lmweber.org
@lmwebr

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

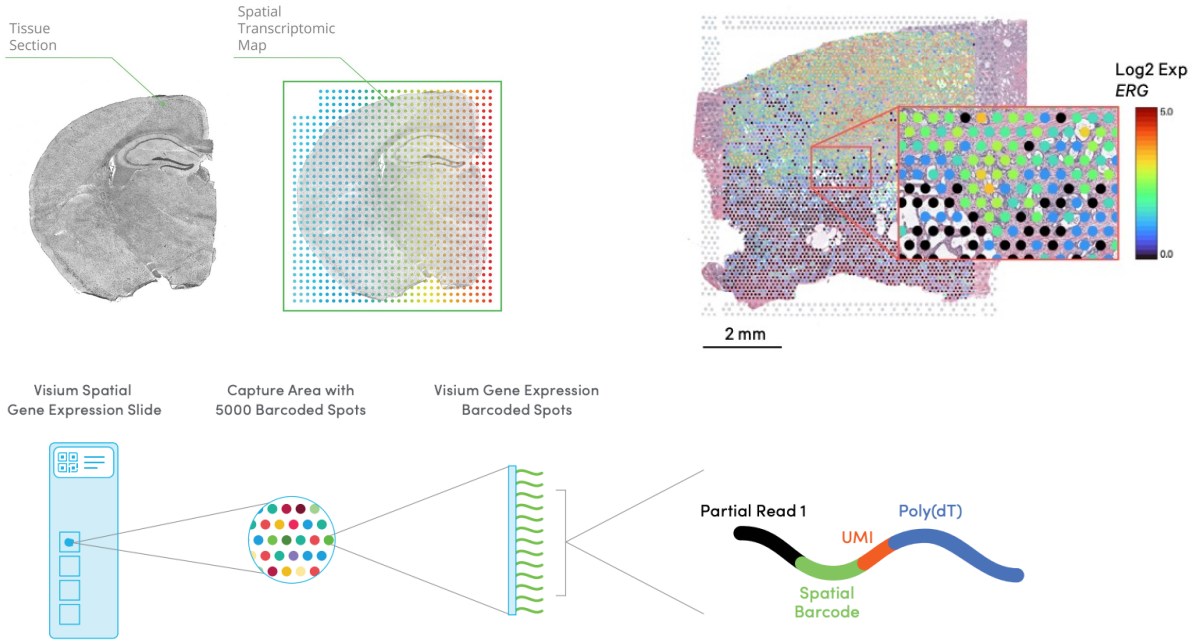


Image source: 10x Genomics

Spatially-resolved transcriptomics

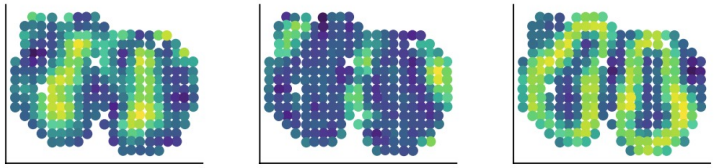
<u>Unsupervised / discovery-based analyses</u>

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

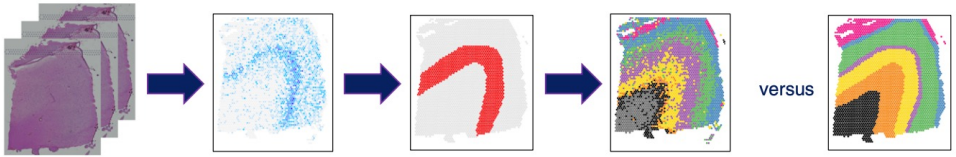
Exploratory analyses

Cell type / state heterogeneity

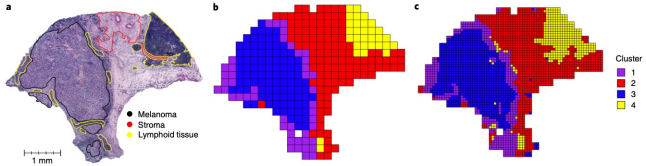
Biologically informative / marker genes



Svensson et al. (2018)



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)

nature neuroscience

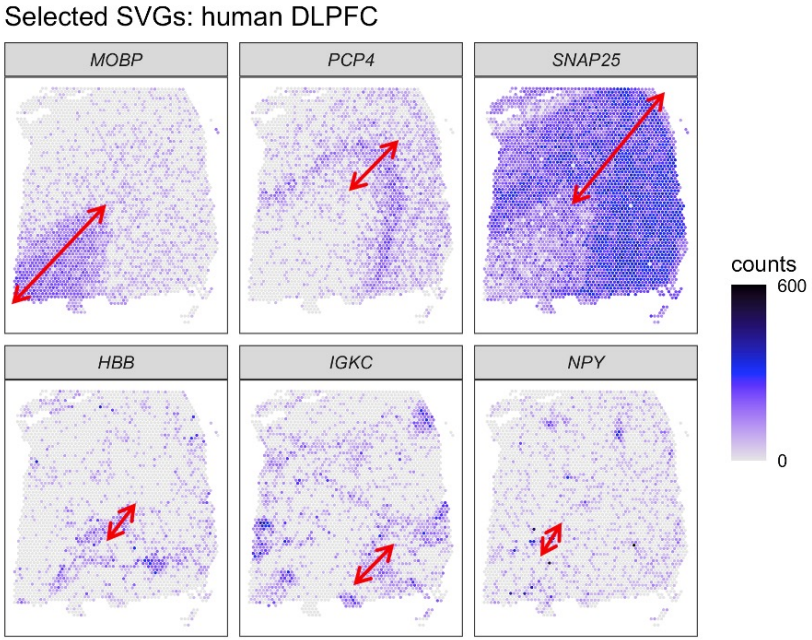
RESOURCE

https://doi.org/10.1038/s41593-020-00787-0

Check for updates

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. Weber[4], Cedric Uytingco[5], Brianna K. Barry[1,6], Stephen R. Williams[5], Joseph L. Catallini II[4], Matthew N. Tran[1,7], Zachary Besich[1,7], Madhavi Tippani[1], Jennifer Chew[5], Yifeng Yin[5], Joel E. Kleinman[1,2], Thomas M. Hyde[1,2,8], Nikhil Rao[5], Stephanie C. Hicks[4], Keri Martinowich[1,2,6] ✉ and Andrew E. Jaffe[1,2,3,4,6,7,9] ✉



Spatially variable genes

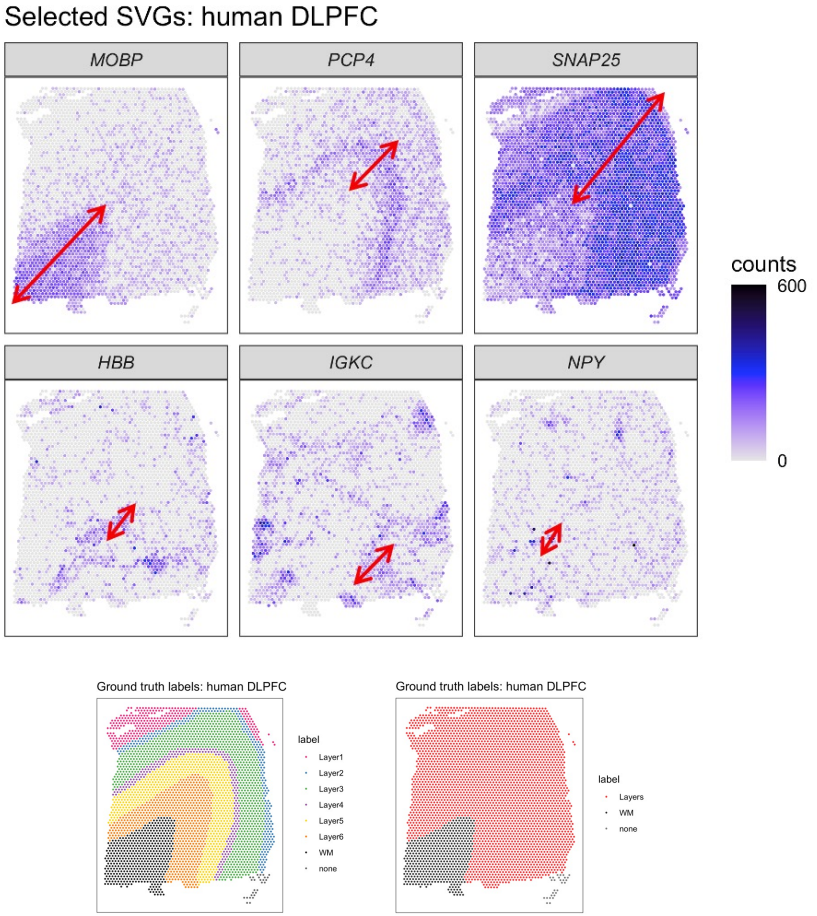
Example

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

nature neuroscience
RESOURCE
https://doi.org/10.1038/s41593-020-00787-0
Check for updates

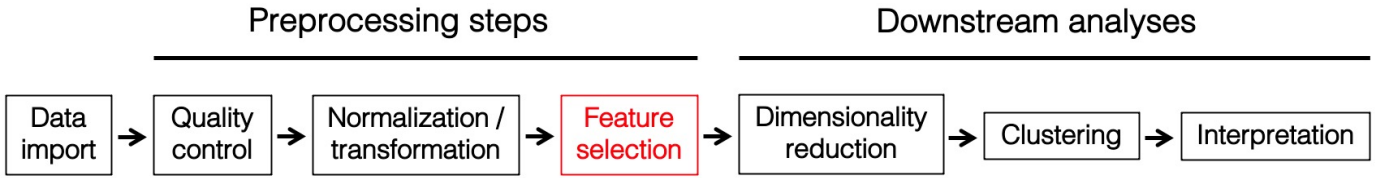
Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex

Kristen R. Maynard[1,2,10], Leonardo Collado-Torres[1,3,10], Lukas M. Weber[4], Cedric Uytingco[5], Brianna K. Barry[1,6], Stephen R. Williams[5], Joseph L. Catallini II[4], Matthew N. Tran[1,7], Zachary Besich[1,7], Madhavi Tippani[1], Jennifer Chew[5], Yifeng Yin[5], Joel E. Kleinman[1,2], Thomas M. Hyde[1,2,8], Nikhil Rao[5], Stephanie C. Hicks[4], Keri Martinowich[1,2,6] ✉ and Andrew E. Jaffe[1,2,3,4,6,7,9] ✉



Why are we interested in spatially variable genes?

Feature selection



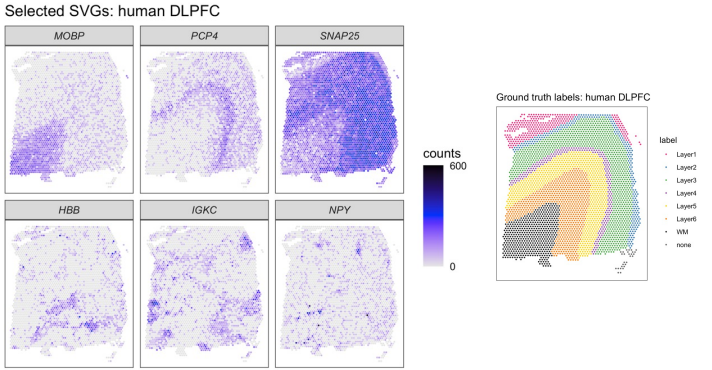
Data preprocessing

- 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

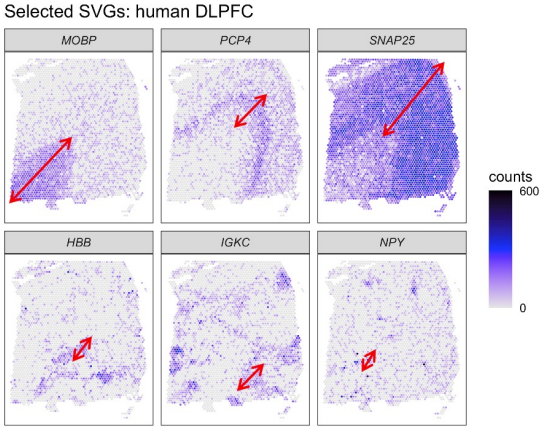
Example: human DLPFC



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

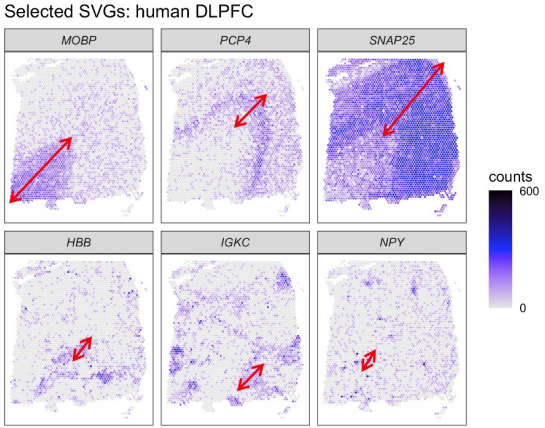
- Highly variable genes (HVGs)
- Moran's I statistic

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

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

Taylor & Francis
Taylor & Francis Group

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

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

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)

nnSVG methodology

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

Taylor & Francis
Taylor & Francis Group

Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets

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

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)

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)

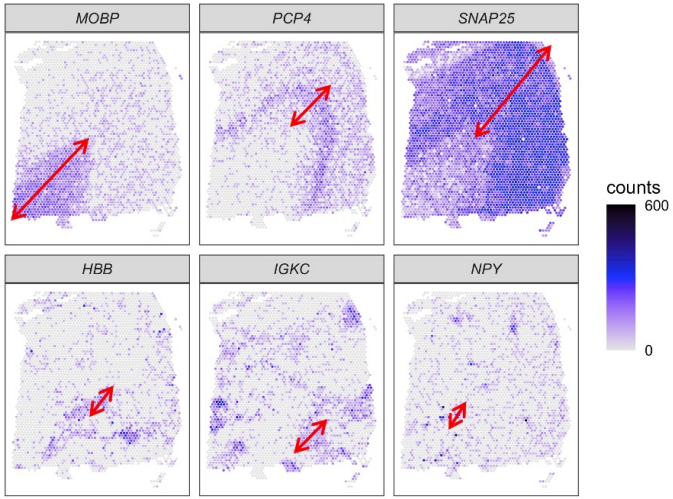
$$y \sim N(X\beta, C(\theta) + \tau^2 I)$$

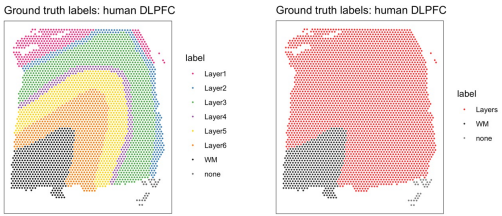
$$C(\theta) = k(s_i, s_j) = \sigma^2 \exp\left(\frac{-||s_i - s_j||}{l}\right)$$

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

nnSVG evaluations / benchmarking

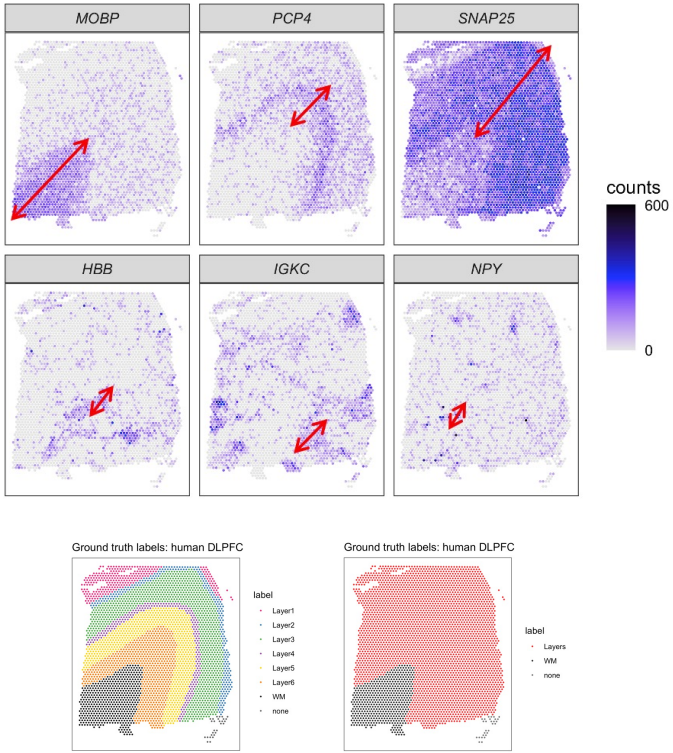
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)

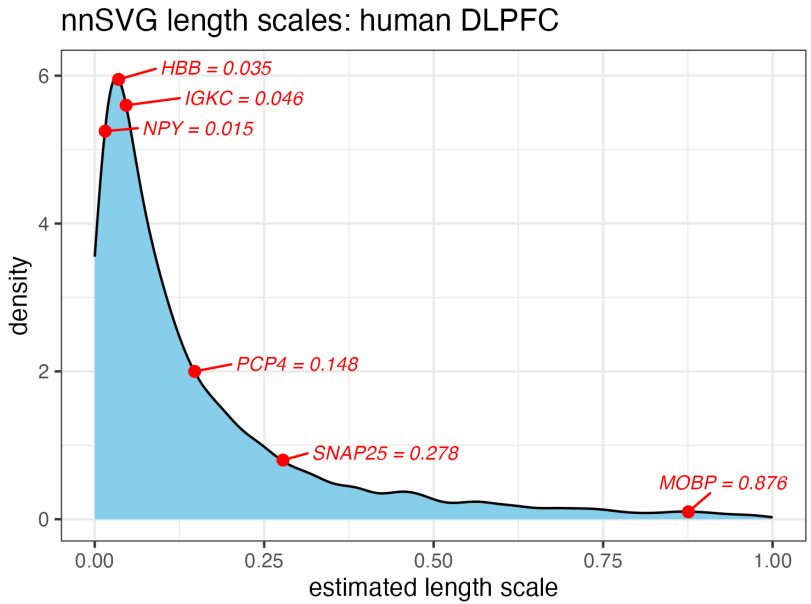




nnSVG evaluations / benchmarking

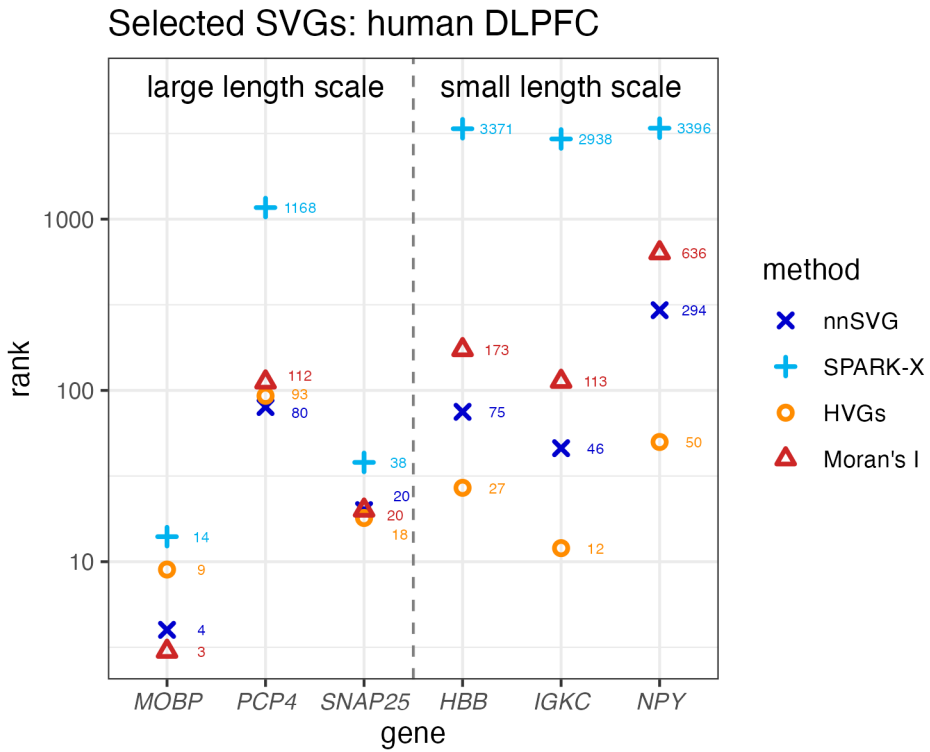
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)





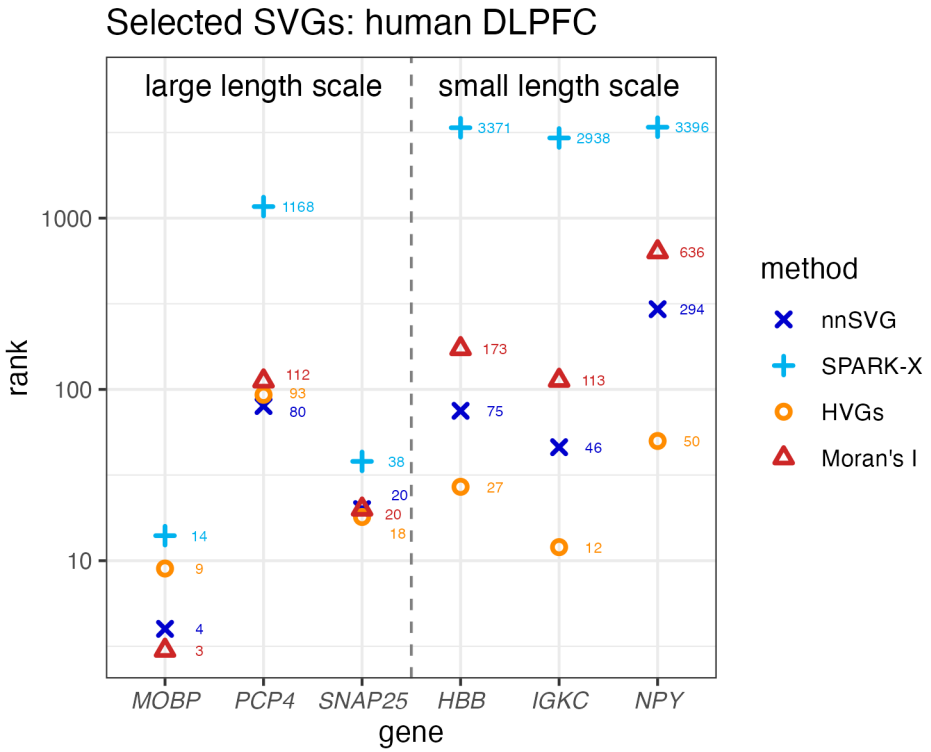
nnSVG evaluations / benchmarking

Human DLPFC dataset



nnSVG evaluations / benchmarking

Human DLPFC dataset

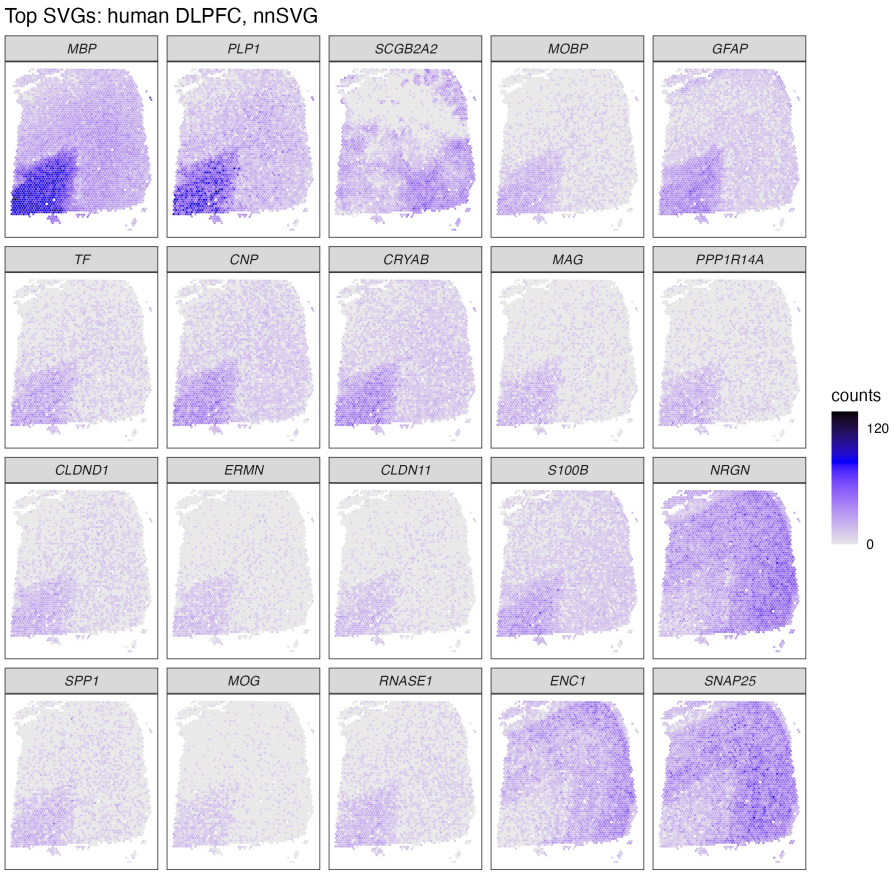


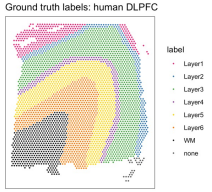
Method performance

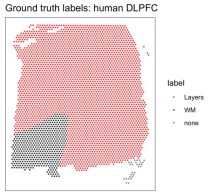
HVGs > nnSVG > Moran's I >> SPARK-X

nnSVG evaluations / benchmarking

Human DLPFC dataset



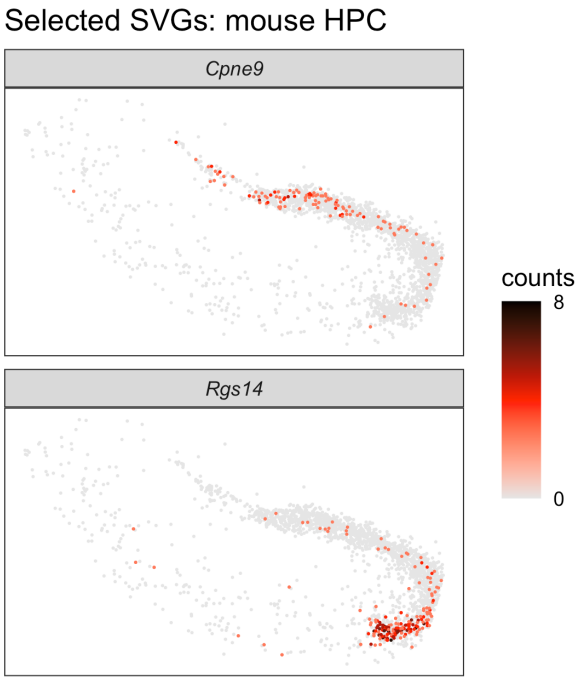


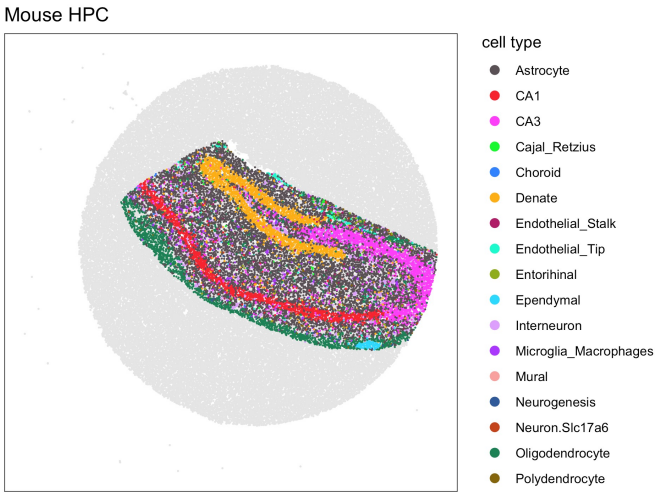


nnSVG evaluations / benchmarking

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

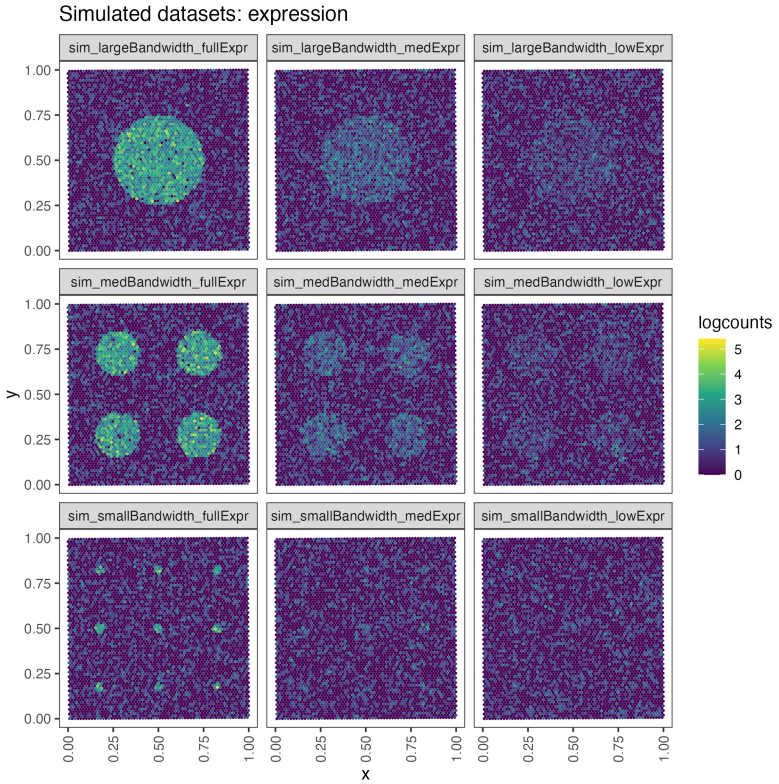
- Identify SVGs within spatial domain

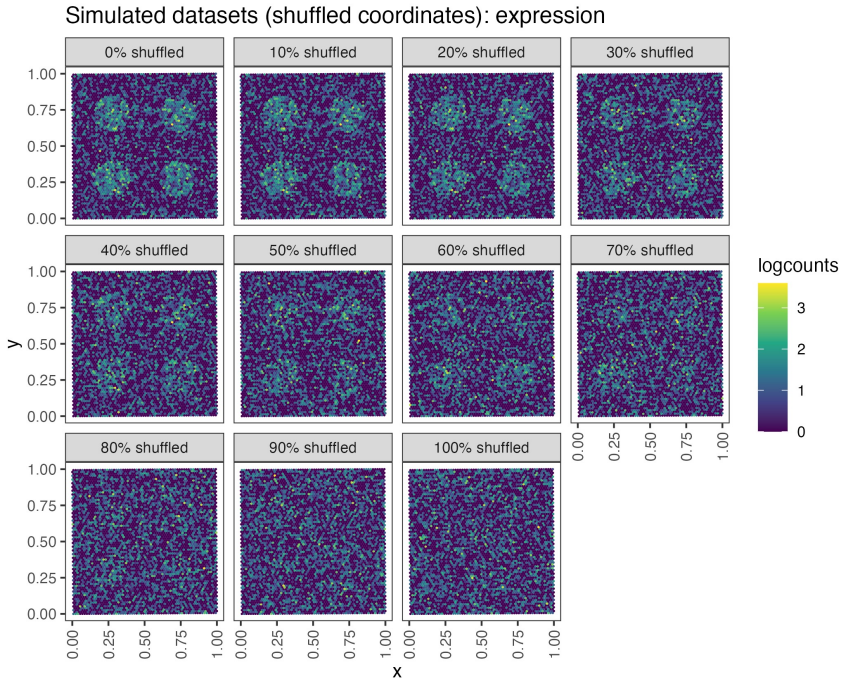




nnSVG evaluations / benchmarking

Simulations

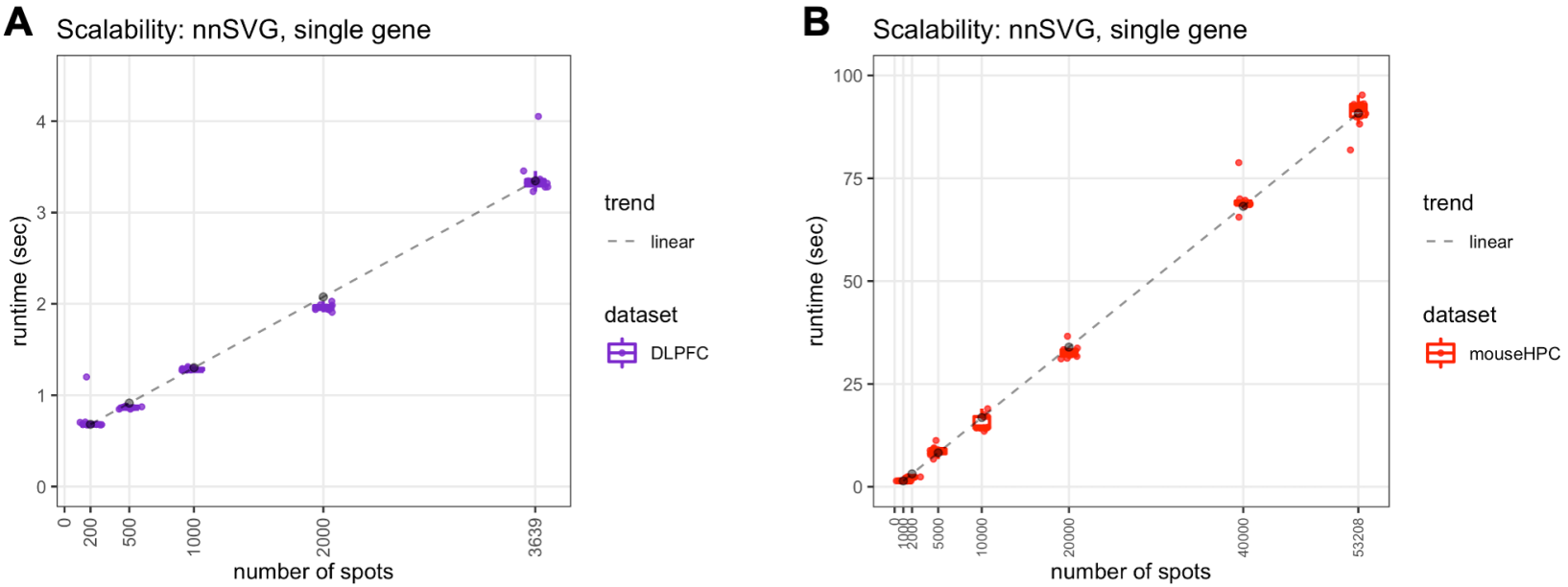




nnSVG evaluations / benchmarking

Computational scalability

- Linear in number of spatial locations

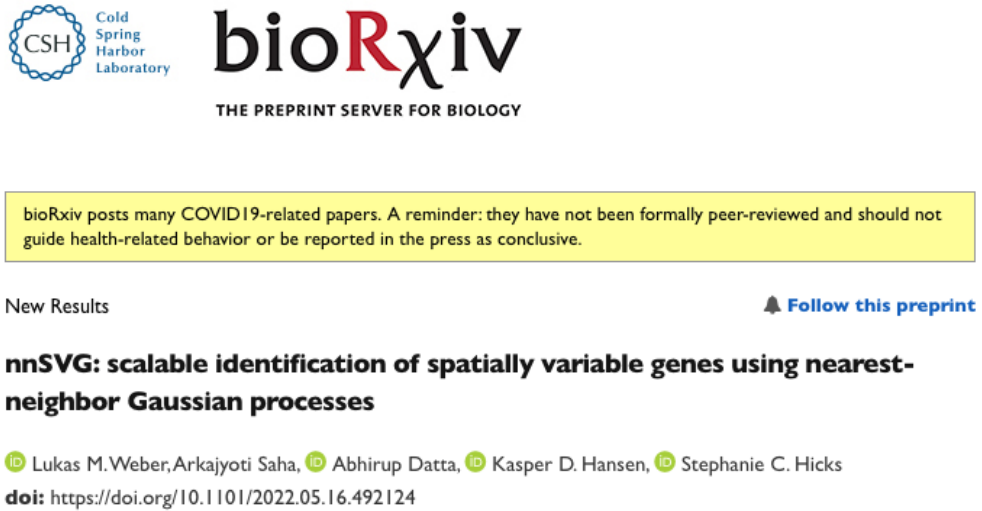


nnSVG summary

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)



nnSVG implementation

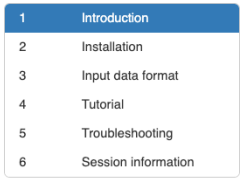
R/Bioconductor package

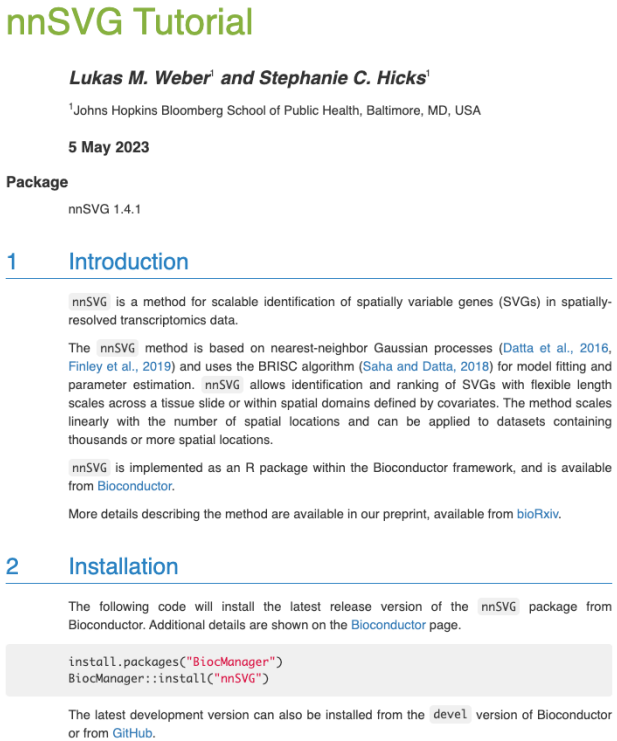
- Open-source software package
- Documentation and tutorial

nnSVG implementation

R/Bioconductor package

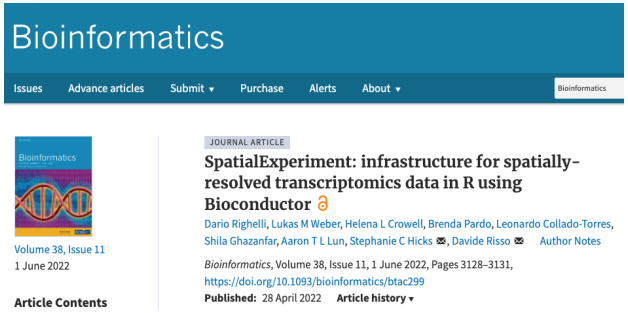
- Open-source software package
- Documentation and tutorial

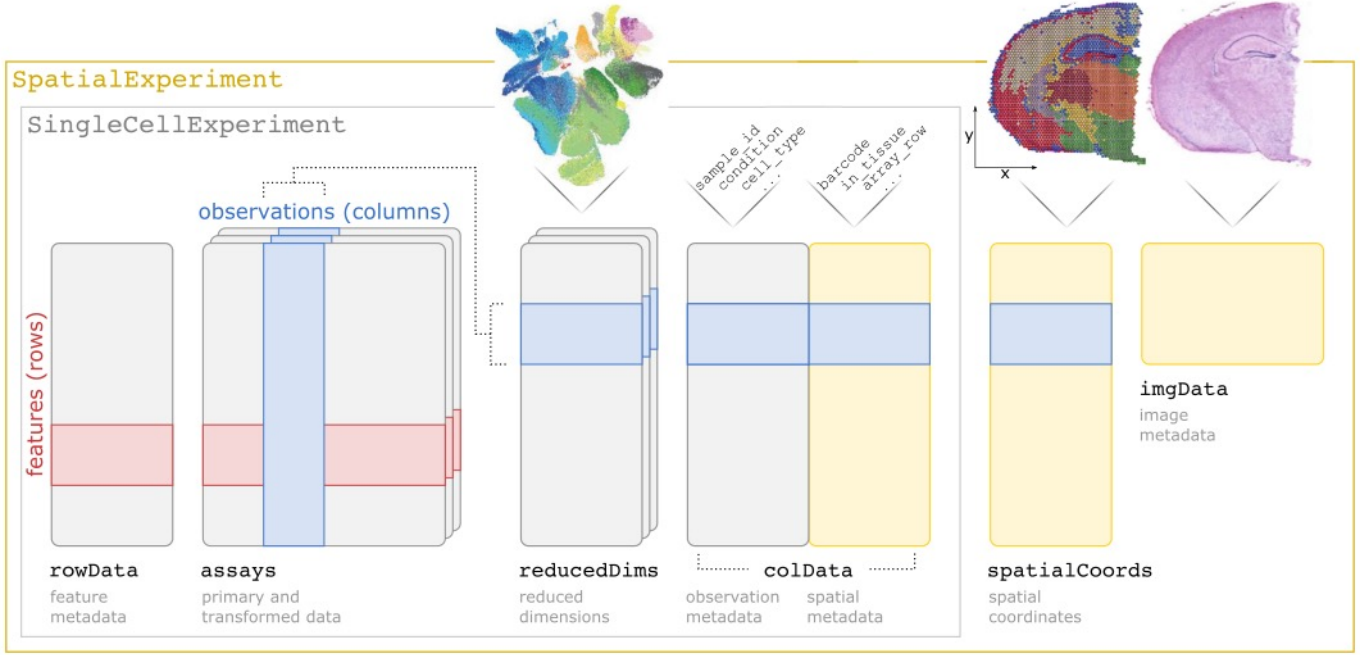




Analysis workflows and tools

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

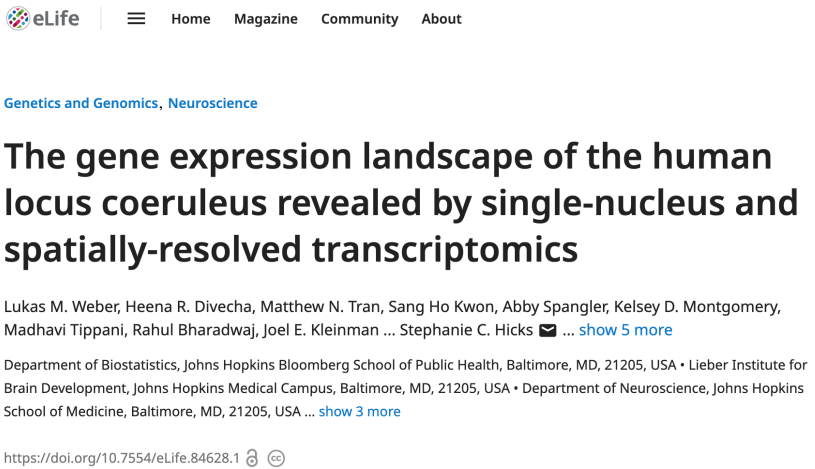




Analysis workflows and tools

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

Data and code available



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

eLife Home Magazine Community About

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

Bioconductor OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

Home Install Help

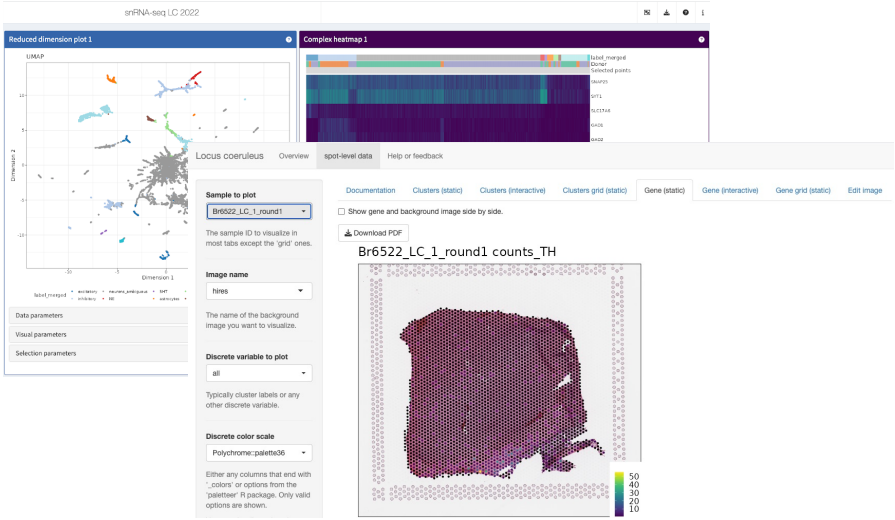
Home » Bioconductor 3.16 » Experiment Packages » WeberDivechaLCdata

WeberDivechaLCdata

platforms all | rank 415 / 416 | support 0 / 0 | build ok

updated < 1 month | dependencies 134

DOI: 10.18129/B9.bioc.WeberDivechaLCdata

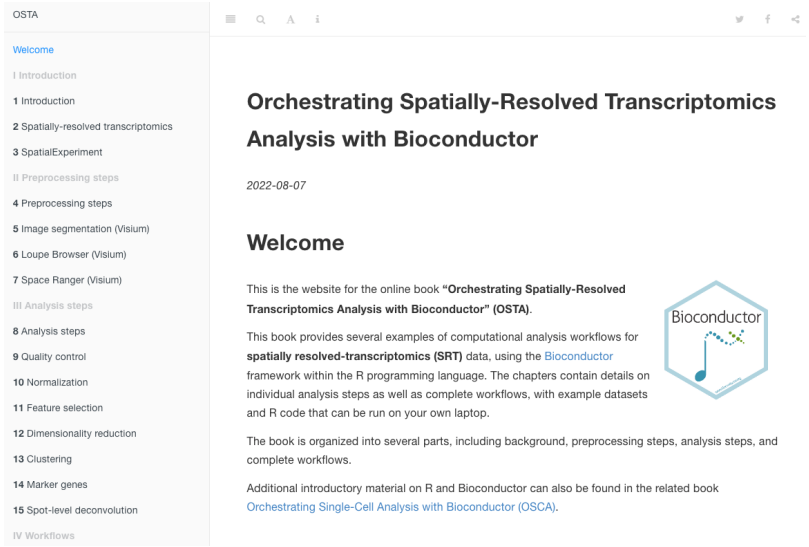


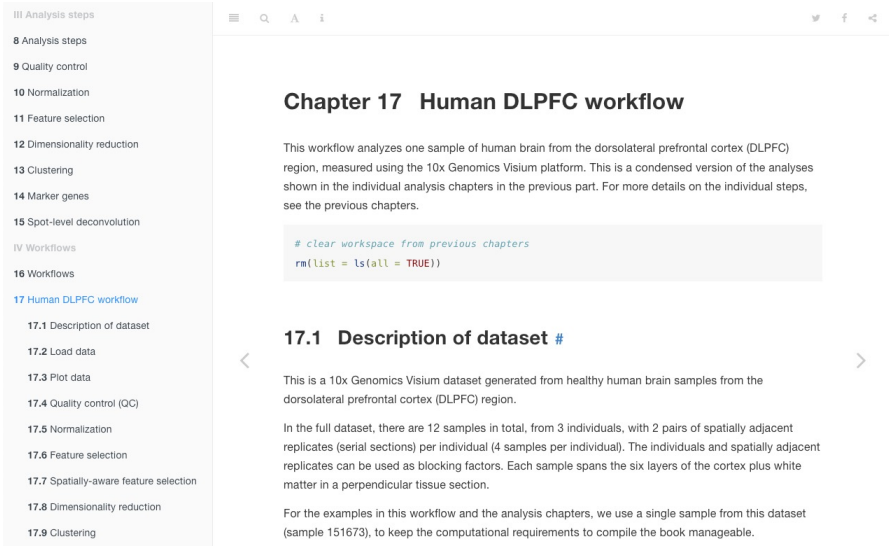
Analysis workflows and tools

Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor (OSTA)

Online book containing example R code and datasets

(Work in progress)

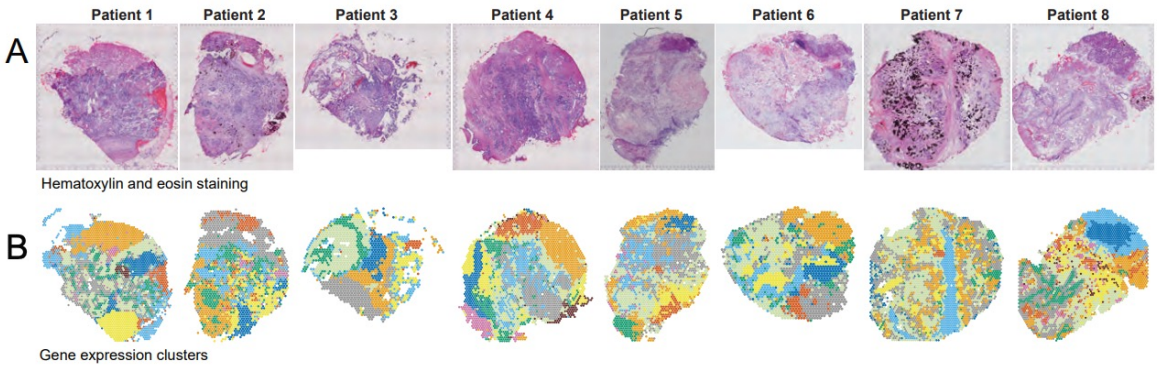




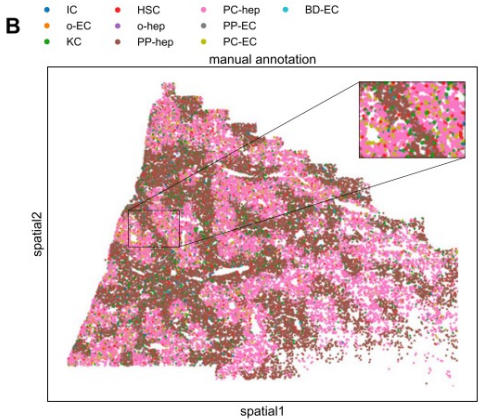
Next steps

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

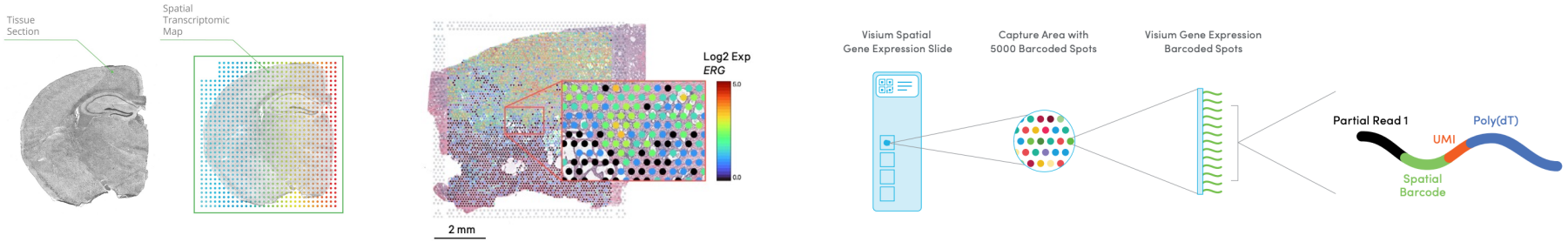


Image source: 10x Genomics

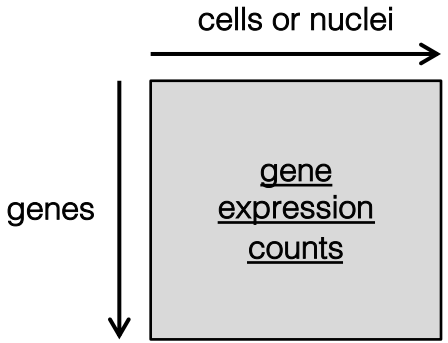
Illustration: cell populations within brain

bulk RNA sequencing → single-cell RNA sequencing → spatially-resolved transcriptomics

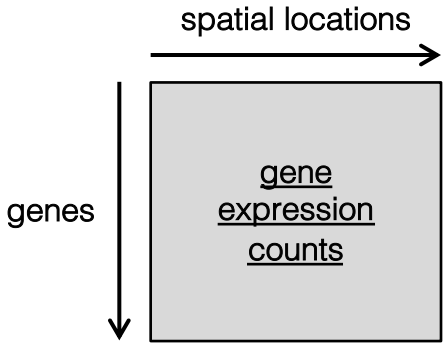
Image credit: Bo Xia
https://twitter.com/BoXia7

Data structure

Single-cell / single-nucleus RNA sequencing



Spatially-resolved transcriptomics



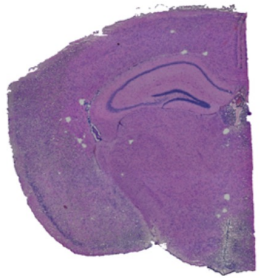
+

spatial coordinates

```
> head(spatialCoords(spe))
                        x    y
AAACAACGAATAGTTC-1 3913 2435
AAACAAGTATCTCCCA-1 9791 8468
AAACAATCTACTAGCA-1 5769 2807
AAACACCAATAACTGC-1 4068 9505
AAACAGAGCGACTCCT-1 9271 4151
AAACAGCTTTCAGAAG-1 3393 7583
```

+

image features



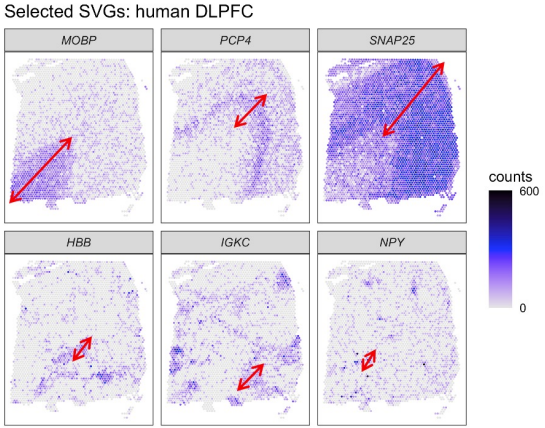
Existing methods are too slow or perform poorly

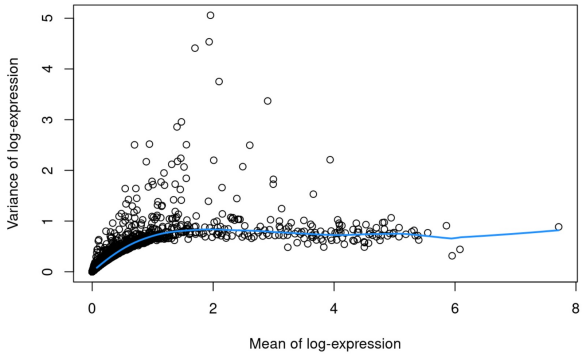
Baseline methods

- 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



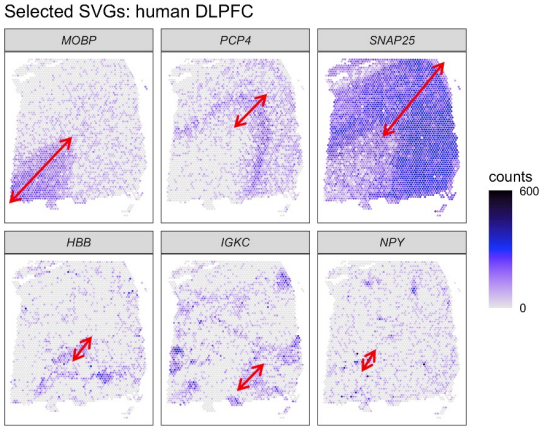


Orchestrating Single-Cell Analysis with Bioconductor
https://bioconductor.org/books/release/OSCA/

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

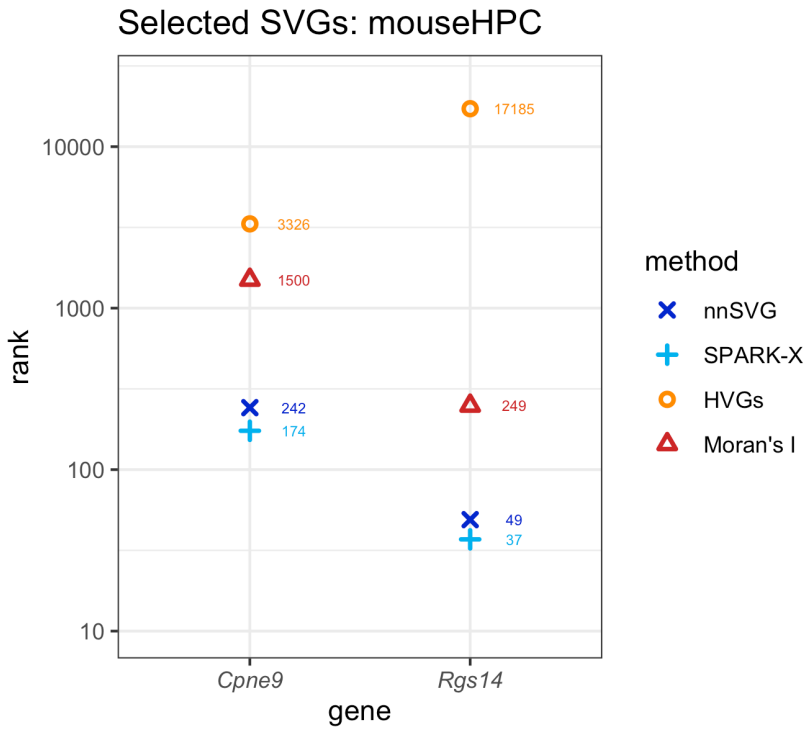
$$P(y \mid \mu, \sigma_s^2, \delta, \Sigma) = N(y \mid \mu \cdot 1, \sigma_s^2 \cdot (\Sigma + \delta \cdot I))$$

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

nnSVG evaluations / benchmarking

Mouse hippocampus dataset

- With covariates for spatial domains

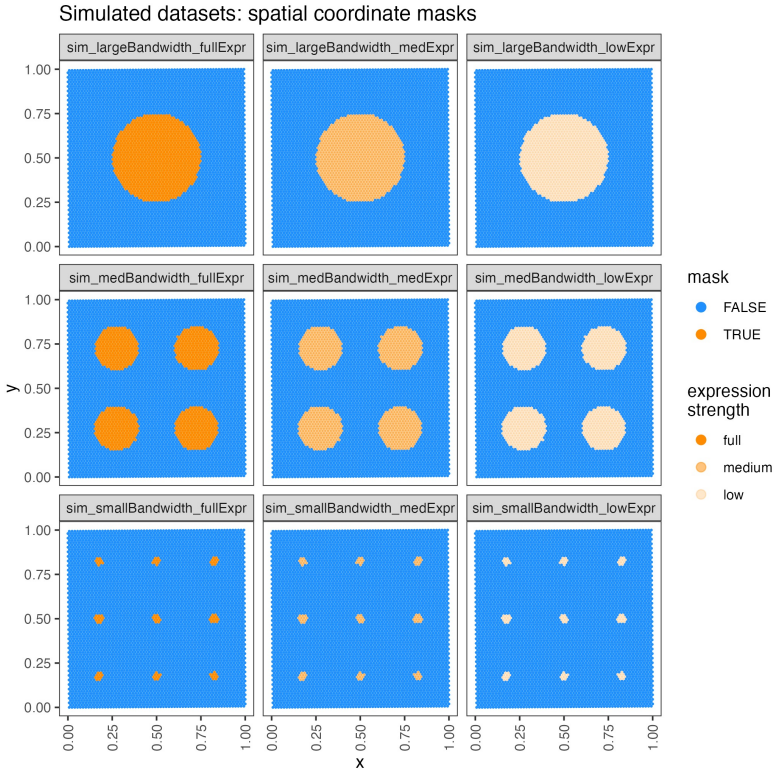


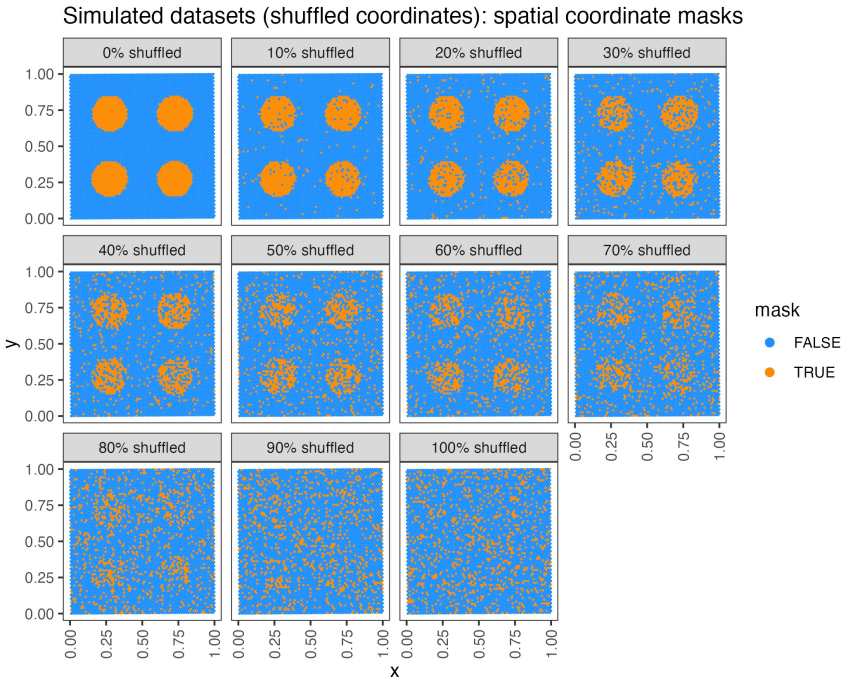
Method performance

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

nnSVG evaluations / benchmarking

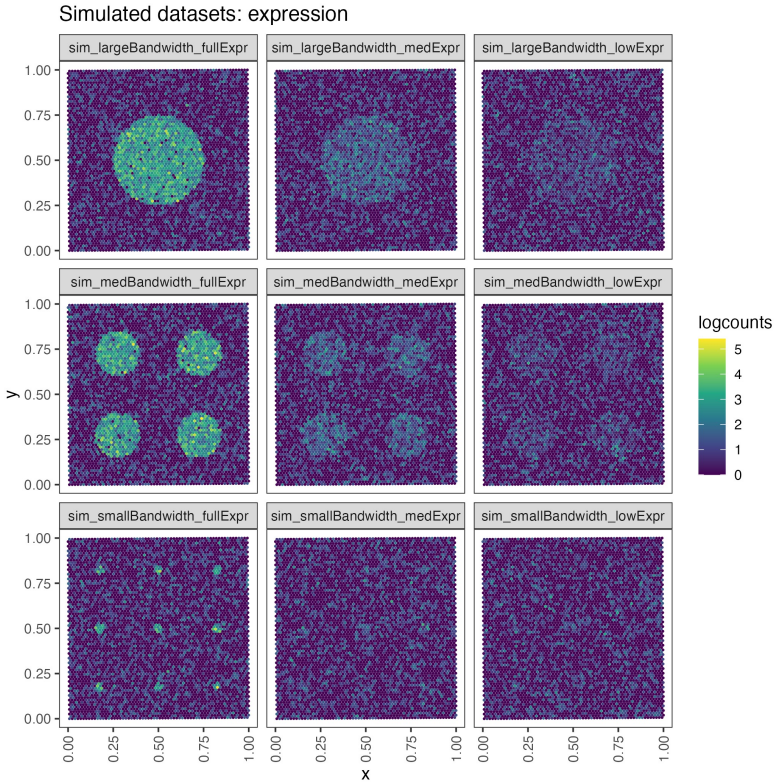
Simulations

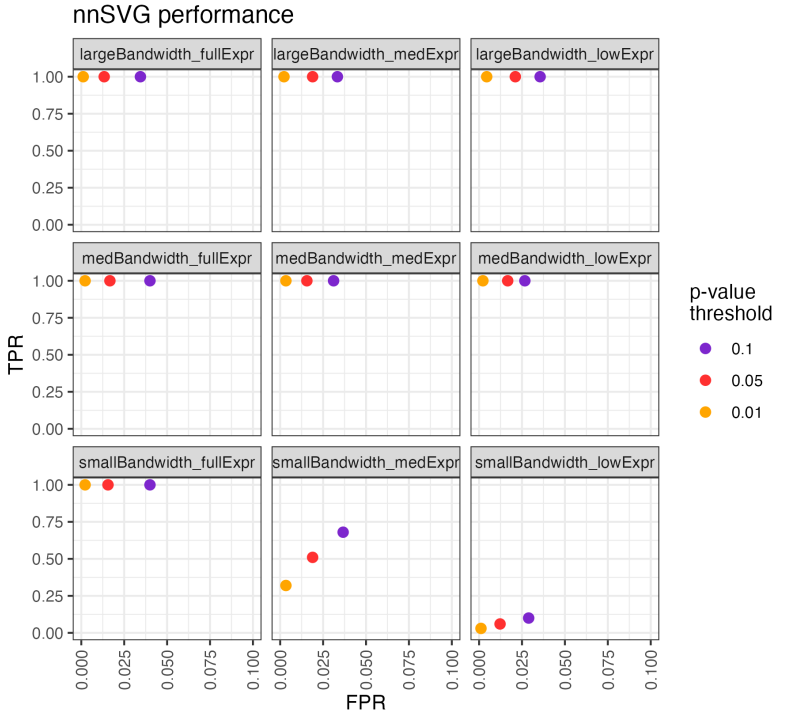




nnSVG evaluations / benchmarking

Simulations





nnSVG evaluations / benchmarking

Simulations