

Unsupervised statistical methods and data-driven analysis workflows for spatially-resolved transcriptomics

Lukas M. Weber

Department of Biostatistics
Johns Hopkins Bloomberg School of Public Health

18 May 2023

2023 Emerging Leaders in Computational Oncology Symposium
Memorial Sloan Kettering Cancer Center

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

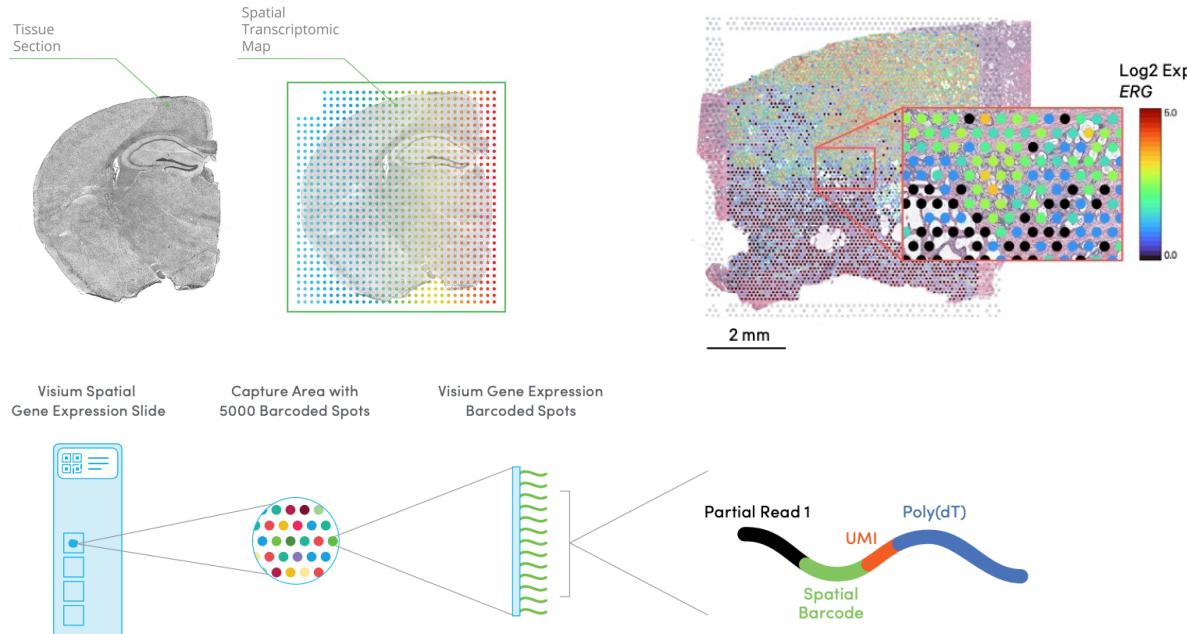
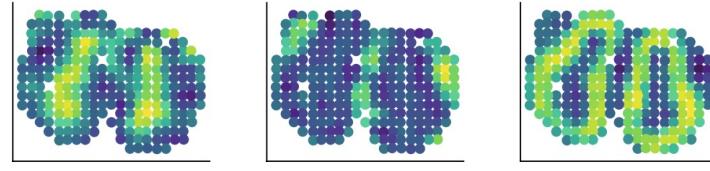


Image source: 10x Genomics

Spatially-resolved transcriptomics

Unsupervised / discovery-based analyses

- feature selection: spatially variable genes
- clustering: spatial domains or spatially distributed cell populations
- differential gene expression

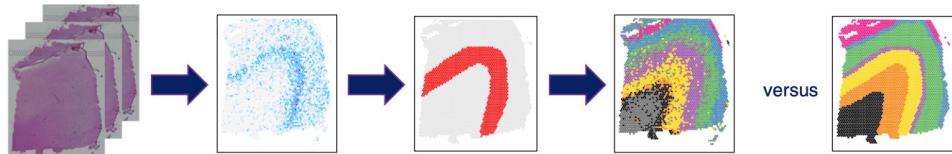


Svensson et al. (2018)

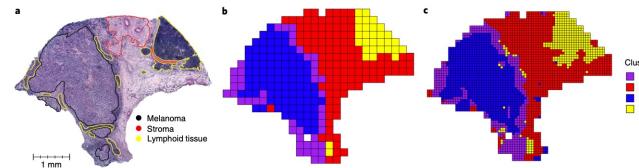
Exploratory analyses

Cell type / state heterogeneity

Biologically informative / marker genes



Maynard and Collado-Torres et al. (2021)



Zhao et al. (2021)

Spatially variable genes

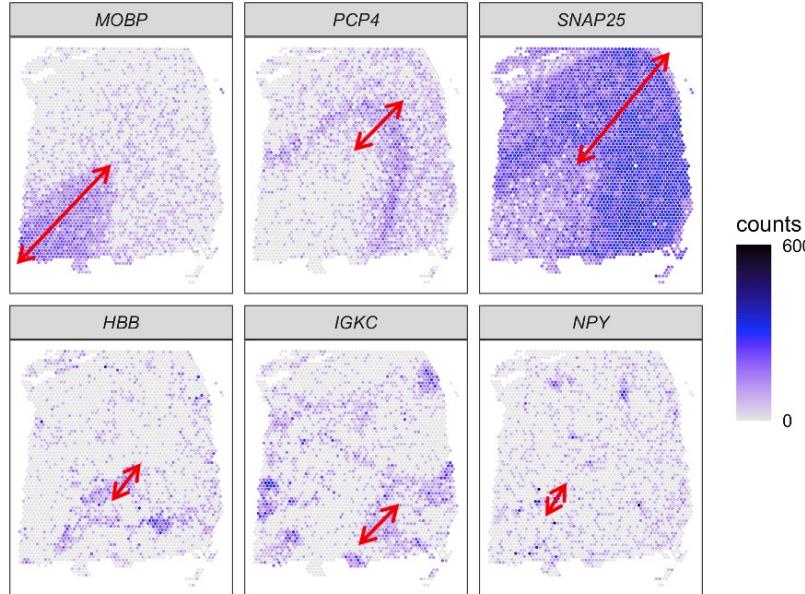
Example

- Dorsolateral prefrontal cortex (DLPFC) in postmortem human brain samples measured with 10x Genomics Visium platform
- Maynard and Collado-Torres et al. (2021)



Kristen R. Maynard^{1,2,10}, Leonardo Collado-Torres^{1,3,10}, Lukas M. Weber⁴, Cedric Uytingco⁵,
Brianna K. Barry^{6,16}, Stephen R. Williams⁵, Joseph L. Catallini II⁴, Matthew N. Tran^{6,17},
Zachary Besich^{1,7}, Madhavi Tippani¹, Jennifer Chew⁵, Yifeng Yin⁵, Joel E. Kleinman^{1,2},
Thomas M. Hyde^{1,2,8}, Nikhil Rao⁵, Stephanie C. Hicks¹⁰, Keri Martinowich^{1,2,6} and
Andrew E. Jaffe^{1,2,3,4,6,7,9}

Selected SVGs: human DLPFC



Spatially variable genes

Example

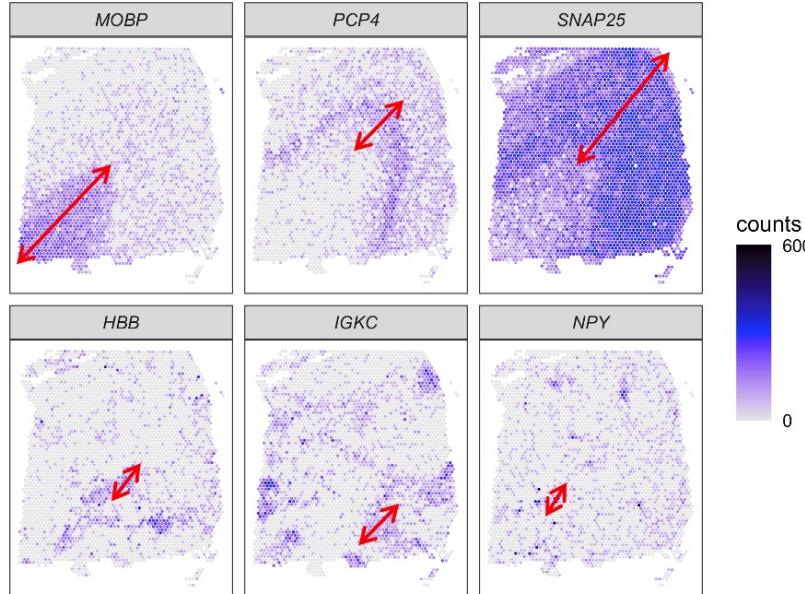
- Dorsolateral prefrontal cortex (DLPFC) in postmortem human brain samples measured with 10x Genomics Visium platform
- Maynard and Collado-Torres et al. (2021)



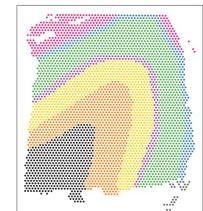
Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex

Kristen R. Maynard^{1,2,10}, Leonardo Collado-Torres^{1,3,10}, Lukas M. Weber⁴, Cedric Uytingco⁵, Brianna K. Barry^{4,6}, Stephen R. Williams⁵, Joseph L. Catallini II⁴, Matthew N. Tran^{1,7}, Zachary Besich^{1,7}, Madhavi Tippani¹, Jennifer Chew⁵, Yifeng Yin⁵, Joel E. Kleinman^{1,2}, Thomas M. Hyde^{1,2,8}, Nikhil Rao⁵, Stephanie C. Hicks^{1,4}, Keri Martinowich^{1,2,6} and Andrew E. Jaffe^{1,2,3,4,6,7,9}

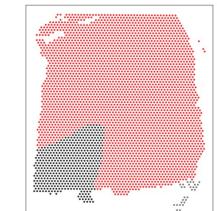
Selected SVGs: human DLPFC



Ground truth labels: human DLPFC

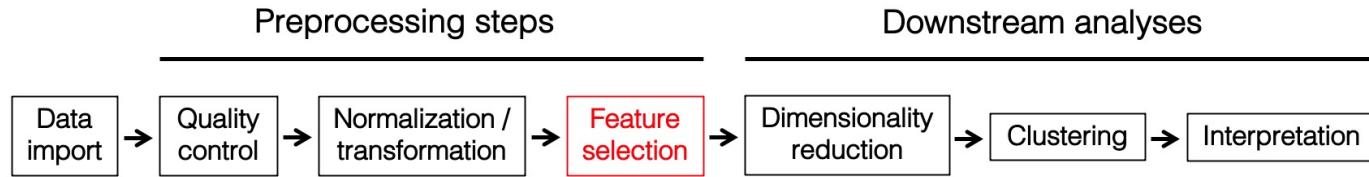


Ground truth labels: human DLPFC



Why are we interested in spatially variable genes?

Feature selection



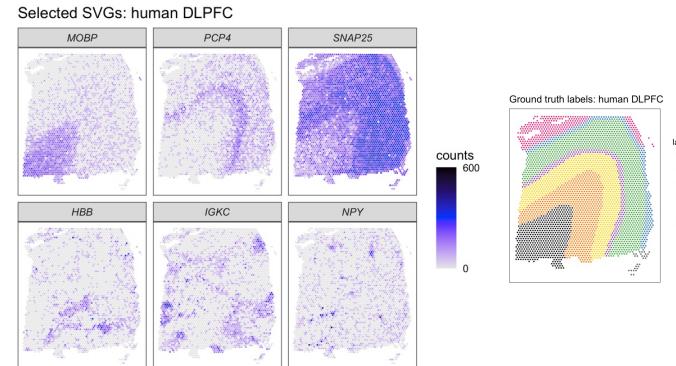
Data preprocessing

- Reduce number of genes (e.g. 20,000 → 1,000) to reduce noise and improve computational performance during downstream analyses

Identify top-ranked genes

- Identify top-ranked genes to further investigate as markers of biological processes

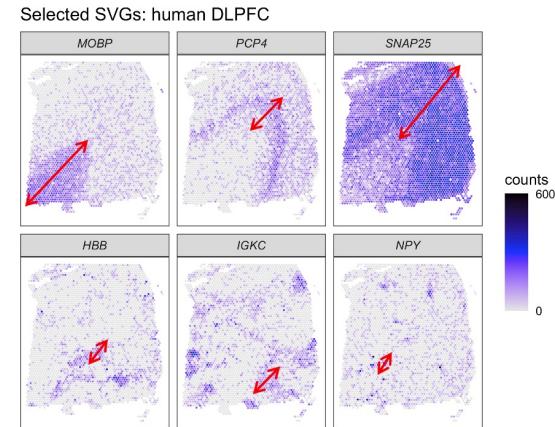
Example: human DLPFC



Existing methods are too slow or perform poorly

Baseline methods

- Highly variable genes (HVGs)
- Moran's I statistic



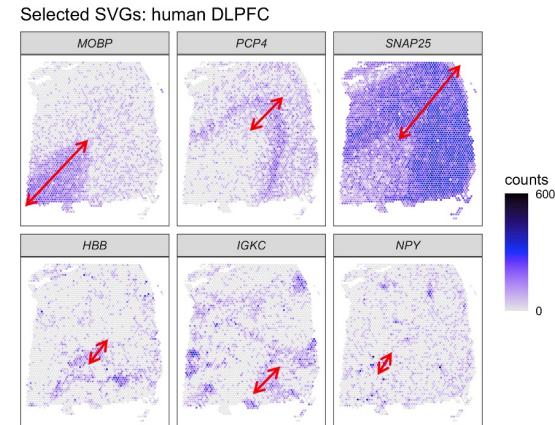
Existing methods are too slow or perform poorly

Baseline methods

- Highly variable genes (HVGs)
- Moran's I statistic

Examples of new methods for spatially variable genes (SVGs)

- SpatialDE (Svensson et al. 2018)
 - Gaussian process regression and likelihood ratio test
 - Scales cubically in number of spatial locations



nnSVG methodology

Nearest-neighbor Gaussian processes

- Datta et al. (2016), Finley et al. (2019)
- Using approximate likelihood (Vecchia 1988) with small set of nearest neighbors (e.g. 10-15) to approximate full data
- Linear scalability with number of spatial locations (due to sparse precision matrix vs. inversion of full covariance matrix)

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand

nnSVG methodology

Nearest-neighbor Gaussian processes

- Datta et al. (2016), Finley et al. (2019)
- Using approximate likelihood (Vecchia 1988) with small set of nearest neighbors (e.g. 10-15) to approximate full data
- Linear scalability with number of spatial locations (due to sparse precision matrix vs. inversion of full covariance matrix)

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2016, VOL. 111, NO. 514, 800–812, Theory and Methods
<http://dx.doi.org/10.1080/01621459.2015.1044091>



Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

Abhirup Datta, Sudipto Banerjee, Andrew O. Finley, and Alan E. Gelfand

nnSVG methodology

- BRISC R package (Saha and Datta, 2018)
- Fit one model per gene and extract maximum likelihood parameter estimates / log-likelihoods
- Exponential covariance function with gene-specific length scale parameter
- Optional covariates for spatial domains
- Likelihood ratio (LR) statistic vs. linear model without spatial terms to rank genes
- Approximate LR test (chi-sq. 2 d.f.) to identify statistically significant SVGs
- Effect size: proportion of spatial variance (Svensson et al. 2018)

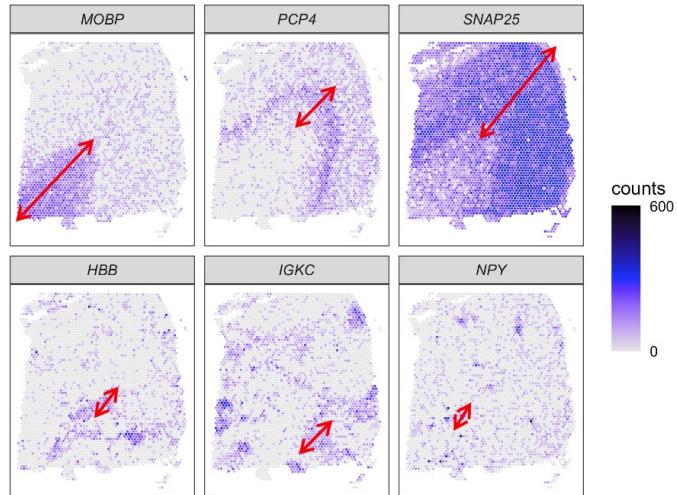
$$y \sim N(X\beta, C(\theta) + \tau^2 I)$$

$$C(\theta) = k(s_i, s_j) = \sigma^2 \exp\left(\frac{-||s_i - s_j||}{l}\right)$$

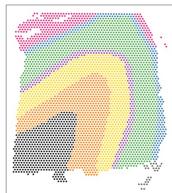
$$propSV = \frac{\sigma^2}{\sigma^2 + \tau^2}$$

nnSVG evaluations / benchmarking

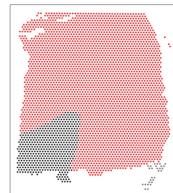
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)



Ground truth labels: human DLPFC

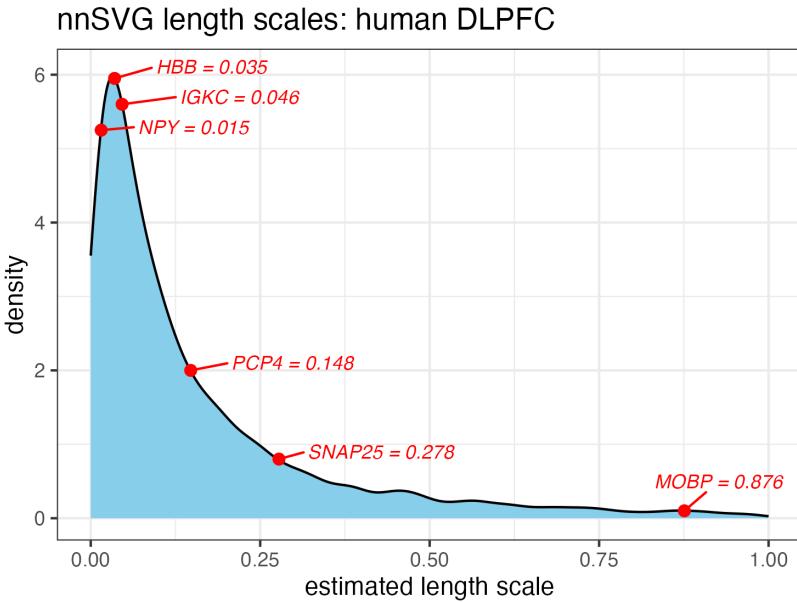
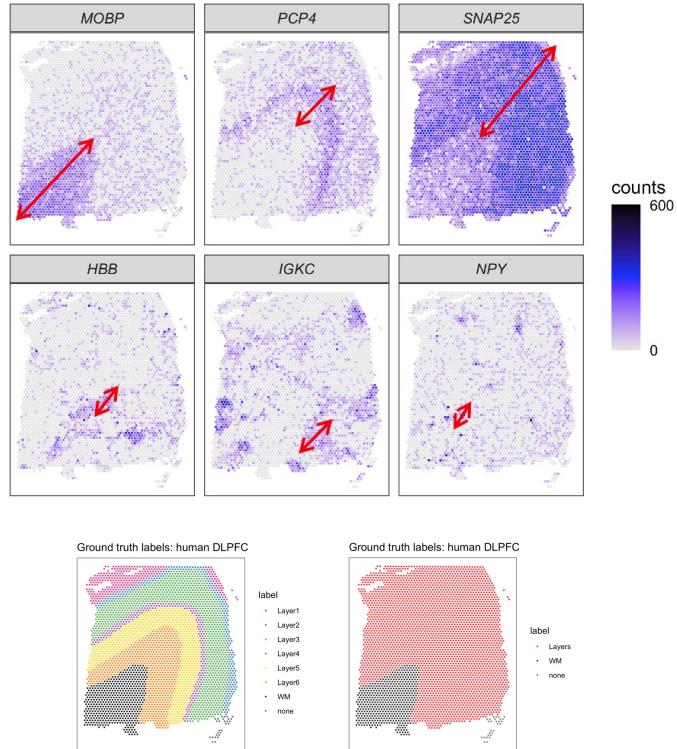


Ground truth labels: human DLPFC



nnSVG evaluations / benchmarking

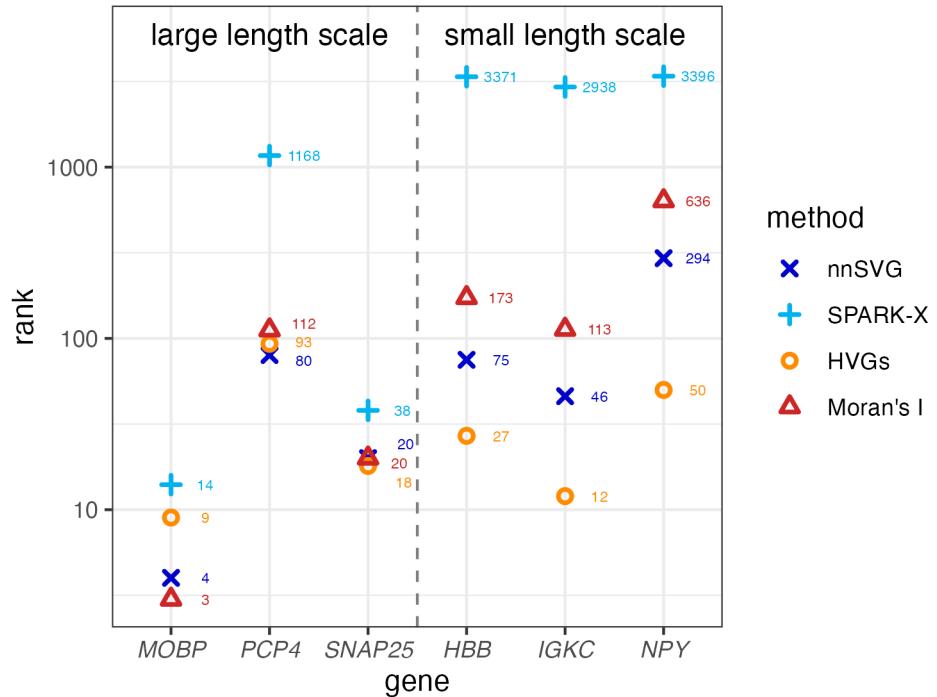
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)



nnSVG evaluations / benchmarking

Human DLPFC dataset

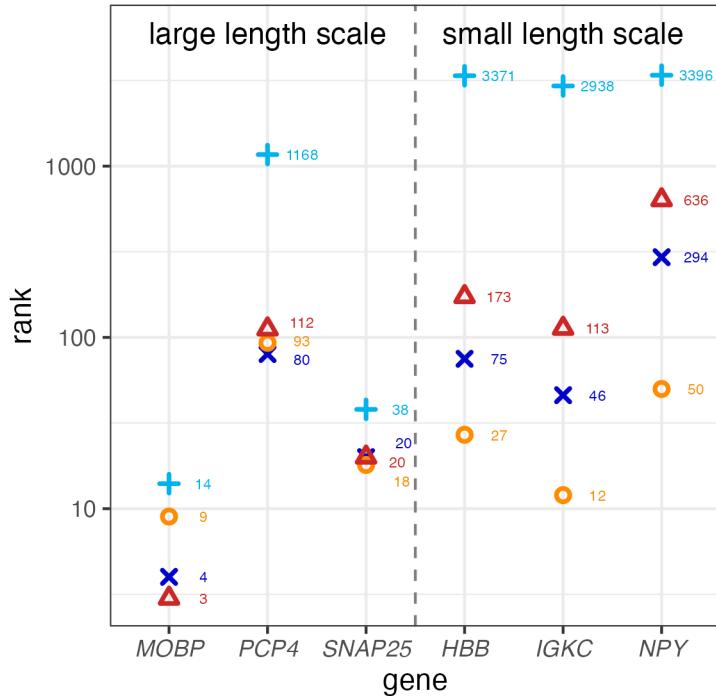
Selected SVGs: human DLPFC



nnSVG evaluations / benchmarking

Human DLPFC dataset

Selected SVGs: human DLPFC



Method performance

HVGs > nnSVG > Moran's I >> SPARK-X

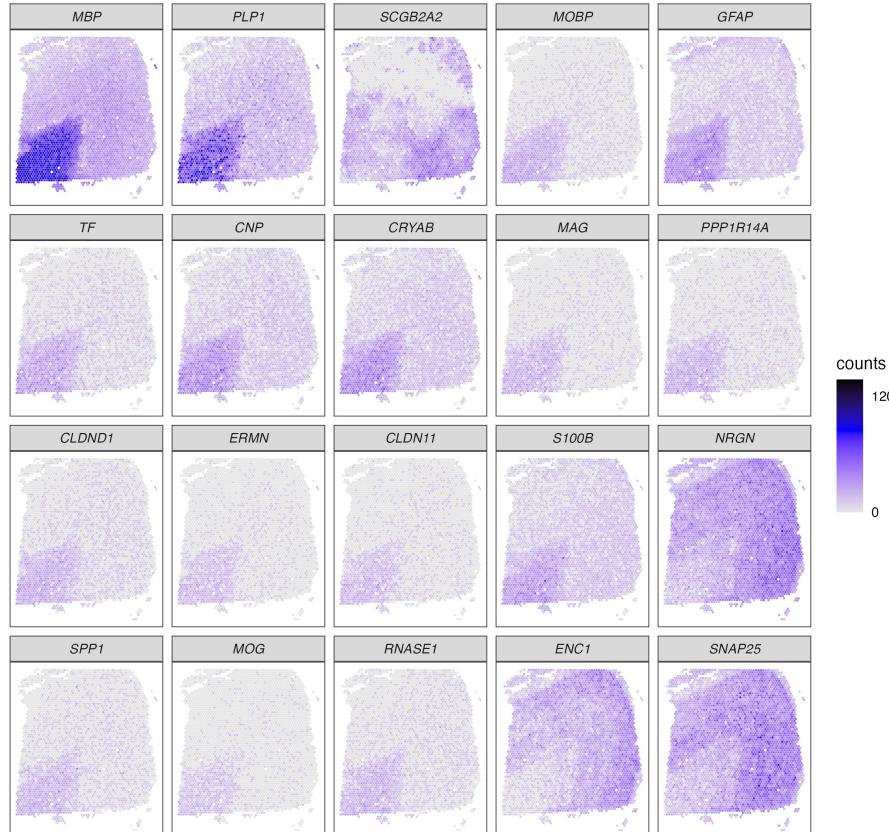
method

- nnSVG (blue X)
- SPARK-X (cyan +)
- HVGs (orange circle)
- Moran's I (red triangle)

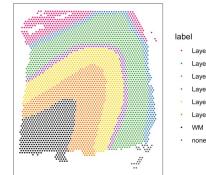
nnSVG evaluations / benchmarking

Human DLPFC dataset

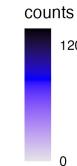
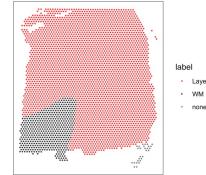
Top SVGs: human DLPFC, nnSVG



Ground truth labels: human DLPFC



Ground truth labels: human DLPFC

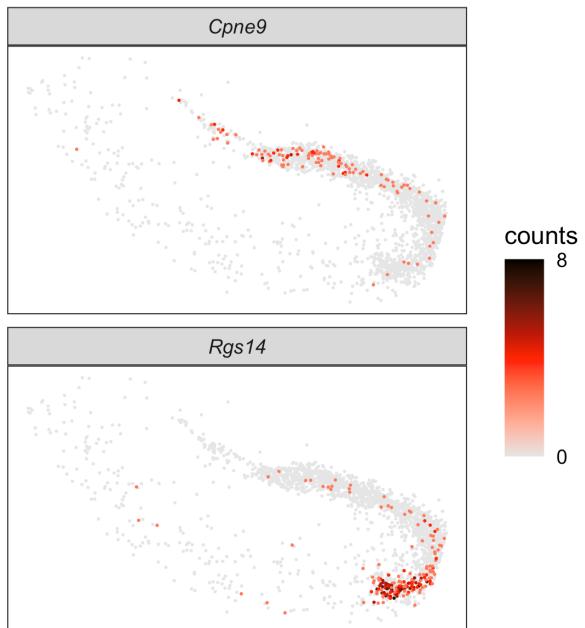


nnSVG evaluations / benchmarking

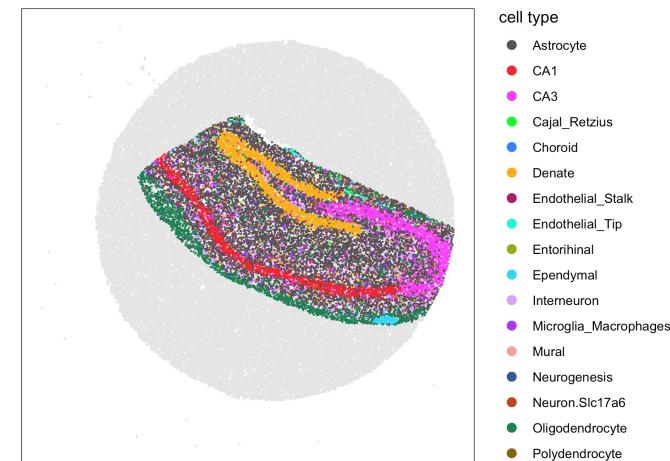
Mouse hippocampus dataset (Slide-seqV2) (Stickels et al. 2020 / Cable et al. 2021)

- Identify SVGs within spatial domain

Selected SVGs: mouse HPC



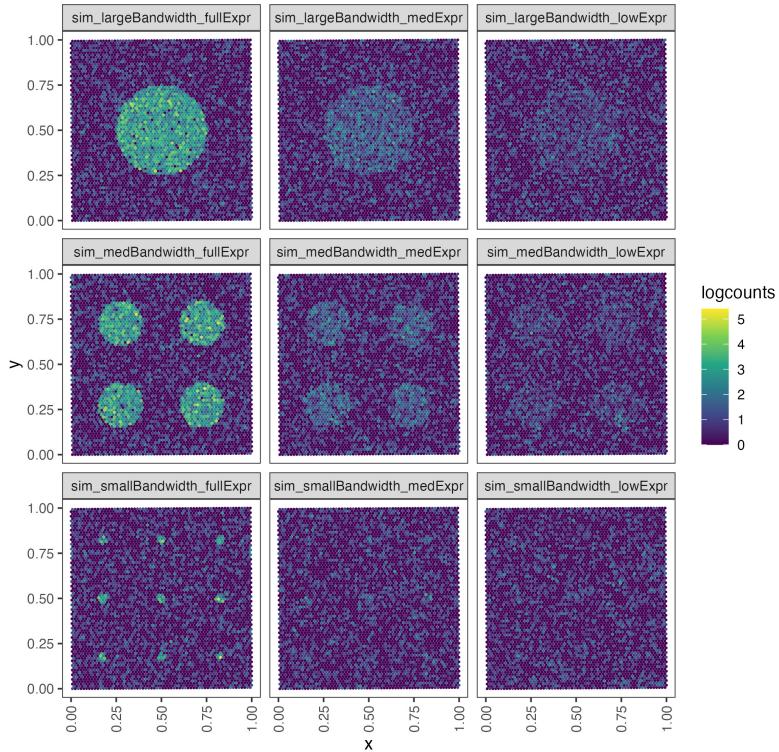
Mouse HPC



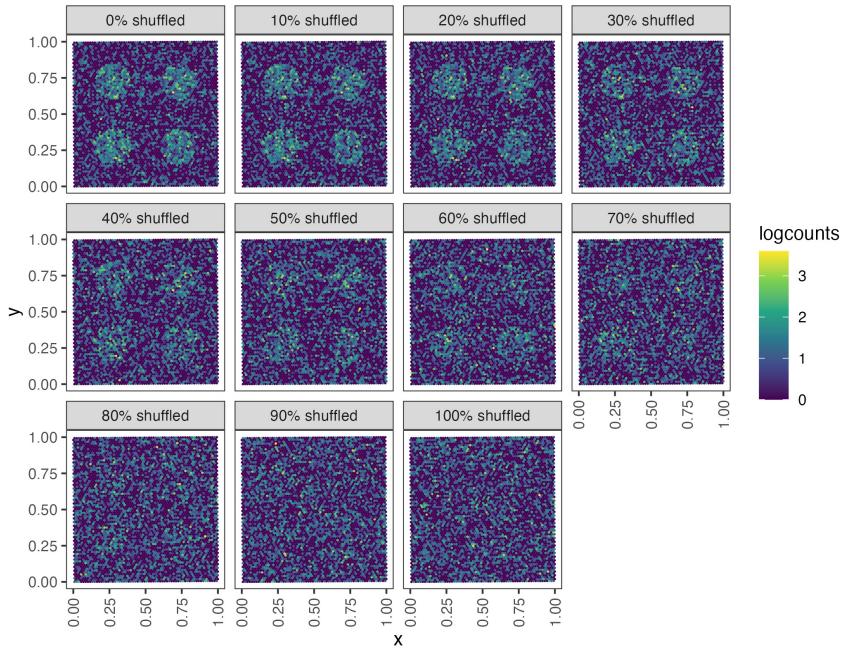
nnSVG evaluations / benchmarking

Simulations

Simulated datasets: expression



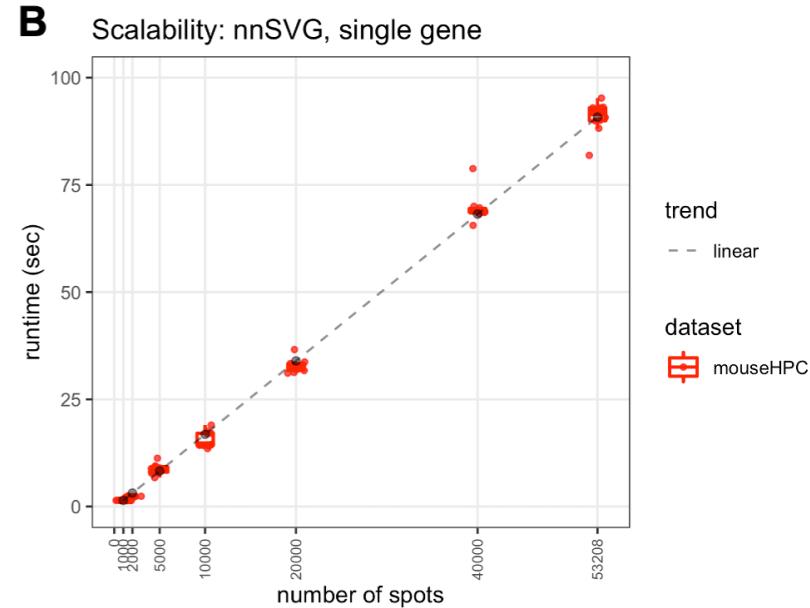
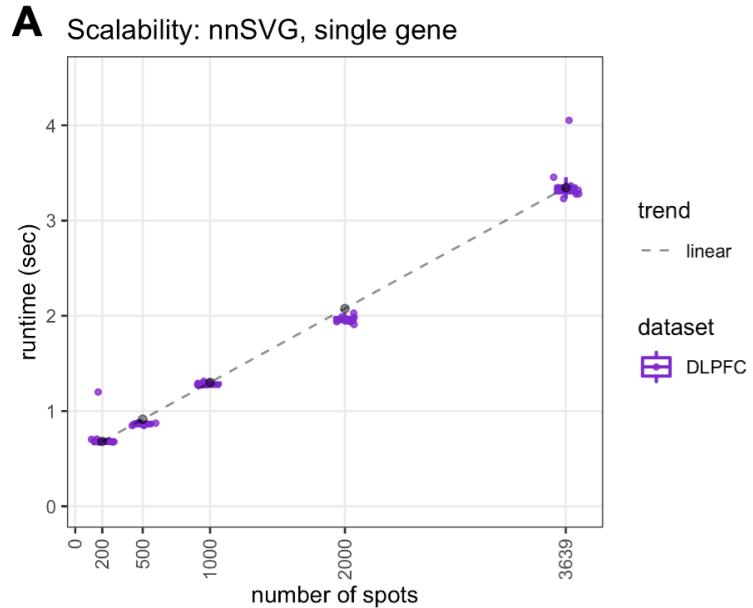
Simulated datasets (shuffled coordinates): expression



nnSVG evaluations / benchmarking

Computational scalability

- Linear in number of spatial locations



nnSVG summary

New method to identify spatially variable genes

- Outperforms existing methods and baseline methods: identifies known SVGs in several datasets
- Sensitivity: flexible length scale parameter, optional covariates for spatial domains
- Linear scalability: can apply to datasets with thousands of spatial locations

Preprint (in revision)



bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

[Follow this preprint](#)

nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes

Lukas M. Weber, Arkajyoti Saha, Abhirup Datta, Kasper D. Hansen, Stephanie C. Hicks

doi: <https://doi.org/10.1101/2022.05.16.492124>

nnSVG implementation

R/Bioconductor package

- Open-source software package
- Documentation and tutorial



Home » Bioconductor 3.16 » Software Packages » nnSVG

nnSVG

platforms	all	rank	1940 / 2183	support	0 / 0	In Bioc	0.5 years
build	ok	updated	before release	dependencies	100		

DOI: [10.18129/B9.bioc.nnSVG](https://doi.org/10.18129/B9.bioc.nnSVG) [f](#) [t](#)

Scalable identification of spatially variable genes in spatially-resolved transcriptomics data

Bioconductor version: Release (3.16)

Method for scalable identification of spatially variable genes (SVGs) in spatially-resolved transcriptomics data. The method is based on nearest-neighbor Gaussian processes and uses the BRISC algorithm for model fitting and parameter estimation. Allows identification and ranking of SVGs with flexible length scales across a tissue slide or within spatial domains defined by covariates. Scales linearly with the number of spatial locations and can be applied to datasets containing thousands or more spatial locations.

nnSVG implementation

R/Bioconductor package

- Open-source software package
- Documentation and tutorial

The Bioconductor logo is located at the top left. To its right is a teal navigation bar with three tabs: "Home", "Install", and "Help".

Home » Bioconductor 3.16 » Software Packages » nnSVG

nnSVG

platforms all | rank 1940 / 2183 | support 0 / 0 | in Bioc 0.5 years
 build ok | updated before release | dependencies 100

DOI: [10.18129/B9.bioc.nnSVG](https://doi.org/10.18129/B9.bioc.nnSVG) [f](#) [t](#)

Scalable identification of spatially variable genes in spatially-resolved transcriptomics data

Bioconductor version: Release (3.16)

Method for scalable identification of spatially variable genes (SVGs) in spatially-resolved transcriptomics data. The method is based on nearest-neighbor Gaussian processes and uses the BRISC algorithm for model fitting and parameter estimation. Allows identification and ranking of SVGs with flexible length scales across a tissue slide or within spatial domains defined by covariates. Scales linearly with the number of spatial locations and can be applied to datasets containing thousands or more spatial locations.

1	Introduction
2	Installation
3	Input data format
4	Tutorial
5	Troubleshooting
6	Session information

nnSVG Tutorial

Lukas M. Weber¹ and Stephanie C. Hicks¹

¹Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA

5 May 2023

Package

nnSVG 1.4.1

1 Introduction

nnSVG is a method for scalable identification of spatially variable genes (SVGs) in spatially-resolved transcriptomics data.

The nnSVG method is based on nearest-neighbor Gaussian processes (Datta et al., 2016, Finley et al., 2019) and uses the BRISC algorithm (Saha and Datta, 2018) for model fitting and parameter estimation. nnSVG allows identification and ranking of SVGs with flexible length scales across a tissue slide or within spatial domains defined by covariates. The method scales linearly with the number of spatial locations and can be applied to datasets containing thousands or more spatial locations.

nnSVG is implemented as an R package within the Bioconductor framework, and is available from Bioconductor.

More details describing the method are available in our preprint, available from bioRxiv.

2 Installation

The following code will install the latest release version of the nnSVG package from Bioconductor. Additional details are shown on the Bioconductor page.

```
install.packages("BiocManager")
BiocManager:::install("nnSVG")
```

The latest development version can also be installed from the `devel` version of Bioconductor or from GitHub.

Analysis workflows and tools

SpatialExperiment: R/Bioconductor infrastructure to store spatially-resolved transcriptomics datasets



Volume 38, Issue 11
1 June 2022

Article Contents

JOURNAL ARTICLE

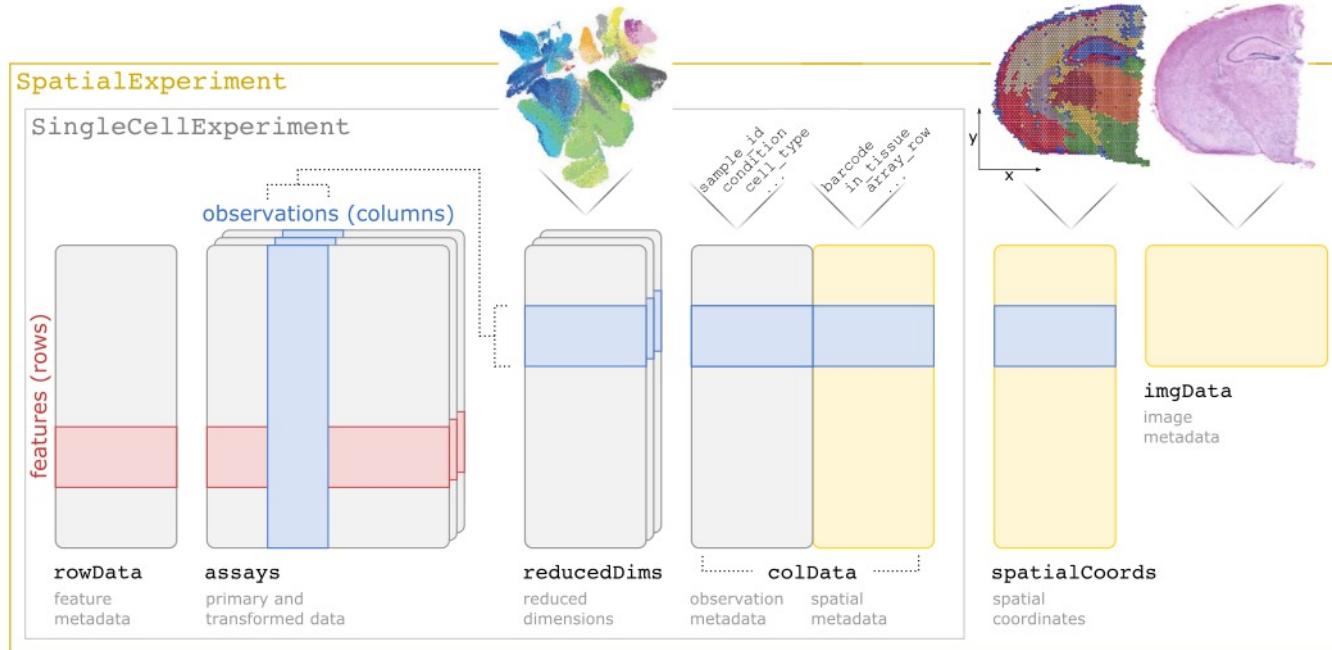
SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor 3

Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T L Lun, Stephanie C Hicks, Davide Risso, Author Notes

Bioinformatics, Volume 38, Issue 11, 1 June 2022, Pages 3128–3131,

<https://doi.org/10.1093/bioinformatics/btac299>

Published: 28 April 2022 Article history ▾



Analysis workflows and tools

Unsupervised analyses of single-nucleus and spatially-resolved landscape of gene expression in locus coeruleus (post-mortem human brain)

Data and code available



Genetics and Genomics, Neuroscience

The gene expression landscape of the human locus coeruleus revealed by single-nucleus and spatially-resolved transcriptomics

Lukas M. Weber, Heena R. Divecha, Matthew N. Tran, Sang Ho Kwon, Abby Spangler, Kelsey D. Montgomery, Madhavi Tippani, Rahul Bharadwaj, Joel E. Kleinman ... Stephanie C. Hicks ... [show 5 more](#)

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA • Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, 21205, USA • Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA ... [show 3 more](#)

<https://doi.org/10.7554/eLife.84628.1>

Analysis workflows and tools

Unsupervised analyses of single-nucleus and spatially-resolved landscape of gene expression in locus coeruleus (post-mortem human brain)

Data and code available



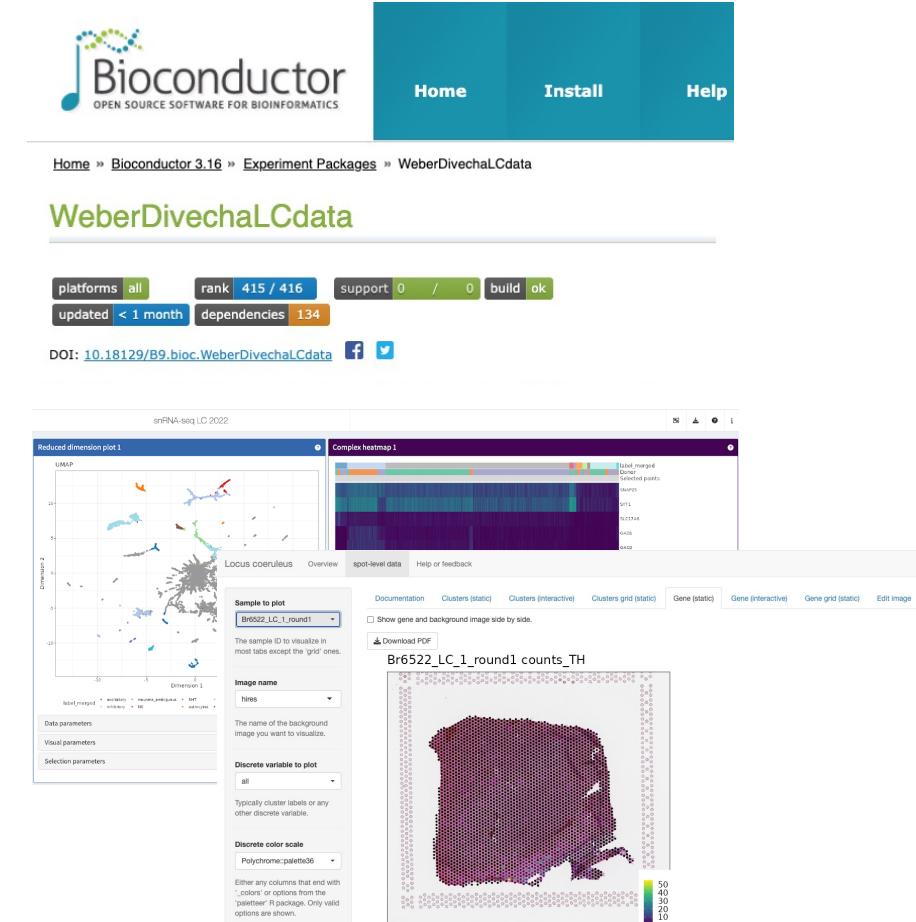
Genetics and Genomics, Neuroscience

The gene expression landscape of the human locus coeruleus revealed by single-nucleus and spatially-resolved transcriptomics

Lukas M. Weber, Heena R. Divecha, Matthew N. Tran, Sang Ho Kwon, Abby Spangler, Kelsey D. Montgomery, Madhavi Tippani, Rahul Bharadwaj, Joel E. Kleinman ... Stephanie C. Hicks ... [show 5 more](#)

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, 21205, USA • Lieber Institute for Brain Development, Johns Hopkins Medical Campus, Baltimore, MD, 21205, USA • Department of Neuroscience, Johns Hopkins School of Medicine, Baltimore, MD, 21205, USA ... [show 3 more](#)

<https://doi.org/10.7554/eLife.84628.1>



Analysis workflows and tools

Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor (OSTA)

Online book containing example R code and datasets

(Work in progress)

The screenshot shows the 'Welcome' page of the OSTA online book. The left sidebar contains a navigation menu with sections I through IV, each with several numbered items. The main content area features the title 'Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor' and a date '2022-08-07'. Below this is a 'Welcome' section with text about the book's purpose and organization, followed by a Bioconductor logo.

OSTA

I Introduction

1 Introduction

2 Spatially-resolved transcriptomics

3 SpatialExperiment

II Preprocessing steps

4 Preprocessing steps

5 Image segmentation (Visium)

6 Loupe Browser (Visium)

7 Space Ranger (Visium)

III Analysis steps

8 Analysis steps

9 Quality control

10 Normalization

11 Feature selection

12 Dimensionality reduction

13 Clustering

14 Marker genes

15 Spot-level deconvolution

IV Workflows

Welcome

This is the website for the online book "Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor" (OSTA).

This book provides several examples of computational analysis workflows for **spatially resolved transcriptomics (SRT)** data, using the **Bioconductor** framework within the R programming language. The chapters contain details on individual analysis steps as well as complete workflows, with example datasets and R code that can be run on your own laptop.

The book is organized into several parts, including background, preprocessing steps, analysis steps, and complete workflows.

Additional introductory material on R and Bioconductor can also be found in the related book *Orchestrating Single-Cell Analysis with Bioconductor (OSCA)*.



The screenshot shows a chapter titled 'Chapter 17 Human DLPFC workflow'. The left sidebar lists various analysis steps. The main content area includes a heading '17.1 Description of dataset #', a block of explanatory text, and a code block starting with '# clear workspace from previous chapters'. The right side of the interface has navigation arrows.

III Analysis steps

8 Analysis steps

9 Quality control

10 Normalization

11 Feature selection

12 Dimensionality reduction

13 Clustering

14 Marker genes

15 Spot-level deconvolution

IV Workflows

16 Workflows

17 Human DLPFC workflow

17.1 Description of dataset

17.2 Load data

17.3 Plot data

17.4 Quality control (QC)

17.5 Normalization

17.6 Feature selection

17.7 Spatially-aware feature selection

17.8 Dimensionality reduction

17.9 Clustering

Chapter 17 Human DLPFC workflow

This workflow analyzes one sample of human brain from the dorsolateral prefrontal cortex (DLPFC) region, measured using the 10x Genomics Visium platform. This is a condensed version of the analyses shown in the individual analysis chapters in the previous part. For more details on the individual steps, see the previous chapters.

```
# clear workspace from previous chapters
rm(list = ls(all = TRUE))
```

17.1 Description of dataset #

This is a 10x Genomics Visium dataset generated from healthy human brain samples from the dorsolateral prefrontal cortex (DLPFC) region.

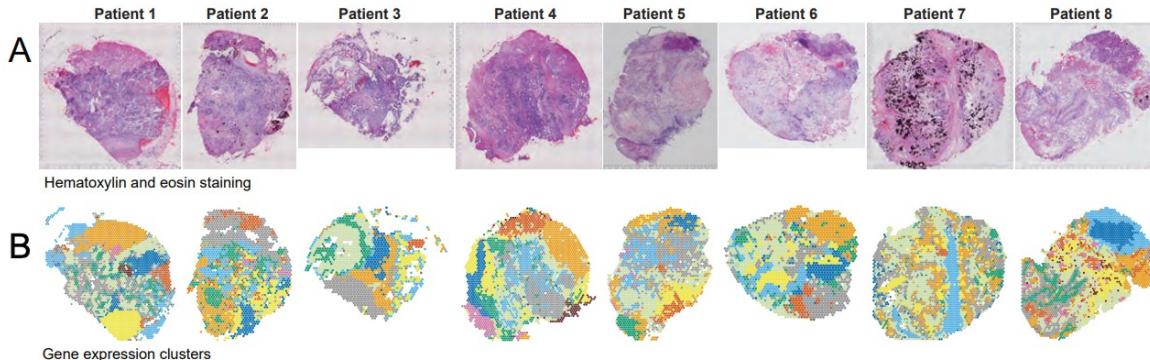
In the full dataset, there are 12 samples in total, from 3 individuals, with 2 pairs of spatially adjacent replicates (serial sections) per individual (4 samples per individual). The individuals and spatially adjacent replicates can be used as blocking factors. Each sample spans the six layers of the cortex plus white matter in a perpendicular tissue section.

For the examples in this workflow and the analysis chapters, we use a single sample from this dataset (sample 151673), to keep the computational requirements to compile the book manageable.

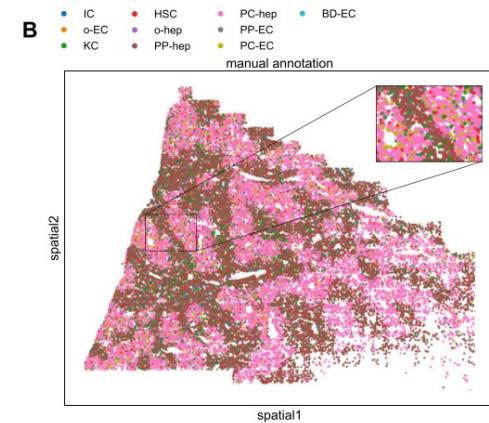
Next steps

New spatially-resolved transcriptomics platforms

- Datasets with multiple samples
- Datasets at single-cell or sub-cellular resolution



Denisenko et al. (2022), *bioRxiv*



Liu and Tran et al. (2022),
Life Science Alliance

Acknowledgments

Johns Hopkins Bloomberg School of Public Health
Department of Biostatistics

Stephanie Hicks
Kasper Hansen
Abhirup Datta
Arkajyoti Saha



Lieber Institute for Brain Development

Keri Martinowich
Kristen Maynard
Leonardo Collado-Torres



SpatialExperiment team

Dario Righelli
Helena Crowell
Davide Risso

Bioconductor community



Funding



Acknowledgments

Johns Hopkins Bloomberg School of Public Health
Department of Biostatistics

Stephanie Hicks
Kasper Hansen
Abhirup Datta
Arkajyoti Saha



Lieber Institute for Brain Development

Keri Martinowich
Kristen Maynard
Leonardo Collado-Torres



SpatialExperiment team

Dario Righelli
Helena Crowell
Davide Risso

Bioconductor community



Funding



Boston University School of
Public Health
Starting as
Assistant Professor in Biostatistics

October 2023

looking for collaborators!

Thank you!

ADDITIONAL SLIDES

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

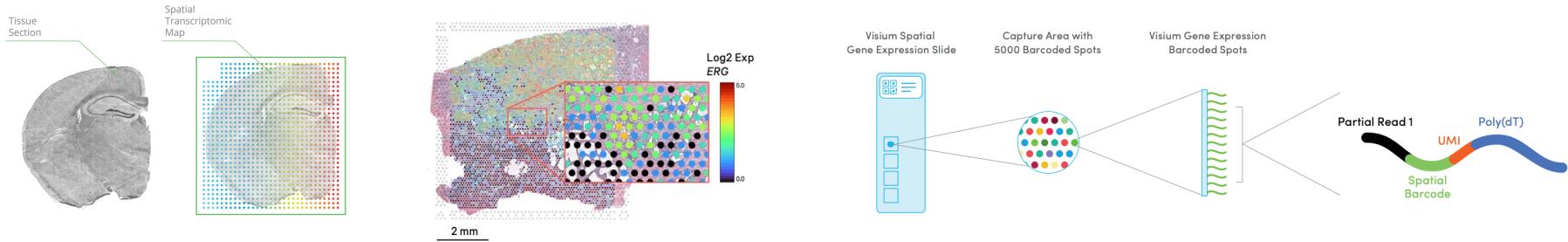


Image source: 10x Genomics

Illustration: cell populations within brain

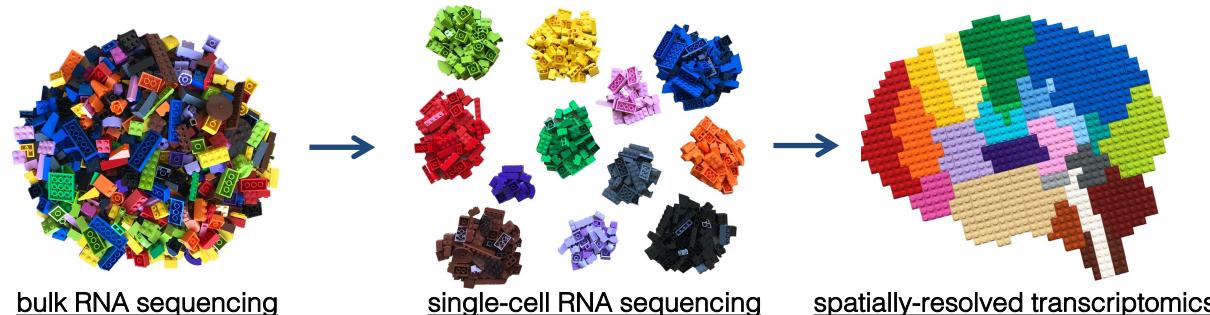
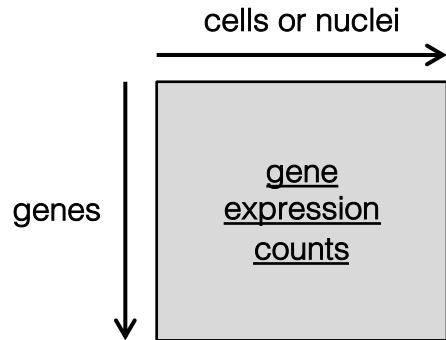


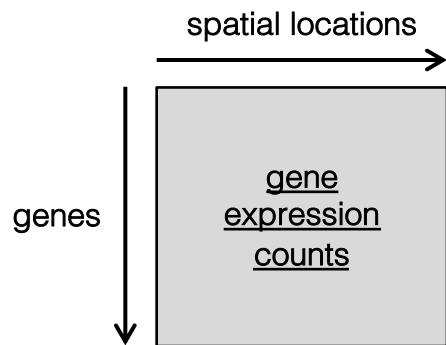
Image credit: Bo Xia
<https://twitter.com/BoXia7>

Data structure

Single-cell /
single-nucleus
RNA sequencing



Spatially-resolved
transcriptomics



+

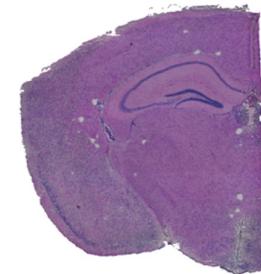
spatial
coordinates

> head(spatialCoords(spe))

	x	y
AAACAACGAATAGTTC-1	3913	2435
AAACAAGTATCTCCCC-1	9791	8468
AAACAATCTACTAGCA-1	5769	2807
AAACACCAATAACTGC-1	4068	9505
AAACAGAGCGACTCCT-1	9271	4151
AAACAGCTTCAGAAG-1	3393	7583

+

image features



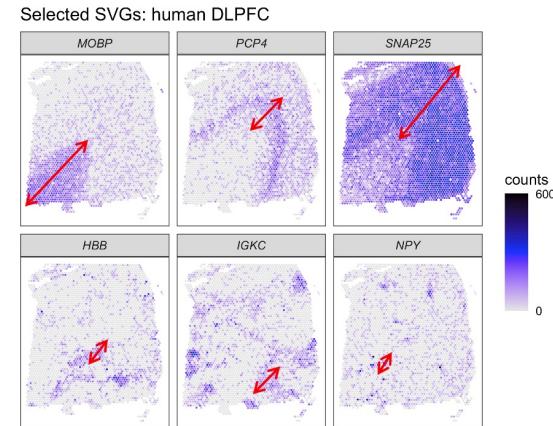
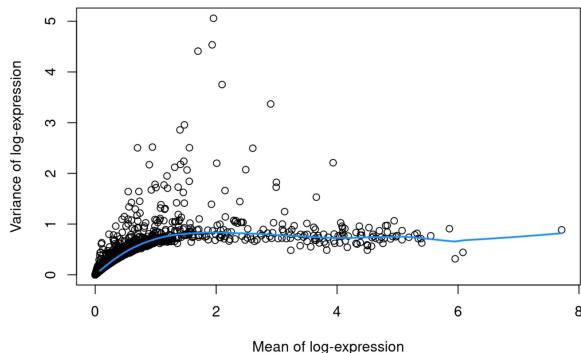
Existing methods are too slow or perform poorly

Baseline methods

- Highly variable genes (HVGs)
- Moran's I statistic

Highly variable genes (HVGs) (non-spatial baseline)

- Rank genes by “excess biological variation” above assumed technical trend (after normalization and log transformation)
- Accounts for mean-variance relationship
- Top 10% or top 1000 HVGs used for downstream analyses



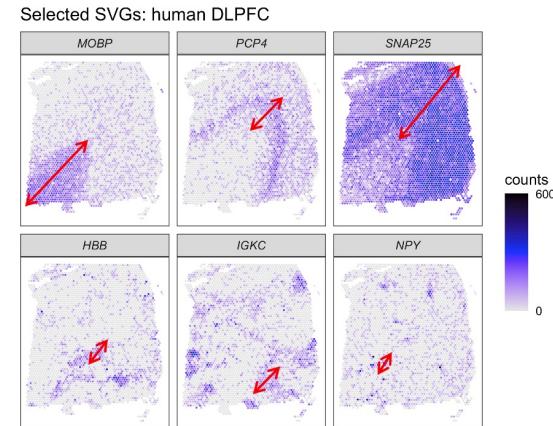
Existing methods are too slow or perform poorly

Baseline methods

- Highly variable genes (HVGs)
- Moran's I statistic

Examples of new methods for spatially variable genes (SVGs)

- **SpatialDE** (Svensson et al. 2018)
Gaussian process regression and likelihood ratio test
Scales cubically in number of spatial locations
- **SPARK-X** (Zhu et al. 2021)
Fast approximation loses sensitivity to spatial patterns with varying length scales



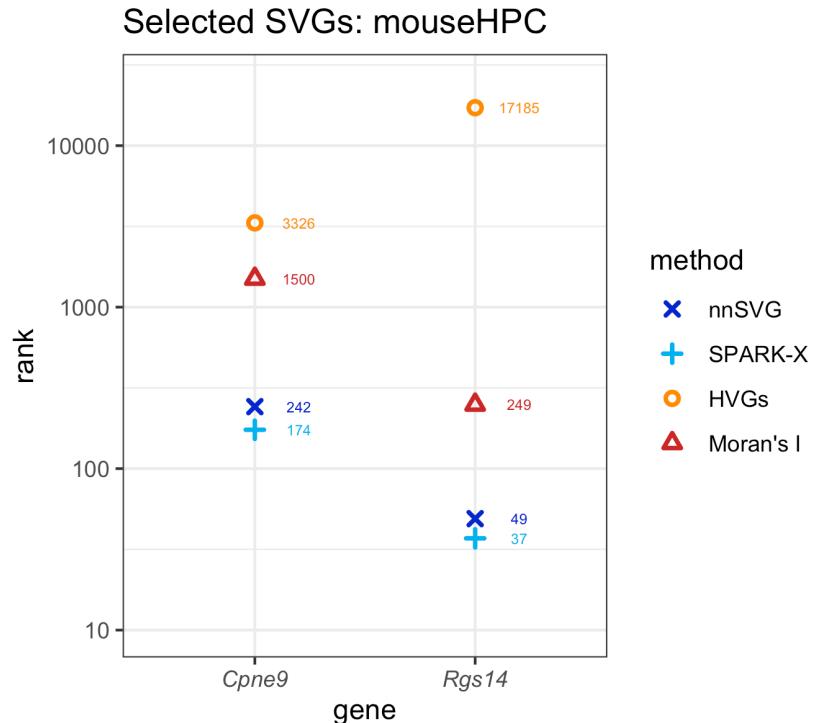
$$P(y | \mu, \sigma_s^2, \delta, \Sigma) = N(y | \mu \cdot 1, \sigma_s^2 \cdot (\Sigma + \delta \cdot I))$$

$$\Sigma_{i,j} = k(x_i, x_j) = \exp\left(-\frac{|x_i - x_j|^2}{2 \cdot l^2}\right)$$

nnSVG evaluations / benchmarking

Mouse hippocampus dataset

- With covariates for spatial domains



Method performance

SPARK-X ~ nnSVG >> Moran's I >> HVGs

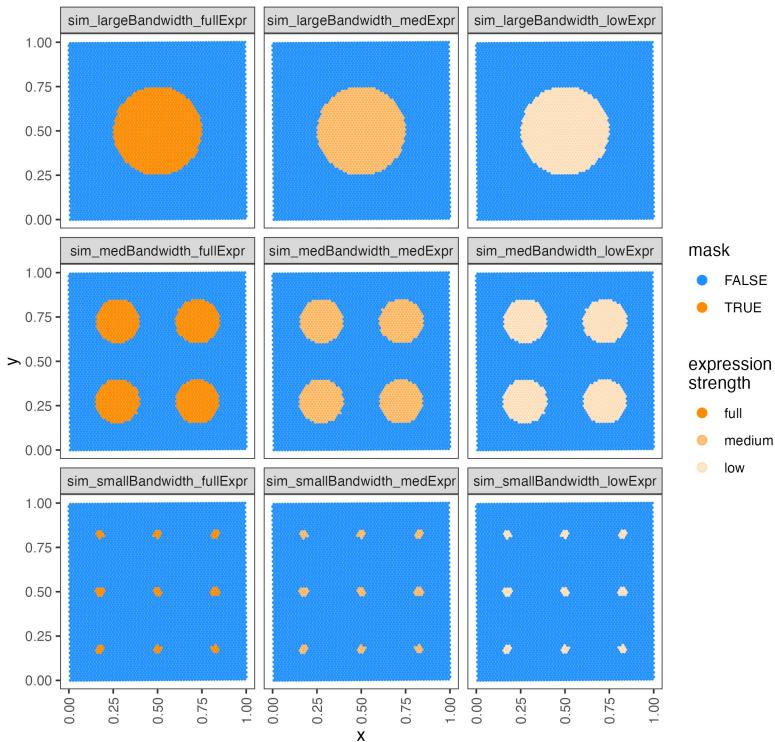
method

- nnSVG
- SPARK-X
- HVGs
- Moran's I

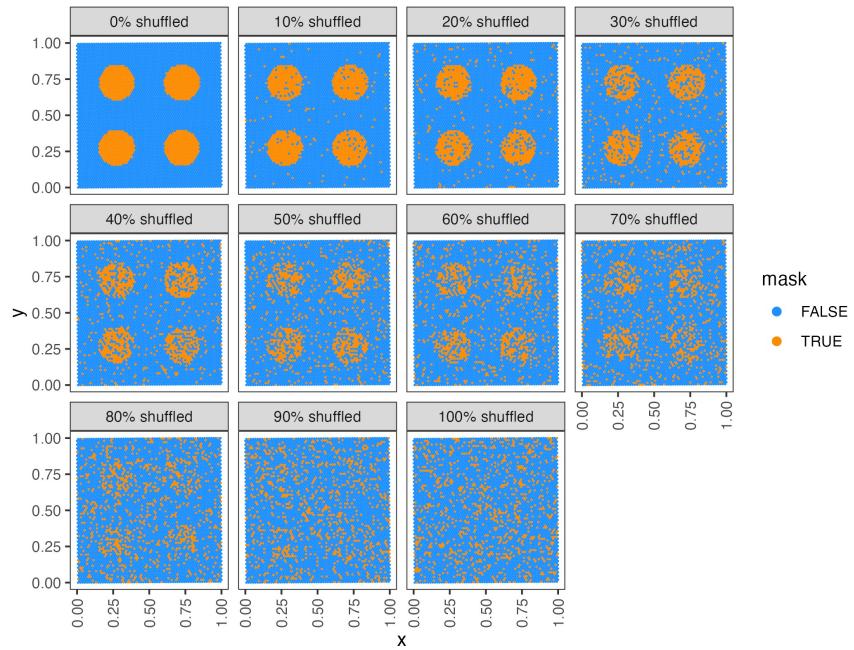
nnSVG evaluations / benchmarking

Simulations

Simulated datasets: spatial coordinate masks

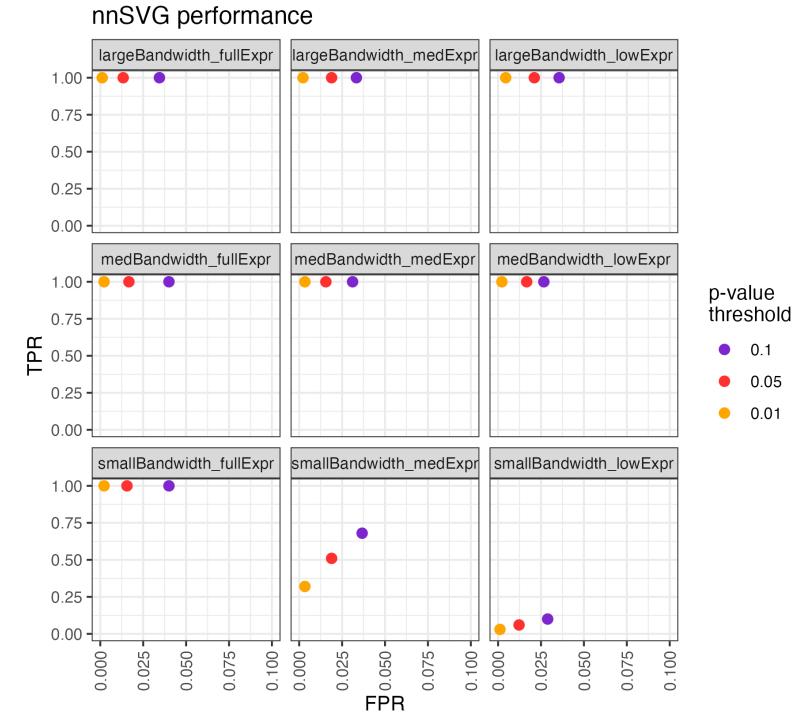
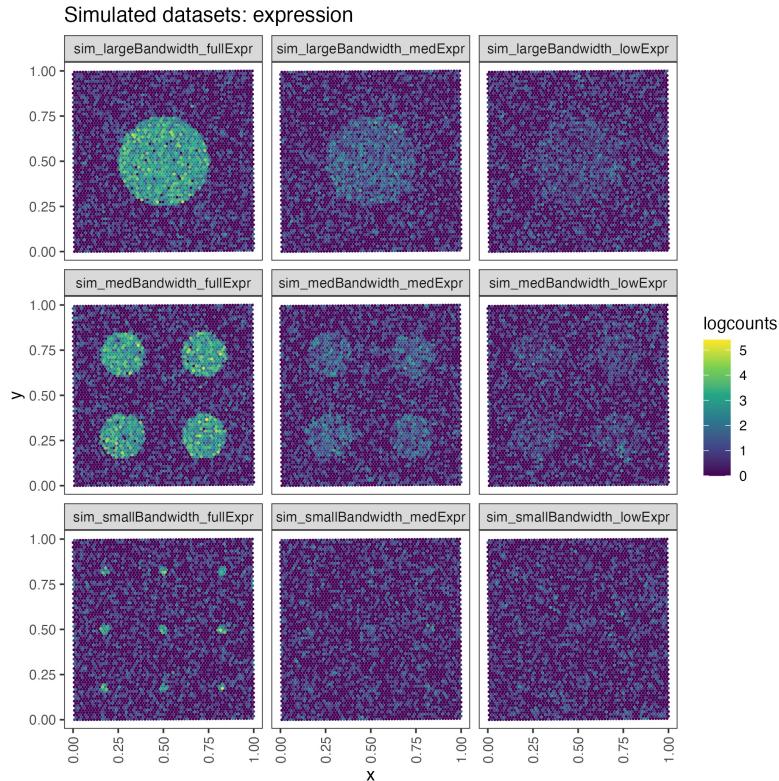


Simulated datasets (shuffled coordinates): spatial coordinate masks



nnSVG evaluations / benchmarking

Simulations



nnSVG evaluations / benchmarking

Simulations

