International Journal Of Public Mental Health And Neurosciences
ISSN: 2394-4668
(An Official publication of Sarvasumana Association)

Unravelling Plasmodium Falciparum's Life Cycle: Insights from Integrated Transcriptomic and Machine Learning Approaches

Shreeya Pahune and Bhaswar Ghosh
Center for Computational Natural Sciences and Bioinformatics
Institute of Information Technology - Hyderabad
Hyderabad, India
shreeya.pahune@research.iiit.ac.in; bhaswar.ghosh@iiit.ac.in

***Abstract*—This study uses a transcriptomic and machine learning-based approach to gain a comprehensive understanding of the genomic landscape of Plasmodium falciparum, a malaria-causing parasite. This enables us to understand this life-threatening disease better and thus identify new treatment targets and intervention methods. We used single-cell RNA sequencing data from the Malarial Cell Atlas with 5,177 genes across 37,624 cells. Our approach combined static analyses, including highly variable gene identification, differential expression analysis, and M3Drop (a feature selection method based on noise filtration), to obtain stage-specific markers along with trajectory analysis (Monocle3) to identify top genes active during stage progression. The feature selection analyses were done for each stage and lifecycle gene, which are genes that remain active throughout the lifecycle. These complementary approaches gave us distinct gene sets: static analysis revealed genes necessary for stage-specific functions, while dynamic trajectory analysis highlighted genes that play an important role in cellular development. A neural network classifier trained on these gene sets achieved high accuracy (Accuracy: 96%; F1-score: 0.96) using trajectory-derived genes, performing better than the model trained on the entire gene set, indicating that the gene set has captured stage-specific signatures. Further, to validate the biological significance of the shortlisted genes, we performed a Gene Ontology analysis. We found that our analysis revealed results that are in line with existing literature. The ring stage genes were linked to immune evasion, trophozoite genes to haemoglobin digestion, schizont genes to merozoite invasion, and gametocyte genes to sexual differentiation. We plan to build on this work by studying the behaviour of important genes using a Poisson-Beta model to infer the kinetics of key genes, which will help us better understand how they change throughout the parasite's life cycle.**

***Index Terms*—Plasmodium falciparum, Single-cell RNA sequencing, Gene expression profiling, Machine learning in parasitology**

I. INTRODUCTION

The Plasmodium species are known to cause Malaria, a life-threatening disease that has been a global health concern, especially in developing nations within sub-Saharan Africa, Asia and Latin America [1]. Within the Plasmodium species, Plasmodium falciparum is the most virulent. It is responsible for severe cases of Malaria and presents a significant challenge in the prevention and control of the disease. Thus, a comprehensive understanding of its life cycle at a genomic level would immensely help develop advanced clinical treatments and specific targets. The complex life cycle of a Plasmodium parasite involves many stages in different hosts, some of which take place in a mosquito, while the rest occur in the mammalian host. Each of the distinct stages, namely, sporozoite, merozoite, ring, trophozoite, schizont, and gametocyte, are all characterised by unique biological and genomic signatures [2].

In particular, with advancements taking place in the single-cell RNA sequencing (scRNA-seq) technology, we are able to perform detailed analyses on the gene expression data, enabling us to gain novel insights into cellular processes and gene expression changes at a single-cell level [3, 4]. Additionally, considering the Plasmodium life cycle, scRNA-seq has enabled the identification of high-variance genes (HVGs) and differentially expressed genes (DEGs) [5, 6]. These genes play a critical role in different stages of the cycle. Moreover, identifying these transcriptional markers in Malaria has the potential to help us understand regulatory mechanisms that are crucial for the parasite's interaction with its host [7]. Some studies have also demonstrated that stagewise expression data in P. falciparum contribute to adaptive mechanisms against host immune responses, especially in the asexual and gametocyte stages [8]. This emphasises the potential for targeted therapies. We can discover more about the underlying biological processes that facilitate the parasite's development and bring about the progression of disease using transcriptome data from every stage of the life cycle.

To address these challenges, we present an analytical approach using transcriptomic techniques along with machine learning (ML) models to add to our understanding of P. falciparum. Using scRNA-seq data, we identify gene sets that capture changes across the different stages of the Plasmodium life cycle. We use three feature selection methods and perform a trajectory analysis to identify key genes that are important during stage-specific as well as stage transition phases. Further, we implement a neural network-based classification model to evaluate these gene sets. This approach allows us to capture stage-specific gene expression signatures that may be promising targets for malaria intervention and further aid our understanding of malaria biology.

II. Results

A. Data Processing

We performed the analysis on single-cell RNA sequencing data obtained from the Malarial Cell Atlas, consisting of 5,177 genes across 37,624 cells. The dataset was well-balanced across the four stages of P. falciparum lifecycle (Table 1).

There was no loss in the number of genes or cells during our processing, as none showed zero variance after RPKM normalization, log transformation, and scaling. After processing, we performed dimensionality reduction using 30 principal components for UMAP. The UMAP in Figure 1 (plotted with just the first two principal components) used the complete processed dataset and resulted in distinct clusters of cells based on the stages as per their label. The UMAP visualization demonstrates distinct transcriptional profiles for each developmental stage, with gametocytes showing the most divergent expression pattern compared to the asexual stages (Figure 1). We observe a continuous trajectory within the asexual stages, indicating the progression of the lifecycle. However, the separation of the gametocytes, the only sexual stage, indicates significant changes in the expression patterns.

TABLE I. Distribution of cells across different stages of *P. falciparum*

Stage	Number of Cells
Trophozoite	13,436
Gametocyte	8,958
Schizont	8,159
Ring	7,071
Total	***37,624***

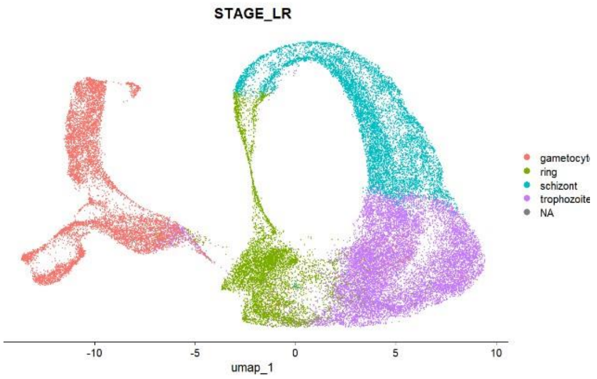

Fig. 1. UMAP visualization of P. falciparum single-cell RNA sequencing data coloured by developmental stage.

B. Feature Selection Identifies Stage-Specific Gene Signatures

To identify essential genes that would help distinguish between the stages, we used three distinct feature selection methods: highly variable genes (HVGs), differentially expressed genes (DEGs), and dropout-based feature selection (M3Drop). We applied these techniques to each stage and the complete dataset. We shortlisted the top 3000 HVGs and DEGs each. Using M3Drop, we obtained 1196 genes for the ring stage, 2088 for schizont, 2588 for trophozoite and 2103 for the gametocyte stage.

For further analysis, we only considered the intersection of all three methods to get a comprehensive gene set that ensured stage characteristics were captured for each method. The intersected list for each stage contained 1146 genes for the ring stage, 1099 genes for schizont, 1602 for trophozoite and 1182 for gametocyte. When we combined these three asexual stages, it resulted in 2394 common intersection genes, which represent the common expression patterns present across the asexual stages. There were 2020 genes in the feature sets for the complete data across all three methods. Additionally, we found 1818 lifecycle genes, which consist of genes operational throughout the lifecycle but do not contain stage-specific signatures. These signify that one-fifth of the genes present in the P. falciparum genome may be contributing to its development irrespective of stage.

C. Trajectory Analysis Reveals Dynamic Gene Expression Patterns

In this analysis, we obtain genes that change dynamically as a function of pseudo-time when the lifecycle transitions from one stage to another. Figure 2 depicts the trajectory plot for the complete dataset, where the time progresses from Day 0 in dark blue to Day 10 in yellow. When the spatial arrangement is mapped to the UMAP (Figure 1), it becomes clear that the trajectory progresses from the asexual stages in the early days and ends with the gametocyte stage. There is a continuous path within the asexual stages, which then separates into the gametocyte phase, thus marking a key change in dynamics.

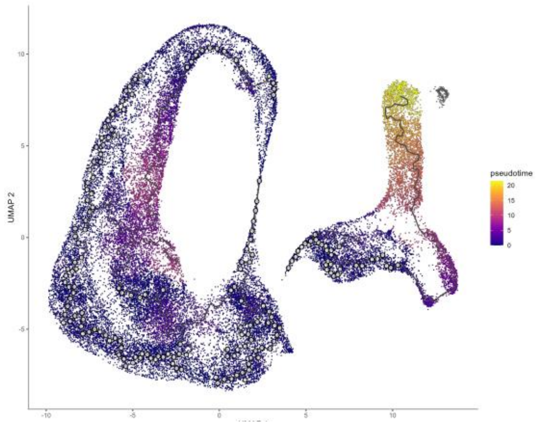

Fig. 2. Pseudo-time trajectory analysis of *P. falciparum* cells showing developmental progression

Stage-wise trajectory plots revealed multiple branches in the ring stage that account for the various development options available to the parasite early in its lifecycle. The schizont plot showed a distinctive divide into two parts suggesting alternative options, whereas we observed a horseshoe pattern in the trophozoite stage. These stage-specific plots help us gain deeper insights into the temporal changes that take place during development.

D. Neural Network Classification Performance

Our neural network classifier demonstrated robust performance across different feature sets (Table 2). Our benchmark model, using all the genes of the dataset, gave us an accuracy of 95.6%. This is marginally superseded when we use the trajectory-based feature set of the top 1000 genes. It achieved the highest overall accuracy of 96%, suggesting that it captured the stage-specific signature the best. We observed trends across stages, and the model demonstrated high reliability in classifying the schizont stage with precision ranging from 0.97 to 0.99, as well as the gametocyte stage (precision: 0.96-0.98).

TABLE II. CLASSIFICATION PERFORMANCE METRICS ACROSS DIFFERENT FEATURE SETS

Input Feature Set	Number of Features	Accuracy	F1-Score
All Genes	5,177	0.956	0.96
Intersection of HVGs, DEGs, M3Drop genes	2,753	0.948	0.95
Top 1000 from trajectory analysis	2,288	0.960	0.96
Top 500 from trajectory analysis	1,170	0.942	0.94
Union set with Top 1000 from trajectory analysis	3,908	0.956	0.96
Union set with Top 500 from trajectory analysis	3,360	0.951	0.95

E. Gene Ontology Enrichment Analysis

We performed a GO enrichment analysis utilizing all the feature sets across each stage. For the ring stage, genes highlighted GO terms related to host cell invasion and protein synthesis, and trophozoite genes highlighted metabolic processes and haemoglobin digestion. For the schizont-specific genes, we observe different terms being associated with them depending on the method used. For the first method (Section 2.2), terms were linked to cell division and merozoite formation, while genes from the trajectory analysis (Section 2.3) enriched for transcriptional regulation and cell cycle progression. Gametocyte-related genes, the only sexual stage, were enriched for sexual development and transmission-related functions. Further, most of the terms associated with feature sets from the trajectory analysis result in themes revolving around dynamic processes involved in stage transitions. These include processes like cell cycle regulation, metabolic switching, etc.

III. METHODS

We aimed to develop an approach that integrates transcriptomic methods along with ML-based models to enhance our understanding of the malaria parasite Plasmodium falciparum using Raw single-cell RNA sequencing (scRNA-seq) data. Gene sets that capture static and dynamic changes across the different stages of the lifecycle were used, and then their biological relevance was explored.

A. Data Processing

Raw single-cell RNA sequencing data from the Malarial Cell Atlas is downloaded for Plasmodium falciparum from [9]. We used data only from the Chromium 10x platform rather than SmartSeq2, as we wanted to focus on mammalian host cells. The data processing was done using the Seurat package (v4.0.0) [10] (R version 4.1.0). Our raw dataset contained 5,177 genes across 37,624 cells and was first normalised using reads per kilobase million (RPKM). This was followed by log normalisation with a factor of 1e6, and then we performed scaling to regress the effects of input variables. We then prepared stage-wise datasets for the four distinct stages present in the dataset: ring, trophozoite, schizont, and gametocyte.

B. Identification of Key Genes by Stage and Lifecycle

We use three different methods to shortlist features from our initial 5,177 genes within each stage and across the dataset. The first method we deployed in our study included finding the most highly variable genes (HVGs). Utilising Seurat's FindVariableFeatures function with the variance stabilising transform (VST), we identified the top 3,000 genes. Next, we found the top differentially expressed genes (DEGs) to identify the most active genes in different stages. Pre-processing steps for this method involved performing dimensionality reduction using the first 30 principal components and clustering with a resolution of 0.5. The third feature selection method fits a model to identify genes with significant dropout patterns in the single-cell RNA-seq data using M3Drop [11]. As our final step, we identified "lifecycle genes". These are genes significant throughout the malarial lifecycle rather than in a particular stage. To obtain this gene set, we subtracted the intersection of the feature set across the dataset with the union of stage-specific genes for every feature selection method. The different feature selection methods allowed us to identify genes that are active within each stage and across the lifecycle.

C. Trajectory Analysis

To map the progression of the cell development, we performed a trajectory analysis using Monoocle3 [12]. The processed scRNA-seq and cell metadata were used as input. As part of the analysis, we began by creating a UMAP, using 30 principal components from PCA, in order to identify essential patterns in gene expression and, at the same time, maintain cell relationships. A trajectory was plotted in pseudo-time using the Day 0 cells as the starting point of reference (root cells). We select two feature sets each, the top 500 and 1000 genes, based on the adjusted p-values. This gave us gene sets that are most associated with the progression of the parasite's gene expression as it transitions from one stage to another. Similar to the feature selection methods, we repeated this analysis within each of the four stages and across the complete dataset.

Performing the trajectory analysis over cells ranging from Day 0 to Day 10, we were able to obtain genes that change with the parasite's lifecycle progression.

D. Neural Network-based Stage Classification

We designed a neural network to classify malaria stages to understand which approach helped us identify stage-wise signatures the best. To see the baseline performance of our model, we trained it with the complete scRNA-seq matrix as the input and used this benchmark model to compare the performance of the rest of the feature sets. We gave the following five feature sets as input:

1. Intersection of the gene sets across the three static feature selection methods used to identify key genes.
2. Top 1000 genes from the trajectory analysis.
3. Top 500 genes from the trajectory analysis.
4. Union of the static gene set (1) and the top 1000 genes from the trajectory analysis.
5. Union of the static gene set (1) and the top 1000 genes from the trajectory analysis.

The goal was to determine if the static, that is, stage-specific focused analysis using three methods, dynamic using trajectory analysis or the former two combined, giving complementary insights, gave us better insights into understanding stages of Malaria. This could aid the development of targeted treatments for Malaria.

The model is a neural network which has an input layer with the dimensionality of the input feature set size, two hidden layers consisting of 256 and 128 nodes that finally classify the gene expression sample into one of four stages. Prior to training, we standardised the data and had a train, validation and test data split of 70%, 10% and 20% each. The performance of the models was evaluated using their accuracy and F1-score, as we did not observe any significant class imbalance.

Our public GitHub repository (https://github.com/SP9144/MalariaInsightsTranscriptomicML) contains all the code.

E. Gene Function Analysis

To understand the biological relevance of our static and dynamic feature analysis, we performed a gene ontology analysis on each feature set using gProfiler [13] with the P. falciparum 3D7 database. A p-value threshold of 0.05 was used to identify significant biological processes.

IV. Discussion

Our integrated approach of combining transcriptomic analysis along with machine learning gave us novel insights into the stage-specific gene expression patterns of P. falciparum. Our neural network model achieved a high accuracy (96%) using only the top 1000 genes that change as a function of pseudo-time. This suggests that the trajectory analysis was successfully able to capture genome-level signatures that are specific to each stage. The performance was better than our baseline model using the full gene set.

As part of our feature analysis, we shortlisted lifecycle genes. These helped us gain insights into the core processes taking place throughout the P. falciparum lifecycle regardless of the stage. Representing nearly one-fifth of the parasite's genome, they maintain essential cellular functions throughout development. This is validated by the proportion of genes inherently expressed genes in Plasmodium based on the study conducted by Bozdech et al. [14].

The Monocle3-based trajectory analysis, in addition to contributing the best-performing feature set, revealed several interesting patterns that occur during the transition of stages. The branching out of the gametocyte stage from the asexual stages, breaking up the continuous trajectory of the latter three stages, is consistent with existing literature [15] and can be attributed to the sexual development of the parasite. Further, in stage-specific analysis, Brancucci et al.'s [16] findings pertaining to environmental sensitivity during early development were supported by the trends of multiple branches we observed in the ring stage. These are the developmental decision points in the earliest life stages that may be critical for the parasite's adaptability.

The GO enrichment analysis results confirmed the potential in our approach by confirming known stage-specific functions while also highlighting new potential targets. The significant association of the ring stage genes with host cell invasion mechanisms aligns with previous literature about the importance of this stage in the initial phases of infection [17]. Metabolic processes enriched by the trophozoite-specific genes supported this stage's already known role in nutrient acquisition and growth of the parasite [18]. Recent research by Josling et al. [19] on the molecular mechanisms driving sexual commitment in

Plasmodium supports the distinctive transcriptional profile of gametocyte-specific genes.

Using these complementary approaches, we achieve a holistic understanding of the genomic landscape and underlying changes during the different stages of the P. falciparum lifecycle. The trajectory analysis captures the complex regulatory networks influencing Plasmodium development, wherein we identified genes involved in stage transitions. In contrast, the static analysis found markers distinct to each stage.

V. Conclusion

Our study demonstrates the effectiveness of applying transcriptomic analyses along with machine learning approaches to understand the complex nature of *P. falciparum* biology. The trajectory analysis identifies genes that alter their expression patterns as a function of pseudo-time, revealing new insights into developmental progression. On the other hand, the high performance of our neural network-based stage classifier and gene ontology analysis of the various feature sets reinforces the potential of our approach and validates its biological correctness and relevance. Potential targets for therapeutic intervention at different stages of the parasite's development can be identified using both stage-specific and lifecycle genes. Stage-specific treatments can be developed using the unique gene patterns revealed for each stage, particularly during the transition points that are highlighted by trajectory analysis. As part of future work, we want to extend the study by modelling the behaviour of key genes using a PoissonBeta model. This will enable a deeper understanding of their kinetics throughout the lifecycle, providing novel insights into the temporal dynamics of gene expression. Our combined static and dynamic analyses could be applied to explore other aspects of *Plasmodium* biology and related parasitic diseases.

References

[1] World Health Organization: World malaria report 2023 (2023)

[2] Vaughan, A.M., et al.: The life cycle of Plasmodium. Nature Reviews Microbiology 16(9), 574–586 (2018)

[3] Macosko, E.Z., et al.: Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161(5), 1202–1214 (2015)

[4] Stoeckius, M., et al.: Cell hashing with barcoded antibodies enables multiplexing and doublet detection for single cell rna sequencing. Nature Methods 14(5), 453– 458 (2017)

[5] Klein, A.M., et al.: Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161(5), 1187–1201 (2015)

[6] Schmitt, T., et al.: Single-cell rna sequencing: A new tool to understand Plasmodium biology. Trends in Parasitology 34(7), 550–563 (2018)

[7] McCarthy, J.S., al.: Integrated analysis of single-cell data reveals a developmental transition in Plasmodium falciparum. Nature Communications 11, 1173 (2020)

[8] Walzer, K.A., al.: Single-cell analysis reveals distinct transcriptional landscapes of human malaria parasites in vivo. Nature Microbiology 3(5), 1010–1017 (2018)

[9] Howick, V.M., Russell, A.J., Andrews, T., Heaton, H., Reid, A.J., Natarajan, K., Butungi, H., Metcalf, T., Verzier, L.H., Rayner, J.C., Berriman, M., HertzFowler, C., Filipe, A.N., Campino, S., Blagborough, A.M.: Single-cell atlas of the malaria parasite Plasmodium falciparum. Science 365, 1110–1114 (2019)

[10] Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., Hao, Y., Stoeckius, M., Smibert, P., Satija, R.: Comprehensive integration of single-cell data. Cell 177, 1888–1902 (2019)

[11] Andrews, T.S., Hemberg, M.: M3drop: dropout-based feature selection for scrnaseq. Bioinformatics 35, 2865–2867 (2019)

[12] Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N.J., Livak, K.J., Mikkelsen, T.S., Rinn, J.L.: The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nature Biotechnology 32, 381–386 (2014)

[13] Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., Vilo, J.: g:profiler: a web server for functional enrichment analysis and conversions of gene lists. Nucleic Acids Research 47, 191–198 (2019)

[14] Bozdech, Z., Llin´as, M., Pulliam, B.L., Wong, E.D., Zhu, J., DeRisi, J.L.:

The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. PLoS biology 1(1), 5 (2003)

[15] Bancells, C., Llor`a-Batlle, O., Poran, A., N¨otzel, C., Rovira-Graells, N., Elemento, O., Kafsack, B.F.C., Cort´es, A.: Revisiting the initial steps of sexual development in the malaria parasite plasmodium falciparum. Nature Microbiology 4(1), 144–154 (2019)

[16] Brancucci, N.M., Bertschi, N.L., Zhu, L., Niederwieser, I., Chin, W.H., Wampfler, R., Freymond, C., Rottmann, M., Felger, I., Bozdech, Z., Voss, T.S.: Probing plasmodium falciparum sexual commitment at the single-cell level. Nature 7, 1–14 (2017)

[17] Cowman, A.F., Healer, J., Marapana, D., Marsh, K.: Malaria: Biology and disease. Cell 167(3), 610–624 (2016)

[18] Crabb, B.S., Koning-Ward, T.F., Gilson, P.R.: Protein export in plasmodium parasites: From the endoplasmic reticulum to the vacuolar export machine. International Journal for Parasitology 40(5), 509–513 (2010)

[19] Josling, G.A., Williamson, K.C., Llin´as, M.: Regulation of sexual commitment and gametocytogenesis in malaria parasites. Annual Review of Microbiology 74, 155–175 (2020)