**BMI 5332 Project Proposal**

Timothy Bi

**Significance**

Sample multiplexing is often used in single-cell RNA sequencing (scRNA-seq) experiments to increase cellular throughput and reduce cost and technical batch effects1–4; downstream analysis necessitates demultiplexing using a variety of bioinformatics tools, of which a rigorous comparison of their performance is lacking. In particular, it is unknown if one tool outperforms the rest, or if different tools excel in different situations, given their varying underlying statistical assumptions. Accurate sample demultiplexing is critical for recovery of cells (important for experiments studying rare cell populations or with low cell numbers) and for eliminating technical artifacts such as multiplets. Demultiplexing errors (i.e. incorrectly assignment of samples) can also significantly confound downstream analysis. Given the high cost of sequencing and the often precious nature of samples being studied, identifying the right tool for the right situation is important for maximizing our ability to extract biologically relevant insights; therefore, we propose to formally benchmark six computational demultiplexing tools: bimodal flexible fitting (BFF)5, HTODemux6, hashedDrops7, GMM-Demux8, DemuxEM9, and demuxmix10 across a variety of simulated experimental conditions. I anticipate that this project will provide the necessary comparisons for researchers to decide which tool best suits their experimental parameters and needs.

**Background**

The concept of multiplexing samples for single cell experiments is not new and has been previously used in a variety of applications ranging from mass11 and flow cytometry12,13 to quantitative proteomics14. Nevertheless, it has been an area of intense study especially in single cell sequencing experiments, where throughput is low, cost is high, and batch effects are significant. The most popular way to barcode samples for multiplexing is to use hashtag oligonucleotides (HTOs), which are oligonucleotide-conjugated monoclonal antibodies that are specific for ubiquitous and highly expressed surface proteins such as β2-microglobulin (a component of the MHC class I molecule, which is expressed on all nucleated cells)15. Another approach is to use oligonucleotide-conjugated lipids, which can incorporate into the cell plasma membranes16. If samples are genetically distinct, single nucleotide polymorphisms can also be used as a natural cell barcode17,18. For this proposal, we will focus our benchmarking efforts on tools that are primarily built to demultiplex HTO data.

In practice, hashtag data in scRNA-seq experiments is often noisy, with a variety of contributing factors such as the presence of multiple cells per droplet (multiplets); cross-contamination of hashtags by dissociation and re-binding once cells are pooled; and ambient, unbound HTOs present in the droplets. As such, computational methods are not only helpful but also necessary for accurate demultiplexing of samples for downstream analysis. The computational tools we will benchmark in this proposal make different statistical assumptions to better differentiate between noise and signal. hashedDrops is the simplest of these tools, assigning hashtags based just on log-fold change of corrected HTO counts. BFF and GMM-Demux both assume a bimodal counts distribution (the former on log-normalized counts, the latter on centered-log ratio (CLR) transformed counts). Demuxmix instead assumes a negative binomial mixture model. DemuxEM uses an expectation-maximization algorithm to assign hashtags. Finally, HTODemux uses a clustering-based approach on CLR transformed values.

**Specific Approach**

We would like to explore the effects of two experimental parameters on the performance of the aforementioned tools – specifically, the effects of varying sample sizes (i.e. number of cells), as well as the quality of the hashtagging. To accomplish this, we will assess algorithm performance on a simulated dataset containing eight samples and eight barcodes.

Generation of simulated dataset: We will generate a total of nine datasets consisting of three sample sizes (4,000, 8,000, and 12,000 cells) across three different sample qualities (high, medium, and low). We define sample quality based on the difference between the means of “positive” and “negative” distributions, with a larger difference being equivalent to higher sample quality. We will divide cells into eight groups of equal size, then generate one random positive HTO vector and seven negative random HTO vectors for each group. We will tentatively use a normal distribution (and rounding) to generate a simulated HTO raw count matrix. To simulate doublets, we will first calculate the multiplet rate based on sample size (using the calculator available here: <https://satijalab.org/costpercell/>), then randomly select that percentage of cells and combine their counts to simulate multiplets. This will give us a matrix of raw counts, the same output that we would obtain from a sequencer. Then, we will normalize and/or transform the data based on the specifications of each tool before running each tool and proceeding to tool performance analysis based on benchmarking metrics described below.

Benchmarking metrics: We will first calculate and compare both the type I (false positive) and type II (false negative) error rates across each tool for both 1) accurate identification of true singlets and 2) identification of true doublets across both different sample sizes and sample qualities. Because doublets can either consist of two cells with the same hashtag (more difficult to accurately identify) or two cells with different hashtags, we will further stratify our assessment of doublet identification rates based on these two categories. We anticipate that most algorithms will perform better at identifying doublets with different hashtags than doublets with the same hashtags. We will also compare how often each tool is unable to assign a cell to a particular hashtag (uncertainty rates).

**Expected Results**

Outcomes of this proposal will provide a comprehensive overview of how sample size and sample quality affect the ability of each tool to accurately identify true singlets and true doublets. Researchers will be able to better understand which tool may better suit their data, especially under non-ideal circumstances where multiplexing quality and/or cell numbers are low. Future best practices may involve exploratory data analysis of hashtagging quality before deciding which demultiplexing tool to use.

We can hypothesize about the performance of each tool under different “experimental” conditions, given the different statistical assumptions that each tool makes. For instance, tools that assume a bimodal distribution (such as BFF) may perform the best on datasets with high signal to noise ratios but may also suffer the greatest decrease in accuracy on datasets with low signal to noise ratios. We also hypothesize that the simplest tool, HashedDrops, will display the highest Type I error rates and will also have the most difficult time identifying doublets due to its relatively unsophisticated assignment of barcodes based on the greatest log-fold change. On the other hand, simple tools like HashedDrops may have better performance in samples with low cell numbers due to their independence from statistical power considerations. We anticipate that more complex algorithms will perform with greater accuracy as sample sizes increase.

**Potential Limitations and Alternative Approaches**

The use of a simulated dataset comes with both advantages and drawbacks. One main advantage is a known ground truth, which can be used to perfectly capture the accuracy of each tool. However, simulating data with simple distributions may oversimplify our model, which means that the findings we report may not necessarily translate well to use in real-world datasets. HTO read counts is likely a function of multiple superimposed distributions, some of which may be related to technical / experimental artifacts, and some of which may be related to biological phenomenon (e.g. cell cycle state). However, elucidating the exact specifics of these distributions is likely beyond the scope of this proposal. As such, if modeling HTO counts using the normal distribution proves to be too reductive, I will explore the use of a negative binomial distribution (which has the added benefit of an integer output similar to an actual raw counts matrix).

Another potential approach is to find real-world datasets from the GEO database, with the caveat that the datasets identified have a “ground truth” of the original sample IDs, either through inherent biological features (e.g. pooling cells from highly distinct cell lines) or somehow via orthogonal labeling of cells. Of course, this may limit the types of analyses we can perform, especially in quantifying accuracy of multiplet assignment when the cells have the same hashtag (no way to use genetic differences to establish ground truth); however, this will also increase the applicability of our findings to use with real data that researchers may generate in the future.

**References**

1. Chen J, Schwarz E. Opportunities and Challenges of Multiplex Assays: A Machine Learning Perspective. In: Guest PC, ed. *Multiplex Biomarker Techniques: Methods and Applications*. Springer New York; 2017:115-122. doi:10.1007/978-1-4939-6730-8\_7

2. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. *Genetics*. 2010;185(2):405-416. doi:10.1534/genetics.110.114983

3. Zhang Y, Xu S, Wen Z, et al. Sample-multiplexing approaches for single-cell sequencing. *Cell Mol Life Sci*. 2022;79(8):466. doi:10.1007/s00018-022-04482-0

4. Brown DV, Anttila CJA, Ling L, et al. A risk-reward examination of sample multiplexing reagents for single cell RNA-Seq. *Genomics*. 2024;116(2):110793. doi:10.1016/j.ygeno.2024.110793

5. Boggy GJ, McElfresh GW, Mahyari E, et al. BFF and cellhashR: analysis tools for accurate demultiplexing of cell hashing data. *Bioinforma Oxf Engl*. 2022;38(10):2791-2801. doi:10.1093/bioinformatics/btac213

6. Hao Y, Hao S, Andersen-Nissen E, et al. Integrated analysis of multimodal single-cell data. *Cell*. 2021;184(13):3573-3587.e29. doi:10.1016/j.cell.2021.04.048

7. Lun ATL, Riesenfeld S, Andrews T, Dao TP, Gomes T, Marioni JC. EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol*. 2019;20(1):63. doi:10.1186/s13059-019-1662-y

8. Xin H, Lian Q, Jiang Y, et al. GMM-Demux: sample demultiplexing, multiplet detection, experiment planning, and novel cell-type verification in single cell sequencing. *Genome Biol*. 2020;21(1):188. doi:10.1186/s13059-020-02084-2

9. Gaublomme JT, Li B, McCabe C, et al. Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat Commun*. 2019;10(1):2907. doi:10.1038/s41467-019-10756-2

10. Klein HU. demuxmix: Demultiplexing oligonucleotide-barcoded single-cell RNA sequencing data with regression mixture models. *BioRxiv Prepr Serv Biol*. Published online January 29, 2023:2023.01.27.525961. doi:10.1101/2023.01.27.525961

11. Muftuoglu M, Li L, Liang S, et al. Extended live-cell barcoding approach for multiplexed mass cytometry. *Sci Rep*. 2021;11(1):12388. doi:10.1038/s41598-021-91816-w

12. Krutzik PO, Clutter MR, Trejo A, Nolan GP. Fluorescent cell barcoding for multiplex flow cytometry. *Curr Protoc Cytom*. 2011;Chapter 6:6.31.1-6.31.15. doi:10.1002/0471142956.cy0631s55

13. Krishhan VV, Khan IH, Luciw PA. Multiplexed microbead immunoassays by flow cytometry for molecular profiling: Basic concepts and proteomics applications. *Crit Rev Biotechnol*. 2009;29(1):29-43. doi:10.1080/07388550802688847

14. Pappireddi N, Martin L, Wühr M. A Review on Quantitative Multiplexed Proteomics. *Chembiochem Eur J Chem Biol*. 2019;20(10):1210-1224. doi:10.1002/cbic.201800650

15. Stoeckius M, Zheng S, Houck-Loomis B, et al. Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol*. 2018;19(1):224. doi:10.1186/s13059-018-1603-1

16. McGinnis CS, Patterson DM, Winkler J, et al. MULTI-seq: sample multiplexing for single-cell RNA sequencing using lipid-tagged indices. *Nat Methods*. 2019;16(7):619-626. doi:10.1038/s41592-019-0433-8

17. Kang HM, Subramaniam M, Targ S, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat Biotechnol*. 2018;36(1):89-94. doi:10.1038/nbt.4042

18. Huang Y, McCarthy DJ, Stegle O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol*. 2019;20(1):273. doi:10.1186/s13059-019-1865-2