

DahShu Short Course 2024, East Lansing, MI

Identify Spatially Variable Genes using BSP and scBSP

Juexin Wang

Department of BioHealth Informatics

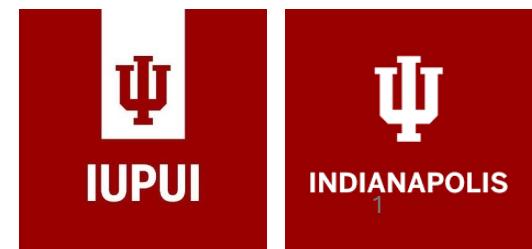
Luddy School of Informatics, computing, and engineering

Indiana University Purdue University Indianapolis

05/16/2024

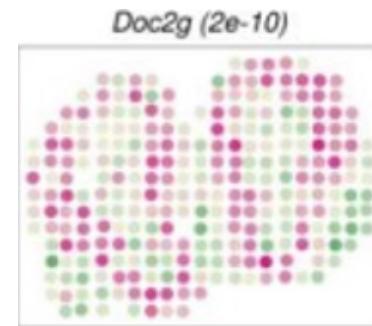
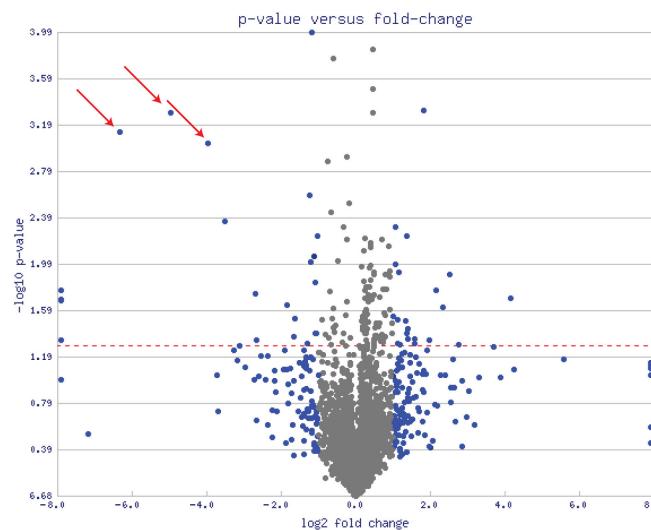
Tutorial:

https://github.com/juexinxwang/Tutorial_DahShu2024



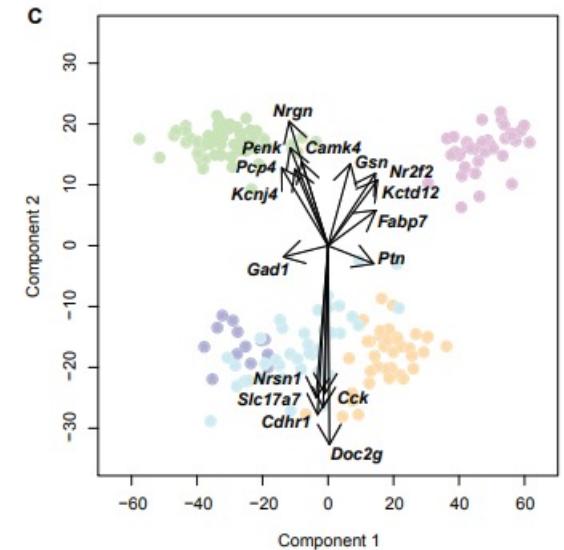
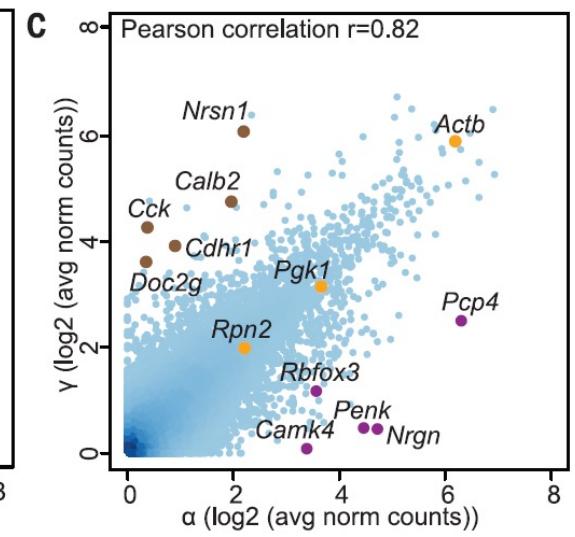
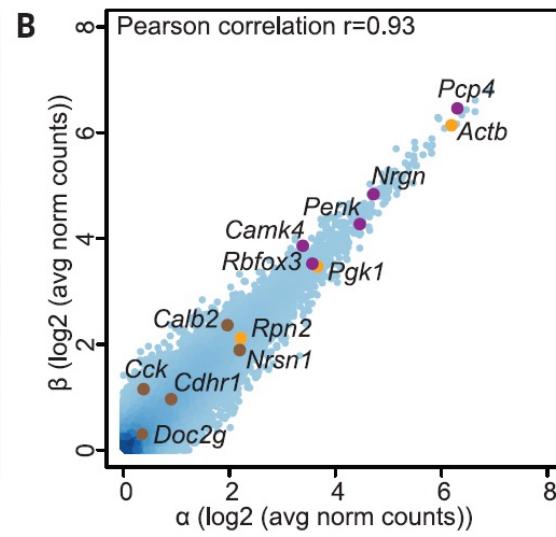
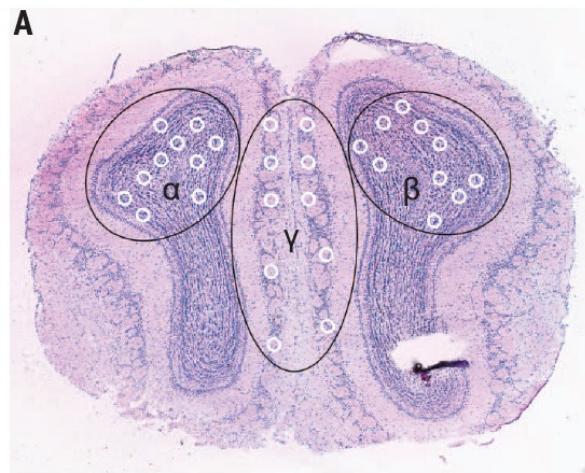
Spatial Variable Genes (SVGs) (spatial context of DEG)

- Differential Expressed Gene -> Spatial Variable Gene
- **Spatial Variable Gene:** Gene differential expressed because of spatial (condition as DEG)
- SVGs are defined as genes with a **highly spatially correlated pattern of expression**, which varies along with the spatial distribution of a tissue structure of interest.
- Spatial expression variation can reflect communication between adjacent cells, position-specific states, or cells that migrate to specific tissue locations to perform their functions.



Earlier days: Ad-hoc methods

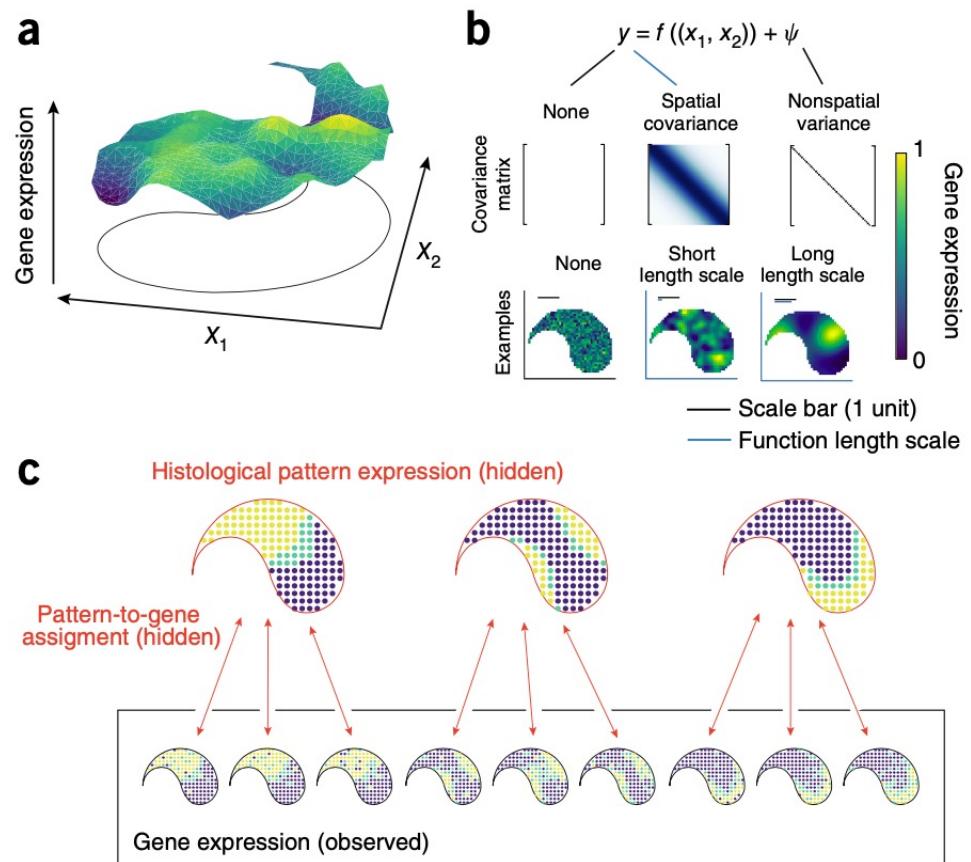
- Union of both PCA and domain correlation
- Ståhl, Patrik L., et al. "Visualization and analysis of gene expression in tissue sections by spatial transcriptomics." *Science* 353.6294 (2016): 78-82.



SpatialDE

- Gaussian process regression; decomposes expression variability into spatial and nonspatial components
- SpatialDE defines spatial dependence for a given gene by using a nonparametric regression model, **testing whether gene expression levels at different locations covary in a manner that depends on their relative location**, and thus are spatially variable.

Svensson, Valentine, Sarah A. Teichmann, and Oliver Stegle. "SpatialDE: identification of spatially variable genes." *Nature methods* 15, no. 5 (2018): 343-346.



Trendsseek

- **Tests for significant dependency between the spatial distributions** of points and their associated marks (**expression levels**) through pairwise analyses of points as a function of the distance r (radius) between them.
- Permutation test

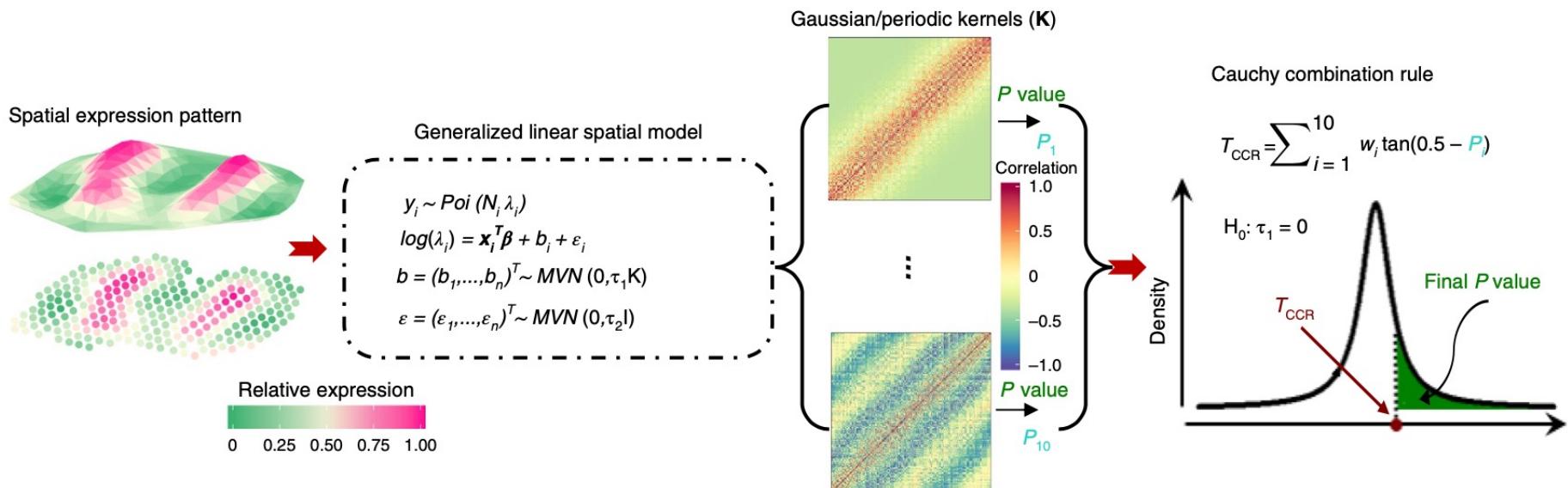
$$M_2(m_1, m_2 | r) = \frac{f_2(m_1, m_2, r)}{f_2(r)}$$

$$M_2(m_1, m_2 | r) \neq M_1(m_1)M_1(m_2)$$

Edsgård, Daniel, Per Johnsson, and Rickard Sandberg.
"Identification of spatial expression trends in single-cell
gene expression data." *Nature methods* 15, no. 5
(2018): 339-342.

Spark

- **Generalized linear spatial model** with a variety of spatial kernels. Chi-square as the exact test statistics distribution; **Cauchy combination rule to combine across multiple spatial kernels**

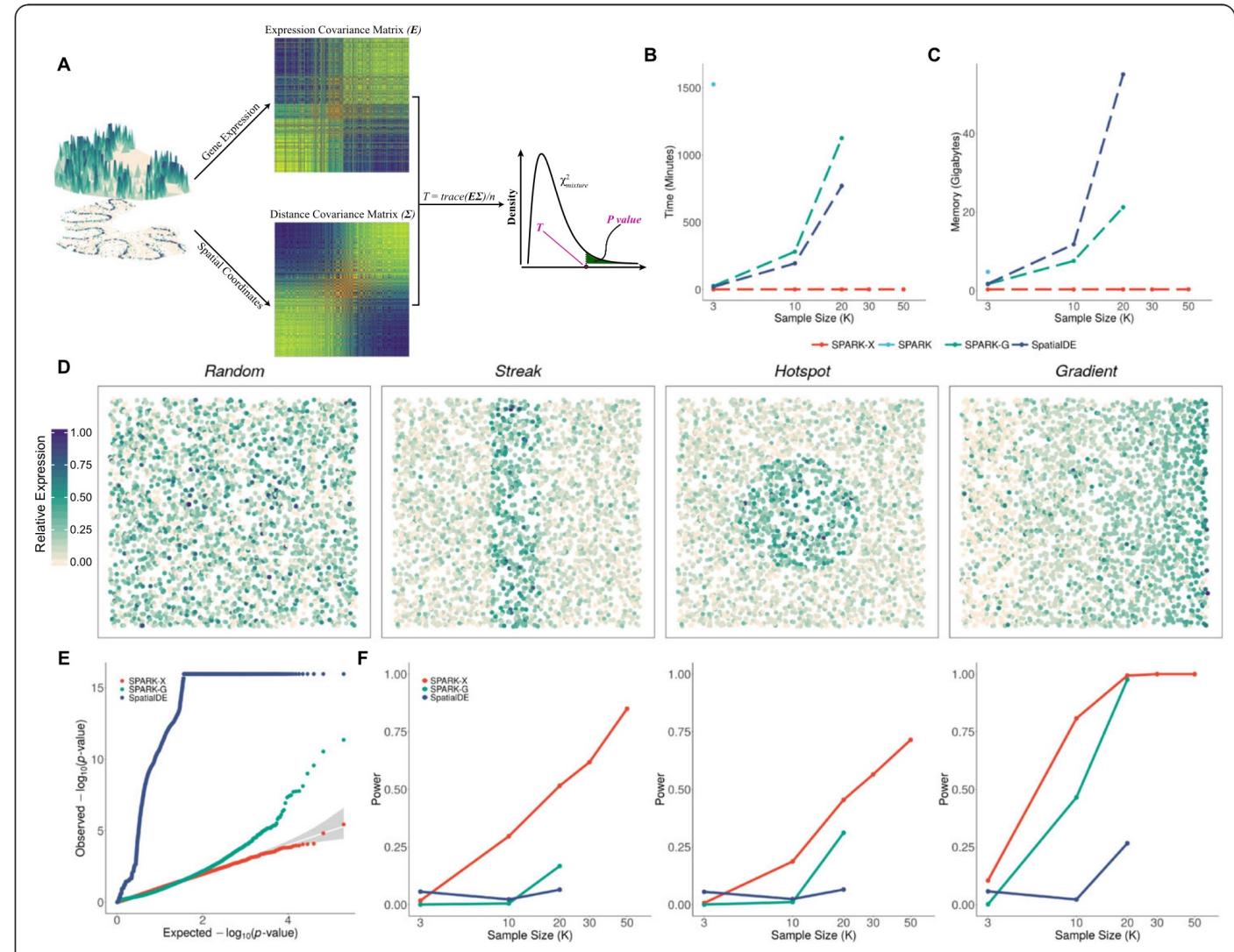


- Sun, Shiquan, Jiaqiang Zhu, and Xiang Zhou. "Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies." *Nature methods* 17, no. 2 (2020): 193-200.

Spark-X

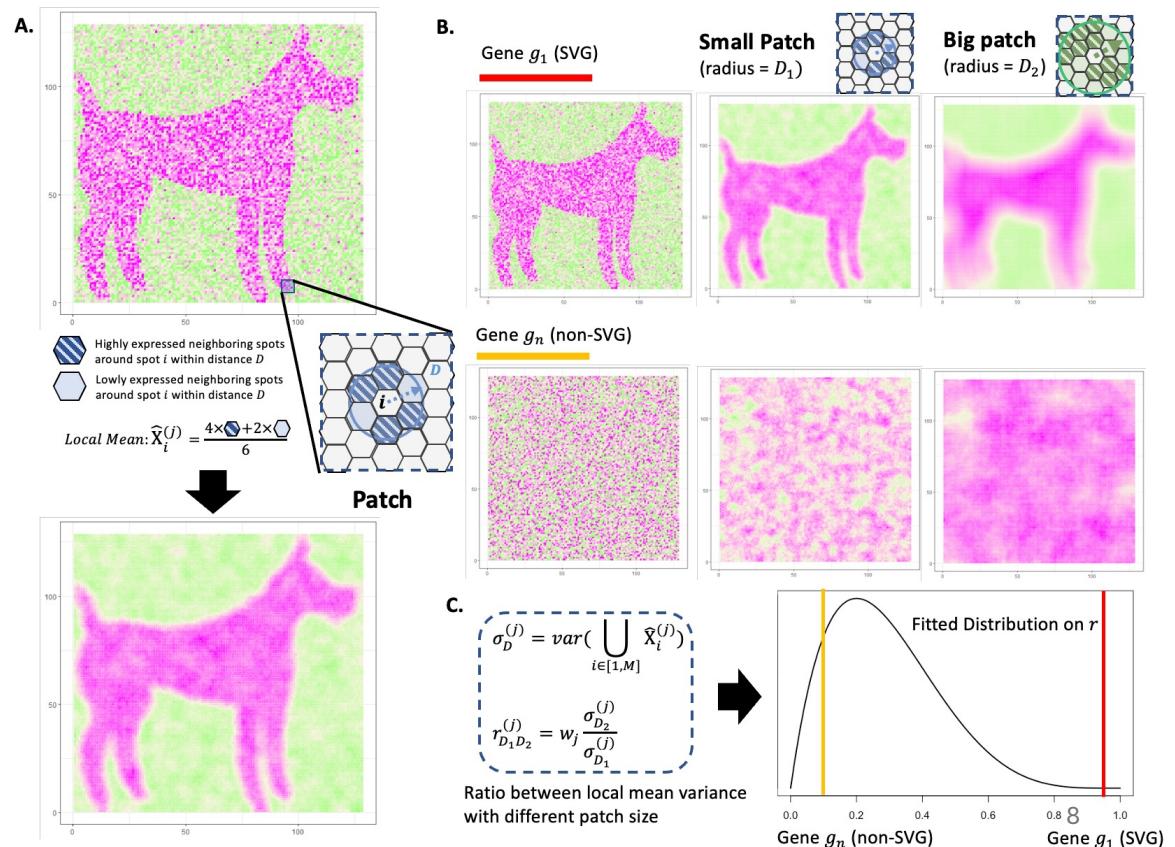
- Non-parametric, check dependence between expression and spatial.
- Simple Statistics
- Designing for Large data

Zhu, Jiaqiang, Shiquan Sun, and Xiang Zhou. "SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies." *Genome Biology* 22, no. 1 (2021): 1-25.



BSP: Dimension-agnostic and granularity-based spatially variable gene identification

- BSP (big-small patch), a non-parametric model by comparing gene expression patterns at two spatial granularities to identify SVGs from two or three-dimensional spatial transcriptomics data in a fast and robust manner.
- **Velocity of change in variances of local means in different granularity can be used to distinguish SVGs and non-SVGs**
- Data driven, model free, any dimension
- Highlighted by Editor's choices

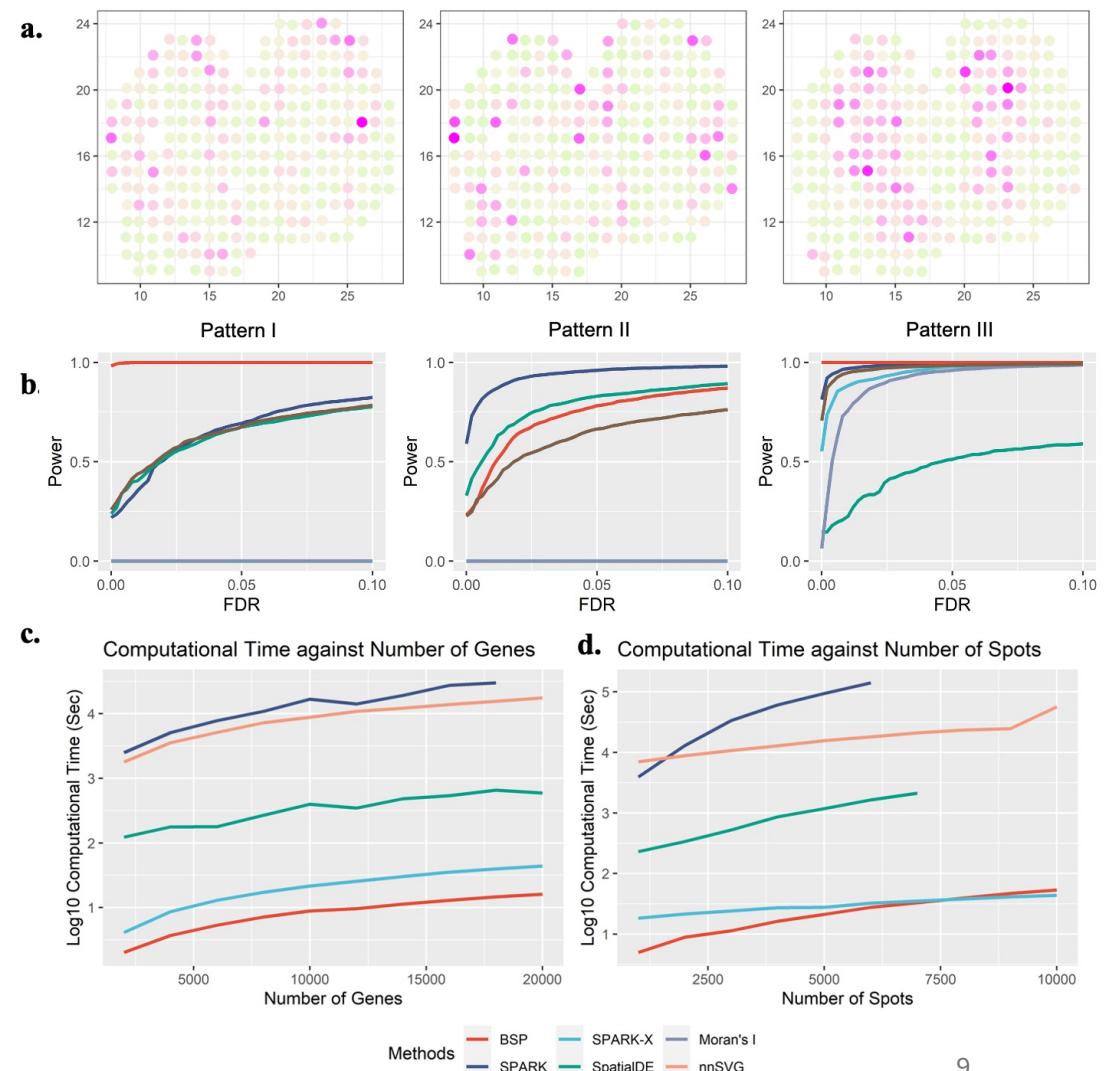


Wang, J., Li, J., Kramer, S.T. et al. Dimension-agnostic and granularity-based spatially variable gene identification using BSP. *Nat Commun* **14**, 7367 (2023).

2D simulations

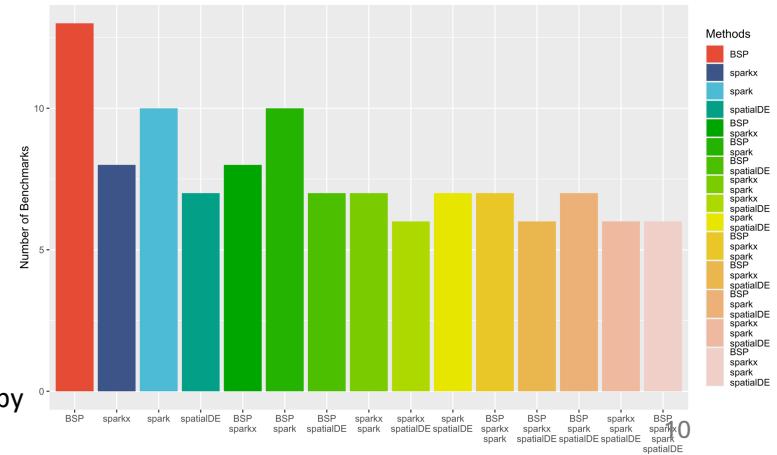
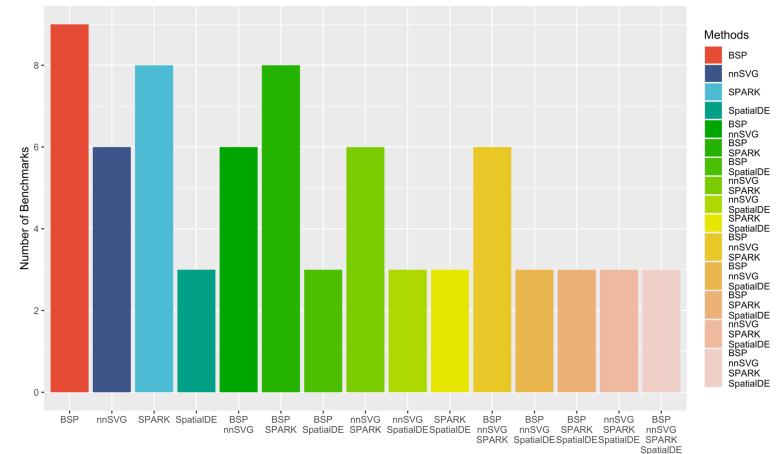
- Better performances on most scenarios varying spatial patterns, size of the pattern, signal strength, and noise level
- Fast and accurate

Mouse olfactory bulb (simulation)



2D Real datasets

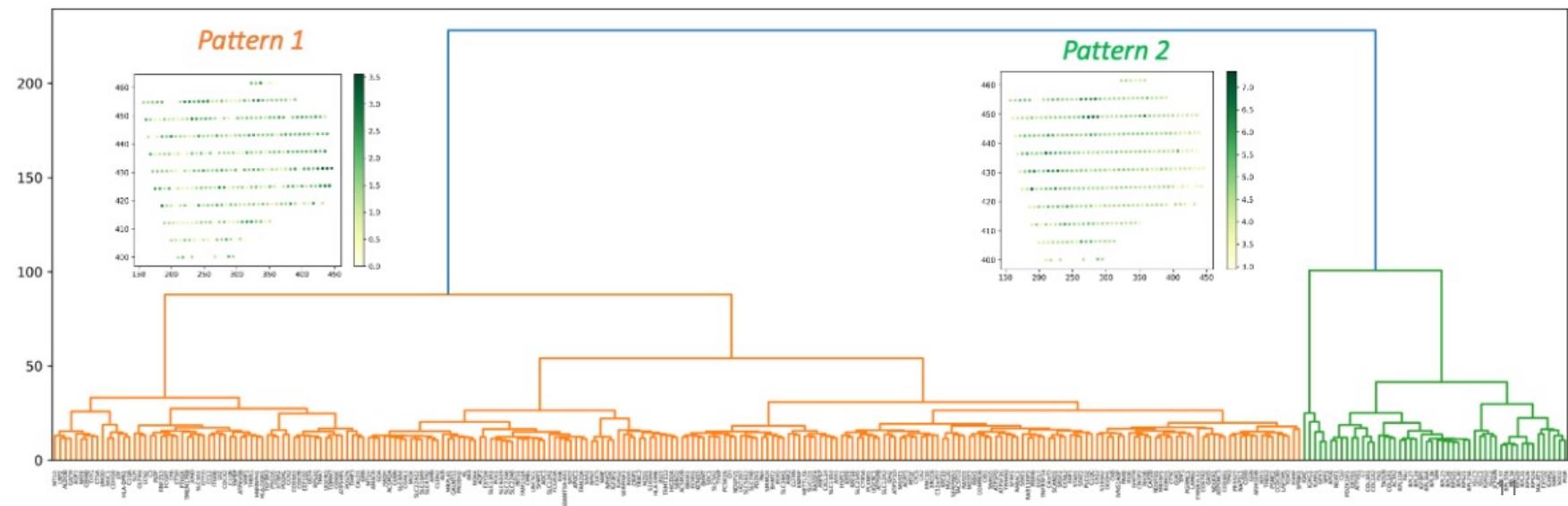
- Mouse factory bulb study
 - ST platform
 - 10 marker genes
 - BSP detected 9
 - SpatialDE detected 3, SPARK detected 8, nnSVG detected 6, and SPARK-X detected 0
- Human breast cancer
 - ST platform
 - 14 marker genes
 - BSP detected 13
 - SpatialDE detected 7, SPARK detected 10, and SPARK-X detected 8



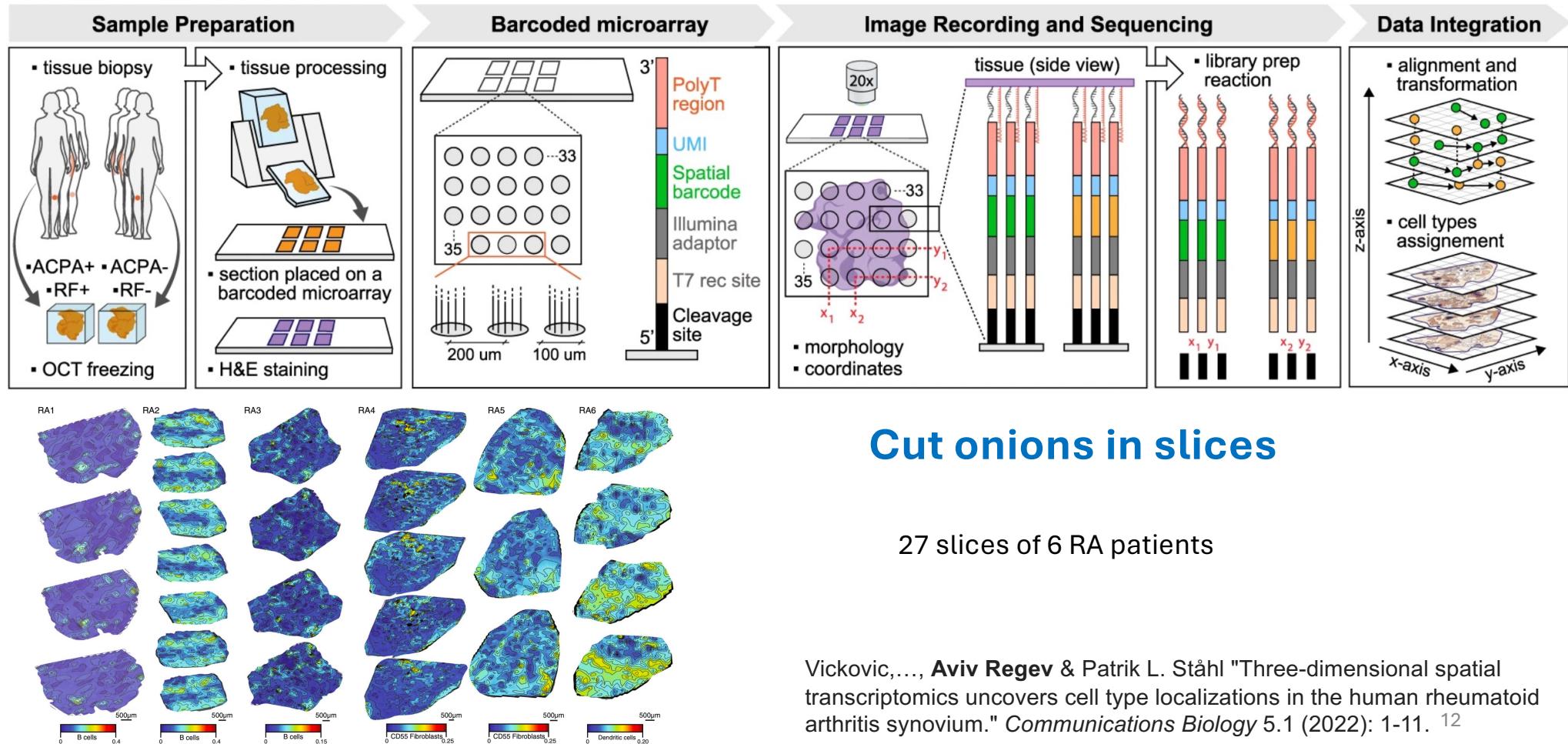
Stahl, P. L. et al. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science* **353**, 78-82 (2016).

2D case study: Acute Kidney Injury (10X Visum)

- Pattern 1
 - Aerobic respiration (q-value 1.04e-09) and oxidative phosphorylation (q-value 4.69e-09), consistent with hypoxic or nephrotoxic acute tubular necrosis, the most common tubulointerstitial form of AKI.
 - Pathways of early recovery were enriched including **kidney development** (q-value 8.56e-07) and **metanephric nephron epithelium development** (q-value 2.30e-06), which included genes like PAX8.
- Pattern 2
 - Humoral immune response (q-value 5.15e-05) and tissue homeostasis (q-value 5.15e-05).
 - Several **immune responses** were activated (q-value 3.51e-04) including B cell receptor signaling pathway (q-value 4.72e-07).



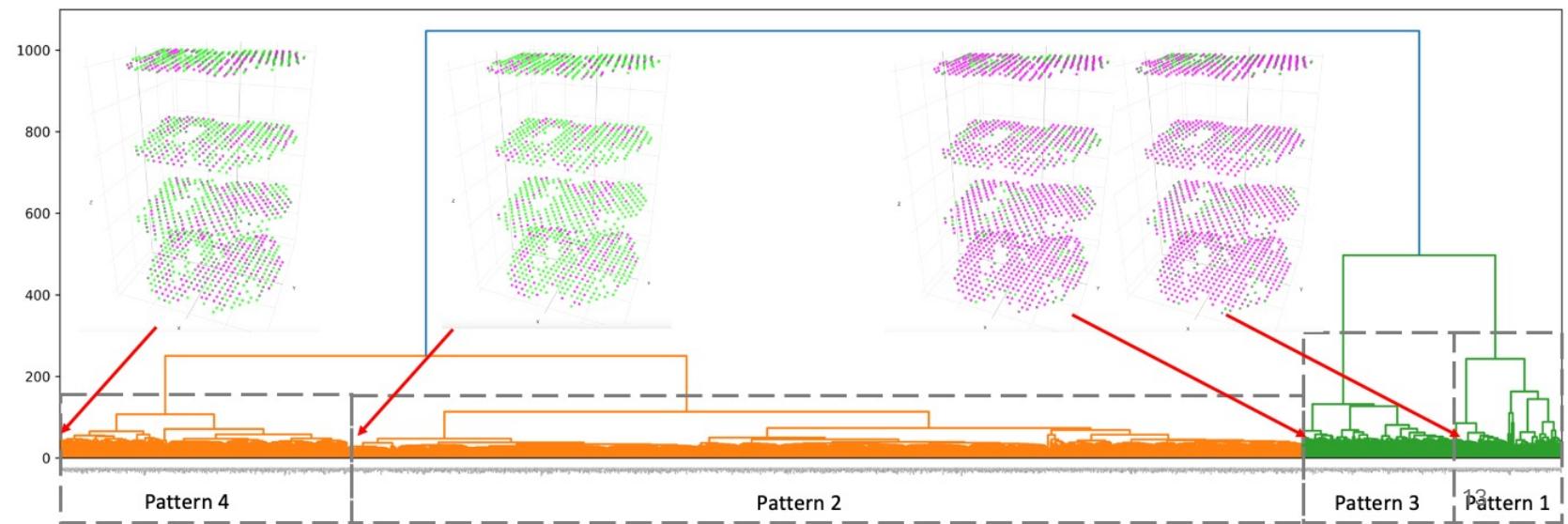
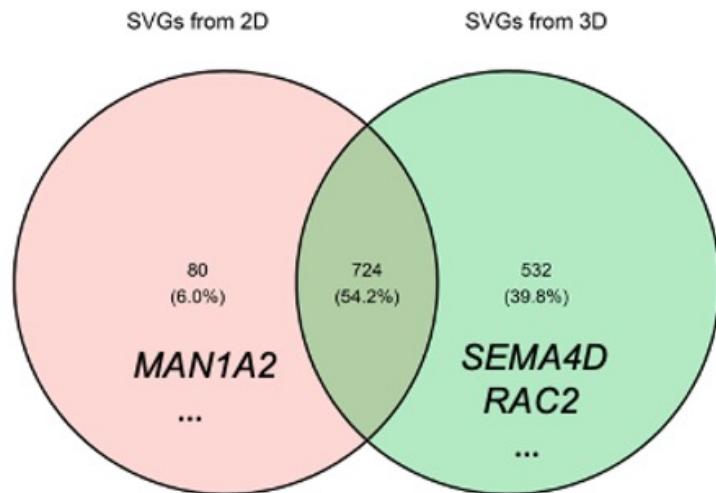
3D spatial transcriptomics is on the way



3D case study: Rheumatoid Arthritis

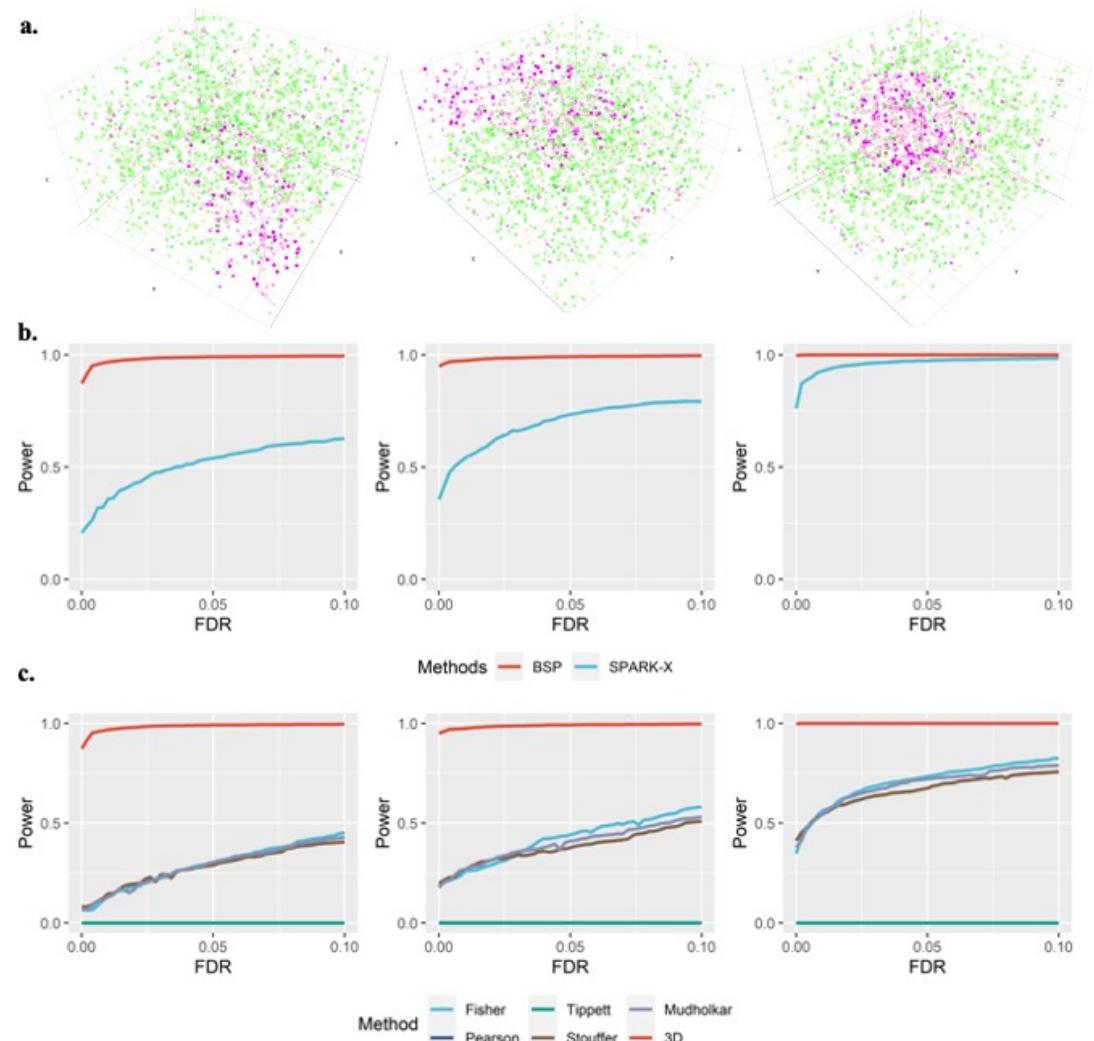
- Identified more biological meaningful genes than 2D meta-analysis

Vickovic, S. et al. Three-dimensional spatial transcriptomics uncovers cell type localizations in the human rheumatoid arthritis synovium.
Commun Biol 5, 129 (2022)

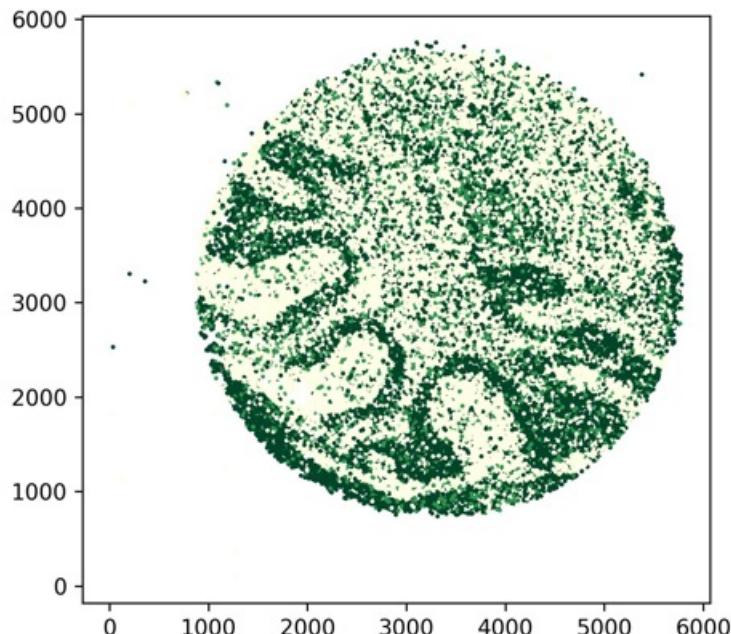


3D simulation

- Three 3D patterns:
- curved stick (Pattern I), thin plate (Pattern II), and irregular lump (Pattern III)
- Power analysis varying pattern size, signal strength, and noise.
- 3D outperforms meta-analysis on 2D slides

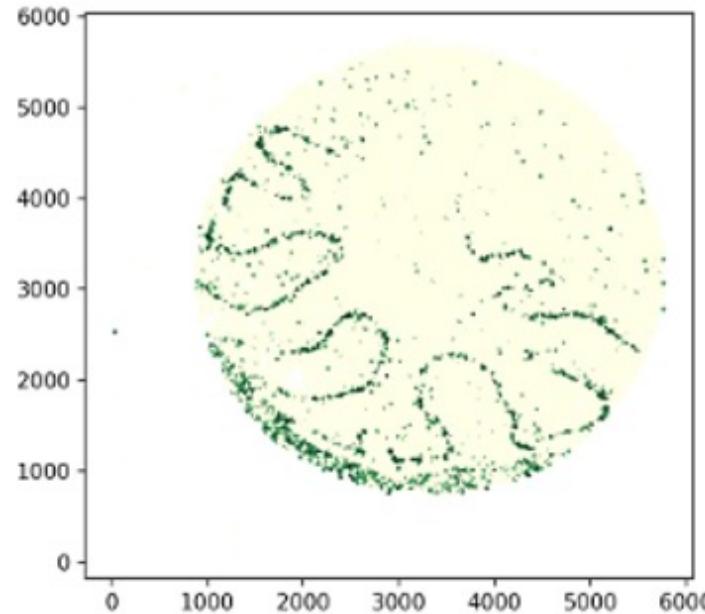


Challenges from large scale data



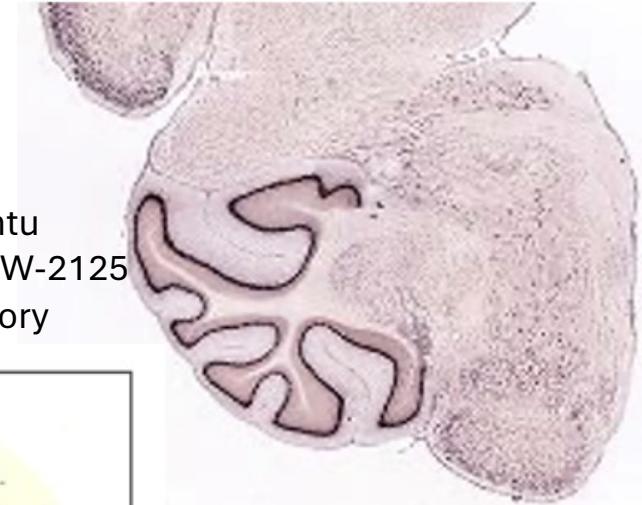
Malat1 (left) and Calb1 (right) as SVG identified by BSP

23,096 genes and 39,496 beads
Running time: **18 min** on an Ubuntu workstation with Intel(R) Xeon(R) W-2125 CPU @ 4.00GHz and 32 GB memory



ISH of *Calb1* in an adult mouse brain from a separate study

2D slide-seq v2



scBSP (BSP 2.0): Improve computational efficiency on BSP

- Using sparse matrix operation to accelerate the computing.
- Targeting subcellular Xenium/CosMx/MERSCOPE data

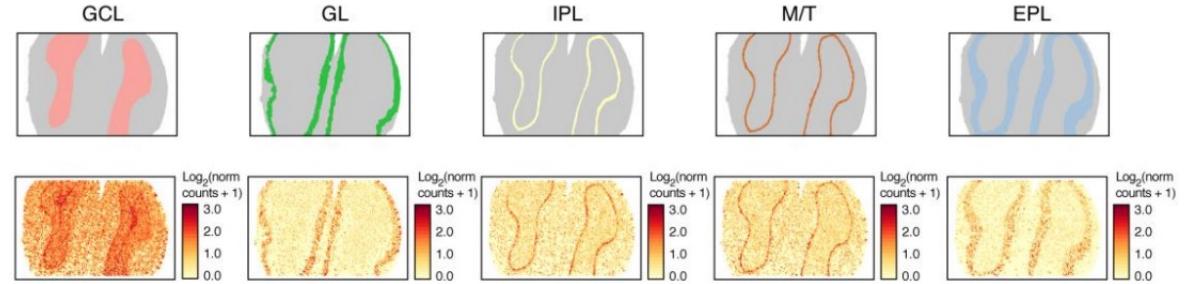
Process	BSP			scBSP		
	Method	Time Complexity	Space Complexity	Method	Time Complexity	Space Complexity
Data Pre-processing	Min-max Scaling	$O(NM)$	$O(NM)$	Maximum Absolute Scaling	$O(M)$	$O(NM)$
Patch Determination	Exhaustive Search	$O(NM^2)$	$O(NM)$	Approximate Nearest Neighbor Search*	$O(NM \log M)$	$O(NM)$
Expression Matrix Calculation	Dense Matrix	$O(M)$	$O(NM)$	Sparse Matrix	$O(SM)$	$O(SNM)$
Local Expression Calculation	Loop Operation	$O(NM)$	$O(NM)$	Matrix Operation	$O(SNM)$	$O(SNM)$

* The Approximate Nearest Neighbor Search in Python is based on HNSW implementation.

Table 1: Time and space complexity of BSP and scBSP. M: number of spots; N: number of genes; S: data sparsity (proportion of non-zero values in the expression matrix).

Li, Jinpu, et al. "scBSP: A fast and accurate tool for identifying spatially variable genes from spatial transcriptomic data." *bioRxiv* (2024): 2024-05.

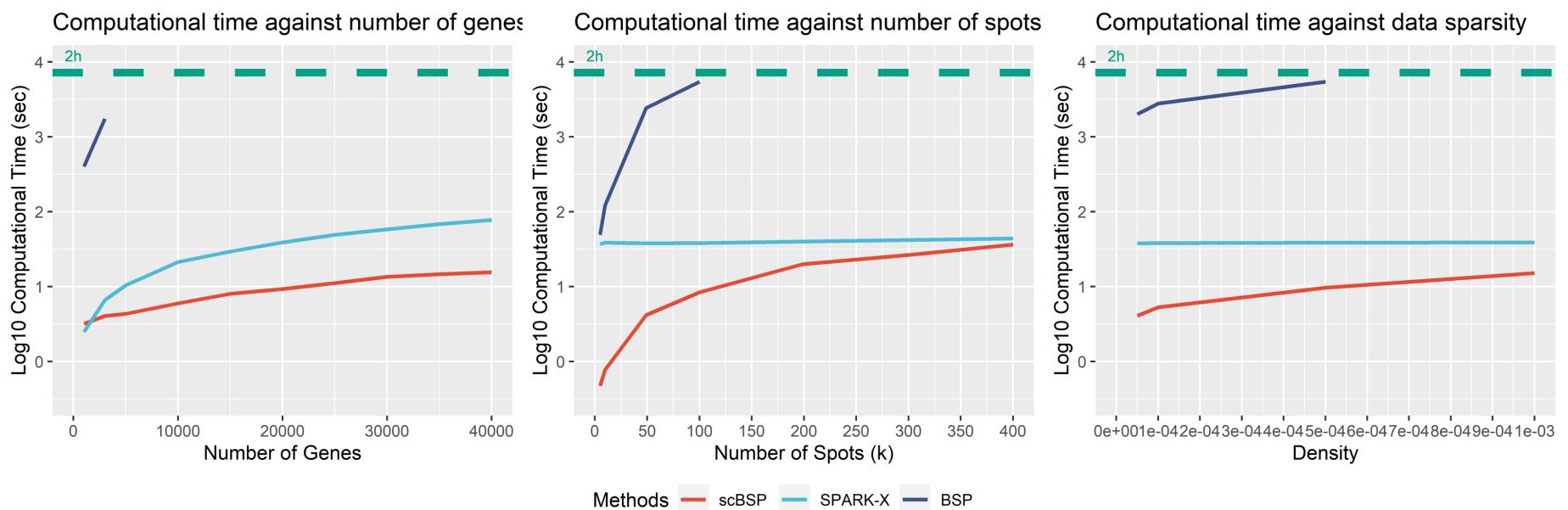
Performance on a laptop



- HDST data:
- ~200,000 spots, 20,000 genes, 0.0003 density
- BSP: 4 hours, 90GB
- scBSP: ~3 sec, 2GB

Vickovic, S., Eraslan, G., Salmén, F. et al. High-definition spatial transcriptomics for *in situ* tissue profiling. *Nat Methods* **16**, 987–990 (2019).

Results – running time



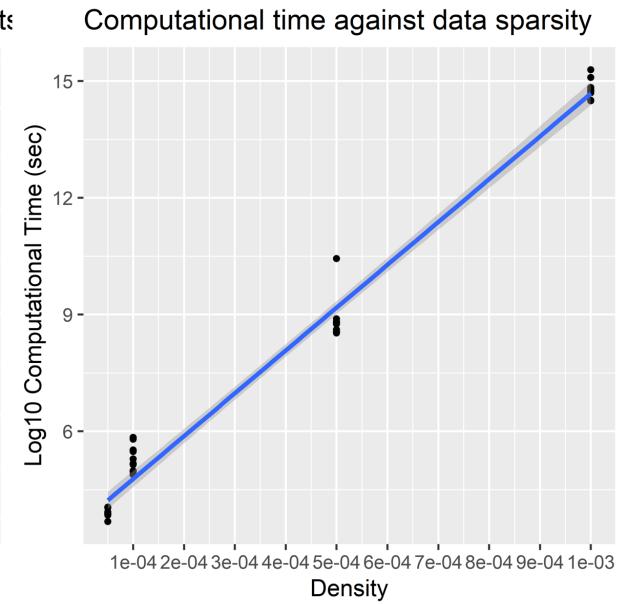
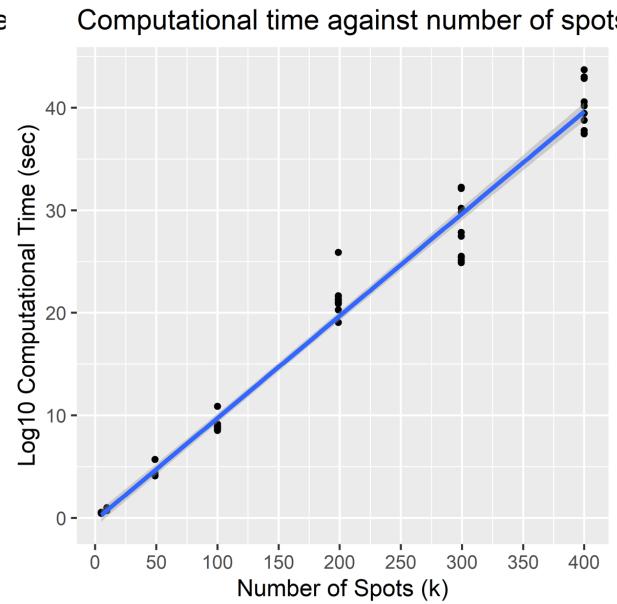
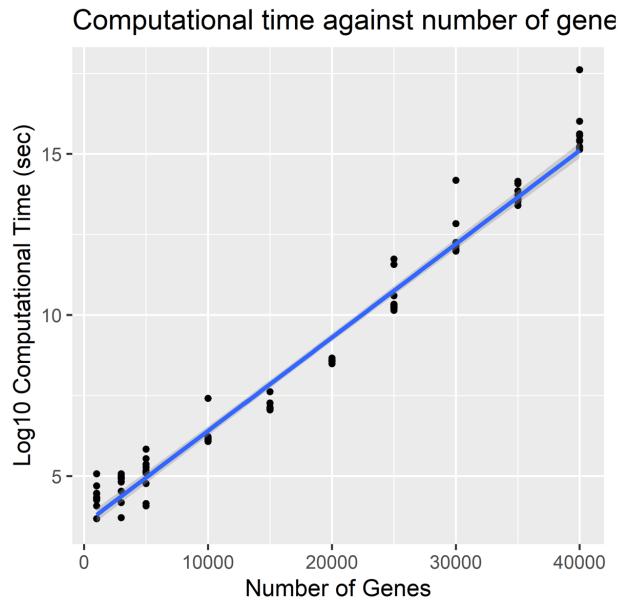
Left: Computational time with an increasing number of genes and fixed 100,000 spots and density of 0.0005.

Middle: Computational time with an increasing number of spots and fixed 20,000 genes and density of 0.0005.

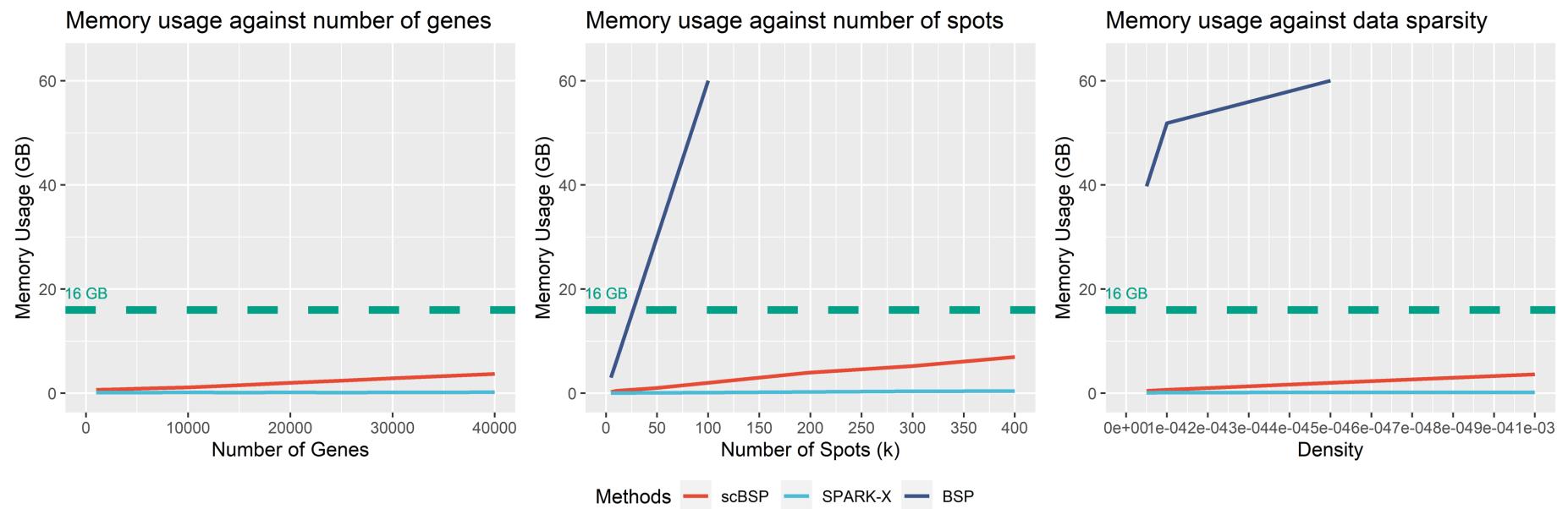
Right: Computational time with an increasing density and fixed 100,000 spots and 20,000 genes.

Results – running time

Computational time scales linearly with number of genes, number of spots, and data density



Results – memory usage



Required RAM memory < 10 GB for most data

Acknowledgement

Lab Member

Mauminah Raina
Shuang Wang

ML and Data analysis

Dr. Dong Xu at University of Missouri

Jinpu Li
Yang Yu
Dr. Fei He
Skyler Kramer
Li Su
...

Dr. Qin Ma at OSU
Dr. Yuzhou Chang
Yi Jiang
...

Dr. Anjun Ma at OSU

Kidney Study

Dr. Michael Eadon at IU SOM
Dr. Krzysztof Kiryluk at Columbia

Tutorial: https://github.com/juexinwang/Tutorial_DahShu2024

R: <https://cran.r-project.org/web/packages/scBSP/index.html>

Python: <https://pypi.org/project/scbsp/>

Funding: NIH/NIDDK R01DK138504

