# PSEUDOTIMEDE ANALYSIS OF SINGLE CELL RNA SEQUENCING MALARIA DATA

by

James Young

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Mathematics

Charlotte

2024

Approved by:

_____
Dr. Shaoyu Li

_____
Dr. Yinghao Pan

_____
Dr. Yanqing Sun

ABSTRACT

JAMES YOUNG. PseudotimeDE Analysis of Single Cell RNA Sequencing Malaria Data.
(Under the direction of DR. SHAOYU LI)

Malaria is a preventable disease that kills hundreds of thousands of people worldwide each year. Focusing on the sporozoite stage of the parasite *Plasmodium berghei*, we identified 561 differentially expressed genes (DEG) across different developmental stages using a dataset published in *Nature*, which contained salivary gland (SG) and midgut (MID) sporozoites. A recently published differential expression method PseudotimeDE was used for the DEG identification. We compared the DEG results with the reference paper, finding that PseudotimeDE identified fewer DEG than the current method, tradeSeq, for multiple possible reasons, and found several vaccine target genes. Finally, we ran Gene Ontology (GO) enrichment and identified additional potential malaria treatment genes, which will require future research to test and confirm.

# ACKNOWLEDGMENTS

I would like to thank Dr. Li for her help and weekly feedback on this thesis. I would also like to thank Dr. Pan and Dr. Sun for taking the time to serve on the committee.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# LIST OF ABBREVIATIONS

| | |
|---|---|
| BP | Biological process |
| CC | Cellular component |
| CS Protein | Circumsporozoite protein |
| DE | Differential expression |
| DEG | Differentially expressed gene(s) |
| FDR | False discovery rate |
| GO | Gene ontology |
| MID | Midgut |
| MF | Molecular function |
| NB-GAM | Negative binomial-generalized additive model |
| PBANKA | *Plasmodium berghei* ANKA |
| PMX | Plasmepsin X |
| RNA | Ribonucleic acid |
| RP | Ribosomal protein |
| SG | Salivary gland |
| scRNA-seq | Single-cell RNA sequencing |
| snRNP | Small nuclear ribonucleaoprotein |
| UMAP | Uniform manifold approximation and projection |
| UIS | Up-regulated in infective sporozoites |
| ZINB-GAM | Zero-inflated negative binomial-generalized additive model |

INTRODUCTION

## Background and Goals

Malaria is an ongoing problem in many countries throughout the world, occurring most commonly in tropical and subtropical climates on or near the equator. The regions most affected are Africa, followed by South and Central America and Asia[1]. Malaria has a whole host of symptoms ranging from the mild symptoms of fever, headache and chills to more severe symptoms of extreme tiredness, difficulty breathing, and jaundice, whereby the eyes and skin of the person become yellowed.

Overall, malaria causes hundreds of thousands of deaths globally every year. In 2022, the number was estimated to be 608,000, 95% of which occurred in Africa, most of which occur in children, due to a lack of previous exposure and immunity to the disease. Adults who get malaria were often exposed to the disease as children and survived, building the necessary immunity or partial immunity to survive subsequent infections[2].

Most cases of malaria in humans are caused by a bite from an infected female mosquito of the *Anopheles* genus. Male mosquitoes do not transmit the disease as they do not bite humans. Other ways malaria spreads includes transfusing blood from an infected person and using a contaminated needle. It can be prevented by avoiding the infectious bite in the first place or by taking one of the two available vaccines. If infection has already occurred, it can be treated using an anti-malaria drug[2].

To understand how malaria infection occurs, it is important to first understand the parasites which cause the infection itself: the organisms of the *Plasmodium* genus. There are three stages in the life cycle of this parasite: gametocytes, sporozoites, and merozoites[3]. When a

mosquito bites an infected host, both male and female gametocytes (the sexual parasite stage) can be taken up by it. The gametocytes then travel to the mosquito Midgut (MID), where they are fertilized and mature into ookinetes. These ookinetes then travel to the hemocoel, where they can develop into oocysts and form into sporozoites. Some of these sporozoites travel to the mosquito's Salivary Glands (SGs), where they can then bite a human and transfer the sporozoites to them. The sporozoites then travel to the liver and become merozoites, which can eventually develop into gametocytes and be taken up by a new mosquito, beginning the process over again[4].

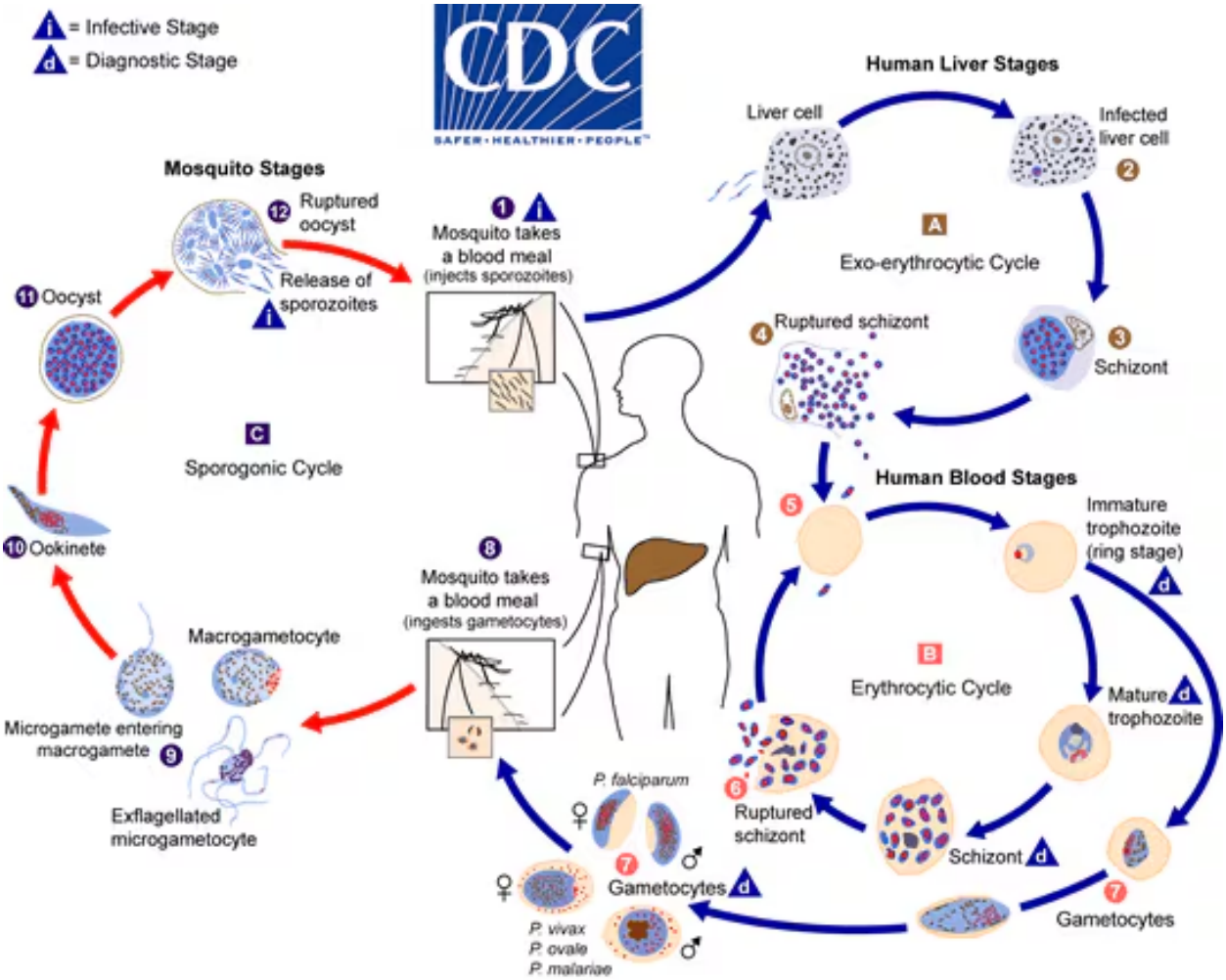



Figure 1: The life cycle of a malaria parasite. Image credit: https://www.cdc.gov/malaria/about/biology/index.html

The primary objective of this paper is to better understand, at a genomic level, the differences between the sporozoites that make it to the SGs and those that do not. In doing so, it is hoped that some insight may be gleaned on how to prevent or treat malaria. This could be accomplished in at least three ways: firstly, by targeting a component of the sporozoite that is more active in those that reach the SGs, damaging them before transmission to humans, secondly, by triggering something in the sporozoites that causes them to remain in the MID, thus stopping the life cycle of the *Plasmodium*, or thirdly, by finding a gene that codes for a protein that could be used as a vaccine target and adding it to a vaccine to train the immune system to identify and combat malaria cells or infected human cells.

**Current Research Overview**

We will present the state of the current research can in three parts: (1) the overall state of malaria genomics research, (2) methods used gather the genomics data and (3) methods to analyze the results of this data.

Around 20 years ago, the first paper was published that included reference genomes for the most deadly malaria parasite and mosquito vector[5]. Since then, the human malaria reference genomes have been published for all malaria parasites and half of the mosquito vectors[5]. Historically, these samples were processed in bulk, but more recent advances in Single-Cell RNA sequencing (scRNA-seq) have allowed it to be used to analyze the parasites at the level of the individual cell.

**Single-Cell RNA Sequencing**

Single-Cell RNA Sequencing (scRNA-seq) is a process where individual cells are isolated, and their transcriptomes are independently sequenced[6]. This allows researchers to assess the individual differences in genetics from a sample of cells.

scRNA-seq is not without its drawbacks, however. When attempting to perform a cluster analysis of genetic data, there are a large number of dimensions and thus it suffers the curse of dimensionality. Therefore, the data must be reduced to only a handful of dimensions. Additionally, the data produced is sparse, meaning that most of the values are zeroes. It is also more challenging to quantify the uncertainty in the data. Finally, it raises the question of whether single-cell is the correct level of analysis, or if it better to analyze this data at the level of the bulk tissue, or based on cell types[7].

It is important to understand the basic steps involved in scRNA-seq to be able to understand and analyze the resulting data. These steps include preparing the sample, generating droplets with unique barcodes and creating the libraries with the sequencing data[6]. The resulting data is split up into by library and the cells in the datasets are identified by their unique 16 letter barcodes.

**Pseudotime, Differential Expression and PseudotimeDE**

Pseudotime is defined as “a time-like variable indicating the relative position a cell takes in a lineage,” and Differential expression (DE) is when a gene’s mean expression changes between different experimental conditions. In this study, particularly, we are interested in finding those genes whose mean expression pattern changes along the pseudotime[8] (Figure 2). An

important implication of this is if there are multiple lineages, every gene will have one DE p-value per lineage it is part of.

There are multiple existing methods to calculate pseudotime and differential expression. Each method has its own strengths and limitations. Three primary limitations of the current pseudotime methods are as follows: Firstly, custom pseudotime inference data cannot be input into DE methods; secondly, uncertainty in pseudotimes cannot be inferred, specifically, it is not possible to infer if there a high certainty that the predicted lineage is the true one; and, finally, these methods may have suboptimal power since it is difficult to detect DEG in sparse data.

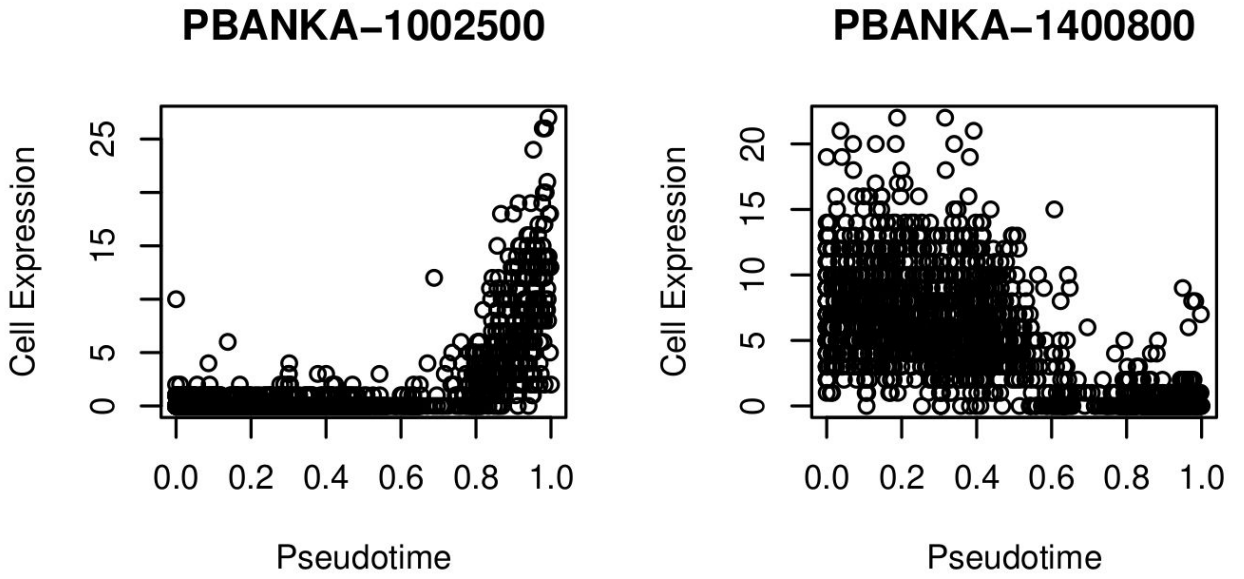



Figure 2: Two examples of differentially expressed genes from the preliminary analysis. The gene on the left, PBANKA_1002500, has a low expression from a (scaled) pseudotime of zero to around 0.7, where it begins increasing until it reaches a maximum at 1. The gene on the right, PBANKA_1400800, begins with a high gene expression, reaches a maximum at around 0.2, reaches a minimum at around 0.7, before increasing slightly.

A recently published statistical method, PseudotimeDE, could potentially address the issues discussed earlier. PseudotimeDE works in four overall steps. In step 1, it creates 80% subsamples of the datasets and permutes the cells. Then, in step 2, it infers pseudotime for the original data and subsamples using the chosen method. In step 3, it fits a zero-inflated negative binomial GAM (ZINB-GAM) to each gene in the original data. Finally, in step 4, it fits the same model from the original data to each subsample by permuting the cell labels "to obtain approximate null values of the gene's test statistic." At the end, either a gamma distribution or a gamma mixture model is fitted and used to compute analytical p-values[8].

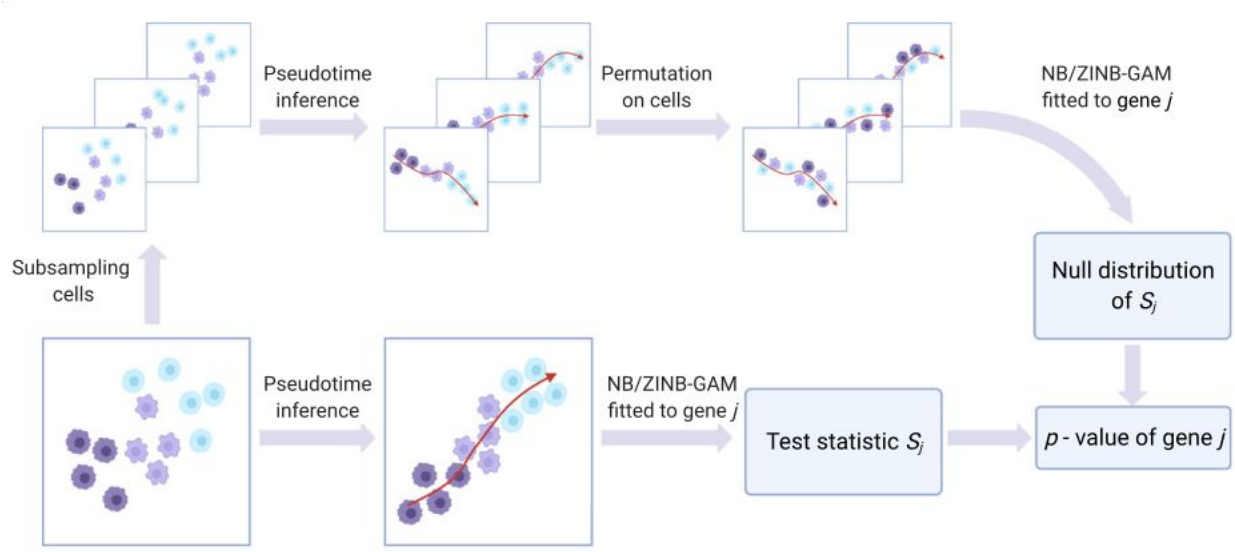



Figure 3: The PseudotimeDE workflow. Image credit: Ruberto, A. A. et al.[8]

From these steps, the advantages of PseudotimeDE can be seen: it allows for any pseudotime method as input (Slingshot, Monocle3-PI, etc.), subsampling allows for uncertainty in inferred psuedotimes to be taken into account, it has higher power to detect DEG, and its ZINB-GAM option to "excludes excess zeroes" and "treat [the zeroes] as non-biological zeroes" provides a better model option for the zero-inflated scRNA-seq data[8].

The primary drawback of PseudotimeDE is that it is computationally intensive: the lineages must be separately calculated for each subsample of the original dataset, and the model that is fitted to each gene in each lineage must also be fitted to that same gene in each subsample, potentially increasing computational requirements by hundreds of times compared to other methods.

**ZINB-GAM Model**

The negative binomial-generalized additive model (NB-GAM) is used to model the relationship between the pseudotime of a cell and the expression of each gene in that cell. The ZINB-GAM builds on this by adding a hidden variable to turn the NB-GAM into a mixture model between the hidden variable and a negative binomial distribution. The full model is as follows[8]:

$$\begin{cases} Z_{ij} \sim Ber\left(p_{ij}\right), \\ Y_{ij} | Z_{ij} \sim Z_{ij} \cdot NB\left(\mu_{ij}, \phi_j\right) + \left(1 - Z_{ij}\right) \cdot 0, \\ \log\left(\mu_{ij}\right) = \beta_{j0} + f_j\left(T_i\right), \\ logit\left(p_{ij}\right) = \alpha_{j0} + \alpha_{j1} \log\left(\mu_{ij}\right) \end{cases}$$

where $Y_{ij}$ is "the read count of gene *j* in cell *i*"[8]

$T_i$ is the normalized pseudotime for cell *i* inferred from the pseudotime method[8]

$Z_{ij}$ is the probability of "a dropout event of gene *j* on cell *i*,"[8] or zero-inflation

$NB(\mu_{ij}, \varphi_j)$ is "the negative binomial distribution with mean $\mu_{ij}$ and dispersion $\varphi_j$"[8]

$Ber(p_{ij})$ is a Bernoulli distribution with probability $p_{ij}$

$f_i(T_i)$ is a cubic spline function with six knots by default[8]

**Gene Ontology Enrichment**

An ontology is “a set of well-defined terms with well-defined relationships”[9], and the goal of the Gene Ontology (GO) Initiative is “to produce a structured, precisely defined, common, controlled vocabulary for describing the roles of genes and gene products in any organism”[9]. The three categories of GO are biological processes, the biological objective of the gene or product of the gene; molecular functions, the biochemical activity of the product of the gene; and cellular components, the resulting location where the product of the gene is active. These categories may be independent of one another, and each gene product may have multiple processes, functions and components[9].

## METHODS

The datasets used in this analysis all came from the malaria reference paper by Ruberto, A. A. et al. published in 2021[10]. In this paper, the authors took two samples of mosquitoes, with 50-80 mosquitoes per sample, allowed them to bite an rodent that was already infected with malaria and dissected the mosquitoes after 21 days. Then, they centrifuged the mosquito parts to separate out the sporozoites from the mosquito, sequenced the sporozoites using scRNA-seq and generated three libraries: two with a 50:50 ratio of sporozoites from the SG and MID and one 90:10[10].

The reference paper generated a total of three data files from the three libraries. The structure of the data files are as follows: each row of the data is a gene label, each column is a unique cell barcode, and the data are the number of times the gene was counted as being expressed for that cell, so called read counts. Low-quality genes and cells were filtered out by the authors of the paper[10].

Both the preliminary analysis and the final comprehensive analysis on the full dataset used similar methods, albeit with some differences. All datasets and scripts referenced were from the provided files in the GitHub linked under the Data Availability section of the malaria reference paper[10]. After filtering for this low-quality data, the final datasets had 922 genes and 2004 cells, 829 genes and 2024 cells and 1535 genes and 4326 cells, respectively, for datasets Pb1, Pb2, and Pb3 (as labeled in the output files). The .rds output files that were used were from the DOI linked under the Outputs section of the same GitHub page.

In this paper, we used Slingshot to infer the pseudotimes of the cells and then used PseudotimeDE for the DEG identification. The decision to use PseudotimeDE was threefold:

Firstly, it infers the uncertainty in the estimated pseudotime by integrating a subsampling step. Secondly, the availability of the ZINB-GAM as an option for the model allows a more suitable fit for the scRNA-seq data under consideration. Lastly, PseudotimeDE facilitates the computation of parametric p-values, which will enhance the robustness of our statistical analysis. We conducted the same analysis for each inferred lineage and performed GO enrichment analysis for the results. We expect more reliable results form our proposed analysis pipeline.

All programming was performed using R. For the preliminary analysis, the beginning .rds file used was *1_PB1_sce.rds*, which was generated from the *Pd_Spz_10X.Rmd* file. Then, adapting the .Rmd files that followed to apply to only Pb1, the necessary variables were created and lineages, principal curves and pseudotimes were inferred by using Slingshot.

This is where the R code began to significantly diverge from the reference code, as the reference paper used tradeSeq for DE analysis[10], while one goal of this paper was to use PseudotimeDE. For each lineage identified (three in the preliminary analysis and four on the full dataset), the data was cleansed and subsamples were generated. The Slingshot data and subsamples were entered into the *runPseudotimeDE* command in batches and the analysis was performed on those results.

The preliminary analysis was performed only on the first of the three datasets; the runPseudotimeDE module was run with a total of 80 subsamples that each contained 80% of the original dataset, and each subsample was generated independently of one another. The DE p-value cutoff had a Bonferroni corrected value of was 0.05/m, where m=992 was the total number of genes tested for significance.

To achieve results that could be compared to the malaria reference paper, *5_Pb_integrated_10X_only.rds* was used as the beginning data file for the comprehensive analysis. This file was output from *Pb_Spz_10X_5.Rmd,* and, from that point forward, the method leading up to the analysis was identical to those used in the preliminary analysis., except that 200 subsamples were generated for each lineage instead of 80, and the DE p-value cutoff was 0.05/m, where m = 1648 was the number of genes contained in the *5_Pb_integrated_10X_only.rds* file.

GO enrichment was performed on the top 50 DEG from each lineage in the comprehensive analysis with the organism parameter *Plasmodium berghei* ANKA and an FDR<0.05. Although the summarized GO enrichment results in Tables 1, 2 and 3 only included names with greater than one result, the analysis included names with exactly one result as well. GO results mentioned in the preliminary analysis were obtained by searching the DEG in an online GO database[15,17].

RESULTS AND DISCUSSION

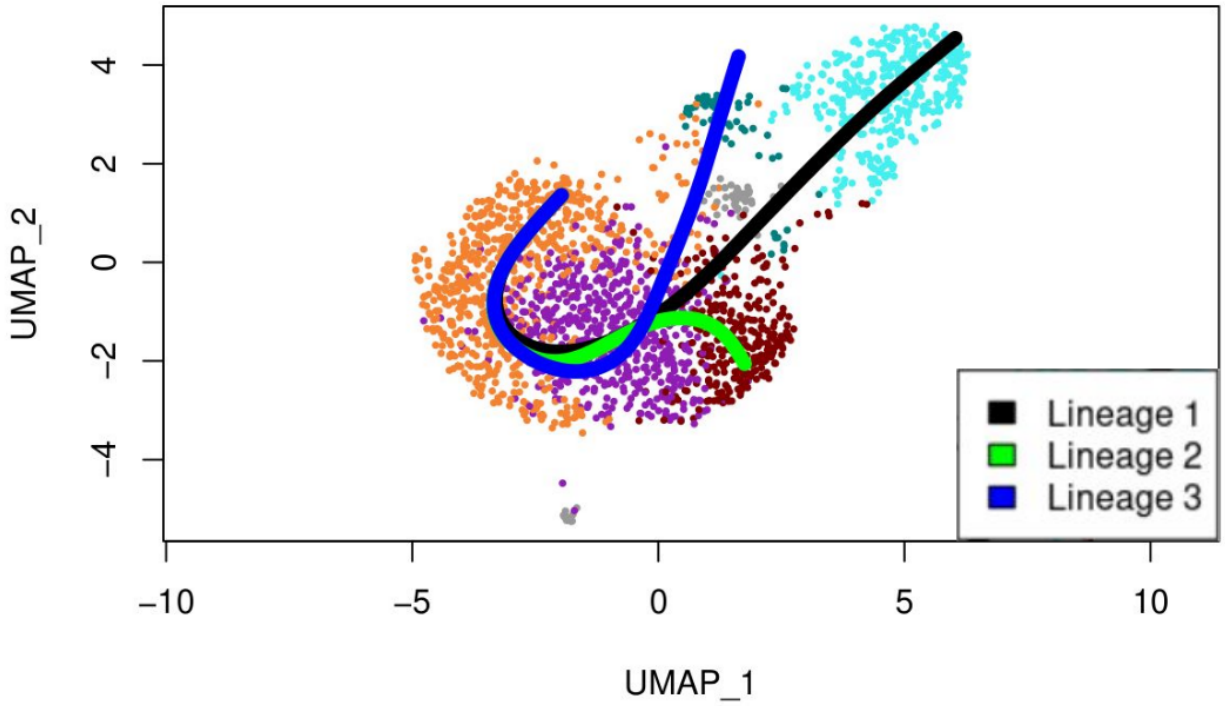
**Preliminary Analysis**

The preliminary analysis consisted of a visual trajectory analysis followed by the PseudotimeDE analysis and DE analysis. The DE analysis focused on the top 10 most significant genes from each lineage, as measured by the smallest parametric p-value. In other words, the focus was on those genes with the lowest probability of a false positive of a gene "changing in the mean expression over pseudotime."

Beginning with the UMAP plot (Figure 4a), all three lineages began on the left side before curving down and back up and to the right. It is at this point that they diverge noticeably. Lineages 1 and 3 continue to move up and to the right, while Lineage 2 moves down and to the right. If these lineage correspond to the end locations of the sporozoites (SG vs MID), then it visually appears that Lineages 1 and 3 end in the SG and Lineage 2 ends in the MID, since Lineages 1 and 3 take the longest path the furthest from the main cluster of cells, although this was not confirmed.

The most notable and potentially relevant result in the preliminary analysis was the detection of the gene PBANKA_0403200. This gene was significant in all three lineages, with a p-value of p = 3.0e-42, p = 3.0e-22, and p = 1.5e-32, for Lineage 1, 2 and 3, respectively. It was detected in the top 10 most DEG in Lineages 2 and 3, but not in Lineage 1. The gene PBANKA_0403200 has an annotation of "circumsporozoite (CS) protein"[11].

a)




b)

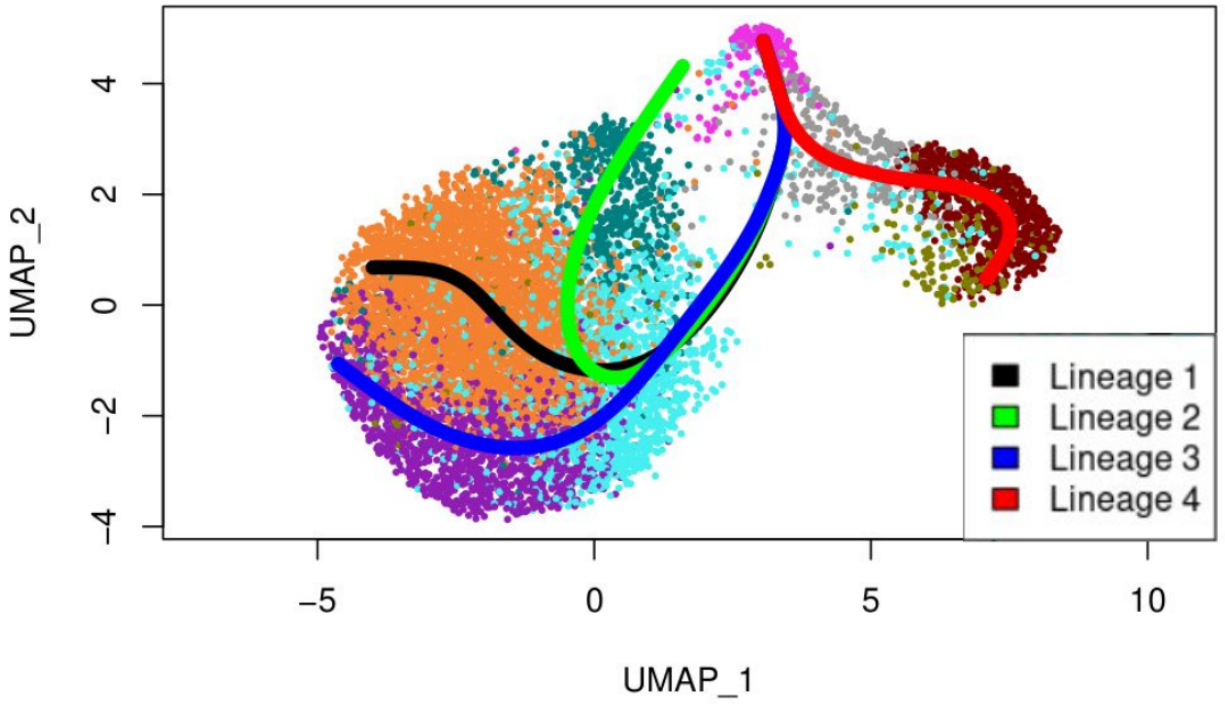



Figure 4: The UMAP plots from the a) preliminary and b) comprehensive analysis. The preliminary analysis focused on the first dataset only. The colored points correspond to different clusters determined by an unsupervised clustering approach.

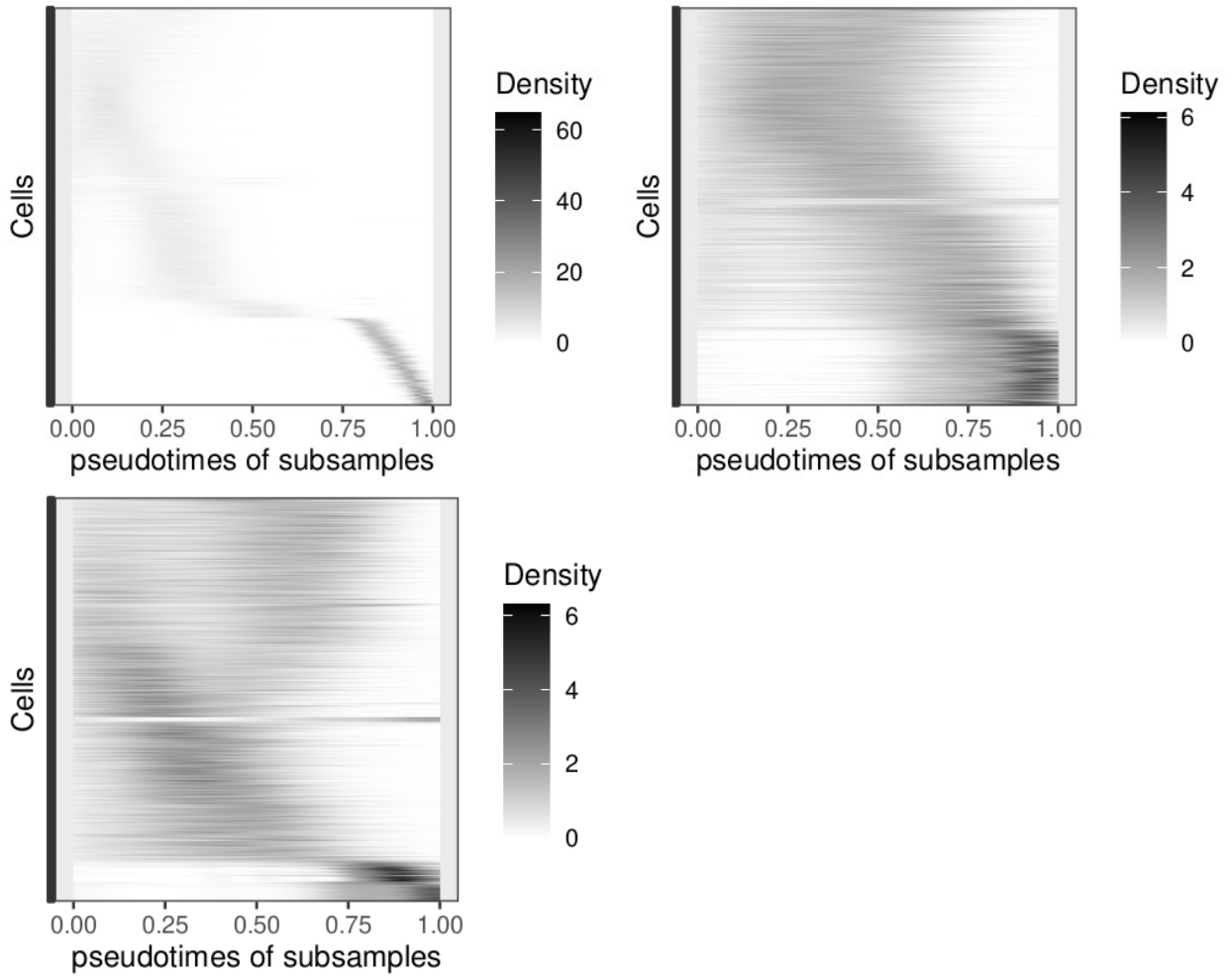
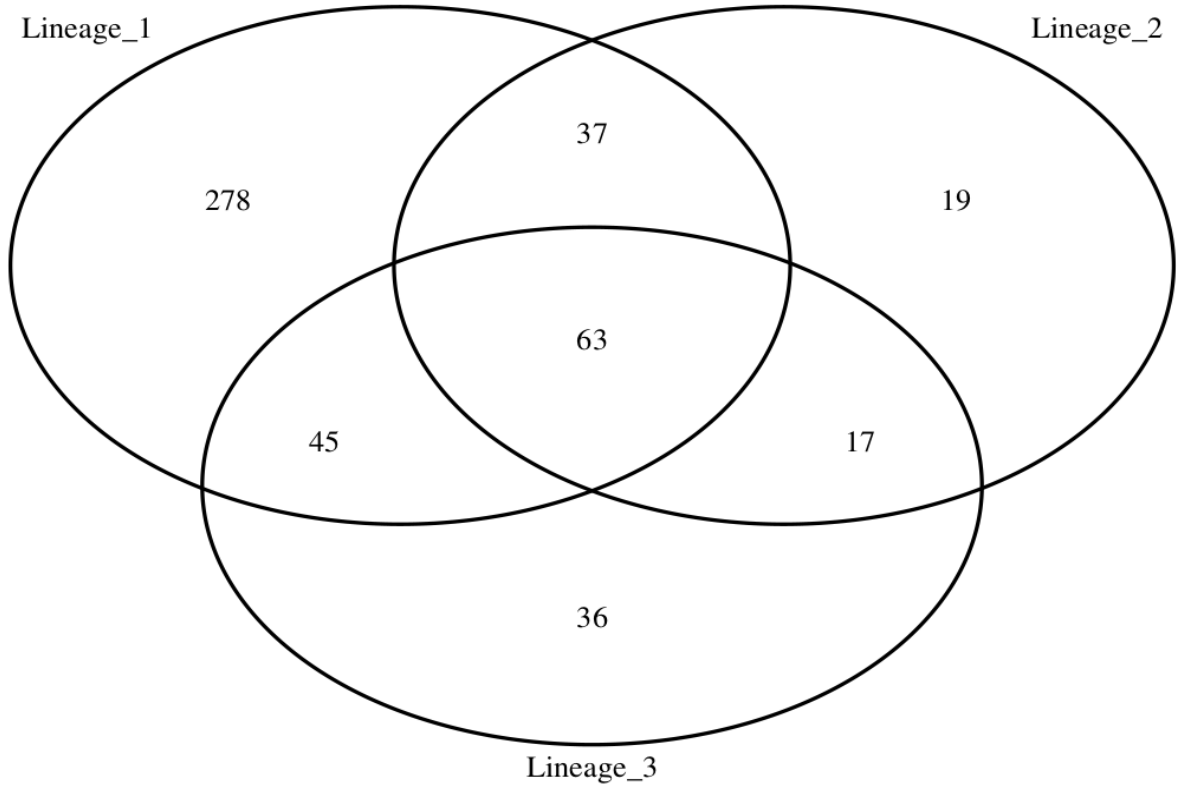



Figure 5: Left to right, top to bottom, uncertainty density plots for Lineages 1, 2 and 3 respectively, from the preliminary analysis. The legends each have their own scale. Lineage 1 had the highest density of any of the plots from a scaled pseudotime of around 0.75 to 1.

a)




b)

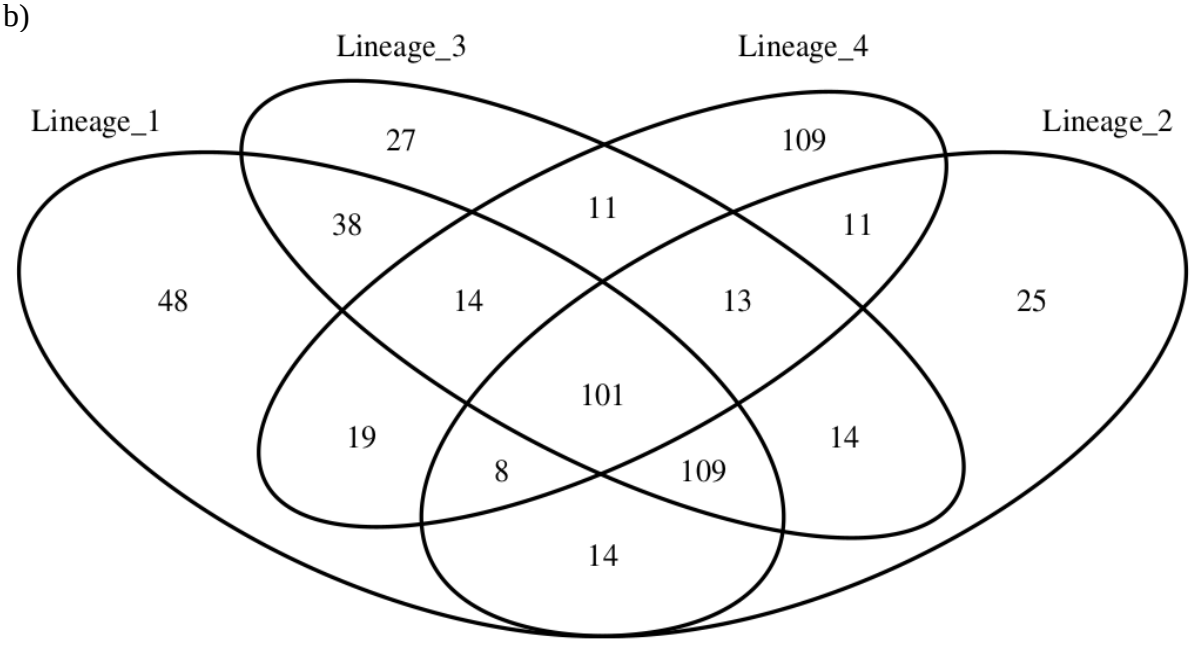



Figure 6: Venn diagrams displaying the total number of DEG detected in the a) preliminary and b) comprehensive analysis. In the preliminary analysis, Lineage 1 had the highest number of unique DEG at 278, while in the final analysis, Lineage 4 had the highest number at 109.

The CS protein is extremely common and is found in many species, including every malaria species. The protein is located on the surface of cell, and it is evenly distributed across the surface[12]. It is used in vaccine research[13]; the RTS,S vaccine, which has already completed Phase III trials, uses a portion of the CS protein from the *P. falciparum* species[14]. R21, a vaccine in development at the University of Oxford, builds on this vaccine and ultimately uses an even larger proportion of the CS proteins by removing free Hepatitis B Antigen that is found in the RTS,S vaccine[13].

The GO Terms related to adhesion to/entry into host cells, and located in the membrane, among others[15], agree with this previous research that used it for a vaccine target. The annotation and GO Terms may help provide an explanation as to why this protein is so important in malaria research. By training the mammal's immune system to detect and fight this protein, these parasites could be detected and killed before they ever adhere to or enter into the host cell, and by being located in the membrane of the parasite, it makes it easy for the mammal's immune system to detect it early on in the process of infection. Additional research could be performed in the future to investigate a drug that inhibits or disables this protein altogether, preventing the sporozoite cell from ever functioning to be able to infect its host.

Gene PBANKA_0501200, which is the only gene to be in the top 10 DEG for all three lineages in our dataset, is annotated as a "early transcribed membrane protein"[11], and our data confirms this, as the cell expression vs pseudotime graphs for each lineage each show a high expression early on in the inferred pseudotime and a dropoff as the pseudotime continues to increase. Its name is UIS4[16], which stands for "Up-regulated in infective sporozoites"[17]. Its GO term is "Membrane"[15]. Note that "Although UIS4 is highly expressed by sporozoites in the

mosquito salivary glands, genetic studies in the rodent malaria parasites have shown that UIS4 is required only after transmission to the mammalian host," and thus the translation of this gene is suppressed in the mosquito salivary glands[18].

In Lineage 3, all late expressed, low expression genes (see Figure 8a) in the top 10 had an annotation that included the words "ribosomal protein" (RP). The ribosome is a structure inside the cell that makes proteins, it does this by surrounding the mRNA molecule on the top and bottom. This allows tRNA to match the mRNA segment by segment and add an amino acid onto the peptide chain[19]. The Ribosomal Proteins, which make up the structure of the ribosome, are used not only to build the amino acid chains, but also for extra-ribosomal functions, such as "activation of... pathways in response to stress, resulting in cell cycle arrest and apoptosis"[20]. This may be useful in an anti-malaria drug by perhaps increasing the number of the RPs relevant for apoptosis, causing the malaria cells to commit it at higher rates.

All mentioned RP genes had GO terms for "structural constituent of ribosome." All genes except PBANKA_0407700 had GO terms for "translation." Additionally, gene PBANKA_1338600 had a GO term for "enables rRNA binding." PBANKA_0407700 also had GO terms for "involved in cytoplasmic translational elongation" and "part of cytosolic large ribosomal subunit"[17].

Gene PBANKA_0911650, a top 10 DEG in Lineage 3 and also significant in Lineage 1 but not 2 ($p = 1e\text{-}26$, $p = 1e\text{-}17$, and $p = 0.104$, respectively), had an interesting shape in the cell expression vs pseudotime graph where all of the high expression cells for this gene had a (scaled) inferred pseudotime of close to 0.8. All of this suggests that it plays an important specific role in that stage of the sporozoite's life. It has an annotation of "U2 spliceosomal RNA"[11]. U2

spliceosomal RNA is found in almost every eukaryotic organism and encodes for U2 small nuclear ribonucleaoprotein (snRNP)[21].

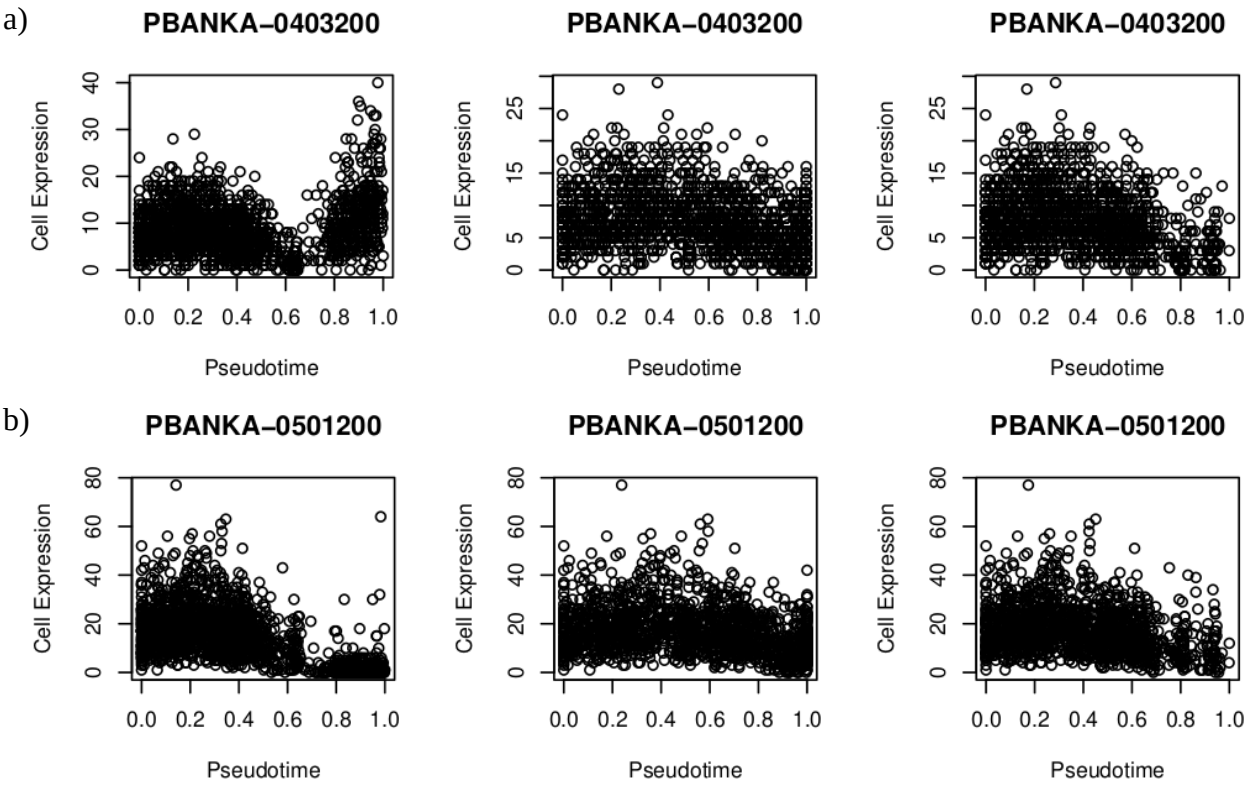
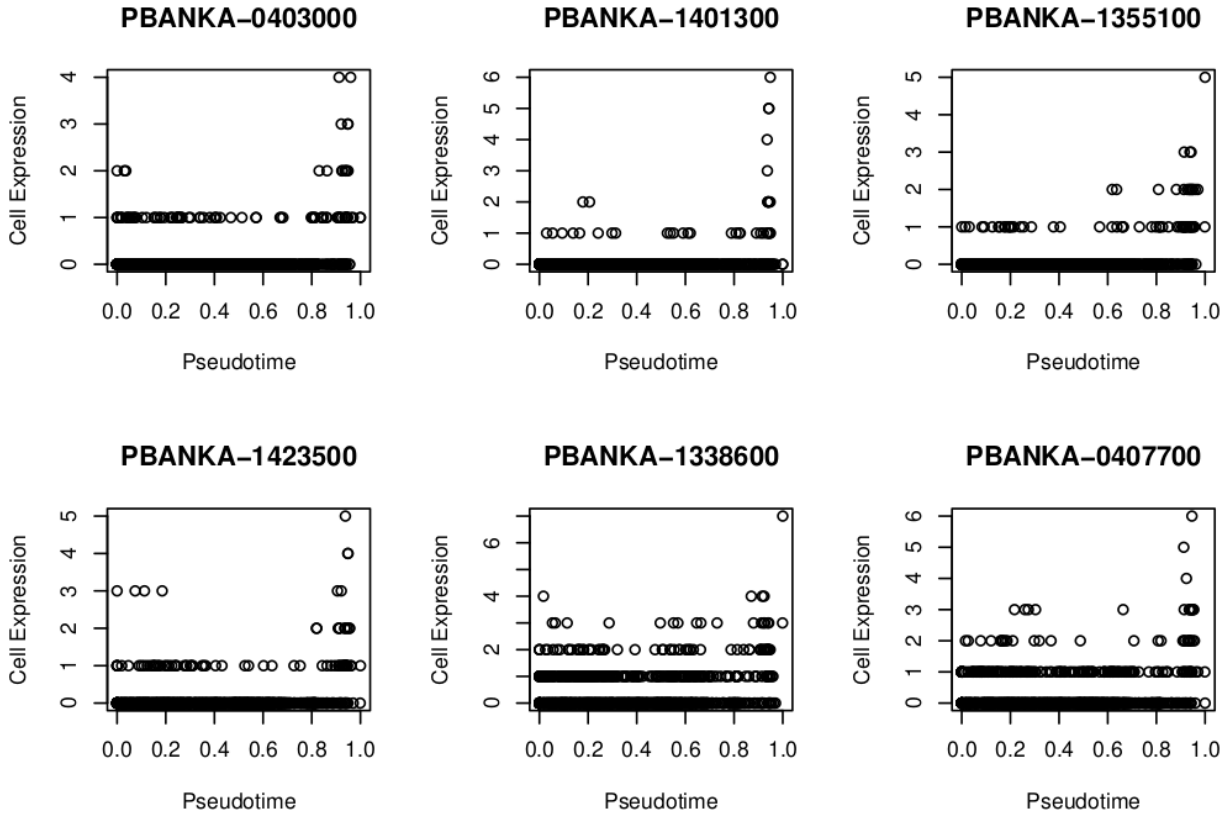



Figure 7: The first two genes in the preliminary analysis. a) PBANKA_0403200, has the same annotation of "Circumsporozoite Protein"[11] as equivalent genes in other malaria species that have been used in multiple vaccines, including the RTS,S vaccine, which has completed Phase III trials[14]. b) PBANKA_0501200, the only gene in the preliminary analysis to be in the top 10 most DEG of all three lineages, has an annotation of "early transcribed membrane protein"[11] and a name of "Up-reguated in infected sporozoites 4"[16].

a)




b)

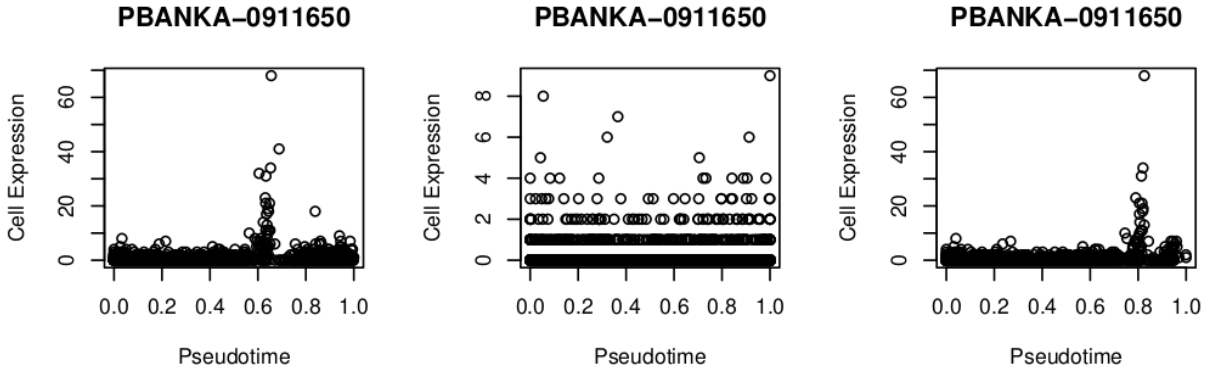



Figure 8: The remaining genes in the preliminary analysis. a) All genes in the top 10 of Lineage 3, in order of decreasing significance, that have a maximum cell expression of <10, where the maximum occurs after a scaled pseudotime of 0.6. The respective annotations are: 60S ribosomal protein L44, putative, 40S ribosomal protein S7, putative, 40S ribosomal protein S6, putative, 60S ribosomal protein L13-2, putative, 60S ribosomal protein L23, putative, 60S acidic ribosomal protein P2, putative[11] . b) PBANKA_0911650 has an annotation of "U2 spliceosomal RNA" and spikes in cell expression in the two lineages it is significant in (Lineage 1 at a scaled pseudotime of 0.6 and Lineage 3 at a scaled pseudotime of 0.8)[11.]

**Proposal and Comprehensive Analysis of the Full Dataset**

The proposal had three aims: Firstly, to use a combined dataset that includes all three datasets and increase the number of subsamples generated from 80 to 200, secondly, to compare the DEG results from the combined dataset to the preliminary analysis and the reference paper, and finally, to identify marker genes that show unique expression patterns at different developmental stages.

In addition to the uncertainty density plots (Figure 9), Venn Diagram (Figure 6b) and cell expression vs pseudotime plots (Figures 10 and 11 and Appendix C), the annotations of the top 20 genes in each lineage were recorded and placed into tables with the parametric p-value (Appendix A) and then placed into six categories: UIS genes, enzymes ending in -ase, ribosomal proteins, genes with unknown functions, other top 20 DEG from Lineages 1, 2 or 3 not significant in Lineage 4 and all other top 20 DEG (Appendix B).

**Preliminary vs Comprehensive Analysis**

On average, the uncertainty density plots for the final analysis (Figure 9) have a higher density, or lower uncertainty, than those generated from the preliminary analysis (Figure 5). This is likely due to the increased number of subsamples, which in turn decreased the uncertainty in the inferred pseudotimes.

Next, the three genes highlighted in the preliminary analysis, PBANKA_0403200, PBANKA_0501200 and PBANKA_0911650 (Figure 7a, Figure 7b and Figure 8b, respectively), were compared to the same plots for the comprehensive analysis (Figure 10).

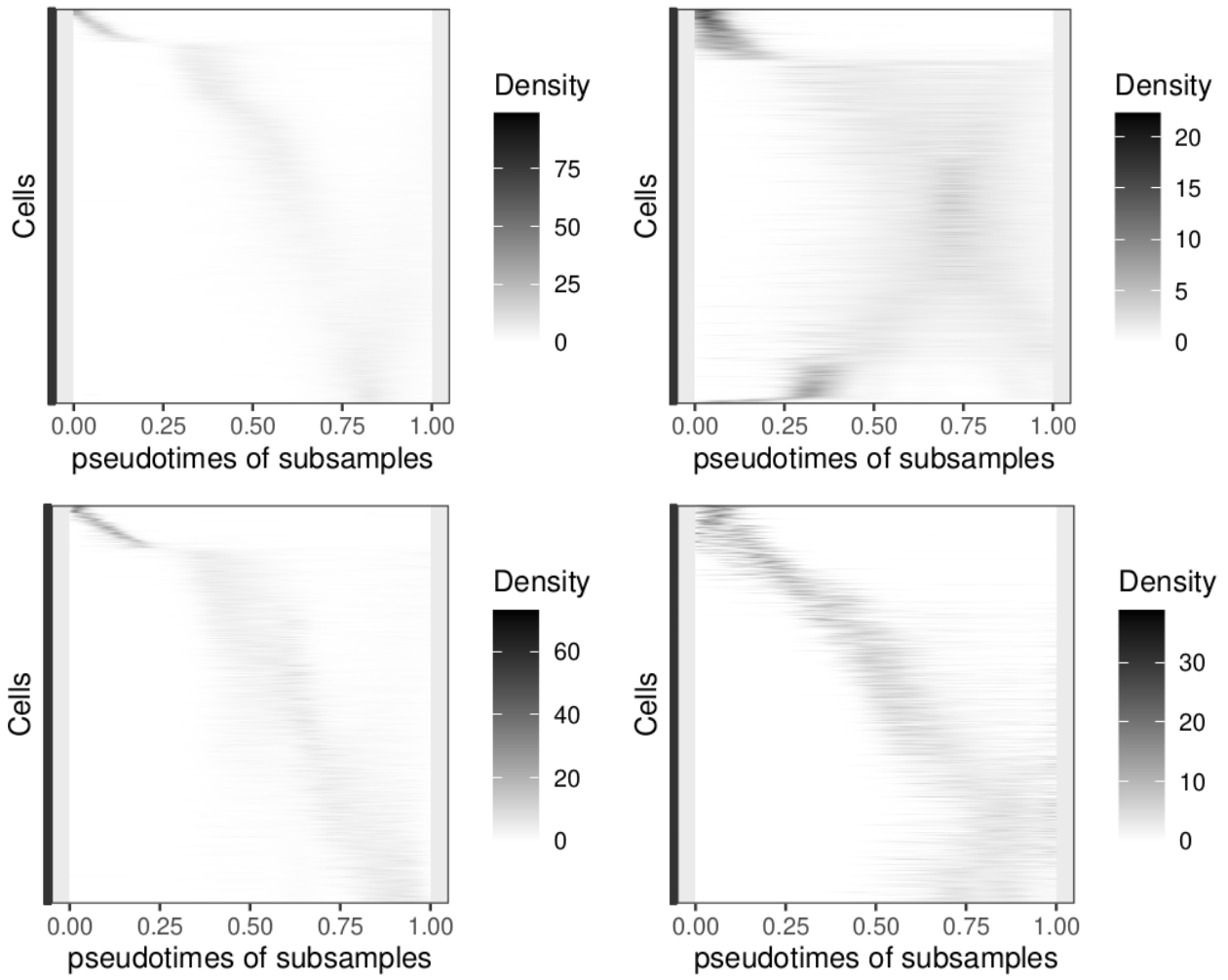



Figure 9: Left to right, top to bottom, uncertainty density plots for Lineages 1, 2, 3 and 4, respectively, as identified in the comprehensive analysis. Each legend has its own scale. Lineage 1 had the highest density of any of the plots from a (scaled) pseudotime of 0 to around 0.2. Lineage 2 had the lowest density, which agrees with the UMAP plot where it “wraps around” back where it started.

In the comprehensive analysis, gene PBANKA_0403200, the CS protein[11], had a similar cell expression vs pseudotime pattern for the first three lineages and a different one for the fourth. The first three lineages are most similar to the Lineage 1 in the preliminary analysis, where the first cluster increases until around 0.2 pseudotime, then decreases until it hits nearly zero at around 0.35 pseudotime (0.7 in the preliminary analysis) and finally increases until it reaches its maximum value.

Gene PBANKA_0501200, which was in the top 10 DEG in all three lineages of the preliminary analysis and has an annotation of early transcribed membrane protein[11] and Up-regulated in infective sporozoites[17], was in the top 20 DEG in the first three lineages of the comprehensive analysis but not the fourth. It also had higher maximum expression in the first three lineages when compared to the fourth.

The expression levels, plots and UIS annotation agree with the malaria paper predicting that the fourth lineage consists of sporozoites that stay in the MID, while the first three lineages end in the SGs. However, the shape of the plots do not necessarily agree with the annotation of "early transcribed membrane protein," since the expression levels do not reach a maximum until around 0.6-0.8 pseudotime for Lineages 1 through 3.

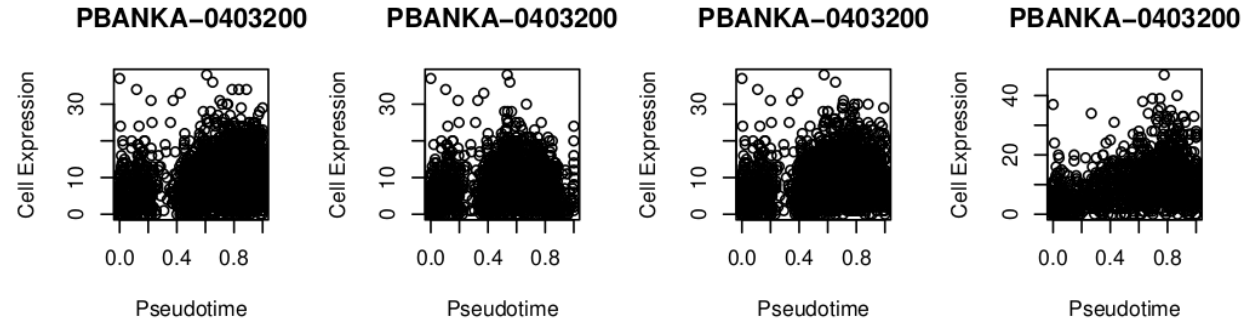
The spike of cell expression for Gene PBANKA_0911650 (U2 spliceosomal RNA)[11] in Lineages 1 and 3 of the preliminary analysis mostly disappeared in the comprehensive analysis, except for in Lineage 2, where there is spike at around 0.85 and Lineage 4, where there are a few (possibly outlier) cells with high expression at around 0.6.

Similarly to how the Ribosomal Protein (RP) DEG in Lineage 3 of the preliminary analysis (Figure 8a) all exclusively followed a similar pattern of low expression, late peak, the

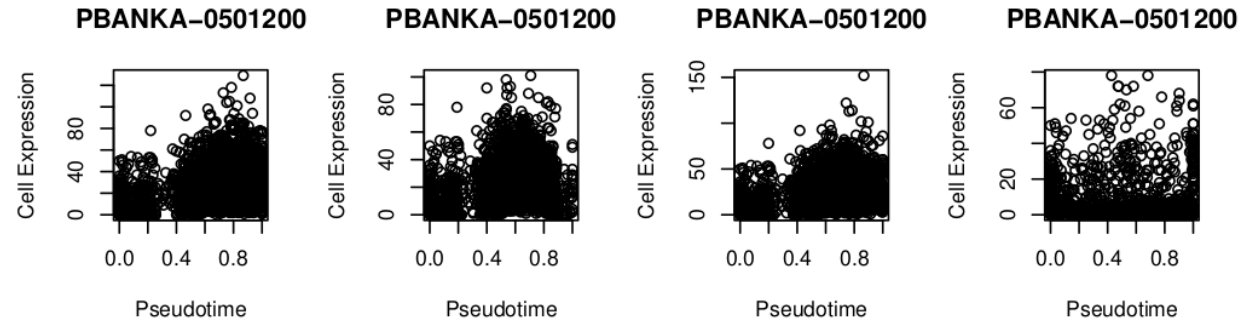
RP DEG in the comprehensive analysis (Appendix B, Table 8; plots in Appendix C), all had low expression, early peak. Specifically, they all had a maximum cell expression of less than 12, where the maximum occurred near a pseudotime of zero, then decreased until around 0.35, where it either leveled off or increased slightly from there.

Additionally, no non-RP gene except PBANKA_1340100 (L-lactate dehydrogenase)[11] had this pattern. Lineage 4 had the fewest RP DEG in the top 20 (8 total compared to 11, 10 and 15 for Lineages 1, 2 and 3, respectively). One way to filter genes with unknown function in the future to search for unannotated RP genes may be to find genes with this early peak, low expression pattern. However, since there were no genes with unknown function (Figure 11) with this pattern, it does not apply here.

a)




b)




c)

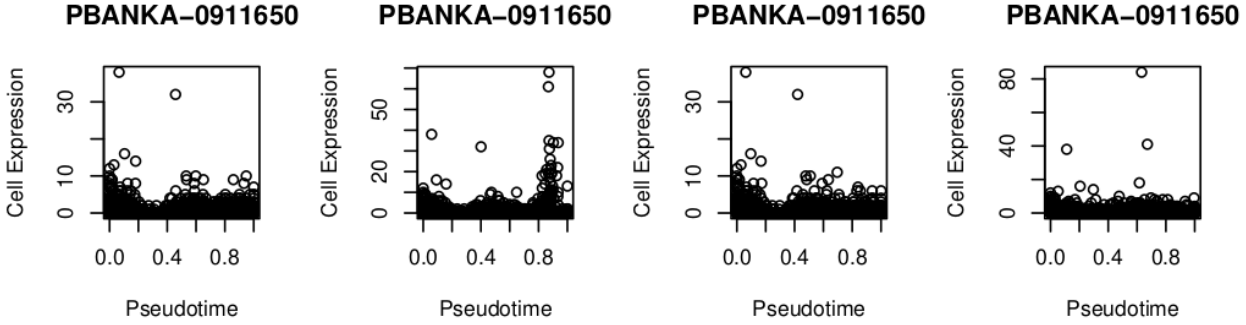



Figure 10: Comprehensive analysis cell expression vs pseudotime plots for the three genes highlighted in the preliminary analysis a) PBANKA_0403200, b) PBANKA_0501200 and c) PBANKA_0911650. PBANKA_0501200 was in the top 20 DEG in Lineages 1, 2 and 3 but not 4, which agrees with its annotations of "up-regulated in infective sporozoites" if the malaria reference paper is correct that Lineages 1 through 3 consists of sporozoites that end in the SG, while Lineage 4 ends in the MID. PBANKA_0911650 only contains the late spike (found in the preliminary analysis) for Lineage 2, which is likely in a cluster of cells near where all of the Lineages began. This is consistent with the fact that the cell expression begins higher on each of the lineages then decreases somewhat.

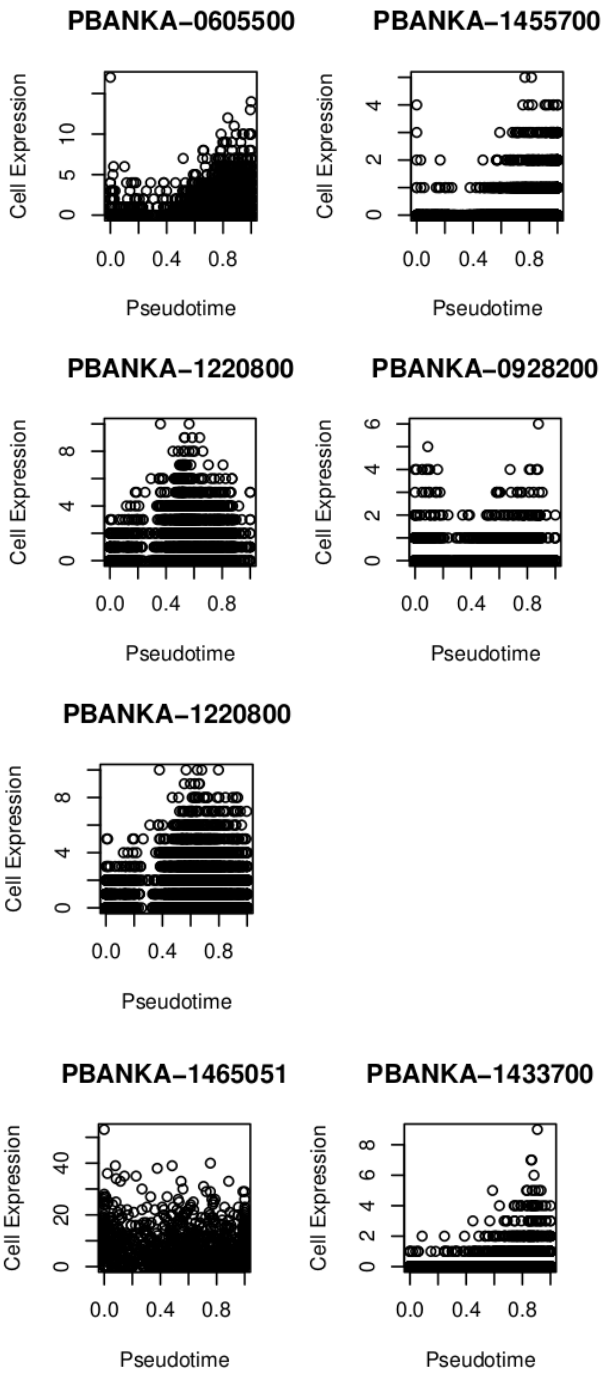



Figure 11: Top to bottom, cell expression vs pseudotime plots for top 20 DEG with unknown functions for Lineage 1, 2, 3 and 4, respectively. None of these plots have the same early peak, low expression as the annotated RP genes.

**Malaria Paper vs Comprehensive Analysis**

Comparing the number of DEG detected between here and the malaria reference paper, the paper used tradeSeq to detect DEG, and found 661 (FDR<0.05)[10] while PseudotimeDE with 200 subsamples detected 561 DEG. We identified exactly 100 fewer genes.

There are three possible reasons that PseudotimeDE identified fewer DEG than tradeSeq. One is that tradeSeq generates more false positives than PseudotimeDE, since PsudotimeDE returns better calibrated p-values than tradeSeq, and therefore has a better FDR control.[8] Another explanation could be that tradeSeq has a higher power on bifurcated datasets when measured using "an actual 5% false discovery proportion (FDP, defined as the proportion of false discoveries among the discoveries in one synthetic dataset) instead of the nominal 5% FDR"[8] and, since the dataset used here is bifurcated, tradeSeq is better at detecting true positives. Finally, the <5% nominal FDR could have resulted in a different cutoff than a Bonferroni cutoff of 0.05/m, making the results not directly comparable.

The malaria reference paper focused its gene analysis at three points: before the trajectory analysis using the UMAP plots as well as Seurat's FindAllMarkers, and after the trajectory analysis using GO. While the focus in our study was on the genes detected during and immediately after the trajectory analysis, the genes mentioned in the paper can still be compared to those detected here to build a bigger picture.

The authors of the malaria paper mention that they "detected various 'up-regulated in infective sporozoite' (UIS) genes, of which PBANKA_1328000 (Serine/threonine protein phosphatase; UIS2), PBANKA_1400800 (UIS3), PBANKA_0501200 (Early transcribed membrane protein; UIS4), and PBANKA_1128100 (Phospholipase, UIS10) were among the

most highly expressed. They also found genes with known expression in MID sporozoites, such as PBANKA_0901300 (Membrane-associated erythrocyte binding-like protein) and PBANKA_1306500 (TRAP-like protein; UOS3).”[10]

Our analysis confirms these results: UIS2, 3, 4 and 10 were found to be the among the most significant DEG in the first three lineages. Additionally, genes PBANKA_0901300 and PBANKA_1306500 were detected as significant only in Lineage 4, and not in any other lineage.

The malaria paper mentions gene PBANKA_0719200, related to puf2 transcripts, although this gene was not found in the top 20 DEG of any lineage here. It also mentions genes with unknown function, which “indicate that other markers linked to sporozoite biology may exist.”[10] The cell expression vs pseudotime plots for genes with unknown function are in Figure 11. These genes all have low expression, except for PBANKA_1465051, which does not have an obvious expression pattern.

**Gene Ontology Enrichment Analysis**

The summarized GO enrichment results are given in Tables 1, 2 and 3. Possible vaccine target genes could include such biological process terms as ATP generation, since stopping this would prevent the parasite from generating usable energy, entry into host, as stopping this would prevent infection form progressing, as well as cellular component terms of extracellular, as these proteins would be more detectable by and vulnerable to the host immune system.

The genes associated with ATP generation are PBANKA_1214300 (enolase, putative[11]) and PBANKA_1340100 (L-lactate dehydrogenase[11]).

Enolase converts 2-phospho-D-glycerate to phosphoenolpyruvate. It is a catalytic enzyme and has been the target of new drug development for *H. pylori*[23]. Lactate dehydrogenase converts lactate to pyruvate. It is an catalytic enzyme and this conversion is reversible[24].

The genes associated with entry into host are PBANKA_0403200 (circumsporozoite (CS) protein[11]), PBANKA_0407700 (60S acidic ribosomal protein P2, putative[11]), PBANKA_1349800 (thrombospondin-related anonymous protein[11]) and PBANKA_1222500 (plasmepsin X (PMX), putative[11]).

Thrombospondin-related anonymous protein is used by sporozoites to glide and enter into the host cell. It has previously been the target of in vivo antibody research which found that these antibodies do not stop infection[25].

Although its structure and full function are the topics of recent research, PMX has been shown to be essential for the formation and activation of multiple proteins, including those required for host cell invasion and rupture in human host cells[26]. One of these proteins is PfRh5, a candidate as a vaccine target[26] undergoing multiple Phase I and II clinical trials[27], so disrupting PMX could also disrupt these proteins and thus the invasion of the parasite. The GO terms of interaction with host and entry into host support that this gene is necessary for, among other things, host cell invasion.

The genes associated with extracellular are PBANKA_0501200 (early transcribed membrane protein[11], Up-regulated in infective sporozoites[17] (UIS4)[9]), PBANKA_0910900 (sporozoite surface protein essential for liver stage development[11]), PBANKA_1003000 (liver specific protein 2[11]), PBANKA_1128100 (phospholipase[11], UIS10[10]), PBANKA_1328000 (serine/threonine protein phosphatase UIS2[11]) and PBANKA_1400800 (protein UIS3[22]),

The only GO result related to the DEG with an unknown function annotation is for gene PBANKA_1465051, which has GO terms of host cell and host cellular component for Lineage 1 cellular component and Lineage 2 cellular component, respectively.

Table 1: GO biological process summary for annotations with more than one total result

| | **Result Count and Percent of Background for Lineage...** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **GO Term (BP)** | **1** | | **2** | | **3** | | **4** | |
| translation | 23 | 15.60% | 26 | 17.70% | 26 | 17.70% | 21 | 14.30% |
| biological process involved in interaction with host | 5 | 6.30% | 4 | 5.00% | 6 | 7.50% | 4 | 5.00% |
| biological process involved in symbiotic interaction | 5 | 6.00% | 4 | 4.80% | 6 | 7.20% | 4 | 4.80% |
| biological process involved in interspecies interaction between organisms | 5 | 6.00% | 4 | 4.80% | 6 | 7.10% | 4 | 4.80% |
| movement in host environment | 4 | 5.70% | 3 | 4.30% | 5 | 7.10% | | |
| glycolytic process | 2 | 14.30% | | | | | | |
| ADP metabolic process | 2 | 14.30% | | | | | | |
| purine nucleoside diphosphate metabolic process | 2 | 14.30% | | | | | | |
| purine ribonucleoside diphosphate metabolic process | 2 | 14.30% | | | | | | |
| ATP generation from ADP | 2 | 14.30% | | | | | | |
| ribonucleoside diphosphate metabolic process | 2 | 14.30% | | | | | | |
| carbohydrate catabolic process | 2 | 13.30% | | | | | | |
| nucleoside diphosphate phosphorylation | 2 | 11.80% | | | | | | |
| nucleotide phosphorylation | 2 | 11.80% | | | | | | |
| nucleoside diphosphate metabolic process | 2 | 11.10% | | | | | | |
| pyruvate metabolic process | 2 | 11.10% | | | | | | |
| exit from host cell | 2 | 10.00% | | | 2 | 10.00% | | |
| exit from host | 2 | 9.50% | | | 2 | 9.50% | | |
| entry into host | | | | | 3 | 5.90% | 3 | 5.90% |
| biological process | | | | | 6 | 3.00% | | |

Table 2: GO molecular function summary for annotations with more than one total result

| | **Result Count and Percent of Background for Lineage...** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **GO Term (MF)** | **1** | | **2** | | **3** | | **4** | |
| structural constituent of ribosome | 24 | 16.80% | 28 | 19.60% | 27 | 18.90% | 22 | 15.40% |
| structural molecule activity | 24 | 14.80% | 28 | 17.30% | 27 | 16.70% | 22 | 13.60% |
| molecular function | 25 | 4.70% | 29 | 5.50% | 30 | 5.60% | 22 | 4.10% |

Table 3: GO cellular component summary for annotations with more than one total result

| | **Result Count and Percent of Background for Lineage...** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **GO Term (CC)** | **1** | | **2** | | **3** | | **4** | |
| ribosome | 24 | 14.40% | 28 | 16.80% | 28 | 16.80% | 24 | 14.40% |
| non-membrane-bounded organelle | 27 | 7.40% | 30 | 8.20% | 30 | 8.20% | 26 | 7.10% |
| intracellular non-membrane-bounded organelle | 26 | 7.20% | 29 | 8.00% | 29 | 8.00% | 25 | 6.90% |
| ribosomal subunit | 6 | 14.60% | 9 | 22.00% | 9 | 22.00% | 7 | 17.10% |
| symbiont-containing vacuole membrane | 5 | 17.90% | 3 | 10.70% | 3 | 10.70% | | |
| host cell | 9 | 6.30% | 7 | 4.90% | 6 | 4.20% | | |
| host cellular component | 9 | 6.30% | 7 | 4.90% | 6 | 4.20% | | |
| host cell part | 8 | 6.70% | 6 | 5.00% | 6 | 5.00% | | |
| cytosolic ribosome | 4 | 20.00% | 7 | 35.00% | 7 | 35.00% | 6 | 30.00% |
| large ribosomal subunit | 4 | 20.00% | 6 | 30.00% | 5 | 25.00% | 3 | 15.00% |
| cell surface | 6 | 9.00% | 3 | 4.50% | 4 | 6.00% | | |
| extracellular membrane-bounded organelle | 6 | 9.00% | 4 | 6.00% | 4 | 6.00% | | |
| symbiont-containing vacuole | 6 | 9.00% | 4 | 6.00% | 4 | 6.00% | | |
| host cell nucleus | 2 | 100.00 % | 2 | 100.00 % | 2 | 100.00 % | 1 | 50.00% |
| extracellular organelle | 6 | 8.80% | 4 | 5.90% | 4 | 5.90% | | |
| extracellular region | 6 | 8.70% | 4 | 5.80% | 4 | 5.80% | | |
| host cell cytoplasm | 7 | 6.60% | 5 | 4.70% | 5 | 4.70% | | |
| host intracellular part | 7 | 6.60% | 5 | 4.70% | 5 | 4.70% | | |
| host intracellular region | 7 | 6.50% | 5 | 4.70% | 5 | 4.70% | | |
| host cell cytoplasm part | 6 | 6.90% | 4 | 4.60% | 4 | 4.60% | | |
| host intracellular membrane-bounded organelle | 2 | 50.00% | 2 | 50.00% | 2 | 50.00% | 1 | 25.00% |
| host intracellular organelle | 2 | 40.00% | 2 | 40.00% | 2 | 40.00% | | |
| cytosolic large ribosomal subunit | 2 | 22.20% | 4 | 44.40% | 3 | 33.30% | 2 | 22.20% |
| microneme | 3 | 9.40% | 3 | 9.40% | 5 | 15.60% | | |
| cytosolic small ribosomal subunit | 2 | 18.20% | 3 | 27.30% | 4 | 36.40% | 4 | 36.40% |
| ribonucleoprotein complex | 6 | 3.60% | 9 | 5.40% | 9 | 5.40% | 7 | 4.20% |
| small ribosomal subunit | 2 | 9.50% | 3 | 14.30% | 4 | 19.00% | 4 | 19.00% |
| cytosol | 5 | 3.30% | 7 | 4.60% | 7 | 4.60% | 6 | 3.90% |
| apical complex | | | | | 5 | 6.00% | | |
| apical part of cell | | | | | 5 | 4.20% | | |

## Discussion

There are many candidate genes for malaria vaccines and therapeutics being researched today. The three most promising genes identified in this paper are PBANKA_0403200 (CS protein[11]), PBANKA_0501200 (early transcribed membrane protein,[11] UIS4[9]) and PBANKA_1222500 (PMX, putative[11]). PBANKA_0403200 is included in a vaccine that has already completed Phase III trials[14] and PBANKA_1222500 is a necessary enzyme for the creation and activation of many genes, notable among them PfRh5. PfRh5 is part of a vaccine that, as of December 2023, is undergoing no fewer than four Phase I and II trials[27]. The disruption of PBANKA_1222500 could therefore disrupt PfRh5 and thus is a possible target for malaria treatment.

REFERENCES


1. Malaria's Impact Worldwide. *Centers for Disease Control and Prevention* https://www.cdc.gov/malaria/malaria_worldwide/impact.html (2021).

2. Fact sheet about malaria. *World Health Organization* https://www.who.int/news-room/fact-sheets/detail/malaria.

3. Plasmodium. *Encyclopædia Britannica* https://www.britannica.com/science/Plasmodium-protozoan-genus (2024).

4. Prudêncio, M., Rodriguez, A. & Mota, M. The silent path to thousands of merozoites: the Plasmodium liver stage. *Nat Rev Microbiol* 4, 849–856 (2006). https://doi.org/10.1038/nrmicro1529

5. Neafsey, D.E., Taylor, A.R. & MacInnis, B.L. Advances and opportunities in malaria population genomics. *Nat Rev Genet* 22, 502–517 (2021). https://doi.org/10.1038/s41576-021-00349-5

6. Clark, S. Single Cell RNA-seq: An introductory overview and tools for getting started. *10x Genomics* https://www.10xgenomics.com/blog/single-cell-rna-seq-an-introductory-overview-and-tools-for-getting-started.

7. Lähnemann, D., Köster, J., Szczurek, E. et al. Eleven grand challenges in single-cell data science. *Genome Biol* 21, 31 (2020). https://doi.org/10.1186/s13059-020-1926-6

8. Song, D. & Li, J. J. PseudotimeDE: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell RNA sequencing data. *Genome Biology* 22:124 (2021).

9. Ashburner, M., Ball, C., Blake, J. et al. Gene Ontology: tool for the unification of biology. Nat Genet 25, 25–29 (2000). https://doi.org/10.1038/75556

10. Ruberto, A. A. et al. Single-cell RNA sequencing reveals developmental heterogeneity among Plasmodium berghei sporozoites. *Nature* 11:4127 (2021).

11. *PlasmoDB* https://plasmodb.org/plasmo/app/record/gene/.

12. Davidson, E. Handbook of Glycomics. *ScienceDirect* https://www.sciencedirect.com/topics/medicine-and-dentistry/circumsporozoite-protein (2010).

13. Schuerman, L. & Ockengouse, C. Plotkin's Vaccines (Eighth Edition). *ScienceDirect* https://www.sciencedirect.com/topics/medicine-and-dentistry/circumsporozoite-protein (2023).

14. Veve, M. P. & Athans, V. Side Effects of Drugs Annual. ScienceDirect https://www.sciencedirect.com/topics/neuroscience/rts-s (2019).

15. GO annotations. *QuickGO* https://www.ebi.ac.uk/QuickGO/

16. Bogale, H. N. *et al.* Transcriptional heterogeneity and tightly regulated changes in gene expression during Plasmodium berghei sporozoite development. *Proceedings of the National Academy of Sciences* 118, (2021).

17. Function GO annotations. *UniProt* https://www.uniprot.org/uniprotkb/

18. 1.Silvie, O., Briquet, S., Müller, K., Manzoni, G. & Matuschewski, K. Post-transcriptional silencing of UIS4 in Plasmodium berghei sporozoites is important for host switch. *Molecular Microbiology* 91**,** 1200–1213 (2014).

19. Ribosome. *Genome.gov* https://www.genome.gov/genetics-glossary/Ribosome.

20. Kang, J., Brajanovski, N., Chan, K.T. et al. Ribosomal proteins and human diseases: molecular mechanisms and targeted therapy. Sig Transduct Target Ther 6, 323 (2021). https://doi.org/10.1038/s41392-021-00728-8

21. U2 spliceosomal RNA. Wikipedia (2021). https://en.wikipedia.org/wiki/U2_spliceosomal_RNA.

22. PBANKA_1400800. *PhenoPlasm* https://phenoplasm.org/singlegene.php?gene=PBANKA_1400800.

23. López-López M.J., *et. al*. Biochemical and Biophysical Characterization of the Enolase from Helicobacter pylori. *Biomed Res Int*. 2018:9538193. doi: 10.1155/2018/9538193 (2018).

24. Farhana A, Lappin SL. *Biochemistry, Lactate Dehydrogenase.* Treasure Island (FL): StatPearls Publishing; May 2023. https://www.ncbi.nlm.nih.gov/books/NBK557536

25. Gantt S., *et. al*. Antibodies against thrombospondin-related anonymous protein do not inhibit Plasmodium sporozoite infectivity in vivo. *Infect Immun.* 68(6):3667-73. doi: 10.1128/IAI.68.6.3667-3673.2000. (2000).

26. Triglia, T., Scally, S.W., Seager, B.A. et al. Plasmepsin X activates the PCRCR complex of Plasmodium falciparum by processing PfRh5 for erythrocyte invasion. *Nat Commun* 14, 2219 (2023). https://doi.org/10.1038/s41467-023-37890-2

27. WHO review of malaria vaccine clinical development. *World Health Organization* https://www.who.int/observatories/global-observatory-on-health-research-and-development/monitoring/who-review-of-malaria-vaccine-clinical-development (2024).

APPENDIX A: TOP 20 DEG TABLES BY LINEAGE

Table 4: Top 20 DEG Lineage 1

| **Gene Label** | **Gene Annotation** | **Parametric p-value** |
|---|---|---|
| PBANKA_1039400 | ribosomal protein S27a, putative[11] | 9.97e-155 |
| PBANKA_1355100 | 40S ribosomal protein S6, putative[11] | 9.41e-88 |
| PBANKA_0501200 | early transcribed membrane protein[11], Up-regulated in infective sporozoites[17] (UIS4)[9] | 2.84e-86 |
| PBANKA_0407700 | 60S acidic ribosomal protein P2, putative[11] | 1.56e-67 |
| PBANKA_1234200 | 40S ribosomal protein S24, putative[11] | 1.75e-52 |
| PBANKA_1401300 | 40S ribosomal protein S7, putative[11] | 4.96e-52 |
| PBANKA_1340100 | L-lactate dehydrogenase[11] | 6.41e-49 |
| PBANKA_1202400 | 60S ribosomal protein L13, putative[11] | 1.22e-42 |
| PBANKA_1312700 | gamete egress and sporozoite traversal protein[11] | 5.34e-42 |
| PBANKA_1135100 | 40S ribosomal protein S15, putative[11] | 3.76e-41 |
| PBANKA_1400800 | protein UIS3[22] | 2.13e-40 |
| PBANKA_1128100 | phospholipase[11], UIS10[10] | 6.69e-37 |
| PBANKA_0605500 | conserved Plasmodium protein, unknown function[11] | 5.85e-35 |
| PBANKA_1231700 | 60S ribosomal protein L2, putative[11] | 8.68e-34 |
| PBANKA_0911650 | U2 spliceosomal RNA[11] | 1.13e-33 |
| PBANKA_1346700 | 60S ribosomal protein L23, putative[11] | 2.15e-33 |
| PBANKA_1328000 | serine/threonine protein phosphatase UIS2[11] | 2.68e-33 |
| PBANKA_1117500 | ribosomal protein L27a, putative[11] | 3.98e-33 |
| PBANKA_1354500 | 60S ribosomal protein L18-2, putative[11] | 2.30e-32 |
| PBANKA_1455700 | conserved Plasmodium protein, unknown function[11] | 2.31e-30 |

Table 5: Top 20 DEG Lineage 2

| **Gene Label** | **Gene Annotation** | **Parametric p-value** |
|---|---|---|
| PBANKA_1117500 | ribosomal protein L27a, putative[11] | 7.78e-69 |
| PBANKA_0501200 | early transcribed membrane protein[11], Up-regulated in infective sporozoites[17] (UIS4)[10] | 1.41e-68 |
| PBANKA_1401300 | 40S ribosomal protein S7, putative[11] | 4.67e-67 |
| PBANKA_0923400 | 60S ribosomal protein L35, putative[11] | 3.83e-65 |
| PBANKA_0622921 | 18S ribosomal RNA[11] | 3.44e-57 |
| PBANKA_0407700 | 60S acidic ribosomal protein P2, putative[11] | 4.73e-53 |
| PBANKA_1354500 | 60S ribosomal protein L18-2, putative[11] | 9.53e-47 |
| PBANKA_1039400 | ribosomal protein S27a, putative[11] | 1.27e-46 |
| PBANKA_0911650 | U2 spliceosomal RNA[11] | 2.08e-42 |
| PBANKA_1220800 | conserved Plasmodium protein, unknown function[11] | 3.45e-42 |
| PBANKA_1400800 | protein UIS3[22] | 6.76e-41 |
| PBANKA_1340100 | L-lactate dehydrogenase[11] | 1.04e-37 |
| PBANKA_0403000 | 60S ribosomal protein L44, putative[11] | 6.85e-37 |
| PBANKA_1312700 | gamete egress and sporozoite traversal protein[11] | 1.54e-35 |
| PBANKA_0622961 | 28S ribosomal RNA[11] | 2.05e-34 |
| PBANKA_1457550 | ACEA small nucleolar RNA U3[11] | 8.01e-33 |
| PBANKA_1135100 | 40S ribosomal protein S15, putative[11] | 2.74e-32 |
| PBANKA_0928200 | conserved protein, unknown function[11] | 1.96e-31 |
| PBANKA_1355100 | 40S ribosomal protein S6, putative[11] | 3.53e-29 |
| PBANKA_0417500 | 60S ribosomal protein L32, putative[11] | 1.50e-27 |

Table 6: Top 20 DEG Lineage 3

| **Gene Label** | **Gene Annotation** | **Parametric p-value** |
|---|---|---|
| PBANKA_0501200 | early transcribed membrane protein[11], Up-regulated in infective sporozoites[17] (UIS4)[10] | 1.25e-60 |
| PBANKA_1401300 | 40S ribosomal protein S7, putative[11] | 6.56e-58 |
| PBANKA_0407700 | 60S acidic ribosomal protein P2, putative[11] | 7.12e-52 |
| PBANKA_1039400 | ribosomal protein S27a, putative[11] | 1.61e-51 |
| PBANKA_1117500 | ribosomal protein L27a, putative[11] | 2.44e-50 |
| PBANKA_1400800 | protein UIS3[22] | 5.93e-49 |
| PBANKA_1231700 | 60S ribosomal protein L2, putative[11] | 1.67e-48 |
| PBANKA_1355100 | 40S ribosomal protein S6, putative[11] | 3.02e-43 |
| PBANKA_1013100 | 60S ribosomal protein L14, putative[11] | 3.09e-38 |
| PBANKA_1220800 | conserved Plasmodium protein, unknown function[11] | 7.30e-38 |
| PBANKA_1338600 | 60S ribosomal protein L23, putative[11] | 7.53e-38 |
| PBANKA_1312700 | gamete egress and sporozoite traversal protein[11] | 2.94e-37 |
| PBANKA_1346700 | 60S ribosomal protein L23, putative[11] | 5.44e-37 |
| PBANKA_0923400 | 60S ribosomal protein L35, putative[11] | 4.05e-36 |
| PBANKA_1103400 | 60S ribosomal protein L31, putative[11] | 1.42e-35 |
| PBANKA_1135100 | 40S ribosomal protein S15, putative[11] | 2.27e-35 |
| PBANKA_1340100 | L-lactate dehydrogenase[11] | 2.42e-34 |
| PBANKA_1234200 | 40S ribosomal protein S24, putative[11] | 3.68e-34 |
| PBANKA_0314500 | 40S ribosomal protein S26, putative[11] | 3.98e-34 |
| PBANKA_0403000 | 60S ribosomal protein L44, putative[11] | 2.21e-33 |

Table 7: Top 20 DEG Lineage 4

| **Gene Label** | **Gene Annotation** | **Parametric p-value** |
|---|---|---|
| PBANKA_1039400 | ribosomal protein S27a, putative[11] | 3.38e-64 |
| PBANKA_1002500 | sporozoite-specific protein S10[11] | 1.05e-62 |
| PBANKA_0901300 | membrane associated erythrocyte binding-like protein[11] | 1.18e-59 |
| PBANKA_1204200 | IMP1-like protein, putative[11] | 7.86e-57 |
| PBANKA_1425200 | sporozoite surface protein 3[11] | 9.51e-44 |
| PBANKA_0403200 | circumsporozoite (CS) protein[11] | 6.47e-39 |
| PBANKA_1465051 | Plasmodium exported protein, unknown function[11] | 1.02e-33 |
| PBANKA_1346700 | 60S ribosomal protein L23, putative[11] | 2.60e-33 |
| PBANKA_1306500 | TRAP-like protein[11] | 1.13e-31 |
| PBANKA_1018600 | 60S ribosomal protein L21, putative[11] | 4.42e-30 |
| PBANKA_0407700 | 60S acidic ribosomal protein P2, putative[11] | 2.22e-29 |
| PBANKA_1106700 | 60S ribosomal protein L4, putative[11] | 1.08e-28 |
| PBANKA_1206800 | zinc finger (CCCH type) protein, putative[11] | 2.42e-26 |
| PBANKA_0615900 | cysteine repeat modular protein 2[11] | 3.69e-26 |
| PBANKA_1123800 | GAS8-like protein, putative[11] | 3.14e-25 |
| PBANKA_0617200 | 40S ribosomal protein S10, putative[11] | 1.45e-24 |
| PBANKA_0417500 | 60S ribosomal protein L32, putative[11] | 5.25e-24 |
| PBANKA_1433700 | conserved Plasmodium protein, unknown function[11] | 7.19e-24 |
| PBANKA_1025700 | inner membrane complex protein 1l, putative[11] | 1.09e-23 |
| PBANKA_0923400 | 60S ribosomal protein L35, putative[11] | 2.15e-23 |

APPENDIX B: TOP 20 DEG TABLES BY ANNOTATION GROUP

Table 8: UIS genes in top 20 DEG

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_0501200 | early transcribed membrane protein[11], Up-regulated in infective sporozoites[17] (UIS4)[10] | 3 | 2 | 1 | |
| PBANKA_1400800 | early transcribed membrane protein (UIS3)[22] | 11 | 11 | 6 | |
| PBANKA_1128100 | phospholipase[11], UIS10[10] | 12 | | | |
| PBANKA_1328000 | serine/threonine protein phosphatase UIS2[11] | 17 | | | |

Table 9: Enzymes ending in -ase in top 20 DEG

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_1340100 | L-lactate dehydrogenase[11] | 7 | | 7 | |
| PBANKA_1128100 | phospholipase[10], UIS10[10] | 12 | | | |
| PBANKA_1328000 | serine/threonine protein phosphatase UIS2[11] | 17 | | | |

Table 10: Ribosomal proteins in top 20 DEG

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_1039400 | ribosomal protein S27a, putative[11] | 1 | 8 | 4 | 1 |
| PBANKA_1355100 | 40S ribosomal protein S6, putative[11] | 2 | 19 | 8 | |
| PBANKA_0407700 | 60S acidic ribosomal protein P2, putative[11] | 4 | 6 | 3 | 11 |
| PBANKA_1234200 | 40S ribosomal protein S24, putative[11] | 5 | | 18 | |
| PBANKA_1401300 | 40S ribosomal protein S7, putative[11] | 6 | 3 | 2 | |
| PBANKA_1202400 | 60S ribosomal protein L13, putative[11] | 8 | | | |
| PBANKA_1135100 | 40S ribosomal protein S15, putative[11] | 10 | 17 | 16 | |
| PBANKA_1231700 | 60S ribosomal protein L2, putative[11] | 14 | | 7 | |
| PBANKA_1346700 | 60S ribosomal protein L23, putative[11] | 16 | | 13 | 8 |
| PBANKA_1117500 | ribosomal protein L27a, putative[11] | 18 | 1 | 5 | |
| PBANKA_1354500 | 60S ribosomal protein L18-2, putative[11] | 19 | 7 | | |
| PBANKA_0923400 | 60S ribosomal protein L35, putative[11] | | 4 | 14 | 20 |
| PBANKA_0403000 | 60S ribosomal protein L44, putative[11] | | 13 | 20 | |
| PBANKA_0417500 | 60S ribosomal protein L32, putative[11] | | 20 | | 17 |
| PBANKA_1013100 | 60S ribosomal protein L14, putative[11] | | | 9 | |
| PBANKA_1338600 | 60S ribosomal protein L23, putative | | | 11 | |
| PBANKA_1103400 | 60S ribosomal protein L31, putative[11] | | | 15 | |
| PBANKA_0314500 | 40S ribosomal protein S26, putative[11] | | | 19 | |
| PBANKA_1018600 | 60S ribosomal protein L21, putative[11] | | | | 10 |
| PBANKA_1106700 | 60S ribosomal protein L4, putative[11] | | | | 12 |
| PBANKA_0617200 | 40S ribosomal protein S11, putative[11] | | | | 16 |

Table 11: Genes with unknown function in top 20 DEG

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_0605500: | conserved Plasmodium protein, unknown function[11] | 13 | | | |
| PBANKA_1455700 | conserved Plasmodium protein, unknown function[11] | 20 | | | |
| PBANKA_1220800 | conserved Plasmodium protein, unknown function[11] | | 10 | 10 | |
| PBANKA_0928200 | conserved protein, unknown function[11] | | 18 | | |
| PBANKA_1465051 | Plasmodium exported protein, unknown function[11] | | | | 7 |
| PBANKA_1433700 | conserved Plasmodium protein, unknown function[11] | | | | 18 |

Table 12: All other top 20 DEG in Lineages 1, 2 or 3 not significant in Lineage 4

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_0911650 | U2 spliceosomal RNA[11] | 15 | 9 | | N/A |
| PBANKA_0622921 | 18S ribosomal RNA[22] | | 5 | | N/A |
| PBANKA_0622961 | 28S ribosomal RNA[11] | | 15 | | N/A |
| PBANKA_1457550 | ACEA small nucleolar RNA U3[11] | | 16 | | N/A |

Table 13: All other top 20 DEG

| | | **Top 20 DEG Significance Ranking in Lineage...** | | | |
|---|---|---|---|---|---|
| **Gene Label** | **Gene Annotation** | **1** | **2** | **3** | **4** |
| PBANKA_1312700 | gamete egress and sporozoite traversal protein[11] | 9 | 4 | 12 | |
| PBANKA_1002500 | sporozoite-specific protein S10[11] | | | | 2 |
| PBANKA_0901300 | membrane associated erythrocyte binding-like protein[11] | | | | 3 |
| PBANKA_1204200 | IMP1-like protein, putative[11] | | | | 4 |
| PBANKA_1425200 | sporozoite surface protein 3[11] | | | | 5 |
| PBANKA_0403200 | circumsporozoite (CS) protein[11] | | | | 6 |
| PBANKA_1306500 | TRAP-like protein[11] | | | | 9 |
| PBANKA_1206800 | zinc finger (CCCH type) protein, putative[11] | | | | 13 |
| PBANKA_0615900 | cysteine repeat modular protein 2[11] | | | | 14 |
| PBANKA_1123800 | GAS8-like protein, putative[11] | | | | 15 |
| PBANKA_1025700 | inner membrane complex protein 1l, putative[11] | | | | 19 |

APPENDIX C: TOP 20 CELL EXPRESSION VS PSEUDOTIME PLOTS FOR RP DEG

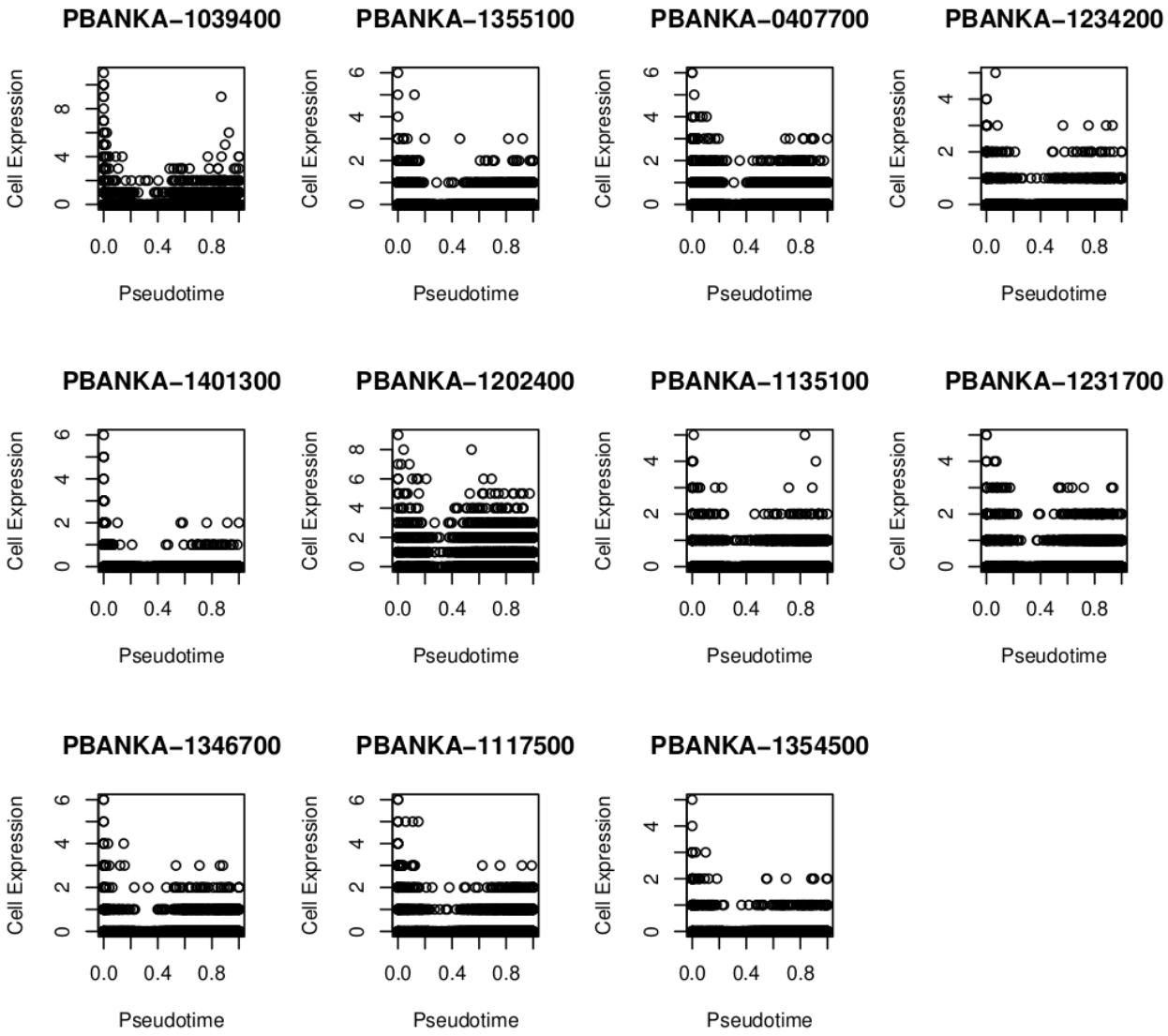



Figure 12: Cell expression vs pseudotime plots for all top 20 RP DEG for Lineage 1

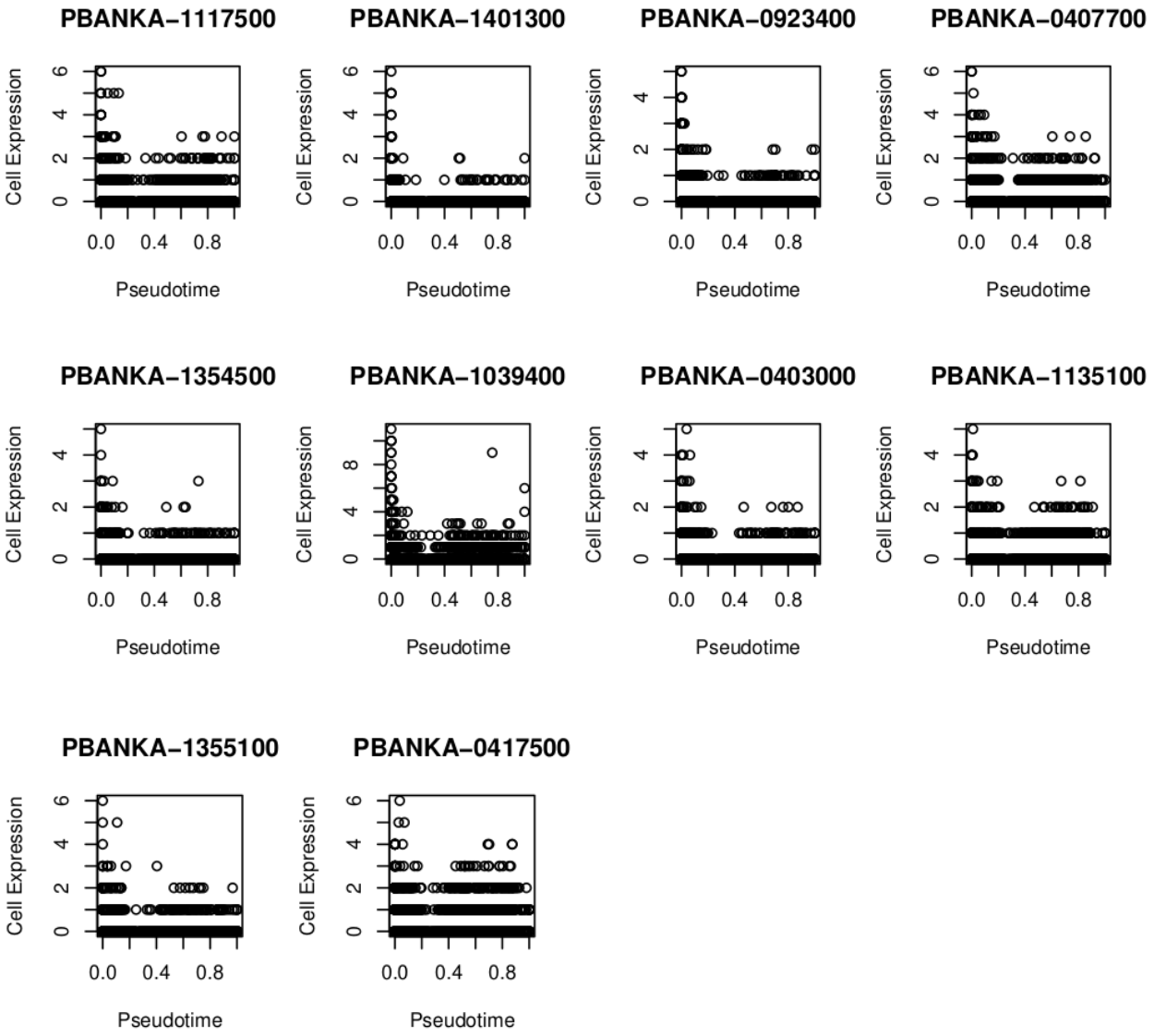



Figure 13: Cell expression vs pseudotime plots for all top 20 RP DEG for Lineage 2

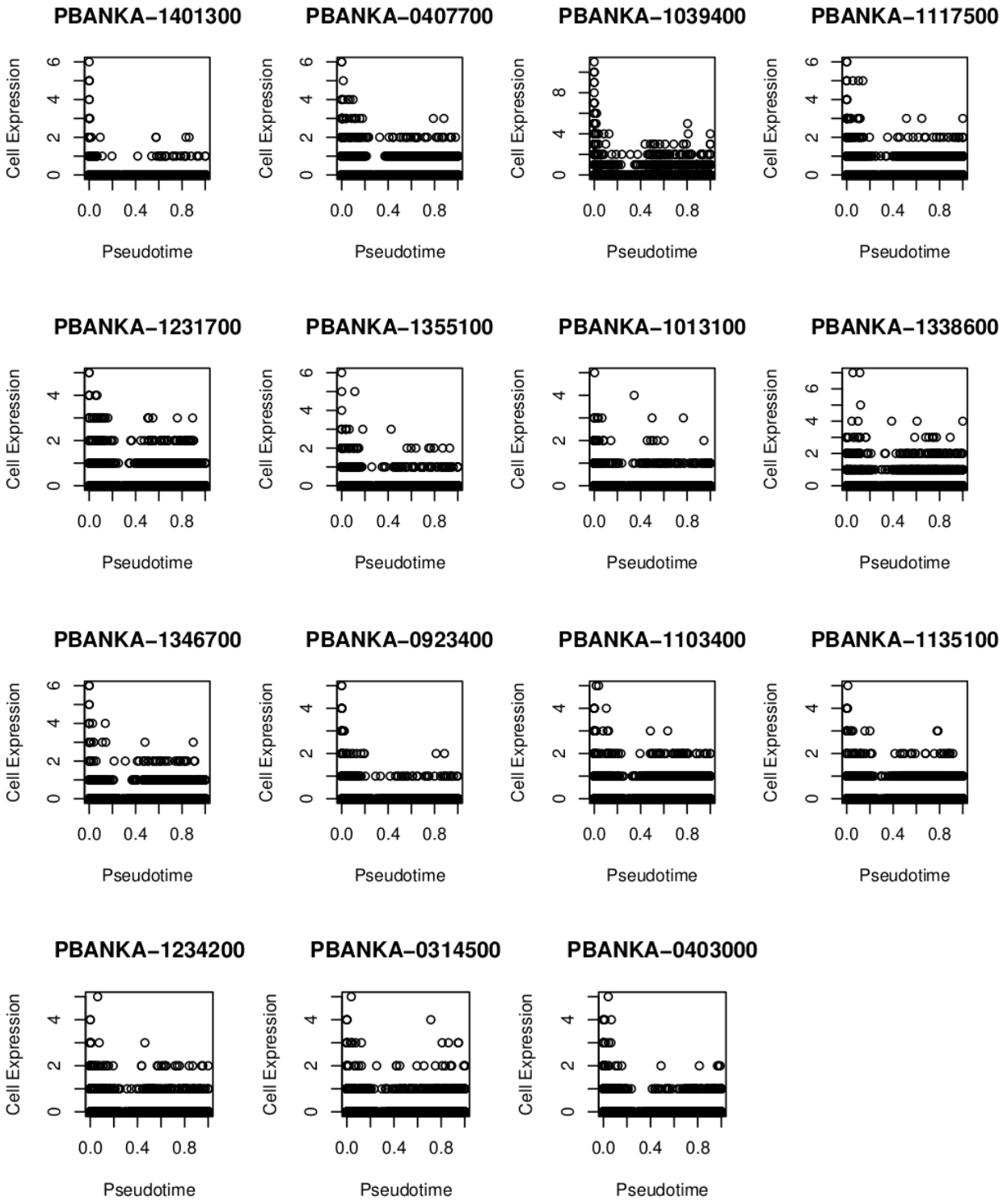



Figure 14: Cell expression vs pseudotime plots for all top 20 RP DEG for Lineage 3

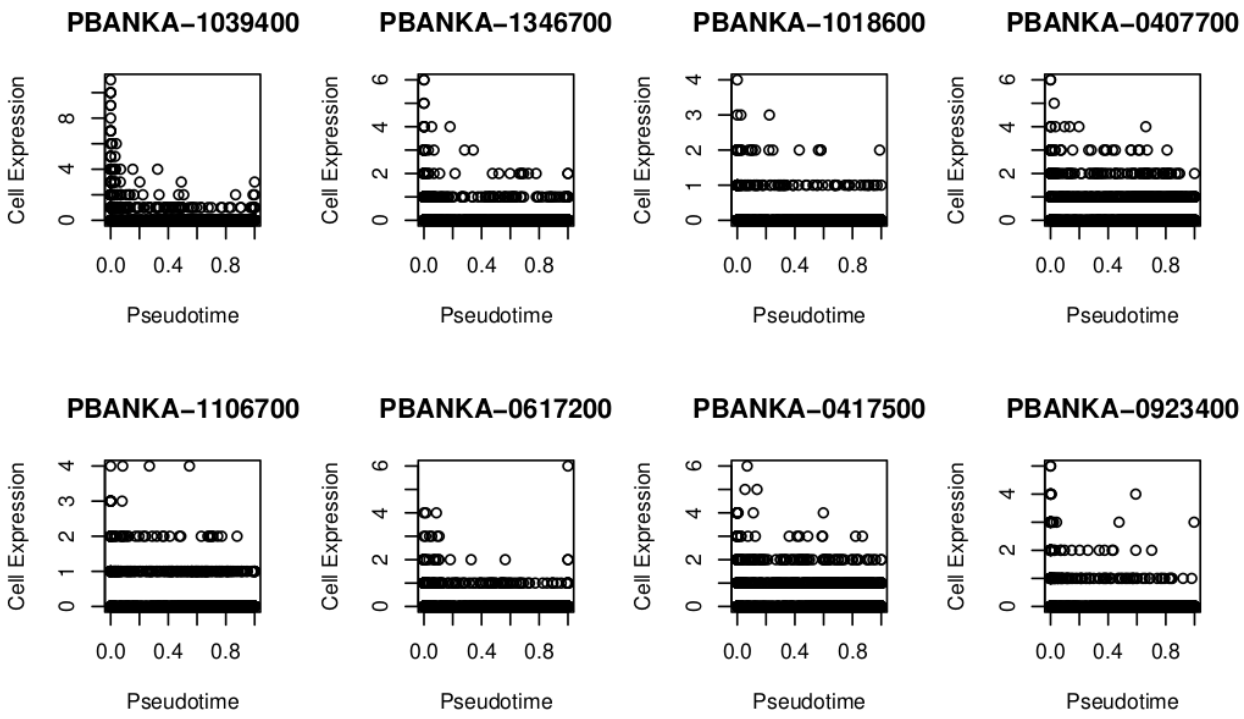



Figure 15: Cell expression vs pseudotime plots for all top 20 RP DEG for Lineage 4