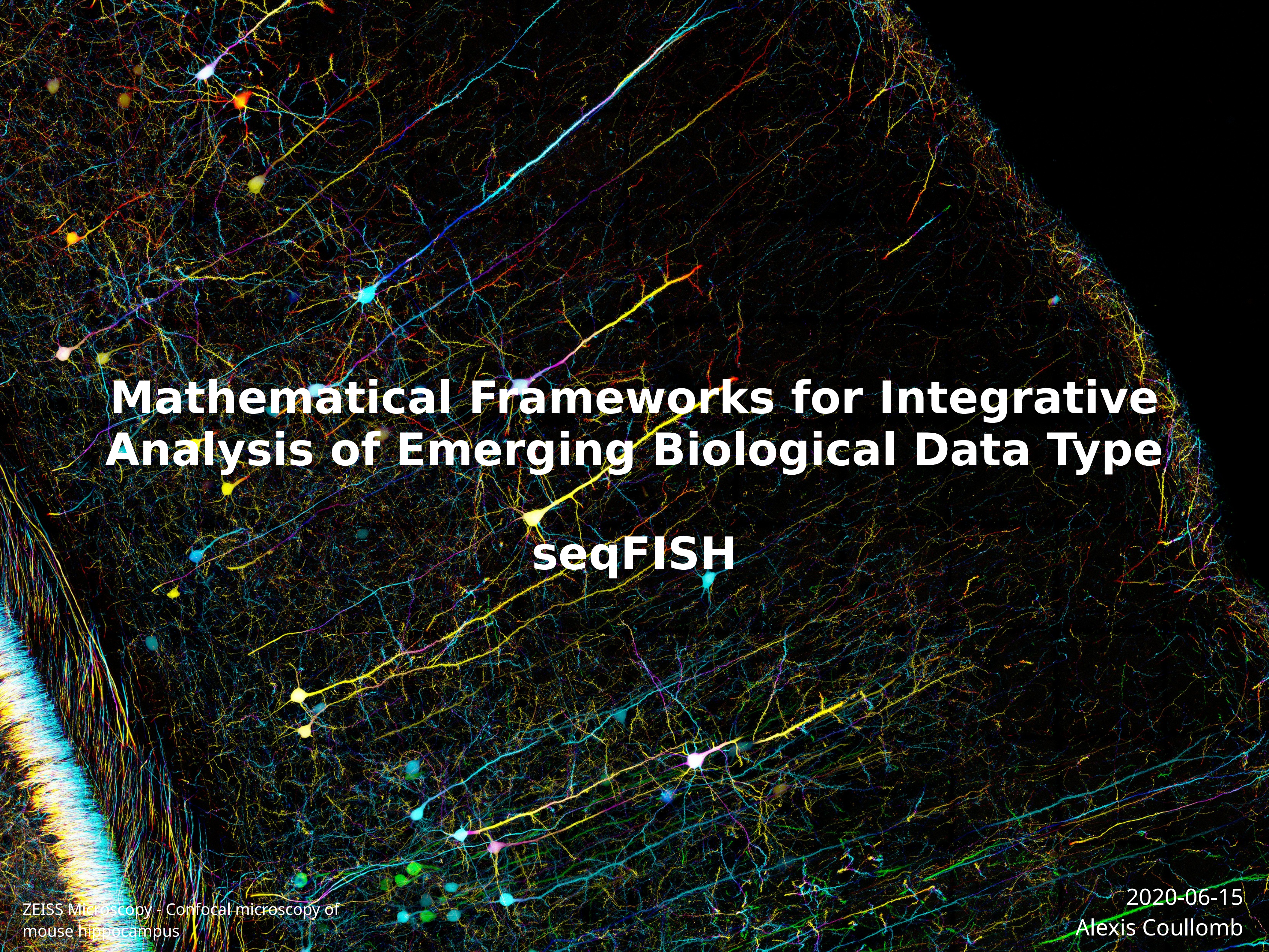


Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Type

seqFISH

ZEISS Microscopy - Confocal microscopy of mouse hippocampus

2020-06-15
Alexis Coullomb

The pyrénées

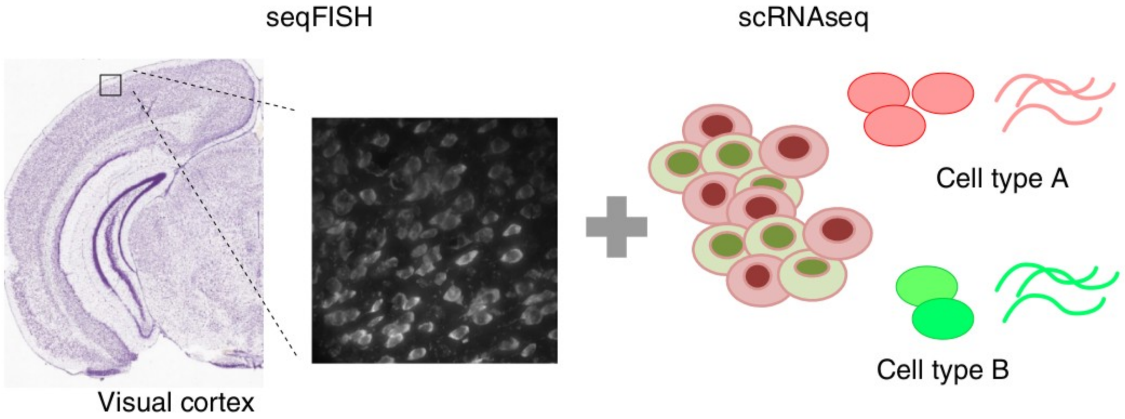
The challenge

Mouse visual cortex

Zhu – Nat. biotechnology - 2018

Tasic – Nat. Neuroscience - 2016

seqfish_theme
Guo-Cheng Yuan
Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health



1597 cells x 113 genes

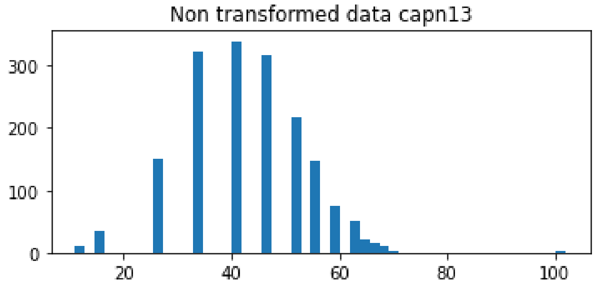
1723 cells x 113 genes

- Can scRNA-seq data be overlaid onto seqFISH for resolution enhancement?
- What is the minimal number of genes needed for data integration?
- Are there signatures of cellular co-localization or spatial coordinates in non-spatial scRNA-seq data?

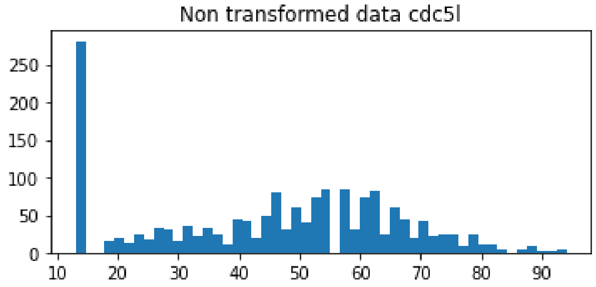
Data transformation

Data already processed for this challenge

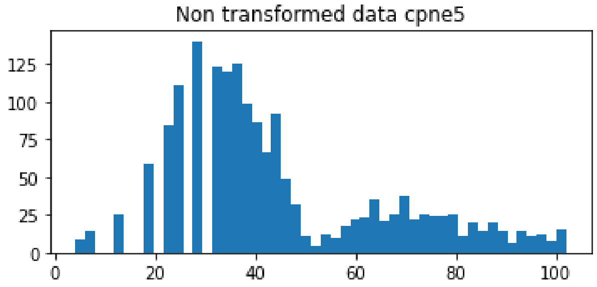
3 types of distributions


Non transformed data capn13

gaussian

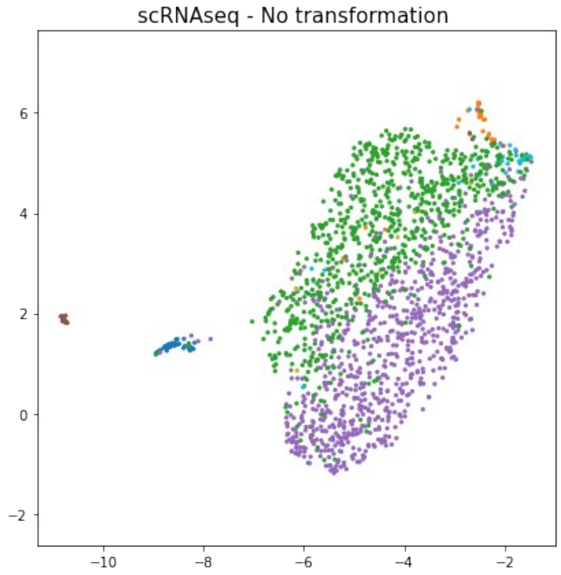

Non transformed data cdc5l

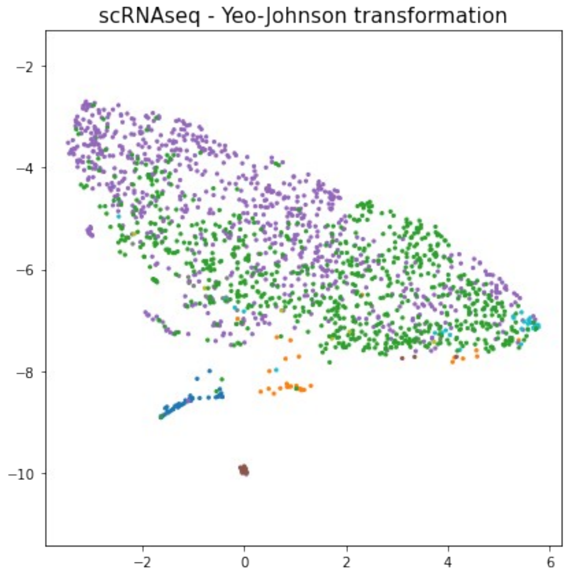
peak at min values

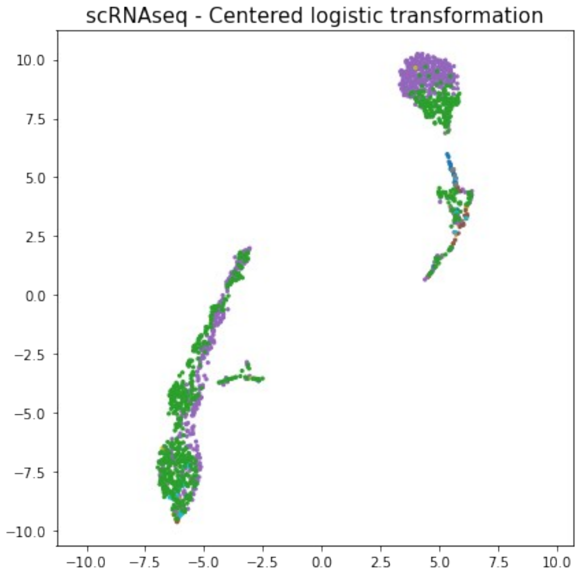

Non transformed data cpne5

bimodal

Caution when transforming data!


scRNAseq - No transformation


scRNAseq - Yeo-Johnson transformation


scRNAseq - Centered logistic transformation

Classification of scRNAseq data

First try

Supervised: Tasic’s labels → gold standard
Model trained on scRNAseq data

As in Zhu et al.: Support Vector Classifier with `C = 10`$^{-6}$, `class_weight = 'balanced'`

Accuracy

	Linear SVC	Kernel SVC
Zhu’s param	0.23	
default	0.57	0.91

Balanced accuracy

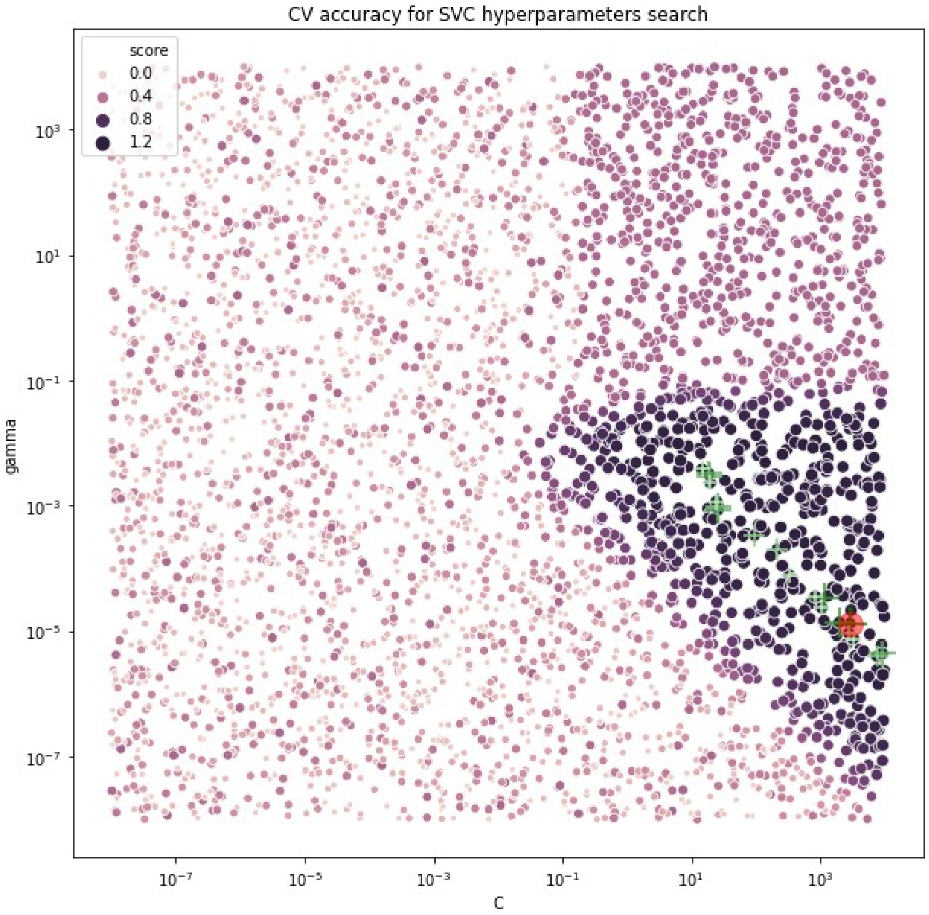
	Linear SVC	Kernel SVC
Zhu’s param	0.10	
default	0.57	0.80

Due to difference in data processing?

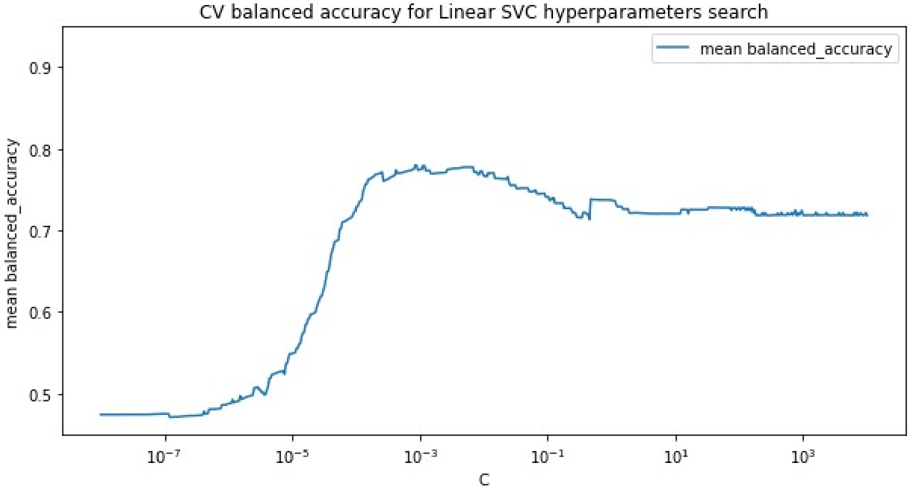
Classification of scRNAseq data

Hyperparameters search

Kernel SVC: randomized search + zoomed search



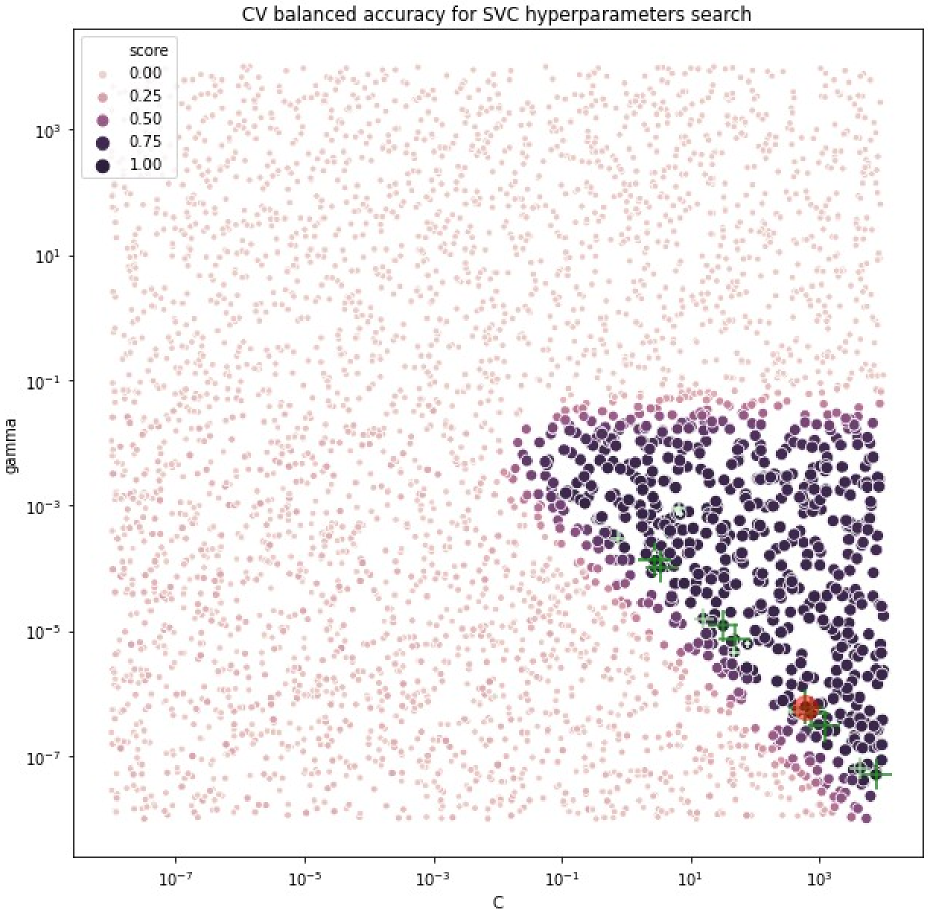
Linear SVC: grid search (1D grid)



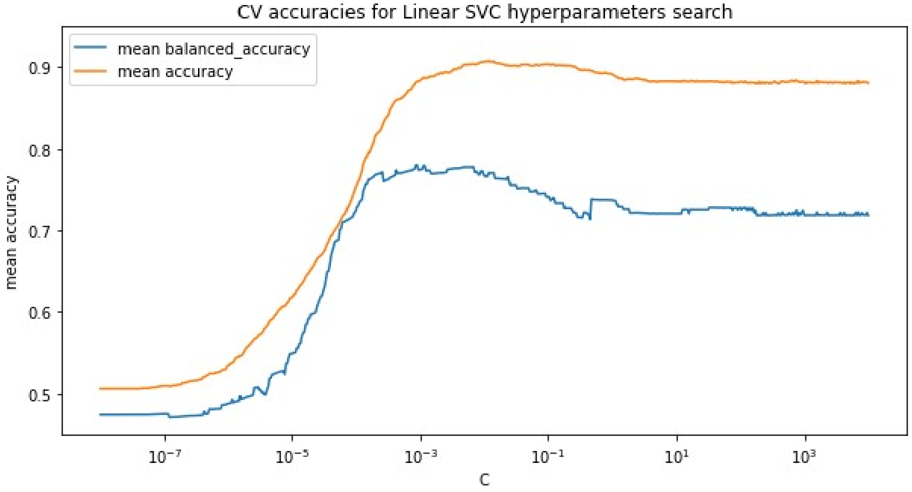
Classification of scRNAseq data

Hyperparameters search

Kernel SVC: randomized search + zoomed search



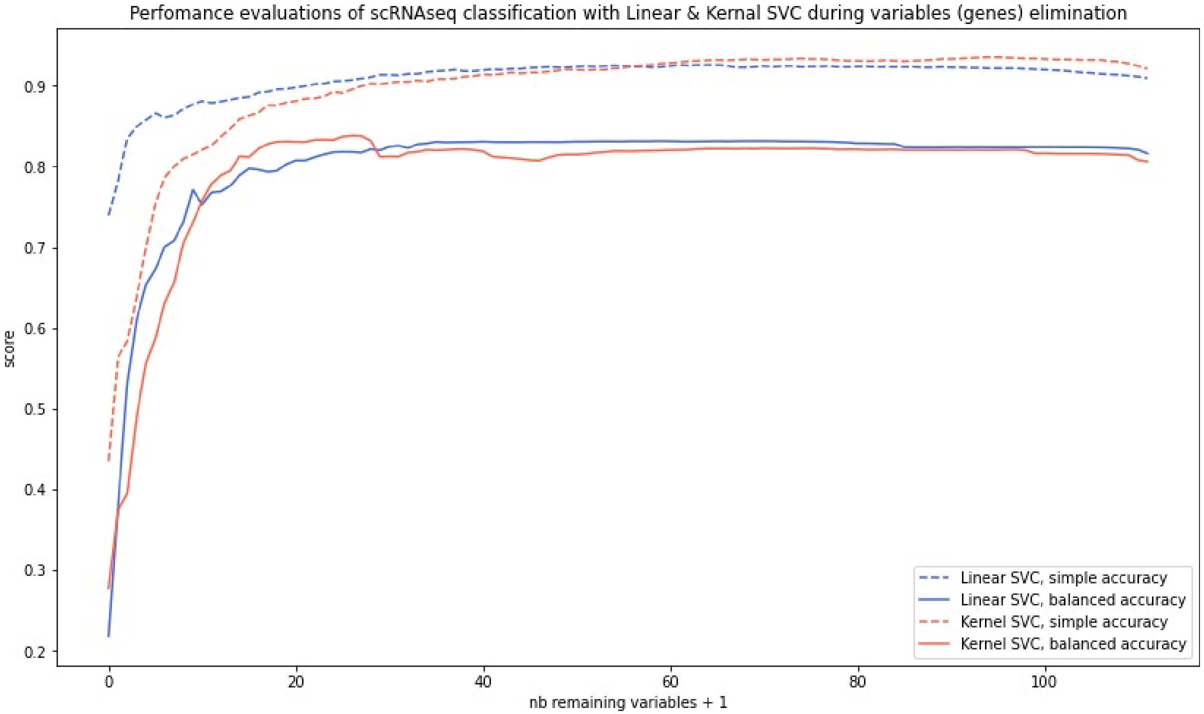
Linear SVC: grid search (1D grid)



Accuracy overestimates classifier performance on imbalanced dataset
Accuracy shift best hyperparameters values

Top-down elimination of variables

```
Initial set of variables (genes):
For each variable:
    - discard it
    - train and test classifier with remaining variables
Drop variable with best score when discarded
Update set of variables
… until only 1 variable left
```



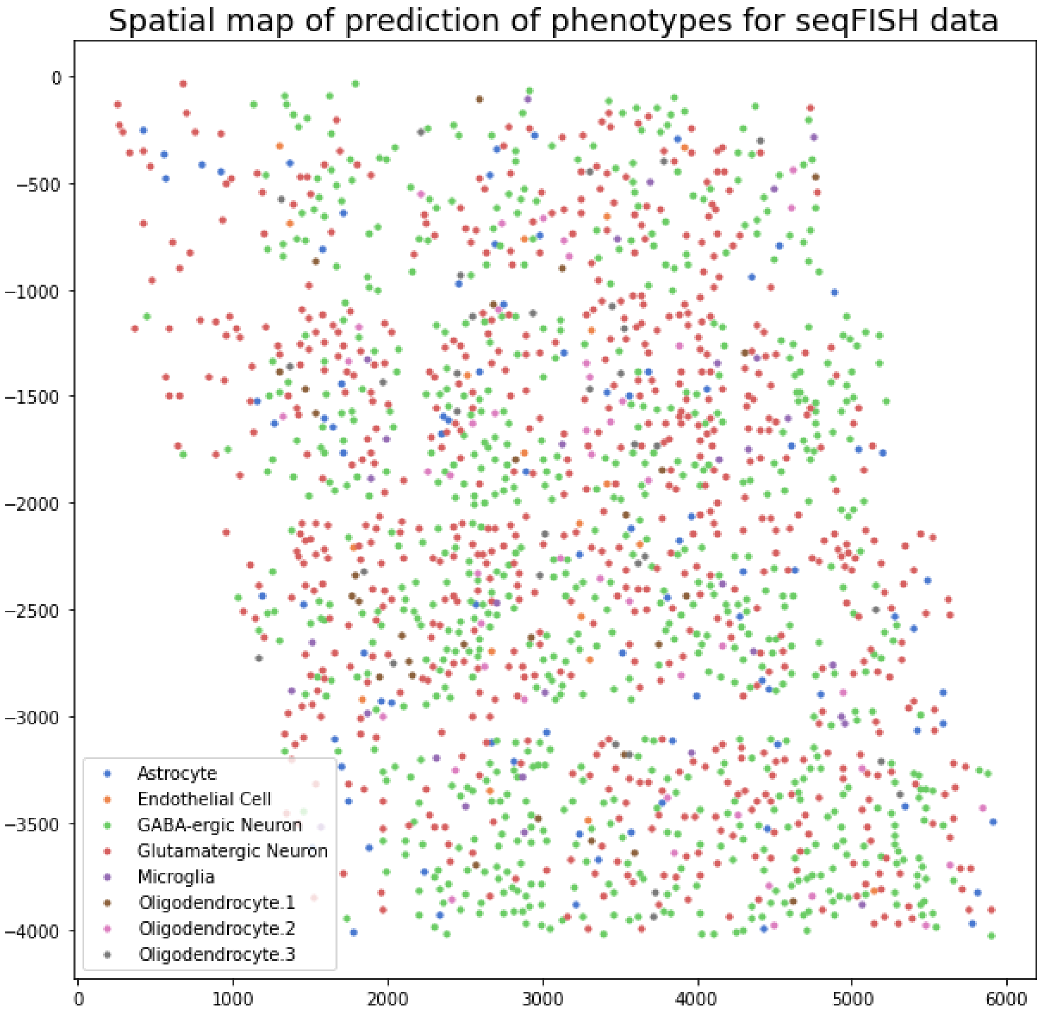
Better balanced accuracy for kernel SVC with fewer genes!

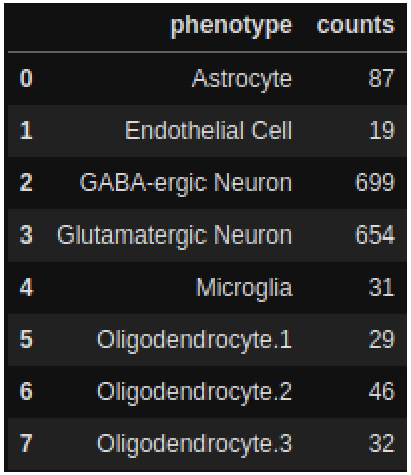
→ due to generalization improvements?

→ role of genes deleted and kept?

Infer cell types from few genes

Re-run 2-steps hyperparameters search with 19 genes



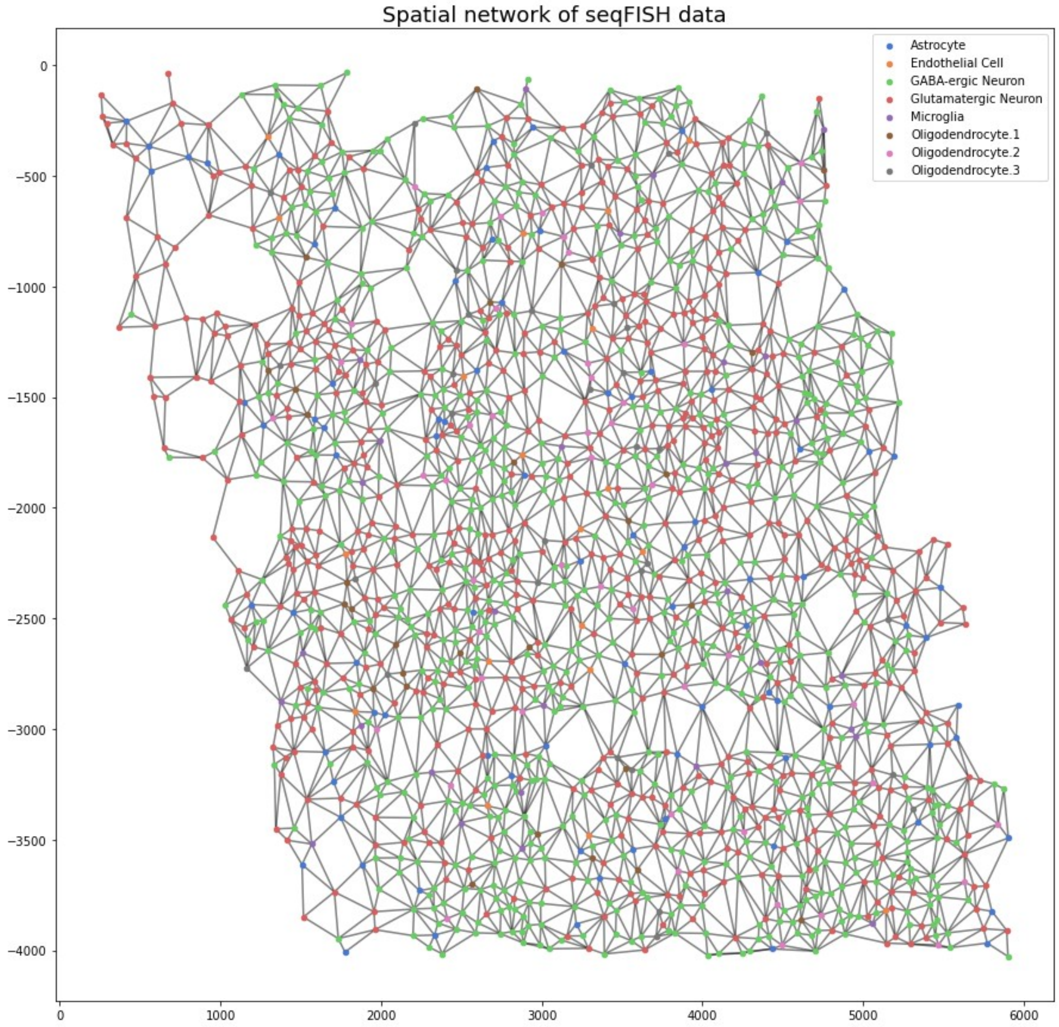


	phenotype	counts
0	Astrocyte	87
1	Endothelial Cell	19
2	GABA-ergic Neuron	699
3	Glutamatergic Neuron	654
4	Microglia	31
5	Oligodendrocyte.1	29
6	Oligodendrocyte.2	46
7	Oligodendrocyte.3	32

Spatial analysis

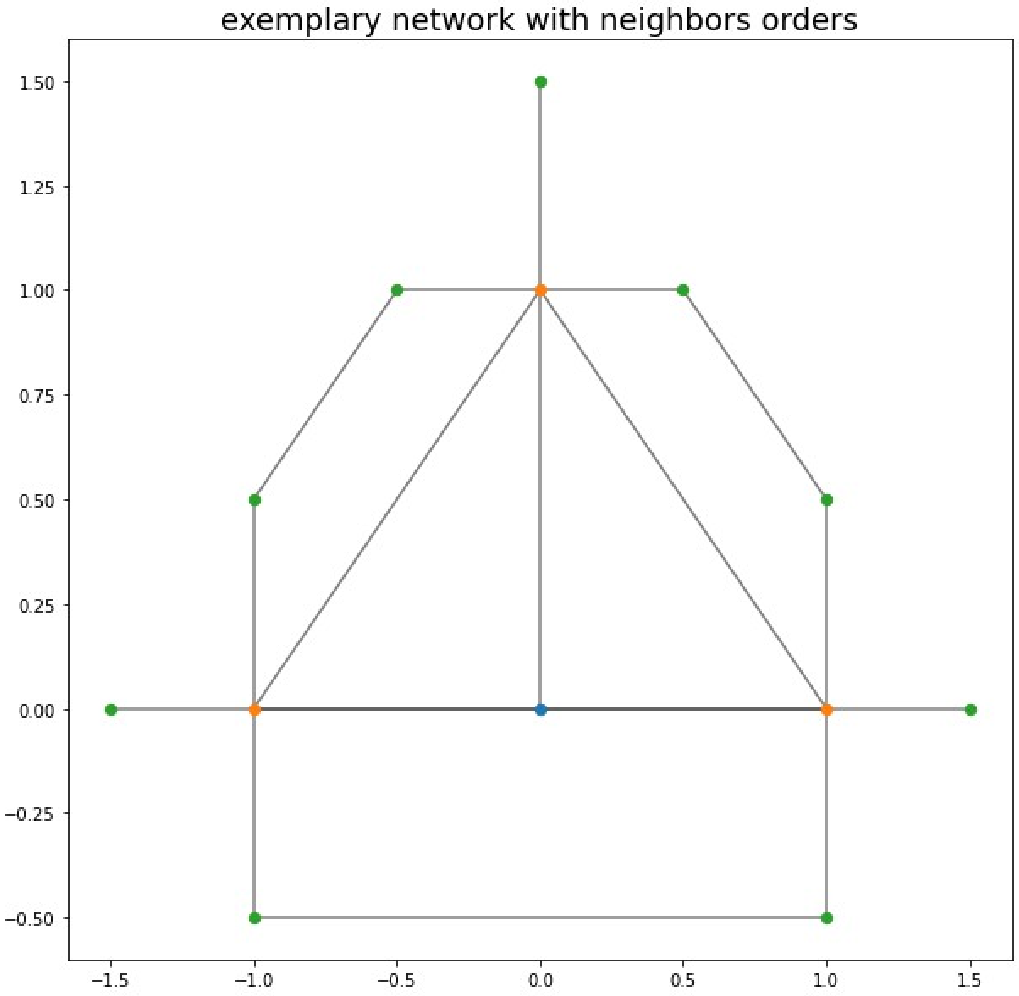
How to define areas?

Network with Voronoi tessellation + distance threshold for artifacts

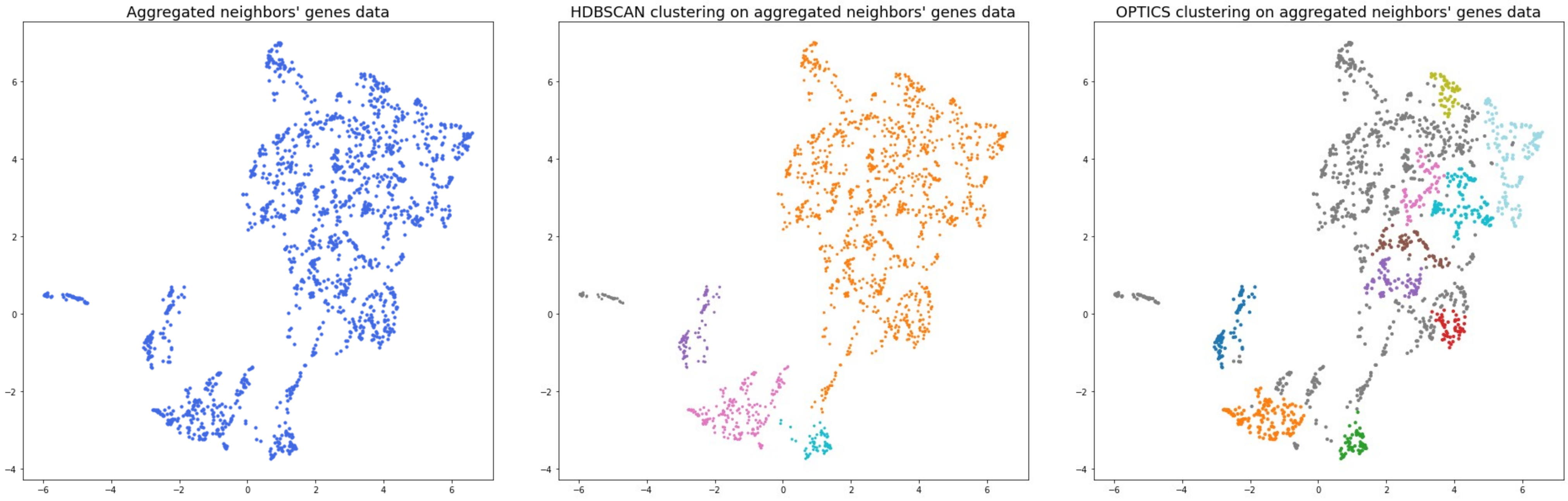


Neighbors gene expression aggregation

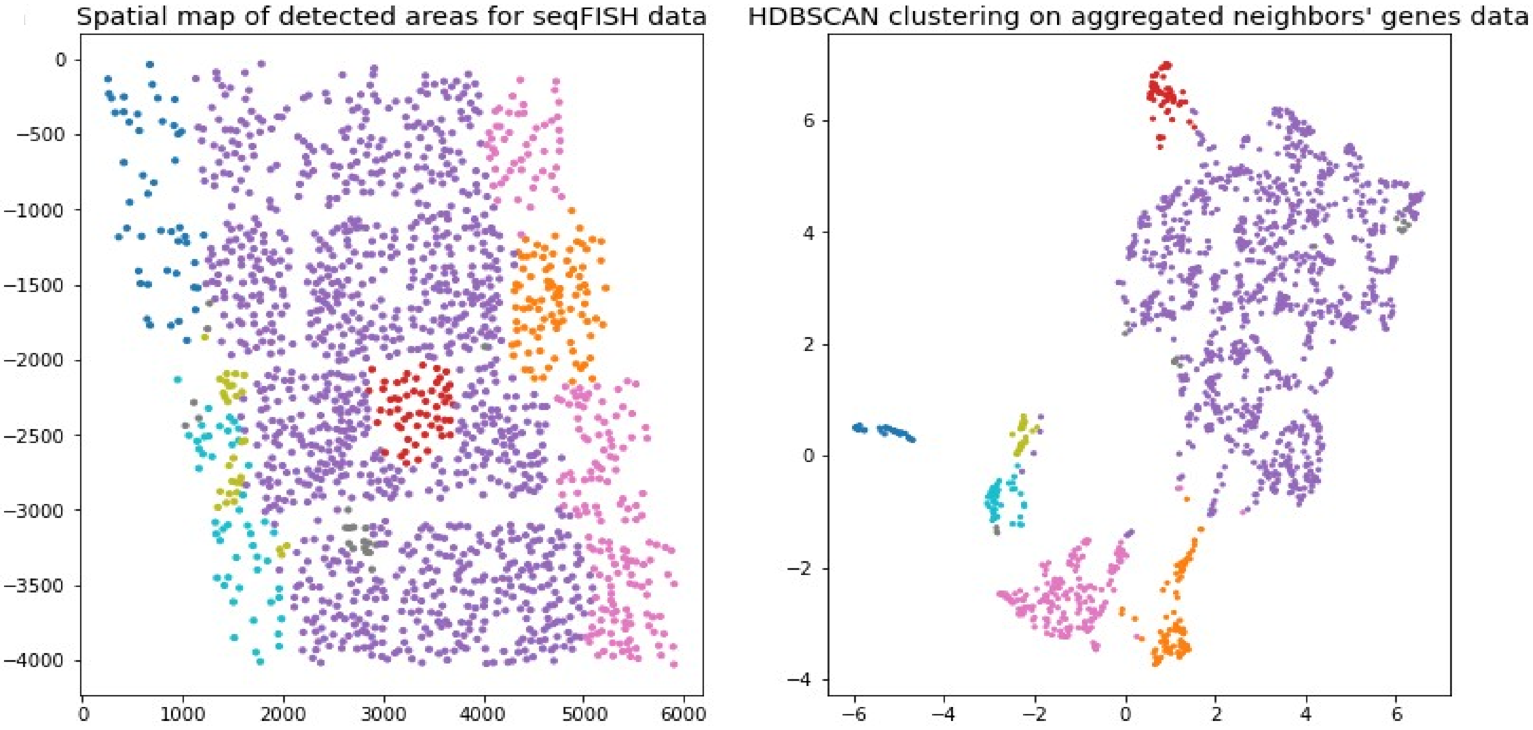
```
For each node:
    - detect all direct neighbors
    - stack all their gene expression data
    - compute some statistics per gene: mean, std, ...
```



UMAP projection of aggregated data



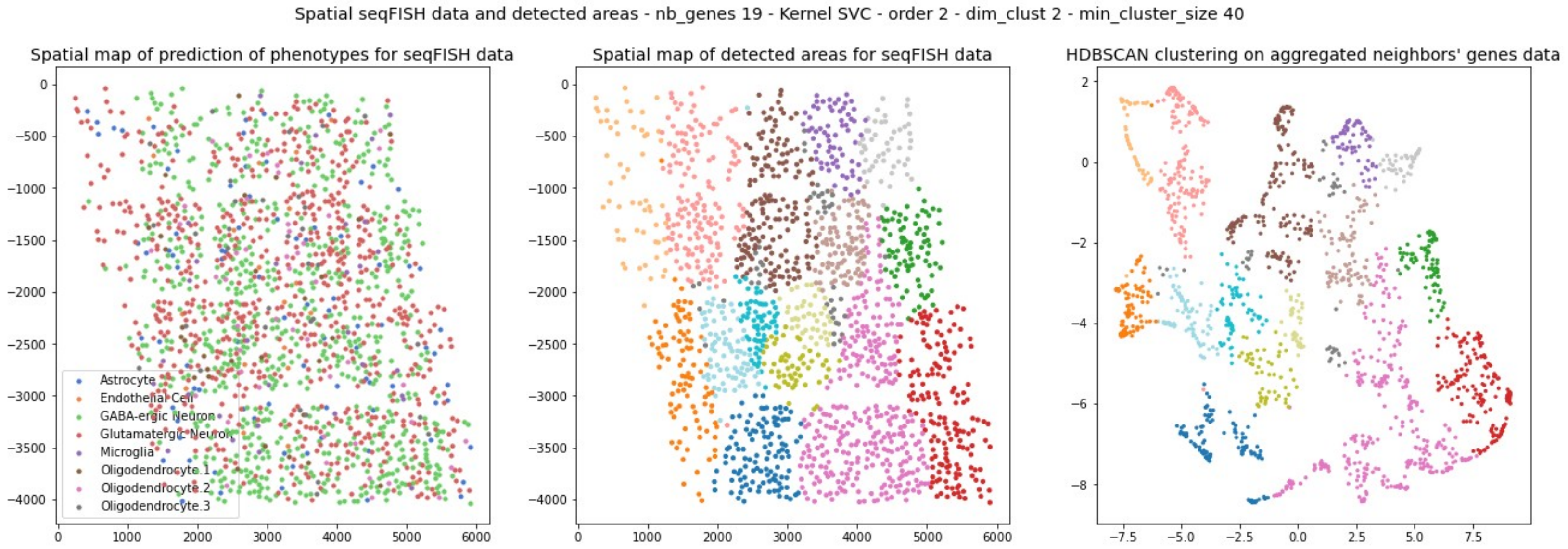
Detected areas

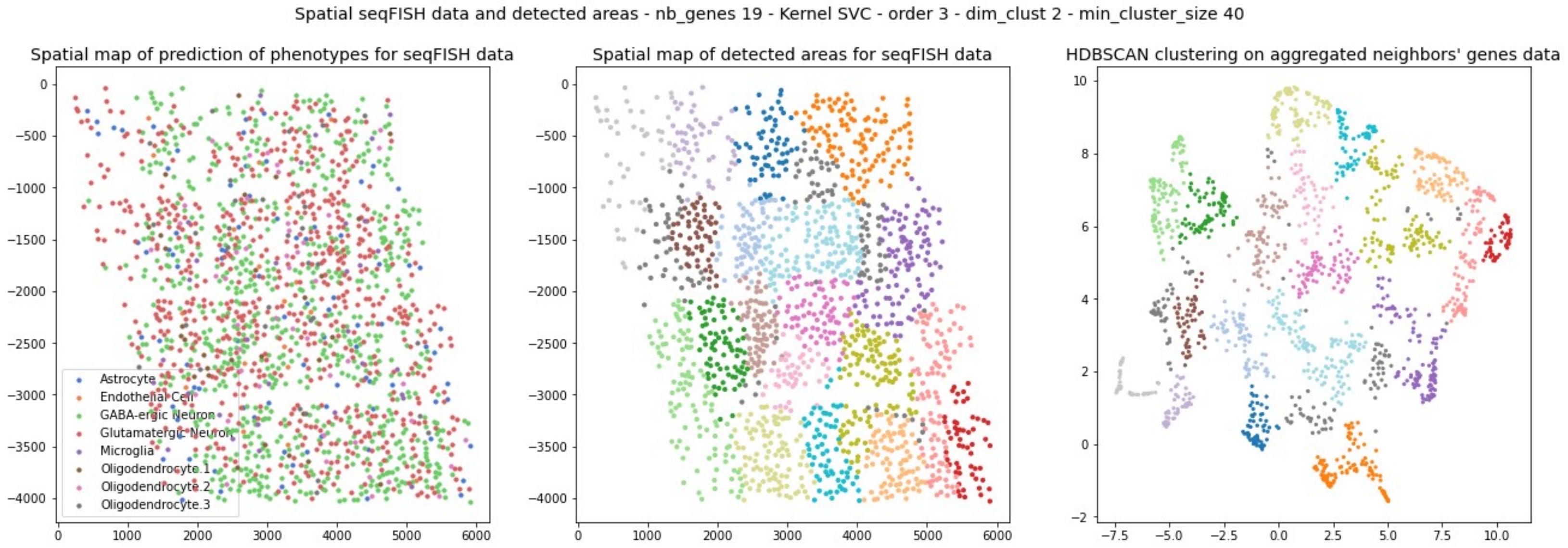


Spatial map of detected areas for seqFISH data

HDBSCAN clustering on aggregated neighbors' genes data

Higher orders neighbors

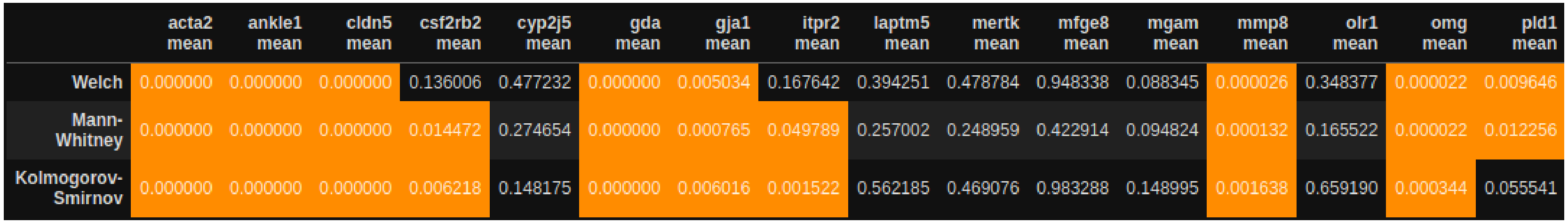




“Differential Expression”

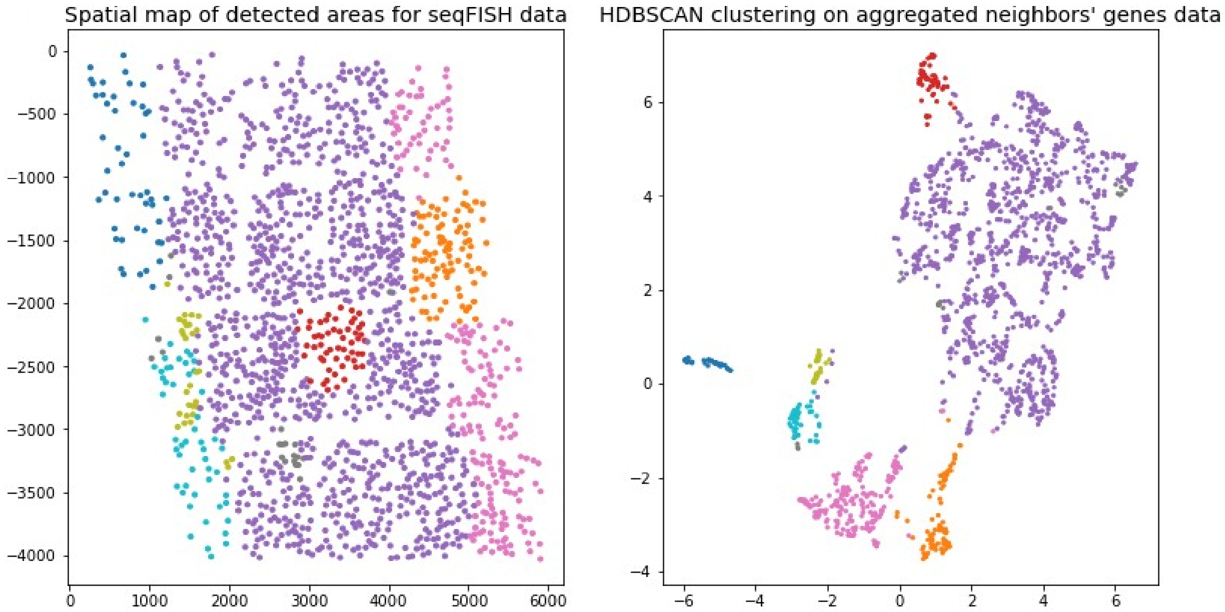
Not really, variables are statistics on aggregated data

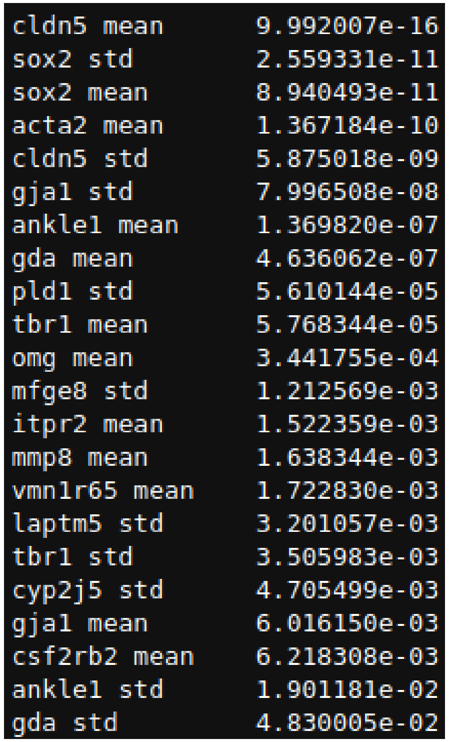
3 statistics for “DE” analysis



	acta2 mean	ankle1 mean	cldn5 mean	csf2rb2 mean	cyp2j5 mean	gda mean	gja1 mean	itpr2 mean	laptm5 mean	mertk mean	mfge8 mean	mgam mean	mmp8 mean	olr1 mean	omg mean	pld1 mean
Welch	0.000000	0.000000	0.000000	0.136006	0.477232	0.000000	0.005034	0.167642	0.394251	0.478784	0.948338	0.088345	0.000026	0.348377	0.000022	0.009646
Mann-Whitney	0.000000	0.000000	0.000000	0.014472	0.274654	0.000000	0.000765	0.049789	0.257002	0.248959	0.422914	0.094824	0.000132	0.165522	0.000022	0.012256
Kolmogorov-Smirnov	0.000000	0.000000	0.000000	0.006218	0.148175	0.000000	0.006016	0.001522	0.562185	0.469076	0.983288	0.148995	0.001638	0.659190	0.000344	0.055541

Compare red spot vs purple area





```
cldn5 mean       9.992007e-16
sox2 std         2.559331e-11
sox2 mean        8.940493e-11
acta2 mean       1.367184e-10
cldn5 std        5.875018e-09
gja1 std         7.996508e-08
ankle1 mean      1.369820e-07
gda mean         4.636062e-07
pld1 std         5.610144e-05
tbr1 mean        5.768344e-05
omg mean         3.441755e-04
mfge8 std        1.212569e-03
itpr2 mean       1.522359e-03
mmp8 mean        1.638344e-03
vmn1r65 mean     1.722830e-03
laptm5 std       3.201057e-03
tbr1 std         3.505983e-03
cyp2j5 std       4.705499e-03
gja1 mean        6.016150e-03
csf2rb2 mean     6.218308e-03
ankle1 std       1.901181e-02
gda std          4.830005e-02
```

“Differential Expression”

It’s a neural zone
Gja1 - … enhancing intercellular electrical and chemical transmission
Vmn1r65 - widespread protein family that includes hormone, neurotransmitter and light receptors
Pld1 - implicated as a critical step in numerous cellular pathways, including signal transduction, membrane trafficking
Itpr2 - release of intracellular calcium

Involved in regeneration?
Omg - Cell adhesion molecule contributing to the interactive process required for *myelination* in the central nervous system
Rtn4r - … mediates axonal growth inhibition and plays a role in regulating *axon regeneration* and neuronal *plasticity*
Sox2 - … controls the expression of a number of genes involved in *embryonic development*
Tbr1 - probable transcriptional regulator involved in *developmental processes*
Laptm5 - may have a special functional role during *embryogenesis* and in adult hematopoietic cells

Conclusion

/!\ data transformation

/!\ code review

Infer cell types with 19 genes

Network-based aggregation of neighboring cells gene expression data

Metrics to capture global tendency (mean) and variability (std)

Clustering on these metrics defines spatially coherent areas

~ DE analysis per area

Perspectives

Develop a multi-output regression model to overlay scRNAseq on seqFISH data

Network-based aggregation and clustering could reveal specific cell states

Apply to larger tissues, with higher order neighbors, decreasing weights

Optimize clusterization jointly on space and attributes

Subtract phenotype contributions to have space-only influence

Questions

If we look at enough genes, aren't we sure to find one that validates our area?
→ importance of comparing to other datasets, like the Allen Brain Atlas

How do you assess the optimal number of clusters? With information theory based criteria? (AIC, BIC, KIC ...)

If one gene is enough to define a cell state, how relevant are these criteria?

For small cell types discovery, could discarding lowly-variable genes be detrimental?

https://github.com/AlexCoul/multiOmics_integration

CRCT
Centre de Recherches en Cancérologie de Toulouse
ONCOPOLE

Inserm
La science pour la santé
From science to health

UNIVERSITÉ TOULOUSE III PAUL SABATIER
Université Fédérale Toulouse Midi-Pyrénées

FONDATION
TOULOUSE CANCER SANTÉ
ONCOPOLE

Pierre Fabre

Thank you