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

1 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)

JOHNS HOPKINS
BLOOMBERG SCHOOL
of PUBLIC HEALTH

University of Zurich[UZH]

ETH zürich

lmweber.org
github.com/lmweber
@lmwebr

Research Theme 1

Unsupervised statistical methods

1

Spatially-resolved transcriptomics: **nnSVG**
- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: **diffcyt**
- Weber et al. (2019), *Comms Biol*

Research Theme 2

Collaborative analyses

2

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

Cancer

Immunology

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

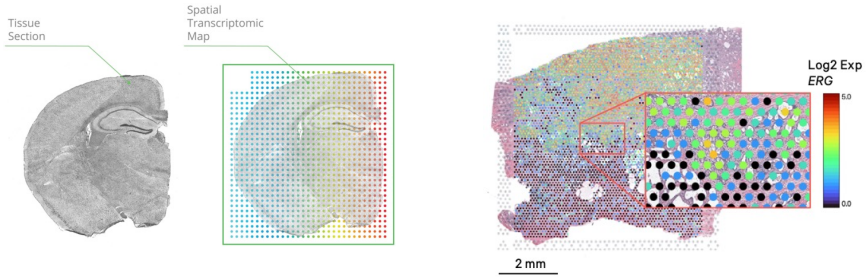
Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform



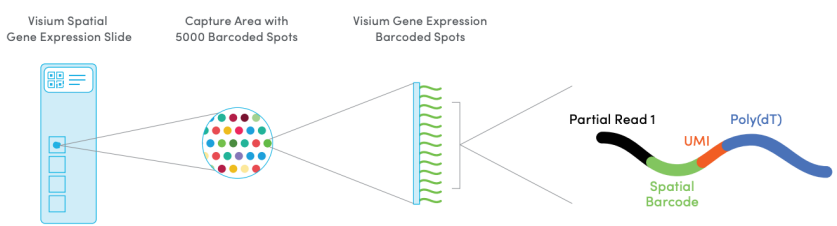


Image source: 10x Genomics

Spatially-resolved transcriptomics

Transcriptome-wide gene expression at spatial resolution

Example: 10x Genomics Visium platform

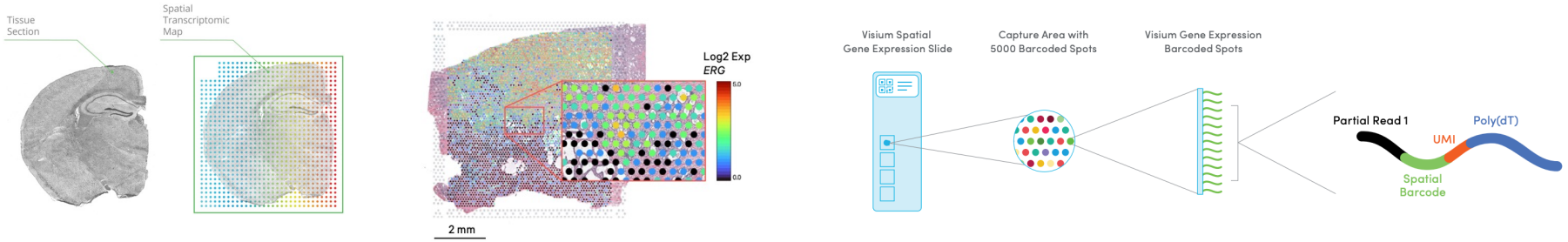


Image source: 10x Genomics

Illustration: cell populations within brain

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

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

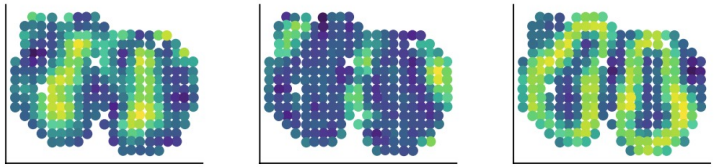
Spatially-resolved transcriptomics

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

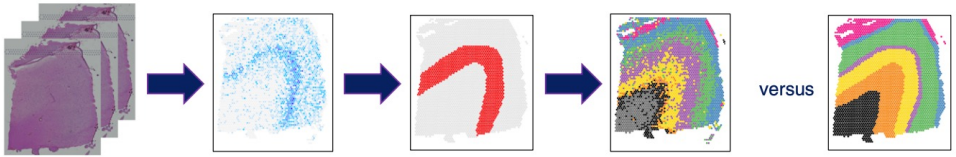
- feature selection: spatially variable genes
- clustering: spatial domains or spatially distributed cell populations
- differential gene expression

Cell type / state heterogeneity

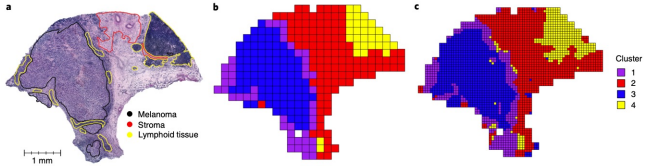
Biologically informative / marker genes



Svensson et al. (2018)



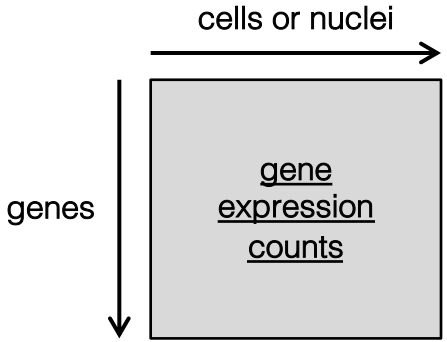
Maynard and Collado-Torres et al. (2021)



Zhao et al. (2021)

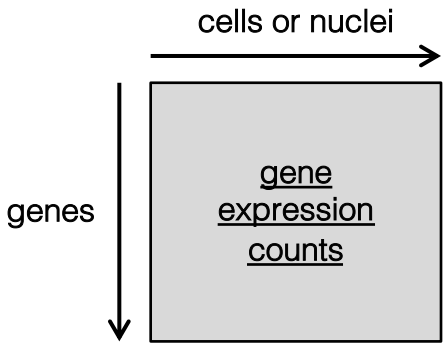
Data structure

Single-cell / single-nucleus RNA sequencing

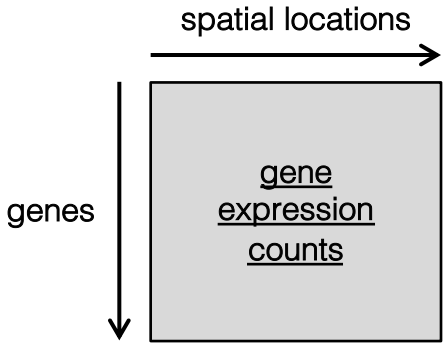


Data structure

Single-cell / single-nucleus RNA sequencing



Spatially-resolved transcriptomics



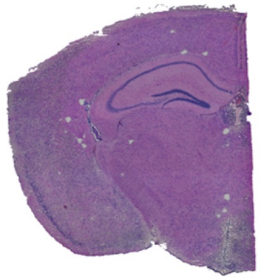
+

spatial coordinates

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

+

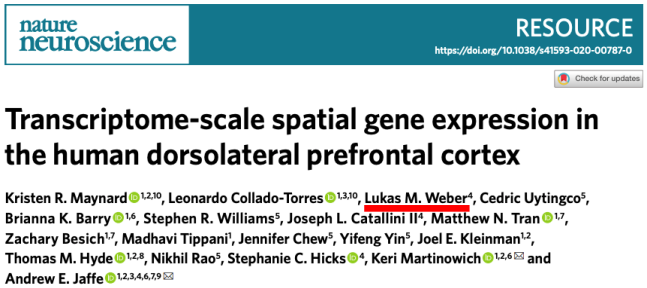
image features



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)



Spatially variable genes

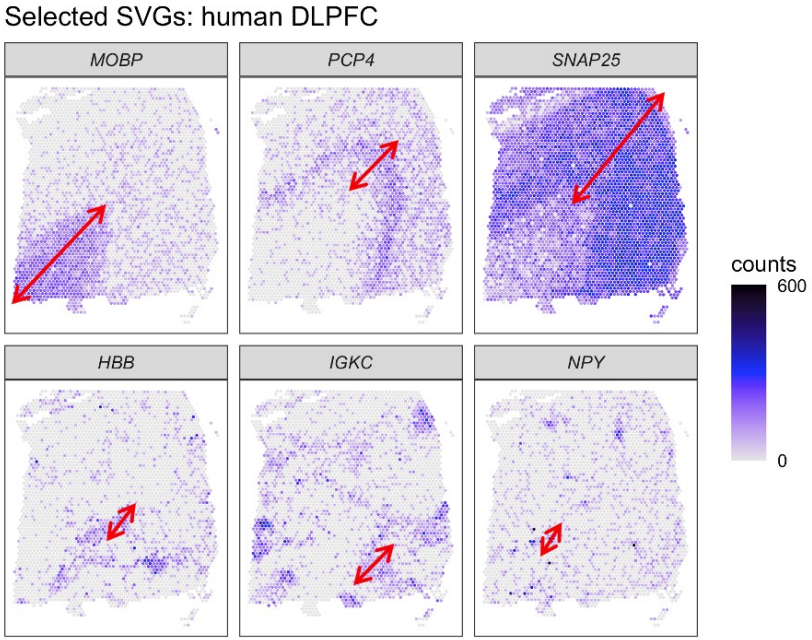
Example

- 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)

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

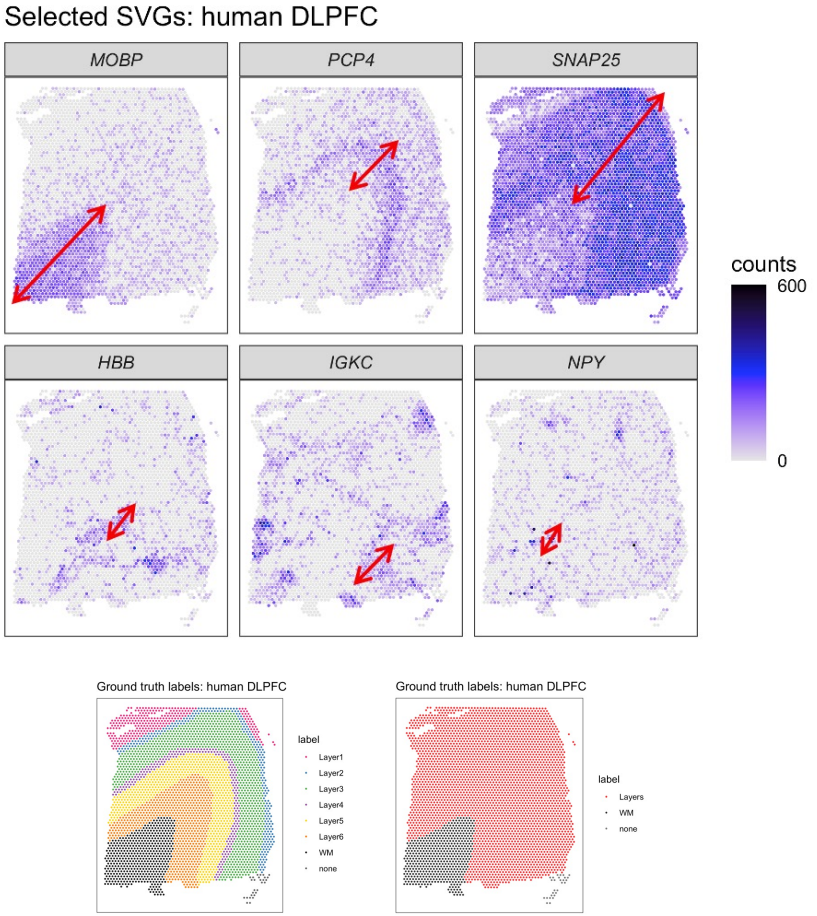
Example

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

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] ✉



Research Theme 1

Unsupervised statistical methods

1

Spatially-resolved transcriptomics: **nnSVG**

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

High-dimensional cytometry: **diffcyt**

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

Research Theme 2

Collaborative analyses

2

Neuroscience

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

Cancer

Immunology

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

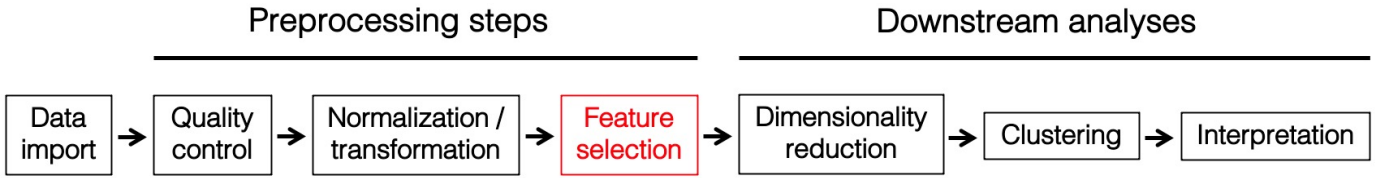
- **R/Bioconductor:** Righelli, Weber, Crowell et al. (2021), *Bioinf*

Open-source software / reproducible research

additional papers on website and Google Scholar

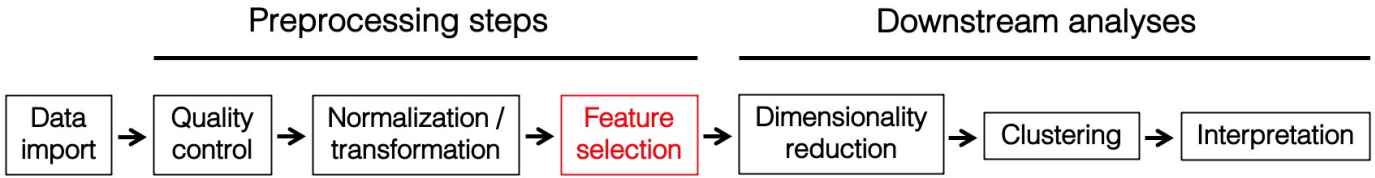
Why are we interested in spatially variable genes?

Feature selection



Why are we interested in spatially variable genes?

Feature selection



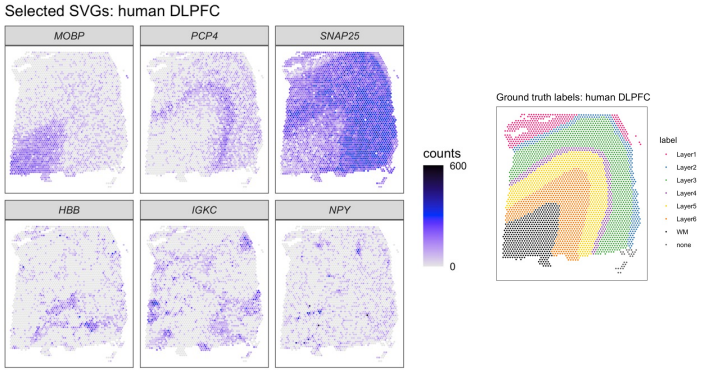
Data preprocessing

- 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

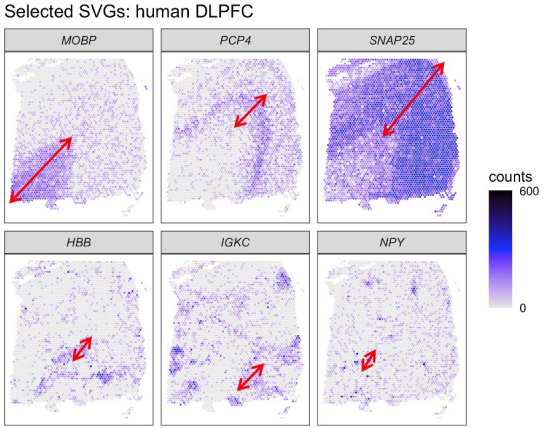
Example: human DLPFC



Existing methods are too slow or perform poorly

Baseline methods

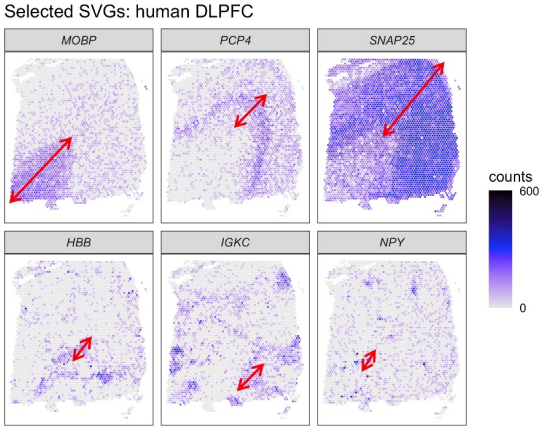
- Highly variable genes (HVGs)
- Moran's I statistic



Existing methods are too slow or perform poorly

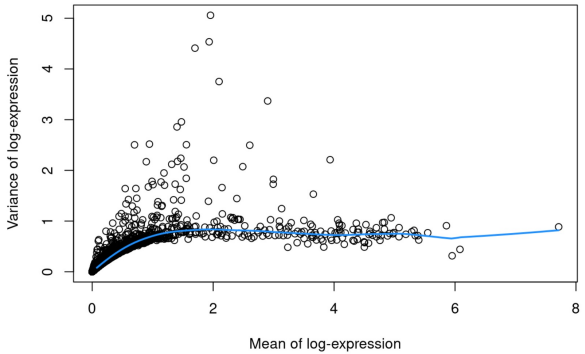
Baseline methods

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



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

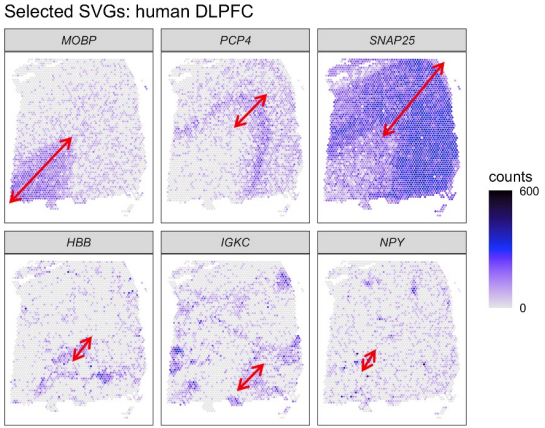
Examples of new methods for spatially variable genes (SVGs)

- **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)

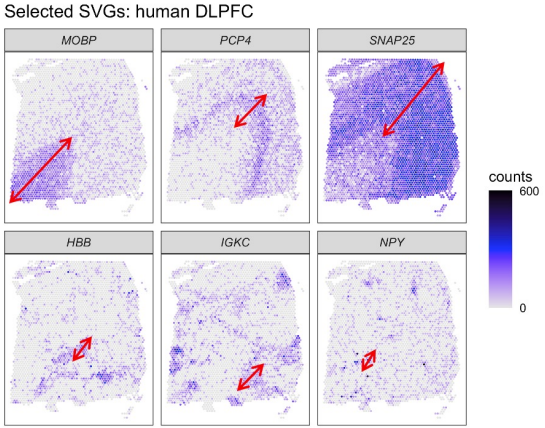
  Fast approximation loses sensitivity to spatial patterns with varying length scales



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 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)

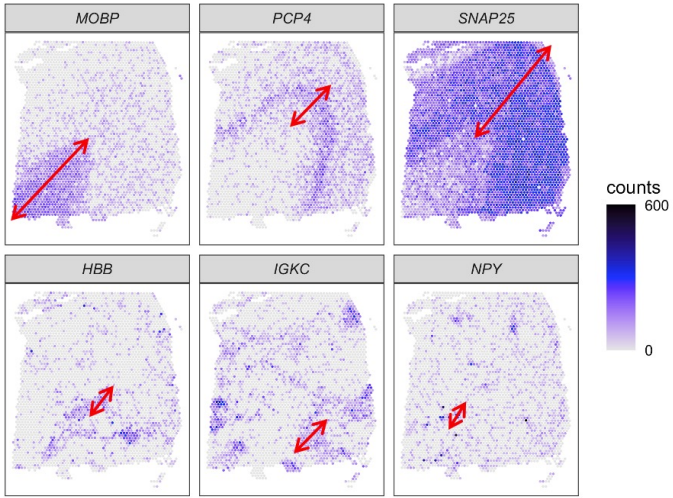
$$y \sim N(X\beta, C(\theta) + \tau^2 I)$$

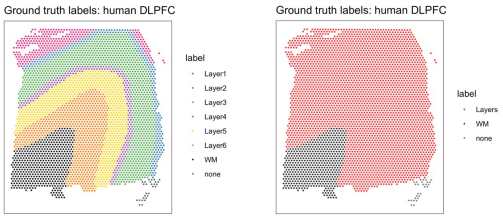
$$C(\theta) = k(s_i, s_j) = \sigma^2 \exp\left(\frac{-||s_i - s_j||}{l}\right)$$

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

nnSVG evaluations / benchmarking

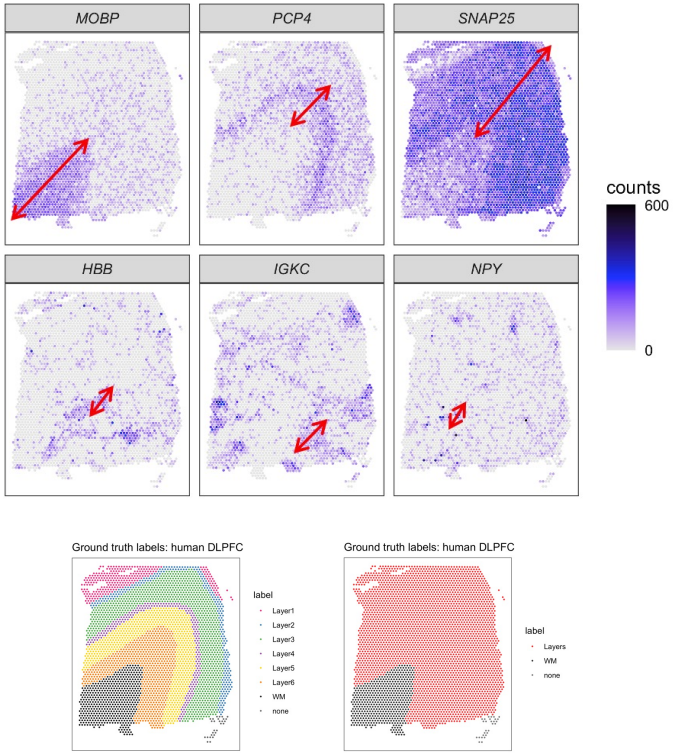
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)

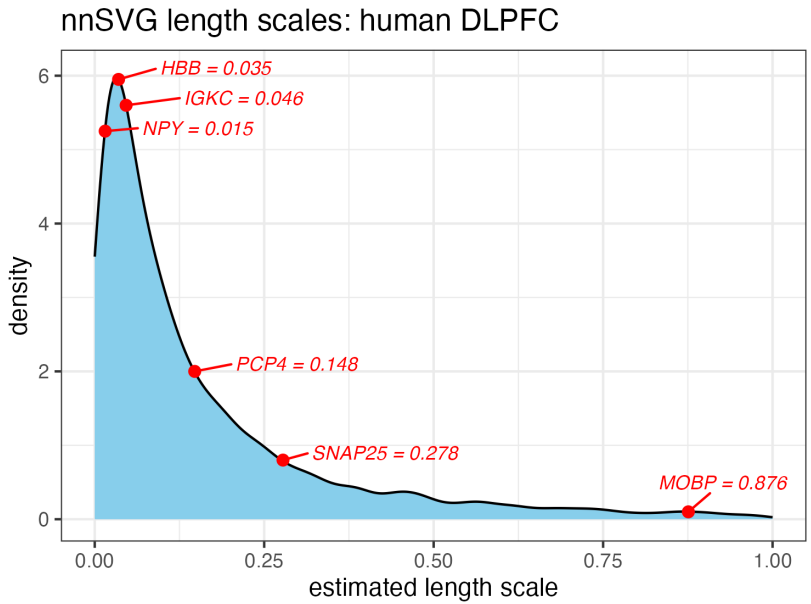




nnSVG evaluations / benchmarking

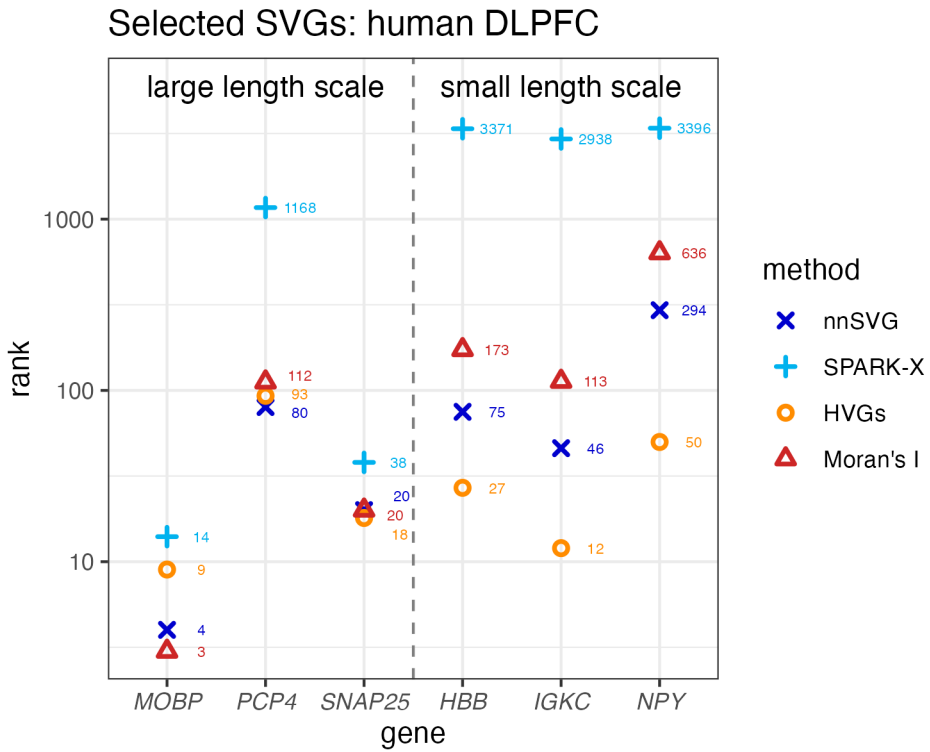
Human DLPFC dataset (10x Genomics Visium) (Maynard and Collado-Torres et al. 2021)





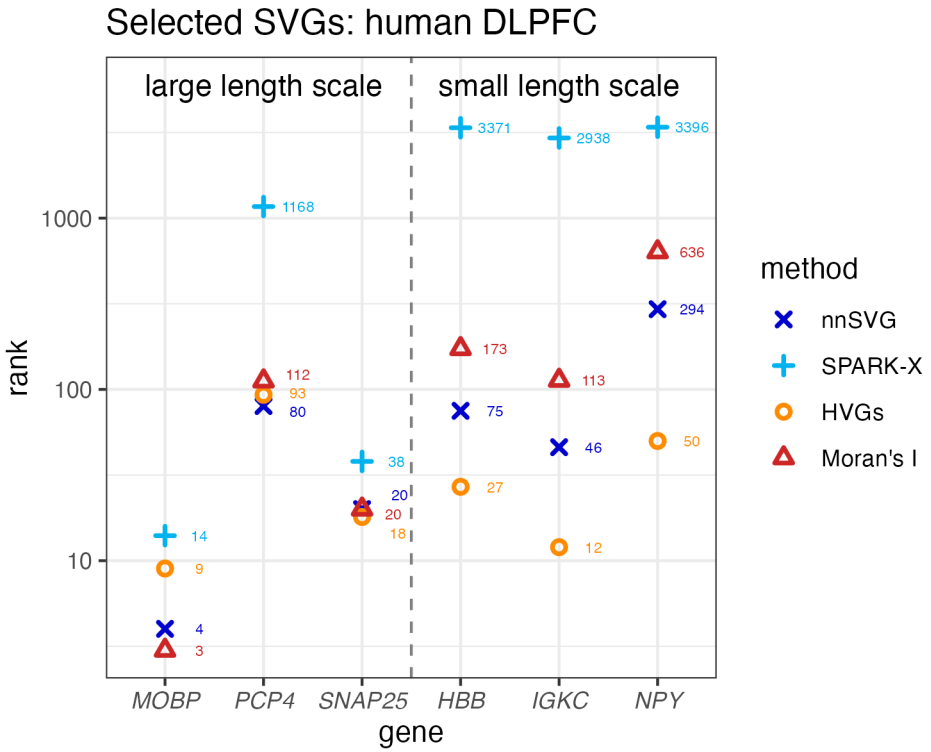
nnSVG evaluations / benchmarking

Human DLPFC dataset



nnSVG evaluations / benchmarking

Human DLPFC dataset

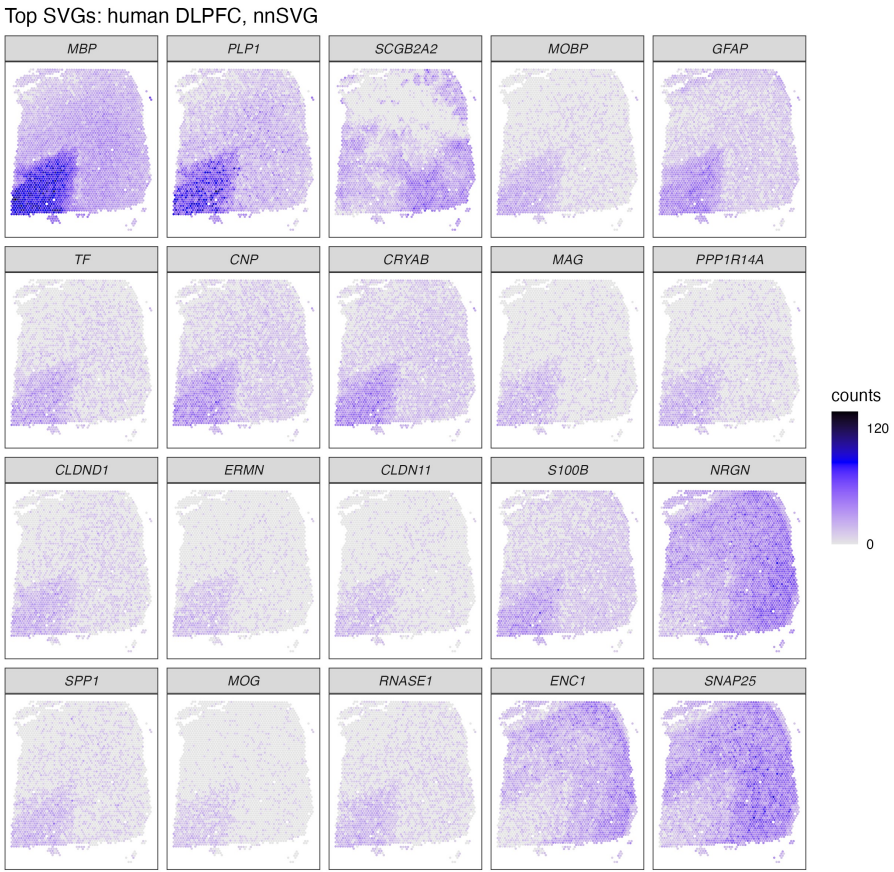


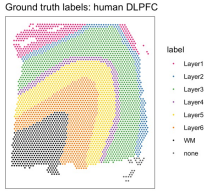
Method performance

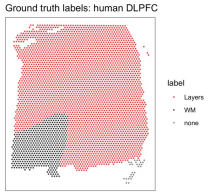
HVGs > nnSVG > Moran's I >> SPARK-X

nnSVG evaluations / benchmarking

Human DLPFC dataset



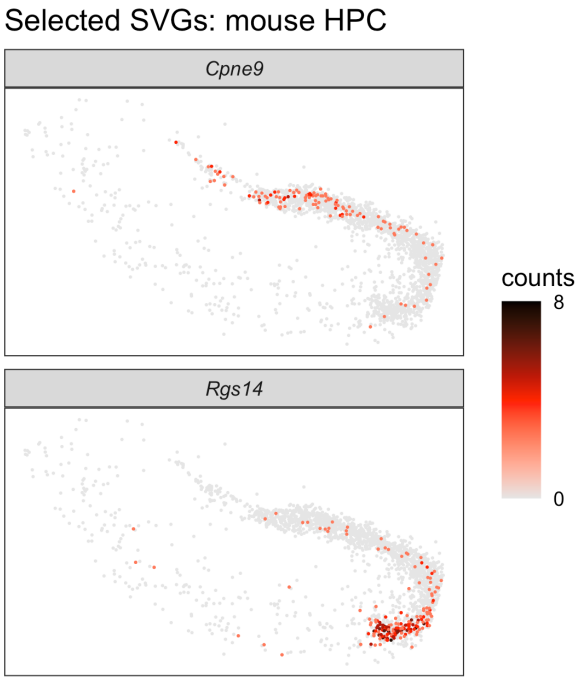


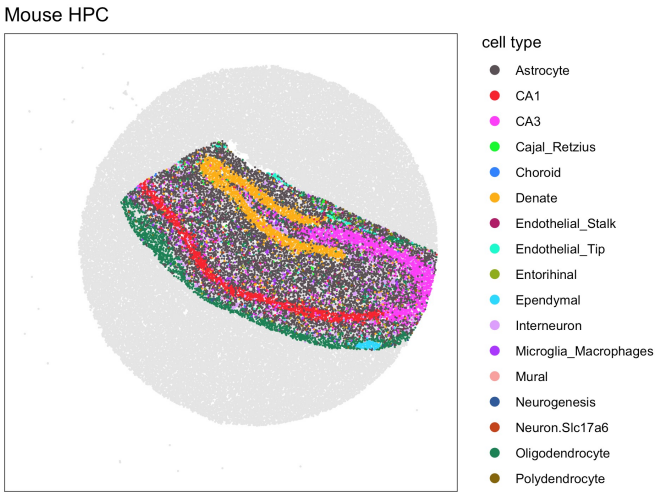


nnSVG evaluations / benchmarking

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

- Identify SVGs within spatial domain

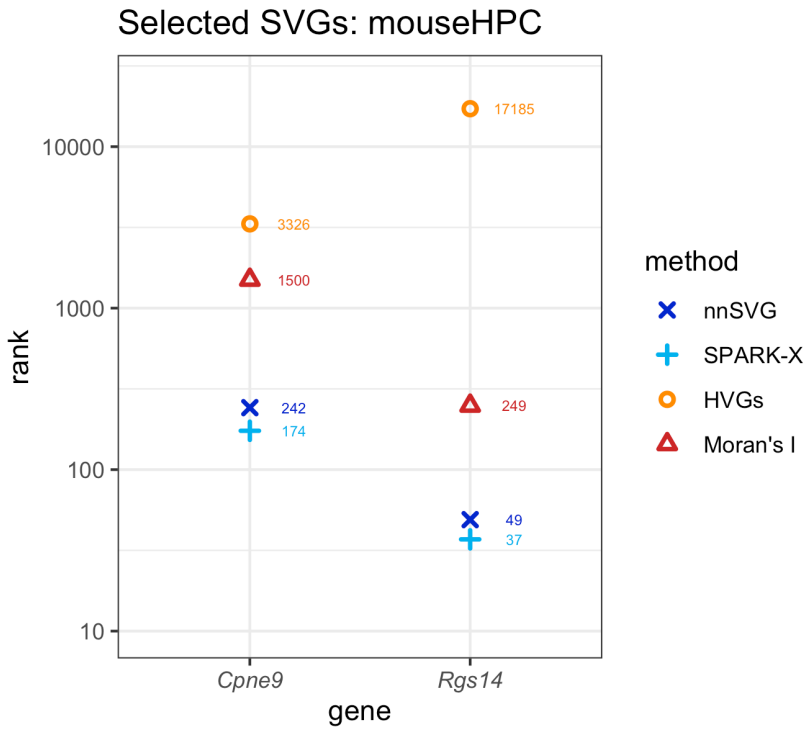




nnSVG evaluations / benchmarking

Mouse hippocampus dataset

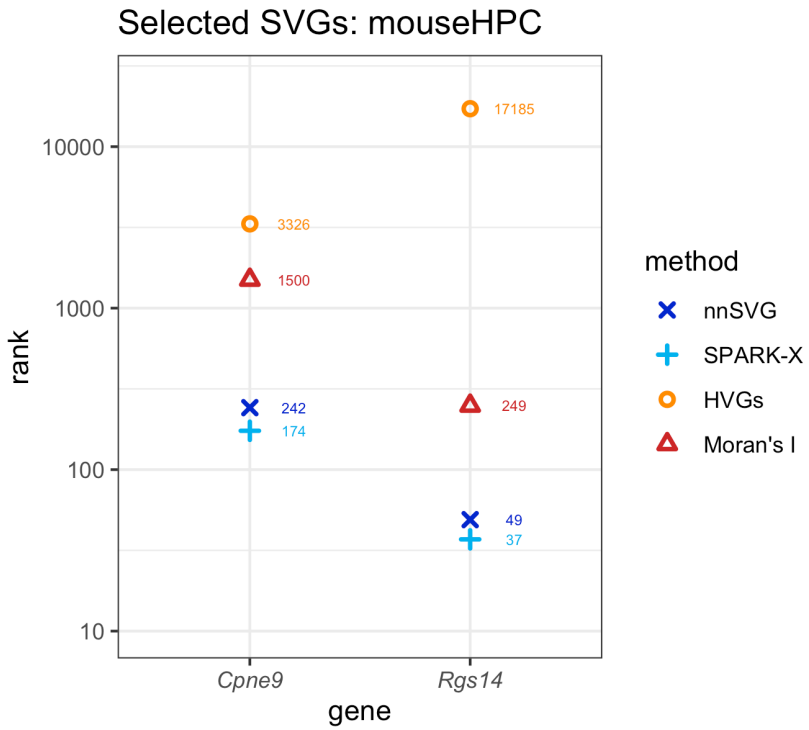
- With covariates for spatial domains



nnSVG evaluations / benchmarking

Mouse hippocampus dataset

- With covariates for spatial domains



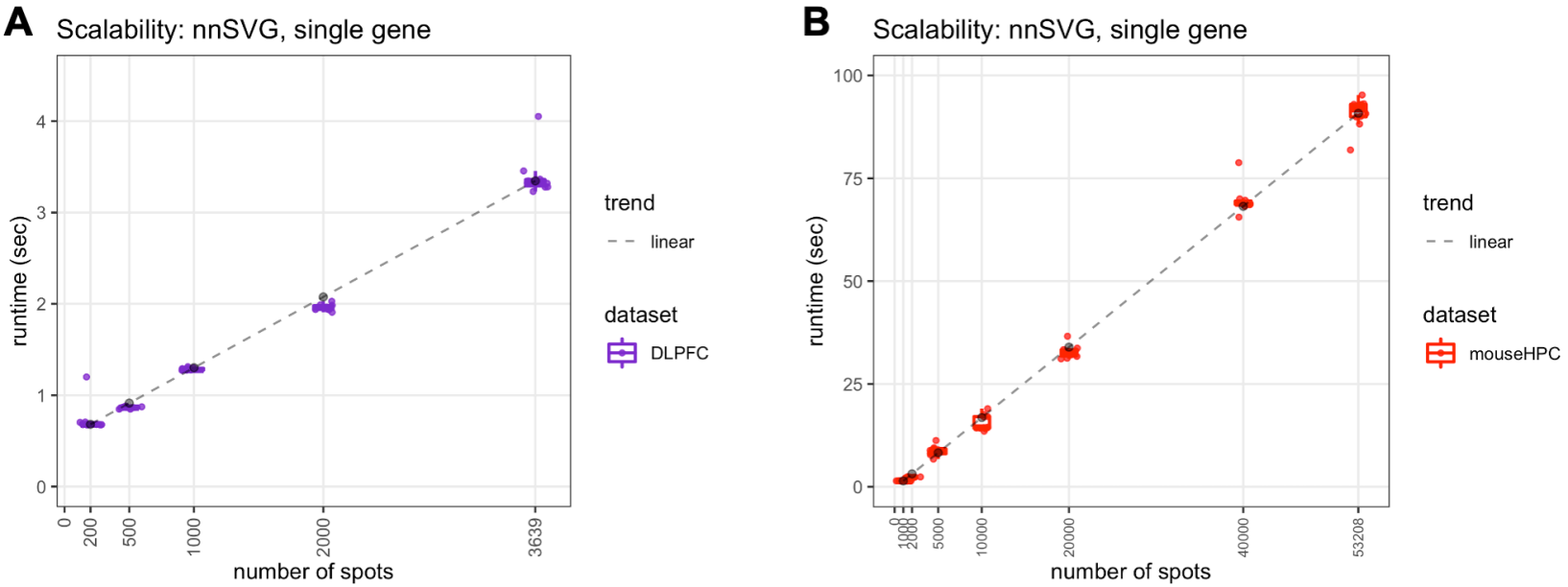
Method performance

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

nnSVG evaluations / benchmarking

Computational scalability

- Linear in number of spatial locations

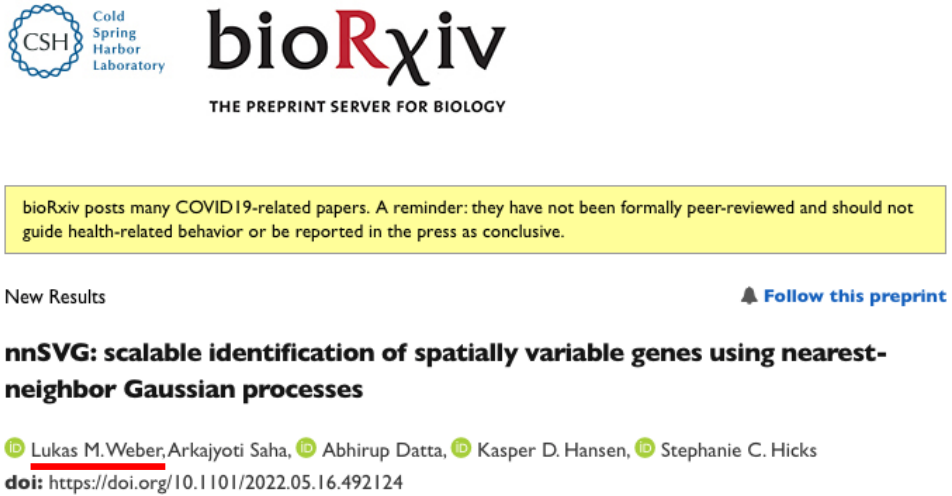


nnSVG summary

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)



nnSVG implementation and reproducibility

R/Bioconductor package

- Open-source software package
- Documentation and tutorial

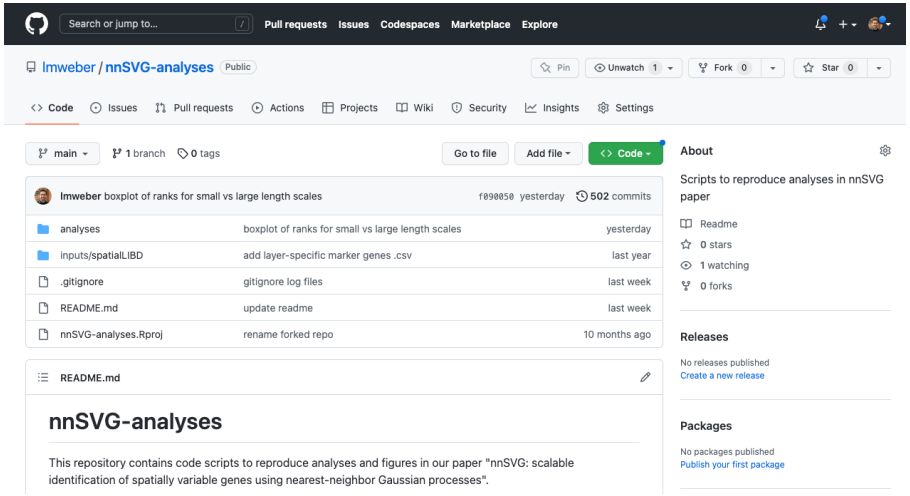
nnSVG implementation and reproducibility

R/Bioconductor package

- Open-source software package
- Documentation and tutorial

GitHub repository

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



Research Theme 1

Unsupervised statistical methods

1

Spatially-resolved transcriptomics: **nnSVG**

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

High-dimensional cytometry: **diffcyt**

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

Research Theme 2

Collaborative analyses

2

Neuroscience

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

Cancer

Immunology

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

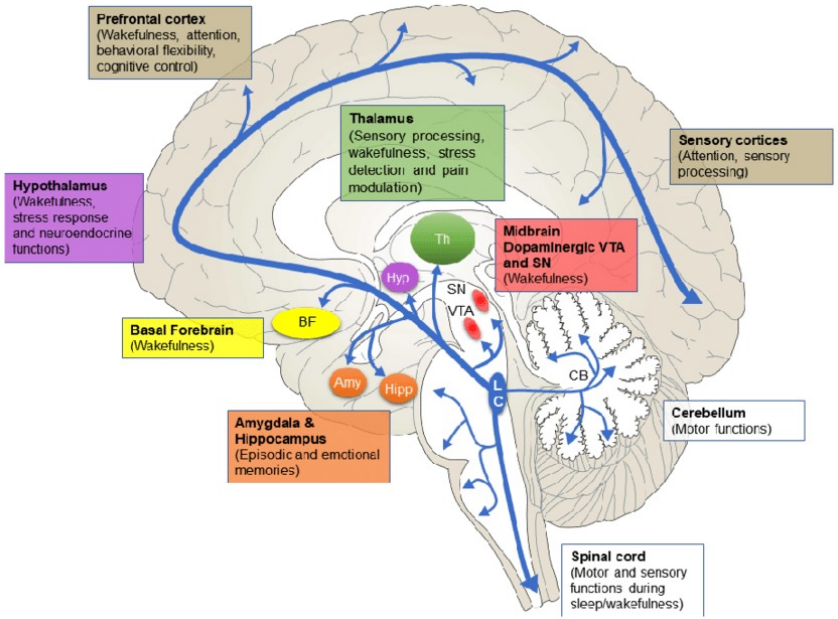
Locus coeruleus (LC)

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

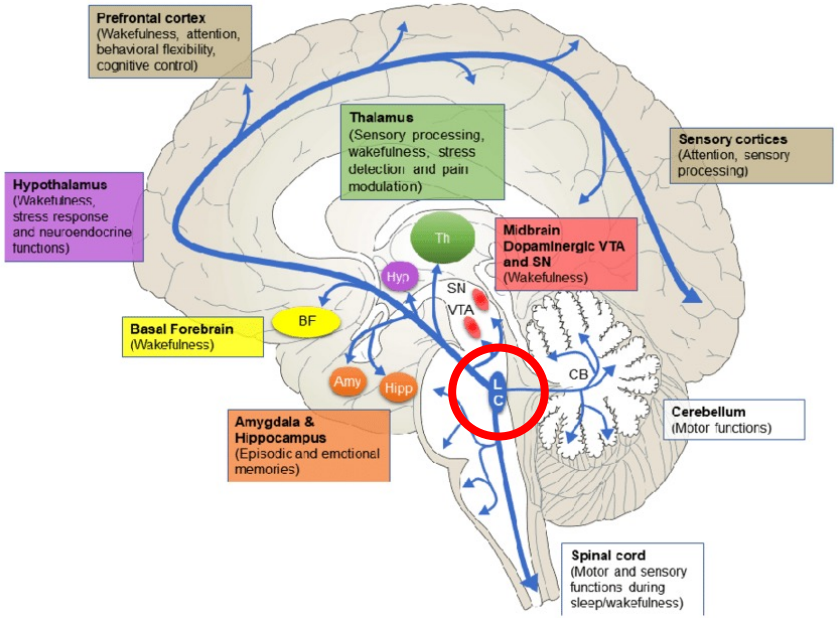


Bari et al. (2020), *Neural Regen Res*

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



Bari et al. (2020), *Neural Regen Res*

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

Study overview

Aim

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

Study overview

Heena Divecha
Lieber Institute for Brain Development

Aim

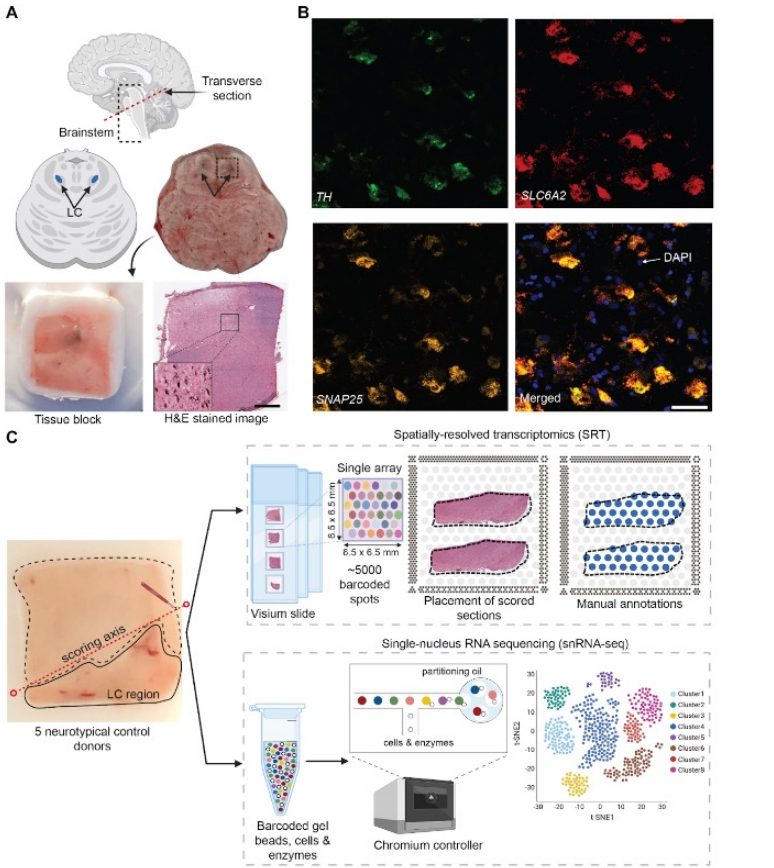
- Characterize transcriptome-wide gene expression landscape of human LC at spatial and single-nucleus resolution

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



Differential gene expression

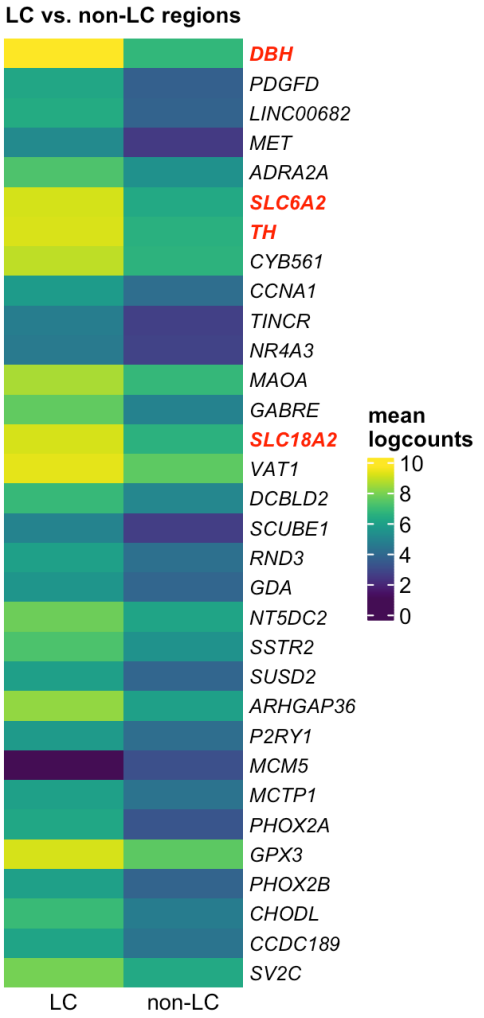
Differential gene expression testing in Visium data

- 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

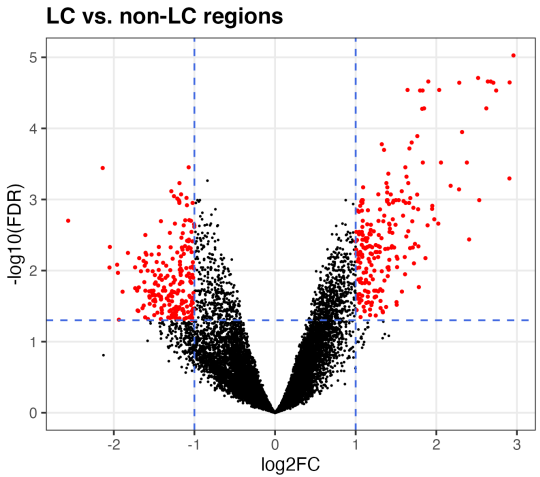
- Manually annotated LC regions
- Recovers known NE neuron markers and additional unsupervised gene list
- 437 statistically significant genes

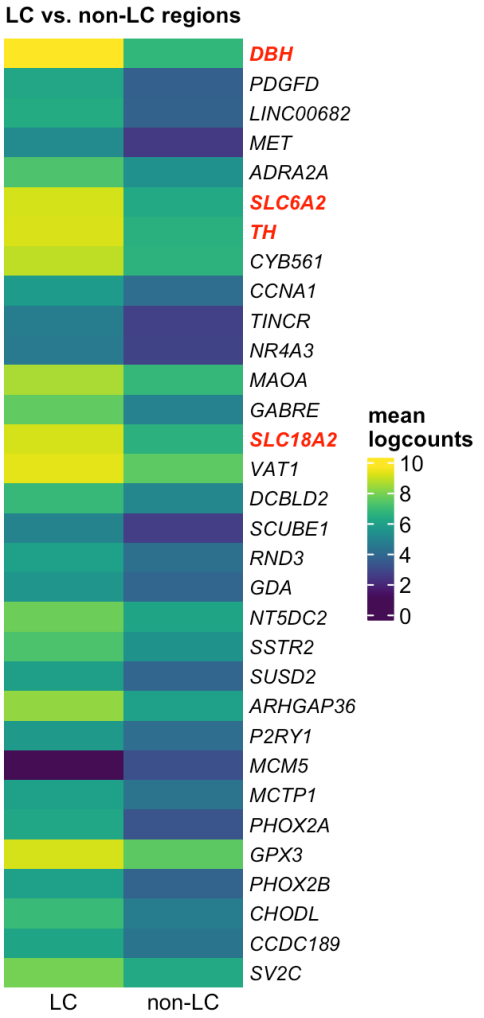


Differential gene expression

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

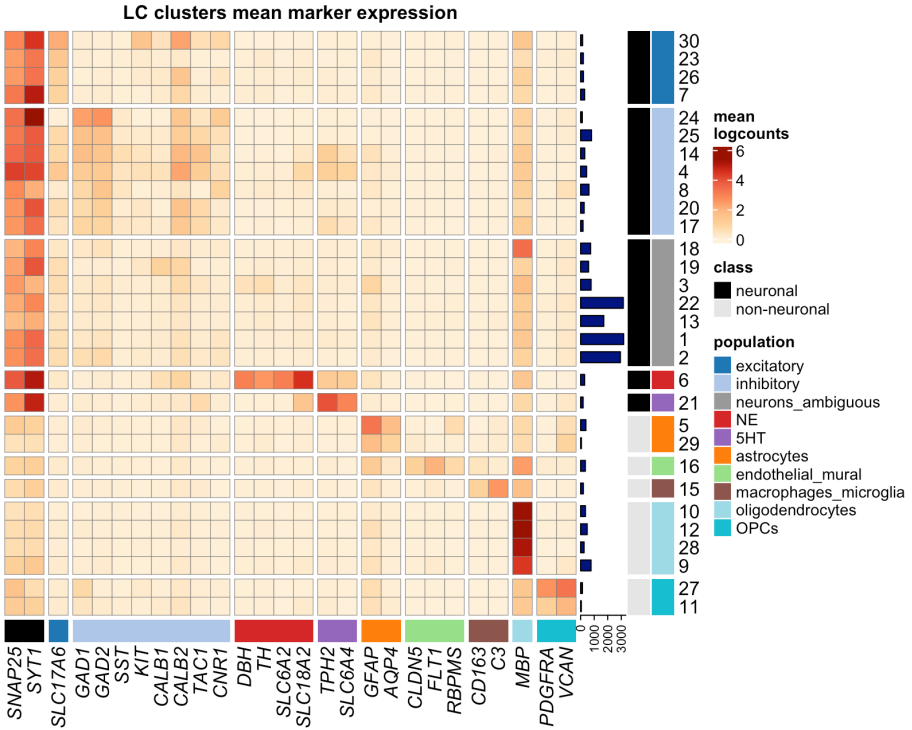
Unsupervised clustering in single-nucleus RNA sequencing data

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

Unsupervised clustering

Unsupervised clustering in single-nucleus RNA sequencing data

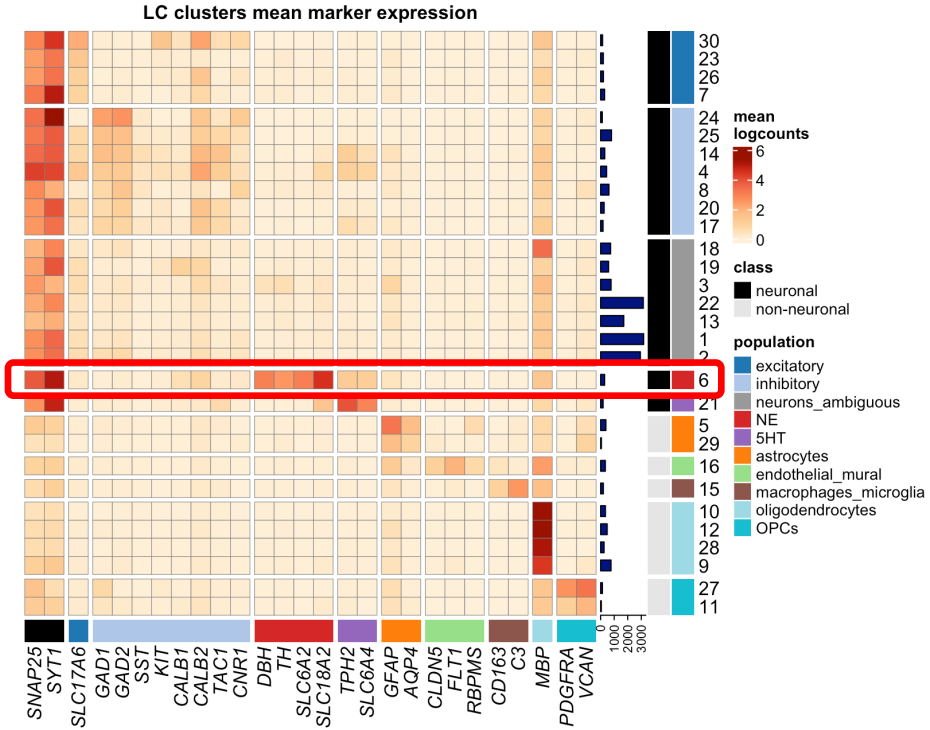
- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei



Unsupervised clustering

Unsupervised clustering in single-nucleus RNA sequencing data

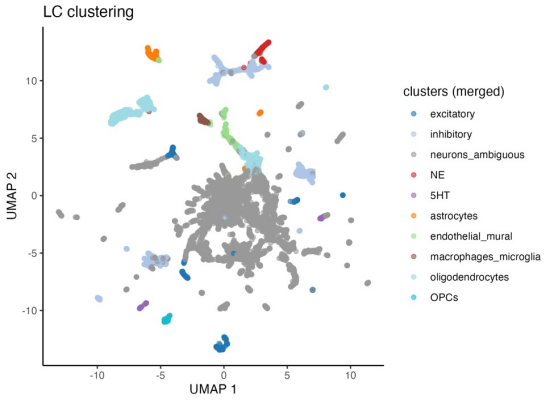
- 30 clusters representing neuronal and non-neuronal cell populations
- NE neuron cluster: 295 nuclei

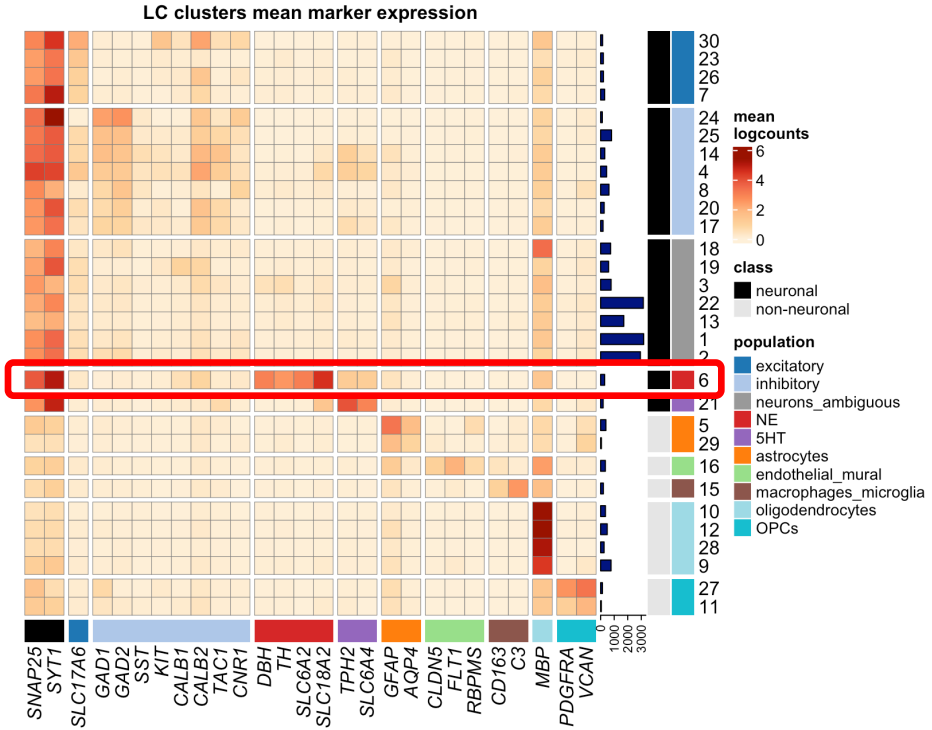


Unsupervised clustering

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

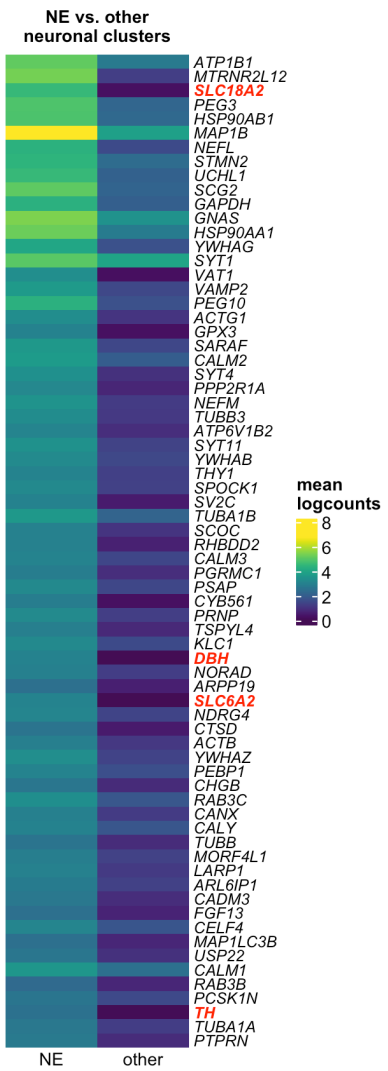
Differential gene expression testing in single-nucleus RNA sequencing data

- 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

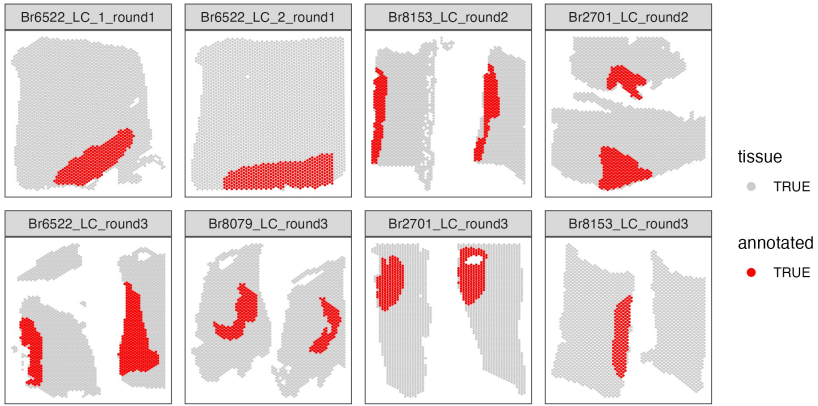
- nnSVG to identify spatially variable genes (SVGs) in Visium samples
- Top SVGs include choroid plexus-associated genes for samples from one donor (Br8079)

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

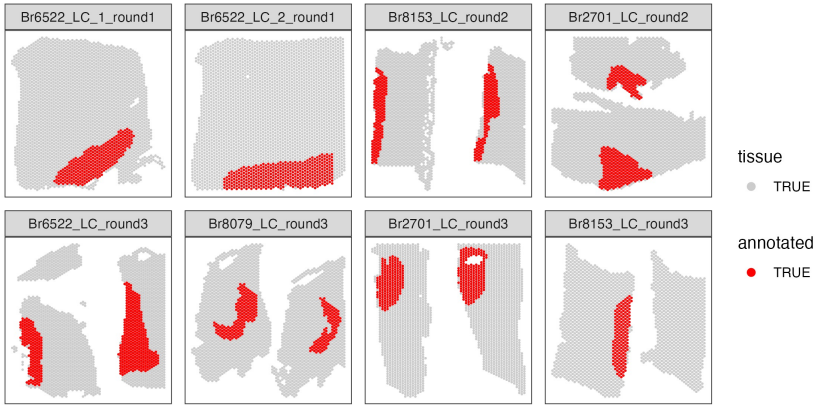


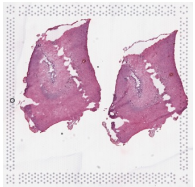
Unsupervised analyses recover additional unexpected results

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





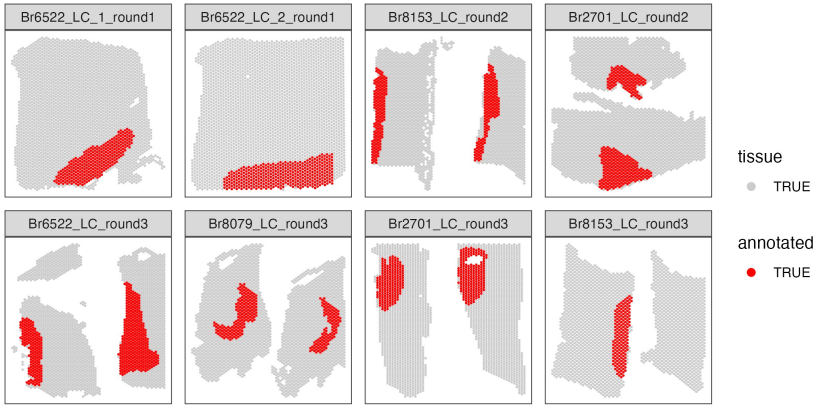
Histology image for donor Br8079

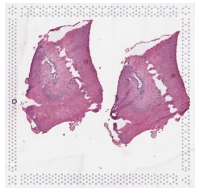
Unsupervised analyses recover additional unexpected results

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)

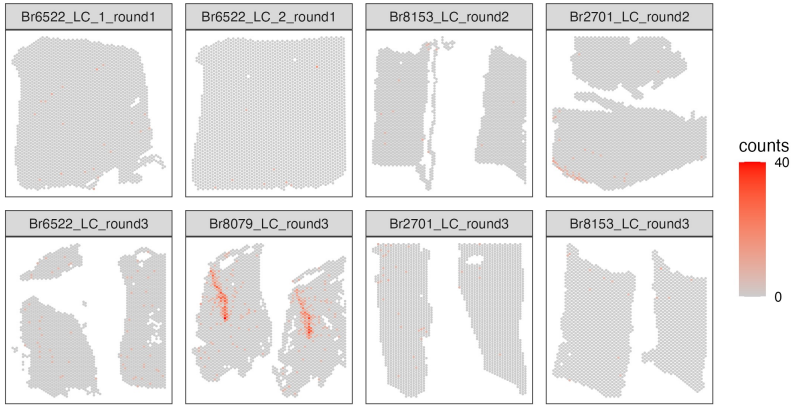
Manual annotations: LC regions





Histology image for donor Br8079

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

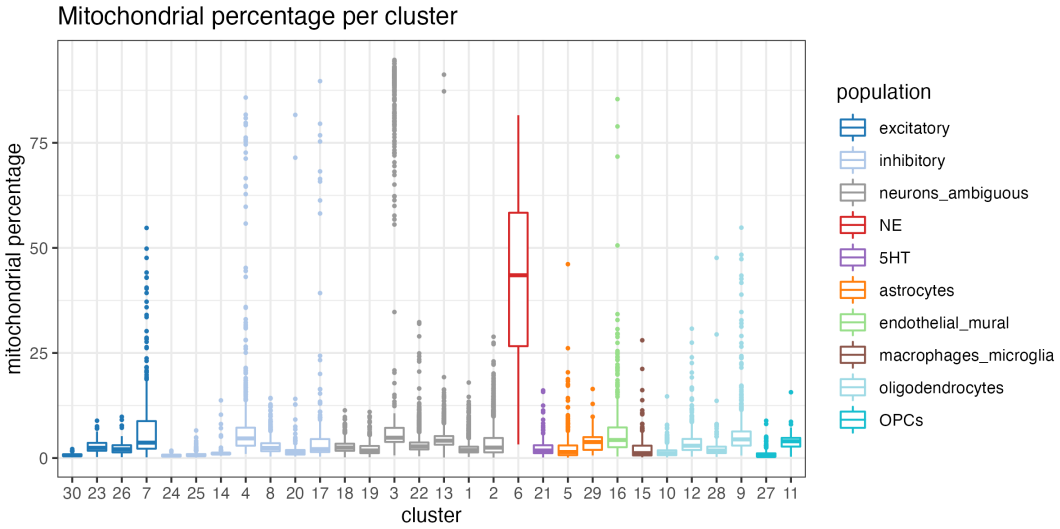
Unsupervised analyses recover additional unexpected results

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!)



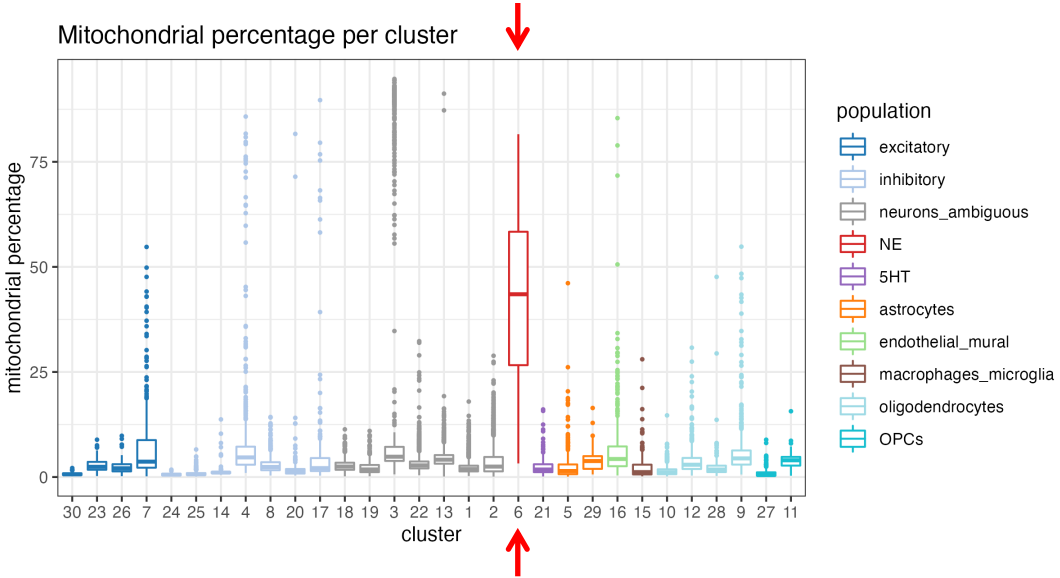
Unsupervised analyses recover additional unexpected results

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!)



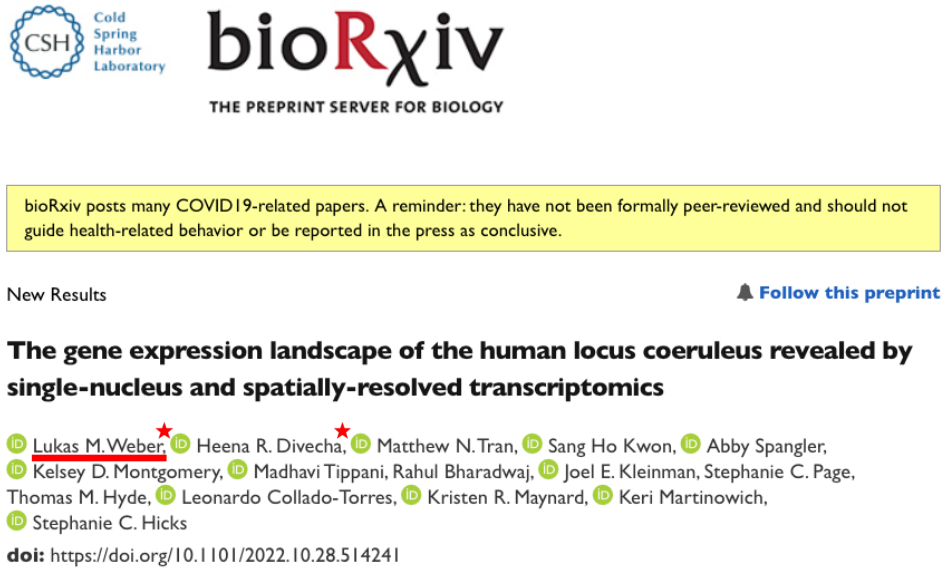
Summary: locus coeruleus (LC) collaboration analysis

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

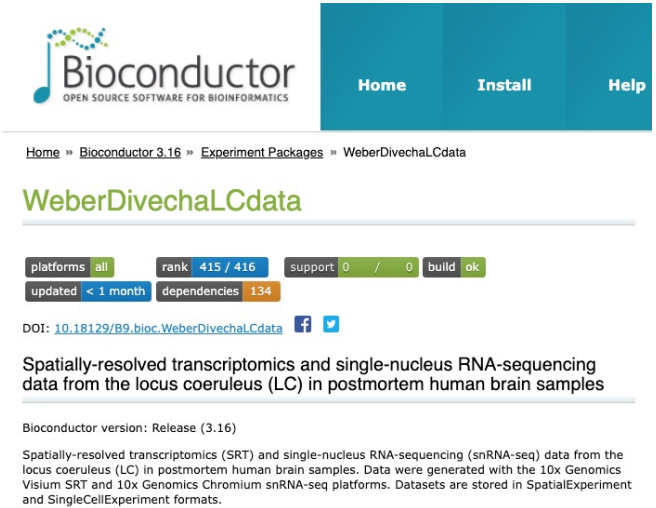
Weber and Divecha et al. (2022)
bioRxiv preprint; accepted (eLife)



Data availability

R/Bioconductor data package

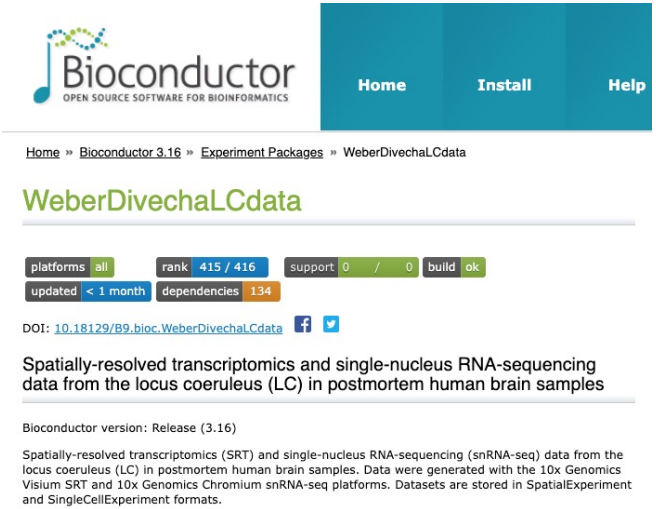
- Downloadable R objects



Data availability

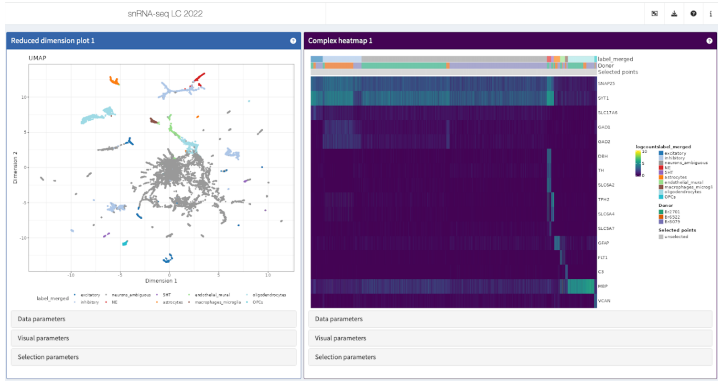
R/Bioconductor data package

- Downloadable R objects



Web apps

- Interactive exploration of data



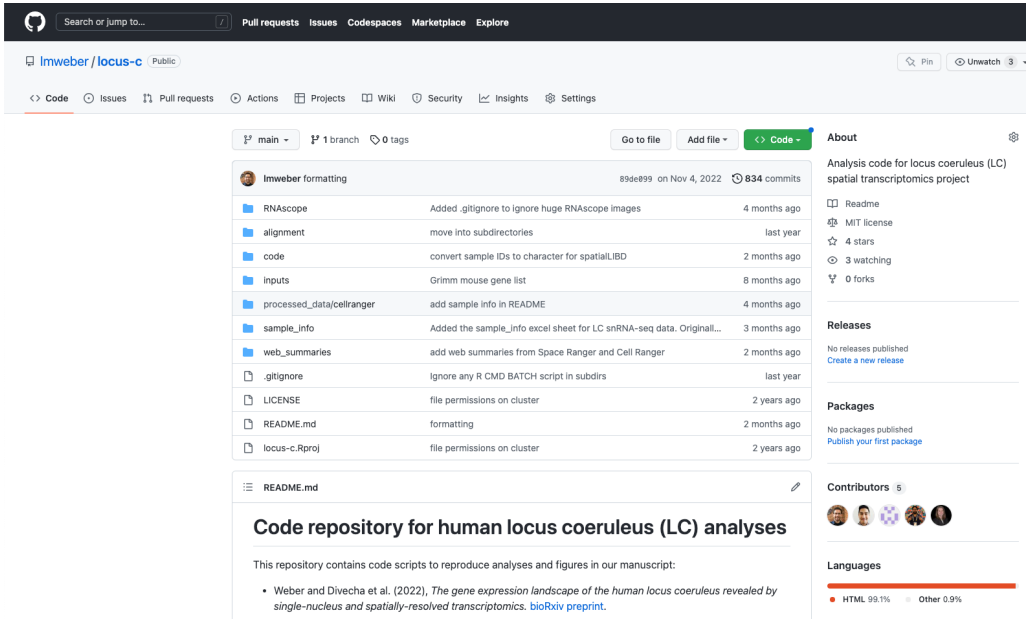
Heena Divecha

Lieber Institute for Brain Development

Reproducibility

GitHub repository

- Code to reproduce analyses and figures



Research Theme 1

Unsupervised statistical methods

1

Spatially-resolved transcriptomics: **nnSVG**
- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: **diffcyt**
- Weber et al. (2019), *Comms Biol*

Research Theme 2

Collaborative analyses

2

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

Cancer

Immunology

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

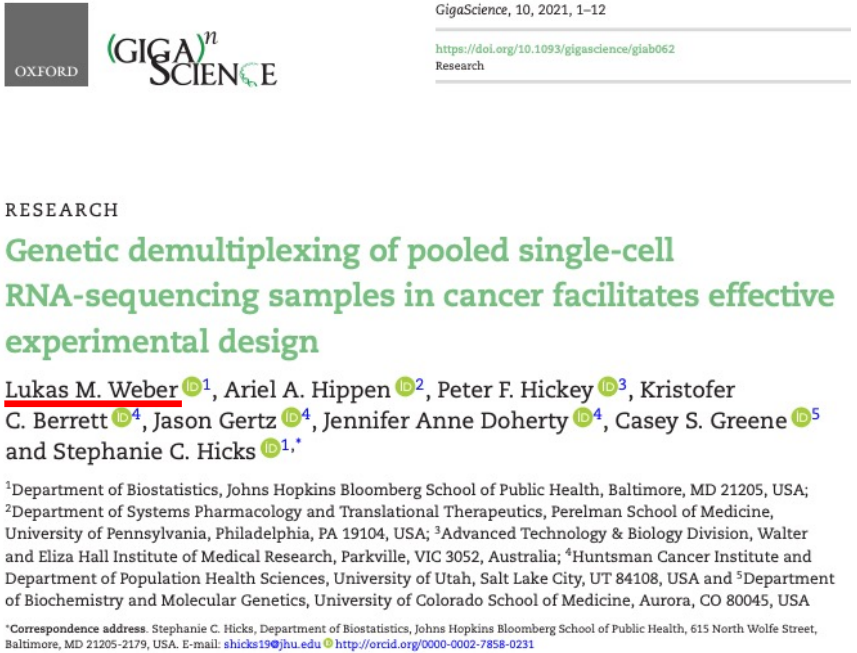
Which methods perform well in which types of data / for which types of analyses?

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



OXFORD (GIGA)n SCIENCE

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,*]

[1]Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, USA; [2]Department of Systems Pharmacology and Translational Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA 19104, USA; [3]Advanced Technology & Biology Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC 3052, Australia; [4]Huntsman Cancer Institute and Department of Population Health Sciences, University of Utah, Salt Lake City, UT 84108, USA and [5]Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, CO 80045, USA

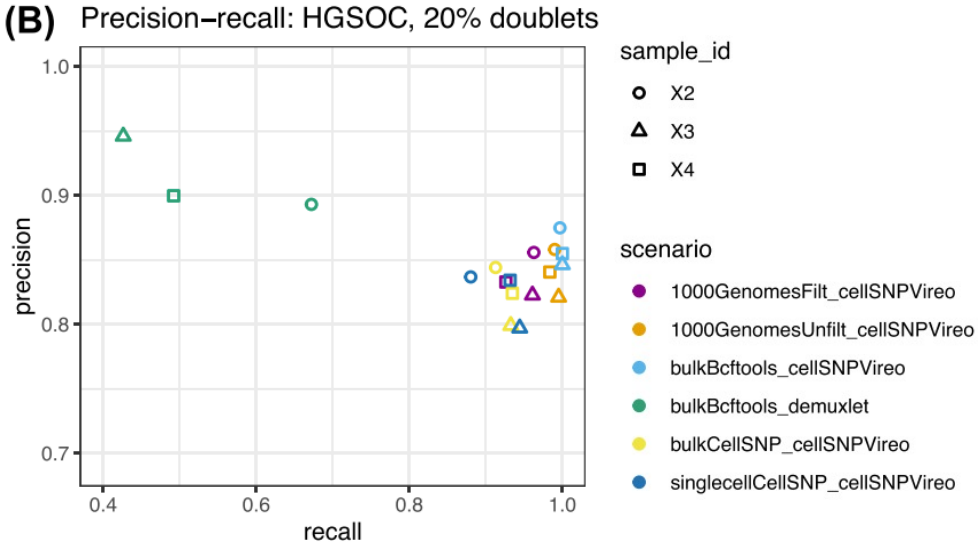
*Correspondence address. Stephanie C. Hicks, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 North Wolfe Street, Baltimore, MD 21205-2179, USA. E-mail: shicks19@jhu.edu http://orcid.org/0000-0002-7858-0231

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

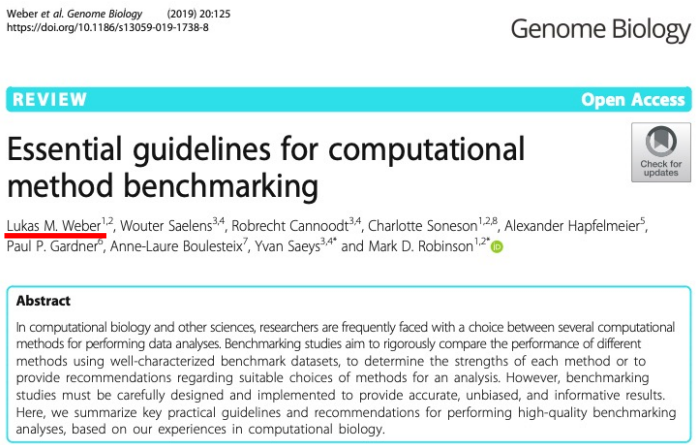
Guidelines / review paper

- 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

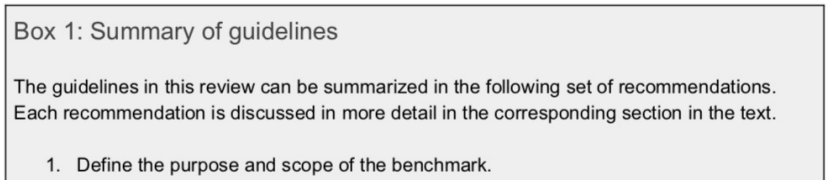
REVIEW Open Access

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 Hapfelmeier[5], Paul P. Gardner[6], Anne-Laure Boulesteix[7], Yvan Saeys[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

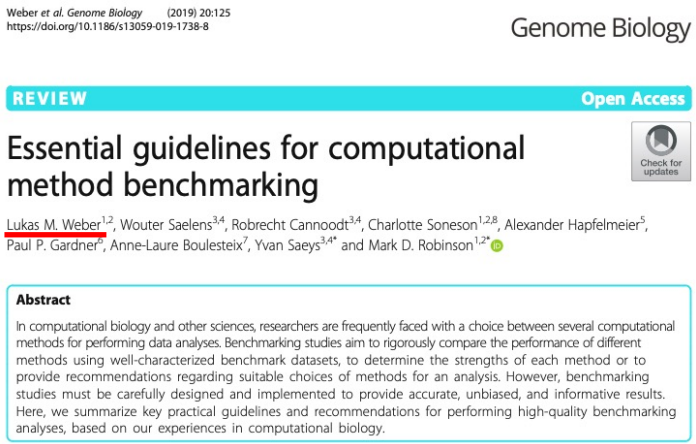
The guidelines in this review can be summarized in the following set of recommendations. Each recommendation is discussed in more detail in the corresponding section in the text.

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

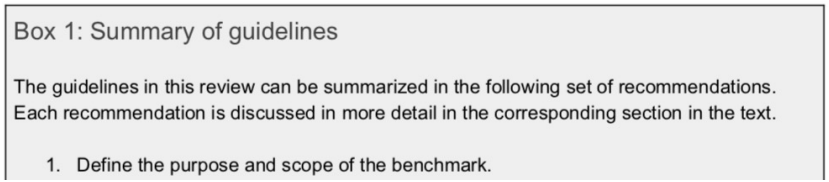
Essential guidelines for computational method benchmarking

Check for updates

Lukas M. Weber[1,2], Wouter Saelens[3,4], Robrecht Cannoodt[3,4], Charlotte Soneson[1,2,8], Alexander Hapfelmeier[5], Paul P. Gardner[6], Anne-Laure Boulesteix[7], Yvan Saeys[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)

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)

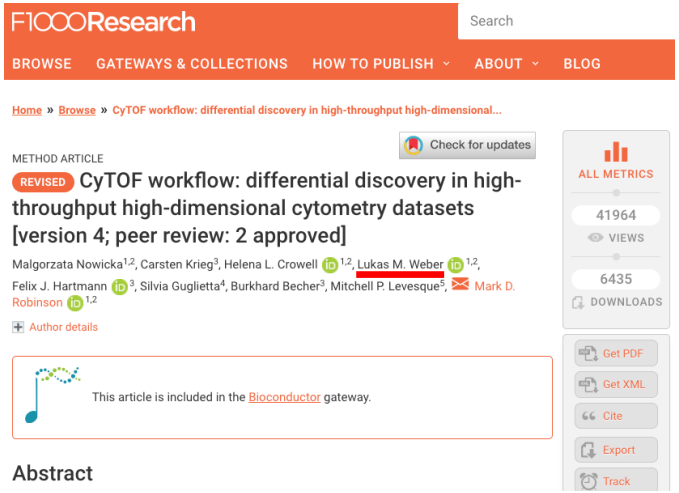
Analysis workflows and tools

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)



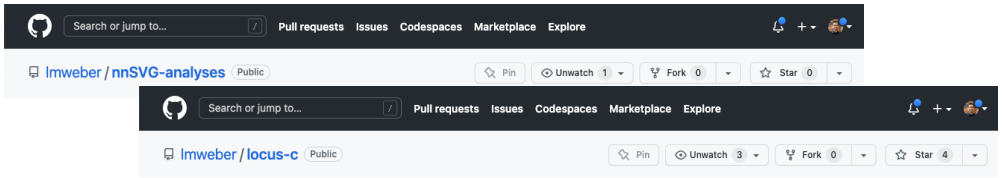
Reproducibility and open science

Freely accessible / open-source software packages

- R/Bioconductor packages (nnSVG, diffcyt)

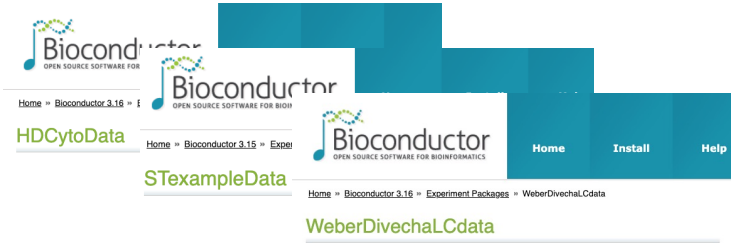
Code availability to reproduce analyses

- GitHub repositories



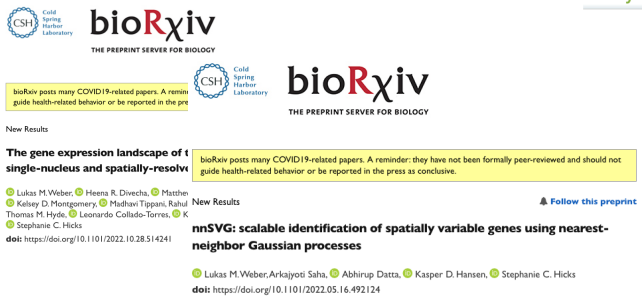
Data availability

- Data packages, managed repositories, web resources



Preprints

- bioRxiv



Reproducibility and open science

Freely accessible / open-source software packages

- R/Bioconductor packages (nnSVG, diffcyt)

Code availability to reproduce analyses

- GitHub repositories

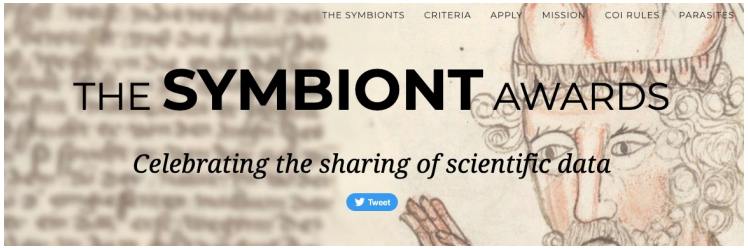
Data availability

- Data packages, managed repositories, web resources

Preprints

- bioRxiv

Research Symbiont Award (2021)



Research Theme 1

Unsupervised statistical methods

1

Spatially-resolved transcriptomics: **nnSVG**
- Weber et al. (2022), *bioRxiv / in revision (Nat Comm)*

High-dimensional cytometry: **diffcyt**
- Weber et al. (2019), *Comms Biol*

Research Theme 2

Collaborative analyses

2

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

Cancer

Immunology

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

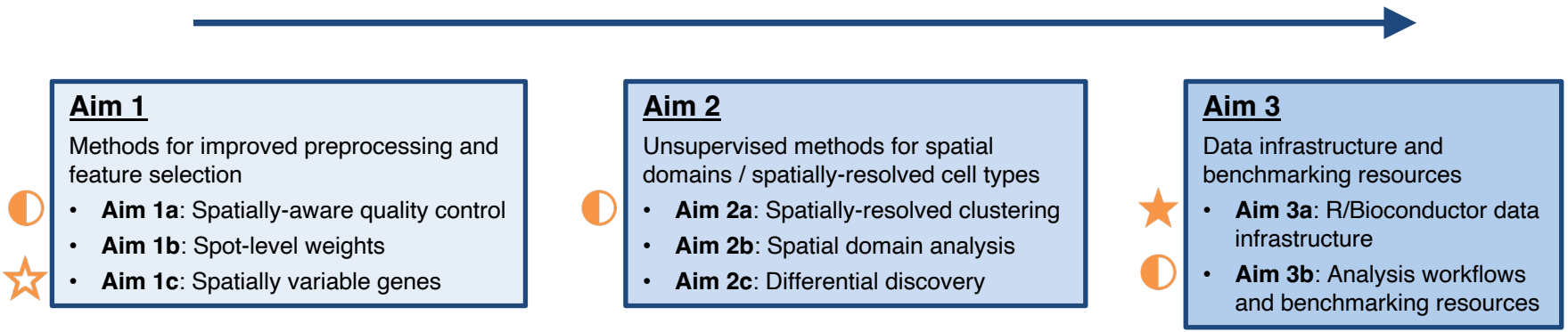
additional papers on website and Google Scholar

Future plans

K99/R00 award

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

Spatially-resolved transcriptomics



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)

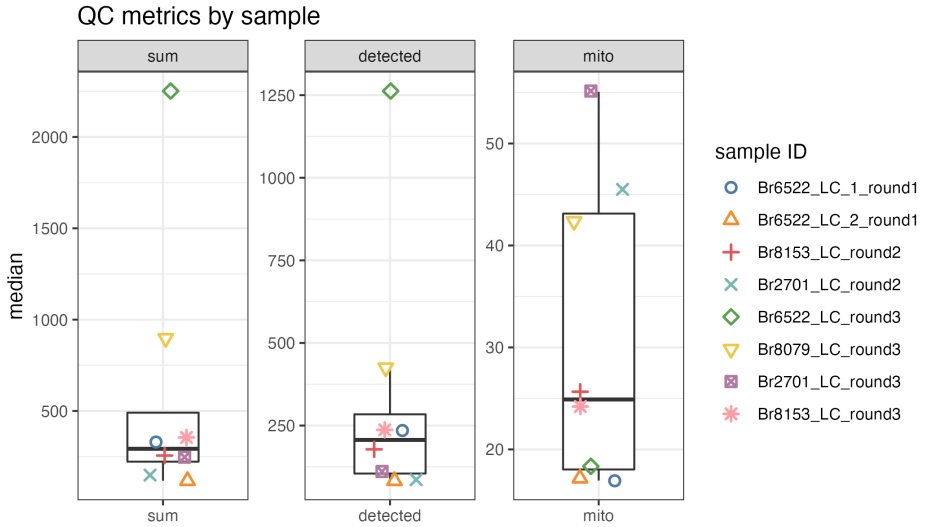
K99/R00 award

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

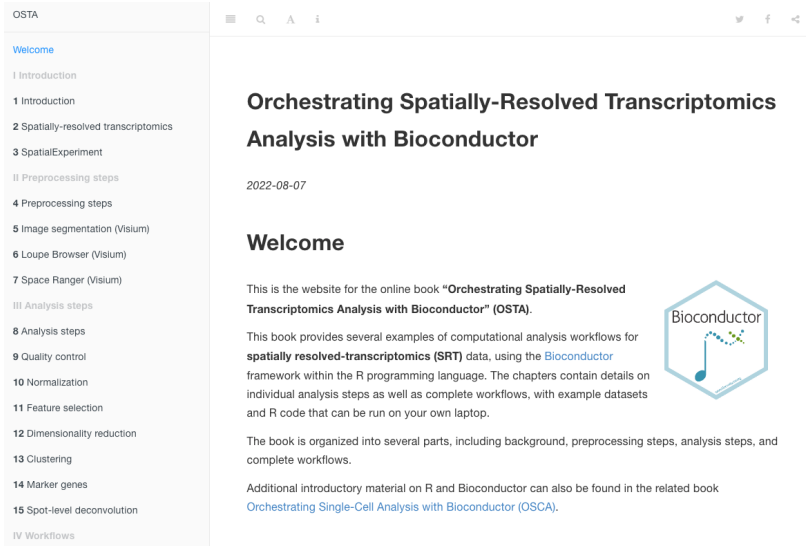
Aim 3b: Analysis workflows for spatially-resolved transcriptomics data

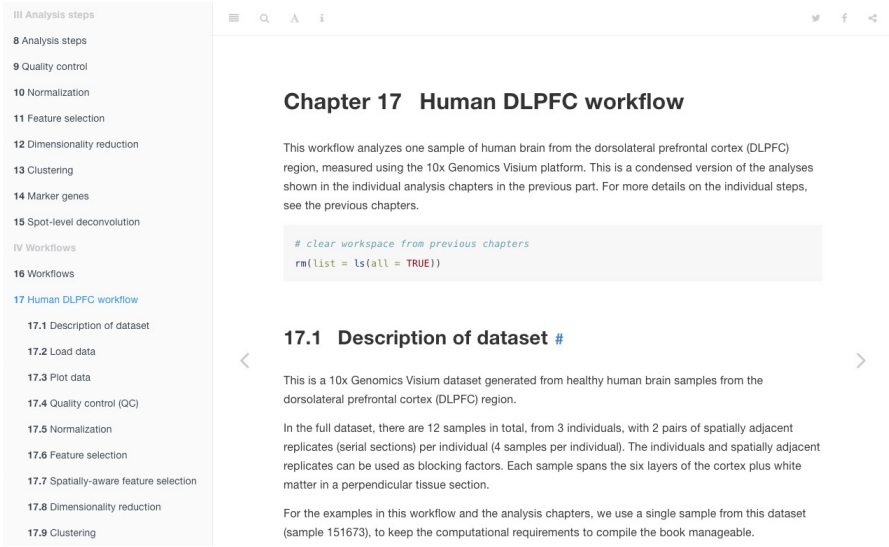
- Orchestrating Spatially-Resolved Transcriptomics Analysis with Bioconductor (OSTA)
- Online book containing example R code and datasets for individual analysis steps and workflows

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





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

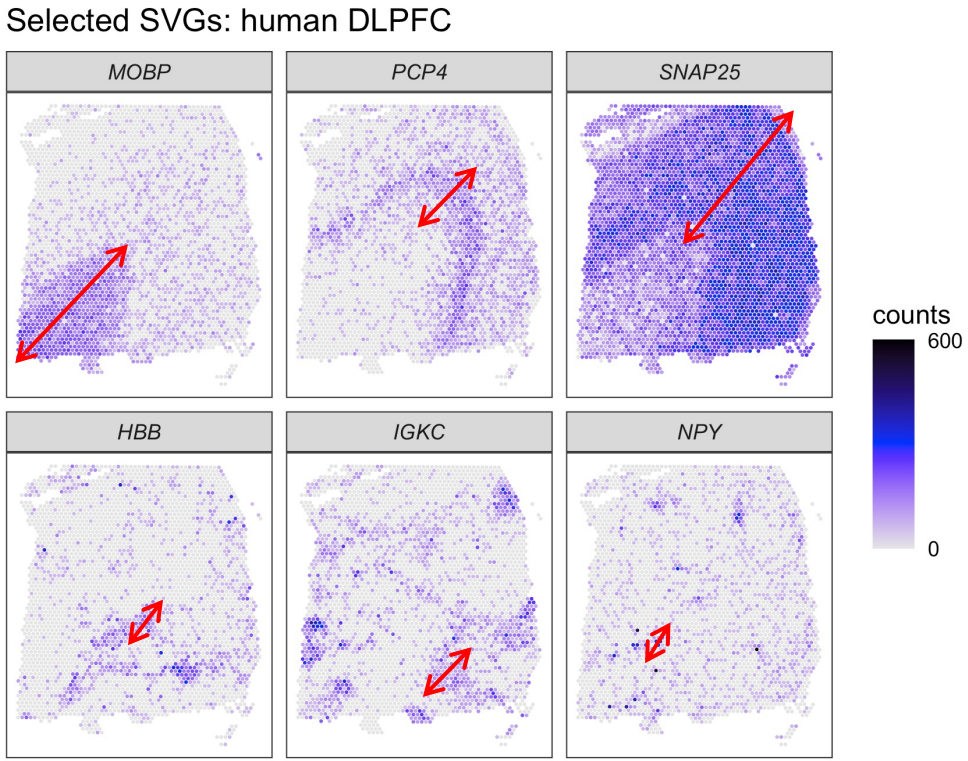
NIH National Human Genome Research Institute

CHAN ZUCKERBERG INITIATIVE

Thank you!

ADDITIONAL SLIDES

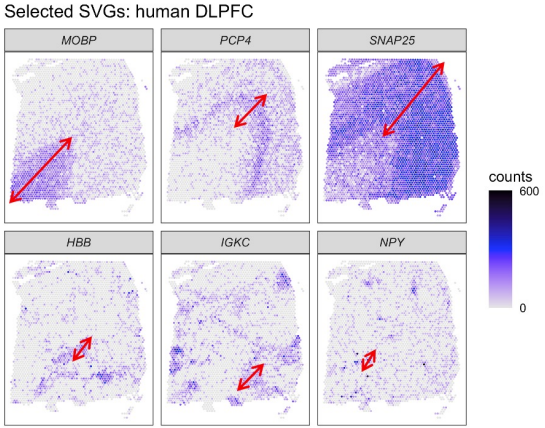
Spatially variable genes



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

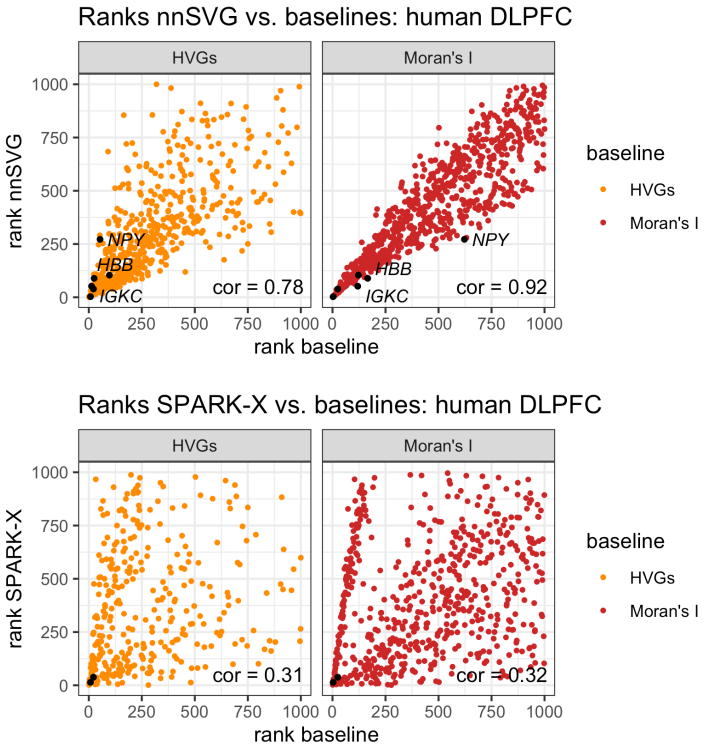
$$I = \frac{N}{W} \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

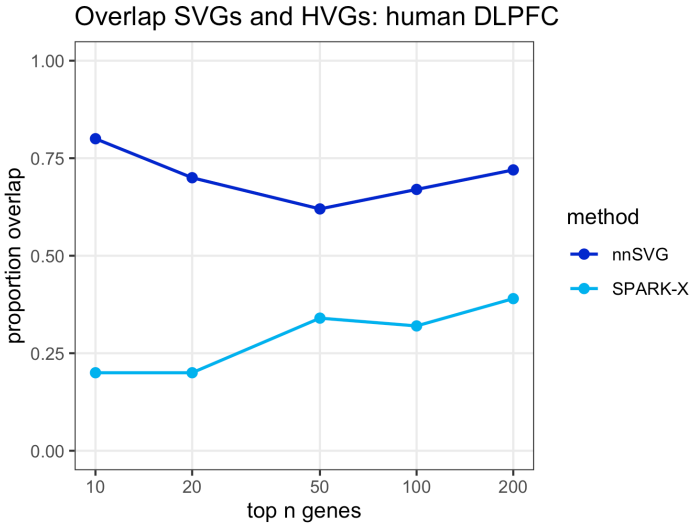
N = number of spatial locations i, j

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

nnSVG evaluations / benchmarking

Human DLPFC dataset

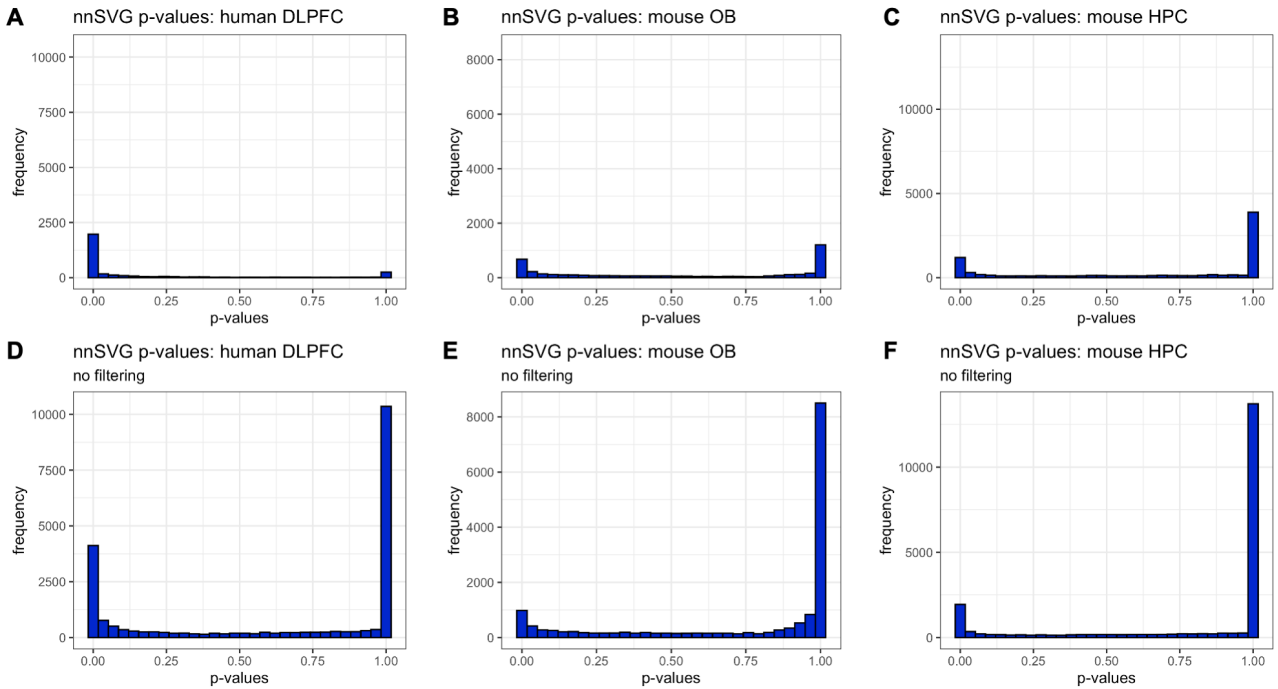




nnSVG evaluations / benchmarking

Human DLPFC dataset

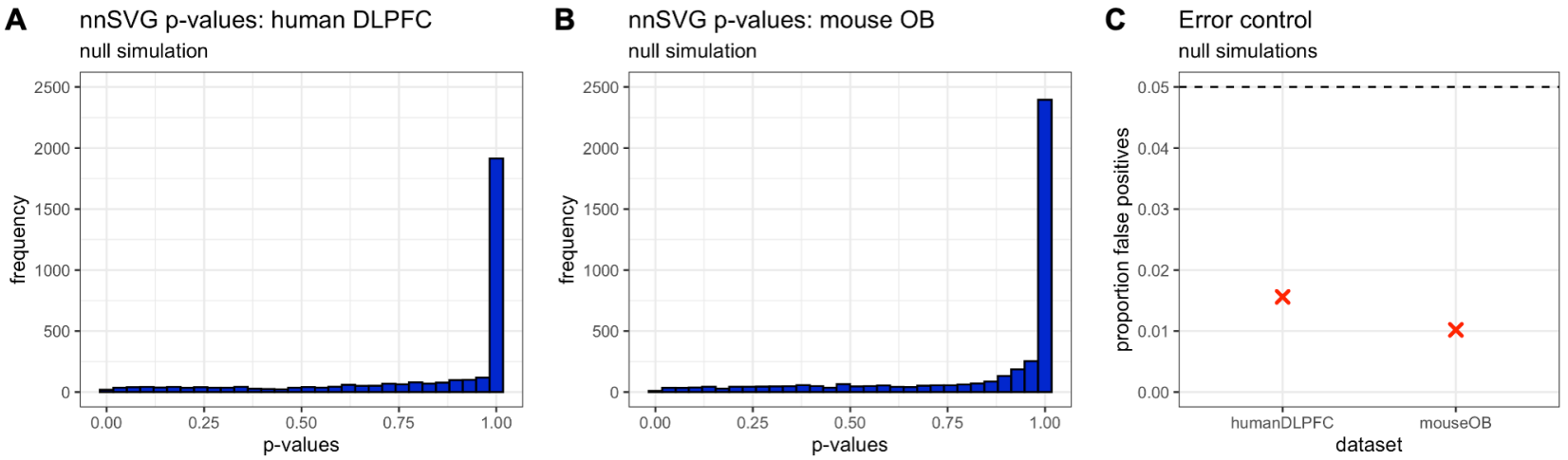
Calibration of p-values



nnSVG evaluations / benchmarking

Human DLPFC dataset

Null simulations

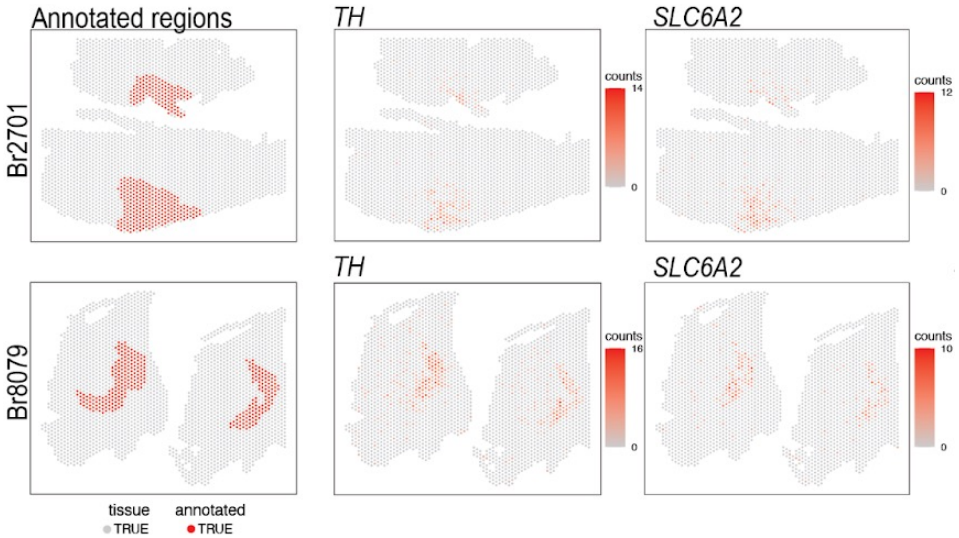


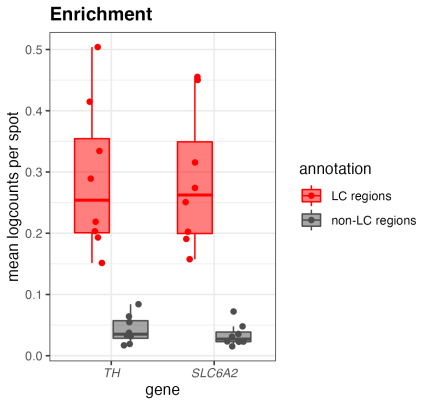
LC: Visium analyses

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





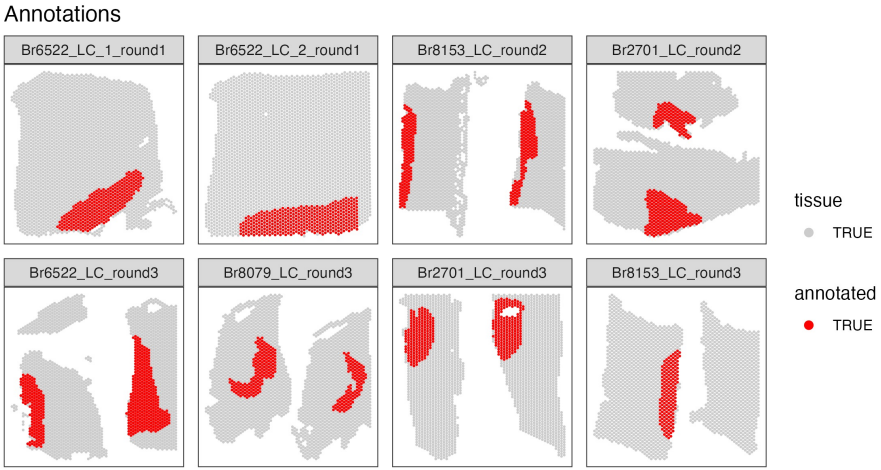
LC: Visium analyses

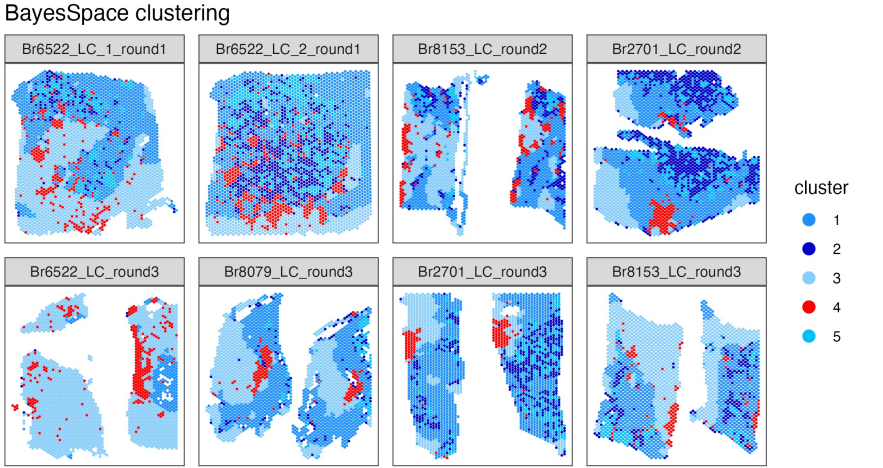
Manual annotations

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

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

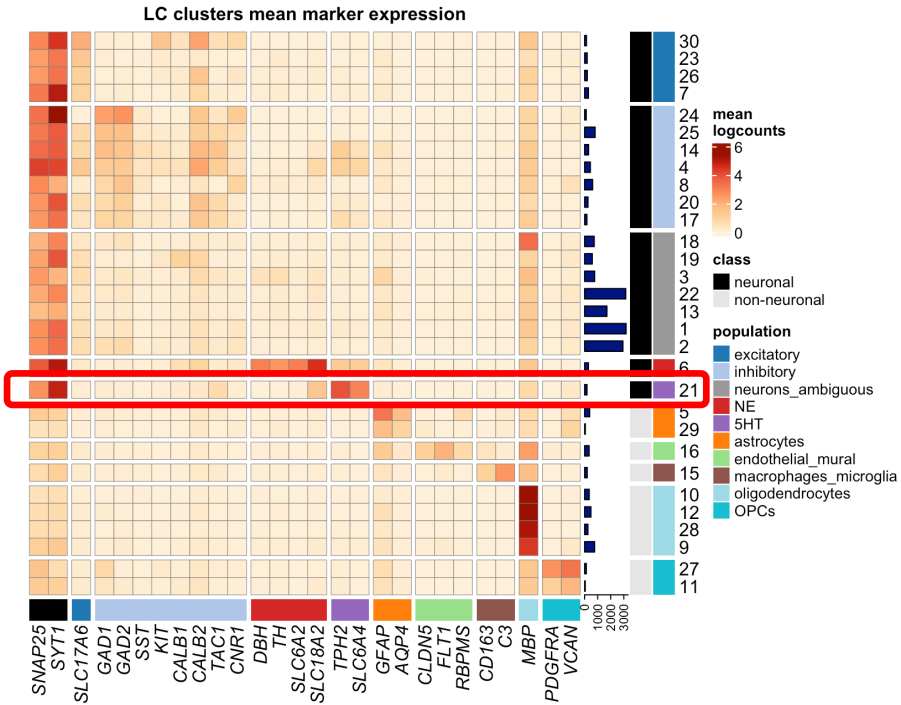




Unsupervised analyses recover additional unexpected results

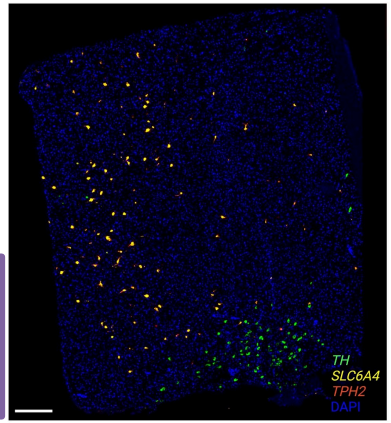
Identification of 5-HT neurons

- 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



scale: 500 µm

Heena Divecha

Lieber Institute for Brain Development

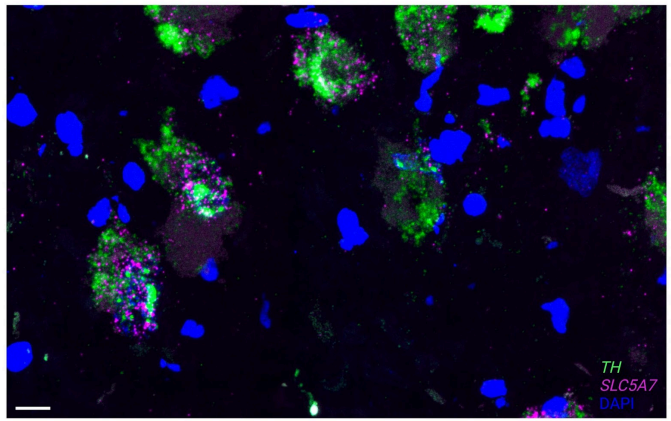
Unsupervised analyses recover additional unexpected results

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

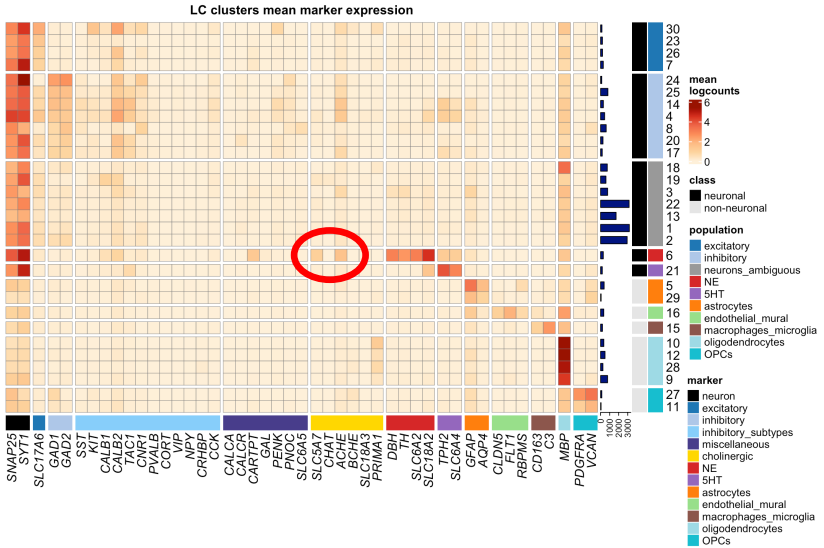
- Unsupervised clustering: transcriptome-wide expression profile of NE neuron cluster

Validation

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



scale: 25 µm

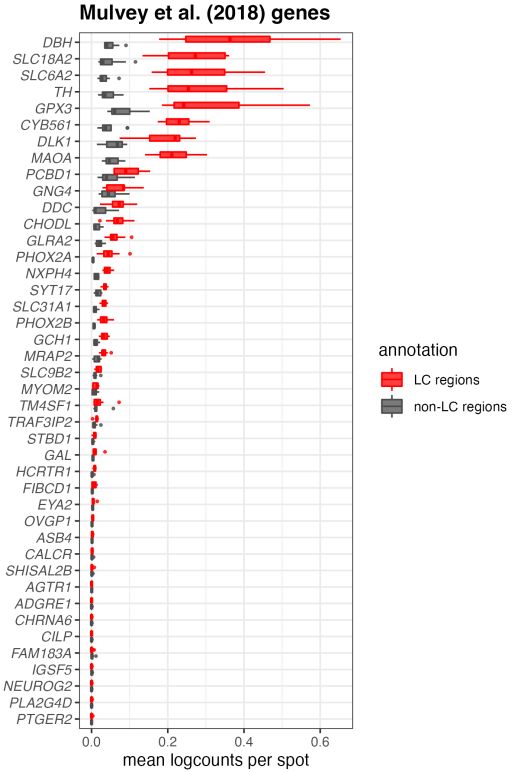


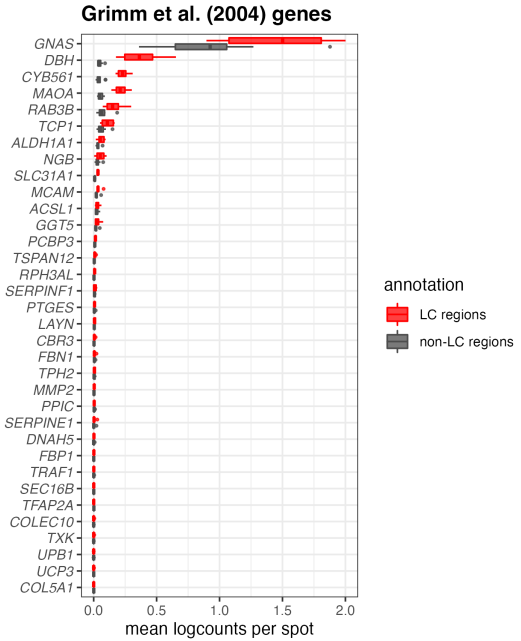
Heena Divecha

Lieber Institute for Brain Development

LC: Visium analyses

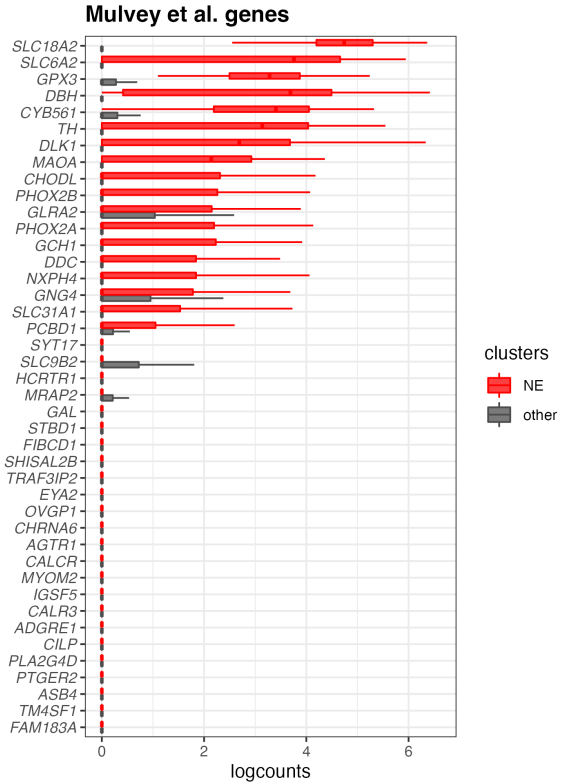
Partial conservation of NE neuron-associated genes across species based on comparison with previous studies in rodents using alternative technological platforms

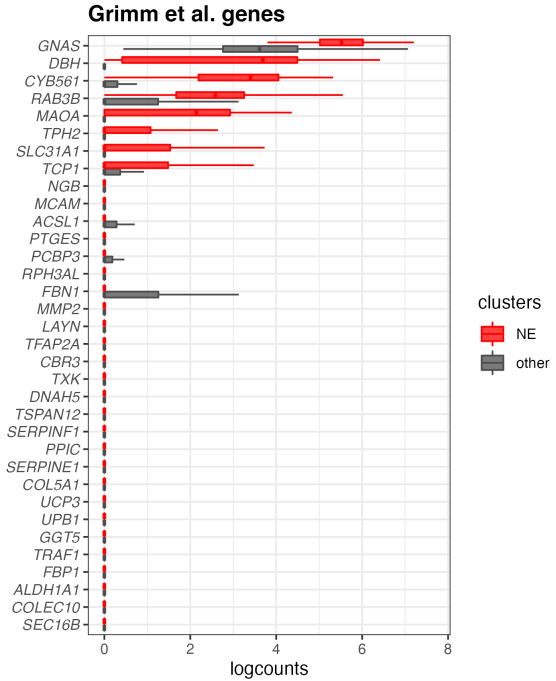




LC: Single-nucleus analyses

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

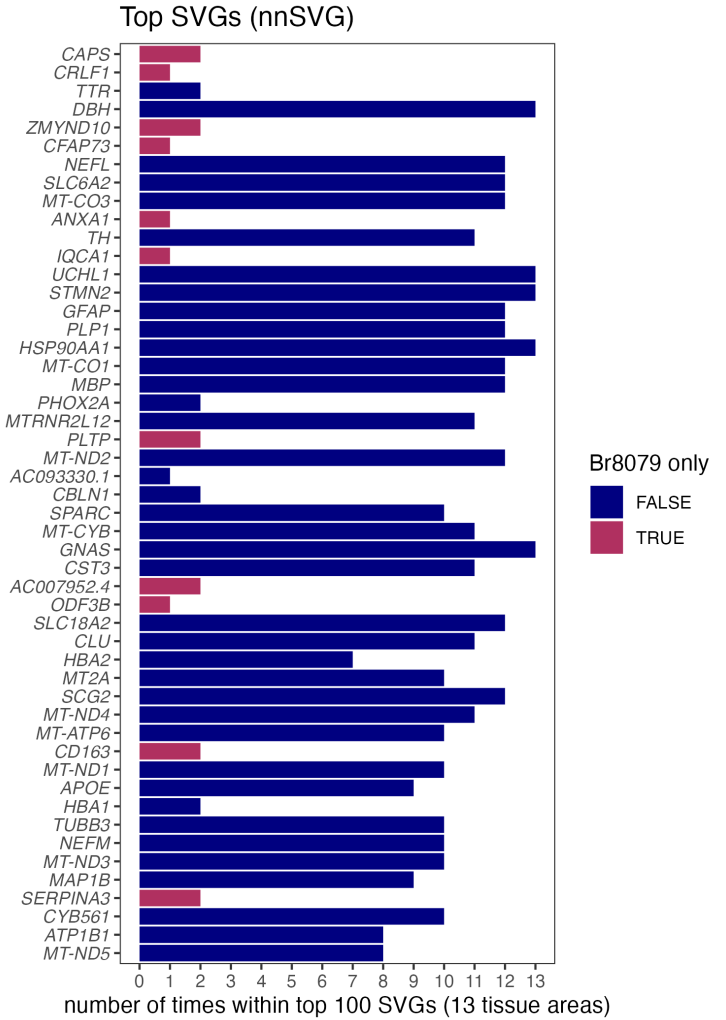




LC: Visium analyses

Visium analyses using nnSVG

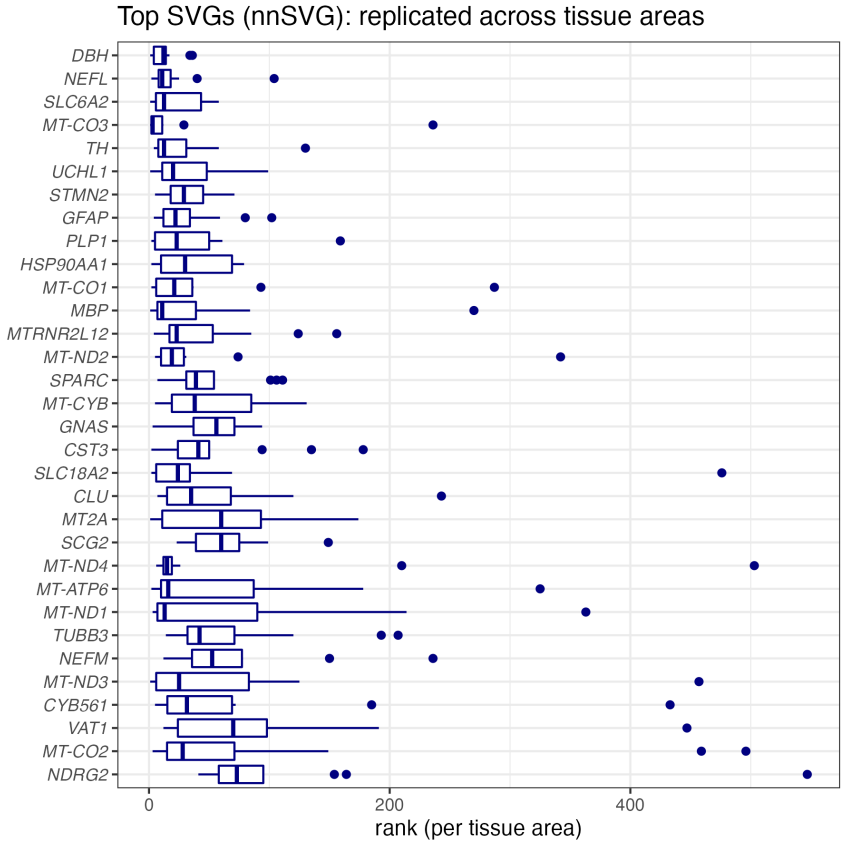
- Identify top LC-associated SVGs replicated across samples
- Number of times each gene is ranked within top 100 SVGs in the 13 tissue areas (Visium samples)
- Donor Br8079: choroid plexus genes



LC: Visium analyses

Visium analyses using nnSVG

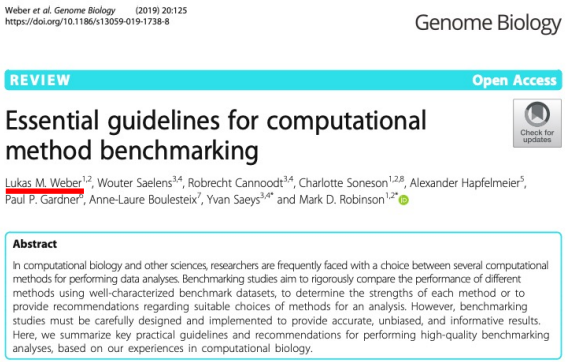
- Excluding choroid plexus genes results in list of top LC-associated SVGs replicated across multiple Visium samples
- Rank per tissue area for genes ranked within top 100 SVGs in at least 10 out of 13 tissue areas



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 Hapfelmeier[5], Paul P. Gardner[6], Anne-Laure Boulesteix[7], Yvan Saeys[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

EDITOR'S CHOICE

Cytometry

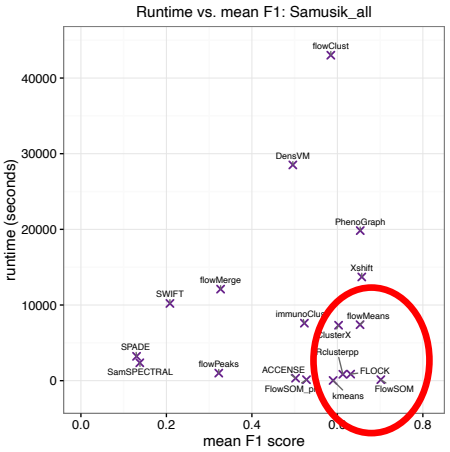
Comparison of Clustering Methods for High-Dimensional Single-Cell Flow and Mass Cytometry Data

Lukas M. Weber,[1,2] Mark D. Robinson[1,2]*

[1]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[2]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland

Received 11 April 2016; Revised 5 September 2016; Accepted 6 November 2016

• Abstract
Recent technological developments in high-dimensional flow cytometry and mass cytometry (CyTOF) have made it possible to detect expression levels of dozens of protein markers in thousands of cells per second, allowing cell populations to be characterized in unprecedented detail. Traditional data analysis by "manual gating" can be inefficient and unreliable in these high-dimensional settings, which has led to the development of a large number of automated analysis methods. Methods designed for unsupervised analysis use specialized clustering algorithms to detect and define cell populations for further downstream analysis. Here, we have performed an up-to-date,



Benchmarking during method development

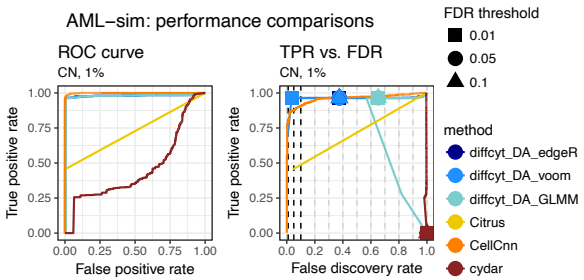
COMMUNICATIONS BIOLOGY

ARTICLE

https://doi.org/10.1038/s42003-019-0415-5 OPEN

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

Lukas M. Weber[1,2], Malgorzata Nowicka[1,2,3], Charlotte Soneson[1,2,4] & Mark D. Robinson[1,2]



bioRχiv

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

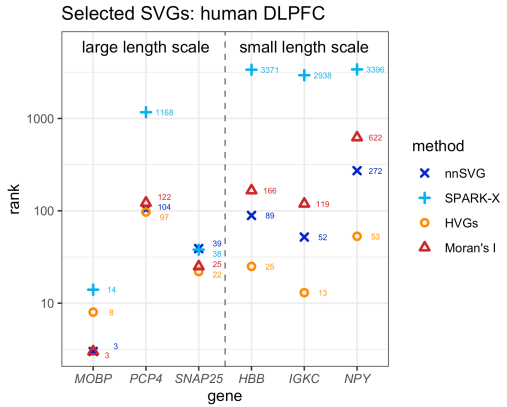
nnSVG: scalable identification of spatially variable genes using nearest-neighbor Gaussian processes

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

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

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



SpatialExperiment

R/Bioconductor class to store spatially-resolved transcriptomics datasets