Comp683: Computational Biology

Lecture 15

March 5, 2024

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/1x9mIJCZAkeogAhmGlpqJkXwuoAKlB_0gewV1LpXZDoU/edit?usp=sharing
- Project proposal materials here, https://github.com/natalies-teaching/Comp683_CompBio_2024/tree/main/Project_Proposal
- Project proposal due March 19

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.

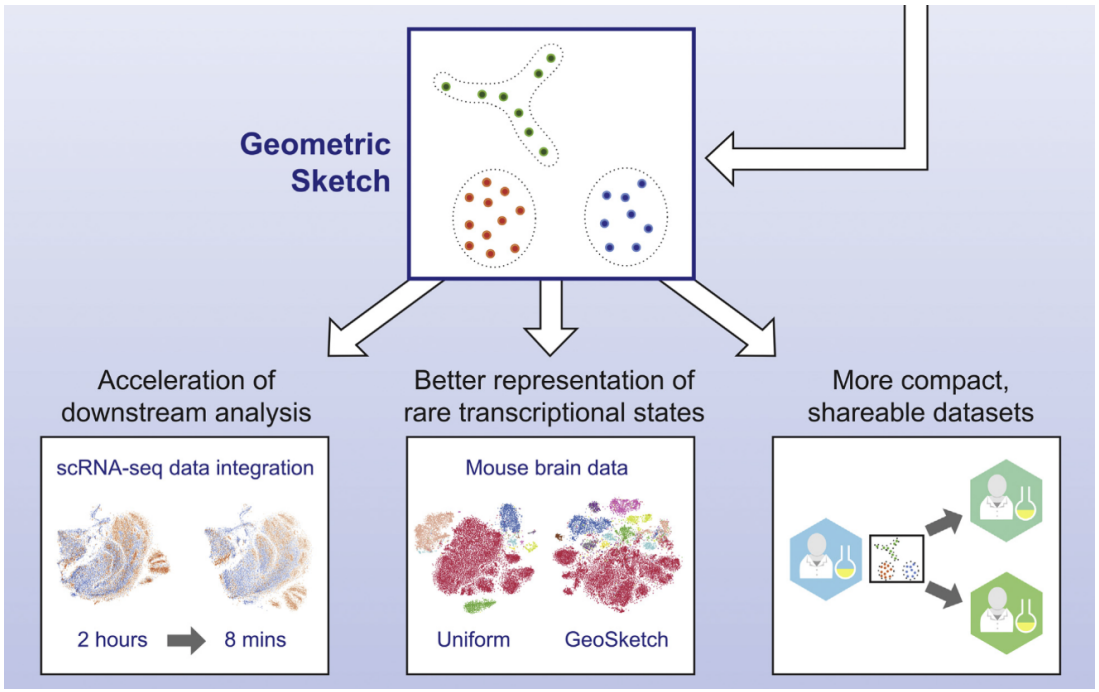


Figure: from Fig. 2 of Hie *et al.* Cell Systems. 2019. Perhaps we need an accurate estimate of cell-population frequencies, etc.

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 Kth 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.

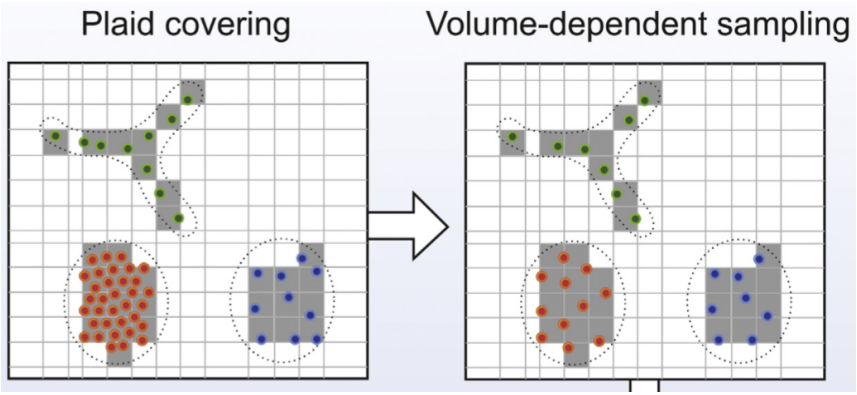


Figure: from Fig. 2

Parameters for Sketching

Geometric Sketching Algorithm Parameters

Parameter	Type	Default Value	Notes
Sketch size (k)	Integer between 0 and total number of cells, inclusive	N/A	The desired sketch size is chosen depending on the amount of compute resources available and the algorithmic complexity of downstream analyses; smaller sketches omit more cells but will accelerate analysis while preserving much of the transcriptional heterogeneity.
Number of covering boxes ($\lvert\mathcal{C}\rvert$)	Integer between 1 and total number of cells, inclusive	Equal to desired sketch size k	Converges to uniform sampling as parameter increases; a number of covering boxes less than k may yield a coarser picture of the transcriptional space, including overrepresentation of rare cell types, at the cost of an increased Hausdorff distance.

Figure: from Hie *et al.* Cell Systems. 2019.

Baseline Downsampling Method, k-means++

- Ref : `https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf`

k-means++ is an approach to better initialize k-means. The downsampling aspect is to find appropriate cluster centers. Cluster centers are added, such that, each addition is different from the previous center added.

which we call **k-means++**.

1a. Choose an initial center c_1 uniformly at random from $\mathcal{X}$.

1b. Choose the next center c_i, selecting $c_i = x' \in \mathcal{X}$ with probability $\frac{D(x')^2}{\sum_{x \in \mathcal{X}} D(x)^2}$.

1c. Repeat Step 1b until we have chosen a total of k centers.

Figure: $D(\cdot)$ is the smallest distance from a point and to any of the previously chosen cluster centers.

Downsampled Cells Facilitate Faster Downstream Tasks

for example : batch effect correction

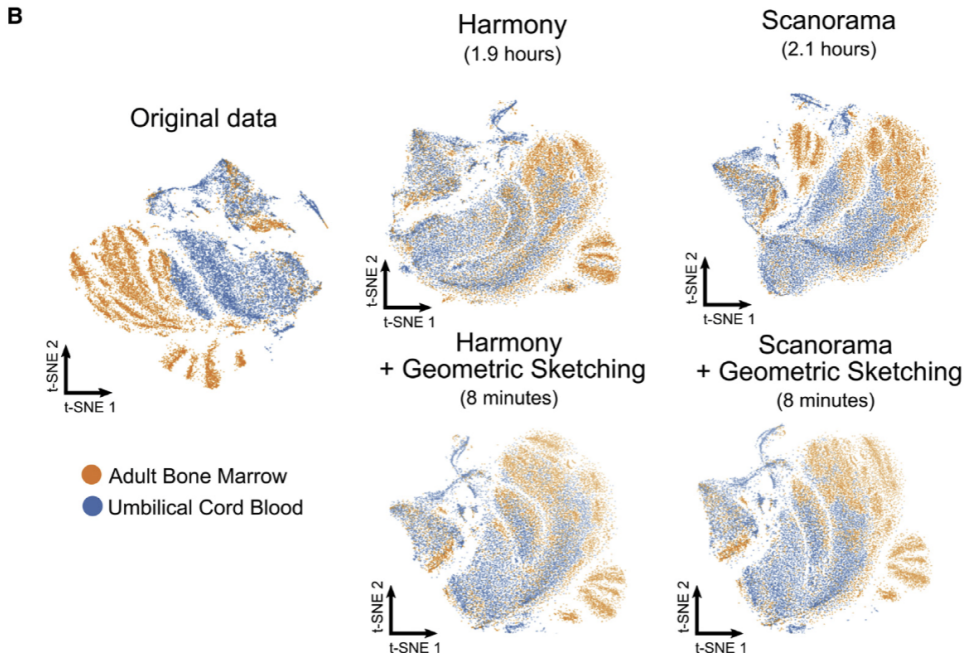


Figure: from Fig. 5

Rarer Populations are Sufficiently Represented

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

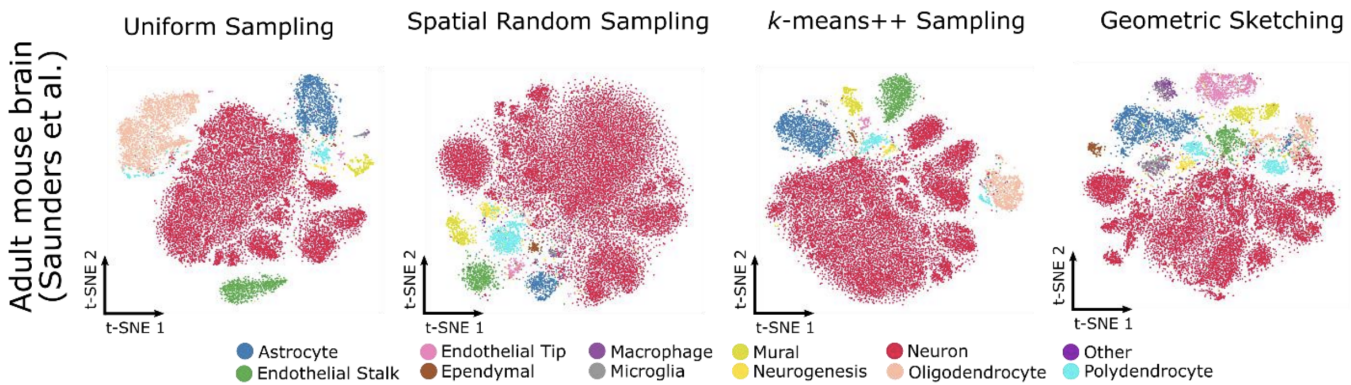


Figure: from Supplementary Figure S1

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.

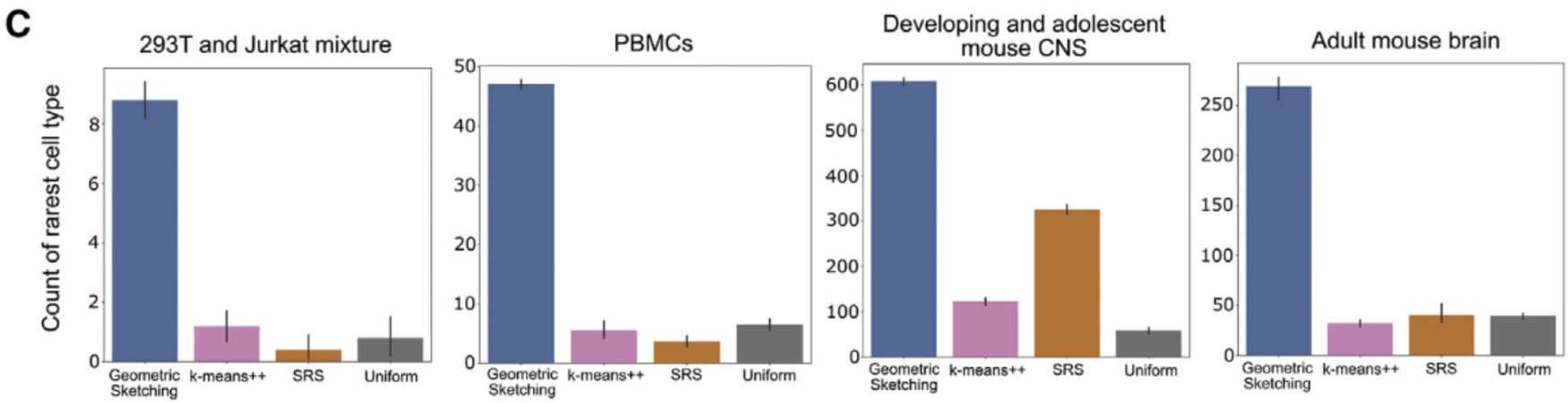


Figure: from Fig. 3

Interpretation wrt Clustering

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

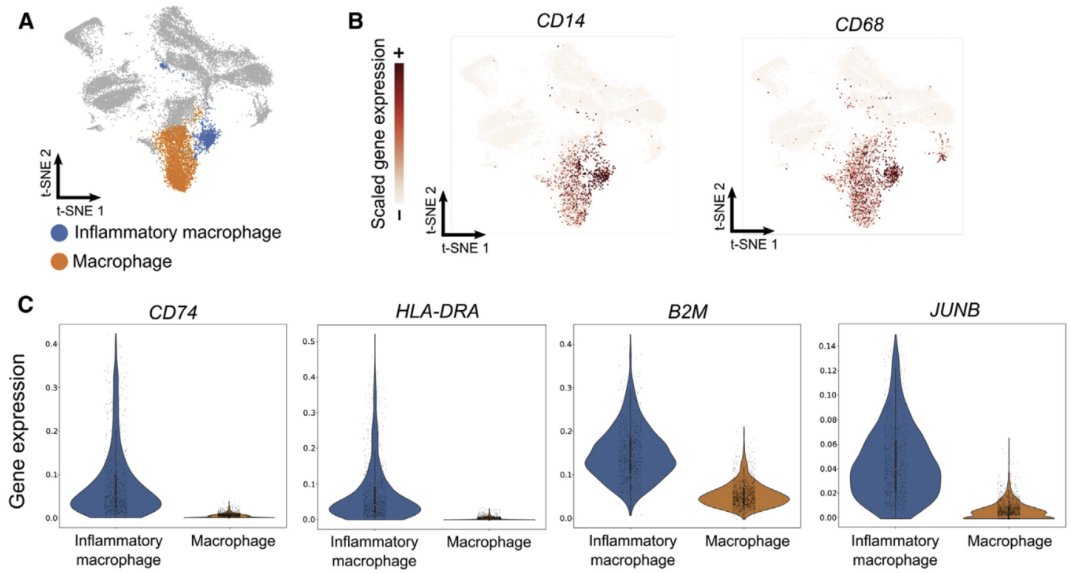


Figure: from Fig. 4

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 a 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 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 $\rightarrow$ 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.

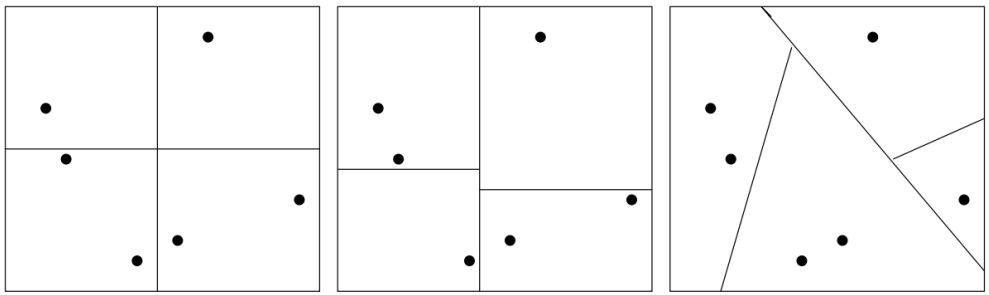


Figure: from `https://arxiv.org/pdf/1205.2609.pdf`

Hausdorff Distances

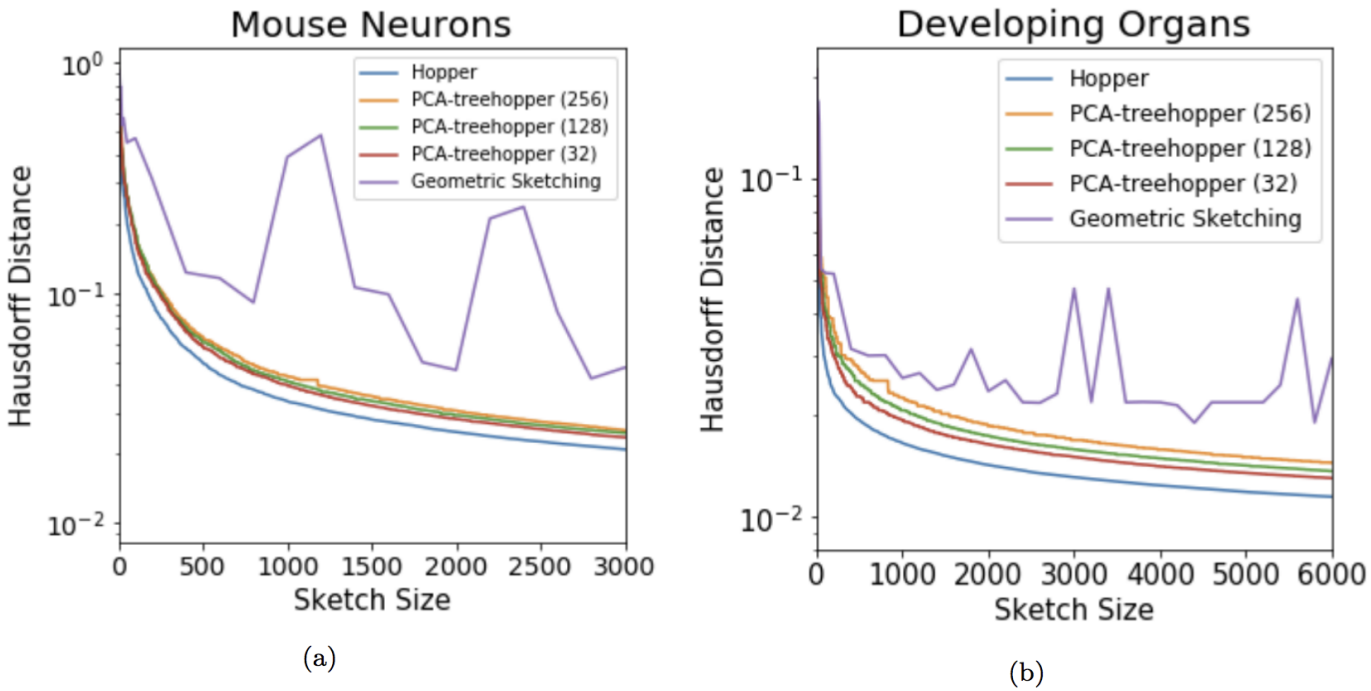
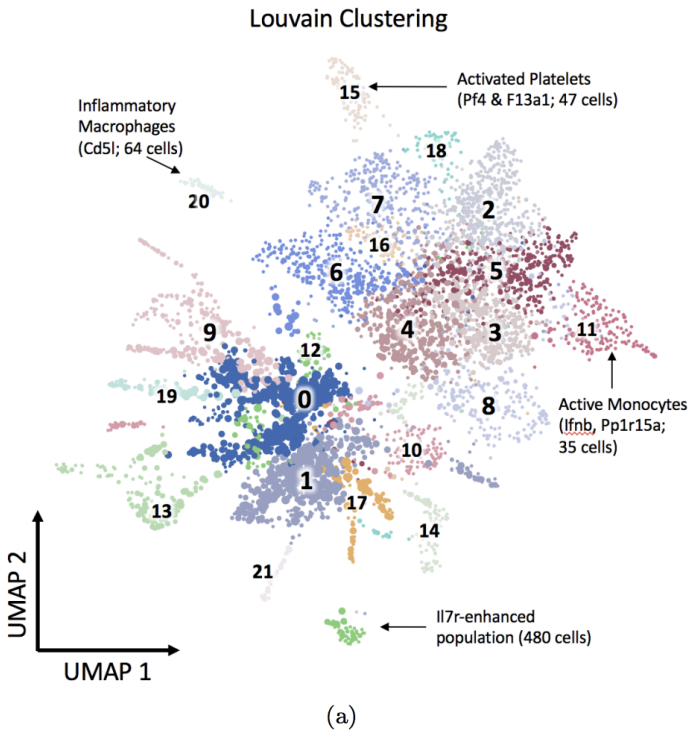


Figure: from Fig. 1 of deMeo *et al.* Bioinformatics. 2019.

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.



(a)

Cluster Number	Cell Count	Top DE Genes
0	1280407	*Ptn, Tuba1a, Tmsmb10*
1	12814	*Igfbp7, Cldn5, Ramp2*
2	635	*Ccl2, Ccl3, Cd83*
3	1144	*P2ry12, Csf1r, Hexb*
4	1738	*Tmsb4x, Rgs10, Aif1*
5	849	*Cd83, Adf3, Zfp36*
6	844	*Spp1, Igf1, Lpl*
7	381	*Apoe, Ms4a6c, Rgs1*
8	383	*Cd52, B2m, Aif1*
9	1853	*Pfn2, Dbi, Zic1*
10	462	*Plac8, Lgals3, S100a6*
11	221	*Ifnb1, Pppr15a, Cxcl10*
12	480	*Il7r, Tmsb10, Actg1*
13	901	*Ptn, Col3a1, Lgals1*
14	176	*Stfa1, S100a8, Gm5483*
15	114	*Pf4, F13a1, Mrc1*
16	115	*Hcar2, Ccl3, Crybb1*
17	1497	*Hba-a2, Hba-a1, Hbb-bs*
18	94	*Cxcl3, Cxcl2, Cd83*
19	221	*Rsph1, Mt2, Meig1*
20	64	*Cd5l, Ctsb, Ftl1*
21	58	*Ppbp, Gp9, Ctla2a*

(b)

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

Full Dataset

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

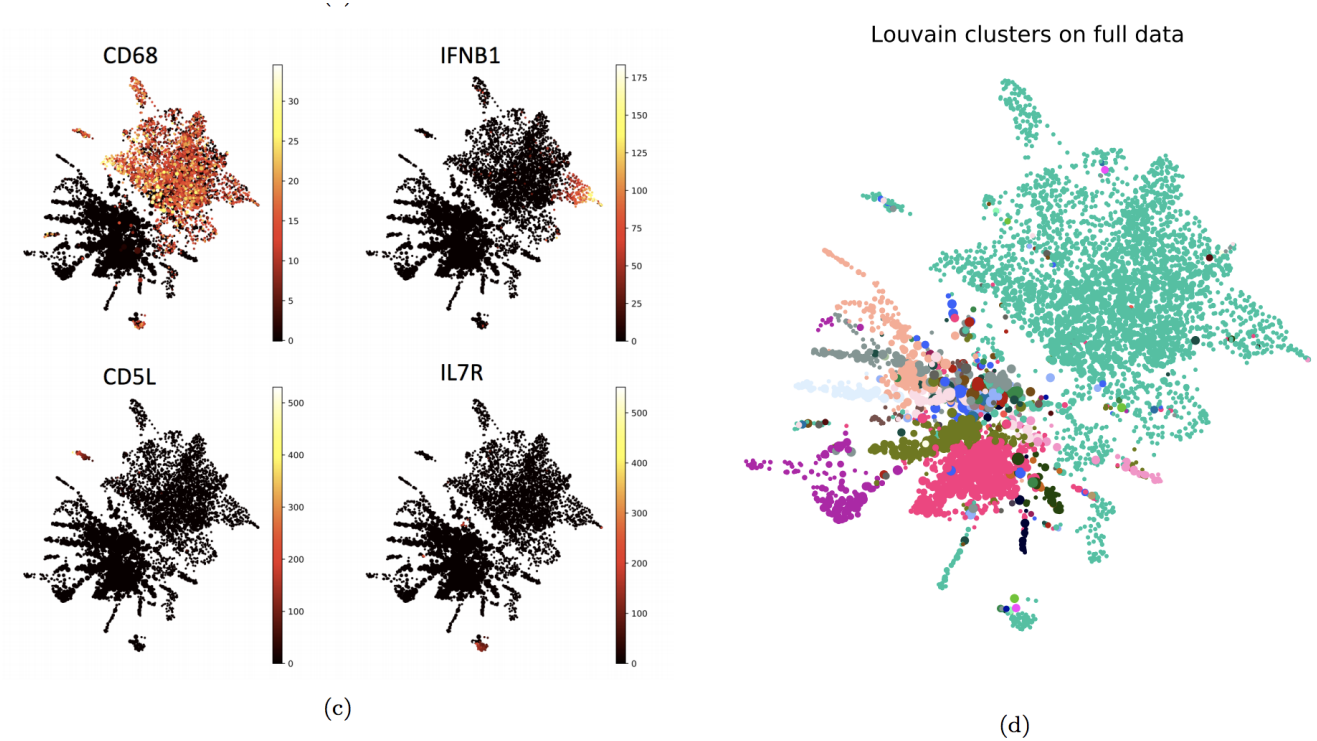


Figure: from Fig. 3 of deMeo *et al.* Bioinformatics. 2019.

Sketching algorithms in the wild....

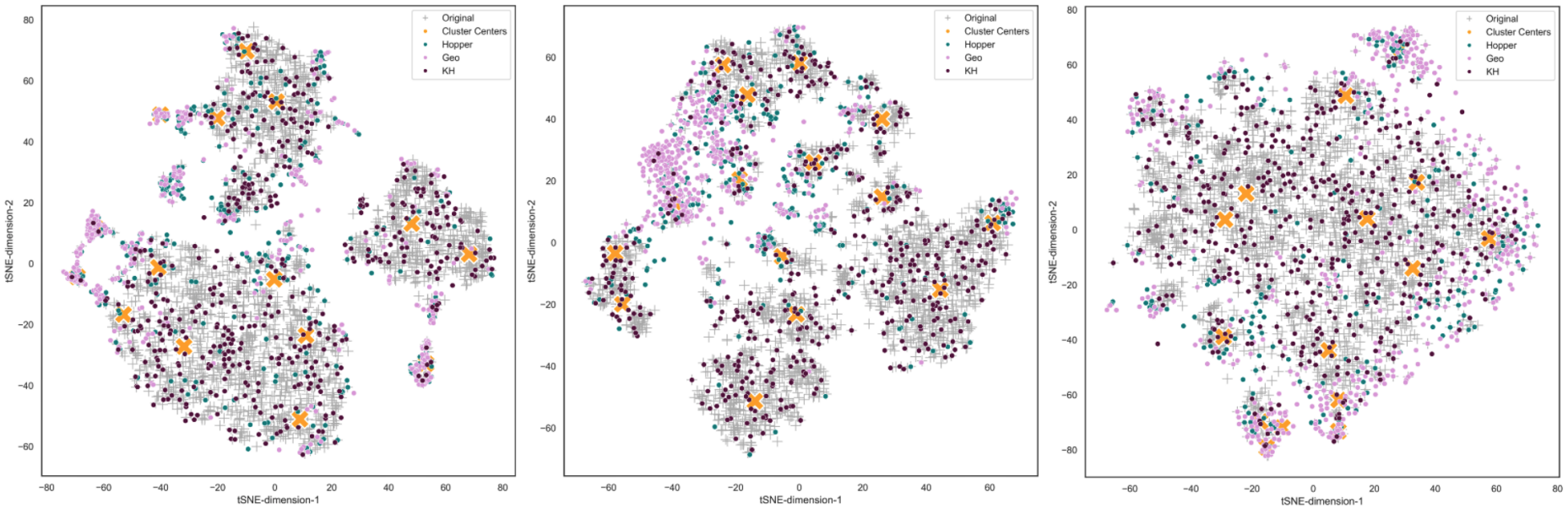


Figure: `https://arxiv.org/pdf/2207.00584.pdf`

Task-driven evaluation

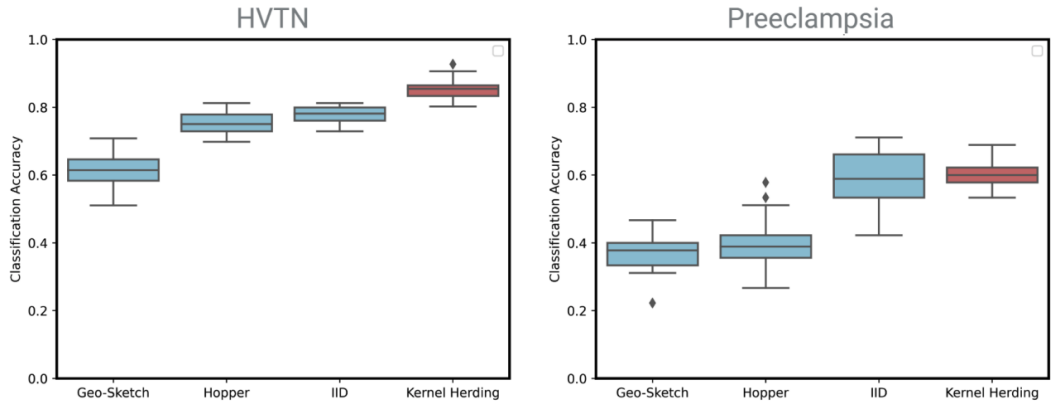


Figure: `https://arxiv.org/pdf/2207.00584.pdf`

Geometric Sketching

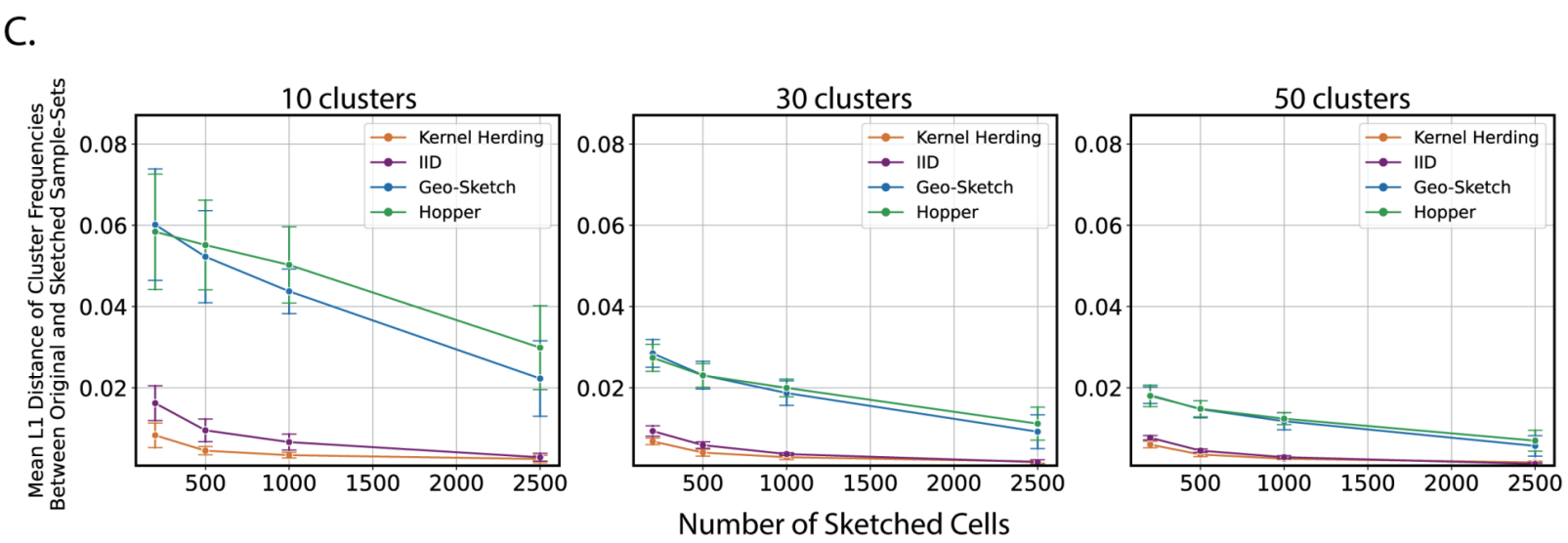


Figure: `https://arxiv.org/pdf/2207.00584.pdf`. Geometric Sketching is overall better at preserving frequencies of cell-types.

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

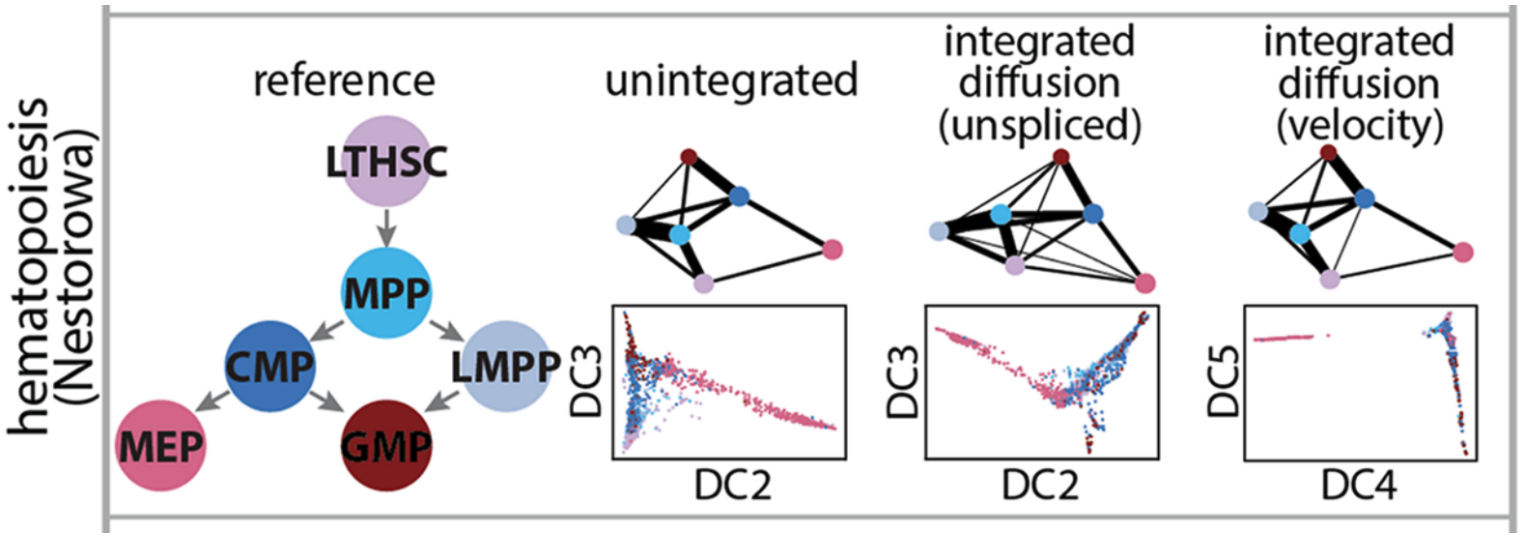


Figure: from Raneke*t al.*Genome Biology. 2022. We would like to establish a chronological ordering of cells that represents their differentiation. This will allow us to uncover gene programs associated with differentiation.

Welcome Back Diffusion Maps

Let's set up some notation as follows :

- Let Ω 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,

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

Here the σ^2 determines the scale over which each cell can randomly diffuse.

Modeling the Probability cell x transitioning to cell y

Define the transition matrix, **P** for all pairs of cells as,

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

Here $Z(\mathbf{x})$ is a partition function providing an estimate for the number of neighbors of **x** in a certain volume defined by σ, 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.

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

$$\tilde{Z}(\mathbf{x}) = \sum_{\mathbf{y}\in\Omega/\mathbf{x}} \frac{\exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right)}{Z(\mathbf{x})Z(\mathbf{y})}$$

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

- $\tilde{\mathbf{P}}$ has n ordered eigenvalues, $\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_{n-1}$
- Corresponding to these eigenvalues are eigenvectors $\boldsymbol{\Psi}_0 \ldots \boldsymbol{\Psi}_{n-1}$
- 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,

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

Unpacking Diffusion Distance

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

- The eigenvector to the largest eigenvalue, λ_0 is a constant vector, $\boldsymbol{\Psi}_0 = \mathbf{1}$. 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

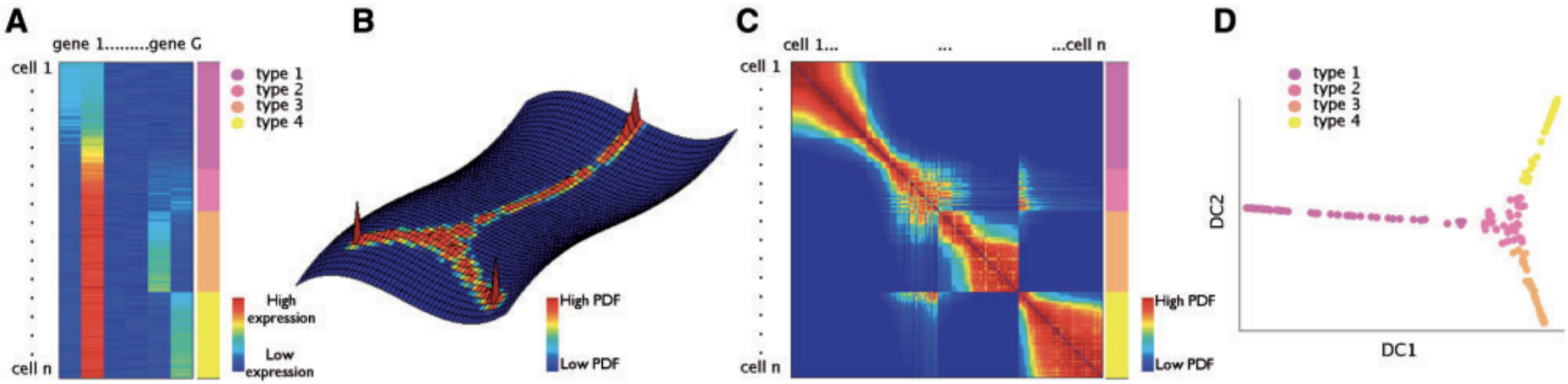


Figure: from Haghverdi *et al.* Bioinformatics. 2015. After projecting cells based on the first two diffusion components, it is still unclear what the time direction is.

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.

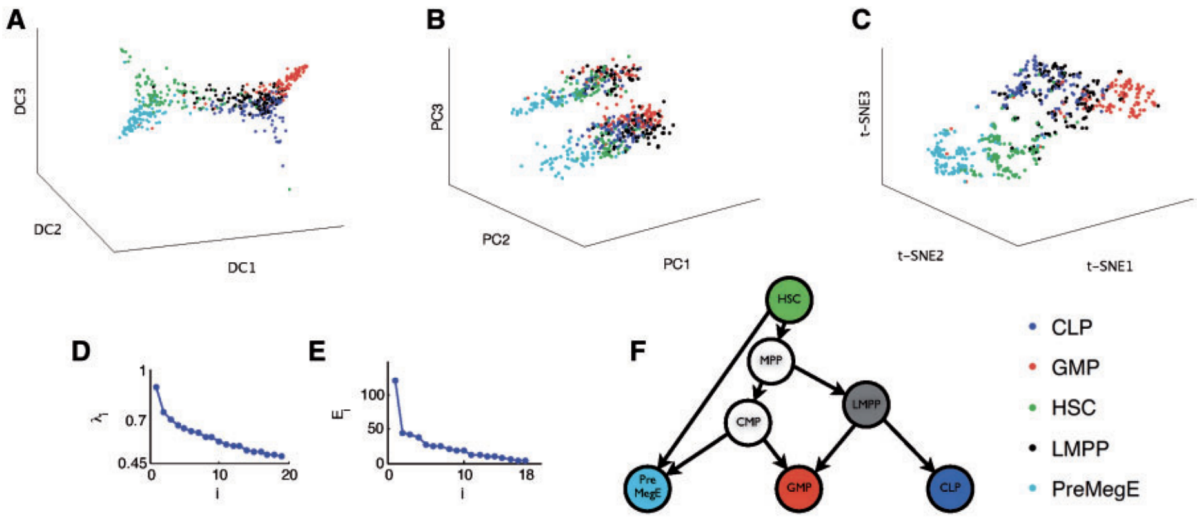


Figure: from Haghverdi *et al.* Bioinformatics. 2015.

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