

# Comp790-166: Computational Biology

## Lecture 13

February 27, 2023

# Today

- Finish geometric sketching
- Sketching with hopper
- Start trajectory inference

## Project Related Announcements

- Please sign up with your group members and topics here,  
[https://docs.google.com/document/d/1x9mIJCZAkeogAhmGlpqJkXwuoAK1B\\_0gewV1LpXZDoU/edit?usp=sharing](https://docs.google.com/document/d/1x9mIJCZAkeogAhmGlpqJkXwuoAK1B_0gewV1LpXZDoU/edit?usp=sharing)
- Writeup due before spring break on March 8.
- Presentations will be the week after spring break. Stay tuned for your date and time.

## Do You Remember Question

- ① What is a sketch of single-cell data? What are some useful properties a sketch should have?
- ② What is the main idea of Geometric Sketching?

# Sketching Algorithms

In general a sketching algorithm takes a dataset and compresses it, such that you can still effectively carry out a query that you wanted to do on the original data.

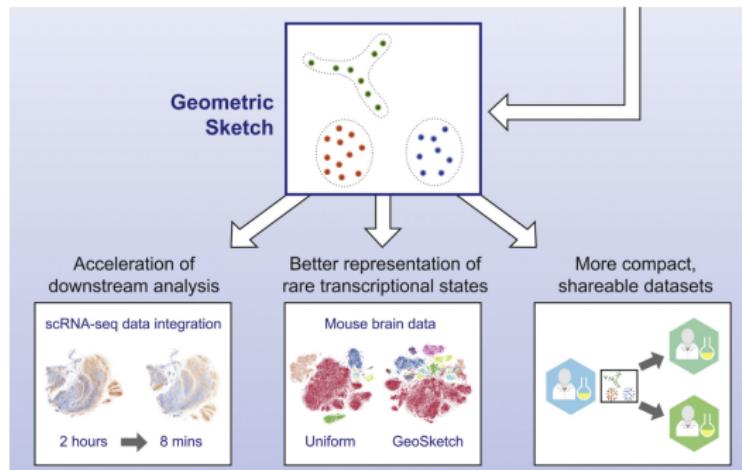


Figure: from Fig. 2 of Hie et al. Cell Systems. 2019. Perhaps we need an accurate estimate of cell-population frequencies, etc.

## Formulation of Sketching Problem

You can measure the quality of a particular sketch,  $\mathcal{S}$  wrt a dataset,  $\mathcal{X}$  through Hausdorff Distance as,

$$d_H(\mathcal{X}, \mathcal{S}) = \max_{\mathbf{x} \in \mathcal{X}} \left\{ \min_{\mathbf{s} \in \mathcal{S}} d(\mathbf{x}, \mathbf{s}) \right\}$$

Here,  $d$  is any distance or dissimilarity measure that you can compute between  $\mathbf{x}$  and  $\mathbf{s}$ .

# A Robust Formulation of Hausdorff Distance

The regular formulation of Hausdorff distance is sensitive to outliers. An alternative partial HD measure is defined as follows,

$$d_{HK}(\mathcal{X}, \mathcal{S}) = K^{\text{th}}_{\mathbf{x} \in \mathcal{X}} \left\{ \min_{\mathbf{s} \in \mathcal{S}} d(\mathbf{x}, \mathbf{s}) \right\}$$

This looks at the  $K$ th largest distance. Define  $q = 1 - K/|\mathcal{X}|$ . The  $q$  can therefore be varied.

# Plaid Covering

The plaid covering,  $\mathcal{C}$  represents a collection of  $m$ -dimensional equal volume hypercubes. The same number of cells are then sampled from each hypercube. In practice, they define the same number of hypercubes as desired cells ( $k$ ) and therefore sample 1 cell per cube.

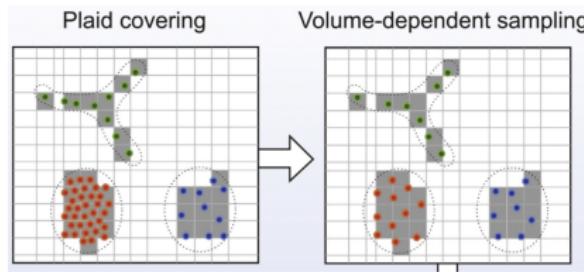


Figure: from Fig. 2

# Downsampled Cells Facilitate Faster Downstream Tasks

for example : batch effect correction

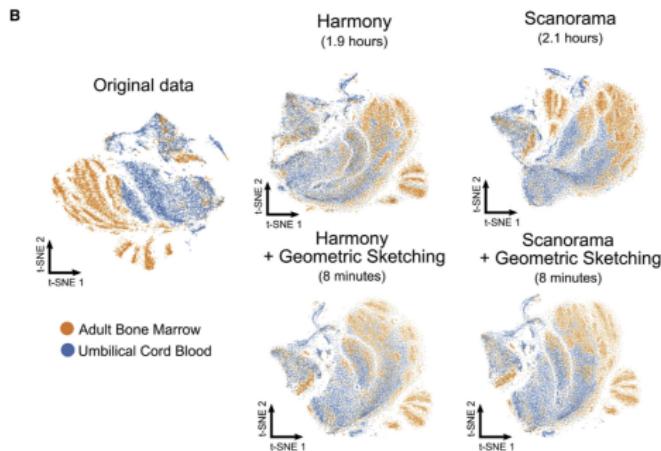


Figure: from Fig. 5

# Rarer Populations are Sufficiently Represented

Check out (for example) Ependymal (cells colored brown)

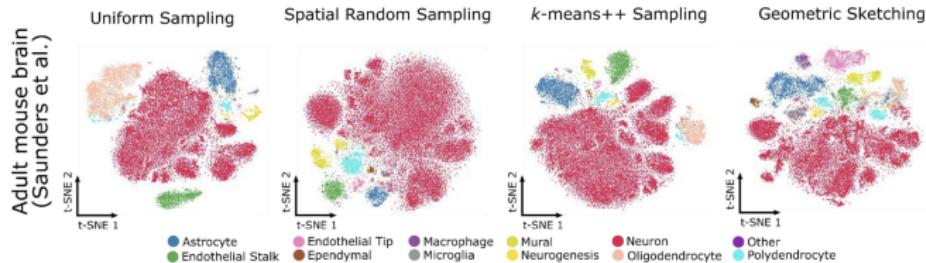


Figure: from Supplementary Figure S1

# Counts of Rarest Cell Type

In sketches containing 2% of the dataset, the methods were compared in terms of their representation of the rarest cell-type.

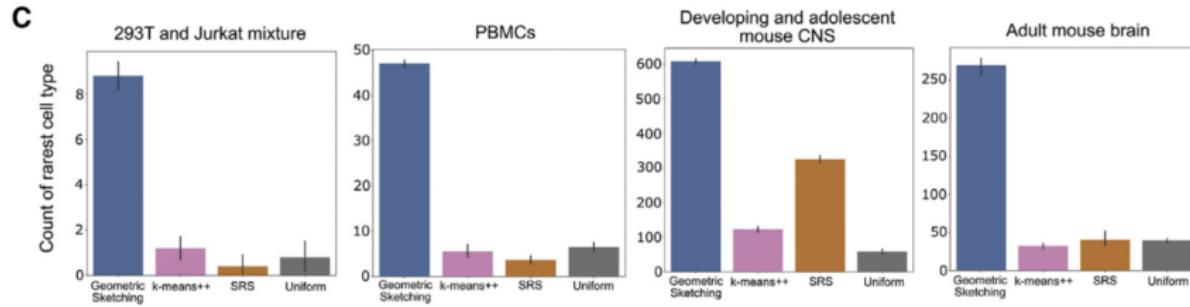


Figure: from Fig. 3

# Interpretation wrt Clustering

Louvain on downsampled cells produced the ability to differentiate between inflammatory macrophage and macrophage, and to observe appropriate gene expression differences.

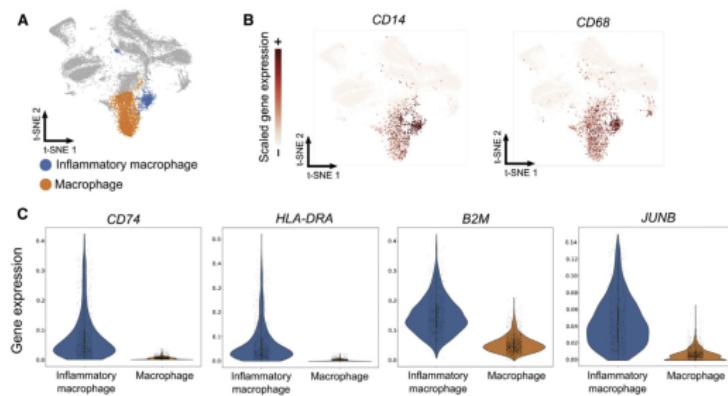


Figure: from Fig. 4

## Recap and Unsolicited Opinions

- Beautiful concept. Similar to what we have seen with Cydar
- Through the plaid covering, a subset of cells can be sampled to create a quality sketch (in terms of Hausdorff Distance)
- The reduced set of cells has good representation across many cell types, including those that are rare.
- **Still Missing:** ??

# Welcome Hopper

Hopper is by the same authors, but takes a slightly different approach. With a goal to find a sketch that minimizes Hausdorff distance, use a greedy approximation (farthest first traversal) to the  $k$ -center problem. Starting with a randomly chosen point, find the point  $p$  such that is furthest away from any of the previously chosen points.

$$p = \arg \max_{x \in X} \left( \min_{s \in S} d(x, s) \right)$$

## Identification of the Time-Consuming Part

- The computationally expensive aspect of the farthest-first traversal is identifying each of the new points (or the ‘ $p$ ’) from the previous slide
- For each  $x \in X$ , we need to compute a distance to each of the points in the set  $S$ .
- Each time a point is added to  $S$  distances must be updated.

Assume that a newly added point,  $p$  has distance  $r$  to its nearest representative in  $S$ . Then by the triangle inequality,

$$r \leq d(s, p) \leq d(s, x) + d(x, p)$$

## Using Triangle Inequality

With  $r \leq d(s, p) \leq d(s, x) + d(x, p)$  by the triangle inequality then,

- In particular  $d(x, p) \leq d(s, x)$ , then  $d(s, x) \geq \frac{r}{2}$
- Therefore, only examine points in  $X$  with distance  $\geq \frac{r}{2}$  to their nearest points in  $S$

# Transition and Contrast to Geometric Sampling

- In geometric sketching, partitions were all hypercubes of the same size and points were drawn from each
- Hopper allows partitions to occupy variable-sized regions of transcriptional space and draws variable number of points from the partitions.

## Further Speedups → TreeHopper

- Points can be split using k-d tree, or your favorite splitting approach
- In the paper, they suggest to sequentially split cells according to the leading PC
- Do Hopper operation on each subset of points
- The more you pre-process via splitting, the faster it will be.

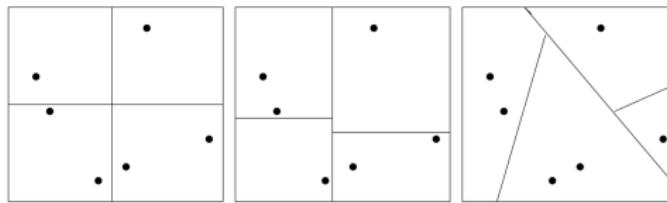


Figure: from <https://arxiv.org/pdf/1205.2609.pdf>

# Hausdorff Distances

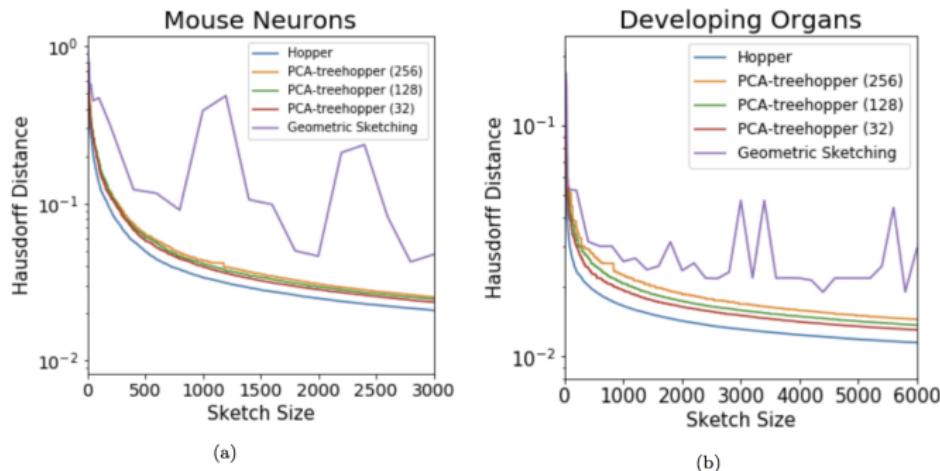
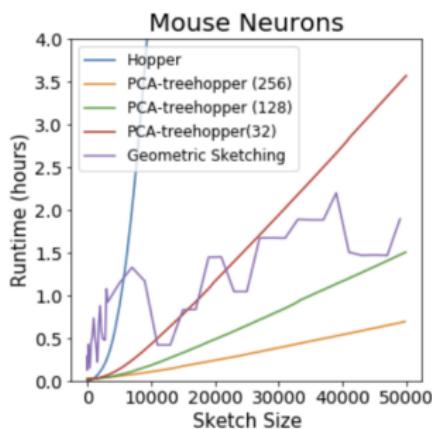


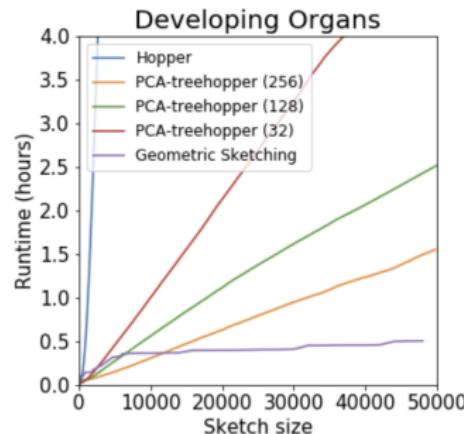
Figure: from Fig. 1 of deMeo *et al.* Bioinformatics. 2019.

# Runtimes

Run-times on two datasets. Geometric sketching is more sensitive to the geometry of the dataset.



(a)



(b)

Figure: from Fig. 2 of deMeo *et al.* Bioinformatics. 2019.

# 5,000 point Hopper Sketch

Even with a sketch containing only 5,000 points, Hopper can still lead to clusters via Louvain corresponding to rare cell-types.

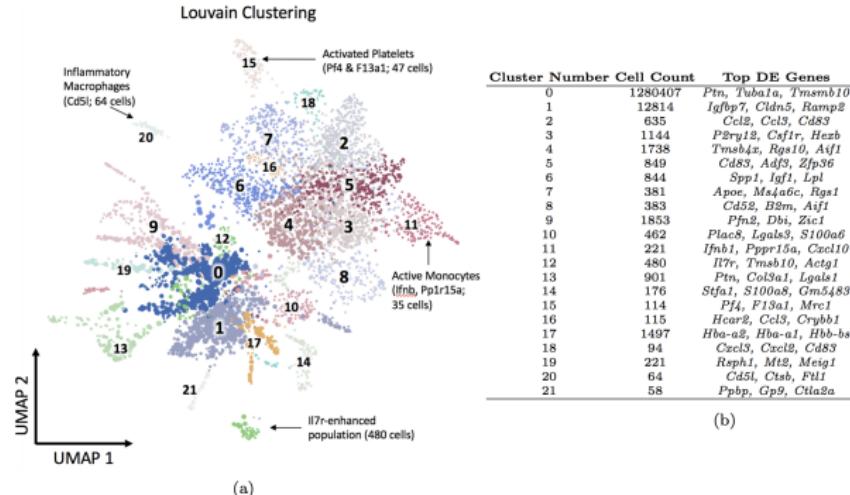


Figure: from Fig. 3 of deMeo et al. Bioinformatics. 2019.

# Full Dataset

Without sketching, some specialized cell types were more obscured (green group).

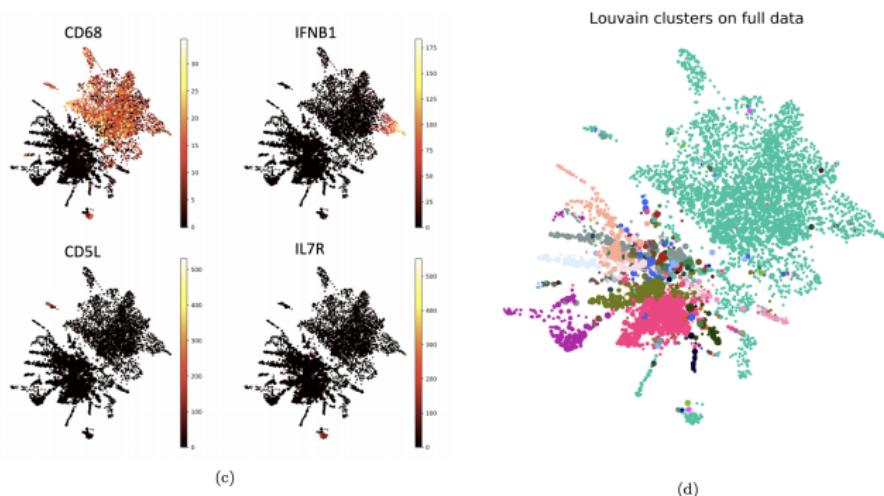


Figure: from Fig. 3 of deMeo *et al.* Bioinformatics. 2019.

# Hopper vs GeoSketch

In comparison to hopper approaches, geometric sketching tends to create more clusters that have occurred at grid intersections.

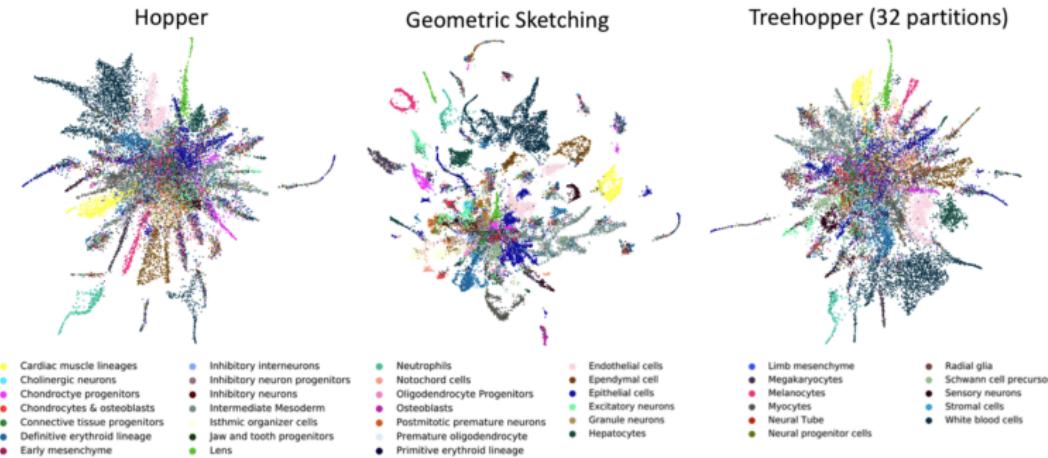


Figure: Fig. 3 of deMeo *et al.* Bioinformatics. 2019. Hopper sketches overall give better classification accuracy than GeoSketch.

# Sketching algorithms in the wild....

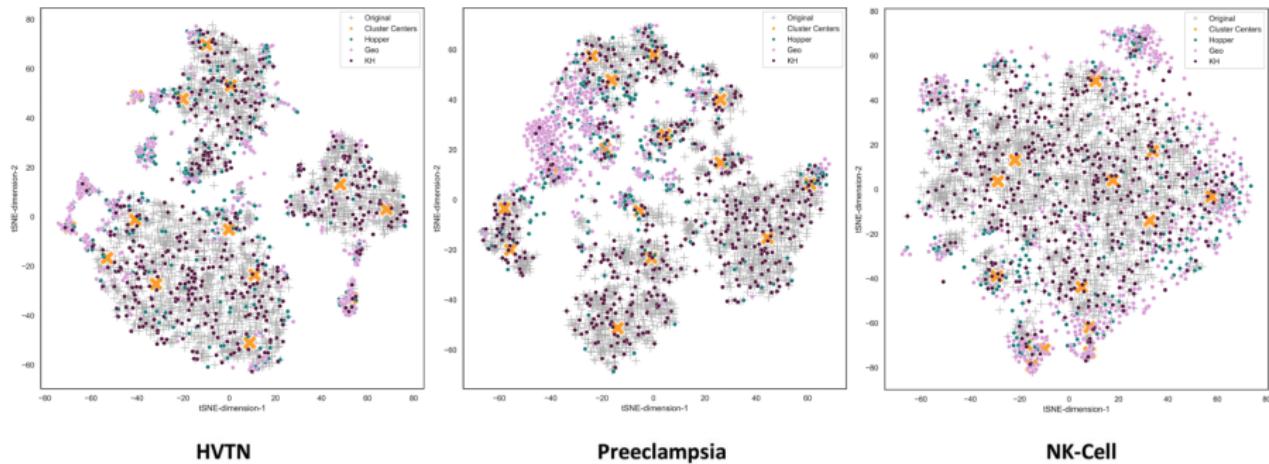


Figure: <https://arxiv.org/pdf/2207.00584.pdf>

# Task-driven evaluation

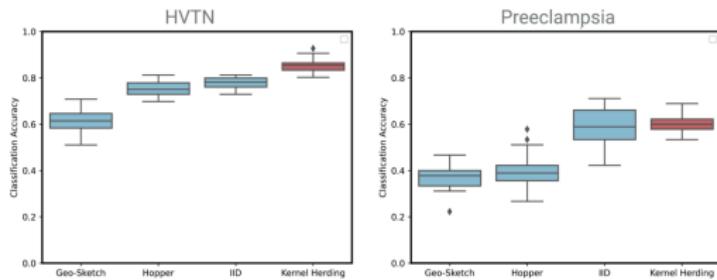


Figure: <https://arxiv.org/pdf/2207.00584.pdf>

# Geometric Sketching

C.

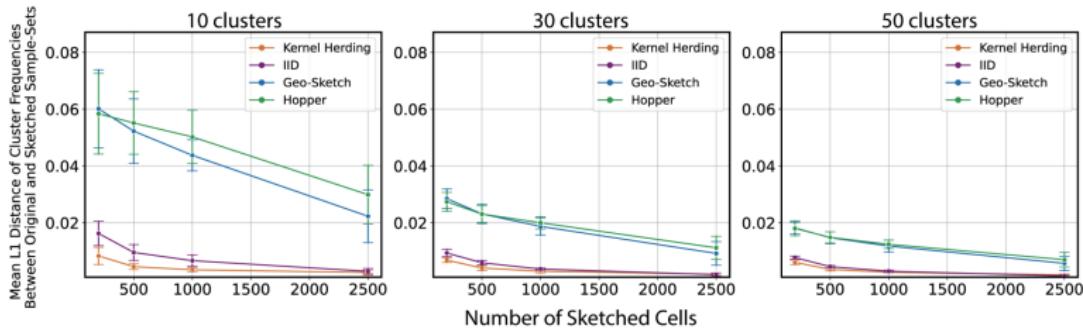


Figure: <https://arxiv.org/pdf/2207.00584.pdf>. Geometric Sketching is overall better at preserving frequencies of cell-types.

## Conclusion and Wrap up

- Hopper uses fastest first traversal to produce sketches. Further accelerated run-time by tree-based splitting
- Both algorithms ensure that rare cell types are still accounted for.
- Evidence that hopper can potentially uncover novel cell-types that were not seen with the entire data.

# Entering Pseudotime and Trajectory Inference

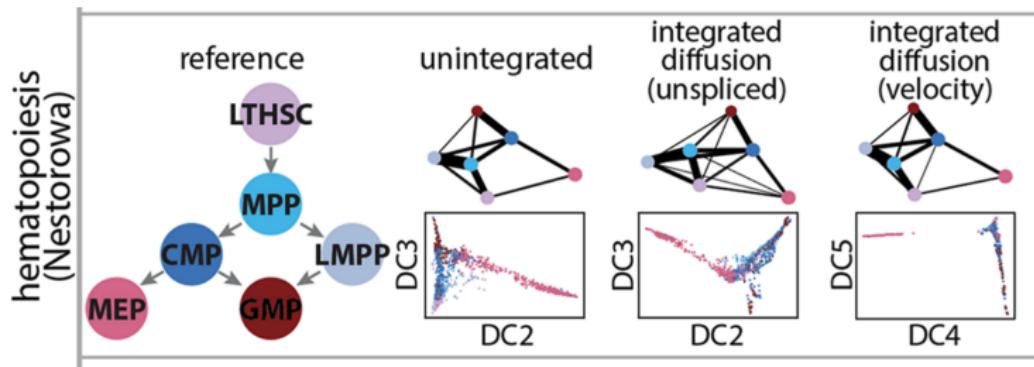


Figure: from Raneket al. *Genome Biology*. 2022. We would like to establish a chronological ordering of cells that represents their differentiation. This will allow us to uncover gene programs associated with differentiation.

# Welcome Back Diffusion Maps

Let's set up some notation as follows :

- Let  $\Omega$  be the set of all measured cells
- Suppose that we have measured  $n$  total cells, each with  $G$  features, so a given  $\mathbf{x}_i \in \mathbb{R}^G$ .

Allow each cell to diffuse around its measured position as,

$$Y_{\mathbf{x}}(\mathbf{x}') = \left( \frac{2}{\pi\sigma^2} \right)^{1/4} \exp \left( -\frac{\|\mathbf{x}' - \mathbf{x}\|^2}{\sigma^2} \right)$$

Here the  $\sigma^2$  determines the scale over which each cell can randomly diffuse.

## Modeling the Probability cell x transitioning to cell y

Define the transition matrix,  $\mathbf{P}$  for all pairs of cells as,

$$P_{xy} = \frac{1}{Z(x)} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$
$$Z(x) = \sum_{y \in \Omega} \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$$

Here  $Z(x)$  is a partition function providing an estimate for the number of neighbors of  $x$  in a certain volume defined by  $\sigma$ , or the density of cells at that proximity.

## Redefine a Density Normalized Transition Probability Matrix, $\tilde{\mathbf{P}}$

We are interested in the transition probabilities between cells and not the on-cell potentials imposed by local density. Therefore, the diagonal of  $\tilde{\mathbf{P}}$  will be set to 0 and  $\mathbf{y} = \mathbf{x}$  will be excluded from the sum in the partition function,  $\tilde{Z}(\mathbf{x})$ . Without some correction, it may look like cells within a very dense region have a small diffusion distance.

$$\tilde{P}_{xy} = \frac{1}{\tilde{Z}(\mathbf{x})} \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)}{Z(\mathbf{x})Z(\mathbf{y})}, \quad \tilde{P}_{xx} = 0$$
$$\tilde{Z}(\mathbf{x}) = \sum_{\mathbf{y} \in \Omega / \mathbf{x}} \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)}{Z(\mathbf{x})Z(\mathbf{y})}$$

# Using Eigenvectors of $\tilde{\mathbf{P}}$

- $\tilde{\mathbf{P}}$  has  $n$  ordered eigenvalues,  $\lambda_0 = 1 > \lambda_1 \geq \dots \geq \lambda_{n-1}$
- Corresponding to these eigenvalues are eigenvectors  $\Psi_0 \dots \Psi_{n-1}$
- As we saw a few weeks ago, powering  $\tilde{\mathbf{P}}$  to  $\tilde{\mathbf{P}}^t$  represents the probability of transitioning between two cells with a walk of length  $t$ .

The diffusion distance between a pair of cells  $\mathbf{x}$  and  $\mathbf{y}$  can be written in terms of the eigenvectors of  $\tilde{\mathbf{P}}$  as,

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} (\psi_i(\mathbf{x}) - \psi_i(\mathbf{y}))^2$$

# Unpacking Diffusion Distance

$$D_t^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} \lambda_i^{2t} (\psi_i(\mathbf{x}) - \psi_i(\mathbf{y}))^2$$

- The eigenvector to the largest eigenvalue,  $\lambda_0$  is a constant vector,  $\Psi_0 = \mathbf{1}$ . Therefore, it contributes 0.
- The eigenvalues of  $\tilde{\mathbf{P}}$  determine the diffusion coefficients in the direction of the corresponding eigenvector
- After the first  $l$  prominent directions, the diffusion coefficients typically drop to a noise level.
- When you find the  $l$  such that there is a large difference between  $l$  and  $l + 1$  eigenvalues (an elbow), you can use the sum up to the  $l$ -th term as an approximation for diffusion distance. The first  $l$  eigenvectors correspond to the diffusion components.

# What we have just defined, illustrated

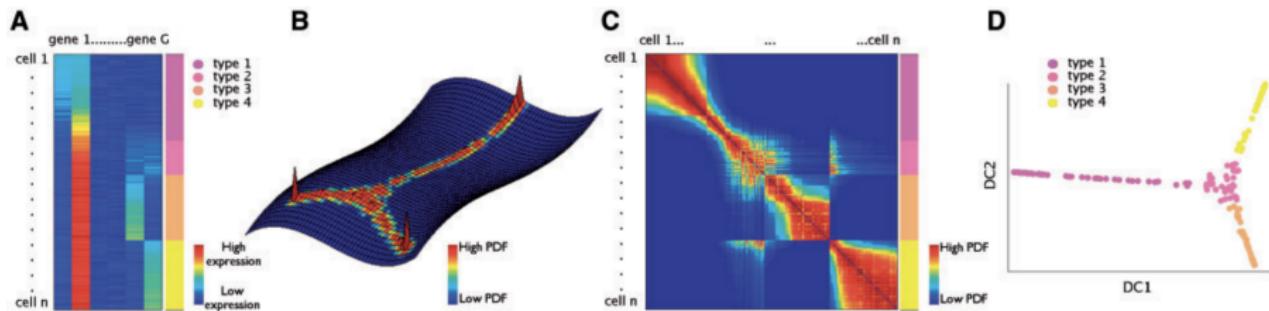


Figure: from Haghverdi *et al.* Bioinformatics. 2015. After projecting cells based on the first two diffusion components, it is still unclear what the time direction is.

# Example Applied to A Dataset of Differentiating Cells

The diffusion map is capturing transitions between cell types that are not reflected in PCA or tSNE.

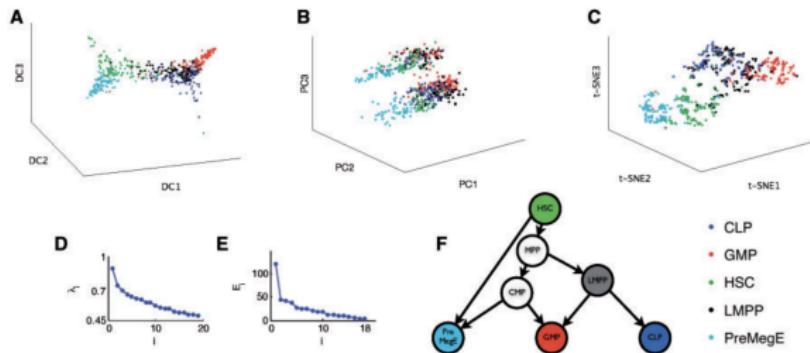


Figure: from Haghverdi et al. Bioinformatics. 2015.

This diffusion map approach was the beginning of thousands of people starting to think about cellular differentiation.....