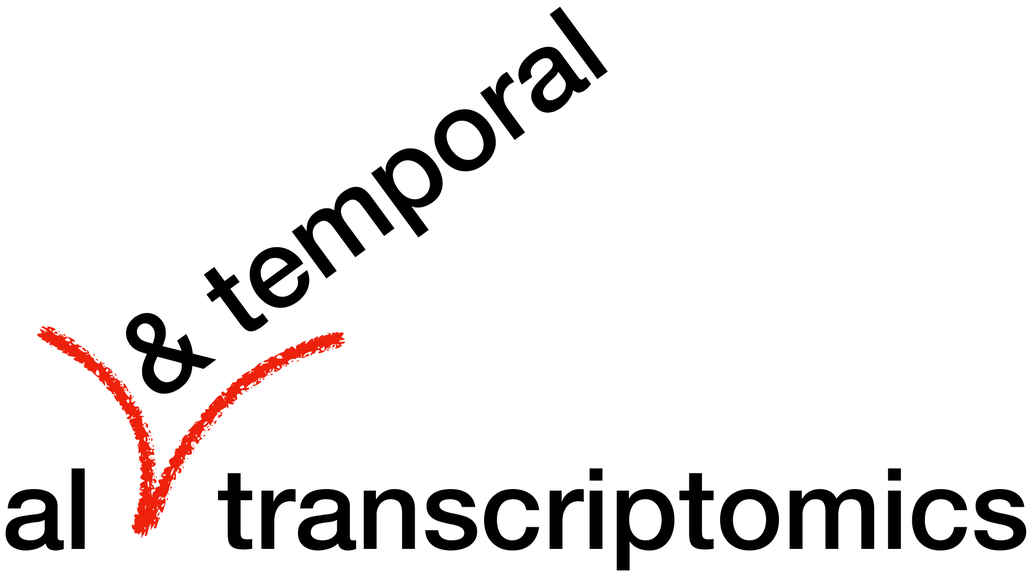
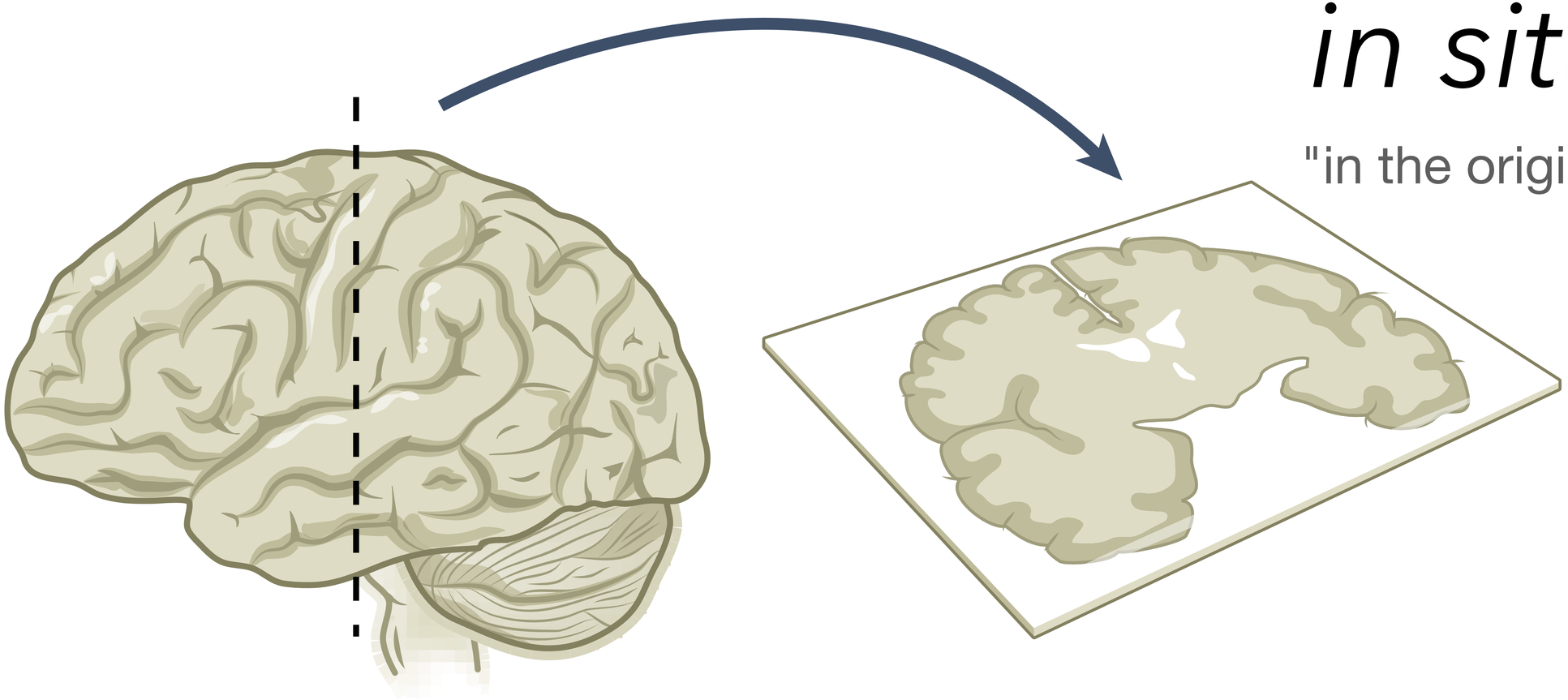
Statistical Methods for High-dimensional Biology



Yongjin Park, UBC Path&Lab, STAT, BC Cancer

Spatial transcriptomics



in situ

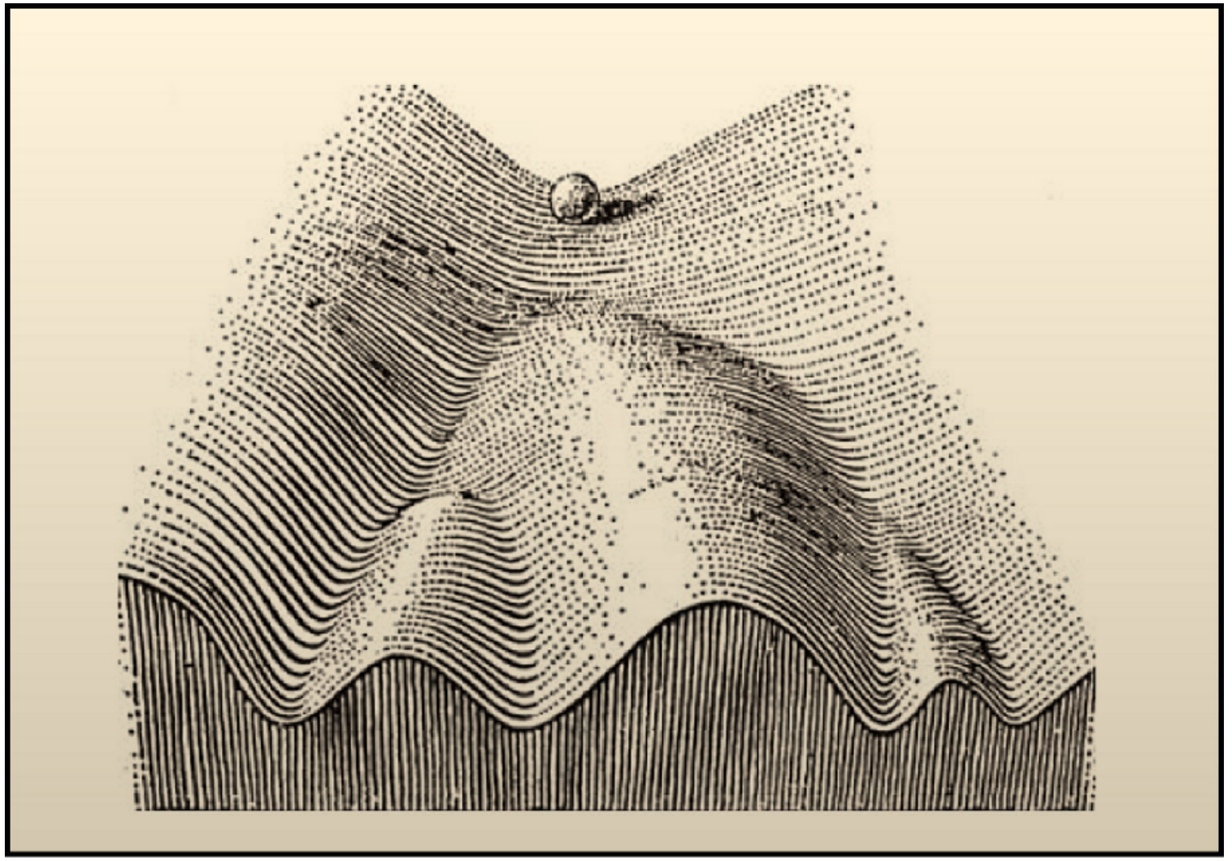
"in the original position"

Tian .. Macosko *Nature Biotech.* (2023)

Why spatial transcriptomics?

- Single-cell to multicellular organism
- Space ~ time
- Contexts: epigenetic landscape ~ cell type differentiation

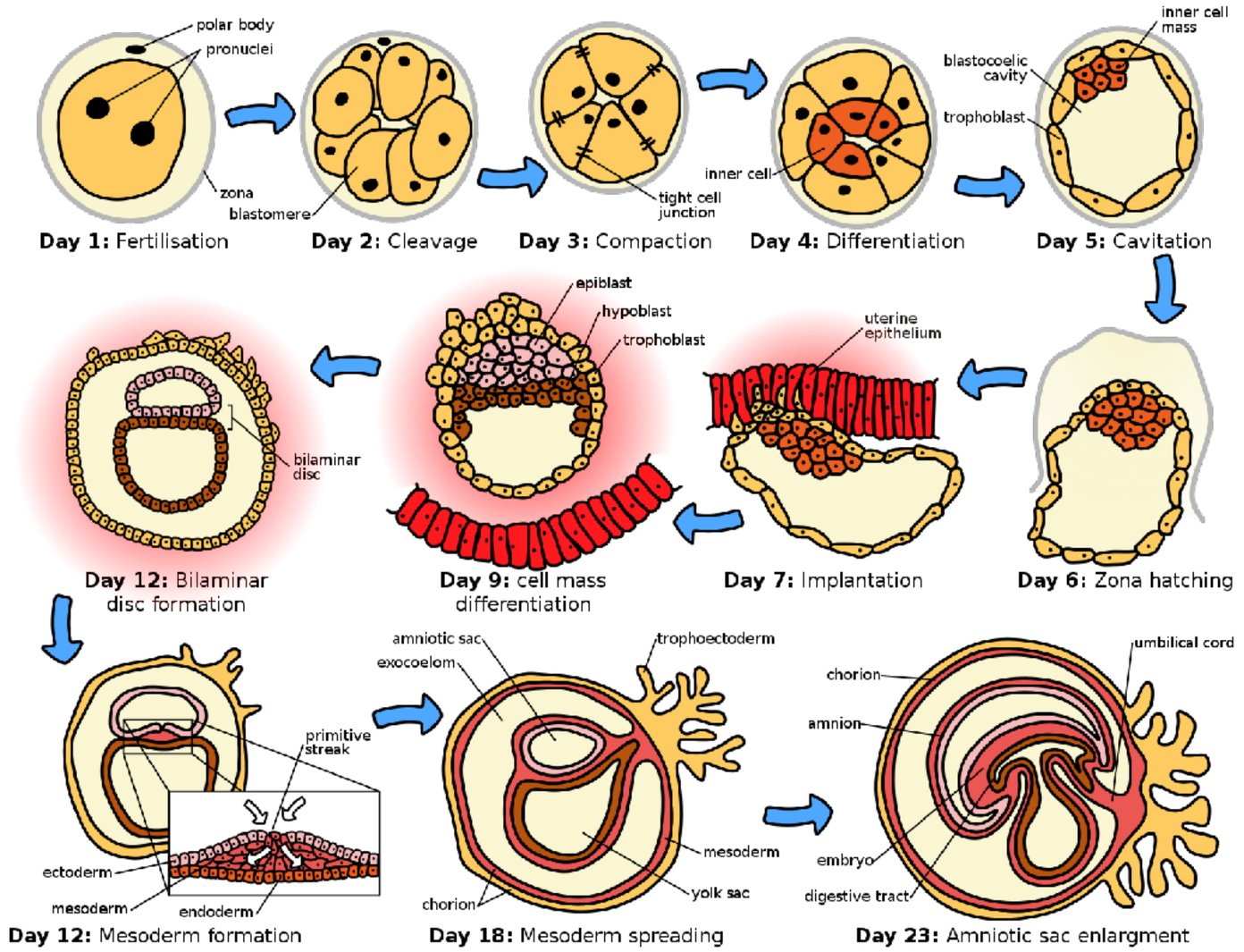
Waddington's Epigenetic Landscape



time

Why spatial transcriptomics?

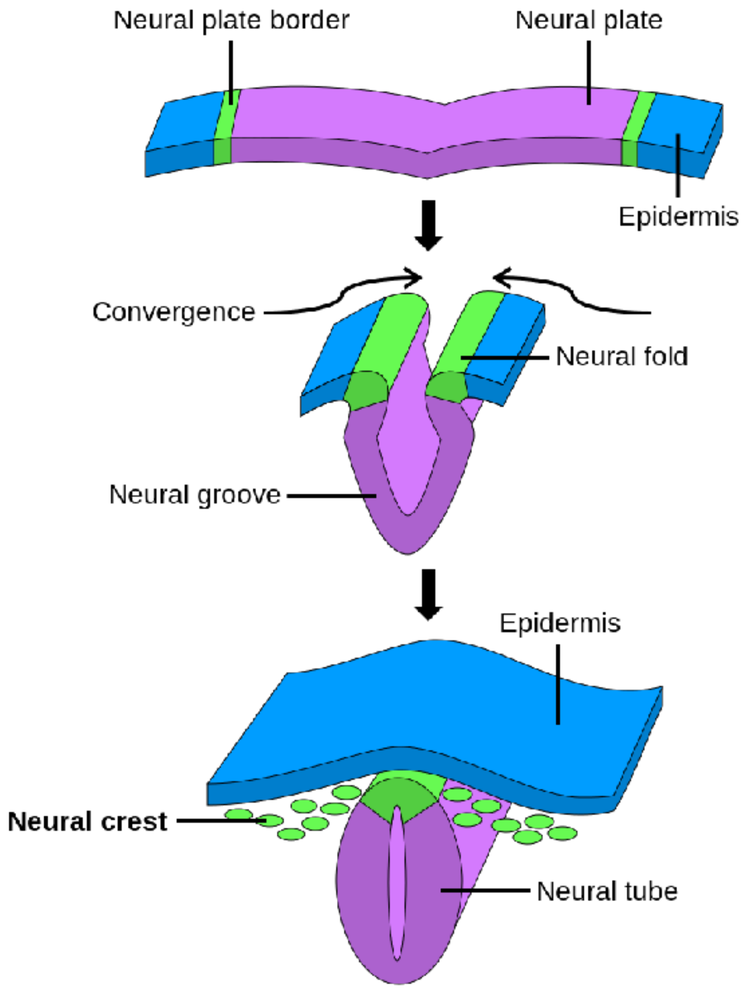
- Single-cell to multicellular organism
- Space ~ time
- Contexts: E.g., tumour immune interactions
- Constraints: little probability of interaction between distant locations



https://en.wikipedia.org/wiki/Human_embryonic_development

Why spatial transcriptomics?

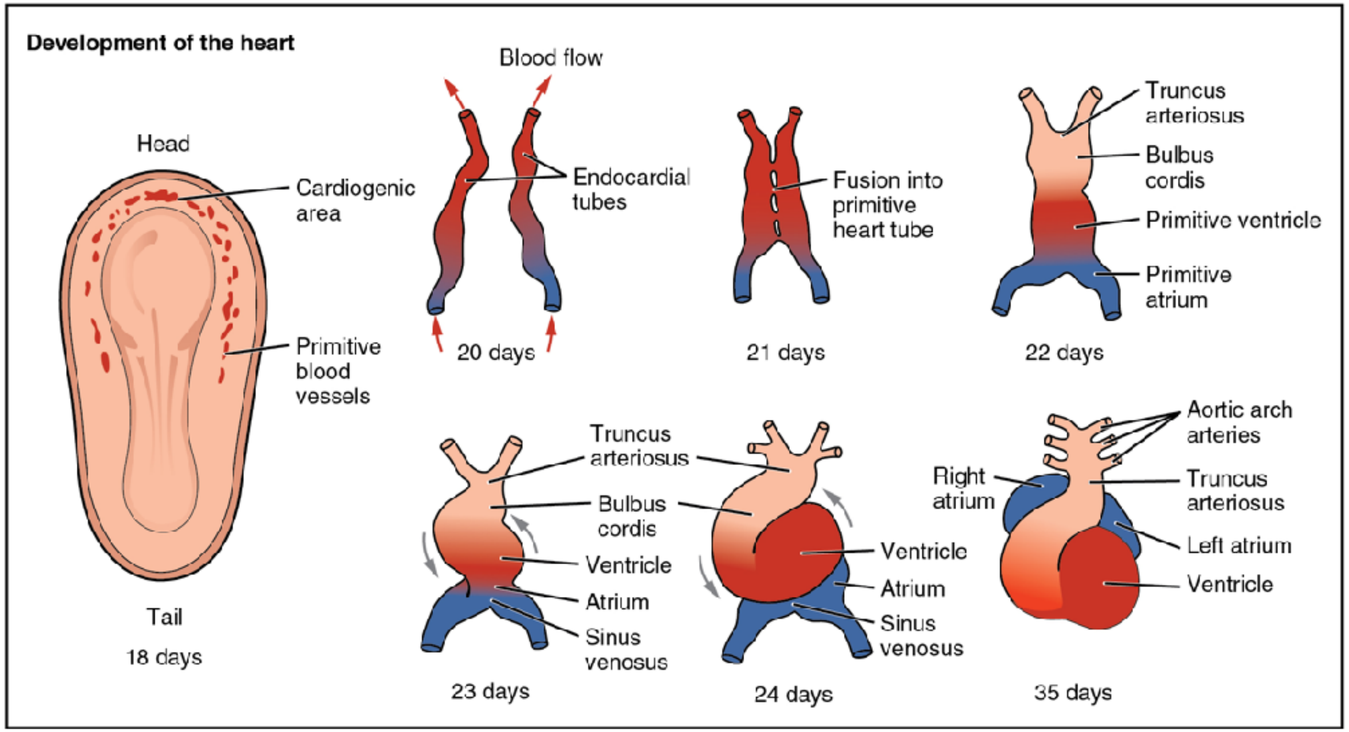
- Single-cell to multicellular organism
- Space ~ time
- Contexts: E.g., tumour immune interactions
- Constraints: little probability of interaction between distant locations



https://en.wikipedia.org/wiki/Human_embryonic_development

Why spatial transcriptomics?

- Single-cell to multicellular organism
- Space ~ time
- Contexts: E.g., tumour immune interactions
- Constraints: little probability of interaction between distant locations



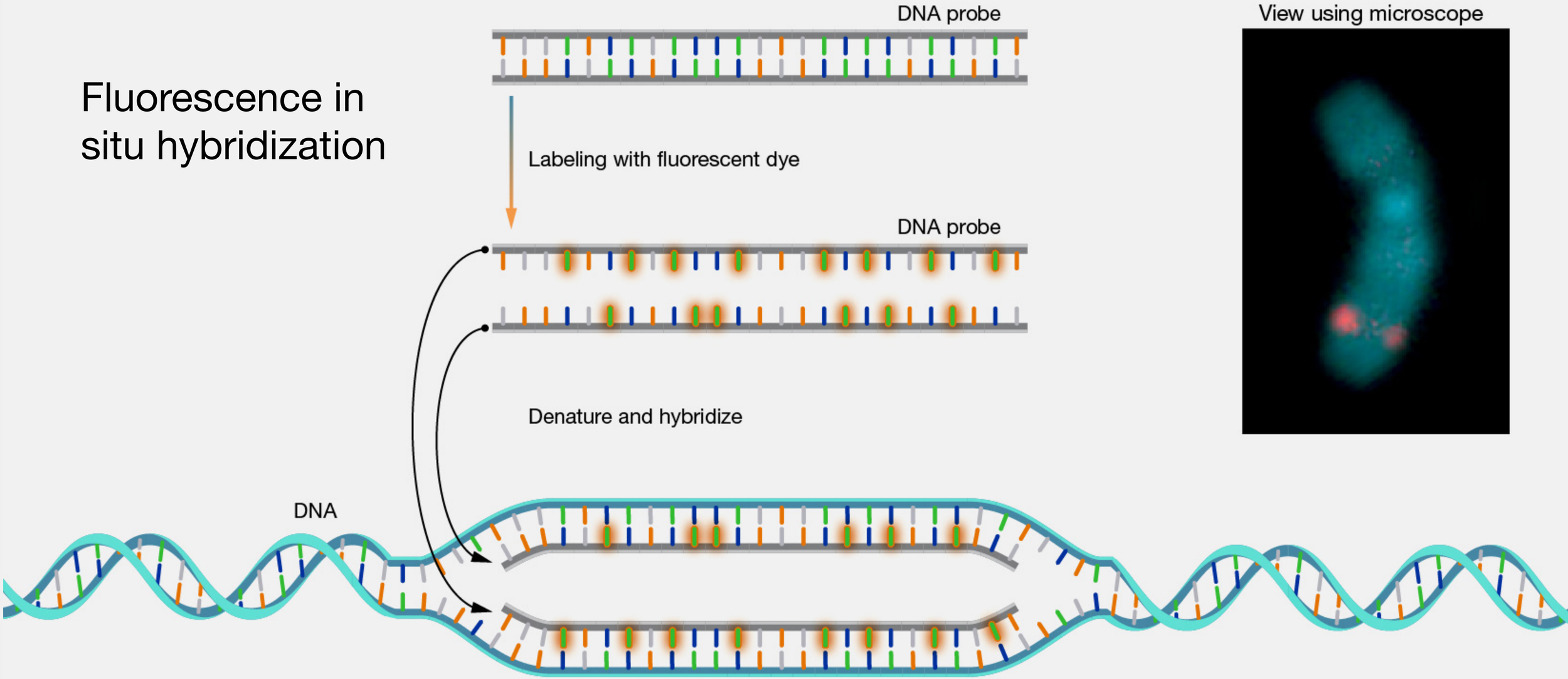
https://en.wikipedia.org/wiki/Human_embryonic_development

Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

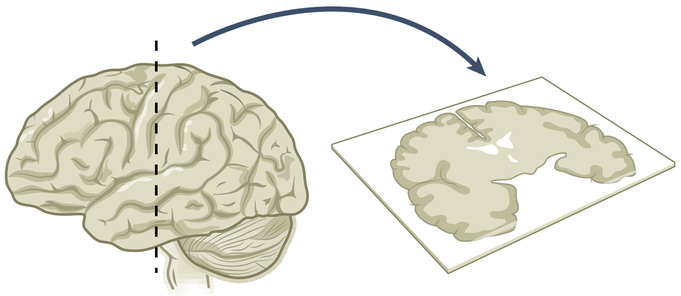
FISH: Spatial transcriptomics in old days

Fluorescence in situ hybridization

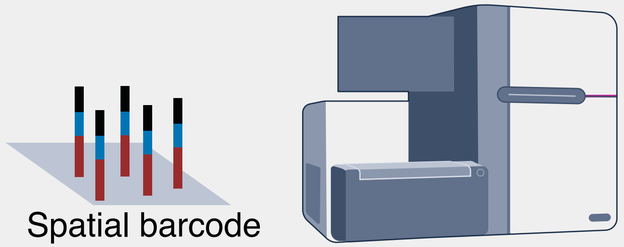


View using microscope

Spatial transcriptomics technologies

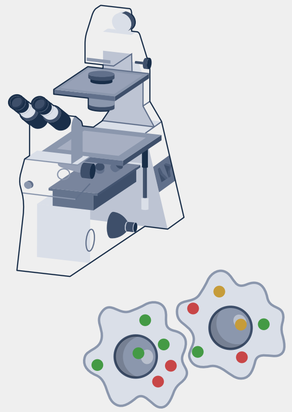


Sequencing-based



ATCGATGGATCGTTTTTT
ATCGATGGTCGGTTTTTT
CCAGTTGGAATGTTTTTT

Imaging-based



Gene expression

	Gene 1	Gene 2	Gene *N*
Spot 1	10	5	6
Spot 2	5	21	20
...	...	...	...
Spot *N*	8	1	2

Spatial information

	x	*y*
Spot 1	141.2	511.4
Spot 2	514.9	219.9
...	...	...
Spot *N*	8	127.4

Tian .. Macosko *Nature Biotech.* (2023)

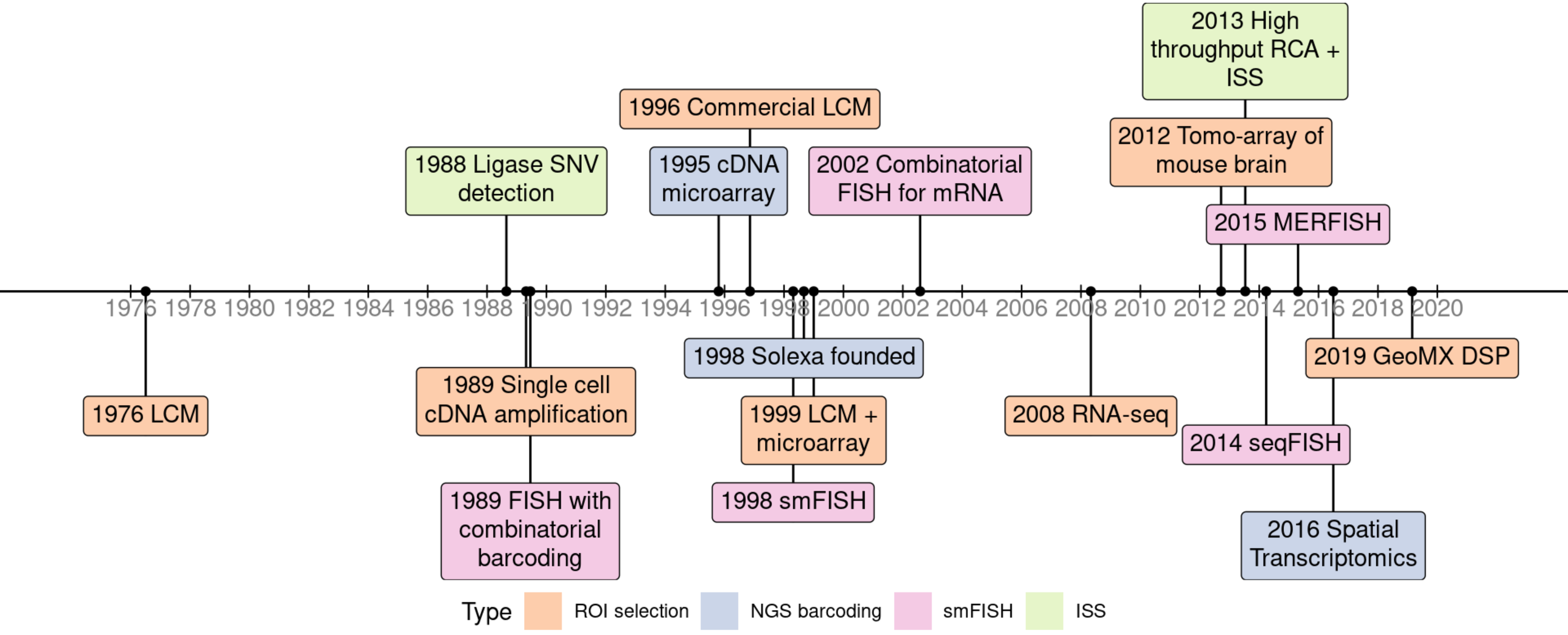
REVIEW ARTICLE
https://doi.org/10.1038/s41592-022-01409-2

nature methods

Museum of spatial transcriptomics

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

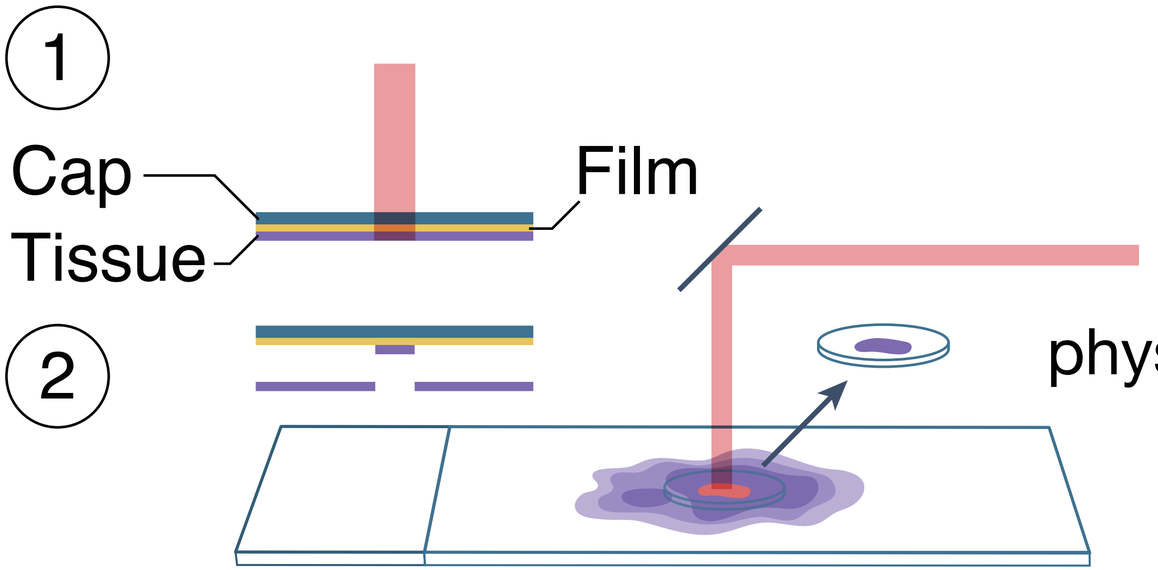
The function of many biological systems, such as embryos, liver lobules, intestinal villi, and tumors, depends on the spatial organization of their cells. In the past decade, high-throughput technologies have been developed to quantify gene expression in space, and computational methods have been developed that leverage spatial gene expression data to identify genes with spatial patterns and to delineate neighborhoods within tissues. To comprehensively document spatial gene expression technologies and data-analysis methods, we present a curated review of literature on spatial transcriptomics dating back to 1987, along with a thorough analysis of trends in the field, such as usage of experimental techniques, species, tissues studied, and computational approaches used. Our Review places current methods in a historical context, and we derive insights about the field that can guide current research strategies. A companion supplement offers a more detailed look at the technologies and methods analyzed: https://pachterlab.github.io/LP_2021/.



https://pachterlab.github.io/LP_2021/current-analysis.html

Imaging-based: How do we know locations?

Infra Red Laser Capture Microdissection



physically dissect out ROI (region of interest)

Moses & Pachter *Nature Methods* (2022)

LCM = Laser Capture Microdissection

Imaging-based: How do we know locations?

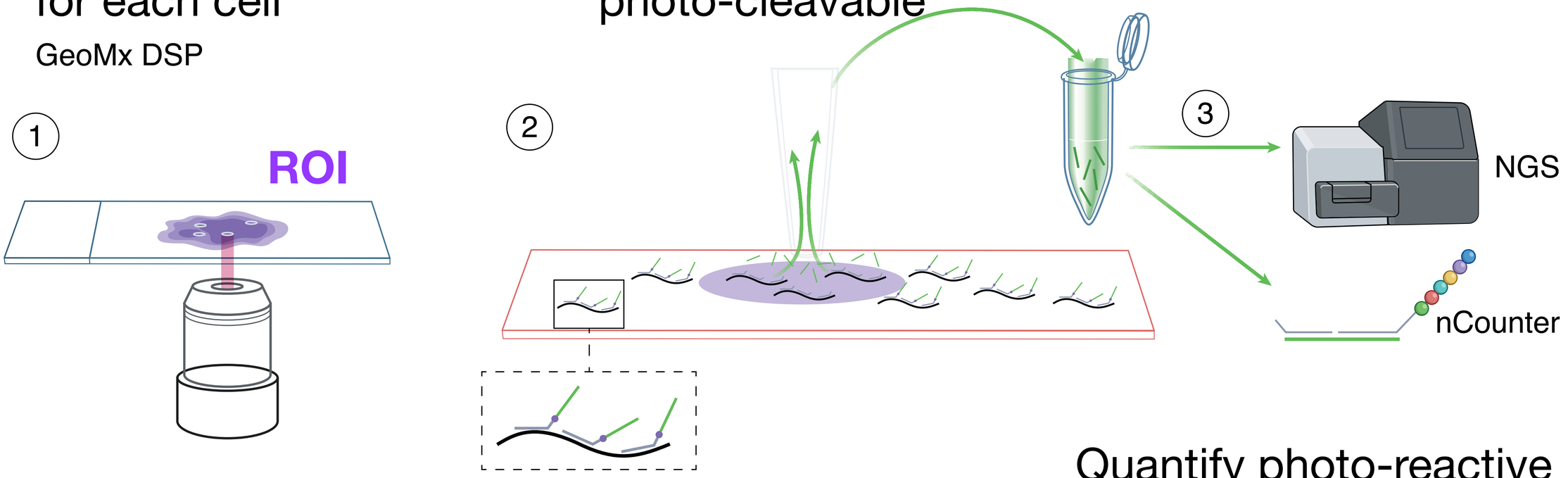
Digital Spatial Profiler

for each cell

GeoMx DSP

Release spatial barcode--

photo-cleavable

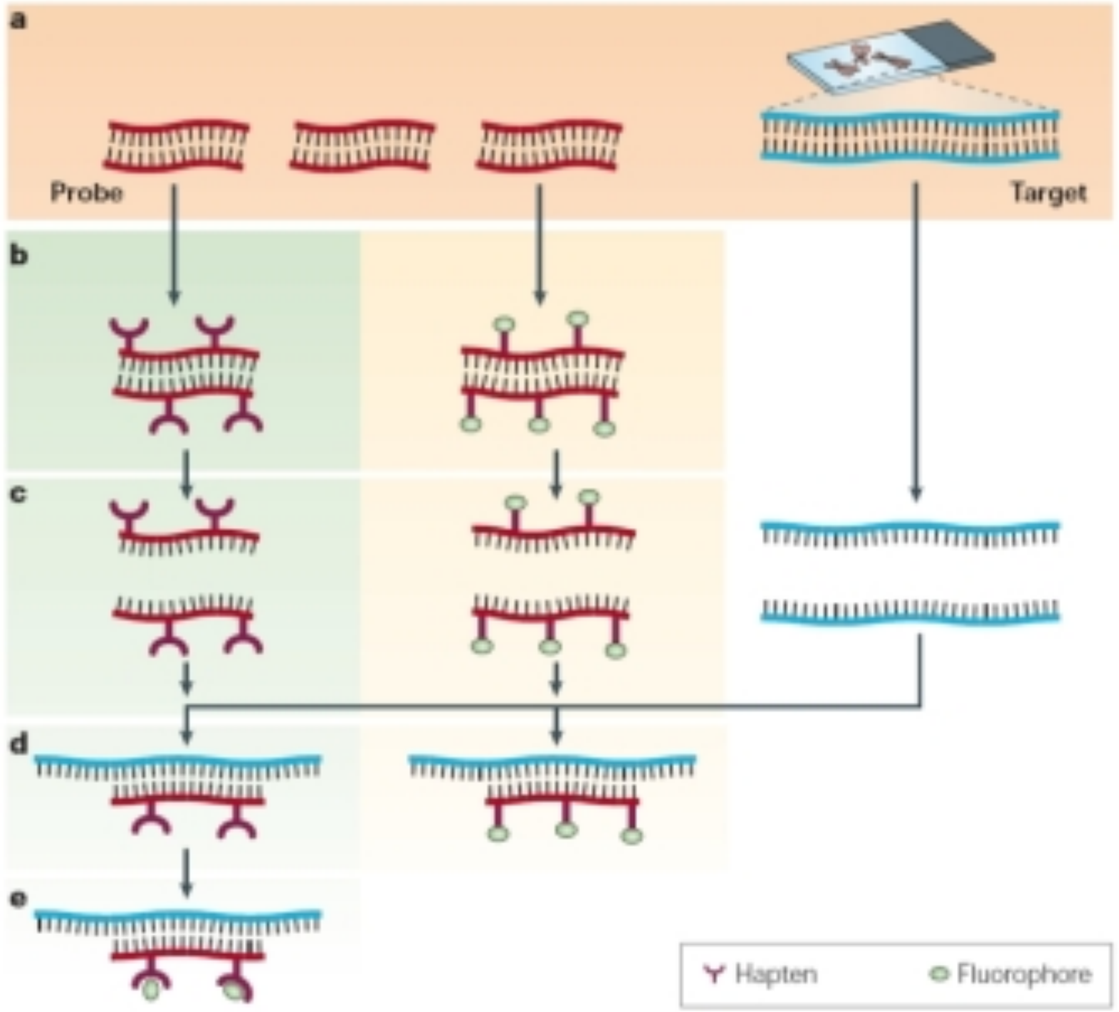


Quantify photo-reactive

barcodes & gene expression

Moses & Pachter *Nature Methods* (2022)

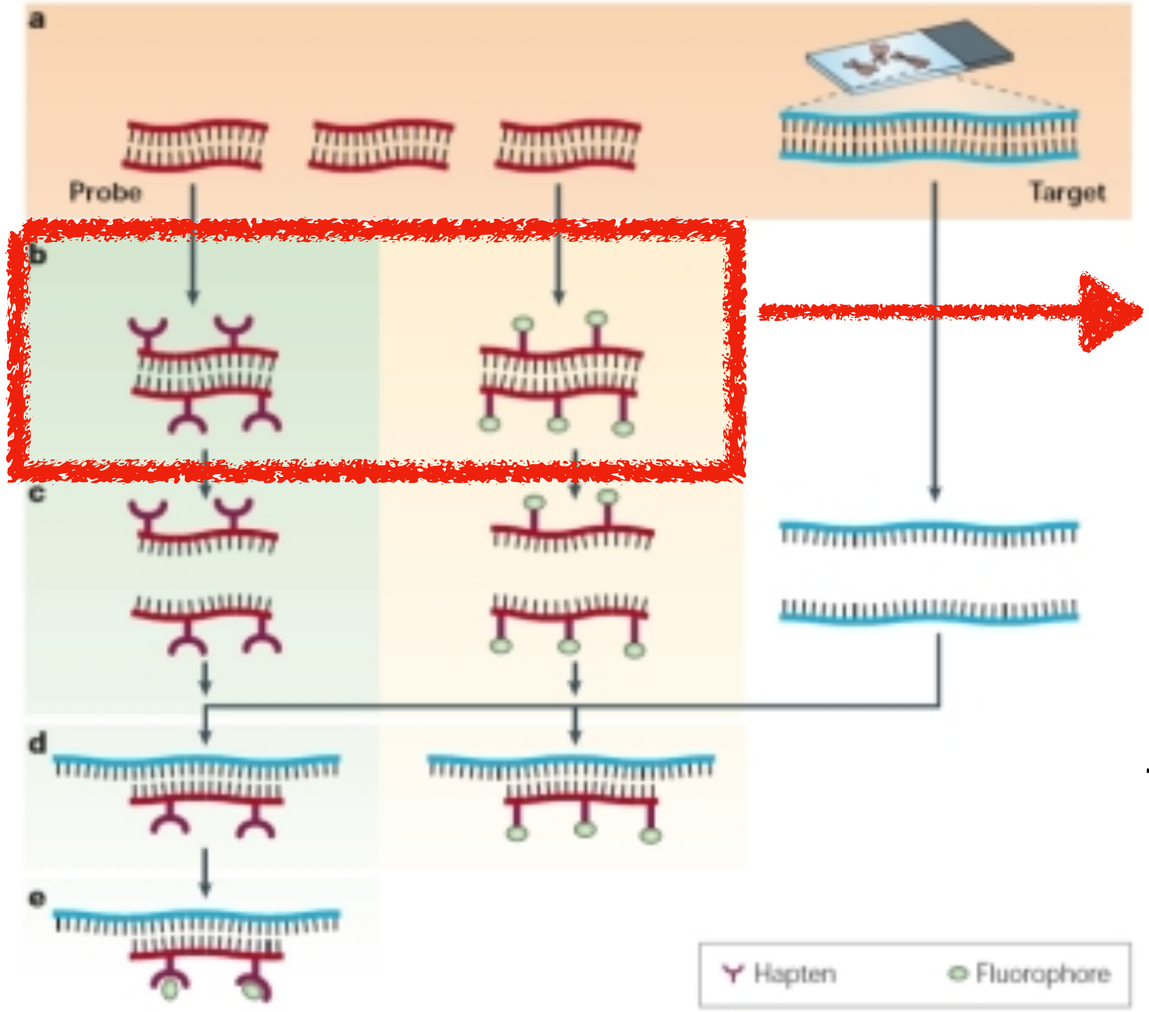
Single-molecule FISH



fluorescence in situ hybridization (FISH)

https://www.nature.com/scitable/topicpage/fluorescence-in-situ-hybridization-fish-327

Single-molecule FISH



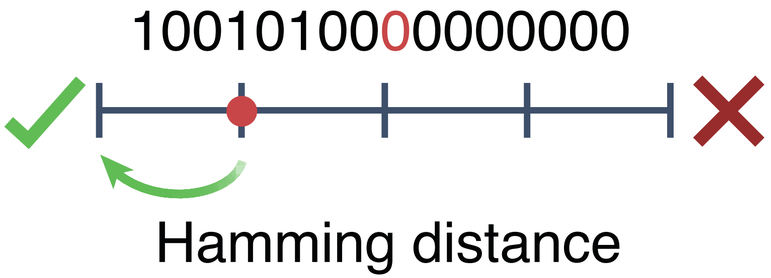
Strategy #1: seqFISH

Strategy #2: MERFISH

(multiplexed error-robust)

1001010010000000

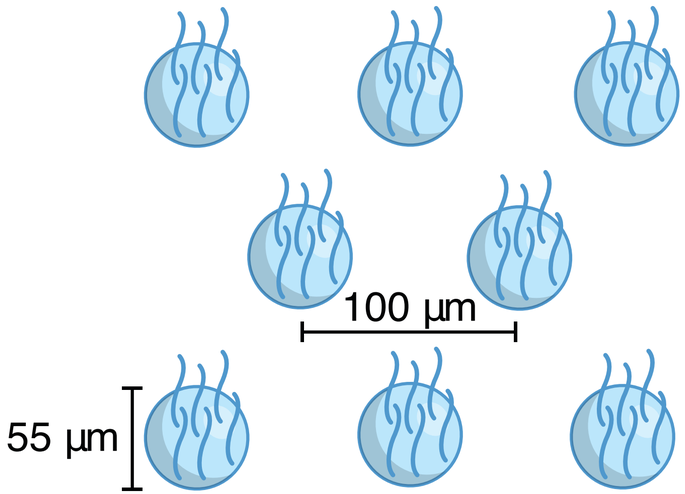
Original



Sequencing + spatial barcoding

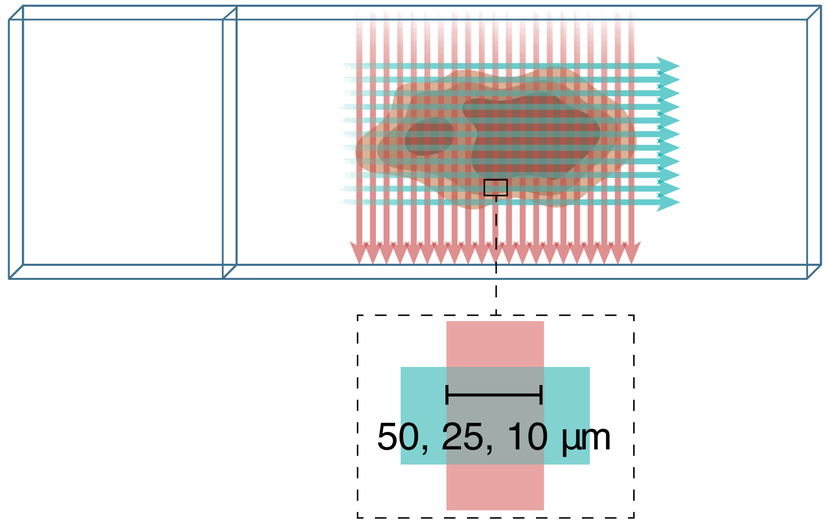
Array-based (3') UMI (unique molecule identifier) per each spot

Visium



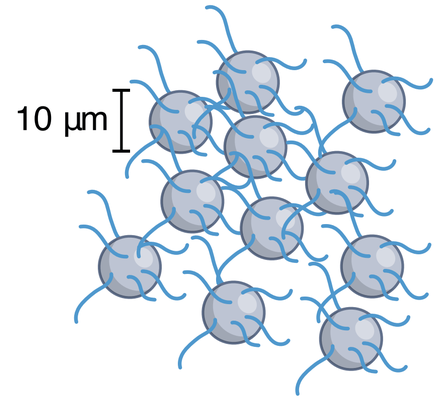
Array generated by microfluidics channels

DBiT-seq

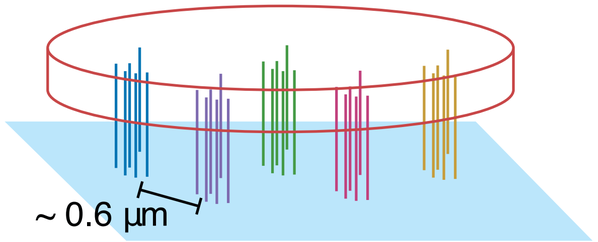


Barcoded beads spread on a slide

Slide-seq

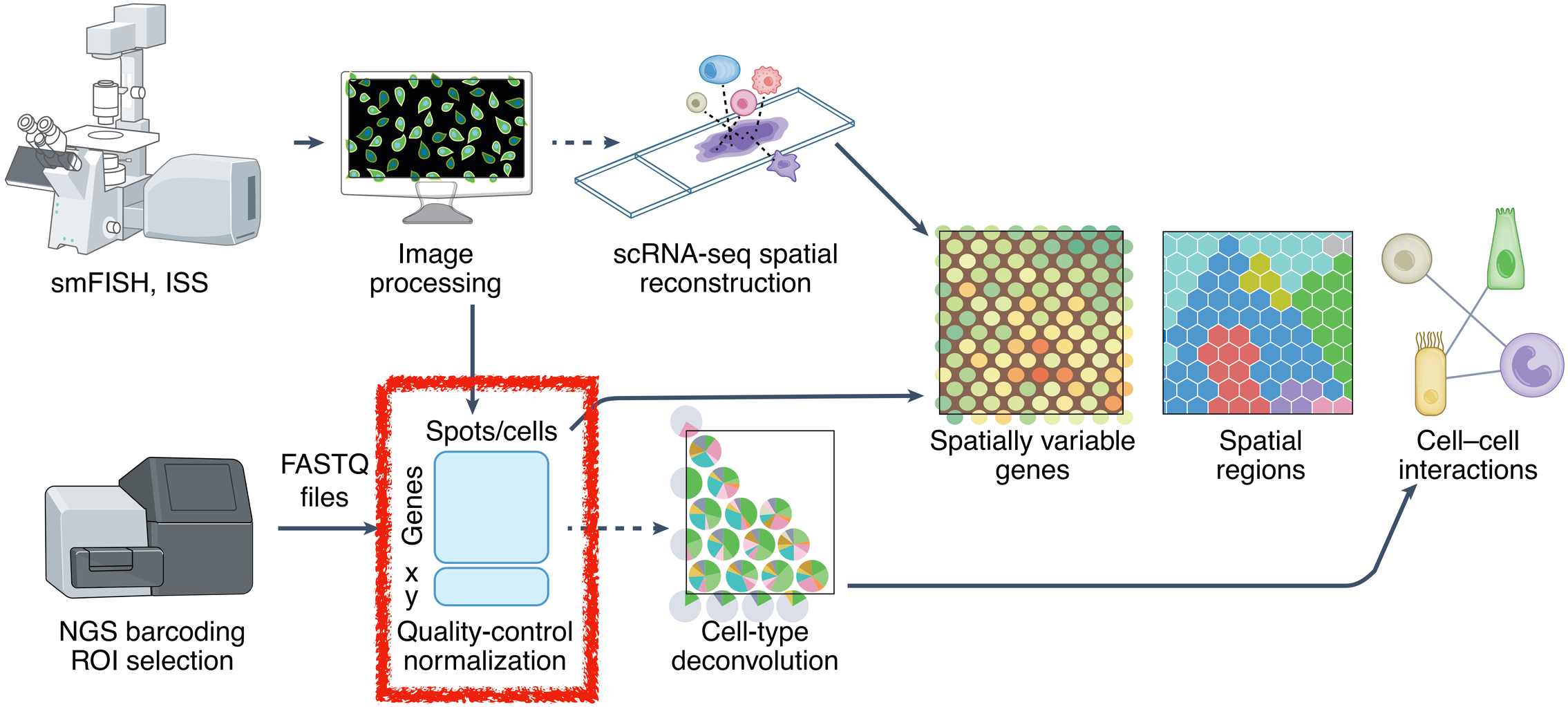


Seq-Scope



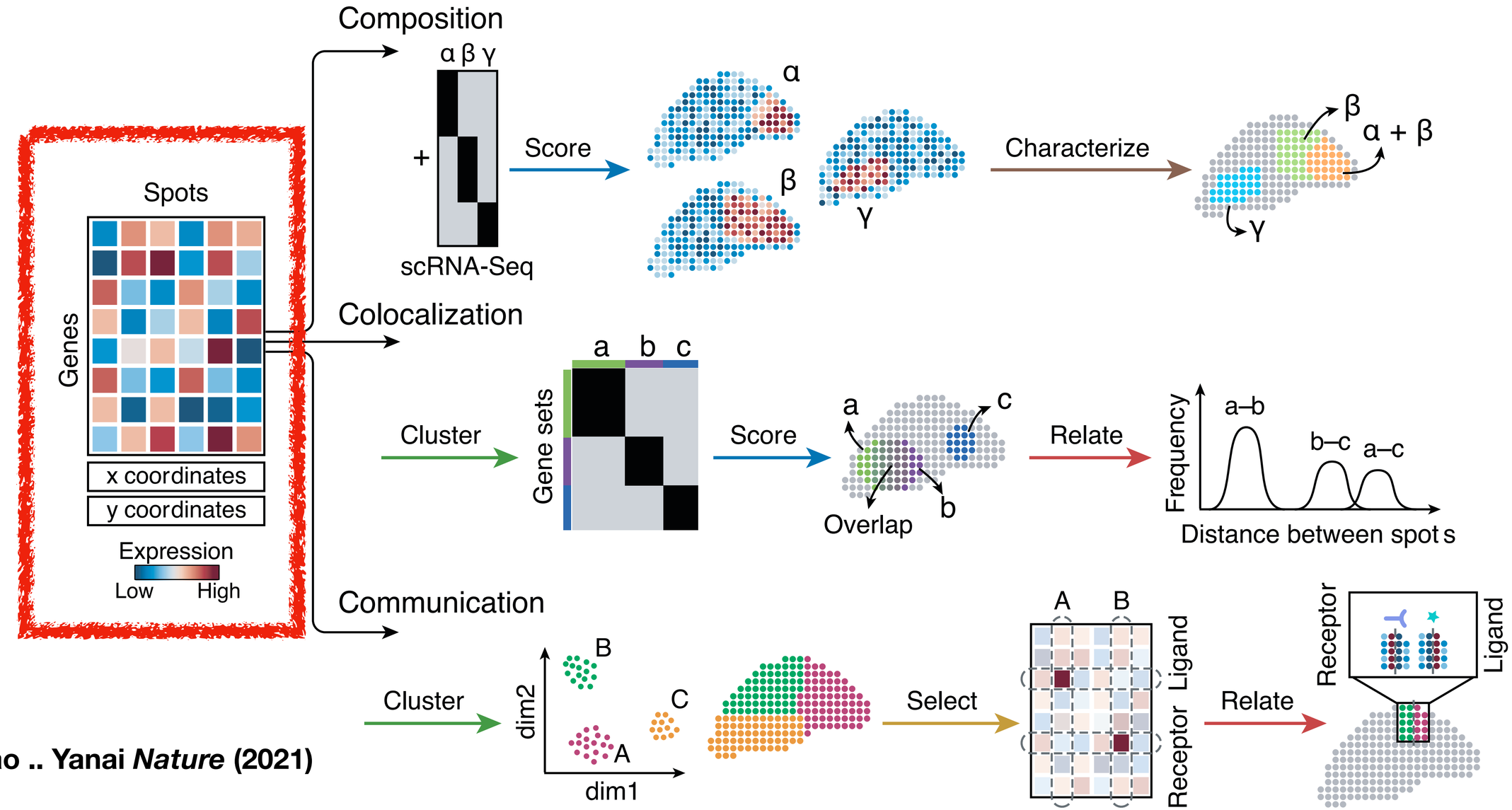
Moses & Pachter *Nature Methods* (2022)

Spatial transcriptomics data analysis



Moses & Pachter *Nature Methods* (2022)

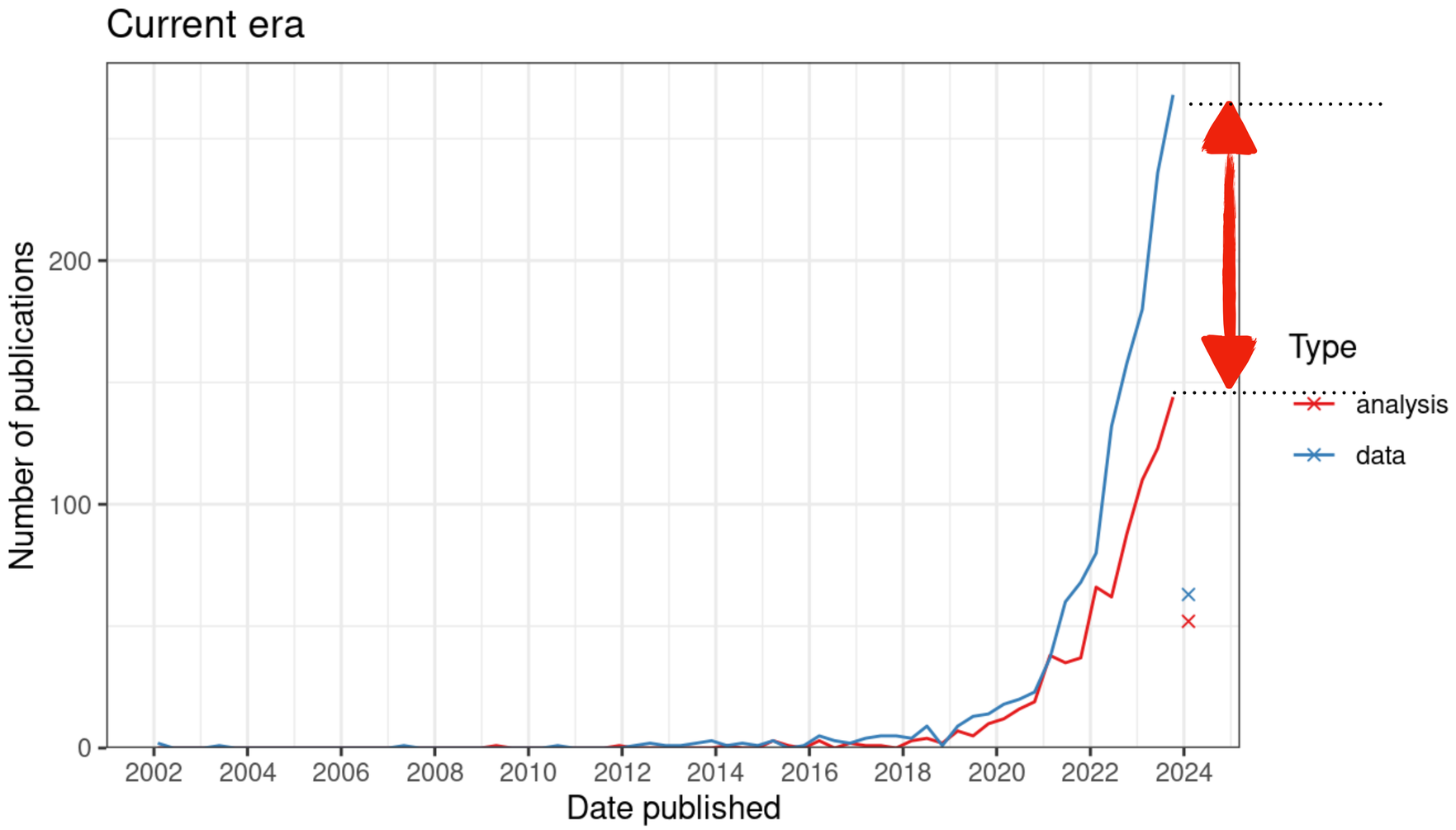
ST analysis pipeline



Rao .. Yanai *Nature* (2021)

https://pachterlab.github.io/LP_2021/current-analysis.html

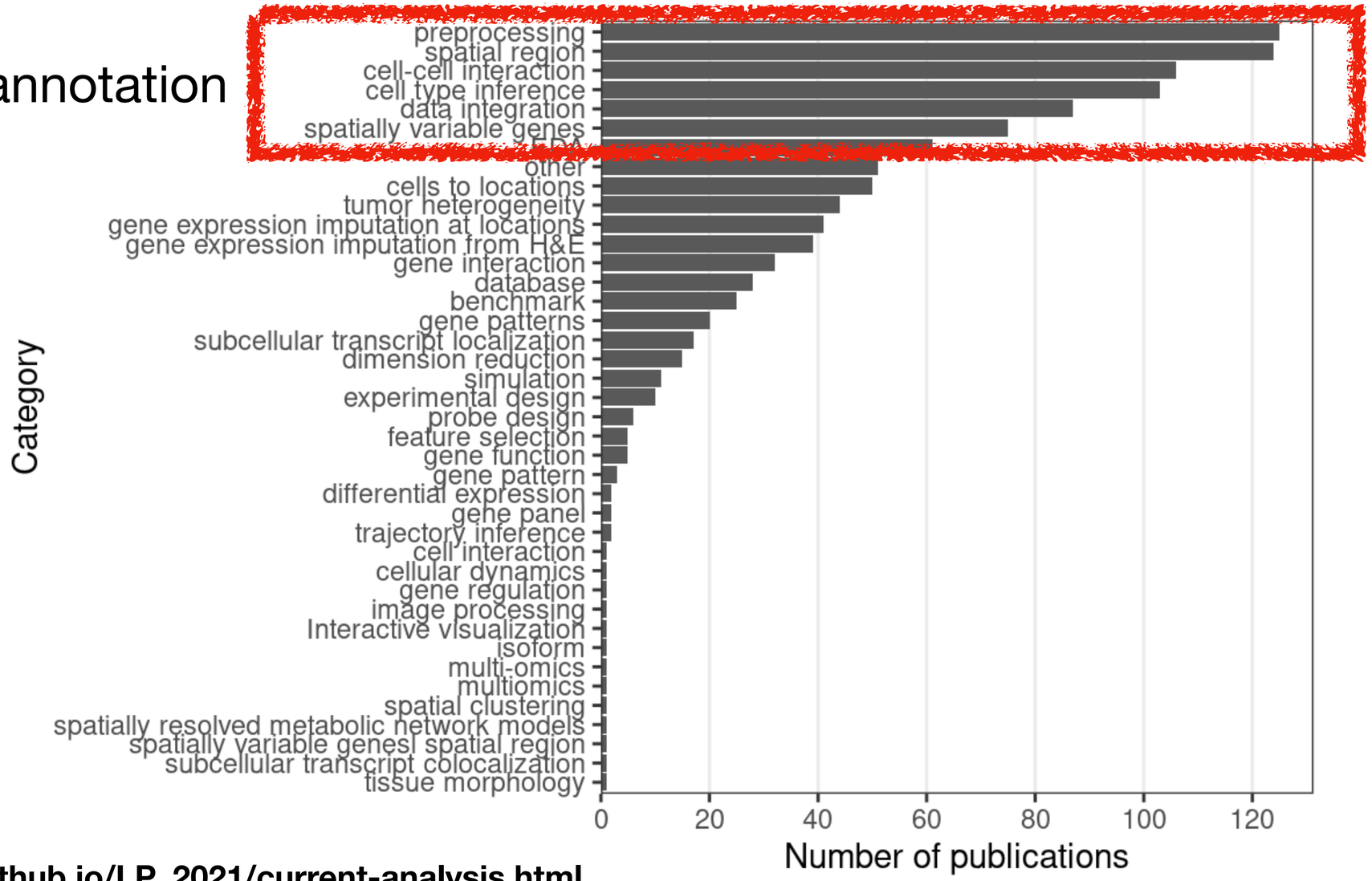
Is ST really a current trend?



https://pachterlab.github.io/LP_2021/current-analysis.html

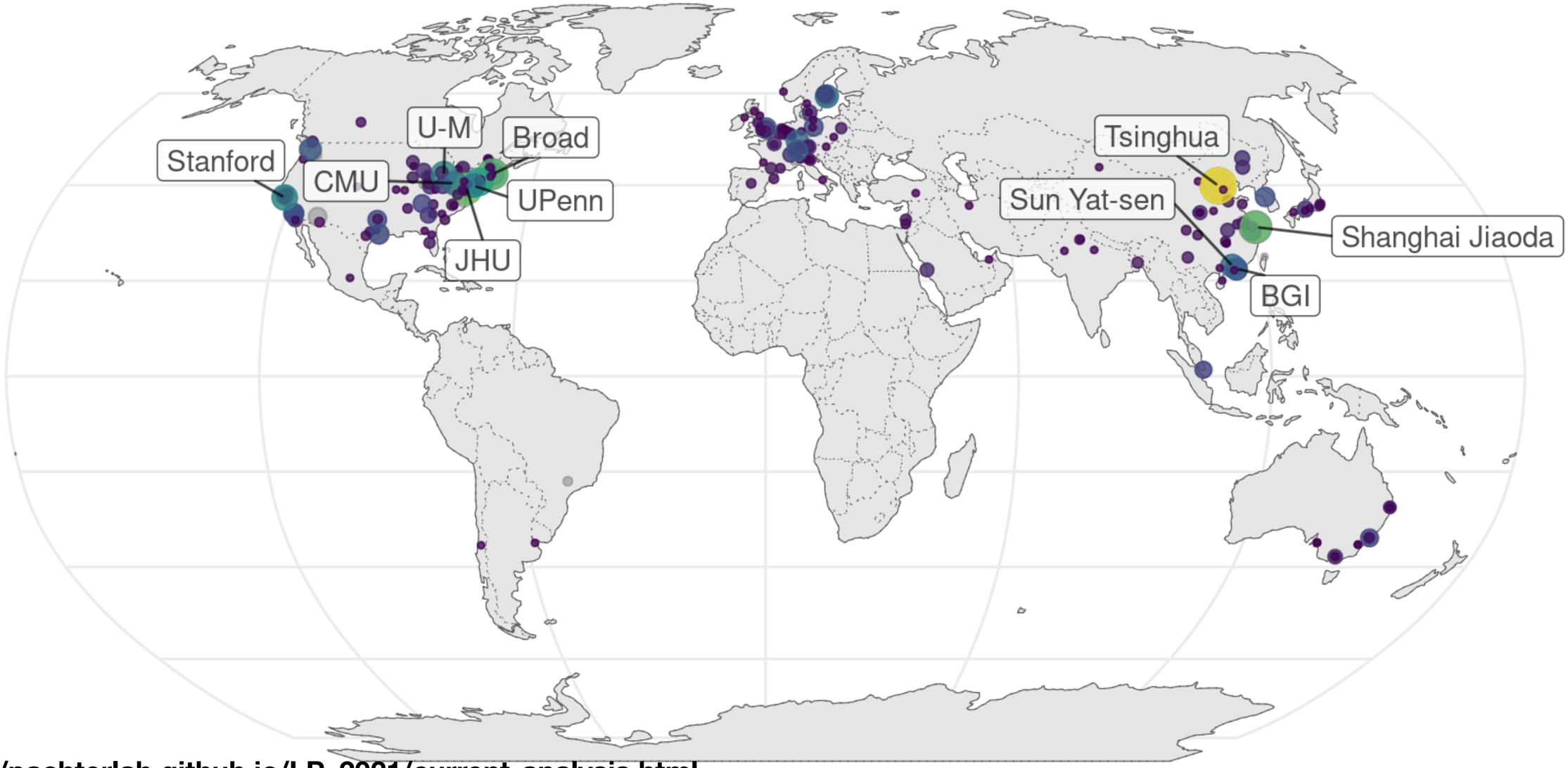
The field is still growing

Cell type annotation



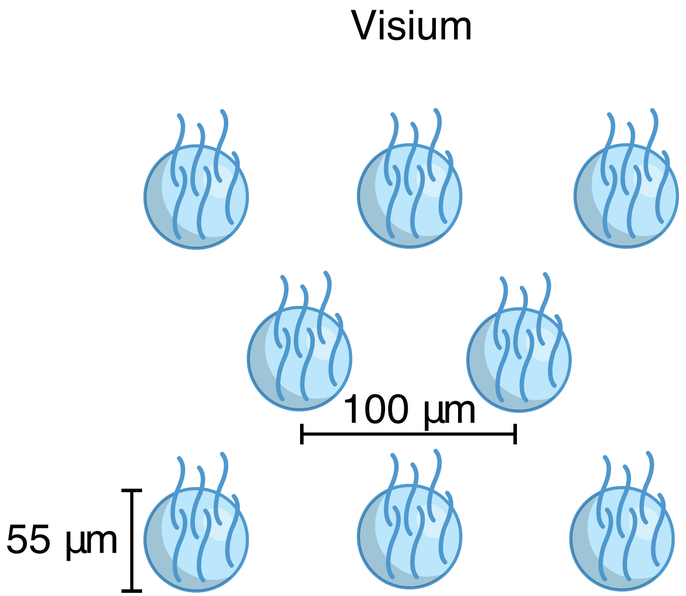
https://pachterlab.github.io/LP_2021/current-analysis.html

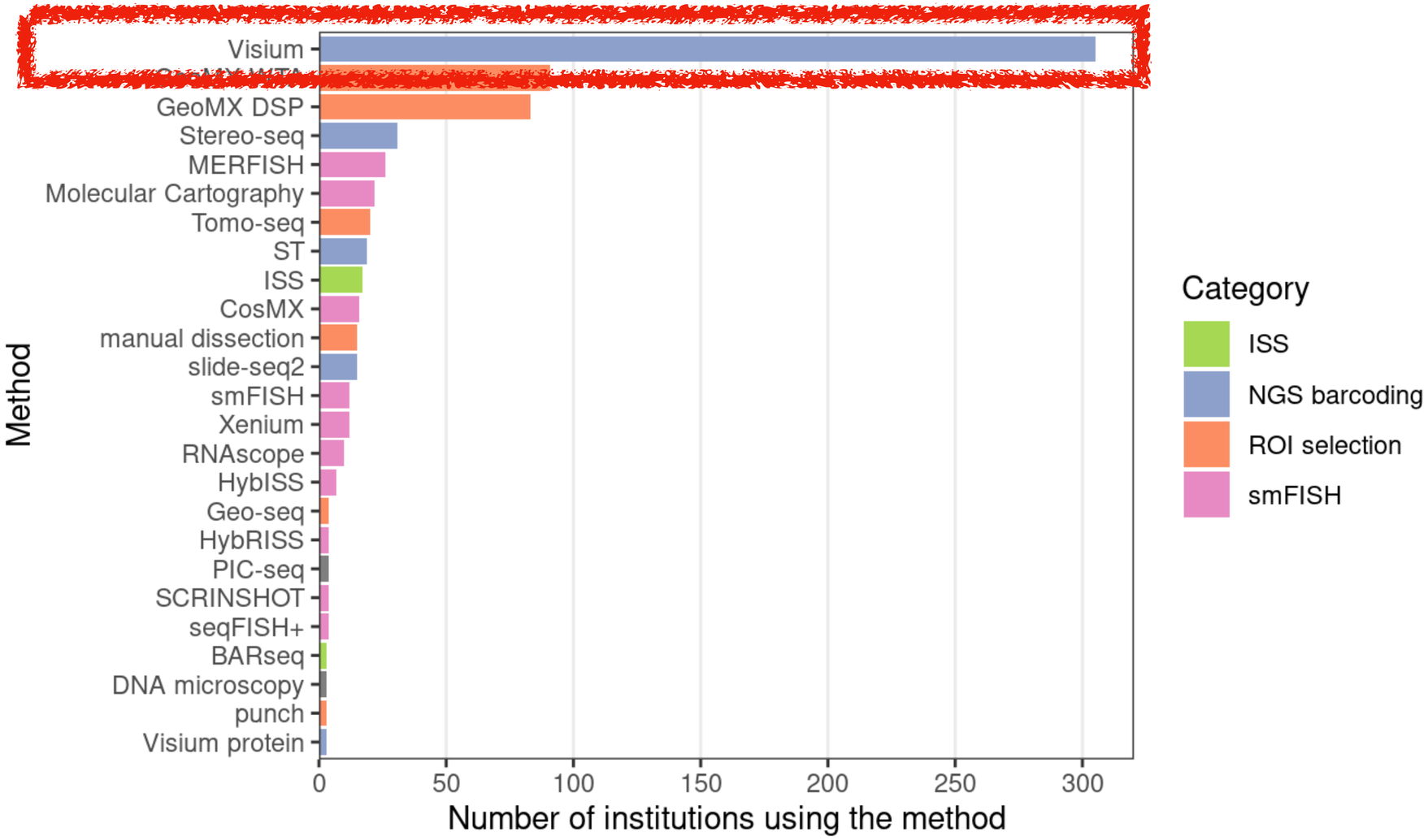
Who are the big players?



https://pachterlab.github.io/LP_2021/current-analysis.html

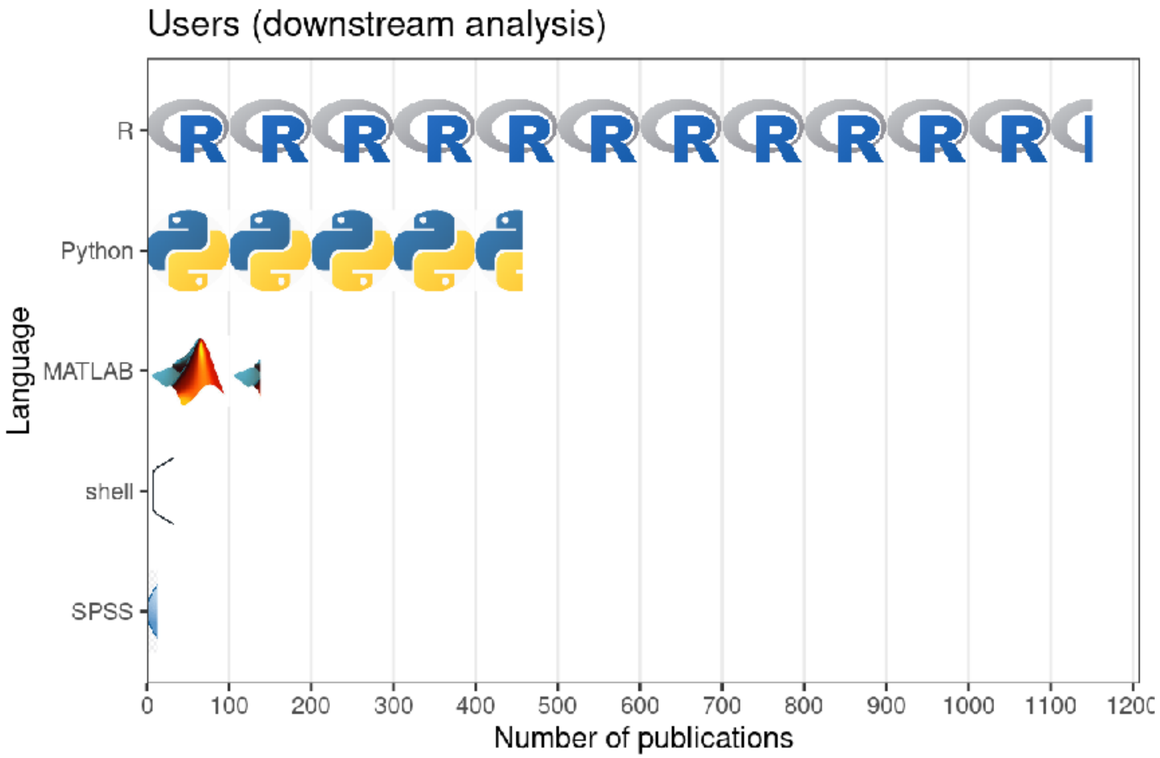
What are the most popular platforms?

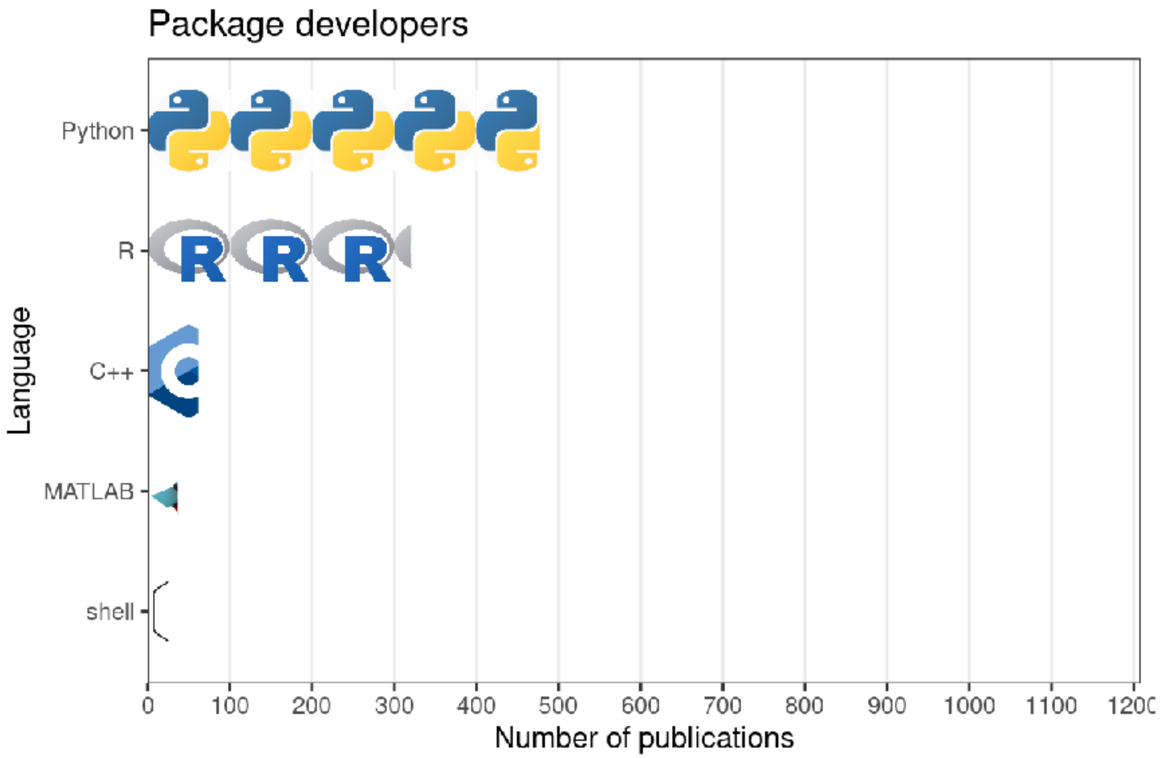




https://pachterlab.github.io/LP_2021/current-analysis.html

In what programming language?





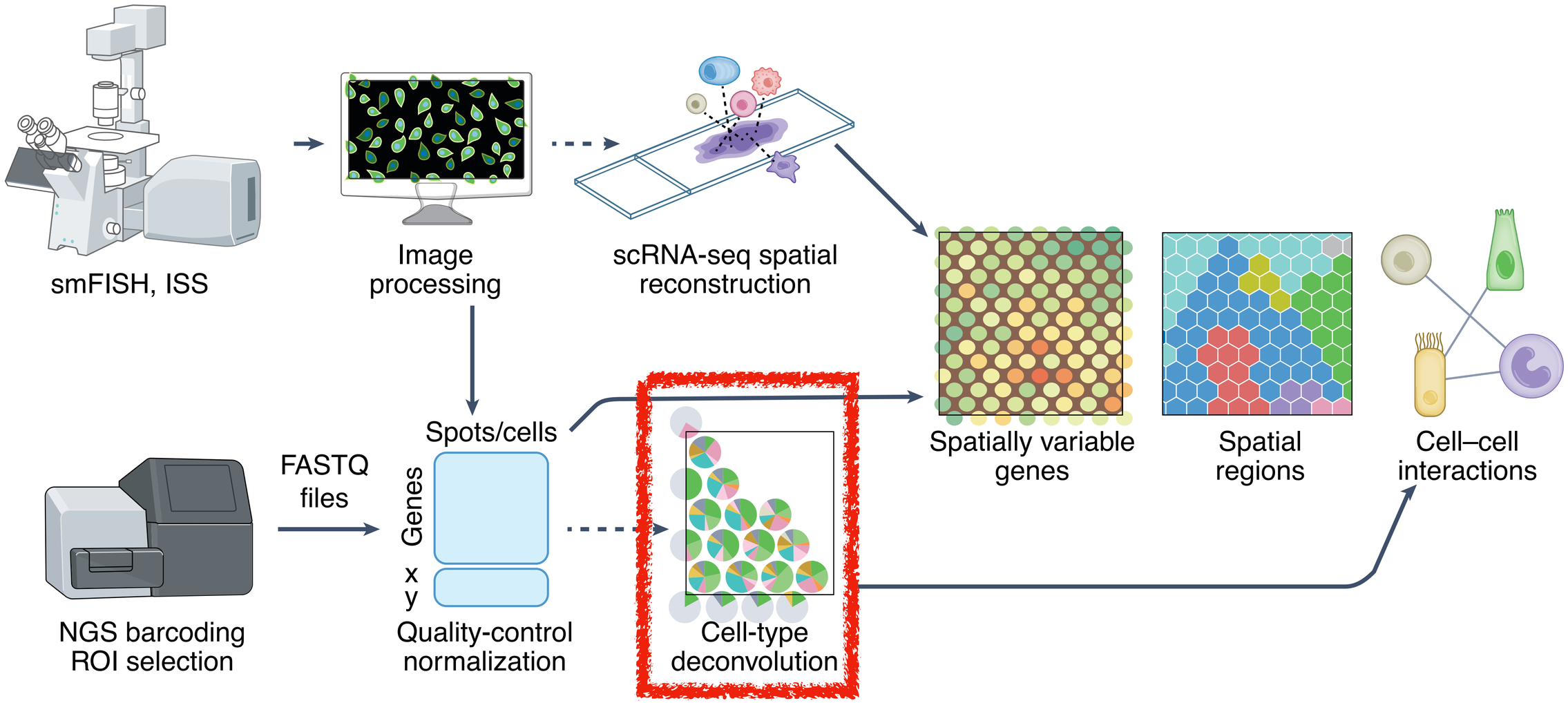
https://pachterlab.github.io/LP_2021/current-analysis.html

Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

Source code available:

`https://github.com/stat540-UBC/lectures`



Moses & Pachter *Nature Methods* (2022)

nature communications

Article

https://doi.org/10.1038/s41467-023-37168-7

A comprehensive benchmarking with practical guidelines for cellular deconvolution of spatial transcriptomics

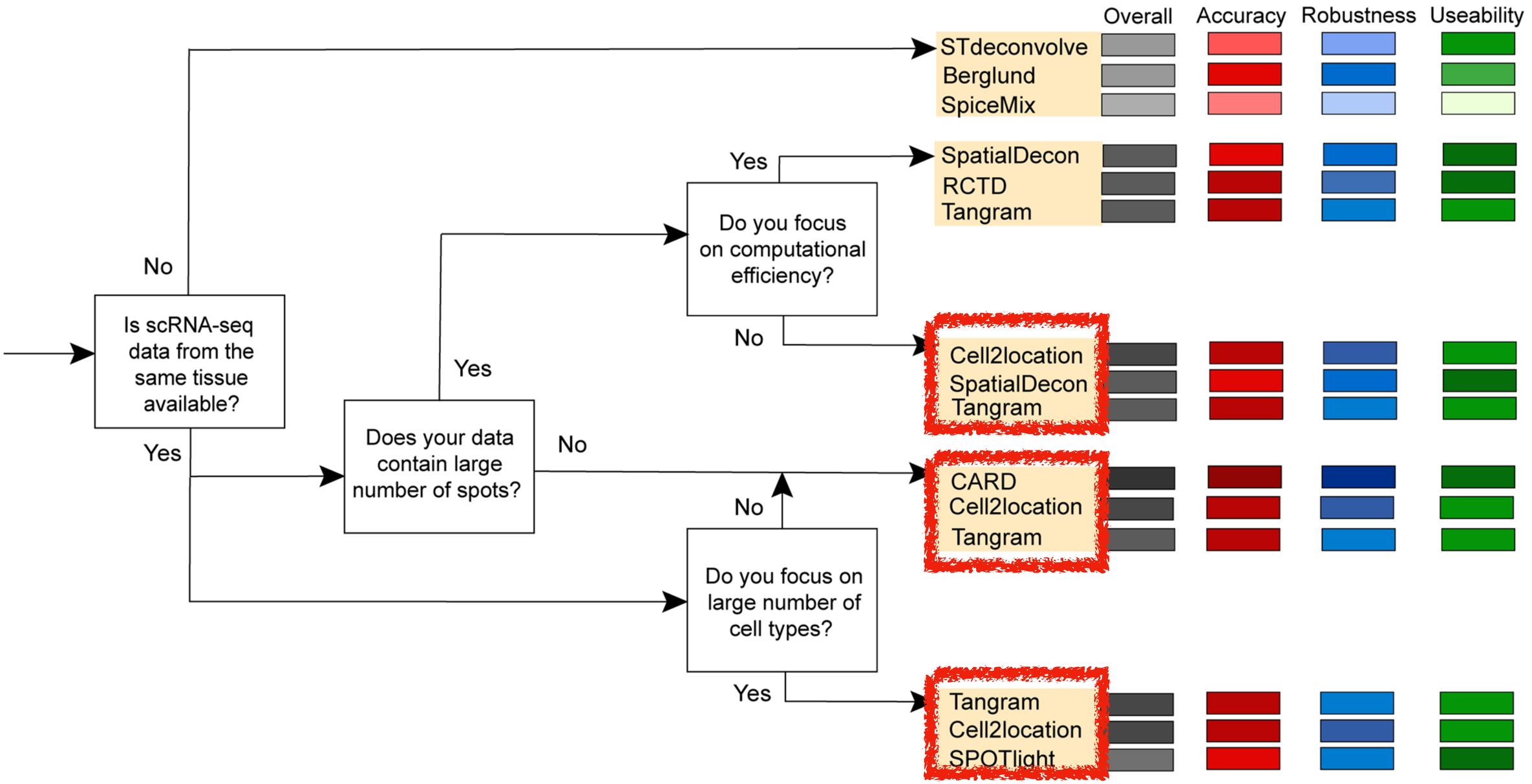
Received: 30 September 2022

Accepted: 3 March 2023

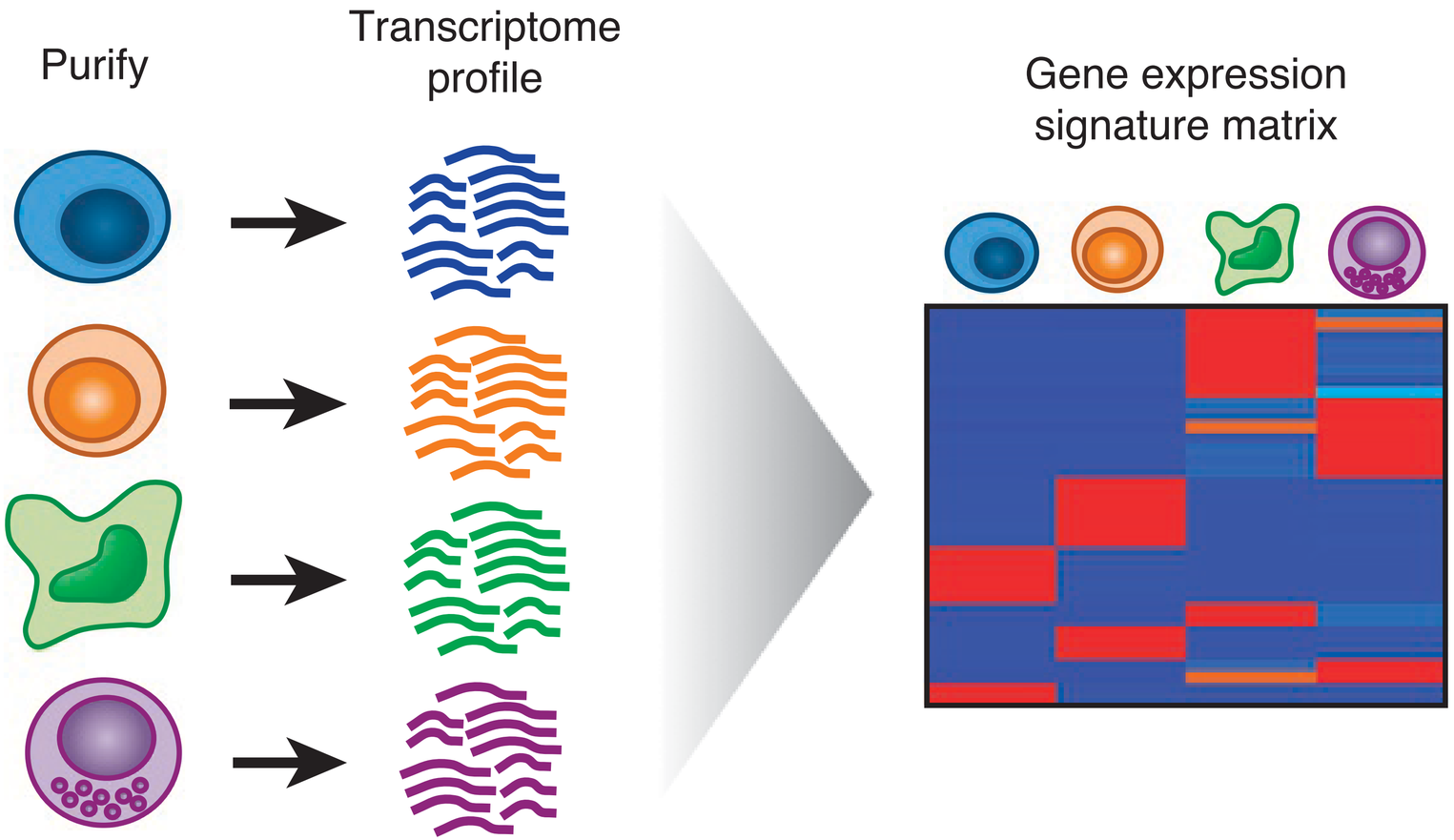
Published online: 21 March 2023

Haoyang Li[1,2,6], Juexiao Zhou[1,2,6], Zhongxiao Li[1,2], Siyuan Chen[1,2], Xingyu Liao[1,2], Bin Zhang[1,2], Ruochi Zhang[3], Yu Wang[3], Shiwei Sun[4,5] & Xin Gao[1,2]

robustness, and usability of the methods. We compare these methods comprehensively using different metrics, resolutions, spatial transcriptomics technologies, spot numbers, and gene numbers. In terms of performance, CARD, Cell2location, and Tangram are the best methods for conducting the cellular deconvolution task. To refine our comparative results, we provide decision-tree-style guidelines and recommendations for method selection and their additional features, which will help users easily choose the best method for fulfilling their concerns.

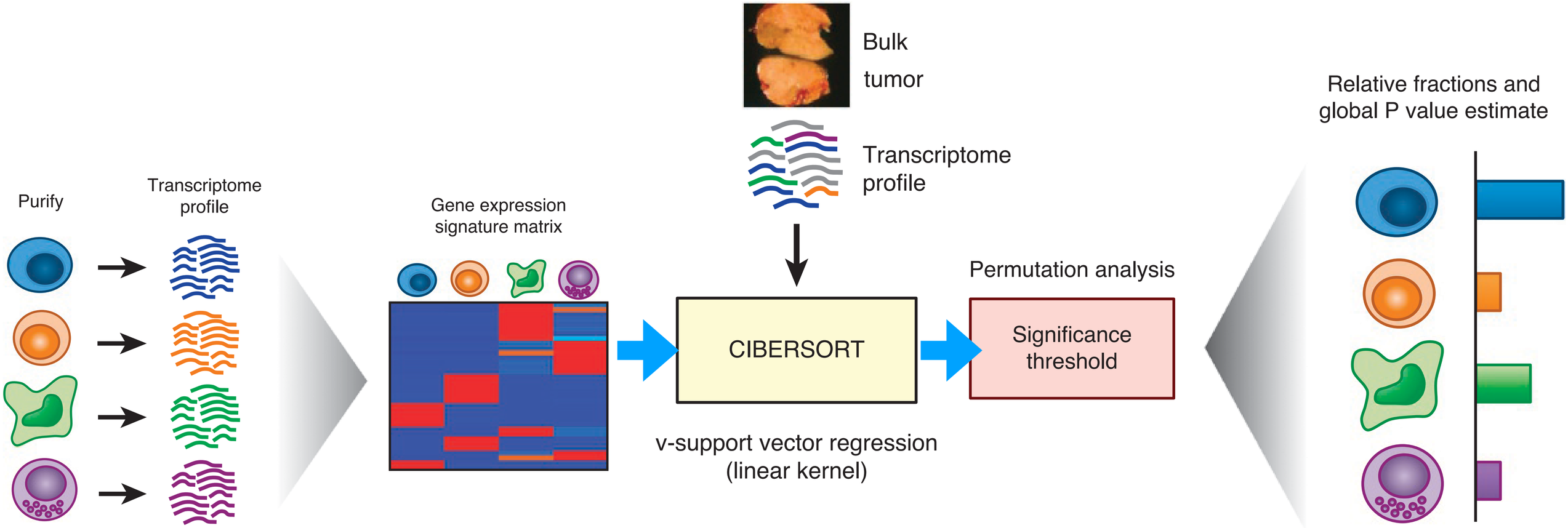

Overall
Accuracy
Robustness
Useability
Is scRNA-seq data from the same tissue available?
No
Yes
Does your data contain large number of spots?
Yes
No
Do you focus on computational efficiency?
Yes
No
Do you focus on large number of cell types?
No
Yes
STdeconvolve
Berglund
SpiceMix
SpatialDecon
RCTD
Tangram
Cell2location
SpatialDecon
Tangram
CARD
Cell2location
Tangram
Tangram
Cell2location
SPOTlight

What is a cell type deconvolution problem?



Newman .. Alizadeh, Nature Methods (2015)

What is a cell type deconvolution problem?



Newman .. Alizadeh, Nature Methods (2015)

What is a cell type deconvolution problem?

$$\mathbb{E}\left[g(Y_{gi})\right] = \sum_{t \in \textbf{cell type}} X_{gt}\pi_{ti}$$

Bulk gene expression

cell-type-specific gene expression

Cell type fraction

Newman .. Alizadeh, Nature Methods (2015)

Deconv Data 1. bulk RNA-seq data

- Human pancreatic islet gene expression data: GSE50244

```
GSE50244.bulk.eset
```

```
## ExpressionSet (storageMode: lockedEnvironment)
## assayData: 32581 features, 89 samples
##   element names: exprs
## protocolData: none
## phenoData
##   sampleNames: Sub1 Sub2 ... Sub89 (89 total)
##   varLabels: sampleID SubjectName ... tissue (7 total)
##   varMetadata: labelDescription
## featureData: none
## experimentData: use 'experimentData(object)'
## Annotation:
```

Deconv Data 2. single-cell RNA-seq data

- Single-cell RNA-seq data in the same tissue: E-MTAB-5061

```
EMTAB.sce
```

```
## class: SingleCellExperiment
## dim: 25453 1097
## metadata(0):
## assays(1): counts
## rownames(25453): SGIP1 AZIN2 ... KIR2DL2 KIR2DS3
## rowData names(1): gene.name
## colnames(1097): AZ_A10 AZ_A11 ... HP1509101_P8 HP1509101_P9
## colData names(4): sampleID SubjectName cellTypeID cellType
## reducedDimNames(0):
## mainExpName: NULL
## altExpNames(0):
```

Cell type deconvolution of the bulk RNA-seq data

Problem definition

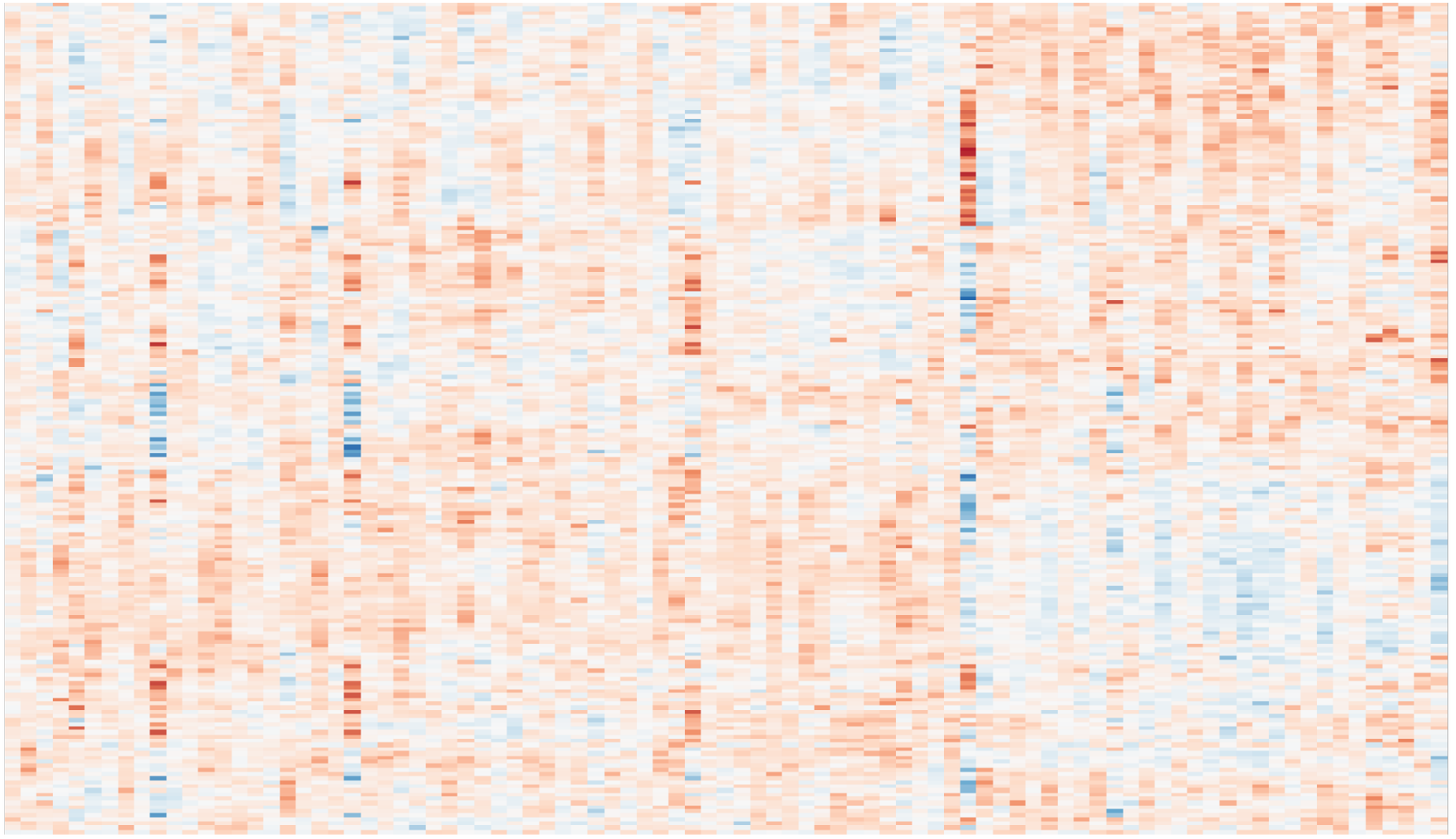
- $\mathbf{y}$: bulk gene expression (gene $\times$ sample)
- X: cell-type-specific single-cell expression (gene $\times$ cell type)

Goal

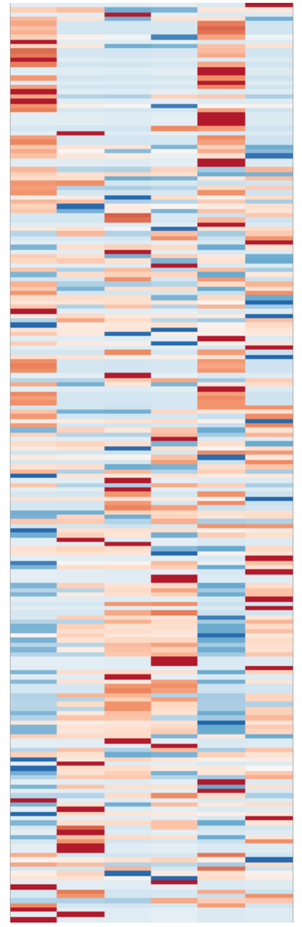
1. Fit a model regressing the bulk profile $\mathbf{y}_i$ of a sample i on the single-cell-type-specific matrix X.
2. What are the estimated cell type fractions in the bulk sample?

100 most expressed genes within each cell type

Y



X



- What are Y and X?

Negative Binomial distribution

Poisson and Gamma distributions are the building blocks of NB.

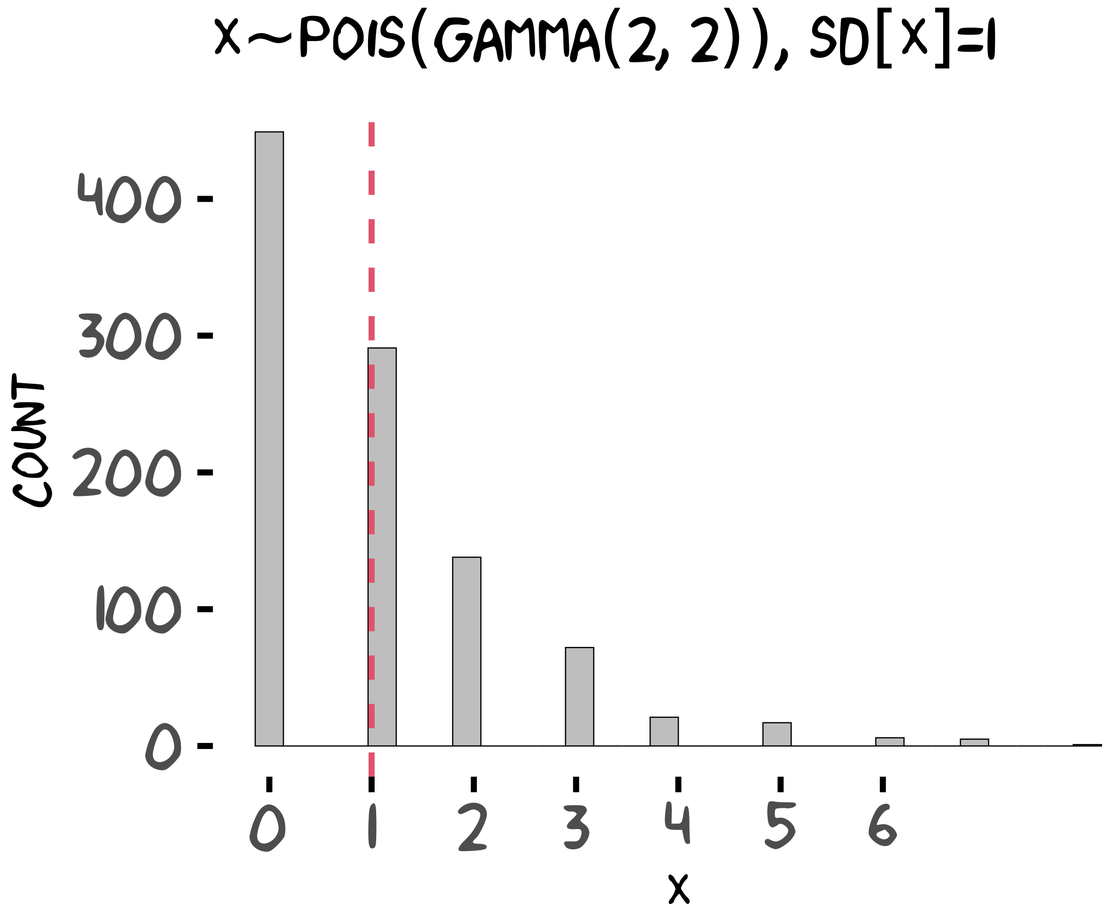
$$\lambda \sim \text{Gamma}(a, b)$$

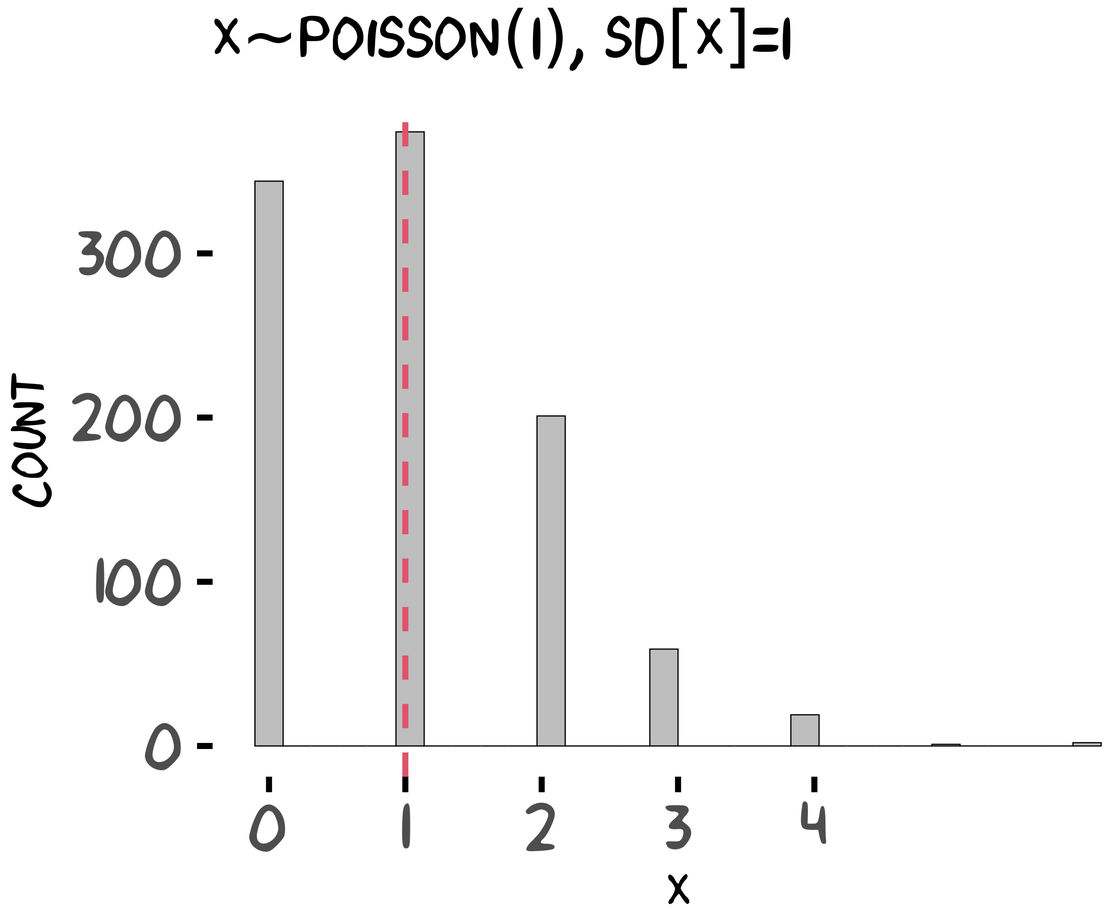
$$Y \sim \text{Poisson}(\lambda)$$

$$Y \sim \text{NB}(\mu = a/b, \phi = 1/a)$$

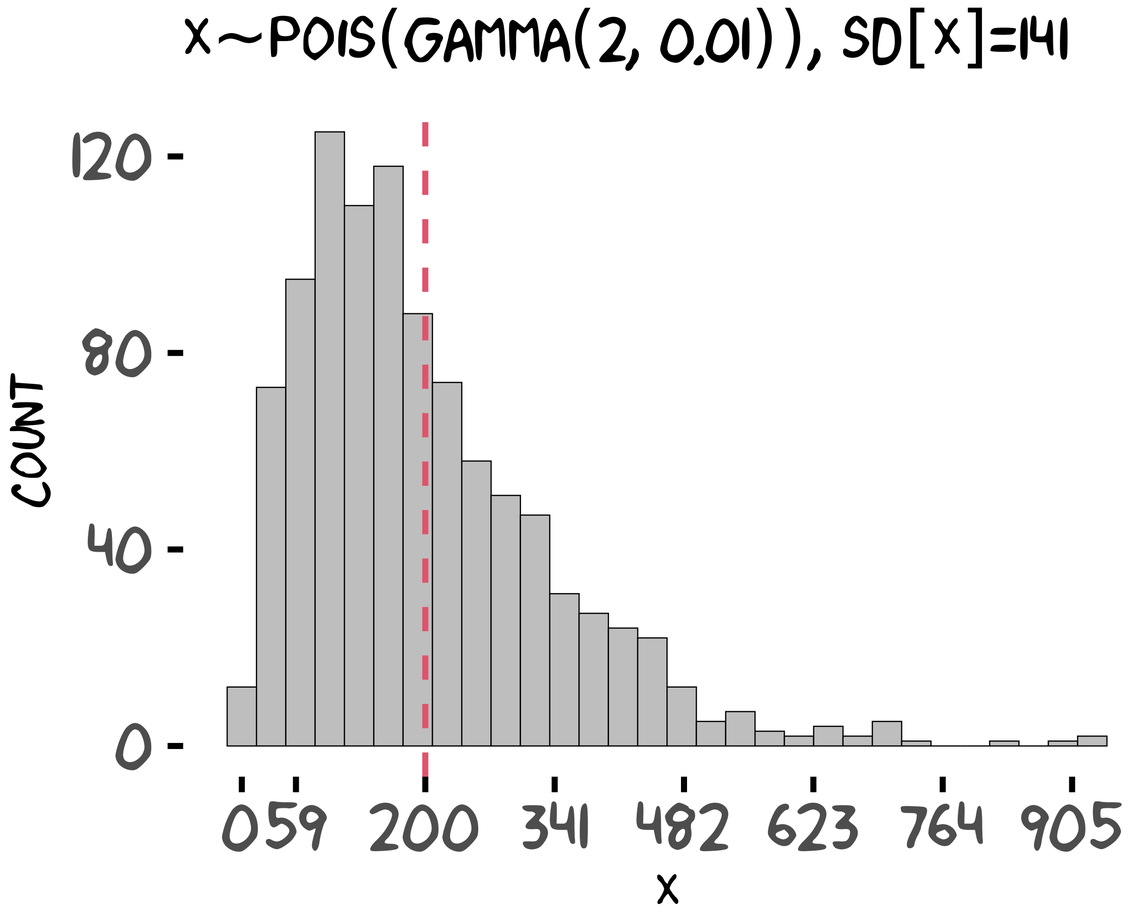
mean over dispersion

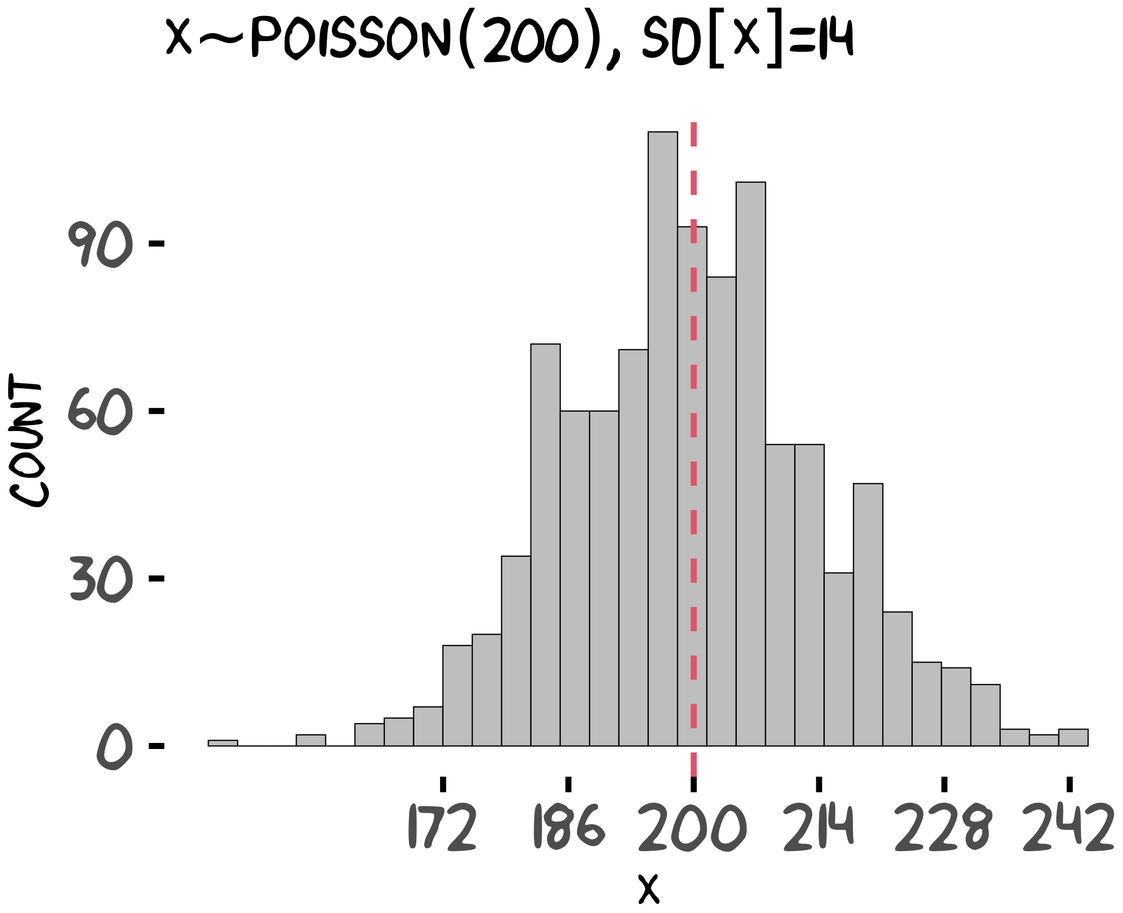
Poisson-Gamma offers flexibility in modelling variance





Poisson-Gamma offers flexibility in modelling variance





Negative Binomial GLM for modelling RNA-seq count

$$p(Y_i|\mu_i,\phi) = \int \overbrace{p(Y_i|\lambda_i)}^{\text{Poisson}} \underbrace{p(\lambda_i|\mu_i,\phi)}_{\text{Gamma}} d\lambda_i$$

where we model

$$\mu_i = \exp\left(\sum_{k=1}^{p} X_{ik}\beta_k\right).$$

Negative Binomial GLM for modelling RNA-seq count

$$
\begin{aligned}
p(Y_i|\mu_i,\phi) &= \int \overbrace{p(Y_i|\lambda_i)}^{\text{Poisson}} \underbrace{p(\lambda_i|\mu_i,\phi)}_{\text{Gamma}} d\lambda_i \\
&= \underbrace{\frac{\Gamma(Y_i+1/\phi)}{\Gamma(Y_i+1)\Gamma(1/\phi)}}_{\text{negative binomial}} \underbrace{\left(\frac{\mu_i}{1/\phi+\mu_i}\right)^{Y_i}}_{\text{success rate}} \underbrace{\left(\frac{1/\phi}{1/\phi+\mu_i}\right)^{1/\phi}}_{\text{failure rate}}
\end{aligned}
$$

where we model

$$\mu_i = \exp\left(\sum_{k=1}^{p} X_{ik}\beta_k\right).$$

Negative Binomial GLM for modelling RNA-seq count

- Y: number of successfully “observed” reads in RNA-seq (~targeting)
- r: number of permitted “dropped” reads until Y observed (~budget)
- ρ: success rate

$$p(Y_i|\mu_i,\phi) = \underbrace{\binom{Y_i+r-1}{Y_i}}_{\text{negative binomial}} \underbrace{\rho_i^{Y_i}}_{\text{success rate}} \underbrace{(1-\rho_i)^r}_{\text{drop rate}}$$

Negative Binomial GLM for modelling RNA-seq count

- Y: number of successfully “observed” reads in RNA-seq (~targeting)
- r: number of permitted “dropped” reads until Y observed (~budget)
- ρ: success rate

$$
\begin{aligned}
p(Y_i|\mu_i,\phi) &= \underbrace{\binom{Y_i+r-1}{Y_i}}_{\text{negative binomial}} \underbrace{\rho_i^{Y_i}}_{\text{success rate}} \underbrace{(1-\rho_i)^r}_{\text{drop rate}} \\
&= \mathsf{NB}(Y_i|r=\phi^{-1}, \rho_i = \mu_i/(\phi^{-1}+\mu_i))
\end{aligned}
$$

Negative Binomial GLM for modelling RNA-seq count

- Y: number of successfully "observed" reads in RNA-seq (~targeting)
- r: number of permitted "dropped" reads until Y observed (~budget)
- ρ: success rate

$$
\begin{aligned}
p(Y_i|\mu_i,\phi) &= \underbrace{\binom{Y_i+r-1}{Y_i}}_{\text{negative binomial}} \underbrace{\rho_i^{Y_i}}_{\text{success rate}} \underbrace{(1-\rho_i)^r}_{\text{drop rate}} \\
&= \text{NB}(Y_i|r=\phi^{-1}, \rho_i=\mu_i/(\phi^{-1}+\mu_i)) \\
\text{or} &= \text{NB}(Y_i|\text{mean}=\mu_i, \text{overdispersion}=\phi)
\end{aligned}
$$

Negative Binomial GLM for modelling RNA-seq count

- Y: number of successfully "observed" reads in RNA-seq (~targeting)
- r: number of permitted "dropped" reads until Y observed (~budget)
- ρ: success rate

$$
\begin{aligned}
p(Y_i|\mu_i,\phi) &= \underbrace{\binom{Y_i+r-1}{Y_i}}_{\text{negative binomial}} \underbrace{\rho_i^{Y_i}}_{\text{success rate}} \underbrace{(1-\rho_i)^r}_{\text{drop rate}} \\
&= \text{NB}(Y_i|r=\phi^{-1}, \rho_i=\mu_i/(\phi^{-1}+\mu_i)) \\
\text{or} &= \text{NB}(Y_i|\text{mean}=\mu_i, \text{overdispersion}=\phi)
\end{aligned}
$$

We can check:

mean: $\mathbb{E}[Y_i|r,\rho] = \rho r/(1-\rho) = \mu_i$

NB GLM is useful for deconvolution

Q: Can we estimate cell type fractions in tissue-level *bulk* data?

$$\overset{\text{bulk}}{\mathbf{y}_i} \sim \text{NB}\left(\text{mean} = \overset{\text{scale factor}}{s_i} \sum_t X_{gt} \overset{\text{GOAL}}{\theta_{ti}}, \text{overdispersion} = \phi\right)$$

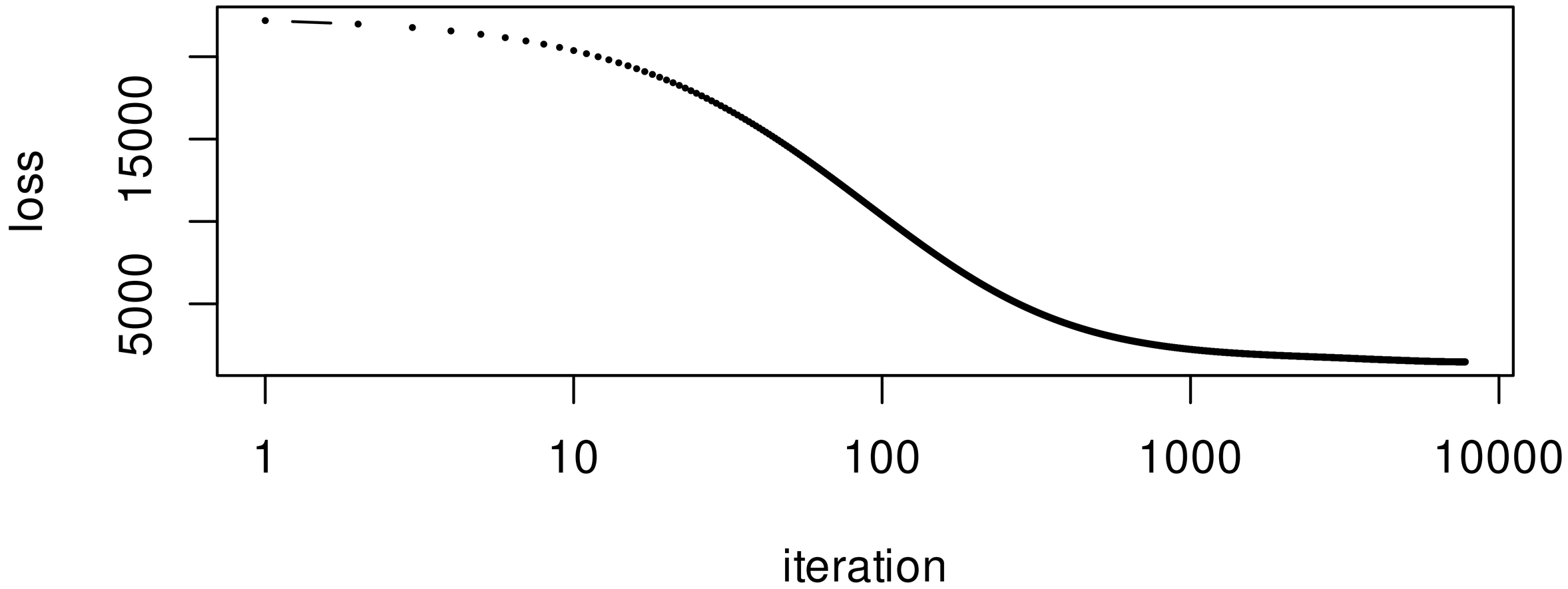
cell-type-sorted

- We use the same data set used in the vignette of MuSic package (Wang *et al.*, Nature Comm., 2019)

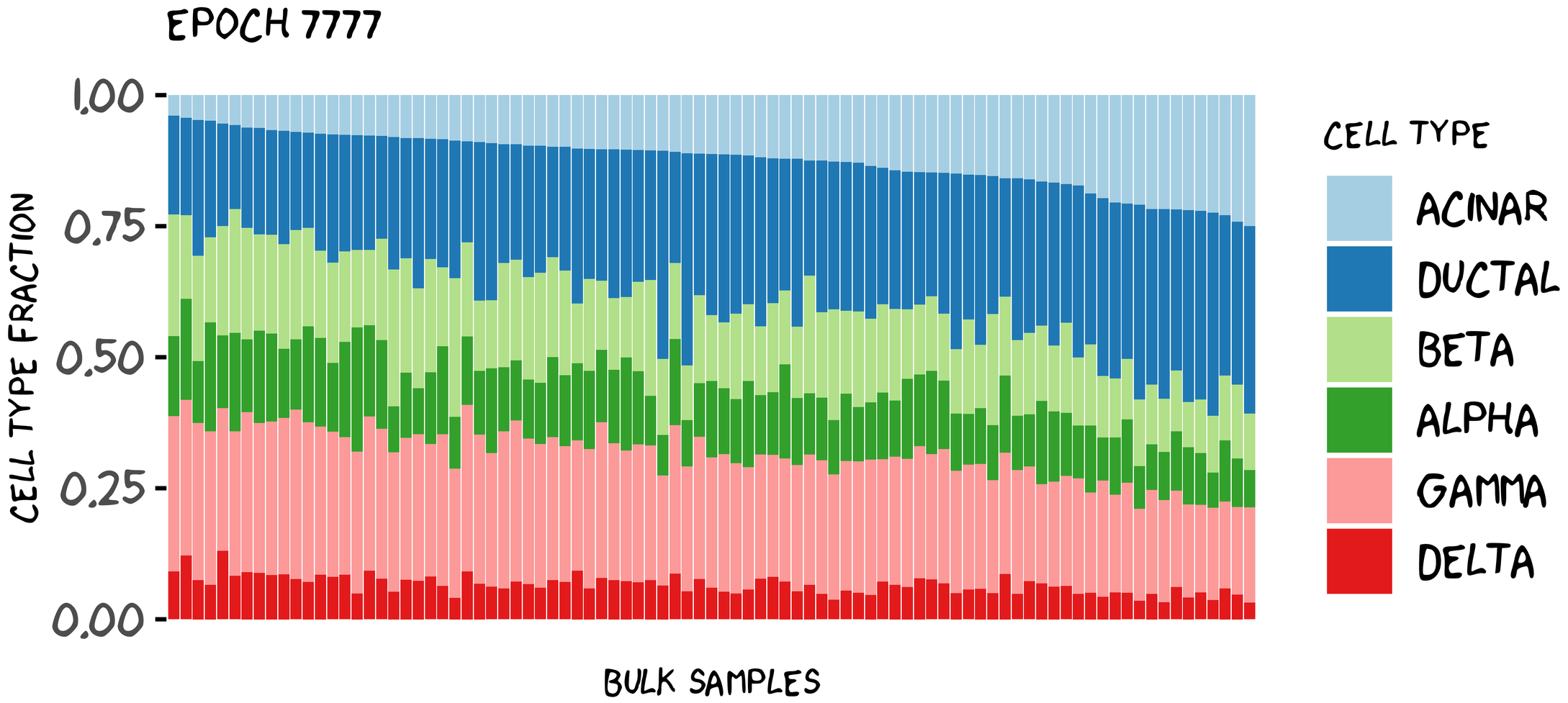
Fit the NB GLM model to the data

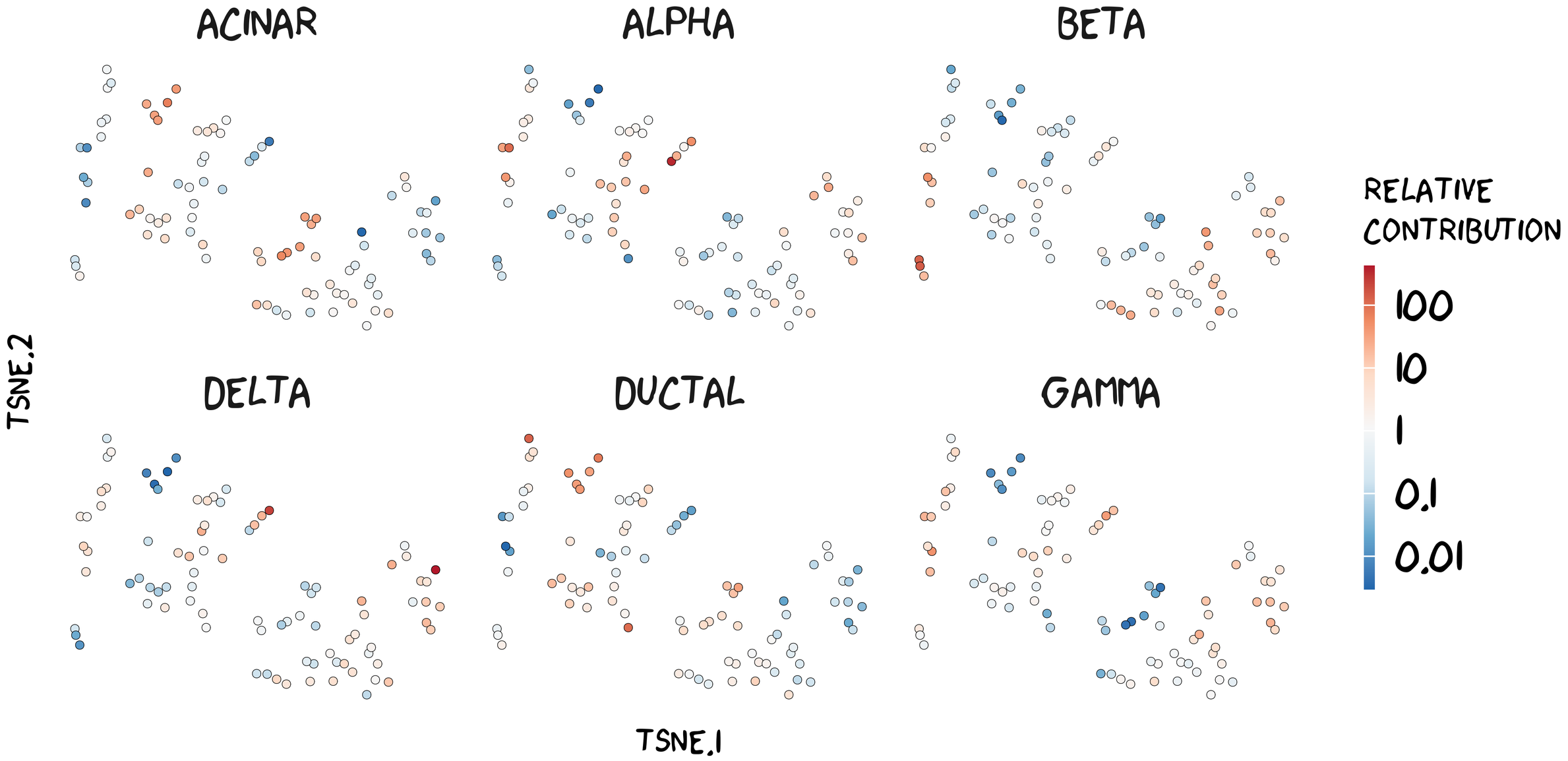
```
nb.glm <- nn_module(
    classname = "NB GLM",
    initialize = function(n, p, m) { # parameters
        self$logit.theta <- nn_parameter(torch_randn(p, m) * 0.01)
        self$log.scale <- nn_parameter(torch_zeros(1, m))
        self$log.phi <- nn_parameter(torch_zeros(n, 1))
    },
    forward = function(xx, yy) { # -log-likelihood
        theta <- torch_exp(nnf_log_softmax(self$logit.theta, dim=1))
        log.mu <- torch_log(torch_mm(xx, theta) * torch_exp(self$log.scale))
        od <- torch_exp(-self$log.phi) + 1e-8

        .term1 <- (torch_lgamma(yy + od) -
                   torch_lgamma(yy + 1) -
                   torch_lgamma(od))
        .term2 <- - nnf_softplus(-log.mu - self$log.phi) * yy
        .term3 <- - nnf_softplus(log.mu + self$log.phi) * od

        llik <- .term1 + .term2 + .term3
        torch_sum(llik, dim=1)
    })
```

Fit the NB GLM model to the data



Show the cell type fraction estimates

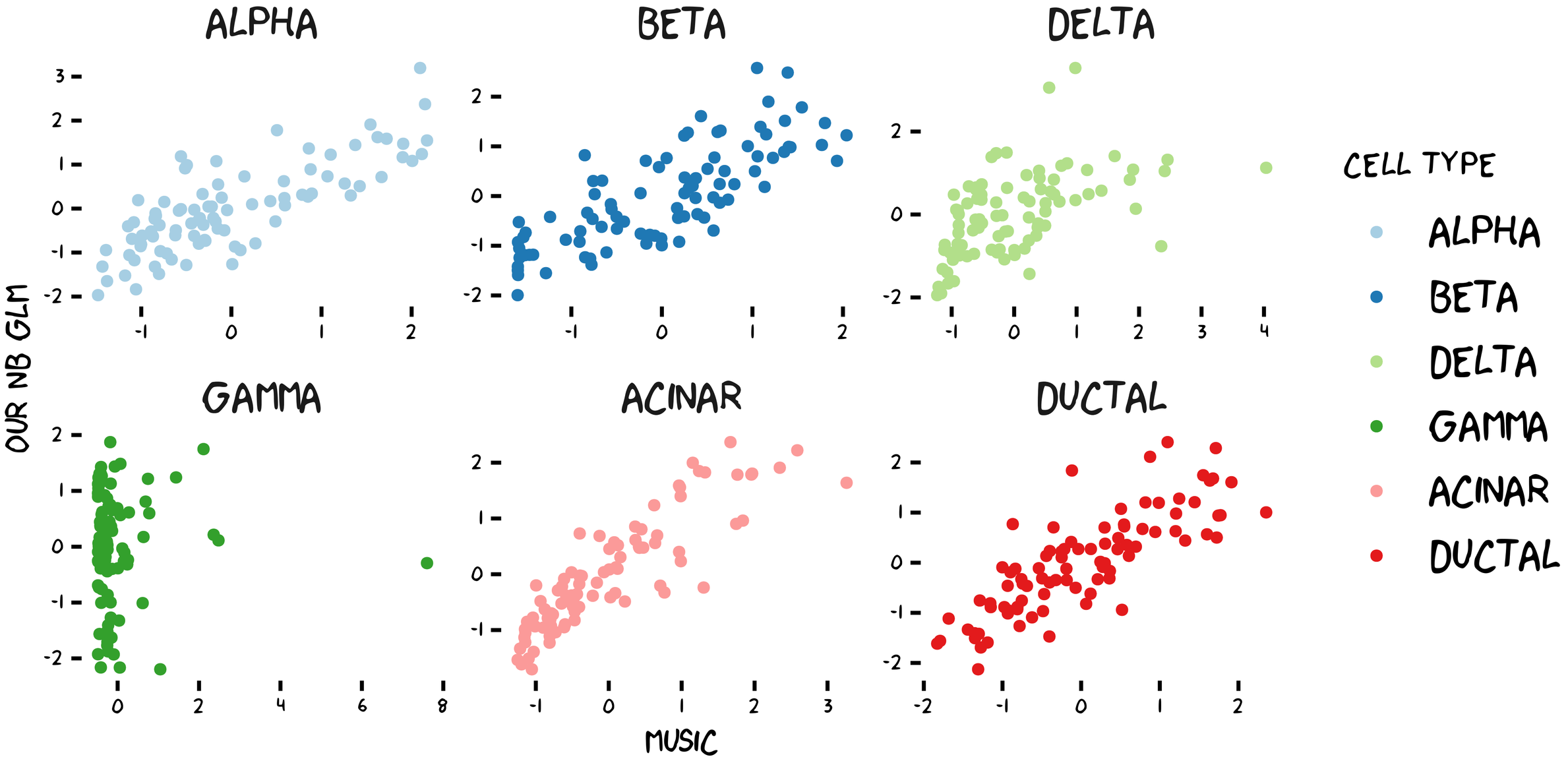



ACINAR
ALPHA
BETA
DELTA
DUCTAL
GAMMA
TSNE.2
TSNE.1
RELATIVE CONTRIBUTION
100
10
1
0.1
0.01

MuSiC: Multi-subject Single-cell Deconvolution

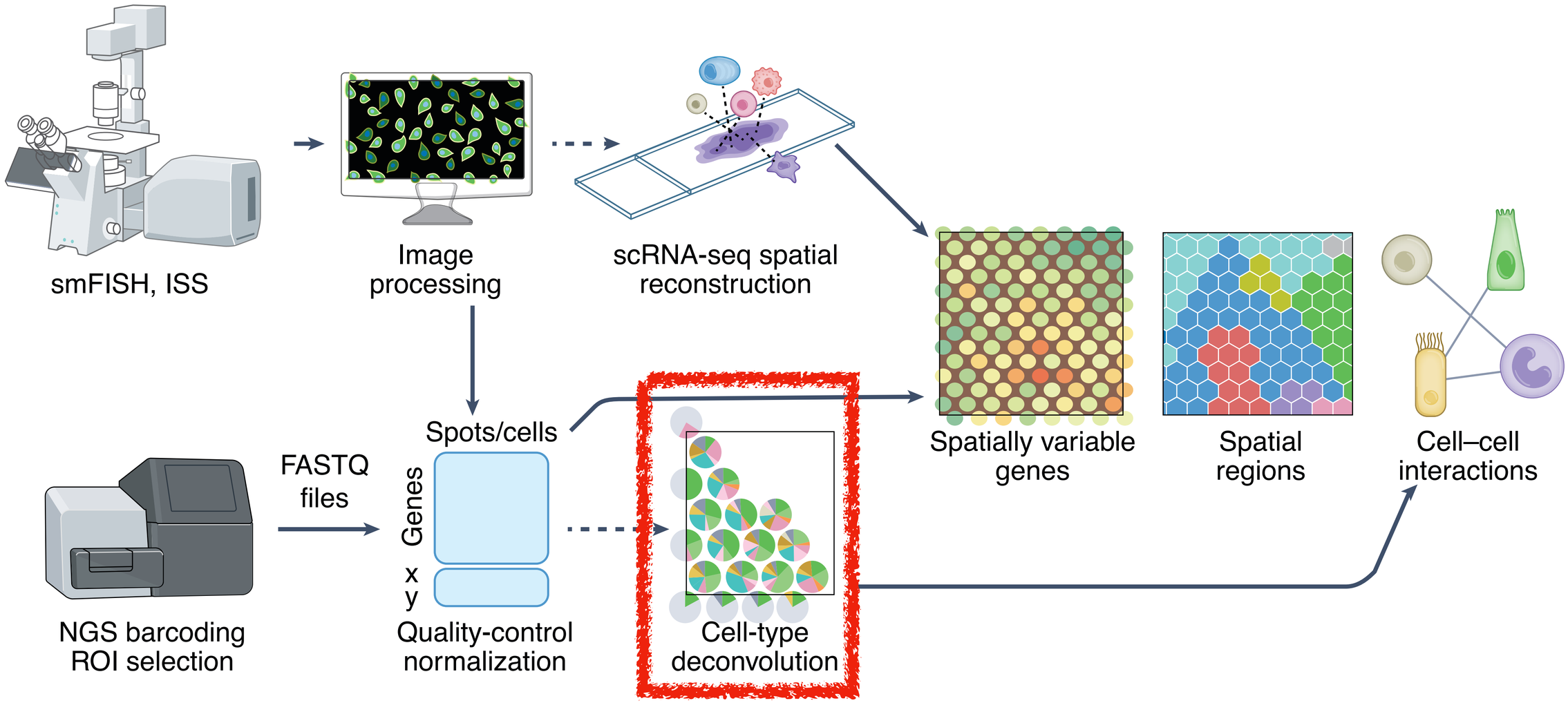
```
library(MuSiC)

bulk.mtx <- exprs(GSE50244.bulk.eset) # Bulk expression

est.prop <- music_prop(bulk.mtx = bulk.mtx,
                       sc.sce = EMTAB.sce,
                       clusters = 'cellType', samples = 'sampleID',
                       select.ct = c('alpha', 'beta', 'delta', 'gamma', 'acinar', 'ductal'),
                       verbose = F)

names(est.prop)
```

```
## [1] "Est.prop.weighted" "Est.prop.allgene"  "Weight.gene"
## [4] "r.squared.full"    "Var.prop"
```


ALPHA
BETA
DELTA
GAMMA
ACINAR
DUCTAL
OUR NB GLM
MUSIC
CELL TYPE
ALPHA
BETA
DELTA
GAMMA
ACINAR
DUCTAL

Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping



Moses & Pachter *Nature Methods* (2022)

Can we bring spatial contexts to deconvolution?

We will take a look at one spatial deconvolution method.

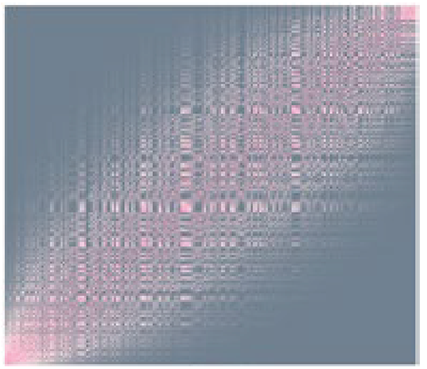
`https://yma-lab.github.io/CARD/`

Conditional autoregressive model-based deconvolution

$$X = \beta\theta^T + E, E_{gi} \sim N(0, \sigma_e^2)$$

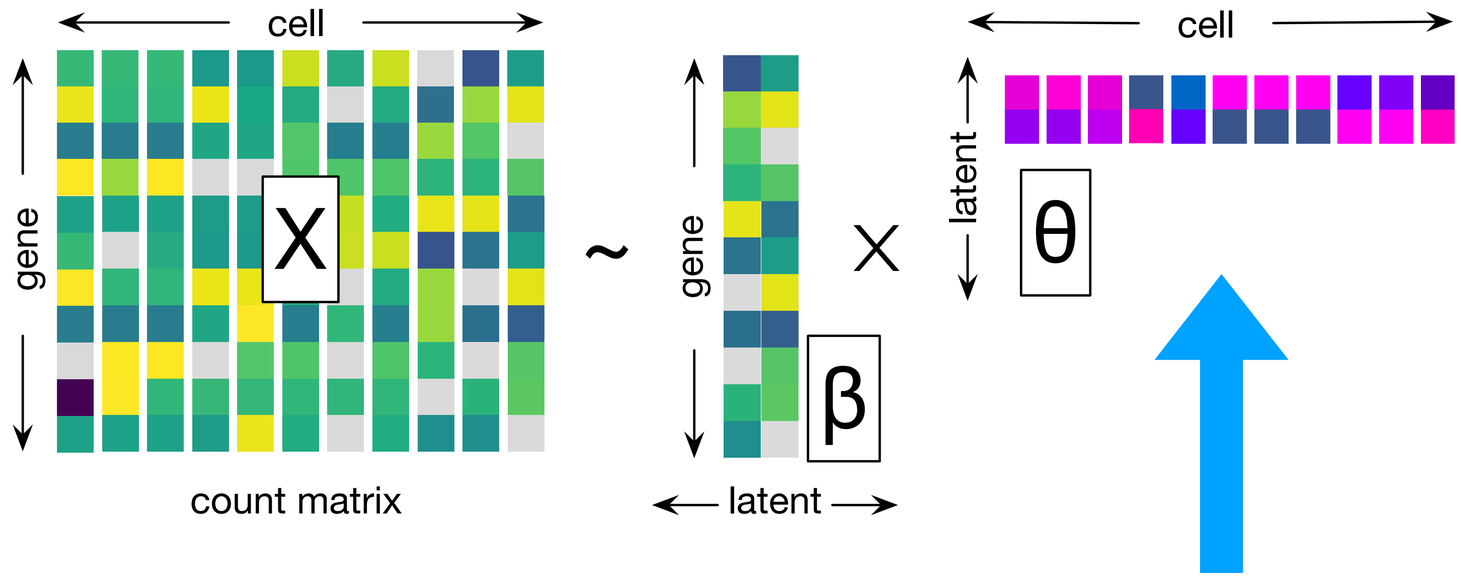
$$\theta_{ik} = b_k + \phi \sum_{j=1, j \neq i}^{n} W_{ij}(\theta_{jk} - b_k) + \varepsilon_{ik} + \varepsilon_{ik} \sim N(0, \sigma_{ik}^2)$$

Gaussian kernel

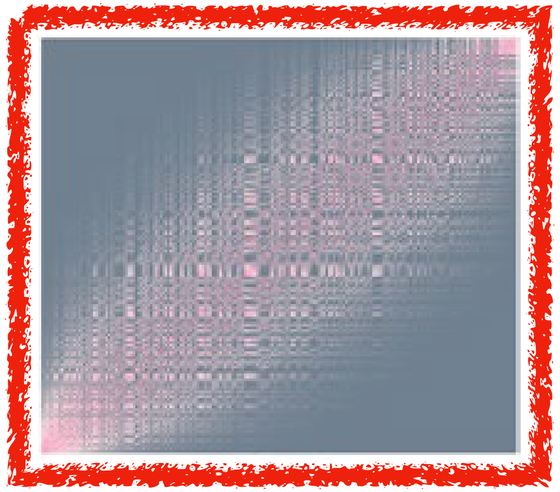


How do we incorporate spatial information?

$$X = \beta\theta^\top + E$$

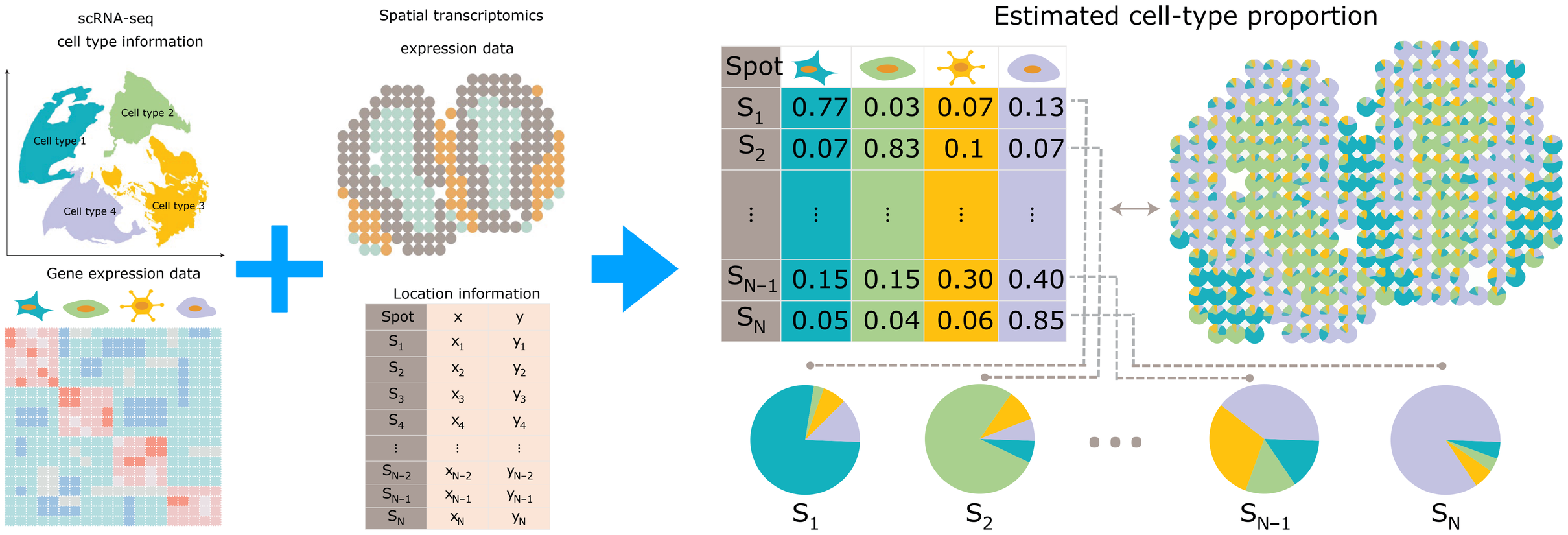


$$\theta_{ik} = b_k + \phi \sum_{j \neq i} W_{ij}(\theta_{jk} - b_k) + \epsilon$$



cell-cell similarity

Conditional AutoRegressive model-based Deconvolution

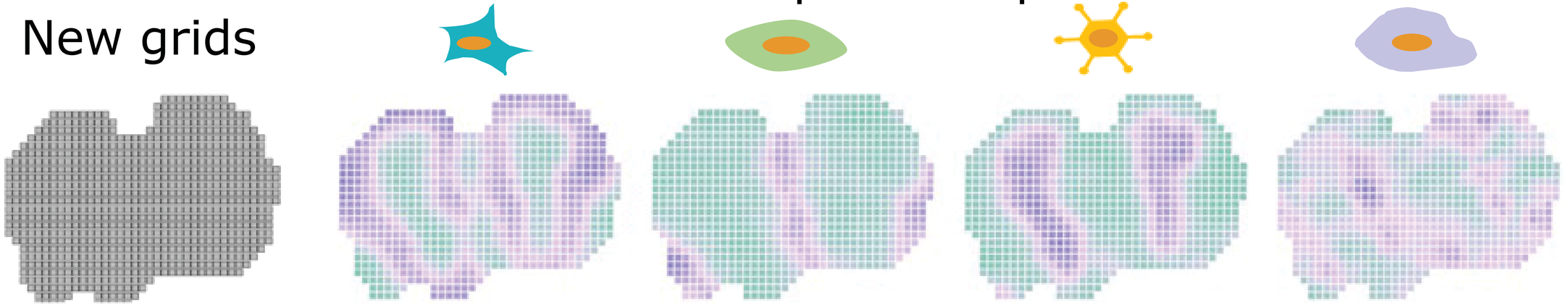


Ma & Zhou, *Nature Biotech.* (2022)

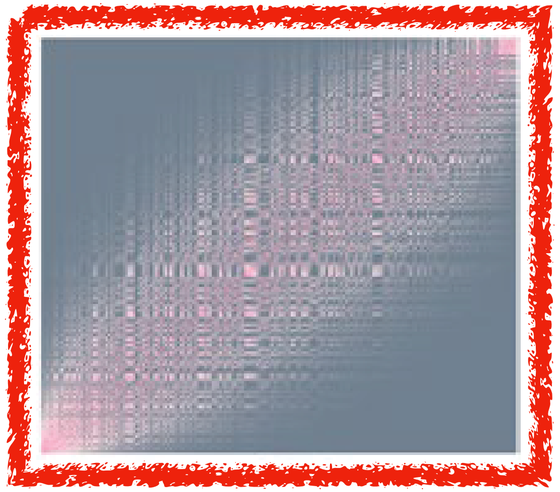
Once we have deconvolution results... Refine it

Refined spatial map

New grids



$$\theta_{ik} = b_k + \phi \sum_{j \neq i} W_{ij}(\theta_{jk} - b_k) + \epsilon$$



cell-cell similarity

Ma & Zhou, *Nature Biotech.* (2022)

What are the data matrices in spatial deconvolution?

1 Count data for spots:

```
dim(spatial_count)
```

```
## [1] 25753   428
```

```
spatial_count[1:3, 1:5]
```

```
## 3 x 5 sparse Matrix of class "dgCMatrix"
##             10x10 10x13 10x14 10x15 10x16
## X5S_rRNA        .     .     .     .     .
## X5_8S_rRNA      .     .     .     .     .
## X7SK            .     .     .     .     .
```

2 Spot coordinates:

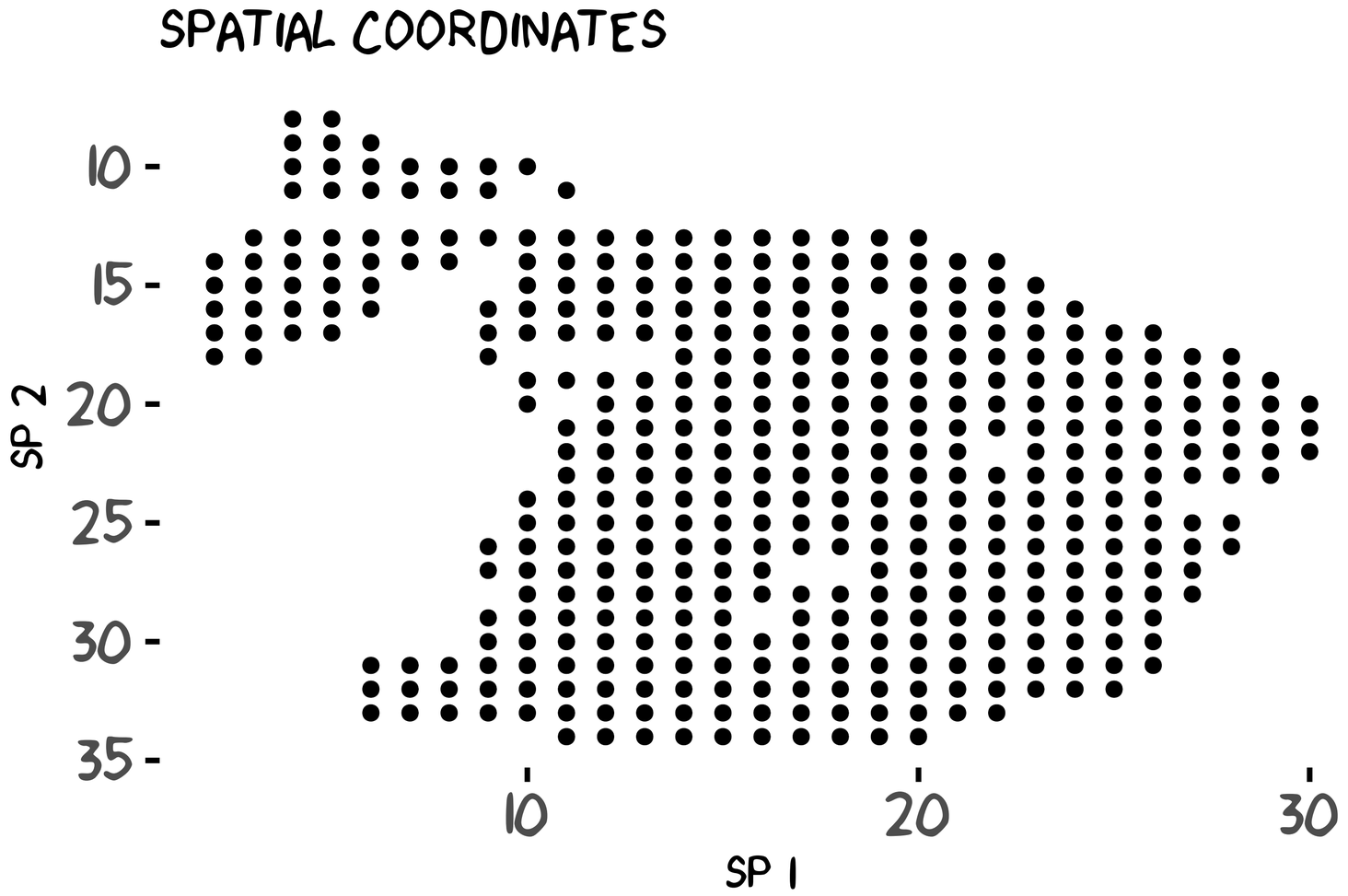
```
dim(spatial_location)
```

```
## [1] 428   2
```

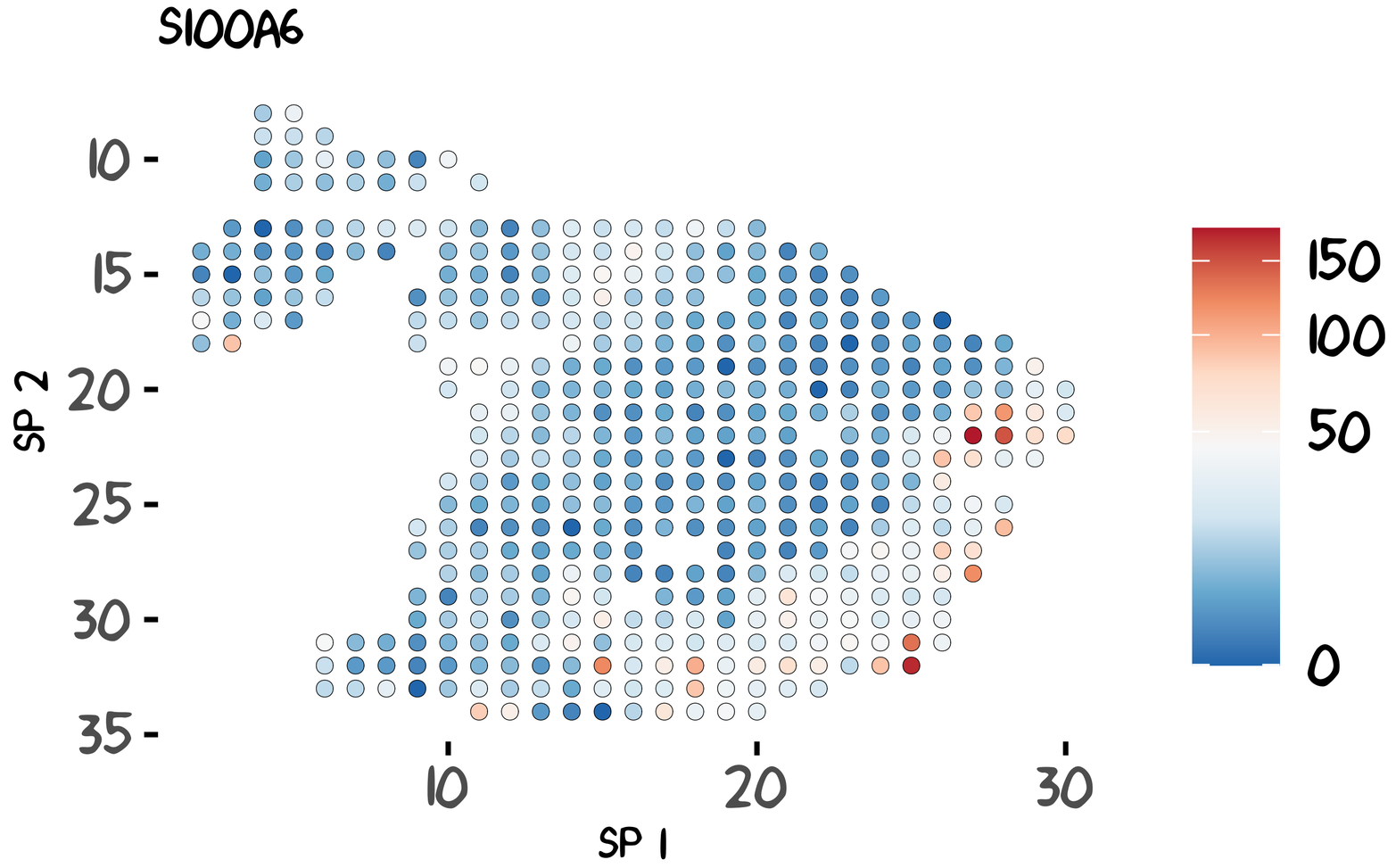
```
spatial_location[1:3,]
```

```
##         x  y
## 10x10 10 10
## 10x13 10 13
## 10x14 10 14
```

Spatial coordinates of spots in 2D



Spatial transcriptomic data = expression within each spot



ST technology also provides scRNA-seq matrix

```
dim(sc_count)
```

```
## [1] 19736  1926
```

```
sc_count[1:5, 1:5]
```

```
## 5 x 5 sparse Matrix of class "dgCMatrix"
##         Cell1 Cell2 Cell3 Cell4 Cell5
## A1BG        .     .     .     .     .
## A1CF        .     .     .     1     .
## A2M         .     .     .     .     .
## A2ML1       .     .     .     .     .
## A3GALT2     .     .     .     .     .
```

```
head(sc_meta)
```

```
##       cellID                                cellType
## Cell1  Cell1                           Acinar_cells
## Cell2  Cell2          Ductal_terminal_ductal_like
## Cell3  Cell3          Ductal_terminal_ductal_like
## Cell4  Cell4 Ductal_CRISP3_high-centroacinar_like
## Cell5  Cell5                       Cancer_clone_A
## Cell6  Cell6                       Cancer_clone_A
```

CARD deconvolves cell types in spatial contexts

```
library(CARD)
CARD_obj = createCARDObject(
    sc_count = sc_count,
    sc_meta = sc_meta,
    spatial_count = spatial_count,
    spatial_location = spatial_location,
    ct.varname = "cellType",
    ct.select = unique(sc_meta$cellType),
    sample.varname = "sampleInfo",
    minCountGene = 100,
    minCountSpot = 5)
```

```
## ## QC on scRNASeq dataset! ...
## ## QC on spatially-resolved dataset! ...
```

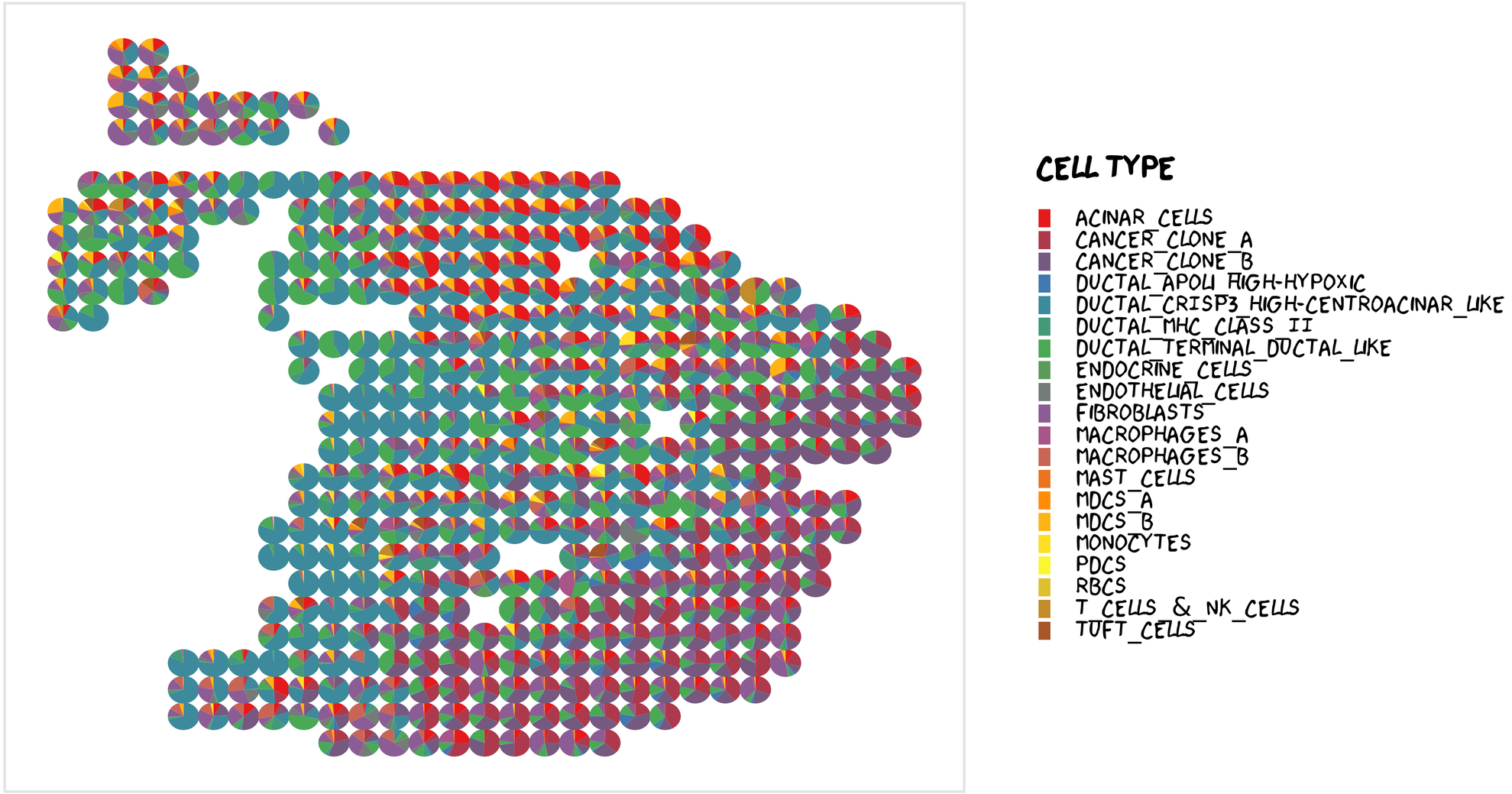
```
CARD_obj = CARD_deconvolution(CARD_object = CARD_obj)
```

```
## ## create reference matrix from scRNASeq...
## ## Select Informative Genes! ...
## ## Deconvolution Starts! ...
## ## Deconvolution Finish! ...
```

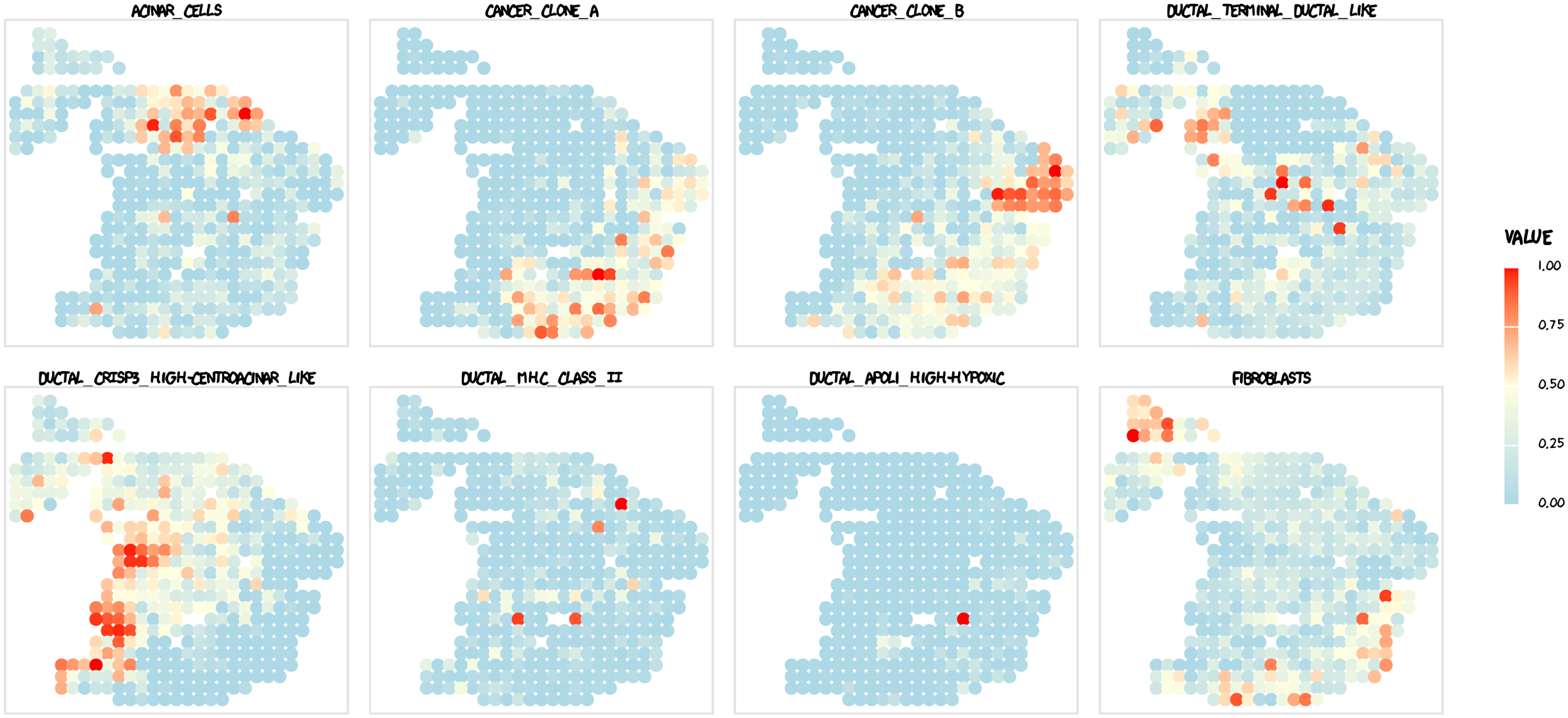
Spatial pie charts using `scatterpie`

```
CARD.visualize.pie(
    proportion = CARD_obj@Proportion_CARD,
    spatial_location = CARD_obj@spatial_location,
    colors = colors,
    radius = 0.52)
```

Spatial pie charts using `scatterpie`



Show multiple cell types' spatial distributions

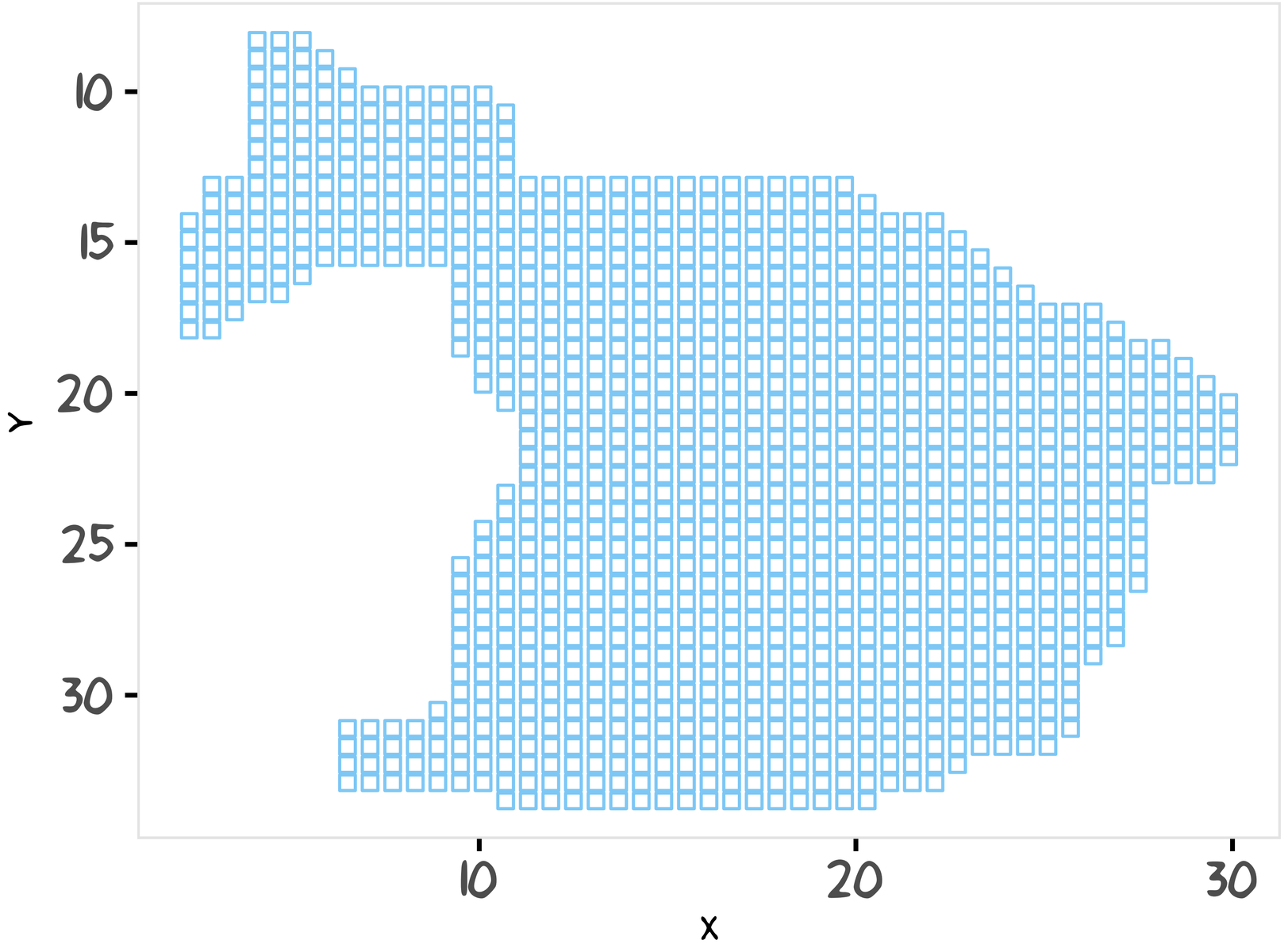


Improve spatial resolution by interpolation

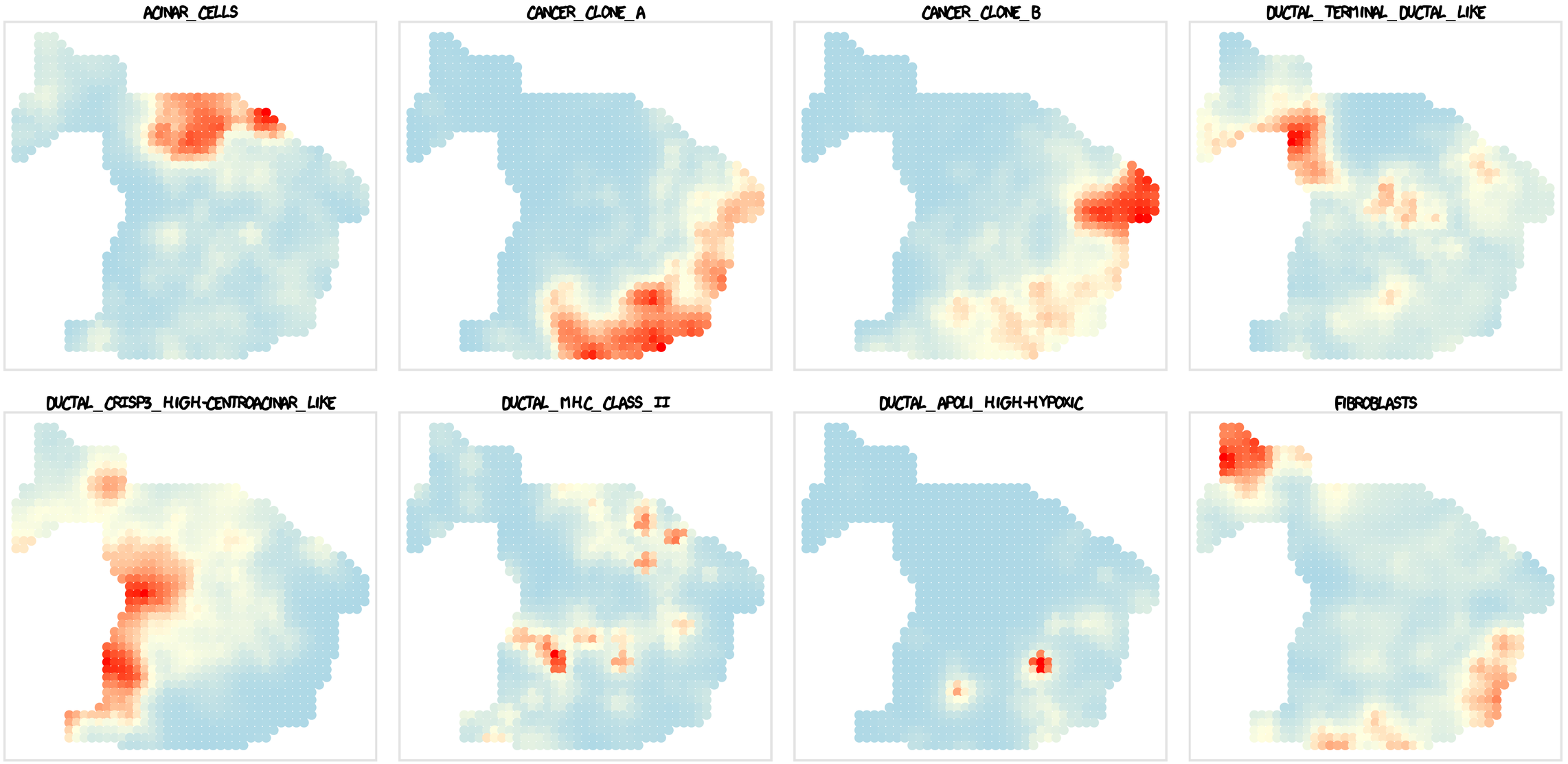
```
CARD_obj <- CARD.imputation(CARD_obj, NumGrids = 2000, ineibor = 10, exclude = NULL)
```

```
## ## The rownames of locations are matched ...
## ## Make grids on new spatial locations ...
```

Improve spatial resolution in a finer grid system



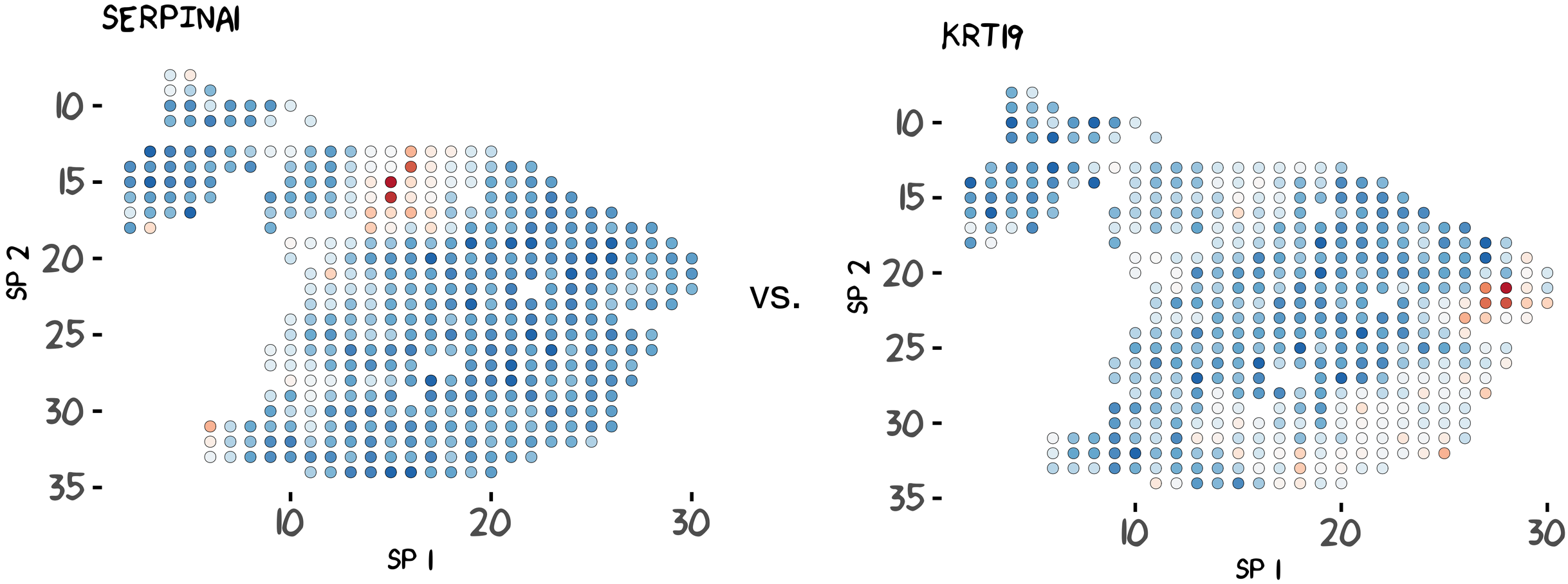
Improve spatial resolution



Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

How can we test significant spatial expressions?

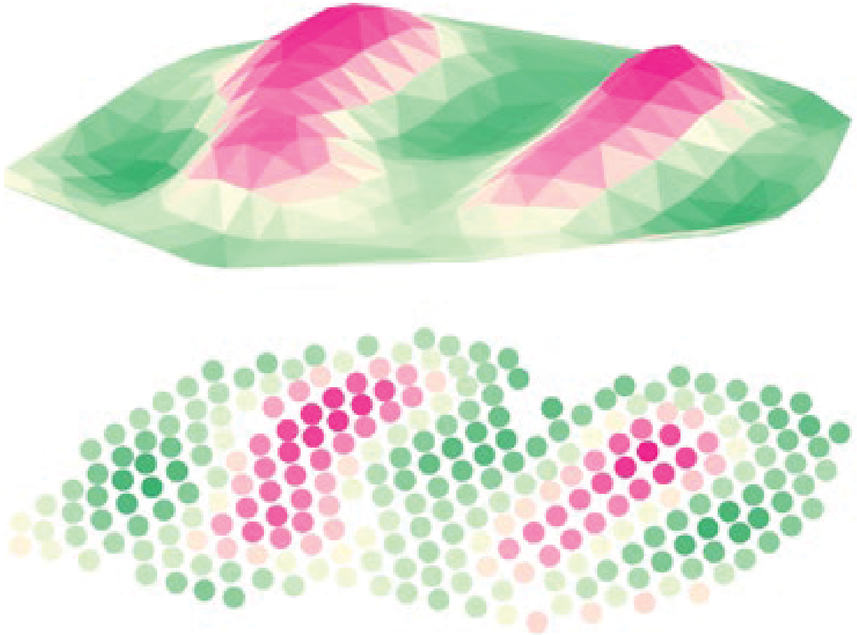


Which one is more significantly enriched in spatial locations?

SPARK: spatial pattern recognition via kernels

similarity scores

Spatial expression pattern



$$y_i \sim \text{Pois}(N_i, \lambda_i)$$

$$\log \lambda_i = \mathbf{x}_i^\top \beta + b_i + \epsilon_i$$

$$\mathbf{b} = (b_1, \ldots, b_n)^\top \sim \text{N}(\mathbf{0}, \tau_1 K)$$

$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim \text{N}(\mathbf{0}, \tau_2 I)$$

Sun .. Zhou, *Nature Methods* (2020)

Generalized linear spatial model

$$\log \lambda_i = \mathbf{x}_i^\top \beta + b_i + \epsilon_i$$

covariates

random spatial effects

$$\mathbf{b} = (b_1, \ldots, b_n)^\top \sim \mathrm{N}(\mathbf{0}, \tau_1 K)$$

$$H_0 : \tau_1 = 0 \quad \text{vs.} \quad \tau_1 > 0$$

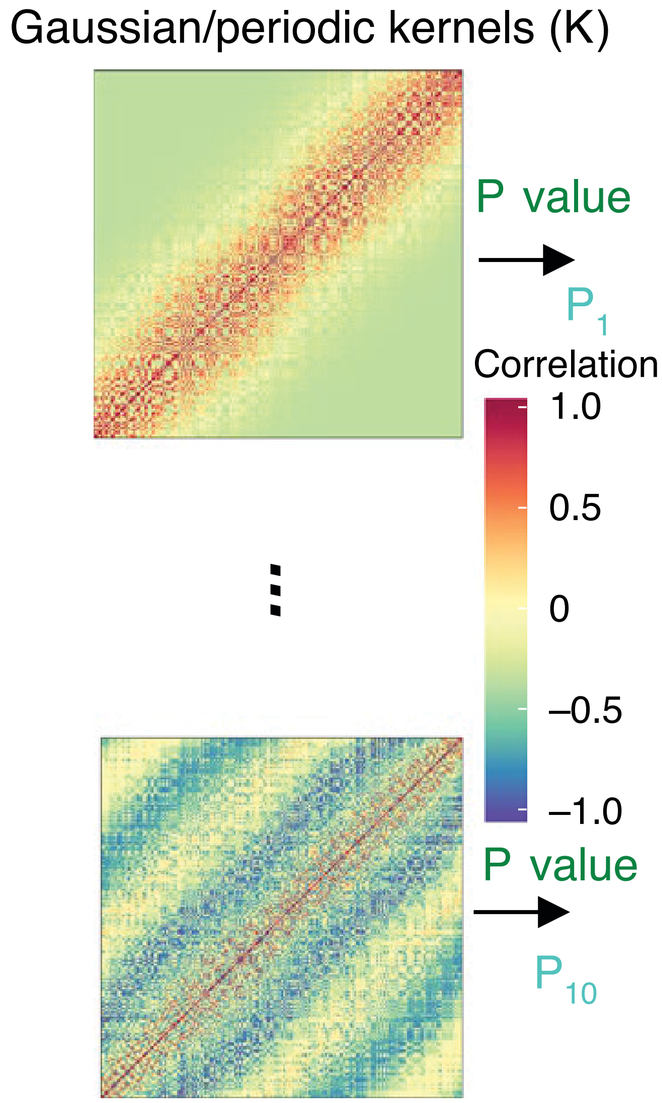
SPARK's generalized linear spatial model

$$y_i \sim \text{Pois}(N_i, \lambda_i)$$

$$\log \lambda_i = \mathbf{x}_i^\top \beta + b_i + \epsilon_i$$

$$\mathbf{b} = (b_1, \ldots, b_n)^\top \sim \text{N}(\mathbf{0}, \tau_1 K)$$

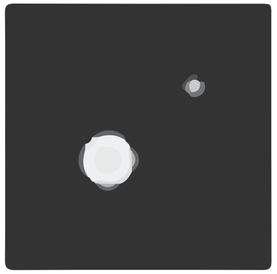
$$\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top \sim \text{N}(\mathbf{0}, \tau_2 I)$$



Sun .. Zhou, *Nature Methods* (2020)

Kernels capture point spread functions

actual objects



two sources

are convoluted

observed image

(blurred)

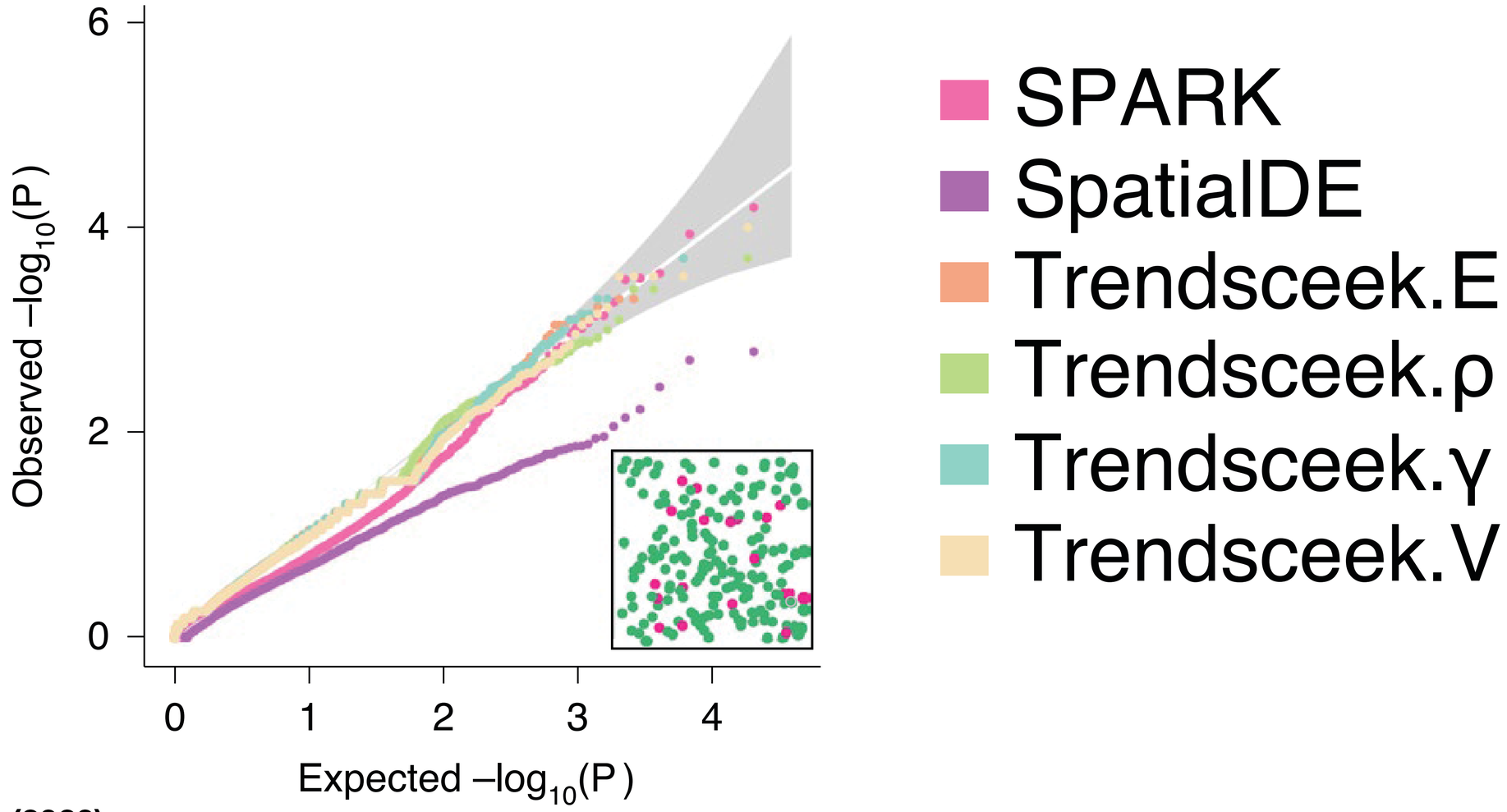
model PSF by a kernel

point spread function

(the degree of blurring)

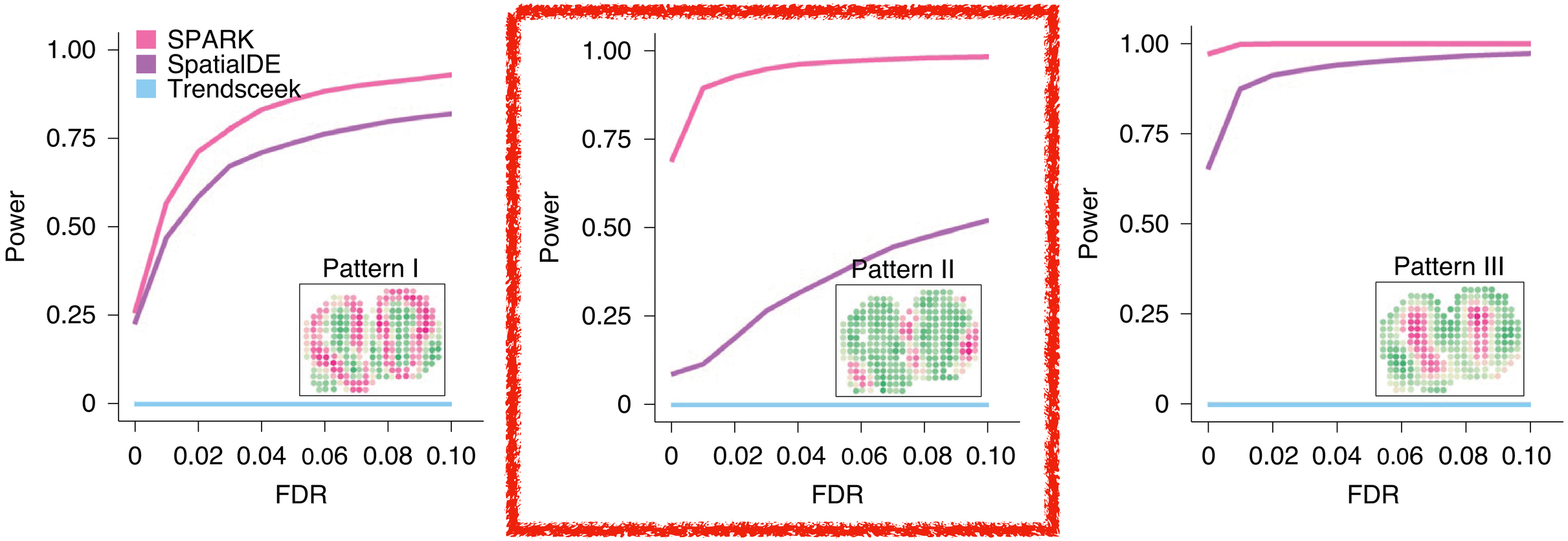
https://en.wikipedia.org/wiki/Point_spread_function

SPARK well-calibrates the null distribution



Sun .. Zhou, *Nature Methods* (2020)

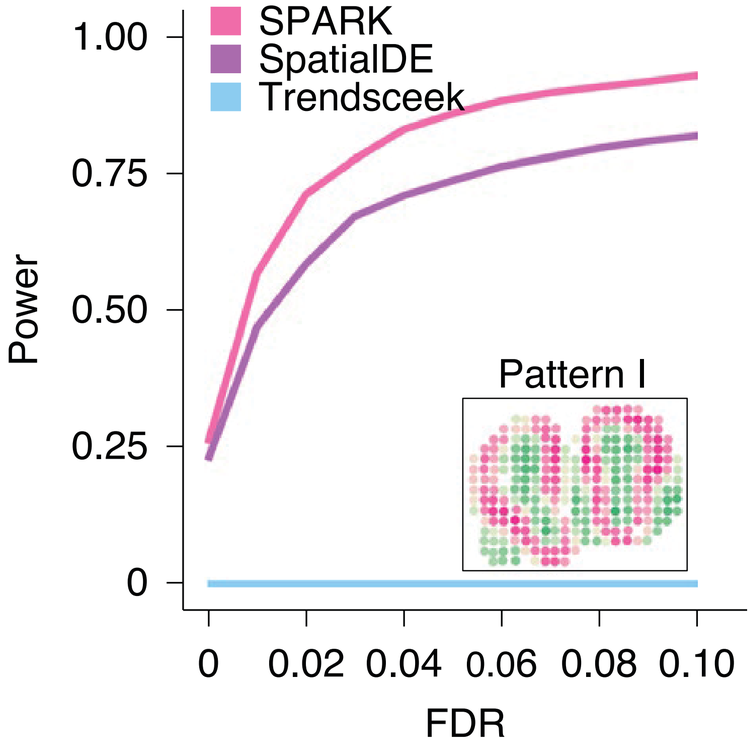
SPARK is more powerful than other DE

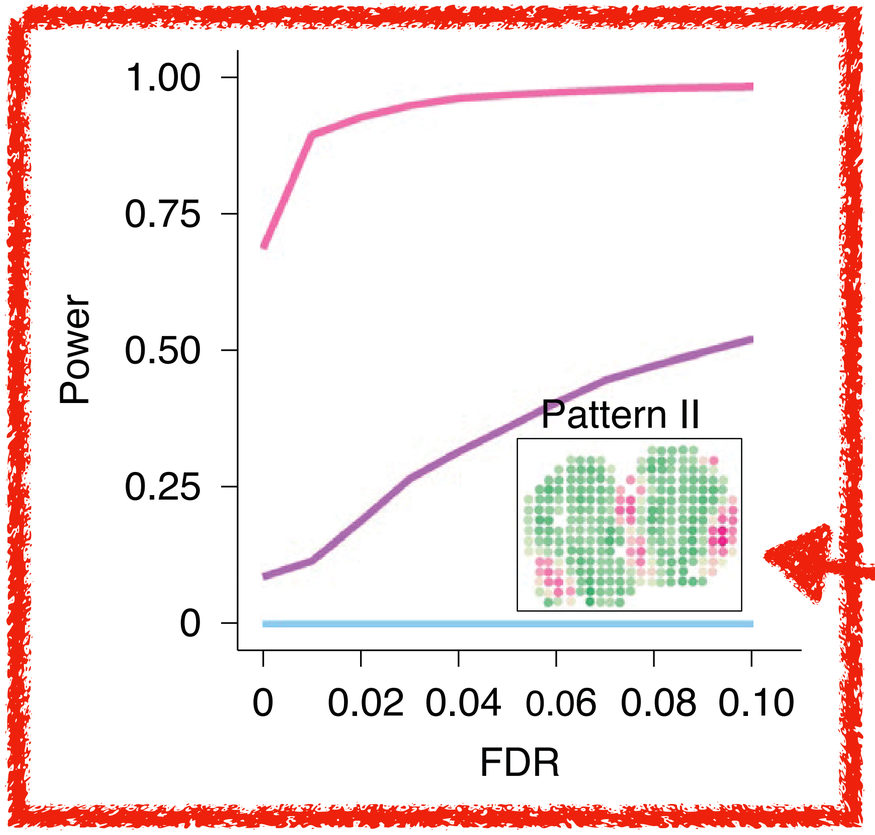


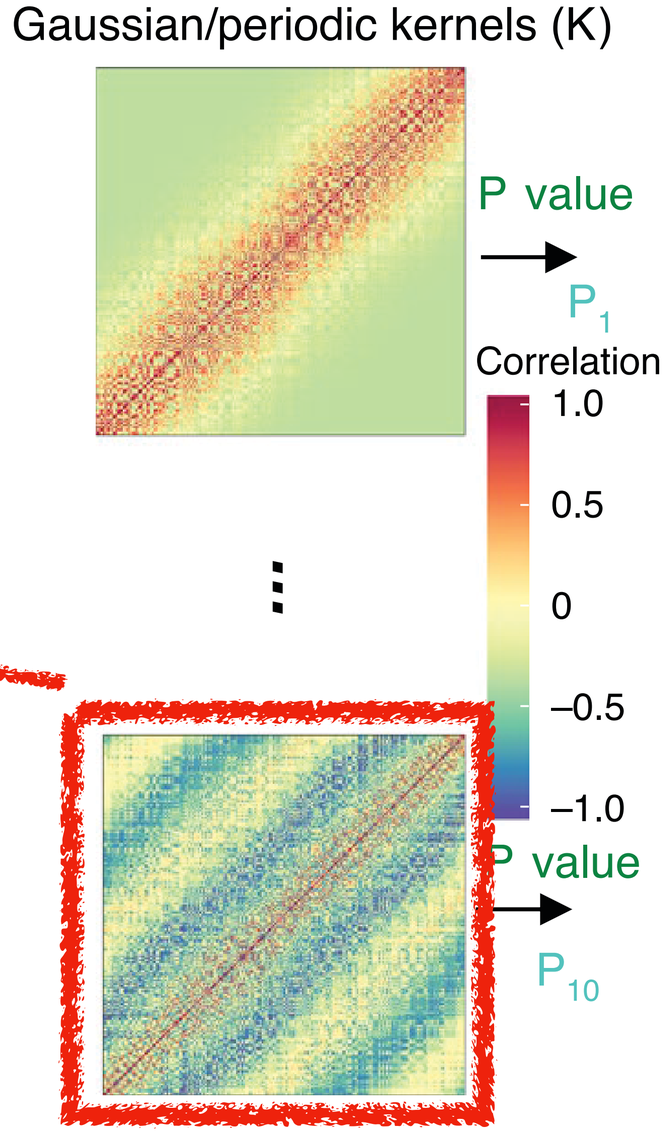
How?

Sun .. Zhou, *Nature Methods* (2020)

SPARK is more powerful than other DE



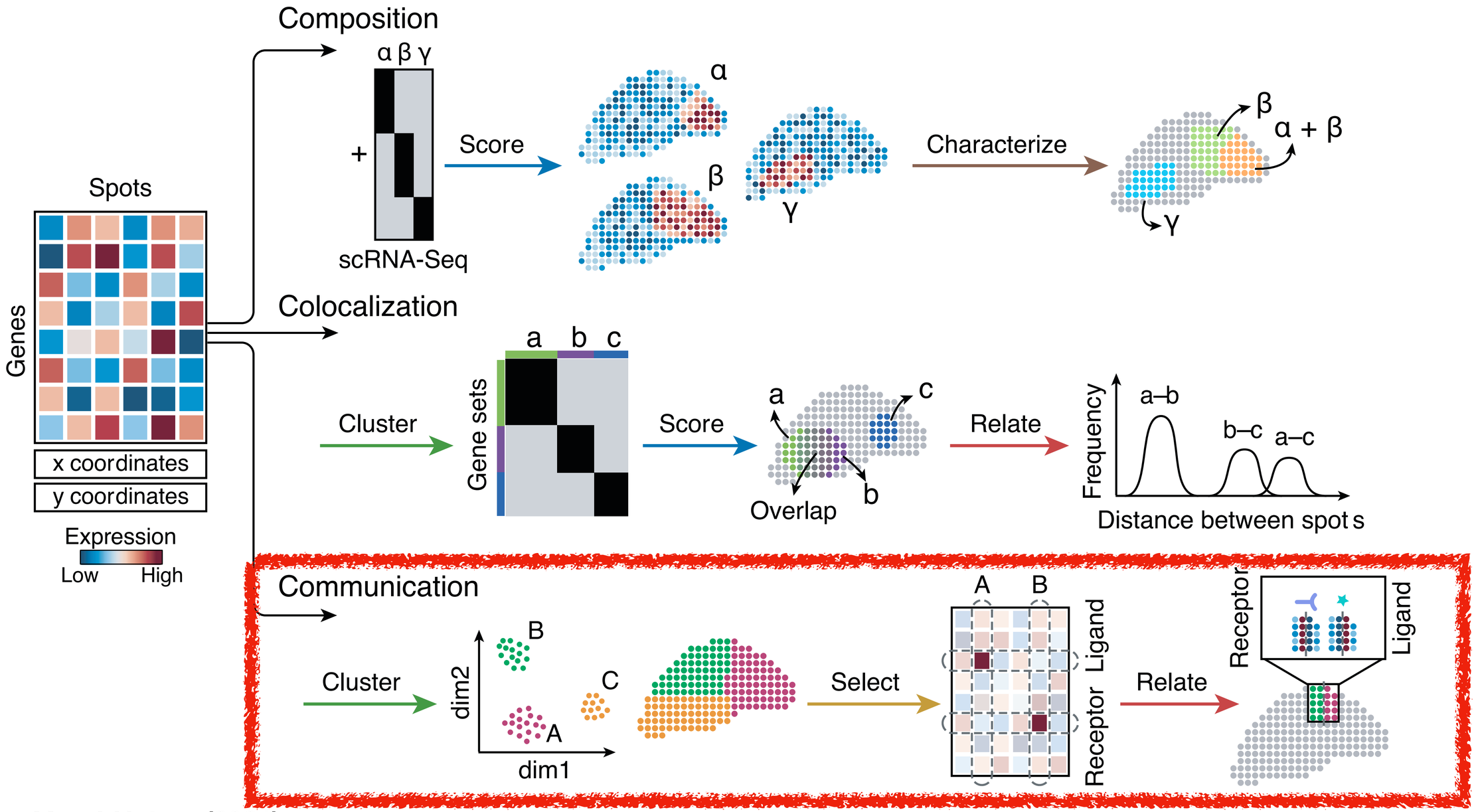




Sun .. Zhou, *Nature Methods* (2020)

Today’s lecture: Spatial Transcriptomics

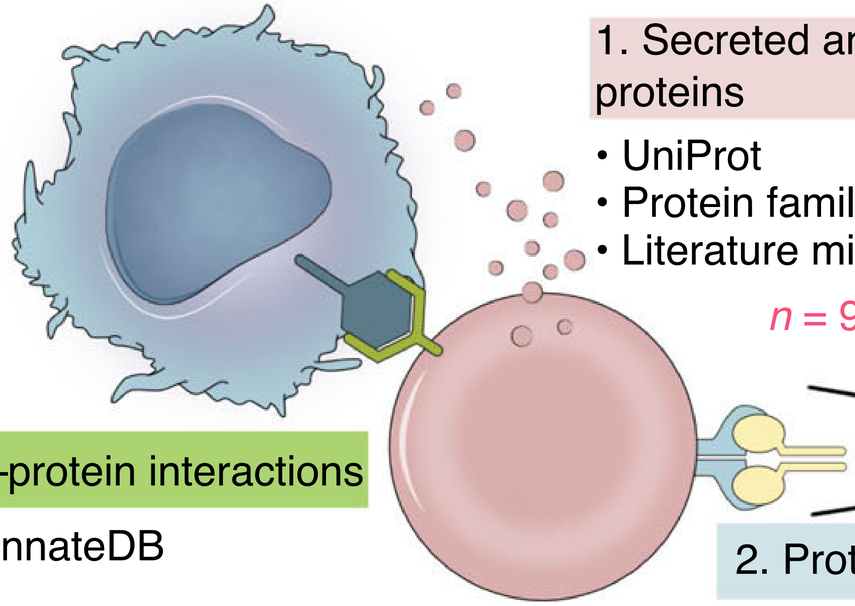
- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping



Rao .. Yanai *Nature* (2021)

Understanding how different cell types interact with each other

CellPhoneDB



1. Secreted and membrane proteins

- UniProt
- Protein families
- Literature mining

n = 978

3. Protein–protein interactions

- IMEx & InnateDB
- IUPHAR
- Literature mining *n* = 878

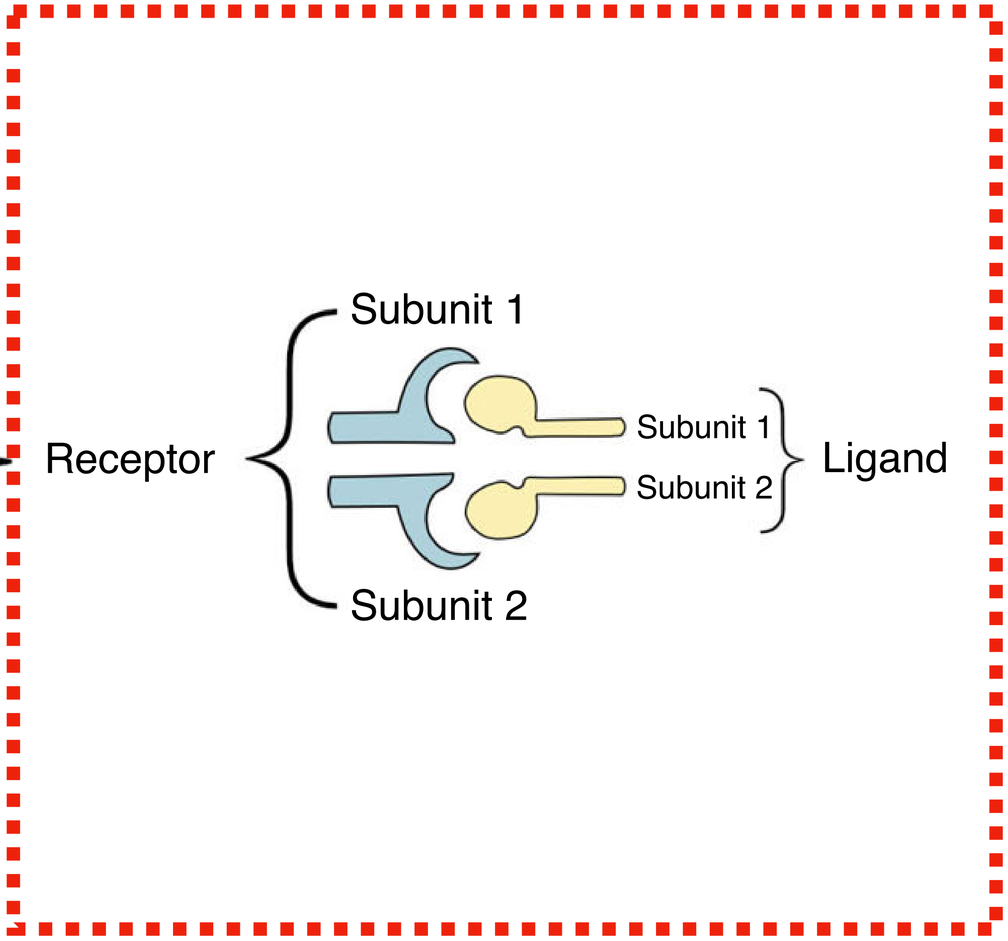
n = 1,396

n = 466 involving protein complexes

2. Protein complexes

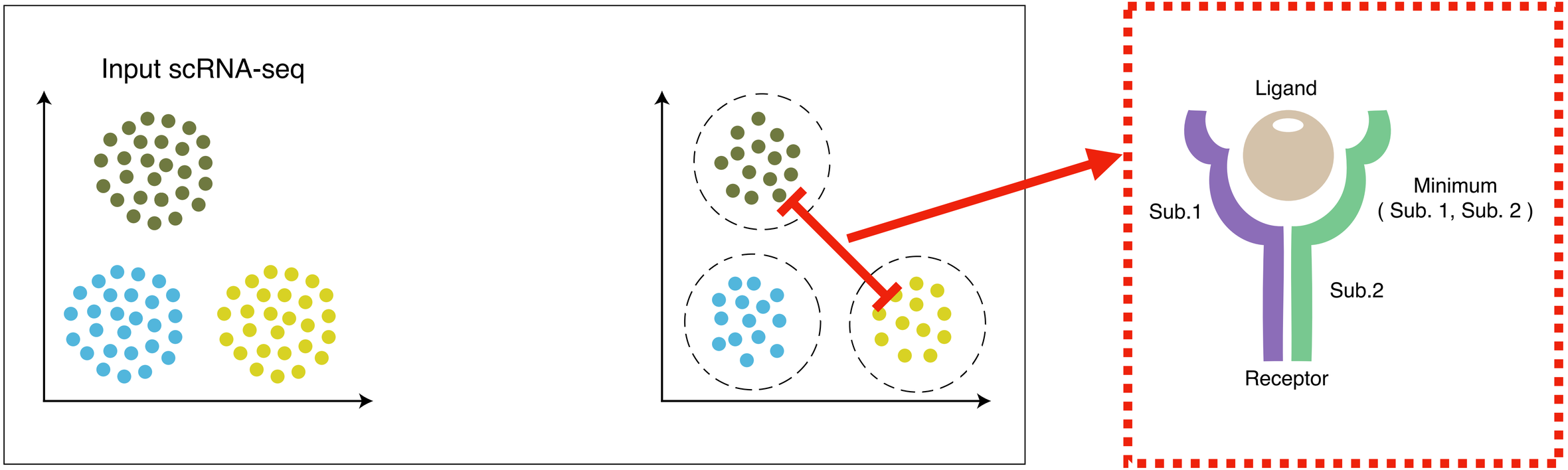
- Literature mining
- PDB

n = 112



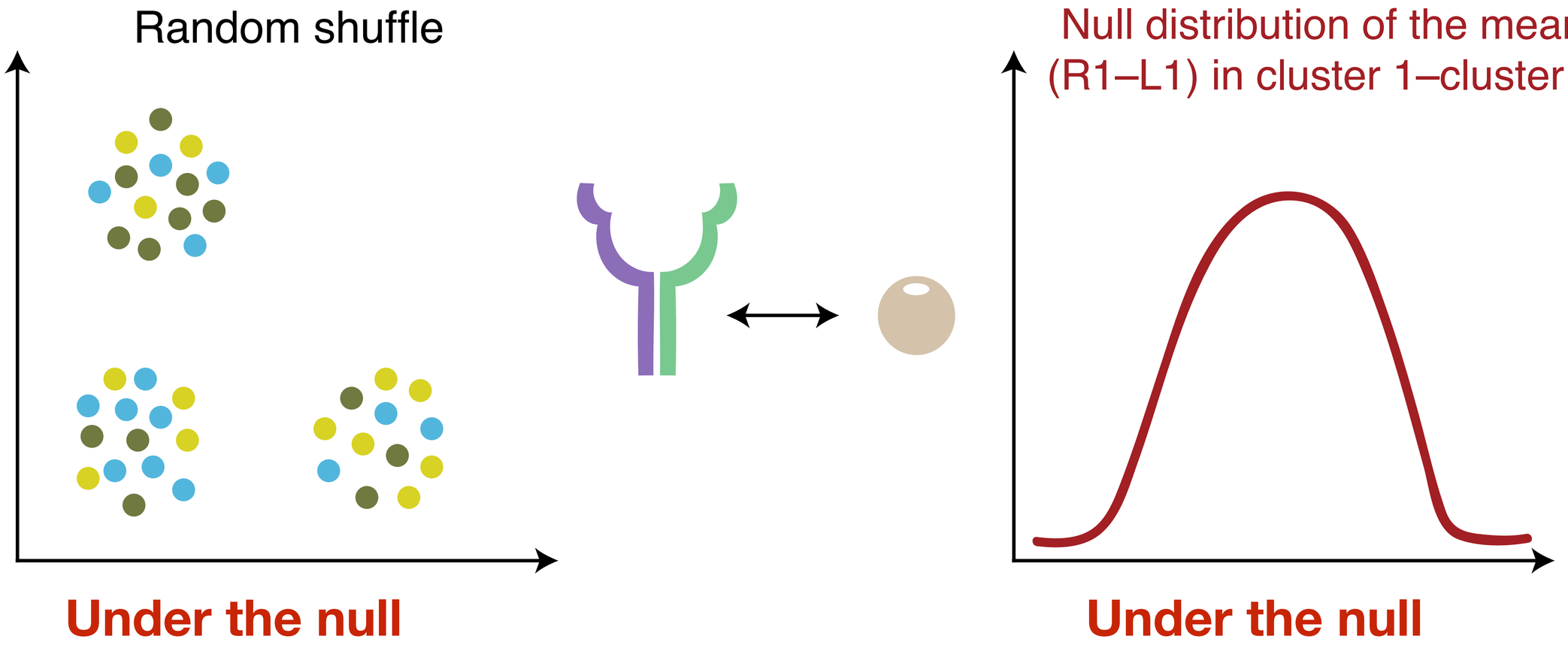
Efremova *et al.* (2020) *Nature Protocols*

CellPhoneDB: How ligand and receptor proteins co-expressed between two adjacent cell types



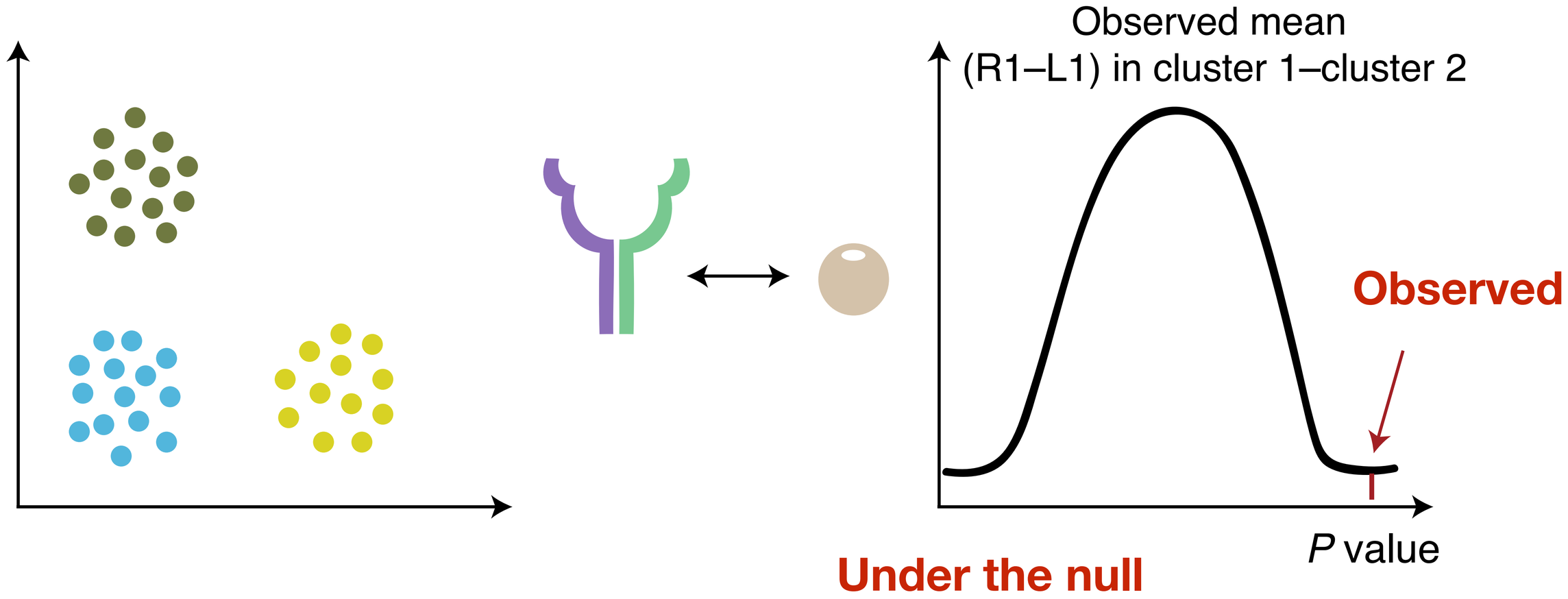
Efremova *et al.* (2020) *Nature Protocols*

How do we know co-expression patterns are statistically significant?



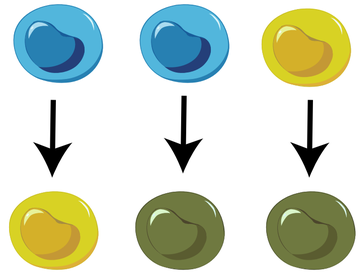
Efremova *et al.* (2020) *Nature Protocols*

How do we know co-expression patterns are statistically significant?



Efremova *et al.* (2020) *Nature Protocols*

Pairwise interaction p-values to construct inter-cellular networks

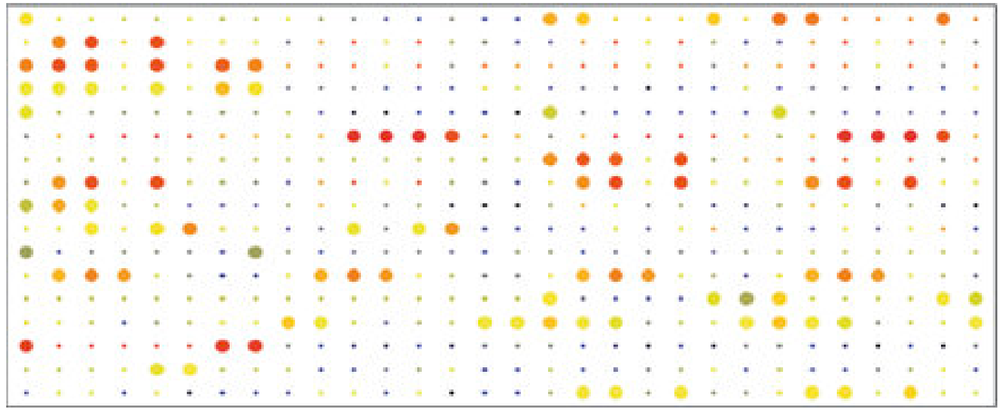


R1–L1	*P* value	*P* value	
R2–L2	*P* value		
R3–L3			

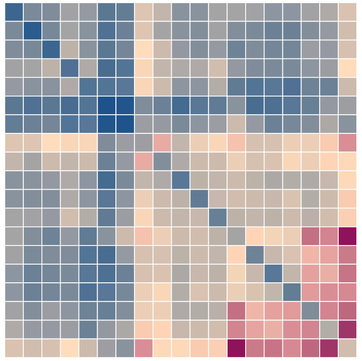
Ranking based on specificity

Receptor-Ligand pairs

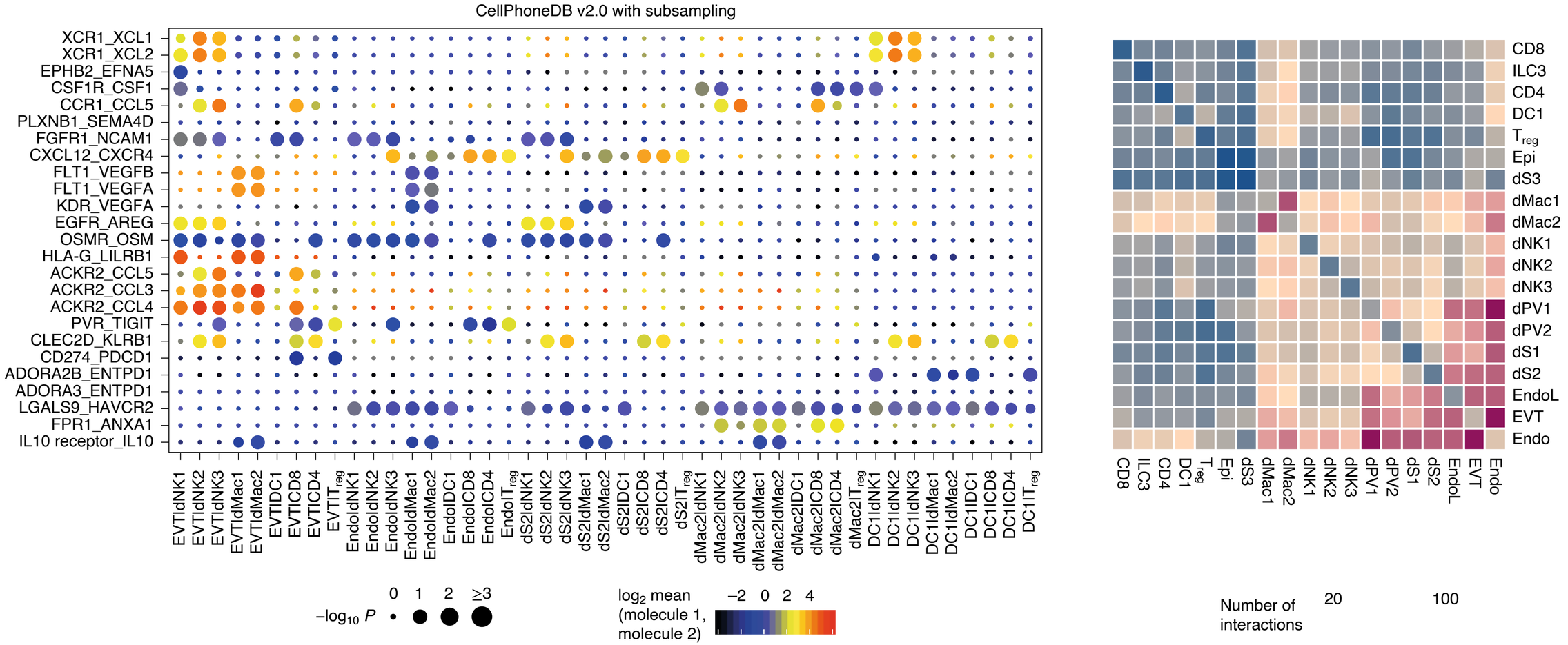
Cell-cell pairs



Cell type cell type similarity

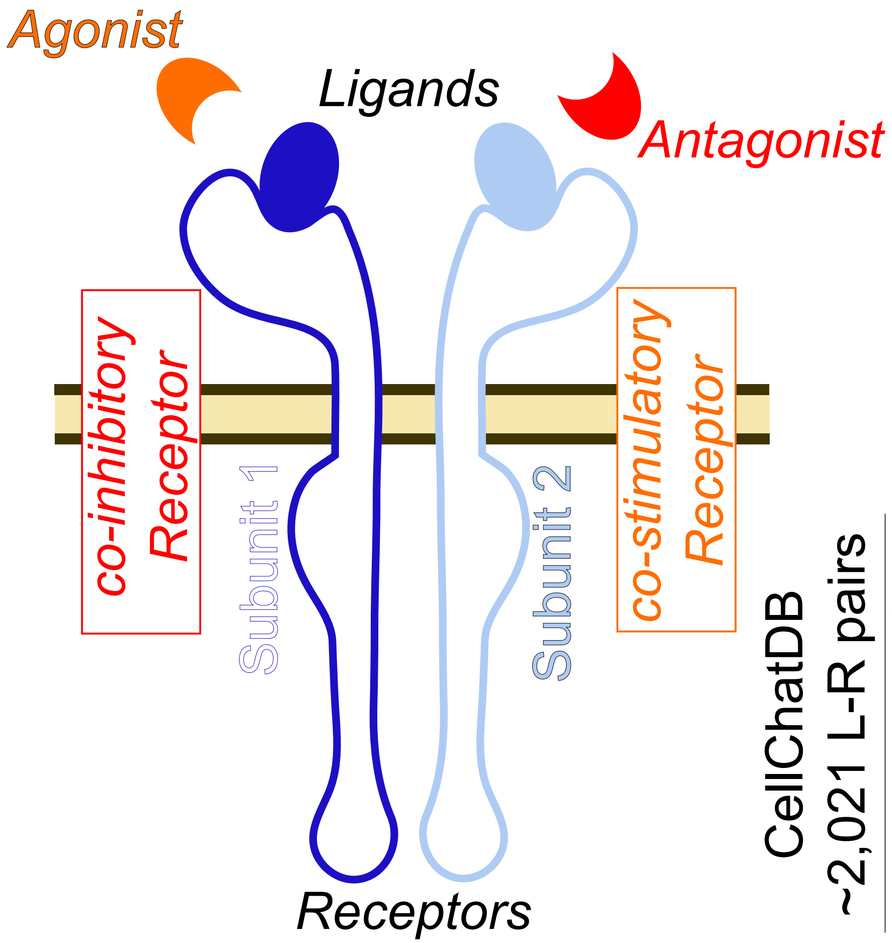


Interacting ligand-receptor pairs in a variety of intercellular contexts



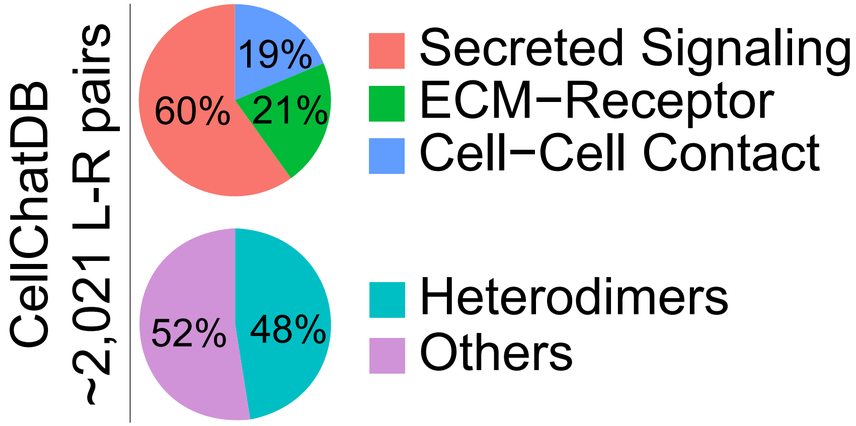
Efremova *et al.* (2020) *Nature Protocols*

CellChat: Inference of cell-cell communications



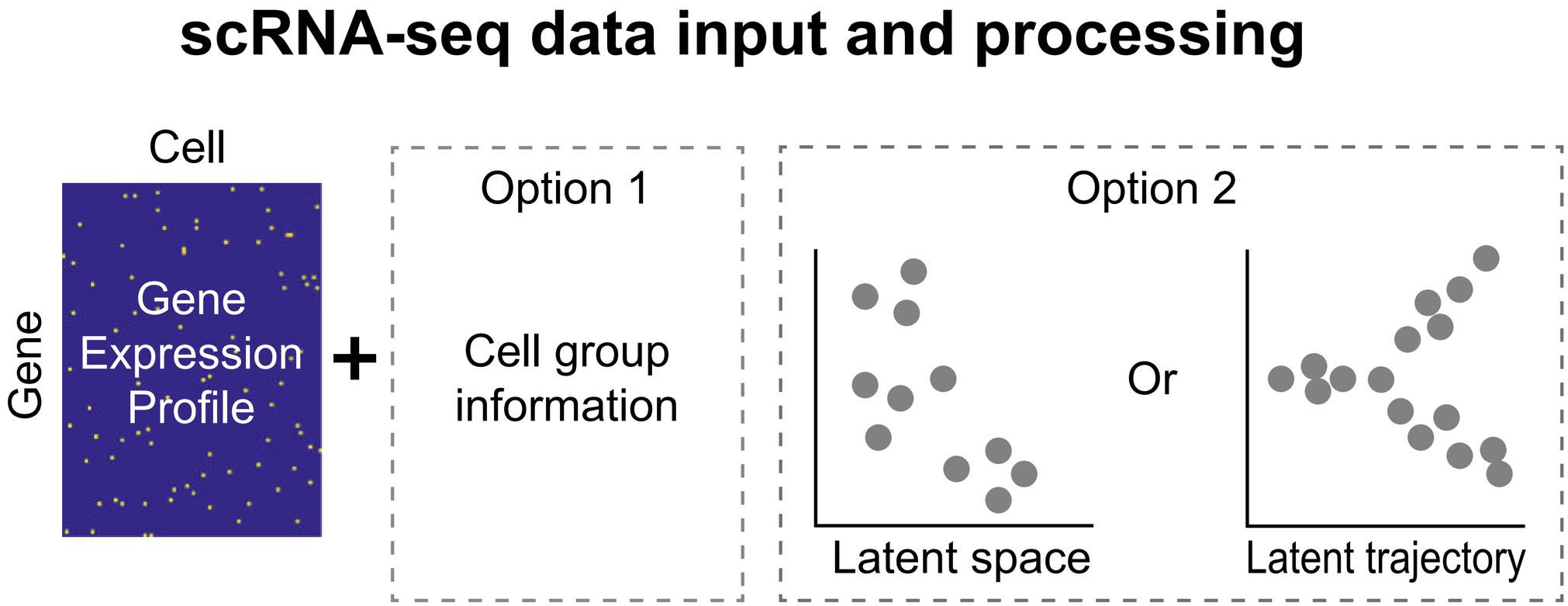
- KEGG base
- Primary literature

CellChat



Jin *et al.* (2021) *Nature Communications*

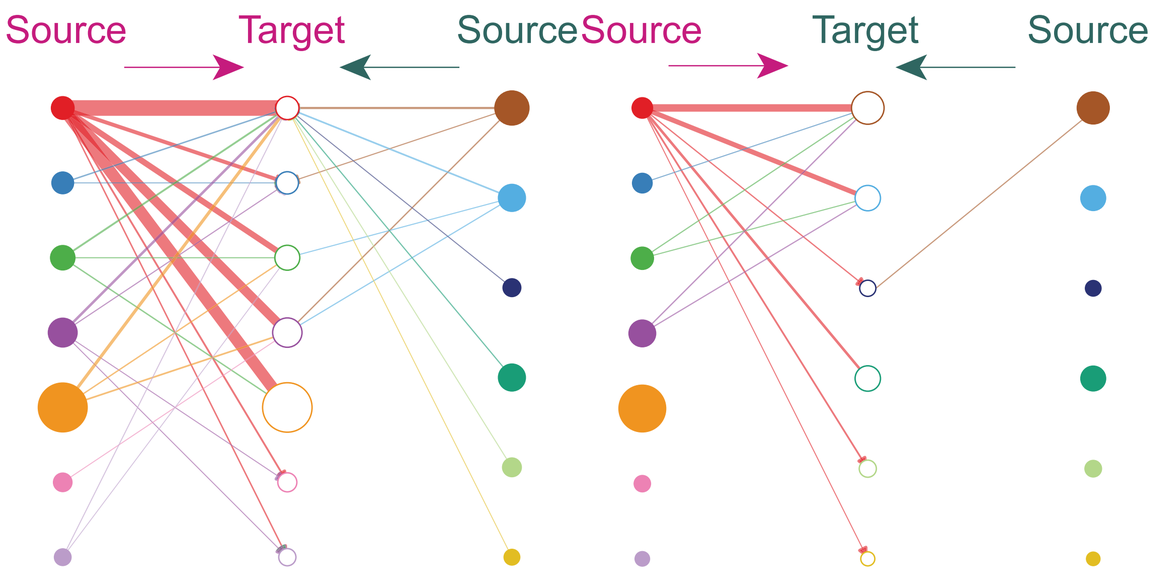
CellChat requires cell “group” information



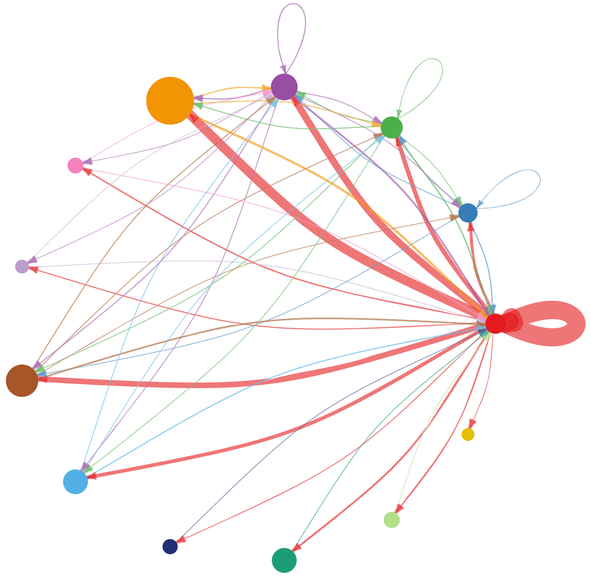
Jin *et al.* (2021) ***Nature Communications***

Visualization tools for Ligand → Receptor naturally-*directed* networks

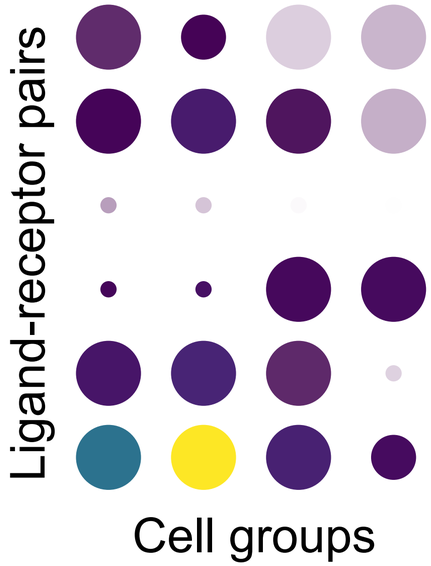
Hierarchy plot



Circle plot

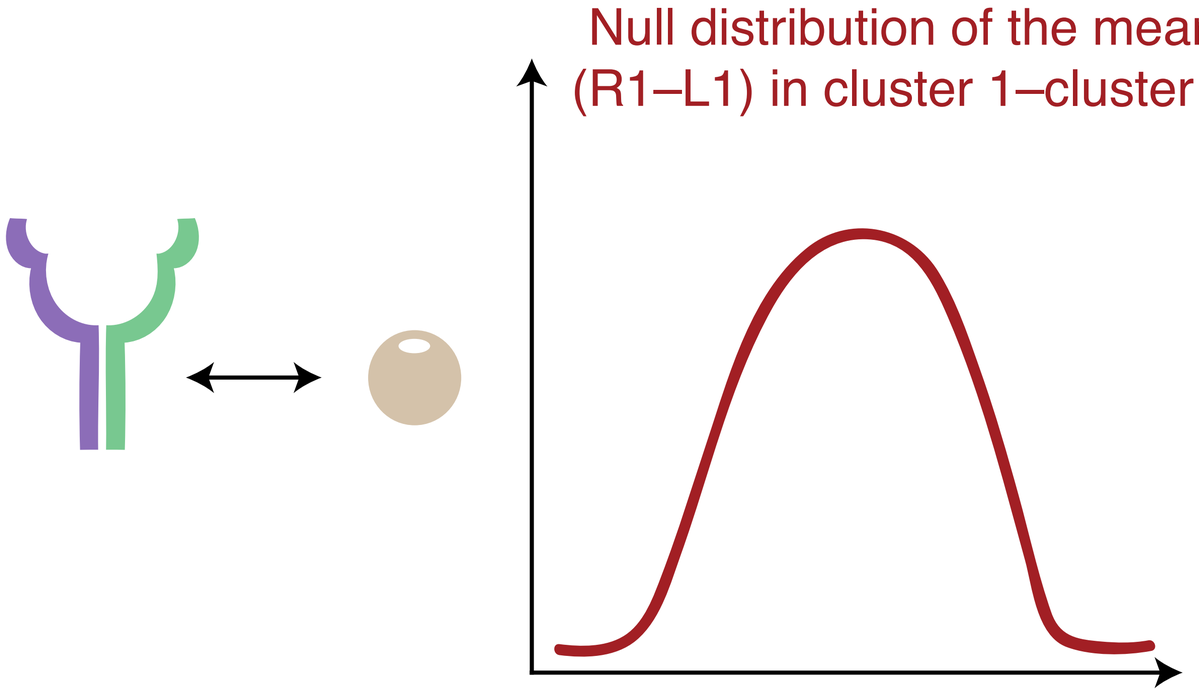


Bubble plot



Jin *et al.* (2021) *Nature Communications*

What statistics should we use?



$$\frac{1}{2}\left(\bar{X}_{lt} + \bar{Y}_{rs}\right)$$

$\bar{X}_{lt}$ **Average expression of a ligand gene/protein *(l)* in a cell type *t***

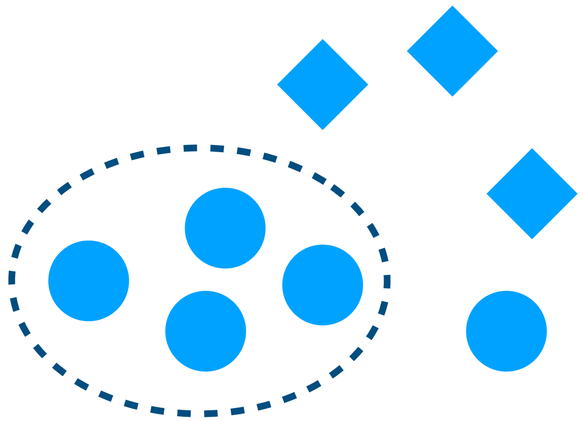
$\bar{Y}_{rs}$ **Average expression of a receptor gene/protein *(r)* in a cell type *s***

Efremova *et al.* (2020) *Nature Protocols*

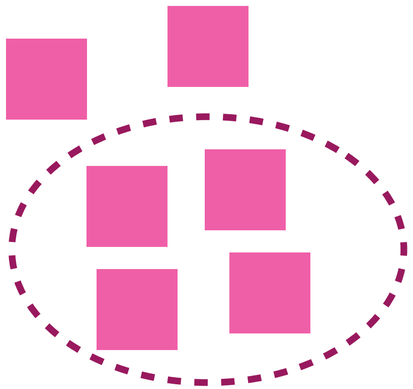
What statistics should we use?

$$P_{i,j}^{l,r} = \frac{X_{li}Y_{rj}}{X_iY_j + K_h} \times \overbrace{\left(1 + \frac{A_i}{K_h + A_i}\right)\left(1 + \frac{G_j}{K_h + Gj}\right)}^{\textbf{agonist}} \times \overbrace{\frac{K_h}{K_h + T_i}\frac{K_h}{K_h + N_j}}^{\textbf{antagonist}}$$

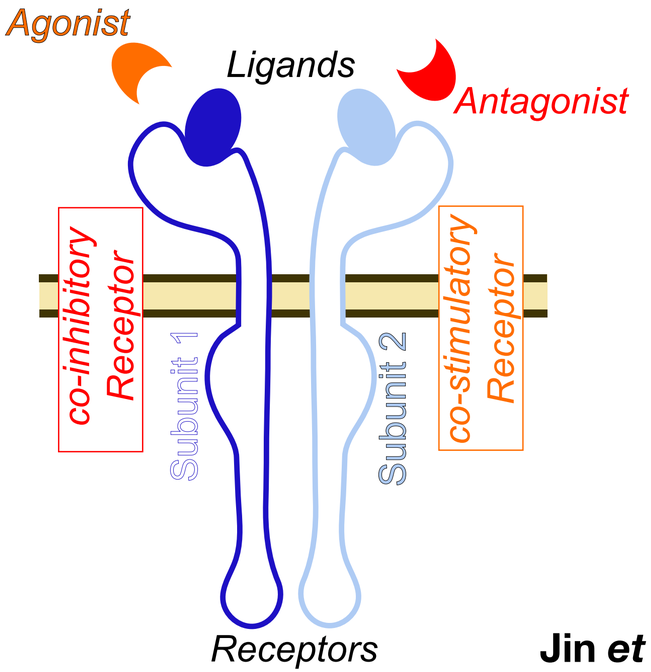
Ligand: l; Receptor r; Hill coefficient K



Cell group *i*



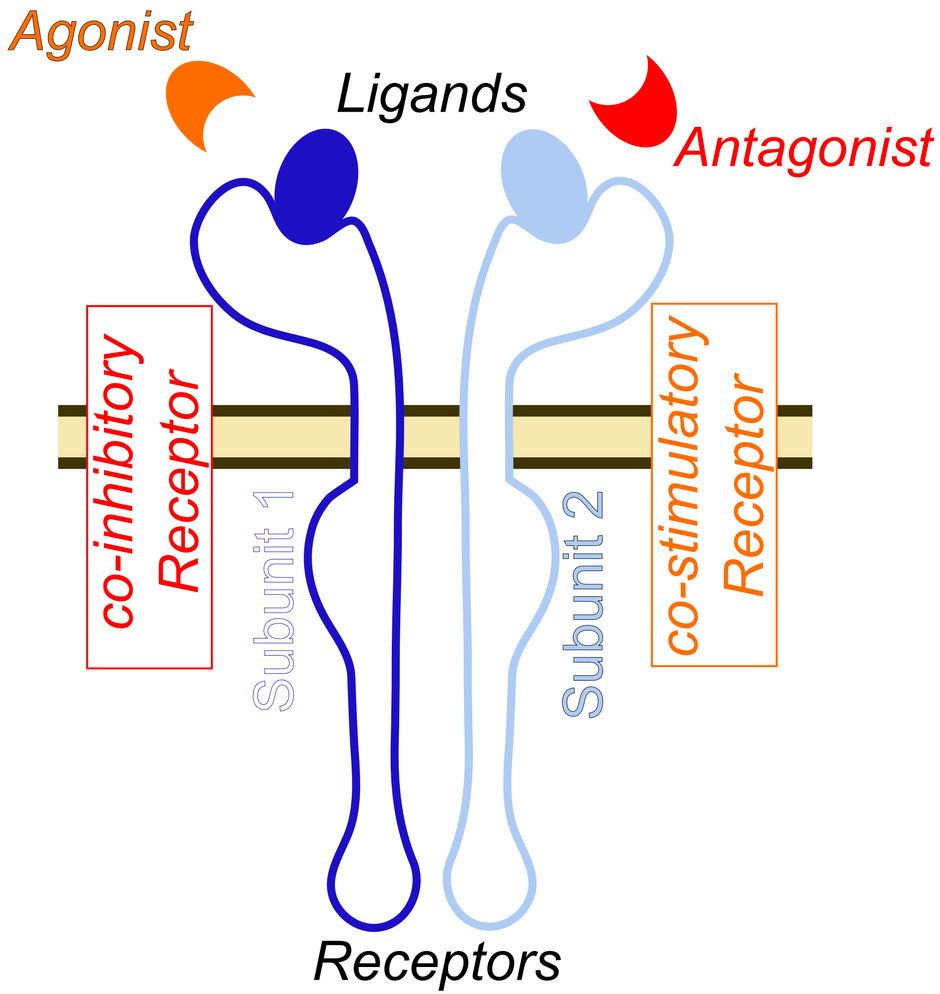
Cell group *j*



Jin *et al.* (2021) *Nature Communications*

Pros and Cons of Ligand-Receptor co-occurrence

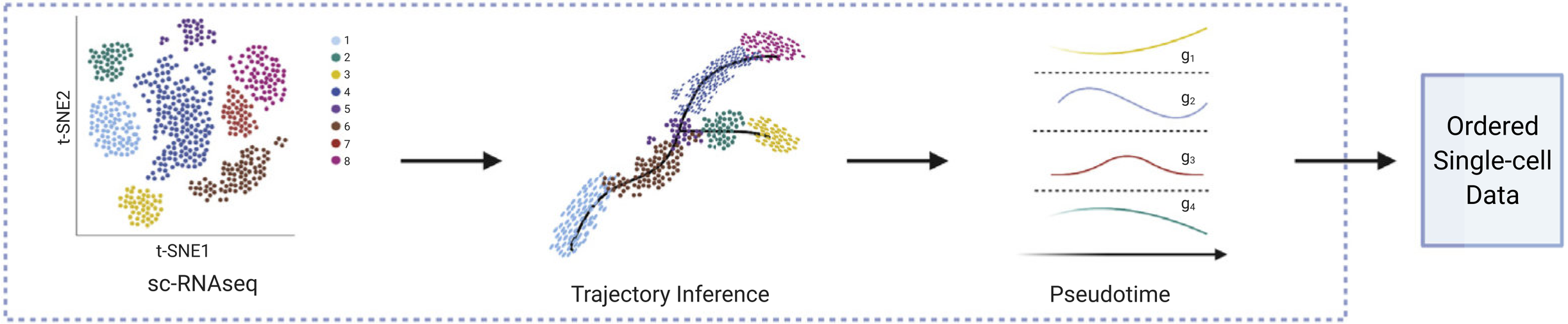
- **Pros**:
  - Intuitive interpretation
  - Biochemical mass-action law
  - Directionality
- **Cons**:
  - Only ligand-receptor pairs
  - Not a causal effect
  - Confounded by cell grouping



Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

If cells are temporally ordered, can we use that information in gene-gene network reconstruction?



Granger causality: Learning causal relationships in time-series data

Autoregressive process

$$\mathbf{x}(t) = \sum_{\tau} A_{\tau}\mathbf{x}(t-\tau) + \epsilon(t)$$

Econometrica, Vol. 37, No. 3 (July, 1969)

INVESTIGATING CAUSAL RELATIONS BY ECONOMETRIC MODELS AND CROSS-SPECTRAL METHODS

BY C. W. J. GRANGER

There occurs on some occasions a difficulty in deciding the direction of causality between two related variables and also whether or not feedback is occurring. Testable definitions of causality and feedback are proposed and illustrated by use of simple two-variable models. The important problem of apparent instantaneous causality is discussed and it is suggested that the problem often arises due to slowness in recording information or because a sufficiently wide class of possible causal variables has not been used. It can be shown that the cross spectrum between two variables can be decomposed into two parts, each relating to a single causal arm of a feedback situation. Measures of causal lag and causal strength can then be constructed. A generalisation of this result with the partial cross spectrum is suggested.

DEFINITION 1: *Causality*. If $\sigma^2(X|U) < \sigma^2(X|\overline{U - Y})$, we say that Y is causing X, denoted by $Y_t \Rightarrow X_t$. We say that Y_t is causing X_t if we are better able to predict X_t using all available information than if the information apart from Y_t had been used.

DEFINITION 2: *Feedback*. If

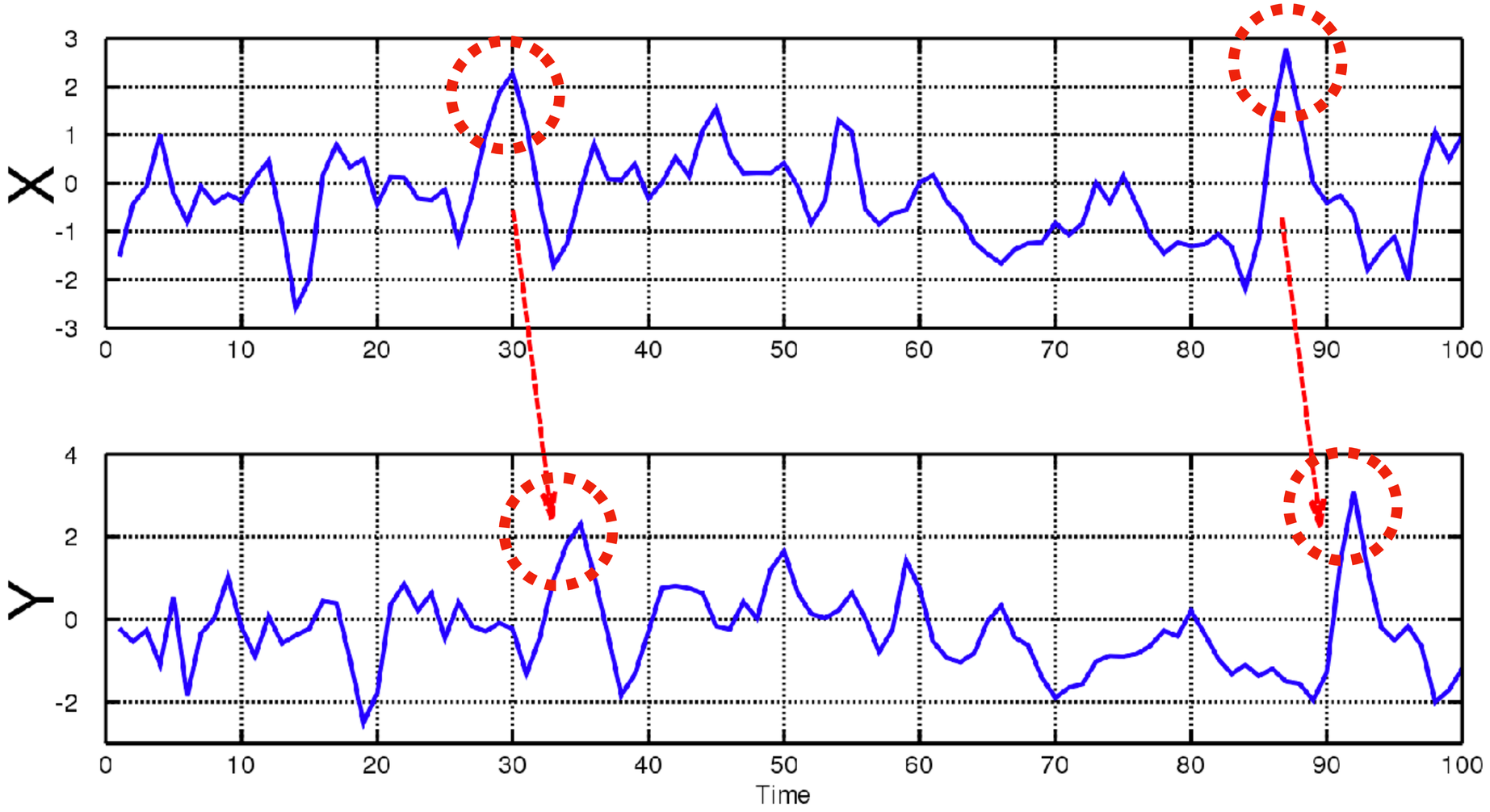
$$\sigma^2(X|\overline{U}) < \sigma^2(X|\overline{U - Y}),$$
$$\sigma^2(Y|\overline{U}) < \sigma^2(Y|\overline{U - X}),$$

we say that feedback is occurring, which is denoted $Y_t \Leftrightarrow X_t$, i.e., feedback is said to occur when X_t is causing Y_t and also Y_t is causing X_t.

(...)

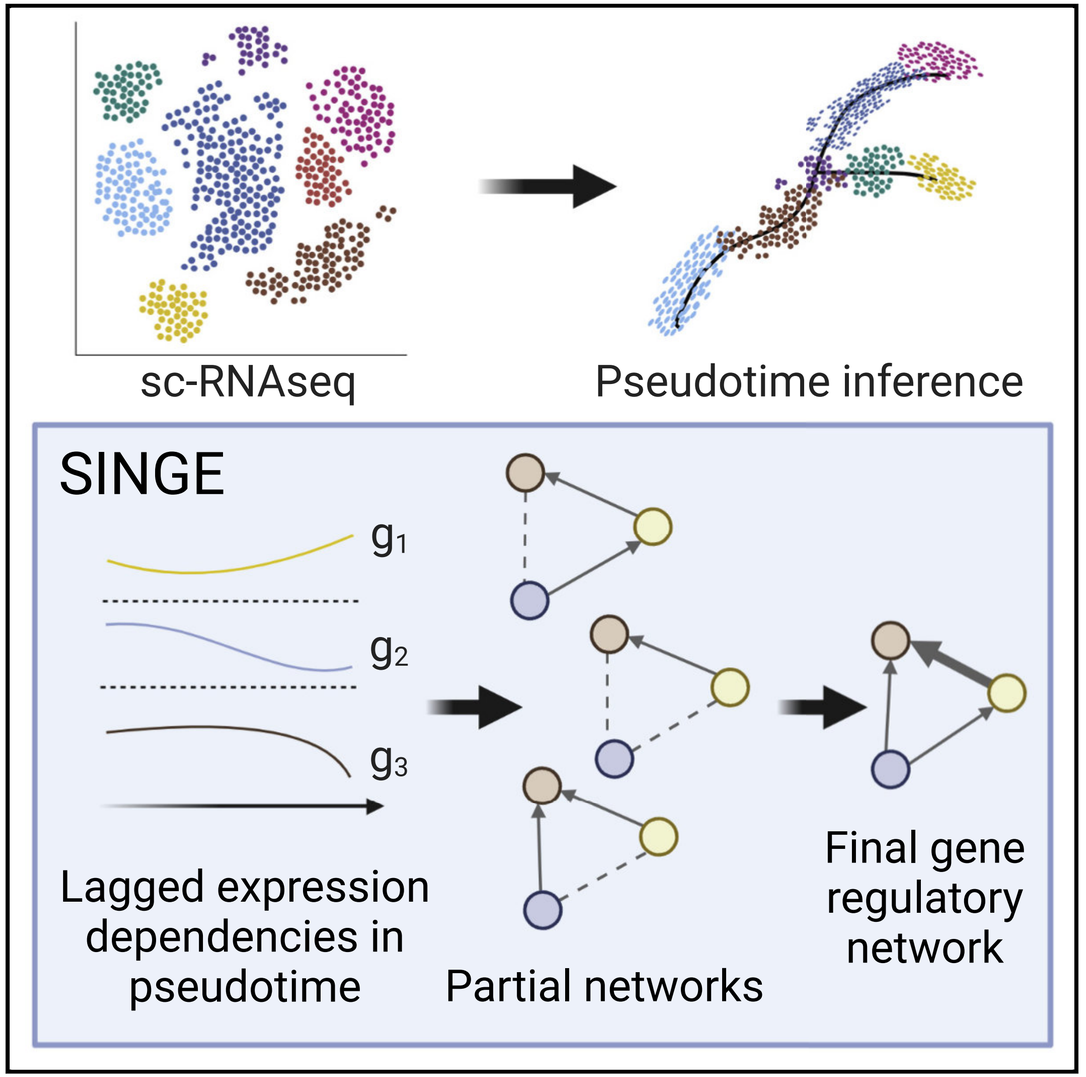
Sir Clive Granger

Granger Causality



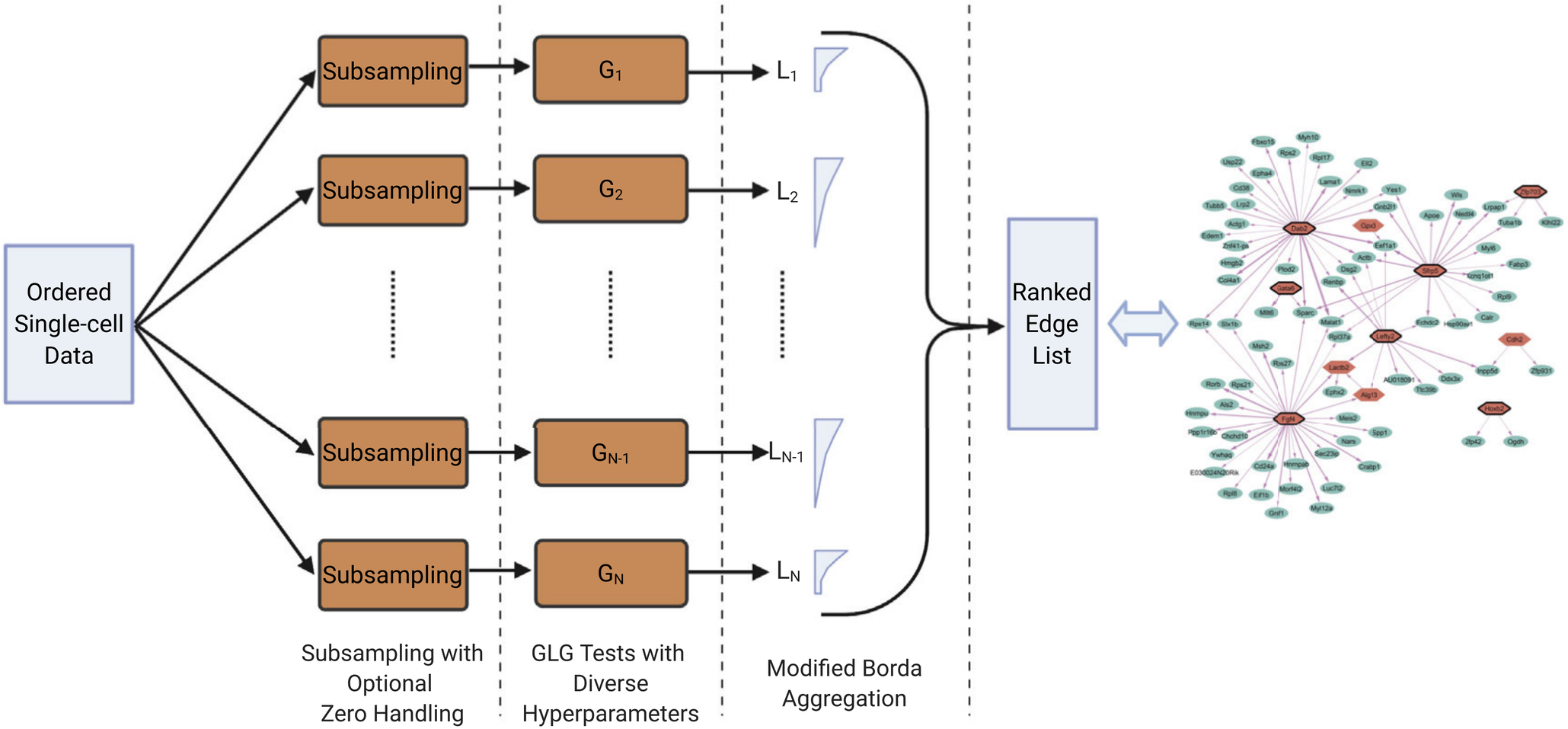
https://en.wikipedia.org/wiki/Granger_causality

SINGE: Single-cell Granger Causality Ensemble



Deshpande … Gitter (2022) *Cell Reports*

GC regression + hyper-parameter ensemble



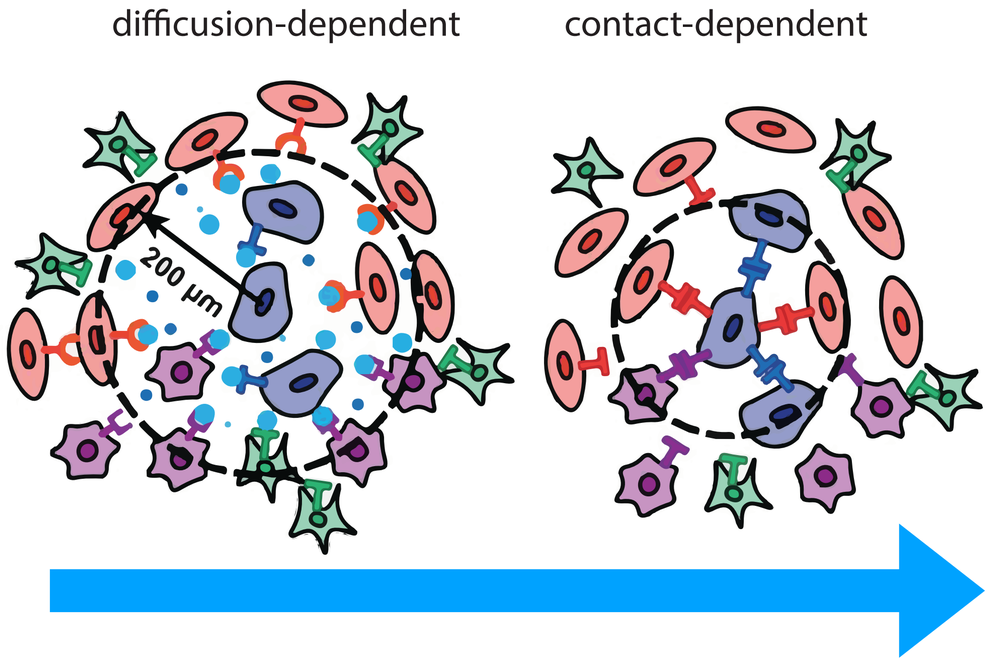
Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

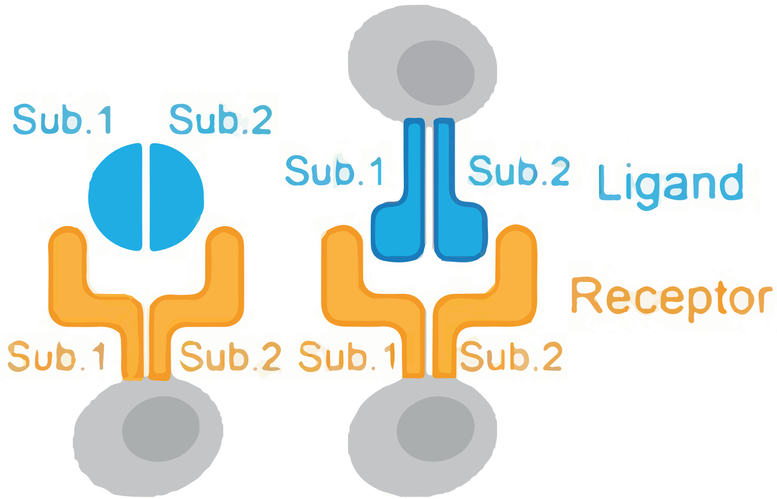
CytoSignal improves LR analysis by imputation

transcripts

spatial position

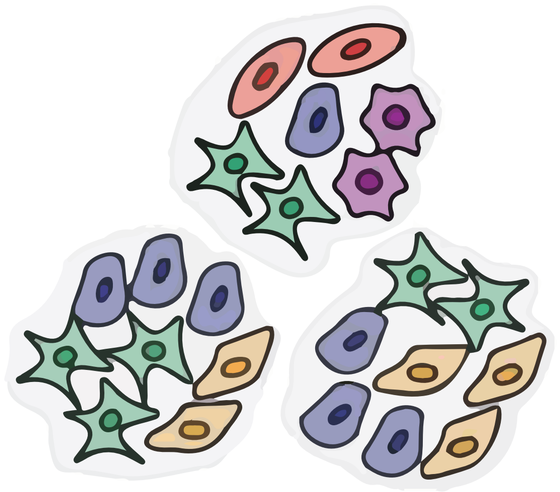


$$S = (L1 + L2) * (R1 + R2)$$



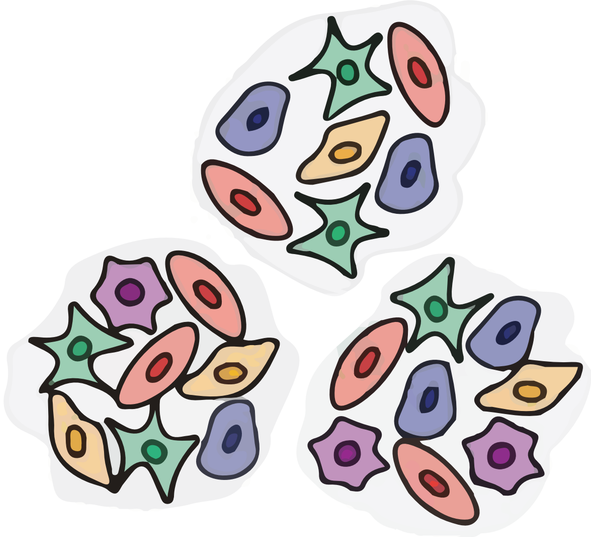
Liu .. Welch, *bioRxiv* (2024)

CytoSignal calibrate the null by permuation

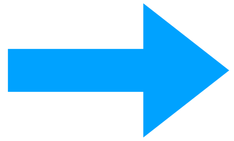


Observed data

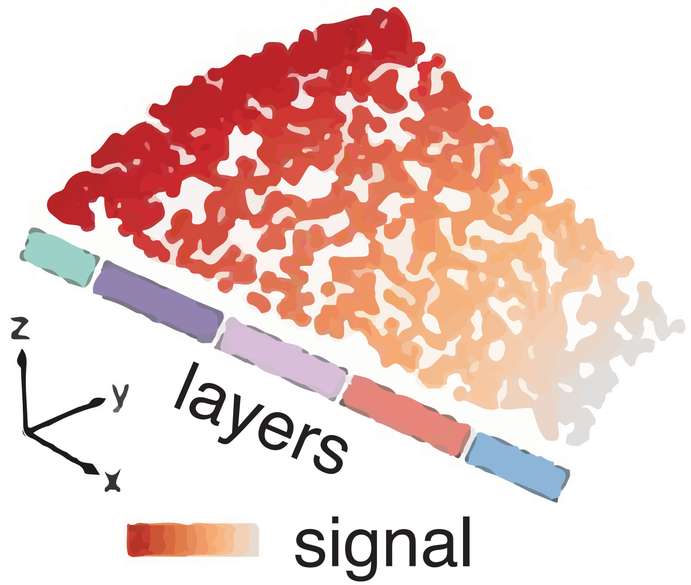
vs.



Cell permutation



LR signal flow



Liu .. Welch, *bioRxiv* (2024)

How do we define cell-cell communications?

1. Contact between cells through ligand and receptor proteins
2. RNA velocity from one cell to the neighbouring cells
3. Spatial-transcriptomic contact map to confirm the interacting cells are in proximity

CytoSignal

https://github.com/welch-lab/CytoSignal

BioRxiv paper:

https://www.biorxiv.org/content/10.1101/2024.03.08.584153v1.full

Toy example data SCP2170

```r
## The RDS file will be loaded into a ready-to-use object
dge <- readRDS("../data/cytosignal/SCP2170_annotated_dgCMatrix.rds")

## The cluster annotation need to be presented as a factor object
cluster <- read.csv("../data/cytosignal/SCP2170_cluster.csv")
cluster <- factor(cluster$cell_type)
names(cluster) <- colnames(dge)

## The spatial coordinates need to be presented as a matrix object
spatial <- as.matrix(read.csv("../data/cytosignal/SCP2170_spatial.csv", row.names = 1))
## Please make sure that the dimension names are lower case "x" and "y"
colnames(spatial) <- c("x", "y")
```

Create `CytoSignal` object

```
library(cytosignal)
library(Matrix)
cs <- createCytoSignal(raw.data = dge,
                       cells.loc = spatial,
                       clusters = cluster)

cs <- addIntrDB(cs, g_to_u, db.diff, db.cont, inter.index)

cs <- removeLowQuality(cs, counts.thresh = 300)

cs <- changeUniprot(cs)

cs
```

```
## An object of class CytoSignal
##  with raw data of dimension:  4623 cells and  15343 genes.
```

Set nearest neighbour search parameters

```
cs <- inferEpsParams(cs, scale.factor = 0.73)
cs <- findNN(cs)
```

```
dim(cs@imputation$DT@imp.data)
```

```
## [1] 0 0
```

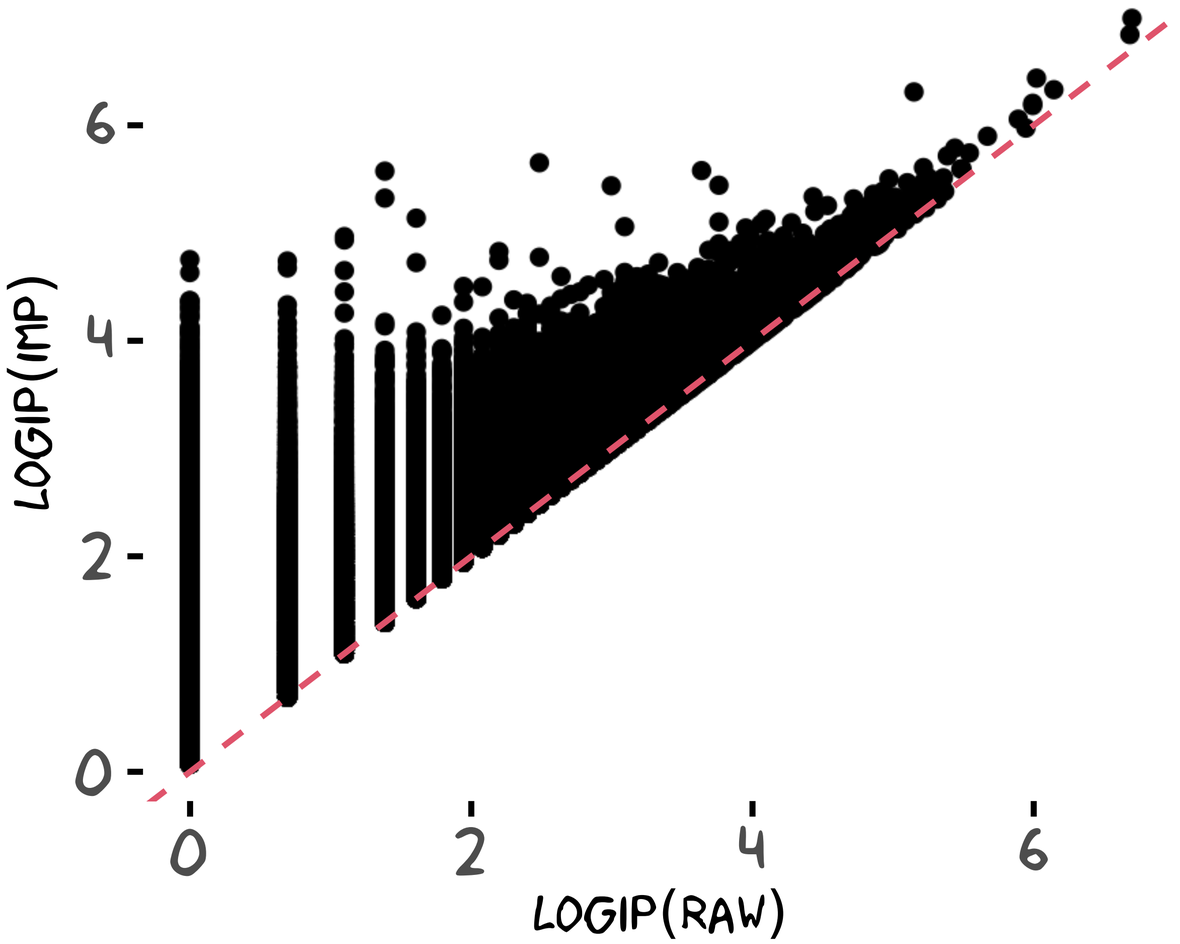
Impute missing ligand receptor expressions

```
cs <- imputeLR(cs)
dim(cs@imputation$Raw@imp.data)
```

```
## [1]  811 4623
```

```
dim(cs@imputation$DT@imp.data)
```

```
## [1]  811 4623
```



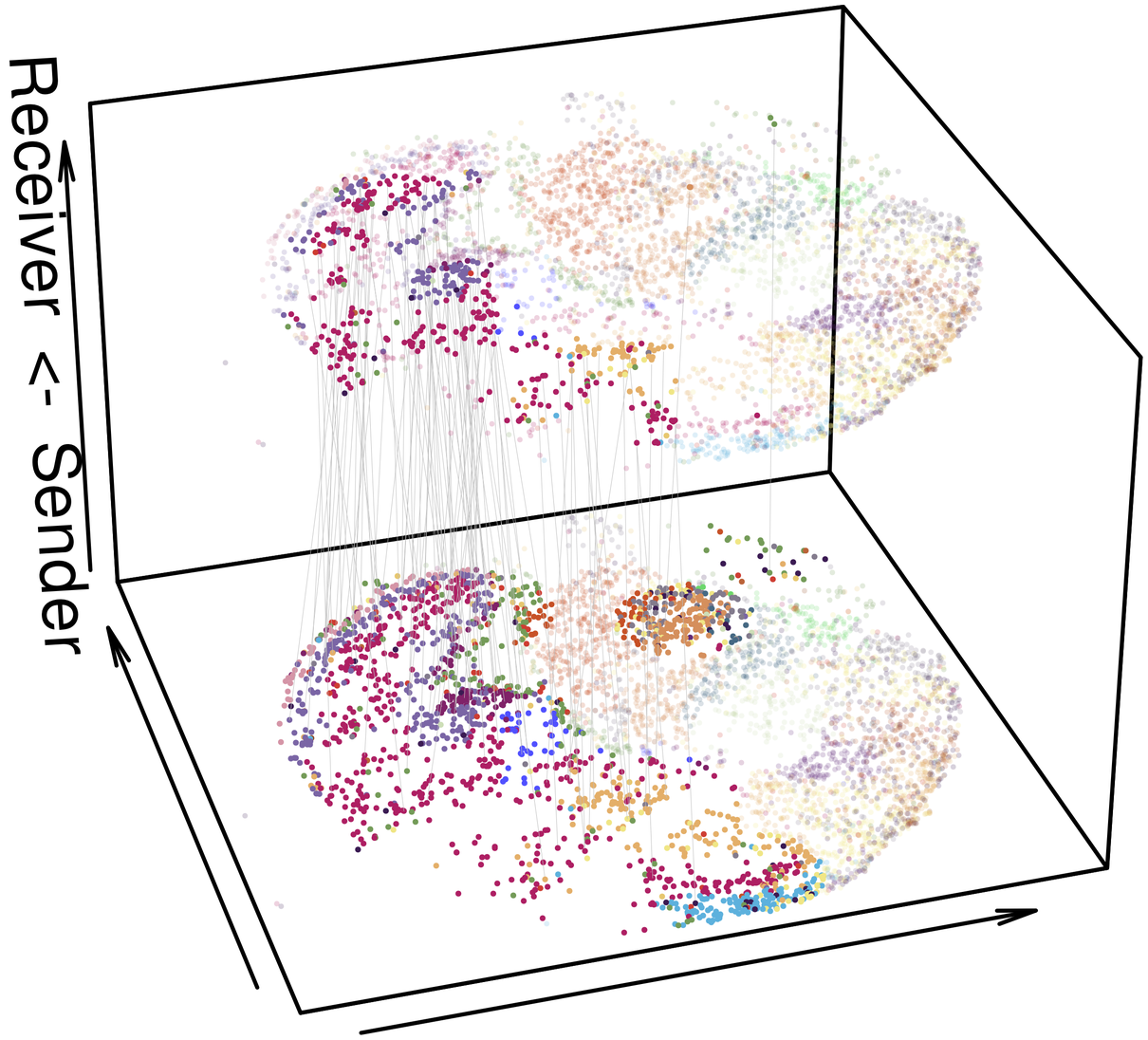
Find significant ligand-receptor by permutation test

```
cs <- inferIntrScore(cs)
cs <- inferSignif(cs, p.value = 0.05, reads.thresh = 100, sig.thresh = 100)
cs <- rankIntrSpatialVar(cs)
cs
```

```
allIntrs <- showIntr(cs, slot.use = "GauEps-Raw",
                     signif.use = "result.spx",
                     return.name = TRUE)
print(head(allIntrs))
```

```
## CPI-SS0659DBE72 CPI-SS04124F4E1 CPI-SS0008137B6 CPI-SS00F4DDF4B CPI-SS0E063192D
##   "EFNA4-EPHA4"     "EPO-EFNB2"   "EFNA4-EPHA5"     "PTN-PTPRS"    "NRG2-ERBB4"
## CPI-SS0C694CB44
##   "VEGFA-NMDE2"
```

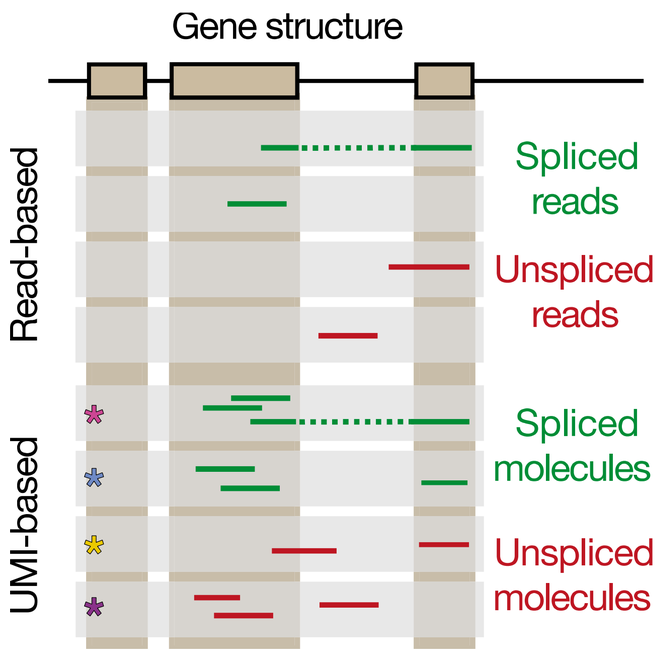
Edge-EFNA4-EPHA4

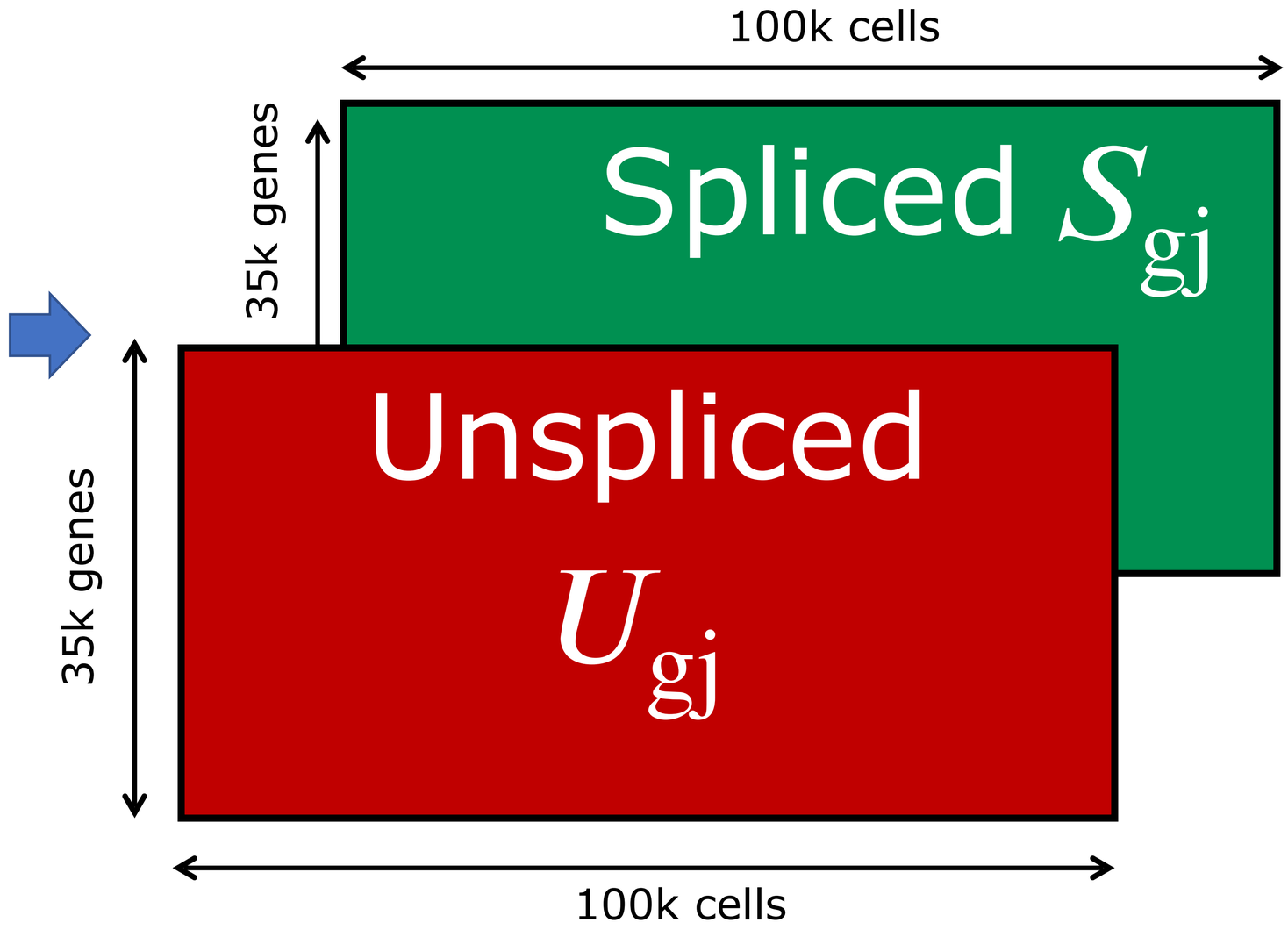


Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

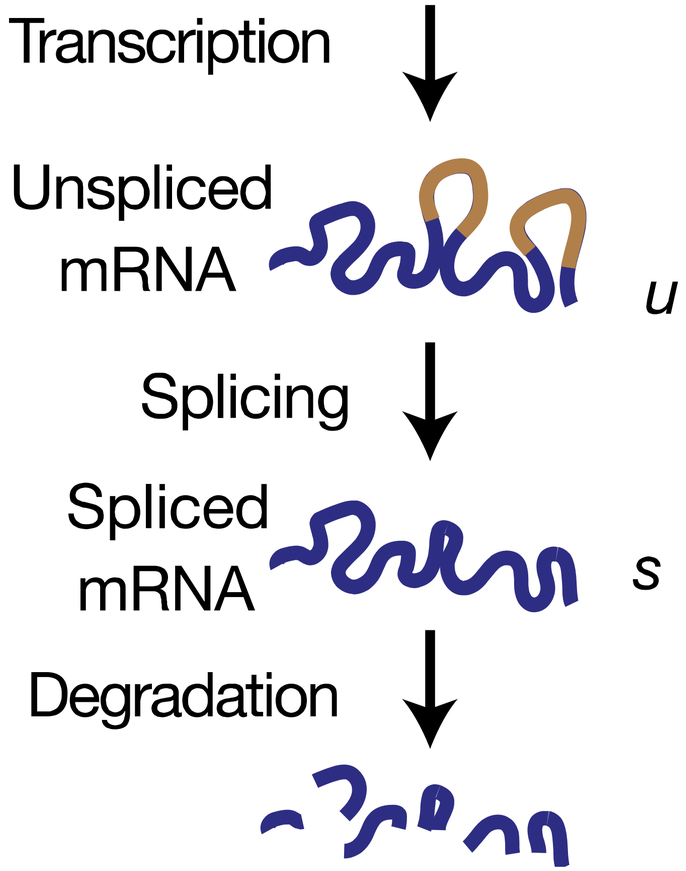
What are the data for RNA velocity problem?





La Manno *et al*. Nature (2018)

Modelling splicing dynamics by mass action law



$$\frac{dU}{dt} = \alpha - \beta U(t)$$

splicing rate

transcription initiation rate

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate mRNA degradation

Phase diagram of RNA velocity

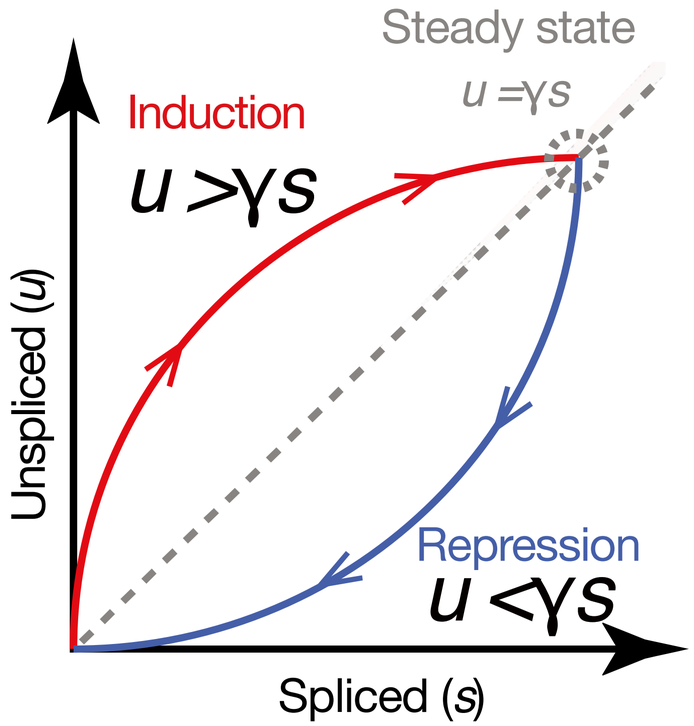
$$\frac{dU}{dt} = \alpha - \beta U(t)$$

splicing rate

transcription initiation rate

$$\frac{dS}{dt} = \beta U(t) - \gamma S(t)$$

splicing rate mRNA degradation



La Manno *et al*. Nature (2018)

Can we incorporate RNA velocity information?

```
## The RDS file will be loaded into a ready-to-use object
dge <- readRDS("../data/cytosignal/SCP815_dgCMatrix.rds")

## The cluster annotation need to be presented as a factor object
cluster <- read.csv("../data/cytosignal/SCP815_cluster.csv", row.names = 1)
cluster <- factor(setNames(unlist(cluster, use.names = F), rownames(cluster)))

## The spatial coordinates need to be presented as a matrix object
spatial <- as.matrix(read.csv("../data/cytosignal/SCP815_cells_loc.csv", row.names = 1))
colnames(spatial) <- c("x", "y")

vcs <- createCytoSignal(raw.data = dge,
                        cells.loc = spatial,
                        clusters = cluster)

vcs <- addIntrDB(vcs, g_to_u, db.diff, db.cont, inter.index)
vcs <- removeLowQuality(vcs, counts.thresh = 300)
vcs <- changeUniprot(vcs)
```

Add the spliced and unspliced counts

The spliced:

```
velo.s <- readRDS("../data/cytosignal/SCP815_190921_19_velo_spliced_matrix.rds")
dim(velo.s)
```

```
## [1]  234 2560
```

The unspliced:

```
velo.u <- readRDS("../data/cytosignal/SCP815_190921_19_velo_unspliced_matrix.rds")
dim(velo.u)
```

```
## [1]  234 2560
```

Add the spliced and unspliced count data:

```
vcs <- addVelo(vcs, velo.s = velo.s, velo.u = velo.u)
```

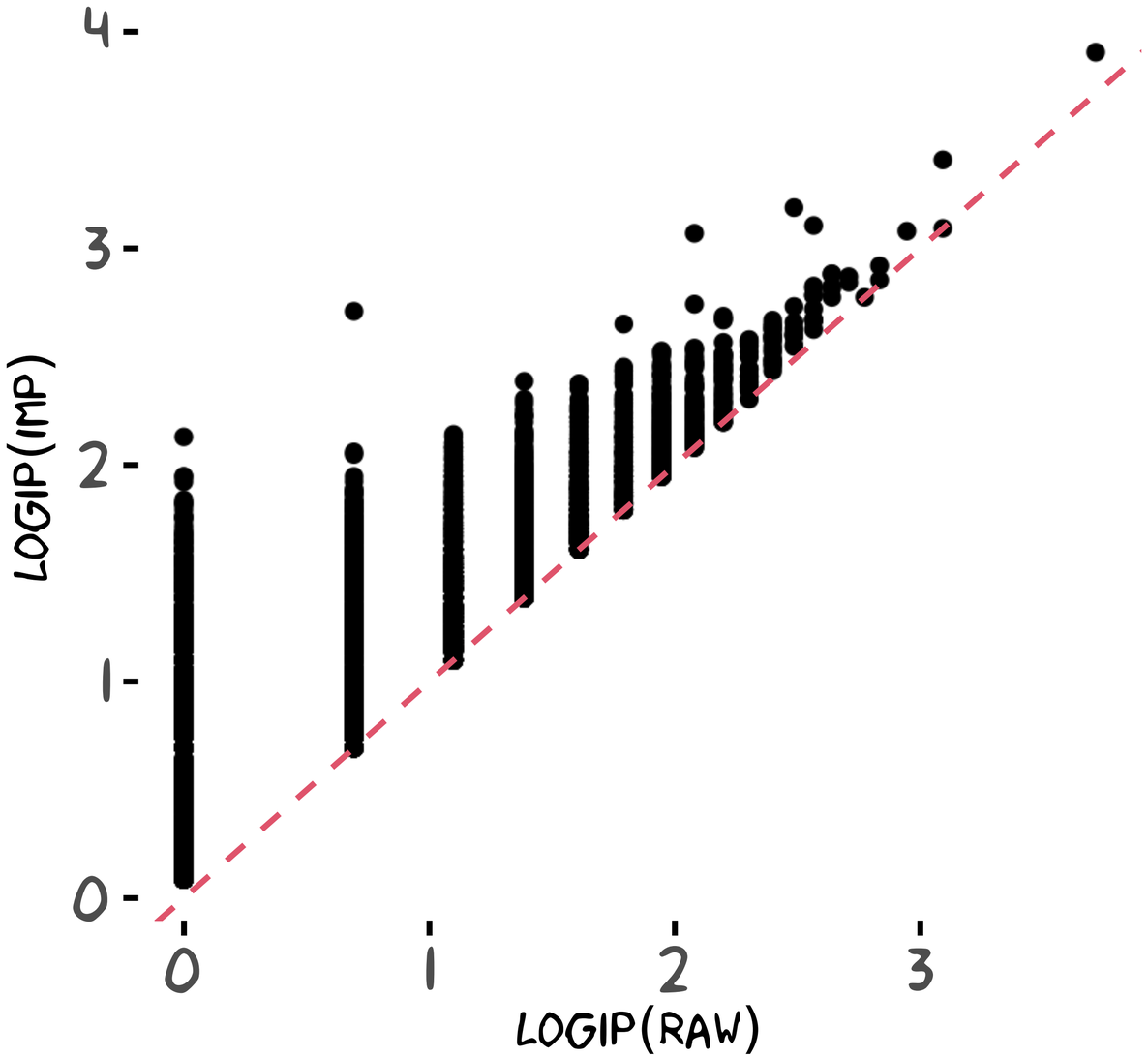
```
## Number of genes in the database: 234
## Number of genes in the database: 234
```

Estimate RNA velocity between ligand and receptor - 1

```
vcs <- inferEpsParams(vcs, scale.factor = 0.73)

## Find nearest neighbours to impute
vcs <- findNN(vcs)

## Impute ligand and receptor
vcs <- imputeLR(vcs)

## Impute velocity LR
vcs <- imputeVeloLR(vcs)
```

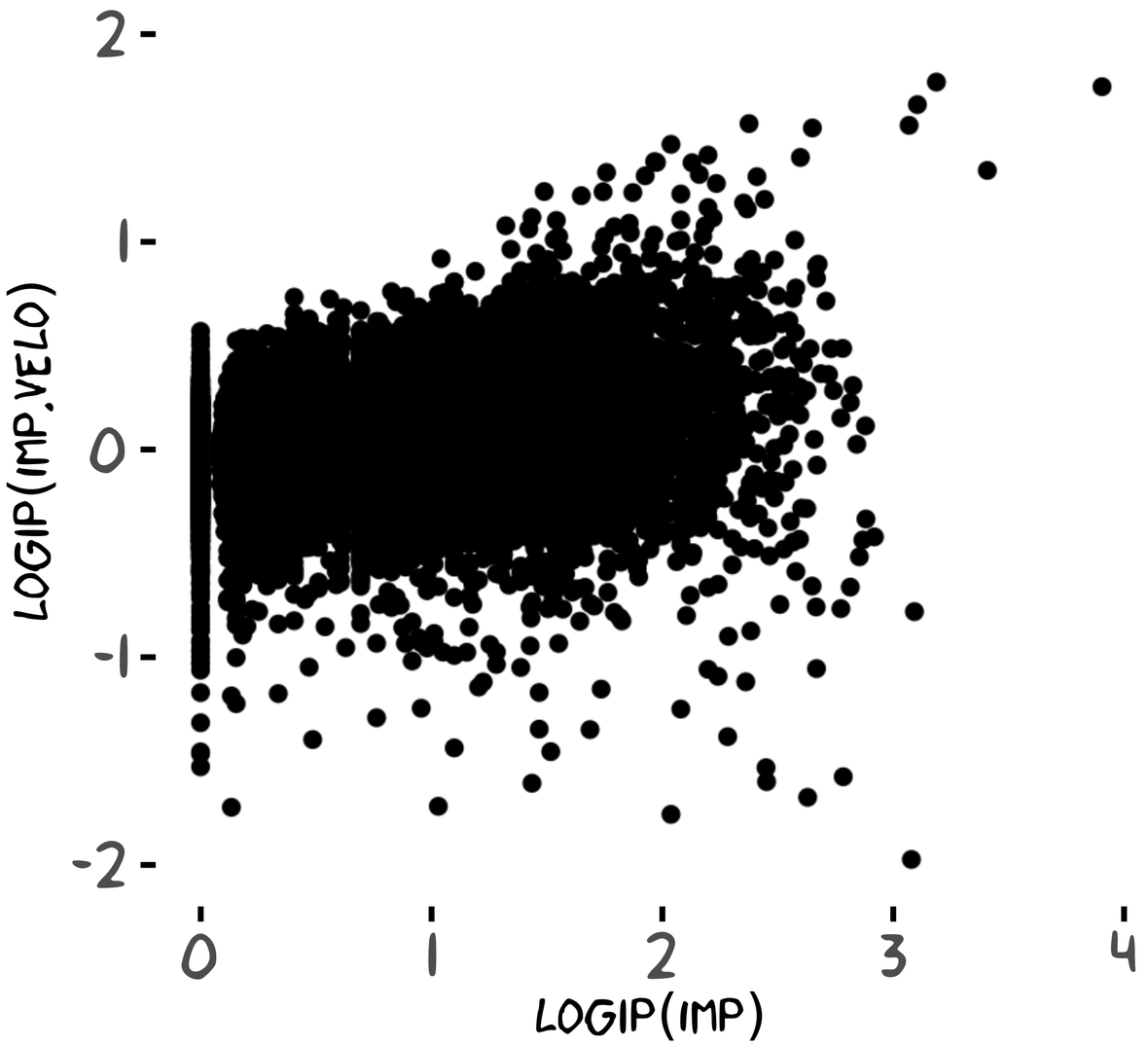
Estimate RNA velocity between ligand and receptor - 1

```
vcs <- inferEpsParams(vcs, scale.factor = 0.73)

## Find nearest neighbours to impute
vcs <- findNN(vcs)

## Impute ligand and receptor
vcs <- imputeLR(vcs)

## Impute velocity LR
vcs <- imputeVeloLR(vcs)
```



Estimate RNA velocity between ligand and receptor - 1

```r
vcs <- inferEpsParams(vcs, scale.factor = 0.73)

## Find nearest neighbours to impute
vcs <- findNN(vcs)

## Impute ligand and receptor
vcs <- imputeLR(vcs)

## Impute velocity LR
vcs <- imputeVeloLR(vcs)
```



Estimate RNA velocity between ligand and receptor - 2

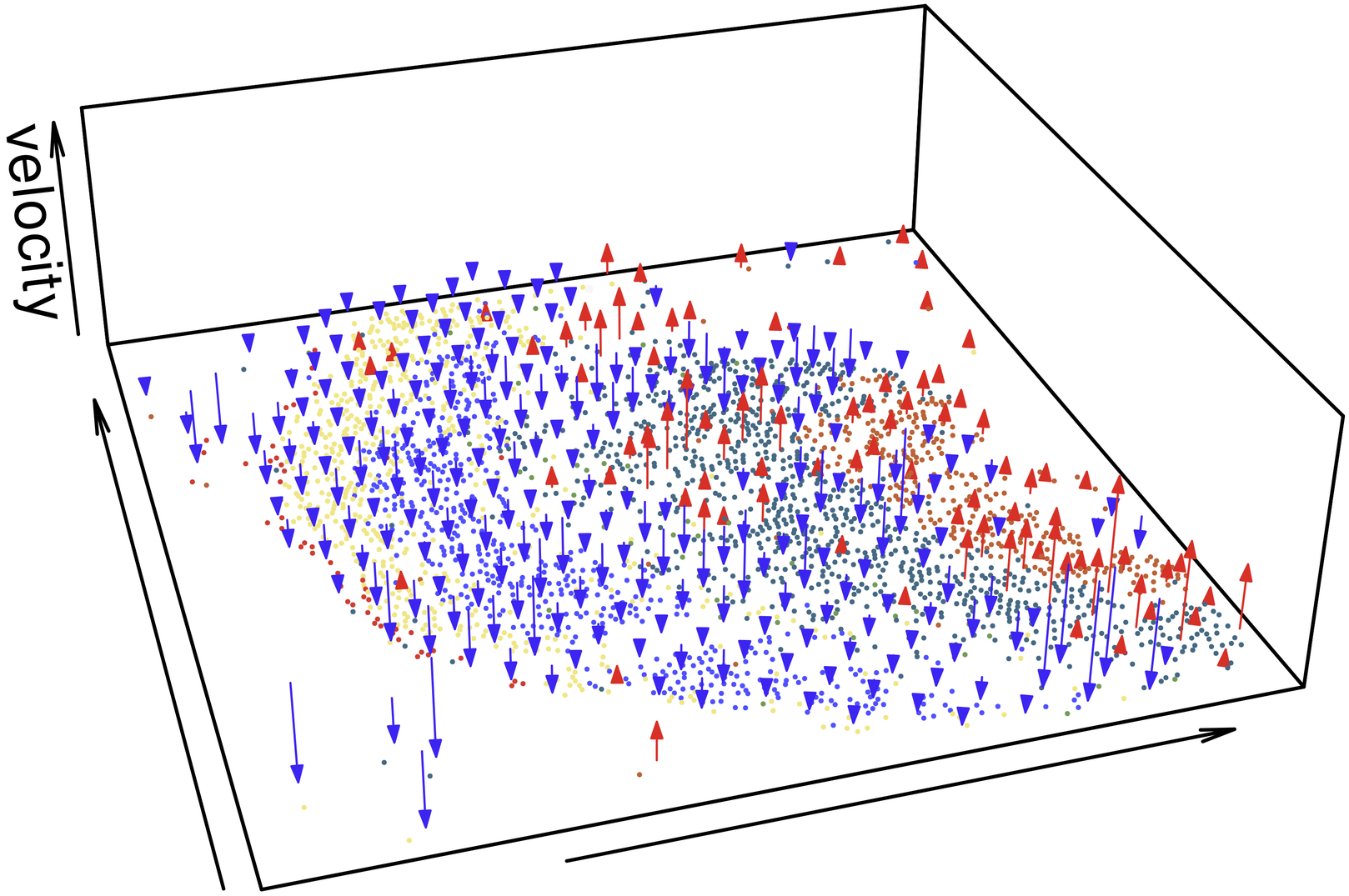
```
vcs <- inferIntrScore(vcs)
vcs <- inferSignif(vcs, p.value = 0.05, reads.thresh = 100, sig.thresh = 100)
vcs <- rankIntrSpatialVar(vcs)
```

```
vcs <- inferIntrVelo(vcs)
```

```
allIntrs <- showIntr(vcs, slot.use = "GauEps-Raw", return.name = TRUE)
print(allIntrs)
```

```
##           CPI-CS0D238C22B           CPI-CS0AA769F5B           CPI-CS0AB19226A
## "SEM3A-PlexinA2_complex1" "SEM3A-PlexinA3_complex1" "SEM3A-PlexinA4_complex1"
```

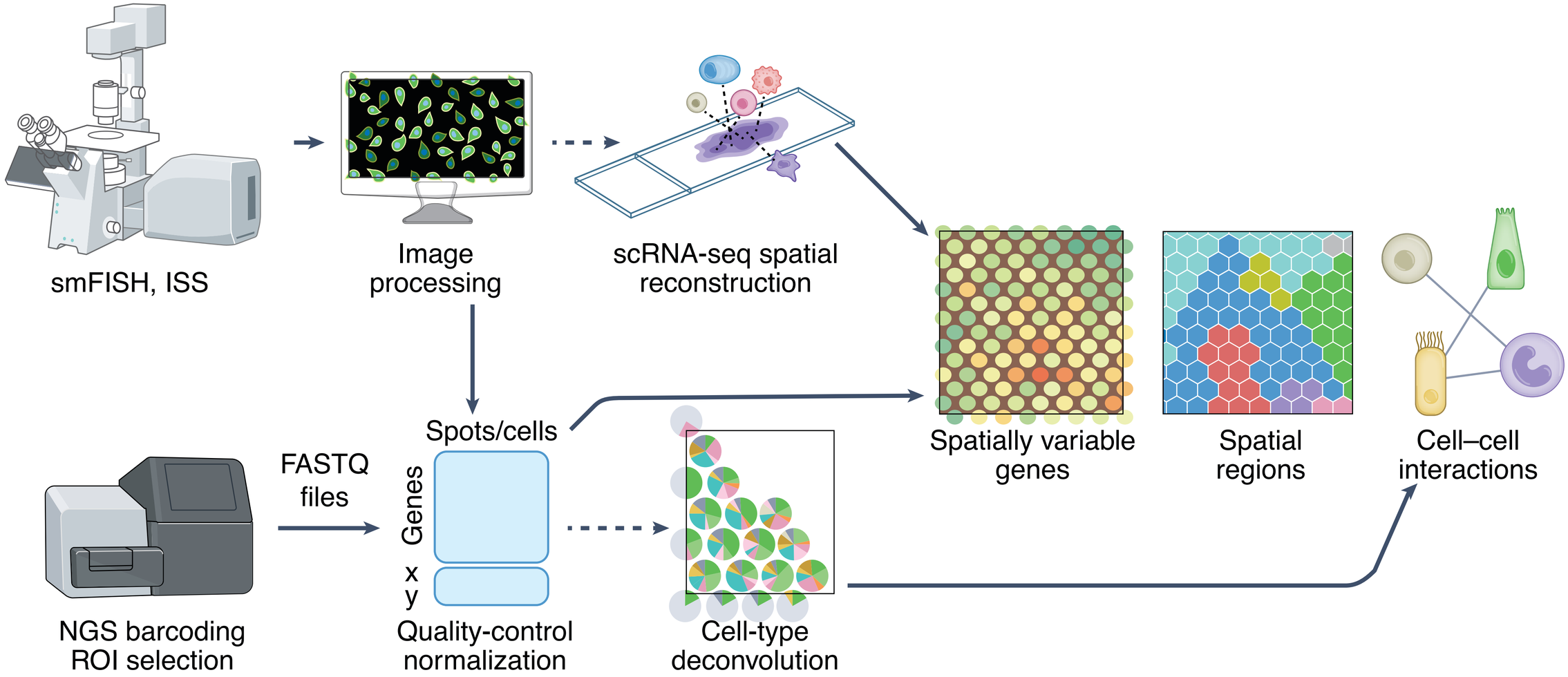
Velo-SEM3A-PlexinA2_complex1



Today’s lecture: Spatial Transcriptomics

- **Technology**
  - Sequencing-based vs. imaging-based
- **Compositional analysis (deconvolution)**
  - Direct deconvolution approach
  - Differential expression analysis
- **Cell-cell communication**
  - Learning ligand-receptor enrichment (mass action law)
  - Spatio-temporal mapping

Spatial transcriptomics data analysis



Moses & Pachter *Nature Methods* (2022)