Statistical methods for spatial omics data

- Overview on the technologies (review)
- (Some) Fundamentals of spatial statistics
  - Point patterns: random, clustered, intensity/correlation
  - Useful summaries / functions
  - models with spatially correlated errors
- Finding spatially-variable genes
- Deconvoluting low-resolution spatial omics data
- Spatial clustering
- Cell-cell communication
- Integration w/ single cell RNA-seq
- (Segmentation, preprocessing)

https://www.nature.com/articles/d41586-022-04426-5

Interesting in the context of this course .. and I was interviewed

TECHNOLOGY FEATURE | 13 December 2022

Which single-cell analysis tool is best? Scientists offer advice

In the fast-paced field of single-cell biology, studies that compare methods can help scientists to pick the right technique for their research.

Amber Dance

Note: link is wrong: omnibenchmark.org is the correct one!

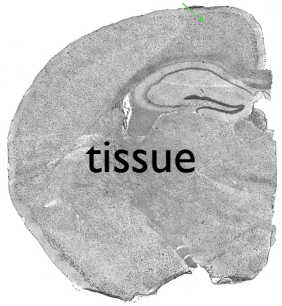
Background

Slide from Helena Crowell

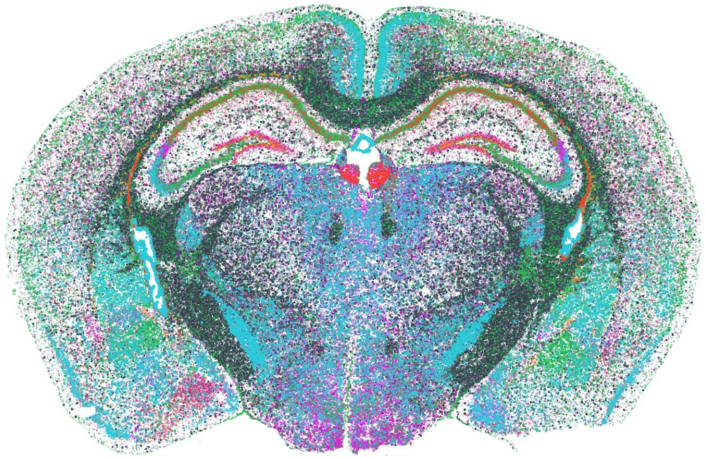
bulk

single-cell

spatial

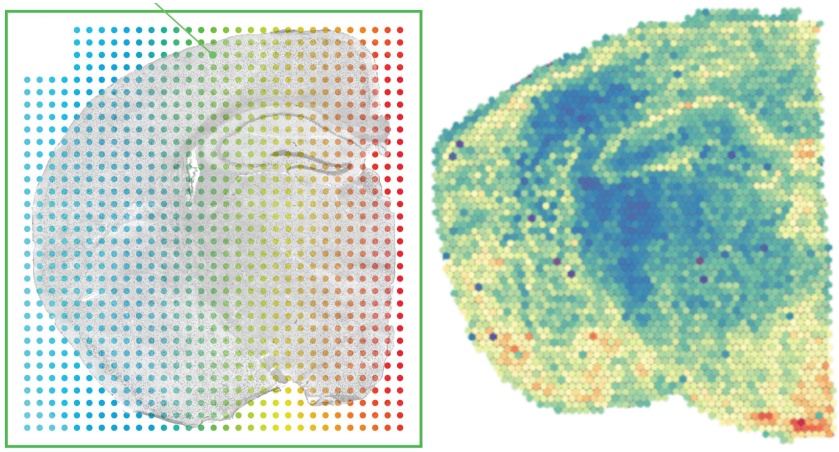


imaging-based



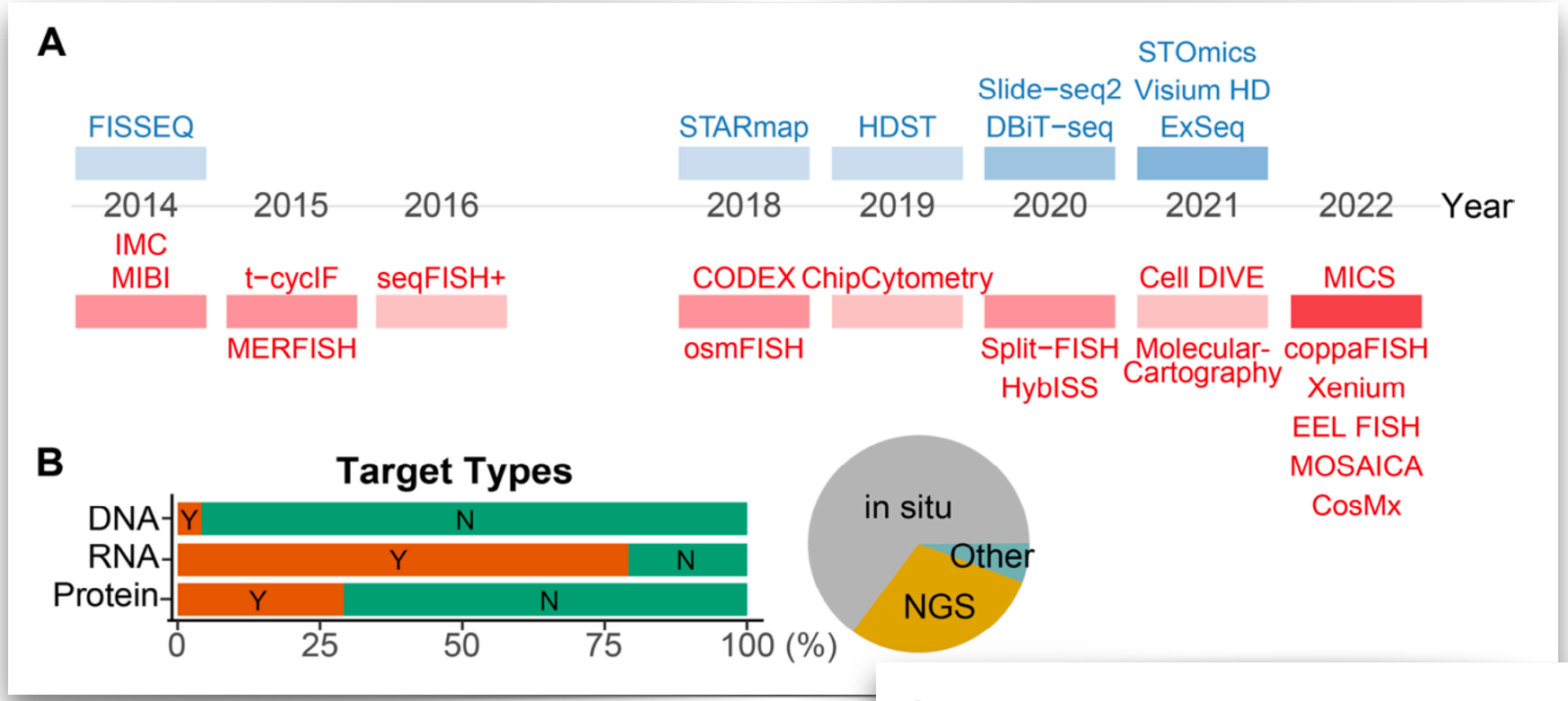
- molecule-level data
- targeted panel (100s of features)
- single-cell resolution requires segmentation

sequencing-based

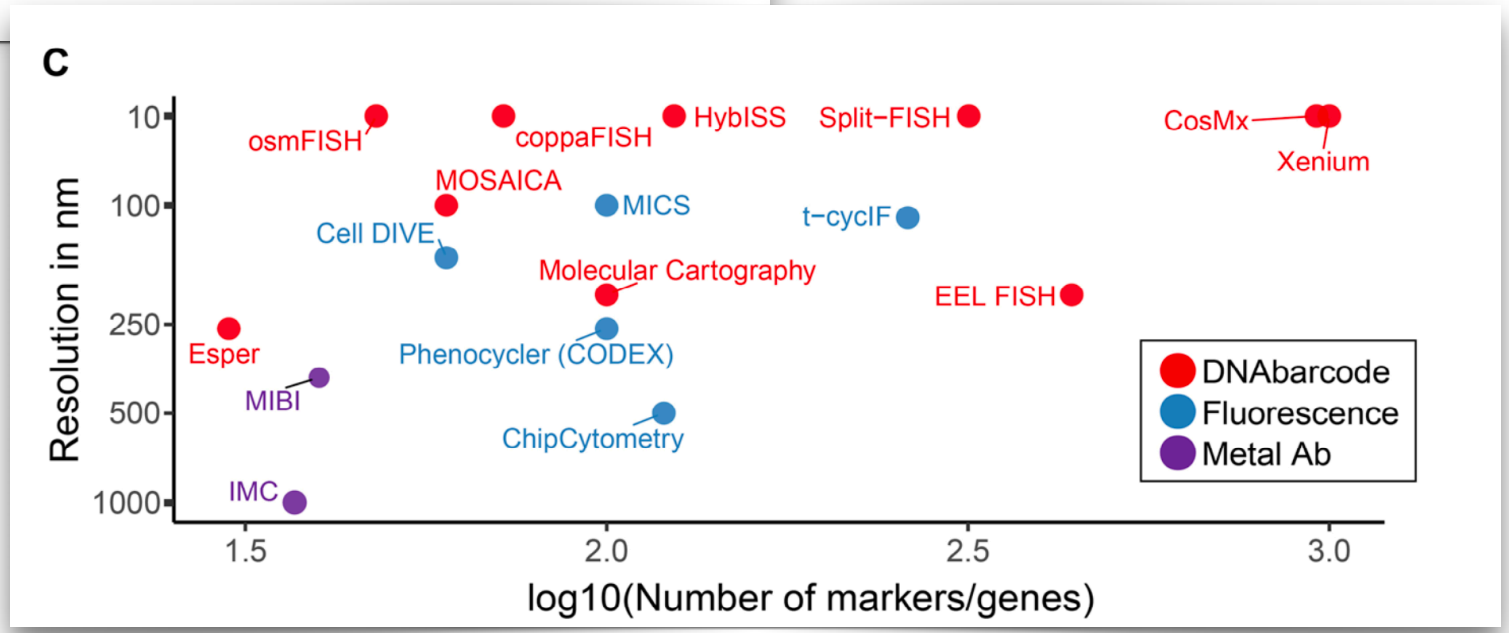


- spot-level data
- whole transcriptome (10,000s of features)
- single-cell resolutions requires aggregation or deconvolution

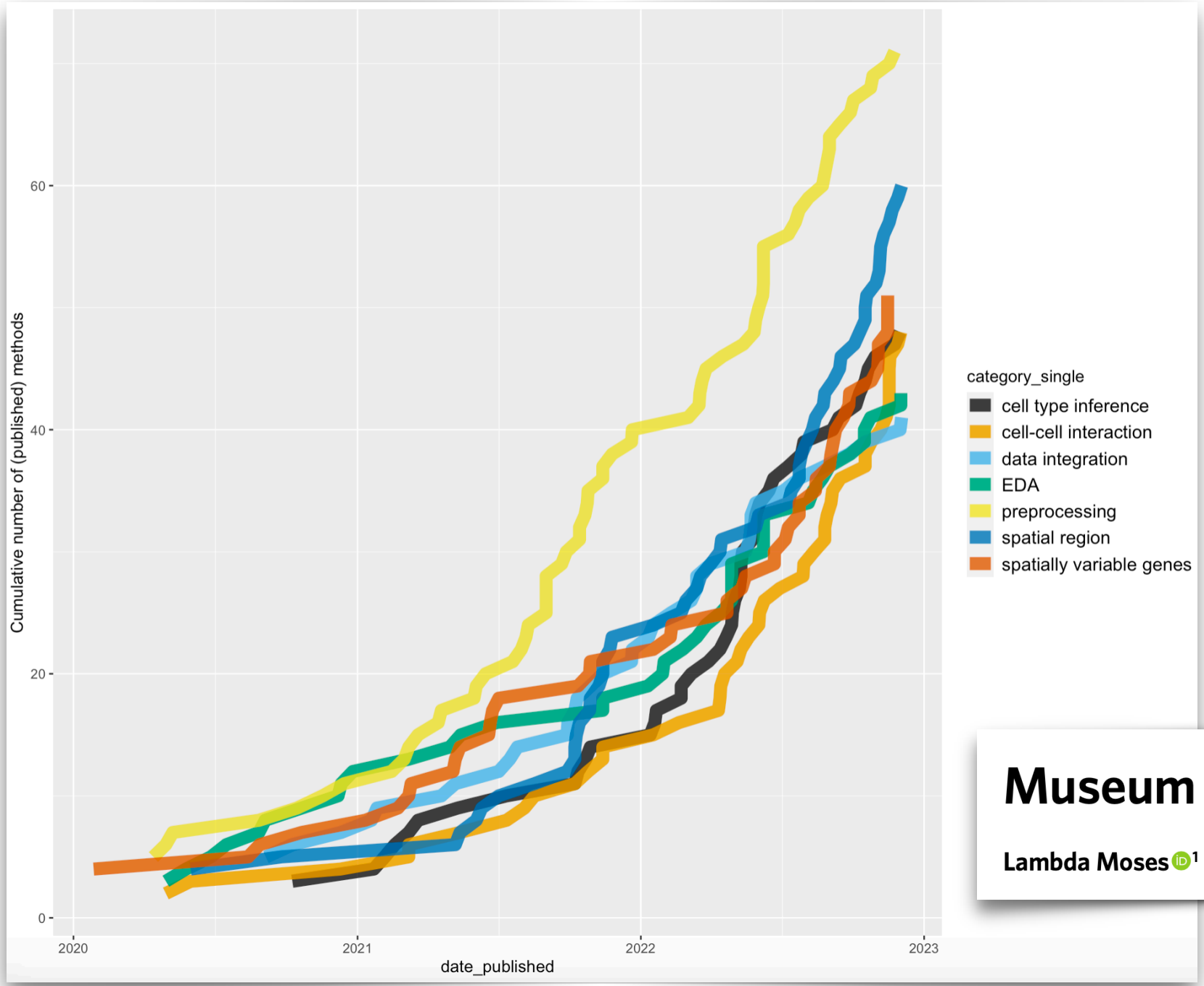
(Spatial omics) Technology explosion



blue colored techniques are sequencing-based while red colored techniques are multiplexed IHC/IF methodologies



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-022-02824-6

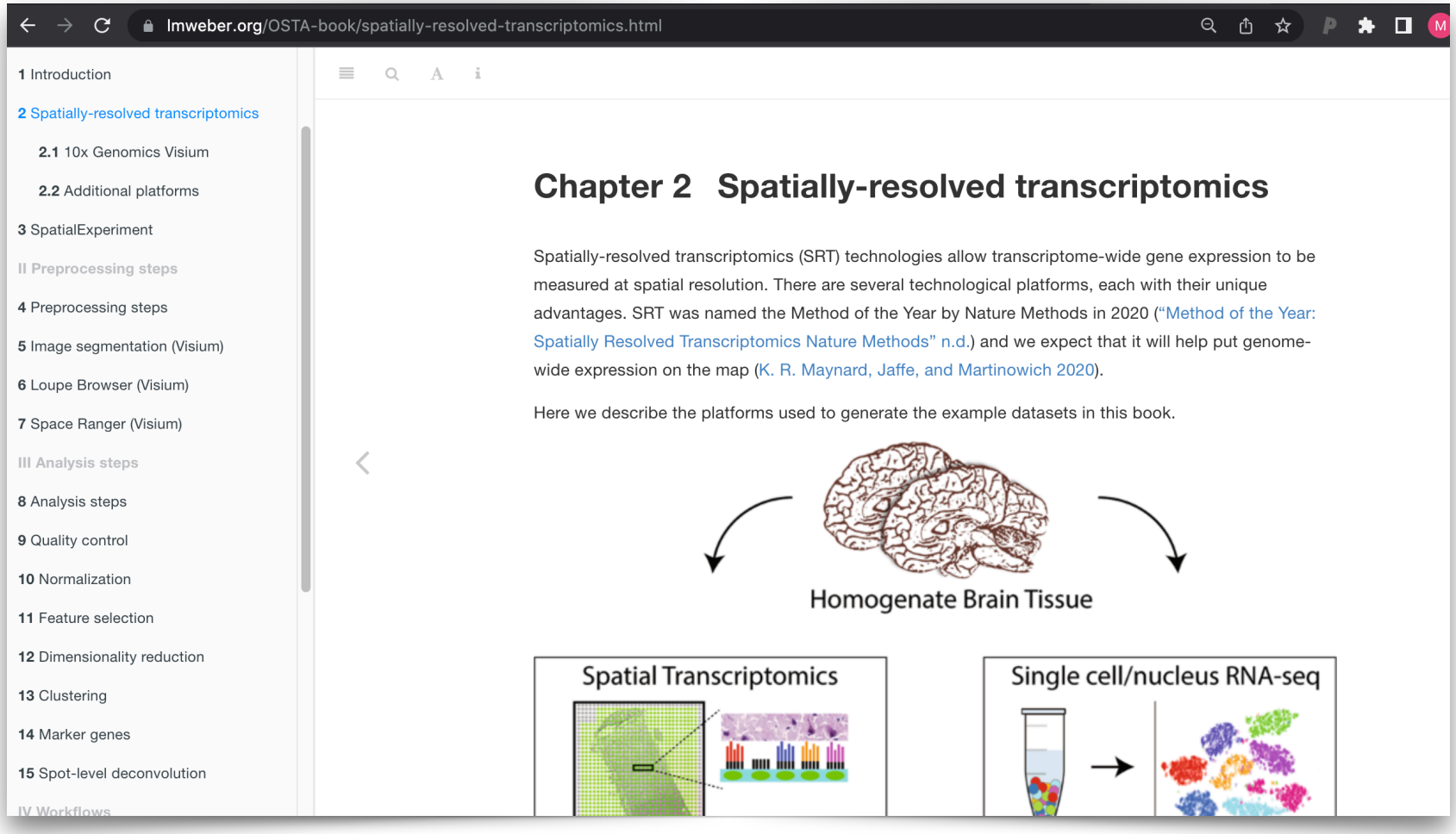


(Spatial omics) computational method explosion

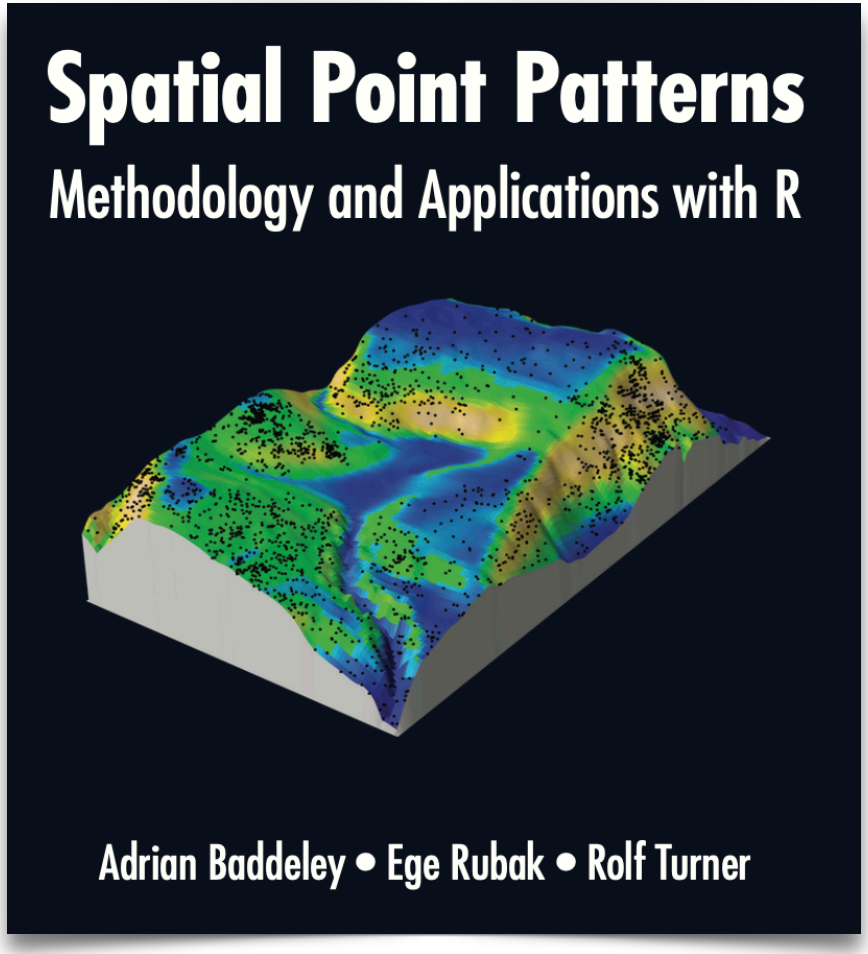
Museum of spatial transcriptomics

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

Data processing for SRT (spatially-resolved transcriptomics) data: OSCA —> OSTA



Statistical methods for spatial omics data



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

- Point patterns (definition): intensity, homogeneity, dependence
- Multi-type point patterns
- Marked point processes
- Statistical summaries
- Covariates versus marks
- (Discrete random fields)

– "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

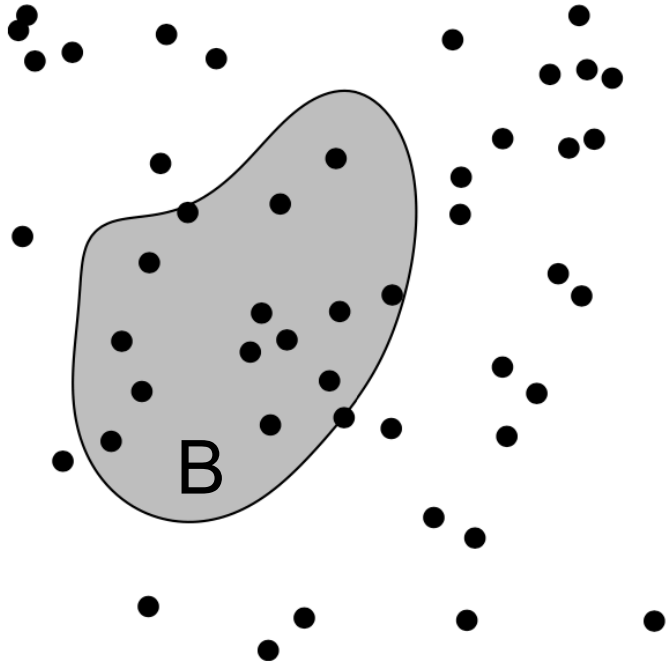
- notation: **X** is the point process; **x** is the (observed) point pattern
- lambda: intensity function

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.

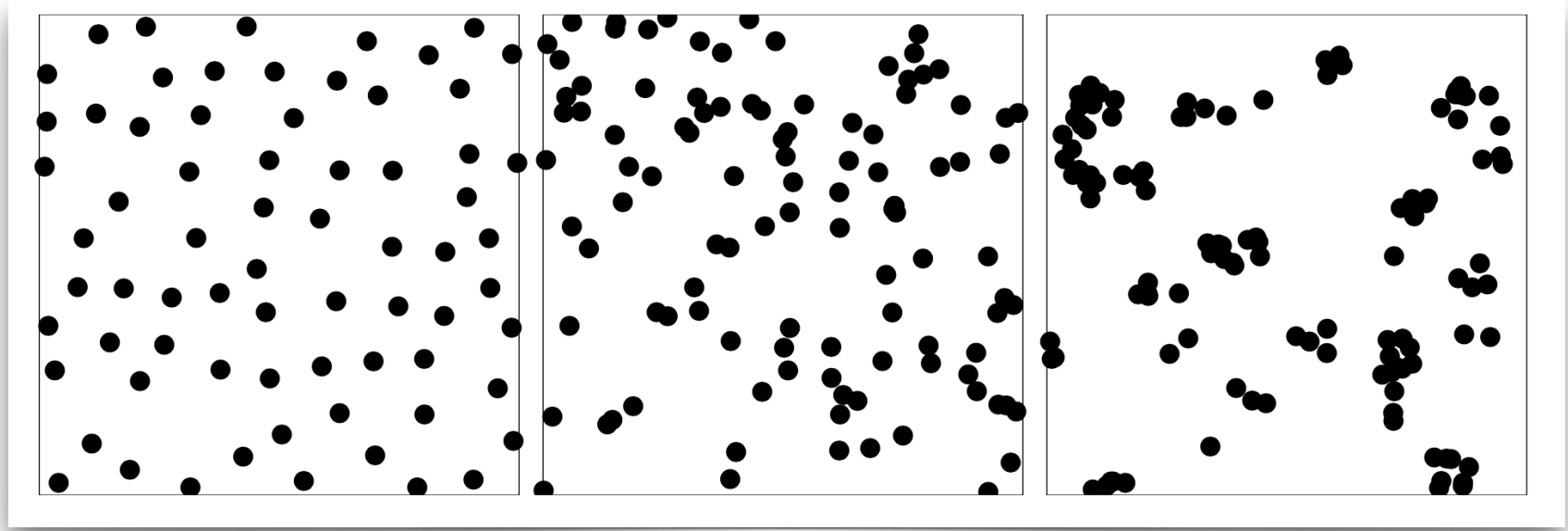
- More specifically:

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;

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?



Couple more definitions

– Inhomogeneity

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

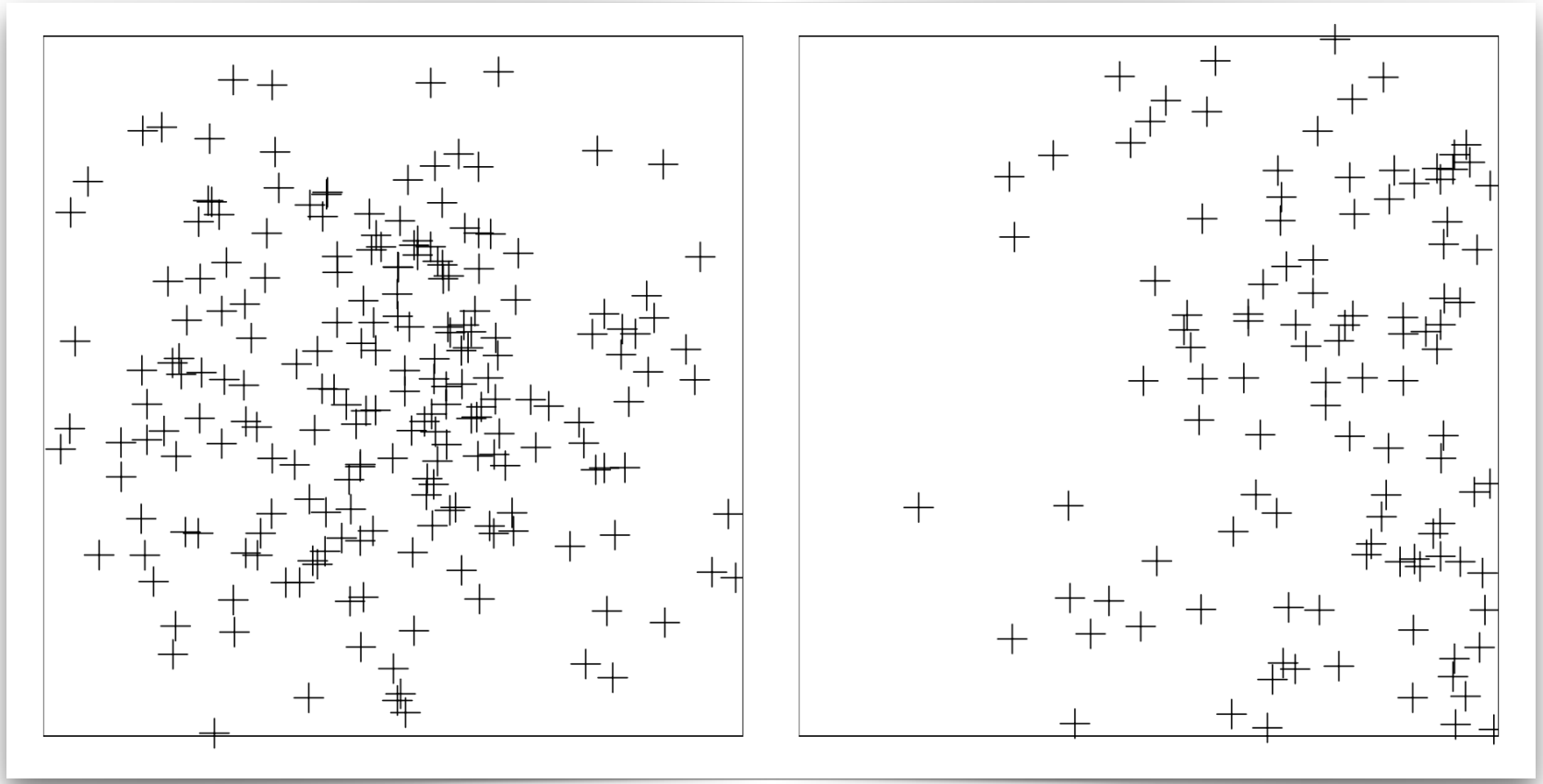
intensity function: the expected number of points falling in a region B is the integral $\mu = \int_B \lambda(u)\,du$ of the intensity function $\lambda(u)$ over the region B;

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;

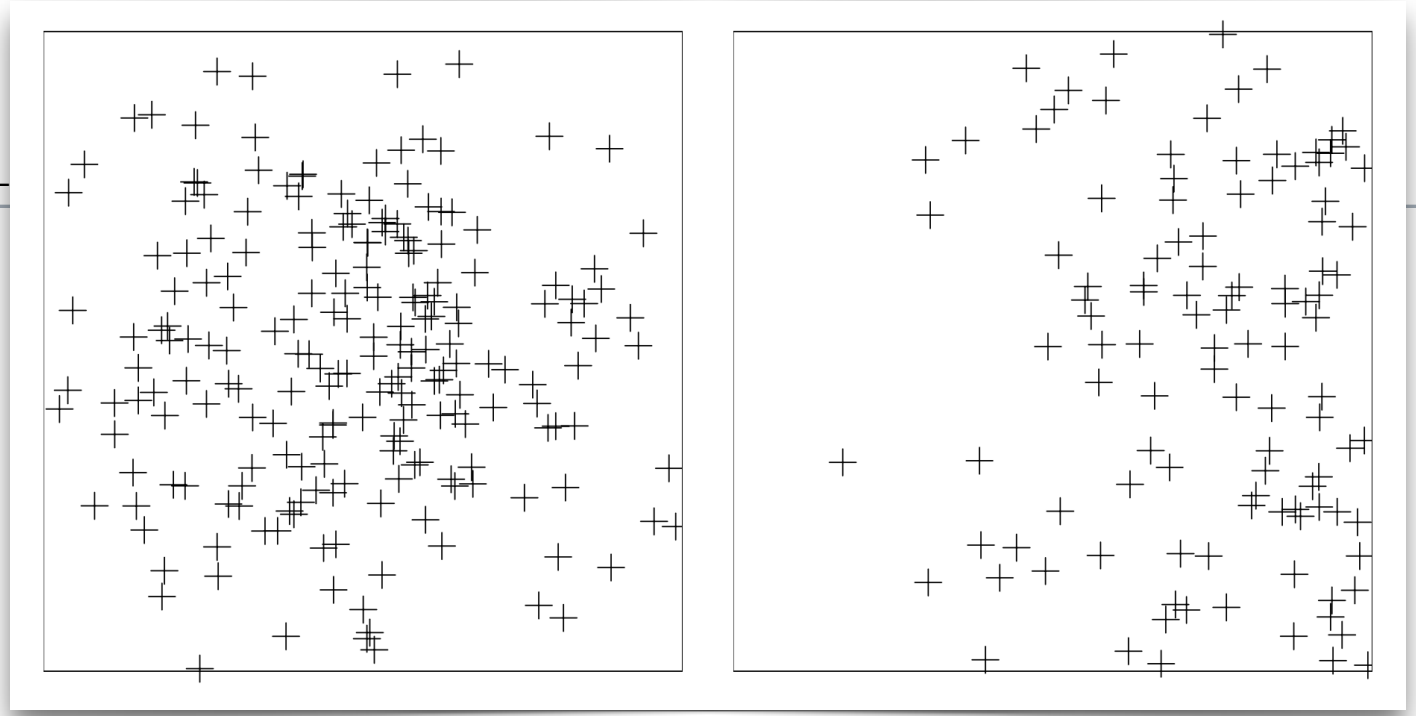
Intensity of a process

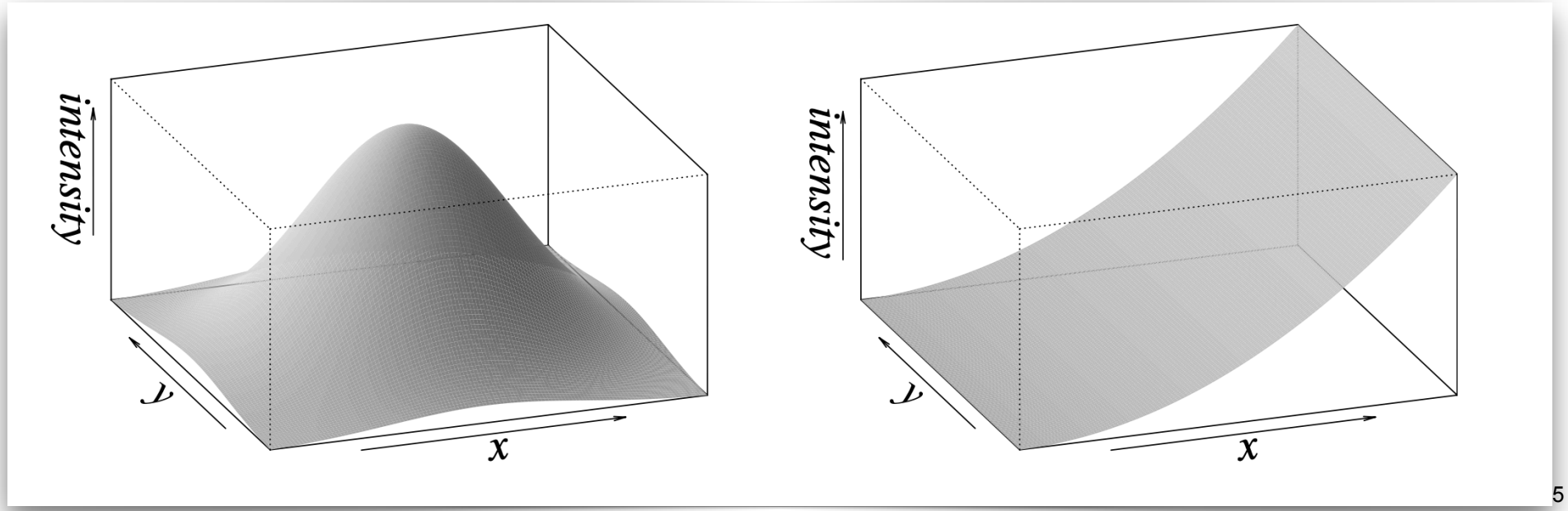
– Which of these processes is homogeneous?



Intensity estimation

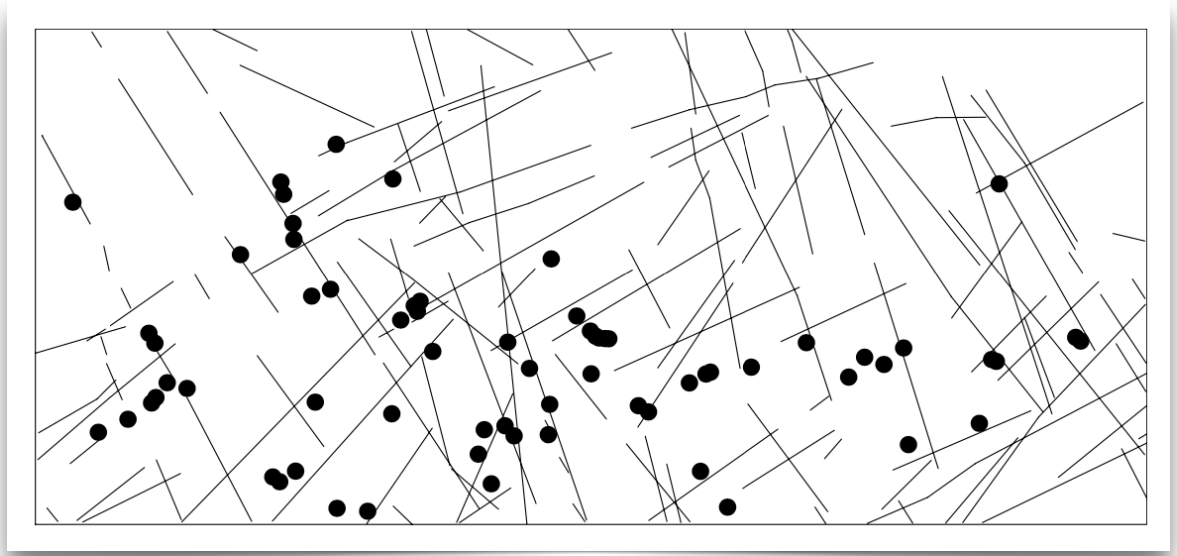
- Methods for intensity estimation

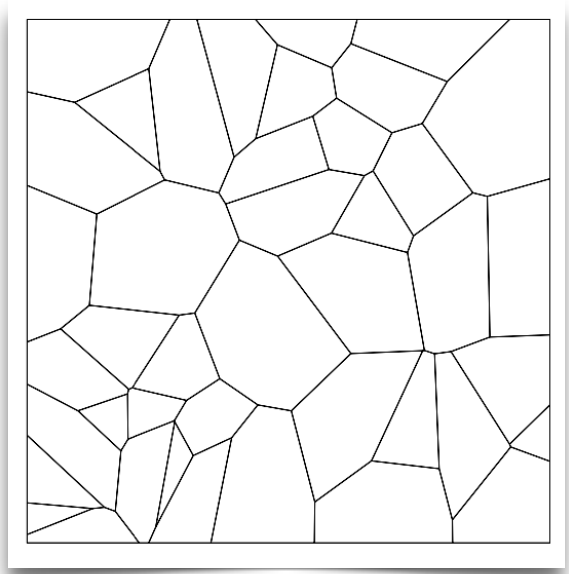




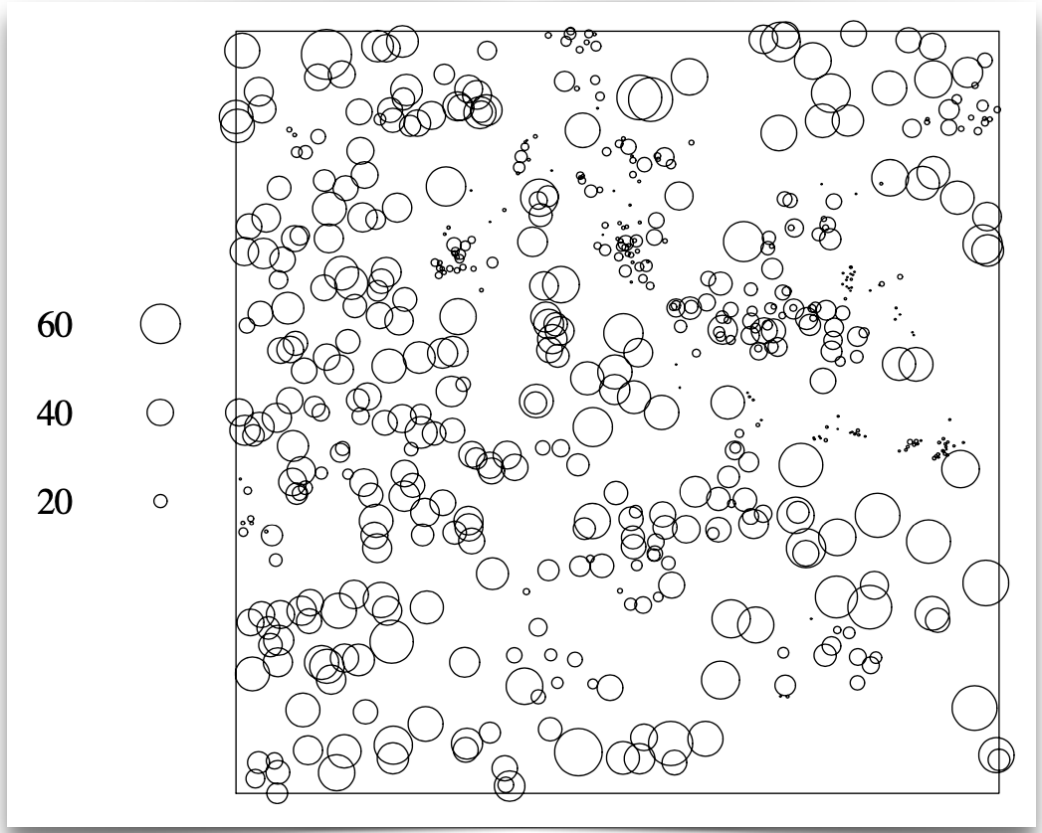
Extensions of point patterns

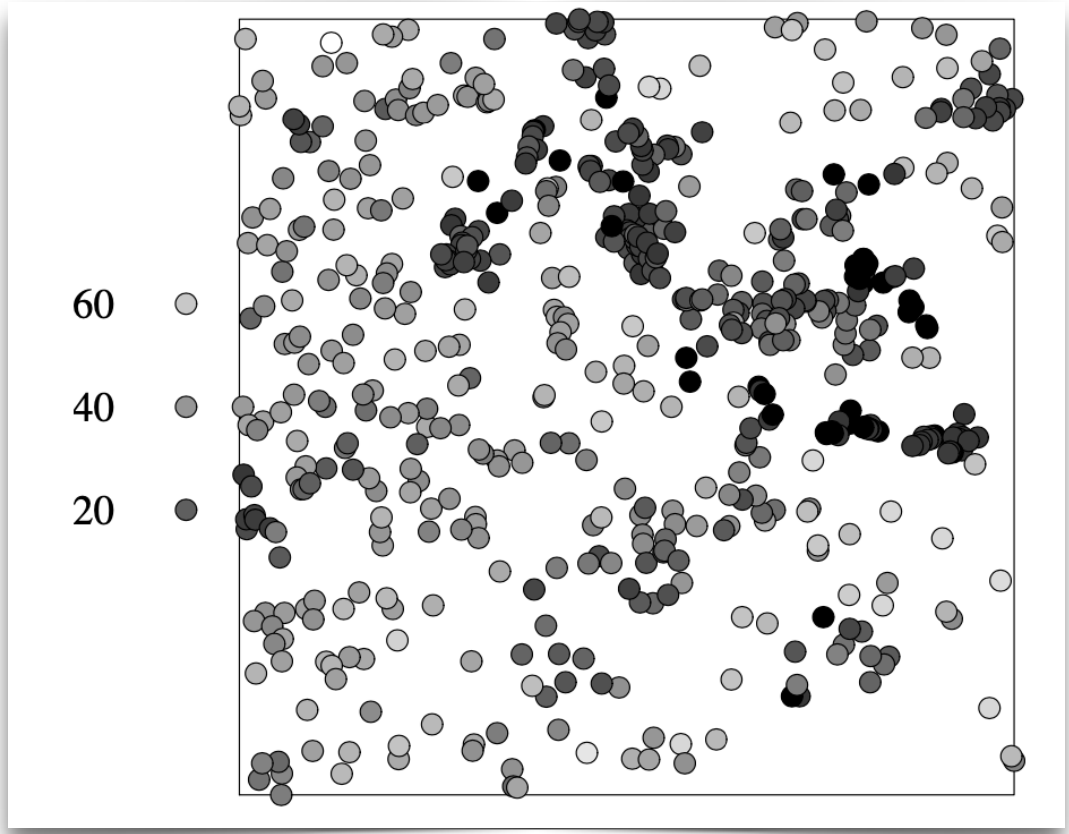
- (patterns can be regions/lines, not just points; can be in higher dimension (e.g., 3D); temporal component .. most of the methods I discuss have extensions; not discussed here)
- marks —> marked point process
- types —> multi-type point process
- covariates





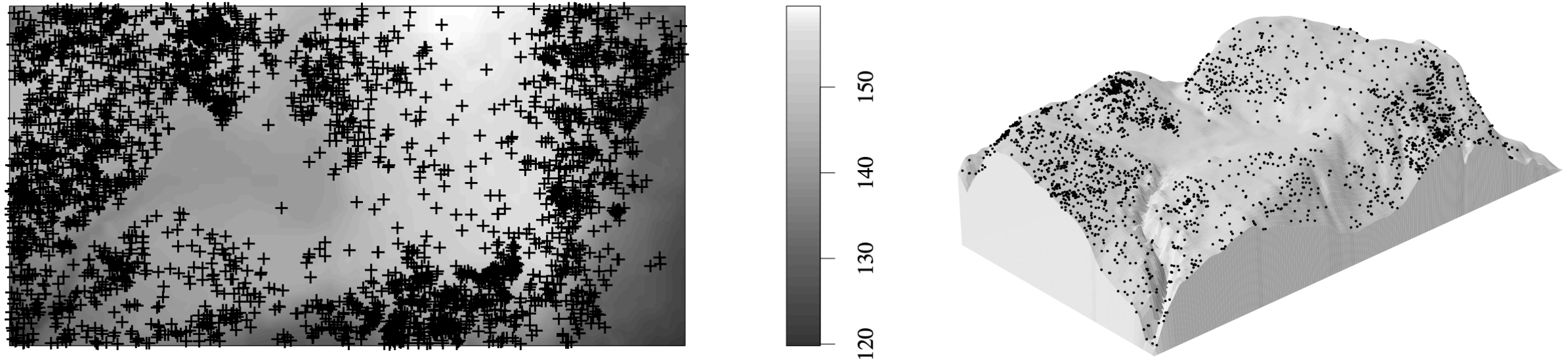
Marks





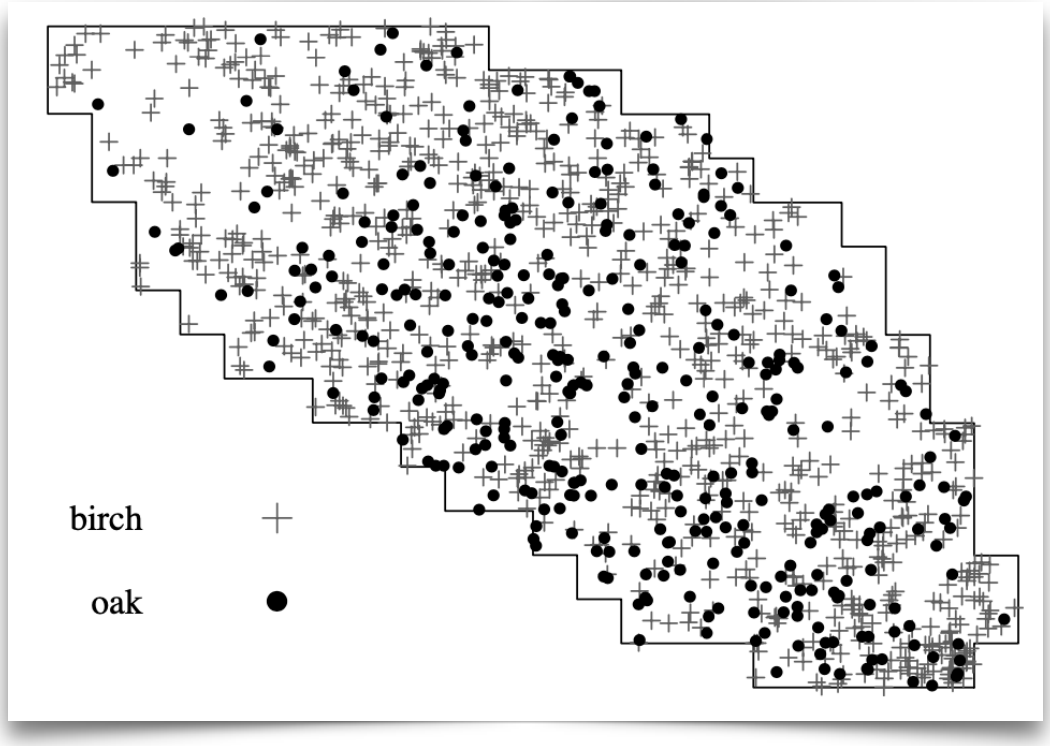
Longleaf pines dataset (location and diameter)

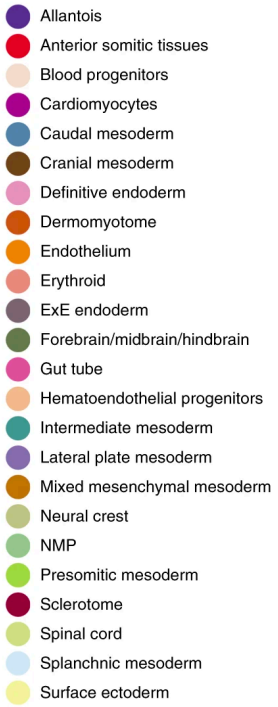
Covariates

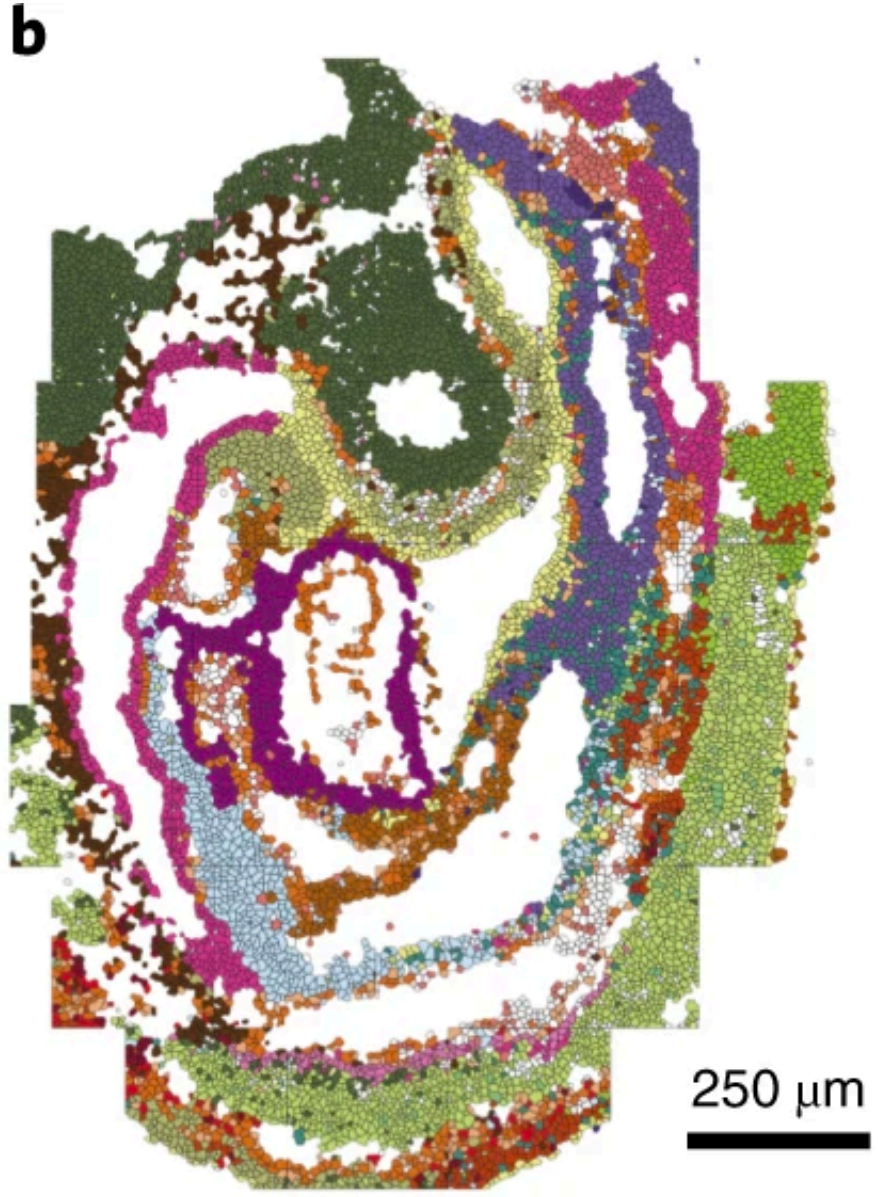


Locations of Beilschmiedia pendula *trees (+) and terrain elevation (greyscale) in a* 1000 × 500 *metre survey plot in Barro Colorado Island.*

Multi-type point patterns



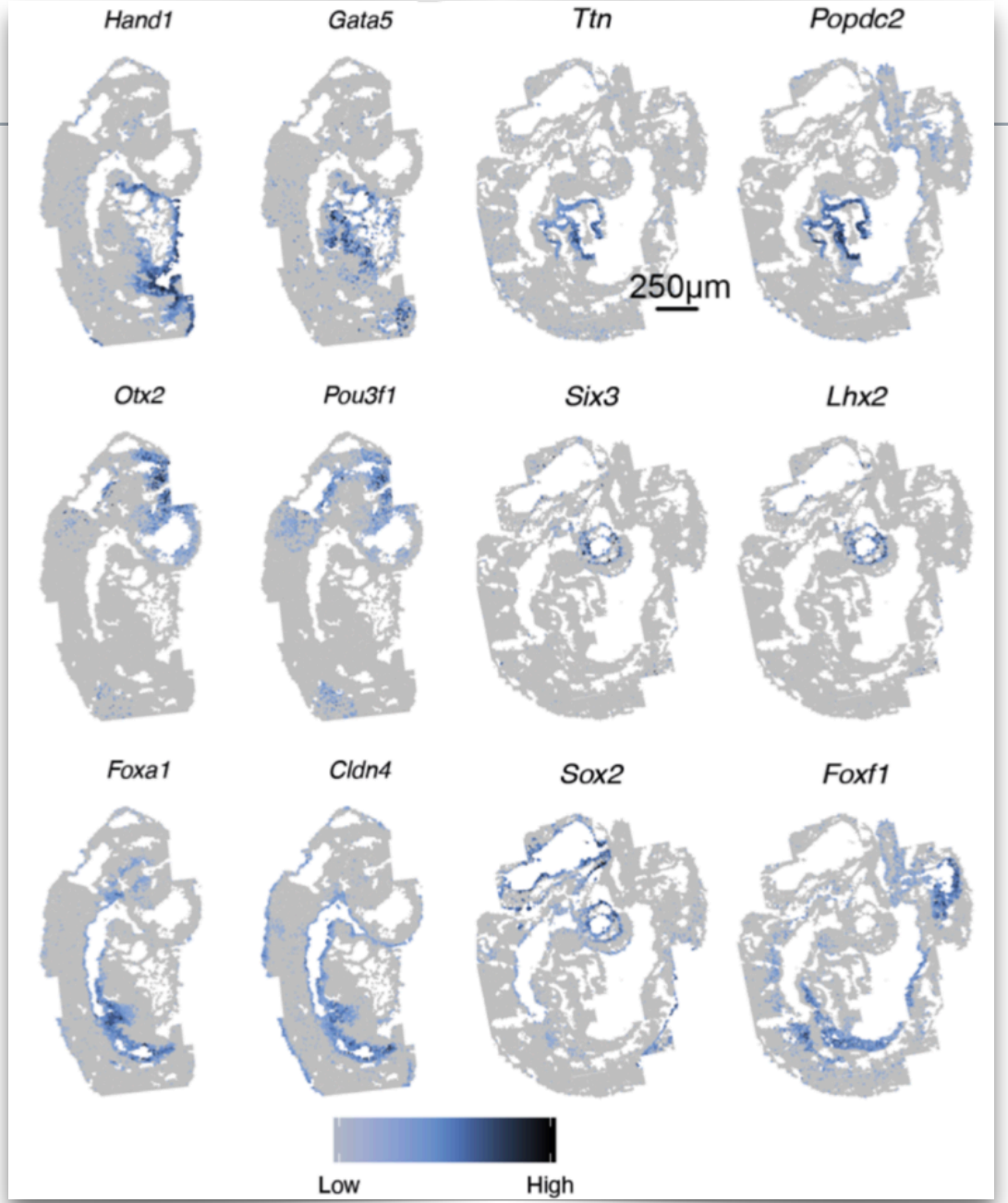




https://www.nature.com/articles/s41587-021-01006-2

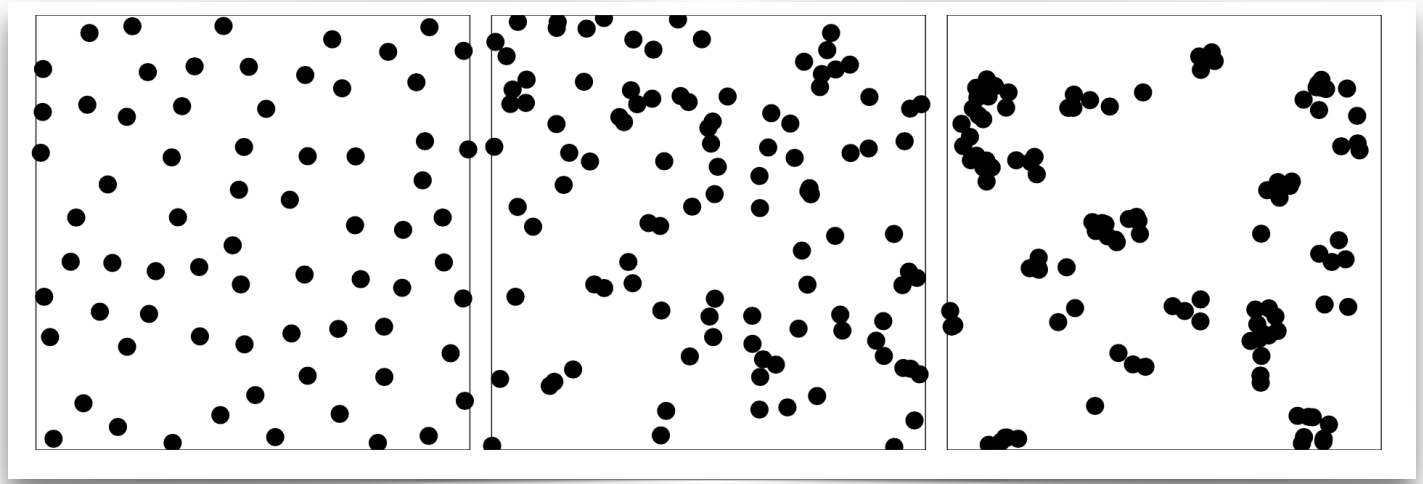
Imagine a spatially-resolved omics dataset

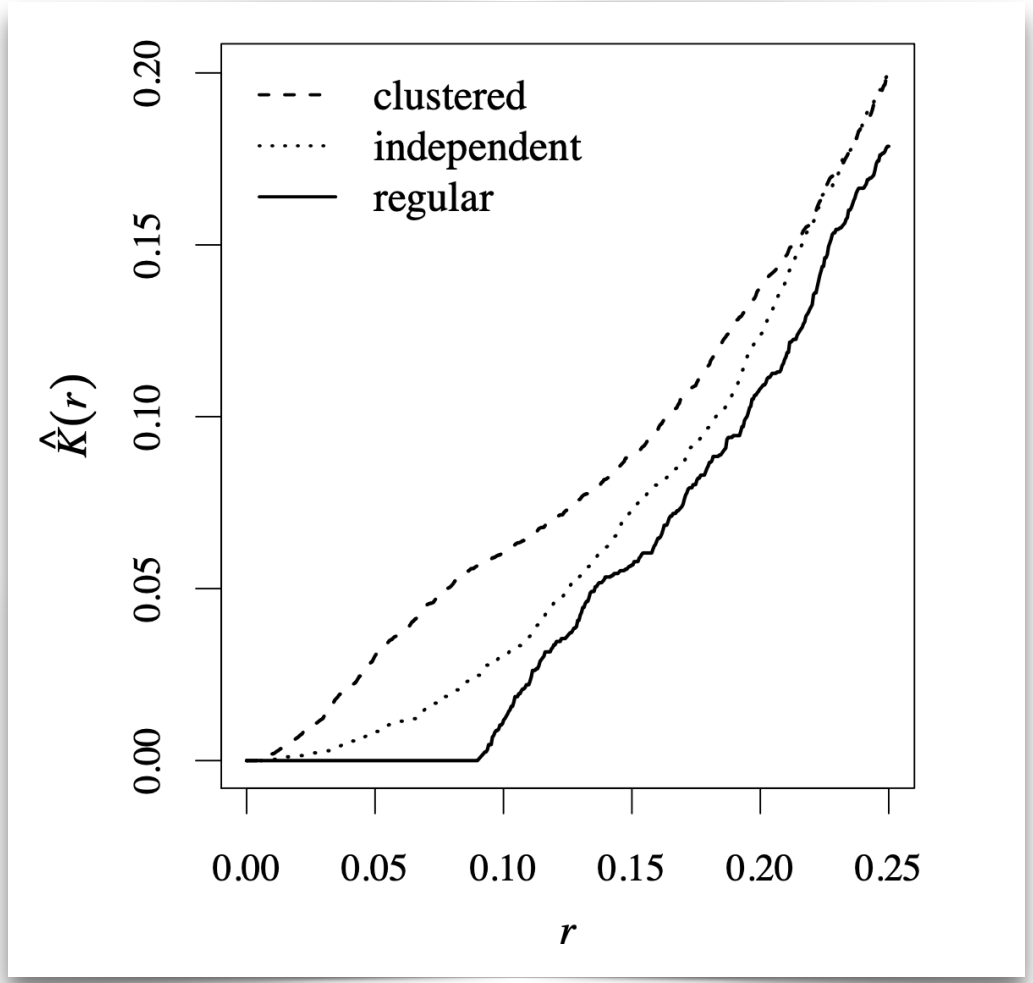
- Say you can segment cells and quantify gene/expression ..
- What are the expression measurements? Types, marks, or covariates



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:

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

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

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

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

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

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

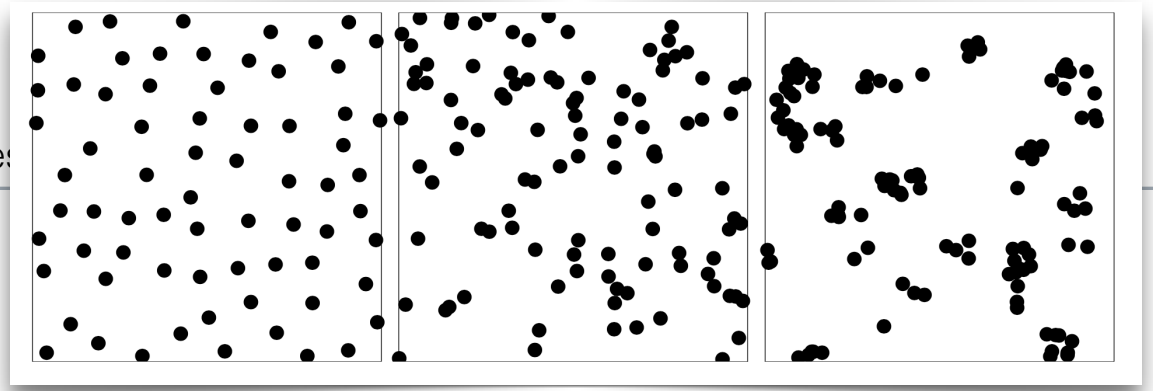
Extensions of the K function (1)

– 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.

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

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

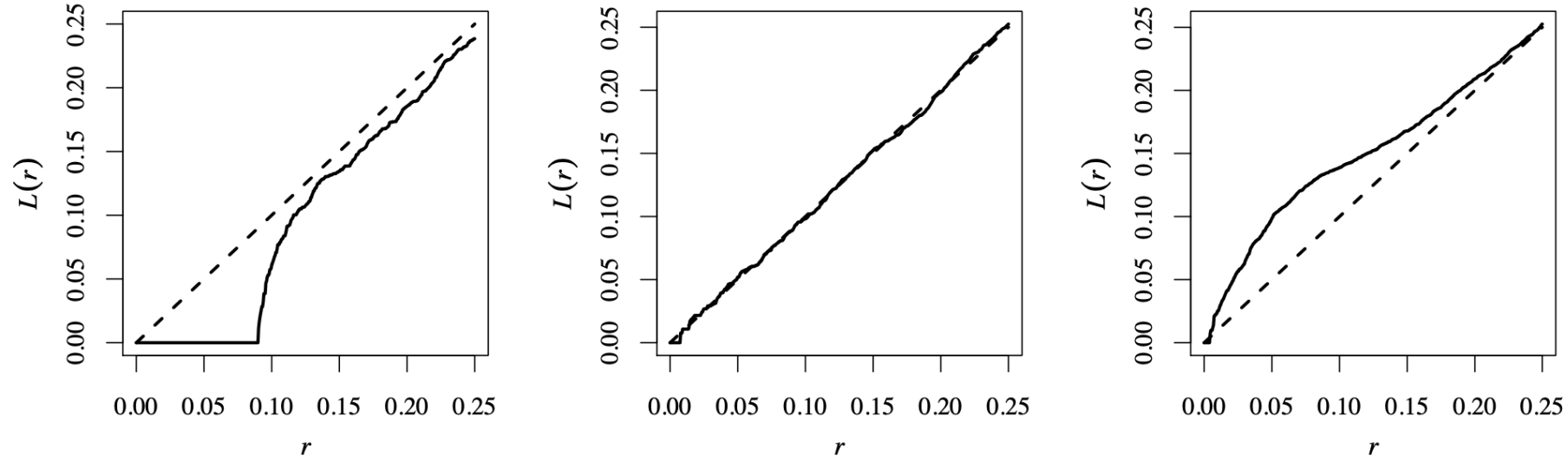
Extensions of the K function (2)

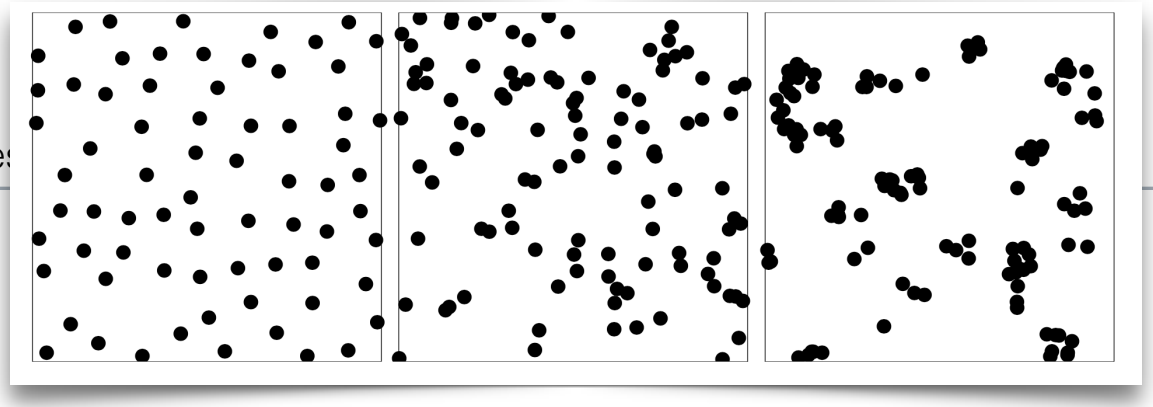


– L and g functions

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

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



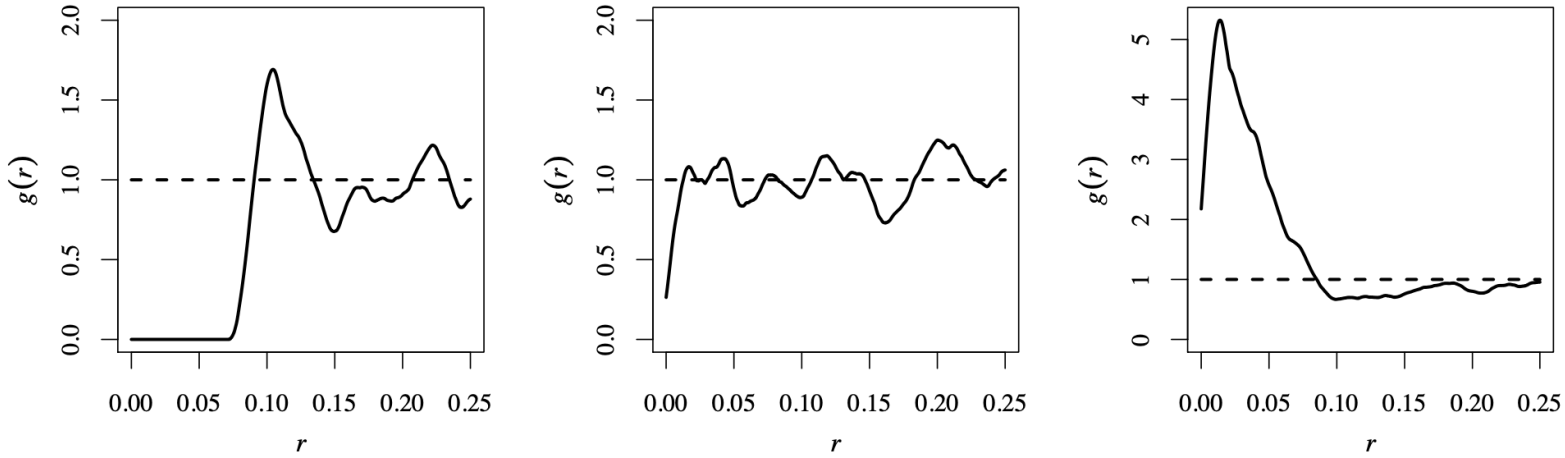


Extensions of the K function (2)

– 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

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



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

Spatial autocorrelation: Moran's I

- Global measure of auto-correlation (correlation to signal nearby in space); assume homogeneity!
- Alternative: Geary's C
- Local measures also exist: LISA (local indicators of spatial association)

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

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

Statistical methods for spatial omics data

- Overview on the technologies (review)
- (Some) Fundamentals of spatial statistics
  - Point patterns: random, clustered, intensity/correlation
  - Useful summaries / functions
  - models with spatially correlated errors
- Finding spatially-variable genes
- Deconvoluting low-resolution spatial omics data
- Spatial clustering
- Cell-cell communication
- Integration w/ single cell RNA-seq
- (Segmentation, preprocessing)

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

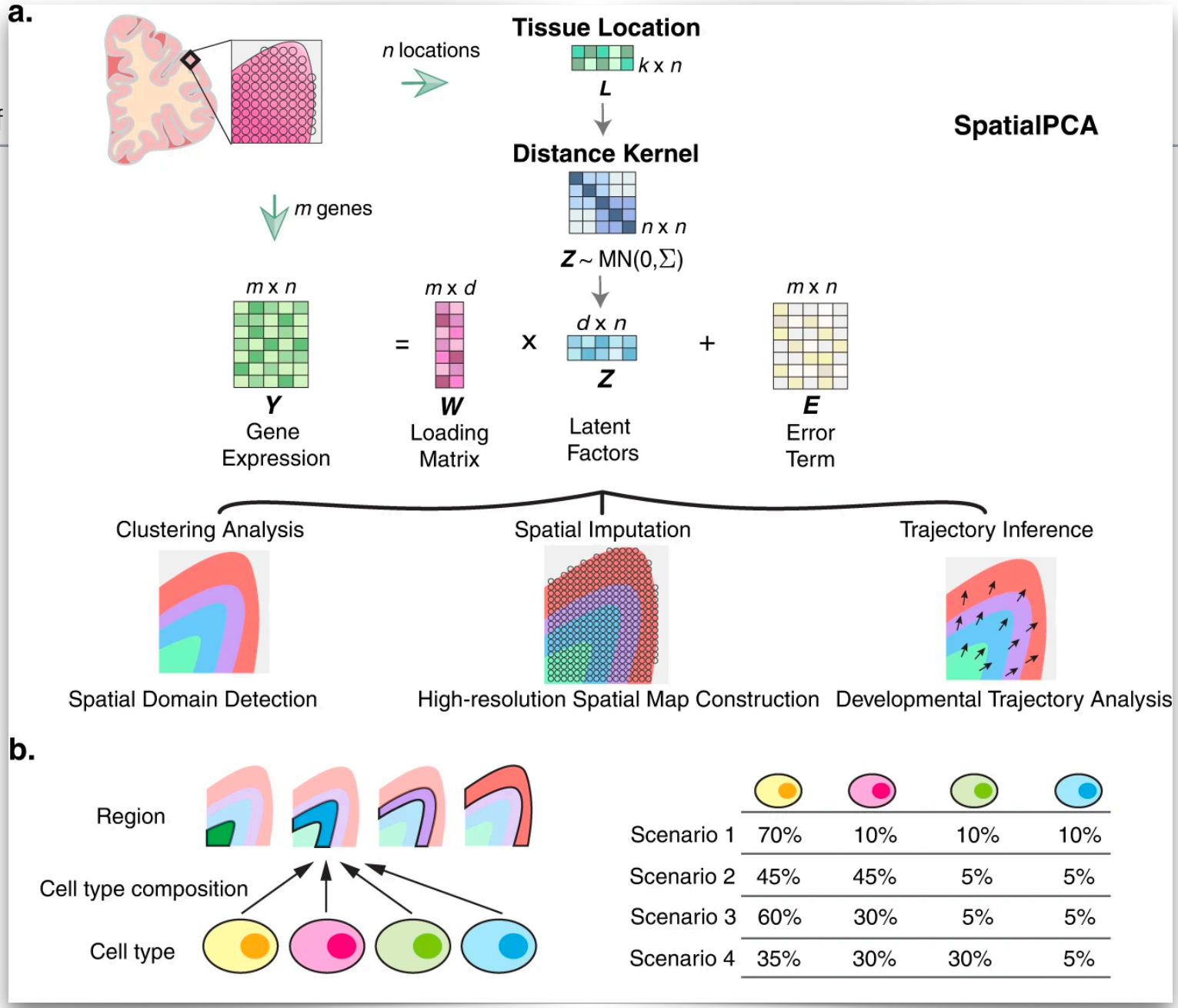
SpatialDE model. SpatialDE models gene expression profiles $y = (y_1, \ldots, y_N)$ for a given gene across spatial coordinates $X = (x_1, \ldots, x_N)$, using a multivariate normal model of the form

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

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:

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

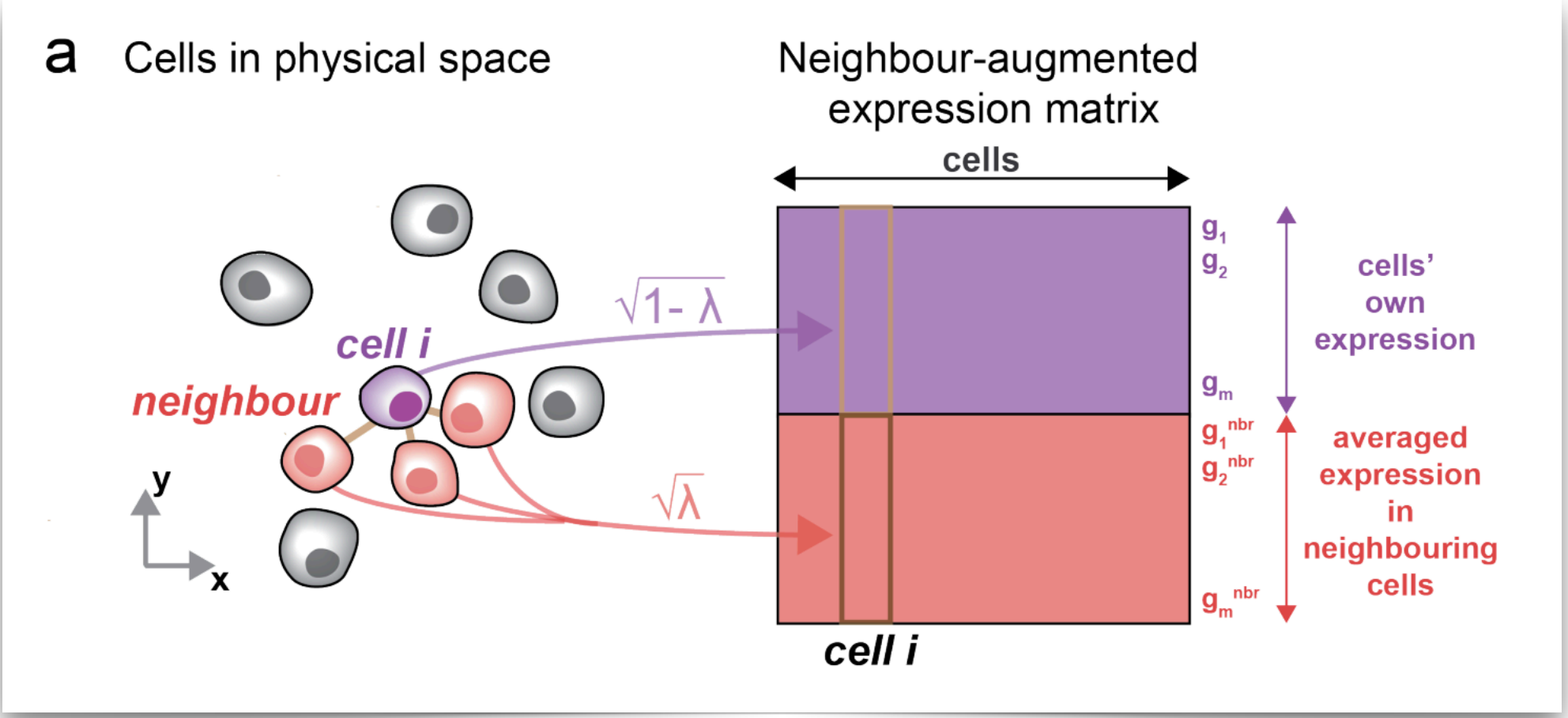
Spatial domains



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

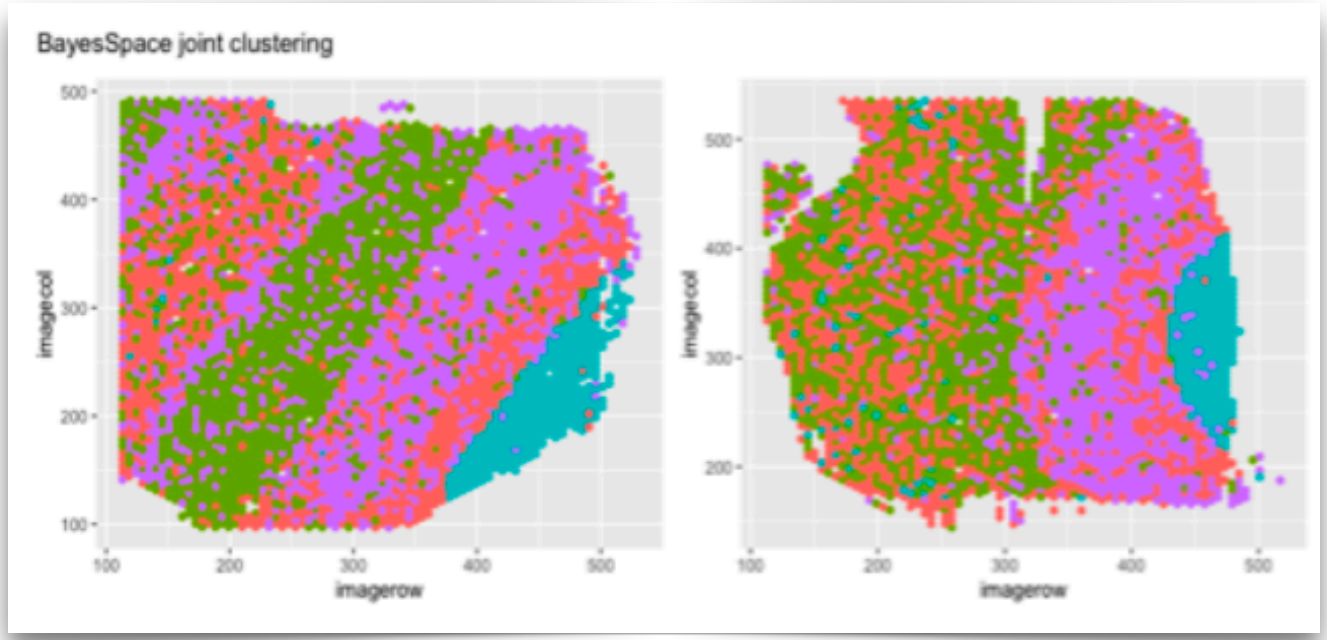
https://www.nature.com/articles/s41467-022-34879-1

Spatial clustering

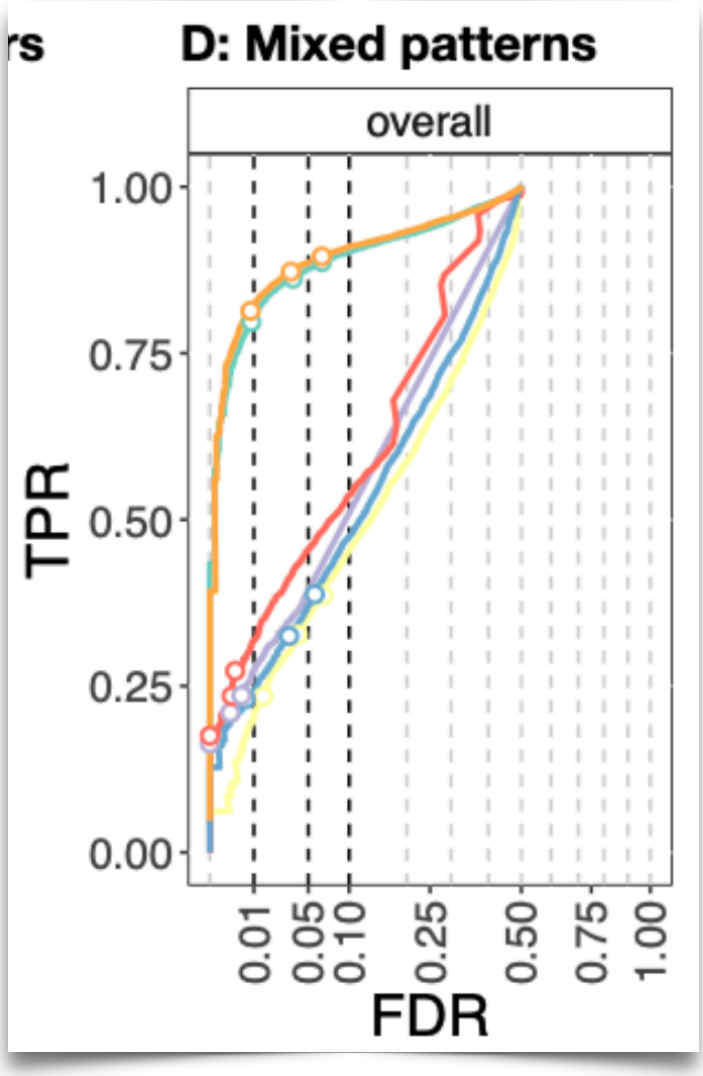


https://www.biorxiv.org/content/10.1101/2022.04.14.488259v1

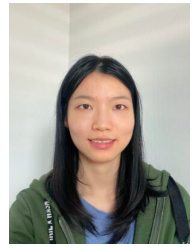
Finding spatially variable features: clustering + DE > spatial statistics



Multi-sample spatial omics simulations; initial goal: find spatially variable genes (SVGs); spatial clustering + classical statistical method works very well for SVG detection.



Simone Tiberi

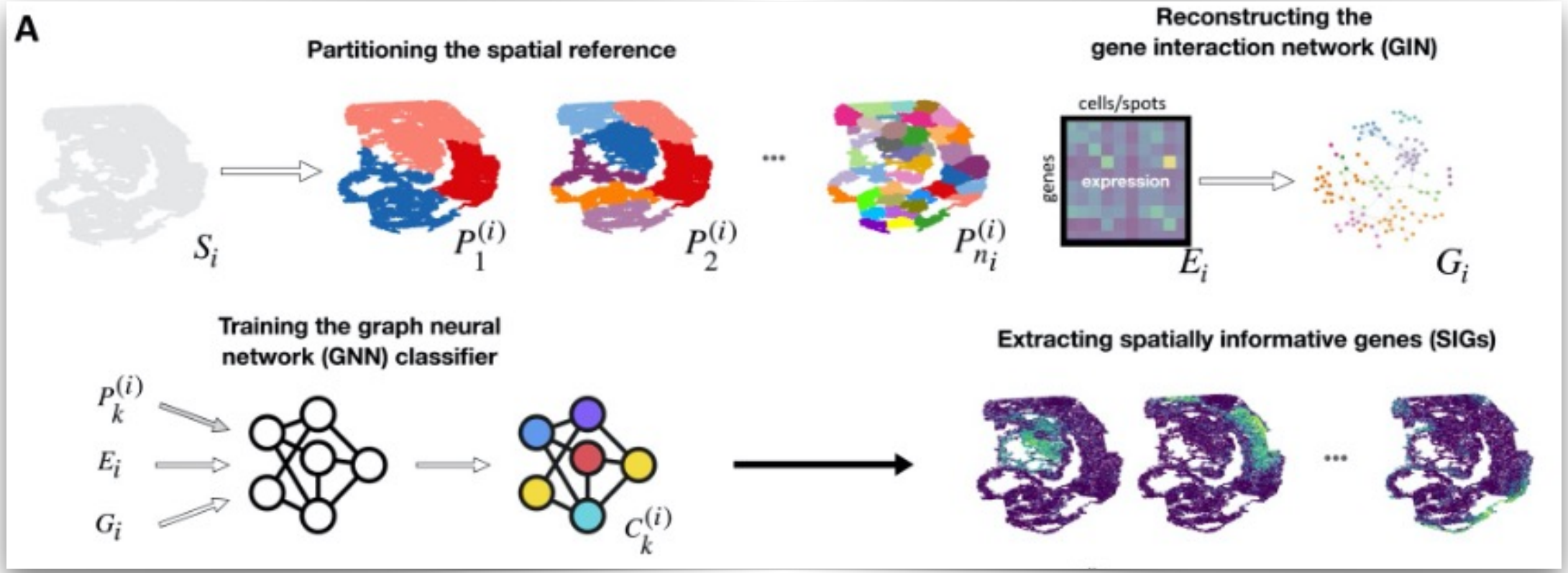


Peiying Cai

Integration of spatial and dissociated single-cell data: class prediction

Supervised spatial inference of dissociated single-cell data with SageNet

Elyas Heidari[1,2,3], Tim Lohoff[4,5,6^], Richard C. V. Tyser[7], John C. Marioni[3,8,9*], Mark D. Robinson[1*], Shila Ghazanfar[3,8*]



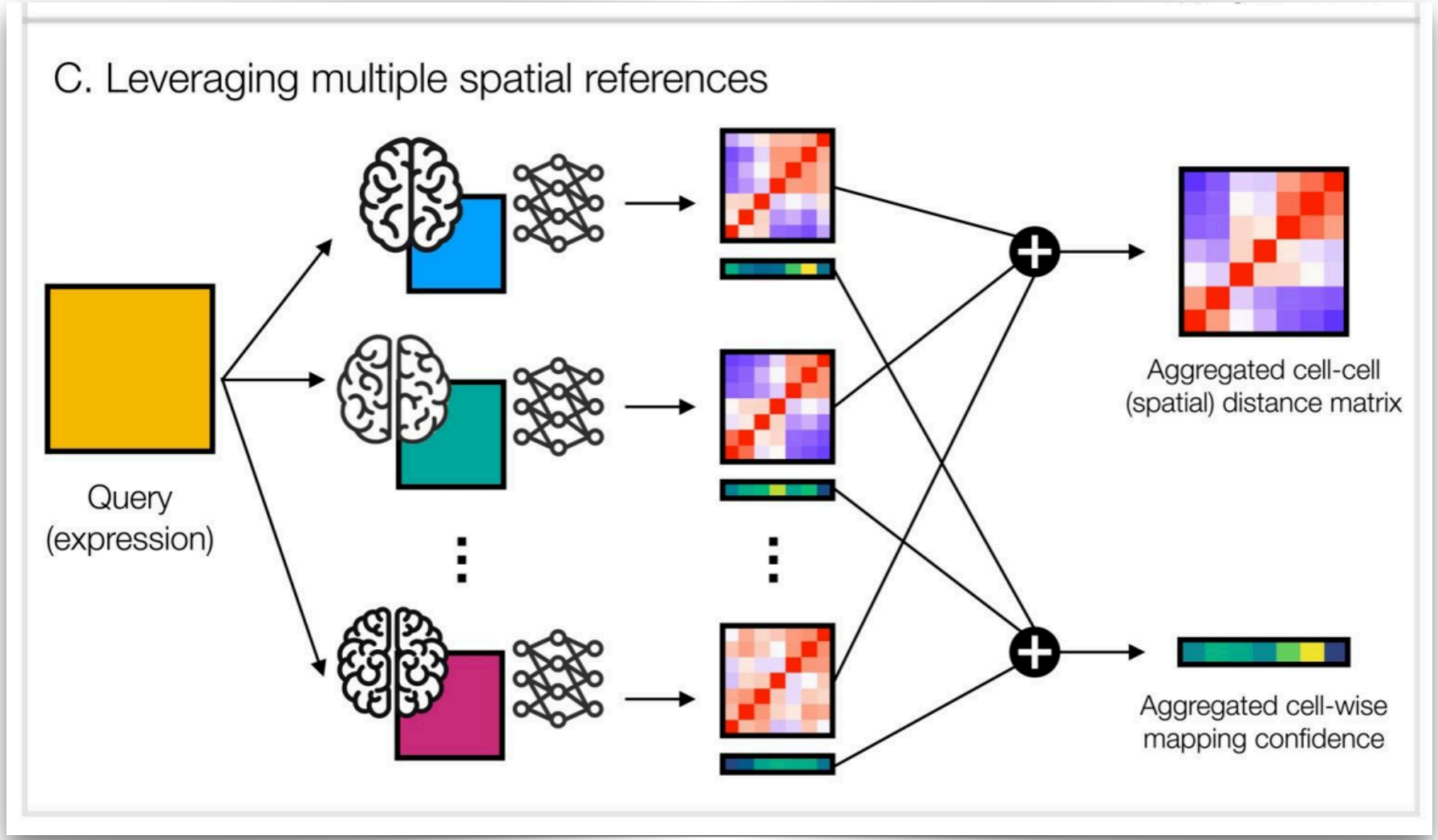
Elyas Heidari

Treat integration of single-cell RNA-seq and spatial transcriptomics as a class prediction (deep learning) problem.

Integration of spatial and dissociated single-cell data: class prediction

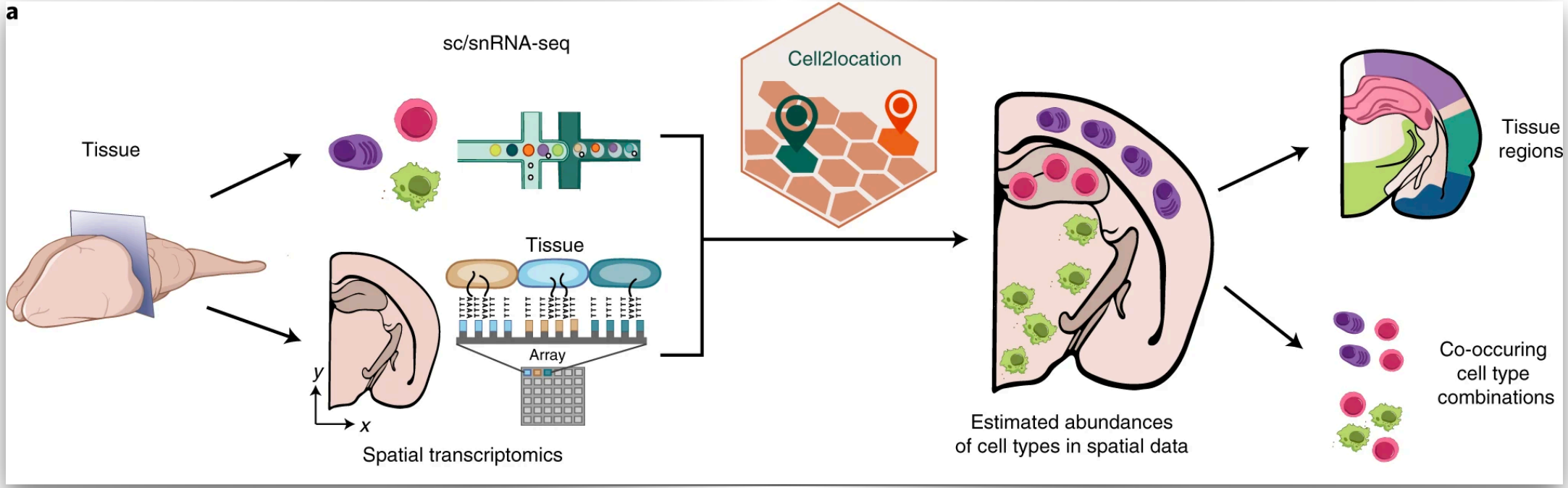
Supervised spatial inference of dissociated single-cell data with SageNet

Elyas Heidari[1,2,3], Tim Lohoff[4,5,6^], Richard C. V. Tyser[7], John C. Marioni[3,8,9*], Mark D. Robinson[1*], Shila Ghazanfar[3,8*]



Deconvoluting low-resolution spatial omics 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 $g_{f,g}$:

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

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

Cell-cell communication

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