Projects

Reminders:

- *project "plan" by **1.12.2025** (a few bullet points)* — done!
- *Private Slack channels (group + Mark, Hubert, optionally collaborators)* — done!
- *Private GitHub repos (of same name)* — done!
- due **9.01.2026** 18.00
- "Office hours" (zoom otherwise Slack): 20.-29.12 Mark is away

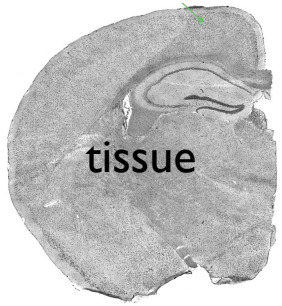
Background

Slide from Helena Crowell

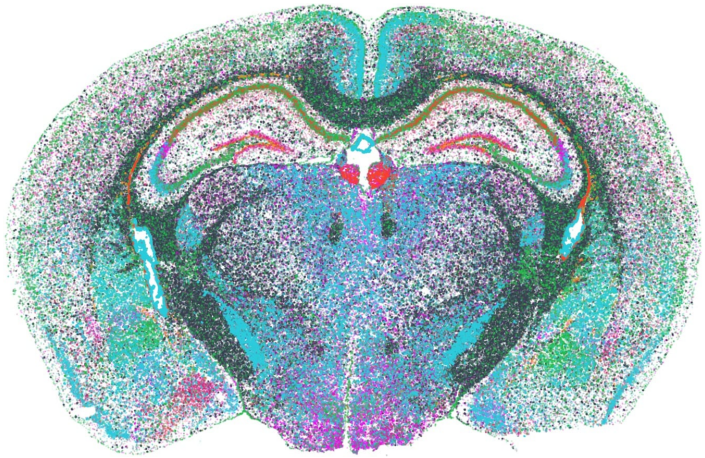
bulk

single-cell

spatial

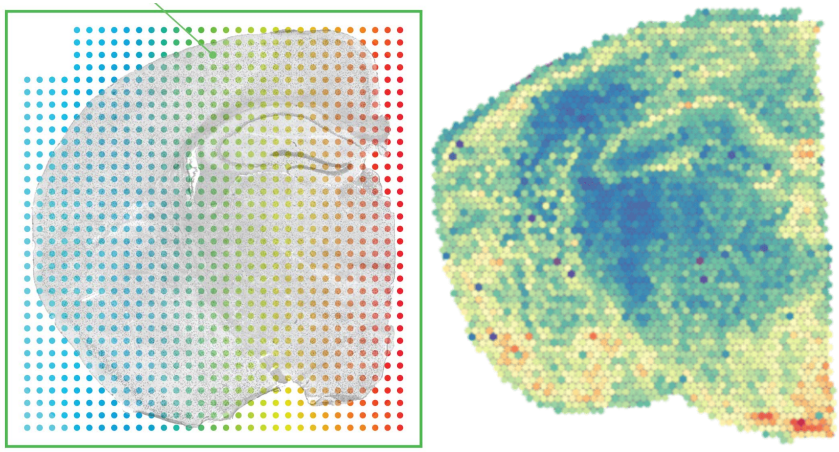


imaging-based



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

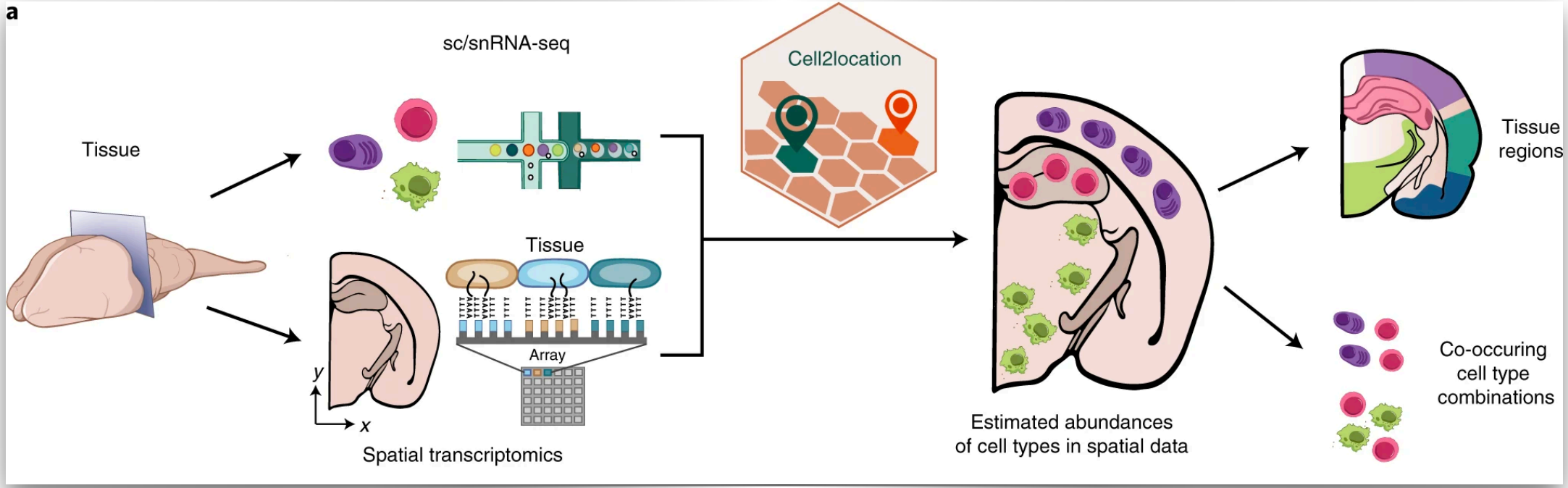
sequencing-based



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

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



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 $\alpha_{e,g}$:

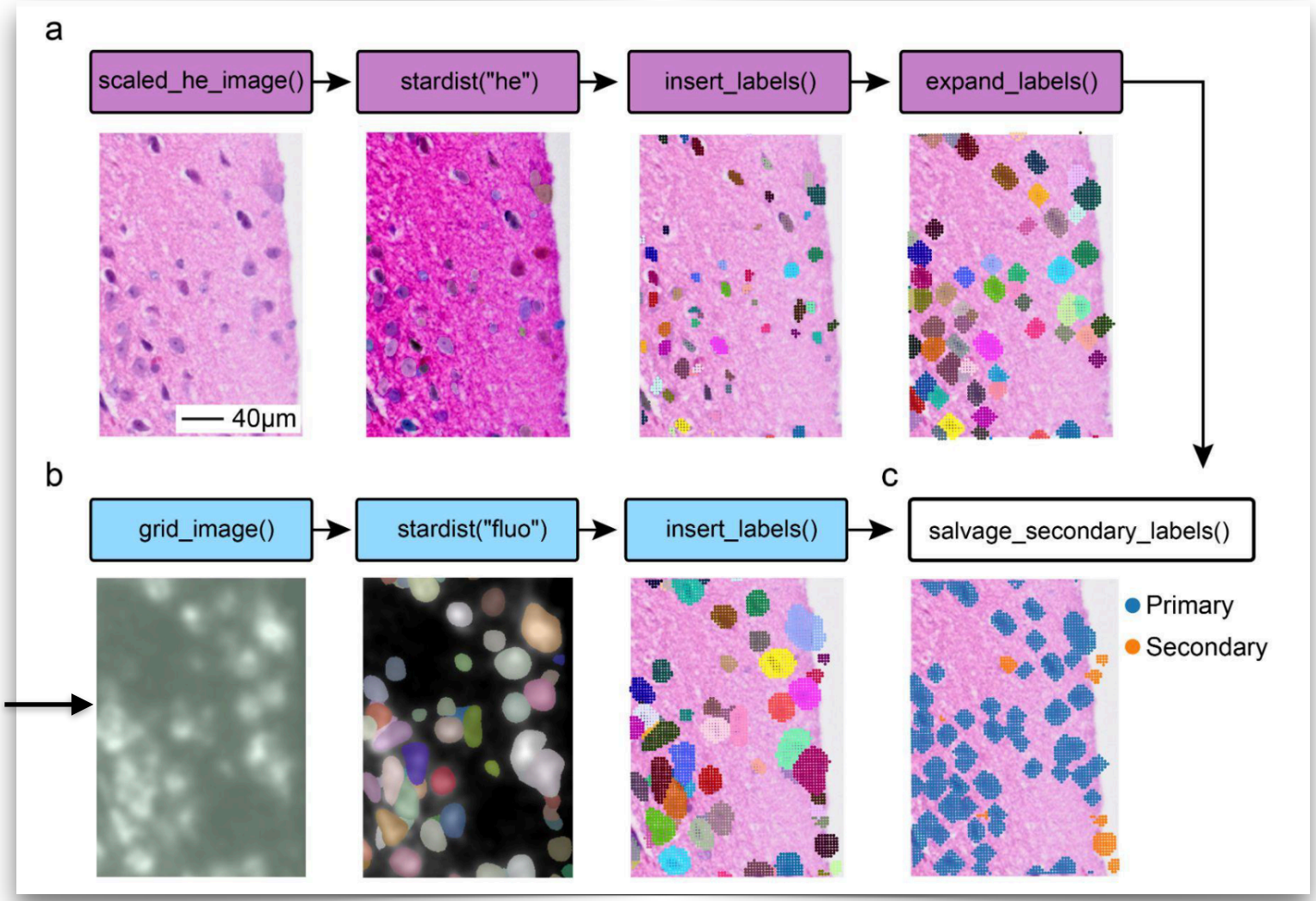
$$d_{s,g} \sim NB\left(\mu_{s,g}, \alpha_{e,g}\right).$$

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 $g_{f,g}$:

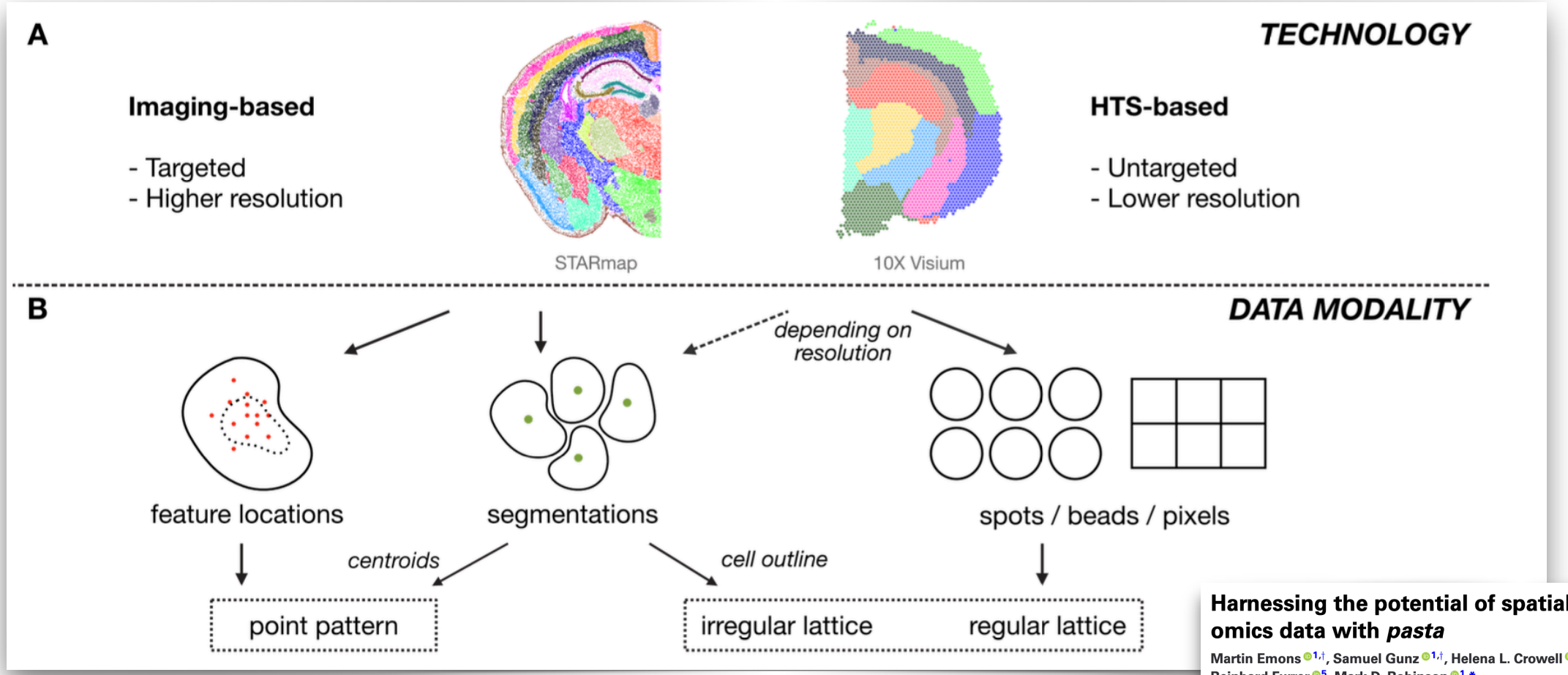
$$\mu_{s,g} = \left(\underbrace{m_g}_{\text{technology sensitivity}} \cdot \underbrace{\sum_f w_{s,f}\, g_{f,g}}_{\text{cell type contributions}} + \underbrace{s_{e,g}}_{\text{additive shift}} \right) \cdot \underbrace{y_s}_{\text{per-location sensitivity}} .$$

Aggregating high-resolution spatial omics (sequencing) data

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



`pasta`: Data representations determine spatial statistics options



Samuel

Martin

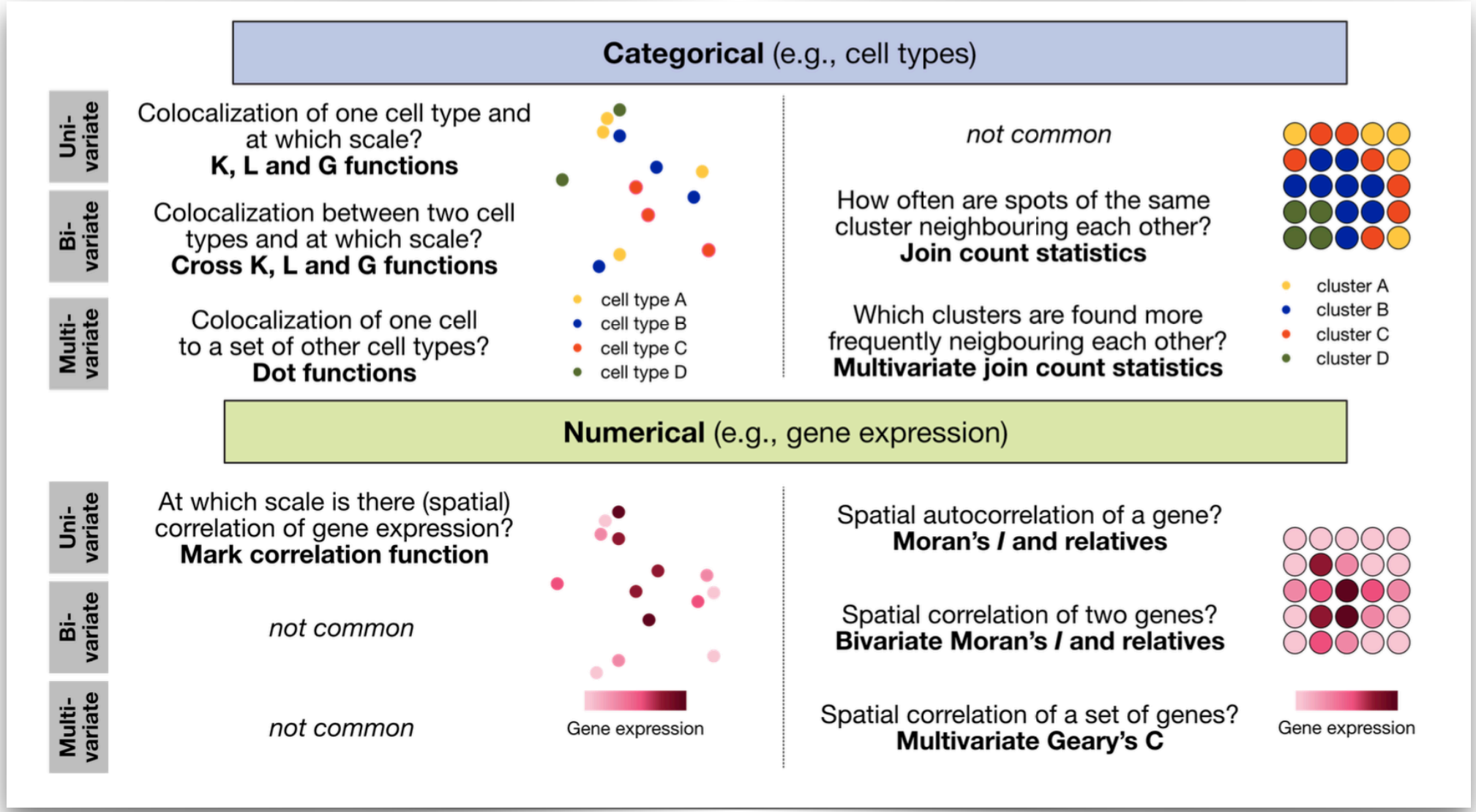
Harnessing the potential of spatial statistics for spatial omics data with *pasta*

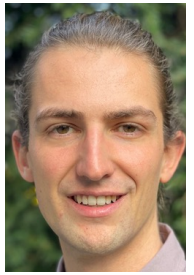
Martin Emons[1,†], Samuel Gunz[1,†], Helena L. Crowell[2], Izaskun Mallona[1], Malte Kuehl[3,4], Reinhard Furrer[5], Mark D. Robinson[1,*]

[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





Samuel

Martin

– "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

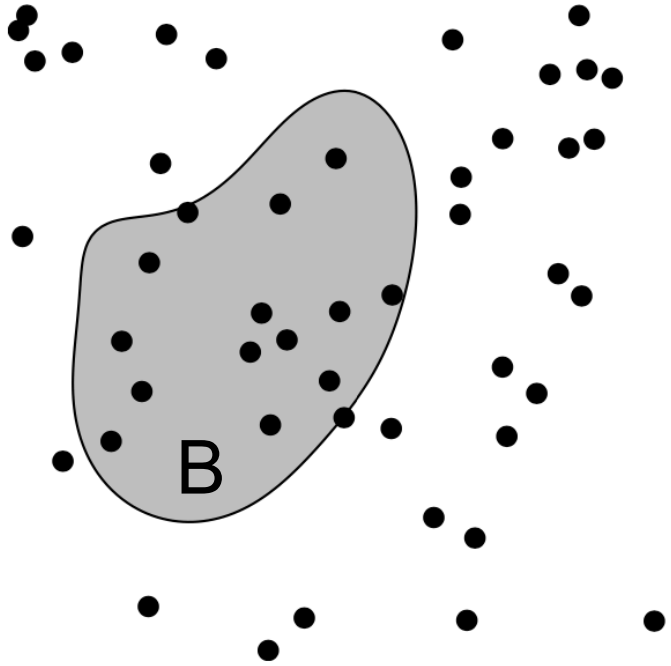
- notation: **X** is the point process; **x** is the (observed) point pattern
- lambda: intensity function

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:

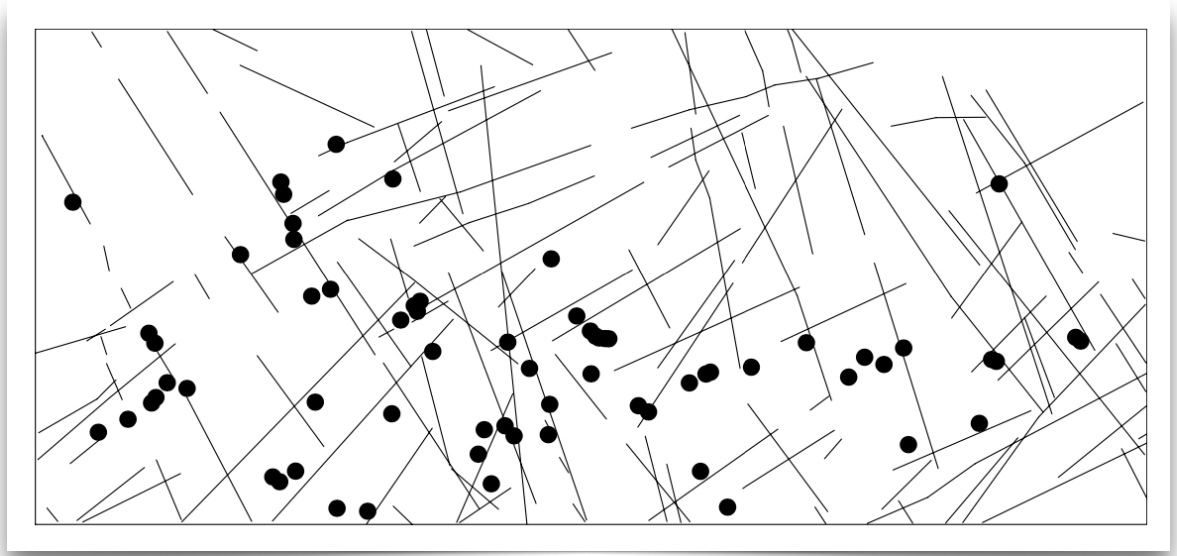
intensity function: the expected number of points falling in a region B is the integral $\mu = \int_B \lambda(u)\,du$ of the intensity function $\lambda(u)$ over the region B;

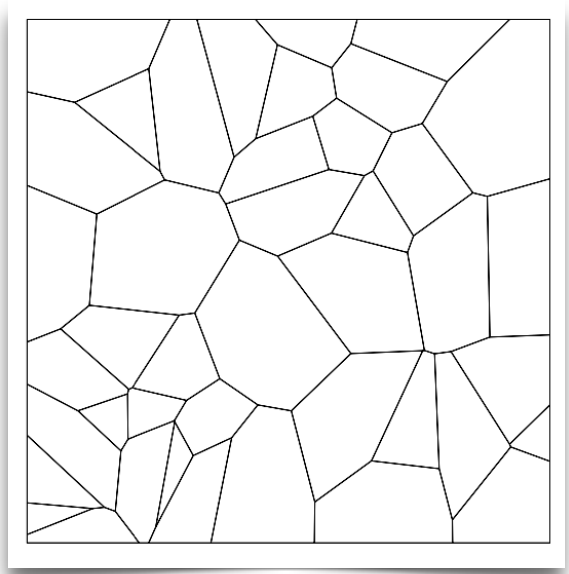
independence: if space is divided into non-overlapping regions, the random patterns inside these regions are independent of each other;

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

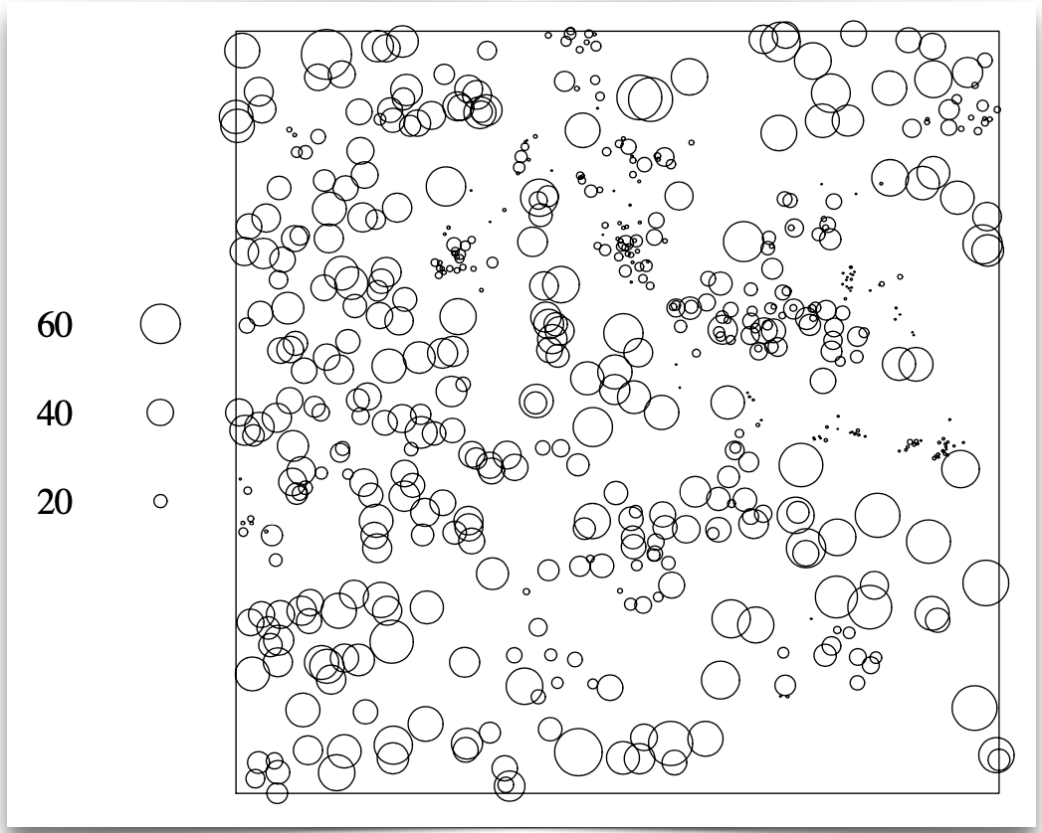
Extensions of point patterns

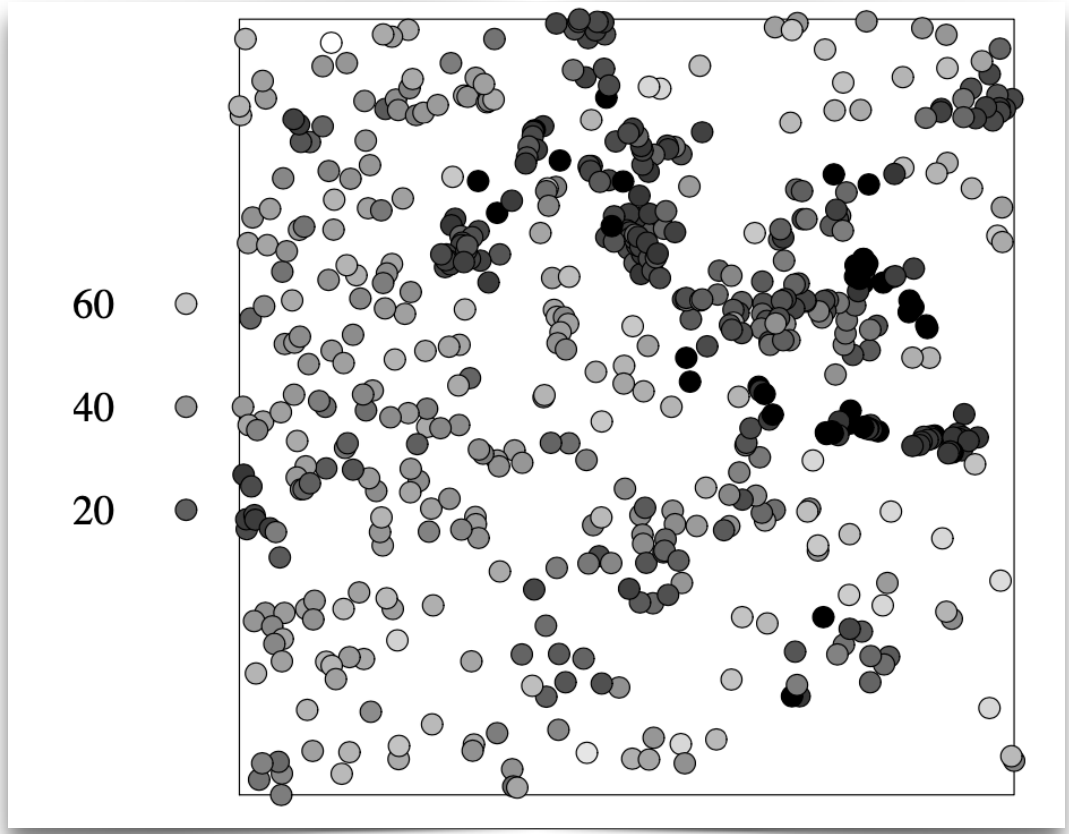
- (patterns can be regions/lines, not just points; can be in higher dimension (e.g., 3D); temporal component .. most of the methods I discuss have extensions; not discussed here)
- marks —> marked point process
- types —> multi-type point process
- covariates





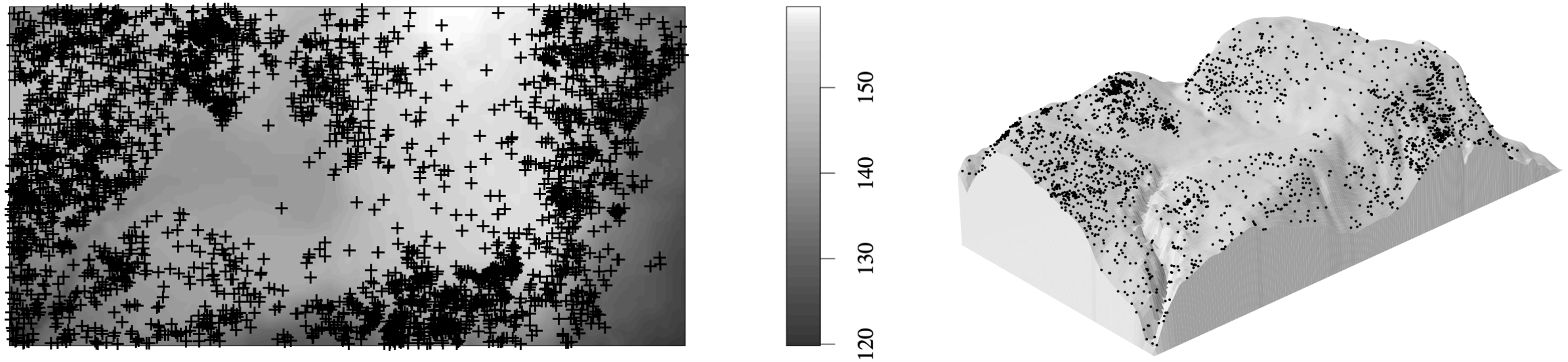
Marks





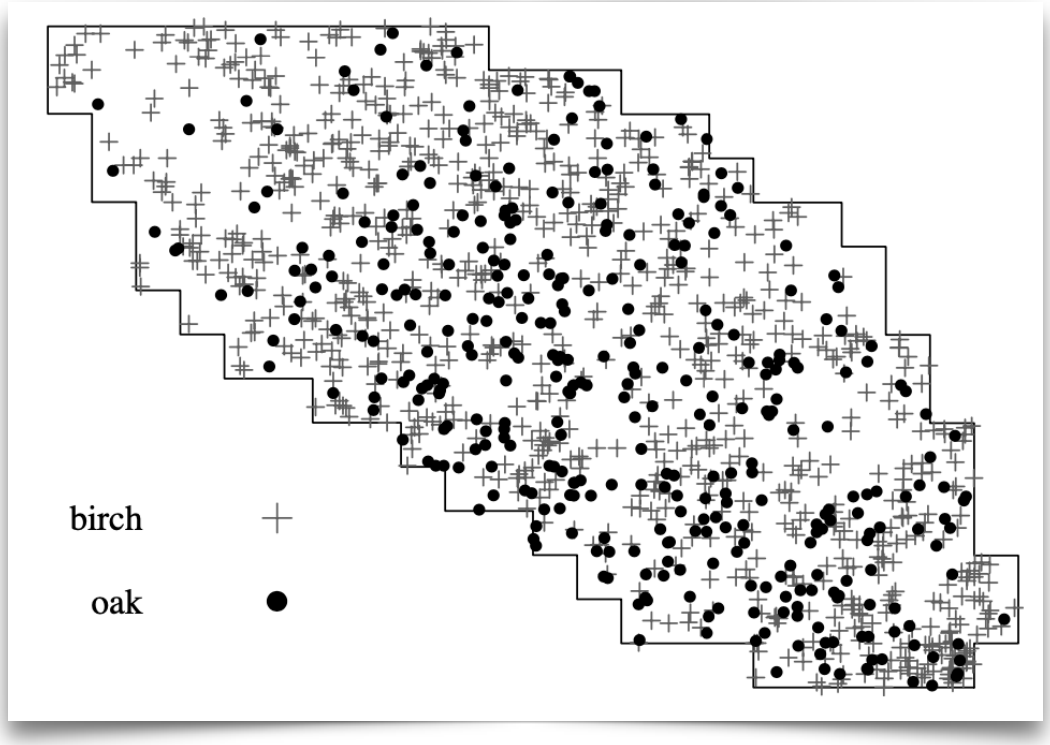
Longleaf pines dataset (location and diameter)

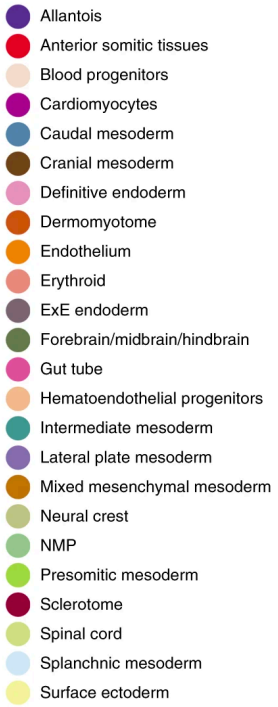
Covariates

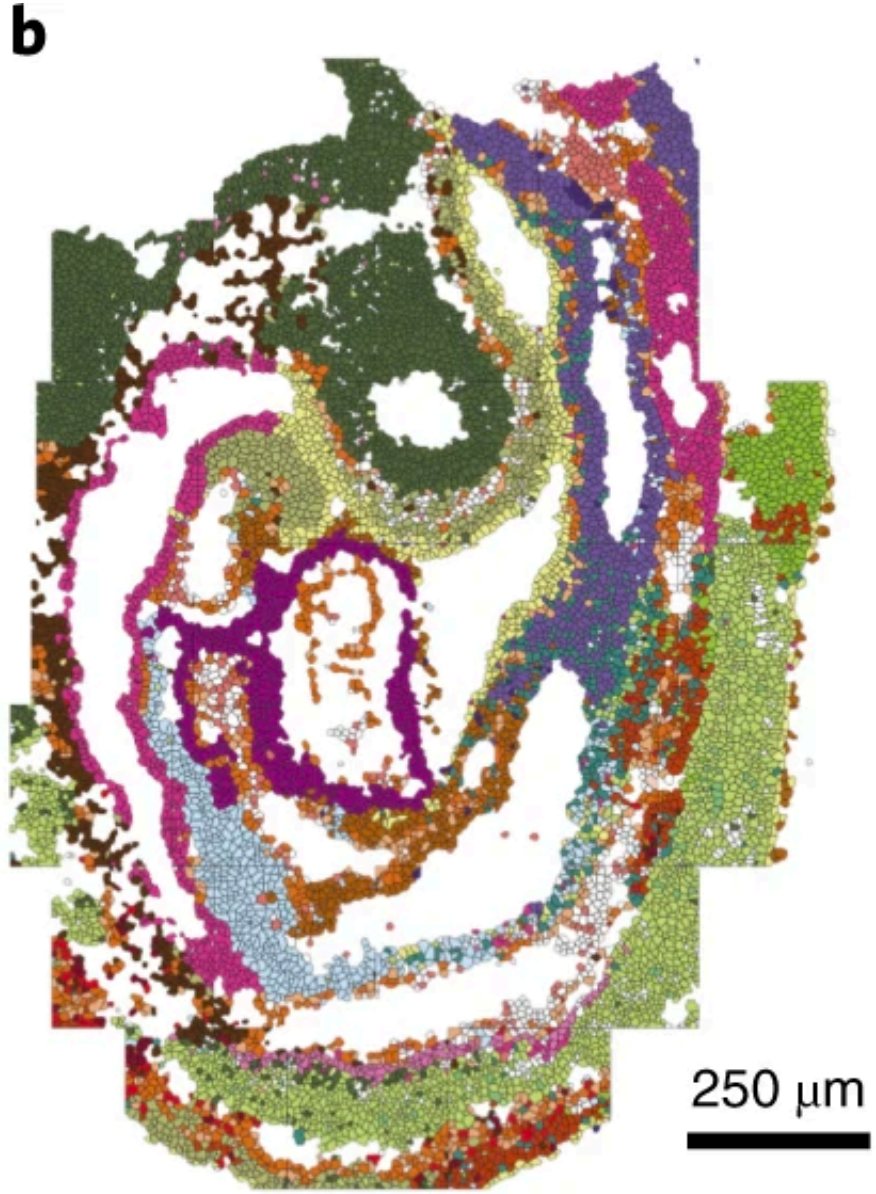


Locations of Beilschmiedia pendula *trees (+) and terrain elevation (greyscale) in a* 1000 × 500 *metre survey plot in Barro Colorado Island.*

Multi-type point patterns



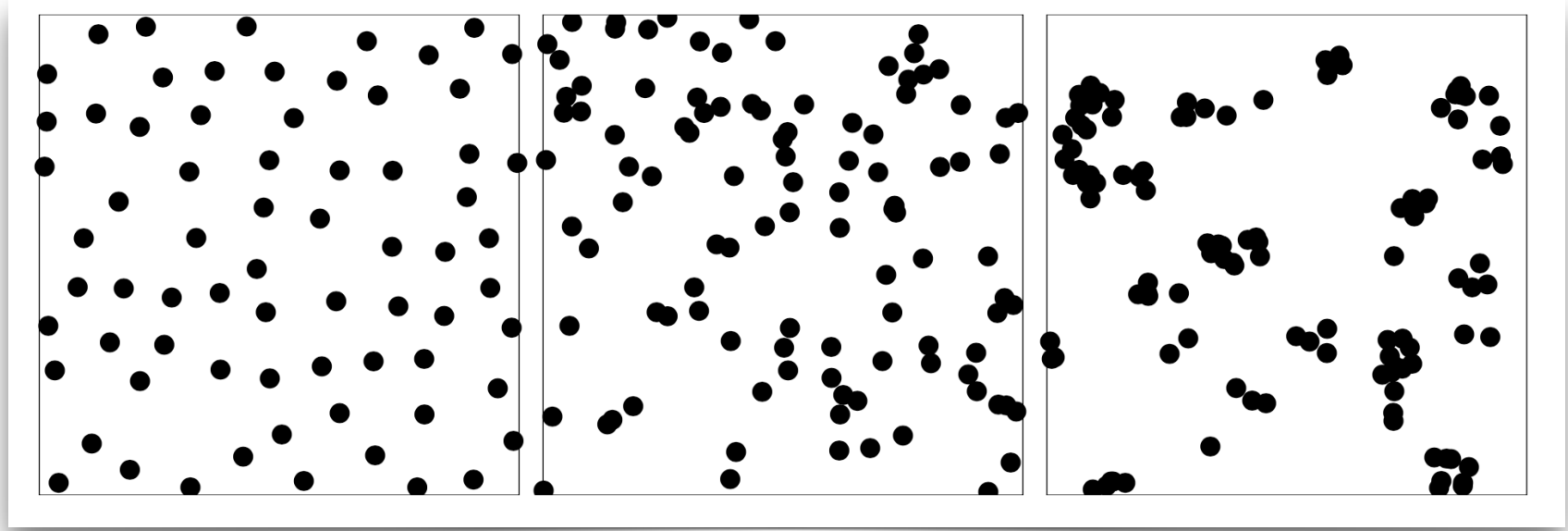




https://www.nature.com/articles/s41587-021-01006-2

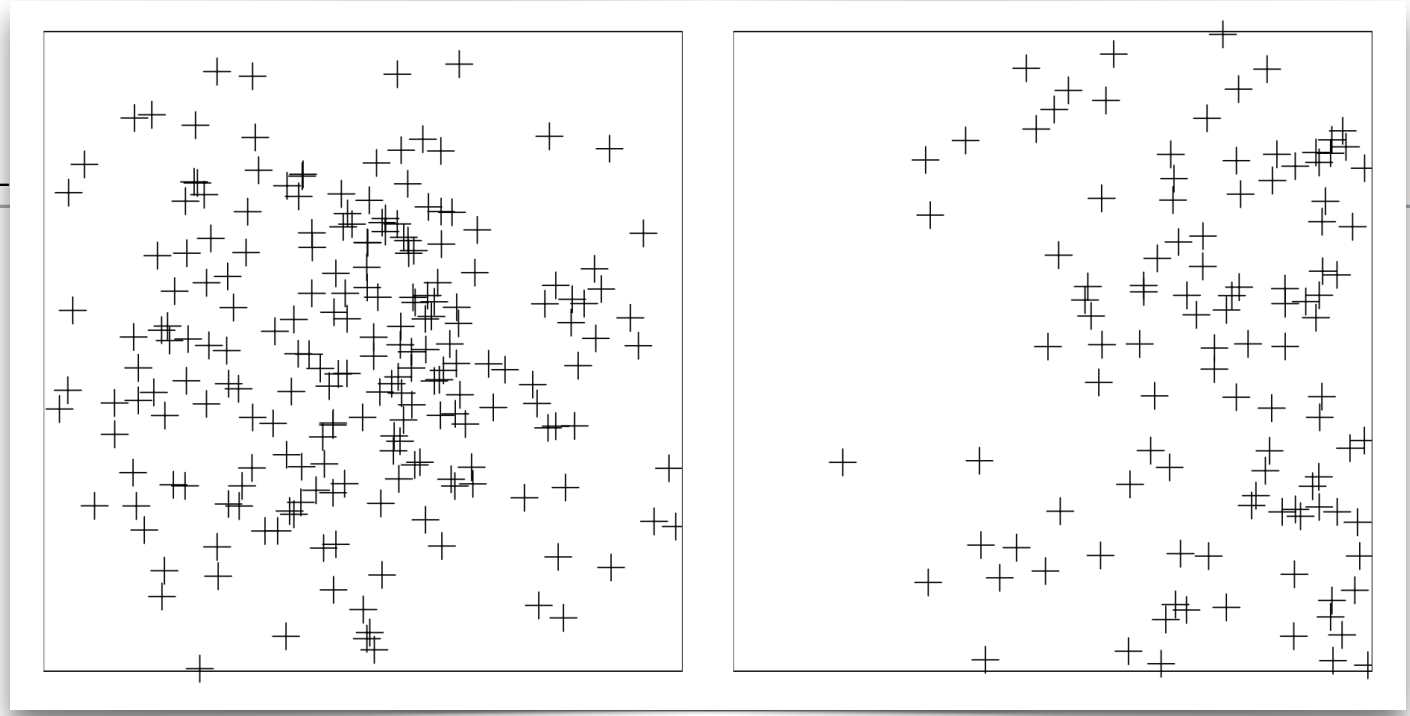
What is a point pattern?

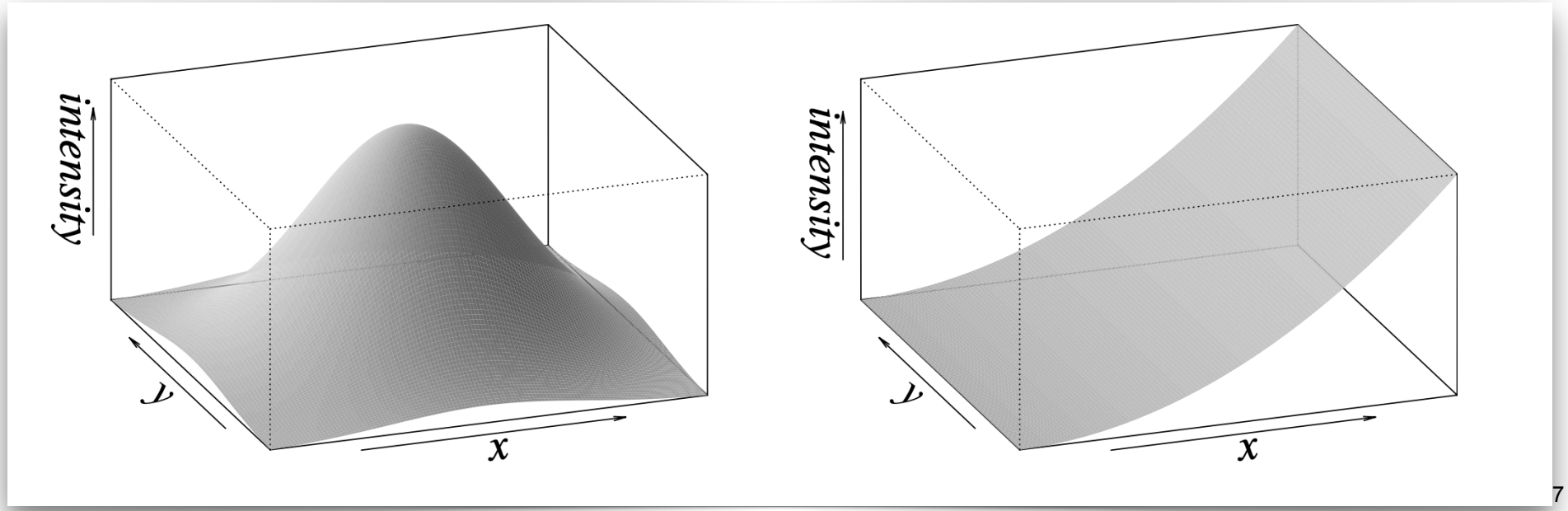
- "a realisation of a spatial point process effectively assumes that the locations of points are not fixed, and that the point pattern is the response or observation of interest."
- Which of these is homogeneous?
- Which of these is completely spatially random (CSR)?
- Which of these is clustered?
- Which of these is not independent?



Intensity estimation

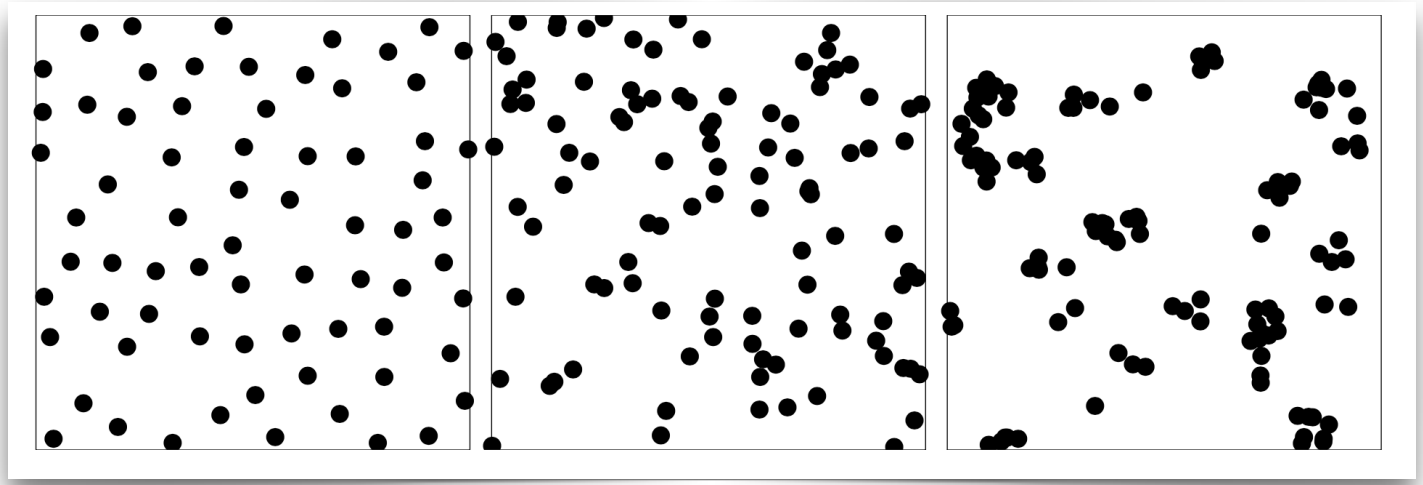
– Methods for intensity estimation

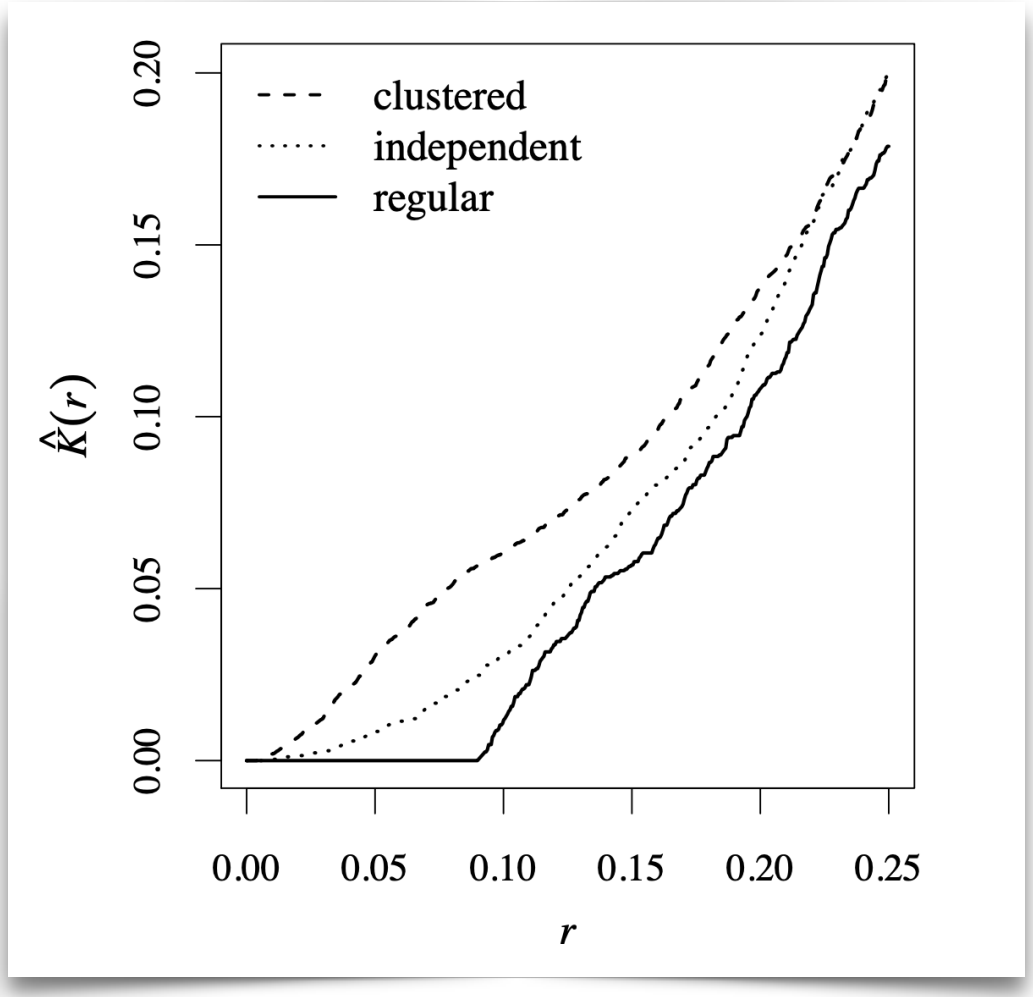


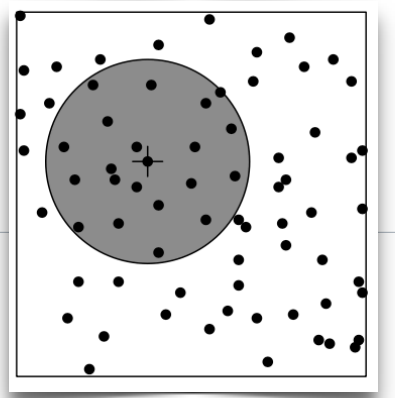


Correlation for point patterns

- 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?

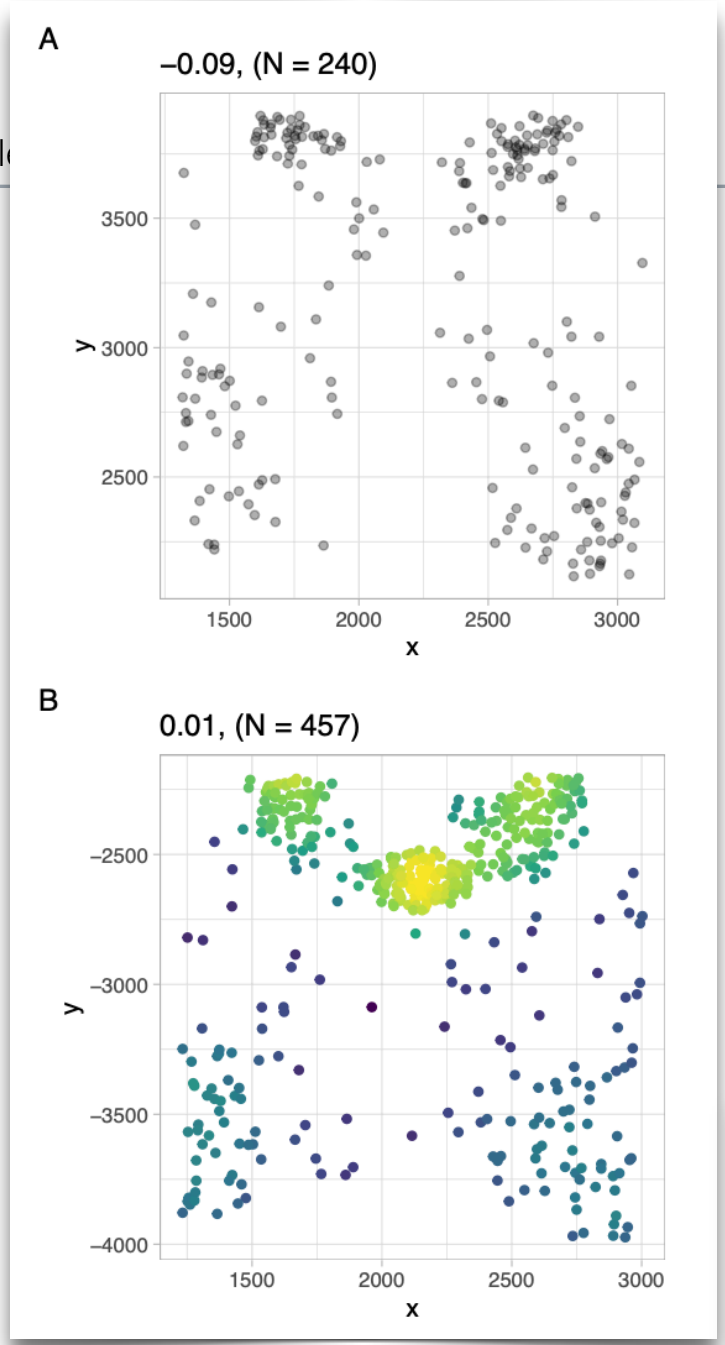
- inhomogeneous correlation functions
- edge correction

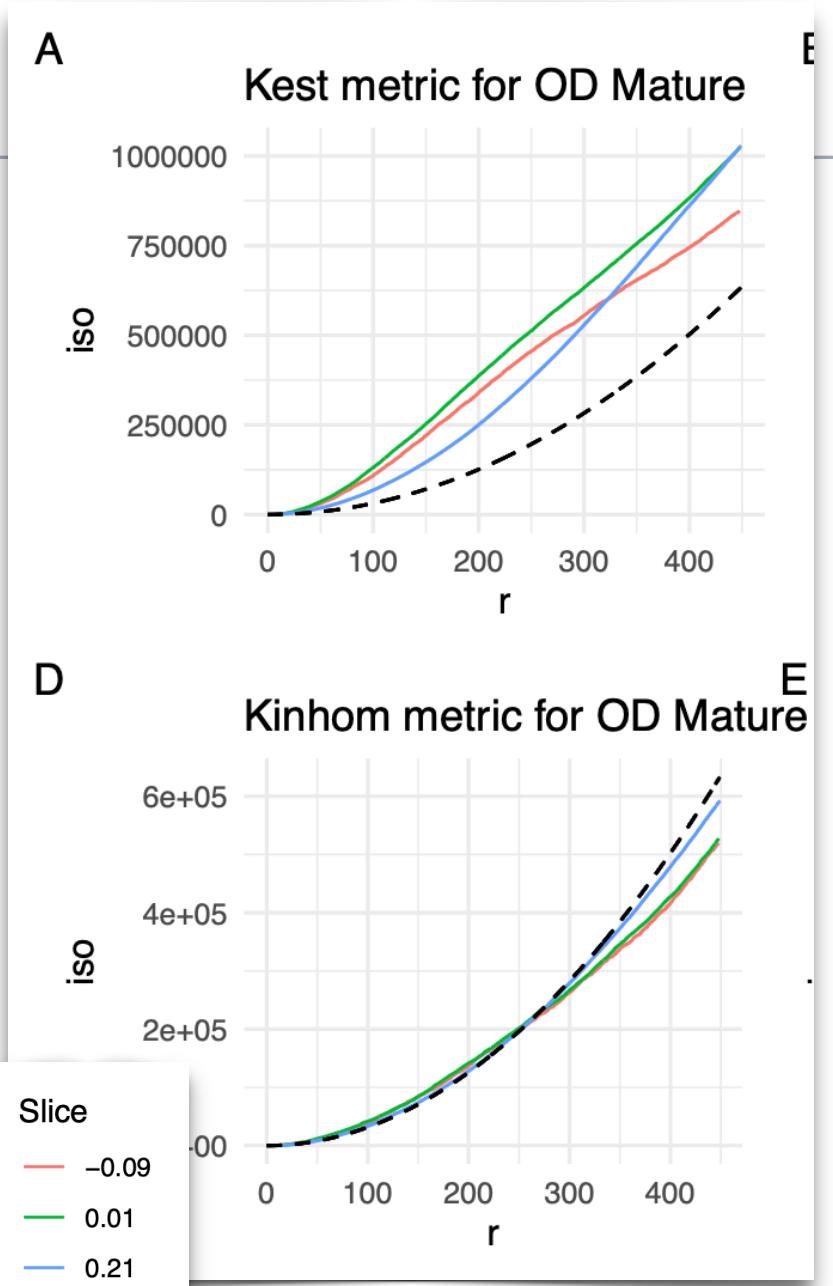
$$\widehat{K}_{inhom}(r) = \frac{1}{D^p|W|}\sum_i\sum_{j\neq i}\frac{\mathbf{1}\{||x_i - x_j|| \leq r\}}{\widehat{\lambda}(x_i)\widehat{\lambda}(x_j)}e(x_i, x_j; r)$$

- 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.





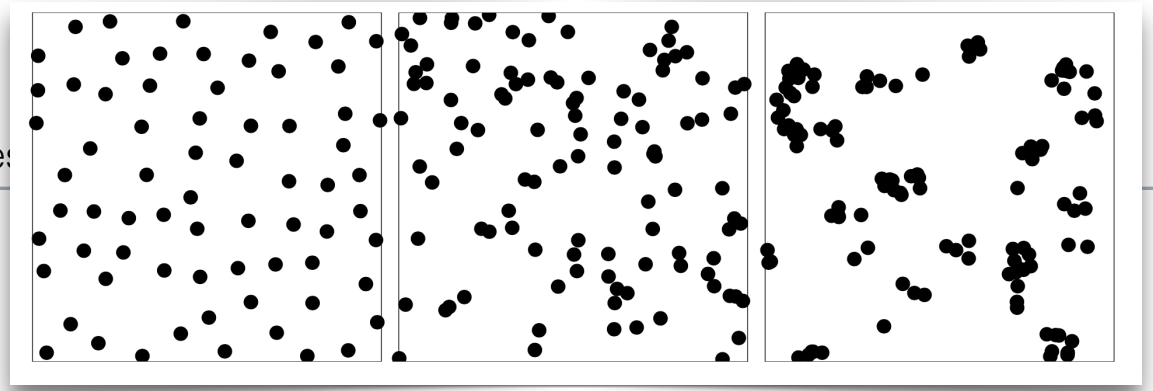
Extensions of the K function (1)

- 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.

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

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

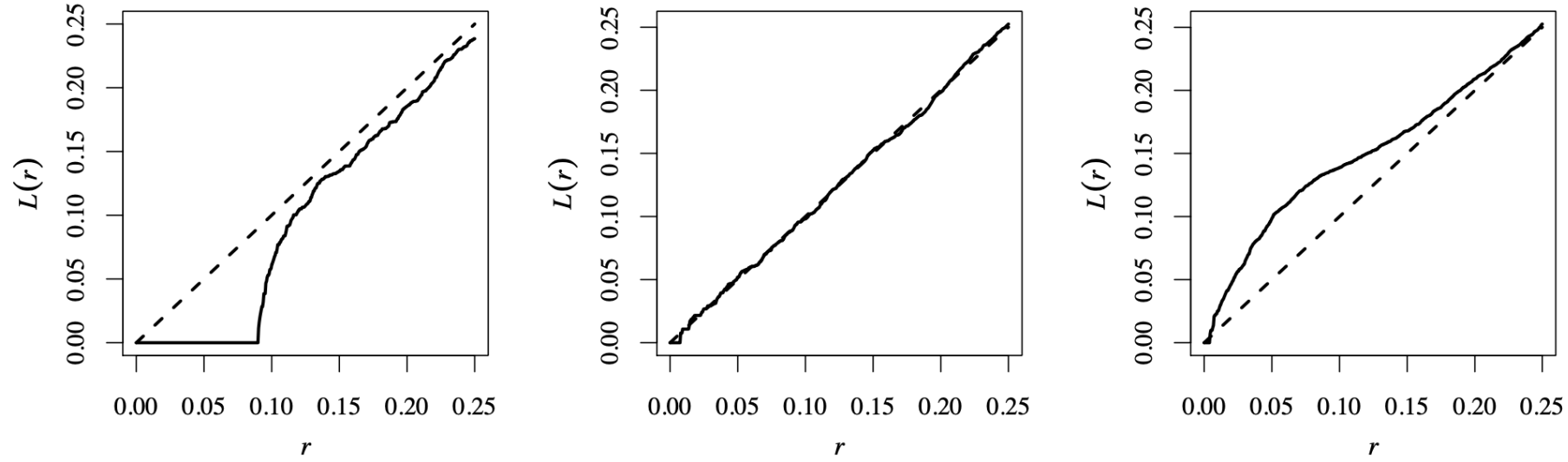
Extensions of the K function (2)

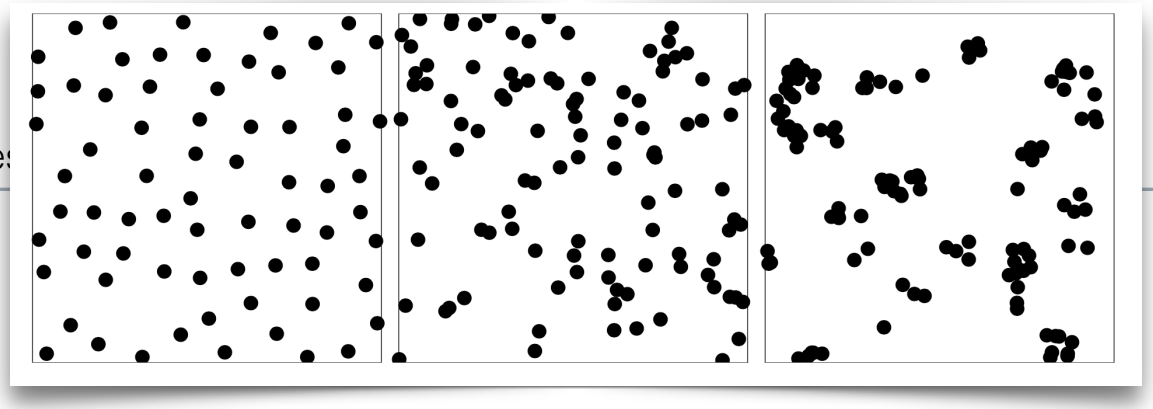


– L and g functions

A commonly used transformation of K proposed by Besag [103] is the **L-function**

$$L(r) = \sqrt{\frac{K(r)}{\pi}}$$



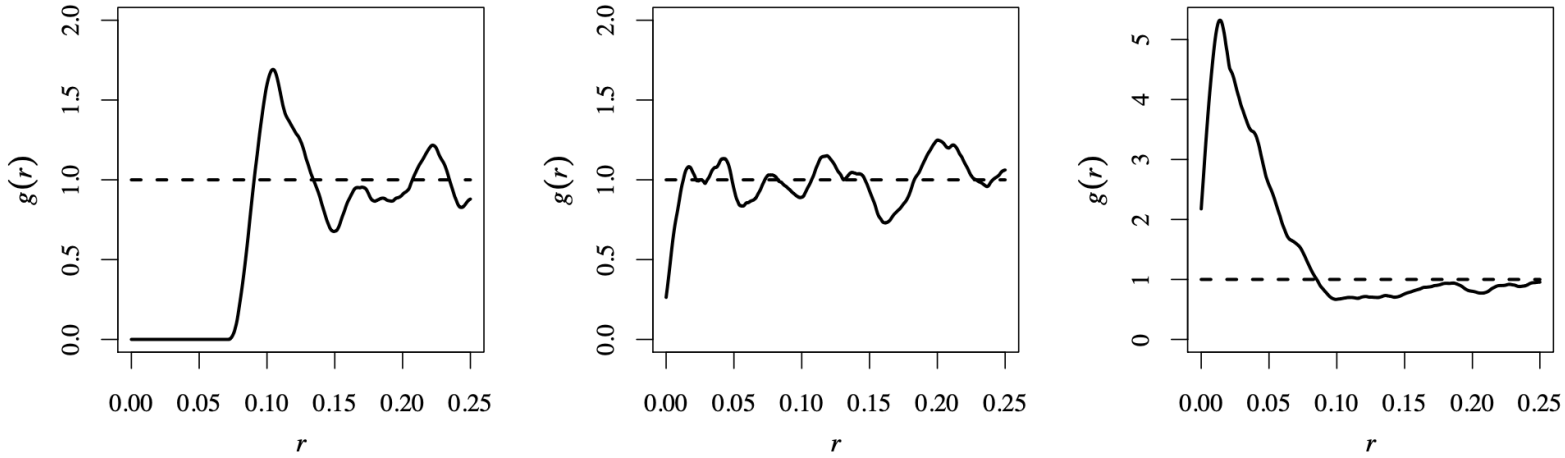


Extensions of the K function (2)

– 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

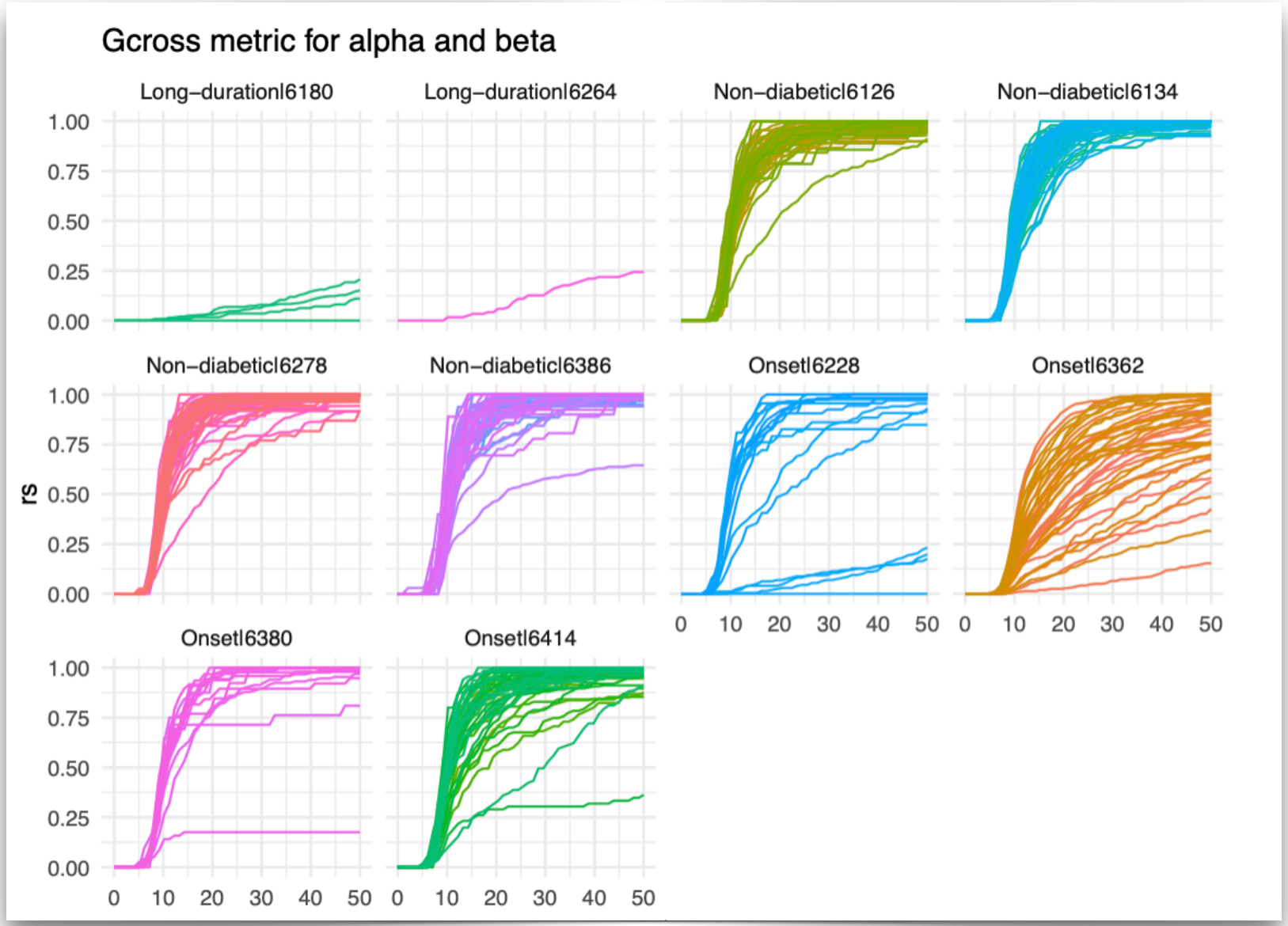
$$g(r) = \frac{K'(r)}{2\pi r} \qquad (7.22)$$



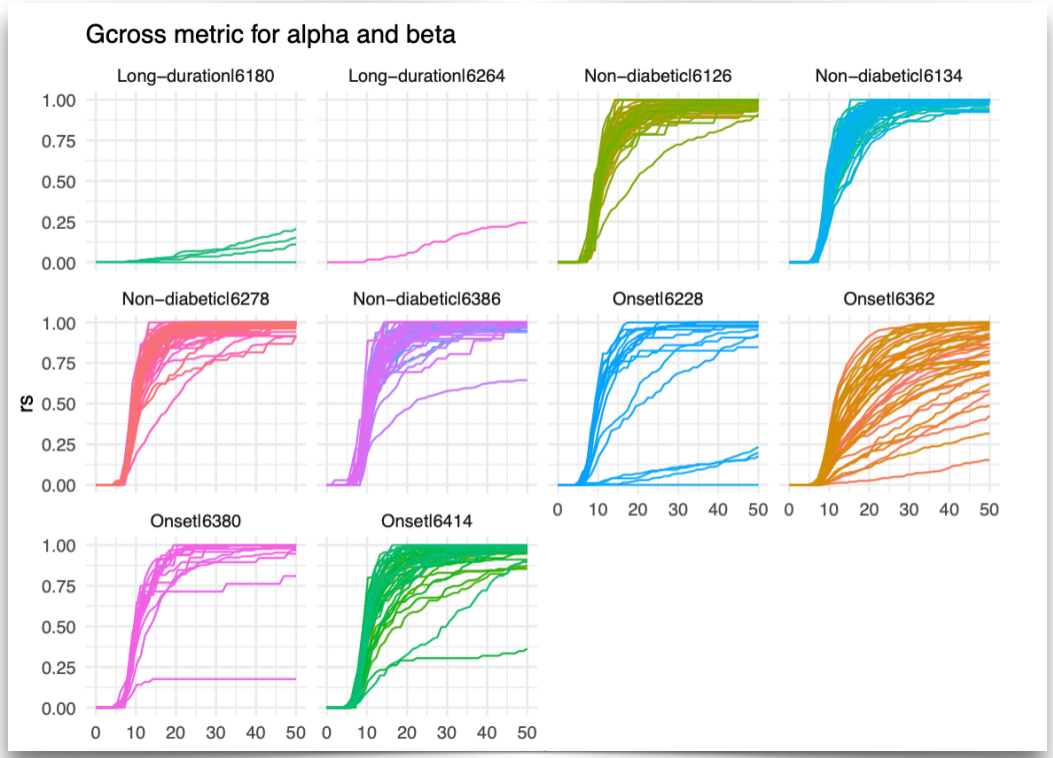
Gcross function across 10 samples (multiple FOVs per patient)

Martin

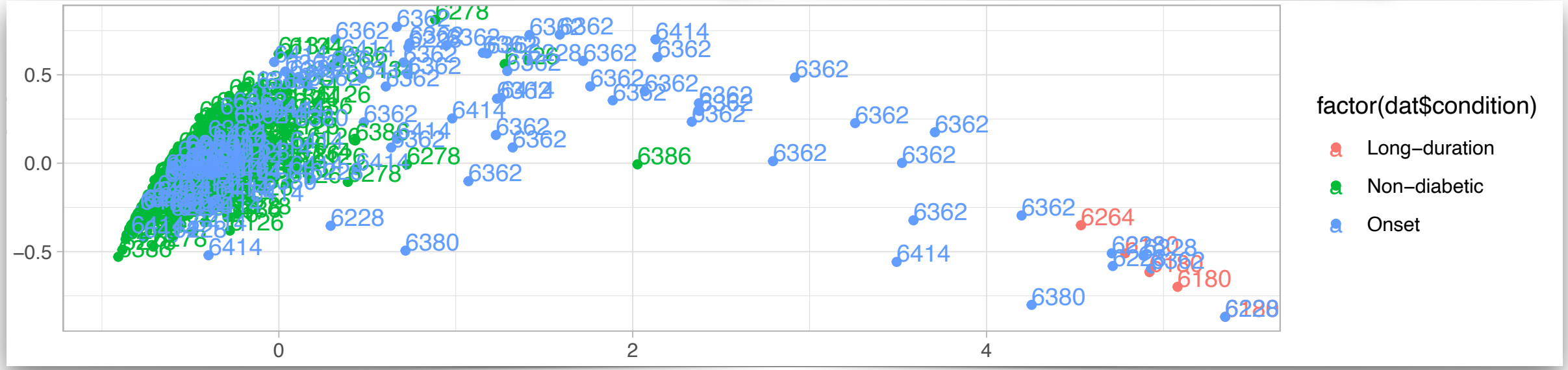
G-cross: represents the probability of finding at least one given beta cell within a radius r of any alpha cell



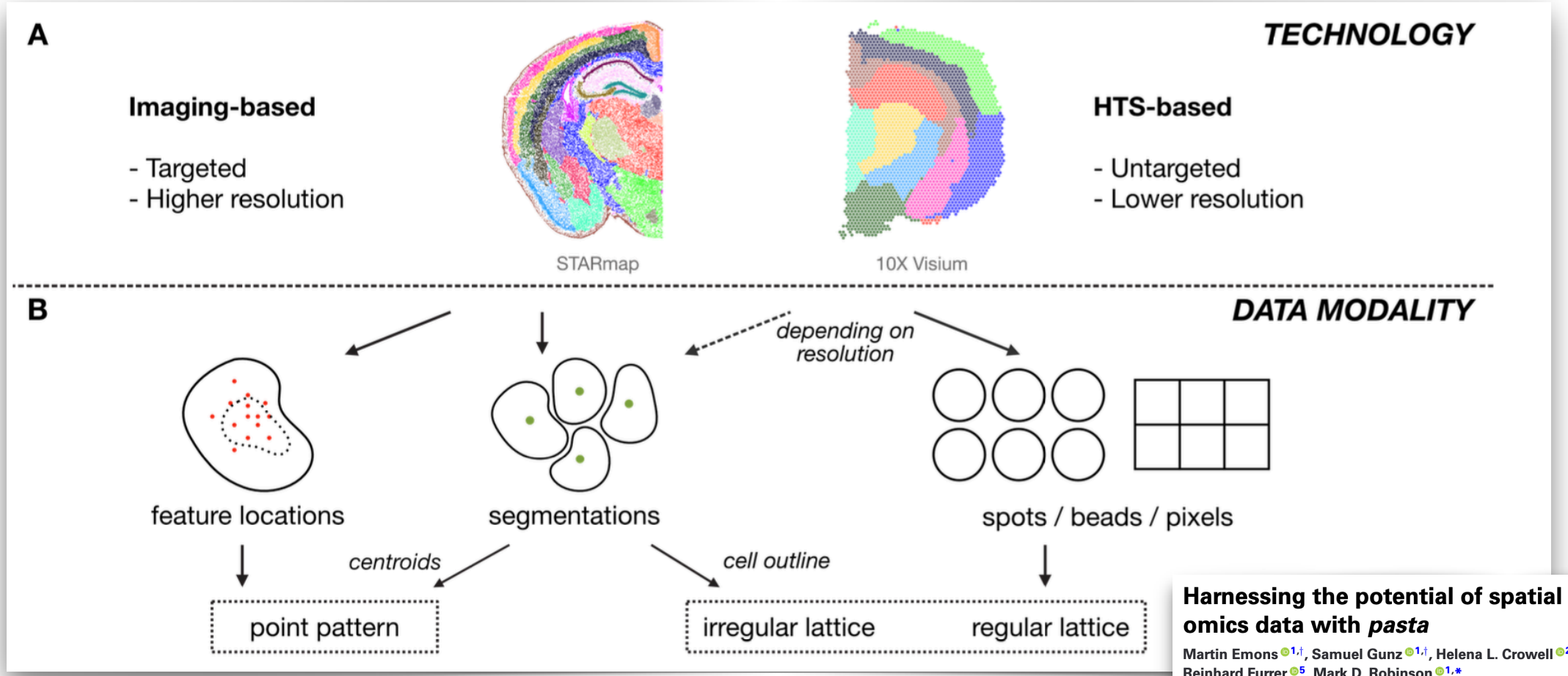
Functional PCA is already useful to decompose FOV-level data



Martin



`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.

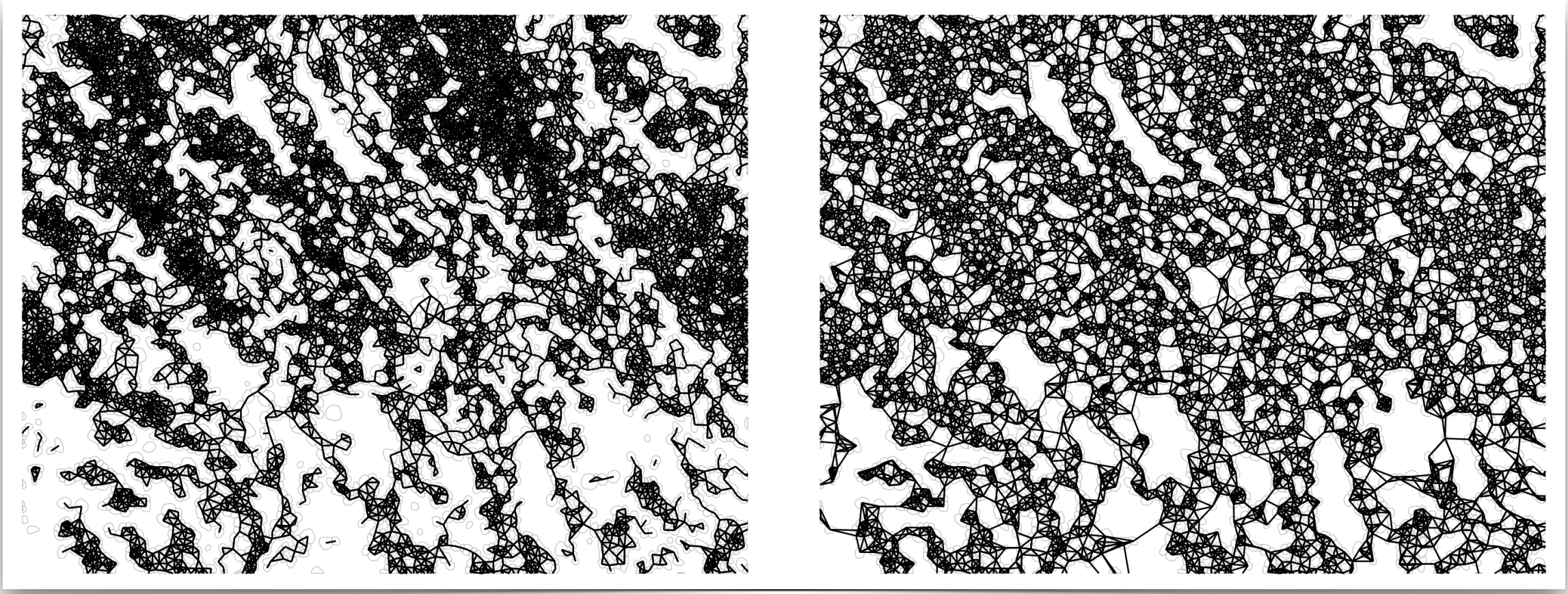
Global autocorrelation measures (Moran's I, Moran's R, Geary's C, etc.)

In general, a global spatial autocorrelation measure has the form of a double sum over all locations i, j

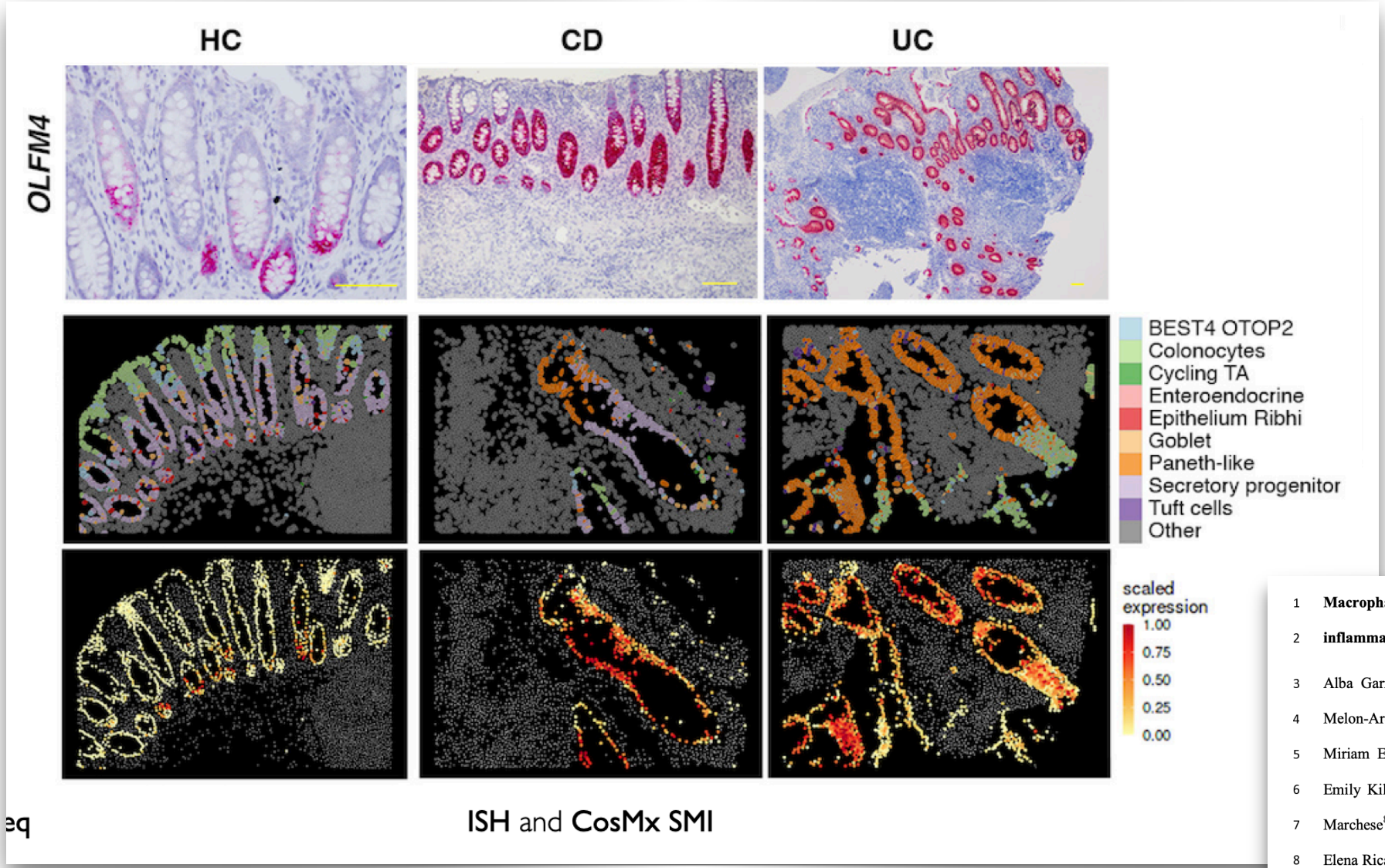
$$\sum_i \sum_j f(x_i, x_j) w_{ij}$$

where $f(x_i, x_j)$ is the measure of association between features of interest and w_{ij} scales the relationship by a spatial weight as defined in the weight matrix W. If i and j are not neighbours, i.e. we assume they do not have any spatial association, the corresponding element of the weight matrix is 0 (i.e., $w_{ij} = 0$).

Weight matrix options



Tissue “structures” are often visible

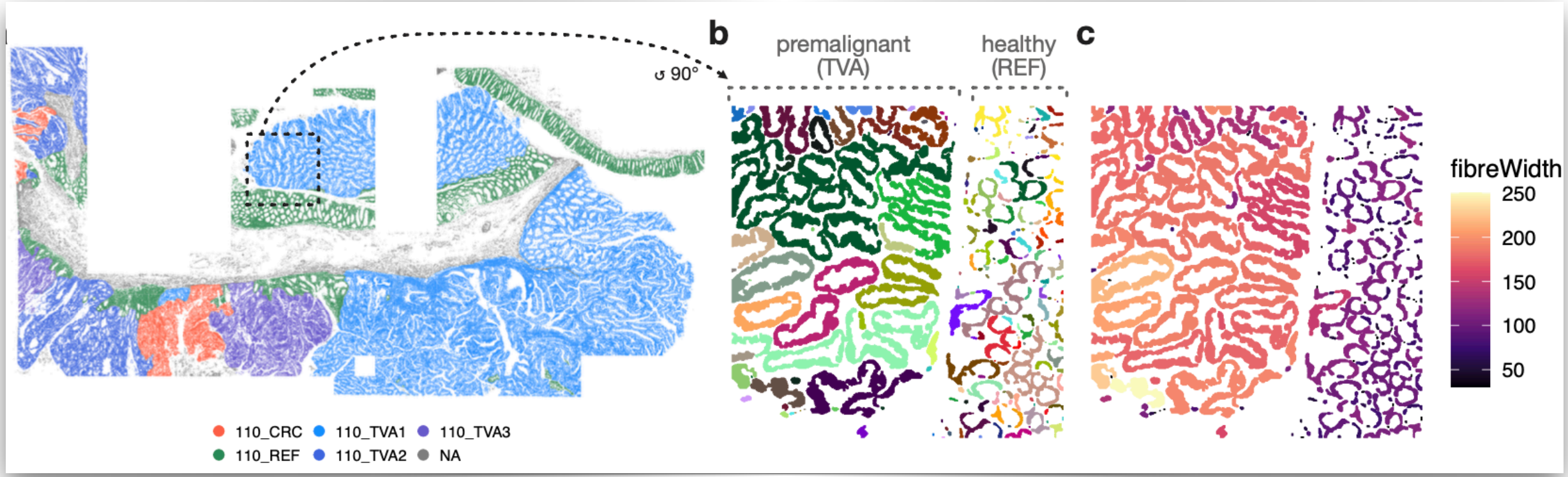


- healthy control (HC)
- Crohn's disease (CD)
- ulcerative colitis (UC)

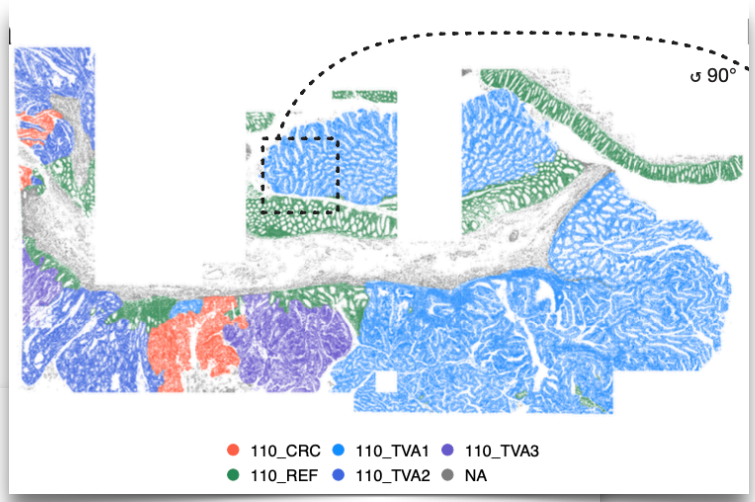
1 **Macrophage and neutrophil heterogeneity at single-cell spatial resolution in**
2 **inflammatory bowel disease**

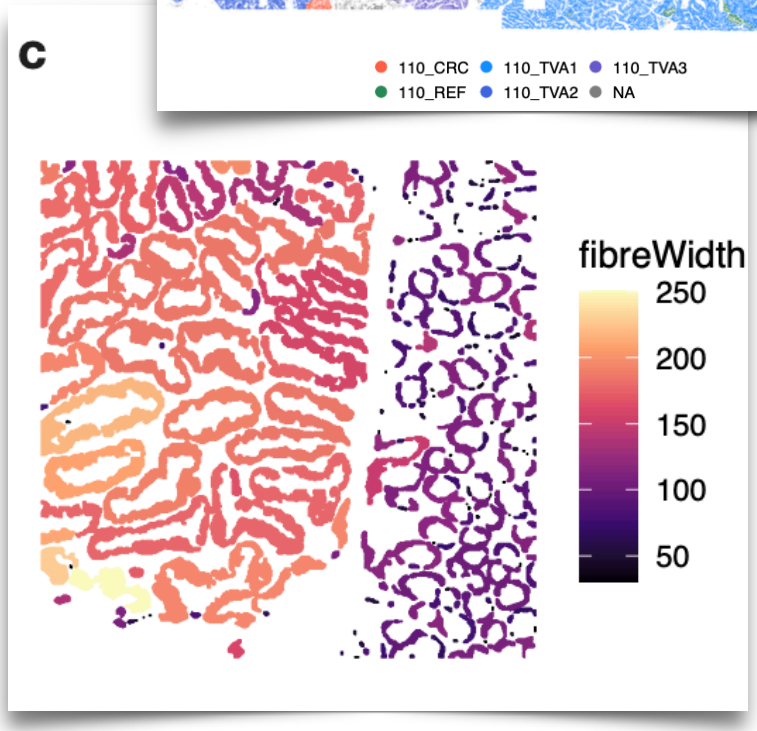
3 Alba Garrido-Trigo[1,2], Ana M. Corraliza[1,2], Marisol Veny[1,2], Isabella Dotti[1,2], Elisa
4 Melon-Ardanaz[1,2], Aina Rill[3], Helena L. Crowell[4], Ángel Corbí[5], Victoria Gudiño[1,2],
5 Miriam Esteller[1,2], Iris Álvarez-Teubel[1,2], Daniel Aguilar[1,2], M Carme Masamunt[1,2],
6 Emily Killingbeck[6],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

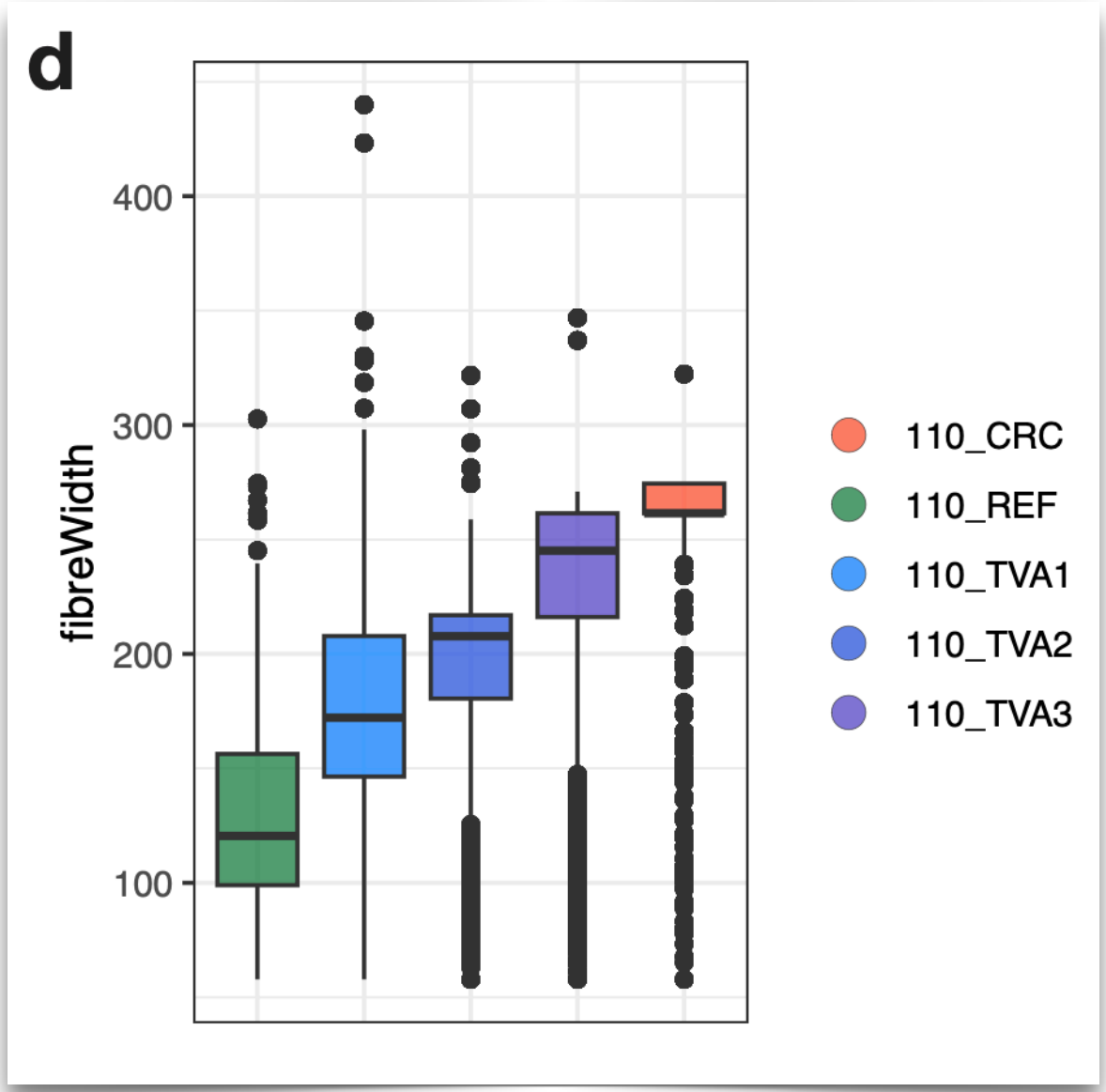
Variation among spatial structures (epithelial example)



Variation among spatial structures (geometric quantifications)







Structures —> Reference axis —> Expression gradients