TRANSCRIPTOMICS OF MALARIA HOST-PATHOGEN INTERACTIONS IN PRIMATES

A Thesis
Presented to
The Academic Faculty

by

Kevin Joseph Lee

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy in the
School of Biology

Georgia Institute of Technology
December 2014

TRANSCRIPTOMICS OF MALARIA HOST-PATHOGEN INTERACTIONS IN PRIMATES

Approved by:

Professor Greg Gibson, Advisor
School of Biology
Georgia Institute of Technology

Professor Jung Choi
School of Biology
Georgia Institute of Technology

Professor I. King Jordan
School of Biology
Georgia Institute of Technology

Professor John McDonald
School of Biology
Georgia Institute of Technology

Professor Mary Galinski
School of Medicine
Emory University

Date Approved: 30 July 2014

To my beautiful wife, Kelly,

who continues to bear the whims

of my academic pursuits.

ACKNOWLEDGEMENTS

I would like to thank all of the individuals that have helped me to learn and grow as a scientific researcher, as well as a person, in the last four years at Georgia Tech. I would like to begin by thanking my committee.

Dr. Jung Choi's *Introduction to Bioinformatics* was the first course I took at Georgia Tech. He first exposed me to the actual molecular details of microarrays and high-throughput sequencing. In his class, the wheels of inquiry began to turn. A world of possibility was opened before me.

Dr. Soojin Yi has been a great mentor to me. Whenever I had a "new" idea, she would help me flesh it out or else inform me that the idea, while interesting, was not "new." She offered a critical eye to my proposals, and guide me to ask the important questions. She would not let me make unfounded assumptions, or at the very least would insist that I be clear about the underlying assumptions that I would make. It was in her Molecular Evolution class that I began forming an idea that would eventually be the entirety of my contribution to an evolve-and-resequence project in *Drosophila*. Due to a scheduling conflict, Dr. Yi was unable to be present for my thesis defence, but her comments and advice during my thesis proposal were greatly appreciated.

In one of the first seminars that I attended at Georgia Tech, Dr. John McDonald inspired me with his presentation on the importance of micro-RNAs in gene expression regulation especially in cancer tissues. His translational medicine approach to research has led to collaborations with medical doctors in the community, including colleagues of my father, an OB/GYN, with the goal of developing early diagnosis markers for ovarian cancer.

Most recently, I began a collaboration with Dr. Mary Galinski, a malaria biologist from the Yerkes Primate Center at Emory University. As the leader of the Malaria Host-Pathogen Interaction Center (MaHPIC), she has greatly affected me and helped mold my research questions especially those concerning parasite gene expression and parasite antigenic variation, in particular. She has demonstrated great poise while leading a large group of talented researchers while always staying grounded in the underlying scientific questions. The lessons I have learned from her go well beyond the field of malaria biology.

Dr. I. King Jordan served as my advisor for my first two years of this program. In my time both in his classes and in his laboratory, I learned the computational tools necessary for performing high-throughput sequence analysis. He also began to plant in me the seeds of understanding of how to perform research in this field of bioinformatics. He would make time to have coffee with me and others in the lab nearly every week to talk about our research but also to teach us life lessons. The training I received are, in no small part, the reason I excelled with Greg.

I rotated with Dr. Greg Gibson my first semester in the program. After spending time in both Dr. Yi's and Dr. Jordan's labs, I came full circle and returned to Greg's lab on the MaHPIC contract. Before we would receive data from MaHPIC, however, I had many months to wait. In the meantime, Greg put me to work on two seemingly disparate projects: evolve-and-resequence data from *Drosophila* and exome sequencing from individuals with a rare cardiomyopathy. The former led to a second-author paper in *Genetics*, and the latter led to a first-author paper that is currently in review in the *American Journal of Human Genetics*. It has been my great pleasure and privilege to work with Greg over the past two years. Both the breadth and depth of research projects on which I have worked has been very fulfilling. I thank Greg for his guidance but also for the freedom that he has given me over the past two years.

Apart from my committee, I would like to acknowledge many other people that

have played an integral role in my formation as a scientist in the past four years. While working on the exome sequencing project, I met Dr. Bahig Shehata, a pathologist from the Emory School of Medicine. From the very start, he and I had a connection: his children were attending my alma mater, Marist School in Atlanta, GA. My time with him has taught me much humility. It has been an honor to know him and to learn from the example of such an amazing man.

My fellow graduate students have also been integral to my formation as a computational biologist. Andy Conley, a graduate from the King lab, taught me an incredible amount of computer science, especially considering that before beginning at Georgia Tech, I had never run a Linux computer or server from the command line nor programmed in perl or R. While much of the coding has to be learned by reading and then doing, Andy was always across the hall willing to answer any question that would come up. And he *always* had the answer. It was a great pleasure to meet and grown in friendship with the other members of the Jordan lab: Lava, Daudi, Deepak, Angela, and Jianrong. All of them helped shape me as a researcher.

In the Gibson lab, I was the only male student until Urko Martinez joined last year. For the past many months, I have enjoyed our long talks about science (specifically human evolution) and life both in English and Spanish. While we often distracted each other from our "real" work, our conversations served as a crucible for new ideas. Urko is also a master of R, and any time I would get stuck, he would give me the simple two-word answer or he would sit with me for a half-hour until we hacked out a solution. He embodies the idea of collaborative research, and I know that he will continue to be a great asset to the members of the Gibson lab. From the Gibson lab, I would also like to thank Dalia Arafat. She has been an amazing lab manager and wet-lab researcher extraordinaire, and without her, there would be no high-throughput sequence data to analyze. It was my great pleasure to meet and grown in friendship with the other members of the Gibson lab: Jing, Monica, Yan, Bee, and Kartik. I

wish them all the best in their continued scientific endeavours.

The entire MaHPIC team has been a great help to me as I have formed my own questions and helped others answer their questions. In Mary's lab, I have worked closely with Stacey Lapp, a world-expert in antigenic variation. Apart from his impressive work ethic and dedication to his research questions, he has taught me that making progress in science requires patience and kindness, especially in multi-institute collaborations. I am indebted to him for his willingness to answer my often-numerous questions and for all of his experimental design expertise and wet-lab work that has made my down-stream analysis possible.

Esmeralda Meyer is yet another amazing researcher with whom I have had the privilege of working. Her job description seems to include everything. She helps coordinate and document meetings for the large MaHPIC contract, is involved in experimental design planning, and works tirelessly in the wet-lab to generate and record important measurements. In the midst of all of her own work, she always makes time to reply to any question I have regarding either experimental design or fundamental malaria biology. She is dedicated to her research questions, and she inspires me to work hard in order to advance the scientific effort in this important field of inquiry.

Dr. Rabindra Tirouvanziam and Chet Joyner have been great collaborators from the immunomics perspective of MaHPIC. I thank them for their conversations and ideas and for always challenging me to ask more nuanced questions in terms of the underlying molecular mechanisms and immunological response to infection.

The informatics core at UGA (Jeremy, Vishal, and Suman) have been a great asset to the MaHPIC team. They catalogue all of the data that comes in, and validate it, and make it available. It was in coordination with them that the integrative portions of this present analysis were performed.

Advances in science require standing upon the shoulders of giants, and Dr. John

Barnwell has been such a giant for me as I have explored questions in malaria. An entire chapter of this present thesis is on antigenic variation in *Plasmodium knowlesi*, a phenomenon first described by him more than 30 years ago (before I was born). Through his leadership and instructive presentations, I have learned much about malaria biology. John has both directly and indirectly guided the path of this thesis.

By convention, I have written most of this thesis in the first person singular, which gives the impression of individual and independent effort. This, of course, is not the case. The results presented herein were made possible by the efforts of many individuals, as described below. I have tired to include attributions in-text, as well.

For chapters 2, 4, and 5 (the 100-day experiments), the staff and veterinarians at Yerkes Primate Center cared for the animals during the course of the experiments with daily feeding, monitoring, and clinical assessment. These individuals also sampled the primates for blood and bone marrow. Important clinical work was supervised and performed by Alberto Moreno, Monica Cabrera-Mora, and Esmeralda Meyer. Experimental design was a collaborative effort with essential roles played by Mary Galinski, Alberto Moreno, Esmeralda Meyer, and many other principal investigators and support staff.

All wet-lab components of gene expression analysis (RNA extraction, purification, library preparation) were performed by Dalia Arafat, and the actual sequencing was performed by the Yerkes Non-human Primate Genomics Core led by Zach Johnson.

Metabolomics data production was led by Dean Jones, and primary data generation and quality control was performed by Karan Uppal and ViLinh Tran.

The life-stage expression deconvolution chapter (Chapter 3) was primarily conceived by Greg Gibson. I received help from Urko Martinez and Isabel Mendizabal in the implementation of STRUCTURE.

This work was funded by contract HHSN272201200031C from the NIH National Institute of Allergy and Infectious Disease, and in part by ORIP/OD P51OD011132

(formerly NCRR P51RR000165).

TABLE OF CONTENTS

DEDICATION . . . . . **iii**

ACKNOWLEDGEMENTS . . . . . **iv**

LIST OF TABLES . . . . . **xv**

LIST OF FIGURES . . . . . **xvi**

LIST OF SYMBOLS OR ABBREVIATIONS . . . . . **xxiii**

GLOSSARY . . . . . **xxv**

SUMMARY . . . . . **xxv**

I INTRODUCTION TO MALARIA **1**

1.1 Malaria - a persistent killer 1

1.2 Human impact of malaria 2

1.3 Parasite life cycle 4

1.4 Parasite gene expression 5

1.5 Diversity of life-history strategies: antigenic variation and hypnozoites 7

1.5.1 Antigenic variation 7

1.5.2 Hypnozoites 8

1.6 Anti-malaria treatments 9

1.7 Malaria Host-Pathogen Interaction Center (MaHPIC) 10

1.8 Specific aims 11

II A SYSTEMS BIOLOGY APPROACH TO DETERMINE THE EFFECT OF PYRIMETHAMINE ON THE NON-HUMAN PRIMATE *MACACA MULATTA* **13**

2.1 Abstract 13

2.2 Introduction 14

2.2.1 Effects of pyrimethamine 14

2.2.2 Motivating Hypotheses 15

2.3 Methods and materials 16

2.3.1 Experimental design and measured outcomes 16
2.3.2 Metabolomics feature quantification 17
2.3.3 Library preparation for RNA-seq 18
2.3.4 Short read mapping 18
2.3.5 Gene expression quantification 19
2.3.6 Variance component analysis 20
2.3.7 Differential gene expression (DGE) 21
2.3.8 Gene set enrichment analysis 21
2.3.9 Blood informative transcript (BIT) axes 22
2.4 Results 23
2.4.1 Subject-specific effects dominate variance components in the transcriptome 23
2.4.2 Hierarchical clustering 28
2.4.3 Covariance-based approach for data-type integration: principal component regression 30
2.4.4 Blood informative transcript (BIT) axes of variation 31
2.4.5 Differential abundance of genes, metabolites, and blood cell counts 34
2.4.6 Gene set enrichment analysis 37
2.5 Discussion 38
2.5.1 Methodology for data analysis and integration 38
2.5.2 Between-subject effects 39
2.5.3 Dysregualtion of gene expression 41
III HOST GENE EXPRESSION ALTERATIONS DURING MALARIA INFECTION 43
3.1 Abstract 43
3.2 Introduction 44
3.2.1 Project overview 44
3.2.2 *P. cynomolgi* as a model for *P. vivax* 44
3.2.3 Hypnozoites and relapse 45

3.2.4 Host gene expression in malaria infection 45
3.2.5 Host response to secondary malaria infections 47
3.2.6 Motivating hypothesis 48
3.3 Methods and materials 49
3.3.1 Experimental design 49
3.3.2 Transcriptome analysis 50
3.3.3 Gene set enrichment analysis 50
3.3.4 Blood informative transcript axis analysis 52
3.3.5 Cell-type-specific gene sets 52
3.4 Results 52
3.4.1 Parasitemia across the 100-day experiment 52
3.4.2 Host blood cell parameters are altered by malaria infection . 53
3.4.3 Primary parasitemia extensively alters host gene expression . 56
3.4.4 Gene set enrichment analysis 58
3.4.5 Cell-type specific expression 62
3.4.6 Blood informative transcript axis analysis 64
3.5 Discussion 66
3.5.1 Gene expression in a relapse 66
3.5.2 Comparison with previous study of *P. cynomolgi* 67
3.5.3 Systemic lupus erythematosus (SLE) and immune related gene set enrichment 68
3.5.4 Hematopoesis 70
3.5.5 Cell type deconvolution of the samples 70
3.5.6 Individual-specific responses to malaria infection 71
3.5.7 Caveats and limitations of this experimental design 72
3.5.8 Future studies 72
3.5.9 Conclusions 73

IV COMPOSITIONAL MODELLING OF *PLASMODIUM FALCIPARUM* ACROSS THE INTRA-ERYTHROCYTIC DEVELOPMENT CYCLE 74

4.1 Abstract . . . 74
4.2 Introduction . . . 74
4.2.1 *Plasmodium* gene expression . . . 74
4.3 Methods and materials . . . 77
4.3.1 Dataset . . . 77
4.3.2 Multi-dimensional scaling . . . 78
4.3.3 Developing a method for life-stage deconvolution of mixed cultures of *Plasmodium* using STRUCTURE . . . 78
4.3.4 Identifying genes specific to each IDC life-stage . . . 80
4.4 Results . . . 81
4.4.1 Description of the dataset . . . 81
4.4.2 Implementation of the probabilistic deconvolution method . . . 81
4.4.3 Genes for each parasite stage . . . 84
4.5 Discussion . . . 85

V GLOBAL SHIFTS IN PARASITE GENE EXPRESSION PROFILES ACROSS AN INFECTION TIME-COURSE . . . 87

5.1 Abstract . . . 87
5.2 Introduction . . . 87
5.2.1 Expression profiling of *Plasmodium* . . . 88
5.2.2 Relapsing malaria . . . 88
5.2.3 Experimental design overview . . . 89
5.2.4 Motivating hypotheses . . . 89
5.2.5 Chapter outline . . . 90
5.3 Methods and materials . . . 90
5.3.1 Whole-genome resequencing for *Plasmodium cynomolgi* B strain 90
5.3.2 Transcriptome analysis . . . 92
5.3.3 Down-sampling methodology . . . 93
5.3.4 Differential gene expression . . . 95
5.3.5 Gene set enrichment analysis . . . 95

5.3.6 Gene group trajectories 96
5.4 Results 97
5.4.1 Whole-genome resequencing for *Plasmodium cynomolgi* B strain 97
5.4.2 100-day experimental summary 99
5.4.3 Clustering of the parasites 101
5.4.4 Differential gene expression across parasitemia peaks 104
5.4.5 Gene set enrichment analysis for *Plasmodium cynomolgi* . . 104
5.4.6 Effect artemether on parasite life-stage abundance 106
5.4.7 Trajectory of selected multigene families 108
5.5 Discussion 109
5.5.1 Sexual stage abundance 110
5.5.2 Parasites from relapse spend less time in the schizont stage . 111
5.5.3 Previously observed effects of AMDs on parasite 112
5.5.4 Caveats of this study 112
5.5.5 Future studies 113
VI CONCLUDING REMARKS 115
REFERENCES 117
VITA 137

LIST OF TABLES

1 **Significant features for the effect of time. For each dataset, the number of significant features identified by the ANOVA test is shown for either fitting the subject effect or not fitting it.** 35

2 **Parasite stage-specific genes.** 84

LIST OF FIGURES

1 **Malaria endemicity and genetic resistance.** Global distribution of (A) *P. falciparum* and (B) *P. vivax* in human populations. (C) The prevalence of G6PD deficiency. Panels A-C from [53, 52, 65], respectively. 2

2 **Life-cycle of *Plasmodium*.** (1) An infected Anophiline mosquito takes a blood-meal from a host and transfers infective sporozoites. (2) The sporozoites enter the bloodstream and eventually enter a hepatocyte. (3a) For some species of *Plasmodium*, the parasite can then enter a dormant (hypnozoite) stage within the hepatocyte. (3b) Other parasites will go through develoment in the exo-erythrocytic development in the liver cell, eventually forming schizonts. (4) The schizonts rupture spilling merozoites into the blood stream. (5) Merozoites enter red blood cells and begin the intra-erythrocytic development cycle (IDC) and undergo many rounds of asexual reproduction. (6) Some of the parasites then switch to a sexual development stage forming gametocytes. (7) These gametocytes are then taken up by a biting mosquito where the parasite undergoes further development, completing its life cycle. Figure and legend adapted from Galinski, Meyer and Barnwell (2013). 6

3 **Overview of the cores of MaHPIC.** 11

4 **Experimental overview and the structure of the data.** CBC measurements were taken daily for the course of the 100 day experiment. RNA for transcriptome analysis was extracted from blood and marrow at seven time points (TP1-7). Blood plasma was sampled at the last 5 time points for metabolomic analysis (TP3-7). The major sampling time points were selected based on the estimates of when sampling would occur in the infection experiments. The light blue time points (TP1 and TP2; pre) are made before the first administration of pyrimethamine. The red arrows (TP3, TP5, TP7; post) are samples taken 7 days after the start of pyrimethamine treatment. The yellow arrows (TP4 and TP6; inter) indicate samples taken more than 30 days after the last pyrimethamine dose. 24

5 **Variance component analysis (VCA) for the three data-types.** The effect of animal (that is, the between-individual variance) is more than 30% of the variance for the two transcriptomes and the CBC data in the full data sets (A). Whereas the animal effect is much less prevalent in the metabolomic datasets. In the reduced datasets (B), the factor of animal explains much more of the variance. Note: a reduced dataset was not generated for the CBC data since it only had 13 measured features. 26

6 **Coefficient of variance changes as a function of mean feature value.** For both of the data-types, there is a relationship between the coefficient of variation and the mean value for each feature, stemming from increases in technical and/or biological variability at the lower end of the measurement spectrum. Each point represents a feature: either a gene or ion peak. Density contours (for every 5%) are included and show that most of the features are maintained after excluding lower abundance features. A line is drawn at x=5 and x=17 for the transcriptome and metabolome, respectively. Features lower than this level were excluded in the reduced dataset used in calculating the VCA. (A) marrow transcriptome; (B) blood transcriptome; (C) metabolome with AE column; (D) metabolome with C18 column. Note: the domain for the transcript average level was trimmed to [0,15] and transcript CV from [0,1] for figure clarity. 27

7 **Hierarchical clustering omics datasets shows lack of correspondence.** (A) Heatmap of the blood transcriptome with each primate uniquely colored. (B) CBC hierarchical cluster for counts of five cell types: RBCs, platelets, monocytes, lymphocytes, and granulocytes. (C) Heat map of metabolomic data for the C18 column. 29

8 **Heatmap of the correlation of the first 10 principal components of the transcriptome and the metabolome.** In the box on the top left (the correlation between the two transcriptome datasets), there are some significant correlations between the PCs, although there is not a clear one to one mapping (e.g. PC1 for marrow does not correspond to PC1 for blood). In the correlation between the metabolomic datasets (bottom right box), the significant correlations lie more clearly on the diagonal, demonstrating that these datasets are capturing similar information. Between the two data-types (large box, top right), there are no significant correlations after Bonferroni adjustment. . . . 32

9 **Blood informative transcript (BIT) axes of bone marrow expression as a function of animal and time.** Panels (A-C) show the BIT axes 3, 5, and 7, respectively, which were highly significantly different across the five macaques. Panels (D-F) show three significantly differentially regulated axes (2, 7, and 9) which appear to show coherent cycling as a function of the anti-malarial drug dosage. . . . 33

10 **Clustering of the significant genes as a function of time.** (A) Marrow and (B) blood gene expression levels clustered by time point. Red (blue) lines indicate the genes that clustered together and were up-(down-)regulated in the two transcriptomes. 36

11 **KEGG gene set enrichment plots for four representative pathways.** The bar that transitions from red to white to blue indicates the value of the t-statistic; red signifies genes that are highly expressed in either the post- or inter-drug treatment, whereas blue signifies genes that are highly expressed in the pre-drug treatment. Each vertical black line is the location of a gene in the specified pathway. (A) One carbon cycling by folate; (B) oxidative phosphorylation; (C) cell cycle; (D) apoptosis. n/s signifies that the enrichment was not significant at FDR=5%. 37

12 **Analytical pipeline of gene expression profiling.** After the 100-day infection cycle, samples from all time points are then used to make paired-end, strand-specific libraries for sequencing on the Illumina Hi-Seq. After sequencing, reads are mapped to a combined reference genome and transcriptome. HTSeq is used to assign expression levels to each annotated gene, and expression levels are then normalized using the method suggested by DESeq. Subsequent down-stream analyses including differential gene expression, cell type profiling, and blood informative transcript axes profiling are then performed on the normalized expression values. 51

13 **Parasitemia levels across the 100-day infection cycle.** Parasitemia across the 100-day experiment. Approximate days of sampling are indicated by arrows. At TP2 (pink), a single-day dose of artemether was administered. At TP3 and TP4 (red), a full eight-day course of artemether was given (see methods for dosage). Primate RFv14 was euthanized on day 23 due to renal failure. Note that a pseudo-count of 1 has been added to the parasitemia counts before log transformation, and therefore a log-parasitemia of zero is zero. This data was collected by the Malaria Core principally led by Alberto Moreno. 54

14 **Abundance of blood cell types across the 100-day infection cycle.** Cell abundances in whole blood of the major blood cell populations: (A) red blood cells (RBCs), (B) white blood cells (WBCs), and (C) platelets. The black line is a kernel-smoothed fitted line showing the trajectory of cell abundances. Data points are colored by animal as in Figure 13. At the peak of the first parasitemia corresponding to days 15-25, there is a precipitous drop in the numbers of all blood cell types, a phenomenon known as pancytopenia. At the two subsequent relapsing parasitemias (occurring around day 60 and day 90), there are smaller dips in the three blood cell populations. (D) The distribution of lymphocyte counts by animal. Center line of the diamond denotes the mean and the ends are the 95% confidence intervals. The (E) mean corpuscular volume and the (F) mean corpuscular hemoglobin across the 100-day experiment. The animal that was euthanized due to severe malaria and renal failure (RFv14, blue triangles) had lower levels of these two blood traits even before patent blood stages. 55

15 **Summary of the blood transcriptome** (A) The variance components of the blood transcriptome. (B) Hierarchical clustering of the samples based on the correlation of their principal component values. 57

16 **Differential gene expression analysis.** (A) A hierarchically clustered heatmap of the significant genes. Volcano plots comparing (B) TP1 versus TP2, and (C) TP1 versus TP4. The dotted line near $y = 2.3$ is the cut-off for FDR=5%. Points represent genes and are colored (if significant) by there direction of effect in the TP1 versus TP2 contrast, up-regulated at TP2 in red, and down-regulated at TP2 in blue. 58

17 **Cycling of the SLE-related pathway.** (A) Gene set enrichment plots of the systemic lupus erythematosus (SLE)-related pathway genes annotated by KEGG. (B) The first principal component of the SLE-related pathway plotted by time. TP2 and TP3 are enriched with genes that are up-regulated in the SLE-related pathway. TP4 there is no significant enrichment. Then at TP5 and TP7 genes in the SLE-related pathway are expressed in the opposite direction, that is they are significantly down-regulated. TP6 is enriched with up-regulated genes. 60

18 **Gene set enrichments.** For a selection of gene sets with significant enrichment in at least one time point compared to control, a heatmap of the normalized enrichment scores (NES), which describe the level of enrichment of each pathway. 61

19 **Cell-type specific gene expression across the experiment.** The general behavior in B-cells (A) and T-cells (B) is qualitatively opposite. CD8+ T-cells (C) do not account for the extreme variation in the T-cell population. 63

20 **Blood informative transcript (BIT) axes over time.** All plots except (A) show the BIT axis trajectory after removing the between-animal effect. (A) Axis 2, hematopoesis-related ($p = 0.0002$). (B) Axis 3, B-cell activation-related ($p = 0.0032$). (C) Axis 5, cytokine receptor activity ($p = 0.0067$). (D) Axis 7, interferon signaling ($p = 0.0085$). (E) Axis 8, RNA-processing ($p = 0.0045$). (F) Axis 9, apoptosis-related ($p = 0.0481$). 65

21 **Expression profiling of the IDC of *Plasmodium falciparum*.** (A) A heatmap of the gene expression levels of genes (y-axis) for samples (x-axis) taken at hourly intervals across the IDC. The extreme red color represents a 64-fold enrichment over the pooled average across the IDC, whereas the extreme green represents a 64-fold reduction over the pooled average. The left column of panel A shows microscopic views of the various life stages. The top part of panel A contains abundance estimates for each of the three unique cell type. Adapted from [23]. (B) and (C) Multi-dimensional scaling (MDS) plots of the samples and the genes, respectively. Samples are labelled by the hour at which they were taken. 82

22 **Compositional modelling of the hourly expression data of the IDC.** The three panels represent the STRUCTURE output from k=3-5 from top to bottom, respectively. Importantly, the k=4 and k=5 populations collapse and converge to essentially the same solution as k=3. This demonstrates that k=3 is a stable solution for the number of unique cell types in the IDC. 83

23 **Integrative Genome Viewer screen capture shows no polymorphisms.** In the reads of this genomic region, there appear to be some sequence polymorphisms which may be due either to acquired mutations due to passaging of the parasite or sequence errors. 98

24 **Parasitemia across the infection and transcriptome read depth.** (A) Parasitemia for each of the animals across the 100-day experiment. A single-day dose of artemether was given to three of the primates (RFa14, RFv14, RMe14) at TP2 (pink box), and full 8-day courses were given to all animals at TP3 and TP4 (red boxes). Parasitemia was much higher during the first infection relative to the relapsing infections. (B) The number of reads mapping to each of the three sources of RNA: the host (macaque), the parasite (*P. cynomolgi*, and the spike-in control RNA. The highest levels of parasite RNA occur at TP2 and TP3, which corresponds to the points of highest parasite density. . . 100

25 **Primary parasitemia expression profiles cluster away from relapse profiles.** (A) A heatmap of ten libraries with sufficient parasite read depth to be considered expressed. The samples from the first parasitemia peak (TP2 and TP3, in red and gold, respectively) hierarchically clustered mostly within time point. The deepest branch in the clustering is between the relapsing parasite profiles (light blue), and the primary parasitemia samples. (B) Same as in (A) but using the down-sampled read counts for each library. 102

26 **Correlation of average expression between experimental groups across all genes.** (A) TP3 versus TP2, (B) secondary versus TP2, and (C) secondary versus TP3. Non-parametric density gradients are overlaid in colored lines. Genes that show a down-regulation in the secondary infection relative to TP2 are colored in black; all other genes colored in grey 103

27 **Volcano plots showing the magnitude and significance of differential gene expression across the three experimental groups.** (A) TP3 versus TP2, (B) secondary versus TP2, and (C) secondary versus TP3. The the significance level of the difference (-log(p-value), y-axis) is plotted against the log-fold change in expression (x-axis). Significant genes are colored by up- or down-regulation (red or blue, respectively) in secondary parasitemias relative to TP2 (B). 105

28 **Life-stage-specific gene set enrichment.** For each pair-wise contrast of the three experimental groups, the enrichment of life-stage for the three asexual IDC forms (ring, trophozoite, and schizont) and the sexual development form (gametocyte) is shown. For each plot, a vertical black line represents a gene specific to the given life-stage. The bottom horizontal line which transitions from red to grey to blue represents the t-statistic for the given contrast. Enrichment of genes on the left side of the GSEA plot indicates coherent up-regulation in the group on the left of the heading label. Enrichment of genes on the right side of the GSEA plot indicates coherent up-regulation in the group on the right of the heading label (e.g. gametocyte genes are up-regulated in TP3 compared to TP2; first column, last row). 107

29 **Life-stage-specific gene expression.** (A-D) Plots of the first two PCs for the life-stage-specific gene sets. In all four plots, the two parasite profiles from TP3 (green inverted triangles) for the macaques that did not receive a sub-curative dose of artemether are closer together than to the parasite profile from the macaque that did receive the sub-curative artemether treatment. Samples from TP2, blue triangles; samples from relapsing time points, gold circles. (E-H) Principal component loading plots. 109

30 **Trajectory of PHIST proteins and tryptophan-rich antigens (TRA) across the 100-day experiment.** Composite measure of (A) PHIST protein and (B) TRA at the three clustered groups. Principal component loading of the (C) PHIST and (D) TRA genes show tight coherence of expression with the first PC explaining 76% and 80%, respectively. 110

LIST OF SYMBOLS OR ABBREVIATIONS

AA	African-American.
AE	anion exchange.
ALS	amylotrophic lateral sclerosis.
AMD	anti-malarial drug.
ANOVA	analysis of variance.
BIT	blood informative transcript.
BM	bone marrow.
BMS	bone marrow suppression.
CBC	complete blood count.
COI	complexity of infection.
CYIR	*P. cynomolgi* interspersed repeat.
DGE	differential gene expression.
DHF	dihydrofolate.
DHFR	dihydrofolate reductase.
EA	European-American.
FDR	false discovery rate.
FY*O	Fy glycoprotein null, duffy antigen negative.
G6PD	glucose-6-phosphate dehydrogenase.
GSEA	gene set enrichment analysis.
GWAS	genome-wide association study.
HbB	hemoglobin B.
HLA	human leukocyte antigen.
IDC	intra-erythrocytic development cycle.
iRBC	infected red blood cell.
KEGG	Kyoto encyclopedia of genes and genomes.

LC/MS	liquid chromatography/mass spectrometry.
MaHPIC	Malaria host-pathogen interaction center.
MCH	mean corpuscular hemoglobin.
MCMC	Markov-chain Monte Carlo.
MCV	mean corpuscular volume.
MDS	multi-dimensional scaling.
MODS	multi-organ dysfunction syndrome.
m/z	mass-to-charge ratio.
NIAID	National Institute for Allergy and Infectious Disease.
NMF	non-negative matrix factorization.
nt	nucleotide.
PB	peripheral blood.
PHIST	*Plasmodium* helical interspersed subtelomeric proteins.
PPAR	peroxisome proliferator-activated receptor.
RBC	red blood cell.
RIN	RNA integrity number.
RNA-seq	RNA sequencing.
SLE	systemic lupus erythematosus.
SNP	single nucleotide polymorphism.
THF	tetrahydrofolate.
TNF	tumor necrosis factor.
TRA	trypophan-rich antigen.
VCA	variance component analysis.
WBC	white blood cell.

SUMMARY

Malaria, a devastating disease that primarily affects individuals in developing nations in the tropics, is caused by parasites from the genus *Plasmodium*. Much time and money has been invested to better understand the underlying biology of both the parasite and the host response to infection with the ultimate goal of improving malaria treatment and prevention. In this thesis, I explored the host response to both anti-malarial drugs as well as to various infection peak types. I found that the host transcriptome and metabolome are greatly altered in response to pyrimethamine, an important anti-malarial drug, and I characterized the specific gene sets that are dysregulated. The nature and persistence of gene expression dysregulation after pyrimethamine treatment raises important questions concerning the prolonged use of this drug.

While many genes are altered in response to the anti-malarial drug, pyrimethamine, the host transcriptome is extensively altered during a primary parasitemia peak in a *P. cynomolgi* infection of *Macaca mulatta*. However, in response to relapsing parasitemia peaks, the host transcriptome behaves in a qualitatively different manner. Specifically, there is extreme response to the primary infection with almost no response to relapsing parasitemia peaks. Some of the dysregulated pathways in the primary malaria infection peak overlapped with previously-identified gene set and suggest a relationship with auto-immune etiology.

After computationally defining parasite life-stage-specific gene sets, I also profiled the parasite transcriptional response to multiple infection peaks. Much like in the

host transcriptional response, parasite gene expression was qualitatively very different between primary and relapsing parasitemias. Specifically, there is a shift away from sexual stage parasites in the relapsing parasite population. I also describe an artemether-induced difference in the parasite transcriptome, which will require further experimental validation.

Lastly, I examine the parasite molecular machinery that is associated with antigenic variation. I found that, as anticipated, *SICAvar* transcription is down-regulated genome-wide in parasites passaged in splenectomized primates. Further, I identified many exported proteins that are also down-regulated and which may play a role in the control of *SICAvar* expression in its broadest sense. Further investigation will help refine the roles of each of these co-regulated genes in the expression of variable antigens, an important parasite feature that contributes to virulence.

CHAPTER I

INTRODUCTION TO MALARIA

1.1 Malaria - a persistent killer

Approximately a quarter of a billion people are stricken with malaria infections each year, and nearly one million deaths result, mostly in children under 5 years of age.[56] Malaria is an infectious disease caused by unicellular eukaryotic parasites of the genus *Plasmodium*. Once bitten by an infectious mosquito, patients will often not experience symptoms for one to four weeks.[42] When the patent blood-stage parasites emerge, patients will experience flu-like symptoms which include headache, fever, chills, vomiting, and muscle pain.[98, 42] Clinically, malaria infection can lead to anemia, and in severe cases, patients may experience coma or death.[69, 68, 92, 98, 42]

Malaria infections are geographically concentrated in the tropical regions of the world: Africa, Latin America, and Southeast Asia and Oceana (Figure 1A,B).[12, 53, 18, 52, 19] In these regions, it is mostly developing countries that bear the majority of malaria cases. This pernicious disease impacts not only human health but also had broader economic impact.[40, 129, 31] Historically, the range of malaria extended well into Southern Europe and the southern United States.[93, 125] Recently, however, vector control programs have limited mosquito populations and thereby eliminated stable transmission of the disease in most developed Western countries.[75, 32] In spite of the fact that nearly half of the world's population is at risk of contracting malaria, there is still no approved and effective vaccine for this disease.[2]

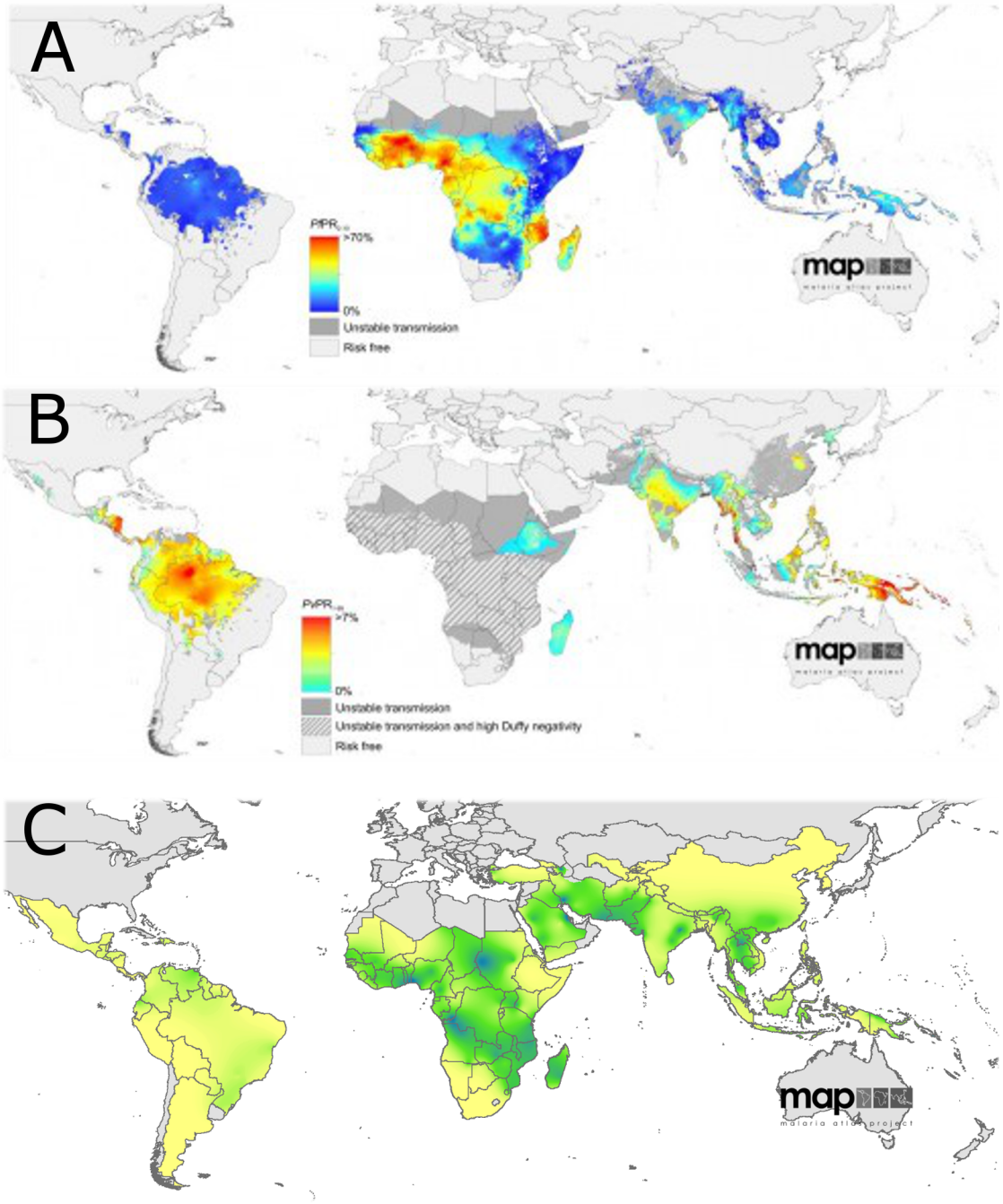

Figure 1: Malaria endemicity and genetic resistance. Global distribution of (A) *P. falciparum* and (B) *P. vivax* in human populations. (C) The prevalence of G6PD deficiency. Panels A-C from [53, 52, 65], respectively.

1.2 Human impact of malaria

Malaria infection has exacted a large fitness cost on the human population.[147, 79, 162, 60, 61, 27] As mentioned above, the majority of malaria-related deaths occur

in children. Furthermore, the anemia caused by high parasite loads may also play a role in reducing host fitness. Based on population genetics, it has been inferred that the numerous polymorphisms that confer resistance to malaria have experienced high levels of positive selection in the past 10,000.[147, 124, 60, 162]

Perhaps the most widely known and referenced human polymorphism of malaria is the sickle-cell mutation (HbS) in the gene that codes for hemoglobin B (HbB). This point-mutation (glutamic acid to valine, E6V) at the sixth amino acid residue of hemoglobin leads to an aggregation of hemoglobin inside the red blood cell when the hemoglobin is not binding oxygen. In spite of the negative consequences of being homozygous for this mutation, the sickle-cell allele is at moderate allele frequency ($\approx$15-20%) in many malaria-endemic regions.[119] Based on haplotype analysis, it is hypothesized that the sickle-cell trait arose independently in at least four populations, a observation that provides evidence for the high selective pressure that malaria has exerted on human genome evolution in the recent past.[79, 1, 61] Furthermore, there are two other novel mutations in the hemoglobin B gene, HbC and HbE, that confer protection against severe malaria.[4]

There are countless other mutations that have been selected for their ability to protect against malaria mortality and morbidity.[147, 79, 162, 60, 61, 27] For example, there are α- and β-thalassemias, glucose-6-phosphate dehyrogenase (G6PD) deficiency, and the Duffy-negative phenotype[1]. Most of these loci show high population-level signatures of positive selection in the recent past (i.e. within the last 5,000 to 10,000 years).[79, 61] These traits all offer the host resistance to the parasite at the level of the red blood cell.

[1]On a more technical note, it should be said that while there is some evidence for protection against malaria for heterozygotes of Duffy negative genotype (i.e. the FY*O allele), it is generally considered recessive. That is, only those individuals who are homozygous for the Duffy negative allele will receive the protection against malaria. This offers evidence that *P. vivax* was not the driving selective force for the near fixation of the Duffy-negative allele in African populations. Other members of *Plasmodium* lineage may have been the cause of selection for the FY*O allele.

In addition to alterations in the red blood cell phenotype, it is also likely that malaria exerted selection pressures that have shifted the host immune response. For example, there is an human leukocyte antigen (HLA) locus that is highly protective against severe malaria, HLA-b53.[62] While many loci that provide protection against severe malaria have been identified, there are likely many others that have yet to be discovered.

And there are at least intimations that the selective pressures imposed by malaria infection have affected hosts in such a way as to make them more susceptible to other immunological disorders.[157] This large number of malaria-resistance loci in the human genome clearly demonstrates the importance of malaria not just in the number of cases and fatalities caused each year, but also in the indelible mark that it has left on the evolution of our species.

1.3 Parasite life cycle

Historically, four (and recently a fifth, *P. knowlesi*) species have been known to cause malaria in humans.[113, 138, 117, 34, 41, 88] Of these four, *Plasmodium falciparum* is the most well-studied.[47, 23, 37] It causes both the greatest number of malaria infections each year, as well as the most deaths. *P. vivax* is the next most highly abundant malaria parasite in terms of human cases. It is the most common malaria parasite outside of Africa and has high prevalence in Latin America and Asia. It is nearly absent from Africa due to the near fixation of the Duffy-negative allele.[160, 86] The other two parasites are *P. malariae* and *P. ovale*, and these two species account for relative few malaria infections (less than 5%). These protozoa are transmitted from human to human via the invertebrate vector, mosquitoes of the genus *Anopheles*. The entire human component of the *Plasmodium* life cycle is haploid. The only diploid stages occur in the mosquito.

After an infective mosquito bites a human (or other host), the injected sporozoites

then migrate to the liver where they invade host hepatocytes (Figure 2). In some *Plasmodium* species like *P. vivax* and *P. cynomolgi*, some of the parasites enter into a dormant liver stage, the hypnozoite stage, which can be reactivated to produce a relapse infection. The remaining parasites begin active growth and division and subsequently produce multiple exo-erythrocytic schizonts that rupture the hepatocyte and form merozoites that will then enter the bloodstream and invade a host red blood cell. Within the red blood cell, the parasite will cycle through the intra-erythrocytic development cycle (IDC), passing through ring, trophozoite, and schizont stages. The parasite will then burst the infected RBC, producing between eight and 32 merozoites that can then infect other red blood cells. Some of the parasites in the blood will leave the asexual development of the IDC and develop into gametocytes, the sexual form of the parasite. Production of gametocytes is an essential part of the transmission process. It is the gametocytes that are taken up by the mosquito and subsequently develop into the gametes which later form the zygote.

1.4 Parasite gene expression

Throughout its various stages of development, the malaria parasite experiences dramatic alterations and cyclic expression of numerous genes across its genome.[23, 24] Perhaps the most well-studied section of the transcriptome of *Plasmodium* is the IDC, which is readily abundant in the host bloodstream.[23, 24, 37, 44, 97, 110] Bozdech and colleagues performed a competitive hybridization gene expression study using microarrays, and elegantly demonstrated the cyclic and tightly regulated waves of expression across the 48-hour IDC of *P. falciparum*.[23] More recently, Bozdech and colleagues showed that the same cyclic regulation in *P. vivax*.[24] In yet another seminal contribution by Bozdech and colleagues, researchers showed that protein abundance of most genes across the IDC was highly correlated with transcript abundance, albeit with an average time delay of 11 hours.[44]

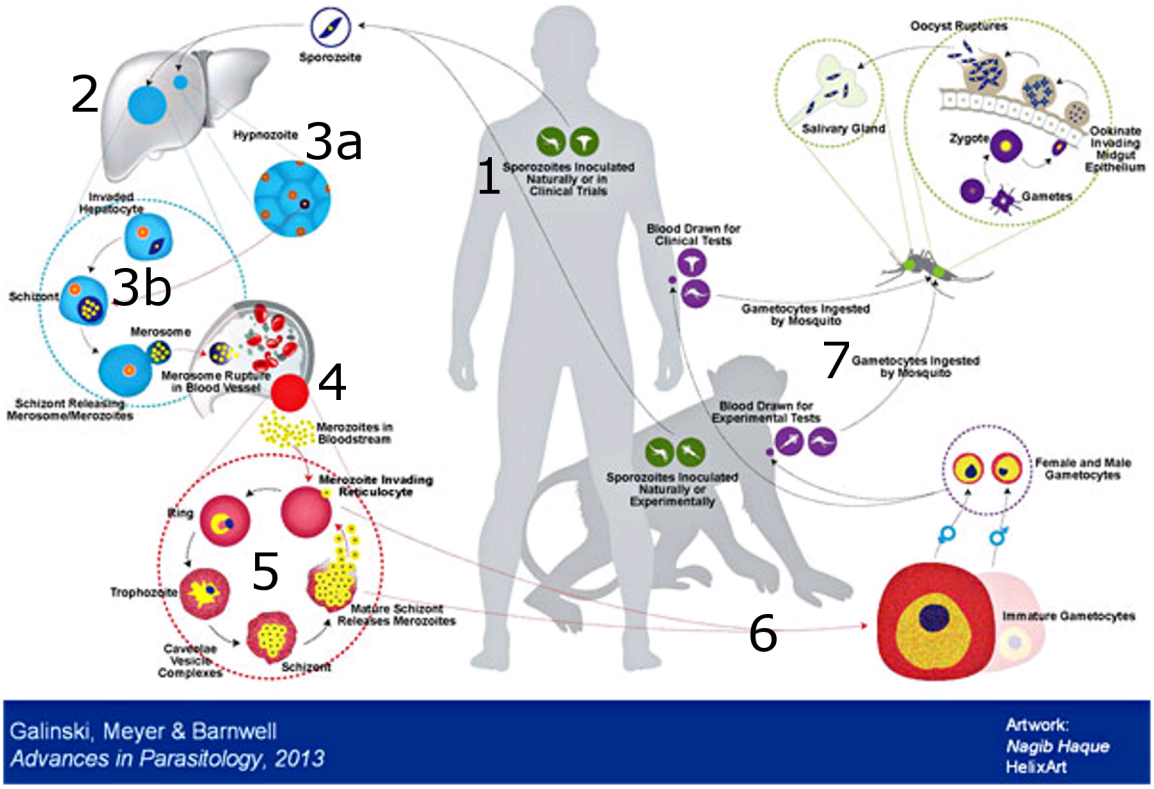

Figure 2: Life-cycle of *Plasmodium*. (1) An infected Anophiline mosquito takes a blood-meal from a host and transfers infective sporozoites. (2) The sporozoites enter the bloodstream and eventually enter a hepatocyte. (3a) For some species of *Plasmodium*, the parasite can then enter a dormant (hypnozoite) stage within the hepatocyte. (3b) Other parasites will go through develoment in the exo-erythrocytic development in the liver cell, eventually forming schizonts. (4) The schizonts rupture spilling merozoites into the blood stream. (5) Merozoites enter red blood cells and begin the intra-erythrocytic development cycle (IDC) and undergo many rounds of asexual reproduction. (6) Some of the parasites then switch to a sexual development stage forming gametocytes. (7) These gametocytes are then taken up by a biting mosquito where the parasite undergoes further development, completing its life cycle. Figure and legend adapted from Galinski, Meyer and Barnwell (2013).

As described above, some parasites exit the asexual development of the IDC and enter into sexual development as gametocytes. The transcriptional patterns of this stage have also been studied, and many gametocyte-specific genes have been identified.[164, 136, 72] In addition to the above-mentioned studied which were all performed *in vitro* in laboratory culture, both *in vivo* and *ex vivo* studies of parasite

transcription in various mammalian hosts have been performed.[37, 87, 36, 83]

While there have been many studies which have examined parasite gene expression, there have not been any published reports (to my knowledge) that investigate alterations in *parasite* gene expression across multiple infection peaks within the same host. The elucidation of the nature of parasite expression changes over various infection peaks may shed light on the mechanisms that dictate the severity of disease in the host.

1.5 Diversity of life-history strategies: antigenic variation and hypnozoites

In spite of many shared features of the life-cycle and expression profile between the numerous species of *Plasmodium*, different species of malaria parasites employ unique mechanisms for avoiding the host immune system and increasing their reproductive success. Two of these strategies are directly interrogated in this thesis and will be discussed in detail in their respective chapters. I briefly introduce antigenic variation and hypnozoites here.

1.5.1 Antigenic variation

Antigenic variation is the process by which some species of malaria parasite (e.g. *P. falciparum*, *P. coatneyi*, and *P. knowlesi*) evade clearance by the host immune system by altering the composition of the cell membrane of the infected RBC. In general, these parasites export one or a few antigenic proteins to the host cell membrane. In response, perhaps, to recognition by the host immune system, the parasite switches the exported antigenic protein to another of the many antigens in the repertoire in their genome.[13, 14, 64, 63] The exact parasite molecular machinery used to accomplish this protein export has not been fully elucidated.[126, 127, 148, 165] Interestingly, it appears that the presence of the spleen or splenic factors are necessary for the expression of the variable antigens by the parasite on the iRBC cell membrane.[13, 81]

The mechanism(s) by which the parasite senses the presence of splenic factors is also unknown.

1.5.2 Hypnozoites

A second important parasite mechanism for immune evasion is the dormant liver stage, the hypnozoite, which is employed by a number of malaria parasite species (e.g. *P. vivax*, *P. cynomolgi*, and *P. ovale*).[66, 45, 15] As described above, after the mosquito injects the sporozoites into the blood stream of its host, the parasite migrates to the liver. Some of the sporozoites will immediately begin development and initiate schizogeny in the hepatocytes. Other sporozoites will enter a dormant stage called a hypnozoite only to re-emerge weeks, months, or even years later. This delayed release of blood stages likely increases the complexity of infection (COI): the number of unique parasite strains in an individual at a given time. A higher COI can result in a much higher level of out-breeding and may facilitate genetic recombination between distinct strains of the same parasite species.[28, 30, 35] Increased recombination makes it easier for *P. vivax* and other hypnozoite-forming species to improve reproductive fitness in the face of external pressures (e.g. anti-malarial drugs). This difference in life-history between *P. falciparum* and *P. vivax* may explain why the latter has a higher genetic diversity and smaller haplotype blocks.[28, 30, 35, 109]

Apart from being a unique component of *P. vivax* biology, hypnozoites present an important hurdle in the control and eradication of malaria. Even after curative doses of co-artemether therapy (Coartem), for instance, the dormant liver stages can still reactivate.[17] The only FDA-approved anti-malarial drug currently shown to have anti-hypnozoite activity is primaquine.[17] Primaquine, unfortunately, has only limited use in many *P. vivax*-endemic areas because it can cause hemolytic anemia in individuals with glucose-6-phosphate dehydrogenase (G6PD) deficiency, a mutation common in many human populations at risk for malaria.[5, 144]

1.6 Anti-malaria treatments

In addition to the two above-mentioned antimalarial drugs (primaquine and artemether), there are many other mechanisms for controlling malaria and treating those who are infected.

The vaccine which has made the most progress toward regulatory approval is RTS-S, which targets the circumsporozoite protein.[2, 123] Another recent study found that using sporozoites dissected from infected mosquitoes salivary glands can yield at least some short-term protection against malaria infection.[130] In spite of the many years of malaria vaccine development, an effective vaccine is still not available. Many difficulties have impeded the development of a malaria vaccine. One such impediment is that a vaccine requires that the host mount an effective immune response, and it continues to be difficult to achieve a strong immune response in most individuals who live in malaria-endemic regions due to immunodeficiency stemming from mal-nutrition and other socio-economic factors. In the absence of an effective vaccine, continued enquiry into the underlying biology of the parasite will assist in the development of other complementary treatment and prevention options, a fact that underscores the continued need for basic and applied malaria research.

Currently, anti-malarial drugs are one of the defences against the development of life-threatening severe malaria. Pyrimethamine, an important anti-malarial drug which is usually paired with a sulfonamide, inhibits the enzymatic conversion of folate to its active form.[137, 16] In spite of the importance and broad usage of this drug in the treatment of malaria, there has been little investigation into its effects on the host at the molecular level. In light of case reports of side effects of this drug, it is likely that it dysregulates many important molecular pathways of the mammalian system.[101, 73, 159] Research into this drug's off-target effects on the host will inform proper medical usage, especially since pyrimethamine has been suggested as a treatment for amyotrophic lateral sclerosis (ALS).[80]

1.7 Malaria Host-Pathogen Interaction Center (MaHPIC)

For most researchers in the malaria field, the ultimate purpose of malaria research is the elimination of malaria.[45] Due to the complex nature of this disease, achieving this goal will require a more nuanced and multi-faceted approach than most other diseases. This difficulty of eradication arises from many factors including the multiple-staged life-cycle of *Plasmodium*, as well as the non-linear effects that come from intervention in the complex ecosystem of interactions involved in disease transmission cycle.[108, 56] Human intervention to reduce transmission may only shift the burden to a different age class, or worse, actually increase mortality and morbidity from the disease.[108, 56] The complexity of this parasite and the host-parasite interaction demand more research into this devastating disease.

Towards the end of eradication and with all of these challenges in mind, an inter-disciplinary team from four institutes across the state of Georgia (Emory University, Georgia Institute of Technology, University of Georgia, and the Centers for Disease Control and Prevention) were awarded a contract from the National Institute for Allergy and Infectious Disease (NIAID) to begin to address the gaps in the malaria knowledge base. As a team, we have generated and will continue to generate and analyze data from a range of different molecular and cellular technologies including transcriptomics, metabolomics, lipidomics, proteomics, as well as host and parasite cell counts and CBCs and more detailed interrogation of immune cell populations using flow cytometry to quantify both numbers and activity levels of cells from the innate an adaptive immune systems (Figure 3). Each of these technologies will be probing a unique part of the host and/or the parasite response. The analysis and integration of these data types from a 100-day control experiment, a 100-day infection experiment, and a complementary *ex vivo* expression profiling experiment are presented in this thesis.

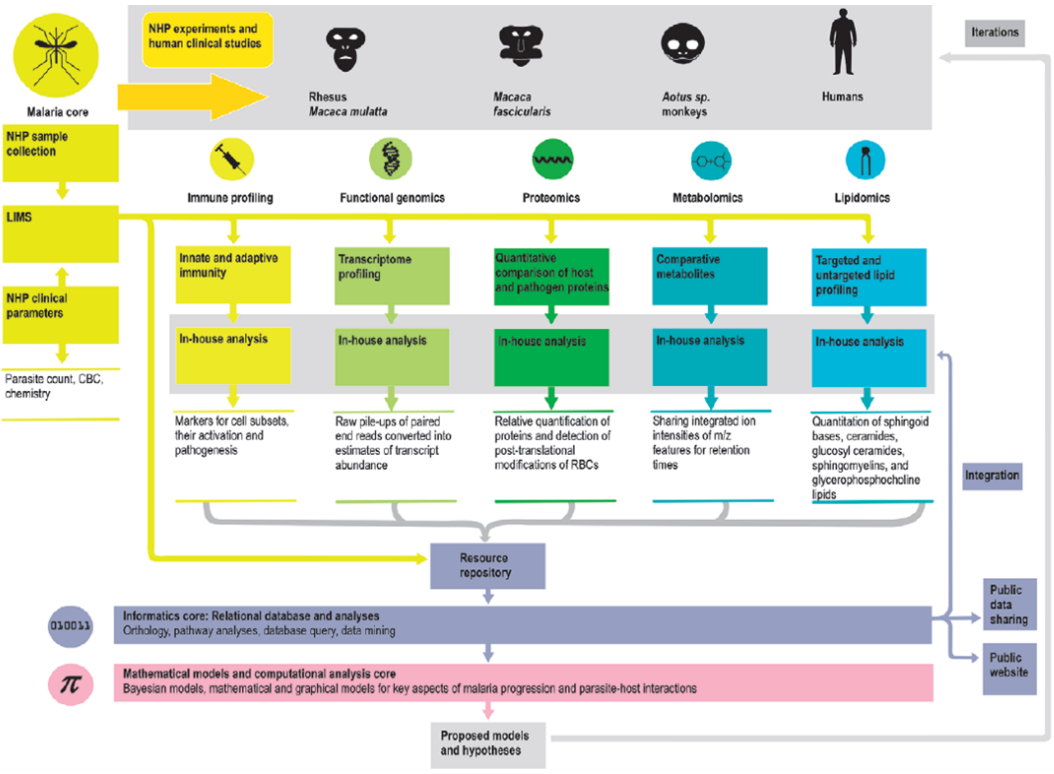

Figure 3: Overview of the cores of MaHPIC.

1.8 Specific aims

In CHAPTER II, I describe the analytical methods and integrative techniques which are used in subsequent infection experiments. I also explore the effects of the anti-malarial drug pyrimethamine on the transcriptome of the rhesus macaque (*Macaca mulatta*).

In CHAPTER III, I investigate the host transcriptional response to malaria infection, specifically interrogating the difference between primary and relapsing parasitemia peaks.

In CHAPTER IV, I implement a novel approach to expression deconvolution to a previously published *P. falciparum* dataset to identify genes whose expression is parasite stage-specific.

In CHAPTER V, I assess the qualitative changes in the parasite transcriptome

between primary and relapsing infection peaks with the goal of identifying and characterizing the qualitative differences between them.

In CHAPTER VI, I conclude the thesis, summarizing my contribution to the scientific body of knowledge of host-pathogen interactions in malaria.

CHAPTER II

A SYSTEMS BIOLOGY APPROACH TO DETERMINE THE EFFECT OF PYRIMETHAMINE ON THE NON-HUMAN PRIMATE *MACACA MULATTA*

2.1 Abstract

The Malaria Host-Pathogen Interaction Center (MaHPIC) was established to bring a integrative systems biology approach to the underlying molecular and cellular processes occurring within both the host and the pathogen during malaria infection. Herein, I present analysis from a control experiment in which rhesus macaques were profiled over a 100-day experiment.

This chapter is composed of two primary parts: the description of a methodology for the integration of multiple data-types including two omics technologies, and the examination of the biological impact on the host of pyrimethamine, an anti-malarial drug. Especially when used in combination with other antifolates, pyrimethamine, a common anti-malarial drug (AMD), has been shown to cause bone marrow suppression (BMS). Also known as myelotoxicity, BMS is the decrease in various blood cell populations (red blood cells, immune cells, and platelets) likely due to pyrimethamine's disruption of folate cycling. Currently unanswered questions exist concerning both the duration of cellular process disruption after pyrimethamine administration as well as the rapidity of clearance in subsequent exposures. In this study, I will examine the effect of pyrimethamine on the transcriptome and metabolome of the rhesus macaque. I further characterize the molecular pathways that are disrupted after administration of this drug.

2.2 Introduction

The Malaria Host-Pathogen Interaction Center (MaHPIC) was established to bring a integrative systems biology approach to the underlying molecular and cellular processes occurring within both the host and the pathogen during malaria infection. To this end, MaHPIC researchers experimentally will infect rhesus macaques with different strains and species of *Plasmodium*, which serves as a well-established primate model for human malaria. As controls for these studies, macaques were inoculated with a mock sporozoite preparation (which contained no *Plasmodium* parasites). These control macaques were then followed over the course of 100 days and given anti-malarial drugs on the days that the infected animals were expected to need treatment. This experiment offers both the opportunity to describe the methodologies that I will employ to integrate the data. The data generated herein will also be leveraged to shed light on the effect of pyrimethamine, an anti-malarial drug (AMD), on the host animals in the absence of malaria infection.

In this chapter, I first lay the foundation of the analytical framework which will be used in subsequent chapters for integrating some of the numerous data-types produced by MaHPIC cores. These include complete blood count (CBC) measures (taken daily), metabolomic data (various samples) and transcriptomic data (taken at seven points). The methods that I will use include variance component analysis, principal component regression, analysis of variance (ANOVA) for differential abundance analysis, and subsequent pathway enrichment. The methods chosen reflect methods used by recent large-scale projects for data integration, but are adapted due to underlying experimental design differences.[38, 29]

2.2.1 Effects of pyrimethamine

In addition to giving a detailed description of the methodologies that will be used in future MaHPIC experiments, I also investigate the effect of the anti-malarial drug

pyrimethamine on the host transcriptional profile at both the level of the individual transcript as well as the pathway scale. Pyrimethamine, a potent and front-line malaria treatment when paired with a sulfonamide, inhibits dihydrofolate reductase (DHFR), an enzyme responsible for the conversion of dihydrofolate (DHF) into tetrahydrofolate (THF).[137, 16] The resulting depletion of THF has important consequences for the cell, especially rapidly dividing cells like those of the parasite.[105] Folate-deficiency-induced reduction in the availability of nucleotide pool and subsequent cell cycle arrest of the parasite is the putative mechanism of action of pyrimethamine.[134] In spite of pyrimethamine's important position as a treatment of malaria, its side-effects on the host have not been fully elucidated.

The host, in this case the macaque, also possesses DHFR, and pyrimethamine will presumably inhibit its functioning, as well, leading to a host cell reduction in THF. THF is a co-enzyme for the production of three of the four deoxyribonucleotides, which are in high demand during the genome duplication of mitosis. THF is also a co-factor for the metabolism and methylation of some amino acids, and a reduction in THF therefore can deplete the amino acid pools of the cell and may also dysregulate post-translational modification of proteins.[9]

Yet another important consideration for pyrimethamine usage is that approximately 40% of cellular folate is present in the mitochondria, reflecting its importance in the functioning of this energy-producing organelle.[132, 85] Based on its underlying molecular mechanism or action, the administration of pyrimethamine is likely to have dramatic effects on the transcriptome of the host. To assess the magnitude and duration of effect of pyrimethamine administration, I will test the following hypotheses.

2.2.2 Motivating Hypotheses

With respect to the host response to pyrimethamine, I anticipate vast alterations in the transcriptional profile after AMD treatment. First, I hypothesize that both the

blood and the bone marrow expression profiles will be affected in the samples taken only seven days after AMD treatment. I anticipate alterations in the folate pathway and anticipate that many such genes will be up-regulated in the "treated" time points as the cell works to restore folate levels. Since pyrimethamine reduces the levels of THF and subsequently purines, a building-block for DNA synthesis, cells that need high levels of nucleotides (i.e. those that are dividing and copying their genome) will likely undergo apoptosis. At a minimum, cell cycle progression will be affected. A more pronounced effect is expected to occur in the bone marrow compared to the whole blood since the bone marrow contains many rapidly-dividing cell types whereas the blood is relatively post-mitotic.

Secondly, I hypothesize that in the two time points that are >30 days after the last AMD treatment ("inter," TP4 and TP6), both blood and marrow expression profiles will have returned to normal. The alternative hypothesis is that more than a month after AMD treatment, gene expression programs in marrow and blood are still dysregulated. A finding of this nature should inform decisions concerning mass administration of AMDs as a way of eradicating malaria versus targeted treatment of confirmed-infected individuals.

In this chapter, I begin by describing the analytical methods which will be used in this and subsequent chapters for integrative analysis. Next, I explore the correlation structure and variance components of the datasets. Finally, I describe findings regarding the transcriptional dysregulation that occurs and persists after pyrimethamine administration.

2.3 Methods and materials

2.3.1 Experimental design and measured outcomes

The Malaria Host-Pathogen Interaction Center (MaHPIC) is a multi-disciplinary investigation that utilizes multiple data types to better understand the systems biology

of the complex host-parasite dynamics in the course of malaria infection. For this control experiment, the design is as follows. Five male rhesus macaques (*Macaca mulatta*) approximately 2 years of age were profiled over the course of a 100-day experiment after being injected with purified mosquito salivary glands on day 0 of this control experiment. Complete blood counts were performed daily by the Malaria core members. Before injection, the timepoint 1 (TP1) samples were taken. Then, at days 20, 26, 53, 59, 89, and 96, members of the Malaria core collected blood and marrow samples for TP2-7, respectively. The transcriptome and metabolome were interrogated at seven (TP1-7) and five (TP3-7) milestones throughout the course of the experiment, respectively (Figure 4).

2.3.2 Metabolomics feature quantification

High resolution metabolomics was performed by the Metabolomics core using a liquid chromatography/mass spectrometry (LC/MS) approach on an Orbitrap Mass Spectrometer. Two different columns were used for the LC separation stage: C18 and anion exchange (AE). Each distinct biological sample was run in triplicate to ensure high reliability of the data, with randomization within batches. MS peaks were called using xMSanalyzer.[154] Standard quality control measures were performed, such that features with greater than 60% missingness or a coefficient of variation within replicates greater than 1 were removed from the analysis. Since the frequency distributions of all samples were comparable, no additional normalization was performed, but an abundance cut-off of 256 peak area units was adopted and all features below this were excluded. All downstream analyses utilized the individual samples rather than attempting to average or otherwise reduce the technical replicates to a single measure per biological sample. Each column generates in excess of 10,000 mass-to-charge (m/z) and retention time features, the majority of which are either not yet annotated or have ambiguous annotation to multiple possible organic compounds.

The m/z features include the majority of known components of central metabolism, as well as xenobiotics.

2.3.3 Library preparation for RNA-seq

Bone marrow (1ml, BM) was collected by the Malaria core into 1.5 ml tubes with EDTA, and the mononuclear cells were purified by density gradient centrifugation on Lymphoprep (Stem Cell Technologies) solution and preserved in RLT buffer (Qiagen) to stabilize mRNA. Whole (peripheral) blood (3 ml, PB) was collected by the Malaria core team members in Tempus tubes (Applied Biosystems) that also preserve mRNA; these samples include erythrocytes, platelets and granulocytes in addition to mononuclear lymphocytes. RNA was extracted from the BM samples using Qiagen RNEasy Mini-Plus kits following the manufacturer-recommended procedures, and from PB samples using Tempus-Spin RNA isolation kits. The quality of all RNA samples was confirmed using a Bioanalyzer, with an RNA Integrity Number (RIN) greater than 8 recorded for all samples.

Approximately 1 g of total RNA per sample was converted to double-stranded cDNA using poly-A beads to enrich for mRNA, using Illumina TruSeq Stranded mRNA Sample Prep kits to generate strand-specific libraries. As a quality control, 96 spike-in RNAs of known concentration and GC composition (ERCC Spike-In Control, Life Technologies) were added to constitute approximately 1% of the total RNA for each library. Adapters were ligated to facilitate 3-plex sequencing on an Illumina HiSeq2000 at the Yerkes Genomics Core, aiming for 80 million paired-end 100-nucleotide (nt) reads per library.

2.3.4 Short read mapping

To quantify gene expression, the RNA-Seq reads were mapped to the most recent available rhesus macaque genome (MacaM assembly, Version 4.0, GenBank accession number PRJNA214746 ID: 214746, created by Aleksey Zimin at the University of

Maryland, Rob Norgren at the University of Nebraska Medical Center, and their colleagues) using Tophat2.[151, 74] Default options were used with the exception that the command –library-type fr-secondstrand was used since the reads were generated using a stranded library preparation method from Illumina. This allowed us to differentiate between sense and antisense transcripts. We also provided an annotated reference transcriptome which was supplied with the *M. mulatta* genome (Version 4.12) which improves the mapping accuracy across splice junctions. Only reads that map to a single location in the genome were included, to ensure high-confidence mapping. All downstream analyses were performed at the level of annotated gene: this study does not consider exon-specific or transcript isoform relative abundance.

Several quality control steps were used to verify the reliability of the data: linear correlation of estimated abundance of ERCC spike-in controls with known concentration; confirmation of 99.9% strand-specificity of the controls; less than 0.1% control fusion transcripts; and absence of 3 bias in the controls was confirmed with RSeqC software. Transcript abundance levels were inferred using HTSeq v0.5.4. HTSeq takes the short-read mapping file (bam) from tophat2 and the gene annotation file which contains the locations of all annotated genes. Since some libraries were sequenced more deeply than others, the libraries were normalized before determining differential gene expression using the gene level expression files with the default parameters of DESeq version 1.10.1.

2.3.5 Gene expression quantification

After quality control steps verified the reliability of the data, we quantified the gene expression levels using htseq v0.5.4.[135] HTSeq takes the short read mapping file (bam) from tophat2 and the gene annotation file which contains the locations of all annotated genes. We obtained the most updated macaque gene annotations from rhesusbase. Since some libraries were sequenced more deeply than others, the libraries

were made comparable (normalized) before determining differential gene expression between libraries. Gene expression normalization was performed using the library size estimation procedure implemented by DESeq version 1.10.1, available in the bioconductor suite in R.[6, 7, 51, 146] Briefly, the software calculates the ratio between gene expression for a given gene against the geometric mean of all samples in the study. It then finds the median value across all genes for each individual and uses it as the library size factor.

2.3.6 Variance component analysis

After data normalization, the transcriptome and metabolome levels were log-2-transformed and imported into JMP Genomics (version 6.0). To determine how much of the variance in each of the datasets is explained by the two measured factors (animal and time), I performed a principal components (PC)-variance component analysis using JMP v6.0 (SAS) for the transcriptome, metabolome, and the CBC data. This consists of the generation of all PC explaining up to 90% of the total variance (12 to 15 for the transcriptomes and 30 for the metabolomes), regressing each PC on Animal or Timepoint, and generating a weighted average of the squared correlation coefficient (percent variance explained) across all of the PC. Since the low abundance features for metabolomics and transcriptomics both have high coefficients of variation, I set thresholds of 5 log2 units for transcripts and 17 log2 units for metabolites (Supplementary Figure 1) and removed lower abundance features to determine their effect(s) on the analysis.

To assess whether the major PC capture similar aspects of the data, the first 10 PC were calculated for the four omics datasets (PB and BM transcriptomes, C18 and AE metabolomes) using JMP. All 780 pairwise correlations of these PC values were determined, and a Bonferroni multiple comparison adjustment was used to assess the significance of each pair of PC. Exploratory partial least square regression analyses

were also performed with MixOmics in attempt to select variables that co-vary, but did not reveal significant associations.

2.3.7 Differential gene expression (DGE)

The next step in the analysis was the identification of genes that are differentially expressed across the experimental conditions. For between-TP differences, an analysis of variance (ANOVA) was performed on each transcript separately using "animal" as a random effect with 5 levels and "timepoint" with 7 levels, or "drug" with 3 levels as the fixed effect. For the drug exposure factor, I define our three experimental conditions as before drug exposure (Pre-drug; TP1 and TP2), 7 days after the most recent dose (Post-drug; TP3, TP5, and TP7), and 30 days after most recent dose and immediately before the next dose (Inter-drug; TP4 and TP6), as shown in Figure 1. A false discovery rate cut-off of 5% was used to define differentially expressed genes. These were examined using hierarchically clustering of the standardized least squares means and volcano plots of significance against fold difference between specific conditions.

2.3.8 Gene set enrichment analysis

Gene set enrichment analysis GSEA is most commonly performed in one of two ways. In the first method, a statistical threshold is set for a list of genes (usually a Bonferroni-corrected $p = 0.05$); any gene with a p-value lower than the threshold is included and a hyper-geometric test for enrichment with other gene sets is calculated.

In the second method, the entire gene list is used, and genes are ranked using some ranking metric, which is often the t-statistic since it measures both magnitude and direction of effect in a two-sample comparison.

The second method has the advantage that its results are not dependent upon the threshold of choice. That is, if one researcher chooses a threshold of $p = 0.05$ whereas another chooses $p = 0.01$, the significance levels for the gene set enrichments

will likely differ. This can effect downstream interpretation.

One of the drawbacks for the second method is that the statistical significance value has to be calculated empirically from permutations of the data. This requires running hundreds or thousands of permutations per gene set to obtain empirical p-values. However, hundreds of gene sets can be interrogated in a matter of minutes with a standard desktop computer, and so the time considerations were not limiting in this case.

To perform this analysis, I chose the ranked gene list method and used the Broad Institute's GSEA v2.0.14 to perform enrichment analysis.[142] For the two contrast of interest (pre-versus-post, and pre-versus-inter), we performed gene set enrichment using the KEGG pathways and GO terms, separately for each tissue. The t-statistic was used as the rank metric and was obtained from the JMP output file. Gene sets with an FDR $< 25\%$ were considered as significant in accordance with the recommendations of the GSEA software manual. Default parameters were used and included the removal of gene sets with more than 500 or less than 15 genes.

To perform GSEA, *a priori* defined gene sets are needed. Since the rhesus macaque is much less well-studied than the human genome, and considering that the majority of genes in the macaque genome have well-conserved syntenic orthologs in the human genome, I used the pre-existing human gene set annotations for this analysis. These genes sets were obtained from the Broad Institute's website.

2.3.9 Blood informative transcript (BIT) axes

In addition to principal component analysis, we employed a second method, blood informative transcript (BIT) axes analysis, which uses an *a priori* defined set of 9 blood axes, which are composed of genes that covary across many blood transcriptional profiles.[114] This method has been described elsewhere.[102] Briefly, we took

the genes that make up each of the axes and calculated the BIT score for each individual for both the blood and the marrow, separately, using the normalized expression data. Then, we performed one-way ANOVA to test for significant differences across both animal and time. Since the between-subject variability was so large, we also performed the analysis with the animal residuals, that is, the error term in the model after fitting the effect of animal. We then examined the dynamics of the axes scores over time as before.

2.4 Results

2.4.1 Subject-specific effects dominate variance components in the transcriptome

After normalization and quality assurance of the datasets from the three data types across the 100-day experiment (Figure 4), I first explored each dataset independent of the others. First, I performed variance component analysis of each of the datasets to attribute the variability of the data to either residual (unexplained) variance or one of the two factors, animal and time. A previous study described high within-subject conservation of expression of a subset of genes from whole blood in humans,[161] but it is unclear how much of the variance in the transcriptome is due to inter-individual variability. I expect to see a moderate to large proportion of the transcriptomic data sets explained by subject effects. To my knowledge, no variance component analysis has been performed for a metabolomic dataset of this size; I, therefore have no *a priori* expectation for metabolites.

For the transcriptomes of both tissues, PB and BM, as well as the CBC data, subject effect accounts for a large component of the variance (more than 30%), which is much larger than the time effect (Figure 5A). That is, greater than 30% of the major variance components are explained by subject-to-subject variation. The within-subject variance in the metabolomics data is much lower (approximately 10%). The smaller proportion of variance explained by the factor of animal in the metabolome

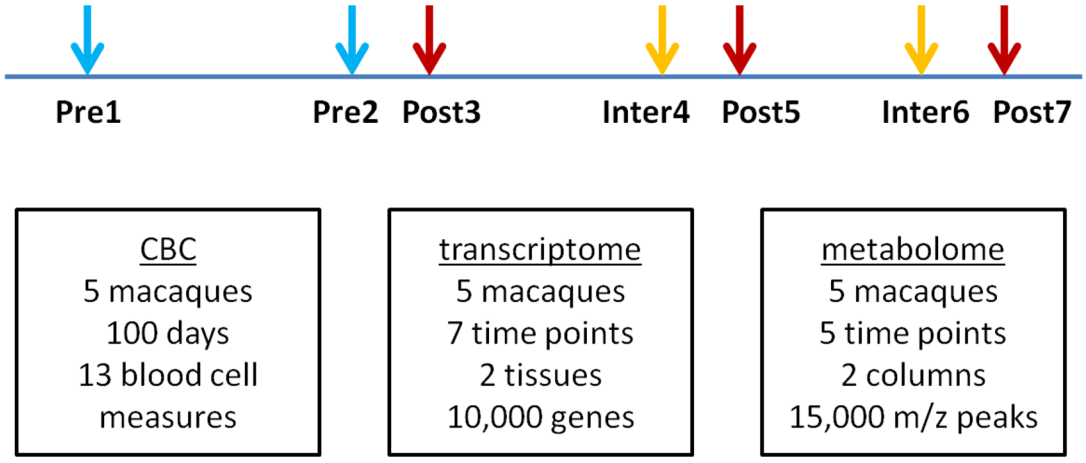

Figure 4: Experimental overview and the structure of the data. CBC measurements were taken daily for the course of the 100 day experiment. RNA for transcriptome analysis was extracted from blood and marrow at seven time points (TP1-7). Blood plasma was sampled at the last 5 time points for metabolomic analysis (TP3-7). The major sampling time points were selected based on the estimates of when sampling would occur in the infection experiments. The light blue time points (TP1 and TP2; pre) are made before the first administration of pyrimethamine. The red arrows (TP3, TP5, TP7; post) are samples taken 7 days after the start of pyrimethamine treatment. The yellow arrows (TP4 and TP6; inter) indicate samples taken more than 30 days after the last pyrimethamine dose.

compared to the transcriptome and the CBC data suggest that there is some buffering of metabolic alterations in spite of between-individual differences at the level of transcript abundance.

A recent study contrasting the variance in the transcript and the proteome across different primate species showed that in spite of alterations at the transcript level, protein levels were often much closer across species. If the same buffering capacity of the cell holds true for within-species differences in transcription, I would expect less variance at the protein level compared to the transcript level with respect to the between-animal differences. It would then follow that more similar levels of protein would lead to more similar metabolite levels between individuals.

For all five datasets, the time effect is close to 10% of the variance. In most of the datasets, more than half of the variance in the data is unexplained by either animal

or treatment. This residual variance could be explained by either technical variance introduced by errors in measurement, unmeasured biological factors, or effects which we are unable to estimate such as subject-by-treatment interaction terms. Specifically, each macaque may be responding uniquely to the drug treatment based on its underlying genetic background. Without replication, however, these effects are statistically impossible to estimate.

As others have noted, the coefficient of variation was dependent upon the mean of the feature for both the transcriptomic and the metabolomic datasets with lower feature abundance tending to have higher coefficients of variation (Figure 6).[152, 6] I suspected that this was due to the fact that low abundance features were poorly estimated, at least in the transcriptomic data. This effect would result in higher technical variance in the lowly expressed features, and therefore a reduced proportion of the variance would be explained by our two factors, animal and time. To test this hypothesis, I removed all genes with expression lower than an average normalized level of five, and for metabolites, the cut-off was set at 17; this appeared to be the close to the elbow-point of the graphs.

After removing lowly expressed genes from consideration and assessing this reduced dataset, the amount of residual variance not explained by the two factors, animal and time, decreases sharply for the transcriptomic datasets; this drop in residual variance is due primarily to an increase in the variance explained by the factor of animal (Figure 5B). This suggests that the variance of the least expressed genes is more heavily influenced by technical variance. In contrast, the metabolomic datasets did not experience such an increase in variance explained after removing low abundance features. Both subject and time explain a similar amount of the variance before and after data reduction for the metabolomic datasets (Figure 5B). This finding shows that the low abundance metabolites compared to high abundance metabolites do not have increased variability due to measurement error. The fact remains, however, that

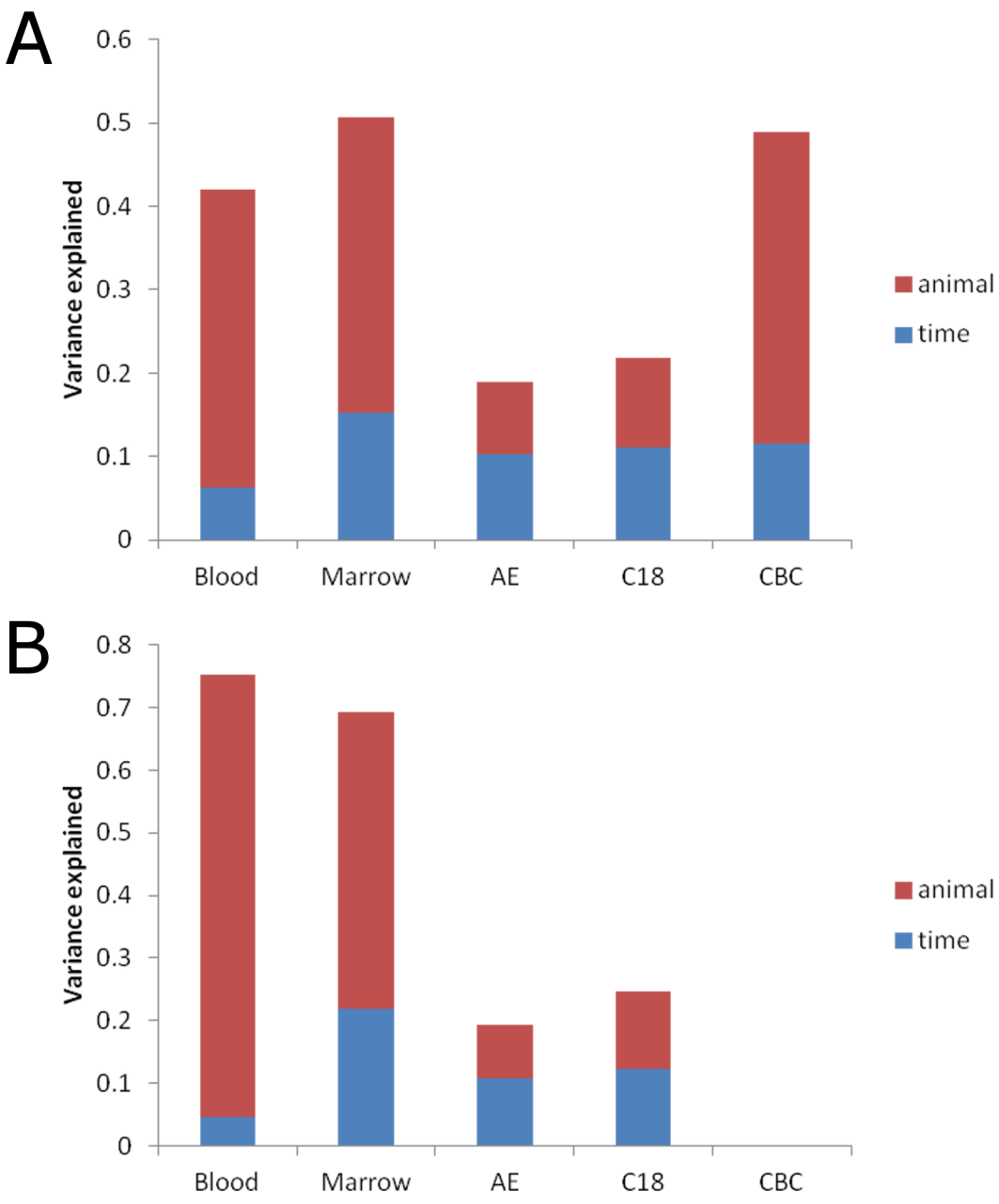

Figure 5: Variance component analysis (VCA) for the three data-types. The effect of animal (that is, the between-individual variance) is more than 30% of the variance for the two transcriptomes and the CBC data in the full data sets (A). Whereas the animal effect is much less prevalent in the metabolomic datasets. In the reduced datasets (B), the factor of animal explains much more of the variance. Note: a reduced dataset was not generated for the CBC data since it only had 13 measured features.

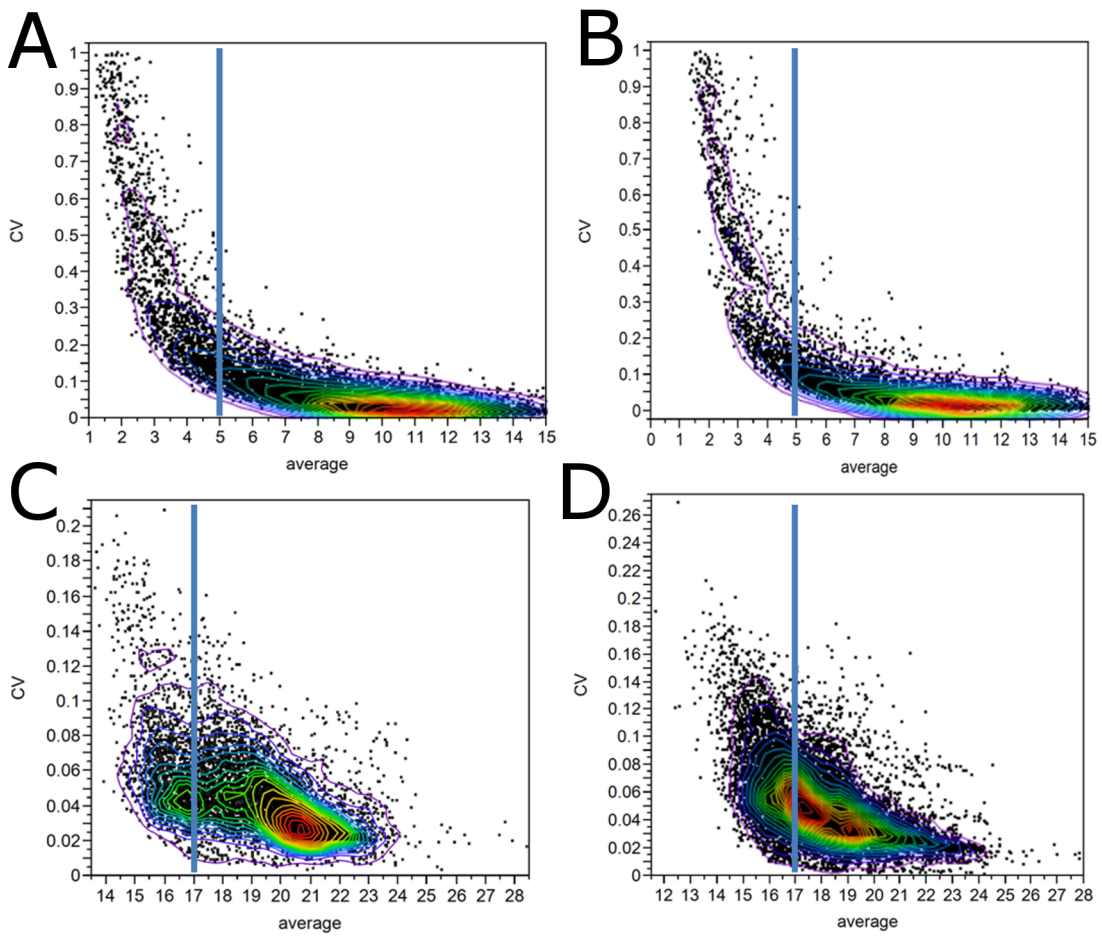

Figure 6: Coefficient of variance changes as a function of mean feature value. For both of the data-types, there is a relationship between the coefficient of variation and the mean value for each feature, stemming from increases in technical and/or biological variability at the lower end of the measurement spectrum. Each point represents a feature: either a gene or ion peak. Density contours (for every 5%) are included and show that most of the features are maintained after excluding lower abundance features. A line is drawn at x=5 and x=17 for the transcriptome and metabolome, respectively. Features lower than this level were excluded in the reduced dataset used in calculating the VCA. (A) marrow transcriptome; (B) blood transcriptome; (C) metabolome with AE column; (D) metabolome with C18 column. Note: the domain for the transcript average level was trimmed to [0,15] and transcript CV from [0,1] for figure clarity.

lower abundance metabolites have a greater variance, which perhaps indicates that higher biological variance (that is not correlated to either of our measured variables) is likely more tolerated for low abundance metabolites.

2.4.2 Hierarchical clustering

As the next step in my top-down integrative approach, I performed hierarchical clustering of the transcriptome and metabolome. Hierarchical clustering allows the visualization of similarity between both subjects and measured variables. To visualize the relationship between the samples in our study, I hierarchically clustered the 70 transcriptomic samples (2 tissues * 5 macaques * 7 time points). The first and deepest division was between tissues: all bone marrow samples clustered together and all whole blood samples clustered together (not shown). Within the blood, each macaque clustered with itself (Figure 7A). That is, the seven time points for a given macaque were more closely related to themselves than to similar time points from other macaques. This is borne out in the principal variance component analysis in which 36% of the variance in the blood is explained by the animal while only 6% is explained by time.

Similar to blood, within the marrow, the within primate samples tended to cluster together. One notable exception was TP4, approximately one month after the first AMD treatment. Many of the samples from TP4 cluster together (not shown). Once again, the difference between blood and marrow is reflected in the principal variance component analysis: within the marrow, time, as a factor, makes up a much large proportion of the total within-tissue variance (more than 15%) than it does within blood (about 6%).

In the blood, there is a tight clustering between the individual macaques. Since each blood cell type has a characteristic expression profile, which allows it to perform its specified role(s), I hypothesized that the macaques that clustered together in the expression profile would also have similar levels of the major cell types. However, upon clustering the samples on the CBC data, I do not observe such a trend (Figure 7B). Therefore, I conclude that the CBC is capturing information about the system that is non-redundant with the transcriptome. This result is particularly striking

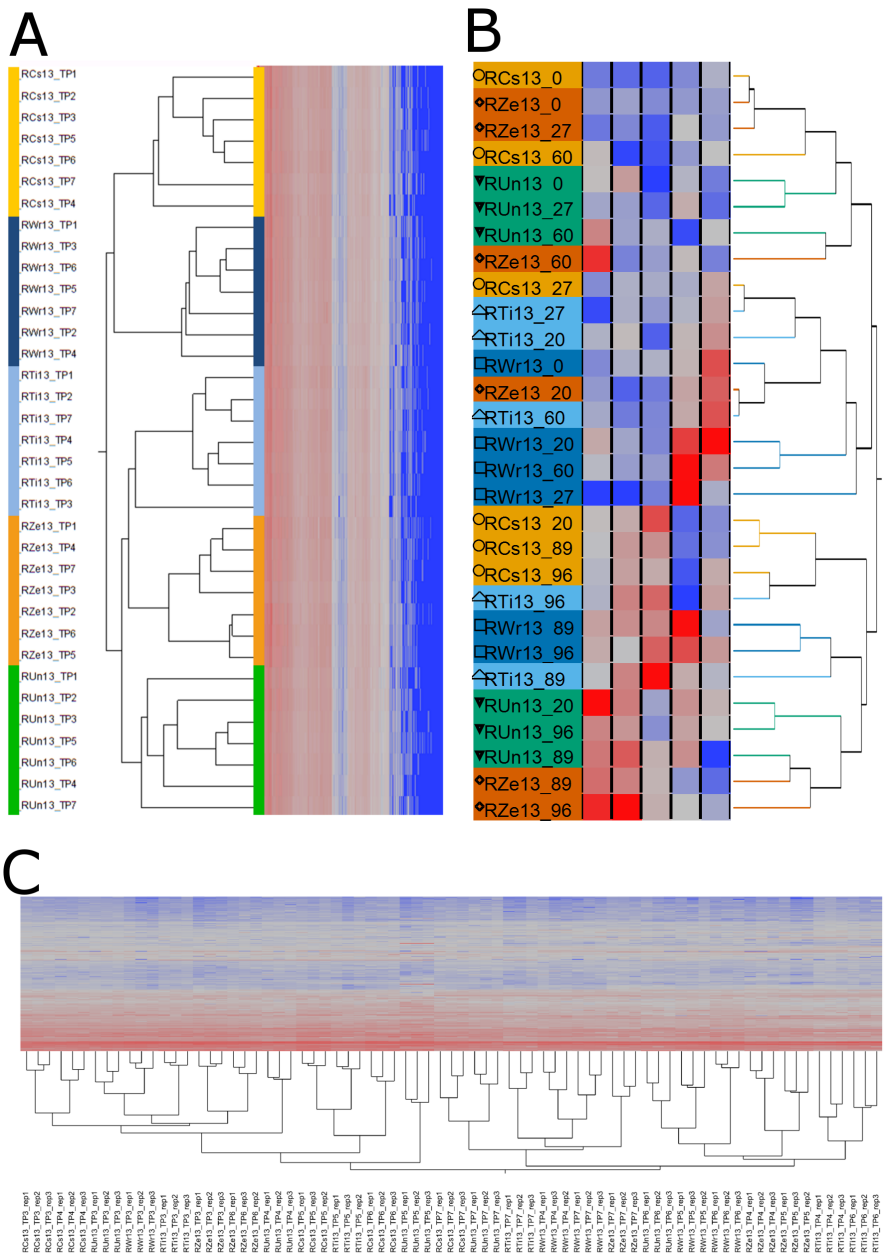

Figure 7: Hierarchical clustering omics datasets shows lack of correspondence. (A) Heatmap of the blood transcriptome with each primate uniquely colored. (B) CBC hierarchical cluster for counts of five cell types: RBCs, platelets, monocytes, lymphocytes, and granulocytes. (C) Heat map of metabolomic data for the C18 column.

when considering that both the transcriptome datasets as well as the CBC dataset have the variance component of animal explaining more than 30% of the variance.

Within the metabolomic datasets (Figure 7C), the clear separation between individuals seen in the transcriptome does not occur. Importantly, each of the three technical replicates for each biological sample cluster together. This finding gives confidence that the machine read-out is capturing reproducible differences that occur between the samples. While some clustering between time and animal appear in the metabolome, it is not as clear as in the transcriptome. This is likely because subject and time effects explain similar levels of the variance in these datasets, about 10% each in both columns.

2.4.3 Covariance-based approach for data-type integration: principal component regression

One commonly used variable reduction strategy is principal component analysis, which finds the orthogonal vectors of maximum variance in a dataset. To assess the amount of information shared between the transcriptomic and metabolomic datasets, I regressed the first 10 principal components of each dataset against each other (Figure 8). The pattern that emerges is informative in many ways.

Firstly, it shows that the two metabolomic datasets are highly correlated (Figure 8, bottom right box). This observation was expected since the two columns measured metabolites from the same plasma sample; the difference between the two datasets is the use of different LC columns which optimized peak resolution across a wider range of metabolites. PC1 and PC2 for the two sets are significantly correlated; many lower PCs are also correlated. Unlike the metabolomic datasets, the two transcriptomic datasets, marrow and blood, do not show as much correlation (Figure 8, top left box). Upon statistical analysis however, we see that PC1, PC2, PC3, and PC4 in blood are significantly correlated with PC2, PC4, PC1, and PC5 of bone marrow, respectively (Bonferroni corrected $p < 0.05$). Such a result is not unexpected considering that the

two tissues have different functions yet one (blood) is composed of cell populations derived from the other (marrow). One difference between the tissues is that the marrow contains many cell types that are rapidly dividing whereas most of the cells in the blood are post-mitotic and terminally differentiated. In some cases the sign of the regression is negative, but this is simply a function of PCA which commonly reverses signs and order of PC due to sampling variance.

Strikingly, there is no significant correlation between the transcriptome PCs and the metabolome PCs after multiple testing correction (Figure 8, top right (large) box). This could be explained by the fact that the transcriptome of these two tissues is contained within the cell whereas the metabolome interrogated the plasma composition. Furthermore, the plasma is not only influenced by metabolites from blood cells, but also receives metabolites from all tissues in the body. As a result, these measurements represent a "whole-body" average, whereas the transcriptomes are tissue-specific, sampled from cells in specific body compartments. This lack of correlation between the major components of the metabolome and the transcriptome demonstrates that these datasets are interrogating different components of the host system.

2.4.4 Blood informative transcript (BIT) axes of variation

As a complementary data reduction method to principal component analysis, I next examined the the between-animal differences and trajectory of previously identified BIT axes of variation.[114] These axes covary with some clinical parameters (e.g. abundance of specific cell types) and are enriched for gene ontology terms (e.g. viral response, and B-cell activation). To reduce the number of features of the transcriptome examined as well as to assess the impact of normalization, I calculated the BIT axes scores for these 9 pre-defined axes for the transcriptome dataset using both the original data and the expression levels after fitting the effect of animal in a linear

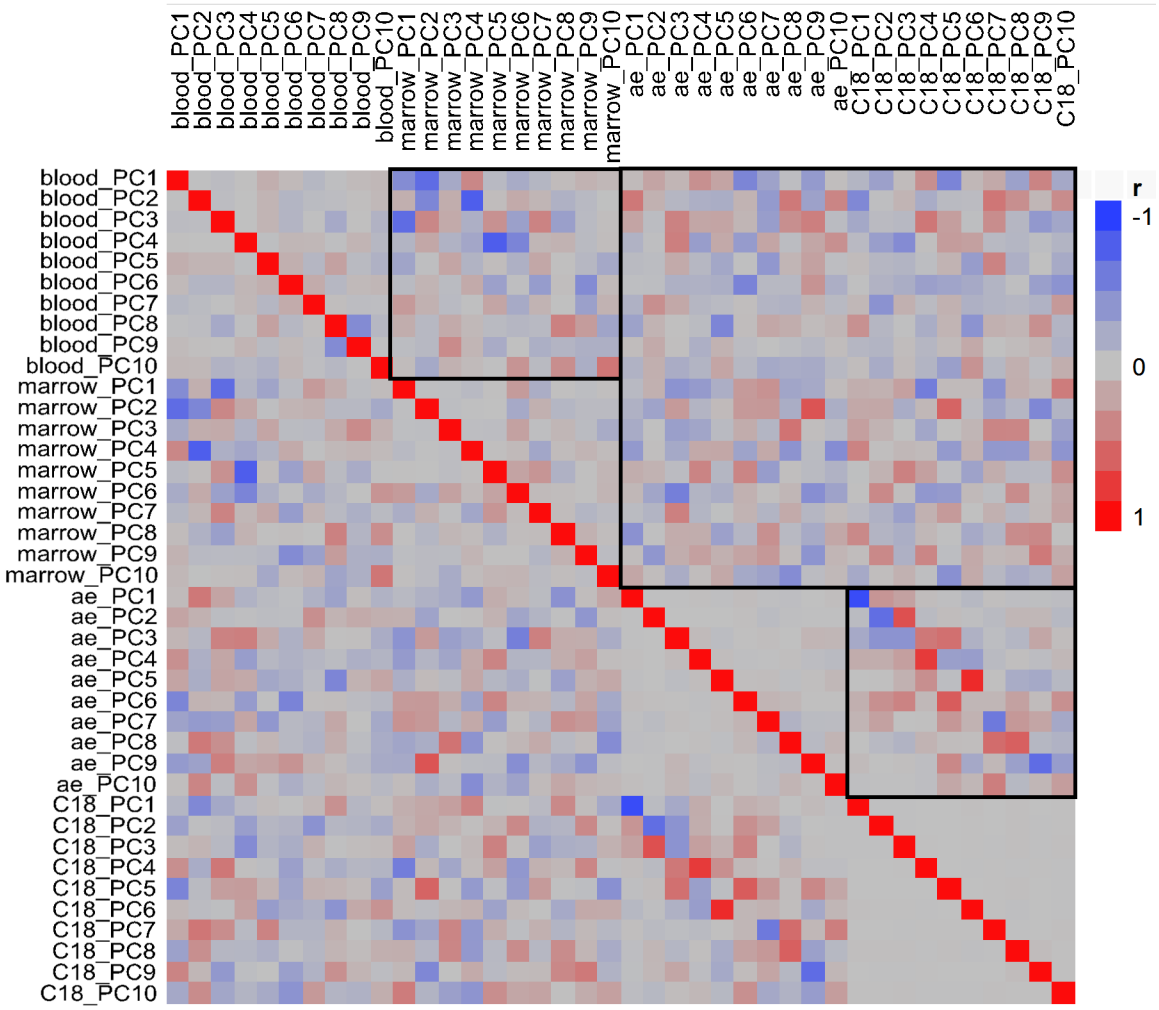

Figure 8: Heatmap of the correlation of the first 10 principal components of the transcriptome and the metabolome. In the box on the top left (the correlation between the two transcriptome datasets), there are some significant correlations between the PCs, although there is not a clear one to one mapping (e.g. PC1 for marrow does not correspond to PC1 for blood). In the correlation between the metabolomic datasets (bottom right box), the significant correlations lie more clearly on the diagonal, demonstrating that these datasets are capturing similar information. Between the two data-types (large box, top right), there are no significant correlations after Bonferroni adjustment.

model. Using these *a priori* defined gene sets, I found that the axes vary significantly across both animal and time point (Figure 9). Axes 2-8 for blood and axes 1-3, 5,7,8, and 9 for marrow were significantly different across the five macaques which shows that BIT axes are relatively stable over time within an individual (Figure 9A-C).

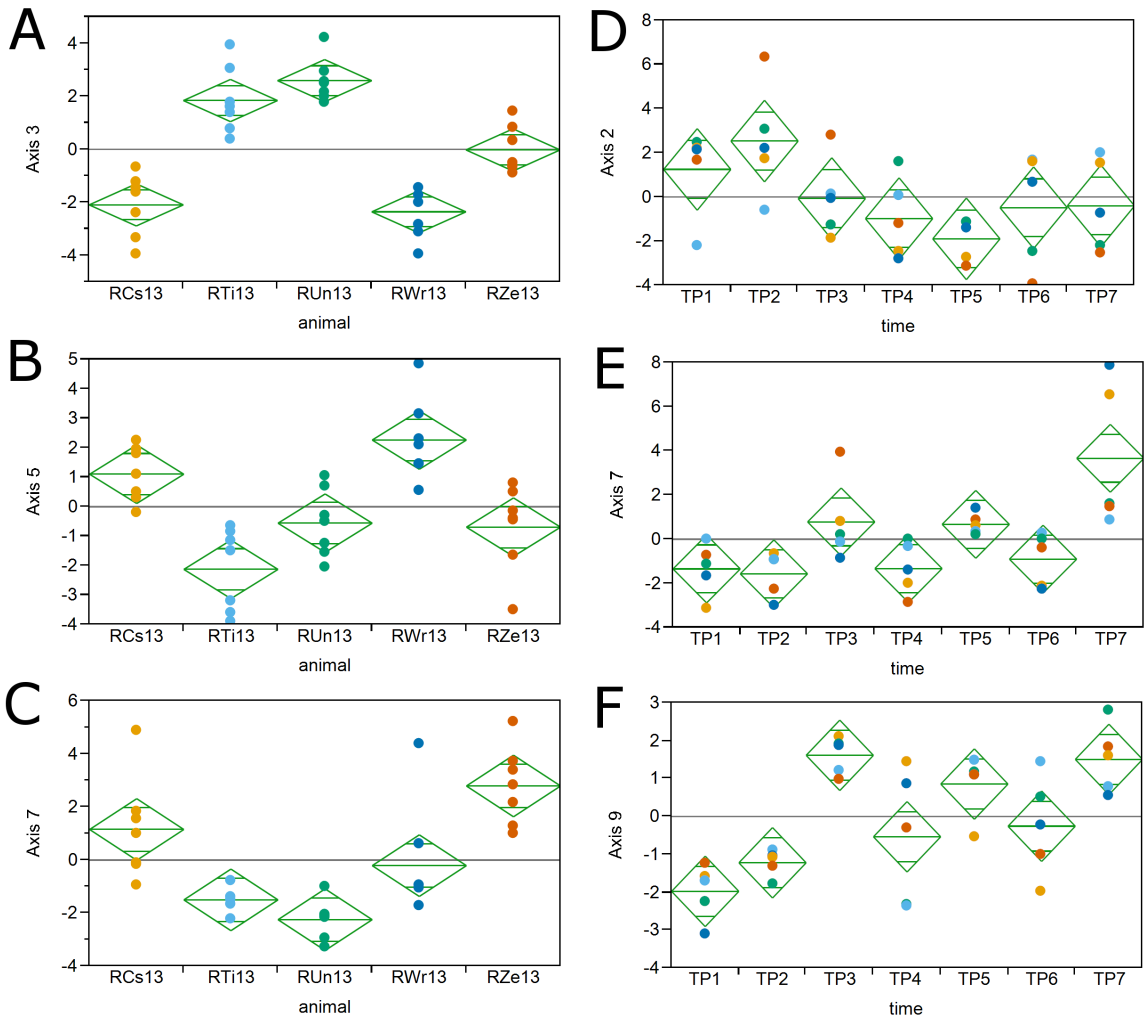

Figure 9: Blood informative transcript (BIT) axes of bone marrow expression as a function of animal and time. Panels (A-C) show the BIT axes 3, 5, and 7, respectively, which were highly significantly different across the five macaques. Panels (D-F) show three significantly differentially regulated axes (2, 7, and 9) which appear to show coherent cycling as a function of the anti-malarial drug dosage.

After removing the effect of animal from the axes, most of the axes change significantly across time in both the marrow and the blood with some of them varying in a manner coherent with the times of treatment with the anti-malarial drug (Figure 9D-F). While it is difficult to assign statistical significance to the coherence of the cycling of axis scores, qualitatively it appears that many of the axes respond to the drug in predictable ways. For instance, in the marrow, I find that axes 2, 7, and 9 are

all behaving in a manner consistent with expectation and thus supporting the value of this data reduction approach for blood and marrow transcriptomics datasets.

Axis 2 genes (Figure 9D), which are related to hematopoesis, were more highly expressed at the first two time points but have decreased after the first pyrimethamine administration, and remained low for the rest of the experiment. Both axis 7 and axis 9, associated with viral response and programmed cell death, respectively, appear to activate in response to pyrimethamine. The later finding is concordant with the gene set enrichment analysis results, which will be present later.

2.4.5 Differential abundance of genes, metabolites, and blood cell counts

To perform gene set enrichment, I first must identify genes that are differentially expressed. There is currently much debate as to the best method for identification of differentially expressed genes for RNA-seq data.[118, 139] Since most RNA-seq studies have a paucity of samples, information about the variance of expression must be borrowed across all genes. In some methods, low variance genes have their variance inflated to account for the possibility of the low variance being due to chance. This adjustment has the property of making the analysis more conservative and therefore less sensitive to differential expression.

In this case, I have many samples, and so herein I use an analysis of variance (ANOVA) approach to determine which features of the transcriptomic and metabolomic datasets are altered over the course of the experiment. To utilize the findings made previously in this study, I explore the impact of including "subject" as a random effect in our significance analysis. I found that there are many genes and metabolites that are differentially up- or down-regulated as a function of time (table 1). Importantly, without including the animal effect in the model there are many fewer features identified as differentially abundant in all of the data-types.

In the examination of the hierarchical clustering of the differentially expressed

Table 1: Significant features for the effect of time. For each dataset, the number of significant features identified by the ANOVA test is shown for either fitting the subject effect or not fitting it.

Data type	tissue	fitting subject	not fitting subject
RNA-seq	whole blood	292	0
RNA-seq	bone marrow	6483	3678
Metabolome (AE)	plasma	3951	1394
Metabolome (C18)	plasma	7113	3487
CBC	whole blood	13	10

genes for the marrow, TP4 is most different from the other time points (Figure 10A). I also noticed a similar trend for TP5 in blood (Figure 10B). As a result, I wondered if the marrow transcriptome at TP4 contributed to the blood transcriptome at TP5. To assess the significance of the overlap, I extracted the genes that were up-regulated in TP4 of marrow. I then performed a sign-test (binomial test with probability of success = 0.5) to see if those genes were more likely to also be up-regulated in the TP5 of blood. The results were highly significant ($p < 2.2 * 10^{-16}$). I performed a similar analysis for the down-regulated genes and found a similarly significant result. As a control, I then took a comparison of TP6 and TP7 in blood for the same up- and down-regulated genes and did not find an enrichment ($p = 0.74$). This shows differential gene expression in marrow is reflected in blood with a time lag. Further, since blood had many fewer significant genes that marrow, I also conclude that gene expression in blood is buffered against differential expression in response to pyrimethamine compared to marrow.

After investigating the changes in expression over the course of the experiment with each time point considered separately, I next performed differential gene expression as a function of drug treatment using the three groupings described in Figure 4: pre, post, and inter. In the marrow and blood, there were 6483 and 0 differentially expressed genes, respectively, at an FDR=5%. This result is qualitatively similar to the analysis performed by time point, where marrow had many more significant genes

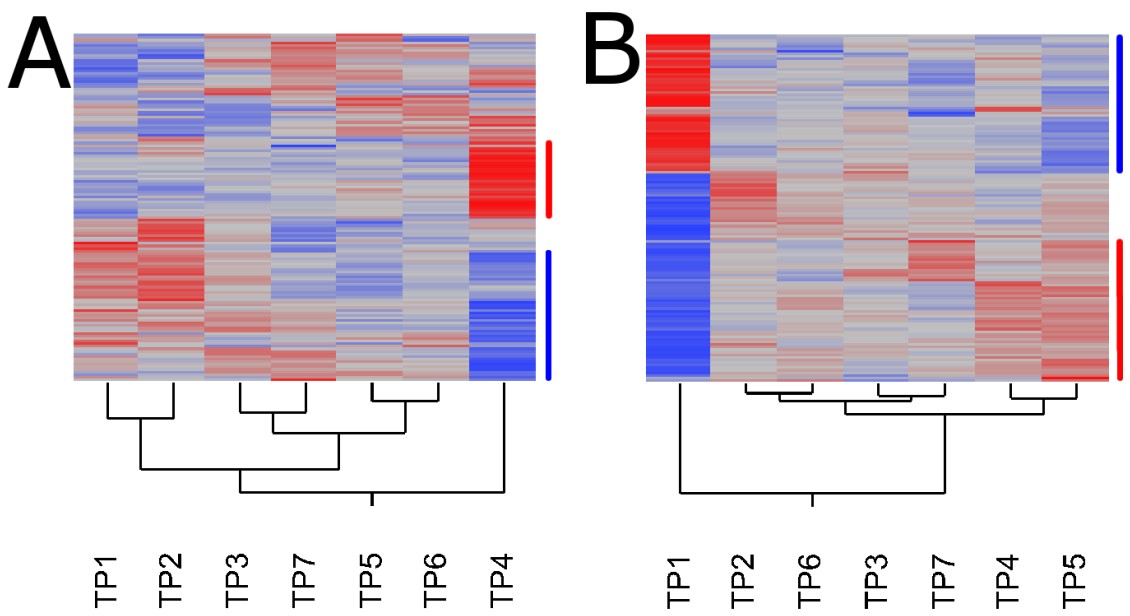

Figure 10: Clustering of the significant genes as a function of time. (A) Marrow and (B) blood gene expression levels clustered by time point. Red (blue) lines indicate the genes that clustered together and were up-(down-)regulated in the two transcriptomes.

that blood.

In the metabolome, the two different columns, AE and C18, had 1,452 and 3,011 differentially abundant features (out of a total of 15,728 and 20,767, respectively). While feature resolution for metabolomics has grown by orders of magnitude in recent years, the ability to assign putative compound names to each of these detected features has lagged behind. As a result, no pathway analysis was performed for these many differentially abundant features. However, the large numbers of adducts that are different across the drug treatment regimen leads me to conclude that the metabolome has been dramatically affected by pyrimethamine. A notable feature of the metabolome is the absence of any adducts with a mass-to-charge (m/z) ratio that corresponded to the administered drug, pyrimethamine.

2.4.6 Gene set enrichment analysis

After identifying the differentially expressed genes, I then performed gene set enrichment analysis on the two datasets, marrow and blood. [1] The most significantly enriched pathway in the marrow for the pre-versus-post comparison was the KEGG pathway of one-carbon cycling by folate; this was also the second-most significant pathway in the pre-versus-inter contrast (Figure 11A). The enrichment of this pathway shows that when the enzyme DHFR is inhibited by pyrimethamine, there is a subsequent down-regulation of the other genes in this pathway. Interestingly, genes in this pathway are not enriched in the blood.

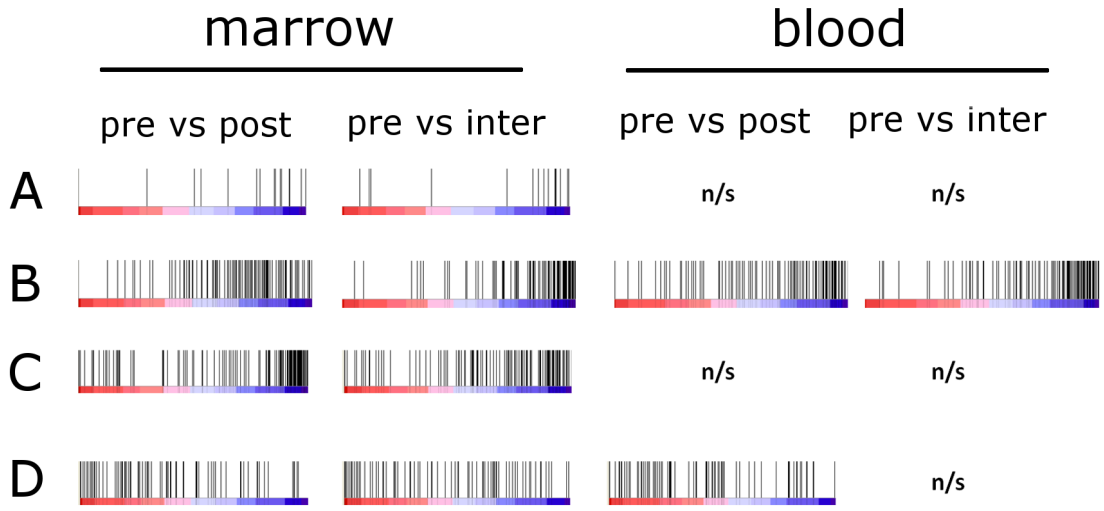

Figure 11: KEGG gene set enrichment plots for four representative pathways. The bar that transitions from red to white to blue indicates the value of the t-statistic; red signifies genes that are highly expressed in either the post- or inter-drug treatment, whereas blue signifies genes that are highly expressed in the pre-drug treatment. Each vertical black line is the location of a gene in the specified pathway. (A) One carbon cycling by folate; (B) oxidative phosphorylation; (C) cell cycle; (D) apoptosis. n/s signifies that the enrichment was not significant at FDR=5%.

Pyrimethamine inhibits DHFR which then leads to decreases in folate cycling

[1] In spite of the fact that there were no significant differentially expressed genes in the blood, gene set enrichment is still possible using the pre-ranked list method of GSEA.

and reductions in the folate pool. Folate, which plays an essential role in nucleotide synthesis, also is needed for energy generation in the mitochondria (Figure 11B). It comes as no surprise, therefore, that genes in the oxidative phosphorylation pathway are also down-regulated after pyrimethamine treatment both seven days as well as 30 days after administration, and the down-regulation occurs not only in the marrow but also in the blood.

As mentioned above, folate is a required cofactor for the synthesis of some nucleotides, as well as the methylation of proteins and amino acids. With insufficient production of nucleotides, the rapidly dividing cells of the bone marrow would be unable to proceed through mitosis at normal rates, or cell division may stop completely. I report that both seven days and 30 days after pyrimethamine treatment, genes related to cell-cycle are down-regulated in the marrow (Figure 11C). Since the blood is mostly a post-mitotic tissue (i.e. most cells are terminally differentiated) there is no significant enrichment of cell-cycle-related genes in this tissue.

Lastly, with an inability to produce sufficient energy and/or a stalling of cellular division, a cell may enter a state of apoptosis. If this were the case, there should be an up-regulation of genes related to programmed cell death, which is what is observed (Figure 11D).

2.5 Discussion

2.5.1 Methodology for data analysis and integration

In this chapter, I laid the groundwork for the statistical and analytical techniques that allowed me to integrate the various data-types available for this experiment. In applying these methods to a group of macaques that will serve as a control, I demonstrated that pyrimethamine, a common anti-malarial drug, has lasting effects on both the blood transcriptome as well as the bone marrow.

In the variance component analysis of this study, I was able to account for much of

the variance in the transcriptome data especially after removing low abundance transcripts (Figure 2B). In spite of the ability to account for such a substantial proportion of the variance, the residual variance is still around 25-30%. This residual variance may be due to many factors. Besides the natural variability in gene expression and other unmeasured biological covariates, this residual variance may be due to primate-specific responses to the AMD treatment; that is, there may be an individual-by-drug-treatment interaction effect in which one primate may have a stronger response to the AMDs due to its underlying genetic background. In a similar experiment in Drosophila, we found that there was a strong genetic main effect (analogous to subject, in this case), as well as a moderate gene-by-environment interaction in the transcriptome. And the effect of the environmental perturbation was relatively small. While the interaction effect cannot be measured in this study due to lack of replicates, this proposed explanation is consistent with the observations in this study.

2.5.2 Between-subject effects

In the exploration of the various data-types, I showed that the variance of both the transcriptomic datasets as well as the CBC data have very strong between subject effects. This finding underscores the importance of repeated measures to increase statistical power in differential gene expression analysis.

The high level of within-individual conservation compared to between-individual differences is under-appreciated in the analysis of many biological data types. The importance of repeated measures (and subsequent inclusion of this term in the statistical model) is shown in my analysis in which I identify differential expression with and without including animal in the model. Perhaps the most data-rich gene expression studies with longitudinal sampling of individuals comes from the Storey lab.[38] Without explicitly describing the inter-individual variation in gene expression profiles, they implicitly control for it by looking at rates of change of profiles over

the course of many days. Unfortunately, these kinds of profiles (and rates of change thereof) depend critically on sufficient temporal resolution so as to be able to determine rates of change. In this case, there was not the required temporal density of data that would have allowed for this kind of analysis. However, explicitly accounting for the inter-individual variation was sufficient to resolve important patterns in the transcriptome.

In spite of this high within-individual conservation of gene expression, there is no analogous individual-wise clustering of the CBC when using the major cell types. This implies that the genes that are differentially expressed across individuals are not differentially expressed across blood cell types.

My approach to integrate the transcriptome and metabolome, principal component regression, yielded a negative result. While I observed good correlations within the datatypes (i.e. blood and marrow transcriptomes were correlated, and C18 and AE columns for the metabolome were correlated), there was minimal correlation between the transcriptomic and metabolomic datasets. This finding was unexpected given the connection between these functional molecular classes. It is important that we do not have metabolomic data from the first two time points (i.e. there is no pre-drug exposure). Since the greatest differences in the transcriptome were between the pre- and post-drug groups, we are clearly missing important data that would have shed light on the effect of anti-malarial drugs.

Between the CBC data and the transcriptome there was a similar lack of a clear relationship between the features despite intuitive expectations otherwise. Upon further consideration however, this lack of strong correlation between different technologies demonstrates that these data types offer non-redundant information about the state of the system, and adding additional data-types has the potential to increase our understanding of the system. If the metabolome could be perfectly recapitulated using only the transcriptome, there would be no need to interrogate multiple data-types.

The usage of BIT axes scores to reduce dimensionality and infer biological structure in RNA-seq data represents a new application of this statistical technique which was originally developed for microarray datasets. Also, this is the first time that this technique has been applied to a tissue other than blood. Based on the high level of variance explained by the each of the first PCs (i.e. the axis scores) for marrow, it appears that this method of data reduction is appropriate for marrow because it is capturing biological signatures that differ across time and drug treatment.

2.5.3 Dysregualtion of gene expression

In the bone marrow, there were thousands of genes that were dysregulated as a function of the drug treatment. Genes with altered expression were enriched in biological functions like one-carbon cycling by folate, oxidative phosphorylation, cell cycle, and apoptosis. For the one-carbon cycling by folate, I anticipated that the host would up-regulate genes in this pathway, especially DHFR, to compensate for the inhibition caused by pyrimethamine. Contrary to my expectation, however, the genes in this pathway were nearly uniformly *down-regulated*, suggesting that the cells sensed the decrease in the substrate tetrahydrofolate, and reduces expression of genes involved in its cycling.

In addition to the mitochondrial disruption that occurs in both the 7 days post-drug and the > 30 days post-drug in BOTH tissues (blood and marrow), there are numerous processes that are down-regulated in the marrow that relate to cell cycle control and mitosis (e.g. DNA replication, condensed chromosome, G1-S transition of mitotic cell cycle). This result is expected considering that the bone marrow has many more cells undergoing cellular division compared to the blood, which has mostly post-mitotic cells.

Through this thorough examination of the effects of pyrimethamine using high-throughput transcriptomics, I have identified numerous molecular pathways that are

dysreguated in response to pyrimethamine treatment. Furthermore, I have shown that this dysregulation of some pathways persists for at least a month after the last drug dose, much after the drug should have been cleared from the blood plasma.

Further questions about the possible dysregulation of host molecular pathways remain unanswered. For instance, pyrimethamine is often given in combination with another AMD usually the sulfonamine sulfadiazine. However, this medicine is often given with folinic acid which helps to reduce the level of folate deficiency and the downstream negative effects. In this study, we did not investigate whether folinic acid supplementation could abate or completely remove the negative host effects produced by pyrimethamine administration.

The results of this study are particularly important given the recent trend of re-purposing pharmaceuticals for the treatment of other diseases. Considering that pyrimethamine has potential as a treatment of amyotrophic lateral sclerosis (ALS), these results and datasets represent an important asset to the scientific community.

CHAPTER III

HOST GENE EXPRESSION ALTERATIONS DURING MALARIA INFECTION

3.1 Abstract

To address gaps in the understanding of host response to relapsing malaria, we performed a 100-day experiment that spans one primary infection and two relapsing infections of *Plasmodium cynomolgi* in the non-human primate *Macaca mulatta*.

Previous studies have examined the differences in gene expression that occur in hosts in response to malaria infection. This project represents that first time that host gene expression has been monitored across multiple peaks of parasitemia using digital gene expression quantification (RNA-seq).

Leveraging the power of this experimental design, I found qualitatively different gene sets enriched between the three parasitemia peaks in light of the vast differences in host clinical parameters between the first and subsequent malaria infections.

After performing differential gene expression followed by gene set enrichment analysis, I found that previously-identified immune pathways are altered in the host in response to the primary parasitemia, but not in the relapsing time points. There are also remarkable differences in the systemic lupus erythematosus (SLE) and heme metabolism pathways. These findings demonstrate that relapsing and primary parasitemias are qualitatively different. I additionally identify differences in host blood traits that covary with infection severity and have the potential to be diagnostic for severe disease susceptibility.

3.2 Introduction

3.2.1 Project overview

The Malaria Host-Pathogen Interaction Center (MaHPIC) is a multi-institute collaboration consisting of a number of different experiments designed to address specific gaps in the understanding of underlying dynamics of malaria infection by performing and integrating a number of unique molecular and cellular analyses. In this first infection experiment, *Macaca mulatta* was infected with *Plasmodium cynomolgi* to understand host and parasite changes across three peaks of parasitemia: a primary parasitemia followed by two relapsing parasitemias.

3.2.2 *P. cynomolgi* as a model for *P. vivax*

Plasmodium vivax, the most common malaria parasite outside of Africa, has received relatively less attention than its more pernicious cousin *P. falciparum*. In spite of its perception as a less-deadly malaria parasite, *P. vivax* is an important cause of morbidity and mortality in humans.[100]

Due to ethical considerations, it is difficult to receive approval to infect humans with *P. vivax* and observe high parasitemias and subsequent relapses of malaria. *P. cynomolgi*, a malaria parasite that also produces hypnozoites, inoculated into *M. mulatta* serves as a good model for relapsing *P. vivax* malaria in humans.

P. cynomolgi is a well established model for *P. vivax*[104, 128, 33, 45] and recently has been used to study host responses to co-infections with malaria and other organisms.[77, 155] Only one previous study in macaques (2005) considered the host response to relapse on a transcriptome-wide level.[163] In this study, they report many genes that respond differently across the various peaks, but due to lack of sufficient sample size, they do not determine statistical significance of their findings. They use a unconventional gene set enrichment approach that only broadly implicates immunological defense responses but not cytoskeleton pathways in response to

infection. Consequently, a clear understanding of the difference in host response to primary compared to relapsing parasitemias is yet to be established. In the intervening years, vast improvements in functional annotations of primate genomes have been achieved and should allow for more detailed analysis of gene set enrichment.

3.2.3 Hypnozoites and relapse

Unlike *P. falciparum*, *P. vivax* can produce dormant liver stage parasites known as hypnozoites. These dormant hypnozoites can be reactivated in response to external cues; parasitemias caused by reactivation of these dormant forms are known as relapses. As a result of having a relapse-causing hypnozoite stage, the parasite prolongs the duration that it is in the blood stage during which it could be transmitted to a mosquito, a necessary step in the completion of its life cycle. While secondary parasitemias increase the probability of continued transmission, another potential evolutionary benefit for the production of relapsing infections is the increasing of the complexity of infection (COI), a measure of the number of different parasite strains within a host at a given time. Having a greater COI offers the opportunity for exchange of genetic material between unique parasite strains during recombination in the mosquito vector.[103] Studies of natural genetic variation of other hypnozoite-forming species have not been performed.

3.2.4 Host gene expression in malaria infection

Previous work in host gene expression has focused on various aspects of the host response: susceptibility and response to severe malaria,[96, 87] gene expression changes during placental malaria,[20, 99] variation in response due to infection history (naive versus persistent infection),[106] differential host outcomes, mild versus severe malaria,[78] and general trends in expression during infection.[67, 57, 163]

In human malaria infections in West Africa, Idaghdour et al reported that the blood transcriptome showed evidence of up-regulation of certain gene sets including

immune-related pathways: TNF pathway, chemokine signalling, NOD-like, Toll-like, and Fc-γ mediated phagocytosis.[67] Also, the PPAR signalling pathway and the insulin receptor signalling pathway were up-regulated suggesting altered host metabolic state in response to infection. Down-regulated were pathways like aminoacyl-tRNA biosynthesis and ribosome, which signifies a reduction in translation in the cell population as a whole.

Kwiatkowski and colleagues reported an increase in expression of neutrophil-related genes during acute malaria.[57] This increase in neutrophil-related genes was accompanied by an increase in the absolute abundance of neutrophils but was not completely explained by it. Type-I interferon has been found to be associated with malaria infection, but reports differ with respect to the direction of effect.[96, 78, 20, 94] While previous expression profiling studies have assessed changes in response to malaria infection, there is a lack of enquiry concerning differences between primary and secondary infections.

Many of the previous studies were either performed in humans or mice. As mentioned above, there are certain experimental limitations when working with human malaria infections. For instance, there is likely to be high variance in expression data due to confounding factors like co-infection with other diseases and different environmental exposures like diet and life-style. Because of the large amount of variance coming from unexplained sources, large sample sizes are needed in human studies. The benefit of using a mouse model is the ability to both tightly control any potential confounding factors as well as allowing the subjects to progress to severe levels of clinical malaria.

Two potential drawbacks of using a mouse model are the genetic homogeneity (due to inbreeding) of the mouse strains and the evolutionary distance between humans and mice. The high genetic homogeneity of the mice introduce the possibility that results are specific only to the strain under study and are not generalizable to

the population, as a whole, not to mention inability to infer human responses. The importance of inter-individual variation due to underlying genetic differences in terms of host response to malaria was underscored in a recent study.[67] In this large expression quantitative trait locus (eQTL) study in an African population, Idaghdour and colleagues identified alleles that altered gene expression in a malaria-status-dependent manner. That is, the allele did not significantly effect the expression of a nearby gene in the absence of malaria, but did effect the expression in the presence of an active malaria infection. This finding demonstrates individual-specific responses to malaria.

A second disadvantage of using mouse models to infer human response to malaria is the evolutionary distance between humans and mice is great. Humans and primates are sufficiently related that some of the same species of *Plasmodium* can infect members of the two groups (e.g. *P. knowlesi* which infects both humans and rhesus macaques, and *P. falciparum* which infects both humans and *Aotus* monkeys). In light of their genetic and immunological similarity, primates may offer greater insight into human malaria response than other model organisms.

3.2.5 Host response to secondary malaria infections

Previous studies have documented great differences between host response to primary versus secondary infection. In rhesus macaques, it has been observed that parasite counts during primary parasitemia are much higher, and clinical measures of disease (anemia, pancytopenia, multi-organ failure, etc.) are much more severe during the first infection; subsequent parasitemias (both relapsing and recrudescent) were well-controlled by the host.[163, 98]

Another observation that leads me to believe that the host response on the transcriptional level will be qualitatively different in the relapsing parasitemias compared to the primary parasitemia is that, in humans, clinical immunity is built over time, further enforcing the belief that the host immune system becomes more adept at

neutralizing the threat of harm from the parasite.

In spite of these observations which demonstrate vast differences between a first and subsequent malaria episodes, a previous expression profiling study only identified a single gene that was differentially expressed between episodes of severe versus subsequent mild malaria, which suggests that in spite of dramatic clinical differences, there may not be differential response to relapsing parasitemias on the host transcriptional level.[78]

3.2.6 Motivating hypothesis

In this chapter, I deeply profile the host transcriptome changes that occur in response to malaria infection and explore whether there are qualitative differences between the host response to primary and relapsing parasitemias.

Using differential gene expression analysis in series with gene set enrichment, I will address the following questions. First, are there many genes differentially regulated between primary and relapsing parasitemias? Based on the results from the aforementioned study, the null hypothesis may be correct. That is, the host transcriptome may not differ between primary and relapsing parasitemias. However, if there are many genes differentially regulated in the primary parasitemia that are not dysregulated in the relapsing time points, that would constitute evidence that there is a difference in magnitude and/or the quality of effect.

Based on the vastly different clinical outcomes of primary versus relapsing parasitemias, I do expect a difference in the magnitude of effect of differential gene expression. The question remains, however, whether the host response is qualitatively similar between the two infection peak types. If the host response is qualitatively similar between primary and relapsing parasitemias, I would expect to see similar gene sets enriched with the differentially expressed genes. If the host transcriptional response is qualitatively different between the two infection peak types, however,

unique gene sets will be enriched.

In addition to gene set enrichment analysis, I also use two additional complementary methods, blood informative transcript (BIT) axis analysis and cell-type specific gene expression, to profile the global alterations that occur in the transcriptome over the course of the experiment. I will specifically contrast the activation of hematopoetic and immune-related pathways across the various infection peaks.

3.3 Methods and materials

3.3.1 Experimental design

The Malaria Host-Pathogen Interaction Center (MaHPIC) is an inter-disciplinary investigation that utilizes multiple data types to better understand the systems biology of the complex host-parasite dynamics in the course of malaria infection. For this malaria infection experiment, the design is as follows. Five male rhesus macaques (*Macaca mulatta*; RFa14, RFv14, RIc14, RMe14, and RSb14) approximately 2 years of age were profiled over the course of a 100-day experiment after being injected with purified sporozoites of the species *P. cynomolgi* on day 0 of this control experiment. Complete blood counts were performed daily by the Malaria core team members. Before injection, the time point 1 (TP1) samples were taken by the Malaria core. Then, on approximately days 20, 26, 53, 59, 89, and 96, blood and marrow samples were collected for TP2-7, respectively (Figure 13). A sub-curative dose (1mg/kg) of artemether was given at TP2 to three of the five animals (RFa14, RFv14, and RMe14) to stem the increases in parasitemia. At TP3 and TP4, all animals received an 8-day course of artemether: day 1 (4mg/kg); days 2-8 (2mg/kg). At the end of the experiment, all animals were given fully-curative doses of chloroquine. All aspects of this study were all approved by the Institutional Animal Care and Use Committee (IUCAC) of Emory University.

3.3.2 Transcriptome analysis

Library preparations, read mapping, and expression quantification were all performed as described in the previous chapter (Figure 12). Briefly, total RNA was extracted from whole blood from samples taken across the 100-day experiment. Due to poor RNA quality, one sample was not sequenced (RFv14 at TP2). mRNA was enriched from total RNA using poly-dT beads. All library preparation was performed by Dalia Arafat. Libraries were bar-coded and sequenced on an Illumina HiSeq2000 in the Yerkes Genomics Core led by Dr. Zach Johnson. After sequencing, reads were mapped to a combined reference genome including host and parasite genomes; both genomes were the most recent genomes and were obtained from the iRODS repository from the MaHPIC project website. HTSeq was used to assign read counts for each annotated gene. Finally, DESeq was used to calculate normalized expression values, which were subsequently used for differential gene expression analysis.

3.3.3 Gene set enrichment analysis

After calculating the normalized expression values, a pseudocount of one was added and the data were then log2-transformed. All genes with an average transformed expression value below 3 were excluded from further analysis. This transformed dataset served as input to the statistical software, JMP. Differential expression was determined using an analysis of variance (ANOVA) test in JMP with animal as a random effect and time as a fixed effect.

To perform gene set enrichment analysis, I chose the ranked gene list method and used the Broad Institute's GSEA v2.0.14 to perform enrichment analysis.[142] For the contrasts of interest (TP1 against each of TP2-7), I performed gene set enrichment using the KEGG pathways, Annotated gene sets, and GO terms. The t-statistic was used as the rank metric and was obtained from the JMP output file. Gene sets with an FDR $< 25\%$ were considered as significant in accordance with the recommendations of

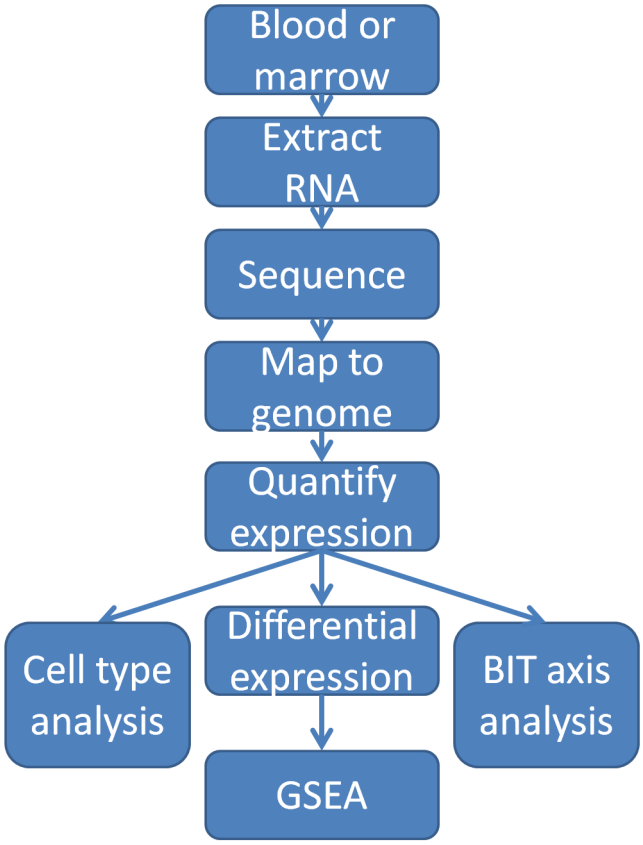

Figure 12: Analytical pipeline of gene expression profiling. After the 100-day infection cycle, samples from all time points are then used to make paired-end, strand-specific libraries for sequencing on the Illumina Hi-Seq. After sequencing, reads are mapped to a combined reference genome and transcriptome. HTSeq is used to assign expression levels to each annotated gene, and expression levels are then normalized using the method suggested by DESeq. Subsequent down-stream analyses including differential gene expression, cell type profiling, and blood informative transcript axes profiling are then performed on the normalized expression values.

the GSEA software manual. Default parameters were used and included the removal of gene sets with more than 500 or less than 15 genes.

To perform GSEA, *a priori* defined gene sets are needed. Since the rhesus macaque is much less well-studied than the human genome, and considering that the majority of genes in the macaque genome have well-conserved syntenic orthologs in the human genome, I used the pre-existing human gene set annotations for this analysis. These genes sets were obtained from the Broad Institute's website.

3.3.4 Blood informative transcript axis analysis

I employed blood informative transcript (BIT) axes analysis, which uses an *a priori* defined set of nine blood axes, which are composed of genes (10 blood informative transcripts per axis) that represent much larger groups of genes that covary across blood transcriptional profiles.[114] This method has been described elsewhere.[102] Briefly, I took the 10 BIT genes for each of the axes and calculated the BIT score for each individual for both the blood and the marrow, separately, using the normalized expression data. Then, I performed ANOVA to test for significant differences across time after fitting the effect of animal. The first PC of most of the BIT axis explained the majority of the variance for the 10 BIT genes.

3.3.5 Cell-type-specific gene sets

To identify the trajectories of specific blood cell types using the transcriptome data, I downloaded lists of genes that were previously identified to be expressed in a cell-type specific manner.[111] Gene sets were available (Additional File 2 from the manuscript) for the following groups of cells: lymphocytes, B-cells, T-cells, CD8+T-cells, and granulocytes. As a single descriptive metric of the trajectory of each sub-population of cells, I performed principal component analysis (PCA) and used the first principal component (PC1) as a relative measure of the abundance of the cell type. I then performed an ANOVA to test for significant differences of cell-type abundance across the experimental time points.

3.4 Results

3.4.1 Parasitemia across the 100-day experiment

Across the course of the 100-day experiment, blood was taken daily to quantify parasitemia using manual counting of infected RBCs on blood smears. The parasitemia count data shows that the primary and subsequent relapse infections occurred at

approximately the anticipated times during the experiment (Figure 13). Macaques were injected with the sporozoites on day 0, and on day 12, all primates showed first positive blood stages. The parasitemias continued to rise, and on day 20 the first time point with evidence of parasitemia (TP2) was taken. To slow the rise of parasitemia, three of the five primates (RFa14, RFv14, and RMe14) were given a single-day dose of artemether to stem the rising parasitemia. Seven days later, the parasitemia levels remained high for most hosts; a sample for TP3 was taken and subsequently all surviving primates were given a full course of artemether, which quickly reduced the blood-stage parasitemias to zero for all macaques. Primate RFv14 experienced severe anemia as well as acute kidney failure, and supervising veterinarians made the decision to euthanize the animal on day 23 per IACUC-approved procedure. Time point 3 for this animal was taken immediately before euthanasia. The relapsing infections occurred around day 60 for three of the four animals, and another relapse occurred near day 90 for all four animals. Parasitemias for the relapsing infections were much lower, a finding consistent with previous studies.[163]

3.4.2 Host blood cell parameters are altered by malaria infection

In addition to parasitemia counts, daily blood samples were used to perform complete blood counts. A severe drop in all three major blood cell types (white blood cells, red blood cells, and platelets) occur during the first and highest parasitemia in all animals. Subsequent relapsing parasitemias are met with a much smaller response from the host. Also note in Figure 14A that around day 20, there are two days where the red blood cell levels for two primates (RFv14 and RMe14, represented by light blue triangles and dark blue squares, respectively) are much lower than all other measured readings. These two animals were given blood transfusions at this time in an effort to prevent complications from severe anemia. As noted above, RMe14 successfully recovered, but RFv14 did not.

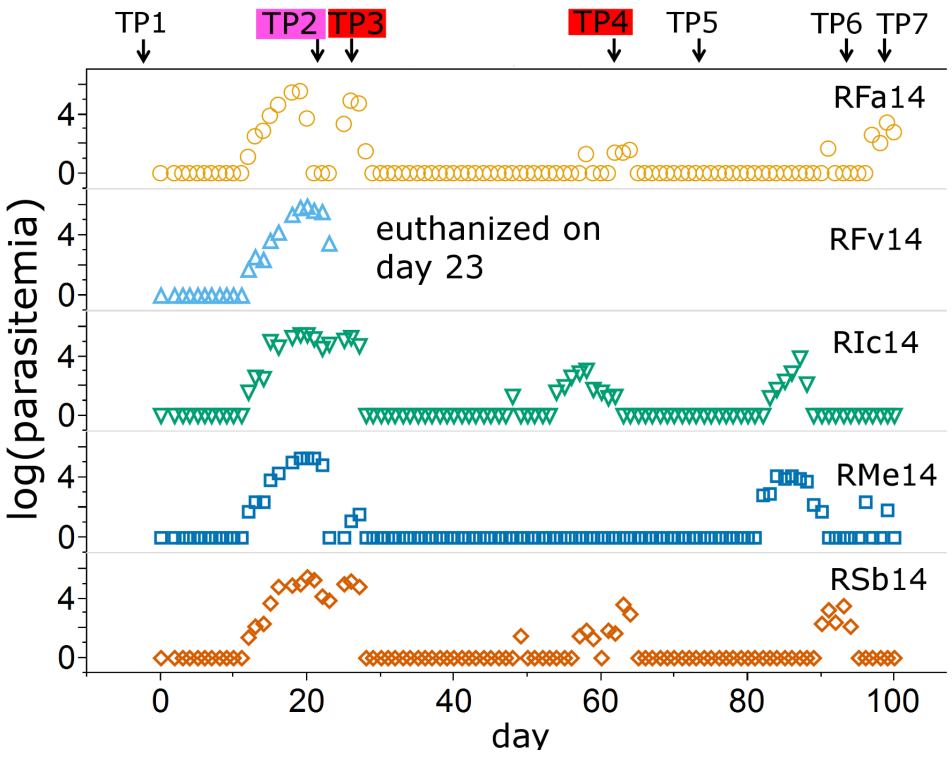

Figure 13: Parasitemia levels across the 100-day infection cycle. Parasitemia across the 100-day experiment. Approximate days of sampling are indicated by arrows. At TP2 (pink), a single-day dose of artemether was administered. At TP3 and TP4 (red), a full eight-day course of artemether was given (see methods for dosage). Primate RFv14 was euthanized on day 23 due to renal failure. Note that a pseudocount of 1 has been added to the parasitemia counts before log transformation, and therefore a log-parasitemia of zero is zero. This data was collected by the Malaria Core principally led by Alberto Moreno.

As a general feature of the CBC data, I report that as in my previous analysis, there are significant differences in many blood cell traits across the primates. Of the 13 blood cell parameters measured, 9 were significantly different across the subjects. Figure 14 shows lymphocyte count, one example of a cell type with differential abundance. Considering that different animals have different blood cell parameters, I wondered if such parameters could contribute to severe disease. For the one primate that experienced severe malaria-induced anemia (RFv14), I noticed that he had lower levels of both mean corpuscular volume (MCV) as well as mean corpuscular hemoglobin (MCH). Since the corpuscle (i.e. RBC) is the specific cell type that

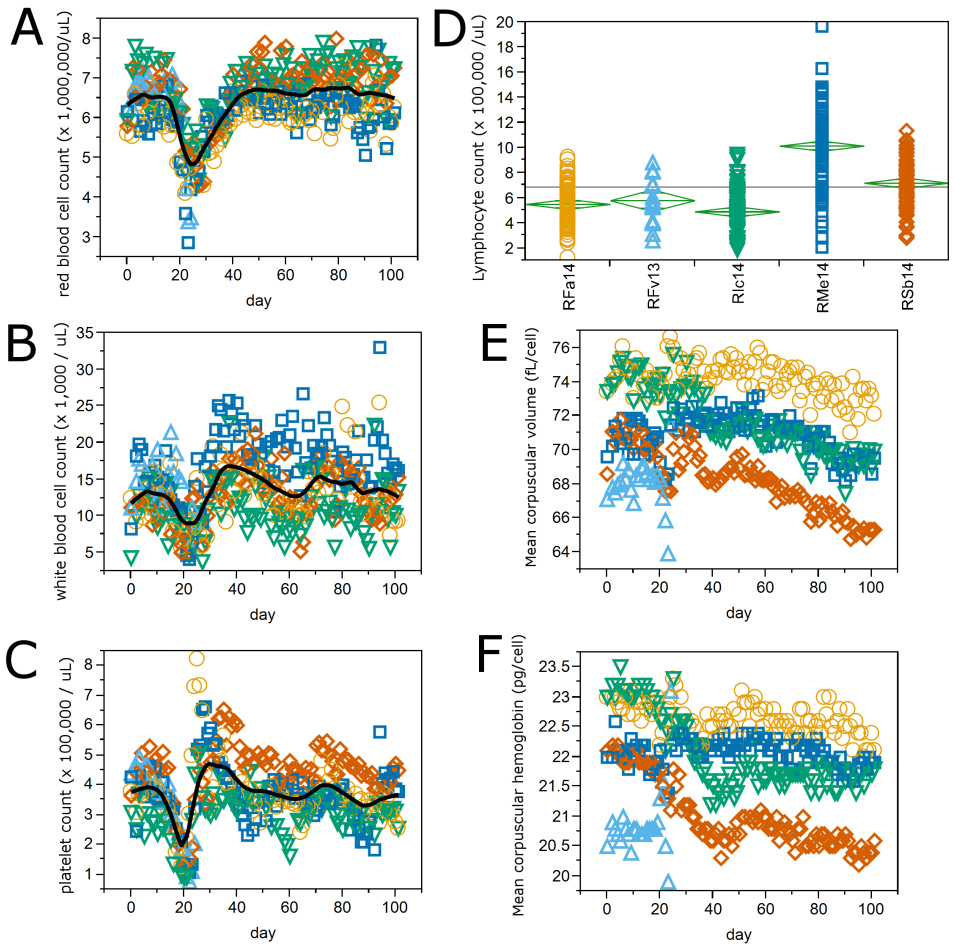

Figure 14: Abundance of blood cell types across the 100-day infection cycle. Cell abundances in whole blood of the major blood cell populations: (A) red blood cells (RBCs), (B) white blood cells (WBCs), and (C) platelets. The black line is a kernel-smoothed fitted line showing the trajectory of cell abundances. Data points are colored by animal as in Figure 13. At the peak of the first parasitemia corresponding to days 15-25, there is a precipitous drop in the numbers of all blood cell types, a phenomenon known as pancytopenia. At the two subsequent relapsing parasitemias (occurring around day 60 and day 90), there are smaller dips in the three blood cell populations. (D) The distribution of lymphocyte counts by animal. Center line of the diamond denotes the mean and the ends are the 95% confidence intervals. The (E) mean corpuscular volume and the (F) mean corpuscular hemoglobin across the 100-day experiment. The animal that was euthanized due to severe malaria and renal failure (RFv14, blue triangles) had lower levels of these two blood traits even before patent blood stages.

the malaria parasite invades and hemoglobin is the substrate upon which it feeds, I suggest that one or both of these parameters may be important in severe malaria pathology. This hypothesis is supported by a recent genome-wide association study (GWAS) performed in humans, which shows that genomic loci that are important in malaria resistance and pathology are enriched with SNPs that explain substantial amounts of the variance in blood cell traits.[39]

3.4.3 Primary parasitemia extensively alters host gene expression

The variance components for this parasite-infection experiment are qualitatively similar to what was observed in the control experiment; the effect of animal explains 40% of the variance while the effect of time point explains slightly more than the control experiment at 25% (Figure 15A). This relationship is born out in the hierarchical clustering of the principal components. Essentially, the samples cluster by animal with the exception of the two time points of the first parasitemia, TP2 and TP3 (Figure 15B).

When the 4,233 significantly differentially expressed genes are hierarchically clustered, TP2 and TP3 do not cluster with the other samples (Figure 16A). Furthermore, nearly all of the differentially expressed genes are either very up- or down-regulated in one of these two time points. The time points sampled during the two relapsing infections (TP4-7) cluster closer to the control time point (TP1) than they do to either of the samples from the first parasitemia (TP2 and TP3). This result offers evidence that there are qualitative differences between the transcriptional profiles of the host response during primary versus relapsing malaria infections. In the volcano plot for the contrast of TP1 versus TP2, there are many genes differentially expressed (Figure 16B). In the TP1 versus TP4 contrast (that is, control versus first relapse parasitemia), however, there are many fewer genes that are differentially expressed (Figure 16C). After coloring the genes that are up- or down-regulated (red and blue,

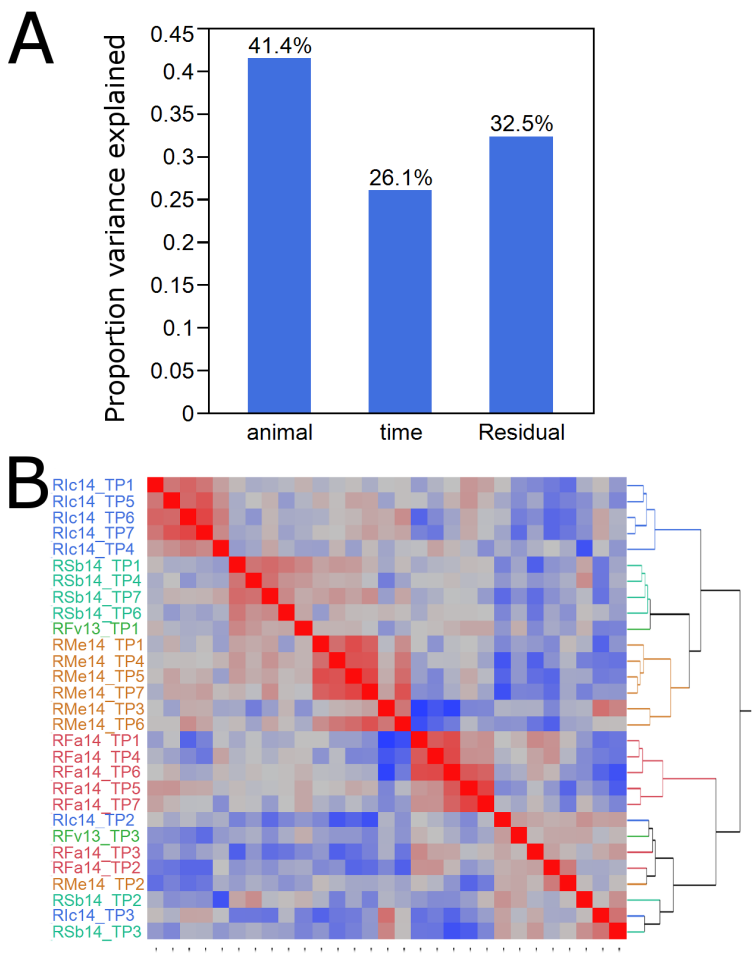

Figure 15: Summary of the blood transcriptome (A) The variance components of the blood transcriptome. (B) Hierarchical clustering of the samples based on the correlation of their principal component values.

respectively) according to the TP1 versus TP2 comparison, it is apparent that there is little coherence in the host transcriptional response with respect to the direction of expression between the primary (TP2) and secondary (TP4) parasitemias. This result supports the alternative hypothesis that primary and relapsing parasitemias elicit not only a different magnitude of response from the host, but also that there is a qualitatively different host response. I note that there are still several hundred genes that are dysregulated in TP4 (relapse) compared to the TP1 (baseline), albeit to a lesser extent than in the first parasitemia (TP2).

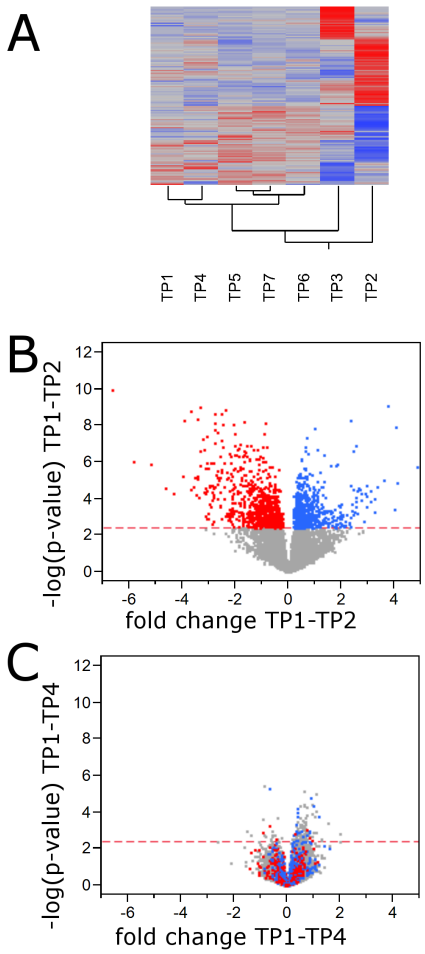

Figure 16: Differential gene expression analysis. (A) A hierarchically clustered heatmap of the significant genes. Volcano plots comparing (B) TP1 versus TP2, and (C) TP1 versus TP4. The dotted line near $y = 2.3$ is the cut-off for FDR=5%. Points represent genes and are colored (if significant) by there direction of effect in the TP1 versus TP2 contrast, up-regulated at TP2 in red, and down-regulated at TP2 in blue.

3.4.4 Gene set enrichment analysis

To determine the biological pathways that are enriched with genes that are differentially expressed across the three parasitemia cycles, I next performed gene set enrichment analysis using pre-ranked gene lists.

The most statistically significant gene set enrichment that I identified between the control time point (TP1) and the time point from the peak of the first infection (TP2), is the up-regulation of genes in the systemic lupus erythematosus (SLE)

related pathway (Figure 17). This pathway contains two predominant gene types: different histone variants and members of the complement pathway, suggesting an alteration in chromatin environment and immune response. The genes in this pathway continue to be elevated at TP3, as well. These two time points correspond to the highest parasitemias of the experiment. The enrichment of this curated KEGG pathway has previously been shown in both mild and severe malaria compared to control blood samples in the same direction as in our study.[67] Extending this finding and further underscoring the importance of repeated measures across the course of infection cycles, I found that genes in this pathway are very significantly enriched with genes that are down-regulated at time points subsequent to relapsing infections (TP5 and TP7). The tight co-regulation and extreme responses of this pathway during malaria infection support the need for further targeted experimental work on these genes.

The heme metabolism pathway (KEGG porphyrin and chlorophyll metabolism) is up-regulated in TP3 compared to TP1, but TP2 does not have this same enrichment (Figure 18). This may be occurring because of one of two non-mutually exclusive mechanisms. First, since artemether, the anti-malarial drug (AMD) given at time point 2, has the property of generating free radicals and reactive oxygen species especially when in proximity to a heme group catalyst, it may be that the artemether is destroying the heme group. The cell would then work to replace the lost heme by synthesizing more of it. A second mechanism that would explain the increase in heme production enzymes would be that the malaria-induced anemia has stimulated the production of more red blood cells. As a result, the red blood cell precursors would have the enzymes necessary for the production of heme. I believe that the latter mechanism is correct for many reasons. Firstly, mature red blood cells lack a nucleus, and so even upon sensing a decrease in the level of functional heme in the cell, they would be unable to increase the abundance of the transcripts in the heme production

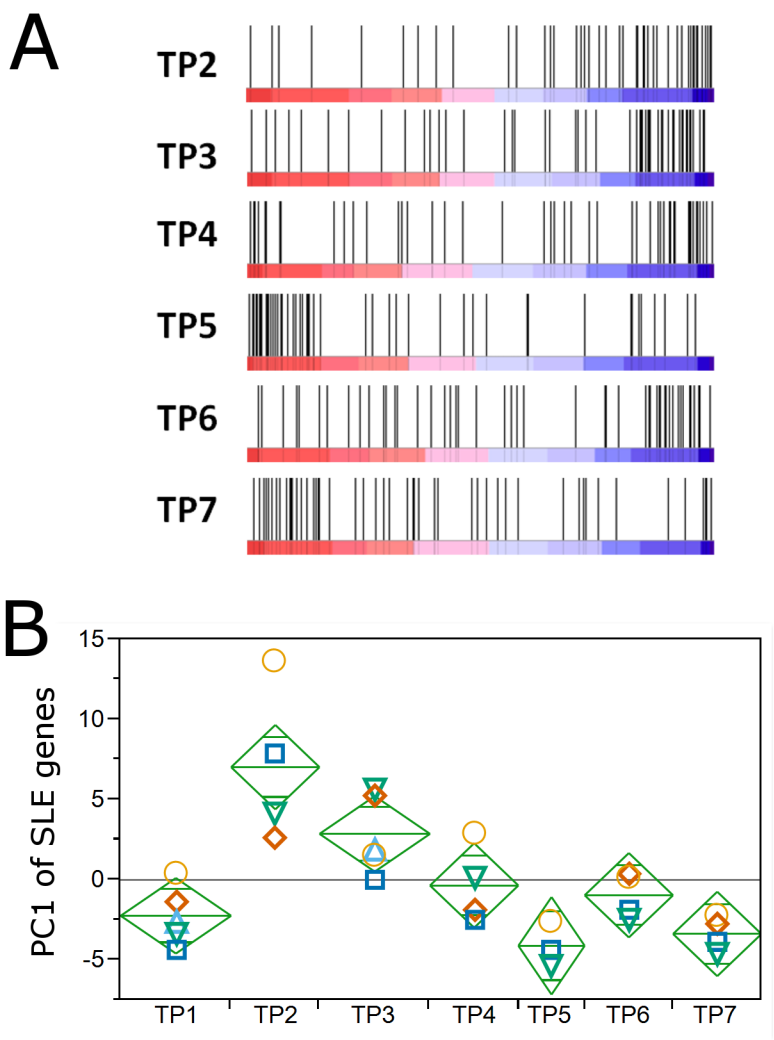

Figure 17: Cycling of the SLE-related pathway. (A) Gene set enrichment plots of the systemic lupus erythematosus (SLE)-related pathway genes annotated by KEGG. (B) The first principal component of the SLE-related pathway plotted by time. TP2 and TP3 are enriched with genes that are up-regulated in the SLE-related pathway. TP4 there is no significant enrichment. Then at TP5 and TP7 genes in the SLE-related pathway are expressed in the opposite direction, that is they are significantly down-regulated. TP6 is enriched with up-regulated genes.

pathway. Secondly, if the single-day dose of artemether at TP2 was strong enough to have an impact on the pathway seven days later, then presumably the eight day course begun at TP4 would have caused the same pathway to be dysregulated at TP5, which is not the case. The CBC data shows malaria-induced anemia in all of the animals

beginning around day 14 and recovery of RBC counts around day 25. TP3 was sampled very close to the turning point; the rising levels of RBCs is consistent with the hypothesis that RBC precursors in the blood are contributing to the transcript pool and leading to the appearance of increased heme pathway transcription.

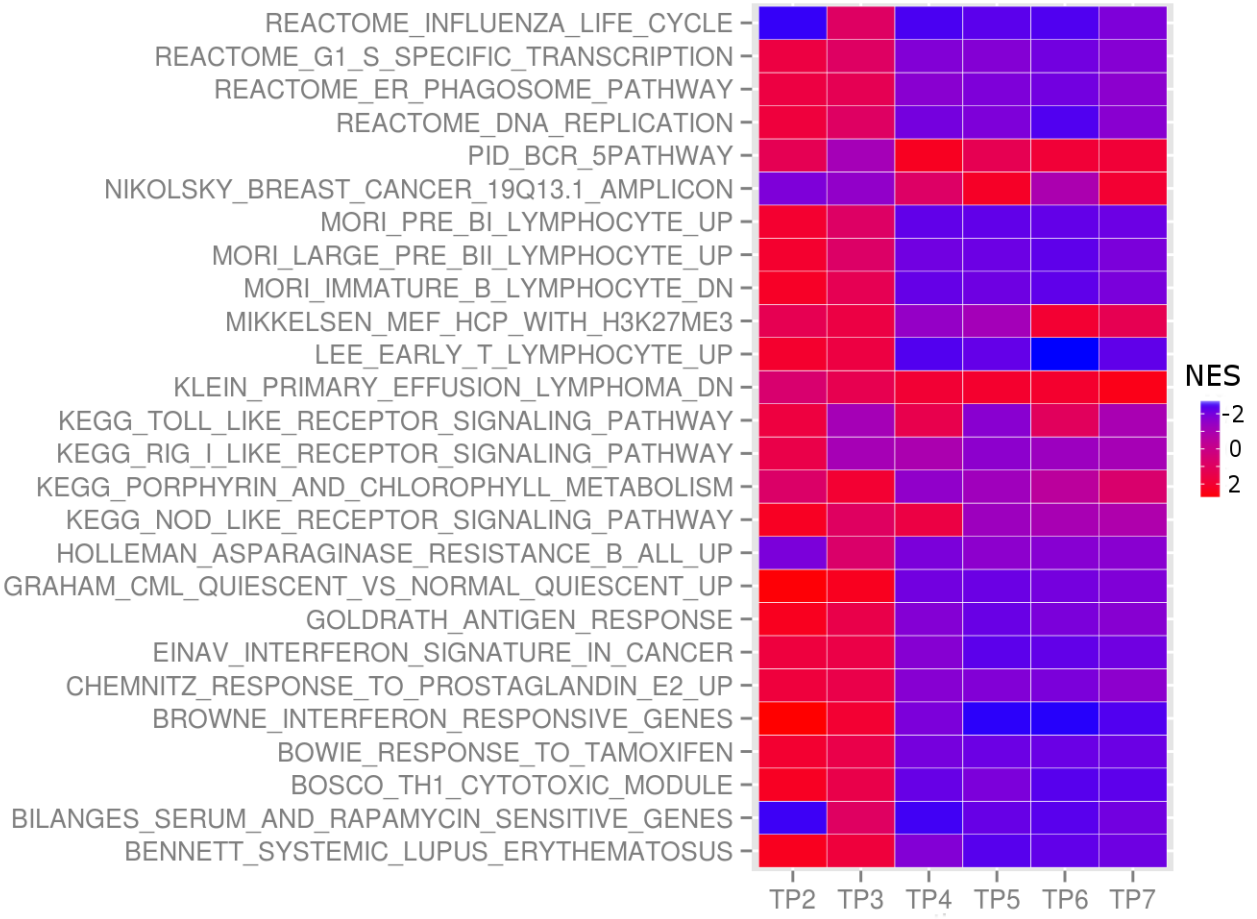

Figure 18: Gene set enrichments. For a selection of gene sets with significant enrichment in at least one time point compared to control, a heatmap of the normalized enrichment scores (NES), which describe the level of enrichment of each pathway.

Other immune-related pathways that are enriched include NOD-like receptor, Toll-like receptor (TLR), and RIG-I receptor pathways. These signalling pathways are most often associated with viral or bacterial infection, and their activation leads to the release of various cytokines including type I interferons. These signalling pathways

have been reported as altered in human malaria infection but this is the first report of genome-wide up-regulation in other primates. The Toll-like receptor pathway is up-regulated at TP2, TP4, and TP6, which corresponds to the peaks of each of the three parasitemias. Axis 5 (discussed in a subsequent section) includes TLR activity, and is also all up at TP2, TP4, and TP6, which is consistent with neutrophilia at the peak of infection.[59]

3.4.5 Cell-type specific expression

While gene set enrichment analysis offers important insights into the underlying host pathways that are activated in response to malaria infections, differences in expression profiles may be, at least in part, due to differences in cell-type abundance. To explore the changes in host immune cell abundances I used a composite metric of cell-type abundance (the first principal component of genes that exhibit cell-type specific expression, see Methods), I estimated the relative changes in cell-type abundance across the course of the experiment.

There were significant changes in both the B-cell and T-cell populations as a function of malaria infection. The B-cell specific genes fell at TP2 and then recovered at TP3 and continued increasing throughout the end of the experiment (Figure19A). T-cell specific genes followed the opposite trajectory: TP2 had a much increased level of T-cells, but at all subsequent time points (TP3-7), T-cell genes were reduced below the uninfected time point (Figure19B). The increase in T-cells does not appear to be related to changes in CD8+ T-cells (also known as killer T-cells) since the genes of this specific sub-type do not follow the same trajectory (Figure19C). This is not surprising considering that CD4+ T-cells (also known as helper T-cells) are usually the more abundant T-lymphocyte; a CD4+ T-cell-specific gene list was not available for this analysis. Yet another possibility for the lack of coherence for the CD8+ T-cell cycling is that activated immune cells often leave the circulation and aggregate

in peripheral tissues.

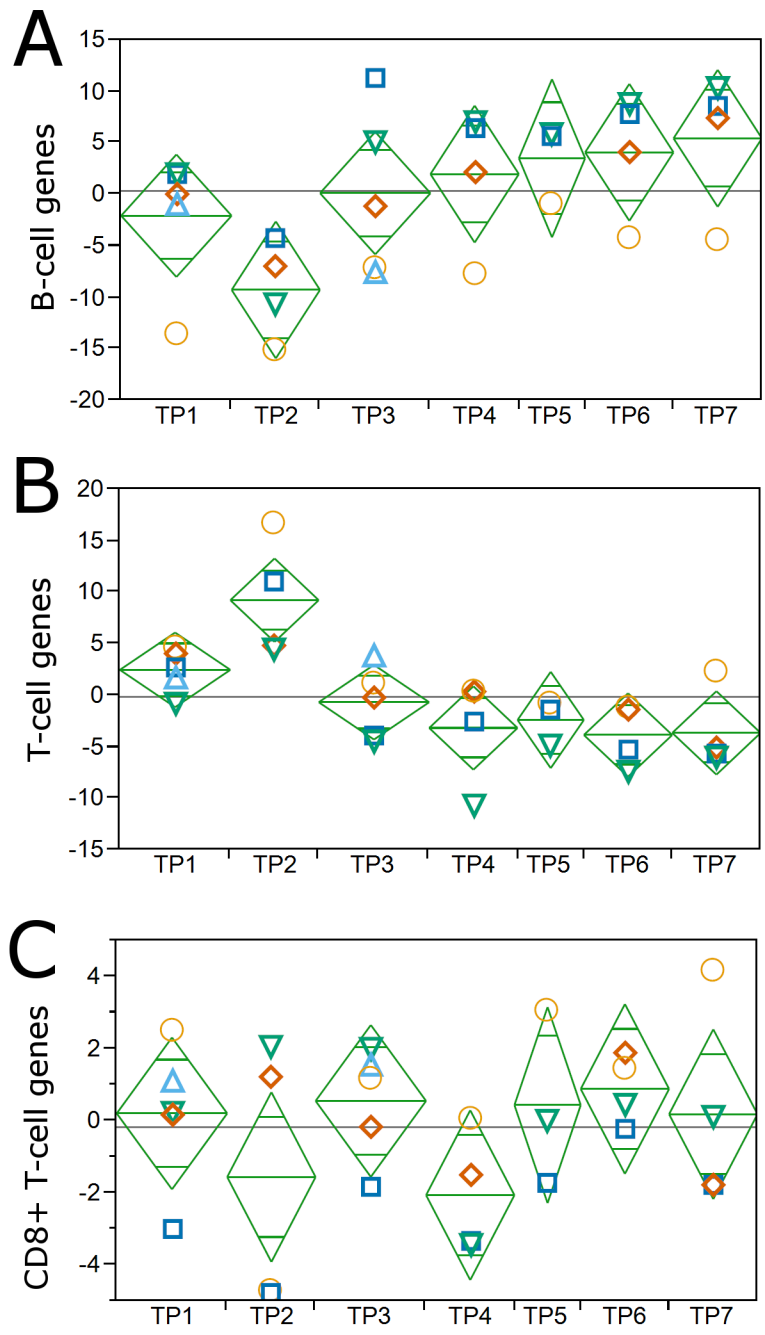

Figure 19: Cell-type specific gene expression across the experiment. The general behavior in B-cells (A) and T-cells (B) is qualitatively opposite. CD8+ T-cells (C) do not account for the extreme variation in the T-cell population.

3.4.6 Blood informative transcript axis analysis

In addition to cell-type specific expression, I performed blood informative transcript (BIT) axes analysis. This BIT axes represent underlying modules of co-expressed genes that have been previously identified in numerous blood transcriptome datasets. Many of the axes are enriched with previously determined functional annotations of gene ontology.

I found that many of the BIT axes showed significant differences across time in this 100-day infection experiment. Axis 2, an axis enriched with genes associated with hematopoesis, is significant across time points without controlling for the effect of animal (Figure 20A).

Axis 2 contains genes that are enriched in GO terms related to oxygen transporter activity and wound healing. The human disease and abnormal mouse phenotypes associated with this axis are hemolytic anemia and hematopoesis, respectively. Up-regulation of axis 2 during the peak of parasitemia and consequently the times of highest parasite-induced hemolysis is expected, and gives further support to the functional coherence of this axis. Importantly, the axis analysis and the cell type specific expression analysis are complementary approaches which have captured distinct aspects of the host response to the infection, hematopoesis and T-cell development, respectively.

After controlling for the effect of animal (see methods), there were significant differences in axes 3, 5, 7, 8, and 9; axis 2 became even more significant when removing the effect of animal.

Another very striking finding is axis 3 (Figure 20B), which is annotated as being associated with B-cell activation, behaves very similarly to the B-cell gene set (Figure 19A). This initial drop in B-cells at TP2 is likely due to the pancytopenia (that is, thrombocytopenia, leukocytopenia, and erythrocytopenia) induced by the infection. This finding lends further support to the ability of BIT axes to capture information

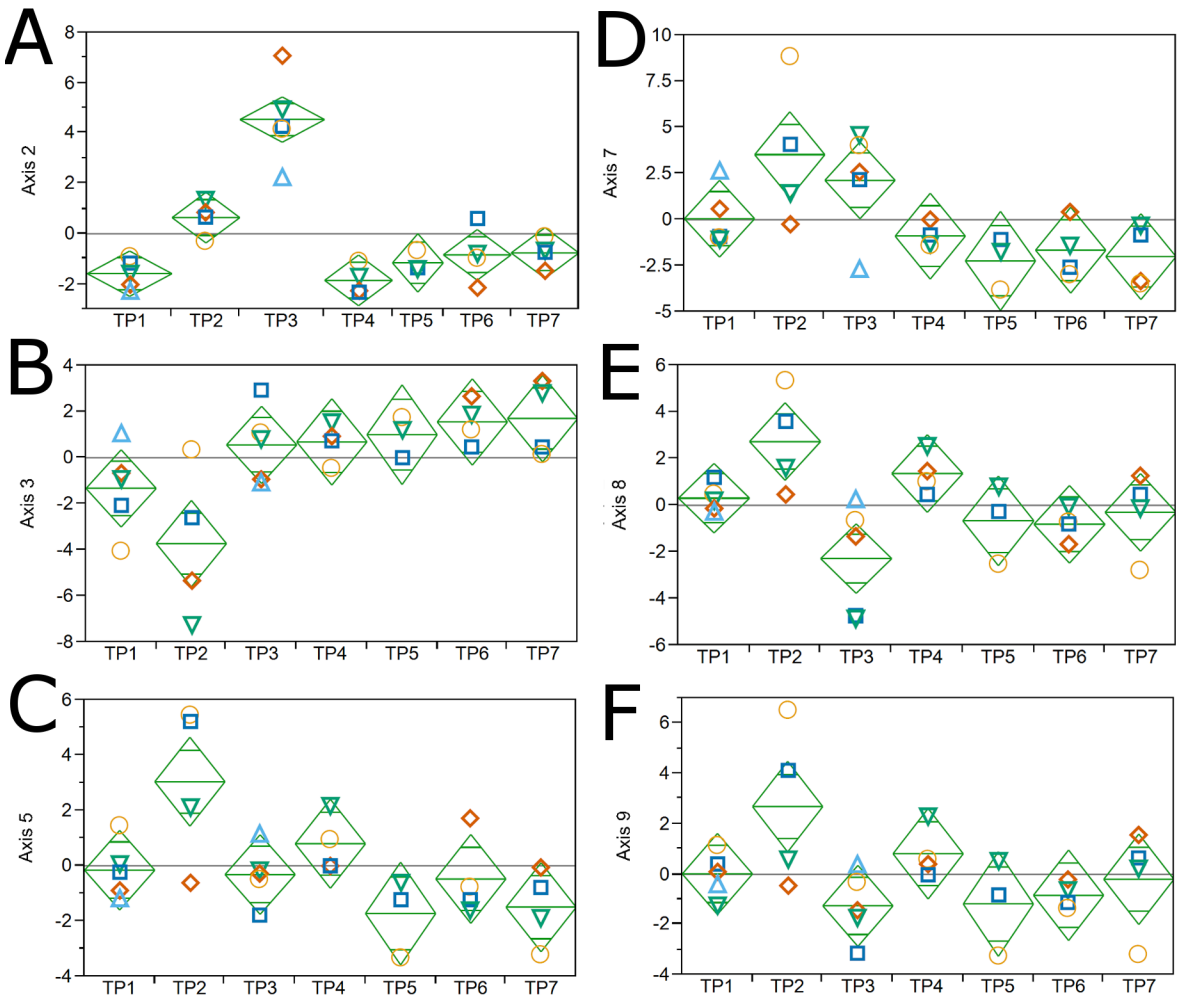

Figure 20: Blood informative transcript (BIT) axes over time. All plots except (A) show the BIT axis trajectory after removing the between-animal effect. (A) Axis 2, hematopoesis-related ($p = 0.0002$). (B) Axis 3, B-cell activation-related ($p = 0.0032$). (C) Axis 5, cytokine receptor activity ($p = 0.0067$). (D) Axis 7, interferon signaling ($p = 0.0085$). (E) Axis 8, RNA-processing ($p = 0.0045$). (F) Axis 9, apoptosis-related ($p = 0.0481$).

about the immunological state of the host using a very small number of transcripts.

Previous studies have identified interferon pathway up-regulation during malaria infection which I also observe here as demonstrated by increases in axis 7 (Figure 20D). Importantly, this interferon axis is down-regulated at TP4-7, further supporting the conclusion that host response to relapsing parasitemias are qualitatively different

from the primary peak. Lastly, axes 5, 8 and 9 appear to follow similar trajectories: high at TP2 relative to the control, followed by a zig-zag pattern: down at TP3 and then back up at TP4, the first relapse (Figure 20C,E,F). Inflammation and apoptosis (axes 5 and 9, respectively) are likely a by-product of the immune response to the malaria infection.

3.5 Discussion

3.5.1 Gene expression in a relapse

The vast differences in the gene set enrichment analysis as well as the BIT and cell-type specific expression analyses demonstrate that there is a qualitatively different host response between primary and relapsing parasitemias in this non-human primate model of *P. vivax* infection in humans.

A previous study investigated the differences in gene expression between a primary and secondary malaria infection in humans.[78] (No baseline transcriptome was reported.) Compared to our analysis, theirs was similar in size, yet they only found one gene differentially expressed at an FDR of 5%. This lack of significantly different expression profiles suggests one of two explanations: 1) either unaccounted for sources of technical or biological variance or both prevented them from finding many significantly differentially expressed genes; or 2) the host response to re-infection is qualitatively different from the host response to reactivation of dormant hypnozoites, that is, a relapse.

The former is most likely. One source of unaccounted for biological variation in the aforementioned human study is differences in diet as well as co-infection with other agents. A source of technical variance could be the difficulty of uniform treatment of samples in the field.

However, the latter may also be true. Re-infection (or recrudescence, as it may have been) with *P. falciparum* may truly elicit a completely difference host response

than a relapsing parasitemia with *P. cynomolgi* or *P. vivax* due to the fact that the parasites re-activated during a relapse are likely to be genetically similar (or identical) to the parasites of the first infection. However, population level diversity of *P. falciparum* is lower than for *P. vivax*, and so this is not the likely mechanism that would explain the difference in response.

The present experimental design measured baseline expression for all genes across the transcriptome and then across three parasitemia peaks for a total of seven time points for each macaque. The previous study lacked an uninfected baseline measure of the transcriptome, a shortcoming that may have decreased their statistical power to detect differential expression. However, if a re-infection behaved similar to our relapsing infection, the difference between the first and subsequent infections would likely have been strong enough to detect.

3.5.2 Comparison with previous study of *P. cynomolgi*

The macaques in our study showed the first evidence of blood stage on day 12 for all individuals with a peak being reached around day 20. While the day of appearance of blood stage parasites is qualitatively similar to a previous study, it took more than a week for parasitemias to reach their peak in our study, whereas peaks were reported to have been reached by day 14 (2-4 days after first appearance) in a previous study.[163] Also, the first relapse for most of our primates occurred between days 50-60, much later than a previous study.

In this study, I was able to perform statistical testing to identify differentially expressed genes, which were plentiful. In the previous study, statistical testing for differential gene expression was not possible.

The hierarchical clustering of all samples in our study was qualitatively different from a previous study in macaques[163]. This may be the result of different platforms

for measuring gene expression (microarray versus RNA-seq). Furthermore, normalization method in microarray has been shown to be important differential expression analysis and subsequently in biological inference.[116]

Unlike the previous study, design included samples both at the peak of relapse as well as seven days after the peak. This allowed me to identify that the SLE pathway, which was up-regulated at the first peak of parasitemia, was actually down-regulated in the recovery periods after the relapses.

Yet another difference between our study and the previous study was that they performed gene set enrichment analysis on the groups of genes that hierarchically clustered together. This has the effect of enriching for genes that are highly correlated, which are also therefore likely to be enriched for the same function if we assume that genes in the same pathway are co-regulated. This kind of gene set enrichment has the effect of artificially reducing the p-values of the enrichments. In spite of this difference in method, we both identify pathways important in hemoglobin metabolism.

3.5.3 Systemic lupus erythematosus (SLE) and immune related gene set enrichment

The recapitulation of the increased expression of SLE-related genes in response to malaria demonstrates that this primate model of malaria is a good representation of what happens in the human body during an infection. However, my identification of a subsequent down-regulation of many of the genes in this pathway at later times (TP5 and TP7) offers a new and important insight with respect to the immune fluctuations that occur in primate hosts during malaria infection and recovery.

In the introduction, I discussed the selection for HLA-b53 locus which provides protection against severe malaria and is highly prevalent in West Africa. But beyond host adaptations to the parasite, we can also consider the host-parasite co-evolution. For instance, SLE, an autoimmune disease, is much more prevalent in African women outside of Africa than it is for European women. Furthermore, there is much more

SLE susceptibility among southern European women (which until recently had higher exposure to malaria) than among northern European women. Importantly, there are relatively very few reported cases of SLE in African women who currently reside in malaria-endemic regions. I hypothesize, therefore, that chronic infection with the parasite modulates the host immune system. And in the absence of such modulation, the over-active immune system leads to the autoimmune destruction of some of the host's own cells.

A recent study in humans found evidence for a genotype-by-malaria-status interaction effect on the gene expression.[67] This provocative result provides a potential mechanism which could explain the high rates of SLE among African-American women compared to European women and African women remaining in malaria-endemic regions. Specifically, malaria infection may be modulating gene expression networks of individuals whose genomes have adapted to a malaria-endemic environment. This finding is important in light of the fact that African-American (AA) women have much higher levels of SLE than European-Americans (EA). I, therefore, hypothesize that the SLE of AA women is possibly due to co-evolution with the malaria parasite in an attempt to modulate the negative impacts on fitness of infection.

Further support of the importance of the SLE pathway in malaria infection is found in a mouse model of SLE.[156, 157] In these studies, mice with susceptibility to lupus (due to genetic polymorphisms) protected mice from cerebral and placental malaria. Both of these finding are consistent with a protective effect of SLE against reduction of reproductive fitness due to severe malaria. Yet another interesting connection between malaria and SLE is the use of anti-malarial drugs in the treatment of SLE.[22, 71]

We are currently pursuing a further bioinformatic analysis to substantiate the link between SLE and malaria. Specifically, we will compare genome-wide association studies (GWAS) from both malaria susceptibility studies and SLE studies to

determine if there are co-enrichments in the significance levels of SNPs or closely linked genetic loci.

3.5.4 Hematopoesis

In my analysis, I observed an up-regulation of the hemoglobin production pathway (KEGG porphyrin and chlorophyll pathway) at TP3. This hemoglobin production could be up-regulated due to either parasite-induced anemia and/or artemether binding to heme groups and causing localized damage. I believe that this up-regulation is due to the parasite-induced anaemia and not due to off-target artemether destruction of RBCs.

Axis 2, which is enriched with genes important in hematopoesis, was most highly up-regulated at TP3. While the animals received a single-day dose of the AMD artemether seven days before this sample was taken, I do not believe that this alteration in RBC genes was related to the drug dose. This is because at TP4, the macaques received a full eight-day course of artemether, and TP5 was taken near the end of the treatment. No similar up-regulation of the hematopoetic axis was apparent in TP5.

3.5.5 Cell type deconvolution of the samples

Blood is a heterogeneous tissue that is composed of many different kinds of cells: RBCs, platelets, and many subtypes of WBCs, each cell type having a characteristic expression profile. In a sample of blood, the proportions of each cell type would have a great impact upon the gene expression profile. One way to determine and account for differences in cellular composition is to perform sample deconvolution, which is a computational method whereby the proportion of cell types are estimated based on expression profiles. In our analysis, we did not perform cell type deconvolution because we lacked RNA-seq profiles for the numerous cell types that would be expected to occur in the blood, and most cell type deconvolution software has only been

written to be compatible with microarray expression data. Advances in this field will allow for estimation of cell type proportions in our samples. Subsequently, we would be able to use the cell type abundance data as a factor as well as a covariate in the model. The infected time points may have differential abundance of certain immune cell types. Additionally, even after accounting for cell type differences, infected time points may have pathways that are differentially regulated. It is with this method, cell type deconvolution, that we can tease apart the alterations that are occurring in the host due to infection.

3.5.6 Individual-specific responses to malaria infection

In our study, we had measurements for only 5 primates, and yet we were able to recapitulate many of the findings in a recent study in humans with more than an order of magnitude more samples, a result that underscores the importance of repeated measures. This increase in power to determine differential expression derives from the ability to estimate individual-specific (and likely genetically controlled) levels of expression.

In light of the differing responses of the hosts to the malaria infection (i.e. one animal had to be euthanized and another also required a blood transfusion), I looked for prominent features that co-varied with the severity of disease. The animal that experienced the severe anemia and subsequent organ failure had the lowest MCH and MCV in the first two weeks of the experiment before blood-stage infection became apparent.

A recent study found that baseline hemoglobin levels were strongly inversely associated with risk for multiple organ dysfunction syndrome (MODS).[158] In this study, the animal that experienced the renal failure did not have the lowest hemoglobin levels at baseline; he was approximately average for the days before evidence of blood-stage parasitemia (days 0-12). As reported above, however, he did have much lower mean

corpuscular hemoglobin as well as lower corpuscular volume. The importance of these blood cell traits, especially MCH, is supported by a recent genome-wide association study (GWAS) performed in humans, which shows that genomic loci that are important in malaria resistance and pathology are enriched with SNPs that explain substantial amounts of the variance in blood cell traits.[39]

3.5.7 Caveats and limitations of this experimental design

While this experiment was able to unveil many important features of the host response to relapsing malaria, there were aspects of the design that prevented more precise dissection of mechanisms of response. For instance, there was a lack of temporal resolution to precisely understand the timing of the hematopoetic activation during the primary infection. From the CBC measurements, it is apparent that the pancytopenia of malaria is most prominent during the first blood-stage infection peak. Future studies could take more dense samples during the course of the first infection peak. Here, we were limited in the number of samples we could take because there was a maximum volume that we could extract from the primates over the course of a given time period, and we were sampling for numerous high-throughput technology cores.

3.5.8 Future studies

Lastly, I have herein left many lines of inquiry unexamined. First, the advantage of using RNA-seq (this study) versus microarrays (nearly all previous studies) to profile gene expression is that we can examine alternative splicing, antisense expression, and differentially abundance of non-coding transcripts. While these kinds of analysis may offer unique insights into the host response to malaria infection, it is difficult to make functional inference about affected pathways. For instance, if there are 100 non-coding transcripts that I detect in the samples and 20 are differentially expressed in the relapse peak but not the primary peak of parasitemia, what further inference

can I make about underlying molecular mechanisms. This kind of analysis would necessitate direct testing of importance of these transcripts by knock-down (say, by RNA interference).

3.5.9 Conclusions

Many pathways that are differentially regulated in this study with macaques after a primary malaria infection are qualitatively similar to those identified in humans using much larger sample sizes; altered pathways (including RIG-I, NOD, Toll-like, and complement pathways) have previously been identified in human studies as being enriched during malaria infection, which supports the use of the model system for the study of malaria in humans.[67]

Relapsing parasitemia peaks induce a much more restrained host transcriptional response demonstrating that primary versus relapsing parasitemias was qualitatively different.

Differential host clinical outcomes (e.g. severe malaria, death) appear to be related to measurable blood cell traits, namely corpuscular volume and hemoglobin.

CHAPTER IV

COMPOSITIONAL MODELLING OF *PLASMODIUM FALCIPARUM* ACROSS THE INTRA-ERYTHROCYTIC DEVELOPMENT CYCLE

4.1 Abstract

Microscopically, *Plasmodium* parasites take on three distinct states during progression in the intra-erythrocytic development cycle (IDC): ring, trophozoite, and schizont. The number of distinct transcriptional states is less clear and is currently still debated.[36, 83] Using a compositional modelling approach of discretized expression data, I estimated expression levels for a range of distinct transcriptional states and simultaneously dissected the proportion of each asexual stage in each hourly sample using previously published high-resolution expression data from Bozdech and colleagues.[23] I discovered that there is a stable solution of three distinct transcriptional states in the IDC, a result concordant with microscopic observations of infected blood samples. Lastly, I used the results from this compositional analysis to identify genes that are highly expressed in each distinct parasite life-stage, which will be used in later analyses.

4.2 Introduction

4.2.1 *Plasmodium* gene expression

Beginning with the publication of the transcriptome of the intra-erythrocytic development cycle (IDC) of *P. falciparum*, much effort has been invested to understanding the transcriptional dynamics of *Plasmodium* using high-throughput omics

technologies.[23] In the first paper that profiled parasite expression in a transcriptome-wide manner, Bozdech and colleagues showed that the majority of the parasite genome is cyclically regulated across the IDC, with each gene having a characteristic peak.[23] Using microarrays and probing the transcriptome hourly across its 48-hour cycle, they were able to identify genes that had their peak at different stages of development (Figure 21A). Subsequently, numerous microarray and RNA-seq studies were published that shed light on gene expression in *Plasmodium*.[122, 25, 58, 133, 43, 145, 24, 110, 89]

One important consideration for these kinds of expression analyses is that they are interrogating a *population* of cells and not one individual cell. And due to natural variability in parasite development times, each sample would reflect the average abundance of expression in a mixed population. A qualitatively similar problem arises when trying to determine the relative contributions of various cell types of a mixed tissue to the transcriptome. To determine the composition of a mixed sample of different cell types, various groups have implemented computational methods to perform sample deconvolution.[121, 131, 54, 49, 55, 84] The underlying statistical methodologies of these approaches are vastly different but have a similar underlying goal: determine the relative proportions of each cell type in a heterogeneous sample. Most of these previously implemented are deterministic and require *a priori* pure cell expression measurements. The former is a short-coming because these models are mathematically over-specified, that is, there are more equations than parameters to estimate. A probabilistic algorithm would allow a greater exploration of the solution space instead of deterministically arriving at a fixed solution. The latter is a difficulty especially when there are an unknown number of cell types in the mixture.

Malaria researchers have implemented similar statistical methods for the mathematical deconvolution of mixed samples. These methods have allowed researchers to both to get a clearer picture of the expression of the pure life stages and also to dissect what is occurring at the level of transcription in the parasite over the course

of an infection.

In a seminal work by Daily and colleagues,[36] researchers investigated the parasite expression profiles of malaria-infected human blood using non-negative matrix factorization (NMF), which is a spectral decomposition method that essentially performs factor analysis on the underlying data.[26, 82] Within this analysis, three distinct in vivo states were identified, and the genes partitioned by these three sets were enriched for significant pathways and molecular functions. One drawback to NMF is that it is not compositional; that is, there is no requirement that the components sum to one. In this sense, it is more similar to a factor analysis decomposition.

A re-analysis of this data by another group accounted for batch effects of microarrays and employed a numerical approach which was more compositional in nature.[83] In the Lemieux et al model, they assumed that all parasites in the blood were either in an asexual or sexual stage and subsequently estimated the proportions of each for each infected individual. Further, they used a maximum likelihood approach to estimate the particular stage of the asexual cycle that most parasites were in. Important to note, *P. falciparum* produces variable antigens which lead to sequestration of later-stage parasites (i.e. trophozoites and schizonts) in the microvasculature. From this re-analysis, all of the samples were abundant with ring stage asexual parasites, as was expected from the biology of this parasite.

The method of Lemieux and colleagues likely produces a reasonable result because the asexual parasites were all in approximately the same stage: rings. In other species like *P. vivax* and *P. cynomolgi*, however, the trophozoite and schizont stages do not sequester in the microvasculature, a behavior prominent in *P. falciparum* due to its exporting of variable antigens, a molecular mechanism first described in *P. knowlesi*.[64, 13, 141] Because their IDC position estimation is also more closely related to a mixture model (which assigns each sample to one of a many possible underlying groups), the implementation of this method would likely lead to very

large confidence interval around the maximum likelihood estimate in a blood sample from a species that did not sequester in later IDC stages. This is because in a species that does not sequester in later stages, the parasites in a sample may be in very different developmental stages.

The purpose of this study is three-fold. First, I explore a range of possible underlying transcriptional states for *P. falciparum* during its progression through the IDC. Next, I identify the number of discrete populations of life-stages in blood-stage *Plasmodium* that optimizes the likelihood of the underlying data. Last, I determine gene expression profile for each of the stages and identify genes that have their peak expression in each stage, which will be used in later down-stream analysis. Based on microscopic observations of the parasite, I hypothesized that there would be three distinct transcriptional states in blood-stage *Plasmodium*, which would correspond to ring, trophozoite and schizont stages. To achieve these aims, I have taken a full compositional modelling approach to dissect the proportion of each asexual stage in a given sample using the data from Bozdech et al, which contains expression measurements from across the entire IDC.[23]

4.3 Methods and materials

4.3.1 Dataset

Bozdech and colleagues reported the hourly gene expression profile across the 48 hour IDC of *P. falciparum*. I downloaded this dataset from the supplementary data available from the publication's website. The expression levels were determined by performing a competitive hybridization of the each time point against a pool of all of the time points. Two time points were not present in the dataset (23h and 29h) for unknown reasons. There were over 3000 genes with expression levels reported. This represents the first and perhaps highest resolution genome-wide expression profiling performed in *Plasmodium* for the *in vitro* IDC to date.

4.3.2 Multi-dimensional scaling

To visualize the cyclical nature of the expression levels in the IDC, I performed a multi-dimensional scaling of both the samples and the genes using R. First, the expression data were downloaded from the supplementary data of the paper. Gene-by-gene the expression levels were z-score transformed. Next, I calculated the Euclidean distances between the feature of interest (either genes or samples). Then I employed the cmdscale function using only two dimensions of scaling. The results were then visualized in R.

```
d.trans.zscore <- dist(trans.data.zscore.Pf)
fit.trans.zscore <- cmdscale(d.trans.zscore,eig=TRUE, k=2)
```

4.3.3 Developing a method for life-stage deconvolution of mixed cultures of *Plasmodium* using STRUCTURE

The aim of this study is to implement a statistical tool that allows for the estimation of the proportion of each blood stage in a mixed population of parasites using gene expression data. This problem is analogous to one encountered in population genetics. In human population genetics, a given individual may be a descendant of one or more historical populations, where each historical population has its own population-specific allele frequencies for single nucleotide polymorphisms (SNPs) and other genetic markers across the genome. Based on the allele frequencies of the historical populations, we can estimate the likelihood that an individual came from a given population using genome-wide genotypic information of the individual and the historical populations.

In most cases, however, there is no access to the genotypes of the historical populations. However, if a sample is large and diverse enough, it is possible to estimate the allele frequencies of the pure populations as well as the percent of the genome coming from the historical population for each sampled (genotyped) individual. Analogously

to gene expression data, we have many samples that are composed of a mixture of cells in different stages of development. Each pure cell type (parasite development stage) has a distinct expression profile with some genes expressed highly in one cell type and at a lower level in other cell types. As in the population genetic problem, we often do not have gene expression data for the pure cell types. In samples that we wish to interrogate, some (unknown) proportion of cells comes from each of the pure cell types. And the expression profile of the samples is a function of its cell type composition multiplied by the expression profile of each pure cell type.

$X_{ij} = \sum_{k=1}^{m} Z_{ik} * P_k$

X denotes the expression levels of genes coming from the parasite for each sample (n*m).

$$X_{n,m} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,m} \\ x_{2,1} & x_{2,2} & \dots & x_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,m} \end{bmatrix}$$

Z denotes the (unknown) vector of cell-type proportions for each of the k populations (n*l).

$$Z_{n,l} = \begin{bmatrix} z_{1,1} & z_{1,2} & \dots & z_{1,l} \\ z_{2,1} & z_{2,2} & \dots & z_{2,l} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n,1} & z_{n,2} & \dots & z_{n,l} \end{bmatrix}$$

P denotes the (unknown) average expression level for each gene from a given population (l*m).

$$P_{l,m} = \begin{bmatrix} y_{1,1} & y_{1,2} & \dots & y_{1,m} \\ y_{2,1} & y_{2,2} & \dots & y_{2,m} \\ \vdots & \vdots & \ddots & \vdots \\ y_{l,1} & y_{l,2} & \dots & y_{l,m} \end{bmatrix}$$

Where n is the number of blood samples, m is the number of genes, and l is the number of pure cell types.

The primary objective of this analysis was to implement a tool that allows for the estimation of the proportion of each blood stage in a mixed population. Since this cell type deconvolution problem is qualitatively similar to the problem of inferring population structure using genome-wide genotype data, I chose to re-purpose a widely-used software, STRUCTURE.[115]

To to make our continuous expression data compatible with STRUCTURE, I performed a tertile discretization. For each gene, the top third of all samples were assigned a value of two; the middle third of all samples were assigned a value of one; and the bottom third of all samples were assigned a value of zero. The discretized dataset (46 hourly samples, 3719 genes) was then input into the STRUCTURE graphical user interface. Another discretization was performed using a 10-80-10 division of samples; qualitatively similar results were obtained.

I ran STRUCTURE for k=3-8 with three independent replicates each, setting a random seed of 123 for reproducibility. Default parameters were used when available. The program was allowed 2000 iterations for burn-in followed by 1000 Markov-chain Monte Carlo (MCMC) iterations.

4.3.4 Identifying genes specific to each IDC life-stage

To find the genes that are specific to each distinct transcriptional state, I performed the following workflow. For every gene, I calculated where its maximum expression level occurred in the IDC. Then I identified the genes that had their maximum expression at each of the hours that had the highest proportion of each of the pure cell types or had their maximum expression within one hour of the highest proportion.

4.4 Results

4.4.1 Description of the dataset

To test the ability of this method to deconvolute samples of mixed cell type, I used a previously published dataset of gene expression profiling across the IDC of *P falciparum.*[23] All of the *Plasmodium* species profiled to date show a similar cyclic progression of gene expression across the IDC (Figure 21A).

In panel B, hours 1 and 2 along with hours 43-48 all cluster in the bottom center. This visual representation makes much more clear the relationship between the late schizonts and the early ring stages in terms of their closeness of gene expression profiles and serves to emphasize the cyclic nature of the IDC. Just as the samples cycle in a predictable way through the IDC, so do the genes (Figure 21C). Note that the coordinate space for multi-dimensional scaling is arbitrary; that is, the axes can rotated around the origin without losing the relative distances between the points.

4.4.2 Implementation of the probabilistic deconvolution method

To achieve the aforementioned aims, I have taken a full compositional modelling approach to dissect the proportion of each asexual stage in a given sample using the data from Bozdech et al, which contains expression measurements from across the entire IDC. I first transform the data from continuous to discrete values and then employ a probabilistic framework to iteratively estimate both the fraction of each life-stage in a given sample and also the expression level for each life stage. The resulting gene sets will be used in subsequent analyses.

After discretization of the expression dataset, I ran STRUCTURE for k=3-8 using three independent replications for each level of population number. For all values of $k > 3$, I found that at least one of the solutions converged to the k=3 solution (Figure 22); that is, the solutions are essentially identical. This is important because given the cyclic nature of gene expression, one of the possible outcomes would have been

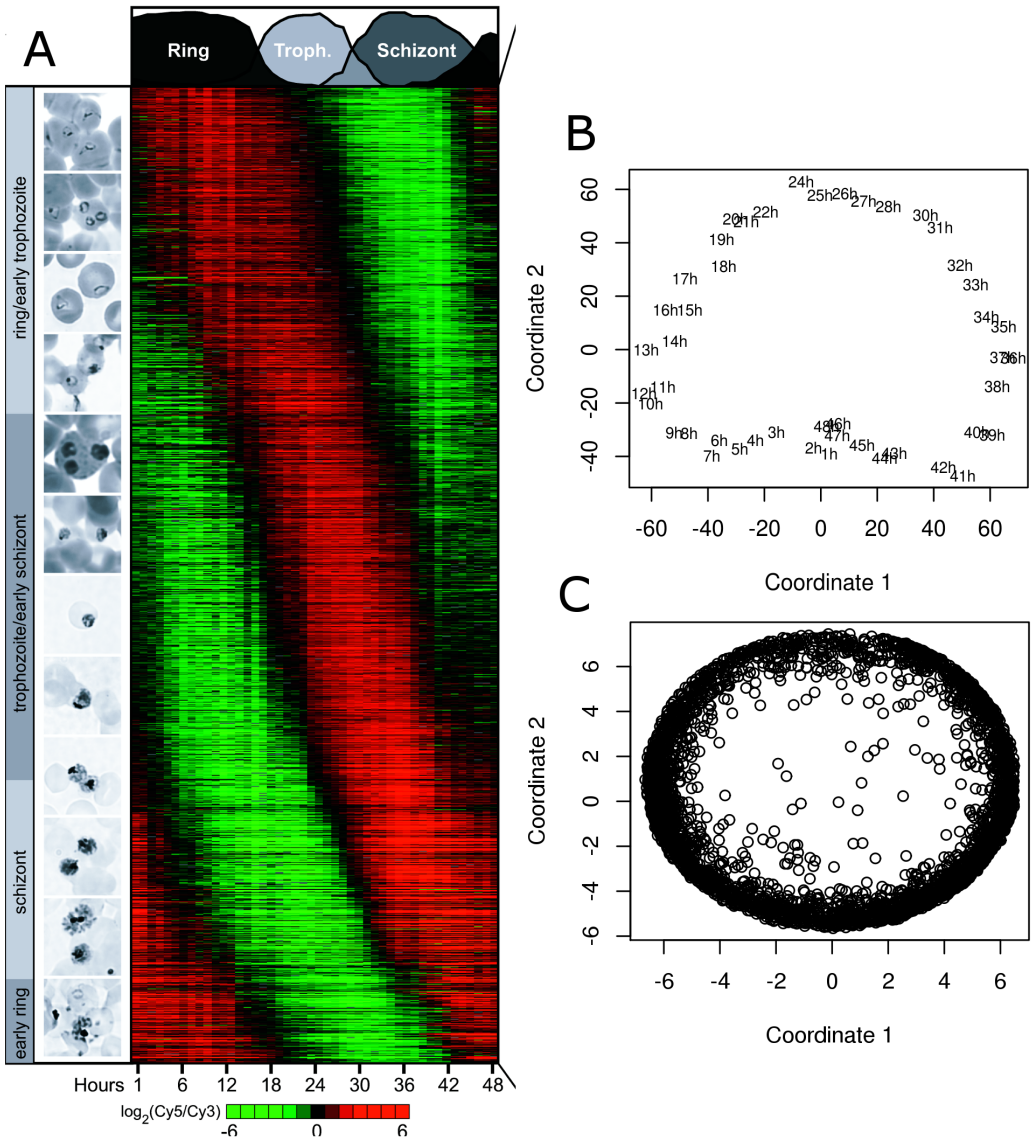

Figure 21: Expression profiling of the IDC of *Plasmodium falciparum*. (A) A heatmap of the gene expression levels of genes (y-axis) for samples (x-axis) taken at hourly intervals across the IDC. The extreme red color represents a 64-fold enrichment over the pooled average across the IDC, whereas the extreme green represents a 64-fold reduction over the pooled average. The left column of panel A shows microscopic views of the various life stages. The top part of panel A contains abundance estimates for each of the three unique cell type. Adapted from [23]. (B) and (C) Multi-dimensional scaling (MDS) plots of the samples and the genes, respectively. Samples are labelled by the hour at which they were taken.

convergence to a k=3 solution but with a phase shift. The fact that all of the k=3 results (as well as many of the k > 3 results) are of the same phase suggests that not

only that k=3 is a stable solution for the number of unique cell types in the IDC but also that we can assign the labels to the pure cell types based on our understanding of the IDC progression.

Interestingly, for all k > 3, at least one of the simulations collapsed to the k=3 solution. This suggests that there are really three distinct populations of parasites in the IDC. This hypothesis is consistent with microscopic observations which led to the naming of three separate phases of the IDC: ring, trophozoite, and schizont. Importantly, these estimates are strikingly similar to the microscopic stage estimates (the top portion of Figure 21A).

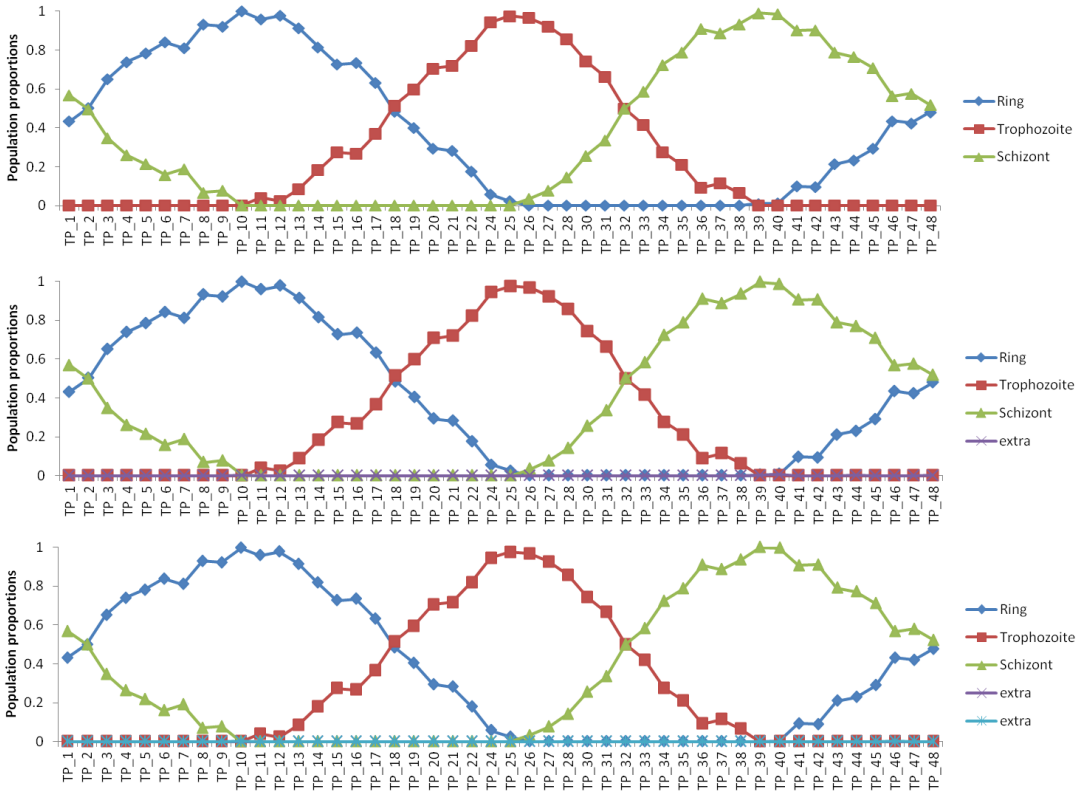

Figure 22: Compositional modelling of the hourly expression data of the IDC. The three panels represent the STRUCTURE output from k=3-5 from top to bottom, respectively. Importantly, the k=4 and k=5 populations collapse and converge to essentially the same solution as k=3. This demonstrates that k=3 is a stable solution for the number of unique cell types in the IDC.

4.4.3 Genes for each parasite stage

After the assigning the proportions of cell types present in each time point, I then determined which sample had the highest abundance of each pure cell type. Hours 10, 25, and 39 had the highest proportion (nearly 100%) for each of the three cell types which I have *post hoc* assigned the labels of ring, trophozoite, and schizont, respectively. I then found the genes that had their peak of expression at those peak samples or the sample time point on either side of it (i.e. 9-11 hours for rings, 24-26 hours for trophozoites, 38-40 hours for schizonts). These genes were considered stage-specific and were compiled into gene sets for use in future analyses.

Table 2: Parasite stage-specific genes.

IDC stage	Number of genes	putative enriched functions
Ring	426	ribosomal and ring-exported proteins
Trophozoite	405	DNA polymerase, helicase, mismatch repair
Schizont	184	glideosome- and rhoptry-associated proteins

There were 426, 405, and 184 genes in each of the stages of ring, trophozoite, and schizont, respectively. The majority of all genes were annotated as coding for conserved proteins of unknown function in plasmodb. While I did not perform a significance analysis of gene set enrichments for these gene lists, the gene names indicate that the genes with peaks in each of the stages are coherent for the biological functions that occur therein (Table 2). For instance, genes with a peak in the ring stage are annotated as ribosomal components and ring-exported proteins. In the ring stage, the parasite has just successfully invaded a new red blood cell and up-regulates the transcriptional program that had been turned off for schizogeny. In the trophozoite stage, genes associated with genome replication and mitochondria-associated proteins were present. In the trophozoite stage, the parasite the energy- and resource-intensive growth and division that will take place in the schizont stage. The schizont stage gene set contains genes that compose the glideosome and rhoptry, which are essential for successful sinvasion of a new RBC after the schizonts rupture

the infected red blood cells (iRBC) and release merozoites.

4.5 Discussion

Herein, I have shown that the asexual IDC of *P. falciparum* has a stable mathematical solution of three distinct life stages, which correspond well to the stages observed microscopically. I further compiled lists of genes that have a peak expression at each of the three stages; these gene lists will be used in subsequent analyses.

While this probabilistic framework successfully recapitulated independent stage assignment and therefore demonstrated its power, there are many caveats to this method. First, it requires dense, high-resolution, time points throughout the IDC to accurately estimate the gene expression of pure populations. Next, it is not easily transferable from microarray to RNA-seq data, or even between different microarray platforms. The Bozdech et al microarray dataset for *P. falciparum* was made using a two-dye competitive hybridization.[23] As a result, it is difficult to use this dataset together with, say, the dataset from Daily et al which used a single dye microarray and therefore captured absolute expression levels.[37]

Concurrent to our development of this discretized method of inferring the relative abundance of parasite stages in a given sample, another group developed and implemented a related method for *Plasmodium*.[70] In their method, they use linear regression and *a priori* information about gene expression at each life stage to predict the relative abundances. Superficially, their method and the one implemented herein appear similar. However, there are many differences, including the primary objectives of the work. In my work, I wanted to determine the number of distinct transcriptional states in the *Plasmodium* IDC. In their study, the goal was to determine the level of each life-stage in a mixed sample. My method has the advantage that it does not make *a priori* assumptions about the number of classes, nor does it require *a priori* information about the expression levels of each gene for the various parasite

blood-stages. However, my method must iteratively estimate both the expression levels of the pure populations and the composition of each of the samples. Additionally, their linear regression method is based on the minimization of a least squares term, which makes it sensitive to outlier genes with extremely different expression, whereas my probabilistic, discretized model is not susceptible to such outlier genes. As a disadvantage of my approach, in the discretization step, an essential part of my implementation, information is lost. However, on aggregate across all genes, sufficient information was preserved to recapitulate microscopic observations.

In conclusion, I have implemented a discretized, probabilistic method for gene expression deconvolution, and subsequently identified the three microscopically observed parasite stages. Using these three populations, I extracted genes whose expression is characteristic of each stage.

CHAPTER V

GLOBAL SHIFTS IN PARASITE GENE EXPRESSION PROFILES ACROSS AN INFECTION TIME-COURSE

5.1 Abstract

Gene expression of parasites from the genus *Plasmodium* has been extensively studied. However, to date, there has been little investigation of parasite expression over the multiple infection cycles in an *in vivo* setting. Basic questions concerning general patterns of gene regulation remain unanswered, especially concerning parasite expression changes in response to anti-malarial drugs (AMDs) and across primary versus secondary parasitemias. Herein, I examine parasite gene expression from *P. cynomolgi* in rhesus macaque (*Macaca mulatta*) hosts. Clustering of parasite gene expression demonstrate a distinct profile in the relapsing parasites compared to the primary parasitemias, which includes a shift away from sexual stages in secondary parasitemias.

5.2 Introduction

The gene expression program of the intra-erythrocytic development cycle (IDC) of malaria parasites from the genus *Plasmodium* has been extensively studied *in vitro* and has been shown to be robust to perturbation.[23, 24, 46, 97, 8] *In vivo*, however, the parasite needs to do more than just grow and divide especially in response to differential host pressures (e.g. immune system). To cope with this dynamic within-host environment, it would benefit the parasite to be capable of responding appropriately. Furthermore, host clinical parameters and outcomes (e.g. anemia and

organ failure) are much less severe in relapsing malaria episodes compared to the primary parasitemia. And whether host outcomes are associated with changing parasite expression programs has yet to be investigated. Herein, I explore the parasite transcriptional response of a *P. cynomolgi* infection of rhesus macaques (*Macaca mulatta*) across a 100-day infection cycle and interrogate the correlation between the host and parasite transcriptomes.

5.2.1 Expression profiling of *Plasmodium*

Bozdech et al paved the way for high-throughput analysis of gene expression in *Plasmodium*.[23] Subsequently, it was shown that even under the selective pressure of antimalarial drugs, the parasite shows an inability to alter its transcriptional program.[46] Whereas most previous studies were *in vitro*, Daily et al looked at *ex vivo* parasite expression from infected humans and found that there are only a few distinct transcriptional states of the parasite in the blood.[36]

5.2.2 Relapsing malaria

While the human parasite *P. falciparum* is considered the most deadly malaria parasite, *P. vivax* causes close to 100 million cases of malaria each year. *P. vivax* is the most common malaria parasite outside of Africa and pervades in tropical regions of the Americas and the Asian Pacific. In spite of its impact on human mortality and morbidity, this important human malaria parasite remains under-studied because of its inability to grow in an *in vitro* culture system.[24] As discussed in a previous chapter, it is not ethical to infect humans with *P. vivax* and follow the infection over multiple peaks of parasitemia. As a result, malaria models in non-human primates are used to gain insight in to the behavior of the human malaria parasites and their within-host dynamics.

In this experiment, we use a *P. cynomolgi* infection of *M. mulatta* as a model for *P. vivax* infection of humans, specifically with the goal of better understanding

parasite expression changes during relapse. A relapse occurs when the dormant hypnozoite form of the parasite, which exists in the host liver, is re-activated and causes a distinct blood-stage parasitemia. The host clinical parameters are usually much more mild during relapse compared to the primary infection. Sometimes the relapsing parasitemia produces no noticeable external symptoms, and the positive blood-smear is the only indication of a relapse.[163]

5.2.3 Experimental design overview

Five male rhesus macaques (*Macaca mulatta*) were profiled over the course of a 100-day experiment after being injected with purified sporozoites of the species *P. cynomolgi* on day 0 of this control experiment. Complete blood counts were performed daily, and microscopic quantification of parasitemia was performed. Before injection, the time point 1 (TP1) samples were taken. Then, on approximately days 20, 26, 53, 59, 89, and 96, blood and marrow samples were collected for TP2-7, respectively. A sub-curative dose (1mg/kg) of artemether was given at TP2 to three animals (RFa14, RFv14, and RMe14) to stem the increases in parasitemia. At TP3 and TP4, all animals received an 8-day course of artemether: day 1 (4mg/kg); days 2-8 (2mg/kg). At the end of the experiment, all animals were given fully-curative doses of primaquine/chloroquine. Animal treatment was all approved by the IUCAC of Emory University.

5.2.4 Motivating hypotheses

In a previous chapter, I showed that the *host* transcriptional response is highly altered between primary and relapsing parasitemias. Now, I will investigate the differences in the parasite transcriptome across multiple parasitemia peaks.

The motivation question for this chapter is: are there differentially expressed parasite genes between primary and relapsing parasitemias? Previous experiments

have shown that even when faced with severe artificial selection pressures (e.g. anti-malarial drug administration), malaria parasites do not alter their transcriptional profile.[46] Given the result from this previous *in vitro* study, I expect that the transcriptional profile of the parasites will not differ over the two parasitemia peak types: primary and relapse.

The alternative hypothesis is that the parasite transcriptional profiles differ between the primary and relapsing parasitemias. Evidence for this alternate hypothesis would suggest that the parasite alters its transcriptional profile as a direct result of having passed through the hypnozoite stage; alternatively the differential parasite transcriptome may be an indirect result of attenuated host immune response to the relapsing parasitemia. Support for the alternative hypothesis would be indicative of yet another layer of complexity of transcriptional regulation in malaria parasites.

5.2.5 Chapter outline

In this chapter, I first describe the whole genome re-sequencing of the *P. cynomolgi* strain that will be used in this experiment to confirm its identity. Next, I qualitatively explore the structure of the expression profiles from the samples of the different time points. Then, I perform differential gene expression analysis between the defined experimental groups, and subsequently perform gene set enrichment to unravel the differences in parasite expression across primary versus relapsing infection peaks. I also explore the trajectories of selected multi-gene families.

5.3 Methods and materials

5.3.1 Whole-genome resequencing for *Plasmodium cynomolgi* B strain

To confirm the identity of the *P. cynomolgi* strain that would be used in the infection experiment, a whole-genome resequencing of infected red blood cells was performed. Genomic DNA was extracted from a whole-blood sample containing infected RBCs. The DNA was sequenced on a HiSeq2000, and generated approximately 100 million

paired-end reads. After inspecting the FastQC output files, there was no detectable contamination of library preparation primers, and subsequently, the raw reads were used in the mapping step.

Bowtie2 was used for mapping the reads to the *P.cynomolgi* reference genome with both the –end-to-end and the –local alignment options, separately. The two methods performed similarly in terms of the number of SNPs and InDels that they called. As expected, the –end-to-end flag (which is more stringent in mapping) mapped fewer reads and called fewer SNPs. However, because of the unfinished nature of the genome, the –local option is most likely the better choice and was used for reporting statistics in this report.

Depth of sequence coverage is an important metric because the deeper a region has been sequenced, the more confident we can be in the genotype calls in that region. [1] Even with only 6 million reads mapping to the *P.cynomolgi* genome, 93.4% (24,457,978/26,180,000*100%) of the nucleotides of the genome are covered with >=20X sequencing depth.

Samtools mpileup was used in series with varscan to call the variants. These are the options for VarScan. In essence, it requires, at least 8 reads of sequencing depth to make a call at a given base. As stated above, nearly all of the genome meets this threshold. Next, there must be at least 2 reads supporting the variant for it to be called. This is an important threshold. Otherwise, sequencing error could dominate the SNPs calls. Even though the error rate is only 0.001, with millions of reads, the likelihood of SNPs would be high. The likelihood of seeing 2 reads support the same SNP call by sequencing error is $0.001^2/3$, [2] and so we would only expect about 8 false positives over the whole genome. This suggests that most of the 1031 SNPs that

[1] Say an allele exists in the population at a frequency of 10%. If a region of the genome is only covered by 10 reads, then the chances of the minor allele being missed is $(.90)^{10} \approx 0.35$; whereas if it had been covered by 20 reads, the chances of missing it are only $(.90)^{20} \approx 0.12$.

[2] The average likelihood of a miss-called base squared for the two occurrences divided by three, the probability of the two miss-called bases being the same (incorrect) base.

we identified are likely true positives; these polymorphisms would be the standing genetic variation in this particular strain.

5.3.2 Transcriptome analysis

Library preparations, read mapping, and expression quantification were all performed as described in a previous chapter. Briefly, total RNA was extracted from whole blood from samples taken across the 100-day experiment, and samples were dosed with spike-in control RNA as described in a previous chapter. Due to poor RNA quality, one sample was not sequenced (RFv14 at TP2). mRNA was enriched from total RNA using poly-dT beads. Libraries were bar-coded and sequenced on an Illumina HiSeq2000 generating approximately 50 million paired-end reads per sample. After sequencing, reads were mapped to a combined reference genome including host and parasite genomes. HTSeq was used to assign read counts for each annotated parasite gene.

After quantification of read counts mapping to annotated genes, I then calculated the total number of reads from each sample that mapped to an annotated gene. The samples taken TP1 before infection should have no parasite reads. The number of reads mapping to annotated parasite genes was, in fact, very low for these uninfected time points (6-11,835; median 3,103). To only include samples with sufficient expression depth of coverage to accurately estimate expression, a minimum of 500,000 reads was required for a sample to be considered as having parasite expression. Ten samples passed this threshold, and the library with the lowest read depth had at least eight reads for 3,980 genes ($\approx 70\%$ of all genes).

In spite of the large variations in the *Plasmodium* transcriptome due to the cyclic nature of expression across the IDC, the high correlation of gene expression across most samples suggested that the robust library size estimation method of DESeq would yield good scaling factors for making libraries comparable. To normalize the

libraries, I used the DESeq method also used in previous chapters.[6]

5.3.3 Down-sampling methodology

Of the 10 samples with more than 500,000 uniquely mapping parasite reads, there was wide variation in the read depth (approximate range 750,000-29,000,000). Reads from all libraries were probabilistically down-sampled to the depth of the library with the fewest reads. Briefly, I first determined the number of reads in the smallest library (749,249). Then, for each library, I determined the number of reads. Then, for every read in each library, the probability of it remaining in the down-sampled library was equal to (number of reads in smallest library) divided by (number of reads in the present library).

The R code is given below:

```
rm(list=ls())

expression.counts = read.table( "/home/kevin/work/Research/malaria/
Experiment04/Ex04_parasite_counts.txt", header=TRUE,sep="\t",row.names=1)

#remove columns (that is, libraries/time points) thhat have low read counts
expression.counts <- expression.counts[which(apply(expression.counts,2,sum)>500000
expression.probs <-t(t(expression.counts)/apply(expression.counts,2,sum))

#initialize the down-sampled expression matrix
new.counts <- expression.probs
for(i in 1:nrow(new.counts))
{
  for(j in 1:ncol(new.counts))
```

```
  {
    new.counts[i,j]<-0
  }
}

#calculate number of reads in each column
total.reads <- apply(expression.counts,2,sum)

#calculate (or input) how many reads to sample
#by default, this finds the library with the fewest number
of reads and sets that as the value; this can be modified to fit needs.
reads.to.sample <- min(total.reads)

#loop over all of the libraries
for( i in colnames(new.counts))
{
  #calculate the ratio of keeping a read for each library
  keep.ratio <- reads.to.sample/total.reads[i]
  print (keep.ratio)

  #run the sampler
  for( j in rownames(new.counts))
{
    if(expression.counts[j,i]>0)
    {
      print(j)
      for(k in 1:expression.counts[j,i])
```

```
      {
        rand.num <- runif(1);
        if(rand.num < keep.ratio)
        {
          new.counts[j,i] <- new.counts[j,i] + 1
        }
      }
    }
}
}
```

5.3.4 Differential gene expression

After observing the clusters of the time points into natural, discrete groups, I allowed the groups to be those used for down-stream analysis. Four samples from TP2 were grouped together (TP2), three samples from TP3 were grouped together (TP3), and all the samples from the relapsing parasitemias (one sample each from TP4, TP6, and TP7) were grouped together (secondary).

Then, one-way ANOVA was performed on these samples using the aforementioned groupings. Unlike in the host expression analysis, the effect of animal was not used as a random effect due to the small number of samples in the analysis.

5.3.5 Gene set enrichment analysis

To perform gene set enrichment analysis of the parasite gene expression data, we had to first establish and populate gene sets. Since the majority of functional studies in *Plasmodium* have been performed for *P. falciparum*, I investigated the *P. falciparum* literature and made two different gene set groups: stage-specific gene sets, and GO term gene sets.

The previously-defined gene sets (Chapter DECONVOLUTION) for the three IDC stages (ring, troph and schizont) were converted from *P. falciparum* to *P. cynomolgi* using syntenic orthologous genes. Similarly, a list of genes that show gametocyte-specific expression were extracted and converted to *P. cynomolgi* gene names.[164] Gene sets of GO terms were downloaded for *P. falciparum* and converted to *P. cynomolgi*, as well.

The following work-flow was used to obtain the syntenic orthologs between these species. In `PlasmoDB`, all gene from *P. cynomolgi* were selected. Then, a step was added: transform by orthology. Only genes that were annotated as syntenic orthologs between these two species were used. Then, the gene-correspondences were downloaded and subsequently used to identify the *P. cynomolgi* orthologs for each gene set.

Similarly, GO annotations for the entire genome were downloaded for *P. falciparum* and the same methodology was used to identify corresponding genes.

After generating the gene sets, the t-statistic from each of the three contrasts were used to find gene sets that were coherently differentially regulated between two or more of the groups. All possible pair-wise contrasts were investigated: TP2 versus TP3, TP2 versus secondary, and TP3 versus secondary.

5.3.6 Gene group trajectories

Plasmodium helical interspersed sub-telomeric (PHIST) proteins play an important role in host-parasite interactions during malaria infection. Some PHIST proteins contain a red blood cell cytoskeleton binding motif, which may facilitate both the trafficking of parasite proteins to the iRBC cell membrane, as well as the remodelling and eventually the rupturing of the iRBC to release the infectious merozoites.[112, 91] Tryptophan-rich antigens are another abundant multi-gene family in *Plasmodium* species and which play a role in RBC invasion of merozoites.

To characterize the expression of these multi-gene families over the time course of the infection, I extracted genes in these families and performed principal component analysis on each gene set, separately. I use the percent variance explained by the first PC as a measure of the coherence of the expression (that is, co-expression) of each family, and plot the value of the first PC for each sample over time.

5.4 Results

5.4.1 Whole-genome resequencing for *Plasmodium cynomolgi* B strain

Integral to the analysis of a transcriptomic dataset (using RNA-seq reads) is the presence of a high-quality reference genome for the species of interest. For *P. cynomolgi* the reference genome was initially published in 2012 by Tanabe and colleagues.[143]. This group sequenced B strain, which is the *P. cynomolgi* reference genome available in PlasmoDB and is the reference genome used herein.[11, 10]

Before the beginning of the 100-day infection experiment, I first undertook the task of determining the identity of the *P. cynomolgi* strain that would be used. The goal was to confirm that it was the same strain as that of the reference genome (B strain), as parasites in culture have been observed to accumulate mutations in certain genes; and sufficiently different genomes could result in poor mapping of transcriptomic reads and therefore mis-quantification of expression.

To confirm the identity of our parasite strain, DNA was extracted from an infected blood sample containing *P. cynomolgi*, and was then sequenced on an Illumina HiSeq2000. Of the 109,375,524 paired-end reads, 6.3% of the reads map to the parasite genome, using the –local flag[3]. To account for the reads that did not map to the *P.cynomolgi* genome, I also performed a mapping of the reads to the *Macaca mulatta*

[3]Using the –end-to-end flag, 4.6% of the reads map to the genome. This discrepancy is occurs because of the fragmented and incomplete nature of the genome assembly with which we are working. The –end-to-end flag requires that the entire read map exactly to the genome. This is difficult for this genome because there are many blocks of N's, which the mapping software will count as mismatches, and subsequently, a mapping, even though it is best, will not be considered because it is below a certain threshold.

genome. Indeed, approximately 94% of the reads map to the primate host's genome, which makes up for the difference.

The next step in confirming the strain's identity is the calling of polymorphisms. Since the parasite was isolated more than 30 years ago and has been maintained through culturing in monkeys and subsequent cryopreservation, I anticipated that there would be some divergence between the reference genome and the genome of the parasites that would be used for inoculation. After mapping the reads to the genome, I called single nucleotide polymorphisms (SNPs) using VarScan.[76] Remarkably, no polymorphisms were detected in the 14 nuclear chromosomes (Figure 23).

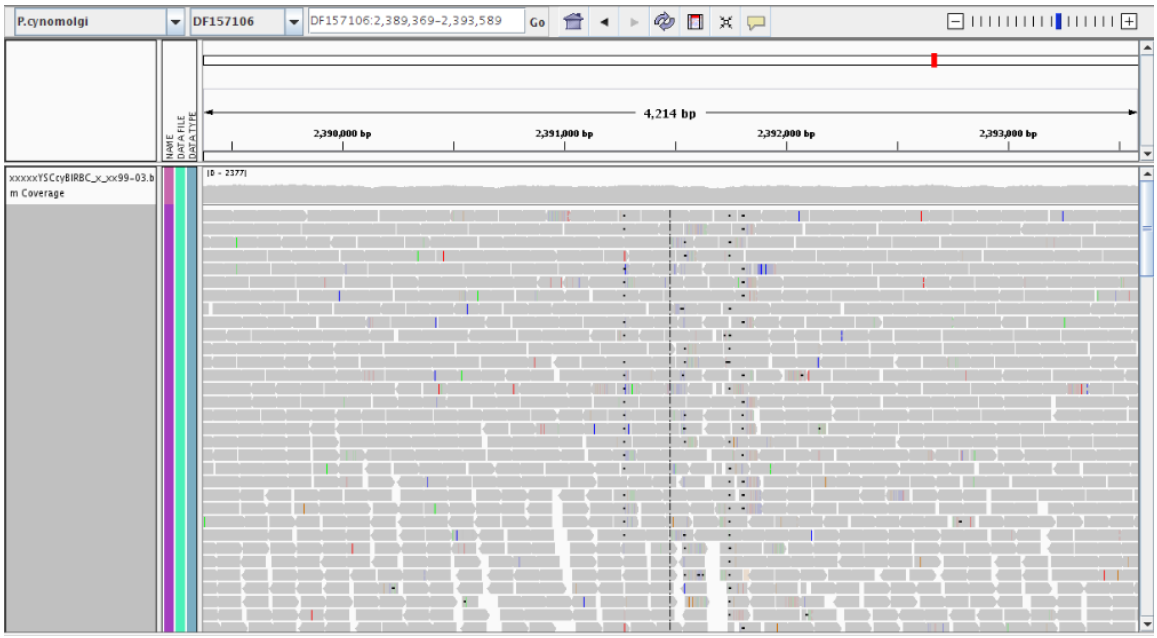

Figure 23: Integrative Genome Viewer screen capture shows no polymorphisms. In the reads of this genomic region, there appear to be some sequence polymorphisms which may be due either to acquired mutations due to passaging of the parasite or sequence errors.

While the *P. cynomolgi* genome is relatively well-assembled, hundreds of large contigs still remain unplaced, and within many of these unplaced contigs, there were some called polymorphisms. Most of these contigs were difficult to place because they are members of high-copy-number gene families and therefore share high levels

of sequence similarity. In general, I do not anticipate that these polymorphisms in unplaced contigs will qualitatively alter conclusions reached from this work. I note that in other species (e.g. *P. falciparum* and *P. knowlesi*), the variable antigen genes, which are responsible for immune evasion, represent a high-copy-number gene family and therefore polymorphisms in unplaced contigs may be important in studies performed in these species. In light of this result, I conclude that the *P. cynomolgi* strain to be used in the 100-day infection experiment is genotypically very close to the *P. cynomolgi* reference genome.

5.4.2 100-day experimental summary

As described in a previous chapter, the experimental infection of *M. mulatta* with *P. cynomolgi* lasted 100 days, and spanned multiple parasitemia peaks for the four macaques that survived the primary infection. Blood samples were taken daily to quantify parasitemia using manual counting of the infected RBCs on blood smears. The parasitemia count data show that the primary and subsequent relapse infections occurred at approximately the same times across all animals supporting the consistency of this model infection.

To stem the increasing parasitemias, a one-day, sub-curative dose of artemether was given to three of the primates (RFa14, RFv14, RMe14) after taking the blood and bone marrow samples for TP2. Seven days later, TP3 was taken (experimental day 26 for all animals except RFv14 which had TP3 taken on experimental day 23 immediately before euthanasia). The relapsing infections occurred around day 60 for three of the four animals (RFa14, RIc14, RSb14), and another relapse occurred near day 90 for all four surviving animals. Parasitemias for the relapsing infections were much lower than the primary parasitemia peak (between $10^{1.5}$ and $10^{4.5}$ and between 10^5 and 10^6 parasites per microliter, respectively). Note that the limit of detection is approximately 10^3 parasites per microliter, and therefore readings below these levels

have a high measurement error.

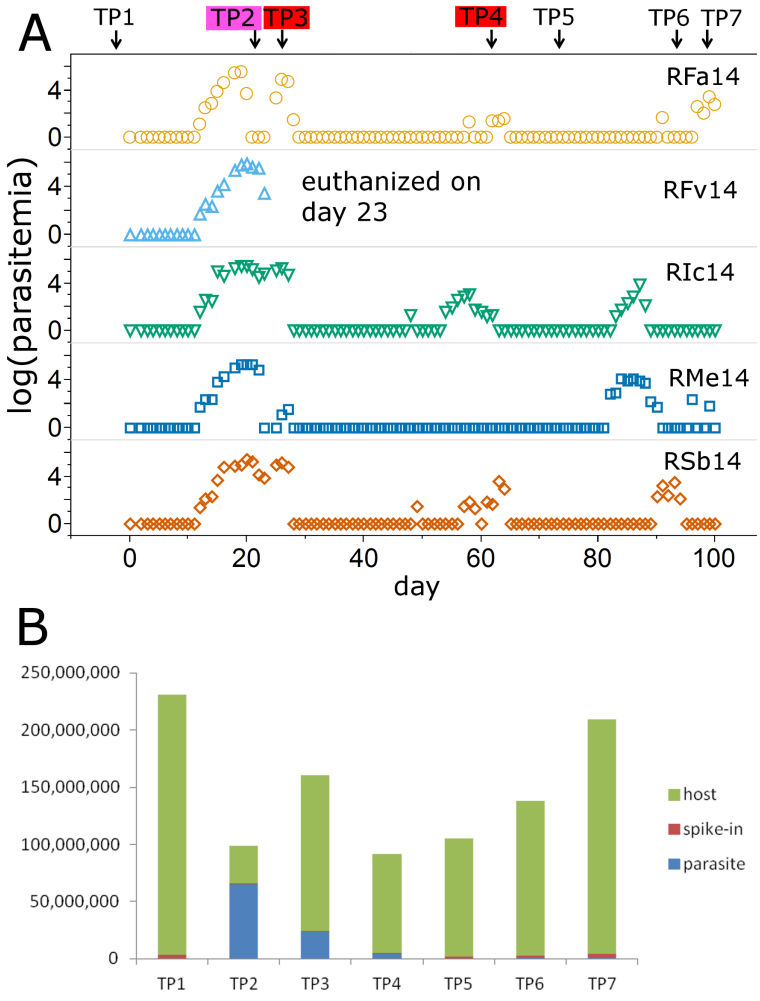

Figure 24: Parasitemia across the infection and transcriptome read depth. (A) Parasitemia for each of the animals across the 100-day experiment. A single-day dose of artemether was given to three of the primates (RFa14, RFv14, RMe14) at TP2 (pink box), and full 8-day courses were given to all animals at TP3 and TP4 (red boxes). Parasitemia was much higher during the first infection relative to the relapsing infections. (B) The number of reads mapping to each of the three sources of RNA: the host (macaque), the parasite (*P. cynomolgi*, and the spike-in control RNA. The highest levels of parasite RNA occur at TP2 and TP3, which corresponds to the points of highest parasite density.

After collecting the samples, RNA was extracted from them and high-throughput sequencing was performed. RNA-seq provides a digital read-out of the transcriptomic

state of a sample. While this technology offers many advantages over other gene expression quantification techniques, it also has an inherent limitation: insufficient sequence read depth prevents accurate quantification of gene expression genome-wide. At low read depths, technical variance dominates, preventing the accurate estimation of biological variation. With lower coverage, the coefficient of variation in the expression measurements increases, subsequently decreasing power to detect differential expression. Furthermore, since most normalization techniques, including the method used herein, are dependent upon the accurate estimation of many genes, I have only included samples with at least a half-million parasite reads to ensure robust inference of differential expression.

Based on this threshold, TP2 and TP3 have enough mapped parasite reads to adequately quantify expression in four and three samples, respectively; in the relapsing time points, there is only one sample each from TP4 (RSb14), TP6 (RSb14) and TP7 (RFa14) with sufficient read depth. As a matter of convention, I refer to the three relapsing time points as secondary parasitemias.

5.4.3 Clustering of the parasites

After setting the sequence depth threshold for inclusion in the analysis (see Methods), I determined the general topology of the relationship between the parasites at each time point. A hierarchical clustering of the samples shows that samples from TP2 and TP3 cluster mostly within time point, and the relapsing samples (one from each of TP4, TP6, and TP7) cluster separately (Figure 25A). Given that parasite sequence depth ranged from about 750,000 reads to more than 29 million and that the relapsing time points had few reads in comparison to TP2 and TP3, sequence read depth was confounded with time point. To verify that the separate clustering of the samples by time point was not due to differences in read depth, I coded a down-sampler in R, which finds the sample with the fewest reads and subsequently probabilistically

reduces the number of reads in all of the other samples' libraries to that level. After down-sampling, the relapsing time points cluster together, and samples from TP2 and TP3 still cluster in a similar way (Figure 25B).

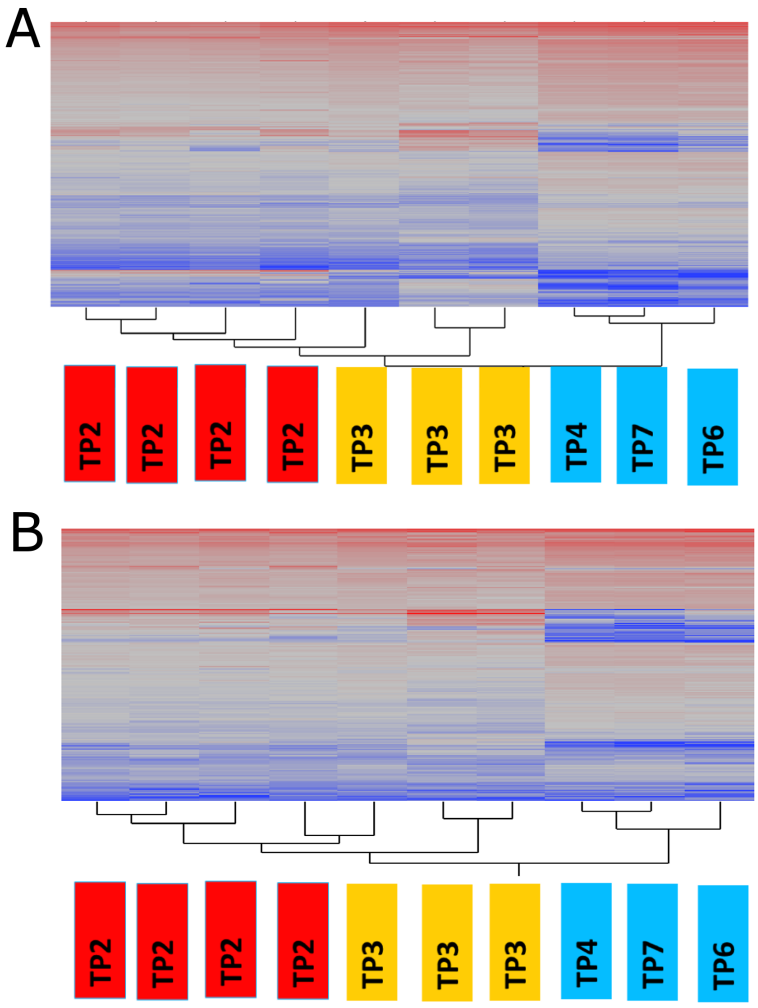

Figure 25: Primary parasitemia expression profiles cluster away from relapse profiles. (A) A heatmap of ten libraries with sufficient parasite read depth to be considered expressed. The samples from the first parasitemia peak (TP2 and TP3, in red and gold, respectively) hierarchically clustered mostly within time point. The deepest branch in the clustering is between the relapsing parasite profiles (light blue), and the primary parasitemia samples. (B) Same as in (A) but using the down-sampled read counts for each library.

After defining the three groups of parasite samples (TP2, TP3, and secondary), I explored the general correlation between genes in each group. In general, all three

groups had strong linear correlations with each other (Figure 26). The secondary time points, however, have a subset of genes that are below line of identity ($x = y$, Figure 26B,C). The genes that are down-regulated in the contrast of secondary versus TP2 (Figure 26B) are colored in black. There is a large overlap of these genes with genes down-regulated in the contrast of secondary versus TP3 (Figure 26C).

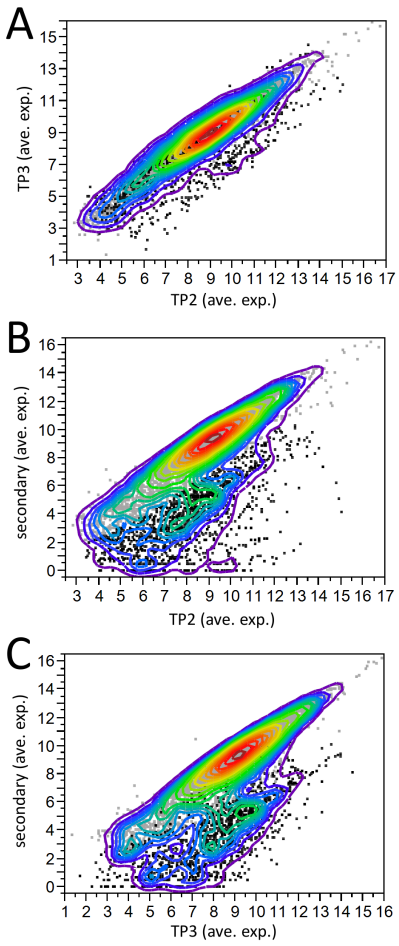

Figure 26: Correlation of average expression between experimental groups across all genes. (A) TP3 versus TP2, (B) secondary versus TP2, and (C) secondary versus TP3. Non-parametric density gradients are overlaid in colored lines. Genes that show a down-regulation in the secondary infection relative to TP2 are colored in black; all other genes colored in grey

5.4.4 Differential gene expression across parasitemia peaks

After the initial exploratory analysis of the parasite expression dataset, I performed differential gene expression to identify genes that are differentially regulated across the three peaks of parasitemia. Between TP2 (peak of parasitemia during first blood-stage infection) and TP3 (still during the first blood-stage infection) there are numerous differentially expressed genes (Figure 27A). However, the vast majority of the genes that are found to be differentially expressed are altered when comparing the secondary infection time points to those of either TP2 or TP3 (Figure 27B,C). Significant genes are colored by up- or down-regulation (red or blue, respectively) in secondary parasitemias relative to TP2 (B). In the plot of secondary versus TP3 (Figure 27C), many of the colored genes maintain both directional coherence and significance. Next, the nature of these genes that are down-regulated in secondary parasitemias was investigated.

5.4.5 Gene set enrichment analysis for *Plasmodium cynomolgi*

The primary objective of this study was to elucidate the differences in the parasite transcriptional state between primary and relapsing peaks of parasitemia. In the previous section, numerous genes were identified that fit this criteria. In this section, I investigated the functional enrichment of the shifts in gene expression by performing a gene set enrichment analysis of both life-stage-specific gene sets as well as annotated GO gene sets.

In the ring and trophozoite life-stages, there is a subtle yet significant up-regulation in the relapsing (secondary) parasitemias compared to either of the primary parasitemia groups (TP2 and TP3) (Figure 28). Much more distinct is the down-regulation in the relapsing time points of both the schizonts and gametocytes. This down-regulation of gametocyte-specific genes in the secondary infection time points

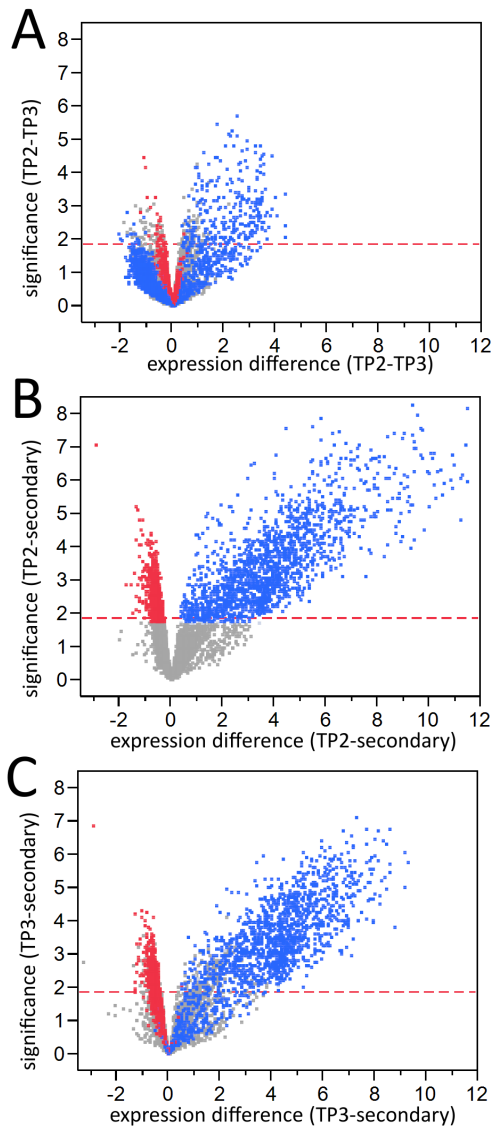

Figure 27: Volcano plots showing the magnitude and significance of differential gene expression across the three experimental groups. (A) TP3 versus TP2, (B) secondary versus TP2, and (C) secondary versus TP3. The the significance level of the difference (-log(p-value), y-axis) is plotted against the log-fold change in expression (x-axis). Significant genes are colored by up- or down-regulation (red or blue, respectively) in secondary parasitemias relative to TP2 (B).

demonstrates that there is a clear difference in the relative abundances of sexually-committed parasites circulating in the blood during relapse.

Additionally, from the bottom panel of the enrichment plots, I infer that gametocyte-specific genes reach their peak at TP3, since they are up-regulated at TP3 with respect to both TP2 and relapsing time points. Importantly, after samples were taken at TP2, three primates were given a sub-curative, one-day dose of artemether. TP3 was then taken seven days later. The two samples (of the three samples at TP3 for which there is parasite expression data) in which the primate did not receive the sub-curative dose of artemether are the ones that have the highest gametocyte expression. The sample from the primate which did receive the sub-curative dose of artemether shows a parasite expression profile much more similar to those of TP2 than of the samples from TP3 that did not receive the AMD. What caused the shift in relative abundance from asexual to sexual stages from TP2 to TP3 is unclear, but may have been due to either sustained high densities of parasites in the blood, or host factors. The exact molecular mechanism by which the parasite controls cell fate determination is yet unknown.

5.4.6 Effect artemether on parasite life-stage abundance

To investigate the effect of the anti-malarial drug artemether on the parasite expression profiles, I used the life-stage specific genes to contrast the samples at TP3: one which came from an animal that was treated with artemether (RFa14), and two from animals that were not treated with artemether (RIc14 and RSb14); animals RFv14 and RMe14 were also treated with the AMD but samples from these animals did not have sufficient parasite read depth for expression quantification (see Methods). As before, I performed principal component analysis using the genes from each of the life-stages. In this case, I use the first two principal components to plot each of the samples in PC space.

In both the schizont- and the gametocyte-specific genes, the outlier point for TP3 is from the primate that received artemether (Figure 29C,D). The two parasite

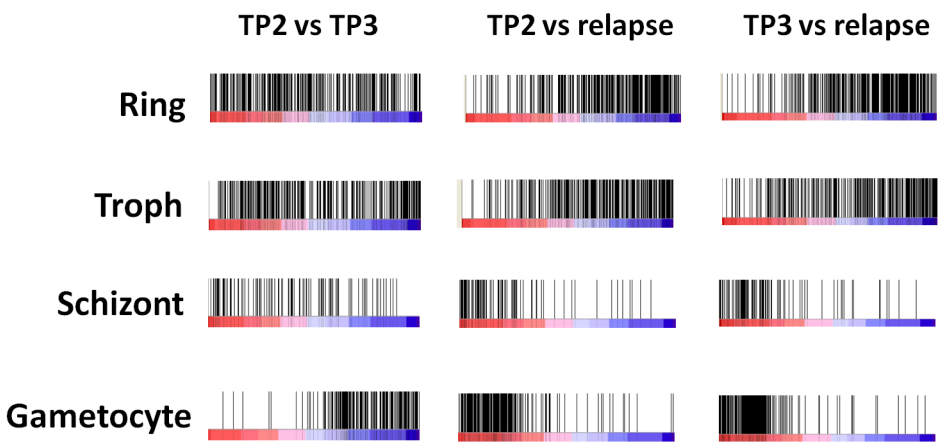

Figure 28: Life-stage-specific gene set enrichment. For each pair-wise contrast of the three experimental groups, the enrichment of life-stage for the three asexual IDC forms (ring, trophozoite, and schizont) and the sexual development form (gametocyte) is shown. For each plot, a vertical black line represents a gene specific to the given life-stage. The bottom horizontal line which transitions from red to grey to blue represents the t-statistic for the given contrast. Enrichment of genes on the left side of the GSEA plot indicates coherent up-regulation in the group on the left of the heading label. Enrichment of genes on the right side of the GSEA plot indicates coherent up-regulation in the group on the right of the heading label (e.g. gametocyte genes are up-regulated in TP3 compared to TP2; first column, last row).

populations from animals that did not receive the AMD have much higher relative abundances of gametocytes, whereas in the parasite profile from the animal that did receive the AMD is much closer to the TP2 profiles (D); the first PC explains 84% of the variance in the gametocyte genes.

Interestingly, for the schizont-specific genes, the parasite profiles from TP3 are all qualitatively similar to those of TP2 in the first principal component direction (x-axis); this PC explains much of the variance and has been used to describe the general trajectory of the genes in a group. However, the parasite profiles from the two animals that did not receive the AMD have a qualitatively different location in the second principal component (y-axis). This result suggests that there is a shift in the underlying schizont gene expression program as the infection progresses, and further, that this shift in schizont-specific genes is ablated by AMD treatment. I

note that these differences are only suggestive of a trend and that I did not perform statistical analysis on the differences between the three TP3 samples due to the lack of replicates.

The PC loading plots of the ring- and trophozoite-specific gene sets show a general lack tight co-regulation of expression of these genes (Figure 29E,F). In contrast, both the schizont- and gametocyte-specific gene sets demonstrate a relatively high level of expression coherence (Figure 29G,H).

5.4.7 Trajectory of selected multigene families

Plasmodium helical interspersed subtelomeric family (PHIST) proteins have recently been shown to contain a erythrocyte cytoskeleton binding domain and consequently may play an important role in the rigidity of the infected red blood cell (iRBC), export of antigens, and formation of cytoadherence knobs, all of which play a role in the virulence of the infection.[90]

Tryptophan-rich antigens (TRAs), another large multigene family present in many *Plasmodium* species, are exported to the parasite cell membrane and may play a role in erythrocyte binding and invasion, an important step in the parasite blood stage progression.[95, 153, 21]

Both of the aforementioned protein families, PHIST and TRA, have a high expression at TP2, with lower expression level at later time points, TP3 and secondary (Figure 30A,B). Furthermore, nearly all members of these two families are tightly co-regulated (Figure 30C,D).

I also examined the expression trajectory of *P. cynomolgi* interspersed repeats (CYIR) genes (the largest multigene family in this parasite species), but there was no significant trend across experimental groups, nor was there expression coherence in the principal component loadings, which shows that transcripts from these genes are not co-regulated (not shown).

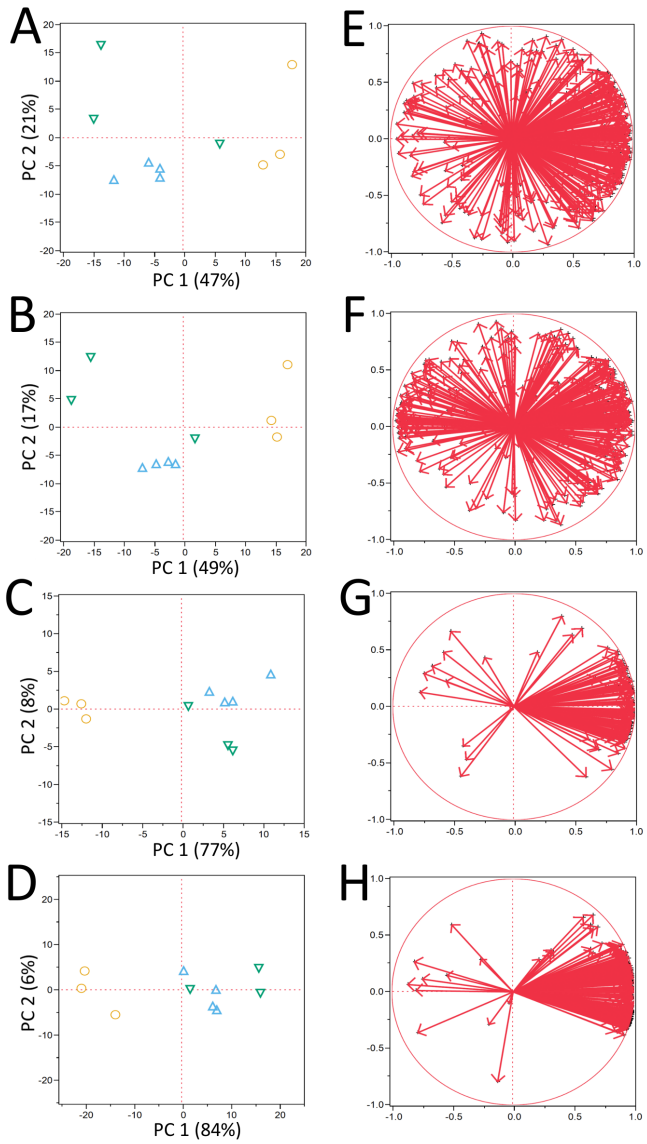

Figure 29: Life-stage-specific gene expression. (A-D) Plots of the first two PCs for the life-stage-specific gene sets. In all four plots, the two parasite profiles from TP3 (green inverted triangles) for the macaques that did not receive a sub-curative dose of artemether are closer together than to the parasite profile from the macaque that did receive the sub-curative artemether treatment. Samples from TP2, blue triangles; samples from relapsing time points, gold circles. (E-H) Principal component loading plots.

5.5 Discussion

In this study, I reported on parasite transcriptome-wide expression profiling across multiple infection peaks, and determined that *P. cynomolgi* has a qualitatively different gene expression profile between relapsing and primary parasitemias. Furthermore,

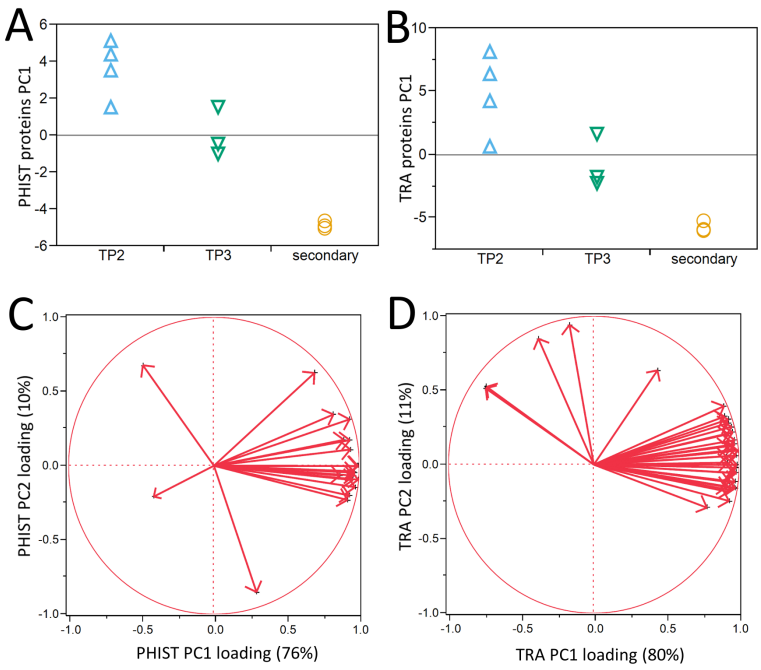

Figure 30: Trajectory of PHIST proteins and tryptophan-rich antigens (TRA) across the 100-day experiment. Composite measure of (A) PHIST protein and (B) TRA at the three clustered groups. Principal component loading of the (C) PHIST and (D) TRA genes show tight coherence of expression with the first PC explaining 76% and 80%, respectively.

the expression differences were life-stage coherent and indicated that the parasite populations of the relapsing time points showed much lower relative abundance of the sexual stage gametocytes.

5.5.1 Sexual stage abundance

A rather striking finding is that gametocyte-specific genes are lowly expressed in the relapsing parasitemias compared to both the peak of parasitemia (TP2) and the later stage of the first parasitemia peak (TP3) of the primary blood stage. This implies that compared to the primary infection, relapsing parasitemias have a much lower relative abundance of gametocytes that circulate in the blood. As suggested elsewhere, transition from asexual to sexual forms may be due to either host or parasite factors such as relative reticulocyte abundance or quorum sensing by microvesicle

or exosome, respectively.[149, 150, 120] But the exact mechanisms controlling the shift from asexual to sexual development are not completely understood.[3] The results from this study are inconclusive since reticulocyte data is not yet available, and microvesicles were not interrogated. Gametocyte formation was qualitatively associated with parasite abundance in the blood-stage, but a greater understanding of gametocyte differentiation requires many more samples.

High gametocytemia in the primary infection and lower gametocyte production in subsequent relapses may be adaptive and increase the fitness of the parasite in the following way. When a naive host is infected with *P. cynomolgi*, there is a strong host immune response. The parasites that most quickly shift from asexual to sexual development and produce viable gametocytes to be taken up by a mosquito will leave the most offspring.

Likewise, in an animal that survived the primary parasitemia, it may be advantageous to the parasite to replicate asexually for a longer period of time. Regardless of the evolutionary reason(s) for this behavior, the mechanism of control is likely to be genetic and regulated in response to the host. As more data is gathered, analyzed, and integrated, a deeper understanding of the mechanisms involved in gametocyte commitment should begin to emerge.

5.5.2 Parasites from relapse spend less time in the schizont stage

In addition to the decreased number of gametocytes produces in the relapsing time points, schizont-specific gene expression is also decreased. TP2 has the highest schizont-specific expression, suggesting that parasites from TP2 –and to a lesser extent TP3– spend more time in the schizont stage, which is when the parasite undergoes asexual division to increase its numbers, compared to relapsing parasite populations.

The first peak of parasitemia is characterized by much higher parasite densities which co-occurs with the longer duration spent in the schizont stage of the IDC. This

association, while tentative based on the small sample size, should be further explored, since controlling parasitemia levels and the resulting severe malarial symptoms is an important goal in the treatment of this disease.

5.5.3 Previously observed effects of AMDs on parasite

Although there is only one sample for which I have parasite expression soon after artemether administration, it appears that artemether reduced the abundance of gametocytes in the blood stage, a finding consistent with previous reports.[48, 107, 50] Interestingly, though, the relative level of gametocyte abundance after AMD treatment is only comparable to that of TP2; in contrast, the relapsing parasite populations have a much lower gametocyte abundances. This finding re-affirms that there are parasite regulatory networks that limit the production of gametocytes. Further investigation into these networks and the genes important in gametocyte development would facilitate drug development to target gametocyte activation pathways, thereby limiting sexual-stage parasites and decreasing malaria transmission.

5.5.4 Caveats of this study

In this study, there were only 10 samples that contained a sufficient number of parasite reads to accurately quantify expression. This low number of samples decreased statistical power to determine differential expression. In spite of this challenge, many genes in coherent gene sets were identified as differentially regulated between the experimental groups. Further, while some libraries had much fewer reads than others, my down-sampling methodology demonstrated that the qualitative differences between the experimental groups were robust to read sequencing depth.

In spite of the high depth of coverage of RNA-seq and the repeated measures taken from the same animals across numerous time points, the causal direction cannot be assigned between host and parasite transcriptome dynamics. That is, while I have found qualitative differences in the parasite transcriptome between primary and

relapse parasitemias, I cannot make the claim that the qualitatively different host clinical parameters (anemia, thrombocytopenia, fever, lethargy, dehydration, etc.) are being driven by the parasite transcriptome changes. Likewise, I cannot say that the host immune response is causing the differences observed in the parasite transcriptome. However, I have identified parasite transcriptional differences that are associated with the severity of the malaria infection. These results should shape the design of future experiments.

5.5.5 Future studies

Using this rich *in vivo Plasmodium* expression dataset, many additional analyses could be performed. Firstly, I suggest that the shift in abundance of expressed alleles should be tracked. That is, over the course of the experiment and in direct response to the artificial selection imposed by the administration of anti-malarial drugs, advantageous mutations may arise in the parasite. These mutations would then be selected for if they offered a fitness advantage to the parasite. While this type of analysis has never been performed previously and could have very important implications in our understanding of the rate of evolution and selection occurring in the parasite genome during an *in vivo* infection within a host, it may not be the best experiment to explore this phenomenon.

In this *P. cynomolgi* infection, a completely curative dose of artemether was used to eliminate the blood stage parasite. Subsequent relapsing parasitemias come from the reactivation of dormant hypnozoites from the liver. These parasites are not affected by the artemether, and so would not have experienced selection against the oxidative stress induced primarily in the red blood cell. I would strongly advise, however, that this type of analysis be performed in subsequent experiments, particularly in the *P. coatneyi* infection, since in this experiment only subcurative doses of

artemether are given. I would expect to see advantageous mutations arise which increase the expression of proteins that can mitigate the increases in oxidative damage to proteins and DNA.

Due to the nature of RNA-seq, the abundance of different isoforms for parasite genes could also be explored. However, the number of *Plasmodium* genes shown to have alternative isoforms is rather low (5% to 10%) compared to, say, *Homo sapiens* (nearly all expressed genes).[140]

CHAPTER VI

CONCLUDING REMARKS

In CHAPTER II, I outlined the analytical methods and integrative techniques which were used in subsequent chapters. I showed also that the anti-malarial drug pyrimethamine dysregulates numerous pathways in both the blood and bone marrow and that many of the effects persist for at least 30 days after the administration of the drug. For the treatment of malaria, which often requires only intermittent administration, these host effects may not be of concern. However, in light of the possibility of re-purposing this drug for the treatment of ALS, which would likely require a more regular usage, further investigations into the safety of this drug need to be performed.

In CHAPTER III, I reported on the host transcriptional response to malaria infection, and detail the qualitatively different host response between primary and relapsing parasitemia peaks. The marked difference in magnitude of host transcriptional changes between the primary and relapsing parasite peaks was unsurprising given the vast difference in clinical parameters and outcomes between the two peak types. However, there were qualitatively different pathways enriched in the relapsing parasitemias. With the sparse sampling of this study, I cannot comment on the day-to-day changes that may have occurred in either the primary or relapsing parasitemias. Follow-up studies with greater sampling rates are planned for later this year and should yield greater insight into host immune response. Furthermore, transcriptional profiling was performed on whole blood. In subsequent experiments, immune cell subsets will be sorted and profiled separately, a study design change which should give greater transcriptional resolution and ability to identify differential expression in important cell types.

In CHAPTER IV, I implemented a novel expression deconvolution method to determine the life-stage composition of samples from a previously-reported *P. falciparum* IDC. Using this method, I was able to identify genes whose expression is parasite stage-specific and used those gene sets in two subsequent chapters which concerned parasite transcriptional profiling. In studies of gene expression profiling, especially in *Plasmodium* where most genes have stage-specific expression, it will be useful to employ expression deconvolution to estimate the abundance of cell types, and then to subsequently use those estimates as both variables as well as covariate in differential gene expression analysis.

In CHAPTER V, I identified qualitative alterations in the parasite transcriptome between primary and relapsing infection peaks across a 100-day infection cycle. I found that gametocytes were more abundant in parasite populations from untreated animals compared to parasites from an animal treated with the anti-malarial drug artemether. Due to a small sample size, it would be inappropriate to conclude that the AMD caused the shift away from the production of the sexual stage of this parasite. However, it does merit further investigation.

REFERENCES

[1] Aggarwal, A., Khurana, P., Mitra, S., Raicha, B., Saraswathy, K. N., Italia, Y. M., and Kshatriya, G. K., "Distribution of beta-globin haplotypes among the tribes of southern Gujarat, India.," *Gene*, vol. 521, pp. 287–292, Jun 2013.

[2] Agnandji, S. T., Lell, B., Soulanoudjingar, S. S., Fernandes, J. F., Abossolo, B. P., Conzelmann, C., Methogo, B. G. N. O., Doucka, Y., Flamen, A., Mordmller, B., Issifou, S., Kremsner, P. G., Sacarlal, J., Aide, P., Lanaspa, M., Aponte, J. J., Nhamuave, A., Quelhas, D., Bassat, Q., Mandjate, S., Macete, E., Alonso, P., Abdulla, S., Salim, N., Juma, O., Shomari, M., Shubis, K., Machera, F., Hamad, A. S., Minja, R., Mtoro, A., Sykes, A., Ahmed, S., Urassa, A. M., Ali, A. M., Mwangoka, G., Tanner, M., Tinto, H., D'Alessandro, U., Sorgho, H., Valea, I., Tahita, M. C., Kabor, W., Oudraogo, S., Sandrine, Y., Guiguemd, R. T., Oudraogo, J. B., Hamel, M. J., Kariuki, S., Odero, C., Oneko, M., Otieno, K., Awino, N., Omoto, J., Williamson, J., Muturi-Kioi, V., Laserson, K. F., Slutsker, L., Otieno, W., Otieno, L., Nekoye, O., Gondi, S., Otieno, A., Ogutu, B., Wasuna, R., Owira, V., Jones, D., Onyango, A. A., Njuguna, P., Chilengi, R., Akoo, P., Kerubo, C., Gitaka, J., Maingi, C., Lang, T., Olotu, A., Tsofa, B., Bejon, P., Peshu, N., Marsh, K., Owusu-Agyei, S., Asante, K. P., Osei-Kwakye, K., Boahen, O., Ayamba, S., Kayan, K., Owusu-Ofori, R., Dosoo, D., Asante, I., Adjei, G., Adjei, G., Chandramohan, D., Greenwood, B., Lusingu, J., Gesase, S., Malabeja, A., Abdul, O., Kilavo, H., Mahende, C., Liheluka, E., Lemnge, M., Theander, T., Drakeley, C., Ansong, D., Agbenyega, T., Adjei, S., Boateng, H. O., Rettig, T., Bawa, J., Sylverken, J., Sambian, D., Agyekum, A., Owusu, L., Martinson, F., Hoffman, I., Mvalo, T., Kamthunzi, P., Nkomo, R., Msika, A., Jumbe, A., Chome, N., Nyakuipa, D., Chintedza, J., Ballou, W. R., Bruls, M., Cohen, J., Guerra, Y., Jongert, E., Lapierre, D., Leach, A., Lievens, M., Ofori-Anyinam, O., Vekemans, J., Carter, T., Leboulleux, D., Loucq, C., Radford, A., Savarese, B., Schellenberg, D., Sillman, M., Vansadia, P., and R. T. S, S. C. T. P., "First results of phase 3 trial of RTS,S/AS01 malaria vaccine in African children.," *N Engl J Med*, vol. 365, pp. 1863–1875, Nov 2011.

[3] Alano, P., "*Plasmodium falciparum* gametocytes: still many secrets of a hidden life.," *Mol Microbiol*, vol. 66, pp. 291–302, Oct 2007.

[4] Allison, A. C., "Genetic control of resistance to human malaria.," *Curr Opin Immunol*, vol. 21, pp. 499–505, Oct 2009.

[5] ALVING, A. S., CARSON, P. E., FLANAGAN, C. L., and ICKES, C. E., "Enzymatic deficiency in primaquine-sensitive erythrocytes.," *Science*, vol. 124, pp. 484–485, Sep 1956.

[6] Anders, S. and Huber, W., "Differential expression analysis for sequence count data.," *Genome Biol*, vol. 11, no. 10, p. R106, 2010.

[7] Anders, S., McCarthy, D. J., Chen, Y., Okoniewski, M., Smyth, G. K., Huber, W., and Robinson, M. D., "Count-based differential expression analysis of RNA sequencing data using R and Bioconductor.," *Nat Protoc*, vol. 8, pp. 1765–1786, Sep 2013.

[8] Andrews, K. T., Gupta, A. P., Tran, T. N., Fairlie, D. P., Gobert, G. N., and Bozdech, Z., "Comparative gene expression profiling of *P. falciparum* malaria parasites exposed to three different histone deacetylase inhibitors.," *PLoS One*, vol. 7, no. 2, p. e31847, 2012.

[9] Appling, D. R., "Compartmentation of folate-mediated one-carbon metabolism in eukaryotes.," *FASEB J*, vol. 5, pp. 2645–2651, Sep 1991.

[10] Aurrecoechea, C., Brestelli, J., Brunk, B. P., Dommer, J., Fischer, S., Gajria, B., Gao, X., Gingle, A., Grant, G., Harb, O. S., Heiges, M., Innamorato, F., Iodice, J., Kissinger, J. C., Kraemer, E., Li, W., Miller, J. A., Nayak, V., Pennington, C., Pinney, D. F., Roos, D. S., Ross, C., Stoeckert, C. J., Treatman, C., and Wang, H., "PlasmoDB: a functional genomic database for malaria parasites.," *Nucleic Acids Res*, vol. 37, pp. D539–D543, Jan 2009.

[11] Bahl, A., Brunk, B., Crabtree, J., Fraunholz, M. J., Gajria, B., Grant, G. R., Ginsburg, H., Gupta, D., Kissinger, J. C., Labo, P., Li, L., Mailman, M. D., Milgram, A. J., Pearson, D. S., Roos, D. S., Schug, J., Stoeckert, C. J., and Whetzel, P., "PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data.," *Nucleic Acids Res*, vol. 31, pp. 212–215, Jan 2003.

[12] Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I., and Nelson, A., "Determining global population distribution: methods, applications and data.," *Adv Parasitol*, vol. 62, pp. 119–156, 2006.

[13] Barnwell, J. W., Howard, R. J., Coon, H. G., and Miller, L. H., "Splenic requirement for antigenic variation and expression of the variant antigen on the erythrocyte membrane in cloned *Plasmodium knowlesi* malaria.," *Infect Immun*, vol. 40, pp. 985–994, Jun 1983.

[14] Barnwell, J. W., Howard, R. J., and Miller, L. H., "Influence of the spleen on the expression of surface antigens on parasitized erythrocytes.," *Ciba Found Symp*, vol. 94, pp. 117–136, 1983.

[15] Barnwell, J. W. and Galinski, M. R., "Malarial liver parasites awaken in culture.," *Nat Med*, vol. 20, pp. 237–239, Mar 2014.

[16] Basco, L. K. and Bras, J. L., "In vitro sensitivity of *Plasmodium falciparum* to anti-folinic agents (trimethoprim, pyrimethamine, cycloguanil): a study of 29 African strains.," *Bull Soc Pathol Exot*, vol. 90, no. 2, pp. 90–93, 1997.

[17] Bassat, Q., "The use of artemether-lumefantrine for the treatment of uncomplicated *Plasmodium vivax* malaria.," *PLoS Negl Trop Dis*, vol. 5, p. e1325, Dec 2011.

[18] Battle, K. E., Gething, P. W., Elyazar, I. R. F., Moyes, C. L., Sinka, M. E., Howes, R. E., Guerra, C. A., Price, R. N., Baird, K. J., and Hay, S. I., "The global public health significance of *Plasmodium vivax.*," *Adv Parasitol*, vol. 80, pp. 1–111, 2012.

[19] Battle, K. E., Karhunen, M. S., Bhatt, S., Gething, P. W., Howes, R. E., Golding, N., Boeckel, T. P. V., Messina, J. P., Shanks, G. D., Smith, D. L., Baird, J. K., and Hay, S. I., "Geographical variation in *Plasmodium vivax* relapse.," *Malar J*, vol. 13, p. 144, 2014.

[20] Boeuf, P., Aitken, E. H., Chandrasiri, U., Chua, C. L. L., McInerney, B., McQuade, L., Duffy, M., Molyneux, M., Brown, G., Glazier, J., and Rogerson, S. J., "*Plasmodium falciparum* malaria elicits inflammatory responses that dysregulate placental amino acid transport.," *PLoS Pathog*, vol. 9, p. e1003153, Feb 2013.

[21] Bora, H., Tyagi, R. K., and Sharma, Y. D., "Defining the erythrocyte binding domains of *Plasmodium vivax* tryptophan rich antigen 33.5.," *PLoS One*, vol. 8, no. 4, p. e62829, 2013.

[22] Borden, M. B. and Parke, A. L., "Antimalarial drugs in systemic lupus erythematosus: use in pregnancy.," *Drug Saf*, vol. 24, no. 14, pp. 1055–1063, 2001.

[23] Bozdech, Z., Llins, M., Pulliam, B. L., Wong, E. D., Zhu, J., and DeRisi, J. L., "The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum.*," *PLoS Biol*, vol. 1, p. E5, Oct 2003.

[24] Bozdech, Z., Mok, S., Hu, G., Imwong, M., Jaidee, A., Russell, B., Ginsburg, H., Nosten, F., Day, N. P. J., White, N. J., Carlton, J. M., and Preiser, P. R., "The transcriptome of *Plasmodium vivax* reveals divergence and diversity of transcriptional regulation in malaria parasites.," *Proc Natl Acad Sci U S A*, vol. 105, pp. 16290–16295, Oct 2008.

[25] Bozdech, Z., Zhu, J., Joachimiak, M. P., Cohen, F. E., Pulliam, B., and DeRisi, J. L., "Expression profiling of the schizont and trophozoite stages of *Plasmodium falciparum* with a long-oligonucleotide microarray.," *Genome Biol*, vol. 4, no. 2, p. R9, 2003.

[26] Brunet, J.-P., Tamayo, P., Golub, T. R., and Mesirov, J. P., "Metagenes and molecular pattern discovery using matrix factorization.," *Proc Natl Acad Sci U S A*, vol. 101, pp. 4164–4169, Mar 2004.

[27] Cagliani, R. and Sironi, M., "Pathogen-driven selection in the human genome.," *Int J Evol Biol*, vol. 2013, p. 204240, 2013.

[28] Carlton, J. M., Das, A., and Escalante, A. A., "Genomics, population genetics and evolutionary history of plasmodium vivax.," *Adv Parasitol*, vol. 81, pp. 203–222, 2013.

[29] Chen, R., Mias, G. I., Li-Pook-Than, J., Jiang, L., Lam, H. Y. K., Chen, R., Miriami, E., Karczewski, K. J., Hariharan, M., Dewey, F. E., Cheng, Y., Clark, M. J., Im, H., Habegger, L., Balasubramanian, S., O'Huallachain, M., Dudley, J. T., Hillenmeyer, S., Haraksingh, R., Sharon, D., Euskirchen, G., Lacroute, P., Bettinger, K., Boyle, A. P., Kasowski, M., Grubert, F., Seki, S., Garcia, M., Whirl-Carrillo, M., Gallardo, M., Blasco, M. A., Greenberg, P. L., Snyder, P., Klein, T. E., Altman, R. B., Butte, A. J., Ashley, E. A., Gerstein, M., Nadeau, K. C., Tang, H., and Snyder, M., "Personal omics profiling reveals dynamic molecular and medical phenotypes.," *Cell*, vol. 148, pp. 1293–1307, Mar 2012.

[30] Chenet, S. M., Tapia, L. L., Escalante, A. A., Durand, S., Lucas, C., and Bacon, D. J., "Genetic diversity and population structure of genes encoding vaccine candidate antigens of plasmodium vivax.," *Malar J*, vol. 11, p. 68, 2012.

[31] Chima, R. I., Goodman, C. A., and Mills, A., "The economic impact of malaria in Africa: a critical review of the evidence.," *Health Policy*, vol. 63, pp. 17–36, Jan 2003.

[32] Coluzzi, M., "Malaria vector analysis and control.," *Parasitol Today*, vol. 8, pp. 113–118, Apr 1992.

[33] Corcoran, K. D., Hansukjariya, P., Sattabongkot, J., Ngampochjana, M., Edstein, M. D., Smith, C. D., Shanks, G. D., and Milhous, W. K., "Causal prophylactic and radical curative activity of WR182393 (a guanylhydrazone) against *Plasmodium cynomolgi* in *Macaca mulatta*.," *Am J Trop Med Hyg*, vol. 49, pp. 473–477, Oct 1993.

[34] Cox-Singh, J., Davis, T. M. E., Lee, K.-S., Shamsul, S. S. G., Matusop, A., Ratnam, S., Rahman, H. A., Conway, D. J., and Singh, B.,

"*Plasmodium knowlesi* malaria in humans is widely distributed and potentially life threatening.," *Clin Infect Dis*, vol. 46, pp. 165–171, Jan 2008.

[35] Cui, L., Escalante, A. A., Imwong, M., and Snounou, G., "The genetic diversity of plasmodium vivax populations.," *Trends Parasitol*, vol. 19, pp. 220–226, May 2003.

[36] Daily, J. P., Scanfeld, D., Pochet, N., Roch, K. L., Plouffe, D., Kamal, M., Sarr, O., Mboup, S., Ndir, O., Wypij, D., Levasseur, K., Thomas, E., Tamayo, P., Dong, C., Zhou, Y., Lander, E. S., Ndiaye, D., Wirth, D., Winzeler, E. A., Mesirov, J. P., and Regev, A., "Distinct physiological states of *Plasmodium falciparum* in malaria-infected patients.," *Nature*, vol. 450, pp. 1091–1095, Dec 2007.

[37] Daily, J. P., Roch, K. G. L., Sarr, O., Ndiaye, D., Lukens, A., Zhou, Y., Ndir, O., Mboup, S., Sultan, A., Winzeler, E. A., and Wirth, D. F., "*In vivo* transcriptome of *Plasmodium falciparum* reveals overexpression of transcripts that encode surface proteins.," *J Infect Dis*, vol. 191, pp. 1196–1203, Apr 2005.

[38] Desai, K. H., Tan, C. S., Leek, J. T., Maier, R. V., Tompkins, R. G., Storey, J. D., Inflammation, and the Host Response to Injury Large-Scale Collaborative Research Program, "Dissecting inflammatory complications in critically injured patients by within-patient gene expression changes: a longitudinal clinical genomics study.," *PLoS Med*, vol. 8, p. e1001093, Sep 2011.

[39] Ding, K., de Andrade, M., Manolio, T. A., Crawford, D. C., Rasmussen-Torvik, L. J., Ritchie, M. D., Denny, J. C., Masys, D. R., Jouni, H., Pachecho, J. A., Kho, A. N., Roden, D. M., Chisholm, R., and Kullo, I. J., "Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study.," *G3 (Bethesda)*, vol. 3, pp. 1061–1068, Jul 2013.

[40] Ettling, M., McFarland, D. A., Schultz, L. J., and Chitsulo, L., "Economic impact of malaria in Malawian households.," *Trop Med Parasitol*, vol. 45, pp. 74–79, Mar 1994.

[41] for Disease Control, C. and (CDC), P., "Simian malaria in a U.S. traveler–New York, 2008.," *MMWR Morb Mortal Wkly Rep*, vol. 58, pp. 229–232, Mar 2009.

[42] for Disease Control, C. and Prevention, "Malaria - about malaria." Electronic, July 2014.

[43] Foth, B. J., Zhang, N., Mok, S., Preiser, P. R., and Bozdech, Z., "Quantitative protein expression profiling reveals extensive post-transcriptional

regulation and post-translational modifications in schizont-stage malaria parasites.," *Genome Biol*, vol. 9, no. 12, p. R177, 2008.

[44] Foth, B. J., Zhang, N., Chaal, B. K., Sze, S. K., Preiser, P. R., and Bozdech, Z., "Quantitative time-course profiling of parasite and host cell proteins in the human malaria parasite *Plasmodium falciparum*.," *Mol Cell Proteomics*, vol. 10, p. M110.006411, Aug 2011.

[45] Galinski, M. R. and Barnwell, J. W., "*Plasmodium vivax*: who cares?," *Malar J*, vol. 7 Suppl 1, p. S9, 2008.

[46] Ganesan, K., Ponmee, N., Jiang, L., Fowble, J. W., White, J., Kamchonwongpaisan, S., Yuthavong, Y., Wilairat, P., and Rathod, P. K., "A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates.," *PLoS Pathog*, vol. 4, p. e1000214, Nov 2008.

[47] Gardner, M. J., Hall, N., Fung, E., White, O., Berriman, M., Hyman, R. W., Carlton, J. M., Pain, A., Nelson, K. E., Bowman, S., Paulsen, I. T., James, K., Eisen, J. A., Rutherford, K., Salzberg, S. L., Craig, A., Kyes, S., Chan, M.-S., Nene, V., Shallom, S. J., Suh, B., Peterson, J., Angiuoli, S., Pertea, M., Allen, J., Selengut, J., Haft, D., Mather, M. W., Vaidya, A. B., Martin, D. M. A., Fairlamb, A. H., Fraunholz, M. J., Roos, D. S., Ralph, S. A., McFadden, G. I., Cummings, L. M., Subramanian, G. M., Mungall, C., Venter, J. C., Carucci, D. J., Hoffman, S. L., Newbold, C., Davis, R. W., Fraser, C. M., and Barrell, B., "Genome sequence of the human malaria parasite plasmodium falciparum.," *Nature*, vol. 419, pp. 498–511, Oct 2002.

[48] Garner, P. and Graves, P. M., "The benefits of artemisinin combination therapy for malaria extend beyond the individual patient.," *PLoS Med*, vol. 2, p. e105, Apr 2005.

[49] Gaujoux, R. and Seoighe, C., "CellMix: a comprehensive toolbox for gene expression deconvolution.," *Bioinformatics*, vol. 29, pp. 2211–2212, Sep 2013.

[50] Gbotosho, G. O., Sowunmi, A., Happi, C. T., and Okuboyejo, T. M., "*Plasmodium falciparum* gametocyte carriage, sex ratios and asexual parasite rates in Nigerian children before and after a treatment protocol policy change instituting the use of artemisinin-based combination therapies.," *Mem Inst Oswaldo Cruz*, vol. 106, pp. 685–690, Sep 2011.

[51] Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., Hornik, K., Hothorn, T., Huber, W., Iacus, S., Irizarry, R., Leisch, F., Li, C., Maechler, M., Rossini, A. J., Sawitzki, G., Smith, C.,

Smyth, G., Tierney, L., Yang, J. Y. H., and Zhang, J., "Bioconductor: open software development for computational biology and bioinformatics.," *Genome Biol*, vol. 5, no. 10, p. R80, 2004.

[52] Gething, P. W., Elyazar, I. R. F., Moyes, C. L., Smith, D. L., Battle, K. E., Guerra, C. A., Patil, A. P., Tatem, A. J., Howes, R. E., Myers, M. F., George, D. B., Horby, P., Wertheim, H. F. L., Price, R. N., Meller, I., Baird, J. K., and Hay, S. I., "A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010.," *PLoS Negl Trop Dis*, vol. 6, no. 9, p. e1814, 2012.

[53] Gething, P. W., Patil, A. P., Smith, D. L., Guerra, C. A., Elyazar, I. R. F., Johnston, G. L., Tatem, A. J., and Hay, S. I., "A new world malaria map: *Plasmodium falciparum* endemicity in 2010.," *Malar J*, vol. 10, p. 378, 2011.

[54] Gong, T., Hartmann, N., Kohane, I. S., Brinkmann, V., Staedtler, F., Letzkus, M., Bongiovanni, S., and Szustakowski, J. D., "Optimal deconvolution of transcriptional profiling data using quadratic programming with application to complex clinical blood samples.," *PLoS One*, vol. 6, no. 11, p. e27156, 2011.

[55] Gong, T. and Szustakowski, J. D., "DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data.," *Bioinformatics*, vol. 29, pp. 1083–1085, Apr 2013.

[56] Griffin, J. T., Ferguson, N. M., and Ghani, A. C., "Estimates of the changing age-burden of *Plasmodium falciparum* malaria disease in sub-Saharan Africa.," *Nat Commun*, vol. 5, p. 3136, 2014.

[57] Griffiths, M. J., Shafi, M. J., Popper, S. J., Hemingway, C. A., Kortok, M. M., Wathen, A., Rockett, K. A., Mott, R., Levin, M., Newton, C. R., Marsh, K., Relman, D. A., and Kwiatkowski, D. P., "Genomewide analysis of the host response to malaria in Kenyan children.," *J Infect Dis*, vol. 191, pp. 1599–1611, May 2005.

[58] Hall, N., Karras, M., Raine, J. D., Carlton, J. M., Kooij, T. W. A., Berriman, M., Florens, L., Janssen, C. S., Pain, A., Christophides, G. K., James, K., Rutherford, K., Harris, B., Harris, D., Churcher, C., Quail, M. A., Ormond, D., Doggett, J., Trueman, H. E., Mendoza, J., Bidwell, S. L., Rajandream, M.-A., Carucci, D. J., Yates, J. R., Kafatos, F. C., Janse, C. J., Barrell, B., Turner, C. M. R., Waters, A. P., and Sinden, R. E., "A comprehensive survey of the *Plasmodium* life cycle by genomic, transcriptomic, and proteomic analyses.," *Science*, vol. 307, pp. 82–86, Jan 2005.

[59] Hayashi, F., Means, T. K., and Luster, A. D., "Toll-like receptors stimulate human neutrophil function.," *Blood*, vol. 102, pp. 2660–2669, Oct 2003.

[60] Hedrick, P. W., "Population genetics of malaria resistance in humans.," *Heredity (Edinb)*, vol. 107, pp. 283–304, Oct 2011.

[61] Hedrick, P. W., "Resistance to malaria in humans: the impact of strong, recent selection.," *Malar J*, vol. 11, p. 349, 2012.

[62] Hill, A. V., Allsopp, C. E., Kwiatkowski, D., Anstey, N. M., Twumasi, P., Rowe, P. A., Bennett, S., Brewster, D., McMichael, A. J., and Greenwood, B. M., "Common west African HLA antigens are associated with protection from severe malaria.," *Nature*, vol. 352, pp. 595–600, Aug 1991.

[63] Howard, R. J. and Barnwell, J. W., "Roles of surface antigens on malaria-infected red blood cells in evasion of immunity.," *Contemp Top Immunobiol*, vol. 12, pp. 127–200, 1984.

[64] Howard, R. J., Barnwell, J. W., and Kao, V., "Antigenic variation of *Plasmodium knowlesi* malaria: identification of the variant antigen on infected erythrocytes.," *Proc Natl Acad Sci U S A*, vol. 80, pp. 4129–4133, Jul 1983.

[65] Howes, R. E., Piel, F. B., Patil, A. P., Nyangiri, O. A., Gething, P. W., Dewi, M., Hogg, M. M., Battle, K. E., Padilla, C. D., Baird, J. K., and Hay, S. I., "G6PD deficiency prevalence and estimates of affected populations in malaria endemic countries: a geostatistical model-based map.," *PLoS Med*, vol. 9, no. 11, p. e1001339, 2012.

[66] Hulden, L. and Hulden, L., "Activation of the hypnozoite: a part of plasmodium vivax life cycle and survival.," *Malar J*, vol. 10, p. 90, 2011.

[67] Idaghdour, Y., Quinlan, J., Goulet, J.-P., Berghout, J., Gbeha, E., Bruat, V., de Malliard, T., Grenier, J.-C., Gomez, S., Gros, P., Rahimy, M. C., Sanni, A., and Awadalla, P., "Evidence for additive and interaction effects of host genotype and infection in malaria.," *Proc Natl Acad Sci U S A*, vol. 109, pp. 16786–16793, Oct 2012.

[68] Idro, R., Kakooza-Mwesige, A., Balyejjussa, S., Mirembe, G., Mugasha, C., Tugumisirize, J., and Byarugaba, J., "Severe neurological sequelae and behaviour problems after cerebral malaria in Ugandan children.," *BMC Res Notes*, vol. 3, p. 104, 2010.

[69] Idro, R., Marsh, K., John, C. C., and Newton, C. R. J., "Cerebral malaria: mechanisms of brain injury and strategies for improved neurocognitive outcome.," *Pediatr Res*, vol. 68, pp. 267–274, Oct 2010.

[70] Joice, R., Narasimhan, V., Montgomery, J., Sidhu, A. B., Oh, K., Meyer, E., Pierre-Louis, W., Seydel, K., Milner, D., Williamson,

K., Wiegand, R., Ndiaye, D., Daily, J., Wirth, D., Taylor, T., Huttenhower, C., and Marti, M., "Inferring developmental stage composition from gene expression in human malaria.," *PLoS Comput Biol*, vol. 9, p. e1003392, Dec 2013.

[71] Jung, H., Bobba, R., Su, J., Shariati-Sarabi, Z., Gladman, D. D., Urowitz, M., Lou, W., and Fortin, P. R., "The protective effect of antimalarial drugs on thrombovascular events in systemic lupus erythematosus.," *Arthritis Rheum*, vol. 62, pp. 863–868, Mar 2010.

[72] Kafsack, B. F. C., Rovira-Graells, N., Clark, T. G., Bancells, C., Crowley, V. M., Campino, S. G., Williams, A. E., Drought, L. G., Kwiatkowski, D. P., Baker, D. A., Corts, A., and Llins, M., "A transcriptional switch underlies commitment to sexual development in malaria parasites.," *Nature*, vol. 507, pp. 248–252, Mar 2014.

[73] KAUFMAN, H. E. and GEISLER, P. H., "The hematologic toxicity of pyrimethamine (Daraprim) in man.," *Arch Ophthalmol*, vol. 64, pp. 140–146, Jul 1960.

[74] Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L., "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.," *Genome Biol*, vol. 14, p. R36, Apr 2013.

[75] Kitron, U. and Spielman, A., "Suppression of transmission of malaria through source reduction: antianopheline measures applied in Israel, the United States, and Italy.," *Rev Infect Dis*, vol. 11, no. 3, pp. 391–406, 1989.

[76] Koboldt, D. C., Chen, K., Wylie, T., Larson, D. E., McLellan, M. D., Mardis, E. R., Weinstock, G. M., Wilson, R. K., and Ding, L., "VarScan: variant detection in massively parallel sequencing of individual and pooled samples.," *Bioinformatics*, vol. 25, pp. 2283–2285, Sep 2009.

[77] Koehler, J. W., Bolton, M., Rollins, A., Snook, K., deHaro, E., Henson, E., Rogers, L., Martin, L. N., Krogstad, D. J., James, M. A., Rice, J., Davison, B., Veazey, R. S., Prabhu, R., Amedee, A. M., Garry, R. F., and Cogswell, F. B., "Altered immune responses in rhesus macaques co-infected with SIV and *Plasmodium cynomolgi*: an animal model for coincident AIDS and relapsing malaria.," *PLoS One*, vol. 4, no. 9, p. e7139, 2009.

[78] Krupka, M., Seydel, K., Feintuch, C. M., Yee, K., Kim, R., Lin, C.-Y., Calder, R. B., Petersen, C., Taylor, T., and Daily, J., "Mild *Plasmodium falciparum* malaria following an episode of severe malaria is associated with induction of the interferon pathway in Malawian children.," *Infect Immun*, vol. 80, pp. 1150–1155, Mar 2012.

[79] Kwiatkowski, D. P., "How malaria has affected the human genome and what human genetics can teach us about malaria.," *Am J Hum Genet*, vol. 77, pp. 171–192, Aug 2005.

[80] Lange, D. J., Andersen, P. M., Remanan, R., Marklund, S., and Benjamin, D., "Pyrimethamine decreases levels of sod1 in leukocytes and cerebrospinal fluid of als patients: a phase I pilot study.," *Amyotroph Lateral Scler Frontotemporal Degener*, vol. 14, pp. 199–204, Apr 2013.

[81] Lapp, S. A., Korir-Morrison, C., Jiang, J., Bai, Y., Corredor, V., and Galinski, M. R., "Spleen-dependent regulation of antigenic variation in malaria parasites: *Plasmodium knowlesi SICAvar* expression profiles in splenic and asplenic hosts.," *PLoS One*, vol. 8, no. 10, p. e78014, 2013.

[82] Lee, C. M., Mudaliar, M. A. V., Haggart, D. R., Wolf, C. R., Miele, G., Vass, J. K., Higham, D. J., and Crowther, D., "Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology.," *PLoS One*, vol. 7, no. 12, p. e48238, 2012.

[83] Lemieux, J. E., Gomez-Escobar, N., Feller, A., Carret, C., Amambua-Ngwa, A., Pinches, R., Day, F., Kyes, S. A., Conway, D. J., Holmes, C. C., and Newbold, C. I., "Statistical estimation of cell-cycle progression and lineage commitment in *Plasmodium falciparum* reveals a homogeneous pattern of transcription in *ex vivo* culture.," *Proc Natl Acad Sci U S A*, vol. 106, pp. 7559–7564, May 2009.

[84] Li, Y. and Xie, X., "A mixture model for expression deconvolution from RNA-seq in heterogeneous tissues.," *BMC Bioinformatics*, vol. 14 Suppl 5, p. S11, 2013.

[85] Lin, B. F., Huang, R. F., and Shane, B., "Regulation of folate and one-carbon metabolism in mammalian cells. III. Role of mitochondrial folylpoly-gamma-glutamate synthetase.," *J Biol Chem*, vol. 268, pp. 21674–21679, Oct 1993.

[86] Liu, W., Li, Y., Shaw, K. S., Learn, G. H., Plenderleith, L. J., Malenke, J. A., Sundararaman, S. A., Ramirez, M. A., Crystal, P. A., Smith, A. G., Bibollet-Ruche, F., Ayouba, A., Locatelli, S., Esteban, A., Mouacha, F., Guichet, E., Butel, C., Ahuka-Mundeke, S., Inogwabini, B.-I., Ndjango, J.-B. N., Speede, S., Sanz, C. M., Morgan, D. B., Gonder, M. K., Kranzusch, P. J., Walsh, P. D., Georgiev, A. V., Muller, M. N., Piel, A. K., Stewart, F. A., Wilson, M. L., Pusey, A. E., Cui, L., Wang, Z., Frnert, A., Sutherland, C. J., Nolder, D., Hart, J. A., Hart, T. B., Bertolani, P., Gillis, A., LeBreton, M., Tafon, B., Kiyang, J., Djoko, C. F., Schneider, B. S., Wolfe, N. D., Mpoudi-Ngole, E., Delaporte, E., Carter, R., Culleton, R. L., Shaw, G. M., Rayner, J. C., Peeters, M., Hahn,

B. H., and Sharp, P. M., "African origin of the malaria parasite plasmodium vivax.," *Nat Commun*, vol. 5, p. 3346, 2014.

[87] Lovegrove, F. E., Pea-Castillo, L., Mohammad, N., Liles, W. C., Hughes, T. R., and Kain, K. C., "Simultaneous host and parasite expression profiling identifies tissue-specific transcriptional programs associated with susceptibility or resistance to experimental cerebral malaria.," *BMC Genomics*, vol. 7, p. 295, 2006.

[88] Luchavez, J., Espino, F., Curameng, P., Espina, R., Bell, D., Chiodini, P., Nolder, D., Sutherland, C., Lee, K.-S., and Singh, B., "Human infections with *Plasmodium knowlesi*, the Philippines.," *Emerg Infect Dis*, vol. 14, pp. 811–813, May 2008.

[89] Lpez-Barragn, M. J., Lemieux, J., Quiones, M., Williamson, K. C., Molina-Cruz, A., Cui, K., Barillas-Mury, C., Zhao, K., and zhuan Su, X., "Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum.*," *BMC Genomics*, vol. 12, p. 587, 2011.

[90] Maier, A. G., Rug, M., O'Neill, M. T., Brown, M., Chakravorty, S., Szestak, T., Chesson, J., Wu, Y., Hughes, K., Coppel, R. L., Newbold, C., Beeson, J. G., Craig, A., Crabb, B. S., and Cowman, A. F., "Exported proteins required for virulence and rigidity of *Plasmodium falciparum*-infected human erythrocytes.," *Cell*, vol. 134, pp. 48–61, Jul 2008.

[91] Marti, M., Baum, J., Rug, M., Tilley, L., and Cowman, A. F., "Signal-mediated export of proteins from the malaria parasite to the host erythrocyte.," *J Cell Biol*, vol. 171, pp. 587–592, Nov 2005.

[92] Medana, I. M., Day, N. P. J., Sachanonta, N., Mai, N. T. H., Dondorp, A. M., Pongponratn, E., Hien, T. T., White, N. J., and Turner, G. D. H., "Coma in fatal adult human malaria is not caused by cerebral oedema.," *Malar J*, vol. 10, p. 267, 2011.

[93] Mendis, K., Sina, B. J., Marchesini, P., and Carter, R., "The neglected burden of *Plasmodium vivax* malaria.," *Am J Trop Med Hyg*, vol. 64, no. 1-2 Suppl, pp. 97–106, 2001.

[94] Miller, J. L., Sack, B. K., Baldwin, M., Vaughan, A. M., and Kappe, S. H. I., "Interferon-mediated innate immune responses against malaria parasite liver stages.," *Cell Rep*, Apr 2014.

[95] Mittra, P., Singh, N., and Sharma, Y. D., "*Plasmodium vivax*: immunological properties of tryptophan-rich antigens PvTRAg 35.2 and PvTRAg 80.6.," *Microbes Infect*, vol. 12, pp. 1019–1026, Nov 2010.

[96] Miu, J., Hunt, N. H., and Ball, H. J., "Predominance of interferon-related responses in the brain during murine malaria, as identified by microarray analysis.," *Infect Immun*, vol. 76, pp. 1812–1824, May 2008.

[97] Mok, S., Imwong, M., Mackinnon, M. J., Sim, J., Ramadoss, R., Yi, P., Mayxay, M., Chotivanich, K., Liong, K.-Y., Russell, B., Socheat, D., Newton, P. N., Day, N. P. J., White, N. J., Preiser, P. R., Nosten, F., Dondorp, A. M., and Bozdech, Z., "Artemisinin resistance in *Plasmodium falciparum* is associated with an altered temporal pattern of transcription.," *BMC Genomics*, vol. 12, p. 391, 2011.

[98] Moreno, A., Cabrera-Mora, M., Garcia, A., Orkin, J., Strobert, E., Barnwell, J. W., and Galinski, M. R., "*Plasmodium coatneyi* in rhesus macaques replicates the multisystemic dysfunction of severe malaria in humans.," *Infect Immun*, vol. 81, pp. 1889–1904, Jun 2013.

[99] Muehlenbachs, A., Fried, M., Lachowitzer, J., Mutabingwa, T. K., and Duffy, P. E., "Genome-wide expression analysis of placental malaria reveals features of lymphoid neogenesis during chronic infection.," *J Immunol*, vol. 179, pp. 557–565, Jul 2007.

[100] Mueller, I., Galinski, M. R., Baird, J. K., Carlton, J. M., Kochar, D. K., Alonso, P. L., and del Portillo, H. A., "Key gaps in the knowledge of *Plasmodium vivax*, a neglected human malaria parasite.," *Lancet Infect Dis*, vol. 9, pp. 555–566, Sep 2009.

[101] MYATT, A. V., HERNANDEZ, T., and COATNEY, G. R., "Studies in human malaria. XXXIII. The toxicity of pyrimethamine (Daraprim) in man.," *Am J Trop Med Hyg*, vol. 2, pp. 788–794, Sep 1953.

[102] Nath, A. P., Arafat, D., and Gibson, G., "Using blood informative transcripts in geographical genomics: impact of lifestyle on gene expression in Fijians.," *Front Genet*, vol. 3, p. 243, 2012.

[103] Neafsey, D. E., Galinsky, K., Jiang, R. H. Y., Young, L., Sykes, S. M., Saif, S., Gujja, S., Goldberg, J. M., Young, S., Zeng, Q., Chapman, S. B., Dash, A. P., Anvikar, A. R., Sutton, P. L., Birren, B. W., Escalante, A. A., Barnwell, J. W., and Carlton, J. M., "The malaria parasite *Plasmodium vivax* exhibits greater genetic diversity than *Plasmodium falciparum.*," *Nat Genet*, vol. 44, pp. 1046–1050, Sep 2012.

[104] Nguyen-Dinh, P., Gardner, A. L., Campbell, C. C., Skinner, J. C., and Collins, W. E., "Cultivation *in vitro* of the vivax-type malaria parasite *Plasmodium cynomolgi.*," *Science*, vol. 212, pp. 1146–1148, Jun 1981.

[105] Nilsson, R., Jain, M., Madhusudhan, N., Sheppard, N. G., Strittmatter, L., Kampf, C., Huang, J., Asplund, A., and Mootha, V. K., "Metabolic enzyme expression highlights a key role for MTHFD2 and

the mitochondrial folate pathway in cancer.," *Nat Commun*, vol. 5, p. 3128, 2014.

[106] Ockenhouse, C. F., chung Hu, W., Kester, K. E., Cummings, J. F., Stewart, A., Heppner, D. G., Jedlicka, A. E., Scott, A. L., Wolfe, N. D., Vahey, M., and Burke, D. S., "Common and divergent immune response signaling pathways discovered in peripheral blood mononuclear cell gene expression patterns in presymptomatic and clinically apparent malaria.," *Infect Immun*, vol. 74, pp. 5561–5573, Oct 2006.

[107] Okell, L. C., Drakeley, C. J., Ghani, A. C., Bousema, T., and Sutherland, C. J., "Reduction of transmission from malaria patients by artemisinin combination therapies: a pooled analysis of six randomized trials.," *Malar J*, vol. 7, p. 125, 2008.

[108] O'Meara, W. P., Bejon, P., Mwangi, T. W., Okiro, E. A., Peshu, N., Snow, R. W., Newton, C. R. J. C., and Marsh, K., "Effect of a fall in malaria transmission on morbidity and mortality in Kilifi, Kenya.," *Lancet*, vol. 372, pp. 1555–1562, Nov 2008.

[109] Orjuela-Snchez, P., Karunaweera, N. D., da Silva-Nunes, M., da Silva, N. S., Scopel, K. K. G., Gonalves, R. M., Amaratunga, C., S, J. M., Socheat, D., Fairhust, R. M., Gunawardena, S., Thavakodirasah, T., Galapaththy, G. L. N., Abeysinghe, R., Kawamoto, F., Wirth, D. F., and Ferreira, M. U., "Single-nucleotide polymorphism, linkage disequilibrium and geographic structure in the malaria parasite plasmodium vivax: prospects for genome-wide association studies.," *BMC Genet*, vol. 11, p. 65, 2010.

[110] Otto, T. D., Wilinski, D., Assefa, S., Keane, T. M., Sarry, L. R., Bhme, U., Lemieux, J., Barrell, B., Pain, A., Berriman, M., Newbold, C., and Llins, M., "New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-Seq.," *Mol Microbiol*, vol. 76, pp. 12–24, Apr 2010.

[111] Palmer, C., Diehn, M., Alizadeh, A. A., and Brown, P. O., "Cell-type specific gene expression profiles of leukocytes in human peripheral blood.," *BMC Genomics*, vol. 7, p. 115, 2006.

[112] Parish, L. A., Mai, D. W., Jones, M. L., Kitson, E. L., and Rayner, J. C., "A member of the *Plasmodium falciparum* PHIST family binds to the erythrocyte cytoskeleton component band 4.1.," *Malar J*, vol. 12, p. 160, 2013.

[113] Perandin, F., Manca, N., Calderaro, A., Piccolo, G., Galati, L., Ricci, L., Medici, M. C., Arcangeletti, M. C., Snounou, G., Dettori, G., and Chezzi, C., "Development of a real-time PCR assay for detection of *Plasmodium falciparum*, *Plasmodium vivax*, and *Plasmodium ovale* for routine clinical diagnosis.," *J Clin Microbiol*, vol. 42, pp. 1214–1219, Mar 2004.

[114] Preininger, M., Arafat, D., Kim, J., Nath, A. P., Idaghdour, Y., Brigham, K. L., and Gibson, G., "Blood-informative transcripts define nine common axes of peripheral blood gene expression.," *PLoS Genet*, vol. 9, no. 3, p. e1003362, 2013.

[115] Pritchard, J. K., Stephens, M., and Donnelly, P., "Inference of population structure using multilocus genotype data.," *Genetics*, vol. 155, pp. 945–959, Jun 2000.

[116] Qin, S., Kim, J., Arafat, D., and Gibson, G., "Effect of normalization on statistical and biological interpretation of gene expression profiles.," *Front Genet*, vol. 3, p. 160, 2012.

[117] Rantala, A.-M., Taylor, S. M., Trottman, P. A., Luntamo, M., Mbewe, B., Maleta, K., Kulmala, T., Ashorn, P., and Meshnick, S. R., "Comparison of real-time PCR and microscopy for malaria parasite detection in Malawian pregnant women.," *Malar J*, vol. 9, p. 269, 2010.

[118] Rapaport, F., Khanin, R., Liang, Y., Pirun, M., Krek, A., Zumbo, P., Mason, C. E., Socci, N. D., and Betel, D., "Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data.," *Genome Biol*, vol. 14, p. R95, Sep 2013.

[119] Rees, D. C., Williams, T. N., and Gladwin, M. T., "Sickle-cell disease.," *Lancet*, vol. 376, pp. 2018–2031, Dec 2010.

[120] Regev-Rudzki, N., Wilson, D. W., Carvalho, T. G., Sisquella, X., Coleman, B. M., Rug, M., Bursac, D., Angrisano, F., Gee, M., Hill, A. F., Baum, J., and Cowman, A. F., "Cell-cell communication between malaria-infected red blood cells via exosome-like vesicles.," *Cell*, vol. 153, pp. 1120–1133, May 2013.

[121] Repsilber, D., Kern, S., Telaar, A., Walzl, G., Black, G. F., Selbig, J., Parida, S. K., Kaufmann, S. H. E., and Jacobsen, M., "Biomarker discovery in heterogeneous tissue samples -taking the in-silico deconfounding approach.," *BMC Bioinformatics*, vol. 11, p. 27, 2010.

[122] Roch, K. G. L., Zhou, Y., Blair, P. L., Grainger, M., Moch, J. K., Haynes, J. D., Vega, P. D. L., Holder, A. A., Batalov, S., Carucci, D. J., and Winzeler, E. A., "Discovery of gene function by expression profiling of the malaria parasite life cycle.," *Science*, vol. 301, pp. 1503–1508, Sep 2003.

[123] RTS, S Clinical Trials PartnershipAlbert Schweitzer Hospital, L., Agnandji, S. T., Lell, B., Fernandes, J. F., Abossolo, B. P., Methogo, B. G. N. O., Kabwende, A. L., Adegnika, A. A., Mordmller, B., Issifou, S., Kremsner, P. G., Sacarlal, J., Aide, P., Lanaspa, M., Aponte, J. J., Machevo, S., Acacio, S., Bulo, H.,

Sigauque, B., Macete, E., Alonso, P., Abdulla, S., Salim, N., Minja, R., Mpina, M., Ahmed, S., Ali, A. M., Mtoro, A. T., Hamad, A. S., Mutani, P., Tanner, M., Tinto, H., D'Alessandro, U., Sorgho, H., Valea, I., Bihoun, B., Guiraud, I., Kabor, B., Sombi, O., Guiguemd, R. T., Oudraogo, J. B., Hamel, M. J., Kariuki, S., Oneko, M., Odero, C., Otieno, K., Awino, N., McMorrow, M., Muturi-Kioi, V., Laserson, K. F., Slutsker, L., Otieno, W., Otieno, L., Otsyula, N., Gondi, S., Otieno, A., Owira, V., Oguk, E., Odongo, G., Woods, J. B., Ogutu, B., Njuguna, P., Chilengi, R., Akoo, P., Kerubo, C., Maingi, C., Lang, T., Olotu, A., Bejon, P., Marsh, K., Mwambingu, G., Owusu-Agyei, S., Asante, K. P., Osei-Kwakye, K., Boahen, O., Dosoo, D., Asante, I., Adjei, G., Kwara, E., Chandramohan, D., Greenwood, B., Lusingu, J., Gesase, S., Malabeja, A., Abdul, O., Mahende, C., Liheluka, E., Malle, L., Lemnge, M., Theander, T. G., Drakeley, C., Ansong, D., Agbenyega, T., Adjei, S., Boateng, H. O., Rettig, T., Bawa, J., Sylverken, J., Sambian, D., Sarfo, A., Agyekum, A., Martinson, F., Hoffman, I., Mvalo, T., Kamthunzi, P., Nkomo, R., Tembo, T., Tegha, G., Tsidya, M., Kilembe, J., Chawinga, C., Ballou, W. R., Cohen, J., Guerra, Y., Jongert, E., Lapierre, D., Leach, A., Lievens, M., Ofori-Anyinam, O., Olivier, A., Vekemans, J., Carter, T., Kaslow, D., Leboulleux, D., Loucq, C., Radford, A., Savarese, B., Schellenberg, D., Sillman, M., and Vansadia, P., "A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants.," *N Engl J Med*, vol. 367, pp. 2284–2295, Dec 2012.

[124] Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S., "Detecting recent positive selection in the human genome from haplotype structure.," *Nature*, vol. 419, pp. 832–837, Oct 2002.

[125] Sallares, R., Bouwman, A., and Anderung, C., "The spread of malaria to Southern Europe in antiquity: new approaches to old problems.," *Med Hist*, vol. 48, pp. 311–328, Jul 2004.

[126] Scherf, A., Hernandez-Rivas, R., Buffet, P., Bottius, E., Benatar, C., Pouvelle, B., Gysin, J., and Lanzer, M., "Antigenic variation in malaria: in situ switching, relaxed and mutually exclusive transcription of var genes during intra-erythrocytic development in plasmodium falciparum.," *EMBO J*, vol. 17, pp. 5418–5426, Sep 1998.

[127] Scherf, A., Lopez-Rubio, J. J., and Riviere, L., "Antigenic variation in plasmodium falciparum.," *Annu Rev Microbiol*, vol. 62, pp. 445–470, 2008.

[128] Schmidt, L. H., "Compatibility of relapse patterns of *Plasmodium cynomolgi* infections in rhesus monkeys with continuous cyclical development and hypnozoite concepts of relapse.," *Am J Trop Med Hyg*, vol. 35, pp. 1077–1099, Nov 1986.

[129] Schultz, L. J., Ettling, M., Chitsulo, L., Steketee, R. W., Nyasulu, Y., Macheso, A., and Nwanyanwu, O. C., "A nation-wide malaria knowledge, attitudes and practices survey in Malawi: objectives and methodology.," *Trop Med Parasitol*, vol. 45, pp. 54–56, Mar 1994.

[130] Seder, R. A., Chang, L.-J., Enama, M. E., Zephir, K. L., Sarwar, U. N., Gordon, I. J., Holman, L. A., James, E. R., Billingsley, P. F., Gunasekera, A., Richman, A., Chakravarty, S., Manoj, A., Velmurugan, S., Li, M., Ruben, A. J., Li, T., Eappen, A. G., Stafford, R. E., Plummer, S. H., Hendel, C. S., Novik, L., Costner, P. J. M., Mendoza, F. H., Saunders, J. G., Nason, M. C., Richardson, J. H., Murphy, J., Davidson, S. A., Richie, T. L., Sedegah, M., Sutamihardja, A., Fahle, G. A., Lyke, K. E., Laurens, M. B., Roederer, M., Tewari, K., Epstein, J. E., Sim, B. K. L., Ledgerwood, J. E., Graham, B. S., Hoffman, S. L., and Team, V. R. C. . S., "Protection against malaria by intravenous immunization with a nonreplicating sporozoite vaccine.," *Science*, vol. 341, pp. 1359–1365, Sep 2013.

[131] Shen-Orr, S. S., Tibshirani, R., Khatri, P., Bodian, D. L., Staedtler, F., Perry, N. M., Hastie, T., Sarwal, M. M., Davis, M. M., and Butte, A. J., "Cell type-specific gene expression differences in complex tissues.," *Nat Methods*, vol. 7, pp. 287–289, Apr 2010.

[132] Shin, Y. S., Chan, C., Vidal, A. J., Brody, T., and Stokstad, E. L., "Subcellular localization of gamma-glutamyl carboxypeptidase and of folates.," *Biochim Biophys Acta*, vol. 444, pp. 794–801, Oct 1976.

[133] Shock, J. L., Fischer, K. F., and DeRisi, J. L., "Whole-genome analysis of mRNA decay in *Plasmodium falciparum* reveals a global lengthening of mRNA half-life during the intra-erythrocytic development cycle.," *Genome Biol*, vol. 8, no. 7, p. R134, 2007.

[134] Sibley, C. H., Hyde, J. E., Sims, P. F., Plowe, C. V., Kublin, J. G., Mberu, E. K., Cowman, A. F., Winstanley, P. A., Watkins, W. M., and Nzila, A. M., "Pyrimethamine-sulfadoxine resistance in *Plasmodium falciparum*: what next?," *Trends Parasitol*, vol. 17, pp. 582–588, Dec 2001.

[135] Simon Anders, Paul T. Pyl, W. H., "HTSeq - A Python framework to work with high-throughput sequencing data," *bioRxiv*, vol. 1, pp. 1–5, February 2014.

[136] Sinha, A., Hughes, K. R., Modrzynska, K. K., Otto, T. D., Pfander, C., Dickens, N. J., Religa, A. A., Bushell, E., Graham, A. L.,

Cameron, R., Kafsack, B. F. C., Williams, A. E., Llins, M., Berriman, M., Billker, O., and Waters, A. P., "A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium.*," *Nature*, vol. 507, pp. 253–257, Mar 2014.

[137] Sixsmith, D. G., Watkins, W. M., Chulay, J. D., and Spencer, H. C., "In vitro antimalarial activity of tetrahydrofolate dehydrogenase inhibitors.," *Am J Trop Med Hyg*, vol. 33, pp. 772–776, Sep 1984.

[138] Snounou, G., Viriyakosol, S., Jarra, W., Thaithong, S., and Brown, K. N., "Identification of the four human malaria parasite species in field samples by the polymerase chain reaction and detection of a high prevalence of mixed infections.," *Mol Biochem Parasitol*, vol. 58, pp. 283–292, Apr 1993.

[139] Soneson, C. and Delorenzi, M., "A comparison of methods for differential expression analysis of RNA-seq data.," *BMC Bioinformatics*, vol. 14, p. 91, 2013.

[140] Sorber, K., Dimon, M. T., and DeRisi, J. L., "Rna-seq analysis of splicing in plasmodium falciparum uncovers new splice junctions, alternative splicing and splicing of antisense transcripts.," *Nucleic Acids Res*, vol. 39, pp. 3820–3835, May 2011.

[141] Su, X. Z., Heatwole, V. M., Wertheimer, S. P., Guinet, F., Herrfeldt, J. A., Peterson, D. S., Ravetch, J. A., and Wellems, T. E., "The large diverse gene family *var* encodes proteins involved in cytoadherence and antigenic variation of *Plasmodium falciparum*-infected erythrocytes.," *Cell*, vol. 82, pp. 89–100, Jul 1995.

[142] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., and Mesirov, J. P., "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.," *Proc Natl Acad Sci U S A*, vol. 102, pp. 15545–15550, Oct 2005.

[143] Tachibana, S.-I., Sullivan, S. A., Kawai, S., Nakamura, S., Kim, H. R., Goto, N., Arisue, N., Palacpac, N. M. Q., Honma, H., Yagi, M., Tougan, T., Katakai, Y., Kaneko, O., Mita, T., Kita, K., Yasutomi, Y., Sutton, P. L., Shakhbatyan, R., Horii, T., Yasunaga, T., Barnwell, J. W., Escalante, A. A., Carlton, J. M., and Tanabe, K., "*Plasmodium cynomolgi* genome sequences provide insight into *Plasmodium vivax* and the monkey malaria clade.," *Nat Genet*, vol. 44, pp. 1051–1055, Sep 2012.

[144] TARLOV, A. R., BREWER, G. J., CARSON, P. E., and ALVING, A. S., "Primaquine sensitivity. glucose-6-phosphate dehydrogenase deficiency: an inborn error of metabolism of medical and biological significance.," *Arch Intern Med*, vol. 109, pp. 209–234, Feb 1962.

[145] Tarun, A. S., Peng, X., Dumpit, R. F., Ogata, Y., Silva-Rivera, H., Camargo, N., Daly, T. M., Bergman, L. W., and Kappe, S. H. I., "A combined transcriptome and proteome survey of malaria parasite liver stages.," *Proc Natl Acad Sci U S A*, vol. 105, pp. 305–310, Jan 2008.

[146] Team, R. D. C., *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. ISBN 3-900051-07-0.

[147] Tishkoff, S. A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., Piro, A., Stoneking, M., Tagarelli, A., Tagarelli, G., Touma, E. H., Williams, S. M., and Clark, A. G., "Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance.," *Science*, vol. 293, pp. 455–462, Jul 2001.

[148] Tonkin, C. J., Carret, C. K., Duraisingh, M. T., Voss, T. S., Ralph, S. A., Hommel, M., Duffy, M. F., da Silva, L. M., Scherf, A., Ivens, A., Speed, T. P., Beeson, J. G., and Cowman, A. F., "Sir2 paralogues cooperate to regulate virulence genes and antigenic variation in plasmodium falciparum.," *PLoS Biol*, vol. 7, p. e84, Apr 2009.

[149] Trager, W. and Gill, G. S., "Enhanced gametocyte formation in young erythrocytes by *Plasmodium falciparum in vitro*.," *J Protozool*, vol. 39, no. 3, pp. 429–432, 1992.

[150] Trager, W., Gill, G. S., Lawrence, C., and Nagel, R. L., "*Plasmodium falciparum*: enhanced gametocyte formation *in vitro* in reticulocyte-rich blood.," *Exp Parasitol*, vol. 91, pp. 115–118, Feb 1999.

[151] Trapnell, C., Pachter, L., and Salzberg, S. L., "TopHat: discovering splice junctions with RNA-Seq.," *Bioinformatics*, vol. 25, pp. 1105–1111, May 2009.

[152] Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L., "Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.," *Nat Biotechnol*, vol. 28, pp. 511–515, May 2010.

[153] Tyagi, R. K. and Sharma, Y. D., "Erythrocyte binding activity displayed by a selective group of *Plasmodium vivax* tryptophan rich antigens is inhibited by patients' antibodies.," *PLoS One*, vol. 7, no. 12, p. e50754, 2012.

[154] Uppal, K., Soltow, Q. A., Strobel, F. H., Pittard, W. S., Gernert, K. M., Yu, T., and Jones, D. P., "xMSanalyzer: automated pipeline for

improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data.," *BMC Bioinformatics*, vol. 14, p. 15, 2013.

[155] van Duivenvoorde, L. M., van der Wel, A. V., van der Werff, N. M., Braskamp, G., Remarque, E. J., Kondova, I., Kocken, C. H. M., and Thomas, A. W., "Suppression of *Plasmodium cynomolgi* in rhesus macaques by coinfection with *Babesia microti*.," *Infect Immun*, vol. 78, pp. 1032–1039, Mar 2010.

[156] Waisberg, M., Lin, C. K., Huang, C.-Y., Pena, M., Orandle, M., Bolland, S., and Pierce, S. K., "The impact of genetic susceptibility to systemic lupus erythematosus on placental malaria in mice.," *PLoS One*, vol. 8, no. 5, p. e62820, 2013.

[157] Waisberg, M., Tarasenko, T., Vickers, B. K., Scott, B. L., Willcocks, L. C., Molina-Cruz, A., Pierce, M. A., yu Huang, C., Torres-Velez, F. J., Smith, K. G. C., Barillas-Mury, C., Miller, L. H., Pierce, S. K., and Bolland, S., "Genetic susceptibility to systemic lupus erythematosus protects against cerebral malaria in mice.," *Proc Natl Acad Sci U S A*, vol. 108, pp. 1122–1127, Jan 2011.

[158] Walton, E., Oliveros, H., and Villamor, E., "Hemoglobin concentration and parasitemia on hospital admission predict risk of multiple organ dysfunction syndrome among adults with malaria.," *Am J Trop Med Hyg*, May 2014.

[159] Waxman, S. and Herbert, V., "Mechanism of pyrimethamine-induced megaloblastosis in human bone marrow.," *N Engl J Med*, vol. 280, pp. 1316–1319, Jun 1969.

[160] Weppelmann, T. A., Carter, T. E., Chen, Z., von Fricken, M. E., Victor, Y. S., Existe, A., and Okech, B. A., "High frequency of the erythroid silent duffy antigen genotype and lack of plasmodium vivax infections in haiti.," *Malar J*, vol. 12, p. 30, 2013.

[161] Whitney, A. R., Diehn, M., Popper, S. J., Alizadeh, A. A., Boldrick, J. C., Relman, D. A., and Brown, P. O., "Individuality and variation in gene expression patterns in human blood.," *Proc Natl Acad Sci U S A*, vol. 100, pp. 1896–1901, Feb 2003.

[162] Wood, E. T., Stover, D. A., Slatkin, M., Nachman, M. W., and Hammer, M. F., "The beta-globin recombinational hotspot reduces the effects of strong selection around HbC, a recently arisen mutation providing resistance to malaria.," *Am J Hum Genet*, vol. 77, pp. 637–642, Oct 2005.

[163] Ylostalo, J., Randall, A. C., Myers, T. A., Metzger, M., Krogstad, D. J., and Cogswell, F. B., "Transcriptome profiles of host gene expression in a monkey model of human malaria.," *J Infect Dis*, vol. 191, pp. 400–409, Feb 2005.

[164] Young, J. A., Fivelman, Q. L., Blair, P. L., de la Vega, P., Roch, K. G. L., Zhou, Y., Carucci, D. J., Baker, D. A., and Winzeler, E. A., "The *Plasmodium falciparum* sexual development transcriptome: a microarray analysis using ontology-based pattern identification.," *Mol Biochem Parasitol*, vol. 143, pp. 67–79, Sep 2005.

[165] Zhang, Q., Huang, Y., Zhang, Y., Fang, X., Claes, A., Duchateau, M., Namane, A., Lopez-Rubio, J.-J., Pan, W., and Scherf, A., "A critical role of perinuclear filamentous actin in spatial repositioning and mutually exclusive expression of virulence genes in malaria parasites.," *Cell Host Microbe*, vol. 10, pp. 451–463, Nov 2011.

VITA

Kevin J. Lee was born in Cincinnati, OH, and was raised in Marietta, GA. After attending the Marist School in Atlanta, he completed his undergraduate degrees in Biology and Ecology at the University of Georgia. He then taught high school sciences for two years before returning for a Master's degree in Biological and Agricultural Engineering. Subsequently he began his PhD at the Georgia Institute of Technology where he has studied many diseases using next generation sequence technology.