Statistical methods for spatial omics data

- Overview on the technologies (review)
- Finding spatially-variable genes
- Deconvoluting low-resolution (or aggregating high-resolution) spatial omics data
- Spatially-aware dimension reduction / clustering
- Cell-cell communication —> co-localization
- Classical spatial statistics
  - Point patterns: random, clustered, intensity/correlation
  - Lattice data: useful summaries / functions
  - models with spatially correlated errors

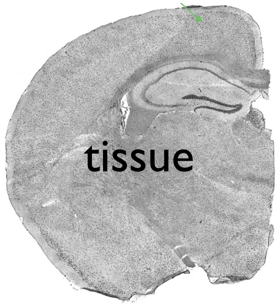
From bulk to single-cell RNA-seq to imaging- & sequencing-based spatially resolved transcriptomics

Background

bulk

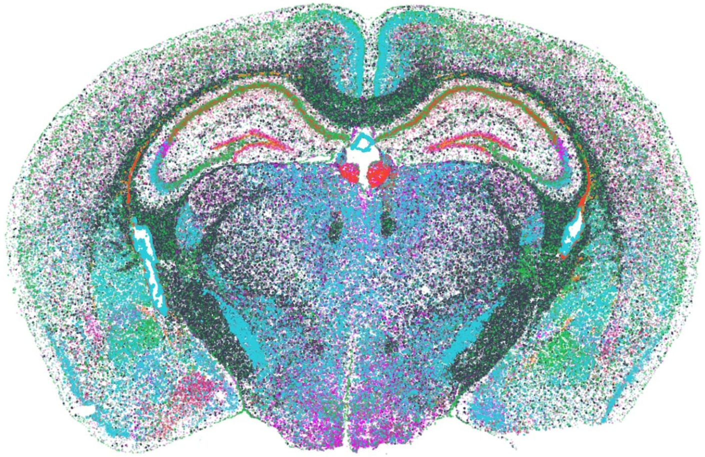
single-cell

spatial



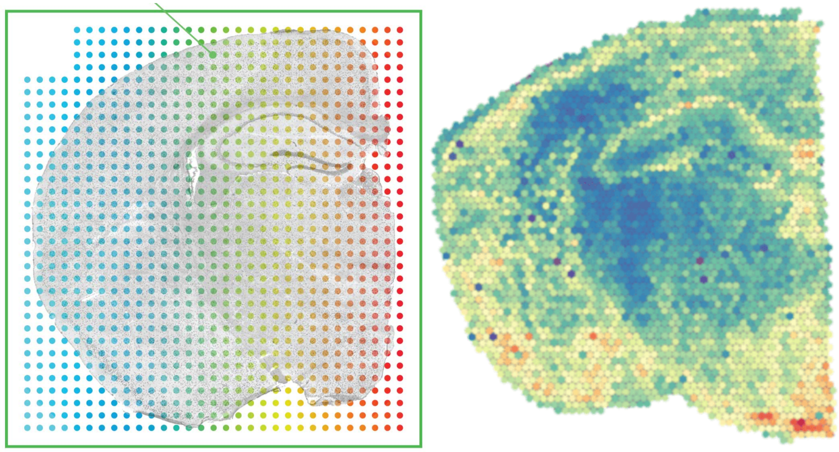
tissue

imaging-based



- molecule-level data
- targeted panel (100s of features; >2024: 1000s)
- single-cell resolution requires segmentation

sequencing-based



- spot-level data
- whole transcriptome (10,000s of features)
- single-cell resolutions requires aggregation or deconvolution

Slide from Helena Crowell

Technology choices: expression table + coordinates

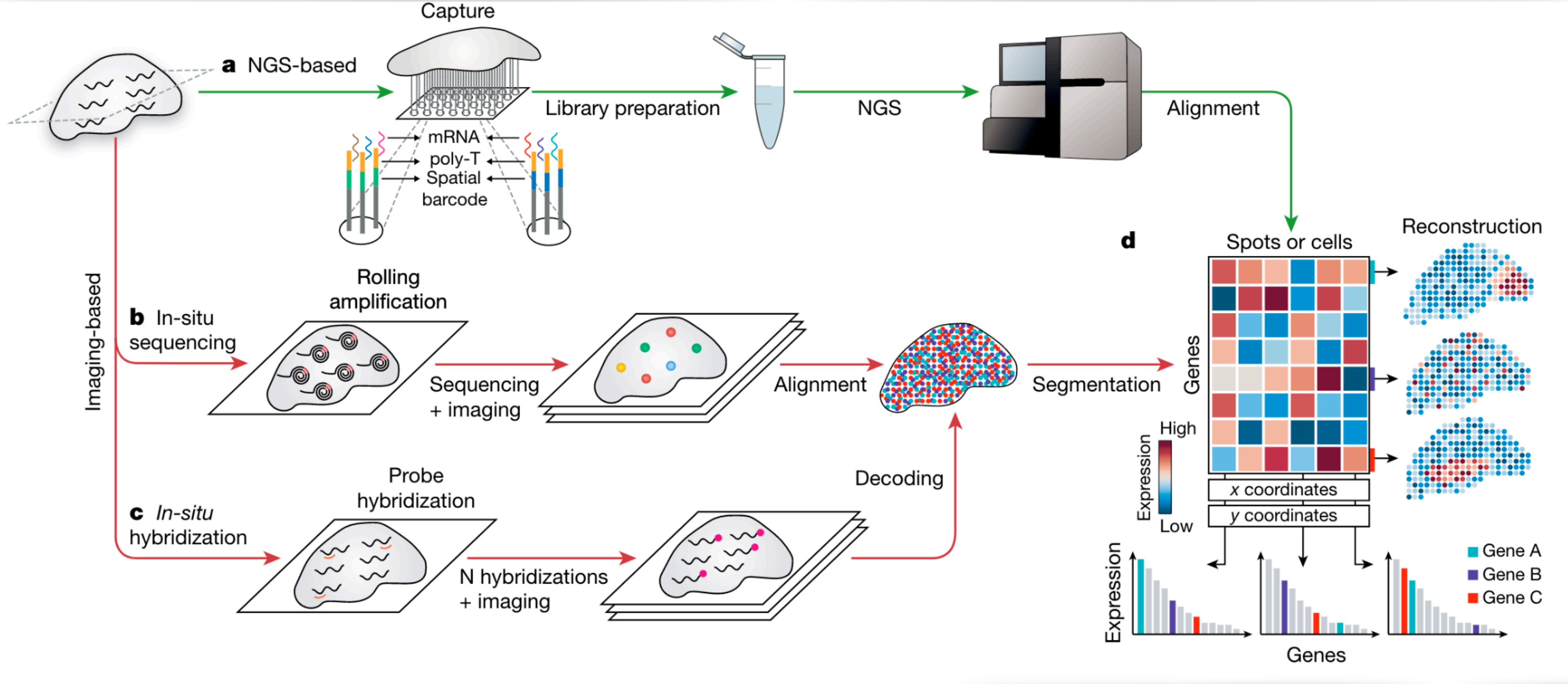
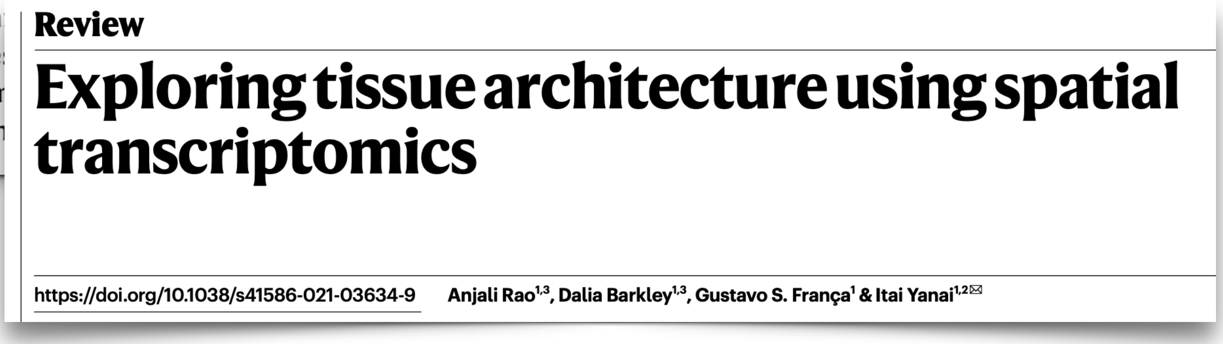


Fig. 1 | The technologies of spatial transcriptomics provide a gene-expression matrix. a, NGS-based spatial transcriptomic methods barcode transcripts according to their location in a lattice of spots. **b**, ISS approaches directly read out the transcript sequence within the tissue. **c**, ISH

methods detect ta
fluorescent probe
gene-expression m
genes and location

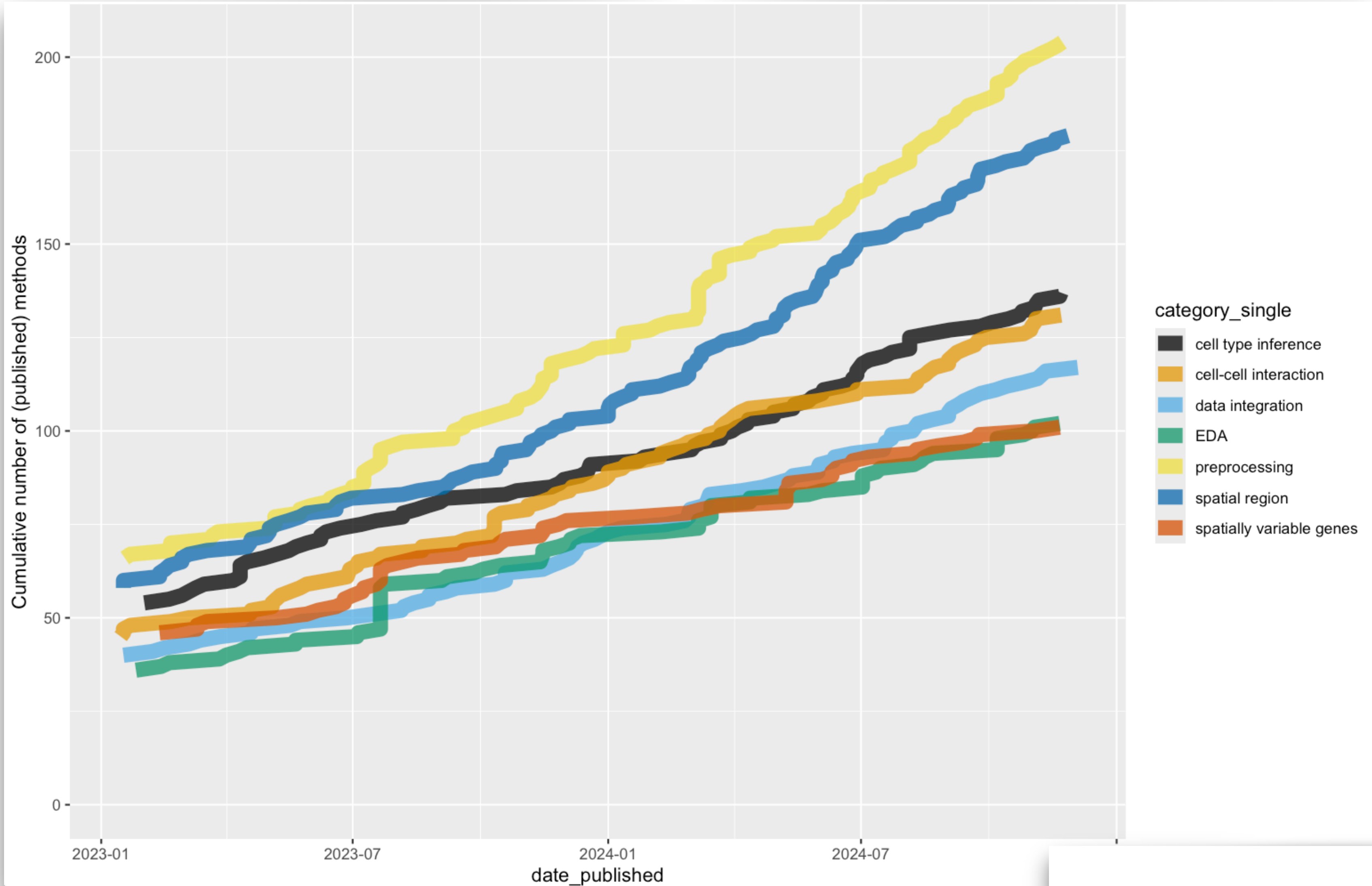


Review

Exploring tissue architecture using spatial transcriptomics

https://doi.org/10.1038/s41586-021-03634-9 **Anjali Rao[1,3], Dalia Barkley[1,3], Gustavo S. França[1] & Itai Yanai[1,2]✉**

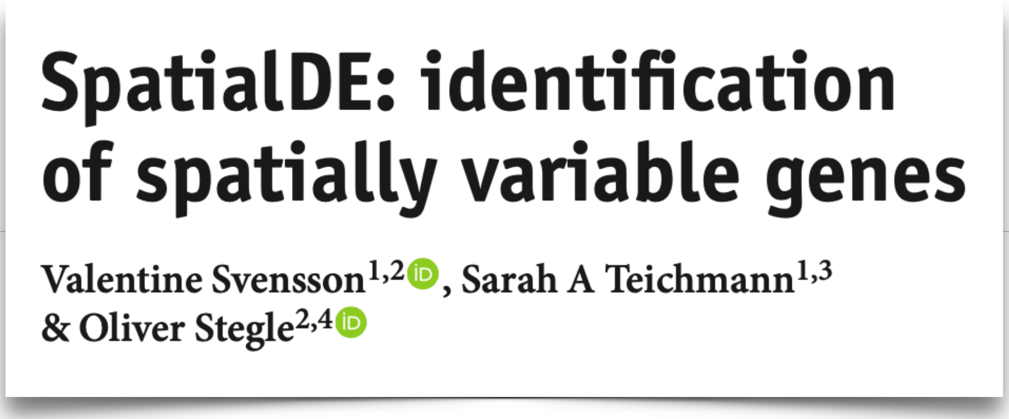
(Spatial omics) computational method explosion



Museum of spatial transcriptomics

Lambda Moses[1] and Lior Pachter[1,2] ✉

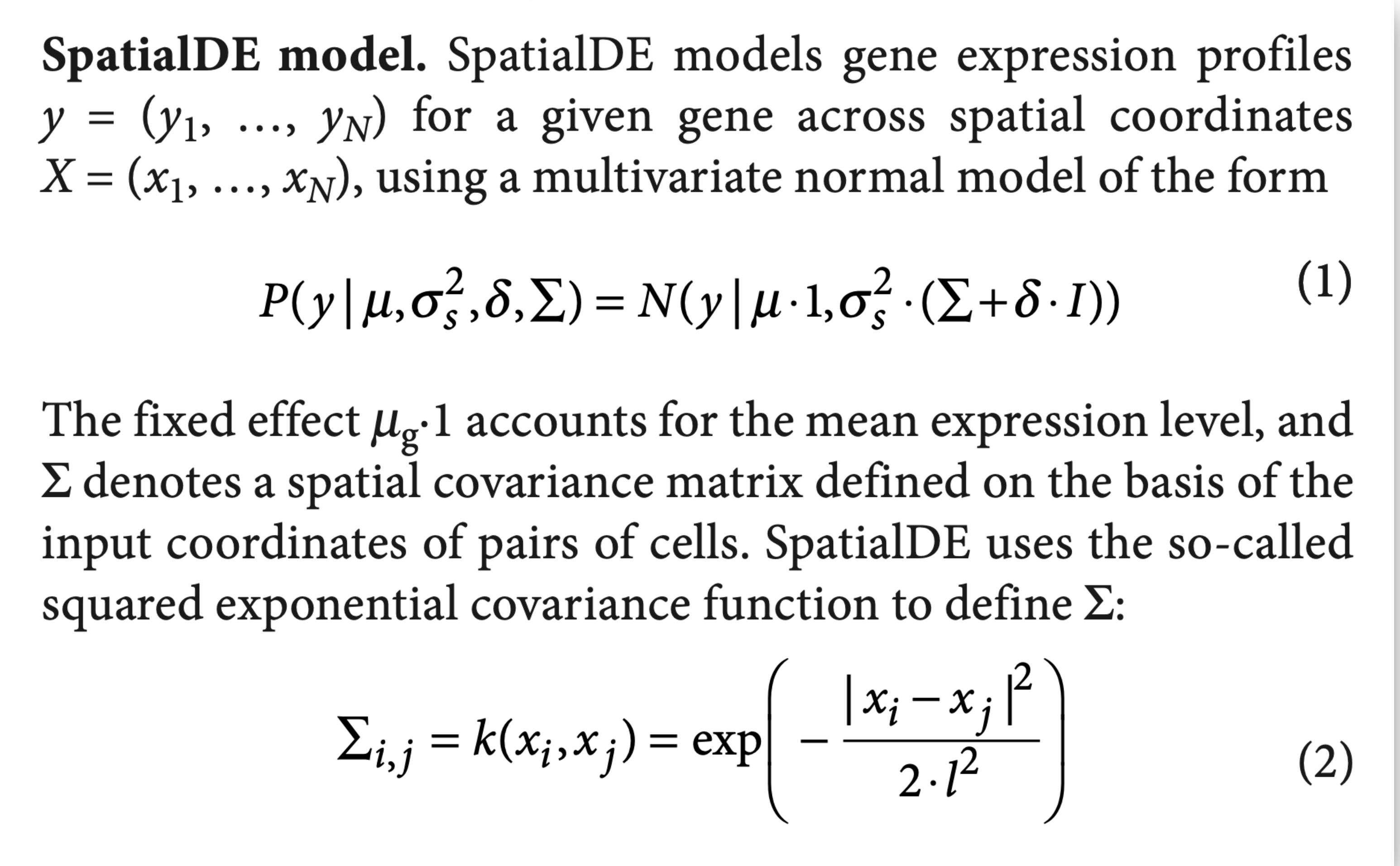
Finding spatially-variable genes: SpatialDE



SpatialDE: identification of spatially variable genes

Valentine Svensson[1,2], Sarah A Teichmann[1,3] & Oliver Stegle[2,4]

- SpatialDE: response = normal distribution with covariance with two components: i) based on distance b/w points - exponential decay; ii) constant non-spatial variance
- Null model: fit just the non-spatial variance (i.e., without sigma)
- Fit 2 models, likelihood ratio test



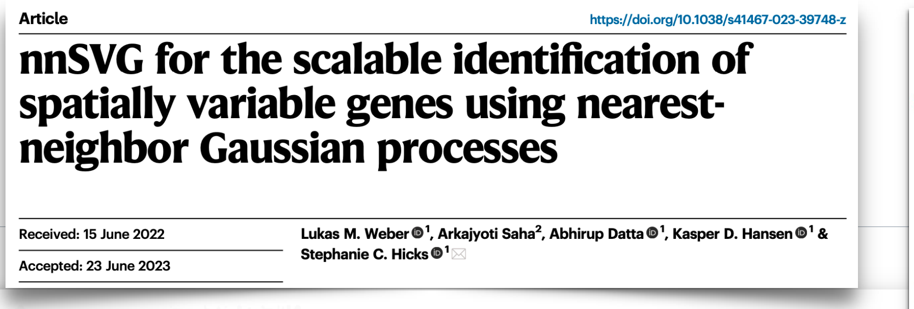
SpatialDE model. SpatialDE models gene expression profiles $y = (y_1, \ldots, y_N)$ for a given gene across spatial coordinates $X = (x_1, \ldots, x_N)$, using a multivariate normal model of the form

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

The fixed effect $\mu_g \cdot 1$ accounts for the mean expression level, and Σ denotes a spatial covariance matrix defined on the basis of the input coordinates of pairs of cells. SpatialDE uses the so-called squared exponential covariance function to define Σ:

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

Spatially variable genes



Article

https://doi.org/10.1038/s41467-023-39748-z

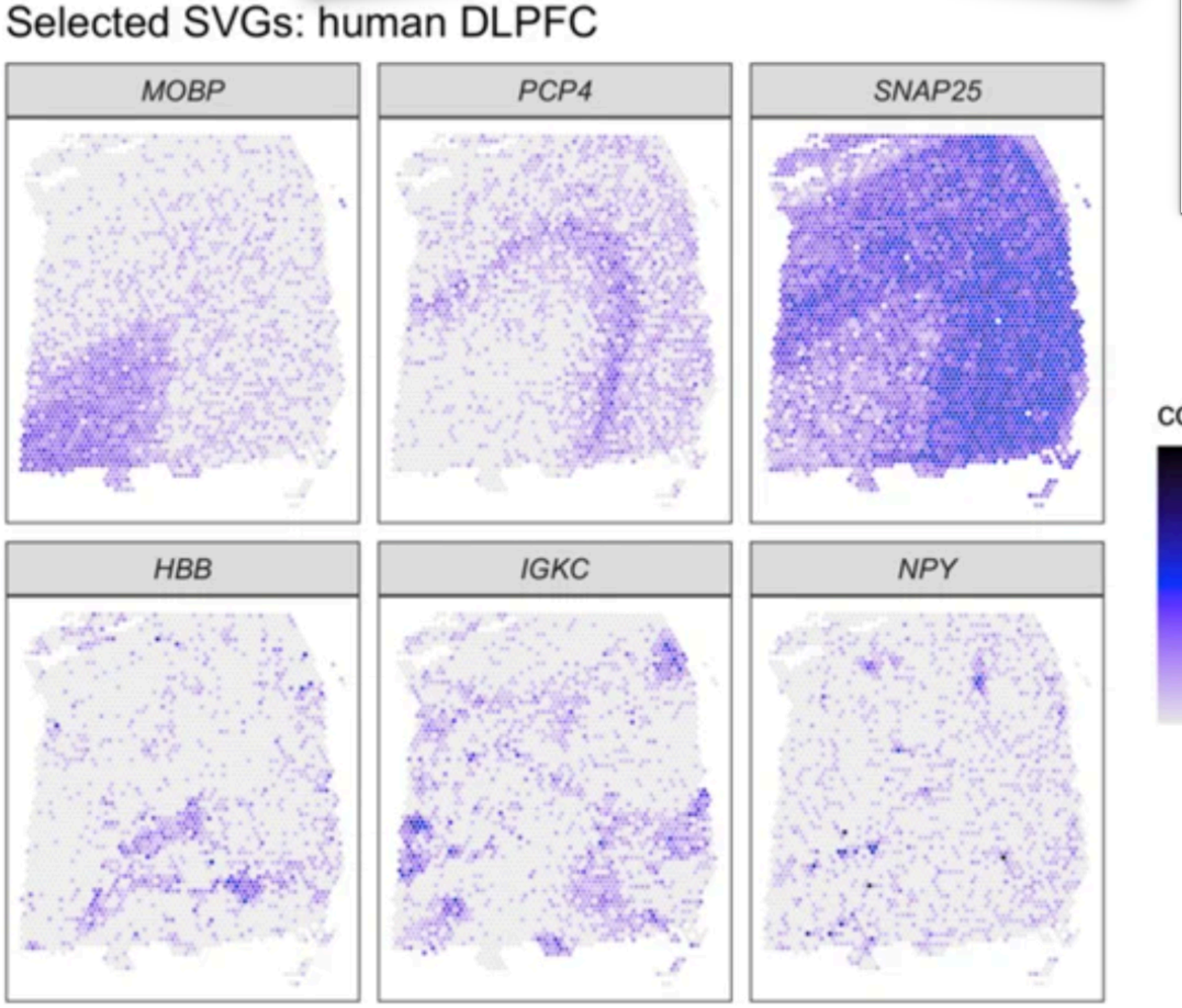
nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes

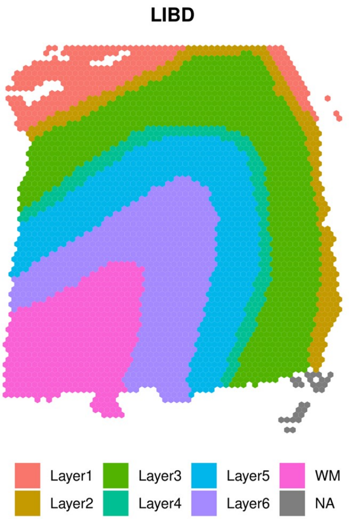
Received: 15 June 2022

Accepted: 23 June 2023

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

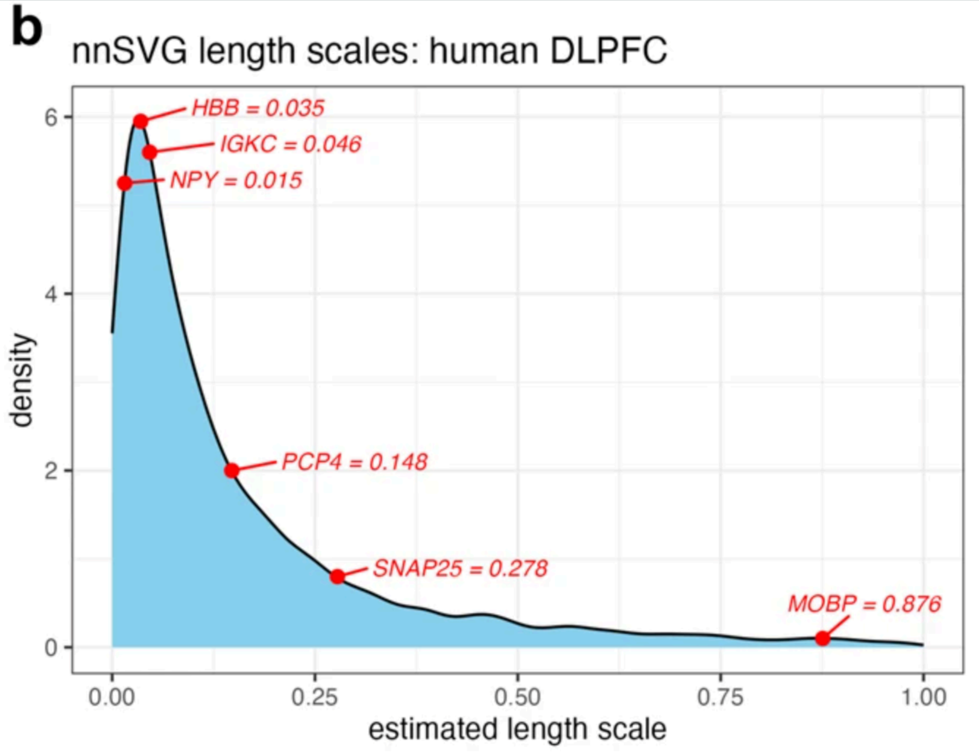
- different types (senses?) of spatially variable genes





Spatially variable genes

$$C_{ij}(\boldsymbol{\theta}) = \sigma^2 \exp\left(\frac{-||\mathbf{s_i} - \mathbf{s_j}||}{l}\right)$$



Article

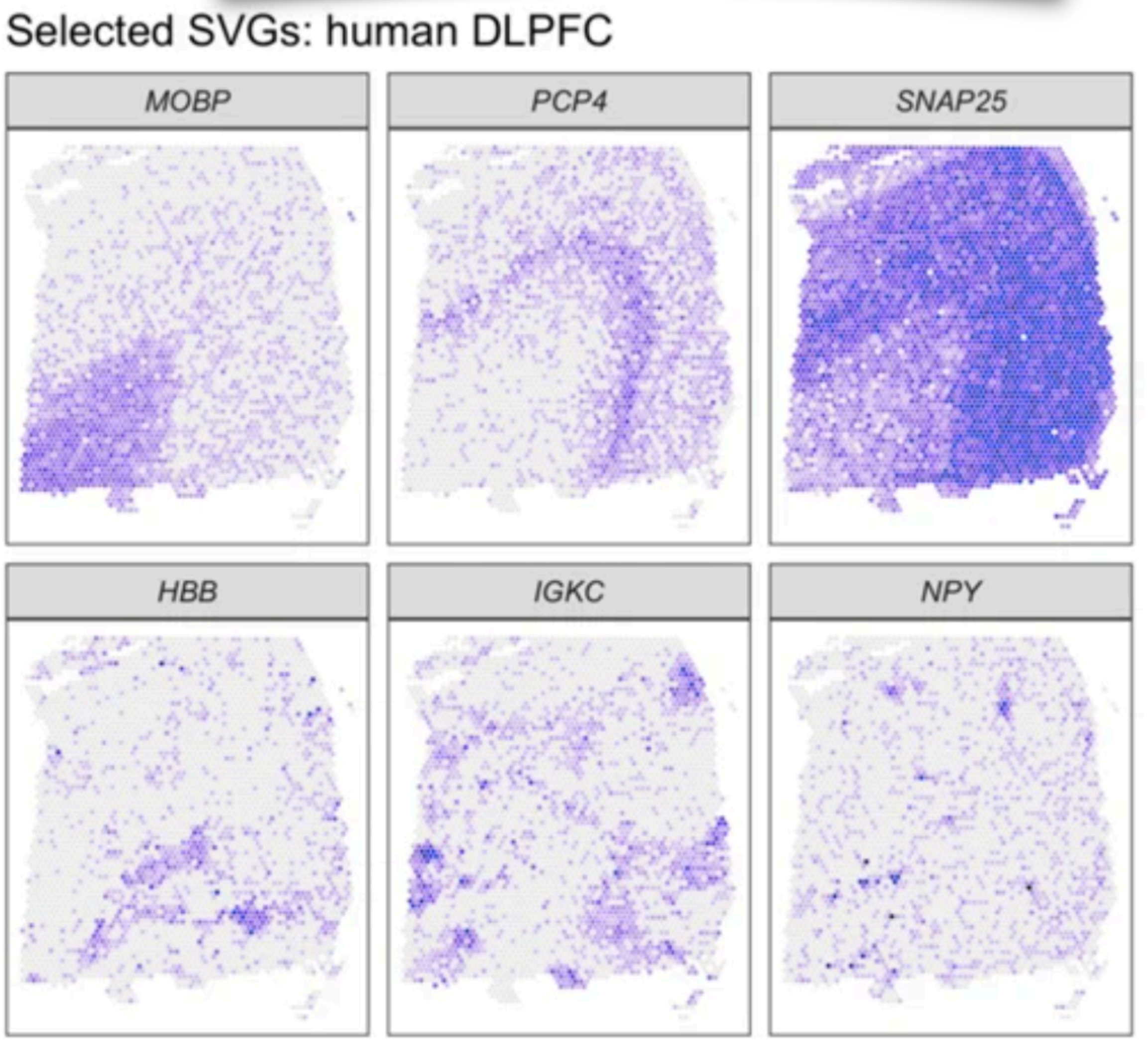
https://doi.org/10.1038/s41467-023-39748-z

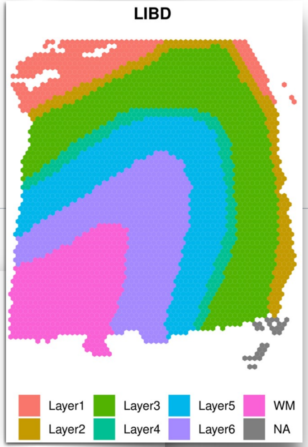
nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes

Received: 15 June 2022

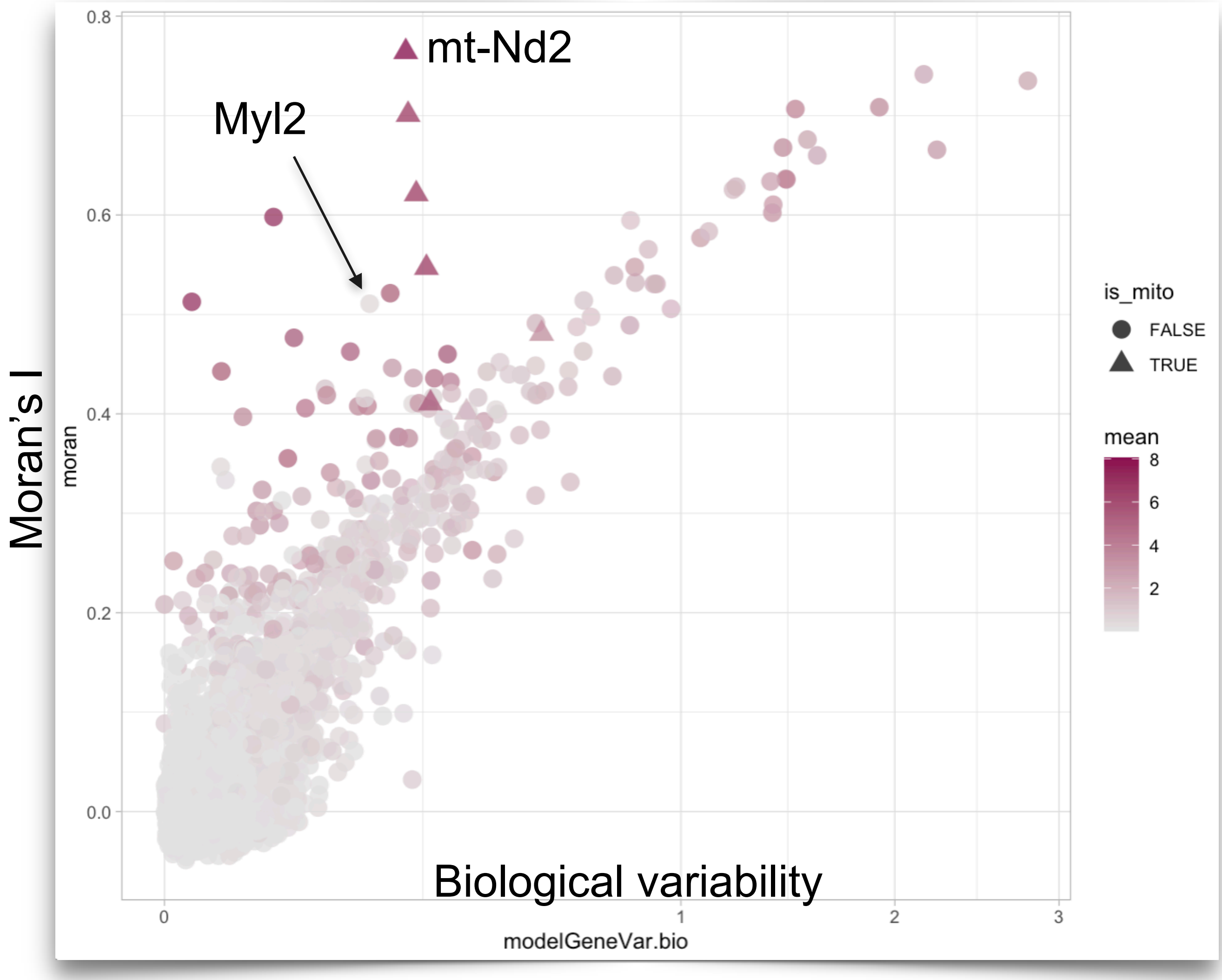
Accepted: 23 June 2023

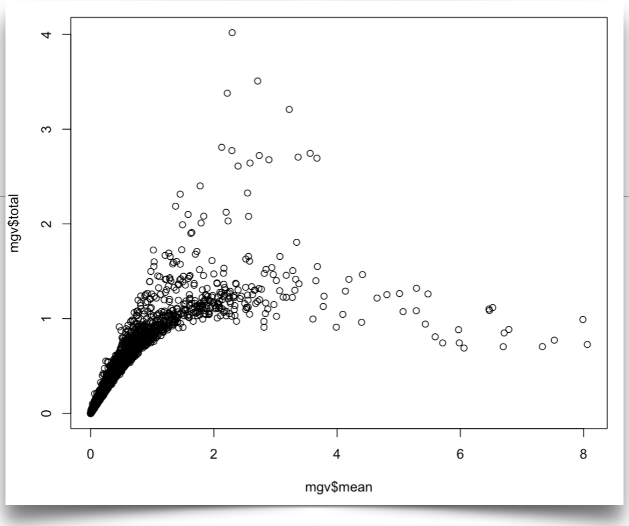
Lukas M. Weber[1], Arkajyoti Saha[2], Abhirup Datta[1], Kasper D. Hansen[1] & Stephanie C. Hicks[1]

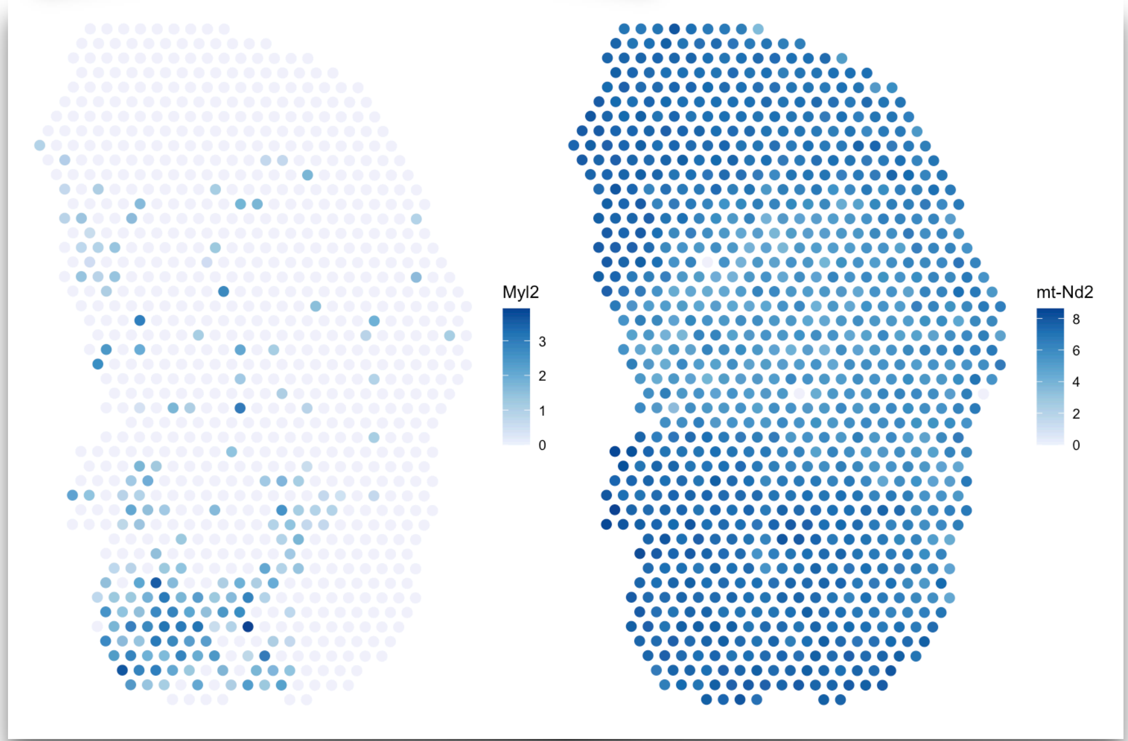




Spatially variable versus highly variable







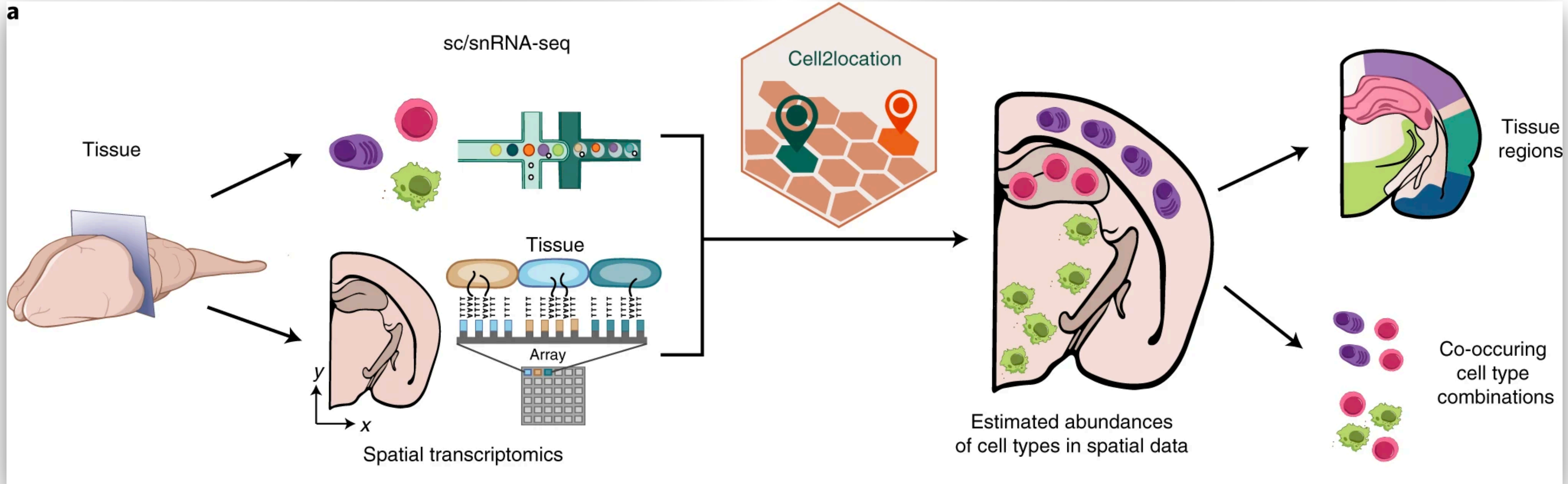
(More mathematical details on Moran's I below)

Statistical methods for spatial omics data

- Overview on the technologies (review)
- Finding spatially-variable genes
- Deconvoluting low-resolution (or aggregating high-resolution) spatial omics data
- Spatially-aware dimension reduction / clustering
- Cell-cell communication —> co-localization
- Classical spatial statistics
  - Point patterns: random, clustered, intensity/correlation
  - Lattice data: useful summaries / functions
  - models with spatially correlated errors

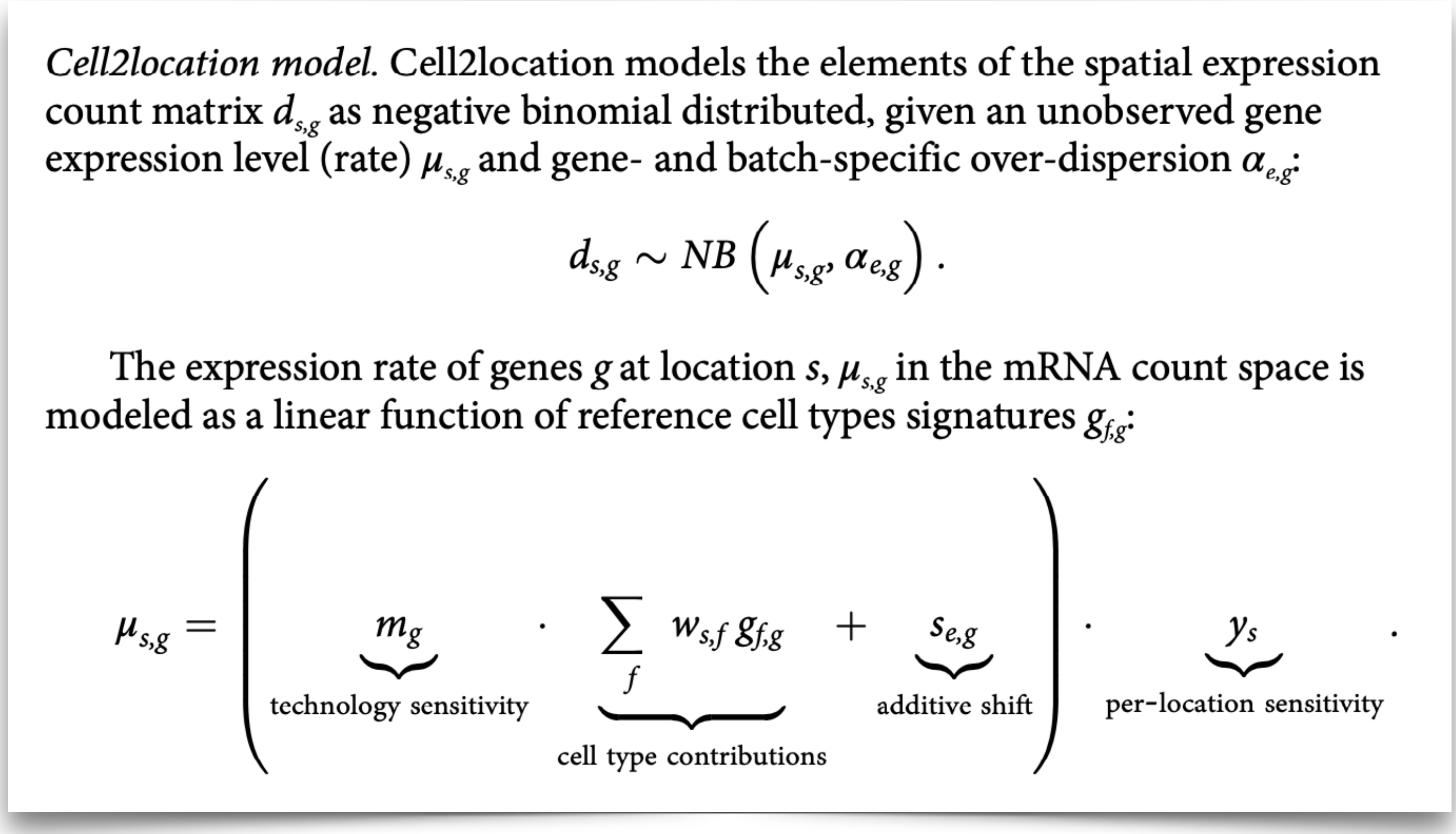
Deconvoluting low-resolution spatial omics (sequencing) data

- Cell2location: negative binomial regression for reference cell type signatures; decompose spot-level mRNA counts into reference cell types


a
sc/snRNA-seq
Cell2location
Tissue
Tissue
Array
y
x
Spatial transcriptomics
Estimated abundances of cell types in spatial data
Tissue regions
Co-occuring cell type combinations

Deconvoluting low-resolution spatial omics data

- Cell2location: negative binomial regression for reference cell type signatures; decompose spot-level mRNA counts into reference cell types



Cell2location model. Cell2location models the elements of the spatial expression count matrix $d_{s,g}$ as negative binomial distributed, given an unobserved gene expression level (rate) $\mu_{s,g}$ and gene- and batch-specific over-dispersion $\alpha_{e,g}$:

$$d_{s,g} \sim NB\left(\mu_{s,g}, \alpha_{e,g}\right) .$$

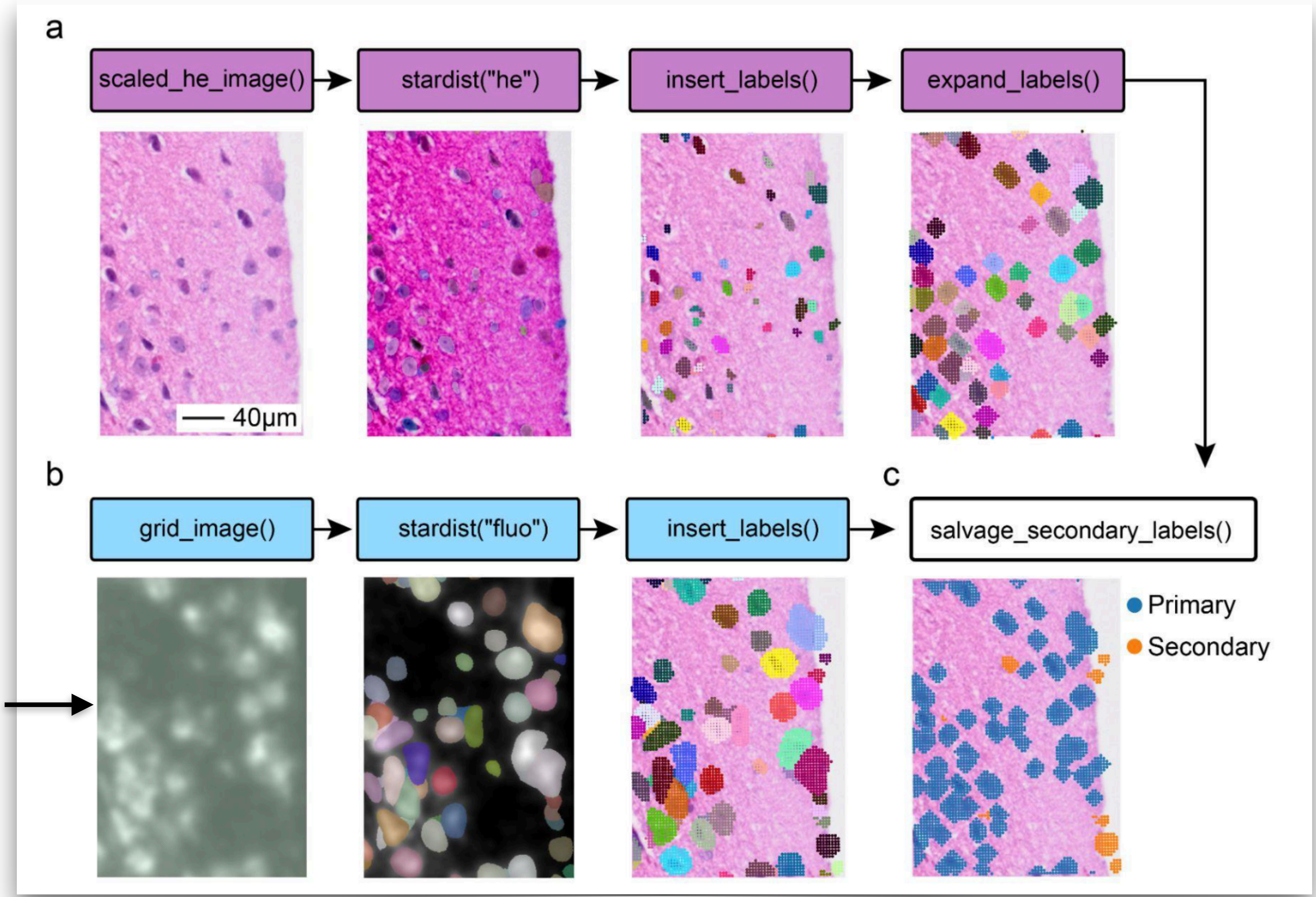
The expression rate of genes g at location s, $\mu_{s,g}$ in the mRNA count space is modeled as a linear function of reference cell types signatures $g_{f,g}$:

$$\mu_{s,g} = \left(\underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}} .$$

Aggregating high-resolution spatial omics (sequencing) data

- bin2cell: combines segmentation on H&E/IF and segmentation on gene expression counts

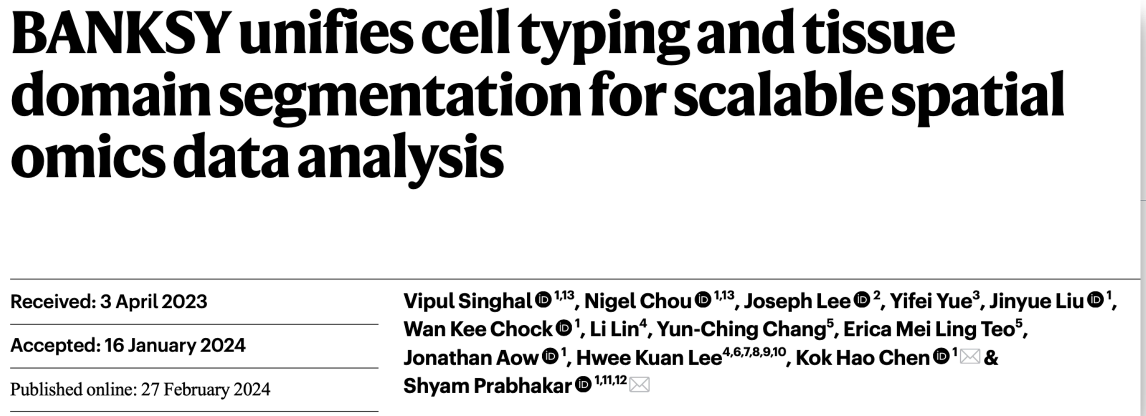
Image of counts per spot (smoothed) →



Statistical methods for spatial omics data

- Overview on the technologies (review)
- Finding spatially-variable genes
- Deconvoluting low-resolution (or aggregating high-resolution) spatial omics data
- Spatially-aware dimension reduction / clustering
- Cell-cell communication —> co-localization
- Classical spatial statistics
    - Point patterns: random, clustered, intensity/correlation
    - Lattice data: useful summaries / functions
    - models with spatially correlated errors

Spatial clustering / domain detection (BANKSY)
—> combine transcription and spatial information



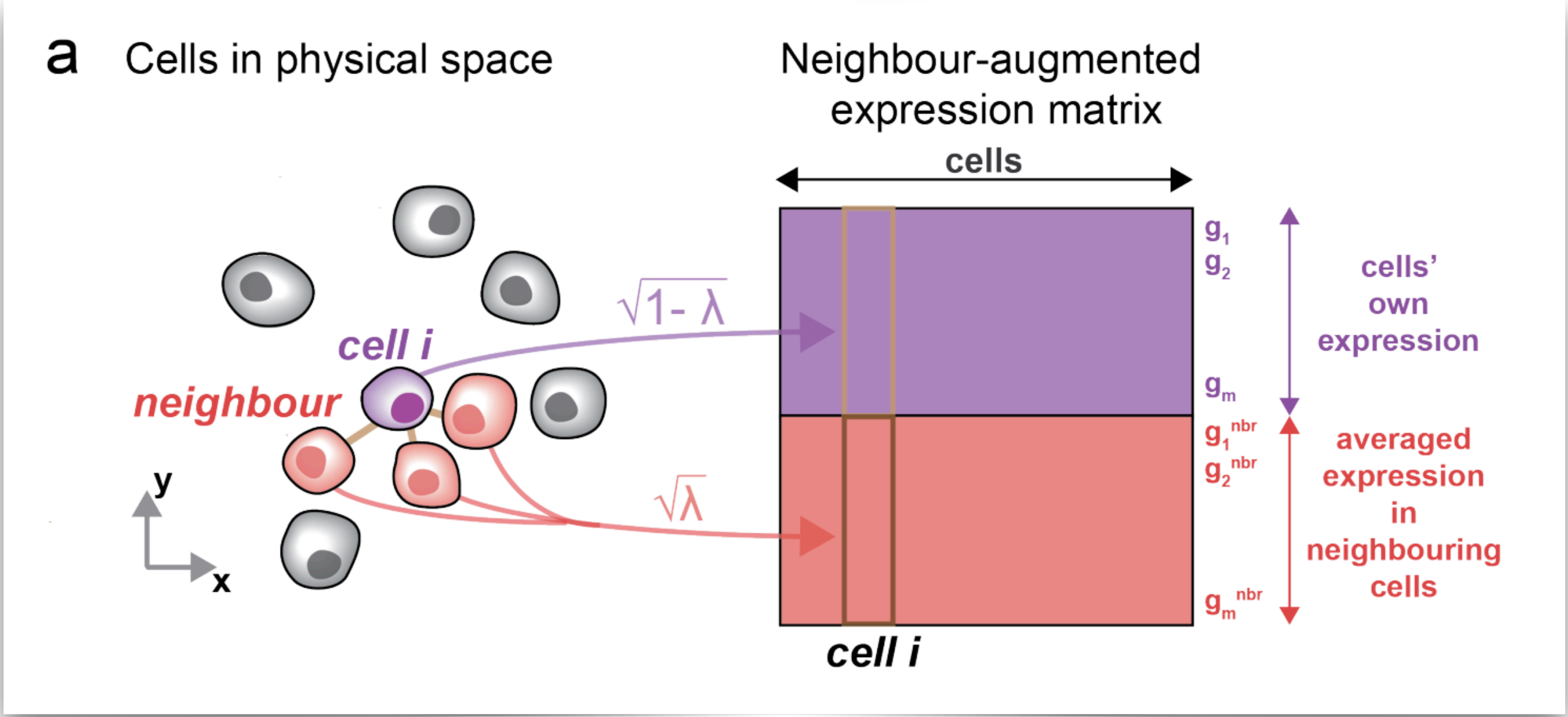
BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis

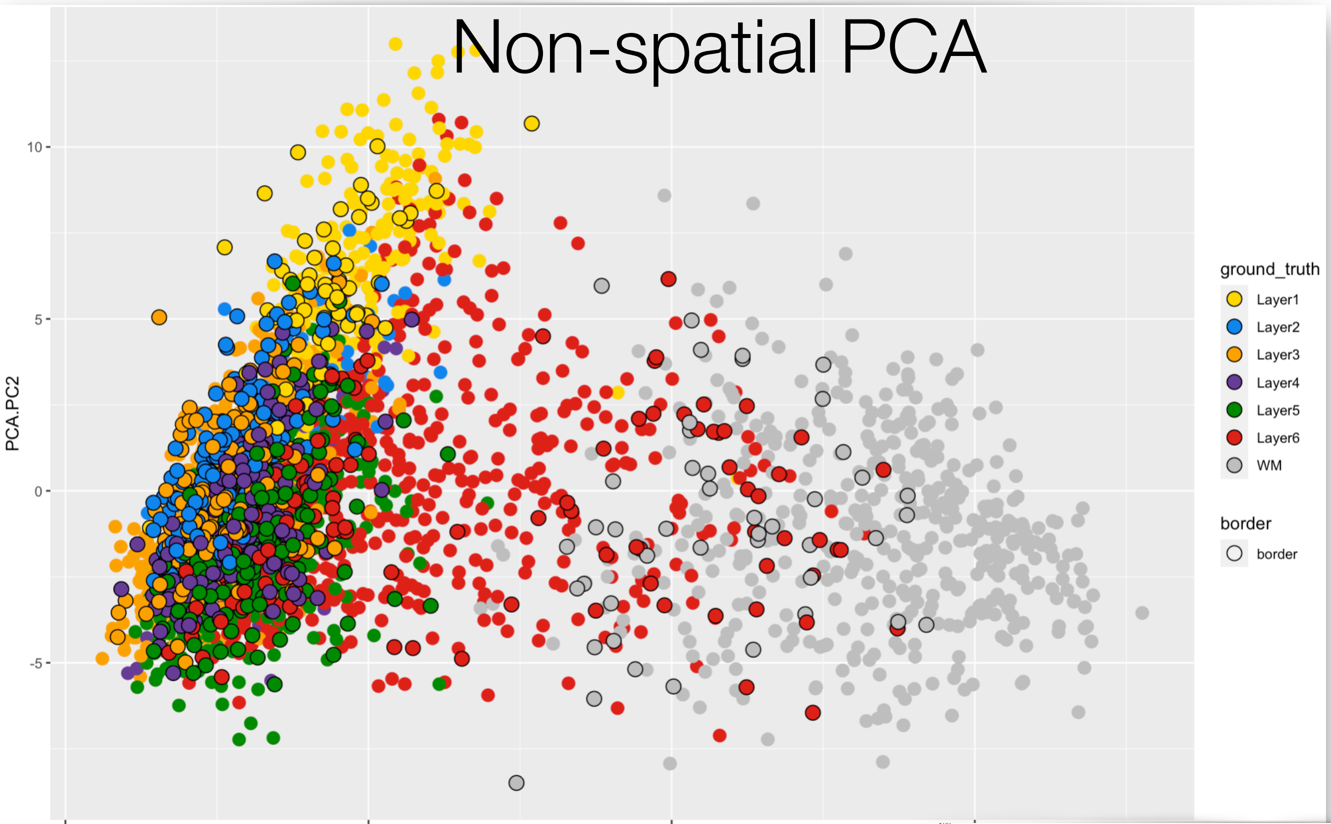
Received: 3 April 2023

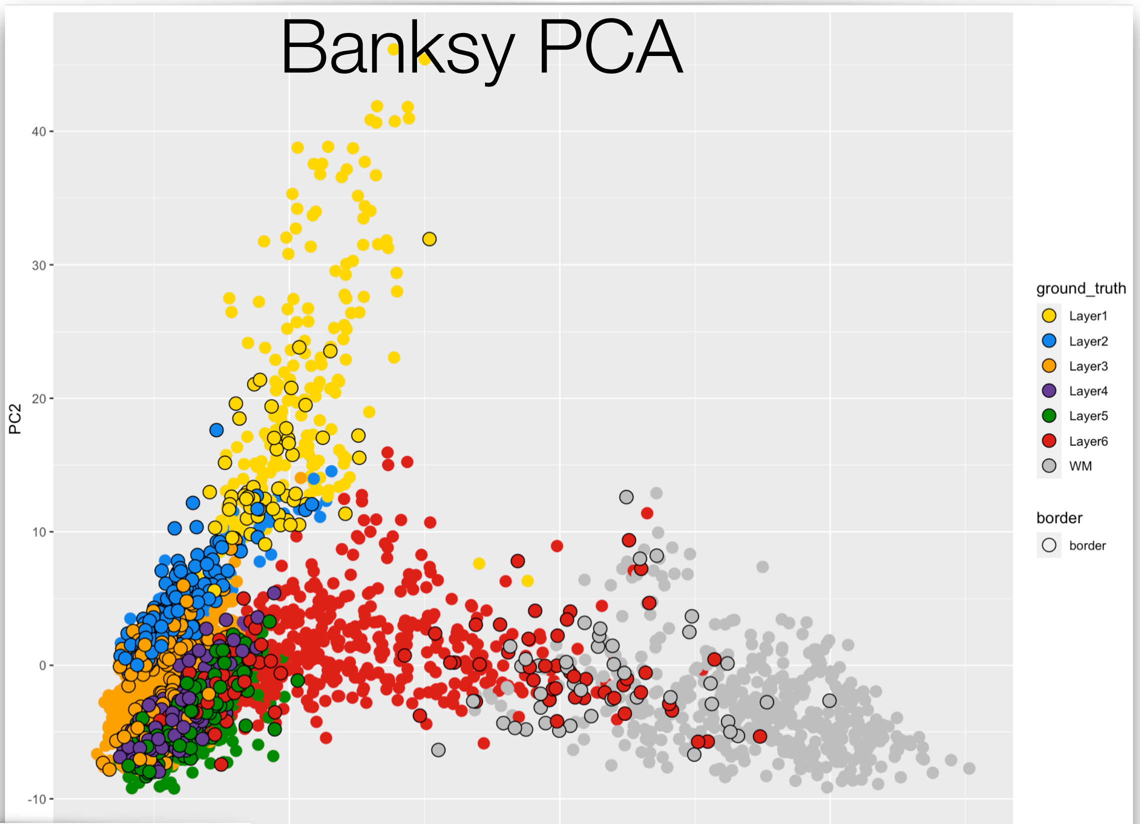
Accepted: 16 January 2024

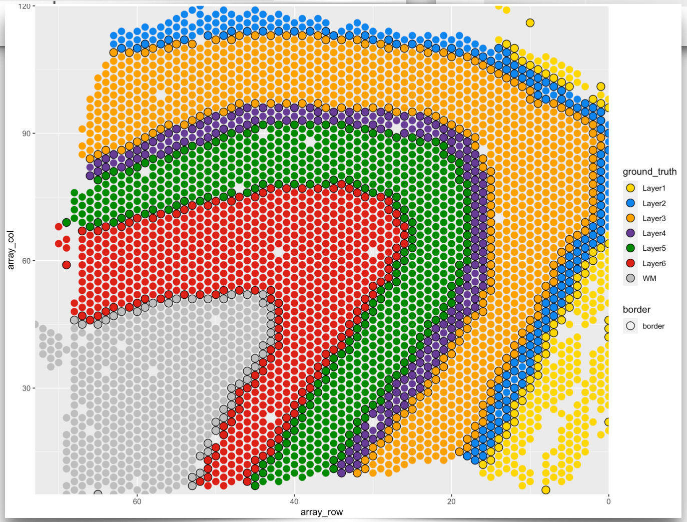
Published online: 27 February 2024

Vipul Singhal[1,13], Nigel Chou[1,13], Joseph Lee[2], Yifei Yue[3], Jinyue Liu[1], Wan Kee Chock[1], Li Lin[4], Yun-Ching Chang[5], Erica Mei Ling Teo[5], Jonathan Aow[1], Hwee Kuan Lee[4,6,7,8,9,10], Kok Hao Chen[1] & Shyam Prabhakar[1,11,12]

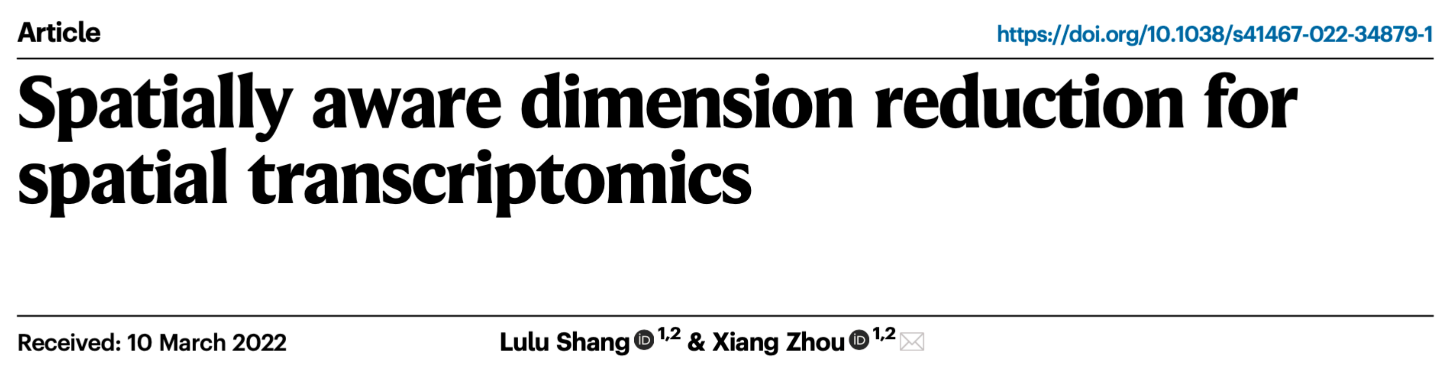



Non-spatial PCA


Banksy PCA



Sample 151673



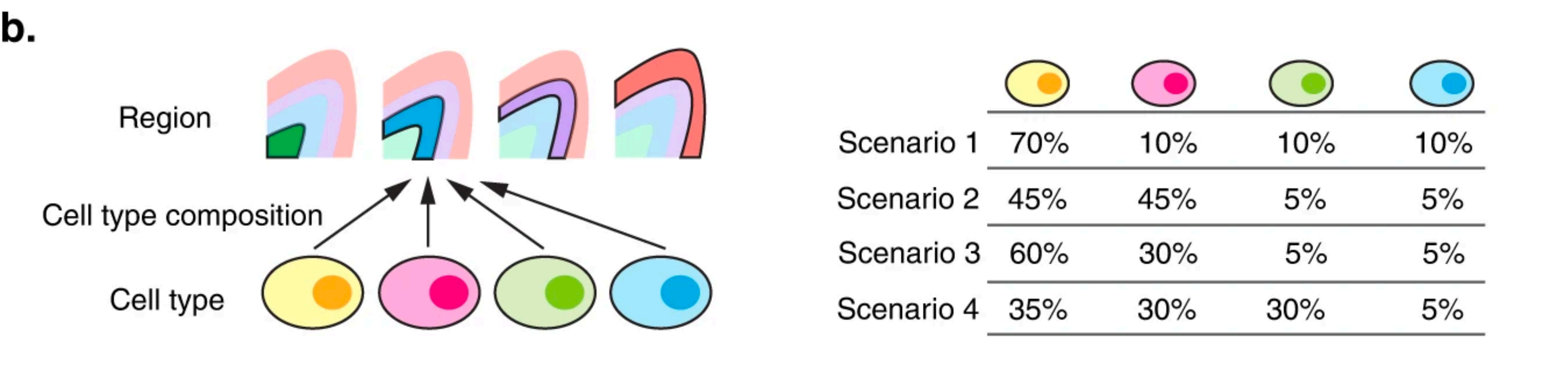
Article https://doi.org/10.1038/s41467-022-34879-1

Spatially aware dimension reduction for spatial transcriptomics

Received: 10 March 2022

Lulu Shang [1,2] & Xiang Zhou [1,2]

Spatial domain detection ~ spatially homogeneous regions ~ spatial niches

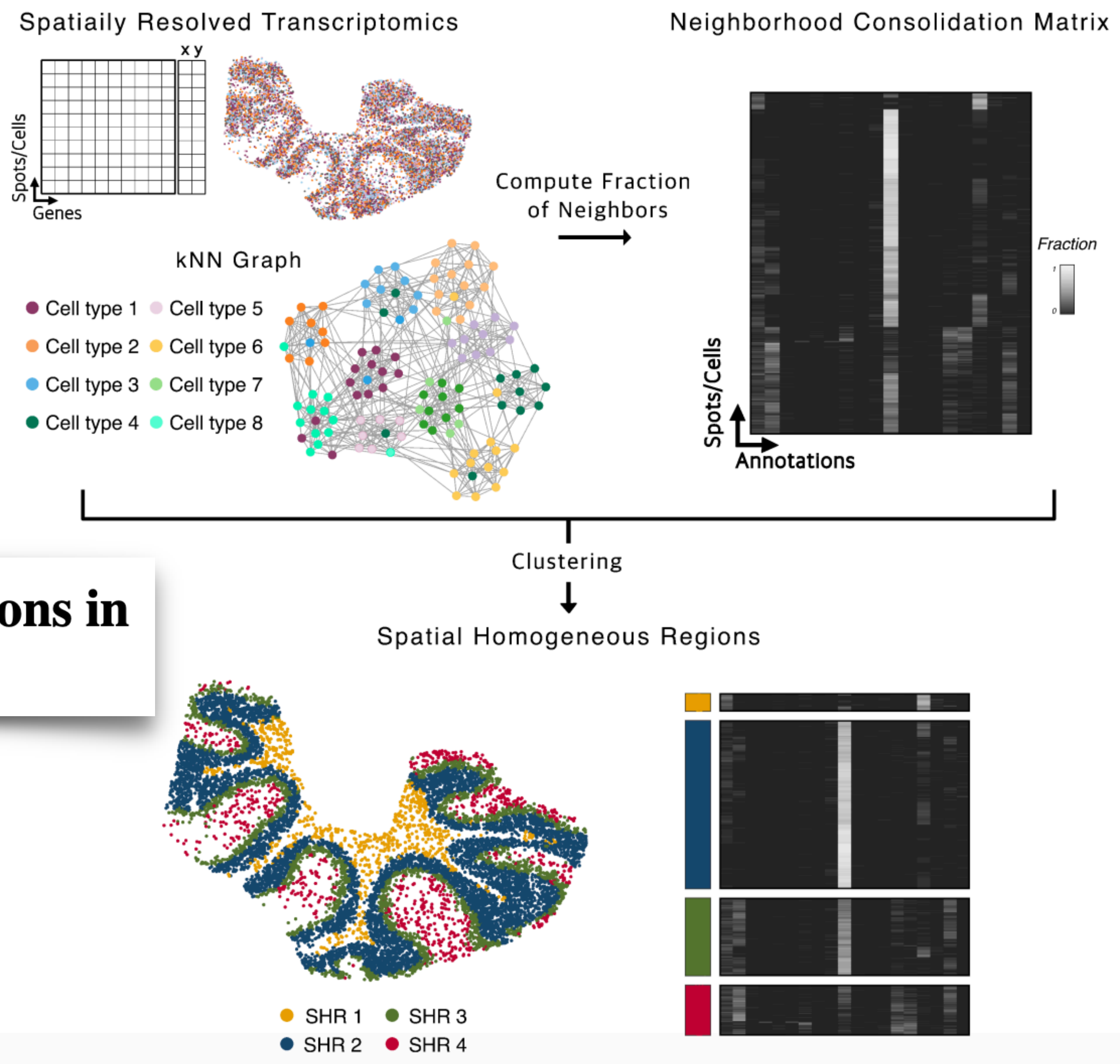


	●	●	●	●
Scenario 1	70%	10%	10%	10%
Scenario 2	45%	45%	5%	5%
Scenario 3	60%	30%	5%	5%
Scenario 4	35%	30%	30%	5%

https://www.nature.com/articles/s41467-022-34879-1

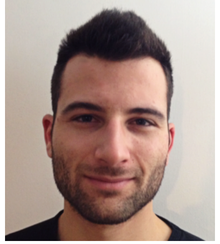
Spatial domain detection ~ spatially homogeneous regions

Identification of spatial homogeneous regions in tissues with concordex

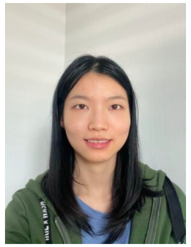


https://www.biorxiv.org/content/10.1101/2023.06.28.546949v2

Alternatively, spatially variable features = DE between domains



Simone Tiberi



Peiying Cai

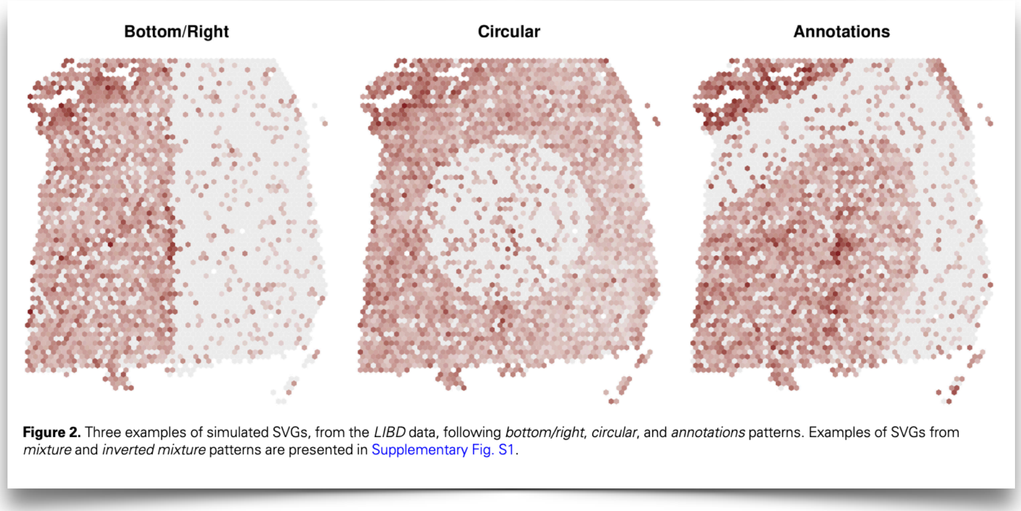
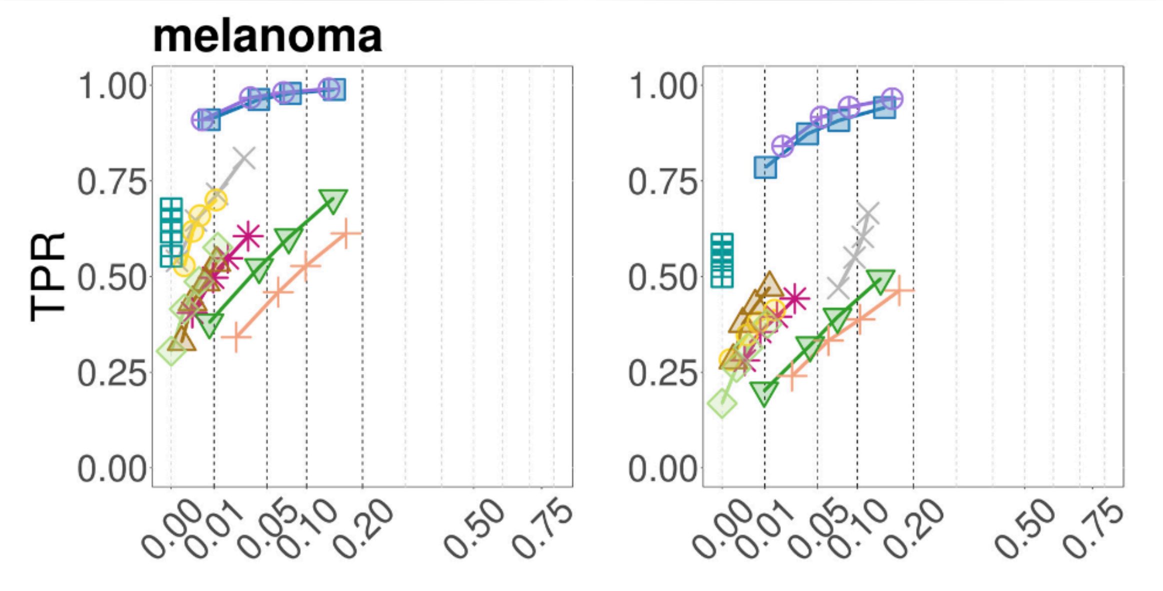
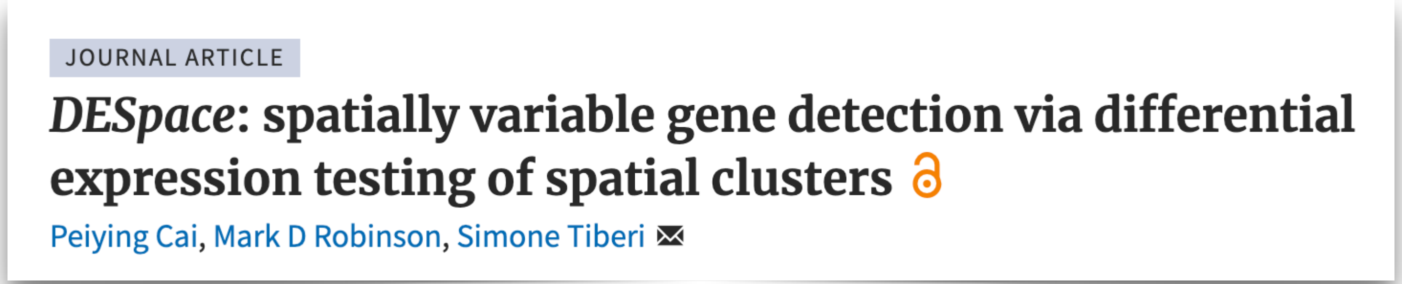


Figure 2. Three examples of simulated SVGs, from the *LIBD* data, following *bottom/right*, *circular*, and *annotations* patterns. Examples of SVGs from *mixture* and *inverted mixture* patterns are presented in Supplementary Fig. S1.



To find spatially variable genes (SVGs); spatial clustering + classical statistical method works quite well



JOURNAL ARTICLE

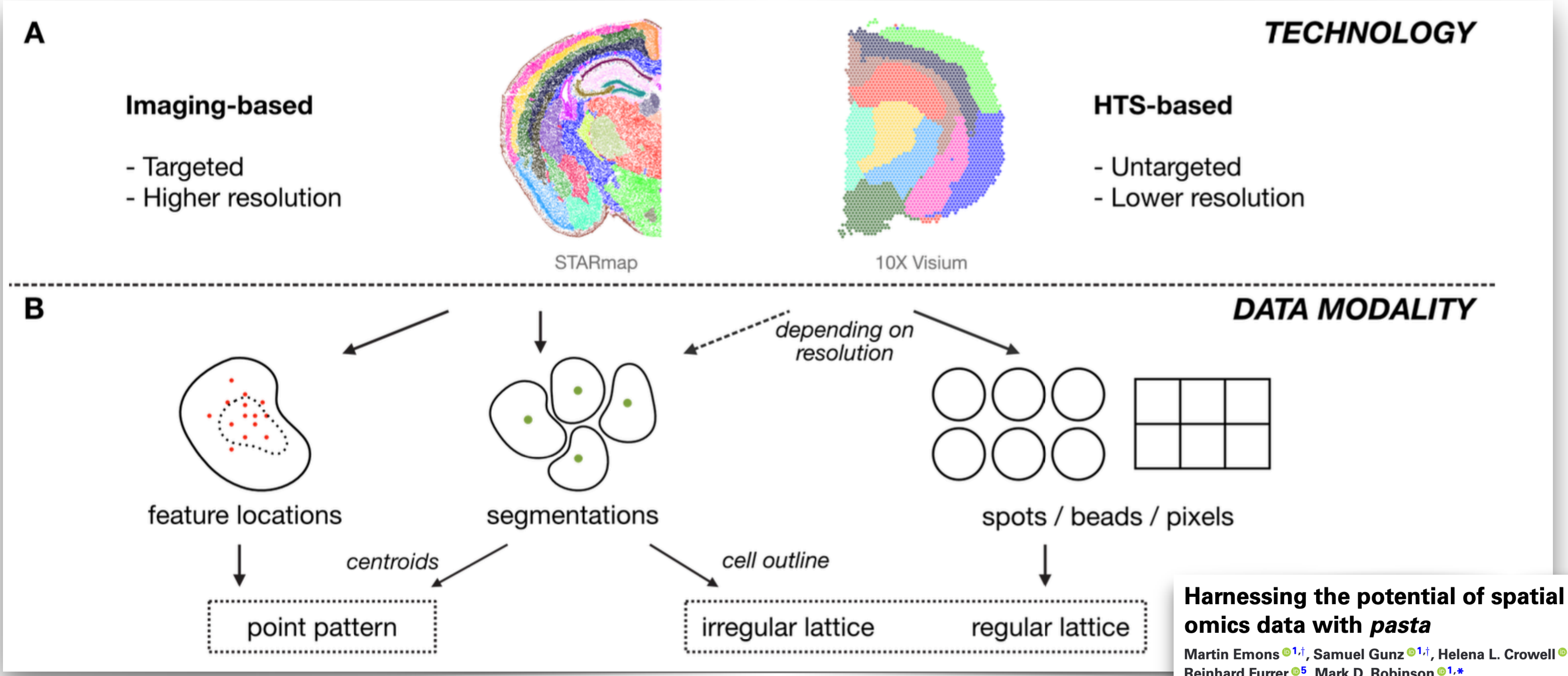
***DESpace*: spatially variable gene detection via differential expression testing of spatial clusters**

Peiying Cai, Mark D Robinson, Simone Tiberi ✉

Statistical methods for spatial omics data

- Overview on the technologies (review)
- Finding spatially-variable genes
- Deconvoluting low-resolution (or aggregating high-resolution) spatial omics data
- Spatially-aware dimension reduction / clustering
- Cell-cell communication —> co-localization
- Classical spatial statistics
  - Point patterns: random, clustered, intensity/correlation
  - Lattice data: useful summaries / functions
  - models with spatially correlated errors

`pasta`: Data representations determine spatial statistics options



Samuel

Martin

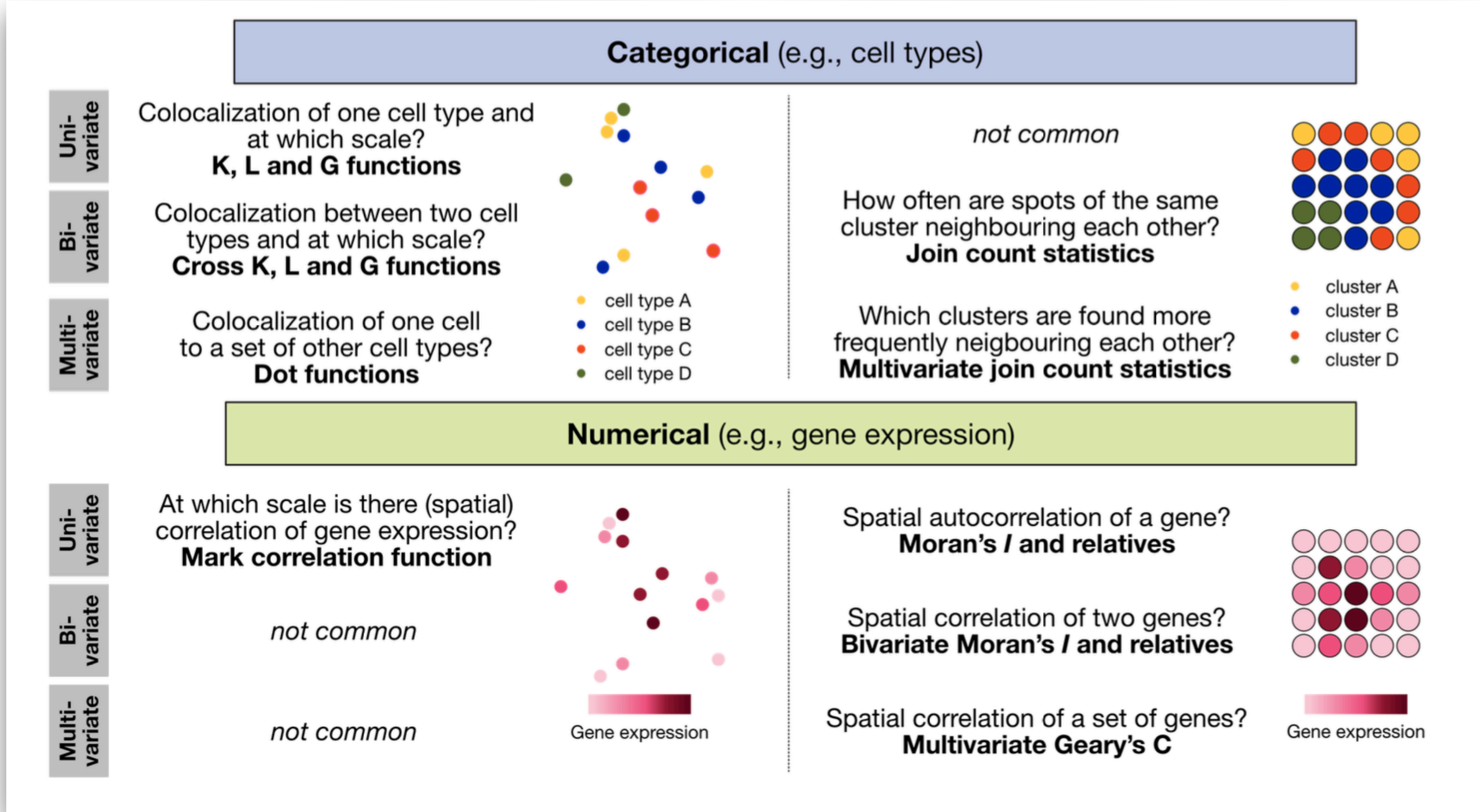
Harnessing the potential of spatial statistics for spatial omics data with *pasta*

Martin Emons[1,†], Samuel Gunz[1,†], Helena L. Crowell[2], Izaskun Mallona[1], Malte Kuehl[3,4], Reinhard Furrer[5], Mark D. Robinson[1,*]

[1] Department of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland
[2] Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain
[3] Department of Clinical Medicine, Aarhus University, 8200 Aarhus N, Denmark
[4] Department of Pathology, Aarhus University Hospital, 8200 Aarhus N, Denmark
[5] Department of Mathematical Modeling and Machine Learning, University of Zurich, 8057 Zurich, Switzerland

[*] To whom correspondence should be addressed. Email: mark.robinson@mls.uzh.ch
[†] The first two authors should be regarded as Joint First Authors.

`pasta`: Data representations determine spatial statistics options



Categorical (e.g., cell types)

Uni-variate: Colocalization of one cell type and at which scale?
K, L and G functions

Bi-variate: Colocalization between two cell types and at which scale?
Cross K, L and G functions

Multi-variate: Colocalization of one cell to a set of other cell types?
Dot functions

not common

How often are spots of the same cluster neighbouring each other?
Join count statistics

Which clusters are found more frequently neigbouring each other?
Multivariate join count statistics

Numerical (e.g., gene expression)

Uni-variate: At which scale is there (spatial) correlation of gene expression?
Mark correlation function

Bi-variate: *not common*

Multi-variate: *not common*

Spatial autocorrelation of a gene?
Moran's *I* and relatives

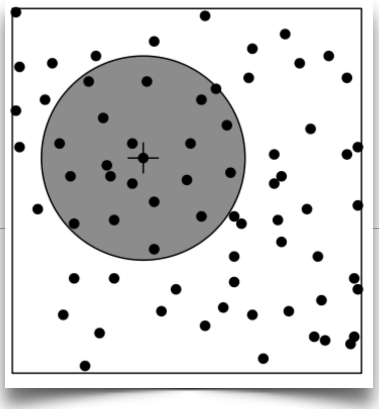
Spatial correlation of two genes?
Bivariate Moran's *I* and relatives

Spatial correlation of a set of genes?
Multivariate Geary's C

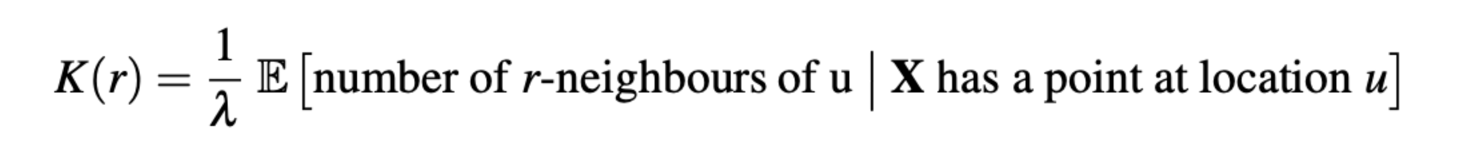
Samuel

Martin

Correlation for **point patterns**

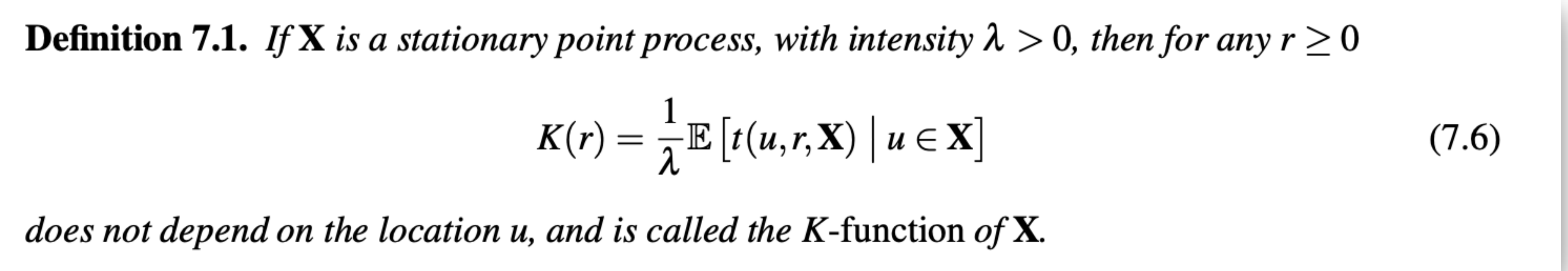


- Ripley's K function
- mathematical definition:



$$K(r) = \frac{1}{\lambda}\,\mathbb{E}\left[\text{number of } r\text{-neighbours of u} \;\middle|\; \mathbf{X} \text{ has a point at location } u\right]$$

$$t(u,r,\mathbf{x}) = \sum_{j=1}^{n(\mathbf{x})} \mathbf{1}\{0 < \|u - x_j\| \le r\}$$



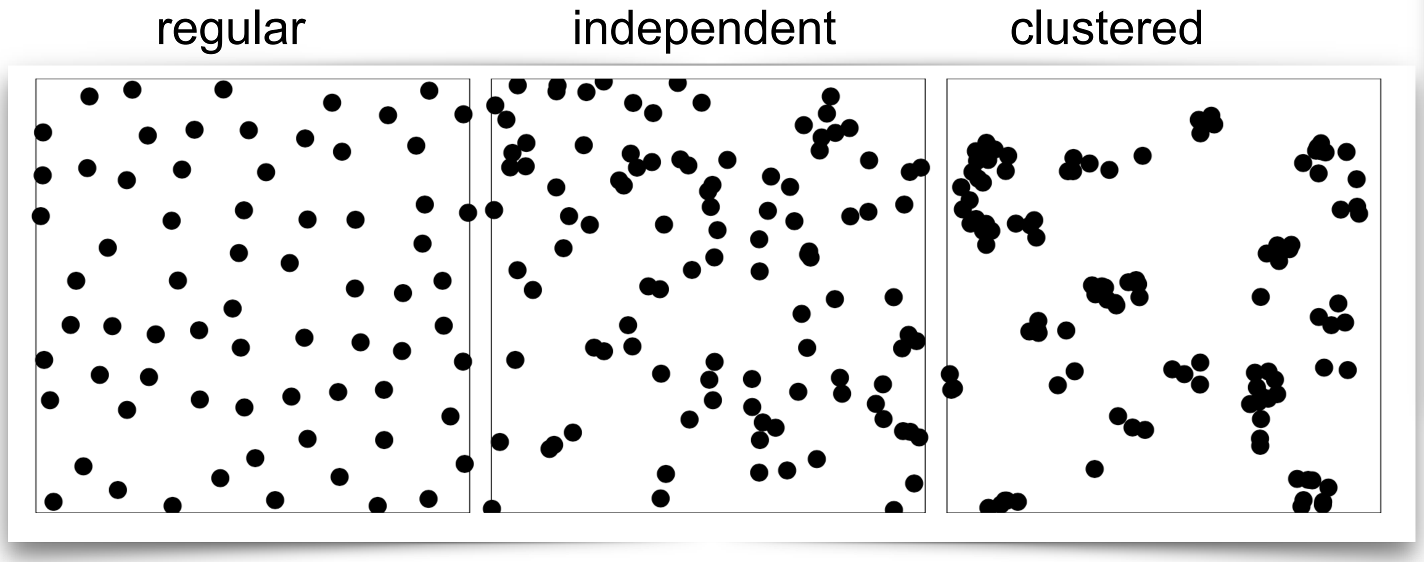
Definition 7.1. *If* $\mathbf{X}$ *is a stationary point process, with intensity* $\lambda > 0$, *then for any* $r \ge 0$

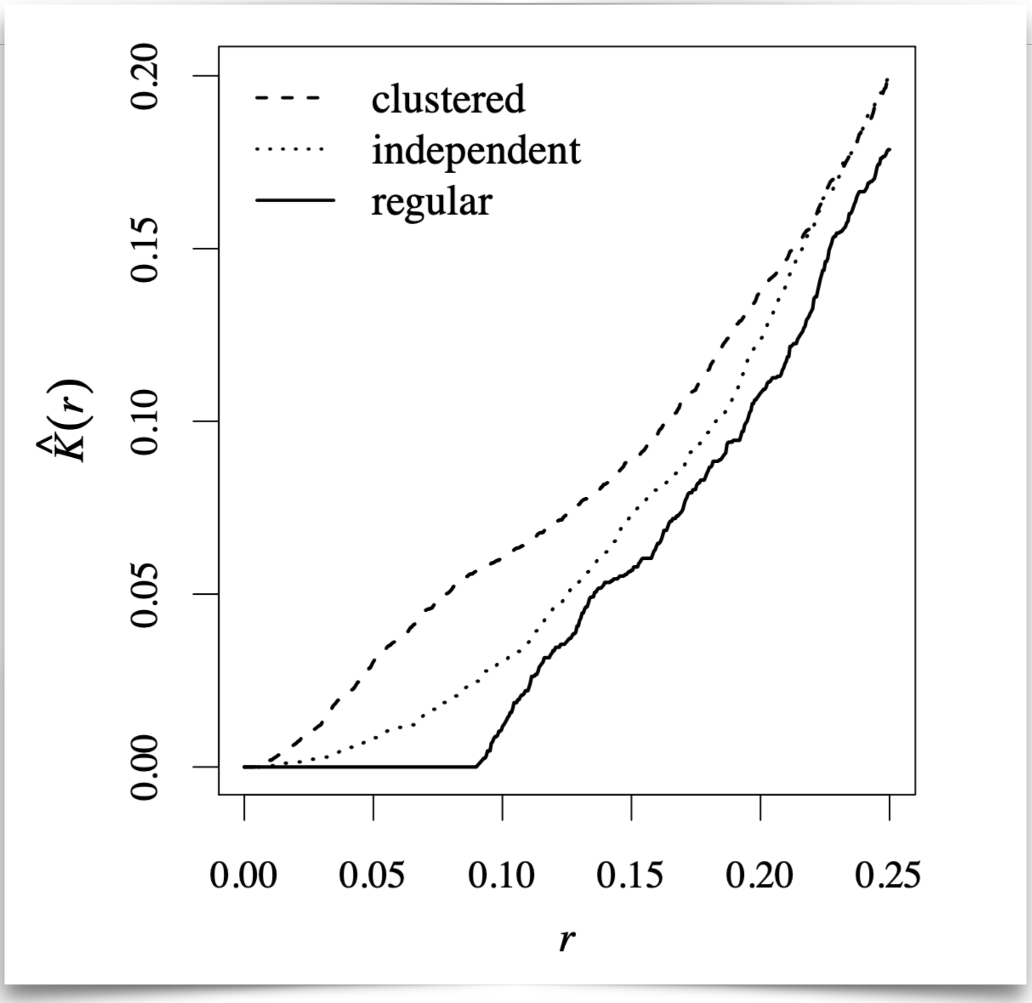
$$K(r) = \frac{1}{\lambda}\mathbb{E}\left[t(u,r,\mathbf{X}) \;\middle|\; u \in \mathbf{X}\right] \qquad (7.6)$$

does not depend on the location u, and is called the K-function *of* $\mathbf{X}$.

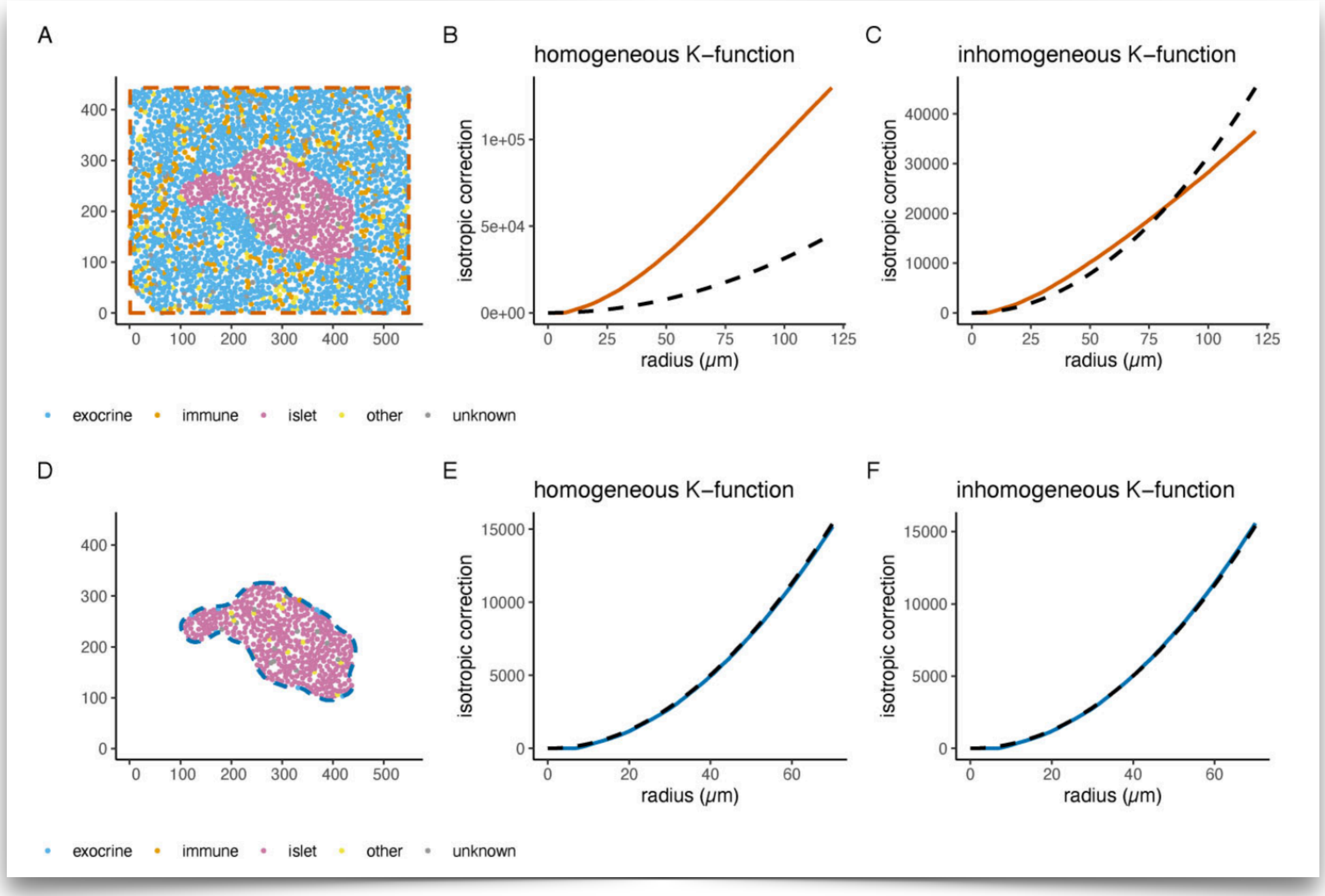
Correlation for **point patterns**

- Ripley's K function
- words definition: *the empirical K-function K(r) is the cumulative average number of data points lying within a distance r of a typical data point*





`pasta`: the ‘gotcha’ of spatial statistics — is it clustering or intensity?

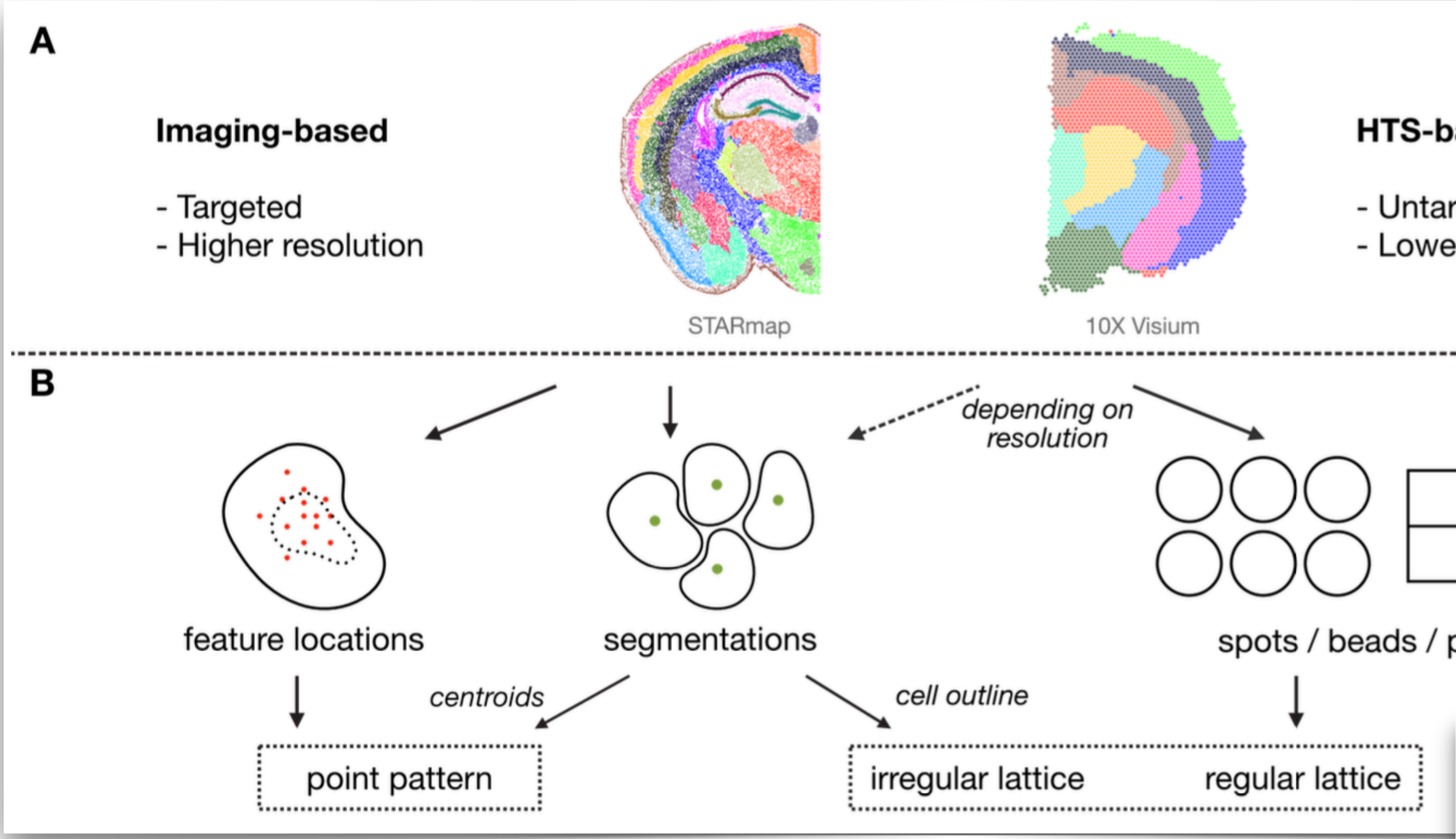


Samuel

Martin

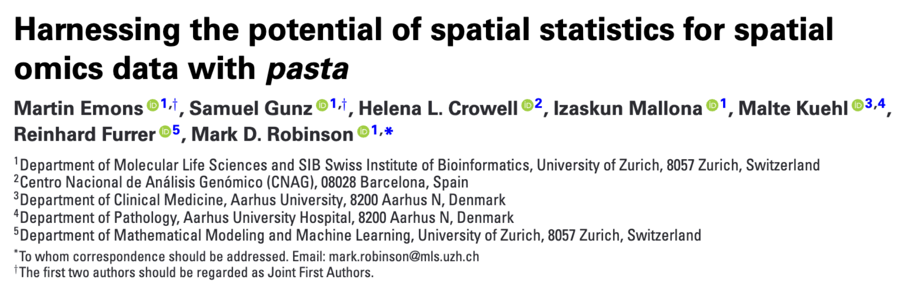
K-functions here: clustering / intensity of pink cells (islets).

pasta: Data representations determine spatial statistics options



Samuel

Martin



Harnessing the potential of spatial statistics for spatial omics data with *pasta*

Martin Emons[1,†], Samuel Gunz[1,†], Helena L. Crowell[2], Izaskun Mallona[1], Malte Kuehl[3,4], Reinhard Furrer[5], Mark D. Robinson[1,*]

[1] Department of Molecular Life Sciences and SIB Swiss Institute of Bioinformatics, University of Zurich, 8057 Zurich, Switzerland
[2] Centro Nacional de Análisis Genómico (CNAG), 08028 Barcelona, Spain
[3] Department of Clinical Medicine, Aarhus University, 8200 Aarhus N, Denmark
[4] Department of Pathology, Aarhus University Hospital, 8200 Aarhus N, Denmark
[5] Department of Mathematical Modeling and Machine Learning, University of Zurich, 8057 Zurich, Switzerland

[*] To whom correspondence should be addressed. Email: mark.robinson@mls.uzh.ch

[†] The first two authors should be regarded as Joint First Authors.

https://en.wikipedia.org/wiki/Moran%27s_I
https://en.wikipedia.org/wiki/Geary%27s_C

Spatial autocorrelation: Global Moran’s I

- Global measure of auto-correlation (correlation to signal nearby in space); assume homogeneity!
- Alternative: Geary’s C

$$I = \frac{1}{\sum_{ij} w_{ij}} \frac{\sum_{ij} w_{ij} \left(X_i - \bar{X}\right)\left(X_j - \bar{X}\right)}{N^{-1} \sum_i \left(X_i - \bar{X}\right)^2}$$

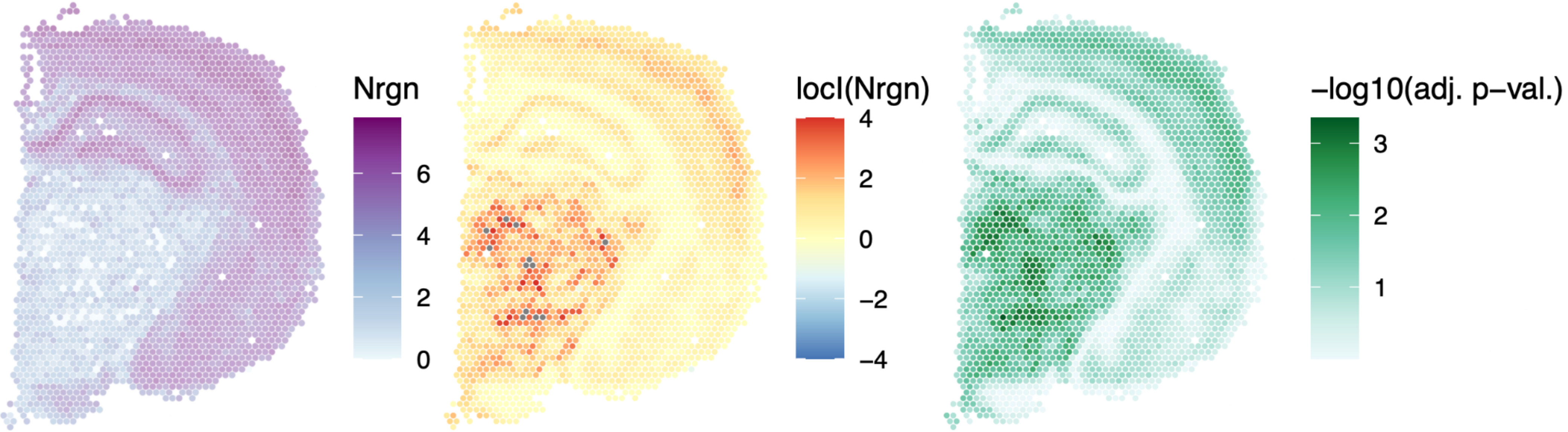
$$C = \frac{(N-1) \sum_i \sum_j w_{ij} (x_i - x_j)^2}{2W \sum_i (x_i - \bar{x})^2}$$

https://en.wikipedia.org/wiki/Moran%27s_I
https://en.wikipedia.org/wiki/Geary%27s_C

Spatial autocorrelation: Local Moran’s I

- Local measure of auto-correlation (correlation to signal nearby in space)

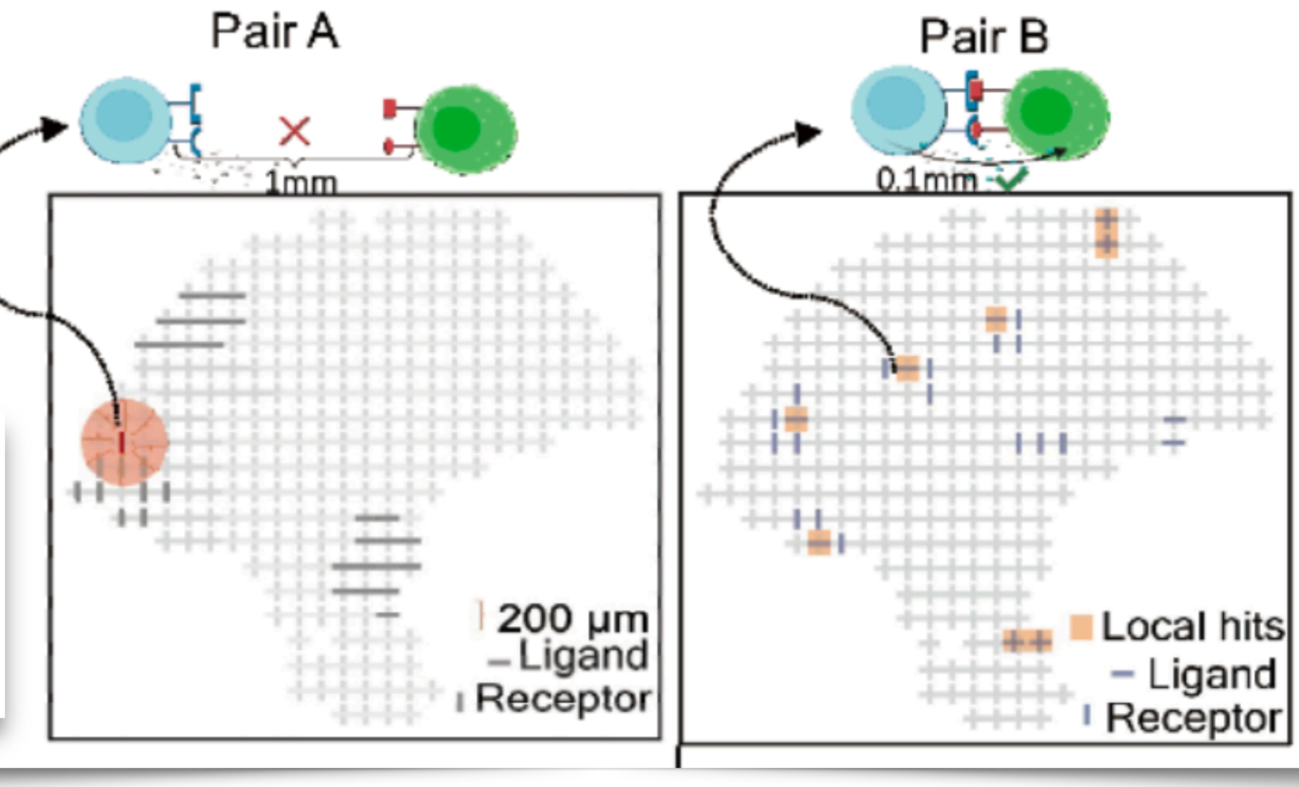
$$I_i = \frac{x_i - \bar{x}}{\sum_{k=1}^{n}(x_k - \bar{x})^2/(n-1)} \sum_{j=1}^{n} w_{ij}(x_j - \bar{x})$$

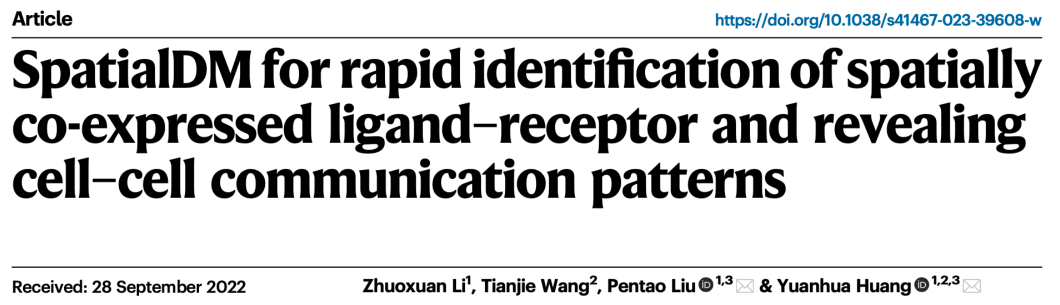


$$\text{Global Moran's } R = \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}},$$

Cell-cell communication

– SpatialDM: Global Moran's R, which is a bivariate version of Moran's I





Article

https://doi.org/10.1038/s41467-023-39608-w

SpatialDM for rapid identification of spatially co-expressed ligand–receptor and revealing cell–cell communication patterns

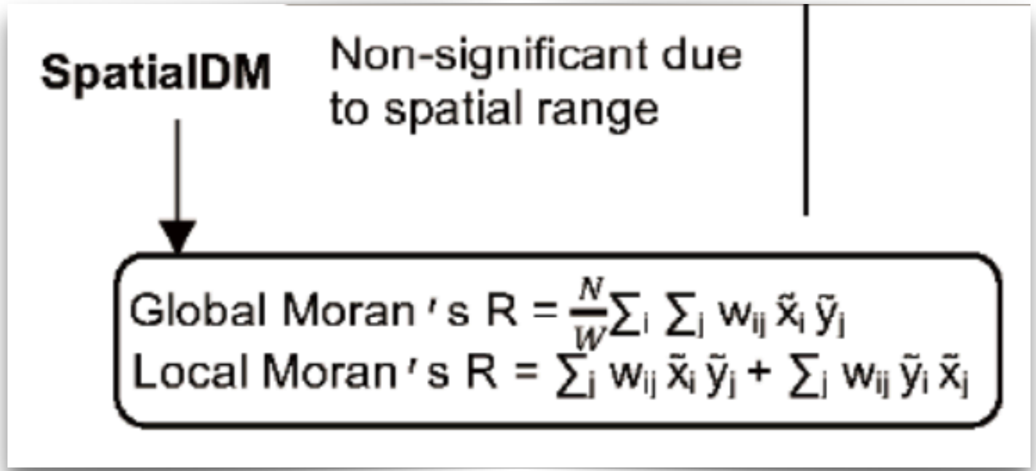
Received: 28 September 2022

Zhuoxuan Li[1], Tianjie Wang[2], Pentao Liu[1,3] & Yuanhua Huang[1,2,3]

Cell-cell communication

$$\text{Global Moran's } R = \frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}},$$

- SpatialDM: Global Moran's R, which is a bivariate version of Moran's I

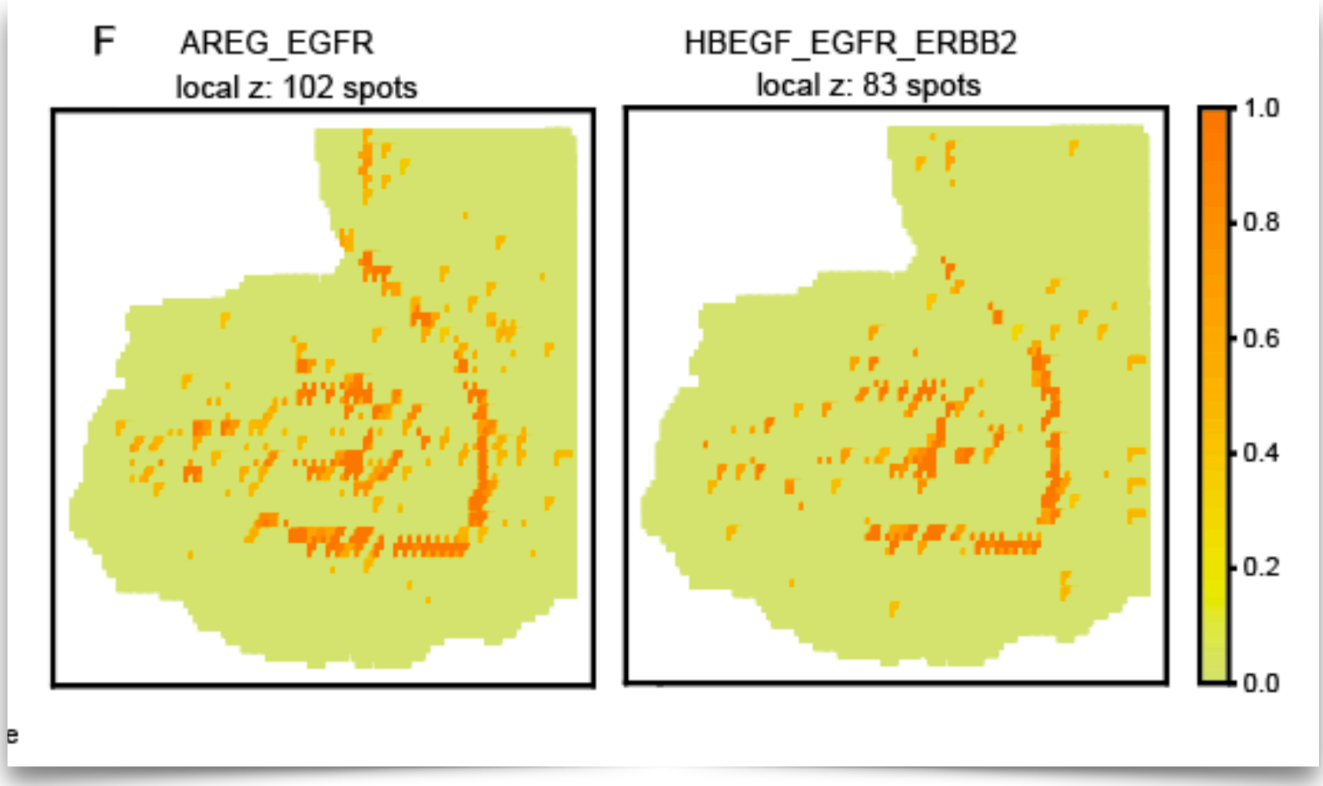


SpatialDM

Non-significant due to spatial range

Global Moran's R = $\frac{N}{W}\sum_i \sum_j w_{ij}\tilde{x}_i\tilde{y}_j$

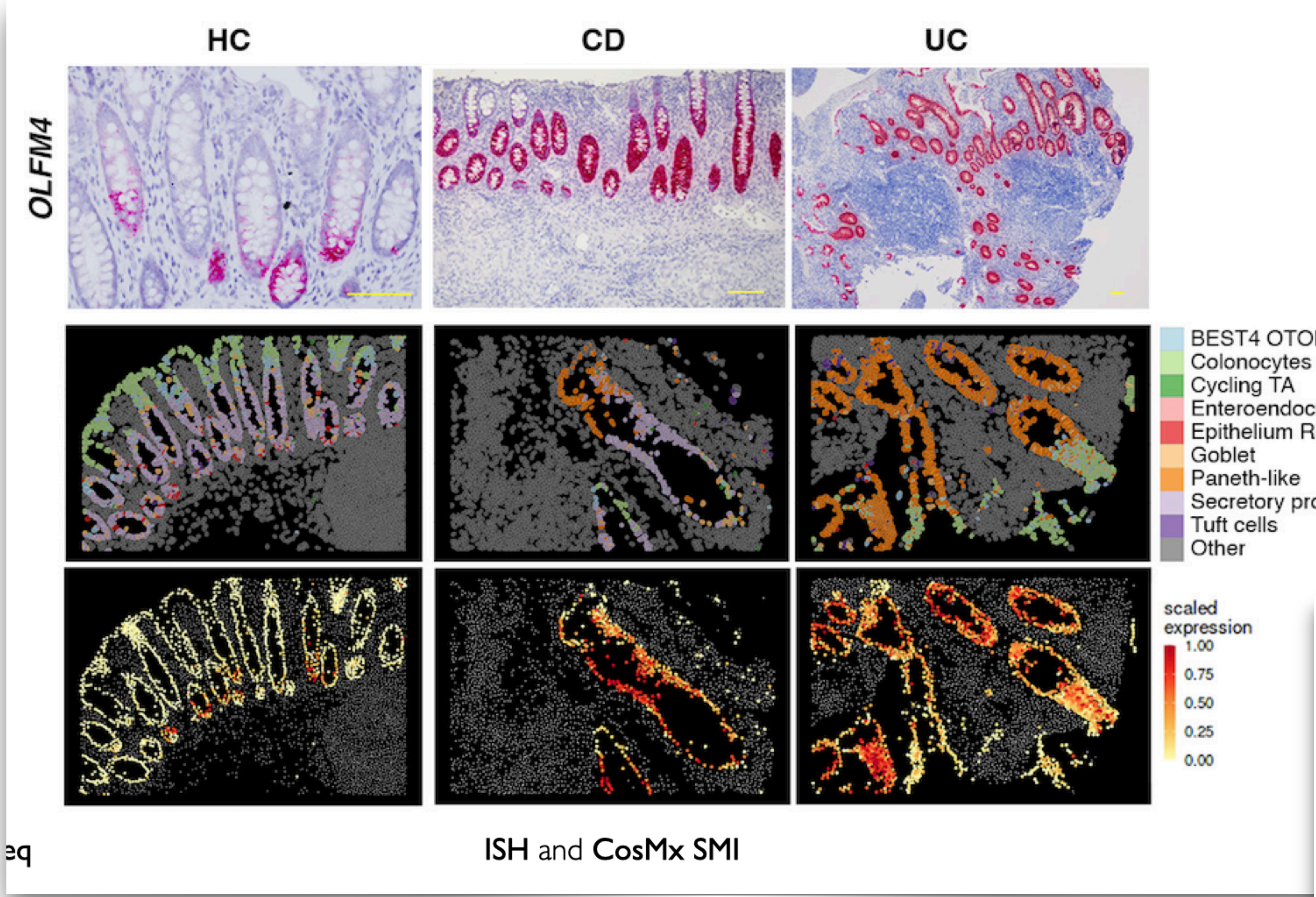
Local Moran's R = $\sum_j w_{ij}\tilde{x}_i\tilde{y}_j + \sum_j w_{ij}\tilde{y}_i\tilde{x}_j$



Research

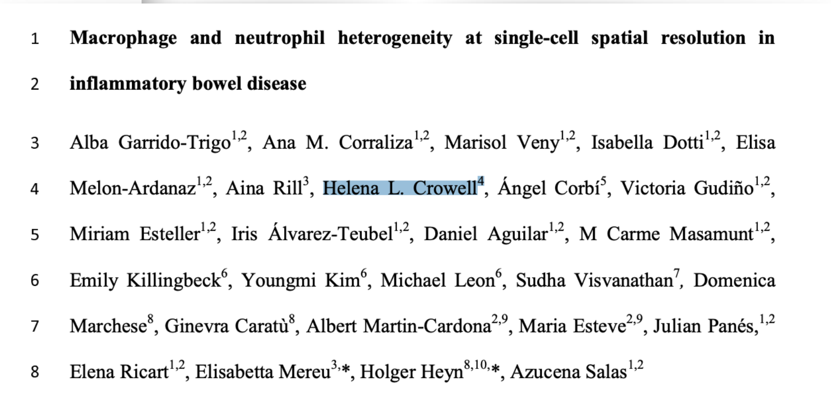
- `spatialFDA`: Flexible modeling of point pattern summaries —> Martin
- `DESpace2`: DE beyond markers/SVGs: “differential spatial patterns” —> Peiying
- **`sosta`: “Spatial structure”-focused analyses**
- OSTA: Orchestrating spatial transcriptomics analysis with Bioconductor
- SpaceHack: using consensus clustering to consolidate domain detection

Tissue “structures” are often visible



ISH and CosMx SMI

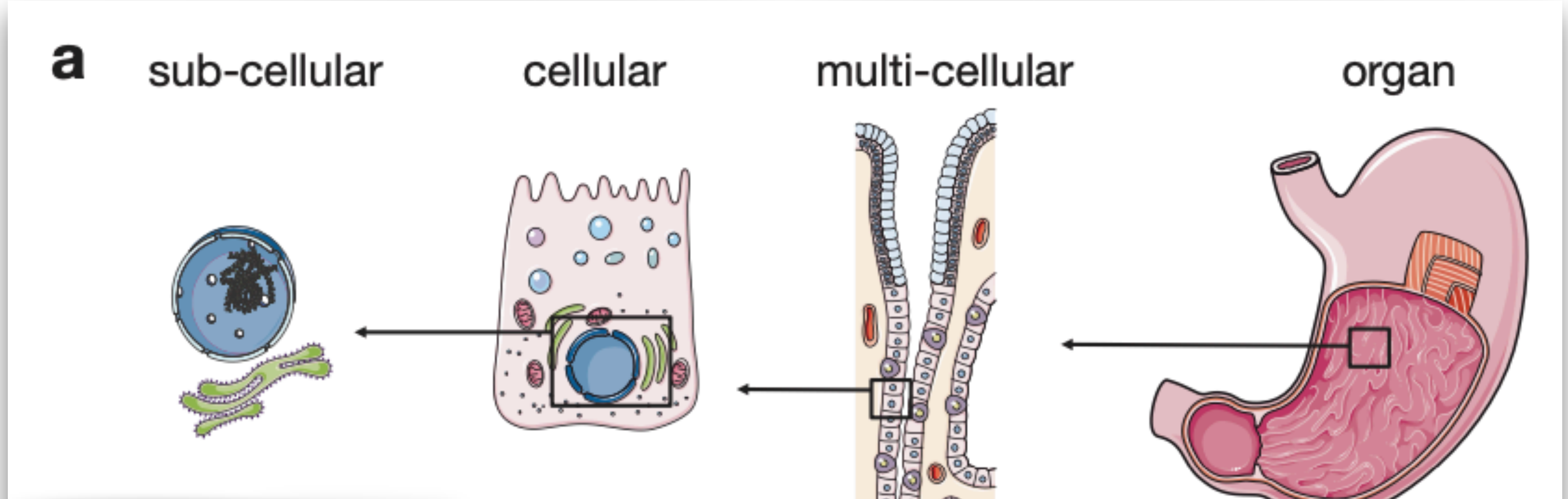
- healthy control (HC)
- Crohn's disease (CD)
- ulcerative colitis (UC)

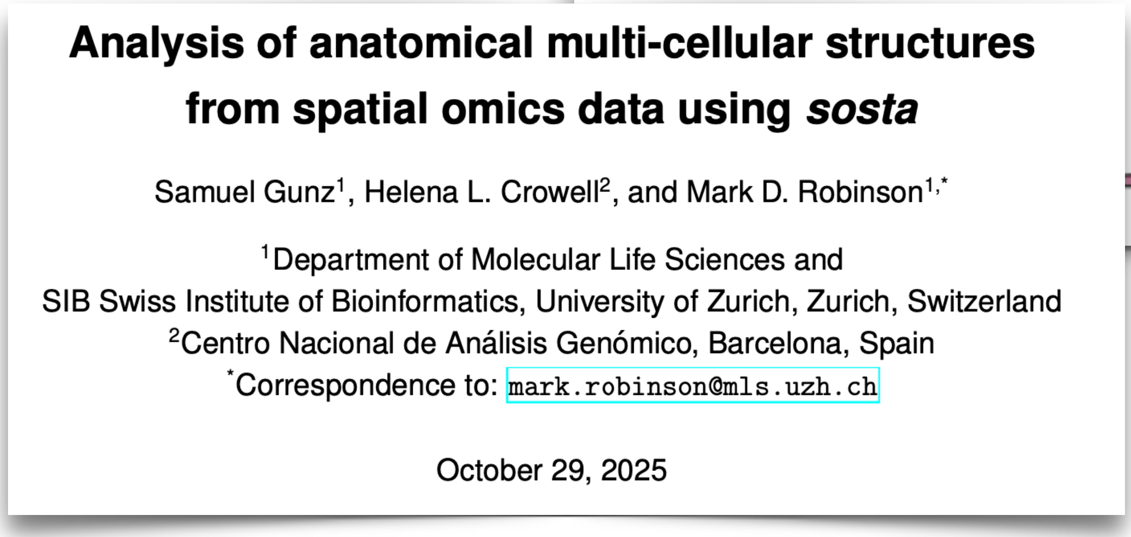


Macrophage and neutrophil heterogeneity at single-cell spatial resolution in inflammatory bowel disease

Alba Garrido-Trigo[1,2], Ana M. Corraliza[1,2], Marisol Veny[1,2], Isabella Dotti[1,2], Elisa Melon-Ardanaz[1,2], Aina Rill[3], Helena L. Crowell[4], Ángel Corbí[5], Victoria Gudiño[1,2], Miriam Esteller[1,2], Iris Álvarez-Teubel[1,2], Daniel Aguilar[1,2], M Carme Masamunt[1,2], Emily Killingbeck[6], Youngmi Kim[6], Michael Leon[6], Sudha Visvanathan[7], Domenica Marchese[8], Ginevra Caratù[8], Albert Martin-Cardona[2,9], Maria Esteve[2,9], Julian Panés,[1,2] Elena Ricart[1,2], Elisabetta Mereu[3,*], Holger Heyn[8,10,*], Azucena Salas[1,2]

Tissue “structures” occur at different scales





Analysis of anatomical multi-cellular structures from spatial omics data using *sosta*

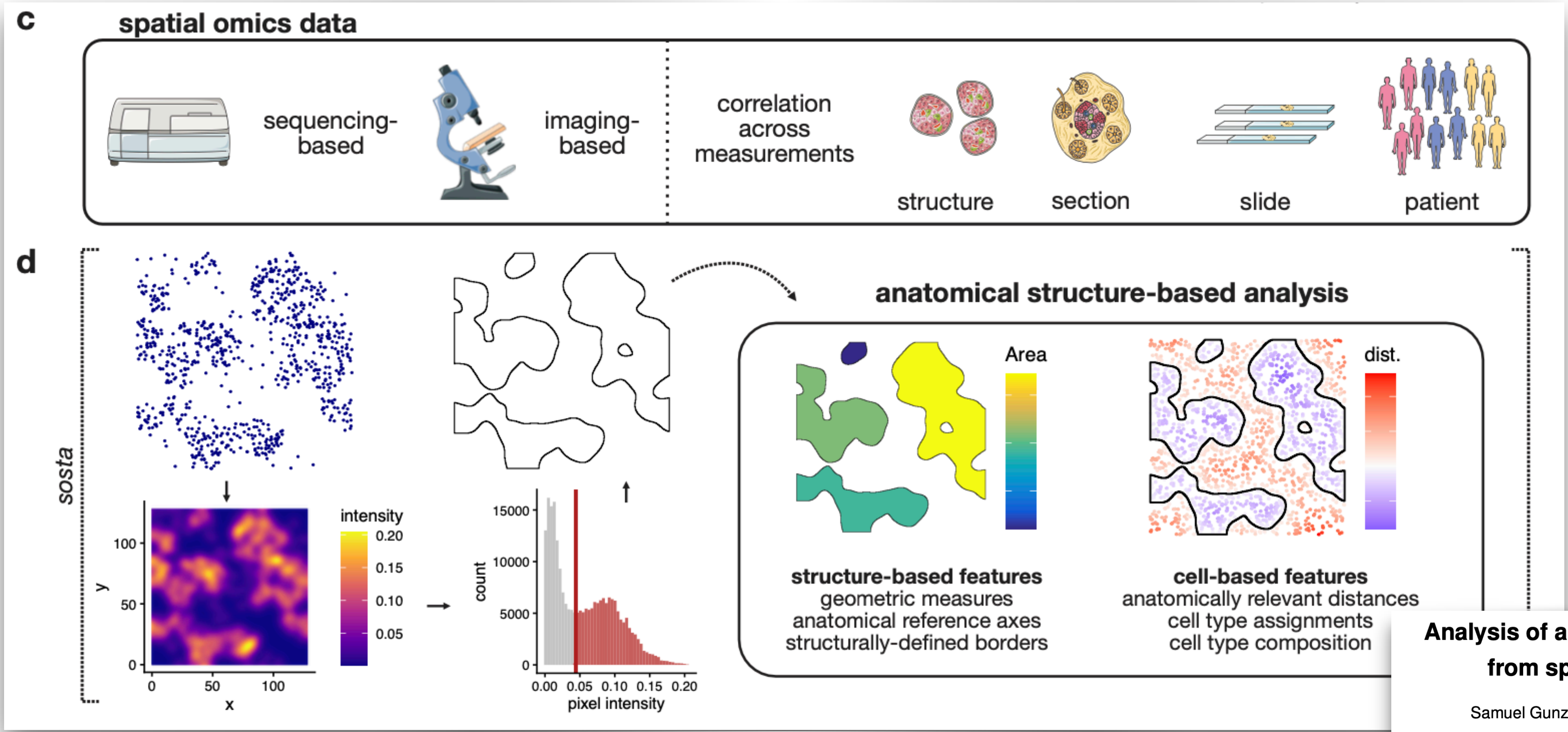
Samuel Gunz[1], Helena L. Crowell[2], and Mark D. Robinson[1,*]

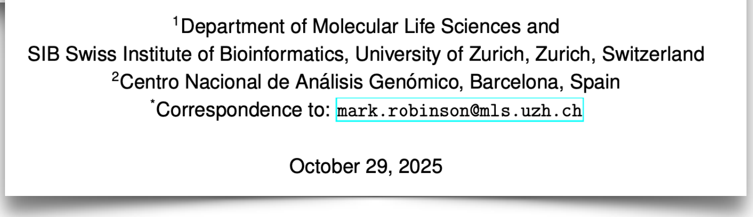
[1]Department of Molecular Life Sciences and
SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[2]Centro Nacional de Análisis Genómico, Barcelona, Spain
[*]Correspondence to: mark.robinson@mls.uzh.ch

October 29, 2025

sosta: extracting spatial “structures” + quantifying metrics + modelling (differential discovery)

Samuel





Analysis of anatomical multi-cellular structures from spatial omics data using *sosta*

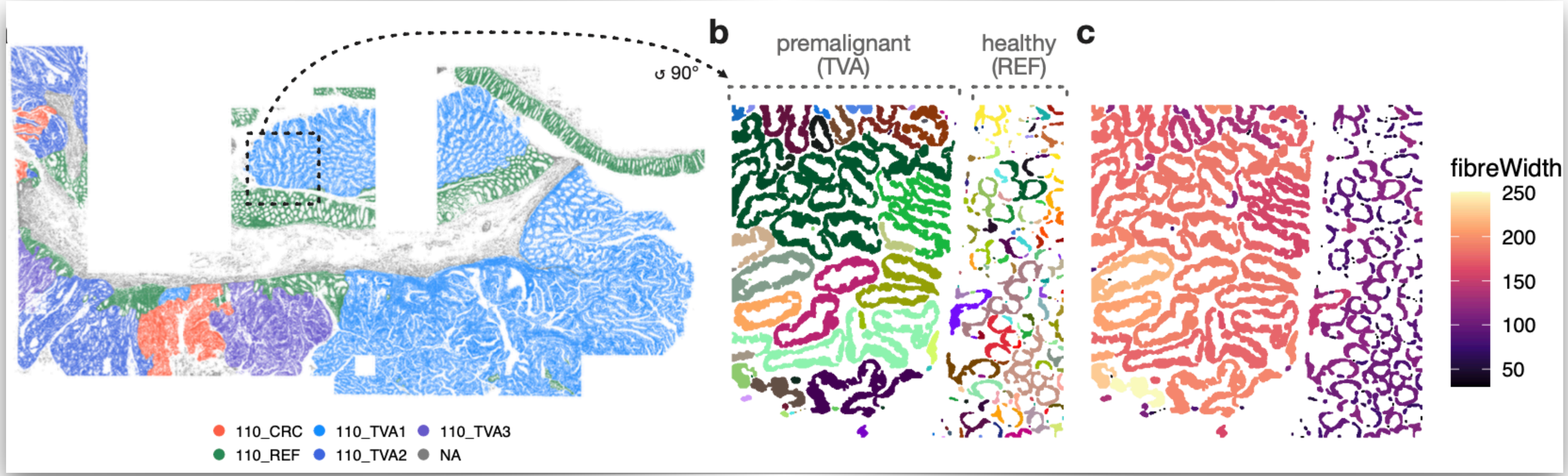
Samuel Gunz[1], Helena L. Crowell[2], and Mark D. Robinson[1,*]

[1]Department of Molecular Life Sciences and
SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[2]Centro Nacional de Análisis Genómico, Barcelona, Spain
[*]Correspondence to: mark.robinson@mls.uzh.ch

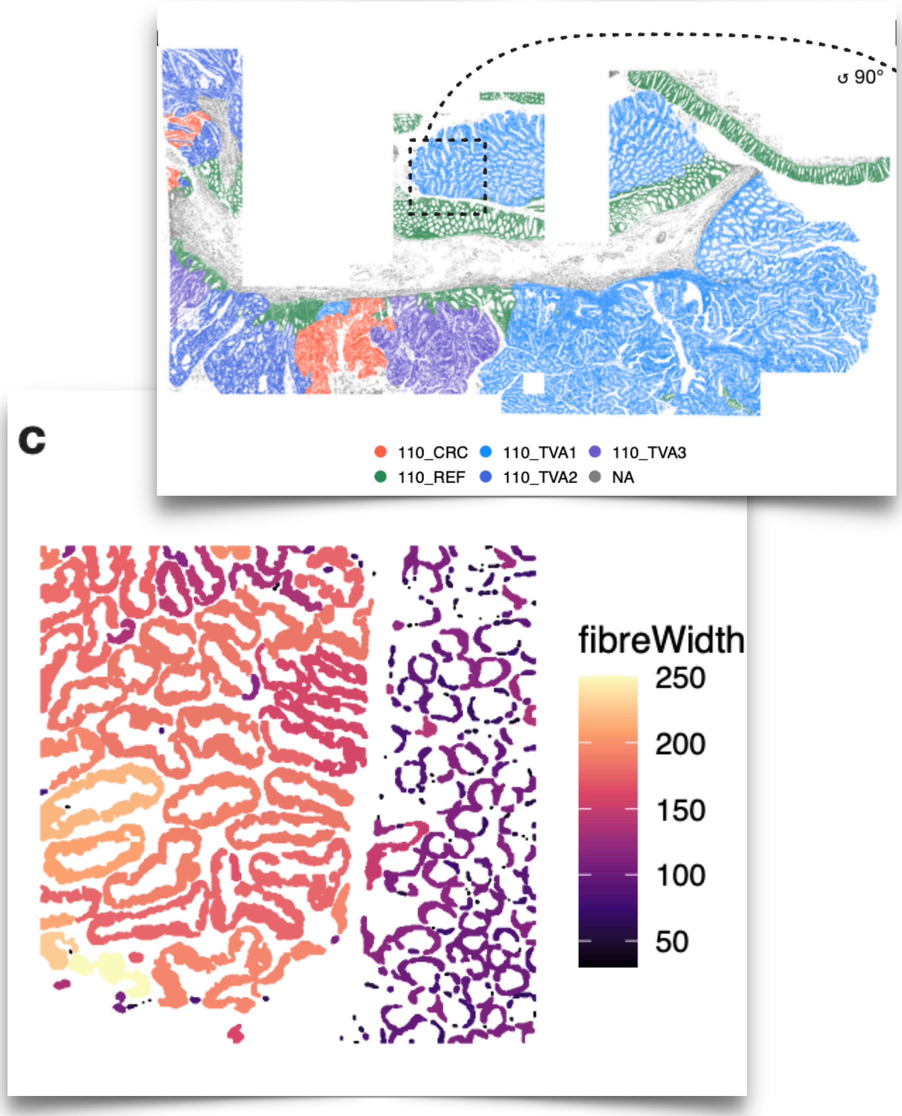
October 29, 2025

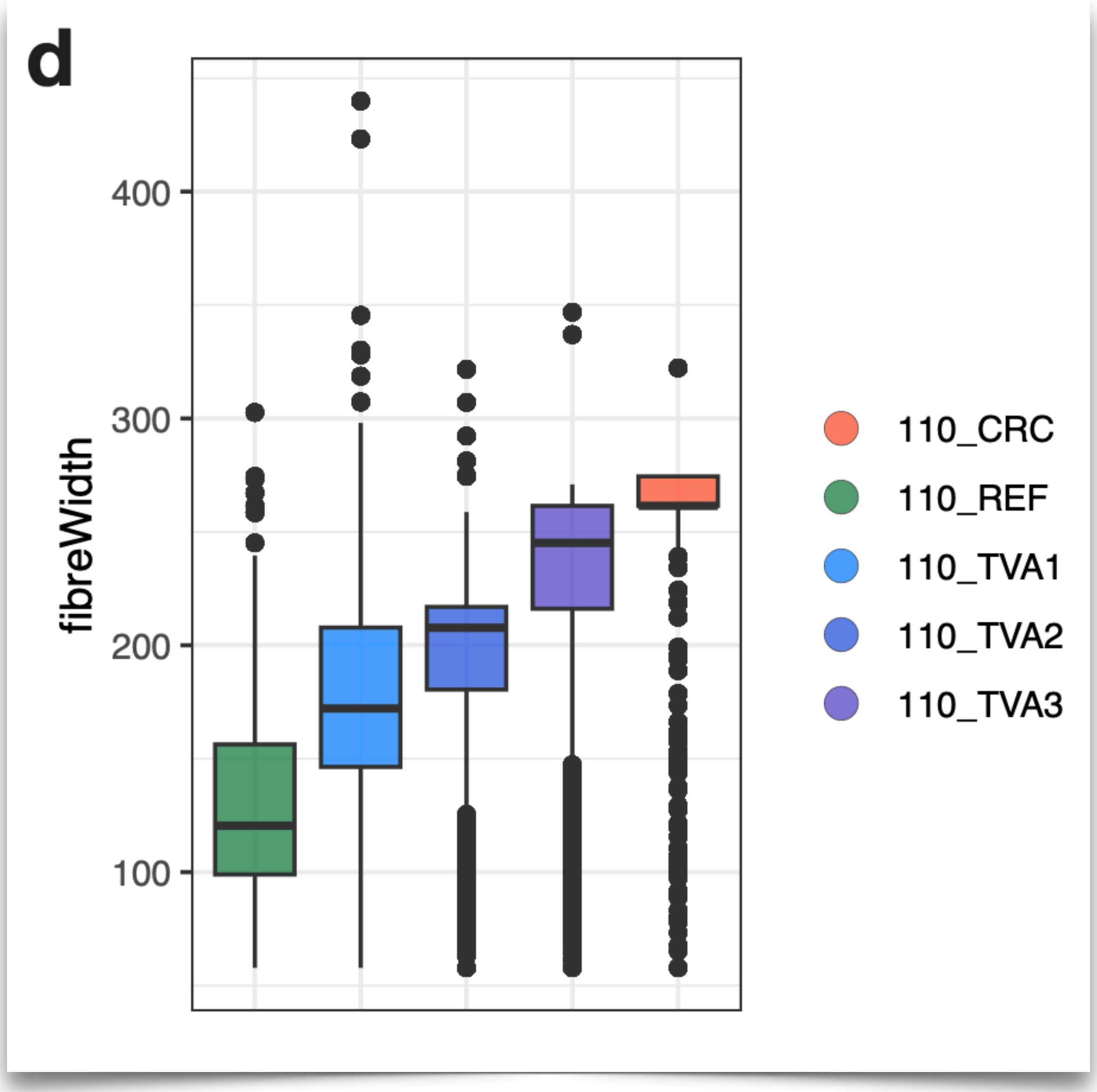
Preprint available.

Variation among spatial structures (epithelial example)

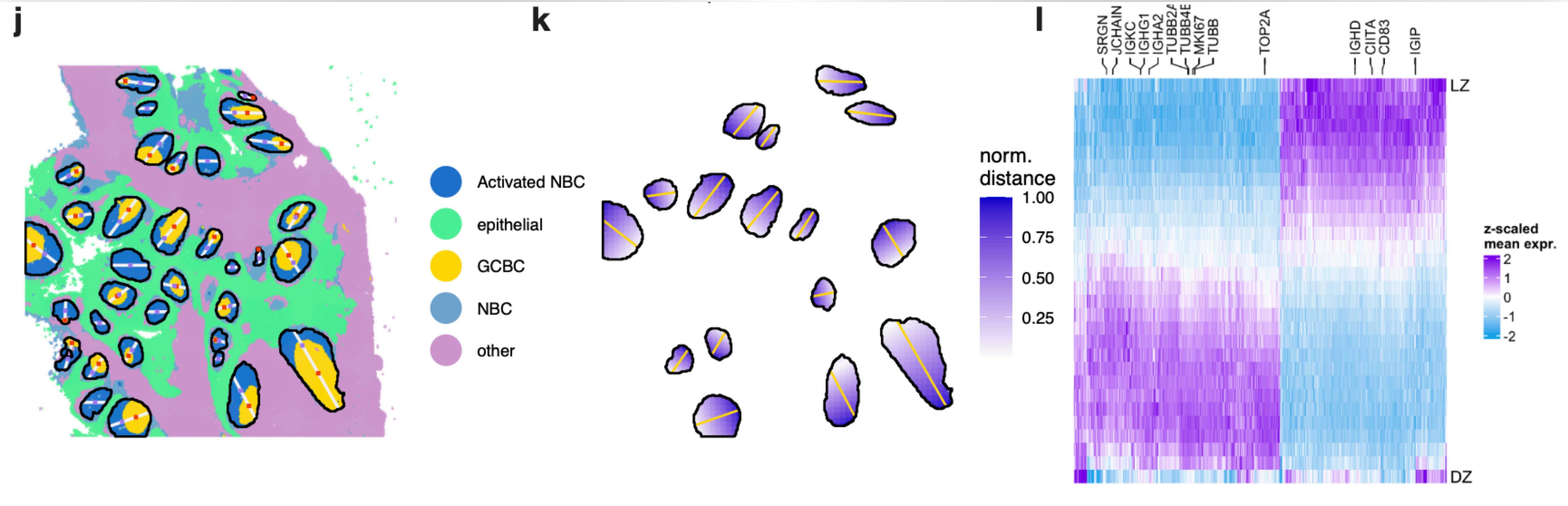


Variation among spatial structures (geometric quantifications)

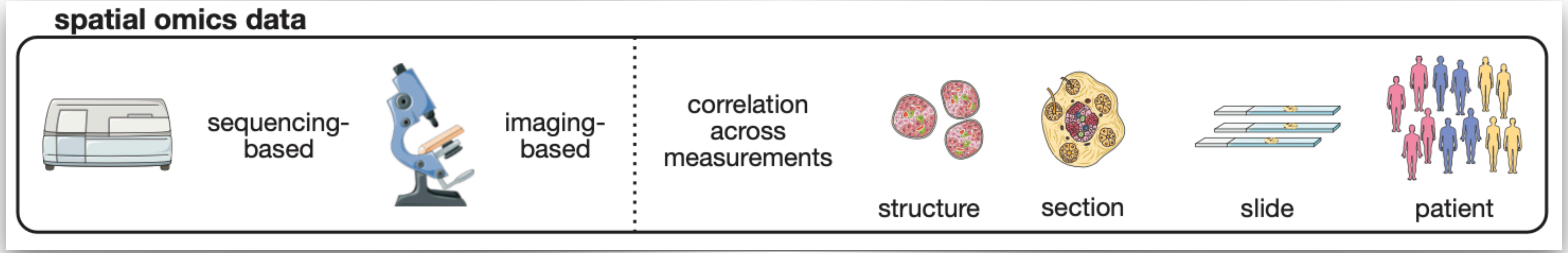




Structures —> Reference axis —> Expression gradients



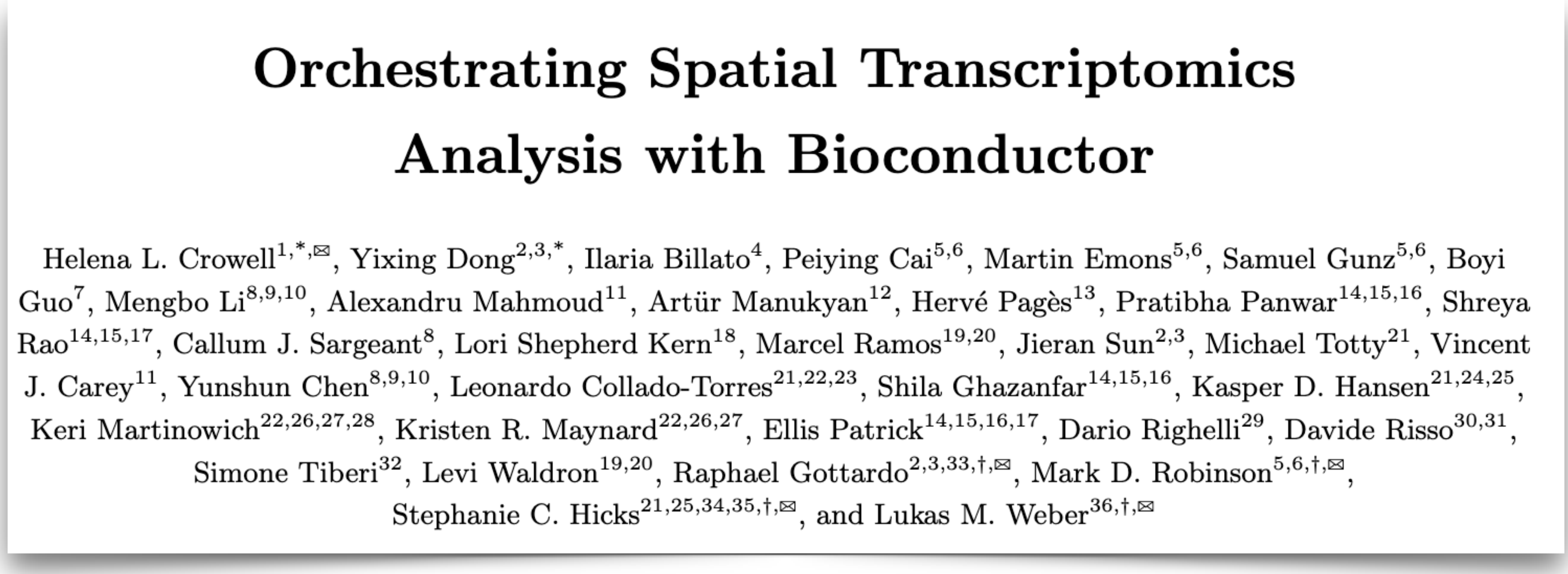
Modeling requires accounting for repeated measurements



Potentially i) multiple structures per tissue slice; ii) multiple slices per patient; iii) replication —> **multiple levels of variability** —> mixed models generally most appropriate

Orchestrating Spatial Transcriptomics Analysis with Bioconductor

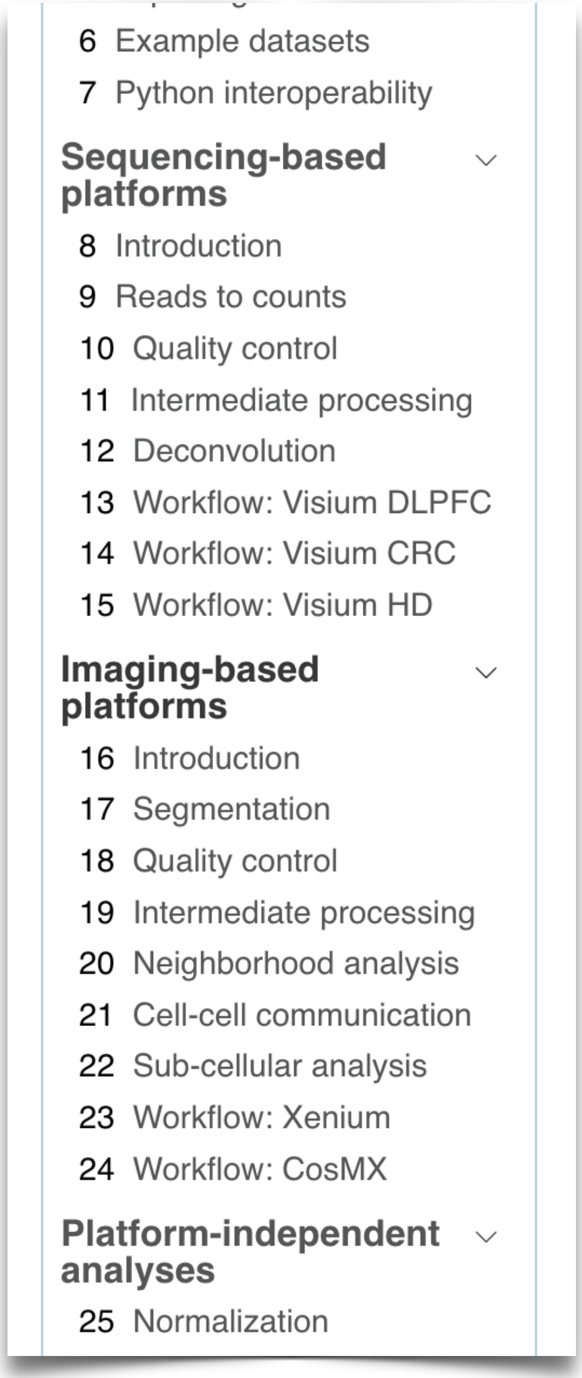
- https://bioconductor.org/books/OSTA



Orchestrating Spatial Transcriptomics Analysis with Bioconductor

Helena L. Crowell[1,*,✉], Yixing Dong[2,3,*], Ilaria Billato[4], Peiying Cai[5,6], Martin Emons[5,6], Samuel Gunz[5,6], Boyi Guo[7], Mengbo Li[8,9,10], Alexandru Mahmoud[11], Artür Manukyan[12], Hervé Pagès[13], Pratibha Panwar[14,15,16], Shreya Rao[14,15,17], Callum J. Sargeant[8], Lori Shepherd Kern[18], Marcel Ramos[19,20], Jieran Sun[2,3], Michael Totty[21], Vincent J. Carey[11], Yunshun Chen[8,9,10], Leonardo Collado-Torres[21,22,23], Shila Ghazanfar[14,15,16], Kasper D. Hansen[21,24,25], Keri Martinowich[22,26,27,28], Kristen R. Maynard[22,26,27], Ellis Patrick[14,15,16,17], Dario Righelli[29], Davide Risso[30,31], Simone Tiberi[32], Levi Waldron[19,20], Raphael Gottardo[2,3,33,†,✉], Mark D. Robinson[5,6,†,✉], Stephanie C. Hicks[21,25,34,35,†,✉], and Lukas M. Weber[36,†,✉]

Book is available. Preprint on bioRxiv.
(Successor of the OSCA book: https://bioconductor.org/books/OSCA/)



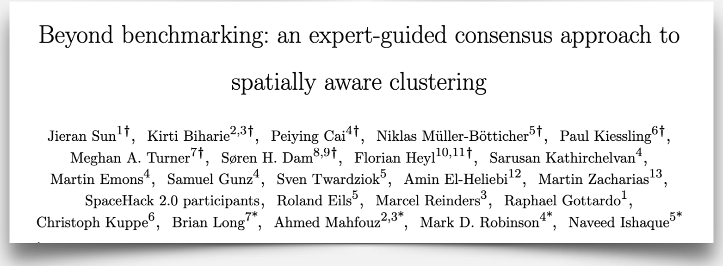
Sequencing-based platforms

Imaging-based platforms

Platform-independent analyses

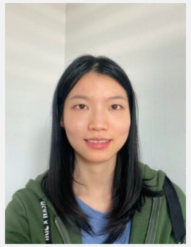
Meta-benchmark

Reported method performances are inconsistent across studies

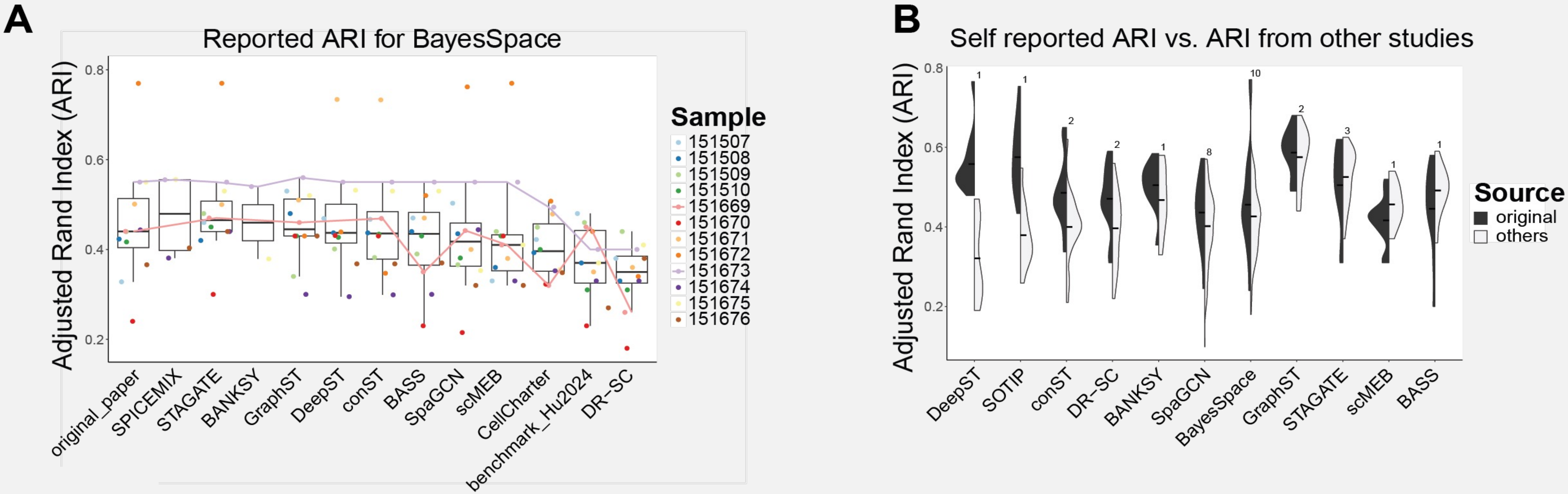


Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering

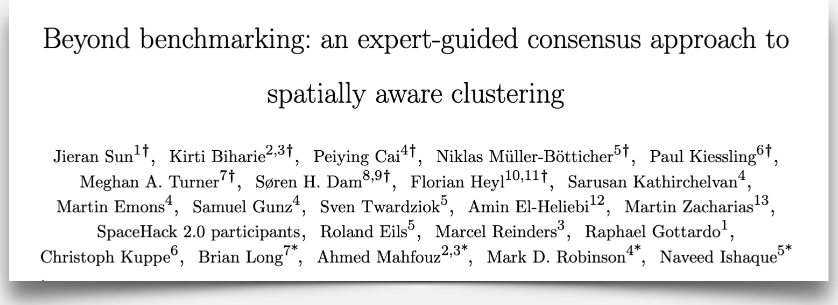
Jieran Sun[1†], Kirti Biharie[2,3†], Peiying Cai[4†], Niklas Müller-Bötticher[5†], Paul Kiessling[6†], Meghan A. Turner[7†], Søren H. Dam[8,9†], Florian Heyl[10,11†], Sarusan Kathirchelvan[4], Martin Emons[4], Samuel Gunz[4], Sven Twardziok[5], Amin El-Heliebi[12], Martin Zacharias[13], SpaceHack 2.0 participants, Roland Eils[5], Marcel Reinders[3], Raphael Gottardo[1], Christoph Kuppe[6], Brian Long[7*], Ahmed Mahfouz[2,3*], Mark D. Robinson[4*], Naveed Ishaque[5*]



Peiying Cai



Ensemble clustering

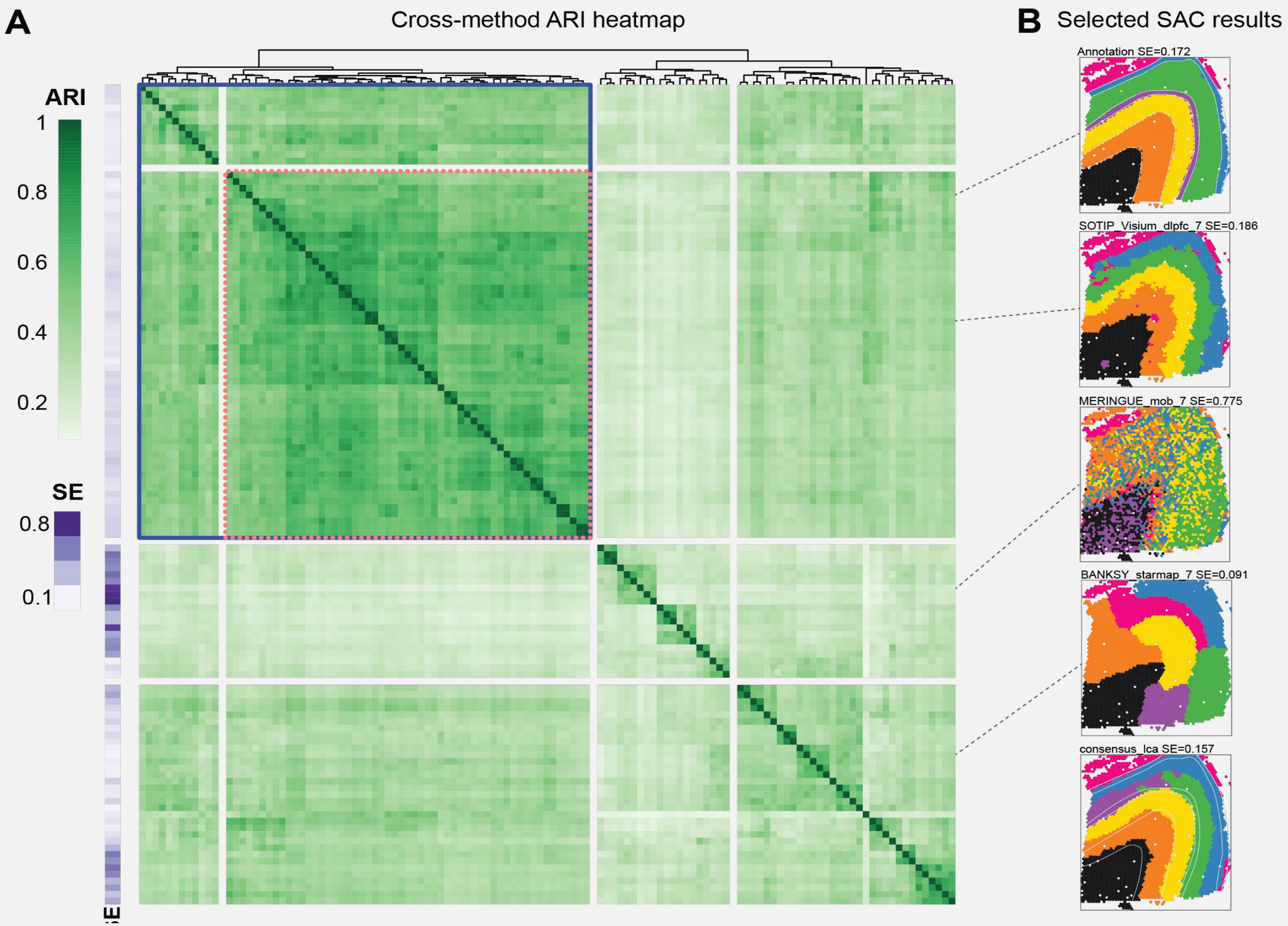


Beyond benchmarking: an expert-guided consensus approach to spatially aware clustering

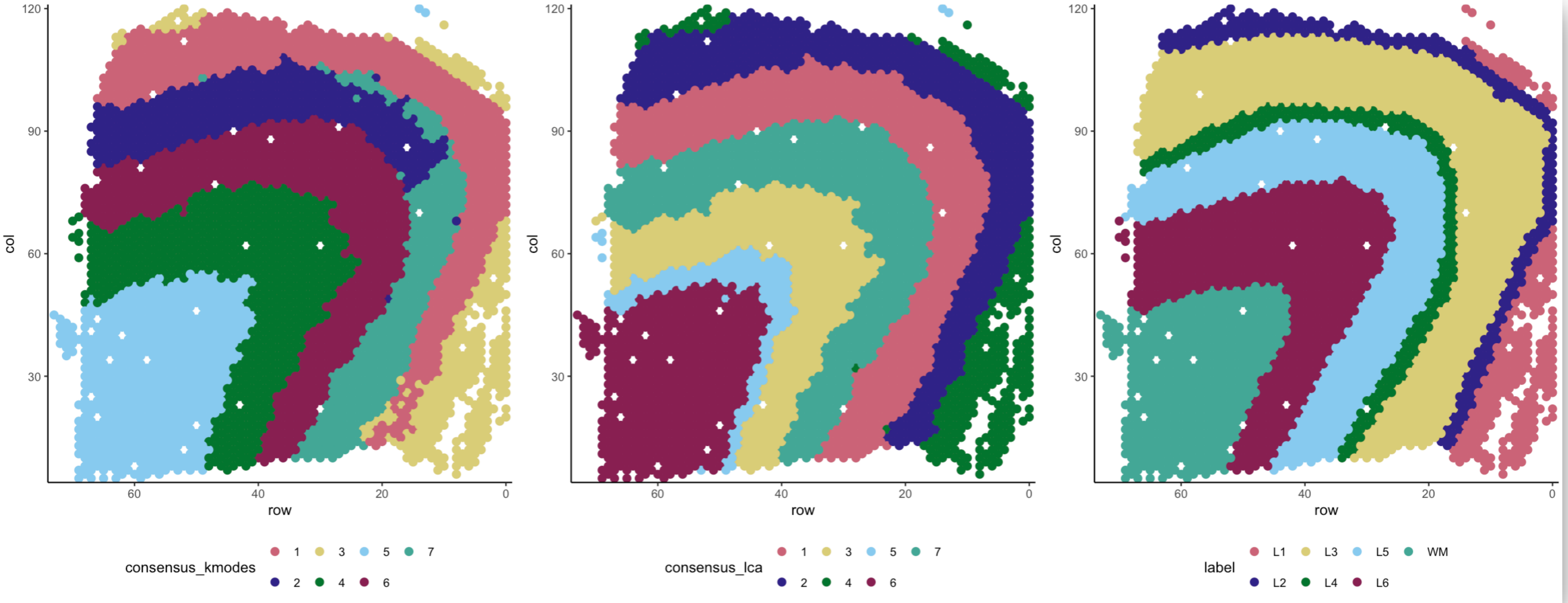
Jieran Sun[1†], Kirti Biharie[2,3†], Peiying Cai[4†], Niklas Müller-Bötticher[5†], Paul Kiessling[6†], Meghan A. Turner[7†], Søren H. Dam[8,9†], Florian Heyl[10,11†], Sarusan Kathirchelvan[4], Martin Emons[4], Samuel Gunz[4], Sven Twardziok[5], Amin El-Heliebi[12], Martin Zacharias[13], SpaceHack 2.0 participants, Roland Eils[5], Marcel Reinders[3], Raphael Gottardo[1], Christoph Kuppe[6], Brian Long[7*], Ahmed Mahfouz[2,3*], Mark D. Robinson[4*], Naveed Ishaque[5*]

Methods are often more similar to each other than to the ground truth.

Smoothness Entropy (Low = smooth)



Consensuses

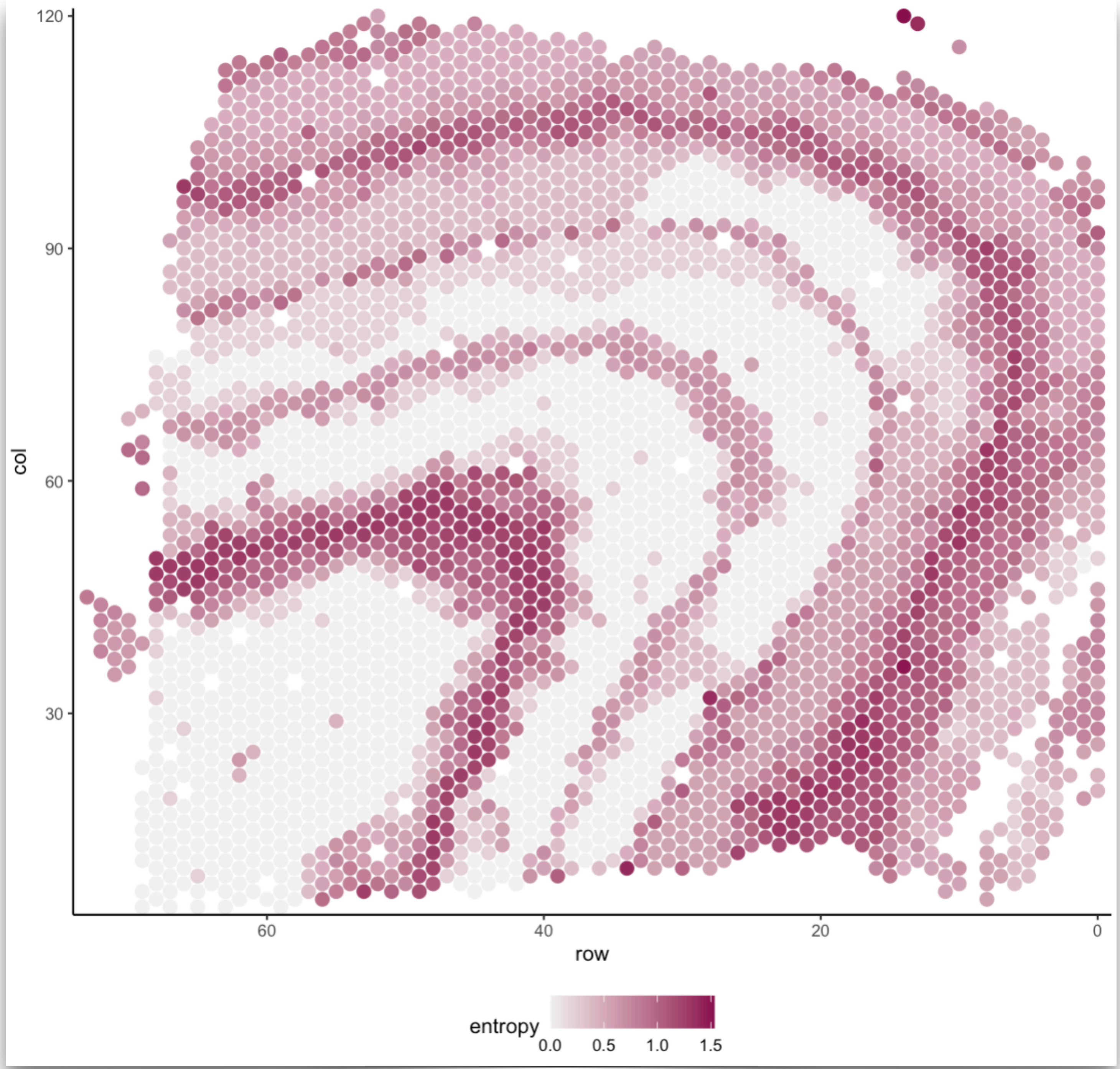


“ground truth”

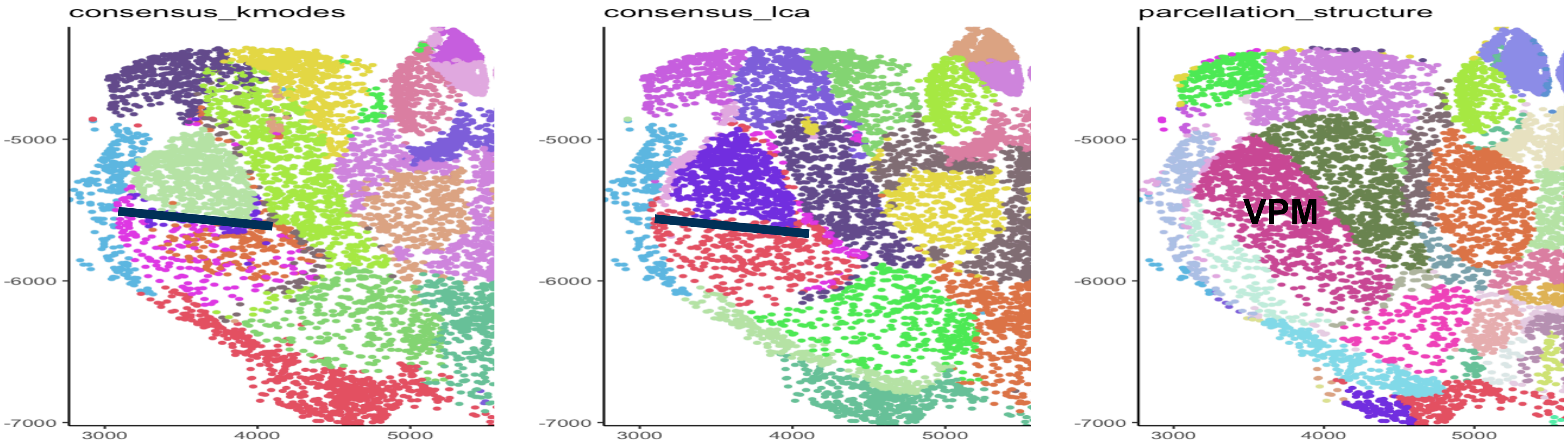
Entropy #2: Understanding spot-level uncertainty (across methods)

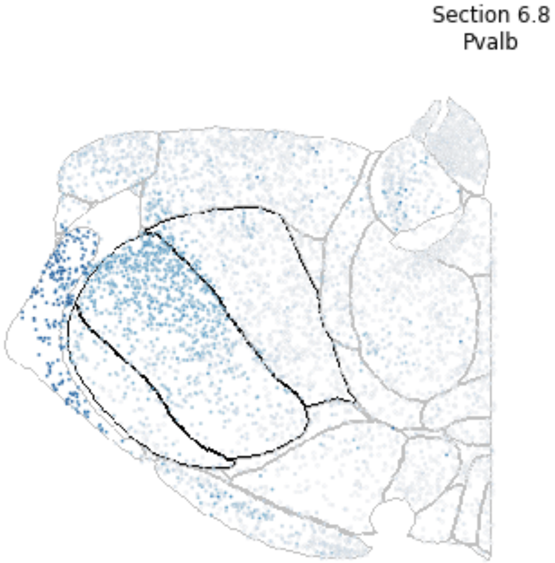
Entropy in the sense of <u>how stable across algorithms</u>

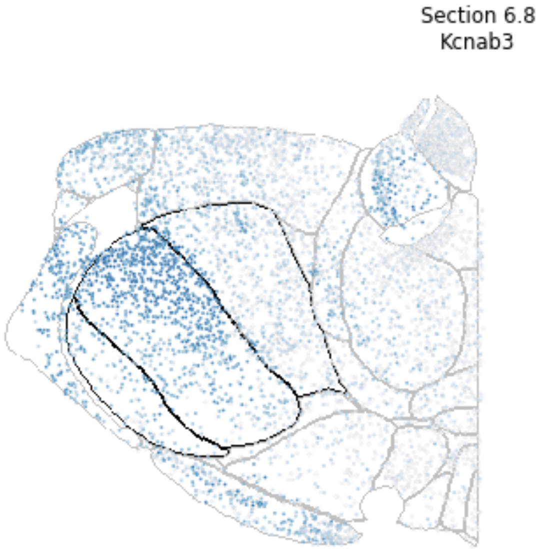
(align the spot-wise cluster labels across methods, entropy across label proportions)

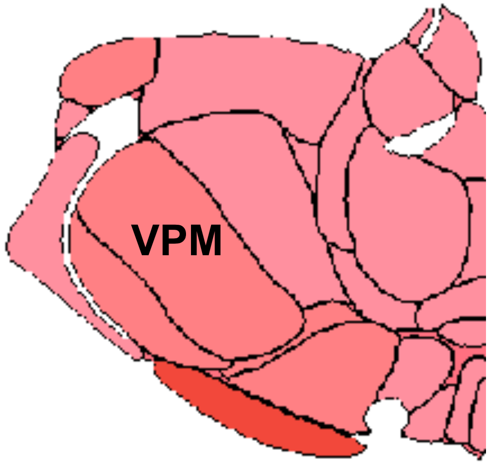


VPM









Slide from
Meghan Turner

Concluding remarks

- You are collecting/analyzing spatial data: what **spatial features** do you want to quantify?
- A few places where (classical) spatial statistics might be useful; data determines: point patterns versus lattice
- Functional data analysis, multi-cellular structure-based analyses, caveats of benchmarking