



Exploring Spatial Transcriptomics

2020-05-28
Alma Andersson



<https://github.com/almaan>



SPATIAL
research
<https://www.spatialresearch.org>

A brief Introduction

- Alma Andersson
- From : Utterbäck, Sweden
 - Population : 69
- Now : Stockholm, Sweden
 - Population : 1,605,030
- 2017-2018 : Delemotte Lab
 - Molecular Dynamics
 - Membrane proteins (Ion Channels)
- 2018-Current : Lundeberg Lab
 - Spatial Transcriptomics (ST)
 - Computational Method Development
- github : almaan
- Small disclaimer : first online teaching experience

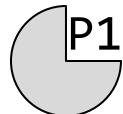


■ ■ | Outline

- Introduction
- Notation
- Background
- Data Processing
- Data Analysis (Overview)
- Break
- Questions
- Cont. Data Analysis (Overview)
- Exercises (Information)
- Questions

■ ■ | Notation

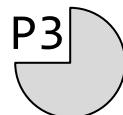
- Exercise session consists of 3 parts
- Symbols below used to indicate when material is included in one of these



- Material in Part 1



- Material in Part 2



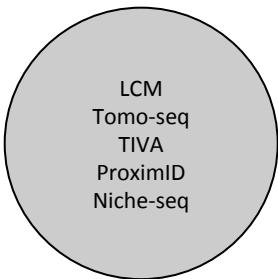
- Material in Part 3

|| ■■ Background ■■||

The spatial space | Overview of techniques

The spatial space | Overview of techniques

Microdissection

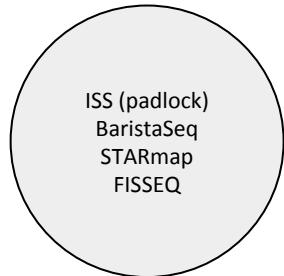


Isolate a region of interest, place isolate in separate well and sequence (either by bulk or single-cell methods).

“Brute Force” approach.

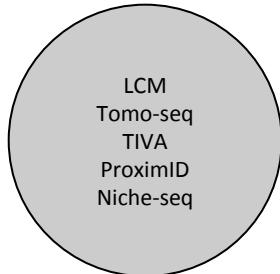
The spatial space | Overview of techniques

In situ sequencing



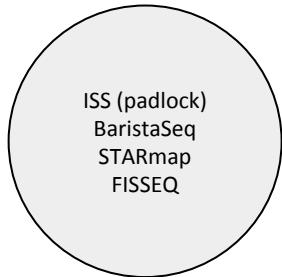
Sequence the transcripts in place. Offer sub-cellular resolution. Tend to rely on gene panels. Need "*a priori*" defined targets.

Microdissection

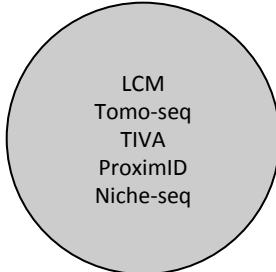


The spatial space | Overview of techniques

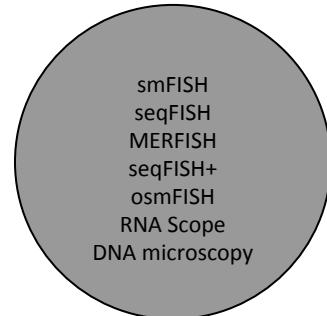
In situ sequencing



Microdissection



In situ hybridization

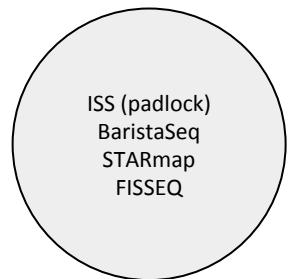
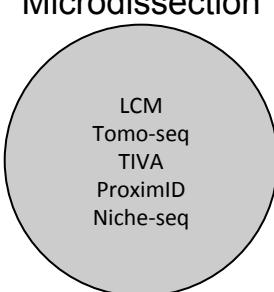


Labeled probes for specific targets, hybridize in place and visualized for spatial information.

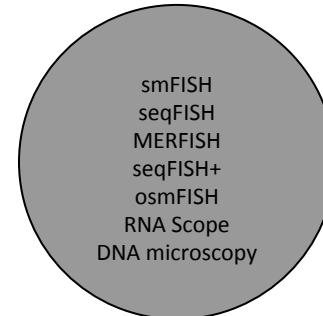


The spatial space | Overview of techniques

In situ sequencing



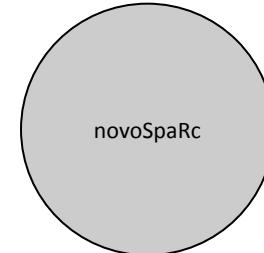
In situ hybridization



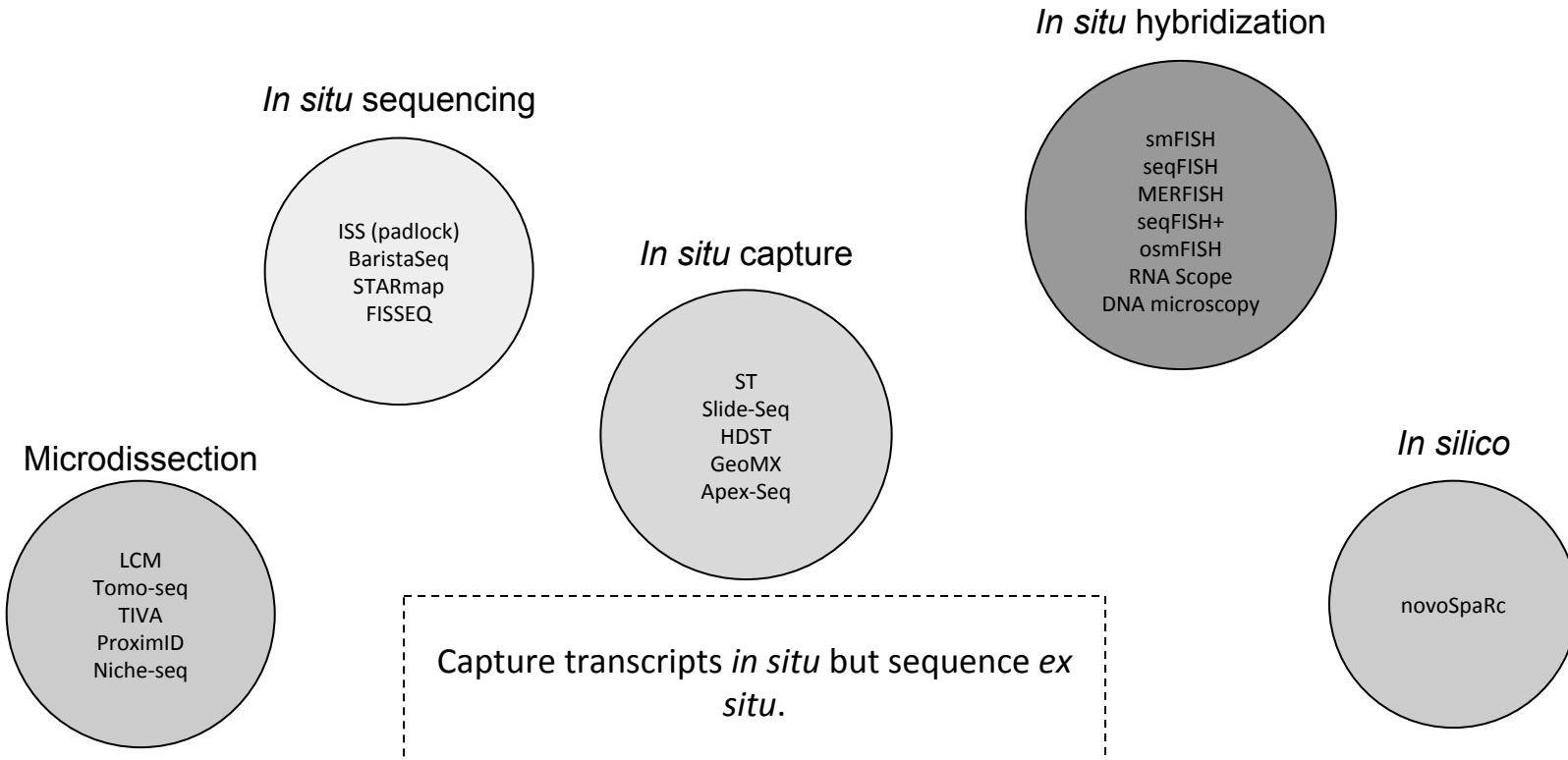
Microdissection

Infer and reconstruct spatial structure from non-spatial data (e.g., single cell).

In silico



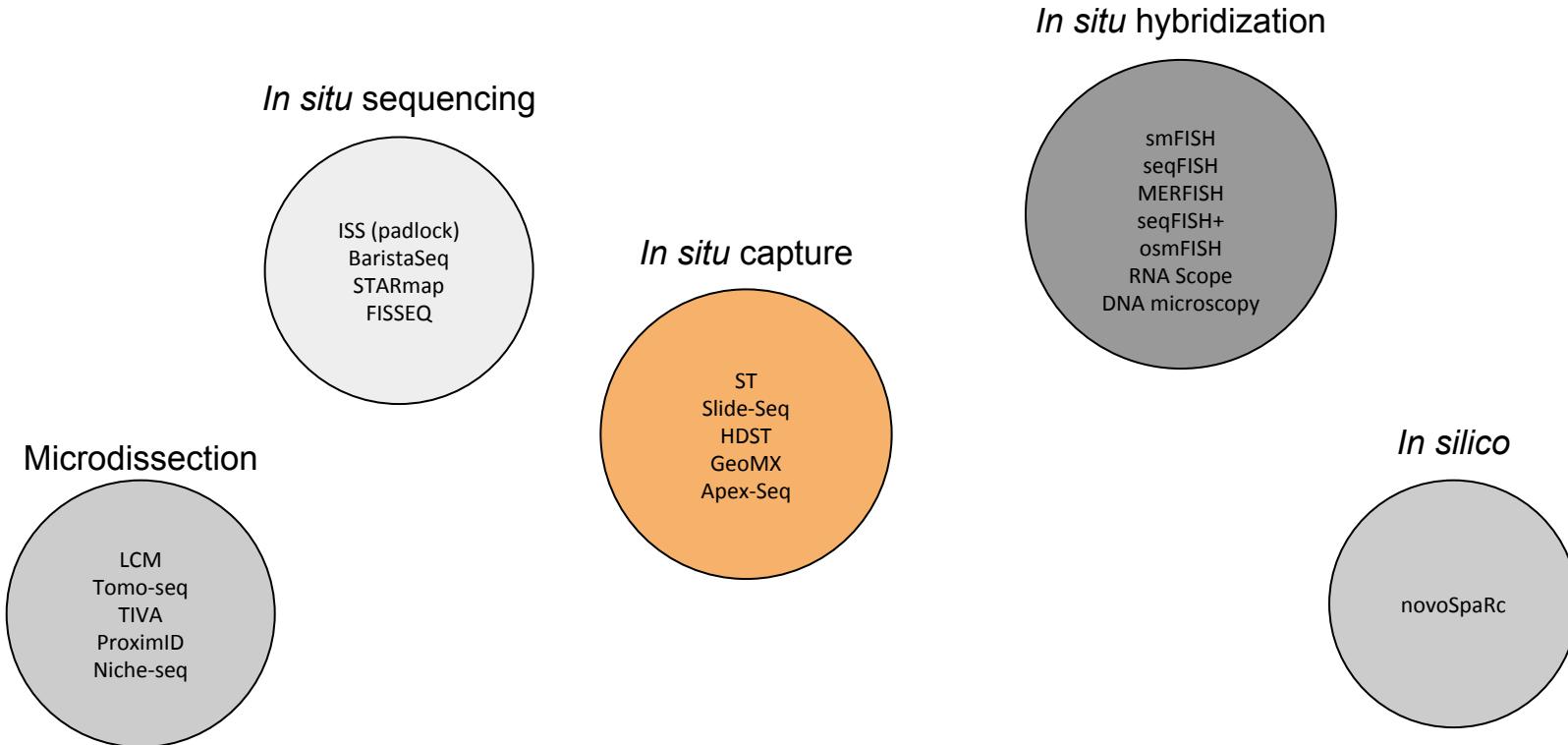
The spatial space | Overview of techniques



Credit to J.Bergenstråhle and M.Asp for categorization of techniques



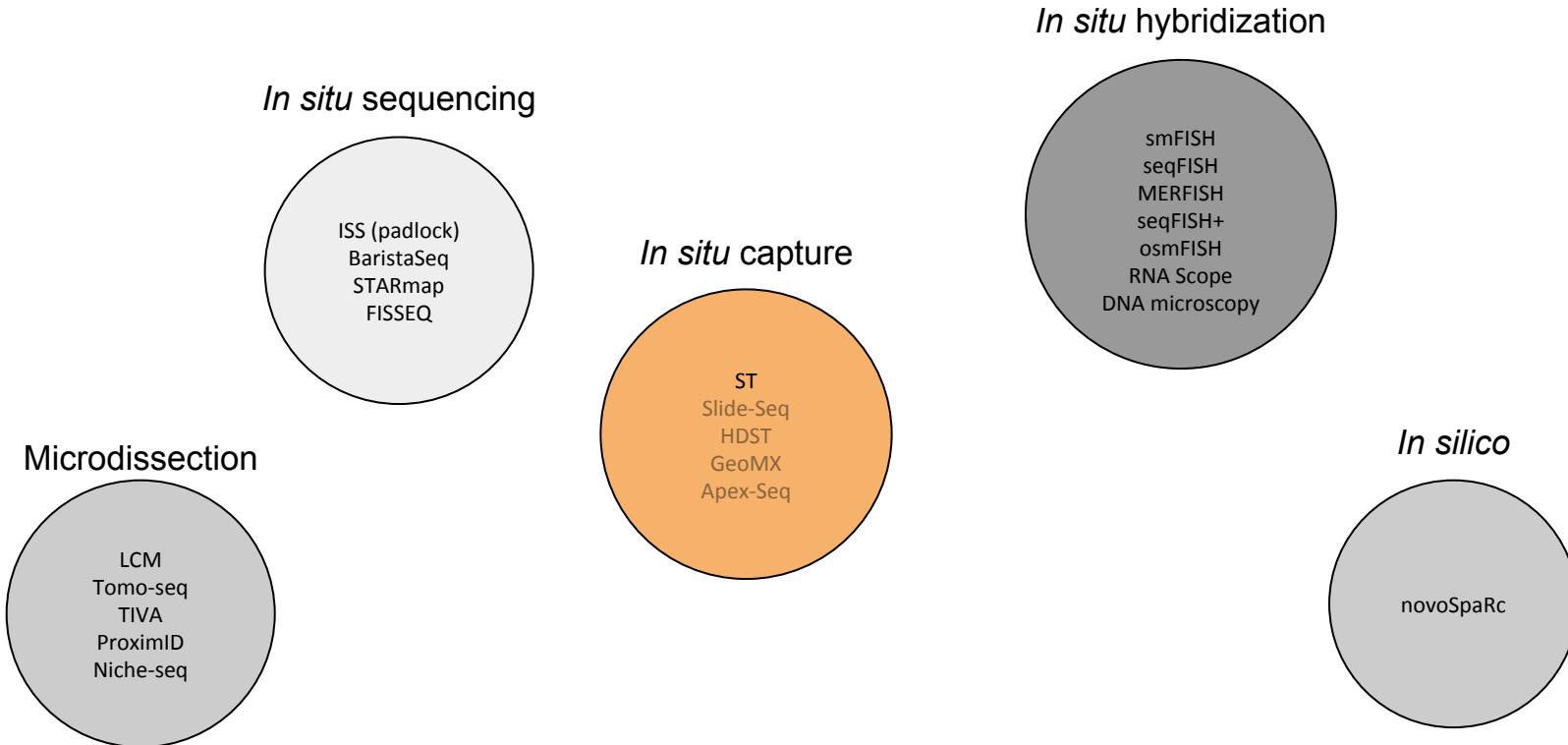
The spatial space | Overview of techniques



Credit to J.Bergenstråhle and M.Asp for categorization of techniques



The spatial space | Overview of techniques



Credit to J.Bergenstråhle and M.Asp for categorization of techniques

Spatial Transcriptomics (ST)

Mid 2016

TRANSCRIPTION

Visualization and analysis of gene expression in tissue sections by spatial transcriptomics

Patric L. Ståhl,^{1,2,*} Fredrik Salminen,^{3,7} Sanja Vicković,⁷ Anna Lundmark,^{2,3} José Fernández-Nebro,^{1,3} Jens Magnusson,¹ Stefania Giannella,⁷ Michaela Aspli,² Joakim Lundeberg,^{1,3} Mikael Lindström,¹ Åsa Sandelin,¹ Stefan Söderström,¹ Simona Codreanu,^{4,5} Åke Borgz,⁶ Fredrik Poulton,⁷ Paul Igiv Coates,⁷ Petter Salminen,² Jan Muster,² Olaf Bergmann,¹ Joakim Lundeberg,^{1,2} Jonas Frisell,²

Analytic of the distribution of proteins or messenger RNAs (mRNAs) in *Neuroscience* tissue sections is a key tool in the molecular research of disease. This study presents the visualization of few proteins or expressed genes at a time. We have developed a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution. By using a combination of two-dimensional positional barcoding and reverse sequencing primers with unique political barcodes, we demonstrate high-quality RNA sequencing data with maintained two-dimensional positional information. The mouse brain tissue sections were used to validate the quantifiable gene expression data and visualization of the distribution of mRNAs within tissue sections and enable novel types of bioinformatics analyses, valuable in research and diagnostics.

Tissue sequencing (RNA-seq) (*i*) of homogenized samples has been used to analyze averaged transcriptome and loss of spatial information. The positional context of gene expression is of key importance to understand-

ing tissue functionality and a-

Several strategies have re-

cently been developed to

analyze the transcriptome in the standard re-

setting of regular histological

samples. However, it is difficult to intro-

duce positional molecular infor-

mation within the context of an

RNAseq. We first use a

modified version of the mRNA (mRNA) in tissue sections o-

modified reverse transcriptase

(RT) to generate cDNA from

adult mouse olfactory bu-

ndles. We then use RT to generate

expression reference data. T

he cDNA is then fragmented,

reverse transcribed, and

fluorescently labeled on nucle-

*These authors contributed equally to this work. These ad-

dress correspondence to: *Corresponding author: E-mail:*

jokim.lundeberg@kth.se

synthesized cDNA (Fig. 1A, and fig. S1). The tissue was then enzymatically removed, which left cDNA coupled to the service oligonucleotides on the slide (Fig. 1B). After this, the cDNA was extracted corresponding to the tissue structure revealed by the general histology (Fig. 1B and C), and the cDNA was stored in aliquots (Fig. 1C, and Fig. 1D, E, F, G, H, I, K, L to G). By comparing the hematoxylin-and-eosin and fluorescence signals, we could measure the average distance between the center of a cell of a cell to 1.7 ± 2 μm (mean ± SD) (Fig. S1, K to H).

We then used this to measure mRNA in tissue sections with minimal diffusion

and maintained positional representation meth-

odology (Fig. 2A), and we denoted this strategy

"spatial transcriptomics."

We deposited ~20 million

sequencing reads onto a

slide with a diameter of 100 μm and a center-to-center distance of 200 μm over an area of 0.1 mm by 0.1 mm (Fig. S2).

We generally sequencing libraries based on

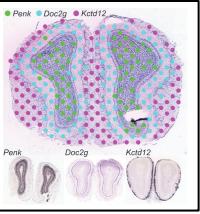
adult mouse brain tissue (Fig. 2A,

and Fig. S2). Comparison with data from RNA

extracted and fragmented in solution revealed

high correlation between the two methods.

The correlation between the surface and in-solution



Late 2018

10x GenomicsTM
acquisition

Science Publication
Ståhl et.al

DOI: 10.1126/science.aaf2403

Spatial Transcriptomics (ST)

Mid 2016

TRANSCRIPTION

Visualization and analysis of gene expression in tissue sections by spatial transcriptomics

Patrik L. Ståhl,^{1,2,*} Fredrik Salminen,^{3,7} Sanja Vicković,⁷ Anna Lundmark,^{2,3} José Fernández Navarro,^{1,3} Jens Magnusson,¹ Stefania Giannella,⁷ Michaela Asplund,² Joakim Lundeberg,^{1,3} Mikael Lindström,¹ Annika Söderström,¹ Simon Cudkovich,^{4,5} Åke Borg,² Fredrik Pontén,² Paul Igro Coates,² Petter Salminen,² Jan Musterer,² Olaf Bergmann,² Joakim Lundeberg,² Jonas Frisell,²

An analysis of the distribution of proteins and messenger RNAs (mRNAs) in *Neurogranin* (Ngn) tissue sections is presented. By using a novel visualization strategy, we can easily monitor the distribution of few proteins or expressed genes at a time. We have developed a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution. By using a combination of standard reverse transcription and barcoded, unique reverse transcription primers with unique political barcodes, we demonstrate high-quality RNA sequencing data with maintained two-dimensional positional information from the mouse brain. By using this strategy, we can now rapidly analyze gene expression patterns and visualization of the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics.

Tissue transcriptomics are typically studied by RNA sequencing (RNA-seq) (*i*) of homogenized samples. This approach has several averaged transcriptome and loss of spatial information. The positional context of gene expression is of key importance to understand-

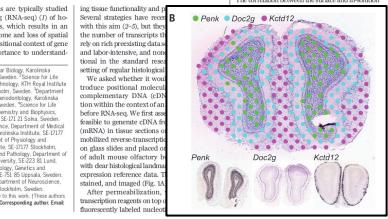
ing tissue functionality and disease.

Several strategies have recently been developed to address this issue. The number of transcripts detected in a single cell is proportional to the number of transcripts in the standard reference setting of regular histological sections (*ii*). However, it is difficult to introduce positional molecular information in this context of an entire RNA-seq. We first used a modified version of the mRNA (mRNA) in tissue sections or mobilized reverse transcriptase (RT)-PCR to generate a library of adult mouse olfactory bulb (OB) tissue sections with spatial gene expression reference data. Tissue sections were fixed, permeabilized, stained, and imaged (Fig. 1A). Then, we used a set of three common transcription reagents on top of fluorescently labeled nucleic

acid probes.

Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *2* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *3* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *4* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *5* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *6* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *7* Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden. *Corresponding author. E-mail: jonas.frisell@ki.se

These authors contributed equally to this work. These authors contributed equally to this work. Corresponding author. E-mail: jokim.lundeberg@ki.se



Science Publication
Ståhl et.al

DOI: 10.1126/science.aaf2403

synthesized cDNA (Fig. 1A, and fig. S1). The tissue was then enzymatically removed, which left cDNA coupled to the service oligonucleotides on the slide (Fig. 1B). After this, the tissue was cut into small pieces corresponding to the tissue structure revealed by the general histology (Fig. 1B and C), and the cDNA was sheared into fragments of approximately 200 bp (Fig. 1C, D to G). By comparing the hematoxylin-and-eosin and fluorescence signals, we could measure the average size of the tissue sections to be approximately 0.4 to 1.7 to 2.1 μm (mean ± SD) (Fig. S1, F to H).

We then used a standard RT-PCR to amplify mRNA in tissue sections with minimal diffusion

and maintained positional representation meth-

odology (Fig. 2A), and we denoted this strategy

"spatial transcriptomics."

We deposited ~20 million

RNA-seq reads in the NCBI Gene Expression

Archive (GSE62900) (Fig. 2B). Comparison of

our general sequencing libraries based on

area and depth in Fig. 2B (Fig. S2,

A and B) (Fig. 2C). Comparison with data from RNA

extraction and fragmentation in solution revealed

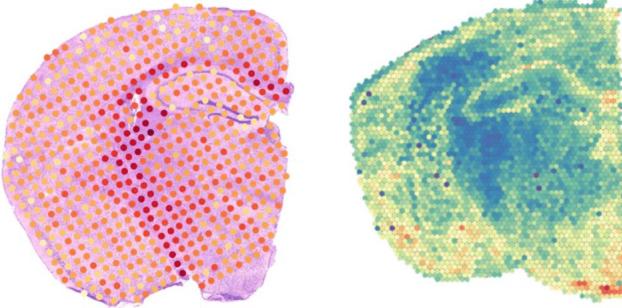
similar results (Fig. 2D). The correlation of the

method was also found with the other (Fig. S2).

The correlation between the surface and in-situ RNA

Late 2018
10x GenomicsTM
acquisition

Late 2019



Launch of **Visium** Spatial Gene
Expression Platform

SciLifeLab



Spatial Transcriptomics (ST) → Visium

Mid 2016

TRANSCRIPTION

Visualization and analysis of gene expression in tissue sections by spatial transcriptomics

Patrik L. Ståhl,^{1,2,*} Fredrik Salhün,^{2,3} Sanja Vicković,^{2,4} Anna Lundmark,^{2,5,6} José Fernández-Nebro,^{1,2} Jens Magnusson,¹ Stefania Giannella,² Michaela Aspli,² Joakim Lundeberg,^{1,2} Mikael Lindström,¹ Åsa Berg,² Anders Sandelin,² Simon Codeluppi,^{2,4} Åke Borg,² Fredrik Pontén,² Paul Igivu Costa,² Petter Salhün,² Jan Muster,² Olaf Bergmann,² Joakim Lundeberg,¹ Jonas Frisell,²

Analysis of the distribution of proteins and messenger RNAs (mRNAs) in *Neurogranin* (Ngn2) sections is a key step in the visualization and research of gene expression. This study provides the distribution of few proteins or expressed genes at a time. We have developed a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with single-cell resolution. By using a combination of spatial transcriptomics and reverse transcription with unique political barcodes, we demonstrate high-quality RNA sequencing data with maintained two-dimensional positional information from the mouse brain. By using the same approach in tissue sections, we demonstrate how spatial transcriptomics can visualize the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics.

Tissue transcriptomes are typically studied by RNA sequencing (RNA-seq) (*J*) of homogenized samples. This approach has averaged transcriptome and loss of spatial information. The positional context of gene expression is of key importance to understand-

ing tissue functionality and disease processes.

Department of Cell and Molecular Biology, Karolinska Institutet, Stockholm, Sweden; *2*KTH Royal Institute of Technology, Division of Biotechnology, KTH Royal Institute of Technology, Stockholm, Sweden; *3*Department of Oral Medicine, Division of Periodontology, Karolinska Institutet, Stockholm, Sweden; *4*Department of Cell and Molecular Biology, KTH Royal Institute of Technology, Stockholm, Sweden; *5*Department of Radiology, Karolinska Institutet, Stockholm, Sweden; *6*Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden; ^{*}These authors contributed equally to this work. [†]These authors contributed equally to this work. Corresponding author: jonasfrisell@kth.se

ing the number of transcripts per cell. Several strategies have recently been developed to maintain the spatial resolution in the standard research setting of regular histological sections (*K*). However, these approaches introduce positional molecular noise and do not allow for a full resolution within the context of an entire RNA-seq. We first used a modified version of the spatial transcriptome (mST) method to analyze mRNA in tissue sections or mobilized reverse transcriptase (mRT) on tissue sections. We then analyzed the distribution of adult mouse olfactory bulb (OB) mRNA expression reference data. Tissue sections were stained, imaged (Fig. 1A), and processed with unique transcriptome transcription reagents on top of fluorescently labeled nucleic acids.

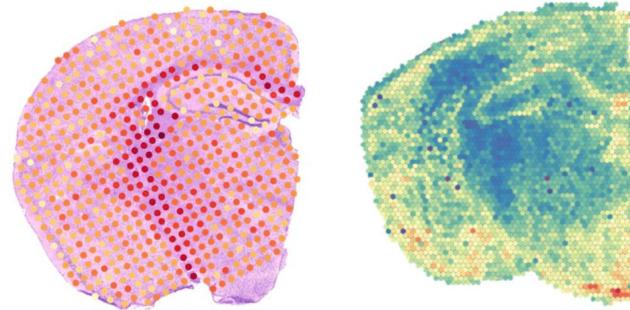


Science Publication
Ståhl et.al

DOI: [10.1126/science.aaf2403](https://doi.org/10.1126/science.aaf2403)

Late 2018
10x Genomics™
acquisition

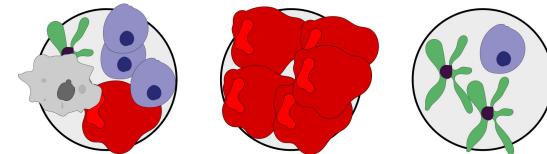
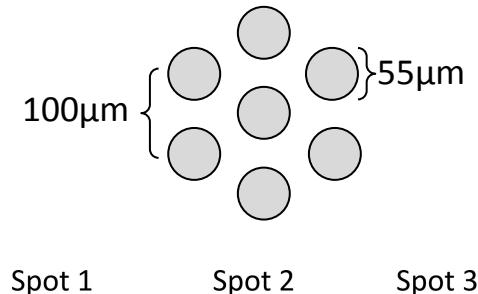
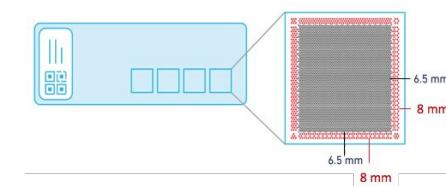
Late 2019



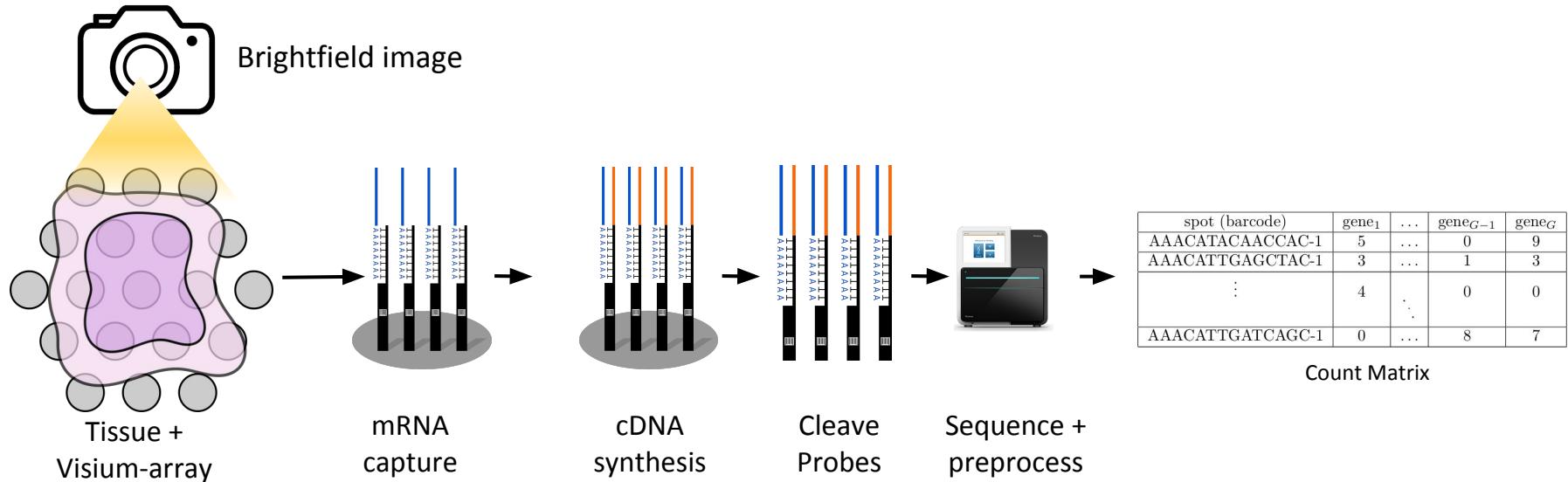
Launch of **Visium** Spatial Gene
Expression Platform

Visium Platform

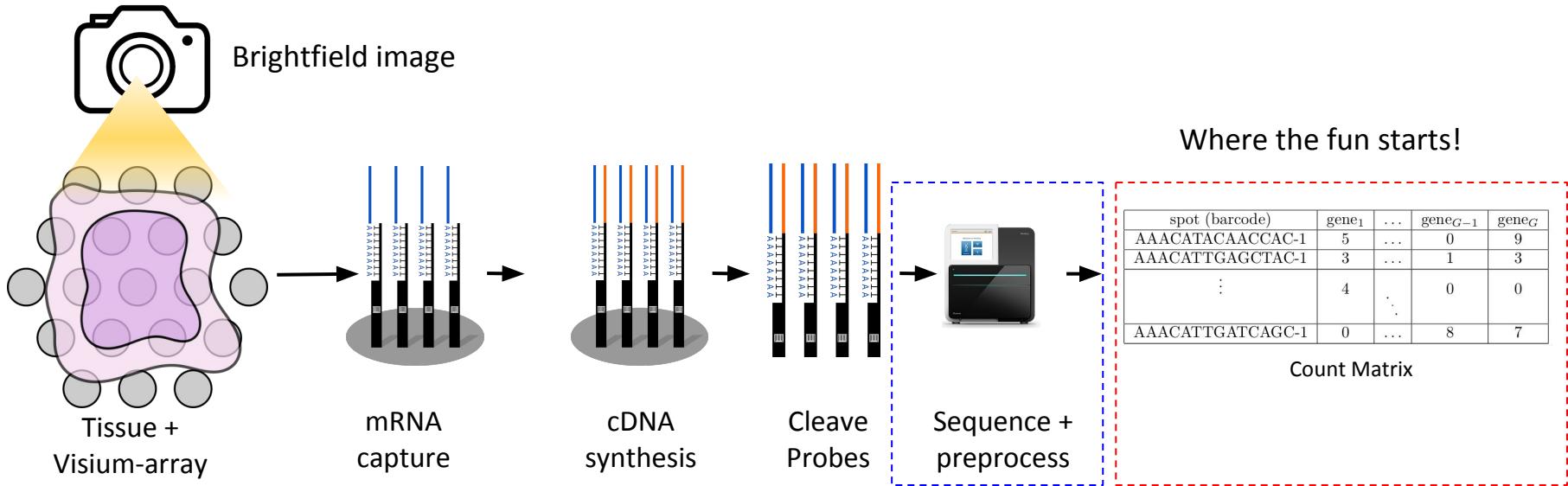
- Array based technique
- 6.5mm x 6.5mm area to put sample on
- 4992 spots arranged in regular grid
- Spot specs:
 - Spot diameter : 55 μm
 - Center to center distance : 100 μm
- Each spot has millions of capture probes
 - spatial barcode
 - polyT sequence
 - captures polyadenylated mRNA
 - Full transcriptome(-ish)
- ~ 1-10 cells contribute to each spot
 - **NOTE :** Not single cell resolution!



The experimental workflow (in a nutshell)



The experimental workflow (in a nutshell)



■ ■ ■ Data Processing ■ ■ ■

After sequencing (brief)

- **spaceranger mkfastq** | BCL files to FASTQ
- **spaceranger count** | tissue detection/alignment, UMI counting



```
-bash-4.2$ tree -L 2
.
├── analysis
│   ├── clustering
│   ├── diffexp
│   ├── pca
│   ├── tsne
│   └── umap
└── cloupe.clope
    ├── filtered_feature_bc_matrix
    │   ├── barcodes.tsv.gz
    │   ├── features.tsv.gz
    │   └── matrix.mtx.gz
    ├── filtered_feature_bc_matrix.h5
    ├── metrics_summary.csv
    ├── molecule_info.h5
    ├── possorted_genome_bam.bam
    ├── possorted_genome_bam.bai
    ├── raw_feature_bc_matrix
    │   ├── barcodes.tsv.gz
    │   ├── features.tsv.gz
    │   └── matrix.mtx.gz
    └── raw_feature_bc_matrix.h5
    └── spatial
        ├── aligned_fiducials.jpg
        ├── detected_tissue_image.jpg
        ├── scaleFactors.json.json
        ├── tissue_hires_image.png
        ├── tissue_lowres_image.png
        └── tissue_positions_list.csv
    └── web_summary.html
```

Automatically generated analysis

filtered = spots under tissue

raw = all spots

For mapping between image coordinates

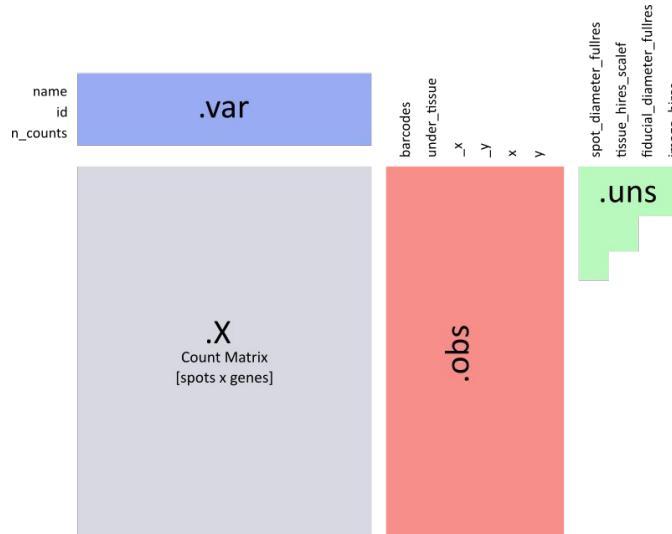
Resized images

Maps barcode to coordinate

Example of **spaceranger count** output

Processed data

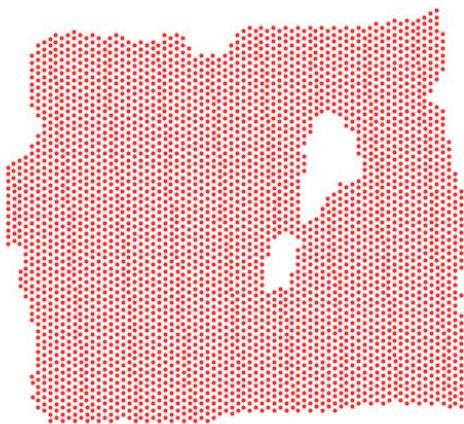
- Either use .mtx files or .h5 files to assemble a data object to work with
 - No standardized format
- Personal preference : convert to .h5ad file (will be using in exercises)
 - scanpy/anndata teams working on - soon to release - their own (similar) format



An initial assessment

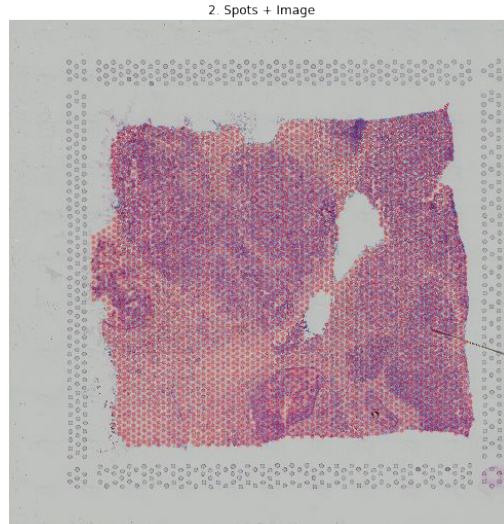
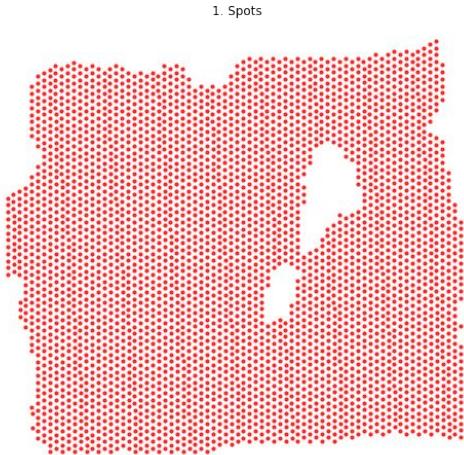
- Example with Human Breast cancer data
 - Public data : Available at 10x website

1. Spots



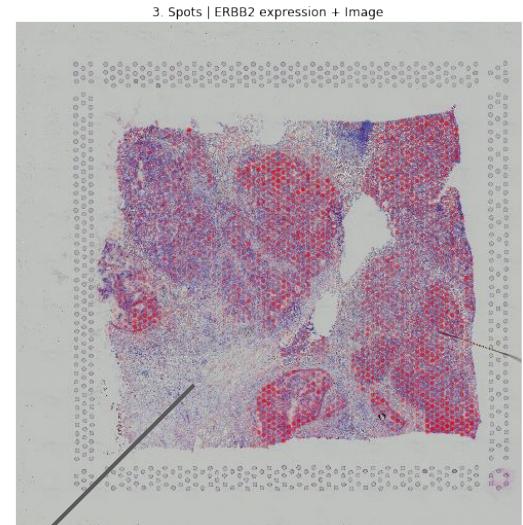
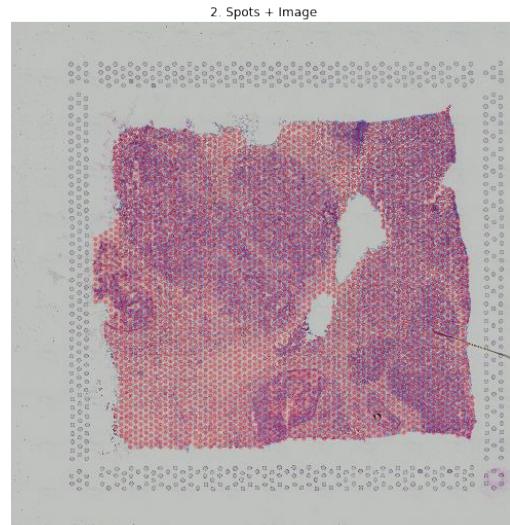
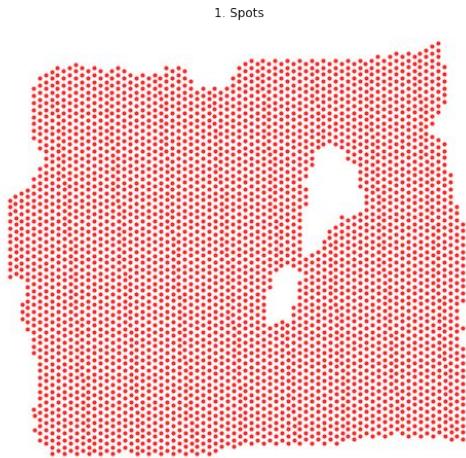
An initial assessment

- Example with Human Breast cancer data
 - Public data : Available at 10x website



An initial assessment

- Example with Human Breast cancer data
 - Public data : Available at 10x website



Facecolor intensity proportional
to gene expression value

■ ■ | Visualizing high dimensional spatial data

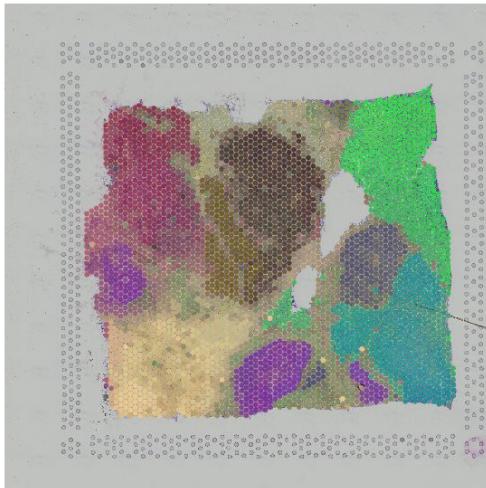
- We visualized one feature/gene (*ERBB2*)
- How do we handle multiple non-mutually exclusive features?

■ ■ ■ Visualizing high dimensional spatial data

- We visualized one feature/gene (*ERBB2*)
- How do we handle multiple non-mutually exclusive features?
- One idea :
 - Embed gene expression data in 3 dimensional space (e.g. using UMAP)
 - Do affine transformation to unit cube
 - Consider values as RGB values (or other colorspace) and color spots accordingly

■ ■ ■ Visualizing high dimensional spatial data

- We visualized one feature/gene (*ERBB2*)
- How do we handle multiple non-mutually exclusive features?
- One idea :
 - Embed gene expression data in 3 dimensional space (e.g. using UMAP)
 - Do affine transformation to unit cube
 - Consider values as RGB values (or other colorspace) and color spots accordingly



Regions with similar colors have similar gene expression.



Data Analysis



■ ■ | Before the analysis | Filtering, Normalization, Batch correction, etc.

- No magic recipe to give
 - How to process your data is very much dependent on the samples and objective
 - Much can be learnt from analysis of single cell data
 - Will give some general advice

■ ■ | Before the analysis | Filtering, Normalization, Batch correction, etc.

- No magic recipe to give
 - How to process your data is very much dependent on the samples and objective
 - Much can be learnt from analysis of single cell data
 - Will give some general advice
- Consider filtering :
 - Genes based on expression levels (total expression > thrs)
 - Genes based on spot presence (#spots gene is observed at > thrs)
 - Spots based on expression levels (total gene expression at spot > thrs) †
 - Ribosomal (RP) and mitochondrial genes (MT) tend to exhibit spurious expression patterns. Exclusion of these is common.

† Only necessary if you expect “defunct” spots under the tissue

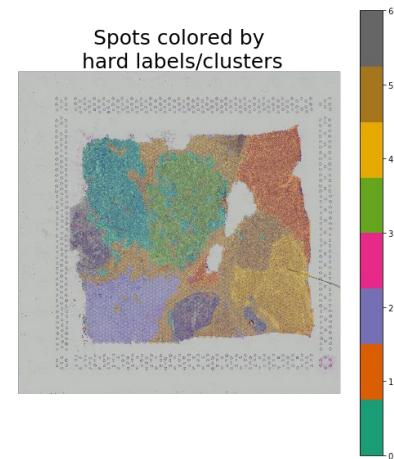
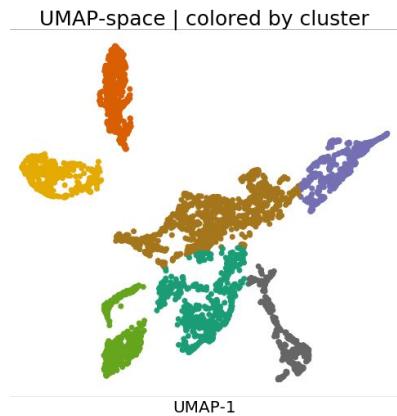
■ ■ ■ Before the analysis | Filtering, Normalization, Batch correction, etc.

- No magic recipe to give
 - How to process your data is very much dependent on the samples and objective
 - Much can be learnt from analysis of single cell data
 - Will give some general advice
- Consider filtering :
 - Genes based on expression levels (total expression > thrs)
 - Genes based on spot presence (#spots gene is observed at > thrs)
 - Spots based on expression levels (total gene expression at spot > thrs) †
 - Ribosomal (RP) and mitochondrial genes (MT) tend to exhibit spurious expression patterns. Exclusion of these is common
- Normalization / batch correction :
 - Recommend to account for spot “library size” - varying cell density
 - Include slide/array as covariate (sometimes big variation is observed)
 - Popular tools for batch correction : sctransform and Harmony

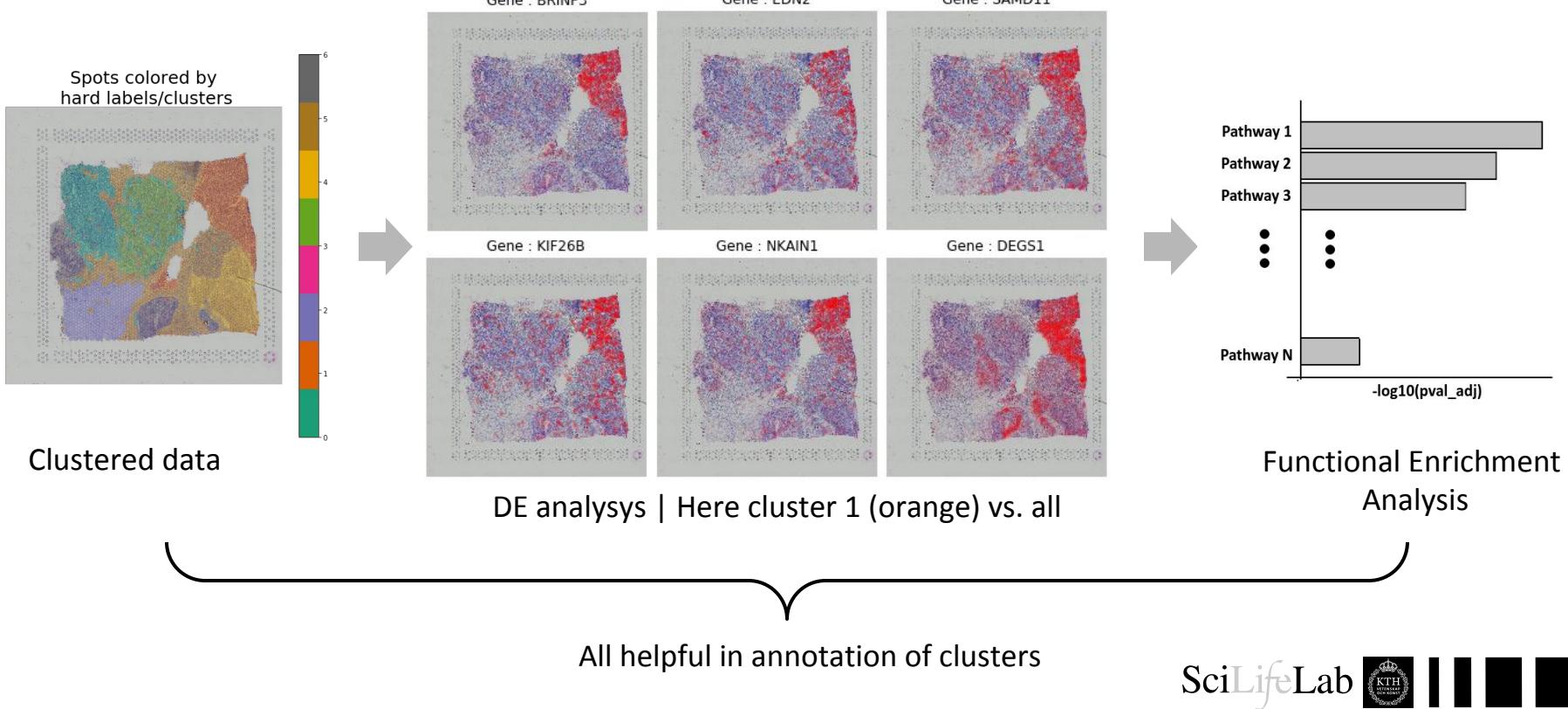
† Only necessary if you expect “defunct” spots under the tissue

Example : Basic Analysis

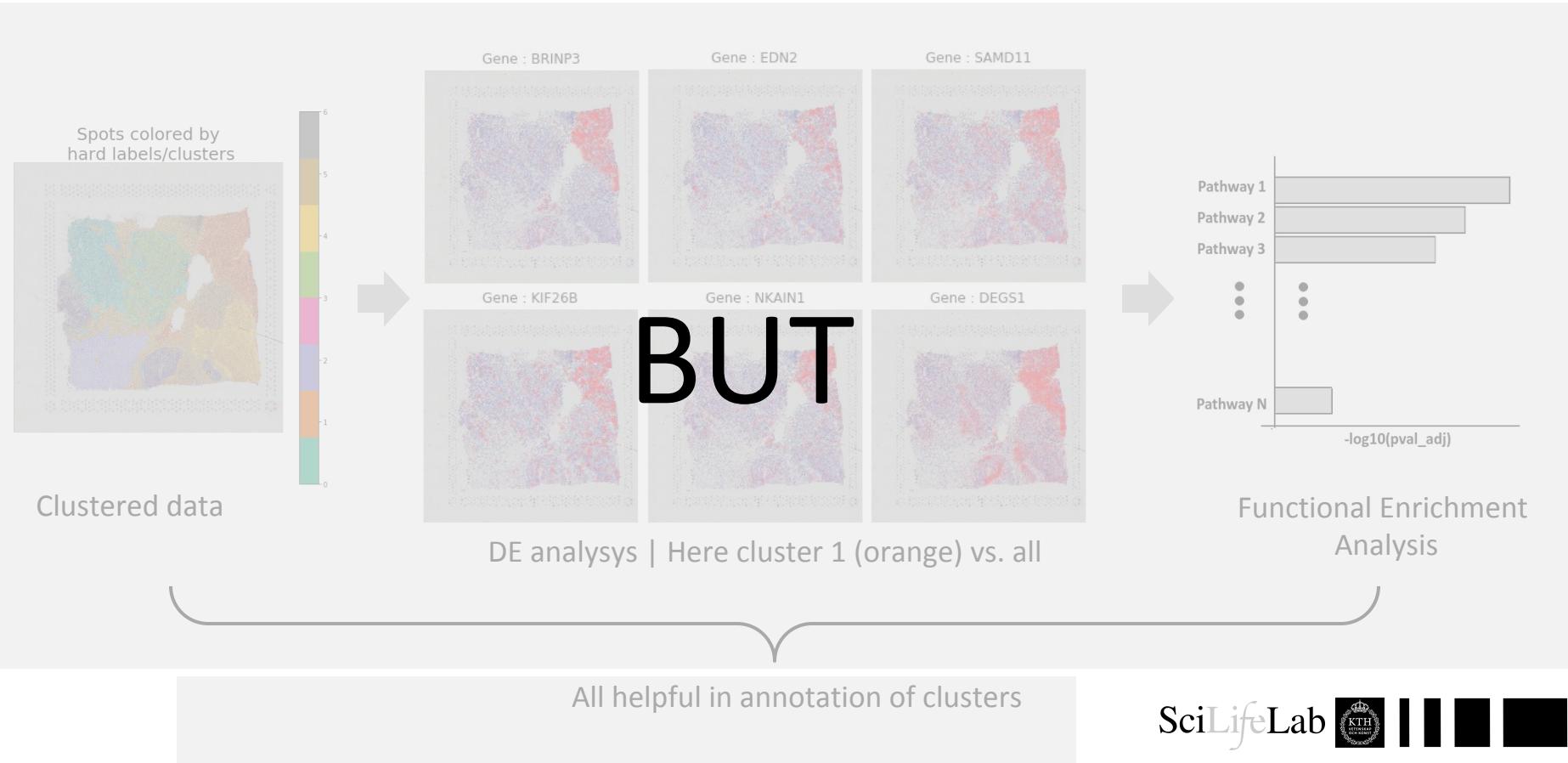
- Cluster the spots based on gene expression
 - Normalize → PCA_{n=20} → UMAP → GMM
- Backmap clusters onto tissue
- Use HE-image as reference
 - Sanity check - does it make sense?
 - Valuable resource
- Next, annotate clusters
 - Find genes associated with cluster
 - Functional enrichment analysis



Example : Basic Analysis



Example : Basic Analysis



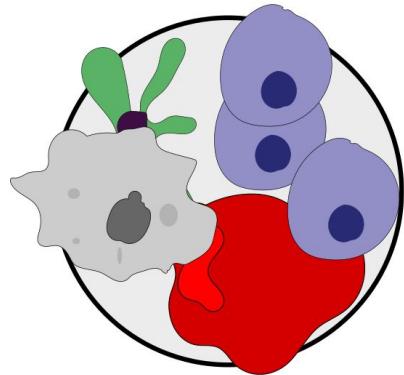


Example : Basic Analysis

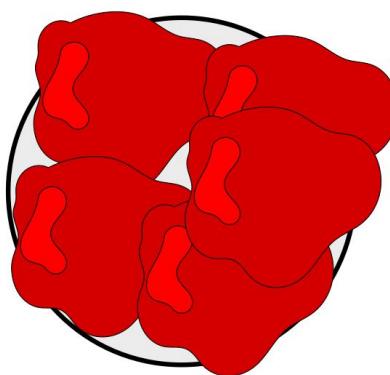
Remember :

*"each spot is a mixture of multiple cells,
i.e., one spot may contain multiple cell types"*

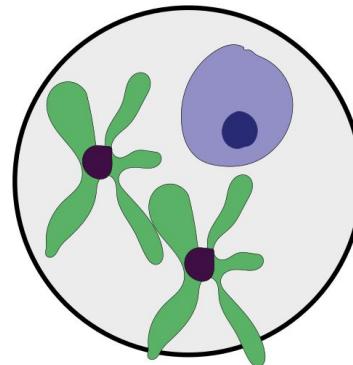
Spot 1



Spot 2

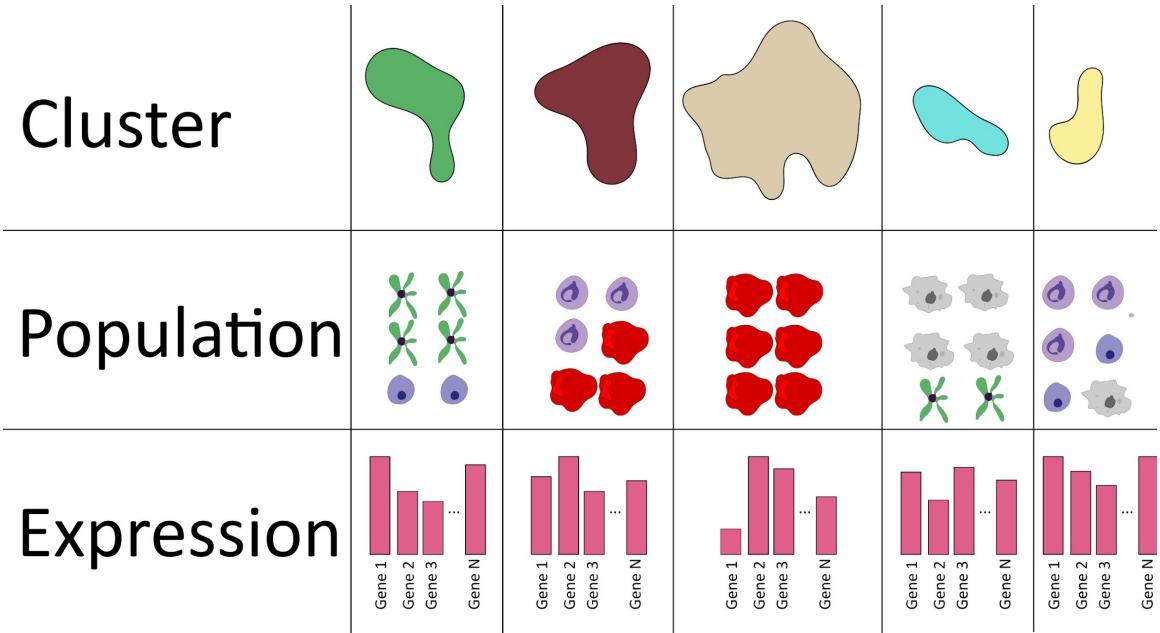
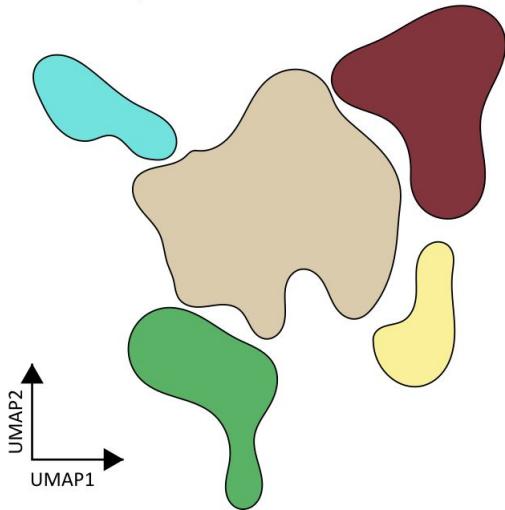


Spot 3



Cluster ≠ Cell type

Clustered Spatial Gene Expression Data



■ ■ | So where are my cell types?

■ ■ || Break

- 10 min Break
- During :
 - Read through questions (for me)
- After :
 - Discuss (some) questions when everyone is back
 - Data Analysis Cont.
 - Information about the exercises
 - Wrap up and questions

(()) C []

■ ■ || Questions

■ ■ | So where are my cell types?

■ ■ || So where are my cell types?

- Marker genes?

So where are my cell types?

- Marker genes? Easy and straightforward, but
 - Requires knowledge of marker genes (not always true)
 - Risk for overlap among marker genes
 - How do we interpret expression values?
 - Lowly expressed markers genes may not always be observed

So where are my cell types?

- Marker genes? Easy and straightforward, but
 - Requires knowledge of marker genes (not always true)
 - Risk for overlap among marker genes
 - How do we interpret expression values?
 - Lowly expressed markers genes may not always be observed
- Alternative solution - Integrate single cell (SC) and spatial data!

So where are my cell types?

- Marker genes? Easy and straightforward, but
 - Requires knowledge of marker genes (not always true)
 - Risk for overlap among marker genes
 - How do we interpret expression values?
 - Lowly expressed markers genes may not always be observed
- Alternative solution - Integrate single cell (SC) and spatial data!
 - Extract information of cell types from SC data and apply to spatial data
 - **Big challenge : deconvolution required (on Visium data)**
 - Multiple approaches have been proposed, e.g., :
 - MIA by Tirosh et al.
 - Seurat's spatial module



Integration of Single Cell and Spatial Data

Single Cell Data
Clusters of cell types

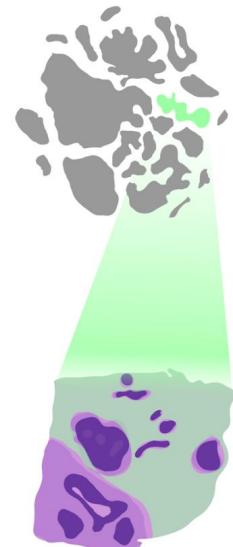
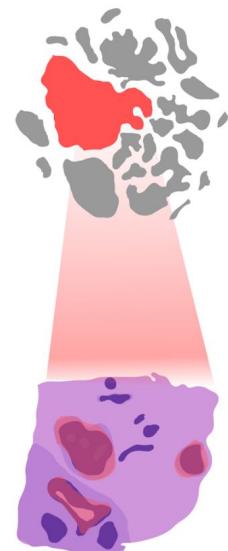


Integration

Spatial Data



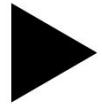
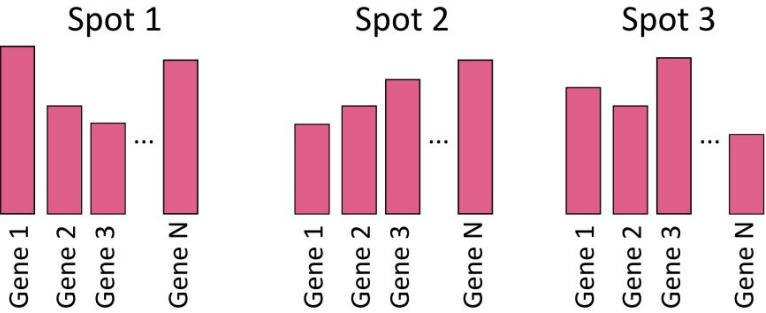
Spatially map each cell type



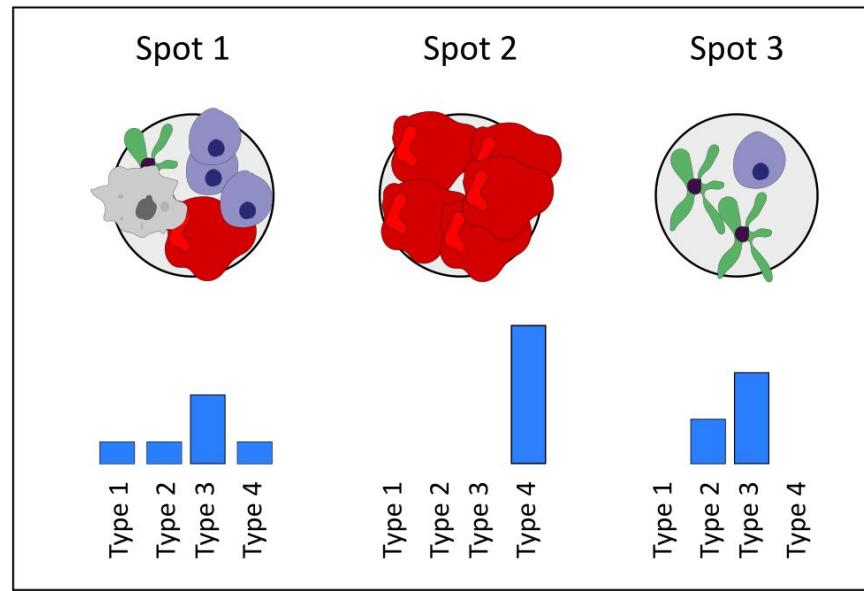


Our objective : deconvolve expression data

From this



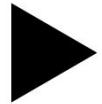
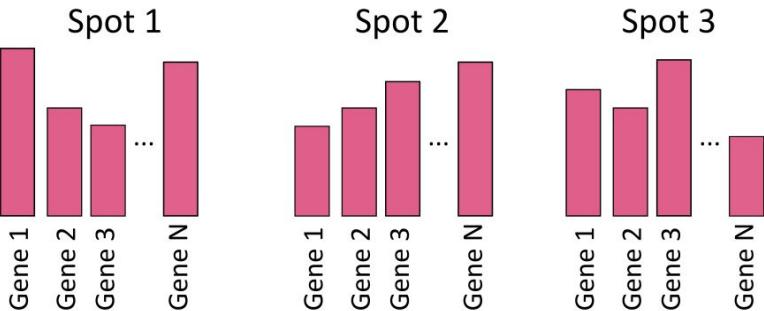
We want this



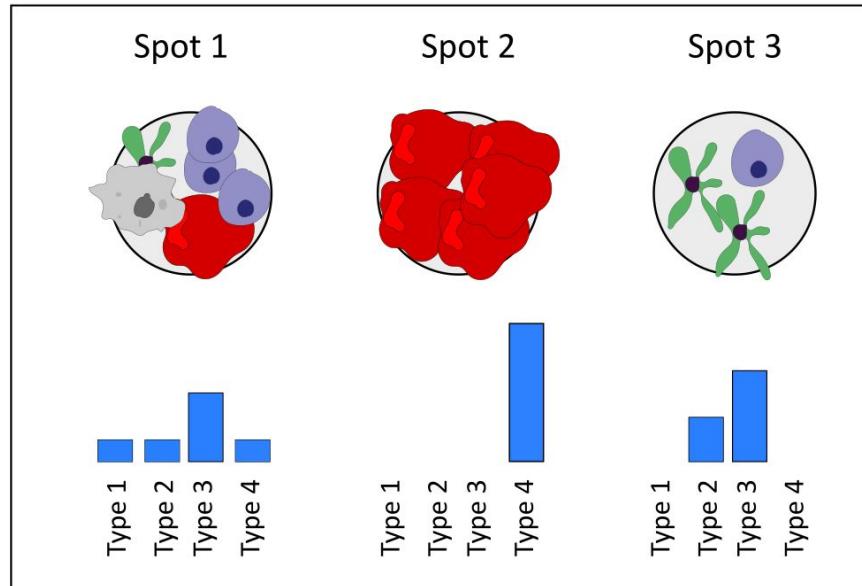


Our objective : deconvolve expression data

From this



We want this

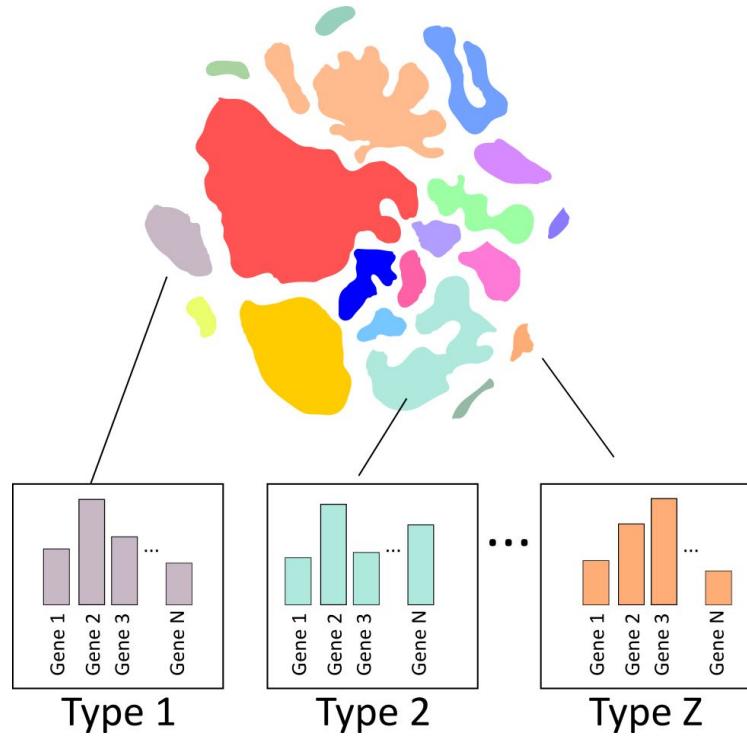


Suggested approach : *Model-based Probabilistic Inference*



It's as easy as 1-2-3

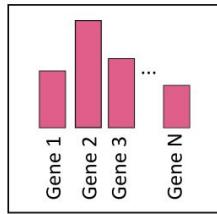
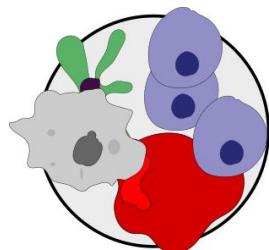
1. Infer cell type expression parameters from SC data





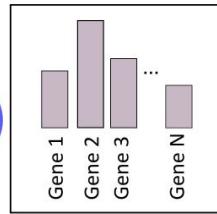
It's as easy as 1-2-3

2. Use inferred parameters to find optimal combination **combination** of cell types in spot



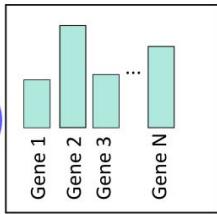
=

W_1



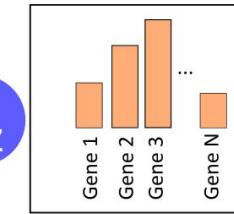
+

W_2



+ ... +

W_Z



Spot 1

Type 1

Type 2

Type Z



It's as easy as 1-2-3

3. Map cell type proportions back onto the tissue



■ ■ | Slightly more complex than adding bar graphs..

- Probabilistic Model



Slightly more complex than adding bar graphs..

- Probabilistic Model

Observed counts for gene g at spot s

$$x_{gs} = \sum_{c \in C_s} x_{gsc},$$

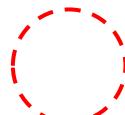
$$x_{gsc} \sim NB(\beta_g \alpha_s r_{gz_c}, p_g)$$

Gene Bias

Scaling factor

Cell type
specific Rate

Success
probability



= Learn from Single Cell Data

$$y_{gc} \sim \mathcal{NB}(s_c r_{gz_c}, p_g)$$

Slightly more complex than adding bar graphs..

- Probabilistic Model

$$x_{gs} \sim NB\left(\alpha_s \sum_{c \in C_s} \beta_g r_{gz_c}, p_g\right)$$



Change Index of summation

$$x_{gs} \sim NB\left(\alpha_s \sum_{z \in Z} \beta_g n_{sz} r_{gz}, p_g\right)$$



Number of cells from type z at spot s

Slightly more complex than adding bar graphs..

- Probabilistic Model

$$x_{gs} \sim NB\left(\alpha_s \sum_{z \in Z} \beta_g n_{sz} r_{gz}, p_g\right)$$



join scaling factor
and cell counts

$$v_{sz} = \alpha_s n_{sz}$$

$$x_{gs} \sim NB\left(\sum_{z \in Z} \beta_g v_{sz} r_{gz}, p_g\right)$$

■ ■ | Slightly more complex than adding bar graphs..

- Probabilistic Model

$$w_{sz} = \frac{v_{sz}}{\sum_{z \in Z} v_{sz}} = \frac{\alpha_s n_{sz}}{\alpha_s \sum_{z \in Z} n_{sz}} = \frac{n_{sz}}{\sum_{z \in Z} n_{sz}}$$

Slightly more complex than adding bar graphs..

- Probabilistic Model

$$w_{sz} = \frac{v_{sz}}{\sum_{z \in Z} v_{sz}} = \frac{\alpha_s n_{sz}}{\alpha_s \sum_{z \in Z} n_{sz}} = \frac{n_{sz}}{\sum_{z \in Z} n_{sz}}$$

Number of cells from cell type z at spot s

n_{sz}

Total number of cells at spot s

Slightly more complex than adding bar graphs..

- Probabilistic Model

$$w_{sz} = \frac{v_{sz}}{\sum_{z \in Z} v_{sz}} = \frac{\alpha_s n_{sz}}{\alpha_s \sum_{z \in Z} n_{sz}} = \frac{n_{sz}}{\sum_{z \in Z} n_{sz}}$$

Number of cells from cell type z at spot s

Proportion of cell type “ z ” at spot “ s ”

Total number of cells at spot s

The diagram illustrates the components of the equation. A red dashed box encloses the term v_{sz} , which is labeled "Proportion of cell type ‘z’ at spot ‘s’". A green dashed box encloses the denominator $\sum_{z \in Z} n_{sz}$, which is labeled "Total number of cells at spot s". An arrow points from the text "Number of cells from cell type z at spot s" to the term n_{sz} in the final fraction.

Slightly more complex than adding bar graphs..

- Probabilistic Model
 - Use MLE estimate to find unadjusted proportions ($V = [v_{sz}]$)
 - Minimize :
$$l(\mathbf{V}, \beta) = -\log[L(\mathbf{V}, \beta | \mathbf{r}, \mathbf{p}, \mathbf{X})]$$
 Likelihood function
 - Stochastic optimization (PyTorch)

Slightly more complex than adding bar graphs..

- Probabilistic Model
 - Assumes single cell and spatial data are both NB distributed
- Tool : *stereoscope*
- Output : [spot] x [cell_type] matrix
 - Elements are proportion of cell belonging to the given cell type at each spot

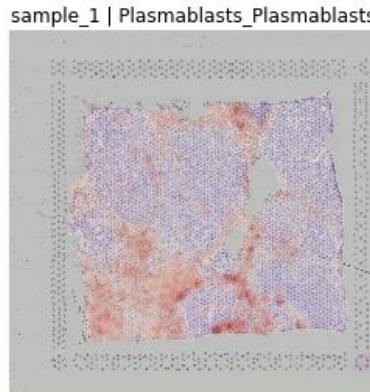
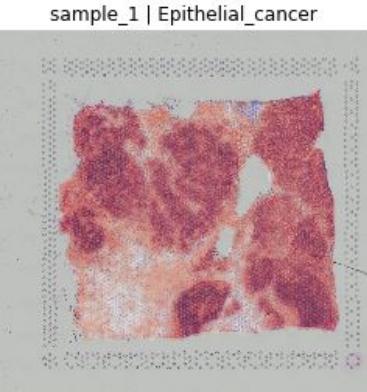
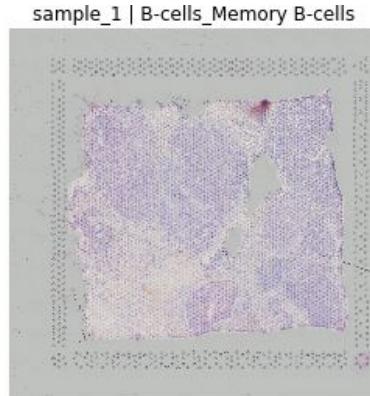
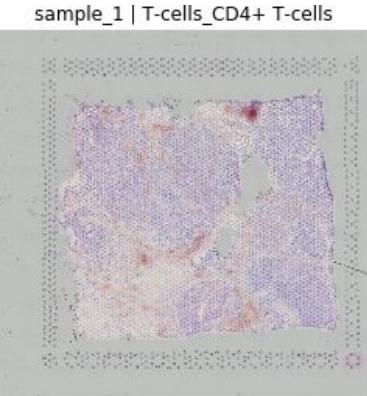


<https://github.com/almaan/stereoscope>

#shameless self-advertising

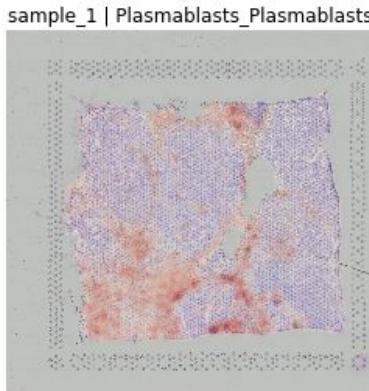
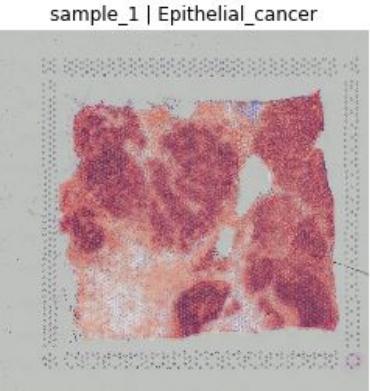
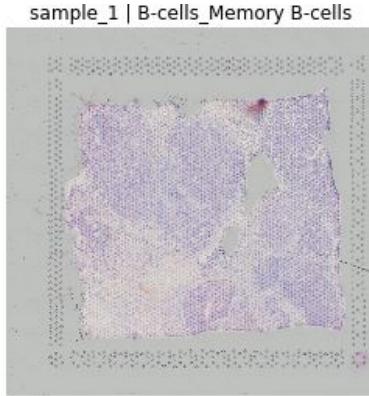
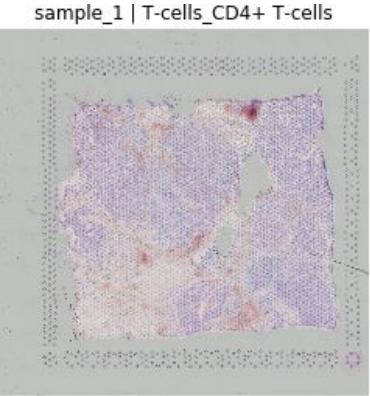
Applying it to our breast cancer data

Proportion estimates

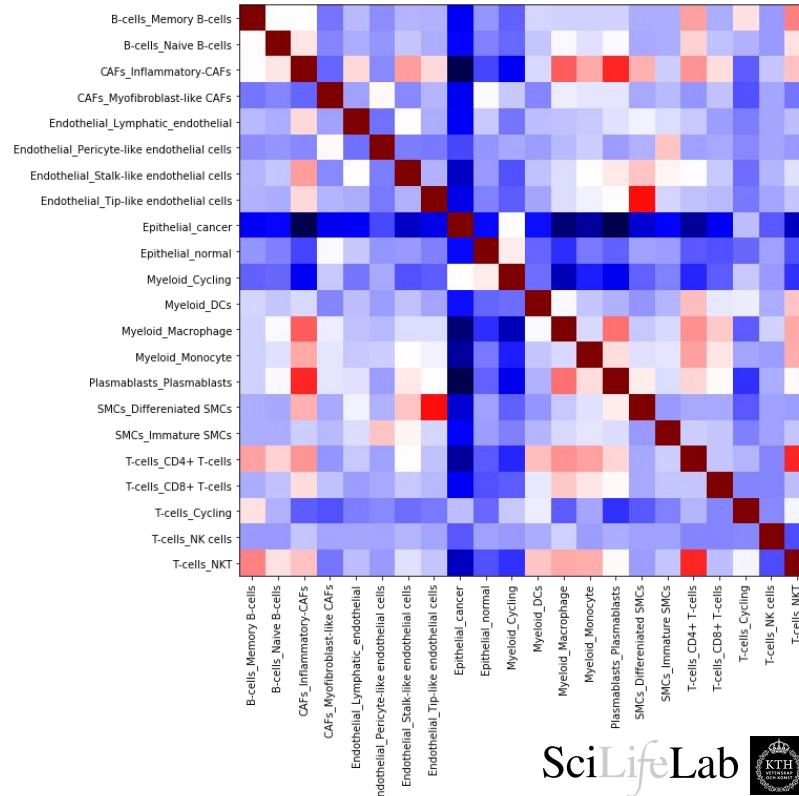


Applying it to our breast cancer data

Proportion estimates



Cell type co-localization



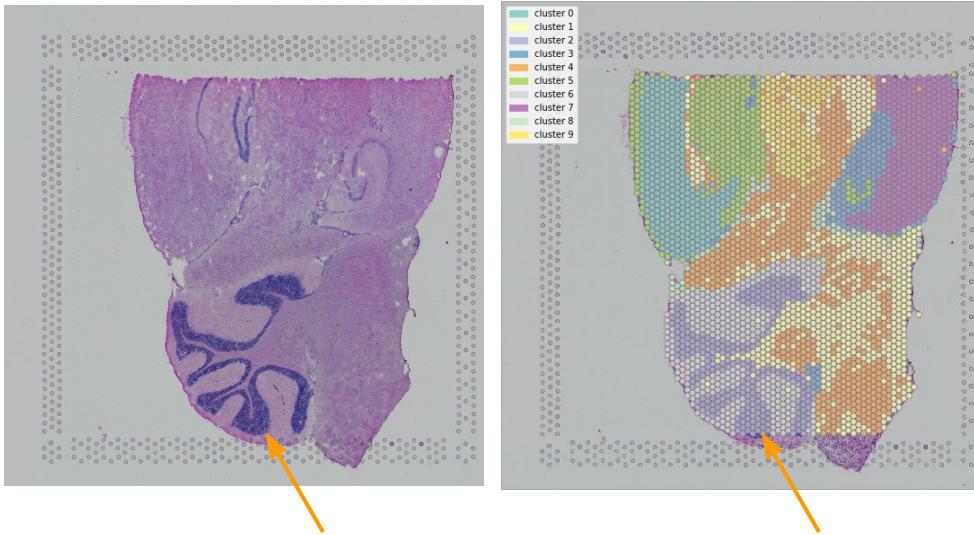


Summary : Integration of single cell and spatial data

- Leverages strengths from respective technique
 - Spatial resolution of well defined cell types
- Can be used as basis in subsequent analyses
 - Patterns of cell type co-localization
- Solution until experimental techniques reach single cell res.
- Atlases are exciting!

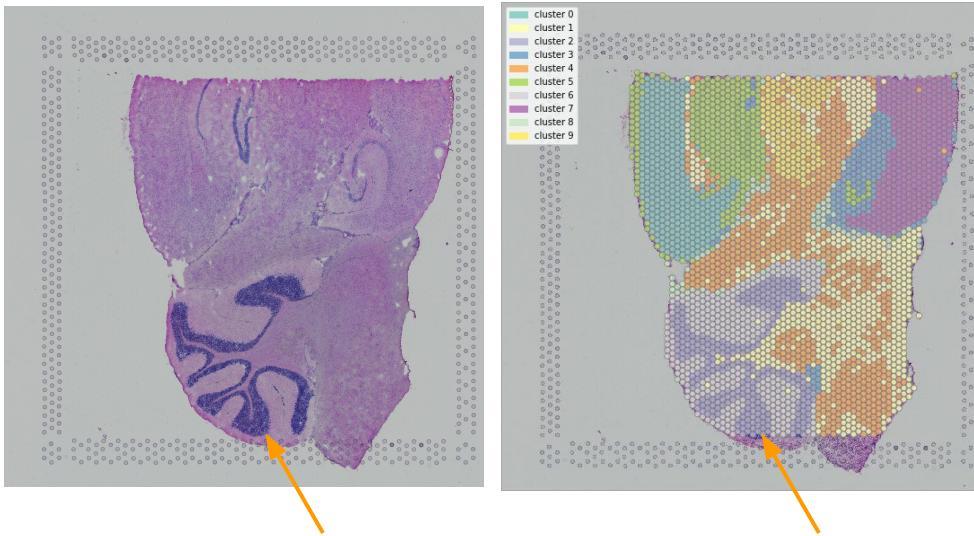
Example Analysis : Expression as a function of distance

- Say we cluster spatial data and find an interesting domain (e.g. cluster 2)



Example Analysis : Expression as a function of distance

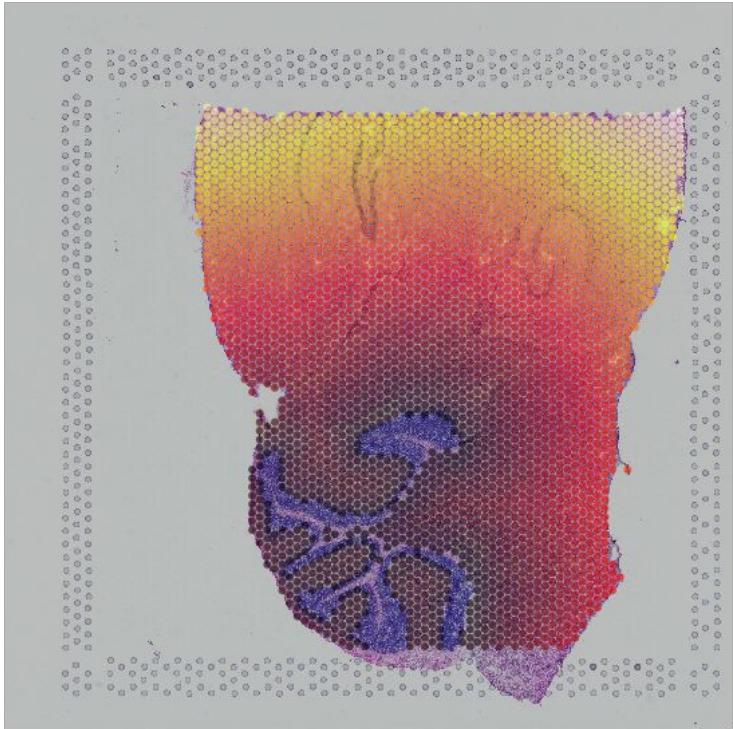
- Say we cluster spatial data and find an interesting domain (e.g. cluster 2)



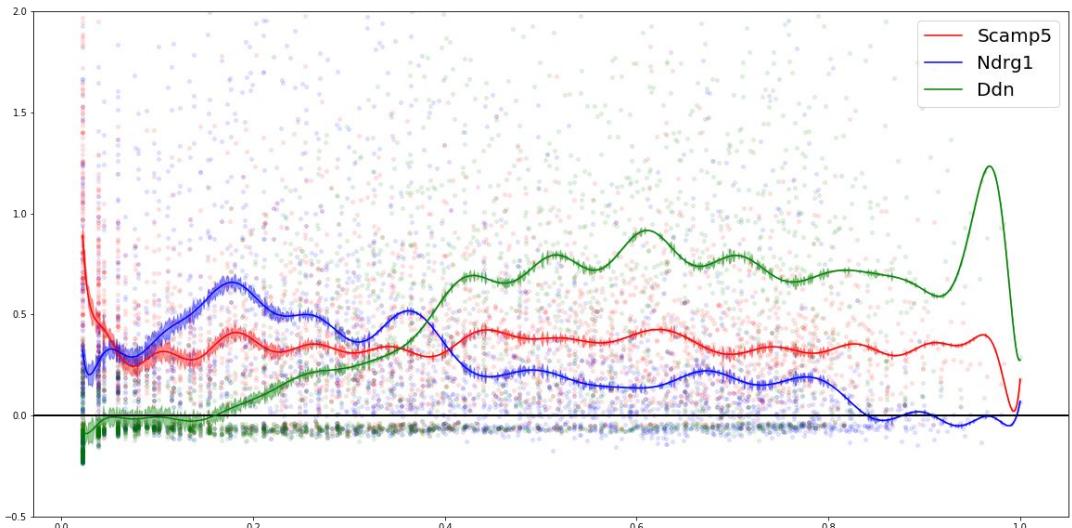
- We may then ask how gene expression changes with the distance to this cluster



Example Analysis : Expression as a function of distance

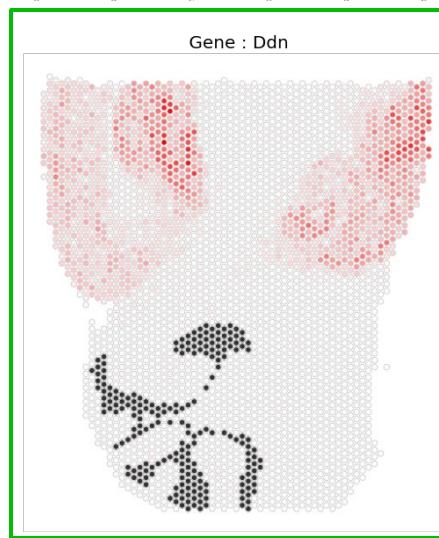
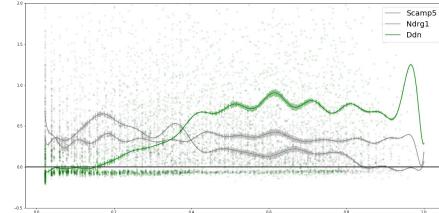
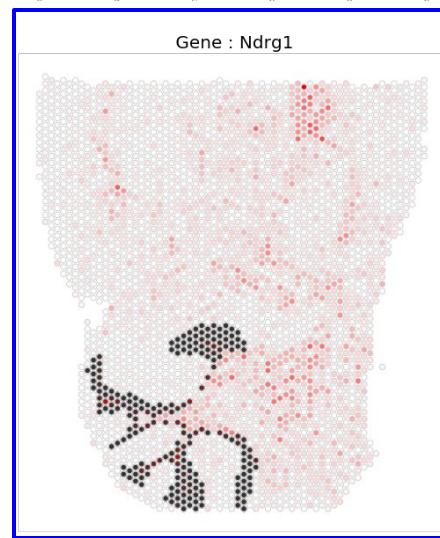
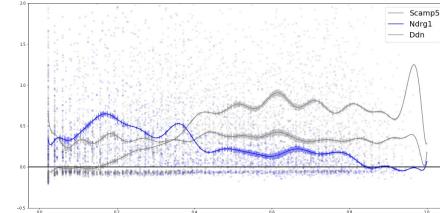
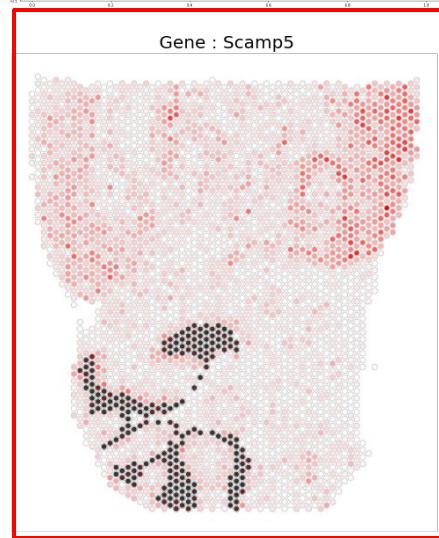
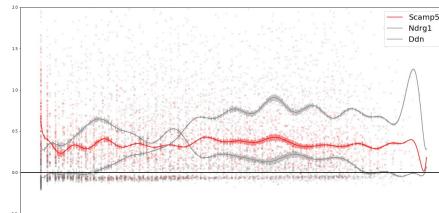


Spots colored by distance do cluster 2



Gene expression as a function of the distance
to cluster 2

Example Analysis : Expression as a function of distance



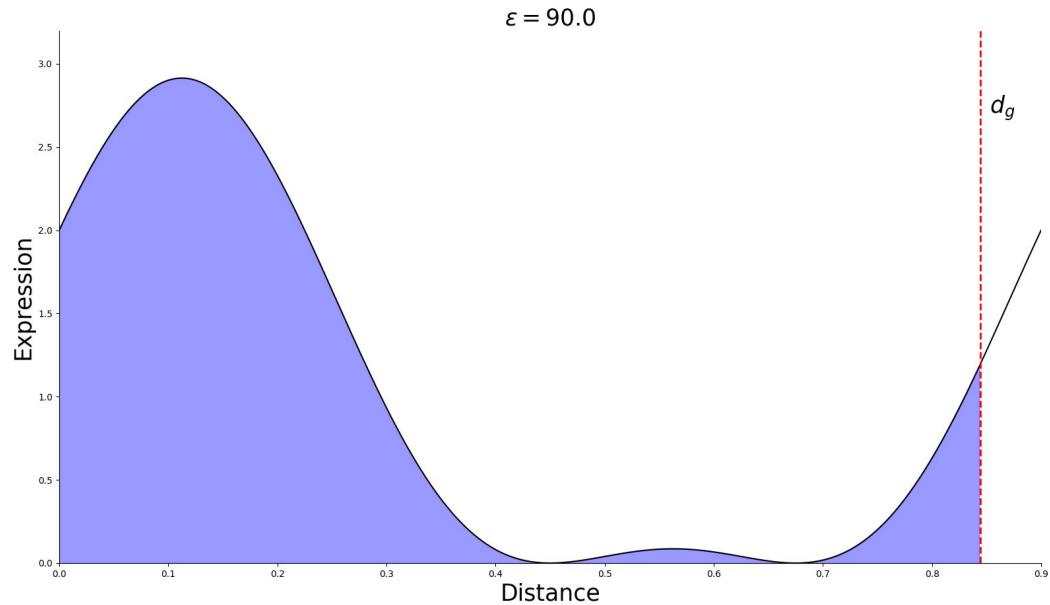
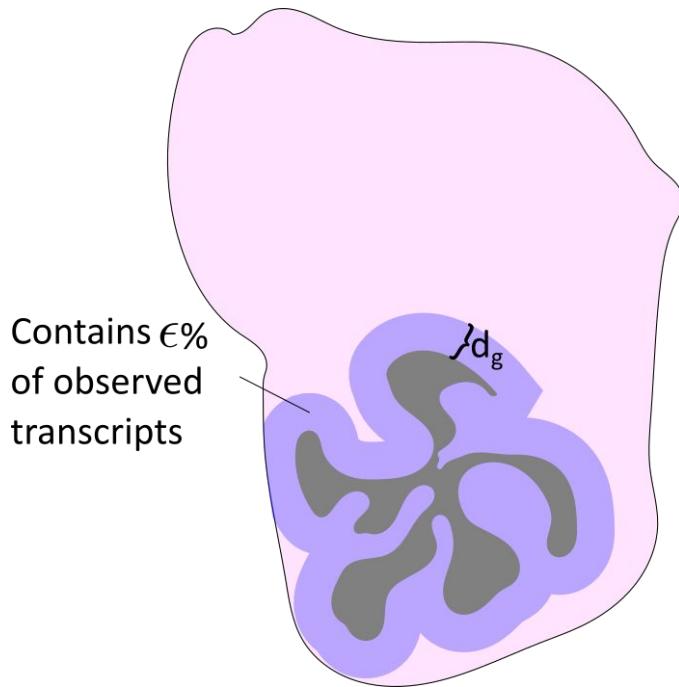
■ ■ ■ Example Analysis : Expression as a function of distance

- Can also ask : “within which distance (d_g) from cluster 2 is ε % of all transcripts from gene g contained?”

$$\varepsilon = 100 \times \frac{\int_0^{d_g} f(x)dx}{\int_0^1 f(x)dx}$$

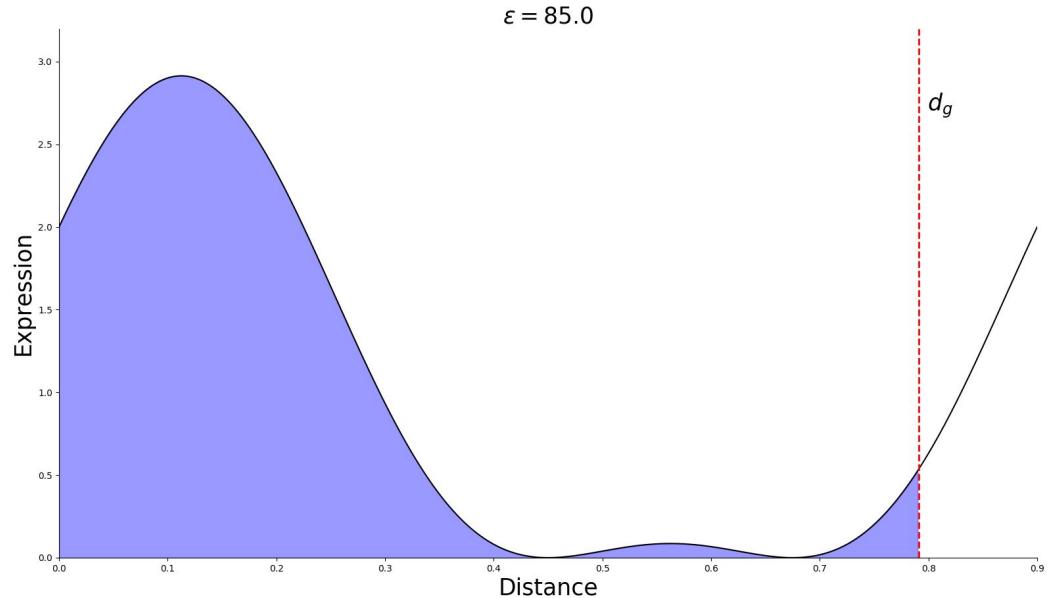
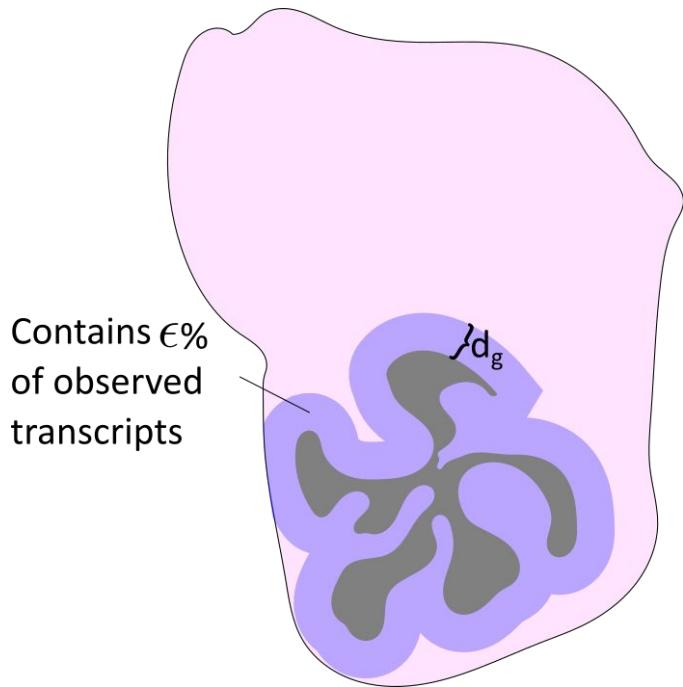
Example Analysis : Expression as a function of distance

- Can also ask : “within which distance (d_g) from cluster 2 is ϵ % of all transcripts from gene g contained?”



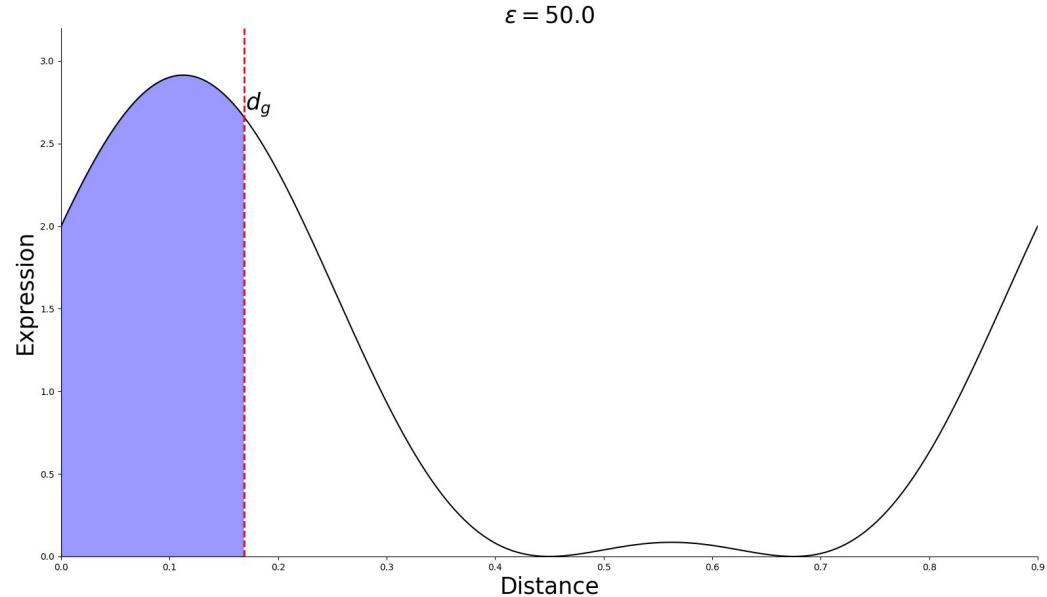
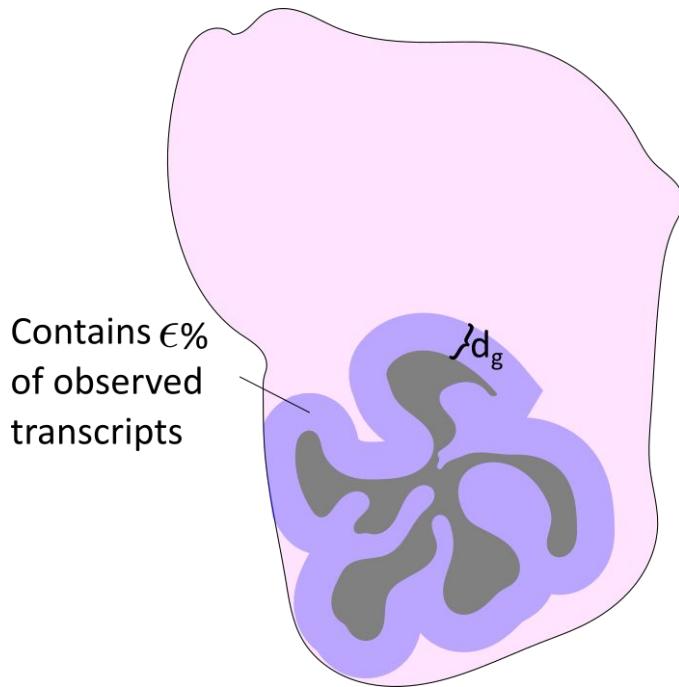
Example Analysis : Expression as a function of distance

- Can also ask : “within which distance (d_g) from cluster 2 is ϵ % of all transcripts from gene g contained?”



Example Analysis : Expression as a function of distance

- Can also ask : “within which distance (d_g) from cluster 2 is ϵ % of all transcripts from gene g contained?”

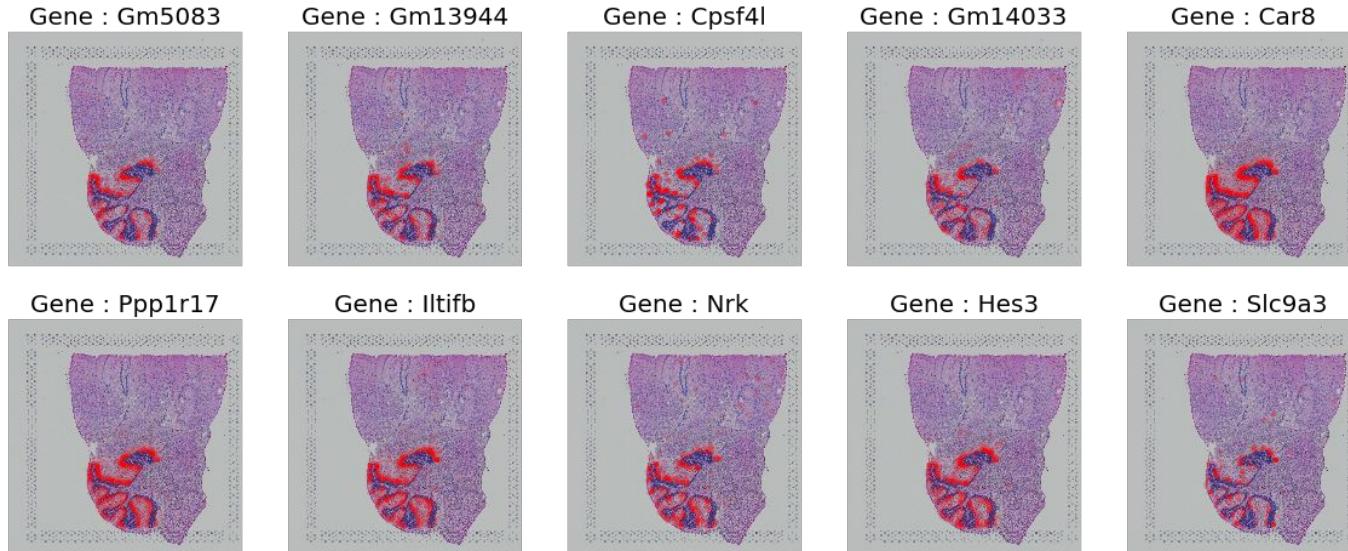


■ ■ ■ Example Analysis : Expression as a function of distance

- Alternatively : “which genes has $\varepsilon\%$ within the shortest distance (d_g) from cluster 2?”

Example Analysis : Expression as a function of distance

- Alternatively : “which genes has $\epsilon\%$ within the shortest distance (d_g) from cluster 2?”



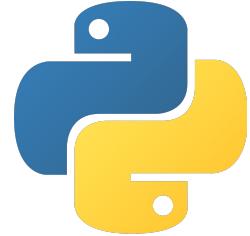


Exercises



Exercise Session

- Aims:
 - Getting familiar with spatial data
 - Overview - concept focused
- Written to be “independent” of lectures
 - Might experience some redundancy
 - but... *repetitio est mater studiorum*
- Three Parts
 - Part 1 - “*Getting Comfy with Spatial Data*”
 - Orienting, Inspecting and visualizing spatial data
 - Basic analysis workflow
 - Part 2 - “*Integrating Single Cell and Spatial RNA-Seq Data*”
 - Working with mapped data
 - Downstream analysis
 - Part 3 - “*Digging deeper into spatial analysis*”
 - Spatial gene set enrichment
 - Expression as a function of distance





Thank you for the attention

■ ■ || Questions