eggplant - building common coordinate frameworks for spatial transcriptomics data

Alma Andersson

SCOG Workshop | May 24th 2022

 @aalmaander

 alma.andersson@differentiable.net

 differentiable.net

Agenda

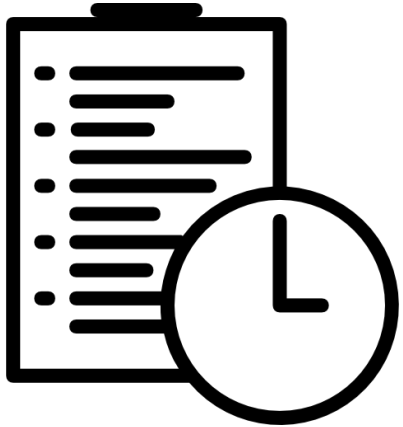
⇒ **Introducing the problem** | explaining why we need a method like this

⇒ **Presenting a solution** | outlining how our approach works

⇒ **Results** | 3 examples highlighting different aspects of *eggplant*

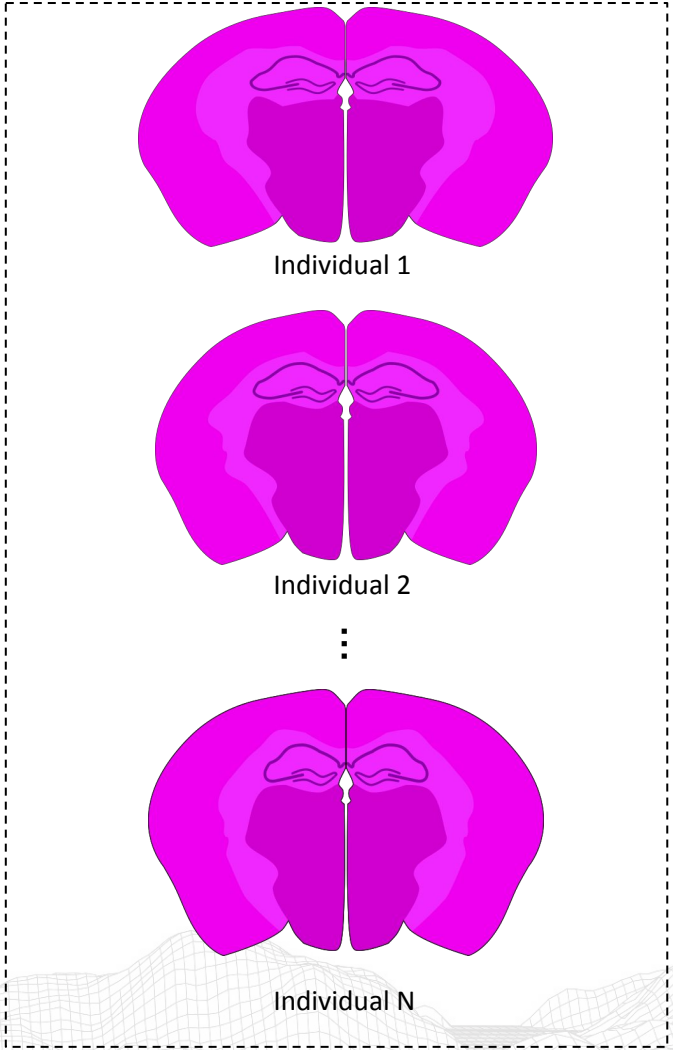
⇒ **Questions** | addressing questions relating to the presentation

⇒ **Notebook walkthrough** | getting familiar with *eggplant*



Setting the scene

- **Scenario:** we have collected tissue samples **representing the same structure** from multiple sources:
  - different conditions
  - time points
  - replicates
- **Objective:** **compare and relate** features between our samples
- **Example projects:** Atlas efforts (e.g., HCA)



The strategy

non-spatial data

Employ strategies for **batch correction** or **data integration**.

Create a **low-dimensional embedding** based on gene expression.

Relate observations in the low-dimensional **gene expression space**.

The strategy

non-spatial data

Employ strategies for **batch correction** or **data integration**.

Create a **low-dimensional embedding** based on gene expression.

Relate observations in the low-dimensional **gene expression space**.

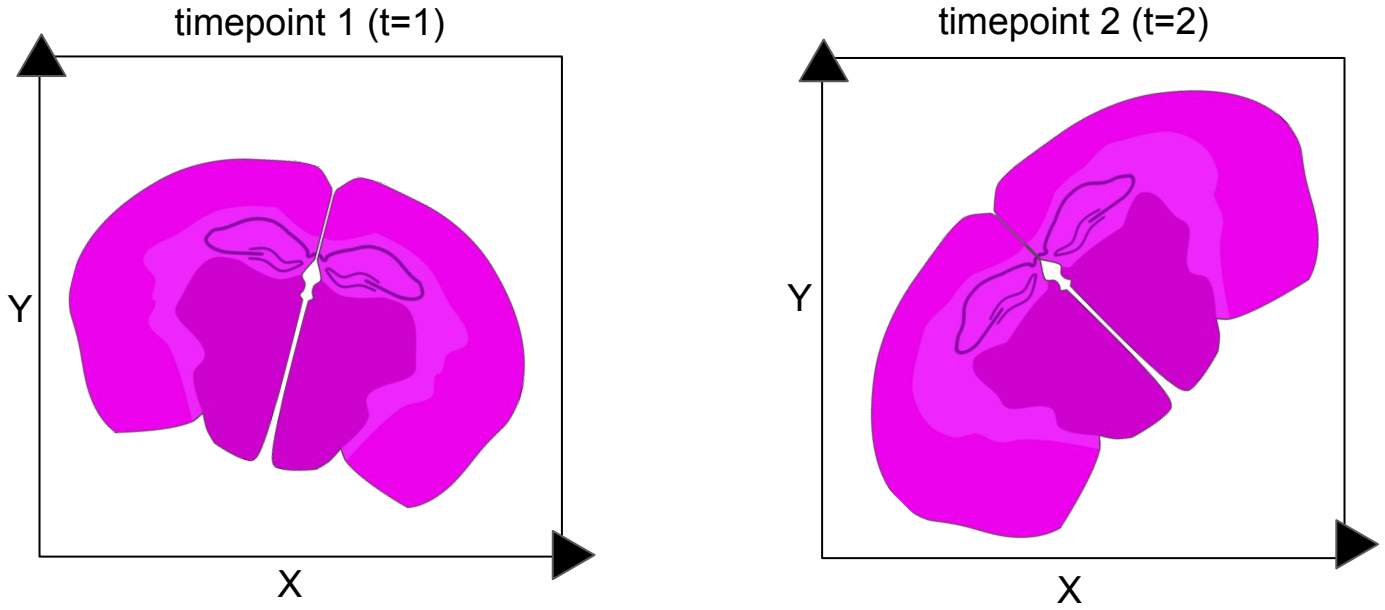
spatial data

Already possess an **inherent low-dimensional representation**: the spatial domain.

Extend idea of integration beyond removal of batch effects, to also **relate observations in the physical domain**.

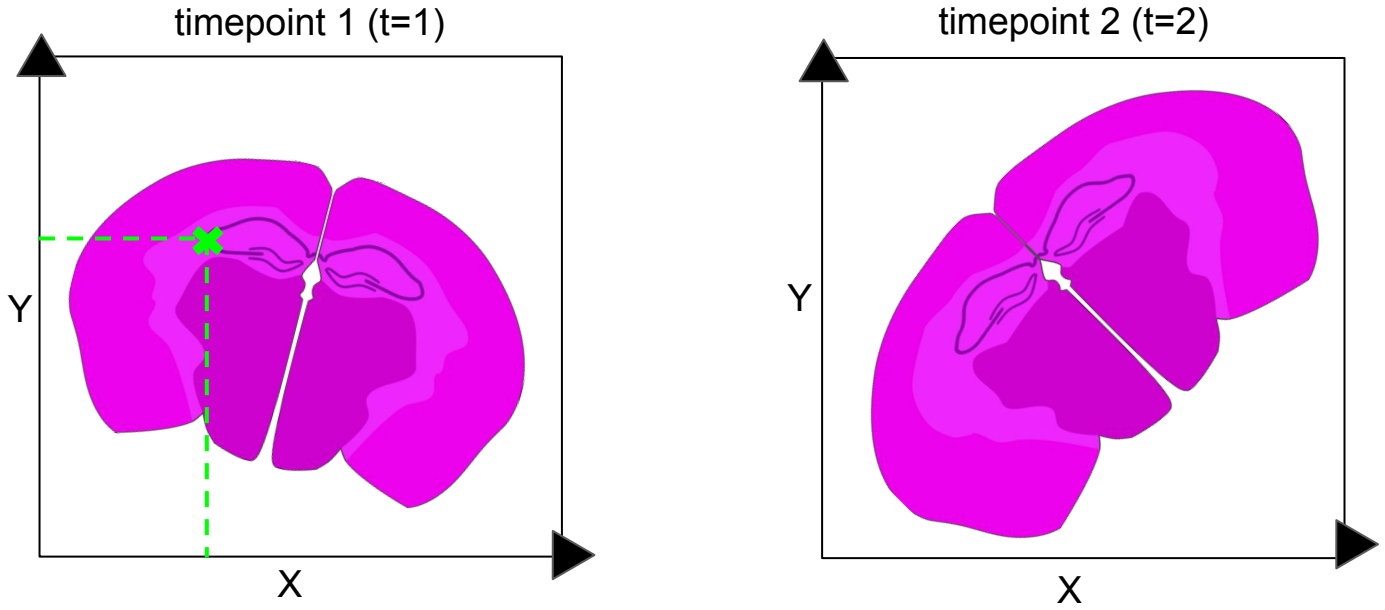
Certain **challenges** to this, biggest one being the issue of **independent coordinate systems**

The challenge :: independent coordinate systems



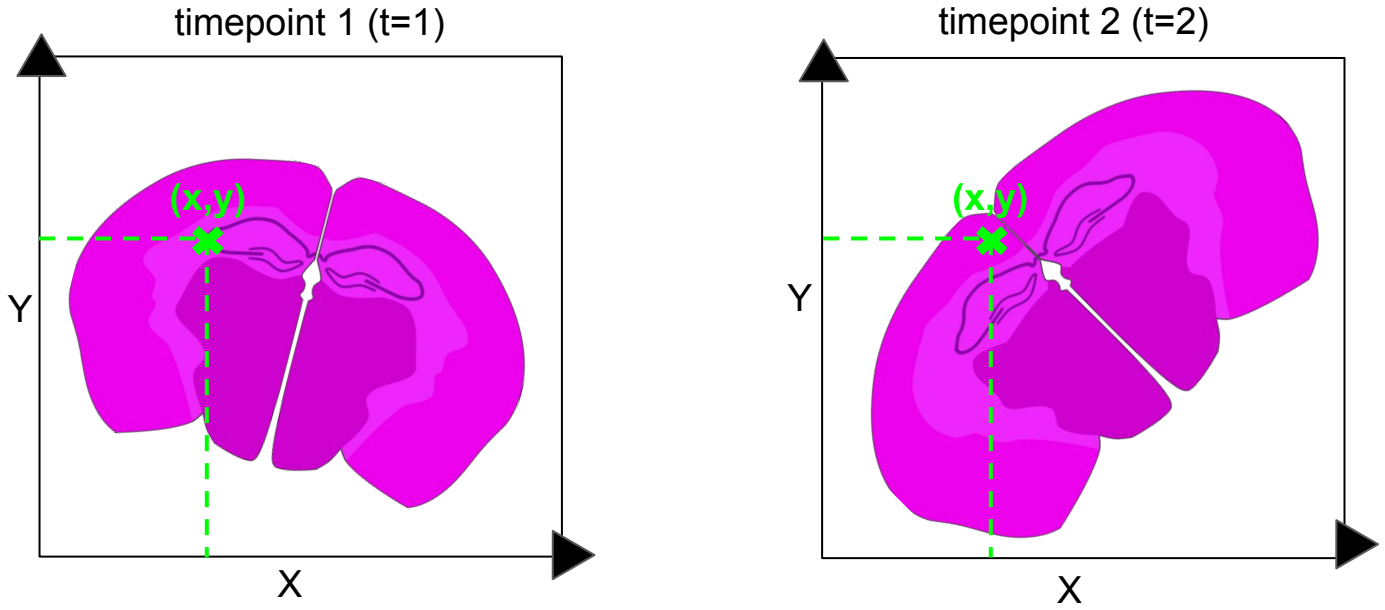
Example: we have two tissue samples from different timepoints, t=1 and t=2

The challenge :: independent coordinate systems



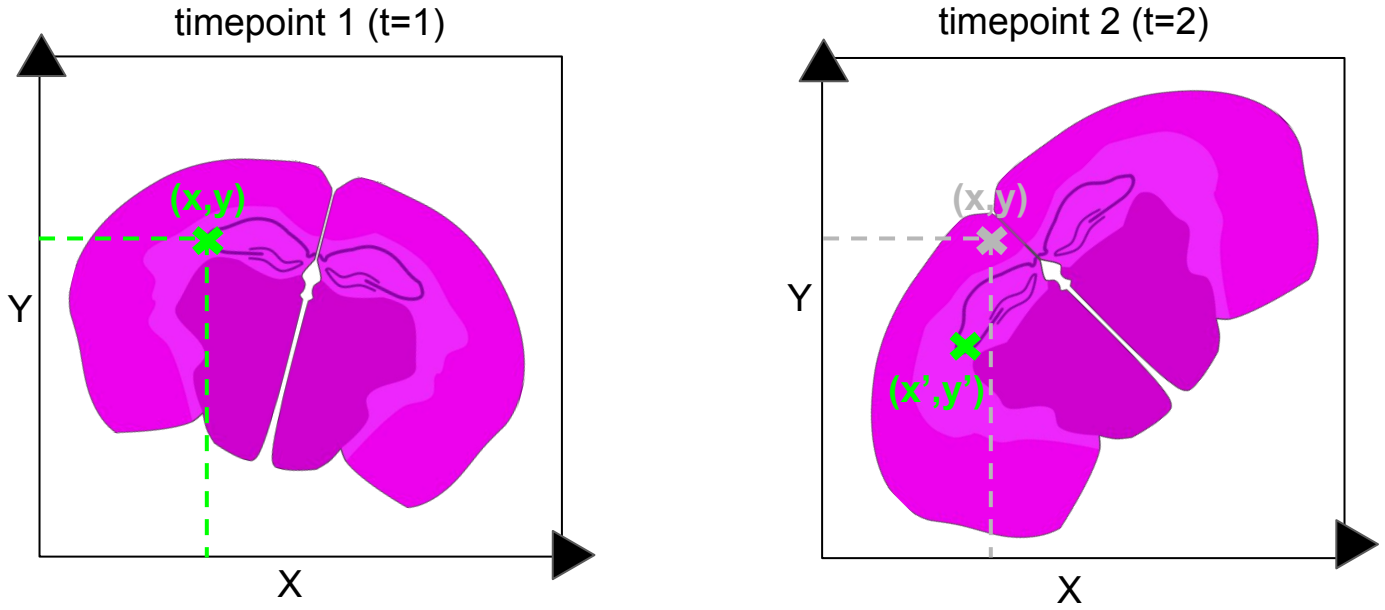
Objective: examine how the gene expression at a certain **position** changes over time

The challenge :: independent coordinate systems



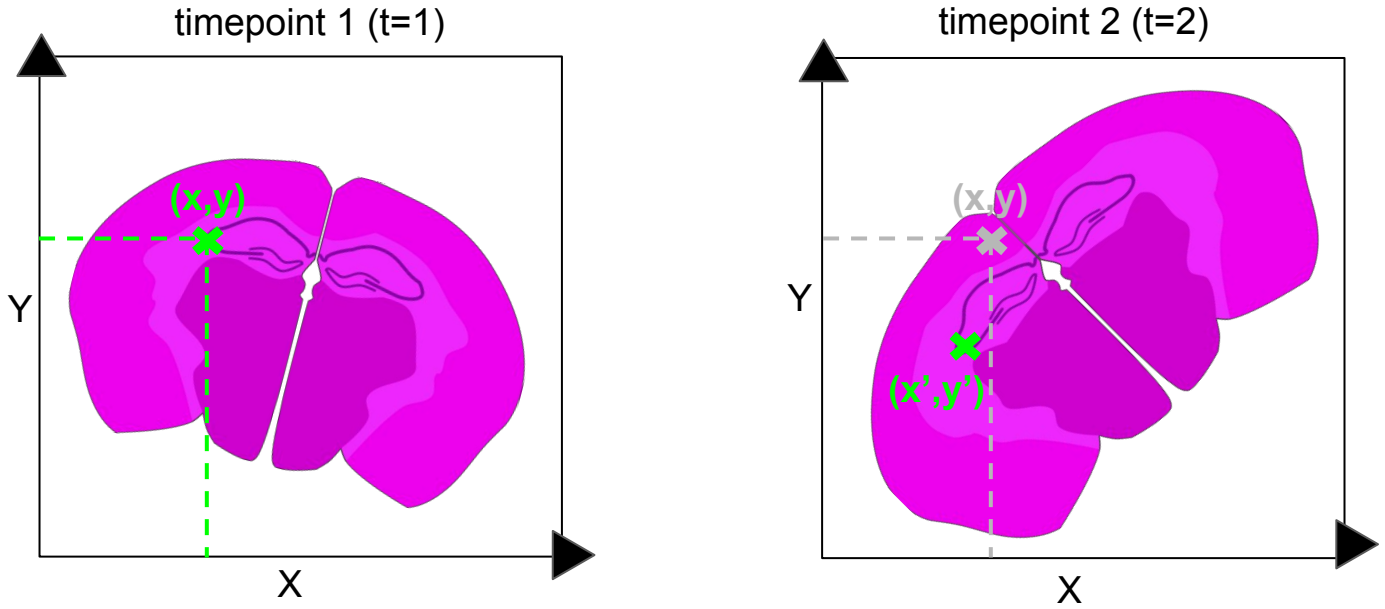
Issue: coordinates **(x,y)** do not represent the same position in t=1 and t=2; independent coordinate systems

The challenge :: independent coordinate systems



Issue: coordinates **(x,y)** do not represent the same position in t=1 and t=2; independent coordinate systems
Correct approach: compare expression at coordinates **(x,y)** at t=1 with expression at **(x',y')** at t=2

The challenge :: independent coordinate systems

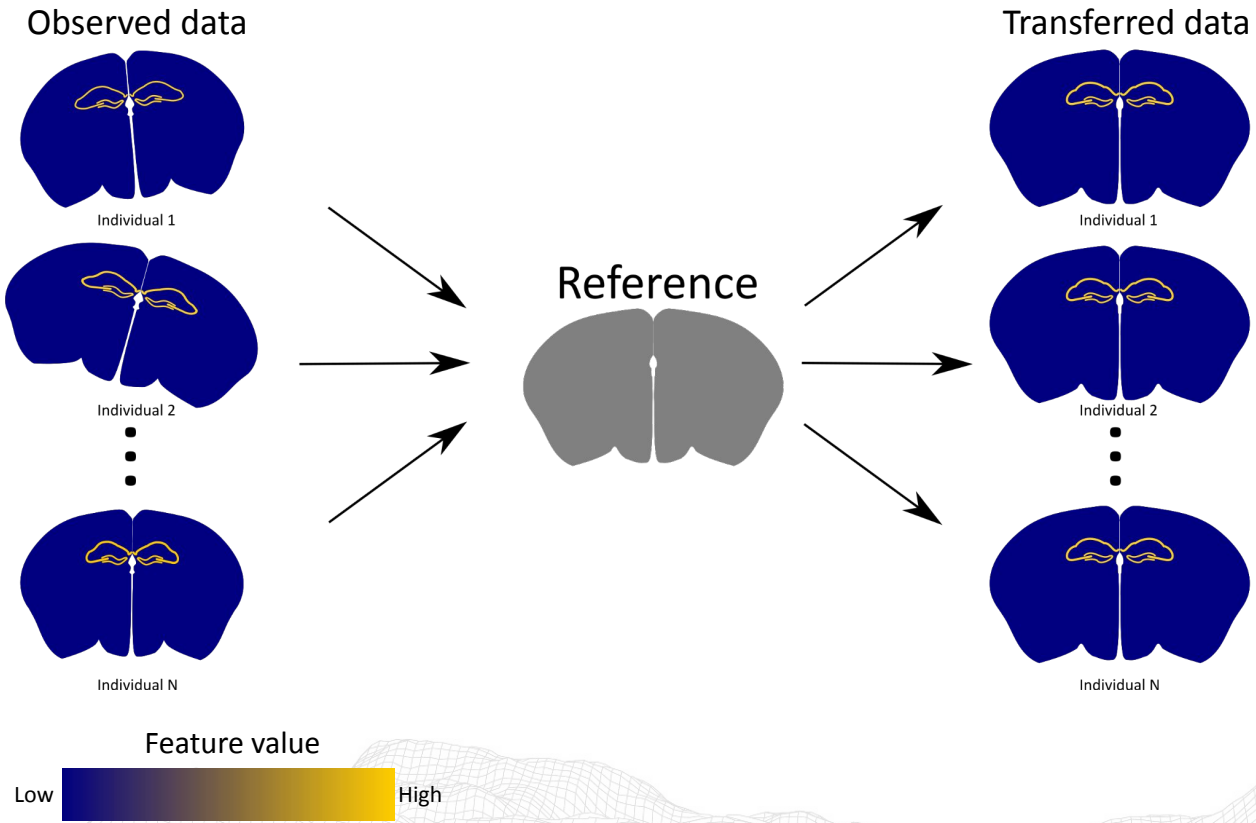


Issue: coordinates **(x,y)** do not represent the same position in t=1 and t=2; independent coordinate systems

Correct approach: compare expression at coordinates **(x,y)** at t=1 with expression at **(x',y')** at t=2

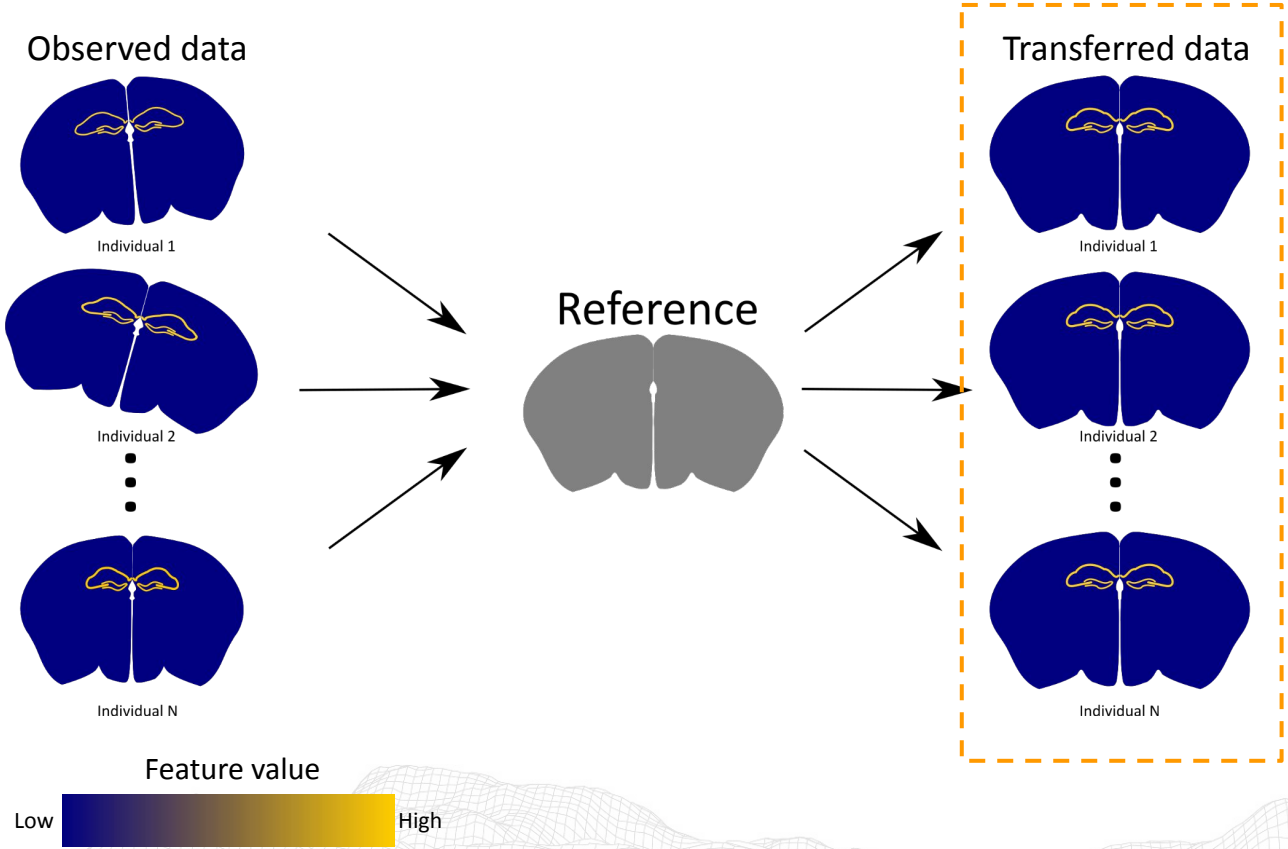
Solution: transfer the expression to a **common coordinate framework** where **(x,y)** represents same position in all samples

Common Coordinate Frameworks (CCFs)



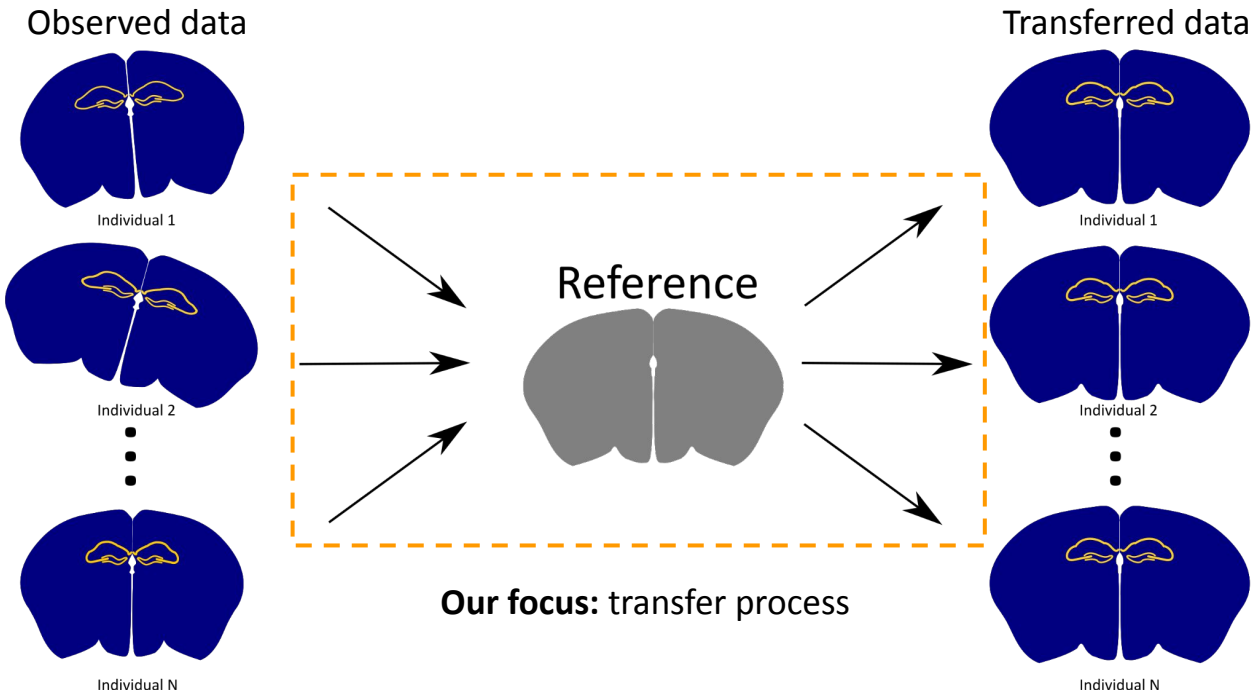
- Transfer information from observed data to reference/CCF
- Information “inhabits” the same spatial domain.
- **More than an alignment**
- **Important:** corrects for non-linear distortions.

Common Coordinate Frameworks (CCFs)



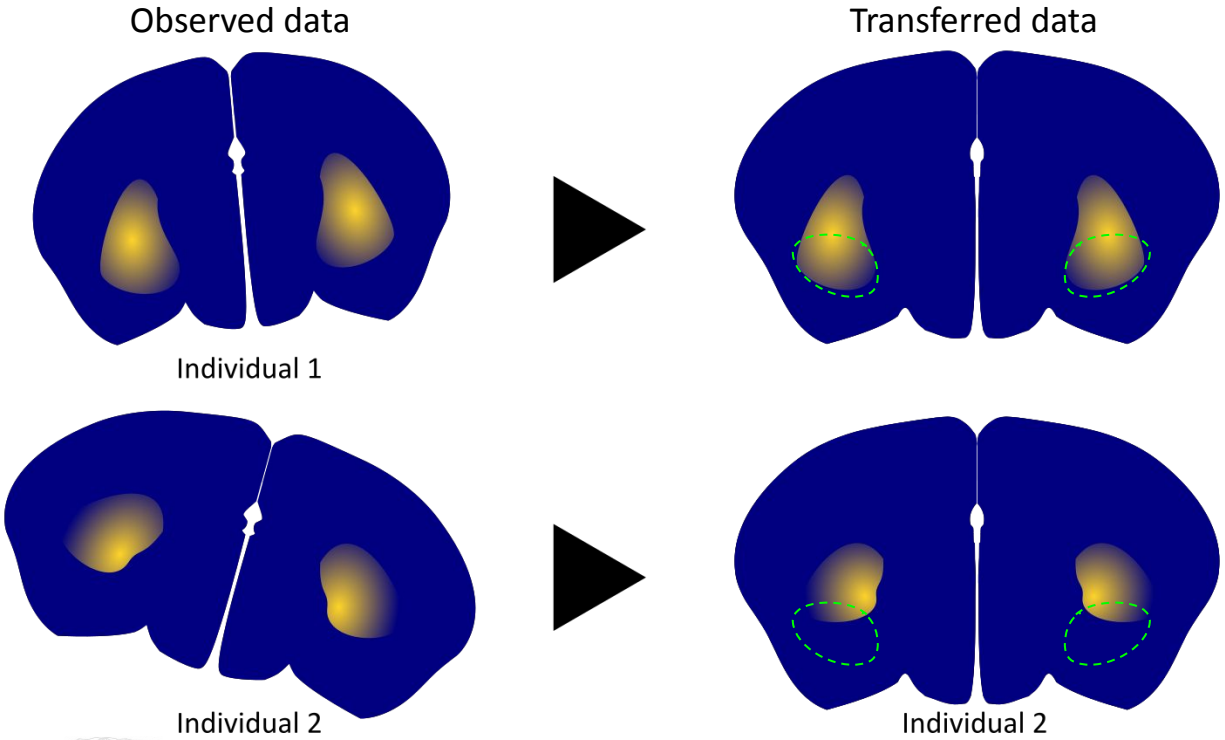
- Transfer information from observed data to reference/CCF
- Information “inhabits” the same spatial domain.
- **More than an alignment**
- **Important:** corrects for non-linear distortions.

Common Coordinate Frameworks (CCFs)



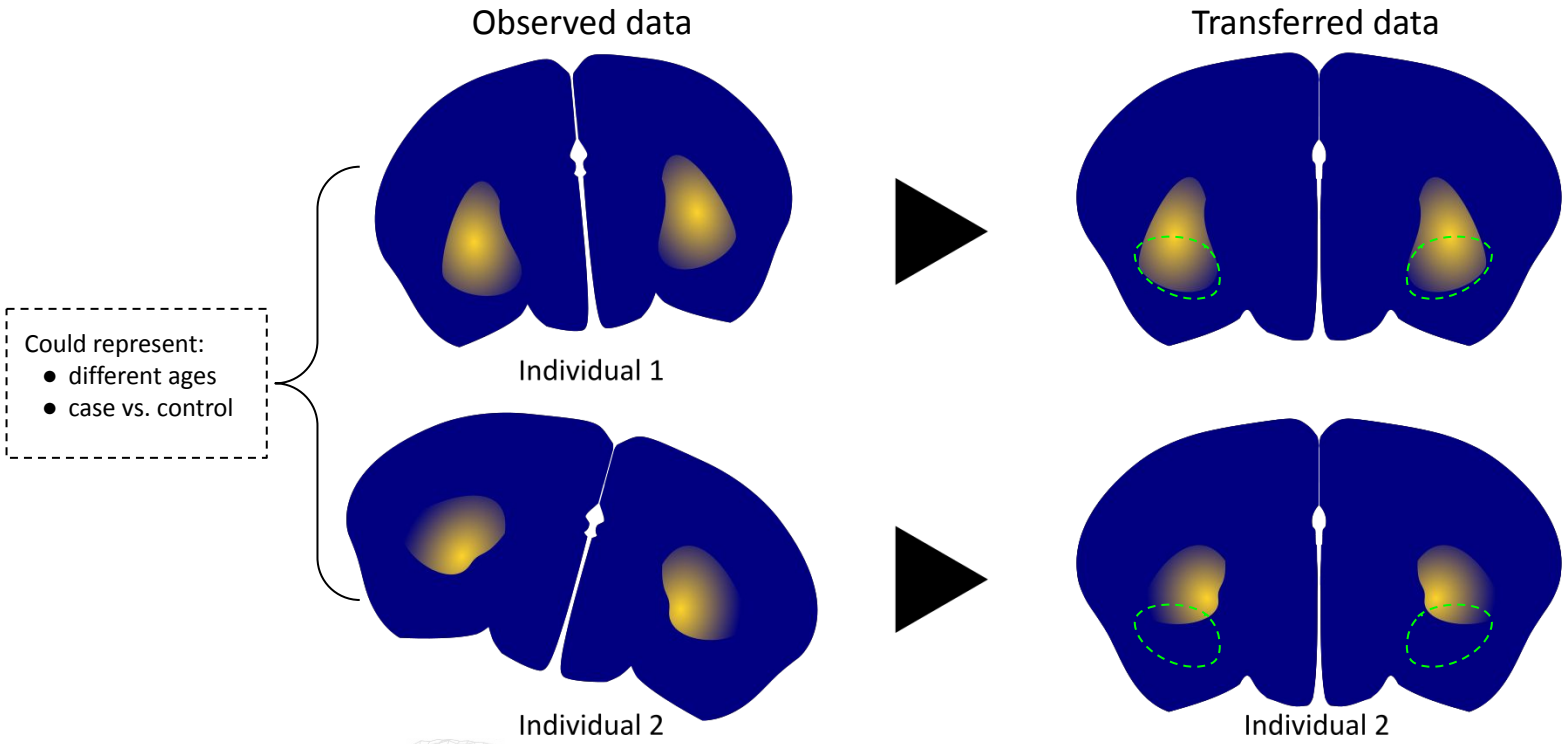
- Transfer information from observed data to reference/CCF
- Information “inhabits” the same spatial domain.
- **More than an alignment**
- **Important:** corrects for non-linear distortions.

Common Coordinate Frameworks (CCFs)



We can **compare** expression **location-by-location** between different samples
Differences can be assessed both visually and computationally!

Common Coordinate Frameworks (CCFs)



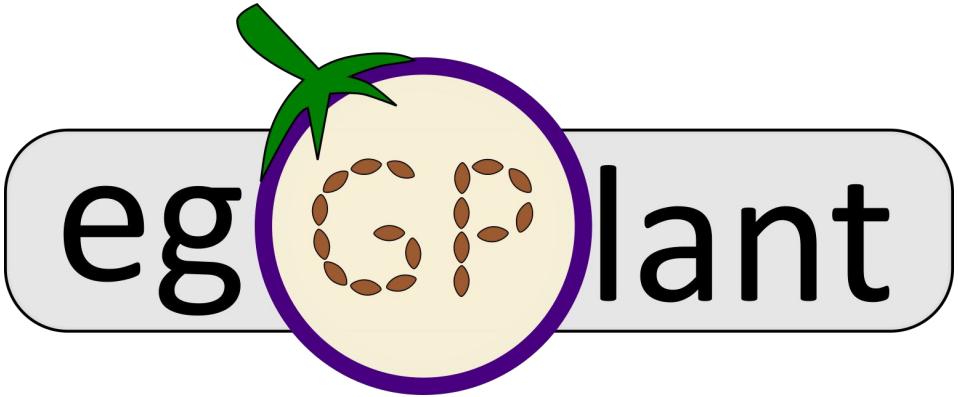
We can **compare** expression **location-by-location** between different samples
Differences can be assessed both visually and computationally!

So what's the purpose?

- Transfer of information to a reference/CCF is **usually not the end goal**.
- **Enables** plenty of downstream analyses:
  - ***Spatiotemporal modeling**:*
    - Analyze and characterize how the expression changes over **time and space**
  - ***Local expression changes***:
    - Assess whether expression is up-or downregulated in certain regions
  - ***Aggregation of data**:*
    - Jointly represent information from multiple samples by **composite representations**
    - Identify **general** spatial expression patterns found in the data
- etc ...

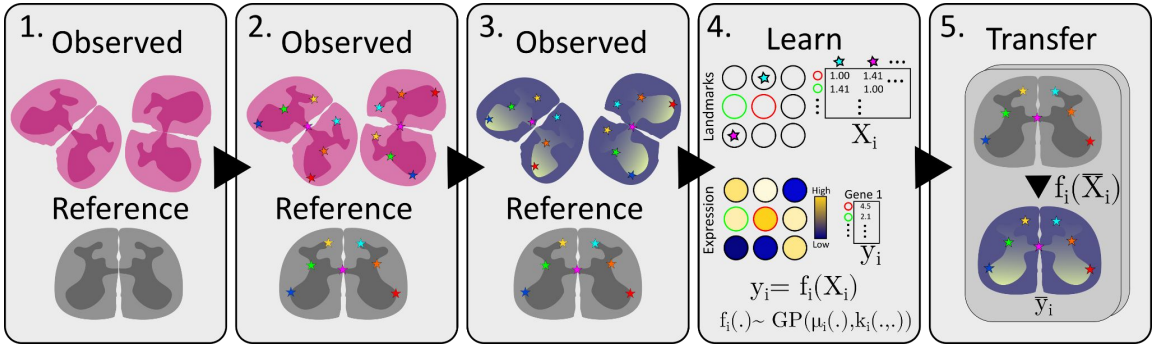
The solution?

~~The~~ Our solution?



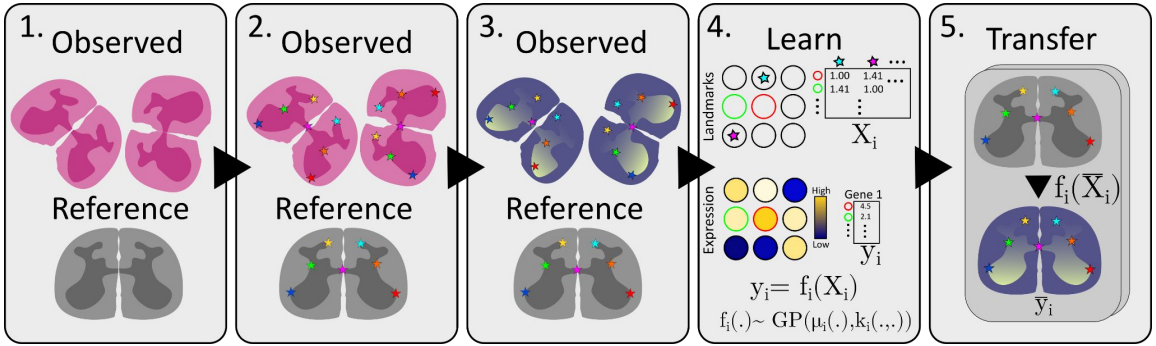
effortless **g**eneric **GP** **l**andmark **t**ransfer

eggplant :: Method overview



1. Construct/choose a reference and select samples to transfer information from
2. *Chart* the **landmarks**, i.e., annotate common landmarks
3. Select feature of interest, e.g., expression of your favorite gene
4. Learn a **transfer function** relating landmark distances to expression
5. **Transfer** information to the reference, **apply transfer function** to each location in the reference

eggplant :: Method overview

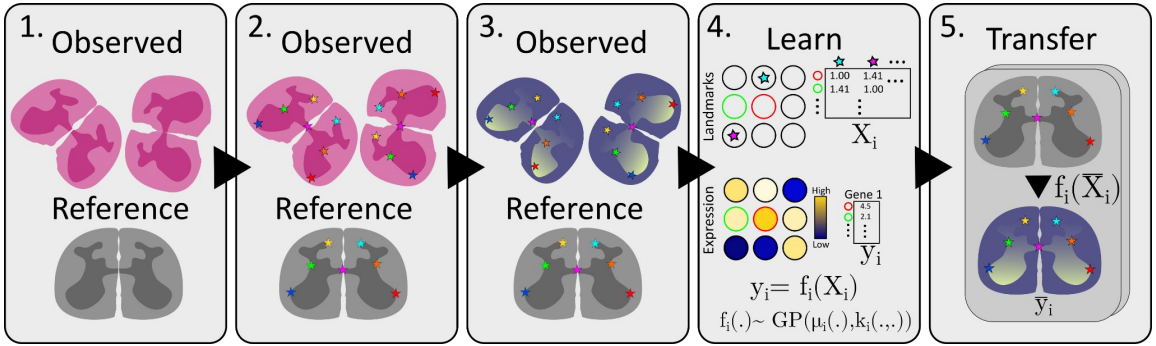


1. Construct/choose a reference and select samples to transfer information from
2. *Chart* the **landmarks**, i.e., annotate common landmarks
3. Select feature of interest, e.g., expression of your favorite gene
4. Learn a **transfer function** relating landmark distances to expression

Will explain in more detail

5. **Transfer** information to the reference, **apply transfer function** to each location in the reference

eggplant :: Method overview



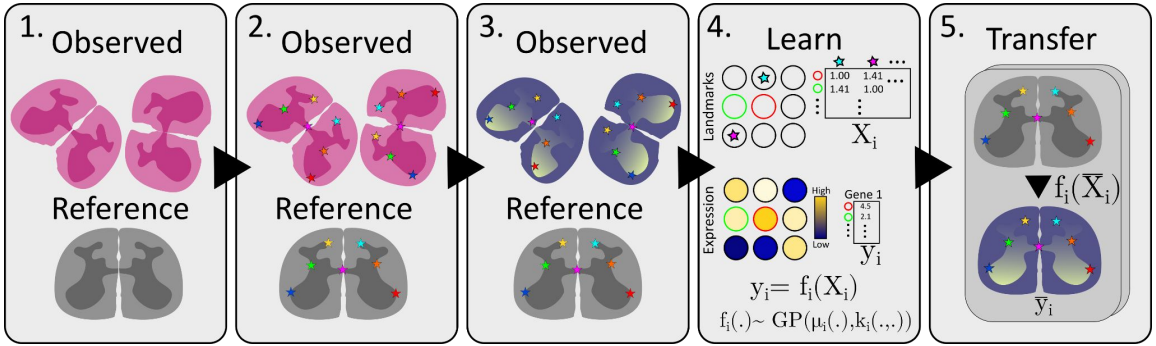
1. Construct/choose a reference and select samples to transfer information from
2. *Chart* the **landmarks**, i.e., annotate common landmarks
3. Select feature of interest, e.g., expression of your favorite gene
4. Learn a **transfer function** relating landmark distances to expression
5. **Transfer** information to the reference, **apply transfer function** to each location in the reference

What is a landmark?

- **Landmarks:** structures, signals or any features **easily identified across samples**
  - definition largely based on *"Toward a Common Coordinate Framework for the Human Body" (Rood et al.)*
- **Key:** a landmark should always represent the **same approximate position**
- Landmarks can, for example, be based on:
  - Morphology (Image)
  - Gene expression
  - Protein signals
- **Important:** *eggplant* only focuses on the **transfer process**
  - Agnostic to annotation method
  - Manual landmark annotation so far, automatic on the way

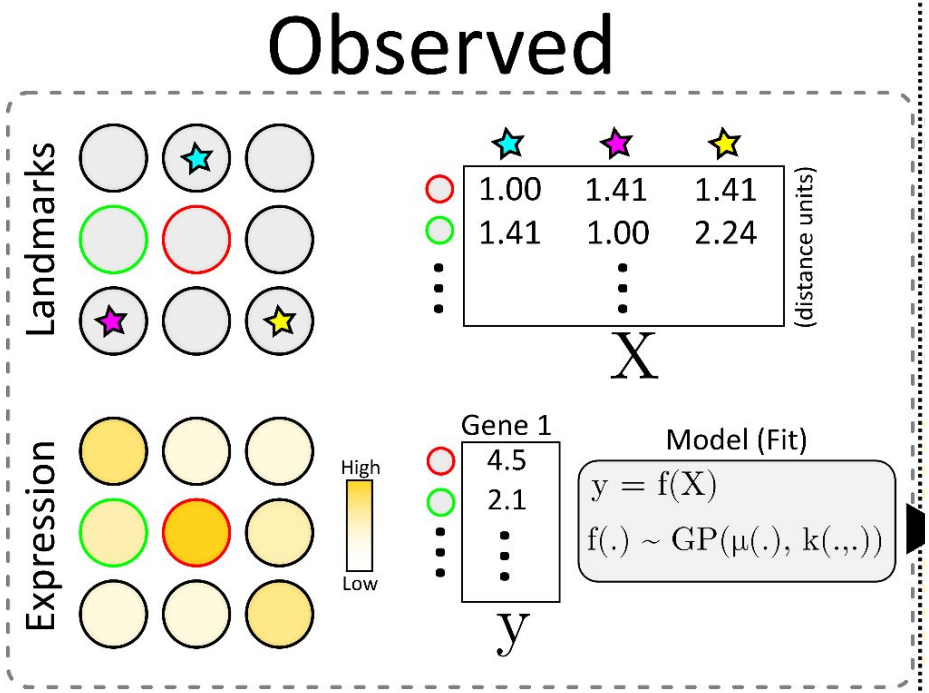
mouse brain (one hemisphere)

eggplant :: Method overview



1. Construct/choose a reference and select samples to transfer information from
2. *Chart* the **landmarks**, i.e., annotate common landmarks
3. Select feature of interest, e.g., expression of your favorite gene
4. Learn a **transfer function** relating landmark distances to expression
5. **Transfer** information to the reference, **apply transfer function** to each location in the reference

Learning the transfer function

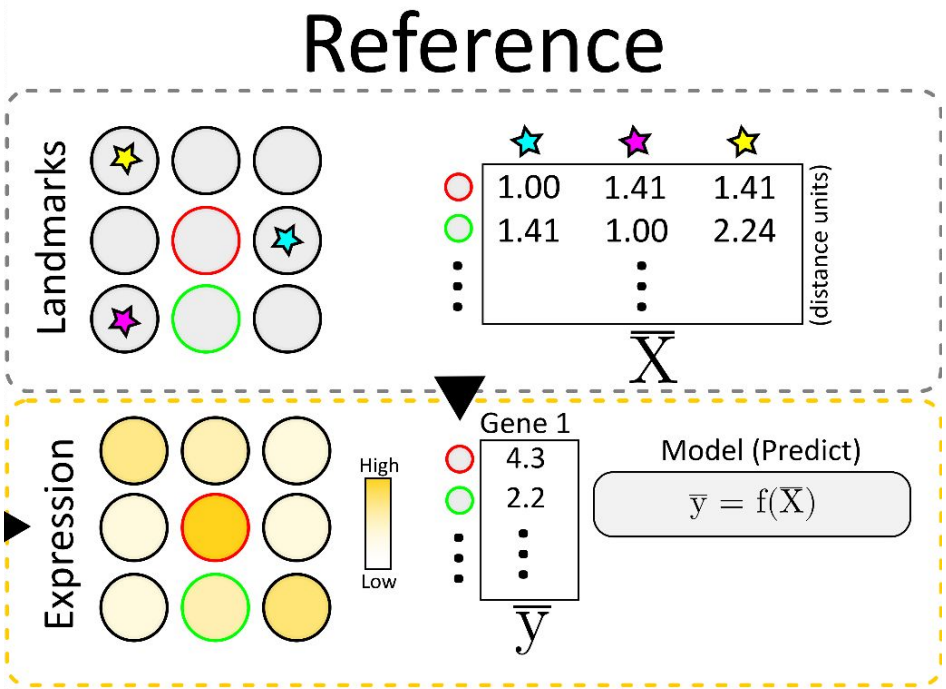


X : observed landmark distance matrix (given)
y : observed gene expression vector (given)

- For each location: measure distance to all landmarks → **distance matrix** (X)
- Assume that gene expression (y) is **related** to landmark distances **via a function**: y = *f*(X)
- Assumes that the transfer function (*f*) follows a **Gaussian Process** (GP) distribution
- Use **Gaussian Process Regression** to determine character of the transfer function

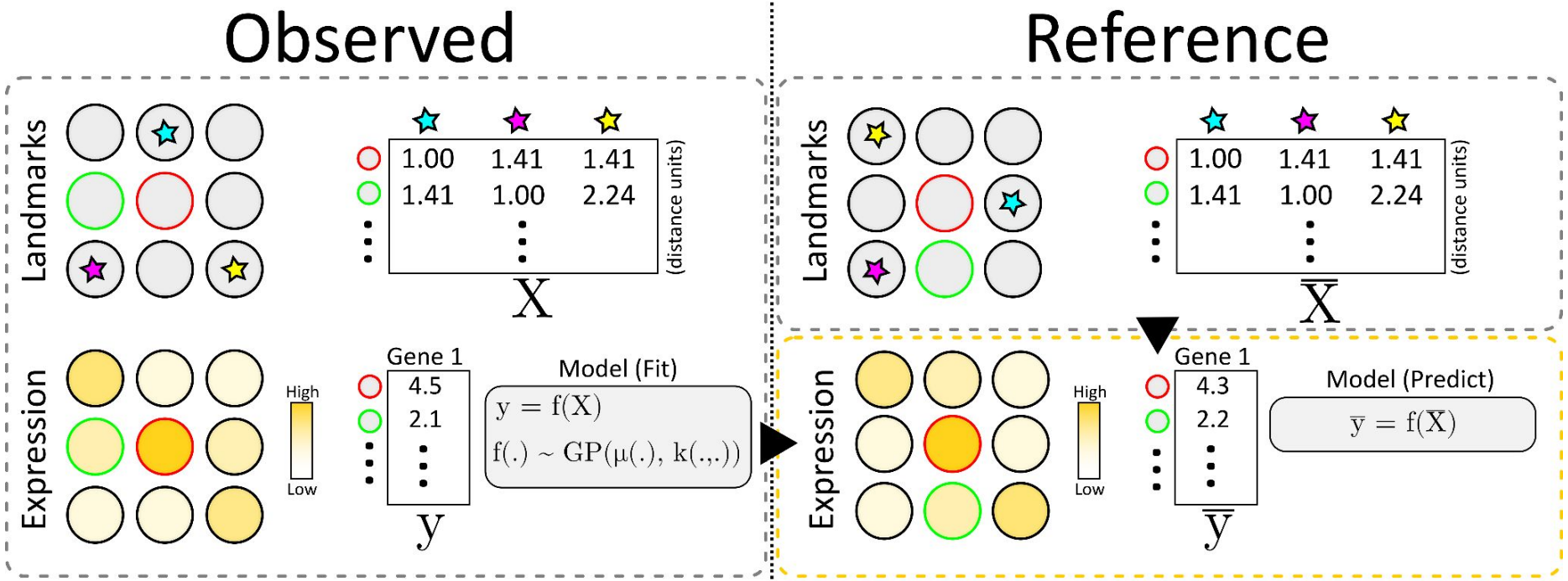
Transferring information to the reference

- Create **distance matrix** ($\bar{X}$) for the reference
- **Transfer** information by **applying** transfer function to $\bar{X}$ (prediction)



$\bar{X}$: reference landmark distance matrix (given)
$\bar{y}$: **predicted** gene expression vector

Transferring information to the reference

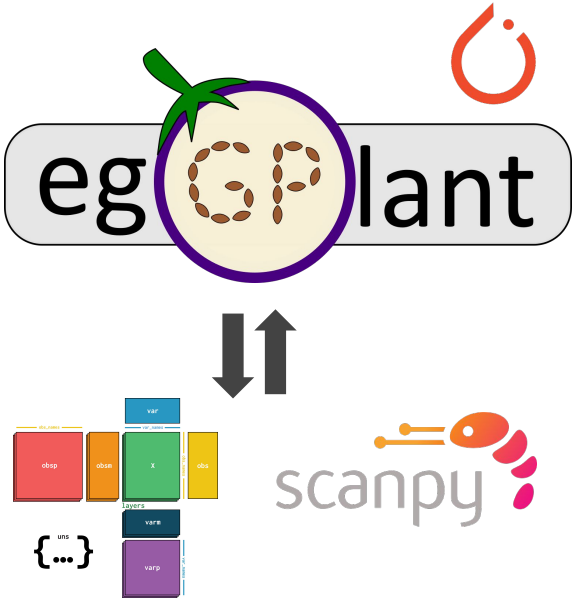


X : observed landmark distance matrix (given)
y : observed gene expression vector (given)

X̄ : reference landmark distance matrix (given)
ȳ : **predicted** gene expression vector

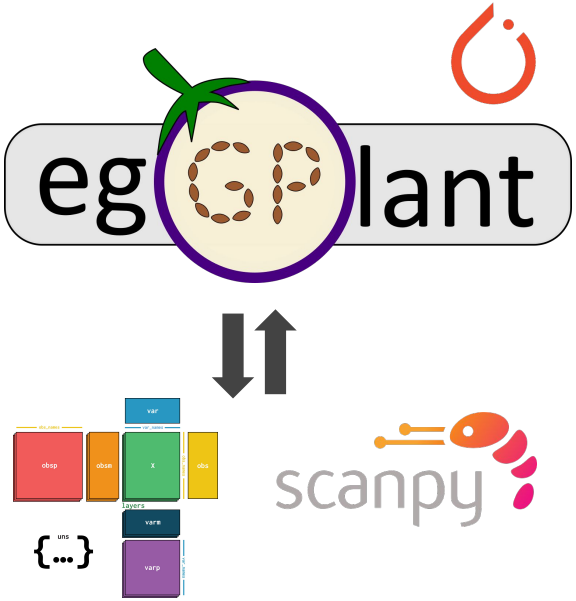
Implementation

- **Python package:**
  - written to be compatible with AnnData objects
  - scanpy-like API
  - can be installed via PyPi and GitHub (recommended)
  - documented on: Read the Docs
- **Backend:** PyTorch/GPyTorch



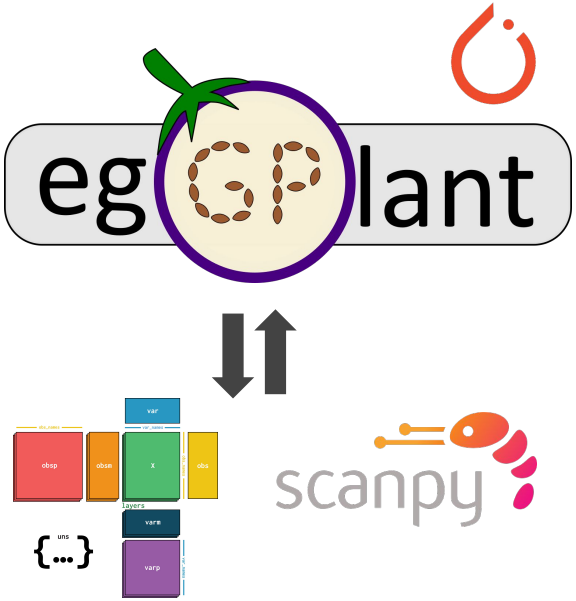
Implementation

- **Python package:**
  - written to be compatible with AnnData objects
  - scanpy-like API
  - can be installed via PyPi and GitHub (recommended)
  - documented on: Read the Docs
- **Backend:** PyTorch/GPyTorch
- **Challenge:**
  - GP's have notoriously bad scaling: O(n^3) where n = number of observations
  - We have plenty of observations, and many features.



Implementation

- **Python package:**
  - written to be compatible with AnnData objects
  - scanpy-like API
  - can be installed via PyPi and GitHub (recommended)
  - documented on: Read the Docs
- **Backend:** PyTorch/GPyTorch
- **Challenge:**
  - GP's have notoriously bad scaling: O(n^3) where n = number of observations
  - We have plenty of observations, and many features.



Implementation : the challenge

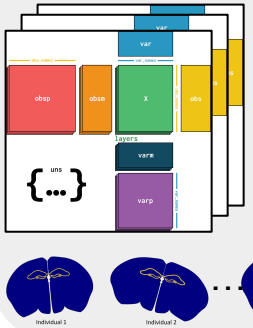
Strategies for scaling the method to large omics data sets:

- **GPU acceleration:**
  - General acceleration of the inference process
  - Handled by PyTorch/GPyTorch
- **Variational Inference:**
  - Uses inducing points
  - reduces complexity w.r.t. Observations : $O(n^3) \rightarrow O(m^2 n)$ (m = #inducing points)
- **Fast approximate transfer:**
  - Uses a data decomposition step
  - Reduces the number of features that has to be transferred
  - Steps:
    1. Use PCA to project full-dimensional data into a K-dimensional subspace
    2. Transfer loadings of each principal component to the reference
    3. Reconstruct full-dimensional data in the reference

Implementation : input and output

Input

anndata objects



- *.X* : features of interest (e.g., gene expression)
- *.obsm["spatial"]* : coordinates of observations (e.g., spots)
- *.uns["curated_landmarks"]* : position of landmarks
  - *.obsm["landmark_distances"]* : distances to landmarks

reference object

reference image

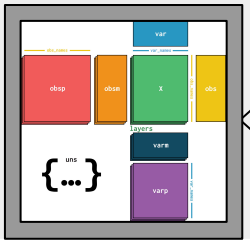
reference landmark coordinates

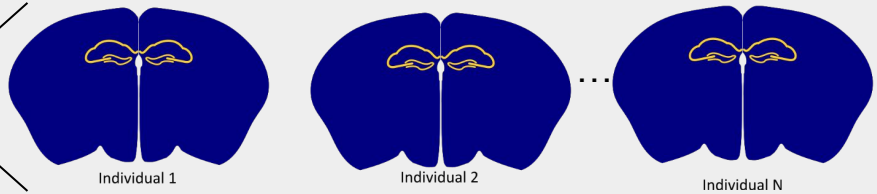
	x_coord	y_coord
Landmark_0	35.938	1160.586
Landmark_1	29.596	63.420
Landmark_2	989.352	822.346
Landmark_3	264.250	1071.798
Landmark_4	304.416	325.556

eggplant

Output

reference object
(anndata object)





Example analyses

- **Human developmental heart** (✣)
  - Basic example
  - Variance in structure between individuals
  - Comparison with alignment
- **Mouse brain** (sagittal)
  - Complex structures
- **Perturbation in mouse brain** (✣)
  - spatial differential expression analysis (SDEA)

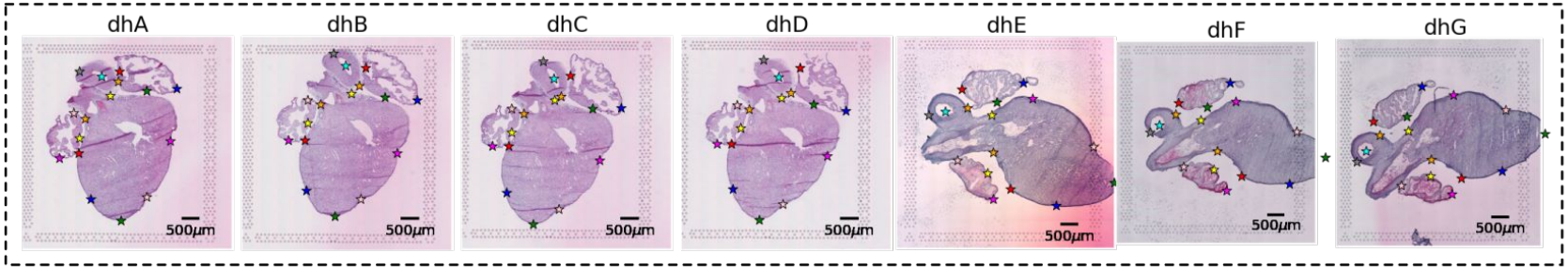
✣ := in the tutorial

Example analyses

- **Human developmental heart** (✣)
  - Basic example
  - Variance in structure between individuals
  - Comparison with alignment
- **Mouse brain** (sagittal)
  - Complex structures
- **Perturbation in mouse brain** (✣)
  - spatial differential expression analysis (SDEA)

✣ := in the tutorial

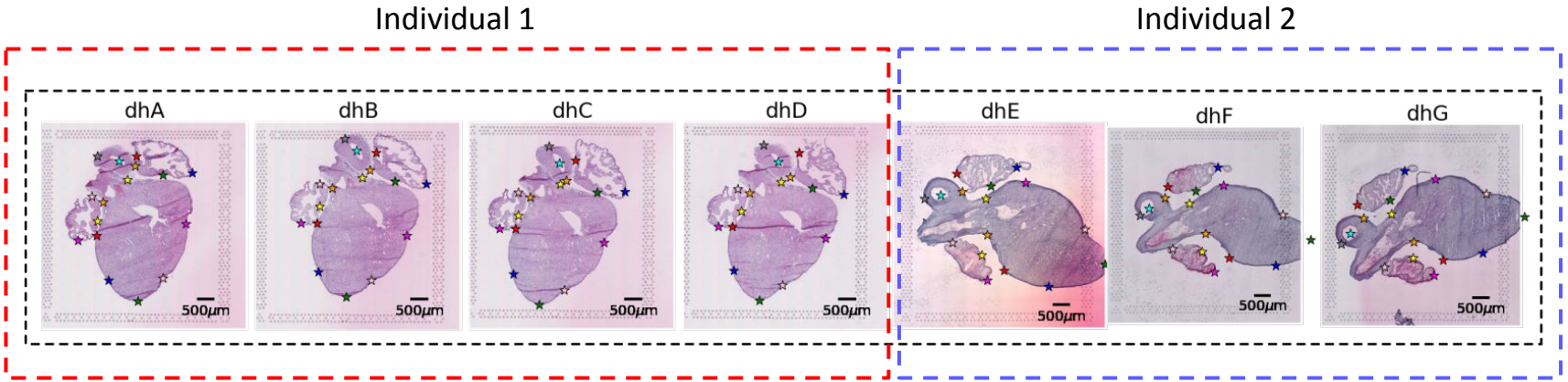
eggplant :: developmental heart analysis



= Landmark

Data from: Žaneta Andrusivová, Xiaofei Li and Erik Sundström

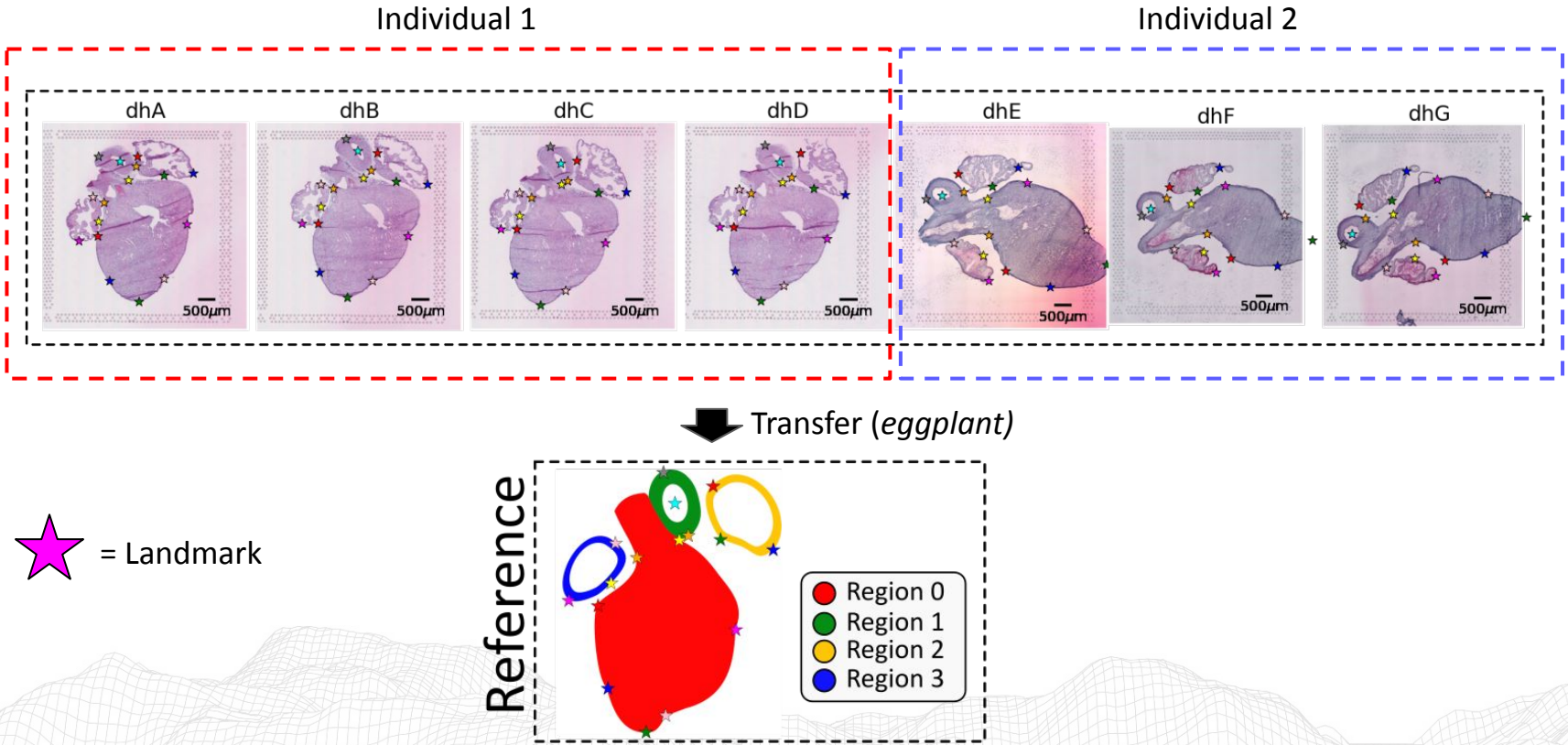
eggplant :: developmental heart analysis



= Landmark

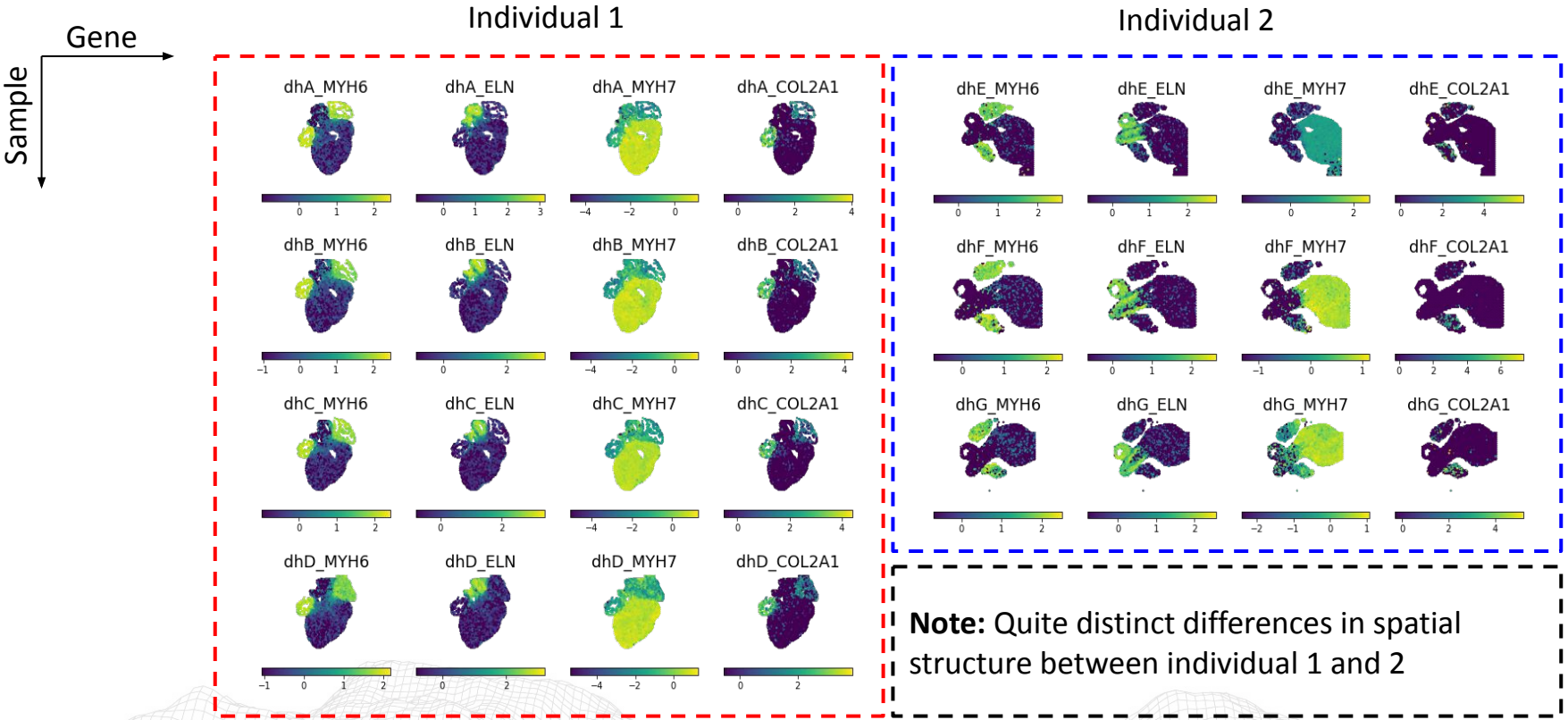
Data from: Žaneta Andrusivová, Xiaofei Li and Erik Sundström

eggplant :: developmental heart analysis



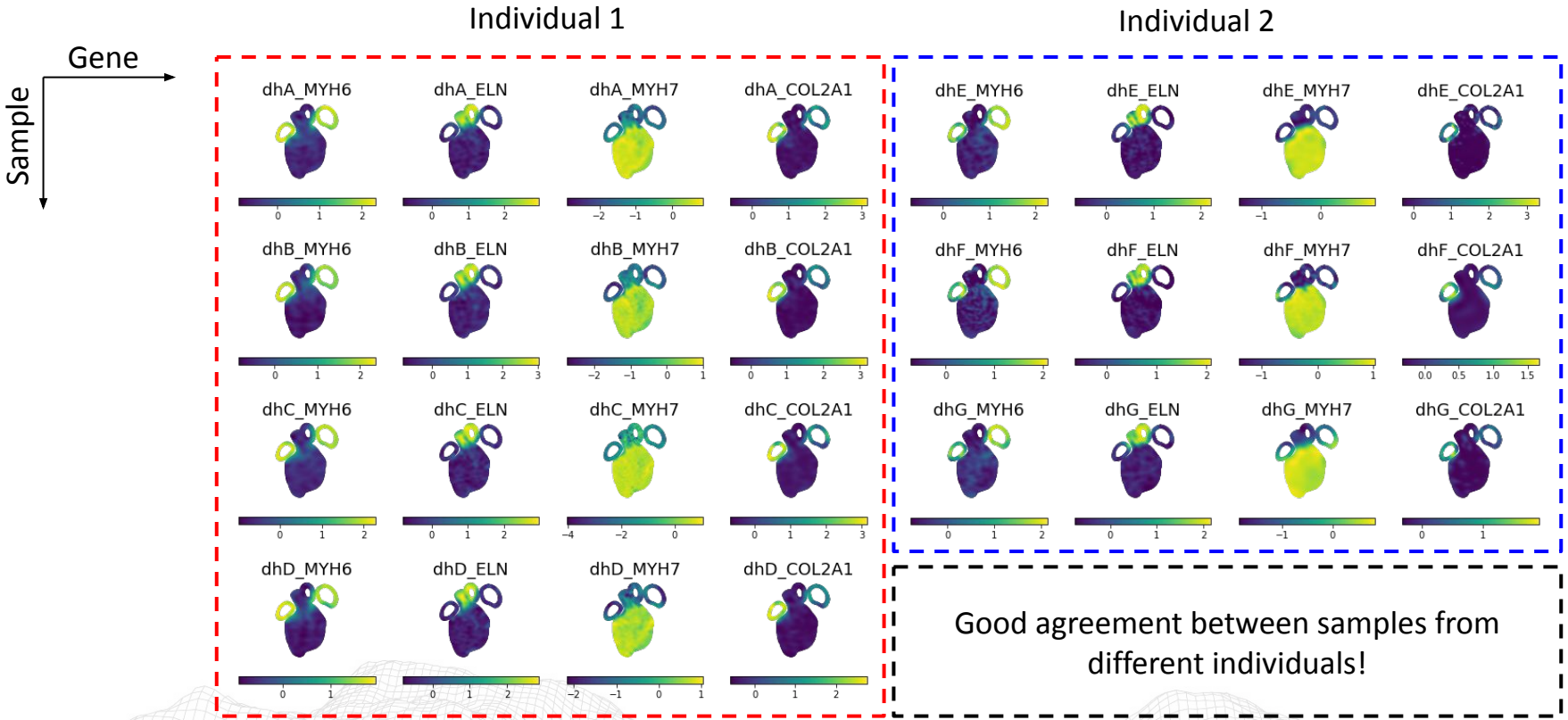
Data from: Žaneta Andrusivová, Xiaofei Li and Erik Sundström

Observed expression (developmental heart)



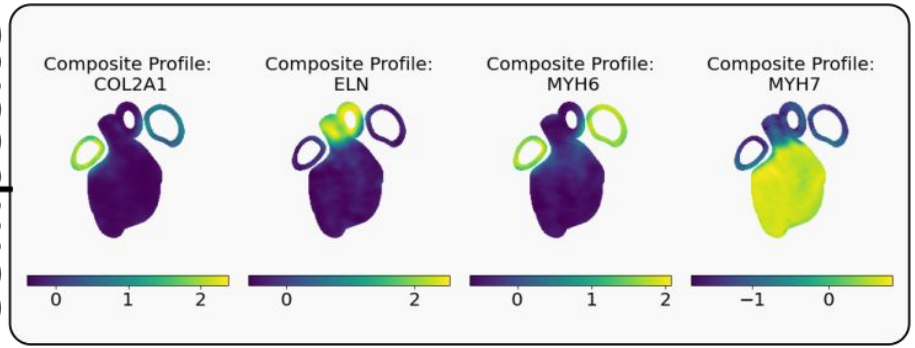
Note: Quite distinct differences in spatial structure between individual 1 and 2

Transferred expression (developmental heart)



eggplant :: Average representations and regional enrichment

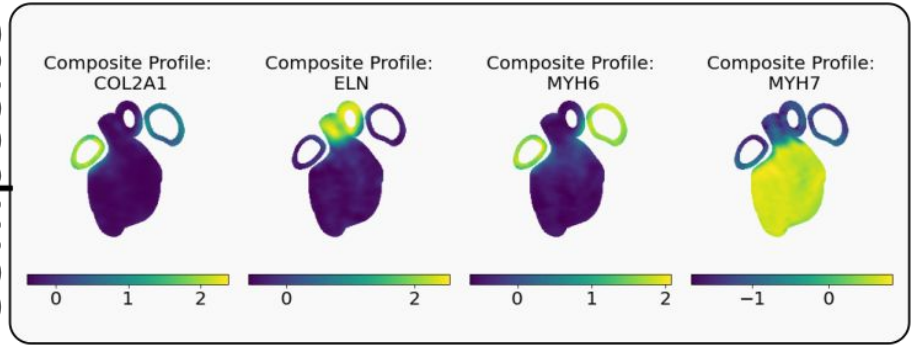
Composite



- Composite representations to **summarize information**

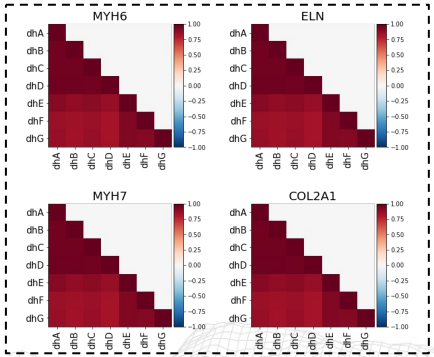
eggplant :: Average representations and regional enrichment

Composite



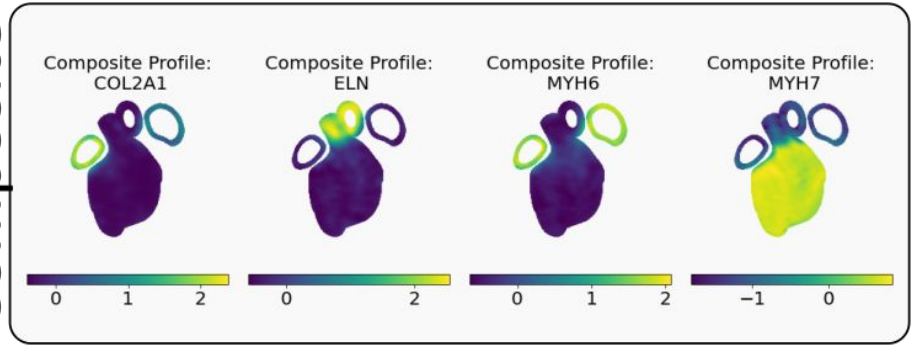
- Composite representations to **summarize information**
- Location-wise correlation across the 7 samples
  - **high correlation** within and **between individuals**

Correlation



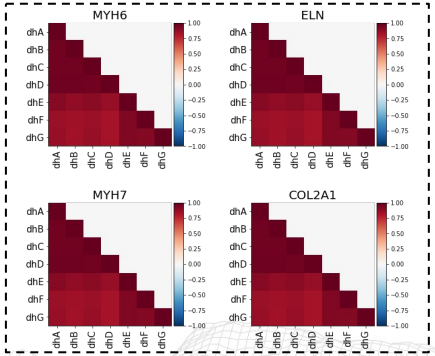
eggplant :: Average representations and regional enrichment

Composite

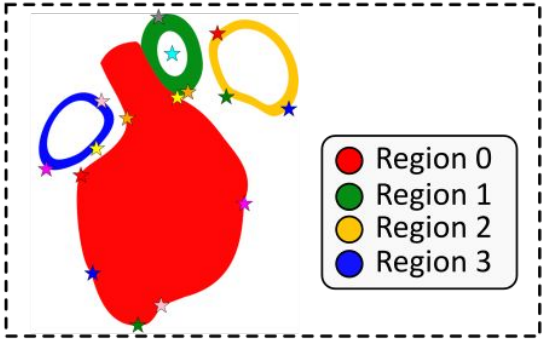


- Composite representations to **summarize information**
- Location-wise correlation across the 7 samples
  - **high correlation** within and **between individuals**
- Use annotated regions to for **enrichment** analysis
  - Assess up/downregulation in regions

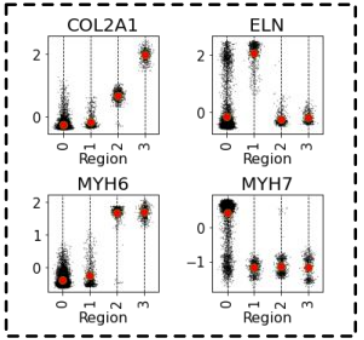
Correlation



Reference

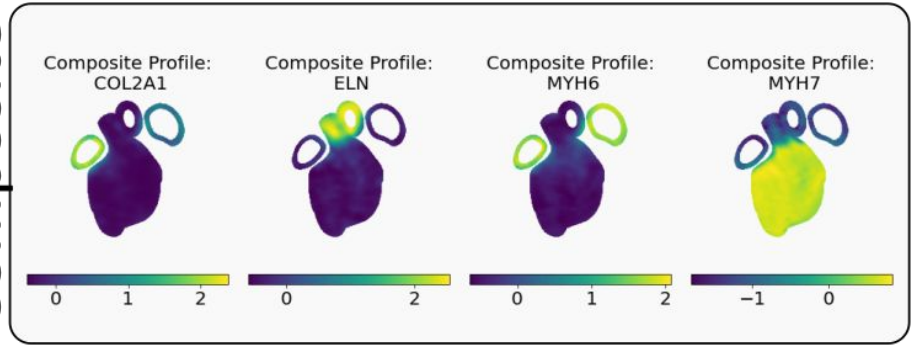


Enrichment



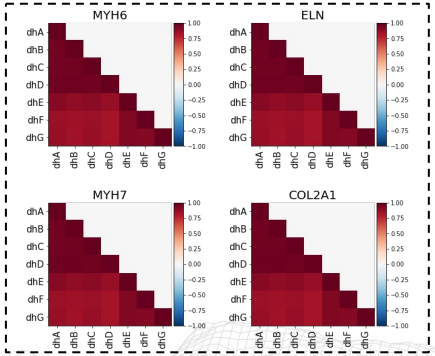
eggplant :: Average representations and regional enrichment

Composite

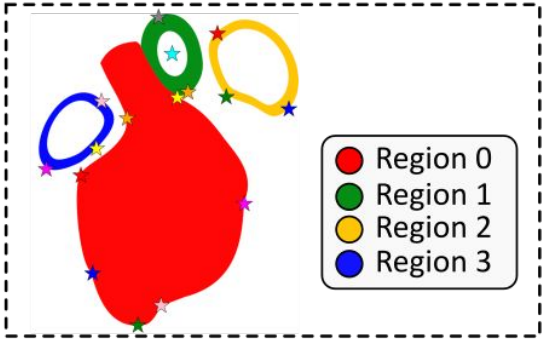


- Composite representations to **summarize information**
- Location-wise correlation across the 7 samples
  - **high correlation** within and **between individuals**
- Use annotated regions to for **enrichment** analysis
  - Assess up/downregulation in regions

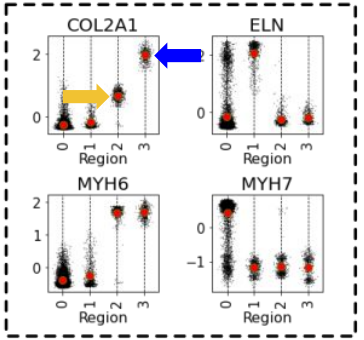
Correlation



Reference

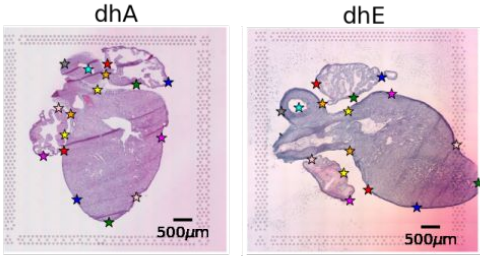


Enrichment

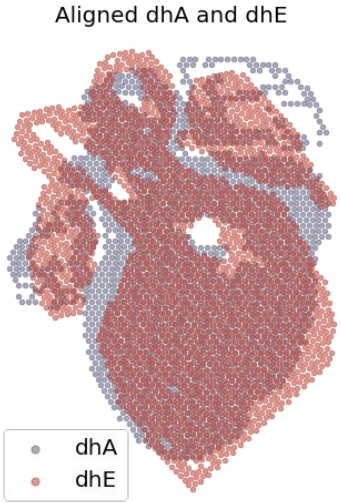


Alignment vs. information transfer

Alignment-based



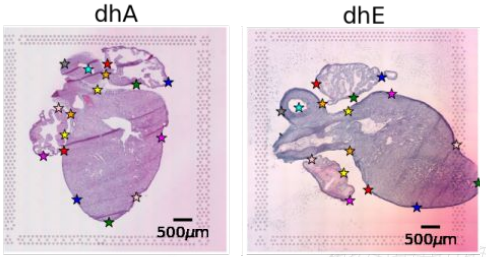
PASTE



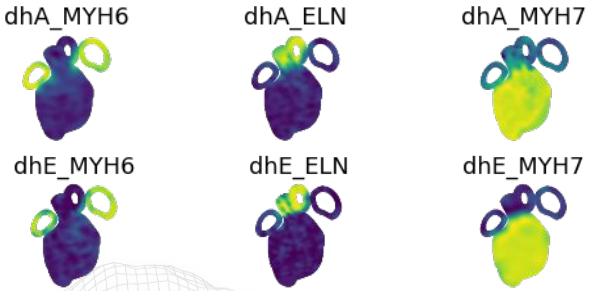
Coordinates still don't correspond to the same position.

PASTE: "*Alignment and integration of spatial transcriptomics data*", Zeira et al. | doi.org/10.1038/s41592-022-01459-6

CCF-based



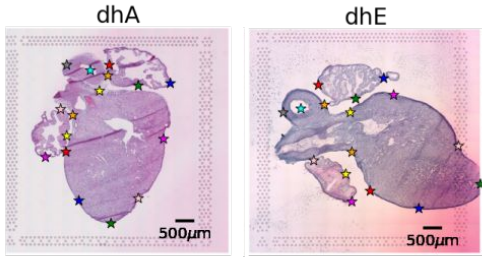
eggplant



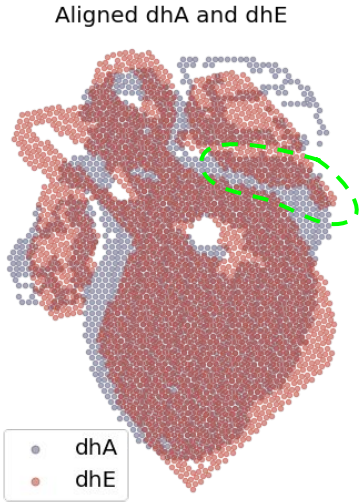
The data inhabits the same space, corresponding coordinates.

Alignment vs. information transfer

Alignment-based



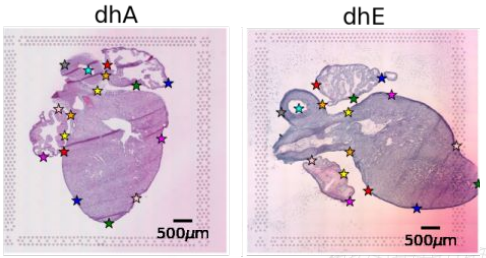
PASTE



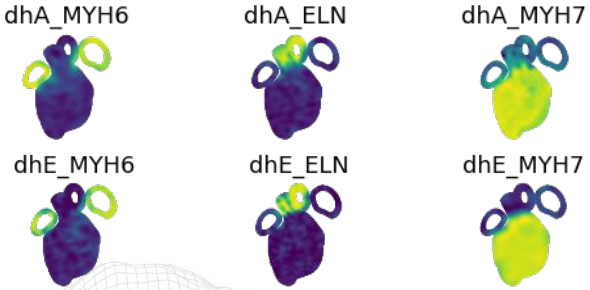
Coordinates still don't correspond to the same position.

PASTE: "*Alignment and integration of spatial transcriptomics data*", Zeira et al. | doi.org/10.1038/s41592-022-01459-6

CCF-based



eggplant



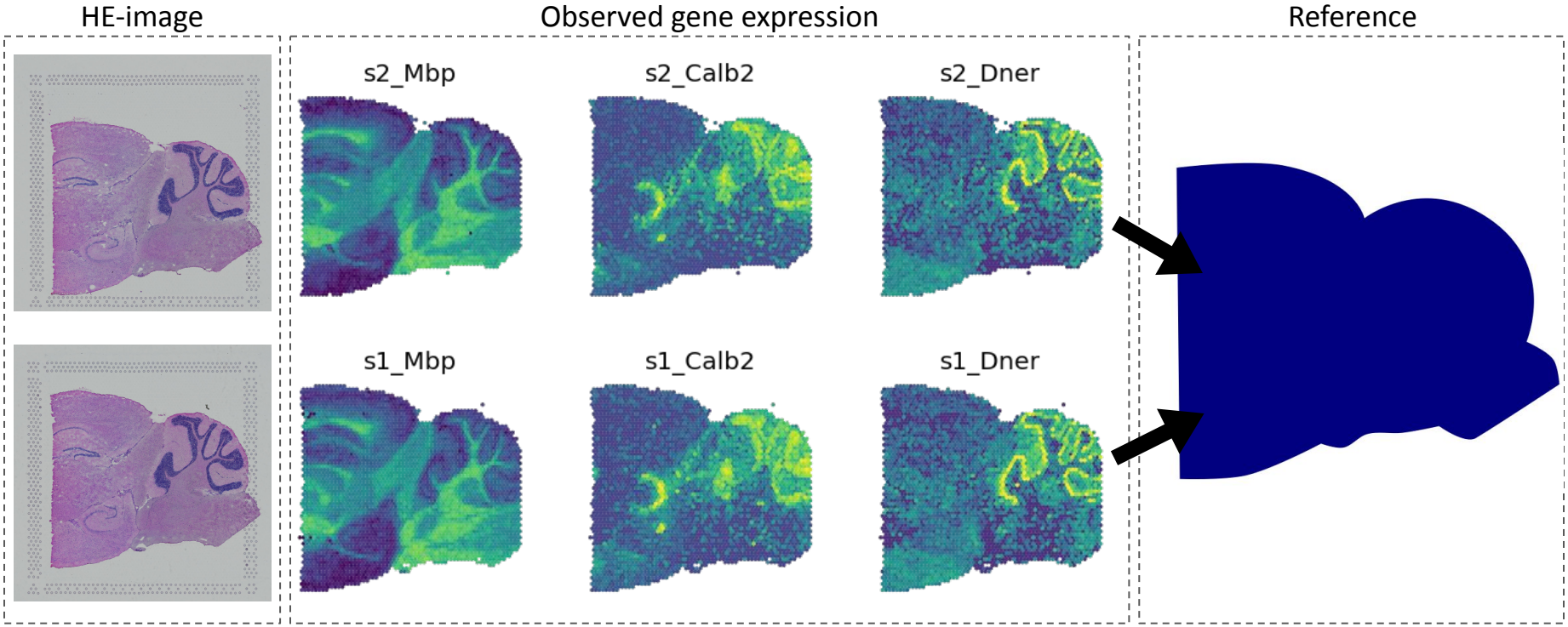
The data inhabits the same space, corresponding coordinates.

Example analyses

- **Human developmental heart** (✣)
  - Basic example
  - Variance in structure between individuals
  - Comparison with alignment
- **Mouse brain** (sagittal)
  - Complex structures
- **Perturbation in mouse brain** (✣)
  - spatial differential expression analysis (SDEA)

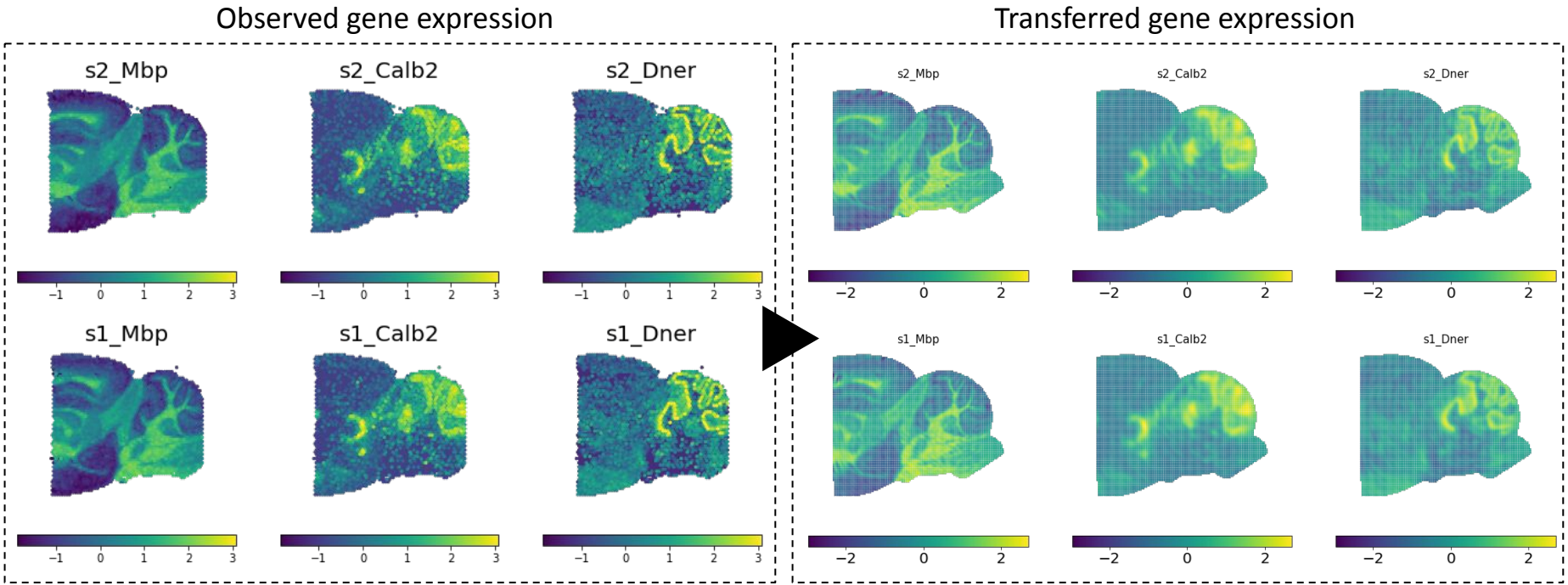
✣ := in the tutorial

Mouse brain (sagittal)



- 2 replicates of mouse brain (sagittal sections)
- Want to transfer: expression of 3 genes with **complex structure**

Mouse brain (sagittal)



Example analyses

- **Human developmental heart** (✣)
  - Basic example
  - Variance in structure between individuals
  - Comparison with alignment
- **Mouse brain** (sagittal)
  - Complex structures
- **Perturbation in mouse brain** (✣)
  - spatial differential expression analysis (SDEA)

✣ := in the tutorial

Perturbation in mouse brain

Free Radical Biology and Medicine 179 (2022) 277–287

Contents lists available at ScienceDirect

Free Radical Biology and Medicine

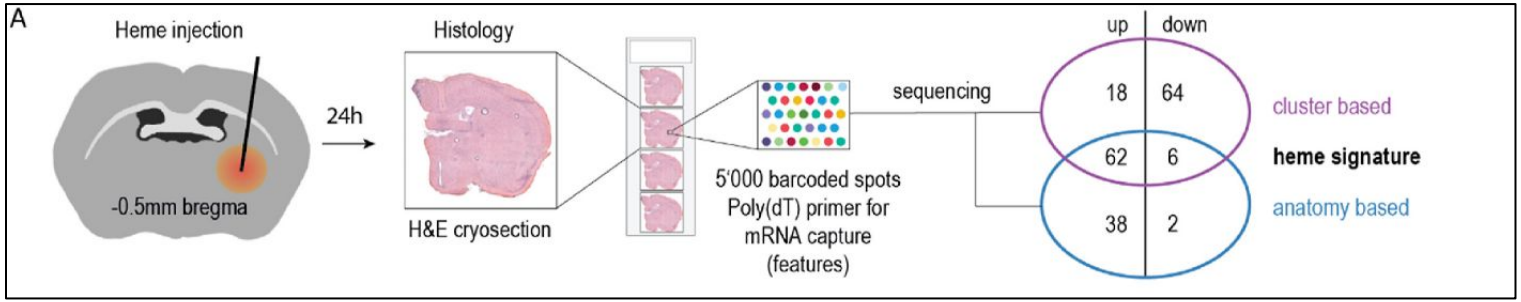
journal homepage: www.elsevier.com/locate/freeradbiomed

Spatial transcriptome analysis defines heme as a hemopexin-targetable inflammatoxin in the brain

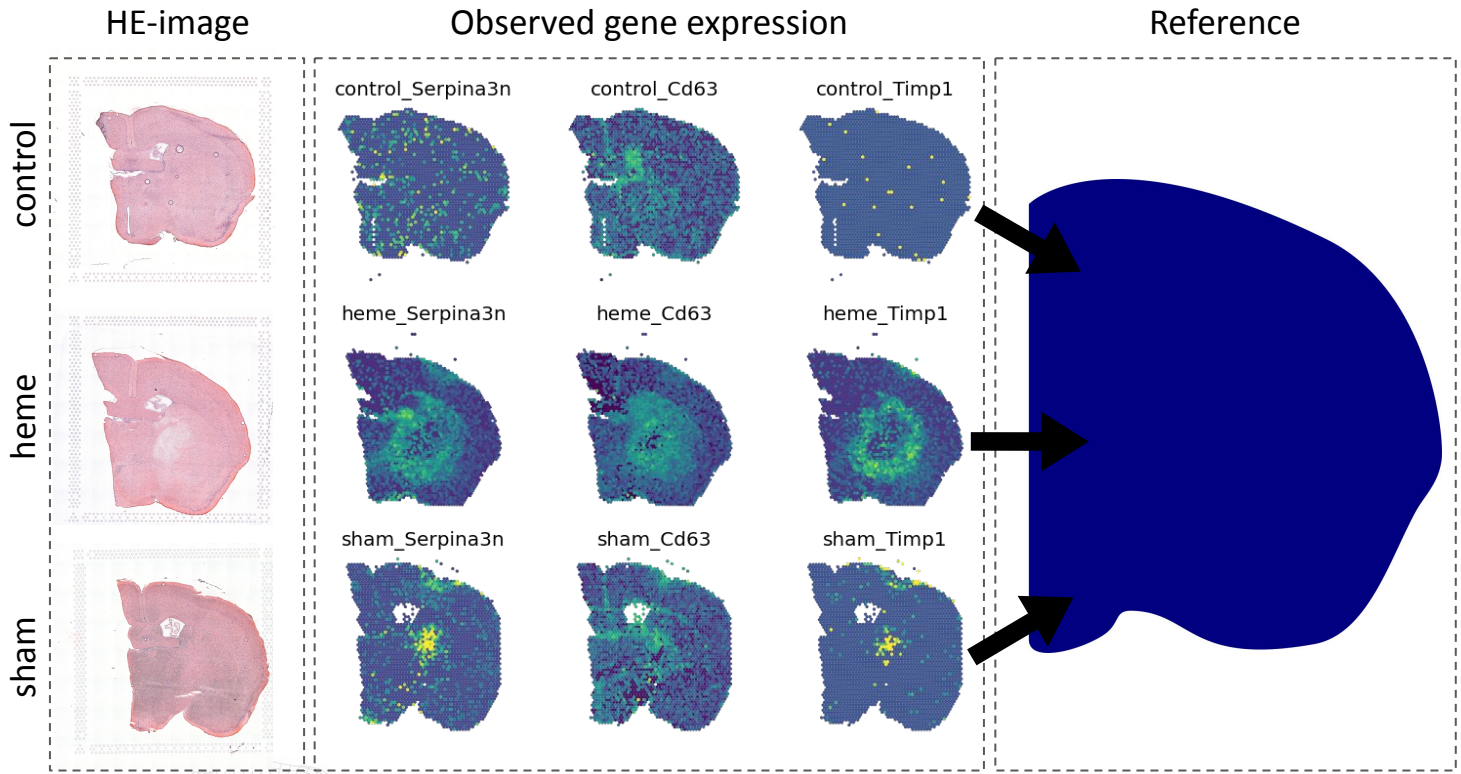
Raphael M. Buzzi [a,1], Kevin Akeret [b,1], Nina Schwendinger [a,b], Jan Klohs [c], Florence Vallelian [a], Michael Hugelshofer [b], Dominik J. Schaer [a,*]

[a] Division of Internal Medicine, Universitätsspital and University of Zurich, Zurich, Switzerland
[b] Department of Neurosurgery, Clinical Neuroscience Center, Universitätsspital und University of Zurich; Zurich, Switzerland
[c] Institute for Biomedical Engineering, University of Zurich and ETH Zurich, Zurich, Switzerland

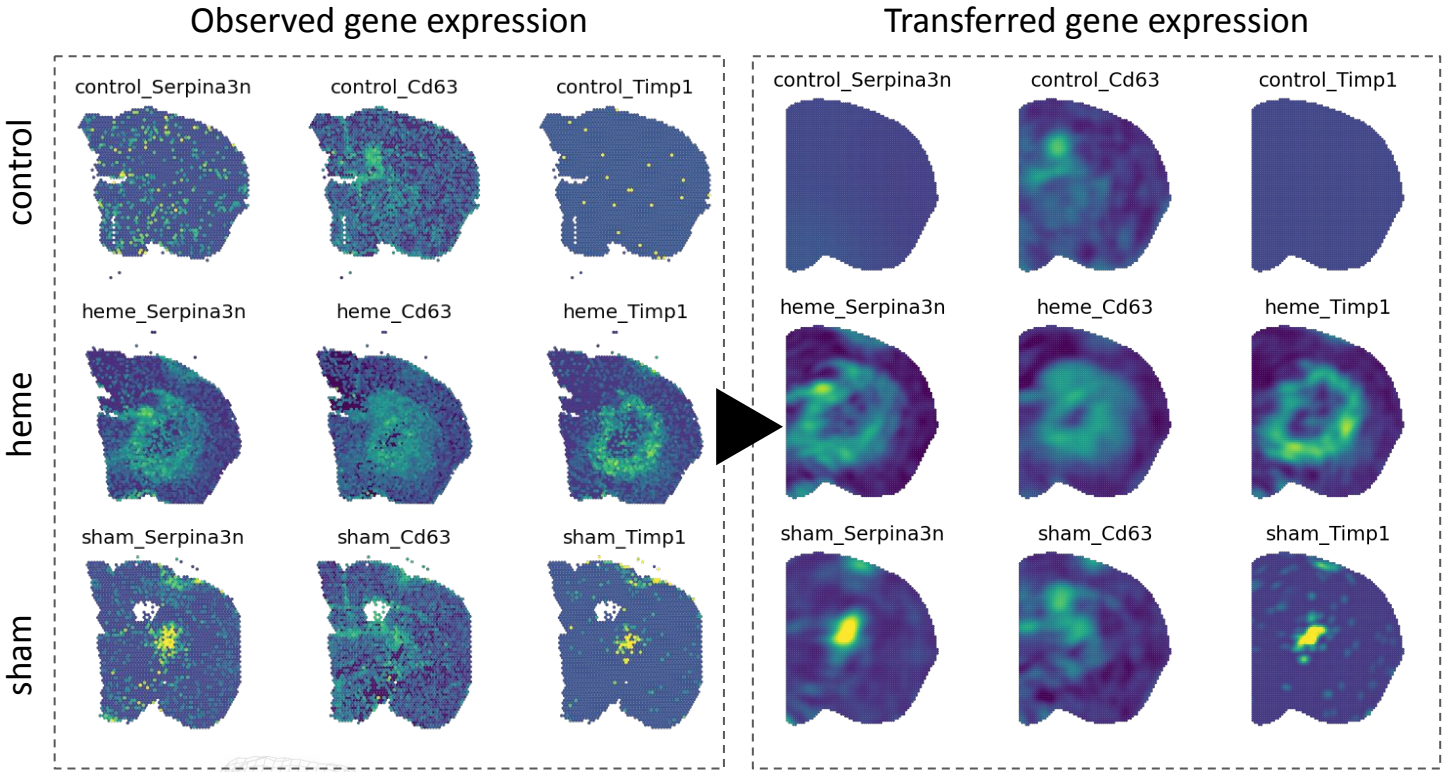
- Already published data
- The authors **injected mice with**:
  - nothing (control)
  - saline (sham)
  - heme
- Then **produced Visium data** from said mice



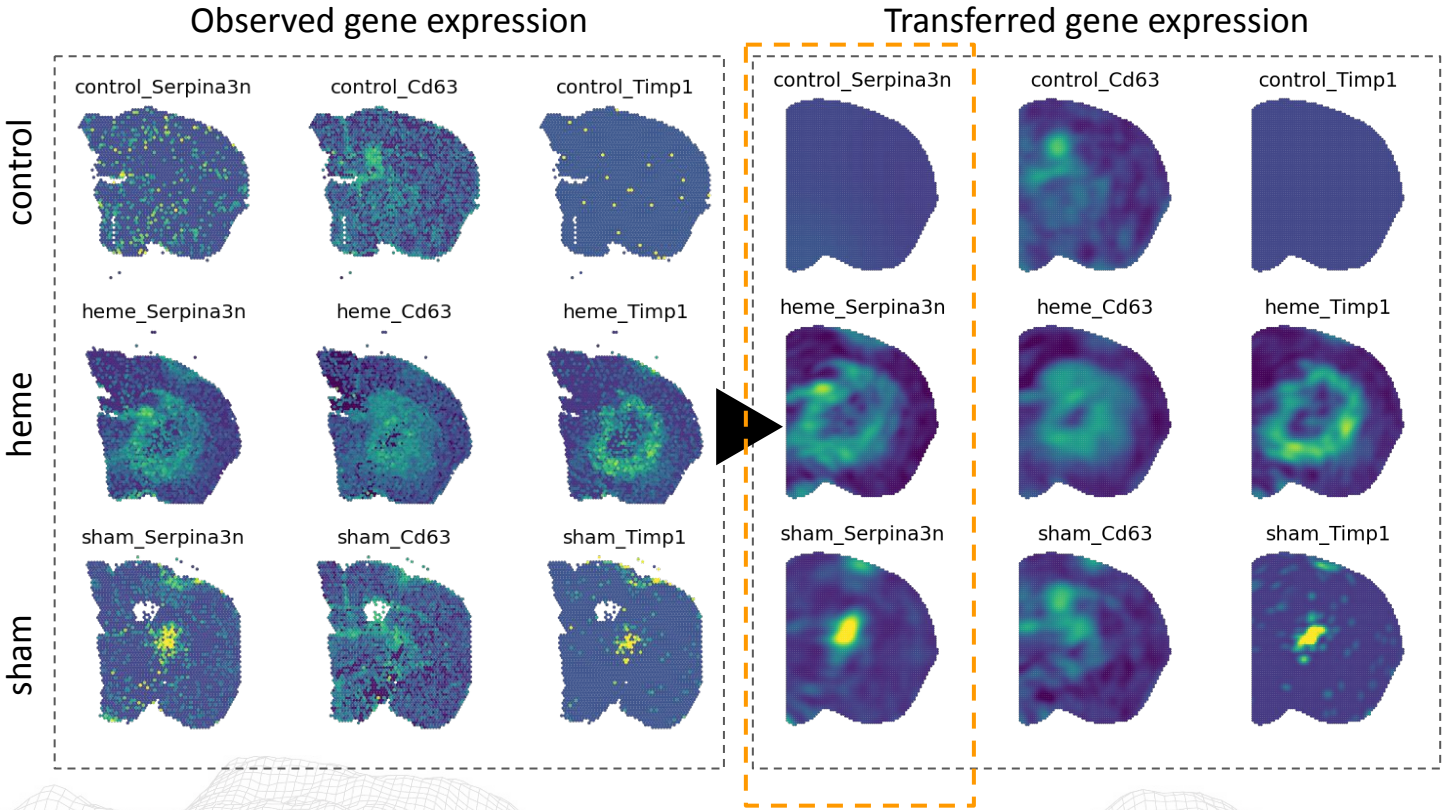
Perturbation in mouse brain



Perturbation in mouse brain



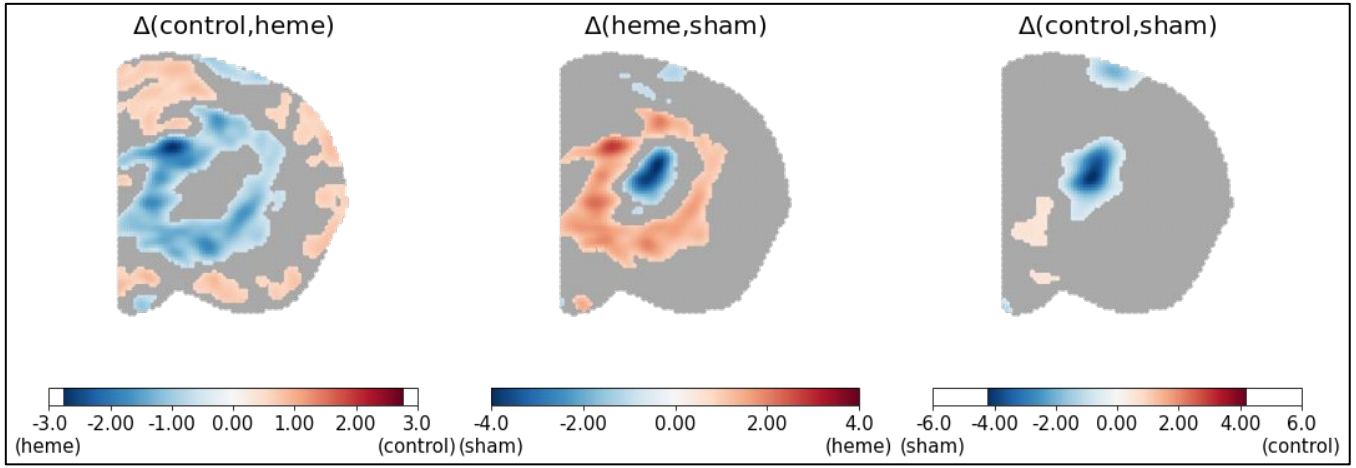
Perturbation in mouse brain



Example of SDEA with *Serpina3n*

Perturbation in mouse brain

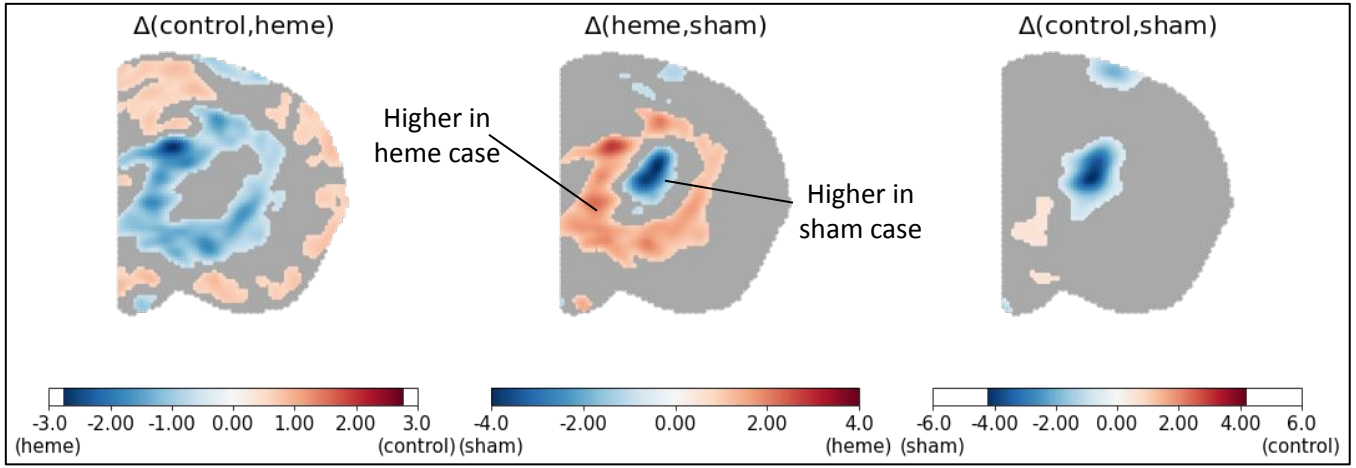
Spatial Differential Expression Analysis (SDEA) of *Serpina3n*



- **Compare** spatial gene expression **between conditions**
- Gray indicates non-DE regions
- Only possible when the information inhabits the same reference/CCF

Perturbation in mouse brain

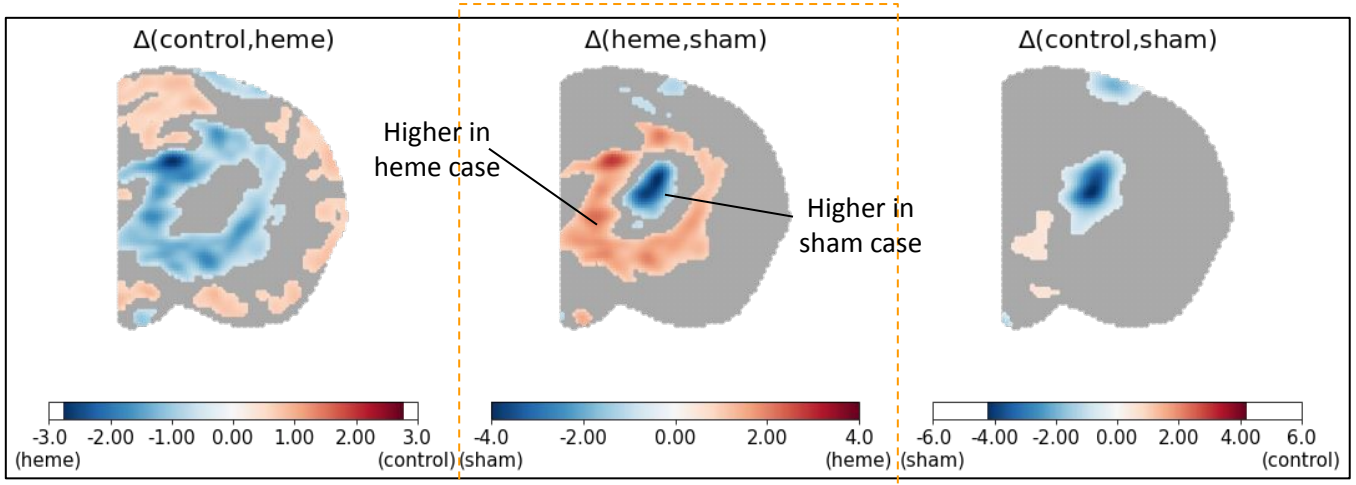
Spatial Differential Expression Analysis (SDEA) of *Serpina3n*

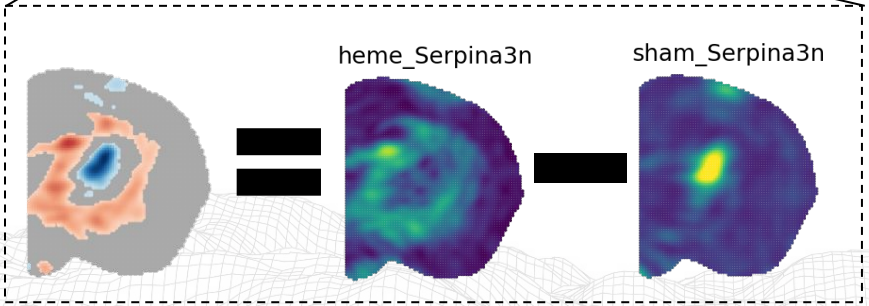


- **Compare** spatial gene expression **between conditions**
- Gray indicates non-DE regions
- Only possible when the information inhabits the same reference/CCF

Perturbation in mouse brain

Spatial Differential Expression Analysis (SDEA) of *Serpina3n*





- **Compare** spatial gene expression **between conditions**
- Gray indicates non-DE regions
- Only possible when the information inhabits the same reference/CCF

- Difference between the two conditions (spatial arithmetics)
- location-by-location comparison

Summary

- Common coordinate frameworks (**CCFs**) represents:
  - a unified framework that allows us to relate observations across samples
  - implicitly, this also means comparisons across **time points and conditions**
- Transferring data to a CCF is **not the same as aligning** samples
- For spatial data, **spatially aware CCFs** can be highly useful
- Our method (*eggplant)* uses:
  - **Landmark annotation**
  - Gaussian Process Regression
- Spatial Differential Expression Analysis (SDEA) is:
  - enabled why having the data in the same reference/CCF
  - a way to compare expression location-by-location

More information

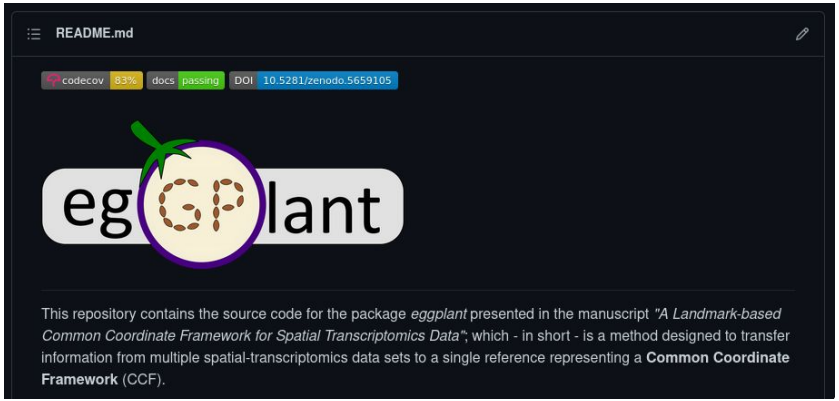
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY

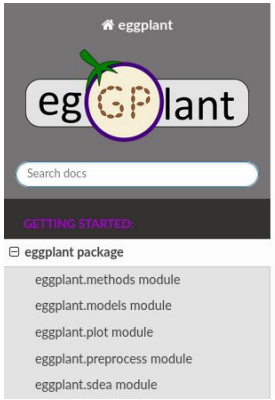
A Landmark-based Common Coordinate Framework for Spatial Transcriptomics Data

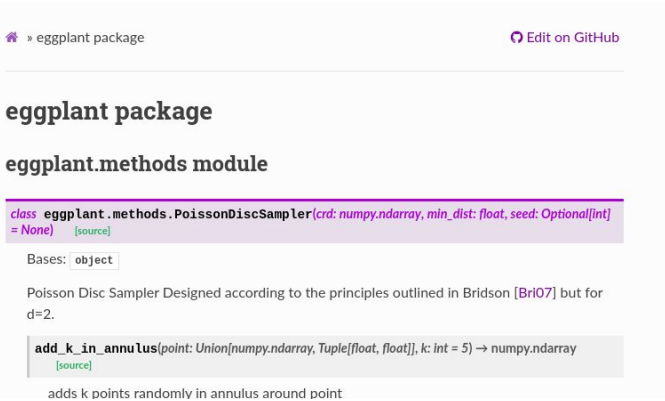
Alma Andersson, Žaneta Andrusivová, Paulo Czarnewski, Xiaofei Li, Erik Sundström, Joakim Lundeberg

doi: https://doi.org/10.1101/2021.11.11.468178

For this course:
> Zulip: #scog_spatial_workshop_2022 | topic eggplant







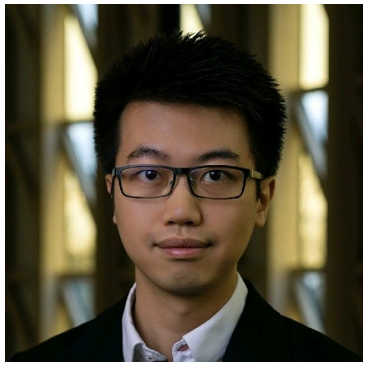
https://github.com/almaan/eggplant

https://spatial-eggplant.readthedocs.io

Acknowledgments

Acknowledgements

Žaneta Andrusivová



Xiaofei Li

Paulo Czarnewski

Erik Sundström

Joakim Lundeberg

Spatial Research Group

Developmental Heart Data:
Zaneta, Xiaofei and Erik

Code beta-testing and discussions:
Paulo Czarnewski

Supervisor:
Joakim Lundeberg

Special Shoutout: Giovanni Palla for excellent feedback!!!

Thank you for the attention!

Questions?

 @aalmaander

 alma.andersson@differentiable.net

 differentiable.net

Notebook walkthrough

Included:

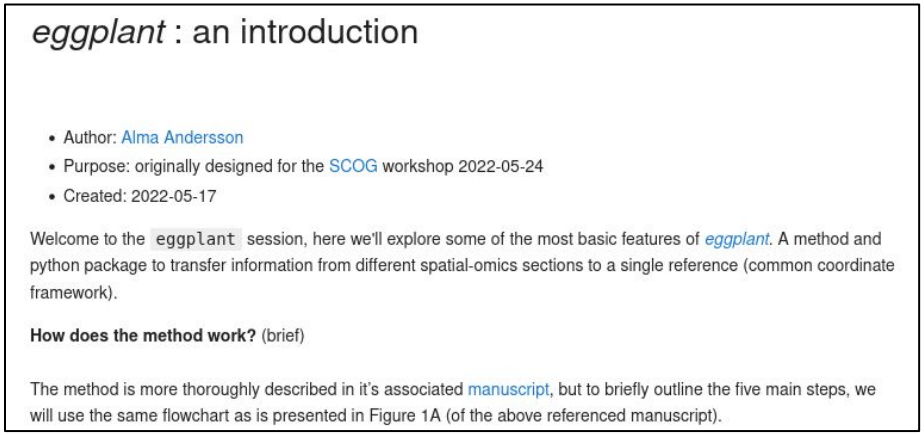
- Load data and reference
- Prepare data and reference for transfer of information
- Transfer of information to reference, using two strategies:
  - exact (but slightly slower) approach
  - approximate (but *significantly* faster) approach
- Creating composite representations
- Regional enrichment analysis
- Spatial differential gene expression analysis (SDEA)

Datasets:

- Human developmental heart
- Perturbed mouse brain

Comment:

- Run notebook in **DEMO** mode to load pre-analyzed data



eggplant : an introduction

- Author: Alma Andersson
- Purpose: originally designed for the SCOG workshop 2022-05-24
- Created: 2022-05-17

Welcome to the `eggplant` session, here we'll explore some of the most basic features of *eggplant*. A method and python package to transfer information from different spatial-omics sections to a single reference (common coordinate framework).

How does the method work? (brief)

The method is more thoroughly described in it's associated manuscript, but to briefly outline the five main steps, we will use the same flowchart as is presented in Figure 1A (of the above referenced manuscript).

https://github.com/theislab/spatial_scog_workshop_2022/tree/main/eggplant/notebooks