

# An orientation in the spatial transcriptomics landscape

Alma Andersson  
2021-01-28



<https://github.com/almaan>



<https://almaan.github.io>



<https://www.spatialresearch.org>

# Who am I?

- **Name** : Alma Andersson
- **Part of** : Lundeberg Lab (PhD Student)
- **Works with** : Computational Method development
  - Mainly focus on spatial data
- **Background** :
  - Engineer by training
  - Molecular Dynamics
    - Ion channels (Delemotte Group)
  - Spatial Transcriptomics
- **Interests** :
  - Statistical modelling
  - Machine learning
  - Evolutionary algorithms
  - Running



# My vision for today

## Experimental spatial transcriptomics techniques

- Broad overview of techniques
- Common themes
- Data produced

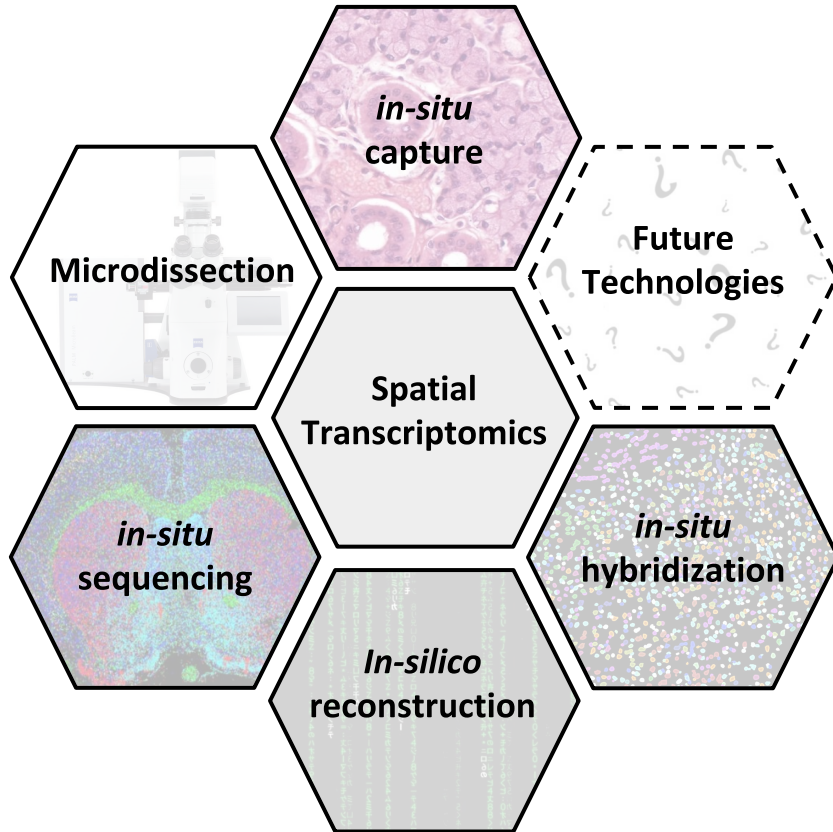
## Computational methods for analysis of spatial transcriptomics data

- Different flavors of methods
- Examples of relevant analysis
- Extra focus on single cell mapping and integration

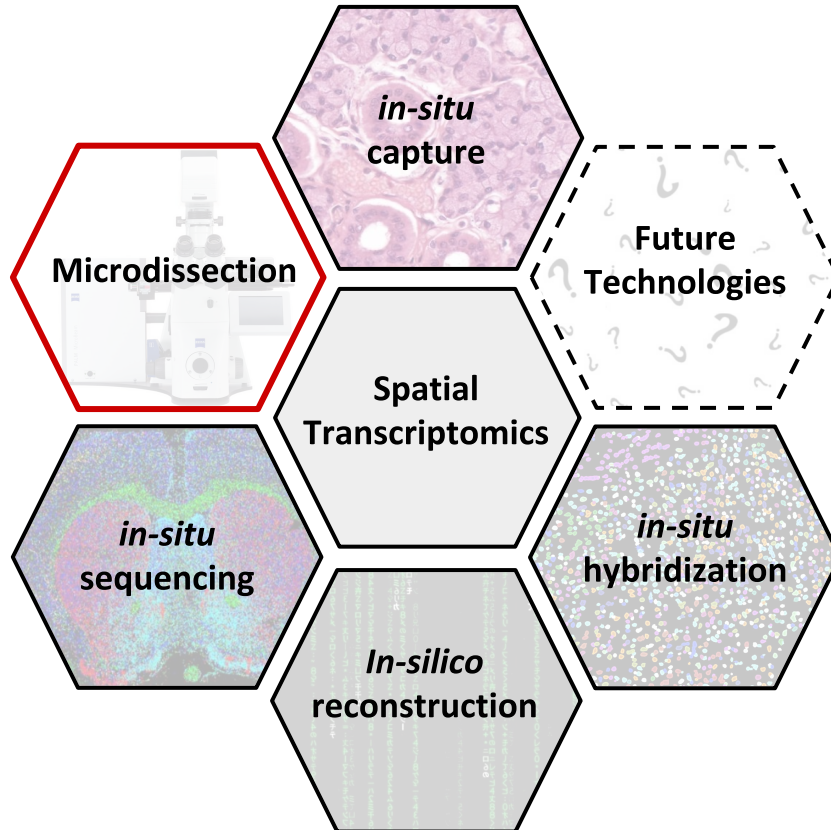
## Visium and Spatial Transcriptomics data

- Clearing up some confusion, ST vs. Visium?
- Visium specs and some brief words of advice on the analysis

# Experimental techniques



# Experimental techniques



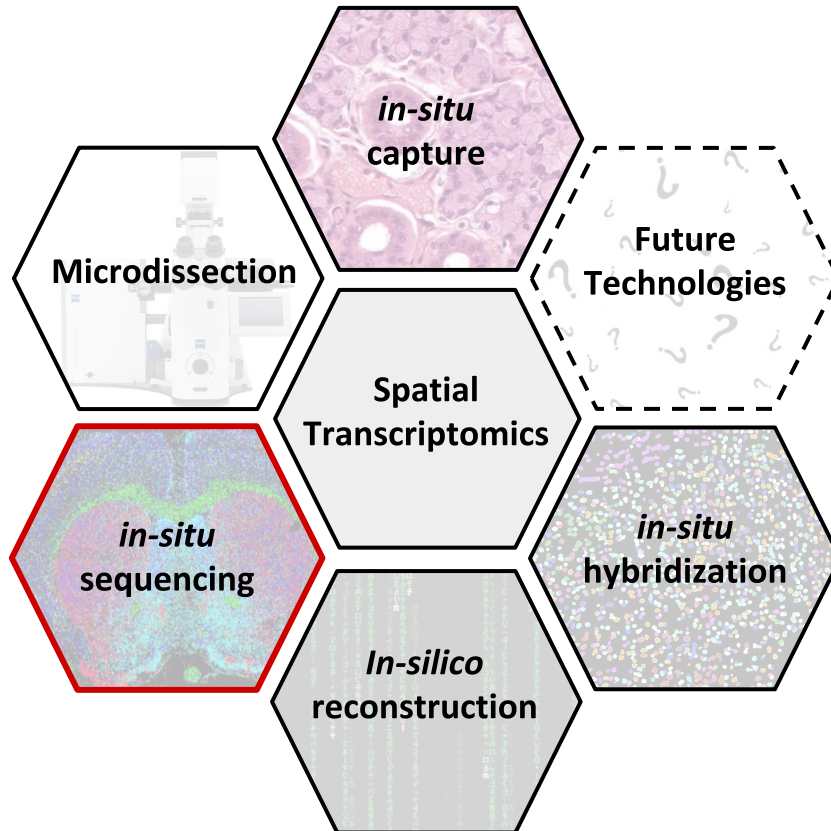
## Microdissection-based technologies

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

A “Brute Force” approach.

**Examples :** LCM, Tomo-seq, TIVA, ProximID, Niche-seq

# Experimental techniques



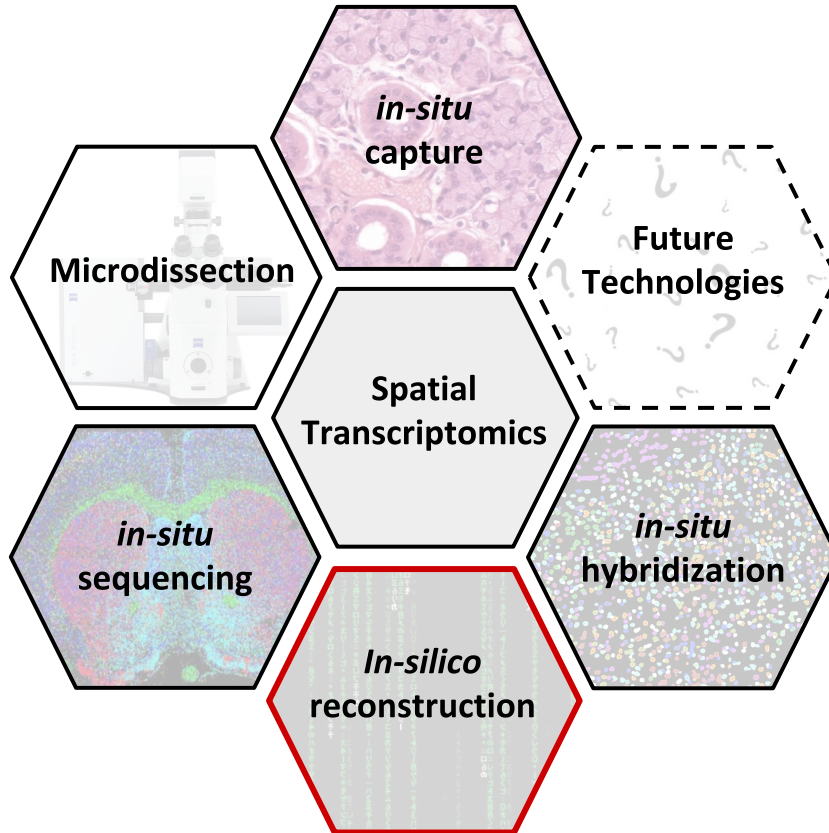
## *In-situ* sequencing based methods

Sequence the transcripts in place.

Offer sub-cellular resolution. Some relies on “*a priori*” defined targets, but not all.

**Examples :** ISS/Cartana (padlock probes), BaristaSeq, STARmap, FISSEQ

# Experimental techniques

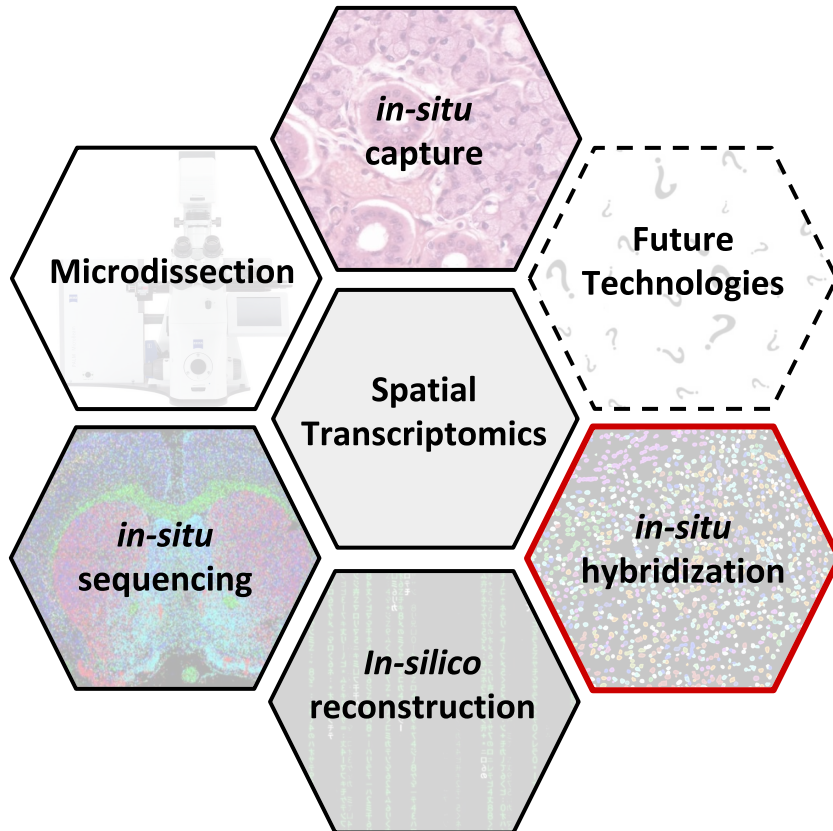


## *In-silico* reconstruction

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

**Examples :** novoSpaRc, CSOmap, Seurat v3

# Experimental techniques



## *In-situ* hybridization based methods

Labeled probes for specific targets, hybridize in place.

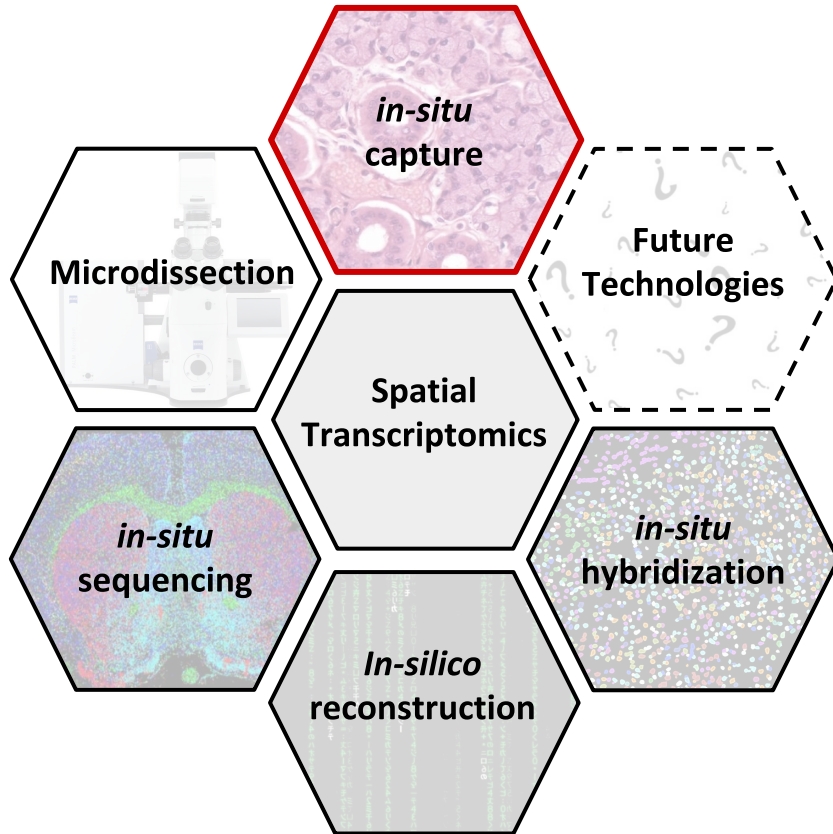
Requires “*a priori*” defined targets.

Expansion strategies and smart decoding scheme has helped to overcome spectral overlap.

**Examples** : smFISH, seqFISH, MERFISH, seqFISH+, osmFISH, RNA Scope, DNA microscopy



# Experimental techniques

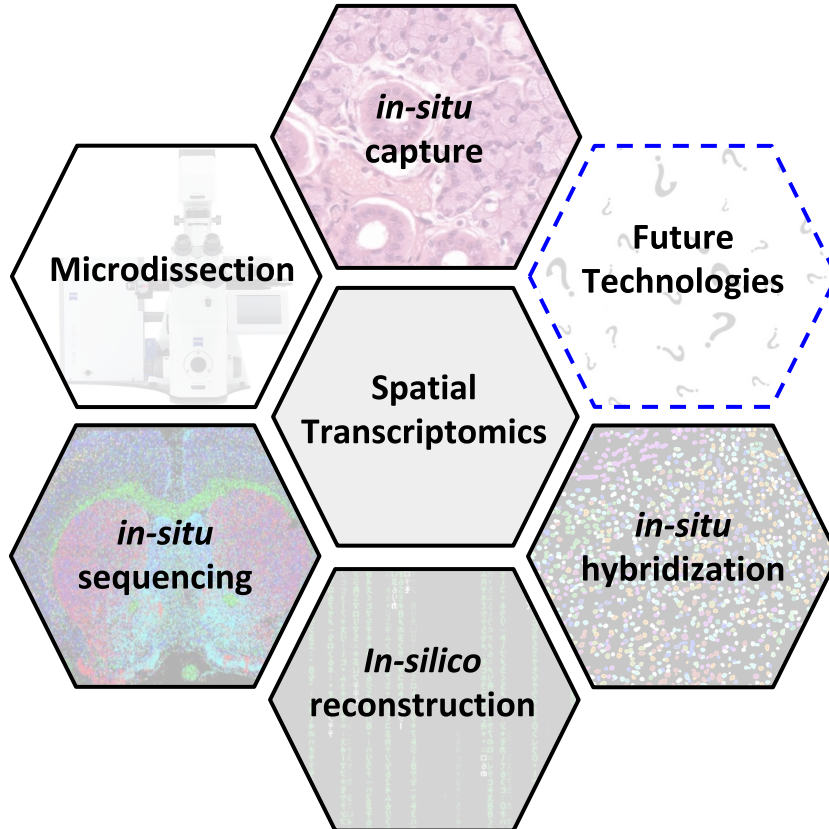


## *In-situ* capture based methods

Capture transcripts *in situ* but sequence *ex situ*.  
Usually less dependent on prior selection of targets.

**Examples** : Visium, ST, Slide-Seq, HDST, GeoMX, Apex-Seq, **Stereo-SEQ**

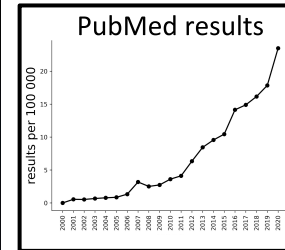
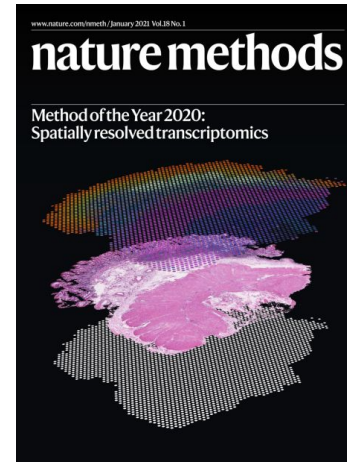
# Experimental techniques



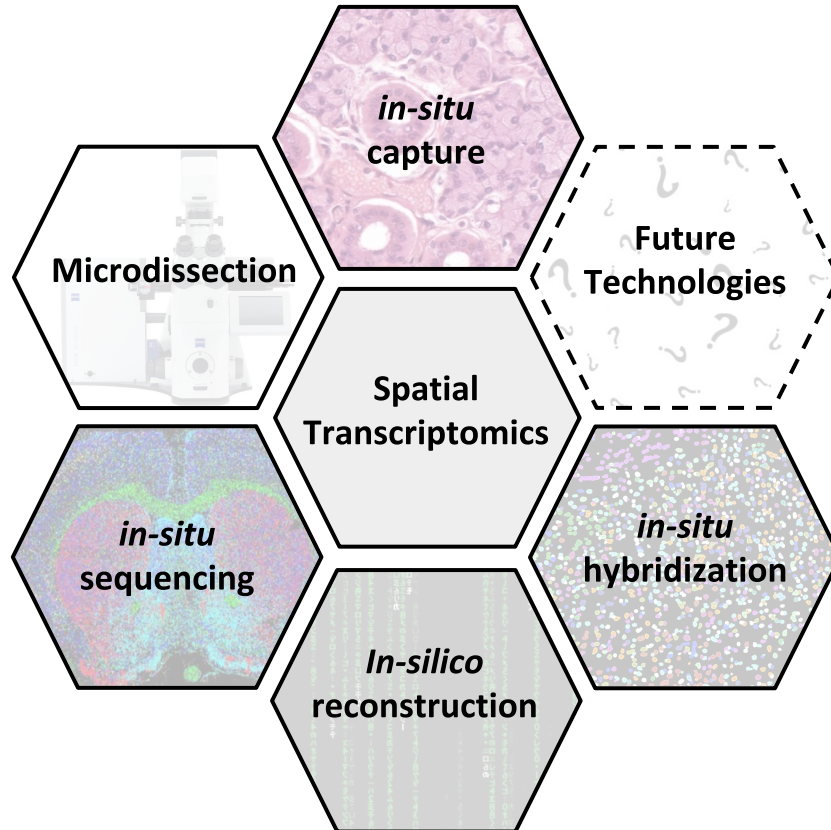
## AGBT 2020: DeciBio Highlights – Spatial Profiling Reloaded & MGI Unleashed

Posted on 02.27.2020 // Categories: [Genomics Market](#), [Life Science Market Research](#), [Uncategorized](#)

**Marco Island, FL February 27, 2020** – AGBT's 20<sup>th</sup> Anniversary didn't disappoint, with many provocative sessions and product launches. In many ways, #AGBT20 mirrored the #AGBT19 vintage! This year, however, the majority of the 56 attendees we interviewed immediately highlighted spatial profiling or MGI as stealing the show. Please refer to our 2019 entry for additional commentary, as these previous trends remained relevant of this conference (e.g., biology takes center stage, NGS continued industrialization).



Spatial Transcriptomics



## Further Readings

***Spatially Resolved Transcriptomes—Next Generation Tools for Tissue Exploration***

**Authors :** Michaela Asp, Joseph Bergensträhle, Joakim Lundeberg

**Published :** 2020-05-04

**DOI:** [10.1002/bies.201900221](https://doi.org/10.1002/bies.201900221)

***Method of the Year 2020: spatially resolved transcriptomics***

**Authors :** Editorial

**Published :** 2021-01-06

**DOI:** [10.1038/s41592-020-01042-x](https://doi.org/10.1038/s41592-020-01042-x)

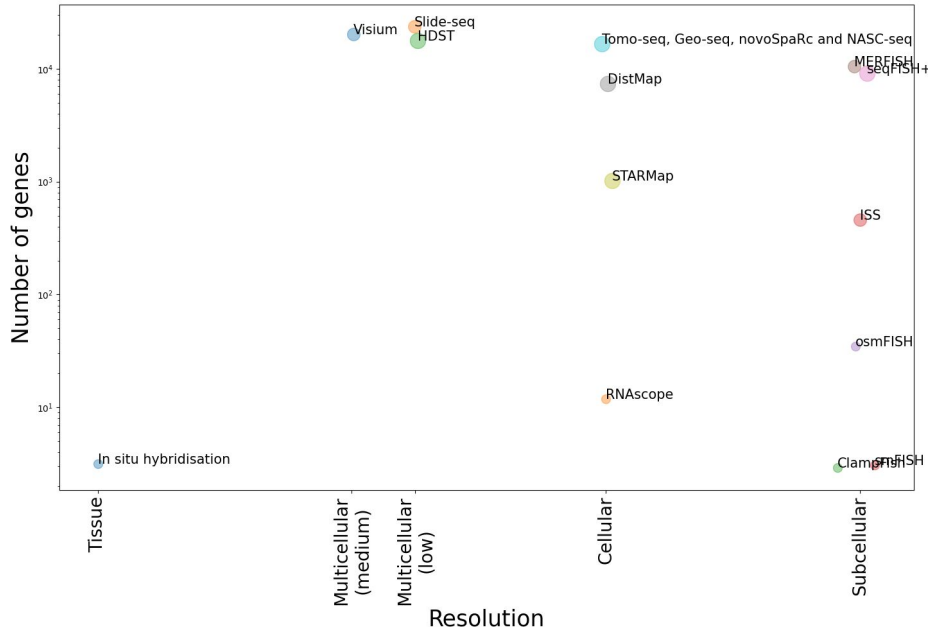
***Spatially resolved transcriptomics adds a new dimension to genomics***

**Authors :** Ludvig Larsson, Jonas Frisé & Joakim Lundeberg

**Published :** 2021-01-06

**DOI:** [10.1038/s41592-020-01038-7](https://doi.org/10.1038/s41592-020-01038-7)

# So which technique is best?



Only 2 sides of a multidimensional coin.

Other things to keep in mind are:

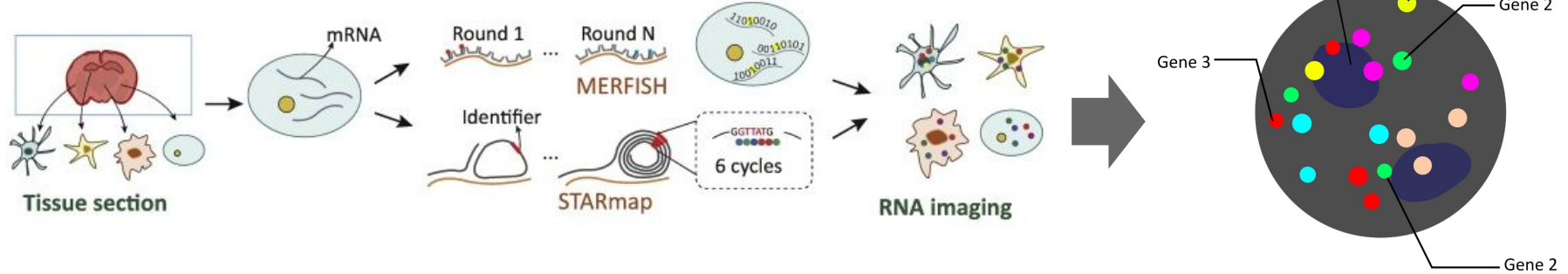
- Area covered
- Targeted or not
- Cost
- Ease of execution
- Time of execution
- Reproducibility

Currently some common rules of thumb:

- Inverse relationship between throughput and resolution
- Commercial products expensive, but robust and fast
- Capture based methods introduce certain spatial bias w.r.t. locations

# What you get in the end

## in-situ sequencing and hybridization



## in-situ capture

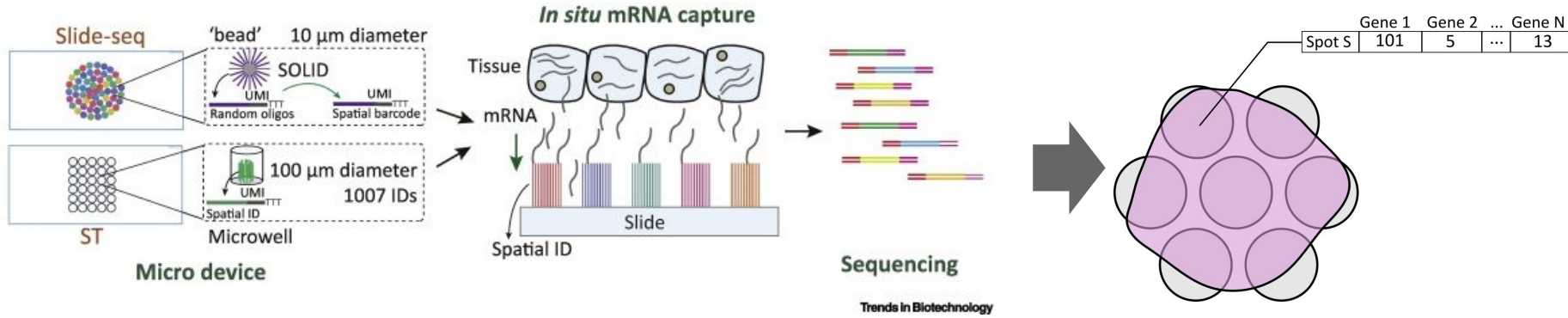
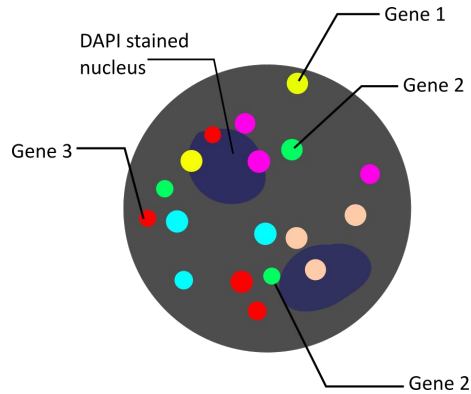


Figure adapted from : *Uncovering an Organ's Molecular Architecture at Single-Cell Resolution by Spatially Resolved Transcriptomics*, Liao et al

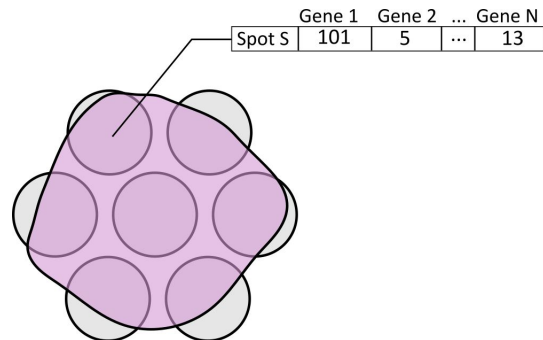
# What you get in the end

## in-situ sequencing and hybridization



- Exact location of targets
- Data Processing often includes :
  - Decoding of signal (which transcript)
  - Cell segmentation
  - Assignment of transcript to cell
  - (Cell type calling)
- Often presented as [cell]x[gene] matrix in the end

## in-situ capture

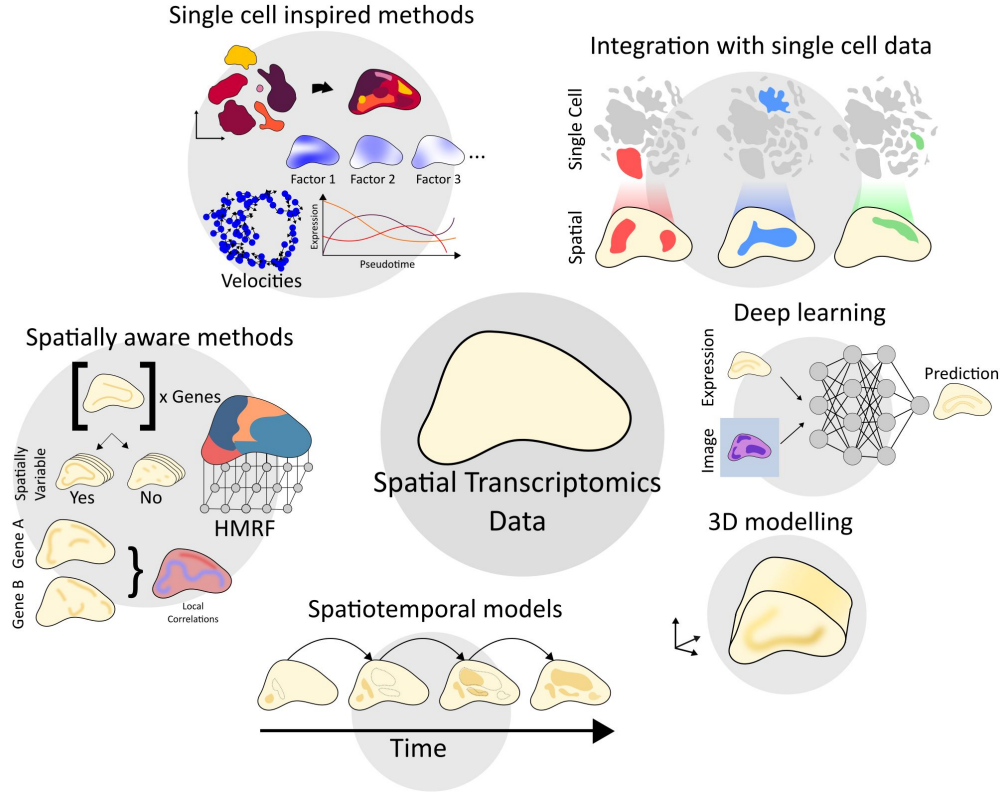


- Mini “bulk” average expression at each spot
- Data Processing often includes :
  - Genome mapping and annotation
  - Spatial barcode demultiplexing
    - Which site does each transcript originate from
- Often presented as [spot] x [gene]

A grayscale topographic map with contour lines and numerical elevation values. The map features a central peak with a value of 1520 and several other peaks and valleys. The text 'Computational Analysis' is centered over the map.

# Computational Analysis

# A motley crew of diverse methods

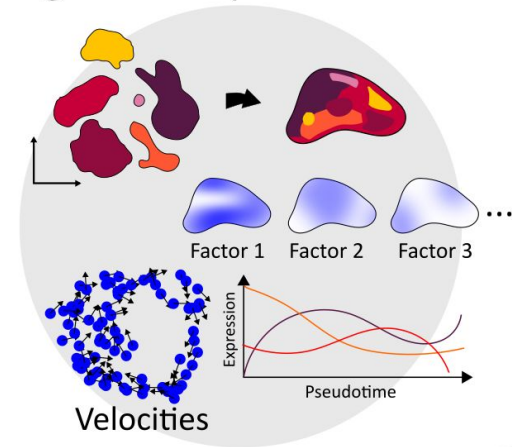




# Single Cell Inspired methods

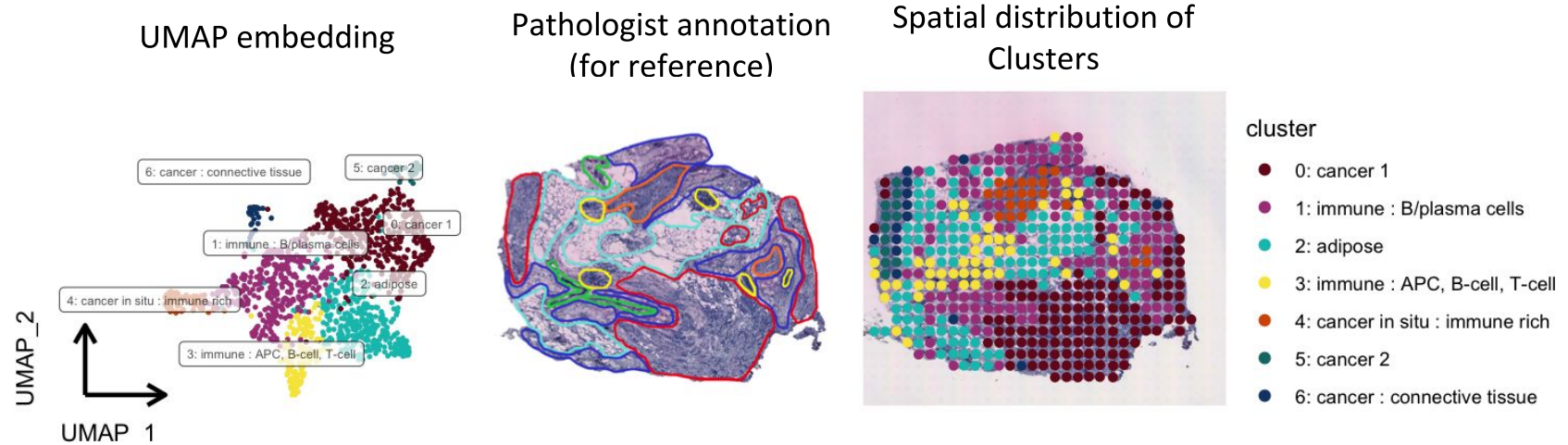
- **Basic idea** : apply existing methods and tools developed for SC data. Fine tune to make them more suitable for spatial data
- **Examples** :
  - Cluster spatial data, show clusters in space
  - Decompose expression profiles using factor models
  - Trajectory Inference :
    - Alt 1 : treat as single cell data
    - Alt 2 : reconstruct algorithm
- **Suites/Tools:**
  - Seurat : added support for spatial data
  - Scanpy : added support for spatial data
  - STUtility : built on Seurat tailored for spatial data
  - stLearn : built on scanpy tailored for spatial data
  - SpatialExperiment : (similar to SingleCellExperiment in R)
  - And many more...

## Single cell inspired methods



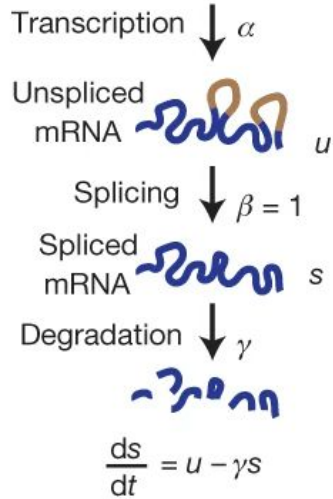
# Example clustering | Human Breast Cancer (ST1K)

G



# Trajectory Inference

Original  
(La Manno et.al)



Modified  
(Xia et.al)

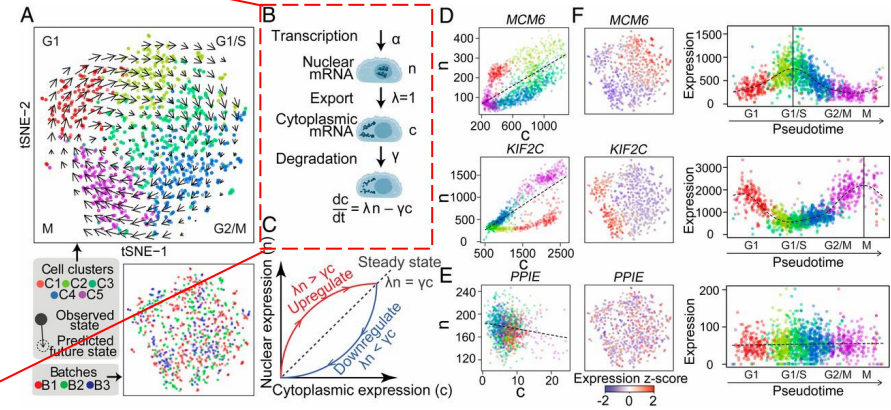
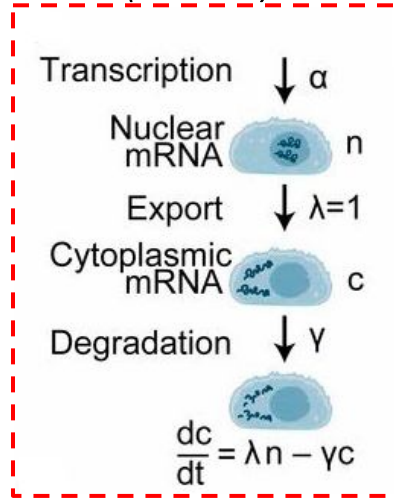
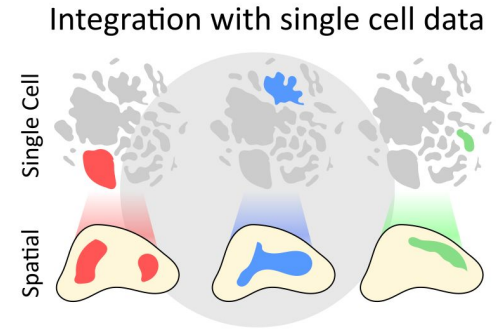


Figure 4 from *Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression*, Xia et.al

- Modified the original velocity algorithm
- Nuclear mRNA vs Cytoplasmic mRNA instead of spliced vs. unspliced
- Infer transient cell states

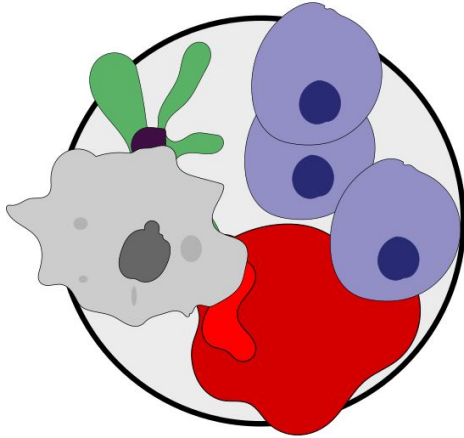
# Integration with single cell data

- **Basic idea** : use SC data as a reference when working with spatial data.
- **Answers** : Where are cell types in SC data found in space?
- **But why?** Two main reasons :
  - **Efficient use of resources**. Leverage extensive annotation work done for single cell data.
  - Problem of **mixed contributions** (some methods)

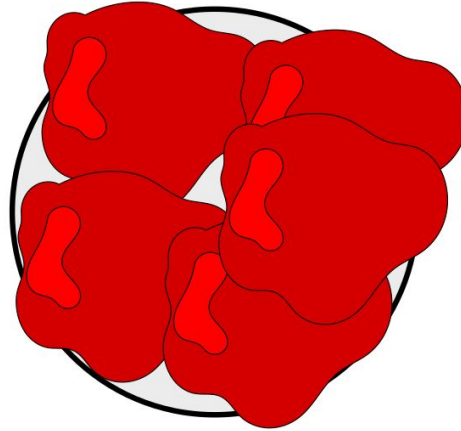


■ ■ ■ ■ Mixed contributions

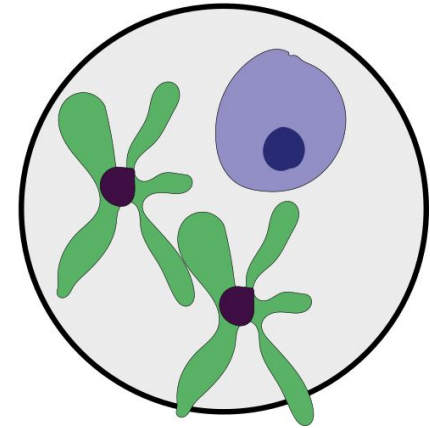
Spot 1



Spot 2

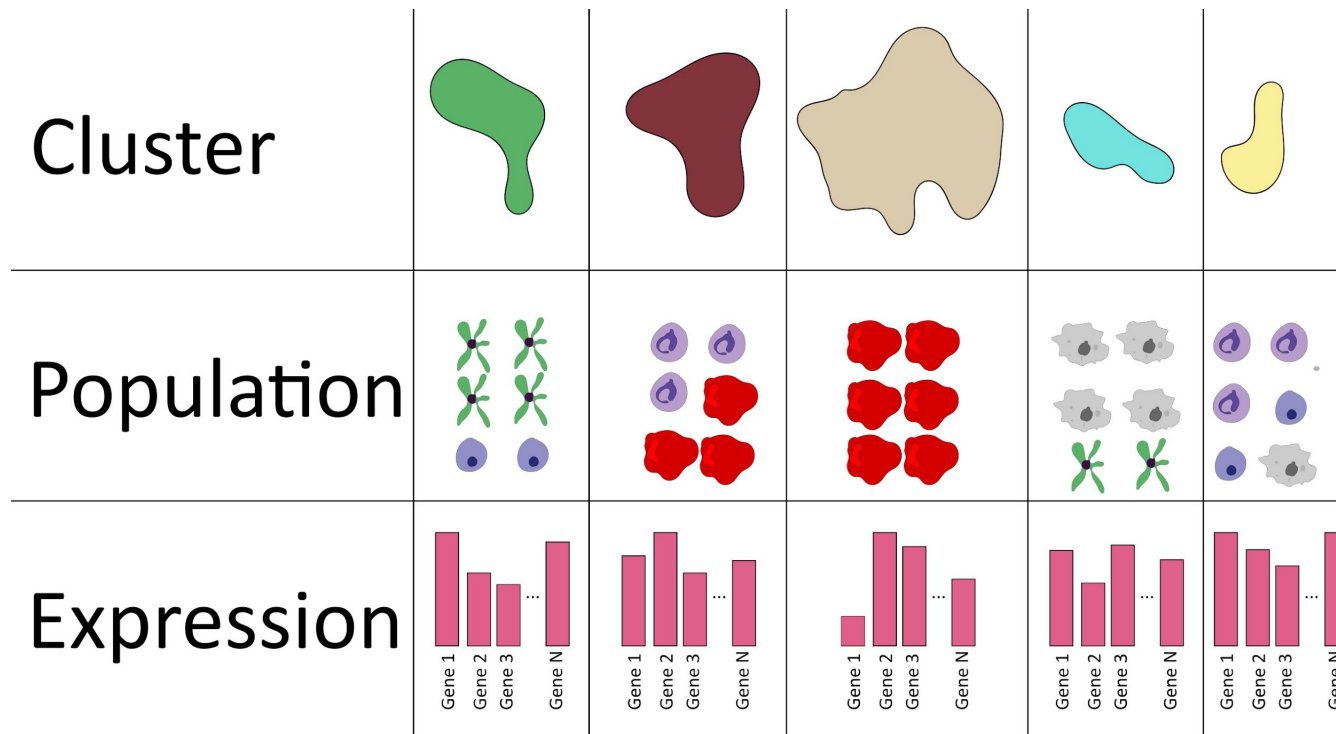


Spot 3



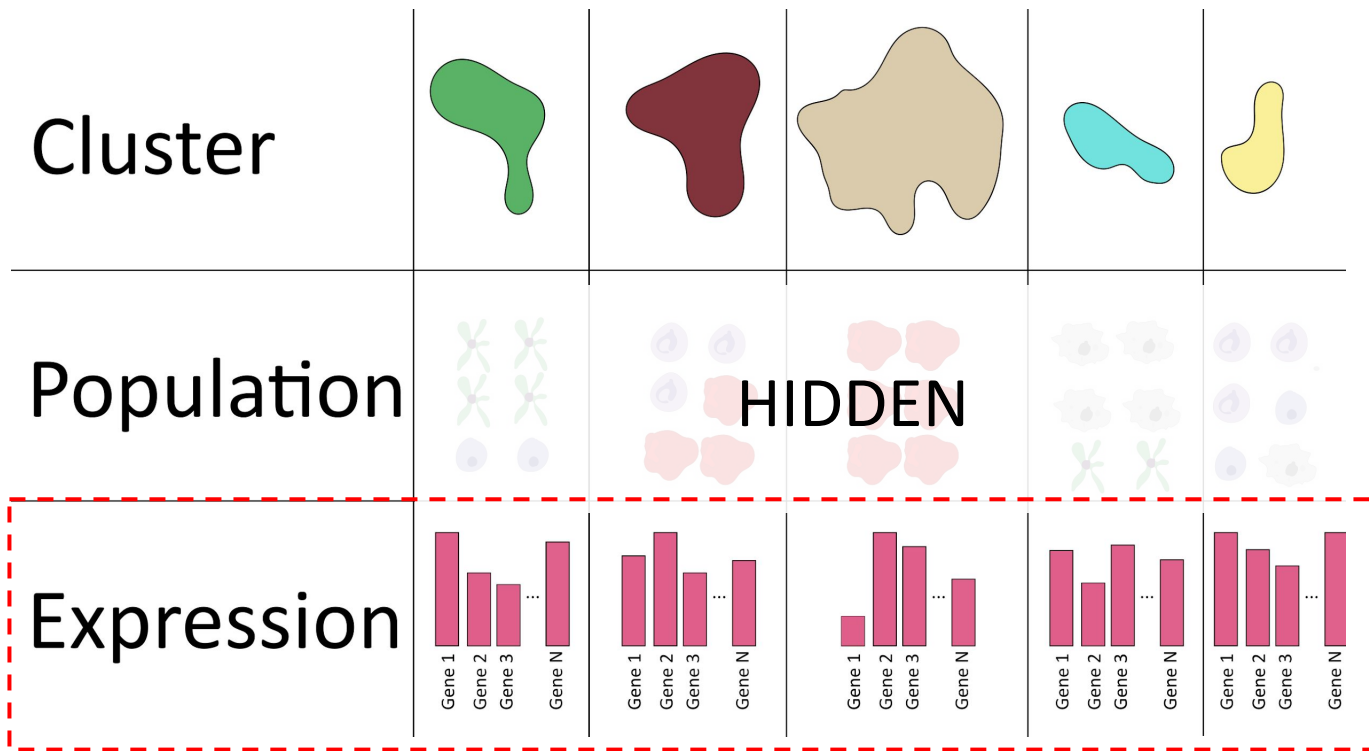
In several of the **capture based techniques** (e.g., Visium and Slide-seq), observed expression values are **contributions from multiple cells**. Not all necessarily of the same type.

# Mixed contributions



- Clusters **do not represent cell types**
- Clusters are more an assembly of spots with **similar composition** of cell types.
- We have no idea what the cell type population looks like. **Only observe expression**

# Mixed contributions



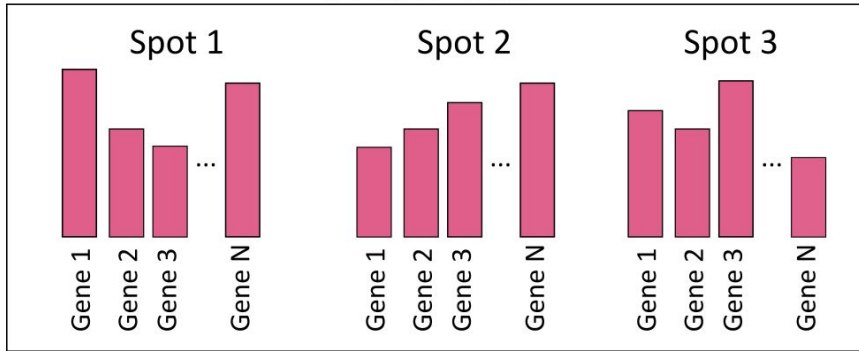
HIDDEN

- Clusters **do not represent cell types**
- Clusters are more an assembly of spots with **similar composition** of cell types.
- We have no idea what the cell type population looks like. **Only observe expression**

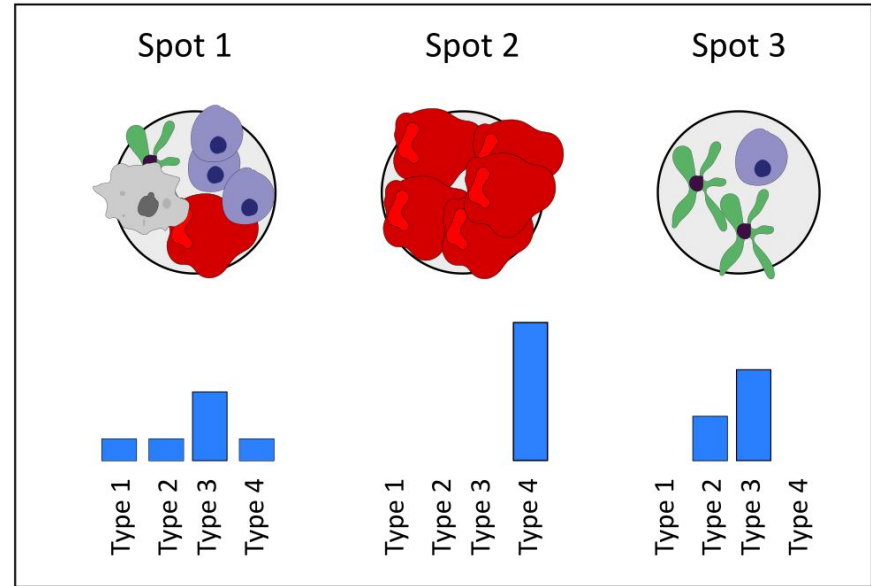
Observed

# Our objective : deconvolve expression data

From this



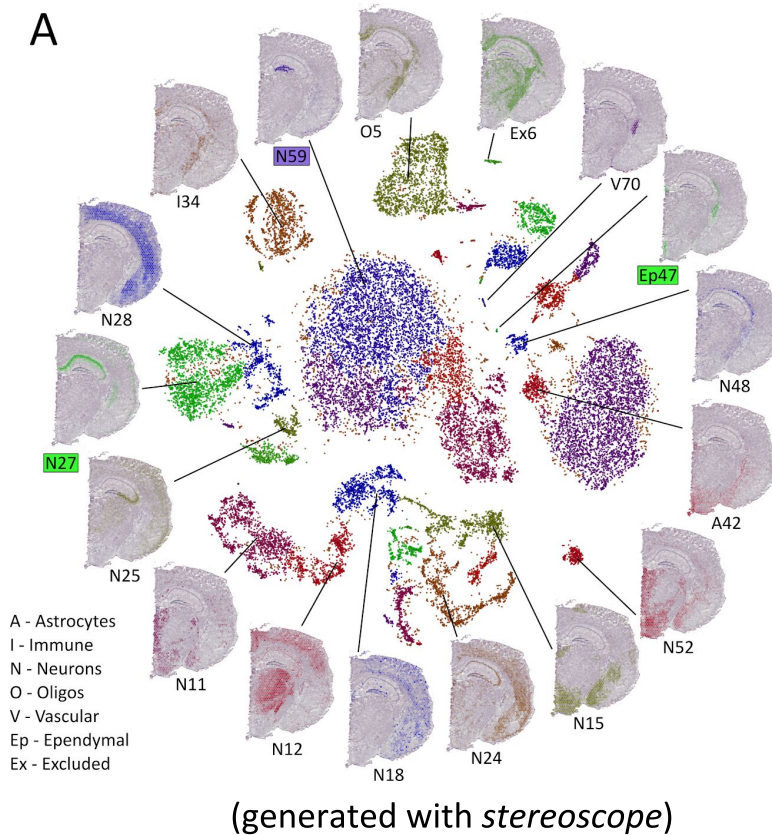
We want this





# Integration with single cell data

A



- **Inner** : Single cell data from mouse brain, gt-SNE embedding. Colored by cluster.
- **Outer** : Visium data of mouse brain. Facecolor intensity indicates proportion value of cluster.

Figure 2 from "Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography", Andersson et al

# Integration with single cell data

## Marker gene based

Extract marker genes (MG) for each cell type from SC data

Compute enrichment score for each set of MGs in spatial locations

Normalize to make scores sum to 1

**Ex:** Itai et.al

## Anchor based

Find anchors between modalities (MNNs). Create correction vector based on differences in expression.

Use correction vectors to remove platform effects. Integrated data sets.

Transfer labels of single cells to spatial data points.

**Ex:** Seurat

## Probabilistic Modelling

Assume gene expression follows certain statistical distributions.

Joint model for SC and spatial data. Learn cell type parameters from SC data, use to deconvolve spatial data (when mixed).

Correct for eventual platform differences

**Ex:** *stereoscope*, RCTD, cell2location

## Optimization based

Find spatial location where each cell is most likely to reside.

Tries to simultaneously optimize terms such as:

- Cell density
- UMI distribution across genes within spots
- gene distribution across spots

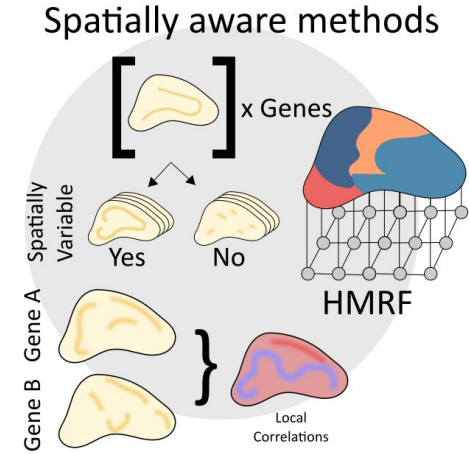
**Ex:** Tangram

# ■ ■ ■ Spatially aware methods

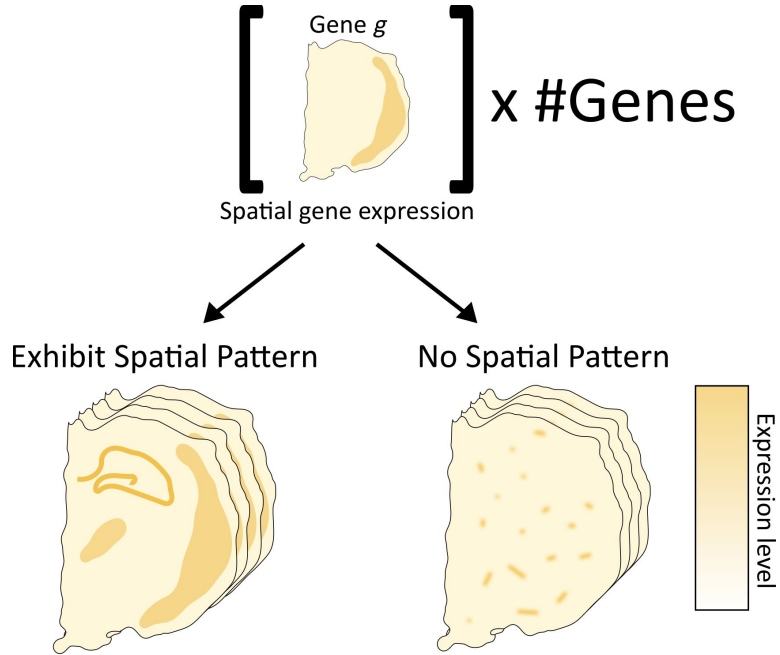
**Basic Idea** : Attempts to include knowledge of spatial structure in the analysis, not only to visualize results.

**Designed for tasks like** :

- Identifying *spatially* variable genes and features
  - Why not just select highly variable genes (**more later**)
- Finding spatially coherent expression domains
- Leverage spatial proximity to increase robustness of inference (e.g., CNA inference)
- Find *local* correlations between features

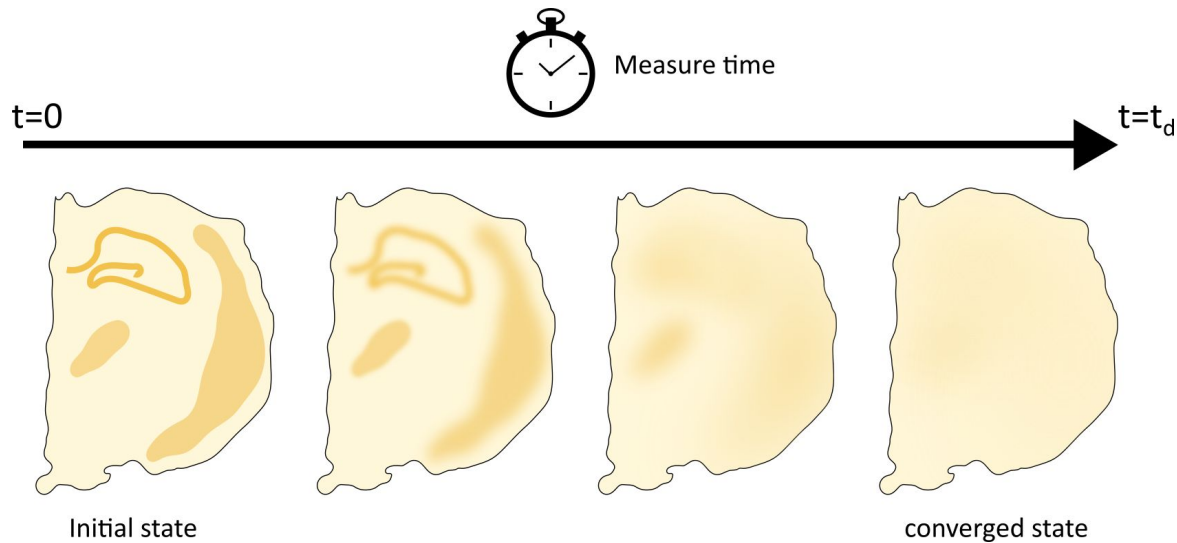


# ■ ■ ■ Spatially Variable Genes



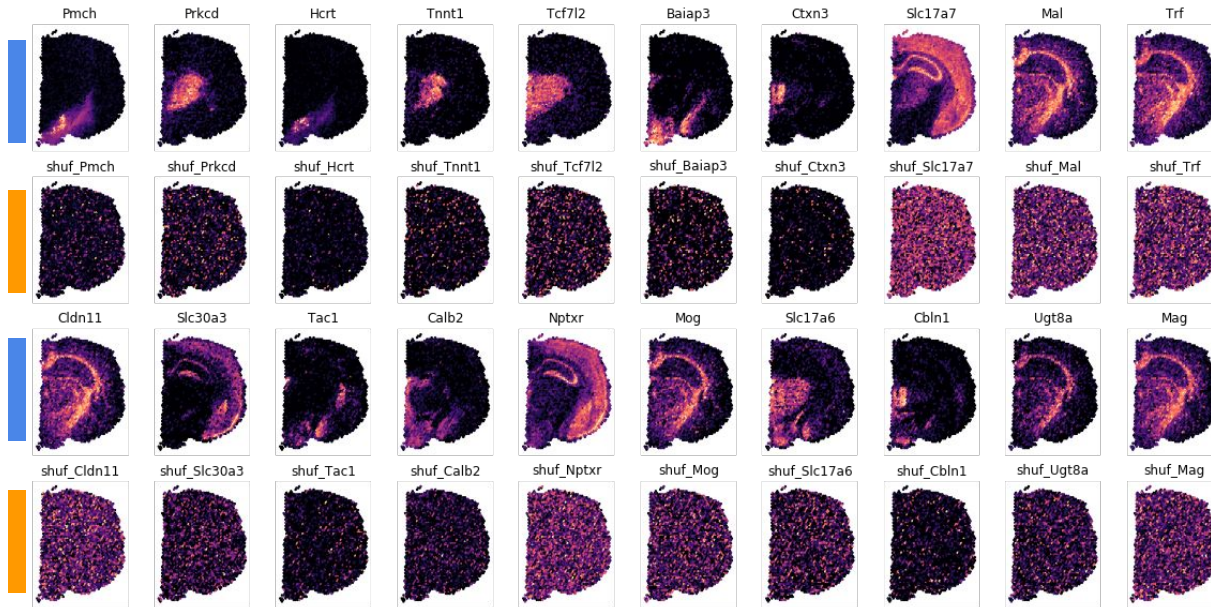
- Sort expression profiles into spatially variable or not.
- SpatialDE, SVCA and SPARK use probabilistic models
- Leverage *Gaussian Processes* (not same thing as multivariate gaussian) to model data
- Essentially, test whether a “spatial” term in the covariance function significantly increase model’s ability to explain data

# ■ ■ ■ Spatially Variable genes



- *sepal* is not probabilistic
- Uses finite differences to simulate diffusion of transcripts.
- Measures time until converges
- Ranks genes by the time it takes to converge.
- Idea : The longer the time, the more structured the initial state.

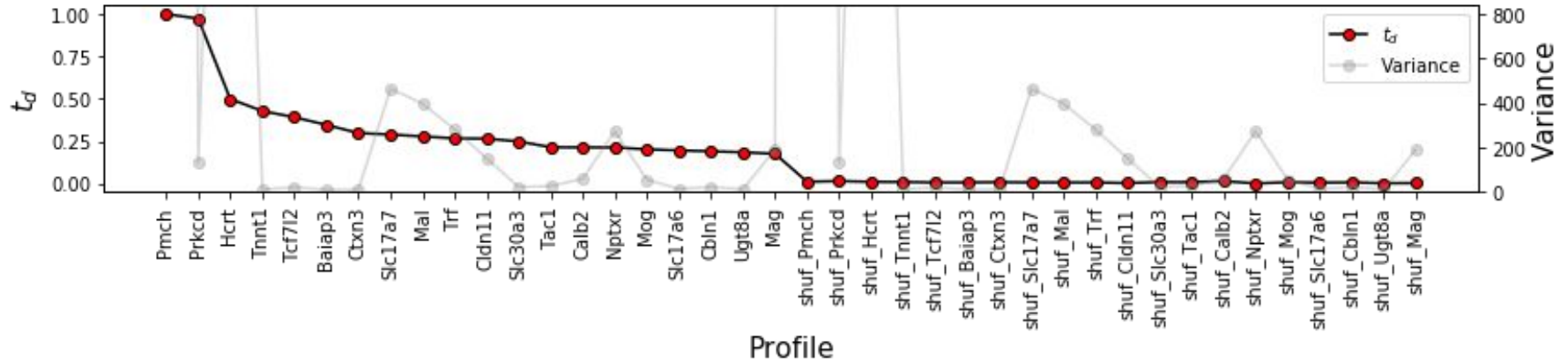
# Spatially Variable Genes



- 20 Expression profiles from mouse brain
- Shuffle spots to get random expression profiles. Has “shuf” prefix.

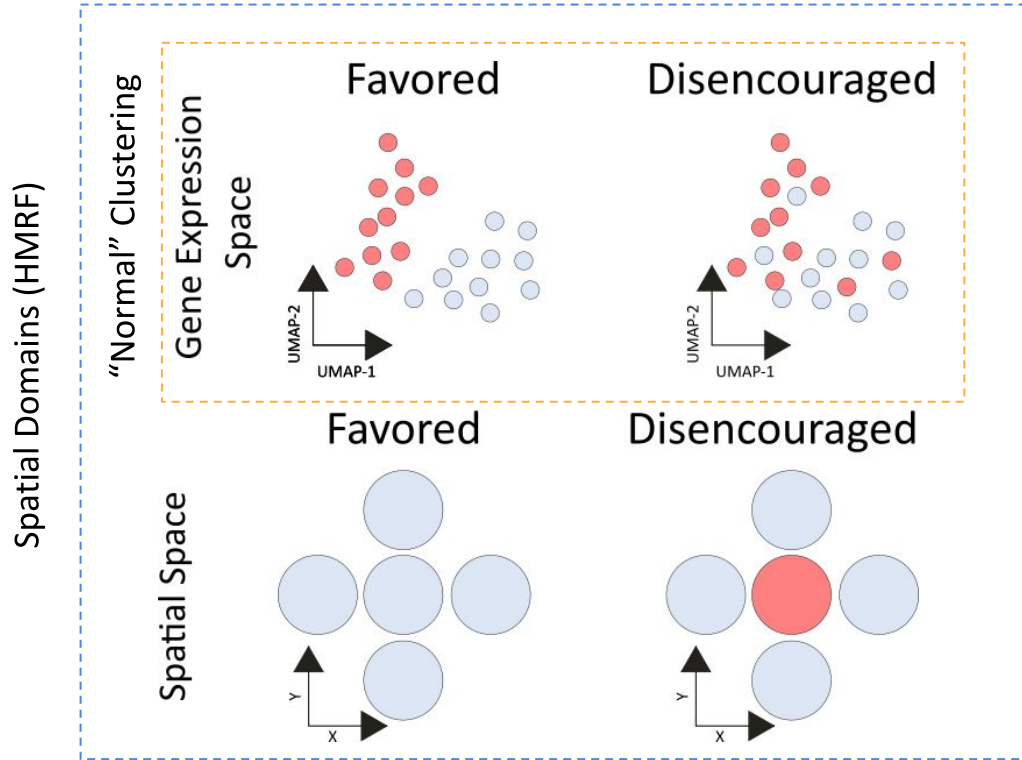
Observed Profiles | Shuffled Profiles

# Spatially Variable Genes



- Variance or dispersion metrics renders exactly the same value (gray) for shuffled and non-shuffled profiles
- *sepal's* ranks real profiles higher than shuffle ones (spatial structure considered)
- Similar results obtained for other methods as well (SpatialDE, SPARK, etc)

# Spatial domain patterns

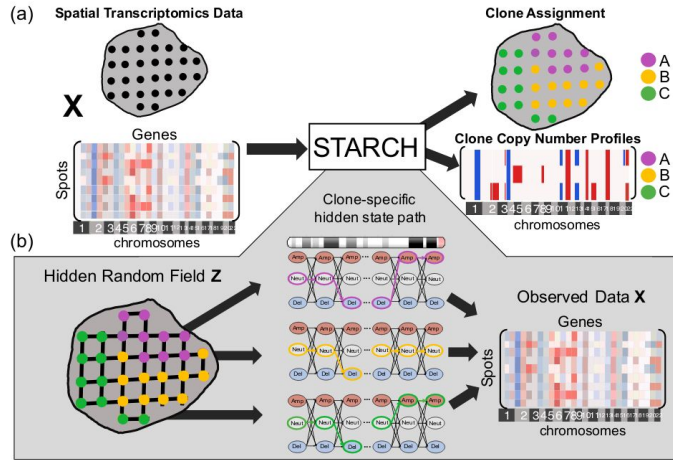


- Normal clustering mainly focus on gene expression
- Leverage spatial information to find spatially coherent clusters (domains)
- Common to use HMRF (Hidden Markov Random Field)
- Construct a graph based on spatial proximity
- Probability of node (cell) belonging to a specific domain depends on:
  - Agreement with domain expression profile
  - Coherence with neighbors

Example : Identification of spatially associated subpopulations by combining scRNAseq and sequential fluorescence in situ hybridization data”, Zhu et.al

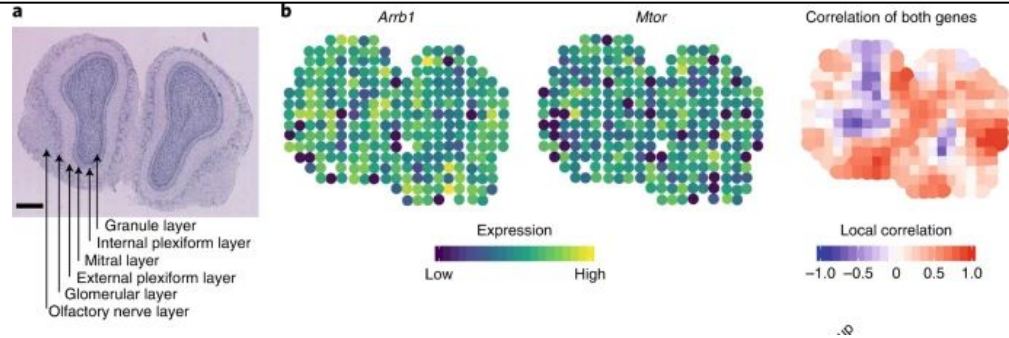


# Spatially aware methods



- **Name** : STARCH
- Infer Copy Number Aberrations (CNA) from spatial transcriptomics data
- Increase robustness of inference by aggregating data in same domains (similar profiles)
- Also uses Hidden Markov Random Fields (HMRF)
- *“STARCH: Copy number and clone inference from spatial transcriptomics data”* by Elyanow et.al

- **Name** : scHOT
- Computes (spatially) weighted correlations to find local correlations.
- *“Investigating higher-order interactions in single-cell data with scHOT”* by Ghazanfar, et.al



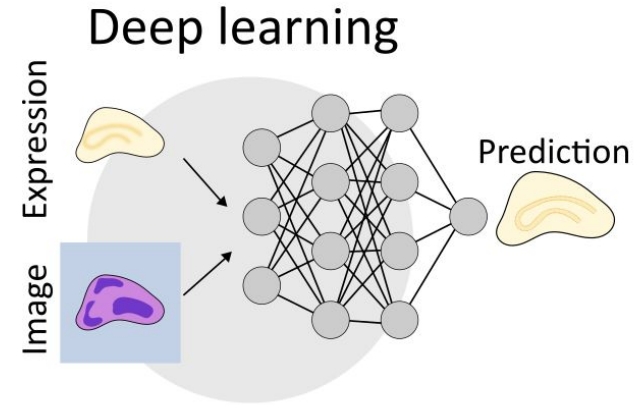
# Deep Learning

**Basic Idea :** Applying deep learning to spatial data (very broad)

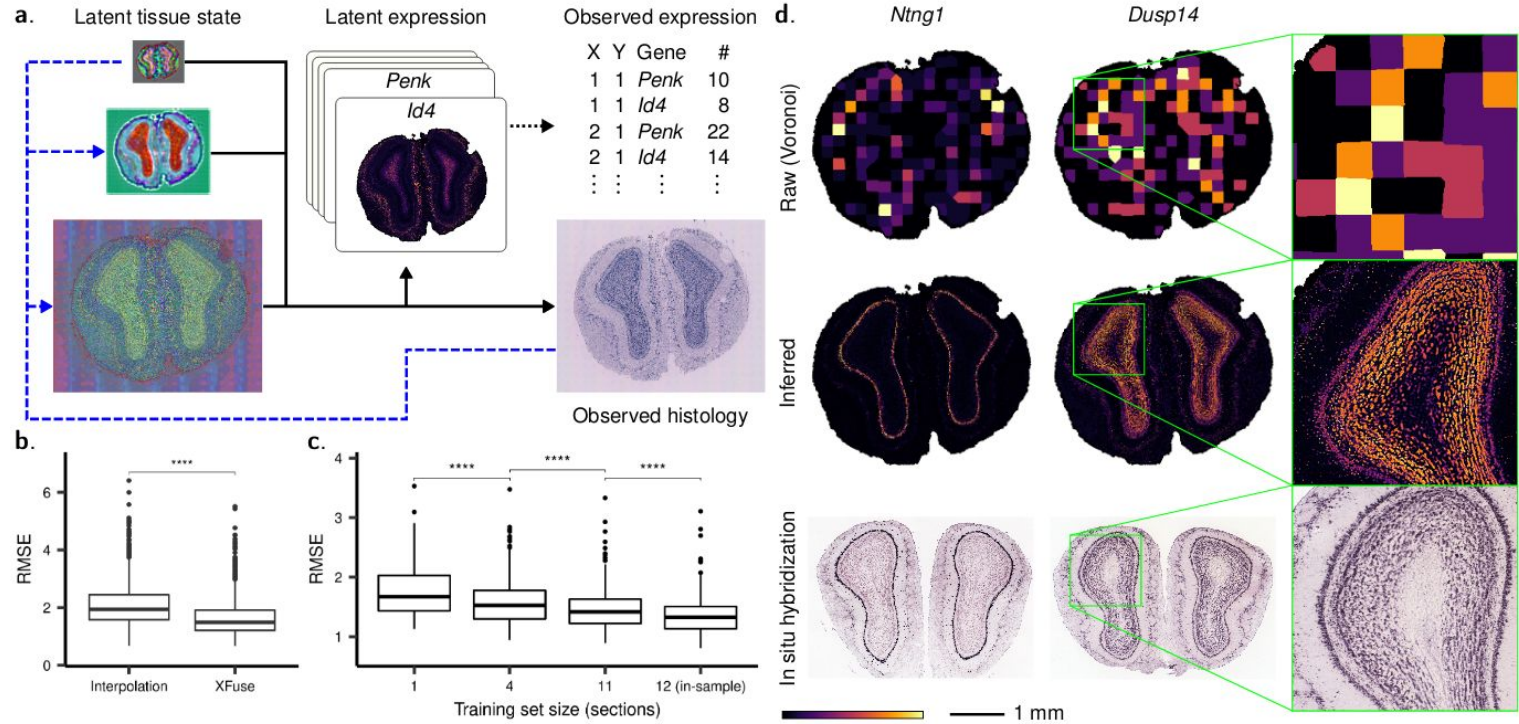
**Nascent :** Relatively few examples. Limited amount of high quality available data. More traditional ML methods have so far been more appropriate to use. This is changing.

**Current examples :**

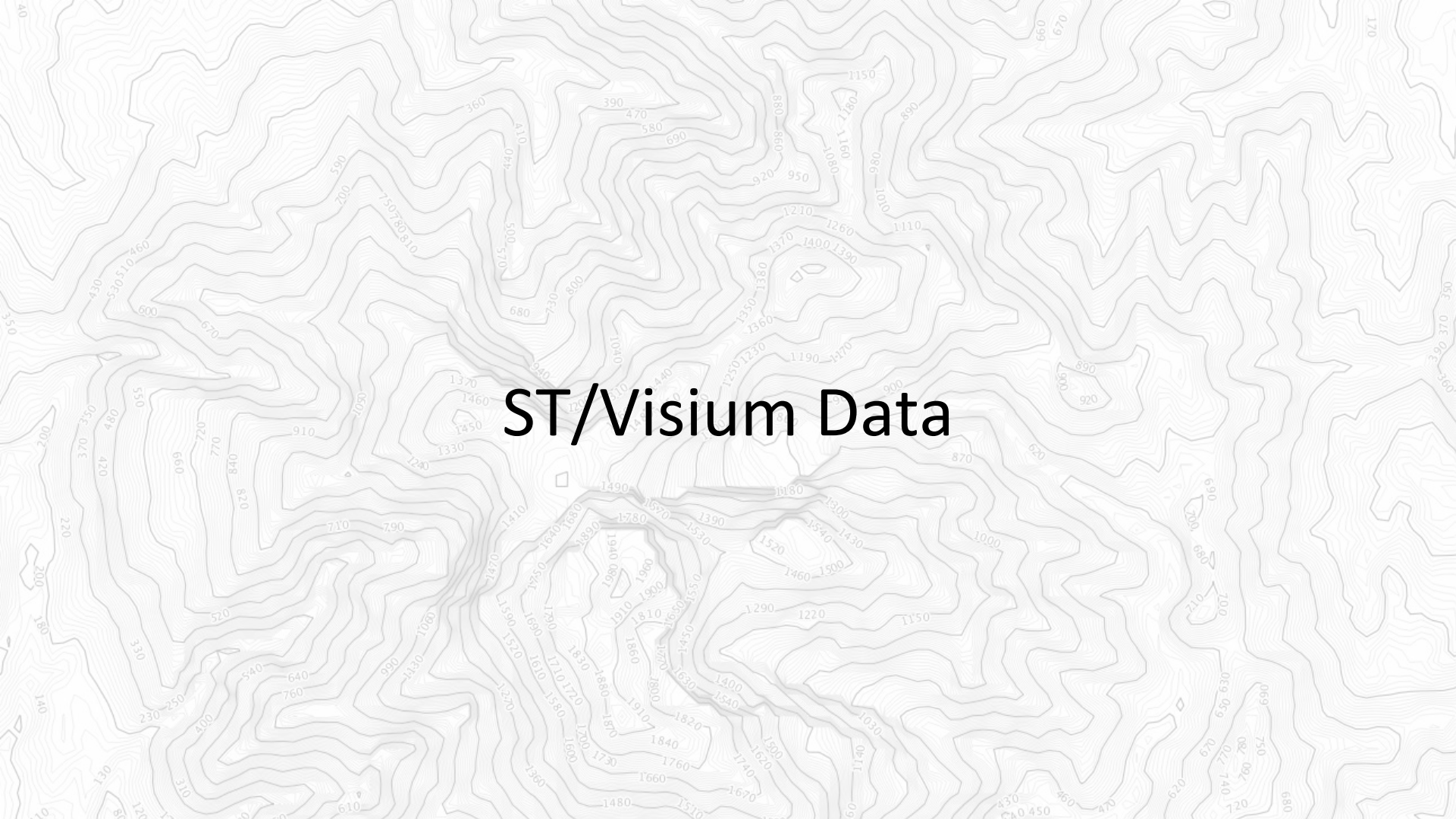
- **stNet** : relate gene expression data to morphology.
- **xFUSE** : “*superresolution*” (pixel) of gene expression by learning joint representation of image and expression data.



# Deep Learning | xFUSE



From : “Super-resolved spatial transcriptomics by deep data fusion”, Bergensträhle et.al (Figure 1)



# ST/Visium Data

Spatial Transcriptomics, ST and Visium,  
what's the deal?



# Spatial Transcriptomics (ST)


Mid 2016

**TRANSCRIPTION**

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

Paula A. Ståhl<sup>1,2\*</sup>, Fredrik Bråte<sup>1,2\*</sup>, David Vickberg<sup>1,2</sup>, Anna Lindvall<sup>1,2,3,4,5</sup>, Jonas Löfdahl<sup>1,2,3,4,5</sup>, Hans Eriksson<sup>1,2,3,4,5</sup>, Anders Carlsson<sup>1,2,3,4,5</sup>, Mikael Häggström<sup>1,2,3,4,5</sup>, Hans Larsson<sup>1,2,3,4,5</sup>, Fredrik Larsson<sup>1,2,3,4,5</sup>, David Lagergren<sup>1,2,3,4,5</sup>, Peter Nilsson<sup>1,2,3,4,5</sup>, and Anders Reijman<sup>1,2,3,4,5</sup> <sup>1</sup>Uppsala University, <sup>2</sup>Uppsala Biomedical Center, <sup>3</sup>Uppsala Genome Center, <sup>4</sup>Uppsala Center for Protein Research, <sup>5</sup>Uppsala Center for RNA Research

Analysis of the pattern of protein expression (RNA-seq) in biological tissue remains a challenge in experimental biology. This difficulty is due to the fact that the spatial distribution of protein expression is not captured by standard RNA-seq. Here, we present a method for visualizing and analyzing gene expression in tissue sections by spatial transcriptomics. This method involves the use of a novel technology for capturing and sequencing RNA in situ. The resulting data are analyzed using a novel computational pipeline. This pipeline allows for the visualization and analysis of gene expression in tissue sections. The resulting data are analyzed using a novel computational pipeline. This pipeline allows for the visualization and analysis of gene expression in tissue sections. The resulting data are analyzed using a novel computational pipeline. This pipeline allows for the visualization and analysis of gene expression in tissue sections.



Mainly used by the Lundeberg and Frisén lab



Late 2018

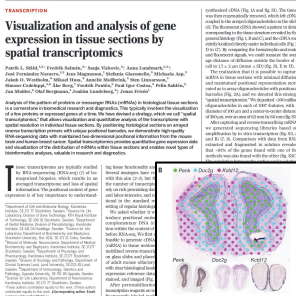
Science Publication  
Ståhl et.al





# Spatial Transcriptomics (ST) Visium

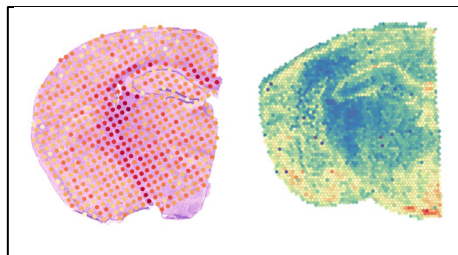
Mid 2016



Science Publication  
 Ståhl et.al

Now referred to as:

- ST
- Legacy ST
- Original ST
- ST1k
- Visium (by unattentive readers..)



Late 2018

Late 2019

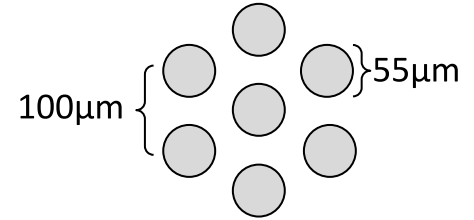
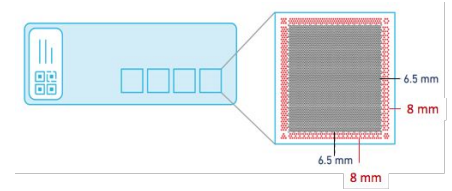
10X  
 GENOMICS®  
 (acquisition)

Launch of **Visium**  
 Spatial Gene  
 Expression Platform



# Visium Platform

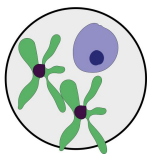
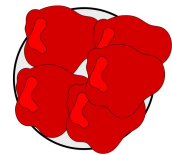
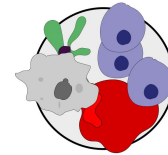
- Array based technique
- 6.5mm x 6.5mm area to put sample on
- 4992 spots arranged in hexagonal grid
- Spot specs:
  - Spot diameter : 55 $\mu$ m
  - Center to center distance : 100  $\mu$ m
- Each spot has millions of capture probes
  - spatial barcode
  - polyT sequence
  - captures polyadenylated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



Spot 1

Spot 2

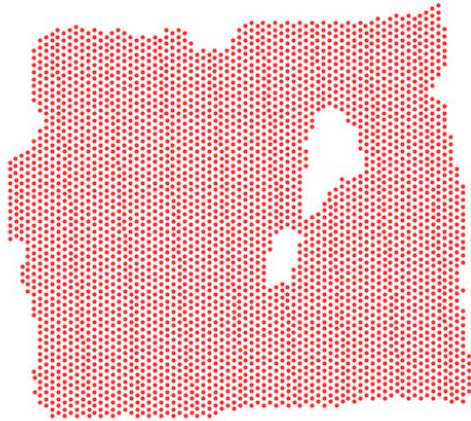
Spot 3



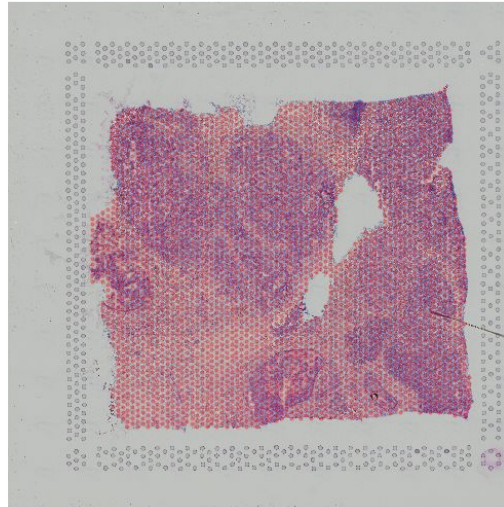
# ■ ■ ■ An example

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

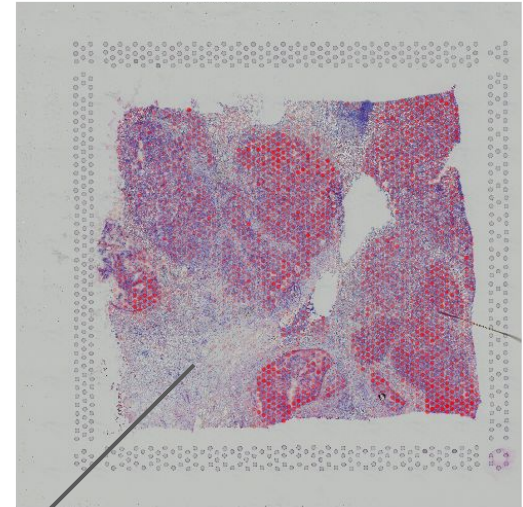
1. Spots



2. Spots + Image



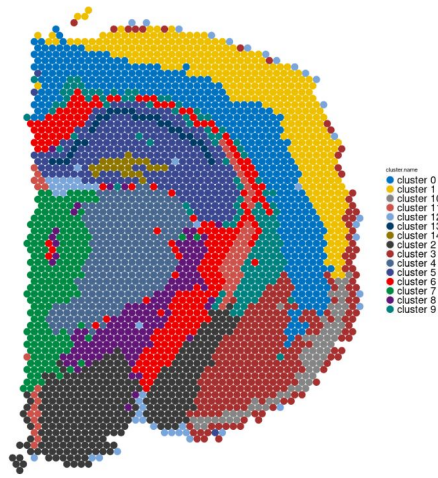
3. Spots | ERBB2 expression + Image



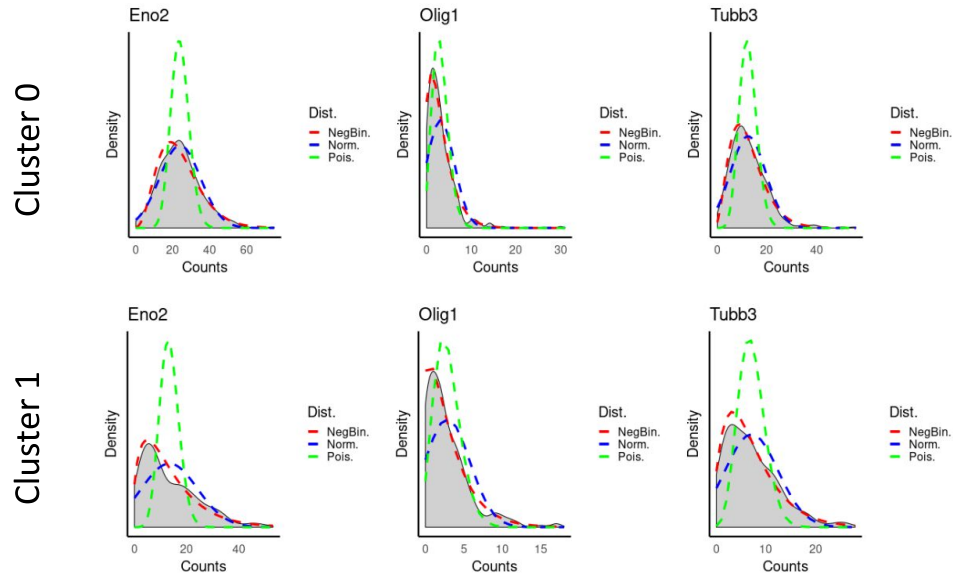
Facecolor intensity proportional  
to gene expression value

# ■ ■ ■ A word on the distribution

- Single cell data usually modelled as overdispersed Poisson distribution (Negative Binomial). Basis for several analysis methods (Normalization, DE, etc.)
- Applicable to ST/Visium data as well



Visium data Mouse Brain  
Colored by cluster



Similar trends for all clusters and genes.

Supports NB distribution, also when corrected for increased parameter number compared to Poisson).

# Some brief words on the ST/Visium analysis

- **Batch effects** between sections are usually observed, try to account for this. SC methods have worked great so far.
- **Cell density** is often not homogeneous across tissue. Good to normalize based on the library size to account for this.
- Keep in mind that expression profiles are **mixtures**, often it makes more sense to analyze them accordingly; looking at factor contributions rather than hard cluster identities.
- Single cell mapping is often **improved by use of HVG** genes or curated lists
- **Trajectory inference is tricky**, no method that I am aware of accounts for the fact that several temporal states might be present at each observation. Incorporation of spatial information has been done fairly heuristically so far.
- Filtering **ribosomal, mitochondrial and Hb-genes** usually have a positive effect on the result. They usually constitute irrelevant sources of variation.
- We have observed some “leakage” around the edges, especially in Visium samples. Diffusion is minimal in tissue, but near borders transcripts might leak a bit. Keep this in mind.

# Summary

- Tons of spatial techniques
  - Only a few commercialized ones
  - Define your question before choosing the method
- Ever increasing repertoire of computational methods
  - Be careful when transferring SC methods, ask yourself if it makes sense.
  - Explore and test
  - Make use of the spatial information for sanity checks
- ST is the old Visium
- Don't just treat spatial data as a different form of SC data, it has much more to offer

# Thank you for the attention!



<https://github.com/almaan>



<https://almaan.github.io>

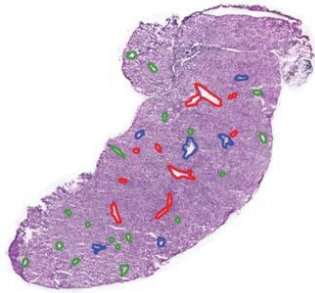


<https://www.spatialresearch.org>

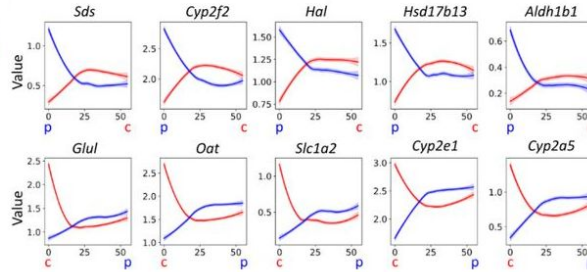




# Expression as function of distance

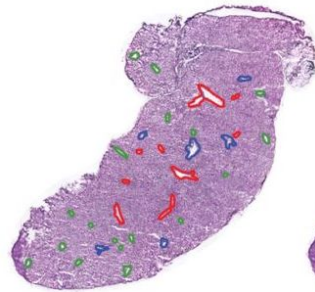


portal vein  
central vein  
ambiguous



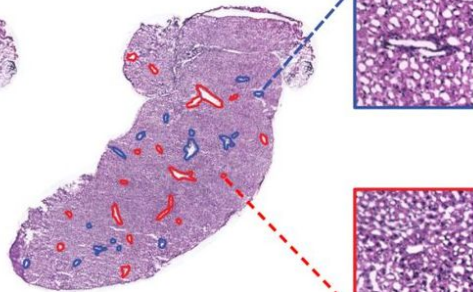
- **Concept** : assess how a feature can be described as a function of the distance to a landmark
- Here we look at expression of genes associated to two different type of veins
  - Central veins : **Red**
  - Portal veins : **Blue**

Visual annotation



portal vein  
central vein  
ambiguous

Computational annotation



P(central) : 0.374

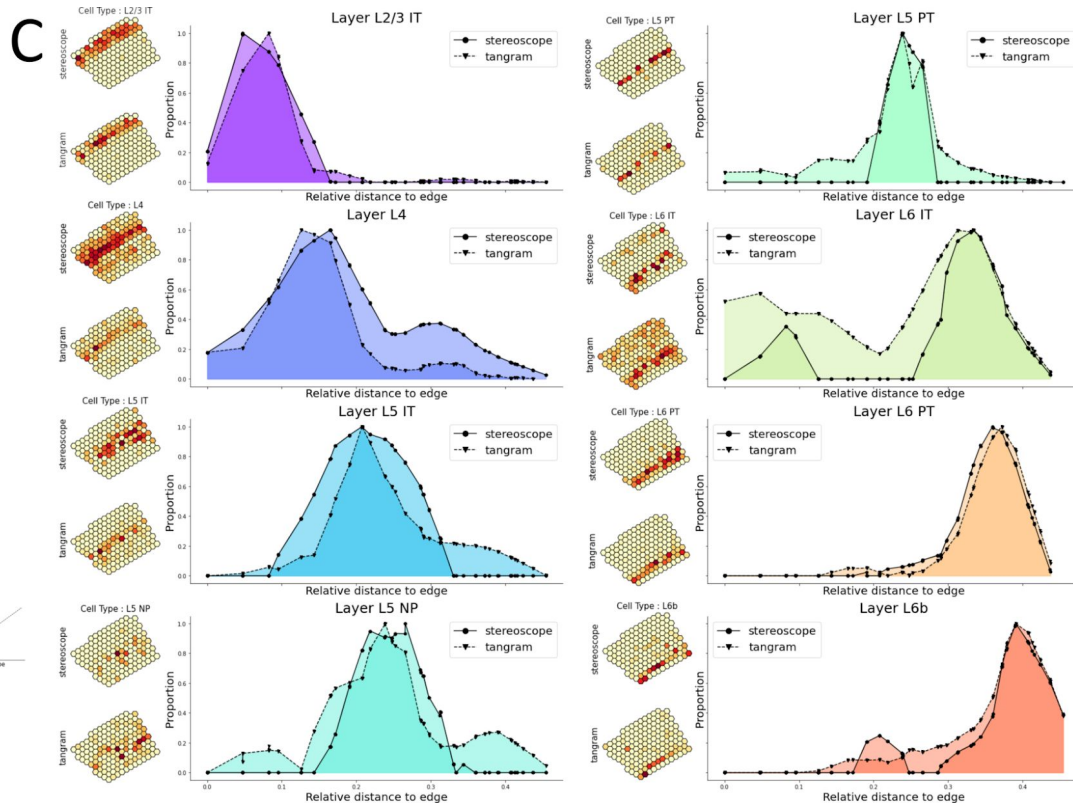
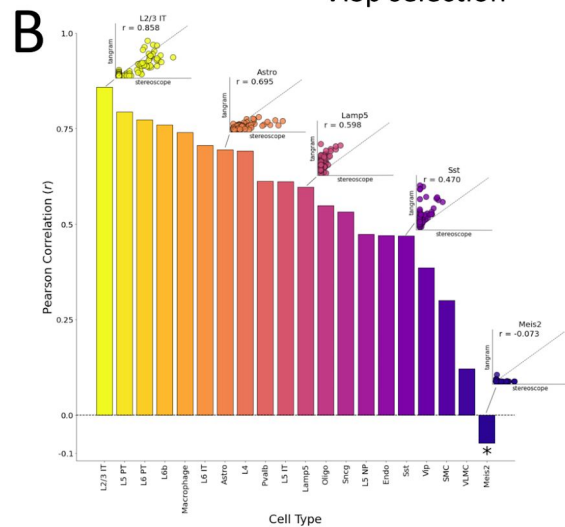
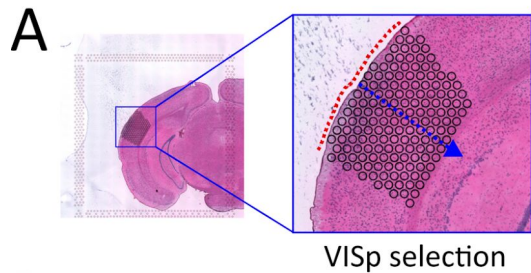
P(portal) : 0.626

P(central) : 0.964

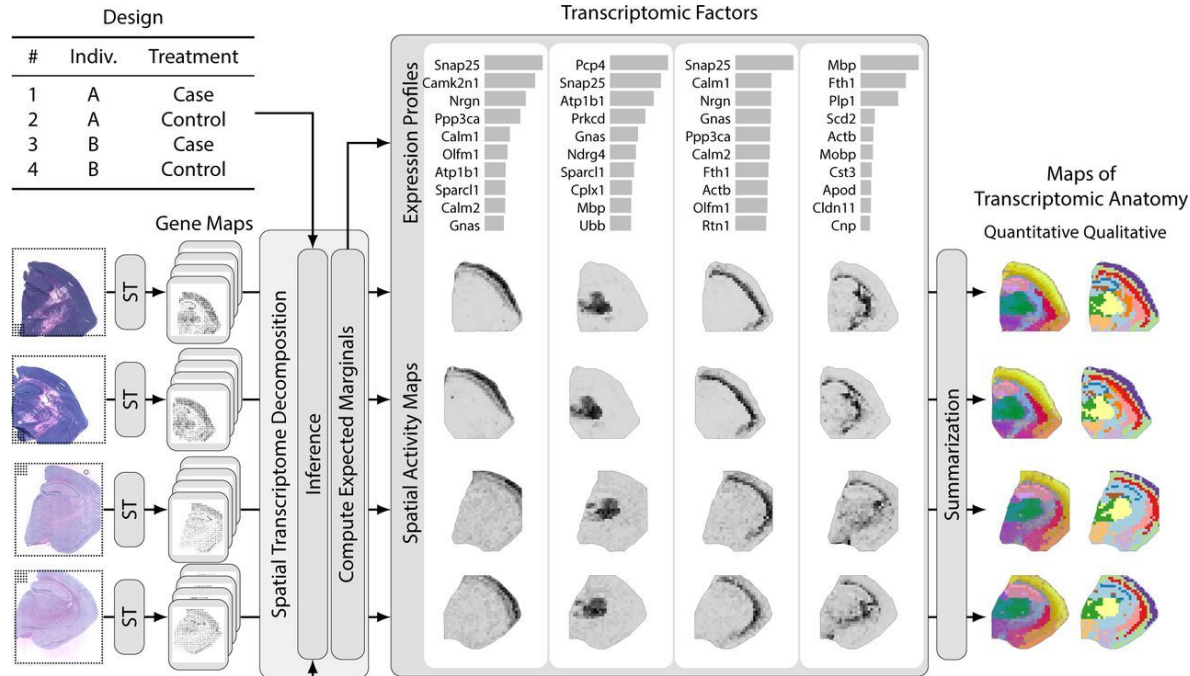
P(portal) : 0.036

- **Predict** vein type based on expression profile of spatial neighborhood
- Trains on expert's annotations, to predict ambiguous (morphological) structures

# Spatial Cell Type Distribution



# Decomposition by factor models



**“Charting Tissue Expression Anatomy by Spatial Transcriptome Decomposition”, Maaskola et.al**

Core Model	Regression	Coefficient Prior Structure		
$x_{gs} = \sum_{i=1}^T x_{gts}$	$\log r_{gts} = r + r_g + r_{gt} + r_t + r_{ts} + r_s$ $+ r_{g\#} + r_{g\text{ indiv}} + r_{g\text{ treat}}$	$r \sim \mathcal{N}(0, 1)$	$r_g \sim \mathcal{N}(0, 1)$	$r_{gt} \sim \mathcal{N}(0, 1)$
		$r_t \sim \mathcal{N}(0, 1)$	$r_{ts} \sim \mathcal{N}(0, 1)$	$r_s \sim \mathcal{N}(0, 1)$
$x_{gts} \sim \text{NB}(r_{gts}, \rho_{gs})$	$\log \rho_{gs} = \rho + \rho_g$	$r_{g\#} \sim \mathcal{N}(0, 1)$	$r_{g\text{ indiv}} \sim \mathcal{N}(0, 1)$	$r_{g\text{ treat}} \sim \mathcal{N}(0, 1)$
		$\rho \sim \mathcal{N}(0, 1)$	$\rho_g \sim \mathcal{N}(0, 1)$	