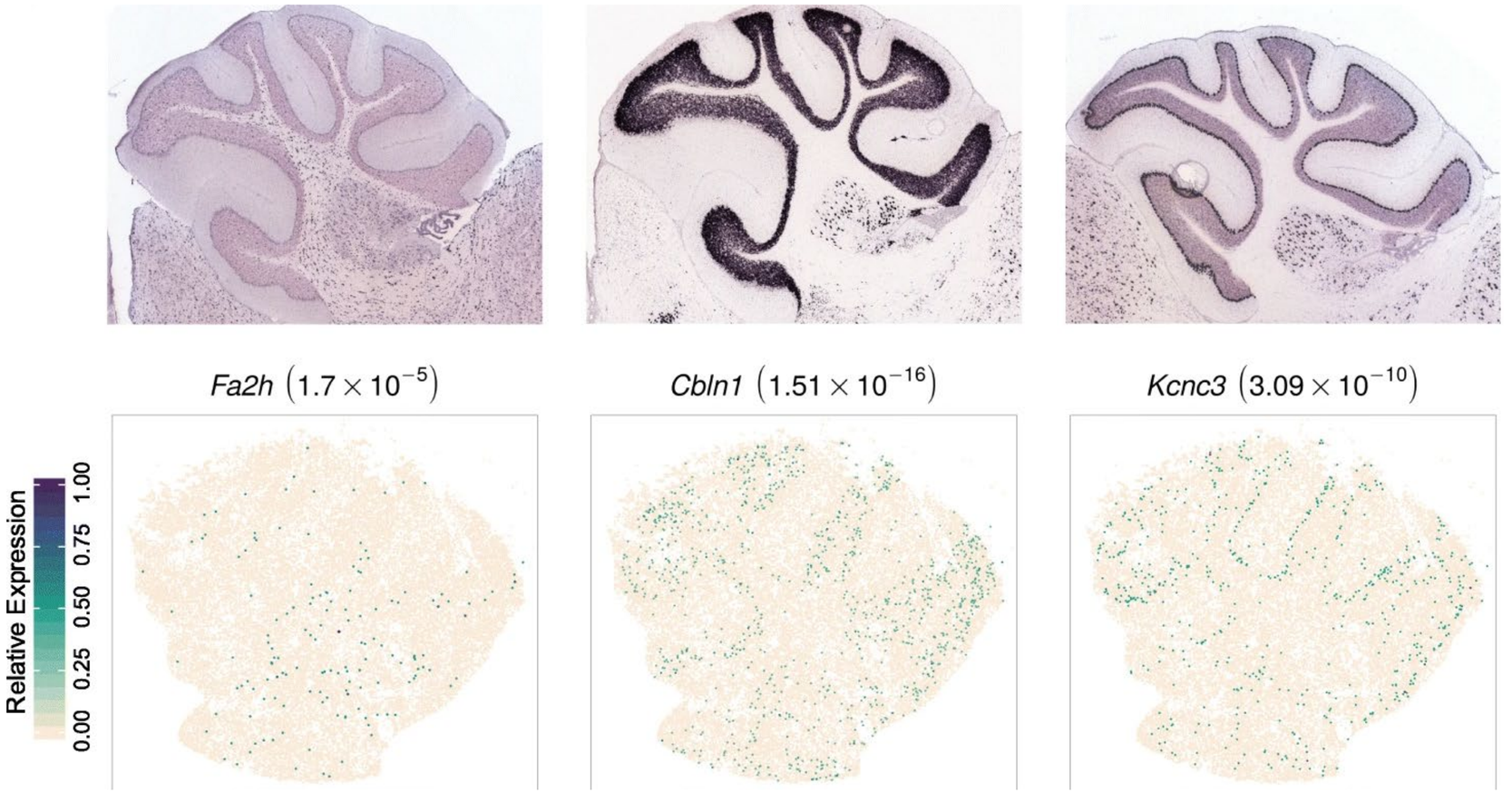
Mapping spatially variable genes at cell type and subcellular resolution in spatial transcriptomics

Xiang Zhou

Department of Statistics and Data Science
Yale University

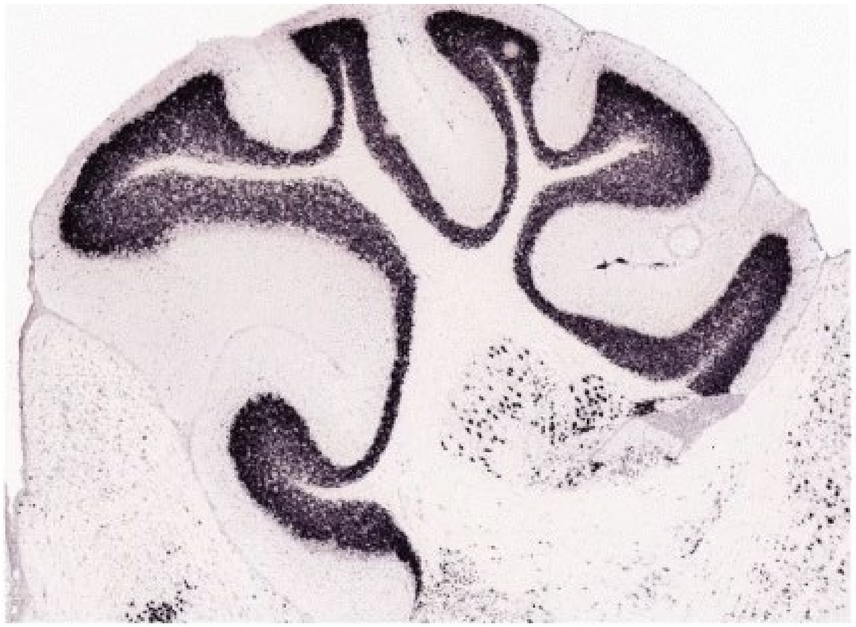
Part I: Mapping Cell Type Specific SVGs

Spatially Variable Genes (SVGs)



Identifying SVGs is important for characterizing the spatial and functional organization of complex tissues.

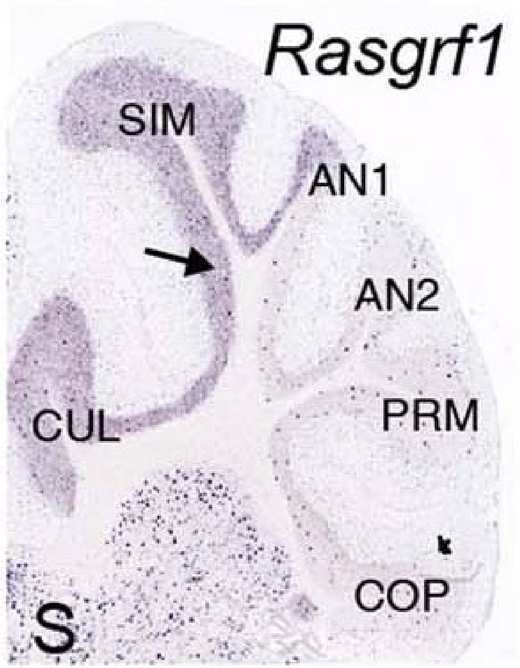
Spatially Variable Genes (SVGs)



Cbln1 $\left(1.51 \times 10^{-16}\right)$

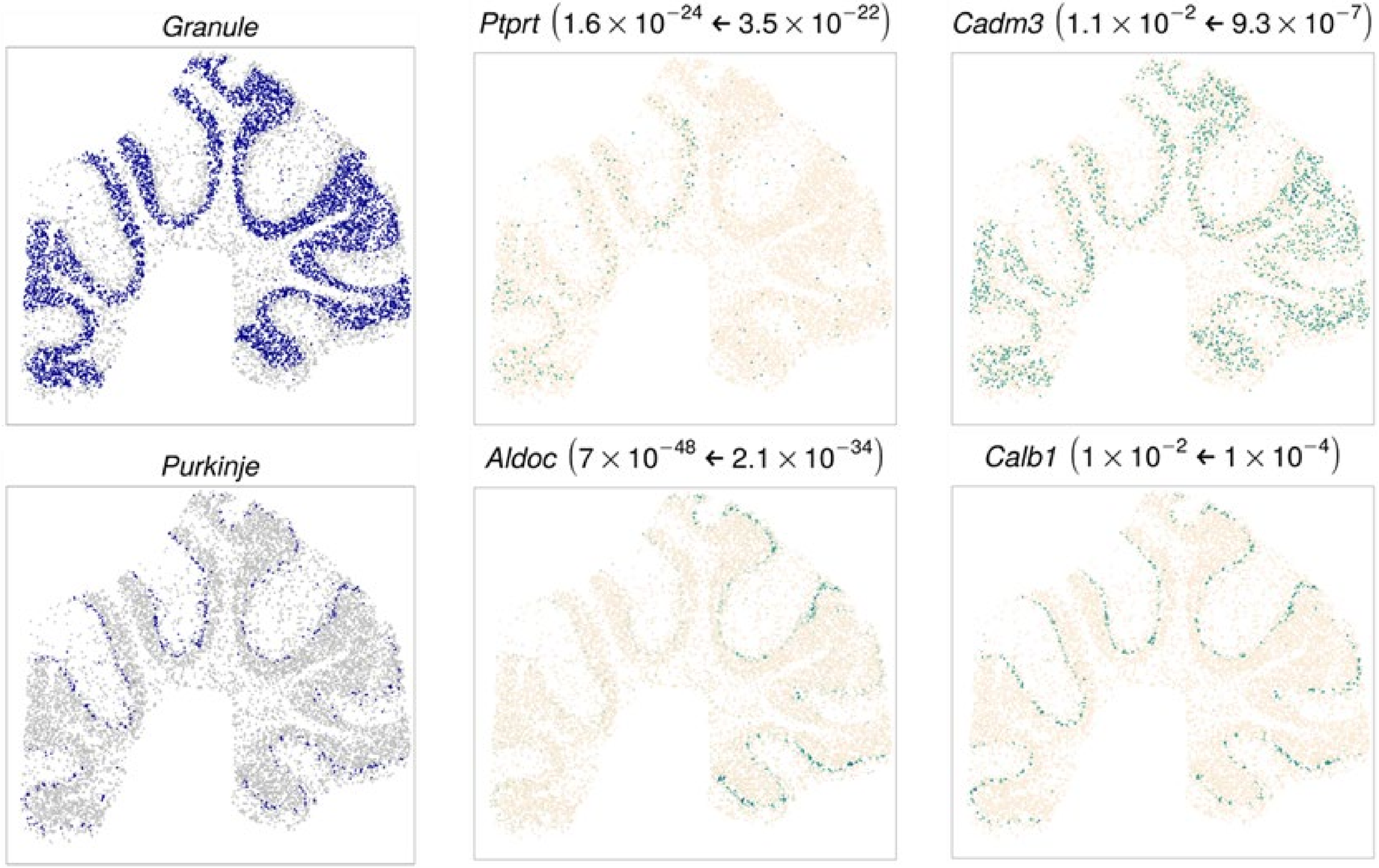
- Many previously identified spatially variable genes are cell type marker genes
- They display spatial expression patterns that mirror the distribution of distinct cell types

Cell Type-Specific SVGs



- However, a substantial fraction of previously discovered SVGs, which we will refer to as cell type-specific spatially variable genes (**ct-SVGs**)
- They display diverse spatial expression patterns within a specific cell type

SVGs vs ct-SVGs



Detecting ct-SVGs

Detecting these ct-SVGs holds the potential to

- delineate the spatial transcriptomic heterogeneity within a particular cell type
- understand the transcriptomic mechanisms underlying cellular heterogeneity

Modifying Existing Approaches to Detect ct-SVGs

Single Cell Resolution Data

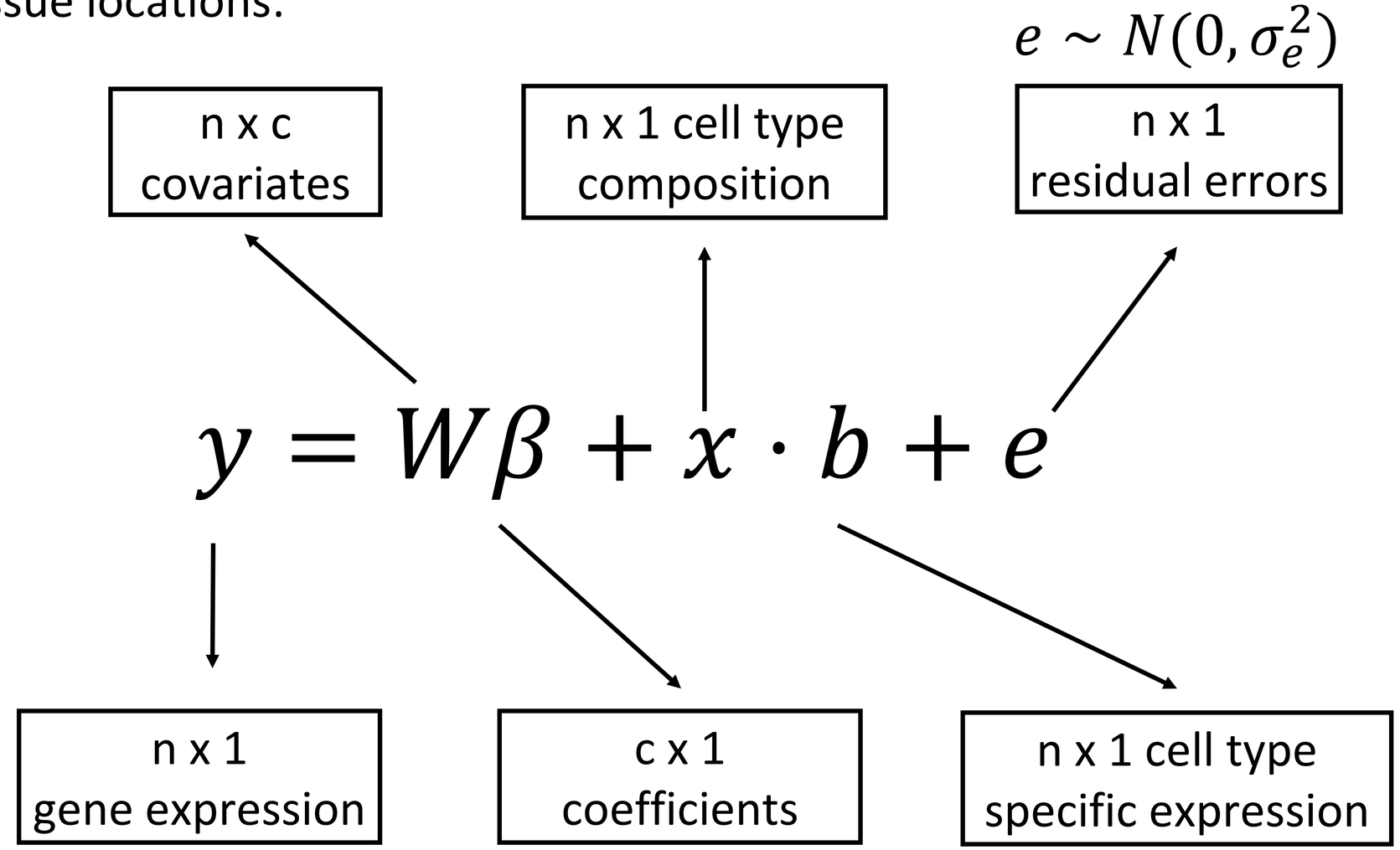
- Directly apply existing SVG detection methods such as SPARK and SPARK-X on cells from a particular cell type
    - Reasonably effective, though power can vary depending on which method is applied

Spot Resolution Data

- Apply existing SVG detection methods while controlling for cell type information
    - Excessive false signals
- Apply CSIDE, which was originally developed in the context of spatial differential expression analysis
    - Analysis failure in a significant proportion of genes
    - Producing p values exactly equaling to one in a substantial fraction of the remaining

Celina: CELl type-specific spatially variable gene IdentificatioN Analysis

For each gene, Celina models the gene's spatial expression pattern with respect to the cell type distribution across tissue locations:



n: number of locations
c: number of covariates

Decompose Cell Type Specific Expression

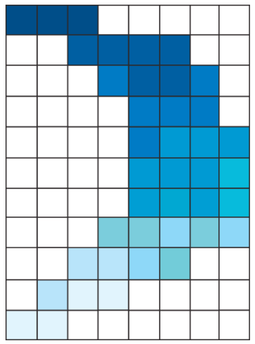
Cell type specific expression is decomposed into three parts

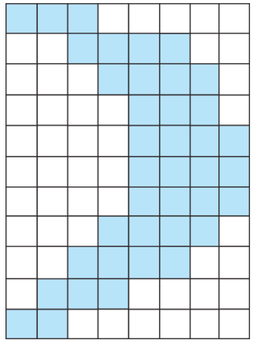
$$\mathrm{b} = 1_n b_0 + b_s + b_r$$

cell type specific mean	n x 1 spatial effect	n x 1 non-spatial effect

$$b_s \sim MVN(0, \sigma_s^2 K) \qquad b_r \sim MVN(0, \sigma_r^2 I)$$

K is a spatial kernel

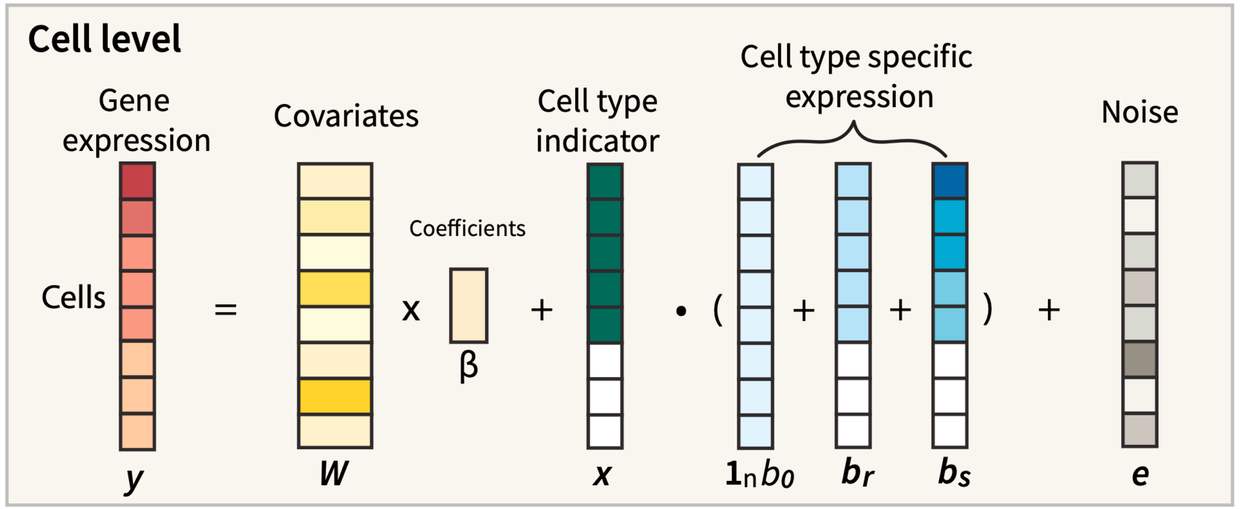


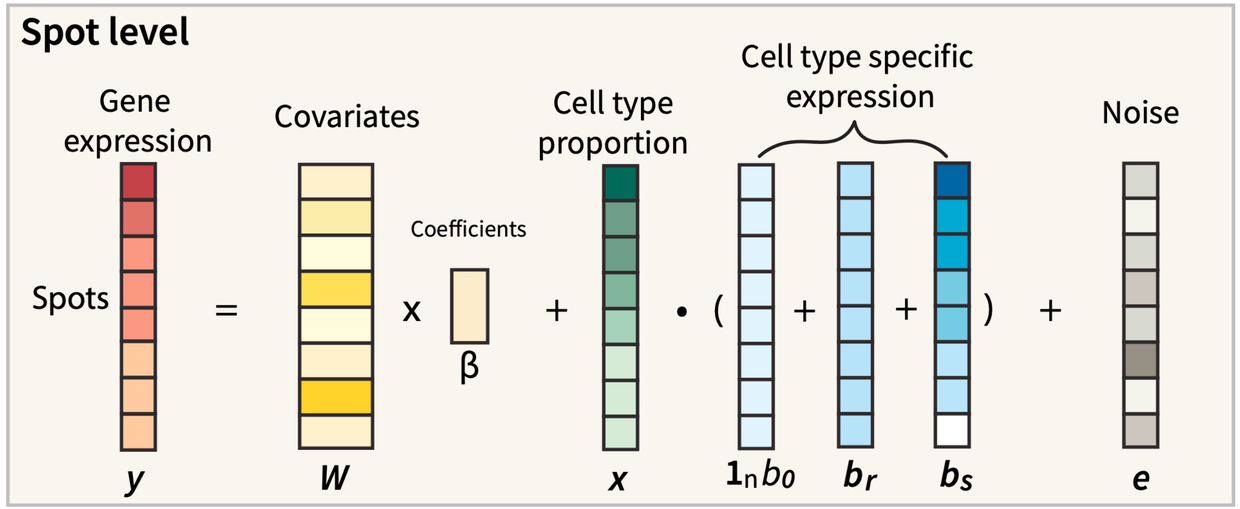


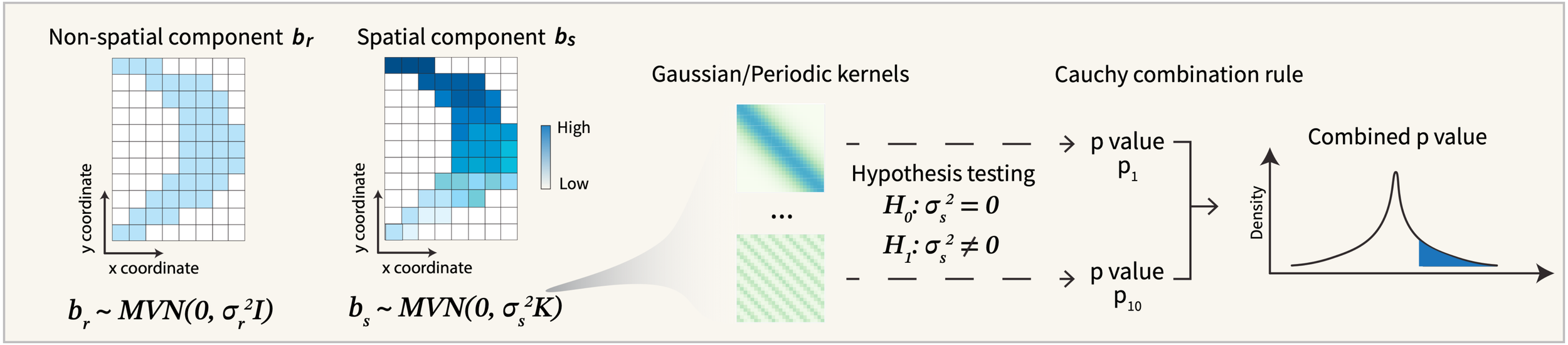
Inference Procedure

- Directing ct-SVGs is equivalent to testing $H_0: \sigma_s^2 = 0$
- Perform inference based on penalized quasi-likelihood followed by an average information algorithm for mixed model
- Use different kernels to capture distinct spatial correlation structure
- Obtain a score test statistic for each kernel
- Calculate the exact p-value based on a mixture of chi-squares
- Combine *k* p-values through the Cauchy combination rule

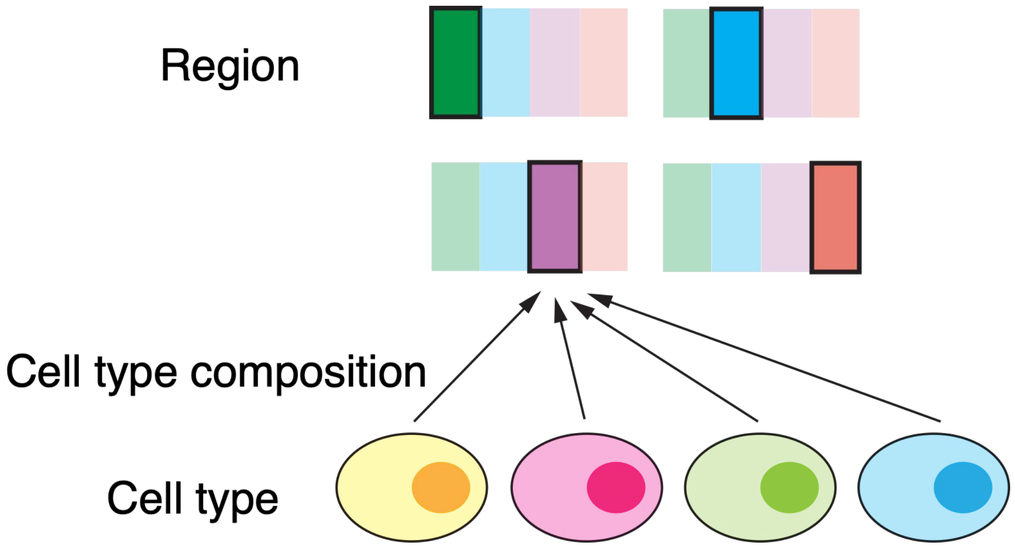
Celina Overview







Simulations

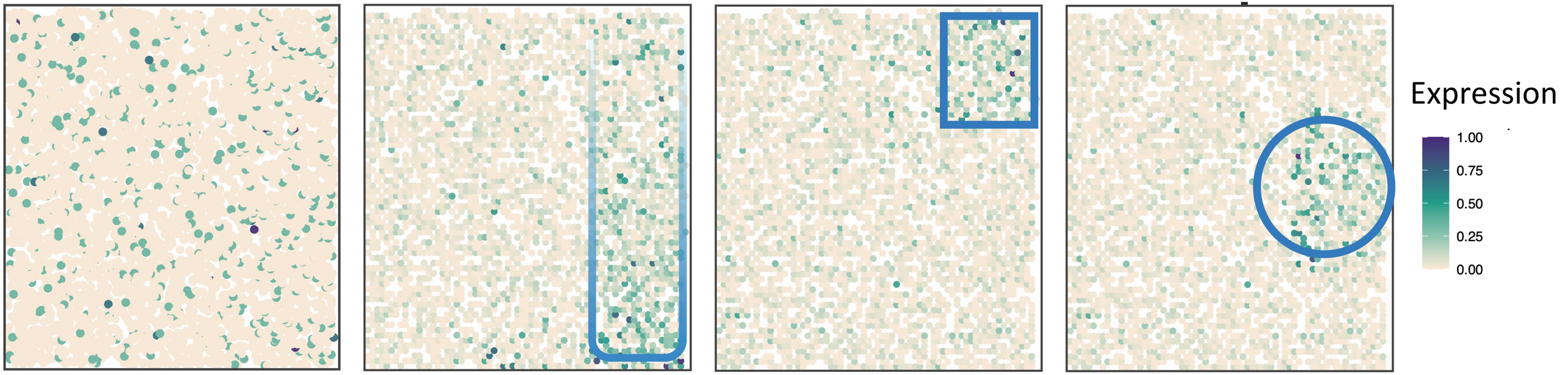


Different regions contain distinct cell type compositions

Scenario 1	100%	0%	0%	0%
Scenario 2	90%	5%	5%	0%
Scenario 3	50%	25%	0%	25%
Scenario 4	33.3%	0%	33.3%	33.3%

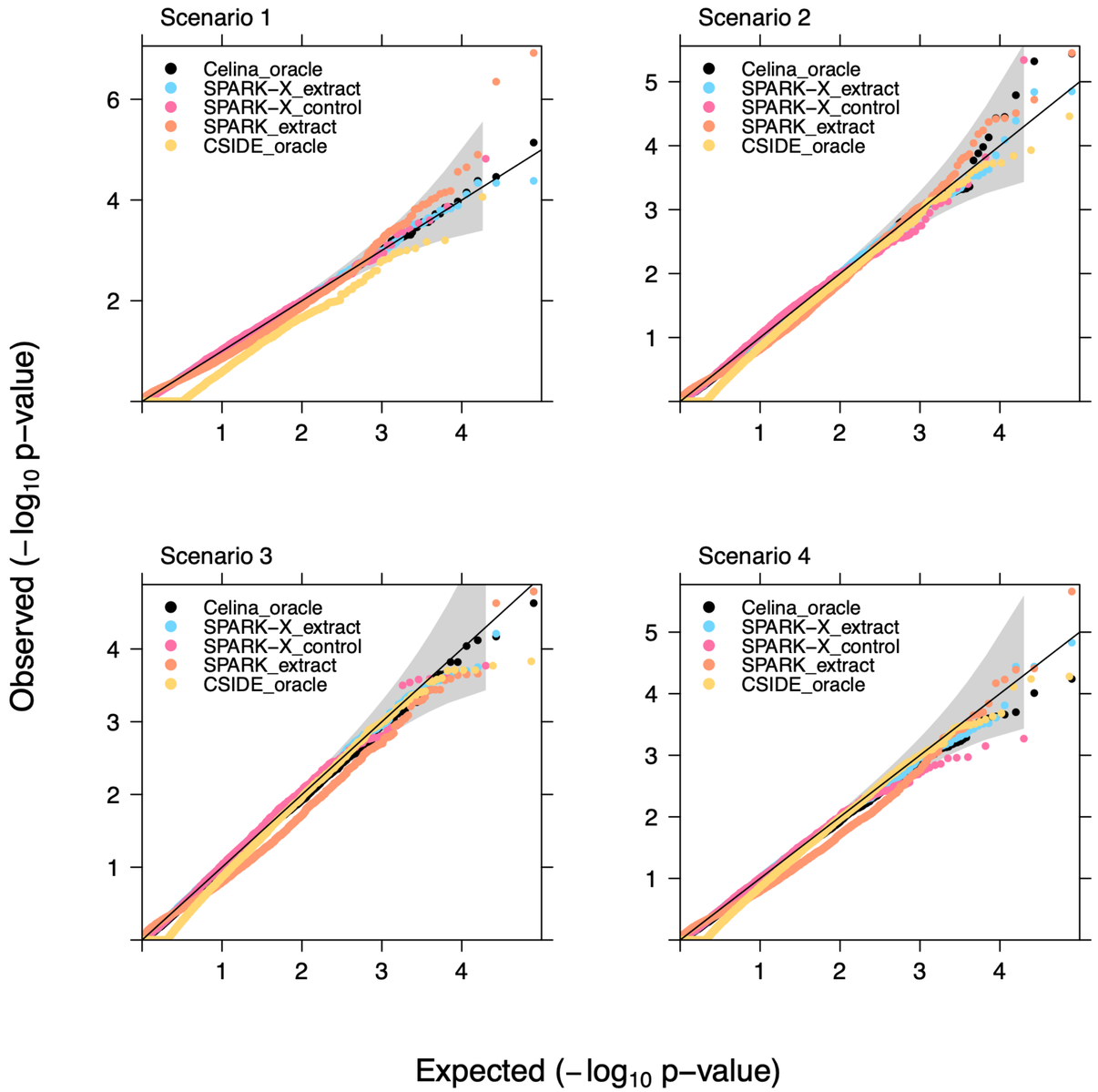
- 20% genes are SVGs
- 70% genes are cell type-specific SVGs (Null)
- 10% genes are cell type-specific SVGs (Alternative)

Simulated Spatial Patterns



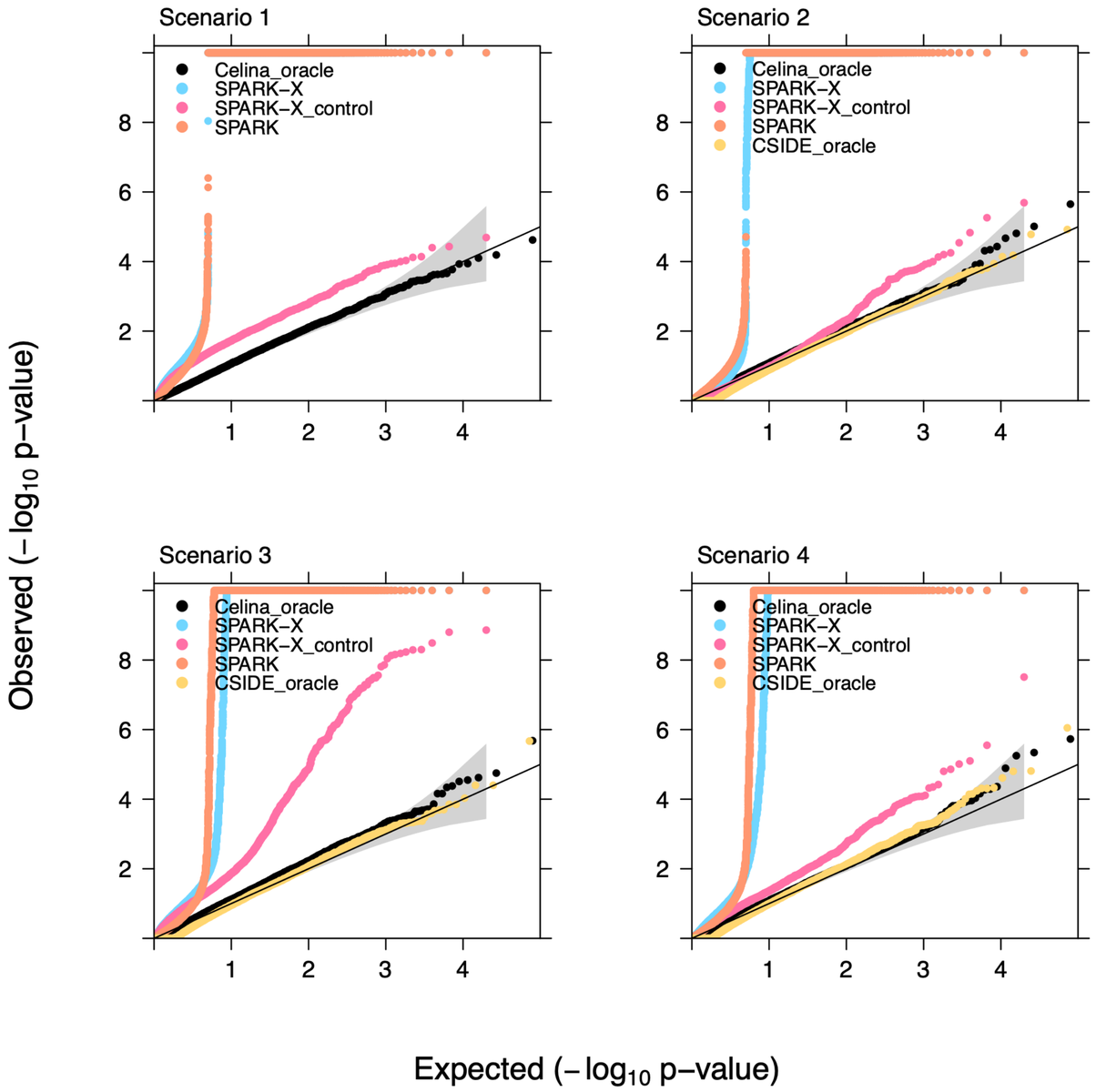
Random Gradient Streak Hotspot

Single Cell Resolution Simulations: Type I Error



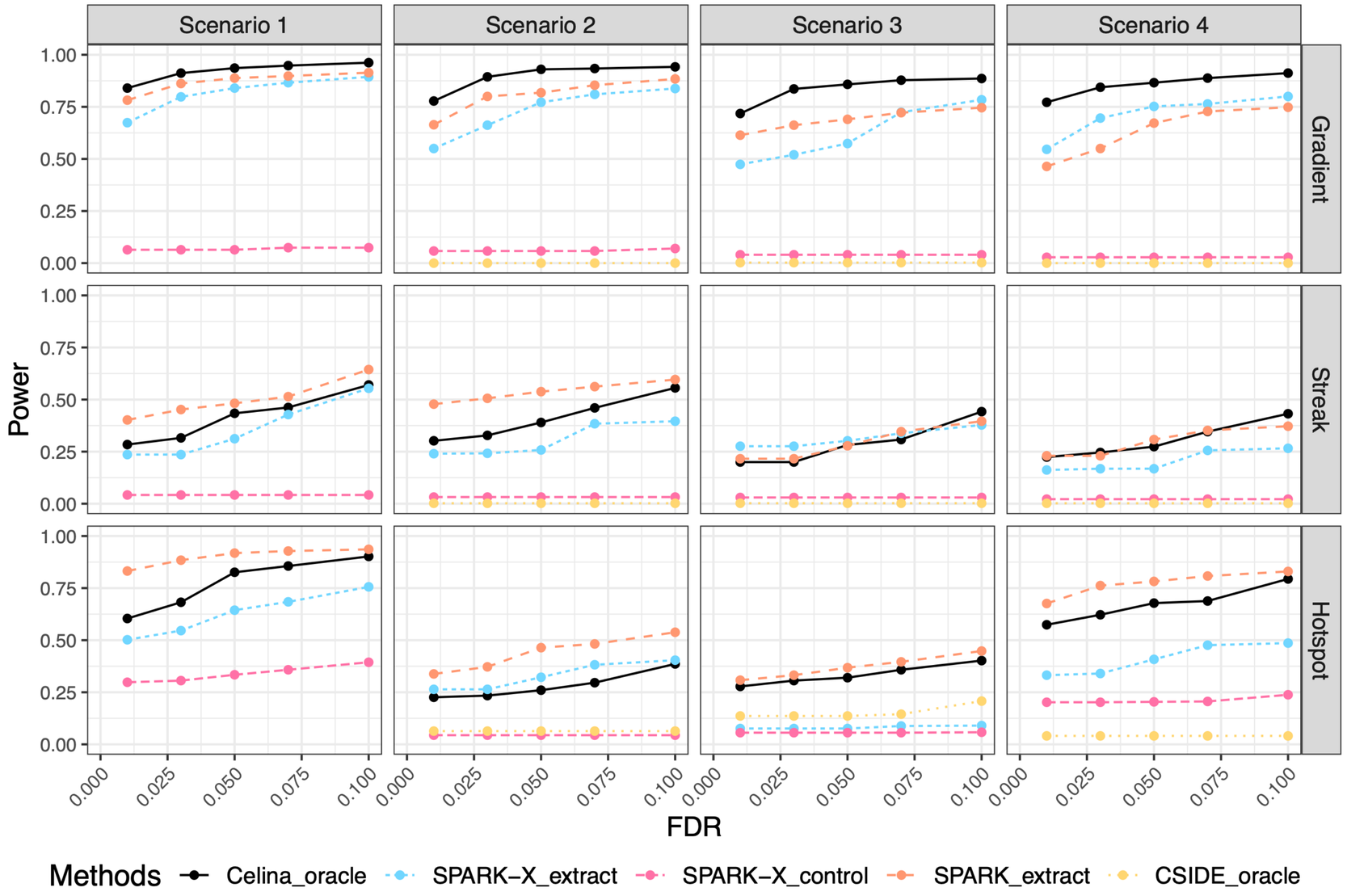
- Celina, SPARK and SPARK-X have calibrated Type I error control
- CSIDE generates a large proportion of p-values = 1

Spot Resolution Simulations: Type I Error



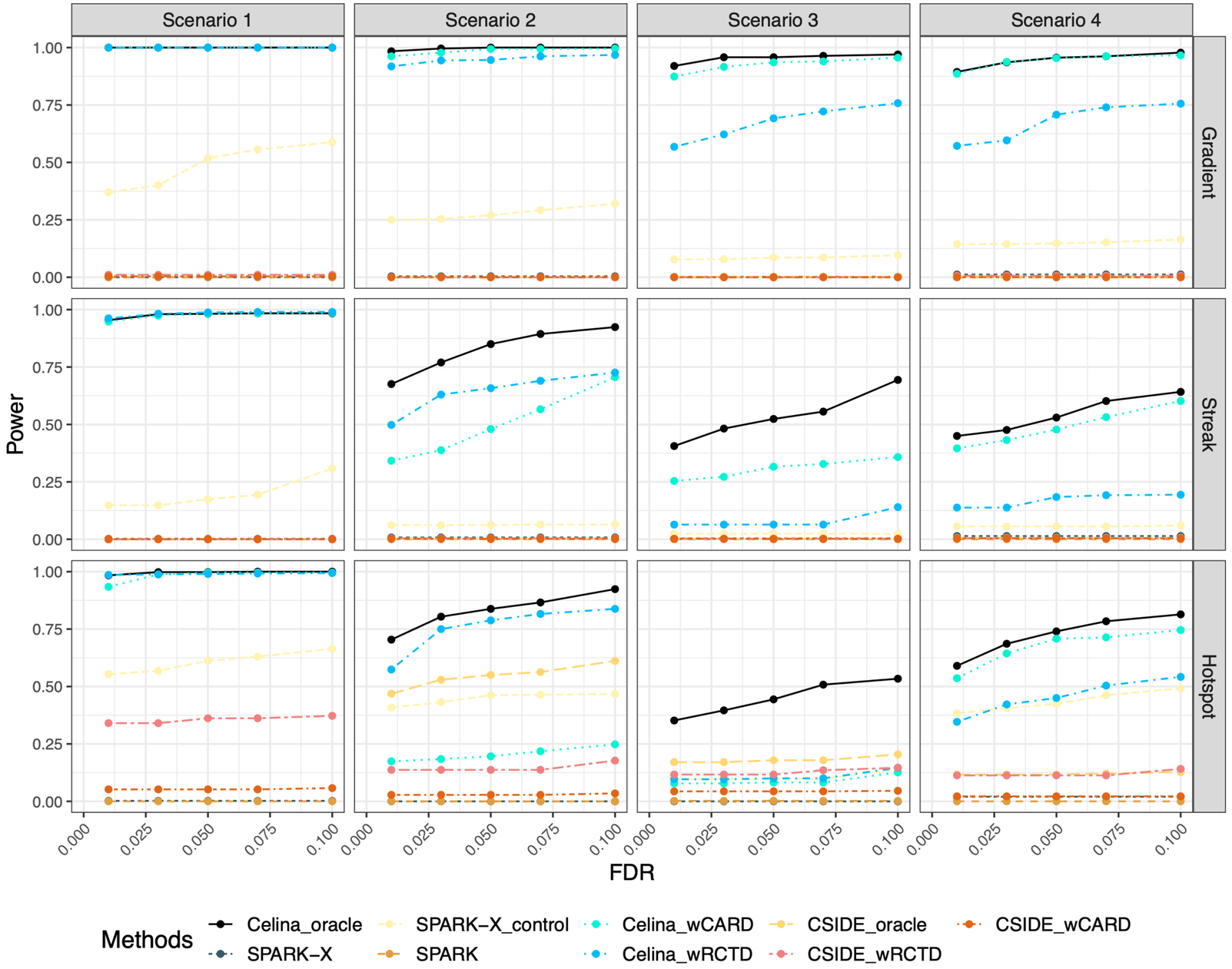
- Celina achieves calibrated Type I error control
- SPARK and SPARK-X produces inflated p-values
- CSIDE generates a large proportion of p-values = 1

Single Cell Resolution Simulations: Power



- Celina is better in detecting the gradient pattern and performs similarly to SPARK in other two patterns.
- CSIDE achieves very low power compared to other three methods.

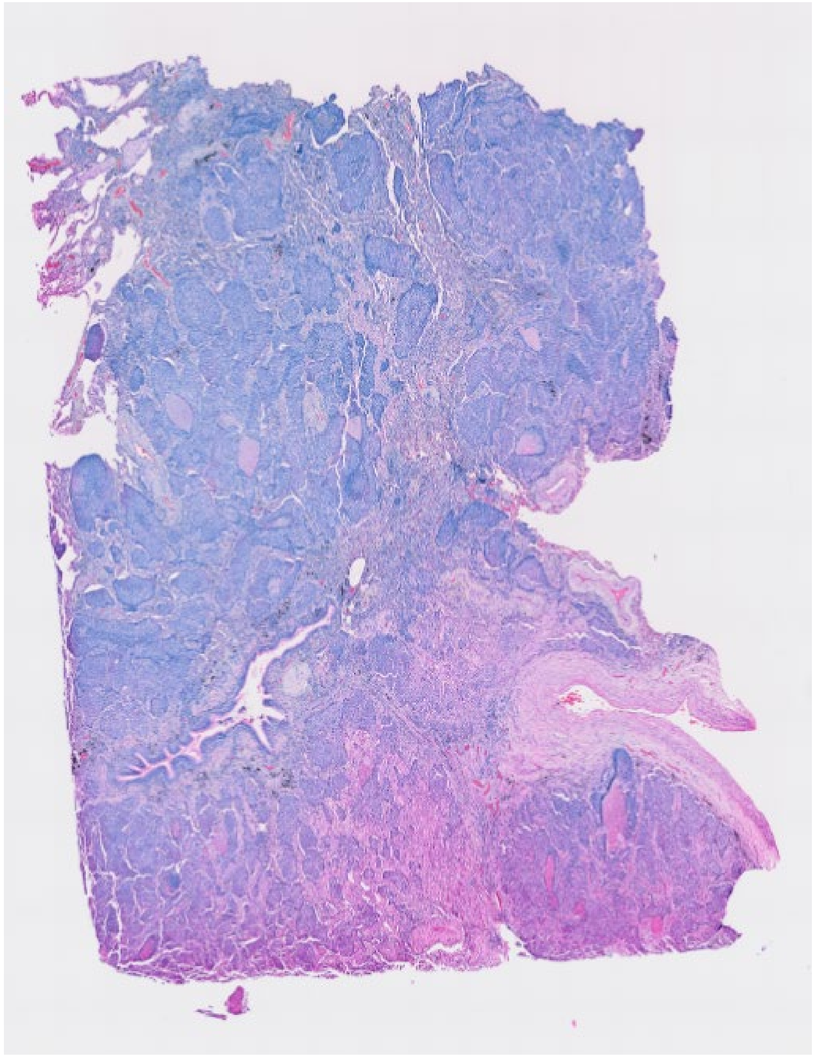
Spot Level Simulations: Power



- Celina outperforms the other methods across all patterns.

- More accurate cell type deconvolution by CARD leads to higher power.

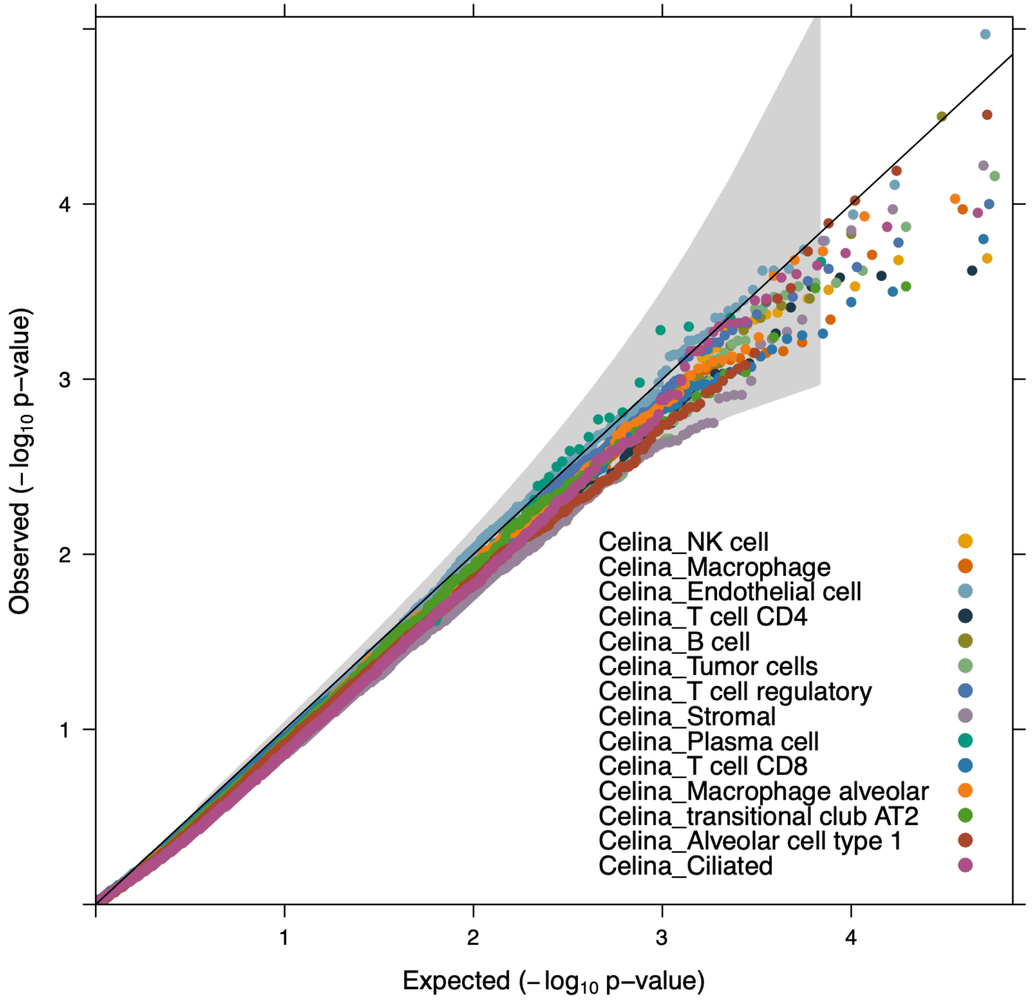
Real Data Application: Human Lung Cancer



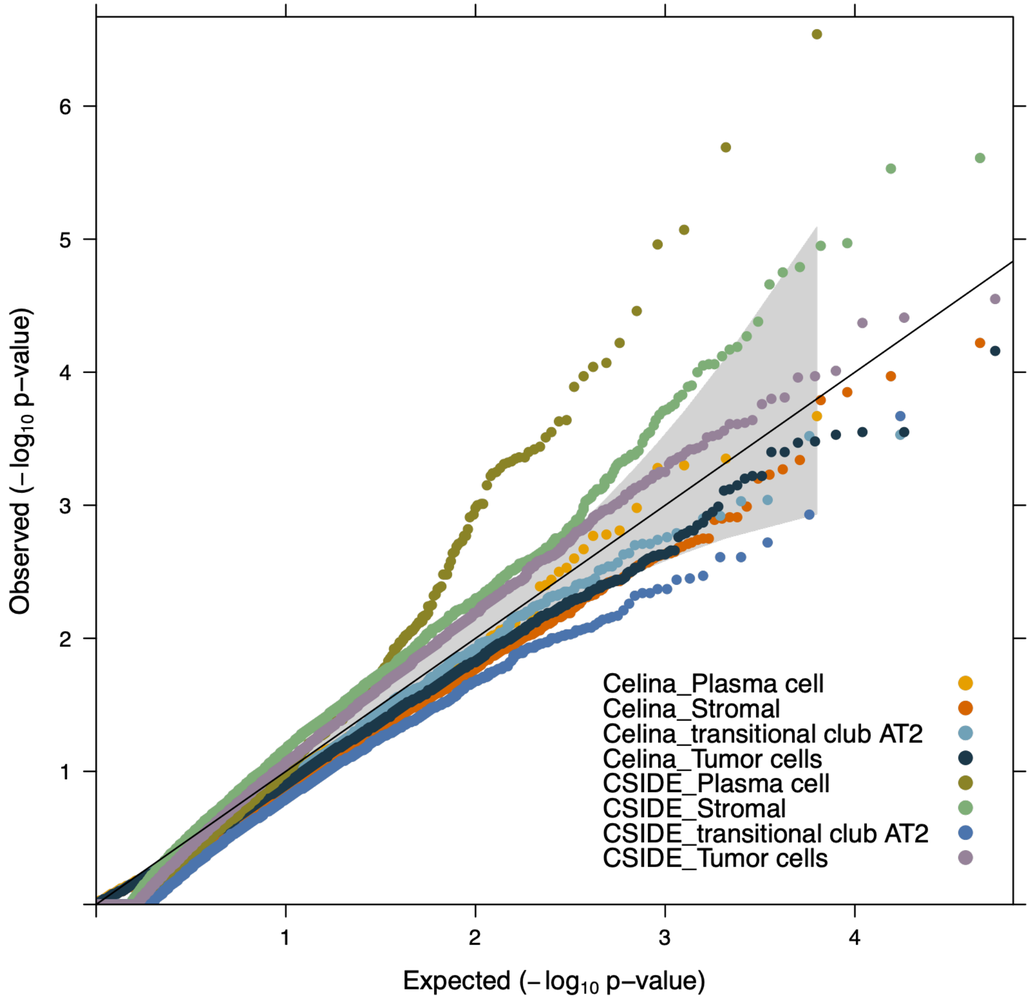
- 10X Visium
- Followed standard filtering, used 17,257 genes measured on 3,813 locations for analysis
- Performed cell type deconvolution on spots using Salcher et al. 2022 as single cell reference

Type I Error Control

QQ plots under the permuted null

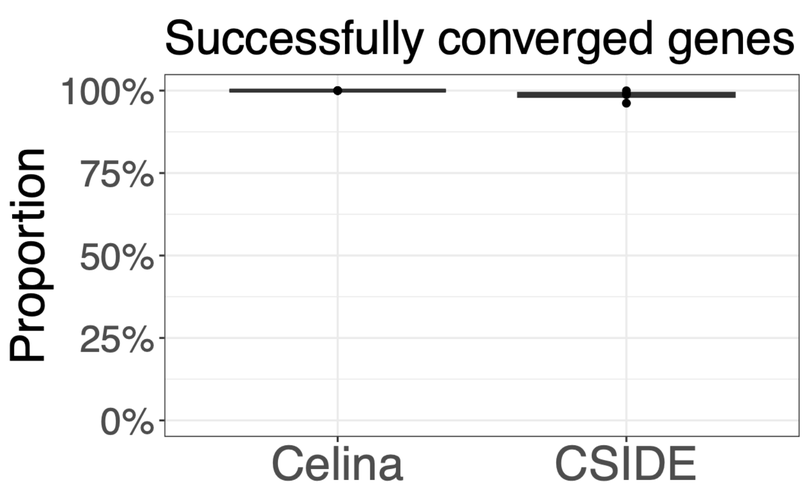


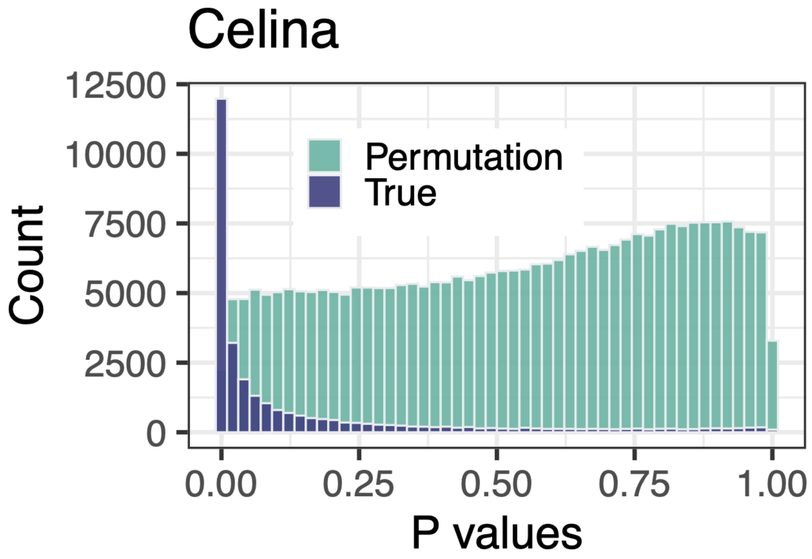
QQ plot: Celina permuted p-values

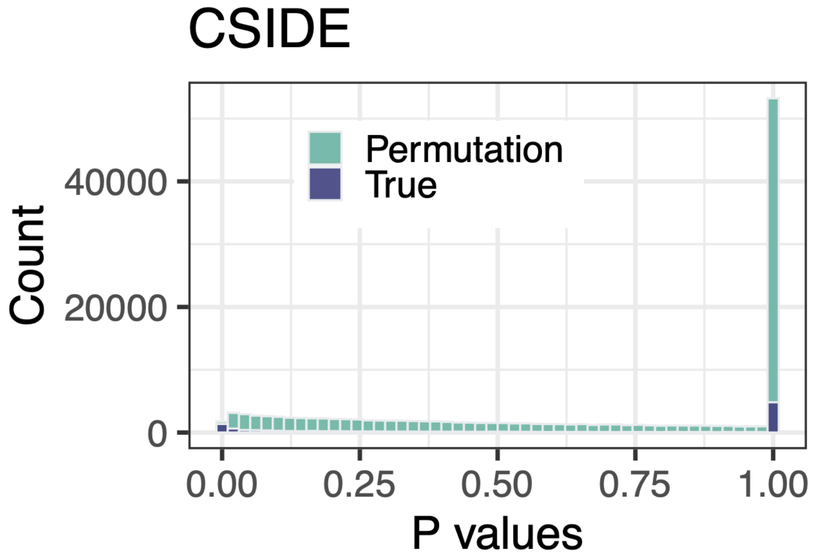


QQ plot: Celina vs CSIDE permuted p-values

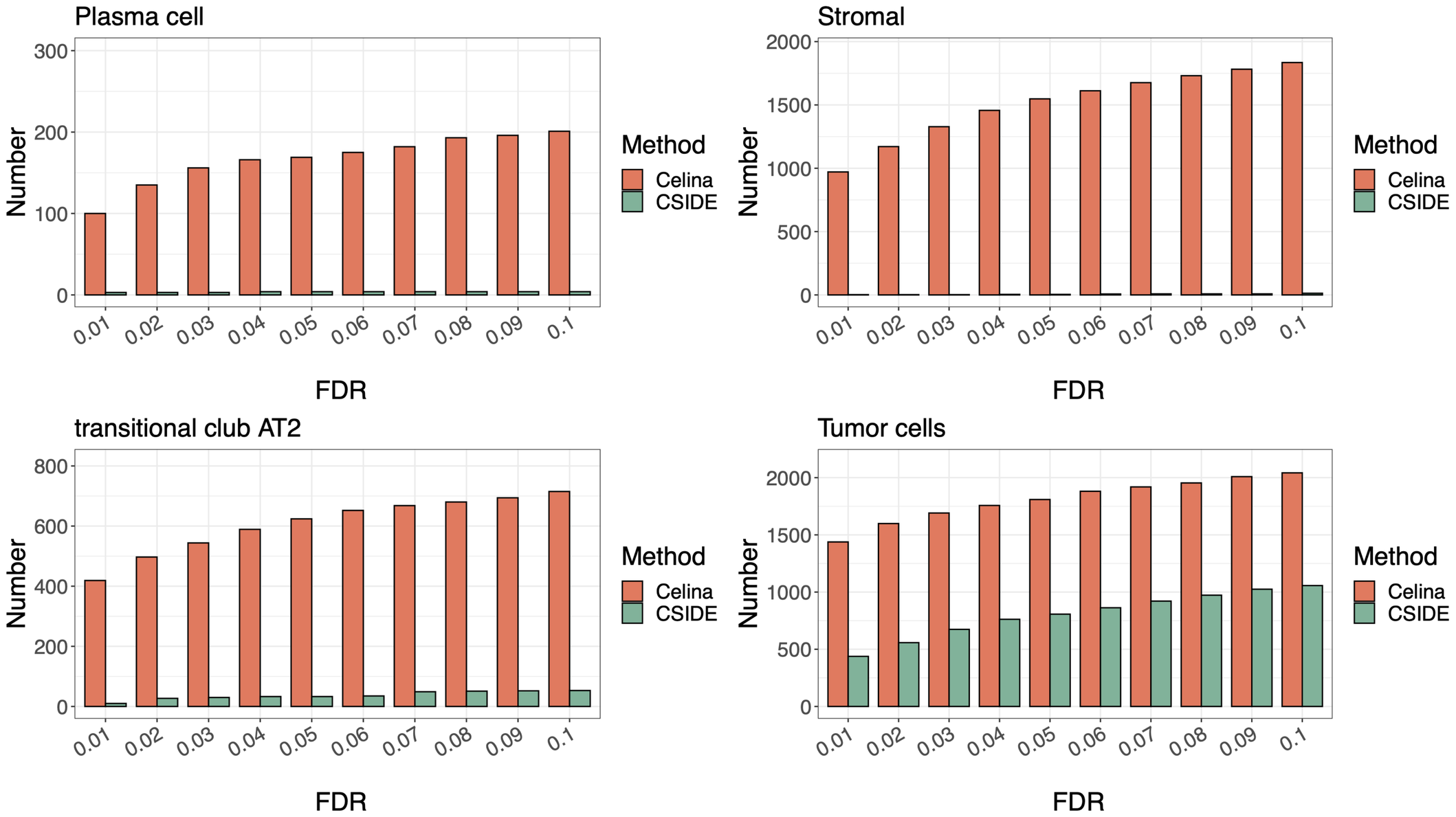
Algorithm Convergence Rates and p-value Distributions



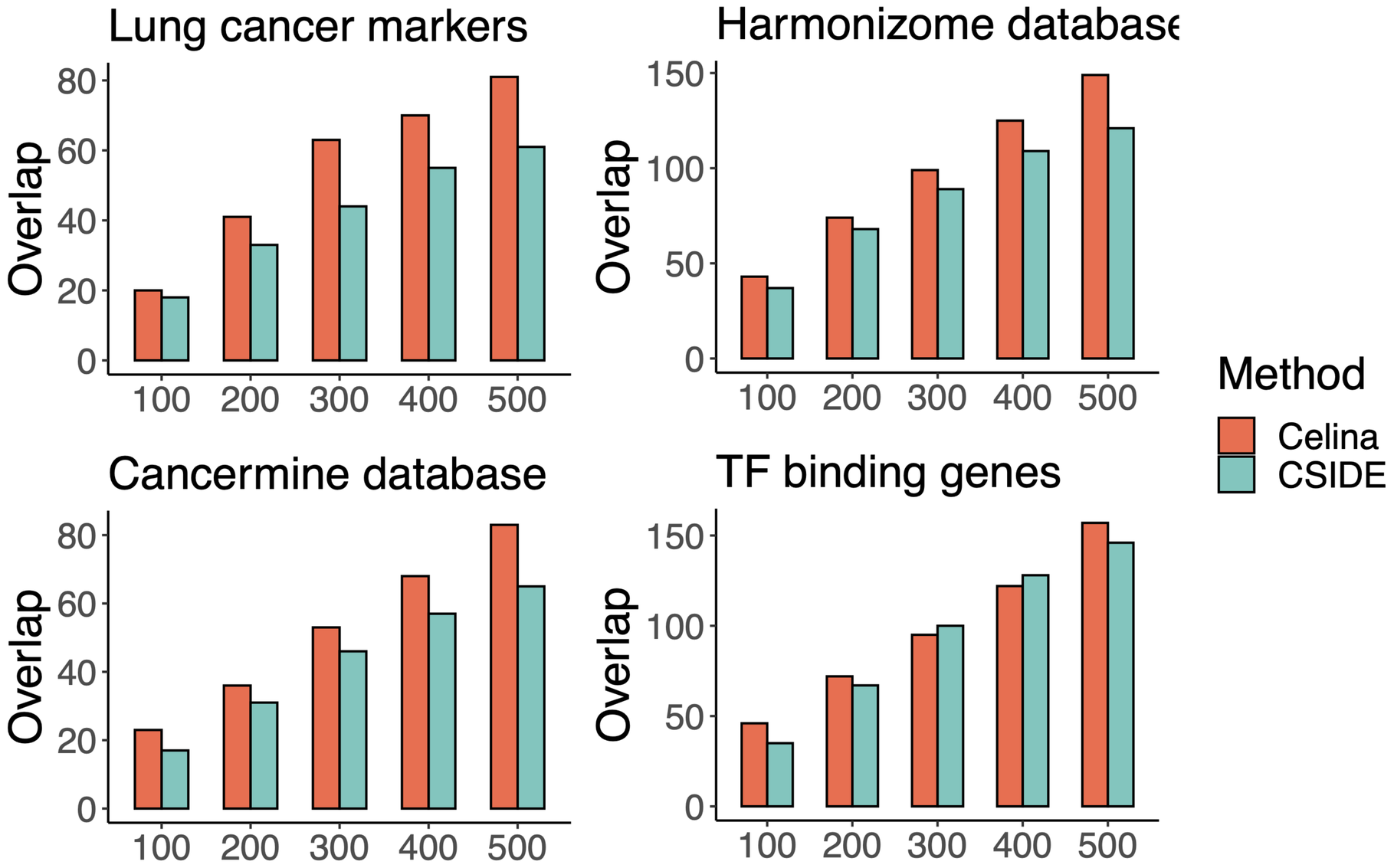




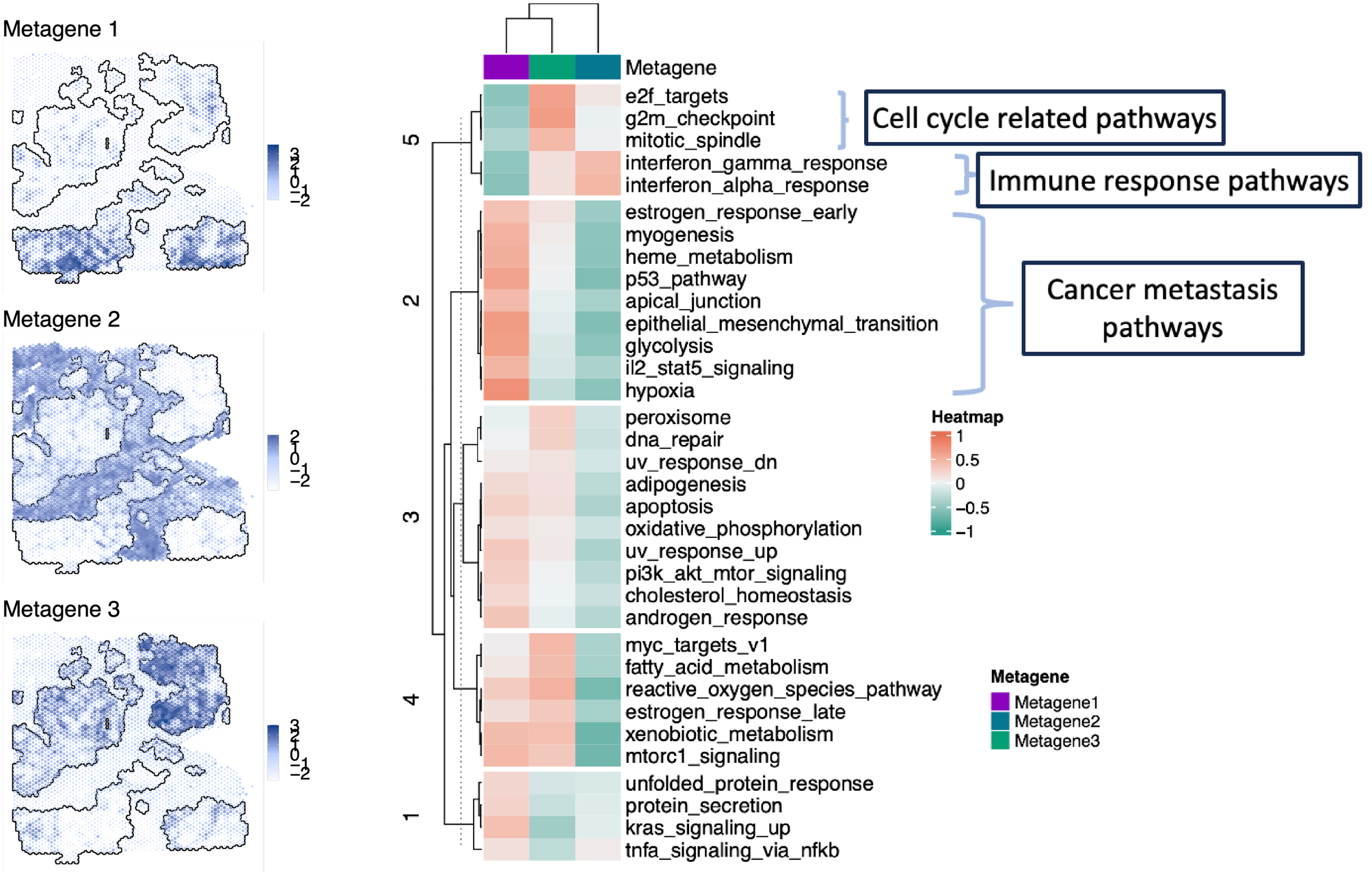
Number of Genes Detected



Enrichment of Top Detected Genes in Existing Functional Databases

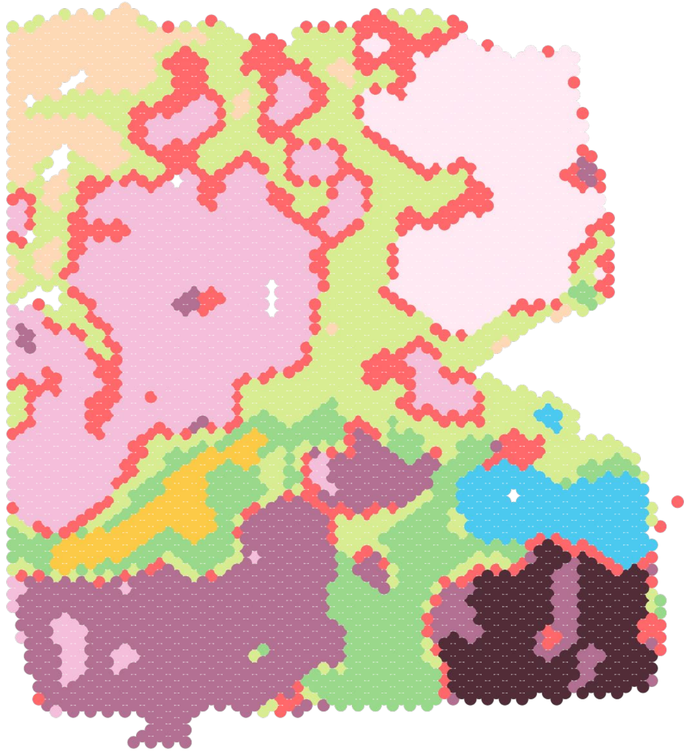


ct-SVGs are Classified into Three Categories

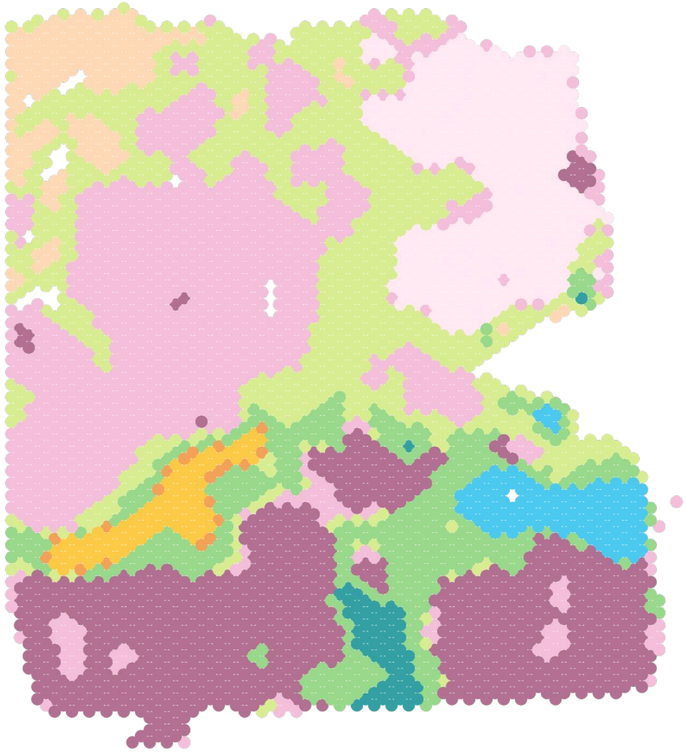


Detection of Tumor Boundary with ct-SVGs

SpatialPCA+Celina



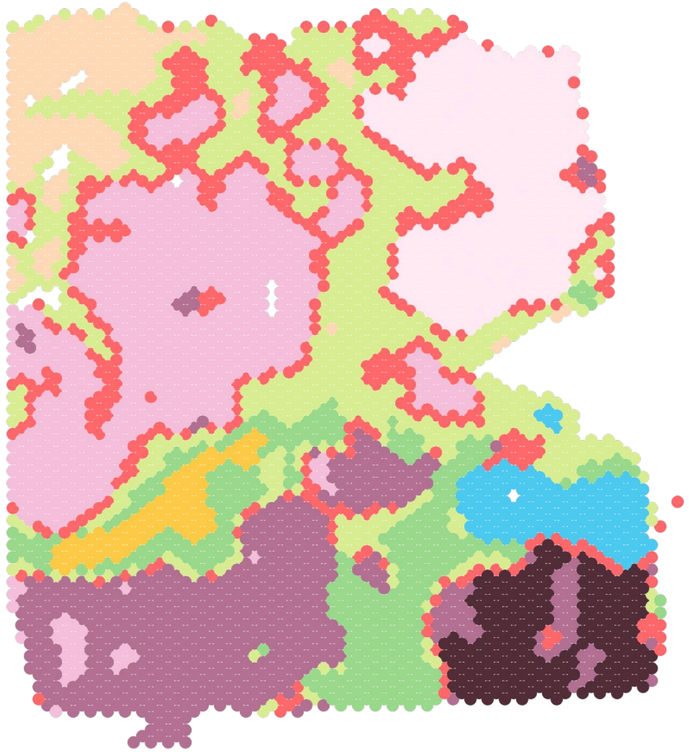
SpatialPCA default

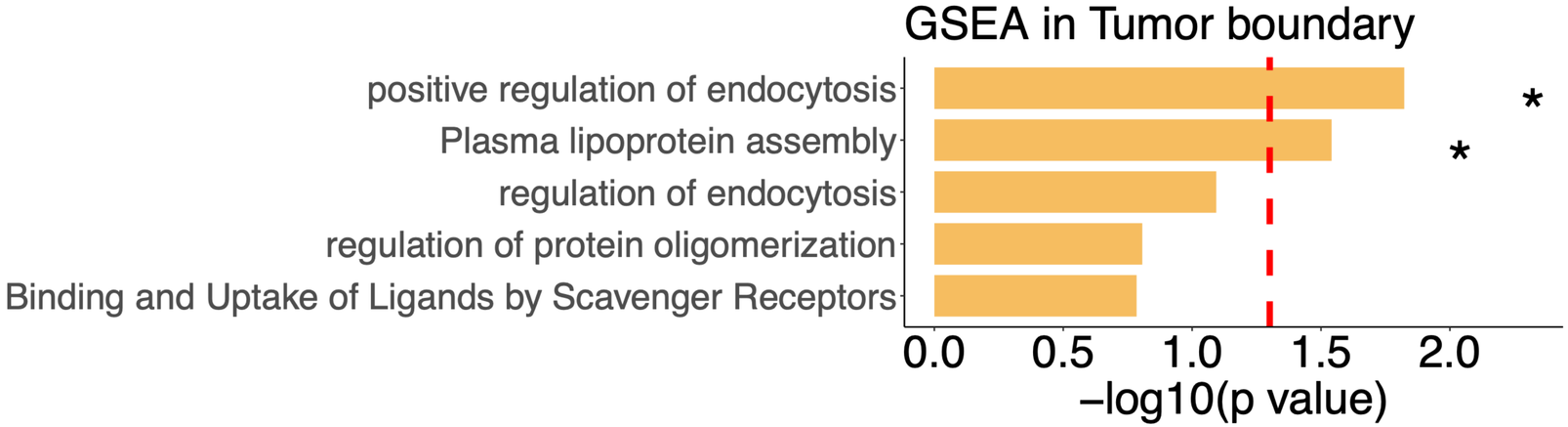


- Tumor subregion1
- Tumor subregion2
- Tumor subregion3
- Tumor subregion4
- Near tumor subregion1
- Near tumor subregion2
- Ciliated/Club
- Stroma
- Alveolar type II
- Tumor boundary

Genes in Tumor Boundary are Enriched in Pathways Related to Tumor Microenvironment

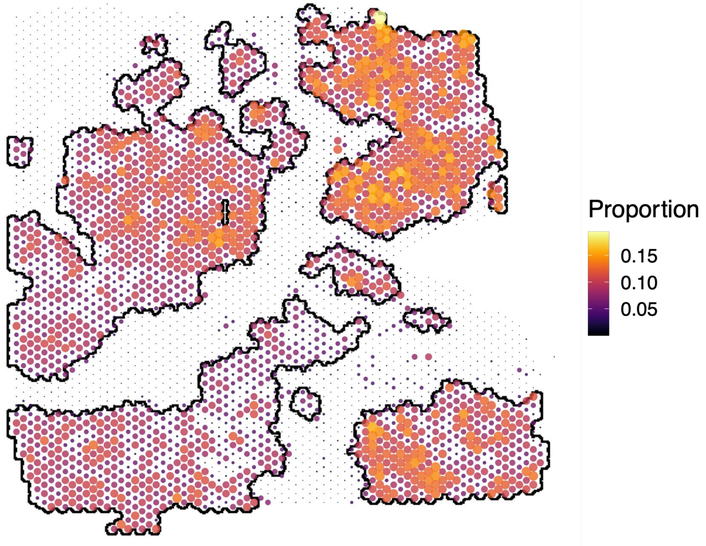
SpatialPCA+Celina



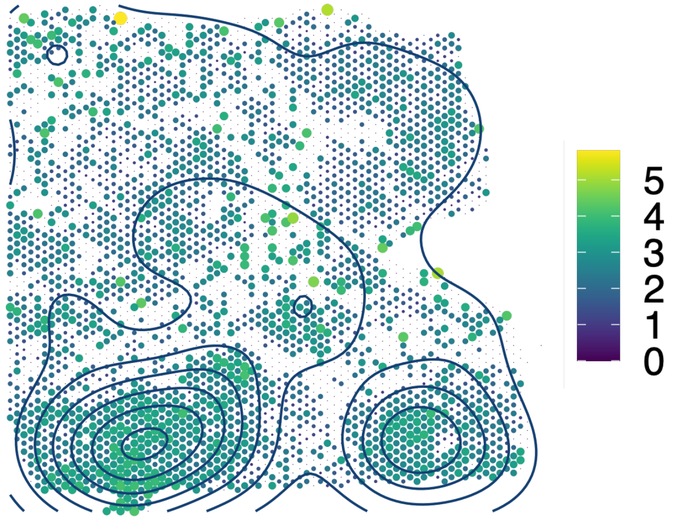


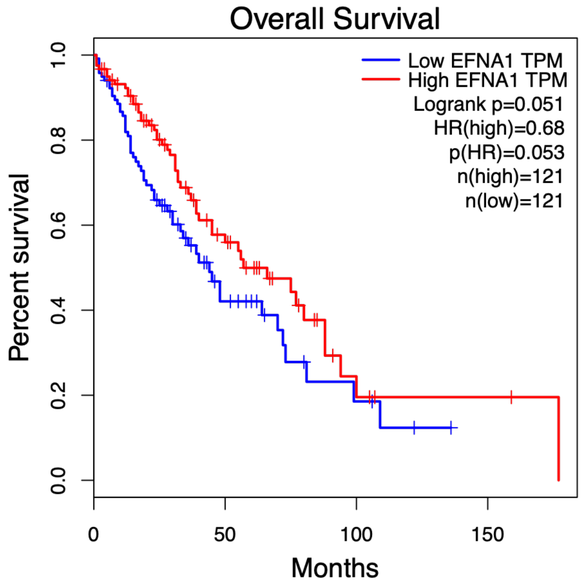
Two Example Tumor ct-SVGs

Tumor cells

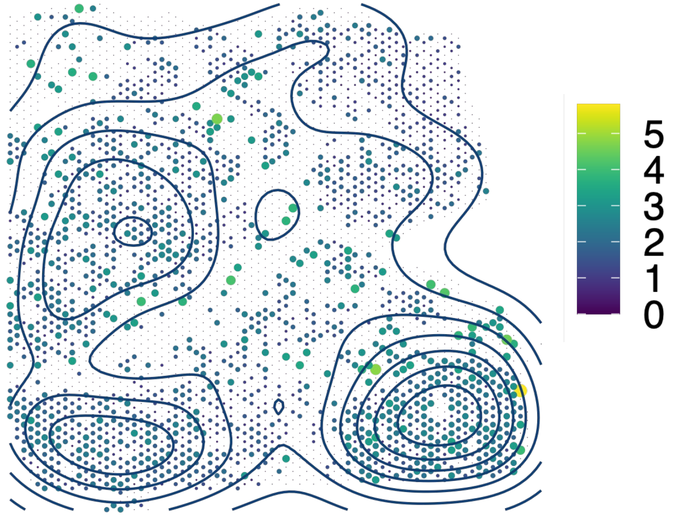


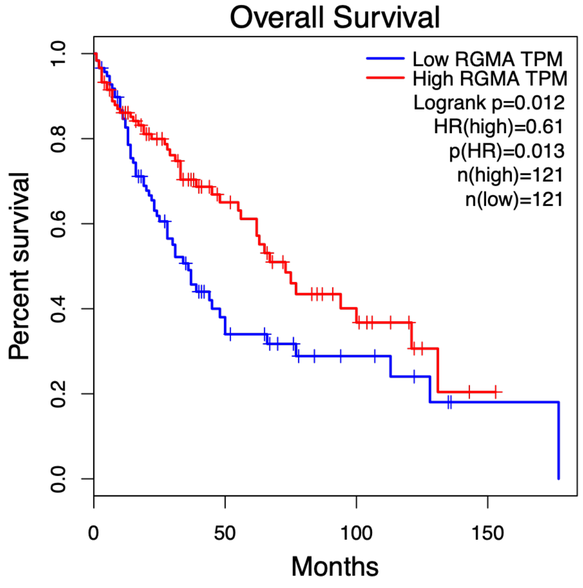
EFNA1





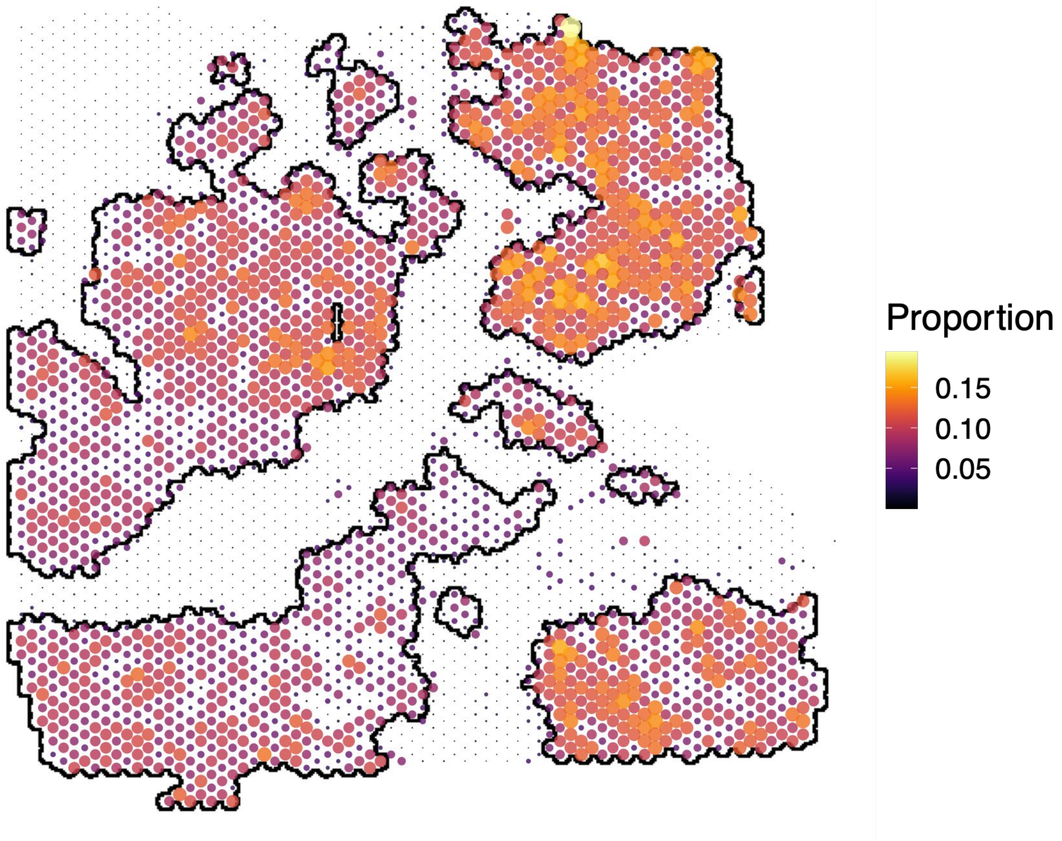
RGMA



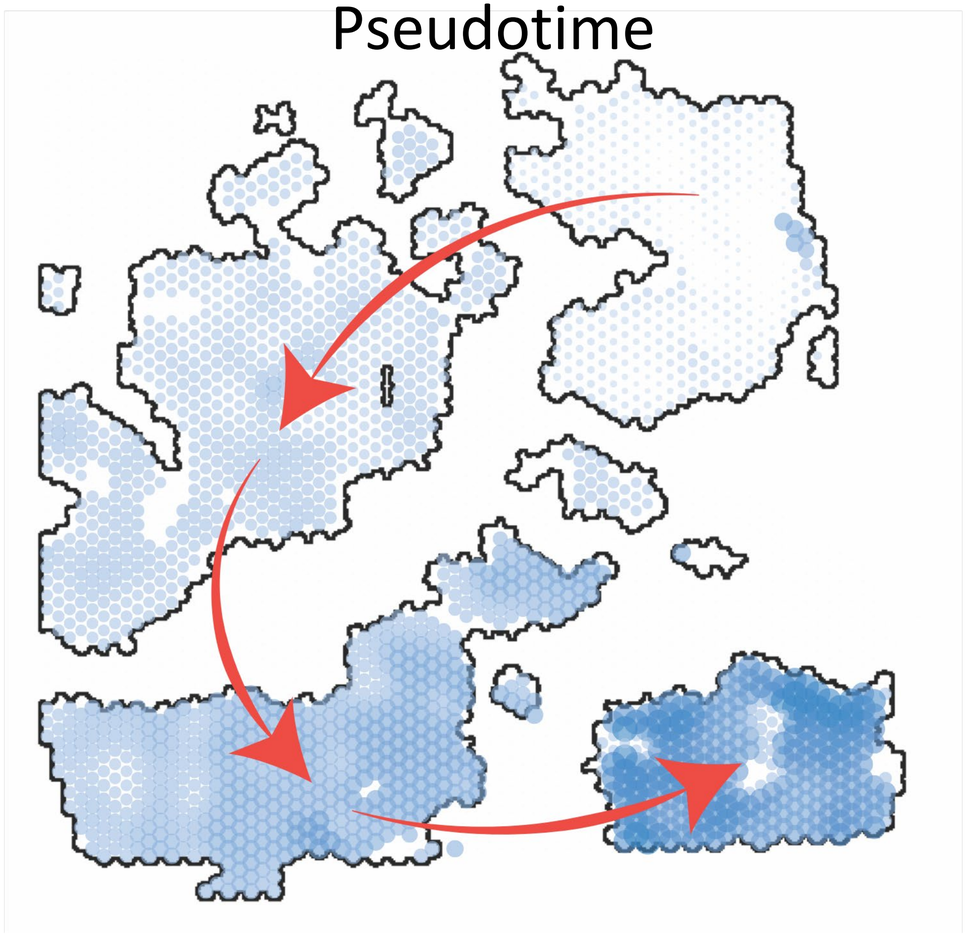


Inferring Trajectory in Tumor

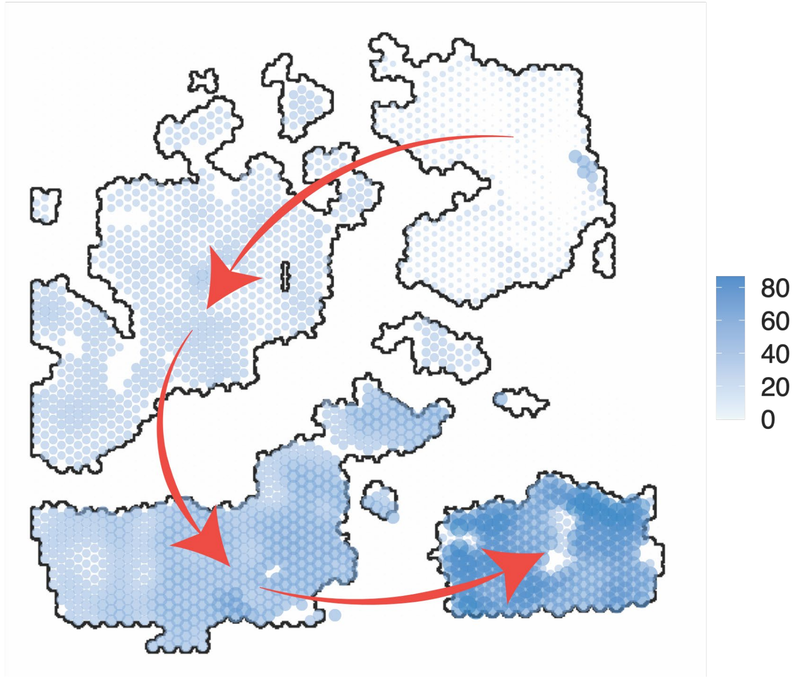
Tumor cells

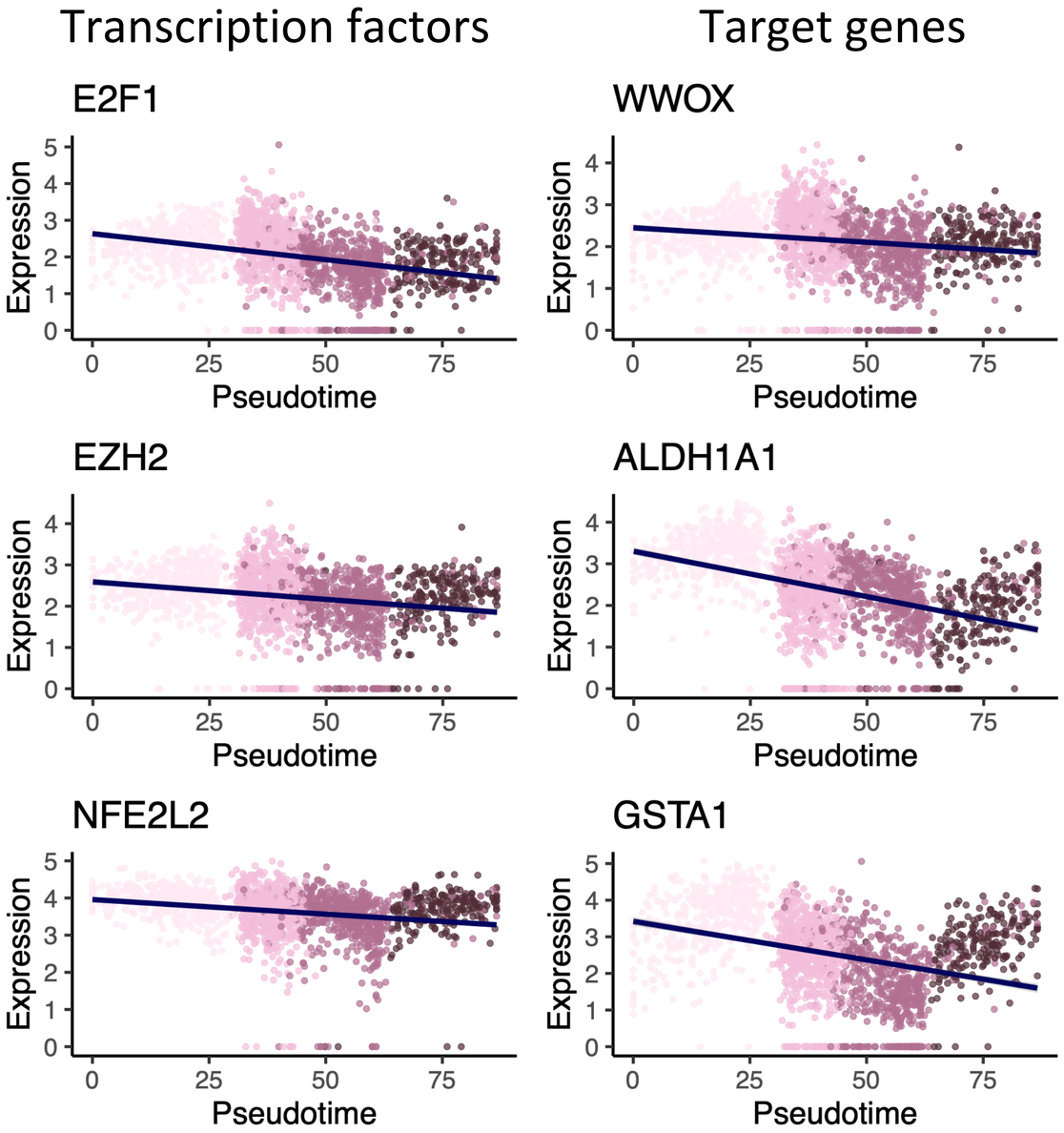


Pseudotime



Many ct-SVGs are Associated with Tumor Trajectory



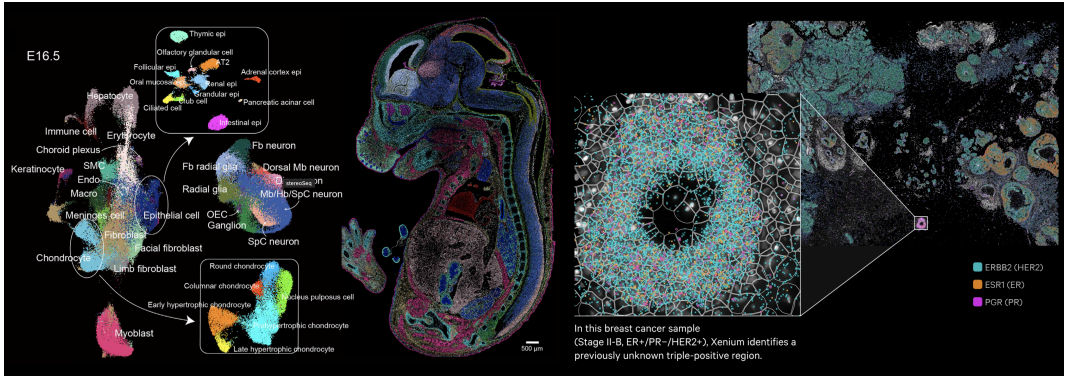


Part II: Mapping Subcellular SVGs

Introduction

Spatial transcriptomics

ST enables transcriptomic profiling on tissues with spatial localization information.

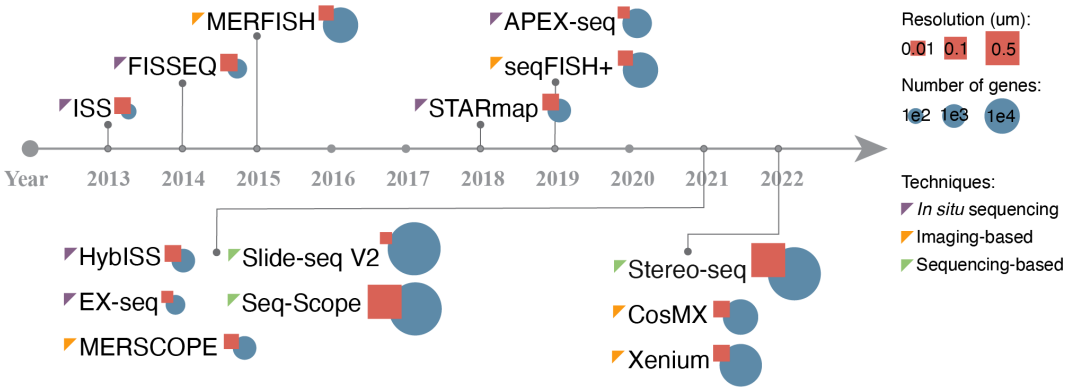


Ref: Stereo-seq (2022), 10xgenomics.com (2024)

Introduction

High-resolution spatial transcriptomics

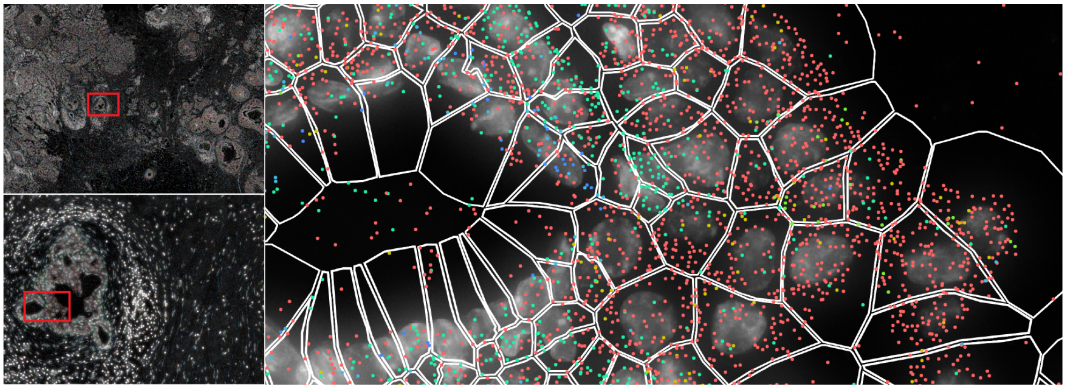
An overview of selected high-resolution ST techniques.



Introduction

High-resolution spatial transcriptomics

High-resolution ST enables precise gene expression measurement at subcellular level.



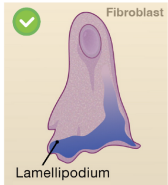
Ref: Xenium.com (2024)

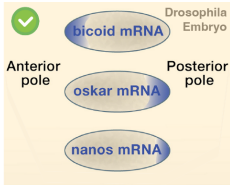
Introduction

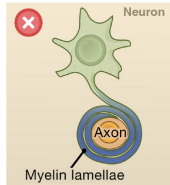
Subcellular localization of mRNAs

Understanding mRNA spatial localization within cells is crucial for unraveling the complexity of cellular structure and function. For example:

- Facilitate the localized protein synthesis – beta actin in fibroblasts.
- Contributes to cellular organization/differentiation – Oskar in drosophila embryo.
- Misplacement can often lead to detrimental effects – Huntington's disease.







Ref: Martin and Ephrussi (2009)

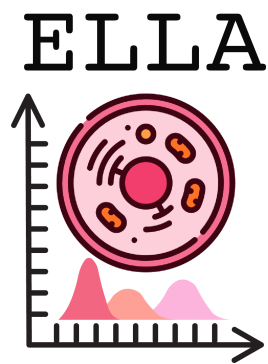
Introduction

Characterizing subcellular mRNA localization

We propose subcellular expression localization analysis (ELLA), a statistical method for modeling the subcellular localization of mRNAs and detecting genes that display spatial variation within cells in high-resolution ST.

The key features of ELLA include:

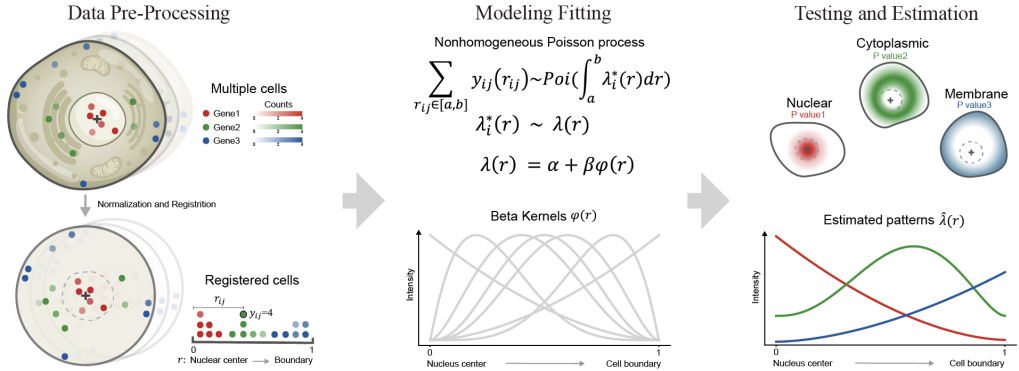
- based on over-dispersed nonhomogeneous Poisson process model;
- estimates various subcellular localization patterns;
- compatible with arbitrary number of cells;
- compatible with diverse ST techniques;
- effective type I errors control and high power;
- scalable to tens of thousands of genes and cells;
- interpretable patterns with pattern-specific mRNA characteristics.



Method

Method overview

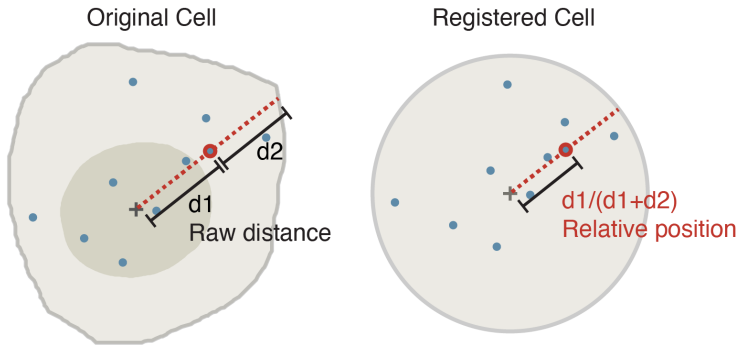
As a method overview, ELLA takes spatial gene expressions, nuclear centers, and cell boundaries as inputs to perform data pre-processing, model fitting, and testing and estimation.



Method

Data pre-processing

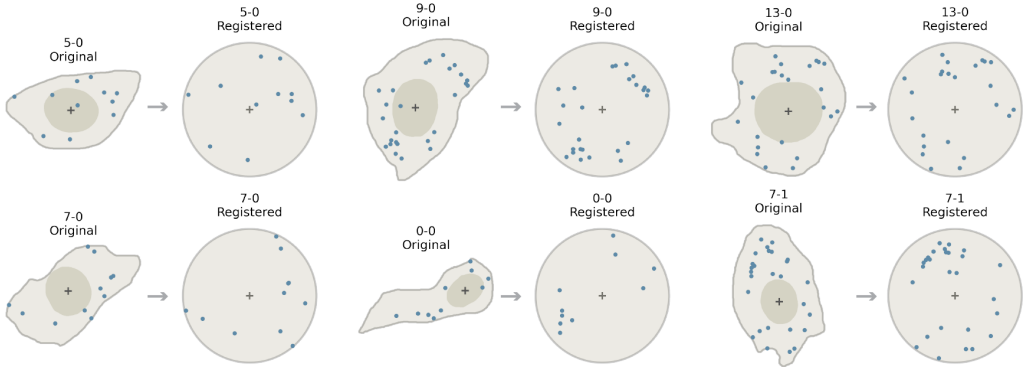
Provided the nuclear center and cell boundary of a cell, the relative positions of the transcripts are calculated.



Method

Data pre-processing

The relative position ranges between 0 and 1 and allows us to create a unified coordinate system across cells, enabling the joint modeling of multiple cells regardless of their sizes and shapes.



Method

Model fitting

We model the mRNA localization within each cell using an over-dispersed one-dimensional nonhomogeneous Poisson process (NHPP) model, which is effectively a tailored Cox Process model. Specifically, we assume that the counts summed cross all relative positions r_{ij} within an interval [a,b] follows a Poisson distribution:

$$\sum_{r_{ij} \in [a,b]} y_{ij}(r_{ij}) \sim Poi(\int_a^b \lambda_i^*(r)\mathrm{d}r),$$

$$\lambda_i^*(r) = c_i s(r) \lambda(r) + \epsilon_i(r).$$

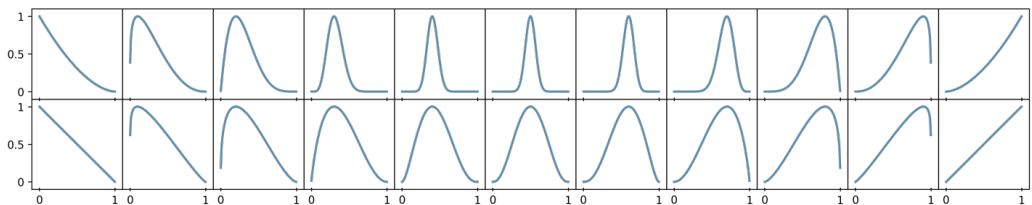
- r_{ij} relative position of location j in cell i;
- y_{ij} mRNA counts corresponding to r_{ij};
- $\lambda_i^*(r)$ NHPP density;
- $\epsilon_i(r)$ follows a normal distribution to model over-dispersion;
- c_i total read depth of cell i;
- $s(r)$ normalizing term $2\pi r$;
- $\lambda(r)$ the subcellular spatial expression intensity function shared across cells.

Method

Model fitting

The unknown intensity function $\lambda(r)$ captures the subcellular spatial expression pattern along the cellular radius. We use k beta kernel functions $\varphi_1, \ldots, \varphi_k$ to capture the intensity function:

$$\lambda(r) = \alpha_l + \beta_l \varphi_l(r),\ l = 1, \ldots, k.$$



We maximize the log likelihood using Adam and obtain the MLE $\hat{\alpha}_l$ and $\hat{\beta}_l$ across kernels.

Method

Testing and estimation

Testing:
Under the proposed NHPP model, identifying genes that display subcellular spatial expression pattern is equivalent to testing whether $\lambda(r)$ is a constant or not. Specifically, for each kernel in turn, we test the null hypothesis H_0: $\beta_l = 0$ using likelihood ratio test and obtain k P values. The k P values are combined using Cauchy combination rule. We control the FDR across genes using the Benjamini-Yekutieli.

$$\{P_1, \ldots, P_k\} \rightarrow P_{\text{combined}} \rightarrow P_{FDR}$$

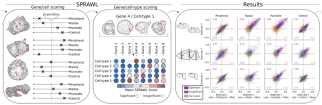
Estimation:
We obtain the k estimated intensity functions across kernels $\hat{\lambda}_l(r) = \hat{\alpha}_l + \hat{\beta}_l \varphi_l(r)$. The $\lambda(r)$ is estimated as the weighted combination in the form of $\hat{\lambda}(r) = \sum_{l=1}^{k} w_l \hat{\lambda}_l(r)$ based on Bayesian model averaging.

$$\{\hat{\lambda}_1(r), \ldots, \hat{\lambda}_k(r)\} \rightarrow \hat{\lambda}(r)$$

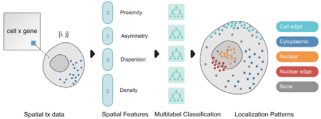
Simulations

Competing methods

SPRAWL (eLife, 2023):



Bento (GB, 2024):



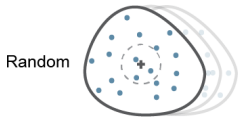
Wilcox (current study):
Nucleus vs cytoplasm

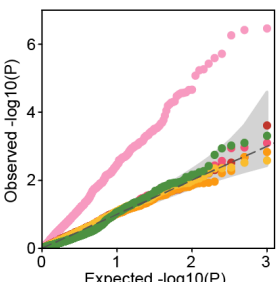
Compatible with	ELLA	SPRAWL	Bento	Wilcox
One cell	✓	✗	✓	✗
Multiple cells	✓	✓	✗	✓
Imaging data	✓	✓	✓	–
Sequencing data	✓	✗	✗	–
P values	✓	–	✗	–
Patterns	Various	4	5	1

Required inputs	ELLA	SPRAWL	Bento	Wilcox
Nuclear center	●			
Nuclear boundary			●	●
Cell centroid		●		
Cell boundary	●	●	●	●

Simulations

Null simulations: compare ELLA with SPRAWL and Wilcox





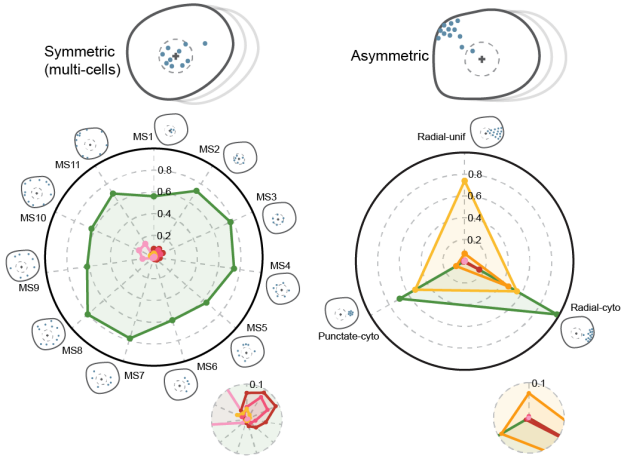
- Sampled n different embryonic fibroblast cells from a seqFISH+ data and simulated expression counts for 1,000 genes to be randomly distributed within these cells.

- Nuclear boundary information was provided to Wilcox.
- P values from ELLA and SPRAWL (of produced) are well calibrated across settings.
- P values from Wilcox are inflated.
- Varying number of cells (n=10-500) and expression level (m=1-100).

ELLA | SPRAWL-peripheral | SPRAWL-central | SPRAWL-radial | SPRAWL-punctate | Wilcox

Simulations

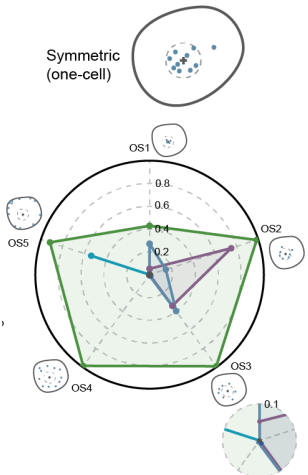
Alternative simulations – multiple cells: compare ELLA with SPRAWL and Wilcox



- Evaluated power based on a fixed false-discovery rate (5%).
- ELLA achieves consistently higher power except for the radial-unif setting.
- ELLA also accurately recovers the true patterns (Appendix).
- Varying number of cells (n=10-500), expression level (m=1-100), and pattern strength (s=0.1-1.0).

Simulations

Alternative simulations – one cell: compare ELLA with Bento

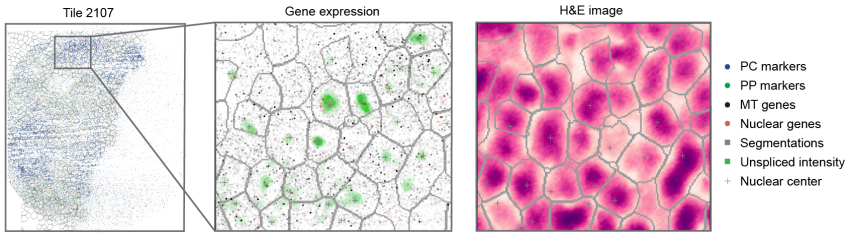


- Five symmetric patterns: gene expression is enriched in nucleus (including 2 patterns), nuclear edge (1), cytoplasm (1), or cellular boundary (1).
- High expression level (m=30) and a high pattern strength (s=9).
- Nuclear boundary information was provided to Bento.
- Because Bento cannot produce P values, we used the prediction probabilities output from Bento to rank genes, with which we measured powers based on FDR.
- ELLA achieves high power and accuracy across all five patterns, consistently outperforming Bento.

Real Data Applications

I Seq-Scope mouse liver data

We first applied ELLA to the (sequencing-based) Seq-Scope mouse liver data.

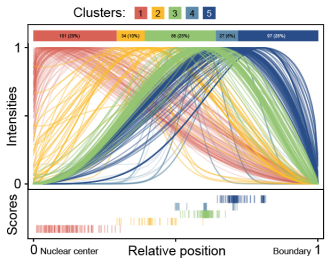


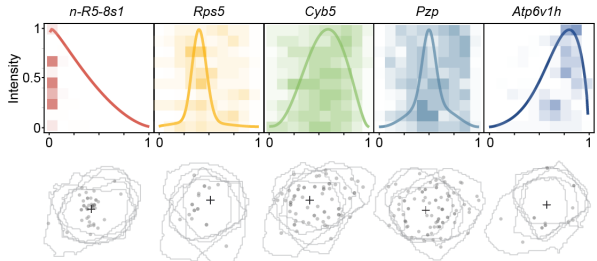
- The data contains ten 1mm-wide circular tiles with a spatial resolution around 0.6 um.
- Cell segmentations were obtained based the H&E images using Cellpose.
- Nuclear centers were identified based on unspliced expression density.
- We applied ELLA to 4 hepatocyte cell types with 497-1,349 genes and 82-276 cells.

Real Data Applications

I Seq-Scope mouse liver data

ELLA identified 345 genes across cell types displaying 5 subcellular expression patterns.



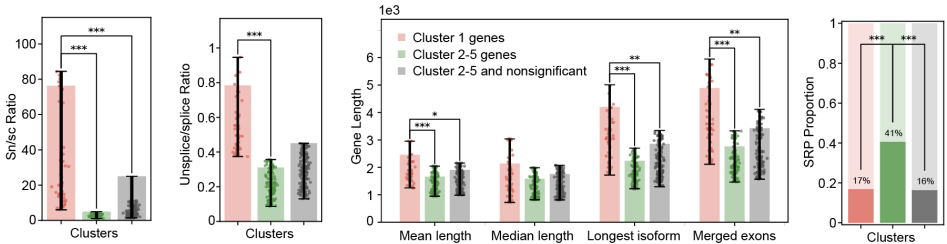


- 101 genes (29%) display a nuclear expression pattern (clusters 1), 34 (10%) genes display a nuclear edge expression pattern (cluster 2), and 210 genes (61%) display one of the three cytoplasmic expression patterns (cluster 3-5).

Real Data Applications

I Seq-Scope mouse liver data

The detected genes under different patterns show different mRNA characteristics.

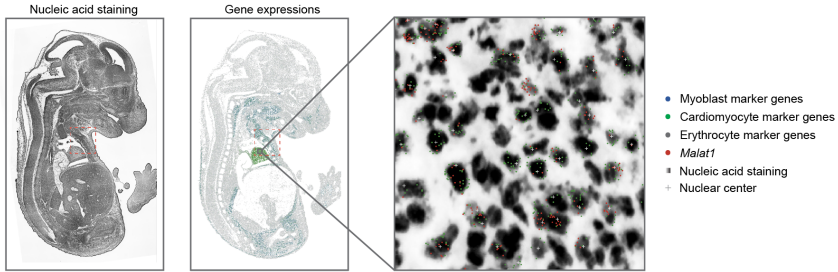


- Nucleus localized (cluster 1) genes have higher snRNA expressions levels.
- Nucleus localized (cluster 1) genes have higher unsplice/splice ratios.
- Nucleus localized (cluster 1) genes have longer gene lengths.
- Cytoplasmic localized genes (clusters 4-5) frequently encode signal recognition peptides (SRPs).

Real Data Applications

II Stereo-seq mouse embryo data

We next applied ELLA to the (sequencing-based) Stereo-seq mouse embryo data E1S3.

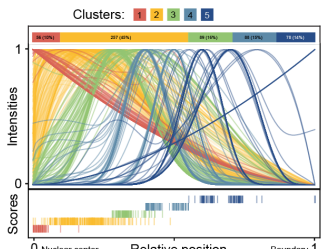


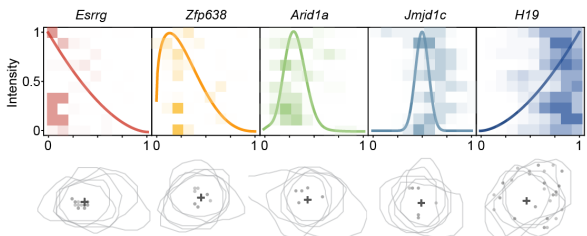
- We focused on two major cell types localized in the cardiothoracic region: myoblasts (596 cells with 2,008 genes) and cardiomyocytes (553 cells with 1,743 genes).
- Cell segmentations were obtained based on nucleic acid staining image using Cellpose.
- Nuclear centers were identified based on unspliced expression density.

Real Data Applications

II Stereo-seq mouse embryo data

ELLA identified 568 genes across cell types displaying 5 subcellular expression patterns.



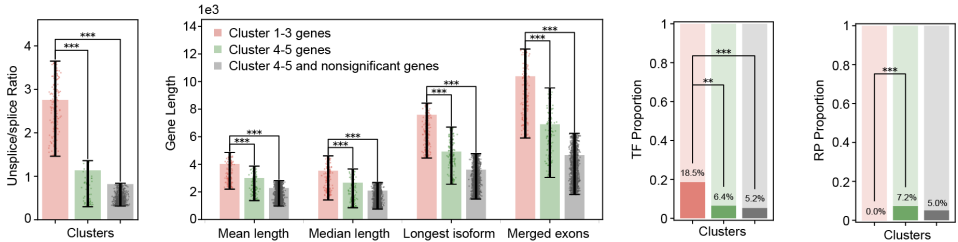


- 56 genes (10%) display a nuclear expression pattern (clusters 1), 346 genes (61%) display one of the two nuclear edge expression patterns (cluster 2-3), and 166 genes (29%) display one of the two cytoplasmic expression patterns (cluster 4-5).

Real Data Applications

II Stereo-seq mouse embryo data

The detected genes under different patterns show different mRNA characteristics.

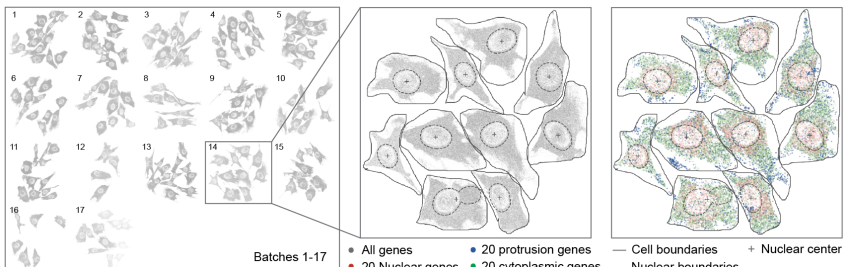


- Cluster 1-3 genes have higher unsplice/splice ratios.
- Cluster 1-3 genes have longer gene lengths.
- Cluster 1-3 genes contain a higher proportion of transcription factors (TFs).
- Cluster 4-5 genes contain a higher proportion of ribosomal protein (RP) genes.

Real Data Applications

III seqFish+ mouse embryonic fibroblast data

We next applied ELLA to the (imaging-based) seqFish+ embryonic fibroblast data.

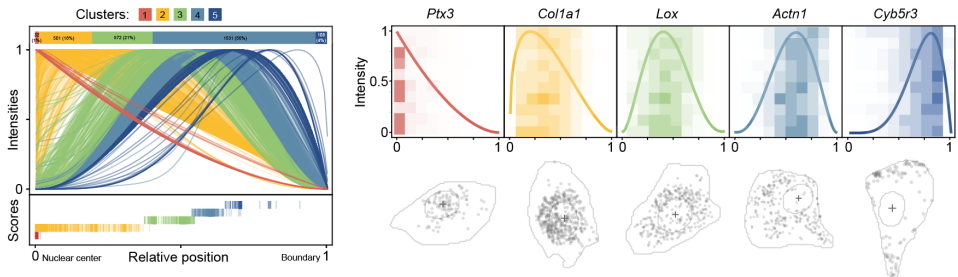


- The data contains 2,747 genes measured on 171 embryonic fibroblast cells.
- Nucleus and cell segmentations are provided and the nuclear center is obtained as the geometric center of all nuclear boundary points.

Real Data Applications

III seqFish+ mouse embryonic fibroblast data

ELLA identified 2,744 genes displaying 5 subcellular expression patterns.

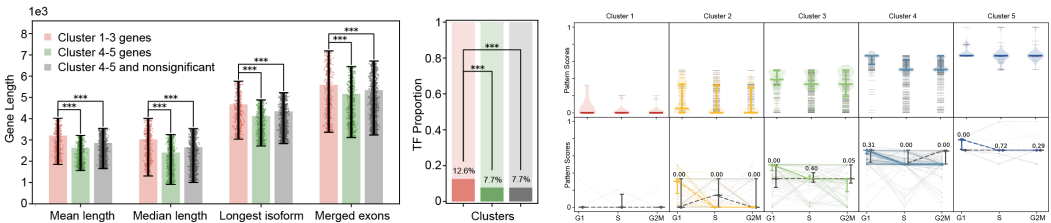


- 32 genes (1%) display a nuclear expression pattern (cluster 1), 1,073 genes (39%) display one of the two nuclear edge expression patterns (clusters 2-3), and 1,639 genes (60%) display one of the two cytoplasmic expression patterns (clusters 4-5).

Real Data Applications

III seqFish+ mouse embryonic fibroblast data

The detected genes with different patterns show different mRNA characteristics.

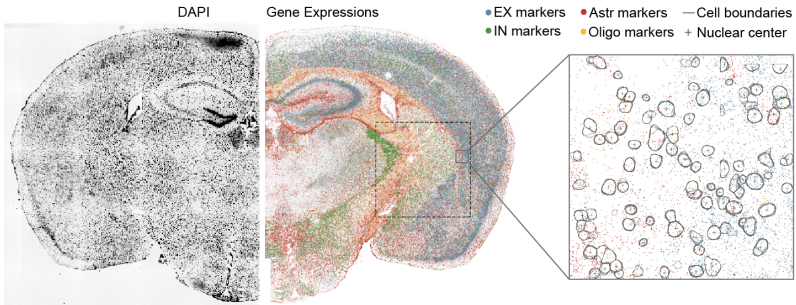


- Cluster 1-3 genes have longer gene lengths.
- Cluster 1-3 genes contain a higher proportion of transcription factors (TFs).
- Genes can exhibit dynamic subcellular localizations during the cell cycle. For example, some genes have decreased nuclear enrichment in G1 phase, while others maintaining their patterns of enrichment regardless of cell cycle phases.

Real Data Applications

IV MERFISH mouse brain data

We last applied ELLA to the (imaging-based) MERFISH adult mouse brain data.

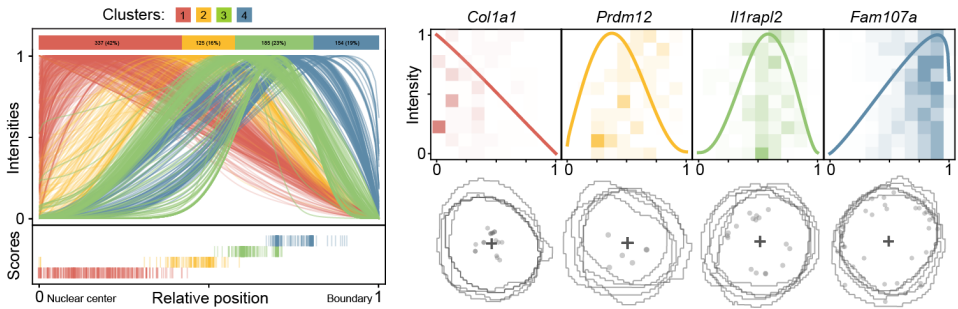


- We focused on 4 major cell types localized in a mid-brain (EX, IN, Astr, Olig) with 557-878 genes and 480-948 cells.
- Cell segmentations are provided by the study.
- Nuclear centers were obtained using the cell centroids from the study.

Real Data Applications

IV MERFISH mouse brain data

ELLA identified 801 genes displaying 4 subcellular expression patterns.

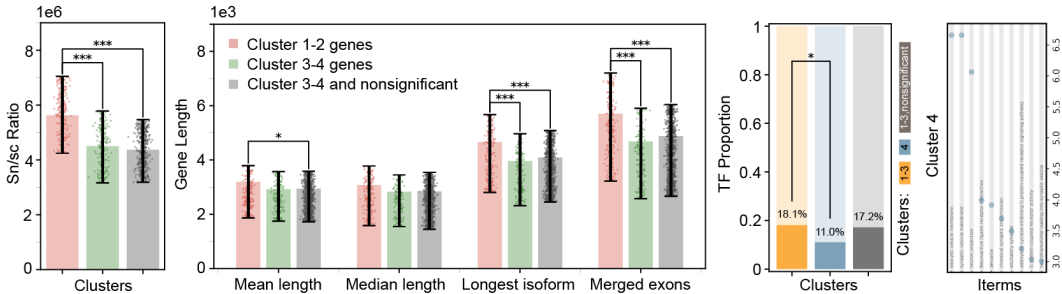


- 337 genes (42%) display a nuclear expression pattern (cluster 1), 125 (16%) genes display a nuclear edge expression pattern (cluster 2), and 339 genes (42%) display one of the two cytoplasmic expression patterns (clusters 3-4).

Real Data Applications

IV MERFISH mouse brain data

The detected genes with different patterns show different mRNA characteristics.



- Cluster 1-2 genes have higher snRNA expression levels.
- Cluster 1-2 genes have longer gene lengths.
- Cluster 4 genes contains a lower portion of transcription factors (TFs).
- Cluster 4 genes are related to dendrites and synaptic transmission and signaling.

Conclusions

- ELLA models and detects spatially variable genes within cells that display various subcellular spatial expression patterns in high-resolution spatial transcriptomics.
- ELLA models the spatial distribution of gene expression measurements along the cellular radius using a nonhomogeneous Poisson process, leverages multiple kernel functions to detect a variety of subcellular spatial expression patterns, and is capable of analyzing a large number of genes and cells.
- ELLA not only identifies genes with distinct subcellular localization patterns but also reveals that these patterns are associated with unique mRNA characteristics.
- Preprint available on bioRxiv (614515)

Summary

- Celina identifies cell type-specific SVGs across a variety of spatial transcriptomics platforms, achieving calibrated type I error control with substantial power gain both single cell and spot resolution spatial transcriptomics.

  -- Lulu Shang*, Peijun Wu*, and Xiang Zhou (2025). Statistical identification of cell type-specific spatially variable genes in spatial transcriptomics. *Nature Communications*. 16: 1059.

- ELLA models and detects spatially variable genes within cells that display various subcellular spatial expression patterns in high-resolution spatial transcriptomics, revealing unique mRNA characteristics that underlie these patterns.

  -- Jade Xiaoqing Wang, and Xiang Zhou (2025). ELLA: Modeling subcellular spatial variation of gene expression within cells in high-resolution spatial transcriptomics. *Nature Communications*. in press.

- Both Celina and ELLA are available on our lab website: https://xiangzhou.github.io/software

Lulu Shang (MD Anderson)
Peijun Wu
Jade Wang (Texas A&M)

R01HG009124
R01GM126553
R01HG011883
R01GM144960

Xiang Zhou Lab