GENOMICS AND COMPUTATIONAL TECHNIQUES FACILITATE UNDERSTANDING

ANTIMALARIAL DRUG MECHANISMS OF ACTION AND MECHANISMS OF RESISTANCE

A Dissertation

Submitted to the Graduate School

of the University of Notre Dame

in Partial Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

by

Katrina A. Button-Simons

__________________________________________

Michael T. Ferdig, Director

Graduate Program in Biological Sciences

Notre Dame, Indiana

July 2017
© Copyright 2017

Katrina A. Button-Simons
Abstract

by

Katrina A. Button-Simons

Malaria mortality has decreased substantially since 2000, largely due to the success of artemisinin combination therapies (ACTs). However, these gains are threatened by the emergence of artemisinin resistance in Southeast Asia. If artemisinin resistance spreads from Southeast Asia to Africa, as has happened for other drugs, malaria deaths due to Plasmodium falciparum would increase dramatically. There is an urgent need to develop new classes of effective antimalarial drugs and to extend the effectiveness of ACTs.

Historically, ACTs contain partner drugs with a long metabolic half-life to counteract the rapid metabolism of artemisinin. Ideally, partner drugs in combination therapies should act synergistically, targeting different pathways. This requires knowledge of drug mechanism of action (MoA) and mechanism of resistance (MoR), which may not be related to MoA. Unfortunately, MoA and MoR are difficult to determine for large numbers of drugs using standard methods. Understanding MoA and
MoR for artemisinin and candidate drugs will enhance our ability to rationally identify optimal ACT partner drugs to overcome existing resistance and to buy time to allow new drugs to progress through the drug development pipeline.

Transcription profiling and chemogenomic profiling are ideal for quickly evaluating MoA of many compounds. Chemogenomic profiling is a powerful tool for determining drug MoA by comparing profiles of susceptibilities to a panel of drugs in a collection of mutant lines representing a wide range of disrupted genes. In our studies, artemisinin functional activity was linked to signal transduction and cell cycle regulation pathways. Transcription profiling of *Plasmodium* parasites perturbed briefly with sub-lethal concentrations of drug can be used to generate a drug pathway fingerprint to infer MoA. Transcriptional signatures of artemisinin resistance can be used to predict novel partner drugs for ACTs to treat artemisinin resistant infections.

Artemisinin MoR is not fully understood. Pathway-based networks analysis of genomic variation data and gene expression data from published studies of artemisinin resistance can provide a more complete understanding of the mechanism underlying artemisinin resistance. Understanding MoA for drug candidates, along with a thorough understanding of artemisinin MoR would enhance efforts to design new combination therapies that can circumvent artemisinin resistance.
This work is dedicated to my family. Mom and Dad, thank you for putting up with my never-ending questions and propensity to come home muddy with various interesting creatures. Ricki and Larry, thank you for all the extra opportunities to explore science and for helping me understand difficult concepts. To my husband, Josh; I would not have been able to finish this dissertation without your love and support. To my wonderful children, Matty and Elizabeth; thank you for all the cuddles and snuggles. It was hard to be too disappointed about lab failures and other frustrations with two adorable children to come home too and endless amounts of hugs. Your boundless energy and curiosity is inspiring.
<table>
<thead>
<tr>
<th>Topic</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.1.1 Rewiring Between Resistant and Sensitive Networks</td>
<td>32</td>
</tr>
<tr>
<td>2.3.1.1.1 Construction of Resistant and Sensitive GCNs</td>
<td>32</td>
</tr>
<tr>
<td>2.3.1.1.2 Examining Global Rewiring between Resistant and Sensitive</td>
<td>32</td>
</tr>
<tr>
<td>Networks</td>
<td></td>
</tr>
<tr>
<td>2.3.1.2 A Network View of Genes Associated with Parasite Clearance</td>
<td>33</td>
</tr>
<tr>
<td>Half-life</td>
<td></td>
</tr>
<tr>
<td>2.3.1.2.1 Construction of Full Dataset GCNs</td>
<td>33</td>
</tr>
<tr>
<td>2.3.1.2.2 Machine Learning Used to Select Genes Associated with</td>
<td>34</td>
</tr>
<tr>
<td>Artemisinin Resistance</td>
<td></td>
</tr>
<tr>
<td>2.3.1.2.3 Full GCNs Used as Tools to Interpret Machine Learning</td>
<td>35</td>
</tr>
<tr>
<td>Results</td>
<td></td>
</tr>
<tr>
<td>2.4.1 Network Statistics</td>
<td>36</td>
</tr>
<tr>
<td>2.4.2 Comparison of Resistant and Sensitive Networks</td>
<td>41</td>
</tr>
<tr>
<td>2.4.3 Kelch13 'Rewiring' between Sensitive and Resistant Networks</td>
<td>42</td>
</tr>
<tr>
<td>2.4.4 Global Rewiring between 'Early Stage' Resistant and Sensitive</td>
<td>47</td>
</tr>
<tr>
<td>Networks</td>
<td></td>
</tr>
<tr>
<td>2.4.5 Predicting Parasite Clearance Half-life from Gene Expression</td>
<td>49</td>
</tr>
<tr>
<td>Data Using Machine Learning</td>
<td></td>
</tr>
<tr>
<td>2.4.5.1 Elastic Net Regression</td>
<td>50</td>
</tr>
<tr>
<td>2.4.5.2 Random Forest Regression</td>
<td>51</td>
</tr>
<tr>
<td>2.4.6 Network View of Machine Learning Results</td>
<td>52</td>
</tr>
<tr>
<td>2.4.7 GWAS Pathway Analysis Results</td>
<td>54</td>
</tr>
<tr>
<td>Discussion</td>
<td>55</td>
</tr>
</tbody>
</table>

Chapter 3: Chemogenomics as a Tool to Study Artemisinin Mechanism of Resistance and Mechanism of Action

3.1 Abstract                                                        | 62   |
3.2 Introduction                                                    | 63   |
3.3 Methods                                                        | 66   |
3.3.1 Chemogenomic Screening                                        | 66   |
3.3.2 Determination of Chemotype-Genotype Association by Phenotypic Clustering | 67   |
3.3.3 Construction of Drug-Drug, Gene-Gene and Drug-Gene Networks   | 68   |
3.3.4 Transcription Analysis of Kelch13 Cluster                     | 69   |
3.3.4.1 Parasite Culture                                            | 69   |
3.3.4.2 RNA Extraction and cDNA Synthesis                          | 70   |
3.3.4.3 cDNA Labelling and Hybridization to Exon Microarrays        | 71   |
3.3.4.4 Microarray Data Processing                                  | 71   |
3.3.5 Time-Series Analysis and Stage Adjustment                     | 71   |
3.3.6 Differential Expression                                       | 72   |
3.3.7 Differential Co-Expression                                    | 72   |
### 3.4 Results

- 3.4.1 Discerning Drug MoA ................................................................. 73
- 3.4.2 Discerning Artemisinin MoR ..................................................... 74
- 3.4.3 Transcription Profiling of *Kelch13* Cluster ................................. 79
  - 3.4.3.1 Differential Expression in NF54 across the IDC .................. 81
  - 3.4.3.2 Differential Expression between *piggyBac* Lines and NF54 ... 83
  - 3.4.3.3 Differential Co-Expression between *piggyBac* Lines and NF54 85

- 3.5 Discussion .................................................................................. 88

### Chapter 4: Transcription Profiles Elucidate Antimalarial Mechanism of Action

- 4.1 Abstract ..................................................................................... 93
- 4.2 Introduction .............................................................................. 93
- 4.3 Methods
  - 4.3.1 Parasite Culturing ................................................................. 97
  - 4.3.2 RNA Extraction and cDNA Synthesis ................................... 98
  - 4.3.3 Labeling, Hybridization and Scanning .................................. 99
  - 4.3.4 Microarray Data Processing .................................................. 99
  - 4.3.5 MoA Signatures ................................................................... 100
- 4.4 Results
  - 4.4.1 Normalization Panel ............................................................. 101
  - 4.4.2 Validation of MoA Profiles .................................................... 105
- 4.5 Discussion ................................................................................ 106

### Chapter 5: Determining Artemisinin Mechanism of Resistance and Choosing Partner Drugs Based on Mechanism of Action

- 5.1 Abstract ..................................................................................... 110
- 5.2 Introduction .............................................................................. 110
- 5.3 Methods
  - 5.3.1 Microarray Data Processing ................................................. 112
  - 5.3.2 Time-Series Analysis and Stage Adjustment ........................... 113
  - 5.3.3 MoR Signatures ................................................................. 114
  - 5.3.4 MoR Signatures Compared to MoA Signatures to Identify Potential Synergies between Drugs .......................... 114
- 5.4 Results
  - 5.4.1 MoR Results ....................................................................... 115
  - 5.4.2 Comparison of MoR and MoA to Predict Drug Synergies ........ 121
- 5.5 Discussion .............................................................................. 121

### Chapter 6: Summary ........................................................................ 125
Appendix A: Additional Tables ................................................................. 129

Bibliography ............................................................................................ 131
FIGURES

Figure 2.1: Node and Edge overlap between the largest connected components between networks (A) Venn diagrams showing overlap between LCC in ‘early stage’ resistant and sensitive networks. (B) Venn diagrams showing overlap between LCC in ‘late stage’ resistant and sensitive networks. (C) Venn diagrams showing overlap between LCC in ‘all stage’ resistant and sensitive networks. (D and E) Venn diagram showing the overlap for the LCC between ‘early stage’, ‘late stage’ and ‘all stage’ full networks.

Figure 2.2: Overlap between largest cluster in the resistant and sensitive ‘early stage’ network. The three largest clusters in the resistant and sensitive network are shown along with the amount of overlap between the corresponding clusters form the networks and enriched biological processes. Cluster 1 and 2 are of similar size have high overlap and share enriched functions. Cluster 3 has less overlap and different enriched functions.

Figure 2.3: Nearest neighbors of kelch13. Shared and rewired neighbors of kelch13 between the resistant and sensitive ‘early stage’ networks are labeled. The color of each node represents the function of the gene. Sensitive and shared neighbors or kelch13 are enriched for cell cycle and resistant neighbors of kelch13 are enriched for processes related to artemisinin MoA and MoR.

Figure 2.4: Sub-network of ‘early stage’ GCN associate with artemisinin resistance. Nodes represent genes and edges represent significant correlations between genes in the full ‘early stage’ dataset. Node color represents pathway information.

Figure 3.1: Clustering of Chemogenomic Signatures for 71 piggyBac lines over a library of 50 drugs. (A) Hierarchical Clustering of Chemogenomic Signatures for 71 piggyBac lines over a library of 50 drugs. Colors in the heatmap represent a piggyBac line’s drug sensitivity relative to NF54 (yellow – sensitive, blue – resistant). (B) Hierarchical Clustering Dendrogram showing clusters of piggyBac mutants ($p \leq 0.05$).

Figure 3.2: Drug-drug and Gene-gene networks. (A) Network of drug correlations, highly significant correlations shown in blue (FDR adj $p$ value < 0.001) and
significant correlation in black (FDR adj p value < 0.05). (B) Network of *piggyBac* correlations, edges represent significant correlations (FDR adj p value < 0.05) with edges denoting clusters of highly connected *piggyBac* lines. ......................... 77

**Figure 4.1:** Examining correlations between MoA profiles and non-specific effects.
Potential drugs for a new normalization panel were chosen from the 31-drug batch based on correlation of their expression profile to the batch average, representing the most non-specific drug effects. Drugs target a variety of biological pathways. Boxes show drugs with highest correlations in each batch, and * indicate drugs used for further study based on correlation and pathway diversity. Drugs from each batch must be used to normalize out batch effects and non-specific drug response across the whole experiment............................ 103

**Figure 4.2:** Exploring optimal normalization panel size. All two, three, four, five and six drug panels were chosen from drugs that had the highest correlations to the batch average in each of the original three batches (5-Fluorouracil, Cerulenin, Chloroquine, Epoximicin, JQ1 and Methotrexate). Three drug panels had the highest correlations to the average across batches.................................................. 104

**Figure 4.3:** Testing possible normalization panels. All possible combinations of drugs for normalization panels are correlated to the batch average. Panels with the highest correlation for each size are shown. Green box shows panels with highest correlation overall, and * indicates that Panel 7 of three drugs has the highest combined correlation rank in the 31-drug batch and 10-drug batch. 5-FU consistently showed up in high ranking panels for both 31- and 10-drug batches. Cerulenin appeared in most highly ranking 10-drug batch panels, while JQ1 appeared in most highly ranked 31-drug batch panels................................. 104

**Figure 5.1:** PCA plot of samples before (A) and after (B) batch correction.................. 115

**Figure 5.2:** Estimating stage for each sample. A) Genome-wide correlations of each sample to 3D7 reference transcriptome. B) Time-point of maximum correlations graphed for the 6 and 24 hpi samples from the same biological replicate and treatment. ................................................................................................. 117

**Figure 5.3:** MoR pathway signatures by time point. Sample grouping affected MoR signatures. A stronger MoR signature was observed for early ring to trophozoite parasites. ................................................................. 120
TABLES

Table 2.1 Gene Co-Expression Network Statistics ................................................................. 39
Table 2.2 Overlap Between Major Clusters in Resistant and Sensitive Networks .......... 45
Table 2.3 Rewiring Network Statistics .............................................................................. 49
Table 2.4 Machine Learning Model R² and MSE Results .................................................. 52
Table 3.1 Sampling Scheme For Replicate PiggyBac Cultures ........................................ 70
Table 3.2 PiggyBac Rewiring Network Statistics ............................................................... 86
Table 4.1 PiggyBac Transposon Insertions and Drugs Targeting Same Pathway .......... 98
Table 4.2 Piggybac Transposon Primarily Perturb Genes in Disrupted Pathways .... 106
Table A.1 Drug Library Information ................................................................................. 129
ACKNOWLEDGMENTS

Many people have been instrumental in helping me complete this work. First, I would like to thank my mentor, Dr. Michael Ferdig. Thank you for your guidance, strong support and allowing me the freedom to forge my own path through my Ph.D. program. In particular, thank you for supporting my efforts to improve my computational skills through courses in computer science and applied math, even though it meant additional coursework.

I would also like to thank my committee members, Dr. Jeanne Romero-Severson, Dr. Michael Pfrender and Dr. Jun Li. You provided invaluable support, with alternative perspectives and technical advice for different aspects of my research projects.

To all the members of the Ferdig lab, both current and former; thank you for providing the training and expertise necessary to complete my research. I would like to thank Lisa Checkley, Richard Pinapati and Dr. Lindsey Turnbull for sharing their lab expertise and skills with malaria culturing and biology. I would like to thank Asako Tan and Dr. Geoffrey Siwo for sharing their computational expertise and knowledge of malaria bioinformatics. I would like to thank Dr. Siwo particularly for his mentorship early in my graduate program. I would also like to give special thanks to Sage Davis and Kate Vendrely, current Ph.D. students, for their work in creating the samples used in
Chapter 5 and Lizzy Heilmann, a former undergraduate student, for her work in producing some of the data used in Chapter 4.

Finally, I am thankful to The Eck Institute of Global Health and The Department of Biological Sciences for fellowship support during my dissertation. I am also thankful for to The Graduate School for providing professional development funds to attend a workshop in statistical genetics.
ABBREVIATIONS

*P. falciparum*  *Plasmodium falciparum*

- ACT  artemisinin combination therapy
- MoR  mechanism of resistance
- MoA  mechanism of action
- *kelch13*  PF3D7_1343700
- SNP  single nucleotide polymorphism
- WHO  World Health Organization
- GDP  gross domestic product
- IRS  indoor residual spraying
- LLIN  long lasting insecticidal bed nets
- malERA  Malaria Eradication Research Agenda

*P. vivax*  *Plasmodium vivax*

- MMV  Medicines for Malaria Ventures
- h  hours
- IC₅₀  inhibitory concentration 50%
- GWAS  genome-wide association studies
- RSA  ring stage assay
- FDR  false discovery rate
- GO  gene ontology
- GSEA  gene set enrichment analysis
- hpi  hours post invasion
- KBPD  BTB/POZ domain of *kelch13*
- DHA  dihydroartemisinin
- CMap  Connectivity Map
- IDC  intra-erythrocytic developmental cycle
- GCN  gene correlation network
- PCA  principal component analysis
- PC  principal component
- LCC  largest connected component
- MAPK  mitogen-activated protein kinase
CHAPTER 1:
INTRODUCTION

1.1 Malaria Control and Eradication Efforts

Malaria is a global health problem that disproportionately affects developing nations. In 2015, there were 212 million malaria infections and 429,000 deaths, with the majority of disease burden (90% of infections and 92% of deaths) in the WHO African region [1]. Malaria mortality is concentrated in young children, who often have much more severe malaria infections. Malaria burden has a negative relationship to gross domestic product (GDP) and growth at the country level. After controlling for other factors such as colonialization history, geographic isolation and tropical location, countries with high malaria burdens had 33% lower incomes than countries without malaria [2].

Through the use of drug treatment and vector control as part of a global effort to control and eliminate malaria, annual malaria mortality has decreased 62% from 2000-2015 [1]. Cumulatively, there were 1.3 billion fewer malaria infections between 2000-2015 and 6.8 million fewer malaria deaths. These reductions in malaria mortality rates have translated to a 1.2 year increase in life expectancy in Africa. The value of this risk
reduction is estimated at 1,810 billion dollars, which equates to 44% of GDP of African Countries [1].

1.2 Malaria Eradication Challenges

Despite the progress being made towards malaria control and elimination there is cause for concern; our main tools to control malaria, insecticides and drugs are failing due to the evolution of insecticide resistance in malaria hosts and drug resistance in malaria parasites. When previous generations of insecticides and drugs failed in the early 1970s malaria rates increased again in many countries and efforts to eradicate malaria in the subtropics were halted [3].

1.2.1 Insecticide Resistance

Malaria is transmitted by Anopheles’ species of mosquitos. Vector control is a major tool for preventing malaria infection. There are two main forms of vector control currently being used as tools to control malaria, indoor residual spraying (IRS) and long lasting insecticidal bed nets (LLIN). There are four classes of insecticides used in vector control: pyrethroids are used in LLIN; organochlorides, carbamates and organophosphates are used primarily in IRS [1]. In Africa, pyrethroid resistance and organochloride resistance are wide-spread with 70% and 65% (respectively) of malaria endemic countries reporting resistance. Resistance to carbamates and organophosphates are less wide-spread (51% and 28%, respectively). Despite these alarming statistics and evidence of the spread of resistance, LLIN and IRS remained effective at preventing infection in a study of five malaria endemic countries in Africa
[4]. Monitoring insecticide resistance and developing new vector control tools is a WHO priority and will be very important in the malaria elimination effort.

1.2.2 Drug Resistance

The malaria parasite species that causes the greatest mortality, *Plasmodium falciparum*, has shown great adaptability and has developed resistance to every drug that has been widely used to treat malaria [5, 6]. Resistance has recently evolved to Artemisinin Combination Therapy (ACT), the mainline treatment for multi-drug resistant malaria [7-11]. Development and spread of resistance to previous widely used antimalarials (chloroquine and sulfadoxine-pyrimethamine) led to large increases in malaria mortality in past decades [12, 13]. If artemisinin resistance spreads through Africa, projections that assume a 30%-70% treatment failure rate indicate that deaths would increase by 110,000-290,000 annually [14].

Recent population genetics studies of parasites isolated from patients in Southeast Asia reveal that artemisinin resistance has emerged independently, multiple times and is spreading throughout this region, a hotspot for emergence of drug resistance [7-9]. In areas where ACT resistance first developed, treatment failure rates are reaching 45% and artemisinin resistant *P. falciparum* is becoming very difficult to treat [15-17]. Furthermore, resistance is developing to the partner drugs used in ACTs as well [10, 11].

The spread of artemisinin resistance and ACT resistance from Southeast Asia to Africa would derail current efforts to decrease malaria mortality that rely of the efficacy
of these drugs [18]. Recent clinical studies have indicated high recrudescence rates in Africa and in South America [19, 20] and severe malaria cases where intravenous artesiminin is ineffective [21]. Mutations associated with resistance from studies of Southeast Asian populations have been observed recently at very low frequency in Africa and South America, but have not yet been linked to resistance [22, 23]. Recently, a single artesiminin resistant isolate of African origin was identified and confirmed through in vitro testing to be resistant and to cluster with other known African isolates [24]. However, studies of ACT efficacy required every two years by the WHO have not yet revealed ACT resistance in any African country. The evolution and spread of artesiminin resistance within Southeast Asia and potential for artesiminin resistance to emerge in Africa or spread to Africa highlights the urgent need for new drug development.

1.3 New Tools for Malaria Eradication

Malaria experts recognize that simply scaling-up current malaria control techniques will not be strong enough to eradicate malaria [25]. The Malaria Eradication Research Agenda (malERA) was established to “identify key knowledge gaps and define the strategies and tools that will result in reducing the basic reproductive rate to less than 1, with the ultimate aim of eradication of the parasites from the human population” [25]. The tools that have been identified as necessary for malaria eradication include vaccines, vector control, diagnostics and drugs. Research is underway to develop these tools.
1.3.1 Vaccines

Creating an effective vaccine for malaria to help reduce malaria incidence and mortality has been a research goal for many years. This problem is especially challenging because the malaria parasite employs many immune evasion tactics. Older children and adults in medium to high malaria transmission areas develop natural immunity that is protective against malaria symptoms and mortality, but does not prevent infection; individuals with natural immunity develop low level, asymptomatic infections [26]. Natural immunity develops over many years and wanes if individuals are no longer exposed to infection [26]. However, the mechanisms that contribute to natural immunity are not well understood. This lack of knowledge regarding natural immunity and the complex immune evasion tactics employed by P. falciparum have made developing a malaria vaccine a daunting challenge.

Previous efforts to develop vaccines have been only marginally successful and there is not yet a vaccine safe and effective enough to be deployed in malaria endemic countries. Phase III clinical trials were completed for the RTS, S/AS01 vaccine in 2015 and showed modest efficacy of 39% after four doses and 18% after three doses [27], the fourth dose falls outside of the normal vaccine schedule making the implementation of the RTS, S/AS01 vaccine difficult. Despite a positive scientific opinion of the vaccine after a review by the European Medicines Agency no regulatory authorities in Africa have approved the vaccine for use [28]. The WHO is beginning a pilot implementation program in three to five sub-Saharan African countries to further evaluate the vaccine. The vaccine will be administered in three doses to infants 5-9 months of age with a
fourth dose 15-18 months later. Based on the results of this study, the RTS, S/AS01 vaccine may be employed alongside current malaria control tools.

For vaccines to be effective tools in malaria eradication they will need to be effective at blocking transmission and will need to target other malaria species, particularly *Plasmodium vivax*, which causes the greatest malaria morbidity outside of sub-saharan Africa. Ideally, a new vaccine would both provide clinical protection to *P. falciparum* and *P. vivax* and target either sexual stages of the parasite or pre-erythrocytic stages of the parasite. A vaccine that establishes long lasting immunity would also be very useful for eradication efforts.

Vaccinating malaria naïve individuals with attenuated *P. falciparum* or *P. vivax* sporozoites via irradiated mosquito bites has been shown to provide sterilizing cross strain immunity that persists for at least 42 weeks [29, 30]. Intravenous administration of cryopreserved, attenuated sporozoites has also been shown to achieve long lasting protective cross strain immunity [31, 32]. Previous studies utilizing intramuscular and subcutaneous administration showed that the cryopreserved, attenuated sporozoites were safe and well tolerated but failed to elicit protective immunity [33]. Three vaccines that use different strategies to attenuate sporozoites (irradiation, chemoattenuation and gene knockout) are currently in clinical trials [32]. Intravenous administration of these vaccines is a barrier to their widespread use and direct venous inoculation has been developed as an alternative that may help facilitate their widespread use while still administering sporozoites directly to the blood [34]. If approved by regulatory agencies,
these vaccines could be an important new tool in the malaria eradication tool kit, providing long lasting, sterilizing immunity to pre-erythrocytic stages of the parasite.

1.3.2 Diagnostics

In high transmission areas, most of the population (excluding young children) has developed non-sterilizing immunity to malaria and tend to have asymptomatic malaria infections [26]. As malaria eradication efforts succeed and malaria incidence decreases there will be a lag before malaria immunity wanes and there will be many asymptomatic infections that will fuel malaria transmission [35]. To continue making progress towards local elimination of malaria, it is necessary to have robust, sensitive and specific methods to detect active malaria infections [36]. Detection, diagnosis and treatment of malaria cases can both reduce mortality and transmission. Current diagnostic techniques that are used in the field include light microscopy and rapid diagnostic tests. Both of these techniques are used to confirm parasitemia in symptomatic infections and can reliably detect parasites at 100-200 parasites/µL [37]. These techniques will both miss many asymptomatic infections and will not be sensitive enough for active case identification and surveillance in pre-elimination and elimination settings where parasitemia can be as low as 1 parasite/µL [36]. PCR based molecular methods are the gold standard and can detect parasitemia as low as 0.5 parasite/µL of blood but require well-equipped labs and technicians with a high level of training [36]. Research and development efforts for novel diagnostic methods are focusing on methods with higher
sensitivity and specificity. One promising new possible method seeks to develop a transcutaneous method to detect hemozoin crystals in blood [36].

1.3.3 Vector Control

One main goal of vector control research is the discovery and development of insecticides with novel modes of action to replace current insecticides in ITN and IRS [38]. However, ITN and IRS are only effective for *Anopheles* species where the adult female rests or feeds indoors. Research on the ecology and behavior of malaria vectors, especially those that rest and/or feed outdoors is necessary to develop effective vector control tools. New tools should target previously neglected areas of the vector life-cycle, such as oviposition site preference, mating behavior and sugar meal secretion. Genetic control techniques to permanently decrease the vectorial capacity of the mosquitoes in high transmission areas is one promising new control measure [39]. Odorant based spatial repellants and attractant traps have been successful in controlling agricultural insect pests and could potentially be very useful as control measures for mosquito species [40].

1.3.4 Drug Development

With ACTs failing in Southeast Asia and treatment failure rates creeping up world-wide we have a small window of opportunity to act to stay on track with controlling and eradicating malaria. If artemisinin resistance spreads to Africa or develops in Africa, mortality rates are projected to increase between 30 and 100% [14]. There remains an urgent need to develop new antimalarial drugs along with the short-
term need for effective combination therapies featuring existing drugs and/or repurposed drugs to continue malaria control and mortality reduction efforts while long term drug development goals are pursued. Long-term drug development efforts are focused on malaria eradication that prioritizes different goals. Drugs are needed that are appropriate for mass administration to cure asymptomatic infections and would be effective against all parasite stages including relapsing liver stages (radical cure) and prevent transmission. To have maximal impact toward malaria eradication, a drug would be a single dose, radical cure with a novel mechanism of action that is effective against all malaria species that infect humans [41].

Medicines for Malaria Ventures (MMV) has spearheaded a public private partnership for antimalarial drug discovery. To date, most research has focused on compounds with efficacy against blood stage parasites as this parasite stage was the easiest to study. High-throughput screens of more than 6 million compounds have identified thousands of compounds with sub-nanomolar blood stage antimalarial activity; however more research is needed to prioritize compounds for advancement into pre-clinical and clinical trials [42, 43]. This effort has culminated in the release of an open source platform for antimalarial drug discovery (Malaria Box) consisting of 400 of these compounds with diverse scaffolds that are freely available to the malaria research community. Additionally, compounds have been developed against very specific targets, such as Plasmodium falciparum dihydroorotate dehydrogenase [44, 45]. The malaria drug development pipeline now contains more than a dozen new drugs and several new combinations of approved drugs are in various stages of pre-clinical and clinical trials
The drugs in later stages of the development pipeline are new ACTs. There is a promising new synthetic artemisinin compound (OZ439) in phase II trials; this drug has significantly longer terminal half-life (46-62 h) that may be able to counteract current artemisinin resistance [47].

Drug discovery priorities are now expanding to include liver stage parasites (specifically hypnozoites) and sexual stage parasites. Drug discovery in these areas requires development of new assays specific to these stages of the parasite life cycle. Research is ongoing to develop efficient, high-throughput screening methods for activity against sexual stage parasites and liver stage parasites, especially hypnozoites [50-52].

Although considerable progress has been made in developing a high-throughput screen for gametocyte activity, there is not yet an established screening method able to screen compounds at the same scale as the blood stage screens [53]. Several promising new assays have been developed based on flow cytometry [50, 51], image based expression analysis of functional gametocyte proteins [52]. So far, the largest of these screens is a SYBR based IC50 assay used to screen current antimalarials, the 400 malaria box compounds and a library of FDA approved drugs from Johns Hopkins [54]. Of the current antimalarials, only artemisinins and 8-aminoquinolines have activity against gametocytes [55]. Several malaria box compound and FDA approved drugs for other indications show strong gametocytocidal activity [54].

Current screening methods for liver stage activity against Plasmodium vivax is based on a string of surrogate assays involving rodent and monkey models. Each sequential assay is used as a filter to narrow down compounds entering the next assay
in the series. Compounds with strong blood stage activity and screened for activity against hypnozoites of *P. berghei* and *P. yoelii* infection in human hepatoma cell lines [56]. Compounds that show activity are then screened in an assay using the most closely related *Plasmodium* species to *P. vivax, P. cynomolgi* infections in rhesus monkeys. Compounds that prevent relapse for 100 days in *P. cynomolgi* infected rhesus monkeys are evaluated with a full preclinical safety panel and then may be cleared for human clinical trials. Replacing these assays with *P. vivax* infection in human hepatocytes would be a huge screening improvement, however this would require advances in *P. vivax* sporozoite production and stable hepatocyte lines that can be infected with *P. vivax* or reliable access to primary human hepatocytes. Progress has been made on both these fronts. Humanized mouse models could potentially offer the ability to create full *P. vivax* infections potentially including all stages of infection. Researchers at the Center for Infectious Disease Research have developed a mouse model with a humanized liver that can be populated with human red blood cells and support a liver stage and blood stage infection with *P. falciparum* [57]. If efforts to replicate this infection using *P. vivax* are successful this would provide an *in vivo* liver stage model to study liver stage infection and hypnozoites and for testing drug efficacy.

1.4 Bridging the Treatment Gap

As outlined above, critical research is being done to develop the next generation tools that will be necessary for malaria eradication. However, it will take time before new vaccines, vector control methods, diagnostic methods and drugs are ready for use.
As our current control methods fail we urgently need new drugs to continue treating malaria infections and prevent a drastic increase in mortality if artemisinin resistance becomes established in Africa. However, it will be several years before novel antimalarial drugs currently in the development pipeline are approved for human use and even longer until we have a drug that provides a radical cure and is suitable for mass drug administrations. We need new treatment options now if we are to avoid a rebound in malaria mortality.

1.5 Current Strategies for Combination Therapies

With no new drugs immediately available to replace artemisinin as the key compound in combination therapies, it is imperative that we understand the mechanism underlying artemisinin resistance and use effective combination therapies to counteract resistance. Artemisinin is very effective at killing parasites, but it has a very short half-life, 40-60 minutes in the human host depending on the artemisinin derivative [58]. The pharmacokinetics of artemisinin make it more vulnerable to the development of drug resistance: artemisinin monotherapy requires 7-10 days to clear malaria infections and non-adherence to treatment regimens is common, a scenario that encourages the evolution of resistance. To protect the efficacy of artemisinin and prevent the development of resistance, artemisinin was paired with partner drugs in ACTs [59]. Traditionally, partner drugs in ACTs have been antimalarials with complimentary pharmacokinetics to artemisinin, particularly a long half-life to counteract the short half-life of artemisinin. Ideally, drug combinations would have synergistic effects, but this
requires consideration of both the pharmacodynamics and pharmacokinetics. In the short-term we should develop new ACTs with a partner drug that counteracts artemisinin resistance. In the best-case scenario, new partner drugs can be found by repurposing drugs that are already approved for treating disease in humans. Getting regulatory approval for the repurposing of drugs with a known preclinical, pharmacokinetic, pharmacodynamic, and toxicity profiles into new combinations and treatment regimens is faster than getting approval for new drugs. Such repurposed combination therapies can rapidly enter phase II/III clinical trials, bypassing many stages of drug development. With the development of resistance to ACT partner drugs and increasing treatment failure rates it is vital that effective new therapies are developed as rapidly as possible.

Recently, researchers demonstrated that proteasome inhibitors act synergistically with artemisinin in resistant parasites [60]. Two FDA approved proteasome inhibitors that are used in cancer chemotherapy, Carfilzomib and Bortezomib, have strong synergistic interactions with artemisinin in resistant parasites from Southeast Asia [60]. However, both these drugs are very expensive and have significant side effects so may not be good candidates for partner drugs in ACT treatments. Chemotherapies often have high toxicity and are not good candidates for drug repurposing. There are ongoing efforts to development proteasome inhibitors specific for Plasmodium with fewer safety issues [61]. Additional classes of drugs that act synergistically with artemisinin in resistant parasites would make better candidates for ACTs.
1.6 Studying Drug Resistance

Potential mechanisms of drug resistance can be studied *in vitro* by pressuring drug sensitive strains with increasing concentrations of drug allowing the parasites to evolve resistance over time. The newly resistant strain is then sequenced and compared to the original strain and the same strain maintained under no drug selection. Non-synonymous mutations are identified and tested to determine if they confer resistance. These experiments can be very time consuming, taking months to years for resistance to develop. According to antimalarial drug development experts from St. Jude Children’s Hospital, resistance only develops in ~50% of compounds tested (K. Guy, personal communication, January 21, 2016). A recent study published in the BioRxiv preprint server [62] describes mutations in 102 resistant clones and their susceptible isogenic parent lines to 37 diverse compounds.

The most direct way to study naturally evolved drug resistance is to measure drug resistance level and genotype isolates from the field and then perform genome wide association scans (GWAS) to determine the mutations associated with increased drug resistance. In the parasites isolated directly from patients (i.e. isolates) that have been sampled there are many mutations present, requiring that multiple testing be accounted for, necessitating a large number of samples to reach statistical significance. If drug resistance levels are not available for isolates, but genetic data is available it is possible to scan the genome for evidence of recent selection using $F_{ST}$ or haplotype homozygosity. Areas under recent selection are likely to be involved in drug resistance.
1.7 Current Understanding of Artemisinin Resistance

In the past year, dramatic strides have been made towards a better understanding of emerging artemisinin resistance; however, much remains unknown. Emerging artemisinin resistance is characterized by delayed clearance of the malaria infection as measured by a parasite clearance half-life of greater than five hours. A molecular marker of artemisinin resistance, Pf3D7_1343700, \((\text{kelch13})\) has been identified in studies of Southeast Asian populations [7]. Mutations in \textit{kelch13} explain much of the heritable variation in parasite clearance half-life. However, artemisinin susceptible parasites with a \textit{kelch13} mutation do not gain full resistance as measured by the Ring Stage Assay (RSA) [63]. RSA measures % malaria proliferation \textit{in vitro} after 72 h in a highly synchronized culture that has been exposed to 700nM dihydroartemisinin (DHA) for 6 h at 0-3 h post invasion (hpi) [64]. RSA was developed as an \textit{in-vitro} phenotype in place of traditional IC50 assays, that do not correlate well with parasite clearance half-life [64]. Several additional genes have been identified that often are associated with artemisinin resistance in Southeast Asia [9], however, the role of these genes in resistance and their potential interactions with \textit{kelch13} are not understood. Although the role of \textit{kelch13} mutations in artemisinin resistance has not been fully characterized, there has been progress in understanding the resistance mechanism(s). Kelch13 binds PI3K, PI3K is then ubiquitinated and degraded [65]. In parasites with the most common \textit{kelch13} mutation C580Y, kelch13 does not bind to PI3K and there is less degradation of PI3K. Mbengue et al [65] further showed that artificially elevating levels of PI3K in WT \textit{kelch13} parasites can elevate RSA to levels typical of naturally resistant
parasites. Artemisinin resistance has been associated with increased expression of unfolded protein response pathways involving the major PROSC and TRiC chaperone complexes [66]. Several studies reported that WT kelch13 isolates that show slow clearance in the field [67] and this phenotype has been confirmed through RSA for some WT kelch13 isolates [68]. Consequently, it is clear that kelch13 mutations are an important determinant of artemisinin resistance; however, evidence indicates that artemisinin resistance is multi-factorial and is not well understood.

1.8 Studying Drug Mechanism of Action

There are several methods for identifying MoA, including in vitro drug pressure followed by sequencing, chemogenetic profiling, transcription response signatures, and several proteomic based methods (affinity chromatography, activity based proteomics and protein microarray technologies) [69, 70]. Of these methods, transcription response signatures and chemogenetic profiling offer the greatest potential for success in malaria at a scale necessary for prioritizing the many compounds with strong antimalarial activity.

Traditional drug pressure methods that identify targets through sequencing reinforce the one-drug-one-target paradigm and are not particularly useful for elucidating complex MoA. Efforts to determine artemisinin MoA and resistance by drug pressure and sequencing implicated kelch13 [7]. This study applied discontinuous artemisinin pressure over a five year period and identified only a few SNPs, one of which was subsequently associated with artemisinin resistance in population studies [9] but
did not provide insights into artemisinin MoA, which is complex [71, 72]. Experts in antimalarial drug development report that such pressure-and-sequence techniques identify targets for about 50% of compounds tested.

The proteome of *P. falciparum* is poorly characterized [73] and few studies have attempted to characterize antimalarial MoA directly by studying direct binding to the proteome [71, 72]. The effort required to investigate direct protein binding in an unbiased manner and the poor characterization of the proteome require significant follow-up experimentation and validation. Although this approach is not amenable to high-throughput applications, proteomic techniques would be very useful for drugs that are further along in drug development or to characterize MoA for current antimalarials, especially drugs with complex polypharmacology of drugs like artemisinin. Recent proteomic studies on artemisinin have shown direct interaction with 124 and 69 proteins, of which 25 proteins are found in both studies [71, 72].

Chemogenetic profiling and transcription profiling offer the best options for determining drug MoA for a large number of compounds. These two techniques offer different but complimentary views of drug mechanisms. Chemogenetic profiling has shown promising results in our ongoing collaboration with Dr. John Adams (University of South Florida) and Dr. Dennis Kyle (University of Georgia) [74]. This approach involves building an IC$_{50}$ profile for a library of drugs with diverse MoA across a library containing a spectrum of mutations induced by *piggyBac* transposon insertion on a shared drug sensitive background. MoA can be inferred for a novel drug based on the similarity of that drug’s profile to a drug with a known target. Drugs also can be linked directly to the
piggyBac mutants they affect most; because each mutant line in the piggyBac library has a single piggyBac insert, drug effects can be attributed to specific genes.

Transcription response signatures are the current best option for a high-throughput technique for predicting drug MoA in malaria. Recent work in cancer biology has relied on transcriptional cellular responses to drug perturbations to identify a set of genes and a pattern of expression that typifies a response to perturbation by a particular drug (MoA signature) [75-77]. As part of the Connectivity Project (CMap), many laboratories have collaborated to build an open source dataset for both drug and genetic perturbations (including disease states) for different cell lines. Recently, a flurry of studies utilized this database to repurpose drugs by matching patterns of gene expression for diseases with anti-correlated patterns of expression for drugs [77, 78]. The CMap database also has the potential to identify off-target effects that could provide useful information on safety concerns for drugs much earlier in the drug development process. Researchers have identified transcriptional signals for drug resistant cancers and then identified drugs with anti-correlated patterns of expression that are able to reverse the drug resistance phenotype of the cancer [79]. Previous work in our lab has applied this method to predict drug MoA for artemisinin and other potential new antimalarials [80].

1.9 Artemisinin Mechanism of Action

Artemisinin MoA has been debated in the literature with various targets proposed. These studies have implicated the inhibition of hemozoin crystal formation
[81], oxidation of lipid membranes [82, 83] and damage to the digestive vacuole [84, 85], endoplasmic reticulum [85, 86] and mitochondrial membrane [85, 87] in artemisinin MoA. However, a specific target of artemisinin that accounts for all the observed effects has not been found.

Our current understanding of artemisinin MoA indicates complex polypharmacology. Artemisinin must first be activated through cleavage of the endoperoxide bond, a step thought to be mediated by reduced iron, most likely from heme [88, 89], a byproduct of hemoglobin digestion; there may be non heme sources of reduced iron as well [90]. Activation of artemisinin creates reactive oxygen species that have been shown to directly damage DNA creating double stranded breaks [91]. Once artemisinin radicals have been activated they can modify proteins and unsaturated membrane lipids and heme [92]. Two independent recent proteomic studies of artemisinin mechanism identified 70 and 125 proteins, respectively, that were modified by artemisinin [71, 72]. Twenty-five of these proteins overlapped between the two studies and have diverse functions, indicating the artemisinin has a complex MoA. Indeed, previous unbiased techniques that involved studying accumulation of radio-labeled drug indicated lipid based membrane effects on multiple organelles [81, 82, 84-87].

1.10 Gaps in Knowledge

It is clear that artemisinin covalently modifies a large number of proteins [71, 72] and causes severe cellular damage in the parasite. However, it is not clear whether
artemisinin’s MoA is mediated through a particular target or whether cell death happens as a result of widespread damage due to effects on many proteins and lipids. Artemisinin resistance is also only partially understood. A clear understanding of artemisinin resistance mechanism would strengthen our effort to identify partner drugs based on their MoA by targeting the MoR specifically, thereby reversing artemisinin resistance. A new ACT partner drug that counteracts artemisinin resistance would extend the life of artemisinin, preventing increases in malaria mortality if the resistance spreads to Africa. However, finding an appropriate partner drug for artemisinin requires broad knowledge of potential drugs’ MoA in malaria and new screening methods to determine MoA would be valuable. Broad knowledge of antimalarial MoA is needed for rational pairing of drugs in combination therapies independent of artemisinin; in this case, a useful focus would be on predicting drug synergies.

1.11 Specific Aims

**Aim 1: Employ network-based analysis techniques to identify SNPs and transcription differences associated with the delayed artemisinin clearance phenotype.**

In this aim, previously published datasets are mined using network-based techniques to provide a more complete and nuanced understanding of artemisinin resistance. Network-based GWAS techniques have greater power to detect SNPs associated with drug resistance and to elucidate relationships among associated SNPs. Network-based analysis of gene expression data provides greater power to detect
associations, allows for exploration of non-linear relationships and relationships among genes whose expression is associated with resistance. In *Plasmodium*, network-based analysis has the added benefit of helping determine relationships among genes, many of which are still unannotated.

**Aim 2:** Mine *P. falciparum piggyBac* mutant chemogenomic profiles spanning a broad range of drugs and small molecules to identify features of artemisinin mechanism of action and artemisinin resistance.

Determining drug MoA is difficult for large numbers of drugs. Chemogenomic profiling has been developed as a means of determining drug MoA for many compounds simultaneously and has been used successfully in other systems. In this aim, we show the utility of chemogenomics in *Plasmodium falciparum*. We also utilize piggyBac mutants to explore the function of several genes whose expression is disrupted in the mutants, including *kelch13*.

**Aim 3:** Optimize and validate a semi high-throughput method for predicting small molecule MoA in *Plasmodium falciparum* using transcription profiles.

In a previous publication, we have shown that transcription profiling is a useful means of determining drug MoA in *Plasmodium falciparum*. This aim presents additional data to validation transcription profiling and optimize transcription profiling for determining drug MoA for novel drugs.

**Aim 4:** Expand the MoA methodology to define artemisinin mechanism of action and mechanism of resistance and to predict new partner drugs that can reverse current artemisinin resistance.
In this aim, transcription profiling of artemisinin resistant and sensitive isolates is used to generate a signature of artemisinin resistance. This signature of resistance can be compared drug MoA signatures generated in Aim 3 to find partner drugs capable of reversing artemisinin resistance.
CHAPTER 2:
STUDYING ARTEMISININ RESISTANCE USING NETWORKS TECHNIQUES

2.1 Abstract

Artemisinin resistance has emerged and spread in Southeast Asia over the past decade. Long-term laboratory-based in vitro drug selections, paired with large scale genomic studies of recent isolates from Southeast Asia, have identified the main genetic feature of artemisinin resistance to be one of several possible mutations focused (mostly) in the propeller and BTB/POZ domains of PF3D7_1343700 (kelch13). Several additional mutations in other genes have also been associated with artemisinin resistance in Southeast Asia. Surveillance in Africa and South America for emerging artemisinin resistance is underway. Parasite recrudescence and day-3 parasitemia has been observed following ACT treatment and SNPs in kelch13 have also been observed; however, definitive evidence of artemisinin resistance in Africa or South America does not exist. Spread of artemisinin resistance to Africa would have devastating consequences for malaria mortality. Despite identifying kelch13 as a major determinant of artemisinin resistance, a molecular understanding of the mechanism of resistance is lacking. Such detailed mechanistic understanding of artemisinin resistance would enhance the identification of optimal partner drugs for counteracting resistance and
restoring artemisinin efficacy while new drugs progress through clinical trials. Network analysis of both GWAS results and transcriptional data have contributed to a better understanding of the molecular mechanisms underlying phenotypes in various organisms and may be useful in building a better understanding of artemisinin resistance. Our network analysis identifies a difference in gene co-expression relationships for kelch13 in resistant vs. sensitive parasites. In sensitive parasites, the kelch13 module is enriched for cell cycle progression and cell redox homeostasis genes, perhaps indicative of its natural function, while in resistant parasites, the kelch13 module is enriched for cell redox homeostasis, protein folding, lipid biosynthesis and metabolism, response to heat, response to stress and response to stimulus. Many of the enriched functions in artemisinin resistant parasites have been implicated in artemisinin resistance and/or artemisinin mechanism of action in various studies. The network analysis presented here provides deeper insights to artemisinin resistance mechanism than traditional analyses that rely on differential expression comparisons.

2.2 Introduction

Resistance to artemisinin has developed in Southeast Asia; P. falciparum infections in areas of Southeast Asia with resistant parasites are becoming increasingly difficult to treat and are recrudescing at a higher rate [15-17]. Evolution of antimalarial resistance in Southeast Asia and subsequent spread to Africa and other regions of the world has been a prominent phenomenon that led to large increases in malaria mortality. If artemisinin resistance spreads to Africa, it is projected to increase mortality
in Africa by 110,000-290,000 deaths annually, depending on the severity of treatment failure rates [14]. Active surveillance for artemisinin resistance in Africa and South America has not yet identified that resistance has spread from Southeast Asia.

A molecular marker gene (PF3D7_1343700; \textit{kelch13}) of artemisinin resistance has more than 20 different coding SNPs, largely focused in the propeller and BTB/POZ domains (KBPD), that have been associated with artemisinin resistance [7, 67]. Introducing several of these point mutations into sensitive genetics backgrounds can confer resistance in the laboratory [63, 93]. \textit{Kelch13} is a major determinant of artemisinin resistance, however, the genetic architecture is more complex. Sensitive lab lines (3D7 and Dd2), genetically modified using Zinc finger endonucleases to contain KBPD SNPs, do not exhibit the same level of resistance as natural isolates with same KBPD SNPs [63]; genetically modified sensitive isolates from Southeast Asia have generally higher levels of resistance. Artemisinin resistant isolates, defined by parasites clearance half-life of > 5 h or parasitemia detected by microscopy on day-3 post-treatment, have been identified that lack a \textit{kelch13} coding mutation [67] and have confirmed \textit{in vitro} resistance phenotypes by RSA [68]. SNPs in other genes have been associated with artemisinin resistance in Southeast Asia [9]. It has been argued that distinct genetic backgrounds between Southeast Asia and African parasites could be limiting the spread of resistance to Africa because these additional associated SNPs could be compensatory mutations and necessary for artemisinin resistance to spread. Although, as outlined below, selective pressures and out crossing are different between Africa and Southeast Asia as well [94].
Surveillance in Africa and South America has not revealed canonical artemisinin resistance. High recrudescence rates post ACT treatment have been observed in Africa [19] and in South America [20], as are mutations in KBPD [22, 23, 95-97]; however, data are sparse and have not yet linked \textit{kelch13} genotype to artemisinin resistance. KBPD SNPs associated with artemisinin resistance in Southeast Asia occur at very low allele frequencies in Africa [22, 98] and South America [23]. A recent report from Rwanda suggest that allele frequencies of KBPD mutations rose from 0\% in 2010 to 2.5\% in 2014 and 4.5\% in 2015, although this suggestive trend is not significant with only three data points ($p = 0.07$). These mutations appear to have evolved locally and it is not known whether they are associated with resistance in Africa and South America because phenotypic data is lacking for these populations. In Kenya, recrudescent infections lack KBPD mutations [99]. In Mali, documented infections with KBPD mutations have been able to clear parasites from the patient rapidly following artemisinin treatment [67, 100]. In 2013, the first confirmed artemisinin resistant strain of African origin was recovered from a patient in China. The patient had recently returned from Guinea and was treated with a course of DHA-piperaquine under direct observation; the patient had positive parasitemia on day 3 and the parasite strain had a RSA survival rate of 2.29\%, significantly higher (less susceptible) than wildtype \textit{kelch13} strains, but lower than RSA survival rates reported for C580Y resistant mutants. Currently, the WHO recommends that all countries using ACTs as frontline treatment perform surveillance for resistance every two years; countries performing recommended surveillance in Africa have reported no widespread evidence for artemisinin resistance.
Further characterization of genetic variation in *kelch13* in Africa and Southeast Asian populations between 2003-2013, paints a portrait of a gene under intense selection in Southeast Asia but little to no selection in African populations [22]. The Southeast Asian population contains many more non-synonymous KBPD mutations than expected, and many of the KBPD mutations in Southeast Asia result in radical amino acid substitutions. In contrast, the non-synonymous KBPD mutations in Africa fall within the expected range of typical non-synonymous mutation rates across the genome and rarely result in radical changes to the predicted 3D shape of kelch13. This apparent difference in selective pressure is consistent with what is known about artemisinin use and *P. falciparum* epidemiology in the two regions:

- Artemisinin has been used longer and more heavily in Southeast Asia;
- There are more immune individuals in Africa with asymptomatic infections who do not receive treatment, leading to a large reservoir of parasites that are not under selective pressure;
- In Africa, multi-genotype infections are more common and leading to sexual recombination of genetically diverse parasites potentially breaking up combinations of advantageous alleles.

It is clear that additional mutations are associated with artemisinin resistance in Southeast Asia, including epistatic interactions between *kelch13* and other genes [94]. The absence of artemisinin resistance in Africa and South America may indicate that other genes associated with artemisinin resistance are required compensatory mutations for artemisinin resistance to spread. The differences in drug use histories and epidemiology of malaria between the regions also play important roles in the current absence of artemisinin resistance in Africa and South America.
Although genes associated with artemisinin resistance have been identified, the MoR for artemisinin is not well understood. The molecular mechanism of *kelch13* mutations in artemisinin resistance has not been fully characterized, however *kelch13* has been shown to bind to PI3K that is subsequently ubiquitinated and degraded [65]. In parasites with C580Y, one of the most common *kelch13* mutations associated with resistance, *kelch13* does not bind to PI3K resulting in less degradation of PI3K. Mbengue et al [65] further demonstrated that artificially elevating levels of PI3K in *kelch13* WT parasites can elevate RSA to resistant levels. Artemisinin resistance has also been associated with increased expression of unfolded protein response [66]. Other studies of artemisinin MoA indicate a complex mechanism involving multiple targets (polypharmacology) [71, 72]; these observations suggest a complex resistance mechanism. A more detailed understanding of artemisinin MoR will enhance the search for optimal partner drugs for counteracting artemisinin resistance and restoring efficacy while new drugs progress through clinical trials.

2.2.1 Reanalyzing Published Datasets

The published datasets underpinning our current understanding of artemisinin resistance have not been fully mined for information about artemisinin MoR. The largest GWAS study to date for artemisinin resistance contained 1,063 samples and 18,322 SNPs, each tested independently for an association with parasite clearance half-life with population structure included as a random effect [9]. That study identified several background mutations and estimated their effect on parasite clearance half-life while
controlling for the presence of specific KBPD mutations. Although the study considered complex genetic architecture, it did not leverage pathway-based GWAS techniques; these methods offer more statistical power to detect important genes and can contribute to a more robust interpretation of results. The largest-scale transcriptomic study of artemisinin resistance [66] in malaria to date provided valuable insights into artemisinin resistance. However, the statistical techniques used in that study to associate changes in transcription with artemisinin resistance were low-powered. Reanalysis of these datasets using networks-based techniques could provide a more complete and nuanced understanding of artemisinin MoR.

2.2.2 Network Analysis Contributes to Understanding of Molecular Mechanisms Underlying Phenotypes

Pathway-based GWAS has improved our mechanistic-level understanding of many different complex human diseases, including Crohn’s disease [101], lupus, rheumatoid arthritis [102], coronary artery disease [103], type II diabetes [104], and schizophrenia [105]. Pathway-based GWAS can test for associations between a phenotype of interest and pathways of functionally related genes. The simplest pathway-based GWAS techniques require only published $p$ values, while more complex methods require analysis of the raw data (phenotypes and SNPs). These techniques improve power to detect small to moderate effects and increase the repeatability of results between GWAS experiments [106].
Networks-based analysis of transcription data through gene co-expression networks (GCNs) paired with phenotypic data analysis often improves the understanding of molecular mechanisms underlying a phenotype [107, 108]. In an example from malaria drug resistance, differences identified in transcription networks between chloroquine resistant and sensitive parasites improved our understanding of that form of drug resistance [109]. Combining network analysis of -omics and machine learning algorithms can improve predictive models of phenotypes in genetic crosses [110]. These networks-based techniques may be helpful in better understanding artemisinin resistance.

2.3 Methods

2.3.1 Network-Based Re-Analysis of Gene Expression Dataset

Published malaria transcriptomes of 1043 clinical samples from Southeast Asia were obtained from NCBI and mapped to relevant metadata such as parasite clearance half-life, geographic location of sample and relevant SNP information published in the supplemental analysis [66]. Hierarchical clustering of these 1043 transcription profiles identified three main clusters that corresponded to stage of the intra-erythrocytic developmental cycle (IDC) [66], as determined by comparison with the reference strain, 3D7, transcriptome [111]. Of these samples, 549 samples were composed primarily of early-stage parasites (early rings) and 272 were composed of late-stage parasites (late ring and trophozoites), as determined by comparison with the reference strain, 3D7, IDC.
transcriptome. Since this dataset was derived directly from patients we do not expect to see schizonts (mature parasites) because as infected red blood cells mature (24-32 hpi) they adhere to the endothelial lining of the microvasculature [112] and are removed by the spleen [113]. Isolates from Africa were excluded from the dataset, leaving 375 early-stage parasites and 173 late-stage parasites. Since gene expression is very different for early ring and late ring to trophozoite stage parasites and could potentially confound our results, we performed three separate analyses based on stage; the 375 early-stage parasites were analyzed together in an ‘early stage’ group and the 173 late-stage parasites were analyzed in a ‘late stage’ group, in addition, stage was ignored and all samples grouped together in an ‘all stage’ group.

Gene correlation networks (GCNs) were constructed and analyzed in two complementary ways. In the first analysis, samples were split into groups based on artemisinin resistance phenotype and a separate GCN was constructed for each phenotype. Genes that have differential co-expression (‘rewired’) between the two networks were identified and compared. The role of kelch13 in each network was evaluated as well. In the second analysis, a single GCN was constructed using all samples within a group. Machine learning techniques were used to predict genes whose expression patterns are associated with artemisinin resistance. The GCN was then analyzed to determine how the genes identified as predictive of artemisinin resistance via machine learning are connected to one another. This second analysis defines genes whose gene expression patterns are associated with artemisinin resistance and how these genes interact with one another within pathways.
2.3.1.1 Rewiring Between Resistant and Sensitive Networks

2.3.1.1.1 Construction of Resistant and Sensitive GCNs

The ‘early stage’, ‘late stage’ and the combined ‘all stage’ datasets of parasites from Southeast Asia were split into artemisinin-resistant and -sensitive parasites using the cut-off of 5 hour established in the literature, samples with a parasites clearance half-life of > 5 hour were classified as resistant and an equal number of parasites with the lowest parasite clearance half-life values were classified as sensitive. A Pearson correlation matrix was computed for each dataset. GCNs were constructed from the correlation matrix using hard thresholding based on the Scale-free topology criterion [114]. The resistant and sensitive GCNs were compared to determine how the gene co-expression network varied between resistant and sensitive parasites. Node mapping between the two networks was done by plasmoDB gene identifiers and presence and absences of edges were compared. Clusters of genes were identified for each network and the enriched pathways for corresponding clusters were compared between the networks. The nearest neighbors of *kelch13* were compared between the resistant and sensitive networks.

2.3.1.1.2 Examining Global Rewiring between Resistant and Sensitive Networks

A rewiring network was constructed to summarize the edge changes observed between the sensitive vs. resistant networks that contains only significantly rewired edges. This analysis was conducted on the weighted GCNs that contained all edges (no
threshold applied) following the methods described in [115]. Briefly, a Fisher transformation was applied to the Pearson correlation between node $i$ and node $j$, $r_{ij}$

$$F(r_{ij}) = \frac{1}{2} \ln \left( \frac{1 + r_{ij}}{1 - r_{ij}} \right)$$

(1.1)

and the rewiring score for the edge between node $i$ and node $j$ was calculated as

$$rewire_{ij} = P \left( |X| \leq \frac{F(r_{ij}^R) - F(r_{ij}^S)}{\frac{1}{\sqrt{n_R-3}} + \frac{1}{\sqrt{n_S-3}}} \right), X \sim N(0,1)$$

(1.2)

To determine which edges were significantly rewired, $p$ values were corrected for multiple testing using the Benjamini Hochberg method [116]. Genes were then ranked by their degree of rewired connections in the rewiring network and this ranked list was evaluated for enriched pathways.

2.3.1.2 A Network View of Genes Associated with Parasite Clearance Half-life

2.3.1.2.1 Construction of Full Dataset GCNs

In the second method, GCNs were constructed for the ‘early stage’, ‘late stage’ and ‘all stage’ datasets from Southeast Asia using Pearson correlation matrixes as described above. These networks were used to interpret the genes that are associated with artemisinin resistance via the machine learning variable selection methods.
2.3.1.2.2 Machine Learning Used to Select Genes Associated with Artemisinin Resistance

Each of the datasets used to construct the GCNs was randomly split into a training set and test set, with the test set containing 30% of the data. For each of these three datasets, four models were fit; two machine learning models were used with and without SNP information included in each model. Random forest regression or elastic net regression were used as variable selection methods to determine which genes expression patterns were associated with artemisinin resistance. Elastic net regression [117] is a penalized linear regression technique that can select sets of correlated variables that are predictive of artemisinin resistance. Random forest regression [110] is an ensemble tree based method that can select variables with non-linear contributions to artemisinin resistance. SNP information was available for five genes that have been previously associated with artemisinin resistance (kelch13, pfcr, pfmdr, fd, arps10). The association of genes with artemisinin resistance was evaluated with and without SNPs included in the models. For each of the five genes the presence of non wild-type SNPs in were coded as dummy variables and included in the models.

To generate $p$ values for the expression of each gene for each model and dataset, the parasite clearance half-life values were permuted 1000 times to construct a null distribution of either variable importance measures (VIM) for random forest or beta values for elastic net regression. For elastic net regression, the null distribution was used to determine $p$ values for each gene with a non-zero beta value. For random forest regression, a $p$ value significance threshold was set at the 95$^{th}$ percentile of the null
distribution. Genes with expression patterns that were significant predictors of parasite clearance half-life were used to predict parasite clearance half-life on the test set and $R^2$ and MSE were calculated to compare models.

2.3.1.2.3 Full GCNs Used as Tools to Interpret Machine Learning Results

Genes whose expression patterns were significant predictors of parasite clearance half-life in the ‘early stage’, ‘late stage’ and ‘all stage’ datasets were used to select a subset of the corresponding full GCN. This allowed us to determine relationships among genes associated with artemisinin resistance and to define the roles of unannotated genes based on their neighborhoods in the full GCNs. Associated genes for each model were analyzed for pathway enrichment in order to compare results from different analyses.

2.3.2 GWAS Pathway-Based Analysis

Miotto et al [9] published a GWAS study of 1063 samples from 13 sites throughout Southeast Asia. In this dataset, 18,322 SNPs were identified with a minor allele frequency > 0.01. Associations between SNPs and parasite clearance half-life were determined using linear regression, with genetic similarity included as a random effect to correct for confounding due to population structure. The $p$ values for each SNP in this analysis were reported and used and are available for subsequent pathway-based GWAS analysis. The published information contains SNP genomic location and $p$ value. In our analysis marginally significant SNPs were determined using the false discovery rate
(FDR) of Benjamini and Hochburg [116] to adjust the reported $p$ values from the GWAS analysis. Marginally significant of SNPs (FDR adj $p$ value < 0.05) were assigned to the nearest gene based on location. A modified GSEA approach [87] was applied to determine enriched pathways.

2.4 Results

2.4.1 Network Statistics

Network-based analyses have the potential to help provide a more mechanistic understanding of complex phenotypes, such as artesiminin resistance, by relying on connections among genes that is missing from standard transcript level analyses. However, for network analyses to be valuable, the underlying network must be sound. Well-established biological networks are known to have scale-free like distributions and high connectivity [118]. In a best case scenario, we would validate our networks against a gold standard network such as a protein-protein interaction (PPI) network, however, the existing PPI network in *P. falciparum* [73] is too sparse to be useful as a gold standard. Methods for building global PPI networks do not work well in *P. falciparum*, but have produced very small, high confidence subsets of interactions.

Three sets of networks were constructed to account for the confounding developmental-stage structure of the data. *P. falciparum* dominated by the characteristic transcriptional cascade throughout the IDC [111, 119], with most genes having a sinusoidal expression pattern with a single peak of expression in the roughly 48 hour IDC. Parasites at different stages in the IDC can have drastically different
expression patterns. As described above in section 2.3.1, the transcription dataset used in this study had a subset of mostly ‘early stage’ parasites (4-10 hpi) and ‘late stage’ parasites (12-30 hpi). To account for this stage structure, three sets of networks were constructed; an ‘early stage’ set of networks (4-10 hpi), a ‘late stage’ set of networks (12-30 hpi), and a set containing all samples or the ‘all stage’ set of networks.

Table 2.1 outlines and compares key statistics of the different networks, including network size, degree distribution, connectivity and module distribution. Consistent with validated biological networks, each GCN had a scale-free degree distribution, high connectivity and were of similar size. ‘Early stage’ networks contained more nodes than ‘late stage’ and ‘all stage’ networks; the number of edges was variable between networks, but did not show a consistent pattern (Table 2.1). The largest connected component in each network contained 88-95% of the nodes in the network.

The largest connected component of the resistant and sensitive networks were most similar for the ‘early stage’ and ‘all stage’ networks, with 76% of nodes and 40.7% of edges shared in the ‘early stage’ networks and 62.9% of nodes and 36.8% of edges shared in the ‘all stage’ networks (Figure 2.1A and C). Only 50.4% of nodes and 20.7% of edges were conserved between the ‘late stage’ resistant and sensitive networks and 37.6% of nodes and 10% of edges between networks of different stages (Figure 2.1 B and D). Overall, it appears that the ‘early stage’ and ‘all stage’ networks had the most consistent nodes and edges. The resistant networks had the most unique nodes in each stage, while the ‘early stage’ network had the most unique nodes between networks. This observed difference in node and edge overlap between stages is interesting
because resistant and sensitive parasites show differences in progression through the life-cycle. Resistant parasites have a prolonged ring stage which might show us as differential co-expression in resistant and sensitive networks for ‘early stage’ parasites.
<table>
<thead>
<tr>
<th></th>
<th>‘Early Stage’ Networks</th>
<th>‘Late Stage’ Networks</th>
<th>‘All Stage’ Networks</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Full</td>
<td>Resistant</td>
<td>Sensitive</td>
</tr>
<tr>
<td>Number of Samples</td>
<td>375</td>
<td>137</td>
<td>137</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>2990</td>
<td>3275</td>
<td>3029</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>55920</td>
<td>52218</td>
<td>46736</td>
</tr>
<tr>
<td>Connected Components</td>
<td>93</td>
<td>69</td>
<td>90</td>
</tr>
<tr>
<td>Largest Connected Component</td>
<td>2761</td>
<td>3125</td>
<td>2815</td>
</tr>
<tr>
<td>Average Number of Neighbors</td>
<td>37.380</td>
<td>31.889</td>
<td>30.859</td>
</tr>
<tr>
<td>Network Diameter</td>
<td>14</td>
<td>12</td>
<td>15</td>
</tr>
<tr>
<td>Power Law R²</td>
<td>0.864</td>
<td>0.883</td>
<td>0.858</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>20</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>Max Cluster Size</td>
<td>1492</td>
<td>1067</td>
<td>1002</td>
</tr>
<tr>
<td>Median Cluster Size</td>
<td>6.5</td>
<td>12.5</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 2.1: Node and Edge overlap between the largest connected components between networks (A) Venn diagrams showing overlap between LCC in ‘early stage’ resistant and sensitive networks. (B) Venn diagrams showing overlap between LCC in ‘late stage’ resistant and sensitive networks. (C) Venn diagrams showing overlap between LCC in ‘all stage’ resistant and sensitive networks. (D and E) Venn diagram showing the overlap for the LCC between ‘early stage’, ‘late stage’ and ‘all stage’ full networks.
2.4.2 Comparison of Resistant and Sensitive Networks

In a GCN groups of highly connected genes form clusters and genes within a cluster tend to have related functions [120]. Comparing the conservation of clusters between networks and functional enrichment of clusters can provide insights into which parts of the networks are conserved and where the main differences originate. Networks were defined using the Community Clustering (GLay) algorithm in cytoscape ClusterMaker app [121]. Figure 2.2 summarizes the overlap between the main clusters in the ‘early stage’ resistant and sensitive network. The two main clusters in the resistant and sensitive ‘early stage’ networks agree relatively well (59.5% and 63.4% of genes shared between the corresponding clusters), however, there is less overlap between the cluster 3 (35.2% overlap) (Figure 2.2). Cluster 1 and cluster 2 are enriched for the same biological processes in both the resistant and sensitive networks (Figure 2.3). Cluster 3 is enriched for cell-cycle in the sensitive ‘early stage’ network and for regulation of transcription in the resistant ‘early stage’ network. This result is also consistent with resistant and sensitive parasites differential progression through the life cycle.

In the ‘late stage’ clusters there is less overlap between clusters and corresponding clusters do not share enrichments (Table 2.2). There is also considerable overlap between the two main clusters in the ‘all stage’ resistant and sensitive networks with clusters also enriched for similar biological processes (Table 2.2).
2.4.3 *Kelch13* ‘Rewiring’ between Sensitive and Resistant Networks

Examining differential co-expression centered on genes of particular interest to a given phenotype, like *kelch13* for artemisinin resistance or *pfcrt* for chloroquine resistance [109] can help further define the role of the gene in the particular phenotype. Indeed, in examining the clusters and neighbors of *kelch13*, we find the *kelch13* appears to switch roles between sensitive and resistant parasites. *Kelch13* is located in cluster 1 in the ‘early stage’ resistant network and cluster 3 in the ‘early stage’ sensitive network. *kelch13* has 200 nearest neighbors in the resistant network and 157 nearest neighbors in the sensitive network, 51 of those partners are present in both *kelch13* neighborhoods (Figure 2.3). These consistent partners of *kelch13* are enriched for biological processes related to regulation of metabolic processes, and cell redox homeostasis. Enrichment for cell redox homeostasis is consistent with the function of KEAP1 a human gene homolog of kelch13 known to be involved in oxidative stress response [7]. Our data indicate that *kelch13* may switch roles in resistant parasites; in the sensitive network, the nearest neighbors of *kelch13* are enriched for cell cycle progression and cell redox homeostasis genes, while in the resistant network, in addition to the shared cell redox homeostasis, *kelch13* expression partners are enriched for protein folding, lipid biosynthesis and metabolism, response to heat, response to stress and response to stimulus. Artemisinin resistant lines are known to delay cell cycle progression in early ring stage [66, 122]. The enriched pathways in the resistant network neighbors associated with artemisinin resistance (protein folding) or artemisinin mechanism (lipid biosynthesis and metabolism, response to heat, response to stress)
[66, 82-85]. Approximately half of kelch13’s nearest neighbors have unknown function of in both networks, but our expanded modular view can implicate these genes that would be missed in traditional approaches.

*Kelch13* is on the periphery of the resistant and sensitive ‘late stage’ networks and not a member of any cluster in either network. *Kelch13* has only six nearest neighbors in the resistant ‘late stage’ network and three in the sensitive ‘late stage’ network. The neighbors of *kelch13* in the ‘late stage’ resistant network are enriched for biological processes related to RNA processing, double-strand break repair, and protein oligomerization. *Kelch13* is absent entirely from the ‘all stage’ networks. These results are consistent with what is known about *kelch13* expression; its peak expression is near erythrocyte reinvasion (depending on the cell line) and low expression over the period between 12-34 hpi.
Figure 2.2: Overlap between largest cluster in the resistant and sensitive ‘early stage’ network. The three largest clusters in the resistant and sensitive network are shown along with the amount of overlap between the corresponding clusters form the networks and enriched biological processes. Cluster 1 and 2 are of similar size have high overlap and share enriched functions. Cluster 3 has less overlap and different enriched functions.
**TABLE 2.2**

OVERLAP BETWEEN MAJOR CLUSTERS IN RESISTANT AND SENSITIVE NETWORKS

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Size</th>
<th>‘Early stage’</th>
<th>‘Late stage’</th>
<th>‘All stage’</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Resistant Cluster 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>1067</td>
<td>1791</td>
<td>822</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Metabolic Processes</td>
<td>none</td>
<td>Metabolic Processes</td>
<td></td>
</tr>
<tr>
<td><strong>Sensitive Cluster 1</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>1002</td>
<td>1027</td>
<td>771</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Metabolic Processes</td>
<td>none</td>
<td>Metabolic Processes</td>
<td></td>
</tr>
<tr>
<td>Cluster 1 overlap</td>
<td>772</td>
<td>724</td>
<td>449</td>
<td></td>
</tr>
<tr>
<td><strong>Resistant Cluster 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>712</td>
<td>1155</td>
<td>539</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Localization, Transport</td>
<td>Ubiquinone metabolic process</td>
<td>Localization, Transport, Protein modification, RNA modification</td>
<td></td>
</tr>
<tr>
<td><strong>Sensitive Cluster 2</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>651</td>
<td>659</td>
<td>694</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Localization, Transport</td>
<td>Localization, Transport</td>
<td>Protein modification</td>
<td></td>
</tr>
<tr>
<td>Cluster 2 overlap</td>
<td>529</td>
<td>451</td>
<td>413</td>
<td></td>
</tr>
<tr>
<td><strong>Resistant Cluster 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>700</td>
<td>153</td>
<td>676</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Regulation of Transcription</td>
<td>Localization, Transport</td>
<td>Chemical homeostasis, protein folding</td>
<td></td>
</tr>
<tr>
<td><strong>Sensitive Cluster 3</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Size</td>
<td>688</td>
<td>372</td>
<td>59</td>
<td></td>
</tr>
<tr>
<td>Enrichment</td>
<td>Cell Cycle</td>
<td>Cell cycle, proton transport</td>
<td>Translation elongation, nitrate metabolic processes, signaling</td>
<td></td>
</tr>
<tr>
<td>Cluster 3 overlap</td>
<td>361</td>
<td>0</td>
<td>18</td>
<td></td>
</tr>
</tbody>
</table>
Figure 2.3: Nearest neighbors of *kelch13*. Shared and rewired neighbors of *kelch13* between the resistant and sensitive ‘early stage’ networks are labeled. The color of each node represents the function of the gene. Sensitive and shared neighbors or *kelch13* are enriched for cell cycle and resistant neighbors of *kelch13* are enriched for processes related to artemisinin MoA and MoR.
2.4.4 Global Rewiring between ‘Early Stage’ Resistant and Sensitive Networks

Networks are also useful for identifying the broader set of connected genes that may play an important role in generating the phenotype of interest and for clarifying the roles of specific genes. In the previous section, we looked for genes linked to *kelch13* that may play a role in resistance. Now, using an unbiased view, we ask what genes are most highly differentially co-expressed between the resistant and sensitive networks by computing a rewiring score (section 2.3.1.1.2 equation 2) for each edge and only including edges representing significantly differently co-expressed gene pairs in the network.

Table 2.3 reports basic network statistics for the rewired networks for each dataset, including network size, degree distribution, and connectivity. The networks had a scale-free like degree distribution and high connectivity, consistent with known properties of biological networks. Our analysis defined 4950 nodes with 37306 significantly rewired edges (FDR adj *p* value < 0.05) between the ‘early stage’ resistant and sensitive networks. In the ‘early stage’ network the top 2% of rewired genes are enriched for biological processes related to DNA repair, mismatch repair, double-strand break repair, DNA metabolic process, DNA catabolic process, response to stress, protein folding, response to unfolded protein, carbohydrate metabolism and biosynthesis, and peptidyl-histidine phosphorylation. The ‘late stage’ networks contained 4604 nodes with 21265 significantly rewired edges (FDR adj *p* value < 0.05). In the ‘late stage’ network, the top 2% of rewired genes were enriched for translocation of peptides into host cell, vacuolar transport, lysyl-tRNA aminoacylation. In the ‘all stage’ network there were
3236 nodes with 28509 significantly rewired edges (FDR adj p value < 0.05). The top 2% of rewired genes were enriched for biological processes related to entry into host cell, modulation of host cell environment, ion transport, fatty acid biosynthesis, protein transport, cell redox homeostasis and vacuolar acidification.

Protein folding and response to unfolded protein were both enriched in genes that were differentially expressed between artemisinin resistant and sensitive parasites [66]. Fatty acid biosynthesis, protein transport, vacuolar transport, vacuolar acidification, ion transport and cell redox homeostasis may represent responses to damage that artemisinin causes damage to the digestive vacuole, endoplasmic reticulum and mitochondrial membrane through lipid membrane oxidation [82-85]. Activation of artemisinin generates reactive oxygen species that damage DNA [91]; the observation of differential co-expression in DNA repair, mismatch repair, double-strand break repair, and DNA metabolism and catabolism genes may represent an effort to mitigate this DNA damage. Our analysis not only corroborates the role of these processes in artemisinin mechanism or resistance, but also expands our understanding to include contextual linkage between genes in the rewiring network.
### TABLE 2.3

REWIRING NETWORK STATISTICS

<table>
<thead>
<tr>
<th></th>
<th>‘Early stage’</th>
<th>‘Late stage’</th>
<th>‘All stage’</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Samples</td>
<td>137</td>
<td>49</td>
<td>186</td>
</tr>
<tr>
<td>Number of Nodes</td>
<td>4950</td>
<td>4604</td>
<td>3238</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>37306</td>
<td>21265</td>
<td>28509</td>
</tr>
<tr>
<td>Connected Components</td>
<td>4</td>
<td>20</td>
<td>61</td>
</tr>
<tr>
<td>Largest Connected Component</td>
<td>4944</td>
<td>4565</td>
<td>3104</td>
</tr>
<tr>
<td>Average Number of Neighbors</td>
<td>15.065</td>
<td>9.229</td>
<td>17.576</td>
</tr>
<tr>
<td>Network Diameter</td>
<td>9</td>
<td>10</td>
<td>14</td>
</tr>
<tr>
<td>Power Law $R^2$</td>
<td>0.917</td>
<td>0.912</td>
<td>0.868</td>
</tr>
<tr>
<td>Maximum degree</td>
<td>307</td>
<td>343</td>
<td>748</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>20</td>
<td>26</td>
<td>23</td>
</tr>
<tr>
<td>Max Cluster Size</td>
<td>1928</td>
<td>1398</td>
<td>1044</td>
</tr>
<tr>
<td>Median Cluster Size</td>
<td>11</td>
<td>19</td>
<td>8</td>
</tr>
</tbody>
</table>

2.4.5 Predicting Parasite Clearance Half-life from Gene Expression Data Using Machine Learning

Machine learning techniques that are designed for high-dimensional analysis are ideal for analyzing –omics data for associations with phenotypes of interest. In –omics datasets variables are often highly correlated and there are many more variables than samples. These two properties of –omics datasets make applying traditional statistical tests difficult and requires multiple testing correction that often does not account appropriately for the structure of the dataset. Applying traditional statistics with multiple testing correction is underpowered. Machine learning techniques for high-
dimensional analysis are better at accounting for correlations among variables and have higher power to detect associations.

2.4.5.1 Elastic Net Regression

Elastic net regression is a linear regression technique particularly suited for analyzing data with highly correlated predictors. This technique can identify sets of correlated variables that contribute to a phenotype, a useful property when trying to predict artemisinin resistance, a phenotype where multiple genes within a single pathway and multiple pathways are thought to have major contributions.

Fifty-one genes were identified as predictive of artemisinin clearance half-life in ‘early stage’ parasites (FDR adj p < 0.001), explaining 47.5% of the variation of this trait. The ‘early stage’ predictive genes were enriched for regulation of gene expression, regulation of RNA splicing, gene silencing, sulfur compound metabolic and biosynthetic processes, biotin metabolic and biosynthetic processes. In ‘late stage’ parasites, 17 genes were identified as predictive of artemisinin clearance half-life explaining 18.7% of the total variation in parasite clearance half-life. These genes are enriched for lipid catabolic processes and protein palmitoylation. In the ‘all stage’ data, 139 genes were identified as predictive of parasite clearance half-life explaining 42.7% of the variation. These genes are enriched for lipid metabolic processes, regulation of gene expression, regulation of translation, regulation of translational initiation, RNA splicing, and regulation of cellular protein metabolic process. Including SNP information for kelch13, pfcrt, pfmdr, fd, and arps10 into the elastic net models did not improve the model fit ($R^2$
or MSE as shown in Table 2.4). Some of these enriched processes have been previously linked to artemisinin MoA or MoR; however, our approach implicates novel processes.

2.4.5.2 Random Forest Regression

Random forest regression adds the ability to define non-linear relationships between variables and phenotypes. This method, in conjunction with permutation testing can be used as a variable selection method and it is an ensemble method, so usually gives stable results.

In analyzing genome-wide expression for the ‘early stage’ samples, the model incorporating SNP information for *kelch13, pfcr, pfmdr, fd, and arps10* contained 65 genes whose expression patterns were significant predictors of artemisinin resistance (FDR adj \( p \) value < 0.05) and explained 55.2% of the variation in parasite clearance half-life (Table 2.4). When SNPs were not included in the random forest model 234 genes were significant predictors of artemisinin resistance in ‘early stage’ dataset (FDR adj \( p \) value < 0.05) explaining only 22.5% of the phenotypic variation. The ‘late stage’ model with SNP information included contained 52 genes whose expression significantly predicted artemisinin resistance and 78 genes when SNPs were not included. In the ‘all stage’ model there were 142 genes whose expression patterns were significant predictors of artemisinin resistance and 271 when SNPs were excluded from the model. In the ‘early stage’ dataset, the genes whose expression predicted artemisinin resistance were enriched for protein palmitoylation, MAPK cascade, and protein folding. In the ‘all stage’ dataset, significant predictors were enriched for protein folding, DNA replication,
carboxylic acid transport, ATP hydrolysis coupled proton transport, and glycerol metabolic process. The ‘late stage’ predictors were enriched for cell cycle, metabolic processes, lipid metabolism, and sulfur compound metabolic processes. Many of the enriched processes have been previously suggested to be involved in artemisinin MoA or MoR; their implications in some of our other analyses further strengthen our predictions of their important biological roles.

<table>
<thead>
<tr>
<th>TABLE 2.4</th>
<th>MACHINE LEARNING MODEL R² AND MSE RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>‘Early stage’</td>
</tr>
<tr>
<td>Elastic Net Regression with GE R²</td>
<td>0.475</td>
</tr>
<tr>
<td>Elastic Net Regression with GE MSE</td>
<td>4.12</td>
</tr>
<tr>
<td>Elastic Net Regression with GE + SNPs R²</td>
<td>0.293</td>
</tr>
<tr>
<td>Elastic Net Regression with GE + SNPs MSE</td>
<td>3.23</td>
</tr>
<tr>
<td>Random Forest Regression with GE R²</td>
<td>0.2252</td>
</tr>
<tr>
<td>Random Forest Regression with GE MSE</td>
<td>4.03</td>
</tr>
<tr>
<td>Random Forest Regression with GE + SNPs R²</td>
<td>0.552</td>
</tr>
<tr>
<td>Random Forest Regression with GE + SNPs MSE</td>
<td>2.33</td>
</tr>
</tbody>
</table>

2.4.6 Network View of Machine Learning Results

Employing both elastic net regression and random forest models to identify genes whose expression was associated with artemisinin resistance allowed us to explore both linear and non-linear relationships. Pathway enrichment generated from
these lists, i.e. the standard approach, can provide some insight to the pathways important to drug resistance, but do not provide information on how genes interact and the results are limited to annotated genes. The full GCNs define relationships among genes and the genes lists of artemisinin resistant genes can be used to identify sub-networks of the GCN that are associated with artemisinin resistance.

For each dataset (‘early stage’, ‘late stage’, and ‘all stage’) the genes whose expression patterns are predictive of artemisinin resistance by random forest and/or elastic net were merged. The combined list for each dataset was used to select a sub-network of the corresponding full GCN creating a sub-network of genes whose expression patterns were related to artemisinin resistance. This sub-network can then be used to understand relationships between genes involved in artemisinin resistance, beyond the pathway annotations available. Applying a network perspective to the machine learning results for all datasets gave similar results to the pathway analysis of the machine learning results but refined the relationships between associated genes and their function (Figure 2.4).
Figure 2.4: Sub-network of ‘early stage’ GCN associate with artemisinin resistance. Nodes represent genes and edges represent significant correlations between genes in the full ‘early stage’ dataset. Node color represents pathway information.

2.4.7 GWAS Pathway Analysis Results

Pathway-based GWAS can provide exciting mechanistic insights and this functional inclusiveness allows for greater reproducibility between studies because SNPs can converge on a function from different genes. However, the greatest benefits of pathway-based analysis are derived from sophisticated GWAS techniques that require that accessible data includes SNPs matched to specific phenotype information. Many studies provide only \( p \) values, rather than SNPs with matching phenotype information, thereby limiting the ability to apply pathway based GWAS techniques. Our analysis identified 181 marginally significant SNPs (FDR adj \( p \) value < 0.05) that mapped to 143...
genes that are enriched for response to drug and drug transport. Sixty-eight (47.5%) of these genes are annotated as genes of unknown function in *P. falciparum*. Using the nearest neighbors in the rewiring networks for each unannotated gene, we were able to predict gene function. Repeating the pathway analysis using these predicted functions, the genes with marginally significant SNPs are enriched for processes related to protein transport and localization and ribonucleoprotein complex subunit organization. These processes are implicated in our other analyses.

2.5 Discussion

GWAS studies of artemisinin resistance have identified many mutations in a single gene, *kelch13* [7, 9], that have been experimentally confirmed to contribute to resistance [63] as well as additional associated mutations. A causal role for these additional mutations has not be confirmed by laboratory manipulation. A transcriptional study identified response to unfolded protein as up-regulated in artemisinin resistance [66]. Increased PI3K and PI3P are elevated in artemisinin resistant lines, with elevation of PI3P conferring resistance in the absence of *kelch13* mutations [65]. However, we do not yet have a detailed molecular understanding of artemisinin resistance.

Network-based analysis techniques have led to a more mechanistic understanding of phenotypes in malaria [109, 123] and in complex human diseases [101-105]. Employing these same techniques to investigate artemisinin resistance allowed us to show that pathways associated with artemisinin MoA are ‘rewired’ (differentially co-expressed) between artemisinin resistant and sensitive parasites.
Moreover, we were able to assign pathways that have been previously associated with artemisinin MoR and MoA to the natural role of *kelch13* in wild-type parasites and/or the role of *kelch13* in artemisinin resistant parasites.

Previous studies implicated lipid membrane oxidation, inhibition of hemozoin crystal formation, damage to the digestive vacuole, damage to the endoplasmic reticulum and damage to the mitochondrial membrane in artemisinin mechanism of action [82-85]. In our network analysis, we found that genes that were differential co-expressed between resistant and sensitive parasites were enriched for fatty acid biosynthesis, protein transport, vacuolar transport, vacuolar acidification, ion transport and cell redox homeostasis. These biological processes may help to mitigate damage to the endoplasmic reticulum, digestive vacuole and mitochondria imposed through lipid membrane oxidation. DNA repair, mismatch repair, double-strand break repair, and DNA metabolism and catabolism genes are ‘rewired’ between artemisinin resistant and sensitive parasites and may play a role in mitigating DNA damage caused by reactive oxygen species produced by artemisinin activation [91]. Protein folding and response to unfolded protein are also enriched in differentially co-expressed genes in agreement with differential expression analysis of this same data [66]. It is interesting that all these biological processes that relate to the different effects of artemisinin on the parasite are ‘rewired’ in parasites without artemisinin treatment. This argues that the process by which artemisinin kills parasites involves a wide range of effects of artemisinin. Thus, artemisinin resistance could be impacted by a broad ‘rewiring’ in resistant parasites in
response to artemisinin’s complex interactions with many different proteins in the parasite.

In the ‘early stage’ resistant network, kelch13 was a member of different modules and had differences in nearest neighbors than in the ‘early stage’ sensitive network. In contrast, kelch13 was at the periphery of the ‘late stage’ networks and not present in the ‘all stage’ networks. These results are consistent with the stage-specificity of artemisinin resistance. Artemisinin resistance is limited to the early ring developmental stage, with later-stage parasites of artemisinin resistant isolates maintaining artemisinin sensitivity. Thus, differences in artemisinin resistant and sensitive networks, particularly with respect to kelch13, were expected in ring stage networks, but not necessarily in late and mixed stage networks. Kelch13 has 51 consistent partners in both the resistant and sensitive ‘early stage’ networks that are enriched for metabolic processes and cell redox homeostasis. Enrichment for cell redox homeostasis is consistent with the function of KEAP1 a human gene homolog to Kelch13. In the sensitive network, the nearest neighbors of kelch13 are enriched for cell redox homeostasis and cell cycle progression. Artemisinin resistant lines are known to delay cell cycle progression in early ring stage [66, 122]. In the resistant network, the nearest neighbors of kelch13 were enriched for other pathways that have been associated with artemisinin resistance (protein folding) or artemisinin mechanism (lipid biosynthesis and metabolism, response to heat, response to stress) [66, 82, 83].

Microscopic analysis of artemisinin resistant parasites, along with transcription studies, have suggested that artemisinin resistant parasites have altered life cycle
progression early in their IDC; this effect is to remain in a prolonged ring phase (the IDC stage at which they are least susceptible to the effects of artemisinin); this is accompanied by a decreased duration of later stages to maintain the essential ∼48 hour IDC [66, 122, 124]. Our results identify differential co-expression for cell-cycle related genes in the ‘early stage’ in both our comparison of major clusters between networks and in our comparison of the neighborhoods of \textit{kelch13}. We identified no ApiAP2 transcription factor genes in the rewired neighborhood of \textit{kelch13}, however we observed differences in epigenetic factors, particularly those involving histone modification, which has been implicated in liver stage cell cycle arrest of other \textit{Plasmodium} species during hypnozoite (quiescent liver-stage) formation [125].

Applying machine learning techniques that are specific for high-dimensional data revealed additional pathways linked to artemisinin resistance in the transcriptional dataset, including protein folding, protein palmitoylation, carboxylic acid transport, ATP hydrolysis couple proton transport, glycerol metabolic process, lipid metabolism, sulfur compound metabolic processes, MAPK cascade and DNA replication. The published analysis of this same dataset identified genes in involved in response to unfolded proteins [66]. The other biological processes are related to artemisinin MoA and have not previously been linked to artemisinin resistance through differential expression of transcription data.

The variables identified by the machine learning techniques were not consistent between datasets indicating stage-specific differences in artemisinin resistance. Protein folding, protein palmitoylation, and MAPK cascade were enriched in the ‘early stage’
dataset, while cell cycle and metabolic processes were enriched in the ‘late stage’ dataset; the ‘all stage’ results were a combination of the results for the distinct datasets. The ‘all stage’ dataset machine learning models identified a larger set of genes whose expression is associated with parasite clearance half-life in the ‘all stage’ network and fewest for the ‘late stage’ network, consistent with differences in sample sizes. There was very little overlap in the features identified. This was expected because elastic net identifies linear relationships while random forest regression can identify non-linear relationships. For each model, the expression of 1-5% of the genome was associated with artemisinin resistance, with fewer genes identified for the ‘late stage’ dataset, consistent with lower sample sizes in this dataset. More than half of the genes we identified as associated with resistance were not annotated. We predicted pathways based on the predominant function of each genes’ network cluster. This network-based analysis allowed us to include more genes, including unannotated genes, in the pathway analysis rather than ignoring them as current enrichment methods do. We were also able to explore interactions between genes that are associated with artemisinin resistance.

Interestingly, one of the genes identified by elastic net regression as a predictor of artemisinin resistance in the ‘early stage’ dataset is PF3D7_0705500, is a putative inositol-phosphate phosphatase involved in phosphatidylinositol dephosphorylation as part of the PI3K/PI3P pathway. The PI3K/PI3P pathway has previously been implicated in artemisinin resistance [65]. A gene fulfilling a similar role was just identified as showing
a *kelch13* like evolutionary trajectory in longitudinal genomic surveillance study of Southeast Asia [94].

In this study, we utilized two different network techniques. In the first technique, we built GCNs from two subsets of each dataset; artemisinin resistant samples were used to construct an artemisinin resistance GCN and artemisinin sensitive samples were used to construct an artemisinin sensitive GCN and compared differential co-expression between the two networks. The second technique involved building a single GCN using both artemisinin resistant and sensitive samples followed by the application of machine learning algorithms to identify genes whose expression predict artemisinin resistance. Both methods revealed characteristics of artemisinin resistance, however, the differential co-expression analysis provided the most useful insights into artemisinin mechanism because it allows us to distinguish the function of Kelch13 in resistant parasites from the normal function of Kelch13 in sensitive parasites. The second method allowed us to identify genes that have both linear and non-linear associations with artemisinin resistance and to predict function for artemisinin-associated genes that are unannotated and to determine relationships between artemisinin resistance associated genes.

Pathway-based GWAS using only published *p* values of SNPs [9] from previous GWAS studies identified additional pathways enriched for genes that were marginally significant. However, 47.5% of these marginally significant genes were unannotated, limiting functional inference. Mapping these marginally significant genes to our artemisinin resistance network allowed us to better evaluate the potential roles of these
unannotated genes and identify additional pathways that may be involved in mitigating oxidative damage to parasites.

Comparing the results from our network-based reanalysis of published GWAS and transcription datasets we corroborated pathways implicated in resistance and expanded on the results of the original studies and revealed insights not evident from traditional GWAS and differential expression analyses.
CHAPTER 3:
CHEMogenomics as a tool to study artemisinin mechanism of resistance and mechanism of action

3.1 Abstract

Methods for studying drug mechanism of action (MoA) and their possible relationships to mechanisms of resistance (MoR) are cumbersome and expensive and do not scale to hundreds or thousands of drugs. Thus, MoA and MoR are both are not accounted for until very late in the drug development process. Chemogenomics is an emerging discipline that offers an alternative way to link drugs to the genes and pathways they perturb. Chemogenomics involves screening libraries of drugs against a panel of genetically diverse cell lines to measure their distinct drug sensitivity profiles. For my study, the drug library contains drugs with both known and unknown MoA and the panel of parasite cell lines contain unique piggyBac transposon insertions, all in a common drug-sensitive genetic background. The insertion mutants are random, affecting both annotated and unannotated genes. This pilot study, with 50 drugs and 71 piggyBac transposon mutagenesis lines, demonstrates how chemogenomics can be used to screen for MoA in P. falciparum. We show that gene function can be inferred
from piggyBac lines with similar drug sensitivity profiles and that transcription profiling of piggyBac mutants can elucidate the function of the disrupted gene.

3.2 Introduction

Chemogenomics is a relatively new discipline that has developed as a nimble alternative to the traditionally cumbersome, expensive and slow drug discovery process. It involves screening large libraries of drugs using modern genomics methodologies to link drugs with biological activity directly to genes that are perturbed by those drugs [126-128]. The NCI60 program at the NIH is an example of a very successful chemogenomics program that has been used to screen more than 140,000 compounds for activity against a panel of cancer cell lines [129]. In that study, for each of the compounds, an IC₅₀ fingerprint was established for a panel of 60 cancer cell lines; the library of compounds contained both compounds with known and unknown mechanisms of action (MoA). Pattern-matching algorithms were then used to impute MoA for novel compounds with unknown MoA.

The NCI-60 panel represents natural diversity in cancer tumors, however, using artificial panels created by introducing well characterized mutations into an isogenic background using forward genetics is a much more powerful approach because it offers a greater specificity in mapping compounds to precise targets due to the defined single mutation for each cell line; this approach includes the added advantage of being able to infer the function of genes based on how its drug sensitivity profile clusters with other
mutants. In yeast, this approach identified genes in the same pathways with highly correlated chemogenomic profiles [130, 131].

Until now, forward genetics techniques have shown limited success in *P. falciparum*. This has changed with the opportunity of *piggyBac* transposon mutagenesis, which has facilitated the construction of an unbiased mutant panel in an isogenic background [132-134]. Our collaborators generated a *P. falciparum* mutant library consisting of single transposon inserts and have characterized the insertion sites and basic growth phenotypes [132, 133]. Insertion site analysis of the initial *piggyBac* transposon library of 177 mutant lines showed that the *piggyBac* transposon inserts randomly into the *P. falciparum* genome with 59% of the insertions in the 5′UTR of annotated genes, 20% in exons or introns, 10% in the 3′ UTR and 11% in intergenic regions [132].

A panel of 71 *piggyBac* lines was chosen from this transposon library to generate a profile across 50 drugs with diverse mechanisms of action [74]. Each of the 71 *piggyBac* lines carries a single *piggyBac* transposon insertion whose location has been identified through next-gen sequencing [134]. These well characterized mutations in a single isogenic background allow for precise determination of genes in which a mutation uniquely influences the susceptibility to drugs, and can include a direct drug target or a targeted pathway; a compilation of these response profiles across mutants can also lead to better functional annotation of the many unannotated genes in *P. falciparum*. When a drug causes a large shift in IC₅₀ for a particular *piggyBac* mutant line we can map the drug to the gene affected by the *piggyBac* insert. The drug library contains several
artemisinin derivatives, giving us the power to investigate artemisinin function through pattern-matching to drugs with known mechanism and by identifying specific mutant lines with large IC\textsubscript{50} shifts for one or more artemisinin derivatives. Gene function can be inferred for unannotated genes based on the clustering of piggyBac mutant lines, i.e. a ‘guilt by association’ approach. The piggyBac mutant panel includes a piggBac mutant (PB58) with an insertion in the promoter region of kelch13 (a causal gene for artemisinin resistant) that causes dysregulation of kelch13 \cite{74}. Clustering of PB58 may allow inferences about the function of kelch13. This preliminary study establishes a proof-of-concept and suggests that predictions of MoA and gene function will improve, perhaps dramatically, with increased scale on both the drug and gene/mutant axes. Furthermore, these data will integrate well with other large-scale approaches and bioinformatics to guide hypotheses about specific genes (piggyBac lines) with interesting phenotype profiles for targeted follow-up investigations.

Transcriptional profiling across the IDC progression in RBCs of mutant lines can reveal function impact(s) of any gene/mutant \cite{123}. Principal Component Analysis (PCA) of transcription profiles across the \textit{P. falciparum} IDC indicated that parasite developmental stage (as measured in hpi) accounts for the major source of biological variation and is well-captured by the first two principal components. It is well known that the expression pattern of most individual genes in \textit{P. falciparum} vary over the cycle in a sinusoidal pattern with a period equal to the cycle duration of the parasite; amplitude and phase within this expression pattern can vary by gene \cite{111, 119} and in response to perturbation \cite{123}. In the context of this strong stage-dependent gene
transcription pattern, it is surprising that few studies have explicitly accounted for parasite stage in efforts to dissect the impact of specific perturbations (e.g. drugs or mutations) [123]. Here, we transcriptionally profile piggyBac mutants that cluster with genes of particular interest with the goal of identifying and clustering the effects of the piggyBac insertions.

3.3 Methods

3.3.1 Chemogenomic Screening

Chemogenomic data were generated by our collaborators as described in Pradhan et al. [74]. In brief, a library of 50 inhibitors (see Appendix A1 for a list of drugs and pathways), at least two per pathway targeting at different levels Plasmodium metabolism were assayed for growth inhibition (IC$_{50}$) against a piggyBac transposon library of 71 lines. Each piggyBac line contained a single well characterized piggyBac transposon insert. Growth assays were run in batches and each batch contained the wild type NF54 line. The growth inhibition concentration affecting 50% and 90% (GI$_{50}$ and GI$_{90}$) parasite growth was calculated by plotting the relative fluorescence unit values (RFU) of the SYBRGreen I added to culture plates. Relative Phenotype Response (RPR) was calculated for each piggyBac line using as follows,

$$RPR = \frac{GI_{50 B}}{GI_{50 WT}}$$

(3.1)
RPRs are described either as a shift towards resistance (i.e., increased IC$_{50}$ from the wild type IC$_{50}$) or sensitivity (decreased IC$_{50}$ value from the wild type). The RPR for each mutant is scored from the wild type assayed along with the mutants.

3.3.2 Determination of Chemotype-Genotype Association by Phenotypic Clustering

A chemotype-genotype association was made by cluster analysis of RPRs. Clustering was executed using a standard agglomerative algorithm described by Eisen et al. [135]. Genes were clustered by average linkage to calculate minimum distance using a scaled, uncentered Pearson correlation matrix. All data were transformed to logarithmic base 2 and hierarchical clustering results were visualized in R using the heatmap.2 function in the gplots package. To assess statistical significance of clusters, the R package pvclust was used to calculate $p$ values for each branch point in the piggyBac and drug dendrograms, the package was modified to allow a Pearson correlation based distance metric. The output is displayed graphically, conveying the clustering and the essential similarity in RPR of the gene functions in presence of an inhibitor. Increased resistance is a positive shift in IC$_{50}$ and vice-versa for sensitivity. A distinct subset of Gene Ontology (GO) functional annotations relevant to the pathways indicated in the study was used to annotate each gene in the chemogenomic interaction dataset for color coding in cluster analysis. Any gene not falling into the defined category has been designated as ‘other’. The genes falling under multiple annotations, we choose to define it as a most probable on the basis of reviews and similarities published concerning the gene. Information on all genes can be found in the
Plasmodium genome resource (http://plasmodb.org/plasmo/), GeneDB (http://www.genedb.org/Homepage/Pfalciparum) and KEGG GENES Database (http://www.genome.jp/kegg/genes.html).

3.3.3 Construction of Drug-Drug, Gene-Gene and Drug-Gene Networks

Correlation between chemogenomic profiles of various drugs was determined using the Pearson correlation coefficient across clusters of highly correlated piggyBac lines. For each drug pair, a permutation test was conducted, where a random distribution of 1000 Pearson correlation coefficients was obtained by permuting the chemogenomic profile of a given drug pair 1000 times, followed by computation of a correlation coefficient in each instance of permutations. Correlation between any drug pair was regarded as significant if the observed correlation coefficient was greater than any of the 1000 correlation coefficients observed in the permutations of the drug pair’s profile. Drug pairs whose correlation met this criterion were regarded as interacting. Visualization of the drug-drug network interactions was performed in cytoscape [136].

Similarly, a piggyBac gene-gene network was constructed using the Pearson correlation co-efficient of each pair of piggyBac mutants across all drugs. For each drug pair, a permutation test was conducted and a correlation between any piggyBac pair was regarded as significant if was greater than any of the 1000 correlation coefficients observed in the permutations of the drug pair’s profiles. Visualization of the gene-gene network interactions and network clustering was performed in cytoscape [136] using the MCODE algorithm, that identifies clusters of highly interconnected nodes [137].
Drug-gene interactions were considered as significant when a specific mutant exhibited at least three fold change in resistance/sensitivity to a given drug. Drug-gene interactions were determined for the five artemisinin compounds.

3.3.4 Transcription Analysis of Kelch13 Cluster

3.3.4.1 Parasite Culture

*P. falciparum* lines NF54 and seven NF54 derived *piggyBac* lines were cultured using standard methods in human red blood cells (Indiana Regional Blood Center, Indianapolis, Indiana) suspended in complete medium (CM) containing RPMI 1640 with L-glutamine (Invitrogen Corp.), 50 mg/L hypoxanthine (Sigma-Aldrich), 25 mM HEPES (Cal Biochem), 0.5% Albumax II (Invitrogen Corp.), 10 mg/L gentamicin (Invitrogen Corp.) and 0.225% NaHCO₃ (Biosource) at 5% hematocrit. Cultures were grown separately in sealed flasks at 37°C under an atmosphere of 5% CO₂, 5% O₂, and 90% N₂.

For each line three replicate cultures were thawed and synchronized using a 5% sorbital at initially as mid ring, again 8 h later and a third time 40 h later. Samples were taken from the three replicates following the sampling scheme shown in Table 3.1, resulting in 144 samples total.
TABLE 3.1

SAMPLING SCHEME FOR REPLICATE PIGGYBAC CULTURES

<table>
<thead>
<tr>
<th>Replicate</th>
<th>2 hpi</th>
<th>6 hpi</th>
<th>10 hpi</th>
<th>14 hpi</th>
<th>18 hpi</th>
<th>22 hpi</th>
<th>26 hpi</th>
<th>30 hpi</th>
<th>34 hpi</th>
<th>38 hpi</th>
<th>42 hpi</th>
<th>46 hpi</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

3.3.4.2 RNA Extraction and cDNA Synthesis

Total RNA was extracted from flash frozen 10 mL pellets of culture using TriZol reagent (Invitrogen, Carlsbad, CA) as described previously [52]. Quality and quantity of RNA was determined using Nanodrop (NanoDrop Technologies). 300ng of RNA was used as starting material for cDNA synthesis using the Sigma WTA2 whole transcriptome amplification kit (Sigma Aldrich, St Louis, MO). The cDNA synthesis reaction was performed in two steps: library synthesis and library amplification. To synthesize the cDNA library, 300ng of sample RNA was incubated with reverse transcriptase and non-self-complimentary primers that contain a quasi-random 3’ end and a universal 5’ end. Primer extension was then performed using WTA2 polymerase to generate an OmniPlex cDNA library consisting of random, overlapping 100 to 1000 b fragments flanked by a universal end sequence. Amplification was then performed using primers targeting the universal 5’ ends. cDNA cleanup was performed using phenol-chloroform precipitation. All batch and reagent information was recorded in a metadata spreadsheet for further analysis.
3.3.4.3 cDNA Labelling and Hybridization to Exon Microarrays

1µg of cDNA was labeled with Cy3 dye using 65% AT rich pre-labeled random hexamers as primers for cDNA synthesis by Klenow fragment of DNA polymerase I. Hybridizations were performed for 17 h to a custom Agilent microarray [Turnbull, in prep] at 65°C at 10 rpm followed by washing of the arrays as described according to standard protocols (Agilent Technologies, Santa Clara, CA). The microarray images were obtained using a 2uM scanner and probe intensity values extracted using NimbleScan software (Roche NimbleGen Inc., Madison, WI) for Nimblegen arrays and Agilent Feature Extraction software Agilent Technologies, Santa Clara, CA) for Agilent arrays.

3.3.4.4 Microarray Data Processing

Probe intensities for all samples were quantile normalized. Samples were visualized using Principle Component Analysis. Combat and SVA packages in R were used to remove batch effects due to RNA processing, cDNA creation, cDNA labeling and microarray hybridization. Transcript expression levels were summarized for each gene by averaging the batch corrected processed signal intensity of all the probes across its exons. Exon expression levels were summarized for each gene by averaging the intensities of all probes within each exon.

3.3.5 Time-Series Analysis and Stage Adjustment

For each piggyBac line, the time-series data were analyzed by comparing the gene expression profiles to the reference 3D7 time course dataset [52] to estimate hour
post invasion (hpi) to assess whether reinvasion time was called consistently between
piggyBac lines and replicates. Any small discrepancies in how reinvasion time was
estimated can be corrected for by using fourier transformation to fit the sinusoidal
expression patterns of the time course data [77]. Peak expression was determined for
each gene and compared to earlier publications on the P. falciparum life-cycle [111, 119].

3.3.6 Differential Expression

Differentially expressed genes between NF54 time points and between each
piggyBac line and NF54 were determined for 6, 26 and 38 hpi by comparing mean
expression via t-test with Benjamani-Hochberg FDR correction applied to account for
multiple testing. The lists of differentially expressed genes for each piggyBac line at each
time point were analyzed for enriched Gene Ontology (GO) Biological Processes using
the PlasmoDB tool [138] to determine what pathways were most affected by the
piggyBac insertion within each piggyBac line.

3.3.7 Differential Co-Expression

Gene co-expression networks (GCN) were built for each piggyBac line and NF54
to determine how the presence of the piggyBac insert alters the gene co-expression
network. GCNs were constructed using a method specific for time series data [139]. To
access rewiring between each piggyBac mutant and NF54, the DTWMIC values for edges
were compared between the networks as described in section 2.3.1.2. A rewired
network was constructed for significantly rewired edges. The gene IDs and enriched
biological processes were compared for the most rewired nodes and the most differentially expressed genes.

3.4 Results

3.4.1 Discerning Drug MoA

Comparing chemogenomic profiles of drugs with unknown MoA to a set of profiles of drugs with known MoA provides an efficient and informative way to establish potential MoA in a semi-high throughput way; while this method can identify simple drug targets, it is especially well-suited to recognize more complex biological effects induced by drugs that include polypharmacology and off-target effects. The drug panel used here contained 50 drugs; 44 drugs with known function and six with unknown function, including five artemisinin derivatives: dihydroartemisinin (DHA); Qinghaosu (QHS); artemether (AM); artesunate (AS); and artelinic acid (AL), and also lumefantrine (LUMF) (see Appendix A1 for full list of drug abbreviations and pathway information).

Using this profile matching approach along with hierarchical clustering of 50 drugs across 71 piggyBac mutant lines (Figure 3.1A), our data indicate that drugs cluster together on the basis of their shared target pathways. We viewed these data as a drug-drug network (Figure 3.2A) and found that the most highly significant links include those between drugs known to target the same pathway.

Included in these data is a particular focus on artemisinin-based profiles as a way to discern possible MoA; we observed that artemisinin derivatives are most highly correlated to drugs that target DNA repair, fatty acid synthesis, calcium ion metabolism
and hemoglobin metabolism. The artemisinin derivatives induce larger IC$_{50}$ shifts for piggyBac mutant lines with transposons affected genes involved in calcineurin signaling, cell cycle and autophagy as additional potential targets.

3.4.2 Discerning Artemisinin MoR

Hierarchical clustering of 71 piggyBac mutant lines across all 50 drugs (Figure 3.1A) identified several clusters of highly correlated piggyBac mutant lines (Figure 3.1B). Similar clusters were identified in a network analysis of the same data (Figure 3.2B). The cluster with the most tightly correlated piggyBac mutants includes one of particular interest due to an insertion in the promotor region of kelch13, a gene that has been associated with artemisinin resistance [7, 9, 67] that affects artemisinin resistance in vitro [63, 93]. The role of kelch13 in P. falciparum and its role in artemisinin resistance is not well understood; consequently, studying the relationships between the piggyBac lines in this cluster (hereafter referred to as the kelch13 cluster) (Figure 3.3) and associations between these piggyBac lines and particular drugs can help elucidate the role of kelch13 in artemisinin resistance. Seven mutant lines in the kelch13 cluster all show increased sensitivity across the panel of drugs, including to dihyrdoartemisinin (DHA), the active form of artemisinin (Figure 3.3A). This change in sensitivity cannot be attributed to a general decrease in growth rates of these piggyBac lines, as inferred from our observation that growth rate is not correlated with IC$_{50}$ across the entire piggyBac panel. Only three of the seven genes in the kelch13 cluster are annotated with a GO biological process or molecular function (Figure 3.3A); these are protein
phosphorylation, cell adhesion and ion transport. There are 159 genes with direct links to *kelch13* in an independent CGN that was constructed using ARACNE with a large drug perturbation dataset on mixed stage cultures that was subsequently pruned to remove indirect correlations between genes [140, 141]. The 159 direct neighbors of *kelch13* in the GCN are enriched for biological processes that have been linked to artemisinin MoA in the literature, including DNA repair, response to DNA damage stimulus, response to stress, and lipid biosynthetic process (Figure 3.2B). Several drugs’ chemogenomic profiles were very highly correlated to DHA over the *piggyBac* lines in the *kelch13* cluster (Figure 3.3C). These drugs have mechanisms that have been tied to artemisinin MoA in the literature including hemoglobin digestion, fatty acid synthesis, DNA synthesis.
Figure 3.1: Clustering of Chemogenomic Signatures for 71 piggyBac lines over a library of 50 drugs. (A) Hierarchical Clustering of Chemogenomic Signatures for 71 piggyBac lines over a library of 50 drugs. Colors in the heatmap represent a piggyBac line’s drug sensitivity relative to NF54 (yellow = sensitive, blue = resistant). (B) Hierarchical Clustering Dendrogram showing clusters of piggyBac mutants (p ≤ 0.05).
Figure 3.2: Drug-drug and Gene-gene networks. (A) Network of drug correlations, highly significant correlations shown in blue (FDR adj p value < 0.001) and significant correlation in black (FDR adj p value < 0.05). (B) Network of piggyBac correlations, edges represent significant correlations (FDR adj p value < 0.05) with edges denoting clusters of highly connected piggyBac lines.
Figure 3.3: Analysis of the kelch13 cluster identifies pathways associated with artemisinin MoA and MoR. (A) Annotations of genes in the kelch13 cluster. (B) Pathway enrichment analysis of genes directly linked to kelch13 in an independent Gene Co-expression Network built using a large dataset of drug perturbations. (C) Drugs correlated with DHA for the piggyBac lines in the kelch13 culture have MoA that have been implicated as artemisinin targets in the literature.
3.4.3 Transcription Profiling of *Kelch13* Cluster

Transcription analysis of medium-resolution time course of the IDC of *P. falciparum* mutant lines has led to better mechanistic understanding of mutant phenotypes [123]. Similar methodology was employed to determine the function of genes disrupted in each of the *piggyBac* mutants in the *kelch13* cluster. Transcriptional profiles were constructed for NF54 and each of the mutants in the *kelch13* cluster, with samples collected at four hour intervals throughout the IDC. The data were corrected for batch effects and stage of each sample was determined by comparing to the 3D7 reference IDC [111]. The data were determined to be high-quality and were used to examine IDC progression, differential expression and differential co-expression between the *piggyBac* lines in the *kelch13* cluster and NF54.

PCA of the *kelch13* cluster gene expression data indicated the presence of batch effects in the raw data. The main source of biological variation in *P. falciparum* time series transcription data is usually parasite developmental stage because of the large cyclic transcriptional changes throughout the parasite IDC. In a PCA plot of the first three principal components, the data points should roughly form a circle. This pattern was evident in the batch corrected data shown in Figure 3.4A, as is shown on PC2 and PC3 in Figure 3.4B. From Figure 3.4, it is evident that we should have collected another set of samples at 50 hpi to ensure the full IDC was observed in the dataset.
Figure 3.4: Principal Component Analysis of *piggyBac* time course samples. The 144 samples in the study are depicted on the first three principal components showing a cascade through the lifecycle with pink representing 2 hpi through 46 hpi in purple.
Gene expression profiles were correlated to a published hourly time course dataset for 3D7; hpi of the maximal correlation was plotted for each sample for each piggyBac line in the kelch13 cluster (Figure 3.5) to ensure that reinvasion time was called consistently between piggyBac lines and their replicates, and also that the piggyBac lines progress through the IDC at roughly the same rate, although with more variation in IDC progression in the middle of the IDC (Figure 3.5).

**Figure 3.5: Progression through the IDC as compared to the reference 3D7 time course.**

3.4.3.1 Differential Expression in NF54 across the IDC

The wild-type NF54 line had 247 significantly differentially expressed genes (FDR adj p value ≤ 0.05, summarized in Figure 3.6). The 6 hpi time point is up-regulated for genes involved in GO Biological Processes known to be important for ‘early stage’
parasites including entry from the host cell, exit from the host cell, locomotion, and modulations of host signal transduction. The 26 hpi time point is up-regulated for processes involved in protein folding, DNA recombination, and DNA integration. The 38 hpi time point is up-regulated for several biological processes including those related to cell division, such as DNA replication, chromosome organization, organelle organization.

Figure 3.6: Differentially expressed genes between time points of NF54. Boxes highlight a summary of enriched GO terms in up-regulated and down-regulated genes.
3.4.3.2 Differential Expression between *piggyBac* Lines and NF54

Enriched GO biological processes are reported for genes differentially expressed between NF54 and each *piggyBac* line for each time point in Supplement A2. All seven *piggyBac* lines were enriched for similar processes across the three time points. Figure 3.7 is a bipartite graph in which each *piggyBac* line (diamond nodes) is connected to biological processes (square nodes) for it is enriched.
Figure 3.7: Bipartite graph showing categories of enriched biological processes for differentially expressed genes between NF54 and the linked PB line.
3.4.3.3 Differential Co-Expression between *piggyBac* Lines and NF54

Table 3.2 outlines the key network statistics for each of the *piggyBac* rewiring networks including network size, degree distribution statistics and connectivity. A power law-like degree distribution and high connectivity, as are observed in these networks, are consistent properties of all validated biological networks. *PiggyBac* lines have similar numbers of differentially co-expressed gene pairs across the *P. falciparum* IDC ranging from 0.04-0.06% of possible edges. Figure 3.8 illustrates the enriched biological functions for the top 2% of rewired genes for each *piggyBac*. The top differentially co-expressed genes for each *piggyBac* line did not overlap with the significantly differentially expressed genes. However, similar pathways were enriched in the differentially expressed genes and most differentially co-expressed genes. The enriched functions for the differentially co-expressed gene lists seemed to be more precise, each *piggyBac* line was enriched for a predominant pathway (Figure 3.8).
# Table 3.2

**PIGGYBAC REWIRING NETWORK STATISTICS**

<table>
<thead>
<tr>
<th></th>
<th>PB58</th>
<th>PB57</th>
<th>PB55</th>
<th>PB54</th>
<th>PB52</th>
<th>PB50</th>
<th>PB35</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Nodes</td>
<td>3926</td>
<td>3737</td>
<td>4118</td>
<td>3852</td>
<td>4310</td>
<td>4296</td>
<td>4158</td>
</tr>
<tr>
<td>Number of Edges</td>
<td>8864</td>
<td>6063</td>
<td>8677</td>
<td>8732</td>
<td>11298</td>
<td>12098</td>
<td>11228</td>
</tr>
<tr>
<td>Connected Components</td>
<td>92</td>
<td>88</td>
<td>80</td>
<td>97</td>
<td>65</td>
<td>56</td>
<td>63</td>
</tr>
<tr>
<td>Largest Connected Component</td>
<td>3690</td>
<td>3486</td>
<td>3898</td>
<td>3589</td>
<td>4155</td>
<td>4164</td>
<td>3983</td>
</tr>
<tr>
<td>Average Degree</td>
<td>4.95</td>
<td>3.25</td>
<td>4.214</td>
<td>4.53</td>
<td>5.243</td>
<td>5.632</td>
<td>5.401</td>
</tr>
<tr>
<td>Maximum Degree</td>
<td>104</td>
<td>78</td>
<td>99</td>
<td>52</td>
<td>69</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>Network Diameter</td>
<td>22</td>
<td>22</td>
<td>20</td>
<td>20</td>
<td>17</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Power Law R²</td>
<td>0.937</td>
<td>0.946</td>
<td>0.859</td>
<td>0.920</td>
<td>0.925</td>
<td>0.922</td>
<td>0.925</td>
</tr>
<tr>
<td>Number of Clusters</td>
<td>45</td>
<td>49</td>
<td>50</td>
<td>44</td>
<td>28</td>
<td>31</td>
<td>42</td>
</tr>
<tr>
<td>Max Cluster Size</td>
<td>632</td>
<td>403</td>
<td>689</td>
<td>709</td>
<td>602</td>
<td>779</td>
<td>813</td>
</tr>
<tr>
<td>Median Cluster Size</td>
<td>15</td>
<td>13</td>
<td>12</td>
<td>9</td>
<td>16.5</td>
<td>8</td>
<td>7.5</td>
</tr>
</tbody>
</table>
Figure 3.8: Bipartite graph showing categories of enriched biological processes for differentially co-expressed genes between NF54 and the linked PB line.
3.5 Discussion

Comparing chemogenomic signatures induced by artemisinin derivatives to those derived from many drugs of known function led to the identification of several functions that had been implicated in artemisinin MoA in the cell. We found that the artemisinin drugs were most highly correlated with drugs that target DNA repair, fatty acid synthesis, calcium ion metabolism and hemoglobin metabolism (Figure 3.2A). Our publication of these results [74] add to the growing evidence that artemisinin has a complex MoA and directly affects many proteins and lipids. This study establishes proof of concept that chemogenomic screening can be used to elucidate MoA for new drugs.

Drug MoA and MoR can be investigated by identifying drugs that specifically impact piggyBac mutant lines, provoking large $IC_{50}$ shifts. Direct connections between artemisinin and piggyBac mutants implicated genes involved in calcineurin signaling, cell cycle and autophagy as additional potential artemisinin targets or mediators of artemisinin resistance. In human cancer cell lines artemisinin has been implicated in autophagy induction through the production of reactive oxygen species, however, it is not known if artemisinins induce autophagy in Plasmodium falciparum [142].

Artemisinin resistant lines are known to arrest progression through the cell cycle in early ring stage [143] and have a prolonged ring stage in comparison to artemisinin sensitive lines [66, 122, 124]. Artemisinin and DHA have been shown to be potent inhibitors of calcineurin in vitro [144].
In our study, each of the artemisinin derivatives have large effects ($\log_2(RPR) \geq \pm 3$) on a relatively small number of piggyBac mutants, indicating that direct cellular effects of particular derivatives of artemisinin may be mediated through a limited number of genes, despite reports that artemisinin binds to and modifies 1-2% of the malaria proteome [71, 72]. An expanded screen of artemisinin derivatives against a larger piggyBac panel will determine whether artemisinin derivatives kill largely through overwhelming activity arising from many drug-target interactions or primarily through a specific target or small combination of targets. QIseq is a new high-throughput method to read out phenotypes of mutants by ascertaining each piggyBac mutant’s relative growth by ascertaining its proportion in large pools of mutants; this same method can be applied to efficient drug-effect phenotyping [134]. Our collaborators have generated thousands of piggyBac mutants, reaching genome-wide saturation; use of these large pools to screen drugs’ impact on growth of piggyBac mutants en masse can be powerfully applied to link drug susceptibility and resistance directly to genes and pathways.

The chemoresponse profile of the piggyBac mutant line with a transposon insertion in the putative promotor region of kelch13 is highly correlated to seven other piggyBac mutant lines, together referred to as the kelch13 cluster. Because only three of these seven lines are annotated (Figure 3.3A) functional inference about kelch13 is limited, a challenge that can be ameliorated by expanding this screen to a larger panel of piggyBac mutant lines. Further analysis of this kelch13 cluster reveals insights into kelch13 function (Figure 3.3B) and artemisinin MoA (Figure 3.3C), some of which bolster
earlier suggestions of artemisinin MoA [74]. Examining the functions of genes linked to
*kelch13* in a GCN built using data from an independent transcription study, we identify
DNA metabolic process, DNA replication, DNA repair, cellular response to stress,
response to DNA damage stimulus, and lipid biosynthetic process. Drugs that
significantly correlate to DHA using the chemogenomic profiles of the *kelch13* cluster
had mechanisms that have been tied to artemisinin MoA in the literature including
hemoglobin digestion, fatty acid synthesis, DNA synthesis [71, 72, 81, 145].

Transcriptional profiling of *kelch13* cluster *piggyBac* mutants by differential
expression analysis allowed us to more specifically identify pathways perturbed by the
*piggyBac* insertion in each line. We compared the expression of each *piggyBac* mutant
to NF54 at 6 hpi, 26 hpi and 38 hpi. Each of the *piggyBac* mutant lines in the *kelch13*
cluster show differential expression of genes related to cell cycle, metabolism and
catabolism of nucleic acids, gene expression, pathogenesis, movement and transport
(Figure 3.7). Many of these processes have been implicated in artemisinin MoA or MoR
[65, 71, 72, 145]. *piggyBac* mutants with highly correlated chemogenomic profiles are
expected to share related functions in their disrupted genes. Our differential co-
expression analysis showed similar enriched processes, however, the profiles were more
specific than the differential expression profiles. For each *piggyBac* mutant, differential
co-expression analysis predominantly implicated a single pathway as affected by the
*piggyBac* insert with only small effects coming from additional pathways. The
differential expression analysis results were more general, showing that each *piggyBac*
insertion affects multiple pathways.
Our experimental design also allowed us to scrutinize parasite developmental stage and differences in the duration of the IDC between piggyBac mutant lines. We were also able to compare our NF54 life-cycle progression to the published malaria transcriptome for 3D7 [111]; 3D7 was cloned from NF54 by limiting dilution [146]. Our enriched biological processes for early ring parasites compliment other studies that have found genes in pathways related to host cell invasion have peak expression in early ring stage [111]. At 26 hpi (during the end of the trophozoite stage) we see an upregulation of protein folding and DNA recombination and integration. At 38 hpi (during the Schizont stage) we see an upregulation of genes involved in biological processes related to cell division. Our piggyBac mutant lines were highly correlated with NF54 through time leading to high confidence that differential gene expression and differential gene co-expression between lines is not due to stage differences.

In conclusion, we note that chemogenomic profiling is a promising tool to link drug phenotypes to genes and pathways and to improve gene annotation. This tool will become much stronger as a result of the new high throughput screening technology QIseq that allows phenotypic screening of thousands of piggyBac mutants simultaneously. Transcriptional profiling, both differential expression and differential co-expression, of PiggyBac lines is also a useful way to link genes to pathways. Differential co-expression analysis provides more specific results but is also more resource intensive. In our experimental design, differential co-expression required double the number of samples and a considerably more difficult sampling regime. However, with the life-cycle differences observed between the different lines, it is also
beneficial to generate a high-resolution time course as it gives confidence that stage differences are not an issue in the differential expression analysis.
4.1 Abstract

High-throughput cellular-based screens have identified thousands of compounds with strong antimalarial activity. These compounds must now be prioritized to identify lead compounds for further drug development. Prioritization normally depends on compound properties such as toxicity and scaffold uniqueness as well as toxicity screens. We propose that drug mechanism of action should also be considered. However, standard methods of determining how drugs kill cells do not readily scale to the level needed. Novel methods include the transcription profiling of drug-perturbed cell lines, an approach that has proven very useful for discerning anti-cancer drug action. We have adapted this method for *P. falciparum* and demonstrated proof of concept using an initial set of 31 drugs. In this chapter, I expand on our initial work by validating new methods to standardize the procedure and allow for increases in throughput.

4.2 Introduction

This project extends our initial development of an unbiased, semi-high-throughput method for predicting mechanisms of action (MoA) of chemicals/drugs with
antimalarial activity [80]. The need for a more efficient drug development pipeline is increasingly urgent due to emerging resistance to artemisinin, the last line treatment for multi-drug resistant malaria [15-17]. World-wide spread of artemisinin resistance would leave a void in effective treatment for multi-drug resistant malaria. With no new antimalarial drugs approved for use in humans, a large increase in malaria mortality would be imminent [12-14]. Thousands of compounds have been identified with antimalarial activity, but a high-throughput method for prioritizing these compounds based on target is lacking [42, 43]. As reviewed in section 1.8 determining drug mechanism is difficult and traditional methods of determining MoA are only viable on a much smaller scale than needed. Proteomic studies that directly demonstrate drug binding to protein targets are the most direct way to study drug MoA, but are difficult to carry out in practice. Pressuring drug sensitive parasites to evolve resistance and then sequencing resistant lines to determine mutations is a popular method of studying drug MoA and MoR. However, pressure and sequence studies are only successful in generating mutations in approximately 50% of drugs. The largest study to date for pressuring a sequencing to determine drug MoA only contains 31 drugs [62]. To be consider MoA earlier in the drug development process, we need a method capable of screening hundreds to thousands of drugs efficiently.

Recent work in cancer biology has generated drug perturbation transcriptional signatures (MoA signature) to determine drug MoA [75-77]. As part of the Connectivity Project (CMap), many laboratories have collaborated to build an open source dataset for drug perturbation signatures for different cells lines and signatures of genetic
perturbations, particularly disease states. A series of recent publications utilized this
database to repurpose drugs by matching patterns of gene expression for diseases with
anti-correlated patterns of expression for drugs [77, 78]. The CMap database also has
the potential to identify off-target effects that could provide useful information on
safety concerns for drugs much earlier in the drug development process.

Motivated by the success of CMap in cancer drug research, we are building
similar tools to predict drug MoA using transcription profiles from *P. falciparum*
perturbed with drugs and small molecule probes with known MoA [80]. In our initial
publication, we demonstrate that a transcriptional response ‘signature’ of *P. falciparum*
to drug perturbation (MoA signature) is highly indicative of the intracellular targets of
those drugs. These MoA signatures can be used to predict drug MoA by comparing the
MoA signature for an unknown drug to a database of MoA signatures for drugs with
known MoA. Additionally, in contrast to the one-drug-one-target paradigm, MoA for a
single drug is often complex, emphasizing the value of an MoA pathway signature that
can provide a direct readout of the pathways affected.

Despite the potential of transcription profiles as a method to determine drug
MoA, this method is difficult to implement in malaria. *P. falciparum* has a subtle
transcriptional response to perturbation; in fact, previous studies have found less than
two fold changes in gene expression in response to drug perturbation [147]. In contrast
to the large fold changes in gene expression seen over the IDC of the parasite [111], the
small fold changes observed in response to perturbation have led to speculation that
the *P. falciparum* transcriptome is hard-wired and largely unresponsive to perturbation.
There is a small, but growing body of evidence that *P. falciparum* does respond transcriptionally to perturbation, but that the response is smaller in magnitude than observed in other species [80, 141, 148, 149]. This subtle transcriptional responsiveness is more readily observed by looking at global transcription signatures rather than focusing on gene-by-gene expression differences. We have shown that MoA signatures provide valuable information on drug MoA [80], however, this study screened only 31 drugs at a single time point, without replication. To maximize the potential of this method, it must be validated and optimized.

*P. falciparum* lines with genetic perturbations (mutations) have changes in transcription in the affected pathway [123]. To validate our perturbation signatures, we compare genetic perturbations and drug perturbations affecting the same pathways. PiggyBac mutants [132] have stable, single gene disruptions and therefore provide well-defined genetic perturbations. Comparing the transcriptional responses of both genetic perturbation and drug perturbation that ‘target’ the same pathways provides a better understanding of the drug perturbation results and a validation of this methodology.

Currently, our MoA signature database contains MoA profiles for 31 compounds [80]. In order to account for non-specific effects of drug perturbation, the experiments to generate the MoA signatures in this database consisted of large batches of drug perturbations using drugs that target different pathways in each batch. The size of these experiments makes it impractical to include many drugs with unknown MoA in any batch; consequently, it would be of great value to have the capacity to screen many drugs with unknown MoA at a time. We explored the use of a normalization panel to
account for non-specific effects of drug perturbation in an effort to optimize the process of generating MoA signatures.

4.3 Methods

4.3.1 Parasite Culturing

*P. falciparum* lines HB3, Dd2, NF54 and nine NF54-derived *piggyBac* lines [56] were cultured as described in section 3.3.4.1. Small molecules were purchased from Sigma-Aldrich with the exception of JQ1 (provided by Dr. James Bradner, Harvard Medical School) and the SJ experimental compounds (provided by Dr. Kiplin Guy, St Jude’s Children’s hospital). Nine single gene knock-out mutants were carefully chosen from a *piggyBac* transposon mutant library [56] and matched with drugs targeting the same pathway (Table 4.1). Drug perturbations were carried out on triple Sorbitol synchronized cultures of HB3, Dd2 and NF54 as specified in Table 1. Ten mL aliquots of HB3, Dd2 and NF54 culture were exposed briefly for 2 h to each compound at a single developmental stage (24 hpi trophozoites) at each compound’s IC_{50} concentration obtained from literature sources or determined by hypoxanthine drug test in our laboratory [36]. The HB3 and Dd2 drug perturbations were done in a single batch with the same culture media and blood sources. The nine *piggyBac* lines and NF54 drug perturbations were grown as a single batch with the same culture media and blood sources, the *piggyBac* lines were collected at 26 hpi to coincide with the NF54 drug perturbations.
### TABLE 4.1

*PIGGYBAC TRANSPOSON INSERTIONS AND DRUGS TARGETING SAME PATHWAY*

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Gene function</th>
<th>Pathway</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF3D7_1368800</td>
<td>DNA repair endonuclease</td>
<td>DNA Repair</td>
<td>Methyl methanesulfonate</td>
</tr>
<tr>
<td>PF3D7_1227800</td>
<td>Histone acetyltransferase</td>
<td>Chromatin Modification</td>
<td>Apicidin</td>
</tr>
<tr>
<td>PF3D7_0831700</td>
<td>HSP70</td>
<td>Protein Folding</td>
<td>Geldanamycin</td>
</tr>
<tr>
<td>PF3D7_0322000</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>Protein Folding</td>
<td>Rapamycin</td>
</tr>
<tr>
<td>PF3D7_1343000</td>
<td>Phosphoethanolamine-N-methyltransferase</td>
<td>Fatty acid synthesis</td>
<td>PPMP</td>
</tr>
<tr>
<td>PF3D7_0110500</td>
<td>Bromodomain mutant</td>
<td>Chromatin Modification</td>
<td>JQ1</td>
</tr>
<tr>
<td>PF3D7_0811300</td>
<td>CAF1 family ribonuclease</td>
<td>RNA Turnover</td>
<td>Apicidin</td>
</tr>
<tr>
<td>PF3D7_1311900</td>
<td>Vacuolar ATP synthase, catalytic subunit a</td>
<td>Ion Transporters</td>
<td>Chloroquine</td>
</tr>
<tr>
<td>PF3D7_1216200</td>
<td>Glycerol-3-phosphate dehydrogenase</td>
<td>Mitochondrial</td>
<td>Atovaquone</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Electron Transport/ Glycolysis</td>
<td></td>
</tr>
</tbody>
</table>

4.3.2 RNA Extraction and cDNA Synthesis

Total RNA was extracted and processed to cDNA as described in 3.3.4.2
4.3.3 Labeling, Hybridization and Scanning

1µg of cDNA was labeled with Cy3 dye using 65% AT rich pre-labeled random hexamers as primers for cDNA synthesis by Klenow fragment of DNA polymerase I. Hybridizations for the HB3 and Dd2 drug perturbation experiments were performed for 17 h to a custom Agilent microarray [Turnbull, in prep] at 65°C at 10 rpm followed by washing of the arrays as described according to standard protocols (Agilent Technologies, Santa Clara, CA). Hybridizations for the NF54 drug perturbations and piggyBac experiment were performed for 22 h to a custom Nimblegen array (as described in Siwo et al. 2015 [36]) followed by washing of the arrays according to standard protocols (Roche NimbleGen Inc., Madison, WI). The microarray images were obtained using a 2uM scanner and probe intensity values extracted using NimbleScan software (Roche NimbleGen Inc., Madison, WI) for Nimblegen arrays and Agilent Feature Extraction software Agilent Technologies, Santa Clara, CA) for Agilent arrays.

4.3.4 Microarray Data Processing

Agilent Feature Extraction software employs advanced algorithms based on replicated control probes placed randomly on the array to correct the reported probe intensity signal for background signal and non-specific binding (Agilent Technologies, Santa Clara, CA). The Nimblegen platform does not contain negative control probes and NimbleScan does not account for and remove background signal and non-specific binding. Thus, control probes were built into the custom Nimblegen microarray to allow for background signal and non-specific binding correction as described in [36]. Briefly, a
set of 10,000 negative control probes with no sequence matches to the \textit{P. falciparum} genome was generated and included on the microarray. For each plex, a background distribution was created from the signal intensities of these 10,000 negative control probes and a minimum threshold was set at the 95\% percentile of this distribution. Any probes on the array that did not meet this minimum threshold were excluded from further analysis. A minimal expression threshold was also set for exons, 1000 exons were simulated by randomly sampling sets of 20 control probes and averaging their values to create a null distribution of exons and set the threshold at the 95\% of this distribution. When probes were averaged for each exon, any exon not meeting this minimal threshold was excluded from further analysis.

To account for plex to plex variation in hybridization, probe intensities for all samples within an experiment were quantile normalized. Transcript levels for each gene were obtained by averaging the processed signal intensity of all the probes across each of its exons. Exon signal intensity for each gene was obtained by averaging the intensities of all probes within each exon.

4.3.5 MoA Signatures

The global perturbation signature for each drug was constructed by averaging the perturbation signatures across two lab lines (HB3 and Dd2). The resulting data was then used to obtain a compound-specific response index for each gene by normalizing the gene’s average transcript level following perturbations in the two clones against its average level across all perturbations within the same experimental batch to obtain a
gene-specific response index. The global transcriptional response to a small molecule was then represented as a vector where each element represents a gene-specific response index. This vector is referred to as the genome-wide response index. Gene Ontology (GO) enrichment analysis was performed on the top 100 and bottom 100 (~2%) most responsive genes in the transcriptome and enriched biological processes (hypergeometric test $P < 0.05$) were determined using the GeneMerge with gene sets defined using the GO annotations in plasmoDBv9.3. For each small molecule, we then represented the transcriptional response as a binary vector whose elements are biological process categories that have a value of 1 (when the process is enriched following the perturbation) and 0 (when the process is not enriched).

4.4 Results

Transcriptional profiling of malaria parasites after brief perturbation with drugs and small molecule probes can be used to discern drug MoA [80]. Our data demonstrate the improvement of the throughput of this approach, along with further validation, key steps needed for this to be a viable method for screening drugs early in the drug development process so that MoA can be considered when prioritizing drugs for further development.

4.4.1 Normalization Panel

Previously, our method required large batches of drugs with many different MoA to remove non-specific effects, inhibiting our ability to test many drugs with unknown MoA simultaneously; in that approach, drugs with unknown MoA were spread among
different batches with drugs with known MoA. We solved this problem by developing a standard normalization panel consisting of a limited number of drugs with known MoA that could be run consistently with large batches of drugs with unknown MoA.

To determine how many drugs a normalization panel should contain, we analyzed subgroups of drugs that target diverse pathways from our full database of 31 drugs [36] to determine which combinations of drugs represent optimal normalization panels. We began by determining which drugs’ perturbation signatures had the highest correlation to the batch average (Figure 4.1). To determine the effectiveness of different mock normalization panels, the entire 31 profiles were combined into a single batch and mock panels of size two, three, four and five drug combinations of the drugs identified in Figure 4.1 were compared to the average profile across all three batches (Figure 4.2), demonstrating that normalization panels with three drugs were most effective.

To test our potential normalization panels, we included the top correlated drugs from Figure 4.1 (5-Fluorouracil, Cerulenin, Chloroquine, Epoximicin, JQ1 and Methotrexate) and several additional drugs with both known and unknown MoA in an experiment with biological replication using both HB3 and Dd2. Each candidate drug replicate was normalized using the panel and by the batch average to produce an MoA signature. The reproducibility of the MoA signature was determined by the degree of similarity in the MoA signatures between the replicates and across the normalization techniques. This experiment confirmed that a normalization panel with three drugs was most highly correlated to the batch average signal ($r = (0.94-0.97)$; Figure 4.3) and signatures were consistent between replicates.
Figure 4.1: Examining correlations between MoA profiles and non-specific effects. Potential drugs for a new normalization panel were chosen from the 31-drug batch based on correlation of their expression profile to the batch average, representing the most non-specific drug effects. Drugs target a variety of biological pathways. Boxes show drugs with highest correlations in each batch, and * indicate drugs used for further study based on correlation and pathway diversity. Drugs from each batch must be used to normalize out batch effects and non-specific drug response across the whole experiment.
Figure 4.2: Exploring optimal normalization panel size. All two, three, four, five and six drug panels were chosen from drugs that had the highest correlations to the batch average in each of the original three batches (5-Fluorouracil, Cerulenin, Chloroquine, Epoximicin, JQ1 and Methotrexate). Three drug panels had the highest correlations to the average across batches.

Figure 4.3: Testing possible normalization panels. All possible combinations of drugs for normalization panels are correlated to the batch average. Panels with the highest correlation for each size are shown. Green box shows panels with highest correlation overall, and * indicates that Panel 7 of three drugs has the highest combined correlation rank in the 31-drug batch and 10-drug batch. 5-FU consistently showed up in high ranking panels for both 31- and 10-drug batches. Cerulenin appeared in most highly ranking 10-drug batch panels, while JQ1 appeared in most highly ranked 31-drug batch panels.
4.4.2 Validation of MoA Profiles

To validate these drug perturbation signatures, drug perturbations and genetic perturbations affecting the same pathways were compared. Nine *piggyBac* mutants (Table 3.1) with well characterized *piggyBac* transposon insertion sites were selected from a *piggyBac* mutant library [132, 133]. Drugs were selected that target the same pathways that are disrupted by the *piggyBac* transposon in the selected *piggyBac* mutant lines (Table 4.1).

All genetic perturbations showed their largest effects in pathways that contained the disrupted gene (Table 4.2). Each drug perturbation was compared to all genetic perturbations and seven of nine (78%) matched most closely to the genetic perturbation targeting the same pathway. For example, the MoA signature for apicidin, a histone deacetylase inhibitor, is most similar to the signature of the genetic perturbation in histone acetyltransferase (PF3D7_1227800). The observed similarity between the drug perturbations and genetic perturbations that target the same pathway indicates that the MoA signatures are accurately capturing the effect of the drug perturbations. These results were consistent when either replicate, or the average of the two replicates were used to generate the MoA signatures.
4.5 Discussion

Currently, MoA is considered late in the malaria drug development process. Large cell-based screens identify compounds with high antimalarial activity in drug susceptible parasite lines such as 3D7 [42, 43]. Basic properties of these compounds and their scaffolds are then compared to prioritize compounds for further drug development. In an example of this process, MMV chose 400 compounds from thousands identified as having high antimalarial activity in these large screens [43]. Many different research groups are now working to determine the MoA of these compounds. A high-throughput methodology is needed to incorporate knowledge of

---

<table>
<thead>
<tr>
<th>PB#</th>
<th>Gene ID</th>
<th>Genetic Perturbation</th>
<th>Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>PB-46</td>
<td>PF3D7_1368800</td>
<td>DNA repair endonuclease</td>
<td>DNA Repair</td>
</tr>
<tr>
<td>PB-115</td>
<td>PF3D7_1227800</td>
<td>Histone acetyltransferase</td>
<td>Chromatin Modification</td>
</tr>
<tr>
<td>PB-118</td>
<td>PF3D7_0831700</td>
<td>HSP70</td>
<td>Protein Folding</td>
</tr>
<tr>
<td>PB-77</td>
<td>PF3D7_0322000</td>
<td>peptidyl-prolyl cis-trans isomerase</td>
<td>Protein Folding</td>
</tr>
<tr>
<td>PB-61</td>
<td>PF3D7_1343000</td>
<td>Phosphoethanolamine-N-methyltransferase</td>
<td>Fatty acid synthesis</td>
</tr>
<tr>
<td>PB-94</td>
<td>PF3D7_0110500</td>
<td>Bromodomain mutant</td>
<td>Chromatin Modification</td>
</tr>
<tr>
<td>PB-1</td>
<td>PF3D7_0811300</td>
<td>CAF1 family ribonuclease</td>
<td>RNA Turnover</td>
</tr>
<tr>
<td>PB-37</td>
<td>PF3D7_1311900</td>
<td>Vacuolar ATP synthase, catalytic subunit a</td>
<td>Ion Transporters</td>
</tr>
<tr>
<td>PB-42</td>
<td>PF3D7_1216200</td>
<td>Glycerol-3-phosphate dehydrogenase</td>
<td>Mitochondrial Electron Transport/ Glycolysis</td>
</tr>
</tbody>
</table>
MoA earlier in the development pipeline to increase both the success rate and cost-effectiveness of moving a drug forward.

Transcription profiling of *P. falciparum* lines after drug perturbations has proven useful in determining MoA [80]. However, that study screened only 31 drugs at a single time point, without replication. To maximize the potential of this method, it must be validated and optimized.

Transcription in *P. falciparum* follows a predictable cascade over the IDC, such that the expression for each gene follows a roughly sinusoidal pattern with peak expression occurring at different times along this IDC continuum for different sets of genes [111, 119]. Perturbation experiments demonstrated that transcriptional responses to perturbation in single genes are small when compared to changes in expression observed across the IDC [147], suggesting that the malaria parasite is unresponsive to perturbation. However, careful experiments with tight synchronization of cultures and careful attention to IDC stage and methodologies that observe the global transcription signatures as a result of perturbation have detected changes in gene expression in response to perturbation [80, 123, 141, 148, 149]. Adding to this growing body of literature showing *P. falciparum* transcriptional response to perturbation, we compared genetic perturbations and drug perturbations targeting the same pathways (Table 4.1). We found that the transcription profiles of *piggyBac* mutant lines showed the largest differences in expression for genes involved in the pathway containing the *piggyBac* transposon (Table 4.2). We observed that the MoA profile of drugs most closely matched the profiles for the *piggyBac* line affecting the same pathway as the
drug. These results were consistent when either replicate or the average of both replicates was used to construct the pathway signatures, indicating high reproducibility of transcriptional response across biological replicates.

Our method required experiments consisting of large batches of drug perturbations using drugs that target different pathways in each batch. The size of these experiments makes it impractical to include many drugs with unknown MoA in any batch; consequently, it would be of great value to have the capacity to screen many drugs with unknown MoA at a time. With this goal in mind, we designed a normalization panel of three drugs that replace the need for a large batch of drugs with different MoA. Initially, the data from our 31 drug database [80] was evaluated to identify drugs with diverse MoA whose perturbation profiles were significantly correlated with the batch averages, that together captured the range of non-specific effects observed in this dataset (Figure 4.2). The size of normalization panel influenced its ability to capture non-specific effects; three drug panels had the highest correlation to the average perturbation signature across the entire 31 drug set (Figure 4.3). In a new experiment designed to test the ability of a three-drug panel to accurately capture non-specific effects, the MoA signatures generated using the normalization panel method were nearly equivalent to MoA signatures generated using the whole batch method and a three-drug panel was superior to two, four and five drug panels (Figure 4.4). Employing our new normalization panel will allow us to test large batches where all the drugs have unknown MoA, greatly improving the throughput of this method by an order of magnitude.
With the successful implementation of a normalization panel, transcription profiling is a viable method for determining MoA for drugs with unknown MoA on the scale of tens to hundreds of drugs. This is roughly comparable to the chemogenomic method of determining MoA described in Chapter 3. However, the chemogenomic method could scale much more easily if the QIseq competitive growth of piggyBac lines in large pools of mutants [134] is adapted to measure growth under drug perturbation. A QIseq based screening method truly has the potential of being a high-throughput method for determining drug MoA. Results of Chapter 2 showed that transcription profiling and chemogenomic signatures provide complimentary results, with transcription profiling augmenting results from the chemogenomic screen. However, results from chemogenomic screens can be influenced by the genetic background of the parent line used to make the piggyBac library. Perhaps, QIseq based chemogenomics can be used as a first stage high-throughput screen and then transcription profiling over the IDC can be used to further refine drug MoA in parasites lines with different genetic backgrounds.
CHAPTER 5:

DETERMINING ARTEMISININ MECHANISM OF RESISTANCE AND CHOOSING PARTNER DRUGS BASED ON MECHANISM OF ACTION

5.1 Abstract

Effective new combination therapies are needed to treat multi-drug resistant malaria as ACT efficacy fails in Southeast Asia. There are new combination therapies in clinical trials, but none are approved for use to treat malaria. Drug repurposing has been used effectively in cancer biology to find effective new combination therapies and has the potential to drastically speed up approval for new combination therapies to treat artemisinin resistant malaria infections. This method relies on determining patterns of differential expression between resistant and sensitive cells and finding anti-correlated drug MoA signatures. Transcription profiles were generated from recent Southeast Asian parasite isolates to determine a signature of artemisinin resistance. The signature of artemisinin resistance was compared to the MoA database to choose partner drugs for new ACT.

5.2 Introduction

Recent work in cancer biology has used transcriptional cellular responses to drug perturbations to identify a set of genes and a pattern of expression that typifies a
response to perturbation by a particular drug (MoA signature) [75, 76, 150]. As part of the Connectivity Project (CMap), many labs are collaborating to build an open dataset for both drug and genetic perturbations (including disease states) for different cell lines. Within the last year, a flurry of studies has utilized this database to repurpose drugs by matching patterns of differential expression for diseases with anti-correlated patterns of expression for drugs [77, 78, 151-159]; consequently, the drugs ‘reverse’ the disease expression patterns and ultimately the phenotype (disease). This process of repurposing drugs can be extended beyond combating diseases to reversing drug resistance [79].

Wei et al. showed that glucocorticoid resistant acute lymphoblastic leukemia had a discernable gene expression pattern and predicted that rapamycin could restore glucocorticoid sensitivity in vitro. The CMap database also has potential to identify off-target effects that could provide useful information on safety concerns for drugs much earlier in the drug development process.

Previous work in our lab developed a CMap-type database for P. falciparum using drug perturbation expression profiles from clonal lab lines HB3 and Dd2 [80]. HB3 and Dd2 represent geographically and phenotypically diverse P. falciparum isolates; HB3 is a drug sensitive isolate from South America and Dd2 is a multi-drug resistant (but artemisinin sensitive) line derived from the Southeast Asian isolate, W2. These two isolates are not particularly relevant in studying clinical artemisinin resistance. In conjunction with Tim Anderson from Texas Biomed, our lab has cloned several field isolates of P. falciparum with varying artemisinin resistance phenotypes. The Anderson lab genotyped the blood samples taken directly from patients to identify the dominant
parasite line in the infection and then genotyped the subsequent lab-adapted clonal lines to confirm the genotypes matched the genotype of the original infections. Our lab has confirmed the slow clearance phenotype of the lab-adapted clonal lines in vitro using RSA. These lab-adapted clonal lines, both artemisinin sensitive and resistant, were used to define a profile of artemisinin resistant parasites (MoR signature).

In 2015, a large study of transcription profiles was published for 1043 in vivo P. falciparum infections (direct from patients) [66]. These transcription profiles form three main clusters based roughly on parasite stage, 549 ring-stage samples, 272 late-stage samples and 222 ring-stages samples with gametocytes. These samples were analyzed to determine if gene expression correlated with clearance rate; over-representation analyses of GO and KEGG pathways were completed for the set of correlated genes to find that artemisinin resistance was associated with increased expression of genes involved in the unfolded protein response pathways. This dataset provides an excellent resource to supplement our in vitro data on differences in gene expression patterns between artemisinin resistant and sensitive lines.

5.3 Methods

In collaboration with my lab mate Sage Davis, global transcription was measured for 20 artemisinin resistant and sensitive culture adapted field isolates from Southeast Asia at 6 and 24 hpi, with and without DHA perturbation, in duplicate. Isolates were grown under standard culture conditions as described in section 3.3.4.1 above. This experiment included a total of 240 samples, from 60 cultures; this was too large a
volume of culture work to be done in a single culture batch. Replicates for each isolate were randomly assigned to six batches. RNA extraction and processing was completed as described in 3.3.4.2 and batches were randomized and recorded in a metadata spreadsheet for analysis of batch effects. Samples were hybridized to our custom Agilent array in randomized batches as described in section 3.3.4.3 and reagent information was recorded in the metadata spreadsheet for analysis of batch effects.

5.3.1 Microarray Data Processing

Probe intensities for all samples were quantile normalized. Samples were visualized using PCA. The R package SVA [160] was used to remove batch effects due to culture date, RNA processing, cDNA creation, cDNA labeling and microarray hybridization. Transcript expression levels were summarized for each gene by averaging the batch corrected processed signal intensity of all the probes across its exons. Exon expression levels were summarized for each gene by averaging the intensities of all probes within each exon.

5.3.2 Time-Series Analysis and Stage Adjustment

For each isolate line, the gene expression profiles were compared to the reference 3D7 IDC dataset [111, 119] to estimate hour post invasion (hpi) to assess whether reinvansion time was called consistently between isolates and replicates. Because of the large differences in gene expression across the IDC, consistently calling reinvansion time is critical when looking for differentially expressed genes. Discrepancies in time zero calling were considered in further analysis.
5.3.3 MoR Signatures

An MoR signature was constructed by comparing expression patterns in artemisinin resistant and sensitive lines for different time points. Global perturbation signatures were averaged across resistant and sensitive isolates and then the ratio of gene expression for resistant isolates to gene expression for sensitive isolates was computed. Enriched GO biological processes were determined for the top 100 up-regulated genes and the top 100 down-regulated genes as described in 4.3.5. The enriched GO biological processes were then represented as a vector, whose elements are biological process categories that have a value of 1 (when the process is enriched in up-regulated genes following perturbation), -1 (when the process is enriched in down-regulated genes following perturbation) and 0 (when the process is not enriched). MoR signatures were compared between time points to determine if stage influences pathways present in MoR signatures.

5.3.4 MoR Signatures Compared to MoA Signatures to Identify Potential Synergies between Drugs

The MoR pathway signatures were compared to the MoA signature database using Pearson correlation. Permutation tests were conducted to identify drug MoA profiles that were significantly correlated to the MoR profile. Drug’s with significant negative correlations were identified as potential drugs to be included in a rationale combination with artemisinin.
5.4 Results

5.4.1 MoR Results

Mok et al. [66] observed differences in expression patterns between isolates from Southeast Asia and Africa. Our previous MoA work in Chapter 4 was done on established lab lines HB3 and Dd2, setting the stage to explore the effect of DHA perturbation in artemisinin sensitive and resistant isolates from Southeast Asia. We perturbed 20 Southeast Asian isolates with DHA and DMSO (as a control) at 5nm for 2 h at two time points (6 and 24 hpi) in duplicate.

This experiment was completed in several batches and metadata was collected for culture batch, RNA extraction batch, cDNA batch and hybridization batch. The R package SVA was used to remove known batch effects. Removal of batch effects improved the distribution of the samples as shown in Figure 5.1, bringing the single outlier closer to the mean of the distribution.

Figure 5.1: PCA plot of samples before (A) and after (B) batch correction
As discussed in previous chapters, parasites at different stages within the IDC of *P. falciparum* have very different expression patterns. When looking for differences in transcriptional responses between parasite lines it is necessary to carefully control for parasite stage. Comparing the gene expression profiles to the reference 3D7 time course showed variation in the estimated stage of the 6 and 24 hpi samples, suggesting a difference in how time zero was called across samples (Figure 5.2). To account for these stage differences between parasite isolates, further analysis was completed in two tracks; track A – assume reinvasion time was called correctly and group samples by 6 and 24 hpi, track B – create groups based on gross stage (ring, trophozoite or schizont) indicated by maximum correlation to reference 3D7 time course.
Figure 5.2: **Estimating stage for each sample.** A) Genome-wide correlations of each sample to 3D7 reference transcriptome. B) Time-point of maximum correlations graphed for the 6 and 24 hpi samples from the same biological replicate and treatment.
5.4.1.1 Expression Signatures of Artemisinin Resistance

Using our control samples, we asked whether artemisinin resistant and sensitive isolates had consistent expression differences. In the set of 20 isolates, there were 6 resistant isolates and 14 sensitive isolates. To account for observed difference in IDC stage of samples, we split groups of samples into three groups; late schizont and early ring (46-10 hpi), late ring and trophozoite (10-30 hpi) and schizont (30-44 hpi). Samples were also analyzed according time point collected (6 and 24 hpi). Within each group expression signatures were determined as described in section 5.3.3. Figure 5.3 shows expression signatures for each group as defined by stage.

The 6 hpi top 100 up-regulated genes and top 100 down-regulated genes are enriched for processes related to response to oxidative stress, cell redox homeostasis, lipid and fatty acid oxidation, nucleoside metabolic processes, protein/lipid localization and transport, and DNA replication. Many of these processes were evident in our Chapter 2 analyses and are related to either artemisinin MoA or MoR. The 24 hpi samples, however, had limited differences between resistant and sensitive parasites in the cell cycle, fatty acid and lipid biosynthesis, localization and transport and instead were dominated by processes related to interaction with host cells, immune evasion and pathogenesis.

Grouping samples by their maximal correlations to the 3D7 IDC gave similar results between the ring and 6 hpi grouping and then trophozoite and 24 hpi groupings. The late schizont to early ring group (46-10 hpi) pathway fingerprints resembled the 6 hpi group, sharing functional categories: response to oxidative stress, cell redox
homeostasis, lipid and fatty acid oxidation, and protein/lipid localization and transport.

The late ring and trophozoite group (10-26 hpi) results were distinct from the 24 hpi group, with changes in gene expression focused in pathways related to nucleotide metabolic processes, fatty acid and lipid biosynthesis and protein/lipid localization to the membrane. The schizont group (26-44 hpi) differentially expressed genes were enriched for processes related to pathogenesis, fatty acid and lipid biosynthesis, cell cycle, and transport. The targeting of proteins and lipids to organelle membranes is a process we did not observe in our previous analysis presented in Chapter 2, perhaps because of the lack of late stage and very early ring samples or differences in analysis techniques.
Figure 5.3: MoR pathway signatures by time point. Sample grouping affected MoR signatures. A stronger MoR signature was observed for early ring to trophozoite parasites.
5.4.1.2 Identifying the DHA Perturbation Signature

We investigated how DHA perturbs gene expression of artemisinin resistant and sensitive Southeast Asian isolates. For each isolate, gene expression differences between DHA perturbed samples and DMSO control samples were determined as described in section 5.3.3. The enriched biological processes for the top 100 up-regulated gene and top 100 down-regulated genes were determined and used to construct pathway signatures as described in 5.3.3. For the DHA perturbation signatures in artemisinin resistant and sensitive isolates, we observed extensive variation in the pathways that were enriched for each isolate, with only pathogenesis and immune evasion being consistently enriched across isolates.

5.4.2 Comparison of MoR and MoA to Predict Drug Synergies

We compared the MoR profile for 6 hpi with the MoA signature database and observed strong negative correlations between the MoR profile and expoxomycin. Positive correlations were observed to chloroquine and atovaquone.

5.5 Discussion

In Chapter 2 and 3 we saw the transcription profiles paired with either differential expression analysis or differential co-expression analysis can discern differences in pathways related to artemisinin MoA and MoR. The datasets used in these chapters include non-drug perturbed whole genome expression profiles. In this chapter, we investigate whether the MoA methodology of finding enriched GO biological processes in the top 100 up-regulated genes and top 100 down-regulated
genes (as defined by relative fold-change) can identify an MoR signature. We further ask whether an artemisinin MoR signature can be used in conjunction with MoA signatures to predict synergistic drug combinations.

The whole genome expression profiles generated for recent parasite isolates from Southeast Asia had high stage variability when compared to the 3D7 reference IDC. When samples were grouped by time point the sample was taken (6 hpi vs. 24 hpi) or by IDC stage (46-10 hpi, 10-30 hpi, 30-46 hpi), we observed different biological processes in the MoR pathway fingerprint for different time points. The MoR pathway fingerprints for early-stage time points, whether grouped by the original 6 hpi time point or the corresponding late schizont and early ring IDC stage from the comparison to the 3D7 reference IDC, were enriched for processes that are related to artemisinin MoA and MoR. These results are consistent with those observed in Chapter 2. The differentially expressed genes at the 24 hpi time point, however, were mostly enriched for processes related to pathogenesis and more general processes and lacked a strong signal related to artemisinin MoA or MoR. When samples were grouped by late ring and trophozoite IDC stage (10-30 hpi) there was more of a signal or artemisinin MoA and MoR present; late ring and trophozoite differentially expressed genes were enriched for pathways related to nucleotide metabolic processes, fatty acid and lipid biosynthesis and protein/lipid localization to the membrane. The schizont grouping did not contain a strong signal of pathways related to artemisinin MoA and MoR. This analysis indicates that sample stage can have a large influence on differential expression results. The 24 hpi sample set contains many late ring samples and many schizont samples, which are
broken into different groups in the stage based analysis. The late ring and trophozoite samples do have a stage specific expression pattern related to artemisinin MoA while the schizont samples do not. This is interesting as we did see a signal related to artemisinin MoA and MoR in the ‘late stage’ samples in Chapter 2, however, there were no schizonts included in this late stage group in Chapter 2. Early and mid schizont parasites have large overall amounts of expression and their signal may be dominating the 24 hpi analysis and masking any differences between artemisinin resistant and sensitive samples at late ring and trophozoite stages.

The targeting of proteins and lipids to organelle membranes is a process we did not observe in our previous analysis. Targeting of proteins and lipids to organelle membranes makes sense as a mechanism of adapting to oxidative damage to organelle membranes observed after treatment with artemisinin. In Chapter 2 we also observed changes in responses to oxidative stress, lipid and fatty acid biosynthesis and metabolism and protein transport which we hypothesized were related to responses to this oxidative membrane damage. Both the analysis techniques and data set differ between the two chapters. It is possible that examining genes with maximal fold-change differences between resistant and sensitive parasites adds information on artemisinin MoR, it is also possible that a data set generated under controlled lab conditions provides different information as well. As an extension of this work, the fold change analysis technique will be applied to the Mok dataset [66].

Fold-change in expression between artemisinin treated and untreated samples were highly variable across isolates. All isolates showed differential expression for genes
involved in pathogenesis and immune evasion. Other processes related to artemisinin MoA were observed, but not consistently across the isolates. This may be in part because we did not follow the MoA procedure outlined in Chapter 4, where we use large batches of drugs with diverse MoA or a normalization panel to control for non-specific effects of drug perturbation. Instead we did a single perturbation and compared it to a control with little replication.

The 6 hpi artemisinin MoR was compared to the MoA signature database [80] and had a negative correlation with epoxomycin. Epoxomycin, a proteasome inhibitor, has been shown to have a synergistic relationship with artemisinin in artemisinin resistant parasites [60]. This is a proof of concept for this method of screening drugs for synergistic relationships with artemisinin. As a further test of this method, we should expand the drugs tested to include other proteasome inhibitors. If this method can act as a screen for identifying potential synergistic partners for artemisinin it would be very helpful. Although synergism with proteasome inhibitors has been demonstrated, these are not good candidates for drug repurposing due to demonstrated human toxicity. A less toxic *P. falciparum* specific proteasome inhibitor would be a better candidate for an ACT partner drug and work to develop one is underway [161].
CHAPTER 6:
SUMMARY

Network analysis of previously published whole genome transcription datasets and GWAS results have provided a more detailed understanding of artemisinin resistance and *kelch13* function. Differential expression analysis and differential co-expression analysis can both identify genes and pathways that are important in artemisinin resistance, however, differential co-expression analysis was able to identify responses in artemisinin resistant lines to all known effects of artemisinin. These additional pathways could also be identified by applying machine learning techniques meant for high-dimensional data analysis to detect differential expressed genes predictive of artemisinin resistance.

Chemogenomic profiles and transcription profiles of drugs with known MoA can effectively help determine MoA of drugs with unknown MoA. We expect the utility of chemogenomics to only improve with the further development of the QIseq competitive growth drug phenotyping method. Differential expression analysis and differential co-expression analysis of *piggyBac* mutants can help determine the function of the gene affected by the *piggyBac* transposon insertion. Transcription profiling of a *piggyBac* mutant with a single *piggyBac* transposon inserted in the putative promoter region of *kelch13* identified effects in genes.
Transcription profiling is a useful tool to determine drug MoA. Results from Chapter 3 and 4 help to validate transcription profiling to define drug perturbation by showing the utility of profiling piggyBac mutants to identify the function of the gene affected by the piggyBac insert. However, the results from chapter 3 and 5 argue that stage specificity of the IDC affect transcription signatures and that multiple time points may be necessary to identify drug MoA. Employing a normalization panel removes some of the experimental difficulty in transcriptional profiling and is a promising technique to increase throughput and experimental complexity. MoR pathway signatures were able to identify many of the same processes as differentially expressed in artemisinin resistant parasites as the analyses in Chapter 2. Comparing the MoR signature to MoA signatures at 6 hpi allows us to identify drugs that are highly correlated with artemisinin and drugs that have been shown in the literature to act synergistically with artemisinin.

The techniques employed in this dissertation provide promising ways to study drug MoA, MoR, gene function and to identify drug synergies. However, a better understanding of gene function would greatly improve the utility of these techniques. In each analysis reported in this dissertation, 40-50% of the relevant genes have unknown function. In fact, 40% of the *P. falciparum* genome is annotated as “gene of unknown function”. This lack of knowledge of gene function for such a large portion of the genome hampers what we can learn from genome-wide analyses.

Traditional methods of determining gene function do not work well in *P. falciparum*. It is difficult to express full *P. falciparum* proteins in bacteria or yeast because of the high AT content and multiple transcriptional start sites. Forward genetic
screens have also been difficult to perform in *P. falciparum*. Predicting gene function informatically is also difficult because of the high AT content and because *P. falciparum* has so many unique proteins. New techniques need to be applied to determine the function of the remaining genes in the *P. falciparum* genome.

GCNs offer a promising way to annotate gene function. In validated GCNs from model systems, genes within a cluster or module have been shown to share functional similarity [120]. Using this information, it is possible to assign function to unannotated genes that are members of a cluster of module. However, this requires a validated GCN constructed using a broad perturbation dataset, such data is limited in *P. falciparum*. GCN validation is usually performed by comparing to a gold standard network, such as a protein-protein interaction (PPI) network. The PPI network for *P. falciparum* is very sparse as many proteins cannot be expressed in yeast to conduct a thorough yeast two hybrid screen. It is possible to validated GCNs by cross validation of enriched cluster functions using genes with known function. However, again because the functional annotation of *P. falciparum* is limited and the missing functional information is not random, this is problematic.

New informatics approaches based on homology of 3D protein structure are a promising technology for functional annotation of unknown genes. I-Tasser is a platform that annotates protein function based on predicted 3D structure and comparing to a database of known 3D structures and functions [162]. In a preliminary study done by a collaborator a small set of 200 proteins containing proteins with known and unknown function, I-Tasser successfully predicted the function of known proteins. This is a very
promising technology, however, unfortunately is computationally very resource intensive and has not yet been completed for the entire *P. falciparum* genome.

The small *piggyBac* transposon library used for chemogenomic pilot study presented in chapter 3 has been extended by our collaborators to reach full genome saturation. They have published a library of 1000 *piggyBac* mutants [134], but have created a much more extensive library. We have shown that even the small *piggyBac* transposon library is useful in improving annotation of gene functions and anticipate that this large library will be invaluable.

The combination of these promising new techniques should allow huge advances in *P. falciparum* functional annotation in the near future that will greatly increase the utility of the whole genome screening technologies used in this dissertation. We look forward to working to help improve community knowledge of gene function in *P. falciparum* and the anticipated improvement in systems biology utility that will result.
APPENDIX A:

ADDITIONAL TABLES

TABLE A.1

DRUG LIBRARY INFORMATION

<table>
<thead>
<tr>
<th>Drug</th>
<th>Target</th>
<th>Metabolic Pathway</th>
</tr>
</thead>
<tbody>
<tr>
<td>DHA</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>SF</td>
<td>Adenosylhomocysteinease</td>
<td>Methionine and Polyamine</td>
</tr>
<tr>
<td>BZD</td>
<td>Tubulin</td>
<td>Merozoite Invasion</td>
</tr>
<tr>
<td>PQ</td>
<td>Farnesyl diphosphate synthase</td>
<td>Methionine and Polyamine</td>
</tr>
<tr>
<td>IB</td>
<td>Farnesyl diphosphate synthase</td>
<td>Methionine and Polyamine</td>
</tr>
<tr>
<td>TTFA</td>
<td>Acetyl-CoA Carboxylase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>VER</td>
<td>Voltage-dependent calcium channels</td>
<td>Membrane transporters/efflux pumps</td>
</tr>
<tr>
<td>DNTA</td>
<td>Tubulin</td>
<td>Merozoite Invasion</td>
</tr>
<tr>
<td>TLM</td>
<td>Oxoacyl-ACP Synthase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>FADE</td>
<td>Adenosylhomocysteinease</td>
<td>Methionine and Polyamine</td>
</tr>
<tr>
<td>QHS</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>E.64</td>
<td>Cysteine proteases</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>MFQ</td>
<td>Hemozoin Polymerization</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>DNR</td>
<td>DNA and RNA synthesis.</td>
<td>DNA/RNA synthesis</td>
</tr>
<tr>
<td>MZB</td>
<td>IMP dehydrogenase</td>
<td>Purine</td>
</tr>
<tr>
<td>TOFA</td>
<td>Acetyl-CoA Carboxylase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>ROT</td>
<td>Glycerol 3-Phosphate dehydrogenase</td>
<td>Mitochondrial Electron Transport Chain</td>
</tr>
<tr>
<td>TCPA</td>
<td>Acetyl-CoA Carboxylase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>TRI</td>
<td>Enoyl-ACP reductase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>Drug</td>
<td>Target</td>
<td>Metabolic Pathway</td>
</tr>
<tr>
<td>--------</td>
<td>--------------------------------------------------</td>
<td>--------------------------------------------------------</td>
</tr>
<tr>
<td>C.MIR</td>
<td>50S ribosomal subunit, inhibition of protein synthesis</td>
<td>Protein synthesis</td>
</tr>
<tr>
<td>T.MIR</td>
<td>50S ribosomal subunit, inhibition of protein synthesis</td>
<td>Protein synthesis</td>
</tr>
<tr>
<td>SDZ</td>
<td>Dihydropteroate synthase</td>
<td>Folate metabolism</td>
</tr>
<tr>
<td>LUMF</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CRL</td>
<td>Oxoacyl-ACP Synthase or fatty-acid synthase</td>
<td>Fatty Acid Synthesis in Apicoplast</td>
</tr>
<tr>
<td>PEP.A</td>
<td>Aspartic acid proteases</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>E.PPMP</td>
<td>Ceramide cholinphosphotransferase, Ceramide glucosyltransferase</td>
<td>Spingomyelin and Ceramide</td>
</tr>
<tr>
<td>AM</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>CUR</td>
<td>Histone acetyltransferases</td>
<td>Transcription Regulation Inhibitors</td>
</tr>
<tr>
<td>TSA</td>
<td>Histone deacetylases</td>
<td>Transcription Regulation Inhibitors</td>
</tr>
<tr>
<td>SP</td>
<td>Unknown</td>
<td>Folate metabolism</td>
</tr>
<tr>
<td>ATQ</td>
<td>Cytochrome bc1 complex</td>
<td>Mitochondrial Electron Transport Chain</td>
</tr>
<tr>
<td>AQ</td>
<td>Hemozoin Polymerization</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>PIP</td>
<td>Hemozoin Polymerization</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>ALLO</td>
<td>Hypoxanthine-guanine phosphoribosyltransferase</td>
<td>Purine</td>
</tr>
<tr>
<td>AS</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>T.PPMP</td>
<td>Ceramide cholinphosphotransferase, Ceramide glucosyltransferase</td>
<td>Spingomyelin and Ceramide</td>
</tr>
<tr>
<td>AL</td>
<td>Unknown</td>
<td>Unknown</td>
</tr>
<tr>
<td>EHNA</td>
<td>Adenosine deaminase</td>
<td>Purine</td>
</tr>
<tr>
<td>CYTD</td>
<td>Actin Polymerization</td>
<td>Merozoite Invasion</td>
</tr>
<tr>
<td>CG</td>
<td>Dihydrofolate reductase</td>
<td>Folate</td>
</tr>
<tr>
<td>CYC</td>
<td>Serine C-palmitoyltransferase</td>
<td>Spingomyelin and Ceramide</td>
</tr>
<tr>
<td>CQ</td>
<td>Hemozoin Polymerization</td>
<td>Hemoglobin Digestion and FP IX polymerization</td>
</tr>
<tr>
<td>MTX</td>
<td>Dihydrofolate reductase</td>
<td>Folate metabolism</td>
</tr>
<tr>
<td>PYR</td>
<td>Dihydrofolate reductase</td>
<td>Folate metabolism</td>
</tr>
<tr>
<td>DAN</td>
<td>Aminodeoxy-chorismate lyase</td>
<td>Shikimate Biosynthesis</td>
</tr>
<tr>
<td>PHEN</td>
<td>Glycerol 3-Phosphate dehydrogenase</td>
<td>Mitochondrial Electron Transport Chain</td>
</tr>
</tbody>
</table>


