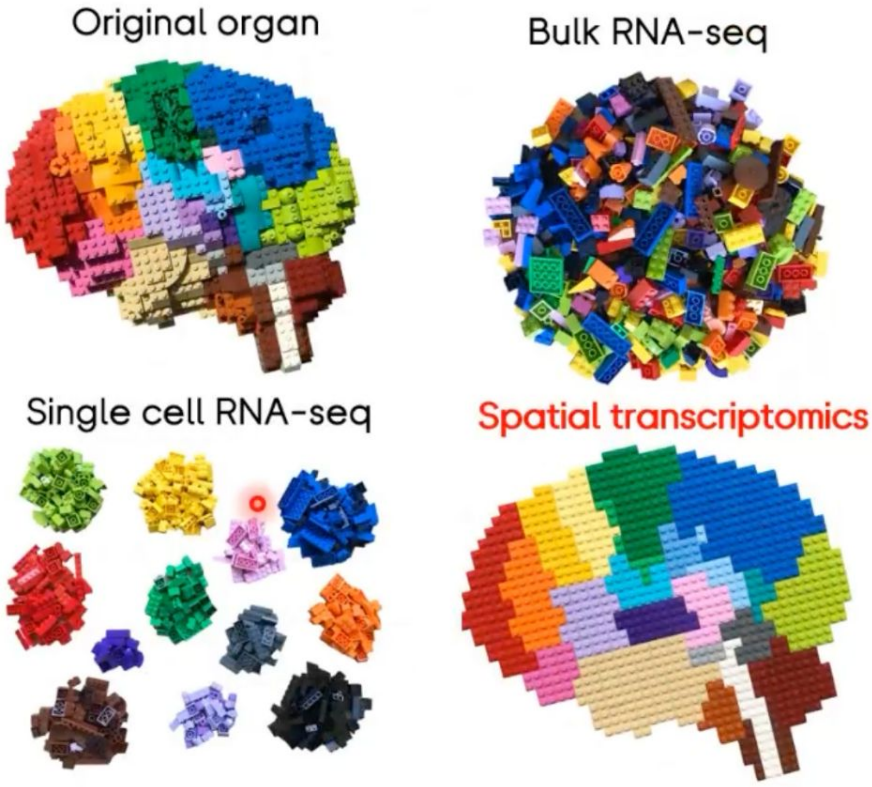
Genome Informatics 2024

Lesson 7 - Spatial transcriptomics

Method of the year 2019/2020:
Single-cell multi-omics and spatially resolved transcriptomics



Images and concept courtesy of Bo Xia, NYU School of Medicine

Editorial | Published: 06 January 2020

Method of the Year 2019: Single-cell multimodal omics

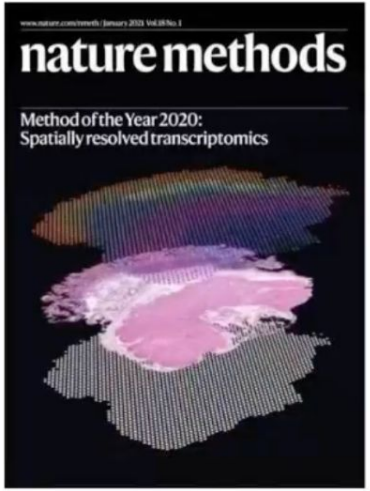
Nature Methods **17**, 1 (2020) | Cite this article

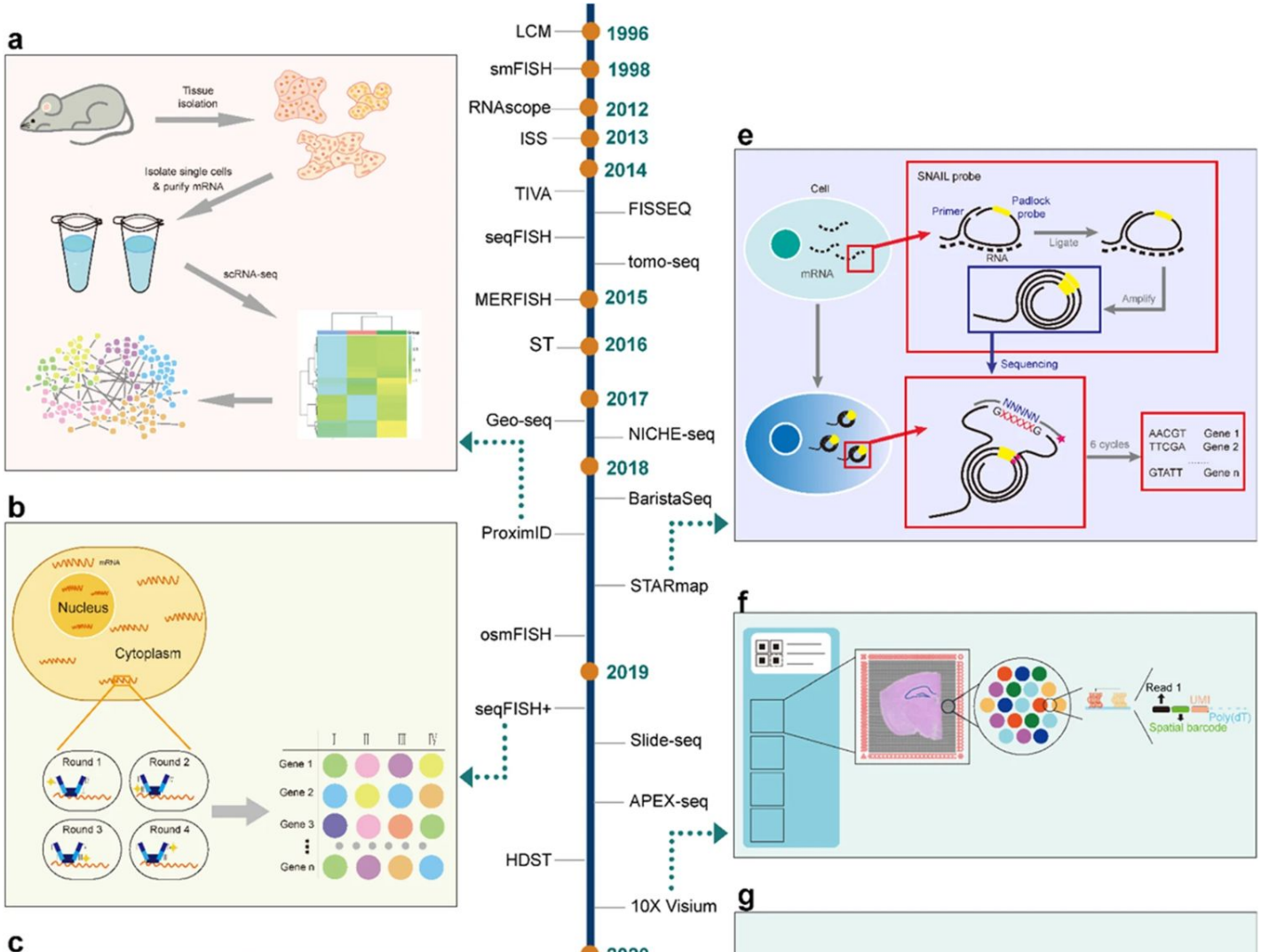
32k Accesses | **128** Altmetric | Metrics

Method of the Year 2020: spatially resolved transcriptomics

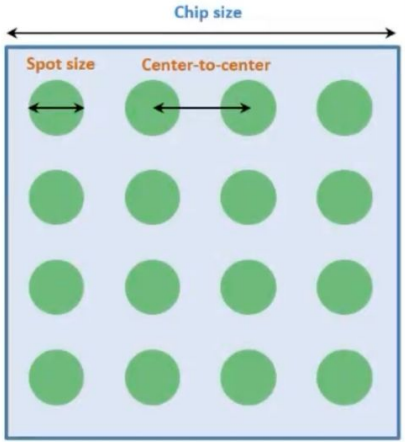
Nature Methods **18**, 1(2021) | Cite this article

13k Accesses | **221** Altmetric | Metrics



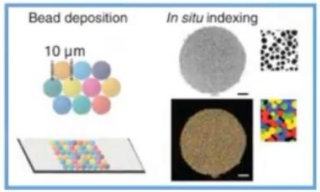

a
Tissue isolation
Isolate single cells & purify mRNA
scRNA-seq
b
mRNA
Nucleus
Cytoplasm
Round 1
Round 2
Round 3
Round 4
Gene 1
Gene 2
Gene 3
Gene n
c
LCM 1996
smFISH 1998
RNAscope 2012
ISS 2013
2014
TIVA
FISSEQ
seqFISH
tomo-seq
MERFISH 2015
ST 2016
2017
Geo-seq
NICHE-seq
2018
BaristaSeq
ProximID
STARmap
osmFISH
2019
seqFISH+
Slide-seq
APEX-seq
HDST
10X Visium
e
Cell
mRNA
SNAIL probe
Primer
Padlock probe
RNA
Ligate
Amplify
Sequencing
NNNNN
GXXXXXG
6 cycles
AACGT Gene 1
TTCGA Gene 2
GTATT Gene n
f
Read 1
UMI
Poly(dT)
Spatial barcode
g

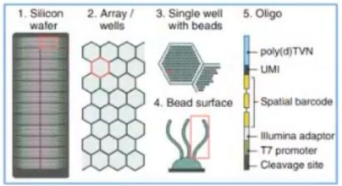
Stereo-Seq

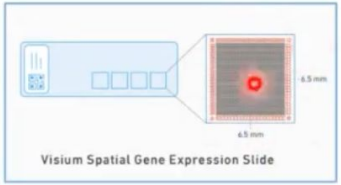


Challenges:
- Resolution
- Capture area

Each spot is 220 nm in diameter and the center-to-center distance between neighboring spots is 500 nm (Stereo-seq)





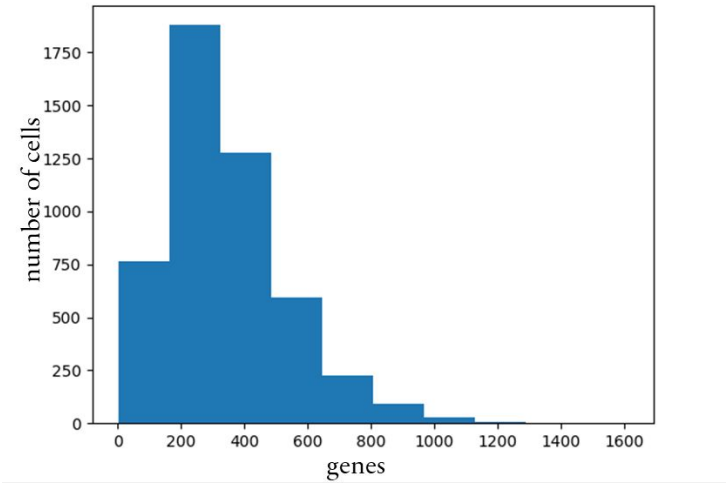


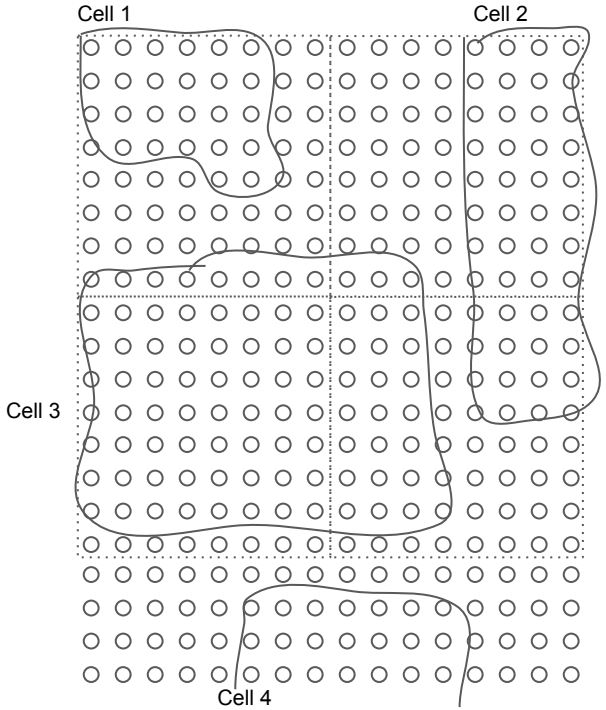
	DBiT-seq	Slide-seq	HDST	Visium	Stereo-seq v1
Spot size (µm)	10	10	2	55	0.22
Center-to-center (µm)	20	10	2	100	0.5
Field of view (mm)	1.0 x 1.0	Φ 3.0	5.7 x 2.4	6.5 x 6.5	132 x 132
Gene number (mean, 100 µm diam)	3,302	4,008	80	1,813	12,661
UMI number (mean, 100 µm diam)	7,604	4,650	100	3,276	133,776

* 1x1 cm slice of Stereo-Seq data costs ~$2000

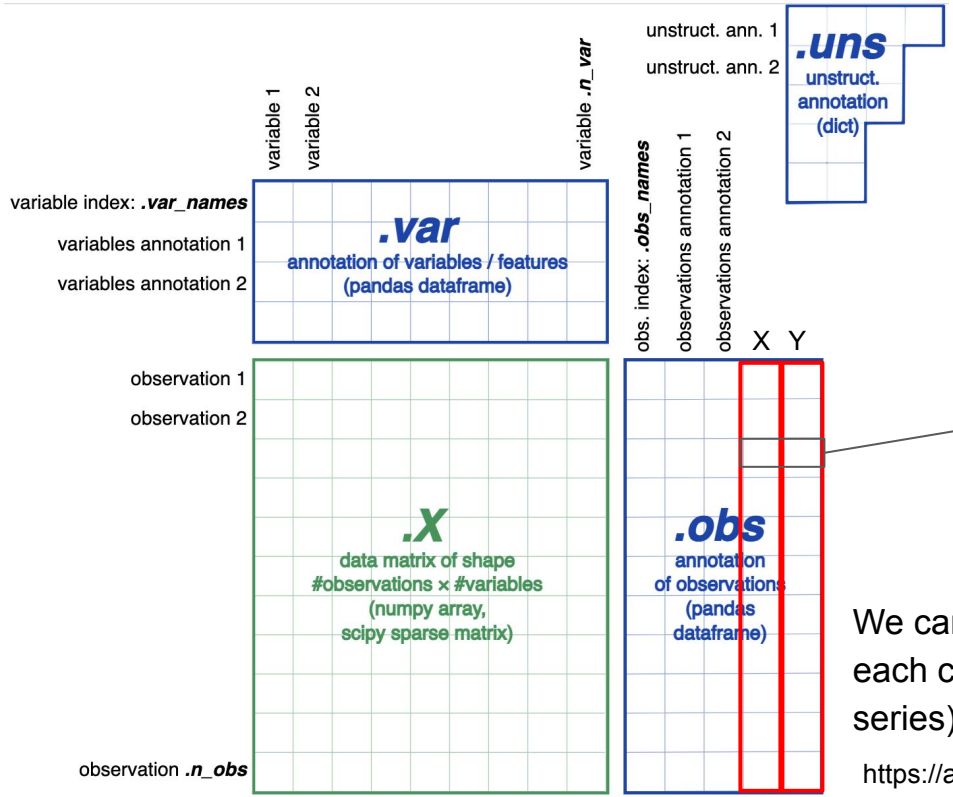
Spatial transcriptomics

- Spatial coordinates for every cell + its gene expression

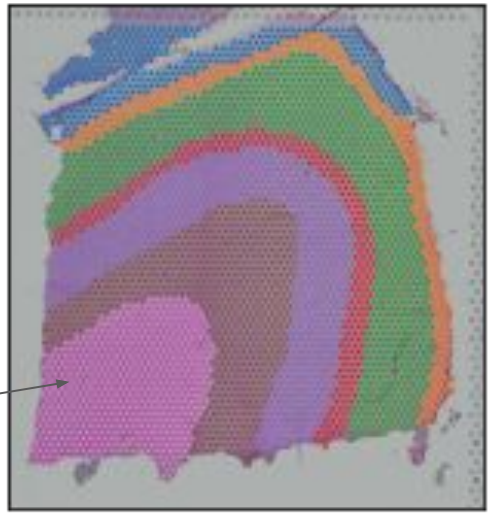




Annotated data object



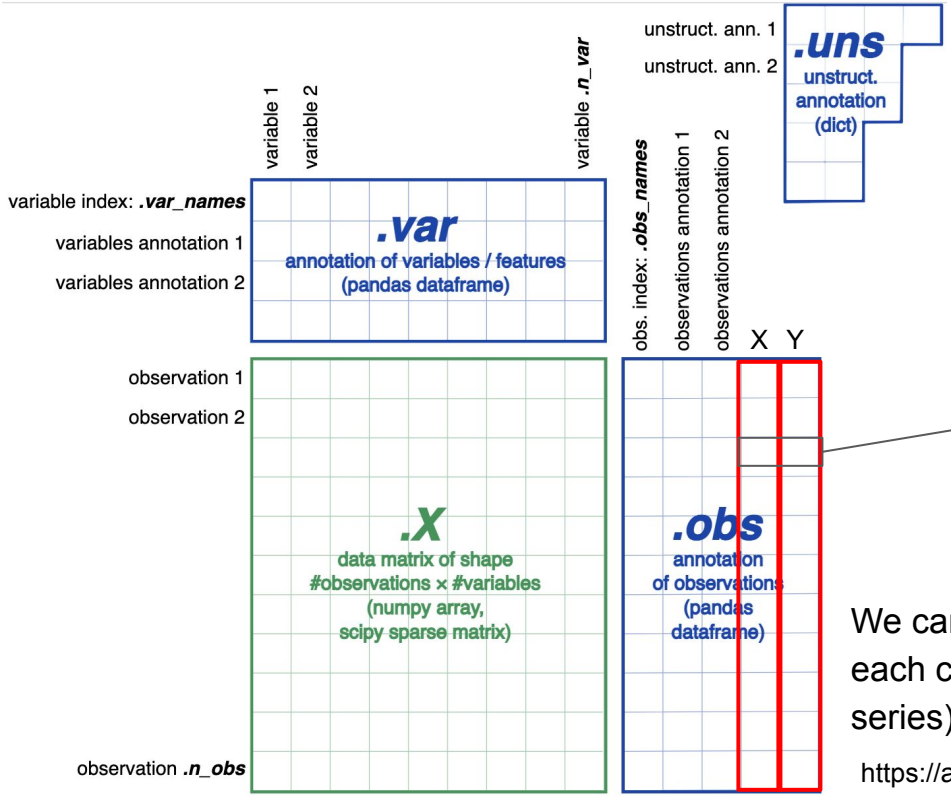
DLPFC tissue blocks



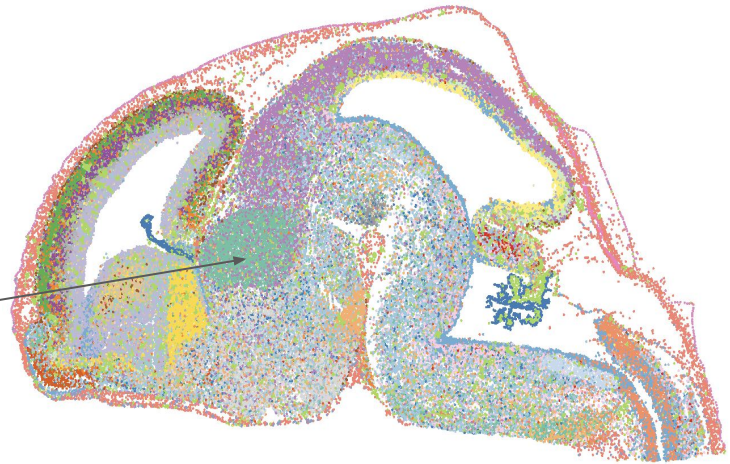
We can fully reconstruct the position of each cell in the tissue (even 3D or time series)

https://anndata-tutorials.readthedocs.io/en/latest/getting-started.html

Annotated data object



Mouse embryo brain, day 16

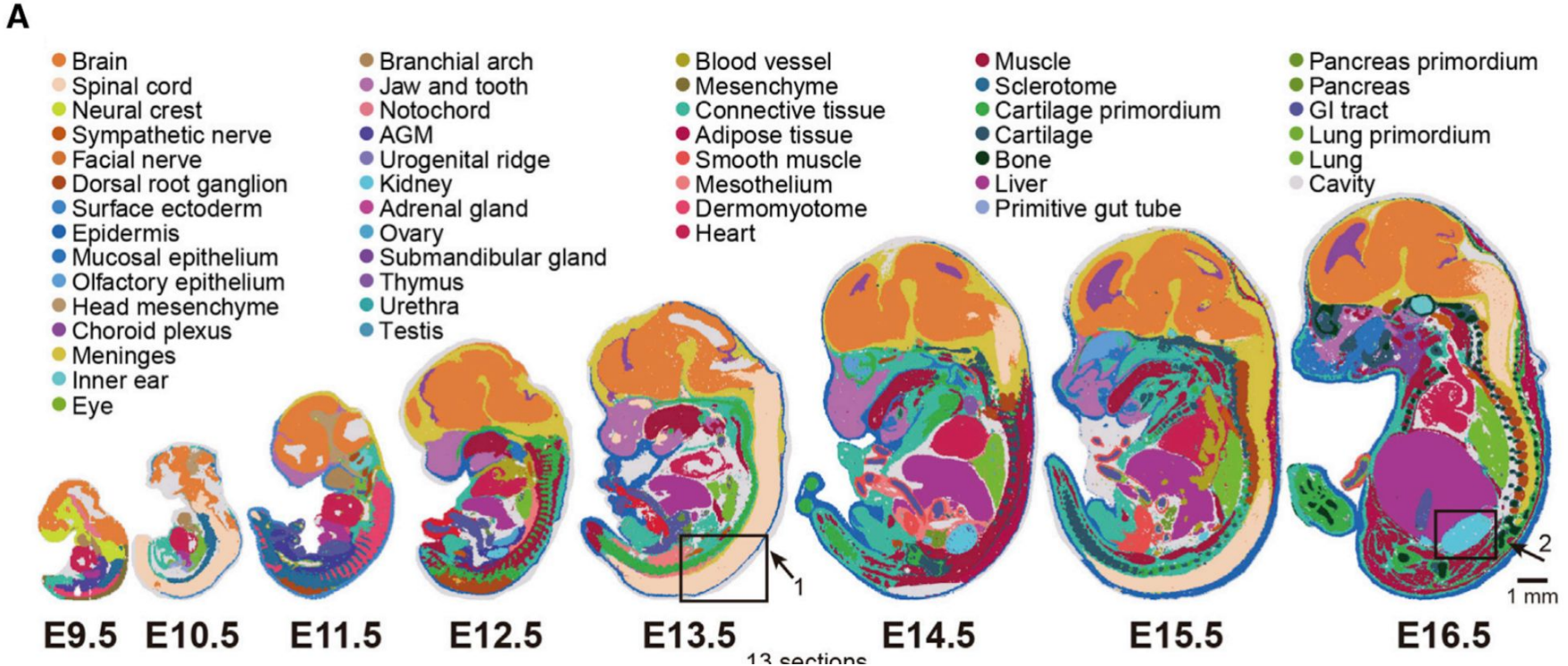


Cere Gran NeuB 216
ChP 658
Corti Glu Neu 1,461
Corti prog 1,305
CortiHippo Glu Neu 1,180
CR 515
Cranium Fibro 202
DA Neu 334
Die GNeu 4,250
DorsHb RGC 1,452
Endo 647
Epidermis 488
Ery

We can fully reconstruct the position of each cell in the tissue (even 3D or time series)

https://anndata-tutorials.readthedocs.io/en/latest/getting-started.html

Stereo-Seq: Spatiotemporal transcriptomic atlas of mouse organogenesis



Source paper. Mosta database

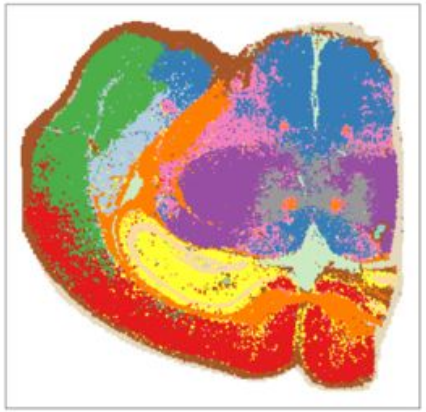
Stereopy library

Spot: 5K
Cell: not single cell resolution

SEURAT R toolkit for single cell genomics

scanpy

Spot: 0.4 – 67.6 billions
Cell: 10,000 to more than one million



- Computation power
- Data storage and IO

- Visualization functionality
- New coding and algorithms

- Stereopy library
- Quick start tutorial

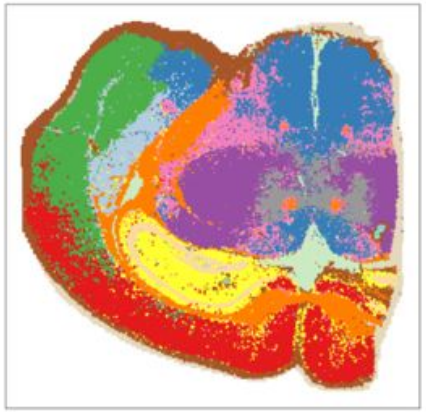
Stereopy library

Spot: 5K
Cell: not single cell resolution

SEURAT R toolkit for single cell genomics

scanpy

Spot: 0.4 – 67.6 billions
Cell: 10,000 to more than one million



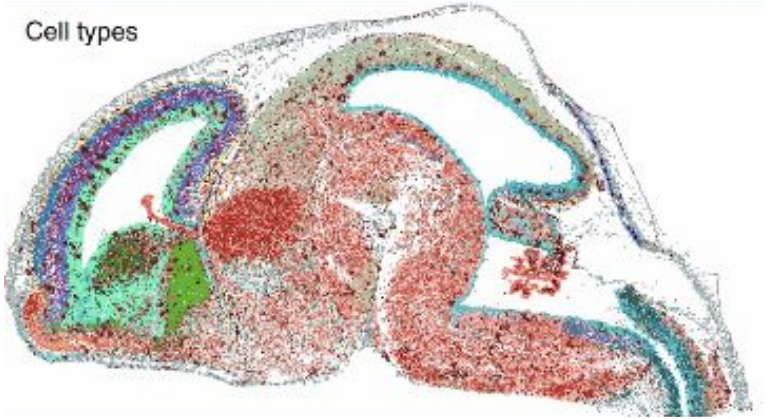
- Computation power
- Data storage and IO

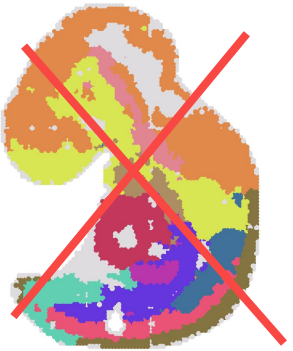
- Visualization functionality
- New coding and algorithms

- Stereopy library
- Quick start tutorial

Cell community detection

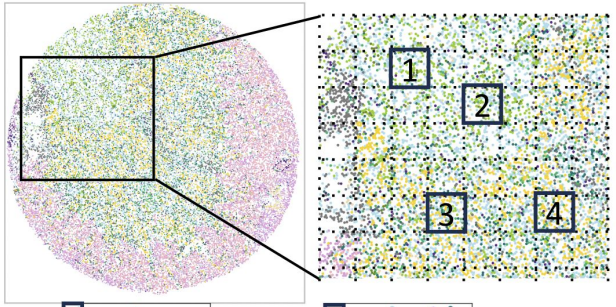
- Get new insight to scientists and biologist about the tissue and its functions based on cell neighborhood
- Detect spatial domains (communities) containing similar percentage of cell types in all of its parts (cell types that co-occur)
- The novelty: Better separability of domains than the existing solutions (SpaGCN, GIOTTO)
- Candidate for Cell Community Detection: Tissue with cell types spread across different parts of it



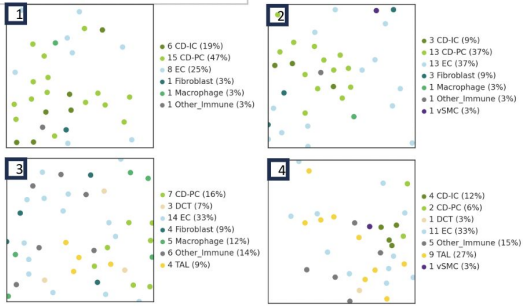


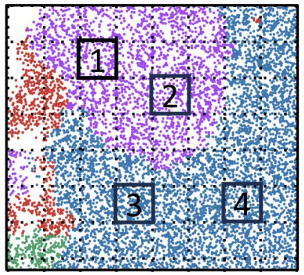
Cell community detection - Under the hood

1. Single size (or multiple size) sliding windows (w) are moved through the surface of the tissue and for each a feature vector is calculated.
   Feature vector contains percentages (p1, p2, ..., pn) of each cell type inside of a window.
2. Feature vectors of all windows and tissue slices are aggregated and clustered using Leiden (or Spectral or Agglomerative) clustering, providing community labels per window
3. Community label of each cell (cell-spot) is defined by majority voting: label with highest occurrence from all windows overlapping the cell-spot is chosen as result



Win	CD-IC	CD-PC	EC	TAL	…	Label
1	19	47	25	0	…	
2	9	37	37	0	…	
3	0	16	33	9	…	
4	0	6	33	27	…	
…	…	…	…	..	…	

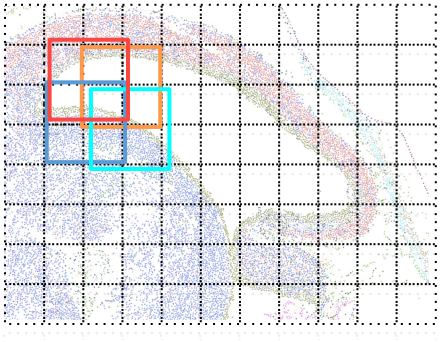




Methods - sliding window & majority voting

- Window's sliding step can be smaller than window size
- One cell can be covered by more than one window (e. g. if sliding step is ½ of window size it will be covered by 4 windows that might have different labels)
- Community label for each cell is determined by majority voting

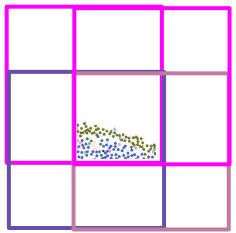
Create cell type window mixture



Cluster cell type mixtures

Type 1	Type 2	Type3	…	Type K	Label
51%	30%	0%		4%	
0%	0%	95%		0%	
0%	55%	5%		12%	
0%	0%	94%		0%	
49%	29%	1%	…	3%	

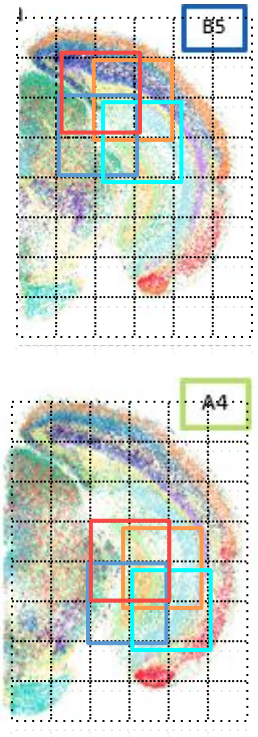
Community calling - majority voting



Methods - Multi-slice processing

- Extract cell-type mixtures from all slices from all windows
- Cluster them all together

Create cell type window mixture



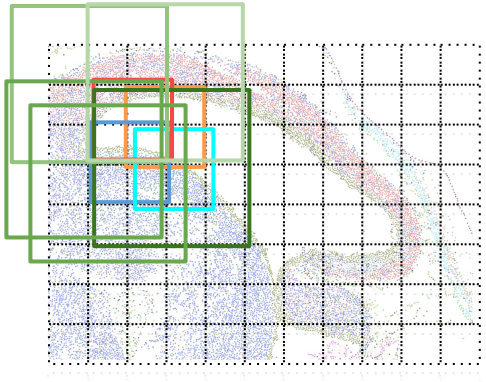
Cluster cell type mixtures

Type 1	Type 2	Type3	...	Type K	Label
51%	30%	0%		4%	
0%	0%	95%		0%	
0%	55%	5%		12%	
0%	0%	91%		0%	
...					
Type 1	Type 2	Type3	...	Type K	Label
51%	30%	0%		4%	
0%	0%	95%		0%	
0%	55%	5%		12%	
0%	0%	91%		0%	

Methods - Multi window size processing

- Window's sliding step can be smaller than window size
- One cell can be covered by more than one window (e. g. if sliding step is ½ of window size it will be covered by 4 windows that might have different labels)
- Community label for each cell is determined by majority voting

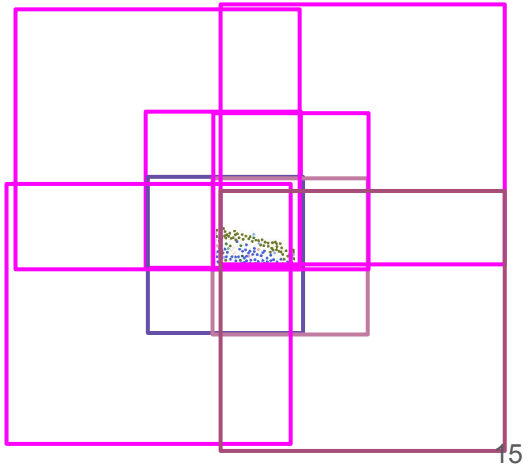
Create cell type window mixture



Cluster cell type mixtures

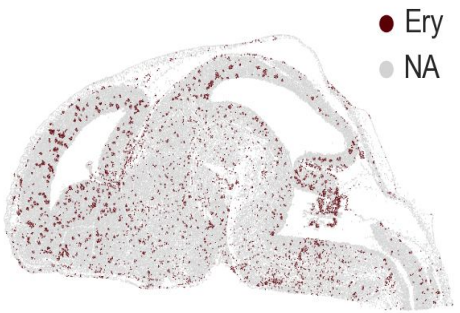
Type 1	Type 2	Type3	…	Type K	Label
51%	30%	0%	…	4%	
0%	0%	95%	…	0%	
0%	55%	5%	…	12%	
0%	0%	94%	…	0%	
49%	29%	1%	…	3%	
51%	30%	0%	…	4%	
0%	0%	95%	…	0%	
0%	55%	5%	…	12%	
0%	0%	94%	…	0%	
49%	29%	1%	…	3%	

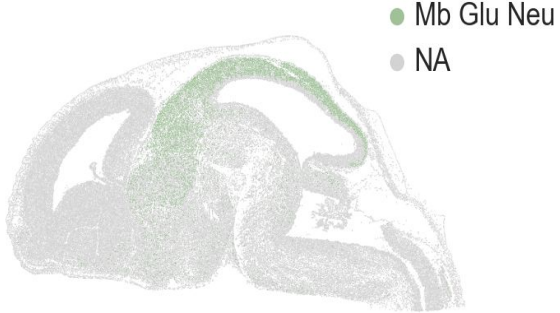
Community calling - majority voting



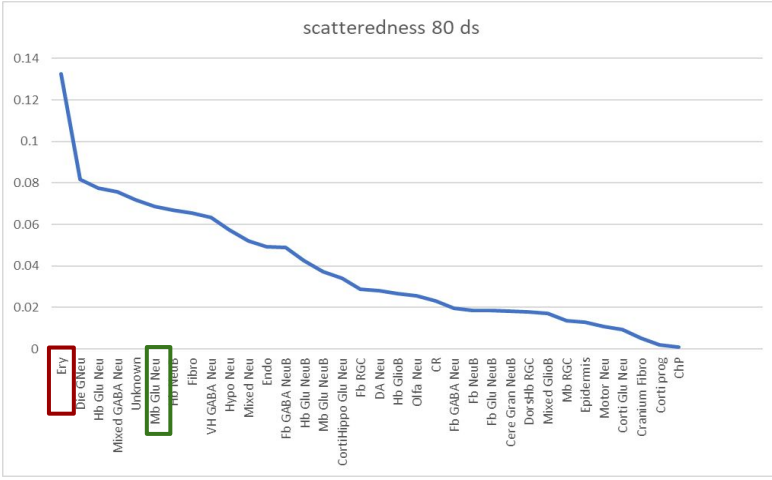
Methods - Preprocessing

- Cell types can be removed from feature vectors based on entropy and scatteredness



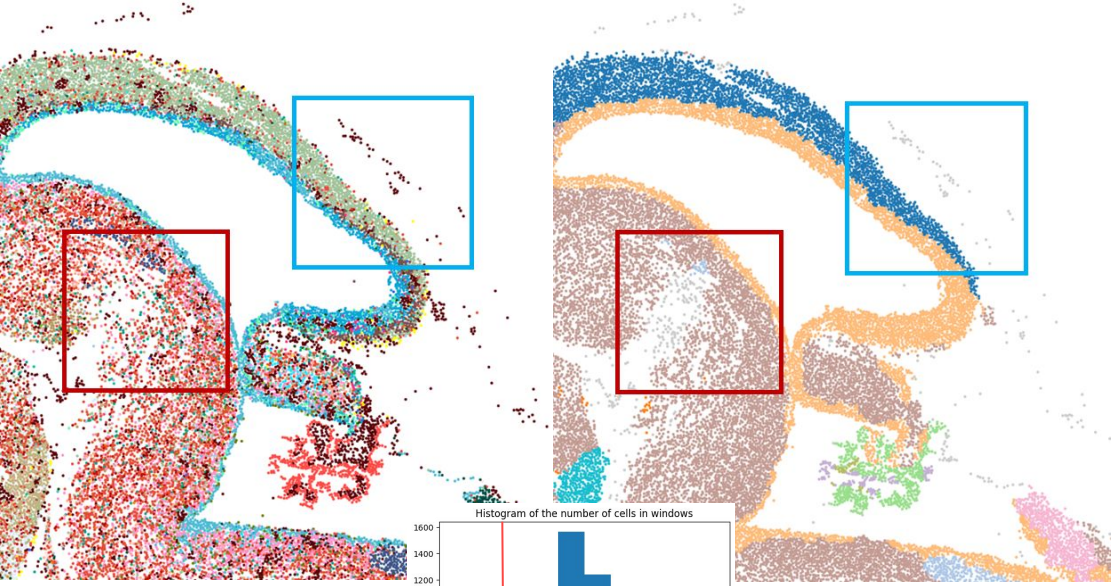


	moransI	entropy
Mixed Neu	0.918228	0.001047
Mb Glu Neu	0.928924	0.000738
Ery	0.77811	0.000659
Fb RGC	0.948998	0.000522
Fibro	1.007095	0.000501
Die GNeu	0.923188	0.0005
Hypo Neu	0.918464	0.000376
Hb NeuB	0.867907	0.000332

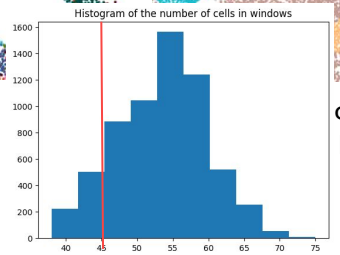


Methods - Preprocessing

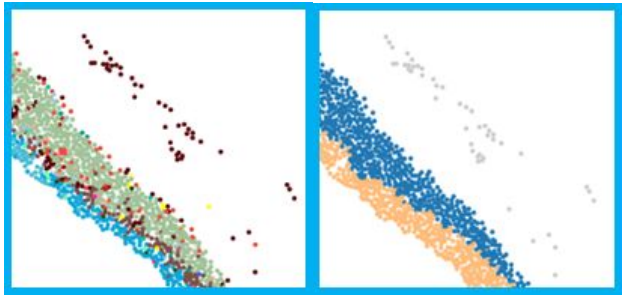
- Windows containing significantly lower number of cells than average are removed from clustering (removed clustering artifacts)

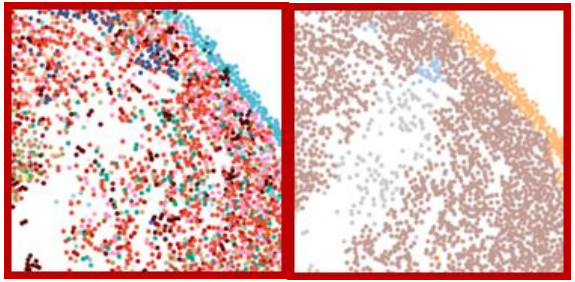


cell type annotation



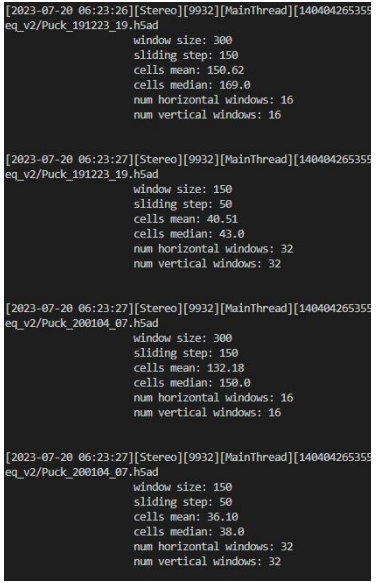
cluster labels





Methods - Estimating the optimal window size

- Optimal window size calculation algorithm
  - Minimum of range on x and y axis is divided by 100 to get initial window size
  - The size is iteratively modified to obtain 30-50 cells per window on average
- Enables users to avoid specifying window size and sliding step
- Gives a good starting point
- Whole mouse brain:
  - window size = 192
  - sliding step = 96
  - mean number of cells = 35.33
  - median number of cells = 39
  - number of horizontal windows = 72
  - number of vertical windows = 29

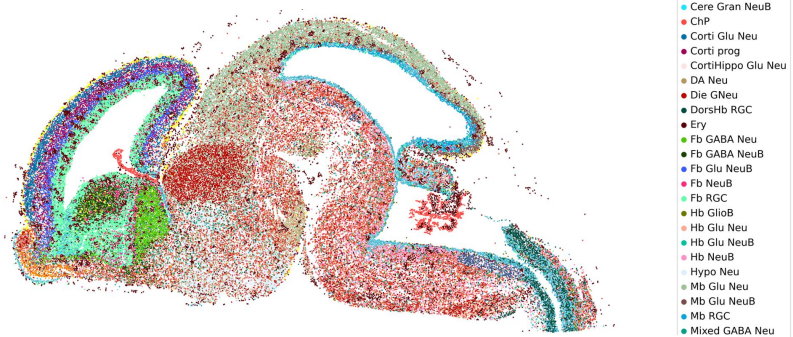


Mouse kidney samples 4

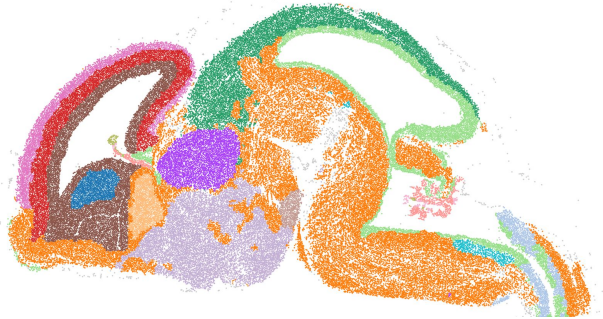
Clustering - Leiden, Spectral, Hierarchical

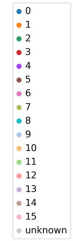
- Leiden uses parameter *r* - resolution for adjusting cluster number.
- Hierarchical clustering is implemented as Agglomerative clustering with ‘ward’ linkage.
- Spectral and Hierarchical clustering use parameter *n_clusters* as predefined value.

cell types (annotation)



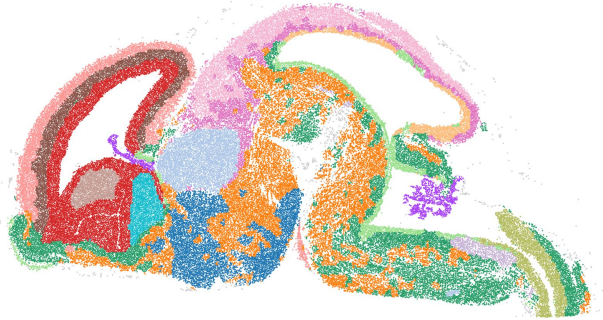
spectral

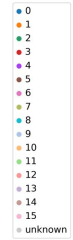




leiden

agglomerative

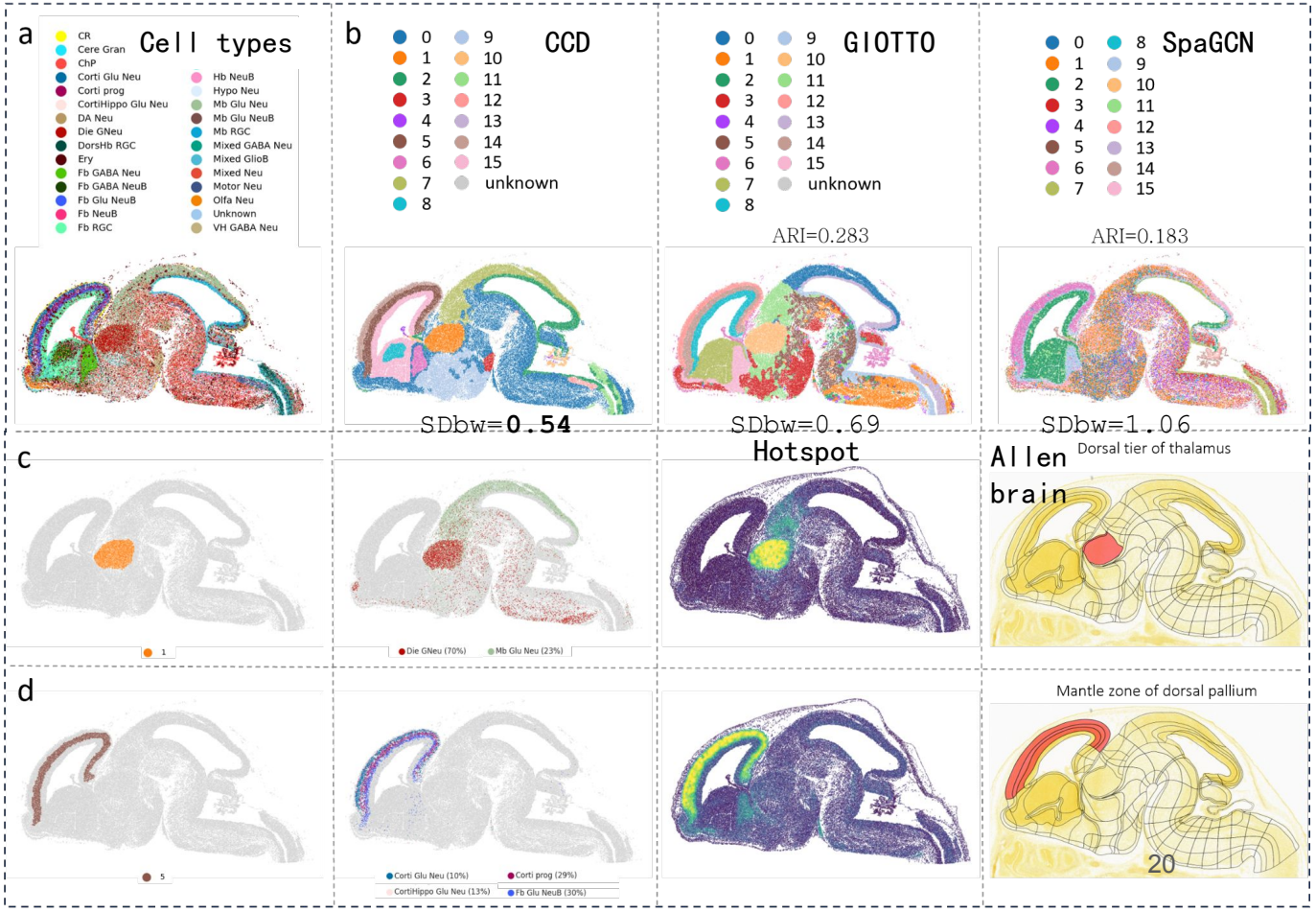




Results - whole brain mouse embryo 16.5

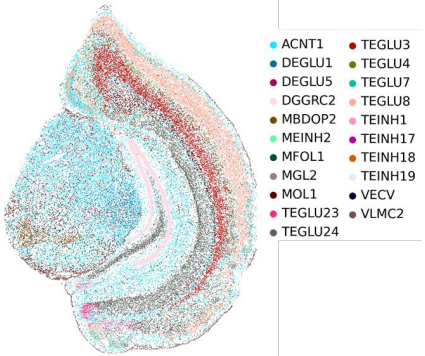
- CCD provides more coherent and separable communities than Giotto's GSDI and spaGCN
- CCD can infer biological function and structure

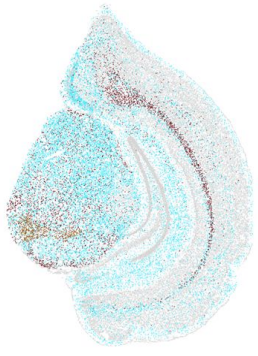
- Cell communities 1 and 5 match in shape and position to the Hotspot modules (implying functional domain) and Allen brain atlas anatomical region (implying structural domain)

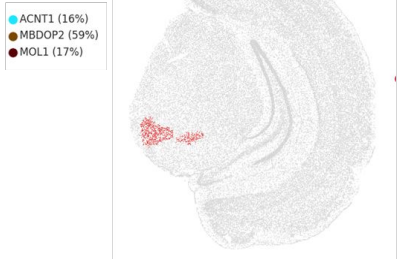


Results - adult mouse brain, single slice

annotation

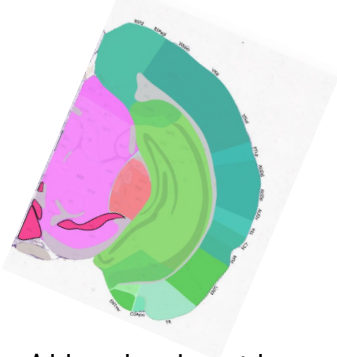




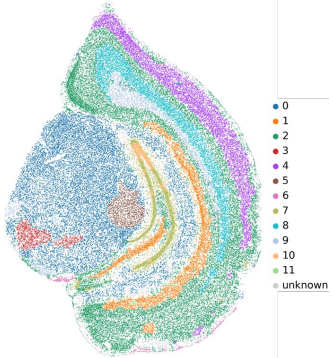


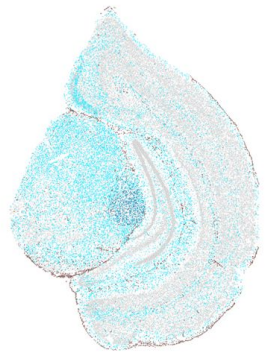
Hotspot

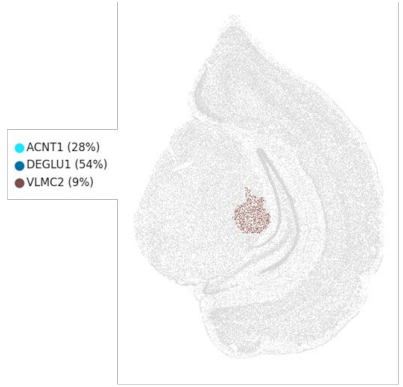
Midbran,
behavioral state related

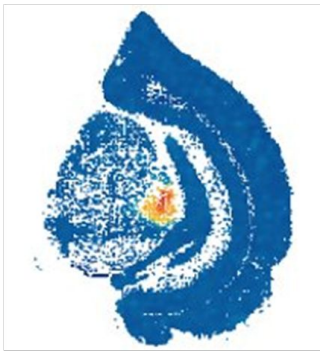


cell communities



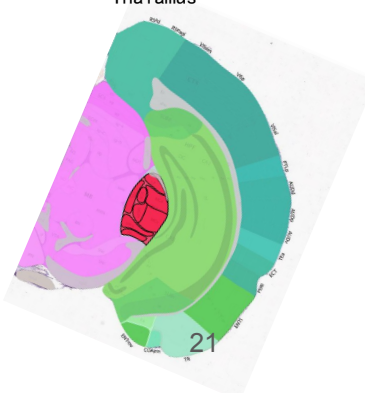






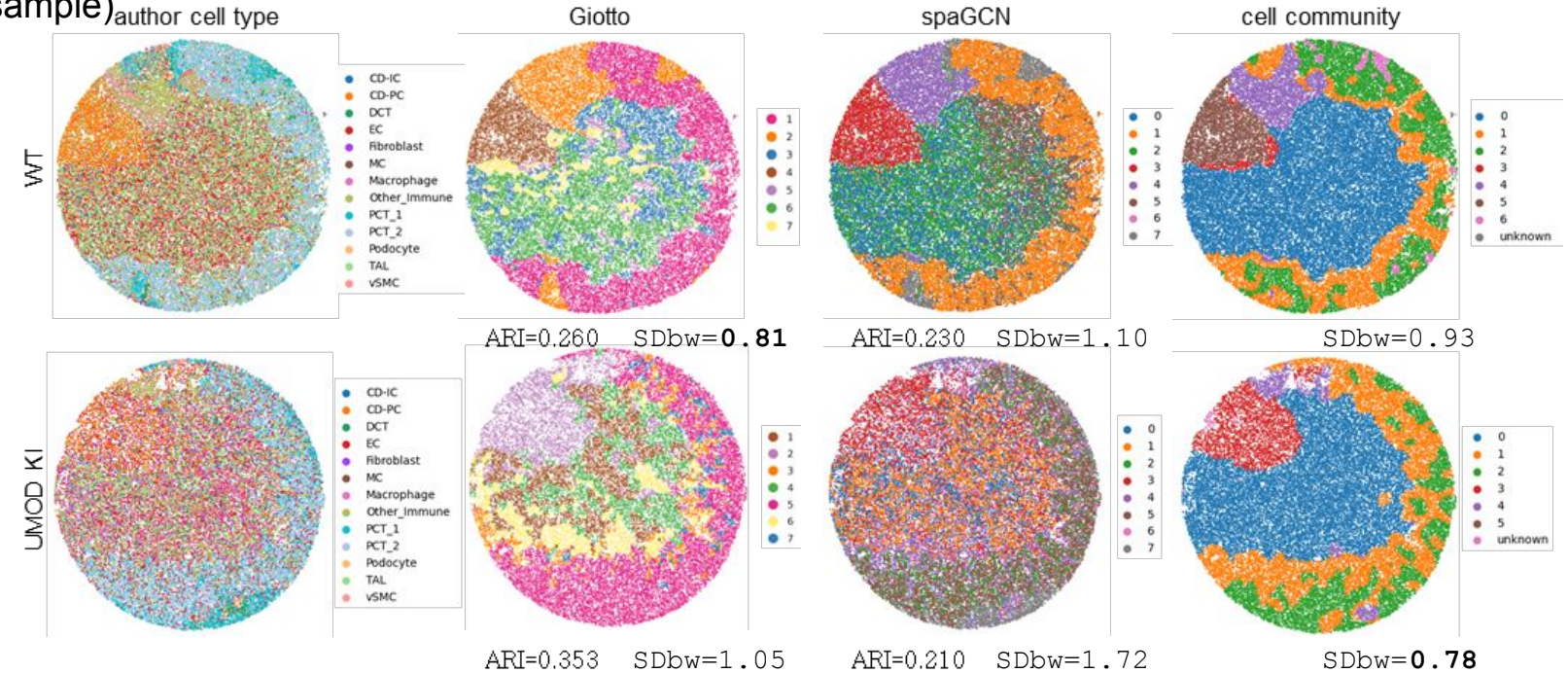
Allen brain atlas

Thalamus



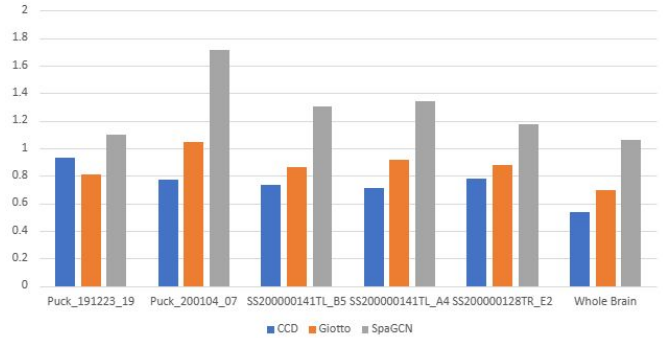
Results - Mouse kidney comparative slices

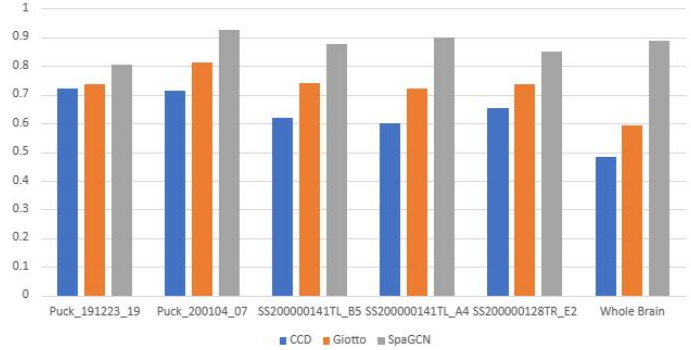
- spaGCN manages to detect several coherent domains in WT sample, but fails in ob/ob
- Giotto's SDI creates coherent domains on slices' borders, but proves unstable in the central medulla section
- CCD uses multi-slice approach which provides unified label for whole medulla area, creating a good base for further analysis of sample (gene expression differences between medulla region in WT and ob/ob sample)



Scatter-Davies–Bouldin and Scatter-Distance clustering marks

- Lower S Dbw (SCATT+DENS_BW) and SD (K*SCATT+DISTANCE) values indicate better



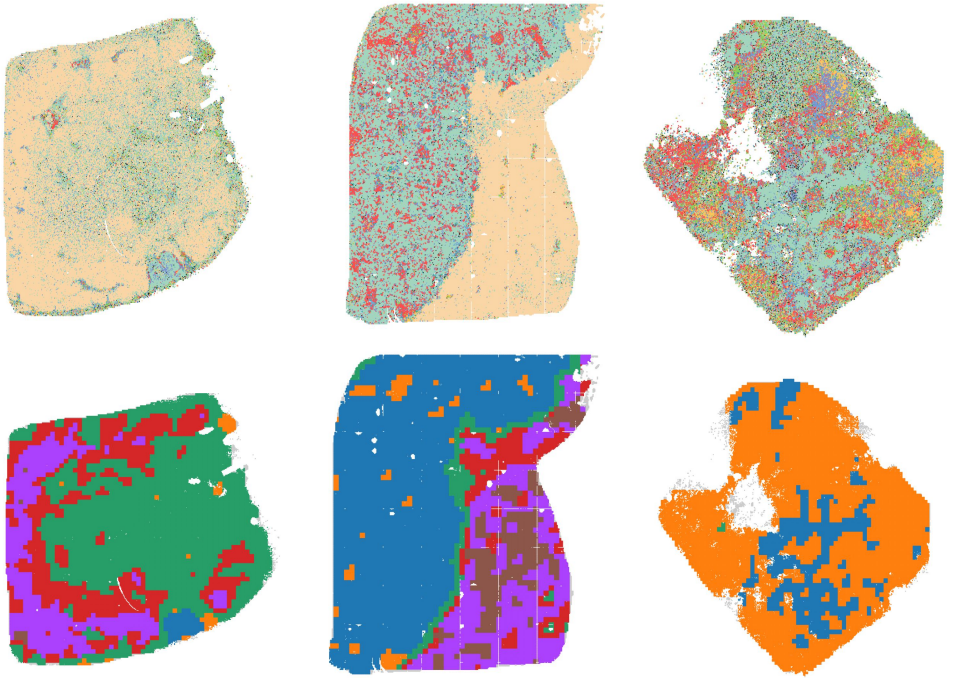


S DBW	Giotto	SpaGCN	CCD
Puck_191223_19	**0.807**	1.097	0.931
Puck_200104_07	1.045	1.717	**0.775**
SS200000141TL_B5	0.863	1.301	**0.736**
SS200000141TL_A4	0.919	1.339	**0.709**
SS200000128TR_E2	0.875	1.173	**0.781**
Whole brain	0.694	1.059	**0.539**

SD	Giotto	SpaGCN	CCD
Puck_191223_19	0.737	0.805	**0.721**
Puck_200104_07	0.813	0.926	**0.714**
SS200000141TL_B5	0.739	0.876	**0.620**
SS200000141TL_A4	0.723	0.898	**0.599**
SS200000128TR_E2	0.737	0.852	**0.653**
Whole brain	0.592	0.889	**0.483**

Results - human liver comparative cancer sample

- Gave some new insights for the margin tissue (Liang et al.)
- Decrease in the number of hepatocytes - indication of tumor invasion



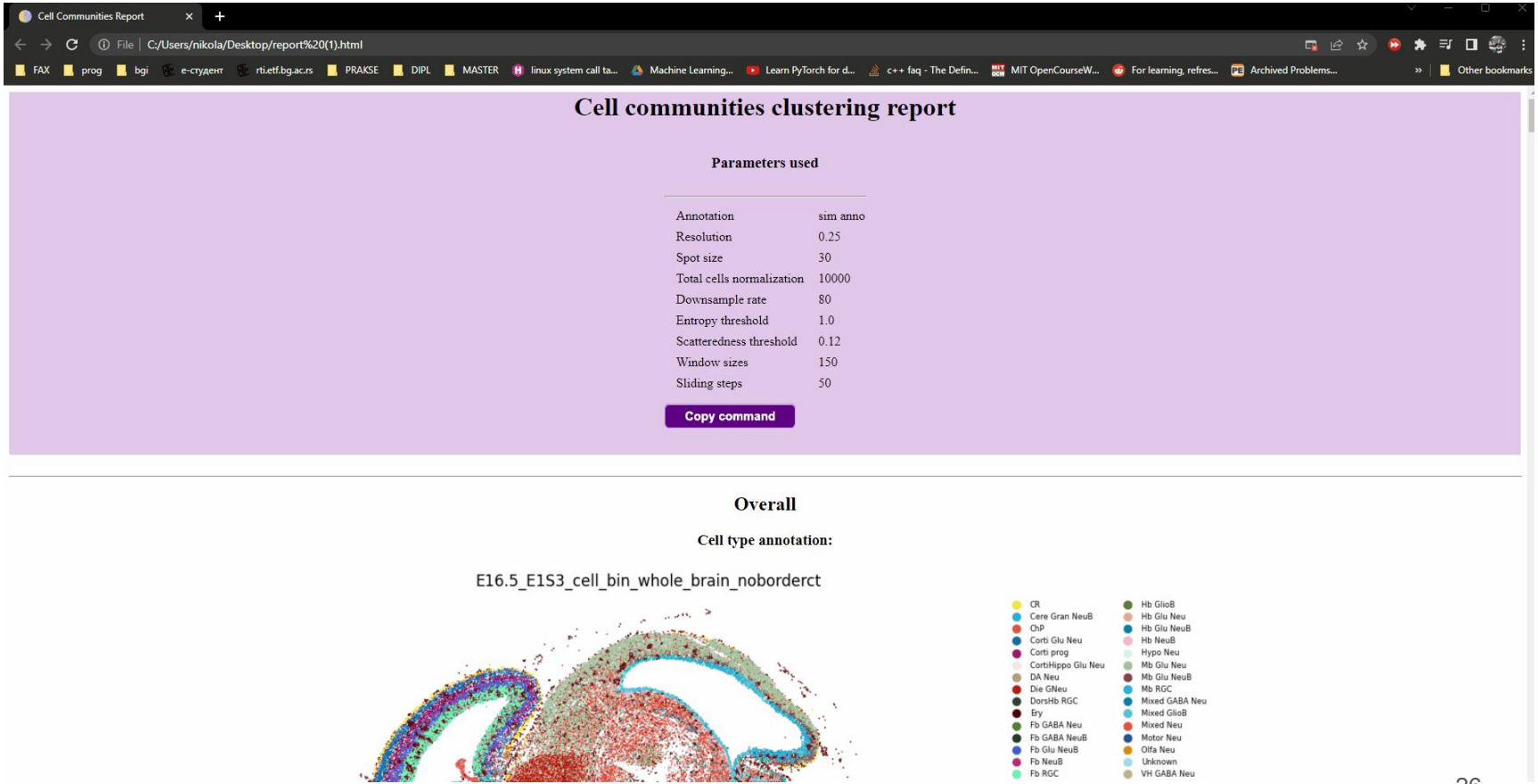
CCD execution time and memory consumption

- The execution time of the CCD is notably faster compared to the GIOTTO and spaGCN, demonstrating a speedup of at least 90 and 35 times, respectively
- The peak memory consumption is affected by the dimensions of the input file, rendering CCD significantly more efficient due to its independence from gene expression matrices

Table S1 Execution time and peak memory consumption for Giotto's SDI spaGCN and Stereopy's CCD.

	Execution time [s]			***Memory consumption [MiB RAM]***		
Tools	Mouse embryo brain (1 slice 59 704 cells)	Adult mouse brain (3 slices 128 281 cells)	UMOD KI (2 slices 67 493 cells)	Mouse embryo brain (1 slice 59 704 cells)	Adult mouse brain (3 slices 128 281 cells)	UMOD KI (2 slices 67 493 cells)
GIOTTO	13366	19854	10974	80632	37884	31675
spaGCN	3696	1342	1869	85445	58445	26534
CCD	**79**	**214**	**31**	**845**	**25716**	**684**

HTML report


Cell Communities Report
File | C:/Users/nikola/Desktop/report%20(1).html
FAX
prog
bgi
e-студент
rti.etf.bg.ac.rs
PRAKSE
DIPL
MASTER
linux system call ta...
Machine Learning...
Learn PyTorch for d...
c++ faq - The Defin...
MIT OpenCourseW...
For learning, refres...
Archived Problems...
Other bookmarks
Cell communities clustering report
Parameters used
Annotation sim anno
Resolution 0.25
Spot size 30
Total cells normalization 10000
Downsample rate 80
Entropy threshold 1.0
Scatteredness threshold 0.12
Window sizes 150
Sliding steps 50
Copy command
Overall
Cell type annotation:
E16.5_E1S3_cell_bin_whole_brain_noborderct
CR
Cere Gran NeuB
ChP
Corti Glu Neu
Corti prog
CortiHippo Glu Neu
DA Neu
Die GNeu
DorsHb RGC
Ery
Fb GABA Neu
Fb GABA NeuB
Fb Glu NeuB
Fb NeuB
Fb RGC
Hb GlioB
Hb Glu Neu
Hb Glu NeuB
Hb NeuB
Hypo Neu
Mb Glu Neu
Mb Glu NeuB
Mb RGC
Mixed GABA Neu
Mixed GlioB
Mixed Neu
Motor Neu
Olfa Neu
Unknown
VH GABA Neu

Visualisations - cell types tables

This table shows percentage of cell types among clusters (like the table on the right). It also shows the size of the cluster as the exact number of cells in it and what percentage of cells it contains. It is generated for every slice separately and for all slices combines.

	B_cell	Cholangiocyte	DC	Endothelial	Fibroblast	Hepatocyte	Macrophage	Malignant	NK	T_cell	total_counts	perc_of_all_cells
0	11	20	3	4	50	0	4	0	4	0	3179	4
1	0	4	0	0	0	91	0	0	0	0	32782	36
2	4	32	1	1	3	45	7	0	2	0	2185	2
3	9	53	3	1	25	0	4	1	1	0	18445	20
4	7	74	1	0	12	0	1	0	0	0	32053	35
5	6	41	3	2	10	17	12	1	3	0	2809	3
total_cells	5109	37828	1359	713	10824	31651	2211	613	1058	87	91453	100

This table shows cell type abundances per cluster (cell communities) and cell type. Colors of column labels are matched with cluster and cell type annotation colors. This table is generated for every slice separately and it can help in detecting the influence of cell types on different communities.

Cell type abundance per cluster (and per cel type set)

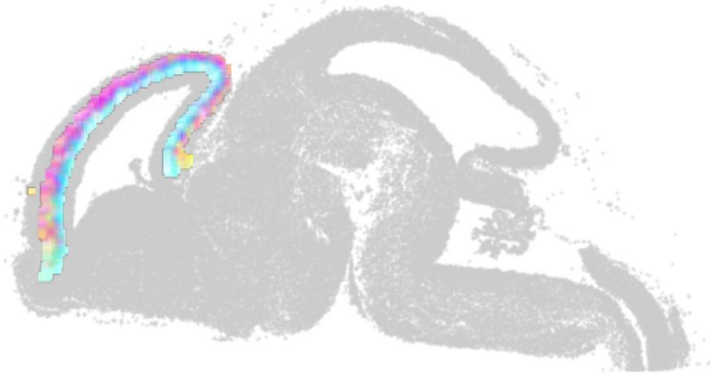
	cluster 0	cluster 1	cluster 2	cluster 3	cluster 4	cluster 5
B.cell	11% (7%)	0% (6%)	4% (1%)	9% (35%)	7% (45%)	6% (3%)
Cholangiocyte	20% (1%)	4% (4%)	32% (1%)	53% (26%)	74% (62%)	41% (3%)
DC	3% (8%)	0% (2%)	1% (2%)	3% (40%)	1% (17%)	3% (7%)
Endothelial	4% (18%)	0% (15%)	1% (5%)	1% (26%)	0% (25%)	2% (8%)
Fibroblast	50% (14%)	0% (0%)	3% (0%)	25% (42%)	12% (38%)	10% (2%)
Hepatocyte	0% (0%)	91% (95%)	46% (3%)	0% (0%)	0% (0%)	17% (1%)
Macrophage	4% (7%)	0% (6%)	7% (8%)	4% (37%)	1% (25%)	12% (16%)
Malignant	0% (2%)	0% (3%)	0% (2%)	1% (30%)	0% (52%)	1% (8%)
NK	4% (12%)	0% (26%)	2% (6%)	1% (24%)	0% (19%)	3% (10%)
T.cell	0% (3%)	0% (64%)	0% (4%)	0% (9%)	0% (6%)	0% (11%)

Visualisations - RGB colorplots

Plots of cell percentages per each window that are generated for every detected community. One plot shows percentages of top three cell types in the community for each window as red, green and blue channels of RGB. It is plotted over the tissue image.

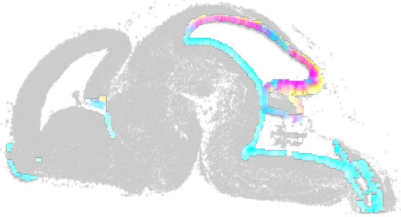
The plot provides visual spatial information on community uniformity and smoothness. Significant difference in cell type percentages provides different colors and shows the possible existence of several communities and need for increase of the clustering resolution (as shown for community 9).

RGB of community 9 win size 150, step 50 - top 3 cell types (Slice_0)



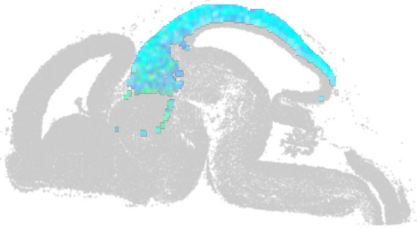
R' - Fb Glu NeuB (30%)
G' - Corti prog (29%)
B' - CortiHippo Glu Neu (13%)

RGB of community 0 win size 150, step 50 - top 3 cell types (Slice_0)



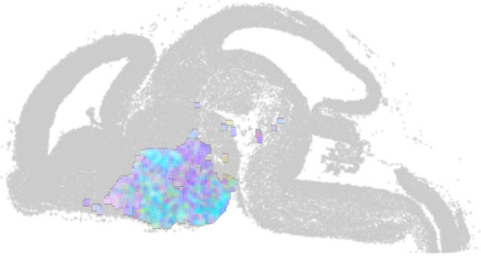
R' - Mixed GlioB (35%)
G' - Mb RGC (19%)
B' - Mb Glu NeuB (9%)

RGB of community 1 win size 150, step 50 - top 3 cell types (Slice_0)



R' - Mb Glu Neu (74%)
G' - Mixed Neu (10%)
B' - Die GNeu (5%)

RGB of community 2 win size 150, step 50 - top 3 cell types (Slice_0)



R' - Hypo Neu (38%)
G' - Mixed Neu (24%)
B' - VH GABA Neu (9%)

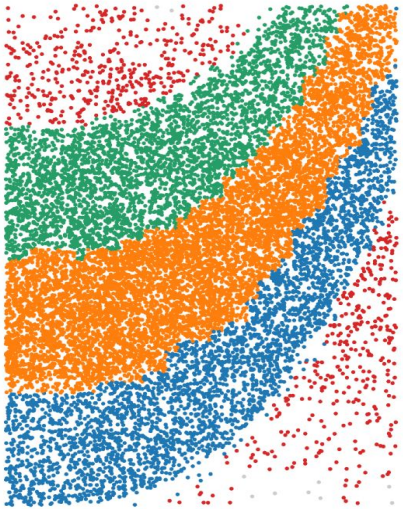
Tutorial notebook - synthetic sample

Default parameters, 14621 observations, 3 “cell types”, auto window size detection (ws164, ss82)

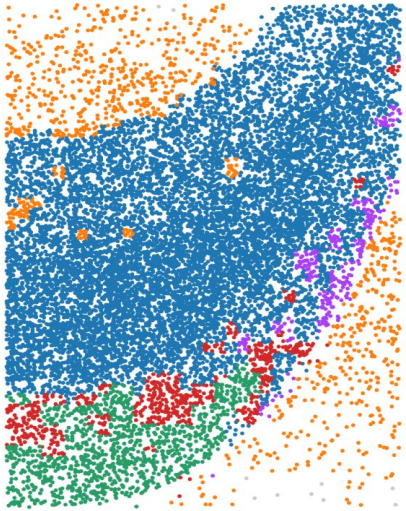
CCD execution time: 1.6s

annotation

Agglomerative n=4
(euclidian dist)



Leiden, r=0.01
nearest neigh



Spectral n=4
nearest neigh