

# Project Summary

---

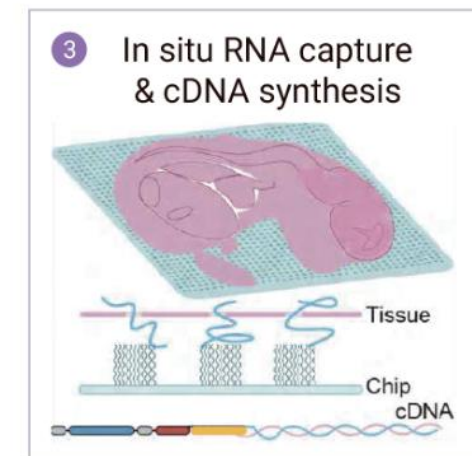
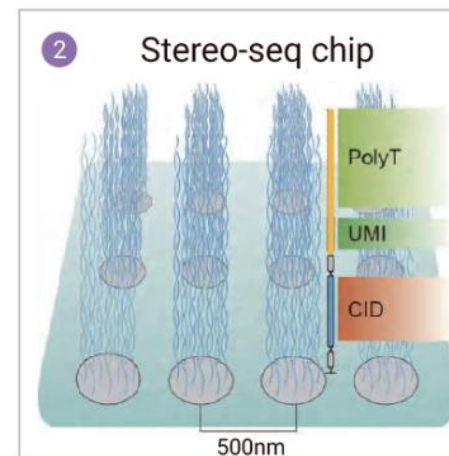
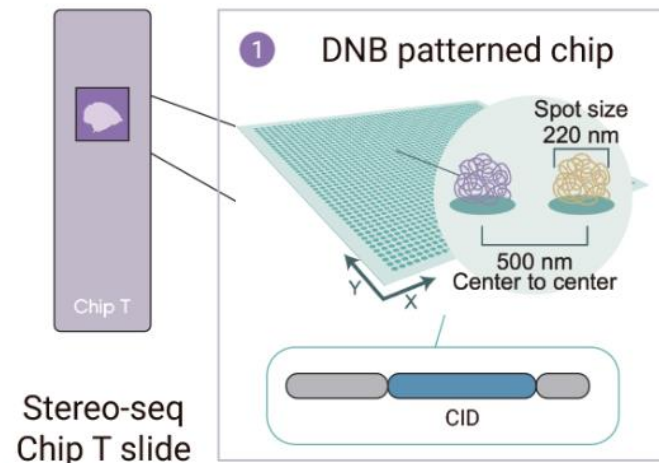
Xinyi Zhang

# Projects

- 1) ML model for neighbour's impact on cell of interest
- 2) ShinyApp - SpIIsoFind Explorer

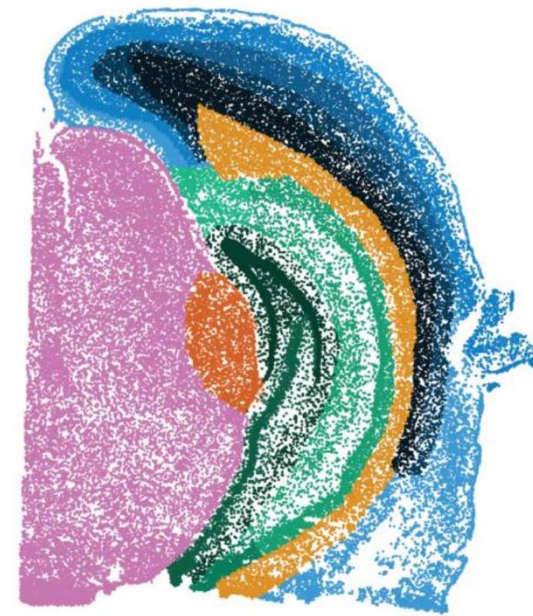
# Introduction

- Spatial transcriptomics
- Spatio-Temporal Enhanced Resolution Omics sequencing (Stereo-Seq)
- Gene expression in tissue sections
- Preserve spatial location of each transcript
- Spl-ISO-Seq2: Stereo-Seq + Long-read sequencing

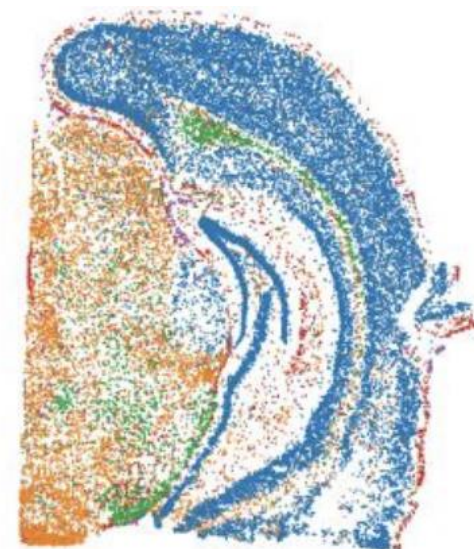


# Model

- Logistic Model: predict the inclusion / exclusion of an exon across all cells
- P56 male mice coronal plane slice
- Hypothesis: Neighbours might impact the alternative splicing pattern of the cell of interest



<b>Cortex</b>	<b>Hippocampus</b>	<b>White Matter</b>
• Cortex	• Subiculum	• Thalamus
• L23	• CA1	• Midbrain
• L4	• CA2	
• L5	• CA3	
• L6	• Dentate Gyrus	



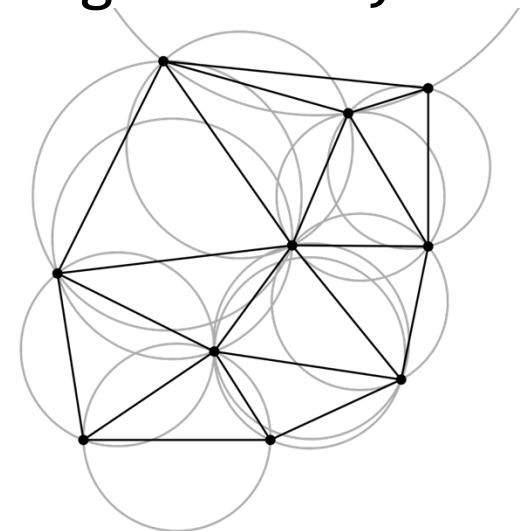
• Excitatory Neuron
• Inhibitory Neuron
• Oligodendrocyte
• Astrocyte
• Vascular

# Model

- X dataframe: rows are cells, columns are genes (91690 \* 3000)
  - Highly variable genes - ScanPy
  - Gene expression level of a cell
- Y dataframe: rows are cells, columns are exons (91690 \* 5001)
  - PSI (Percent Splice-In) value: proportion of mRNAs that include the exon
    - $$\frac{N \text{ of mRNA that include the exon}}{N \text{ of all mRNA}}$$
  - [0, 1]

# Model

- Three models
  - Cell + Neighbour
  - Cell
  - Neighbour
- How to define a neighbour?
  - `sq.gr.spatial_neighbors(data, coord_type="generic", delaunay = True)`



# Model

For each exon:

- Drop the NaN rows
- Filter out: too few samples, strongly imbalanced inclusion/exclusion ratios
- Stratified 10-Fold CV: the class balance is preserved in each fold
- Create logistic regression classifier: if  $PSI \geq 0.5$ : Inclusion (1); if  $PSI < 0.5$ : Exclusion (0)
- Run 10-fold CV predictions
- Measure the coefficients of the model (which genes have the most impact)
- Calculate accuracy, F1 score (inclusion, exclusion, mean)

# Result

Top 5 performing exons based on F1 mean

- Moderate – good predictive power
- Cell + neighbour
- Intrinsic + Neighbour influence

## Cell + Neighbour

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr12_112578982_112579038_ENSMUSG000000037679.10_+	0.788501	0.690691	0.839314	0.765002
chr1_80481636_80481711_ENSMUSG000000038608.16_-	0.809756	0.654867	0.868687	0.761777
chr7_44250293_44250352_ENSMUSG000000062785.15_+	0.700787	0.756410	0.612245	0.684328
chr2_130351750_130351776_ENSMUSG000000027303.19_+	0.685714	0.731707	0.620690	0.676198
chr11_83930482_83930517_ENSMUSG000000034940.16_+	0.780702	0.489796	0.860335	0.675066

## Cell

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr12_112578982_112579038_ENSMUSG000000037679.10_+	0.724846	0.601190	0.789969	0.695580
chr1_80481636_80481711_ENSMUSG000000038608.16_-	0.743902	0.516129	0.825871	0.671000
chr2_71063993_71064052_ENSMUSG000000027012.16_+	0.772727	0.440433	0.857406	0.648919
chr19_40298025_40298777_ENSMUSG000000025006.19_-	0.665217	0.729825	0.560000	0.644912
chr7_28594546_28594611_ENSMUSG000000054808.16_-	0.704433	0.491525	0.791667	0.641596

## Neighbour

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr1_80481636_80481711_ENSMUSG000000038608.16_-	0.821951	0.681223	0.876481	0.778852
chr12_112578982_112579038_ENSMUSG000000037679.10_+	0.776181	0.668693	0.831008	0.749850
chr7_44250293_44250352_ENSMUSG000000062785.15_+	0.700787	0.759494	0.604167	0.681830
chr7_28594546_28594611_ENSMUSG000000054808.16_-	0.724138	0.562500	0.798561	0.680531
chr7_28594857_28594937_ENSMUSG000000054808.16_-	0.733668	0.816609	0.513761	0.665185

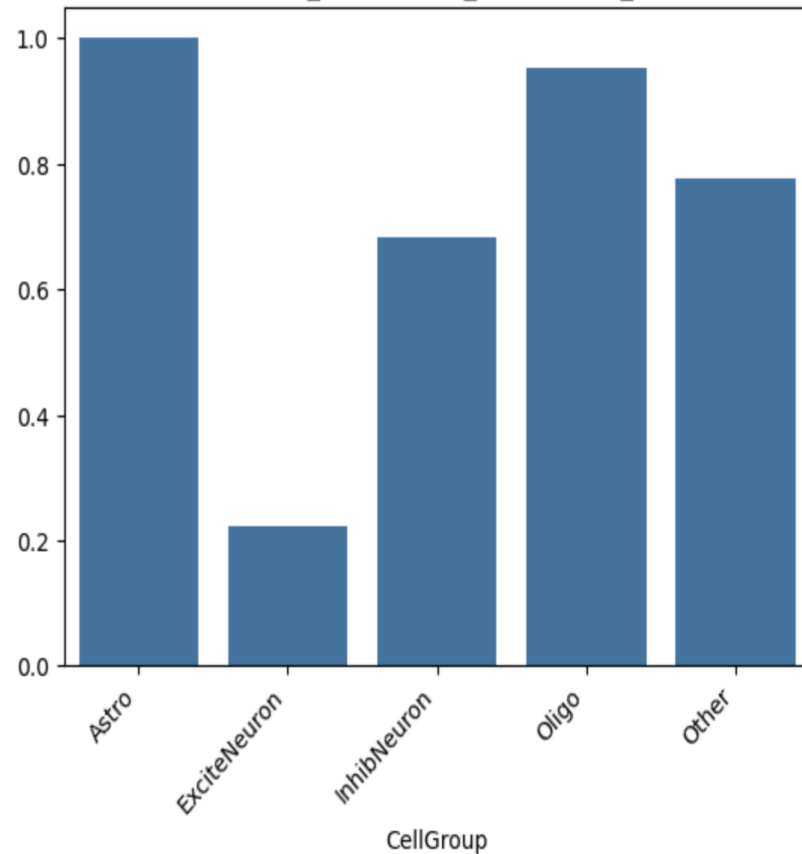


# Result

- Check PSI score per cell type for exons with F1 inclusion and exclusion  $> 0.5$

Example: Cell + Neighbour

Average PSI per Cell Type for chr12\_112578982\_112579038\_ENSMUSG000000037679.10\_+



CellGroup	TP	FP	FN	TN
ExciteNeuron	5	13	19	71
Astro	2	0	0	0
Oligo	20	1	0	0
InhibNeuron	12	5	3	2
Other	230	32	30	42

- PSI vary by cell type

# Result

Top 20 genes contributing to the classification of exons

- ENSMUSG00000053310:  
neurogranin – neuronal signalling
- ENSMUSG00000074129:  
Large ribosomal subunit protein uL13, rplM – protein assembly

Cell +  
Neighbour

Gene	Exon number
ENSMUSG00000053310	167
ENSMUSG00000074129	102
ENSMUSG00000093672	81
ENSMUSG00000078942	78
ENSMUSG00000021609	72

Cell

Gene	Exon number
ENSMUSG00000053310	496
ENSMUSG00000074129	285
ENSMUSG00000031765	262
ENSMUSG00000092341	133
ENSMUSG00000026568	116

Neighbour

Gene	Exon number
ENSMUSG00000053310	188
ENSMUSG00000074129	110
ENSMUSG00000093672	80
ENSMUSG00000078942	77
ENSMUSG00000097712	73

# Conclusion

- Exon inclusion / exclusion is mainly controlled by cell-intrinsic splicing regulators with some/little influence from microenvironment
- Cell type dependent alternative splicing
- Neuronal activation and metabolic state influence splicing dynamics
  - Microenvironment signalling
- Next steps: train on individual cell type

# Model – Excitatory neuron only

- X dataframe: rows are cells, columns are genes (29800 \* 3000)
  - Highly variable genes
  - Gene expression level of a cell
- Y dataframe: rows are cells, columns are exons (29800 \* 5001)
  - PSI (Percent Splice-In) value: proportion of mRNAs that include the exon
- Three models
  - Cell + Neighbour
  - Cell
  - Neighbour

# Result

Top 5 performing  
exons based on F1  
mean

- Cell + Neighbour  
the best
- Worse than non  
cell type specific  
models

## Cell + Neighbour

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr3_90409423_90409560_ENSMUSG00000001017.17_-	0.698630	0.765957	0.576923	0.671440
chr11_4861435_4861470_ENSMUSG000000034274.12_+	0.680000	0.500000	0.764706	0.632353
chr1_127869331_127869351_ENSMUSG000000036104.13_+	0.630631	0.623853	0.637168	0.630511
chr9_13723929_13724006_ENSMUSG000000031922.13_-	0.635135	0.689655	0.557377	0.623516
chr14_52241430_52241573_ENSMUSG000000004562.17_+	0.620690	0.620690	0.620690	0.620690

## Cell

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr11_95182346_95182435_ENSMUSG000000038909.18_-	0.694915	0.780488	0.500000	0.640244
chr18_42668150_42668212_ENSMUSG000000024498.17_+	0.662791	0.539683	0.733945	0.636814
chr11_64896280_64896331_ENSMUSG000000033389.17_-	0.676923	0.764045	0.487805	0.625925
chr12_8749064_8749188_ENSMUSG000000020594.15_+	0.694915	0.790698	0.437500	0.614099
chr19_27900608_27900666_ENSMUSG0000000040929.18_-	0.623377	0.674157	0.553846	0.614002

## Neighbour

	Accuracy	F1 Exclusion	F1 Inclusion	F1 mean
Exon				
chr1_80481636_80481711_ENSMUSG000000038608.16_-	0.676923	0.758621	0.511628	0.635124
chr1_127869331_127869351_ENSMUSG000000036104.13_+	0.621622	0.625000	0.618182	0.621591
chr2_165694788_165694799_ENSMUSG000000039671.19_-	0.622642	0.642857	0.600000	0.621429
chr9_13723929_13724006_ENSMUSG000000031922.13_-	0.621622	0.674419	0.548387	0.611403
chr14_52241430_52241573_ENSMUSG000000004562.17_+	0.609195	0.622222	0.595238	0.608730

# Result

Top 20 genes contributing to the classification of exons

- Cell + Neighbour
- Cell
- Neighbour
- ENSMUSG00000053310: neurogranin – neuronal signalling
  - ENSMUSG00000049630: C1ql3 - regulate the number of excitatory synapses that are formed on hippocampus neurons
  - ENSMUSG00000074129: Large ribosomal subunit protein uL13, rplM – protein assembly

Gene	Exon number
ENSMUSG00000053310	74
ENSMUSG00000049630	59
ENSMUSG00000114131	42
ENSMUSG00000058470	41
ENSMUSG00000064023	37

Gene	Exon number
ENSMUSG00000053310	265
ENSMUSG00000074129	83
ENSMUSG00000092341	65
ENSMUSG00000049630	62
ENSMUSG00000025658	52

Gene	Exon number
ENSMUSG00000053310	86
ENSMUSG00000049630	43
ENSMUSG00000114131	41
ENSMUSG00000079042	40
ENSMUSG00000035456	37

# Conclusion

- Inside one cell type, gene expression variation is much smaller → not good for logistic model (need Ridge/LASSO)
  - Some splicing differences are cell-type specific. But in one cell type, the difference is excluded, only internal / microenvironment influence is preserved
- Fewer training samples: higher variance in coefficient estimates
- But same genes are dominating the prediction of inclusion / exclusion in excitatory neurons as full dataset

# ShinyApp - SpIIsoFind Explorer

- Purpose: Allow users to freely discover the spatial isoform data for 2 samples (whole-exome, enriched)
- Instruction page
- Explorer tab
  - SpIIsoFind function from Python
  - Input Ensembl ID / Name, auto suggestion
  - If no information in one sample: Transcript xxx not found in this sample
  - If data/number of cell  $< 10$  for selected cell type/region: No data for transcript xxx in cell type xxx (in region xxx)