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

Sean Connelly **Ryan Videgar-Laird**

Stephanie Ting

1 Introduction

The biology that underlies disease is heterogeneous, with multiple cell types and pathways within those cell types that contribute to the development of disease. All cells have DNA, the genetic code, that is transcribed into messenger RNA (mRNA), called gene expression. After these genes are expressed, the mRNA is then translated into proteins, which go on to perform different functions. mRNA can be expressed at different levels, depending on which proteins are needed by the given cell type. These different levels of mRNA expression can characterize cells and allow further analysis of their contribution to disease biology [1].

The mRNA produced can be quantified at the cellular level, called single cell RNA sequencing (scRNAseq), to obtain a 'fingerprint' of different cell types and can be compared between normal or disease conditions to find how these cell types change. scRNAseq data is 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 [2].

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 Wilcoxon 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. Many recommended pipelines utilize standard methods, such as logistic regression, support vector machine or random forest models [3].

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 [5] to COVID-19 response in non-human

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

primates [6] to single cell RNA sequencing of other parasitic diseases [7, 8, 9], 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

Two previous studies have introduced methods that successfully use attention based models to predict cell type in scRNAseq. The first one, TOSICA, uses a three part approach of cell embedding, multi-head self attention, and then cell type classification [10]. TOSICA is highly accurate for both mouse and human data. This model, however, relies on curated expert knowledge in the cell embedding stage to provide pathway information critical to identifying cell types. The second model, scBERT, uses an adaptation of BERT, a natural language processing model on scRNAseq where the transformer architecture is replaced with a performer in order to better handle larger scale input data [11]. This model is trained with data from millions of cells from a variety of different biological contexts. scBERT not only predicts cell type with very high accuracy, it also has interpretability in that the prediction rediscovers known biomarker genes of each cell type. However, scBERT does require an immense amount of training data and has only been tested on human data. It is also a complex model that is computationally intensive.

Our approach is to use these methods as a foundation to build a simpler model that is less computationally expensive and works on non-human, non-mouse datasets. Our model will also not rely on manual curation of knowledge, which is often inconsistently annotated and formatted.

2 Methods

Anndata [12] and Scanpy [13] were used to create efficient data objects and follow best practices for scRNA-seq analysis. This includes: filtering out cells that express less than 100-200 genes, removing genes that are observed in less than 3 cells, and applying a shifted logarithm transformation. These quality control (QC) steps help reduce noise from experimental prep from sources such as low quality cells and environmental contamination, and stabilize variance across cells [3].

Let y represent total gene counts for a given cell, and s_c represent the size factor for that cell. Counts are then shifted by:

$$f(y) = log\left(\frac{y}{s_c} + y_0\right)$$

The size factor for a given cell designed to account for variation in cell size (larger cells typically contain more genetic transcripts) and sampling effects. Let L represent the median raw count depth across all cells. Then s_c is calculated across all genes, g, per cell:

$$s_c = \frac{\sum_g y_{gc}}{L}$$

More information on developing a transformer model in Future Work section.

3 Preliminary Results

3.1 Traditional Methods

We began with a scRNAseq dataset (*P. falciparum*) from the Malaria Cell Atlas [14] to test out various supervised classification models. We began with this dataset as it is small and therefore easy to use on local machines while gaining familiarity with the methods. In addition, standard techniques have been shown to perform well on it [15], so it serves as a good baseline for later comparisons.

Principal Component Analysis was performed to visualize how the given cells vary compared to one another. The cells are well stratified by developmental stage (Figure 1). To test the performance of two classic models for classification tasks, we used logistic regression (LR) and support vector machines (SVM), implemented through Scikit-learn [16]. To incorporate more graph based information in our model, we used an implementation of a graph convolutional network in the Spektral package [17], based on the work of Kipf and Welling [18]. The data was split into a 70% training and 30% test split. The input to the model was the filtered and normalized gene expression matrix along with the cell type labels. For each model, we labeled the cells in the PCA as correct (blue) or incorrect (orange). Incorrectly predicted cells were spread throughout the PCA and did not concentrate within a certain cell type across all models (Figure 2).

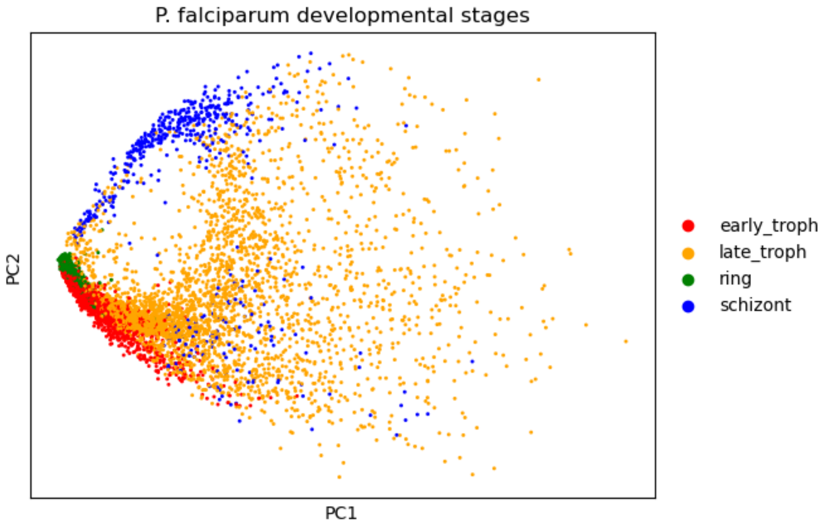
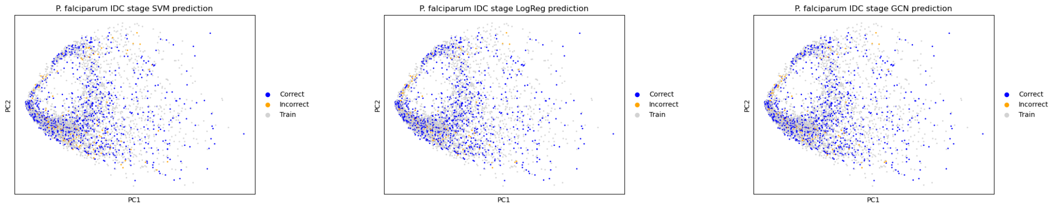


Figure 1: Principal components analysis of *P. falciparum* single cell RNA sequencing data, colored by lifecycle stage.



(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.

In addition, the F1 score was calculated per cell type. All cell type prediction methods were above 80% and very similar among the three models (Figure 3).

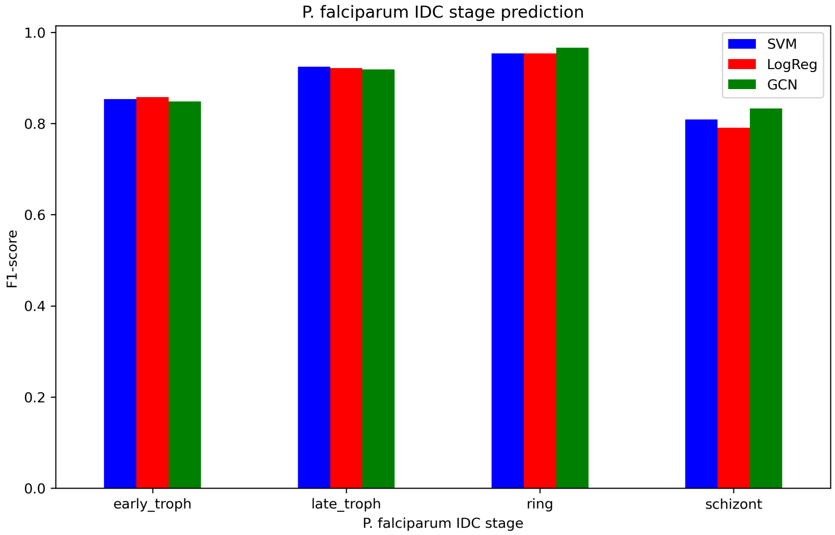


Figure 3: Prediction accuracy of models on *P. falciparum* single cell RNA sequencing data.

3.2 Additional Datasets

We collated a list of scRNAseq datasets, which represent non-human, non-mouse datasets from a wide array of organisms and reference human and mouse datasets (Table 1). The organisms include malaria, parasitic worm *Schistosoma mansoni*, the cause of African sleeping sickness *Trypanosoma brucei*, tick borne parasite *Babesia microti*, zebrafish, and non-human primates. Through training and testing our model on a variety of organisms, we hope to create a model that transcends the need for manual knowledge and can find biomarkers of different cell types to better characterize these diseases.

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

Host	**Source**	**Reference**	**Cells**
Parasite	GitHub	Howick et al. [14]	6,737
Parasite	GEO: GSE146737	Wendt et al. [8]	43,642
Parasite	Zenodo: 5163554	Briggs et al. [7]	8,599
Parasite	Github	Rezvani et al. [9]	8,719-12,910
Human	TENxPBMCData R Package	Hansen et al. [19]	3,000
Human	scRNAseq R Package	Risso and Cole [20], Lawlor et al. [21]	978
Human	GEO: GSE151530	Ma et al. [22]	56,721
Mouse	scRNAseq R Package	Risso and Cole [20], Zeisel et al. [23]	3,005
Zebrafish	GEO: GSE106587	Farrell et al. [5]	38,731
Non-human primate	GEO: GSE156755	Speranza et al. [6]	100,795

4 Future Plans

4.1 Continued Data Curation

The datasets represented in Table 1 have to be accessed from their data repositories to obtain the cell x gene matrix, which will undergo quality control filtering and normalization.

4.2 Cross Validation

Rather than simple test/train splits, we will implement cross-validation across all tested models.

4.3 Transformer Model

We are working on implementing a minimal single head attention network using Pytorch [24], based on the defining paper Attention is All you Need [25]. Yet there are several challenges we must address in order to generalize 'standard' models for use with gene expression data. First, gene expression, as currently measured, is inherently unordered, i.e. there is no need for positional encoding of each token (gene). Therefore it would be ideal to not limit the input sequence length so global gene-gene interactions can be captured. However, the memory and time complexity scale quadratically with input length, which limits standard models to an input length of around 500 in practice. There have been several publications, including Reformer and Performer, that each aimed to improve transformer efficiency [26, 27, 28, 29, 30, 31]. We are working to adapt the Performer method to our simpler model.

In the human and mouse datasets that can be input into scBERT and TOSICA, we will implement these methods and classical methods, such as logistic regression, random forest and support vector machines, to compare the classification accuracy of our method.

4.4 Reproducibility

Although there is a growing awareness of the 'reproducibility crisis' in research, many computational biological studies continue to only make part of their code and analysis publicly available. We are striving to proactively organize our analysis in an open and easy-to-reproduce structure. This includes, but is not limited to, using: pinned dependency and dataset versions, Docker, Git, and Make/Snakemake. Our final analysis will be available at: https://github.com/RyanVidegar-Laird/gettention (bad name subject to change).

4.5 Division of Work

We will split the preparation of the remaining 9 datasets in thirds, where Sean will prepare the three parasite datasets, Ryan will prepare the three human datasets and Stephanie will prepare the mouse, zebrafish, and other non-human primate datasets.

For the final report, Stephanie will expand on the introduction and related work. Ryan will write up the methods and some of the results. Sean will help write up the results and conclusion. We will all contribute to editing the final report.

References

[1] Ignacio San Segundo-Val and Catalina S. Sanz-Lozano. Introduction to the Gene Expression Analysis. *Methods in Molecular Biology (Clifton, N.J.)*, 1434:29–43, 2016. ISSN 1940-6029. doi: 10.1007/978-1-4939-3652-6_3.

[2] Aisha A. AlJanahi, Mark Danielsen, and Cynthia E. Dunbar. An Introduction to the Analysis of Single-Cell RNA-Sequencing Data. *Molecular Therapy. Methods & Clinical Development*, 10: 189–196, September 2018. ISSN 2329-0501. doi: 10.1016/j.omtm.2018.07.003.

[3] Lukas Heumos, Anna C. Schaar, Christopher Lance, Anastasia Litinetskaya, Felix Drost, Luke Zappia, Malte D. Lücken, Daniel C. Strobl, Juan Henao, Fabiola Curion, Herbert B. Schiller, and Fabian J. Theis. Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, 24(8):550–572, August 2023. ISSN 1471-0064. doi: 10.1038/s41576-023-00586-w.

[4] World Health Organization. *World Malaria Report 2022*. World Health Organization, December 2022. ISBN 978-92-4-006489-8.

[5] 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.

[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] 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.

[8] 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.

[9] 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.

[10] Jiawei Chen, Hao Xu, Wanyu Tao, Zhaoxiong Chen, Yuxuan Zhao, and Jing-Dong J. Han. Transformer for one stop interpretable cell type annotation. *Nature Communications*, 14(1):223, January 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-35923-4.

[11] 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.

[12] Isaac Virshup, Sergei Rybakov, Fabian J. Theis, Philipp Angerer, and F. Alexander Wolf. Anndata: Annotated data, December 2021.

[13] F. Alexander Wolf, Philipp Angerer, and Fabian J. Theis. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology*, 19(1):1–5, December 2018. ISSN 1474-760X. doi: 10.1186/s13059-017-1382-0.

[14] 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.

[15] Swarnim Shukla, Soham Choudhuri, Gayathri Priya Iragavarapu, and Bhaswar Ghosh. Supervised learning of Plasmodium falciparum life cycle stages using single-cell transcriptomes identifies crucial proteins. *Journal of Bioinformatics and Systems Biology*, 06(01), 2023. ISSN 26885107. doi: 10.26502/jbsb.5107047.

[16] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928.

[17] Daniele Grattarola and Cesare Alippi. Graph Neural Networks in TensorFlow and Keras with Spektral, June 2020.

[18] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, February 2017.

[19] Kasper Hansen, Davide Risso, and Stephanie Hicks. TENxPBMCData. Bioconductor, 2018.

[20] Davide Risso and Michael Cole. scRNAseq. Bioconductor, 2017.

[21] Nathan Lawlor, Joshy George, Mohan Bolisetty, Romy Kursawe, Lili Sun, V. Sivakamasundari, Ina Kycia, Paul Robson, and Michael L. Stitzel. Single-cell transcriptomes identify human islet cell signatures and reveal cell-type–specific expression changes in type 2 diabetes. *Genome Research*, 27(2):208–222, February 2017. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.212720.116.

[22] Lichun Ma, Sophia Heinrich, Limin Wang, Friederike L. Keggenhoff, Subreen Khatib, Marshonna Forgues, Michael Kelly, Stephen M. Hewitt, Areeba Saif, Jonathan M. Hernandez, Donna Mabry, Roman Kloeckner, Tim F. Greten, Jittiporn Chaisaingmongkol, Mathuros Ruchirawat, Jens U. Marquardt, and Xin Wei Wang. Multiregional single-cell dissection of tumor and immune cells reveals stable lock-and-key features in liver cancer. *Nature Communications*, 13 (1):7533, December 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-35291-5.

[23] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler, and Sten Linnarsson. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science*, 347(6226):1138–1142, March 2015. doi: 10.1126/science.aaa1934.

[24] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, August 2023.

[26] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention, August 2020.

[27] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, June 2019.

[28] Stephen Merity. Single Headed Attention RNN: Stop Thinking With Your Head, November 2019.

[29] Krzysztof Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Davis, Afroz Mohiuddin, Lukasz Kaiser, David Belanger, Lucy Colwell, and Adrian Weller. Rethinking Attention with Performers, November 2022.

[30] Nikita Kitaev, Łukasz Kaiser, and Anselm Levskaya. Reformer: The Efficient Transformer, February 2020.

[31] Apoorv Vyas, Angelos Katharopoulos, and François Fleuret. Fast Transformers with Clustered Attention, September 2020.