# Impact of Preprocessing and Cell Exclusion on Stability of scRNA-seq Clustering: Implications for Single-Cell Genomic Analysis

Diego Garcia[1] , Juan Vergara Najar[1], Tim Hamilton[2,3], Eric Deeds[5]

1 BIG Summer Program, Institute for Quantitative and Computational Biosciences, UCLA, 2 Department of Integrative Biology and Physiology, UCLA

3 Institute for Quantitative and Computational Biosciences, UCLA, 5 Bioinformatics Interdepartmental Ph.D. Program, UCLA

## Abstract

Our study explores the effects of preprocessing and clustering on single-cell RNA sequencing (scRNA-seq) data, a revolutionary technology in cellular diversity and disease research. Specifically, this project analyzes whether excluding certain cells, be it the smallest cluster or a random selection would affect the stability of the clustering results as measured by the Adjusted Rand Index (ARI). We found that the ARI values between clusters created before and after the removal of certain cells indicated a high divergence between the two. This finding was consistent across multiple parameter values and datasets analyzed. These discrepancies could lead to errors in cell type identification, amplifying the need for improved clustering and dimensionality reduction algorithms. As we probe the expanding realm of single-cell genomics, our research underscores the need for effective, reliable, and interpretable analysis pipelines for single-cell data.

## Background and Objectives

- **How stable is the clustering process?**
- **How can we ensure observed differences or patterns in gene expression between cells are more likely to be biologically meaningful?**

**Objective:**

- **<u>Quantify the impact of preprocessing and visualization decisions on cell-type identification in scRNA-seq data.</u>**

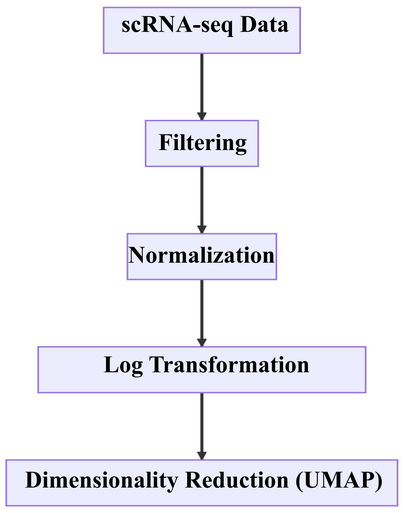



**Fig. 1**:The standard pipeline for pre-processing for our scRNA-seq data.

## Methods

**Fig. 2**: Methods

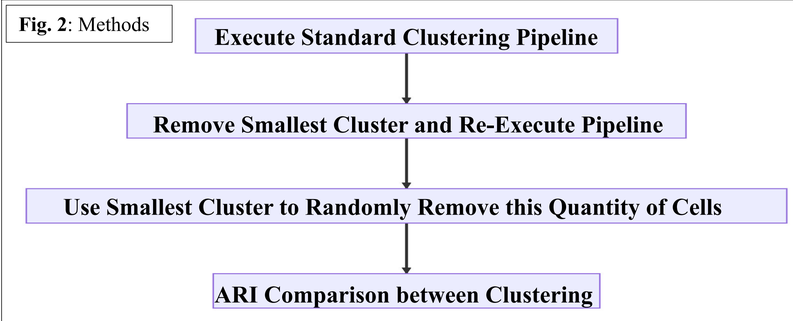
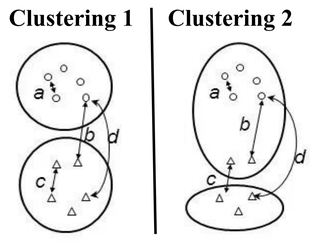



**Fig. 3: ARI –** Statistical measure to evaluate the similarity between two sets of clustered data; ranges from zero to one, with zero equating to random labelling and one when the clusters are identical.




Agreement: $a, d$

Disagreement: $b, c$

$$RI(P,G) = \frac{a+d}{a+b+c+d}$$

$$ARI = \frac{RI - E(RI)}{1 - E(RI)}$$

## Results

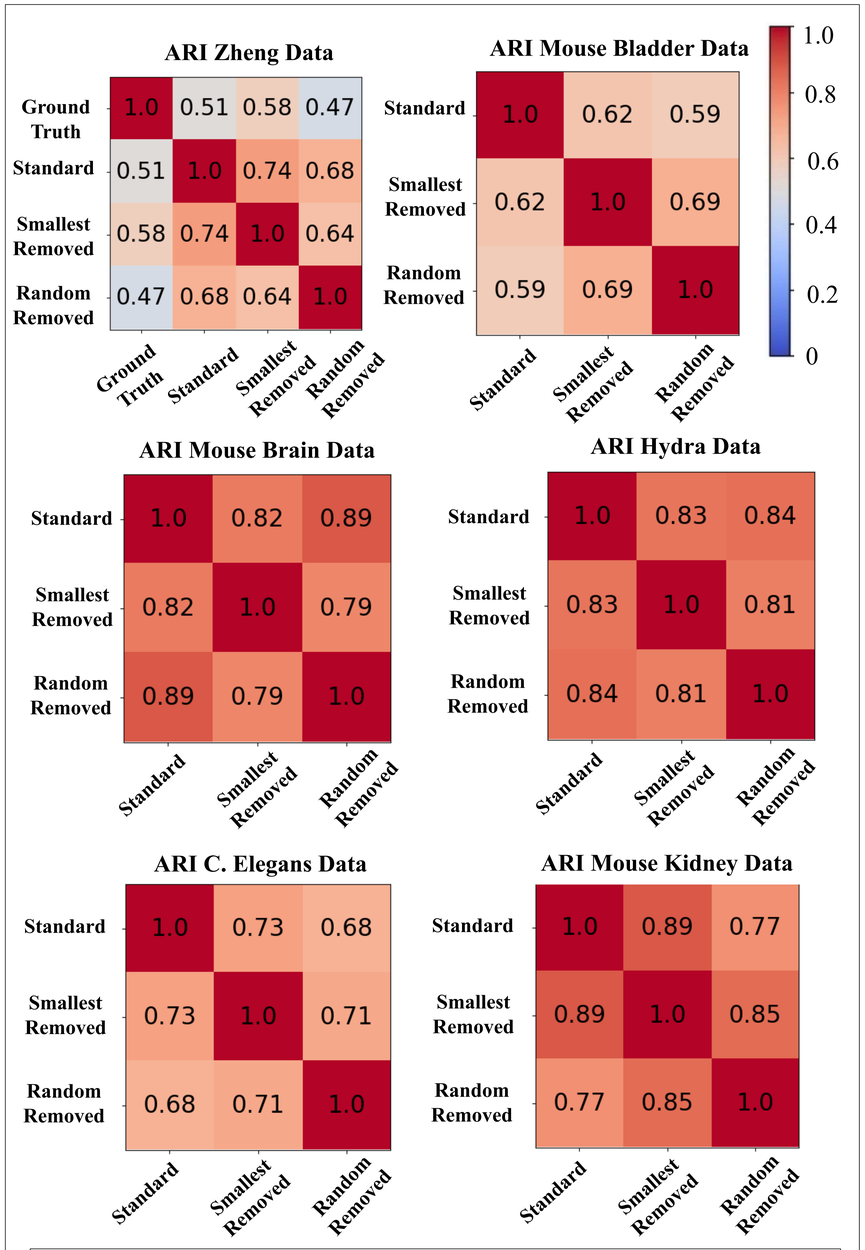



**Fig. 4**: Heatmaps of ARIs. Low to moderate ARIs are observed across all data sets and runs indicating the impact of preprocessing on analysis outcomes.

**Resolution –** Threshold within the Leiden clustering algorithm that allows differing clustered groups to join based on the modularity.

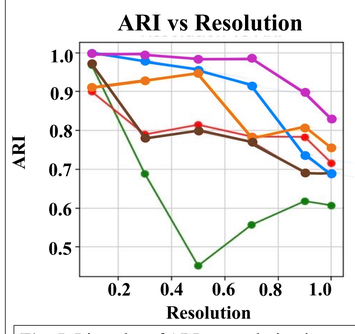



**Fig. 5**: Line plot of ARI as resolution increases showcasing data-dependent decreases in ARI.


**Fig. 6**: Line plot of total clusters increasing as resolution increases comparing all data sets.

## Discussion

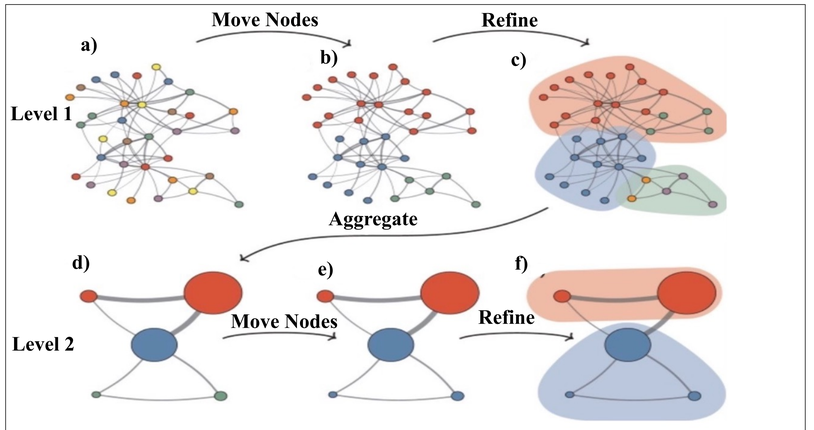



**Fig. 7** : Illustrating the use of modularity in Leiden clustering

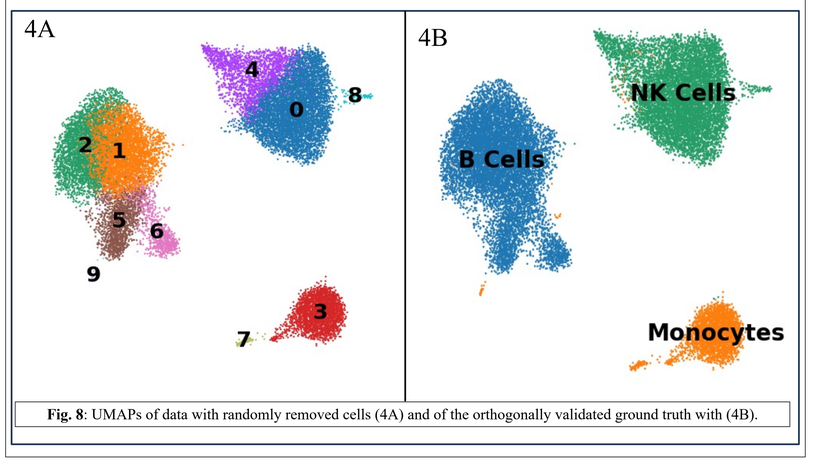



**Fig. 8**: UMAPs of data with randomly removed cells (4A) and of the orthogonally validated ground truth with (4B).

## Future Work

- Significant impact of preprocessing on scRNA-seq data analysis.
- Continuous evolution in single-cell genomics.
- Importance of rigorous data analysis.
- Need for accurate and interpretable single-cell analysis.
- Potential of single-cell genomics in understanding complex systems.

## References

1. Satija, R., Farrell, J., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial reconstruction of single-cell gene expression data. Nature Biotechnology, 33(5), 495–502. https://doi.org/10.1038/nbt.3192
2. Duò, A., Robinson, M. D., & Soneson, C. (2018). A systematic performance evaluation of clustering methods for single-cell RNA-seq data. F1000Research, 7 , 1141. https://doi.org/10.12688/f1000research.15666.3
3. Siebert, S., Farrell, J. A., Cazet, J. F., Abeykoon, Y., Primack, A. S., Schnitzler, C. E., & Juliano, C. E. (2019). Stem cell differentiation trajectories in Hydra resolved at single-cell resolution. Science (New York, N.Y.), 365(6451), eaav9314.
4. Medium. (n.d.). Medium. Retrieved August 8, 2023, from https://medium.com/@anushka.datascoop/evaluation-metrics-in-machine-learning-101-accc3cd35af9
5. Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019). https://doi.org/10.1038/s41598-019-41695-z

## Acknowledgements

We would like to acknowledge the B.I.G. Summer Program and the Deeds Lab for their unwavering and invaluable support throughout the duration of the program.