Discovering Spatial Gene Expression Patterns

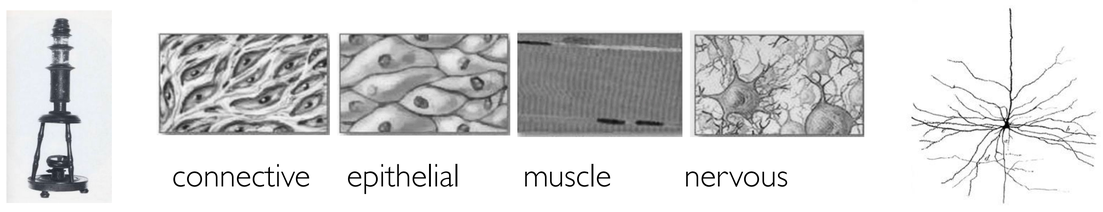
STANLEY CENTER FOR PSYCHIATRIC RESEARCH AT BROAD INSTITUTE

BROAD INSTITUTE

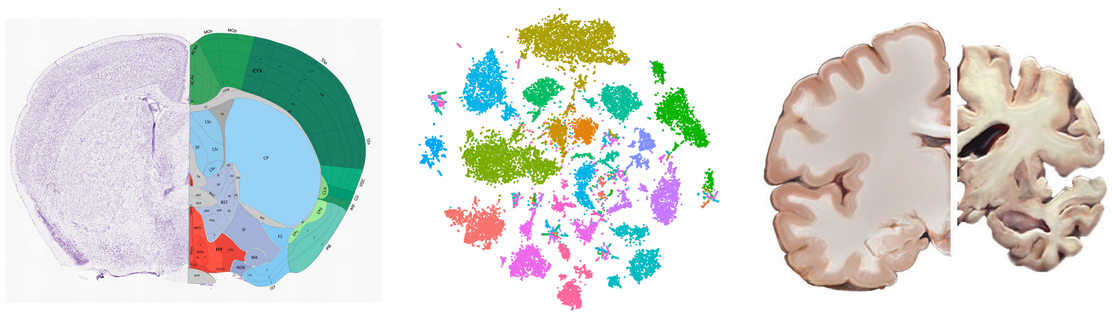
Aleksandrina Goeva • Macosko Lab • Broad Institute of MIT and Harvard

Why does this matter?

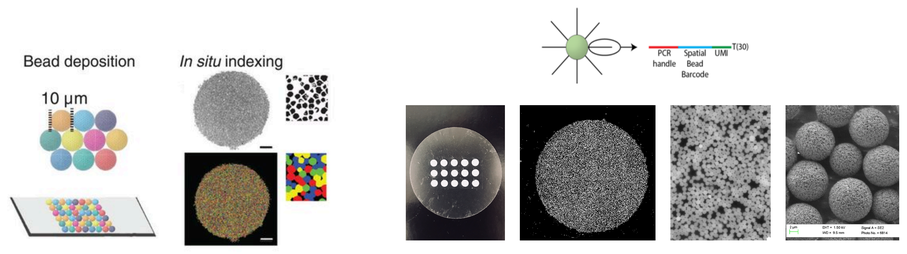
The organization of tissues is not spatially random. Each cell type caries a specialized function and paints a distinct pattern in space.

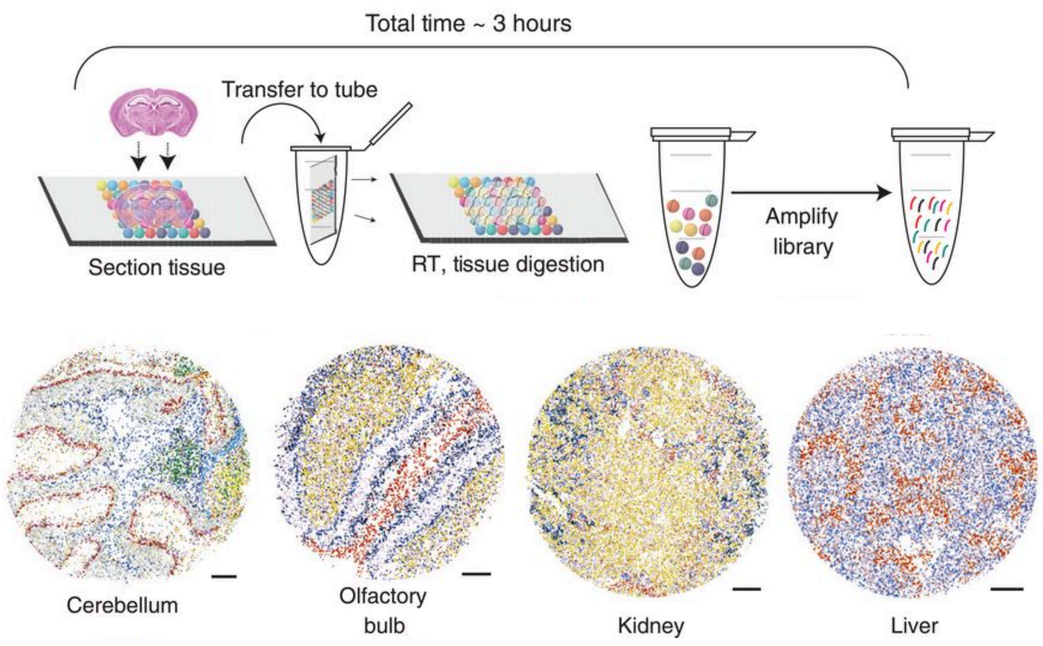


Nowadays, we can perform single-cell RNA-seq on a tissue of interest and get molecular knowledge about both healthy and disease affected cell types. However, the spatial context is lost.



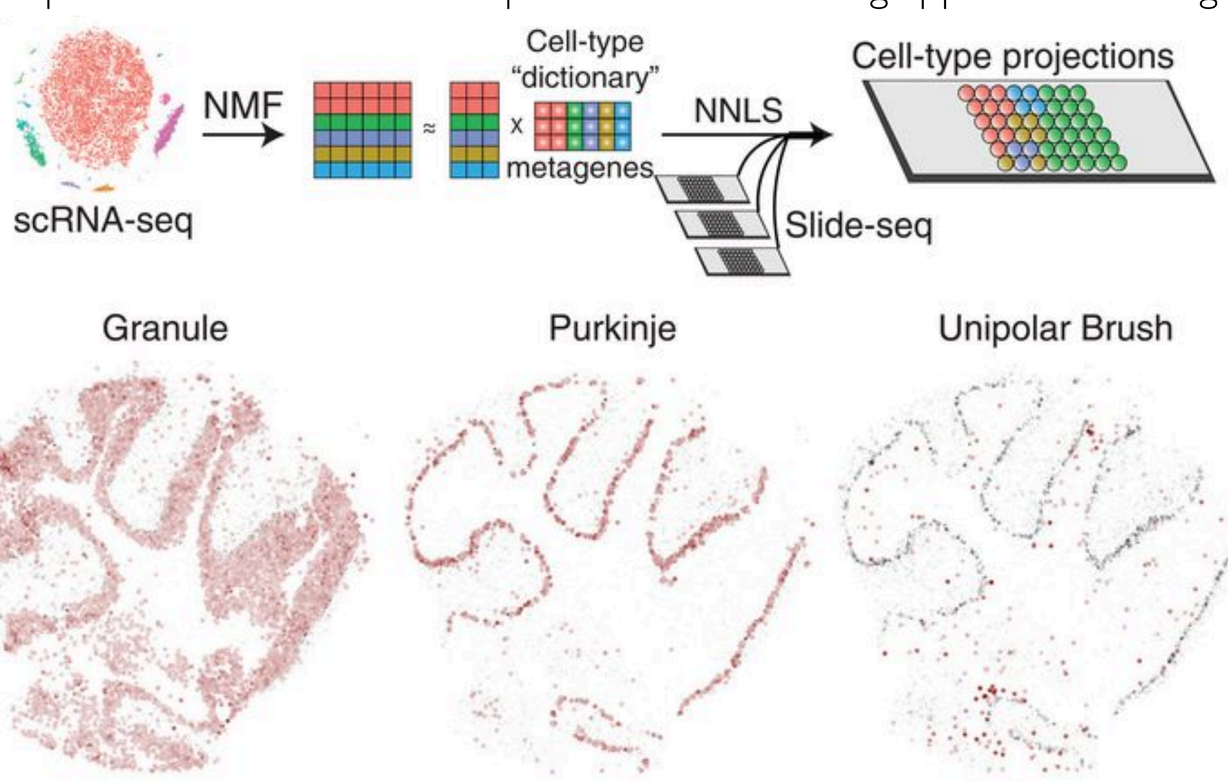
Slide-seq Data[1]





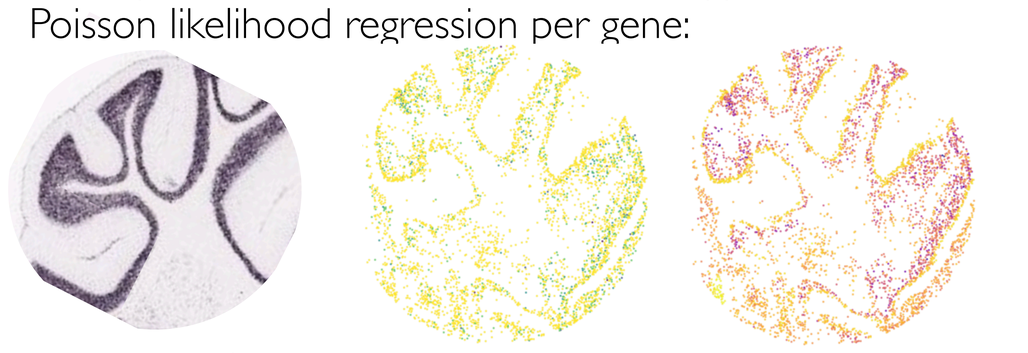
Where in space are the known cell types?

Slide-seq beads contain potentially composite measurements of gene expression from multiple neighboring cells. Often, we have access to a corresponding labeled single-cell reference. To deconvolve the Slide-seq measurements we developed a transfer learning approach: NMFreg.



Gaussian process regression

A GP prior with smoothness over cell type and Poisson likelihood regression per gene:



Multi-output GP regression

$$f_k : \mathbb{R}^2 \to \mathbb{R}, \quad k = 1, \ldots, K$$

$$f_k \sim GP$$

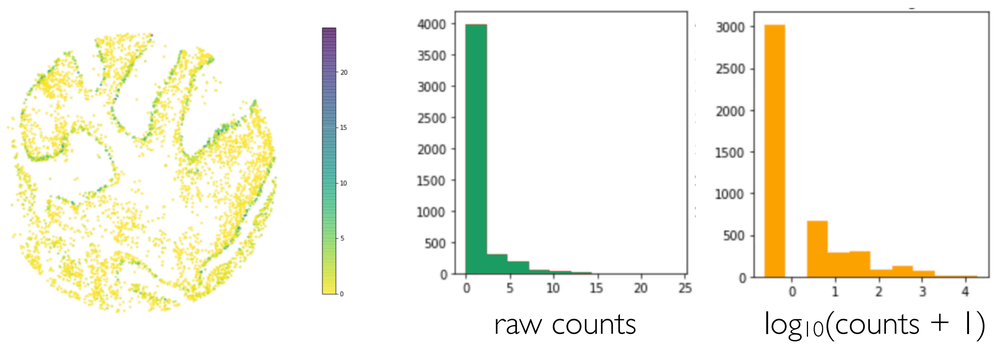
$$\underset{G\times 1}{\vec{y}} \approx \underset{G\times K}{W} \cdot \underset{K\times 1}{\vec{f}}$$

Learn K spatially smooth functions and a mixing matrix W. The non-Gaussianity and the high number of genes necessitate the use of SVGPs.

De novo unsupervised discovery of gene patterns and cell clusters

There are several challenges:

- the data is **sparse**, and **non-Gaussian**



- **high-dimensional**, and **large**: 20K genes across ~50K beads;
- existing models for clustering gene expression, such as matrix factorization, do **not** take the spatial information into account;
- the problem of defining spatial gene expression patterns is highly unsupervised, with **limited ground truth** domain knowledge for validation making it hard to efficiently evaluate the quality of putative models.

What else can we try?

Can we integrate the spatial information into common DL and RL methods while aiming for approaches that start from simple principles and have biologically-interpretable structure?

- VAEs?
- CNNs?
- RL to incorporate domain knowledge?
- other?

References and Contact

1. Rodriques, S.G., Stickels, R.R., Goeva, A., Martin, C.A., Murray, E., Vanderburg, C.R., Welch, J., Chen, L.M., Chen, F. and Macosko, E.Z., 2019. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, *363*(6434), pp. 1463-1467.