

Statistical Methods for Spatial Transcriptomics

Xiang Zhou

Department of Statistics and Data Science
Yale University

Spatial transcriptomics

Spatial transcriptomics may be viewed as a pointillist painting.

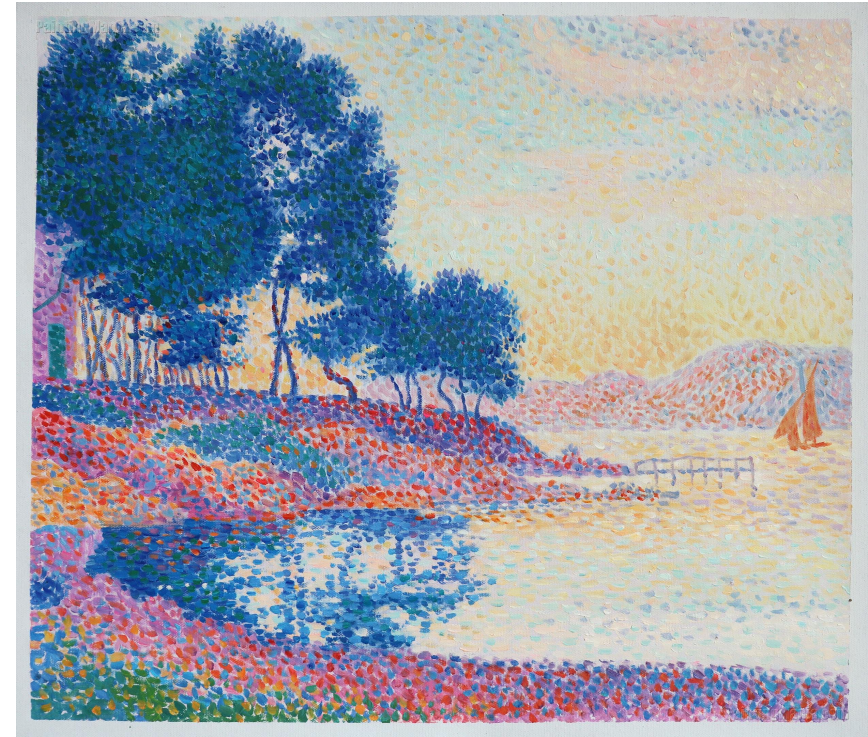
Bulk RNAseq



Single Cell RNA-seq

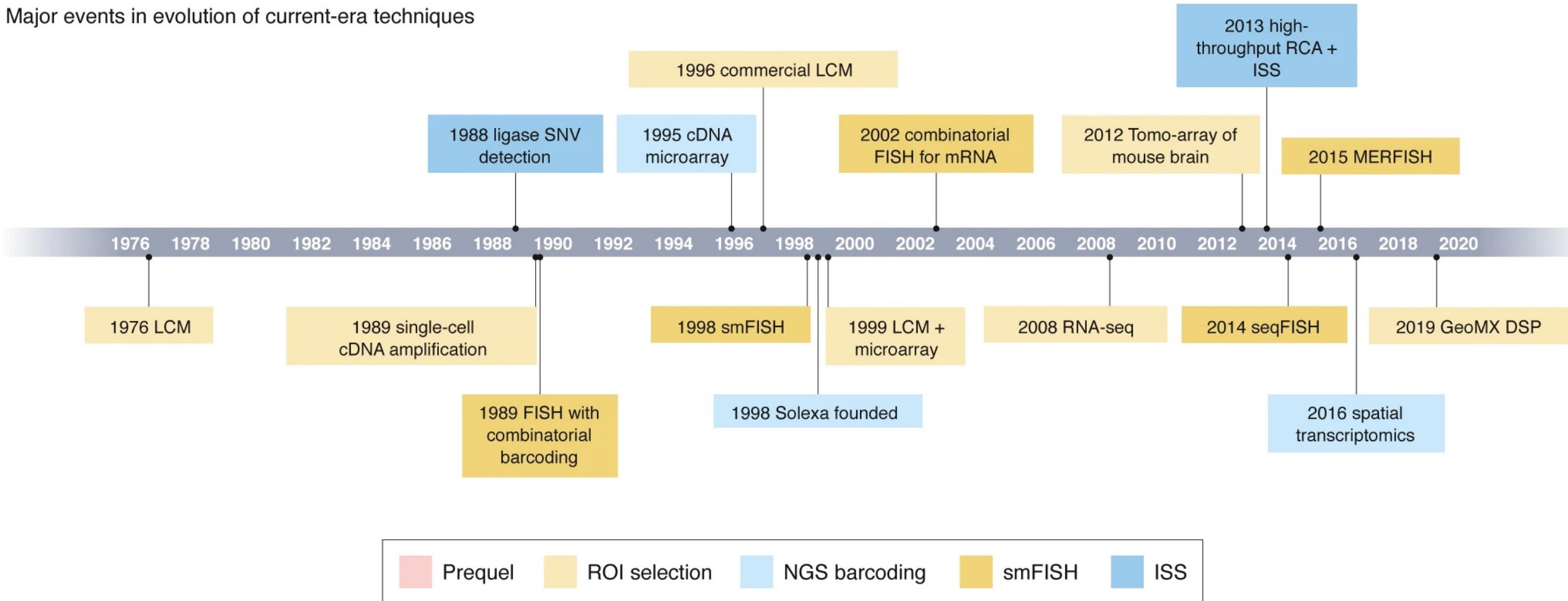


Spatial Transcriptomics



Spatial Transcriptomics

Major events in evolution of current-era techniques



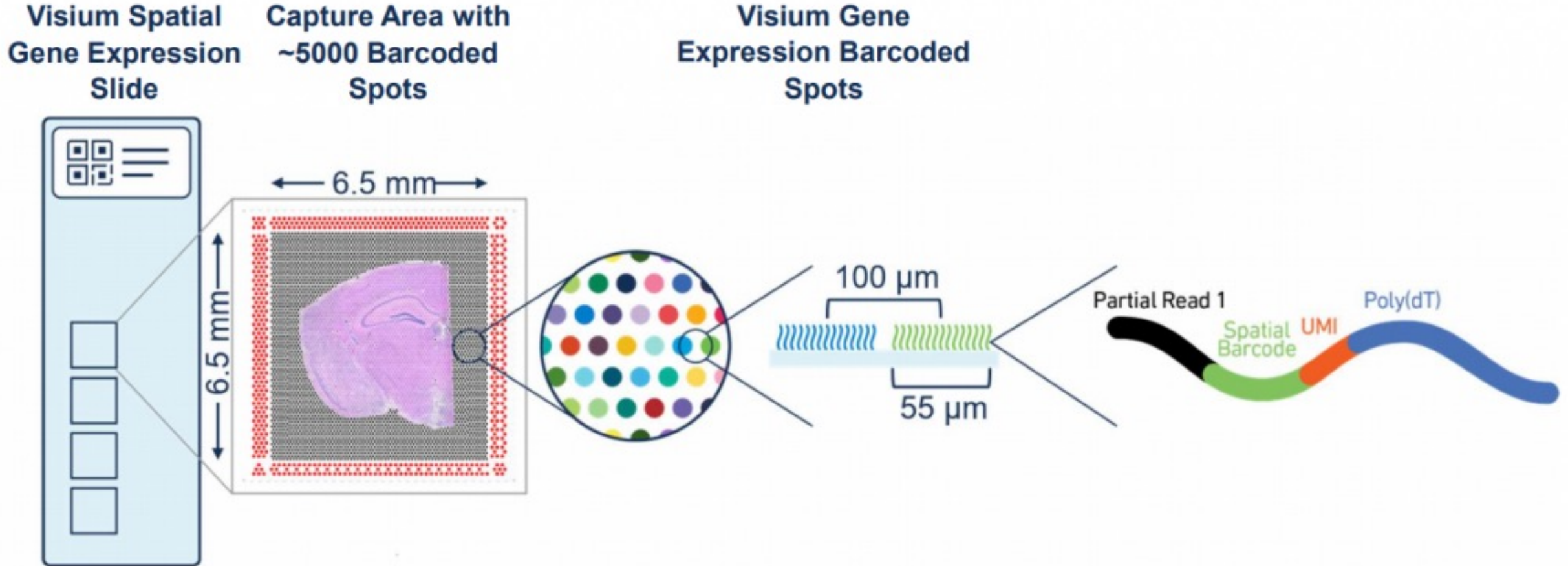
Categorization of Spatial transcriptomics

- Gene expression profiling on tissue sections with spatial localization information.
- Some techniques are based on sequencing:
 - Spatial transcriptomics (10x genomics visium)
 - GeoMx
 - Slide-seq (v1, v2)
 - Seq-Scope
- Some techniques are based on smFISH:
 - Multiplexed error-robust FISH (MERFISH)
 - Sequential FISH (seqFISH, seqFISH+)
 - CosMx
 - 10x Xenium

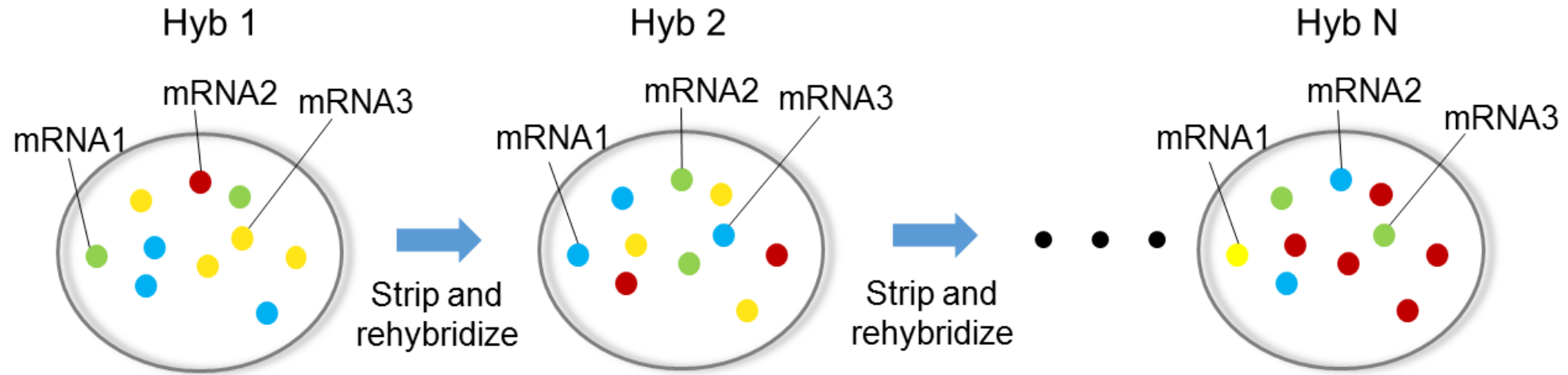
Resolution of Spatial transcriptomics

- Common sequencing-based techniques, such as 10x Visium, are of spot level resolution, with each spot containing 2-20 cells.
- Recent sequencing-based techniques, such as Seq-Scope, VisiumHD, Open-ST, and Stereo-seq, offer resolution of 0.5-2 μm .
- *In situ* imaging techniques, such as MERFISH, SeqFISH+, MERSCOPE, CosMx, and 10X Xenium, provide spatial resolutions as fine as 0.1-0.2 μm .
- Regardless of technologies, two output matrices are generated (n : number of locations; p : number of genes):
 - an n by p count matrix of gene expression
 - an n by 2 location matrix

10x Genomics Visium



SeqFISH+



Barcode # scales as F^N

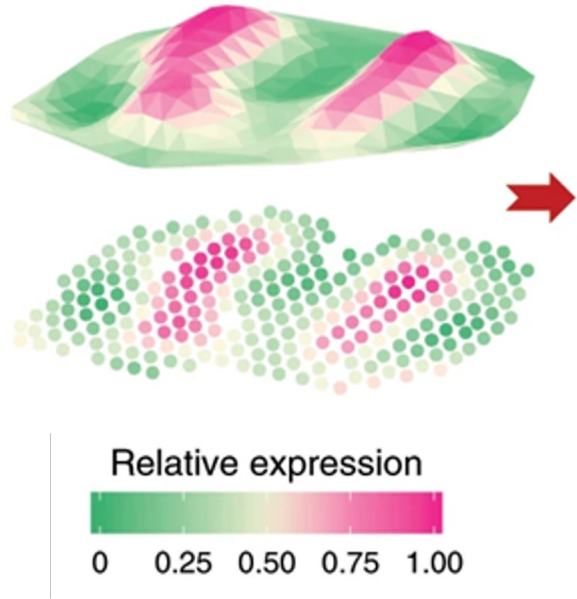
$$5^3=125$$

$$5^6=15,625$$

Barcode Key: (for N=6)

mRNA1	●	●	●	●	●	●
mRNA2	●	●	●	●	●	●
mRNA3	●	●	●	●	●	●

Common Analytic Tasks



Data Features:
Noisy Measurements
Non-linearity

Spatially Variable Gene Detection

Cell Type Deconvolution

Tissue Segmentation and Spatial Domain Detection

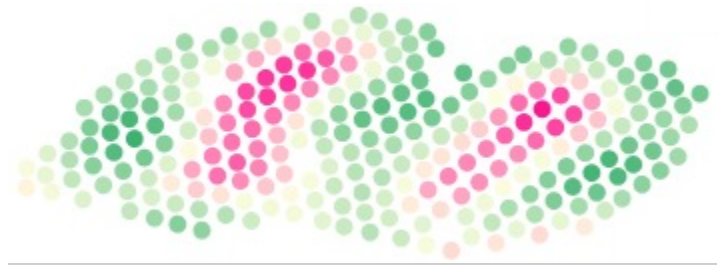
Cell Type Clustering

Cell Cell Communication

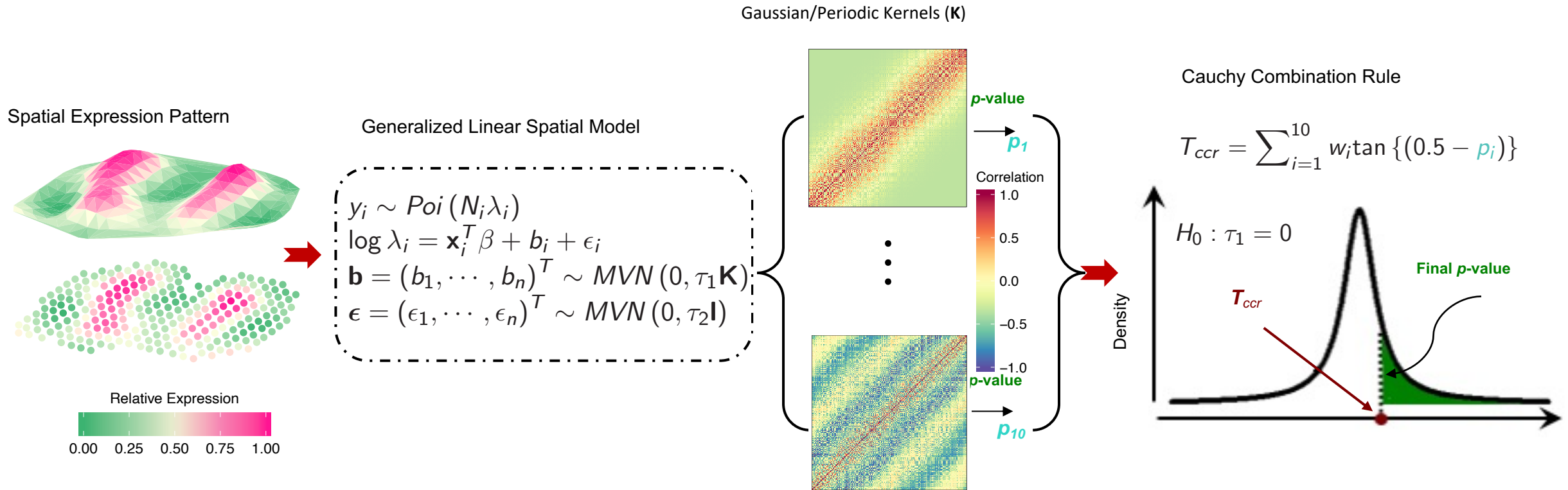
Multi-Sample Analysis

Cell Boundary Detection

Detecting Spatially Variable Genes



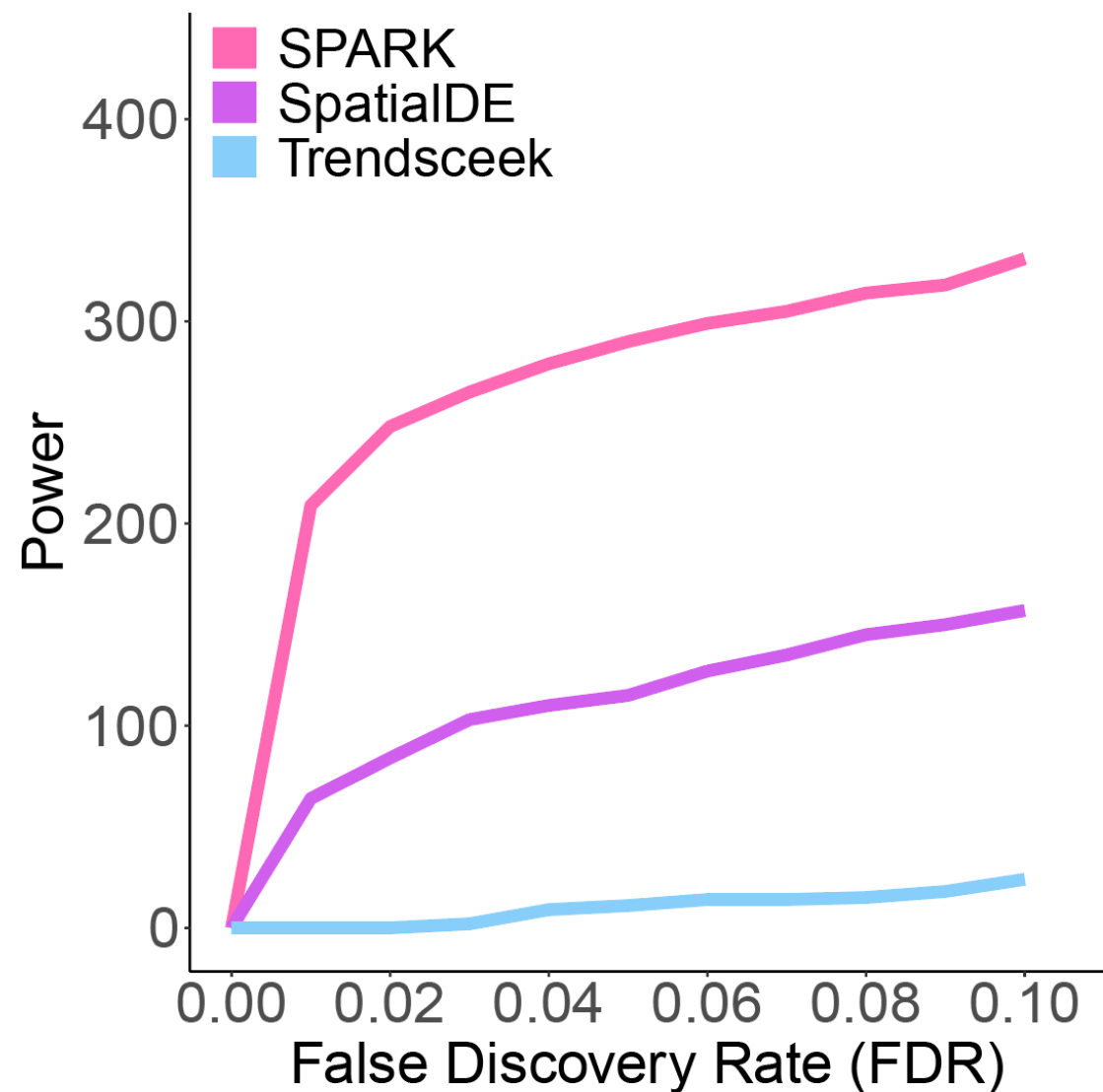
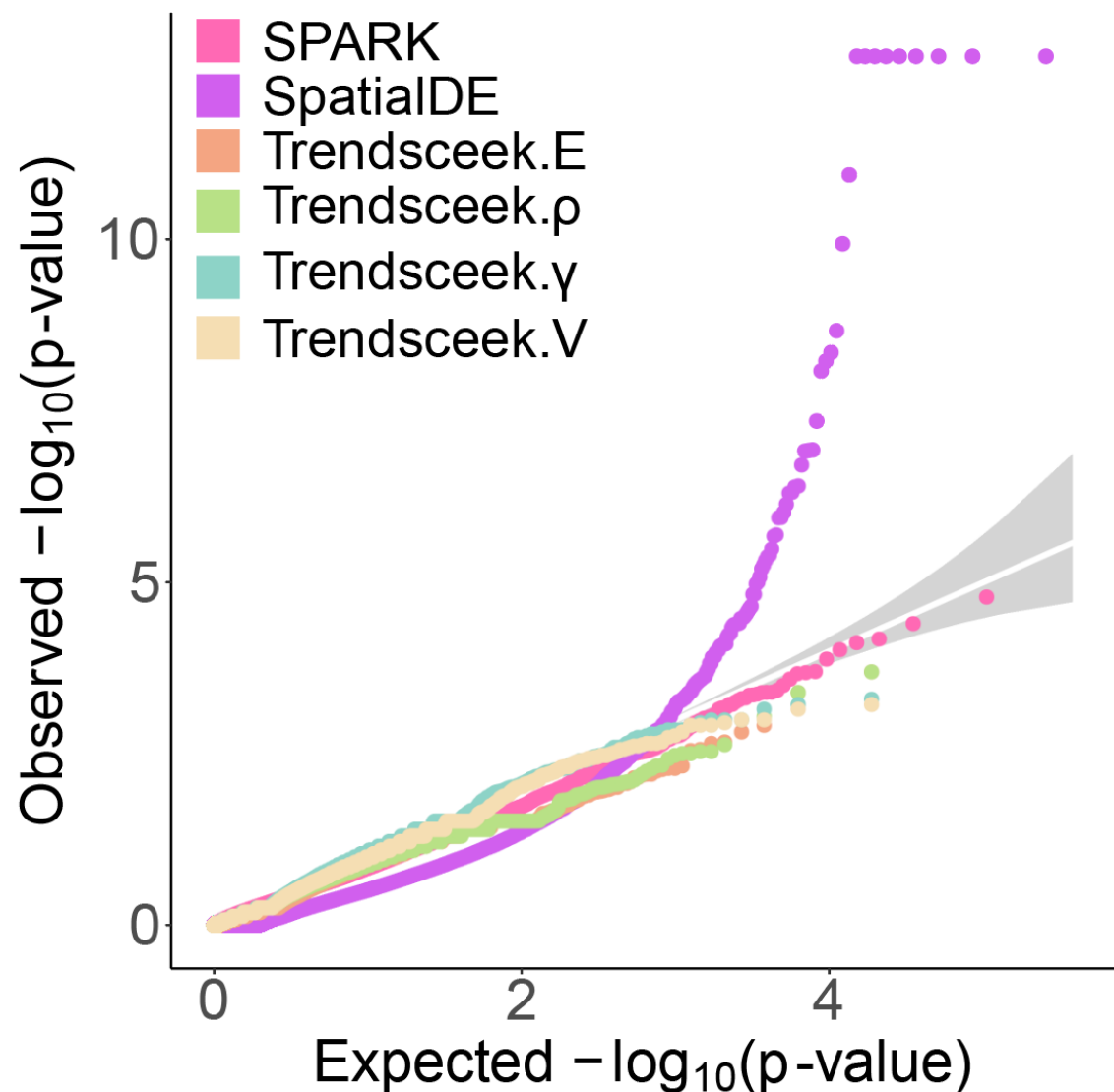
SPARK Schematic



Human Breast Cancer Data

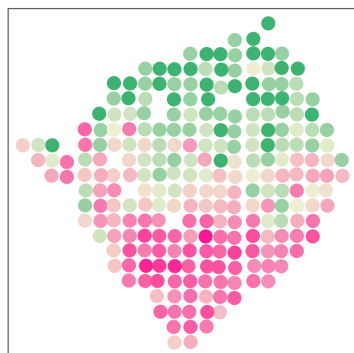
- Spatial transcriptomics sequencing technique (Stahl 2016).
- Filtered out genes that are expressed in less than 10% of the array spots and selected spots with at least 10 total reads.
- 5,262 genes measured on 250 spots.

Type I Error Control and Power

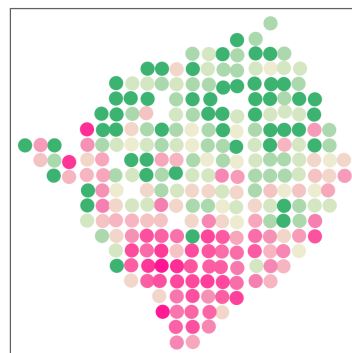


SPARK Detected All 10 Known Cancer Genes

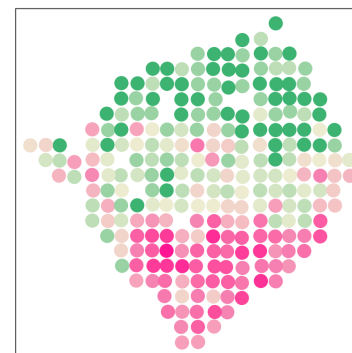
POSTN ($6e-17$)



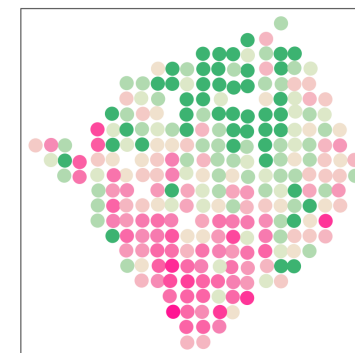
SPARC ($6e-17$)



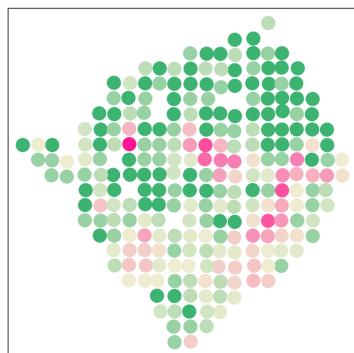
FN1 ($6e-17$)



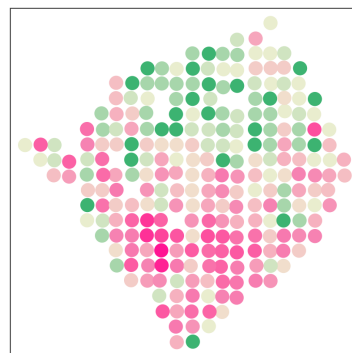
DCN ($2e-15$)



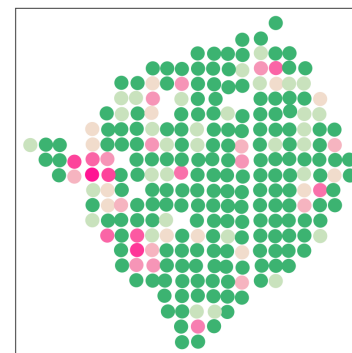
IGFBP5 ($9e-11$)



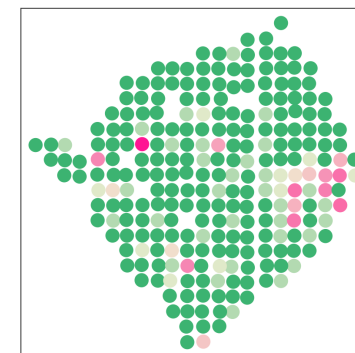
VIM ($2e-09$)



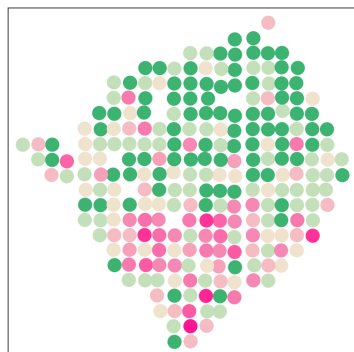
KRT17 ($2e-08$)



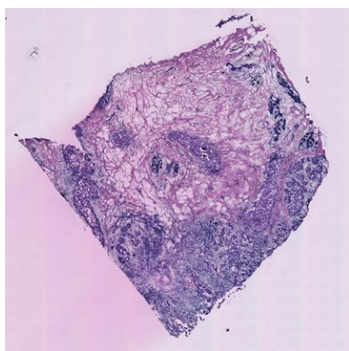
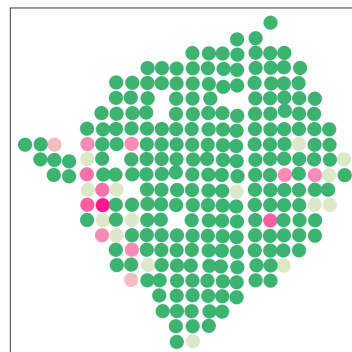
SCGB2A2 ($5e-06$)



MMP14 ($3e-05$)

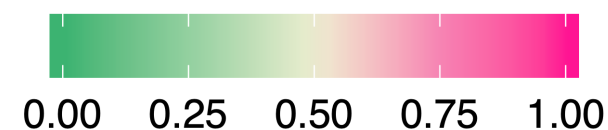


PEG10 ($1e-04$)

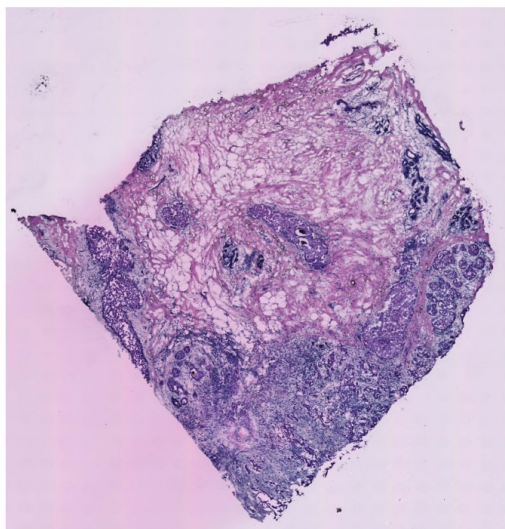


Other two methods
detected 7

Relative Expression



A Few New Examples



HLA-B (5.6×10^{-17})

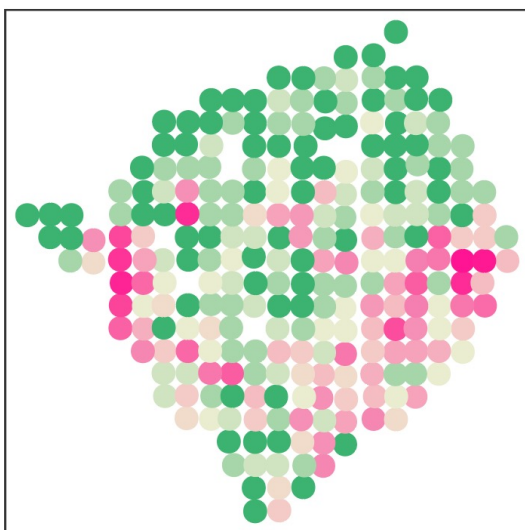


EEF1A1 (3×10^{-7})

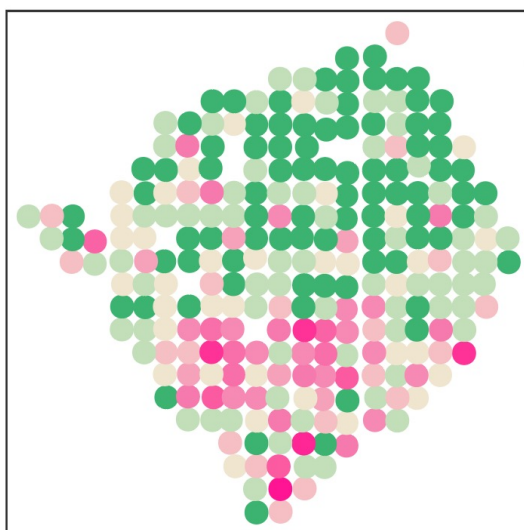


4 mm

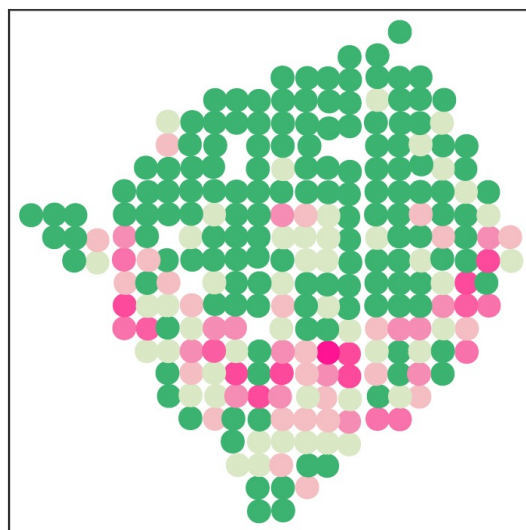
ERBB2 (2.9×10^{-6})



MMP14 (2.5×10^{-5})



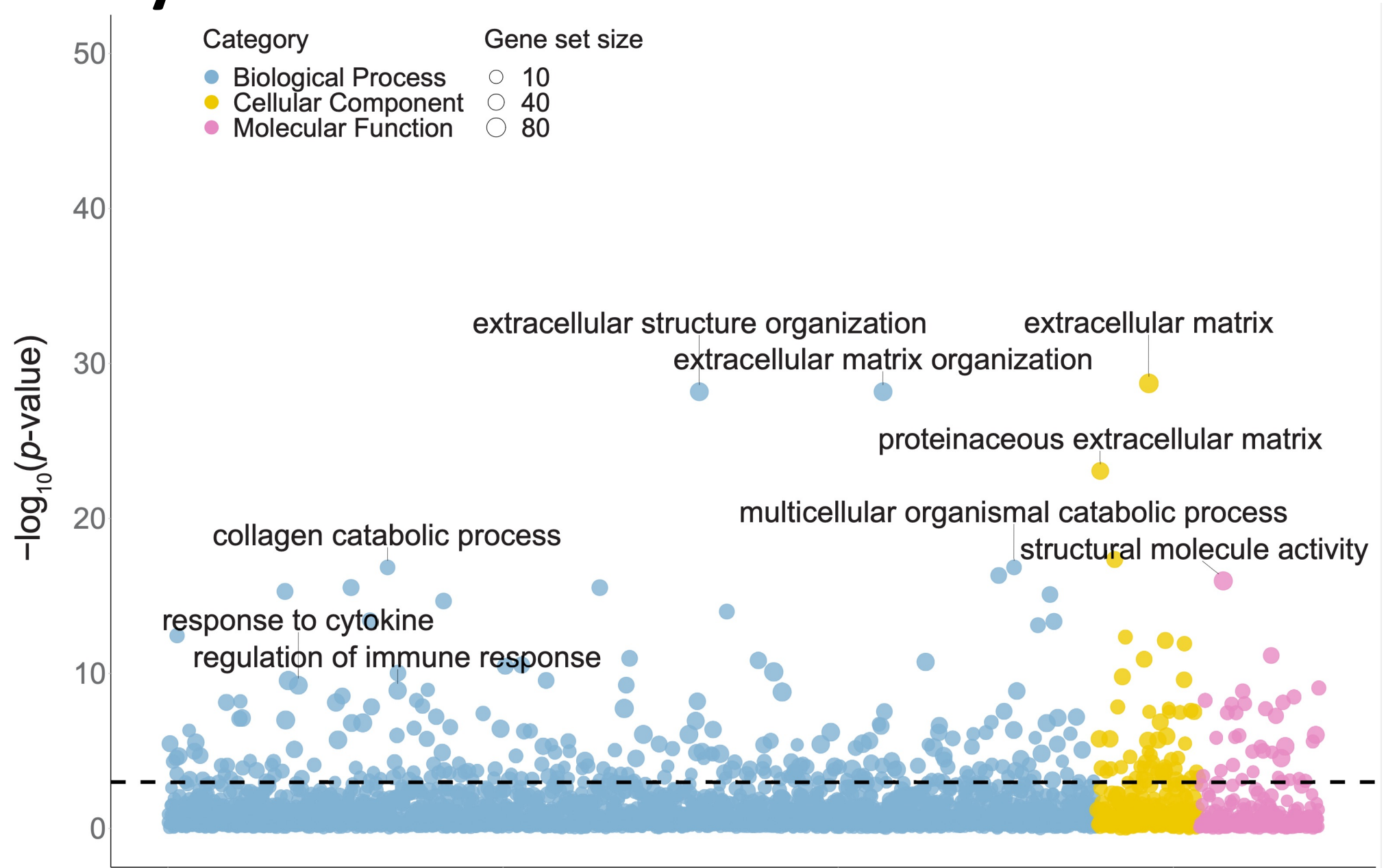
CD44 (1.9×10^{-4})



4 mm

Relative Expression
1.00
0.75
0.50
0.25
0.00

GSEA Analysis



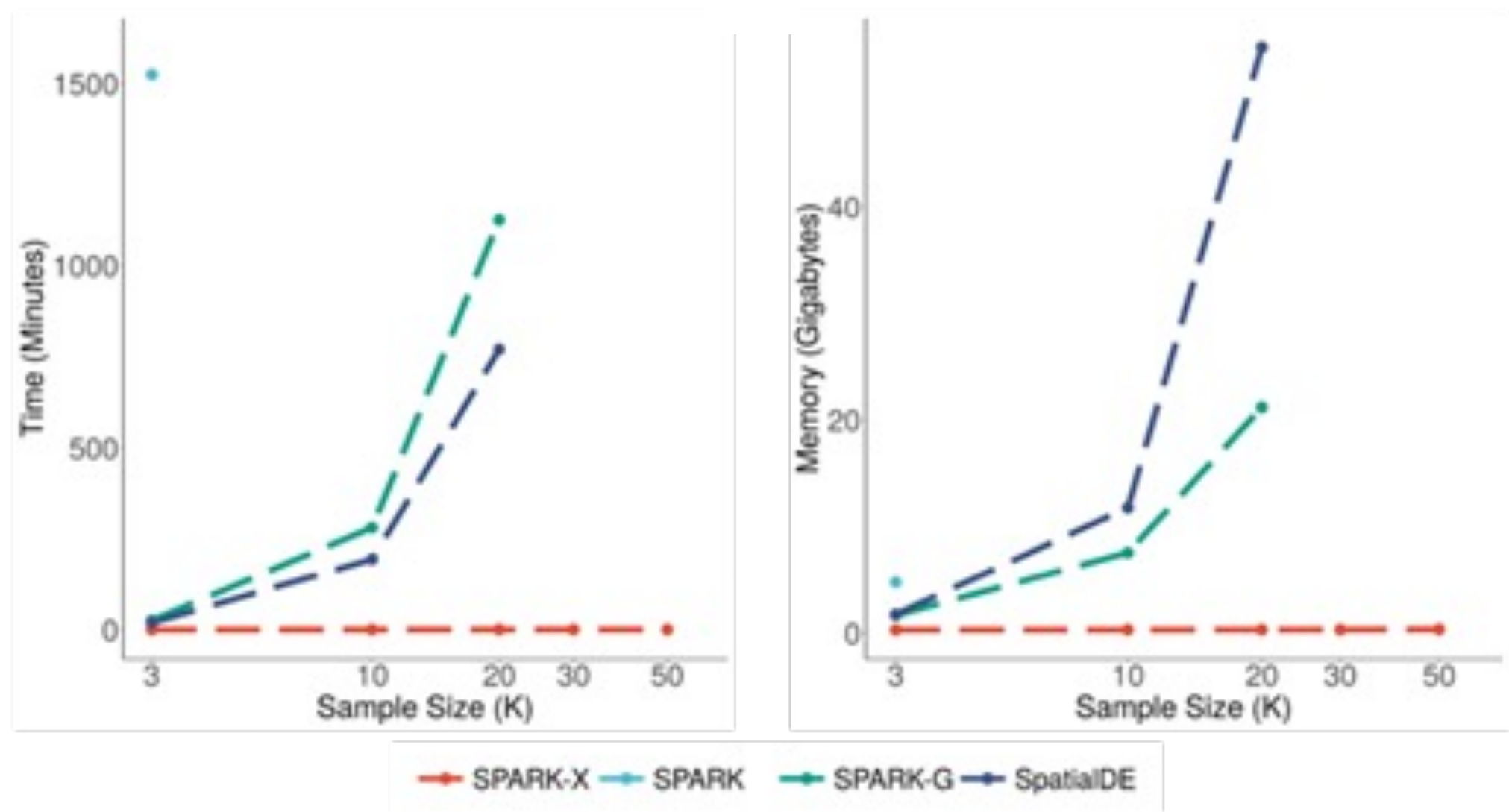
New Data Types and New Challenges

- Increased sample size
 - Recent techniques, such as Slide-seq and HDST, are of high resolution and measure tens to hundreds of thousands of locations.
- Low counts and prevalence of zeros
 - The resulting data, limited by sequencing depth, are also in sparse forms, with a prevalence of low counts and zeros.
- Computational and statistical challenges.

SPARK-X: SPARK eXpedited

- Based on a general class of covariance tests, including the Hilbert-Schmidt independence test and distance covariance test, for non-parametric SE analysis.
- Procedure:
 - Construct an expression kernel matrix $\mathbf{E}$.
 - Construct a distance kernel matrix $\mathbf{\Sigma}$.
 - Test the similarity between the two matrices by $T = \text{trace}(\mathbf{E}\mathbf{\Sigma})/n$.

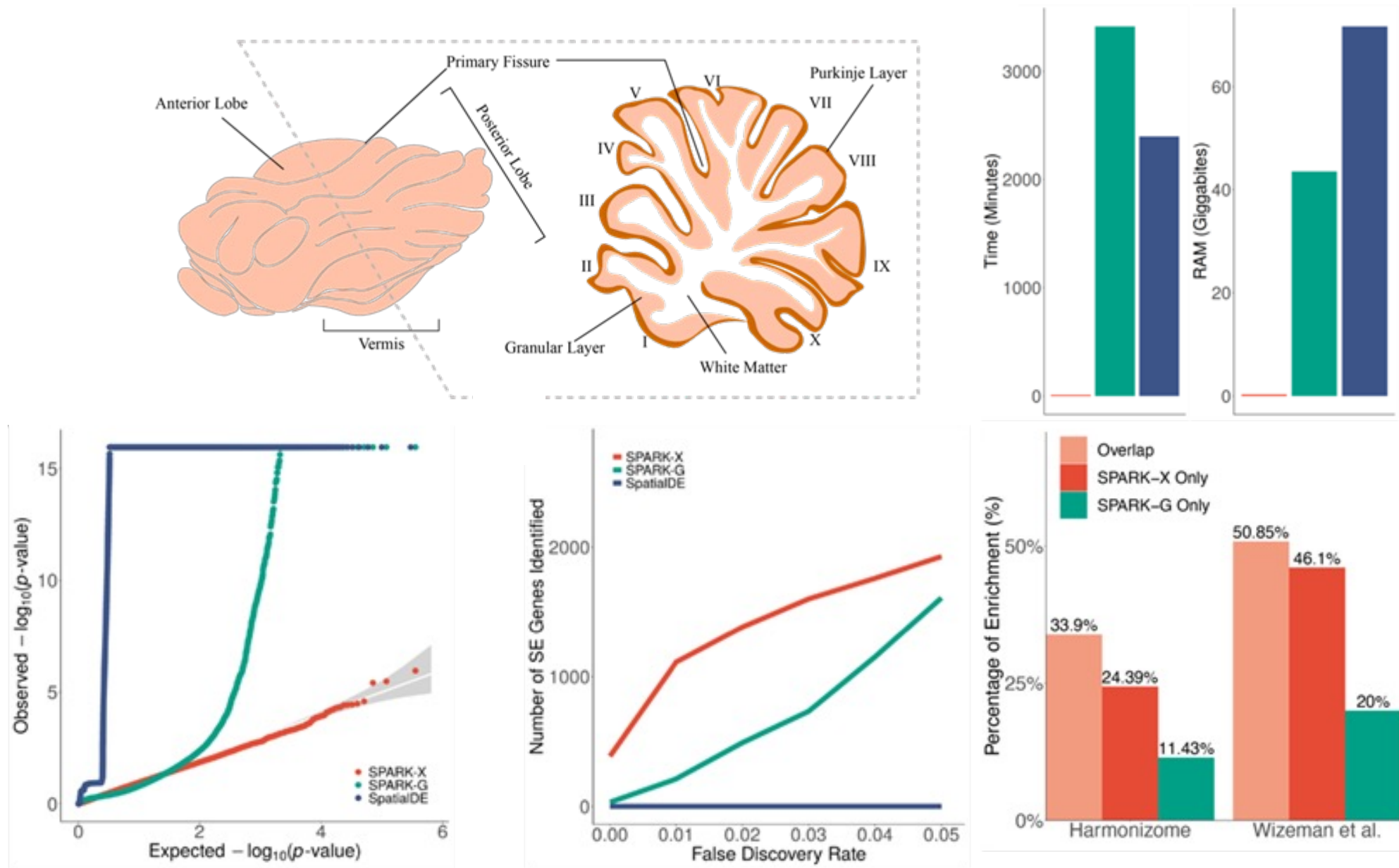
Computational Gains



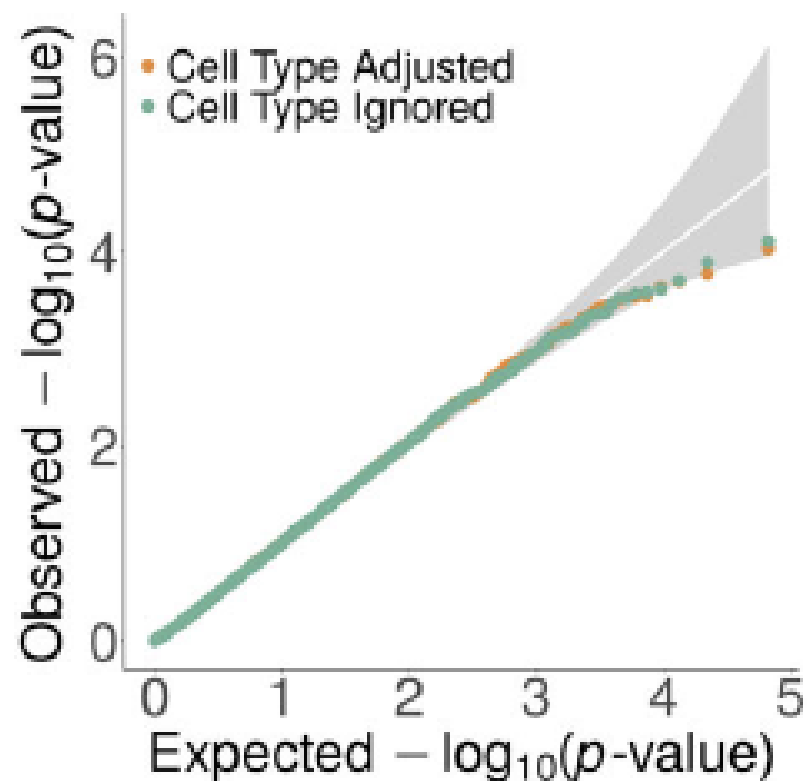
Mouse Cerebellum Data

- Slide-seq technique (Rodriques 2019).
- 11,729 genes measured on 25,551 spatial locations.

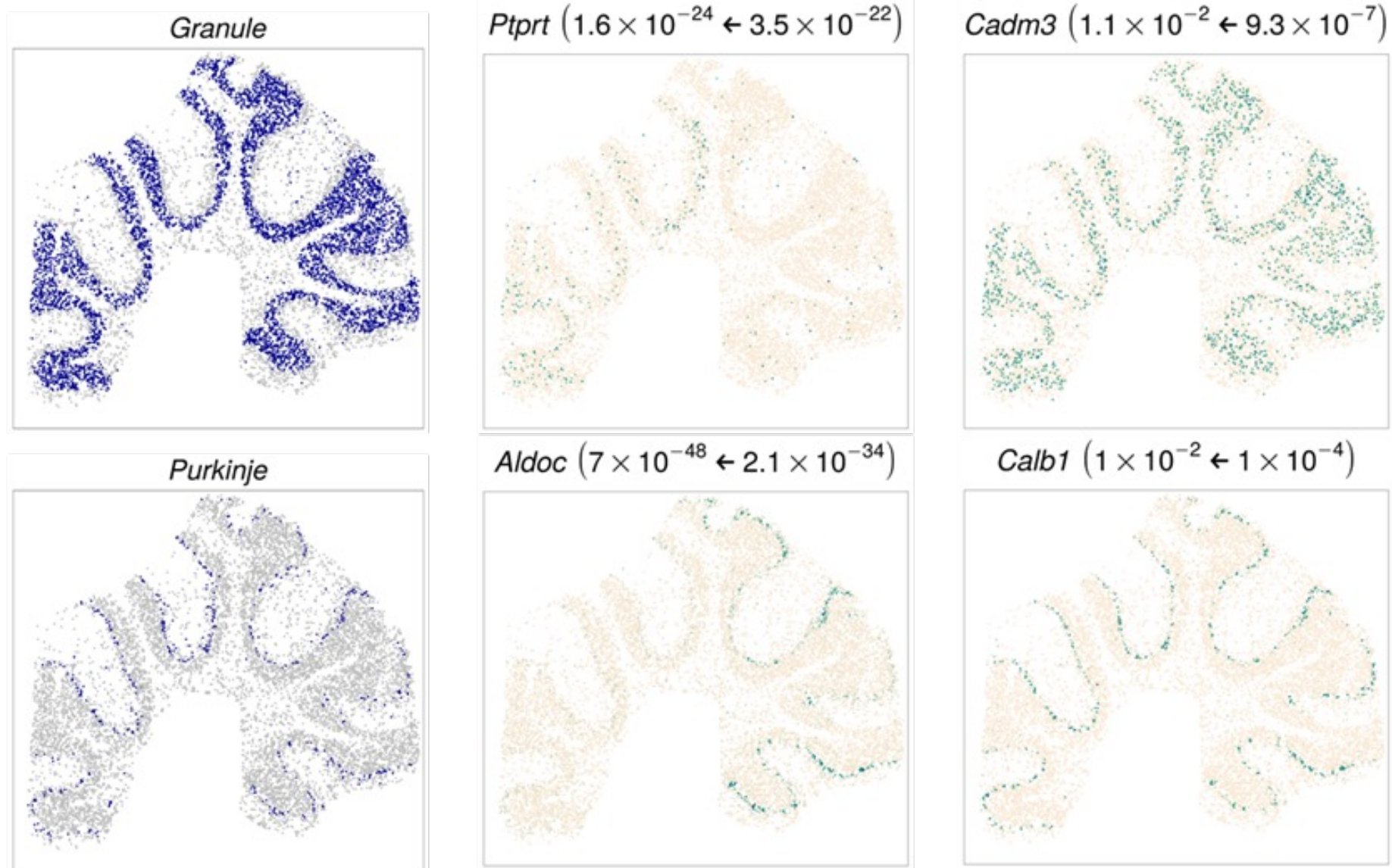
Computation Cost, Type I Error Control and Power



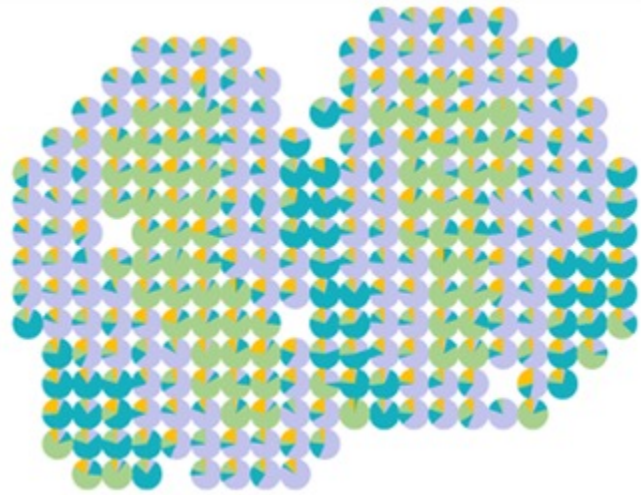
Analysis that Adjusts for Cell Types



Example SVGs after Adjusting for Cell Types



Cell Type Deconvolution



CARD Overview



scRNA-seq

Cell Type Information

Gene Expression Data

Conditional AutoRegressive Model-based Deconvolution

$$X = BV^T + E, E_{gi} \sim N(0, \sigma_e^2)$$
$$V_{ik} = b_k + \phi \sum_{j=1}^n W_{ij}(V_{jk} - b_k) + \epsilon_{ik} \quad + \quad \text{Gaussian Kernel}$$
$$\epsilon_{ik} \sim N(0, \sigma_{ik}^2)$$

Spatial Transcriptomics

Expression Data

Estimated Cell Type Proportion

Spot				
S_1	0.77	0.03	0.07	0.13
S_2	0.07	0.83	0.1	0.07
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
S_{N-1}	0.15	0.15	0.30	0.40
S_N	0.05	0.04	0.06	0.85

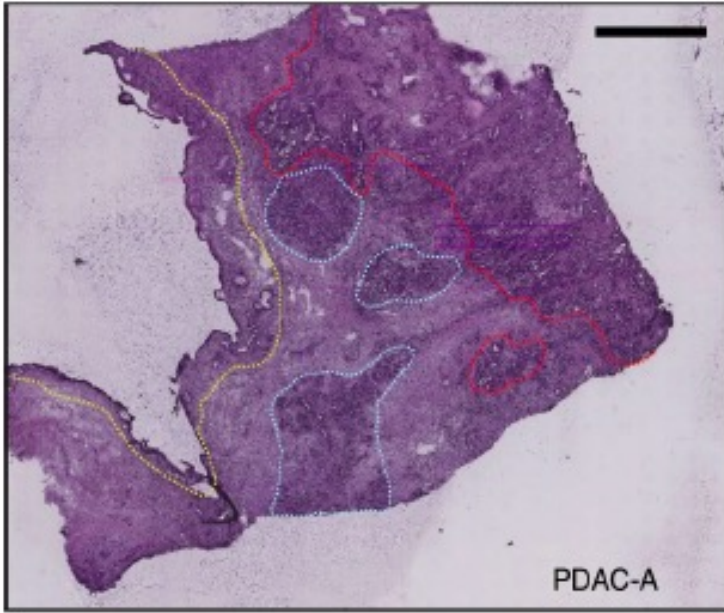
S_1 S_2 S_{N-1} S_N

Refined Spatial Map

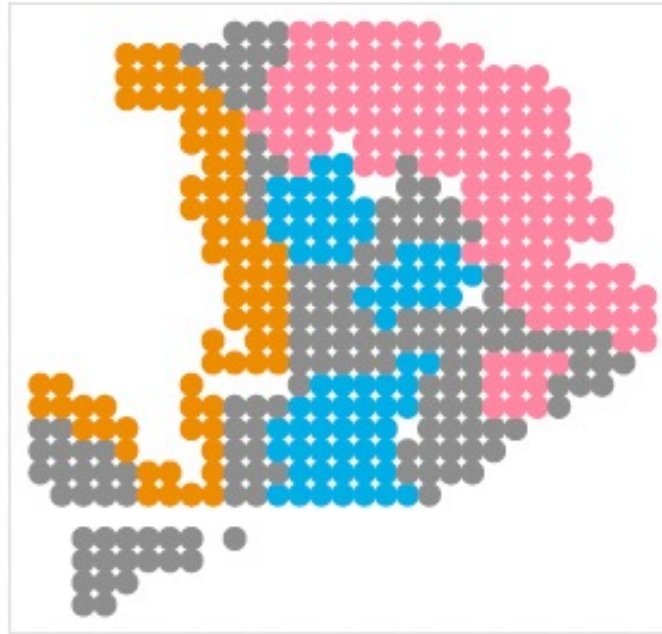
Location Information

Spot	x	y
S_1	x_1	y_1
S_2	x_2	y_2
S_3	x_3	y_3
S_4	x_4	y_4
$\vdots$	$\vdots$	$\vdots$
S_{N-2}	x_{N-2}	y_{N-2}
S_{N-1}	x_{N-1}	y_{N-1}
S_N	x_N	y_N

Human Pancreatic Ductal Adenocarcinomas (PDAC)



Moncada et al., 2021



Spatial Transcriptomics data

By spatial transcriptomics

25k genes, 428 locations

scRNAseq data

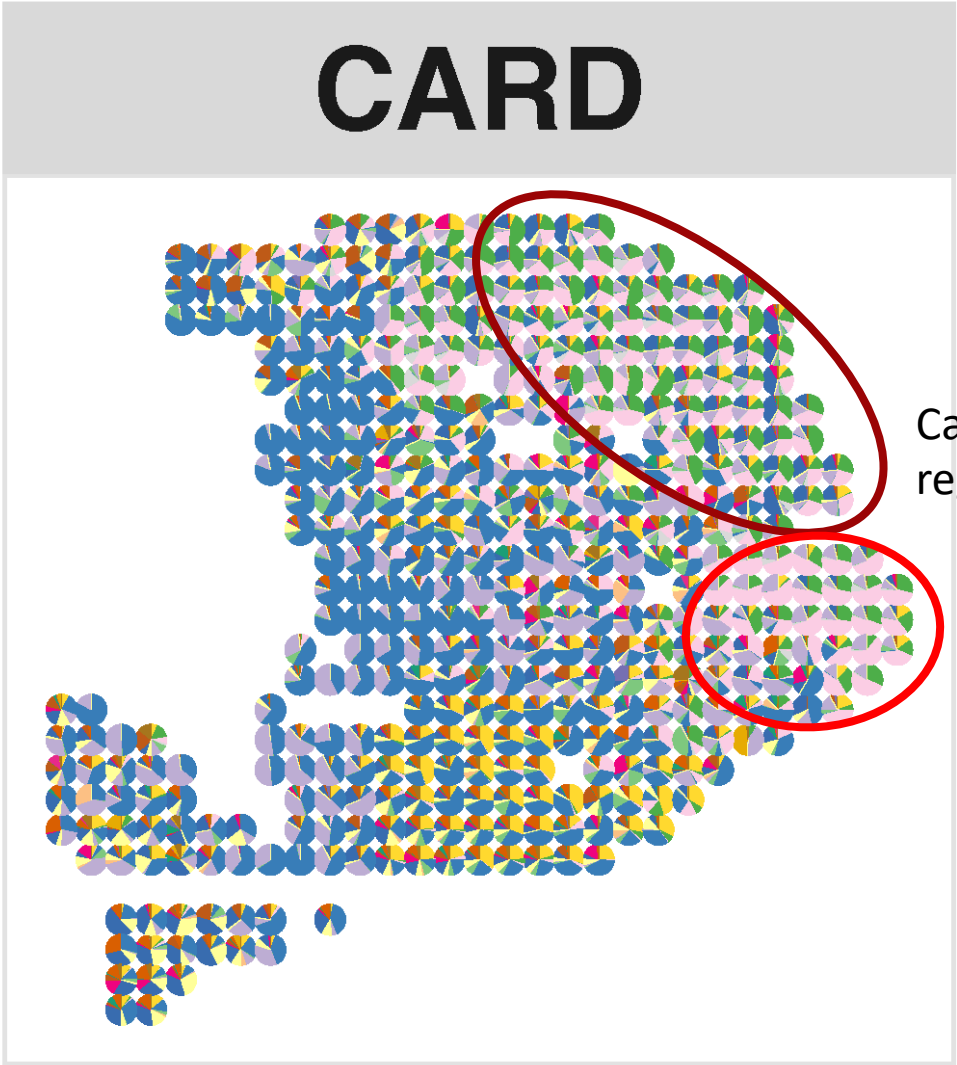
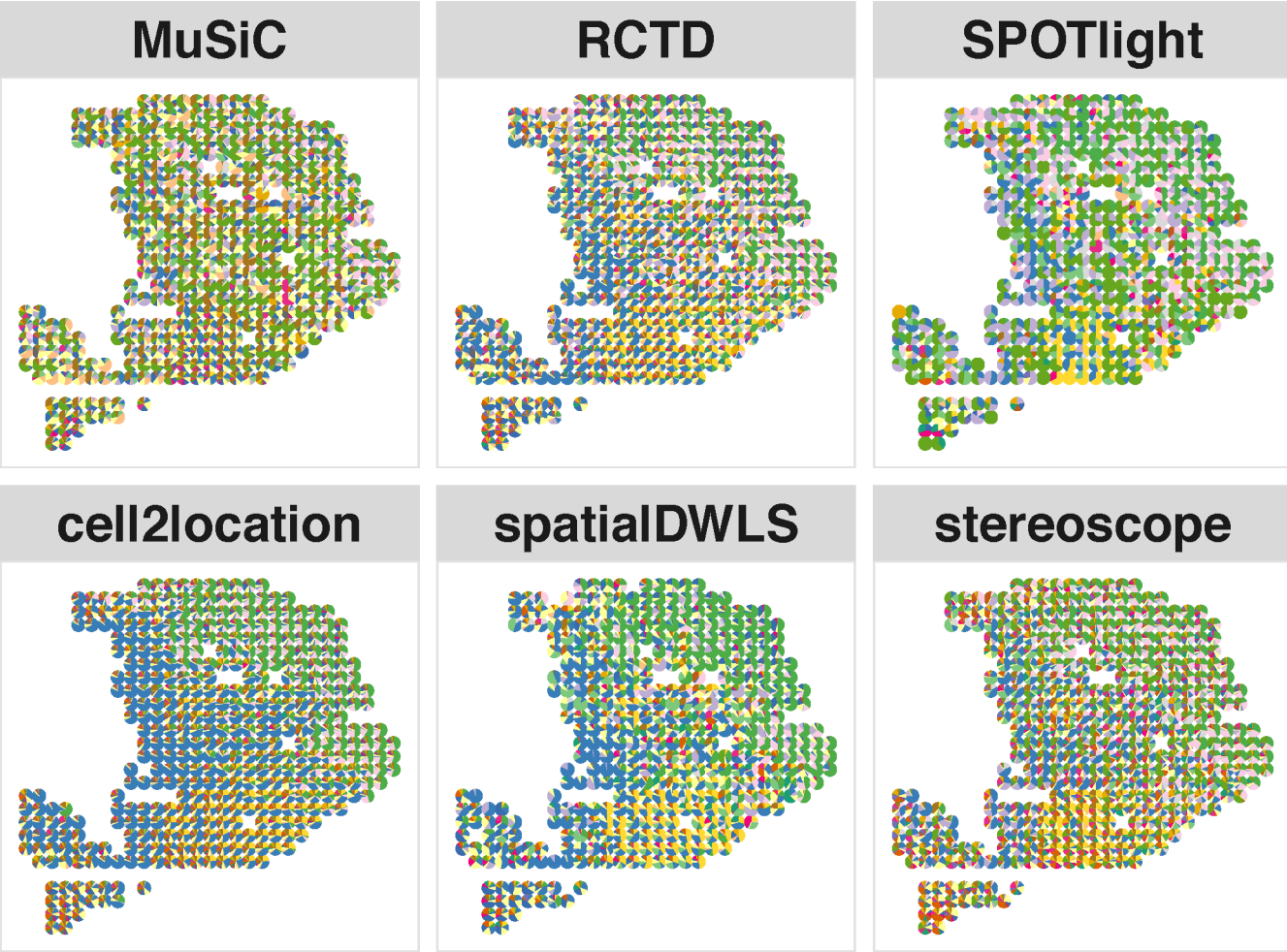
By inDrop

19k genes, 1926 cells, 20 cell types

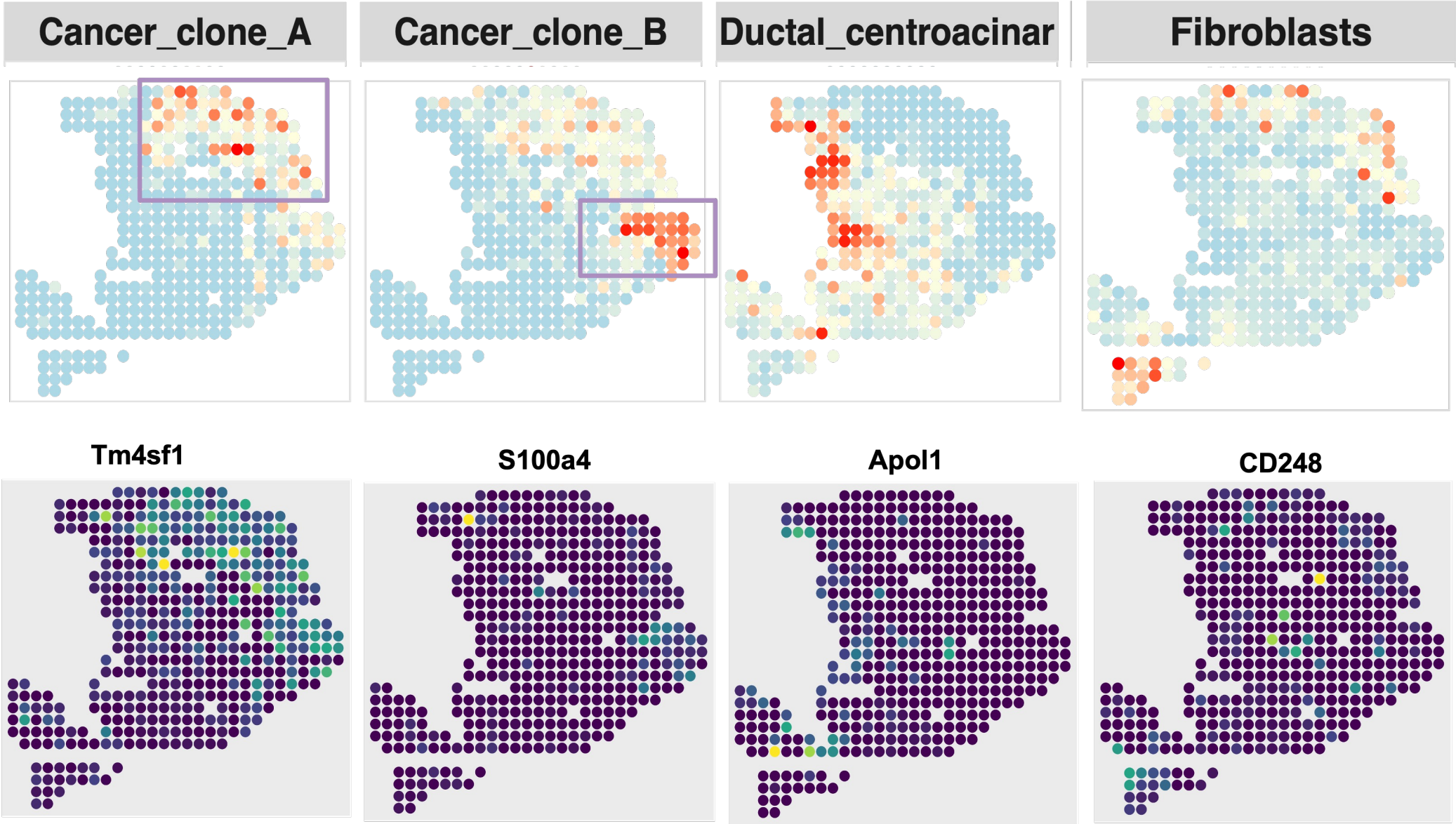
Region • Cancer • Pancreatic • Duct Epithelium • Stroma

- ❖ PDAC is the most prevalent neoplastic disease of the pancreas
- ❖ Complex tumor microenvironment and intra-tumor heterogeneity

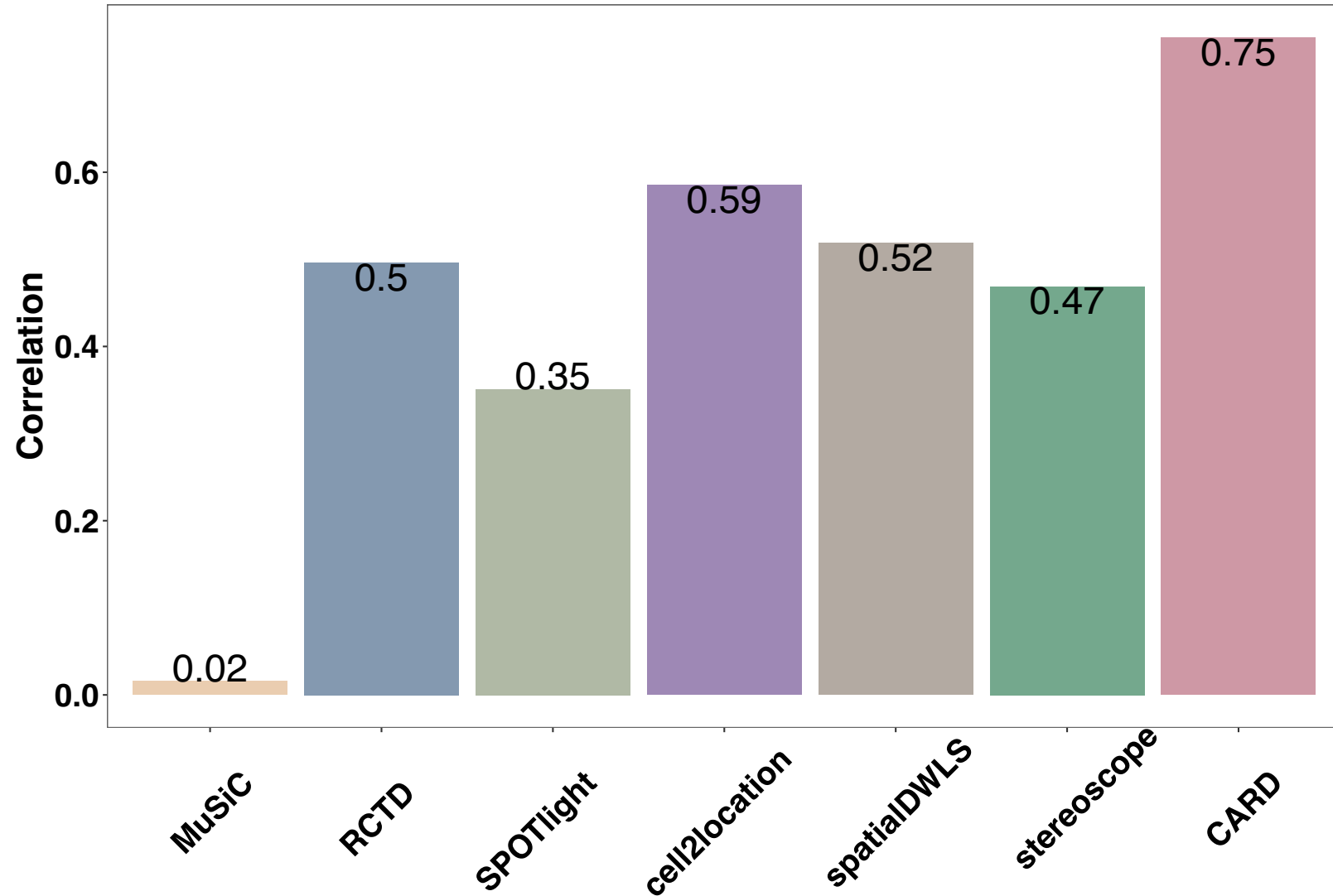
Estimated Cell Type Composition



Spatial Distribution of Cell Type Proportion

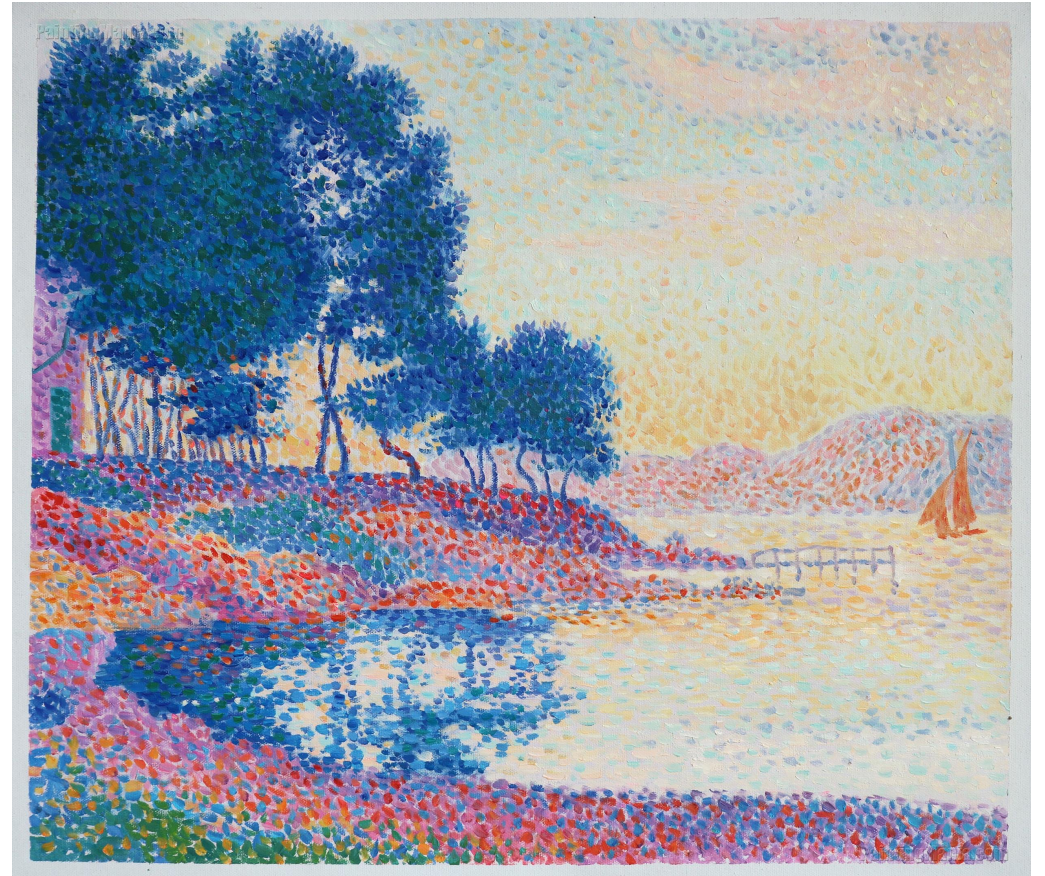
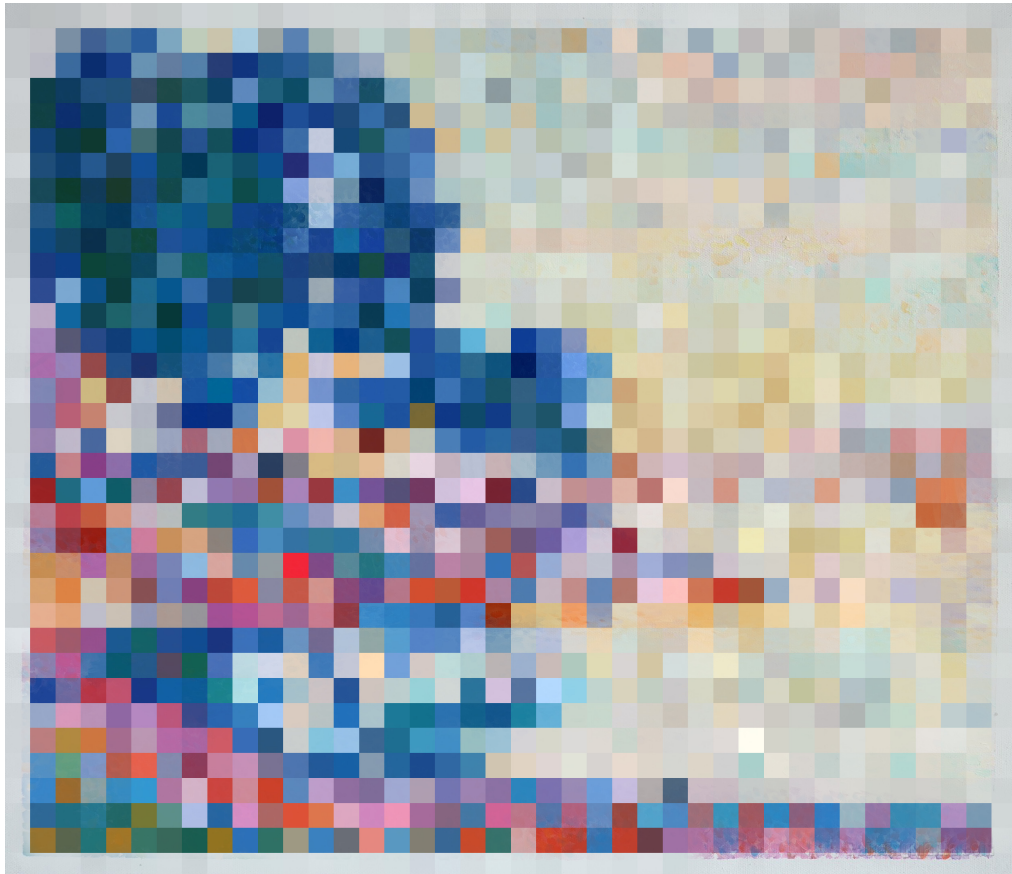


Mean Cell Type Correlation with scRNA-seq Reference

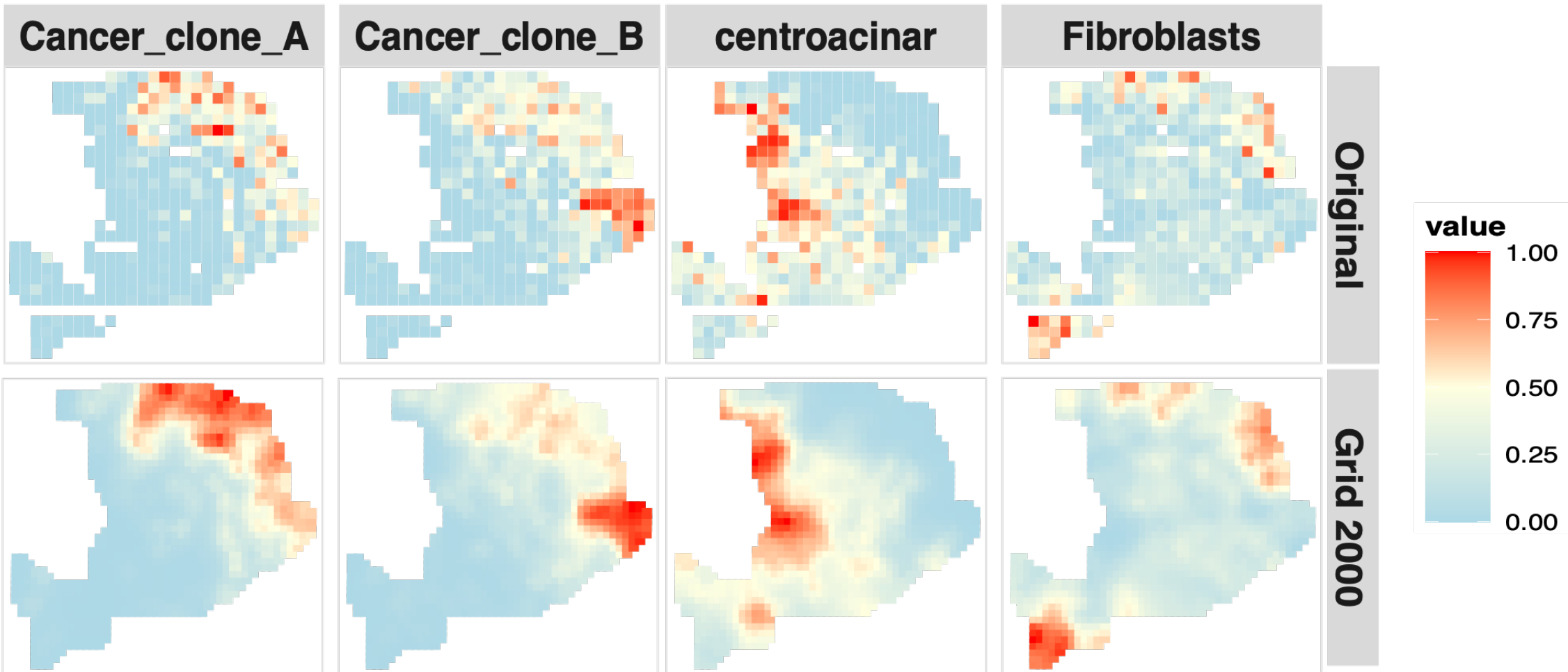


Resolution Refinement

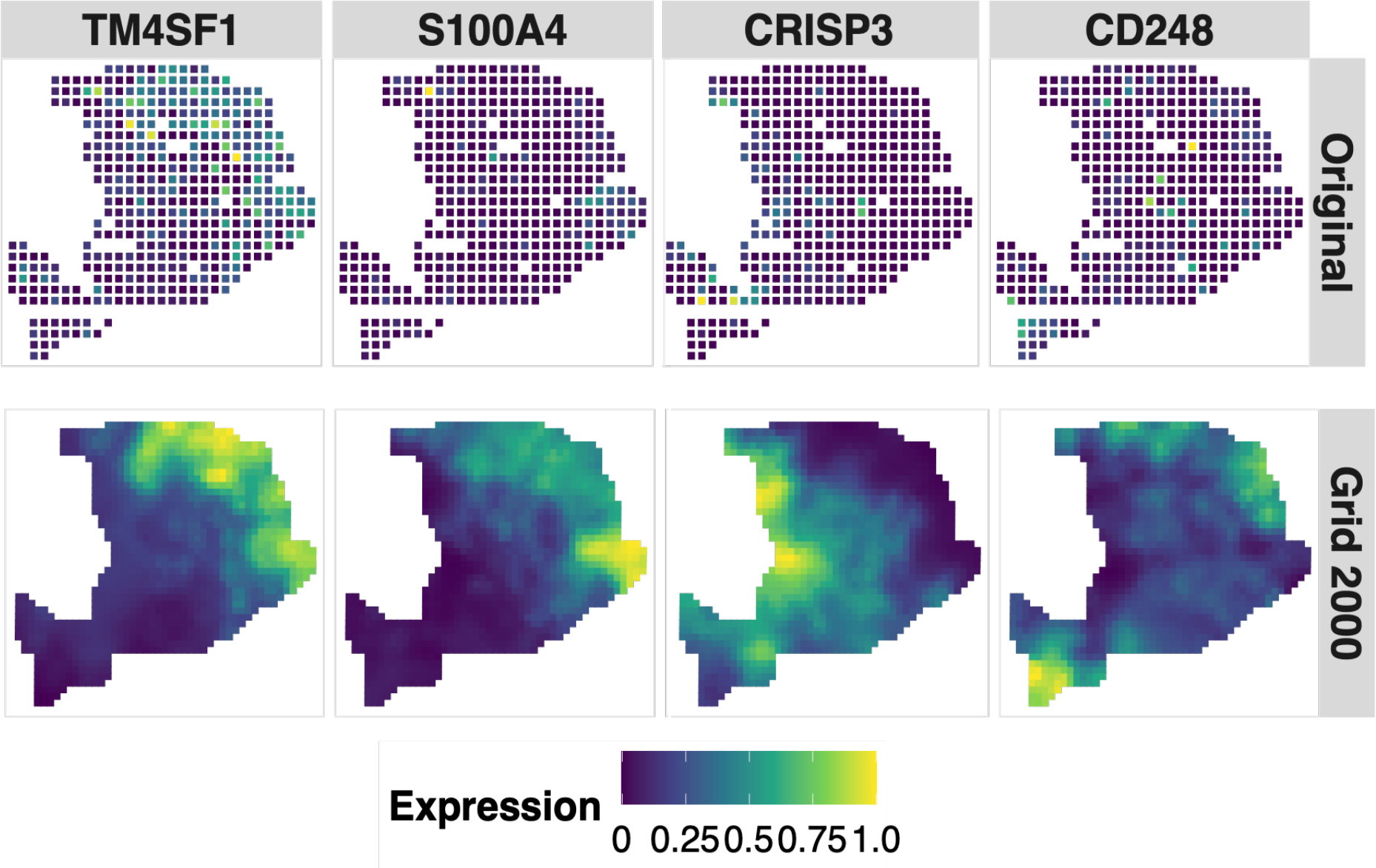
The spatial correlation modeled by CAR on cell type composition enables spatial interpolation and imputation of cell type composition and gene expression on unmeasured locations, thus enhancing spatial resolution.



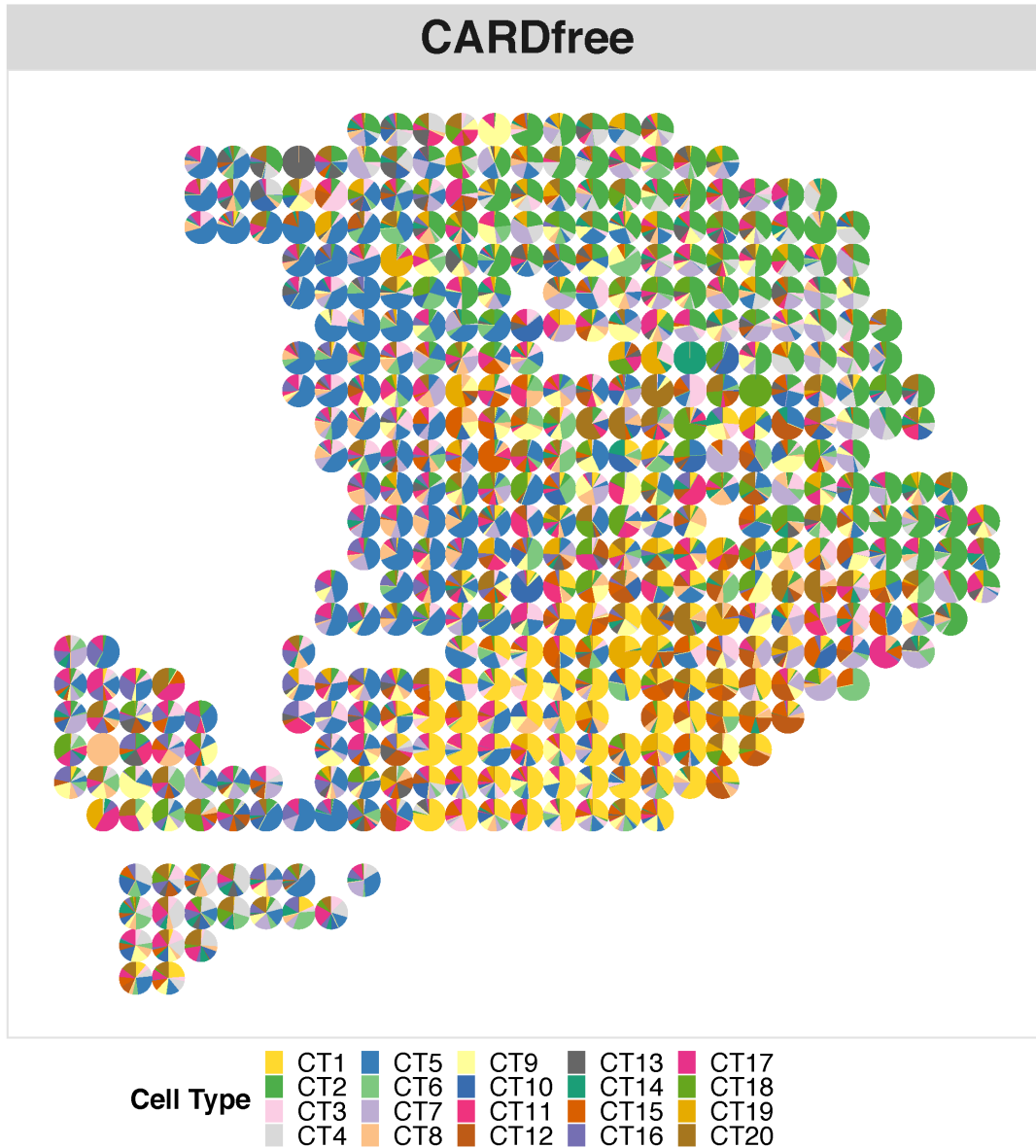
Resolution Refinement: Cell Type Composition



Resolution Refinement: Gene Expression



Reference-free Deconvolution with CARD

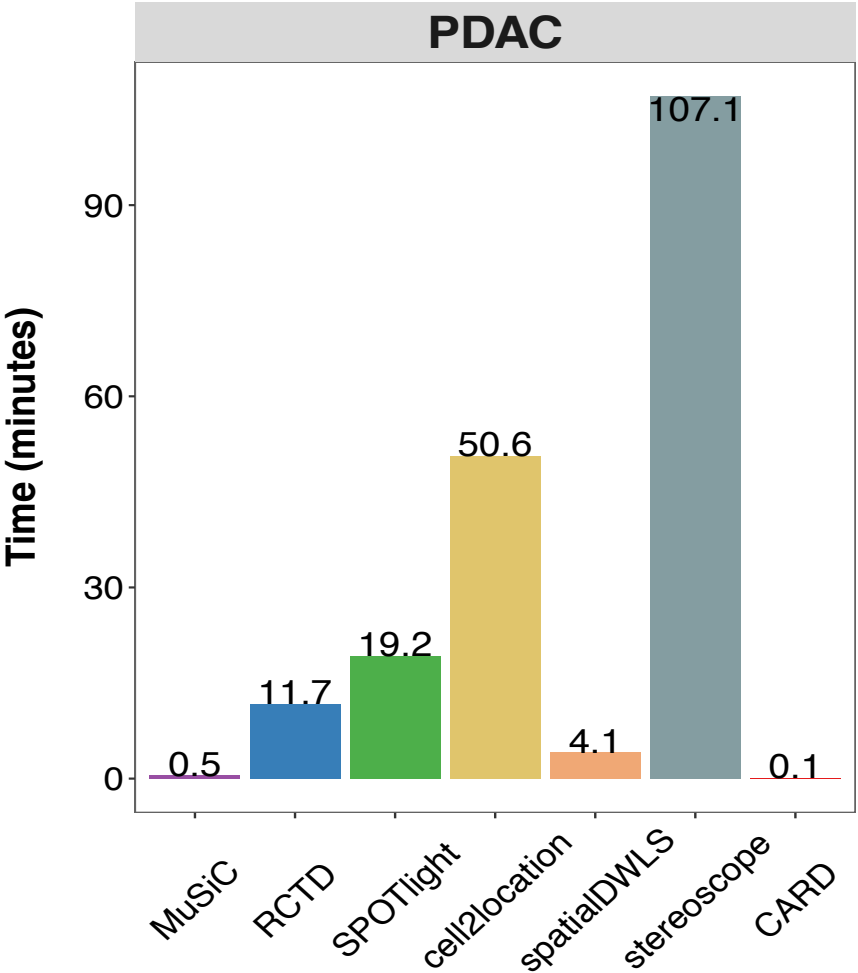


Limitations:

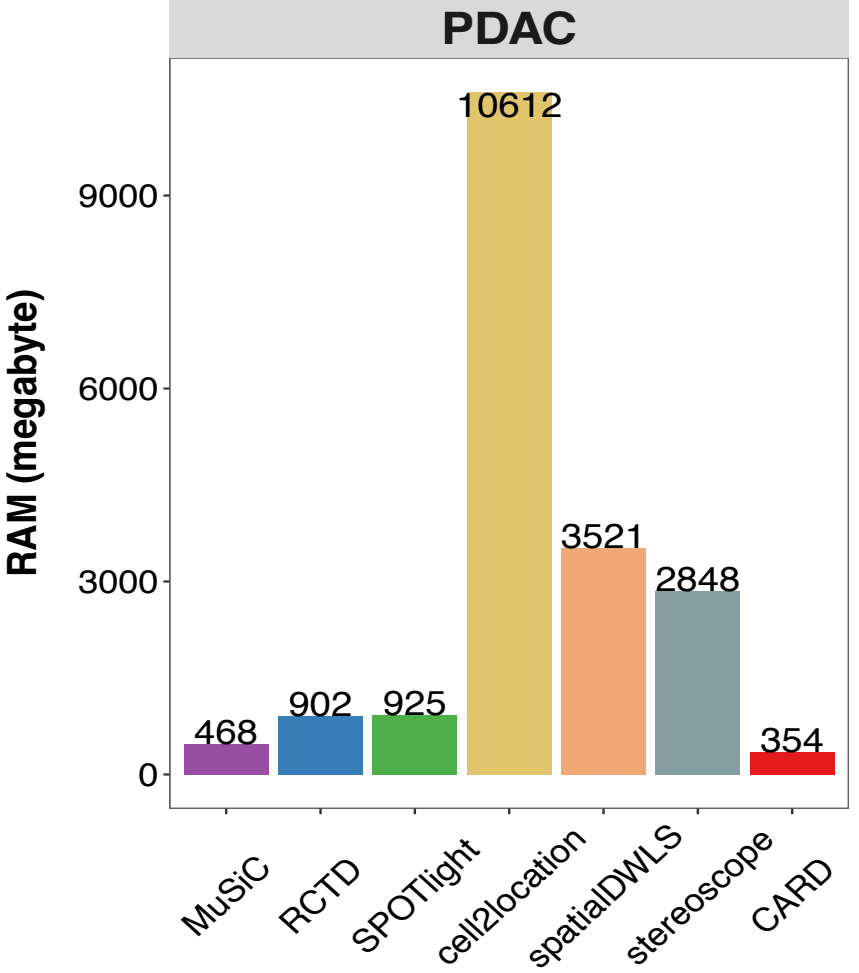
- ❖ Uses less information for deconvolution
- ❖ A post-processing step is usually required to correctly label the inferred cell types.

Computation Efficiency

Time



Memory

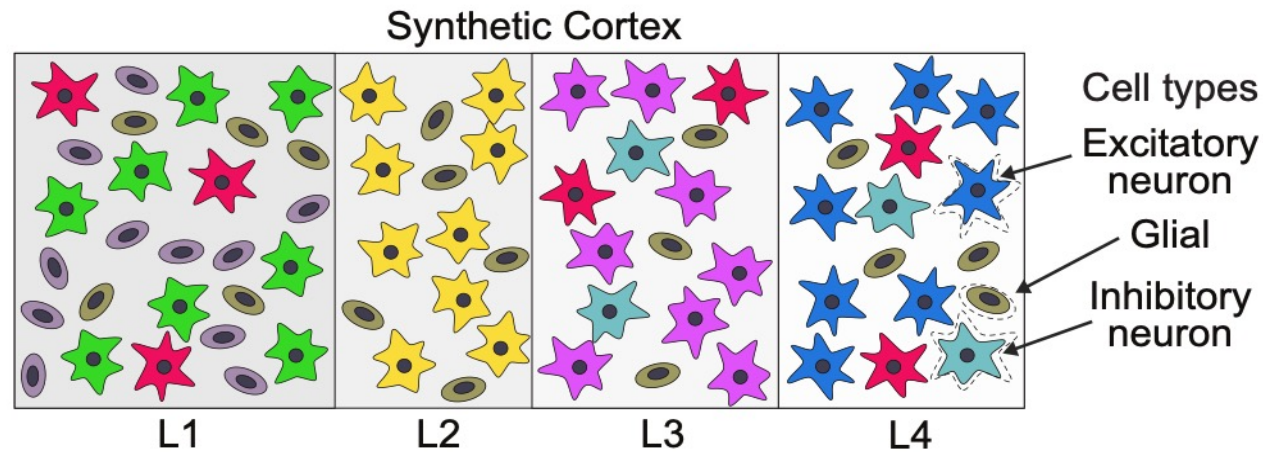


Spatial Domain Detection



Cell Type Composition and Spatial Domains

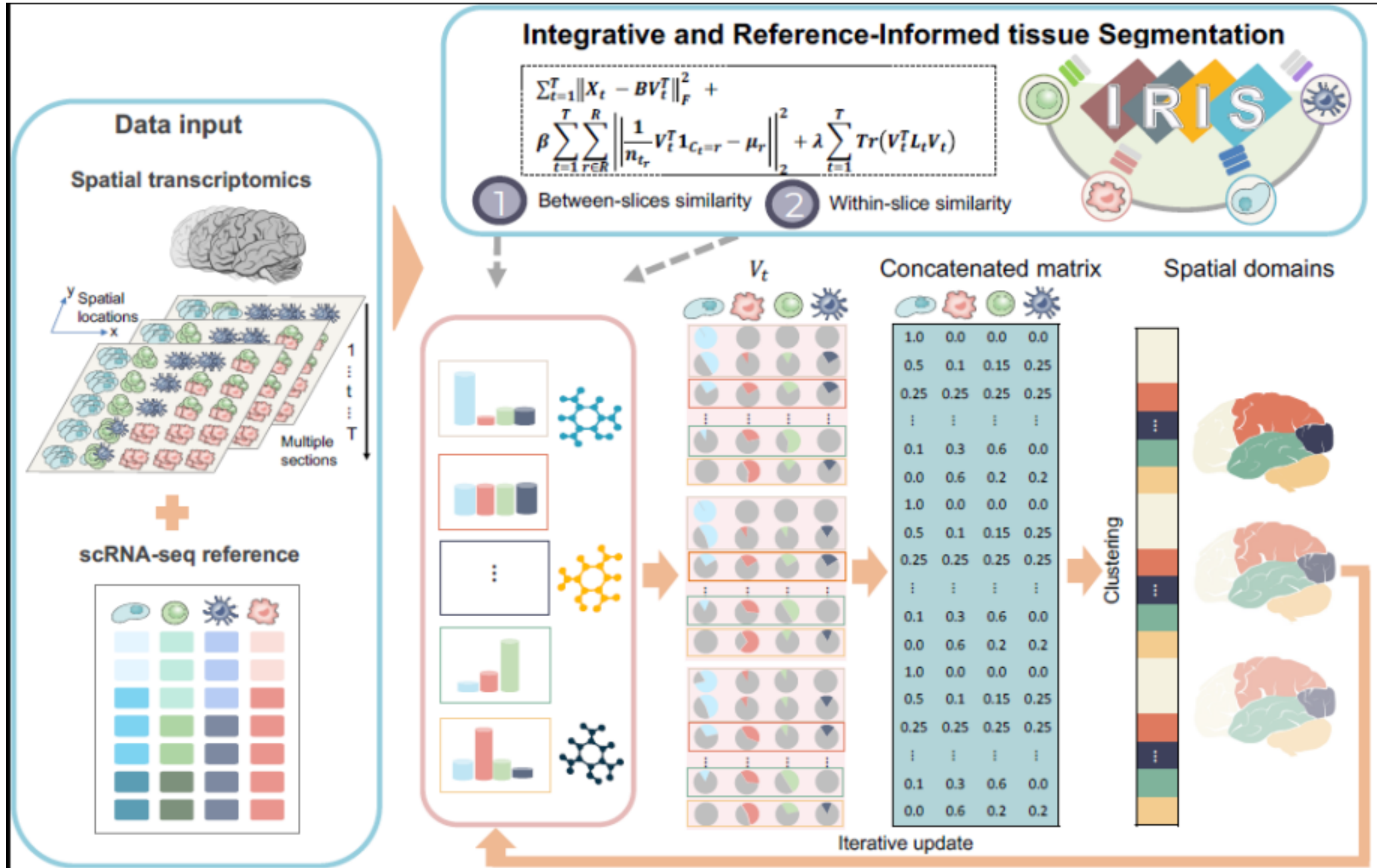
- Spatial domains are tissue regions that are spatially coherent in both gene expression and histology.
- They are the building blocks underlying the spatial and functional organization of complex tissues.
- Each spatial domain is characterized by unique cell type composition.



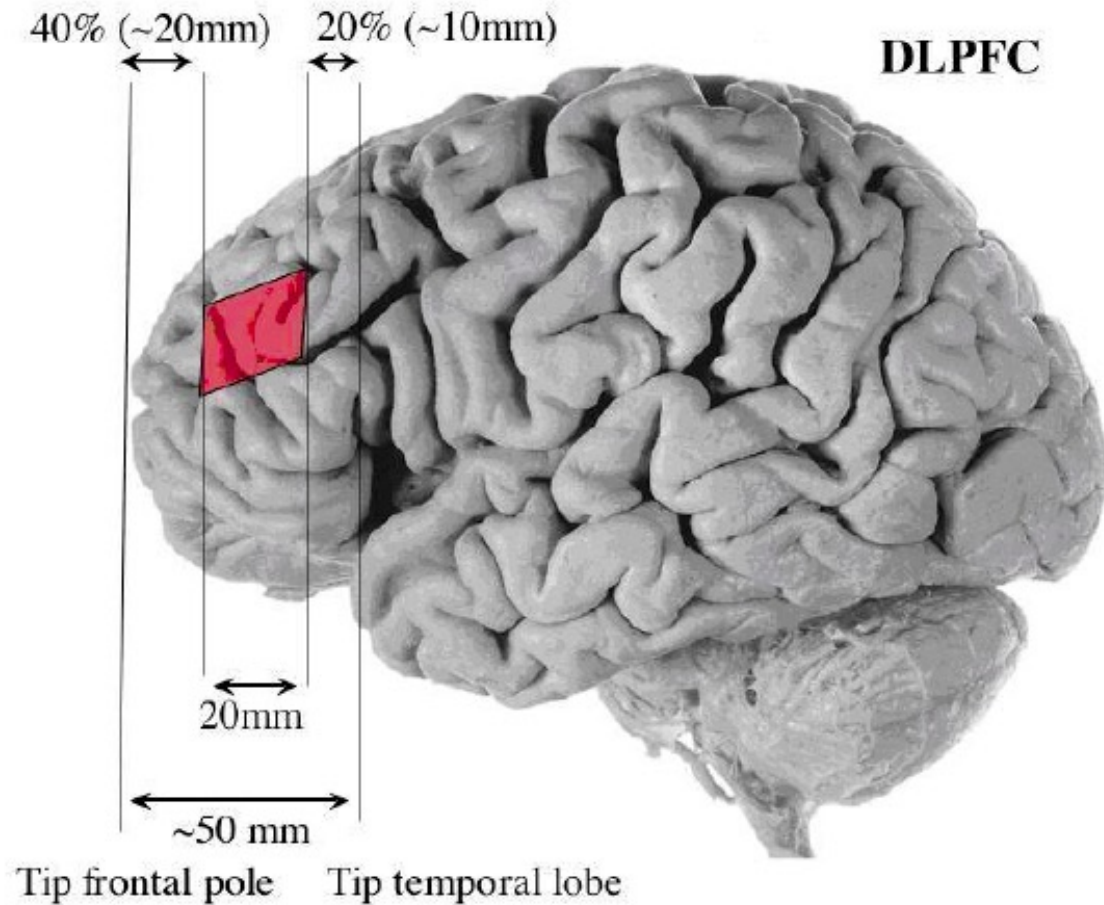
IRIS

- IRIS (Integrative and **R**eference-**I**nformed domain detection for **S**patial transcriptomics): Directly model the cell type compositional heterogeneity to segment the tissue into multiple biologically relevant spatial domains, each characterized by a distinct composition of cell types.
- Key features:
 - Reference-informed: utilize information from scRNA-seq to substantially improve performance
 - Integrative analysis: jointly analyze multiple slices
 - Computationally efficient: applicable to very large-scale ST datasets

IRIS overview



Benchmarking dataset: Human DLPFC 10x Visium data



Al-Hakim et al., Medical Imaging 2006

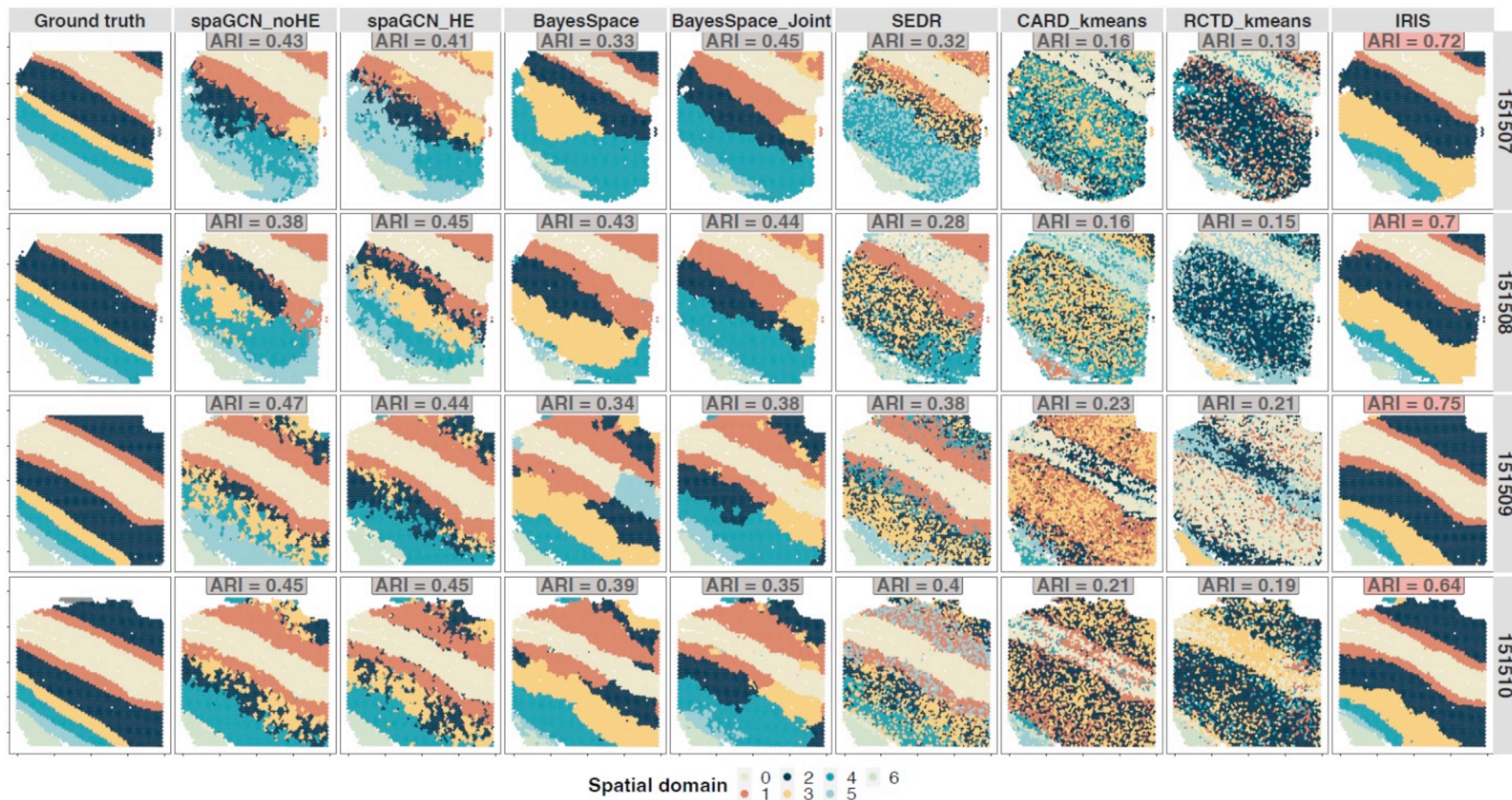
Spatial Transcriptomics data

- ~ Maynard et al, Nature Neuroscience 2021
- ~ Human dorsolateral prefrontal cortex
- ~ Six neuronal layers plus white matter
- ~ 12 samples from 3 individuals
- ~ 33k genes, 3,498 – 4,789 locations

snRNAseq data

- ~ Mathy et al., Nature 2019
- ~ Human post-mortem brain tissue
- ~ 17k genes, 70,634 cells, 44 cell types

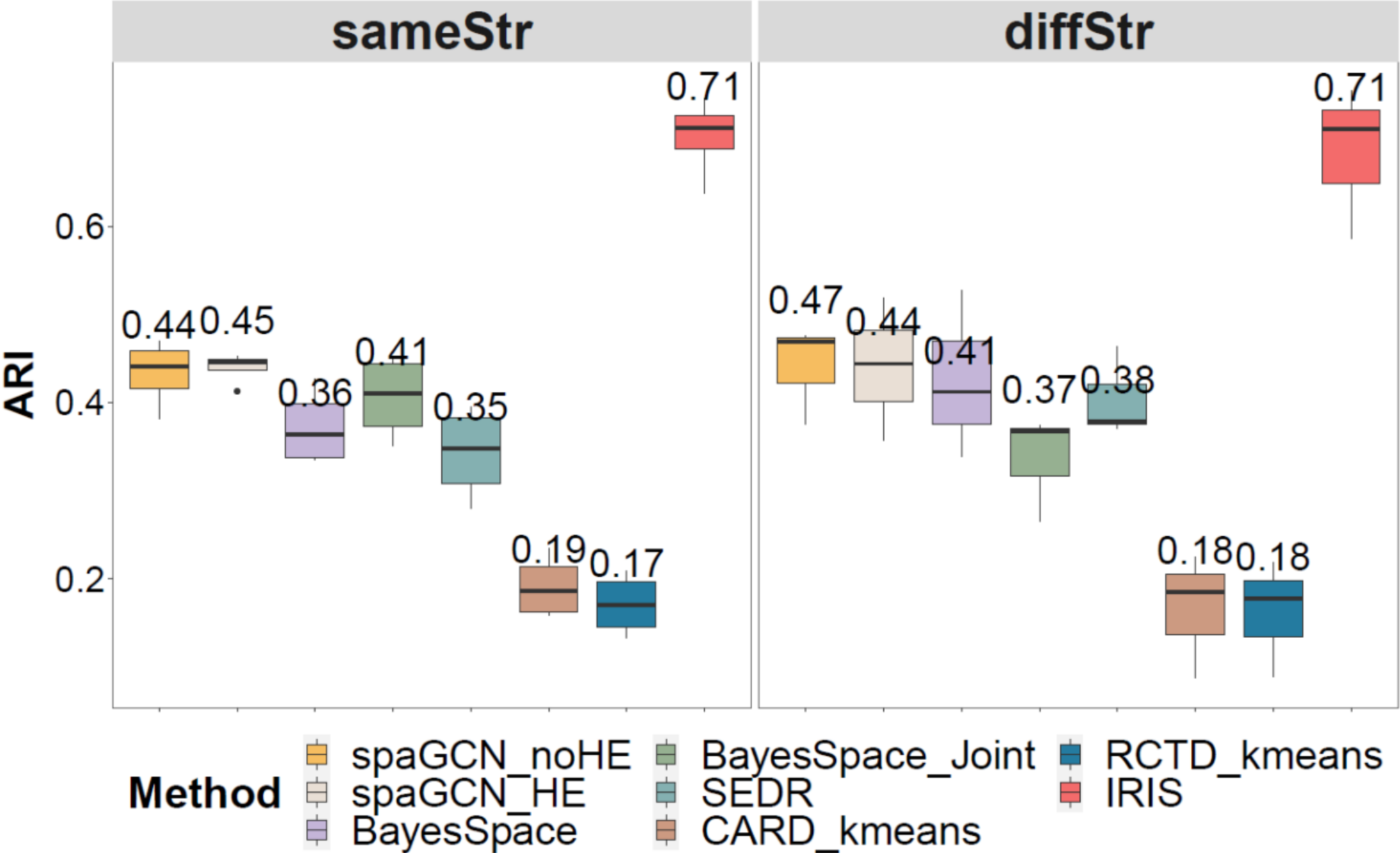
Human DLPFC 10x Visium data: baseline setting



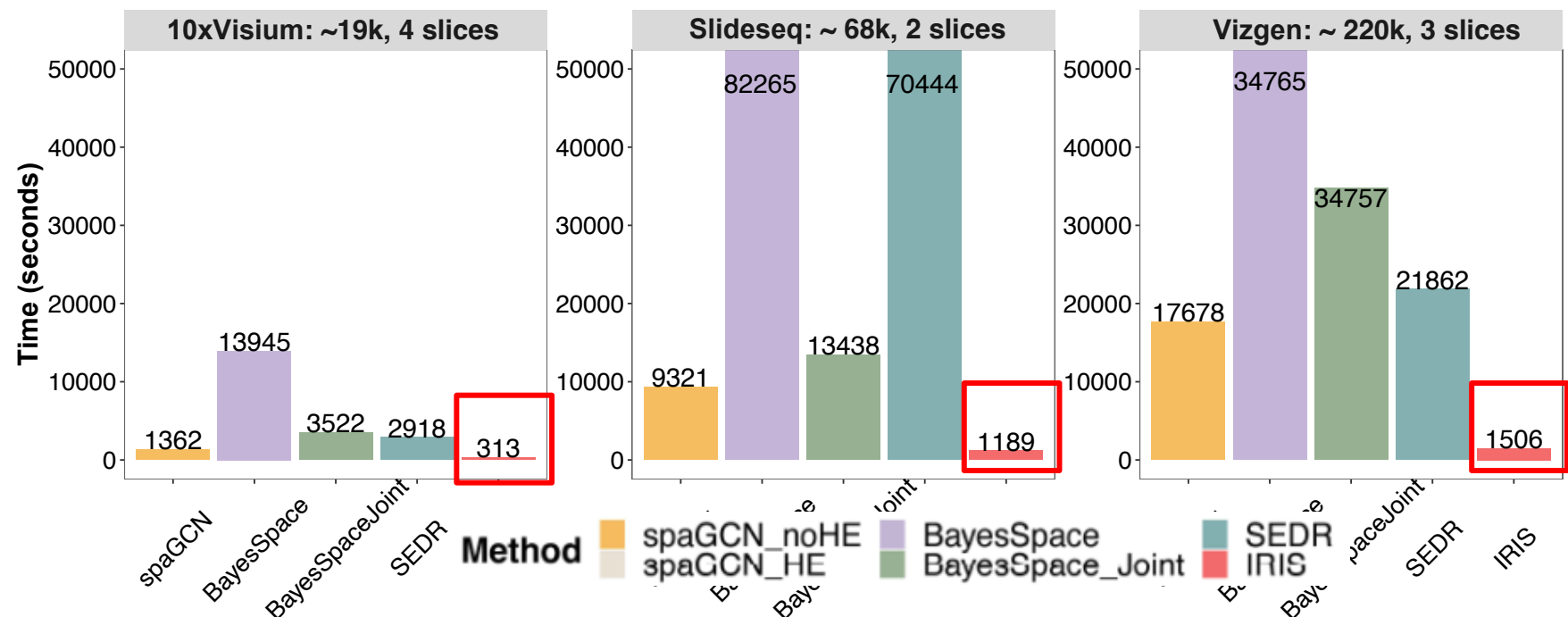
Clustering Accuracy

sameStr: 4 tissue slices from the same donor

diffStr: 3 tissue slices from 3 different donors



Computational Efficiency



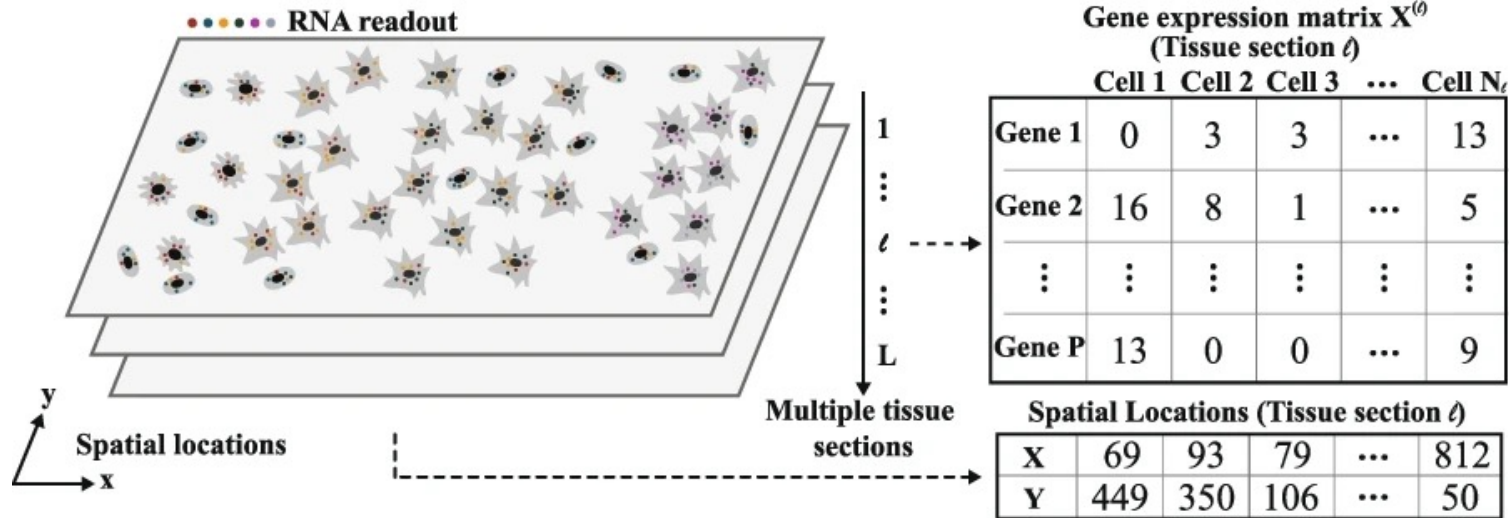
10x Xenium BC (~ 300k spatial locations)		
Metric	IRIS	Other methods
Time	26 min	Not Applicable
Memory	7.6 GB	Not Applicable

Stereo-seq MOB (~ 200k spatial locations)		
Metric	IRIS	Other methods
Time	28 min	Not Applicable
Memory	26.1 GB	Not Applicable

Multi-Scale and Multi-Sample Analysis

BASS: Multi-scale and multi-sample Analysis

Spatial Transcriptomics Data



BASS: Bayesian Analytics for Spatial Segmentation

Pre-processing:

- Library size normalization
- Log-transformation
- Dimension reduction
- Batch adjustment

$$X_i^{(l)} | c_i^{(l)} = c \sim MVN(\mu_c, \Sigma)$$

$$c_i^{(l)} | z_i^{(l)} = r \sim Cat(\pi_r)$$

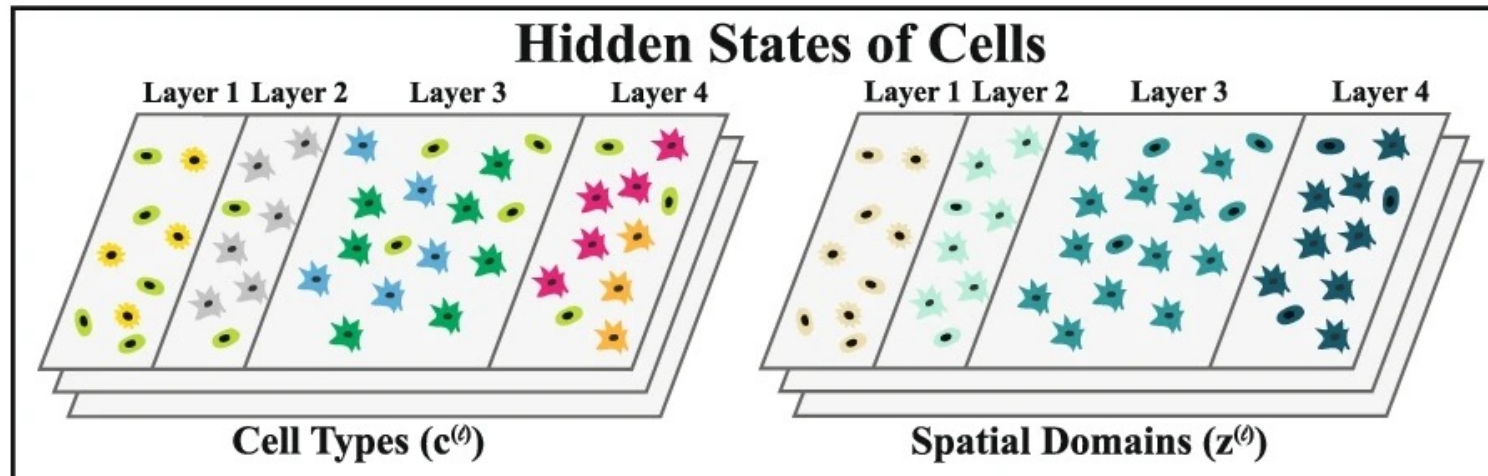
$$z^{(l)} | V^{(l)} \sim Potts(V^{(l)}, \beta)$$

Output

1. Cell type cluster labels (c)
2. Spatial domain labels (z)
3. Cell type proportions (π)

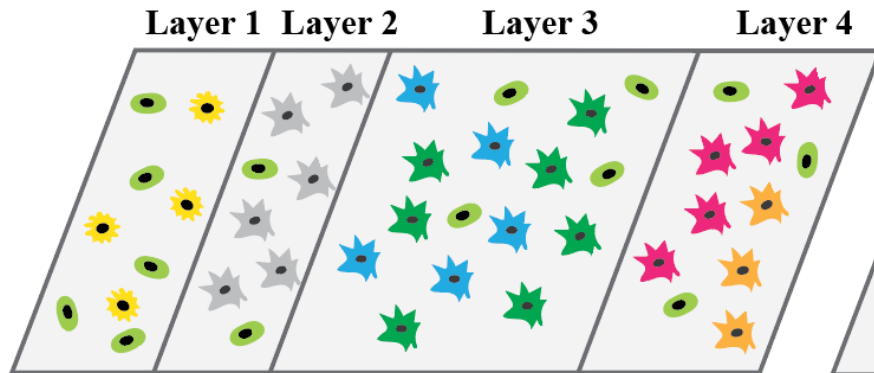
Layer 1 Layer 2 Layer 3 Layer 4

	Layer 1	Layer 2	Layer 3	Layer 4
	0.60	0.25	0.25	0.27
	0.40	0	0	0
	0	0.75	0	0
	0	0	0.44	0
	0	0	0.31	0
	0	0	0	0.45
	0	0	0	0.28

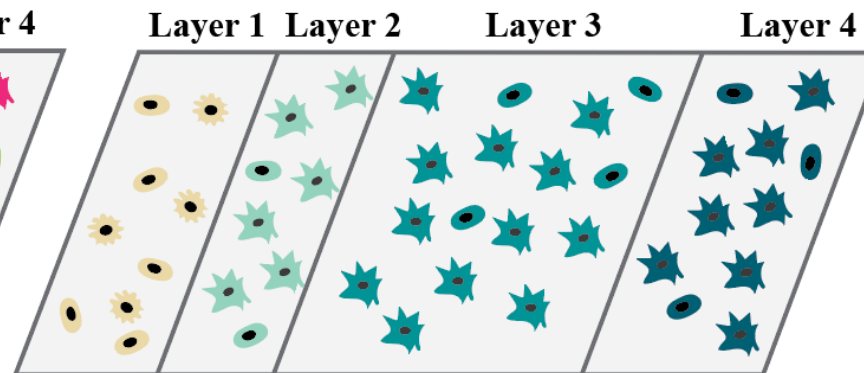


BASS output

1. Cell type cluster labels



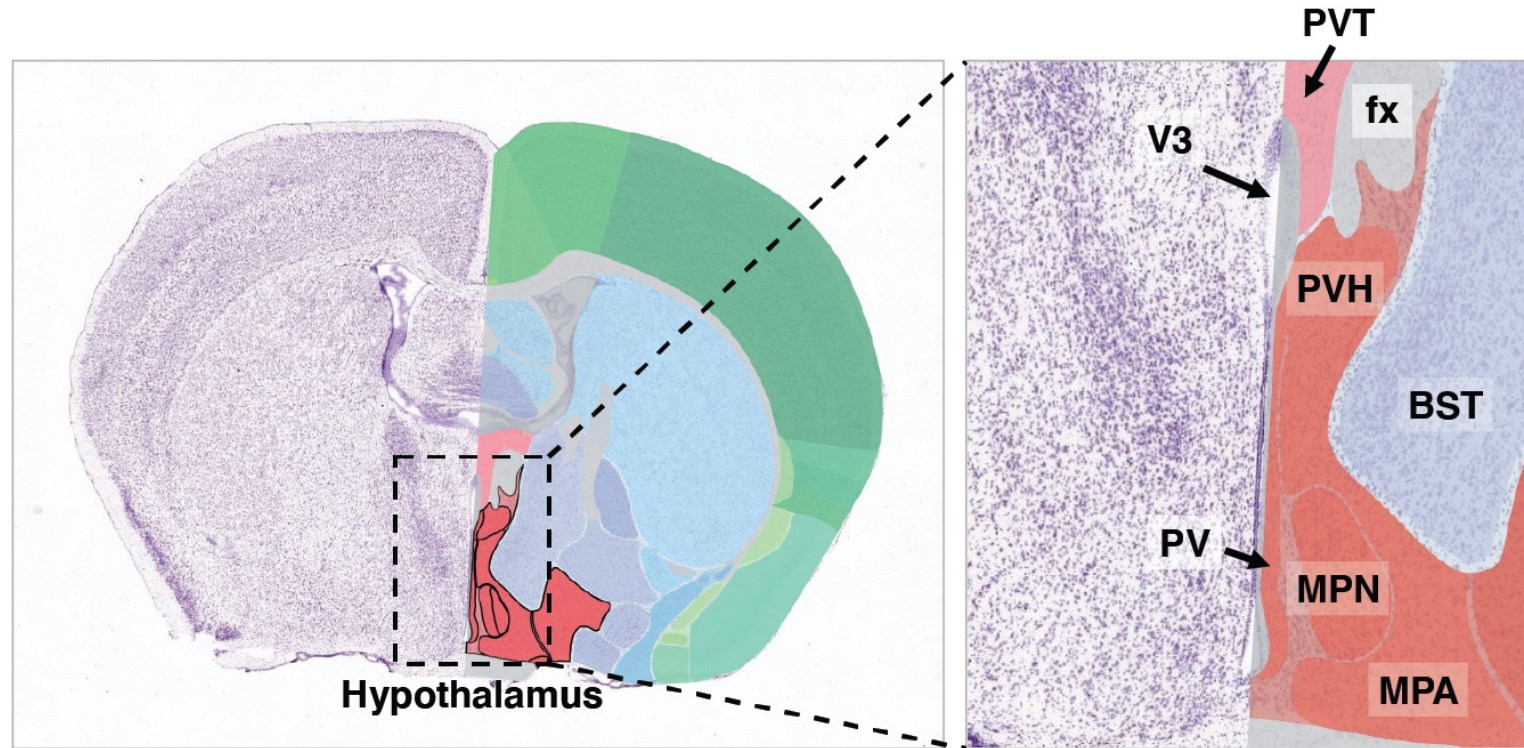
2. Spatial domain labels



3. Cell type compositions

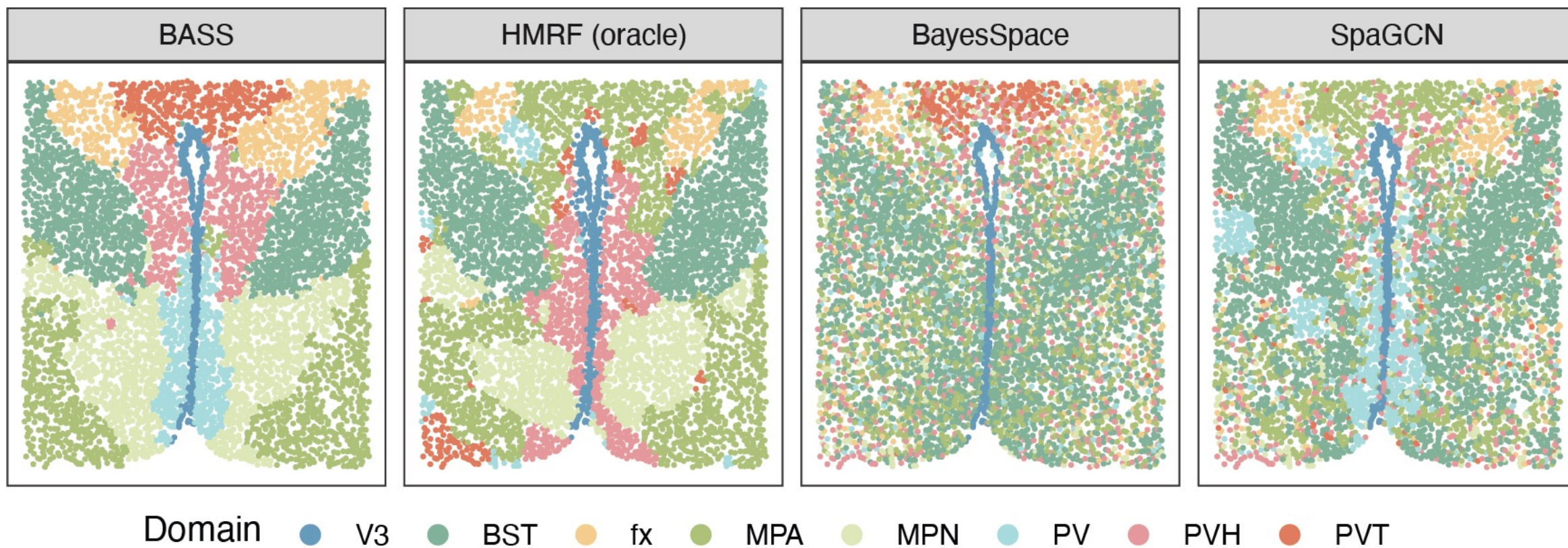
	Layer 1	Layer 2	Layer 3	Layer 4
	0.60	0.25	0.25	0.27
	0.40	0	0	0
	0	0.75	0	0
	0	0	0.44	0
	0	0	0.31	0
	0	0	0	0.45
	0	0	0	0.28

Mouse hypothalamus data by MERFISH



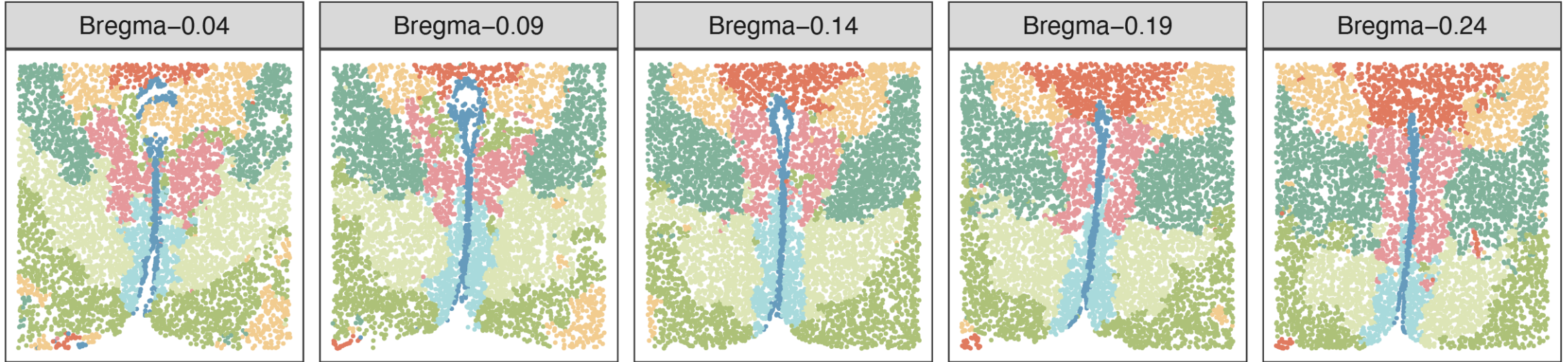
- Five tissue sections
- ~ 6,000 single-cells for each tissue section
- 155 genes (cell type markers genes)
- Eight symmetric spatial domains

Spatial domain detection

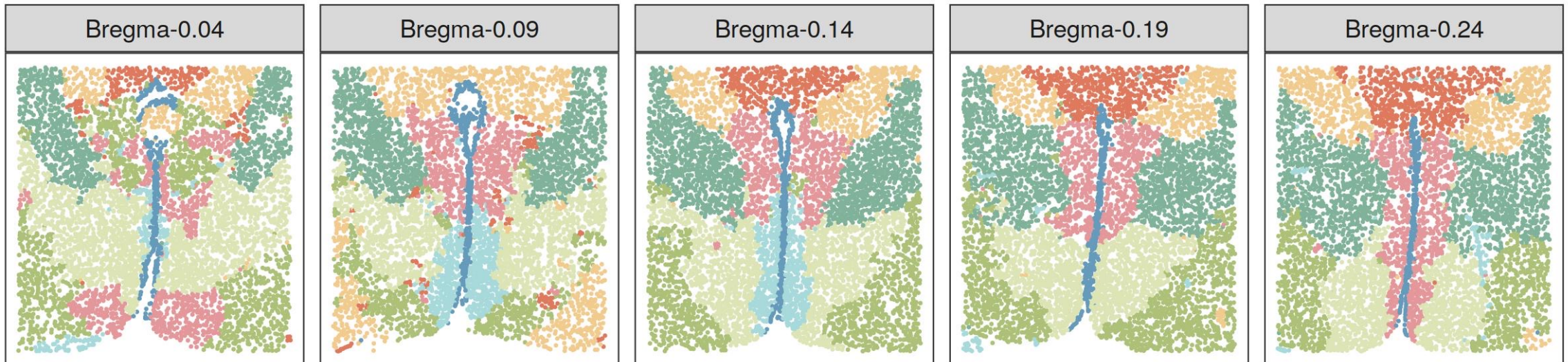


Spatial domain detection

Multi-sample analysis

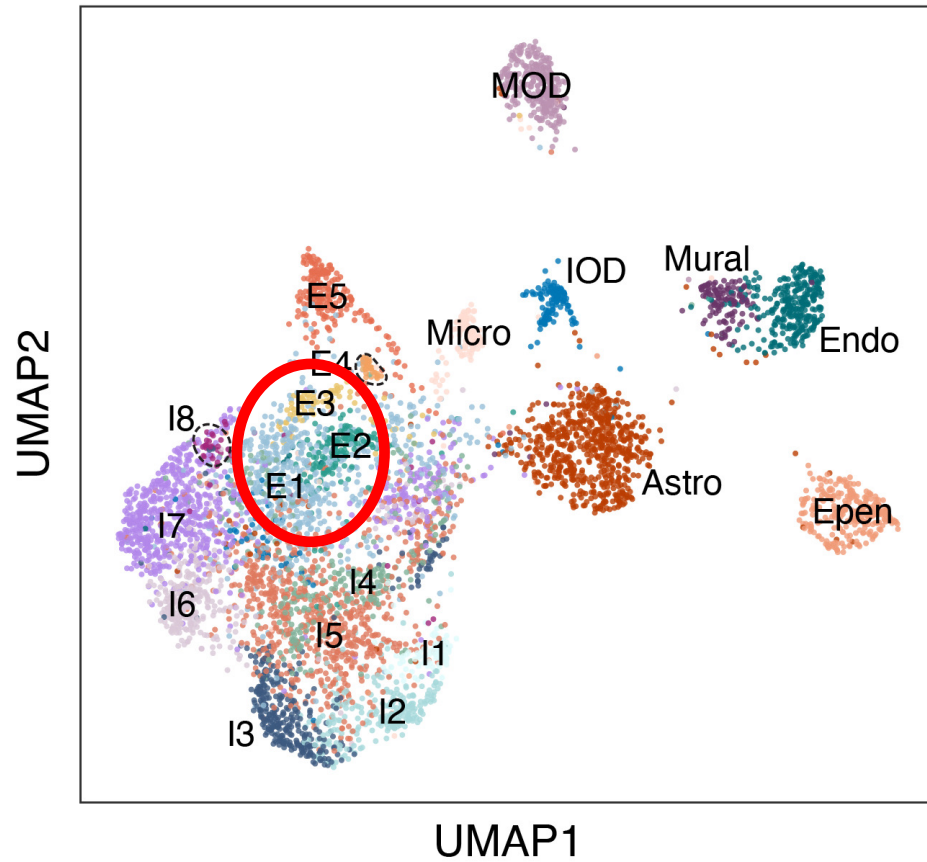


Single-sample analysis

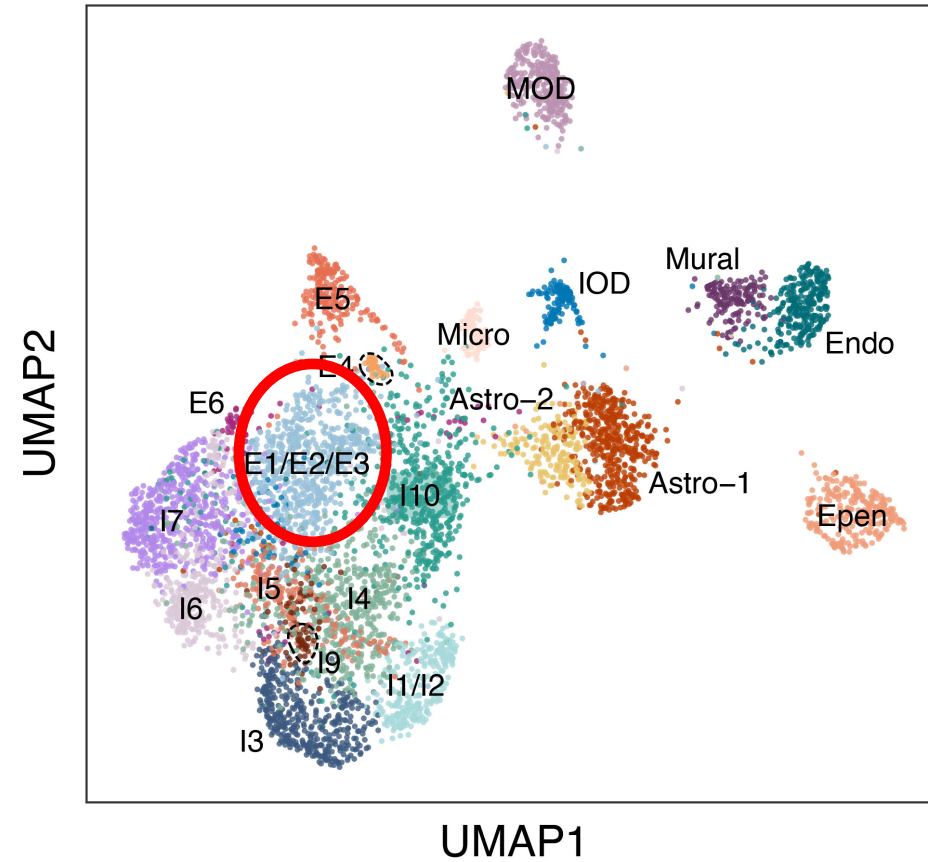


Cell type clustering

BASS (Single:Bregma-0.14)

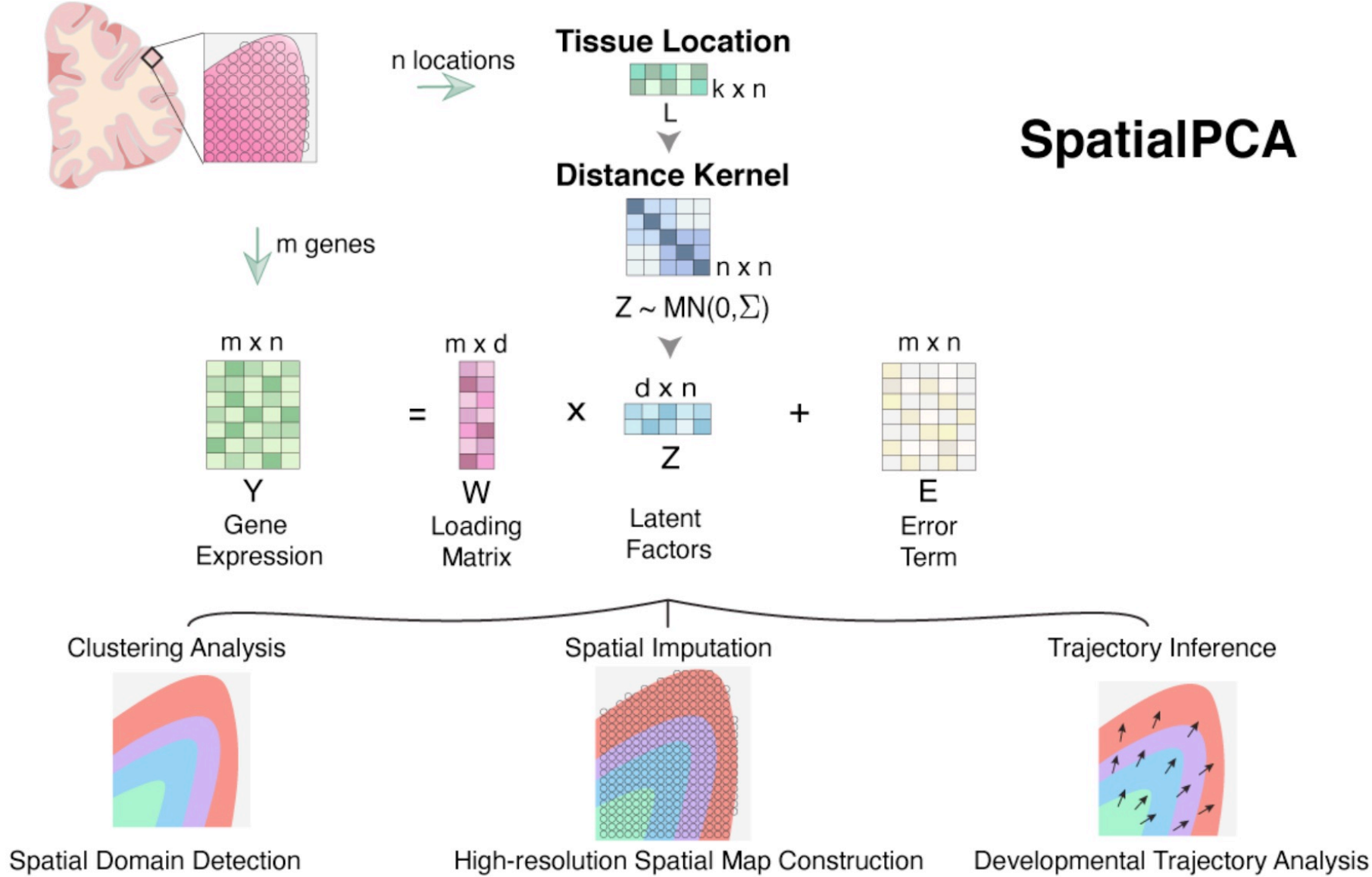


Seurat (Single:Bregma-0.14)

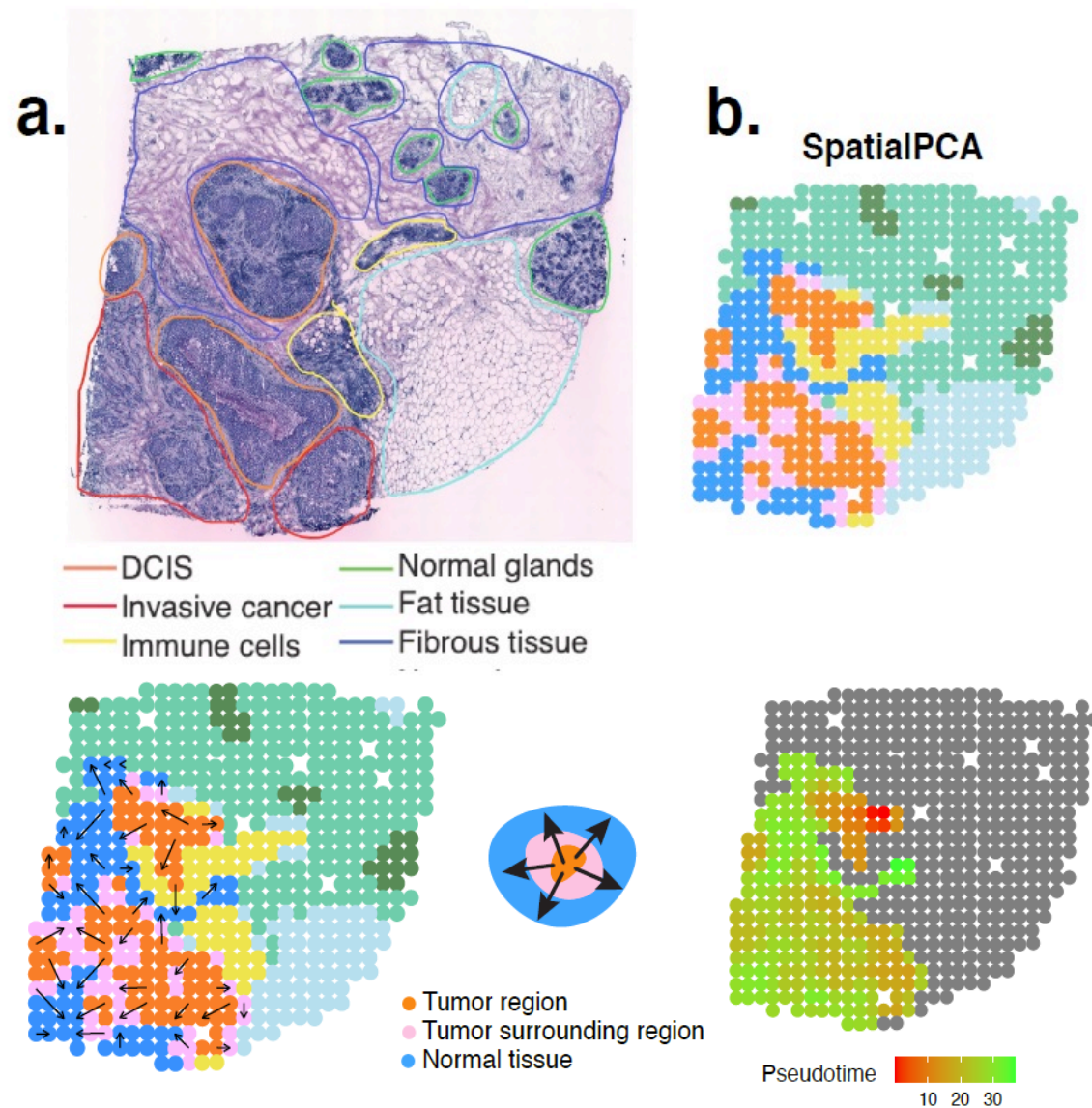


Dimension Reduction

SpatialPCA for spatially aware dimension reduction



Spatial domain detection and trajectory inference

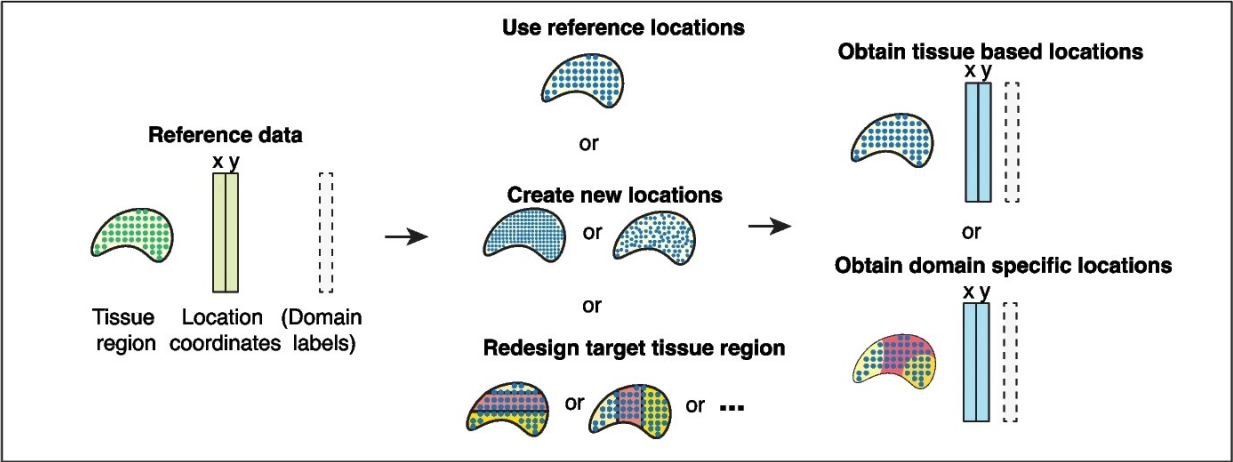


Simulator for Spatial Transcriptomics

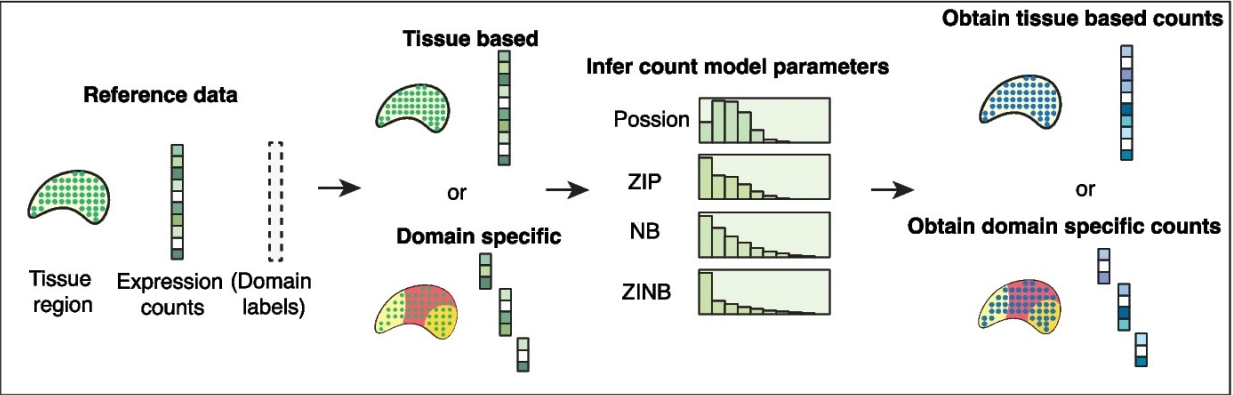
SRTsim: Simulator for spatial transcriptomics

Reference based

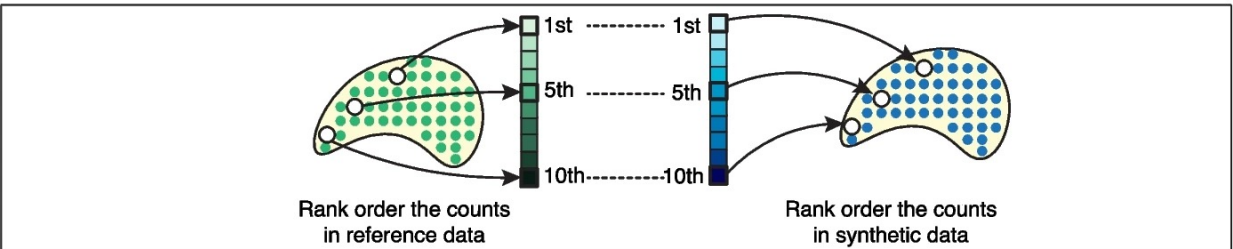
Step 1. Obtain or create location coordinates.



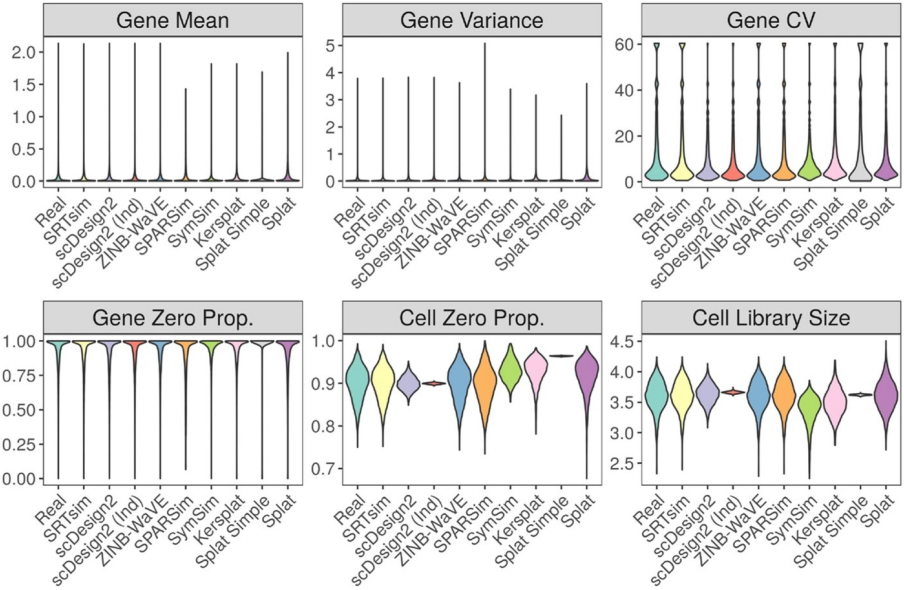
Step 2. Obtain gene-specific expression counts.



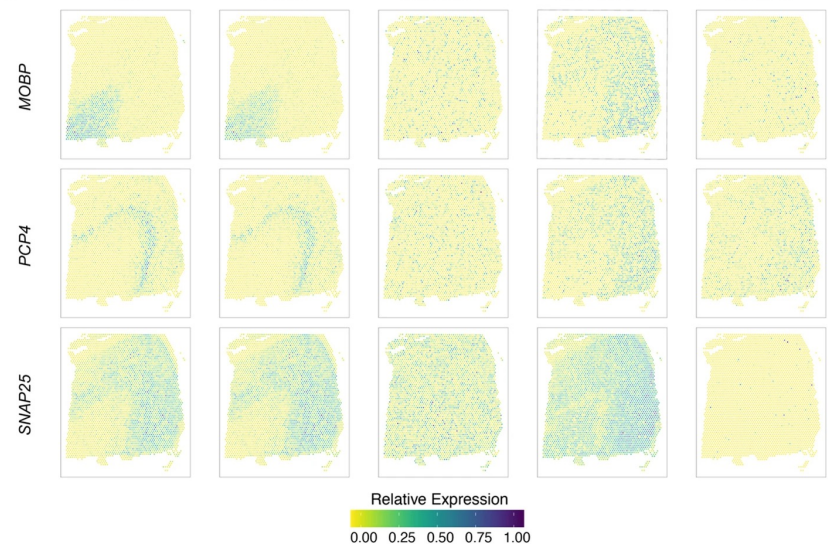
Step 3. Assign generated gene expression counts to the locations.



A



B



Web-Server for SRT Analysis

SRT-Server: Powering the analysis of spatial transcriptomics

SRT-Server

+

←

→

↻

🔒

🌐

https://spatialtranscriptomicsanalysis.com/home/homePage

📄

📌

☆

🔍

👤

🔖

☰

M

UNIVERSITY OF MICHIGAN



南京医科大学

NANJING MEDICAL UNIVERSITY

SRT-Server

Home

Tutorial

Example Data

Help

Contact

Sign up

Login

Quick Start

SRT-Server

Spatially Resolved Transcriptomics Data Analysis Service

Sign up now

Login

Quick Start

The most user-friendly and comprehensive server to analyze SRT data.



Upload your SRT file.
All the uploaded SRT data are secured.SRT-Server is compatible to four specific formats and the generic format.



Build pipeline by yourself.
All the analysis pipelines are built by drag modules and connect line. You can set different parameters even in the same module.



Download your result.
SRT-Server supplies figures and R object for different modules. All the files will be deleted after seven days.

Summary

- **SPARK**: Calibrated and powerful analysis to detect spatially variable genes (SVGs) that display spatial expression patterns
- **SPARK-X**: Scalable detection of SVGs and cell type specific SVGs
- **CARD**: Cell type deconvolution/mapping and resolution refinement
- **SpatialPCA**: Dimension reduction for spatial domain detection and trajectory inference on the tissue
- **IRIS**: Effective and scalable integrative spatial domain detection
- **BASS**: Multi-scale and multiple-sample analysis for cell type clustering and spatial domain detection
- **SRTsim**: Spatial pattern preserving simulations
- **SRT-Server**: Server for spatial transcriptomics analysis

References

- **SPARK**: Sun S, Zhu J, and Zhou X. **Nature Methods** (2020) 17: 193-200.
- **SPARK-X**: Zhu J, Sun S, and Zhou X. **Genome Biology** (2021) 22: 184.
- **CARD**: Ma Y, and Zhou X. **Nature Biotechnology** (2022) 40: 1349-1359.
- **SpatialPCA**: Shang L, and Zhou X. **Nature Communications** (2022) 13: 7203.
- **IRIS**: Ma Y, and Zhou X. **Nature Methods** (2024). in press.
- **BASS**: Zheng L, and Zhou X. **Genome Biology** (2022). 23: 168.
- **SRTsim**: Zhu J, Shang L, and Zhou X. **Genome Biology** (2023). 24: 39
- **SRT-Server**: Yang S, and Zhou X. **Genome Medicine** (2024). 16: 18
- All software packages are available at: <https://xiangzhou.github.io>

Acknowledgement



R01HG009124
R01GM126553
R01HG011883
R01GM144960



Xiang Zhou Lab

