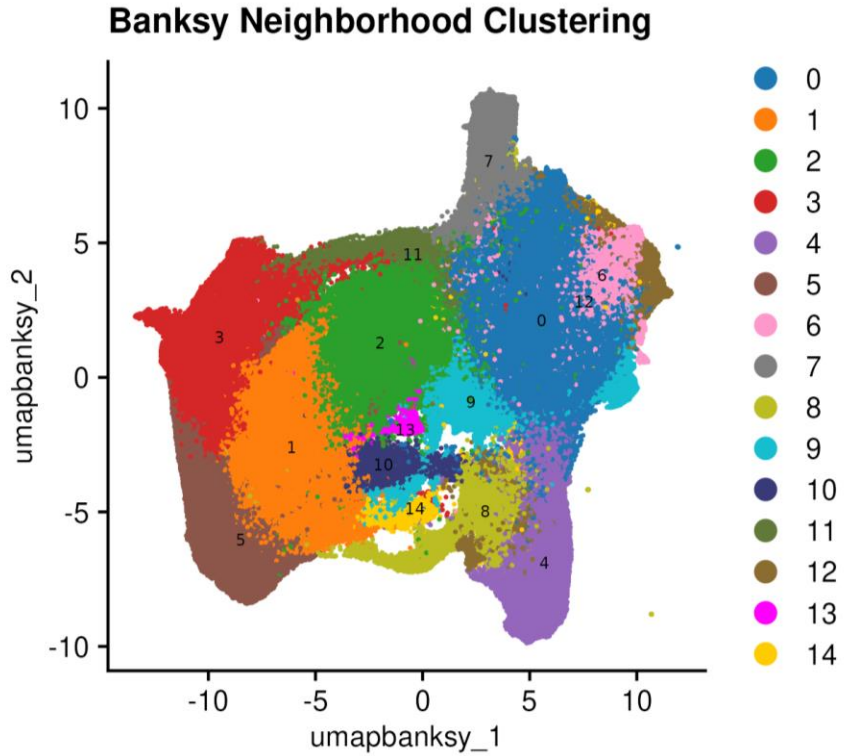
```
##########################################
##########    Banksy Clustering     ############
##########################################

DefaultAssay(object) <- "Spatial.008um"
object <- NormalizeData(object)
object <- FindVariableFeatures(object)

# Run Banksy
object <- RunBanksy(object,
                    lambda = 0.8, verbose = TRUE,
                    assay = "Spatial.008um", slot = "data", features = "variable",
                    k_geom = 24)

DefaultAssay(object) <- "BANKSY"
object <- RunPCA(object, assay = "BANKSY", reduction.name = "pca.banksy", features = rownames(object), npcs = 20)
ElbowPlot(object, reduction = "pca.banksy")
object <- FindNeighbors(object, reduction = "pca.banksy", dims = 1:15)
object <- FindClusters(object, cluster.name = "banksy_cluster", resolution = 0.3)
object <- RunUMAP(object, reduction = "pca.banksy", reduction.name = "umap.banksy", return.model = T, dims = 1:15)
```



Cancer Research Institute™







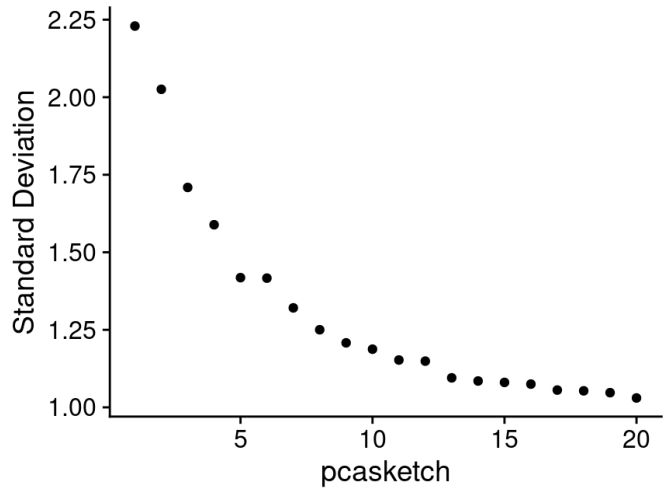
Seurat Unsupervised Clustering: Sketching and Projection (single object)

object	S4 [4000 x 593942] (SeuratOl	S4 object of class Seurat
assays	list [2]	List of length 2
Spatial.008um	S4 [18085 x 593942] (Seurat(	S4 object of class Assay5
BANKSY	S4 [4000 x 593942] (SeuratOl	S4 object of class Assay
meta.data	list [593942 x 8] (S3: data.fra	A data.frame with 593942 rows and 8 columns
active.assay	character [1]	'BANKSY'
active.ident	factor	Factor with 15 levels: "0", "1", "2", "3", "4", "5", ...
graphs	list [2]	List of length 2
neighbors	list [0]	List of length 0
reductions	list [2]	List of length 2
images	list [1]	List of length 1
project.name	character [1]	'SeuratProject'
misc	list [0]	List of length 0
version	list [1] (S3: package_version,	List of length 1
commands	list [6]	List of length 6
tools	list [0]	List of length 0

object	Large Seurat (52.4 GB)

CRI Cancer Research Institute™

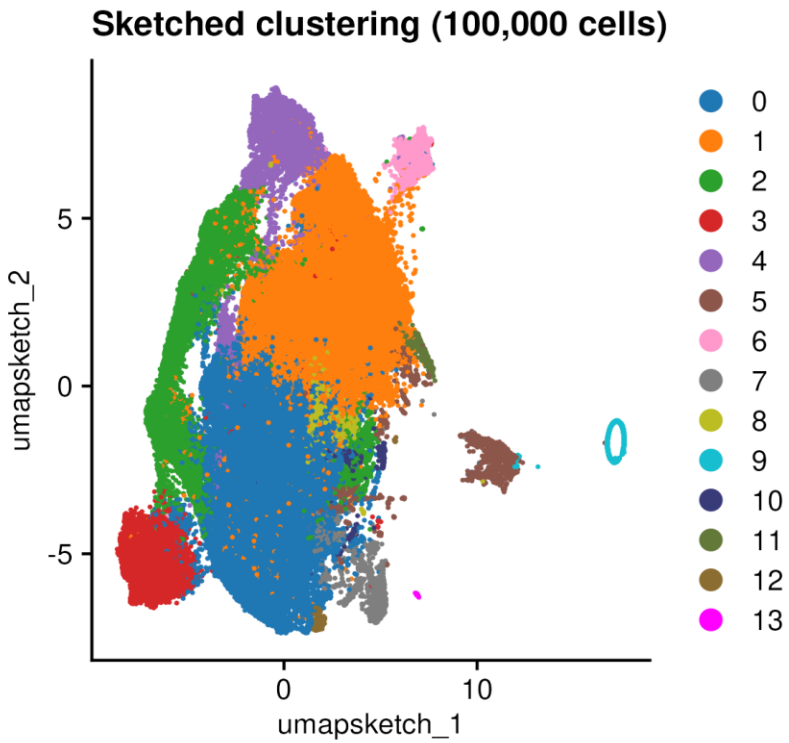
```
##############################################################################
############################################## Sketching followed by Projection ####
##############################################################################

# note that data is already normalized but not scaled
DefaultAssay(object) <- "Spatial.008um"
object <- FindVariableFeatures(object)
object <- ScaleData(object)

# we select 100,0000 cells and create a new 'sketch' assay
object <- SketchData(
  object = object,
  ncells = 100000,
  method = "LeverageScore",
  sketched.assay = "sketch"
)

# switch analysis to sketched cells
DefaultAssay(object) <- "sketch"

# perform clustering workflow
object <- FindVariableFeatures(object)
object <- ScaleData(object)
object <- RunPCA(object, assay = "sketch", reduction.name = "pca.sketch")
ElbowPlot(object, reduction = "pca.sketch")

object <- FindNeighbors(object, assay = "sketch", reduction = "pca.sketch", dims = 1:15)
object <- FindClusters(object, cluster.name = "seurat_cluster.sketched", resolution = 0.2)
object <- RunUMAP(object, reduction = "pca.sketch", reduction.name = "umap.sketch", return.model = T, dims = 1:15)
```

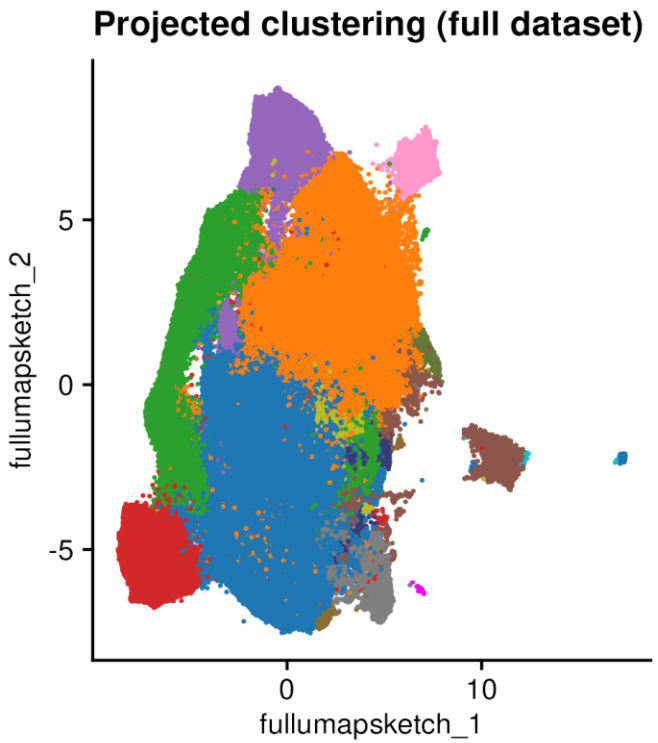


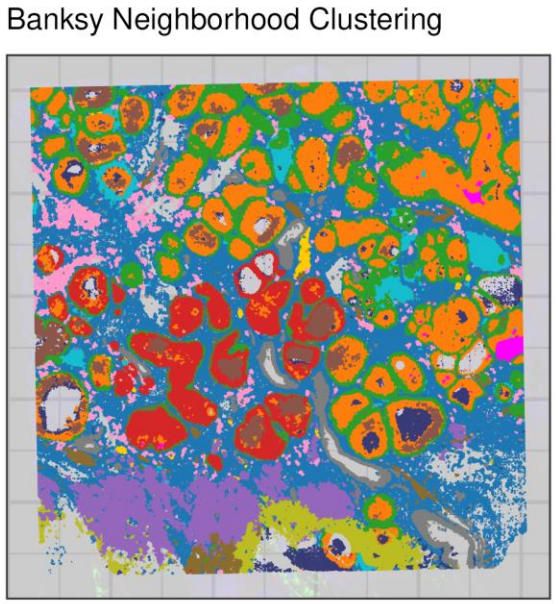
```r
options(future.globals.maxSize = 3 * 1024^3)

object <- ProjectData(
  object = object,
  assay = "Spatial.008um",
  full.reduction = "full.pca.sketch",
  sketched.assay = "sketch",
  sketched.reduction = "pca.sketch",
  umap.model = "umap.sketch",
  dims = 1:15,
  refdata = list(seurat_cluster.projected = "seurat_cluster.sketched")
)
```

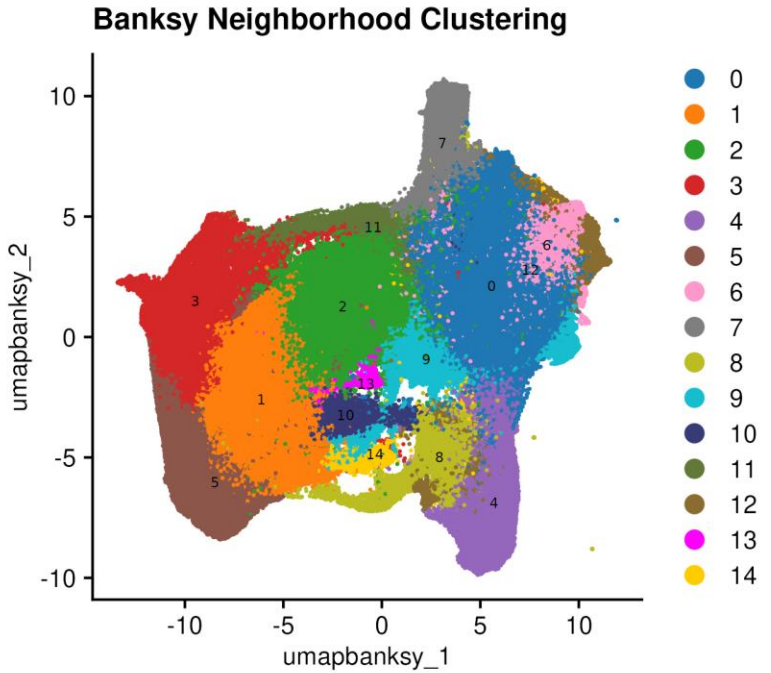
object	S4 [18085 x 593942] (Seurat(	S4 object of class Seurat
assays	list [3]	List of length 3
Spatial.008um	S4 [18085 x 593942] (Seurat(	S4 object of class Assay5
BANKSY	S4 [4000 x 593942] (SeuratOl	S4 object of class Assay
sketch	S4 [18085 x 1e+05] (SeuratO	S4 object of class Assay5

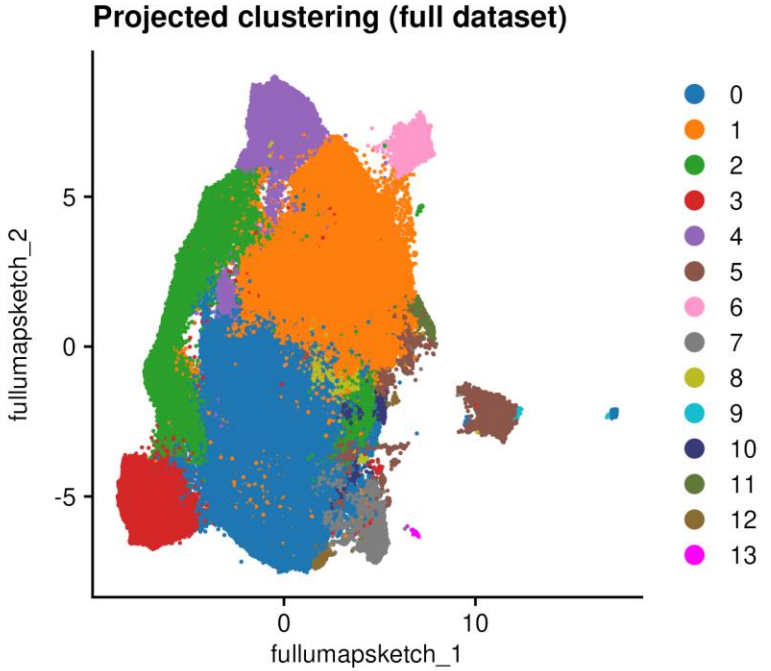
object	Large Seurat	(53.5 GB)











Banksy vs Sketching and Projection:

Banksy unsupervised clustering takes into account neighborhood, shared transcriptional signatures, and rare cell types

Seurat's built-in unsupervised clustering only takes into account shared transcriptional signatures only

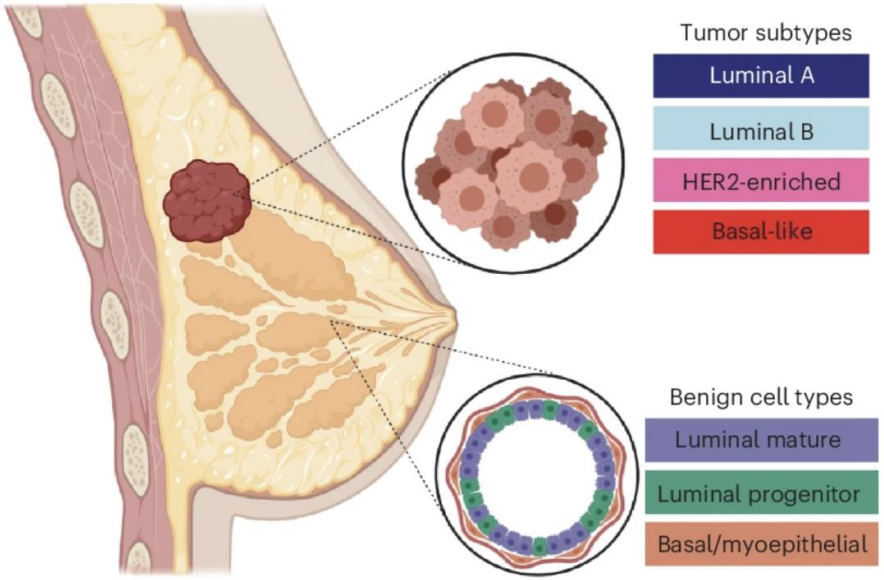
Cell Typing, CNV, and Trajectory Analysis

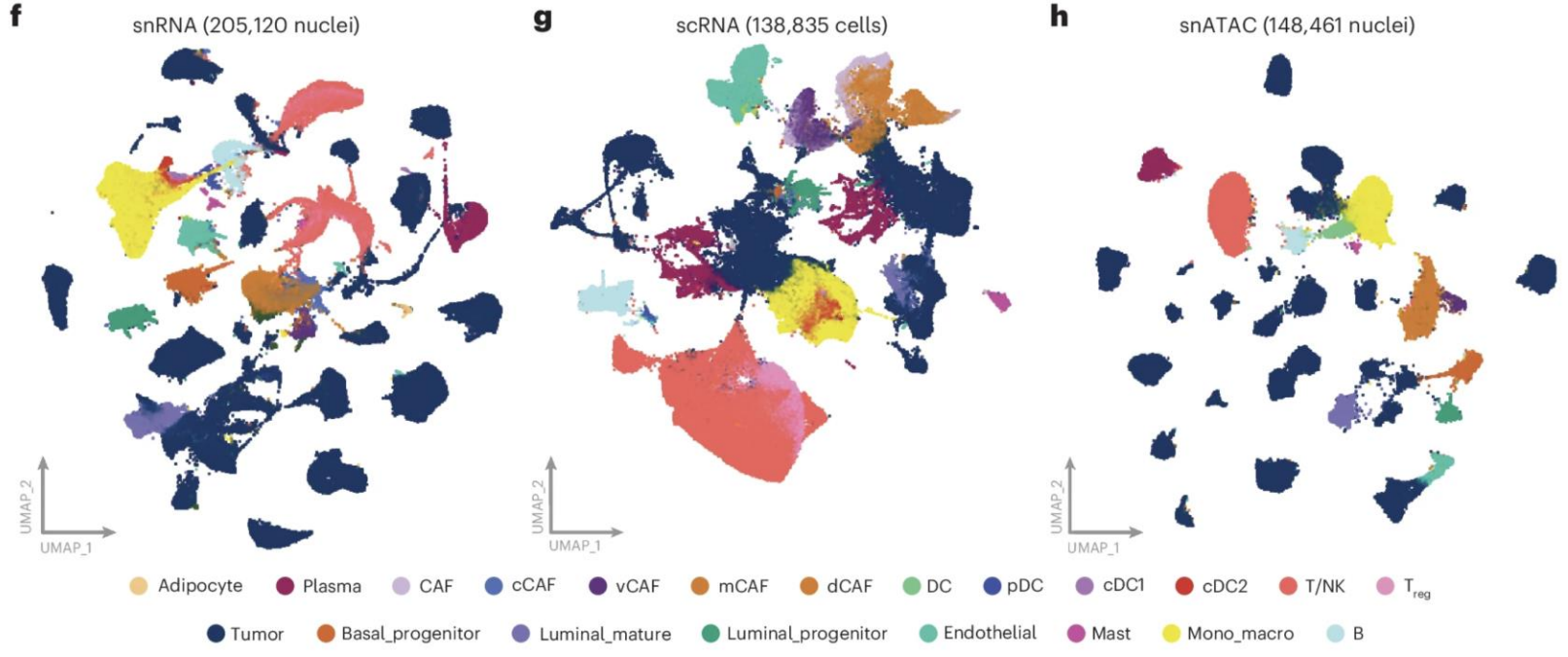
Gene-set Supervised Cell Type Annotation

Differential chromatin accessibility and transcriptional dynamics define breast cancer subtypes and their lineages

Michael D. Iglesia, Reyka G. Jayasinghe, Siqi Chen, Nadezhda V. Terekhanova, John M. Herndon, Erik Storrs, Alla Karpova, Daniel Cui Zhou, Nataly Naser Al Deen, Andrew T. Shinkle, Rita Jui-Hsien Lu, Wagma Caravan, Andrew Houston, Yanyan Zhao, Kazuhito Sato, Preet Lal, Cherease Street, Fernanda Martins Rodrigues, Austin N. Southard-Smith, André Luiz N. Targino da Costa, Houxiang Zhu, Chia-Kuei Mo, Lisa Crowson, Robert S. Fulton, ... Li Ding + Show authors

Nature Cancer **5**, 1713–1736 (2024) | Cite this article



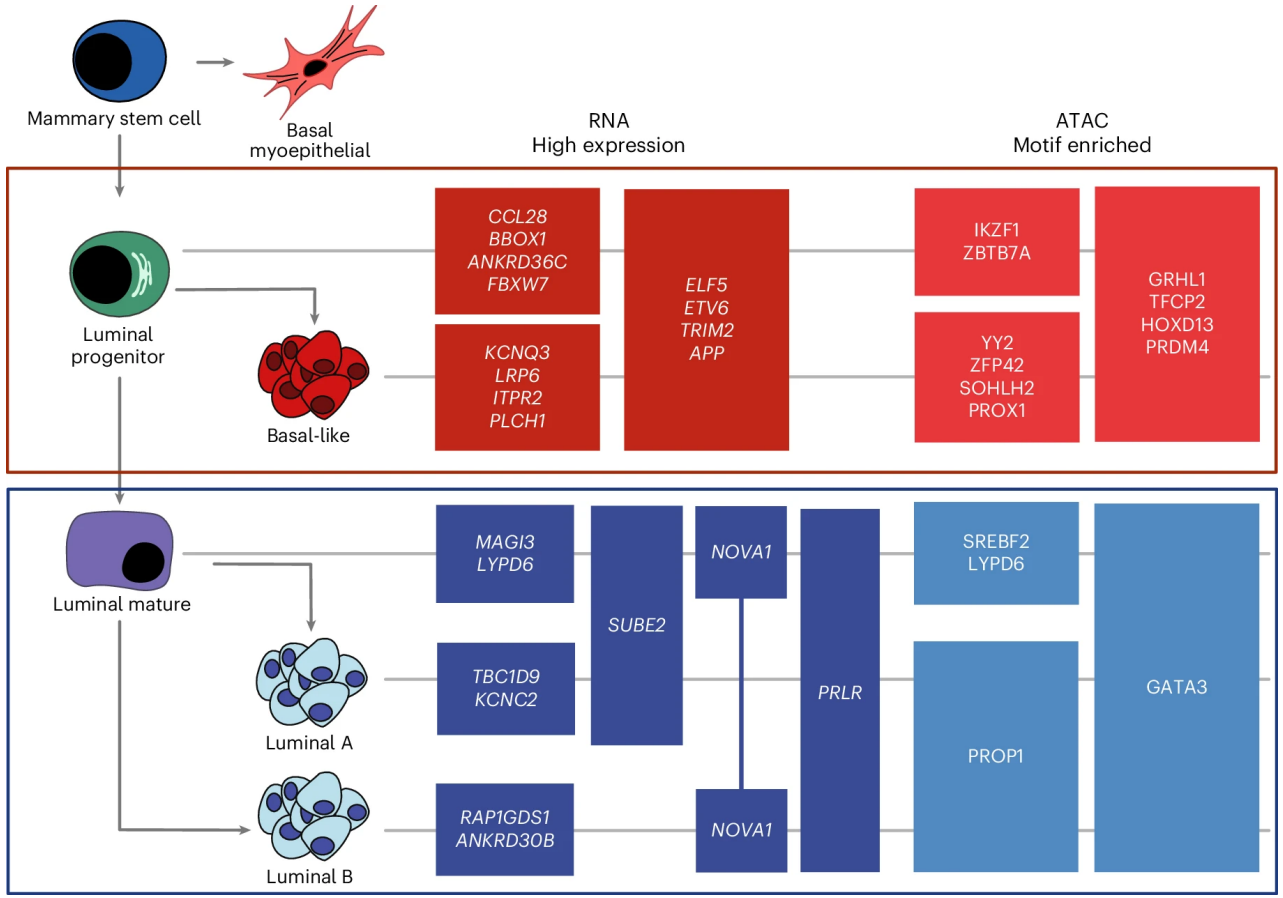


Input = Supervised gene lists

Output= Seurat Object (.RDS)

Step3: Supervised cell type assignment using well defined gene sets with AddModuleScore

```
## Breast cancer gene set from Iglesia MD ... Naser Al Deen N et al. Nature Cancer. 2024 Nov;5(11):1713-36.

########################################################################
#######################   DCIS Major Cell types   ######################
########################################################################

library(writexl)
library(openxlsx)
library(readxl)

gene_lists <- list(
  CD8.T.cells = c("CD8B", "CD8A", "CD3E", "CD3D"),
  CD4.T.cells = c("CD4", "CD3E", "CD3D", "SELL", "CCR7", "IL7R", "TCF7", "LEF1"),
  NK.cells = c("XCL2", "XCL1", "KLRF1", "KIR2DL3", "IL2RB", "CD7", "KLRB1", "KLRD1",
               "GZMA", "PRF1", "CD160", "NCAM1"),
  B.cell = c("BANK1", "CD79A", "CD74", "MS4A1", "MEF2C", "CD19", "CD79B"),
  Plasma = c("IGKC", "IGLC3", "IGLC2", "IGHG1", "IGHA1", "IGHG3", "IGHG4", "IGHA2","FCRL5", "TNFRSF17"),
  Mast.cells = c("TPSB2", "TPSAB1", "CPA3", "MS4A2", "HPGDS", "KIT", "ENPP3"),
  Macrophages = c("ITGAM", "LGALS3", "CD68", "CD163", "LYZ", "ADGRE1", "LAMP2"),
  Monocytes = c("CD14", "FCGR3A", "FCGR1A"),
  DCs = c("PTGDS", "FCHSD2", "GPR183", "NR3C1", "TCF4"),
  Endothelial.cells = c("FABP4", "VWF", "ACKR1", "LDB2", "PECAM1"),
  CAF = c("CFD", "DCN", "GSN", "EBF1", "PRKG1"),
  Basal_Myoepithelial = c("KRT14", "DST", "MMP7", "MIR205HG", "MT1X", "OXTR", "KRT17", "FST"),
  Epithelial.cells = c("EPCAM", "AMBP", "MUC1"),
  Fibroblasts = c("TIMP1", "FN1", "POSTN", "ACTA2", "BST2", "LY6D", "COL6A1", "SLC20A1","COL6A2", "KRT16", "CD9",
                  "S100A4", "EMP1", "LRRC8A", "EPCAM", "PDPN", "ITGB1", "PDGFRA", "THY1"),
  Adipocytes = c("PNPLA2", "CAV1", "FABP4", "PPARG", "CEBPA", "LEP", "CIDEA", "SHOX2", "SLC7A10", "SLC36A2", "P2RX5"),
  Tumor = c("MYC", "ESR1", "AR", "PGR", "CDH1", "AKT1", "ERBB2", "EPCAM", "KRT8","KRT18", "KRT19"),
  Luminal.progenitor = c("MGP", "SCGB2A2", "SLPI", "LTF", "PTN", "KIT", "ALDH1A3"),
  Luminal.mature = c("MUCL1", "MGP", "ERBB4", "ANKRD30A", "AZGP1", "AGR2", "STC2"))

# Convert the list into a data frame
genes_df <- do.call(rbind, lapply(names(gene_lists), function(cell_type) {
  data.frame(Gene = gene_lists[[cell_type]], Cell.type = cell_type, stringsAsFactors = FALSE)
}))

# Write the data frame to an Excel file
write_xlsx(genes_df, "/data1/shahs3/users/naseran1/work/projects/CRI_Spatial/objects/DCIS_genes.xlsx")
```



Iglesia MD et al. Nature Cancer. 2024 Nov;5(11):1713-36.

Supervised (gene-list informed) cell type annotation using AddModuleScore

CRI Cancer Research Institute™

CD8.T.cells

CD4.T.cells

B.cell

Plasma

Monocytes

Macrophages

DCs

Endothelial.cells

CAF

Basal_Myoepithelial

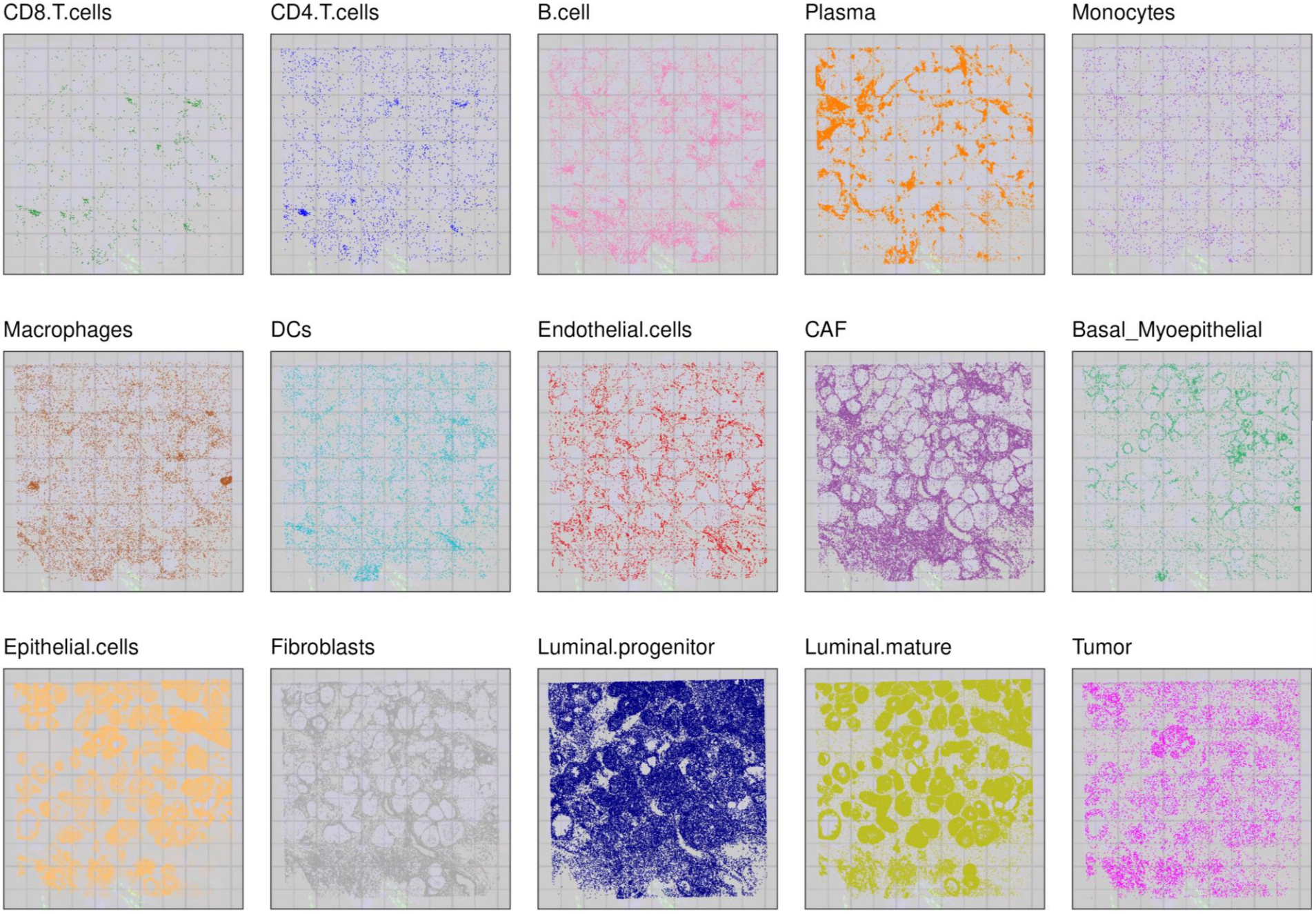
Epithelial.cells

Fibroblasts

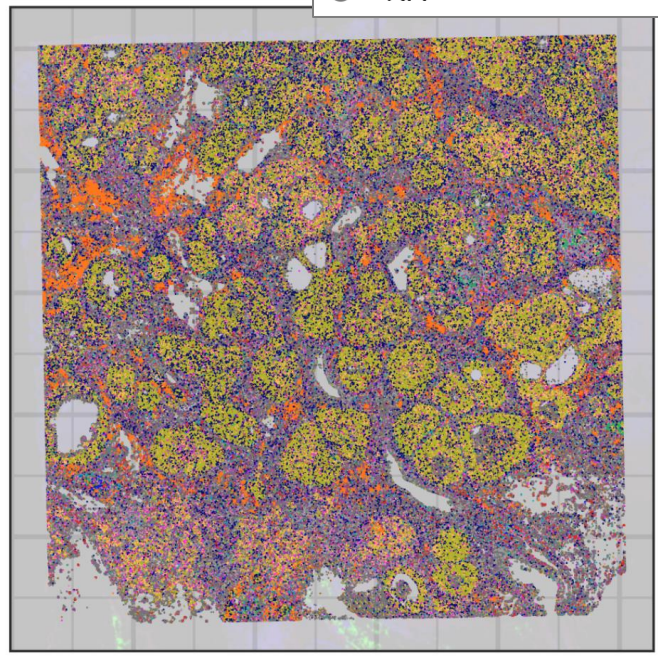
Luminal.progenitor

Luminal.mature

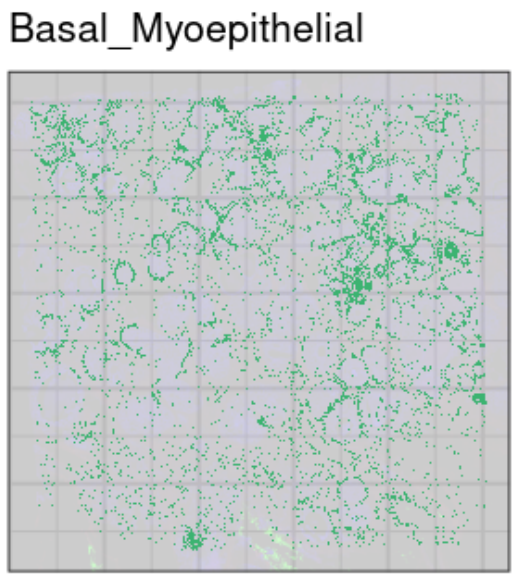
Tumor

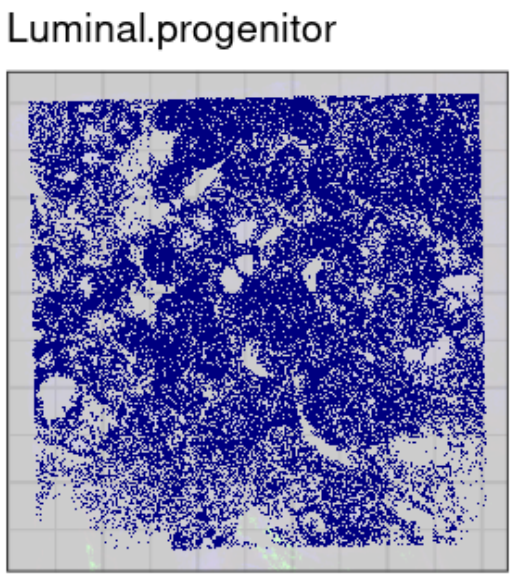


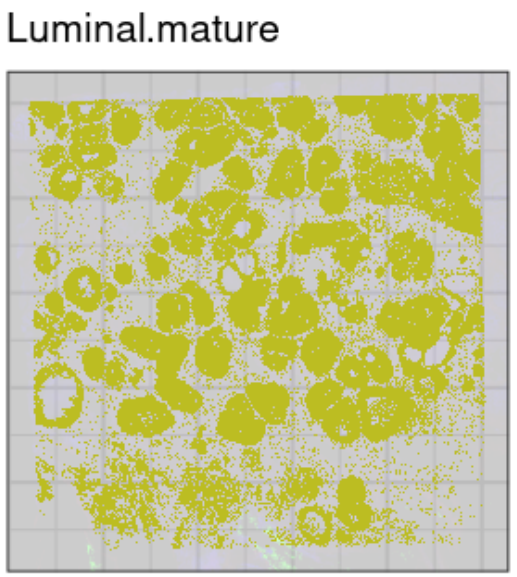
Major Cell Types

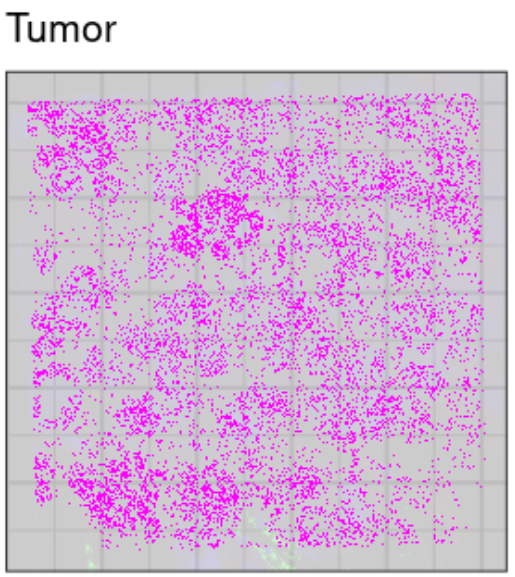


Subsetting into epithelial cells only and re-clustering based on Banksy neighborhood

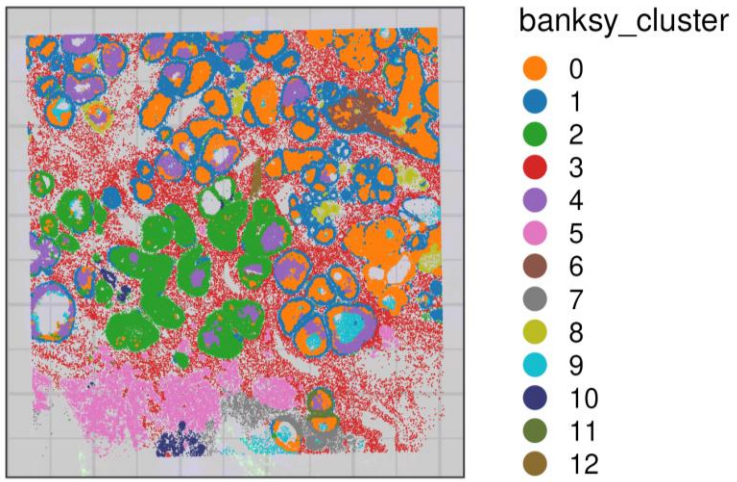


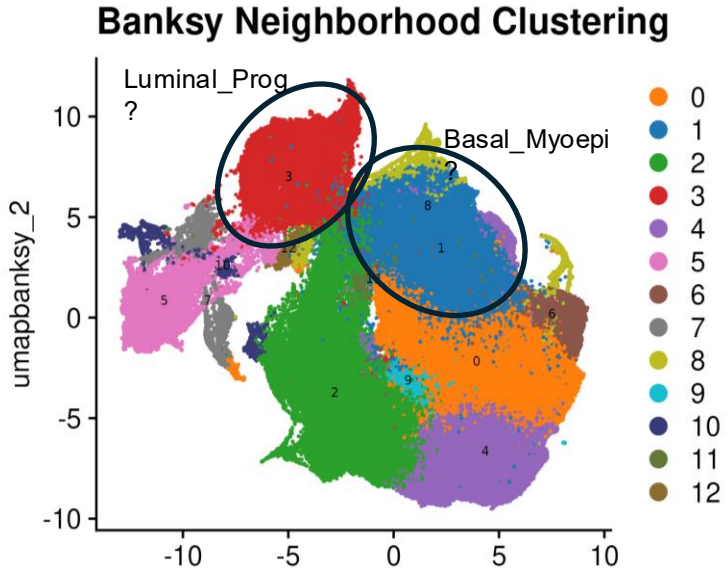


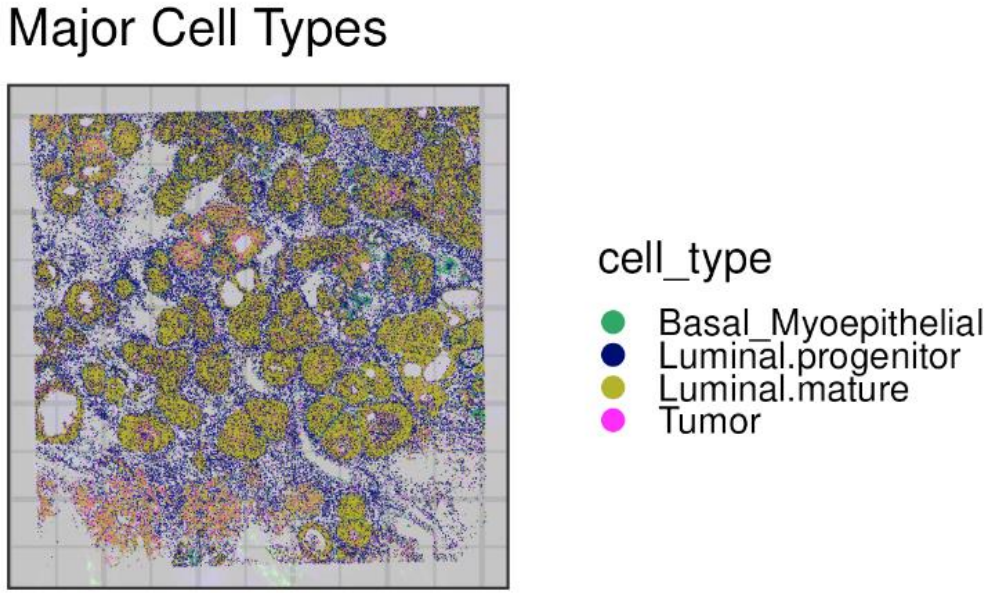




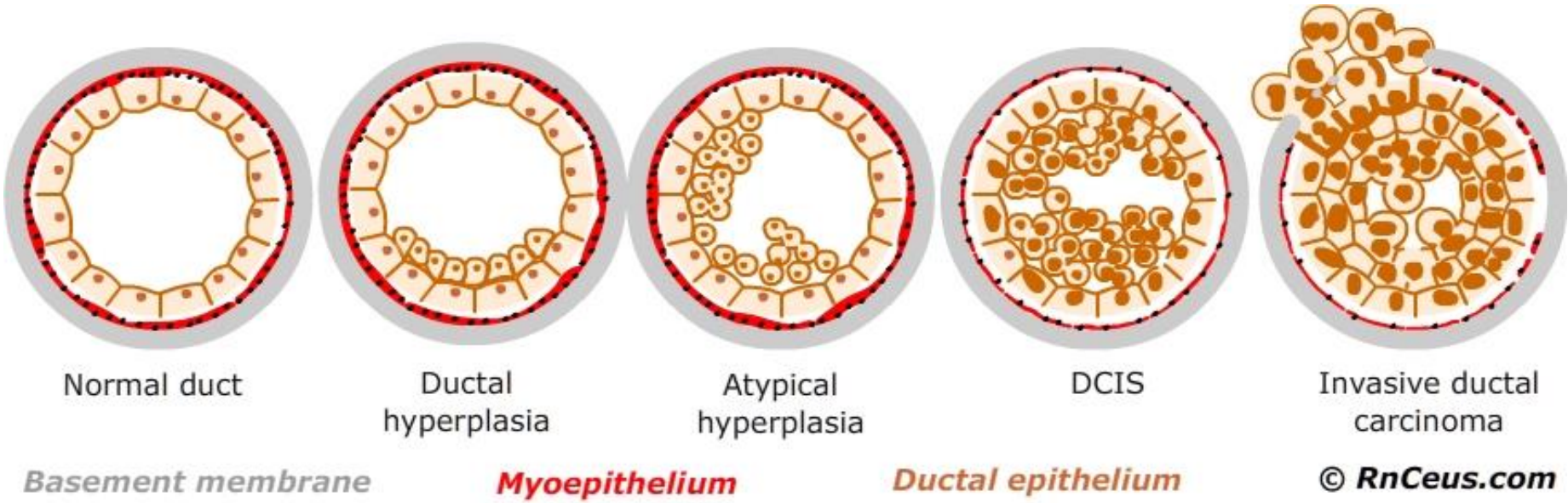
The rest is a mix of all cell types except for the Basal_Myoepithelial – Is cluster 5 IDC (doesn't seem to have an intact myoepithelium) - check DEGs and hallmark of cancer pathways





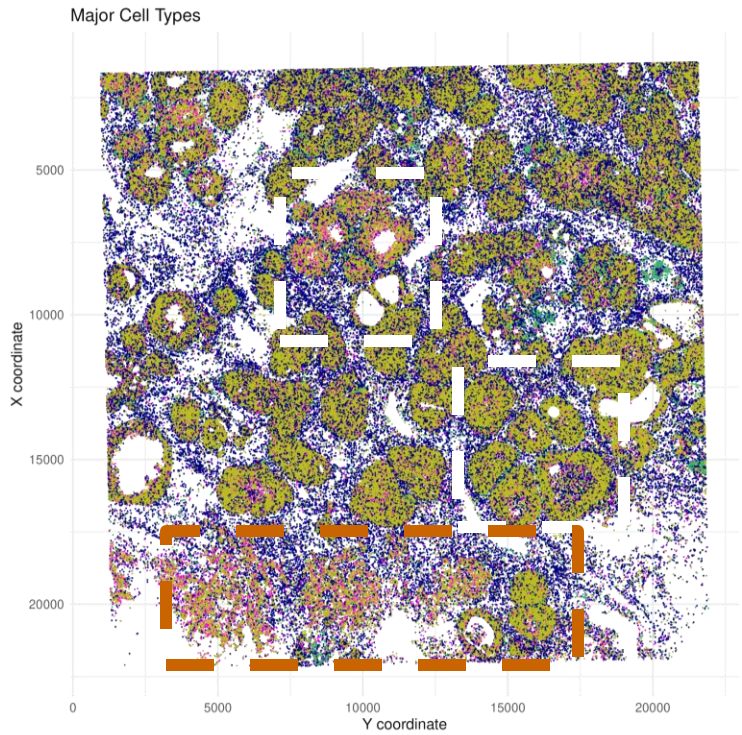


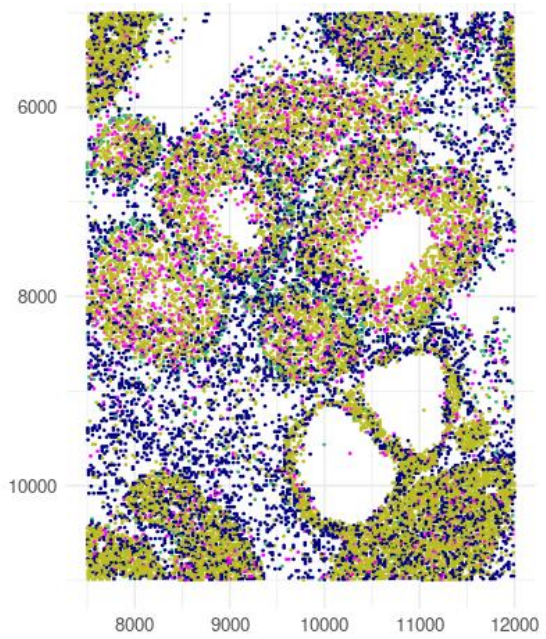
Ductal carcinoma in situ progression

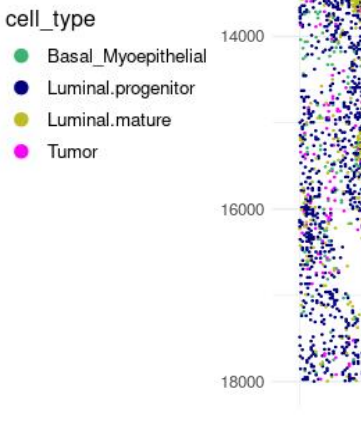


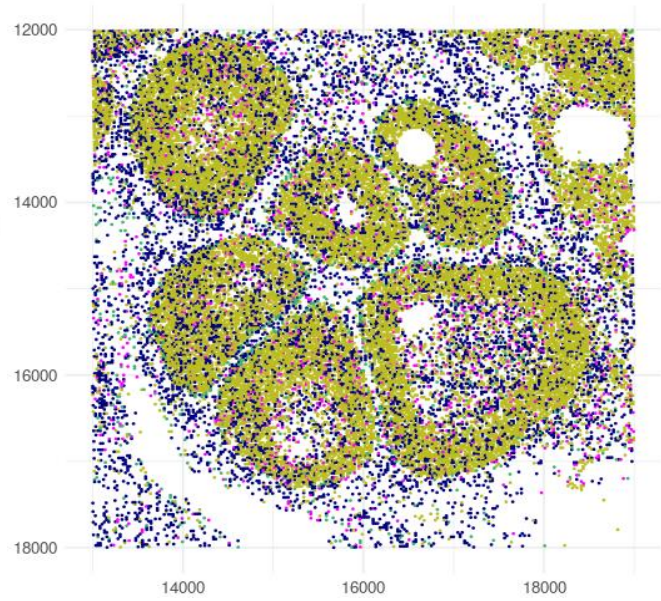
Zooming into certain regions of interest (ROIs)

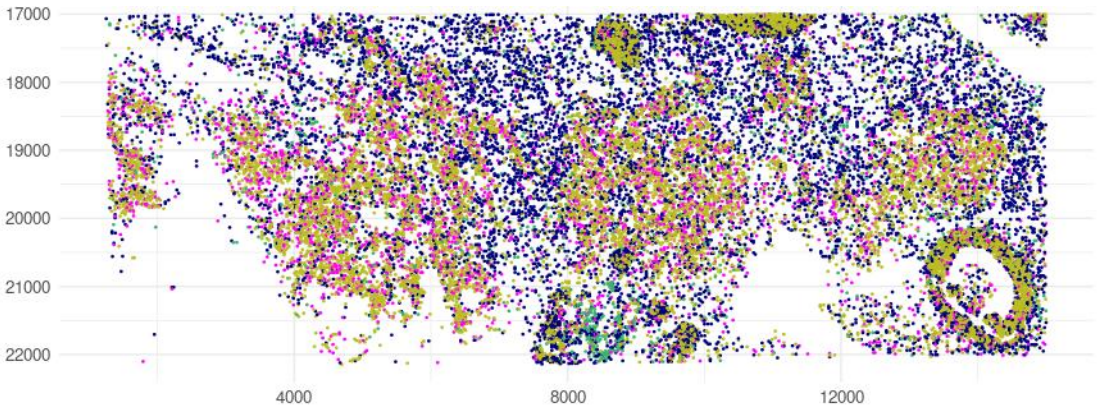
coordinates <- GetTissueCoordinates(object)

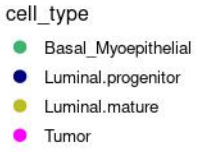


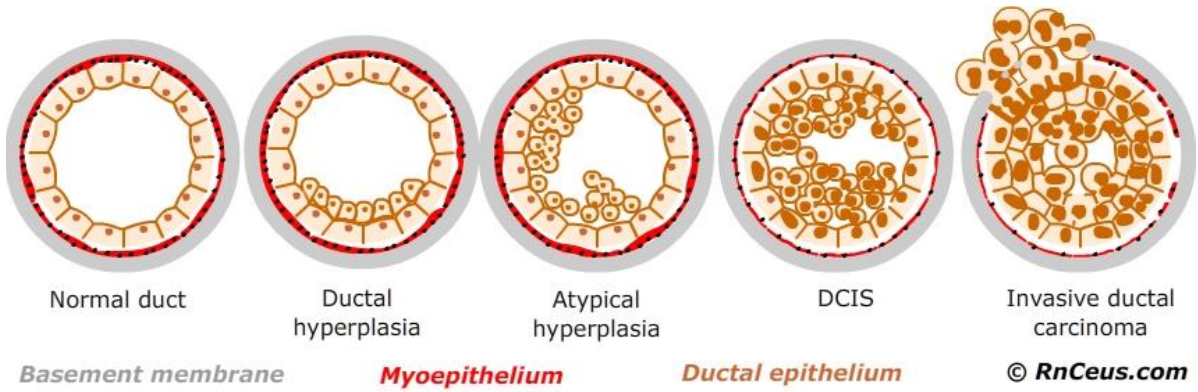




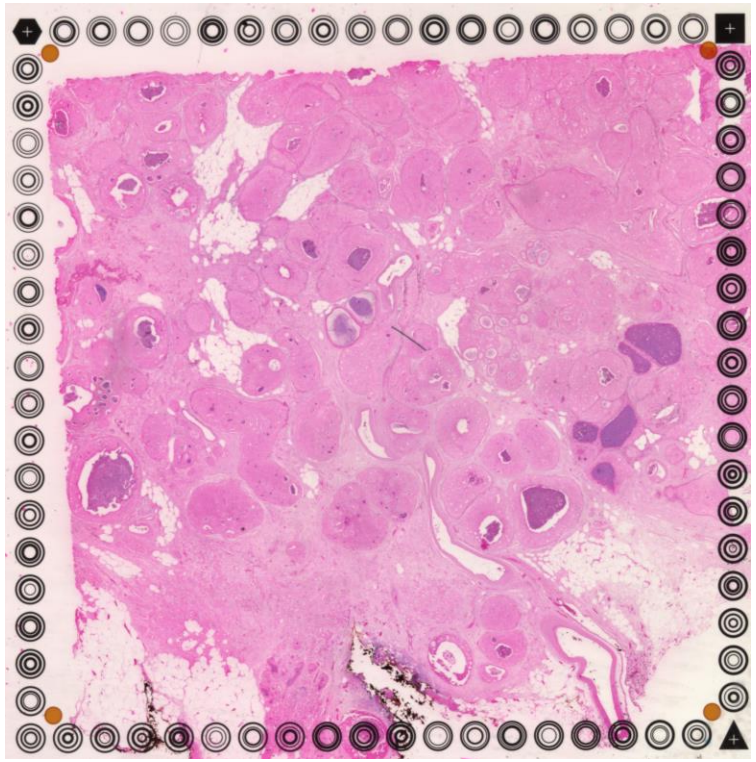




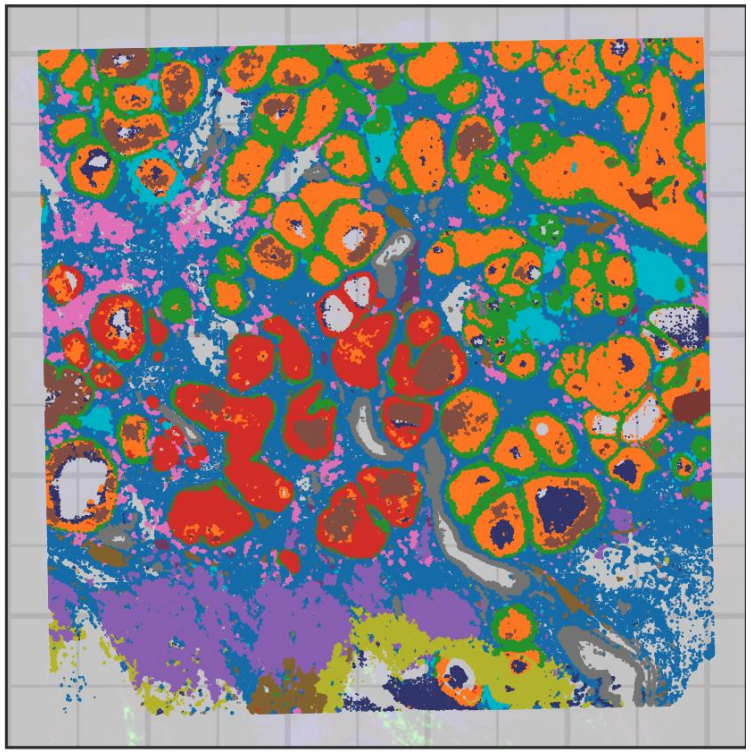




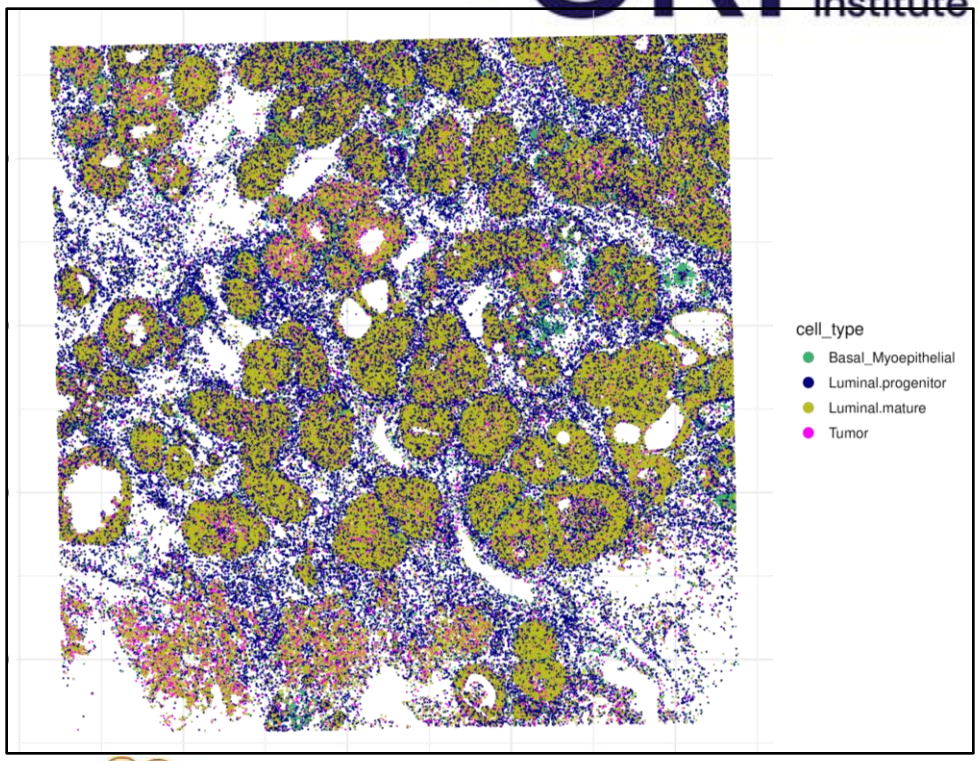
H&E CytAssist Image

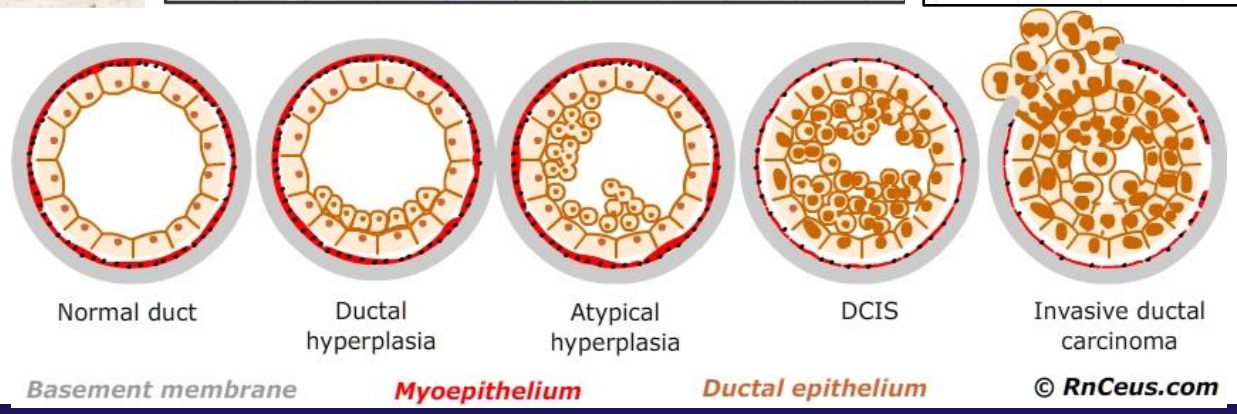


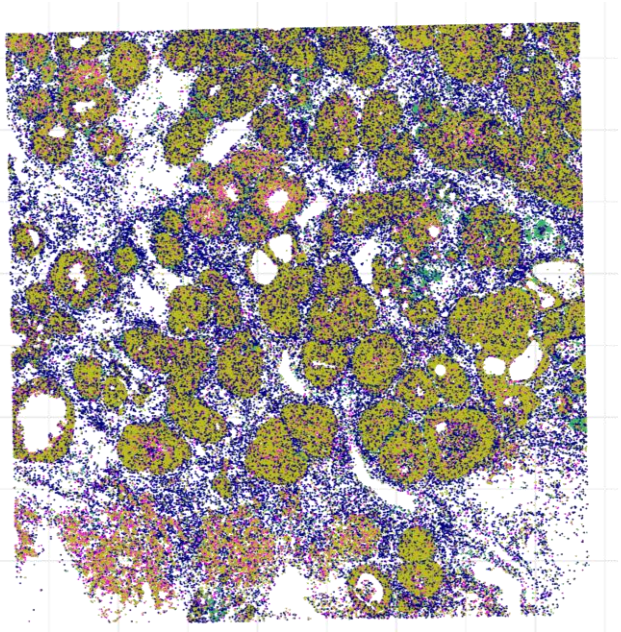
Full Object Banksy Clustering

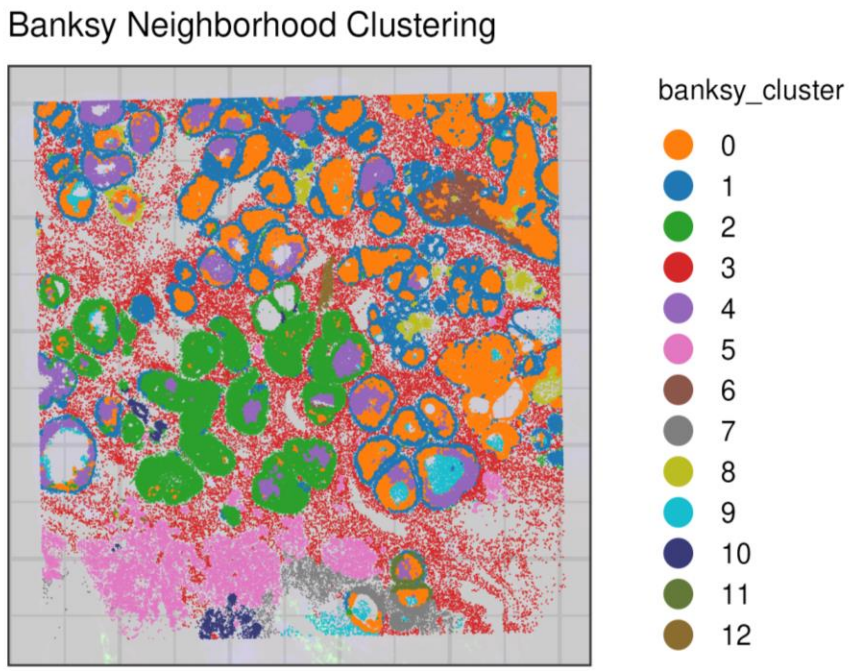


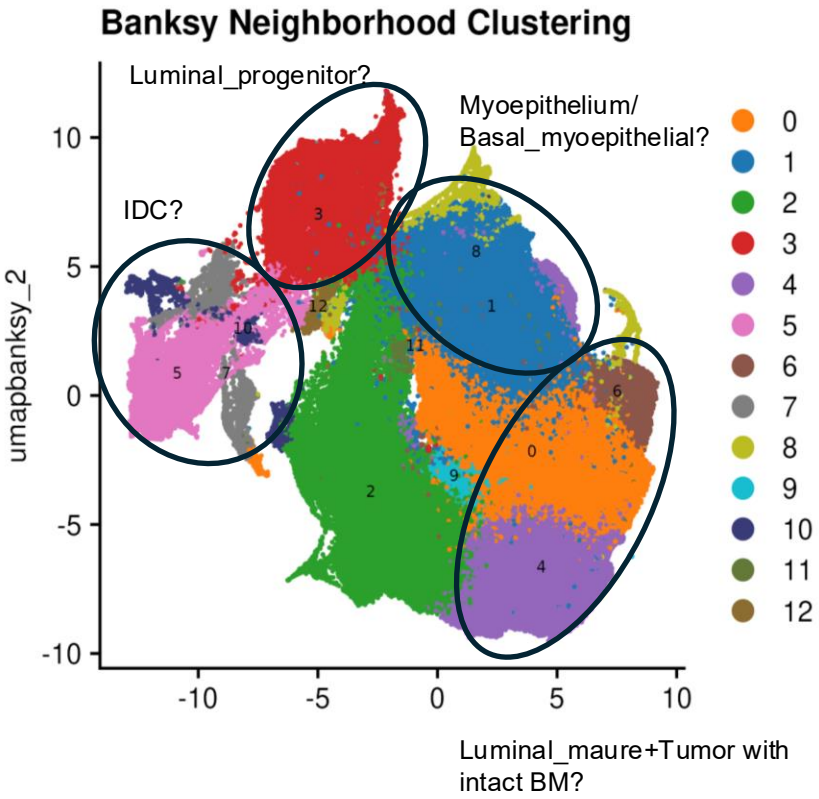
Epithelial Cells only



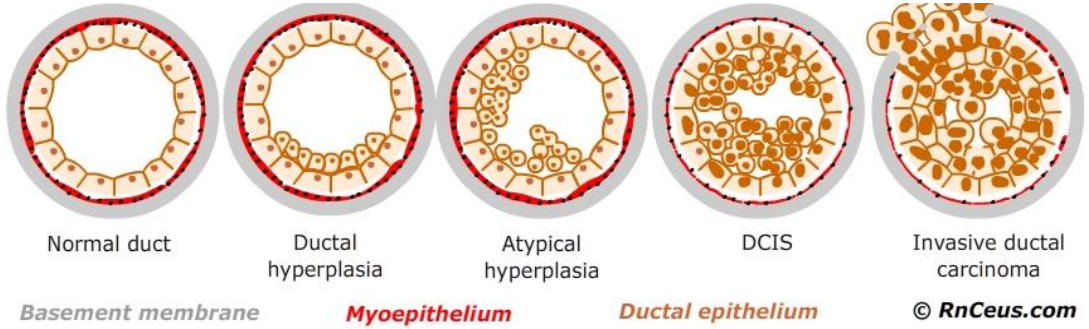


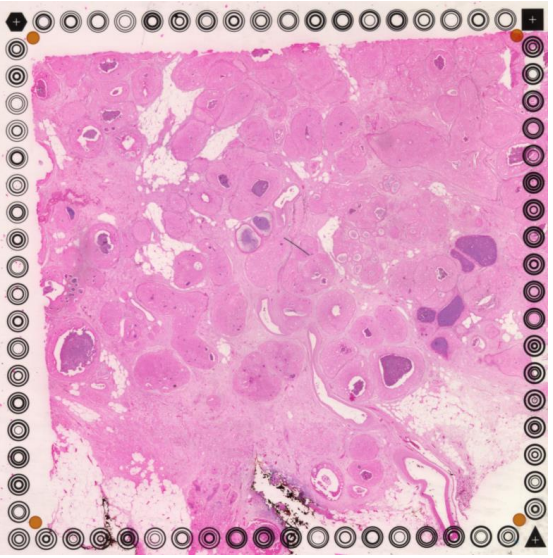




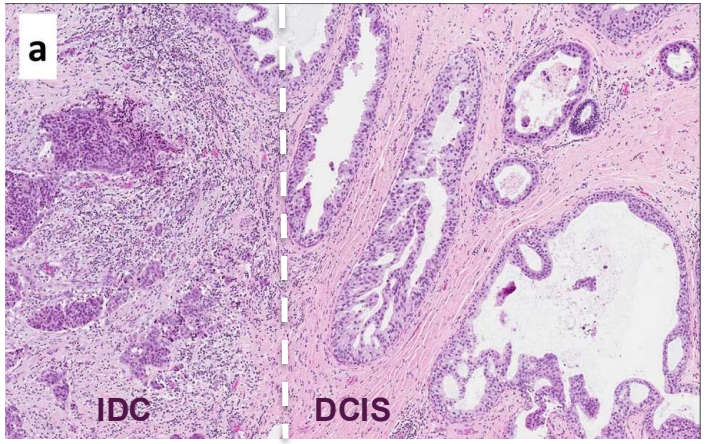


Luminal_maure+Tumor with intact BM?



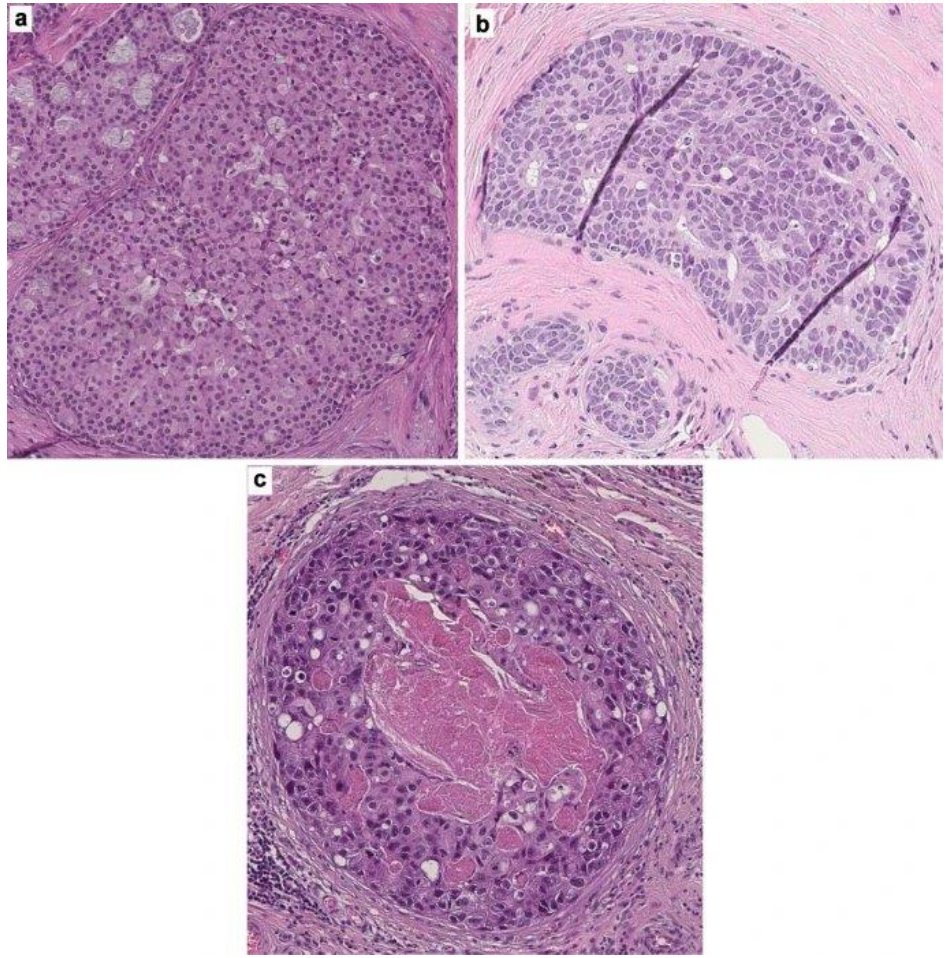


Hanna, W.M., et al. (2019) Mod Pathol 32, 896–915

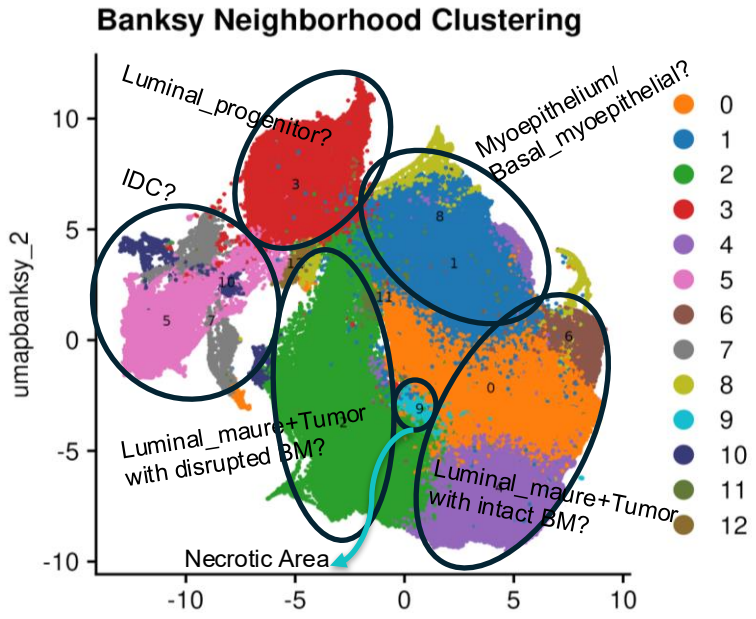


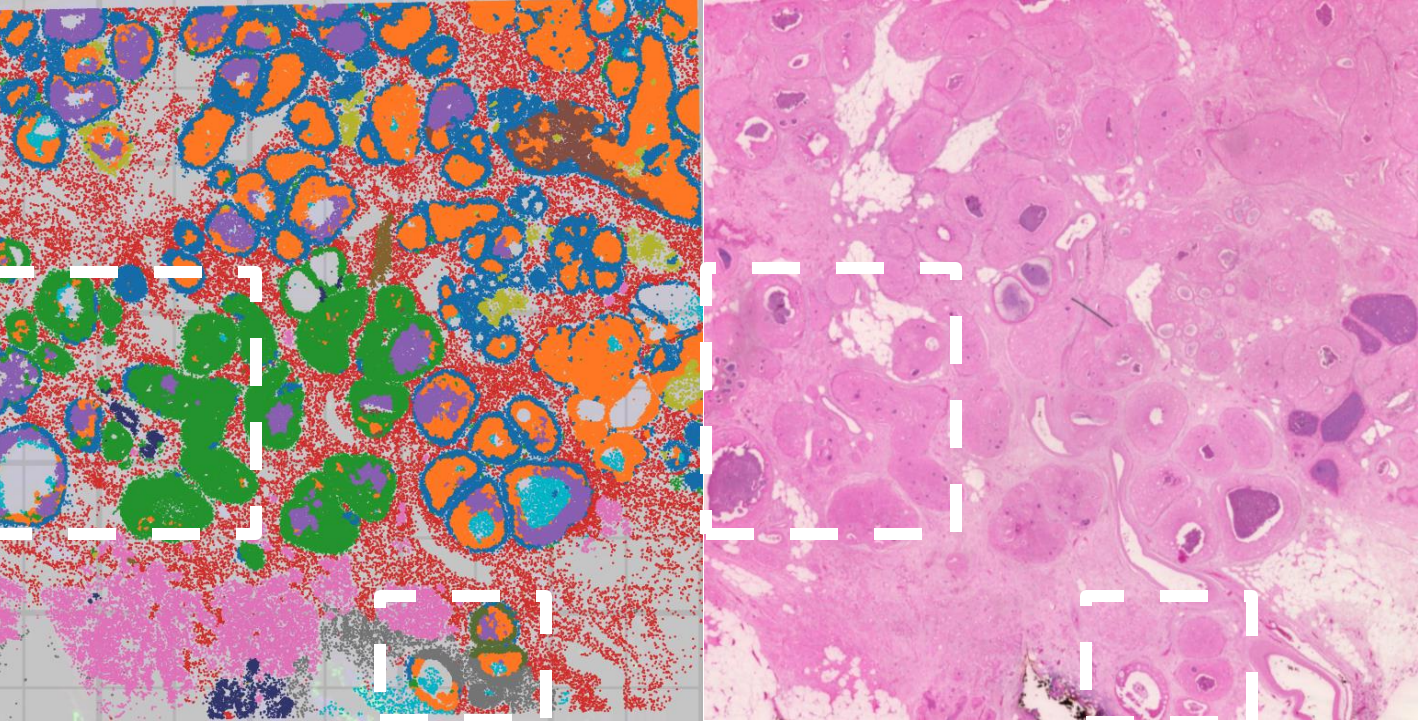
(a) DCIS associated with invasive ductal carcinoma

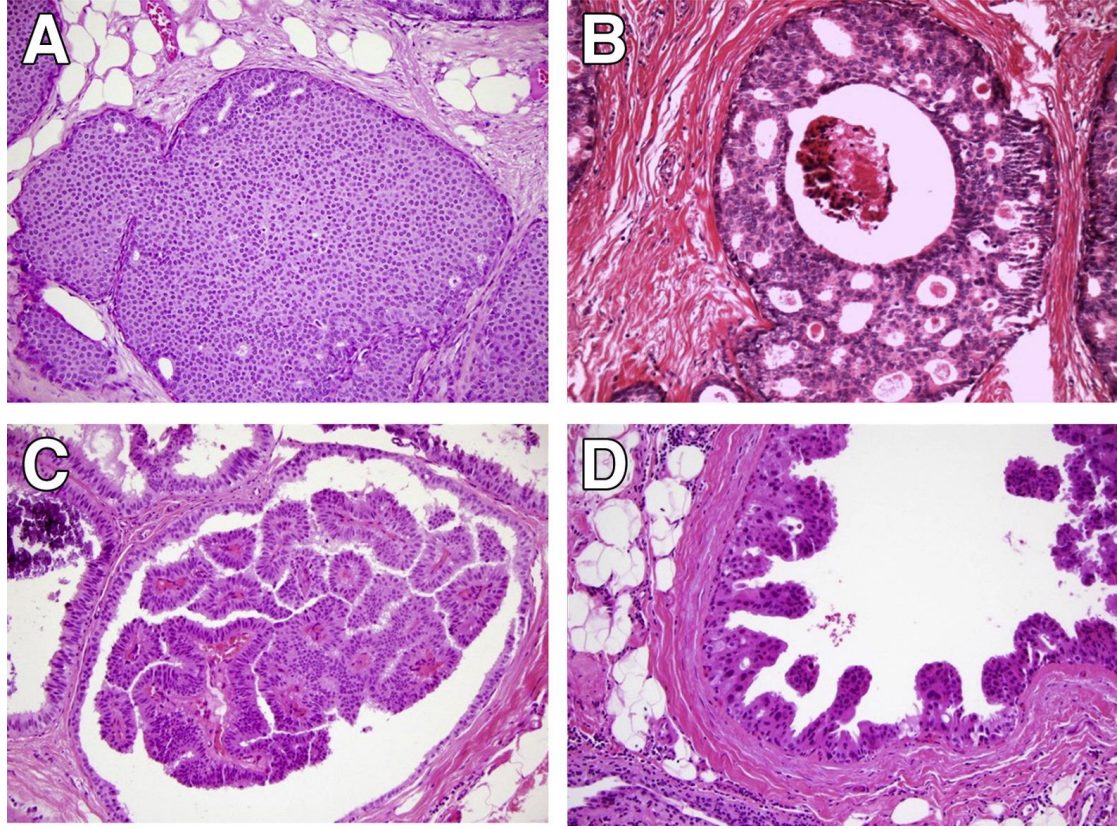
(Castellano, I., Metovic, J. (2018). Encyclopedia of Pathology. Springer, Cham)



(a) Low-grade DCIS is characterized by small, monomorphic, well-polarized cells, with uniform size. **(b) Intermediate-grade DCIS** consists of cells variable in size, shape, and placement and presents occasional mitotic figures and coarse chromatin. **(c) High-grade DCIS** is composed of highly atypical cells, large in size, with pleomorphic and poorly polarized irregular nuclei and presence of **necrosis** (H&E, 20x)

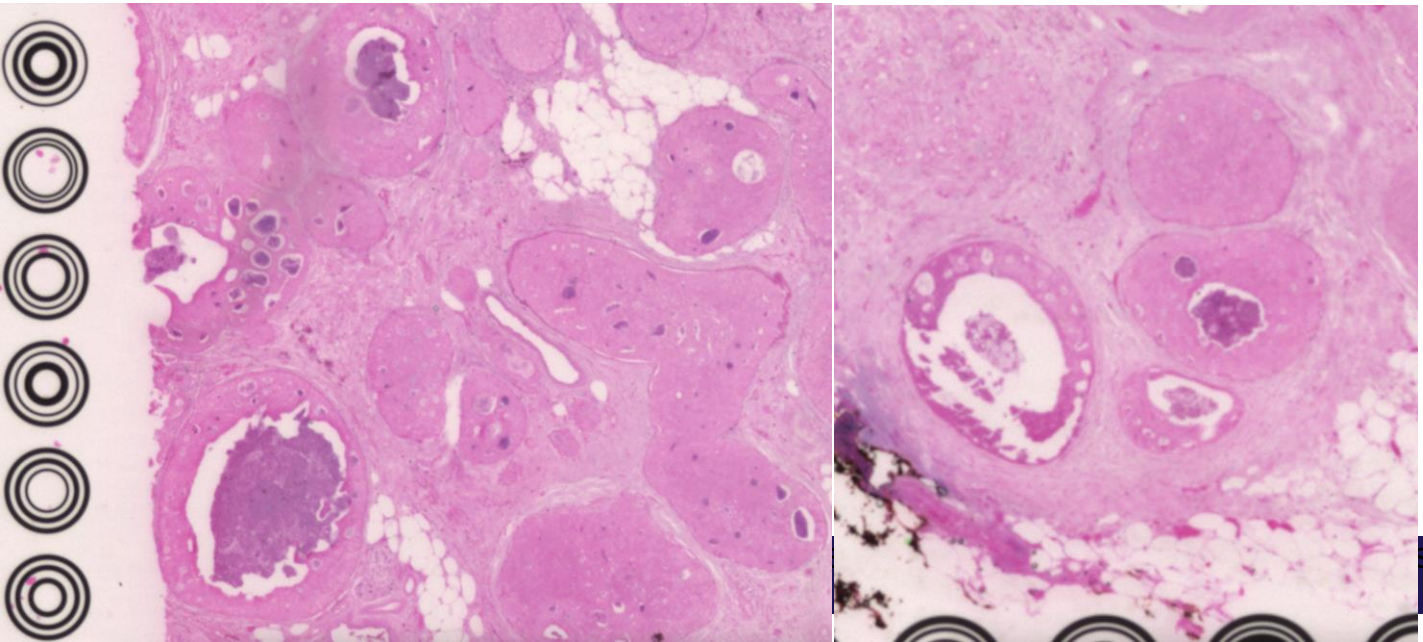




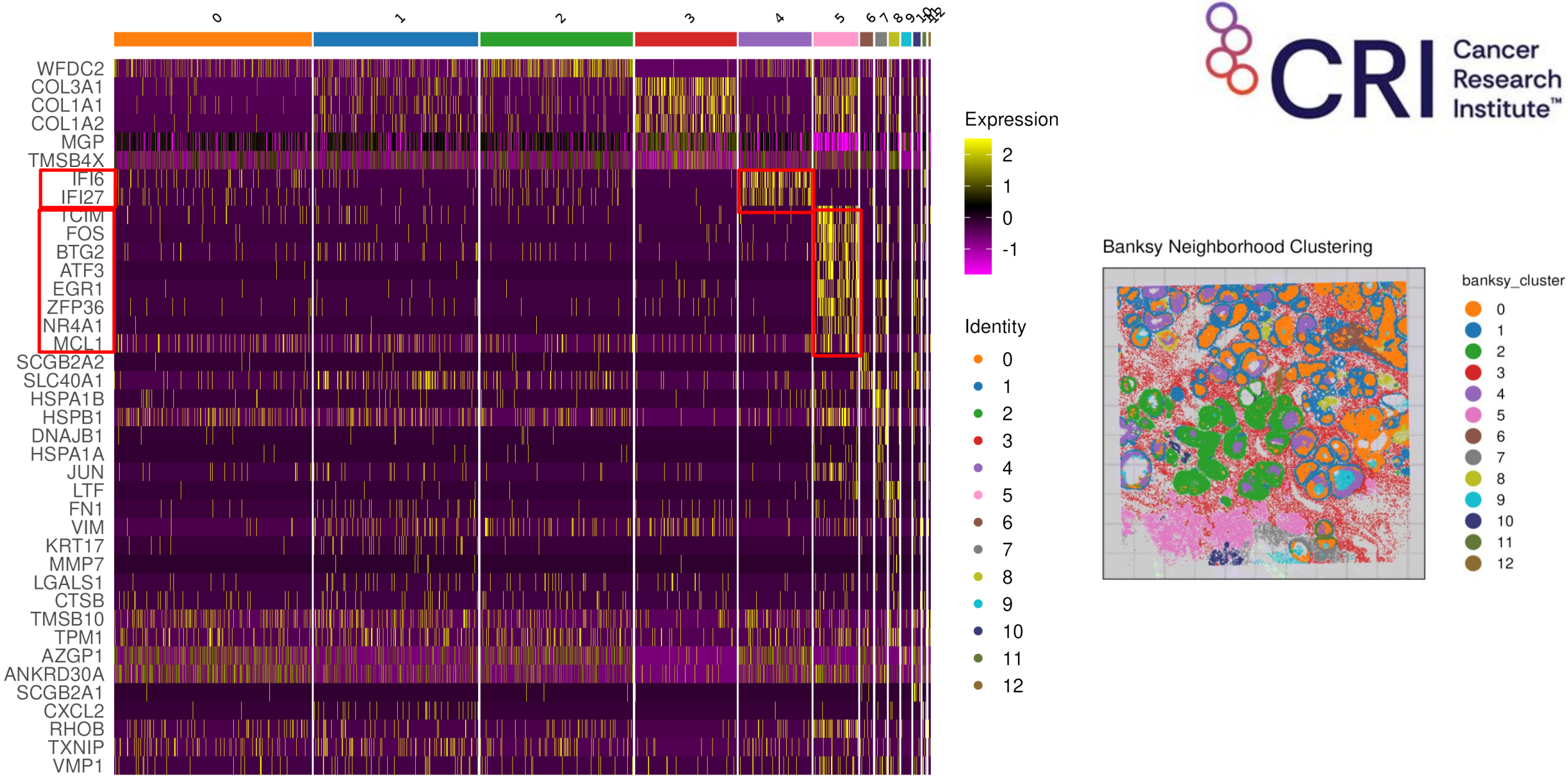


A: Ductal carcinoma *in situ* with solid growth pattern. **B:** Ductal carcinoma *in situ* with cribriform growth pattern and central comedo necrosis. **C:** Ductal carcinoma *in situ* with papillary growth pattern. **D:** Ductal carcinoma *in situ* with micropapillary growth pattern. Hematoxylin and eosin stain was used (**A–D**). Original magnification, ×20

(Sanati S. The American journal of pathology. 2019;189(5):946-55)

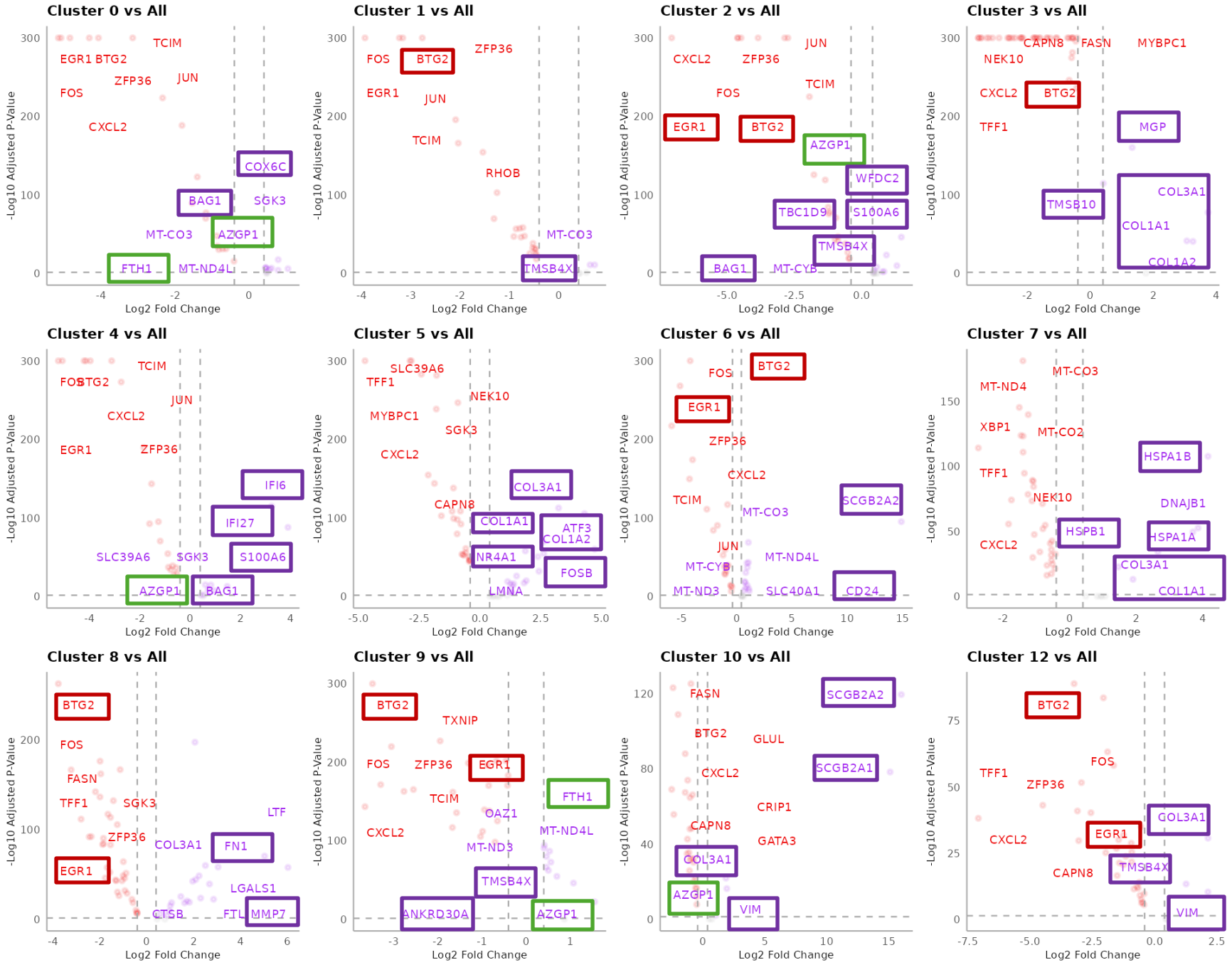


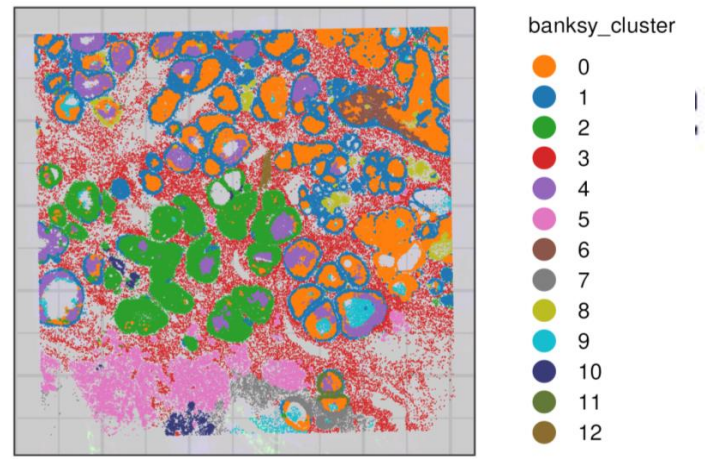
some IDC and very few "normal" ducts with a single layer of lumen

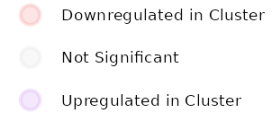


CRI Cancer Research Institute™

Banksy Neighborhood Clustering







In summary, **cluster 0**, **cluster 4**, and **cluster 9** have significant over-expression of DCIS markers as well as some tumor-suppressive genes (FTH1, AZGP1)

and have a less "aggressive" signature than **cluster 2** and **cluster 6**, which in addition over-express genes involved in DCIS invasion, tumor progression, and metastasis (SA100A6, WFDC2, CD24).

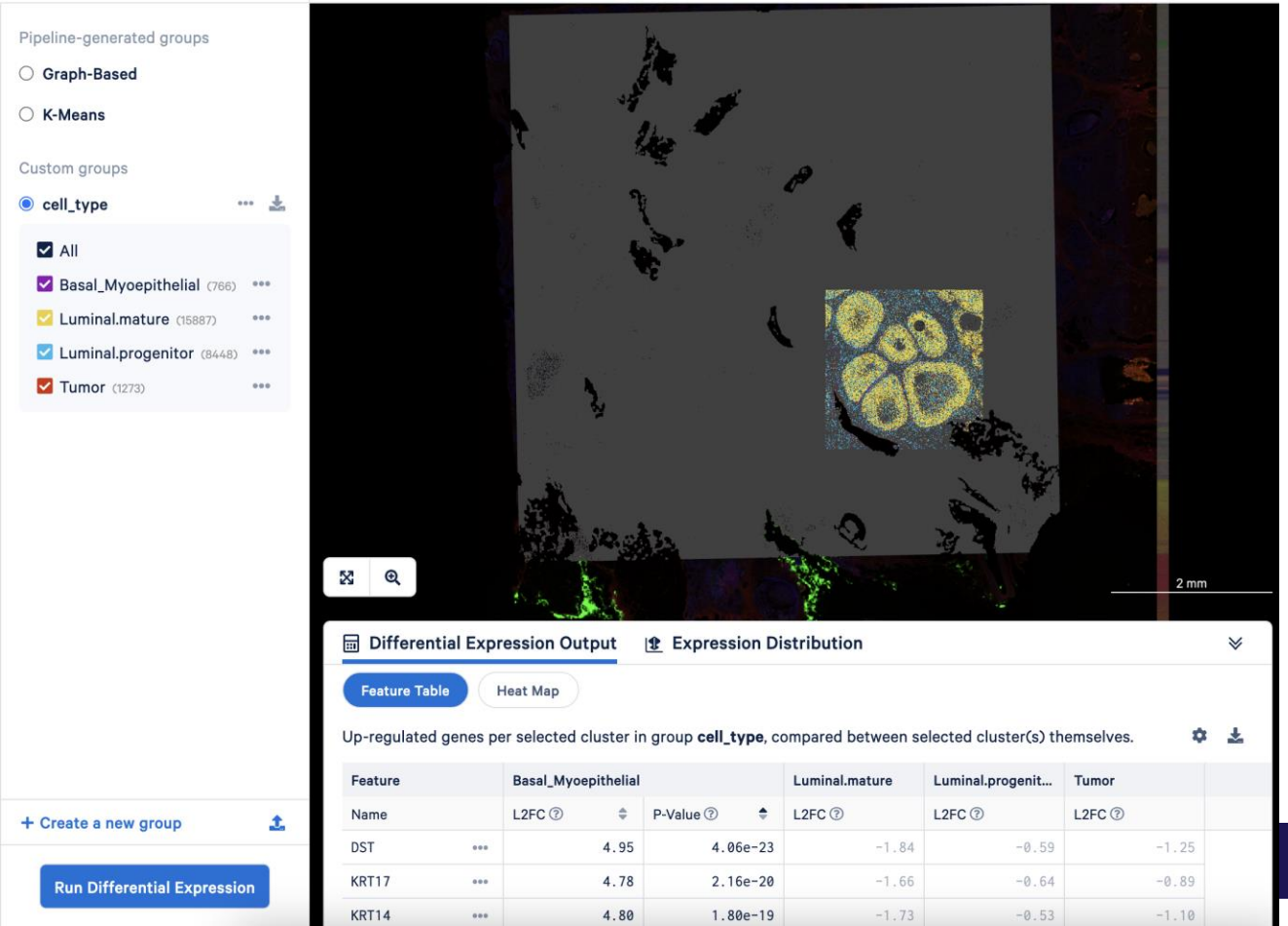
Cluster 3, **Cluster 5**, **cluster 7**, **cluster 10** and **cluster 12** all express markers of IDC, metastasis, and progression (VIM, COLs, TMSB4X, FOSB, SCGB2A1)

Cluster 8, and **cluster 1** over-express markers of myoepithelial cells (FN1, MMP7), and downregulates tumor suppressor genes (BTG2, EGR1)

Re-assign cell type annotation after AddModuleScore followed by DEG analysis

Importing cluster and cell type annotation to Loupe Browser

```
#############################################################################################
############ To extract csv files of the cell types and clusters to add to Loupe Browser ###############
#############################################################################################

metadata <- object@meta.data[, c("seurat_cluster.projected", "banksy_cluster", "cell_type")] %>%
  tibble::rownames_to_column(var = "CellID")

export_cluster_proj <- metadata[, c("CellID", "seurat_cluster.projected")]
export_cluster_banksy <- metadata[, c("CellID", "banksy_cluster")]
export_celltype <- metadata[, c("CellID", "cell_type")]

# Write to CSV
write.csv(export_cluster_proj, file = file.path(sample_dir, paste0(sample_names, "_cluster_proj.csv")), row.names = FALSE)
write.csv(export_cluster_banksy, file = file.path(sample_dir, paste0(sample_names, "_cluster_banksy.csv")), row.names = FALSE)
write.csv(export_celltype, file = file.path(sample_dir, paste0(sample_names, "_cell_type.csv")), row.names = FALSE)
```



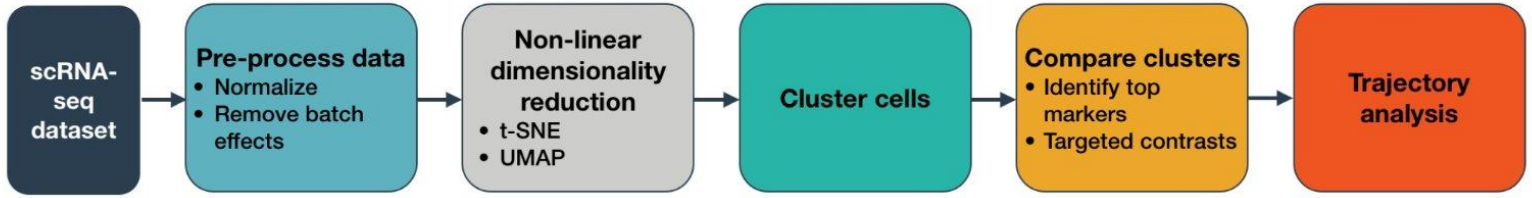
To identify a rare cell type with no prior gene sets in the literature, we can transfer the cluster label to Loupe Browser (or manually select an area annotated by a pathologist) and Run DEG analysis to define a sub-type

Monocle3 Single Cell Trajectory Analysis

Home / Docs / Getting started

Getting started with Monocle 3

Workflow steps at a glance



The cell_data_set class

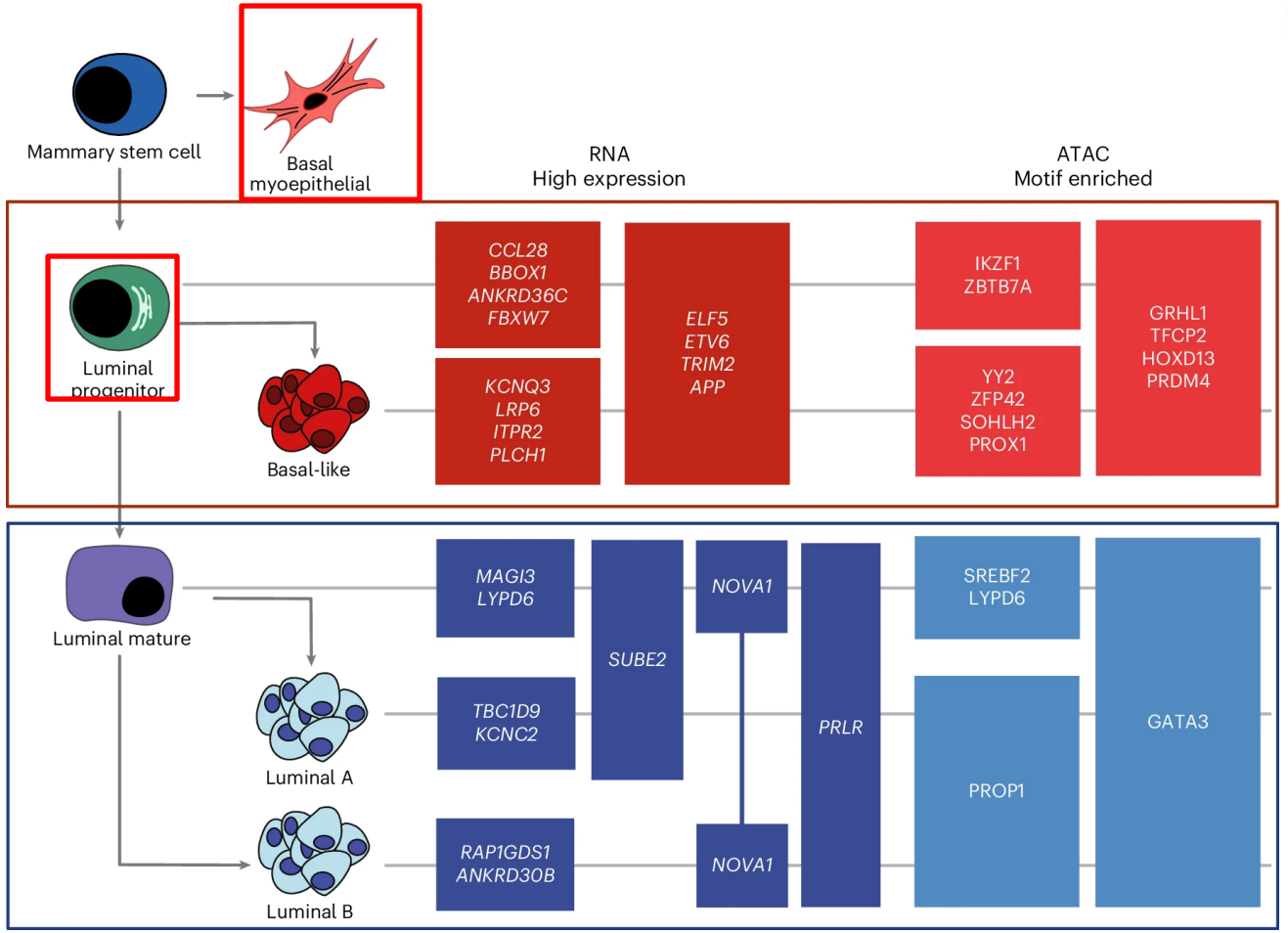
Monocle holds single-cell expression data in objects of the `cell_data_set` class. The class is derived from the Bioconductor `SingleCellExperiment` class, which provides a common interface familiar to those who have analyzed other single-cell experiments with Bioconductor. The class requires three input files:

- `expression_matrix`, a numeric matrix of expression values, where rows are genes, and columns are cells
- `cell_metadata`, a data frame, where rows are cells, and columns are cell attributes (such as cell type, culture condition, day captured, etc.)
- `gene_metadata`, an data frame, where rows are features (e.g. genes), and columns are gene attributes, such as biotype, gc content, etc.

Rather than purifying cells into discrete states experimentally, Monocle uses an algorithm to learn the sequence of gene expression changes each cell must go through as part of a dynamic biological process. Once it has learned the overall "trajectory" of gene expression changes, Monocle can place each cell at its proper position in the trajectory.

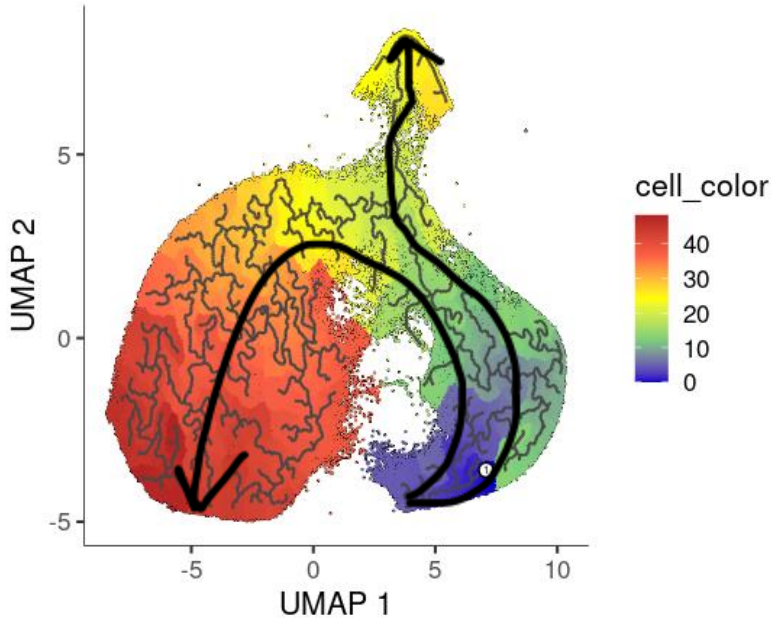
If there are multiple outcomes for the process, Monocle will reconstruct a "branched" trajectory. These branches correspond to cellular "decisions", and Monocle provides powerful tools for identifying the genes affected by them and involved in making them.

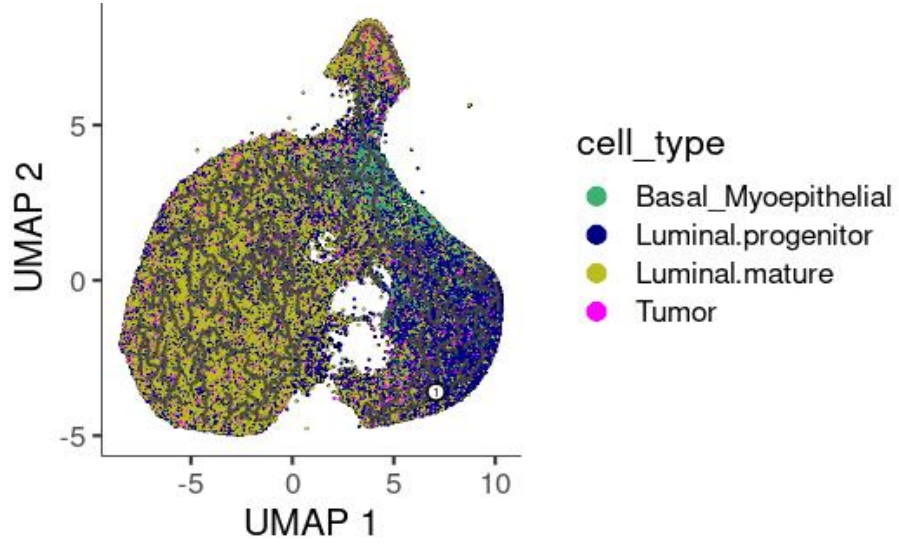
https://cole-trapnell-lab.github.io/monocle3/docs/getting_started/

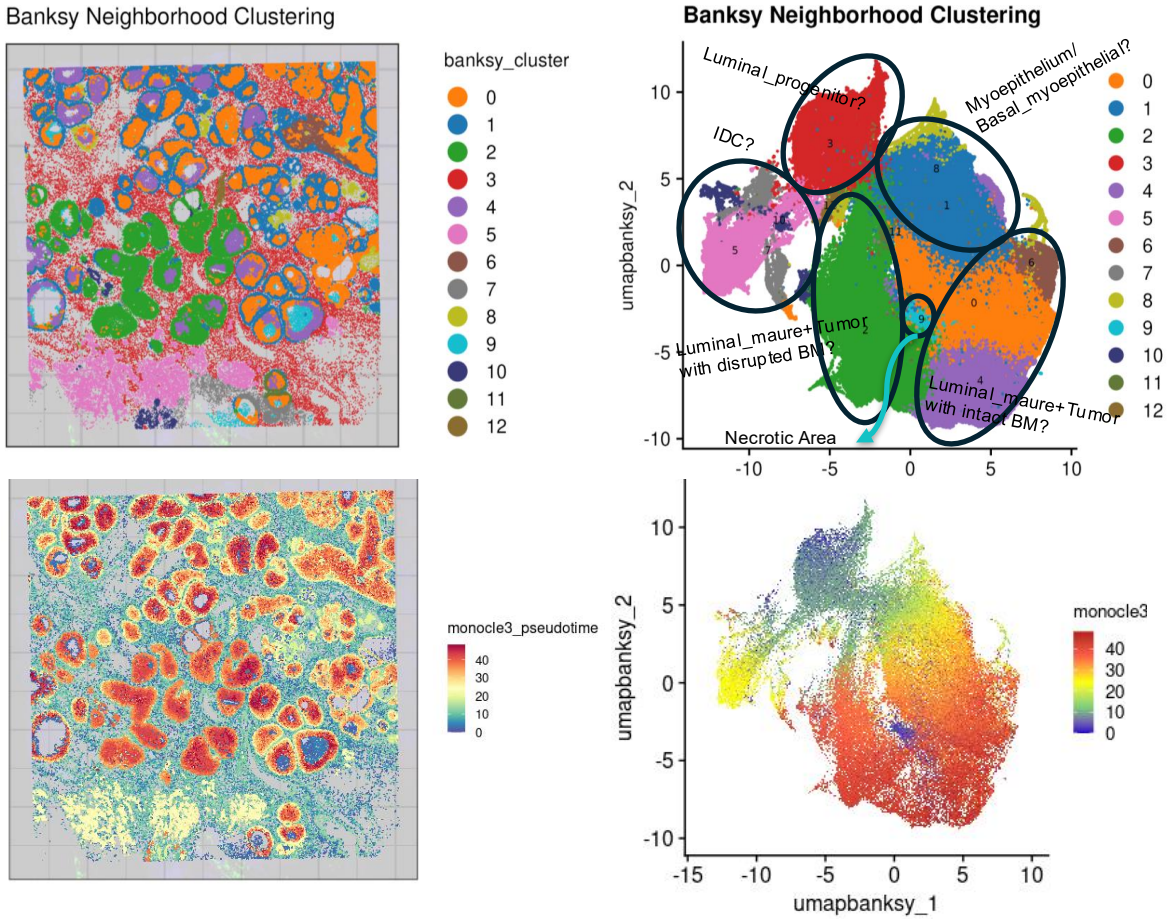


Iglesia MD et al. Nature Cancer. 2024 Nov;5(11):1713-36.

Using both **Luminal.progenitor** and **Basal_Myoepithelial** as the earliest_principal_node







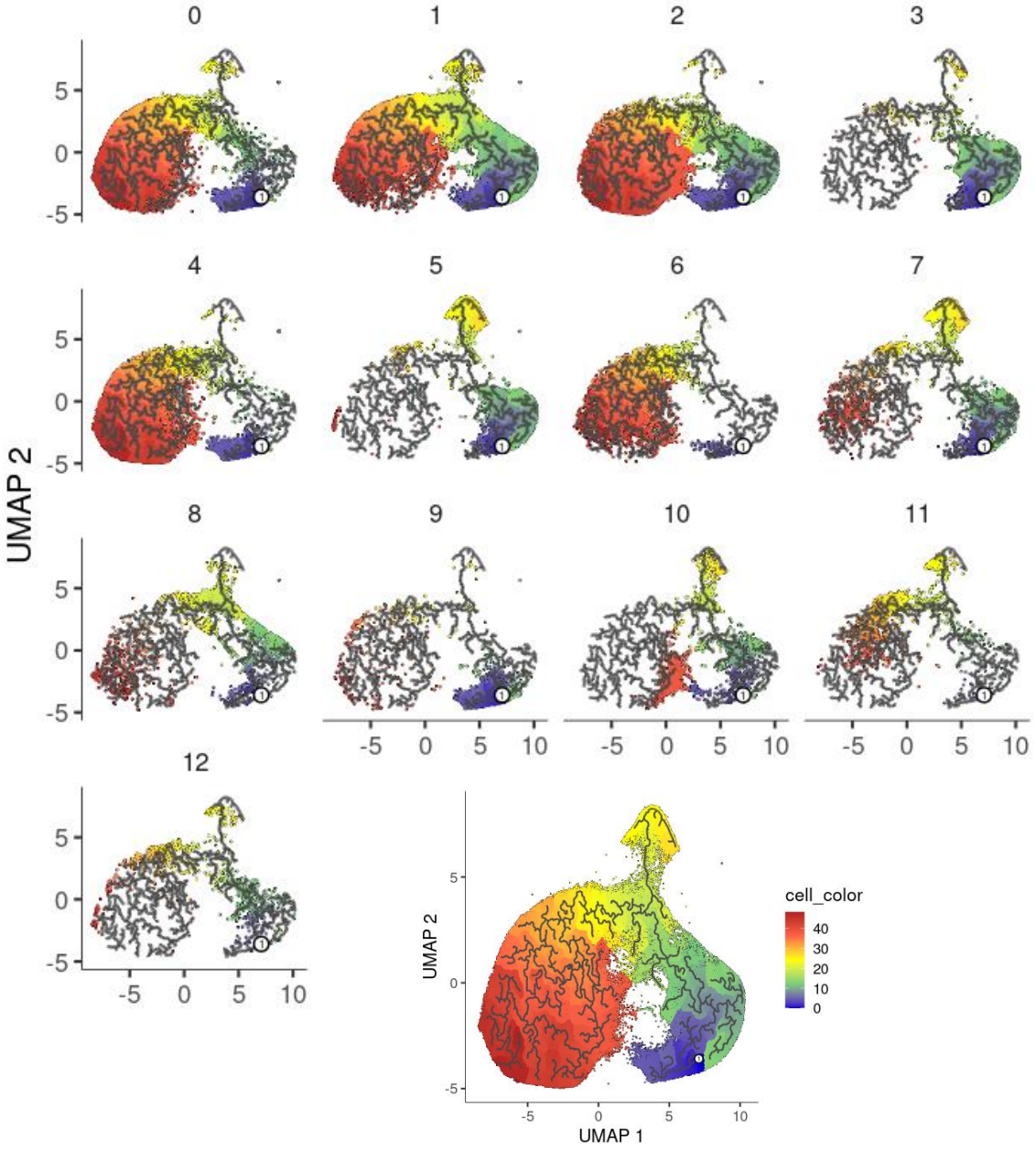
0, 1, 2, 4, 6 seem to be the trajectory from progenitor to well-differentiated DCIS

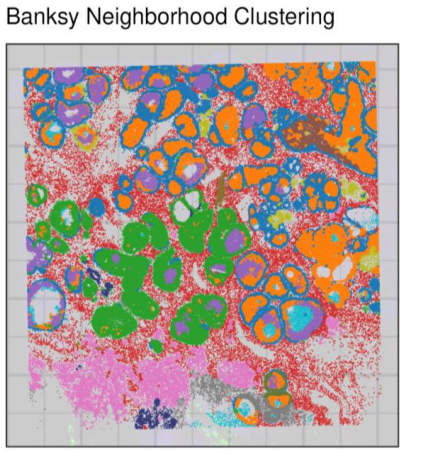
9 seems to be a poorly-differentiated DCIS

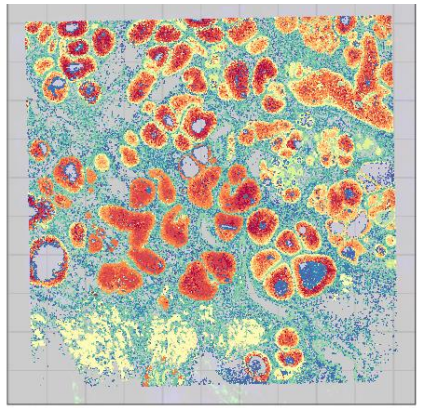
5, 7, 10, 11, 12 seem to be moderately-differentiated IDC?

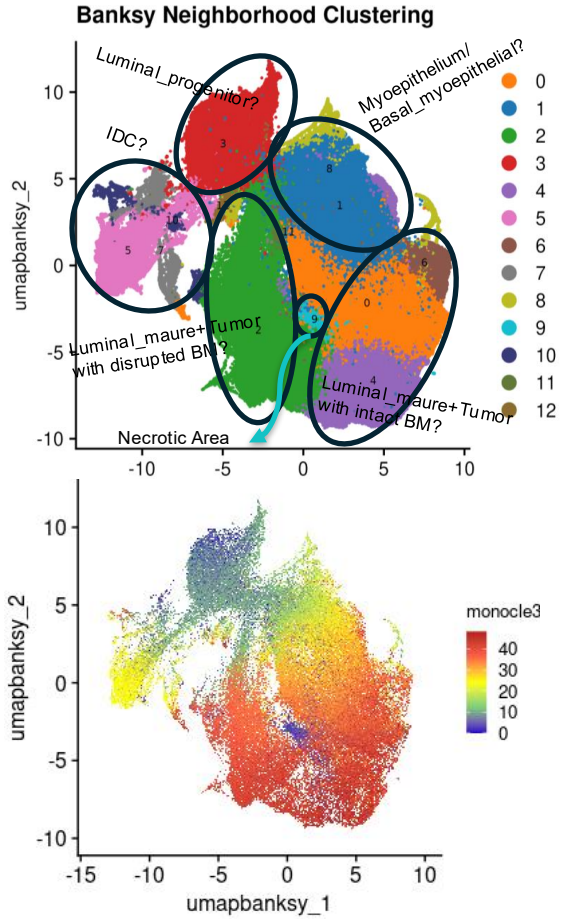
3 seems to be basal_myoepithelial, and part of **8** seems to be luminal.progenitor (confirm with DEG)

Using both Luminal.progenitor and Basal_Myoepithelial as the earliest_principal_node









0, 1, 2, 4, 6 seem to be the trajectory from progenitor to well-differentiated DCIS

9 seems to be a poorly-differentiated DCIS

5, 7, 10, 11, 12 seem to be moderately-differentiated IDC?

3 seems to be basal_myoepithelial, and part of **8** seems to be luminal.progenitor (confirm with DEG)

Hands-on: Unsupervised clustering, cell typing, loupe browser (2 hrs)

Summary/Questions

- Sample QC before, during, and after spatial transcriptomics run is very important to ensure good data quality
- Understanding the abundance of the tissue (block/USS/archived slides), the important questions you would like answer (discovery vs validation), plexity needed (whole transcriptome vs custom panel), resolution needed can help decide which tool to use
- Explore the Visium HD data in Loupe Browser, understand the pathology and biology of the tissue, sometimes filtering out the bins with the highest expression might remove important bins (for example tumor areas with WGD, examine each sample carefully before performing the same QC on all samples)
- For Visium HD, merging several samples at 8x8µm bin level can be very computationally intensive. Try to subset to tumor cells/epithelial cells before merging, immune, stroma, etc...
- For CNV analysis, try several aggregate sized (pooling a few adjacent 8x8µm bins, since the data can be very messy at 8x8)

Summary/Questions

Some resources for **Visium HD Spatial Transcriptomics** prior to the session:

- For sample pre-processing, the 10X Genomics Visium HD Analysis with spaceranger count explains how to run SpaceRanger with automatic and manual alignment. For manual alignment of the Visium HD CytAssist image, please make sure to download Loupe Browser 8.1.2 (Nov 18, 2024).
- To get started with how to perform analysis, visualization, and integration of Visium HD spatial datasets with Seurat, please check this vignette, which also includes information on how to run Banksy unsupervised neighborhood clustering
- For information on how to construct single-cell trajectories/pseudotime trajectories using Monocle3, please check this vignette
- For using InferCNV with the Visium HD data, this vignette illustrates inferring copy number alterations from tumor single cell RNA-Seq data
- For nuclei Segmentation and Custom Binning of Visium HD Gene Expression Data, please refer to this vignette by 10X. This is the only step that requires working in Python, all the rest of the scripts work with Rstudio. To convert the Seurat object (R) to work with Python (AnnData) please follow this vignette.

Closing and Feedback (15 min)