**BMI 5332 Project Progress Report**

Timothy Bi

**Significance and Proposed Methods**

Sample multiplexing is often used in single-cell sequencing experiments to reduce cost and account for technical batch effects1–3. Two major technologies are used for sample multiplexing: staining with oligonucleotide-labeled monoclonal antibodies against ubiquitous cell surface markers (like β2 microglobulin)4; or with oligonucleotide-labeled lipids, which can spontaneously incorporate into cell plasma membranes5. Computational tools are necessary for accurate demultiplexing and downstream analysis, as well as cell population recovery. Because each tool has different underlying statistical assumptions, I hypothesized that each tool’s performance would vary under different experimental conditions; therefore, to better understand how various effects (such as sample size, multiplet rate, and staining quality) change demultiplexing performance, I proposed to formally benchmark six tools / algorithms: bimodal flexible fitting (BFF)6, HTODemux7, hashedDrops8, GMM-Demux9, DemuxEM10, and demuxmix11. Of these tools, BFF and GMM-Demux both assume a bimodal counts distribution; demuxmix assumes a negative binomial mixture model; DemuxEM uses an expectation-maximization algorithm; and HTODemux uses a clustering-based approach on center log ratio transformed values.

To benchmark these demultiplexing tools, I proposed to simulate multiple experimental conditions of varying sample sizes (three conditions – 4000, 8000, and 12000 cells) and staining quality (“high”, “medium”, and “low”, with a lower difference of means representing lower staining quality). This would be accomplished by modeling counts using two rounded normal distributions representing positive staining and negative staining. We would assume that each barcoded sample were evenly divided by eight hashtags – thus, for . Multiplet rate would be estimated from the calculator available at <https://satijalab.org/costpercell/>, and counts for multiplets would be calculated by adding together the simulated counts for each cell in the multiplet. Benchmark metrics include Type I and Type II error rates for both single cell and multiplets, as well as rates of assignment uncertainty.

**Expected Results and Current Results**

I anticipate that our findings from this proposal will provide a comprehensive overview of how sample size and staining quality affect each tool’s ability to classify true singlets and true doublets. Specifically, I believe that this project will help me understand which statistical models are best suited for which conditions. I hypothesize that bimodal and simple thresholding methods will perform adequately for conditions of higher staining quality (greater separation between distributions) but will struggle when separation is less pronounced. I also hypothesize that HTODemux, the only clustering-based demultiplexing algorithm, will perform better as sample size increases due to increased cluster sizes and resolution.

As of today, I have successfully implemented a function in R for simulating the experimental data. Count data is modeled by sampling either a positive or a negative distribution. The negative distribution is defined as a negative binomial with mean of 20 and a dispersion parameter of 10. The positive distribution is defined as a negative binomial with a mean of 100, 80, or 60 for high, medium, or low staining quality respectively, and a dispersion parameter of 10. Histograms illustrating separation for hashtag 1 (HTO-1) across staining qualities at 12,000 cells are shown below.

A graph of a number of different sizes

Description automatically generated with medium confidence

In the high- and medium-quality staining, it is apparent that there is a distinct population of cells that are positive for HTO-1, while the positive population for low-quality staining is much less distinct.

We can also plot two simulated hashtags against each other to crudely visualize staining quality. See below example of HTO-1 vs. HTO-2, both raw and log-transformed:

A graph of a number of cells

Description automatically generated A graph of a number of cells

Description automatically generated

We might hypothesize that cells highlighted by the red boxes above (drawn simply by eye) represent doublets that are either composed of two cells tagged with HTO-2, one cell tagged by HTO-1 and one cell tagged by HTO-2, or two cells tagged with HTO-1 (from left to right on the untransformed plot) due to their higher levels of expression. If we color that same plot by said multiplets, however, we observe that actual classification of multiplets can be much trickier than it seems (and that simpler algorithms like the simple threshold used by HashedDrops will perform worse in discriminating between singlets and doublets).

A graph of a number of cells

Description automatically generatedA graph of a number of cells

Description automatically generated

**Deviations from the Original Proposal and Future Directions**

In the original proposal, we proposed to model counts data by sampling a normal distribution and rounding the resulting data to approximate integer count data. Instead, I am using the negative binomial distribution, which is able to more closely approximate real-life HTO count data and which bypasses the necessity of rounding. Otherwise, my approach has not deviated from the original proposal thus far.

For the future, I will begin to implement each of the benchmarking tools and assess their performance across the nine simulated datasets I have generated thus far (low\_4k, low\_8k, low\_12k; med\_4k, med\_8k, med\_12k; and high\_4k, high\_8k, and high\_12k; with “low”, “medium”, or “high” corresponding to different staining qualities, and “4k”, “8k”, and “12k” corresponding to the number of thousands of cells in the sample). I will analyze trends in Type I and Type II error rates across staining quality and sample size to better assess tool behavior under different experimental conditions.

**References**

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

2. 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

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. 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

5. 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

6. 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

7. 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

8. 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

9. 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

10. 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

11. 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