Efficient Sketch-Based Analysis of Single-Cell RNA-Seq Datasets Using Geometric Sketching.

Arth Banka
Computational Biology Department, Carnegie Mellon University
abanka@andrew.cmu.edu

1. Introduction

Advancements in single-cell RNA sequencing (scRNA-seq) have enabled profiling of millions of individual cells, offering deep insights into cellular heterogeneity. However, the scale of these datasets poses significant computational challenges for analysis and interpretation. To address this, we explore sketch-based methods that select representative subsets of cells while preserving biological diversity. Using the 10x Genomics 68k PBMC dataset, we assess each method's ability to maintain transcriptional heterogeneity, capture rare populations, and support downstream analyses such as clustering and differential expression. Our study compares **uniform sampling, geometric sketching, K-Means sketching, kernel herding, hashing-based, and PCA-based methods**, evaluating their performance in terms of runtime and biological fidelity.

2. Sketching Methods

Uniform Sampling: Randomly selects N cells from the full dataset. This approach is fast and unbiased, but may miss rare cell populations or underrepresent the dataset's diversity.

Geometric Sketching: Projects cells into PCA space and performs farthest-point sampling to select a subset that evenly covers the geometry of the data. This preserves transcriptional heterogeneity by favoring diverse spatial coverage.

K-Means Sketching: Runs k-means clustering on PCA-reduced data and selects one representative (medoid) per cluster. This ensures each major data region is represented but may miss small or rare clusters.

Kernel Herding: Iteratively selects cells that minimize the discrepancy between the sketch and full dataset in a reproducing kernel Hilbert space (RKHS). Offers high fidelity but is computationally expensive.

Hashing-Based Sampling: Applies a hash function to deterministic cell identifiers and selects the top N sorted entries. This method is extremely fast and reproducible, though biologically agnostic like random sampling.

PCA-Weighted Sampling: Computes variance in PCA space and assigns selection probabilities proportional to a cell's contribution to high-variance components, enriching for cells in more variable regions.

3. UMAP visualization of Selected Samples

We visualize the sketching outputs using UMAP, a dimensionality reduction technique that embeds high-dimensional gene expression data into a 2D space while preserving local neighborhood structure. Each plot overlays the full dataset (gray) with the subset selected by a given sketching method (colored by cluster identity from the full data).

- **Uniform Sampling** covers the data evenly but may miss rare or tightly clustered groups.
- **Geometric Sketching** distributes points across underrepresented regions, preserving structural diversity.
- **K-Means Sketching** concentrates samples in high-density areas,but the resulting plot has good coverage.
- **Kernel Herding** provides coverage to some clusters that are not visible in other methods but misses out on several common clusters too.
- **Hashing-Based** sampling is very similar to random sketch as expected.
- **PCA-Weighted Sampling** focuses on high-variance regions, potentially amplifying transcriptional extremes.

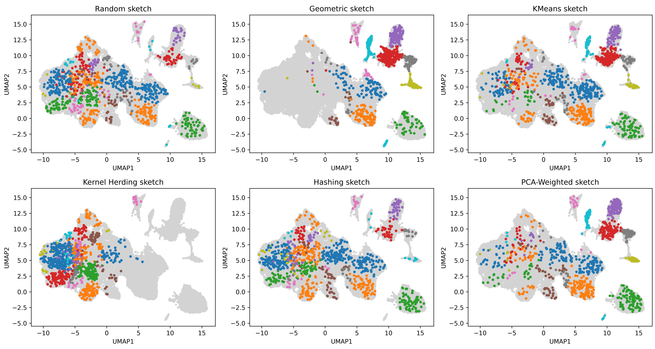

Fig 1: UMAP for each method against the grayed out complete set

4. Time Complexity Analysis

Random, Hashing, and PCA-based methods are fastest. Kernel Herding, while theoretically powerful, is computationally costly.

- **Random Sampling:** $\mathcal{O}(N)$ — Direct random selection of N cells.
- **Hashing-Based Sampling:** $\mathcal{O}(n \log n)$ — Compute hash and sort n cells, where n is the total number of cells.
- **PCA-Weighted Sampling:** $\mathcal{O}(N+n)$ — Sample N cells proportional to variance scores.
- **Geometric Sketching:** $\mathcal{O}(Nn)$ — Greedy farthest-point sampling in PCA space.
- **K-Means Sketching:** $\mathcal{O}(knt)$ — k clusters, n data points, t iterations (usually $t \ll n$).
- **Kernel Herding:** $\mathcal{O}(Nn^2)$ — Requires dense kernel matrix and greedy selection over N iterations.

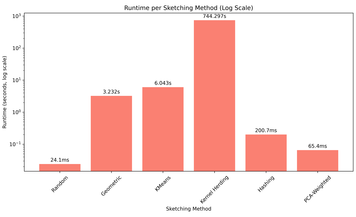

Fig 2:Log scaled time taken in seconds for each method for n=1000.

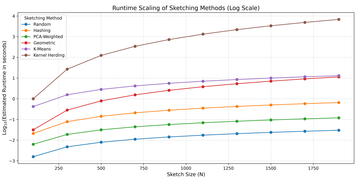

Fig 3:Log scaled time taken in seconds for each method as a function of n.

5. Biological Fidelity

This panel evaluates each method's ability to preserve rare and diverse cell populations. The top plot shows how many cells from the rarest group (with 193 cells) were captured, while the bottom plot reflects how many distinct clusters are represented in the sketch. Ideal methods retain both rare and common populations. Geometric, Random and PCA-weighted sketches perform well in balancing these goals, while surprisingly kernel herding performs poorly in both aspects.

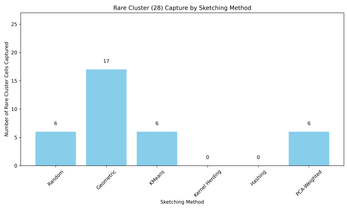

Fig 4: Rare cluster cell count across methods.

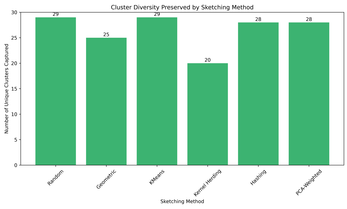

Fig 5: No. of clusters caught across methods.

6. Conclusions

- Fastest methods: Random, Hashing, and PCA-weighted sampling showed low runtimes and good scalability.
- Geometric & PCA-weighted sketches preserved both rare cell types and transcriptional diversity. K-Means captured diverse clusters but missed rarer populations.
- Kernel Herding failed to provided strong coverage fidelity while having high computational cost.
- **Method choice involves trade-offs between runtime efficiency, diversity preservation, and rare cell recovery.**

7. References

[1] Bryson B. Berger B. Hie, B. Geometric sketching compactly summarizes the single-cell transcriptomic landscape. *Cell Systems*, 8(6):483–493, 2019.

[2] Hao S. Andersen-Nissen E. et al. Hao, Y. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587, 2021.

[3] Welling M. Smola A. Chen, Y. Super-samples from kernel herding. *UAI*, 36(1):109–116, 2010.