Comp790-166: Computational Biology

Lecture 20

April 19, 2021

Announcements

- Homework 2 is online. Due April 23. `https://github.com/stanleyn/Comp790-166-Comp-Bio/tree/main/Homework2`
- Project Signup Sheet `https://docs.google.com/spreadsheets/d/1bSZByipJJxlRGXNqECi02VHxbS39WV6-cXL0akbA2kI/edit?usp=sharing`
- Comments on projects? How are they going?

Homework Comment

The edgelists in the homework had some repeated edges. I removed repeated edges from the edgelists. (Thanks to DJ for pointing this out). Note that the nodenames are a bit ugly because they are shortened names provided by the authors.

Project Writeup Template

The project writeup is structured just like a typical conference-style research paper.

- PDF `https://github.com/stanleyn/Comp790-166-Comp-Bio/blob/main/Projects/Project_writeup.pdf`
- TeX `https://github.com/stanleyn/Comp790-166-Comp-Bio/blob/main/Projects/Project_writeup.tex`

Guideline for Figures

- Overview figure
  - Schematic illustration of your method and what you did + data, etc,
- 1-2 figures and or tables is a reasonable set of results

Example

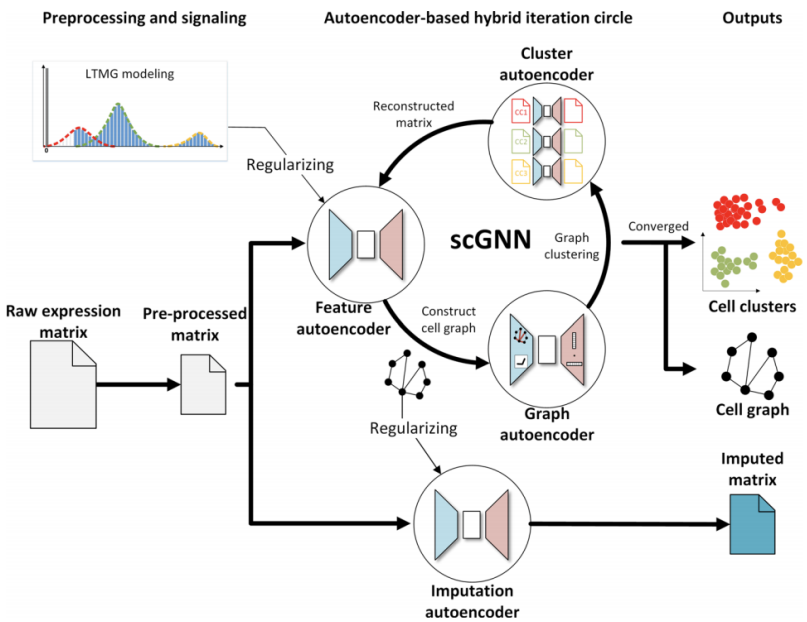


Figure: example from `https://www.nature.com/articles/s41467-021-22197-x.pdf`. Clearly shows inputs/outputs and what is happening under the hood.

Today

- Geometric Sketching for Single-Cell Data
- Hopper and Single Cell Dataset Examples

Issues We Have Seen

Consider a mass cytometry or scRNA seq (:D) dataset from a huge clinical cohort (hundreds of samples, each with 100s of thousands of cells)

- We can't deal with all of these cells at once (clustering, visualization, etc)
- Downsampling is biased towards the higher frequency cell-types
- **Compression:** What if we want to take a representative subsample of the entire dataset?

An Important Task from a Practical Viewpoint: Patient Outcome Prediction

If you want to build a regression model, get a prediction for each cell and pool predictions for cells across patients, you are limited by the number of cells

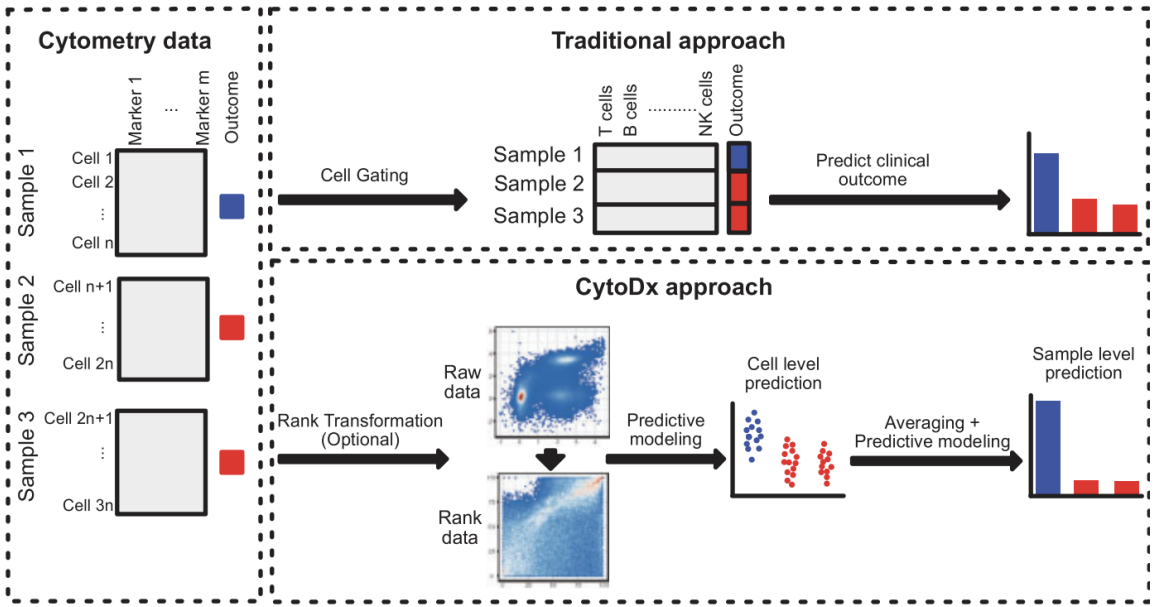


Figure: from Hu *et al.* Bioinformatics. 2018.

Previously Seen in Spade

The probability of keeping cell i was computed as,

$$\text{prob(keep cell } i) = \begin{cases} 0, \text{ if } LD_i \leq OD \\ 1, \text{ if } OD < LD_i \leq TD \\ \dfrac{TD}{LD_i}, \text{ if } LD_i > TD \end{cases}$$

- Local density (LD_i) is the number of cells within a particular cell i
- (TD) is target density
- (OD) is outlier density

Visualization of Spade Downsampling

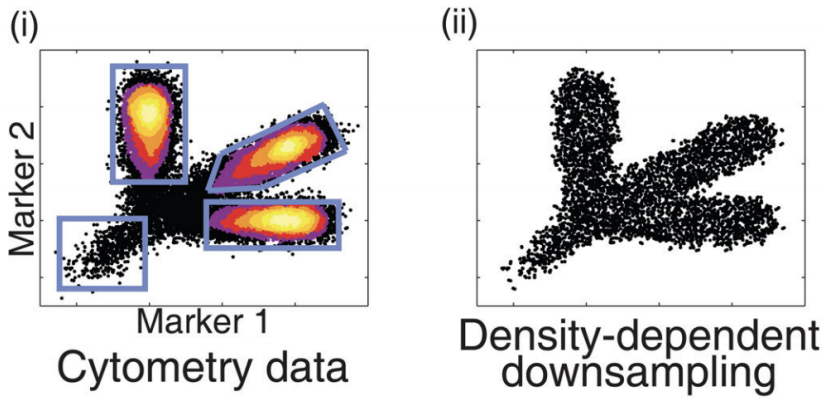


Figure: from Qiu *et al.* Nature Biotech. 2012.

Seeding a Graph

If you wanted to quickly grab some nodes of a graph that were well-distributed, what would you do? How would you figure out how many seeds?

Welcome Geometric Sketching

Instead of randomly selecting a subset of points for visualization, etc., why not do something a bit more smart, according to the inherent data geometry.

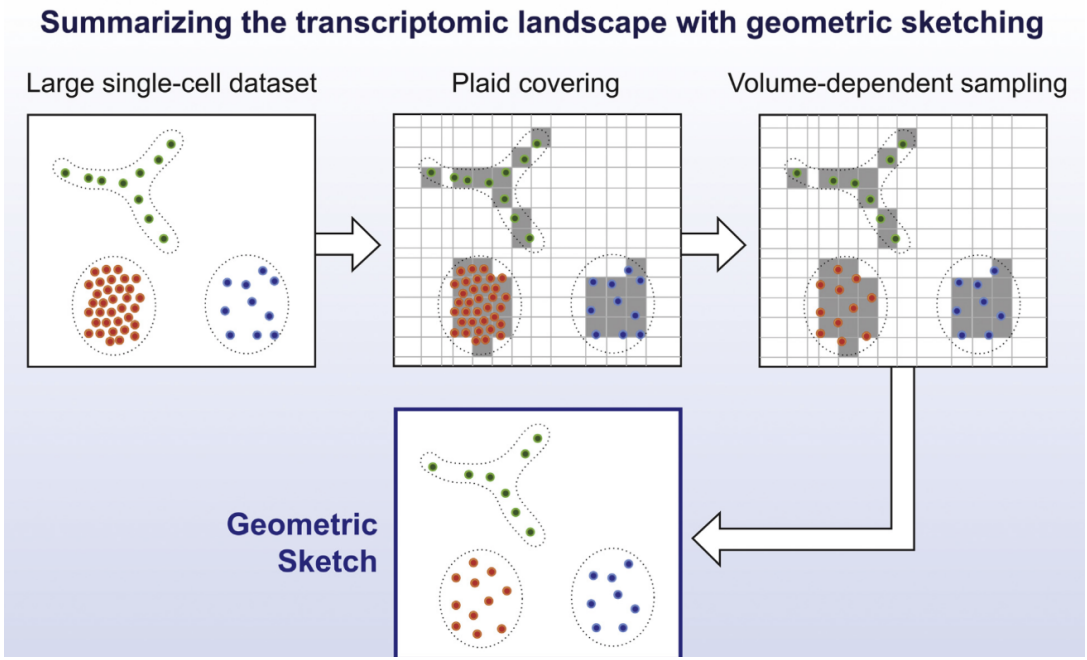


Figure: from Hie *et al.* Cell Systems. 2019

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.

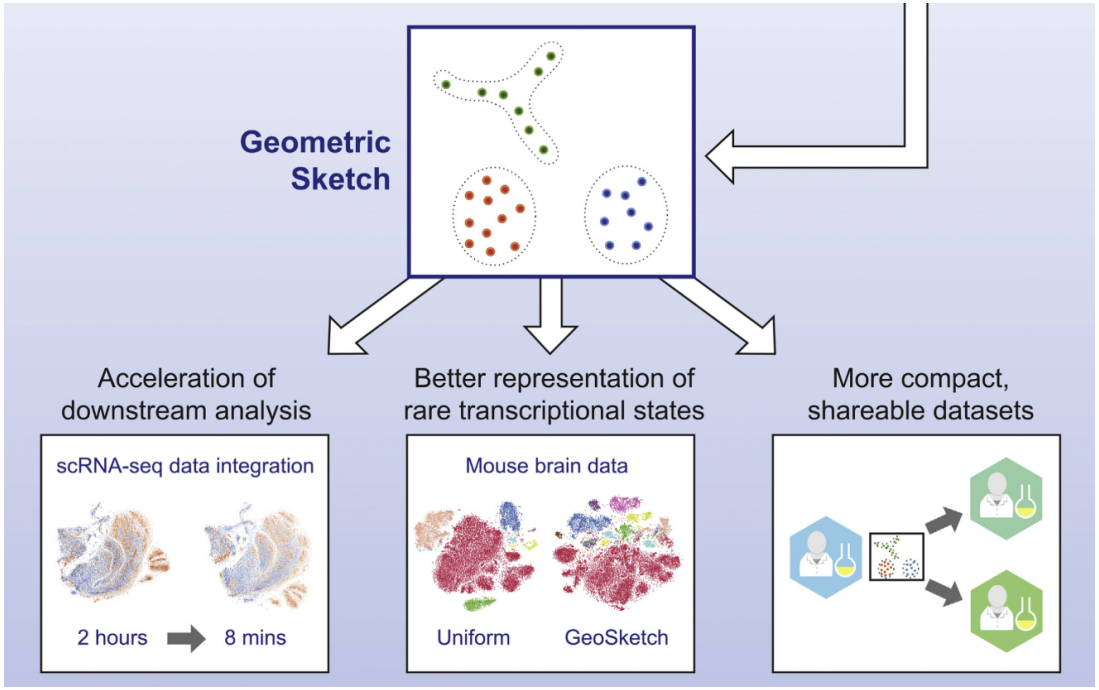


Figure: from Fig. 2 of Hie *et al.* Cell Systems. 2019. Perhaps we need an accurate estimate of cell-population frequencies, etc.

Sketching Algorithms

For a brilliant tutorial and description of sketching, refer to the following tutorial, `https://nips.cc/virtual/2020/public/tutorial_7bef20627bb50052e352b9653c3bca53.html`

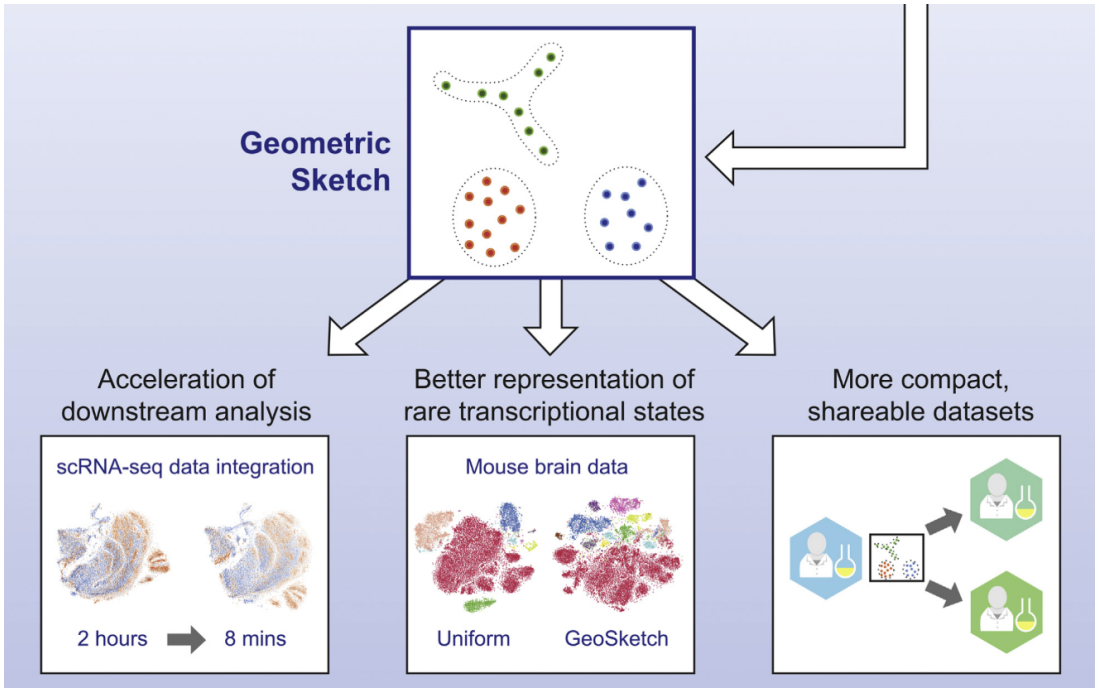


Figure: from Fig. 2 of Hie *et al.* Cell Systems. 2019.

Formulation of Sketching Problem

- Let $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ be a collection of m-dimensional cells ($\mathbf{x}_i \in \mathbb{R}^m$).
- We seek a down-sampled subset, $S \subset \mathcal{X}$, which can be used for downstream analysis, specifically to understand the salient characteristics of $\mathcal{X}$

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 Formuation 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.

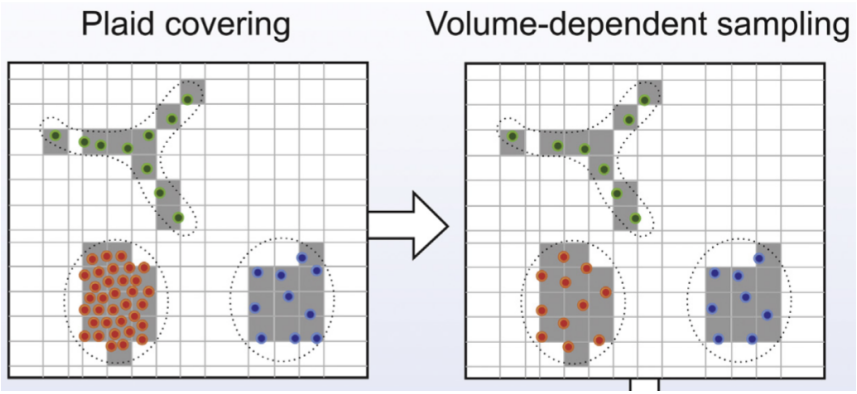


Figure: from Fig. 2

Parameters for Sketching

Geometric Sketching Algorithm Parameters

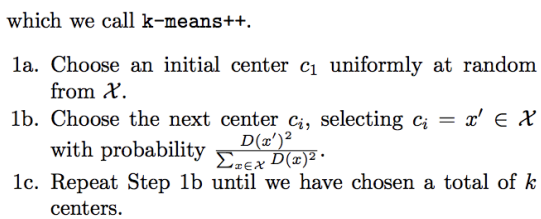
Parameter	Type	Default Value	Notes
Sketch size (k)	Integer between 0 and total number of cells, inclusive	N/A	The desired sketch size is chosen depending on the amount of compute resources available and the algorithmic complexity of downstream analyses; smaller sketches omit more cells but will accelerate 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 distance from a point and the previously chosen cluster center.

Baseline Downsampling Method, Spatial Random Sampling (SRS)

- Ref : `https://arxiv.org/pdf/1705.03566.pdf`
- A limited number of points are sampled based on their proximity to randomly sampled points on the unit sphere.

Results: Hausdorff Distance

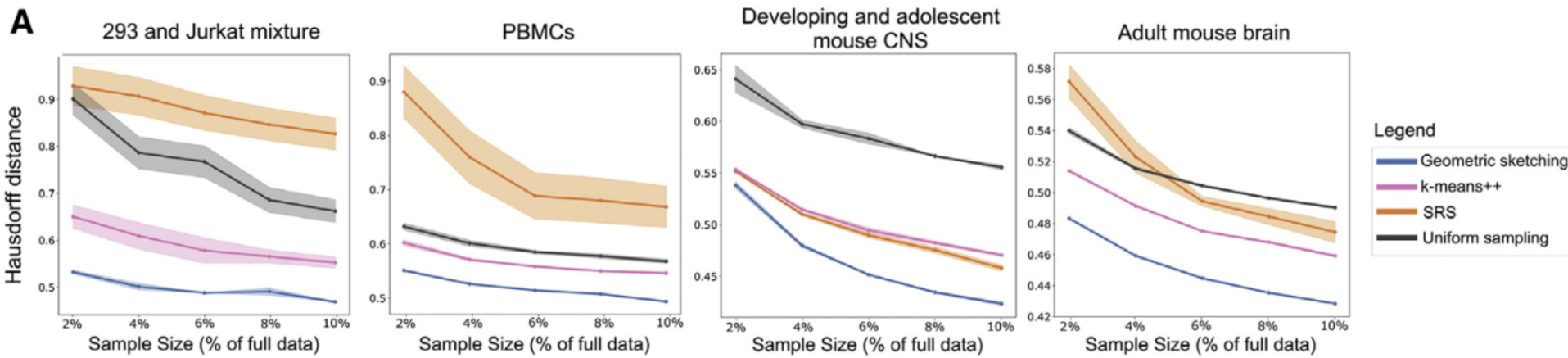


Figure: from Fig. 3. Note the superior performance of geometric sketching!

Using Robust Hausdorff

Recall the interplay between q and k for computing the robust hausdorff distance.

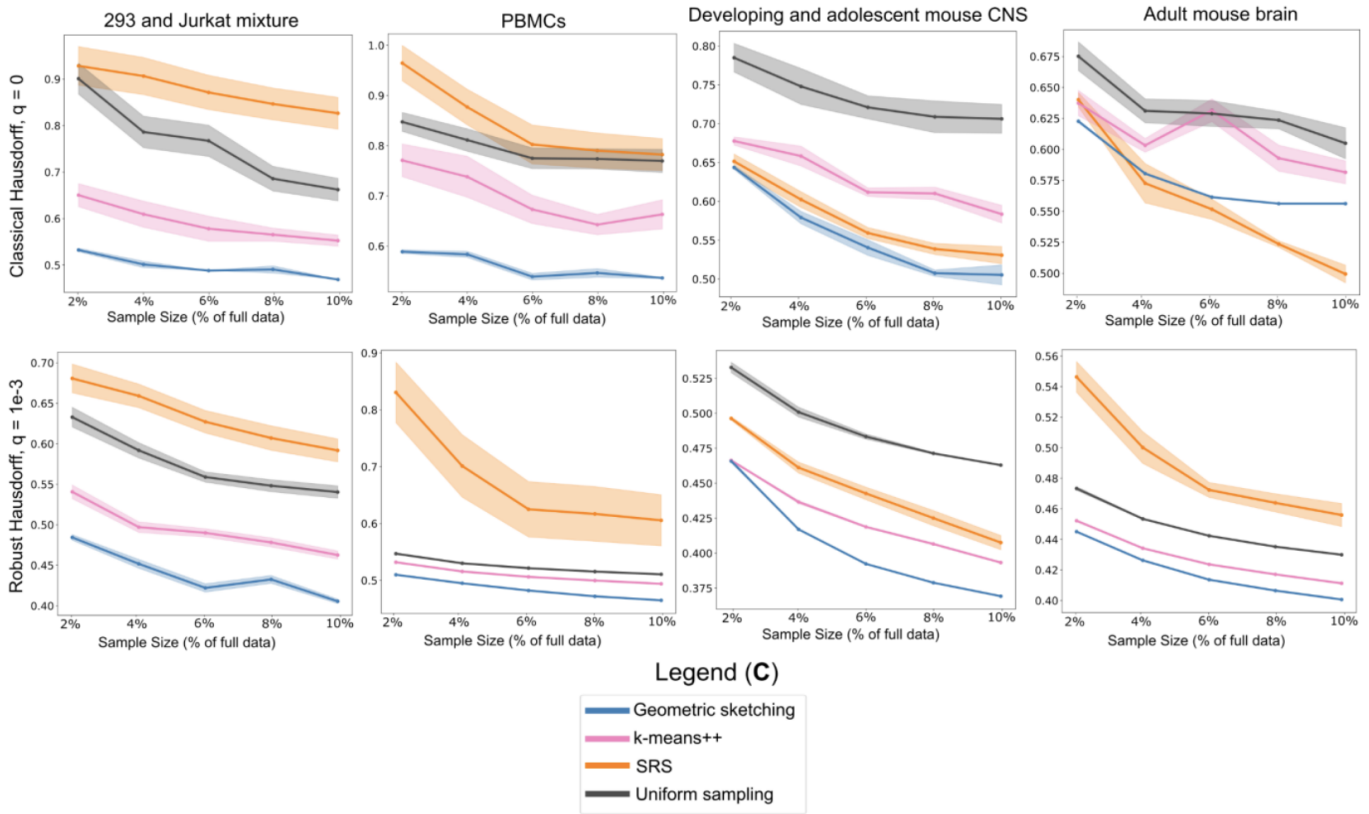


Figure: from Supplementary Figure S5.

Downsampled Cells Facilitate Faster Downstream Tasks

for example : batch effect correction

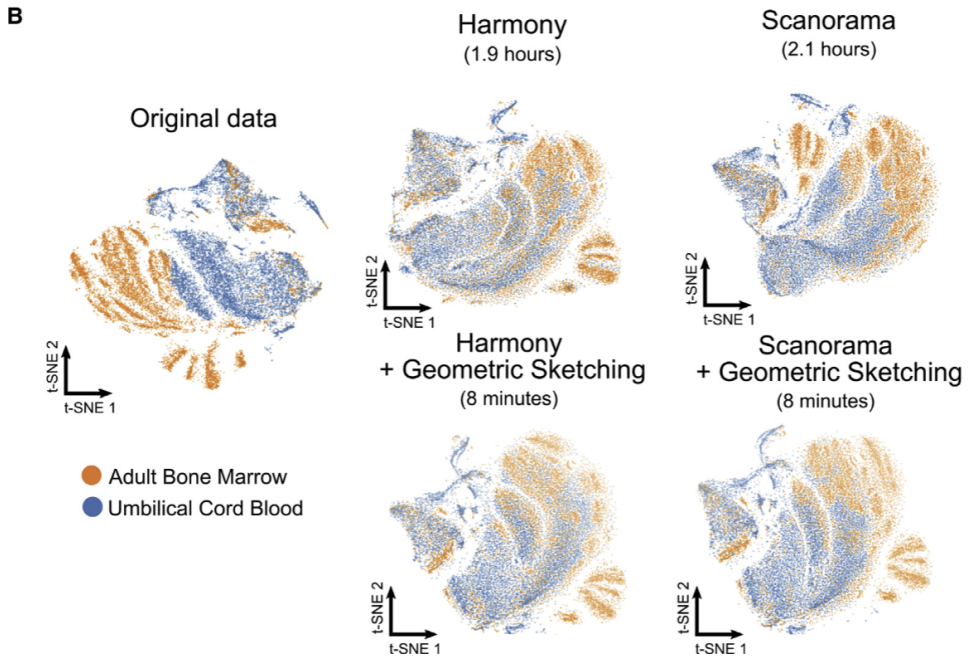


Figure: from Fig. 5

Rarer Populations are Sufficiently Represented

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

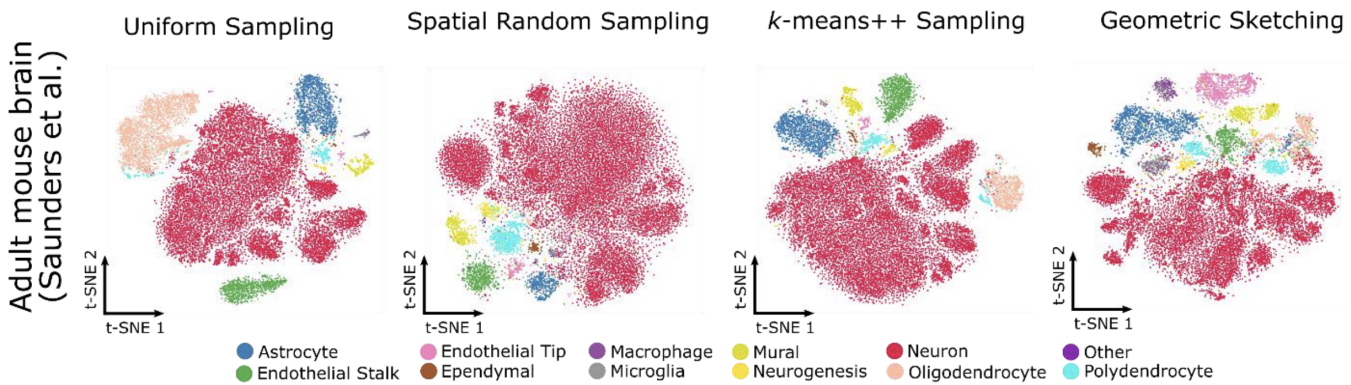


Figure: from Supplementary Figure S1

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.

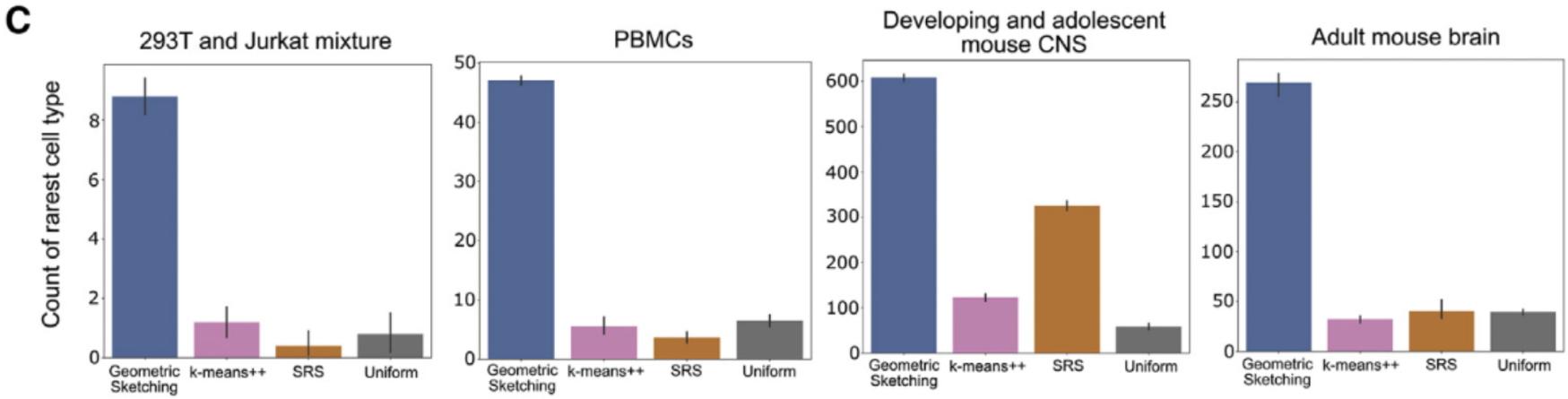


Figure: from Fig. 3

Interpretation wrt Clustering

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

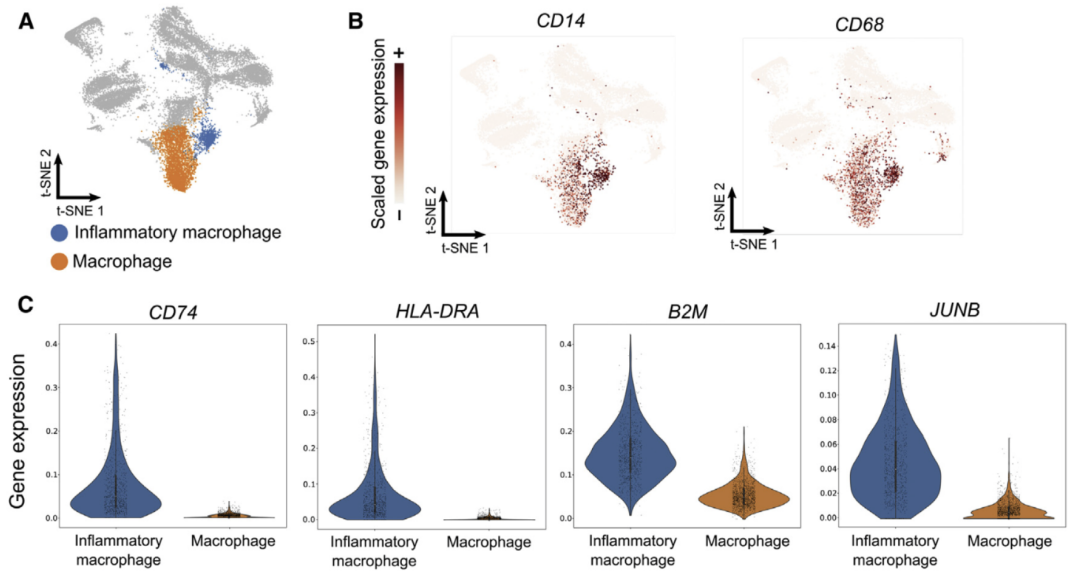


Figure: from Fig. 4

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:** (for class discussion only, because it's a research idea)

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,

- Since $d(x,p) \leq d(s,x)$, 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
- Quickly find these points by sorting X by their distances to the nearest point of S. If $d(s,p) \geq 2r$ with $x \in X$ closes to x, the triangle inequality gives, $d(s,x) \geq d(s,p) - d(x,p) \geq r$.

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.

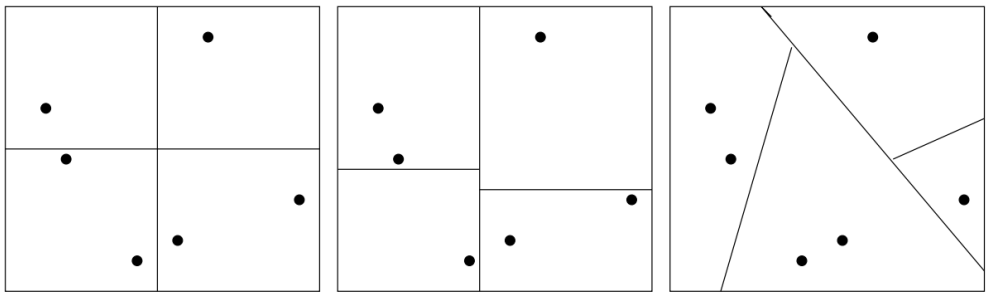


Figure: from `https://arxiv.org/pdf/1205.2609.pdf`

Hausdorff Distances

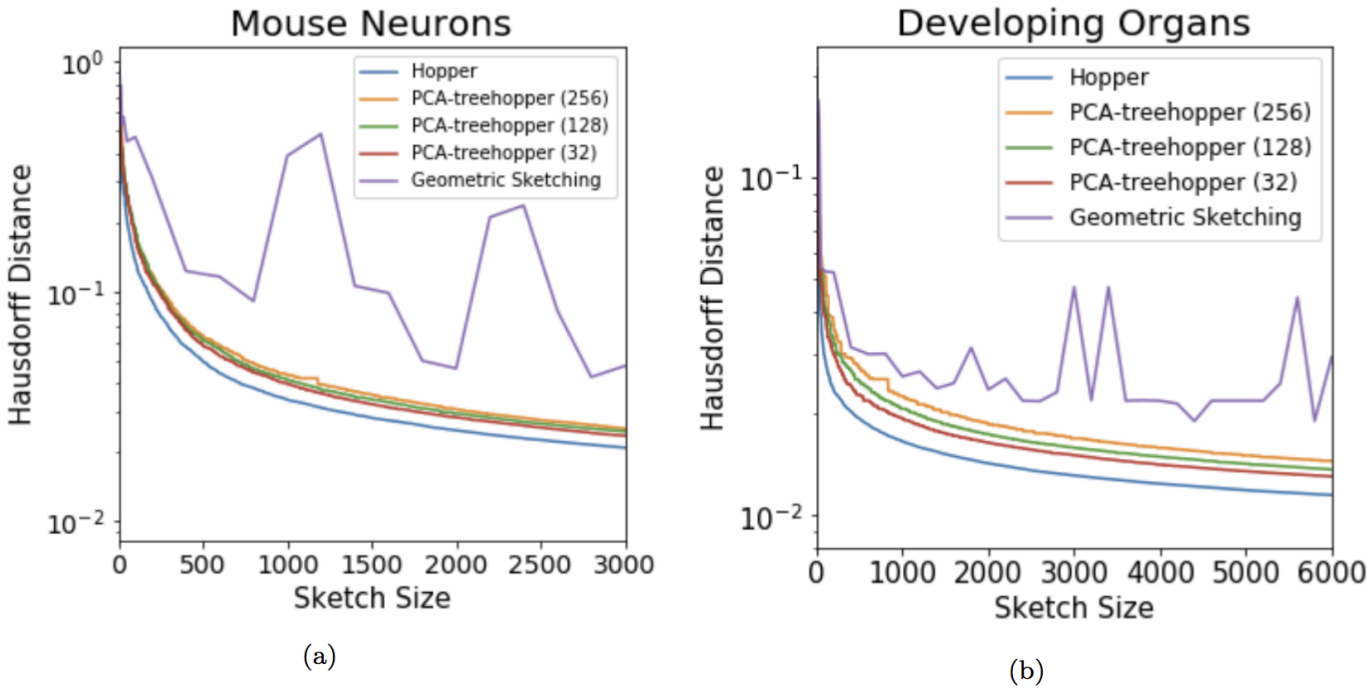


Figure: from Fig. 1 of deMeo *et al.* Bioinformatics. 2019.

Runtimes

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

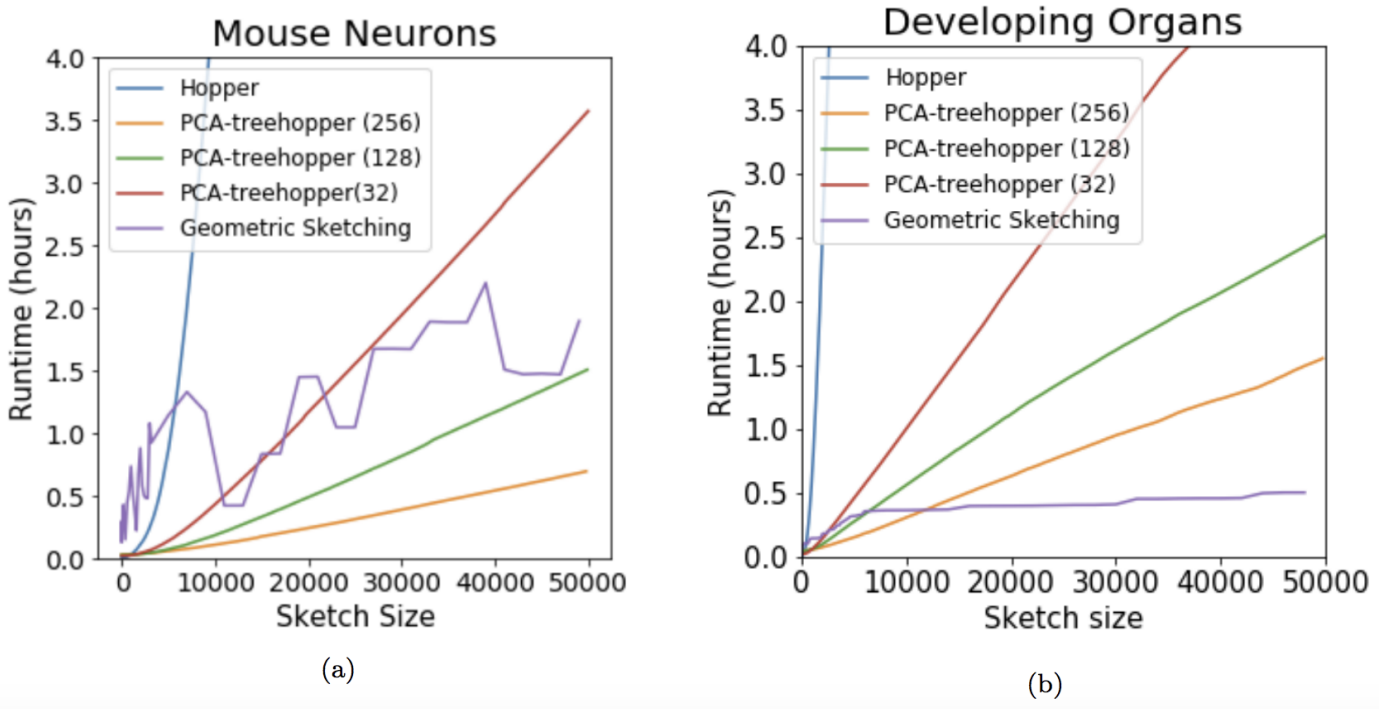
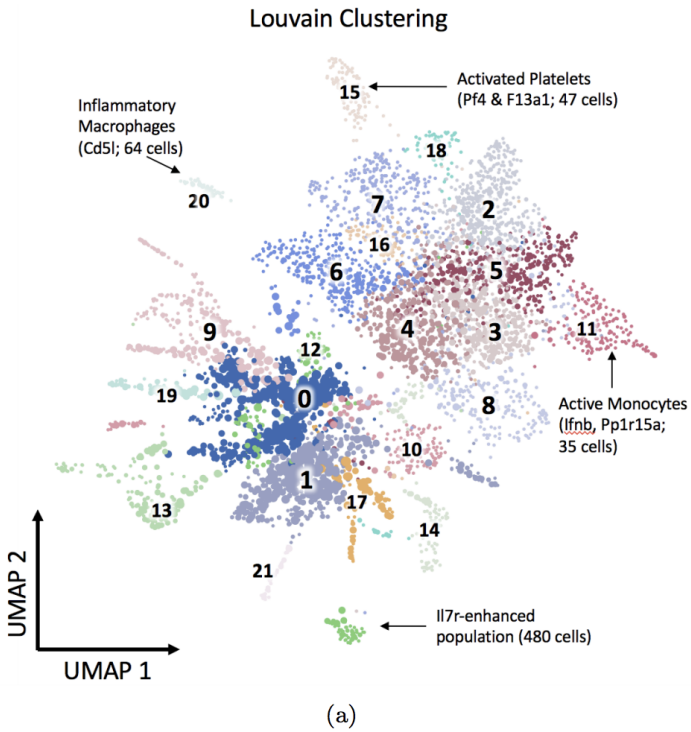


Figure: from Fig. 2 of deMeo *et al.* Bioinformatics. 2019.

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.



(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).

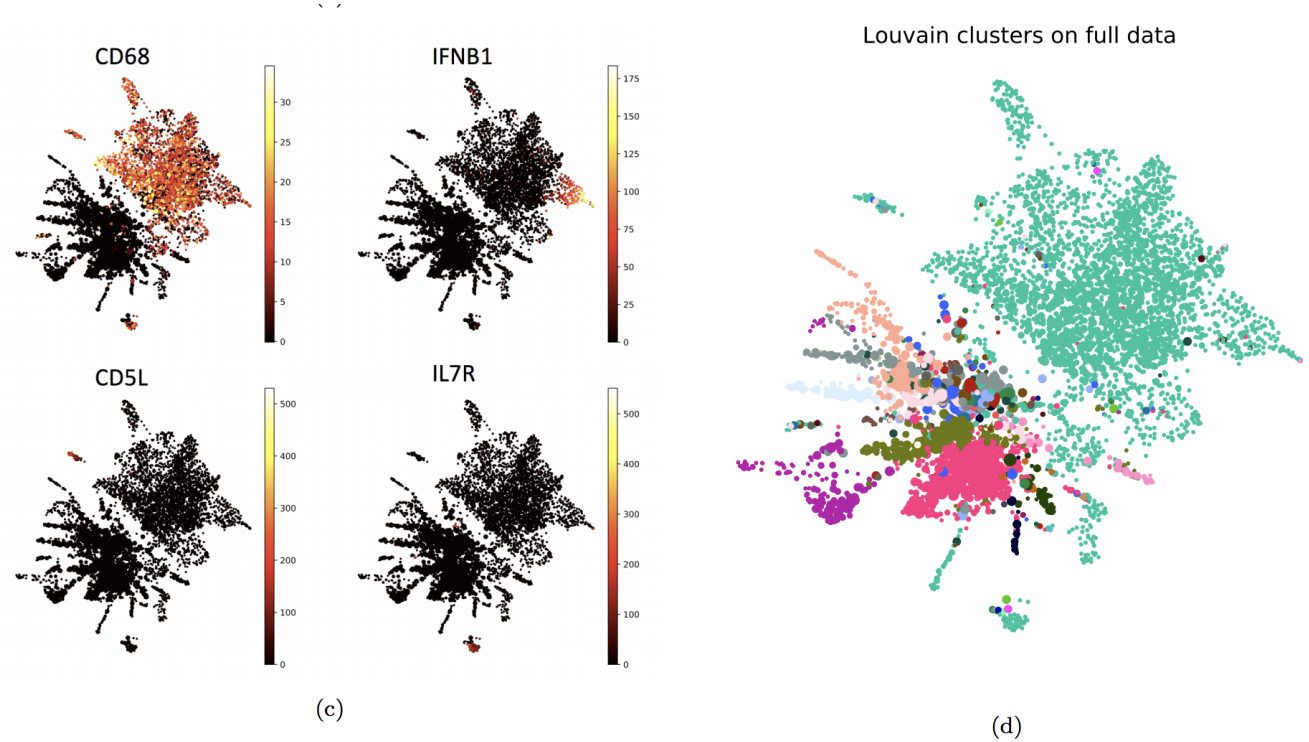


Figure: from Fig. 3 of deMeo *et al.* Bioinformatics. 2019.

Comparing Hopper to Geometric Sketching

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

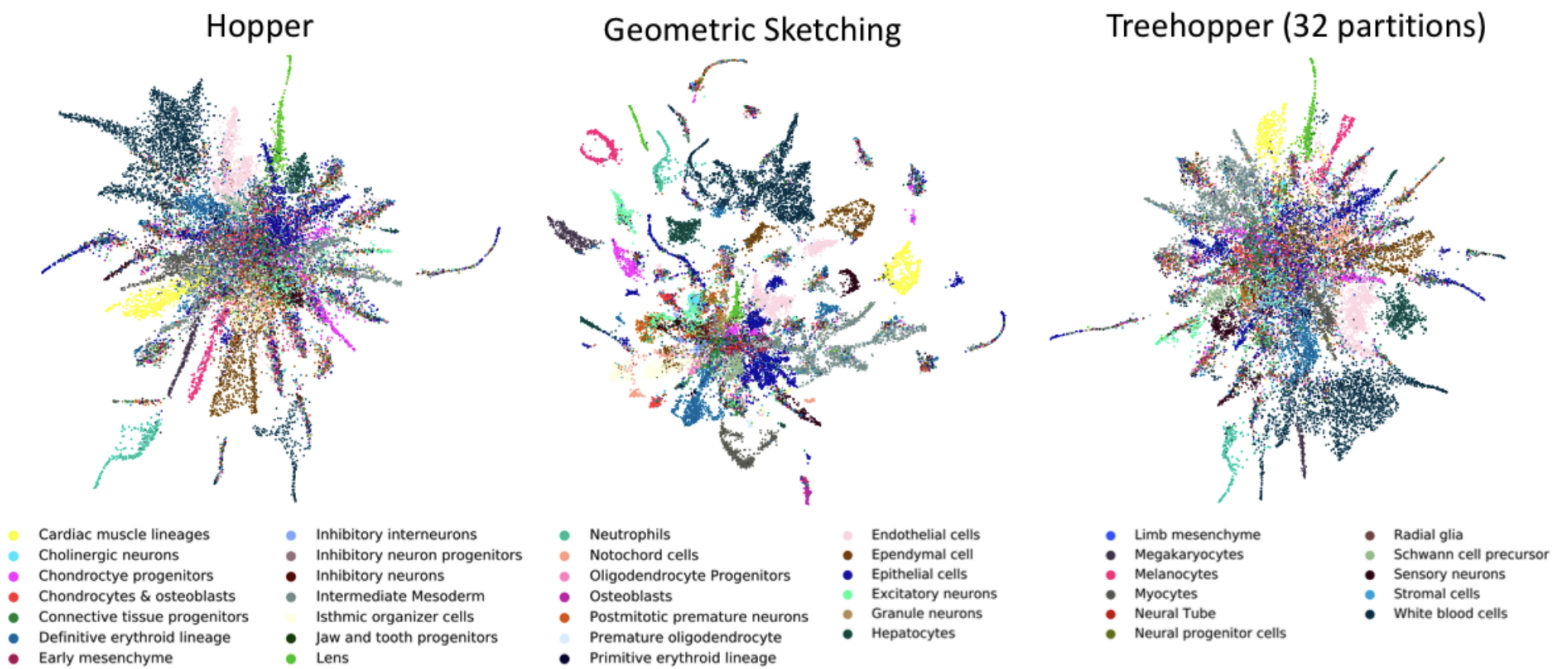


Figure: Fig. 3 of deMeo *et al.* Bioinformatics. 2019.

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.