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

Poster # 139

Lukas M. Weber[1], Arkajyoti Saha[2], Abhirup Datta[1], Kasper D. Hansen[1], Stephanie C. Hicks[1,*]

[1] Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [2] Department of Statistics, University of Washington, Seattle, WA, USA.

* Corresponding author

ABSTRACT

Feature selection to identify spatially variable genes or other biologically informative genes is a key step during analyses of spatially-resolved transcriptomics data. Here, we propose nnSVG, a scalable approach to identify spatially variable genes based on nearest-neighbor Gaussian processes. Our method (i) identifies genes that vary in expression continuously across the entire tissue or within *a priori* defined spatial domains, (ii) uses gene-specific estimates of length scale parameters within the Gaussian process models, and (iii) scales linearly with the number of spatial locations. We demonstrate the performance of our method using experimental data from several technological platforms and simulations. A software implementation is available at https://bioconductor.org/packages/nnSVG.

SPATIALLY-RESOLVED TRANSCRIPTOMICS

- Transcriptome-wide gene expression at spatial resolution
- Example: 10x Genomics Visium platform

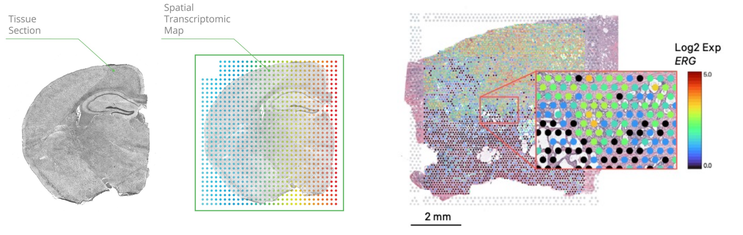
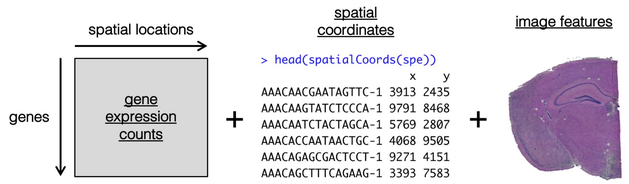


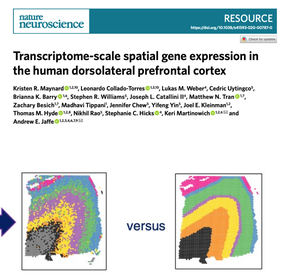
Image source: 10x Genomics

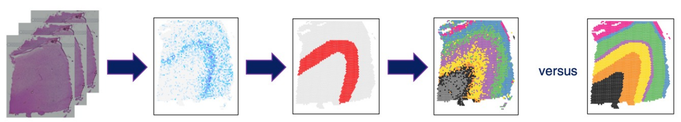
- Data structure



Unsupervised / discovery-based analyses

- Feature selection: spatially variable genes
- Clustering: spatial domains / spatially distributed cell populations
- Differential gene expression



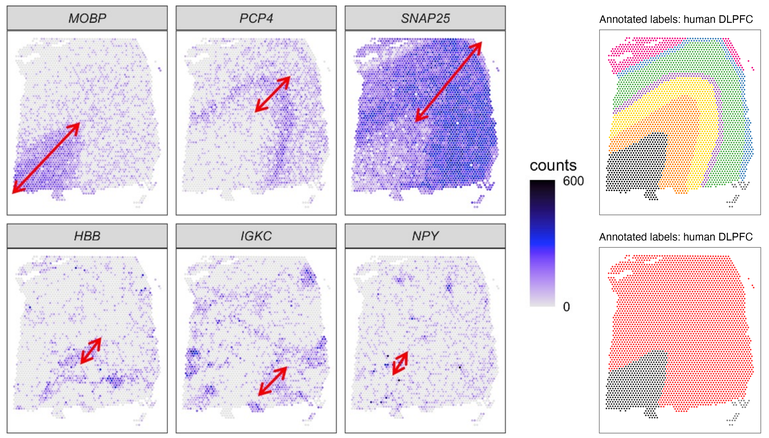


Maynard and Collado-Torres et al. (2021)

SPATIALLY VARIABLE GENES

- Biologically informative genes with spatially defined expression patterns
- Example: dorsolateral prefrontal cortex (DLPFC) in human brain samples (10x Genomics Visium)

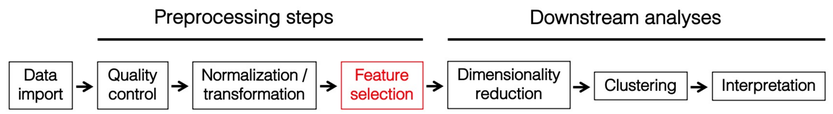
Selected SVGs: human DLPFC



ANALYSIS WORKFLOWS

Spatially variable genes

- Data preprocessing: reduce noise and improve computational performance
- Identify top-ranked genes for further investigation as markers of biological processes



METHODOLOGY

Motivation: existing methods perform poorly or do not scale computationally

Baseline methods: highly variable genes (HVGs), Moran's I statistic

nnSVG methodology

- Nearest-neighbor Gaussian processes (NNGP) (Datta et al. 2016)
- Linear scalability in number of spatial locations
- Exponential covariance function with gene-specific length scale parameter
- Optional covariates for spatial domains
- Fit one model per gene, calculate likelihood ratio (LR) statistic compared to linear model without spatial terms
- Rank genes by LR statistic, identify statistically significant SVGs by approximate LR test
- Effect size: proportion of spatial variance (Svensson et al. 2018)

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

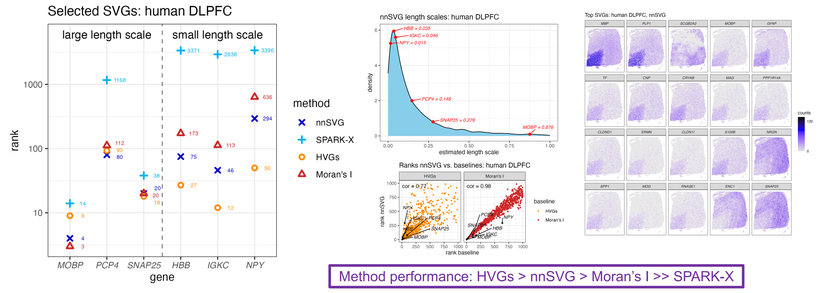
$$C(\theta) = k(s_i, s_j) = \sigma^2 \exp\left(\frac{-||s_i - s_j||}{l}\right)$$

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

EVALUATIONS

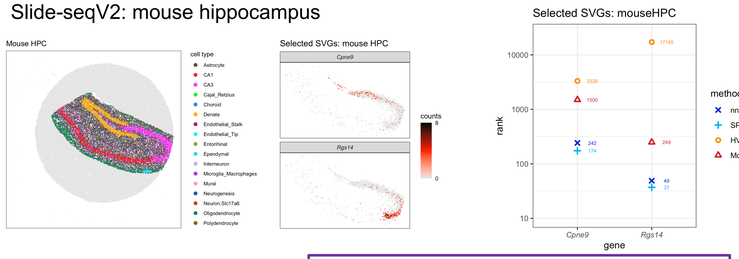
Evaluations on spatially-resolved transcriptomics datasets from several platforms

- 10x Genomics Visium: dorsolateral prefrontal cortex (DLPFC) in human brain



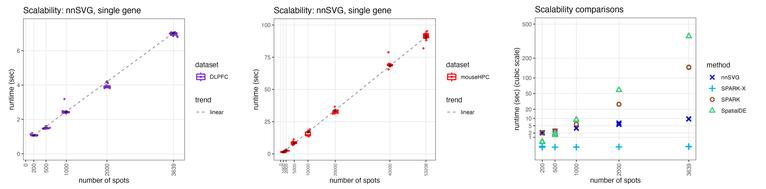
Method performance: HVGs > nnSVG > Moran's I >> SPARK-X

- Slide-seqV2: mouse hippocampus

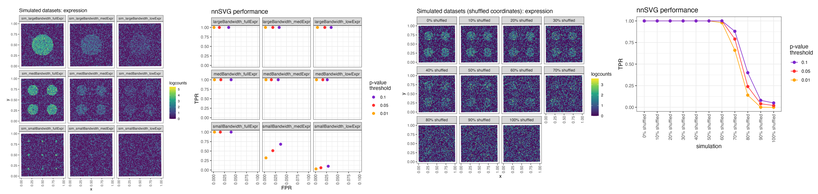


Method performance: SPARK-X ~ nnSVG >> Moran's I >> HVGs

- Computational scalability



- Simulations



- Additional datasets and results in preprint

PREPRINT AND LINKS

Preprint:
https://www.biorxiv.org/content/10.1101/2022.05.16.492124v2

Software package:
https://bioconductor.org/packages/nnSVG
https://github.com/lmweber/nnSVG

Code to reproduce analyses:
https://github.com/lmweber/nnSVG-analyses/

Datasets:
https://bioconductor.org/packages/STexampleData

Contact details:
- lukas.weber@jhu.edu
- https://lmweber.org/
- https://twitter.com/lmwebr