Exploring Spatial Transcriptomics

2020-05-28

Alma Andersson

https://github.com/almaan

SPATIAL research

https://www.spatialresearch.org

SciLifeLab

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

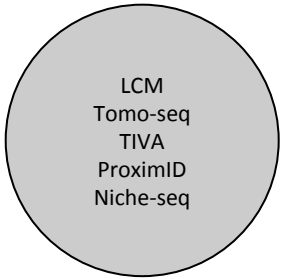
- Material in Part 3

Background

The spatial space | Overview of techniques

The spatial space | Overview of techniques

Microdissection

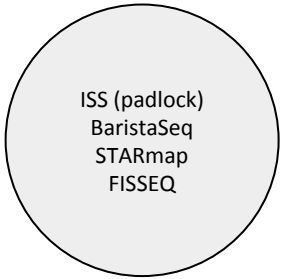


Isolate a region of interest, place isolate in separate well and sequence (either by bulk or single-cell methods).

“Brute Force” approach.

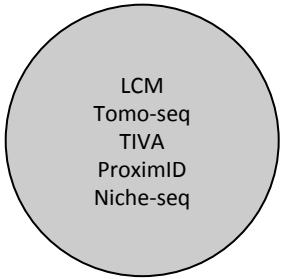
The spatial space | Overview of techniques

In situ sequencing



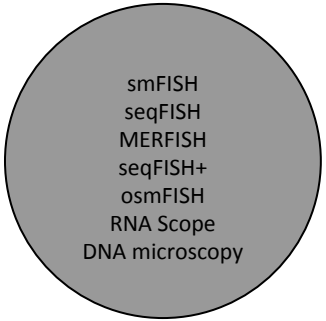
Sequence the transcripts in place. Offer sub-cellular resolution. Tend to rely on gene panels. Need "*a priori*" defined targets.

Microdissection

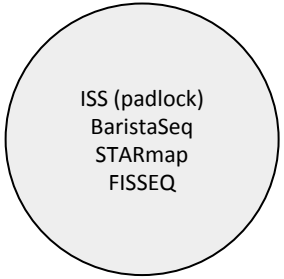


The spatial space | Overview of techniques

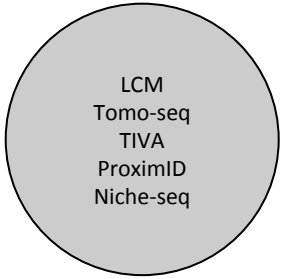
In situ hybridization



In situ sequencing



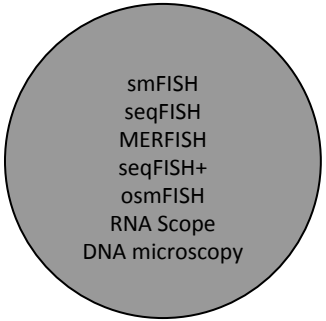
Microdissection



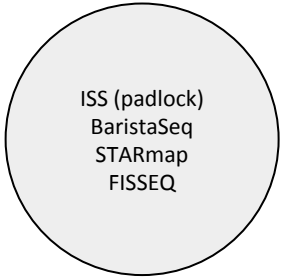
Labeled probes for specific targets, hybridize in place and visualized for spatial information.

The spatial space | Overview of techniques

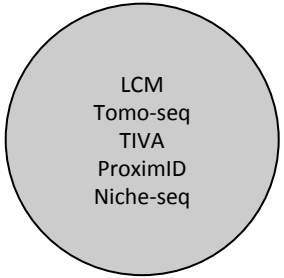
In situ hybridization



In situ sequencing

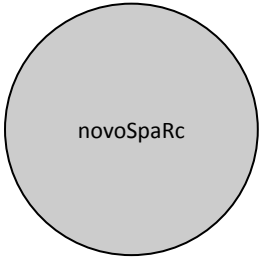


Microdissection

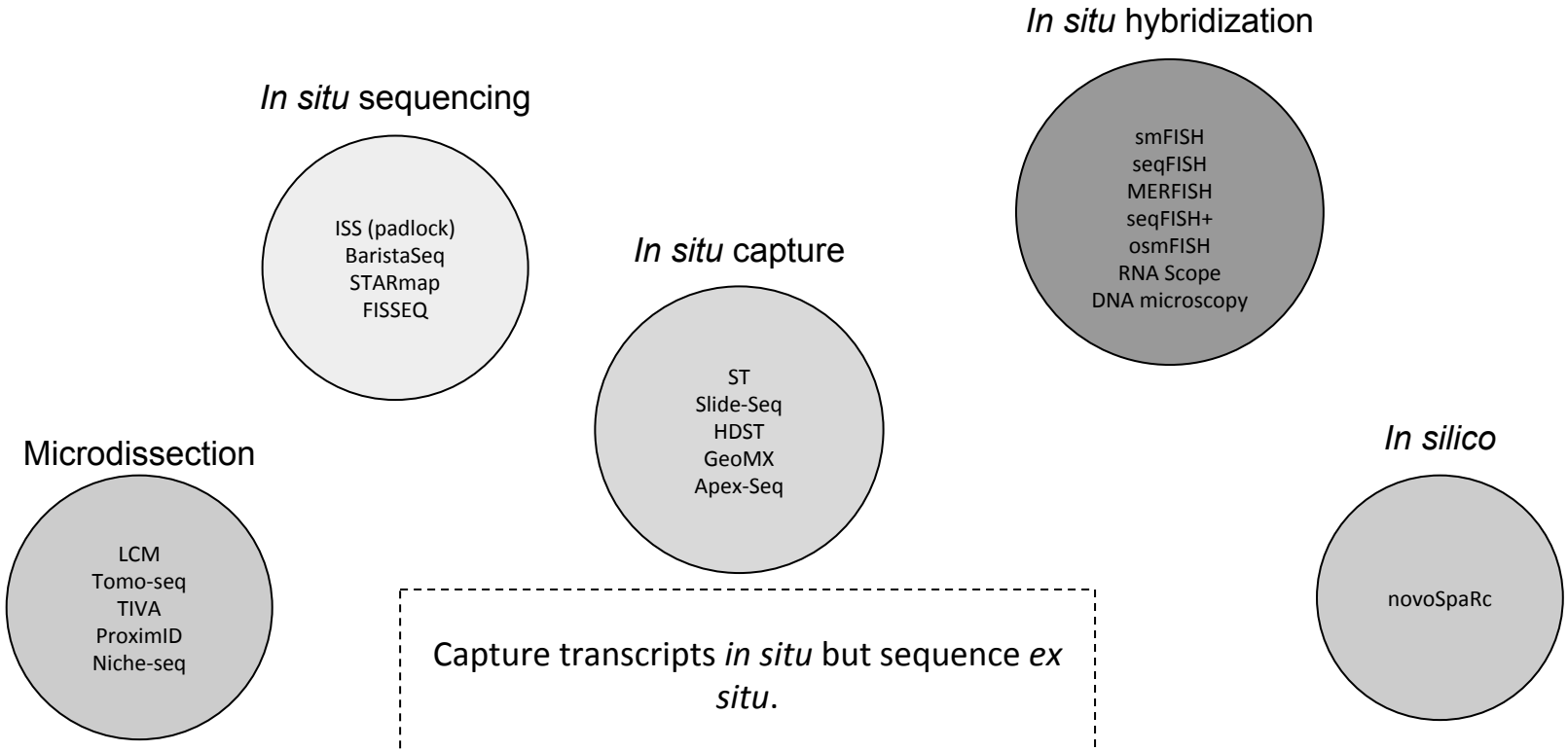


Infer and reconstruct spatial structure from non-spatial data (e.g., single cell).

In silico

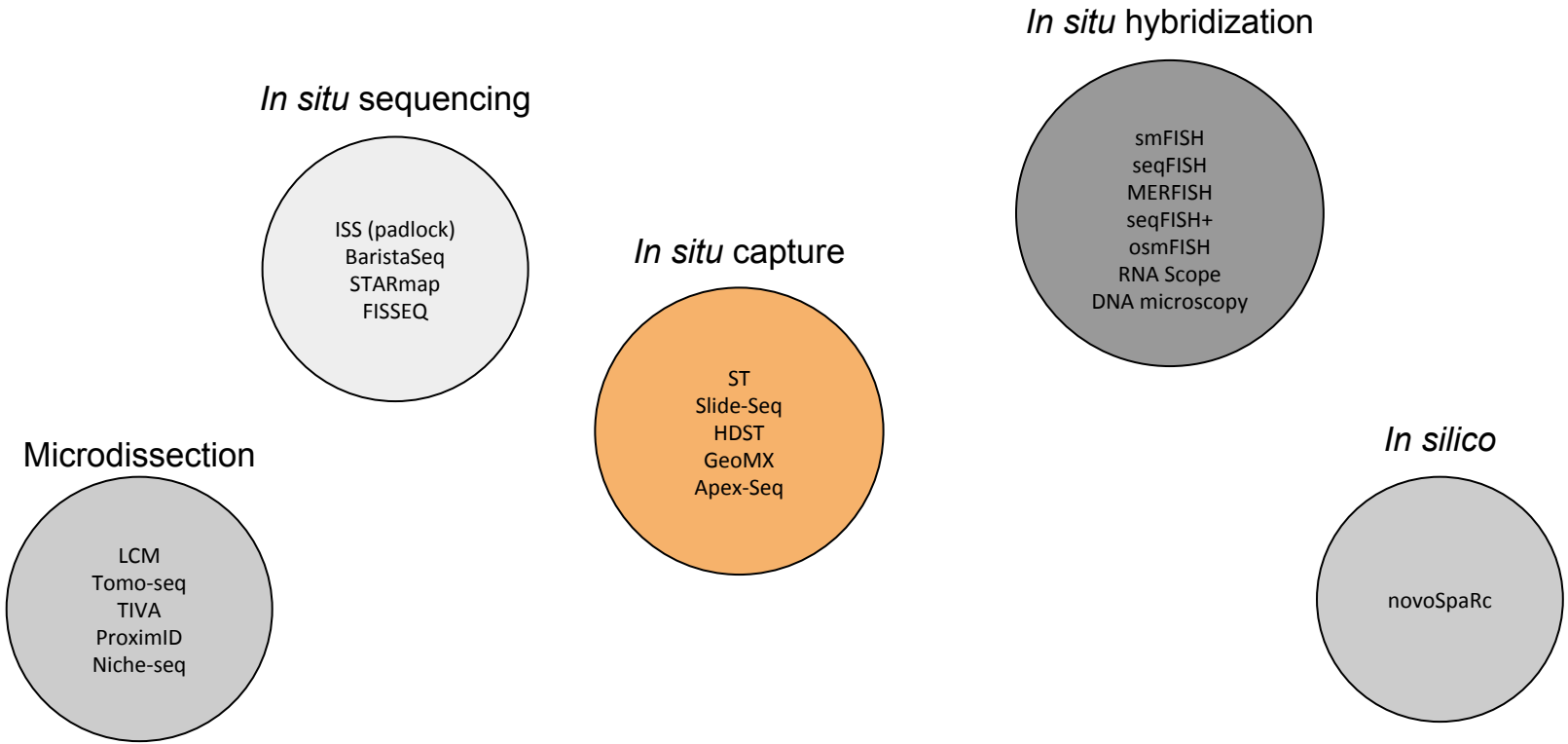


The spatial space | Overview of techniques



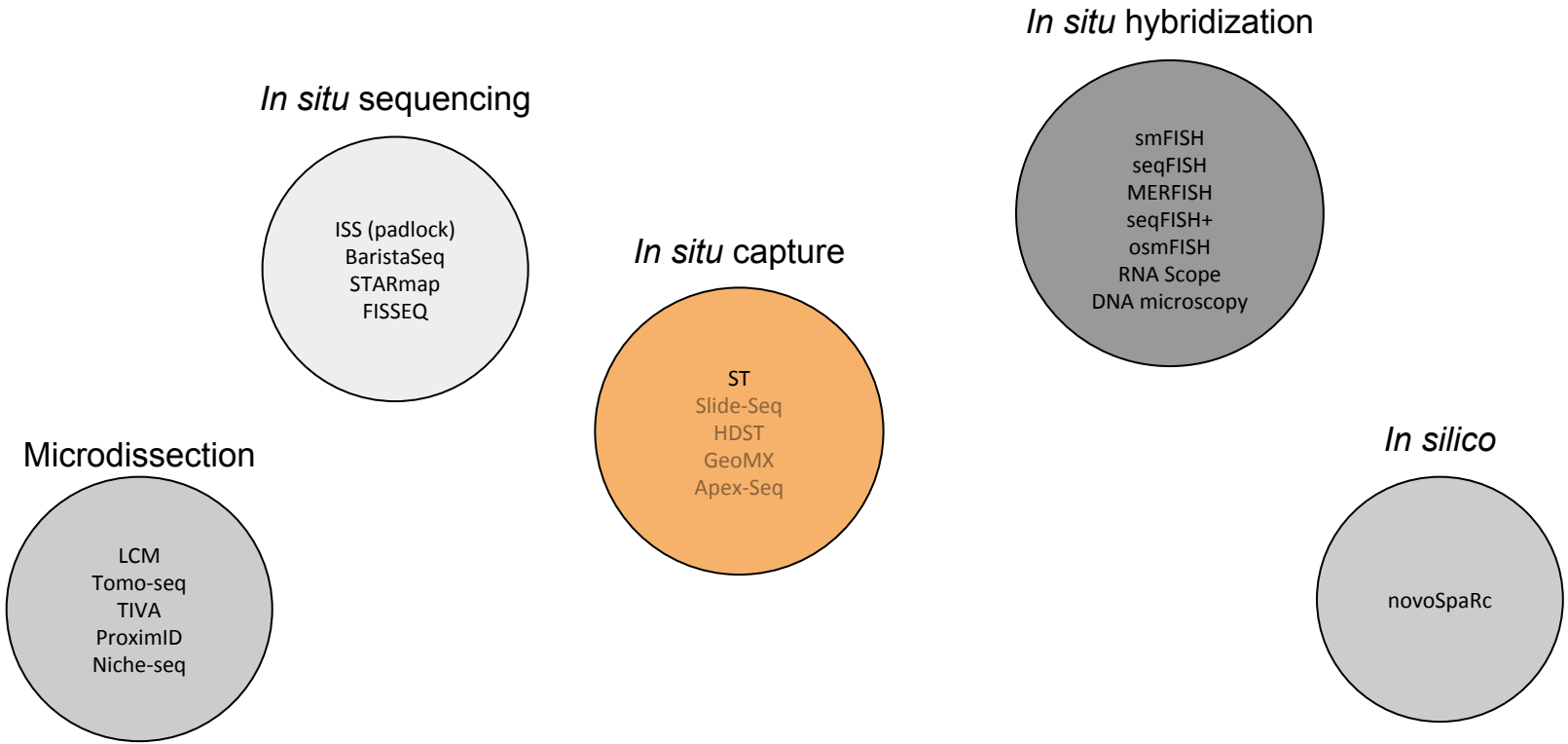
Credit to J.Bergenstråhle and M.Asp for categorization of techniques

The spatial space | Overview of techniques



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

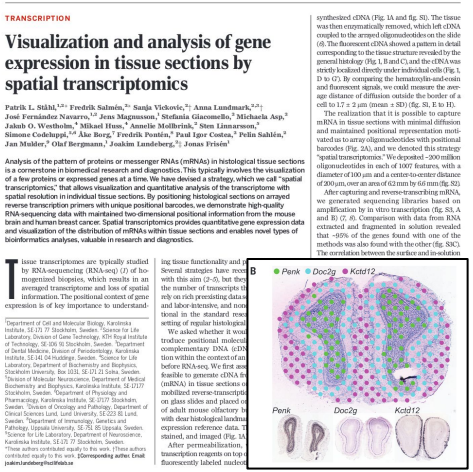
The spatial space | Overview of techniques



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

Patrik L. Ståhl,[1,2,*] Fredrik Salmén,[2,*] Sanja Vickovic,[2,*] Anna Lundmark,[2,*†] José Fernández Navarro,[1,2] Jens Magnusson,[2] Stefania Giacomello,[2] Michaela Asp,[2] Jakub O. Westholm,[3] Mikael Huss,[4] Annelie Mollbrink,[2] Sten Linnarsson,[3] Simone Codeluppi,[2,5] Åke Borg,[6] Fredrik Pontén,[4] Paul Igor Costea,[2] Pelin Sahlén,[2] Jan Mulder,[4] Olaf Bergmann,[1] Joakim Lundeberg,[2†] Jonas Frisén[1]

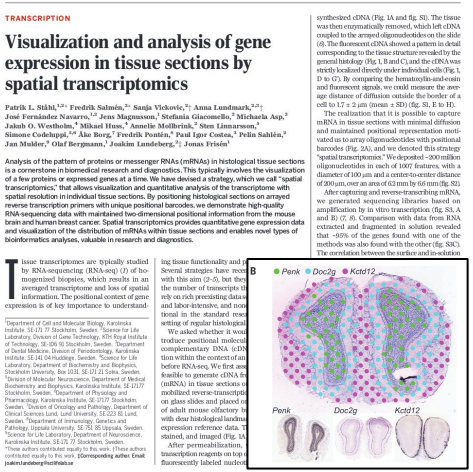
Analysis of the pattern of proteins or messenger RNAs (mRNAs) in histological tissue sections is a cornerstone in biomedical research and diagnostics. This typically involves the visualization of a few proteins or expressed genes at a time. We have devised a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue sections. By positioning histological sections on arrayed reverse transcription primers with unique positional barcodes, we demonstrate high-quality RNA-sequencing data with maintained two-dimensional positional information from the mouse brain and human breast cancer. Spatial transcriptomics provides quantitative gene expression data and visualization of the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics.

Science Publication
Ståhl et.al
DOI: 10.1126/science.aaf2403

SciLifeLab

Spatial Transcriptomics (ST)

Mid 2016



TRANSCRIPTION

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

Patrik L. Ståhl,[1,2,†] Fredrik Salmén,[2,†] Sanja Vickovic,[2,†] Anna Lundmark,[3,4,†] José Fernández Navarro,[1,5] Jens Magnusson,[1] Stefania Giacomello,[3] Michaela Asp,[2] Jakub O. Westholm,[3] Mikael Huss,[4] Annelie Mollbrink,[2] Sten Linnarsson,[3] Simone Codeluppi,[3,6] Åke Borg,[7] Fredrik Pontén,[8] Paul Igor Costea,[3] Pelin Sahlén,[2] Jan Mulder,[9] Olaf Bergmann,[1] Joakim Lundeberg,[2,‡] Jonas Frisén[1]

Analysis of the pattern of proteins or messenger RNAs (mRNAs) in histological tissue sections is a cornerstone in biomedical research and diagnostics. This typically involves the visualization of a few proteins or expressed genes at a time. We have devised a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue sections. By positioning histological sections on arrayed reverse transcription primers with unique positional barcodes, we demonstrate high-quality RNA-sequencing data with maintained two-dimensional positional information from the mouse brain and human breast cancer. Spatial transcriptomics provides quantitative gene expression data and visualization of the distribution of mRNAs within tissue sections and enables novel types of bioinformatics analyses, valuable in research and diagnostics.

Late 2018

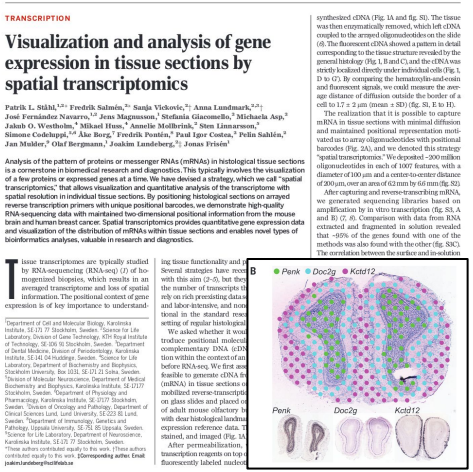
10x Genomics™ acquisition

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 Salmén,[2,*] Sanja Vickovic,[2,*] Anna Lundmark,[3,4,†] José Fernández Navarro,[1,5] Jens Magnusson,[1] Stefania Giacomello,[6] Michaela Asp,[2] Jakub O. Westholm,[3] Mikael Huss,[4] Annelie Mollbrink,[2] Sten Linnarsson,[3] Simone Codeluppi,[3,4] Åke Borg,[7] Fredrik Pontén,[8] Paul Igor Costea,[9] Pelin Sahlén,[2] Jan Mulder,[6] Olaf Bergmann,[1] Joakim Lundeberg,[2,‡] Jonas Frisén[1]

Analysis of the pattern of proteins or messenger RNAs (mRNAs) in histological tissue sections is a cornerstone in biomedical research and diagnostics. This typically involves the visualization of a few proteins or expressed genes at a time. We have devised a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue sections. By positioning histological sections on arrayed reverse transcription primers with unique positional barcodes, we demonstrate high-quality RNA-sequencing data with maintained two-dimensional positional information from the mouse brain and human breast cancer. Spatial transcriptomics provides quantitative gene expression data and visualization of 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) (*1*) of homogenized biopsies, which results in an averaged transcriptome and loss of spatial information. The positional context of gene expression is of key importance to understand-

[1]Department of Cell and Molecular Biology, Karolinska Institute, SE-171 77 Stockholm, Sweden. [2]Science for Life Laboratory, Division of Gene Technology, KTH Royal Institute of Technology, SE-106 91 Stockholm, Sweden. [3]Department of Dental Medicine, Division of Periodontology, Karolinska Institute, SE-141 04 Huddinge, Sweden. [4]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Box 1031, SE-171 21 Solna, Sweden. [5]Division of Molecular Neuroscience, Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-17177 Stockholm, Sweden. [6]Department of Physiology and Pharmacology, Karolinska Institute, SE-17177 Stockholm, Sweden. [7]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, SE-223 81 Lund, Sweden. [8]Department of Immunology, Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden. [9]Science for Life Laboratory, Department of Neuroscience, Karolinska Institute, SE-171 77 Stockholm, Sweden.
*These authors contributed equally to this work. †These authors contributed equally to this work. ‡Corresponding author. Email: joakim.lundeberg@scilifelab.se

synthesized cDNA (Fig. 1A and fig. S1). The tissue was then enzymatically removed, which left cDNA coupled to the arrayed oligonucleotides on the slide (*6*). The fluorescent cDNA showed a pattern in detail corresponding to the tissue structure revealed by the general histology (Fig. 1, B and C), and the cDNA was strictly localized directly under individual cells (Fig. 1, D to G'). By comparing the hematoxylin-and-eosin and fluorescent signals, we could measure the average distance of diffusion outside the border of a cell to 1.7 ± 2 μm (mean ± SD) (fig. S1, E to H).

The realization that it is possible to capture mRNA in tissue sections with minimal diffusion and maintained positional representation motivated us to array oligonucleotides with positional barcodes (Fig. 2A), and we denoted this strategy "spatial transcriptomics." We deposited ~200 million oligonucleotides in each of 1007 features, with a diameter of 100 μm and a center-to-center distance of 200 μm, over an area of 62 mm by 65 mm (fig. S2).

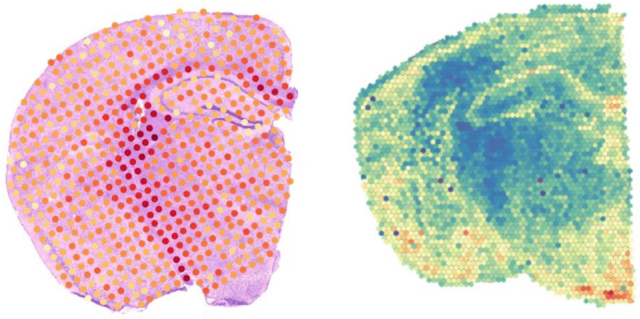
After capturing and reverse-transcribing mRNA, we generated sequencing libraries based on amplification by in vitro transcription (fig. S3, A and B) (*7*, *8*). Comparison with data from RNA extracted and fragmented in solution revealed that ~95% of the genes found with one of the methods was also found with the other (fig. S3C). The correlation between the surface and in-solution

Science Publication
Ståhl et.al
DOI: 10.1126/science.aaf2403

Late 2018

10x Genomics™ acquisition

Late 2019

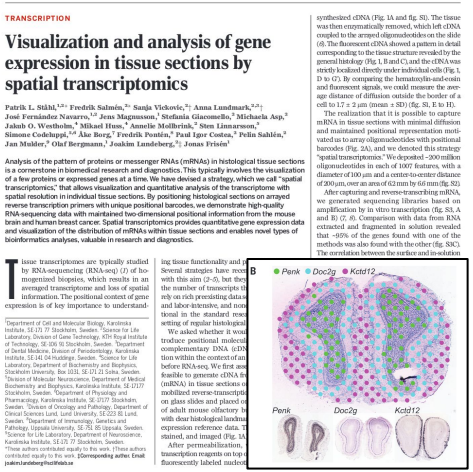


Launch of **Visium** Spatial Gene Expression Platform

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 Salmén,[2,*] Sanja Vickovic,[2,*] Anna Lundmark,[3,4,†] José Fernández Navarro,[1,5] Jens Magnusson,[1] Stefania Giacomello,[3] Michaela Asp,[2] Jakub O. Westholm,[3] Mikael Huss,[4] Annelie Mollbrink,[2] Sten Linnarsson,[3] Simone Codeluppi,[3,6] Åke Borg,[7] Fredrik Pontén,[8] Paul Igor Costea,[3] Pelin Sahlén,[2] Jan Mulder,[9] Olaf Bergmann,[1] Joakim Lundeberg,[2,‡] Jonas Frisén[1]

Analysis of the pattern of proteins or messenger RNAs (mRNAs) in histological tissue sections is a cornerstone in biomedical research and diagnostics. This typically involves the visualization of a few proteins or expressed genes at a time. We have devised a strategy, which we call "spatial transcriptomics," that allows visualization and quantitative analysis of the transcriptome with spatial resolution in individual tissue sections. By positioning histological sections on arrayed reverse transcription primers with unique positional barcodes, we demonstrate high-quality RNA-sequencing data with maintained two-dimensional positional information from the mouse brain and human breast cancer. Spatial transcriptomics provides quantitative gene expression data and visualization of 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) (*1*) of homogenized biopsies, which results in an averaged transcriptome and loss of spatial information. The positional context of gene expression is of key importance to understand-

[1]Department of Cell and Molecular Biology, Karolinska Institute, SE-171 77 Stockholm, Sweden. [2]Science for Life Laboratory, Division of Gene Technology, KTH Royal Institute of Technology, SE-106 91 Stockholm, Sweden. [3]Department of Dental Medicine, Division of Periodontology, Karolinska Institute, SE-141 04 Huddinge, Sweden. [4]Science for Life Laboratory, Department of Biochemistry and Biophysics, Stockholm University, Box 1031, SE-171 21 Solna, Sweden. [5]Division of Molecular Neuroscience, Department of Medical Biochemistry and Biophysics, Karolinska Institute, SE-17177 Stockholm, Sweden. [6]Department of Physiology and Pharmacology, Karolinska Institute, SE-17177 Stockholm, Sweden. [7]Division of Oncology and Pathology, Department of Clinical Sciences Lund, Lund University, SE-223 81 Lund, Sweden. [8]Department of Immunology, Genetics and Pathology, Uppsala University, SE-751 85 Uppsala, Sweden. [9]Science for Life Laboratory, Department of Neuroscience, Karolinska Institute, SE-171 77 Stockholm, Sweden. *These authors contributed equally to this work. †These authors contributed equally to this work. ‡Corresponding author. Email: joakim.lundeberg@scilifelab.se

synthesized cDNA (Fig. 1A and fig. S1). The tissue was then enzymatically removed, which left cDNA coupled to the arrayed oligonucleotides on the slide (*6*). The fluorescent cDNA showed a pattern in detail corresponding to the tissue structure revealed by the general histology (Fig. 1, B and C), and the cDNA was strictly localized directly under individual cells (Fig. 1, D to G′). By comparing the hematoxylin-and-eosin and fluorescent signals, we could measure the average distance of diffusion outside the border of a cell to 1.7 ± 2 μm (mean ± SD) (fig. S1, E to H).

The realization that it is possible to capture mRNA in tissue sections with minimal diffusion and maintained positional representation motivated us to array oligonucleotides with positional barcodes (Fig. 2A), and we denoted this strategy "spatial transcriptomics." We deposited ~200 million oligonucleotides in each of 1007 features, with a diameter of 100 μm and a center-to-center distance of 200 μm, over an area of 62 mm by 65 mm (fig. S2).

After capturing and reverse-transcribing mRNA, we generated sequencing libraries based on amplification by in vitro transcription (fig. S3, A and B) (*7*, *8*). Comparison with data from RNA extracted and fragmented in solution revealed that ~95% of the genes found with one of the methods was also found with the other (fig. S3C). The correlation between the surface and in-solution

ing tissue functionality and
Several strategies have recei
with this aim (*2*–*5*), but thes
the number of transcripts th
rely on rich preexisting data s
and labor-intensive, and non
tional in the standard reac
setting of regular histological

We asked whether it woul
troduce positional molecul
complementary DNA (cDN
tion within the context of an
before RNA-seq. We first ass
feasible to generate cDNA fr
(mRNA) in tissue sections a
mobilized reverse-transcripti
on glass slides and placed o
of adult mouse olfactory b
with clear histological landma
expression reference data. T
stained, and imaged (Fig. 1A

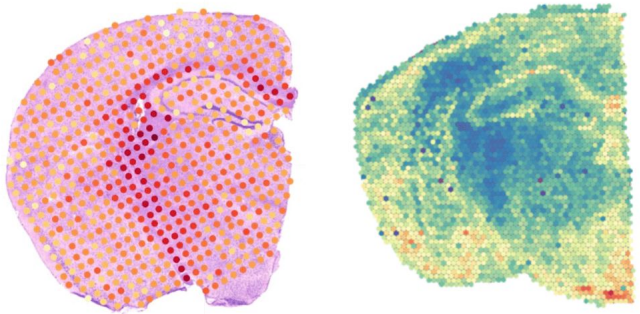
After permeabilization,
transcription reagents on top
fluorescently labeled nucleot

Science Publication
Ståhl et.al
DOI: 10.1126/science.aaf2403

Late 2018

10x Genomics™ acquisition

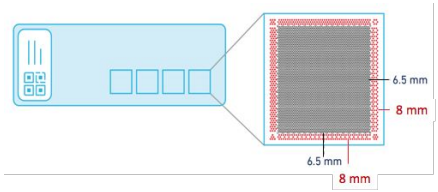
Late 2019

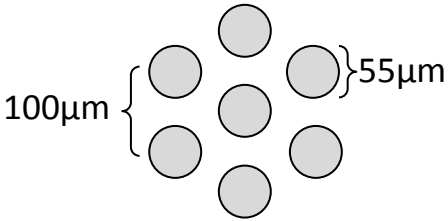


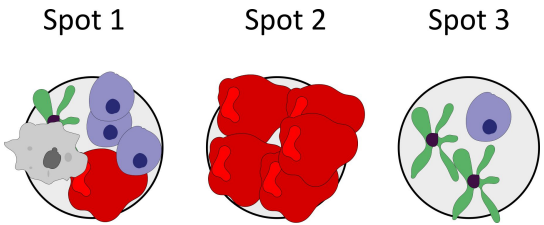
Launch of **Visium** Spatial Gene Expression Platform

Visium Platform

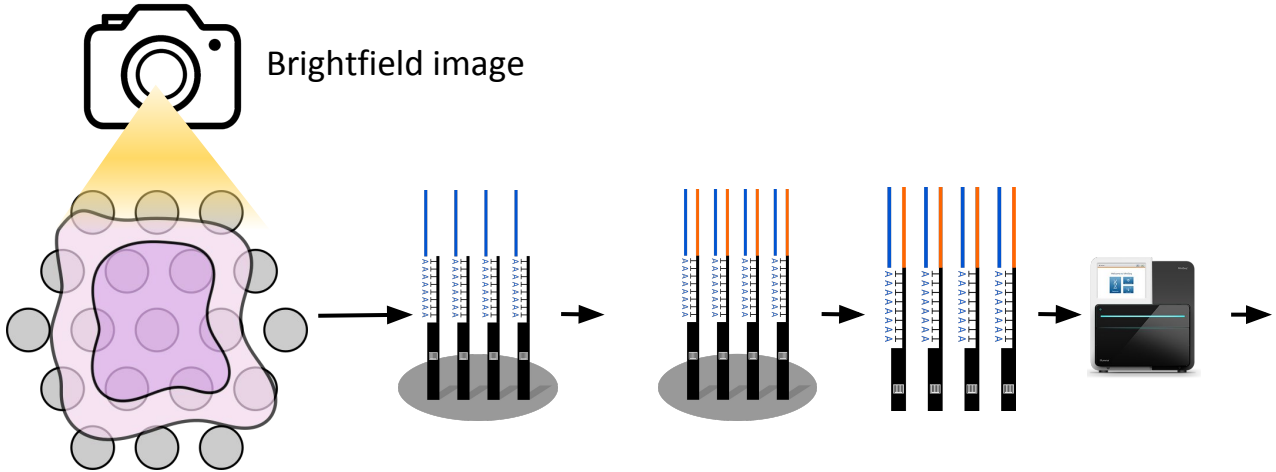
- Array based technique
- 6.5mm x 6.5mm area to put sample on
- 4992 spots arranged in hexgonal grid
- Spot specs:
  - Spot diameter : 55µm
  - Center to center distance : 100 µm
- Each spot has millions of capture probes
  - spatial barcode
  - polyT sequence
  - captures polyadenylated mRNA
  - Full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







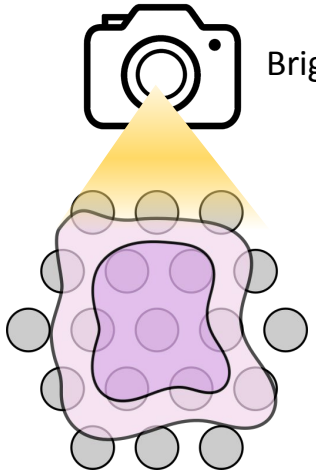
The experimental workflow (in a nutshell)



spot (barcode)	gene_1	...	gene_{G-1}	gene_G
AAACATACAACCAC-1	5	...	0	9
AAACATTGAGCTAC-1	3	...	1	3
⋮	4	⋱	0	0
AAACATTGATCAGC-1	0	...	8	7

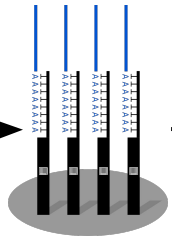
Count Matrix

The experimental workflow (in a nutshell)

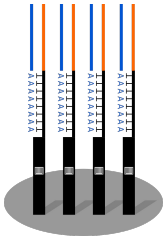


Brightfield image

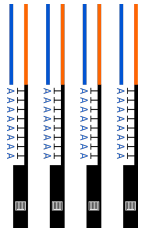
Tissue +
Visium-array



mRNA
capture



cDNA
synthesis



Cleave
Probes

Sequence +
preprocess

Where the fun starts!

spot (barcode)	gene_1	...	gene_{G-1}	gene_G
AAACATACAACCAC-1	5	...	0	9
AAACATTGAGCTAC-1	3	...	1	3
⋮	4	⋱	0	0
AAACATTGATCAGC-1	0	...	8	7

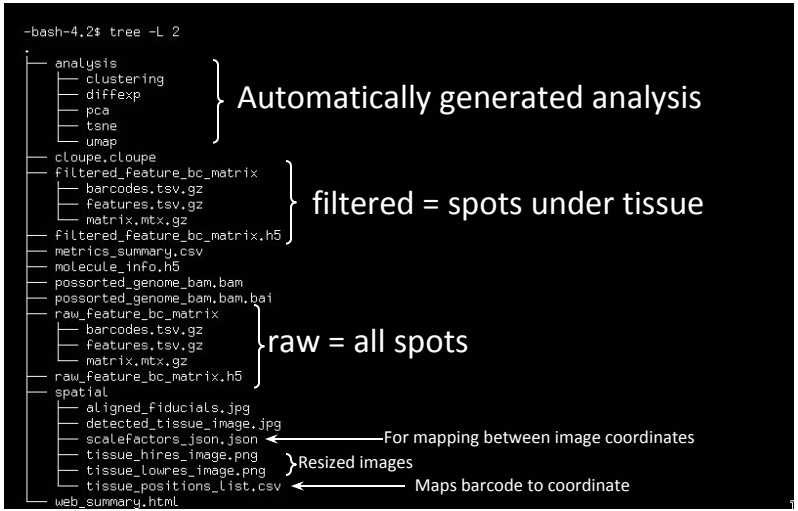
Count Matrix

SciLifeLab KTH

Data Processing

After sequencing (brief)

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

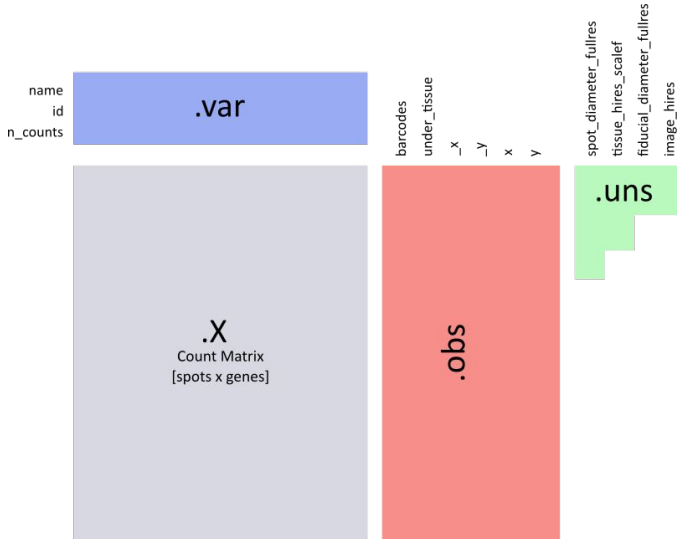


Example of spaceranger count output

Processed data

P1

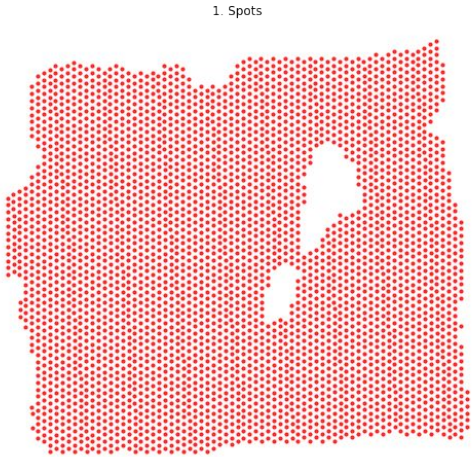
- Either use .mtx files or .h5 files to assemble a data object to work with
  - No standardized format
- Personal preference : convert to .h5ad file (will be using in exercises)
  - scanpy/anndata teams working on - soon to release - their own (similar) format



An initial assessment

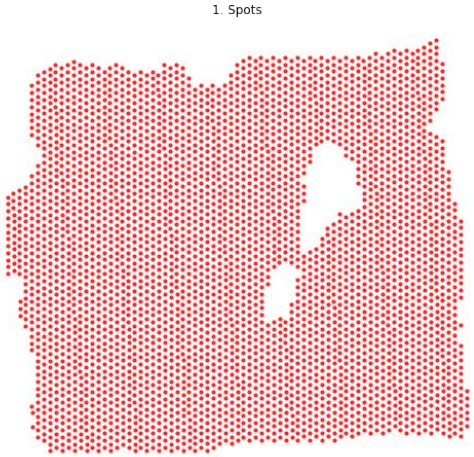
P1

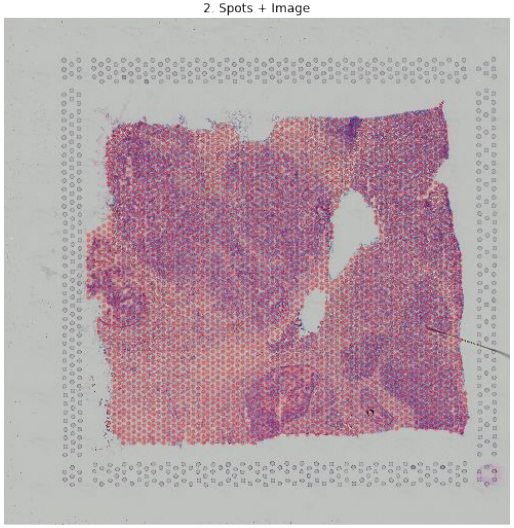
- Example with Human Breast cancer data
  - Public data : Available at 10x website



An initial assessment

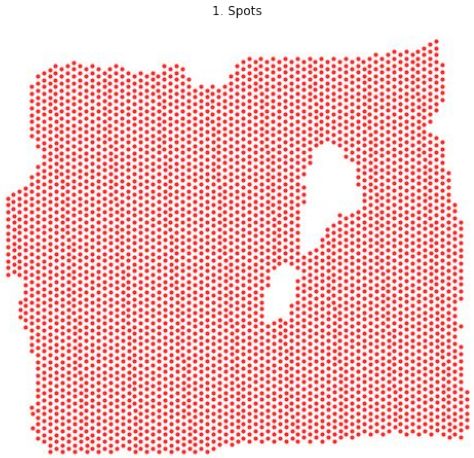
- Example with Human Breast cancer data
  - Public data : Available at 10x website

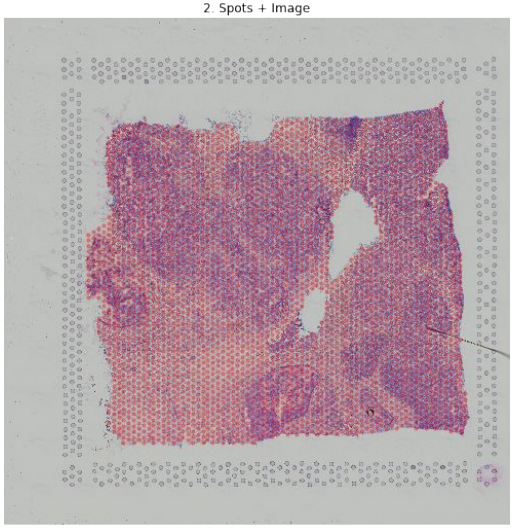


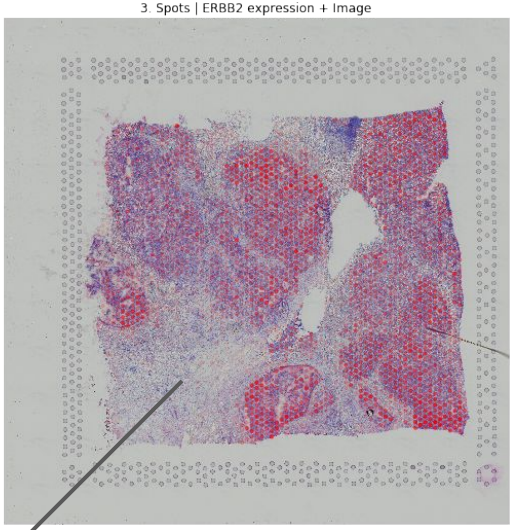


An initial assessment

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







Facecolor intensity proportional to gene expression value

Visualizing high dimensional spatial data

(Tips)

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

Visualizing high dimensional spatial data

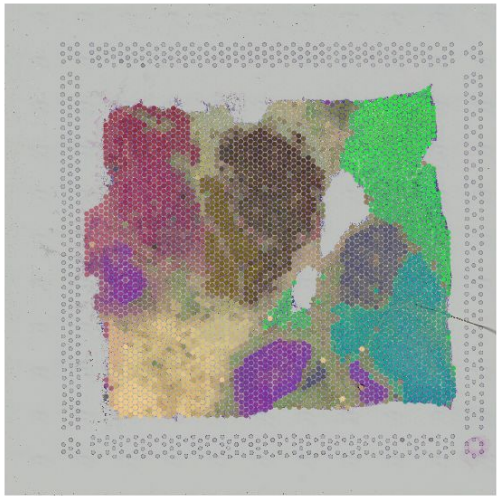
(Tips)

- 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

(Tips)

- 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.

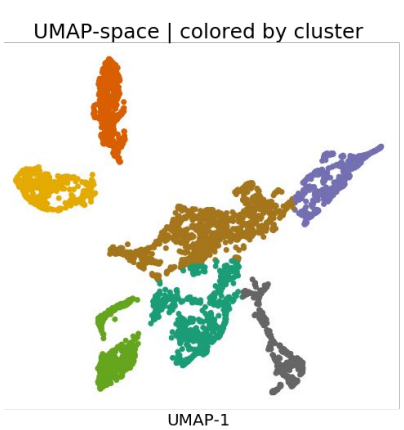
- No magic recipe to give
  - How to process your data is very much dependent on the samples and objective
  - Much can be learnt from analysis of single cell data
  - Will give some general advice
- Consider filtering :
  - Genes based on expression levels (total expression > thrs)
  - Genes based on 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

P1

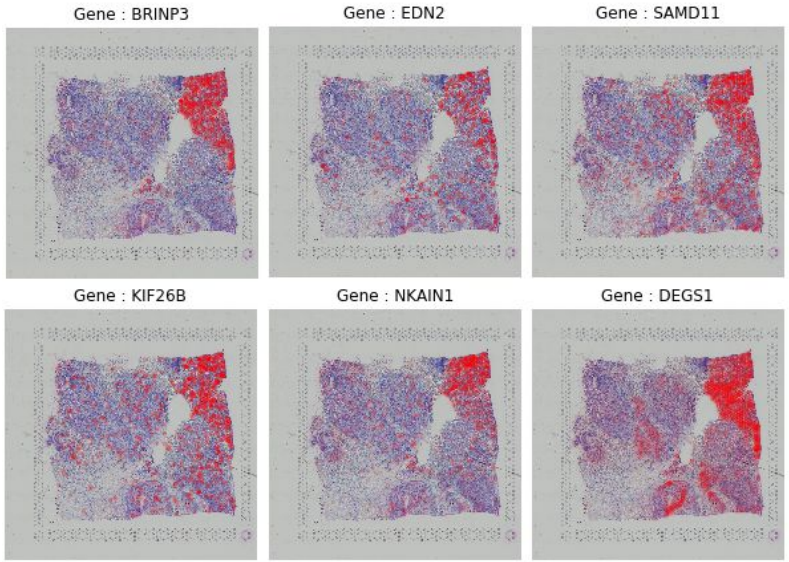
- Cluster the spots based on gene expression
  - Normalize → $PCA_{n=20}$ → UMAP → GMM
- Backmap clusters onto tissue
- Use HE-image as reference
  - Sanity check - does it make sense?
  - Valuable resource
- Next, annotate clusters
  - Find genes associated with cluster
  - Functional enrichment analysis



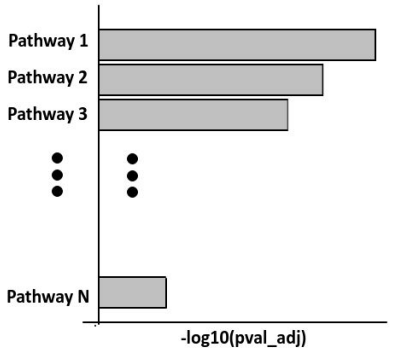
SciLifeLab

Example : Basic Analysis

Clustered data



DE analysys | Here cluster 1 (orange) vs. all

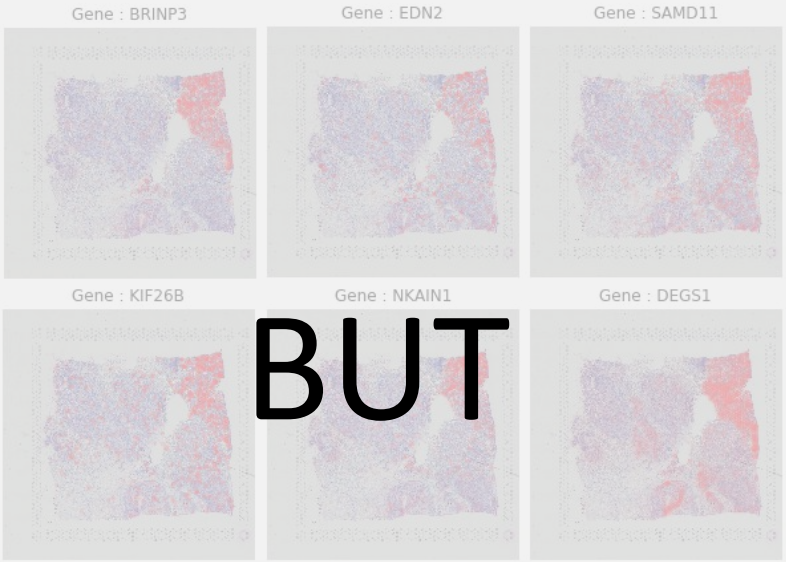


Functional Enrichment Analysis

All helpful in annotation of clusters

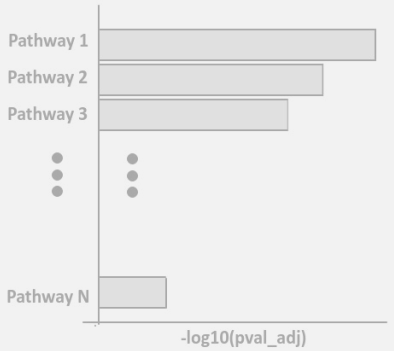
Example : Basic Analysis

Clustered data



BUT

DE analysys | Here cluster 1 (orange) vs. all



Functional Enrichment Analysis

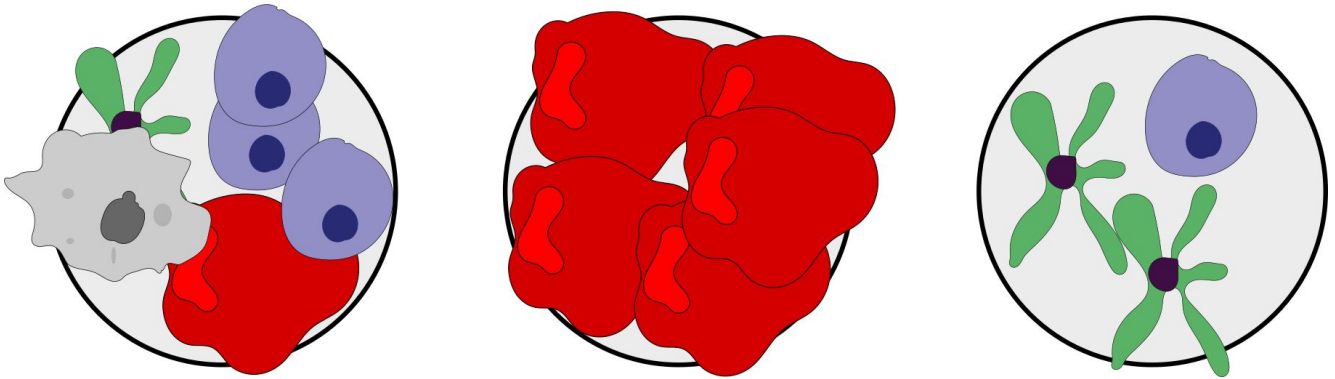
All helpful in annotation of clusters

Example : Basic Analysis

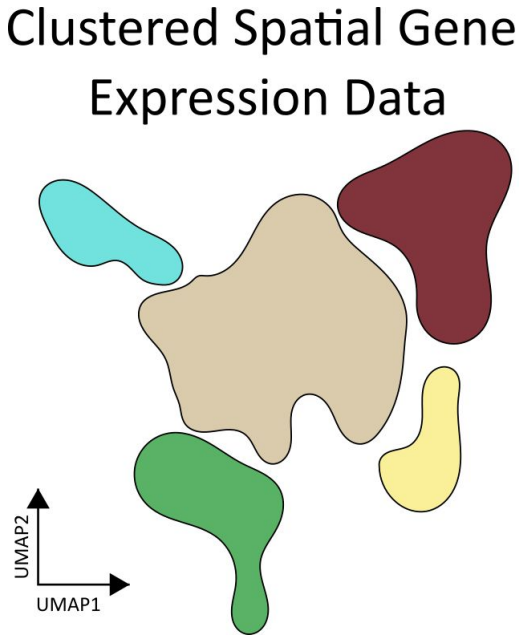
Remember :

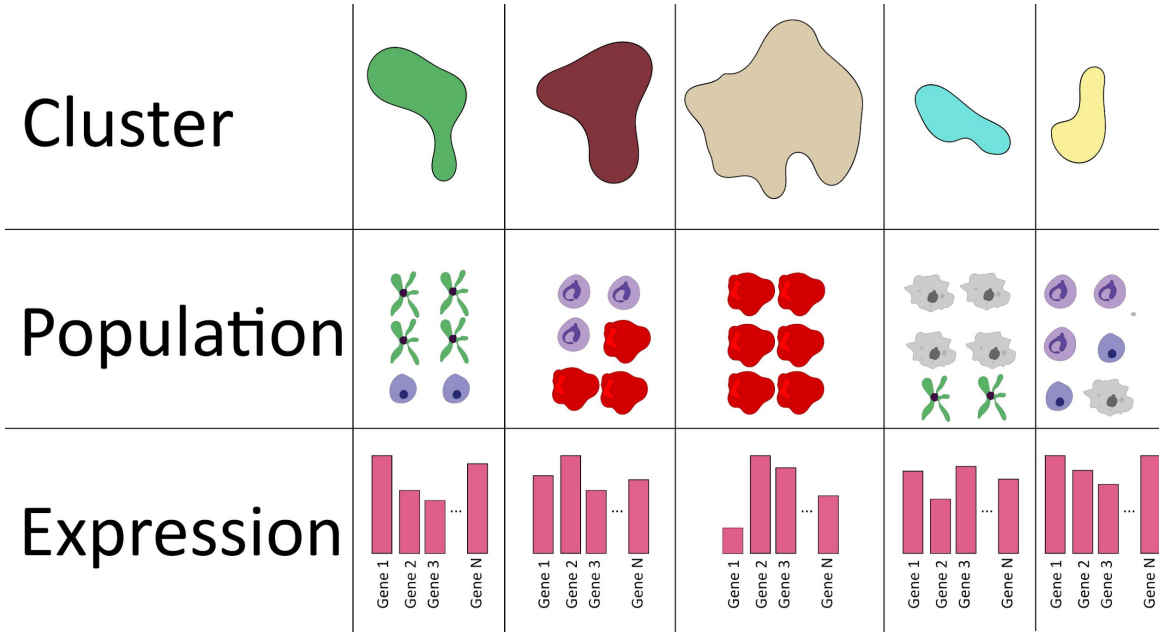
"each spot is a mixture of multiple cells,
i.e., one spot may contain multiple cell types"

Spot 1 Spot 2 Spot 3



Cluster ≠ Cell type

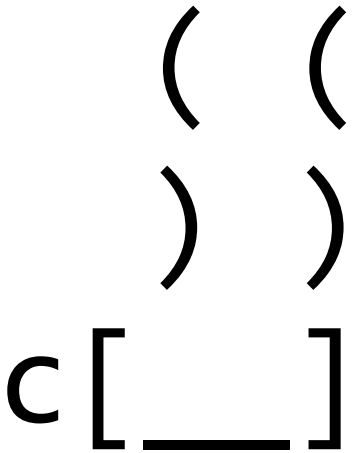




So where are my cell types?

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



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

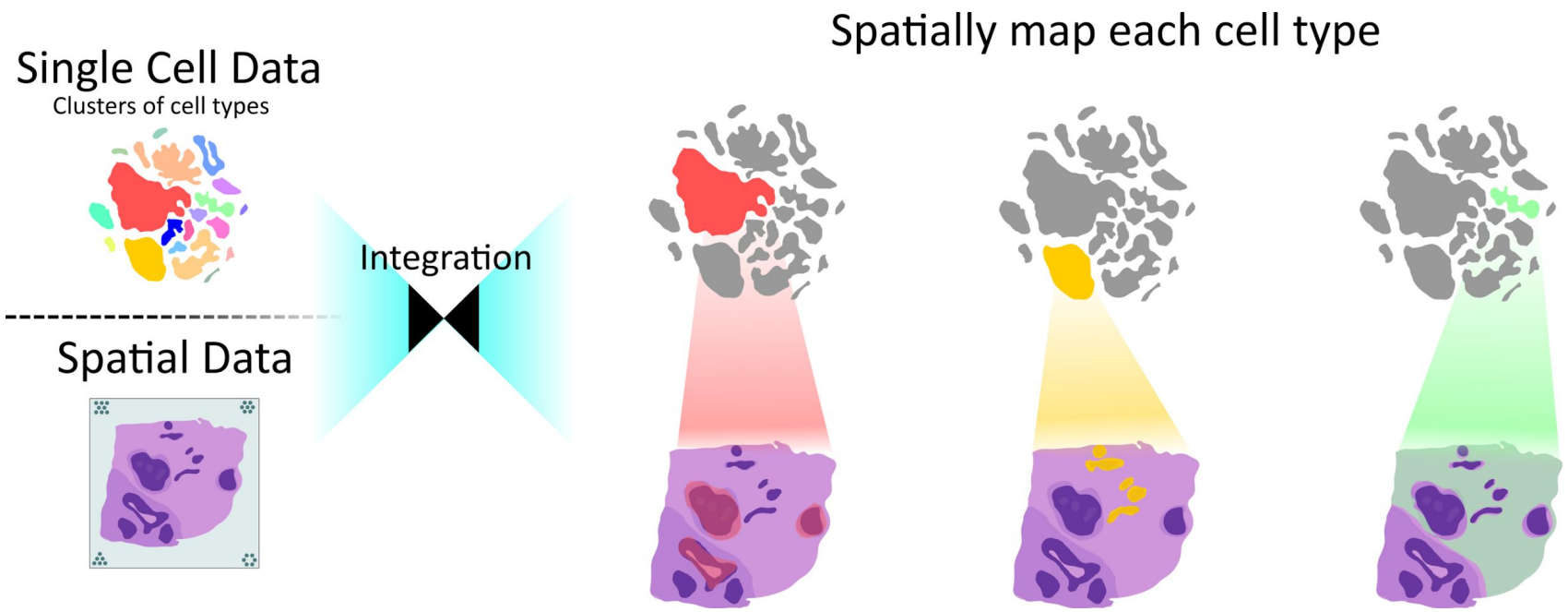
So where are my cell types?

- 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!

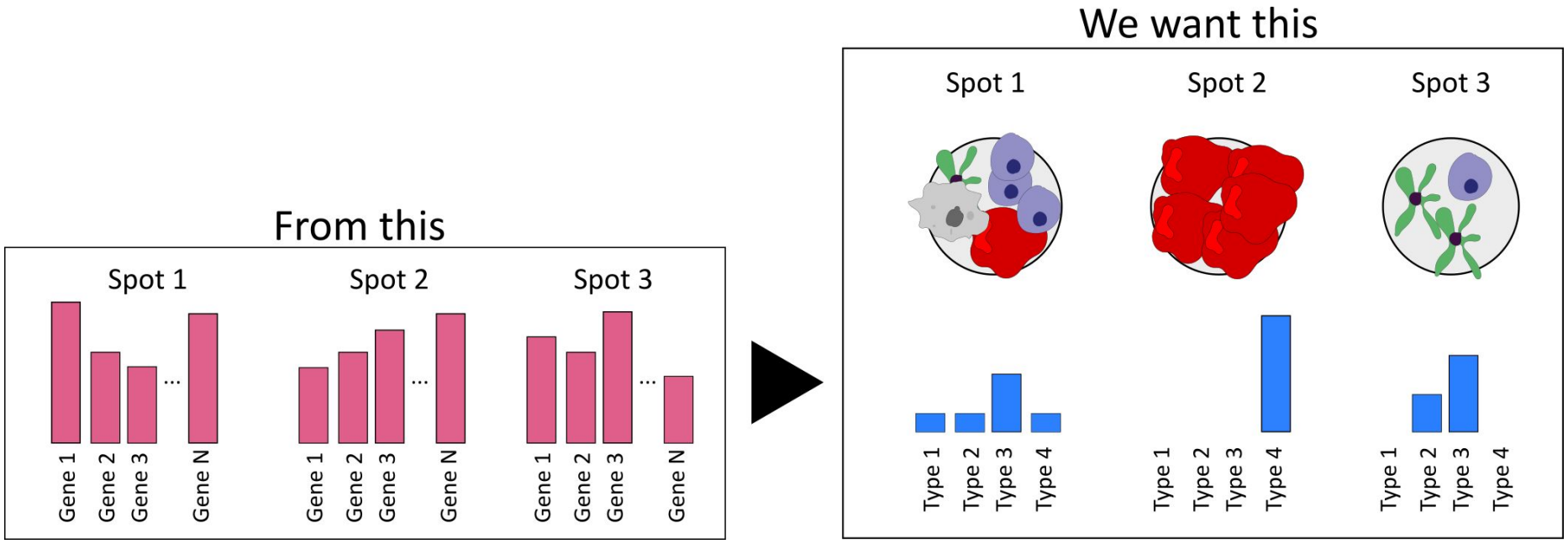
So where are my cell types?

- 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

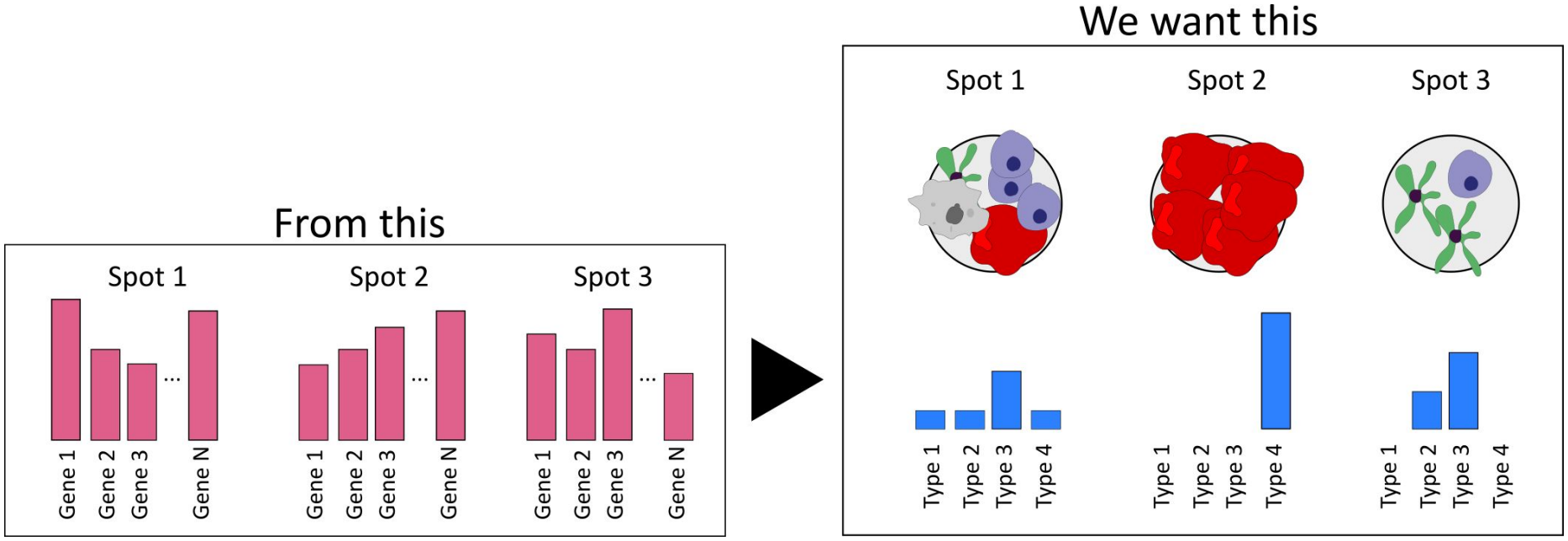
Integration of Single Cell and Spatial Data



Our objective : deconvolve expression data



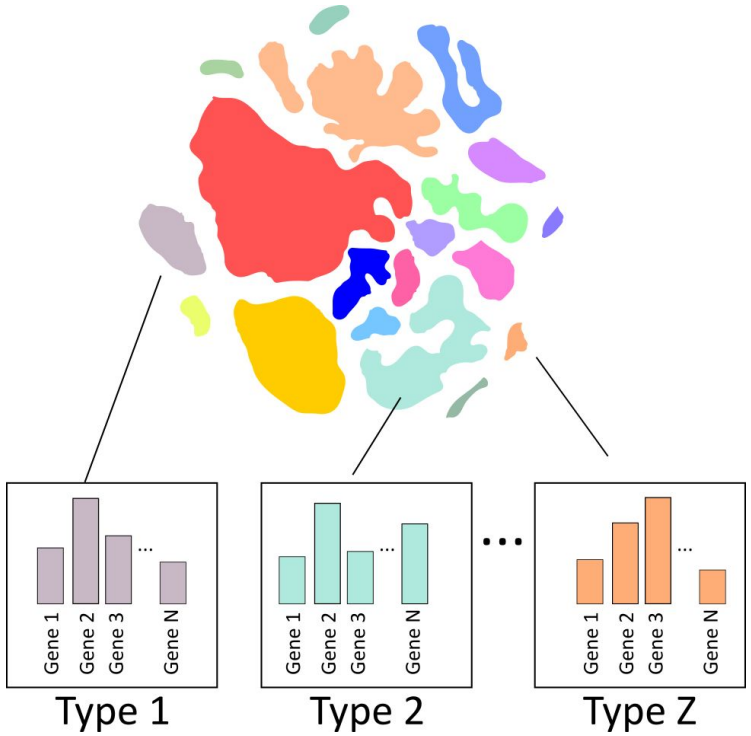
Our objective : deconvolve expression data



Suggested approach : *Model-based Probabilistic Inference*

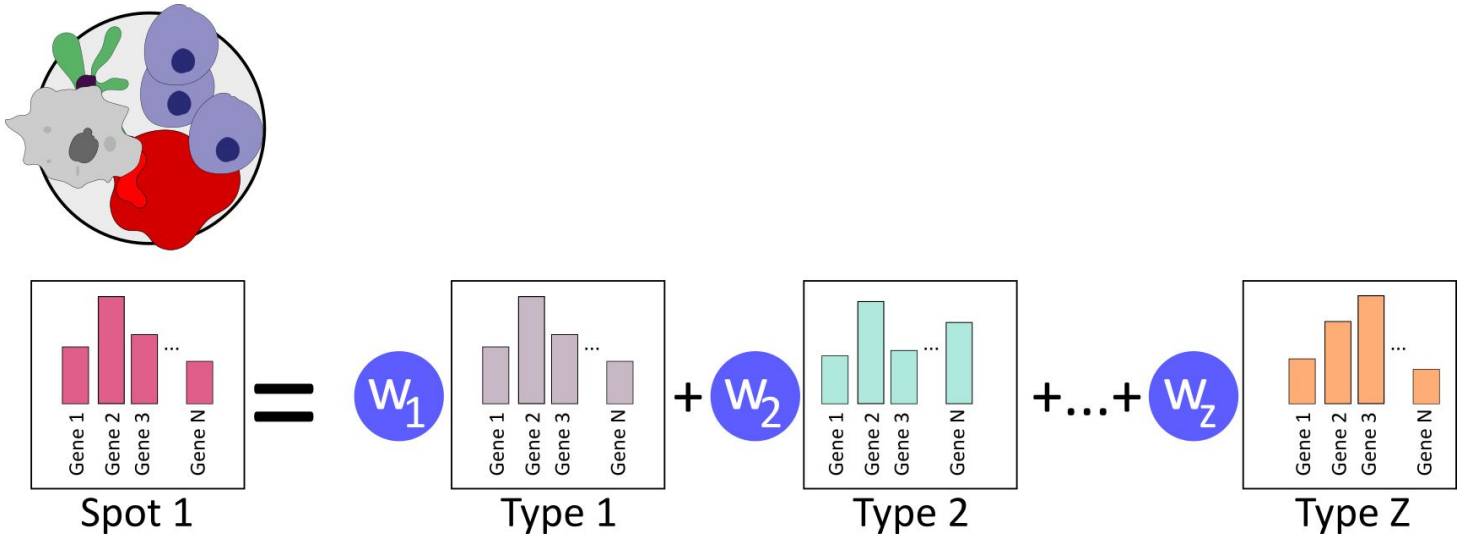
It’s as easy as 1-2-3

1. Infer cell type expression parameters from SC data



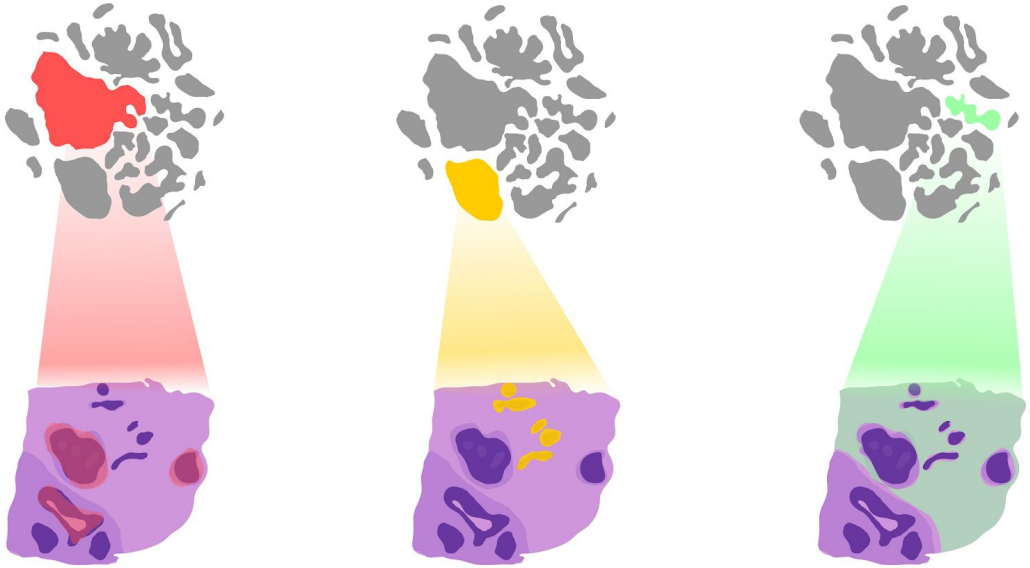
It’s as easy as 1-2-3

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



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

Scaling factor

Gene Bias

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

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

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

Change Index of summation

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

Number of cells from type z at spot s

Slightly more complex than adding bar graphs..

- Probabilistic Model

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

join scaling factor and cell counts

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

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

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

Total number of cells at spot s

Slightly more complex than adding bar graphs..

- Probabilistic Model

Number of cells from cell type z at spot s

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

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

Total number of cells at spot s

Slightly more complex than adding bar graphs..

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

Likelihood function

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

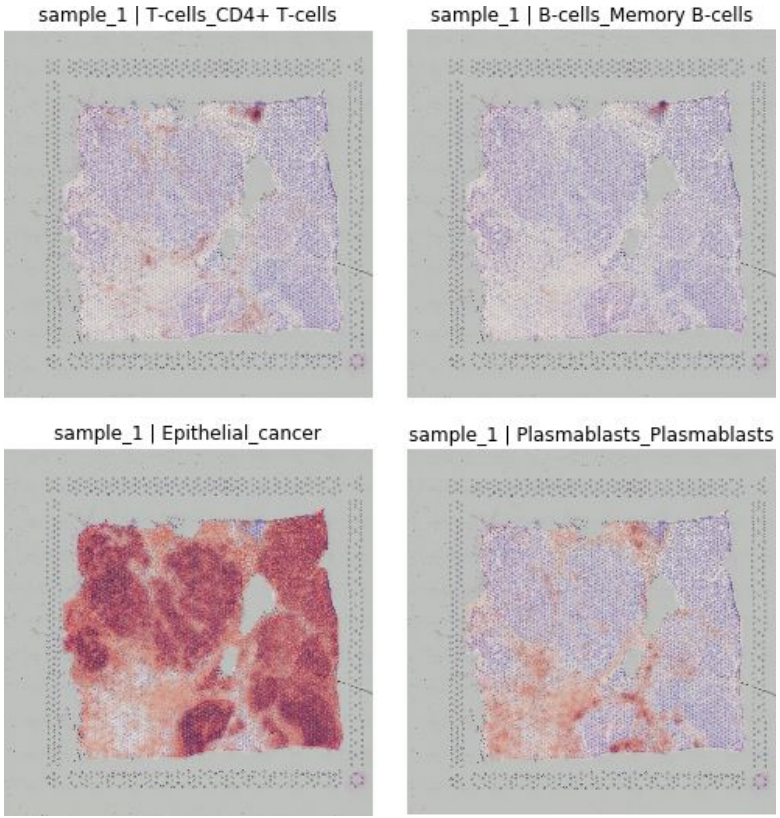
https://github.com/almaan/stereoscope

#shameless self-advertising

SciLifeLab

Applying it to our breast cancer data

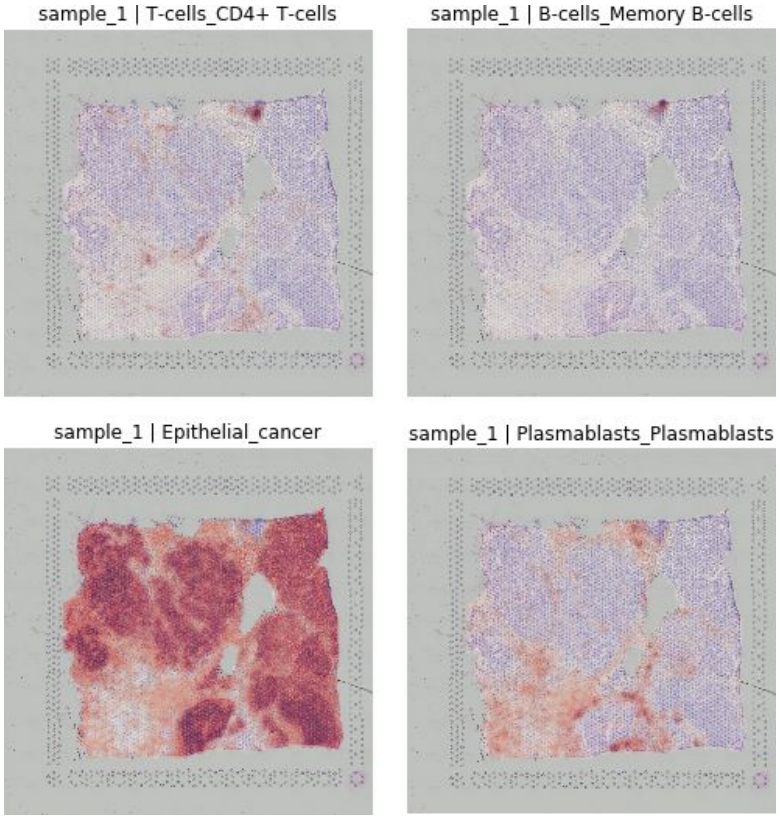
Proportion estimates



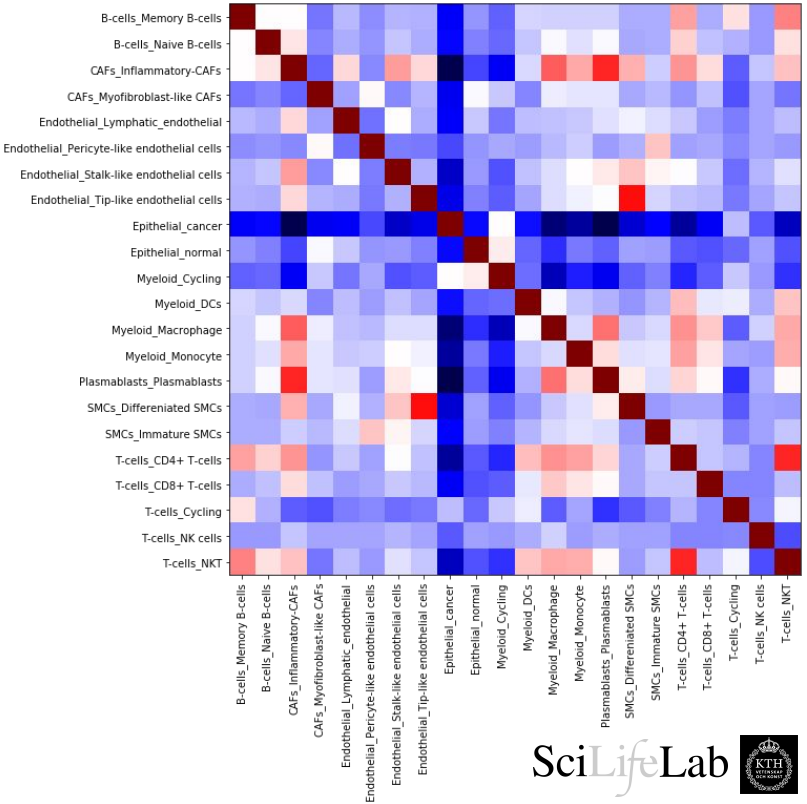
Applying it to our breast cancer data

P2

Proportion estimates



Cell type co-localization



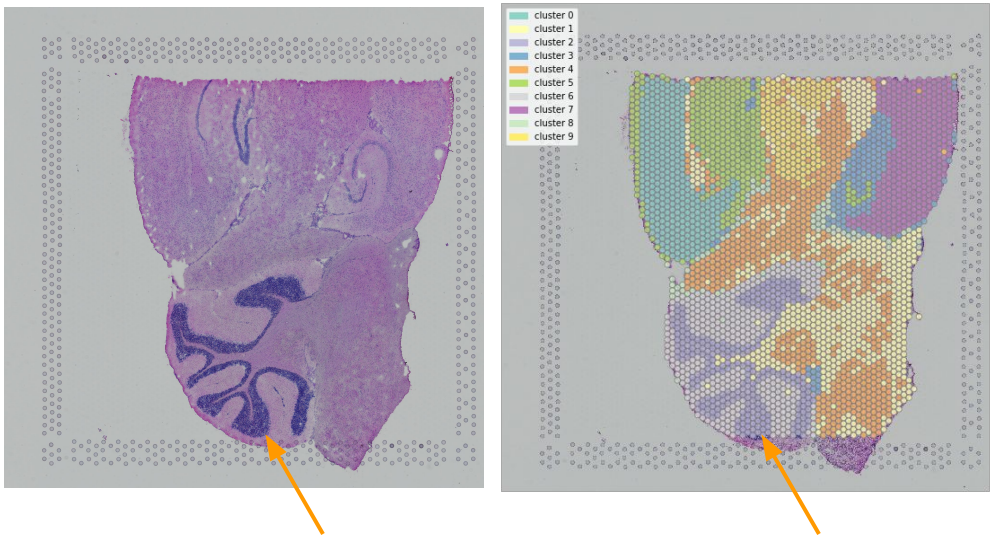
SciLifeLab KTH

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!

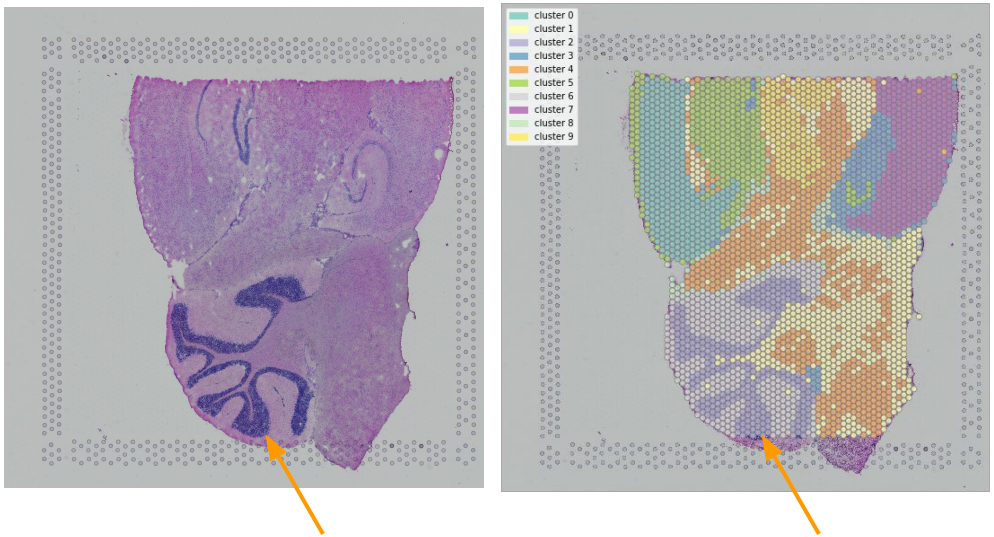
Example Analysis : Expression as a function of distance

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



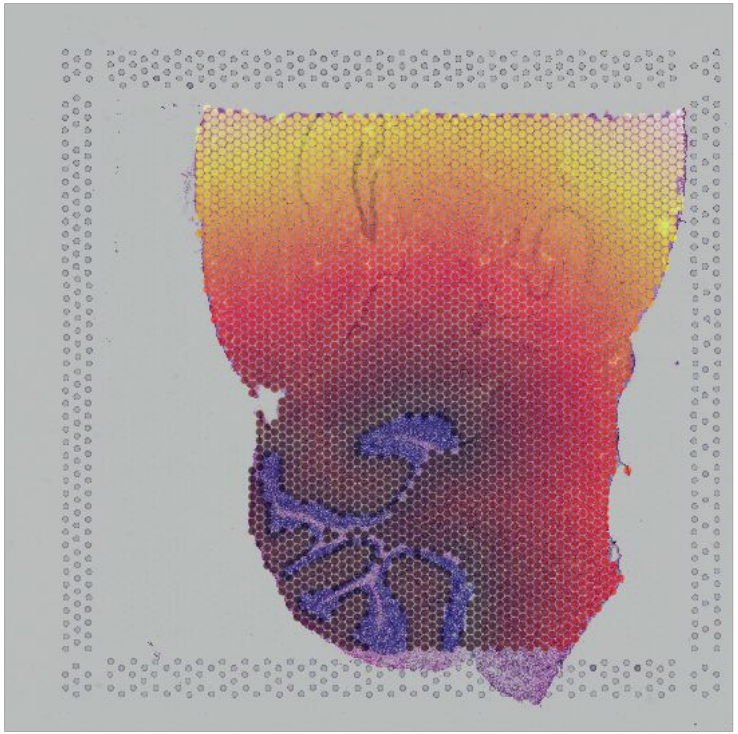
Example Analysis : Expression as a function of distance

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

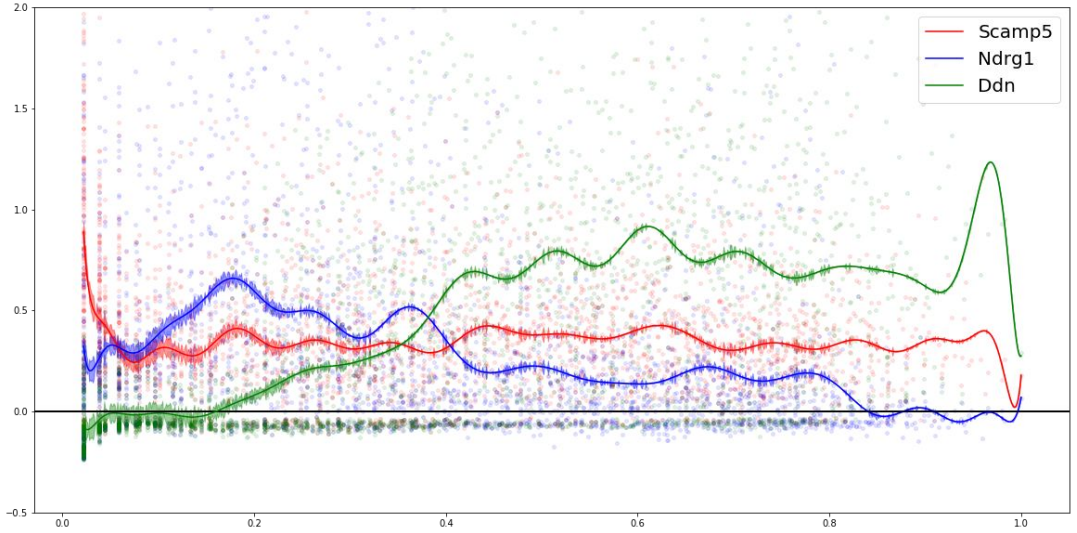


- We may then ask how gene expression changes with the distance to this cluster

Example Analysis : Expression as a function of distance

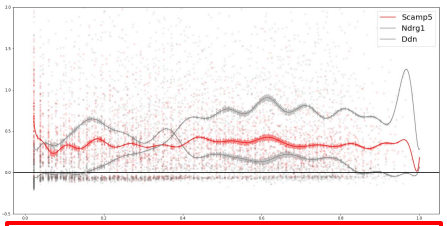


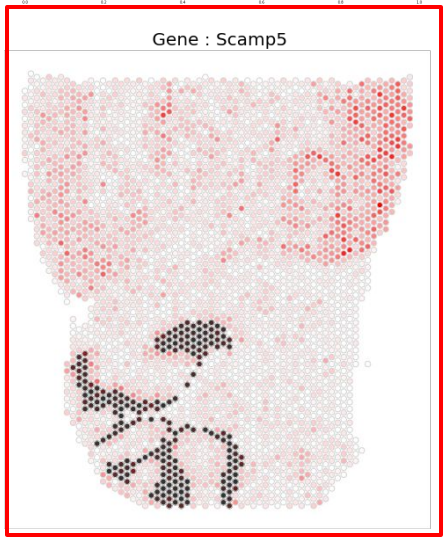
Spots colored by distance do cluster 2

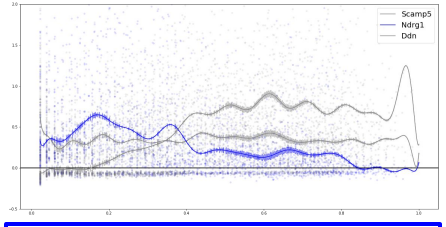


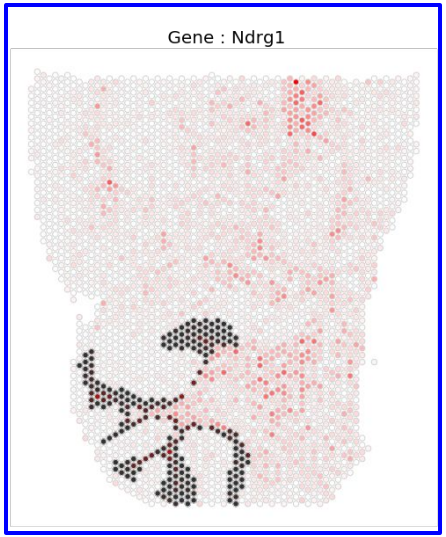
Gene expression as a function of the distance to cluster 2

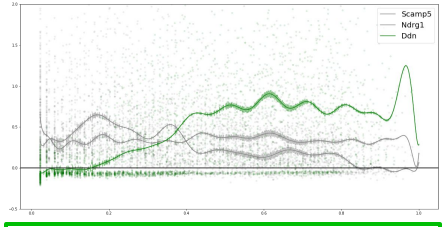
Example Analysis : Expression as a function of distance

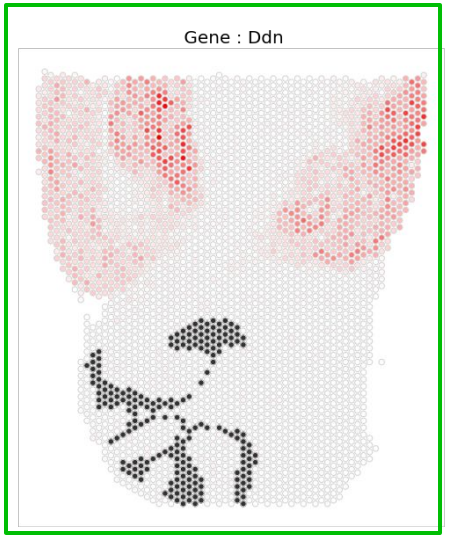












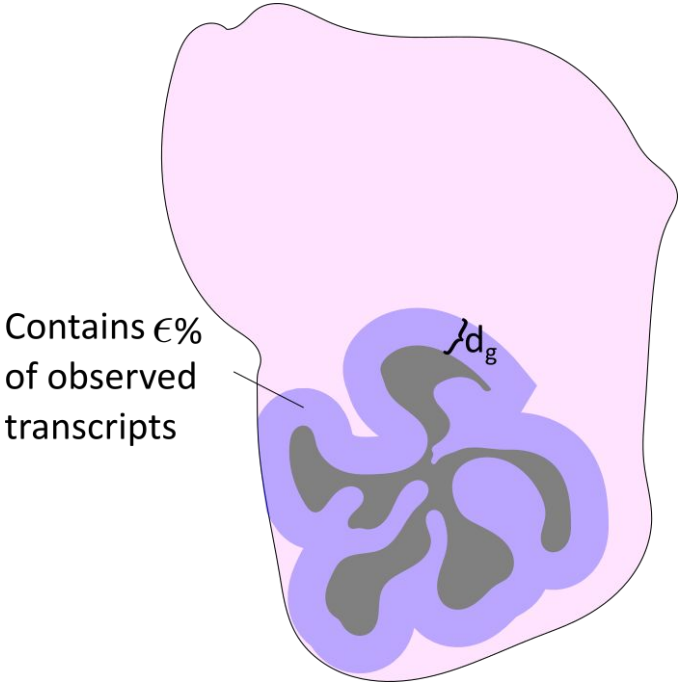
Example Analysis : Expression as a function of distance

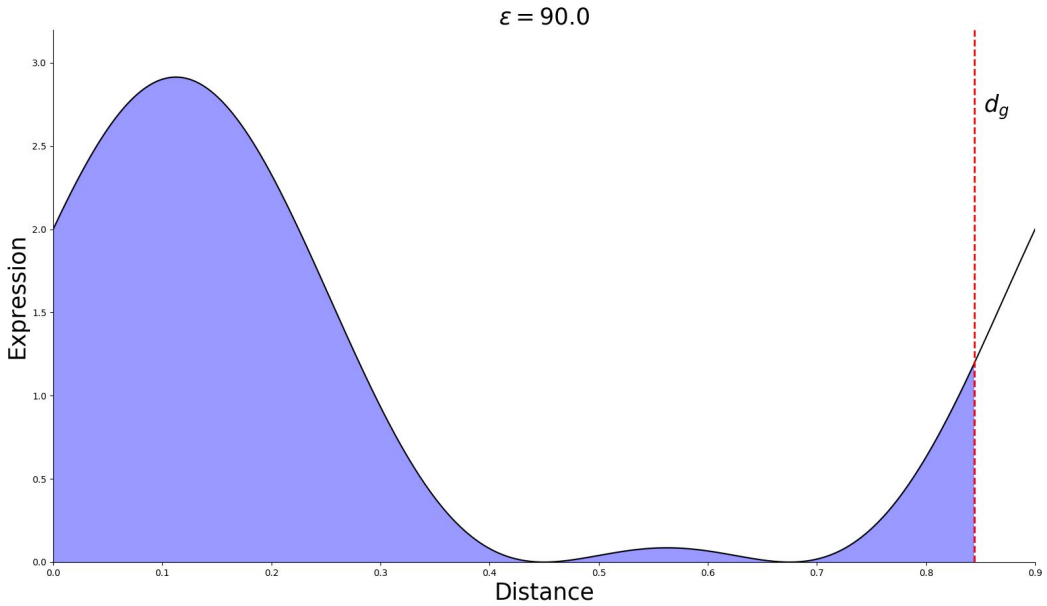
- Can also ask : "*within which distance (d_g) from cluster 2 is ε % of all transcripts from gene g contained?*"

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

Example Analysis : Expression as a function of distance

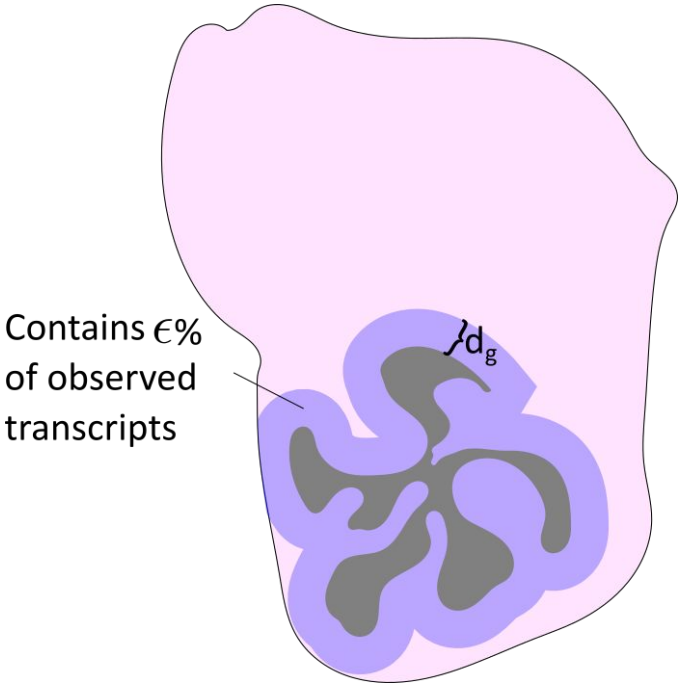
- Can also ask : "*within which distance (d_g) from cluster 2 is ε % of all transcripts from gene g contained?*"

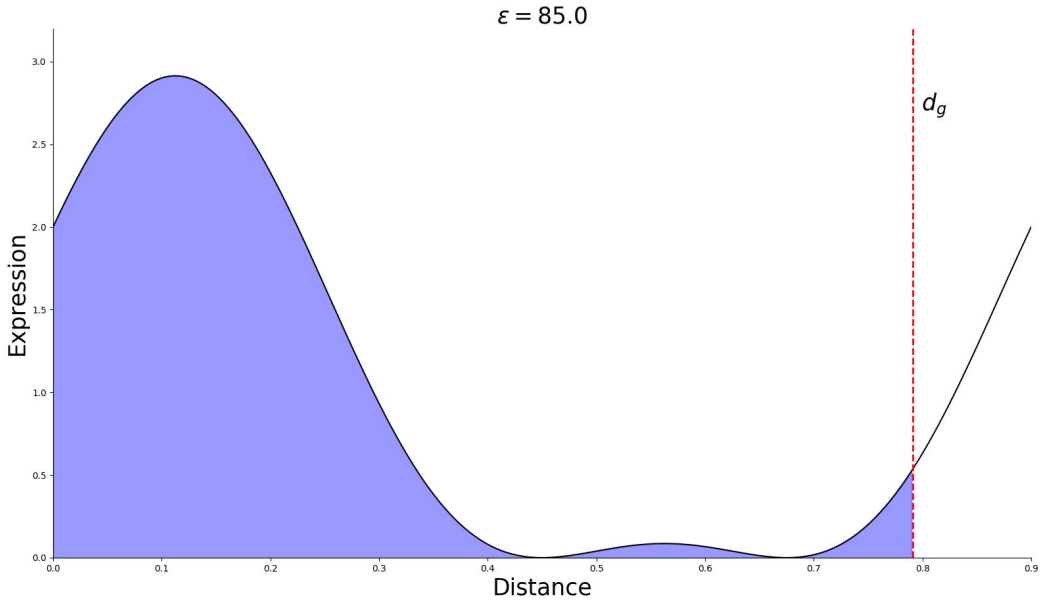




Example Analysis : Expression as a function of distance

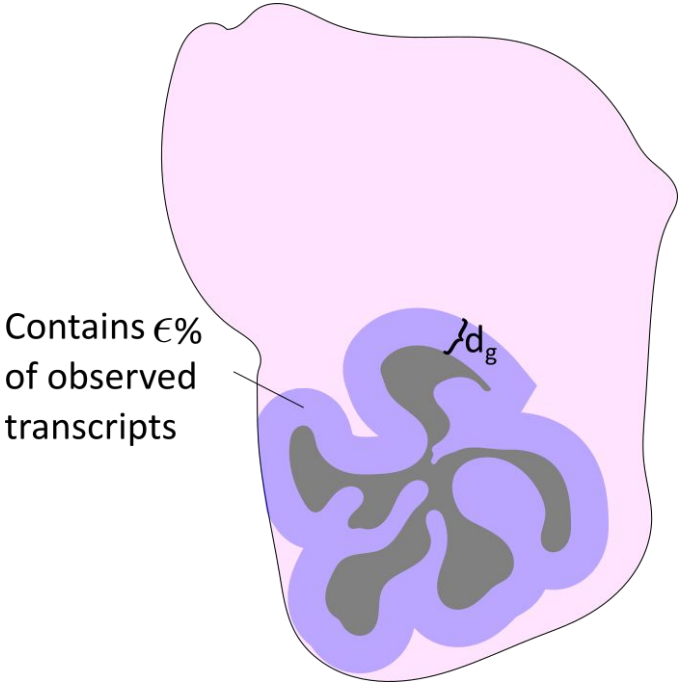
- Can also ask : "*within which distance (d_g) from cluster 2 is ε % of all transcripts from gene g contained?*"

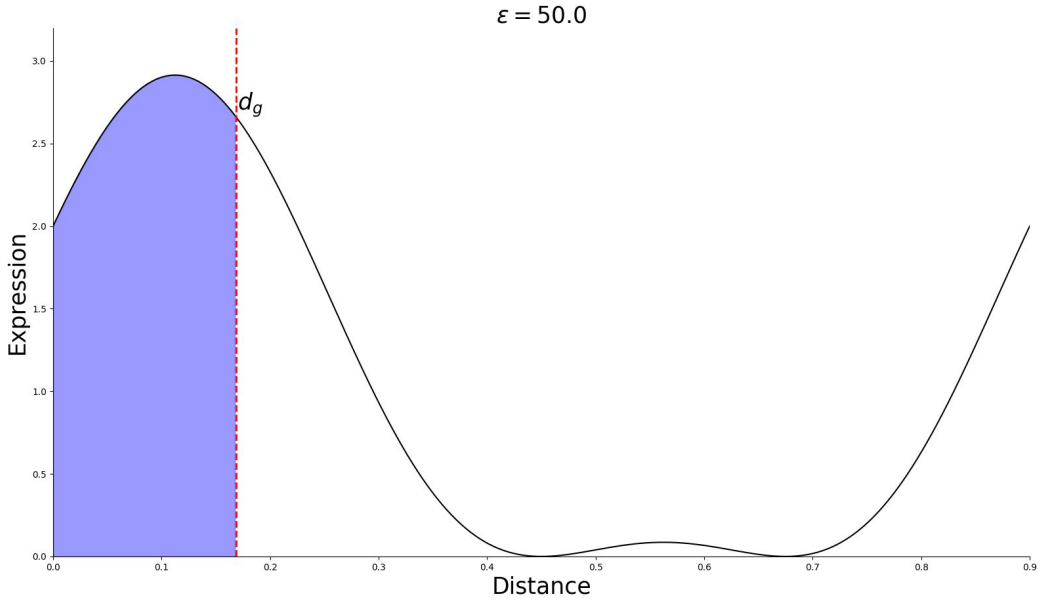




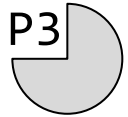
Example Analysis : Expression as a function of distance

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





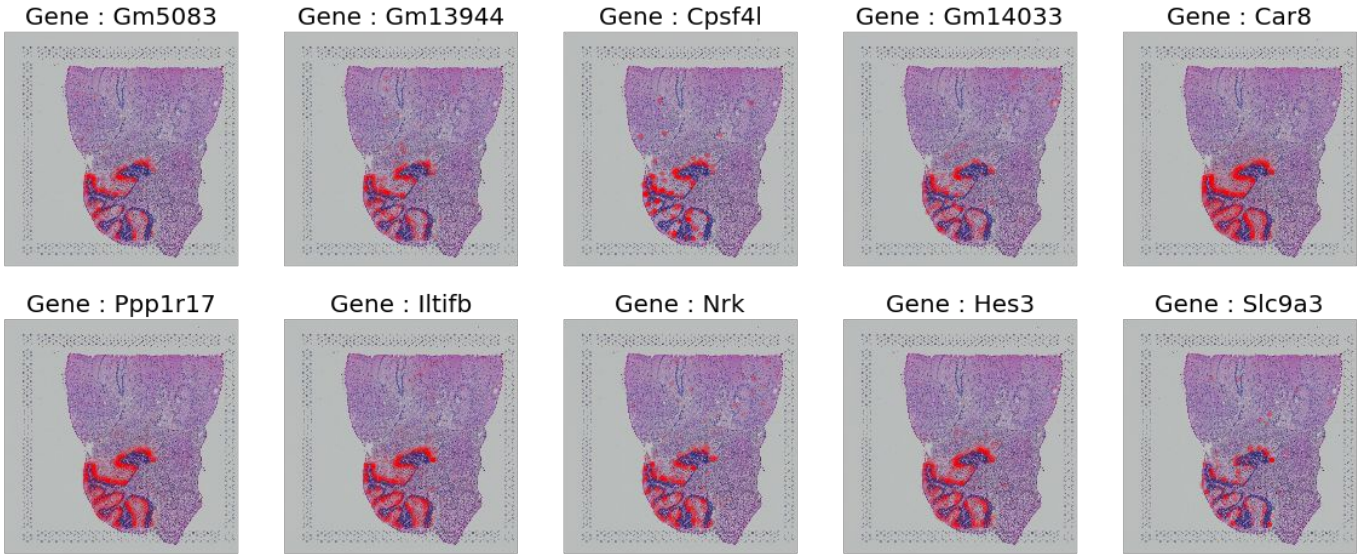
Example Analysis : Expression as a function of distance



- Alternatively : "*which genes has ε % within the shortest distance (d_g) from cluster 2?*"

Example Analysis : Expression as a function of distance

- Alternatively : "*which genes has ε % 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