

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

16 February 2023

Short bio



Applied biostatistician working on methodological development and collaborative analyses for high-throughput genomic data with focus on single-cell and spatially-resolved transcriptomics

Background

Postdoctoral Fellow (since 2019)

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

PhD in Biostatistics, University of Zurich, Switzerland (2019)

MSc in Statistics, ETH Zurich, Switzerland (2014)



ETH zürich

lmweber.org

github.com/lmweber

[@lmwebr](https://twitter.com/lmwebr)

Research Theme 1

Unsupervised statistical methods

Spatially-resolved transcriptomics: [nnSVG](#)

- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: [diffcyt](#)

- Weber et al. (2019), *Comms Biol*

1

Research Theme 2

Collaborative analyses

Neuroscience

- Weber and Divecha et al. (2022), *bioRxiv / in revision (eLife)*
- Maynard and Collado-Torres et al. (2021), *Nat Neur*

Cancer

Immunology

2

Methodological development and collaborative analyses for high-throughput genomic data

Technological platforms: spatially-resolved transcriptomics, single-cell / single-nucleus RNA sequencing, high-dimensional cytometry

K99/R00 Award

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data
(1K99HG012229-01)



Research Theme 3

Benchmarking

- Cancer: Weber et al. (2021), *GigaScience*
- Review / guidelines: Weber et al. (2019), *Genome Biol*
- Independent benchmarking: Weber et al. (2016), *Cyt Part A*

Analysis workflows and tools

- R/Bioconductor: Righelli, Weber, Crowell et al. (2021), *Bioinf*

Open-source software / reproducible research

additional papers
on website and
Google Scholar



Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

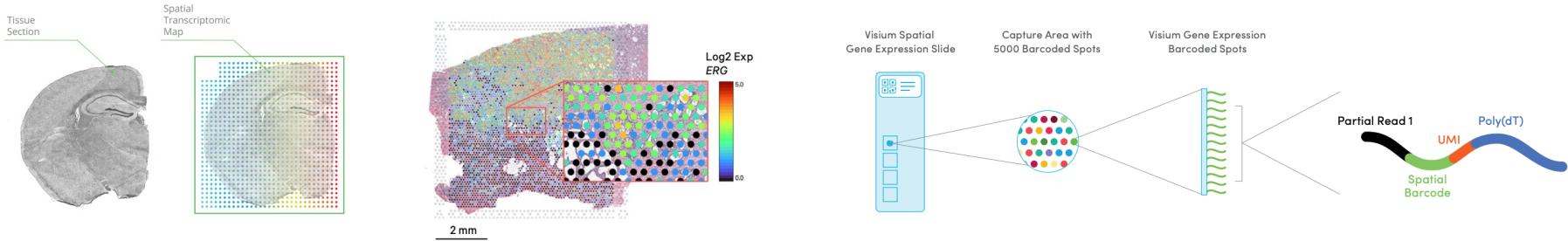


Image source: 10x Genomics

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

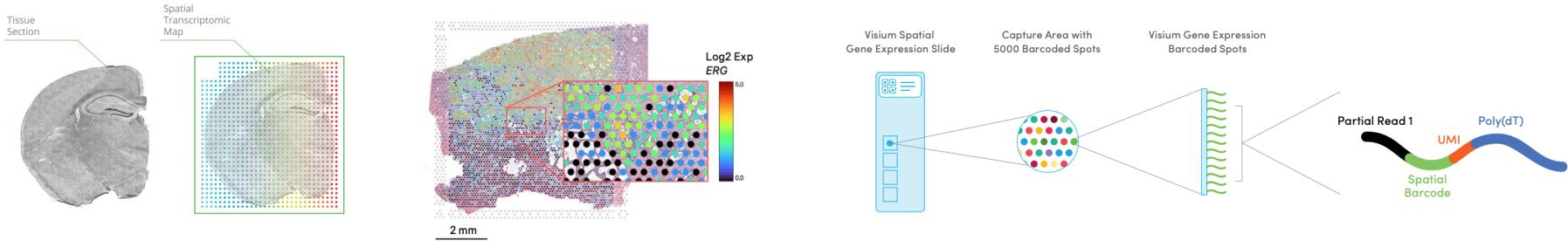


Image source: 10x Genomics

Illustration: cell populations within brain

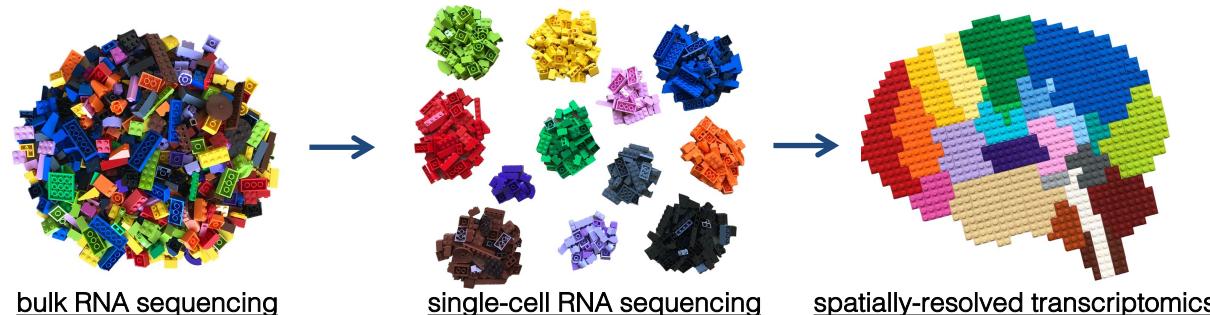


Image credit: Bo Xia
<https://twitter.com/BoXia7>

Spatially-resolved transcriptomics

Unsupervised / discovery-based analyses

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

Cell type / state heterogeneity

Biologically informative / marker genes

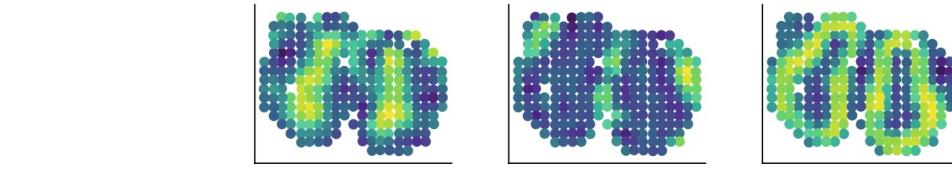
Spatially-resolved transcriptomics

Unsupervised / discovery-based analyses

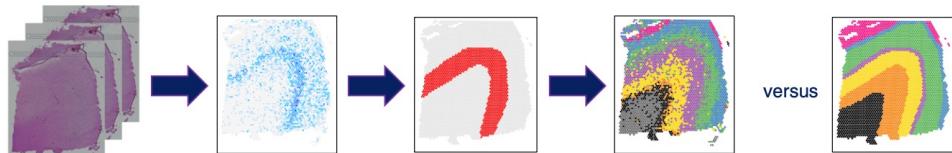
- feature selection: spatially variable genes
- clustering: spatial domains or spatially distributed cell populations
- differential gene expression

Cell type / state heterogeneity

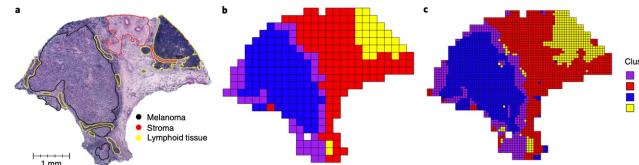
Biologically informative / marker genes



Svensson et al. (2018)



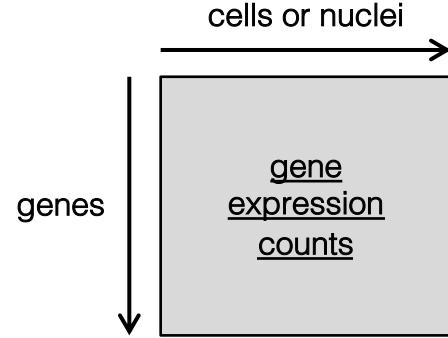
Maynard and Collado-Torres et al. (2021)



Zhao et al. (2021)

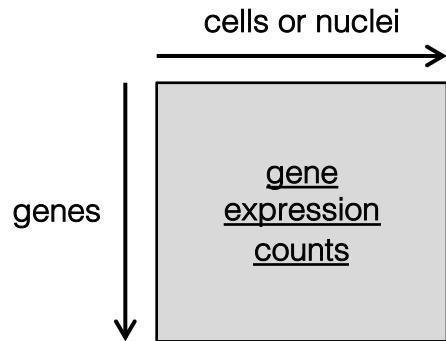
Data structure

Single-cell /
single-nucleus
RNA sequencing

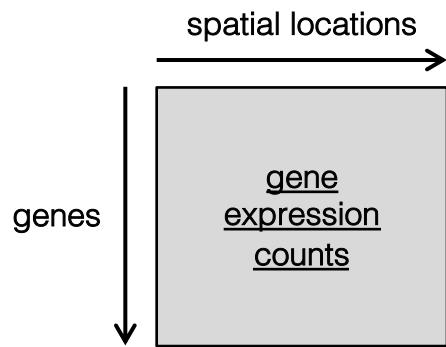


Data structure

Single-cell /
single-nucleus
RNA sequencing



Spatially-resolved
transcriptomics



+

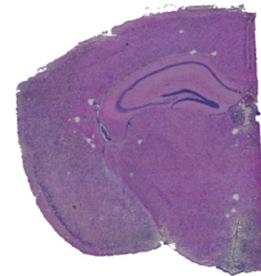
spatial
coordinates

> 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



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)
(my role: unsupervised analysis workflow)



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^{1,6}, 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^{1,9}, Keri Martinowich^{1,2,6} and Andrew E. Jaffe^{1,2,3,4,6,7,9}

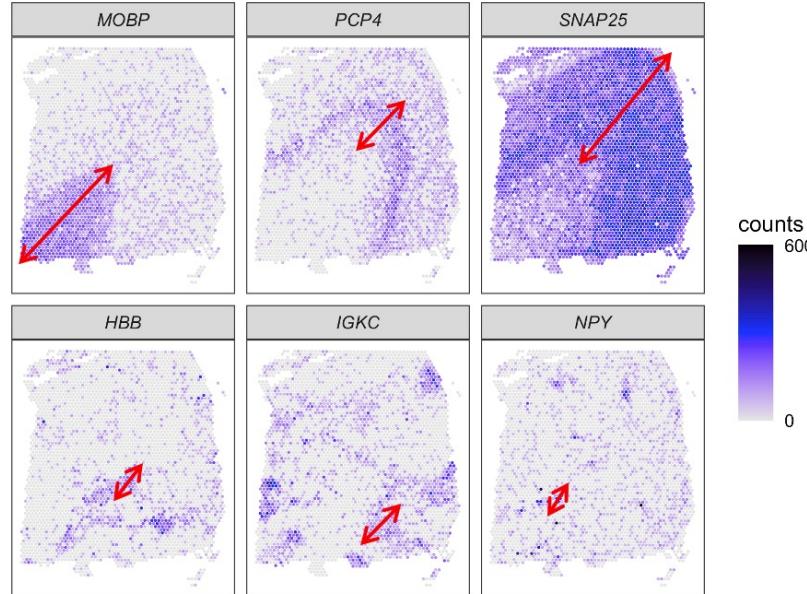
Spatially variable genes

Example

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



Selected SVGs: human DLPFC



Spatially variable genes

Example

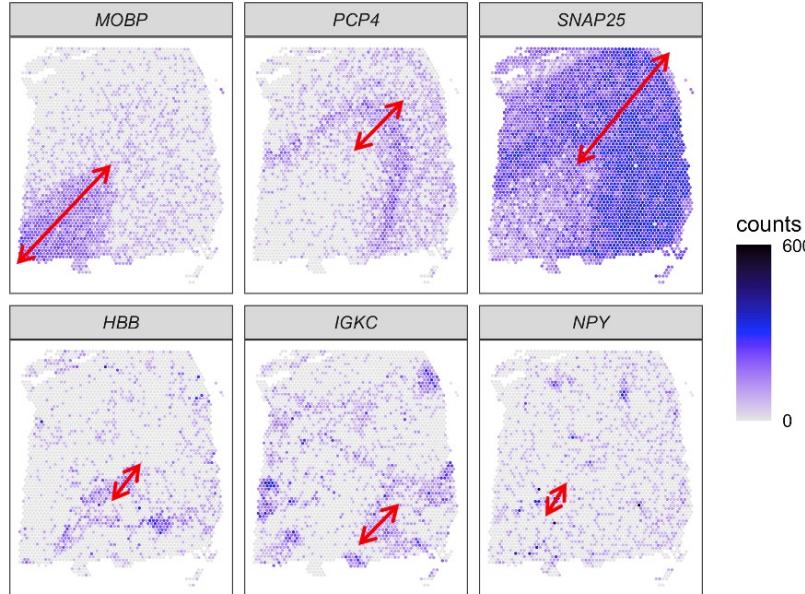
- Dorsolateral prefrontal cortex (DLPFC) in postmortem human brain samples measured with 10x Genomics Visium platform
- Maynard and Collado-Torres et al. (2021) (my role: unsupervised analysis workflow)



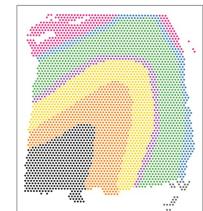
Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex

Kristen R. Maynard^{1,2,10}, Leonardo Collado-Torres^{1,3,10}, Lukas M. Weber⁴, Cedric Uytingco⁵, Brianna K. Barry^{1,6}, 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}

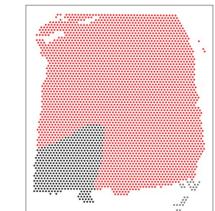
Selected SVGs: human DLPFC



Ground truth labels: human DLPFC



Ground truth labels: human DLPFC



Research Theme 1

Unsupervised statistical methods

Spatially-resolved transcriptomics: [nnSVG](#)

- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: [diffcyt](#)

- Weber et al. (2019), *Comms Biol*

1

Research Theme 2

Collaborative analyses

Neuroscience

- Weber and Divecha et al. (2022), *bioRxiv / in revision (eLife)*
- Maynard and Collado-Torres et al. (2021), *Nat Neur*

Cancer

Immunology

2

Methodological development and collaborative analyses for high-throughput genomic data

Technological platforms: spatially-resolved transcriptomics, single-cell / single-nucleus RNA sequencing, high-dimensional cytometry

K99/R00 Award

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data
(1K99HG012229-01)



Research Theme 3

Benchmarking

- Cancer: Weber et al. (2021), *GigaScience*
- Review / guidelines: Weber et al. (2019), *Genome Biol*
- Independent benchmarking: Weber et al. (2016), *Cyt Part A*

Analysis workflows and tools

- R/Bioconductor: Righelli, Weber, Crowell et al. (2021), *Bioinf*

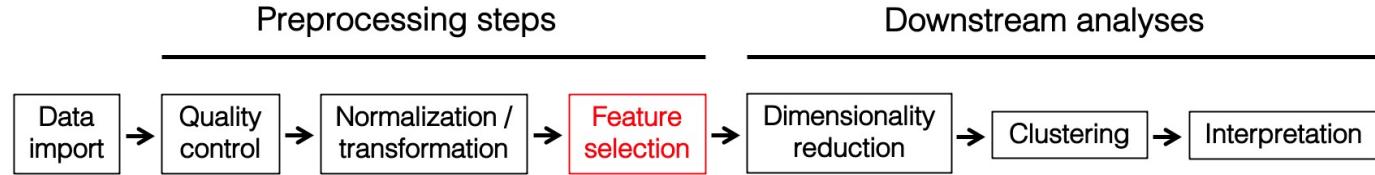
Open-source software / reproducible research

additional papers
on website and
Google Scholar



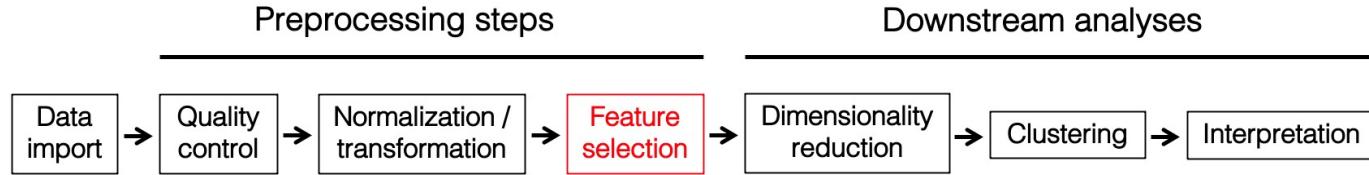
Why are we interested in spatially variable genes?

Feature selection



Why are we interested in spatially variable genes?

Feature selection



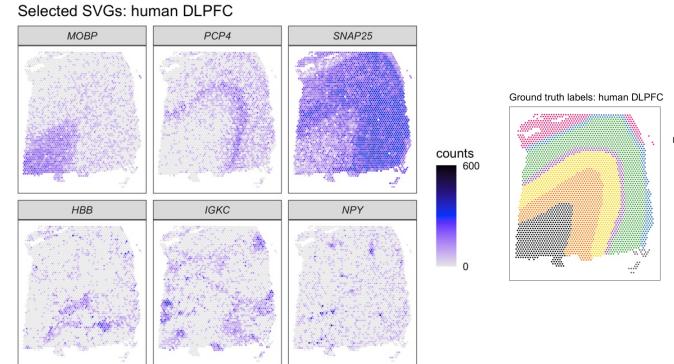
Data preprocessing

- 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

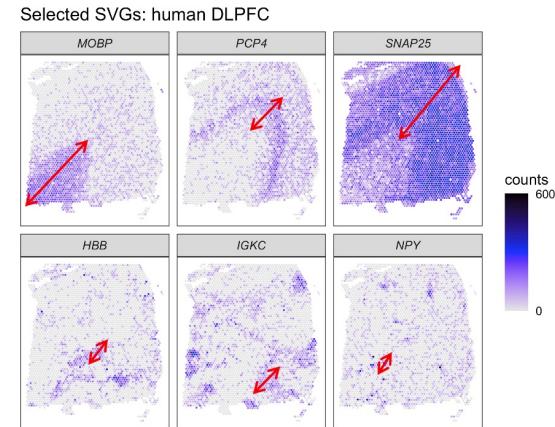
Example: human DLPFC



Existing methods are too slow or perform poorly

Baseline methods

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



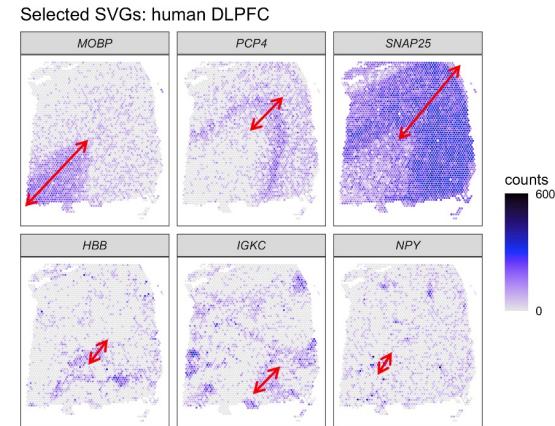
Existing methods are too slow or perform poorly

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
- **SPARK-X** (Zhu et al. 2021)
Fast approximation loses sensitivity to spatial patterns
with varying length scales



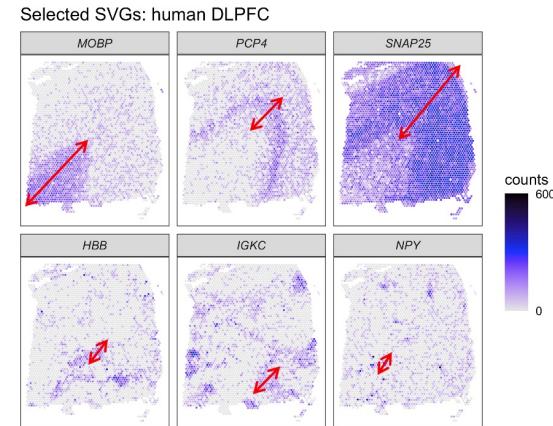
Existing methods are too slow or perform poorly

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 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)

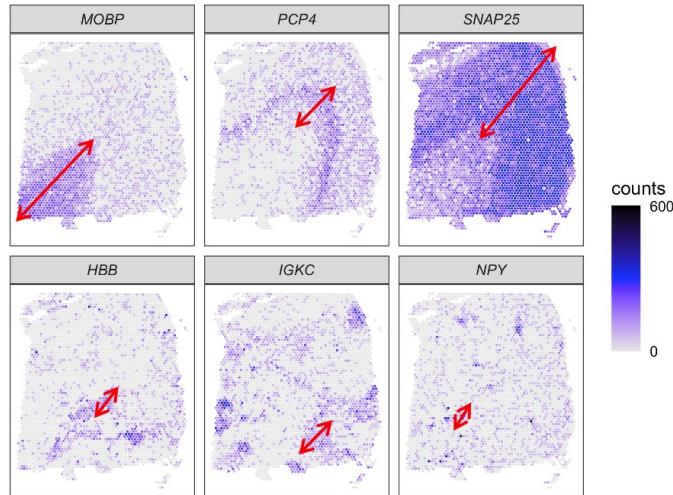
$$y \sim N(X\beta, C(\theta) + \tau^2 I)$$

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

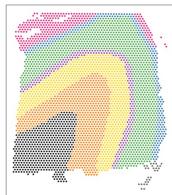
$$propSV = \frac{\sigma^2}{\sigma^2 + \tau^2}$$

nnSVG evaluations / benchmarking

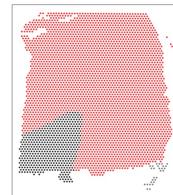
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)



Ground truth labels: human DLPFC

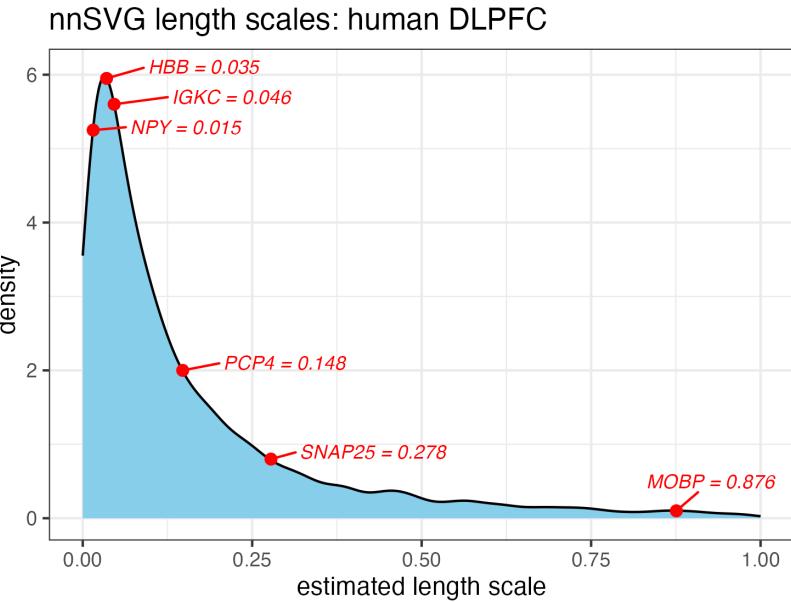
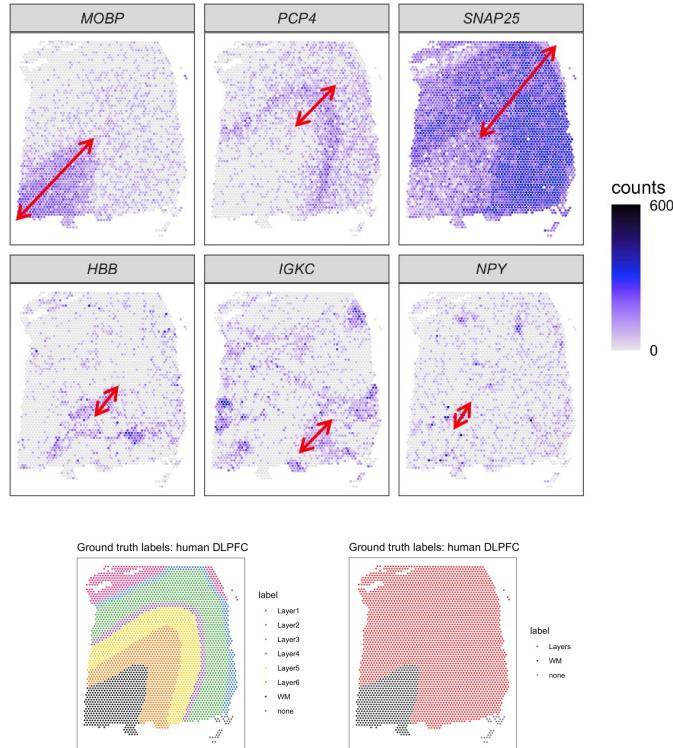


Ground truth labels: human DLPFC



nnSVG evaluations / benchmarking

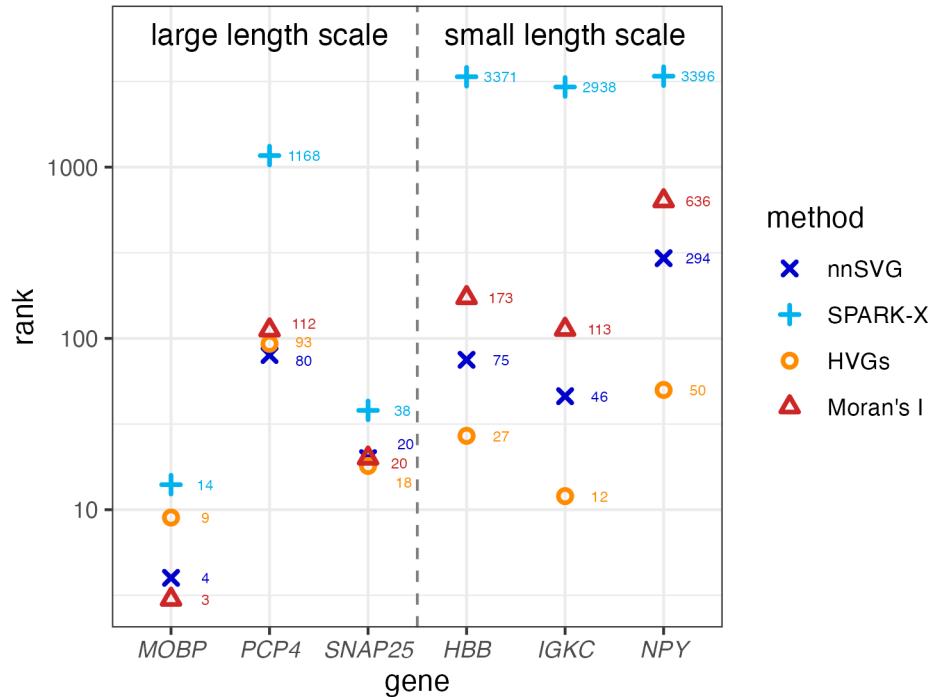
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)



nnSVG evaluations / benchmarking

Human DLPFC dataset

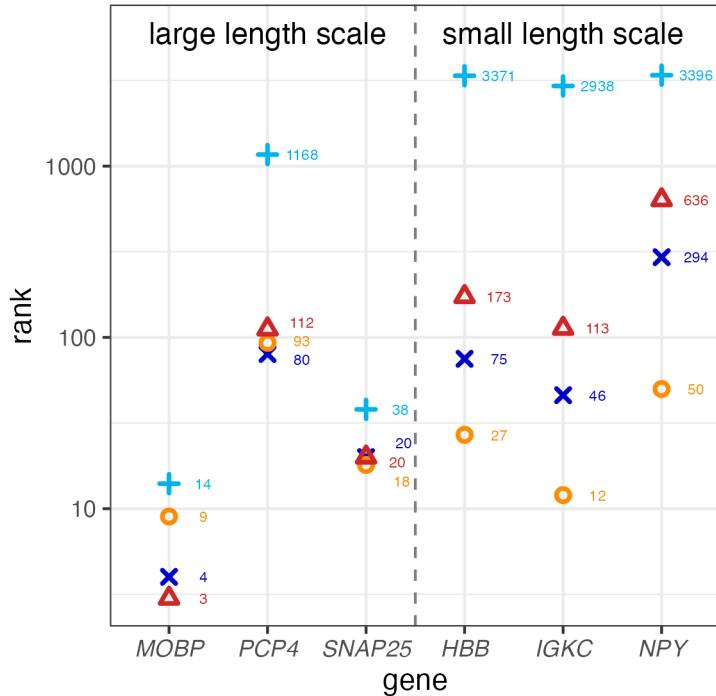
Selected SVGs: human DLPFC



nnSVG evaluations / benchmarking

Human DLPFC dataset

Selected SVGs: human DLPFC



Method performance

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

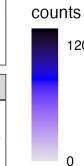
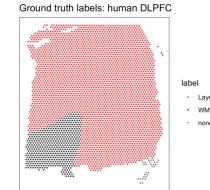
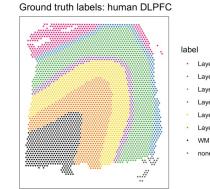
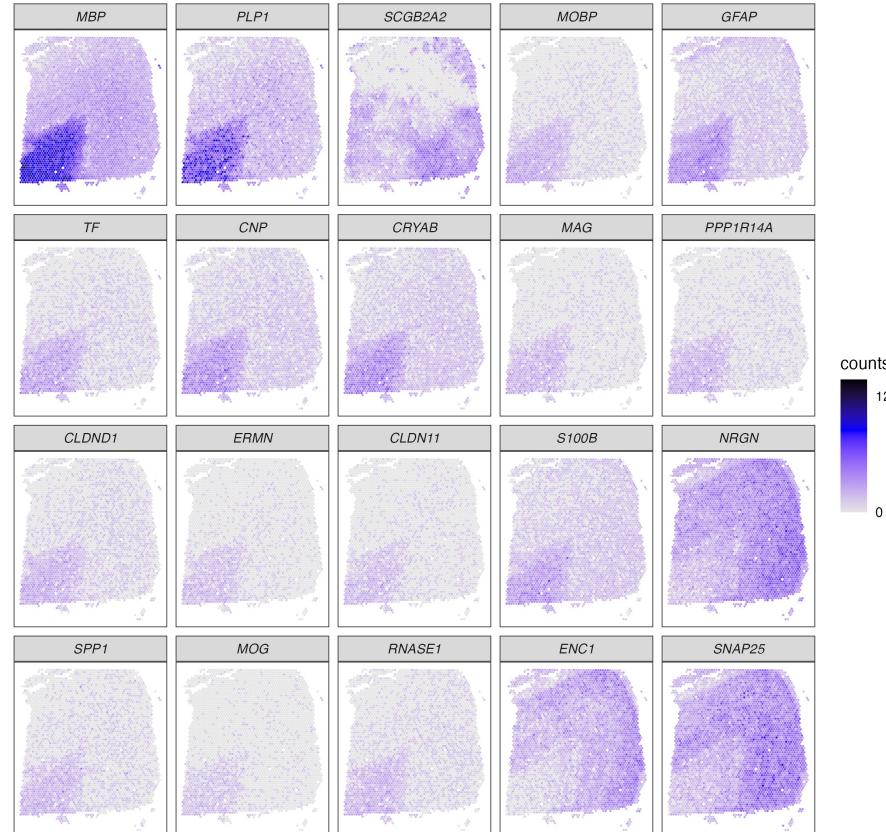
method

- 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

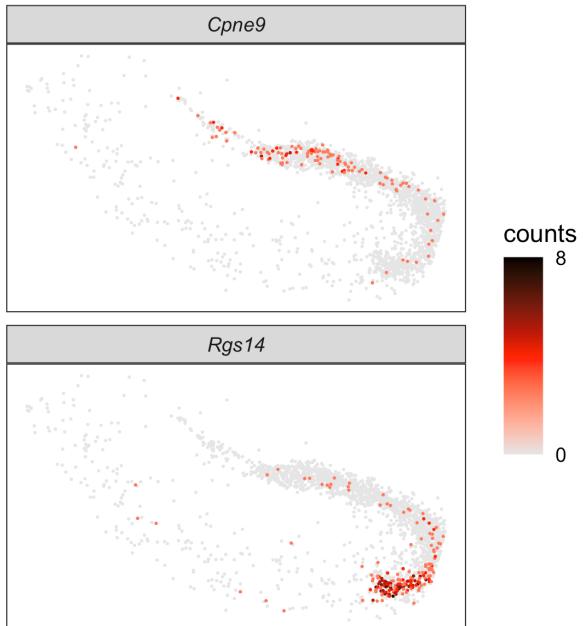


nnSVG evaluations / benchmarking

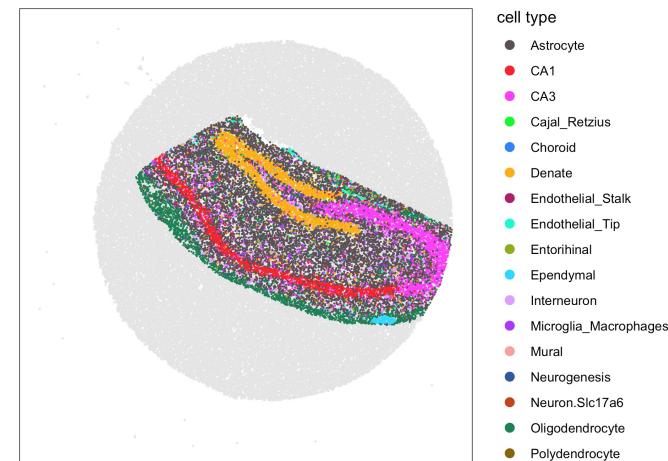
Mouse hippocampus dataset (Slide-seqV2) (Stickels et al. 2020 / Cable et al. 2021)

- Identify SVGs within spatial domain

Selected SVGs: mouse HPC



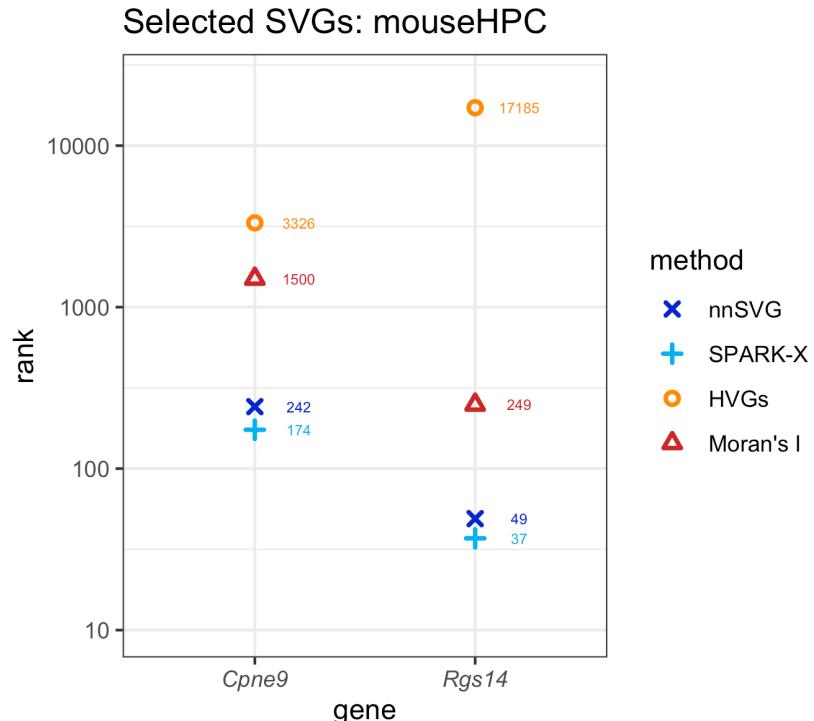
Mouse HPC



nnSVG evaluations / benchmarking

Mouse hippocampus dataset

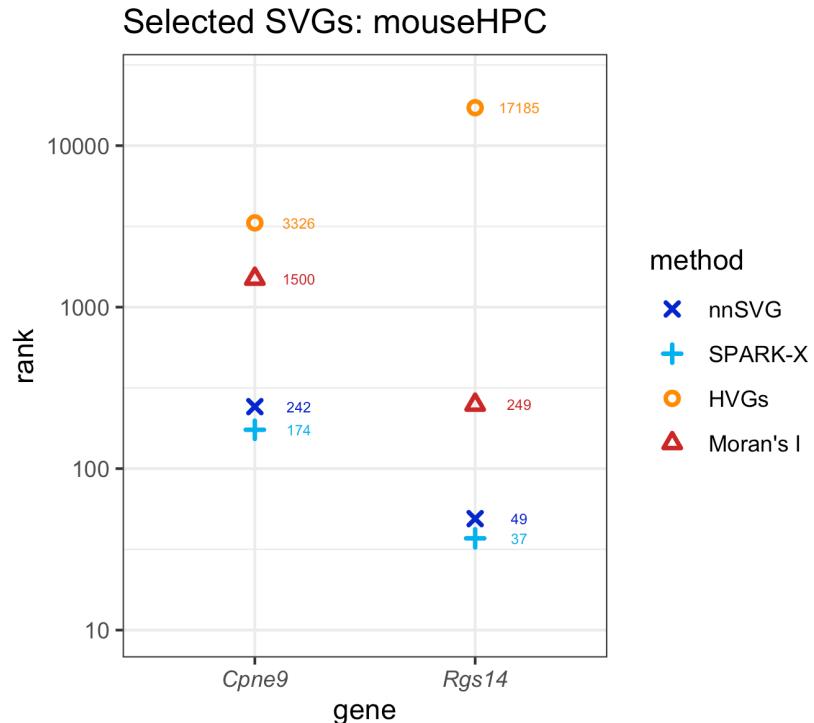
- With covariates for spatial domains



nnSVG evaluations / benchmarking

Mouse hippocampus dataset

- With covariates for spatial domains



Method performance

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

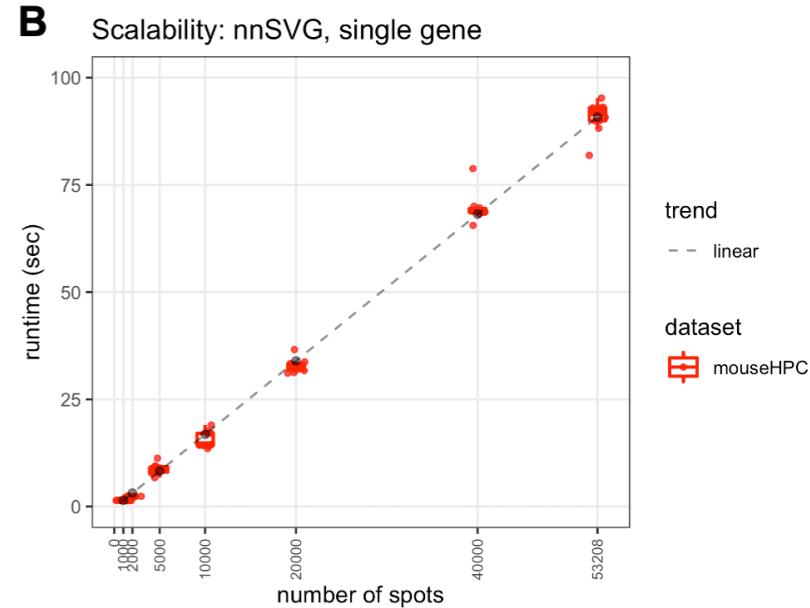
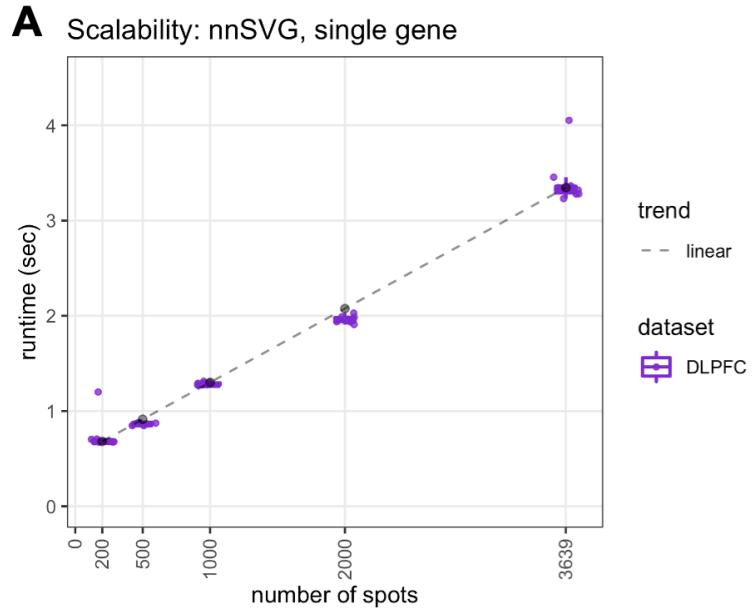
method

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

nnSVG evaluations / benchmarking

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

Weber et al. (2022)

bioRxiv preprint; in revision (Nat Comm)



THE PREPRINT SERVER FOR BIOLOGY

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 and reproducibility

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)

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 and reproducibility

R/Bioconductor package

- Open-source software package
- Documentation and tutorial

The screenshot shows the Bioconductor website with the nnSVG package page. At the top, there's a navigation bar with the Bioconductor logo, Home, Install, and Help buttons. Below the navigation, the URL is Home > Bioconductor 3.16 > Software Packages > nnSVG. The main content area has a green header "nnSVG". Below it, there's a summary table with metrics: platforms (all), rank (1940 / 2183), support (0 / 0), build (ok), updated (before release), dependencies (100). A DOI link (10.18129/B9.bioc.nnSVG) and social media links (Facebook, Twitter) are also present. A descriptive paragraph about the package's purpose follows.

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.

GitHub repository

- Code to reproduce analyses and figures
- Data availability (STexampleData package)

The screenshot shows the GitHub repository page for nnSVG-analyses. The repository is public, created by Imweber. It has 502 commits, 1 branch, and 0 tags. The code tab is selected, showing a list of recent commits:

Commit	Message	Date
main	boxplot of ranks for small vs large length scales	yesterday
branch	add layer-specific marker genes.csv	last year
.gitignore	gitignore log files	last week
README.md	update readme	last week
nnSVG-analyses.Rproj	rename forked repo	10 months ago

On the right side, there are sections for About (scripts to reproduce analyses in nnSVG paper), Releases (no releases published, Create a new release), and Packages (no packages published, Publish your first package).

nnSVG-analyses

This repository contains code scripts to reproduce analyses and figures in our paper "nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes".

Research Theme 1

Unsupervised statistical methods

Spatially-resolved transcriptomics: [nnSVG](#)

- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: [diffcyt](#)

- Weber et al. (2019), *Comms Biol*

1

Research Theme 2

Collaborative analyses

Neuroscience

- Weber and Divecha et al. (2022), *bioRxiv / in revision (eLife)*
- Maynard and Collado-Torres et al. (2021), *Nat Neur*

Cancer

Immunology

2

Methodological development and collaborative analyses for high-throughput genomic data

Technological platforms: spatially-resolved transcriptomics, single-cell / single-nucleus RNA sequencing, high-dimensional cytometry

K99/R00 Award

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data
(1K99HG012229-01)



Research Theme 3

Benchmarking

- Cancer: Weber et al. (2021), *GigaScience*
- Review / guidelines: Weber et al. (2019), *Genome Biol*
- Independent benchmarking: Weber et al. (2016), *Cyt Part A*

Analysis workflows and tools

- R/Bioconductor: Righelli, Weber, Crowell et al. (2021), *Bioinf*

Open-source software / reproducible research

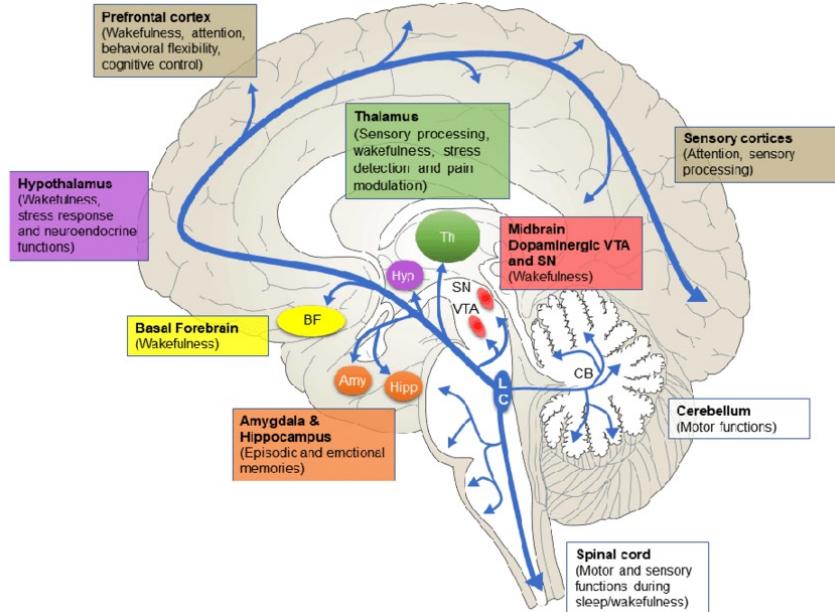
additional papers
on website and
Google Scholar



Locus coeruleus (LC)

Background

- Norepinephrine (NE) neurons within LC project widely throughout central nervous system
- Critical roles in arousal, mood, components of cognition including attention, learning, memory
- Implicated in neurological and neuropsychiatric disorders, sensitive to degeneration in Alzheimer's and Parkinson's



Study overview

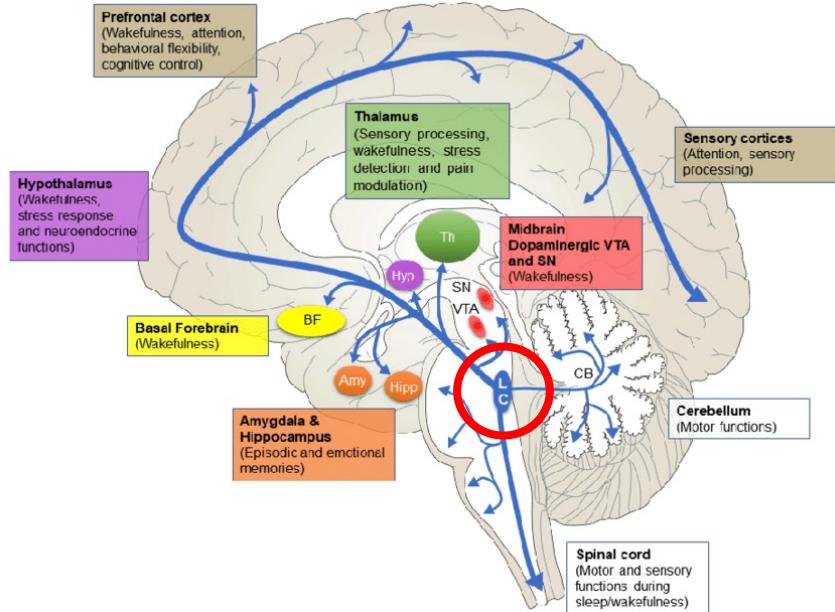
- LC is relatively understudied in humans due to small size and inaccessibility within brainstem
- Characterize transcriptome-wide gene expression landscape of human LC at spatial and single-nucleus resolution

Bari et al. (2020),
Neural Regen Res

Locus coeruleus (LC)

Background

- Norepinephrine (NE) neurons within LC project widely throughout central nervous system
- Critical roles in arousal, mood, components of cognition including attention, learning, memory
- Implicated in neurological and neuropsychiatric disorders, sensitive to degeneration in Alzheimer's and Parkinson's



Study overview

- LC is relatively understudied in humans due to small size and inaccessibility within brainstem
- Characterize transcriptome-wide gene expression landscape of human LC at spatial and single-nucleus resolution

Bari et al. (2020),
Neural Regen Res

Study overview

Aim

- Characterize transcriptome-wide gene expression landscape of human LC at spatial and single-nucleus resolution

Study overview

Aim

- Characterize transcriptome-wide gene expression landscape of human LC at spatial and single-nucleus resolution

Experimental design

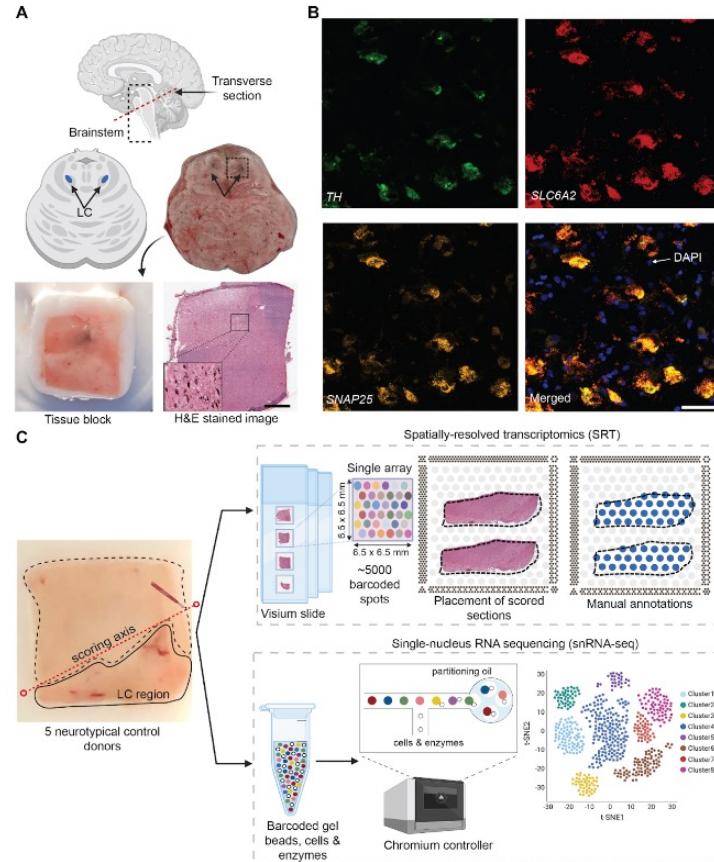
- 5 neurotypical adult human brain donors
- Spatially-resolved transcriptomics: 8 samples from 4 donors (after quality control) using 10x Genomics Visium platform
- Single-nucleus RNA sequencing: 20,191 nuclei from 3 donors (after quality control) using 10x Genomics Chromium platform
- Identification of regions containing LC in Visium samples by neuroanatomical landmarks and pigmented neurons, validated by single-molecule fluorescence *in situ* hybridization (smFISH / RNAscope)

Analyses

- Unsupervised / discovery-driven analysis workflow

Heena Divecha

Lieber Institute for Brain Development



Differential gene expression

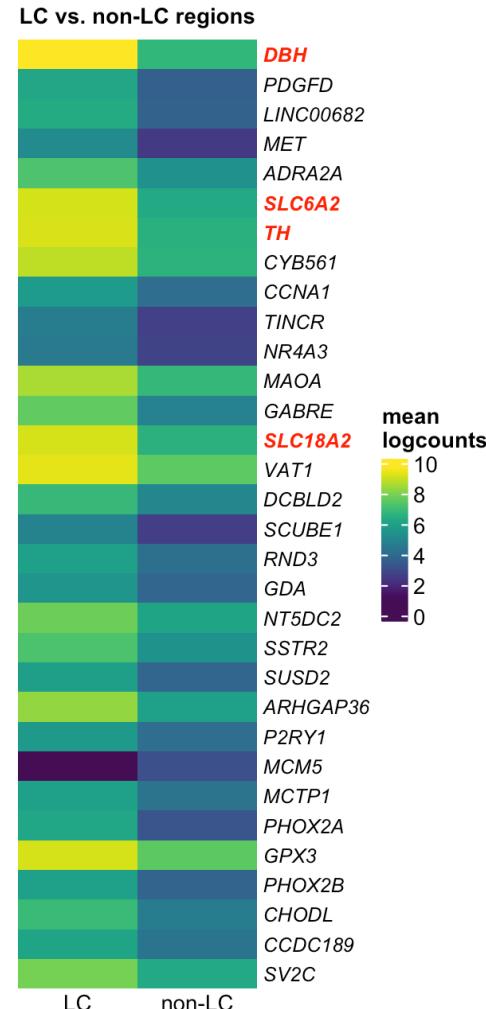
Differential gene expression testing in Visium data

- Manually annotated LC regions
- Recovers known NE neuron markers and additional unsupervised gene list
- 437 statistically significant genes

Differential gene expression

Differential gene expression testing in Visium data

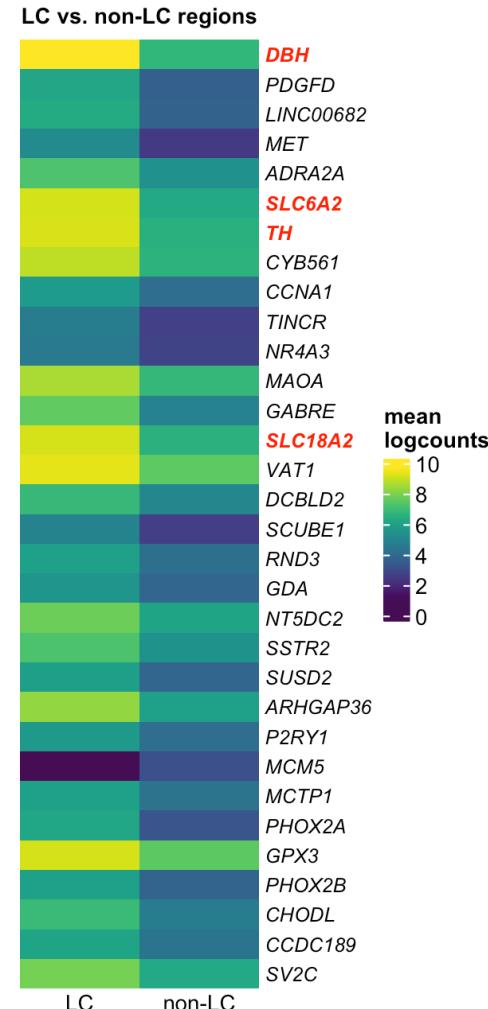
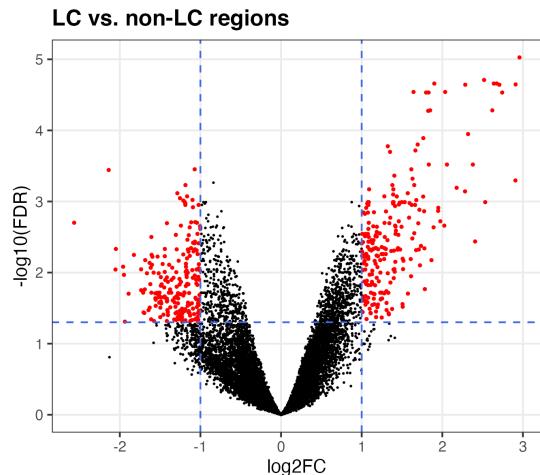
- Manually annotated LC regions
- Recovers known NE neuron markers and additional unsupervised gene list
- 437 statistically significant genes



Differential gene expression

Differential gene expression testing in Visium data

- Manually annotated LC regions
- Recovers known NE neuron markers and additional unsupervised gene list
- 437 statistically significant genes



Unsupervised clustering

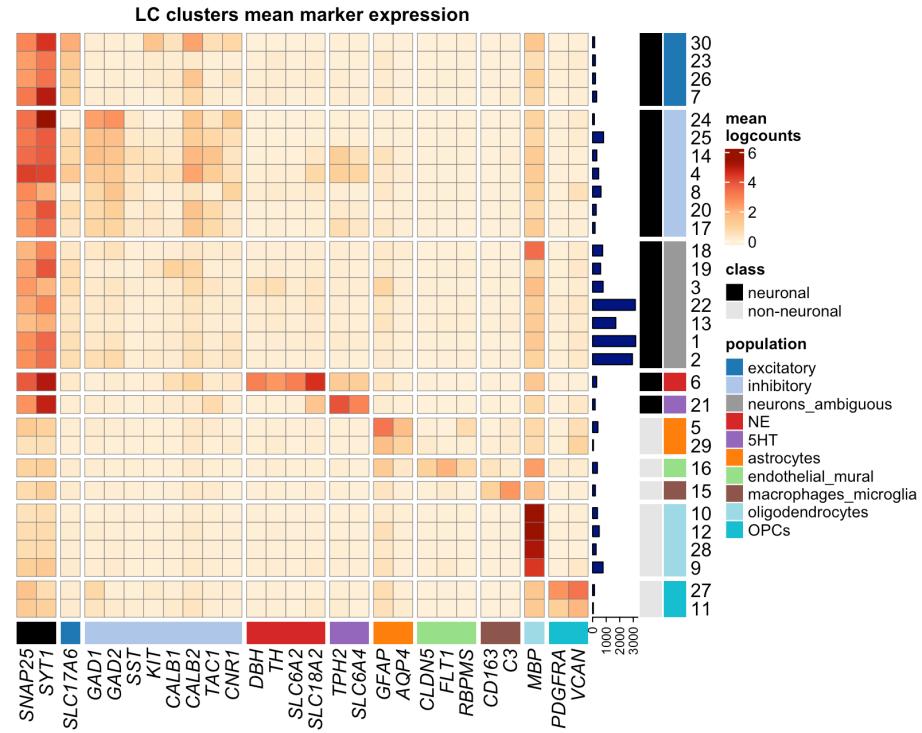
Unsupervised clustering in single-nucleus RNA sequencing data

- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei

Unsupervised clustering

Unsupervised clustering in single-nucleus RNA sequencing data

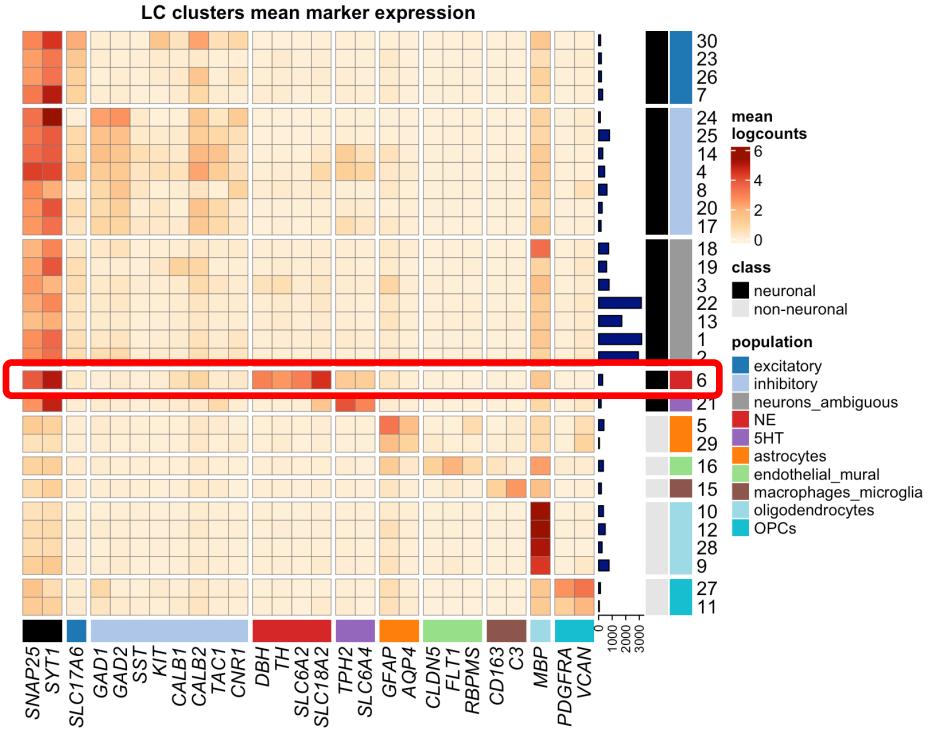
- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei



Unsupervised clustering

Unsupervised clustering in single-nucleus RNA sequencing data

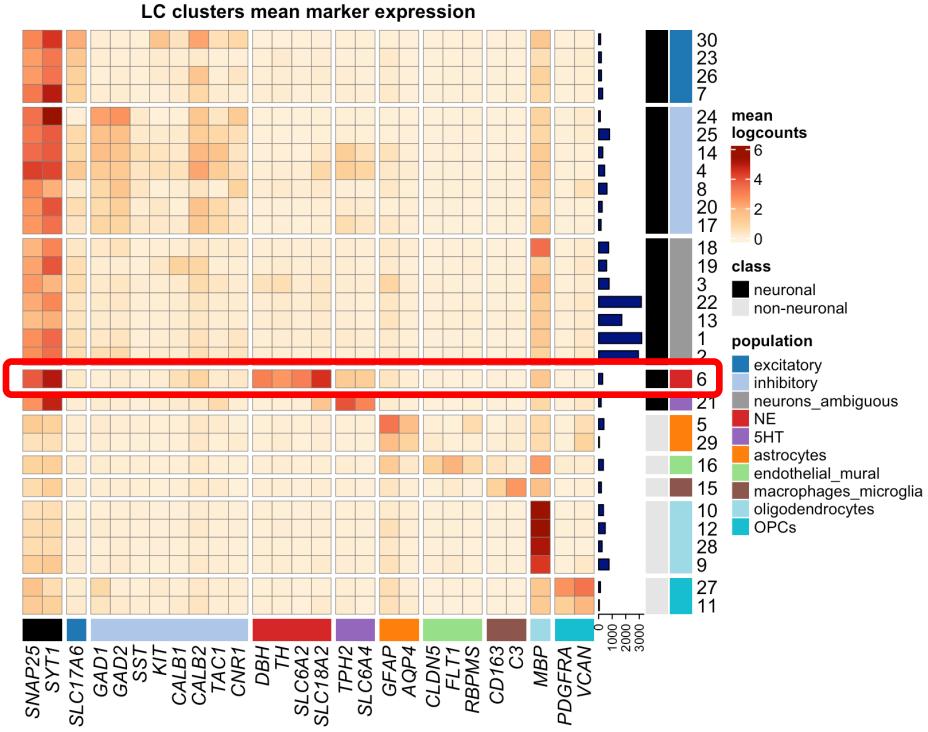
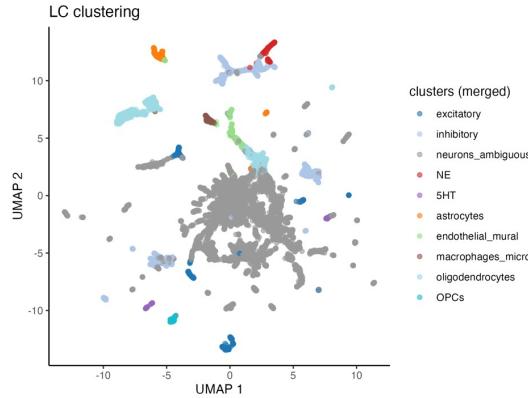
- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei



Unsupervised clustering

Unsupervised clustering in single-nucleus RNA sequencing data

- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei



Differential gene expression

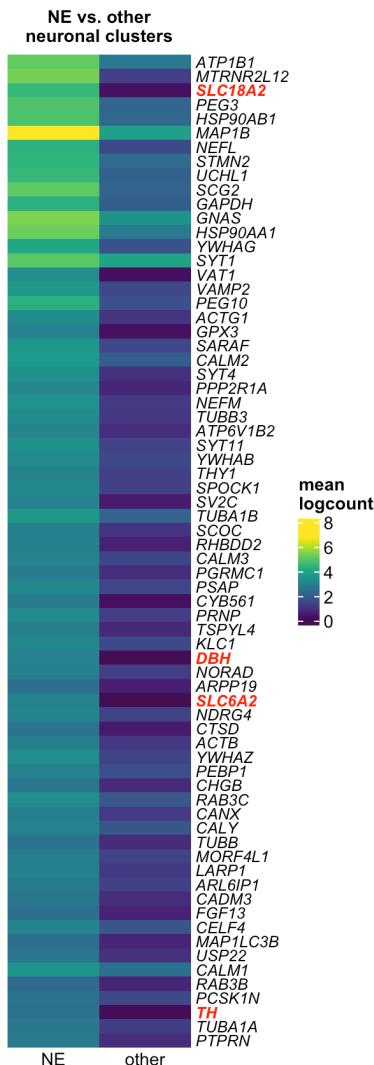
Differential gene expression testing in
single-nucleus RNA sequencing data

- NE neuron cluster vs. all other neuronal clusters
- Recovers known NE neuron markers and
additional unsupervised gene list
- 327 statistically significant genes

Differential gene expression

Differential gene expression testing in
single-nucleus RNA sequencing data

- NE neuron cluster vs. all other neuronal clusters
- Recovers known NE neuron markers and
additional unsupervised gene list
- 327 statistically significant genes



Unsupervised analyses recover additional unexpected results

Section of choroid plexus within LC sample from one donor

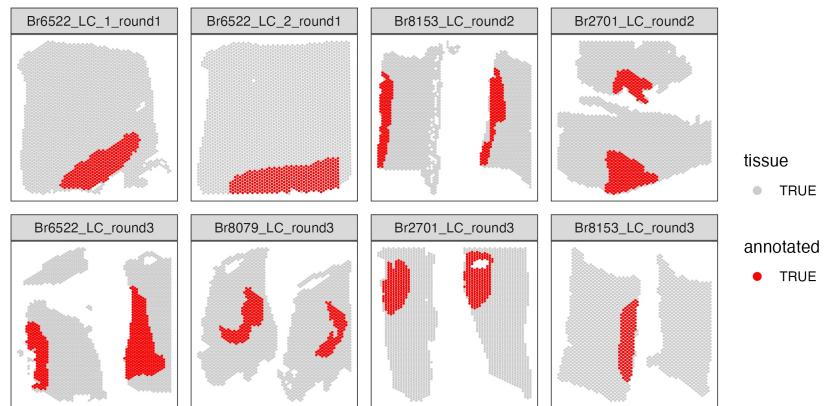
- nnSVG to identify spatially variable genes (SVGs) in Visium samples
- Top SVGs include choroid plexus-associated genes for samples from one donor (Br8079)

Unsupervised analyses recover additional unexpected results

Section of choroid plexus within LC sample from one donor

- nnSVG to identify spatially variable genes (SVGs) in Visium samples
- Top SVGs include choroid plexus-associated genes for samples from one donor (Br8079)

Manual annotations: LC regions

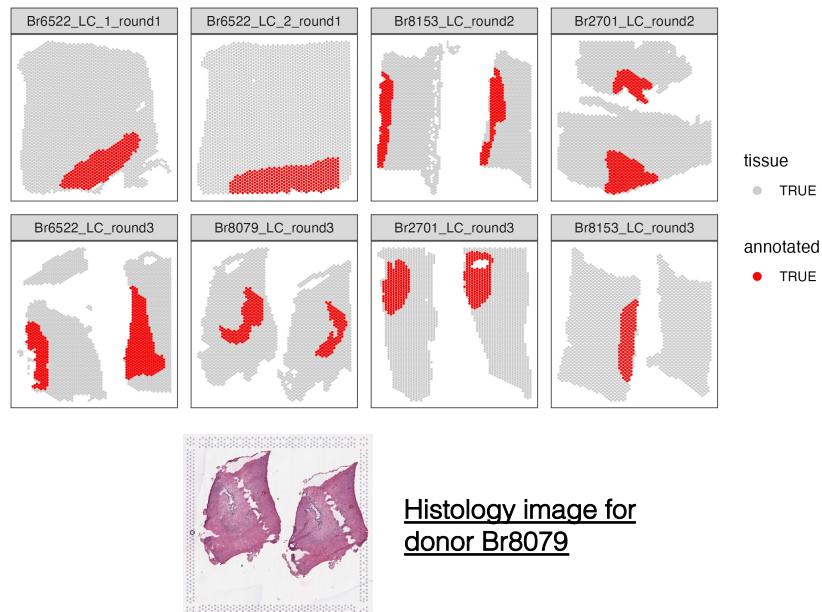


Unsupervised analyses recover additional unexpected results

Section of choroid plexus within LC sample from one donor

- nnSVG to identify spatially variable genes (SVGs) in Visium samples
- Top SVGs include choroid plexus-associated genes for samples from one donor (Br8079)

Manual annotations: LC regions

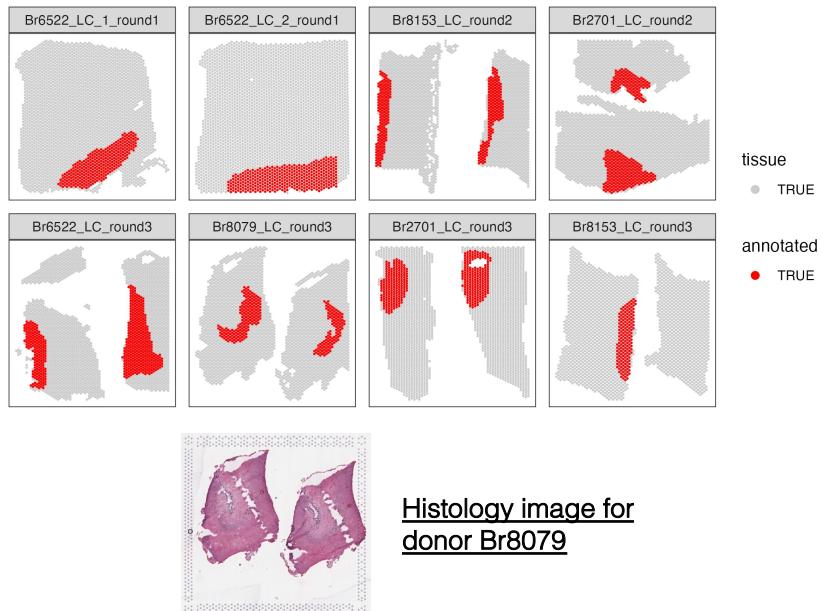


Unsupervised analyses recover additional unexpected results

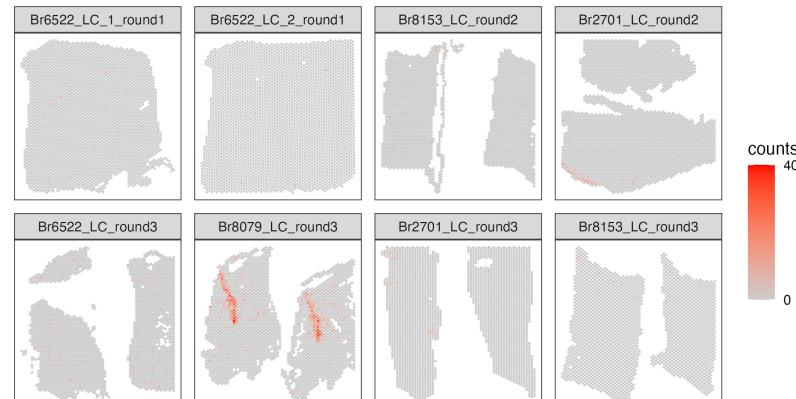
Section of choroid plexus within LC sample from one donor

- nnSVG to identify spatially variable genes (SVGs) in Visium samples
- Top SVGs include choroid plexus-associated genes for samples from one donor (Br8079)

Manual annotations: LC regions



Expression of choroid plexus-associated gene *CAPS*



Unsupervised analyses recover additional unexpected results

High proportion of mitochondrial reads within NE neuron cluster (unsupervised clustering)

- **Biological effect:** elevated expression of mitochondrial genes due to high metabolic demand for NE neurons
- **Technical effect:** capture of mitochondrial RNA within droplets / attached to nuclei

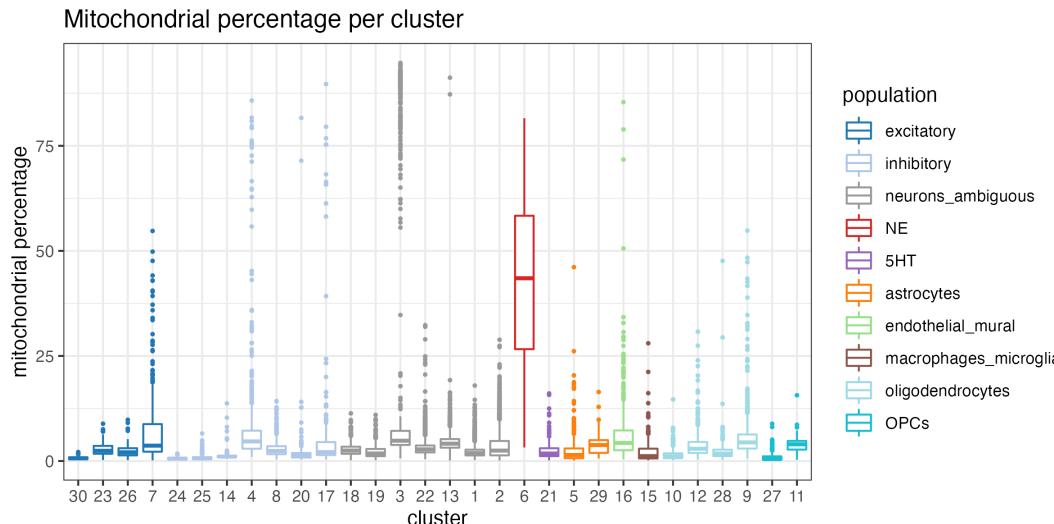
Unsupervised analyses recover additional unexpected results

High proportion of mitochondrial reads within NE neuron cluster (unsupervised clustering)

- **Biological effect:** elevated expression of mitochondrial genes due to high metabolic demand for NE neurons
- **Technical effect:** capture of mitochondrial RNA within droplets / attached to nuclei

Affects preprocessing analyses in analysis workflow

- Standard quality control metrics would remove NE neuron nuclei (main population of interest!)



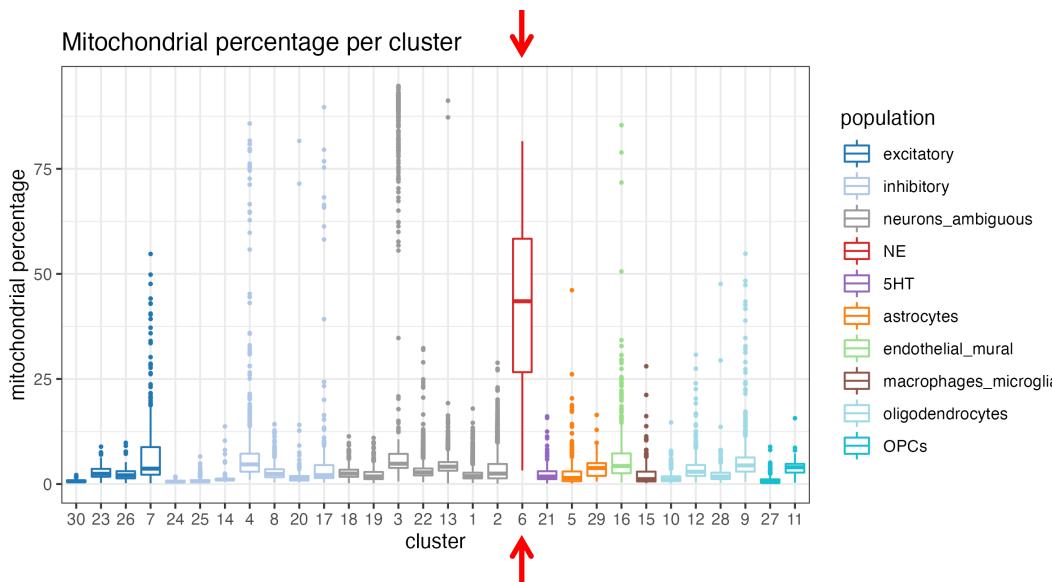
Unsupervised analyses recover additional unexpected results

High proportion of mitochondrial reads within NE neuron cluster (unsupervised clustering)

- **Biological effect:** elevated expression of mitochondrial genes due to high metabolic demand for NE neurons
- **Technical effect:** capture of mitochondrial RNA within droplets / attached to nuclei

Affects preprocessing analyses in analysis workflow

- Standard quality control metrics would remove NE neuron nuclei (main population of interest!)



Summary: locus coeruleus (LC) collaboration analysis

Characterization of transcriptome-wide gene expression landscape in human LC at spatial and **single-nucleus** resolution

Unsupervised analyses recover known marker genes, extended gene list, and unexpected results

- Further results: identification of 5-HT (5-hydroxytryptamine, serotonin) neurons, cholinergic marker gene (*SLC5A7*) expression within NE neurons

Weber and Divecha et al. (2022)
bioRxiv preprint; in revision (eLife)



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

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 Tippanni, Rahul Bharadwaj, Joel E. Kleinman, Stephanie C. Page, Thomas M. Hyde, Leonardo Collado-Torres, Kristen R. Maynard, Keri Martinowich, Stephanie C. Hicks

doi: <https://doi.org/10.1101/2022.10.28.514241>

Data availability

R/Bioconductor data package

- Downloadable R objects

The screenshot shows the Bioconductor package page for 'WeberDivechaLCdata'. At the top left is the Bioconductor logo with the text 'OPEN SOURCE SOFTWARE FOR BIOINFORMATICS'. To the right is a navigation bar with 'Home', 'Install', and 'Help' buttons. Below the navigation bar, the package name 'WeberDivechaLCdata' is displayed in green. A breadcrumb trail shows 'Home > Bioconductor 3.16 > Experiment Packages > WeberDivechaLCdata'. The main title 'WeberDivechaLCdata' is in bold green text. Below the title are several status indicators in colored boxes: 'platforms all', 'rank 415 / 416', 'support 0 / 0', 'build ok', 'updated < 1 month', and 'dependencies 134'. Below these, a DOI link 'DOI: 10.18129/B9.bioc.WeberDivechaLCdata' is followed by social media sharing icons for Facebook and Twitter. The main description reads: 'Spatially-resolved transcriptomics and single-nucleus RNA-sequencing data from the locus coeruleus (LC) in postmortem human brain samples'. At the bottom, it says 'Bioconductor version: Release (3.16)' and provides a detailed description of the dataset: 'Spatially-resolved transcriptomics (SRT) and single-nucleus RNA-sequencing (snRNA-seq) data from the locus coeruleus (LC) in postmortem human brain samples. Data were generated with the 10x Genomics Visium SRT and 10x Genomics Chromium snRNA-seq platforms. Datasets are stored in SpatialExperiment and SingleCellExperiment formats.'

Data availability

R/Bioconductor data package

- Downloadable R objects



Home » Bioconductor 3.16 » Experiment Packages » WeberDivechaLCdata

WeberDivechaLCdata



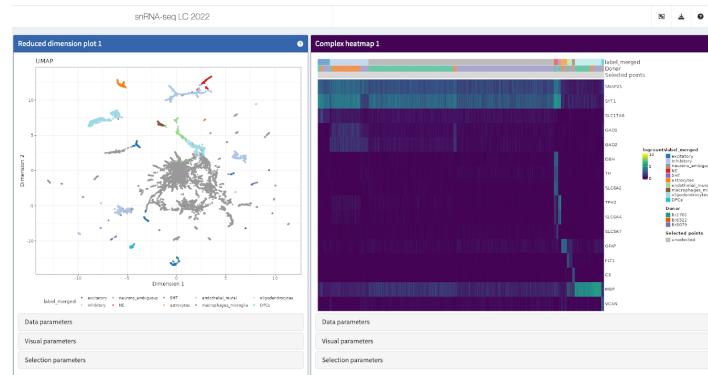
Spatially-resolved transcriptomics and single-nucleus RNA-sequencing data from the locus coeruleus (LC) in postmortem human brain samples

Bioconductor version: Release (3.16)

Spatially-resolved transcriptomics (SRT) and single-nucleus RNA-sequencing (snRNA-seq) data from the locus coeruleus (LC) in postmortem human brain samples. Data were generated with the 10x Genomics Visium SRT and 10x Genomics Chromium snRNA-seq platforms. Datasets are stored in SpatialExperiment and SingleCellExperiment formats.

Web apps

- Interactive exploration of data



Heena Divecha
Lieber Institute
for Brain
Development

Reproducibility

GitHub repository

- Code to reproduce analyses and figures

The screenshot shows a GitHub repository page for the user 'lmweber' with the repository name 'locus-c'. The page includes a navigation bar with links to Pull requests, Issues, Codespaces, Marketplace, and Explore. Below the navigation is a search bar and a repository header with the name 'lmweber/locus-c' and a 'Public' status. The main content area displays a list of recent commits:

Commit	Message	Date	Commits
RNAscope	Added .gitignore to ignore huge RNAscope images	4 months ago	89de099
alignment	move into subdirectories	last year	
code	convert sample IDs to character for spatialIBD	2 months ago	
inputs	Grimm mouse gene list	8 months ago	
processed_data/cellranger	add sample info in README	4 months ago	
sample_info	Added the sample_info excel sheet for LC snRNA-seq data. Originally...	3 months ago	
web_summaries	add web summaries from Space Ranger and Cell Ranger	2 months ago	
.gitignore	Ignore any R CMD BATCH script in subdirs	last year	
LICENSE	file permissions on cluster	2 years ago	
README.md	formatting	2 months ago	
locus-c.Rproj	file permissions on cluster	2 years ago	

Below the commits is a 'About' section with a brief description: 'Analysis code for locus coeruleus (LC) spatial transcriptomics project'. It includes links to 'Readme', 'MIT license', '4 stars', '3 watching', and '0 forks'. The 'Releases' section indicates 'No releases published' and 'Create a new release'. The 'Packages' section shows 'No packages published' and 'Publish your first package'. The 'Contributors' section lists five contributors with their profile icons. The 'Languages' section shows a chart where 'HTML' accounts for 99.1% and 'Other' for 0.9%. The 'README.md' file content is also shown at the bottom.

Research Theme 1

Unsupervised statistical methods

Spatially-resolved transcriptomics: [nnSVG](#)

- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: [diffcyt](#)

- Weber et al. (2019), *Comms Biol*

1

Research Theme 2

Collaborative analyses

Neuroscience

- Weber and Divecha et al. (2022), *bioRxiv / in revision (eLife)*
- Maynard and Collado-Torres et al. (2021), *Nat Neur*

Cancer

Immunology

2

Methodological development and collaborative analyses for high-throughput genomic data

Technological platforms: spatially-resolved transcriptomics, single-cell / single-nucleus RNA sequencing, high-dimensional cytometry

K99/R00 Award

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data
(1K99HG012229-01)



Research Theme 3

Benchmarking

- Cancer: Weber et al. (2021), *GigaScience*
- Review / guidelines: Weber et al. (2019), *Genome Biol*
- Independent benchmarking: Weber et al. (2016), *Cyt Part A*

Analysis workflows and tools

- R/Bioconductor: Righelli, Weber, Crowell et al. (2021), *Bioinf*

Open-source software / reproducible research

additional papers
on website and
Google Scholar



Benchmarking evaluations

Which methods perform well in which types of data / for which types of analyses?

Benchmarking evaluations

Which methods perform well in which types of data / for which types of analyses?

Example

- Weber et al. (2021), *GigaScience*
- Genetic demultiplexing algorithms for pooled single-cell RNA-sequencing data from multiple donors for high-grade serous ovarian cancer and lung adenocarcinoma
- Algorithms (e.g. Vireo; Huang et al. 2019) perform well despite additional somatic mutations



GigaScience, 10, 2021, 1–12

<https://doi.org/10.1093/gigascience/giab062>
Research

RESEARCH

Genetic demultiplexing of pooled single-cell RNA-sequencing samples in cancer facilitates effective experimental design

Lukas M. Weber ¹, Ariel A. Hippen ², Peter F. Hickey ³, Kristofer C. Berrett ⁴, Jason Gertz ⁴, Jennifer Anne Doherty ⁴, Casey S. Greene ⁵ and Stephanie C. Hicks ^{1,*}

¹Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA;

²Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; ³Advanced Technology & Biology Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; ⁴Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, UT 84108, USA and ⁵Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

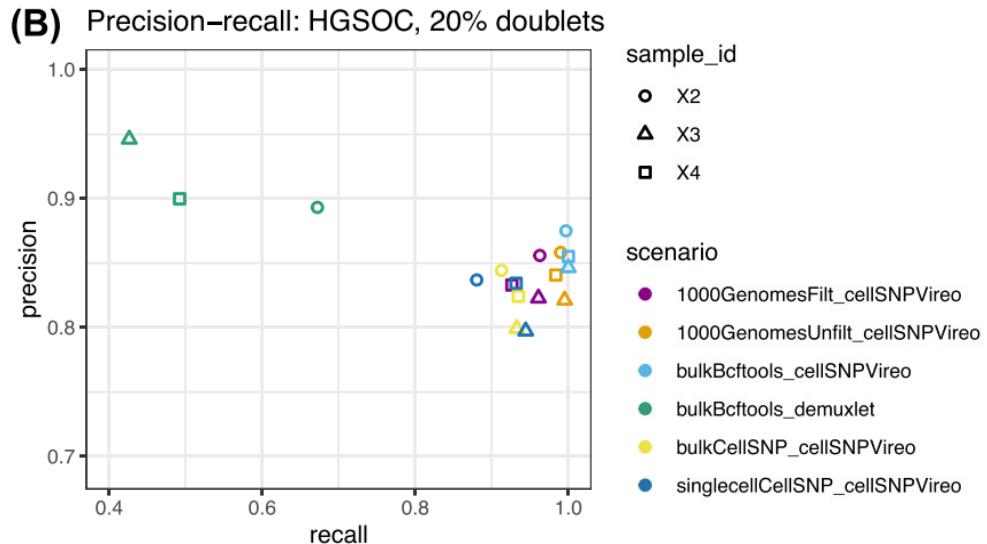
*Correspondence address. Stephanie C. Hicks, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205-2179, USA. E-mail: shicks19@jhu.edu <http://orcid.org/0000-0002-7858-0231>

Benchmarking evaluations

Which methods perform well in which types of data / for which types of analyses?

Example

- Weber et al. (2021), *GigaScience*
- Genetic demultiplexing algorithms for pooled single-cell RNA-sequencing data from multiple donors for high-grade serous ovarian cancer and lung adenocarcinoma
- Algorithms (e.g. Vireo; Huang et al. 2019) perform well despite additional somatic mutations



Benchmarking studies

Guidelines / review paper

- How to design / perform rigorous benchmarking studies?
- Independent benchmarking vs. benchmarking during method development

Benchmarking studies

Guidelines / review paper

- How to design / perform rigorous benchmarking studies?
- Independent benchmarking vs. benchmarking during method development

Weber *et al.* *Genome Biology* (2019) 20:125
<https://doi.org/10.1186/s13059-019-1738-8>

Genome Biology

REVIEW **Open Access**

Essential guidelines for computational method benchmarking



Lukas M. Weber^{1,2}, Wouter Saelens^{3,4}, Robrecht Cannoodt^{3,4}, Charlotte Soneson^{1,2,8}, Alexander Hapfelmeier⁵, Paul P. Gardner⁶, Anne-Laure Boulesteix⁷, Yvan Saey^{3,4*} and Mark D. Robinson^{1,2*} 

Abstract

In computational biology and other sciences, researchers are frequently faced with a choice between several computational methods for performing data analyses. Benchmarking studies aim to rigorously compare the performance of different methods using well-characterized benchmark datasets, to determine the strengths of each method or to provide recommendations regarding suitable choices of methods for an analysis. However, benchmarking studies must be carefully designed and implemented to provide accurate, unbiased, and informative results. Here, we summarize key practical guidelines and recommendations for performing high-quality benchmarking analyses, based on our experiences in computational biology.

Box 1: Summary of guidelines

The guidelines in this review can be summarized in the following set of recommendations. Each recommendation is discussed in more detail in the corresponding section in the text.

1. Define the purpose and scope of the benchmark.

Benchmarking studies

Guidelines / review paper

- How to design / perform rigorous benchmarking studies?
- Independent benchmarking vs. benchmarking during method development

Weber et al. *Genome Biology* (2019) 20:125
<https://doi.org/10.1186/s13059-019-1738-8>

Genome Biology

REVIEW

Open Access



Essential guidelines for computational method benchmarking

Lukas M. Weber^{1,2}, Wouter Saelens^{3,4}, Robrecht Cannoodt^{3,4}, Charlotte Soneson^{1,2,8}, Alexander Hapfelmeier⁵, Paul P. Gardner⁶, Anne-Laure Boulesteix⁷, Yvan Saey^{3,4*} and Mark D. Robinson^{1,2*}

Abstract

In computational biology and other sciences, researchers are frequently faced with a choice between several computational methods for performing data analyses. Benchmarking studies aim to rigorously compare the performance of different methods using well-characterized benchmark datasets, to determine the strengths of each method or to provide recommendations regarding suitable choices of methods for an analysis. However, benchmarking studies must be carefully designed and implemented to provide accurate, unbiased, and informative results. Here, we summarize key practical guidelines and recommendations for performing high-quality benchmarking analyses, based on our experiences in computational biology.

Box 1: Summary of guidelines

The guidelines in this review can be summarized in the following set of recommendations. Each recommendation is discussed in more detail in the corresponding section in the text.

1. Define the purpose and scope of the benchmark.

Independent benchmarking studies

- Weber et al. (2021), *GigaScience*
- Weber et al. (2016), *Cyt Part A*

Benchmarking during method development

- nnSVG: Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*
- diffcyt: Weber et al. (2019), *Comms Biol*

Analysis workflows and tools

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

Analysis workflows and tools

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

- Righelli, Weber, Crowell et al. (2022)

The screenshot shows a journal article from the Bioinformatics journal. The header includes links for 'Issues', 'Advance articles', 'Submit', 'Purchase', 'Alerts', and 'About'. A 'Bioinformatics' button is also present. The main content area features a thumbnail of the journal cover, which has a DNA double helix on it. Below the cover, the text reads 'Volume 38, Issue 11' and '1 June 2022'. A link 'Article Contents' is provided. To the right, the article title is displayed in a box labeled 'JOURNAL ARTICLE': 'SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor'. The authors listed are Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T Lun, Stephanie C Hicks, Davide Risso, and Author Notes. The DOI is https://doi.org/10.1093/bioinformatics/btac299. The article was published on 28 April 2022.

Bioinformatics

Issues Advance articles Submit ▾ Purchase Alerts About ▾ Bioinformatics

JOURNAL ARTICLE

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

Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T 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

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

- Righelli, Weber, Crowell et al. (2022)

R/Bioconductor-based analysis workflows for high-dimensional cytometry

- Nowicka et al. (2019)

The screenshot shows a journal article page from the Bioinformatics journal. The header includes links for 'Issues', 'Advance articles', 'Submit', 'Purchase', 'Alerts', and 'About'. A 'Bioinformatics' button is also present. The main content area features a thumbnail of the journal cover, which has a DNA double helix on it. Below the cover, the text reads 'Volume 38, Issue 11' and '1 June 2022'. A link to 'Article Contents' is provided. To the right, the article title is displayed in a box labeled 'JOURNAL ARTICLE': 'SpatialExperiment: infrastructure for spatially-resolved transcriptomics data in R using Bioconductor'. The authors listed are Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T Lun, Stephanie C Hicks, Davide Risso, and Author Notes. The DOI is https://doi.org/10.1093/bioinformatics/btac299. The article was published on 28 April 2022.

Analysis workflows and tools

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

- Righelli, Weber, Crowell et al. (2022)

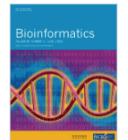
R/Bioconductor-based analysis workflows for high-dimensional cytometry

- Nowicka et al. (2019)



Bioinformatics

Issues Advance articles Submit ▾ Purchase Alerts About ▾ Bioinformatics



Volume 38, Issue 11
1 June 2022

Article Contents

JOURNAL ARTICLE

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

Dario Righelli, Lukas M Weber, Helena L Crowell, Brenda Pardo, Leonardo Collado-Torres, Shila Ghazanfar, Aaron T 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 ▾

F1000Research

Search

BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ▾ ABOUT ▾ BLOG

Home » Browse » CyTOF workflow: differential discovery in high-throughput high-dimensional...

CHECK FOR UPDATES

METHOD ARTICLE REVISED CyTOF workflow: differential discovery in high-throughput high-dimensional cytometry datasets [version 4; peer review: 2 approved]

Malgorzata Nowicka^{1,2}, Carsten Krieg³, Helena L. Crowell^{1,2}, Lukas M. Weber^{1,2}, Felix J. Hartmann^{1,2}, Silvia Guglietta⁴, Burkhard Becher³, Mitchell P. Levesque⁵, Mark D. Robinson^{1,2} AUTHOR DETAILS

ALL METRICS 41964 VIEWS 6435 DOWNLOADS

Get PDF Get XML Cite Export Track

This article is included in the [Bioconductor](#) gateway.

Abstract

Reproducibility and open science

Freely accessible / open-source software packages

- R/Bioconductor packages (nnSVG, diffcyt)



The screenshot shows the Bioconductor website's navigation bar with links for Home, Install, and Help. Below the navigation, there are two package pages: 'diffcyt' and 'nnSVG'. The 'diffcyt' page has a green header, while the 'nnSVG' page has a blue header.

Code availability to reproduce analyses

- GitHub repositories

The screenshot shows three GitHub repository pages for the user 'Imweber'. The first repository, 'nnSVG-analyses', has a green header. The second and third repositories, both named 'locus-c', have blue headers. Each repository page includes standard GitHub navigation and repository details.

Data availability

- Data packages, managed repositories, web resources

The screenshot shows two bioRxiv preprint server pages. Both pages feature a yellow banner at the top stating: '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.' The first preprint is titled 'The gene expression landscape of t single-nucleus and spatially-resolved' and the second is titled 'nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes'. Both pages include author information and a DOI link.

The screenshot shows the Bioconductor website's navigation bar with links for Home, Install, and Help. Below the navigation, there are two package pages: 'HDCytoData' and 'STexampleData'. The 'HDCytoData' page has a green header, while the 'STexampleData' page has a blue header.

Preprints

- bioRxiv

Reproducibility and open science

Freely accessible / open-source software packages

- R/Bioconductor packages (nnSVG, diffcyt)

Code availability to reproduce analyses

- GitHub repositories

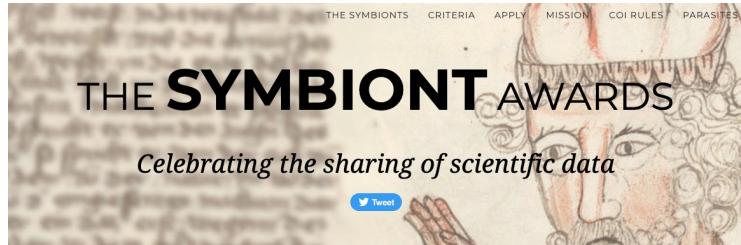
Data availability

- Data packages, managed repositories, web resources

Preprints

- bioRxiv

Research Symbiont Award (2021)



AWARD RECIPIENTS

Exemplars of research symbiosis.



Lukas Weber
2021 Junior Symbiont

Research Theme 1

Unsupervised statistical methods

Spatially-resolved transcriptomics: [nnSVG](#)

- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: [diffcyt](#)

- Weber et al. (2019), *Comms Biol*

1

Research Theme 2

Collaborative analyses

Neuroscience

- Weber and Divecha et al. (2022), *bioRxiv / in revision (eLife)*
- Maynard and Collado-Torres et al. (2021), *Nat Neur*

Cancer

Immunology

2

Methodological development and collaborative analyses for high-throughput genomic data

Technological platforms: spatially-resolved transcriptomics, single-cell / single-nucleus RNA sequencing, high-dimensional cytometry

K99/R00 Award

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data
(1K99HG012229-01)



Research Theme 3

Benchmarking

- Cancer: Weber et al. (2021), *GigaScience*
- Review / guidelines: Weber et al. (2019), *Genome Biol*
- Independent benchmarking: Weber et al. (2016), *Cyt Part A*

Analysis workflows and tools

- R/Bioconductor: Righelli, Weber, Crowell et al. (2021), *Bioinf*

Open-source software / reproducible research

additional papers
on website and
Google Scholar



Future plans

K99/R00 award (NHGRI)

Unsupervised Statistical Methods for Data-driven Analyses in Spatially Resolved Transcriptomics Data (1K99HG012229-01)

Spatially-resolved transcriptomics



Aim 1

Methods for improved preprocessing and feature selection

- **Aim 1a:** Spatially-aware quality control
- **Aim 1b:** Spot-level weights
- **Aim 1c:** Spatially variable genes



★ published

☆ preprint

● work in progress

Aim 2

Unsupervised methods for spatial domains / spatially-resolved cell types

- **Aim 2a:** Spatially-resolved clustering
- **Aim 2b:** Spatial domain analysis
- **Aim 2c:** Differential discovery



Aim 3

Data infrastructure and benchmarking resources

- **Aim 3a:** R/Bioconductor data infrastructure
- **Aim 3b:** Analysis workflows and benchmarking resources

K99 funding start date: Feb 2022

K99/R00 award

Aim 1a: Spatially-aware quality control

- Spatially-resolved transcriptomics datasets with multiple samples
(e.g. multiple Visium capture areas / slides)

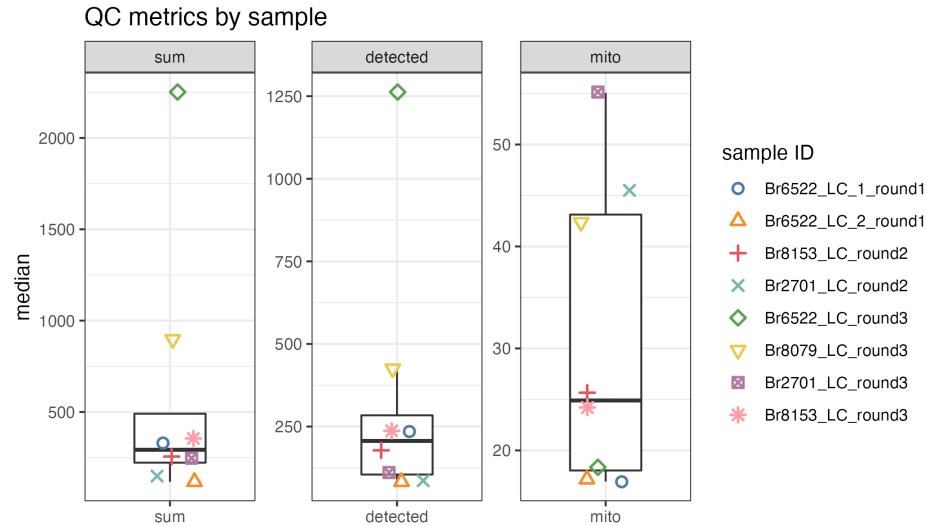
K99/R00 award

Aim 1a: Spatially-aware quality control

- Spatially-resolved transcriptomics datasets with multiple samples
(e.g. multiple Visium capture areas / slides)

Example: locus coeruleus (LC) dataset

- Variation in read depth across samples
- Data-driven / adaptive method to select quality control metrics per sample



K99/R00 award

Aim 3b: Analysis workflows for spatially-resolved transcriptomics data

- Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor ([OSTA](#))
- Online book containing example R code and datasets for individual analysis steps and workflows

K99/R00 award

Aim 3b: Analysis workflows for spatially-resolved transcriptomics data

- Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor ([OSTA](#))
- Online book containing example R code and datasets for individual analysis steps and workflows

The screenshot shows the homepage of the "Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor" (OSTA) website. The header includes a search bar and navigation icons. The main content area features a title card with the book's name and subtitle, the date (2022-08-07), and a "Welcome" section. The "Welcome" section contains a brief introduction to the book, mentioning it is an online book for computational analysis workflows for spatially resolved transcriptomics (SRT) data using the Bioconductor framework in R. It also notes that the book is organized into several parts, including background, preprocessing steps, analysis steps, and complete workflows. A Bioconductor logo is displayed. The sidebar on the left lists the book's structure: I Introduction, II Preprocessing steps, III Analysis steps, IV Workflows.

The screenshot shows the "Chapter 17 Human DLPFC workflow" page. The header includes a search bar and navigation icons. The main content area features a title card with the chapter name and subtitle, the date (2022-08-07), and a "Description of dataset" section. The "Description of dataset" section contains a brief introduction to the 10x Genomics Visium dataset generated from healthy human brain samples from the dorsolateral prefrontal cortex (DLPFC) region. It notes that the full dataset has 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. A code block at the bottom shows R code for clearing the workspace:

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

The sidebar on the left lists the chapter's structure: III Analysis steps, IV 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.

Long-term plans

Statistical and machine learning methodological development driven by biological collaborations in neuroscience and cancer, with focus on single-cell and spatially-resolved transcriptomics

New technological platforms

Funding experience and plans

- Currently funded by K99/R00 award
- Applied for additional funding from Chan Zuckerberg Initiative (funding decisions Mar 2023)
- Previously applied to Swiss National Science Foundation (SNSF) (2018, 2019) (2x applications; 1 successful)
- Plan to apply for R01 funding from NIH during initial years as tenure-track faculty



Acknowledgments

Postdoc advisor

Dr. Stephanie Hicks
Johns Hopkins Bloomberg School of Public Health



PhD advisor

Dr. Mark Robinson
University of Zurich,
Switzerland



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
Heena Divecha



SpatialExperiment team

Dario Righelli
Helena Crowell
Davide Risso

Bioconductor community



Funding

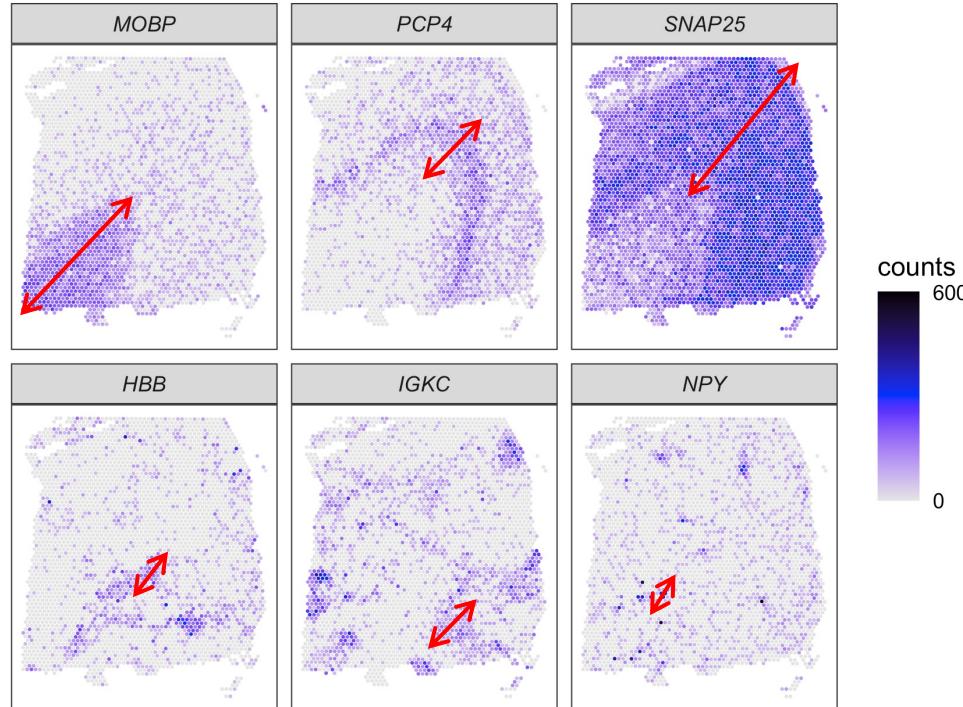


Thank you!

ADDITIONAL SLIDES

Spatially variable genes

Selected SVGs: human DLPFC



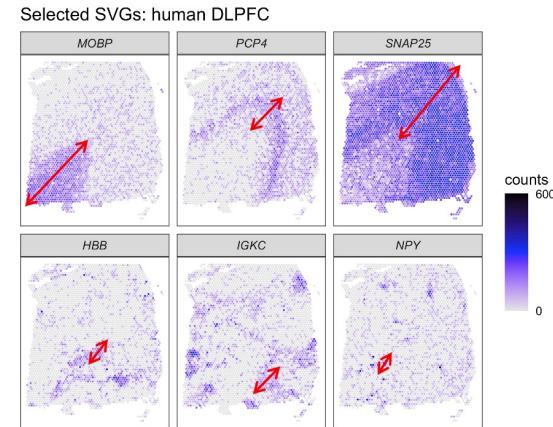
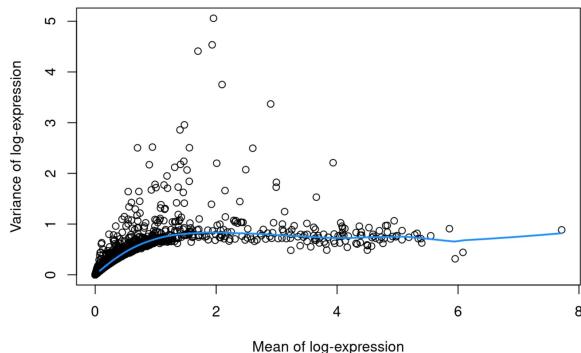
Existing methods are too slow or perform poorly

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

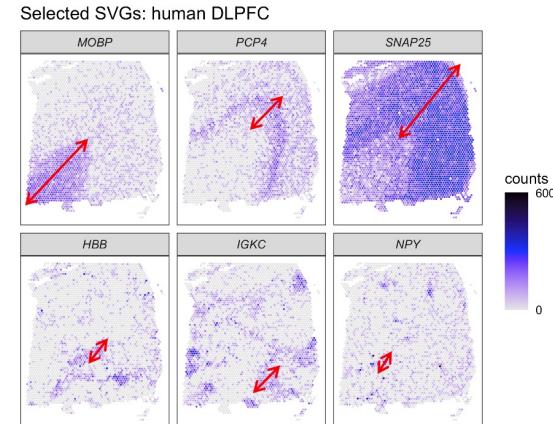
Moran's I statistic (spatial baseline)

- Rank genes by observed spatial autocorrelation
- Values range from -1 to +1

$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

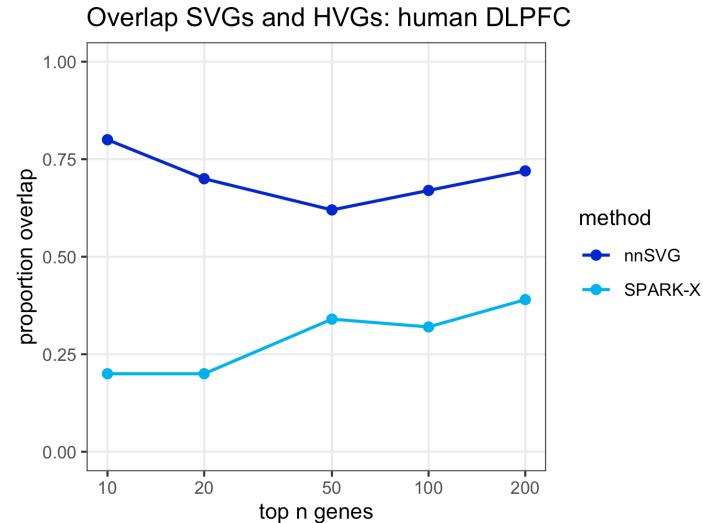
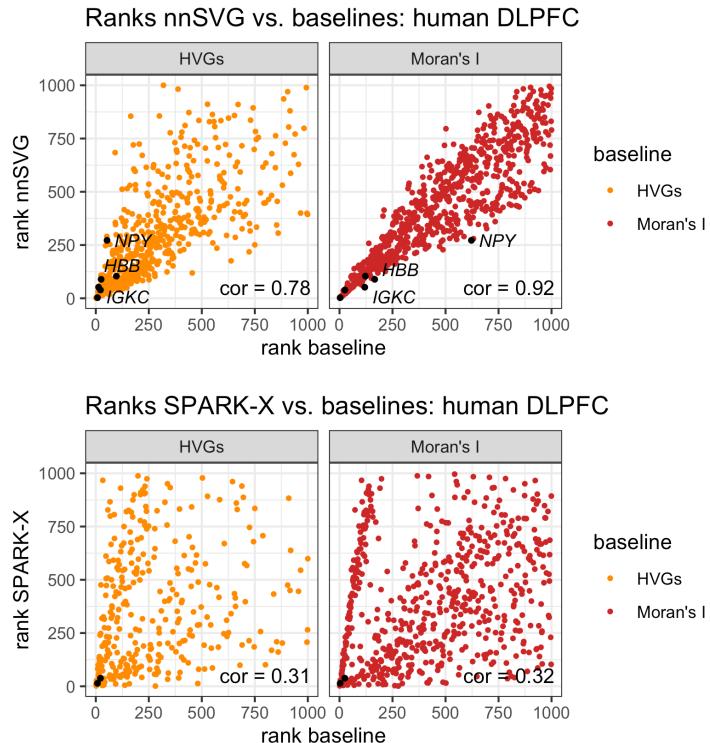
N = number of spatial locations i, j

W = sum of weights w_{ij} defined by e.g. inverse squared Euclidean distances



nnSVG evaluations / benchmarking

Human DLPFC dataset

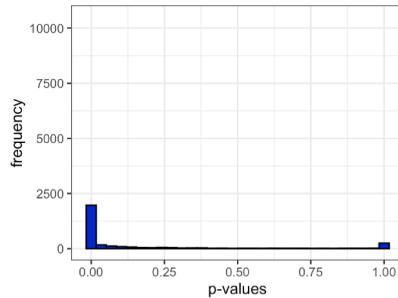


nnSVG evaluations / benchmarking

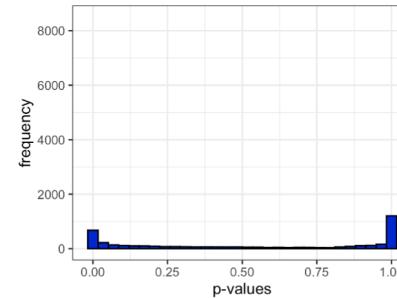
Human DLPFC dataset

Calibration of p-values

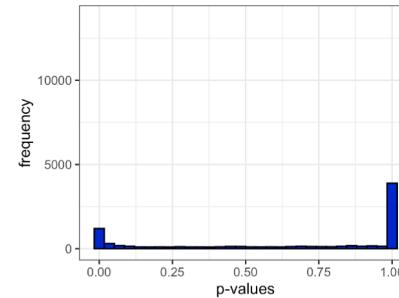
A nnSVG p-values: human DLPFC



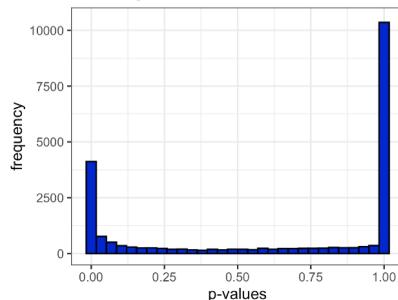
B nnSVG p-values: mouse OB



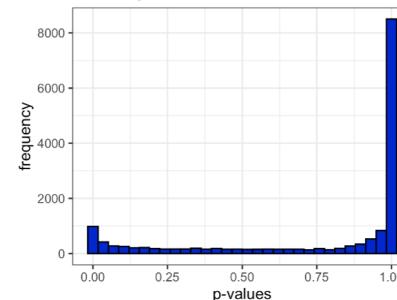
C nnSVG p-values: mouse HPC



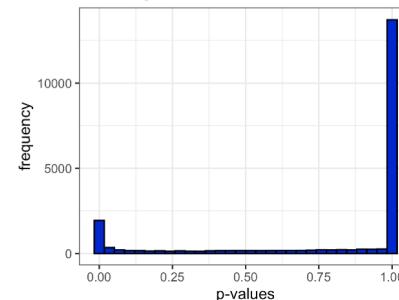
D nnSVG p-values: human DLPFC
no filtering



E nnSVG p-values: mouse OB
no filtering



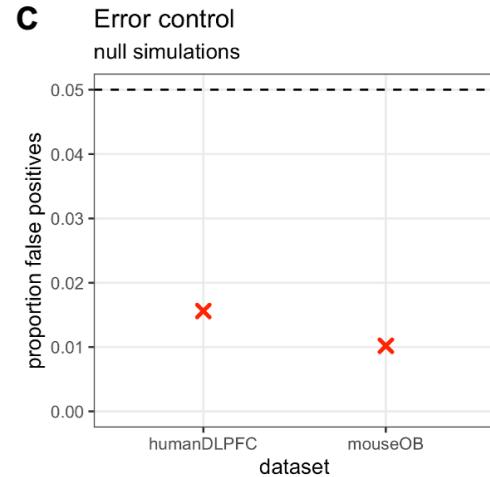
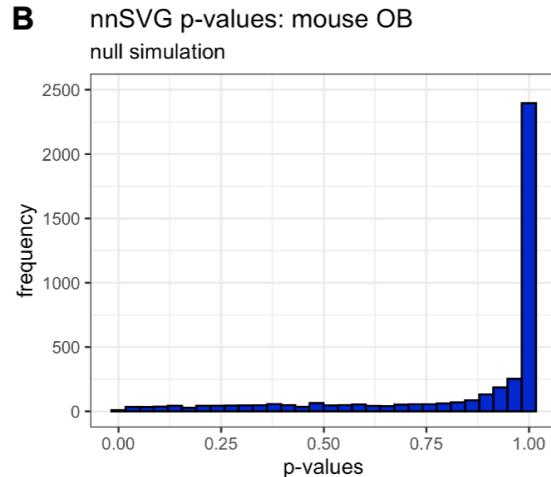
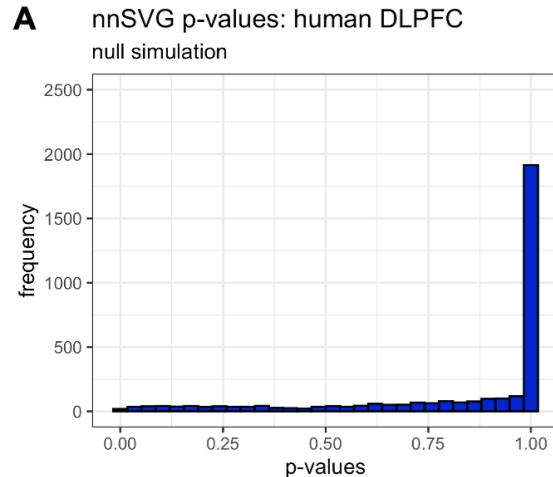
F nnSVG p-values: mouse HPC
no filtering



nnSVG evaluations / benchmarking

Human DLPFC dataset

Null simulations



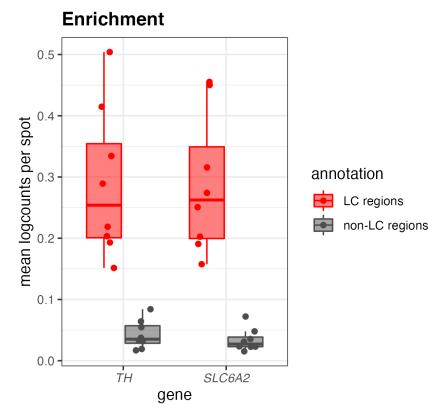
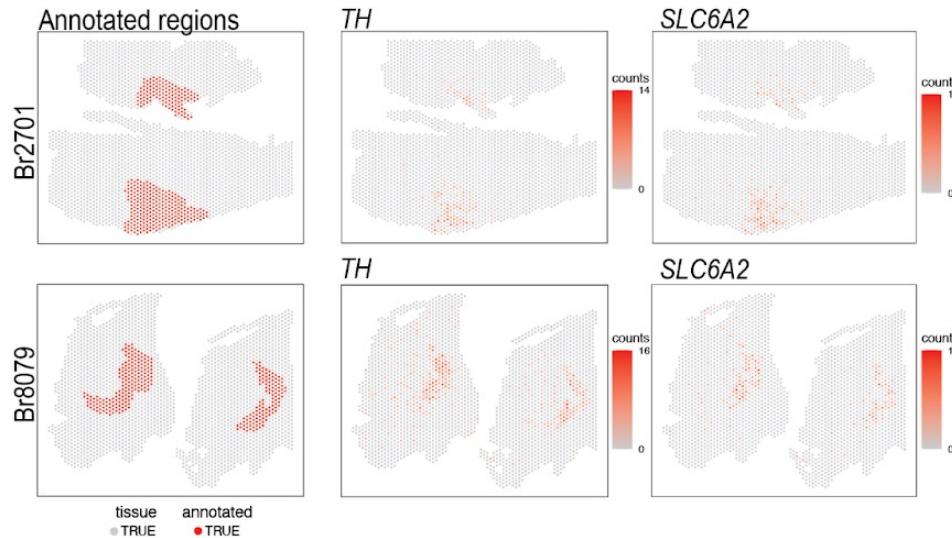
LC: Visium analyses

Heena Divecha
Kristen Maynard

Lieber Institute for Brain Development

Manually annotated regions used as input for computational analyses

Annotations validated by confirming expression of known norepinephrine (NE) marker genes

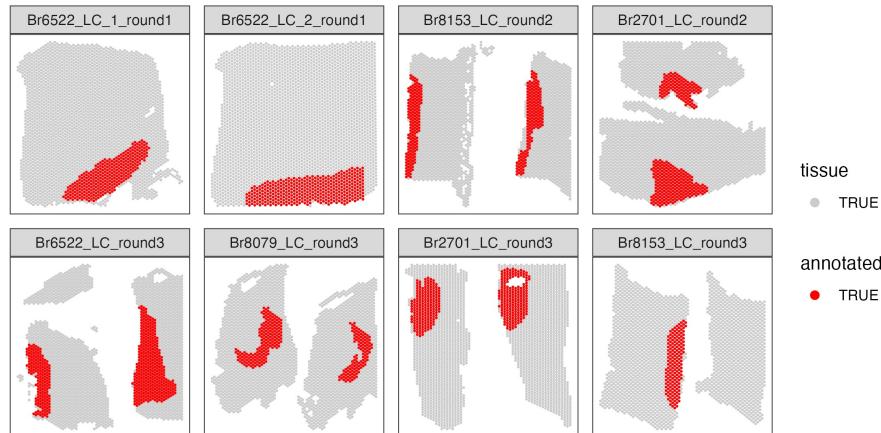


LC: Visium analyses

Manual annotations

- 13 annotated LC regions after quality control
- 124 to 395 Visium spots per annotated LC region

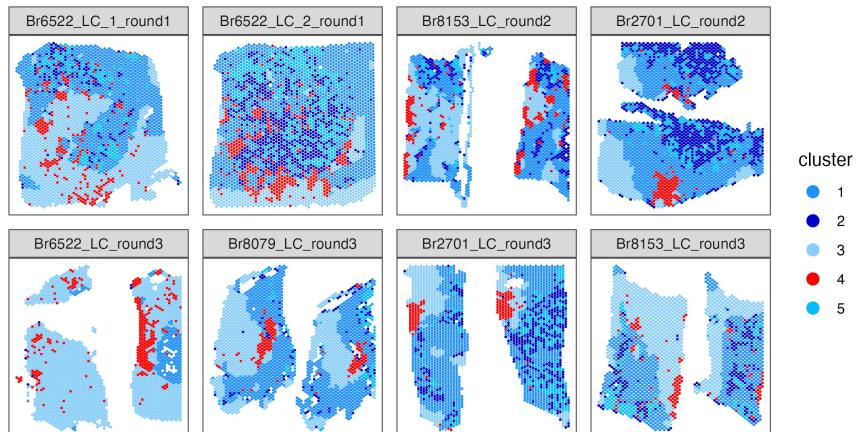
Annotations



Spatial clustering

- Spatial clustering using BayesSpace (Zhao et al. 2021) does not sufficiently recover manually annotated regions, so we rely on manual annotations for downstream analyses

BayesSpace clustering



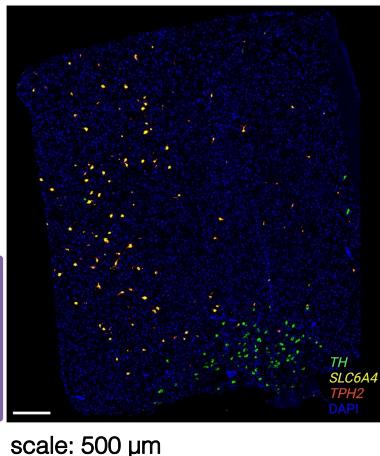
Unsupervised analyses recover additional unexpected results

Identification of 5-HT neurons

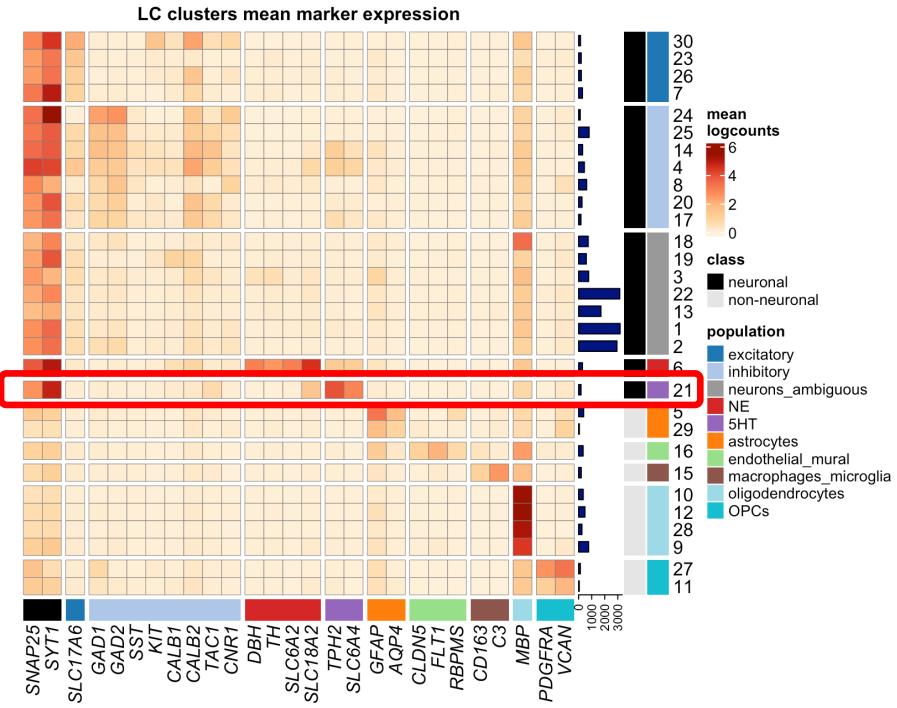
- Unsupervised clustering recovers a cluster of 5-hydroxytryptamine / serotonin (5-HT) neurons
- Dorsal raphe nucleus / adjacent to LC

Validation

- Single-molecule in situ hybridization (smFISH / RNAscope) and high-magnification confocal imaging



Heena Divecha
Lieber Institute
for Brain
Development



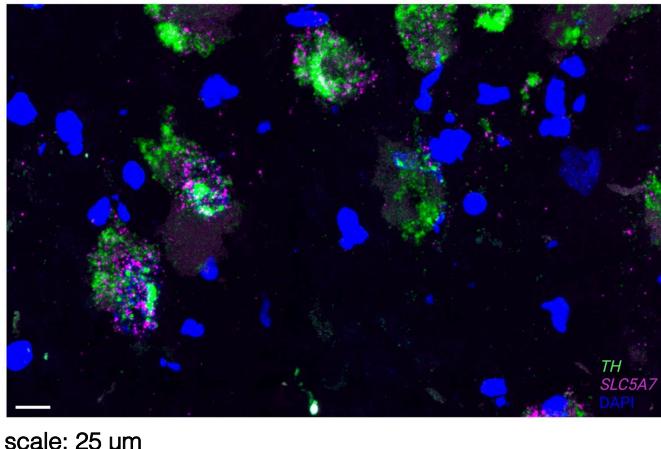
Unsupervised analyses recover additional unexpected results

Expression of cholinergic marker genes (*SLC5A7*) within NE neuron cluster

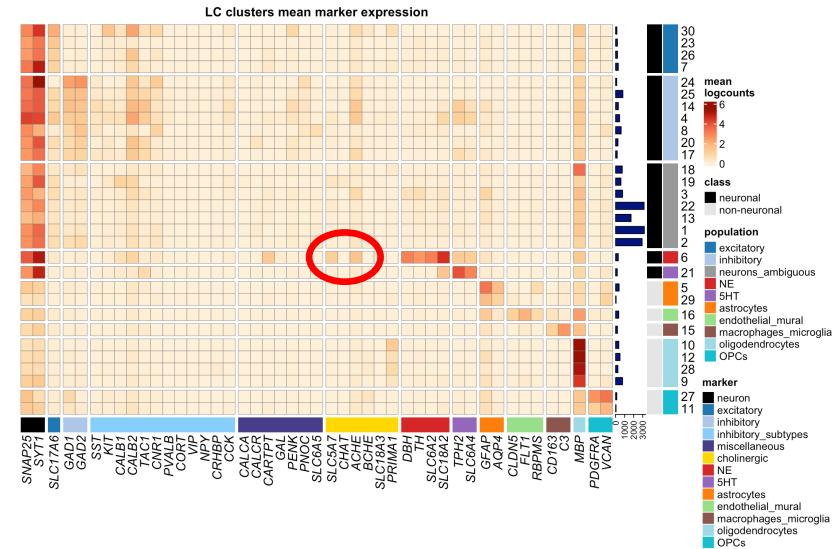
- Unsupervised clustering: transcriptome-wide expression profile of NE neuron cluster

Validation

- Single-molecule in situ hybridization (smFISH / RNAscope) and high-magnification confocal imaging

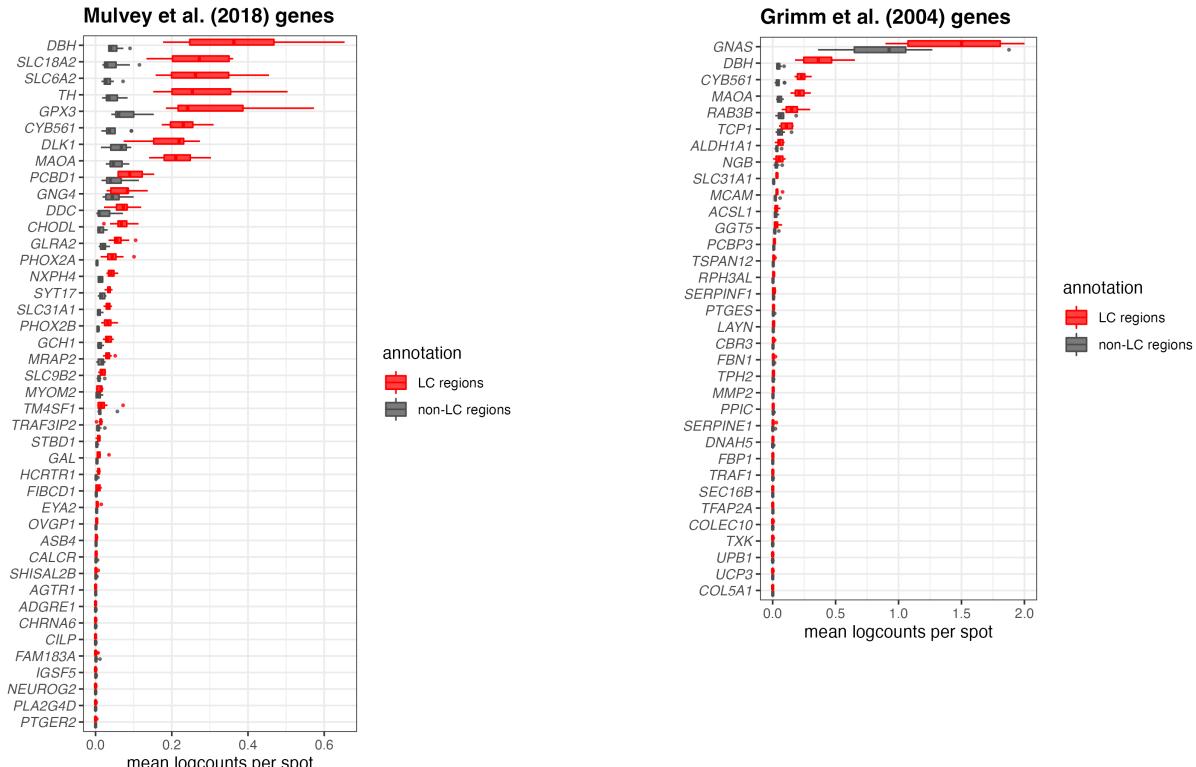


Heena Divecha
Lieber Institute
for Brain
Development



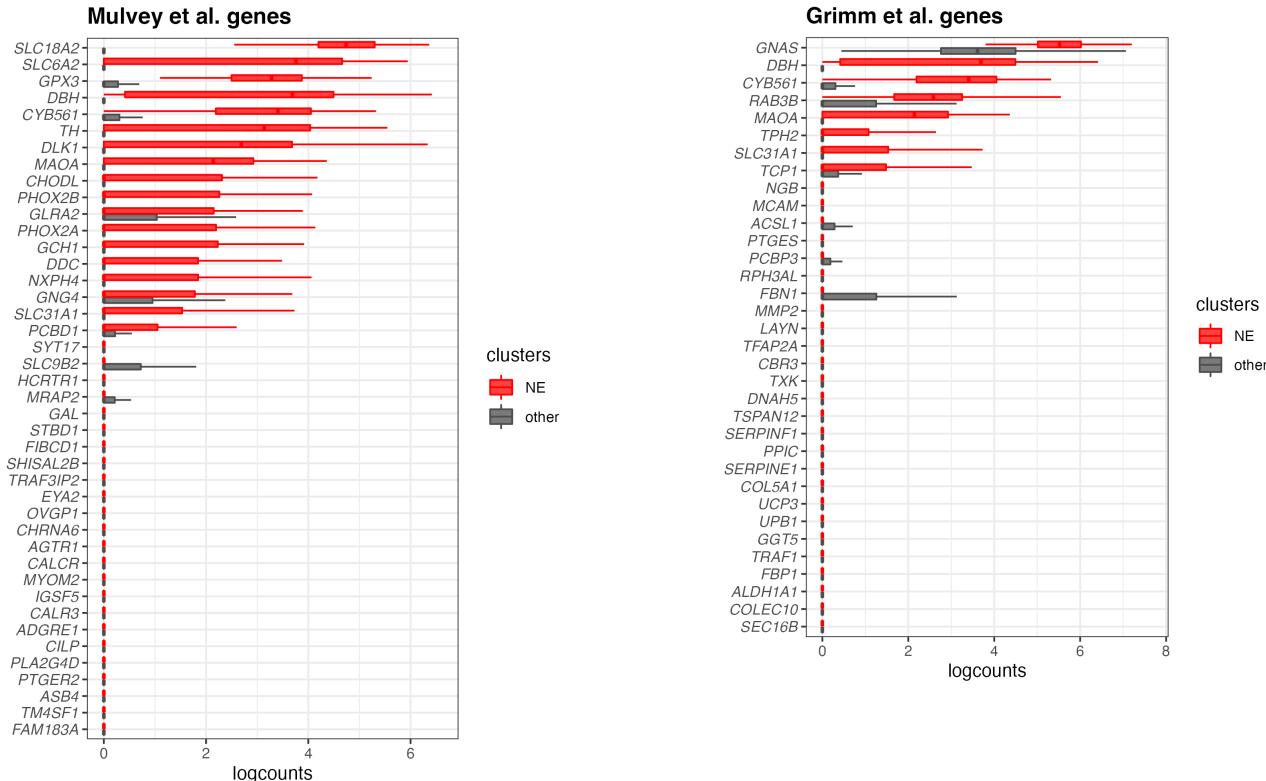
LC: Visium analyses

Partial conservation of NE neuron-associated genes across species based on comparison with previous studies in rodents using alternative technological platforms



LC: Single-nucleus analyses

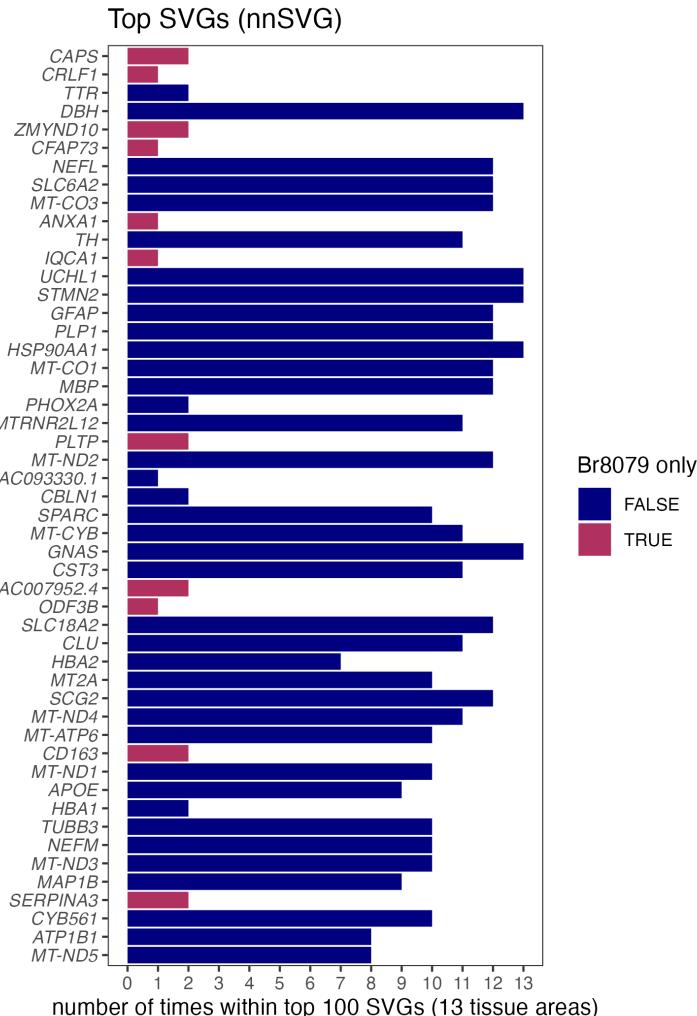
Partial conservation of NE neuron-associated genes across species based on comparison with previous studies in rodents using alternative technological platforms



LC: Visium analyses

Visium analyses using nnSVG

- Identify top LC-associated SVGs replicated across samples
- Number of times each gene is ranked within top 100 SVGs in the 13 tissue areas (Visium samples)
- Donor Br8079: choroid plexus genes

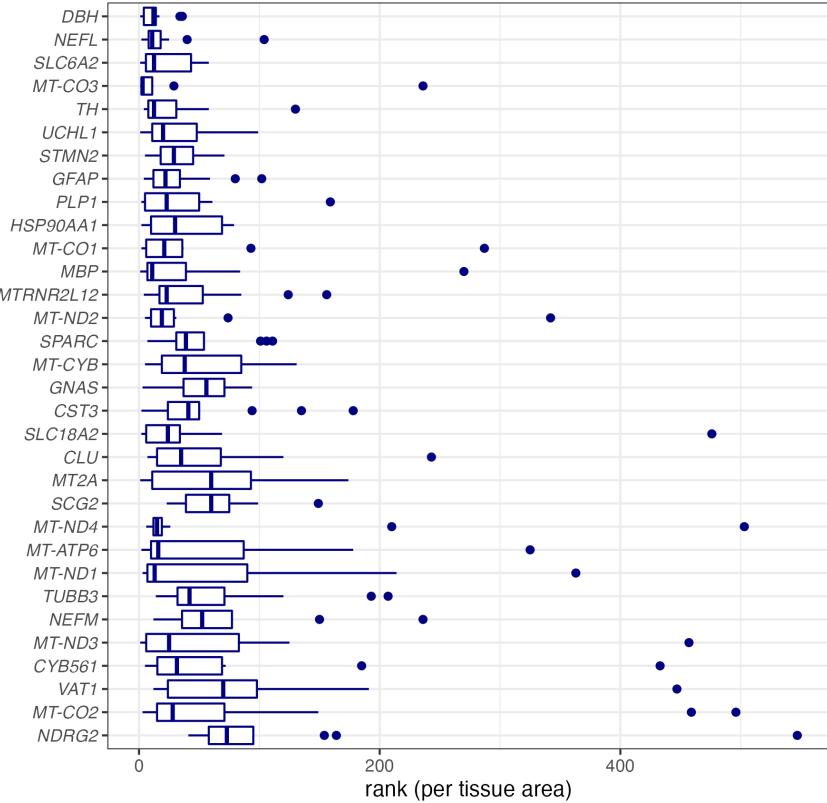


LC: Visium analyses

Visium analyses using nnSVG

- Excluding choroid plexus genes results in list of top LC-associated SVGs replicated across multiple Visium samples
- Rank per tissue area for genes ranked within top 100 SVGs in at least 10 out of 13 tissue areas

Top SVGs (nnSVG): replicated across tissue areas



Benchmarking studies

Guidelines / review paper

- How to design / perform rigorous benchmarking studies?
- Independent benchmarking vs. benchmarking during method development

Weber et al. *Genome Biology* (2019) 20:125
<https://doi.org/10.1186/s13059-019-1738-8>

Genome Biology

REVIEW

Open Access



Essential guidelines for computational method benchmarking

Lukas M. Weber^{1,2}, Wouter Saelens^{3,4}, Robrecht Cannoodt^{3,4}, Charlotte Soneson^{1,2*}, Alexander Hapfelmeier⁵, Paul P. Gardner⁶, Anne-Laure Boulesteix⁷, Yvan Saey^{3,4*} and Mark D. Robinson^{1,2}

Abstract

In computational biology and other sciences, researchers are frequently faced with a choice between several computational methods for performing data analyses. Benchmarking studies aim to rigorously compare the performance of different methods using well-characterized benchmark datasets, to determine the strengths of each method or to provide recommendations regarding suitable choices of methods for an analysis. However, benchmarking studies must be carefully designed and implemented to provide accurate, unbiased, and informative results. Here, we summarize key practical guidelines and recommendations for performing high-quality benchmarking analyses, based on our experiences in computational biology.

Box 1: Summary of guidelines

The guidelines in this review can be summarized in the following set of recommendations. Each recommendation is discussed in more detail in the corresponding section in the text.

1. Define the purpose and scope of the benchmark.
2. Include all relevant methods.
3. Select (or design) representative datasets.
4. Choose appropriate parameter values and software versions.
5. Evaluate and rank methods according to key quantitative performance metrics.
6. Evaluate secondary measures including runtimes and computational requirements, user-friendliness, code quality, and documentation quality.
7. Interpret results and provide guidelines or recommendations from both user and method developer perspectives.
8. Publish and distribute results in an accessible format.
9. Design the benchmark to enable future extensions.
10. Follow reproducible research best practices, in particular by making all code and data publicly available.

Independent benchmarking studies

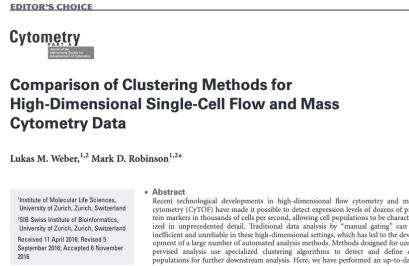
- Weber et al. (2021), *GigaScience*
- Weber et al. (2016), *Cyt Part A*

Benchmarking during method development

- nnSVG: Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*
- diffcyt: Weber et al. (2019), *Comms Biol*

Benchmarking examples / papers

Independent benchmarking



Benchmarking during method development



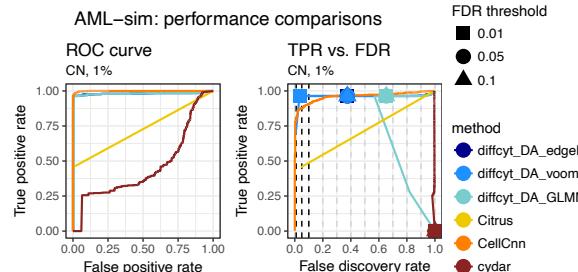
ARTICLE

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

OPEN

diffcyt: Differential discovery in high-dimensional cytometry via high-resolution clustering

Lukas M. Weber^{1,2}, Małgorzata Nowicka^{1,2,3}, Charlotte Soneson^{1,2,4} & Mark D. Robinson^{1,2}



bioRxiv posts many COVID-19-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.

[Follow this preprint](#)

New Results

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

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

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

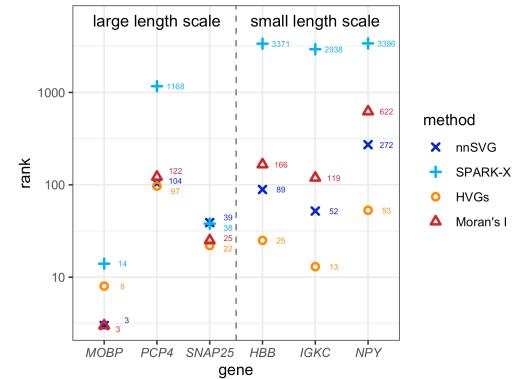
This article is a preprint and has not been certified by peer review (what does this mean?).



Abstract Full Text Info/History Metrics

Preview PDF

Selected SVGs: human DLPFC



SpatialExperiment

R/Bioconductor class to store spatially-resolved transcriptomics datasets

