



Exploring Spatial Transcriptomics

2020-05-28
Alma Andersson



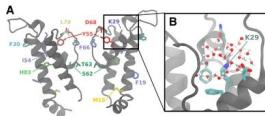
<https://github.com/almaan>



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

A brief Introduction

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



 SPATIAL
research

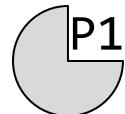


■ ■ | Outline

- ~~Introduction~~
- Notation
- Background
- Data Processing
- Data Analysis
 - Basic Analysis (Clustering and Visualization)
- Break
 - Questions
- Data Analysis
 - Mapping of cell types
 - More “spatially” focused
- Exercises (Info)
- Questions

Notation

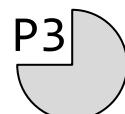
- Exercise session consists of 3 parts
- Symbols below used to indicate when material is included in one of these



- Material in Part 1



- Material in Part 2



- Material in Part 3

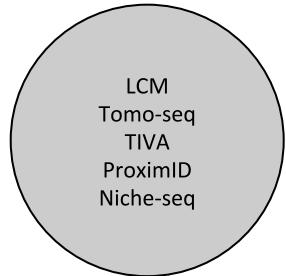
Background

■ ■ | The spatial space | Overview of techniques

What is currently out there?

The spatial space | Overview of techniques

Microdissection-based

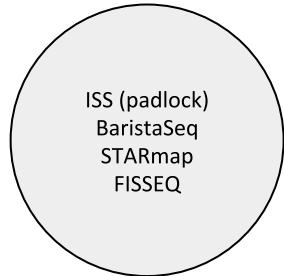


Isolate a region of interest, place isolate in separate well and sequence (either by bulk or single-cell methods).

“Brute Force” approach.

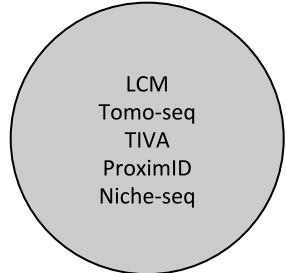
The spatial space | Overview of techniques

In situ sequencing



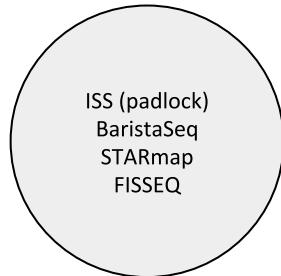
Sequence the transcripts in place. Offer sub-cellular resolution. Some relies on "*a priori*" defined targets, but not all.

Microdissection-based

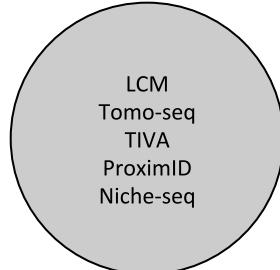


The spatial space | Overview of techniques

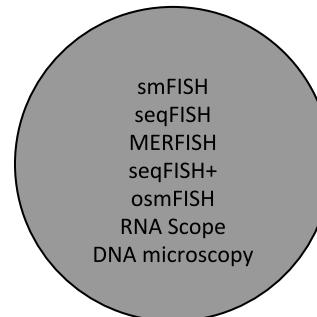
In situ sequencing



Microdissection-based



In situ hybridization

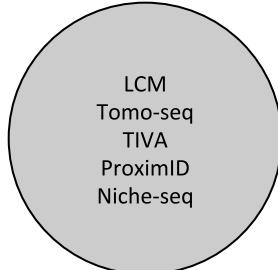


Labeled probes for specific targets, hybridize in place. Requires "*a priori*" defined targets. Expansion strategies and decoding scheme has helped to overcome spectral overlap.

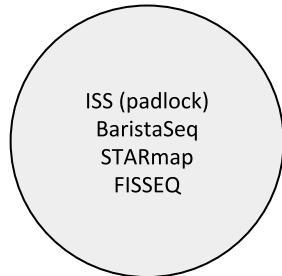


The spatial space | Overview of techniques

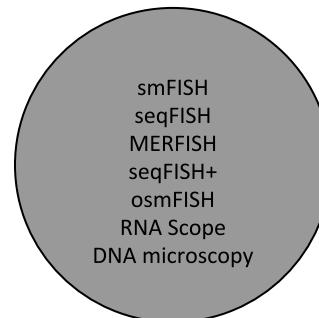
Microdissection-based



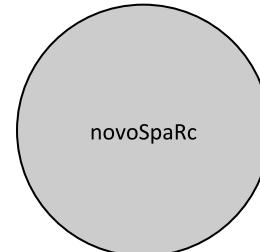
In situ sequencing



In situ hybridization

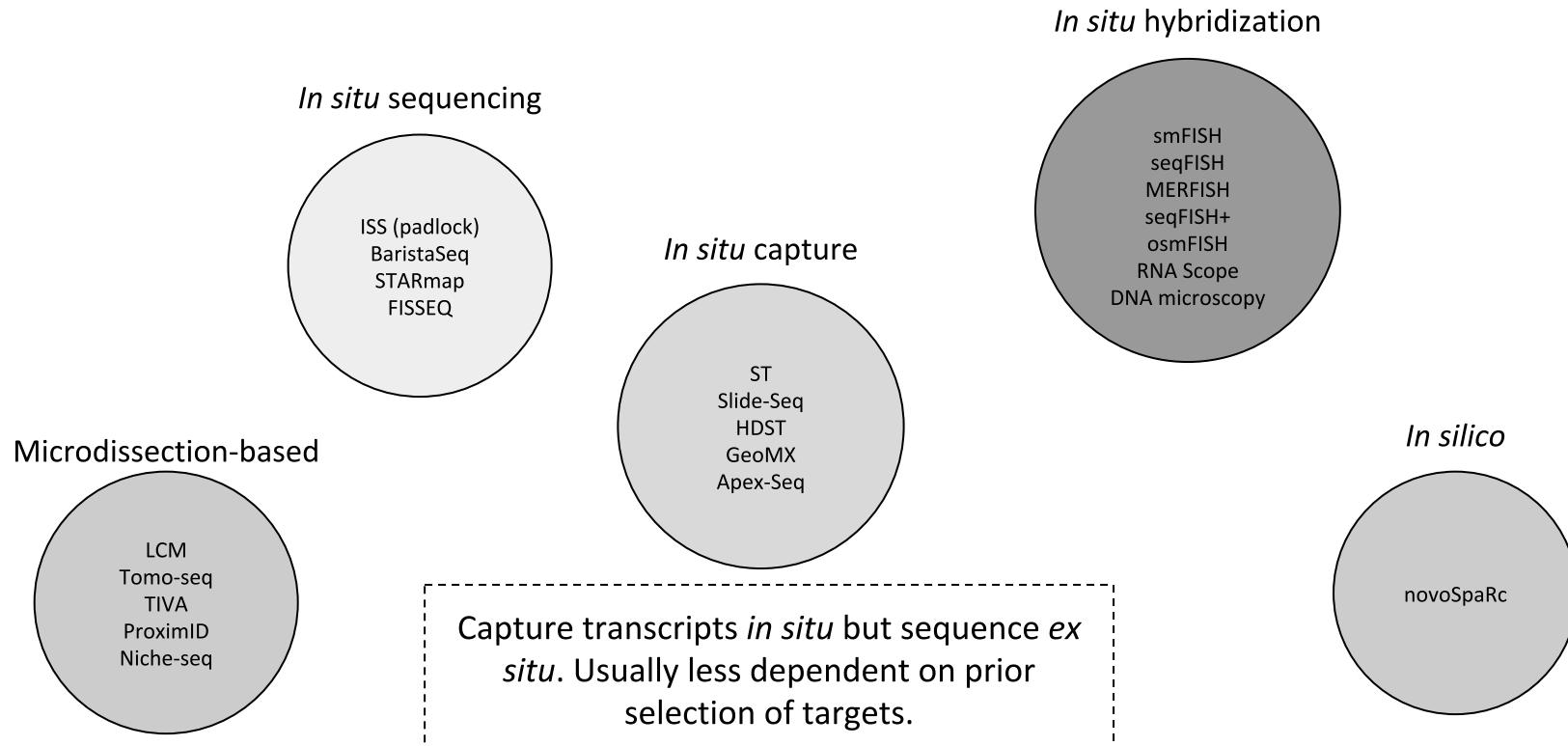


In silico



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

The spatial space | Overview of techniques

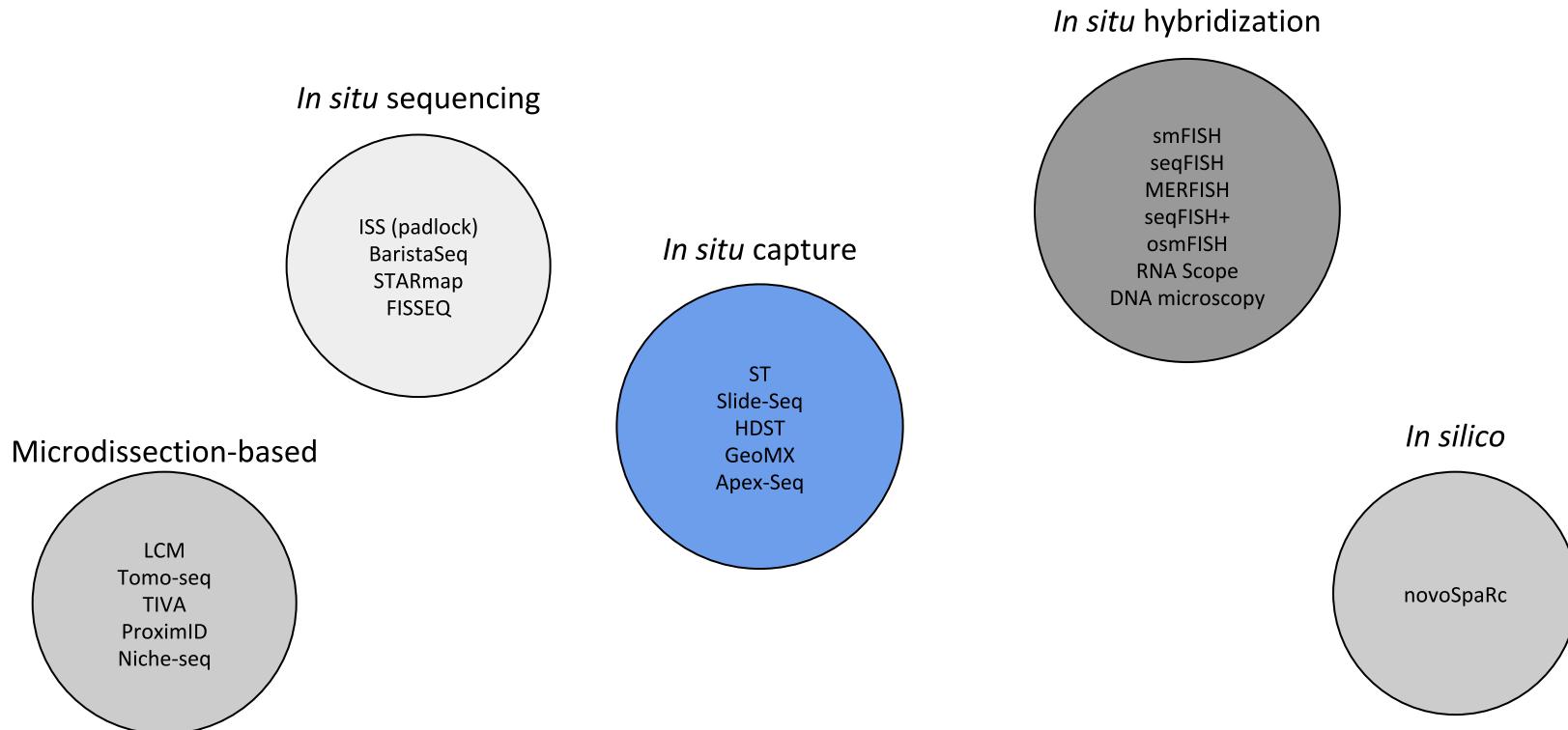




The spatial space | Overview of techniques

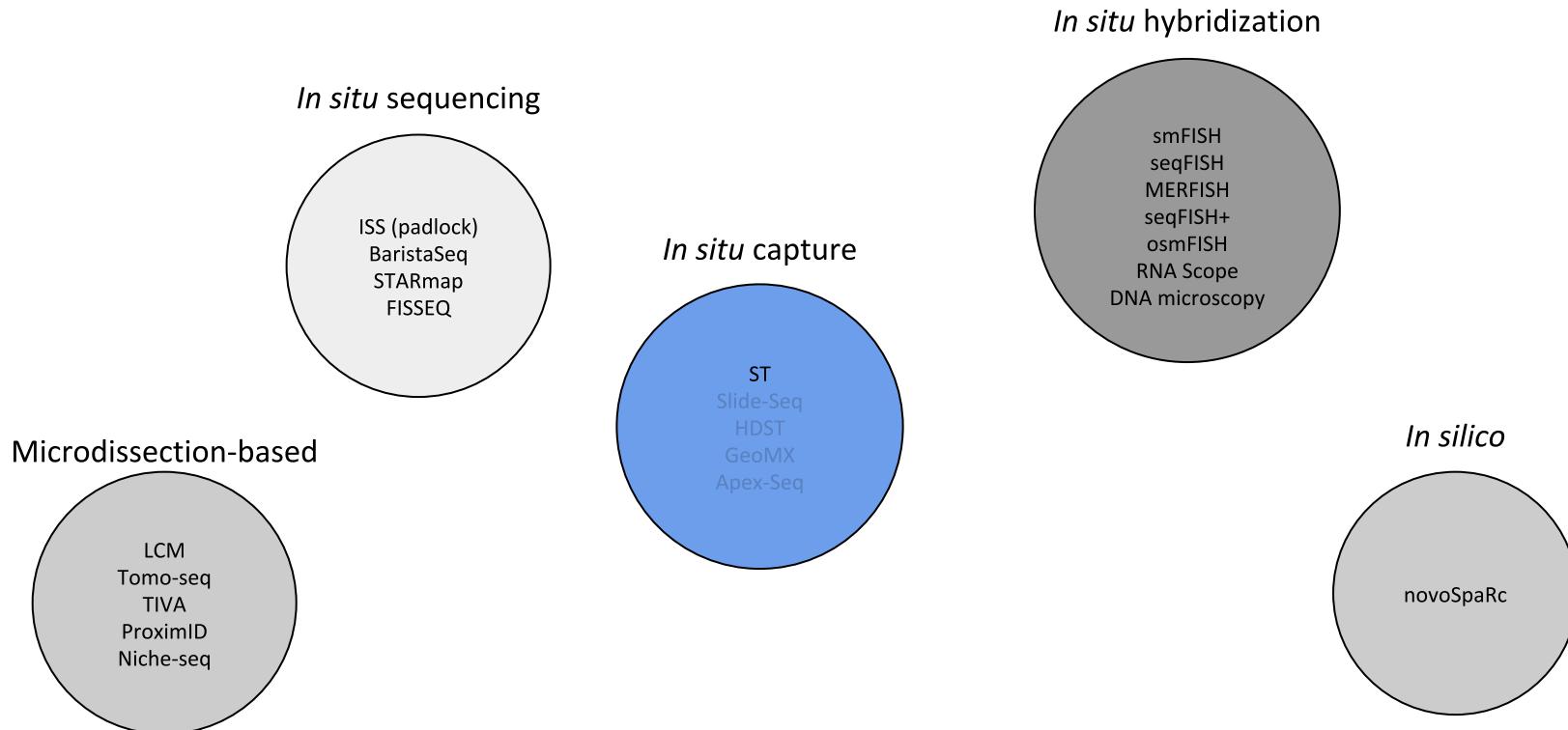


The spatial space | Overview of techniques



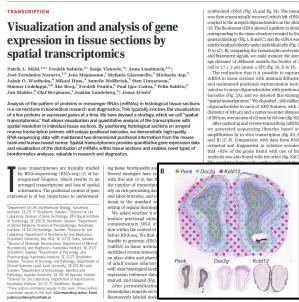


The spatial space | Overview of techniques



Spatial Transcriptomics (ST)

Mid 2016



“Spatially barcoded arrays”

Technique/Method Name : *Spatial Transcriptomics*

Science Publication

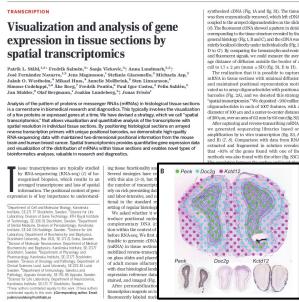
Ståhl et.al

DOI: 10.1126/science.aaf2403

SciLifeLab |

Spatial Transcriptomics (ST)

Mid 2016



Science Publication
Ståhl et.al

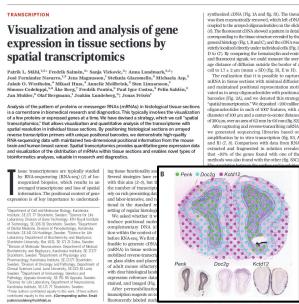
DOI: [10.1126/science.aaf2403](https://doi.org/10.1126/science.aaf2403)

Late 2018

10X
GENOMICS®
(acquisition)

Spatial Transcriptomics (ST)

Mid 2016



Science Publication

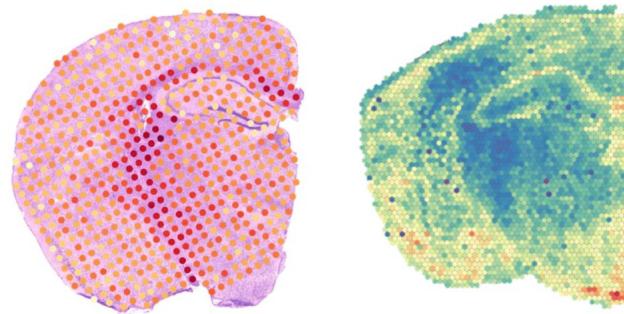
Ståhl et.al

DOI: 10.1126/science.aaf2403

Late 2018

The logo for 10x Genomics, featuring the number "10x" in a large blue font where the "x" is composed of three colored lines (green, yellow, and pink), followed by the word "GENOMICS" in a smaller blue font, and "(acquisition)" in a black font below it.

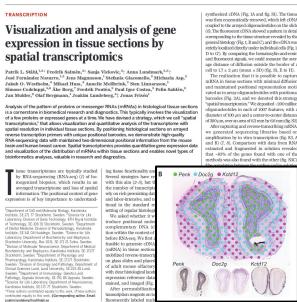
Late 2019



Launch of Visium Spatial Gene Expression Platform

Spatial Transcriptomics (ST) Visium

Mid 2016

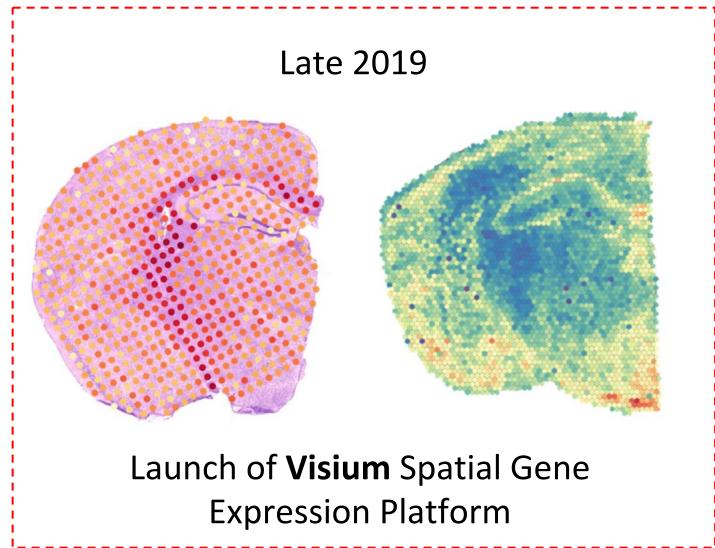


Science Publication
Ståhl et.al

DOI: 10.1126/science.aaf2403

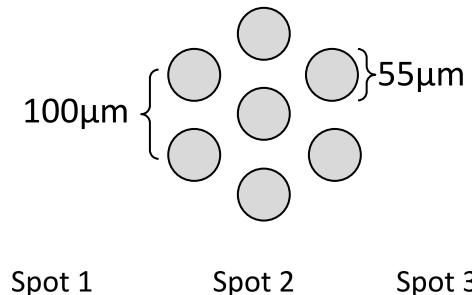
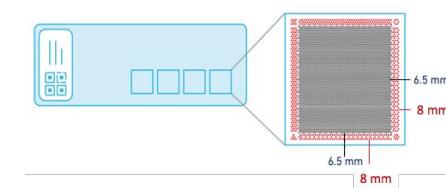
Late 2018

The logo for 10x Genomics consists of the number "10" in a large, dark blue serif font, followed by a stylized "X" composed of four colored segments (green, yellow, red, blue) forming an "X" shape. Below this, the word "GENOMICS" is written in a smaller, dark blue sans-serif font, with a registered trademark symbol (®) at the end. Underneath the main text, the word "(acquisition)" is written in a smaller, dark blue sans-serif font.

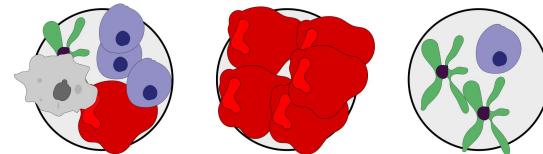


Visium Platform

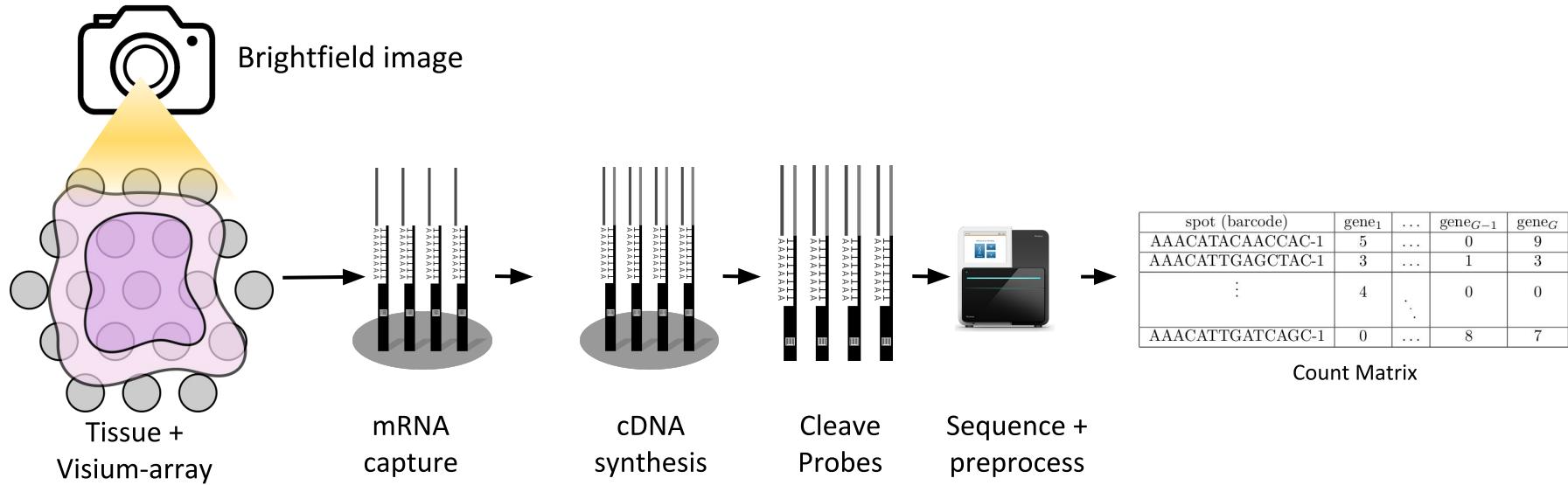
- Array based technique
- 6.5mm x 6.5mm area to put sample on
- 4992 spots arranged in hexagonal grid
- Spot specs:
 - Spot diameter : 55 μm
 - Center to center distance : 100 μm
- Each spot has millions of capture probes
 - spatial barcode
 - polyT sequence
 - captures polyadenylated mRNA
 - Full transcriptome(-ish)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



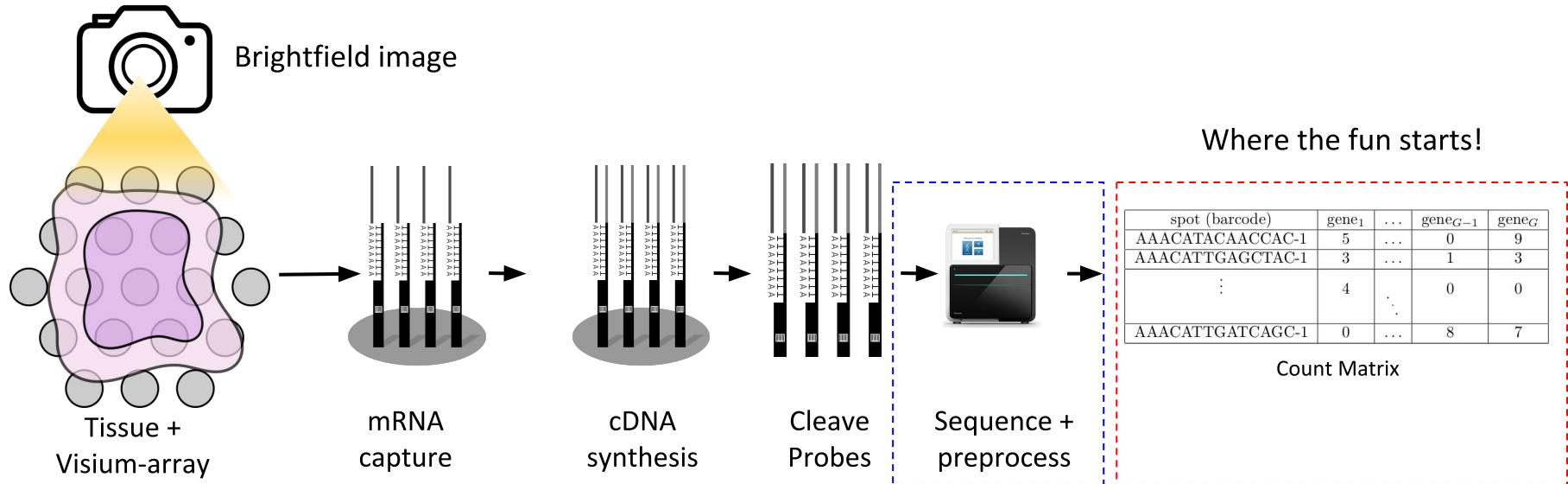
Spot 1 Spot 2 Spot 3



The experimental workflow (in a nutshell)



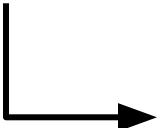
The experimental workflow (in a nutshell)



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.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
```

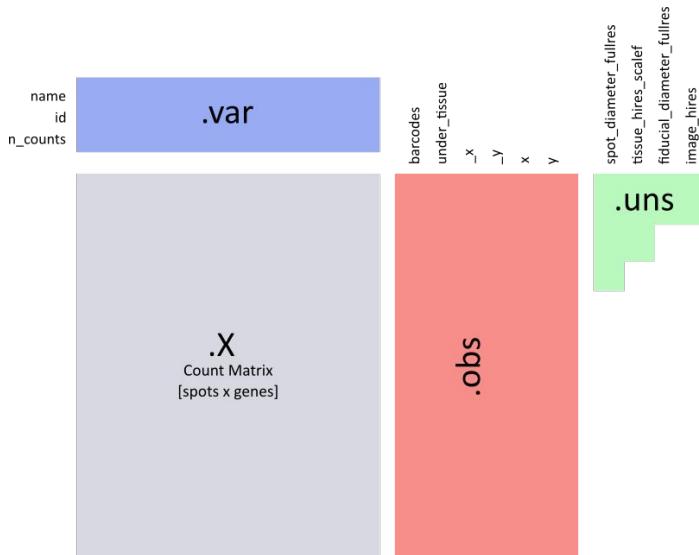
Count matrices as MEX and HDF5
- *raw* : all spots
- *filtered* : spots under tissue

Example of spaceranger count output

<https://support.10xgenomics.com/spatial-gene-expression/software/pipelines/latest/what-is-space-ranger>

Processed data

- Convert to more convenient data object to work with
 - No standardized format
- Personal preference : convert to .h5ad file (will be using in exercises)
 - *scanpy/anndata* teams have their own (similar) format



In short:

- .var - holds gene identifiers
- .obs - spot identifiers and coordinates
- .uns - image and scaling factors

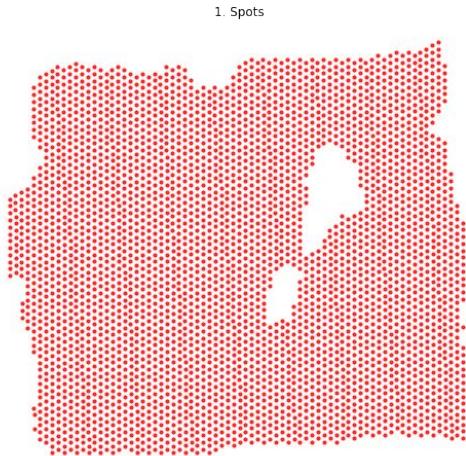
<https://github.com/almaan/space2h5ad/>

An initial assessment

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

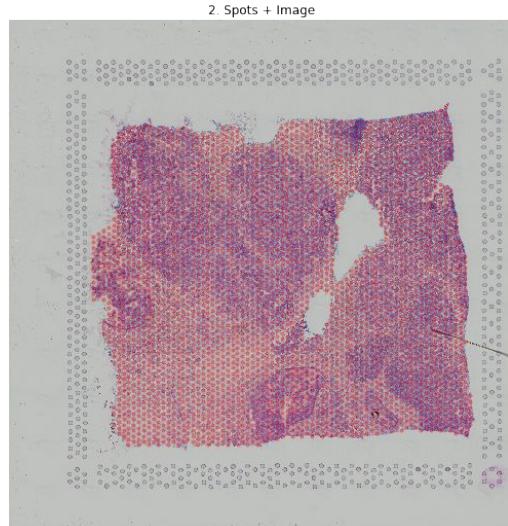
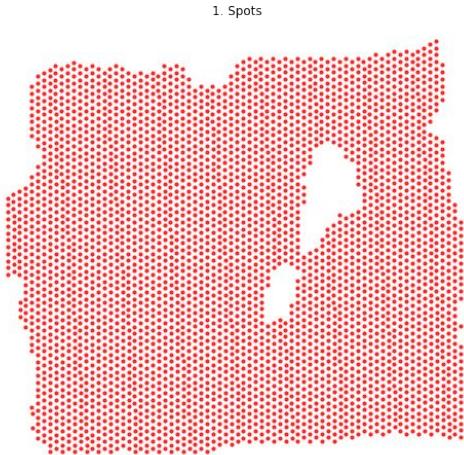
An initial assessment

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



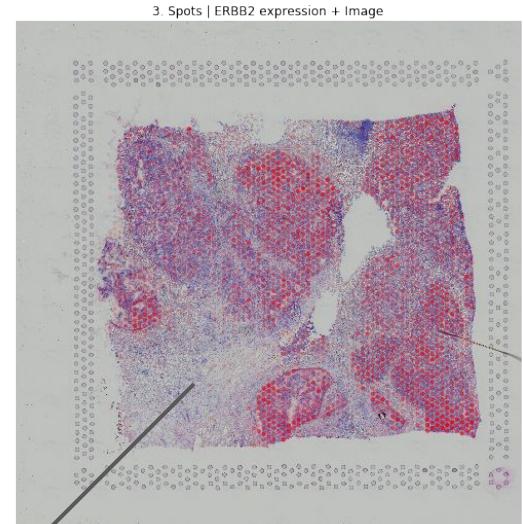
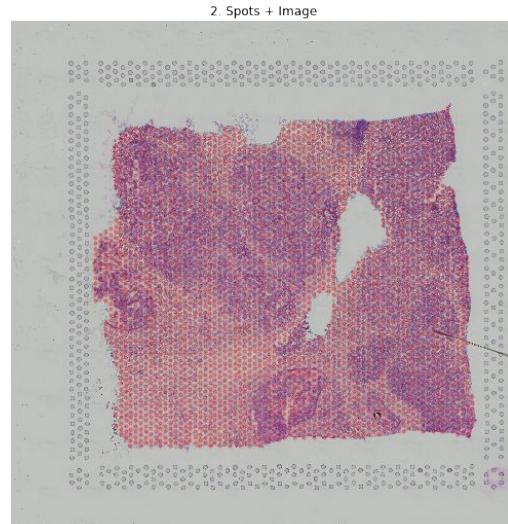
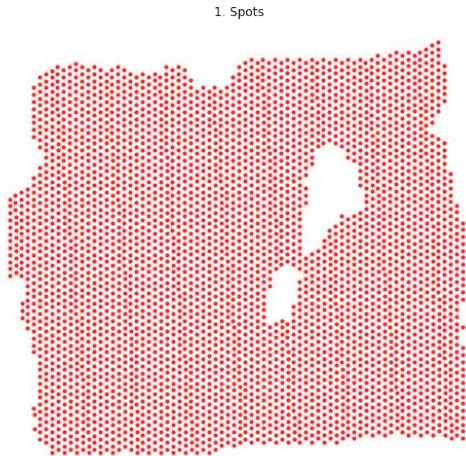
An initial assessment

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



An initial assessment

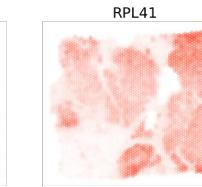
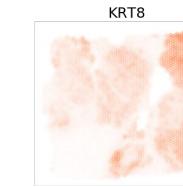
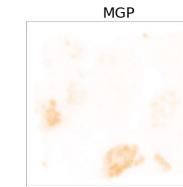
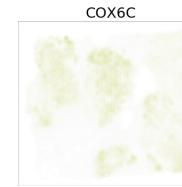
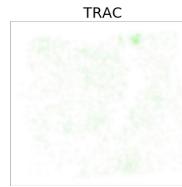
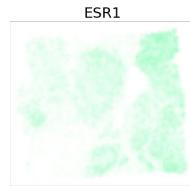
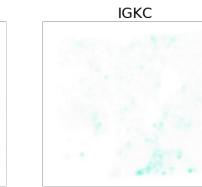
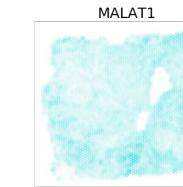
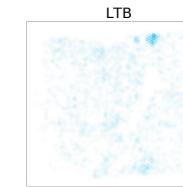
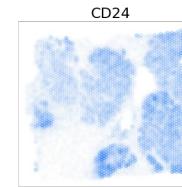
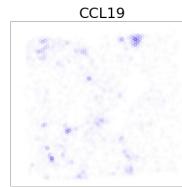
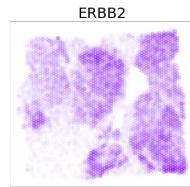
- Example with Human Breast cancer data
 - Public data : Available at 10x website



Facecolor intensity proportional
to gene expression value

■ ■ ■ Visualizing high dimensional spatial data

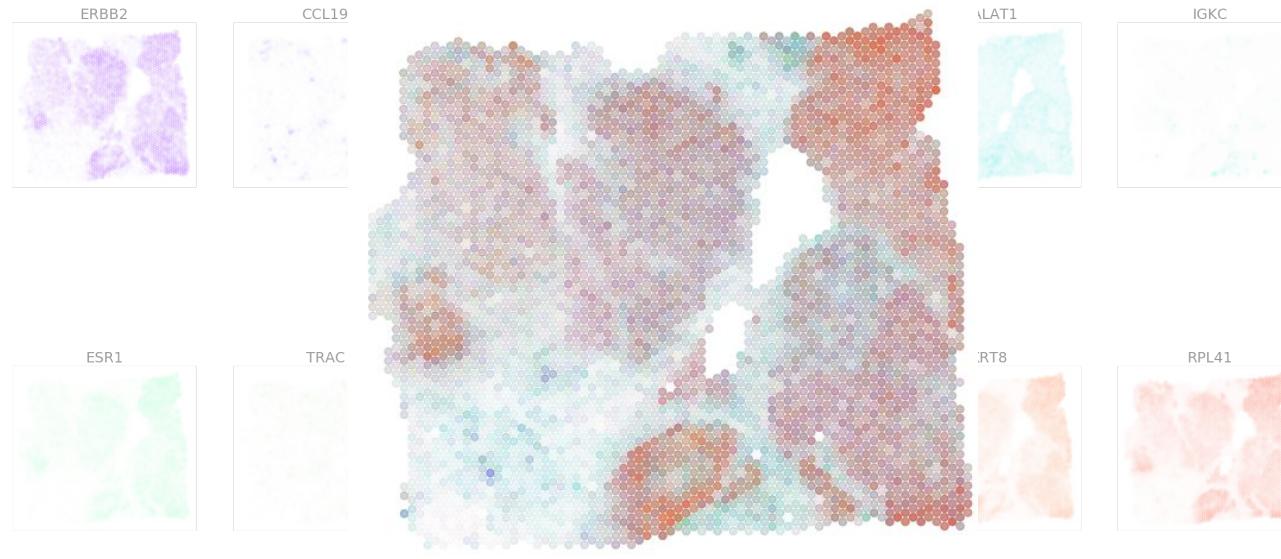
- We visualized one feature/gene (*ERBB2*)
- Q: How do we “visualize” ~20000 features?



⋮

■ ■ ■ Visualizing high dimensional spatial data

- We visualized one feature/gene (*ERBB2*)
- Q: How do we “visualize” ~20000 features?



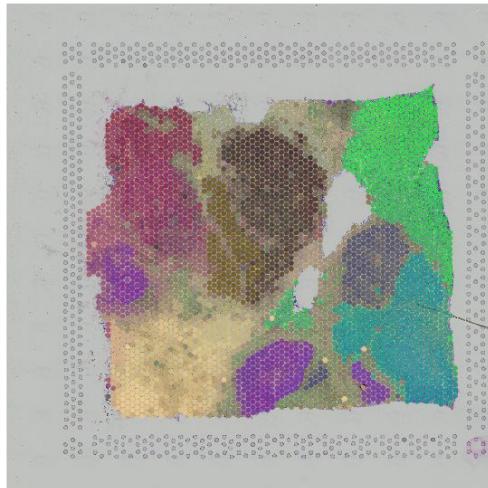
All in one image

■ ■ | Visualizing high dimensional spatial data

- We visualized one feature/gene (*ERBB2*)
- **Q:** How do we “visualize” ~20000 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*)
- Q: How do we “visualize” ~20000 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

■ ■ | 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



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 and mitochondrial genes tend to exhibit spurious expression patterns.

† Not always necessary

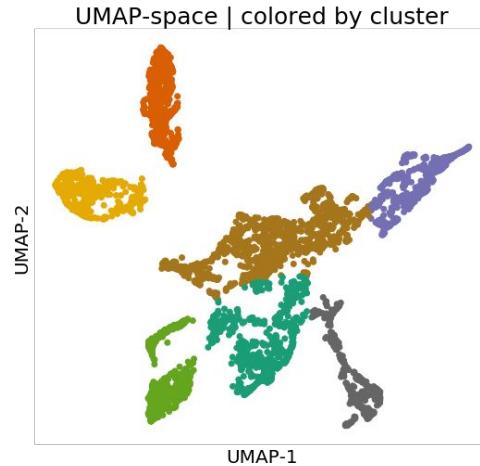
■ ■ | Filtering, Normalization, Batch correction, etc.

- 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 and mitochondrial genes tend to exhibit spurious expression patterns.
- Normalization / batch correction :
 - Recommend to account for spot “library size” - varying cell density
 - Include slide/array as covariate (sometimes big variation is observed)
 - Tools that have performed well : [sctransform](#) and [Harmony](#)

† Not always necessary

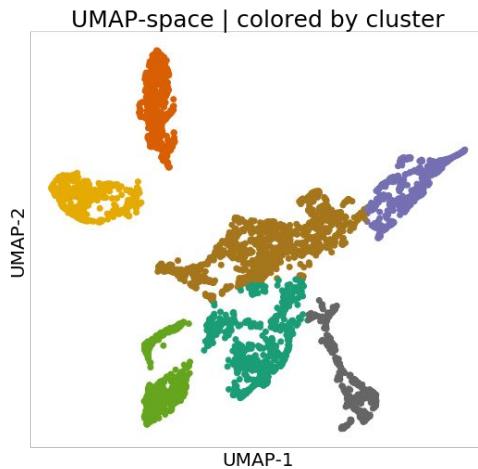
Example : Basic Analysis

- Cluster the spots based on gene expression



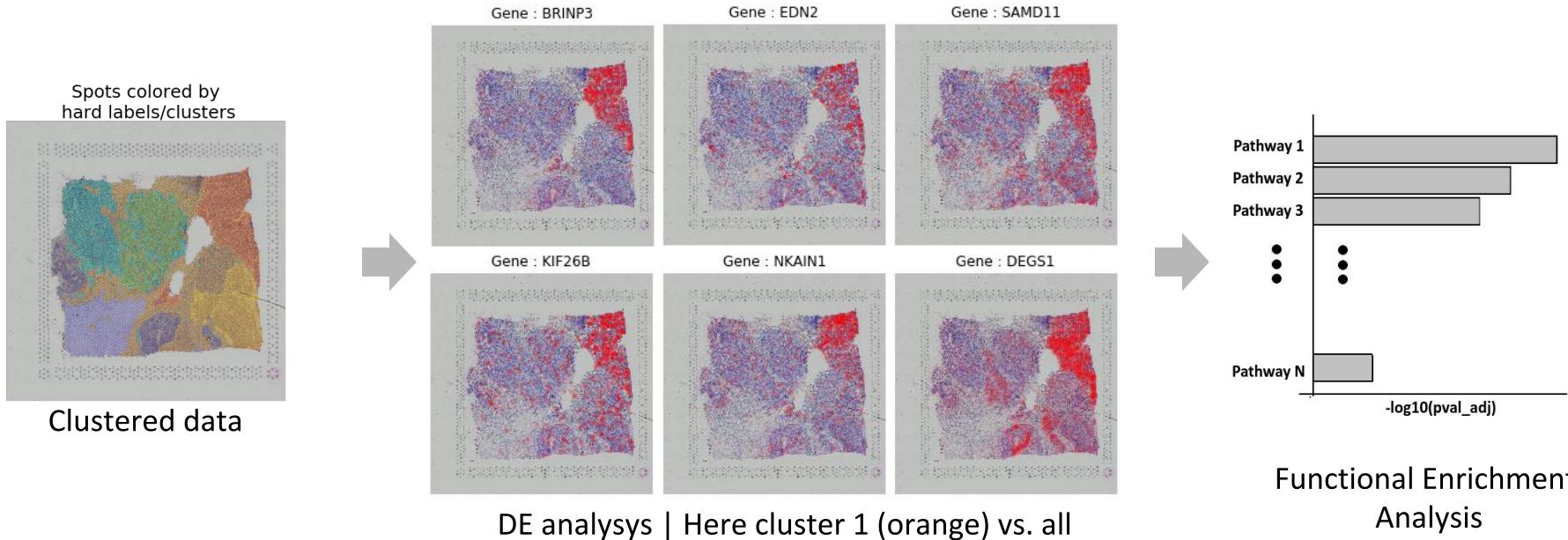
Example : Basic Analysis

- Cluster the spots based on gene expression
- Backmap clusters onto tissue
- Use HE-image as reference
 - Sanity check - does it make sense?
 - Valuable information resource
- Next, what do these clusters represent?



Example : Basic Analysis

A circular icon with a vertical line through the center, containing the text "P1".



Functional Enrichment Analysis

Example : Basic Analysis

Spots colored by hard labels/clusters



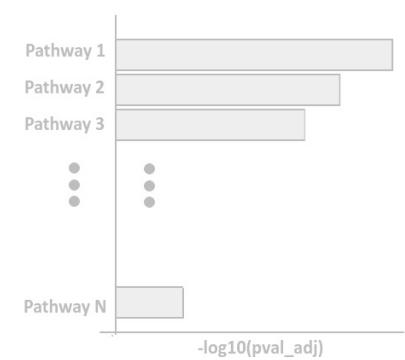
Clustered data

Spots colored by hard labels/clusters



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

All helpful in annotation of clusters



Functional Enrichment
Analysis

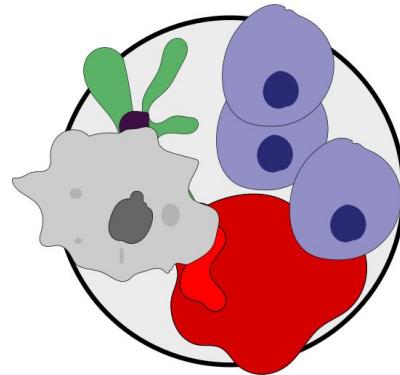


Example : Basic Analysis

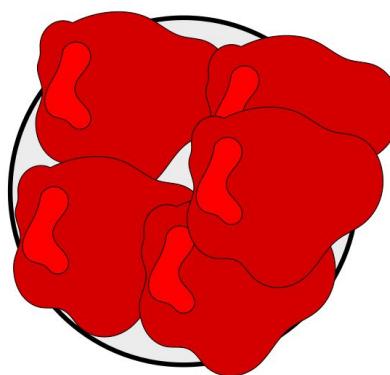
Remember :

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

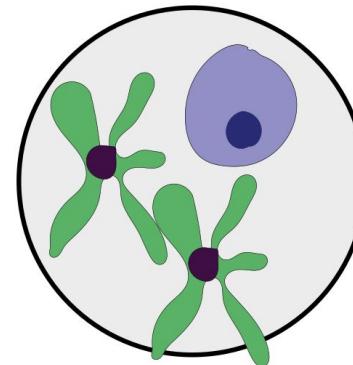
Spot 1



Spot 2

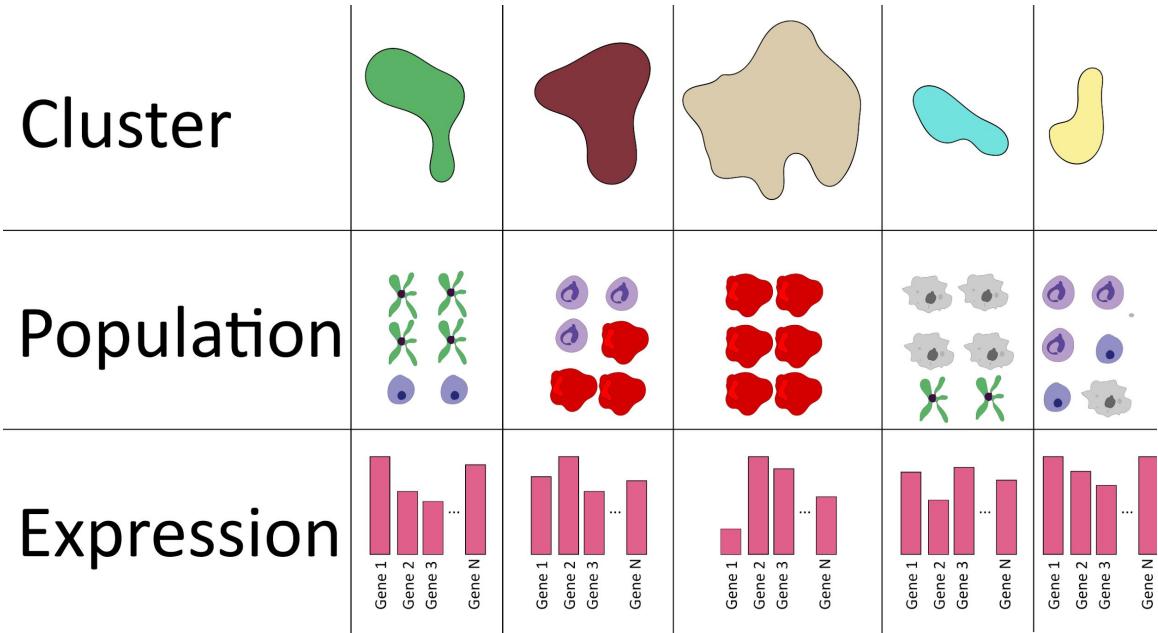
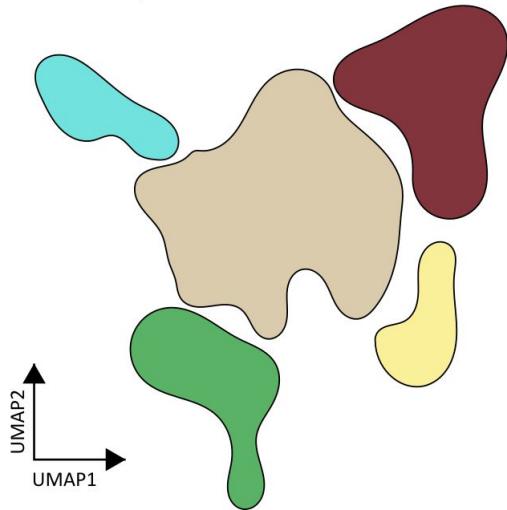


Spot 3



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?

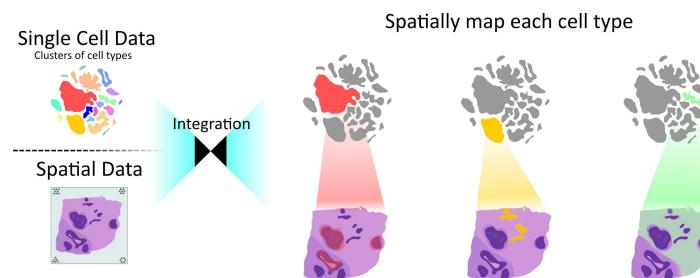
- Marker genes?

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!
 - Extract information of cell types from SC data and apply to spatial data



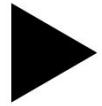
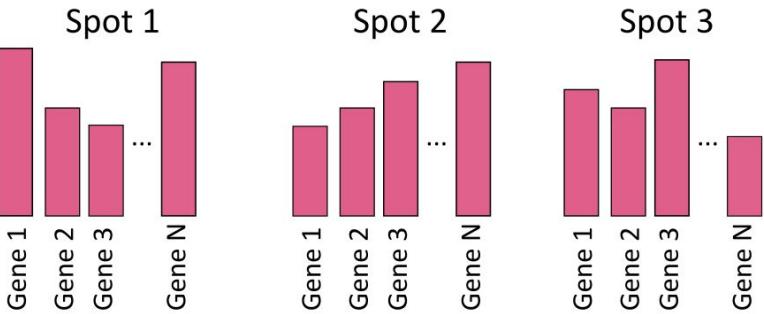
So where are my cell types?

- 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)**
 - How “much” of each cell type at each spot

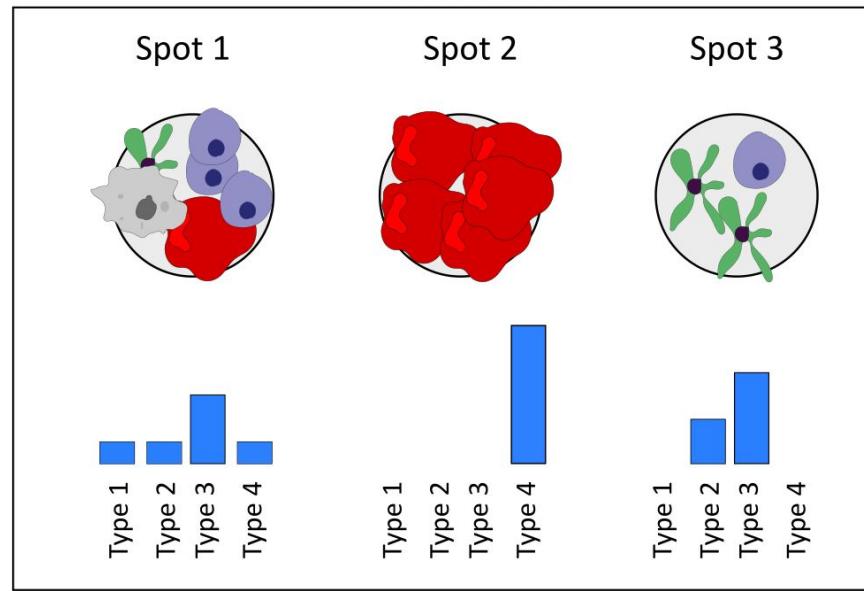


Our objective : deconvolve expression data

From this



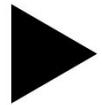
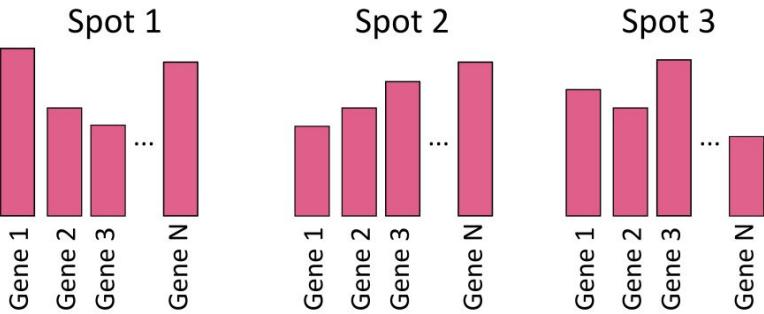
We want this



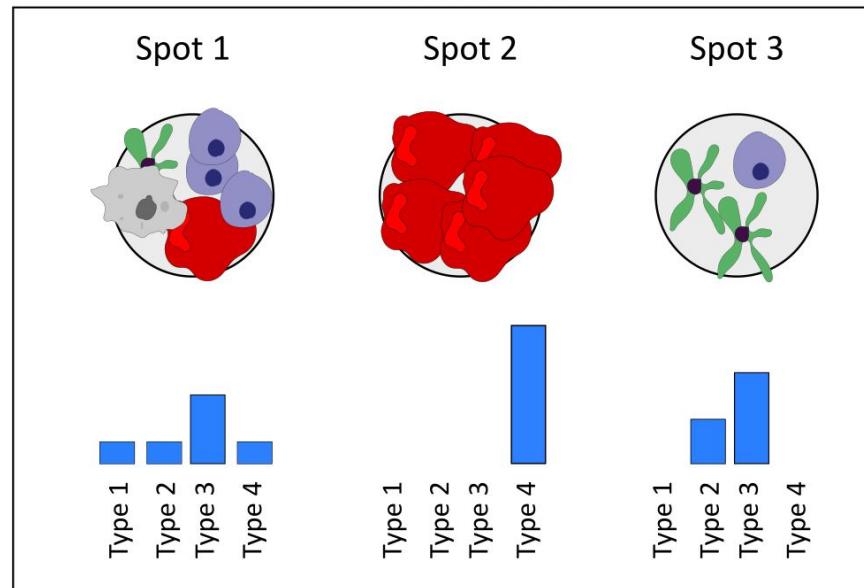


Our objective : deconvolve expression data

From this



We want this

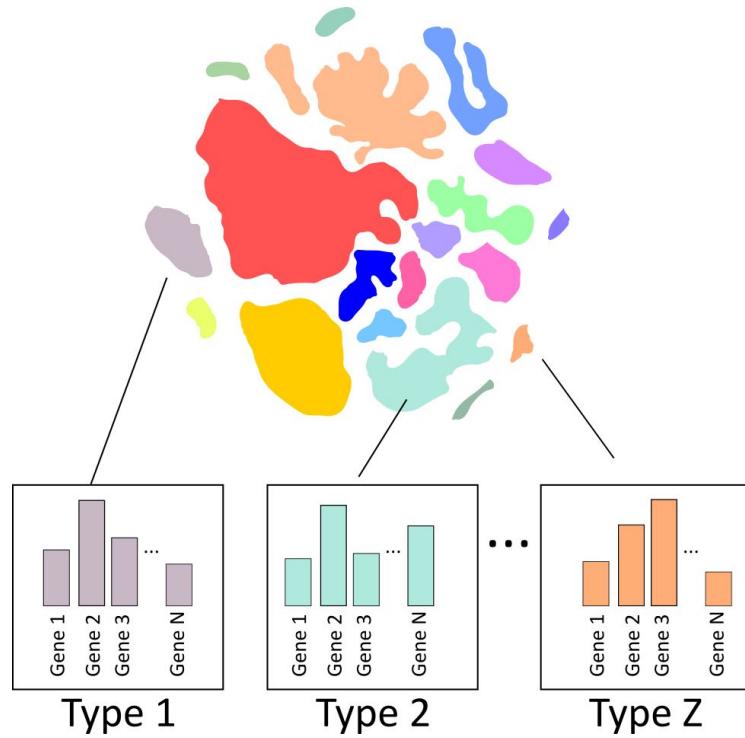


Suggested approach : *Model-based Probabilistic Inference*



It's as easy as 1-2-3

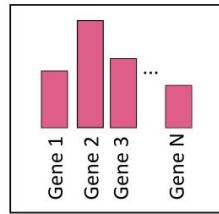
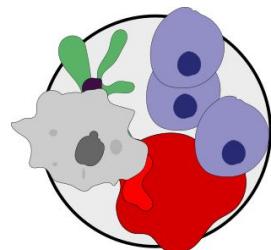
1. Infer cell type expression parameters from SC data





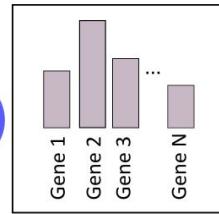
It's as easy as 1-2-3

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



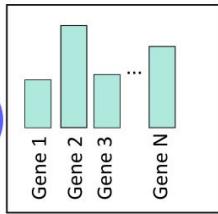
=

W_1



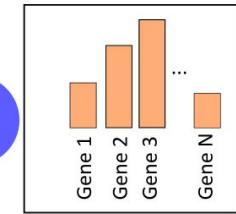
+

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



The machinery behind it

Underlying Assumption :

Single Cell and Spatial Transcriptomics Data can be modeled as
Negative Binomial Distributed

$$x \sim \mathcal{NB}(r, p)$$

$$\Pr(x = k) = \frac{\Gamma(k + r)}{\Gamma(k + 1)\Gamma(r)}(1 - p)^r p^k$$

The machinery behind it

Single Cell

Scaling factor

Cell type and gene specific rate

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

Expression of gene “g” in cell “c”.

“c” is of cell type “z”

The machinery behind it

Single Cell

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

Expression of gene “ g ” in cell “ c ”.

“ c ” is of cell type “ z ”

Spatial

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

Expression of gene “ g ” at spot “ s ”
from cell “ c ”.

“ c ” is of cell type “ z ”

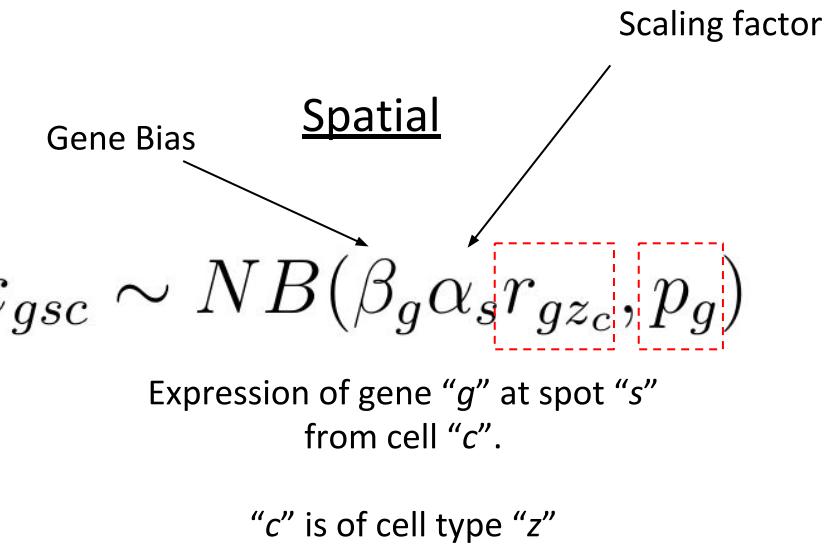
The machinery behind it

Single Cell

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

Expression of gene “ g ” in cell “ c ”.

“ c ” is of cell type “ z ”



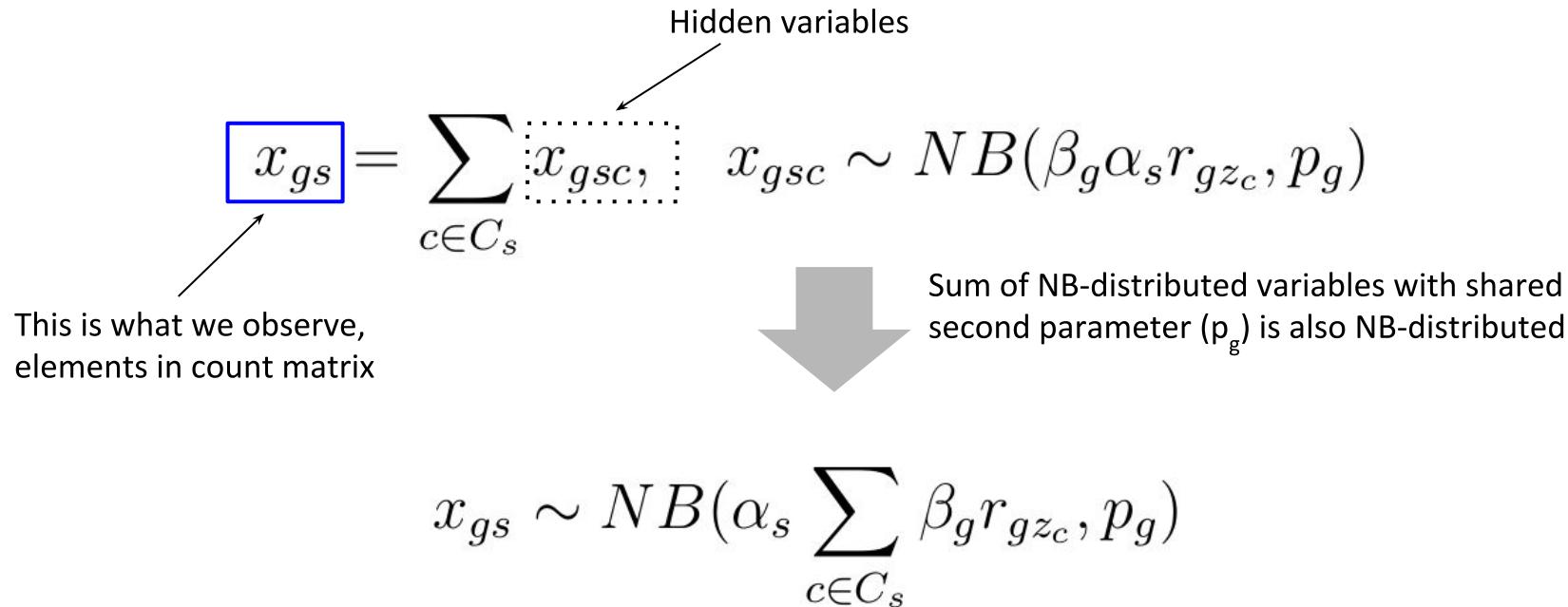
The machinery behind it

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

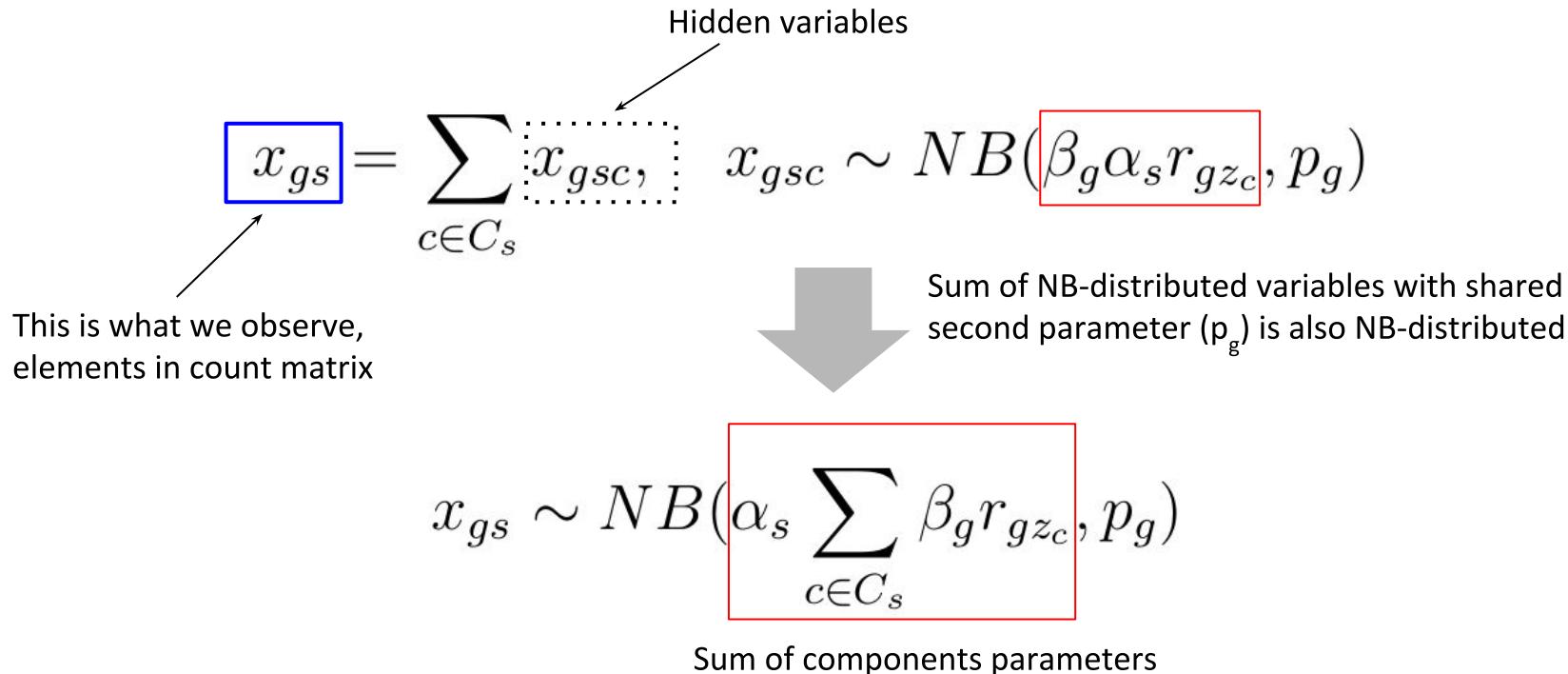
Hidden variables

This is what we observe,
elements in count matrix

The machinery behind it



The machinery behind it



The machinery behind it

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



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”

The machinery behind it

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



join scaling factor (α_s) and cell counts (n_{sz}).
unadjusted proportions
 $v_{sz} = \alpha_s n_{sz}$

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

The machinery behind it

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

Use MLE to get estimates of cell type parameters



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

Given the cell type parameters, use MLE to get values of gene scaling factors (β_g) and unadjusted proportions (v_{sz})

The machinery behind it

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

The machinery behind it

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

Number of cells from cell type “z” at spot “s”

n_{sz}

Total number of cells at spot “s”

A diagram illustrating the components of the formula. A blue dashed box highlights the term n_{sz} , with an arrow pointing to it from the text "Number of cells from cell type ‘z’ at spot ‘s’". A green dashed box highlights the entire denominator $\sum_{z \in Z} n_{sz}$, with an arrow pointing to it from the text "Total number of cells at spot ‘s’".

The machinery behind it

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

Proportion of cell type "z" at spot "s"

The machinery behind it

- Probabilistic Model : models data as NB distributed
- Tool : *stereoscope*
- Output : [spot] x [cell_type] matrix
 - Elements are proportion of cell belonging to the given cell type at each spot
- Preprint : <https://doi.org/10.1101/2019.12.13.874495>

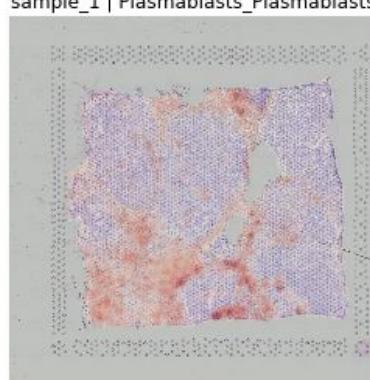
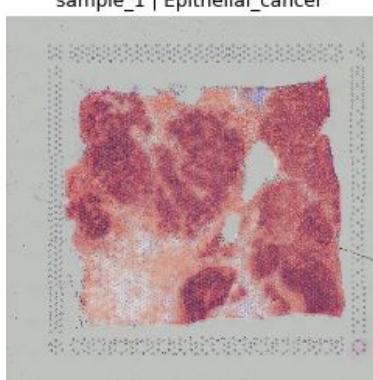
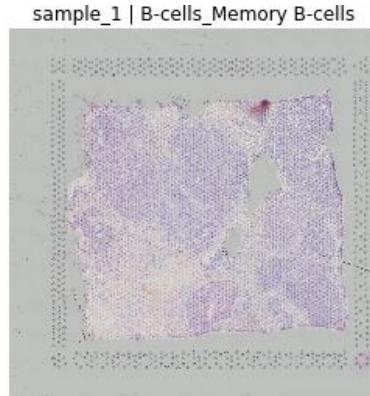
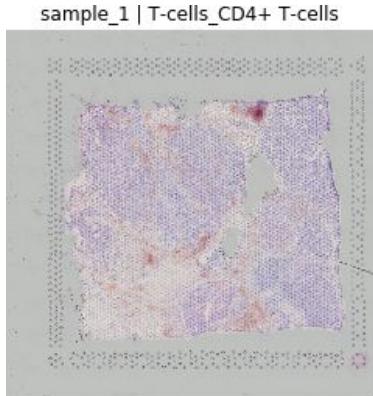


<https://github.com/almaan/stereoscope>

#shameless self-advertising

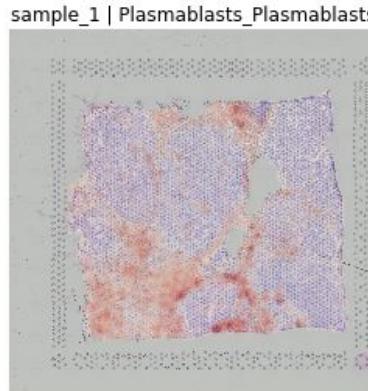
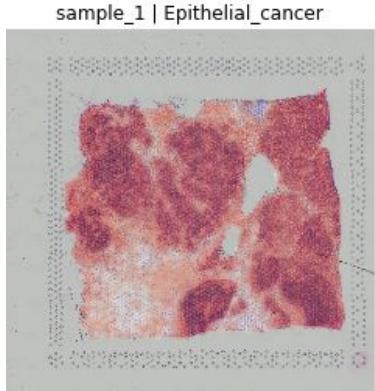
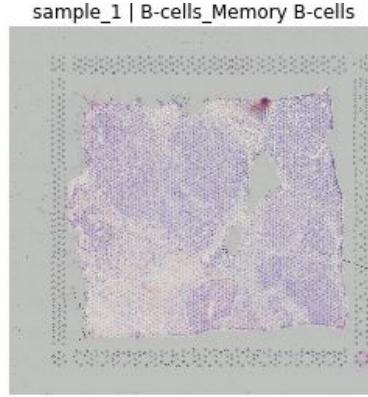
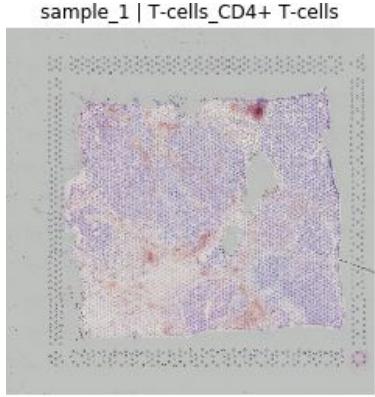
■ ■ ■ Applying it to our breast cancer data

Proportion estimates

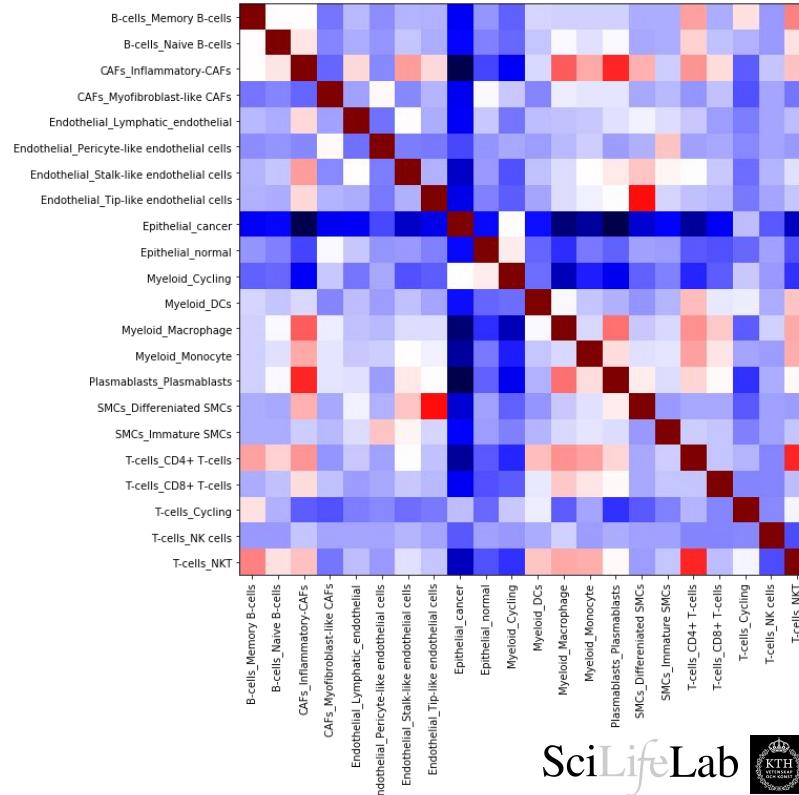


Applying it to our breast cancer data

Proportion estimates



Cell type co-localization



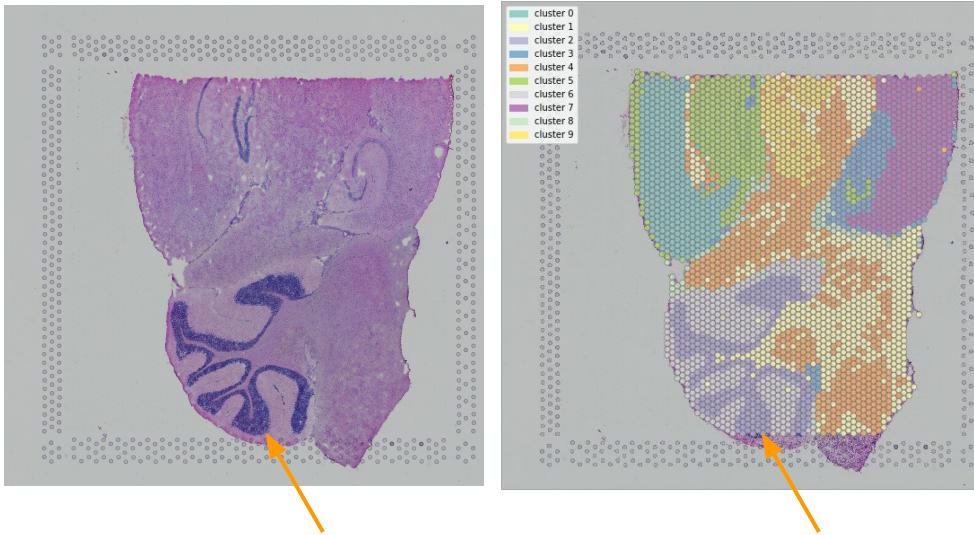


Summary : Integration of single cell and spatial data

- 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
- Until experimental techniques reach single cell resolution
- Atlases are exciting!

Example Analysis : Expression as a function of distance

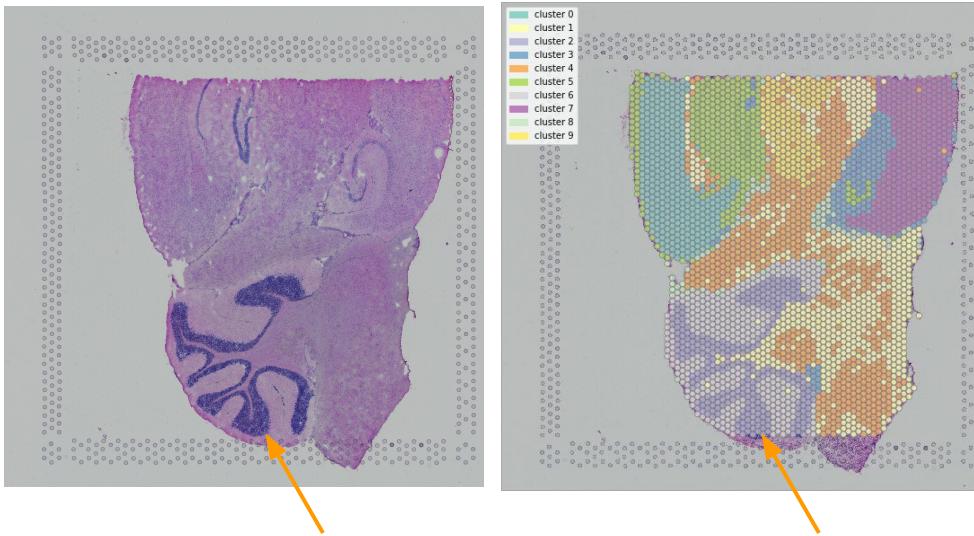
- Say we cluster spatial data and find an interesting domain (e.g. cluster 2)



Data from 10X Genomics website

Example Analysis : Expression as a function of distance

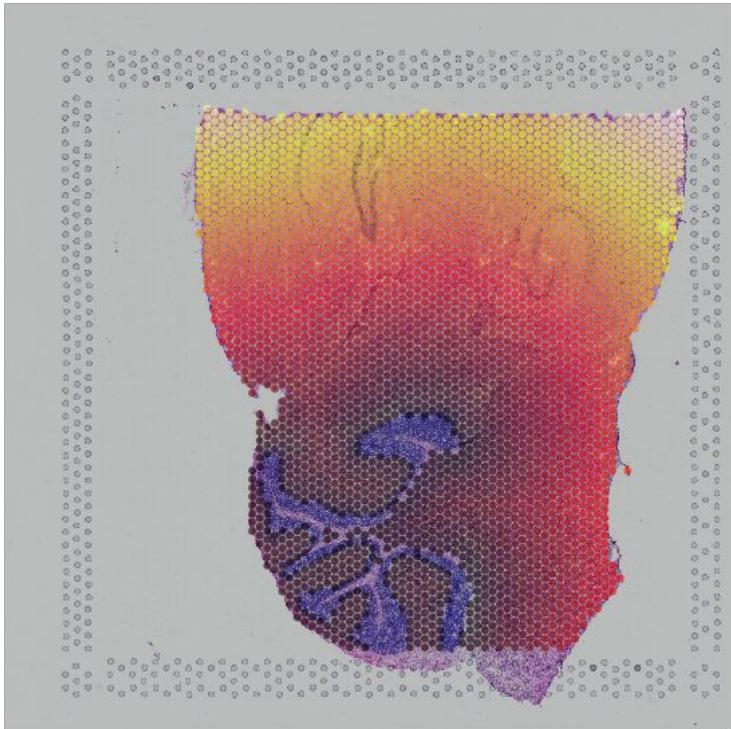
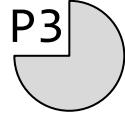
- Say we cluster spatial data and find an interesting domain (e.g. cluster 2)



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



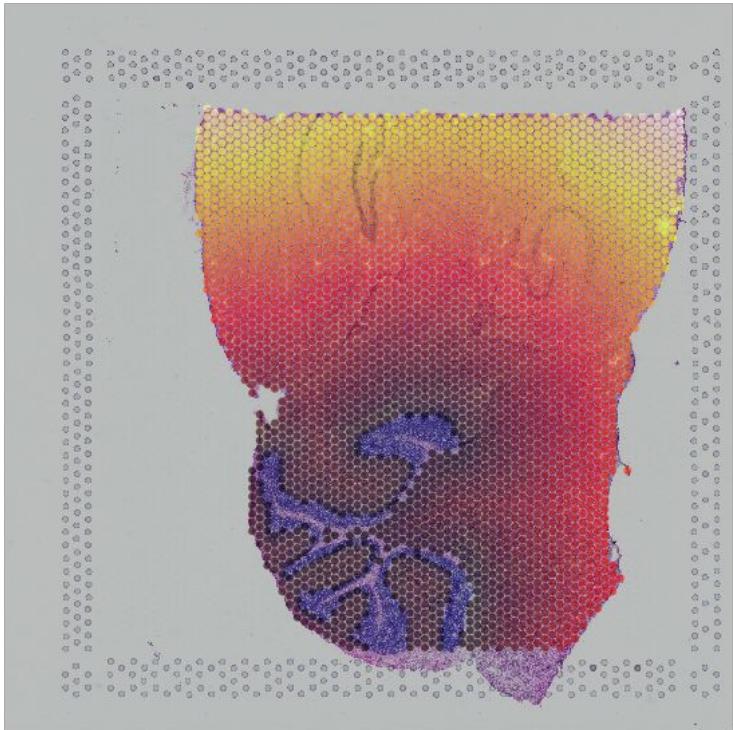
Example Analysis : Expression as a function of distance



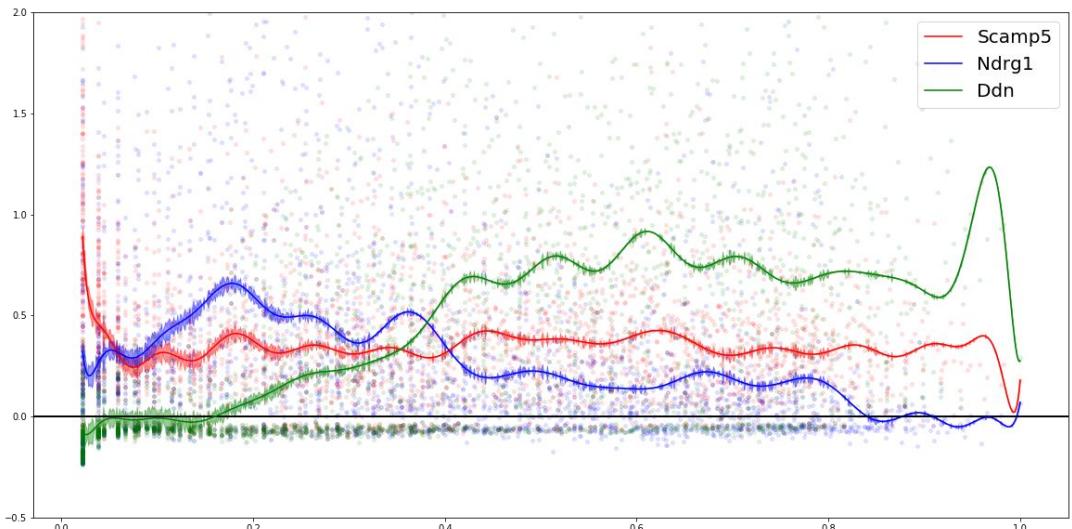
Spots colored by distance do cluster 2



Example Analysis : Expression as a function of distance

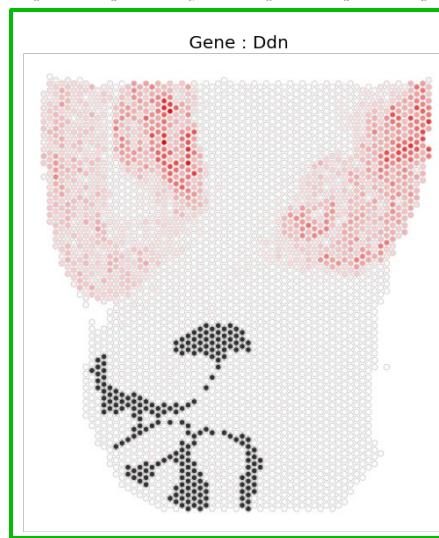
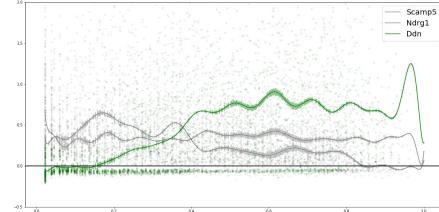
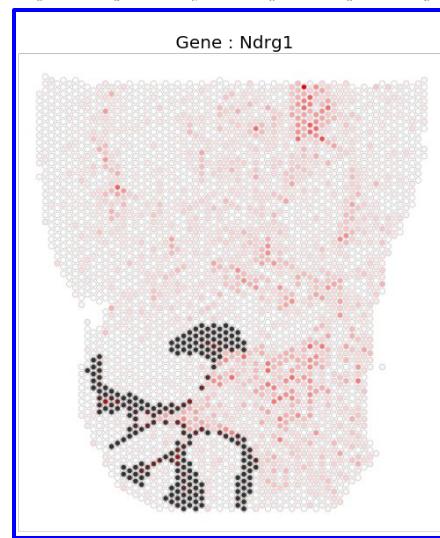
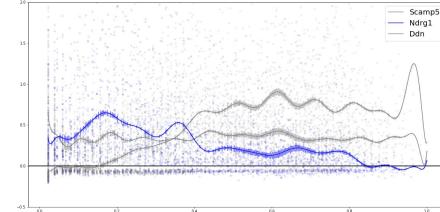
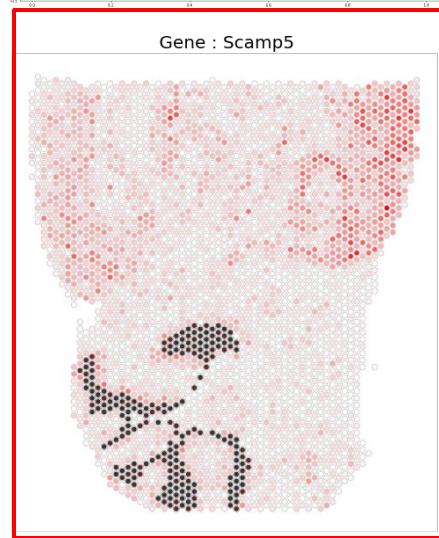
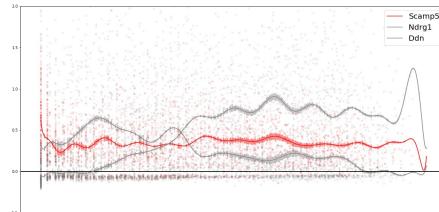


Spots colored by distance do cluster 2



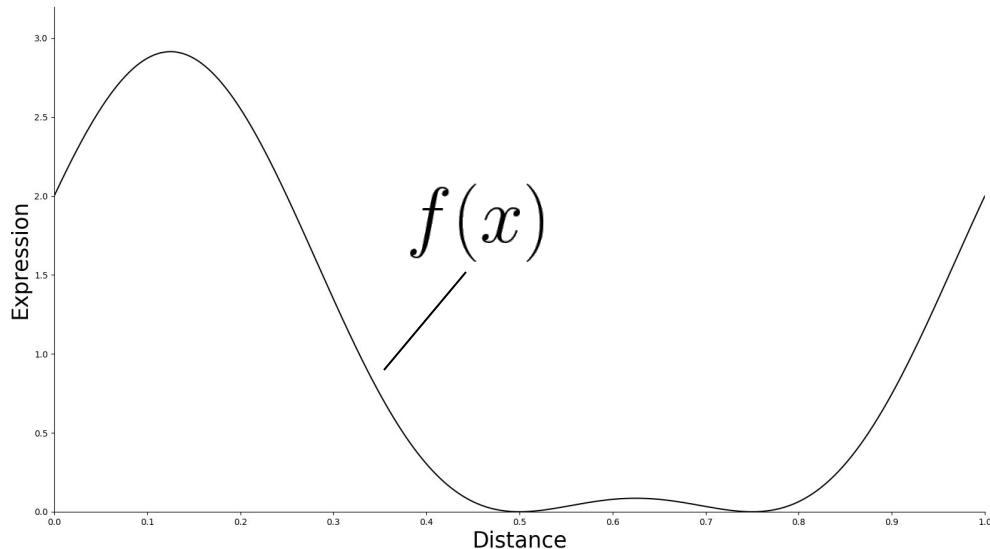
Gene expression as a function of the distance
to cluster 2

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?”



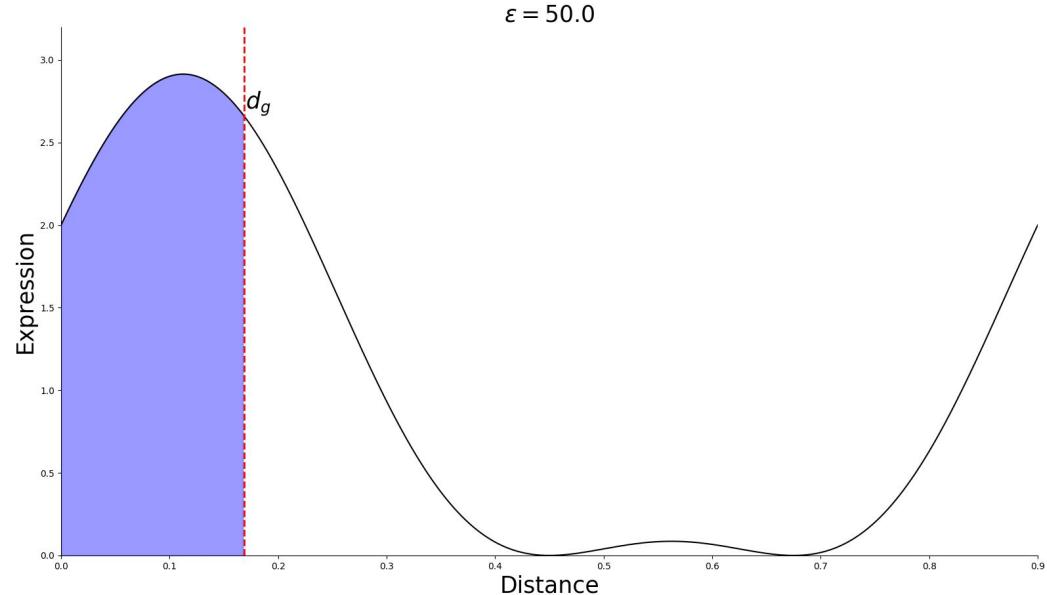
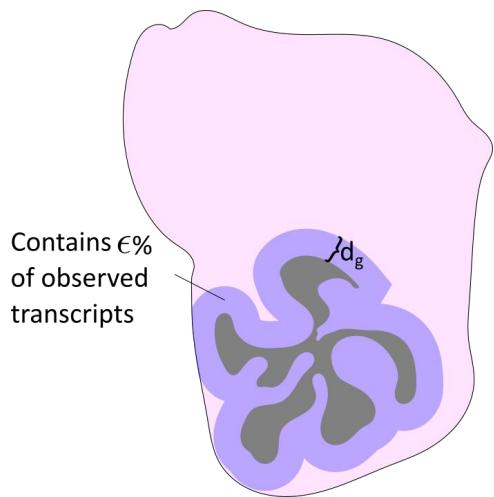
Equivalent to solving :

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

w.r.t. d_g

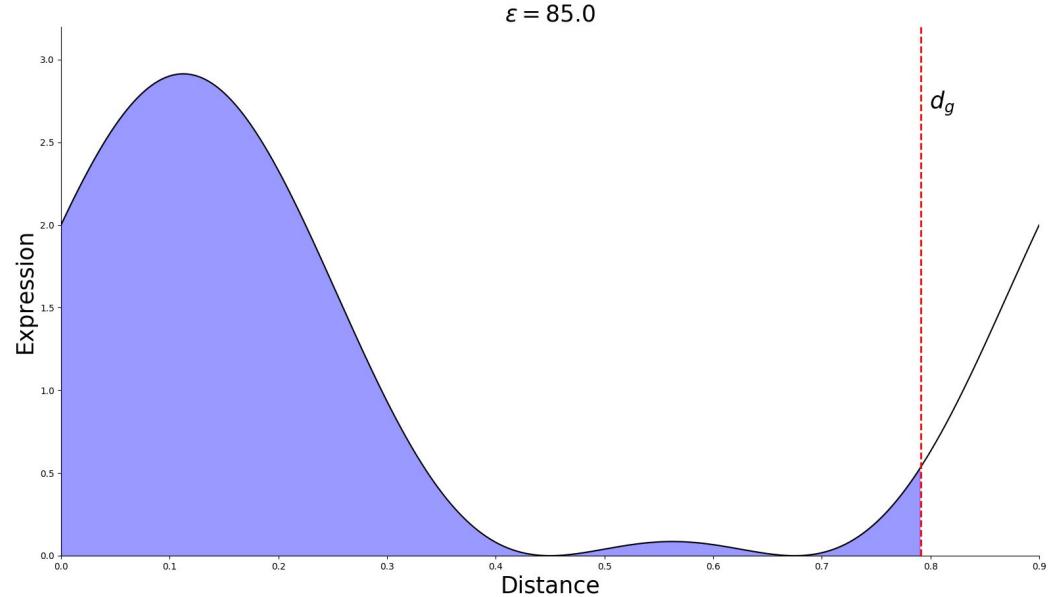
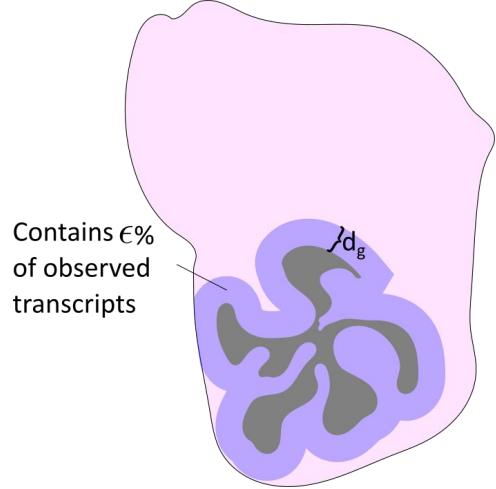
Example Analysis : Expression as a function of distance

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



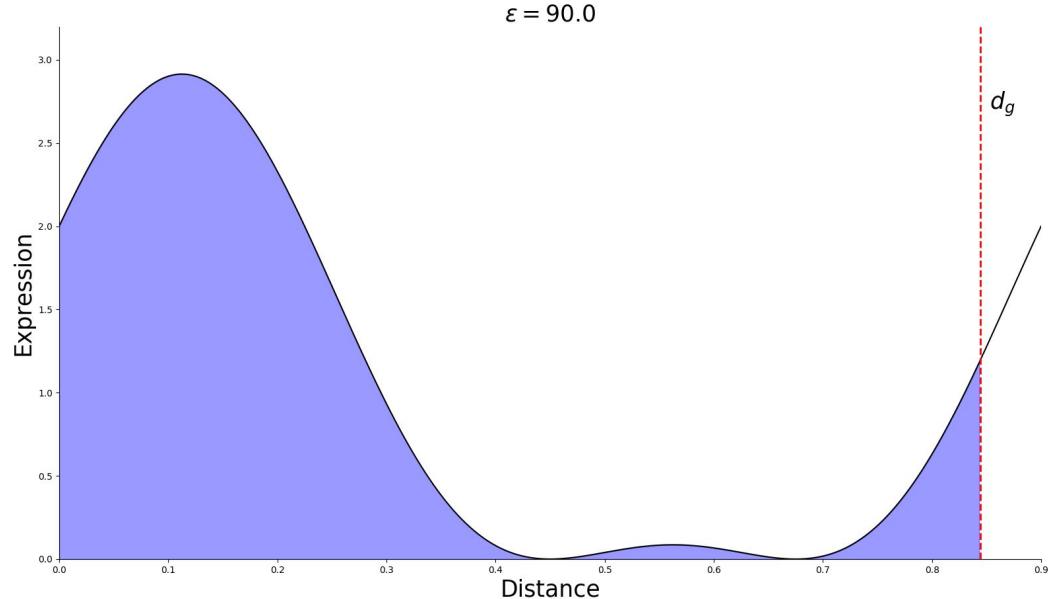
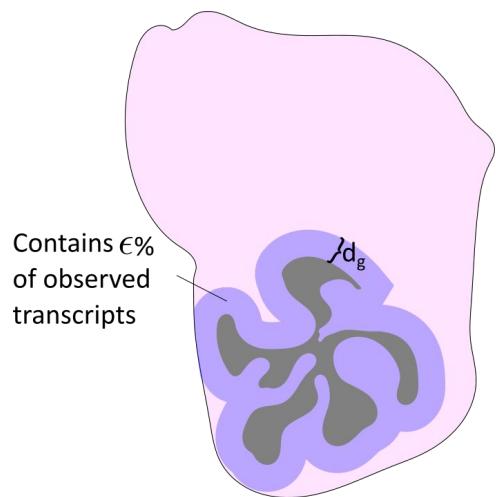
Example Analysis : Expression as a function of distance

- 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?”

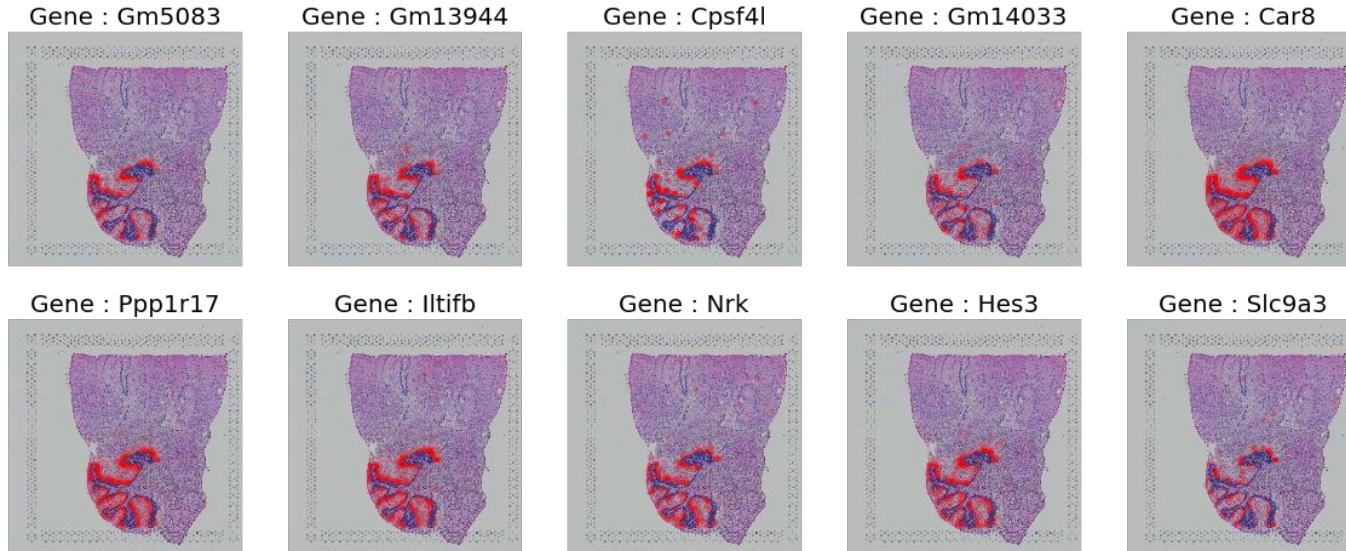


■ ■ ■ Example Analysis : Expression as a function of distance

- 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
- 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