(Spatial) Statistics for spatial omics data

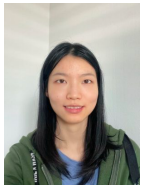
Mark D. Robinson
Statistical Bioinformatics Group, DMLS@UZH+SIB

@markrobinsonca
https://robinsonlabuzh.github.io/


Samuel


Martin


Helena


Peiying


Simone

First name	Last name	Program_time
Ana	Conesa	Monday early morning
Matt	Ritchie	Monday late morning
Jessica Jingyi	Li	Monday early afternoon
Yue	Cao	Tuesday early morning
Hongkai	Ji	Tuesday late morning
Gregory	Grant	Tuesday early afternoon
Salvador	Capella-Gutierrez	Wednesday early morning
Malte	Lücken	Wednesday early morning
Izaskun	Mallona	Wednesday early morning
Alice	McHardy	Thursday early morning
Moritz	Herrmann	Thursday late morning

Benchmarking microbial differential abundance analysis methods by employing only real data

Benchmarking Single-Cell Multiomics Integration to Enhance Cellular and Subpopulation Resolution

Invited

Leveraging Multi-Omics Data for the Evaluation of Kinase Activity Inference and Network Contextualisation

Benchmarking cell segmentation methods for fluorescent images in spatial omics

Benchmarking pretrained language models for modeling antigen-specific TCR-CDR3 regions

11 keynotes, ~20 contributed talks, ~25 posters

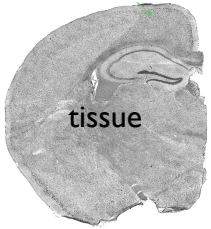
From bulk to single-cell RNA-seq to imaging- & sequencing-based spatially resolved transcriptomics

Slide from Helena Crowell

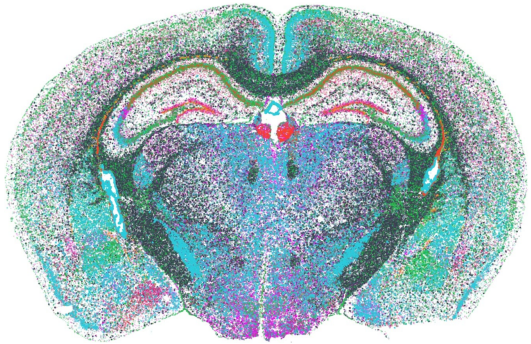
bulk

single-cell

spatial

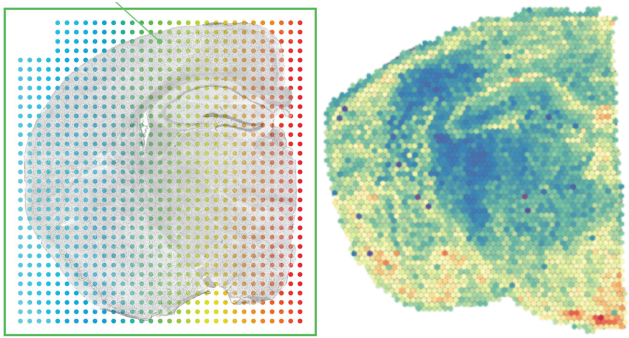


imaging-based



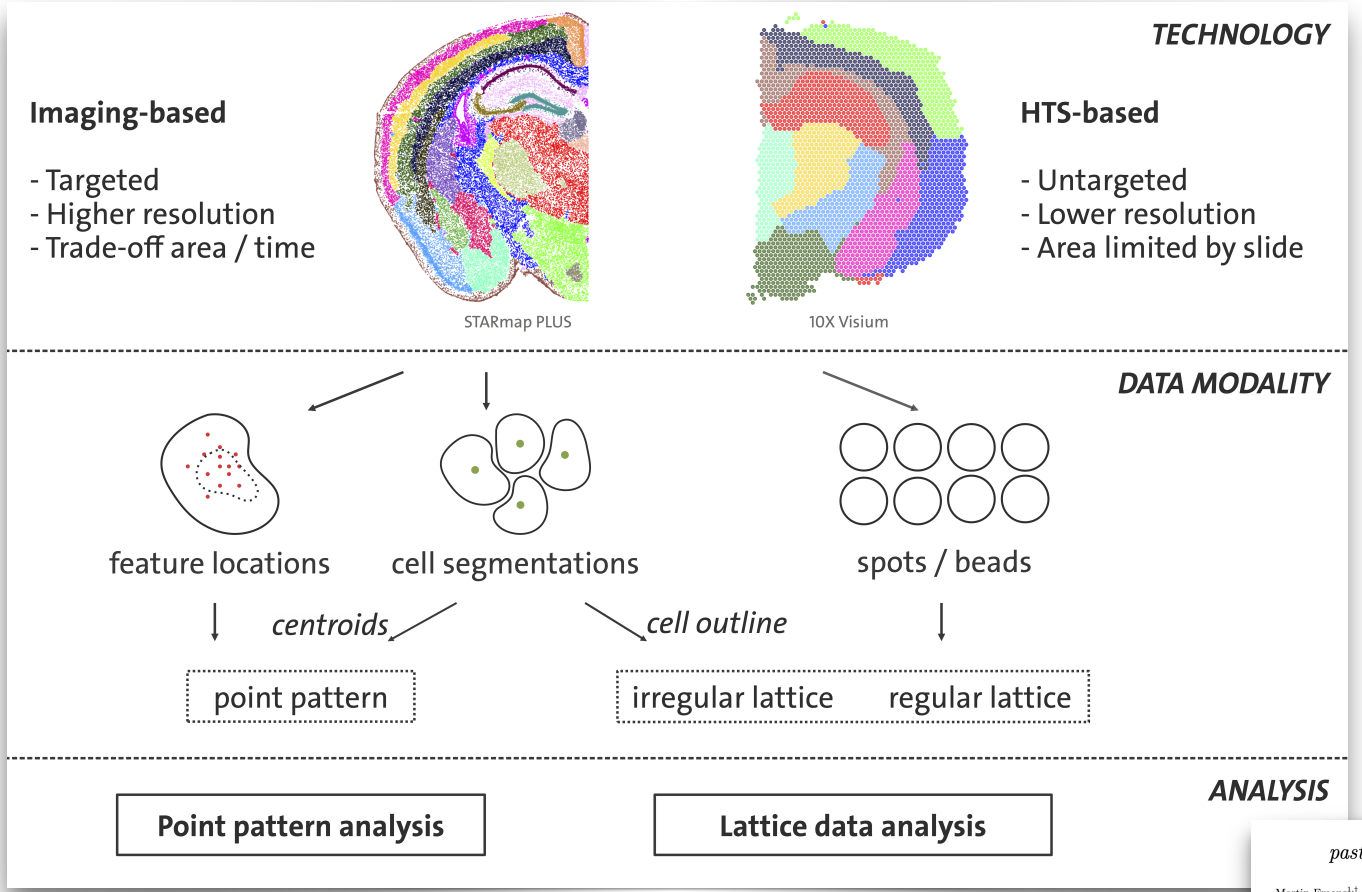
- molecule-level data
- targeted panel (100s-1000s of features)
- single-cell resolution requires segmentation

sequencing-based



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

Main concept: different data representations



Samuel

Martin

pasta: Pattern Analysis for Spatial Omics Data

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

What (statistical) analyses are done on spatial omics datasets?

- ~~Deconvolution~~
- Finding spatially-variable “features”
- Point patterns
- Lattice data
- Cell-cell communication (simple) —> co-localization

Tobler’s “first law of geography”: everything is related to everything else, but near things are more related than distant things.

Finding spatially-variable genes: SpatialDE

- 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: identification of spatially variable genes

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

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

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

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:

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

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

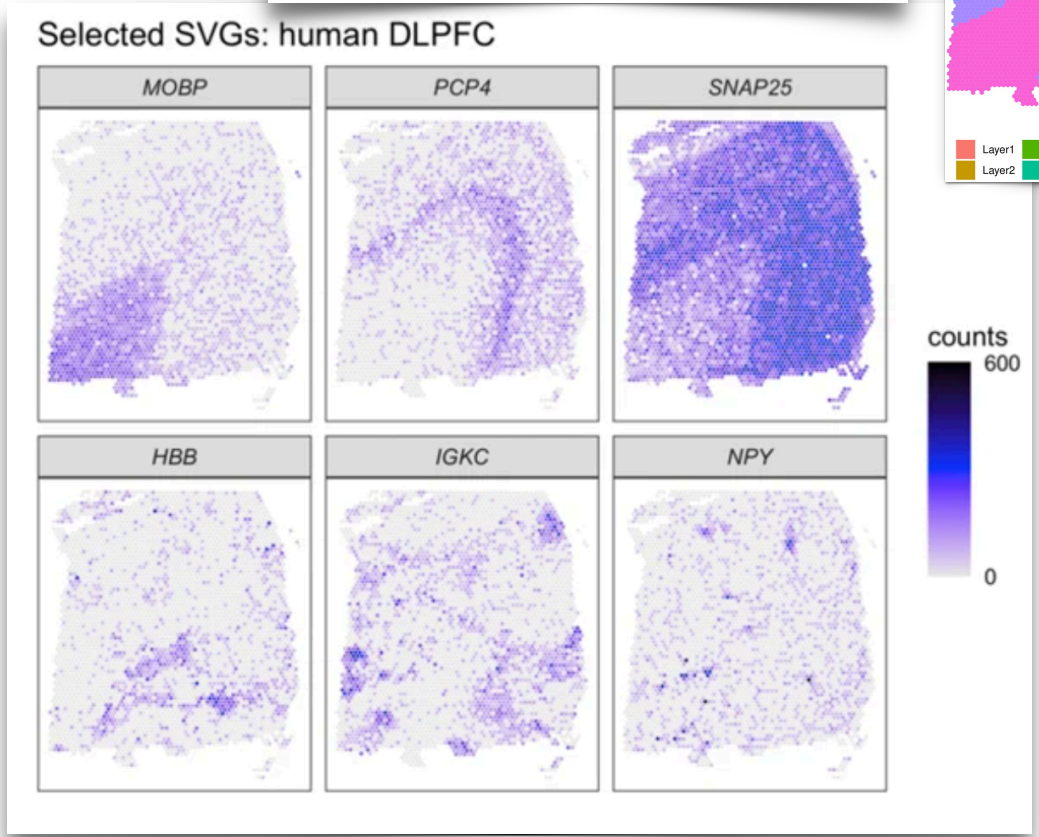
nnSVG for the scalable identification of spatially variable genes using nearest-neighbor Gaussian processes

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 genes

- different types (senses?) of spatially variable genes

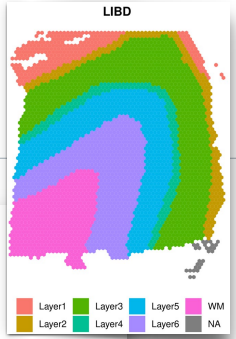


Article
https://doi.org/10.1038/s41467-023-39748-z

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

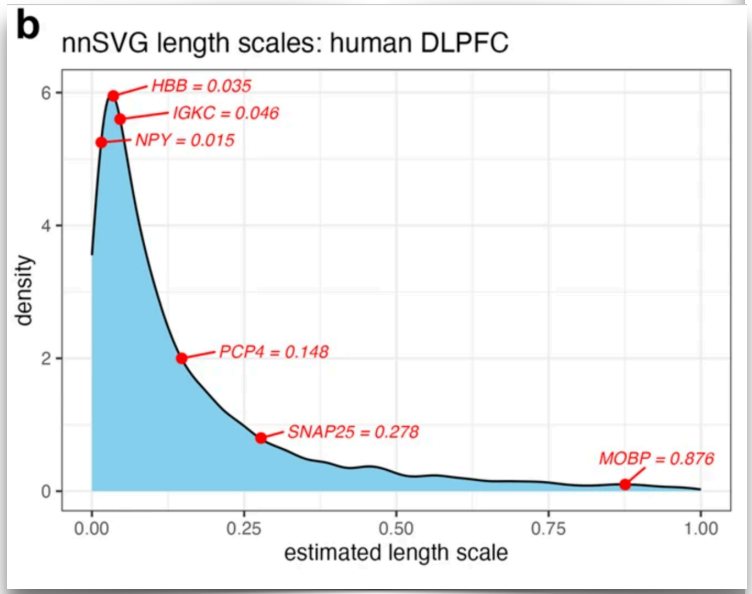
Received: 15 June 2022
Accepted: 23 June 2023

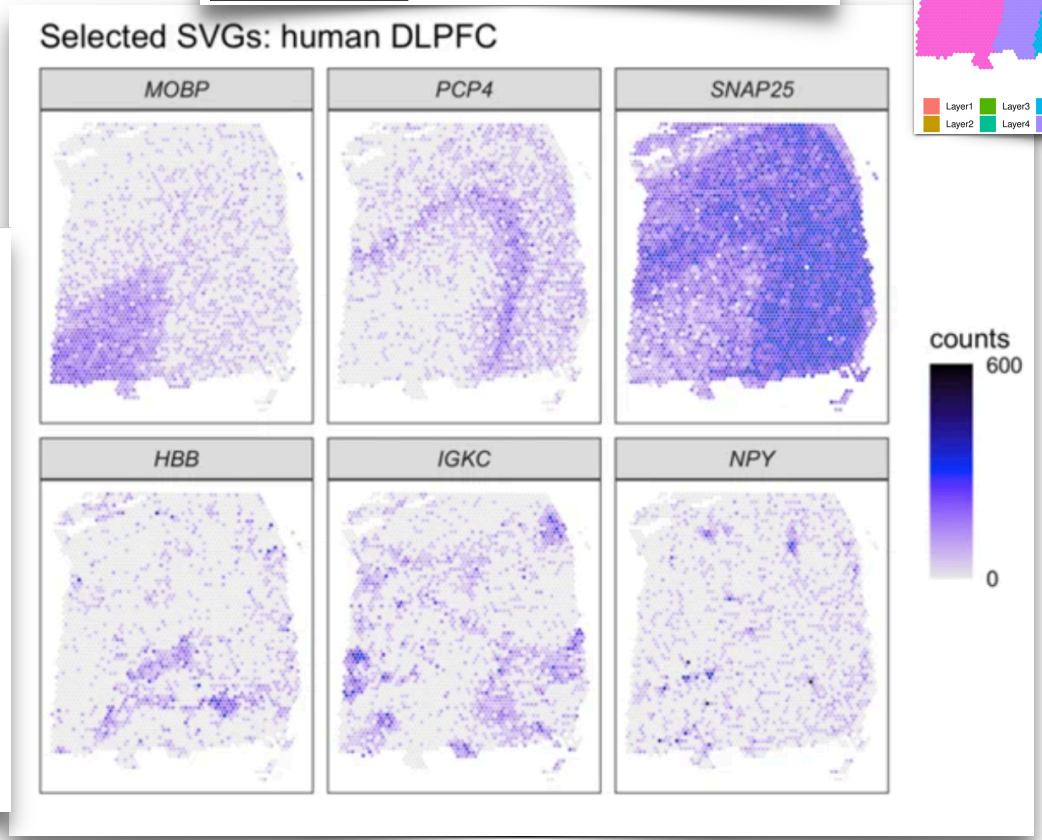
Lukas M. Weber[1], Arkajyoti Saha[2], Abhirup Datta[1], Kasper D. Hansen[1] & Stephanie C. Hicks[1]



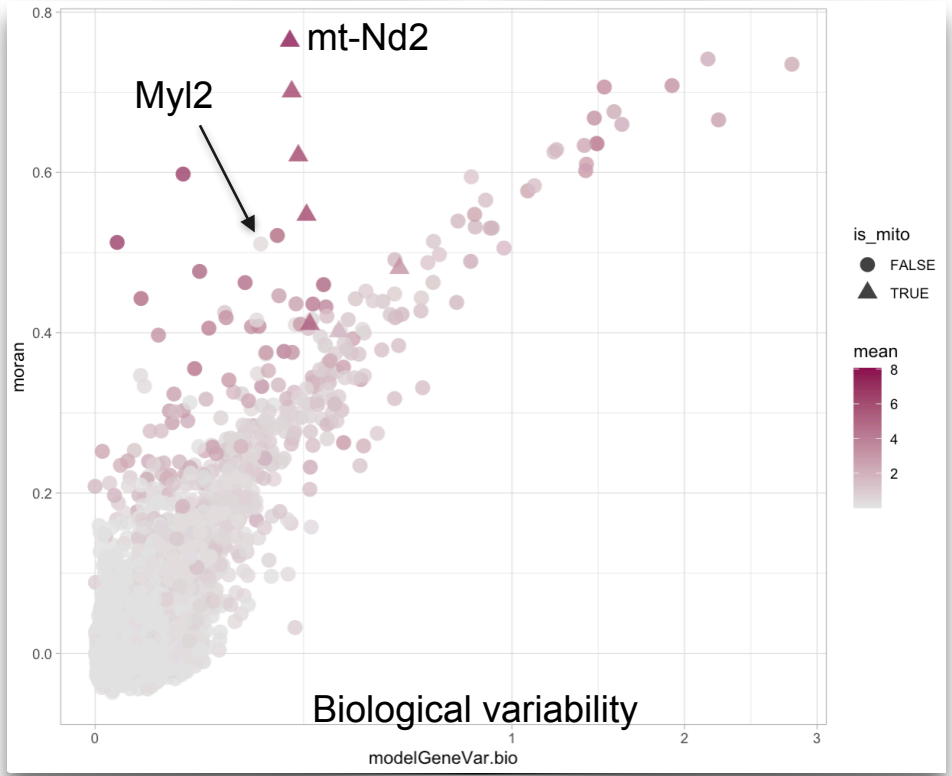
Spatially variable genes

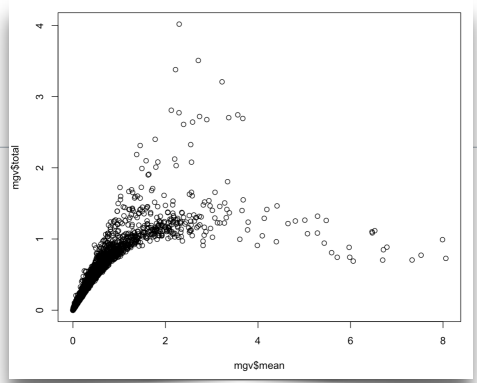
$$C_{ij}(\boldsymbol{\theta}) = \sigma^2 \exp\left(\frac{-||\mathbf{s_i} - \mathbf{s_j}||}{l}\right)$$

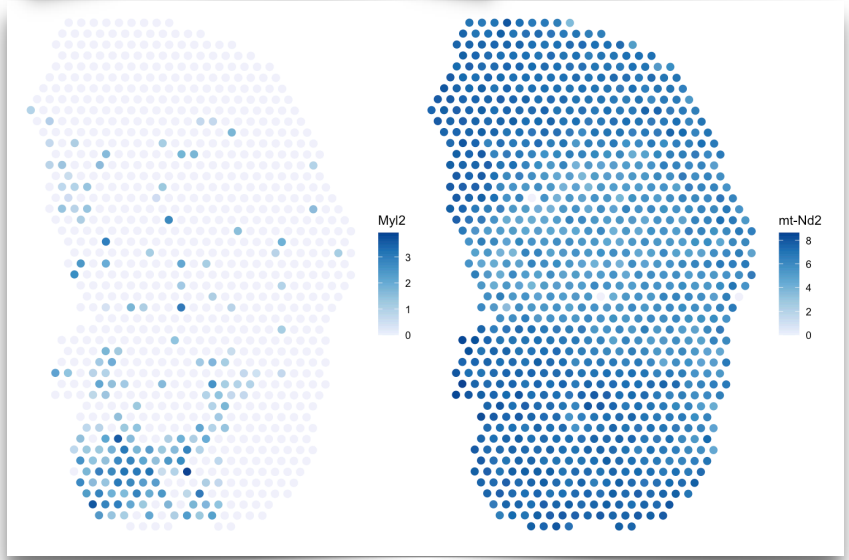




Spatially variable versus highly variable



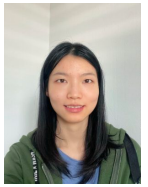




(More mathematical details on Moran's I below)

Alternatively, spatially variable features = DE between domains

Simone Tiberi



Peiying Cai

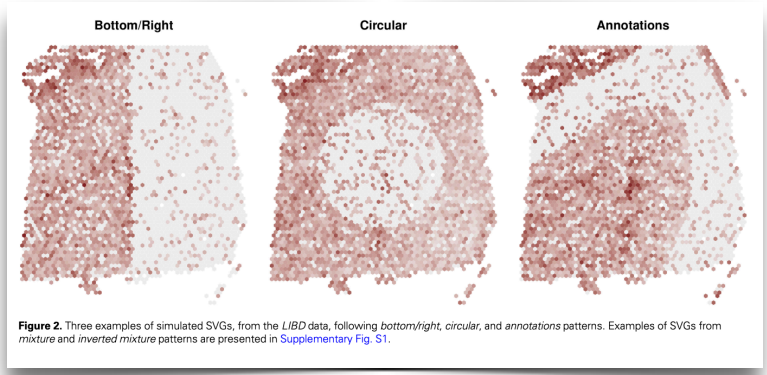
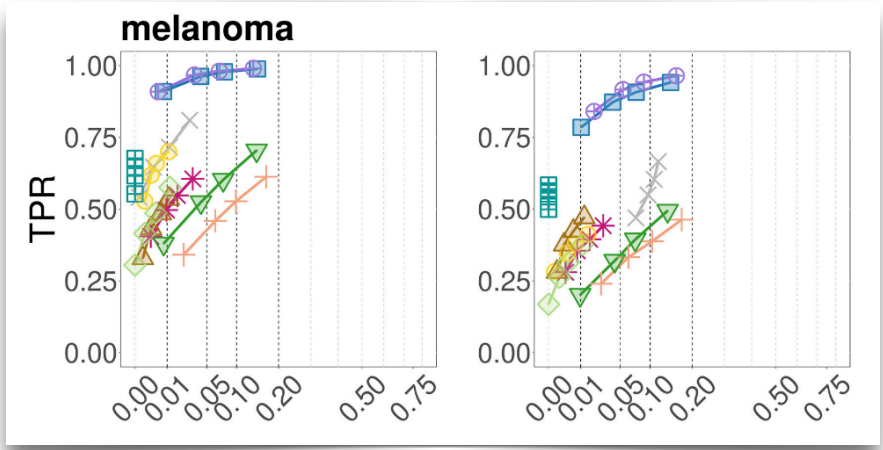


Figure 2. Three examples of simulated SVGs, from the *LIBD* data, following *bottom/right*, *circular*, and *annotations* patterns. Examples of SVGs from *mixture* and *inverted mixture* patterns are presented in Supplementary Fig. S1.



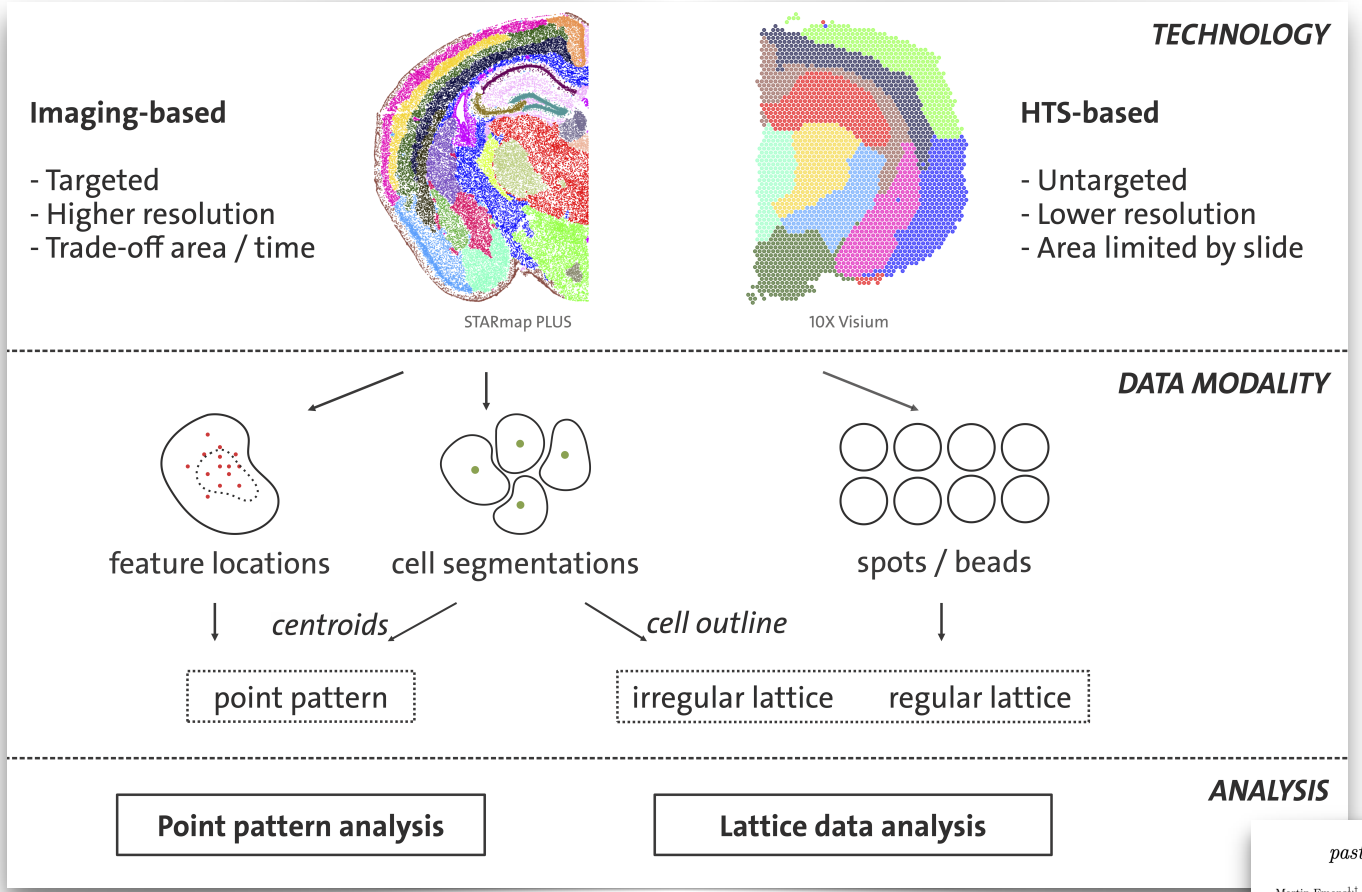
To find spatially variable genes (SVGs); spatial clustering + classical statistical method works quite well

JOURNAL ARTICLE

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

Peiying Cai, Mark D Robinson, Simone Tiberi

Main concept: different data representations



Samuel

Martin

pasta: Pattern Analysis for Spatial Omics Data

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

Fundamentals of Spatial Statistics (the subset that is useful for spatial omics data)

- **Point patterns**
  - **Definitions: intensity, homogeneity, dependence**
  - **Multi-type point patterns**
  - **Marked point processes**
  - **Statistical summaries**
- Lattice data
  - Definitions: lattice, regularity, neighbourhood matrix
  - Univariate global spatial autocorrelation
  - Univariate local spatial autocorrelation
  - Multivariate options

What are point patterns?

- data with spatial locations of objects (cells .. maybe transcripts) are generated by a stochastic process
- lattice data is not a point pattern!
- point pattern analysis gives tools for both discrete (cell types) and continuous (gene expression) “marks”
- We mostly focus here on discrete categories, i.e., patterns within and between cell types

- "a realisation of a spatial point process effectively assumes that the locations of points are not fixed, and that the point pattern is the response or observation of interest."

Scenario 14.1. *A weather map for Europe displays a symbol for each major city indicating the expected type of weather (e.g., sunny, cloudy, storms).*

Scenario 14.2. *An optical astronomy survey records the sky position and qualitative shape (elliptical, spiral, etc.) of each galaxy in a nearby region of space.*

Scenario 14.3. *Trees in an orchard are examined and their disease status (infected/not infected) is recorded. We are interested in the spatial characteristics of the disease, such as contagion between neighbouring trees.*

Some definitions

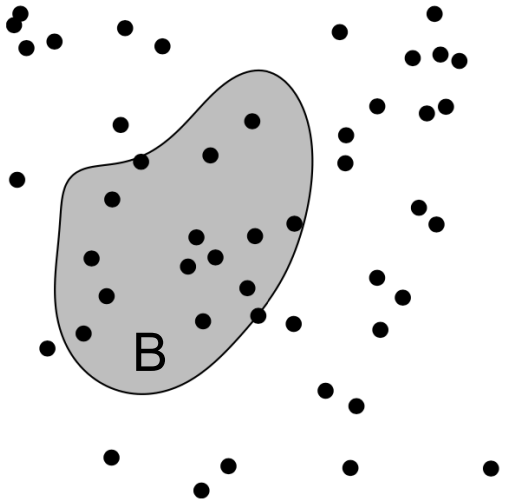
- notation: **X** is the point process; **x** is the (observed) point pattern
- lambda: intensity function

A point pattern is denoted by a bold lower case letter like **x**. It is a set

$$\mathbf{x} = \{x_1, x_2, \ldots, x_n\}$$

of points x_i in two-dimensional space $\mathbb{R}^2$. The number $n = n(\mathbf{x})$ of points in the pattern is not fixed in advance, and may be any finite nonnegative number *including zero*. In practice, the data points are obviously recorded in some order $x_1, \ldots, x_n$; but this ordering is artificial, and we treat the pattern **x** as an unordered set of points.

$$\mathbb{E}[n(\mathbf{X} \cap B)] = \int_B \lambda(u)\,\mathrm{d}u$$



Definitions

- **X** is the point process; **x** is the (observed) point pattern
- lambda: intensity function
- Complete spatial randomness (CSR) has two properties:

homogeneity: the points have no preference for any spatial location;

independence: information about the outcome in one region of space has no influence on the outcome in other regions.

- More specifically:

homogeneity: the number $n(\mathbf{X} \cap B)$ of random points falling in a test region B has mean value $\mathbb{E}n(\mathbf{X} \cap B) = \lambda |B|$;

independence: for test regions $B_1, B_2, \ldots, B_m$ which do not overlap, the counts $n(\mathbf{X} \cap B_1), \ldots, n(\mathbf{X} \cap B_m)$ are independent random variables;

Couple more definitions

– Inhomogeneity

The *inhomogeneous Poisson point process* with intensity function $\lambda(u)$ is defined by the following properties:

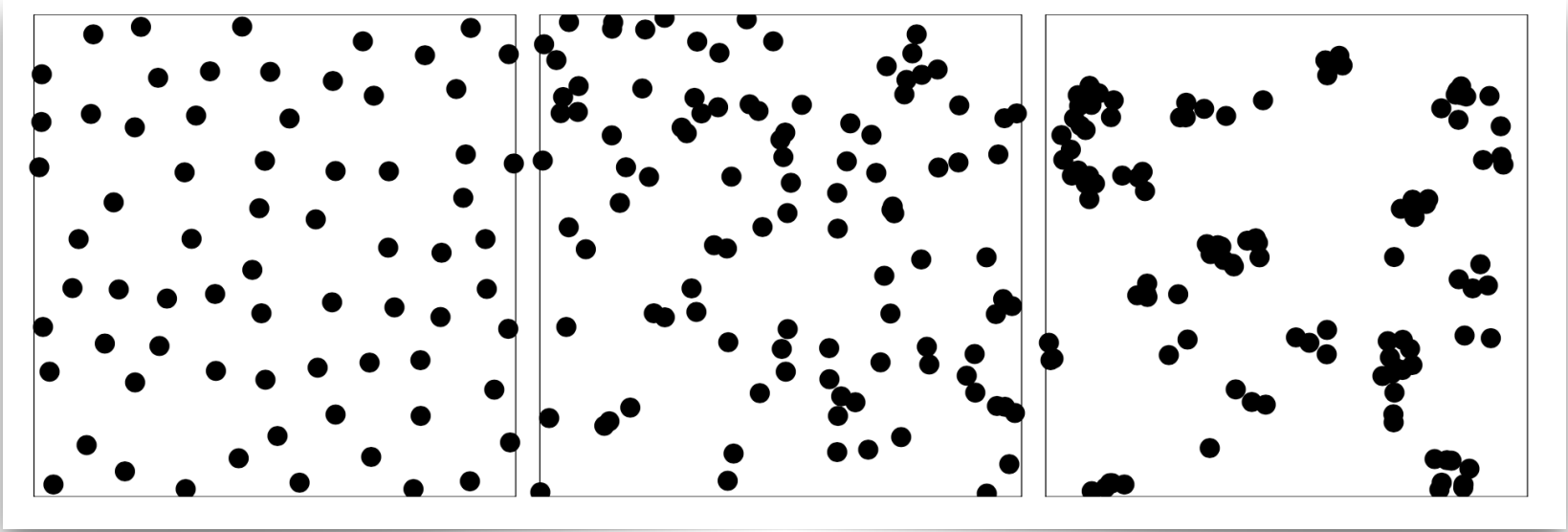
intensity function: the expected number of points falling in a region B is the integral $\mu = \int_B \lambda(u)\,du$ of the intensity function $\lambda(u)$ over the region B;

independence: if space is divided into non-overlapping regions, the random patterns inside these regions are independent of each other;

Poisson-distributed counts: the random number of points falling in a given region has a *Poisson* probability distribution;

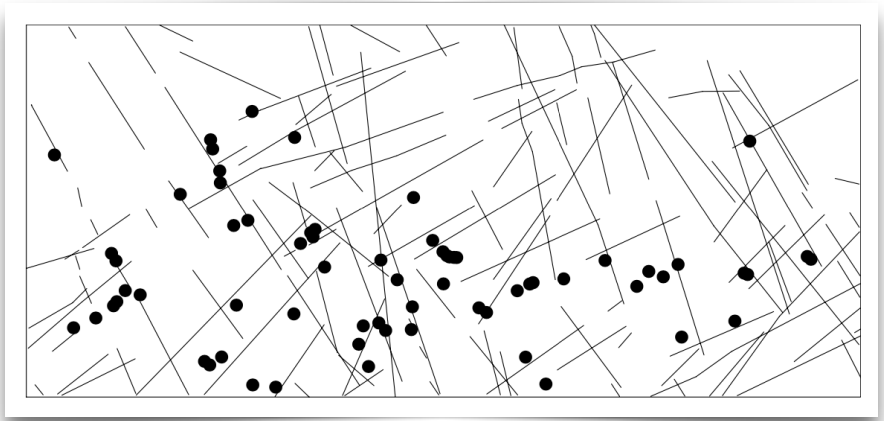
Properties of point patterns

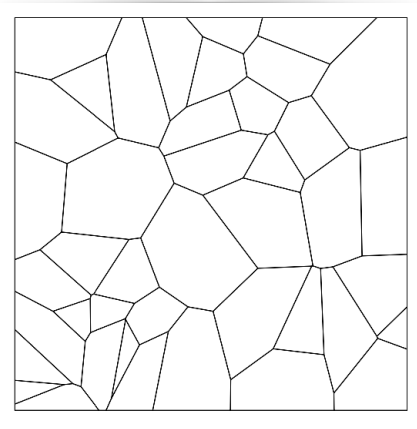
- Which of these is homogeneous?
- Which of these is completely spatially random (CSR)?
- Which of these is clustered?
- Which of these is not independent?



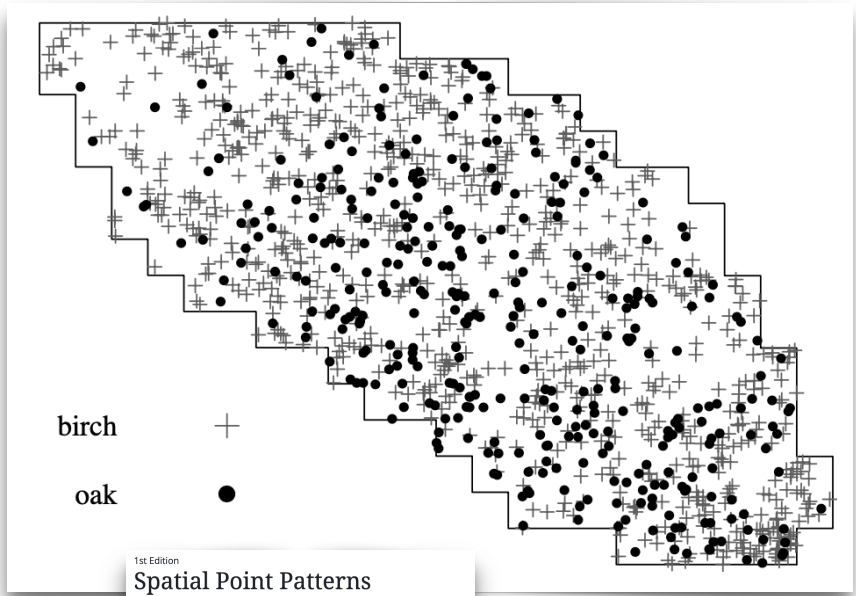
Extensions of point patterns

- (patterns can be regions/lines, not just points; can be in higher dimension (e.g., 3D); temporal component .. most of the methods I discuss have extensions; not discussed here)
- marks —> marked point process
- types —> multi-type point process (contrast with 'multivariate')
- covariates



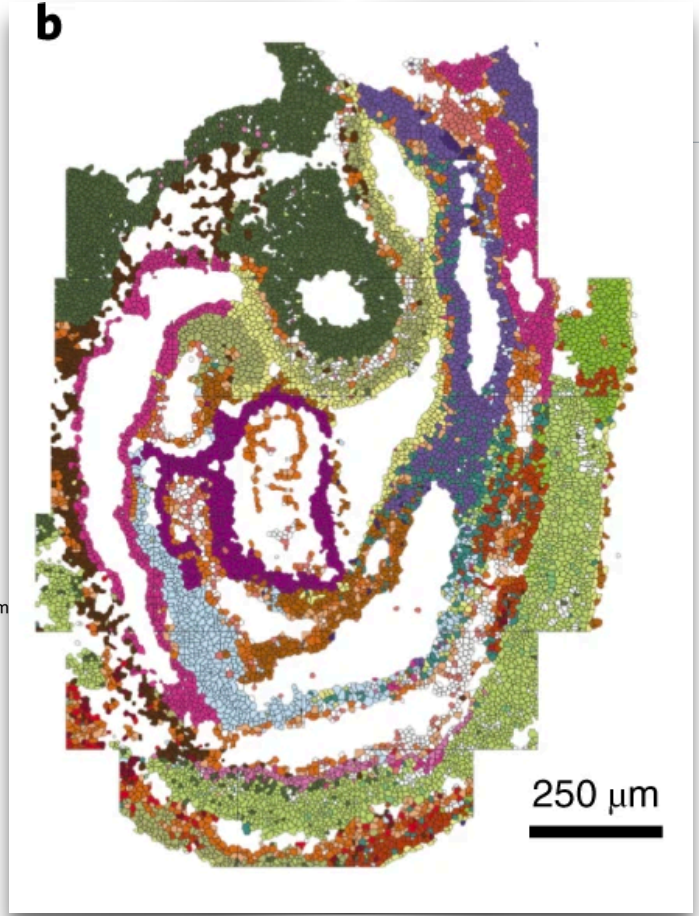


Multi-type point patterns



1st Edition
Spatial Point Patterns
Methodology and Applications with R
By Adrian Baddeley, Ege Rubak, Rolf Turner
Copyright 2016

- Allantois
- Anterior somitic tissues
- Blood progenitors
- Cardiomyocytes
- Caudal mesoderm
- Cranial mesoderm
- Definitive endoderm
- Dermomyotome
- Endothelium
- Erythroid
- ExE endoderm
- Forebrain/midbrain/hindbrain
- Gut tube
- Hematoendothelial progenitors
- Intermediate mesoderm
- Lateral plate mesoderm
- Mixed mesenchymal mesoderm
- Neural crest
- NMP
- Presomitic mesoderm
- Sclerotome
- Spinal cord
- Splanchnic mesoderm
- Surface ectoderm

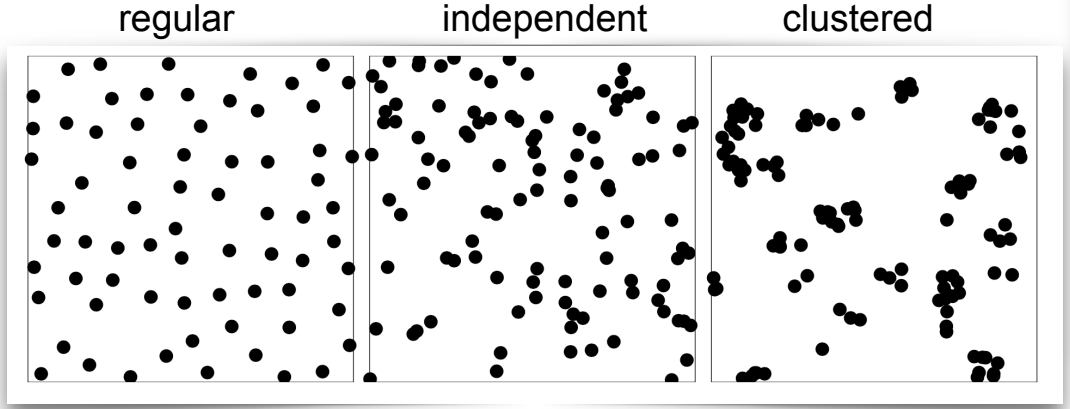


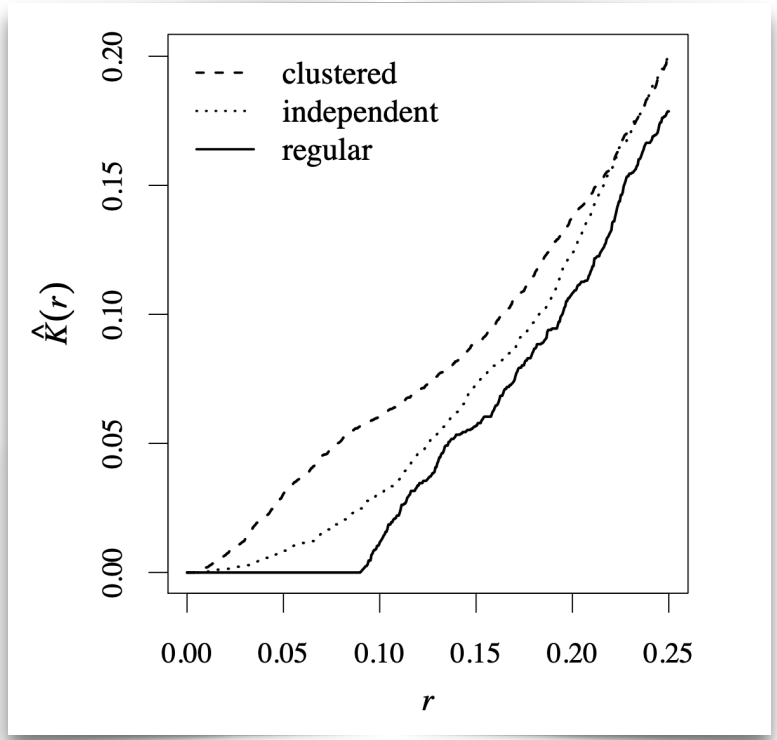
https://www.nature.com/articles/s41587-021-01006-2

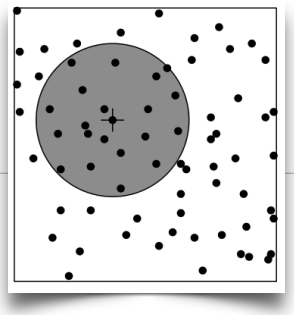
Subtlety here: multi-type point patterns versus multivariate point patterns

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*







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* $\mathbf{X}$.

What about correlation and intensity together?

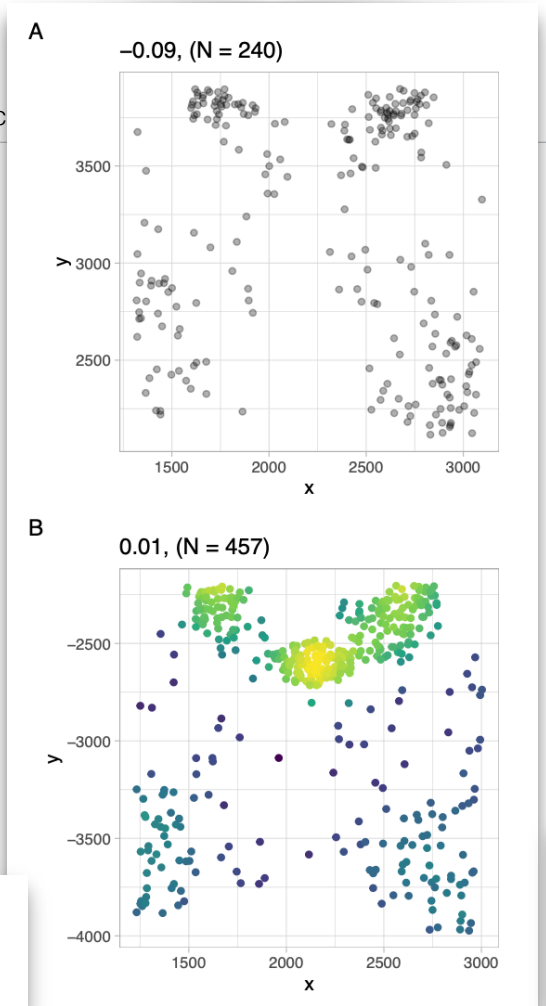
- **inhomogeneous** correlation functions
- edge correction

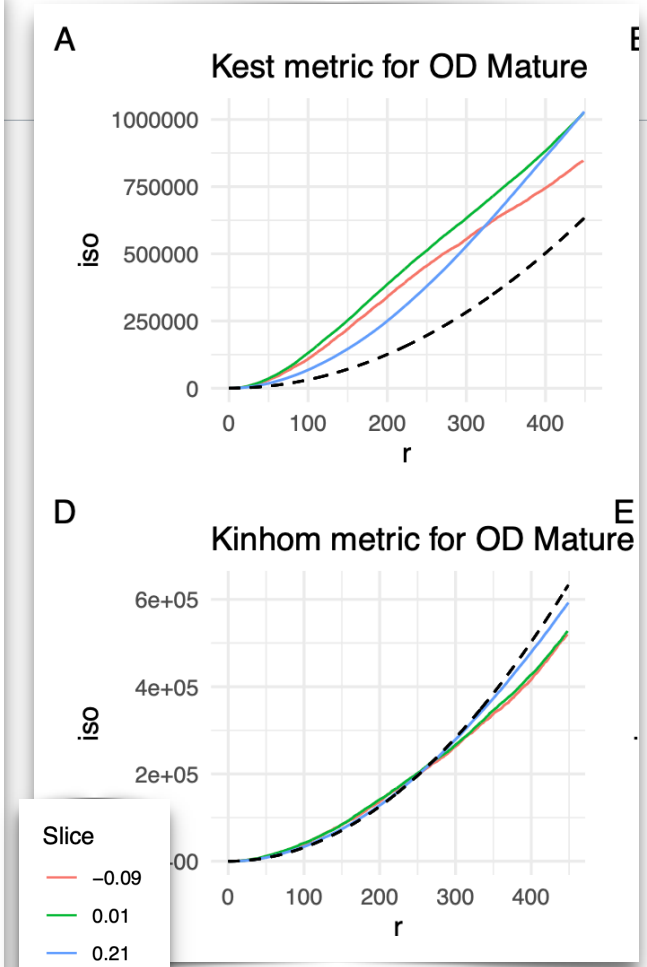
$$\widehat{K}_{inhom}(r) = \frac{1}{D^p|W|}\sum_i\sum_{j\neq i}\frac{\mathbf{1}\{||x_i - x_j|| \leq r\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)}e(x_i, x_j; r)$$

- n.b. confounding of correlation and intensity (next slide)

Confounding between clustering and intensity

- Whether you assume homogeneity or not (in the K-function calculation) can have a big impact on the estimated curves
- Are these cells clustered or have different intensity? Hard to tell.

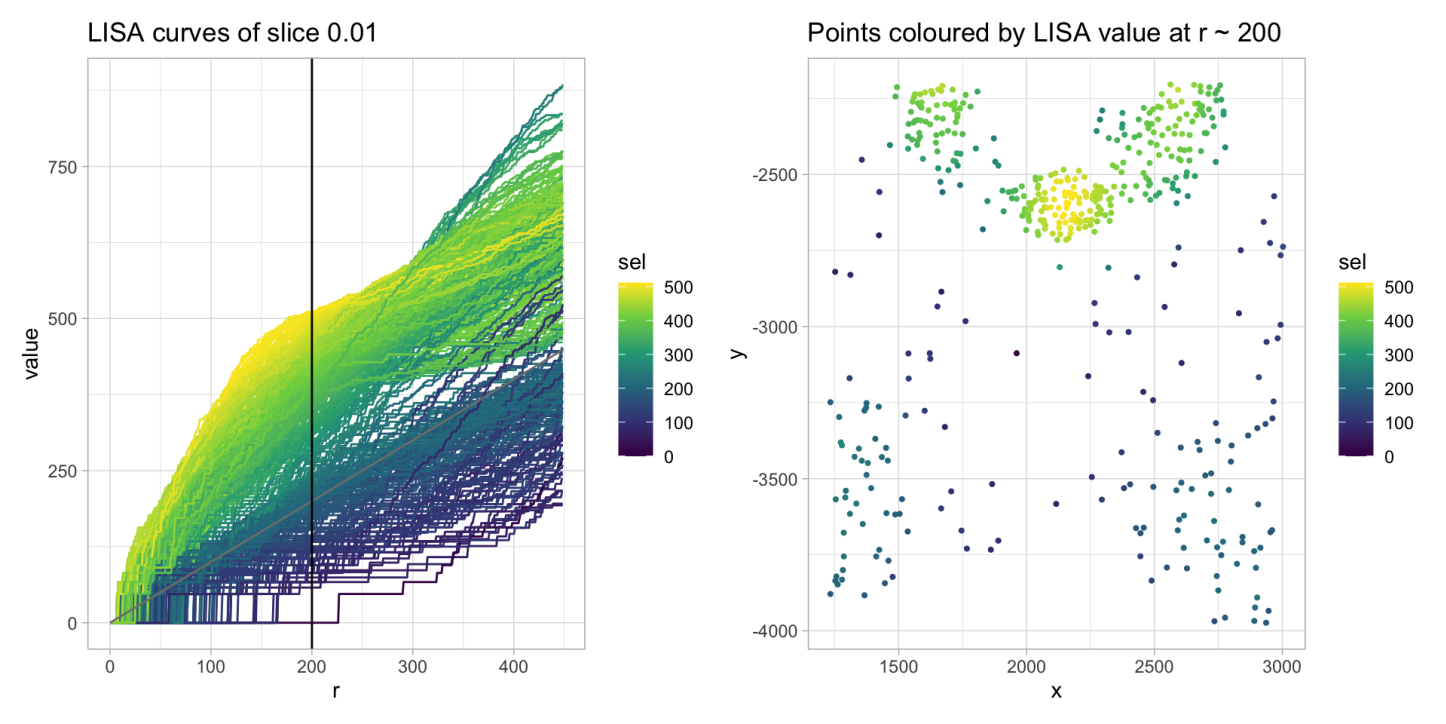




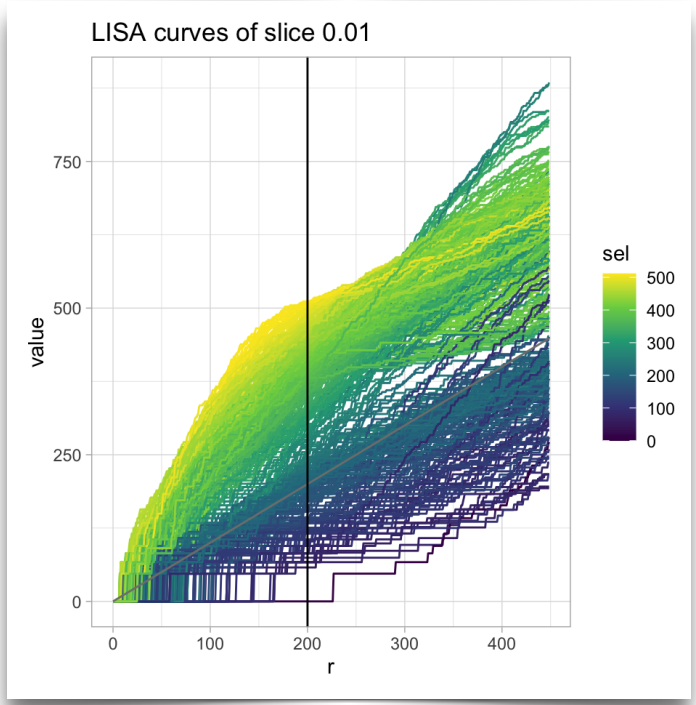
pasta: Pattern Analysis for Spatial Omics Data

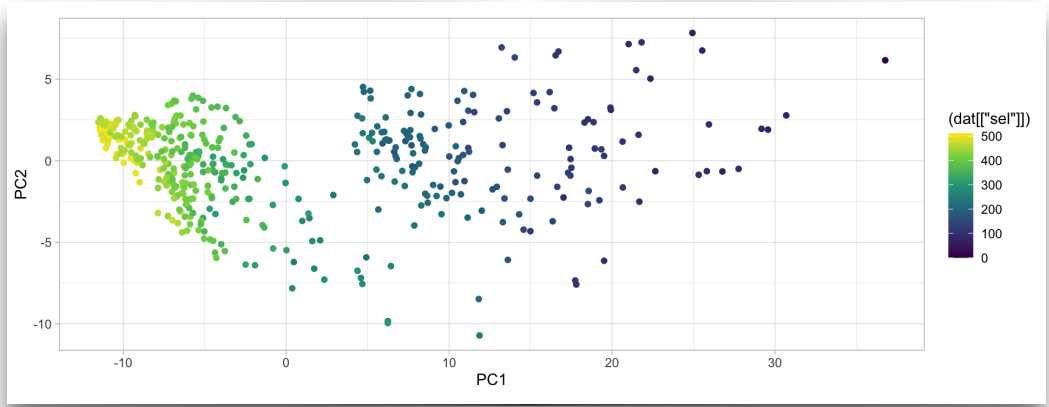
Martin Emons[1,†], Samuel Gunz[1,†], Helena L. Crowell[2], Izaskun Mallona[1], Reinhard Furrer[3], and Mark D. Robinson[1,*]

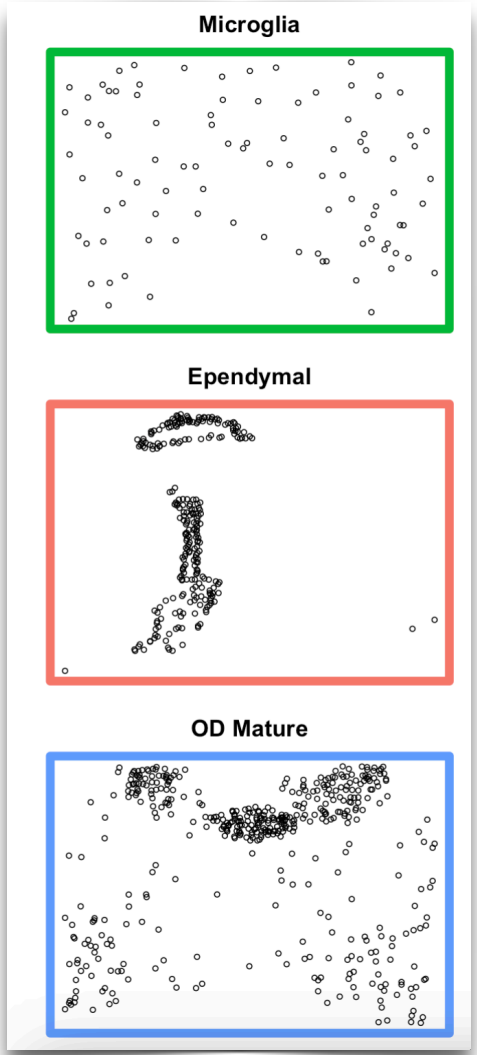
“Local” indicators of spatial association = LISA



Functions/Curves —> Functional PCA

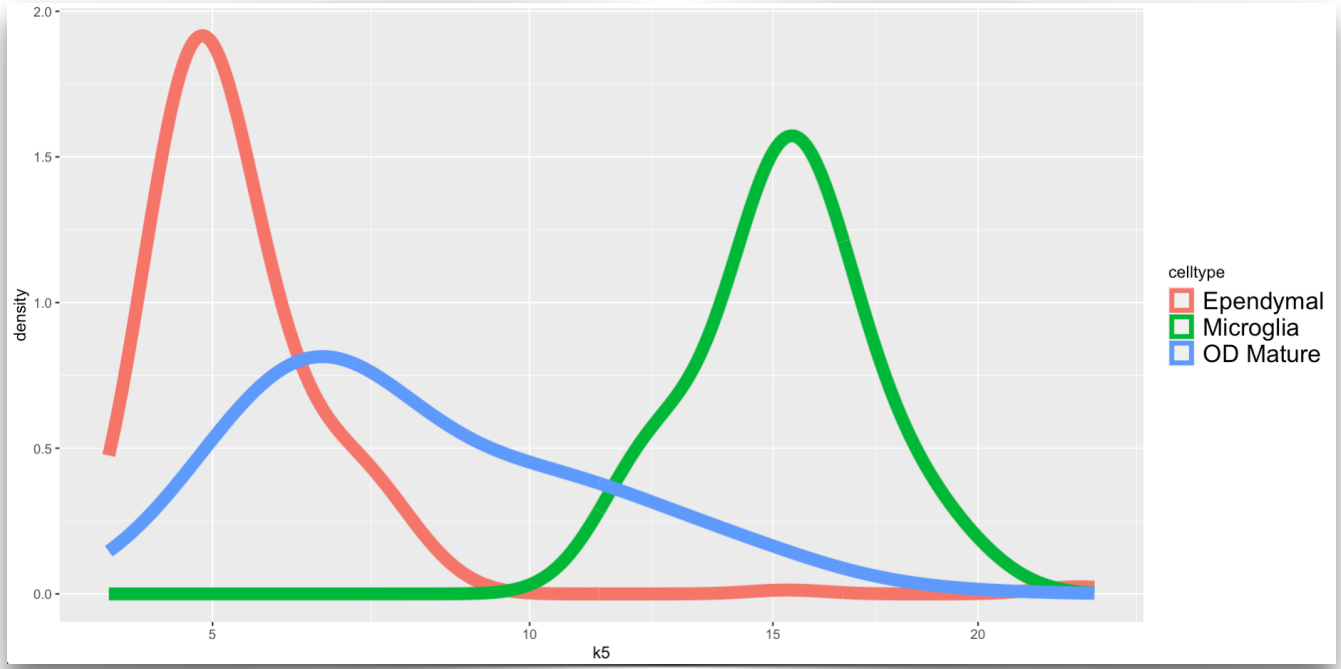






Other senses of spatial heterogeneity (again, functions!)

Martin



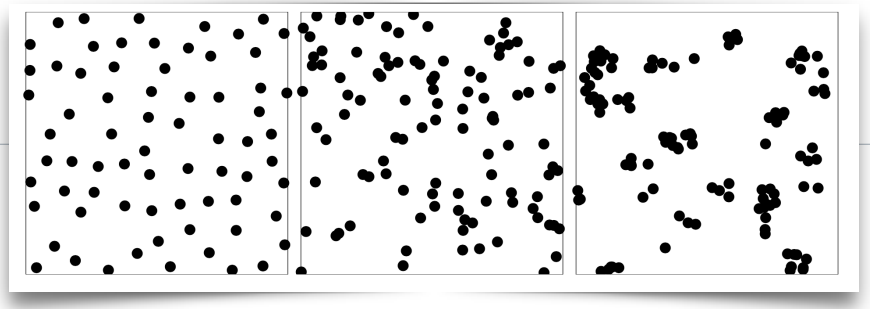
Distribution of distance to the 5th nearest neighbour

Extensions of the K function (1)

– multitype K-function Kij(r), also called the bivariate or cross-type K-function, is **the expected number of points of type j lying within a distance r of a typical point of type i**, standardised by dividing by the intensity of points of type j.

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

$$K_{ij}(r) = \frac{1}{\lambda_j}\mathbb{E}\left[t(u,r,\mathbf{X}^{(j)}) \mid u \in \mathbf{X}^{(i)}\right]$$

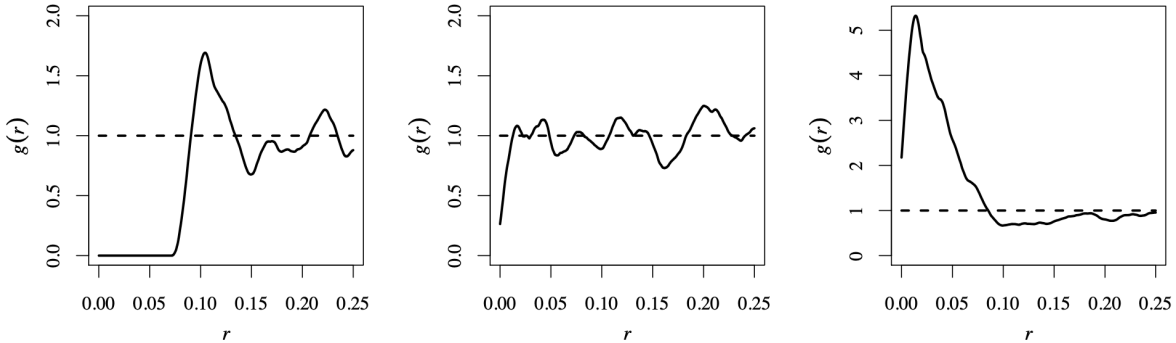


Extensions of the K function (2)

– L and g functions

An alternative tool is the **pair correlation function** $g(r)$ which contains contributions only from interpoint distances *equal to* r. In two dimensions, it can be defined by

$$g(r) = \frac{K'(r)}{2\pi r} \qquad (7.22)$$



Other point pattern summaries

Summary statistics for a point pattern:

`Fest`	empty space function F
`Gest`	nearest neighbour distribution function G
`Jest`	J-function $J = (1-G)/(1-F)$
`Kest`	Ripley's K-function
`Lest`	Besag L-function
`Tstat`	Third order T-function
`allstats`	all four functions F, G, J, K
`pcf`	pair correlation function
`Kinhom`	K for inhomogeneous point patterns
`Linhom`	L for inhomogeneous point patterns
`pcfinhom`	pair correlation for inhomogeneous patterns
`Finhom`	F for inhomogeneous point patterns
`Ginhom`	G for inhomogeneous point patterns
`Jinhom`	J for inhomogeneous point patterns
`localL`	Getis-Franklin neighbourhood density function
`localK`	neighbourhood K-function
`localpcf`	local pair correlation function
`localKinhom`	local K for inhomogeneous point patterns
`localLinhom`	local L for inhomogeneous point patterns
`localpcfinhom`	local pair correlation for inhomogeneous patterns
`Ksector`	Directional K-function
`Kscaled`	locally scaled K-function
`Kest.fft`	fast K-function using FFT for large datasets
`Kmeasure`	reduced second moment measure
`envelope`	simulation envelopes for a summary function
`varblock`	variances and confidence intervals for a summary function
`lohboot`	bootstrap for a summary function

Fundamentals of Spatial Statistics (the subset that is useful for spatial omics data)

- Point patterns
  - Definitions: intensity, homogeneity, dependence
  - Multi-type point patterns
  - Marked point processes
  - Statistical summaries
- **Lattice data**
  - **Definitions: lattice, regularity, neighbourhood matrix**
  - **Univariate global spatial autocorrelation**
  - **Univariate local spatial autocorrelation**
  - **Multivariate options**

Lattice data analysis

- Lattice: spatial units $D = \{A_1, A_2, \ldots, A_n\}$ that are not overlapping
- Lattice data: Random variable along the lattice $Y_i = Y(A_i)$
- Regularity: All spatial units have the same size, shape and observations are placed on a regular grid
- Neighbourhood matrix: $W = w_{ij}$ defines the spatial relationships

Spatial Autocorrelation Spatial autocorrelation measures take the form $w_{ij}U_{ij}$, which uses similarity measure U_{ij} weighted by the strength of the connection w_{ij} [66, 26].

Zuur et al. 2007

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

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

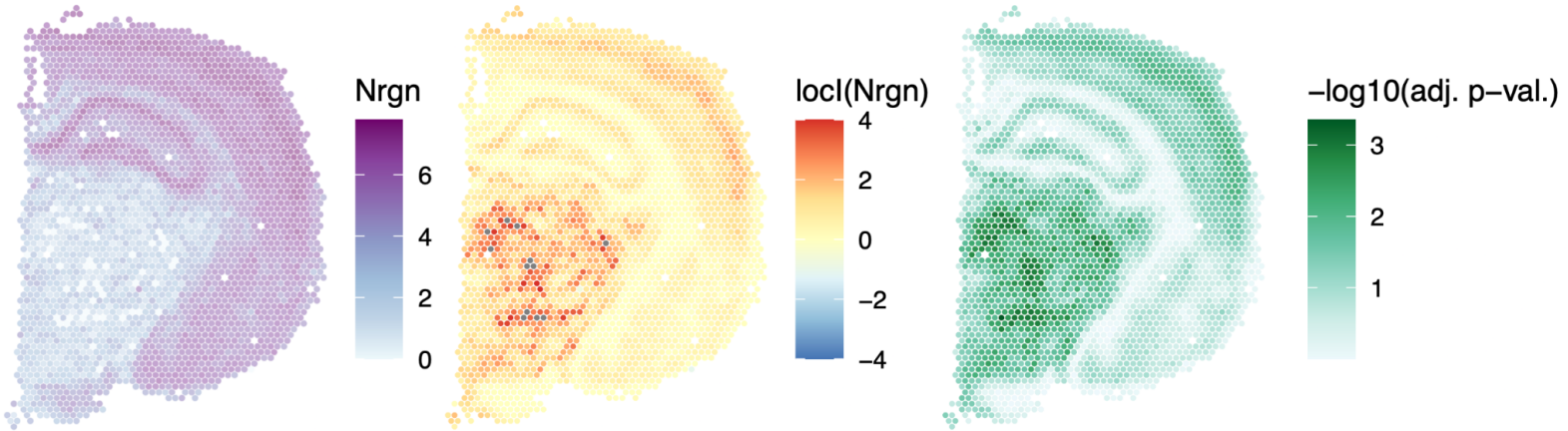
$$C = \frac{(N-1)\sum_i \sum_j w_{ij}(x_i - x_j)^2}{2W\sum_i (x_i - \bar{x})^2}$$

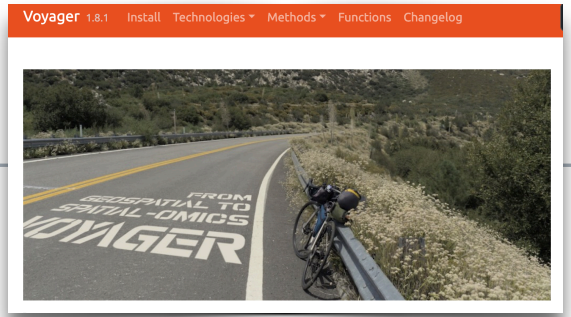
https://en.wikipedia.org/wiki/Moran%27s_I
https://en.wikipedia.org/wiki/Geary%27s_C

Spatial autocorrelation: Local Moran's I

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

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



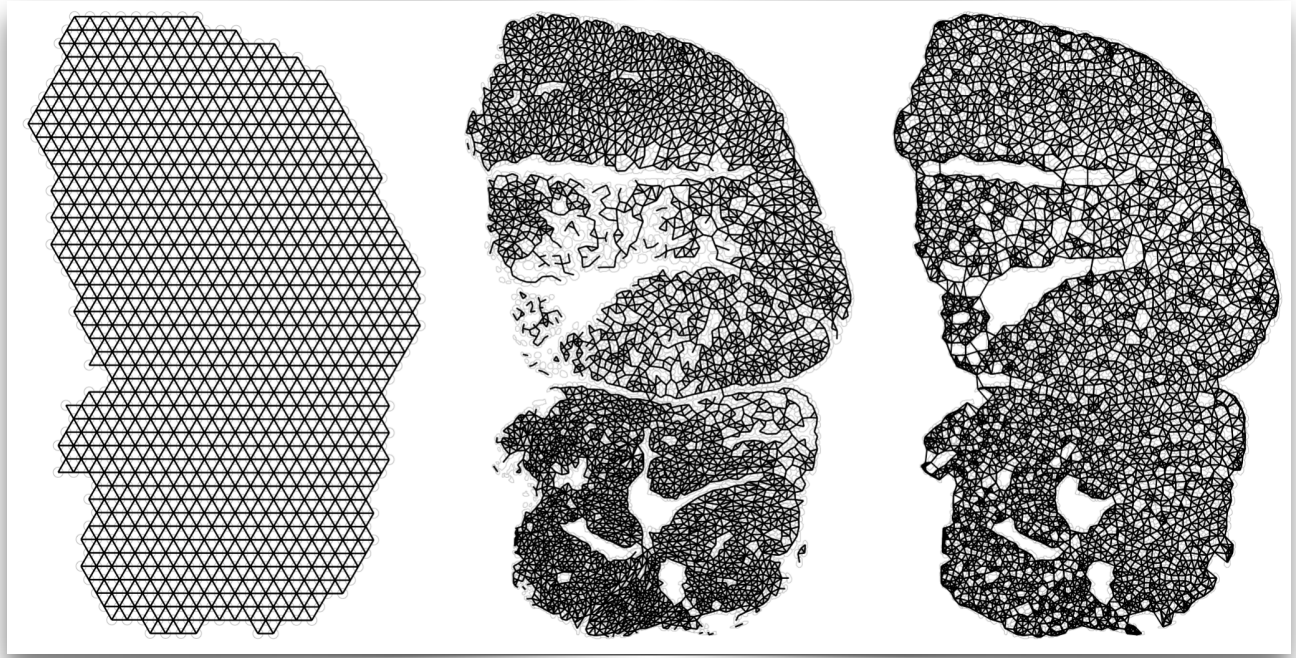


Spatial autocorrelation measures: weights!

Adjacent spots

H&E neighbors

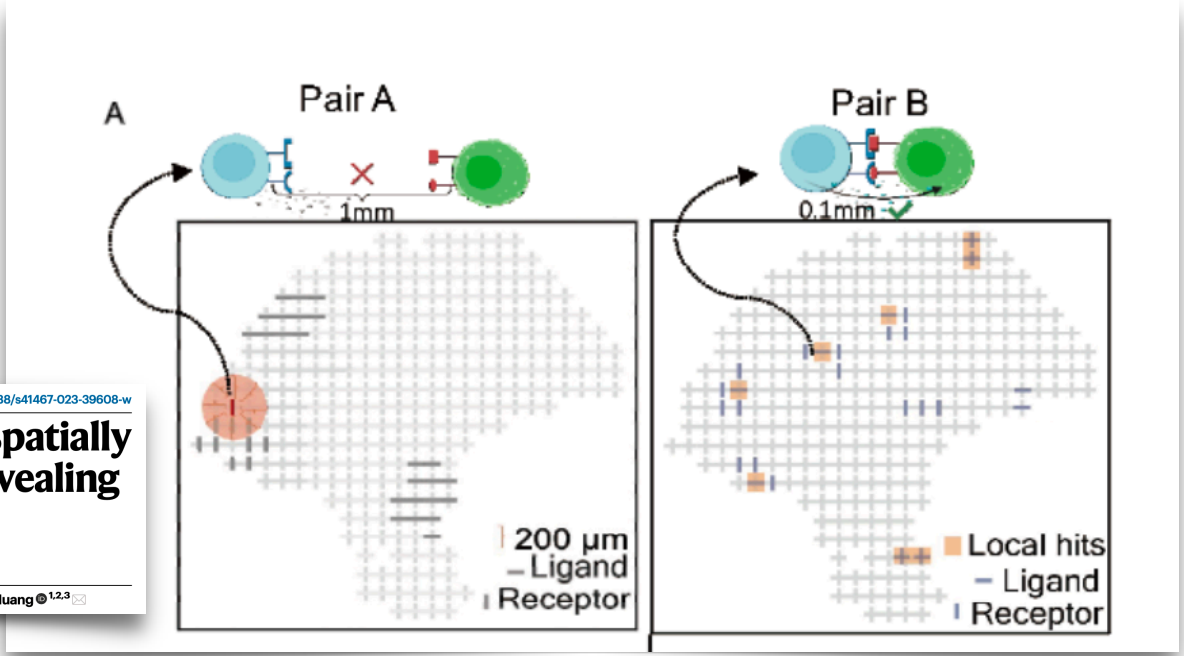
k=5 NNs



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

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]

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

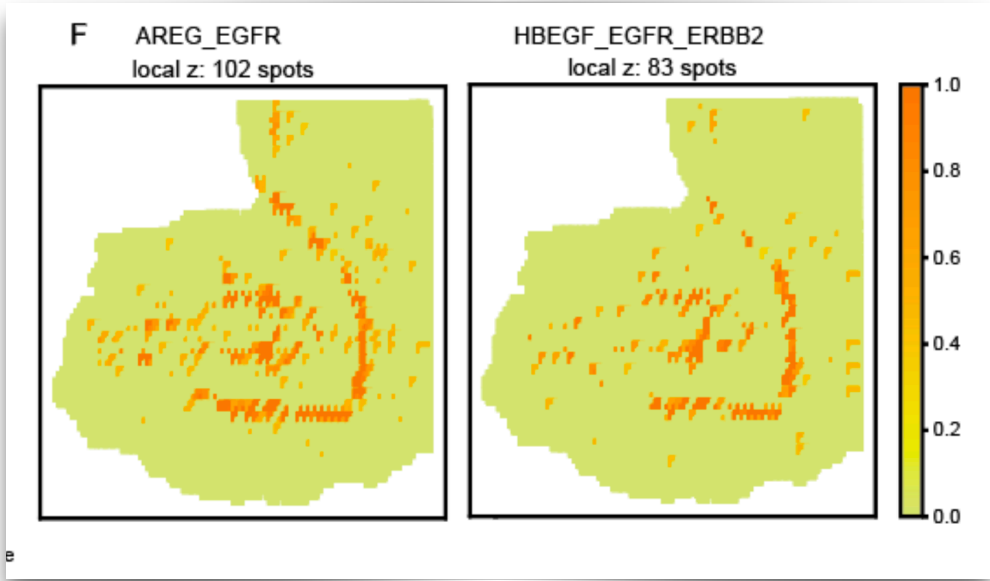
Cell-cell communication

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

SpatialDM Non-significant due to spatial range

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

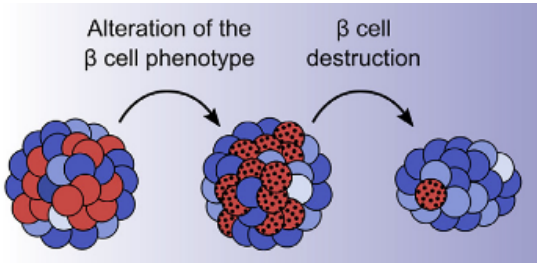
$$\text{Local Moran's R} = \sum_j w_{ij}\,\tilde{x}_i\,\tilde{y}_j + \sum_j w_{ij}\,\tilde{y}_i\,\tilde{x}_j$$

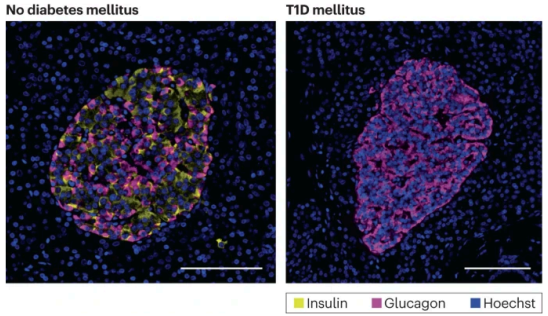


Example: Alpha and beta cells

Martin

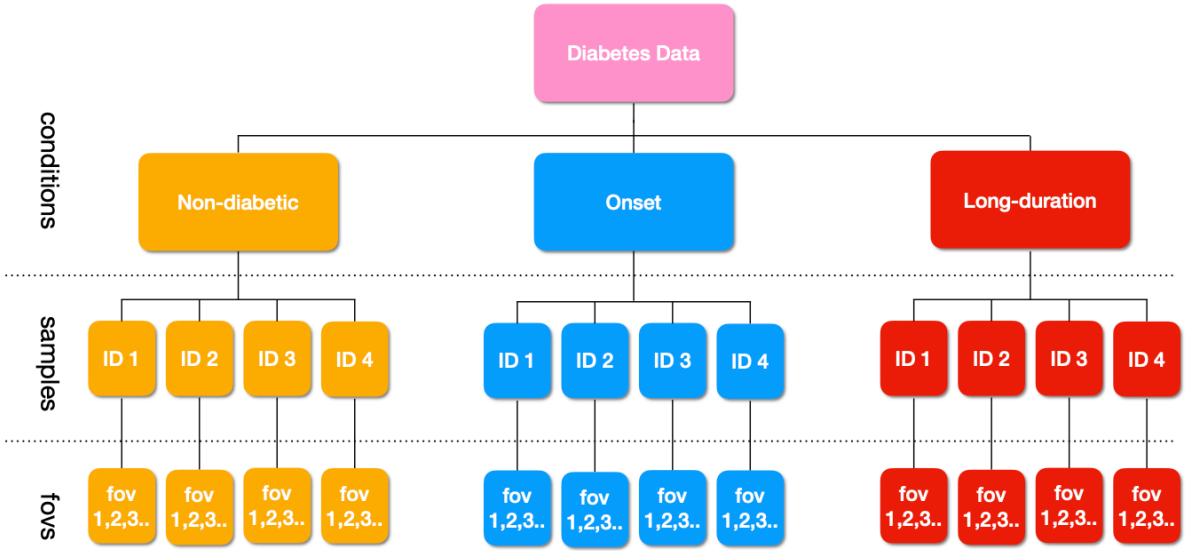
- “feature of type 1 diabetes (T1D) is that the immune system destroys pancreatic β-cells but not neighbouring α-cells, even though both β-cells and α-cells are dysfunctional.”
- Calculated **Gcross** metric between alpha and beta cells
    - if the alpha cells remain the same and beta cells get progressively destroyed, the distances should increase over the disease





`spatialFDA` can be applied to multi-sample/condition experiments

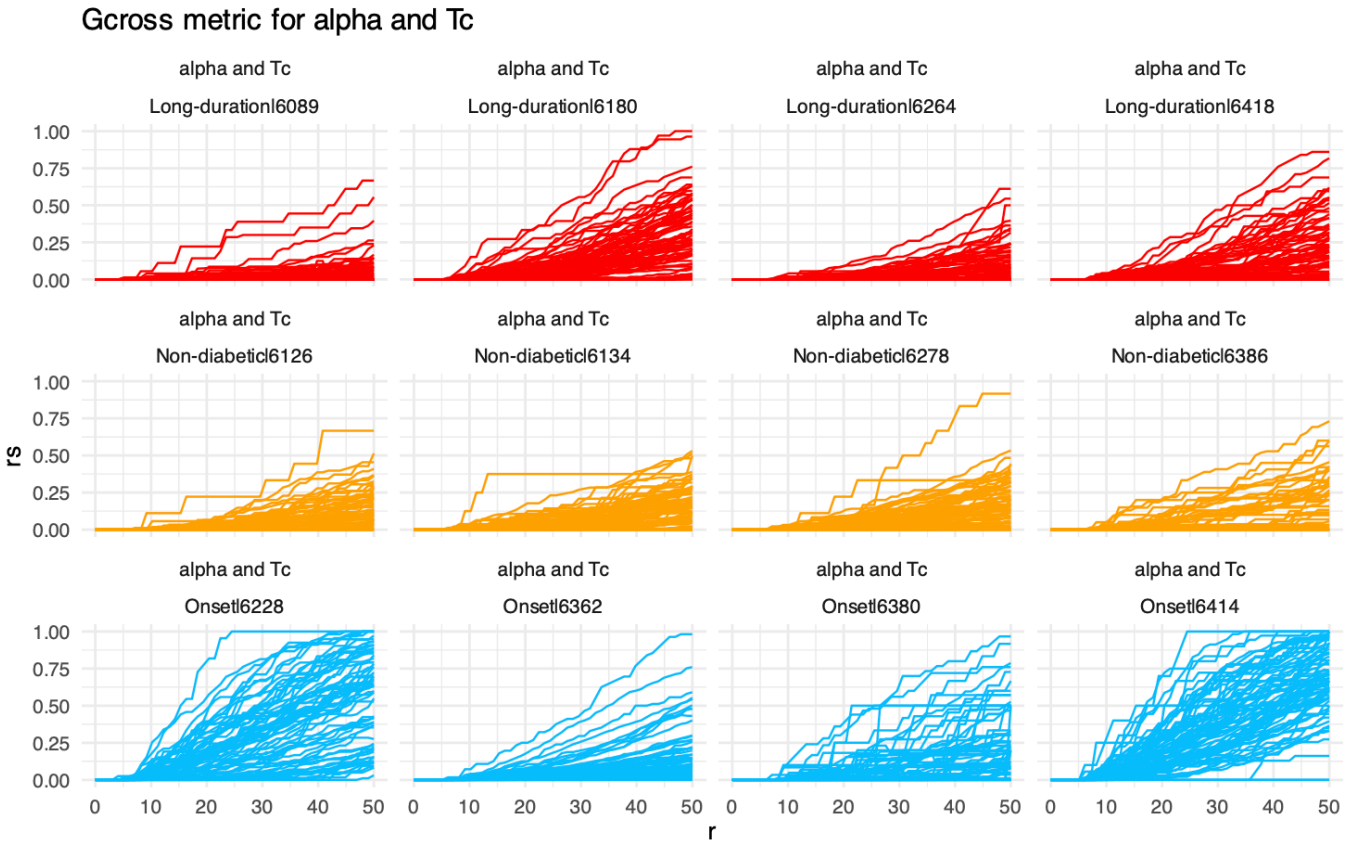
- Diabetes islet IMC dataset [Damond et al., 2019]
- Analyse recruitment of cytotoxic T cells to the islets during the disease → responsible for β cell destruction



Martin

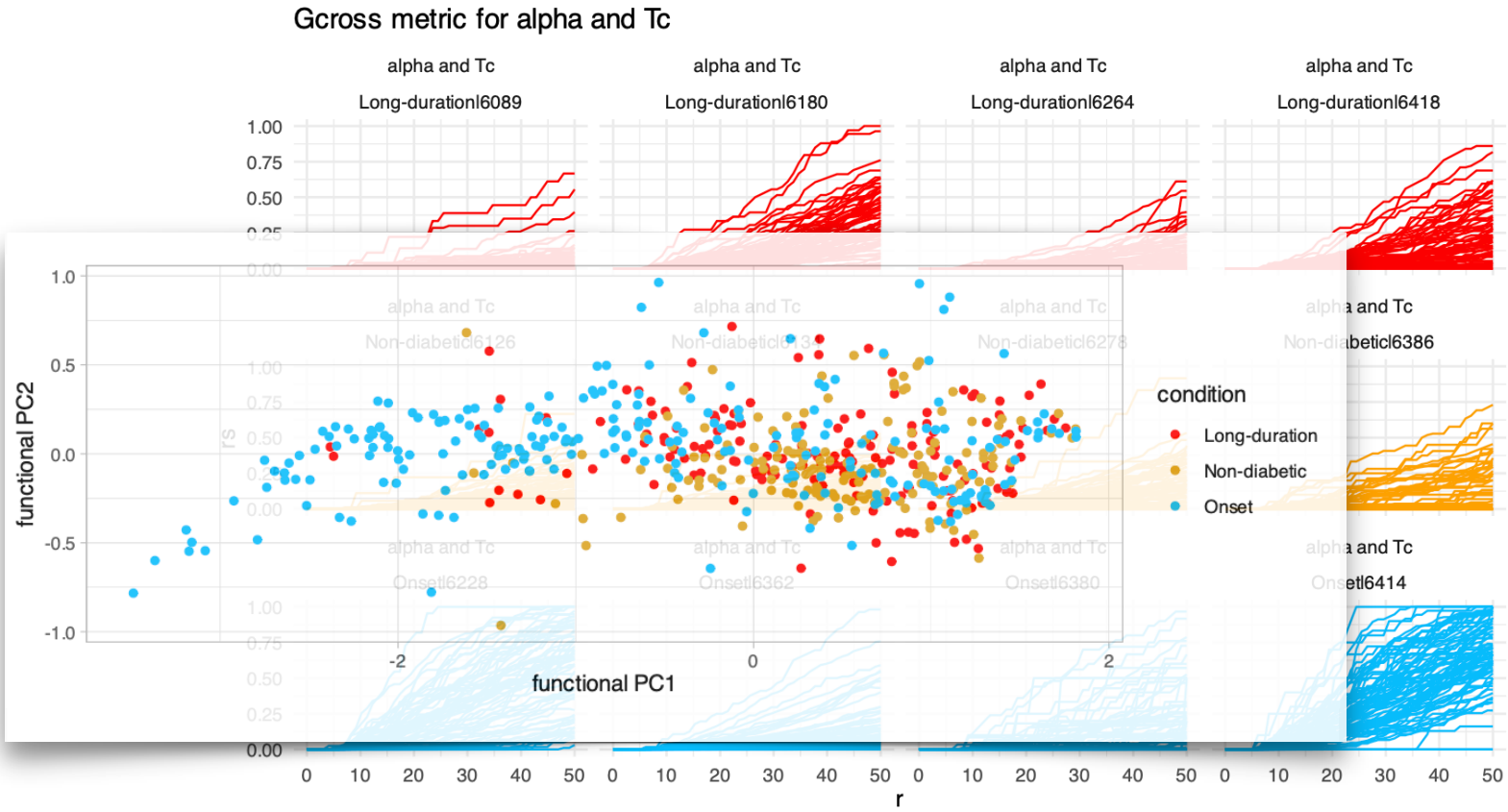
Spatial metrics summarise differential colocalisation

Martin



Spatial metrics summarise differential colocalisation

Martin



Functional general additive mixed model

Martin

- Compare not a summary statistic between conditions but the function over the domain r
- General additive mixed model with a functional response

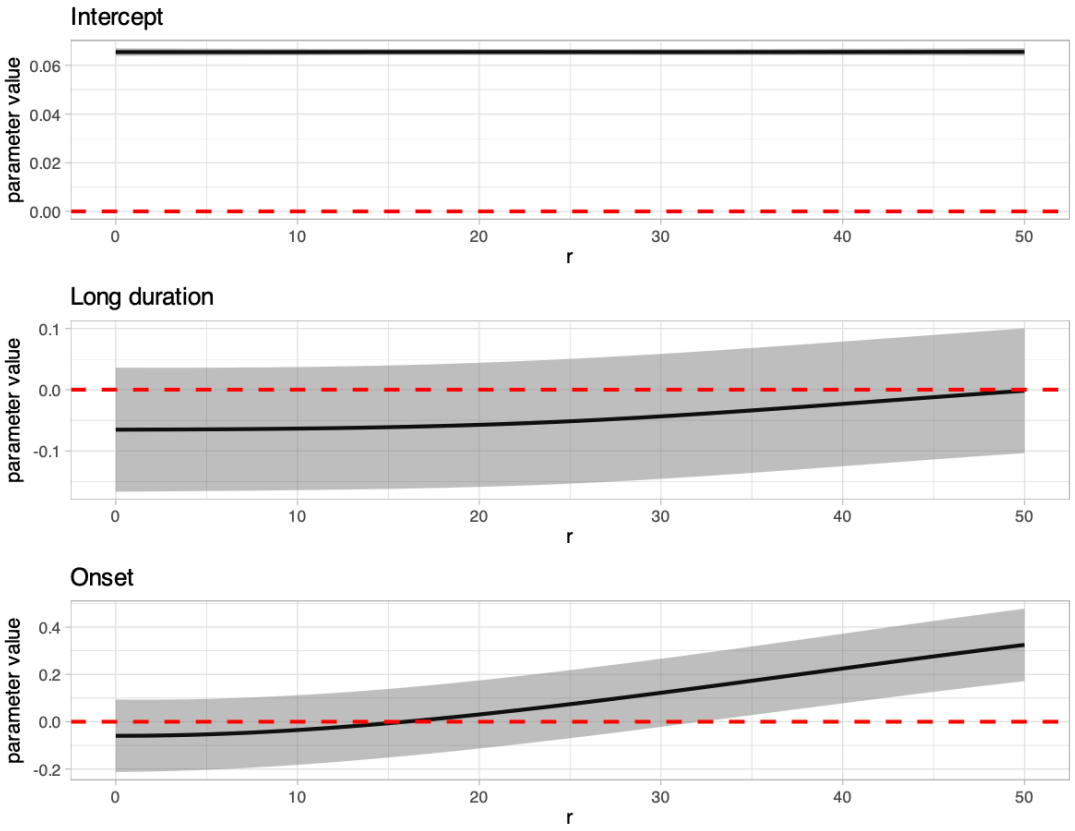
$$\mathbb{E}[y_i(r)] = g(\alpha(r) + \beta_{0,g(i)}(r) + \sum_{j=1}^{J} f_j(X_{ji}, r) + \epsilon_i(r))$$

- Compare non-diabetic curves as reference to onset and long-duration.
  [Scheipl et al., 2015, Scheipl et al., 2016]

`spatialFDA` Bioconductor package

fGAM results - parameters over a radius *r*

Martin



Other work in this space

SpaceANOVA: Spatial Co-occurrence Analysis of Cell Types in Multiplex Imaging Data Using Point Process and Functional ANOVA

Souvik Seal,* Brian Neelon, Peggi M. Angel, Elizabeth C. O'Quinn, Elizabeth Hill, Thao Vu, Debashis Ghosh, Anand S. Mehta, Kristin Wallace, and Alexander V. Alekseyenko

Bioimage informatics

spicyR: spatial analysis of *in situ* cytometry data in R

Nicolas P. Canete [1,2], Sourish S. Iyengar[3], John T. Ormerod[3], Heeva Baharlou [1,2], Andrew N. Harman[1,2] and Ellis Patrick [1,3,*]

search, University of Sydney, Westmead, NSW, Australia, [2]Sydney Medical School, University
School of Mathematics and Statistics, University of Sydney, Sydney, NSW, Australia

STExplorer: Navigating the Micro-Geography of Spatial Omics Data

Posted January 22, 2025.

Download PDF

Eleftherios Zormpas, Nikolaos I. Vlachogiannis, Anastasia Resteu, Adrienne Unsworth, Simon Tual-Chalot, Birthe Dorgau, Rachel Queen, Majlinda Lako, Dina Tiniakos, Alexis Comber, Quentin M. Anstee, Antonis Giakountis, Simon J. Cockell

doi: https://doi.org/10.1101/2025.01.17.633539

RESEARCH ARTICLE

FunSpace: A functional and spatial analytic approach to cell imaging data using entropy measures

Thao Vu*, Souvik Seal, Tusharkanti Ghosh, Mansooreh Ahmadian, Julia Wrobel¤, Debashis Ghosh

Concluding remarks

- Most spatial data can be represented as lattice data; some data can be represented as point patterns
- Point pattern analysis offers a few tools for exploring / inferring spatial heterogeneity —> represent summaries as **functions** and thus functional data analysis tools could also be useful
- Multi-sample analysis: functional PCA, functional GAM modelling = flexible framework
- Gotchas: confounding b/w intensity and correlation, weighting matrices

Statistical Bioinformatics Group, DMLS, UZH

Helena Samuel Martin

Simone Peiying

Universität Zürich[UZH]

University of Zurich[UZH]

URPP Evolution in Action: From Genomes to Ecosystems

FNSNF

Fonds national suisse
Schweizerischer Nationalfonds
Fondo nazionale svizzero
Swiss National Science Foundation

SIB Swiss Institute of Bioinformatics