STAT 35510
Lecture 12

Spring, 2024

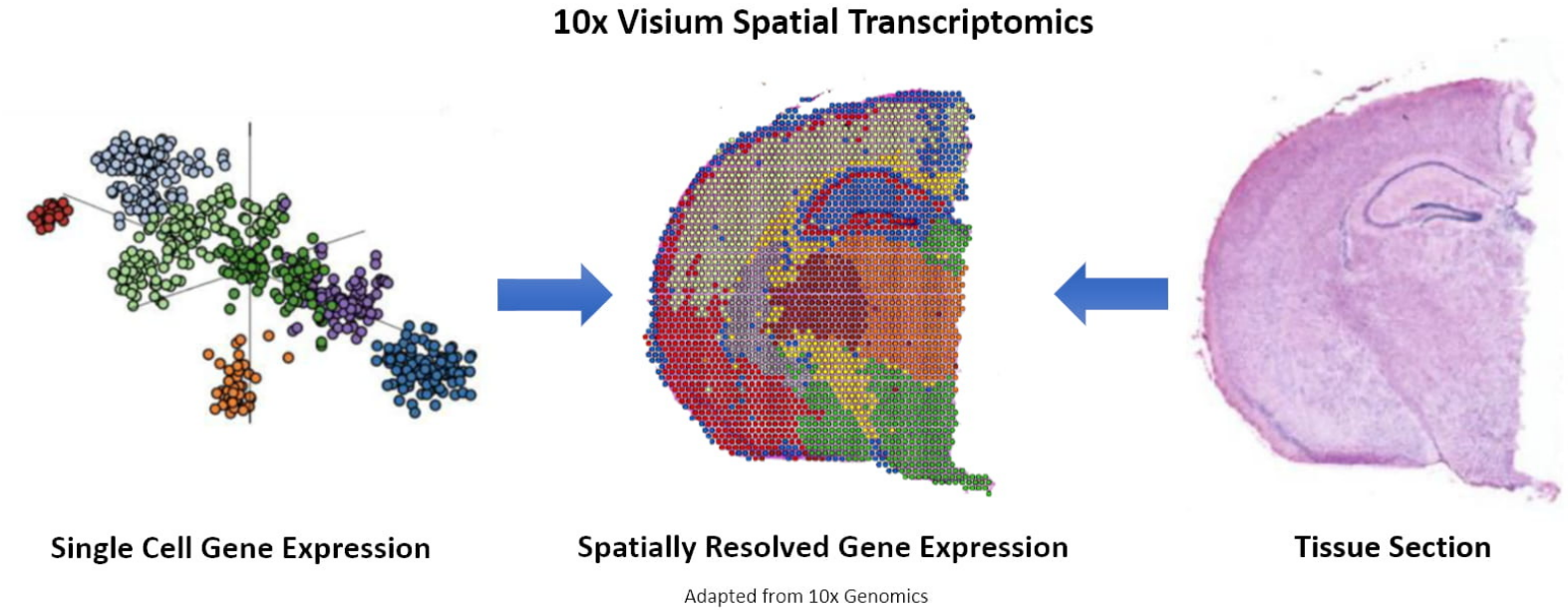
Jingshu Wang

Outline

- Spatial transcriptomics
  - Histology
  - Image-based and sequencing-based technologies
- Spatial domain detection
  - Spatial statistics-based methods and GNN methods
  - Integration of multiple slices

What is spatial transcriptomics?

- Spatial transcriptomics measure both transcriptomics (gene expression levels across the whole genome) and spatial information
  - Many genes need to be properly regulated in space for the system to function
  - Understand spatial patterns of gene expressions



Histology

- Histology: spatial transcriptomics data often have an associated histology image
  - Microscopic anatomy of biological tissues
  - Staining provides colors:
    - H&E stain: stains the nuclei purplish-blue and cytoplasm and other tissues in various stains of pink
  - Can be used to diagnose cancer and other diseases

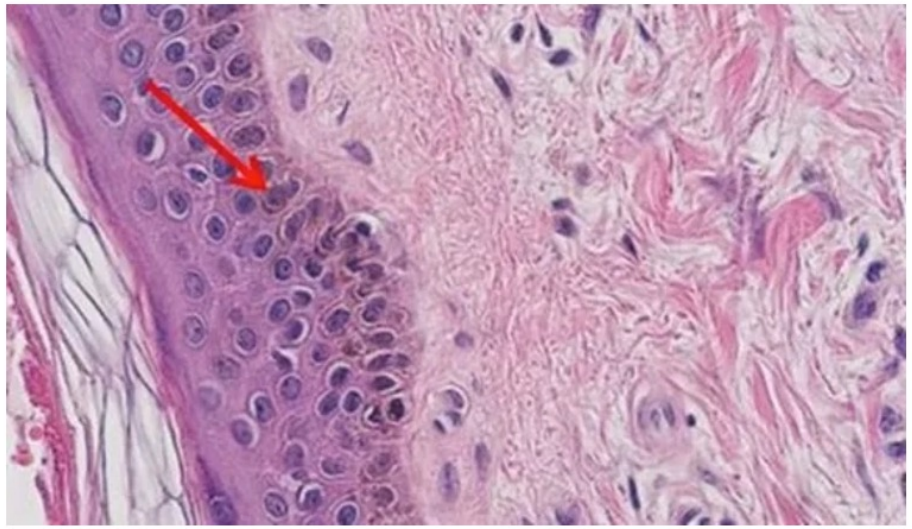
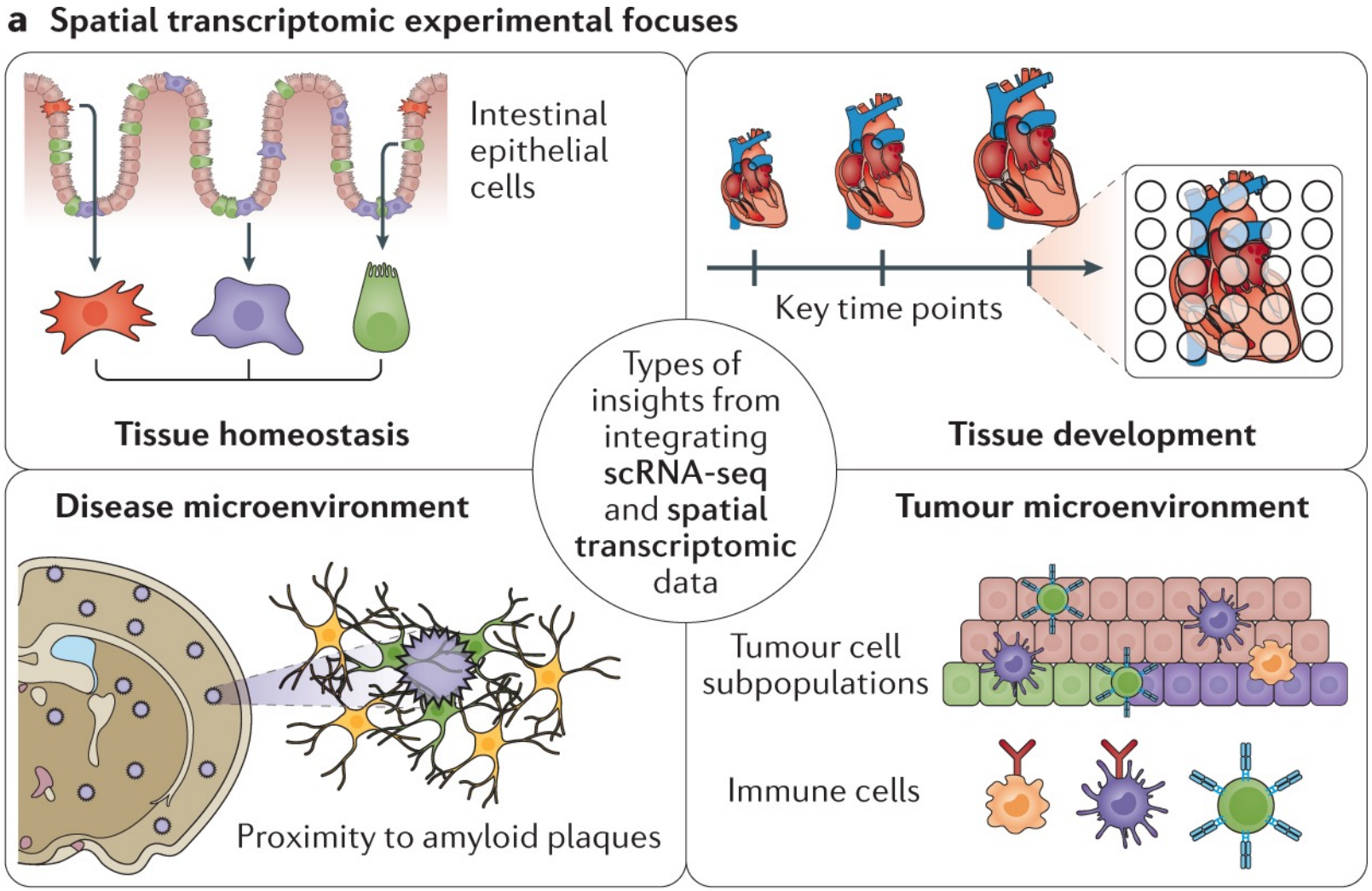


Fig 1: Skin H&E. Note the balanced coloration in this section of skin. The nuclei are stained purple, while the cytoplasmic components are pink.

https://www.leicabiosystems.com/us/knowledge-pathway/he-staining-overview-a-guide-to-best-practices/

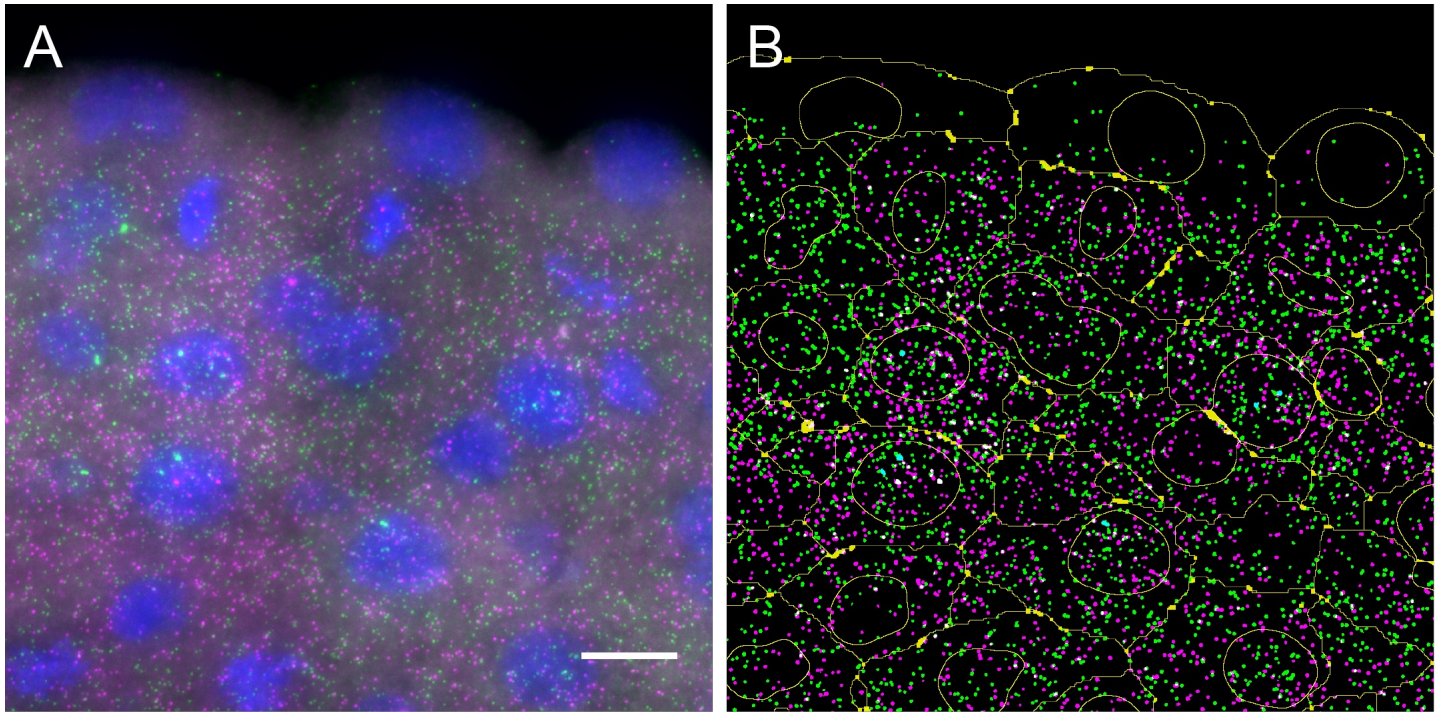
Why spatial transcriptomics?



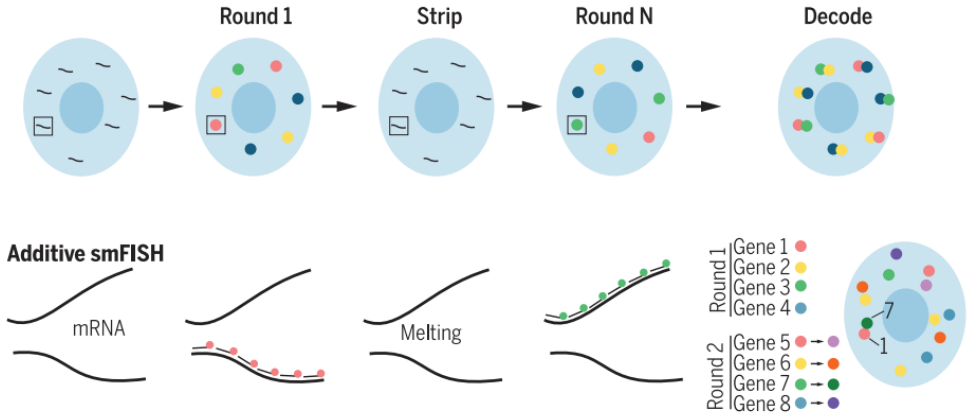
(Longo et. al., Nature Reviews Genetics 2021)

RNA-Fluorescence in situ hybridization (FISH)

- FISH is a technique using fluorescently labeled probe to detect specific DNA/RNA sequence
  - Keep the location of the cells but can only detected a limited number of genes



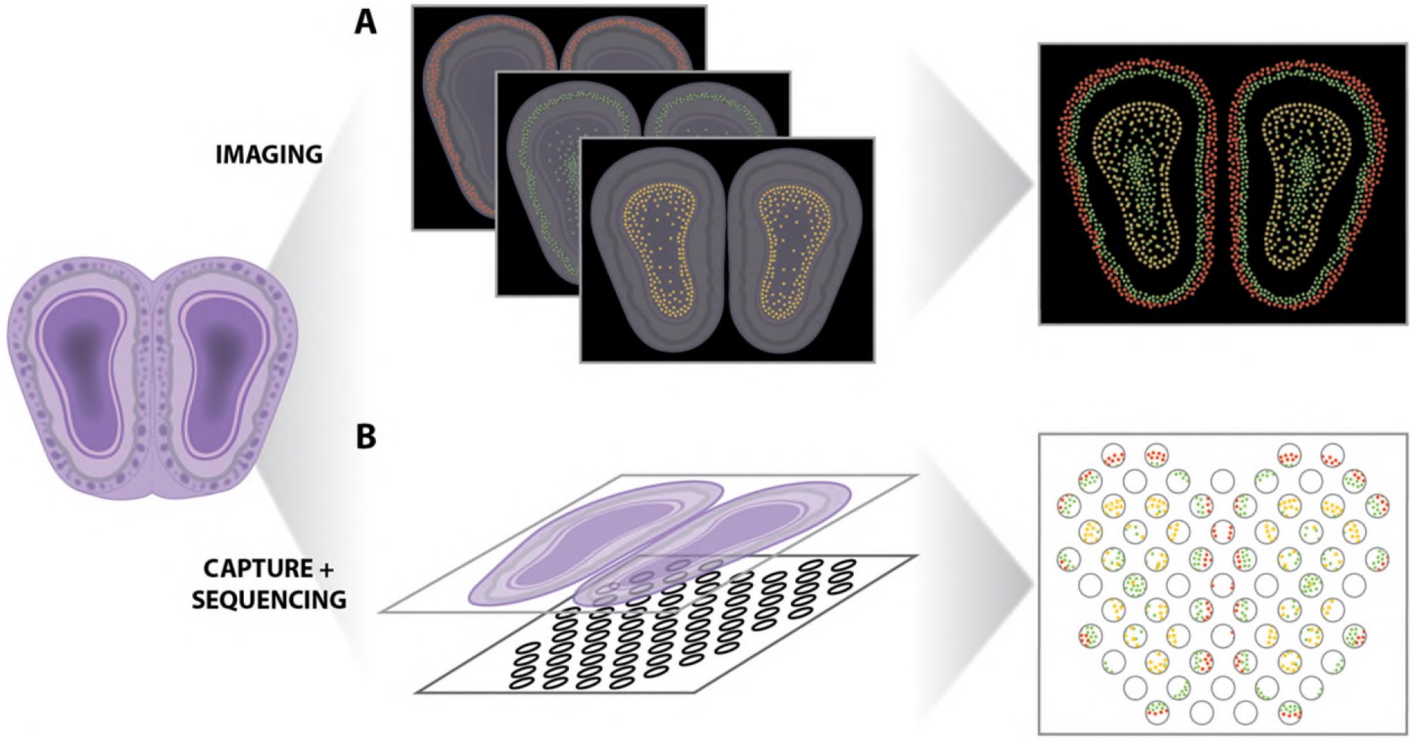
Transcripts detected by smFISH. (**A**) Original smFISH image. Blue: nuclei, magenta: smFISH for *apoeb*, green: smFISH for *aldob*. (**B**) Results of the smFISH analysis pipeline when applied to the image shown in (A). Yellow: outlines of cells and nuclei, magenta: detected *apoeb* transcripts, white: detected transcription foci for *apoeb*, green: detected *aldob* transcripts, cyan: detected transcription foci for *aldob*. Scale bar: 10 μm.



https://thenode.biologists.com/fishing-fish-2/resources/

Two types of spatial transcriptomics technologies

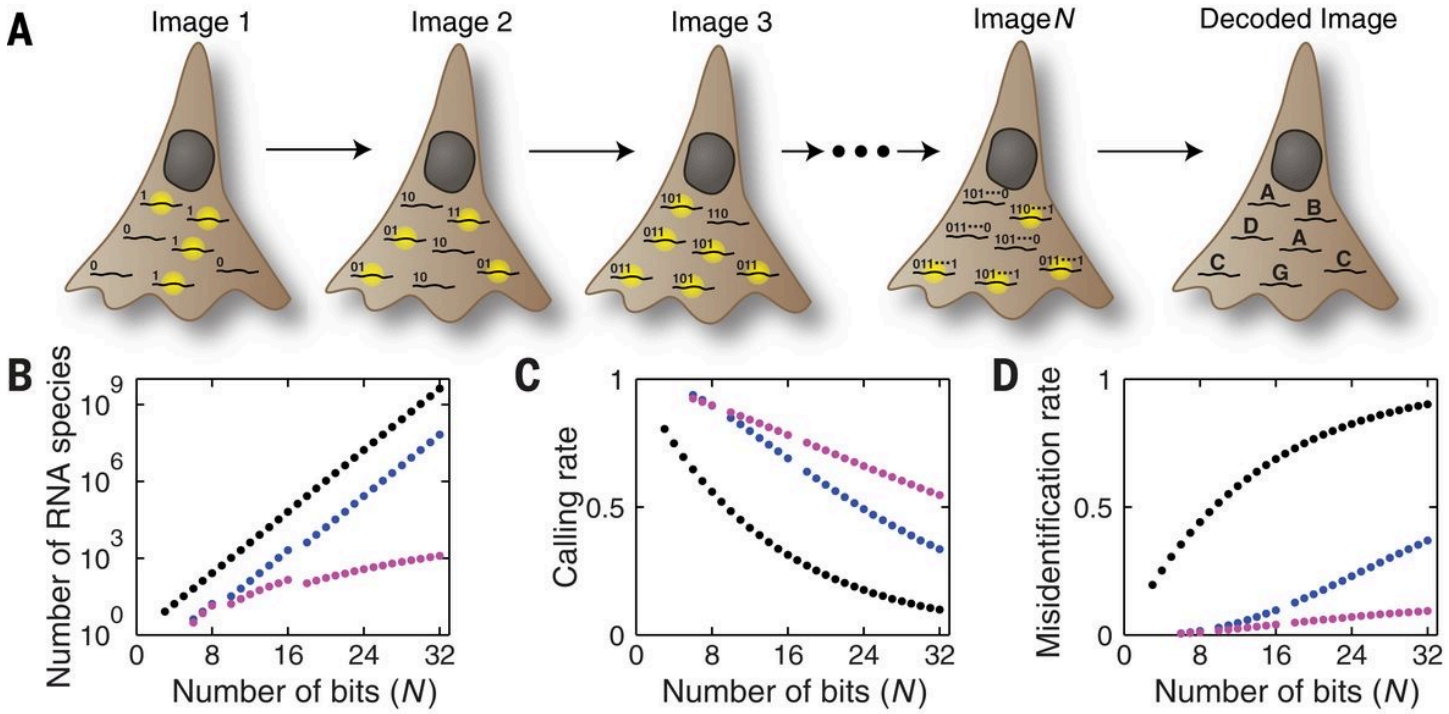
- Sequencing based spatial transcriptomics
  - Use scRNA-seq techniques to measure transcriptomics profiles for each spatial spot
- Image-based spatial transcriptomics
  - Use FISH techniques, increase the number of genes detected to a few hundreds



(Atta and Fan, Nature Comm, 2021)

MERFISH (Chen et. al., Science 2015)

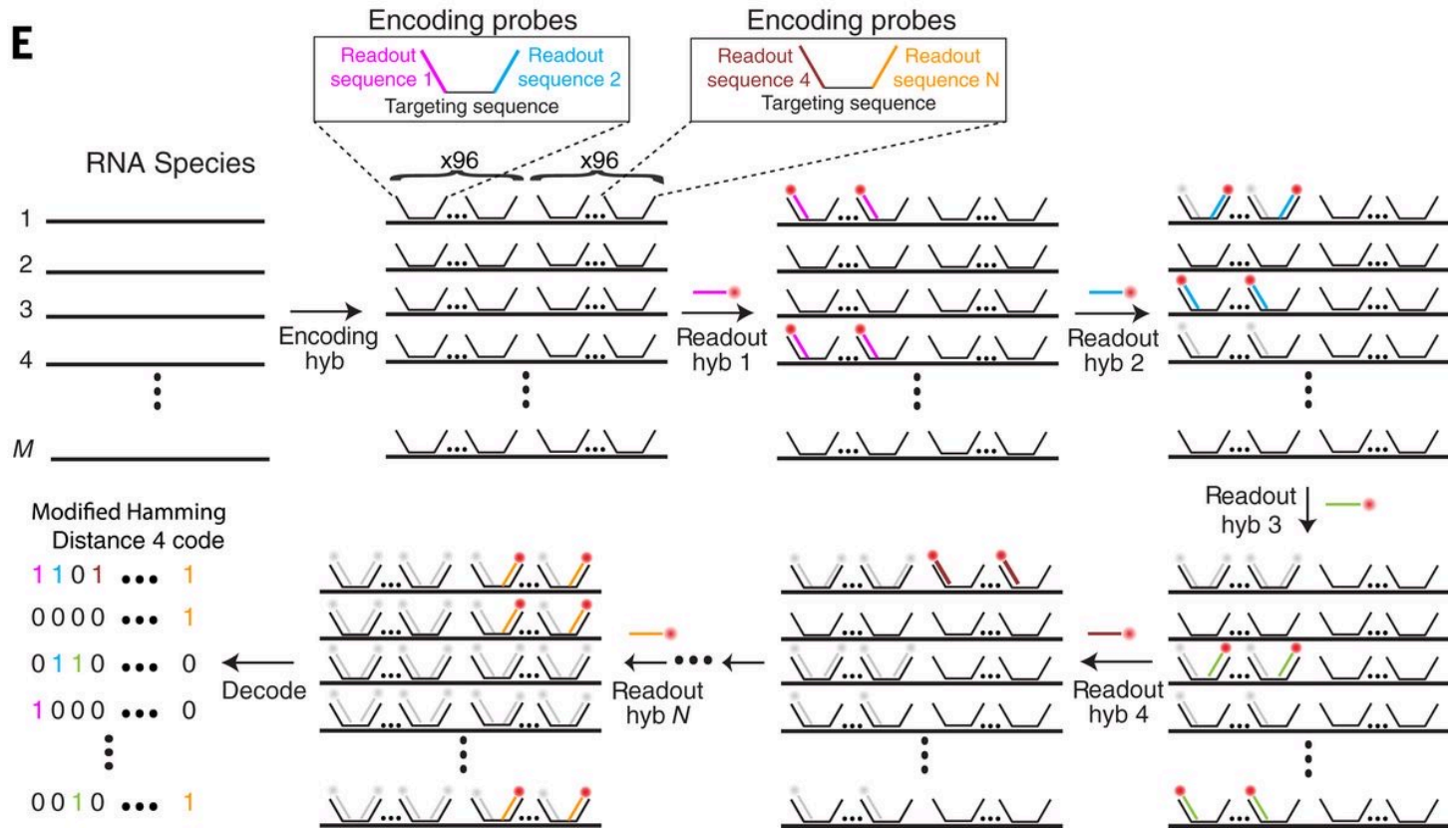
- Multiplex error-robust FISH that can measure 100-1000 genes
- smFISH: K round→ measure K gene
- Combinatorial barcoding of the genes: K round → measure $2^K - 1$ genes at most
  - Problem: calling rate also has an exponential decay (black dots)
    - Assume 1 -> 0 error p_1, 0 -> 1 error p_2, the code has m 1s, recall rate will be $(1-p_1)^m(1-p_2)^{K-m}$



MERFISH (Chen et. al., Science 2015)

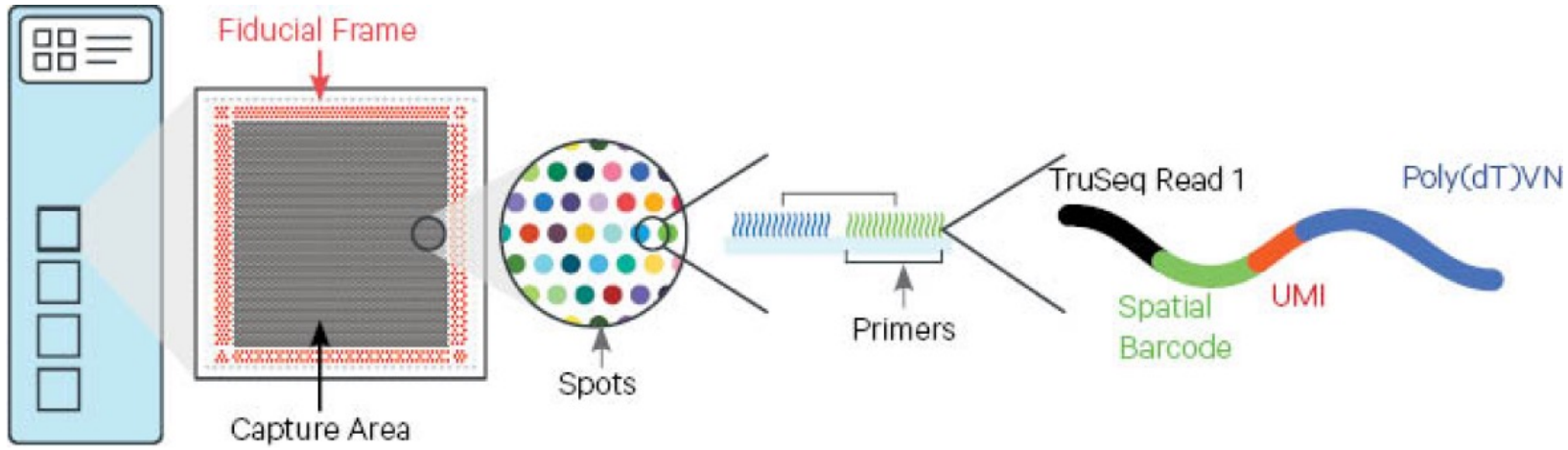
Solution: error-robust coding

- Encode each gene so that the barcode Hamming distance is at least 4
- Each gene barcode has exactly 4 1s to increase recall rate (as $p_1 > p_2$)

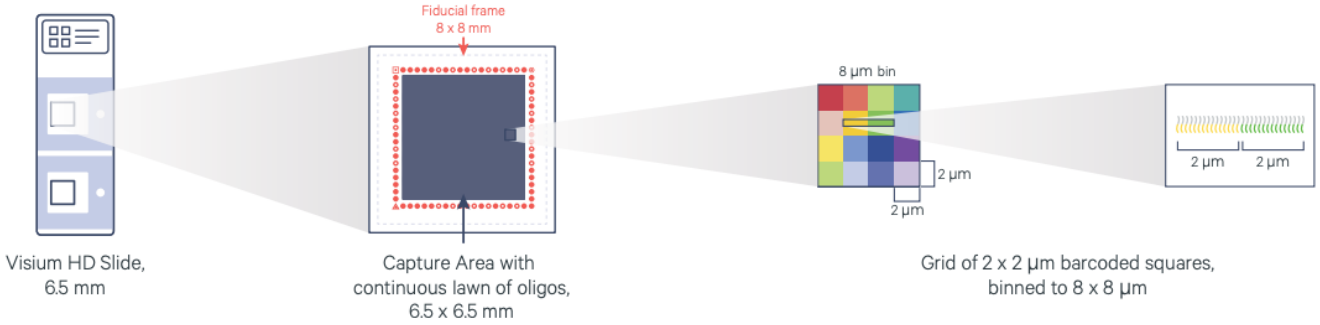


10X Visium

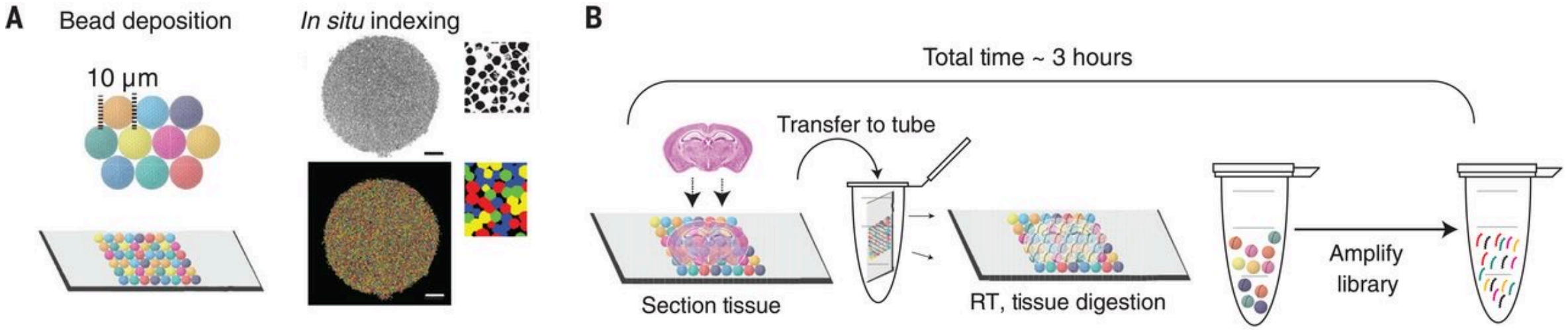
- Resolution: 55 µm spot (Stahl et. al., Science 2016)
- Typical human cell dimension: 10-15 µm in diameters, depend on the cell type



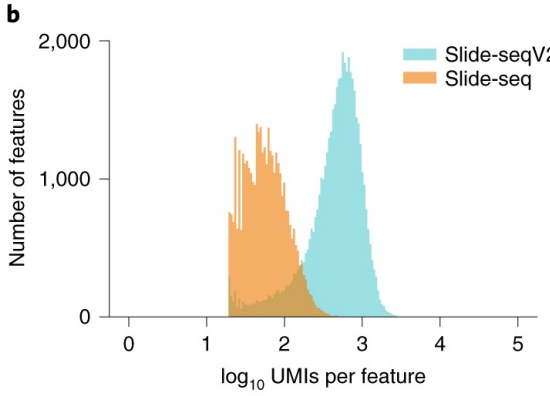
- Visium HD: 3 µm resolution, binned to 8 * 8 µm bins as a starting point
  - Much more expensive

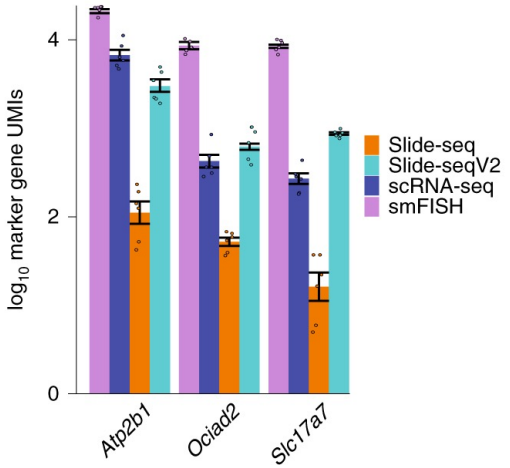


Slide-seq (Rodrigques et. al. Science 2019) & Slide-seqV2 (Stickels et. al., Nature Biotech 2021)

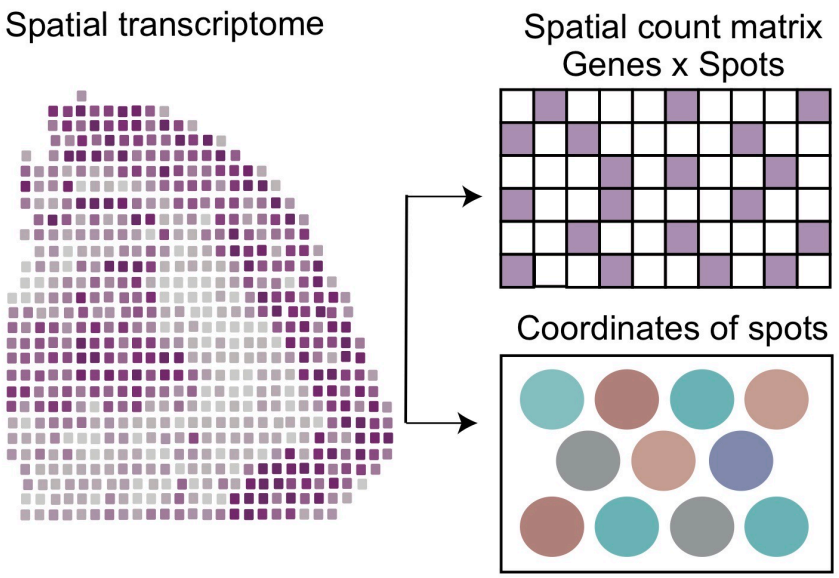


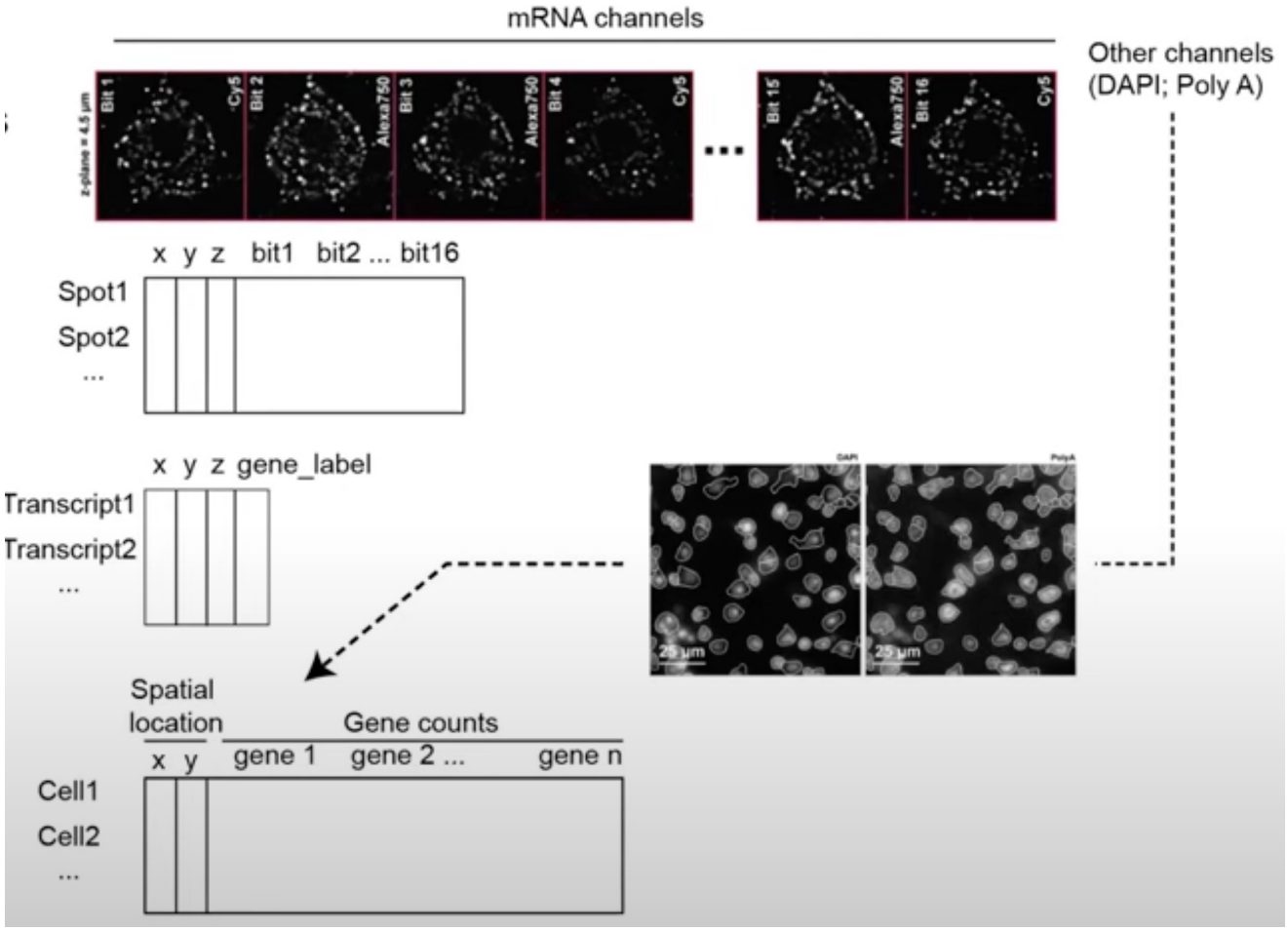
- Slide-seqV2 keeps the 10 µm resolution but has much higher mRNA capture efficiency





Data from spatial transcriptomics





Common types of downstream analyses:

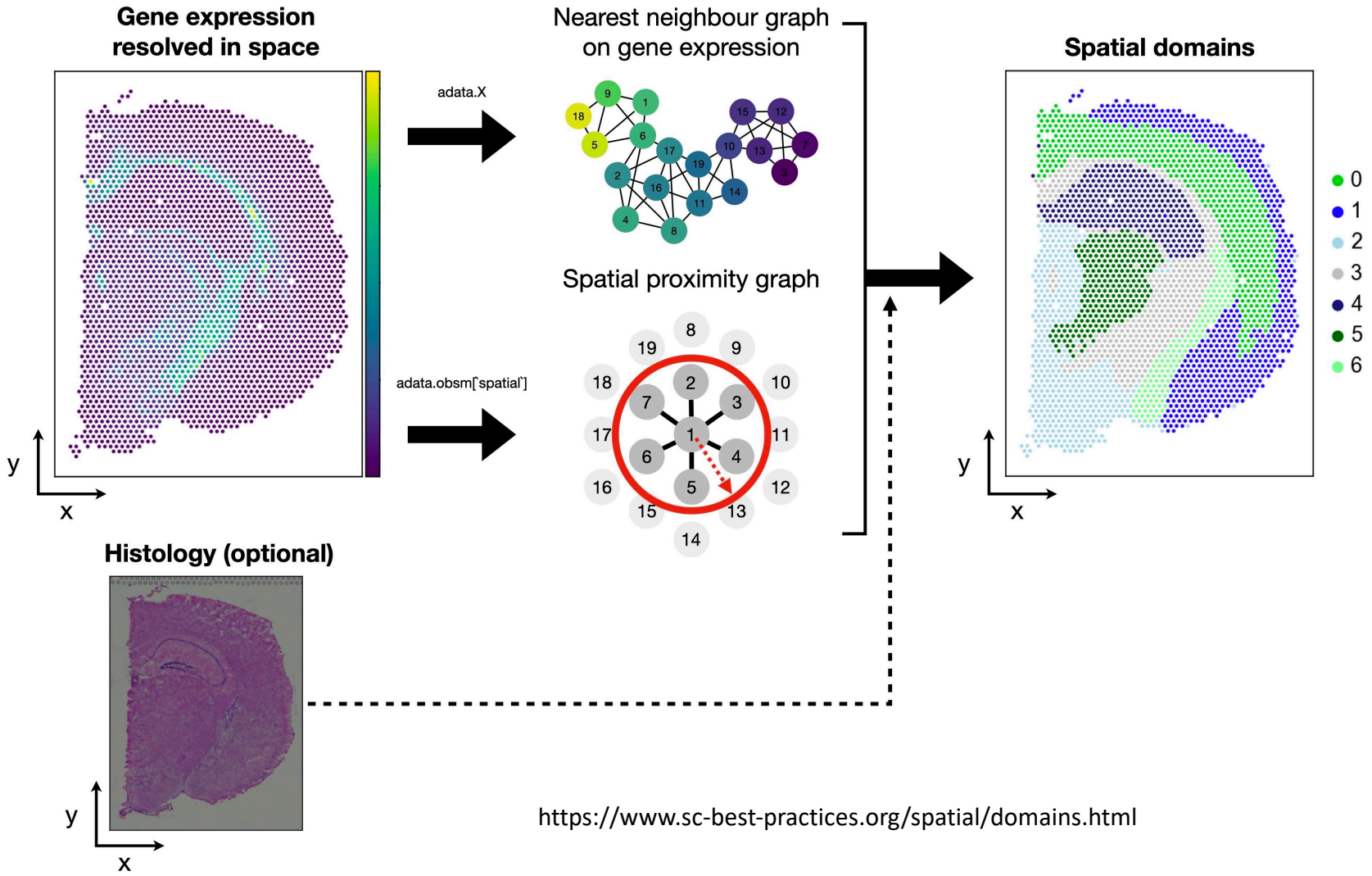
- Spatial domain detection
- Deconvolution and cell type annotation
- Imputation with external data
- Finding spatially variable genes
- Understand cell-cell interactions

https://qcb.ucla.edu/collaboratory/workshops/w31-spatial-transcriptomics/

- Detecting cell boundaries can be a challenge
  (Prabhakaran, Bioinformatics advances, 2022)

Spatial domain detection

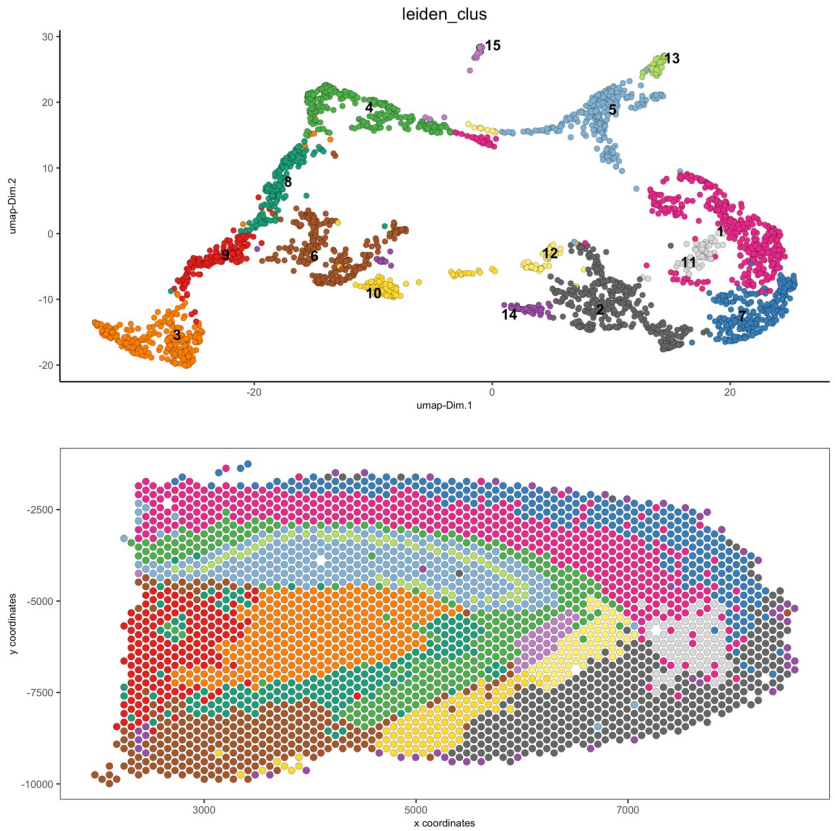
- How to perform clustering of the cells/spots taking spatial coordinates into consideration?



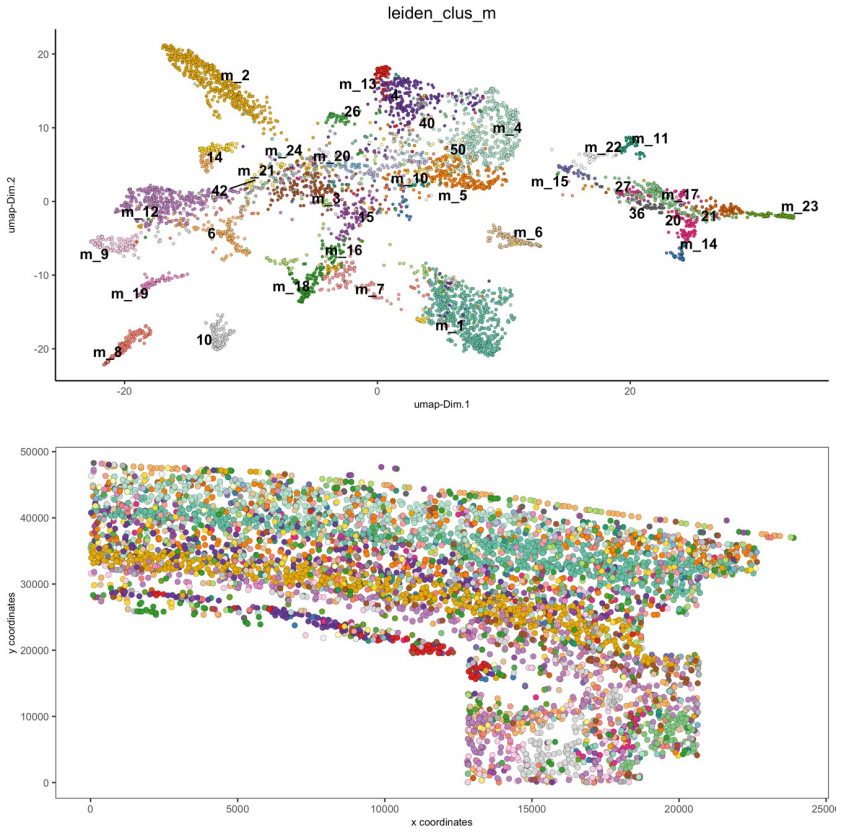
https://www.sc-best-practices.org/spatial/domains.html

Giotto (Dries et. al., Genome Biology, 2021)

- Spatial domain detection in Giotto uses hidden-Markov random field (HMRF)
- Clustering without using spatial information seems not too bad



Visium Brain data



osmFISH mouse SS cortex data

Giotto (Dries et. al., Genome Biology, 2021)

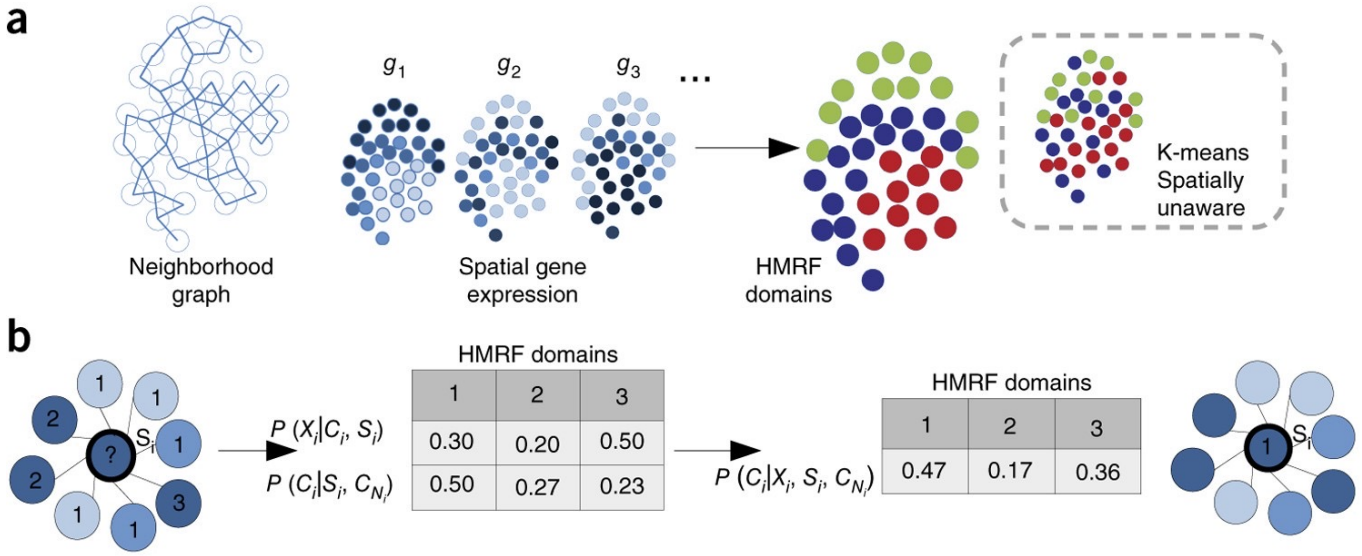
- hidden-Markov random field (HMRF) -> two-dimensional hidden Markov model
- Key assumptions (Zhu et. al., Nature Biotech, 2018):
  - For a spot/cell i, gene expressions given the hidden state c_i are independent across i

$$p(y \mid x, \theta) = \prod_{i \in \mathcal{S}} \mathcal{N}(y_i | x_i = k, \mu_k, \Sigma_k)$$

  - The hidden state c_i depends on hidden states of spatially nearby points (Potts model)

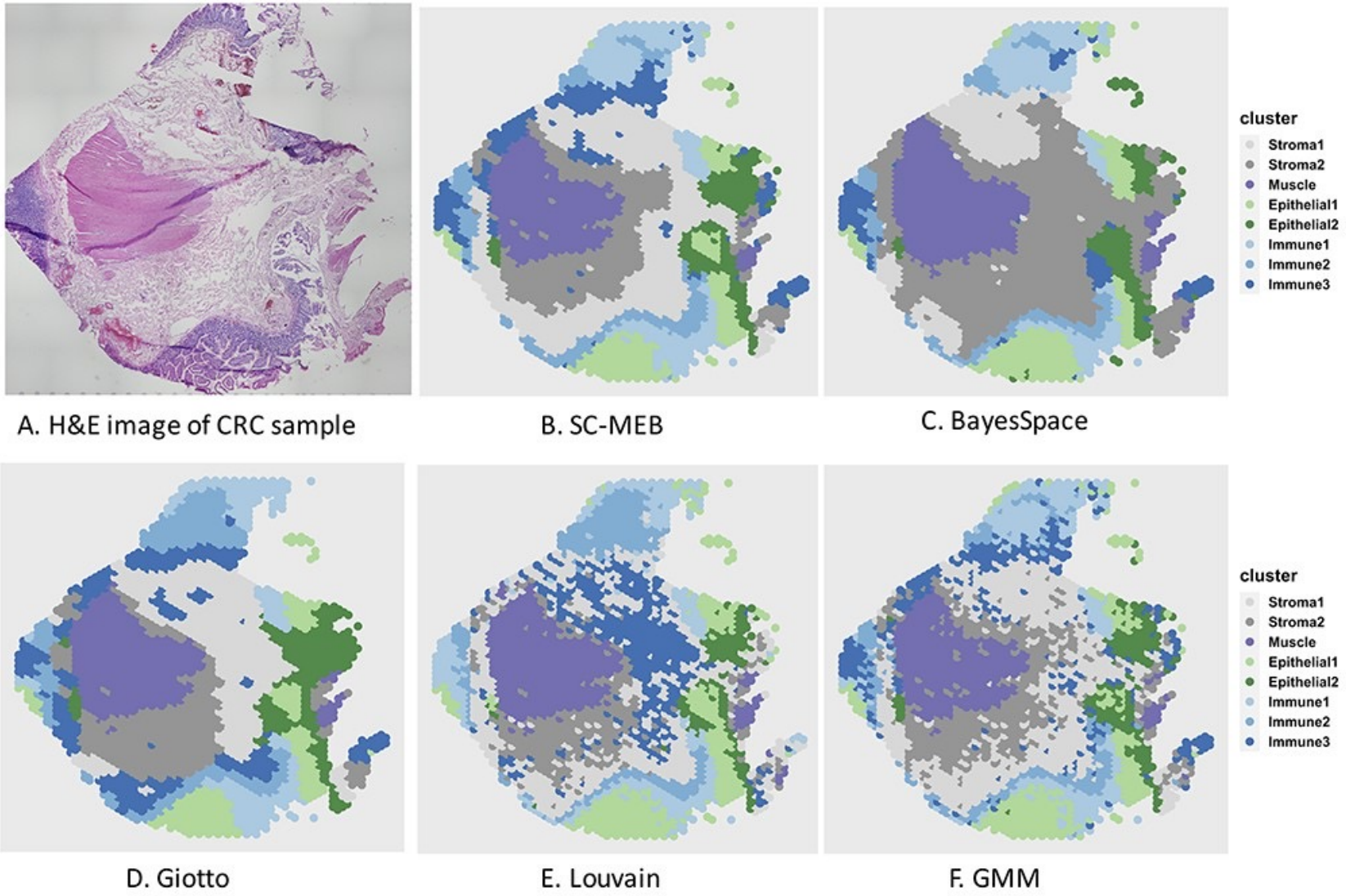
$$P(x; \beta) = \frac{1}{Z_\beta} \exp\{-U(x)\} \qquad U(x) = \sum_{i,i' \in \mathcal{N}_i} \beta[1 - \delta(x_i, x_i')]$$

  - Assign spatial domains / clusters based on the posterior probability of the hidden states



Giotto (Dries et. al., Genome Biology, 2021)

- Performance of HMRF models (SC-MEB, Yang et. al. Briefings in Bioinformatics 2021)

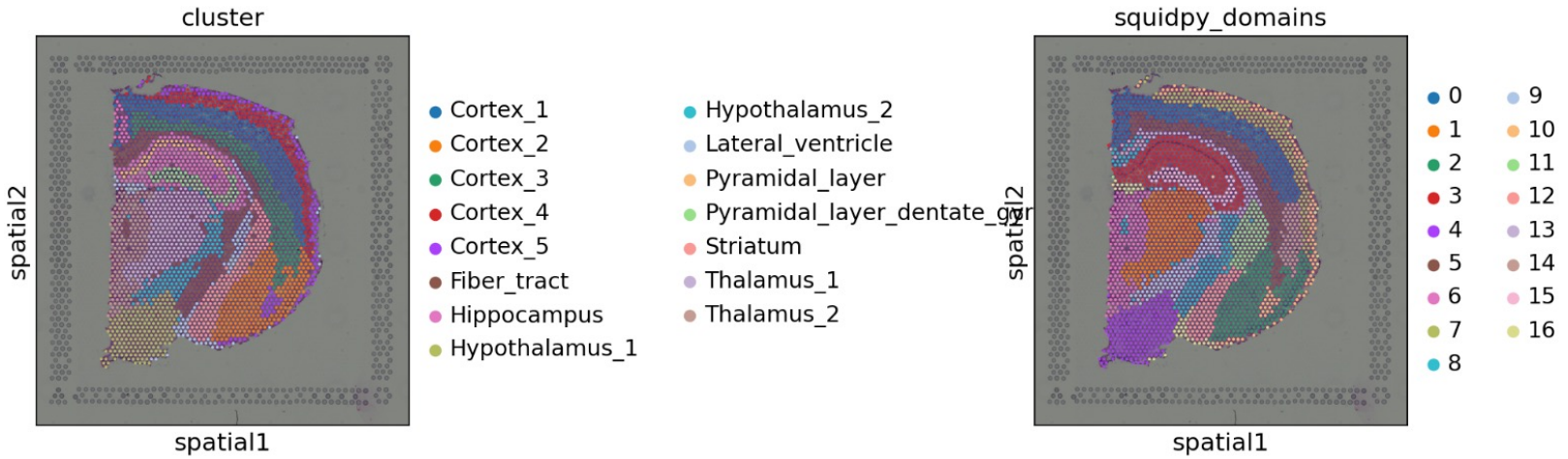


A simple weighted graph method

- Illustrated using Squidpy (Palla et. al., Nature Methods 2022): https://www.sc-best-practices.org/spatial/domains.html#id555
- Idea: spatial smoothing in clustering
  - Compute cell-cell connectivity graph using both graphs:
    - Nearest neighbor graph based on gene expression PCA
    - Nearest neighbor graph based on spatial coordinates
    - Weighted average to create a new graph for clustering

```
alpha = 0.2

joint_graph = (1 - alpha) * nn_graph_genes + alpha * nn_graph_space
sc.tl.leiden(adata, adjacency=joint_graph, key_added="squidpy_domains")
```



SpaGCN (Hu et. al., Nature Methods, 2021)

- Use both histology image and spatial locations to build connectivity graph between two spots
  - Convert histology RGB values to a single value and treat it as a 3rd dimension when calculating cell-cell distances

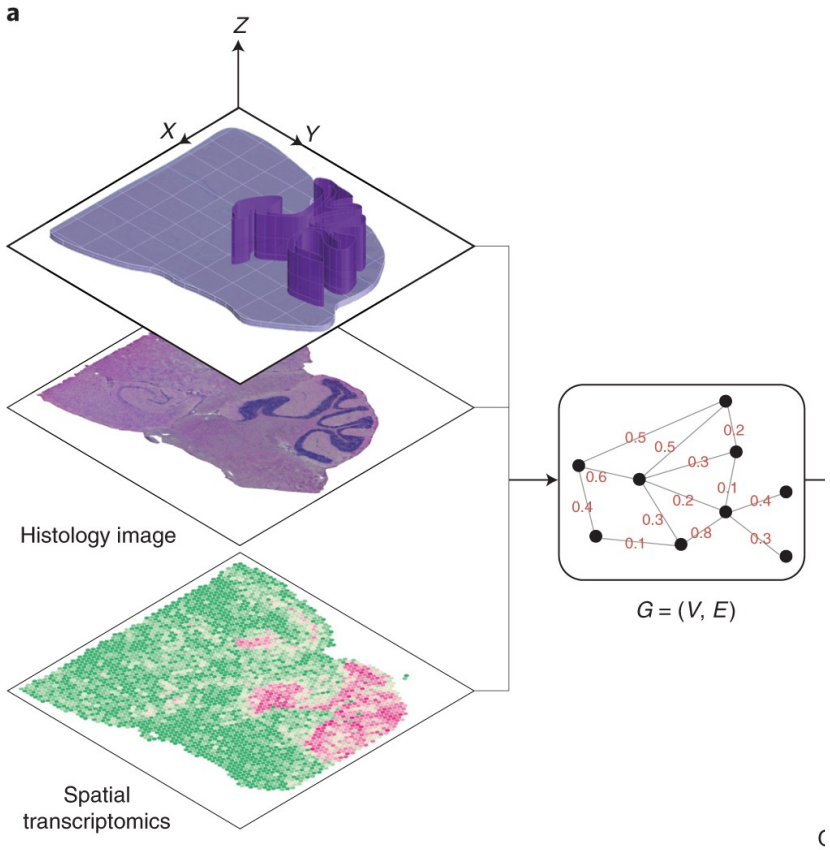
$$z_v = \frac{r_v \times V_r + g_v \times V_g + b_v \times V_b}{V_r + V_g + V_b}$$

$$z_v^* = \frac{z_v - \mu_z}{\sigma_z} \times \max(\sigma_x, \sigma_y) \times s$$

$$d(u,v) = \sqrt{(x_u - x_v)^2 + (y_u - y_v)^2 + (z_u^* - z_v^*)^2}$$

  - Compute cell-cell similarity matrix A
    - Edge weights

$$w(u,v) = \exp\left(-\frac{d(u,v)^2}{2l^2}\right)$$



SpaGCN (Hu et. al., Nature Methods, 2021)

- Use graph convolutional layer to perform smoothing
  - Use the top PCs as input X

$$f(X, A) = \delta(AXB)$$

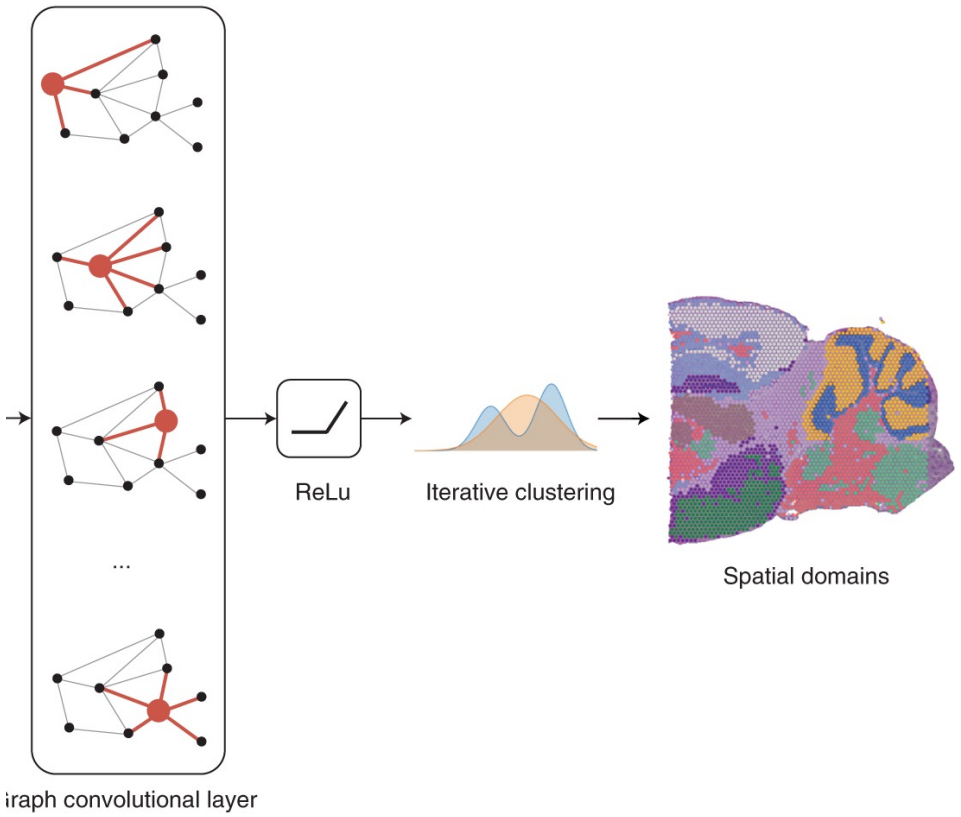
  - Loss function: measuring the clustering performance
    - Perform Louvain clustering on based on the output of the graph convolutional layer
    - Calculate "assignment probability" assuming t-distributions

$$q_{ij} = \frac{\left(1 + h_i - \mu_j^2\right)^{-1}}{\sum_{j'=1}^{K}\left(1 + h_i - \mu_{j'}^2\right)^{-1}}$$

    - Minimize the loss to encourage q_{ij} to be close to 0 or 1

$$L = \mathrm{KL}(P||Q) = \sum_{i=1}^{N}\sum_{j=1}^{K} p_{ij}\log\frac{p_{ij}}{q_{ij}}$$

$$p_{ij} = \frac{q_{ij}^2 / \sum_{i=1}^{N} q_{ij}}{\sum_{j'=1}^{K}\left(q_{ij'}^2 / \sum_{i=1}^{N} q_{ij'}\right)}$$

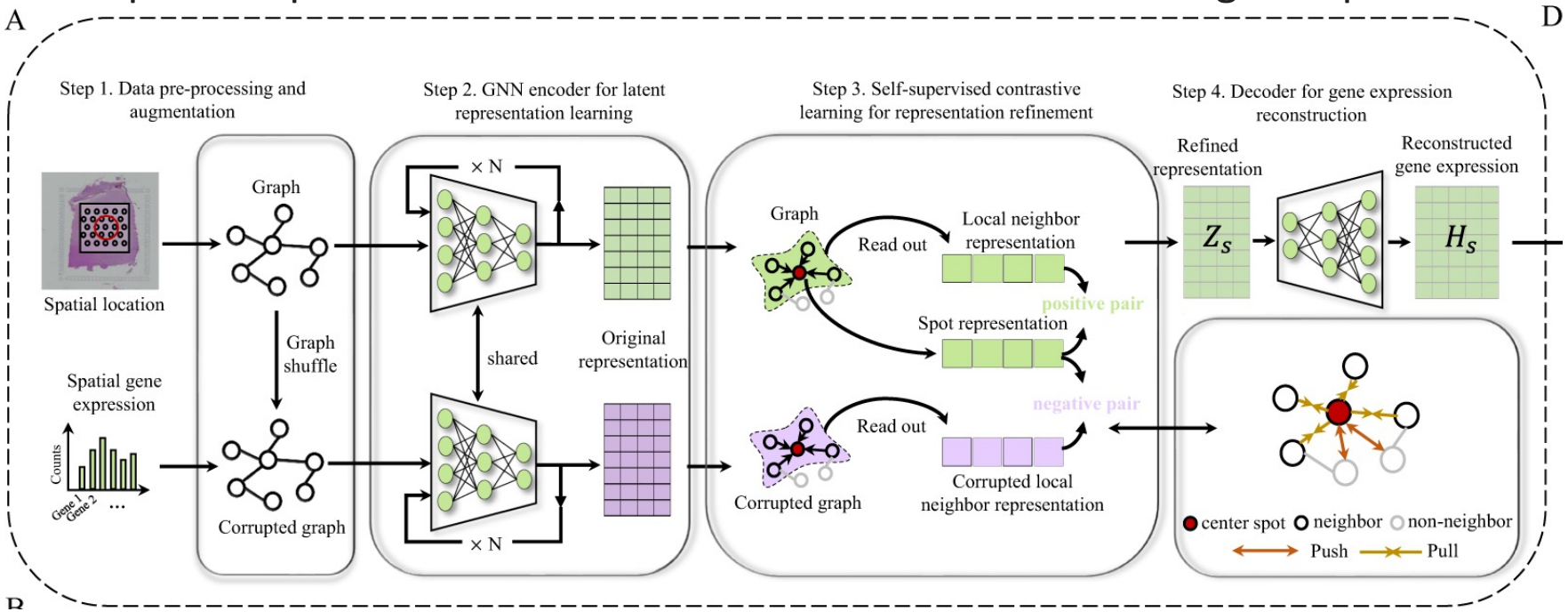


GraphST (Long et. al. Nature Comm, 2023)

- Main idea: GNN + self-supervised contrastive learning
  - Build KNN graph using spatial locations and obtain adjacency matrix A
  - Use graph convolutional network to build the encoder

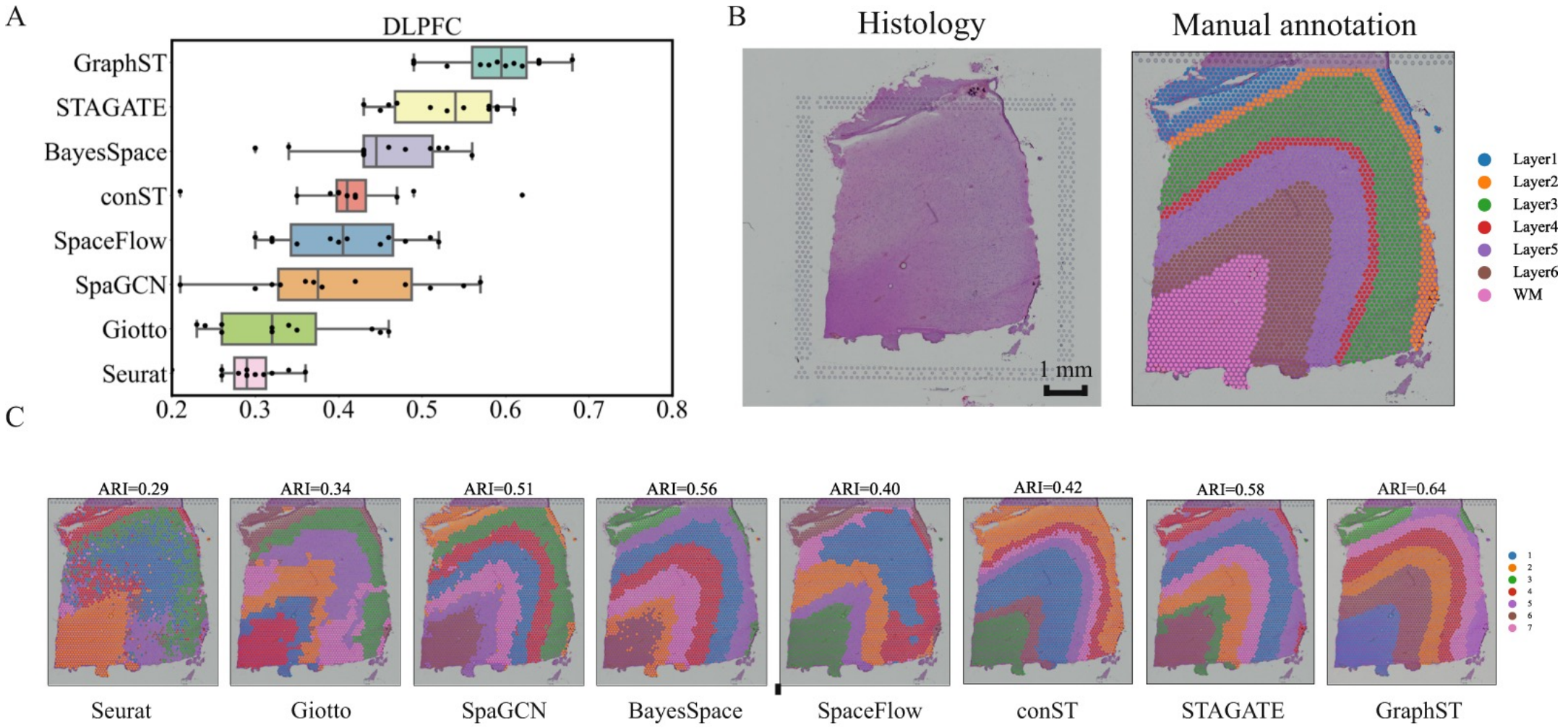
$$Z_s^l = \sigma\left(\tilde{A} Z_s^{l-1} W_e^{l-1} + b_e^{l-1}\right) \quad \tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$$

  - Contrastive learning : generate corrupted graph by randomly permute cell labels
    - Make positive pairs more similar to each other and contrast negative pairs



  - Perform standard clustering on reconstructed gene expression matrix + surrounding refinement

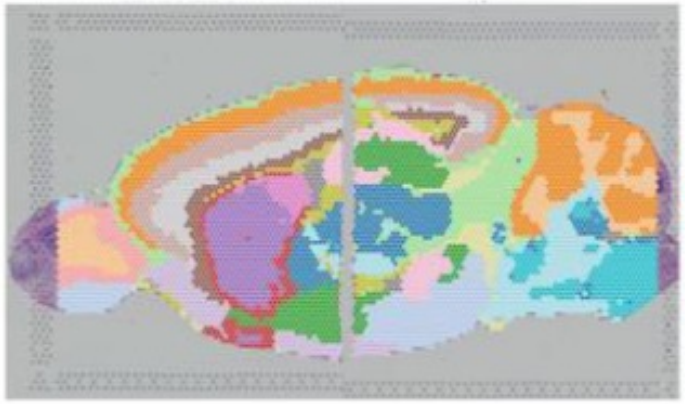
GraphST (Long et. al. Nature Comm, 2023)



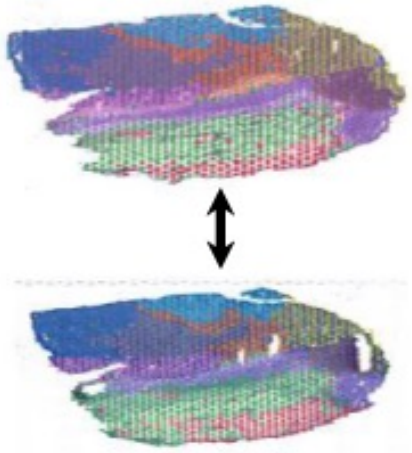
Integration of spatial transcriptomics data

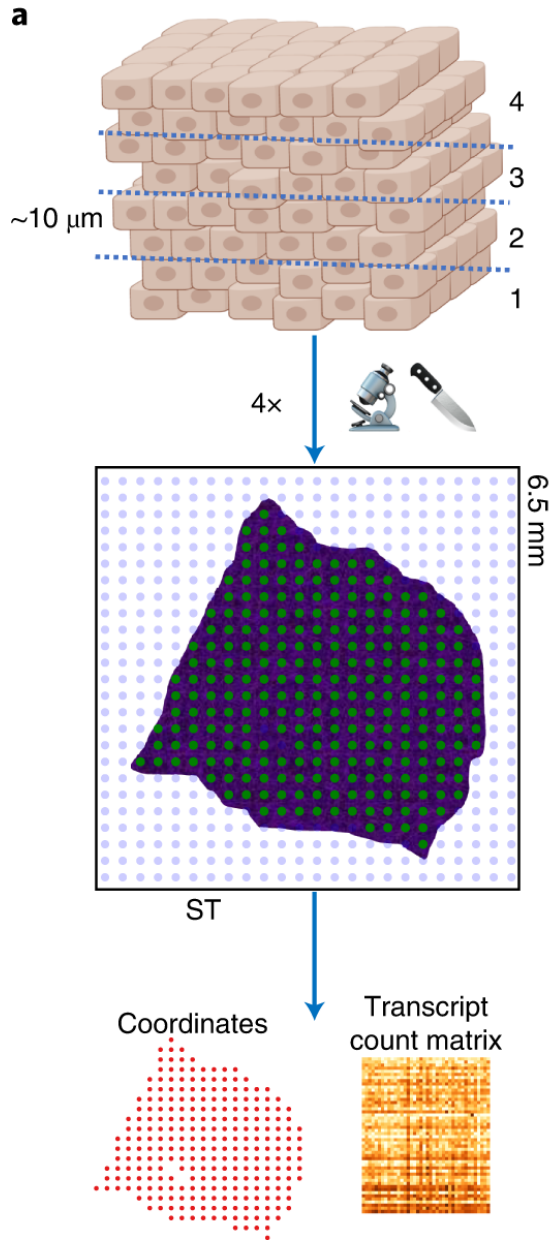
- Tissue sample can be dissected into multiple sections
- Serial tissue slices can be used to infer 3D information

b) Horizontal Integration



a) Vertical Integration





- Extract 3D spatial domains

- Challenge: placement and orientation of the tissue on the array can be arbitrary

PASTE (Zeira et. al., Nature Methods 2022)

- Pairwise alignment of ST slices
  - Convert spatial coordinate matrix to spatial distance matrix between any two spots on the same slice s
  - Define alignment matrix $\mathbf{\Pi} = [\pi_{ij}] \in \mathbb{R}_{+}^{n \times n'}$
  - Minimize the transport cost

$$F(\Pi;\, X, D, X', D', c, \alpha) = (1-\alpha)\sum_{i,j} c(x_{\cdot i}, x'_{\cdot j})\pi_{ij} + \alpha \sum_{i,j,k,l} (d_{ik} - d'_{jl})^2 \pi_{ij}\pi_{kl} \quad (1)$$

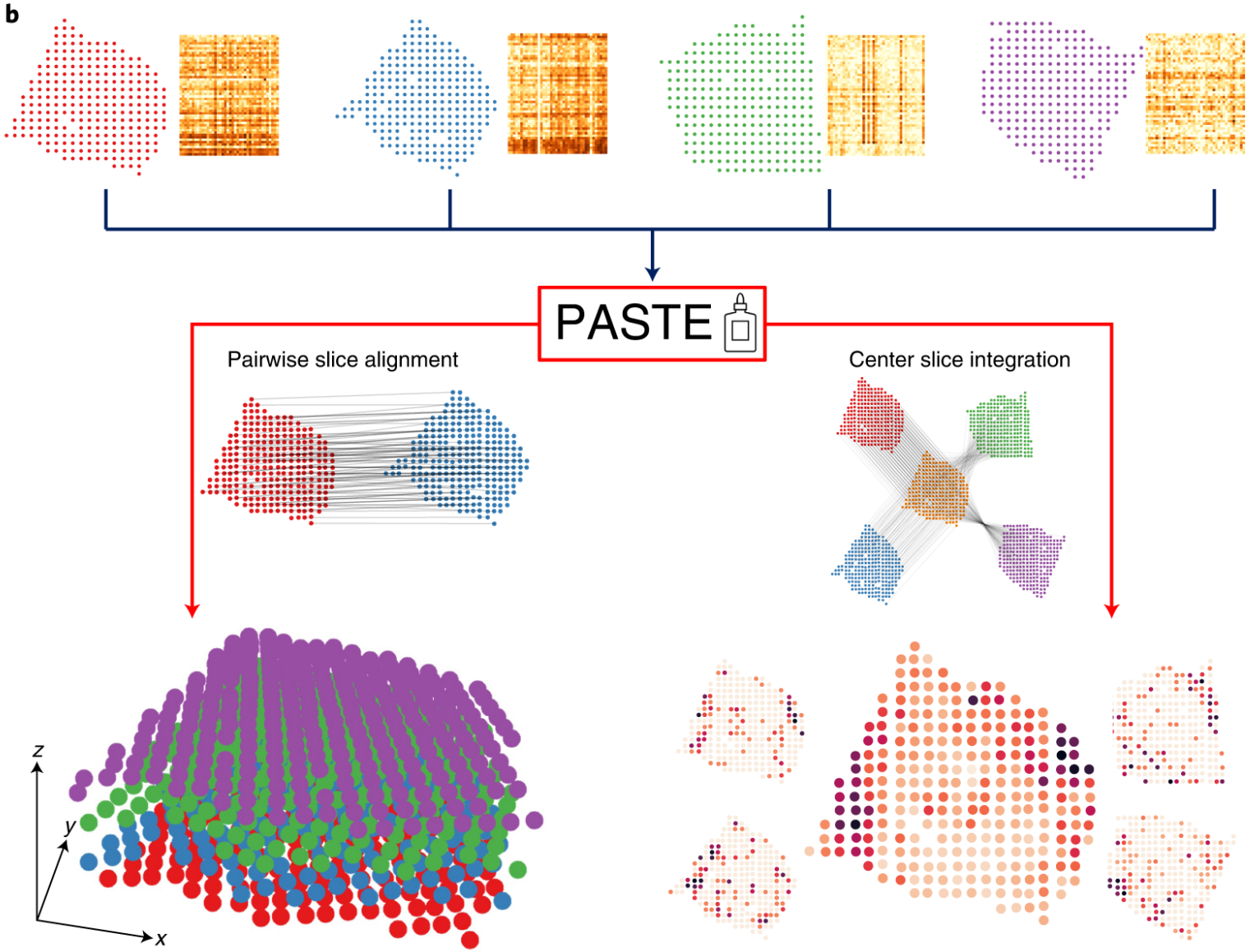
  - Computational cost: $O(n^2 n' + n n'^2)$

- Reconstruct stacked 3D spatial representation
  - Obtain pairwise alignment matrix between adjacent slices
  - Estimate a shared rotation matrix and translation vector across slices

$$\hat{R}, \hat{v} = \min_{\substack{R \in \mathbb{R}^{2\times 2}, v \in \mathbb{R}^2 \\ R^T R = I}} \sum_{i,j} \pi_{ij}^{(k)} ||z_{\cdot i}^{(k)} - R z_{\cdot j}^{(k+1)} - v||^2$$

- Construct a center slice to represent all slices if the slices are similar to each other

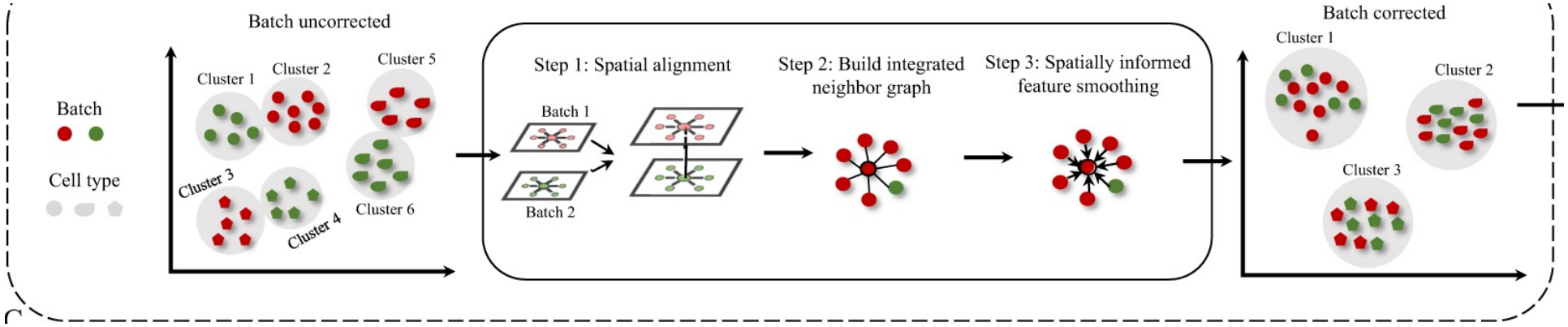
PASTE (Zeira et. al., Nature Methods 2022)



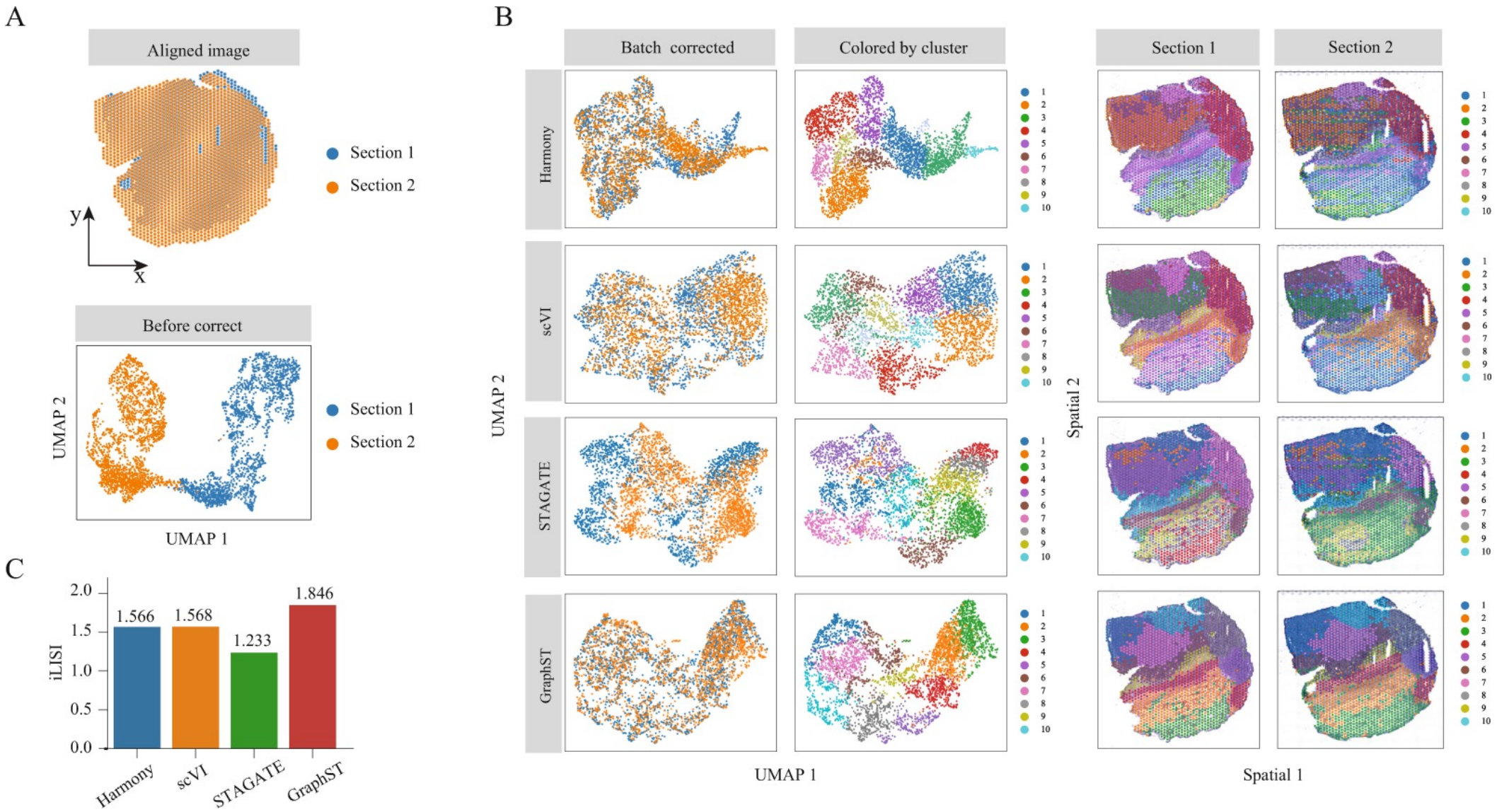
Integrative domain detection using GraphST

- Align the spatial locations
  - Horizontal integration
    - Align the two histological image to ensure slices are adjacent in space
  - Vertical integration
    - Use PASTE to align the coordinates

- Joint neighborhood construction
  - Construct neighborhood graph including both intra-slice and inter-slice adjacent spots
- Train all slices together
  - Implicitly removes batch effects



Integrative domain detection using GraphST



Related papers

- Longo, S. K., Guo, M. G., Ji, A. L., & Khavari, P. A. (2021). Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nature Reviews Genetics*, *22*(10), 627-644.
- Atta, L., & Fan, J. (2021). Computational challenges and opportunities in spatially resolved transcriptomic data analysis. *Nature Communications*, *12*(1), 5283.
- Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). Spatially resolved, highly multiplexed RNA profiling in single cells. *Science*, *348*(6233), aaa6090.
- Ståhl, P. L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., ... & Frisén, J. (2016). Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78-82.
- Rodriques, S. G., Stickels, R. R., Goeva, A., Martin, C. A., Murray, E., Vanderburg, C. R., ... & Macosko, E. Z. (2019). Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, *363*(6434), 1463-1467.
- Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., ... & Chen, F. (2021). Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2. *Nature biotechnology*, *39*(3), 313-319.
- Prabhakaran, S. (2022). Sparcle: assigning transcripts to cells in multiplexed images. *Bioinformatics Advances*, *2*(1), vbac048.
- Dries, R., Zhu, Q., Dong, R., Eng, C. H. L., Li, H., Liu, K., ... & Yuan, G. C. (2021). Giotto: a toolbox for integrative analysis and visualization of spatial expression data. *Genome biology*, *22*, 1-31.
- Zhu, Q., Shah, S., Dries, R., Cai, L., & Yuan, G. C. (2018). Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data. *Nature biotechnology*, *36*(12), 1183-1190.
- Yang, Y., Shi, X., Liu, W., Zhou, Q., Chan Lau, M., Chun Tatt Lim, J., ... & Liu, J. (2022). SC-MEB: spatial clustering with hidden Markov random field using empirical Bayes. *Briefings in bioinformatics*, *23*(1), bbab466.
- Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A. C., Kuemmerle, L. B., ... & Theis, F. J. (2022). Squidpy: a scalable framework for spatial omics analysis. *Nature methods*, *19*(2), 171-178.
- Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., ... & Li, M. (2021). SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nature methods*, *18*(11), 1342-1351.
- Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., ... & Chen, J. (2023). Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nature Communications*, *14*(1), 1155.
- Zeira, R., Land, M., Strzalkowski, A., & Raphael, B. J. (2022). Alignment and integration of spatial transcriptomics data. *Nature Methods*, *19*(5), 567-575.