A computational approach to analyzing gene expression in *Plasmodium falciparum*.

Faustino Cortina

Abstract

In an effort to gain a better understanding of the genes responsible for the virulence and drug resistance of the *Plasmodium falciparum* parasite, a large quantity of transcript expression data has been published. However, it is not always easy to establish conclusive results when analyzing individual genes since it is hard to link one's findings to the context of a larger biological process. Additionally, there is a significant amount of unannotated genes within the *P. falciparum* genome, limiting the number of genes that can accurately be used for transcript expression experiments. In order to further explore gene expression data for the *P. falciparum* parasite, I used a method called Gene Set Enrichment Analysis (GSEA). With this method, user-defined gene sets representing different biological processes were used to analyze published gene expression data. I built these gene sets based on experimental data on subcellular localization and gene expression during the intraerythrocytic and gametocytic stages of the *P. falciparum* life cycle. I then used these gene sets to analyze biological datasets with GSEA. Due to the widespread mutational resistance of *P. falciparum* to antimalarial drugs, I focused my research on analyzing gene expression experiments on *P. falciparum* resistance to the antimalarial drugs thiostrepton and chloroquine (CQ). The GSEA results showed potential correlations in CQ resistance in seven different localizations. The results also suggest that *P. falciparum* responds to thiostrepton by increasing expression of genes located in the mitochondrion and apicoplast. The computational method used in this paper can be easily replicated with different gene expression datasets and/or gene sets to provide valuable insight about the biological mechanisms within the *P. falciparum* parasite.

Background

The *Plasmodium falciparum* parasite is responsible for causing malaria, a disease responsible for 214 million cases of infection and 438,000 deaths in 2015 [1]. Of the five known *Plasmodium* species capable of infecting humans, *P. falciparum* is the most lethal species and the most prevalent in Africa, where the majority of malaria infections and deaths occur [2]. Although there are various antimalarial drugs that can be used to treat malaria, the *P. falciparum* parasite has continuously developed resistance to these drugs, limiting treatment options. For instance, *P. falciparum* has developed widespread resistance to chloroquine (CQ), a drug that has been heavily used to combat malaria for decades [2].

Transcriptomics, a method for measuring gene expression, has become crucial to understanding the biological mechanisms that allow *P. falciparum* to resist antimalarial drugs. A significant area of focus for many transcriptomic studies in *P. falciparum* is the intraerythrocytic developmental stage of the *P. falciparum* life cycle, a recurring 48-hour cycle post-infection that makes malaria symptomatic. The intraerythrocytic developmental stage occurs in the bloodstream after the parasite invades the liver cells of the infected human host to produce merozoites. During the ring phase of the intraerythrocytic cycle, merozoites invade red bloods cells (erythrocytes) and develop into a young, ring-shaped trophozoite. In the trophozoite stage, the trophozoite grows in size and finishes its development within the erythrocyte. Finally, during the schizont stage, the trophozoite undergoes mitotic divisions to form 16-18 merozoites within the erythrocyte, and the erythrocyte ruptures to re-introduce the merozoites into the bloodstream and restart the cycle [1]. Since most antimalarial drugs target *P. falciparum* during this intraerythrocytic cycle, understanding how the parasites alter processes that occur during this cycle in response to drug treatment is very important to malaria research and finding out why parasites become drug resistant. Another important phase of the *P. falciparum* life cycle is the gametocytic stages in which sexual development occurs in the mosquito host. When a female *Anopheles* mosquito has a blood meal on a human infected with *P. falciparum*, the mosquito ingests specialized gametocyte

parasites that develop in the mosquito's midgut into sporozoites ready to be inoculated in a new host. The process of gametocytic development in the mosquito's midgut typically occurs over 10-12 days [1]. By understanding what happens during gametocyte development, researchers can gain insight to the biological processes that enable *P. falciparum* to perpetuate its life cycle.

Interpreting the vast amounts of data generated from transcriptomics in a meaningful way is a difficult task. In initial gene expression analyses, one can identify genes that have statistically significant gene expression patterns between two conditions. However, this analysis cannot determine any broader connections that these genes have in relation to the parasite's broader biological processes. Gene Set Enrichment Analysis (GSEA), a computational method for analyzing gene expression data [3], enables researchers to gain a more comprehensive analysis of their gene expression results. Gene sets corresponding to different biological processes are compiled and used to determine whether a gene set representing a whole pathway or biological process has any significant correlation with a phenotype or biological state. Rather than looking at the expression level of individual genes, the entire pathway is considered. The paper by Croken et al. [4] analyzing the genome of *Toxoplasma gondii* using GSEA served as a good model for my research, as its method of analyzing published gene expression data with user-defined gene sets is the same approach I took in my research.

In my experiment, I applied GSEA to gene expression datasets of *P. falciparum* relating to drug treatment. Using gene sets based on intraerythrocytic development, gametocyte formation, and localization, I was able to gain valuable insight into the effect that antimalarial drugs have on different biological processes within the *P. falciparum* parasite.

Methods

Gene Set Enrichment Analysis (GSEA)

GSEA is a computational method by the Broad Institute website (http://www.broadinstitute.org/gsea/index.jsp) that determines whether genes have a correlation between two biological states [3]. The GSEA results in this paper were obtained using the Java Desktop Application implementation of GSEA. GSEA uses an algorithm to determine the correlation a gene set has with a biological state by calculating an enrichment score (ES). The ES is then normalized to account for the size of the gene set to create a normalized enrichment score (NES). The higher the magnitude of the NES, the stronger the enrichment of the gene set with the biological state specified [3].

Developing Gene Sets

In order to use GSEA to gain a broader understanding of how *P. falciparum* functions, gene sets must be created that each contain a biological process in common. Gene sets were created according to localization and peak gene expression during intraerythrocytic development and gametocyte formation as described below. I aimed to create gene set lists of 15 to 500 genes, the default gene set range of GSEA.

Localization Gene Sets

The ApiLoc website (http://apiloc.biochem.unimelb.edu.au) is a curated database recording protein subcellular localization in apicomplexan parasites. Annotated *P. falciparum* gene data was taken from the Github repository of the ApiLoc database (https://github.com/wwood/ApiLoc/tree/master/raw_data). This raw data was refined using the R programming language to a data frame providing each gene with a brief description of its subcellular localization. From this table, gene sets according to localization were created by searching for key words in each gene's localization description. For instance, to create a gene set for genes that are localized in the nucleus, genes containing the word "nucleus" in their localization description were used. Some localizations, such as apical organelles, needed multiple key words, like "rhoptry

bulb" and "rhoptry neck", to create a complete gene set. Also, many localization descriptions use key words to explicitly mention that they are not affiliated with the gene, so I needed to filter out the expressions such as "not nucleus" or "not apical" when making the gene sets. Since the gene set sizes for the endoplasmic reticulum and the golgi apparatus localizations were below the 15 gene minimum to be used for GSEA (and those compartments cannot easily be resolved by microscopy), I combined the two localizations together.

Peak Gene Expression Gene Sets

Gene sets for intraerythrocytic peak gene expression were created using data from PlasmoDB, a community genome database for the plasmodium genus [5]. I used gene expression data from five published experiments: three were from the intraerythrocytic developmental stage [6,7,8], one was for the gametocytic stages [9], and one was for both the intraerythrocytic and gametocytic stages [10]. For clarity, the intraerythrocytic developmental stage gene expression experiments conducted by Bozdech et al. [6], Bartfai et al. [7], and Hoeijmakers [8] will be referred to as Intraerythrocytic-1, Intraerythrocytic-2, and Intraerythrocytic-3, respectively, for the duration of this paper. The gametocyte gene expression experiments conducted by Young et al. [9] will be referred to as Gametocyte-1, and the experiment conducted by Lopez-Barragan et al. [10] will be referred to as Intraerythrocytic-Gametocyte-1. All of the experimental data was obtained from 3D7 strain *P. falciparum* parasites. For each gene, peak expression was defined as the mean of its expression values across every timepoint plus 1.5 standard deviations. Gene sets were created at each timepoint containing genes that satisfied the peak expression criteria.

The gene names included in the ApicLoc database are outdated. In order for the localization gene sets to be compatible for GSEA, these outdated gene names were updated to the current nomenclature for *P. falciparum*. A gene alias file in http://plasmodb.org/ containing old and current gene symbol names for each gene in the 3D7 strain of *P. falciparum* was used to re-annotate these genes.

Test Datasets

In order to test the compiled gene sets with GSEA, I used two published microarray data sets from experiments examining how *P. falciparum* gene expression is affected by antimalarial drugs. All the datasets contained 3 biological replicates, the minimum number of replicates required for GSEA to run.

The dataset GSE28701 analyzed the effects of ring-phase *P. falciparum* when exposed to the thiostrepton [14], and the dataset GSE10022 analyzed the effects of chloroquine on mutant lines of *P. falciparum* [11]. Both of these datasets were generated using the Affymetrix Plasmodium/Anopheles Genome Array (http://www.affymetrix.com/catalog/131485/AFFY/Plasmodium+Anopheles+Genome+Array#1_3).

Expression data from these two data sets were compiled using the ExpressionFileCreater module from the GenePattern platform available on the Broad Institute website (http://www.broadinstitute.org/cancer/software/genepattern/). This module converts the raw CEL data files into the GCT format that is then used directly in GSEA for analysis.

So that the GSEA program can determine the genes that each probe on the microarray corresponds with, **t**he probe labels in the Affymetrix Plasmodium/Anopheles Genome Array needed to be compatible. A CSV file containing the most up to date probe annotations for the Plasmodium/Anopheles Genome Array was used for this purpose (http://www.affymetrix.com/catalog/131485/AFFY/Plasmodium+Anopheles+Genome+Array#1_3). The file was then manually modified to follow the file format for a CHIP file as shown in the File Format Guide found on the GenePattern platform of the Broad Institute website (http://software.broadinstitute.org/cancer/software/genepattern/file-formats-guide).

Results and discussion

Profiling gene expression in *Plasmodium falciparum* containing *pfcrt* gene mutations

Mutations in the *P. falciparum* chloroquine-resistant transporter (*pfcrt*) at amino acid position 76 cause chloroquine (CQ) resistance in *P. falciparum* [11]. In the expression experiment conducted by Jiang et al. (GSE10022) [11], gene expression data was taken from two parasite profiles with different *pfcrt* mutations at the 76 position, $106/1^{76I}$ and $106/1^{76I\text{-}352K}$, and their parental line, $106/1^{K76}$, which has no mutation at position 76. For clarity, the $106/1^{K76}$ parent line will be abbreviated to 106, the $106/1^{76I}$ to 176I, and the $106/1^{76I\text{-}352K}$ mutant to 352K. The 176I mutant was derived from the 106 line after it was exposed to a lethal dose of CQ, selecting for a mutation at amino acid 76. The resulting 176I mutant is resistant to CQ but extremely sensitive to quinine (QN). The 352K mutant was derived from the 176I parasite through a lethal dose of QN. The 352K line is resistant to QN but sensitive to CQ. All three parasite lines have the same growth rate if not treated with chloroquine [11].

Since GSEA compares two phenotypes at a time, GSEA was run for six different phenotype comparisons. The three untreated parasite lines were compared with one another (176I vs. 106, 352K vs. 106, and 352K vs. 176I) and the treated and untreated profiles of each line (106 vs. 106CQ, 176I vs. 176ICQ, and 352K vs. 352KCQ). After running GSEA, gene sets with a FWER (family wise error rate) q-value greater than 0.05 were not considered statistically significant and were assigned normalized enrichment scores (NES) of 0.

Results were plotted as a heat map (Figure 1). Untreated mutant parasite lines show similar growth kinetics [11]. In the comparisons between the 352K line with the 106 parent line and the 176I line with the 106 line, the gene sets of peak expression during the intraerythrocytic cell cycle were enriched in the first phenotype compared (352K or 176I) during the late ring and trophozoite stages and enriched in the 106 phenotype during the early ring and schizont stages. The lack of enrichment of cell activity during

the late ring and trophozoite stages of the two profiles with a *pfcrt* mutation would explain why studies have found that the *pfcrt* gene is most sensitive to drug resistance during the trophozoite stage [12,13] since a reduction in cell activity would likely make the parasite more vulnerable to external pressures such as antimalarial drugs.

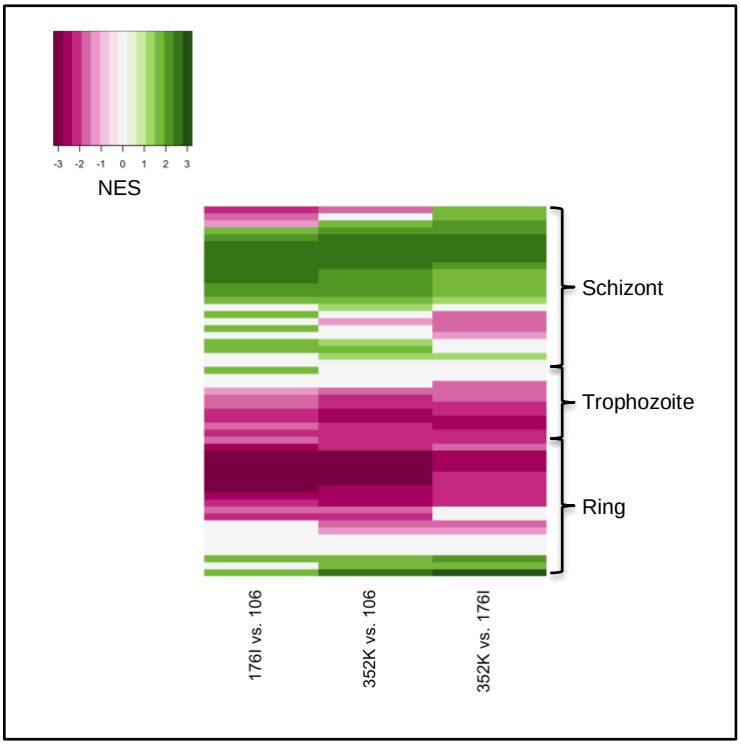


Figure 1. Heatmap representing the GSEA results comparing the three *P. falciparum* profiles from the experiment conducted by Jiang et al. [11] with the Intraerythrocytic-1 gene set collection. Green indicates a positive normalized enrichment score (NES), which means the first of the two phenotypes compared for each comparison is enriched. Red indicates a negative NES, which means that the second phenotype listed in each comparison is enriched. White represents no enrichment or a statistically insignificant result. The gene sets on the vertical axis are arranged in chronological order, and each gene set is an hourly interval of gene expression during the intraerythrocytic developmental cycle. All three comparisons exhibit the same pattern of enrichment in the second phenotype compared during the late ring and trophozoite stages and enrichment in the first phenotype compared during the early ring and schizont stages.

Although both *pfcrt* mutants had similar changes in gene expression relative to the parent 106 line, they exhibited significant differences when compared with one another. Comparing 352K with 176I, I noticed a similar pattern: gene sets were associated with

late ring and trophozoite stages were underrepresented and gene sets associated early ring and schizont stages were enriched.

Roles of peak gene expression and subcellular localization in chloroquine drug resistance

We compared gene expression in untreated and CQ treated parasite lines using GSEA. For each comparison, there was a significant difference observed when parasites were treated with CQ (Figure 2). For 176I, which is chloroquine resistant, many gene sets are enriched in the untreated line – suggesting that though it is resistant, there is a persistent growth defect in the mutant strain under CQ treatment. In the parental strain 106, which is chloroquine sensitive, very few gene sets were enriched in the untreated line compared to treated, which presumably is because the parasites die under CQ treatment. Finally, with the double mutant 352K line, in which CQ sensitivity is restored, I was surprised that many gene sets are enriched during CQ treatment. This suggests there are complex transcriptional changes in CQ treated 352K parasites which may lead to reemergence of resistance in that strain.

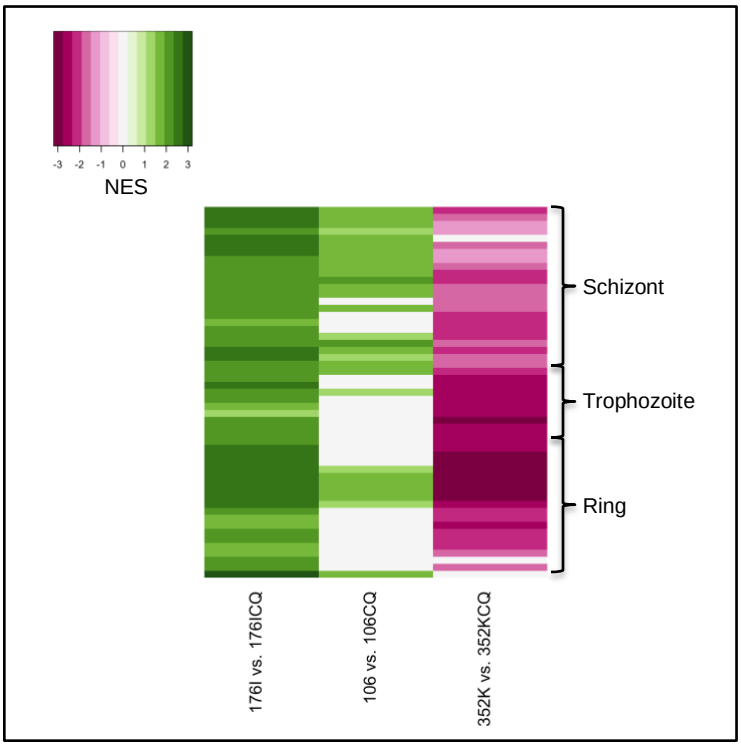


Figure 2. Heatmap representing the GSEA results comparing the untreated and CQ treated lines of the three *P. falciparum* profiles from the experiment conducted by Jiang

et al. [11] with the Intraerythrocytic-1 gene set collection. Green indicates a positive NES and enrichment in the untreated parasite profile, and red indicates a negative NES and enrichment with the CQ treated parasite profile. White represents no enrichment or a statistically insignificant result. The gene sets on the vertical axis are arranged in chronological order, and each gene set is an hourly interval of gene expression during the 48-hour intraerythrocytic developmental cycle. Although 106 and 352K are both sensitive to CQ, they exhibit different enrichment patterns, suggesting complex transcriptional changes in 352K that may relate to reemergence of CQ/QN resistance.

We next used GSEA to see the effects CQ treatment had on the gametocytic cycle. I found the similar results to what I had seen in the intraerythrocytic cycle: there was enrichment in the untreated phenotype in the 176I vs. 176ICQ and 106 vs. 106CQ comparisons and enrichment in the treated phenotype in the 352K vs. 352KCQ comparison (Figure 3). The similarity of the results from both these stages of the *P. falciparum* life cycle suggests that antimalarial drugs such as CQ would have lingering effects on *P. falciparum* gene expression even after it leaves the host's body and undergoes sexual development.

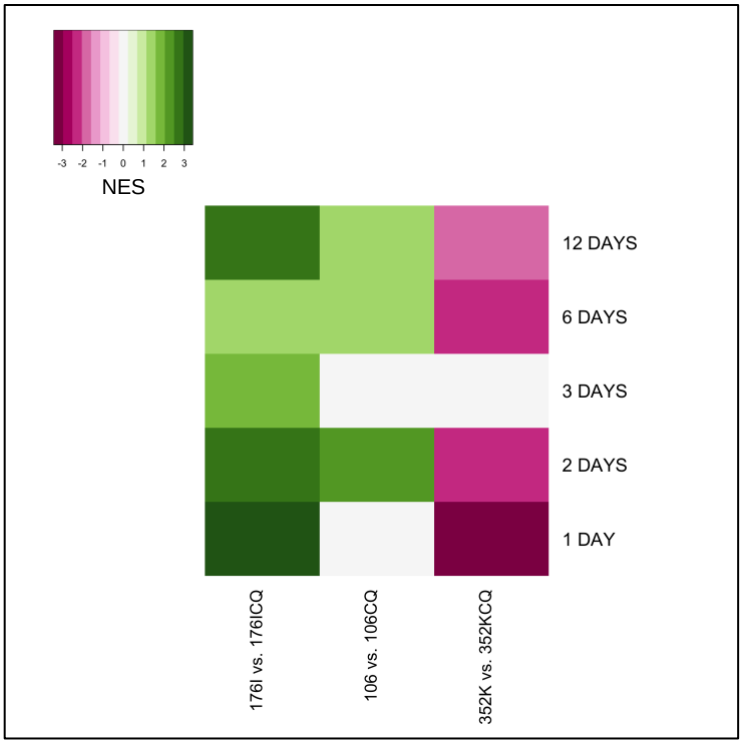


Figure 3. Heatmap representing the GSEA results comparing the untreated and CQ treated lines of the three *P. falciparum* profiles from the experiment conducted by Jiang et al. [11] with the Gametocyte-1 gene set collection. Green indicates a positive NES

and enrichment in the untreated parasite profile, and red indicates a negative NES and enrichment with the CQ treated parasite profile. The gene sets on the vertical axis represent gene expression during the gametocytic cycle. Similar to Figure 2, the 176I vs. 176ICQ and 106 vs. 106CQ comparisons exhibit enrichment in the untreated versus treated phenotypes, and the 352K vs. 352KCQ comparison exhibits enrichment in the 352KCQ phenotype.

In order to determine whether localization has a significant effect on the CQ drug resistance in the *pfcrt* gene, I ran GSEA with the same phenotype comparisons as above, only with gene sets relating to localization. Many gene sets relating to localization were not significantly enriched in any comparison. For the 176I vs. 176ICQ comparison, all of the statistically significant gene sets were enriched in untreated parasites, and for the 352K vs. 352KCQ comparison, all but one of the statistically significant gene set results were enriched the treated line. Seven gene sets were enriched in 176I compared to 176ICQ, and in 352KCQ compared to 352K (Figure 4). Since the 352K mutant is CQ sensitive and the 176I mutant is CQ resistant, the genes in one or more of these seven gene sets may play a direct role in determining CQ resistivity in *P. falciparum* parasites. Unfortunately, it is impossible to know which, if any, of these subcellular localizations have a direct correlation to CQ drug sensitivity. If the GSEA results from the 106 vs. 106CQ comparison had been statistically significant enough to consider reliable data, the third comparison would likely have narrowed down the number of gene sets with potential correlations to CQ resistance.

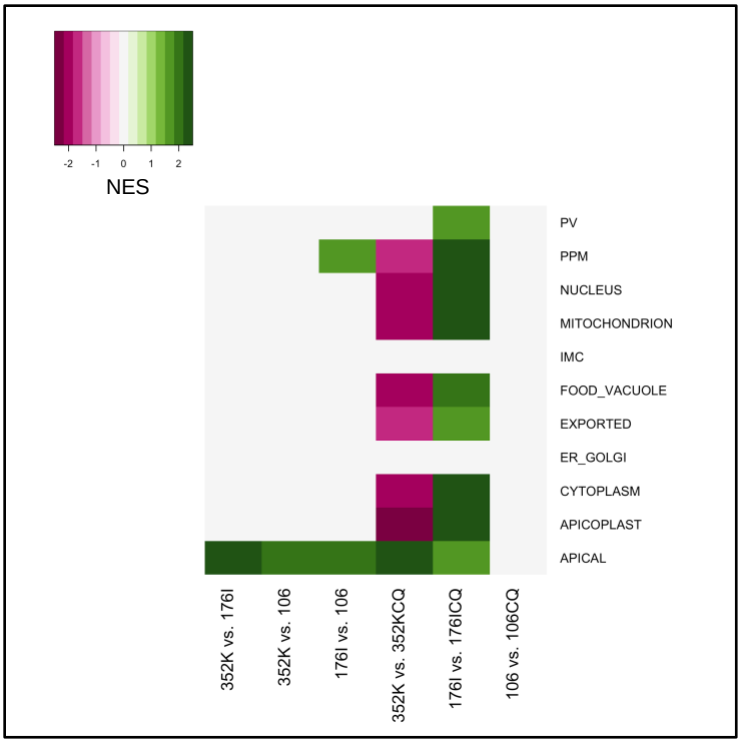
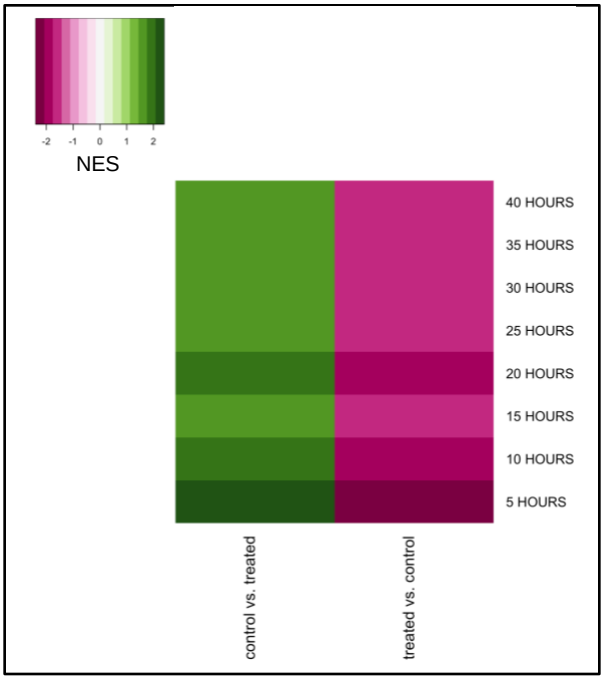


Figure 4. GSEA results of six different phenotype comparisons from the datasets in the Jiang et al. experiment [11] tested for enrichment with gene sets based on subcellular localization. Only the 352K vs. 352KCQ and 176I vs. 176ICQ comparisons exhibited more than one gene set with statistically significant enrichment.

Cell cycle response to thiostrepton

In their gene expression experiment, Tarr et al. [14] used transcriptomics to shed light on how parasites respond to thiostrepton. They measured the gene expression of *P. falciparum* when treated with a non-lethal dose of thiostrepton and when treated with a DMSO-control that contained no thiostrepton. To see whether thiostrepton treatment exhibited any trends during intraerythrocytic development, I conducted GSEA using the intraerythrocytic gene sets. Since there were only two phenotypes in the dataset, the DMSO-control and thiostrepton treated, GSEA was only run once to compare the treated parasite with the control. Two of the peak gene expression gene sets, Intraerythrocytic-2 and Intraerythrocytic-3, showed enrichment in the control versus the treated phenotype throughout the cell cycle (Figure 5a). One possible factor for this genome-wide enrichment in the control is that the treated parasites experienced some irregularities in their cell cycle progression. Despite appearing phenotypically normal 24

hours into the treatment, not all of the parasites had progressed into the next cell cycle 49-52 hours into the treatment [14]. This delay in cell cycle progression likely caused gene expression to appear under-expressed for the parasites influenced by thiostrepton, causing negative enrichment for the treated phenotype. However, when I analyzed the Jiang et al. experiment with the other two data sets containing peak gene expression in the intraerythrocytic cell cycle, Intraerythrocytic-1 and Intraerythrocytic-Gametocyte-1, enrichment of trophozoite stage gene sets was associated with the treated parasites (Figure 5b). Since the Intraerythrocytic-1 gene set collection contains gene expression data at hourly intervals while the other three experiments recorded at intervals of 5 hours or more, it is most likely that the result from Intraerythrocytic-1 showing enrichment in the treated parasites during the trophozoite stage is the most accurate. One possibility why the thiostrepton treated parasites were enriched during the trophozoite stage is that some of the parasites had not progressed past the trophozoite stage, a result that occurs when *P. falciparum* is treated with high doses of thiostrepton [17]. Even though the parasites were not treated with a high enough thiostrepton dose to cause this effect to occur in all the parasites, there is a chance that a small percentage of these parasites were suspended in the trophozoite stage. If this were the case, then it would make sense that the gene sets of peak expression during the trophozoite phase were enriched in treated parasites since there would be a larger than expected percentage of parasites in the trophozoite stage when treated with thiostrepton. Although Tarr et al. looked for abnormalities in cell progression 24 hours and 49-52 hours into the treatment, there was no mention of observations of cell cycle progression after the trophozoite phase and before the end of the first cell cycle. As a result, there is no evidence for or against the presence of parasites suspended in the trophozoite stage during the first cycle of the treatment. However, given that Tarr et al. did find abnormal cell cycle progression at the end of the first cycle, it would not be surprising if thiostrepton had also delayed the parasite's life cycle in an earlier intraerythrocytic stage.



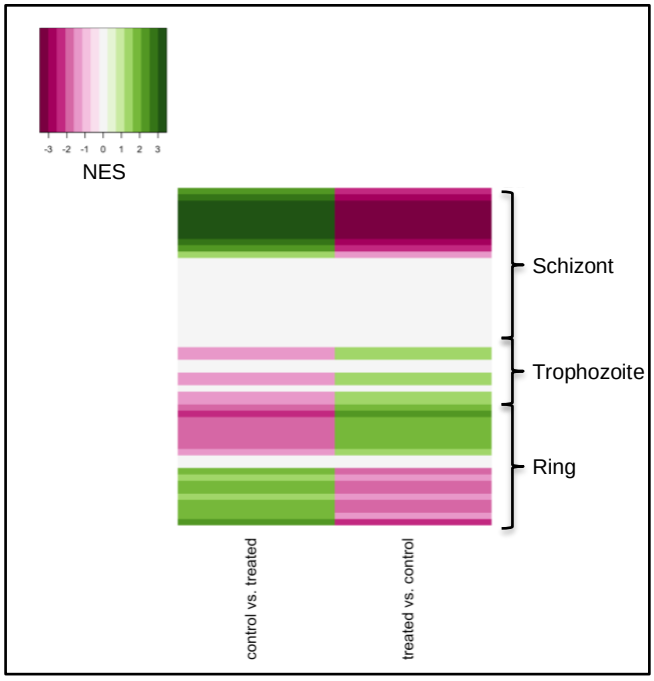
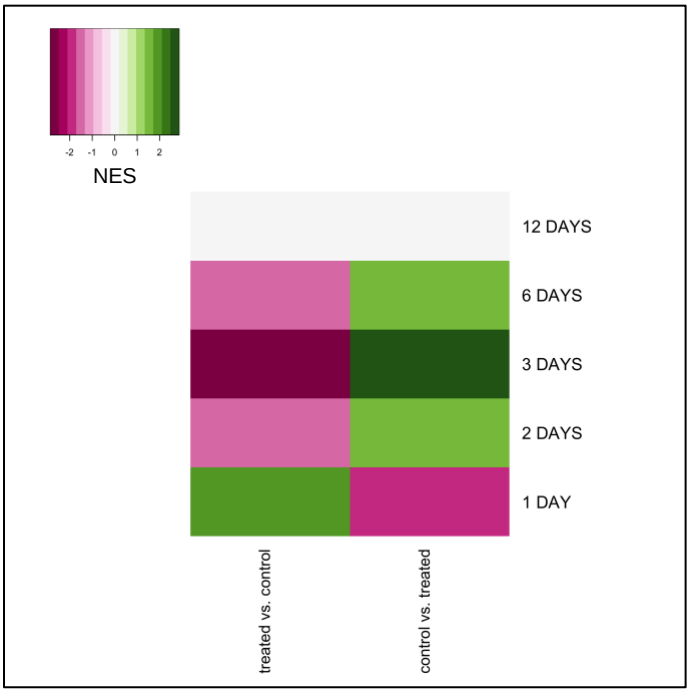


Figure 5a (left). One of two gene set collections (Intraerythrocytic-2) showing enrichment in the control parasites throughout the intraerythrocytic developmental. The datasets represented by the horizontal labels recorded peak gene expression in the intraerythrocytic cell cycle every 5 hours.

Figure 5b (right). One of two gene set collections (Intraerythrocytic-1) showing enrichment of *P. falciparum* when treated with thiostrepton during the trophozoite stage. Since Intraerythrocytic-1 collected gene expression data most regularly out of the other three gene expression data set collections for the intraerythrocytic cycle, it is most likely that the enrichment in the trophozoite stage of the thiostrepton treated parasites found in this experiment what actually occurs in the *P. falciparum* parasite.

Figure 5c (bottom). Heatmap showing the Gametocytic-1 gene sets NES values when *P. falciparum* is treated with thiostrepton. Three of the four statistically significant gene sets are enriched in the control phenotype, and the Day 1 gene set is enriched in the treated phenotype.

(note that the two columns in each heatmap are the same comparison, only flipped. Heat maps created in R typically require at least two columns)



When GSEA was run with the Gametocyte-1 gene set collection and thiostrepton dataset, there was enrichment in the control parasites in all but one of the statistically significant gene sets. Similar to the 176I vs. 176ICQ comparison in the Jiang et al. experiment [11], the widespread enrichment in the untreated phenotype may be a result of a growth defect caused by the drug. The only gene set with positive enrichment in the treated parasites was the gene set representing gene expression on the first day of the gametocytic stage. This result shows the possibility that *P. falciparum* exhibits a transcriptional response to thiostrepton early into its sexual development in the mosquito.

Response to thiostrepton in the in the apicoplast and mitochondrial localizations

Although thiostrepton is widely believed to target the ribosomes in the apicoplast organelle of *P. falciparum* [16,17], Tarr et al. showed that thiostrepton has a more significant effect on mitochondrial protein synthesis and this suggests that there may be a mitochondrion-nucleus signaling pathway involved. In order to analyze these findings further, I ran this dataset through GSEA with the subcellular localization gene sets. Between untreated and treated parasites, the mitochondrion and apicoplast gene sets were not statistically significantly enriched (using a cutoff of FWER = 0.05). However, if one ignores the FWER q-value of all the localization gene sets, the only two gene sets enriched in treated parasites are the mitochondrion and apicoplast gene sets (Table 1a). Although the results of these gene sets should be treated with caution due to the high probability of them being false positive findings, it is a very unlikely coincidence that the two positively-enriched gene sets happened to be the two localizations previously found to be linked to thiostrepton treated parasites in other studies.

Table: Gene sets enriched in phenotype treated (3 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	APICOPLAST	Details ...	28	0.35	1.10	0.334	0.657	0.771	1092	tags=25%, list=8%, signal=27%
2	MITOCHONDRION	Details ...	30	0.36	1.09	0.329	0.334	0.776	1510	tags=33%, list=10%, signal=37%

Table 1a (top). GSEA output table showing the two gene sets with positive normalized enrichment scores (NES) for thiostrepton treatment: apicoplast and mitochondrion. Although these positive NES values support published evidence of a correlation between the apicoplast and mitochondrion with thiostrepton responses in *P. falciparum*, the FWER q-value in both gene sets is too high to be considered statistically significant data.

Table 1b (bottom). GSEA output table showing the remaining seven localization gene sets with negative NES values for thiostrepton treatment. Only two of these gene sets have high enough FWER q-values to be considered statistically insignificant. Since the majority of the localization gene sets have negative enrichment, it is likely that the abnormal cell cycle progression of some of the thiostrepton treated parasites caused a genome-wide decrease in gene expression for the treated parasites.

Table: Gene sets enriched in phenotype control (3 samples) [plain text format]

	GS follow link to MSigDB	GS DETAILS	SIZE	ES	NES	NOM p-val	FDR q-val	FWER p-val	RANK AT MAX	LEADING EDGE
1	APICAL	Details ...	26	-0.96	-2.59	0.000	0.000	0.000	528	tags=92%, list=4%, signal=96%
2	PV	Details ...	20	-0.81	-2.12	0.000	0.000	0.000	793	tags=60%, list=5%, signal=63%
3	CYTOPLASM	Details ...	67	-0.60	-1.95	0.000	0.001	0.003	1745	tags=46%, list=12%, signal=52%
4	NUCLEUS	Details ...	48	-0.56	-1.74	0.011	0.015	0.044	1759	tags=48%, list=12%, signal=54%
5	EXPORTED	Details ...	31	-0.61	-1.74	0.011	0.012	0.044	758	tags=35%, list=5%, signal=37%
6	PPM	Details ...	18	-0.66	-1.69	0.033	0.017	0.070	1865	tags=44%, list=13%, signal=51%
7	FOOD_VACUOLE	Details ...	18	-0.45	-1.12	0.329	0.315	0.801	1705	tags=39%, list=12%, signal=44%

Of the other seven localization gene sets run through GSEA, five were enriched in the control versus the treated phenotype and the other two had statistically insignificant q-values but were also enriched in the control phenotype (Table 1b). This genome-wide enrichment matched the findings from Tarr et al., as they found that only a small proportion of the differentially expressed genes from their study came from the mitochondrion and apicoplast-related genes they were analyzing, but rather from various different localizations. A likely reason why the treated parasites exhibited so much enrichment in the control versus the treated phenotype may have to do with the effects thiostrepton has on slowing *P. falciparum* progression throughout the intraerythrocytic cell cycle.

Since high doses of thiostrepton prevent *P. falciparum* from progressing past the trophozoite stage [19], the parasites were treated with lower non-lethal doses of thiostrepton. However, as mentioned earlier, some of the parasites did not progress past the first intraerythrocytic cycle [14]. As a result, the parasites with delayed cell cycle progression most likely affected the GSEA results of the parasites undergoing normal cell cycle progression by exhibiting lower gene expression. Assuming that delayed cell cycle progression does in fact correspond with a decrease in genome-wide gene expression, it is not surprising that the majority of the localization gene sets showed enrichment in the control parasites when treated with thiostrepton. Moreover, the gene expression data from the parasites with abnormal cell cycle progression may have contributed to the inaccurate FWER q-value of the mitochondrion and apicoplast gene sets' positive enrichment scores.

Using the findings from the peak gene expression GSEA results in combination with the localization GSEA results, I can predict when and where *P. falciparum* exhibits a response to thiostrepton. Assuming that the positive enrichment in the mitochondrion and apicoplast gene sets were not false positive findings I can postulate that *P. falciparum* responds to thiostrepton during the trophozoite stage in biological processes involving the apicoplast and mitochondrion. According to Bozdech et al., during the late trophozoite-early schizont stage, genes are expressed that are related to the components of mitochondrial and/or the apicoplast plasmid translational machinery [6]. This theory goes along with Tarr et al. and their suggestion of mitochondrion-nucleus signaling within the parasite since many of the genes expressed by the apicoplast need to be imported from the nucleus [14,18]. Since the observations from the aforementioned papers match perfectly with my GSEA results, it is very likely that the enrichment in the mitochondrion and apicoplast localizations were not false positive findings.

Conclusion

A large portion of the *Plasmodium falciparum* genome is not fully characterized, creating a lot of gaps in researchers' understanding of how the parasite functions. Although many gene expression data sets have been published to further analyze the function of specific genes within the parasite, few studies attempt to bridge together the individual expressions of genes into broader biological processes. Through GSEA, I gained valuable insight about the drug responses of *P. falciparum* using only two published datasets and several gene set collections. This work shows that GSEA can help researchers gain a deeper understanding of the *P. falciparum* biology from genomic studies. Understanding how antimalarial drugs affect parasites is one of the most important fields in the biology of malaria. In order to prevent the spread of malaria through enhanced drugs, researchers need to understand why malaria has developed resistance to drugs already in existence. Although my GSEA results developed some insight towards the processes that are and aren't responsible for drug resistance in malaria, my findings were limited by the relatively narrow range of biological processes covered by my gene sets and datasets, which could be expanded on in future. By running GSEA with more data in the future, it would be possible to enhance the accuracy and breadth of the findings from this paper to make more headway into understanding malaria.

Acknowledgements

The research conducted in this paper was supported by Dr. Kami Kim's lab in the Albert Einstein College of Medicine. I acknowledge the assistance I received from Dr. Kami Kim and her research associates, Dr. Natalie Silmon de Monerri and Dr. Inessa Gendlina in conducting the research in this paper. More specifically, I received guidance from Natalie Silmon de Monerri in developing the procedure for producing the gene sets and researching the data sets needed for GSEA, and I received guidance from Inessa Gendlina on how to make my data compatible to be run through GSEA. I would also like to acknowledge the work done by Croken et al. [4] as their approach towards analyzing gene expression data with GSEA served as a model for this research project.

References

1. "Malaria." *World Health Organization*. World Health Organization, 2016. Web. 06 Nov. 2016.
2. "Malaria." *Centers for Disease Control and Prevention*. Centers for Disease Control and Prevention, 01 Nov. 2016. Web. 06 Nov. 2016.
3. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102:15545–15550. doi: 10.1073/pnas.0506580102.
4. Croken MM, Qiu W, White MW, Kim K (2014) Gene Set Enrichment Analysis (GSEA) of Toxoplasma gondii expression datasets links cell cycle progression and the bradyzoite developmental program. BMC Genomics 15: 515 doi: 10.1186/1471-2164-15-515.
5. PlasmoDB: a functional genomic database for malaria parasites. Nucleic Acids Res. 2008 Oct 31. Aurrecoechea C, et al.
6. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, et al. (2003) The Transcriptome of the Intraerythrocytic Developmental Cycle of Plasmodium falciparum. PLoS Biol 1(1): e5. doi: 10.1371/journal.pbio.0000005.
7. Bártfai R, Hoeijmakers WAM, Salcedo-Amaya AM, Smits AH, Janssen-Megens E, et al. (2010) H2A.Z Demarcates Intergenic Regions of the Plasmodium falciparum Epigenome That Are Dynamically Marked by H3K9ac and H3K4me3. PLoS Pathog 6(12): e1001223. doi: 10.1371/journal.ppat.1001223
8. Hoeijmakers, Radboud University Nijmegen, Jan 2015.
9. Young J.A., Fivelman Q.L., Blair P.L., de la Vega P., Le Roch K.G., Zhou Y., Carucci D.J., Baker D.A., Winzeler E.A. The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification. Mol. Biochem. Parasitol. 2005;143:67–79. doi: 10.1016/j.molbiopara.2005.05.007.
10. Lopez-Barragan MJ, Lemieux J, Quinones M, Williamson KC, Molina-Cruz A, Cui K, Barillas-Mury C, Zhao K, Su XZ.Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics* 2011;12:587. doi: 10.1186/1471-2164-12-587.
11. Jiang H, Patel J, Yi M, Mu J, Ding J, Stephens R, et al. Genome-wide compensatory changes accompany drug-selected mutations in the *Plasmodium falciparum* crt gene. PLoS ONE. 2008;3:e2484. doi: 10.1371/journal.pone.0002484.
12. Gligorijevic B, Purdy K, Elliott D, Cooper RA, Roepe PD. STAGE INDEPENDENT CHLOROQUINE RESISTANCE AND CHLOROQUINE TOXICITY REVEALED VIA SPINNING DISK CONFOCAL MICROSCOPY

.*Molecular and biochemical parasitology*. 2008;159(1):7-23. doi:10.1016/j.molbiopara.2007.12.014.
13. Caillard V, Beaute-Lafitte A, Chabaud A, Ginsburg H, Landau I. Stage sensitivity of Plasmodium vinckei petteri to quinine, mefloquine, and pyrimethamine. J Parasitol. 1995;81:295–301.
14. Tarr SJ, Nisbet RE, Howe CJ. 2011. Transcript-level responses of *Plasmodium falciparum* to thiostrepton. Mol. Biochem. Parasitol. 179, 37–41. doi:10.1016/j.molbiopara.2011.05.004.
15. Goodman CD, Su V, McFadden GI. The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. Mol Biochem Parasitol 2007;152(2):181–91.
16. Clough B, Strath M, Preiser P, Denny P, Wilson I. Thiostrepton binds to malarial plastid rRNA. FEBS Lett 1997;406:123–5.
17. Goodman CD, Su V, McFadden GI. The effects of anti-bacterials on the malaria parasite Plasmodium falciparum. Mol Biochem Parasitol 2007;152(2):181–91. doi: 10.1016/j.molbiopara.2007.01.005.
18. Martin W., Herrmann R. G. 1998 Gene transfer from organelles to the nucleus: how much, what happens, and why? *Plant Physiol.* 118, 9–17. doi:10.1104/pp.118.1.9.