Computational Analysis of Spatial Transcriptomics data

BrainOmics 2.0 - Day 3

Agenda

- Spatial dataset and data format

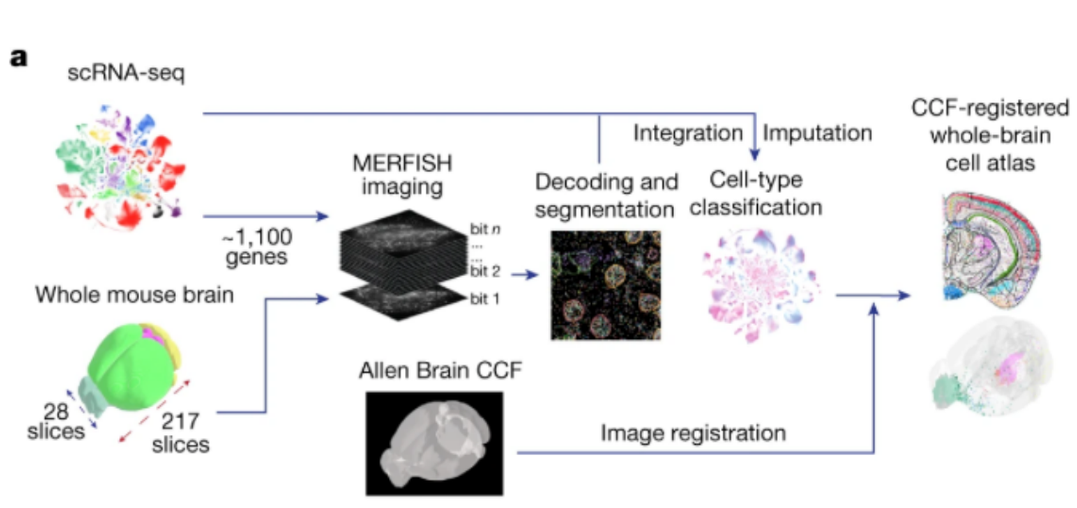
- Unmeasured spatial genes imputation
  - with SpaGE
  - with Tangram

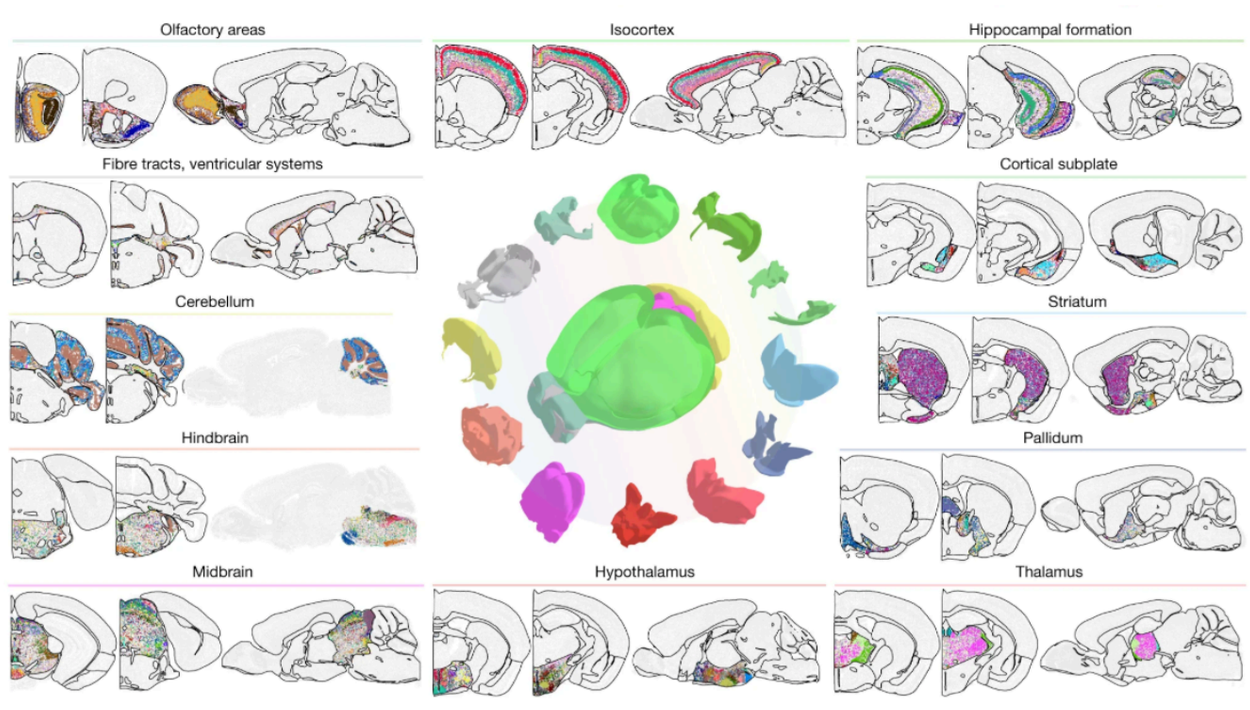
- Spatial clustering with BANKSY

- Access image in AnnData object

Dataset: MERFISH spatial transcriptomics of adult mouse brain

- Obtained with Vizgen MERFISH technology
  - High resolution and sensitivity
  - Targeted spatial transcriptomics = limited gene panel
- Zhuang-ABCA-1: 147 sections with a 1122 gene panel and 2.8M spots



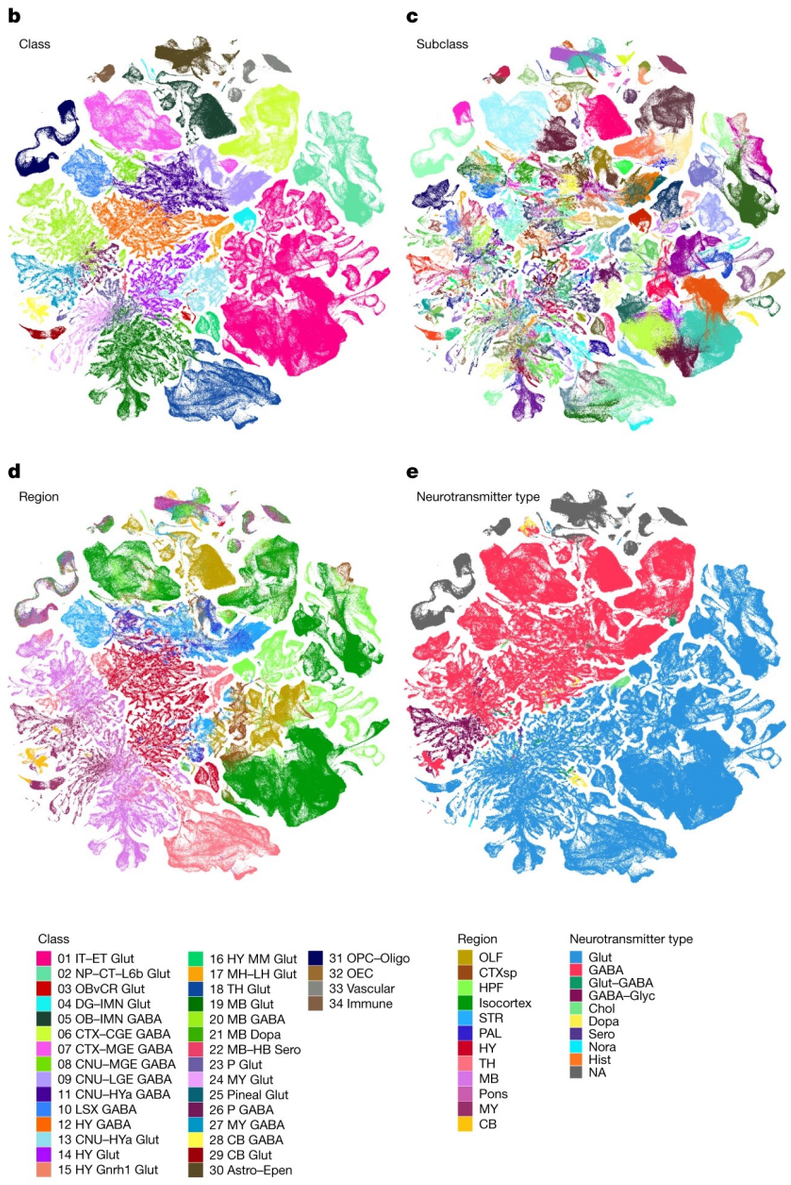


Molecularly defined and spatially resolved cell atlas of the whole mouse brain (Zhang et al. 2023, Nature)

Dataset: MERFISH spatial transcriptomics of adult mouse brain

Whole-brain single-cell RNA-sequencing

- 4 million cells and 32.285 genes
- Annotated per region, class, subclass and neurotransmitter type
- will be used as a reference to impute gene expression of unmeasured genes in the spatial dataset



A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain (Yao et al. 2023, Nature)

Dataset access: Allen Brain Cell (ABC) Atlas

Each directory corresponds to a dataset; check the available datasets with `abc_cache.list_directories`

```
['Allen-CCF-2020',
 'MERFISH-C57BL6J-638850',
 'MERFISH-C57BL6J-638850-CCF',
 'MERFISH-C57BL6J-638850-imputed',
 'MERFISH-C57BL6J-638850-sections',
 'WHB-10Xv3',
 'WHB-taxonomy',
 'WMB-10X',
 'WMB-10XMulti',
 'WMB-10Xv2',
 'WMB-10Xv3',
 'WMB-neighborhoods',
 'WMB-taxonomy',
 'Zhuang-ABCA-1',
 'Zhuang-ABCA-1-CCF',
 'Zhuang-ABCA-2',
 'Zhuang-ABCA-2-CCF',
 'Zhuang-ABCA-3',
 'Zhuang-ABCA-3-CCF',
 'Zhuang-ABCA-4',
 'Zhuang-ABCA-4-CCF']
```

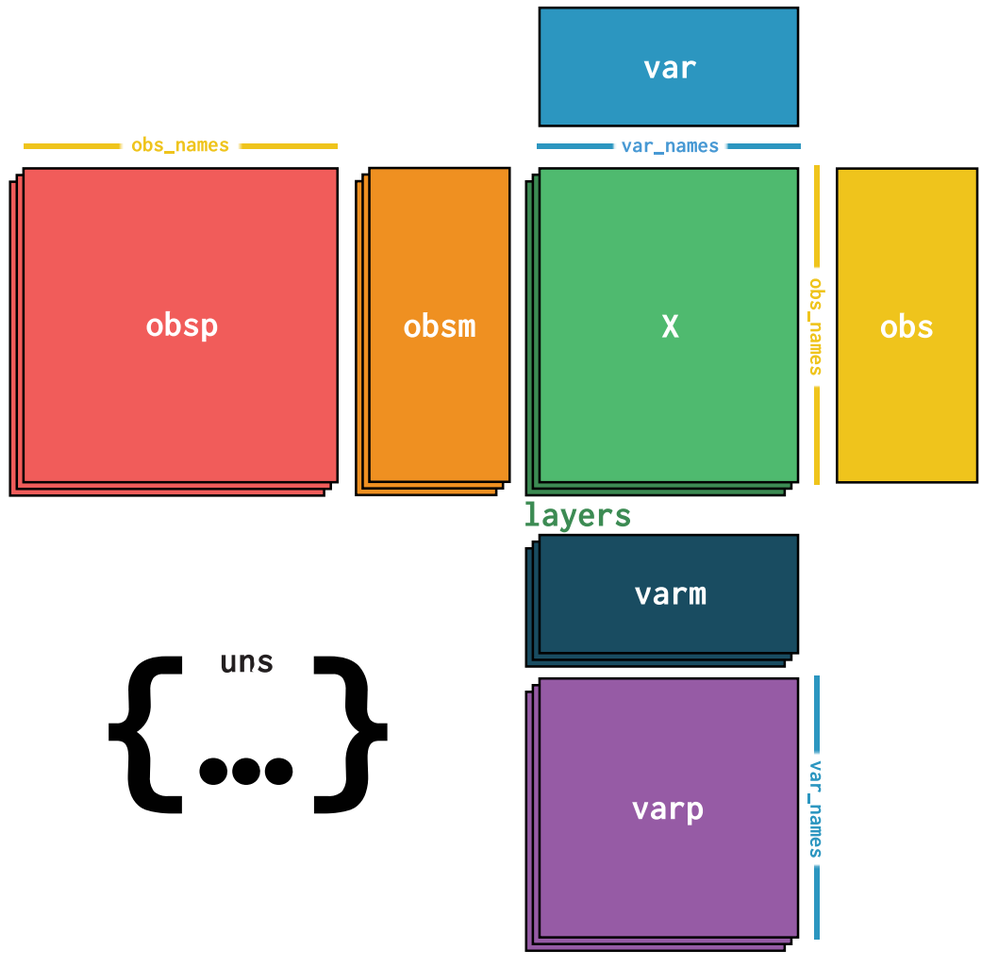
Single-cell (mouse) reference datasets

Spatial dataset we will use

Once you decide on which dataset to use check which files are available with `abc_cache.list_data_files` and `abc_cache.list_metadata_files`

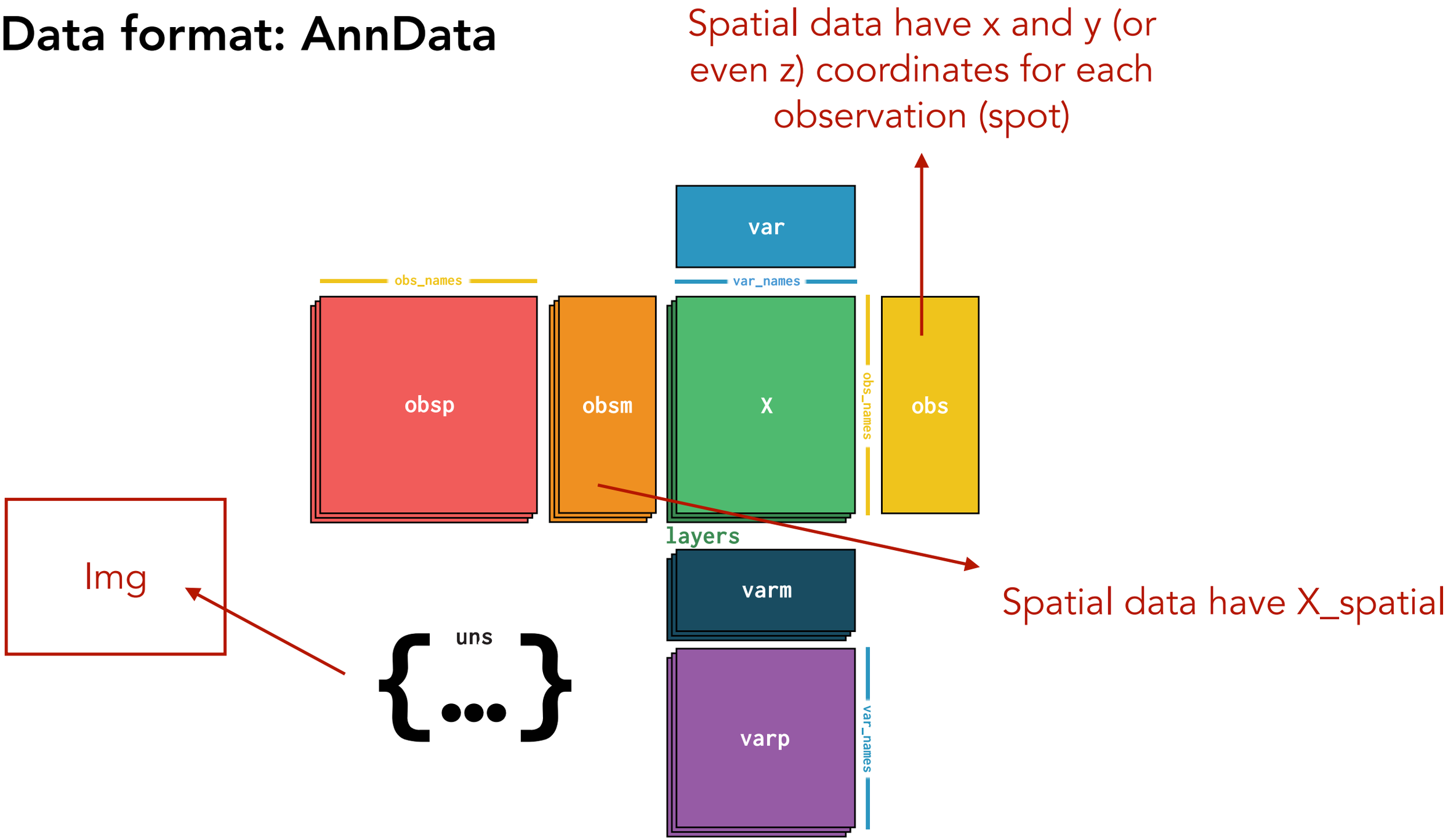
Data format: AnnData

- **X** is the main data matrix
  - **.layer**
- **obs** is a pandas dataframe that stores metadata and annotations for each observation
- **var** is a pandas that contains metadata for each variable,
- **uns** is a set of dictionary for storing unstructured or auxiliary data that doesn't fit neatly into the obs, var or X matrices.
- **obsm** and **varm** are for metadata that have many dimensions to it



anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

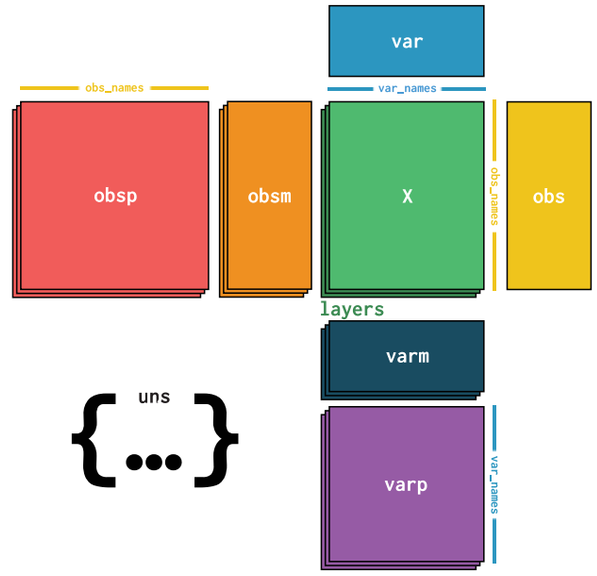
Data format: AnnData



anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read an AnnData object?



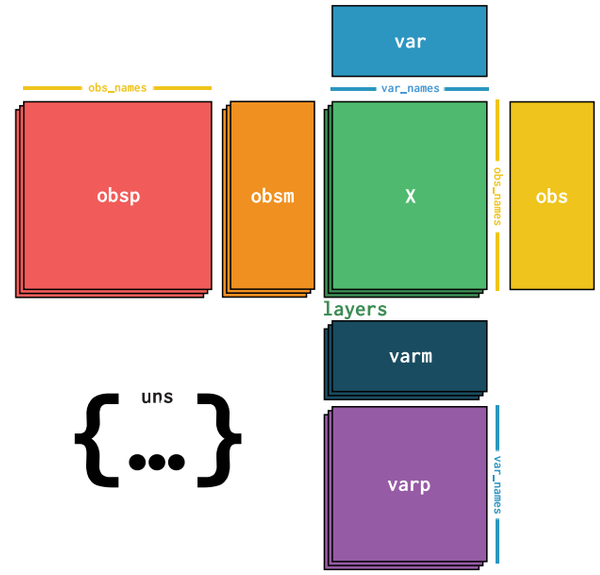
anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read an AnnData object?
  func: ad.read_h5ad

EX1

EX2



- Add the raw AnnData to the ‘raw' layer and the log-transformed AnnData to the ‘log2’ layer

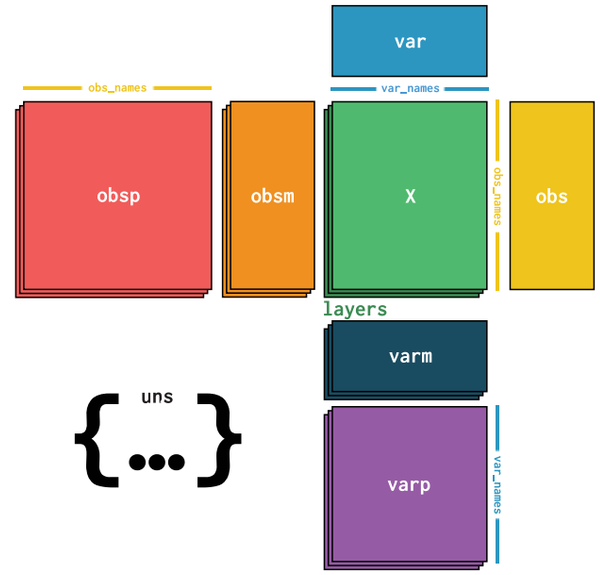
anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read an AnnData object?
  func: ad.read_h5ad

EX1

EX2



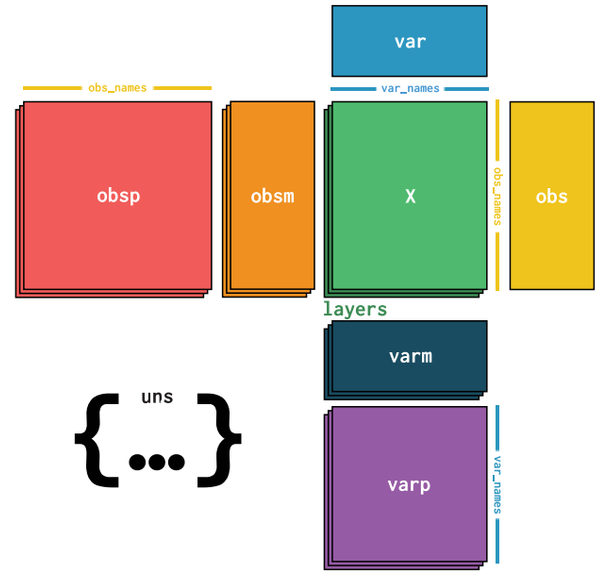
- Add the raw AnnData to the ‘raw' layer and the log-transformed AnnData to the ‘log2’ layer

EX3

anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read a csv file?
  func: pd.read_csv



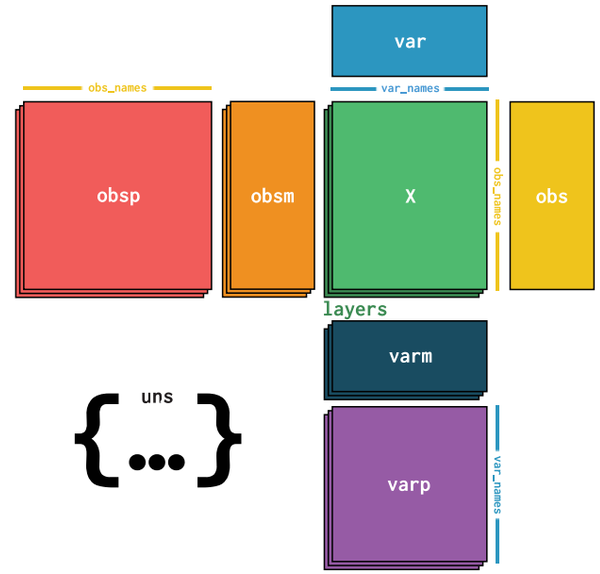
anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read a csv file?
  func: pd.read_csv

EX4

EX5



- Add genes to the right field in the AnnData object

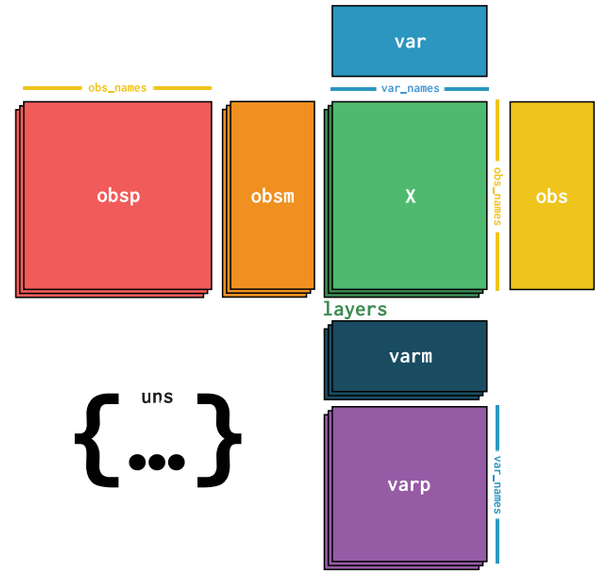
anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Read data

- How to read a csv file?
  func: pd.read_csv

EX4

EX5



- Add genes to the right field in the AnnData object
  adata.var = genes

EX6

anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Data format: AnnData

It's possible to subset a AnnData object to keep only specific observations (subsetting obs) or specific genes (subsetting var)

We will do this to keep only one out of the 147 sections in the whole-brain.

```
adata.obs
```

cell_label	brain_section_label	feature_matrix_label	donor_label	donor_genotype	donor_sex	cluster_alias	x	y
182941331246012878296807398333956011710	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	704	0.682522	3.366483
221260934538535633595532020856387724686	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	5243	0.667690	3.442241
222287926068147815332409556230309437 08	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.638731	3.474328
272043042552227961220474294517855477150	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.653425	3.433218

anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Data format: AnnData

It’s possible to subset a AnnData object to keep only specific observations (subsetting obs) or specific genes (subsetting var)

We will do this to keep only one out of the 147 sections in the whole-brain.

```
adata.obs
```

cell_label	brain_section_label	feature_matrix_label	donor_label	donor_genotype	donor_sex	cluster_alias	x	y
182941331246012878296807398333956011710	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	704	0.682522	3.366483
221260934538535633595532020856387724686	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	5243	0.667690	3.442241
222287926068147815332409556230309437 08	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.638731	3.474328
272043042552227961220474294517855477150	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.653425	3.433218

How? Remember that .obs and .var are (pandas) DataFrames and you can use all the pandas methods on them

```
df_subset = df[df[‘column_name’] == ‘something’] (.OBS)

df_subset = df[:, df[‘column_name’] == ‘something’] (.VAR)
```

anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Data format: AnnData

EX 7

It’s possible to subset a AnnData object to keep only specific observations (subsetting obs) or specific genes (subsetting var)

We will do this to keep only one out of the 147 sections in the whole-brain.

```
adata.obs
```

cell_label	brain_section_label	feature_matrix_label	donor_label	donor_genotype	donor_sex	cluster_alias	x	y
182941331246012878296807398333956011710	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	704	0.682522	3.366483
221260934538535633595532020856387724686	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	5243	0.667690	3.442241
222287926068147815332409556230309437 08	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.638731	3.474328
272043042552227961220474294517855477150	Zhuang-ABCA-1.089	Zhuang-ABCA-1	Zhuang-ABCA-1	wt/wt	F	14939	0.653425	3.433218

How? Remember that .obs and .var are (pandas) DataFrames and you can use all the pandas methods on them

```
df_subset = df[df[‘column_name’] == ‘something’] (.OBS)

df_subset = df[:, df[‘column_name’] == ‘something’] (.VAR)

adata_section = adata[adata.obs[‘brain_section_label’] == ‘Zhuang-ABCA-1.0XX’]
```

anndata: Access and store annotated data matrices (Virshup et al. 2024, JOSS)

Quality Control and EDA

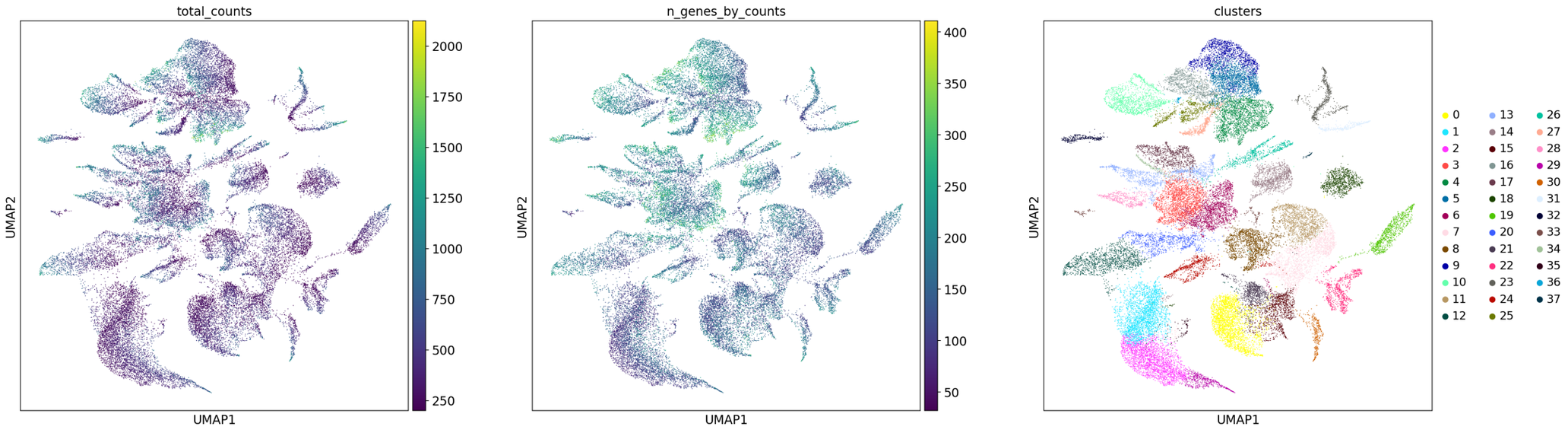
Filter cells:

- by minimum counts
- by maximum counts
- on mitochondrial gene expression

Filter genes:

- by minimum cells

use sc.pp.filter_genes and minimum cells = 20



Quality Control and EDA

EX 8

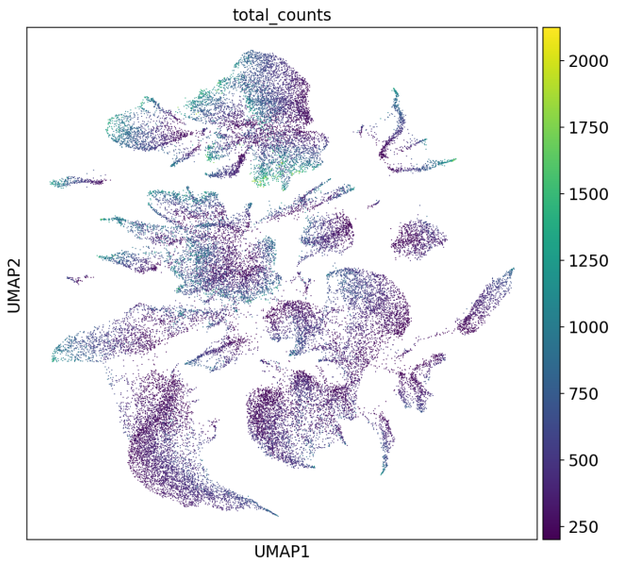
Filter cells:

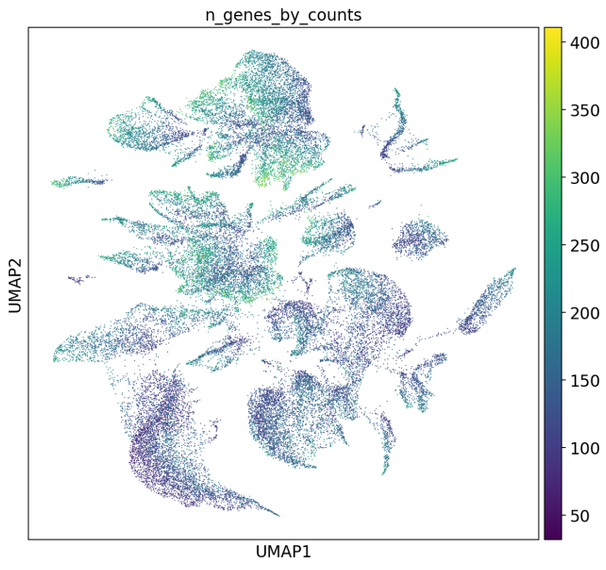
- by minimum counts
- by maximum counts
- on mitochondrial gene expression

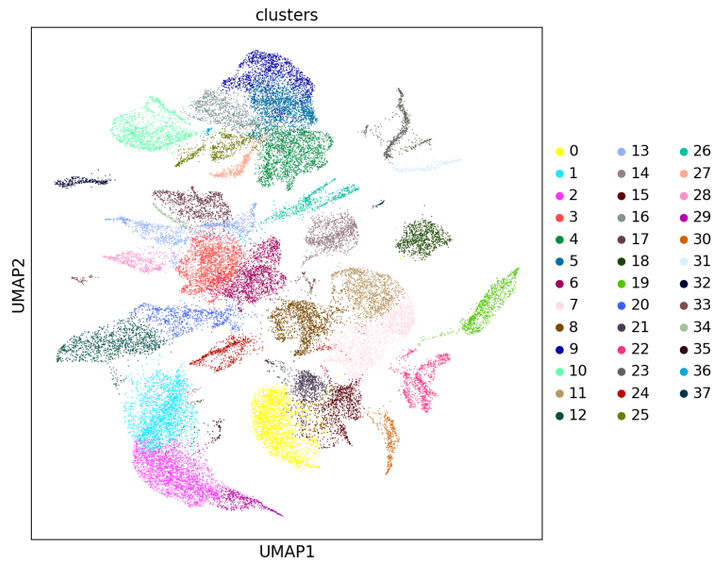
Filter genes:

- by minimum cells

use sc.pp.filter_genes and minimum cells = 20







Spatial plotting

EX 9

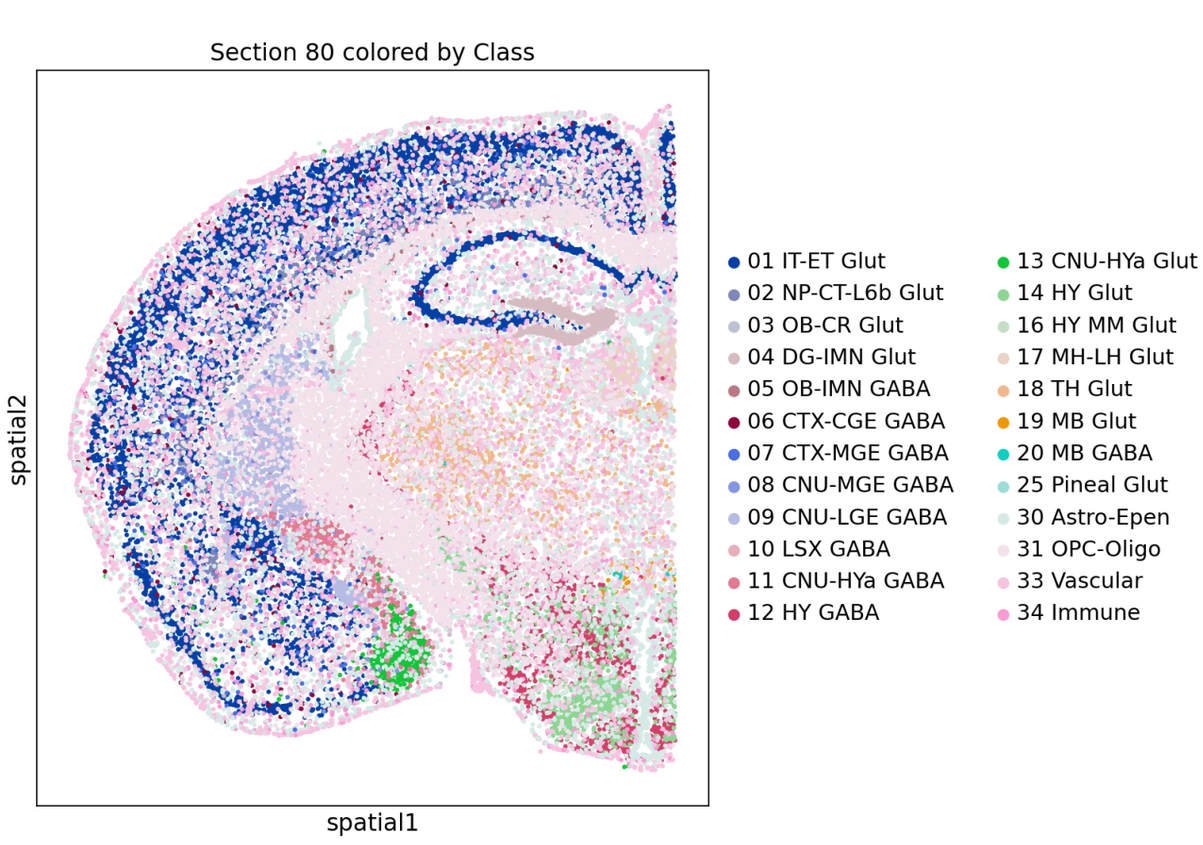
Add X_spatial to the correct field; which one is it?

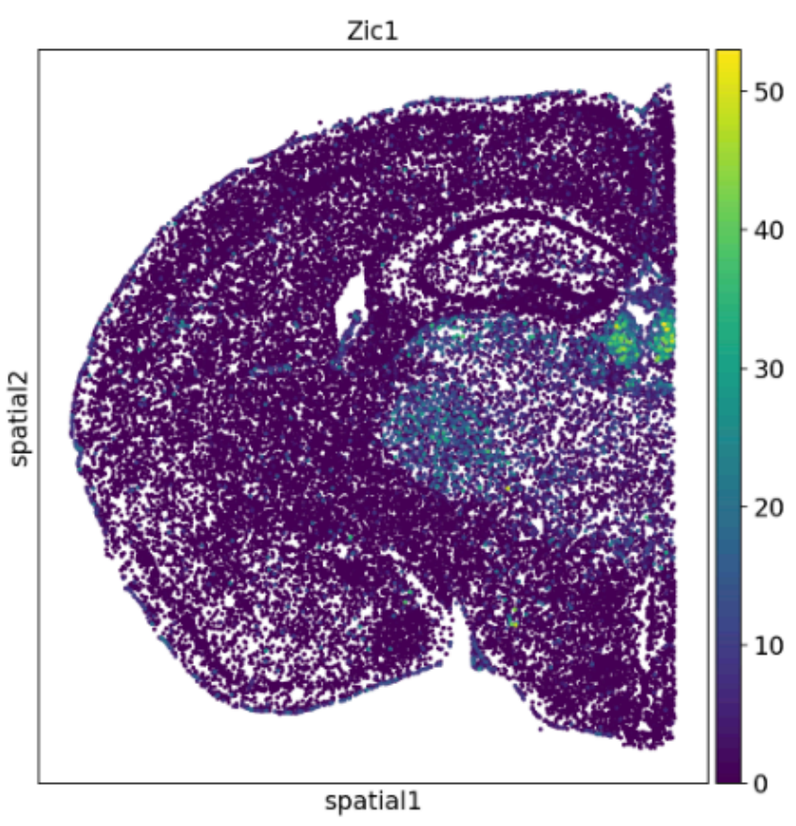
Spatial plotting

EX 9

Add X_spatial to the correct field; which one is it?

Now we can plot the class distribution in space and the spatial gene expression with with **sc.pl.spatial**



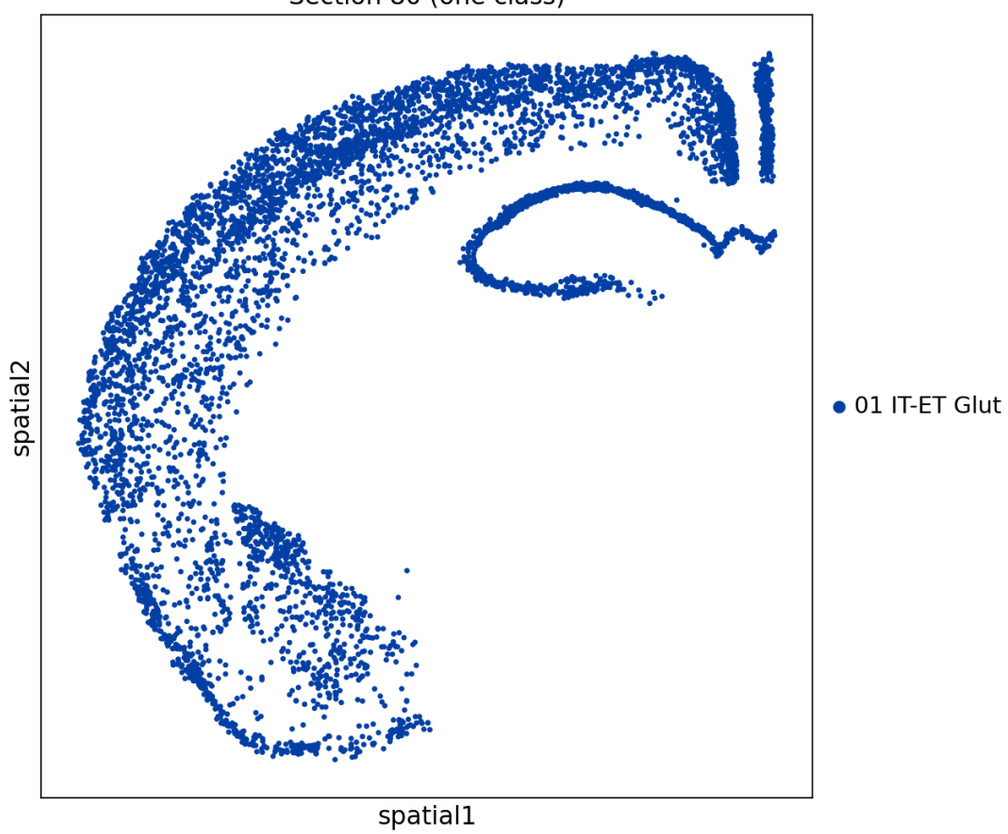


Spatial plotting

EX 10

Subset the AnnData object to keep only one class type for better visualisation

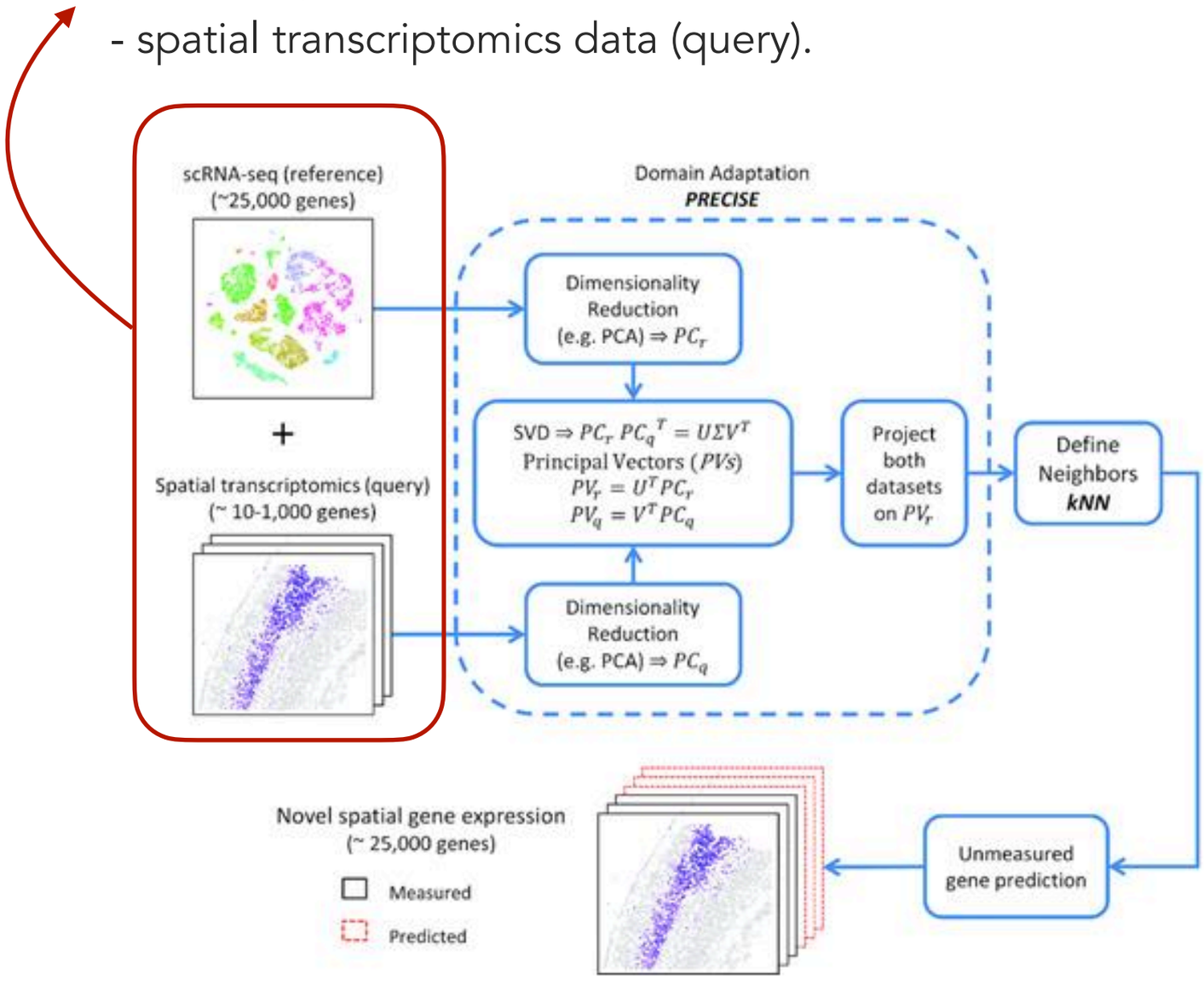
Which obs column should you use?



Unmeasured gene imputation with SpaGE

Input of SpaGE algorithm:

- gene expression matrices corresponding to the scRNA-seq data (reference)
- spatial transcriptomics data (query).

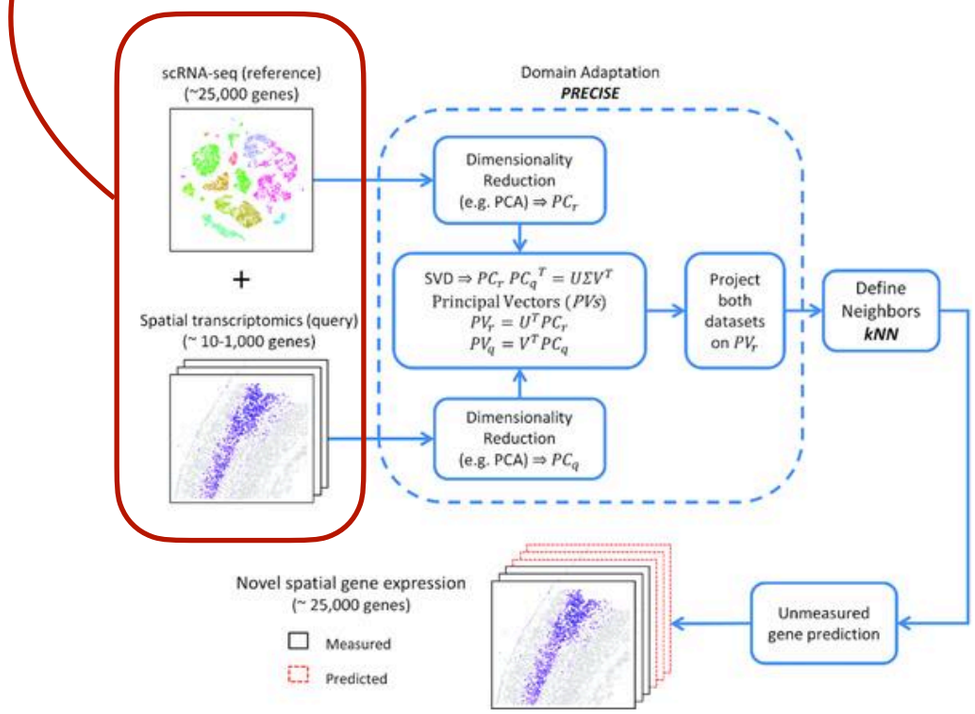


SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with SpaGE

Input of SpaGE algorithm:

- gene expression matrices corresponding to the scRNA-seq data (reference)
- spatial transcriptomics data (query).



Single cell reference

```
print(f"Size of WMB-10Xv2 data: {abc_cache.get_directory_data_size('WMB-10Xv2')}")
```

```
Size of WMB-10Xv2 data: 104.16 GB
```

```
print(f"Size of WMB-10Xv3 data: {abc_cache.get_directory_data_size('WMB-10Xv3')}")
```

```
Size of WMB-10Xv3 data: 176.41 GB
```

```
print(f"Size of WMB-10X Multi data: {abc_cache.get_directory_data_size('WMB-10XMulti')}")
```

```
Size of WMB-10X Multi data: 211.28 MB
```

Subsetted for you with obtain_scref.py:

1. Download files (280GB)
2. Using metadata, filter out cells that belong to clusters not present in the spatial data
3. Subset to keep only a percentage

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

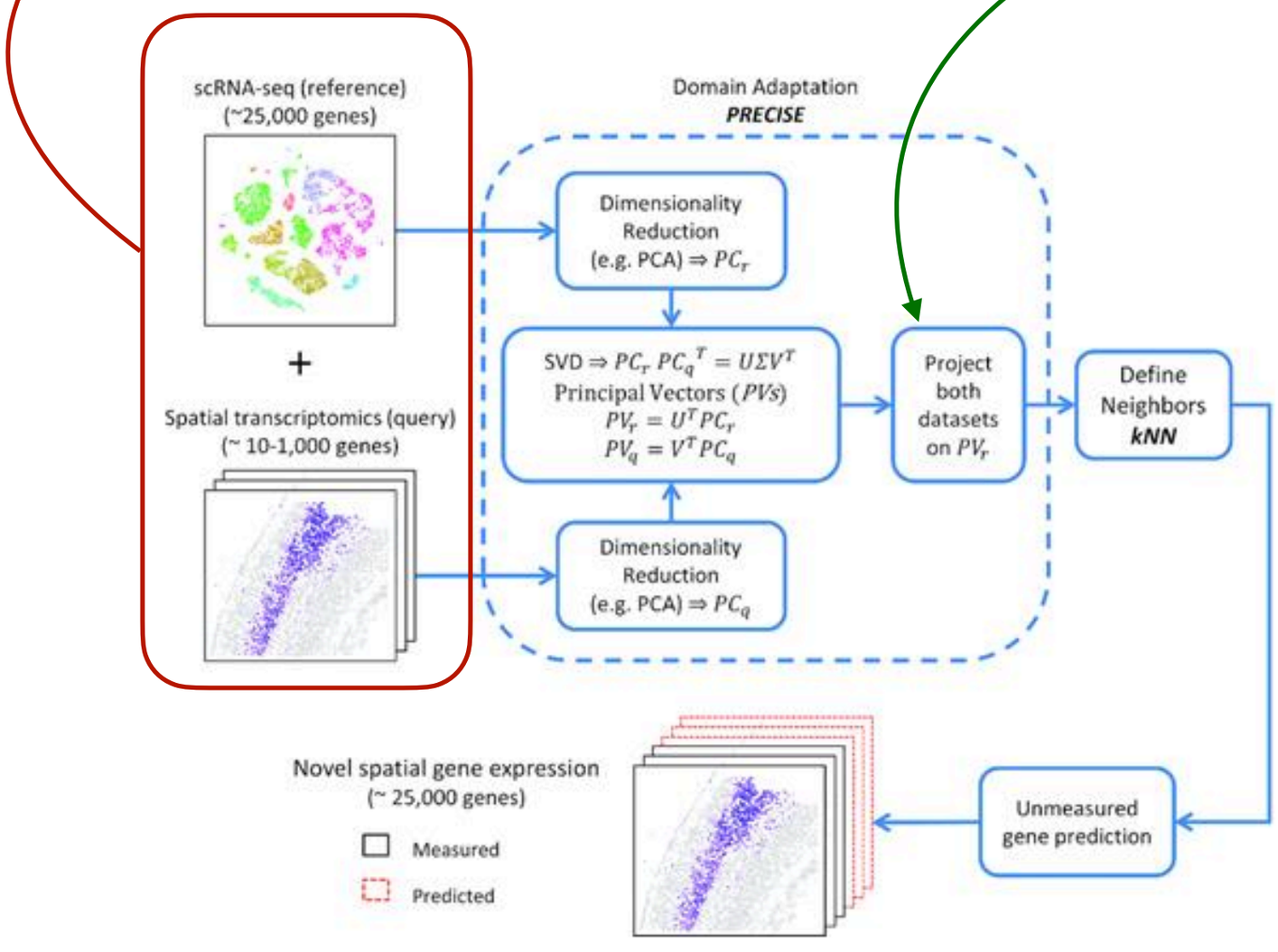
Unmeasured gene imputation with SpaGE

Input of SpaGE algorithm:

- gene expression matrices corresponding to the scRNA-seq data (reference)
- spatial transcriptomics data (query).

Two major steps:

(1) alignment of the two datasets: project both datasets into a common latent space using the gene expression of the shared set of genes.



EX 11 **How many genes in common do we have in this dataset?**

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

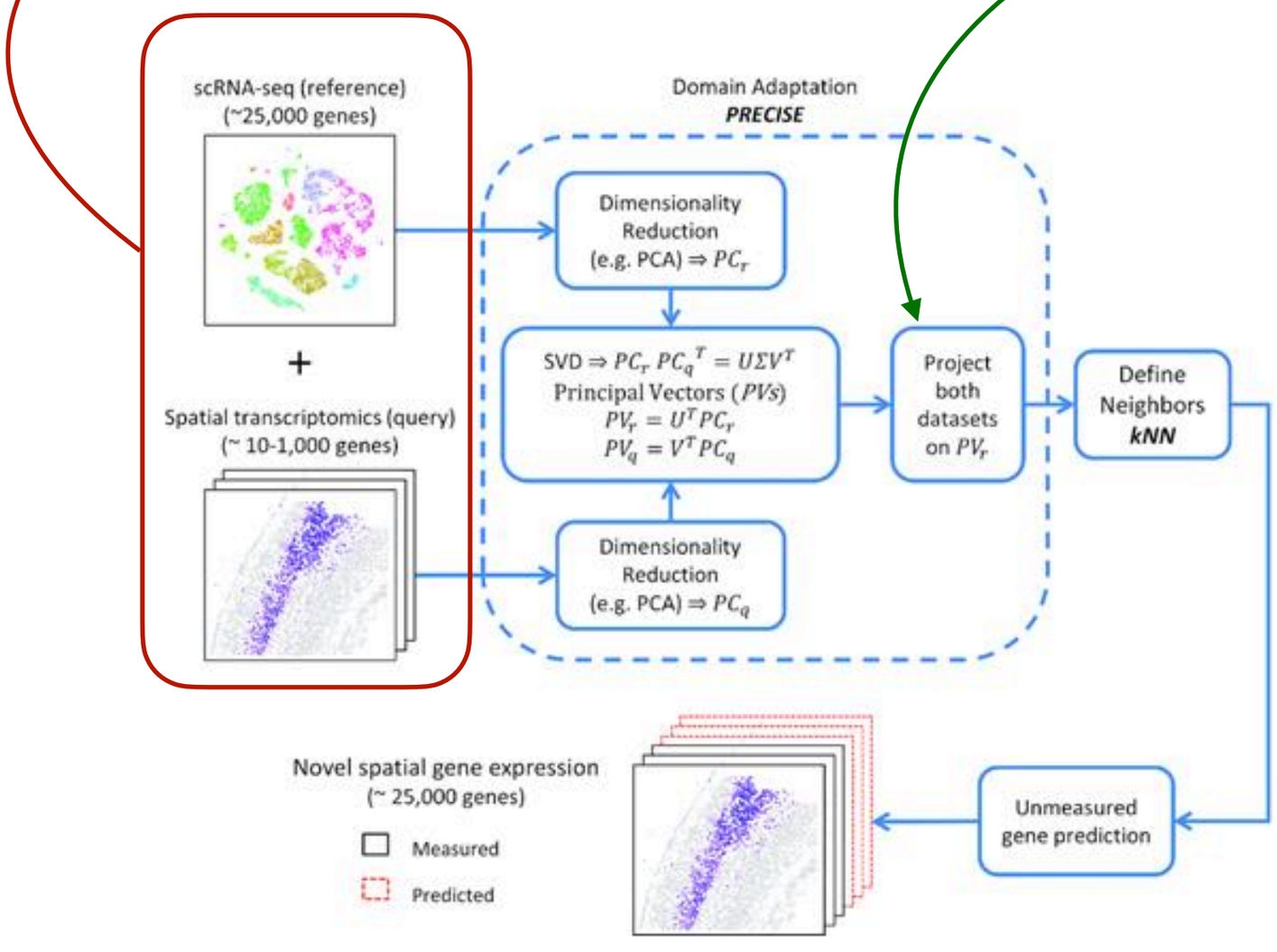
Unmeasured gene imputation with SpaGE

Input of SpaGE algorithm:

- gene expression matrices corresponding to the scRNA-seq data (reference)
- spatial transcriptomics data (query).

Two major steps:

(1) alignment of the two datasets: project both datasets into a common latent space using the gene expression of the shared set of genes.



EX 11 **How many genes in common do we have in this dataset?**

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

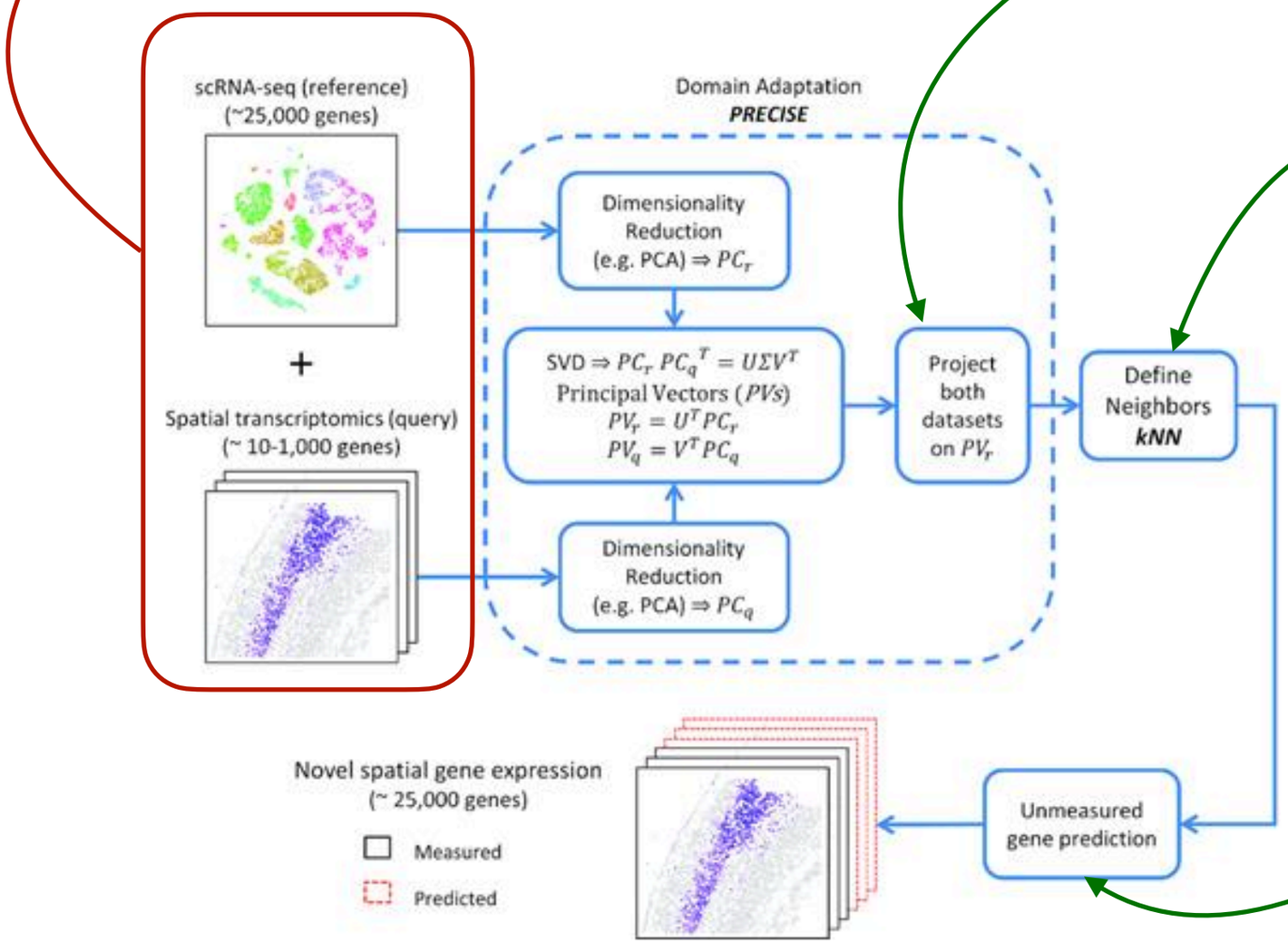
Unmeasured gene imputation with SpaGE

Input of SpaGE algorithm:

- gene expression matrices corresponding to the scRNA-seq data (reference)
- spatial transcriptomics data (query).

Two major steps:

(1) alignment of the two datasets: project both datasets into a common latent space using the gene expression of the shared set of genes.



(2) gene expression prediction using *k*-nearest-neighbor regression: for each spatial cell, define the k-nearest neighbors from the scRNAseq reference cells, using cosine distance.

gene expression of the unmeasured genes = weighted average of the nearest neighbors dissociated cells that have positive cosine similarity.

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with SpaGE

Two modalities:

(1) Leave-one-out cross validation:

Leave out one gene in the shared set (gene_intersection list) and predict its expression; then compare it to the real expression

(2) Predict unmeasured gene expression:

Predict the expression of a gene in the unmeasured spatial genes (genes_not_intersection)

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with SpaGE

Two modalities:

(1) Leave-one-out cross validation:

Leave out one gene in the shared set (gene_intersection list) and predict its expression; then compare it to the real expression

(2) Predict unmeasured gene expression:

Predict the expression of a gene in the unmeasured spatial genes (genes_not_intersection)

→ How can we validate this scenario?

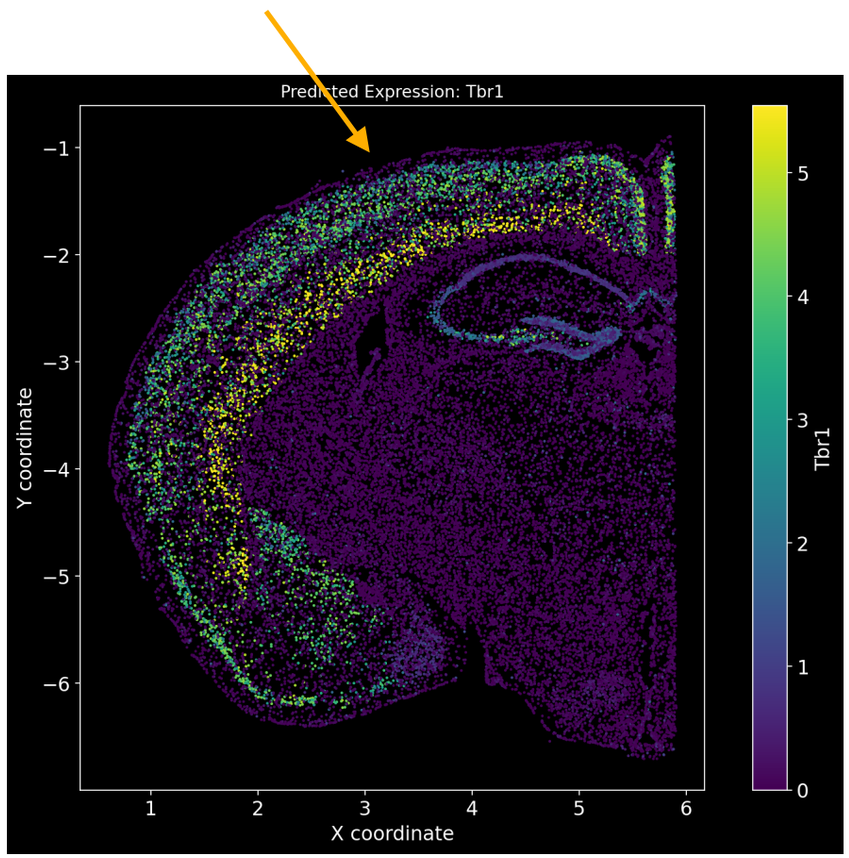
Plot brain region-specific genes and verify whether their expression aligns with the expected brain regions.

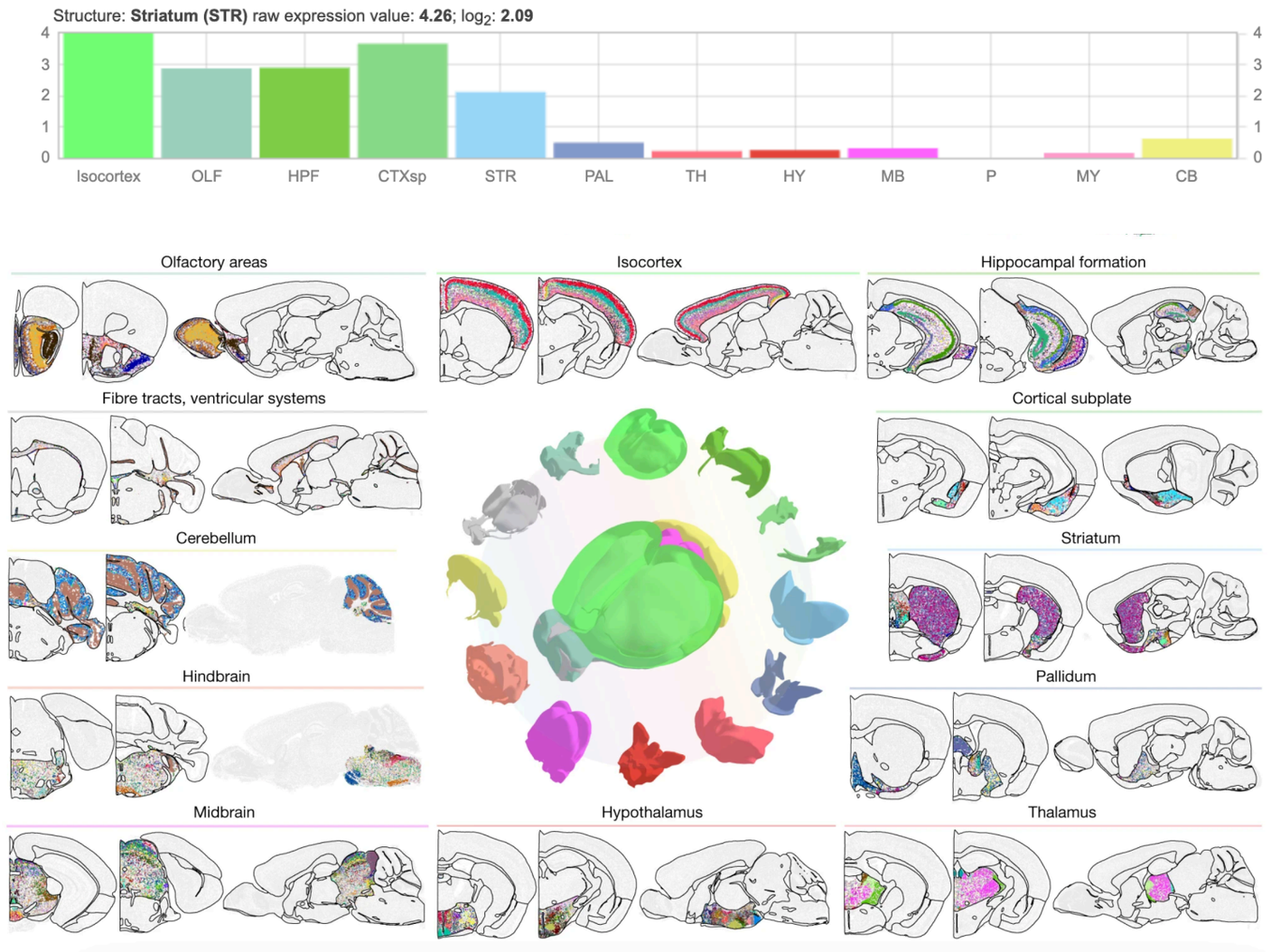
For example, Tbr1 is known to be Isocortex-specific; does its expression show higher levels in the Isocortex? Similarly, Foxp2 is Striatum-specific; does it exhibit a Striatum-specific expression pattern?

SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with SpaGE

Tbr

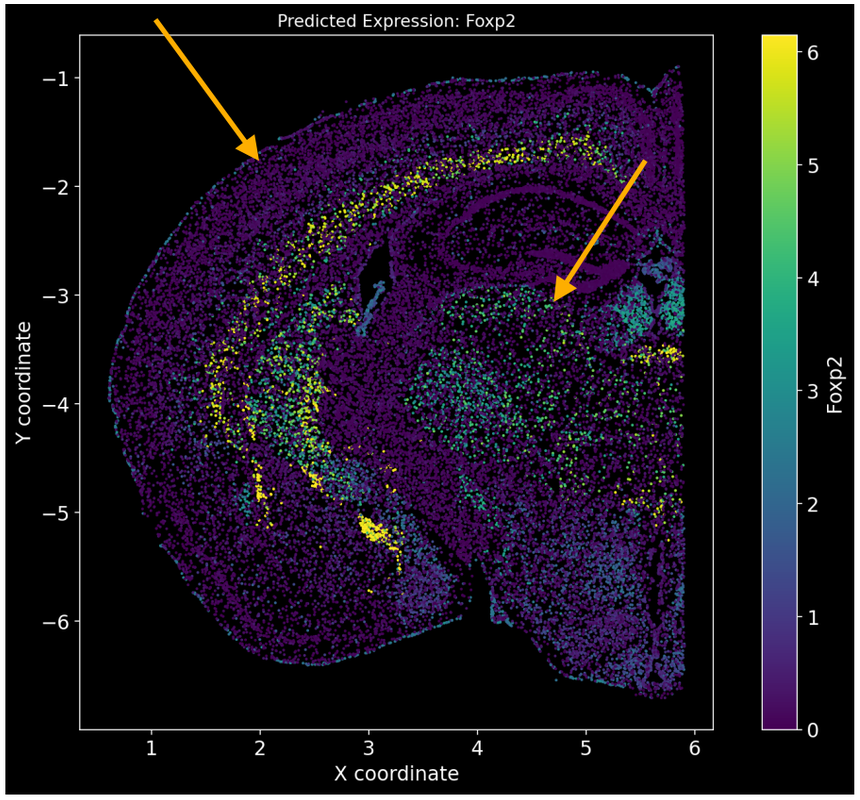


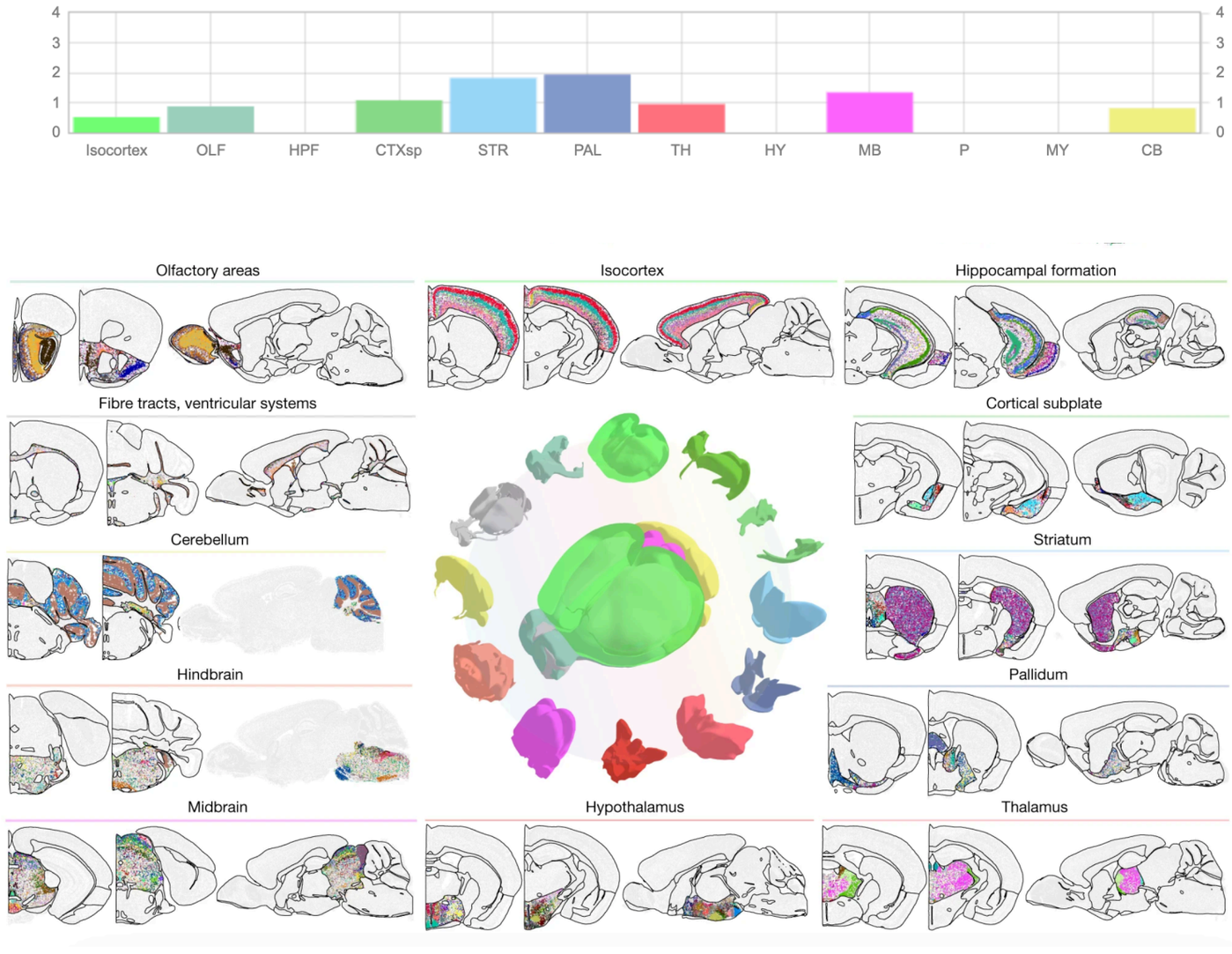


SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with SpaGE

Foxp2



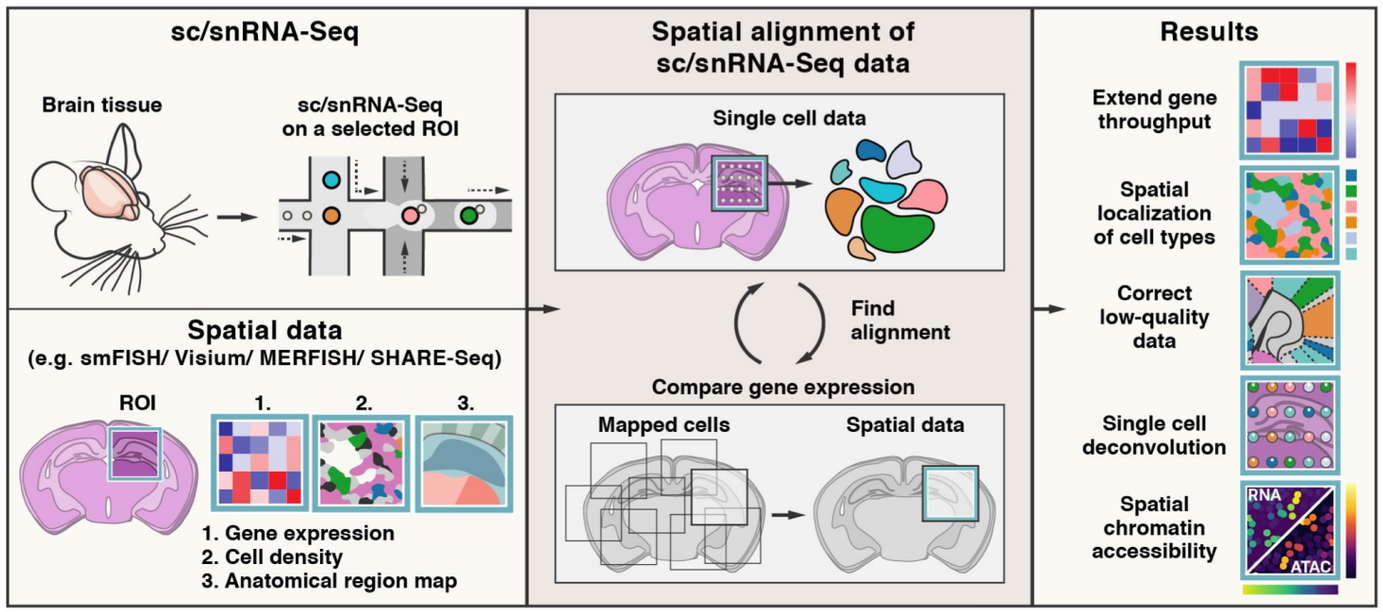


SpaGE: Spatial Gene Enhancement using scRNA-seq (Abdelaal et al. 2020, NAR)

Unmeasured gene imputation with Tangram

Probabilistic method:

1. Learn an alignment between the sc cells and the spatial voxels
2. Learn probability sc i to be in voxel j (`ad_map`)
    ad_map: n sc cells x n spatial voxels
3. Extend the transcriptome-wide expression to the spatial data (`ad_ge`)
    ad_ge: n spatial voxels x n sc genes



Deep learning and alignment of spatially resolved single-cell transcriptomes with Tangram (Biancalani et al. 2021, Nature Methods)

Tangram - SpaGE

Which tool results in higher correlation between the true-measured gene expression and the predicted expression of the genes of interest?

Tangram - SpaGE

Which tool results in higher correlation between the true-measured gene expression and the predicted expression of the genes of interest?

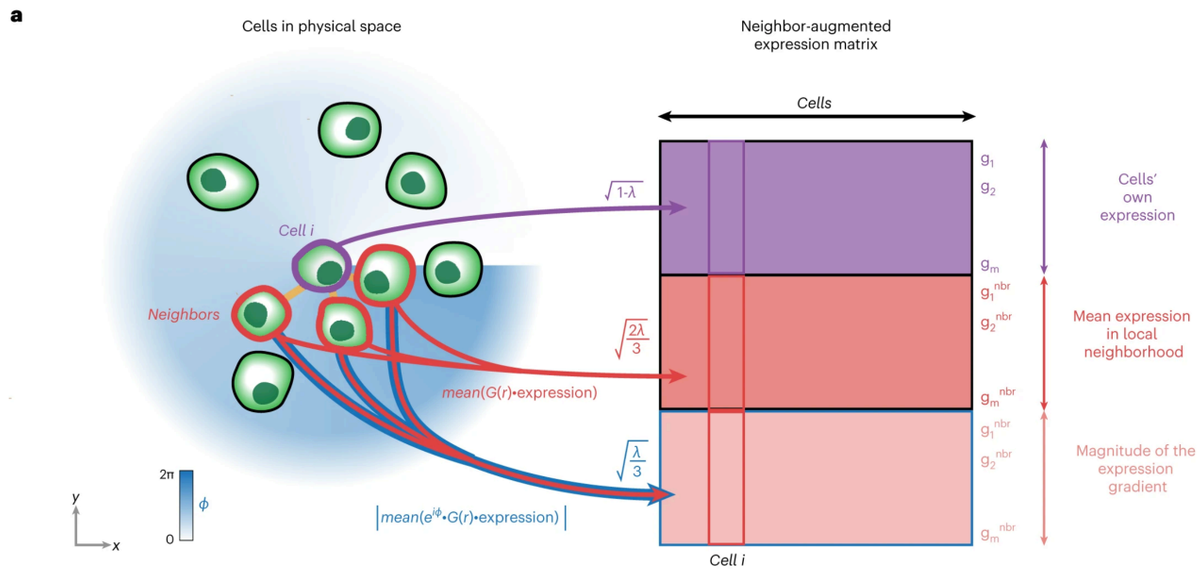
	correlations_spage	correlations_tangram
Satb2	0.569600	0.521339
Penk	0.462047	0.404159
Dlx6	0.402179	0.367113
Cux2	0.623565	0.520271
Zic1	0.560255	0.438475
Bcl11b	0.476320	0.465106
Fezf2	0.427884	0.363526

Spatial clustering with Banksy

Steps:

1. From the original count matrix (matrix **C**):

    a. Construct a neighborhood graph between cells in physical space using *k*-nearest neighbors (Matrix **M**) and compute the **mean expression** in the neighborhod of a cell



BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

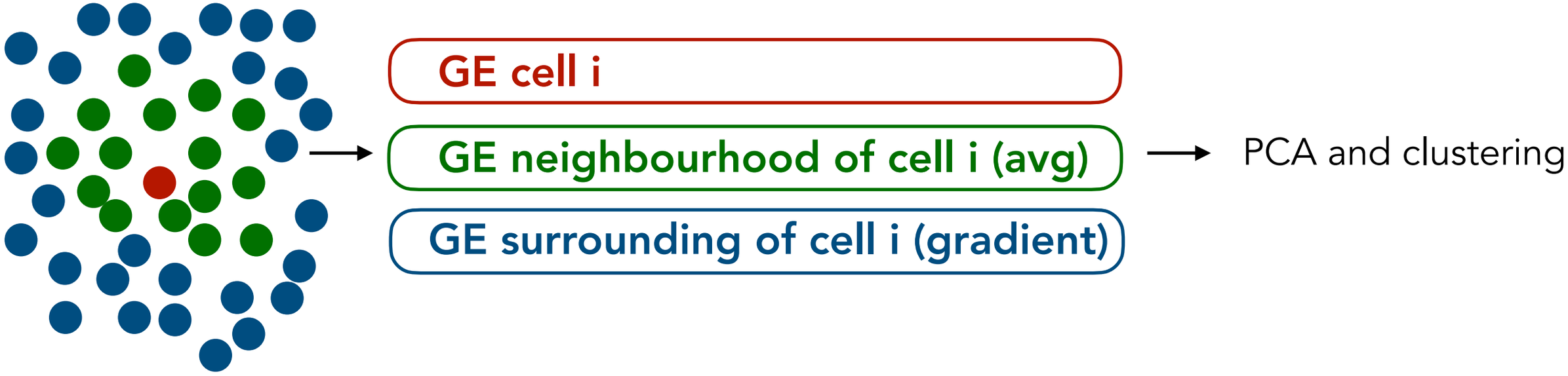
Spatial clustering with Banksy

Steps:

1. From the original count matrix (matrix **C**):

   a. Construct a neighborhood graph between cells in physical space using *k*-nearest neighbors (Matrix **M**) and compute the **mean expression** in the neighborhod of a cell

   b. AGF: measures gradients in gene expression in each neighborhood (2*k). (Matrix **G**)

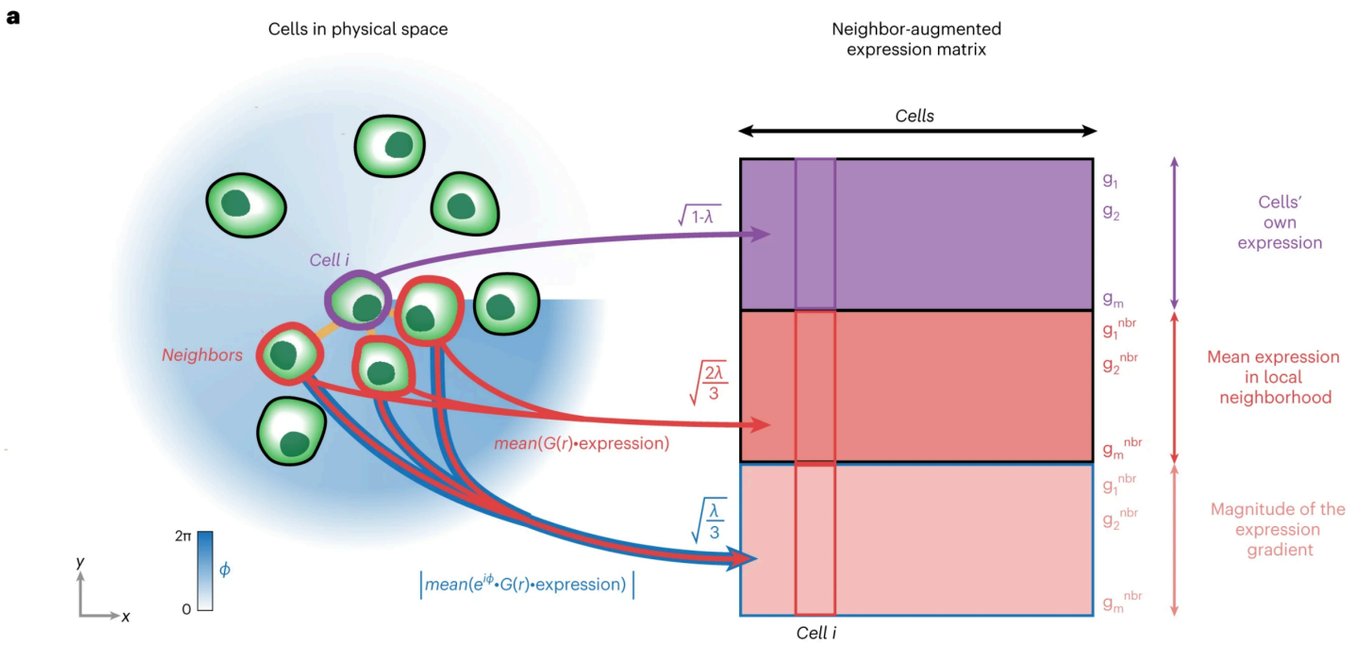


BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

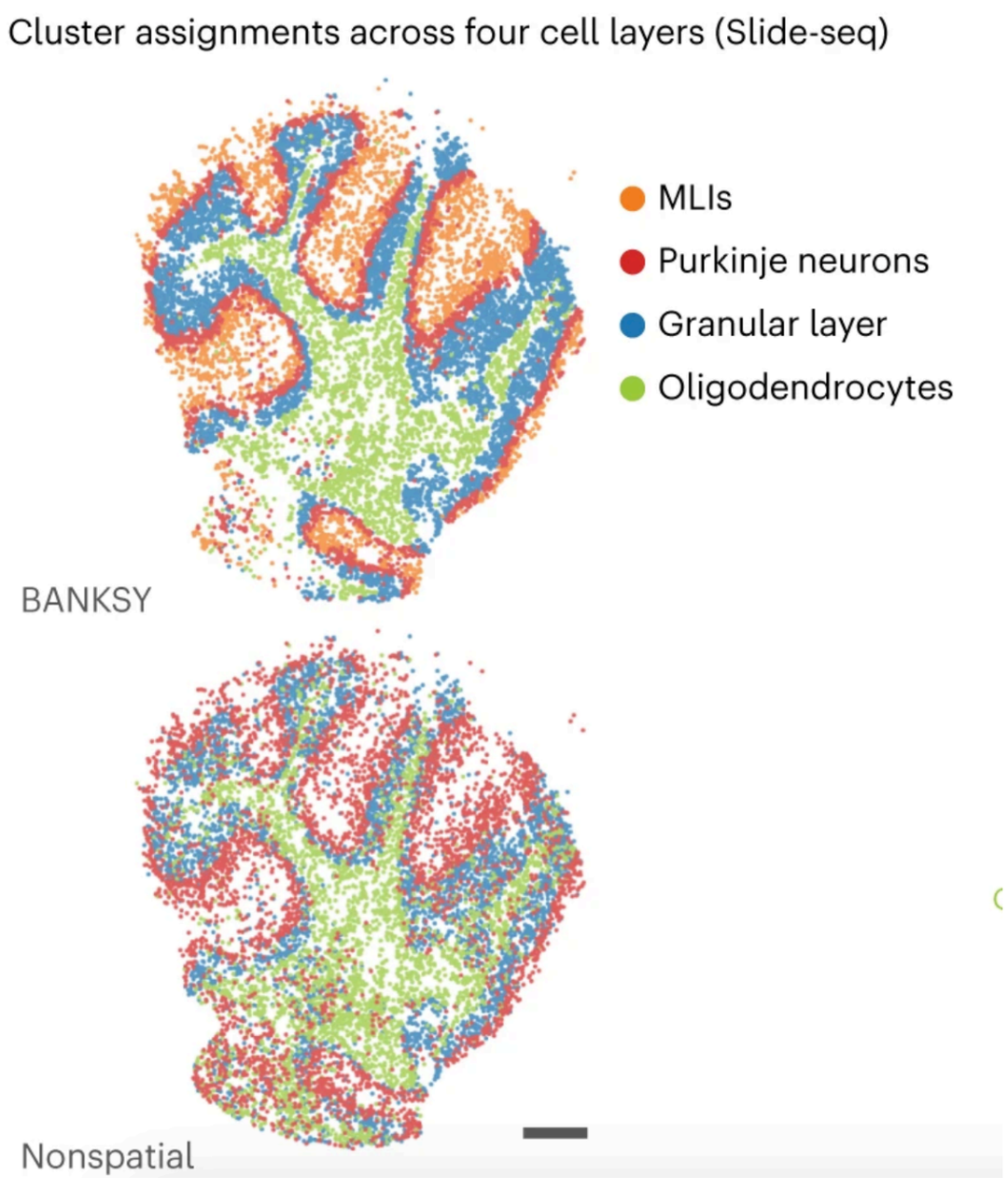
Steps:

1. From the original count matrix (matrix **C**):
   a. Construct a neighborhood graph between cells in physical space using *k*-nearest neighbors (Matrix **M**) and compute the **mean expression** in the neighborhod of a cell
   b. AGF: measures gradients in gene expression in each neighborhood. (Matrix **G**)
2. Matrices M and G are **scaled based on a mixing parameter λ**, which controls their relative weighting.
3. **Concatenate** matrices M, G and C
4. PCA
5. Graph-based clustering



BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

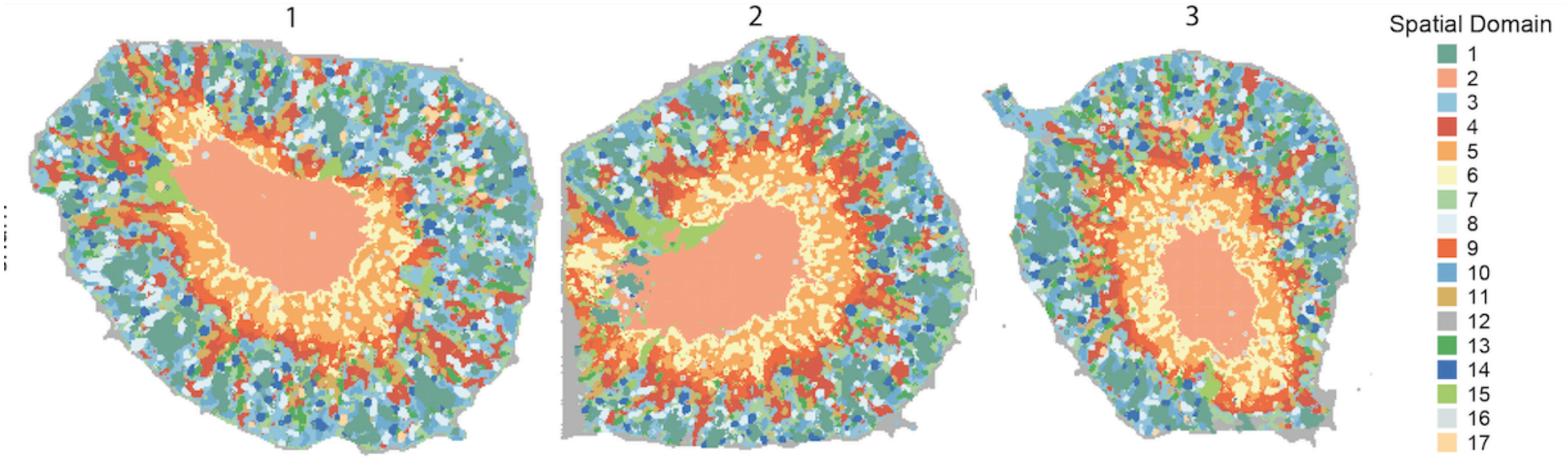
Spatial clustering with Banksy



Spatial map of four cell type clusters in the Slide-seq dataset (granular layer, Purkinje neurons, MLIs and the oligodendrocytes and polydendrocytes cluster). The MLI cluster is absent in nonspatial clustering ($\lambda = 0$)

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy



Same spatial domains at different locations in the tissue will have same transcriptomics profile but also similar (cellular) neighbourhood and similar SD neighborhood

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

Step by step

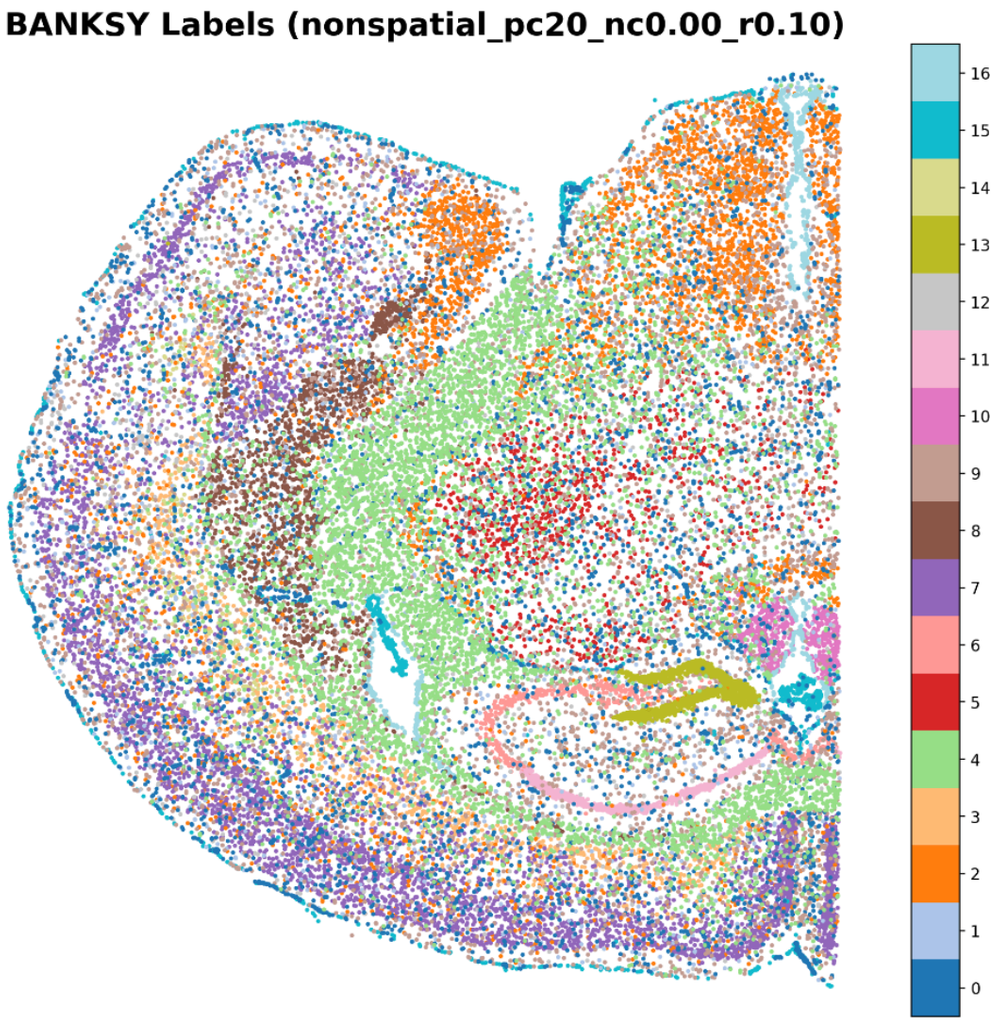
1. Normalise data
2. Construction of k-nearest-neighbors based on physical distance
3. Generate spatial weights from distance
4. Generate Banksy matrix
    A. Matrix multiply sparse CSR weights matrix with cell-gene matrix (if AGF true)
    B. Z-score both matrices along genes
    C. Multiply each matrix by a weighting factor lambda (neighbourhood contribution parameter)
    D. Concatenate the matrices (along the genes dimension)
5. Run PCA
6. Run Leiden

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

lambda=0

lambda=0.6

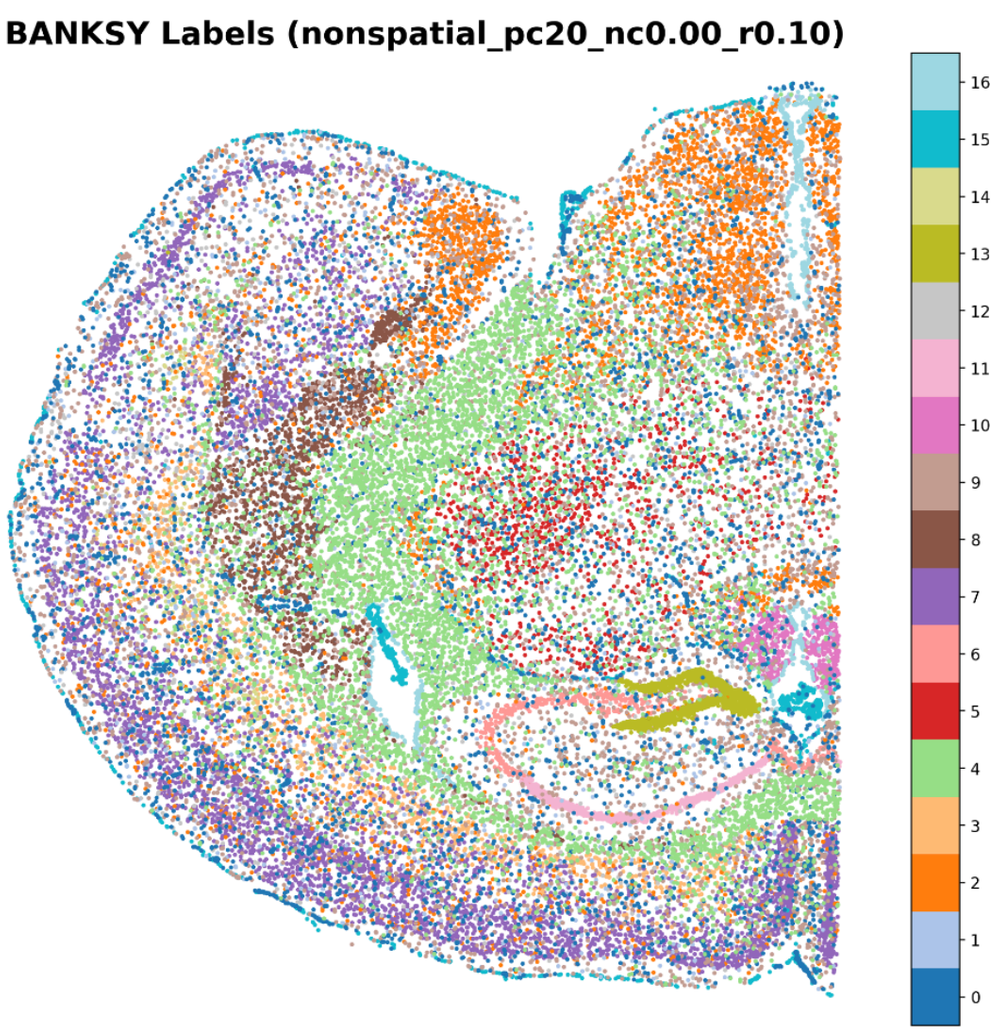


What do you think the spatial plot will look like when lambda=1?

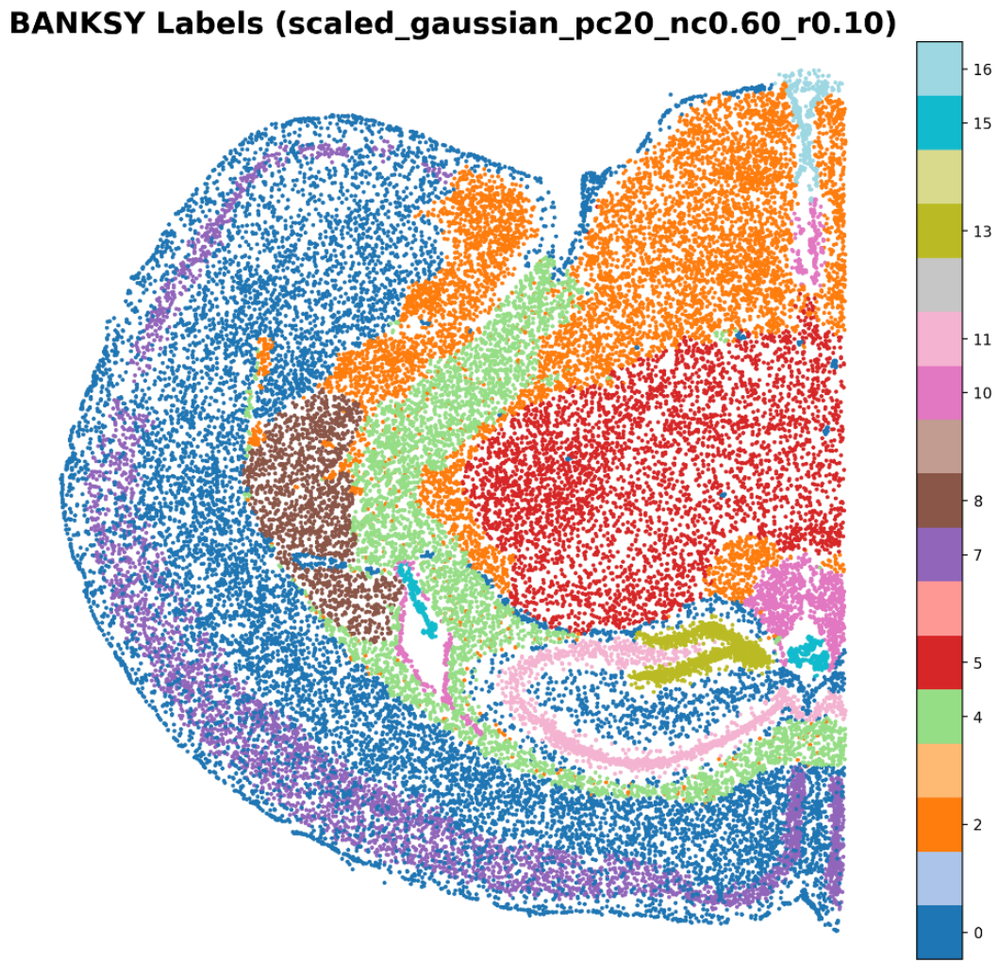
BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

lambda=0



lambda=0.6



What do you think the spatial plot will look like when lambda=1?

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

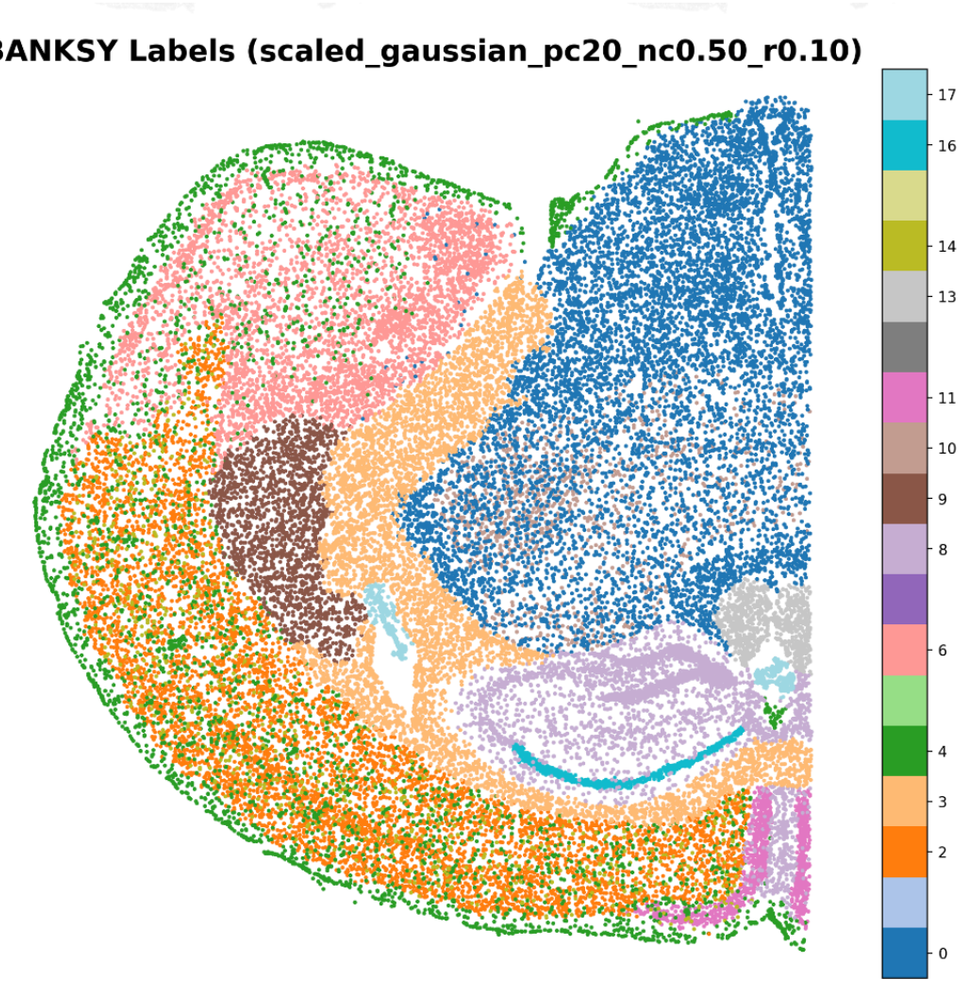
What will happen when you increase k geom? What is an optimal k?

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

EX 12

What will happen when you increase k geom? What is an optimal k?



k=80

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

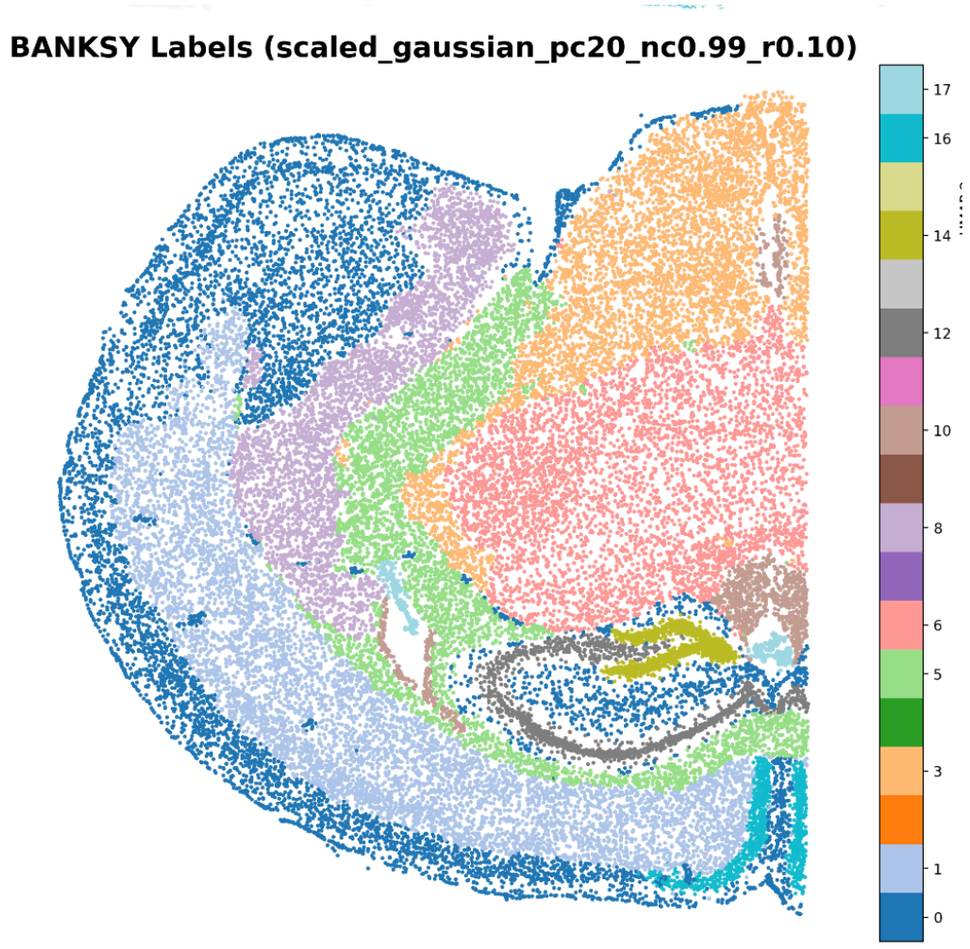
What do you think the spatial plot will look like when lambda=1?

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

EX 13

What do you think the spatial plot will look like when lambda=1?



lambda=1

More “net" clusters

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

EX 15

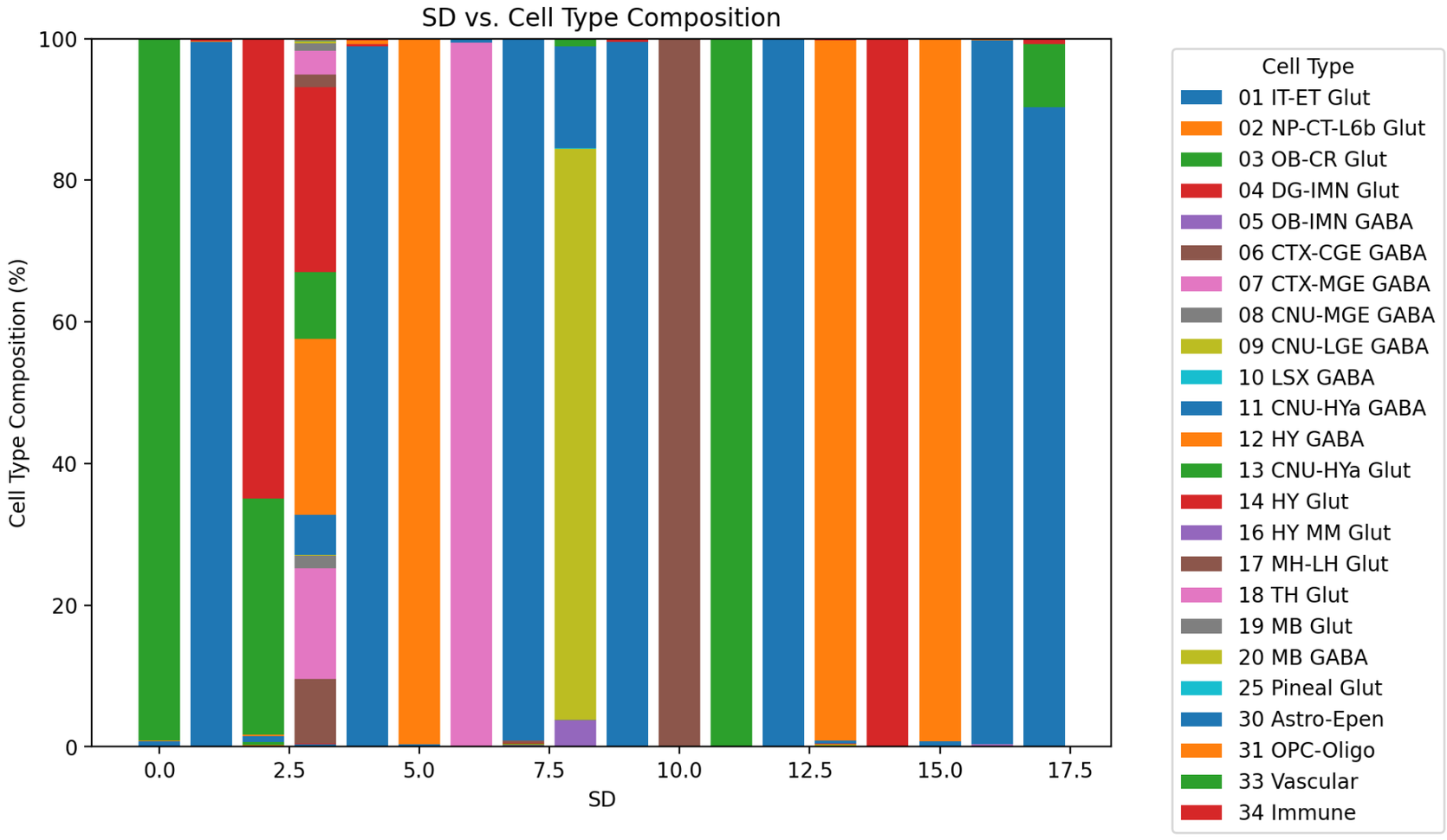
How do you expect the Cell Type composition plot to look like when lambda=0?

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)

Spatial clustering with Banksy

EX 15

How do you expect the Cell Type composition plot to look like when lambda=0?



lambda=0

Spatial Domains resemble Cell Types

BANKSY unifies cell typing and tissue domain segmentation for scalable spatial omics data analysis (Signal et al. 2024, Nature Genetics)