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.

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 [world malaria report]. 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 [cite] to COVID-19 response in non-human primates [cite] to single cell RNA sequencing of other parasitic diseases [citations], we would like to build a powerful, simple and efficient model that can classify cell types on this source of data.

1.1 Related Work

1.2 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

2 Preliminary Results

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

Host	**Source**	**Reference**	**Cells**
Parasite	GitHub	Howick et al. 2019	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

Yang et al. [2022]

References

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, October 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z.

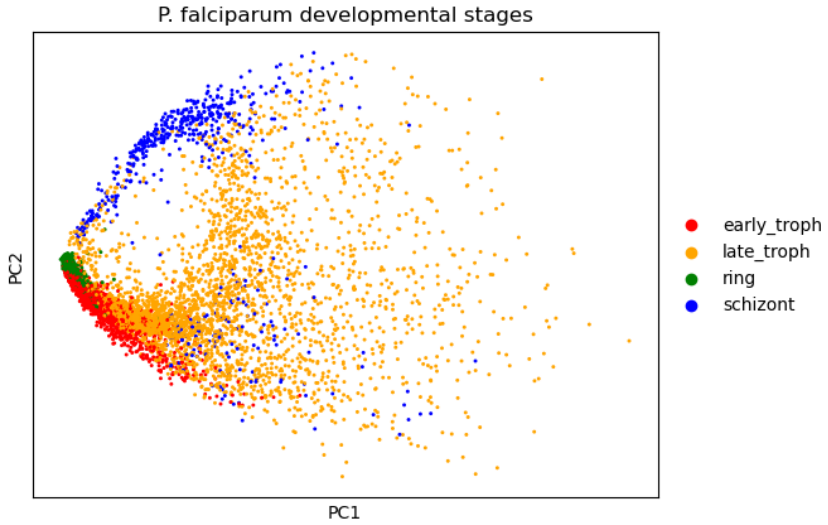
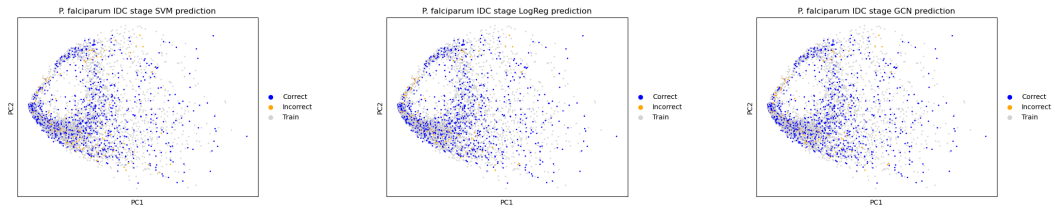


Figure 1: Principal components analysis of *P. falciparum* single cell RNA sequencing data, colored by lifecycle stage.



(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.

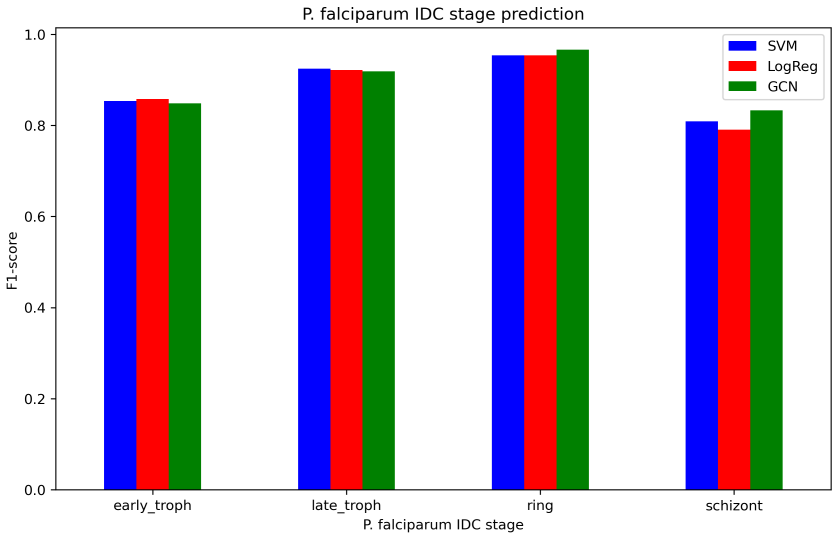


Figure 3: Prediction accuracy of models on *P. falciparum* single cell RNA sequencing data.