Predicting Cell Types in Non-Human Single Cell RNA Sequencing Data

Sean Connelly **Ryan Videgar-Laird**

Stephanie Ting

1 Introduction

Single cell RNA sequencing data is eventually processed into a gene expression matrix, which has dimensions of m cells x n genes. After performing quality control steps, normalizing and scaling this gene expression matrix, this matrix undergoes dimensionality reduction to compress this dataset that contains many features (genes) to see how these cells relate to one another. The patterns of how cells vary compared to one another can be clustered using any algorithm of one's choosing. The resulting clusters represent certain types of cells that need to be characterized and classified.

Classifying single cells has mainly focused on utilizing three categories of techniques. The first class of techniques annotate these clusters through comparing the differential expression (greater number of counts) of the genes in a given cluster with every other cluster through statistical tests, like the Wilcoxian rank sum test. These differentially abundant genes in each cluster can be compared to experimentally validated genes that characterize cell types to identify the label for that cluster. The second class of techniques correlates the newly generated dataset with a reference dataset to annotate cell clusters. This reference dataset can be other single cell RNA sequencing data or bulk RNA sequencing of specific cell types individually or over a time series. Lastly, the third class of techniques uses supervised learning, which utilizes previously generated data to train a model to classify cell types. Some examples of potential models to implement could be logistic regression, support vector machine or random forest, to name a few.

1.1 Related Work

Many of these tasks have been focused on human data, where there has been much focus on diseases, like different forms of cancer, that affect many people domestically, but not much focus on diseases that have broad impacts globally. One such disease is malaria, a parasitic infection that resulted in 247 million cases in 2021 and causes a large burden of morbidity and mortality globally [4]. Malaria has a complex life cycle, which starts in the mosquito vector to result in infection of a human host. When individuals are sick with malaria, the parasite replicates in the blood cells and causes these blood cells to lyse when the parasite replicates. These replication stages have unique gene expression patterns and single cell RNA sequencing provides a key benefit to further understand how these genes change per cell stage. These same concepts can be applied to other neglected diseases and hosts. From looking at developmental trajectories in zebrafish [2] to COVID-19 response in non-human primates [6] to single cell RNA sequencing of other parasitic diseases [1, 3, 5], [7], we would like to build a powerful, simple and efficient model that can classify cell types on this source of data.

2 Methods

- QC dataset to normalization
- use processed gene expression matrix to perform one hot encoding as our embedding and learn with the one hot encoding and the processed gene exp matrix

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

- implement the single head attention network
- supervised
  - input: normalized gene expression matrix and the one hot encoding
  - single head attention
  - output: classification of cell type
- unsupervised
  - input: normalized gene expression matrix and the one hot encoding
  - single head attention
  - output: embedding for each gene for each cell
  - weights: what are updated during your loss function, can pull and look at them

3 Preliminary Results

3.1 Dataset Preprocessing

We are pre-processing and organizing existing datasets...

Table 1: Summary of datasets that will be used in this study.

Host	**Source**	**Reference**	**Cells**
Parasite	GitHub	Howick et al. [3]	6,737
Parasite	GSE146737	Wendt et al. 2020	43,642
Parasite	Zenodo	Briggs et al. 2021	8,599
Parasite	GitHub	Rezvani et al. 2022	8,719-12,910
Human	TENxPBMCData R package (Kasper et al. 2021)	10X Genomics	3,000
Human	scRNAseq package (Risso and Cole, 2021)	Lawlor et al. 2017	978
Human	GSE151530	Ma et al. 2021 and Ma et al. 2022	56,721
Mouse	scRNAseq package (Risso and Cole, 2021)	Zeisel et al. 2015	3,005
Zebrafish	GSE106587	Farrell et al. 2018	38,731
Non-human primate	GSE156755	Speranza et al. 2021	100,795

3.2 Initial Modeling

As our first run-through, we are using the small P. falciparium dataset from ...

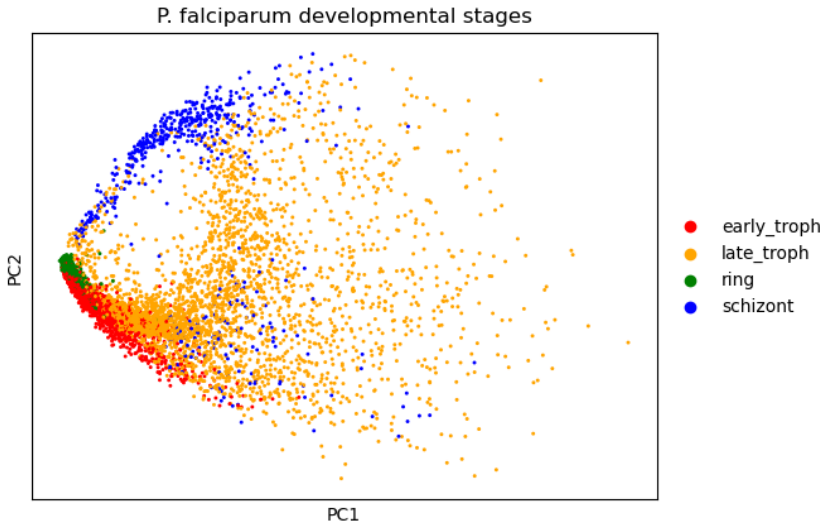
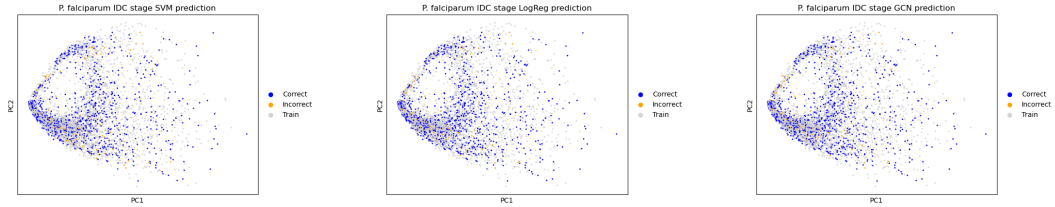


Figure 1: Principal components analysis of *P. falciparum* single cell RNA sequencing data, colored by lifecycle stage.

This dataset is simple enough that most predictors perform quite well...



(a) PCA of predictions from SVM model.

(b) PCA of predictions from logistic regression model.

(c) PCA of predictions from graph convolutional network model.

Figure 2: Model performance visualized on PCA plots.

Each IDC stage can be predicted with good precision and recall.

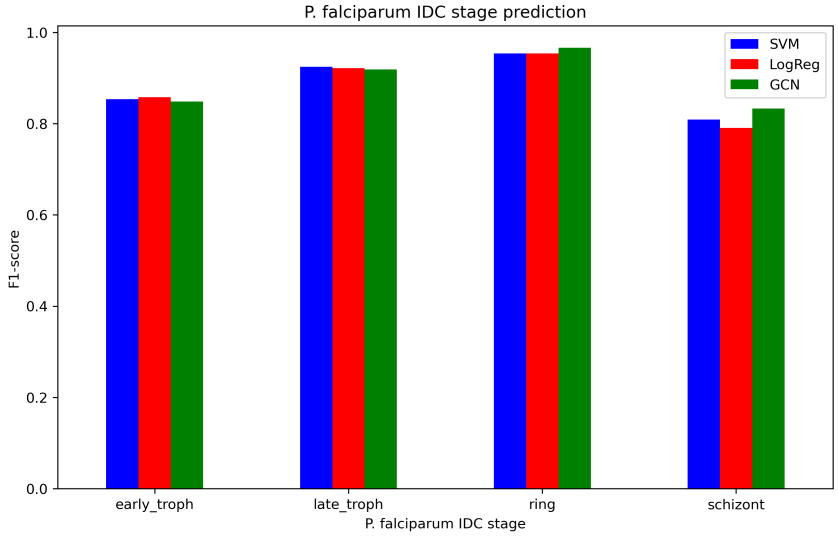


Figure 3: Prediction accuracy of models on *P. falciparum* single cell RNA sequencing data.

4 Future Work

References

[1] Emma M. Briggs, Federico Rojas, Richard McCulloch, Keith R. Matthews, and Thomas D. Otto. Single-cell transcriptomic analysis of bloodstream Trypanosoma brucei reconstructs cell cycle progression and developmental quorum sensing. *Nature Communications*, 12(1):5268, September 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-25607-2.

[2] Jeffrey A. Farrell, Yiqun Wang, Samantha J. Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F. Schier. Single-cell reconstruction of developmental trajectories during zebrafish embryogenesis. *Science*, 360(6392):eaar3131, June 2018. doi: 10.1126/science.aar3131.

[3] Virginia M. Howick, Andrew J. C. Russell, Tallulah Andrews, Haynes Heaton, Adam J. Reid, Kedar Natarajan, Hellen Butungi, Tom Metcalf, Lisa H. Verzier, Julian C. Rayner, Matthew Berriman, Jeremy K. Herren, Oliver Billker, Martin Hemberg, Arthur M. Talman, and Mara K. N. Lawniczak. The Malaria Cell Atlas: Single parasite transcriptomes across the complete Plasmodium life cycle. *Science*, 365(6455):eaaw2619, August 2019. doi: 10.1126/science.aaw2619.

[4] April Monroe, Nana Aba Williams, Sheila Ogoma, Corine Karema, and Fredros Okumu. Reflections on the 2021 World Malaria Report and the future of malaria control. *Malaria Journal*, 21 (1):154, May 2022. ISSN 1475-2875. doi: 10.1186/s12936-022-04178-7.

[5] Yasaman Rezvani, Caroline D. Keroack, Brendan Elsworth, Argenis Arriojas, Marc-Jan Gubbels, Manoj T. Duraisingh, and Kourosh Zarringhalam. Comparative single-cell transcriptional atlases of Babesia species reveal conserved and species-specific expression profiles. *PLoS biology*, 20 (9):e3001816, September 2022. ISSN 1545-7885. doi: 10.1371/journal.pbio.3001816.

[6] Emily Speranza, Brandi N. Williamson, Friederike Feldmann, Gail L. Sturdevant, Lizzette Pérez-Pérez, Kimberly Meade-White, Brian J. Smith, Jamie Lovaglio, Craig Martens, Vincent J. Munster, Atsushi Okumura, Carl Shaia, Heinz Feldmann, Sonja M. Best, and Emmie de Wit. Single-cell RNA sequencing reveals SARS-CoV-2 infection dynamics in lungs of African green monkeys. *Science Translational Medicine*, 13(578):eabe8146, January 2021. ISSN 1946-6242. doi: 10.1126/scitranslmed.abe8146.

[7] George Wendt, Lu Zhao, Rui Chen, Chenxi Liu, Anthony J. O'Donoghue, Conor R. Caffrey, Michael L. Reese, and James J. Collins. A single-cell RNA-seq atlas of Schistosoma mansoni identifies a key regulator of blood feeding. *Science (New York, N.Y.)*, 369(6511):1644–1649, September 2020. ISSN 1095-9203. doi: 10.1126/science.abb7709.