

An orientation in the spatial transcriptomics landscape

2021 :: Summer School Advanced Topics in Single Cell Analysis



<https://github.com/almaan>



<https://almaan.github.io>



@aalmaander

Who am I?

- **Name** : Alma Andersson
- **Part of** : Lundeberg Lab (PhD Student)
- **Works with** : Computational Method Development
 - Mainly focus on spatial transcriptomics data
- **Background** :
 - Engineer by training
 - **Before**: Molecular Dynamics
 - **Now**: Spatial Transcriptomics
- **Work** :
 - Single cell and spatial transcriptomics data integration (*stereoscope*)
 - Model to find spatially variable genes (*sepal*)
 - Spatial characterization of HER2 breast cancer samples
 - Common coordinate frameworks for spatial data
- **Non-scientific Interests**
 - Trail/Ultrarunning, Hiking, Outdoor stuff



Introduction

- Broad overview of experimental spatial transcriptomics techniques
- A Recap on Visium
- Data character - what are we working with?

Computational methods and frameworks

- Different flavors of currently available methods
- Example methods
- Extra focus on single cell mapping and integration
- squidpy : a framework for handling spatial data

Observations from the wild

- General advice
- Example : A spatial survey of HER2-positive breast cancer
- Example : Spatial gene expression dynamics in the mouse liver

Introduction

- Broad overview of experimental spatial transcriptomics techniques
- A Recap on Visium
- Data character - what are we working with?

Computational methods and frameworks

- Different flavors of currently available methods
- Example methods
- Extra focus on single cell mapping and integration
- squidpy : a framework for handling spatial data

Observations from the wild

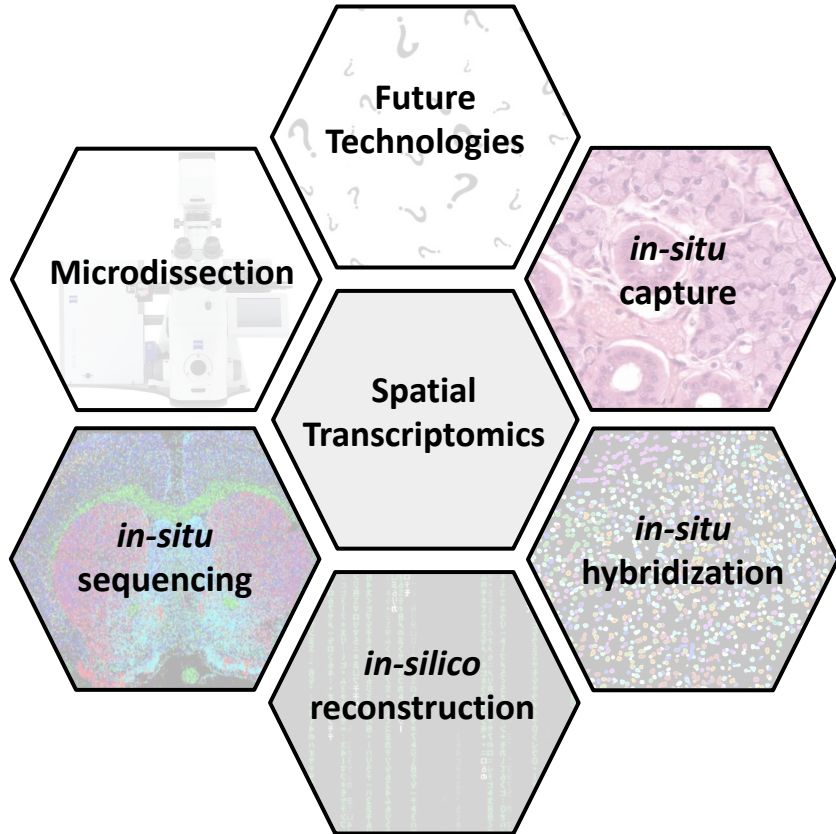
- General advice
- Example : A spatial survey of HER2-positive breast cancer
- Example : Spatial gene expression dynamics in the mouse liver

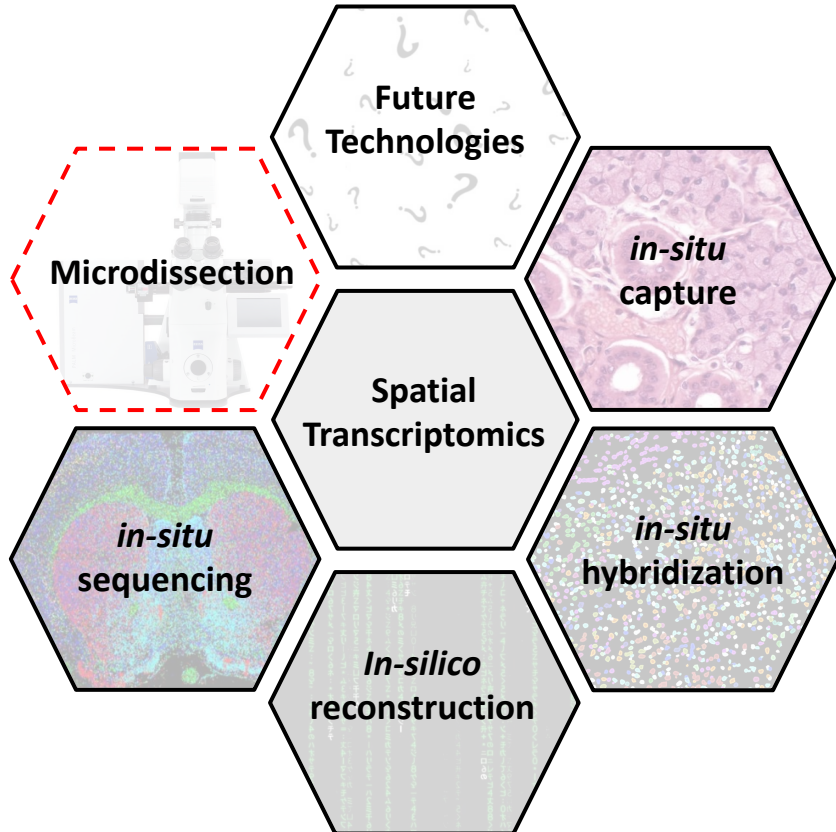
Slides ::

lectures/spatial_transcriptomics.pdf

Links ::

<https://almaan.github.io/extras/advsc-info/>



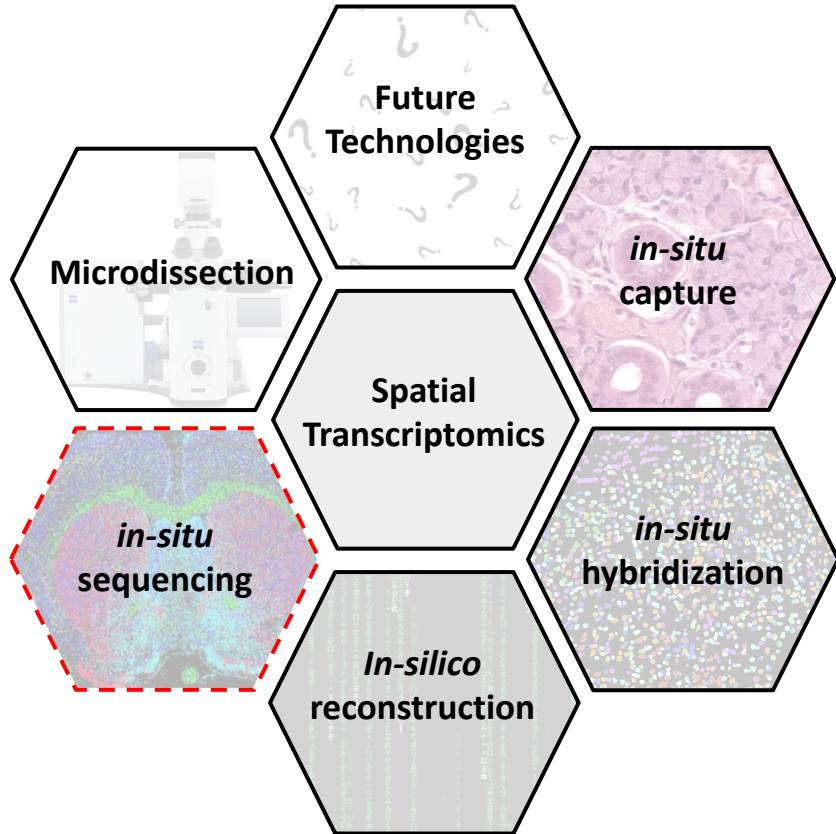


Microdissection-based technologies

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

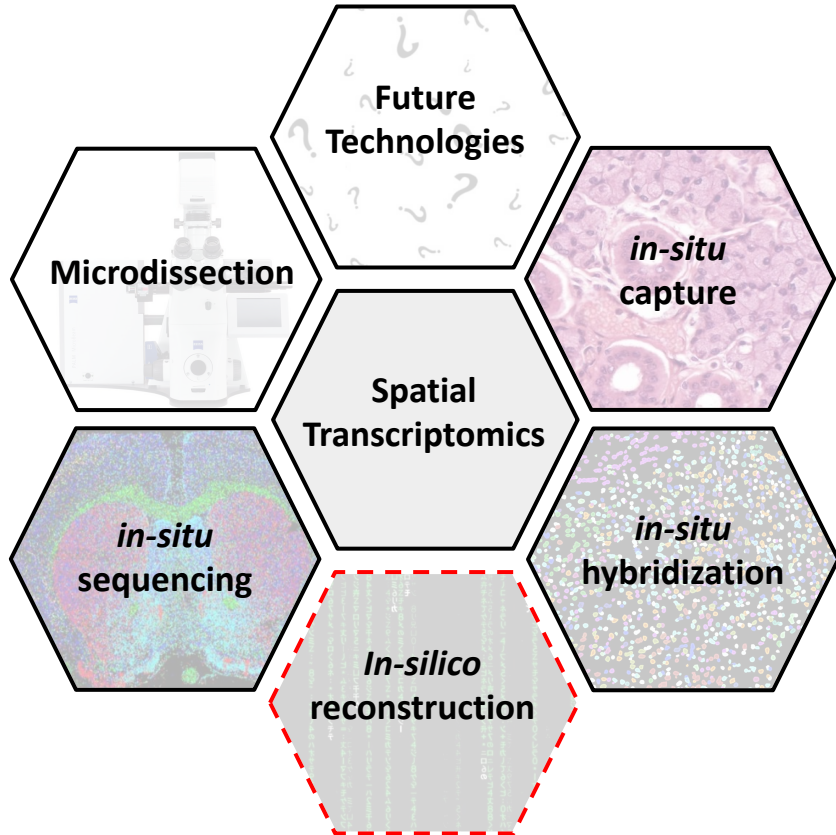


In-situ sequencing based methods

Sequence the transcripts in place.

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

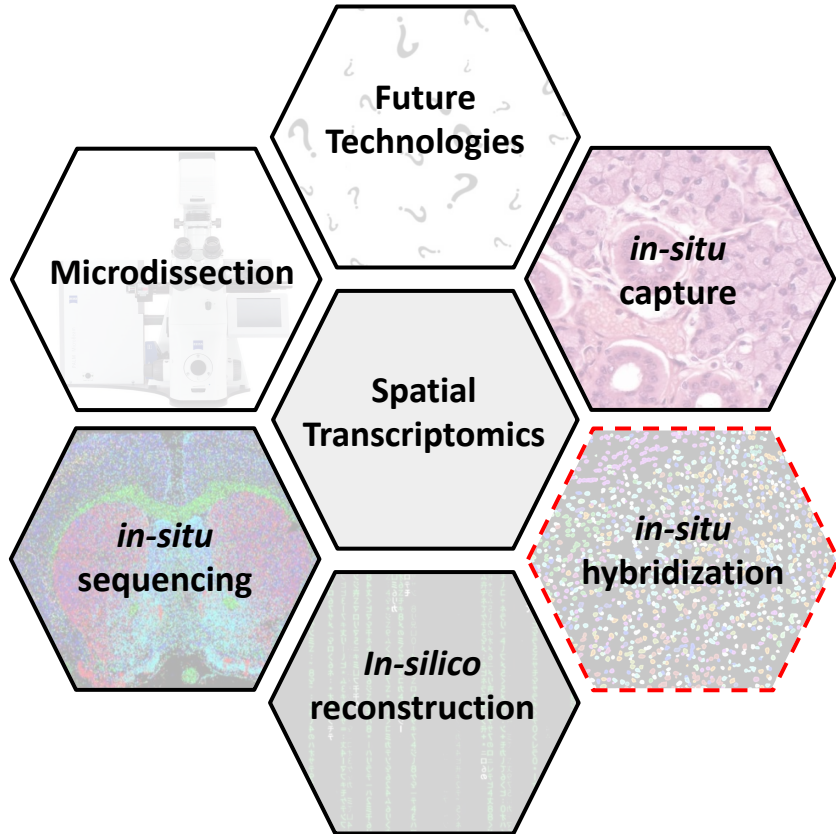
Examples : ISS/Cartana (padlock probes), BaristaSeq, STARmap, FISSEQ



In-silico reconstruction

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

Examples : novoSpaRc, CSOmap, Seurat v3



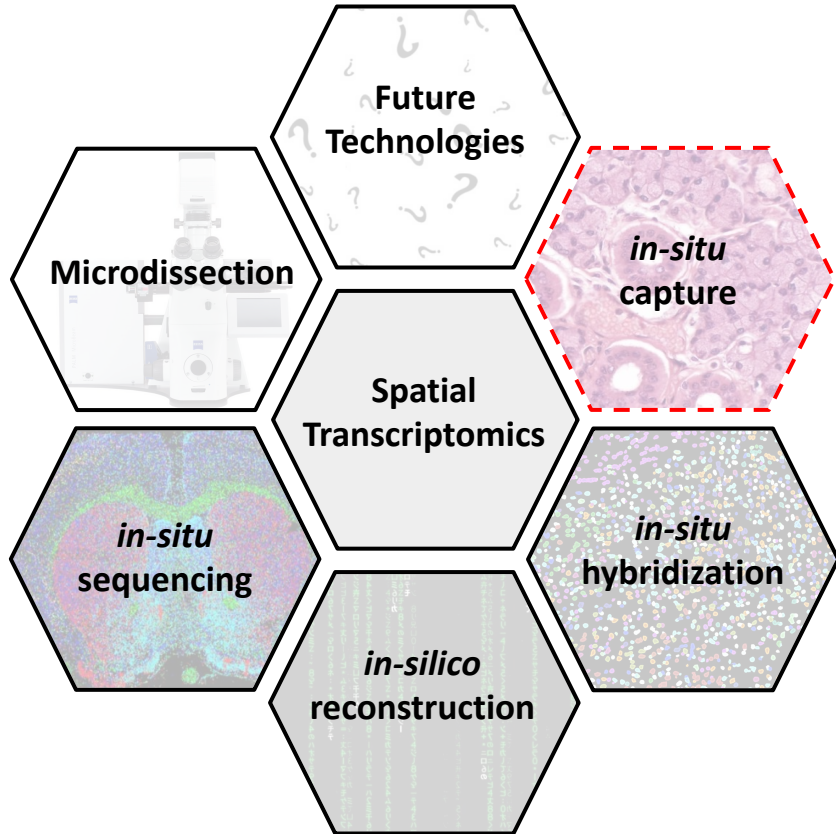
In-situ hybridization based methods

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.

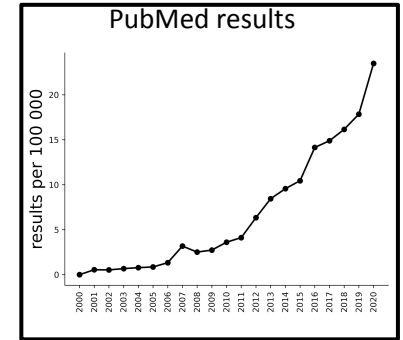
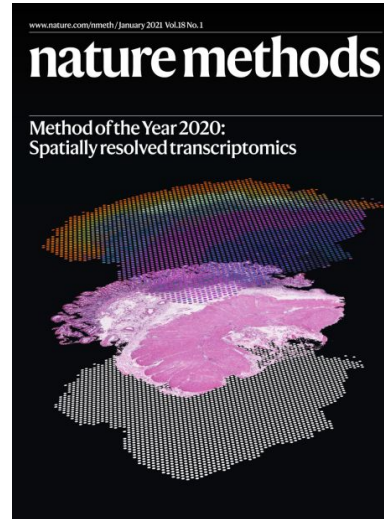
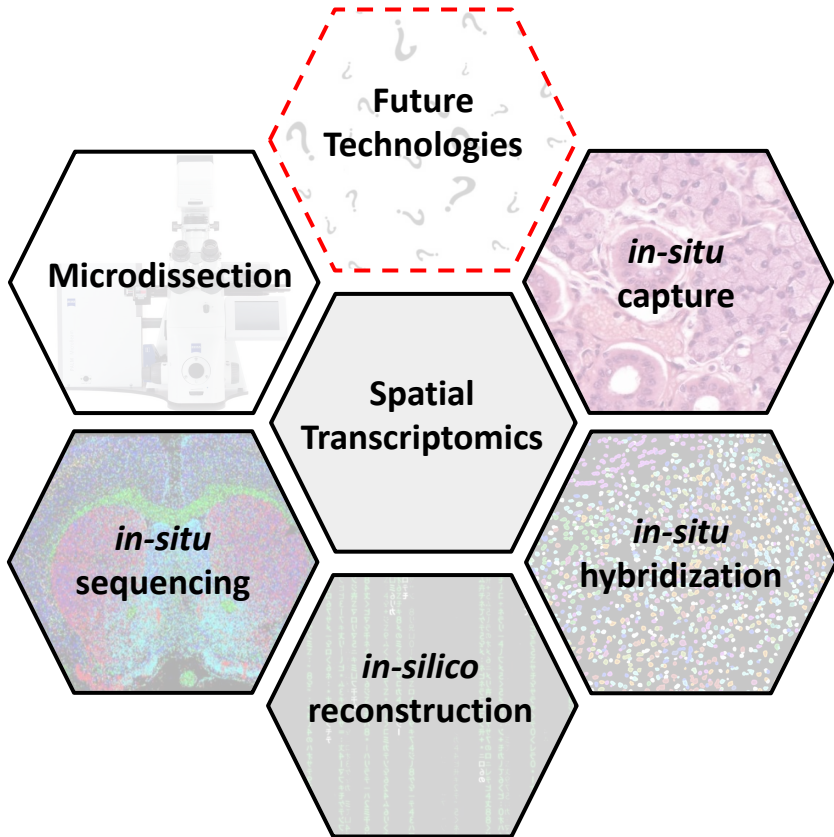
Examples : smFISH, seqFISH, MERFISH, seqFISH+, osmFISH, RNA Scope, DNA microscopy



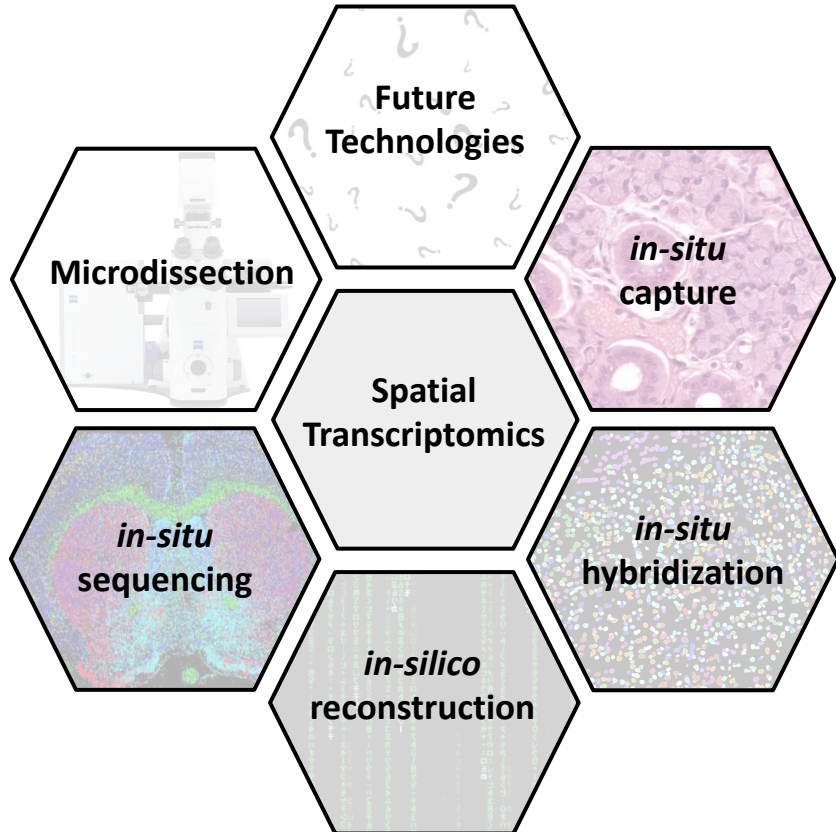
In-situ capture based methods

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



Search: 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)

Spatially resolved transcriptomics adds a new dimension to genomics

Authors : Ludvig Larsson, Jonas Frisé & Joakim Lundeberg

Published : 2021-01-06

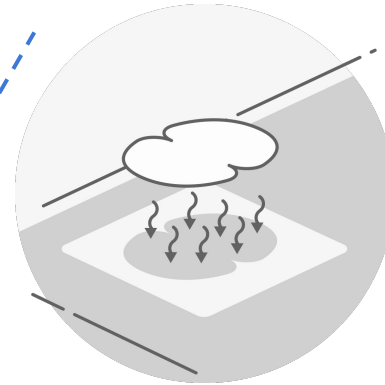
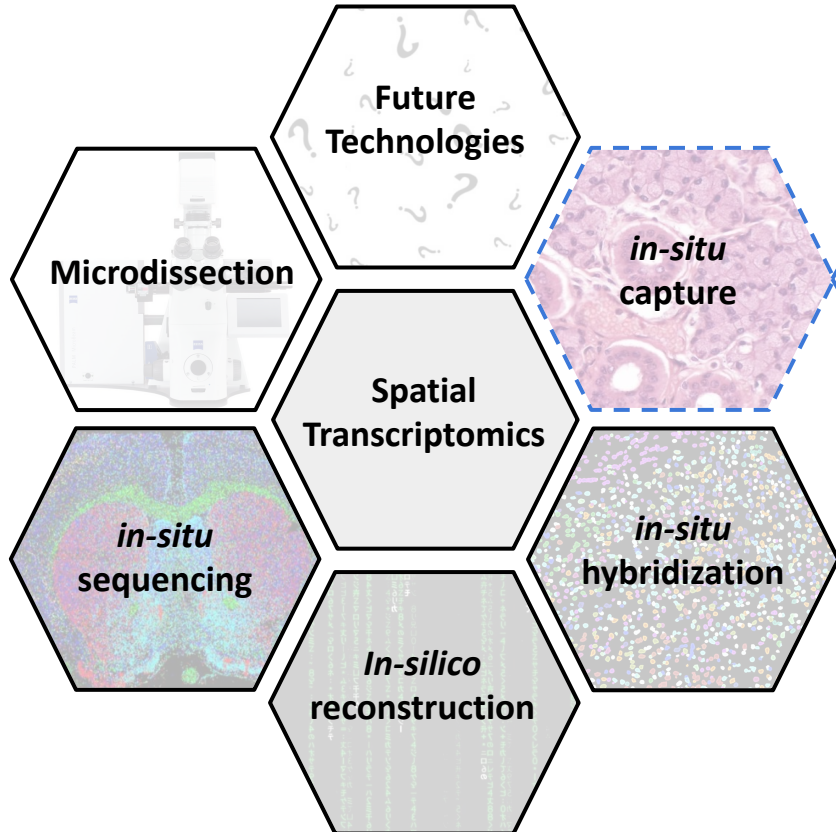
DOI: [10.1038/s41592-020-01038-7](https://doi.org/10.1038/s41592-020-01038-7)

Museum of Spatial Transcriptomics

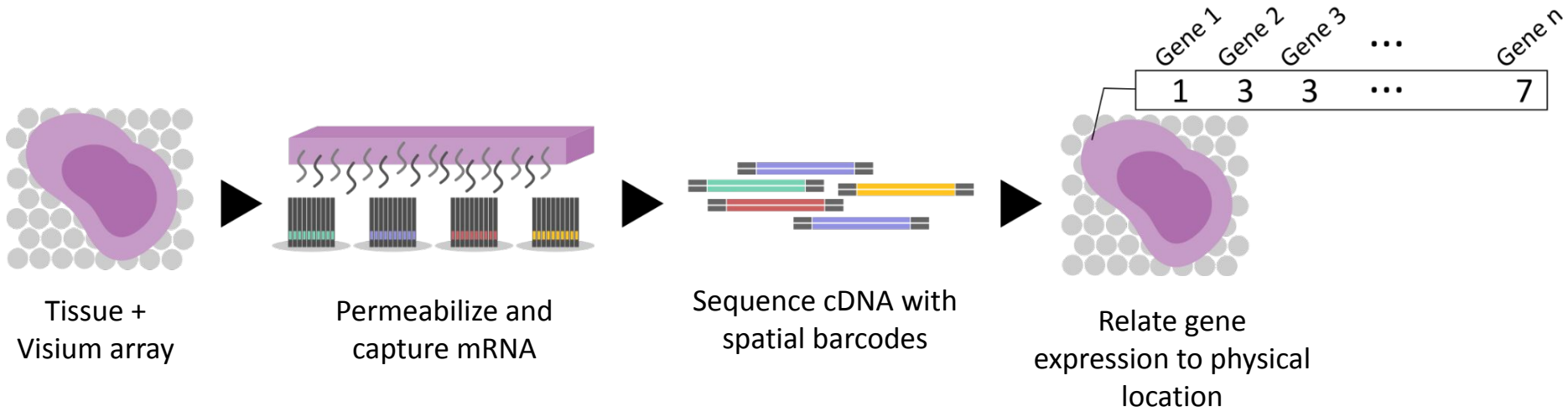
Authors : Lambda Moses and Lior Pachter

Published : 2021-05-12

Link: https://pachterlab.github.io/LP_2021/

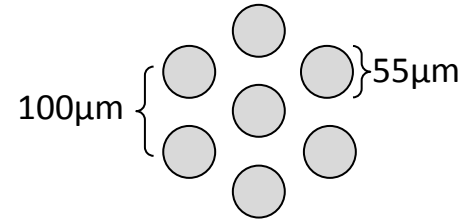
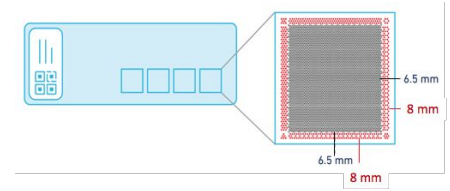


Visium

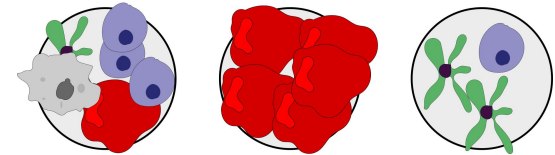


Visium :: Recap

- Array based technique
- 6.5mm x 6.5mm area to put sample on
- 4992 spots arranged in hexagonal grid
- Array specs:
 - Spot diameter : 55 μ m
 - Center to center distance : 100 μ m
- Successor to Spatial Transcriptomics (ST)
- Data processing often includes :
 - Genome mapping and annotation
 - Spatial barcode demultiplexing
- Approx. 1-10 cells contribute to each spot
 - **NOTE** : Not single cell resolution!
- Data represented as [spot] x [gene] matrix
- You also get HE-image of the **same** tissue

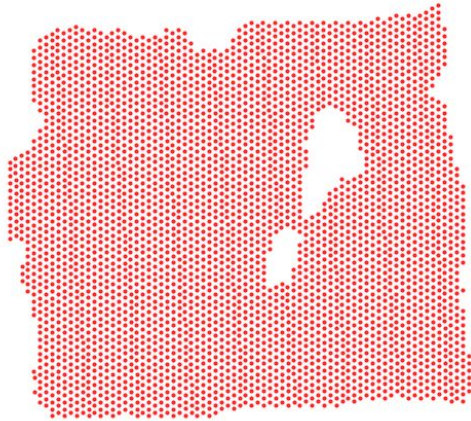


Spot 1 Spot 2 Spot 3

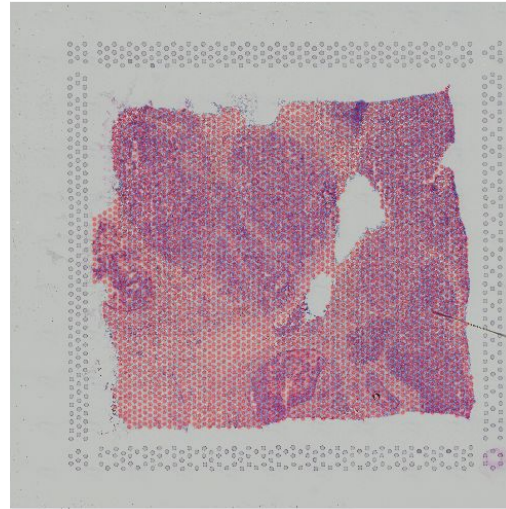


- Example with Human Breast cancer data
 - Public data : Available at 10x website

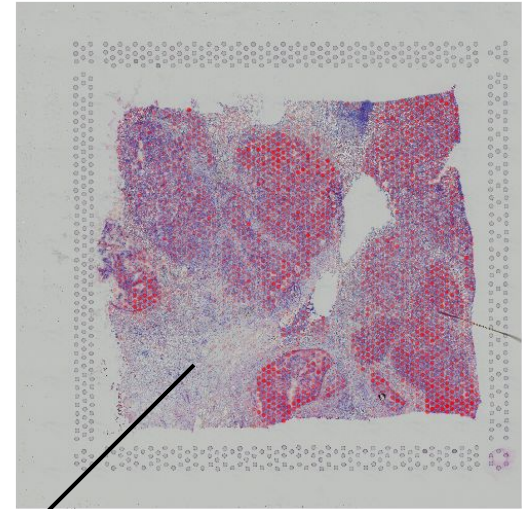
Spots Only



Spots + HE image



Spots + ERBB2 expression + HE image

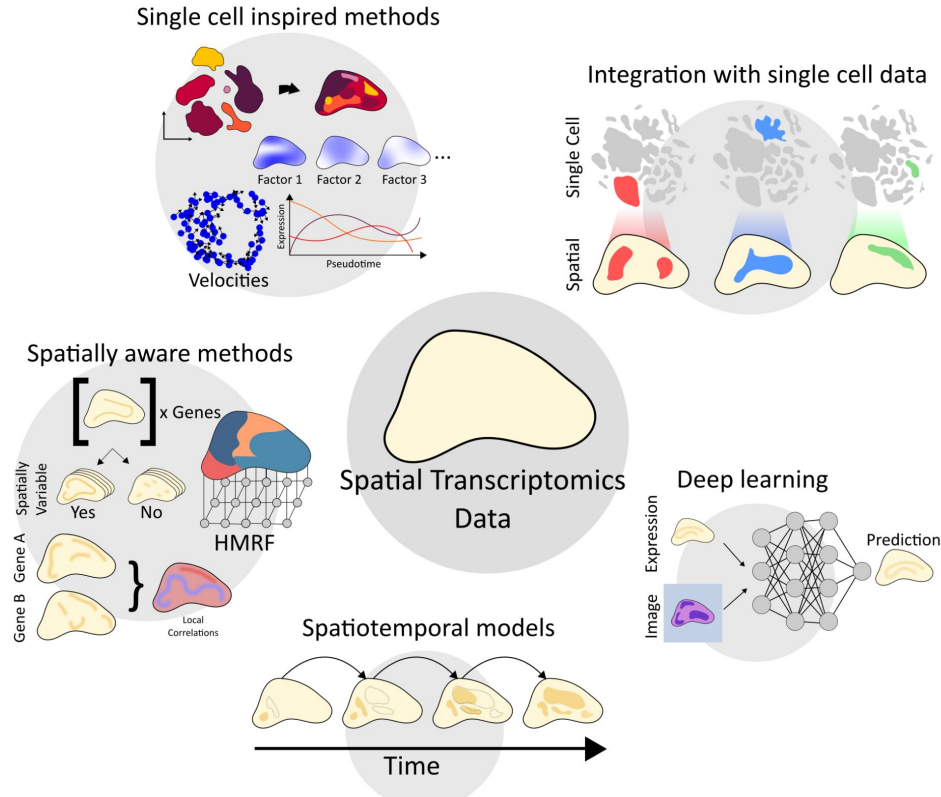


Facecolor intensity proportional
to gene expression value

A grayscale topographic map with contour lines and elevation values. The map shows a complex terrain with numerous peaks and valleys. The text "Computational Analysis" is centered over the map.

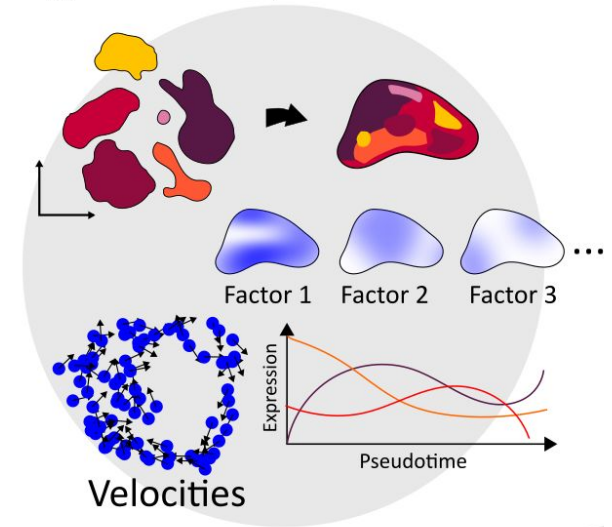
Computational Analysis

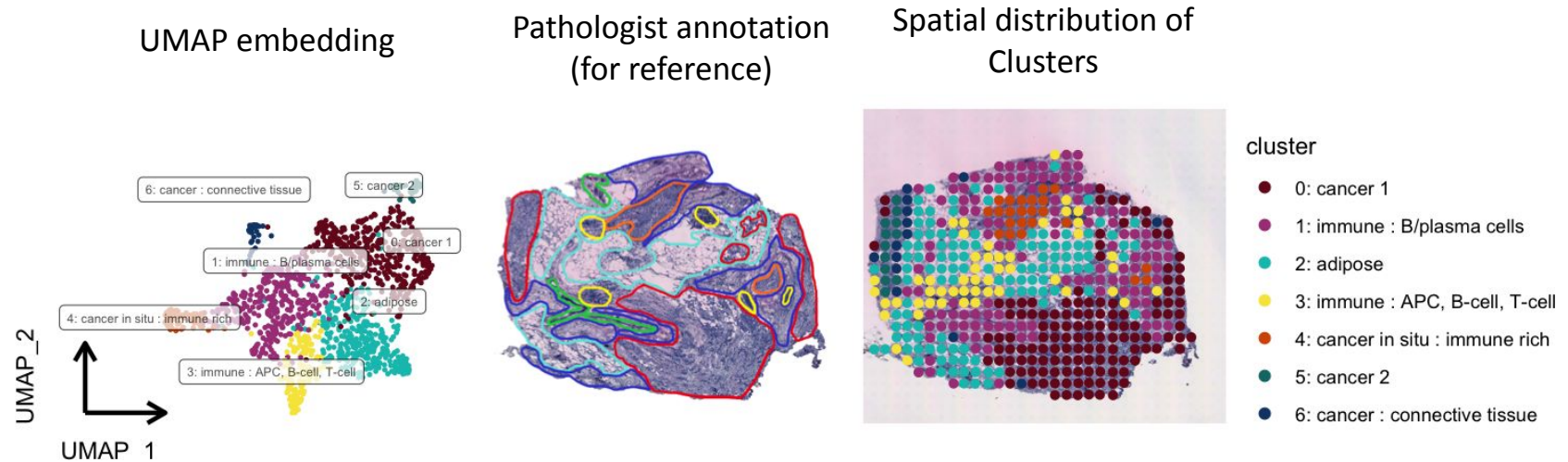
A motley crew of diverse methods



- **Basic idea** : apply existing methods and tools developed for single cell data.
- **Examples** :
 - Cluster spatial data, show clusters in space
 - Factor models for data decomposition
 - Trajectory Inference
- **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)
 - And many many more...

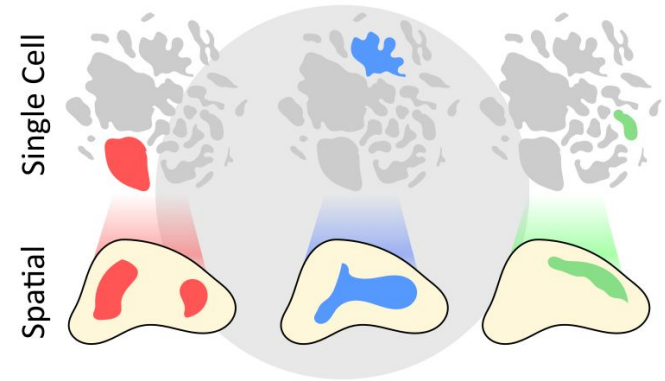
Single cell inspired methods



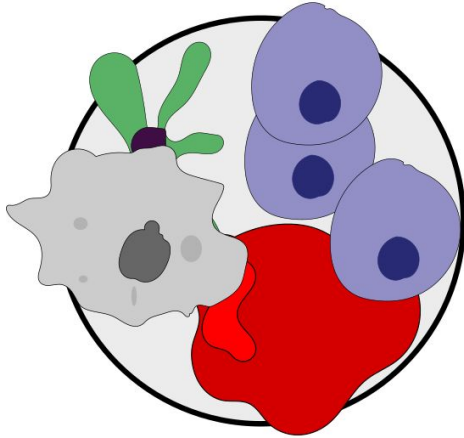


- **Basic idea** : use single cell data as a *reference* when working with spatial data.
- **Answers** : Where are cell types in SC data found in space?
- **But why?** Two main reasons :
 - **Efficient use of resources**. Leverage extensive annotation work done for single cell data.
 - Problem of **mixed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

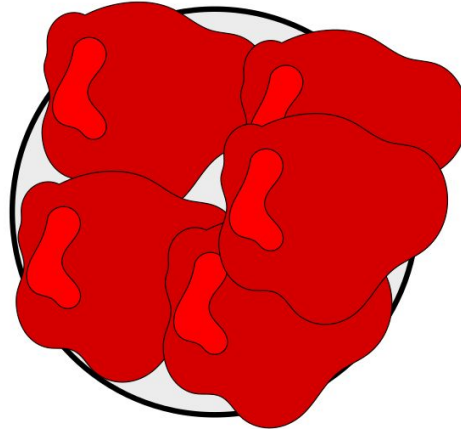
Integration with single cell data



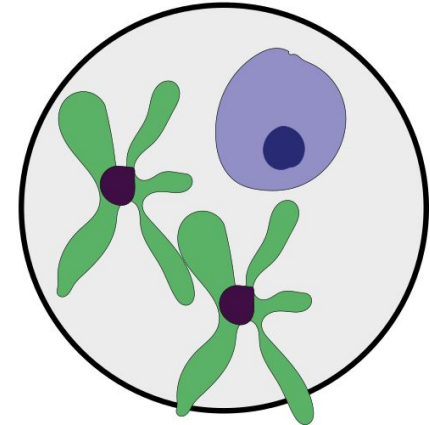
Spot 1



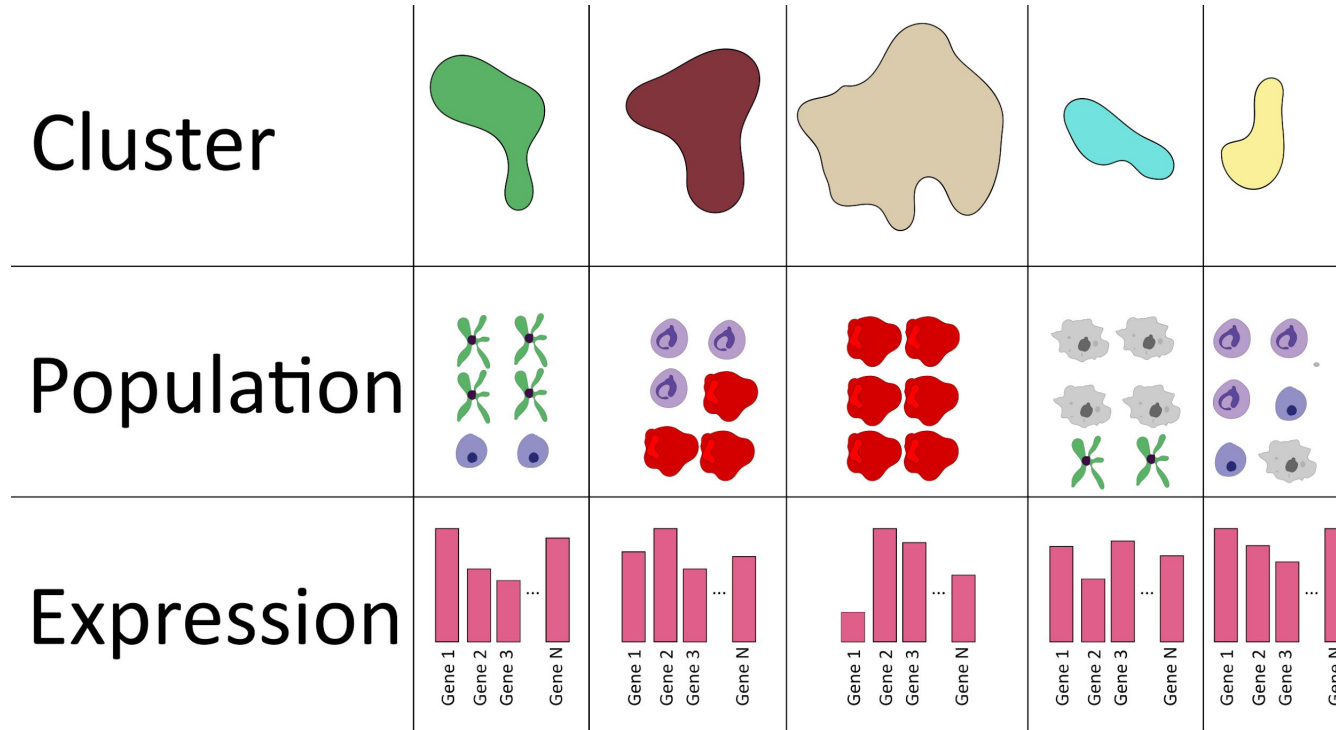
Spot 2



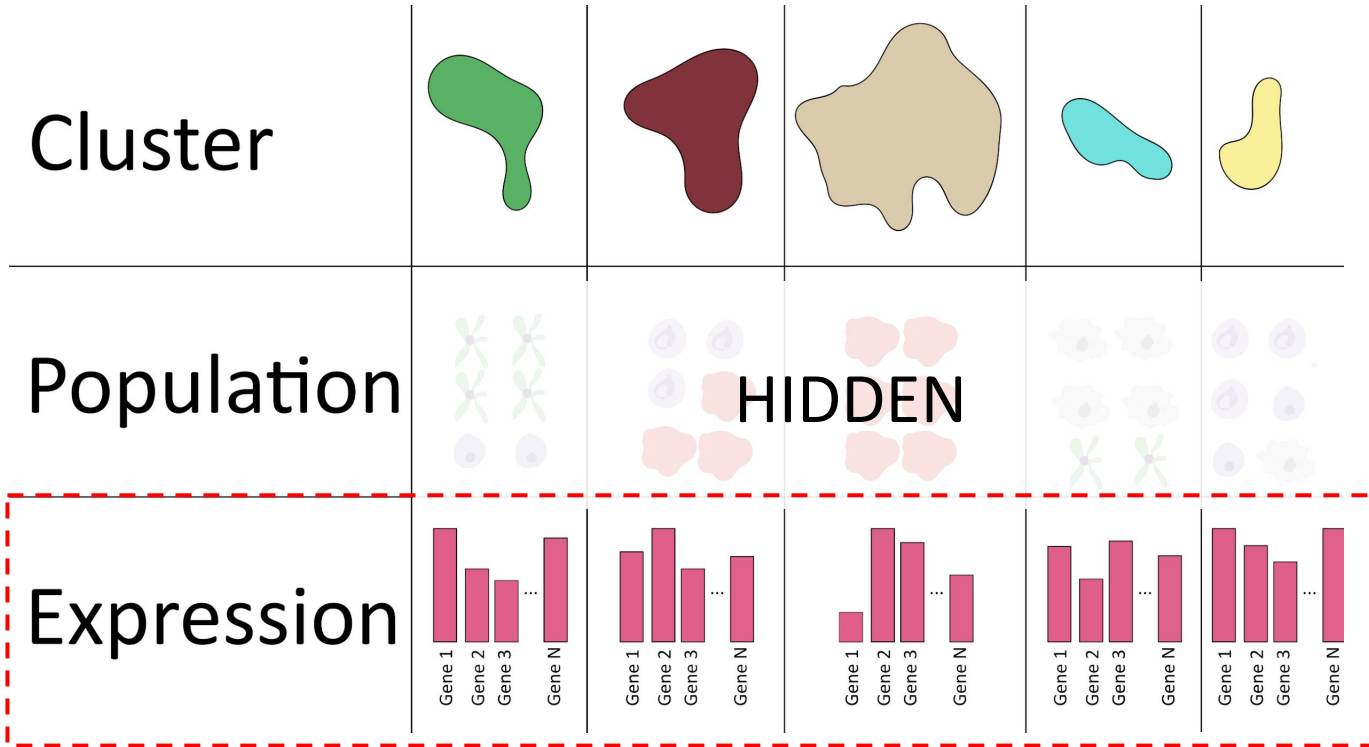
Spot 3



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.



- 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**

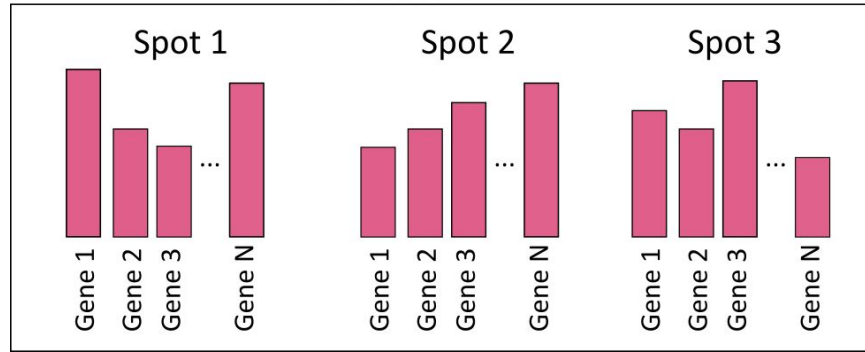


HIDDEN

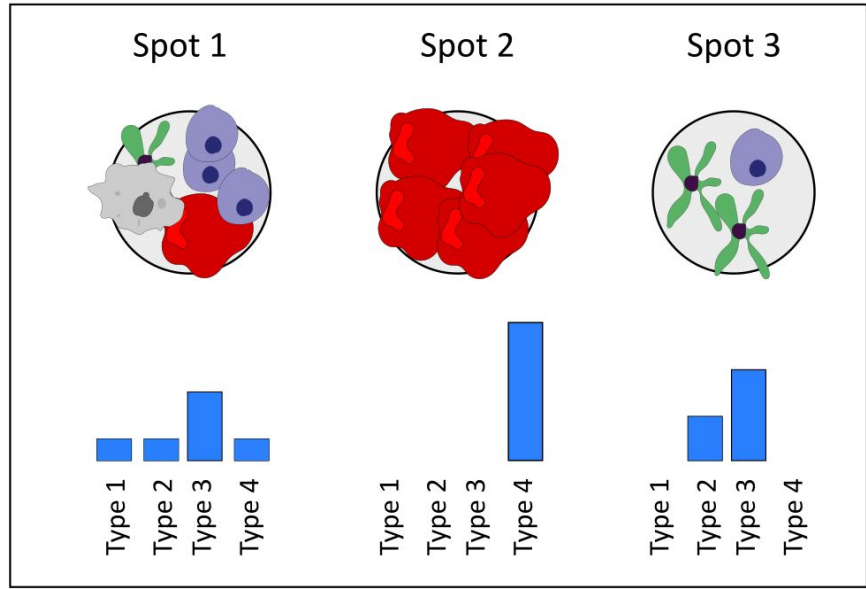
- Clusters **do not represent cell types**
- Clusters are more an assembly of spots with **similar composition** of cell types.
- We have no idea what the cell type population looks like. **Only observe expression**

Observed

From this



We want this



Integration with single cell data

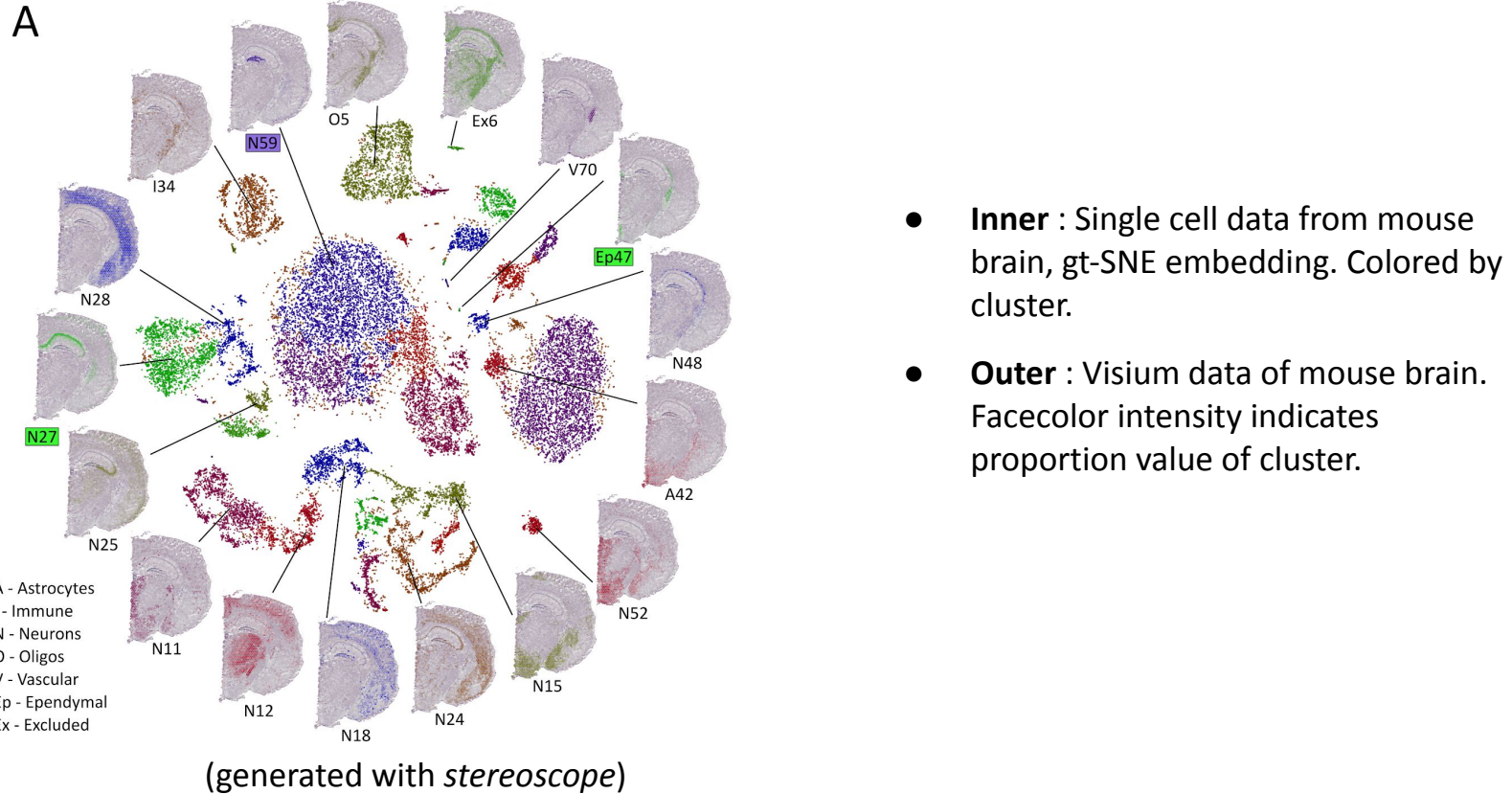


Figure 2 from “Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography”, Andersson et al.

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: Moncada 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.

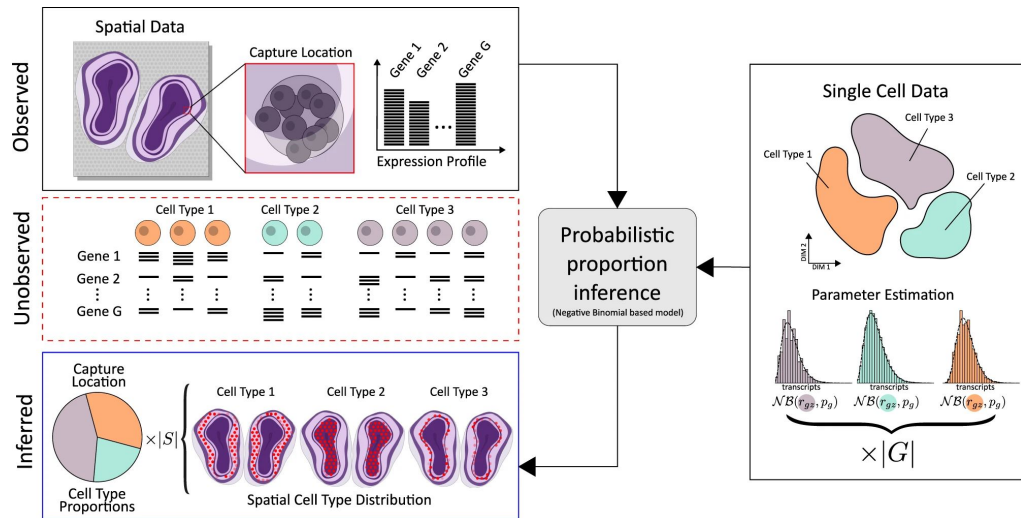
Tries to simultaneously optimize terms such as:

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

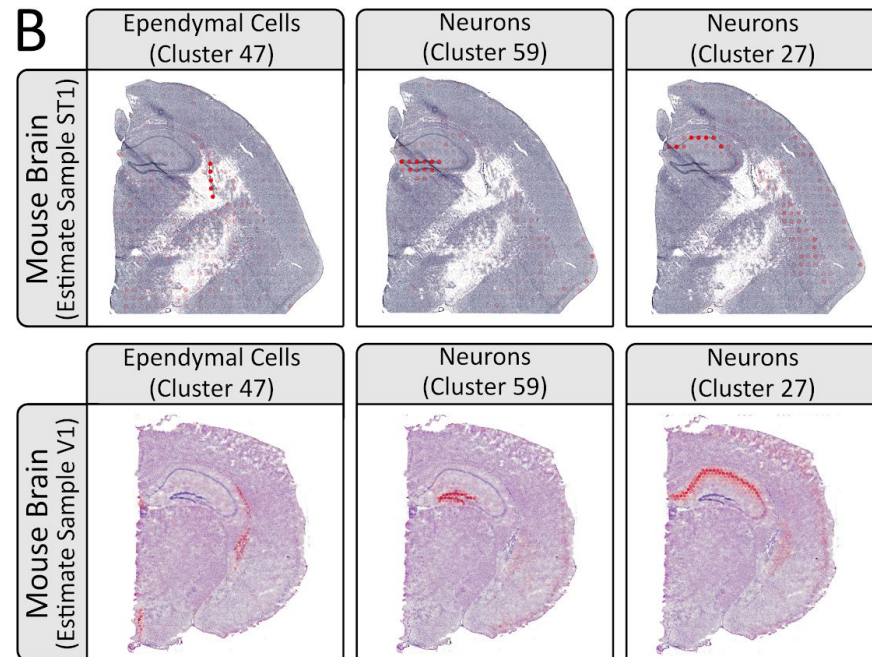
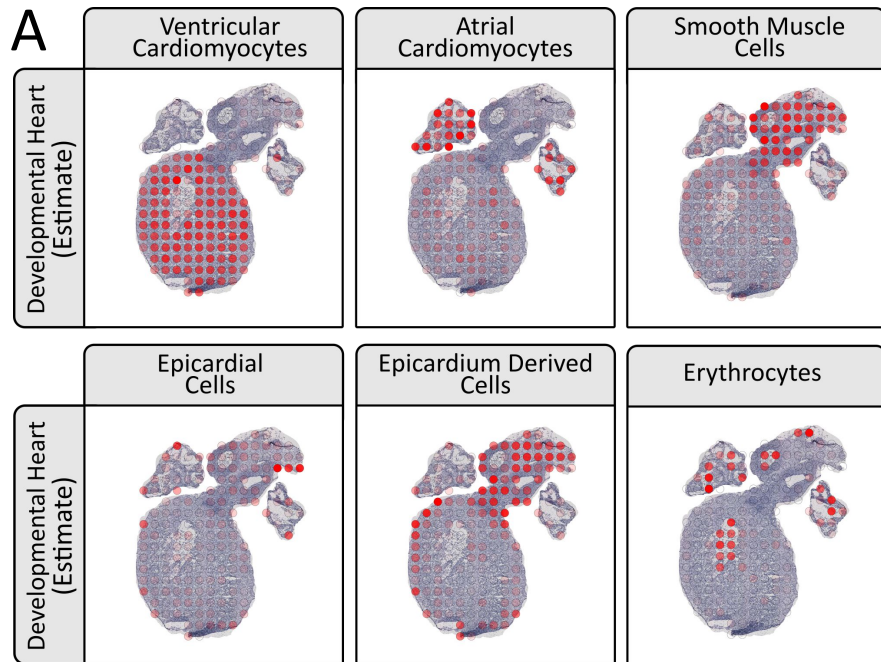
Ex: Tangram

Integration with single cell data :: *stereoscope*

- Probabilistic, models single cell and spatial transcriptomics data with **negative binomial** distribution
- Two-step process:
 1. Learn parameters from sc-data
 2. Infer proportions in spatial data
- Parameters from single cell data can be reused, cut computational time in half.
- Accounts for missing cell types by including a “dummy cell types”
- “Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography”, Communications Biology, Andersson et al.



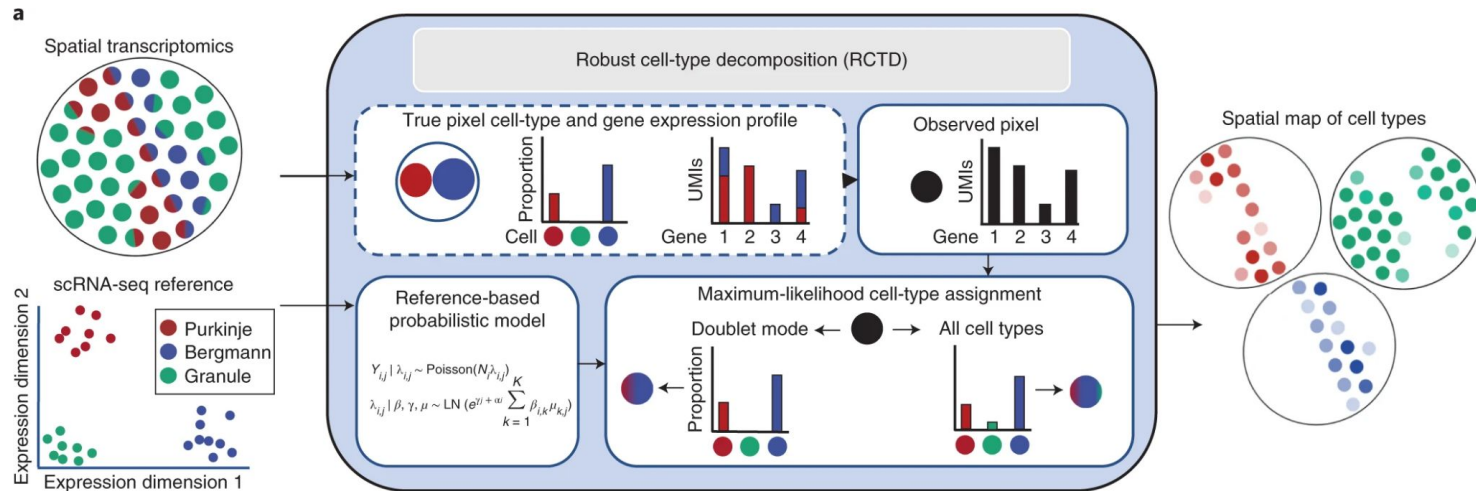
<https://github.com/almaan/stereoscope>



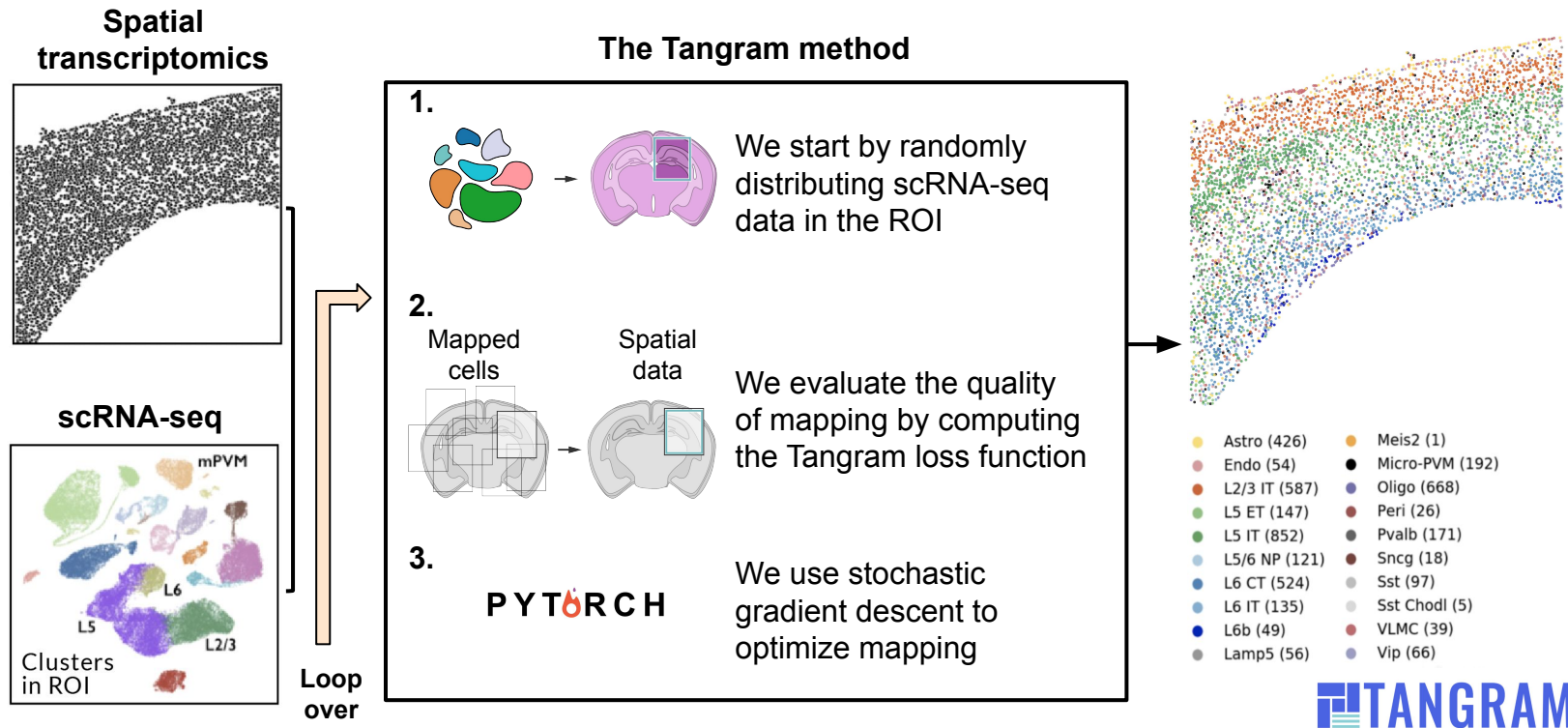
Developmental heart : DOI: [10.1016/i.cell.2019.11.025](https://doi.org/10.1016/i.cell.2019.11.025)

Mouse Brain : 10X Genomics website + mousebrain.org

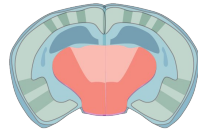
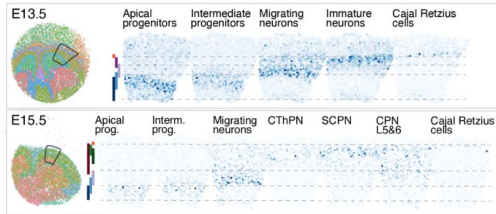
Integration with single cell data :: RCTD



- Probabilistic model for **inferring cell types in spatial transcriptomics data**, supervised with a labeled single-cell RNA-seq reference.
- Infers **platform effects** (or technical differences across sequencing platforms) in order to correct for differences between the single-cell reference and the spatial target dataset.
- RCTD uses maximum likelihood estimation to **identify cell types present on each spatial transcriptomics spot, in addition to estimating cell type proportions**.
- Robust decomposition of cell type mixtures in spatial transcriptomics, Nat. Biotech, Cable et al.



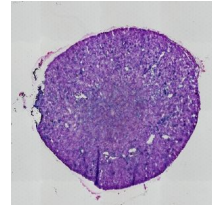
Spatial maps of cell types in developmental mouse brain



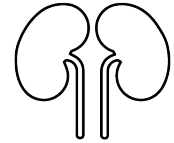
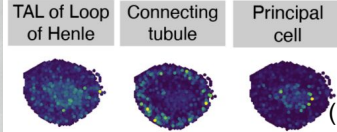
with Paola Arlotta lab
(Nature 2021 in press)

Assessing cross-species conservation in kidney

Mouse kidney histology



Cell type maps in kidney



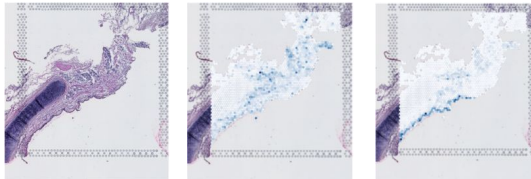
with Aviv Regev lab
(Nature Methods 2021 in press)

Localization of epithelial cell types in human lung

Histology

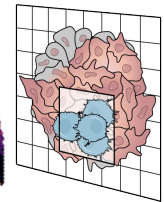
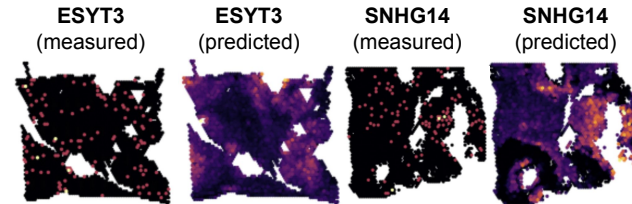
Non-epithelial cell map

Epithelial cell map



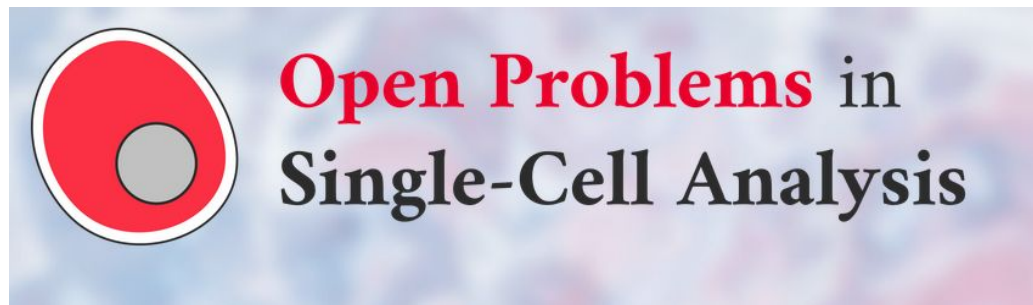
with Jay Rajagopal lab
(in preparation)

Correction of gene expression in colorectal cancer



with Fred De Sauvage

- Initiative to formalize problems in single cell (and spatial) analysis. Includes *proper* definition.
- Provide datasets for unbiased evaluation of data, and define metrics to be used.
- Build framework for said evaluation.
- Allows you to make informed choice.
- <https://openproblems.bio/>



[new task] Spatial decomposition #309

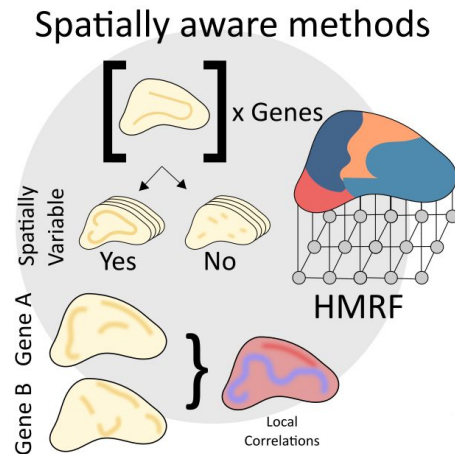


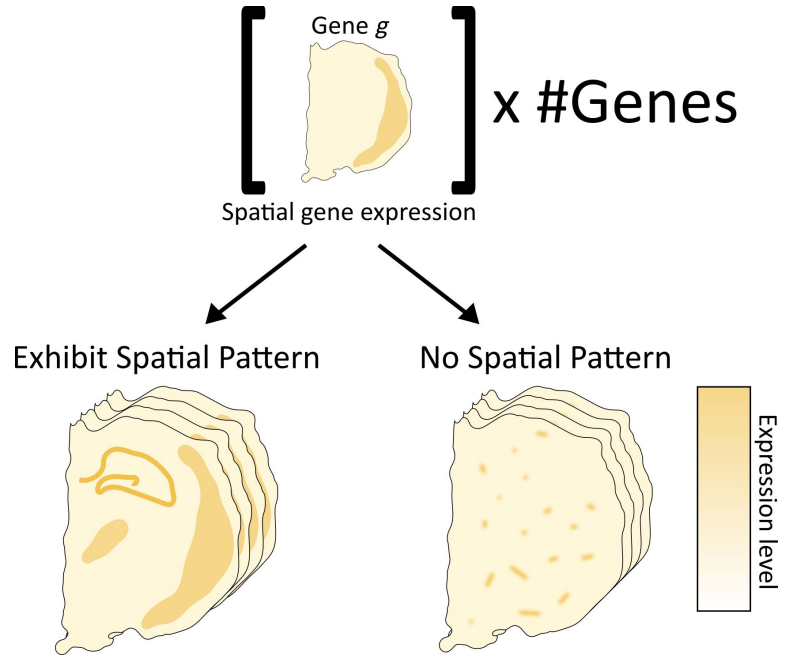
giovp wants to merge 47 commits into `openproblems-bio:main` from `giovp:master`

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

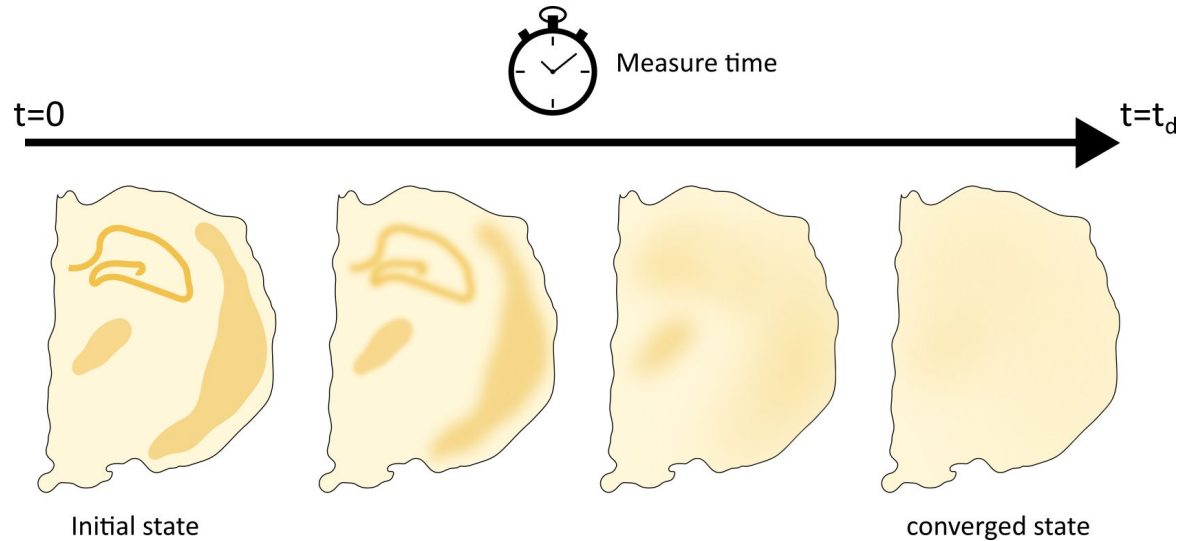
Designed for tasks like :

- 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

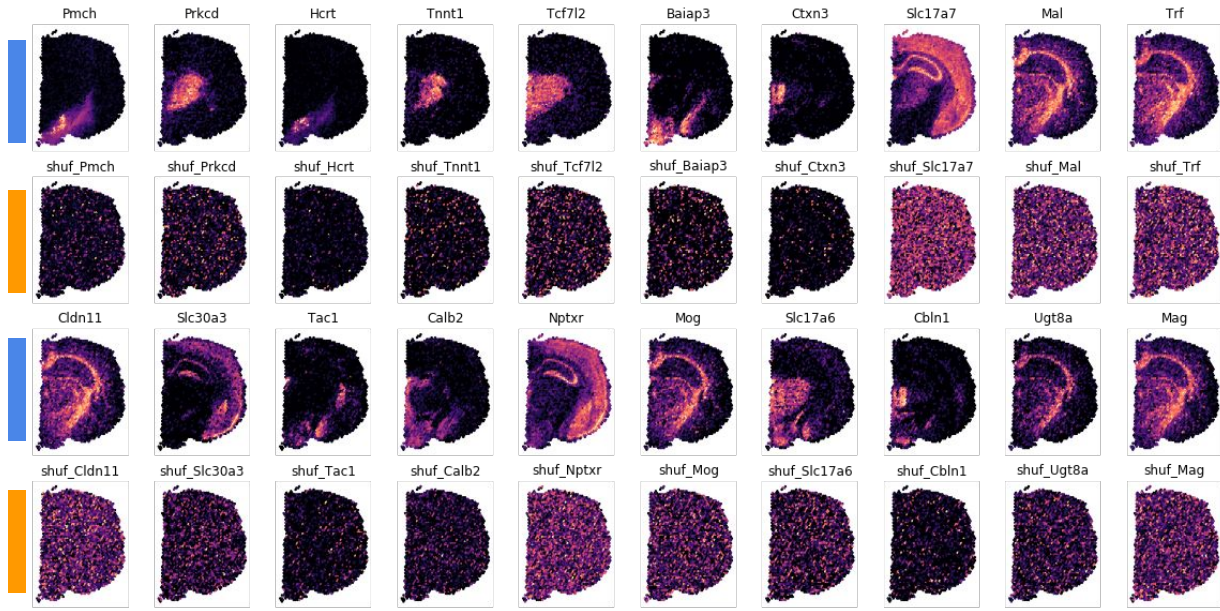





- Sort expression profiles into spatially variable or not.
- SpatialDE, SVCA and SPARK use probabilistic models
- Leverage *Gaussian Processes* to model data
- Essentially, test whether a “spatial” term in the covariance function significantly increase model’s ability to explain data

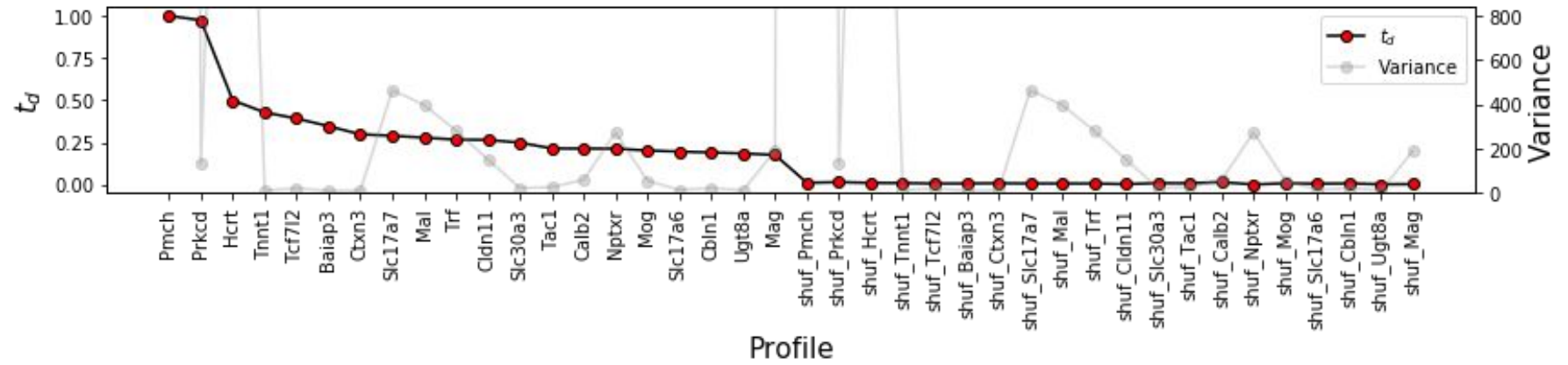


- *sepal* is not probabilistic
- Uses finite differences to simulate diffusion of transcripts.
- Measures time until converges
- Ranks genes by the time it takes to converge.
- **Key Idea** : The longer the time, the more structured the initial state.

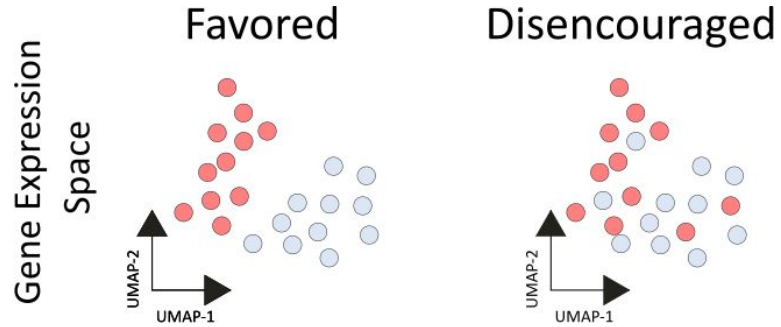


- 20 Expression profiles from mouse brain
- Shuffle spots to get random expression profiles. Has “shuf” prefix.

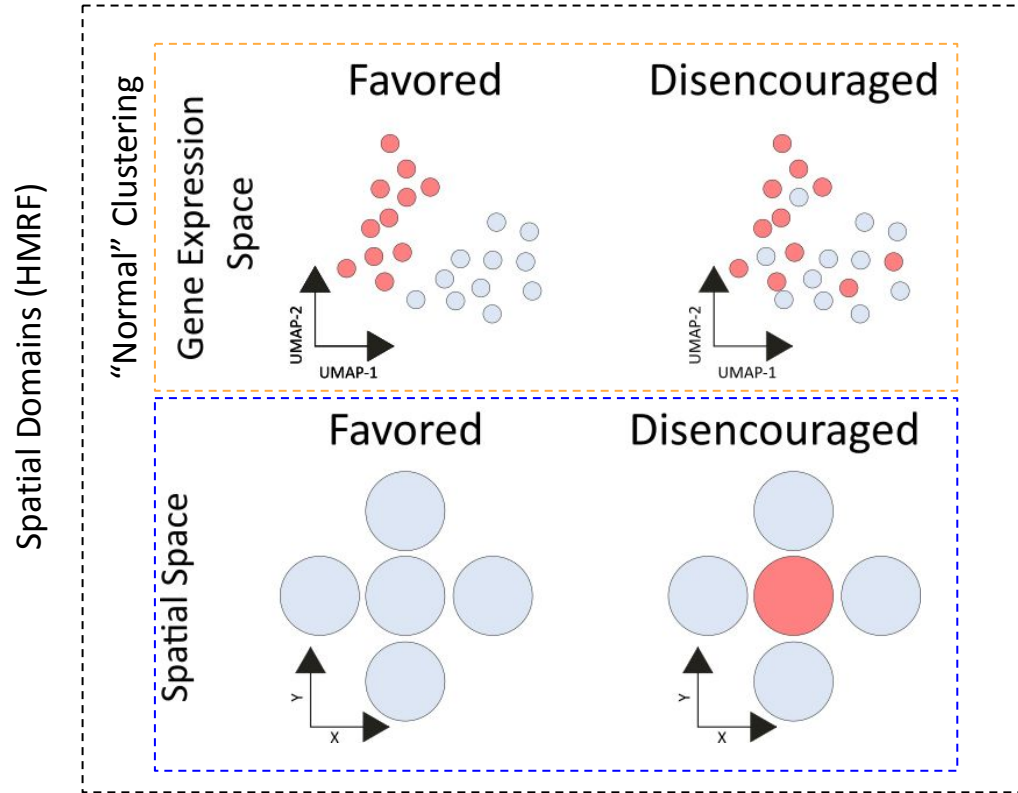
 Observed Profiles  Shuffled Profiles



- 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)

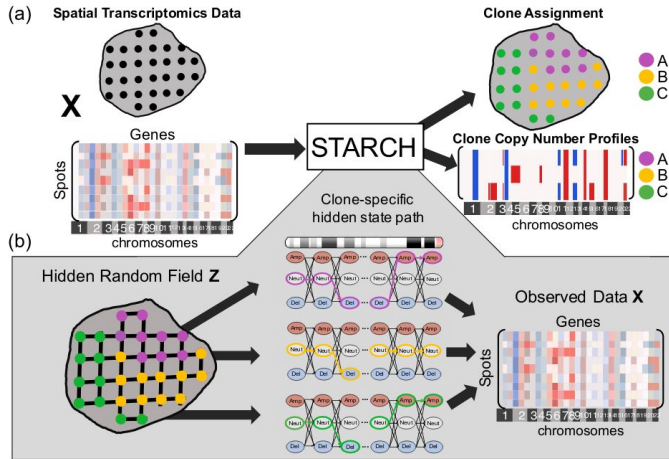


- Normal clustering mainly focus on gene expression

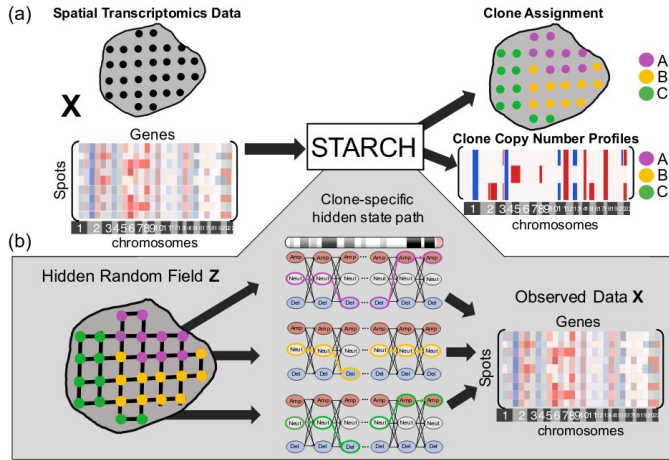


- Normal clustering mainly focus on gene expression
- Leverage spatial information to find spatially coherent clusters (domains)
- Common to use HMRF (Hidden Markov Random Field)
- Construct a graph based on spatial proximity
- Probability of node (spot) belonging to a specific domain depends on:
 -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

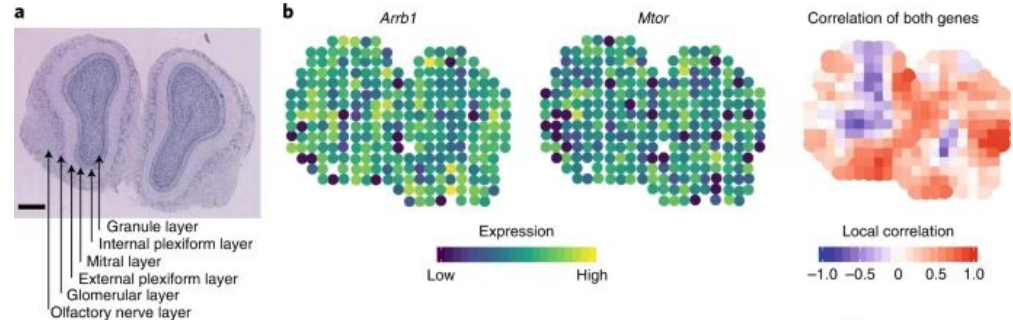


- **Name** : STARCH
- Infer Copy Number Aberrations (CNA) from spatial transcriptomics data
- Increase robustness of inference by aggregating data in same domains (similar profiles)
- Also uses Hidden Markov Random Fields (HMRF)
- *“STARCH: Copy number and clone inference from spatial transcriptomics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

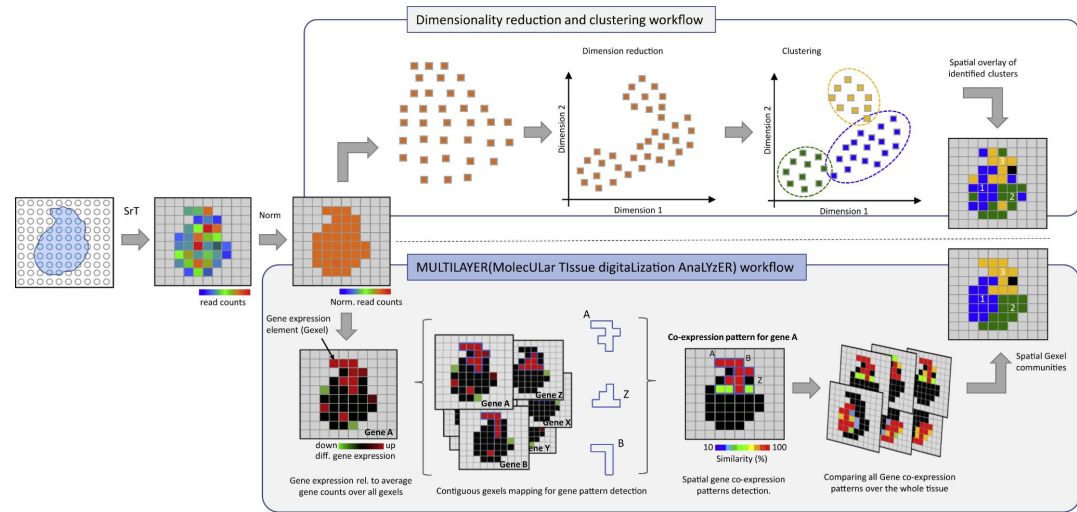


- **Name** : STARCH
- Infer Copy Number Aberrations (CNA) from spatial transcriptomics data
- Increase robustness of inference by aggregating data in same domains (similar profiles)
- Also uses Hidden Markov Random Fields (HMRF)
- *“STARCH: Copy number and clone inference from spatial transcriptomics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

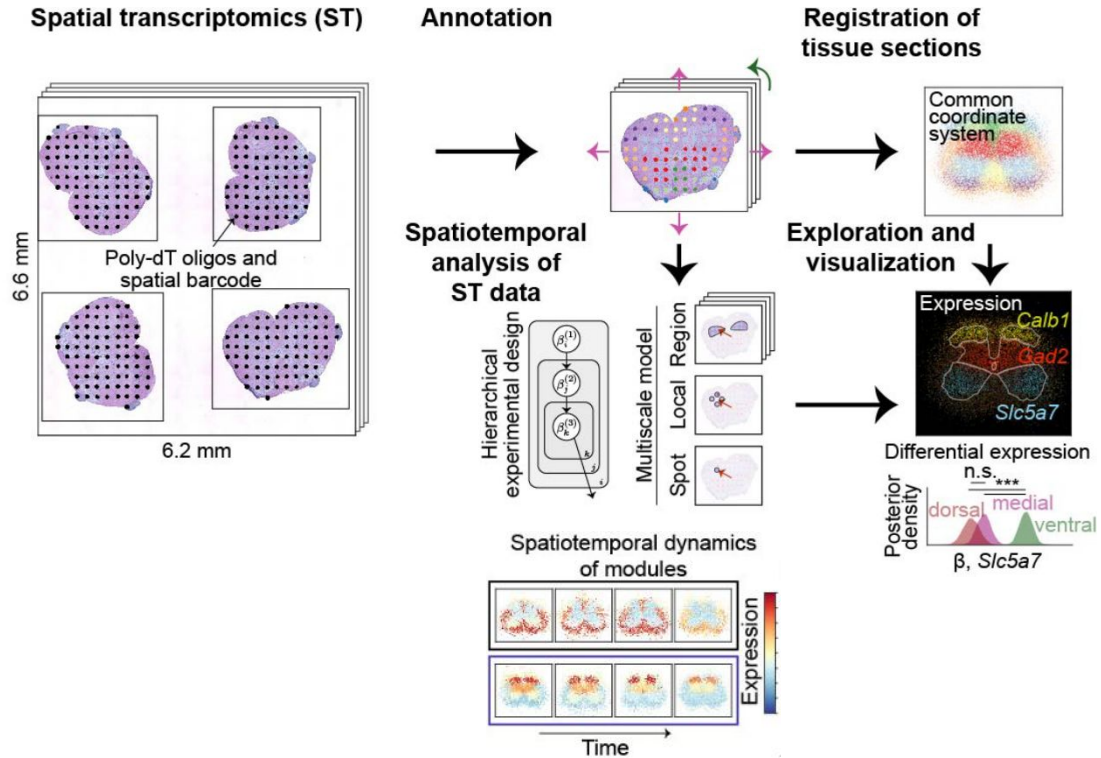
- **Name** : scHOT
- Computes (spatially) weighted correlations to find local correlations.
- *“Investigating higher-order interactions in single-cell data with scHOT”, Ghazanfar et al.*



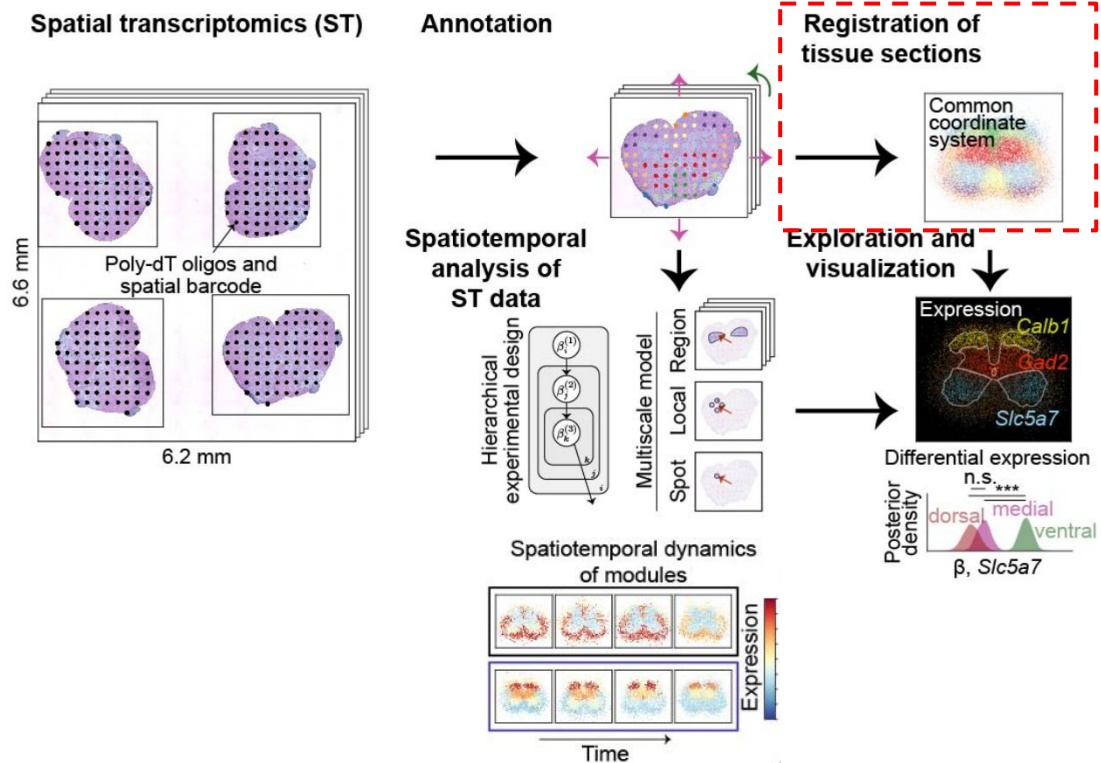
- Inspired by digital image processing. Introduces terminology of “gexels”.
- Looks at relative gene expression of each gexel compared to the rest. Find locally up and down-regulated genes.
- Uses agglomerative clustering to find contiguous patterns that share similar structures (co-expression modules)
- Extracts communities (clusters) from co-expression modules by using Louvain clustering.

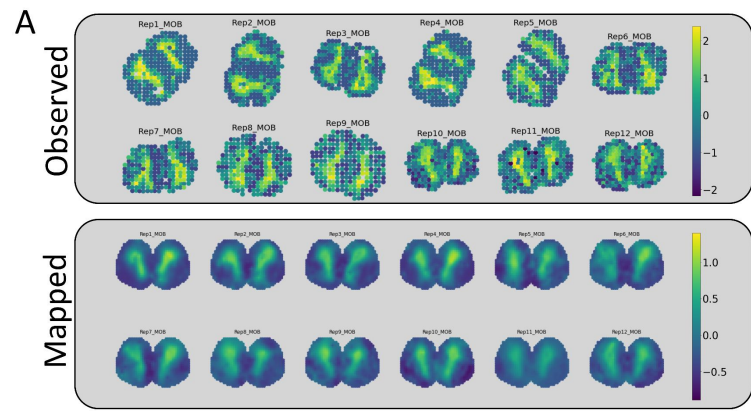


- Hierarchical generative probabilistic model for analyzing Spatial Transcriptomics data
- Uses Zero Inflated Poisson (ZIP) regression model to account for:
 - Tissue region context
 - Local components
 - Spot effects
- Also aligns sections
- Can identify genes that changes over both space and time

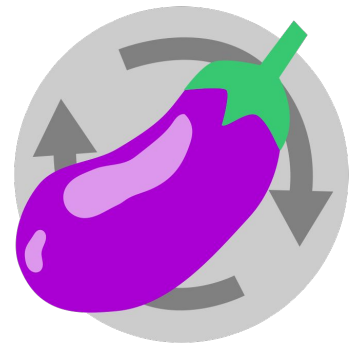
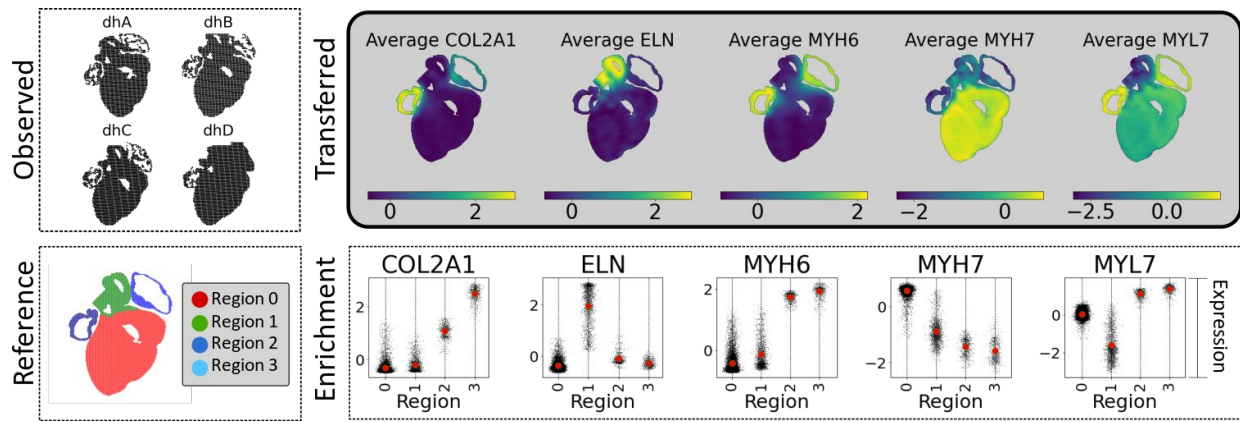


- Hierarchical generative probabilistic model for analyzing Spatial Transcriptomics data
- Uses Zero Inflated Poisson (ZIP) regression model to account for:
 - Tissue region context
 - Local components
 - Spot effects
- Also aligns sections
- Can identify genes that changes over both space and time





- New method (*eggplant*) that maps features (gene expression, cell type proportions) to a *common coordinate framework (CCF)*
- Allows user to define a reference and then transfer feature values to it
- Enables spatiotemporal modeling and facilitates construction of atlases

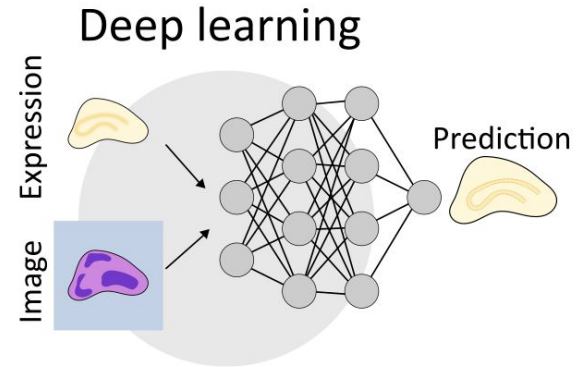


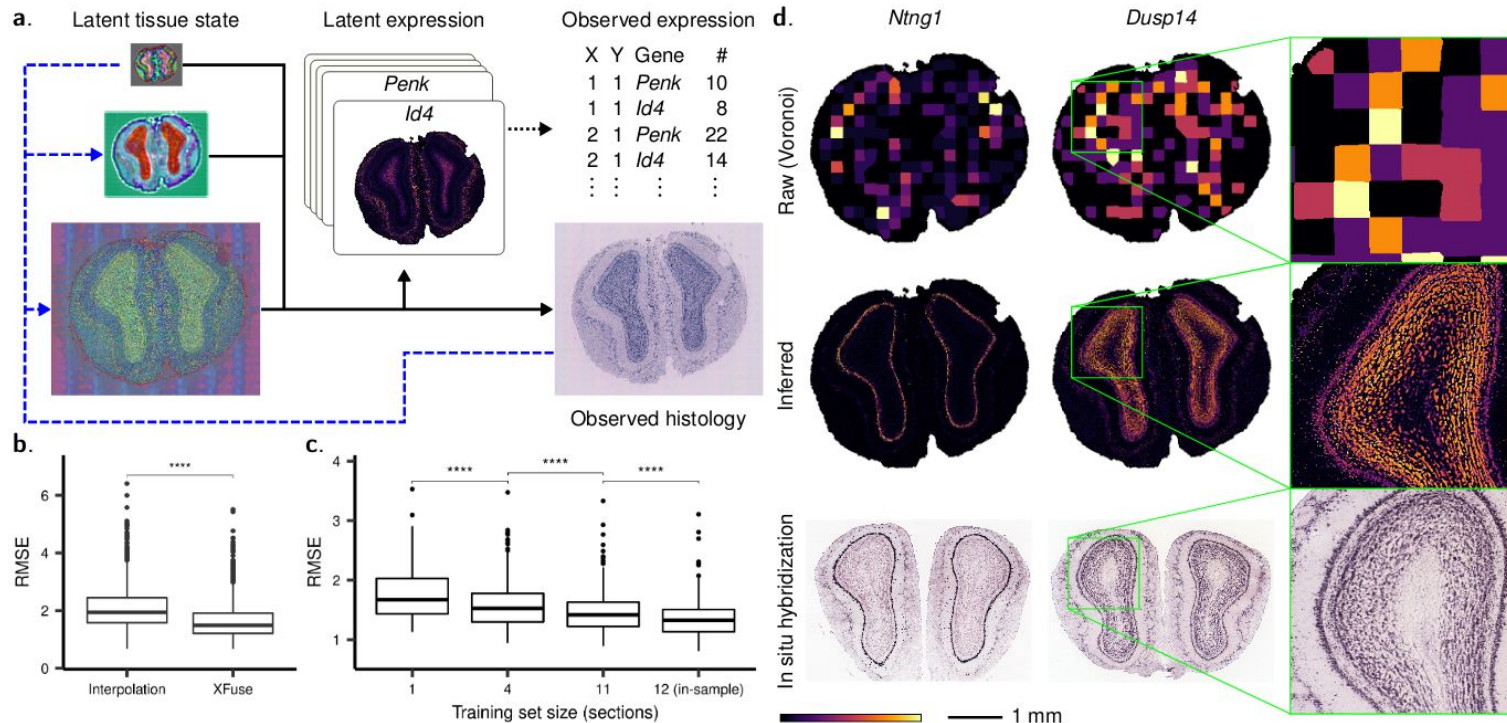
Basic Idea : Apply deep learning to spatial data (very broad)

Fairly nascent : Relatively few examples. Limited amount of high quality available data.

Current examples :

- **XFuse** : “*superresolution*” (pixel) of gene expression by learning joint representation of image and expression data.
- **stPlus** : Uses scRNA-seq data and autoencoders to enhance spatial transcriptomics data
- **SpaGCN** : simultaneous domain and SVG detection using graph convolution layers
- **RESEPT** : Uses graph convolutional network to embed spatial data in RGB space, then uses a CNN to segment data into spatially coherent tissue domains



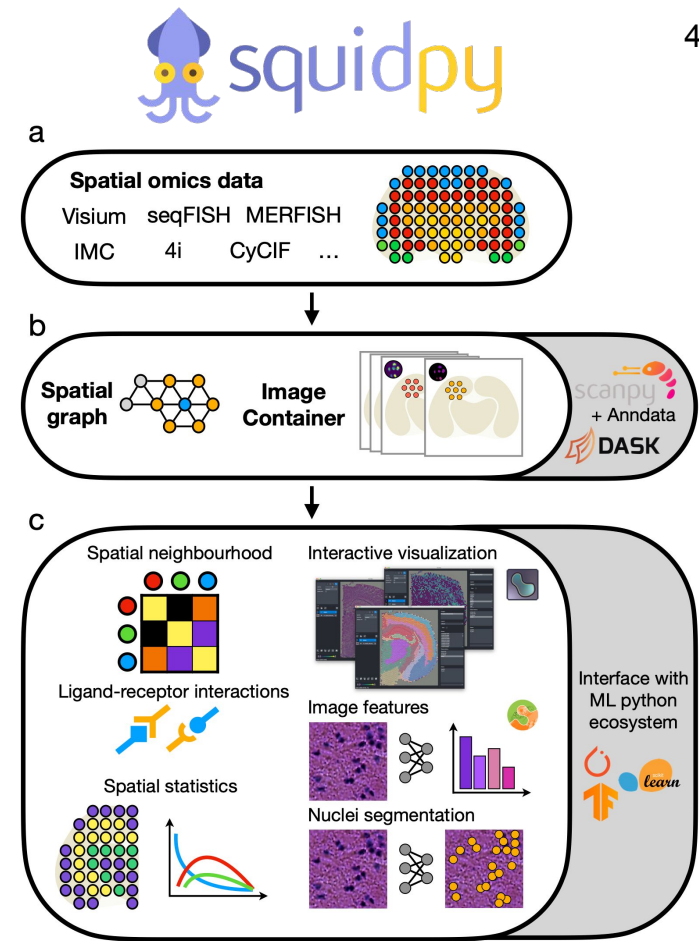


From : “Super-resolved spatial transcriptomics by deep data fusion”, Bergenstr hle et al. (in press Nature Biotechnology)

Computational suites :: squidpy

“One framework to rule them all, one framework to find them...”

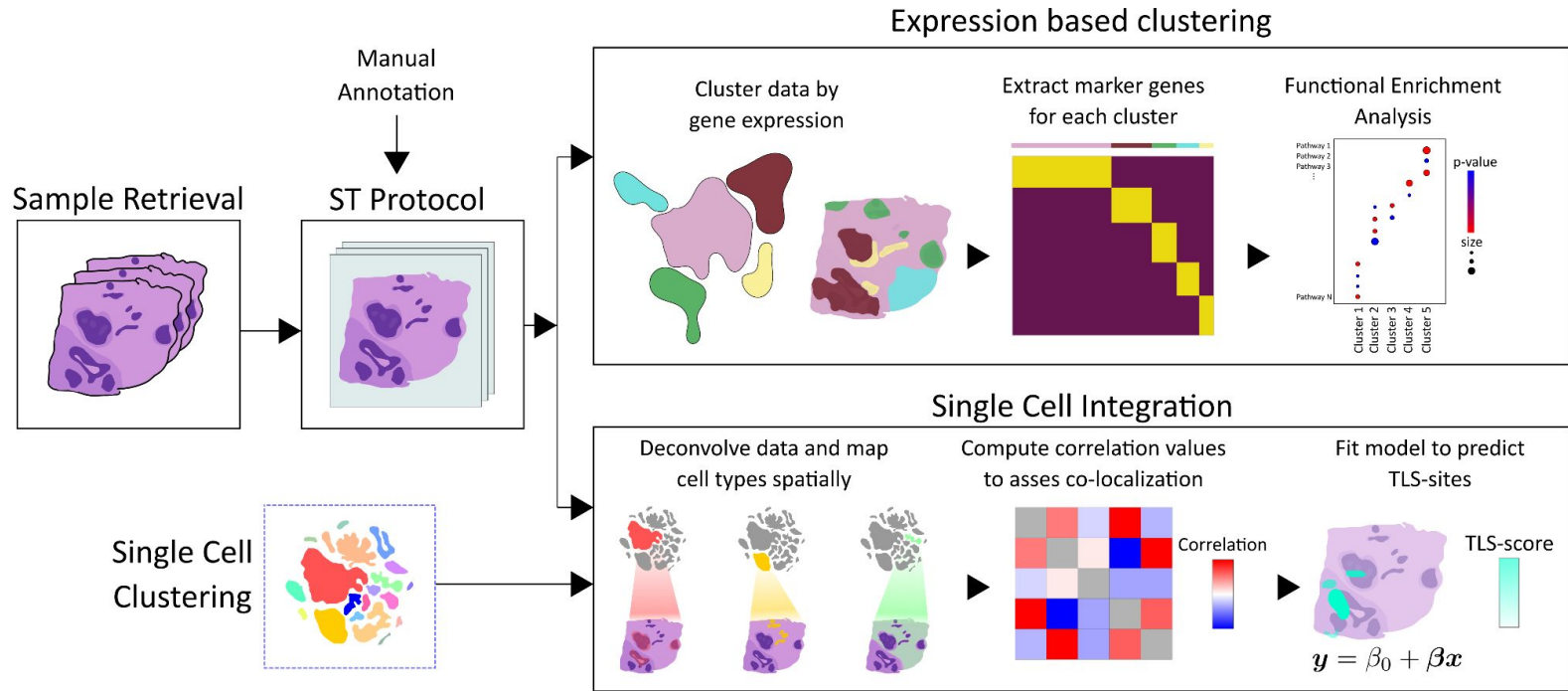
- Similar philosophy as scanpy, uses same kind of API, built on AnnData objects
- Tailored towards spatial data with support for multiple different experimental platforms (not only Visium)
- Easy to construct spatial graphs and perform graph operations
- Has great interface with ML ecosystems such as PyTorch, TensorFlow and sklearn
- Simplified my life a lot and something I tend to use now in method development
- (Also has sepal integrated into the suit)



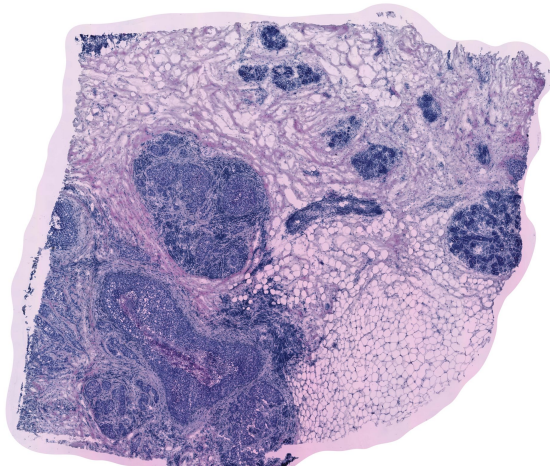


Observations from the wild

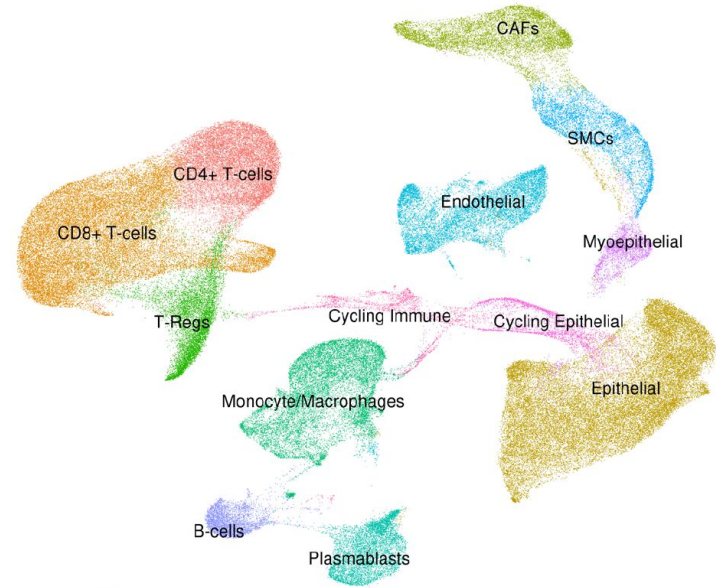
- **Batch effects** between sections are usually observed, try to account for this. Single cell methods have worked great so far.
- **Cell density** is often not homogeneous across tissue. Good idea to normalize based on the library size to account for this.
- Keep in mind that expression profiles are **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'm 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. However, keep them if relevant!
- Use the image to visualize and inspect your data, one of the best quality checks there is. Always ask yourself: "does this make sense"?



“Spatial deconvolution of HER2-positive Breast cancer delineates tumor-associated cell type interactions”, Andersson et al. (in press Nature Communications)

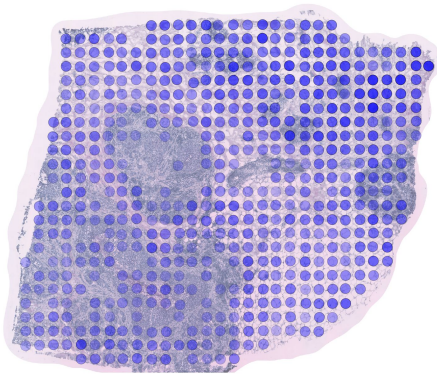


Spatial Data
Human Breast Cancer
HER2-positive

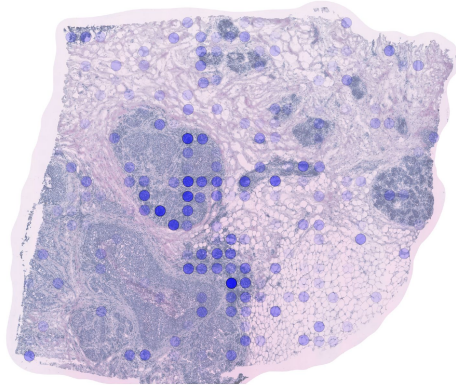


SC Data
Human Breast Cancer
Multiple types (incl. HER2)

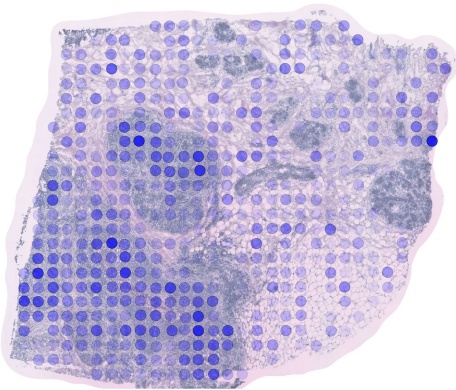
Plasma Cells



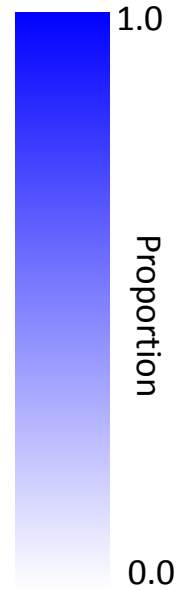
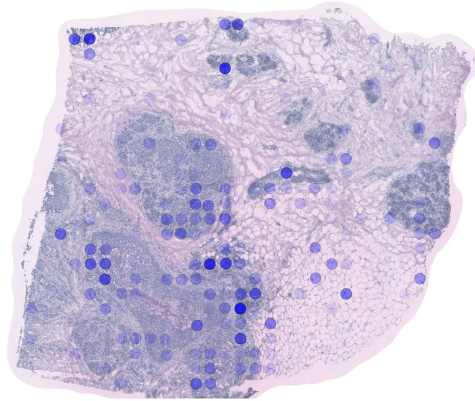
CD4+ T-cells



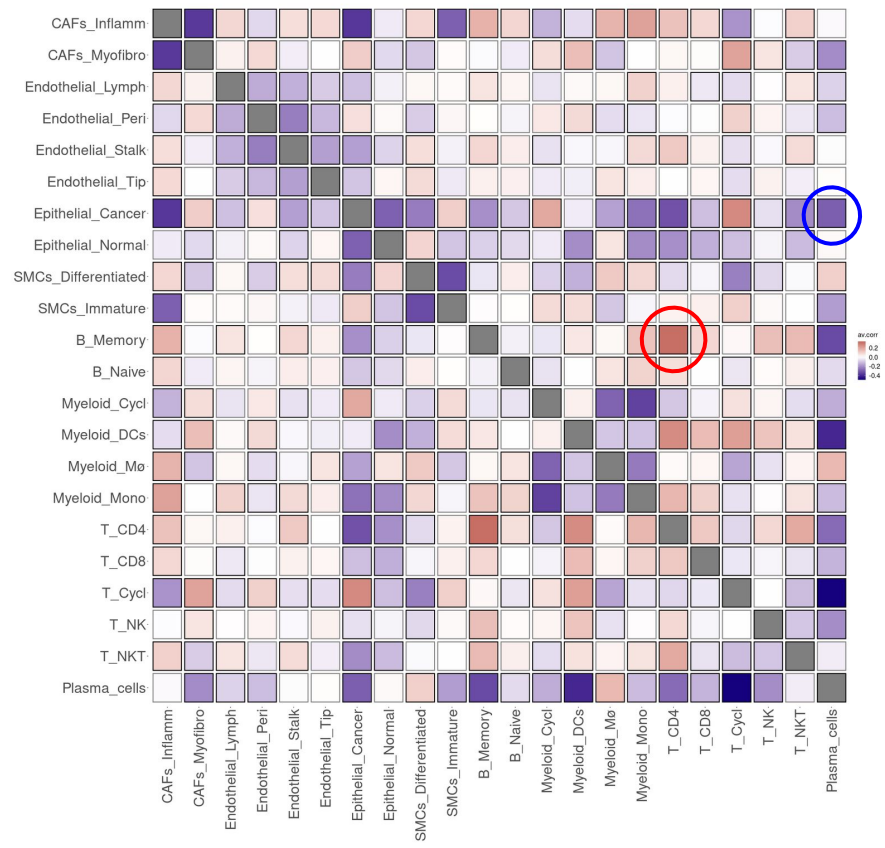
Epithelial Cancer



Memory B-cells



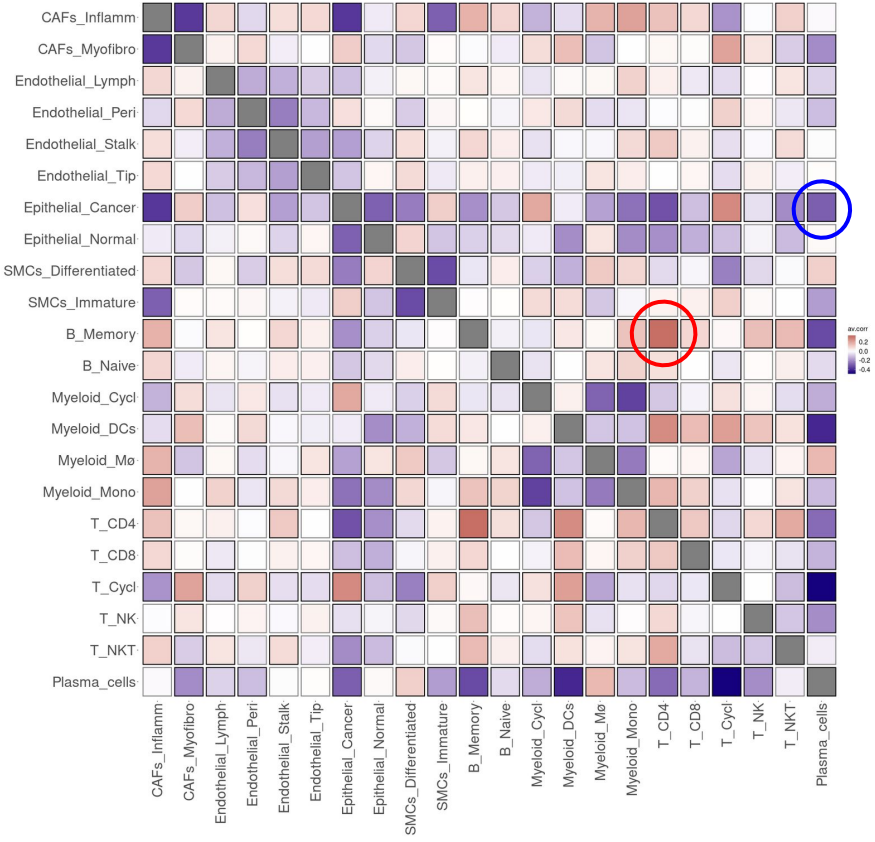
Cell type co-localization



- Find cell types with similar spatial distributions
- Confirms previous observations

- Plasma cells and Epithelial Cancer anticorrelate
- Memory B-cells and CD4+ T-cells co-localize

Cell type co-localization



- Find cell types with similar spatial distributions
- Confirms previous observations

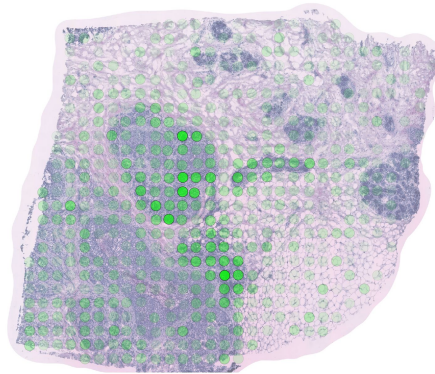
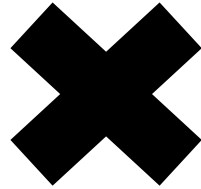
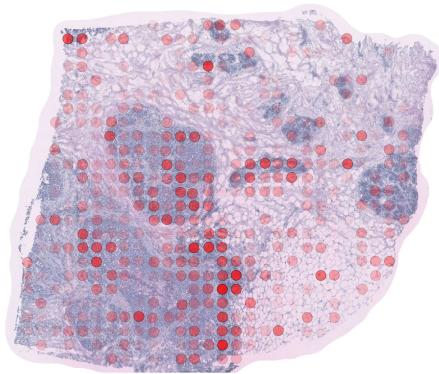
● Plasma cells and Epithelial Cancer
anticorrelate

● Memory B-cells and CD4+ T-cells co-localize

- Characterized by high presence of **B** and **T**-cells
- Interesting for several reasons
 - Partially dictates degree of TILs (Tumor Infiltrating Lymphocytes)
 - Implications in tumor treatment and outcome
- **Question** : Can we locate TLSs in our samples?

B-cells

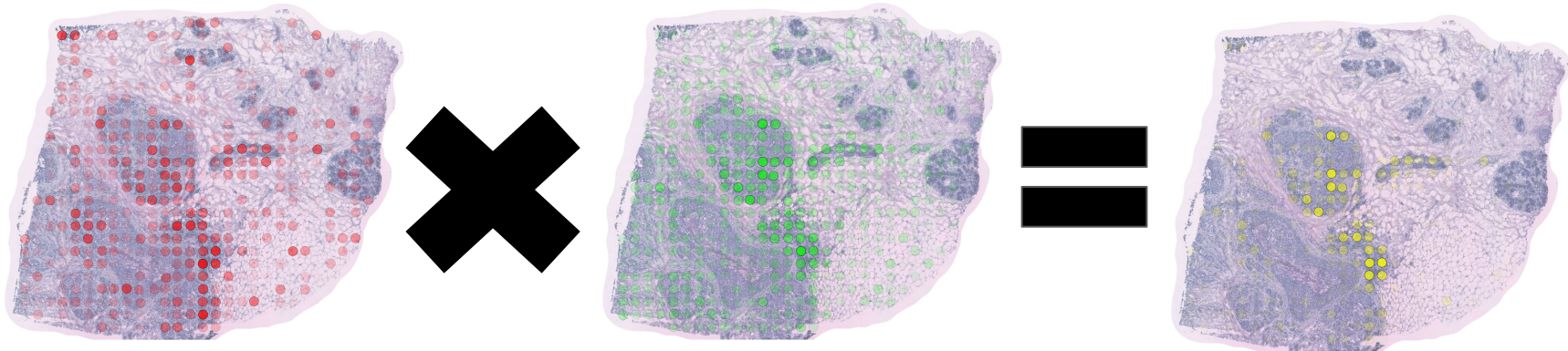
T-cells



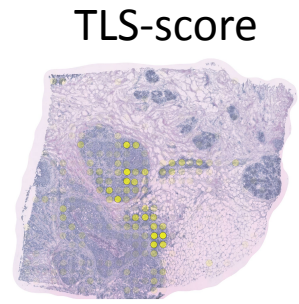
B-cells

T-cells

TLS



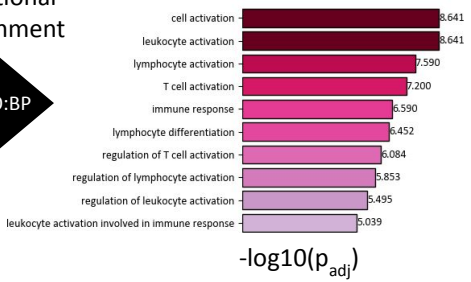
$$\beta_0 + \beta_1[\text{Gene 1}] + \dots + \beta_N[\text{Gene N}] =$$



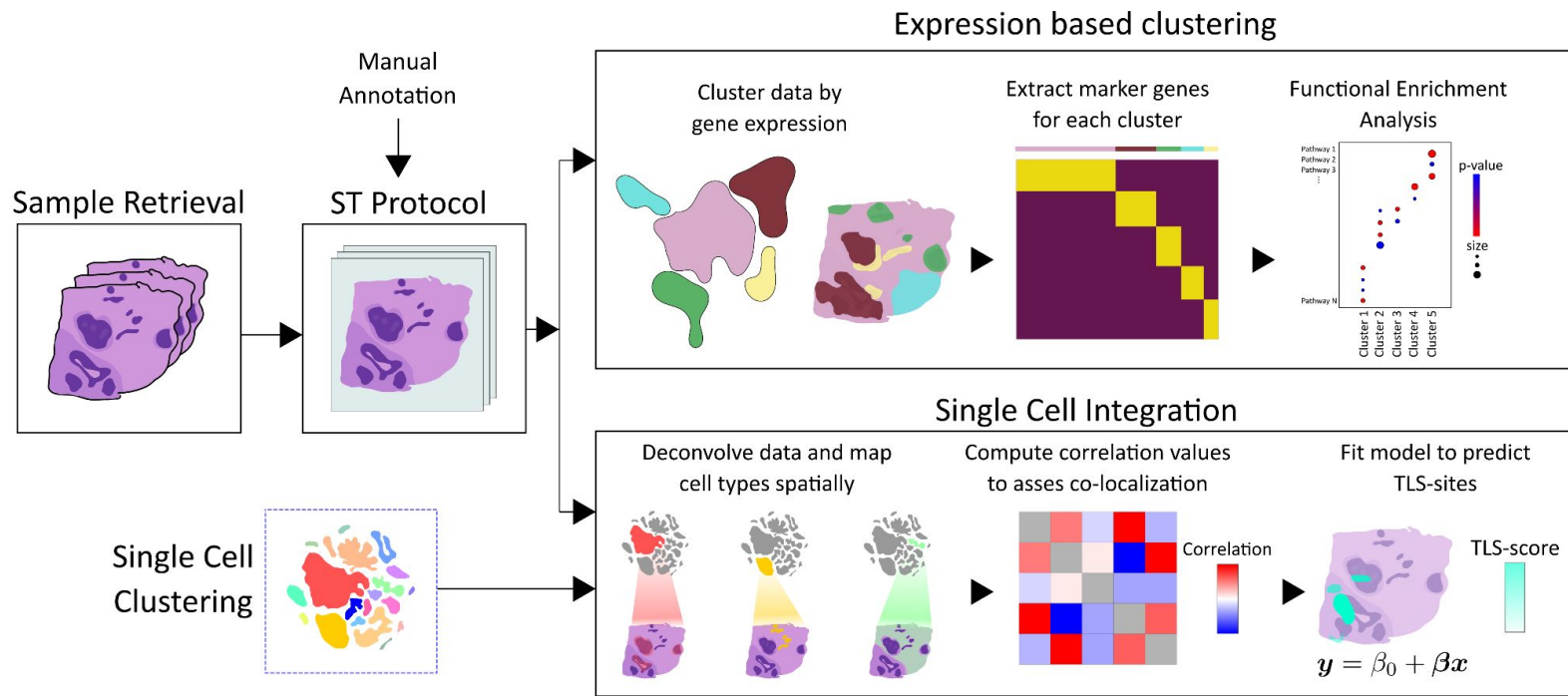
Set of
TLS-related genes:
CXCL13
LTB
CXCR5
...



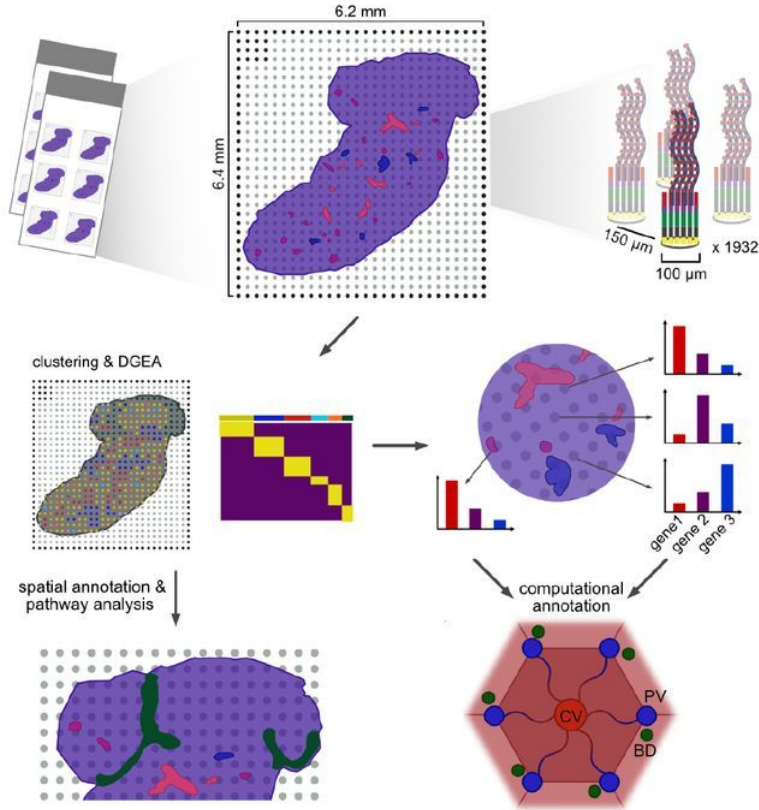
Functional
Enrichment



A spatial survey of HER2-positive breast cancer

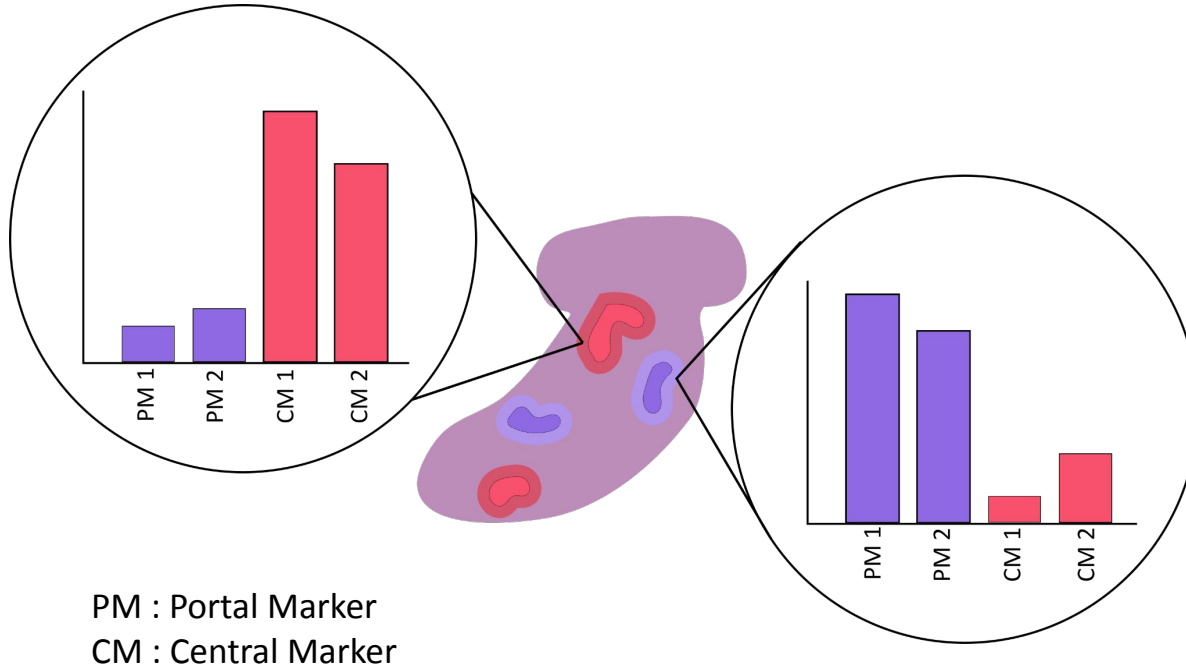


“Spatial deconvolution of HER2-positive Breast cancer delineates tumor-associated cell type interactions”, Andersson et al. (in press Nature Communications)

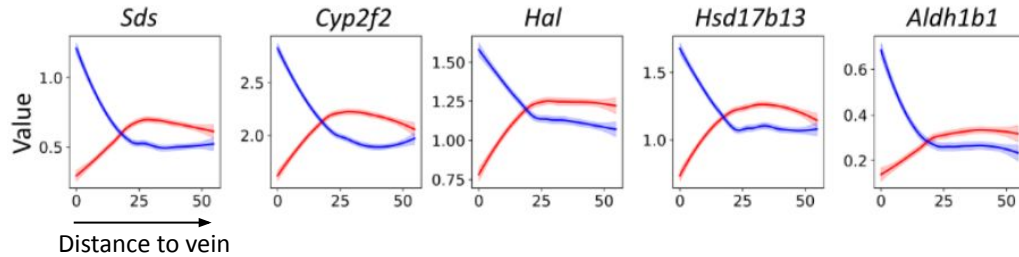


“Spatial Transcriptomics to define transcriptional patterns of zonation and structural components in the liver”, Hildebrandt and Andersson et al.

bioRxiv <https://doi.org/10.1101/2021.01.11.426100>



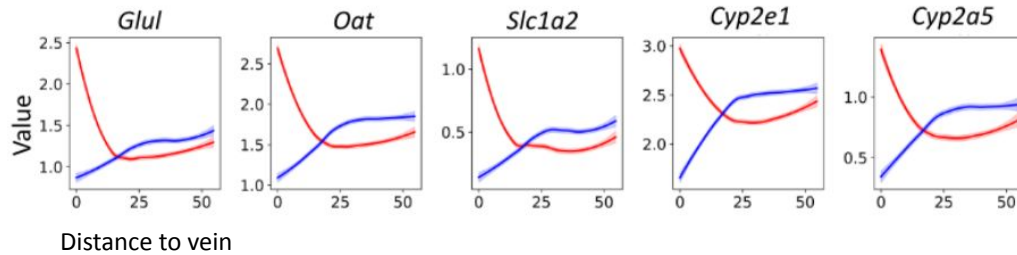
- **Portal** and **central** veins have certain marker genes associated with them
- **Key concept** : Marker gene expression is dependent on distance to the veins



- Model gene expression as a **function of the distance** to respective vein

Blue curves : expression as a function of distance to portal veins

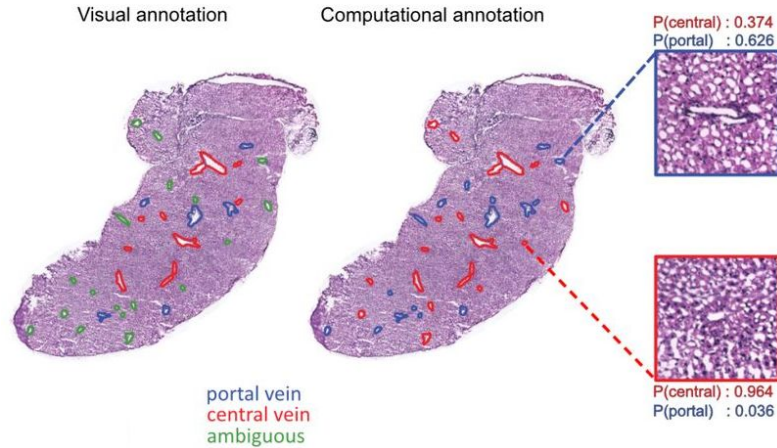
Red curves : expression as a function of distance to central veins



- Model gene expression as a ***function of the distance*** to respective vein

Blue curves : expression as a function of distance to portal veins

Red curves : expression as a function of distance to central veins



- **Objective** : Unsupervised classification of vein types
- **Implementation** :
 - Construct neighborhood expression profile (NEP).
 - Train logistic classifier on NEPs from (expert annotated) veins
 - Predict vein type based on NEP for un-annotated veins. Gives probabilistic assignment.

- Tons of spatial techniques
 - I'm very fond of Visium, but you should always pick whatever is best for you!
- Ever increasing repertoire of computational methods!
 - Be **careful** when transferring single cell methods, make sure the methods' assumptions are valid
 - A lot of tools out there, but sometimes beneficial to construct **custom solutions**
- **Don't just treat spatial data as a different form of single data**, it has much more to offer



Spatial research group : www.spatialresearch.org

Organizers : NBIS and SIB

External slide contributions : Tommaso Biancalani and Dylan Cable

Thank **you** for the attention!



<https://github.com/almaan>

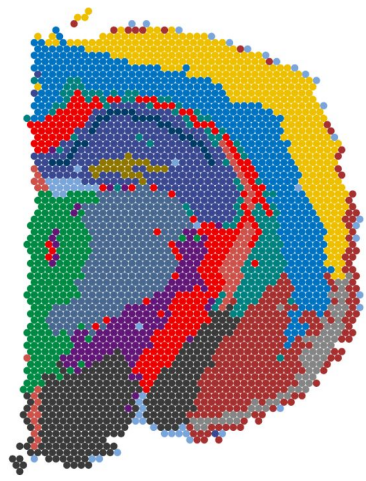


<https://almaan.github.io>

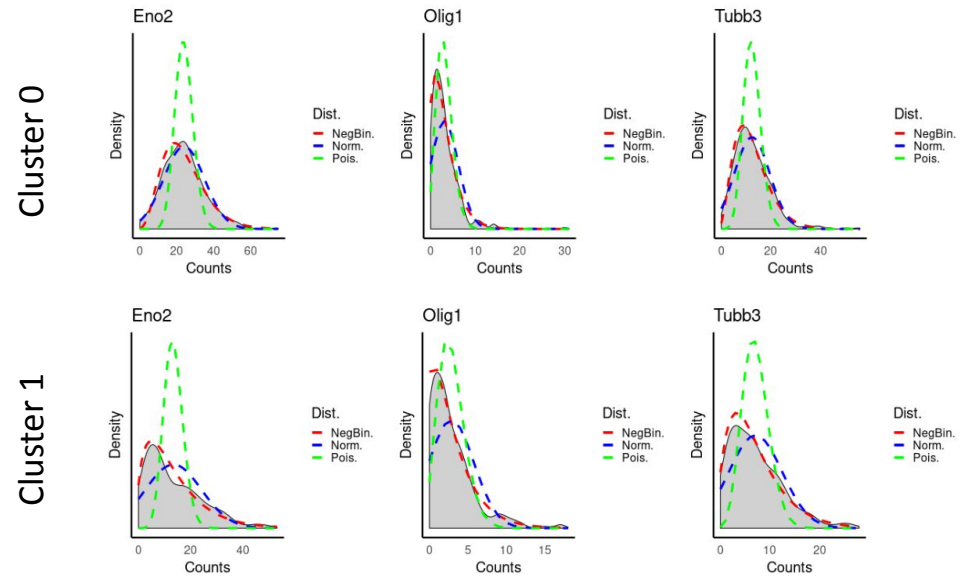


@aalmaander

- 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

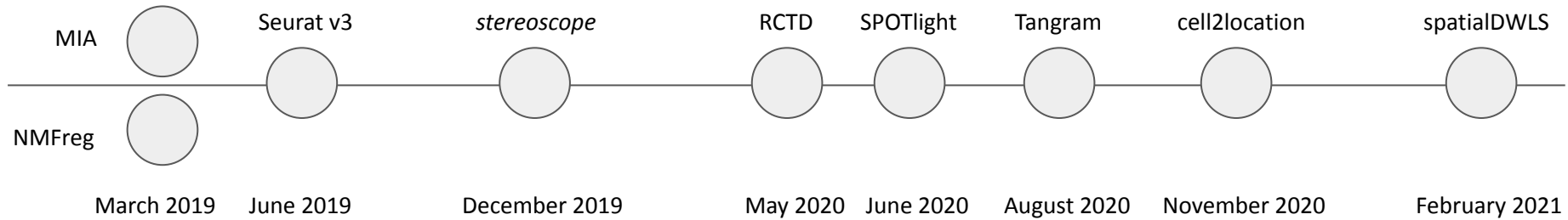


Visium data Mouse Brain
Colored by cluster



Similar trends for all clusters and genes.
Supports NB distribution, also when corrected for increased parameter number compared to Poisson).

A non-comprehensive overview based on bioRxiv releases



A non-comprehensive overview based on bioRxiv releases

