Multi-omic and Multi-scale Data Integration for the Characterization of Malaria Infection in Non-Human Primates

by

Yi Heng Yan

(Under the Direction of Juan B. Gutierrez)

Abstract

Plasmodium parasites were identified as the cause of malaria more than 200 years ago. However, malaria remains a public health burden responsible for approximately 400,000 (236,000 $\sim$ 635,000) death in 2015. Severe malaria is responsible for the majority of malaria mortality, yet the understanding of mechanisms of host responses underlying severe malaria pathology remains incomplete. The objective of this project is to identify and characterize host transcriptomic, cellular and cytokine responses that are associated with malaria severity.

To quantify the removal of healthy red blood cells (hRBCs) by the host, we created a novel mathematical model that could capture the various outcomes of malaria infection. This model was fitted to the Malaria Host-Pathogen Interaction Center (MaHPIC) time series data set of five *Macaca mulatta* infected with Plasmodium cynomolgi. Using the fitted model, our group discovered association of the loss of healthy red blood cells and pro-inflammatory cytokines and CD-8 T cell population.

Furthermore, our group also created novel statistical tools for the identification of differential networks. Through the application of both traditional bioinformatics analysis tools and

differential network analysis, our group characterized severe malaria infection with differential transcriptional up-regulation of genes linked with response to the pathogen-associated molecular pattern (PAMP) and pro-inflammatory cytokines.

Through a combined approach of mathematical modeling, differential network analysis and traditional bioinformatics analysis, we were able to identify host transcriptomic, cellular and cytokine responses that are associated with both malaria severity and host removal of healthy red blood cells. This project provides novel insight into the molecular and cellular basis for the development of severe malaria.

INDEX WORDS: Malaria, *Plasmodium cynomolgi*, Bioinformatics, Partial Differential Equation Model, Differential Network Analysis

Multi-omic and Multi-scale Data Integration for the Characterization of Malaria Infection in Non-Human Primates

by

Yi Heng Yan

B.S., Rutgers, The State University of New Jersey, 2010

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2017

Multi-omic and Multi-scale Data Integration for the Characterization of Malaria Infection in Non-Human Primates

by

Yi Heng Yan

Approved:

Major Professor: Juan B. Gutierrez

Committee: Jonathan Arnold
Julie Moore
Ying Xu

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2017

Dedication

This thesis is dedicated to my advisor Dr. Juan B. Gutierrez for his exemplary role as a Phd advisor.

Acknowledgments

I would like to thank Elizabeth Trippe for helping me proofread my thesis and Diego Moncada for biological insights. I would also like to thank every one involved in the Malaria Host-Pathogen Interaction Center for collecting some of the data used in this thesis.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction and Literature Review

1.1 Background and Problem Statement

Malaria is a major public health burden, responsible for approximately 400,000 (236,000 $\sim$ 635,000) deaths in 2015 [33]. Out of the five human *Plasmodium* species capable of causing malaria, *P. falciparum* and *P. vivax* account for the majority of human malaria infections. *Plasmodium* is a large genus of parasitic protozoa (unicellular eukaryotic organisms), with complex genomes and sophisticated life cycles. The genome, behavior and epidemiological characterization of Plasmodium are orders of magnitude more complex than that of viruses or bacteria. The clinical manifestation of infection by Plasmodium, the disease malaria, has a broad spectrum of symptoms, varying from asymptomatic to highly severe. Malaria affects birds, reptiles, and some mammals (mostly rodents and primates).

The Plasmodium life cycle is comprised of several stages, as demonstrated in Figure 1.1. The infection process in humans starts with the injection of sporozoites by female anopheline mosquitoes into the skin of the host. This is followed by the liver stage, in which the inoculated sporozoites grow and multiply asexually within hepatocytes for 1-2 weeks to produce merozoites. The newly produced merozoites emerge from the liver and enter the bloodstream. The blood-stage infection starts immediately after the hepatic stage; merozoites invade red blood cells (RBCs) where they also reproduce asexually. In some instances, the parasites develop into the sexual stage form called gametocytes. Gametocyte infected red blood cells (iRBCs) infect newly feeding anopheline mosquitoes and through sexual reproduction, followed by many more rounds of asexual multiplication, are ultimately transformed

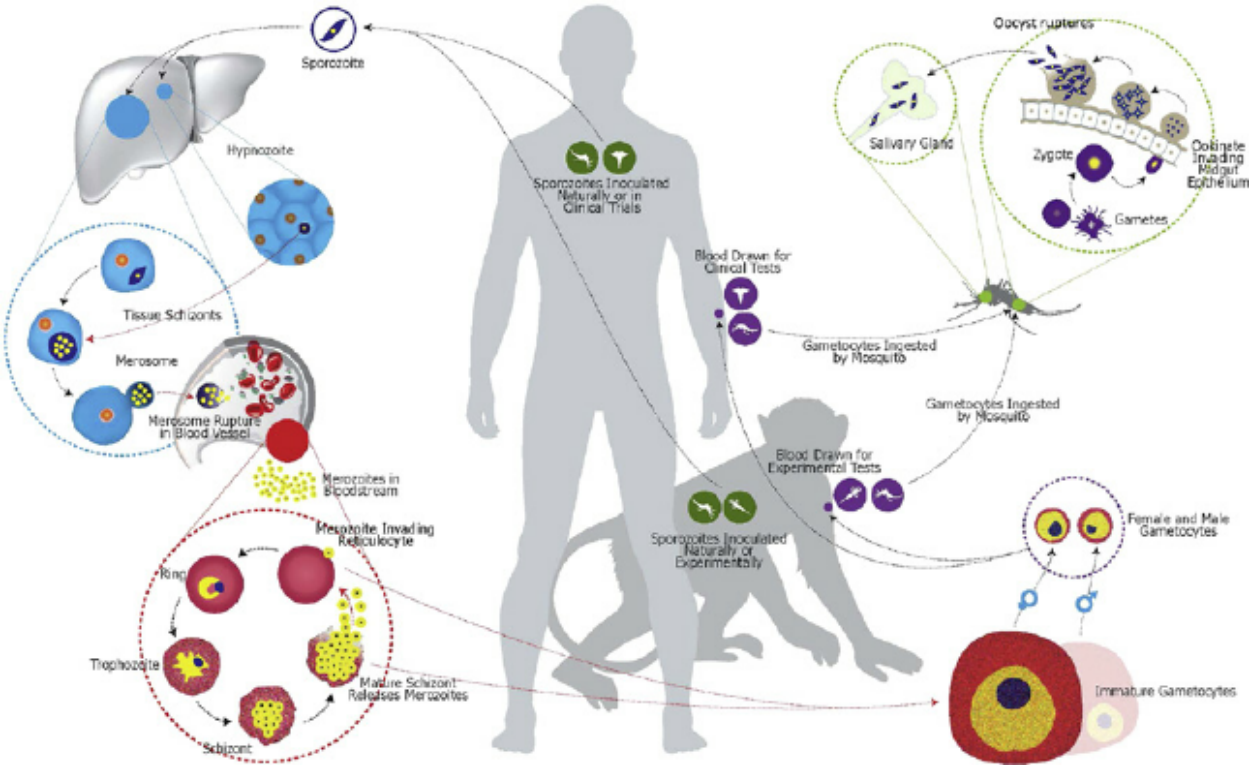

Figure 1.1: Schematic Diagram Describing the Malaria Life Stages. Original graphic generated by Galinski *et al.* [16]

into sporozoites, thus completing the infection cycle between an Anopheline mosquito and its host.

The parasite's blood stage infection in both human and non-human primates has a regular cycle of 24 or 72 hours depending on the species of the *Plasmodium* parasite [6, 11]. The parasites invade healthy RBCs and replicate asexually, remodeling and ultimately destroying the RBCs in the process. The destruction of RBCs during blood-stage malaria infection sometimes results in severe anemia, which is one of the major complications of malaria and a leading cause of mortality. Anemia caused by malaria infection is a complicated process not only involving the malaria parasites and host RBCs, but also the innate and adaptive immune system of the host [26, 35].

The proposed research aims to understand better the development of severe malaria symptoms with an emphasis on severe anemia through the multi-omic and multi-scale characterization of the host response to malaria infection associated with severe malaria.

1.2 Objective, Hypothesis, and Specific Aims

Development of Mathematical Model Describing Cellular Dynamics During Malaria Infection

Our first aim is to develop a mathematical model that simulates the with-in host dynamics of healthy red blood cells and infected red blood cells. One of the earlier models describing the interaction among red blood cells, malaria parasite and host immune response is published by Anderson *et al.* in 1989 [2]. The Anderson model included four compartments: healthy red blood cells, infected red blood cells, free merozoite and immune effector response. The model was used to demonstrate that merozoite based intervention is less effective than intervention aimed against infected red blood cells. One important aspect of malaria infection which the Anderson model failed to capture was the possible disruption of erythropoietic process [13, 32, 35]; erythropoiesis refers to the differentiation and maturation of hematopoietic stem cells (HSCs) through committed cell lineages culminating in the production of erythrocytes/RBCs.

Multiple other models have been proposed loosely based on the Anderson model. In the model published by Hellriegel *et al.* [21], the author explored the scenario of superinfection, where multiple malaria parasite species co-infects a host and concluded that in the absence of host immune response, the more virulent parasite strain would dominate the parasite population. Similar to the Anderson model, the Hellriegel model failed to incorporate parasite and red blood cell age structures and variable erythropoiesis process. Later models by Grosvenor *et al* [18] included age compartment for the parasites. Models including the age structure of red blood cells have also been proposed [29]. Additionally, taking advantage of

the regularity of malaria life cycle and possible synchronicity of parasite populations during blood stage infection [37], several discrete time step models has also been proposed [23, 31].

Our model aims to include age structure for both healthy red blood cells and infected red blood cells. The inclusion of age structure allows the model to accommodate preferential infection of young red blood cells by *P. vivax* and varying degrees of synchronicity of the parasite. Furthermore, the model captures the three import factors contributing to severe anemia caused by malaria infection: (i) The destruction of RBCs directly by the parasite, (ii) the destruction of RBCs by a non-parasitic factor, possibly the host immune system, and (iii) the disruption of the host's erythropoietic process.

Classic models of erythropoiesis rest on three variables representing the precursor/progenitor population of immature erythrocytes, the population of mature RBCs, and the concentration of erythropoietin (Epo), a signaling molecule critical to erythrocyte production. Mackey and Milton (1990) represented these three variables with a system of ordinary differential equations (ODEs) with one delay accounting for the time for EPO-dependent maturation of progenitors into erythrocytes (about six days) [27]. In 1995, Belair *et al.* developed an "age-structured" partial differential equation (PDE) model for erythropoiesis [4]. Belair *et al.*'s model characterizes the proliferation and aging of the precursor and mature erythrocyte populations as nonlinear, first-order PDEs and includes an ODE to represent EPO dynamics over time. Building on Belair's 1995 model, Mahaffy *et al.* (1998) investigated the addition of state-dependent delays to the model [28].

Due to the difficulty involved in measuring EPO accurately over an extended period, we formulated the rate at which mature RBCs enter circulation as only dependent upon the current number of RBCs and the number of RBCs at equilibrium. Such a formulation has previously been shown by Savil *et al* in 2009. To capture the erythropoiesis process using data derived from a murine model [39]. As the only formulation of an erythropoiesis model without an EPO component that has been validated by time course data, we believe that

such a formulation would also be able to capture the fluctuation in the erythropoietic process due to loss of RBCs within our hemodynamic system.

To capture the complex interactions between different age groups of iRBCs, RBCs and various immune cells during blood stage malaria [30], we propose a coupled age-structured PDE model of both iRBCs and host RBCs. Our model borrows from the erythropoiesis PDE model [1, 45] and expands upon it by adding the parasite age structure system and the effect of immune cells on both RBCs and iRBCs. We hypothesize that such a model will be able to recreate the three clinical outcomes of malaria infection: the death of the host, clearance of parasites and chronic disease where the parasite and the host co-exists. Additionally, the model can be used to quantify the rate of the removal of healthy red blood cells.

Development of Novel Statistical Tools for the Integrative Analysis of Multi-Scale Data

Our second aim is to develop novel statistical tools for the integrative analysis of multi-scale and multi-omics data. The advent of high-throughput molecular technologies has allowed us to measure a large number of variables simultaneously. Transcriptomic experiments typically measure the abundance of 20,000 or so transcripts. The number of variables measured by other "omic" technologies such as lipidomic, proteomic, glycomic and metabolomic studies is of similar magnitude. Time series -omic data refers to molecular snapshots taken using these -omic technologies on a time trajectory. Time series -omic data captures time-dependent molecular dynamics and is a powerful tool in the study of disease progression [3], developmental processes [38] and vaccination [34].

A variety of tools has been developed for the analysis of time series -omic data. Methods such as MESS [5] and EDGE [41] aim to discover individual gene expression time series that are significantly different between two experimental conditions. Other methods such as TcGSA [20], CAMERA [51], and GSEA-TS [43] seek to find time series of pre-defined gene

sets that are significantly different among groups. Furthermore, clustering based time series analysis tools [14, 42] have also been developed.

Current time series analysis tools focus on finding gene expression time series or predefined gene sets that most likely have changed between groups but ignore the changes in pairwise gene dynamics. Pairwise dynamics between genes can be quantified using correlation, mutual information or a distance metric. Differential correlation has been used to study gene association with the clinical outcome of lung cancer [40] and estrogen receptor modulation in hormonal cancers [22].

Differential network analysis (DiNA) refers to a recent school of algorithms focused on identifying differences in network topology between states. Unlike traditional differential analysis, DiNA identifies changes in the pairwise dynamics of genes rather than the shift in abundance of individual genes. DiNA has been successfully used for the identification of transcriptional regulator [8] and estrogen modulated genes in cancer [22]. Various methods have been proposed to conduct DiNA, such as modulator inference by network dynamics (MINDy) [48], differential network analysis in genomics (DINGO) [19] and modulated gene interaction (MAGIC) analysis [22]. Each algorithm adapts its own metric to detect differential dynamics between genes, MINDy uses mutual information based measurement of dynamics between two genes, MAGIC uses Spearman correlation, and DINGO utilizes a Gaussian graphical model. Each metric aims to characterize pair-wise dynamics based on a derived quantity, and each metric has its own limitations, for example, Pearson's correlation coefficient only detects changes in linear dependency between two variables.

We aim to develop a novel metric to apply differential network analysis to time series data that overcomes the limitation of correlation measurement. Furthermore, we aim to develop an ensemble differential network analysis tool to leverage the strength of multiple network quantification methods. Our group hypothesizes that our novel differential network analysis can reveal underlying host response that is associated with malaria severity.

Characterization of Host Response to Malaria Infection Associated with Disease Severity in *Macaca mulatta* infected with *Plosmodium cynomolgi*

Our third aim is to apply the developed model and statistical tools to characterize the association between host response and malaria severity. Specifically in *Macaca mulatta* infected with *P. cynomolgi*. Our goal is to identify differentially regulated genes, cell types and immune functions that are unique to hosts experiencing severe malaria. The identification of these entities will allow us to generate novel hypothesis explaining the underlying mechanism of severe malaria.

Despite the vast potential of host transcriptome data to help elucidate the molecular mechanisms underlying malaria pathology, very few primate host transcriptome studies have been conducted. In 2005, Joni Ylostalo *et al* generated the first transcriptome time series of two *Plasmodium cynolomolgi* infected *M. mulattas* [54]. In 2014, Junya Yamagishi *et al* conducted RNA-seq analysis of 116 Indonesian patients infected with *Plasmodium falciparum* and discovered sets of host genes that correlate with the severity of malaria infections [53].

Plasmodium cynolmogi is a non-human primate parasite that infects old world monkeys and is capable of recapitulate clinical and histopathological findings of vivax malaria patients[12, 15, 44]. It is both genetically and physiologically similar to *P. vivax* [44, 50]. For instance, both parasites exhibit 48-h erythrocytic cycle during blood stage infection [9], preferential infection of reticulocytes [49] and form hypnozoites, which are dormant, liver stage forms that can activate and cause relapse infections [25]. Due to the difficulty of studying *P. vivax* pathogenesis, *P. cynomolgi* infection of rhesus monkeys (*M. mulatta*) has being used to better understand hypnozoite caused relapse [10].

Host clearance of malaria parasites without complication requires a concerted effort between inflammatory and anti-inflammatory cytokines; their balance and timing are critical in determining clinical outcome [17]. The association of cytokine, transcriptomic and immune response and clinical outcome has been extensively studied in *P. falciparum* infection [7, 36, 46, 47, 52], in comparison, much less is known for *P. vivax* infection.

With the aim to better characterize *P. vivax* infection in humans, a time series experiment where five *M. mulatta* were infected with *P. cynomolgi* was conducted as a part of the Malaria Pathogen-Host Interaction Center (MaHPIC) project [24]. This experiment captured host transcriptomic, cellular and cytokine response to *P. cynomolgi*. The subjects within this study responded to the infection in different manners and resulted in one death, two cases of severe malaria, and two cases of mild malaria. The presence of both severe malaria and mild malaria provided us the opportunity to characterize host responses that associated with clinical outcome.

We hypothesize that the temporal immune and cytokine profiles of the two subjects experiencing severe malaria and the two subjects experiencing mild malaria are significantly different. Identification of the specific genes, cellular population and cytokine response uniquely associated with severe malaria should be enriched in biological functions that can shed light on the underlying molecular mechanisms associated with severe malaria pathology. Furthermore, the rate of healthy red blood cell removal for all four subjects will be estimated using the model described in chapter 2, and we expect to identify sets of genes, cellular population, and cytokine that are highly correlated with the rate of healthy red blood cell removal. These findings will provide insight into the mechanisms of malaria induced anemia.

1.3 References

Bibliography

[1] A. S. Ackleh, B. Ma, and J. J. Thibodeaux. A second-order high resolution finite difference scheme for a structured erythropoiesis model subject to malaria infection. *Mathematical biosciences*, 2013.

[2] R. Anderson, R. May, and S. Gupta. Non-linear phenomena in hostparasite interactions. *Parasitology*, 99(S1):S59–S79, 1989.

[3] C. Bécavin, N. Tchitchek, C. Mintsa-Eya, A. Lesne, and A. Benecke. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*, 27(10):1413–1421, 2011.

[4] J. Belair, M. C. Mackey, and J. M. Mahaffy. Age-structured and two-delay models for erythropoiesis. *Mathematical Biosciences*, 128(1):317–346, 1995.

[5] M. Berk, C. Hemingway, M. Levin, and G. Montana. Longitudinal analysis of gene expression profiles using functional mixed-effects models. In *Advanced Statistical Methods for the Analysis of Large Data-Sets*, pages 57–67. Springer, 2012.

[6] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS biology*, 1(1):e5, 2003.

[7] H. Chavale, J. R. Santos-Oliveira, A. M. Da-Cruz, and S. Enosse. Enhanced t cell activation in plasmodium falciparum malaria-infected human immunodeficiency virus-1 patients from mozambique. *Memórias do Instituto Oswaldo Cruz*, 107(8):985–992, 2012.

[8] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005.

[9] G. Coatney, W. Collins, M. Warren, and P. Contacos. The primate malarias. 1971. *Washington: US Government Printing Office Google Scholar*.

[10] F. B. Cogswell. The hypnozoite and relapse in primate malaria. *Clinical microbiology reviews*, 5(1):26–35, 1992.

[11] W. E. Collins, M. Warren, J. S. Sullivan, and G. G. Galland. Plasmodium coatneyi: observations on periodicity, mosquito infection, and transmission to macaca mulatta monkeys. *The American journal of tropical medicine and hygiene*, 64(3):101–110, 2001.

[12] G. A. Deye, M. Gettayacamin, P. Hansukjariya, R. Im-erbsin, J. Sattabongkot, Y. Rothstein, L. Macareo, S. Fracisco, K. Bennett, A. J. Magill, et al. Use of a rhesus plasmodium cynomolgi model to screen for anti-hypnozoite activity of pharmaceutical substances. *The American journal of tropical medicine and hygiene*, 86(6):931–935, 2012.

[13] N. M. Douglas, N. M. Anstey, P. A. Buffet, J. R. Poespoprodjo, T. W. Yeo, N. J. White, R. N. Price, et al. The anaemia of plasmodium vivax malaria. *Malar J*, 11(135.10):1186, 2012.

[14] J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168, 2005.

[15] M. R. Galinski and J. W. Barnwell. Plasmodium vivax: who cares? *Malaria Journal*, 7(1):S9, 2008.

[16] M. R. Galinski, E. Meyer, and J. W. Barnwell. Plasmodium vivax: modern strategies to study a persistent parasites life cycle. *Adv Parasitol*, 81:1–26, 2013.

[17] R. M. Gonçalves, N. F. Lima, and M. U. Ferreira. Parasite virulence, co-infections and cytokine balance in malaria. *Pathogens and global health*, 108(4):173–178, 2014.

[18] M. B. Gravenor, A. L. Lloyd, P. G. Kremsner, M. A. Missinou, M. English, K. Marsh, and D. Kwiatkowski. A model for estimating total parasite load in falciparum malaria patients. *Journal of theoretical biology*, 217(2):137–148, 2002.

[19] M. J. Ha, V. Baladandayuthapani, and K.-A. Do. Dingo: differential network analysis in genomics. *Bioinformatics*, 31(21):3413–3420, 2015.

[20] B. P. Hejblum, J. Skinner, and R. Thiébaut. Time-course gene set analysis for longitudinal gene expression data. *PLoS Comput Biol*, 11(6):e1004310, 2015.

[21] B. Hellriegel. Modelling the immune response to malaria with ecological concepts: short-term behaviour against long-term equilibrium. *Proceedings of the Royal Society of London B: Biological Sciences*, 250(1329):249–256, 1992.

[22] T.-H. Hsiao, Y.-C. Chiu, P.-Y. Hsu, T.-P. Lu, L.-C. Lai, M.-H. Tsai, T. H.-M. Huang, E. Y. Chuang, and Y. Chen. Differential network analysis reveals the genome-wide landscape of estrogen receptor modulation in hormonal cancers. *Scientific reports*, 6, 2016.

[23] G. Jakeman, A. Saul, W. Hogarth, and W. Collins. Anaemia of acute malaria infections in non-immune patients primarily results from destruction of uninfected erythrocytes. *Parasitology*, 119(02):127–133, 1999.

[24] C. Joyner, A. Moreno, E. V. Meyer, M. Cabrera-Mora, J. C. Kissinger, J. W. Barnwell, and M. R. Galinski. Plasmodium cynomolgi infections in rhesus macaques display clinical and parasitological features pertinent to modelling vivax malaria pathology and relapse infections. *Malaria Journal*, 15(1):451, 2016.

[25] W. Krotoski, R. Bray, P. Garnham, R. Gwadz, R. Killick-Kendrick, C. Draper, G. Targett, D. Krotoski, M. Guy, L. Koontz, et al. Observations on early and late post-sporozoite tissue stages in primate malaria. ii. the hypnozoite of plasmodium cynomolgi bastianellii from 3 to 105 days after infection, and detection of 36-to 40-hour pre-erythrocytic forms. *The American journal of tropical medicine and hygiene*, 31(2):211–225, 1982.

[26] T. J. Lamb and J. Langhorne. The severity of malarial anaemia in plasmodium chabaudi infections of balb/c mice is determined independently of the number of circulating parasites. *Malaria journal*, 7(1):68, 2008.

[27] M. Mackey and J. Milton. *Feedback, delays and the origin of blood cell dynamics*. University of Minnesota. Institute for Mathematics and Its Applications, 1990.

[28] J. M. Mahaffy, J. Belair, and M. C. Mackey. Hematopoietic model with moving boundary condition and state dependent delay] applications in erythropoiesis. *Journal of Theoretical Biology*, 190(2):135–146, 1998.

[29] P. G. McQueen and F. E. McKenzie. Age-structured red blood cell susceptibility and the dynamics of malaria infections. *Proceedings of the National Academy of Sciences of the United States of America*, 101(24):9161–9166, 2004.

[30] C. Metcalf, A. Graham, S. Huijben, V. Barclay, G. Long, B. Grenfell, A. Read, and O. Bjørnstad. Partitioning regulatory mechanisms of within-host malaria dynamics using the effective propagation number. *Science*, 333(6045):984–988, 2011.

[31] L. Molineaux, H. Diebner, M. Eichner, W. Collins, G. Jeffery, and K. Dietz. Plasmodium falciparum parasitaemia described by a new mathematical model. *Parasitology*, 122(04):379–391, 2001.

[32] A. Moreno, M. Cabrera-Mora, A. Garcia, J. Orkin, E. Strobert, J. W. Barnwell, and M. R. Galinski. Plasmodium coatneyi in rhesus macaques replicates the multisystemic

dysfunction of severe malaria in humans. *Infection and immunity*, 81(6):1889–1904, 2013.

[33] W. H. Organization et al. World malaria report 2016. *Geneva: WHO. Embargoed until*, 13, 2016.

[34] R. E. Palermo, L. J. Patterson, L. D. Aicher, M. J. Korth, M. Robert-Guroff, and M. G. Katze. Genomic analysis reveals pre-and postchallenge differences in a rhesus macaque aids vaccine trial: insights into mechanisms of vaccine efficacy. *Journal of virology*, 85(2):1099–1116, 2011.

[35] D. J. Perkins, T. Were, G. C. Davenport, P. Kempaiah, J. B. Hittner, and J. M. Ong'echa. Severe malarial anemia: innate immunity and pathogenesis. *International journal of biological sciences*, 7(9):1427, 2011.

[36] D. Prakash, C. Fesel, R. Jain, P.-A. Cazenave, G. C. Mishra, and S. Pied. Clusters of cytokines determine malaria severity in plasmodium falciparum–infected patients from endemic areas of central india. *Journal of Infectious Diseases*, 194(2):198–207, 2006.

[37] I. M. Rouzine and F. E. McKenzie. Link between immune response and parasite synchronization in malaria. *Proceedings of the National Academy of Sciences*, 100(6):3473–3478, 2003.

[38] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

[39] N. J. Savill, W. Chadwick, and S. E. Reece. Quantitative analysis of mechanisms that govern red blood cell age structure and dynamics during anaemia. *PLoS computational biology*, 5(6):e1000416, 2009.

[40] K. Shedden and J. Taylor. Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In *Methods of Microarray Data Analysis*, pages 121–131. Springer, 2005.

[41] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[42] J. Straube, A.-D. Gorse, B. E. Huang, K.-A. Lê Cao, et al. A linear mixed model spline framework for analysing time course omics data. *PloS one*, 10(8):e0134540, 2015.

[43] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[44] S.-I. Tachibana, S. A. Sullivan, S. Kawai, S. Nakamura, H. R. Kim, N. Goto, N. Arisue, N. M. Palacpac, H. Honma, M. Yagi, et al. Plasmodium cynomolgi genome sequences provide insight into plasmodium vivax and the monkey malaria clade. *Nature genetics*, 44(9):1051–1055, 2012.

[45] J. J. Thibodeaux. Modeling erythropoiesis subject to malaria infection. *Mathematical biosciences*, 225(1):59–67, 2010.

[46] K. J. Torres, E. Villasis, J. Bendezú, J. Chauca, J. M. Vinetz, and D. Gamboa. Relationship of regulatory t cells to plasmodium falciparum malaria symptomatology in a hypoendemic region. *Malaria journal*, 13(1):108, 2014.

[47] T. M. Tran, M. B. Jones, A. Ongoiba, E. M. Bijker, R. Schats, P. Venepally, J. Skinner, S. Doumbo, E. Quinten, L. G. Visser, et al. Transcriptomic evidence for modulation of host inflammatory responses during febrile plasmodium falciparum malaria. *Scientific Reports*, 6, 2016.

[48] K. Wang, M. Saito, B. C. Bisikirska, M. J. Alvarez, W. K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. A. Margolin, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human b cells. *Nature biotechnology*, 27(9):829–837, 2009.

[49] M. Warren, J. Skinner, and E. Guinn. Biology of the simian malarias of southeast asia. i. host cell preferences of young trophozoites of four species of plasmodium. *The Journal of parasitology*, pages 14–16, 1966.

[50] A. P. Waters, D. G. Higgins, and T. McCutchan. Evolutionary relatedness of some primate models of plasmodium. *Molecular biology and evolution*, 10(4):914–923, 1993.

[51] D. Wu and G. K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

[52] M. N. Wykes, J. M. Horne-Debets, C.-Y. Leow, and D. S. Karunarathne. Malaria drives t cells to exhaustion. *Frontiers in microbiology*, 5:249, 2014.

[53] J. Yamagishi, A. Natori, M. E. Tolba, A. E. Mongan, C. Sugimoto, T. Katayama, S. Kawashima, W. Makalowski, R. Maeda, Y. Eshita, et al. Interactive transcriptome analysis of malaria patients and infecting plasmodium falciparum. *Genome research*, 24(9):1433–1444, 2014.

[54] J. Ylostalo, A. C. Randall, T. A. Myers, M. Metzger, D. J. Krogstad, and F. B. Cogswell. Transcriptome profiles of host gene expression in a monkey model of human malaria. *Journal of Infectious Diseases*, 191(3):400–409, 2005.

Chapter 2

Mathematical model of susceptibility, resistance, and resilience in the within-host dynamics between a *Plasmodium* parasite and the immune system[1]

[1]Yi H. Yan, Brian Adam, Mary Galinski, Jessica Kissinger, Alberto Moreno and Juan B. Gutierrez. Accepted by *Mathematical Biosciences,* 10/23/2015. Reprinted here with permission of publisher, 6/19/2017

2.1 Abstract

We developed a coupled age-structured partial differential equation model to capture the disease dynamics during blood-stage malaria. The addition of age structure for the parasite population, with respect to previous models, allows us to better characterize the interaction between the malaria parasite and red blood cells during infection. Here we prove that the system we propose is well-posed and there exist at least two global states. We further demonstrate that the numerical simulation of the system coincides with clinically observed outcomes of primary and secondary malaria infection. The well-posedness of this system guarantees that the behavior of the model remains smooth, bounded, and continuously dependent on initial conditions; calibration with clinical data will constrain domains of parameters and variables to physiological ranges.

2.2 Introduction

The goal of this article is to formulate a mathematical model to characterize the within-host dynamics present in the disease malaria, between different host cell types and pathogens of the *Plasmodium* species. *Plasmodium* is a large genus of parasitic protozoa (unicellular eukaryotic organisms), with complex genomes and sophisticated life cycles. The genome, behavior, and epidemiological characterization of *Plasmodium* are orders of magnitude more complex than that of viruses or bacteria. The clinical manifestation of infection by *Plasmodium*, the disease malaria, has a wide spectrum of symptoms, varying from asymptomatic to highly severe. Malaria affects birds, reptiles, and some mammals (mostly rodents and primates).

The *Plasmodium* life cycle is comprised of several stages. The infection process in humans starts with the injection of sporozoites by female Anopheline mosquitoes into the skin of the host. This is followed by the liver stage, in which the inoculated sporozoites grow and multiply asexually within hepatocytes for 1-2 weeks to produce merozoites. The newly produced

merozoites emerge from the liver and enter the blood stream. The blood-stage infection starts immediately after the hepatic stage; merozoites invade red blood cells (RBCs) where they also reproduce asexually. In some instances, the parasites develop into the sexual stage form called gametocytes. Gametocyte-infected red blood cells (iRBCs) infect newly feeding Anopheline mosquitoes and through sexual reproduction followed by many more rounds of asexual multiplication are ultimately transformed into sporozoites, thus completing the infection cycle between an Anopheline mosquito and its host.

The parasite's blood-stage infection in both human and non-human primates generally has a regular cycle of 24 to 72 hours depending on the species of the *Plasmodium* parasite [3, 4]. The parasites invade healthy RBCs and replicate asexually, remodeling and ultimately destroying the RBCs in the process. The destruction of RBCs during blood-stage malaria infection sometimes results in severe anemia, which is one of the major complications of malaria and a leading cause of mortality. Human and non-human primate RBCs have a normal life span of 120 to 100 days, respectively, where afterwards RBCs are cleared rapidly [11]. Both the age structure of RBCs and iRBCs play an important role in the hematodynamics of the host during malaria infection. It has been shown previously that *Plasmodium vivax* infects RBCs of different age groups at different rates and the interaction between the life cycle of the malaria parasite and the immune system can lead to synchrony of the parasite [12, 14]. Anemia caused by malaria infection is a complex process not only involving the malaria parasites and host RBCs, but also the innate and adaptive immune system of the host [7, 13]. It has been previously postulated that there are at least three factors that contribute to severe anemia resulting from blood stage malarial infection: (i) The destruction of RBCs directly by the parasite, (ii) the destruction of RBCs by a non-parasitic factor, possibly the host immune system, and (iii) the disruption of the host's erythropoietic process [6, 11, 13]; erythropoiesis refers to the differentiation and maturation of hematopoietic stem cells (HSCs) through committed cell lineages culminating in the production of erythrocytes/RBCs.

Classic models of erythropoiesis rest on three variables representing the precursor/progenitor population of immature erythrocytes, the population of mature RBCs, and the concentration of erythropoietin (Epo), a signaling molecule critical to erythrocyte production. Mackey and Milton (1990) represented these three variables with a system of ordinary differential equations (ODEs) with one delay accounting for the time for Epo dependent maturation of progenitors into erythrocytes (about 6 days) [8]. In 1995, Belair *et al.* developed an "age-structured" partial differential equation (PDE) model for erythropoiesis [2]. Belair *et al.*'s model characterizes the proliferation and aging of the precursor and mature erythrocyte populations as nonlinear, first order PDEs and includes an ODE to represent Epo dynamics over time. Building on Belair's 1995 model, Mahaffy *et al.* (1998) investigated the addition of state-dependent delays to the model [9].

Due to the difficulty involved in measuring Epo accurately over an extended period of time, we formulated the rate at which mature RBCs enter circulation as only dependent upon the current number of RBCs and the number of RBCs at equilibrium. Such a formulation has previously been shown by Savil *et al.* to capture the erythropoiesis process using data derived from a murine model [15]. As the only formulation of an erythropoiesis model without an Epo component that has been validated by time course data, we believe that such a formulation would also be able to capture the fluctuation in the erythropoietic process due to loss of RBCs within our hematodynamic system.

Most recent and most relevant to our study, Thibodeaux *et al.* published two papers (2010, 2013) modifying Belair, Mackey, and Mahaffy's 1995 model in an attempt to simulate and analyze erythropoietic dynamics subject to malaria infection [1, 16]. They conjoined an ODE-based model of malaria infection with an age-structured PDE model of erythropoiesis, and examined the dynamics of this new system to simulate the effects of hemozoin (Hz) on the suppression of erythropoiesis.

In order to capture the complex interactions between different age groups of iRBCs, RBCs and various immune cells during blood stage malaria [10], we propose a coupled

age-structured PDE model of both iRBCs and host RBCs. Our model borrows from the erythropoiesis PDE model [1, 16] and expands upon it by adding the parasite age structure system and the effect of immune cells on both RBCs and iRBCs.

This paper is organized as follows: Section 2 contains the detailed derivation and description of our hematodynamic model. Section 3 proves the well posedness of the system and the existence of at least two global behaviors of biological interest. Section 4 provides numerical simulation results of our model. Section 5 offers some conclusions, and presents future directions.

2.3 Model Formulation

Let $u(a,t)$ be a function which approximates the concentration of RBCs of age a at time point t. RBCs of age $> a_{max}$ are rapidly cleared from the circulation, thus we assume that $u(a,t) = 0$ when $a > a_{max}$. Additionally, Let $v(\alpha,t)$ be a function that approximates the concentration of iRBCs of age α at time t. iRBCs of age α_{max} burst to produce merozoites, which in turn infect other RBCs, thus we assume that $v(\alpha,t) = 0$ when $\alpha > \alpha_{max}$. Assuming that the loss of RBCs during blood stage malarial infection is only due to parasite invasion, innate immune cells, adaptive immune cells, random loss of RBCs during aging and the rapid clearance of RBCs with age $> a_{max}$, and the destruction of RBCs by innate and adaptive immune cells follows the law of mass action, then the difference between the concentrations

of RBCs across all ages between two time points, t_1 and t_2 is

$$\begin{aligned}\int_{a_0}^{a_{max}} u(a,t_2)\,da \;&= \int_{a_0}^{a_{max}} u(a,t_1)\,da \;+ \int_{t_1}^{t_2} u(a_0,t)\,dt \;- \int_{t_1}^{t_2} u(a_{max},t)\,dt \\ &- \gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u(a,t))\,dt\,da \\ &- \sum_{i=1}^{T} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} w_i(t)\theta_i u(a,t)\,dt\,da \\ &- \sum_{i=1}^{Q} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} s_i(t)\psi_i u(a,t)\,dt\,da \\ &- \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} h(a) u(a,t)\,dt\,da\ , \end{aligned} \tag{2.1}$$

where

$$p(u(a,t)) = \frac{u(a,t)}{\int_{a_0}^{a_{max}} u(a,t)}$$

is a probability density function such that $\int_{a_0}^{a_{max}} p(u(a,t))da = 1$.

The left hand side (LHS) term of (2.1) and the first three terms of the right hand side (RHS) of (2.1),

$$\int_{a_0}^{a_{max}} u(a,t_2)\,da \;= \int_{a_0}^{a_{max}} u(a,t_1)\,da \;+ \int_{t_1}^{t_2} u(a_0,t)\,dt \;- \int_{t_1}^{t_2} u(a_{max},t)\,dt,$$

account for the change of the concentration of RBCs from t_1 to t_2 due to the production of new RBCs and the rapid clearance of RBCs with age $> a_{max}$.

The term

$$-\gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u(a,t))\,dt\,da\ ,$$

accounts for the loss of RBCs due to infection from bursting iRBCs. $v(\alpha_{max},t)$ is the concentration of bursting iRBCs at time t. $r(a)$ is the success rate of merozoites infecting RBCs of age a. γ is the expected number of merozoites each bursting iRBC produces. $p(u(a,t))$ is the expected percentage of merozoites infecting RBCs of specific age a at time t. The concentration of new iRBCs produced at time t is

$$\gamma \int_{a_0}^{a_{max}} v(\alpha_{max},t) r(a) p(u(a,t))\,da. \tag{2.2}$$

The term

$$-\sum_{i=1}^{T}\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} w_i(t)\theta_i u(a,t)\,dt\,da,$$

accounts for the loss of RBCs due to T different kinds of innate immune cells. $w_i(t)$ is the concentration of the ith kind of innate immune cells at time t. θ_i represents the ith kind of innate immune cell's effectiveness at destroying RBCs. The term

$$\sum_{i=1}^{Q}\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} s_i(t)\psi_i u(a,t)\,dt\,da,$$

accounts for the loss of RBCs due to Q different kinds of adaptive immune cells. $s_i(t)$ is the concentration of the ith kind of adaptive immune cells at time t and ψ_i represents the ith kind of adaptive immune cell's effectiveness at destroying RBCs.

The last term

$$\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} h(a)u(a,t)\,dt\,da\ ,$$

accounts for the random loss of RBCs during aging. $h(a)$ is the natural death rate of RBCs of different ages.

Using the fundamental theorem of Calculus, it is clear that:

$$\int_{a_0}^{a_{max}} u(a,t_2)\,da\ -\int_{a_0}^{a_{max}} u(a,t_1)\,da\ =\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} u_t\,dt\,da, \tag{2.3}$$

and

$$\int_{t_1}^{t_2} u(a_0,t)\,dt-\int_{t_1}^{t_2} u(a_{max},t)\,dt=-\int_{t_1}^{t_2}\int_{a_0}^{a_{max}} u_a\,da\,dt. \tag{2.4}$$

Substituting (2.3) and (2.4) into (2.1), we obtain:

$$\begin{aligned}
&\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} u_t\,dt\,da\ +\int_{t_1}^{t_2}\int_{a_0}^{a_{max}} u_a\,da\,dt\ =\\
&-\gamma\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} v(\alpha_{max},t)r(a)p(u(a,t))\,dt\,da\\
&-\sum_{i=1}^{T}\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} w_i(t)\theta_i u(a,t)\,dt\,da\\
&-\sum_{i=1}^{Q}\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} s_i(t)\psi_i u(a,t)\,dt\,da\\
&-\int_{a_0}^{a_{max}}\int_{t_1}^{t_2} h(a)u(a,t)\,dt\,da,
\end{aligned} \tag{2.5}$$

which can be rearrange to the following equation because all the terms have the same integral:

$$\begin{aligned}
&\int_{a_0}^{a_{max}} \int_{t_1}^{t_2} (u_t + u_a)\ da\, dt\ = \\
&-\gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max}, t) r(a) p(u(a,t)) dt\, da \\
&-\int_{a_0}^{a_{max}} \int_{t_1}^{t_2} \left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a) \right) u(a,t) dt\, da. \qquad (2.6)
\end{aligned}$$

If the function $u(a,t)$ and its partial derivatives are continuous functions then the integrands on both sides of (2.6) should be equal, which leads us to the following partial differential equation:

$$\begin{aligned}
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = &- \left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a) \right) u(a,t) \\
&- \gamma v(\alpha_{max}, t) r(a) p(u(a,t)). \qquad (2.7)
\end{aligned}$$

(2.7) is subject to the boundary condition:

$$u(0,t) = f(t, \varphi(t)), \qquad (2.8)$$

where

$$f(t,\varphi(t)) = \begin{cases} \varsigma,\ t < T_d \\ \varsigma e^{\varepsilon(\varphi_0 - \varphi(t))},\ t > T_d. \end{cases}$$

This boundary condition reflects the erythropoietic response to the change in concentration of RBCs. $\varphi(t)$ is the concentration of RBCs at time point t, ς is the normal rate at which RBCs enter the peripheral blood, T_d reflects a one time lag in the erythropoietic response to the change in concentration of RBCs, as the erythropoietic process only responds to the infection after a time period of T_d. φ_0 is the normal concentration of RBCs in a healthy host and ε is a parameter that controls the maximum amount of erythropoietic response. During malarial infection, as the concentration of RBCs decreases due to malaria parasites, the rate of new RBCs entering into circulation increases.

The age structured model of iRBCs can be similarly derived. Let $v(\alpha, t)$ be a function that approximates the concentration of iRBCs of age α at time t. iRBCs of age α_{max} bursts

to produce merozoites, which in turn infect other RBCs, thus we assume that $v(\alpha, t) = 0$ when $\alpha > \alpha_{max}$. Assuming the loss of iRBCs is only due to the bursting of iRBCs of age α_{max}, innate immune cells and adaptive immune cells, then the difference between the concentration of iRBCs across all ages between two time points, t_1 and t_2 is

$$\begin{aligned}\int_{\alpha_0}^{\alpha_{max}} v(\alpha, t_2)\, d\alpha \; &= \int_{\alpha_0}^{\alpha_{max}} v(\alpha, t_1)\, d\alpha \; + \int_{t_1}^{t_2} v(\alpha_0, t)\, dt \; - \int_{t_1}^{t_2} v(\alpha_{max}, t)\, dt \\ &\quad - \sum_{i=1}^{T} \int_{\alpha_0}^{\alpha_{max}} \int_{t_1}^{t_2} w_i(t)\phi_i v(\alpha, t)\, dt\, d\alpha \\ &\quad - \sum_{i=1}^{Q} \int_{\alpha_0}^{\alpha_{max}} \int_{t_1}^{t_2} s_i(t) b_i(t) v(\alpha, t)\, dt\, d\alpha\end{aligned} \tag{2.9}$$

Using the same technique we used to derive (2.7), we arrive at the following PDE for $v(\alpha, t)$:

$$\frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} = -\left(\sum_{i=1}^{T} w_i(t)\phi_i + \sum_{i=1}^{Q} s_i(t) b_i(t)\right) v(\alpha, t), \tag{2.10}$$

where V is the speed at which the parasite ages, which is subject to the unit of α and the term

$$b_i(t) = \frac{\nu_i}{1 + \exp\left(-\lambda_i(\int_{t_0}^{t} v(0, t)\, dt - \xi_i)\right)},$$

describes the adaptive immune cell i's effector strength against iRBCs. The function $b_i(t)$ captures the ability of adaptive immune cells to increase their effect against iRBCs through increased exposure, which increases, as the exposure to new iRBCs, characterized by $\int_{t_0}^{t} v(0, t)$ increases past threshold ξ_i. The effector strength has a maximum value of ν_i as the amount of exposure to iRBCs becomes substantially large.

The term

$$-\left(\sum_{i=1}^{T} w_i(t)\phi_i\right) v(\alpha, t),$$

describes the destruction of iRBC by different innate immune cells each with constant strength ϕ_i. The term

$$-\left(\sum_{i=1}^{Q} s_i(t)b_i(t)\right)v(\alpha,t),$$

describes the destruction of iRBCs by different adaptive immune cells each with strength $b_i(t)$ that changes according to the total exposure to iRBCs. Additionally, this PDE is subject to the boundary condition:

$$v(\alpha_0,t) = \gamma\int_{a_0}^{a_{max}} v(\alpha_{max},t)r(a)p(u(a,t))da,$$

which describes the creation of new iRBCs. Because the life span of an iRBC is between 24 and 72 hours, and the infection process is within minutes, we assume that such process is instantaneous within our model.

Additionally, due to the presence of iRBCs, the following equation is proposed to describe the change in concentration of innate immune cells:

$$\frac{dw_i}{dt} = o_i(t) - \beta_i w_i(t) - \tau_i w_i\int_{\alpha_0}^{\alpha_{max}} v(\alpha,t)d\alpha, \tag{2.11}$$

and

$$o_i(t) = \varpi_i + \frac{\epsilon_i - \varpi_i}{1+\exp\left(-\eta_i(\int_{\alpha_0}^{\alpha_{max}} v(\alpha,t)d\alpha - M_i)\right)}. \tag{2.12}$$

We assume that each innate immune cell population $w(t)_i$ decays at a rate of β_i. The production of innate immune cells increases as the concentration of iRBCs increases past a threshold M_i and reaches a maximum level of $\epsilon_i - \varpi_i$. Additionally, the term τ_i describes the loss of functionality of innate immune cells upon contact with iRBCs. (2.12) describes the change in the rate of production of innate immune cells in response to the change in concentration of iRBCs.

The change in adaptive immune cell population in response to increased concentration of iRBCs is similarly formulated as the following:

$$\frac{ds_i}{dt} = l_i(t) - \delta_i s_i(t) - \vartheta_i s_i\int_{\alpha_0}^{\alpha_{max}} v(\alpha,t)d\alpha, \tag{2.13}$$

and

$$l_i(t) = \sigma_i + \frac{\varrho_i - \sigma_i}{1 + \exp\left(-\omega_i(\int_{\alpha_0}^{\alpha_{max}} v(\alpha, t)d\alpha - R_i)\right)}. \tag{2.14}$$

We assume that each adaptive immune cell population $s(t)_i$ decays at a rate of δ_i. The production of adaptive immune cells increases as the concentration of iRBCs increases past a threshold R_i and reaches a maximum level $\varrho_i - \sigma_i$. Additionally, the term ϑ_i describes the loss of functionality of adaptive immune cells upon contact with iRBCs. (2.14) describes the change in the the rate of production of adaptive immune cells in response to the change in concentration of iRBCs.

Putting (2.7), (2.8), (2.10), (2.11), (2.11), and (2.13) together we arrive at the following PDE system:

$$\begin{aligned}
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} =& -\left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a)\right) u(a,t) \\
& - \gamma v(\alpha_{max}, t) r(a) p(u(a,t)), \\
\frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} =& -\left(\sum_{i=1}^{T} w_i(t)\phi_i + \sum_{i=1}^{Q} s_i(t) b_i(t)\right) v(\alpha, t), \\
\frac{dw_i}{dt} =& o_i(t) - \beta_i w_i(t) - \tau_i w_i(t) \int_{\alpha_0}^{\alpha_{max}} v(\alpha, t)d\alpha, \\
\frac{ds_i}{dt} =& l_i(t) - \delta_i s_i(t) - \vartheta_i s_i(t) \int_{\alpha_0}^{\alpha_{max}} v(\alpha, t)d\alpha,
\end{aligned} \tag{2.15}$$

subject to the following initial and boundary conditions:

$$\begin{aligned}
u(a,0) =& g(a), \\
u(0,t) =& f(t, \varphi(t)), \\
v(\alpha,0) =& c(\alpha), \\
v(0,t) =& \gamma \int_{a_0}^{a_{max}} v(\alpha_{max}, t) r(a) p(u(a,t))\, da,
\end{aligned}$$

where $g(a)$ is the initial RBC age distribution and $c(\alpha)$ is the initial iRBC age distribution.

Since only *Plasmodium vivax* has been shown to infect RBCs of different ages at different rates [12], we consider $r(a)$ to be a constant κ within our model where $\kappa \in [0, 1]$. Furthermore,

due to a lack of data regarding the loss of innate and adaptive immune function upon interaction with iRBC, we consider ϑ_i and τ_i to be 0 within our model. Additionally, according to the erythropoiesis model developed by Savil *et al* [15], $h(a)$ was also set to be a constant. Lastly, we assume that the total number of iRBCs is much smaller than the total number of RBCs. Therefore, the crowding effect on the production of iRBCs is negligible. This assumption is justified if the host becomes resilient or completely removes the infection. The simplified model is:

$$\begin{aligned}
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} =& -\left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a)\right) u(a,t) \\
& - \gamma\kappa v(\alpha_{max}, t) p(u(a,t)), \\
\frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} =& -\left(\sum_{i=1}^{T} w_i(t)\phi_i + \sum_{i=1}^{Q} s_i(t) b_i(t)\right) v(\alpha, t), \\
\frac{dw_i}{dt} =& o_i(t) - \beta_i w_i(t), \\
\frac{ds_i}{dt} =& l_i(t) - \delta_i s_i(t),
\end{aligned} \tag{2.16}$$

subject to the following initial and boundary conditions:

$$\begin{aligned}
u(a,0) =& g(a), \\
u(0,t) =& f(t, \varphi(t)), \\
v(\alpha,0) =& c(\alpha), \\
v(0,t) =& \gamma\kappa \int_{a_0}^{a_{max}} v(\alpha_{max}, t) p(u(a,t))\, da.
\end{aligned} \tag{2.17}$$

From this point on we will be working with the above simplified version of our model.

2.4 System Behavior

2.4.1 Immune Saturation

Given an large parasite exposure, the adaptive immune strength $b_i(t)$ becomes constant:

$$\begin{aligned}\lim_{\int_{t_0}^{t} v(0,t)\,dt\to+\infty} b_i(t) =& \lim_{\int_{t_0}^{t} v(0,t)\,dt\to+\infty} \frac{\nu_i}{1+\exp\left(-\lambda_i(\int_{t_0}^{t} v(0,t)\,dt-\xi_i)\right)} \\ =&\nu_i.\end{aligned} \tag{2.18}$$

Additionally, the production rate of both innate and adaptive immune cells also reaches a constant value given a large iRBC population in the system:

$$\begin{aligned}\lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} o_i(t) =& \lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} \varpi_i \\ &+ \frac{\epsilon_i-\varpi_i}{1+\exp\left(-\eta_i(\int_{\alpha_0}^{\alpha_{max}} v(\alpha,t)d\alpha-M_i)\right)}, \\ =&\epsilon_i,\end{aligned}$$

and

$$\begin{aligned}\lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} l_i(t) =& \lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} \sigma_i \\ &+ \frac{\varrho_i-\sigma_i}{1+\exp\left(-\omega_i(\int_{\alpha_0}^{\alpha_{max}} v(\alpha,t)d\alpha-R_i)\right)}, \\ =&\varrho_i.\end{aligned}$$

Because the production rate of both the innate and adaptive immune cells becomes constant given a large parasite population, the population of both the innate and adaptive immune cells also become constant:

$$\lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} w_i(t) = \frac{\epsilon_i}{\beta_i} \tag{2.19}$$

and

$$\lim_{\int_{\alpha_0}^{\alpha_{max}} v(a,t)d\alpha\to+\infty} s_i(t) = \frac{\varrho_i}{\delta_i}. \tag{2.20}$$

Assuming immune saturation by substituting (2.19) and (2.20),(2.7) becomes:

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial a}=-\left(\sum_{i=1}^{T}\left(\frac{\epsilon_i}{\beta_i}\right)\theta_i+\sum_{i=1}^{Q}\left(\frac{\varrho_i}{\delta_i}\right)\psi_i+h(a)\right)u(a,t)$$
$$-\gamma\kappa v(\alpha_{max},t)p(u(a,t)),$$

where $h(a)$ is a constant decay term describing the random loss of RBCs over time and ϵ_i, β_i, θ_i, ϱ_i, δ_i, ψ_i are all constants. The system further simplifies to:

$$\frac{\partial u}{\partial t}+\frac{\partial u}{\partial a}=C_2u(a,t)-\gamma\kappa v(\alpha_{max},t)p(u(a,t)). \tag{2.21}$$

Additionally, assuming immune saturation by substituting (2.19),(2.20) and (2.18), (2.9) becomes

$$\frac{\partial v}{\partial t}+V\frac{\partial v}{\partial \alpha}=-\left(\sum_{i=1}^{T}\left(\frac{\epsilon_i}{\beta_i}\right)\phi_i+\sum_{i=1}^{Q}\left(\frac{\varrho_i}{\delta_i}\right)\nu_i\right)v(\alpha,t), \tag{2.22}$$

where ϵ_i, β_i, ϕ_i, ϱ_i, δ_i and ν_i are all positive constants. (2.22) further simplifies to:

$$\frac{\partial v}{\partial t}+V\frac{\partial v}{\partial \alpha}=C_1v(\alpha,t). \tag{2.23}$$

Corollary 2.4.1. *There exist an unique solution for $v(\alpha,t)$ under immune saturation conditions.*

Proof. Using the fact that α is a function of time, $v(\alpha,t)$ can be expressed as $v(\alpha(t),t)$, which transforms (2.23) into the following Equation:

$$\frac{d}{dt}v=\frac{\partial v}{\partial t}+\frac{d\alpha}{dt}\frac{\partial v}{\partial \alpha}=C_1v,$$

which reduces (2.23) to the following ODE system:

$$\begin{cases}\frac{d\alpha}{dt}=V,\\ \\ \frac{dv}{dt}=C_1v,\end{cases}$$

which states that along the curves given by $\frac{d\alpha}{dt}=V$, $\frac{dv}{dt}=C_1v$. Solving the above system, we arrive at:

$$\begin{cases}\hat{\alpha}(t)=Vt+a_0,\\ \\ \hat{v}(t)=v(\hat{t}_0)\exp(C_1t),\\ \\ \hat{v}(t_0)=v(a_0,0).\end{cases} \tag{2.24}$$

However, due to the fact that our system is bounded by $t \in [0, \infty]$ and $\alpha \in [0, \alpha_{max}]$, a_0 does not exist when $t > \frac{\hat{\alpha}(t)}{V}$. But the boundary condition of $v(0,t)$ allows us to express (2.24) as the following

$$\begin{cases} t_0 = t - \frac{\alpha(t)}{V}, \\ \hat{v}(t) = v(0, t_0) \exp(C_1(t - t_0)), \end{cases} \tag{2.25}$$

which essentially traces the characteristic curve to the boundary $v(0, t_0)$ if $t > \frac{\alpha(t)}{V}$. Consider that the solution to $v(\alpha, t)$ is a system of ODEs with the form $\frac{dv}{dt} = -Cv$, which is Lipschitz continuous. Thus using the Picard-Lindelöf theorem, each ODEs within the system have an unique solution, therefore, there exist an unique solution to the PDE $v(\alpha, t)$.

□

Corollary 2.4.2. *If $C_1 = \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$, then the boundary $v(0,t)$ exhibits periodic behavior with a period of $\frac{\alpha_{max}}{V}$. If $C_1 < \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$, $\lim_{t\to+\infty} v(0,t) = \infty$. If $C_1 > \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$, $\lim_{t\to+\infty} v(0,t) = 0$.*

Proof. According to Corollary 2.4.1, The characteristic curve for $v(\alpha, t)$ intersects the boundary $v(0,t)$ when $(t - t_0) = \frac{\alpha_{max}}{V}$, which leads to the following expression:

$$v(\alpha_{max}, t) = v(0, t - (\frac{\alpha_{max}}{V})) \exp C_1(\frac{\alpha_{max}}{V}), \tag{2.26}$$

which describes the constant decay of v along the characteristic curve. Using the simplified model where $r(a) = \kappa$ then (2.2) simplifies to:

$$v(0,t) = \gamma\kappa v(\alpha_{max}, t), \tag{2.27}$$

if

$$C_1 = \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}. \tag{2.28}$$

Substituting (2.28) into (2.26) we get:

$$v(\alpha_{max}, t) = \frac{1}{\gamma\kappa} v(0, t - (\frac{\alpha_{max}}{V})), \tag{2.29}$$

Substituting (2.27) into (2.29) we finally arrive at:

$$v(0,t) = v(0, t - (\frac{\alpha_{max}}{V})). \tag{2.30}$$

Meaning that $v(\alpha, t)$ repeat itself with a period of $\frac{\alpha_{max}}{V}$. Consider that $v(\alpha_{max}, t) = v(0, t - (\frac{\alpha_{max}}{V})) \exp\left(-C_1(\frac{\alpha_{max}}{V})\right)$, if $v(0, t)$ is repeating as shown in (2.30) then $v(\alpha_{max}, t)$ is also periodic. Therefore, if $C_1 = \ln \frac{1}{\gamma\kappa} / \frac{\alpha_{max}}{V}$ then $v(\alpha, t)$ is bounded. If

$$C_1 < \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V},$$

then

$$v(\alpha_{max}, t) > \frac{1}{\gamma\kappa} v(0, t - (\frac{\alpha_{max}}{V})),$$

and if

$$C_1 > \frac{\ln 1/\gamma\kappa}{\alpha_{max}/V},$$

then

$$v(\alpha_{max}, t) < \frac{1}{\gamma\kappa} v(0, t - (\frac{\alpha_{max}}{V})).$$

□

Corollary 2.4.3. *$u(a, t)$ can be reduced to a system of ODEs using the method of characteristics.*

Proof. Using the fact that a is a function of time, $u(a, t)$ can be expressed as $u(a(t), t)$ which transforms (2.21) into:

$$\frac{d}{dt}u = \frac{\partial u}{\partial t} + \frac{da}{dt}\frac{\partial u}{\partial a} = C_2 u - \gamma\kappa v(\alpha_{max}, t) p(u(a, t)),$$

reducing the PDE (2.21) to an ODE system:

$$\begin{cases} \frac{da}{dt} = 1, \\ \frac{d\hat{u}}{dt} = C_2\hat{u} - \gamma\kappa v(\alpha_{max}, t) p(\hat{u}, t). \end{cases} \tag{2.31}$$

□

Corollary 2.4.4. *$u(0, t)$ is bounded provided $0 < \inf_t u(a, t)$ for all t and a.*

Proof. If $0 < \inf_t u(a,t)$, then $\int_0^{a_{max}} u(a,t)da > 0$. Which leads to:

$$\sup_t u(0,t) \leq \varsigma e^{\varepsilon\varphi_0}.$$

and

$$\inf_t u(0,t) \geq \varsigma e^{\varepsilon(\varphi_0 - a_{max}\varsigma e^{\varepsilon\varphi_0})} > 0$$

Therefore $u(0,t) \in (0, \varsigma e^{\varepsilon\varphi_0}]$. □

Corollary 2.4.5. *There exists a set of parameters and initial conditions such that* $0 < \inf_{a,t} u(a,t) < \inf_t u(0,t)$.

Proof. Let $0 < N \leq \inf_{a,t} u(a,t) < \inf_t u(0,t)$. Equation (2.31) shows that $u(a,t)$ is monotonically decreasing along the characteristic curves, if there exists a set of parameters and initial conditions such that $0 < \inf_{a,t} u(a,t) < \inf_t u(0,t)$, then the following inequality will hold:

$$\inf\Big(\hat{u}(0) + C_2\int_0^{a_{max}} \hat{u}(t)dt - \gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\Big) \geq N,$$

where $p(\hat{u},t) = \frac{\hat{u}(t)}{\int_0^{a_{max}} u(a,t)da}$ and $C_2 < 0$. Assume $u(a,t) \geq N > 0$:

$$\inf C_2\int_0^{a_{max}} \hat{u}(t)dt \geq C_2 a_{max}\hat{u}(0).$$

Let $|C_2| < \frac{1}{a_{max}}$, then $\hat{u}(0) + C_2 a_{max}\hat{u}(0) = (1 + C_2 a_{max})\hat{u}(0) > 0$.

$$\inf\Big((1 + C_2 a_{max})\hat{u}(0) - \gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\Big) \geq N$$

Let $\hat{u}(0) = \inf \hat{u}(0)$ and $0 < N < (1 + C_2 a_{max}) \inf \hat{u}(0) < \inf \hat{u}(0)$, then:

$$\inf\Big(-\gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\Big) \geq N - (1 + C_2 a_{max}) \inf \hat{u}(0)$$

Because both sides of the inequality are negative, we can rearrange the inequality in the following form:

$$\inf\Big(\gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\Big) \leq (1 + C_2 a_{max}) \inf \hat{u}(0) - N \tag{2.32}$$

Because $\sup_t p(\hat{u},t) < \frac{\sup_t \hat{u}(0)}{a_{max}N} = M$ and let $v(\alpha_{max},t)$ either be oscillating or decreasing as shown in Corollary 2.4.2, then

$$\sup\left(\gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\right) \leq \gamma\kappa a_{max} \sup_{t_0\leq t\leq t_0+a_{max}} v(\alpha_{max},t)M$$

Let $\sup_{t_0\leq t\leq t_0+a_{max}} v(\alpha_{max},t) = V_{max}$, then

$$\sup\left(\gamma\kappa\int_{t_0}^{t_0+a_{max}} v(\alpha_{max},t)p(\hat{u},t)dt\right) \leq \gamma\kappa a_{max}V_{max}M.$$

Divide (2.32) by $\gamma\kappa a_{max}$ M:

$$V_{max} \leq \frac{(1+C_2a_{max})\inf\hat{u}(0) - N}{\gamma\kappa a_{max}M}$$

Therefore, if

$$V_{max} \leq \frac{(1+C_2a_{max})\inf\hat{u}(0) - N}{\gamma\kappa a_{max}M},$$
$$|C_2| < \frac{1}{a_{max}}$$

then

$$0 < N < \inf_{a,t}(u(a,t)) < \inf_t(u(0,t))$$

. □

Corollary 2.4.6. *There exist an unique solution of $u(a,t)$.*

Proof. As stated in Corollary 2.4.3 the solution of $u(a,t)$ is a system of ODEs with the form:

$$\frac{d\hat{u}}{dt} = C_2\hat{u} - \gamma\kappa v(\alpha_{max},t)p(\hat{u},t) = F(\hat{u},t).$$

Let $\hat{u}_1$ and $\hat{u}_2$ be different, then:

$$|F(\hat{u}_1,t) - F(\hat{u}_2,t)| = |C_2(\hat{u}_1-\hat{u}_2) - \gamma\kappa v(\alpha_{max},t)(p(\hat{u}_1,t)-p(\hat{u}_2,t))|. \tag{2.33}$$

Because $p(\hat{u}_n,t) = \frac{\hat{u}_n}{E_n(t)}$, where $E_n(t) = \int_{a_0}^{a_{max}} u_n(a,t)da$.(2.33) further simplifies to:

$$|F(\hat{u}_1,t) - F(\hat{u}_2,t)| = \left|C_2(\hat{u}_1-\hat{u}_2) - \gamma\kappa v(\alpha_{max},t)\Big(\frac{\hat{u}_1}{E_1(t)} - \frac{\hat{u}_2}{E_2(t)}\Big)\right|.$$

As a result, it follows that:

$$
\begin{aligned}
|F(\hat{u}_1,t)-F(\hat{u}_2,t)| &= \left|C_2(\hat{u}_1-\hat{u}_2)-\gamma\kappa v(\alpha_{max},t)\frac{\hat{u}_1E_2(t)-\hat{u}_2E_1(t)}{E_1(t)E_2(t)}\right| \\
&\leq |C_2||\hat{u}_1-\hat{u}_2|+\left|\gamma\kappa v(\alpha_{max},t)\frac{\hat{u}_1E_2(t)-\hat{u}_2E_1(t)}{E_1(t)E_2(t)}\right| \\
&\leq |C_2||\hat{u}_1-\hat{u}_2|+|\gamma\kappa v(\alpha_{max},t)|\left|\frac{\hat{u}_1E_2(t)-\hat{u}_2E_1(t)}{E_1(t)E_2(t)}\right|
\end{aligned} \tag{2.34}
$$

In the previous Corollary, we have proved that under suitable initial conditions and parameters, $0 < N < \inf_t u(a,t) < u(a,t) < \sup_t u(a,t)$. Which implies that $E_n(t)$ is bounded. Let $A = \inf_t E_n(t)$ and $B = \sup_t E_n(t)$. Without loss of generality. assume $E_1(t) \leq E_2(t)$. It follows that (2.34) is bounded above by:

$$
\begin{aligned}
&\sup_t |C_2||\hat{u}_1-\hat{u}_2|+\sup|\gamma\kappa v(\alpha_{max},t)|\frac{\sup_t E_2(t)}{\inf_t E_1(t)\inf_t E_2(t)}|\hat{u}_1-\hat{u}_2| \\
&\leq |C_2||\hat{u}_1-\hat{u}_2|+\sup|\gamma\kappa v(\alpha_{max},t)|\frac{B}{A^2}|\hat{u}_1-\hat{u}_2| \\
&= \Big(|C_2|+\sup|\gamma\kappa v(\alpha_{max},t)|\frac{B}{A^2}\Big)|\hat{u}_1-\hat{u}_2|.
\end{aligned}
$$

If we define:

$$\Omega = \Big(|C_2|+\sup|\gamma\kappa v(\alpha_{max},t)|\frac{B}{A^2}\Big),$$

then we have established that:

$$|F(\hat{u}_1,t)-F(\hat{u}_2,t)| \leq \Omega|\hat{u}_1-\hat{u}_2|.$$

We have demonstrated that each of the infinite number of ODEs composing the solution of $u(a,t)$ is Lipschitz continuous. Applying the Picard-Lindelöf theorem, we arrive at the conclusion that there exist an unique solution for each of the ODEs composing the solution of $u(a,t)$ and therefore there exist an unique solution of $u(a,t)$. □

Corollary 2.4.7. *The unique solution for $v(\alpha,t)$ is stable for a set of parameters and initial conditions.*

Proof. Assume that there are two solutions of the system $v_1(\alpha,t)$ and $v_2(\alpha,t)$ with different boundary conditions:$v_1(0,t)$,$v_1(\alpha,0)$,$v_2(0,t)$ and $v_2(\alpha,0)$. Recall Corollary 2.4.1 which states

that the solution of $v(\alpha, t)$ is a system of ODEs with the following form:

$$\begin{cases} \hat{\alpha}(t) = Vt + a_0 \\ \hat{v}(t) = v(\hat{t_0}) \exp(C_1 t) \\ \hat{v}(t_0) = v(a_0, 0). \end{cases}$$

Using (2.24), it is clear that, $v_1(\alpha, t) = v_1(t - Vt, 0) \exp(C_1 t)$ on its characteristic curves and $v_2(\alpha, t) = v_2(t - Vt, 0) \exp(C_1 t)$ on its characteristic curves. Therefore, the difference between $v_1(\alpha, t)$ and $v_2(\alpha, t)$ at any given α and t can be expressed as the following if $t < \alpha_{max}$:

$$|v_1(\alpha, t) - v_2(\alpha, t)| = |(v_1(t - Vt, 0) - v_2(t - Vt, 0)) \exp(C_1 t)| . \tag{2.35}$$

If $t > \alpha_{max}$, using (2.25), the difference between $v_1(\alpha, t)$ and $v_2(\alpha, t)$ at any given α and t can be expressed as the following:

$$|v_1(\alpha, t) - v_2(\alpha, t)| = |(v_1(0, t_0) - v_2(0, t_0)) \exp(C_1(t - t0))| , \tag{2.36}$$

where

$$t_0 = t - \frac{\alpha}{V}.$$

Recall Corollary 2.4.2, which states that there exist a set of parameters such that $v(\alpha, t)$ is bounded. Therefore, according to (2.35) and (2.36) it is clear that the solution of $v(\alpha, t)$ depends continuously on its initial and boundary conditions, thus $v(\alpha, t)$ is stable. □

Corollary 2.4.8. *The unique solution for $u(a, t)$ is stable.*

Proof. Assume that there are two solutions of the system $u_1(a, t)$ and $u_2(a, t)$ with different boundary conditions: $u_1(0, t)$, $u_1(a, 0)$, $u_2(0, t)$ and $u_2(a, 0)$. Then the difference between the total numbers of RBCs in $u_1(a, t)$ and $u_2(a, t)$ at an arbitrary time point t can be expressed

in the following form as demonstrated in (2.1):

$$\begin{aligned}\int_{a_0}^{a_{max}} u_1(a,t_2) - u_2(a,t_2)\,da \;&= \int_{a_0}^{a_{max}} u_1(a,t_1) - u_2(a,t_1)\,da \\ &+ \int_{t_1}^{t_2} u_1(a_0,t) - u_2(a_0,t)\,dt \\ &- \int_{t_1}^{t_2} u_1(a_{max},t) - u_2(a_{max},t)\,dt \\ &- \gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u_1(a,t))\,dt\,da \\ &+ \gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u_2(a,t))\,dt\,da \\ &- \sum_{i=1}^{T} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} w_i(t)\theta_i u_1(a,t) - u_2(a,t)\,dt\,da \\ &- \sum_{i=1}^{Q} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} s_i(t)\psi_i u(a,t) - u_2(a,t)\,dt\,da \\ &- \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} h(a) u_1(a,t) - u_2(a,t)\,dt\,da\ . \end{aligned} \tag{2.37}$$

Let $W = u_1(a,t) - u_2(a,t)$, then (2.37) simplifies to:

$$\begin{aligned}\int_{a_0}^{a_{max}} W(a,t_2)\,da \;&= \int_{a_0}^{a_{max}} W(a,t_1)\,da \;+ \int_{t_1}^{t_2} W(a_0,t)\,dt \\ &- \int_{t_1}^{t_2} W(a_{max},t)\,dt \\ &- \gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u_1(a,t))\,dt\,da \\ &+ \gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u_2(a,t))\,dt\,da \\ &- \sum_{i=1}^{T} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} w_i(t)\theta_i W(a,t)\,dt\,da \\ &- \sum_{i=1}^{Q} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} s_i(t)\psi_i W(a,t)\,dt\,da \\ &- \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} h(a) W(a,t)\,dt\,da\ . \end{aligned} \tag{2.38}$$

Assuming $r(a)$ is a constant,then the following is true:

$$\gamma \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) p(u_1(a,t))\,dt\,da = \gamma \int_{t_1}^{t_2} v(\alpha_{max},t) r(a) dt. \tag{2.39}$$

Substituting (2.39) into (2.38) we arrive at the following system:

$$\begin{aligned}\int_{a_0}^{a_{max}} W(a,t_2)\,da \;=& \int_{a_0}^{a_{max}} W(a,t_1)\,da \;+ \int_{t_1}^{t_2} W(a_0,t)\,dt \\ &- \int_{t_1}^{t_2} W(a_{max},t)\,dt \\ &- \sum_{i=1}^{T} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} w_i(t)\theta_i W(a,t)\,dt\,da \\ &- \sum_{i=1}^{Q} \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} s_i(t)\psi_i W(a,t)\,dt\,da \\ &- \int_{a_0}^{a_{max}} \int_{t_1}^{t_2} h(a)W(a,t)\,dt\,da\ . \end{aligned} \tag{2.40}$$

If $W(a,t)$ and its partial derivatives are continuous functions then (2.40) can be rearranged into an transport system as demonstrated in (2.21):

$$\frac{\partial W}{\partial t} + \frac{\partial W}{\partial a} = C_2 W(a,t), \tag{2.41}$$

subject to the following initial and boundary condition:

$$\begin{aligned} W(a,0) =& u_1(a,0) - u_2(a,0), \\ W(0,t) =& u_1(0,t) - u_2(0,t). \end{aligned} \tag{2.42}$$

As shown in Corollary 2.4.1,the solution to (2.41) is a system of ODEs with the following form:

$$\begin{cases} \frac{da}{dt} = 1, \\ \frac{d\hat{W}}{dt} = C_2\hat{W}, \end{cases} \tag{2.43}$$

which states that alone the characteristic curve given by $\frac{da}{dt}$, $\frac{dW}{dt} = \exp(C_2 t)$. The solution to (2.43) along the characteristic curve is:

$$\hat{W}(t) = \hat{W}(0)\exp(C_2 t), \tag{2.44}$$

where $\hat{W}(0)$ is the initial condition for the ODE on the characteristic curve and lies either on $W(a,0)$ or $W(0,a)$. Recall Corollary 2.4.5 and Corollary 2.4.4, which state that there exists a set of parameters such that $u(a,t)$ is bounded. Therefore, according to (2.43) and

(2.44), the solution to $u(a,t)$ depends continuously on its initial and boundary conditions, thus $u(a,t)$ is stable.

□

Lemma 2.4.9. *The PDE system* (2.15) *is well-posed for a set of parameters and initial conditions.*

Proof. Corollary 2.4.1 and Corollary 2.4.3 demonstrate that the solution of the two transport equation within (2.15) are systems of first order ODEs. Corollary 2.4.6 and Corollary 2.4.1 proved the existence and uniqueness of the solution to the first order ODEs composing the solution to the two transport equation within our system. Additionally, Corollary 2.4.7 and Corollary 2.4.8 demonstrate that the solution to (2.15) depends continuously on the initial and boundary conditions. Therefore, the PDE system (2.15) is well-posed. □

2.5 Numerical Solution

Recall that our system is a transport system with the form:

$$\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} = -m(a,t)u(a,t), \tag{2.45}$$

where

$$m(a,t) = \left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a) + \frac{\gamma v(\alpha_{max},t) r(a)}{\int_{a_0}^{a_{max}} u(a,t)} \right).$$

Using the first order forward finite difference scheme at $u(a,t)$, where the time interval is divided into N steps of size k (indicated with the superindex j), and the age interval is divided into M segments of size h (indicated with the subindex i). it is clear that:

$$\frac{\partial u}{\partial t} = \frac{u_i^{j+1} - u_i^j}{k},$$

and

$$\frac{\partial u}{\partial a} = \frac{u_i^j - u_{i-1}^j}{h}.$$

Substituting the above Equation into (2.45), and let $r = k/h$, the system becomes:

$$\frac{u_i^{j+1} - u_i^j}{k} + \frac{u_i^j - u_{i-1}^j}{h} = -m_i^j u_i^j,$$

resulting in the following difference equation:

$$\begin{cases} u_i^{j+1} = -km_j^i u_i^j + (1-r)u_i^j + ru_{i-1}^j, \\ u_i^0 = g(i), \\ u_0^j = f(j). \end{cases} \tag{2.46}$$

Using the above schema one can iteratively approximate all $u(a,t)$ at $t = k$ then $t = 2k$ etc.

Lemma 2.5.1. *The above finite difference schema is numerically stable if* $r < 1$ *and* $\max\{\left|m_i^j\right| : \forall i, j\}k < (2-2r)$.

Proof. Let $\hat{u}_i^j$ be the numerical solution of (2.46). Then the error $e_i^j = \hat{u}_i^j - u_i^j$ satisfies:

$$\begin{cases} e_i^{j+1} = -km_j^i e_i^j + (1-r)e_i^j + re_{i-1}^j, \\ e_i^0 = 0, \\ e_0^j = 0. \end{cases}$$

Let $e^j = max\{\left|e_i^j\right| : \forall i\}$ for each $j \geq 0$. Then

$$\begin{aligned} \left|e_i^{j+1}\right| &= \left|(1 - r + km_j^i)\right| + re_{i-1}^j \\ &\leq (\left|1 - r + km_j^i\right| + r)e^j \end{aligned}$$

Thus $e^{j+1} \leq (\left|1 - r - km_j^i\right| + r)e^j$ for all $j \geq 0$. Let $\hat{m} = max\{\left|m_i^j\right| : \forall i, j\}$ for each $i \geq 0$ and $j \geq 0$. For $r < 1$ and $k\hat{m} < (2-2r)$, we have $e^{j+1} \leq e^j$ for all $j \geq 0$. This means that the error e^j is not increasing when j is increasing. Therefore, (2.46) is stable.

The same schema is also applied to $v(\alpha, t)$. Recall that:

$$\frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} = -n(\alpha, t)v(\alpha, t),$$

where

$$n(\alpha,t)=\left(\sum_{i=1}^{T}w_i(t)\phi_i+\sum_{i=1}^{Q}s_i(t)b_i(t)\right).$$

Using the same finite difference schema one can iteratively approximate all $v(\alpha,t)$ at $t=k$ then $t=2k$ etc.

$$\begin{cases} v_i^{j+1}=-kn_j^iv_i^j+(1-rV)v_i^j+rVv_{i-1}^j \\ v_i^0=c(i) \\ v_0^j=d(j) \end{cases}$$

Which is also stable for $rV<1$ and $\max\{n_j^i\}k<(2-2rV)$. □

Additionally, using Taylor expansion, the value of $w(t)$ and $s(t)$ are approximated as following:

$$\begin{cases} w_i(t+k)=w(t)+k\frac{dw_i}{dt}(t), \\ \frac{dw_i}{dt}(t)=o_i(t)-\beta_iw_i(t)-\tau_iw_i\int_{\alpha_0}^{\alpha_{max}}v(a,t)d\alpha, \\ w_i(0)=A, \end{cases}$$

and

$$\begin{cases} s_i(t+k)=s(t)+k\frac{ds_i}{dt}(t), \\ \frac{ds_i}{dt}(t)=l_i(t)-\delta_is_i(t)-\vartheta_is_i\int_{\alpha_0}^{\alpha_{max}}v(a,t)d\alpha, \\ s_i(0)=B. \end{cases}$$

2.6 Results

2.6.1 RBC Equilibrium Condition

The erythropoietic process responds to events altering the concentration of RBCs and ensures that the number and age distribution of RBCs return to normal level upon perturbation. Figure 2.1 shows that using an arbitrary initial age distribution of RBCs, the system eventually returns to the basal RBC age distribution overtime. This result demonstrates that our system reflects the actual biological property of erythropoietic control, where the host

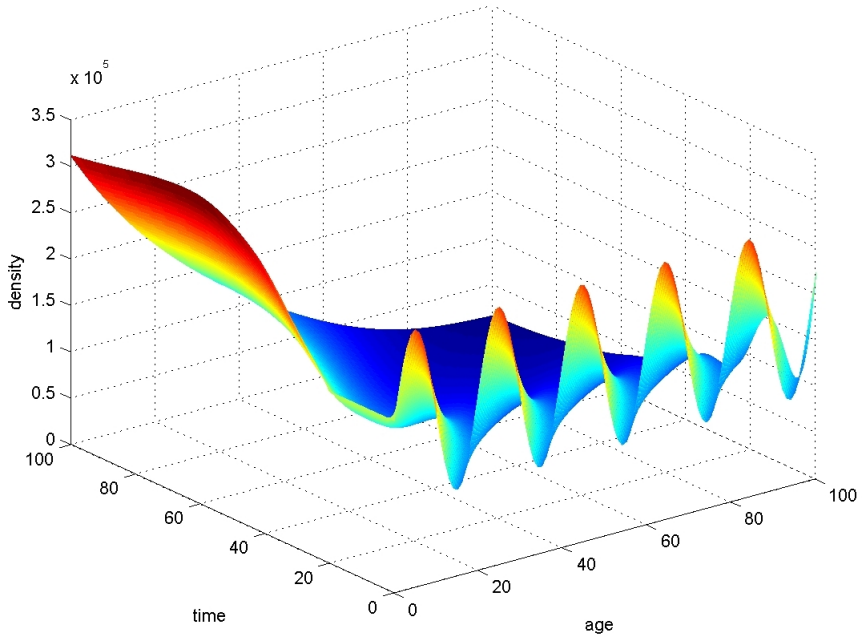

Figure 2.1: Using an arbitrary initial condition, the system returns to Equilibrium condition after 80 days.

responds to a perturbation in concentration of RBCs by altering accordingly, the rate of RBC production in the bone marrow, and the rate of RBCs entering circulation. This regulatory effect can be seen with the fluctuation of $u(0,t)$ in Figure 2.1, which first responds to the lower than normal number of RBCs with a increase of RBCs entering circulation after a delay of 20 days.

2.6.2 Susceptibility

As stated in Corollary 2.4.2, if the immune strength is weak, eventually the iRBC population grows without bounds and RBCs are depleted, resulting in severe anemia (Figure 2.2). Depletion of RBCs is characteristic of first time malaria infection which without treatment could result in mortality of the host. Within our model, this behavior can be achieved under at least two parameter sets. The first one is where C_1, the combined effector strength upon

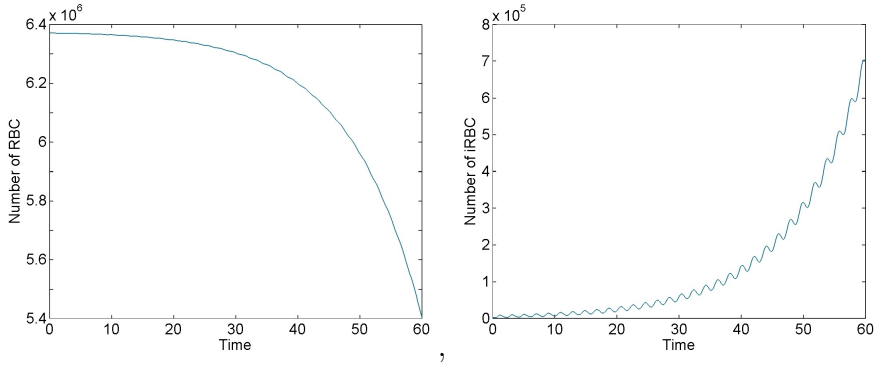

Figure 2.2: The iRBC population grows exponentially due to weak immune strength

saturation of both innate and adaptive immune cells is weak, which allows for the depletion of RBCs. The second set is where C_1 is large enough to halt the exponential growth of iRBCs, but the parameters controlling the speed at which innate immune cells, adaptive immune cells and adaptive immune effector strength growth are small, resulting in the depletion of RBCs before C_1 can reach its maximum value. The second scenario reflects the case where a host is susceptible to severe anemia during primary infection, but during repeated infections, C_1 had enough time to reach maximum value, and the host is resilient to severe anemia [11].

2.6.3 Resilience

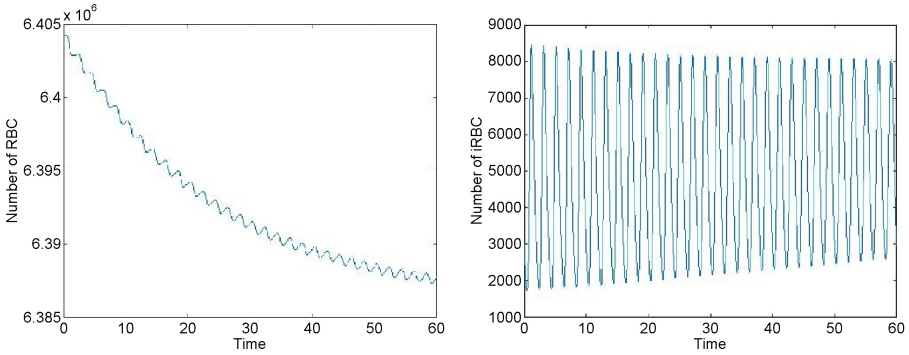

Figure 2.3: iRBC and RBC co-exist when total immune strength is close to $\frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$

If the immune effector strength is close to $\frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$, the iRBC and RBC population coexist for a long period of time as shown in Figure 2.3. This state of resilience is characteristic of repeated malaria infection, where the host has already built up an immune strength against the parasite. However, sustained periodic coexistence is almost never observed in the real world, which is expected due to the fixed maximum immune strength required for sustained coexistence.

2.6.4 Resistance

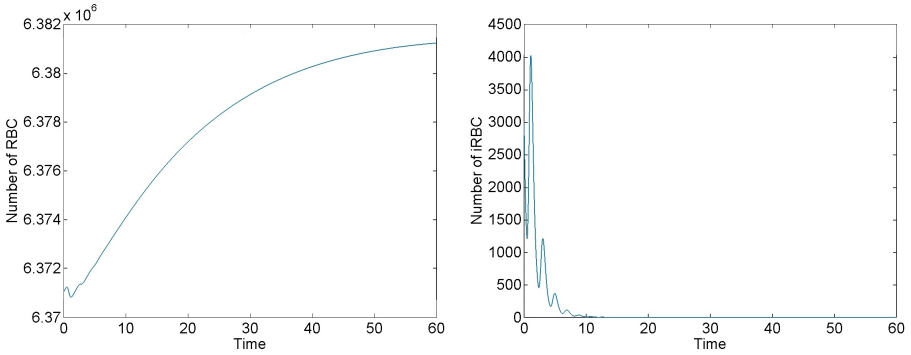

Figure 2.4: iRBC and RBC co-exist when total immune strength is greater than $\frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$

As stated in Corollary 2.4.2, if the immune strength reaches a greater value than $\frac{\ln 1/\gamma\kappa}{\alpha_{max}/V}$, then iRBC population decays constantly until the complete removal of iRBC, as seen in Figure 2.4. This state of resistance is rarely observed without medical intervention during first infection; however, it is feasible to develop an immune response against malaria upon continuous and repeated exposure high enough to clear the parasite [5].

2.7 Discussion

In this paper we present a theoretical model to capture the complex interaction between iRBCs, RBCs and the host immune system during the course of a *Plasmodium* infection. Despite not knowing many of the parameters regarding the immune system's effect on RBCs and iRBCs, our model was able to capture the general behavior of two major outcomes of

infection with *Plasmodium*: (i) the depletion of host RBCs in a naive individual and (ii) the coexistence of RBCs, iRBCs, and the immune system in individuals with previous exposure.

Our model states that the immune effector has to be equal to an exact number for sustained coexistence of the system, which on first look seems rather impossible as biological systems are rarely precise. This further coincides with known malaria infection dynamics where the oscillation of iRBC populations is rarely sustained or regular. To produce sustained oscillation over weeks or months, the immune effector strength only needs to lie in a range rather than being fixed.

It is worth mentioning, that one of the major simplifications of the model was the assumption of $r(a)$, the infection probability of RBCs of different age, to be constant. For sufficiently regular $r(a)$, the well-posedness results can be derived in a similar fashion. However, the rigid trichotomy which conveniently characterizes the system is not inherited. The simplified system in (2.16) results in uncoupled dynamics for iRBCs. This could imply that infected cells may appear even in the absence of uninfected cells; however, this condition would not be physiological. Given realistic initial and boundary conditions, the simplified system permits analysis and approximates well the reality of a malaria infection.

We plan to calibrate our model in the future using time course data from the Malaria Host-Pathogen Interaction Center (MaHPIC), which is a large collaborative project between Emory University, The University of Georgia, The Georgia Institute of Technology, and the Centers for Disease Control and Prevention. The MaHPIC project aims to obtain time courses of -omic data (transcriptomic, metabolomic, lipidomic, proteomic) from non-human primates infected with several different species of *Plasmodium* parasites during multiple experiments, with initial experiments lasting for about 100 days. Using data produced by the MaHPIC group, which include daily RBC counts, iRBC counts and immune cell counts over the 100-day period, we plan to calibrate the full model, including the different infection rates of different age groups of RBCs and the different effector strengths against different age groups of iRBCs. We expect to find discrepancies between our model prediction and the

actual experimental data; such discrepancy will not only provide biological insight regarding the hematodynamics of malaria infection, but will also allow us to re-iterate our model.

Our current PDE model is by no means a complete reflection of the complex within-host dynamics of malaria due to its vast simplification of the immune system's effect on the iRBC and RBC populations along with its complete disregard of the within immune system interaction. However, the model could be expanded to also contain these features and serve as a liaison where the molecular features of the system can be reflected at the cellular level. We plan to use the molecular level data generated by MaHPIC to further expand our model to incorporate the more subtle aspects of the within-host interaction of malaria infection.

Acknowledgments

The authors would like to thank Jacob A. Grey for his help in refining the corollaries 3.4 through 3.7, and also Stacey Lapp, Tracey Lamb, Rabindra Tirouvanziam, Chet Joyner, and Esmeralda Meyer for fruitful discussions. Two anonymous reviewers helped to improve this manuscript significantly. This research has been supported Malaria Host-Pathogen Interaction Center - MaHPIC, NIH's NIAID contract HHSN272201200031C, and the International Centers for Excellence in Malaria Research - Center for non-Amazonian regions of Latin America (CLAIM), NIH's NIAID cooperative agreement U19AI089702.

2.8 Table of Parameters and Variables

2.9 References

Table 2.1: Table of Functions and Variables

	Description	Unit	Value	Source
Independent Variables				
t	Time.	Day	n/a	n/a
a	Age of RBC.	Day	n/a	n/a
α	Age of iRBC.	$10^{-}1Day$	n/a	n/a
Dependent Variables				
$u(a,t)$	Circulating RBC age distribution.	$cells/ul/day$	n/a	n/a
$v(\alpha,t)$	iRBC age distribution.	$cells/ul/day$	n/a	n/a
$\varphi(t)$	Circulating RBC concentration.	$cells/ul$	n/a	n/a
$w_i(t)$	Circulating innate immune cell concentration.	$cells/ul$	n/a	n/a
$s_i(t)$	Circulating adaptive immune cell concentration.	$cells/ul$	n/a	n/a
Functions				
$h(a)$	Percentage of RBC that leaves circulation per day, usually a constant.	1/day	0.02 0.03	MaHPIC
$f(t,\varphi_t)$	Rate at which RBC enters circulation.	$cells/ul/day$	n/a	n/a
$r(a)$	Success rate of merozoite invading RBC of age. a	$1/day$	[0, 1]	Literature
$p(u(a,t))$	Probability of a merozoite infecting a RBC of age. a at time t	dimensionless	n/a	n/a
$b_i(t)$	Rate at which adaptive immunity cell i destroys iRBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
$o_i(t)$	Rate at which innate immune cell enters circulation.	$cells/ul/day$	n/a	Estimated
$l_i(t)$	Rate at which adaptive immune cell enters circulation.	$cells/ul/day$	n/a	Estimated

Table 2.2: Table of Parameters

	Description	Unit	Value	Source
	Parameters			
V	Speed at which parasite ages, equal to $1/\alpha$.	$1/(day)$	10	n/a
φ_0	Normal circulating RBC concentration, equal to $\varphi(0)$.	$cells/ul$	6400000	MaHPIC
γ	Average number of merozoite produced by a single iRBC.	$dimensionless$	25	Literature
θ_i	Rate at which innate immunity cell i destroys RBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
ψ_i	Rate at which adaptive immunity cell i destroys RBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
ϕ_i	Rate at which innate immunity cell i destroys iRBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
β_i	Rate at which innate immunity cell decays.	$1/day$	n/a	Estimated
δ_i	Rate at which adaptive immunity cell decays.	$1/day$	n/a	Estimated
τ_i	Rate at which innate immunity cell decay due to contact with iRBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
ϑ_i	Rate at which adaptive immunity cell decay due to contact with iRBC.	$1/(day \cdot cells/ul)$	n/a	Estimated
λ_i	Coefficient for change of adaptive immunity effector strength.	$1/cells/ul$	n/a	Estimated
ϖ_i	Normal rate of production of innate immune cell i.	$cells/ul/day$	n/a	Estimated
σ_i	Normal rate of production of adaptive immune cell i.	$cells/ul/day$	n/a	Estimated
ϵ_i	Maximum rate of production of innate immune cell i.	$cells/ul/day$	n/a	Estimated
ϱ_i	Maximum rate of production of adaptive immune cell i.	$cells/ul/day$	n/a	Estimated
ν_i	Maximum adaptive immunity effector i strength.	$1/day/cells/ul$	n/a	Estimated
η_i	Coefficient for change of production of innate immune cell i.	$1/cells/ul$	n/a	Estimated
ω_i	Coefficient for change of production of adaptive immune cell i.	$1/cells/ul$	n/a	Estimated
M_i	Amount of parasite where the production of innate immune cell i. increase the most	$cells/ul$	n/a	Estimated
R_i	Amount of parasite where the production of adaptive immune cell i. increase the most	$cells/ul$	n/a	Estimated
xi_i	Amount of parasite where the production of adaptive immune cell i. effector strength increase the most	$cells/ul$	n/a	Estimated
ε	Coefficient for change of erythropoiesis.	$1/cells/ul$	n/a	Estimated
ς	Number of RBC entering blood stream.	$cells/ul/day$	8000	MaHPIC
T_d	Delay of hematopoietic response.	day	10 20	Literature

Bibliography

[1] A. S. Ackleh, B. Ma, and J. J. Thibodeaux. A second-order high resolution finite difference scheme for a structured erythropoiesis model subject to malaria infection. *Mathematical biosciences*, 2013.

[2] J. Belair, M. C. Mackey, and J. M. Mahaffy. Age-structured and two-delay models for erythropoiesis. *Mathematical Biosciences*, 128(1):317–346, 1995.

[3] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS biology*, 1(1):e5, 2003.

[4] W. E. Collins, M. Warren, J. S. Sullivan, and G. G. Galland. Plasmodium coatneyi: observations on periodicity, mosquito infection, and transmission to macaca mulatta monkeys. *The American journal of tropical medicine and hygiene*, 64(3):101–110, 2001.

[5] D. L. Doolan, C. Dobano, and J. K. Baird. Acquired immunity to malaria. *Clinical microbiology reviews*, 22(1):13–36, 2009.

[6] N. M. Douglas, N. M. Anstey, P. A. Buffet, J. R. Poespoprodjo, T. W. Yeo, N. J. White, R. N. Price, et al. The anaemia of plasmodium vivax malaria. *Malar J*, 11(135.10):1186, 2012.

[7] T. J. Lamb and J. Langhorne. The severity of malarial anaemia in plasmodium chabaudi infections of balb/c mice is determined independently of the number of circulating parasites. *Malaria journal*, 7(1):68, 2008.

[8] M. Mackey and J. Milton. *Feedback, delays and the origin of blood cell dynamics.* University of Minnesota. Institute for Mathematics and Its Applications, 1990.

[9] J. M. Mahaffy, J. Belair, and M. C. Mackey. Hematopoietic model with moving boundary condition and state dependent delay] applications in erythropoiesis. *Journal of Theoretical Biology*, 190(2):135–146, 1998.

[10] C. Metcalf, A. Graham, S. Huijben, V. Barclay, G. Long, B. Grenfell, A. Read, and O. Bjørnstad. Partitioning regulatory mechanisms of within-host malaria dynamics using the effective propagation number. *Science*, 333(6045):984–988, 2011.

[11] A. Moreno, M. Cabrera-Mora, A. Garcia, J. Orkin, E. Strobert, J. W. Barnwell, and M. R. Galinski. Plasmodium coatneyi in rhesus macaques replicates the multisystemic dysfunction of severe malaria in humans. *Infection and immunity*, 81(6):1889–1904, 2013.

[12] D. A. Moreno-Pérez, J. A. Ruíz, and M. A. Patarroyo. Reticulocytes: Plasmodium vivax target cells. *Biology of the Cell*, 105(6):251–260, 2013.

[13] D. J. Perkins, T. Were, G. C. Davenport, P. Kempaiah, J. B. Hittner, and J. M. Ong'echa. Severe malarial anemia: innate immunity and pathogenesis. *International journal of biological sciences*, 7(9):1427, 2011.

[14] I. M. Rouzine and F. E. McKenzie. Link between immune response and parasite synchronization in malaria. *Proceedings of the National Academy of Sciences*, 100(6):3473–3478, 2003.

[15] N. J. Savill, W. Chadwick, and S. E. Reece. Quantitative analysis of mechanisms that govern red blood cell age structure and dynamics during anaemia. *PLoS computational biology*, 5(6):e1000416, 2009.

[16] J. J. Thibodeaux. Modeling erythropoiesis subject to malaria infection. *Mathematical biosciences*, 225(1):59–67, 2010.

Chapter 3

A Method for Massively Parallel Analysis of Time Series[1]

[1]Yi H. Yan, Elizabeth D. Trippe and Juan B. Gutierrez. Submitted to *ARXIV* 12/27/2016.

3.1 ABSTRACT

Quantification of system-wide perturbations from time series -omic data (i.e. a large number of variables with multiple measures in time) provides the basis for many downstream hypothesis generating tools. Here we propose a method, Massively Parallel Analysis of Time Series (MPATS) that can be applied to quantify transcriptome-wide perturbations. The proposed method characterizes each individual time series through its ℓ_1 distance to every other time series. Application of MPATS to compare biological conditions produces a ranked list of time series based on their magnitude of differences in their ℓ_1 representation, which then can be further interpreted through enrichment analysis. The performance of MPATS was validated through its application to a study of IFNα dendritic cell responses to viral and bacterial infection. In conjunction with Gene Set Enrichment Analysis (GSEA), MPATS produced consistently identified signature gene sets of anti-bacterial and anti-viral response. Traditional methods such as Empirical Analysis of Digital Gene Expression (EDGE) and Gene Set Enrichment Analysis for Time Series (GSEA-TS) failed to identify the relevant signature gene sets. Furthermore, the results of MPATS highlighted the crucial functional difference between STAT1/STAT2 during anti-viral and anti-bacterial response. In our simulation study, MPATS exhibited acceptable performance with small group size (n = 3), when the appropriate effect size is considered. This method can be easily adopted for other -omic data types.

3.2 Introduction

The advent of high-throughput molecular technologies has allowed us to measure a large number of variables simultaneously. Transcriptomic experiments typically measure the abundance of 20,000 or so transcripts. The number of variables measured by other "omic" technologies such as lipidomic, proteomic, glycomic and metabolomic studies is of similar magnitude. Time series -omic data refers to molecular snapshots taken using these -omic tech-

nologies on a time trajectory. Time series -omic data captures time-dependent molecular dynamics and is a powerful tool in the study of disease progression [2], developmental processes [12] and vaccination [11].

A variety of tools has been developed for the analysis of time series -omic data. Methods such as MESS[3] and EDGE[14] aim to discover individual gene expression time series that are significantly different between two experimental conditions. Other methods such as TcGSA [8] ,CAMERA [19] , and GSEA-TS aim to find time series of pre-defined gene sets that are significantly different among groups. Furthermore, clustering based time series analysis tools [5, 15] have also been developed.

Current time series analysis tools focus on finding gene expression time series or pre-defined gene sets that most likely have changed between groups but ignore the changes in pairwise gene dynamics. Pairwise dynamics between genes can be quantified using correlation, mutual information or a distance metric. Differential correlation has been used to study gene association with the clinical outcome of lung cancer [13] and estrogen receptor modulation in hormonal cancers [10]. In this paper, we propose a novel method, Massive Parallel Analysis of Time Series (MPATS), for the analysis of time series data based on the detection of differential pairwise ℓ_1 distances. MPATS is based on the use of a linear mixed model for gene expression time series. By characterizing each gene expression time series through its ℓ_1 distance to all other time series within the system, MPATS allows for the clear quantification of the impact of perturbation on each time series in the context of the biological system.

A simulation study was conducted to demonstrate the statistical performance of our method and a motivational study was used to validate the biological relevance of the analysis result produced by MPATS. In the motivational study, time series transcriptomic data were used to characterize the responses of IFNα dendritic cells to different antigens [1]. In contrast with traditional time series analysis methods such as EDGE and GSEA-TS, MPATS was able to identify signature gene modules that distinguish between anti-viral and anti-bacterial responses of IFNα dendritic cells.

This paper is organized as follows: Section 2 contains the derivations, theoretical justifications and description of our method. Section 3 provides the assessment of the statistical power of our method based on simulated data. Section 4 contains the application of our method to a recent study of IFNα dendritic cells. Section 5 provides the discussion of the performance of our method and future directions.

3.3 Method Description

3.3.1 Linear Mixed Model of Time Series

Each individual time series is denoted $G_{ij}(t)$ and represents the abundance of gene i from individual j at time point t, where $i \in \{1, 2, ...n\}$, $j \in \{1, 2, ...m\}$ and $t \in \{1, 2...h\}$. We used a mixed linear model [14] for the time series where:

$$G_{ij}(t) = g_{ij}(t) + \nu_t + \epsilon. \tag{3.1}$$

The above model implies that the observed entity $G_{ij}(t)$ can be explained by the additive effect of its time dependent mean response $g_{ij}(t)$, individual and time based variation ν_t, and instrumentation error ϵ. We assume that both ν_t and ϵ follow Gaussian distributions, $\nu_t \sim \mathcal{N}(0, \nu_t)$ and $\epsilon \sim \mathcal{N}(0, \epsilon)$. Equation (3.1) can be further simplified to:

$$G_{ij}(t) \sim \mathcal{N}\left(g_{ij}(t), \nu_t + \epsilon\right).$$

3.3.2 ℓ_1 Distance Between Two Time Series

Let $\ell_1(p, q, j)$ denote the ℓ_1 distance between the mean response curves of genes p and q for subject j.

$$\ell_1(p, q, j) = \sum_{t=1}^{h} |g_{pj}(t) - g_{qj}(t)|. \tag{3.2}$$

Let $X(p, q, j)$ be the observed ℓ_1 distance between the two genes p and q for subject j.

$$X(p, q, j) = \sum_{t=1}^{h} x(p, q, j, t),$$

where

$$x(p,q,j,t) = |G_{pj}(t) - G_{qj}(t)| \sim |\mathcal{N}\left(\mu_t, \sigma_t^2\right)|,$$

and

$$\begin{aligned} \mu_t &= g_{pj}(t) - g_{qj}(t), \\ \sigma_t^2 &= 2(\nu_t + \epsilon) \end{aligned} \tag{3.3}$$

Through numerical investigation, Tsagris et al [18] saw that for any folded normal distribution F, where

$$F \sim |\mathcal{N}(\mu, \sigma^2)|,$$

if $\mu \geq 3\sigma^2$, then F can be well approximated with a normal distribution with mean $|\mu|$ and variance σ^2.

Assuming that for any t, $|\mu_t| \geq 3\sigma_t^2$, then

$$X(p,q,j) \sim \sum_{t=1}^{h} \mathcal{N}(\mu_t, \sigma_t^2).$$

Which means

$$X(p,q,j) \sim \mathcal{N}(\sum_{t=1}^{h} |\mu_t|, \sum_{t=1}^{h} \sigma_t^2).$$

According to equation (3.3) and (3.2) $\sum_{t=1}^{h} |\mu_t| = \ell_1(p,q,j)$, and

$$X(p,q,j) \sim \mathcal{N}(\ell_1(p,q,j), \sum_{t=1}^{h} \sigma_t^2).$$

Which implies that given sufficiently small variance for all t, the observed ℓ_1 distance between two gene expression time series follows an approximate normal distribution with mean equal to the true ℓ_1 distance.

In the case that $|\mu_t| < 3\sigma_t^2$, the mean of $X(p,q,j)$ does not necessarily equal to $\ell_1(p,q,j)$.

Let c be a constant, such that

$$c > \max 10\sigma_t^2 \quad \forall \quad t \in \{1,2,3...h\}.$$

Then let

$$y_1(p,q,t,c,j) = G_{pj}(t) - G_{qj}(t) + c,$$
$$\sim |\mathcal{N}\left(\mu_1, \sigma_t^2\right)|,$$

where

$$\mu_1 = g_{pj}(t) - g_{qj}(t) + c.$$

Let

$$y_2(p,q,t,c,j) = G_{qj}(t) - G_{pj}(t) + c,$$
$$\sim |\mathcal{N}\left(\mu_2, \sigma_t^2\right)|,$$

where

$$\mu_2 = g_{qj}(t) - g_{pj}(t) + c.$$

Because $c > \max 10\sigma_t^2 \quad \forall \quad t \in \{1,2,3...h\}$,

$$y_1(p,q,t,c,j) \sim \mathcal{N}(\mu_1, \sigma_t^2),$$
$$y_2(p,q,t,c,j) \sim \mathcal{N}(\mu_2, \sigma_t^2).$$

Furthermore,

$$\max(\mu_1, \mu_2) = |g_{pj}(t) - g_{qj}(t)| + c.$$

Let

$$z(p,q,t,c,j) = \begin{cases} y_1(p,q,t,c,j) & \text{if } \mu_1 \geq \mu_2, \\ y_2(p,q,t,c,j) & \text{if } \mu_1 < \mu_2. \end{cases}$$

Let

$$Z(p,q,c,j) = \sum_{t=1}^{h} z(p,q,t,c),$$

then,

$$Z(p,q,c,j) \sim \mathcal{N}(\ell_1(p,q,j) + hc, \sum_{t=1}^{h} \sigma_t^2). \tag{3.4}$$

Which implies that $Z(p,q,c,j)$ follows a normal distribution with mean equal to the true ℓ_1 distance plus some known constant.

3.3.3 Detection of Significantly Changed L_{pq} Between Groups

Each individual gene expression time series can be characterized by its ℓ_1 distances to every other gene expression time series. A perturbation of the system (e.g. disease progression) shifts the mean response curve of each gene expression time series with an unknown magnitude and direction.

The impact of the perturbation on the system can be quantified by the number of ℓ_1 distances that have changed due to perturbation. The impact of the perturbation on each individual gene time series can be quantified by the number of pairwise distances to this gene that have changed.

In the case of a two-group experiment, we define two sets of subjects, where $S_1 = \{1, 2, 3, 4...a\}$ and $S_2 = \{a+1, a+2...m\}$. For a given pair of entities p and q, let $A_{pqc} = \{Z(p, q, j, c) : j \in S_1\}$ and $B_{pqc} = \{Z(p, q, j, c) : j \in S_2\}$. As shown in (3.4), both A_{pqc} and B_{pqc} follow a normal distribution with mean corresponding to the ℓ_1 distance between gene p and q within each group. Welch's test is used to determine if A_{pqc} and B_{pqc} have the same mean.

Let $P(p, q)$ be equal to the p-value from Welch's test. The results of conducting Welch's test for every pair of p and q can be captured in a n by n matrix M, where

$$M_{pq} = \begin{cases} 1, P(pq) < \delta \\ 0, P(pq) \geq \delta, \end{cases}$$

where δ is the threshold p value.

For a gene p, its contribution to the perturbation can be quantified by its perturbation score (PS), where:

$$PS = \sum_{q=1}^{n} M_{pq}.$$

The genes are then ranked according to their perturbation score from highest to lowest.

Table 3.1: Categories of Effect Size

Category	ES
Minimal	$0 < ES \leq 0.25$
Low	$0.25 < ES \leq 0.5$
Medium	$0.5 < ES \leq 1$
High	$1 < ES$

3.4 Power Assessment Using Simulated Data

A simulation study was conducted to evaluate the statistical power of MPATS. The simulation framework was chosen based on the motivational study. For each simulation setup, 5000 gene expression time series were simulated and each time series contained 5-time points. The simulated values of each gene at each time point were drawn from a normal distribution with mean and variance estimated from the original data. 10 percent of the gene expression time series were perturbed in the treatment group. Receiver operating characteristic (ROC) curves were generated to explore the statistical performance of the proposed method for different group sizes and effect sizes. The effect size (ES) refers to the true difference in ℓ_1 distance, and the category of effect sizes can be found in Table 3.1. Statistical performance of the proposed method is presented in Figure 3.1. At a fixed specificity of 90%, MPATS has a statistical power greater than 0.7 with $n = 3$ for medium and high effect sizes. With $n = 10$, MPATS can reach a power greater than 0.7 with a fixed specificity of 90% for low effect size.

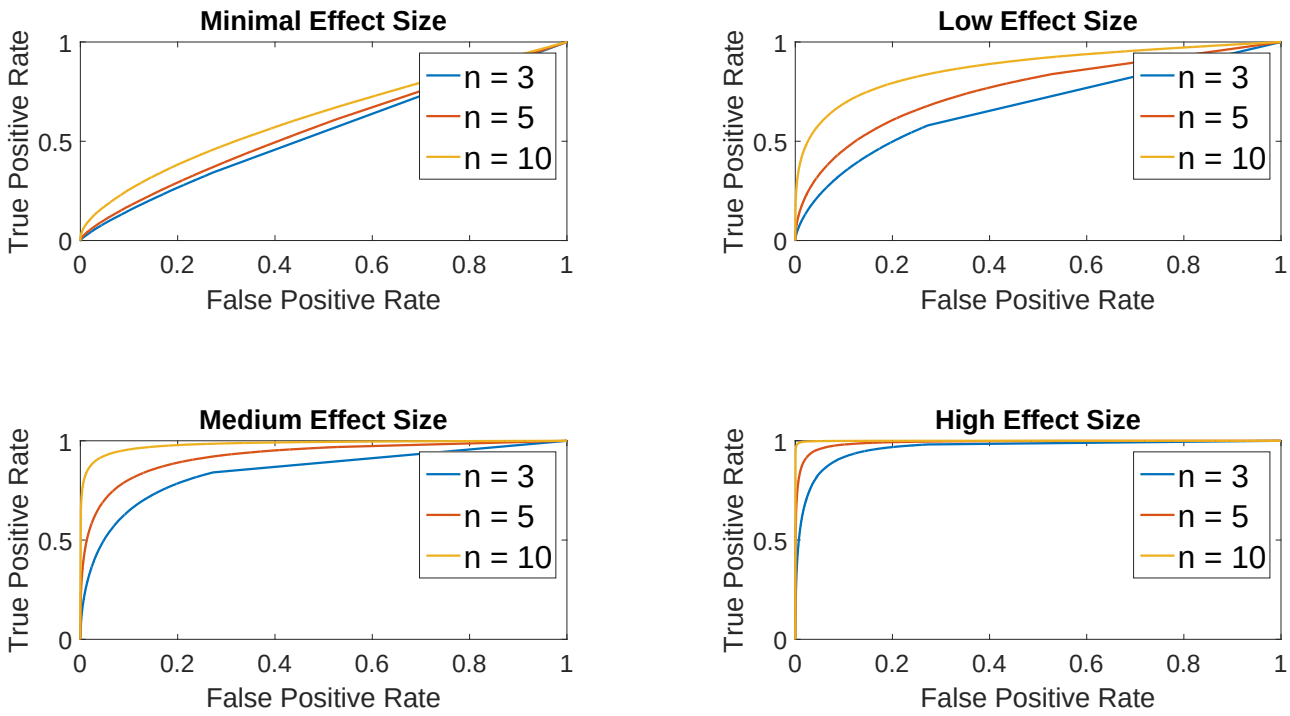

Figure 3.1: ROC curves were generated for 4 different effect sizes. For each effect size, three different sample sizes were tested, $n = 3$, $n = 5$ and $n = 10$. For each simulation, 5000 gene expression time series were generated.

3.5 Experimental Results

Motivating Study: IFNα dendritic cell response to antigen

As part of a study to understand the different responses of dendritic cells to various antigens, microarray time series data were collected for IFNα dendritic cells and IL4 dendritic cells challenged with different antigens. Data were collected at (1,2,6,12,24) hours after challenge. At each time point, three biological replicates were taken. The detailed description of the study can be found in Banchereau et al [1].

For the purpose of the present paper, we focused on the time series of IFNα dendritic cells grown on media alone (MEDIA), challenged with H1N1 virus (H1N1) and heat-killed *Staphylococcus aureus* (HKSA). The data used in this study can be found on Gene Omnibus (GSE44720).

Application of MPATS, EDGE, and GSEA-Time Series

MPATS and EDGE were used to conduct pairwise comparisons among the two treatment groups and the control group. GSEA-TS was applied to the time series data of the two treatment groups alone. During the application of MPATS, a q-value of 0.25 was used as the threshold to determine whether a pairwise ℓ_1 distance had changed significantly between groups. The ranked gene list produced by MPATS was analyzed using GSEA preranked [16]. The results of EDGE consisted of q-values for each gene. A ranked list was produced by ranking the genes according to their q-values in ascending order. The ranked gene list produced by EDGE was also analyzed using GSEA preranked. The gene sets tested for enrichment included: GSEA Hallmark gene sets, GO biological process gene sets, KEGG gene sets, and the signature gene modules described by Banchereau et al [1].

3.5.1 Clustering of pairwise ℓ_1 distance representation of samples

Hierarchical clustering of the samples based on their pairwise ℓ_1 distance representation reveals differences in biological conditions (Figure 3.2). Furthermore, the projection of the data into the first and second principle component space also demonstrate clustering based on biological condition (Figure 3.2). The hierarchical clustering based on biological condition can be reproduced using only the top 100 ℓ_1 distances with the largest variance across all three biological conditions (Figure 3.3).

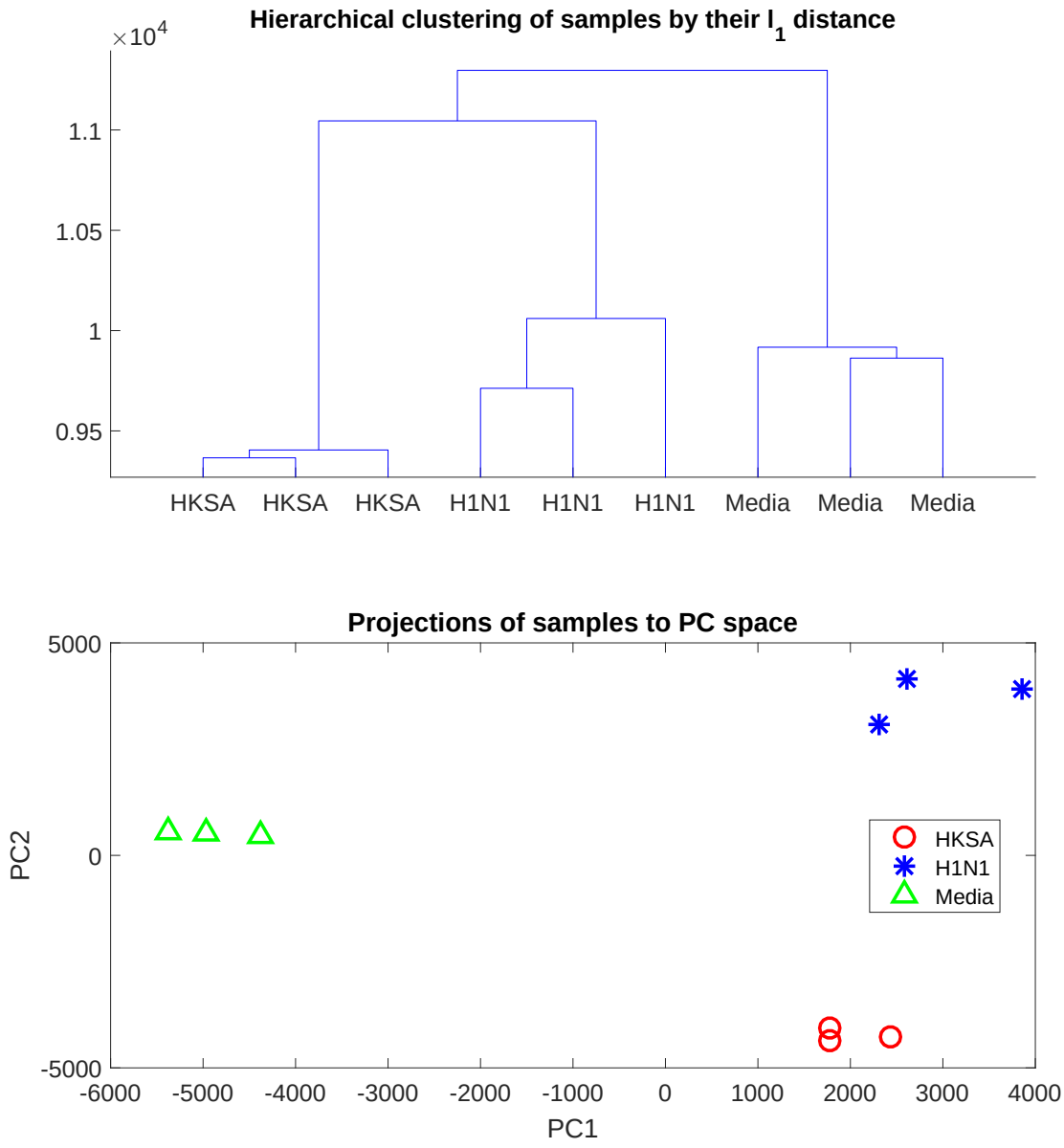

Figure 3.2: Both hierarchical clustering of pairwise ℓ_1 distance and projection into principle component space reveal underlying biological conditions.

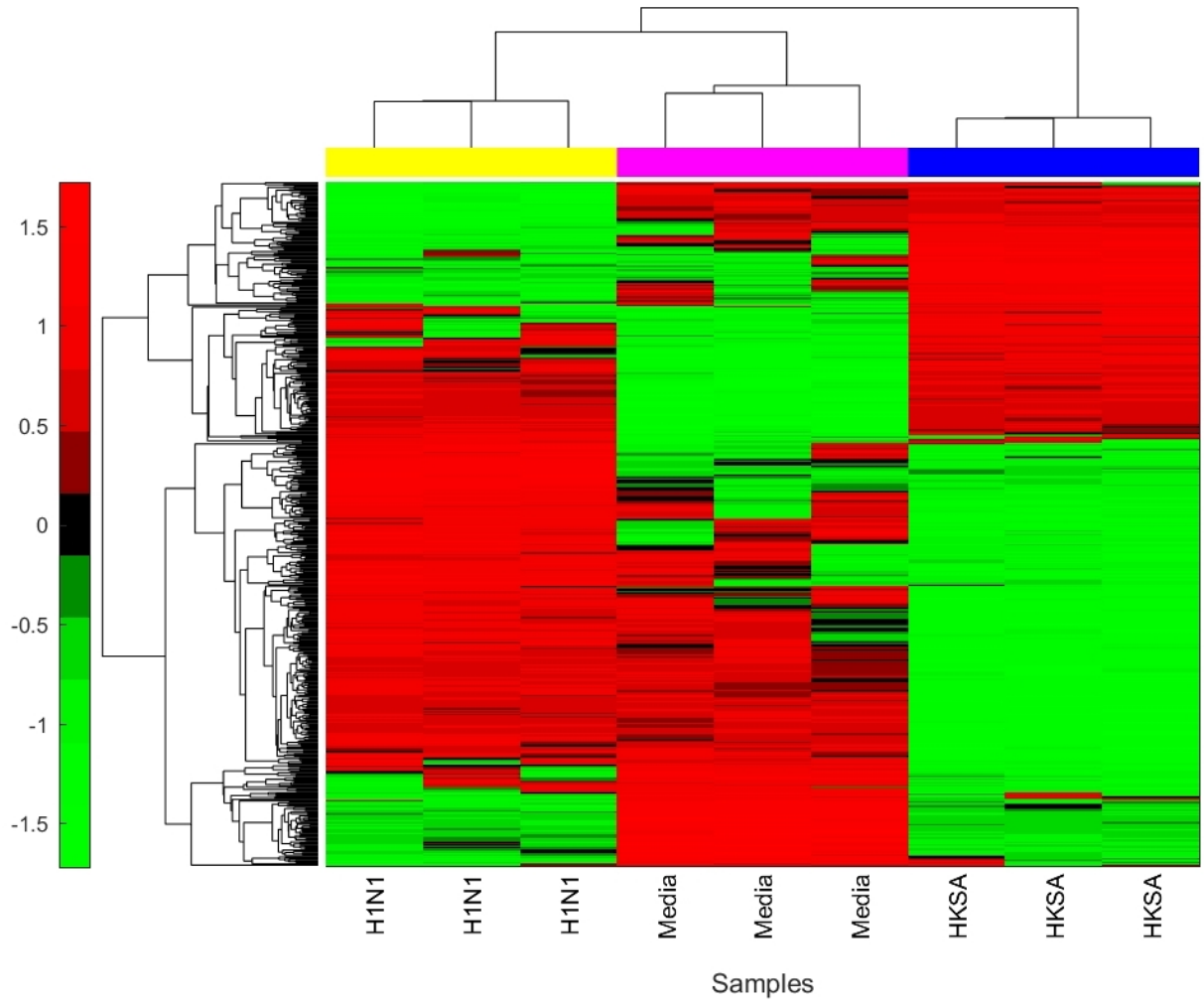

Figure 3.3: Top 100 ℓ_1 distances with the largest variance can be used to cluster the samples based on biological condition. Each row refers to the ℓ_1 distance between two genes. Each column represents a sample.

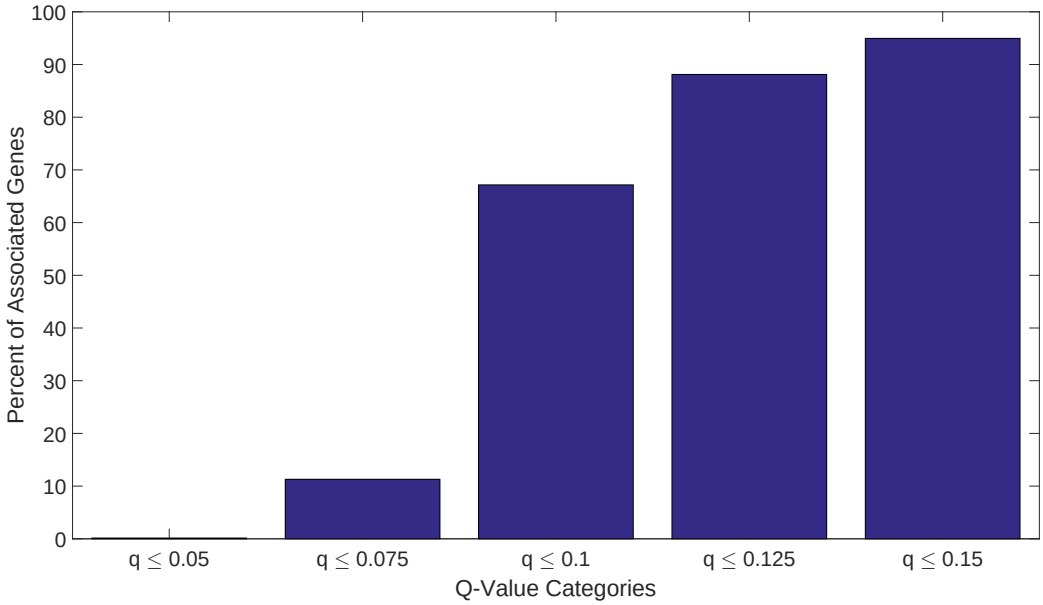

Figure 3.4: Percentage of genes associated with at least one ℓ_1 distance smaller than a threshold (q).

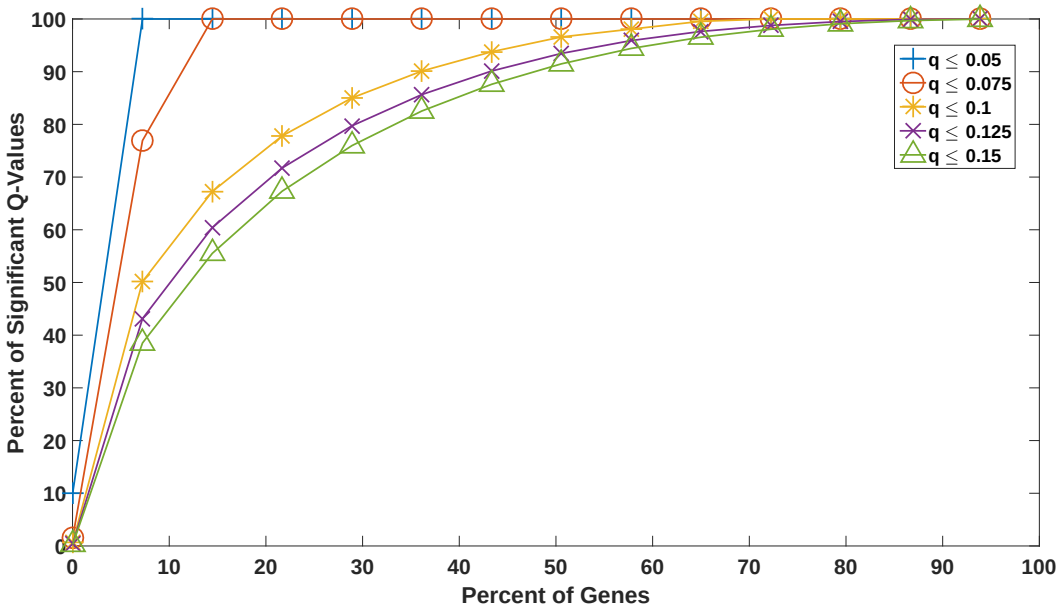

Figure 3.5: Highly significantly changed ℓ_1 distance are associated with a small sub-portion of the total genes.

Only a subset of genes is associated with ℓ_1 distances that have a small q-value. For the comparison between the H1N1 group and the MEDIA group, only 10% of the genes are associated with a ℓ_1 distance with a q-value smaller than 0.075 (Figure 3.4). Furthermore, the highly significantly changed ℓ_1 distances are associated with a small portion of the genes as shown in Figure 3.5. For the comparison the between the H1N1 group and the MEDIA group, 70% of the ℓ_1 distances with $q - value \leq 0.1$ are associated with 20% of the genes. This non-uniform distribution of these significantly changed ℓ_1 distances provides the basis for downstream enrichment analysis. An example of a significantly changed ℓ_1 distance between two genes is shown in Figure 3.6. The ranking of each gene is based on the number of significantly changed ℓ_1 distance associated with that gene. Figure 3.7 shows that the ranking of each gene is strongly associated with the change in the time series of the gene itself.

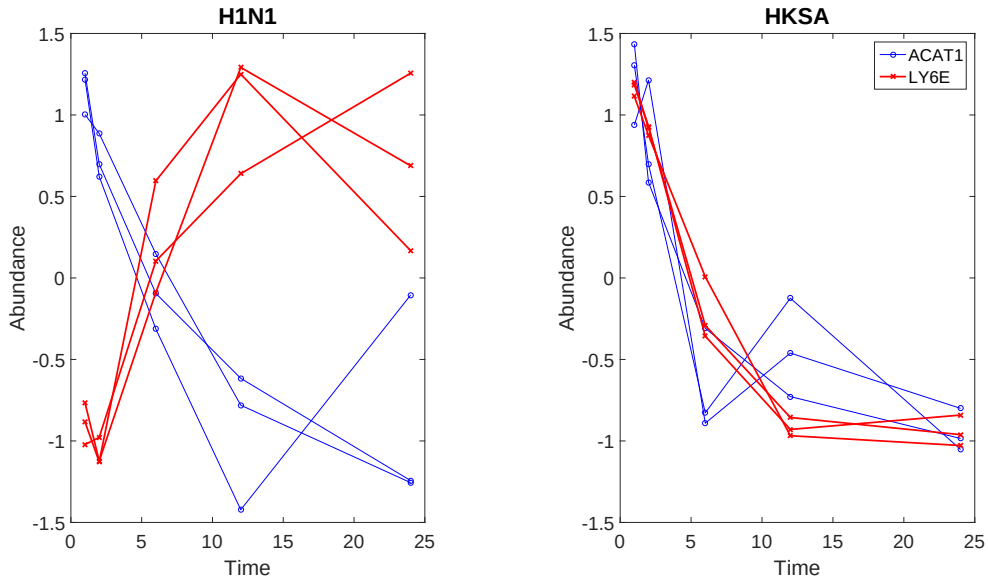

Figure 3.6: ℓ_1 distance between ACAT1 and LYSE changed significantly ($q \leq 0.005$) between the H1N1 challenged group and the HKSA challenged group. Time series from all three biological replicated are displayed.

MPATS Identifies Signature Gene Sets

MPATS identified the greatest number of perturbed gene sets out of all three methods. The number of perturbed gene sets identified by each method is reported in Table 3.2.The differences in the number of perturbed gene sets identified by the three methods highlight the sensitivity of MPATS. This increased sensitivity is due to the fact that MPATS ranks the genes based on their contribution to system-wide dynamic differences between biological conditions. The large number of gene sets perturbed highlights the tremendous impact of antigen challenge on the cell culture. A majority of gene sets and pathways identified are interferon response related pathways. The enrichment results are included in the supplementary material.

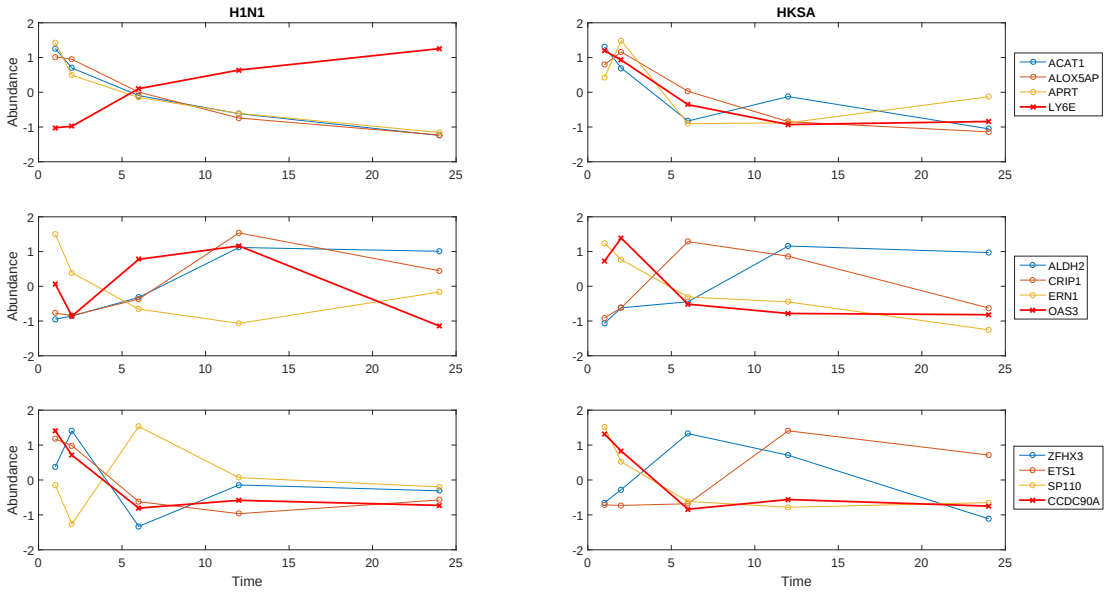

Figure 3.7: The red solid lines are the time series expression profile of the gene of interest in each row. The time series of LY6E, OAS3, and CCDC90A are shown. Three most significantly changed ℓ_1 distances associated with each gene are also plotted. LY6E has rank 1, OAS3 has rank 100, and CCDC90A has rank 1000. The ranking of genes strongly associates with the changes in their temporal expression pattern.

Table 3.2: Three pairwise comparisons were conducted. Comparison between the control group and the treatment group challenged by SA (SA). Comparison between the control group and the treatment group challenged by H1N1 (H1N1). Comparison between the two treatment groups (H1N1\SA). Number of perturbed gene sets discovered by each method is shown in the table

	HKSA	H1N1	H1N1\SA
MPATS	326	348	270
EDGE	24	39	26
GSEA-TS	48	42	NA

In the original study, a list of signature modules was identified for INF cell response to different antigens. Enrichment analysis of both MPATS and EDGE results captured 9 out of 10 signature modules of INF response to the H1N1 challenge, where GSEA-TS only captured 5. All three methods performed worse for the analysis of signature response modules of anti-bacterial responses, capturing only 30% of the signature modules. This could be attributed to the low signal strength of these modules themselves in comparison to the signal strength of the H1N1 modules.

Table 3.3: MPATS results are enriched in the signature modules of INF cells challenged with different antigens. Number of enriched signature modules detected by each method is shown in the table

	SA	H1N1
MPATS	11/30	9/10
EDGE	9/30	9/10
GSEA-TS	12/30	5/10

The analyses comparing IFNα response to H1N1 and HKSA are enriched in interferon-related gene sets, such as response to interferon gamma, TNF signal of beta kappa B and mTORC signaling. Analysis of the top 30 genes using String10 [17] shows an enrichment in protein-protein interactions (PPIs) with $p - value < 0.00001$. Furthermore, the PPIs are clustered around STAT1 and STAT2 (Figure 3.8). This result agrees with recent findings of the crucial differences in the role of STAT1/STAT2 in anti-viral and anti-bacterial responses [4, 6, 7, 9].

3.6 Discussion

In this paper, we presented a novel framework to quantify the perturbation of time series for two group experiments. The quantification of perturbation of time series provides context for downstream functional analysis. We were able to apply this framework to explore anti-viral and anti-bacterial responses of IFNα dendritic cells. We found specific pathways that

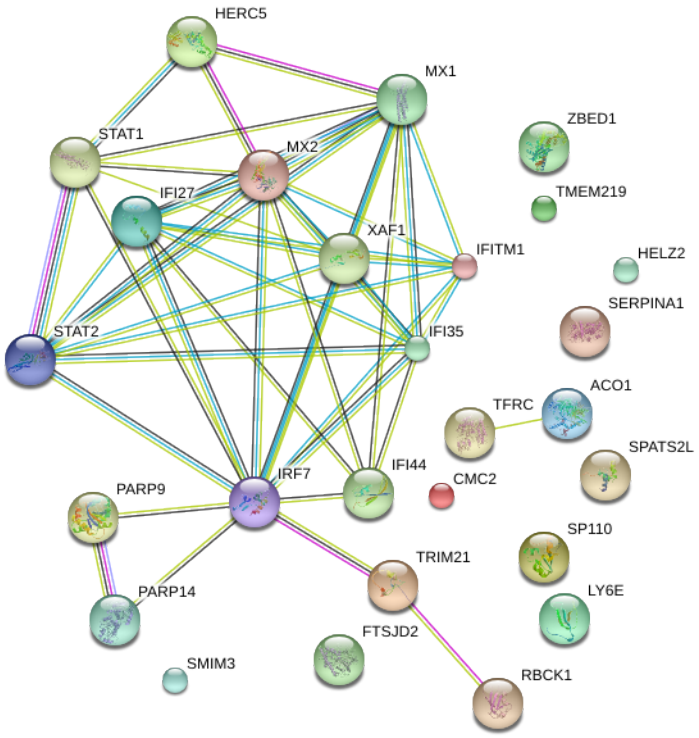

Figure 3.8: PPI Map of the top 30 genes for the comparison of H1N1 and HKSA.

differentiate between the viral and bacterial challenge of IFNα dendritic cells and identified STAT1/STAT2 as important regulatory elements differentiating these responses.

MPATS quantifies system-wide perturbations and individual gene perturbations by characterizing each time series by its ℓ_1 distance to every other time series. The biological significance of this intuitive method was demonstrated through analysis of the motivational study. The top ranked genes are not only enriched in PPIs but are also enriched in signature gene sets. The performance of MPATS remained stable through all three analyses, whereas EDGE and GSEA-TS identified varying numbers of perturbed gene sets. This is probably due to the fact that ranking by the p-value of change is not biologically informative because a change of small magnitude with low variance can generate a low p-value. Ranking time series using variance normalized effect size should produce results similar to MPATS. The

framework of MPATS is based on a linear mixed model and assumes normally distributed variance, and the use of pairwise dynamics to quantify the magnitude of change for each individual entity can be easily expanded to other -omic data types.

In conclusion, MPATS complements existing time series analysis methods by providing a more biologically informative ranked list of genes of interest in addition to detecting genes that have experienced large perturbations but are overlooked by existing methods.

Acknowledgments

This project was funded in part by Federal funds from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract #HHSN272201200031C (PI: Mary Galinski), which supports the Malaria Host-Pathogen Interaction Center (MaHPIC).

3.7 References

Bibliography

[1] R. Banchereau, N. Baldwin, A.-M. Cepika, S. Athale, Y. Xue, I. Y. Chun, P. Metang, A. Cheruku, I. Berthier, I. Gayet, et al. Transcriptional specialization of human dendritic cell subsets in response to microbial vaccines. *Nature communications*, 5, 2014.

[2] C. Bécavin, N. Tchitchek, C. Mintsa-Eya, A. Lesne, and A. Benecke. Improving the efficiency of multidimensional scaling in the analysis of high-dimensional data using singular value decomposition. *Bioinformatics*, 27(10):1413–1421, 2011.

[3] M. Berk, C. Hemingway, M. Levin, and G. Montana. Longitudinal analysis of gene expression profiles using functional mixed-effects models. In *Advanced Statistical Methods for the Analysis of Large Data-Sets*, pages 57–67. Springer, 2012.

[4] K. Blaszczyk, A. Olejnik, H. Nowicka, L. Ozgyin, Y.-L. Chen, S. Chmielewski, K. Kostyrko, J. Wesoly, B. L. Balint, C.-K. Lee, et al. Stat2/irf9 directs a prolonged isgf3-like transcriptional response and antiviral activity in the absence of stat1. *Biochemical Journal*, 466(3):511–524, 2015.

[5] J. Ernst, G. J. Nau, and Z. Bar-Joseph. Clustering short time series gene expression data. *Bioinformatics*, 21(suppl 1):i159–i168, 2005.

[6] B. Hahm, M. J. Trifilo, E. I. Zuniga, and M. B. Oldstone. Viruses evade the immune system through type i interferon-mediated stat2-dependent, but stat1-independent, signaling. *Immunity*, 22(2):247–257, 2005.

[7] S. Hambleton, S. Goodbourn, D. F. Young, P. Dickinson, S. M. Mohamad, M. Valappil, N. McGovern, A. J. Cant, S. J. Hackett, P. Ghazal, et al. Stat2 deficiency and susceptibility to viral illness in humans. *Proceedings of the National Academy of Sciences*, 110(8):3053–3058, 2013.

[8] B. P. Hejblum, J. Skinner, and R. Thiébaut. Time-course gene set analysis for longitudinal gene expression data. *PLoS Comput Biol*, 11(6):e1004310, 2015.

[9] M. J. Hofer, W. Li, P. Manders, R. Terry, S. L. Lim, N. J. King, and I. L. Campbell. Mice deficient in stat1 but not stat2 or irf9 develop a lethal cd4+ t-cell-mediated disease following infection with lymphocytic choriomeningitis virus. *Journal of virology*, 86(12):6932–6946, 2012.

[10] T.-H. Hsiao, Y.-C. Chiu, P.-Y. Hsu, T.-P. Lu, L.-C. Lai, M.-H. Tsai, T. H.-M. Huang, E. Y. Chuang, and Y. Chen. Differential network analysis reveals the genome-wide landscape of estrogen receptor modulation in hormonal cancers. *Scientific reports*, 6, 2016.

[11] R. E. Palermo, L. J. Patterson, L. D. Aicher, M. J. Korth, M. Robert-Guroff, and M. G. Katze. Genomic analysis reveals pre-and postchallenge differences in a rhesus macaque aids vaccine trial: insights into mechanisms of vaccine efficacy. *Journal of virology*, 85(2):1099–1116, 2011.

[12] S. Roy, J. Ernst, P. V. Kharchenko, P. Kheradpour, N. Negre, M. L. Eaton, J. M. Landolin, C. A. Bristow, L. Ma, M. F. Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

[13] K. Shedden and J. Taylor. Differential correlation detects complex associations between gene expression and clinical outcomes in lung adenocarcinomas. In *Methods of Microarray Data Analysis*, pages 121–131. Springer, 2005.

[14] J. D. Storey, W. Xiao, J. T. Leek, R. G. Tompkins, and R. W. Davis. Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences of the United States of America*, 102(36):12837–12842, 2005.

[15] J. Straube, A.-D. Gorse, B. E. Huang, K.-A. Lê Cao, et al. A linear mixed model spline framework for analysing time course omics data. *PloS one*, 10(8):e0134540, 2015.

[16] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[17] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, page gku1003, 2014.

[18] M. Tsagris, C. Beneki, and H. Hassani. On the folded normal distribution. *Mathematics*, 2(1):12–28, 2014.

[19] D. Wu and G. K. Smyth. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic acids research*, 40(17):e133–e133, 2012.

Chapter 4

Ensemble Differential Network Analysis (eDiNA)[1]

[1]Yi H Yan, Elizabeth D. Trippe and Juan B Gutierrez. To be submitted to *Nature Scientific Report*

4.1 ABSTRACT

Differences in the transcriptome between malaria infected and healthy individuals have the potential to reveal molecular mechanisms underlying host response to malaria infection. Differential network analysis (DiNA) is a recent class of algorithms designed to identify differences in network topology between states. Here we propose an ensemble based differential network analysis (eDiNA) tool to conduct comparative analyses of transcriptome profiles between *Plasmodium cynomolgi* infected *Macaca mulatta* and healthy controls. The analysis was carried out for both bone marrow and whole blood transcriptome profiles separately. Of particular interest, our analysis reveals that the differential topology of the bone marrow transcriptome is enriched in oxidative phosphorylation and mitochondrial specific genes. Furthermore, the analysis of the whole blood transcriptome exhibits enrichment of antigen processing and presentation, spliceosome, and T-cell signaling pathway related genes. Our analysis provides novel insights into host transcriptomic responses to malaria infection, which can be used to gain a better understanding of the mechanistic causes underlying malaria pathology.

4.2 Introduction

Human malaria is a major public health burden and host response to malaria infection plays a major role in malaria pathology [1, 7, 11]. There were 212 million (UI: 148-304 million) reported cases and 429,000 (UI: 235,000 - 639,000) estimated deaths in 2015 [10]. Despite the vast potential of host transcriptome data to help elucidate the molecular mechanisms underlying malaria pathology, very few primate host transcriptome studies have been conducted. In 2005, Joni Ylostalo *et al.* generated the first transcriptome time series of two *Plasmodium cynolomolgi* infected *Macaca mulattas* [17]. In 2014, Junya Yamagishi *et al.* conducted RNA-seq analysis of 116 Indonesian patients infected with *Plasmodium falciparum* and discovered sets of host genes that correlate with the severity of malaria infections [15]. Most recently,

the Malaria Pathogen Interaction Center (MaHPIC) consortium published a novel data set containing the time series transcriptome profiles of five *P cynolmogi* infected *M mulatta* and five healthy hosts.

Differential network analysis (DiNA) refers to a recent school of algorithms focused on identifying differences in network topology between states. Unlike traditional differential analysis, DiNA identifies changes in the pairwise dynamics of genes rather than the change of abundance of individual genes. DiNA has been successfully used for the identification of transcriptional regulator [2] and estrogen modulated genes in cancer [5]. Various methods have been proposed to conduct DiNA, such as modulator inference by network dynamics (MINDy) [14], differential network analysis in genomics (DINGO) [4] and modulated gene interaction (MAGIC) analysis [5]. Each algorithm adapts its own metric to detect differential dynamics between genes, MINDy uses mutual information based measurement of dynamics between two genes, MAGIC uses Spearman correlation and DINGO utilizes a Gaussian graphical model. Each individual metric aims to characterize pair-wise dynamics based on a derived quantity, and each metric has its own limitations, for example, Pearson's correlation coefficient only detects changes in linear dependency between two variables.

In this paper, we present an ensemble differential network analysis (eDiNA) method that provides a flexible framework to combine multiple metrics for the detection of differential dynamics. A simulation study was done to demonstrate the statistical power of our proposed method and the method was applied to the above-mentioned MahPIC data set. The analysis was carried out for both bone marrow and whole blood transcriptome data separately. Of particular interest, our analysis reveals that the differential topology of the bone marrow transcriptome is enriched in oxidative phosphorylation and mitochondrial specific genes. Furthermore, the analysis of the whole blood transcriptome demonstrates enrichment of antigen processing and presentation, spliceosome, and T-cell signaling pathway related genes.

This paper is organized as follows: Section 3 contains the detailed theoretical background and derivation of our method. Section 4 provides the assessment of the statistical power of

our method based on simulated data. Section 5 contains the application of our method to the MahPIC data set. Section 6 provides the discussion of the performance of our method, the biological interpretation of the result and future directions.

4.3 Method

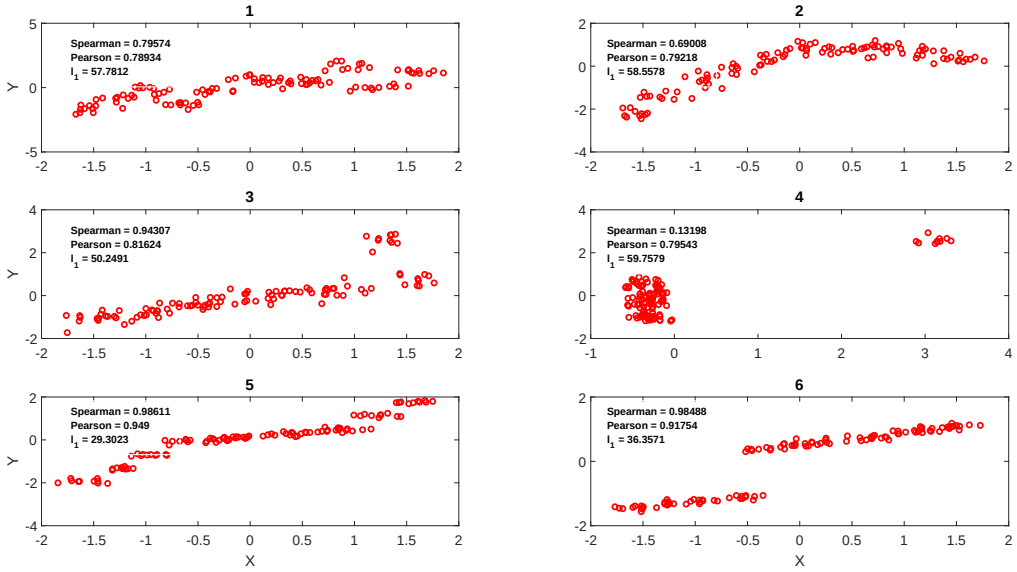

Figure 4.1: Examples of differential pairwise dynamics between two variables, X and Y. Subplot 1 through 4 is the modified Anscombe's quartet, which differs in Pearson's correlation by 10^{-3}. Subplot 1 through 4 can be separated by their Spearman's correlation and ℓ_1 distances. Subplot 4 and 5 demonstrates the inability of both Pearson's correlation and Spearman's correlation to distinguish between a monotonically increasing relationship between X and Y and a step-function like increase. ℓ_1 distance can be used to separate between the behavior exhibited by subplot 5 and 6

4.3.1 Method Description

eDiNA characterizes pairwise dynamics of genes using four measurements: Pearson's Correlation, Spearman's Correlation, ℓ_1 distance and χ^2 statistics. Figure 4.1 provides examples of different pairwise dynamics that are indistinguishable using only one of the above-mentioned measurements. For example, the first 4 sub-figures of Figure 4.1 all have the same Pearson's

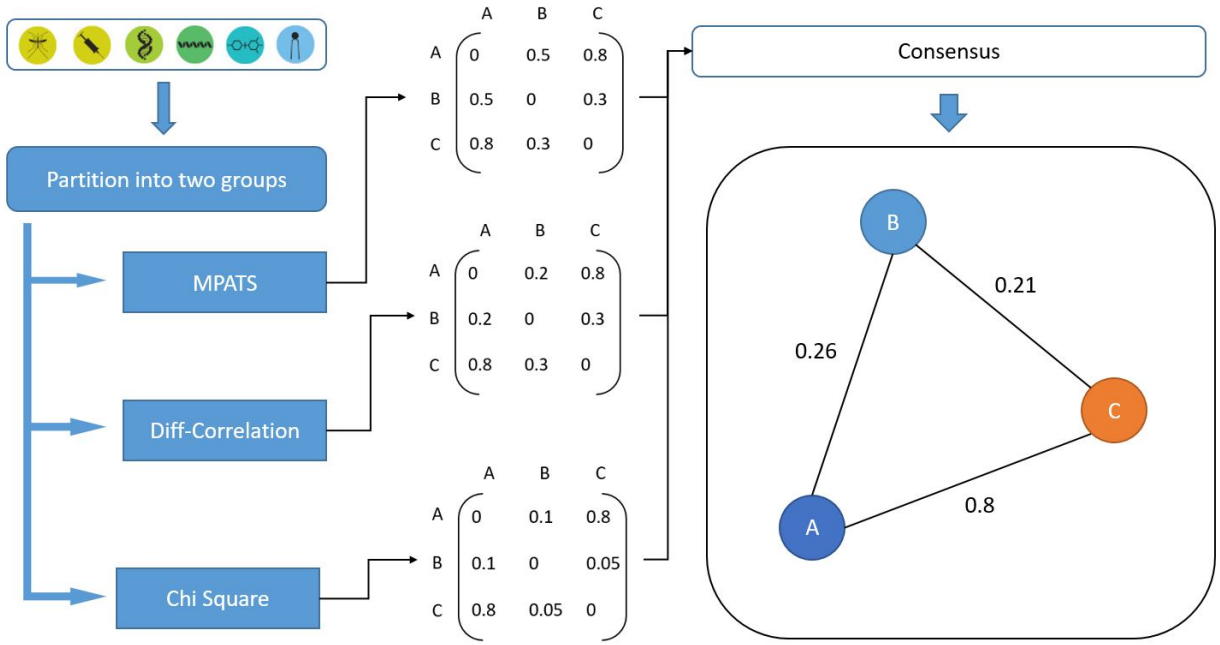

Figure 4.2: eDiNA pipeline is composed of two major components. Individual statistical testings are carried out to detect differences of pairwise dynamics in correlation, ℓ_1 distance and χ^2 statistics. The overall null hypothesis, that the pairwise dynamics between two genes did not differ in any of the proposed measurement is then carried out using Fisher's combined p-value test. The final result can be visualized as a graph, where each node represents a gene, and the edge weight denotes the result of the Fisher's combined p-value test.

Correlation coefficient despite having vastly different dynamics. Figure 1.5 and 1.6 differs most significantly in their pairwise ℓ_1 distance.

An overview of eDiNA is presented in Figure 4.2. eDiNA tests the consensus null hypothesis (H_0): For a given pair of genes, X and Y, none of the four measurements of pairwise dynamics differ between condition 1 and condition 2. A statistical test is conducted to detect significant differences in each of the four measurements. Fisher's transformation was used to test for differential Pearson's and Spearman's correlation, MPATS was used to detect differential ℓ_1 distance and Pearson's test of independence was used to detect differences in the joint distribution of X and Y after discretization. Each of these statistical tests produces a p-value to reject the individual null hypothesis. Fisher's combined probability test is then used to test the consensus null hypothesis that the pairwise dynamics between X and Y does not differ in any of the measurements. The resulting p-values are then adjusted for multiple hypothesis testing [12].

4.3.2 Fisher's Transformation and Differential Correlation

Taking n paired observations of variables X and Y from condition 1, and taking m paired observations of X and Y from condition 2 and let r_1 and r_2 denote the correlation coefficient between X and Y under condition 1 and 2 respectively. Fisher's transform [3] is then used to generate Z_1 and Z_2, where

$$\begin{cases} Z_1 &= \frac{1}{2}\ln\frac{1+r_1}{1-r_1}, \\ Z_2 &= \frac{1}{2}\ln\frac{1+r_2}{1-r_2}. \end{cases}$$

A z-test is then conducted where

$$Z = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n-3} + \frac{1}{m-3}}}.$$

$Z \sim \mathbb{N}(0,1)$ under the null hypothesis that $r_1 = r_2$. The p-value is then calculated accordingly.

4.3.3 Pearson's test of independence

To apply Pearson's test of independence, continuous gene expression data is first discretized. Setting the maximum number of quantization levels to be 6, a Gaussian Mixture model was constructed for each variable with predefined 1 to 6 kernels. Bayesian information criterion was then used to determine the optimal level of quantization, n. Each continuous variable is then discretized using k-means clustering where $k = n$.

A contingency table for each gene pair is then constructed using the discretized gene expression. The columns are condition 1 and condition 2. Each row represents a possible combination of the discretized expression level of gene X and Y. The combination of discretization with zero observations is then removed from the contingency table. Pearson's test of independence is then used to test whether the discrete joint distribution of X and Y are independent of the condition, and a p-value is calculated.

4.3.4 MPATS

MPATS [16] utilizes a linear mixture model for gene time series. Where:

$$G_{ij}(t) = g_{ij}(t) + \nu_t + \epsilon.$$

The observed abundance of gene j in individual i at time point t can be explained by the additive effect of its time dependent mean response $g_{ij}(t)$, individual and time based variation ν_t, and instrumentation error ϵ.

MPATS is designed to test whether the ℓ_1 distance between two genes, a and b, sampled for n time points where

$$\ell_1 = \sum_{t=1}^{t=n} |g_{ia}(t) - g_{ib}(t)|$$

differs significantly between two conditions. A detailed description of the method can be found in the publication by Yan *et al.* [16].

Table 4.1: Categories of Effect Size

Category	ES
Low	$0 < ES \leq 0.6$
Medium	$0.6 < ES \leq 1.2$
High	$1.2 < ES$

4.3.5 Fisher's Combined Probability Test

Fisher's combined probability test is used to test the hypothesis that each independent null hypothesis is true. (For a given pair of genes X and Y, their Pearson's correlation, Spearman's correlation, ℓ_1 distance and discretized joint distributions are the same between condition 1 and condition 2)

Let p_i denote the p-value from test i. Under the assumption that the p-values are independent, then

$$-2\sum_{i=1}^{k}\log(p_i) \sim \chi^2_{2_k}$$

if the k individual null hypothesis tested are all true. A p-value is then generated from a χ^2 distribution with $2k$ degrees of freedom.

4.3.6 Enrichment Analysis and GSEA

For a gene X, a perturbation score (PS) is assigned. Where the score is the sum of the χ^2 statistics produced by Fisher's combined probability test for each pairwise dynamics involving X. A ranked list of genes is produced by arranging the genes in descending perturbation score and analyzed using GSEA [13].

4.4 Simulation Study

A simulation study was conducted to evaluate the statistical power of eDiNA. The simulated data set were constructed to closely resemble the MaHPIC dataset. 4,000,000 pairs of gene time series were simulated. The simulated values of each gene time series at each time point were drawn from a normal distribution with mean and variance estimated from the MaHPIC study. 20 percent of the paired gene time series were perturbed in the simulated treatment group. Receiver operating characteristic (ROC) curves were generated to explore the statistical power of eDiNA for different group sizes and effect sizes. The effect size (ES) refers to the true difference in normalized ℓ_1 distance, Pearson's correlation or Spearman's correlation. ℓ_1 distances are normalized to have a range of $[0, 2]$. The categories of ES can be found in Table 4.1. The ES of a differential pairwise dynamics is determined by the maximum difference among normalized ℓ_1 distance, Pearson's correlation or Spearman's correlation. Statistical performance of the proposed method is presented in Figure 4.3. At a fixed specificity of 90%, eDiNA has a statistical power greater than 0.7 with $n = 5$ for medium and high effect sizes. The consensus method outperforms all individual method for the detection of medium and high effect sizes independent of group sizes.

4.5 Analysis of MaHPIC Data

4.5.1 Study Design

The MaHPIC data set is composed of 2 studies, E13 and E04. E13 is a study of five *M. mulatta* without infection over 100 days. Transcriptomic data of these five hosts were collected at 7-time points from both the bone marrow and whole blood. A detailed description of the study can be found in the paper by Kevin *et al.* [8]. E04 is a study of five *M. mulatta* infected with *P. cynolmogi*. Transcriptomic data of these five hosts were collected at 7-time points from both the bone marrow and whole blood. This study resulted in two severe malaria infections, two

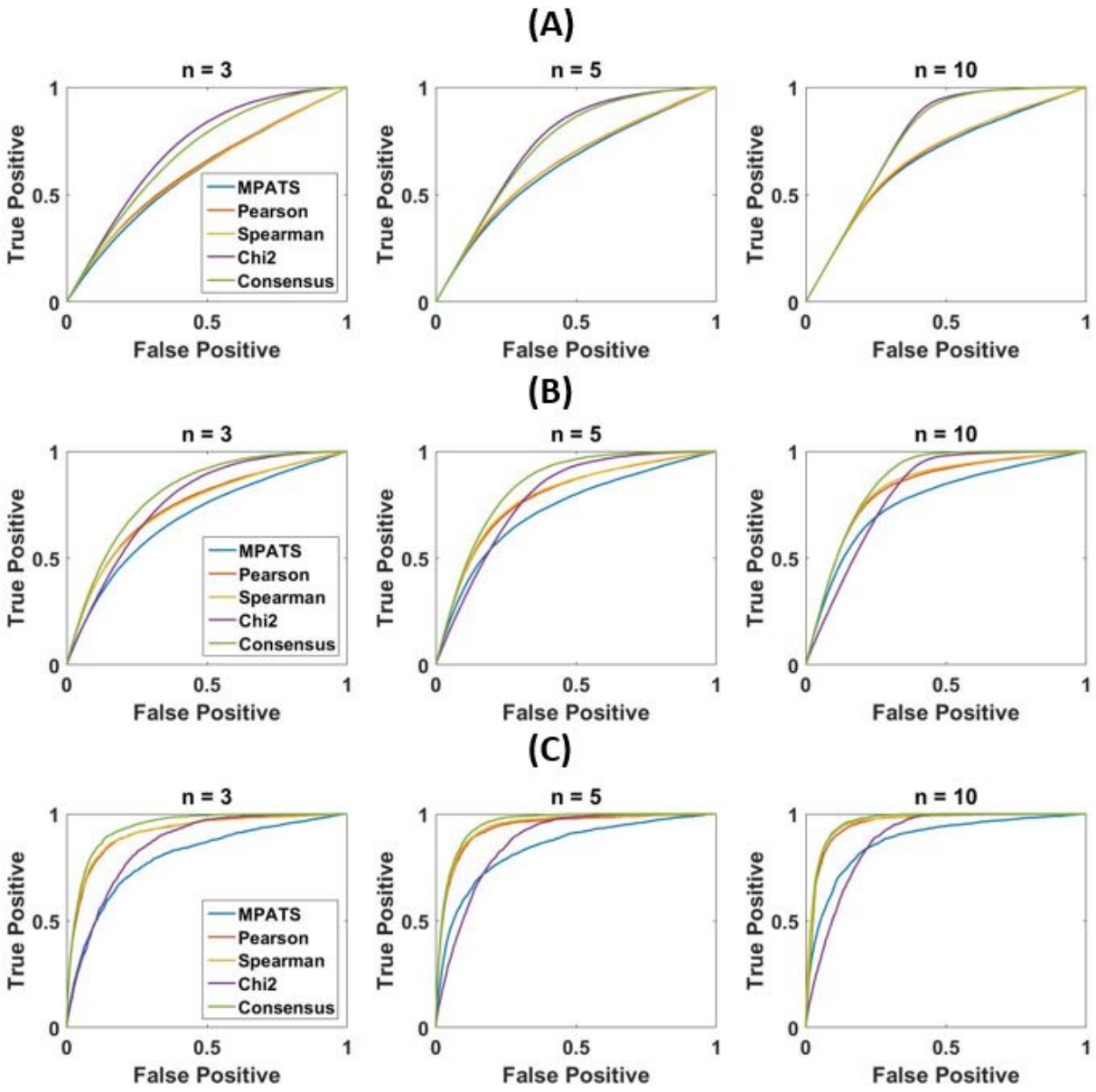

Figure 4.3: ROC curve of individual methods and consensus method for a different number of subjects and ES categories.(A) Low effect size. (B) Medium effect size. (C) Large Effect Size

mild malaria infections, and one death. A detailed explanation of the experimental scheme can be found in the publication by Chester *et al.* [6].

4.5.2 Result Overview

Hierarchical clustering based on Pearson's correlation, Spearman's Correlation, pairwise ℓ_1 distance and ensemble distance was conducted using transcriptomic data from the bone marrow of all 9 subjects Figure 4.4. Hierarchical clustering of all four measurements except for Spearman's correlation clustered the healthy hosts (H) closer to other healthy hosts than to the disease host, regardless of mild (M) or severe clinical outcome. Principal component analysis (PCA) was also applied and the result was then projected into the first 2 principal component (PC) space to be inspected visually. The first 2 PC explained 32.2 percent of the total variation and showed clear separation of both the healthy and diseased hosts in addition to separation of mild infection and severe infection.

Hierarchical clustering and PCA were also applied to whole blood transcriptomic data Figure 4.5. The resulting clustering and projection onto the PC space showed a similar pattern to that of bone marrow data. Specific to whole blood data, hierarchical clustering based on Spearman's correlation clustered the data correctly based on healthy and disease state. In both cases, using any of the pairwise dynamics measures allows for the separation of healthy and disease state, demonstrating that pairwise gene dynamics are in fact different between the two states where between states differences are greater than within state differences. This finding provides the basis for the down-stream application of eDiNA to identify the particular gene-pairs that exhibit such differences in pairwise dynamics and how their biological function are related to host response to malaria infection.

4.5.3 Comparative Analysis of Bone Marrow Transcriptome Between Healthy and Infected Hosts

eDiNA was applied to identify significantly different pairwise gene dynamics between transcriptomic data from the bone marrow of healthy host and disease host. Genes with

low signal, library size normalized read count $< log_2(2.5)$ were filtered out. 10,073 genes were analyzed for differential dynamics. After adjustment for multiple hypothesis testing, 14 percent of the pairwise dynamics had a q-value less than 0.05. An overview of differential pairwise gene dynamics with the lowest q-values are shown in Figure 4.6. The gene pair that had the lowest q-value was SLC14A1-PAQR9, as shown in Figure 4.7.

The distribution of PS is visualized using a histogram in 4.8. The ranked PS plot demonstrates an asymptotic behavior, where the change of PS across rank decreases after the first few hundred genes. Interestingly, subnetwork composed of the top 300 genes have a much higher average edge weight and shows a significantly different distribution of edge weight than that of the entire set of genes. Differential dynamics with q-value $< 10^{-23}$ are visualized in Figure 4.6. This non-uniform distribution of PS and edge weight are then explored for biological significance using GSEA.

GSEA-Preranked was applied to the ranked list of genes according to PS, the weighted scoring scheme was selected to emphasize the distribution of PS. Gene set tested includes the hallmark gene set, GO gene set, annotated gene sets, motif gene sets, and immunological signature gene sets. Out of 12,423 gene sets tested for positive enrichment, 492 gene sets had q-value < 0.05. Leading edge analysis reveals the clustering of immune-related gene sets, specifically wnt signaling, oxidative phosphorylation, and purine biosynthesis.

4.5.4 Comparative Analysis of Whole Blood Transcriptome Between Healthy and Infected Hosts

eDiNA were applied to identify significantly different gene pairwise dynamics between transcriptomic data from the whole blood of healthy host and disease host. Genes with low signal, library size normalized read count $< log(2.5)$ were filtered out. 9640 genes were-were analyzed for differential dynamics. After adjustment for multiple hypothesis testing, 12

percent of the pairwise dynamics had a q-value less than 0.05. An overview of differential pairwise gene dynamics with the lowest q-values are shown in Figure 4.9. The gene pair that had the lowest q-value was ARAP2-RNF125, as shown in Figure 4.10.

The distribution of edge weight and PS follows the same pattern as the analysis result of bone marrow data4.11. However, GSEA result showed a different pattern. Out of 12423 gene sets tested for positive enrichment, 65 gene sets had $q - value < 0.05$. Leading edge analysis reveals the clustering of the spliceosome, oxidative phosphorylation and protein transport.

Table 4.2: Top Whole Blood Unique Gene Sets

NAME	SIZE	FDR q-val
GSEA34205_RSV_VS_FLU_INF_INFANT_PBMC_UP	97	0.006
CATALYTIC_STEP_2_SPLICEOSOME	76	0.008
U2_TYPE_SPLICEOSOMAL_COMPLEX	23	0.011
SPLICESOME	113	0.013
U12_SPLICESOMAL_COMPLEX	25	0.015

The result of GSEA for both bone marrow analysis and whole blood analysis was compared. 6 gene sets were enriched in both the bone marrow and whole blood. Oxidative phosphorylation related gene sets and ribosomal related gene sets constitute the majority of genes sets enriched in both whole blood transcriptome response and bone marrow transcriptome response. 17 gene sets were uniquely enriched in whole blood transcriptome response, the top 5 gene sets according to q-value are shown in Table 4.2. 7 out of these 17 unique gene sets are related with spliceosome function. Association of malaria infection with the change in spliceosome related genes has recently been reported by Yamagishi *et al.* [15]. Top unique gene sets of bone marrow transcriptome are shown in Table 4.3.

Table 4.3: Top Bone Marrow Unique Gene Sets

NAME	SIZE	FDR q-val
CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	42	0.000
GSE2405_0H_VS_9H_A_PHOGOCYTOPHILUM_STIM_NEUTROPHIL_DN	177	0.000
3_UTR_MEDIATED_TRANSLATIONAL_REGULATION	96	0.000
PEPTIDE_CHAIN_ELONGATION	79	0.000
SRP_DEPENDENT_COTRANSLATIONAL_PROTEIN_TARGETING_TO_MEMBRANE	101	0.000

4.6 Discussion

In this paper, we presented a novel differential network analysis tool. The identification of differential networks provides context for downstream functional analysis. We were able to apply this framework to explore differential transcriptomic network resulting from *P. cynolmogi* infection. We discovered specific pathways that characterize bone marrow and whole blood transcriptional response to *P. cynolmogi* infection.

eDiNA discovers differential networks through the ensemble of four measurements: Pearson's Correlation, Spearman's Correlation, χ^2 statistics and ℓ_1 distance. Each individual measurement only describes one aspect of pair-wise dynamics, the ensemble method leverages the strength of each. The ability of eDiNA to provide biological insight of transcriptional response to malaria was demonstrated through its application to the MaHPIC study. Our analysis identified 17 gene sets that were uniquely enriched in whole blood transcriptome response to malaria infection. The splicesome related perturbation identified by our study has also been independently reported by Yamagishi *et al.* [15]. Furthermore, our analysis of bone marrow transcriptional response suggests the activation of the immune response in the bone marrow during *P. cynolmogi* infection. Interestingly, mitochondrial

and oxidative phosphorylation-related gene sets are perturbed in both the bone marrow and whole blood, suggesting critical function of mitochondria during host response to parasite invasion [9].The framework of eDiNA is based on Fisher's combined p-value test and can be easily expanded to other -omic data types and to include other measurements of pairwise dynamics.

In conclusion, eDiNA provides the framework to leverage the strength of many different differential network analysis methods and allow for the identification of perturbed pairwise dynamics that are overlooked by a singular quantification method.

Acknowledgments

This project was funded in part by Federal funds from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract #HHSN272201200031C (PI: Mary Galinski), which supports the Malaria Host-Pathogen Interaction Center (MaHPIC).

4.7 References

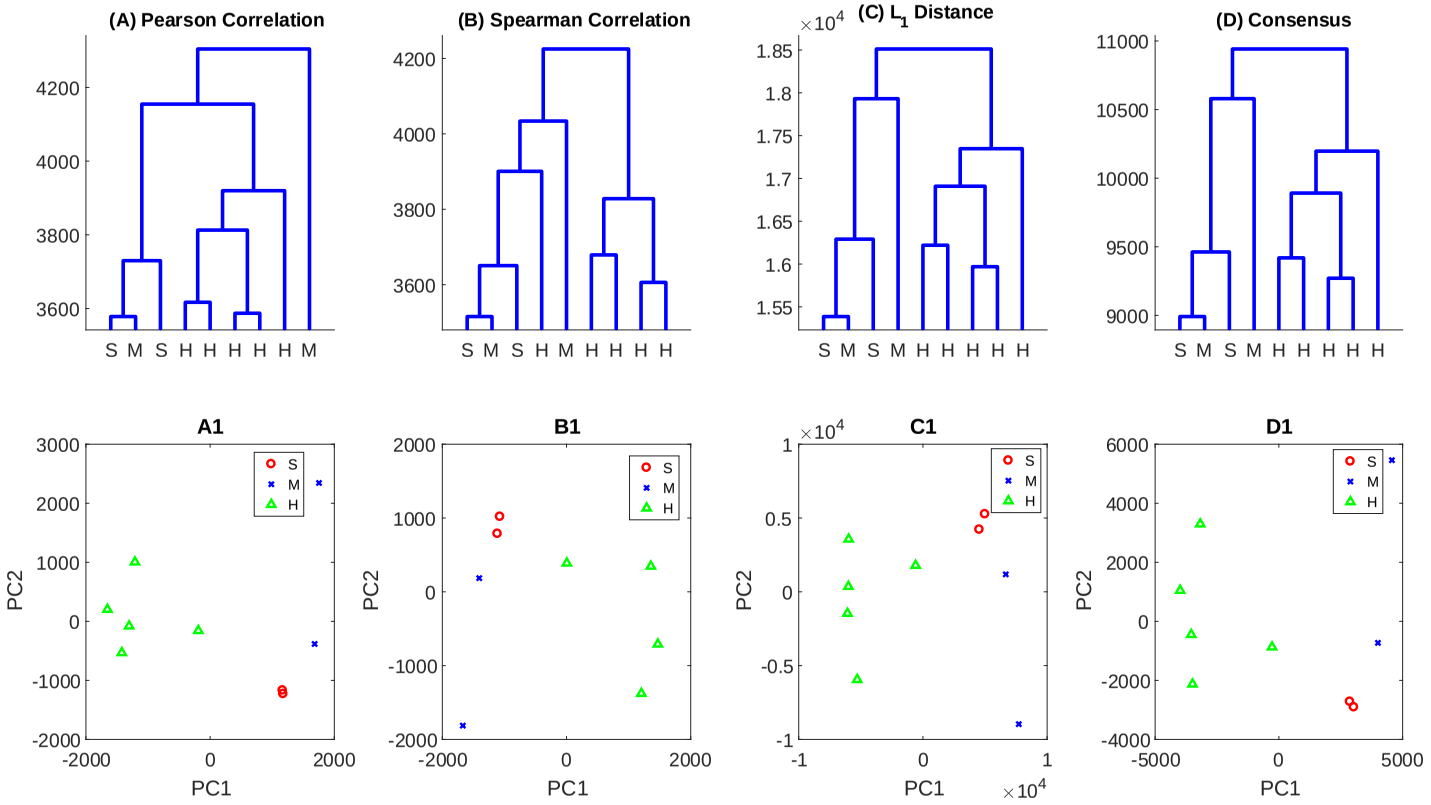

Figure 4.4: Hierarchical Clustering and PCA analysis of bone marrow transcriptomic data separate healthy and disease states. The three states are healthy (H), severe disease (S) and mild disease (M).

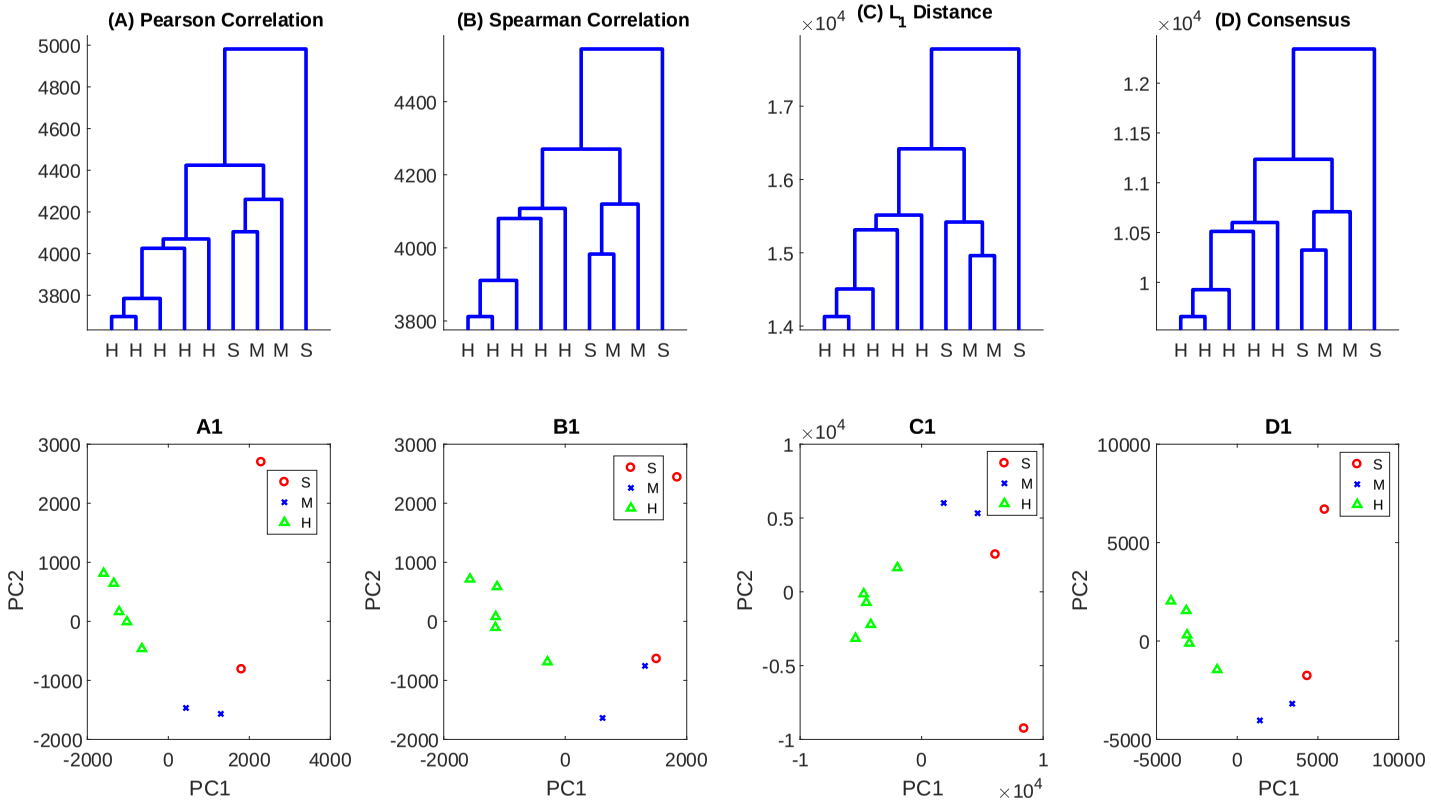

Figure 4.5: Hierarchical Clustering and PCA analysis of whole blood transcriptomic data separate healthy and disease states. The three states are healthy (H), severe disease (S) and mild disease (M).

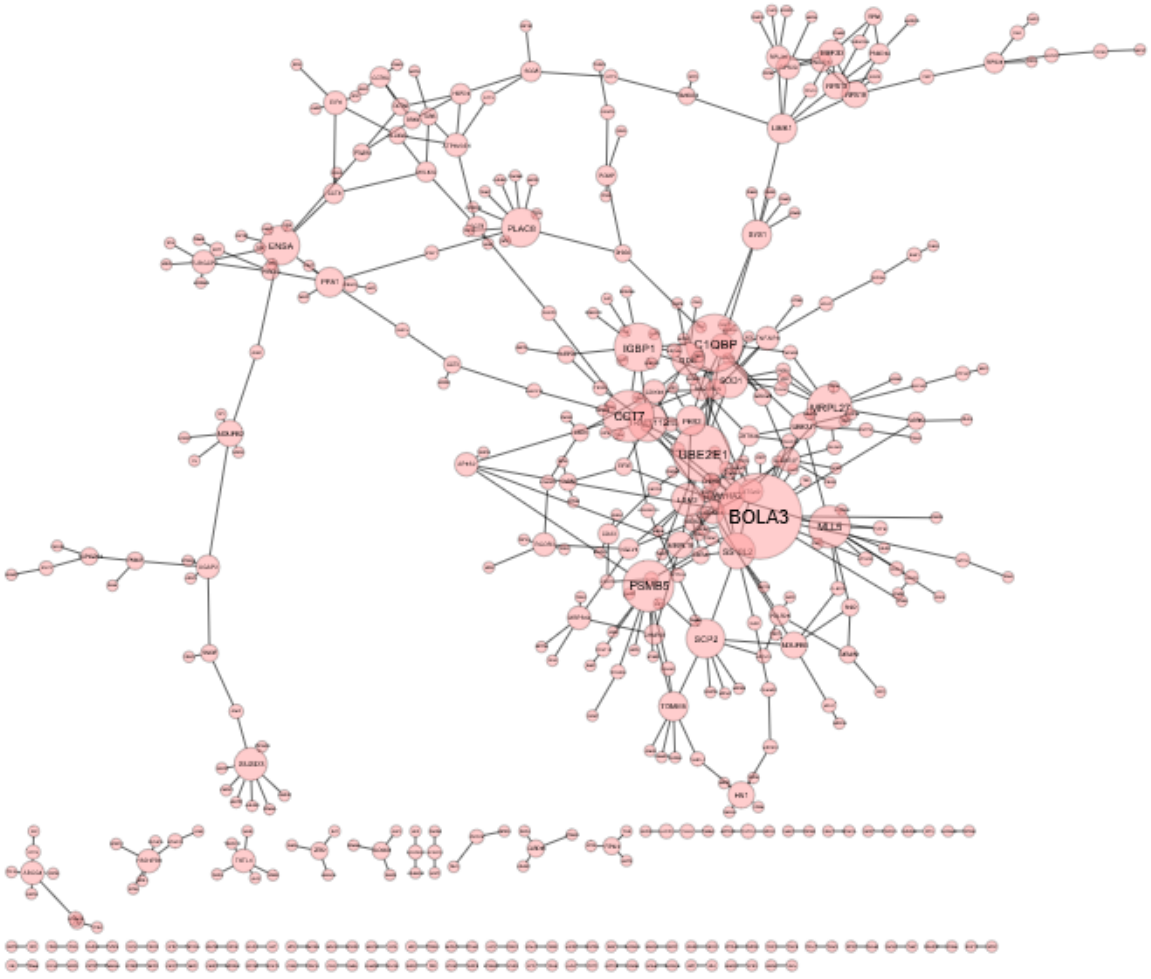

Figure 4.6: Visualization of differential pairwise gene dynamics with q value $< 10^{-23}$. Each edge represents a differential pairwise gene dynamic. Each node represents a gene and the size of a node is proportional to the degree of the node.

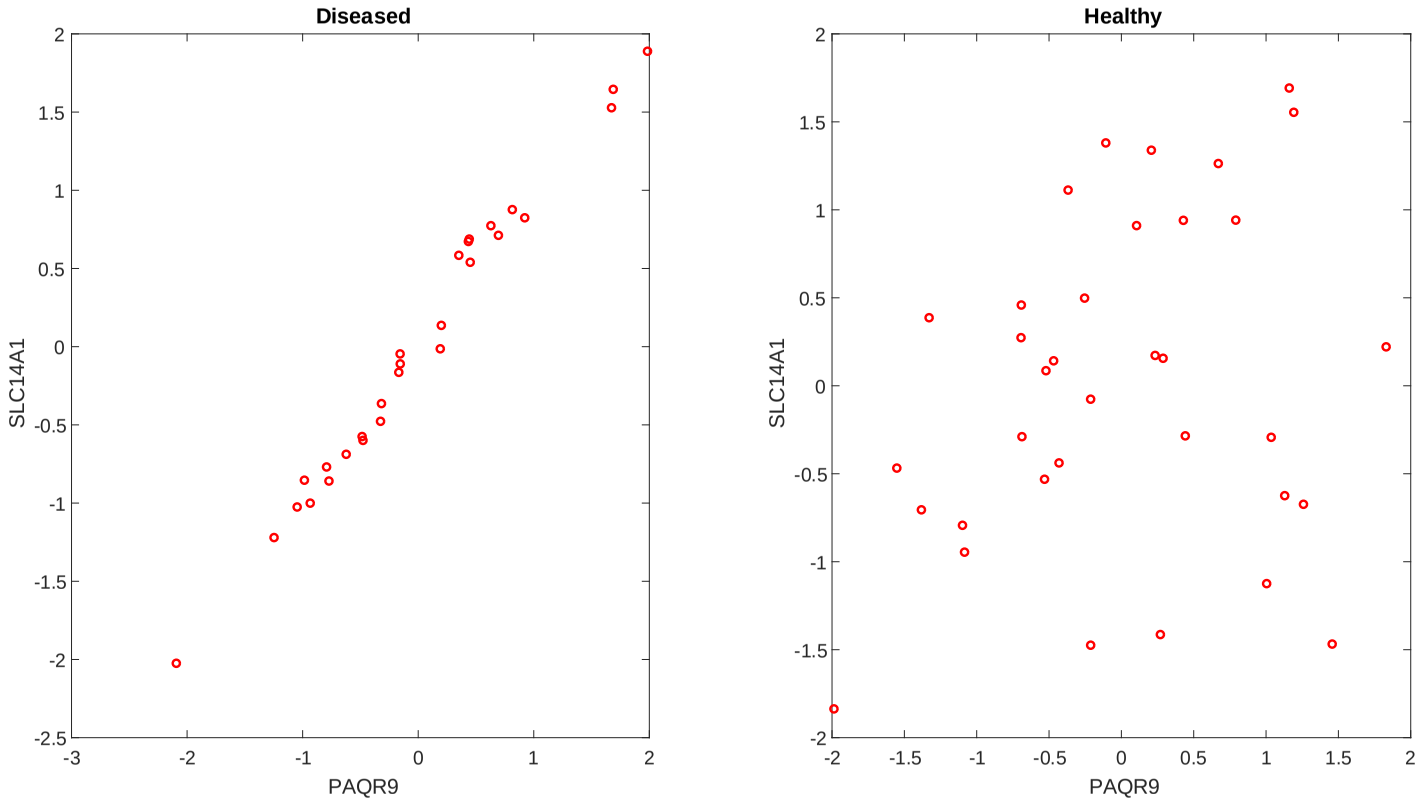

Figure 4.7: Differential dynamics of SLC14A1-PAQR9 between healthy and disease hosts.

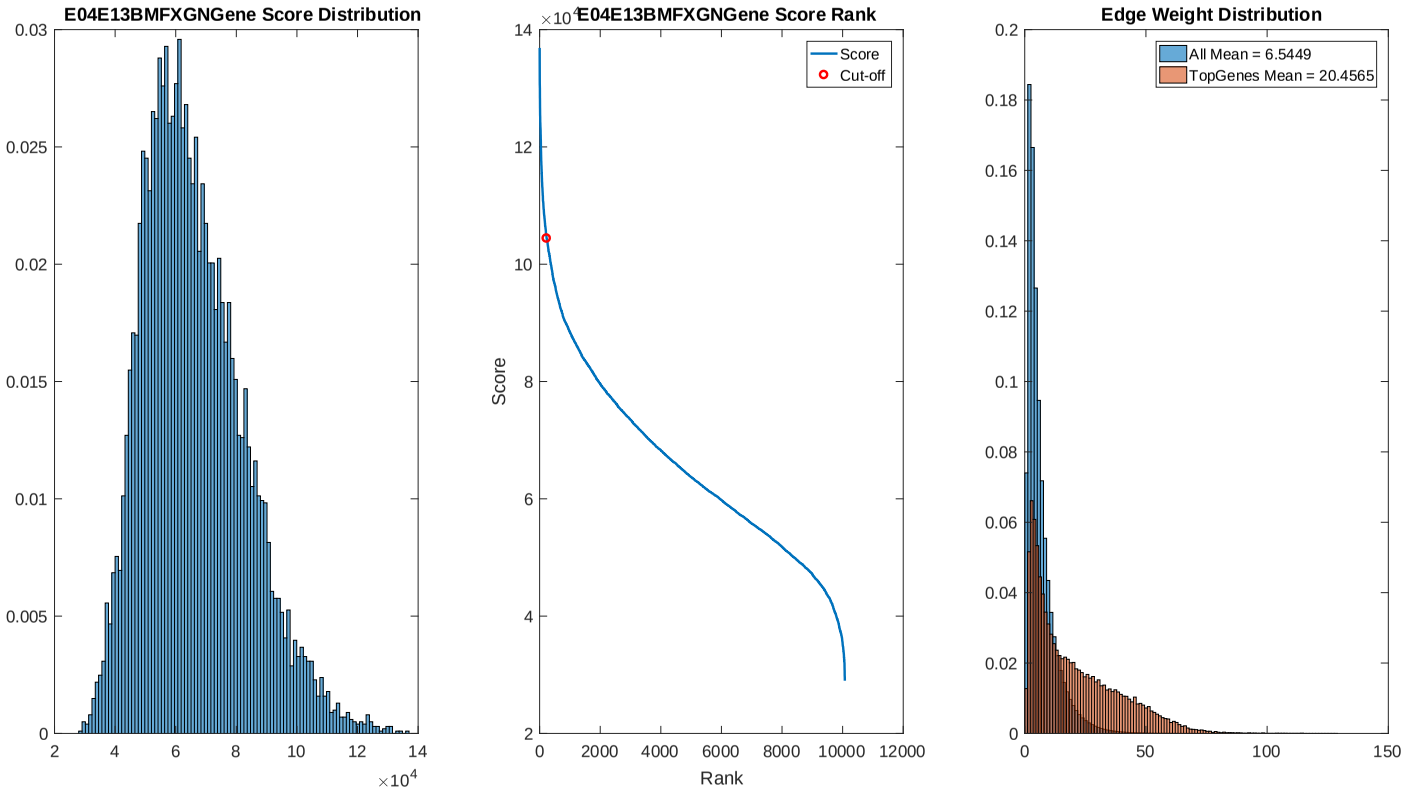

Figure 4.8: A: Histogram of PS, B: Ranked PS, C: Global Edge Score Distribution and Edge Score Distribution among Representative Genes.

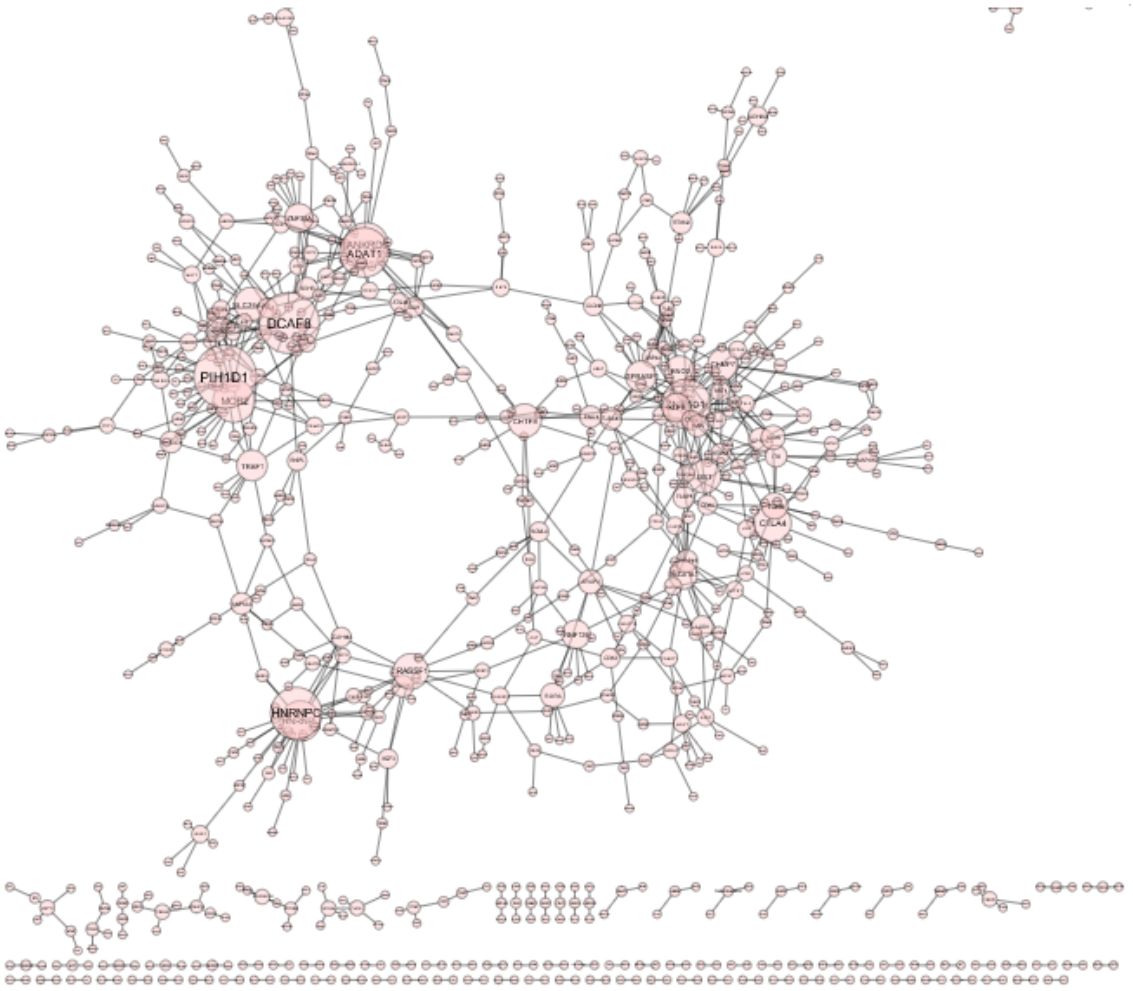

Figure 4.9: Visualization of differential pairwise gene dynamics with q value $< 10^{-23}$. Each edge represents a differential pairwise gene dynamic. Each node represents a gene and the size of a node is proportional to the degree of the node.

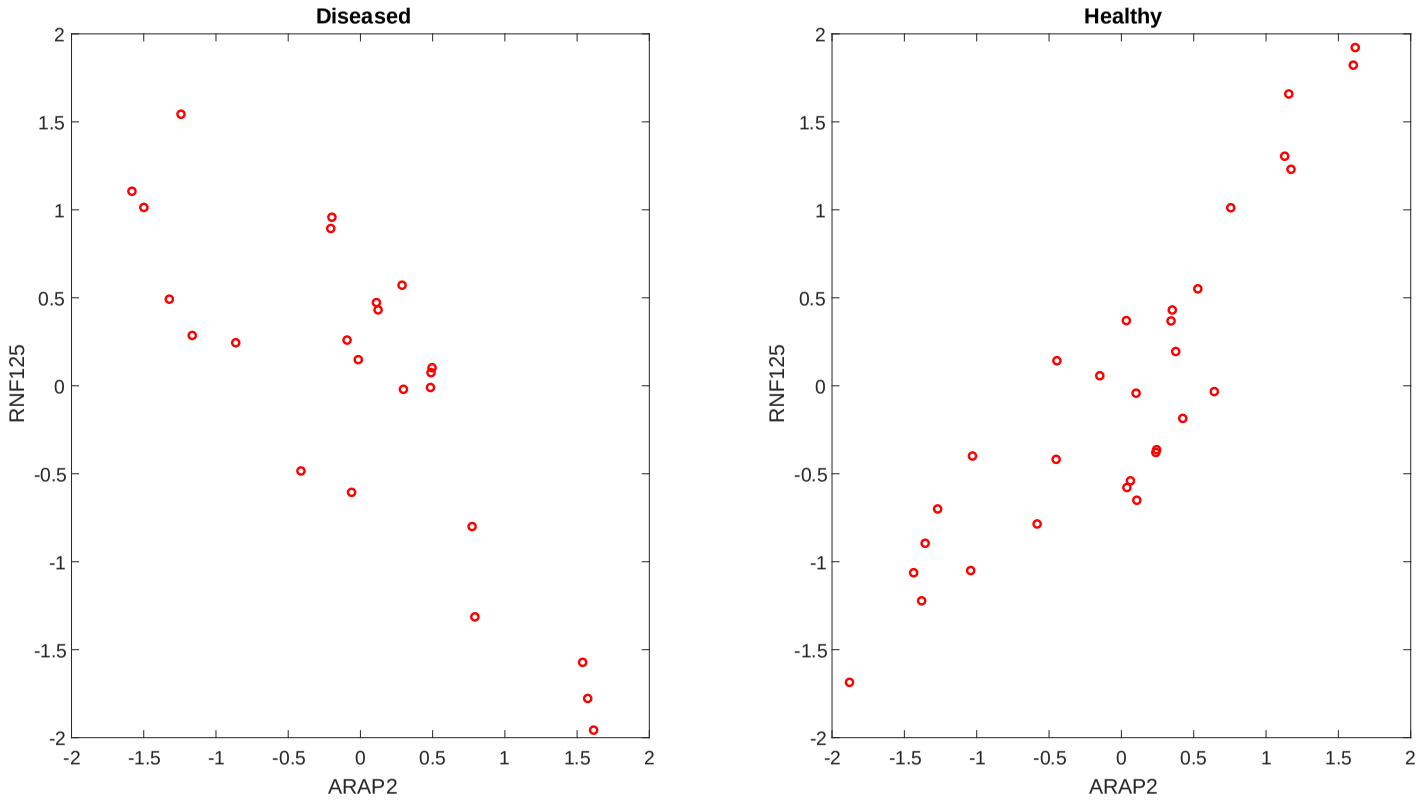

Figure 4.10: Differential dynamics of ARAP2-RNF125 between healthy and disease hosts.

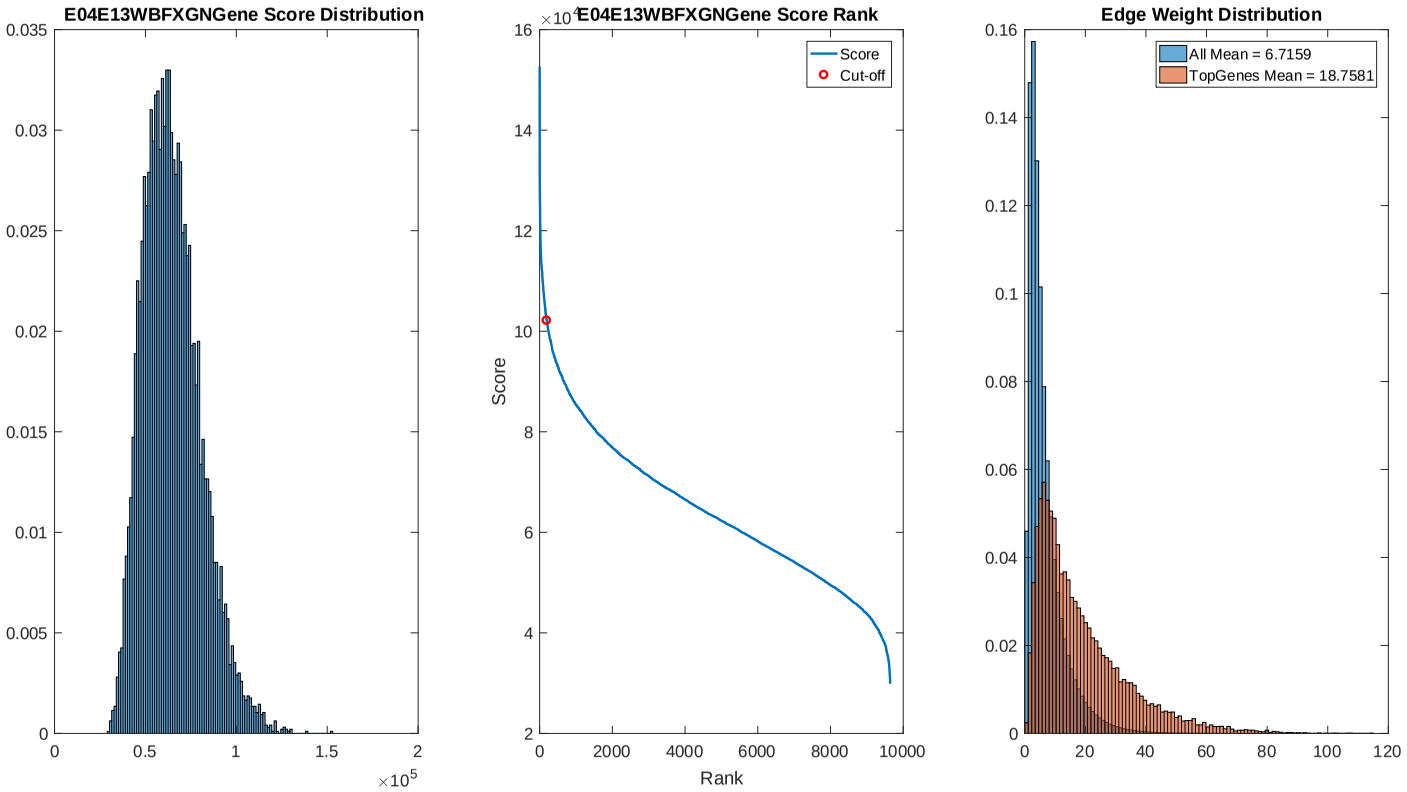

Figure 4.11: A: Histogram of PS, B: Ranked PS, C: Global Edge Score Distribution and Edge Score Distribution among Representative Genes.

Bibliography

[1] K. Artavanis-Tsakonas, J. Tongren, and E. Riley. The war between the malaria parasite and the immune system: immunity, immunoregulation and immunopathology. *Clinical & Experimental Immunology*, 133(2):145–152, 2003.

[2] J. K. Choi, U. Yu, O. J. Yoo, and S. Kim. Differential coexpression analysis using microarray data and its application to human cancer. *Bioinformatics*, 21(24):4348–4355, 2005.

[3] R. A. Fisher. On the probable error of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[4] M. J. Ha, V. Baladandayuthapani, and K.-A. Do. Dingo: differential network analysis in genomics. *Bioinformatics*, 31(21):3413–3420, 2015.

[5] T.-H. Hsiao, Y.-C. Chiu, P.-Y. Hsu, T.-P. Lu, L.-C. Lai, M.-H. Tsai, T. H.-M. Huang, E. Y. Chuang, and Y. Chen. Differential network analysis reveals the genome-wide landscape of estrogen receptor modulation in hormonal cancers. *Scientific reports*, 6, 2016.

[6] C. Joyner, A. Moreno, E. V. Meyer, M. Cabrera-Mora, J. C. Kissinger, J. W. Barnwell, and M. R. Galinski. Plasmodium cynomolgi infections in rhesus macaques display clinical and parasitological features pertinent to modelling vivax malaria pathology and relapse infections. *Malaria Journal*, 15(1):451, 2016.

[7] J. Langhorne, F. M. Ndungu, A.-M. Sponaas, and K. Marsh. Immunity to malaria: more questions than answers. *Nature immunology*, 9(7):725–732, 2008.

[8] K. J. Lee, W. Yin, D. Arafat, Y. Tang, K. Uppal, V. Tran, M. Cabrera-Mora, S. Lapp, A. Moreno, E. Meyer, et al. Comparative transcriptomics and metabolomics in a rhesus macaque drug administration study. 2014.

[9] S. Luckhart, N. Pakpour, and C. Giulivi. Host–pathogen interactions in malaria: cross-kingdom signaling and mitochondrial regulation. *Current opinion in immunology*, 36:73–79, 2015.

[10] W. H. Organization et al. World malaria report 2016. *Geneva: WHO. Embargoed until*, 13, 2016.

[11] D. J. Perkins, T. Were, G. C. Davenport, P. Kempaiah, J. B. Hittner, and J. M. Ong'echa. Severe malarial anemia: innate immunity and pathogenesis. *International journal of biological sciences*, 7(9):1427, 2011.

[12] J. D. Storey, J. E. Taylor, and D. Siegmund. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: a unified approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1):187–205, 2004.

[13] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[14] K. Wang, M. Saito, B. C. Bisikirska, M. J. Alvarez, W. K. Lim, P. Rajbhandari, Q. Shen, I. Nemenman, K. Basso, A. A. Margolin, et al. Genome-wide identification of post-translational modulators of transcription factor activity in human b cells. *Nature biotechnology*, 27(9):829–837, 2009.

[15] J. Yamagishi, A. Natori, M. E. Tolba, A. E. Mongan, C. Sugimoto, T. Katayama, S. Kawashima, W. Makalowski, R. Maeda, Y. Eshita, et al. Interactive transcriptome

analysis of malaria patients and infecting plasmodium falciparum. *Genome research*, 24(9):1433–1444, 2014.

[16] Y. H. Yan, E. D. Trippe, and J. B. Gutierrez. A method for massively parallel analysis of time series. *arXiv preprint arXiv:1612.08759*, 2016.

[17] J. Ylostalo, A. C. Randall, T. A. Myers, M. Metzger, D. J. Krogstad, and F. B. Cogswell. Transcriptome profiles of host gene expression in a monkey model of human malaria. *Journal of Infectious Diseases*, 191(3):400–409, 2005.

Chapter 5

Correlates of severity of disease in *Macaca mulatta* infected with *Plasmodium cynomolgi*[1]

[1]Yi H. Yan, Diego M. Moncada, Elizabeth D. Trippe, and Juan B. Gutierrez. To be submitted to *Nature Scientific Report*

5.1 Abstract

Characterization of host responses associated with severe malaria through an integrative approach is necessary to understand the dynamics of a *Plasmodium cynomolgi* infection. In this study, we conducted temporal immune profiling, cytokine profiling and transcriptomic analysis of five *Macaca mulatta* infected with *P. cynomolgi*. This experiment resulted in two severe infections, and two mild infections. Our analysis reveals that differential transcriptional up-regulation of genes linked with response to pathogen-associated molecular pattern (PAMP) and pro-inflammatory cytokines is characteristic of hosts experiencing severe malaria. Furthermore, our analysis discovered associations of transcriptional differential regulation unique to severe hosts with specific cellular and cytokine responses. The combined data provide a molecular and cellular basis for the development of severe malaria during *P. cynomolgi* infection.

5.2 Introduction

Malaria remains a public health challenge, responsible for approximately 400,000 (236,000635,000) death in 2015 [19]. Out of the five human *Plasmodium* species capable of causing malaria, *Plasmodium falciparum* and *Plasmodium vivax* account for the majority of human malaria infections. *P. falciparum* is responsible for the majority of malaria-related mortality and is most prevalent in sub-Saharan Africa [19]. Although *P. vivax* infection does not typically result in mortality, this parasite causes substantial morbidity outside of sub-Saharan Africa and exhibits the propensity to cause severe disease. The mechanisms that underly vivax malaria pathogenesis, however, remain poorly understood partially due to experimental constraints such as the lack of an *in vitro* culture system and a rodent model of *P. vivax* infection [7, 16, 17].

Plasmodium cynomolgi is a non-human primate parasite that infects old world monkeys and is capable of recapitulating clinical and histopathological findings of vivax malaria

patients[5, 6, 23]. It is both genetically and physiologically similar to *P. vivax* [23, 27]. For instance, both parasites exhibit 48-h erythrocytic cycle during blood stage infection [3], preferential infection of reticulocytes [26] and form hypnozoites, which are dormant, liver stage forms that can activate and cause relapse infections [12]. Due to the difficulty of studying *P. vivax* pathogenesis, *P. cynomolgi* infection of rhesus monkeys (*Macaca mulatta*) has being used to better understand hypnozoite-caused relapse [4].

Host clearance of malaria parasites without complication requires a concerted effort between inflammatory and anti-inflammatory cytokines; their balance and timing are critical in determining clinical outcomes [8]. The association between cytokine, transcriptomic and immune response, and clinical outcomes has been extensively studied in *P. falciparum* infections [2, 20, 24, 25, 28], in comparison, much less is known for *P. vivax* infections.

With the aim to better characterizing *P. vivax* infection in humans, a time series experiment where four *Macaca mulatta* were infected with *P. cynomolgi* was conducted as a part of the Malaria Pathogen Host Interaction Center (MaHPIC) project [10]. This experiment captured host transcriptomic, cellular and cytokine response to *P. cynomolgi* infection. The subjects within this study responded to the infection in different manners and resulted in two cases of severe malaria, and two cases of mild malaria. The presence of both severe malaria and mild malaria provided us the opportunity to characterize host responses associated with clinical outcomes.

Here we present the combined analysis of cellular, transcriptome and cytokine data collected during this experiment with an emphasis on comparing host responses during severe malaria and mild malaria. Our analysis shows associations of severe malaria with differential up-regulation of pro-inflammatory signals independent of adaptive immune response, whereas moderate malaria is associated with transcriptomic up-regulation of complement and heme metabolism related genes. These distinct characteristics demonstrate that *P. cynomolgi* induced severe malaria is strongly associated with up-regulation of inflammatory signaling through transcriptional regulation.

5.3 Results

5.3.1 Study Design

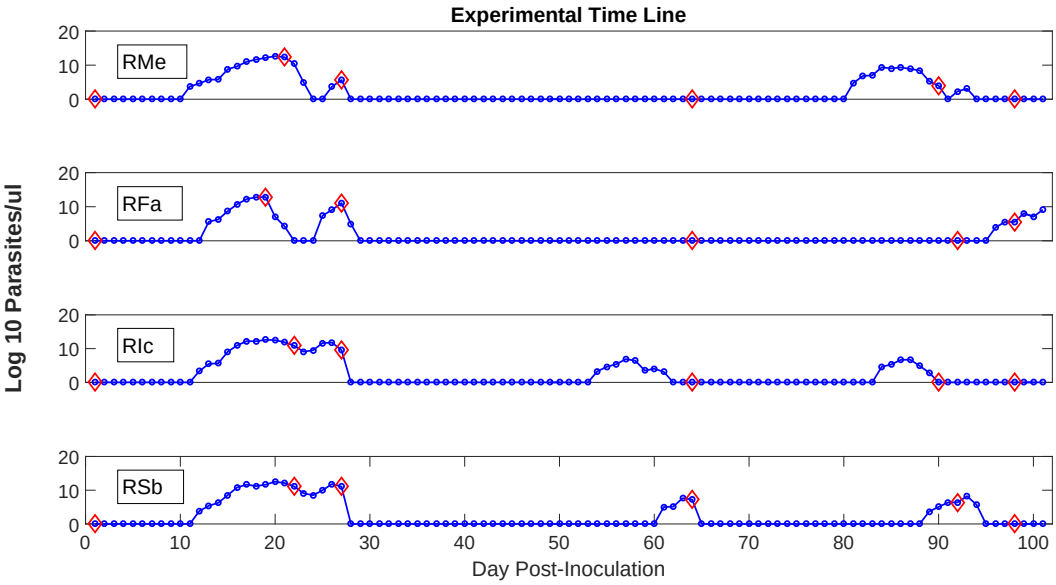

Figure 5.1: Parasite count over the entire experiment. Transcriptome, cytokine, and immune profiling were conducted during the six time points marked with red diamonds. Subject S1 and S2 experienced severe malaria symptom. Subject M1 and M2 experienced mild malaria symptom.

The experimental design is shown in Figure 5.1. Transcriptomic, cellular and cytokine data were measured from peripheral blood samples collected during six time points. Time point 1 (TP1) corresponds to the pre-infection baseline for each subject. Time point 2 (TP2) corresponds to the acute primary infection when parasite counts peaked. At time point 2, two subjects, S1 and S2 developed signs of severe malaria including anemia and thrombocytopenia, which required each animal to be sub-curatively treated to prevent possible severe complications. The other two subjects M1 and M2 experienced mild disease manifestations and did not receive sub-curative treatment. All animals received curative blood-stage treatment with artemether on Day 28 to ensure that subsequent parasitemias were due to relapses and not recrudescing parasitemias. A detailed description of the experimental set-up can be found in the publication by Joyner *et al.* [10].

5.3.2 Differential Gene Regulation during Primary Infection

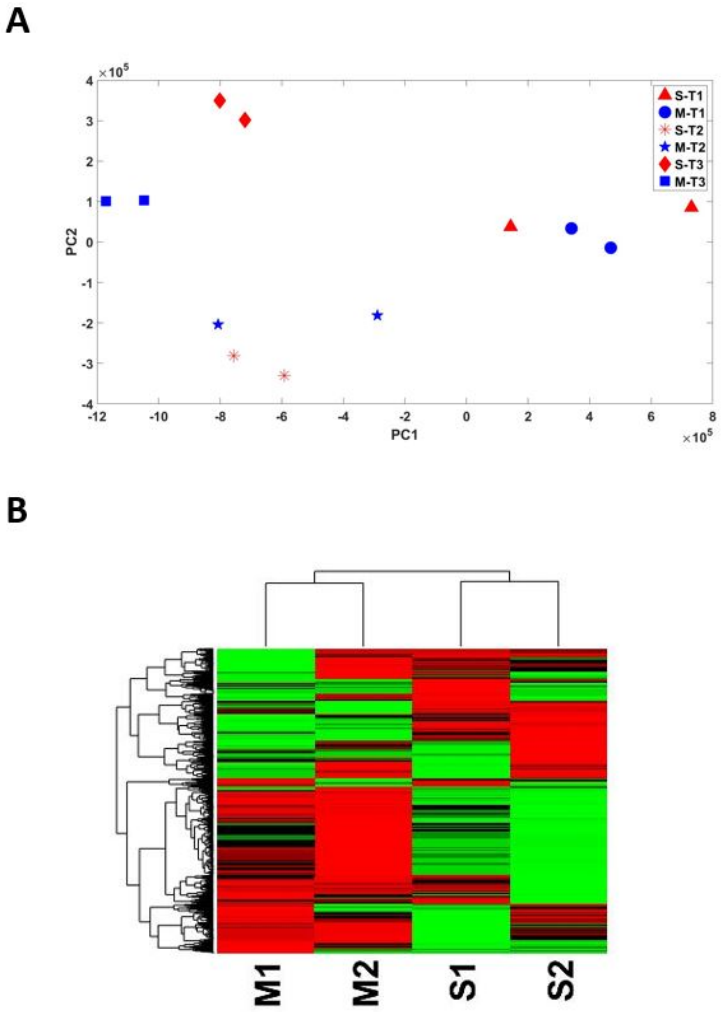

Figure 5.2: Clinical classes were defined as in Figure 5.1. (A) Principal component analysis of the combined -omic data (Transcriptome, cytokine and immune cell counts). (B) An unsupervised hierarchical heat map of the top 50% most variable genes at time point 2. Subjects experiencing severe malaria symptoms are labeled S1 and S2, subjects experiencing mild malaria symptoms are labeled M1 and M2.

To characterize the differences of transcriptomic responses induced by *P. cynomolgi* infection between the subjects experiencing severe malaria (severe hosts) and the ones experiencing mild malaria (mild hosts), we conducted RNA-Seq analysis for each clinical group between TP1 and TP2 using the DESeq algorithm [1]. Furthermore, between-group transcriptome analysis was also carried out for TP2. The transcriptome profiles differ the most

from the healthy baseline (TP1) at TP2 and TP3 shown by principal component analysis (Figure 5.2). Additionally, the transcriptome profiles of the severe hosts were more similar to each other than to the mild hosts at TP2 as shown by both principal component analysis and hierarchical clustering using the top 50% most variably expressed genes (Figure 5.2).

Mild hosts showed similar degrees of transcriptome perturbation during the onset of primary infection (698 DEGs using an absolute fold-change threshold of > 1.5 and q-value < 0.05) relative to severe hosts (511 DEGs). 148 genes were identified by both analyses. At TP2, 1337 genes were differentially expressed between the mild hosts and severe hosts. 380 of these DEGs were temporally differentially expressed in severe hosts between TP1 and TP2. 220 DEGs were temporally differentially expressed in mild hosts between TP1 and TP2.

To identify differentially regulated transcriptome responses to *P. cynomolgi* between severe hosts and mild hosts, DEGs were classified into four groups: differentially up-regulated genes in severe hosts (DUGs-S), differentially up-regulated genes in mild hosts (DUGs-M), differentially down-regulated genes in severe hosts (DDGs-S) and differentially down-regulated genes in mild hosts (DDGs-M). Genes were classified into each of these groups based on their between group differential expression at TP2 and their with-in group differential expression between TP1 and TP2. The specific classification scheme is shown in Figure 5.3 and Figure 5.4.

5.3.3 Biological Pathways Perturbed During Primary Infection

To identify the biological pathways associated with each of the four groups of differentially regulated genes, we used MySigDB to conduct enrichment analysis [13]. Using a q-value cut-off of 0.05, DUGs-S are enriched in 205 pathways with the most significantly enriched pathway being cytokine signaling in the immune system. DDGs-S are enriched in only 24 pathways (Figure 5.3) with the most significantly enriched pathway being immunoregulatory interactions between lymphocyte cells (LCS) and non-lymphocyte cells (Non - LCS). DUGs-M is enriched in 51 pathways with the most significantly enriched pathway being

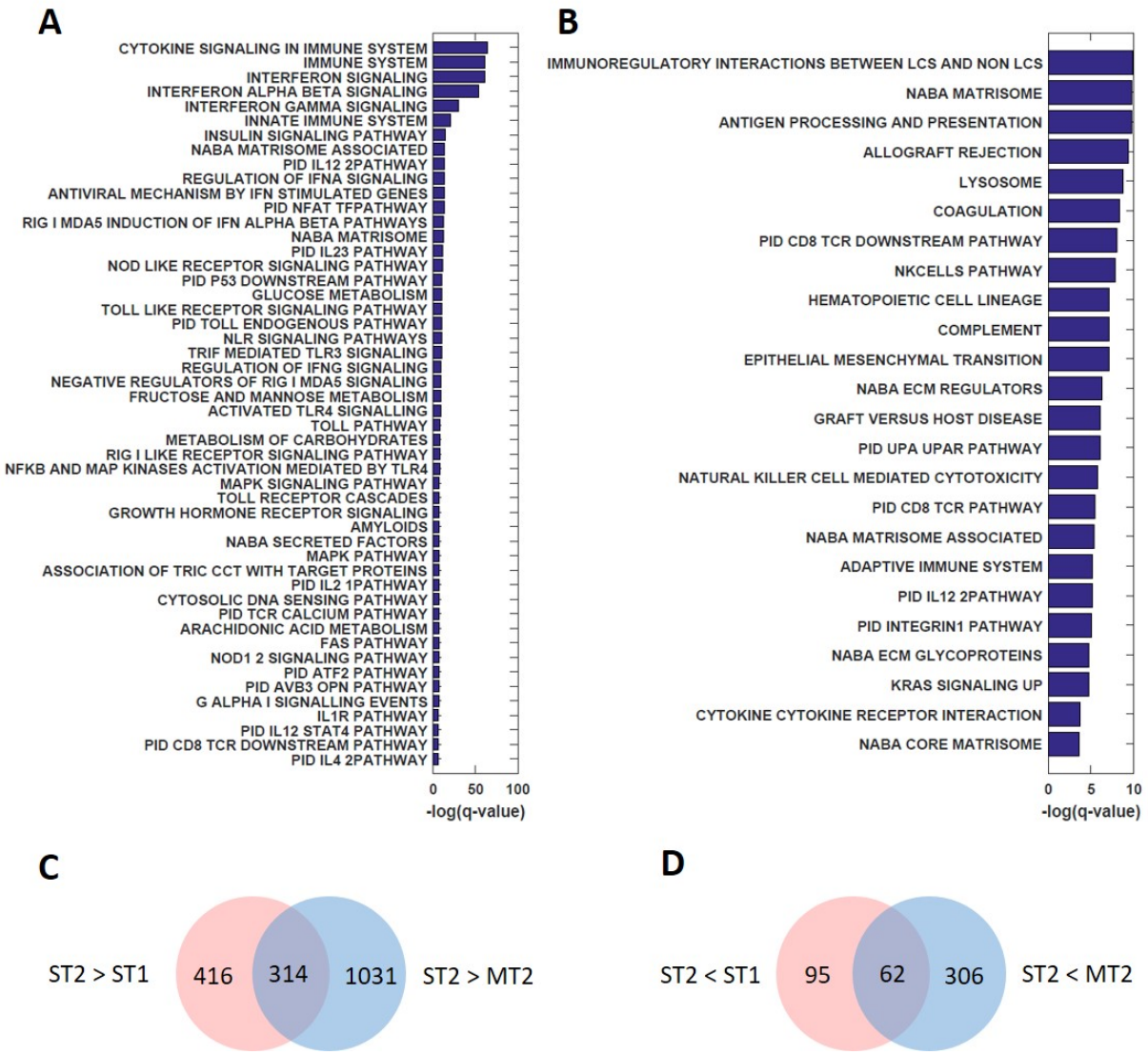

Figure 5.3: (A) Enrichment analysis using DUGs-S with q-value < 0.05. (B) Enrichment analysis using DDGs-S with q-value < 0.05. (C) and (D) Venn diagrams showing the identification of DUGs-S and DDGs-S. $ST2 > ST1$ refers to genes within severe hosts that are up-regulated at TP 2 in comparison to TP 1. $ST2 > MT2$ refers to genes that are up-regulated in severe hosts in comparison to mild hosts at TP2. $ST2 < ST1$ refers to genes within severe hosts that are down-regulated at TP 2 in comparison to TP 1. $ST2 < MT2$ refers to genes that are down-regulated in severe hosts in comparison to mild hosts at TP2.

heme metabolism, and DDGs-M are enriched in 6 pathways. The number of enriched pathways within each gene set is proportional to the gene set size. Both DUGs-S and DUGs-M

show enrichment in innate immune-related response and interferon gamma related response. However, complement and coagulation-related genes are differentially up-regulated in the mild hosts and differentially down-regulated in the severe hosts.

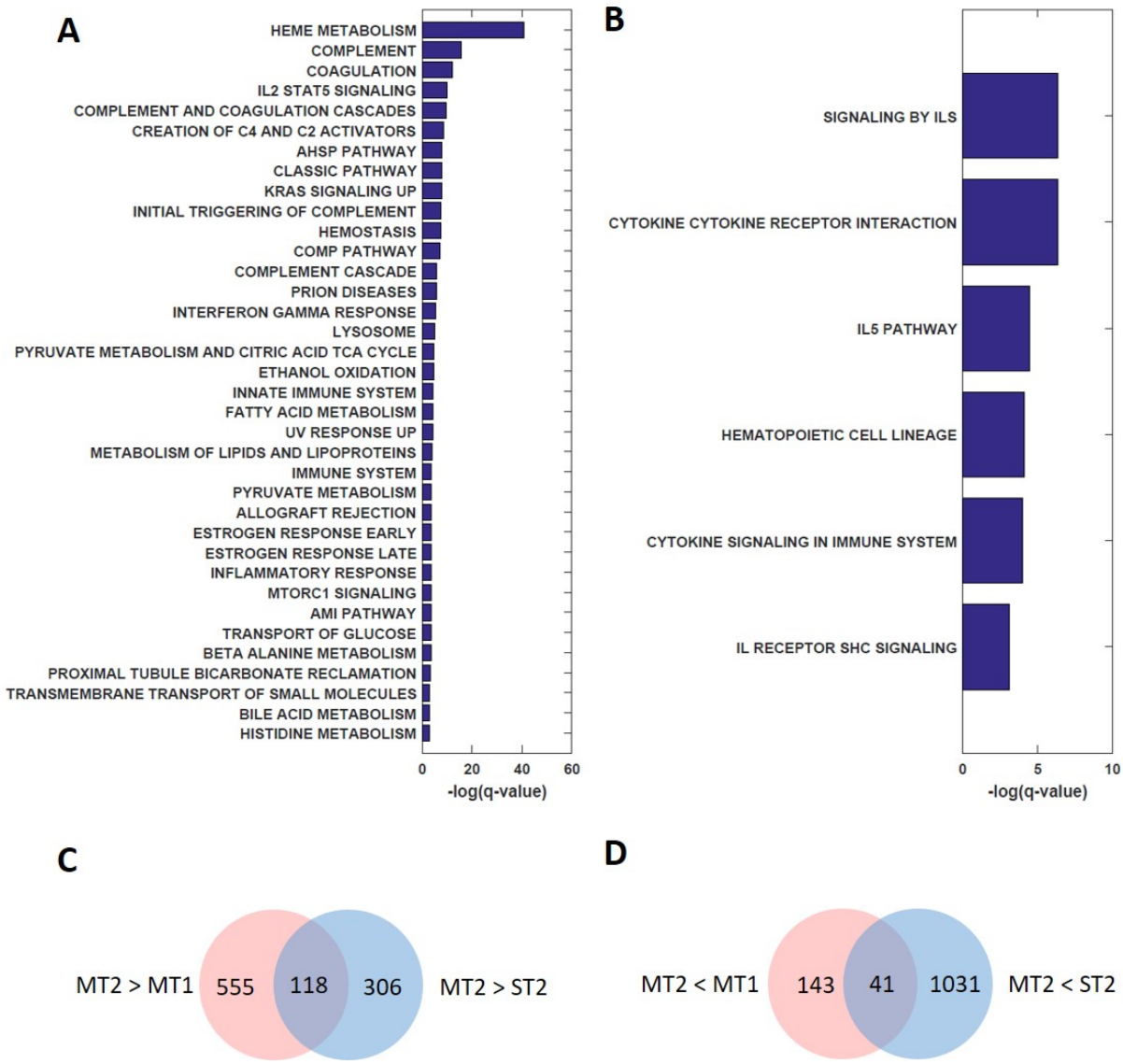

Figure 5.4: (A) Enrichment analysis using DUGs-M with q-value < 0.05. (B) Enrichment analysis using DDGs-M with q-value < 0.05. (C) and (D) Venn diagrams showing the identification of DUGs-M and DDGs-M. $MT2 > MT1$ refers to genes within mild hosts that are up-regulated at TP 2 in comparison to TP 1. $MT2 > ST2$ refers to genes that are up-regulated in mild hosts in comparison to severe hosts at TP2. $MT2 < MT1$ refers to genes within mild hosts that are down-regulated at TP 2 in comparison to TP 1. $MT2 < ST2$ refers to genes that are down-regulated in mild hosts in comparison to severe hosts at TP2.

5.3.4 Differentially Regulated Genes Demonstrate Distinct Correlation Pattern with Cell Populations and Cytokine Abundance

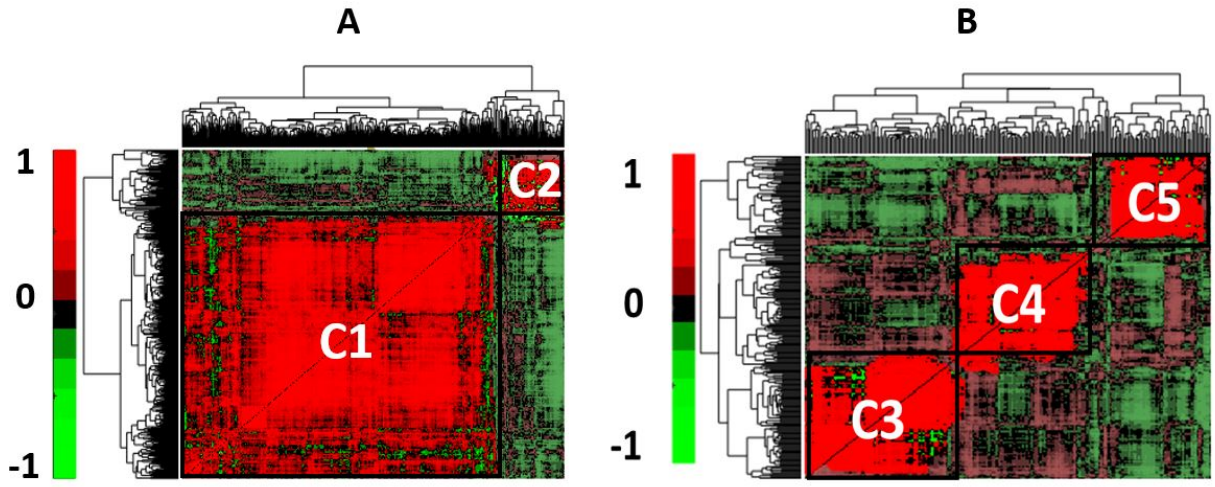

Figure 5.5: (A) An unsupervised hierarchical heat map of the pairwise Spearman's correlation among DUGs-S, cell populations, and cytokine abundances. (B) Heat map of the pairwise Spearman's correlation among DUGs-M, cell populations, and cytokine abundances.

To identify the similarities among temporal profiles of DUGs-S, DUGs-M, cytokine abundances and cell populations, we constructed two correlation matrices and identified 5 clusters (Figure 5.5). DUGs-S and DUGs-M exhibited distinct correlation patterns with respect to cytokine abundances and cell populations. Two clusters were identified for DUGs-S, all but 11 genes were clustered into C1 and the rest into C2. C1 contained the majority of cytokine abundances along with a few cell populations such as the percentage of CD14 CD16 monocytes out of total monocyte, the percentage of granulocytes out of total lymphocytes and parasite concentration. Furthermore, we performed enrichment analysis of transcriptional factors for C1 and identified a list of enriched transcriptional factor targets. The top 3 transcriptional factor targets enriched are ISRE, IRF7, and STAT5. Incidentally, IRF7 was also observed to be transcriptionally differentially up-regulated in severe hosts. Unlike C1, C2 contained most of the adaptive and innate immune cell populations but very few genes. In fact, the majority ($> 60\%$) of the correlations between the genes in C1 and the adaptive immune cell populations in C2 are negative.

Three clusters were identified for DUGs-M. C3 contained 37 genes and the majority of the pro-inflammatory cytokines along with reticulocyte population and parasite concentration. Enrichment analysis of C3 reveals that it is enriched in heme metabolism related genes. C4, on the other hand, contained the rest of DUGs-M along with caspase 3 CD 8 and CD 4 T cell populations, monocyte and granulocyte populations. C4 is enriched in complement and coagulation-related genes. C5 is composed entirely of adaptive immune cell types such as B cell and traditional T cell populations. Transcription factor analysis of C3 and C4 did not reveal any enrichment of transcription factor targets. Despite the fact that both heme metabolism and complement-related genes are differentially up-regulated at TP2 in mild hosts, their expression patterns are associated with different cell populations and cytokine abundances. Furthermore, their temporal expressions differ over the entire experiment, in contrast to the uniform temporal expression patterns of DUGs-S.

5.3.5 Cytokine Perturbation during Primary Infection

To characterize the differences of cytokine responses induced by *P. cynomolgi* infection between the clinical groups, two-way ANOVA was used to identify significantly changed cytokines. 4 cytokines out of all 44 cytokines measured displayed significantly different abundances between the two clinical groups. All 4 cytokines have a higher abundance in severe hosts at TP2 (Figure 5.6).

To determine whether the differences in cytokine abundances can be attributed to transcription levels, Pearson's correlation coefficients between the time series of each of the cytokines and their respective transcripts were calculated (Figure 5.6, 5.7). The correlations of IL-6, MIG, MIP-1 and SICAM1 (soluble ICAM1) to their transcripts are 0.94, 0.77, 0.83 and 0.48 respectively. The correlation between SICAM1 and its transcript within the severe hosts is 0.59, which is significantly higher than that of the mild hosts (correlation of 0.17). Additionally, MIG, SICAM1, and MIP-1 only differ in cytokine abundances but not transcriptional abundances. The transcription regulators of IL-6, MIG, MIP-1 and SICAM1

were identified using the TRRUST database [9] and analyzed for differential up-regulation in severe hosts. One transcription factor of IL-6, JUN, and one transcription factor of ICAM1, TWIST2, were identified as differentially up-regulated in severe hosts at TP2.

Interestingly, 13 cytokines exhibited negative correlation between their protein abundances and transcriptional abundances. One of these 13 cytokines, IL-2 is transcriptionally differentially up-regulated in mild hosts at TP2. Additionally, all the genes belonging to DUGs-S showed positive correlations to their protein abundances. MCP1 and IL1RA exhibit the highest correlation between transcript abundances and protein abundances. Both IL1-beta and IL1RA showed differential up-regulation in severe hosts at the transcriptional level, however, their transcript-protein correlation differs. IL-1-beta showed a 5 fold transcriptional change which is associated with a 4 fold increase protein abundance. IL-1RA showed 7 fold increase in transcripts but $\sim$ 8000 fold change in protein abundance. The ratio of IL1 to IL1RA is lower in severe hosts (0.0045) in comparison to mild hosts (0.04).

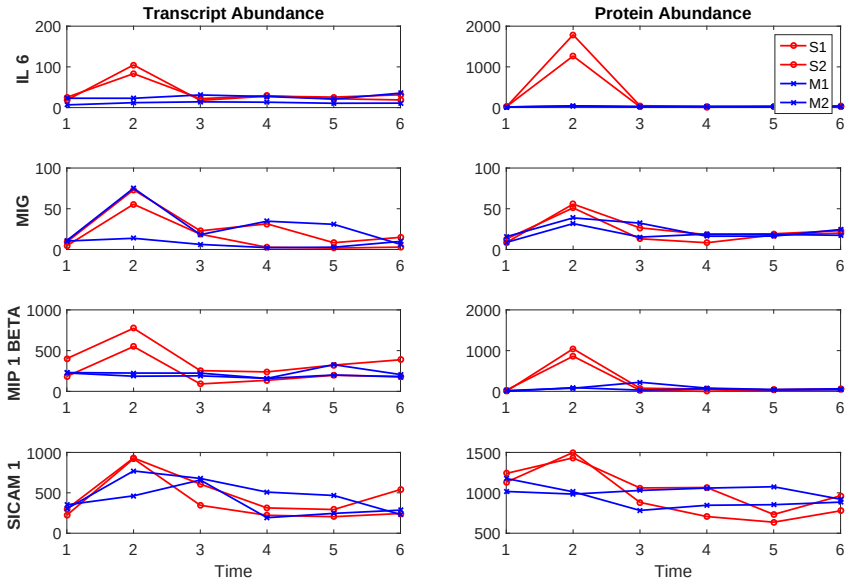

Figure 5.6: Time series data of significantly different cytokines between severe hosts and mild hosts at time point 2. The first column indicates transcript abundance and the second column indicates protein abundance.

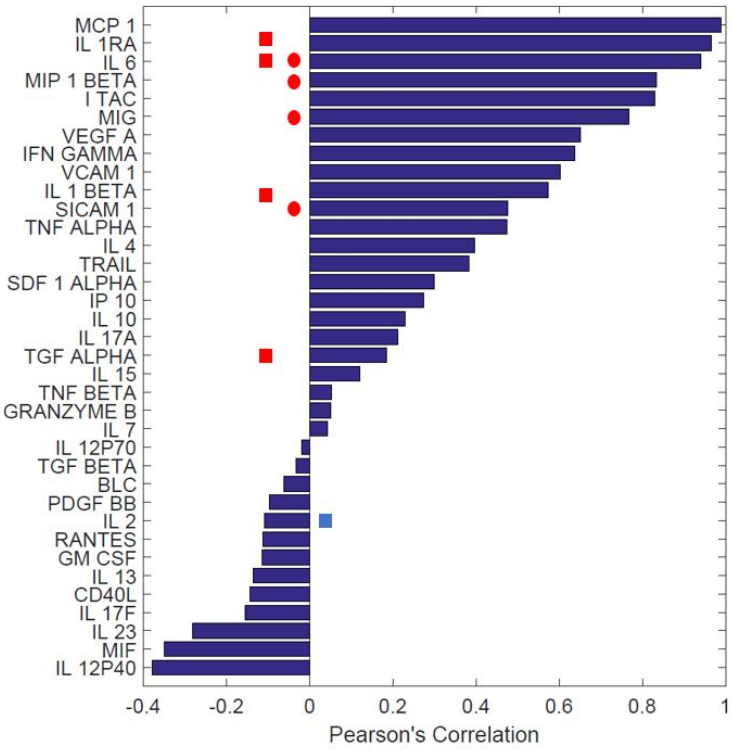

Figure 5.7: Pearson's correlation between cytokine concentrations and their transcript abundances. A red circle indicates protein level differential up-regulation in severe hosts. A red square indicates transcriptional differential up-regulation in severe hosts. A blue square indicates transcriptional differential up-regulation in mild hosts.

5.4 Discussion

In this study, we analyzed temporal transcriptomic, cellular and cytokine profiles of four *M. mulatta* infected with *P. cynomolgi* to characterize the host response to infection. We leveraged the fact that two of the subjects experienced severe malaria during the primary infection and the other two subjects experienced mild malaria to understand the association of host responses with *P. cynomolgi* disease severity. The diverse data types allow us to cross-validate transcriptional response with actual protein abundance in relation to cellular populations.

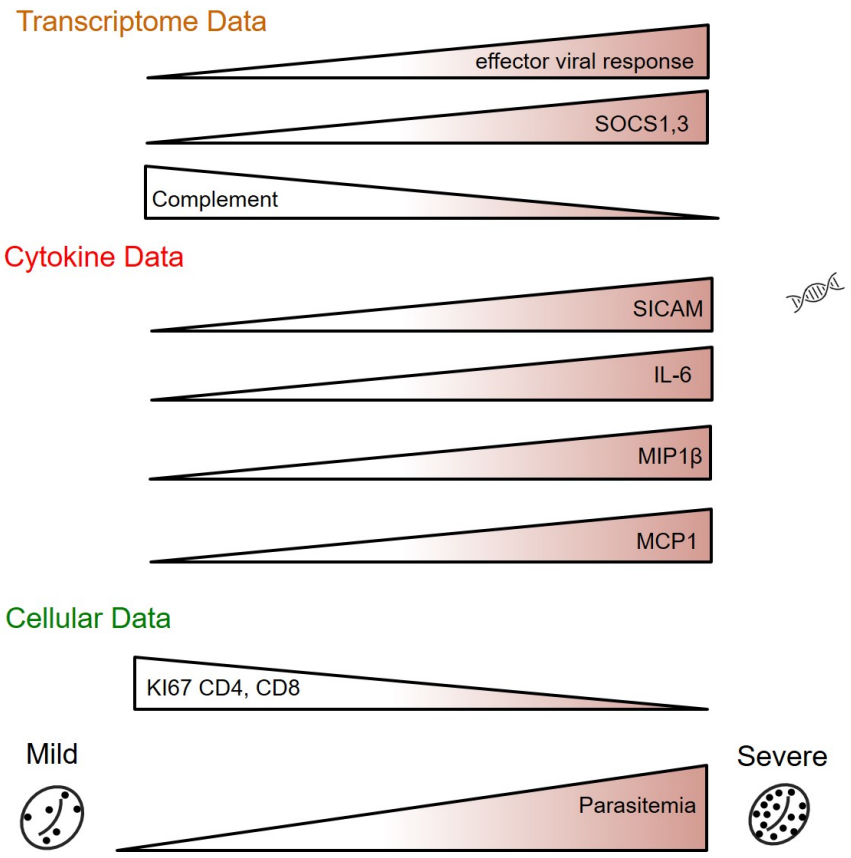

Figure 5.8: Overview of differential regulation of transcriptome, cytokine, and cellular profiles between subject groups.

The identification of differentially up-regulated cytokines and transcripts in the severe hosts combined with the downstream pathway analysis demonstrate that severe malaria within our experiment is associated with elevated pro-inflammatory cytokine levels (MIP, IL6, MIG and SICAM1) and the up-regulation of innate immunity and anti-viral responses related gene sets (TLR Pathway, NOD Pathway, and RIG-I Pathway). Additionally, the positive correlations between the abundance of MIP, IL6, MIG and SICAM1 and their respective transcripts suggest that the differential up-regulation of these four cytokines are associated with transcriptional up-regulation. On the other hand, only two transcription factors that activate the production of these cytokines have been identified to be differentially up-

regulated. The lack of detection of other transcription factors associated with these cytokines could be attributed to the lack of statistical power due to small sample size, or that the up-regulation of other transcription factors happened prior to our sampling. Additionally, most of the transcriptional regulators that control the expression of the acute phase inflammatory cytokines identified here are controlled by post-translational modifications, and thus, an increase in transcription is not alway observed even though signaling may be occurring. Our characterization of severe malaria caused by malaria infection through elevated IL-6 levels and up-regulation of innate and anti-viral related response have also been observed in studies of *P. falciparum* infection severity [14, 15, 20, 25], suggesting likely shared mechanism of recognition of foreign dsRNA and DNA material by the innate immune cells. Correlation analysis of DUGs-S, cell population, and cytokine abundance demonstrates that differential up-regulation of innate immunity and anti-viral responses related gene sets are accompanied by the increase in pro-inflammatory cytokine abundances. The high correlation between cytokine abundance and DUGs-s suggest that the increase in pro-inflammatory cytokines are not independent of transcriptional up-regulation. The negative correlations between DUGs-S and adaptive immune cell types suggest possible innate immune suppression by the adaptive immune system [11, 22] or cross-talk between innate immune response and adaptive immune response.

On the other hand, the mild hosts are characterized by the differential up-regulation of the complement, coagulation and heme metabolism related genes. Interestingly, the complement and coagulation pathways are differentially down-regulated in the severe hosts. Complement activation has been previously linked to children experiencing severe malaria during *P. falciparum* infection [18, 21], however, our result indicate that within the context of *P. cynomolgi* infected *M. mulatta*, elevated levels of complement related transcripts are associated with mild malaria. Correlation analysis of DUGs-M, cell population, and cytokine abundance demonstrates that heme metabolism and complement-related genes are associated with different cell populations and cytokine signals. Complement-related genes are associ-

ated with both innate and adaptive immune cells and heme metabolism related genes are associated with parasitemia and inflammation. Interestingly, SICAM1 which is differentially up-regulated in severe hosts at time point 2 shows a significantly higher transcript-cytokine correlation within the severe hosts than that of the mild hosts, which suggest that the abundance of SICAM1 can be further attributed to post-transcriptional activities within the mild hosts such as less cellular damage, inflammation or endothelial activation. Furthermore, IL-1B was observed to be transcriptionally differentially up-regulated in severe hosts, but the large difference of IL-1/IL1RA ratio between the two clinical groups suggest that IL-1 signaling might actually have been hindered in severe hosts.

In summary, we provided transcriptomic, cellular and cytokine evidence associated with severe disease outcome during *P. cynomolgi* infection in *M. mulatta*. We observed that severe malaria is associated with differential up-regulation of innate viral response related genes and pro-inflammatory cytokine, specifically elevated level of IL-6. Mild malaria is associated with up-regulation of the complement pathway strongly associated with monocyte and neutrophil populations. Our analysis provided a molecular and cellular basis for the development of severe malaria during *P. cynomolgi* infection. Larger and more frequent sampling studies are needed to validate our findings and specifically determine the importance of innate immune activation and the role of complement in controlling disease severity.

5.5 Material and Methods

5.5.1 Experimental Setup and Data Collection

A detailed description of the experimental set up and the generation of cytokine and immune profiles can be found in the publication by Joyner *et al.* [10]. The procedure for RNA-seq data collection and the library size normalized data can be found on Gene Omnibus under series number GSE99486. The daily clinical data of the experiment is stored in PlasmoDB using the identifier *E04MalariaClinical*. The cytokine and immune profiling data are deposited on ImmPort under the identifier SDY1015.

5.5.2 Bioinformatics Analysis

Differential expression analysis was carried out using MATLAB's implementation of DESeq algorithm on library size normalized read counts. Enrichment analysis of identified DEG was conducted using MySigDB web service. The gene sets tested for enrichment included canonical pathways, Hallmark pathways, BioCarta pathways and KEGG pathways. A q-value cut-off of 0.05 was used. Clustering of transcriptomic, cellular population and cytokine abundances was done using k-means clustering. The number of clusters used was determined by the silhouette value of different numbers of clusters (1-6).

5.6 References

Bibliography

[1] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome biology*, 11(10):R106, 2010.

[2] H. Chavale, J. R. Santos-Oliveira, A. M. Da-Cruz, and S. Enosse. Enhanced t cell activation in plasmodium falciparum malaria-infected human immunodeficiency virus-1 patients from mozambique. *Memórias do Instituto Oswaldo Cruz*, 107(8):985–992, 2012.

[3] G. Coatney, W. Collins, M. Warren, and P. Contacos. The primate malarias. 1971. *Washington: US Government Printing Office Google Scholar*.

[4] F. B. Cogswell. The hypnozoite and relapse in primate malaria. *Clinical microbiology reviews*, 5(1):26–35, 1992.

[5] G. A. Deye, M. Gettayacamin, P. Hansukjariya, R. Im-erbsin, J. Sattabongkot, Y. Rothstein, L. Macareo, S. Fracisco, K. Bennett, A. J. Magill, et al. Use of a rhesus plasmodium cynomolgi model to screen for anti-hypnozoite activity of pharmaceutical substances. *The American journal of tropical medicine and hygiene*, 86(6):931–935, 2012.

[6] M. R. Galinski and J. W. Barnwell. Plasmodium vivax: who cares? *Malaria Journal*, 7(1):S9, 2008.

[7] M. R. Galinski, E. Meyer, and J. W. Barnwell. Plasmodium vivax: modern strategies to study a persistent parasites life cycle. *Adv Parasitol*, 81:1–26, 2013.

[8] R. M. Gonçalves, N. F. Lima, and M. U. Ferreira. Parasite virulence, co-infections and cytokine balance in malaria. *Pathogens and global health*, 108(4):173–178, 2014.

[9] H. Han, H. Shim, D. Shin, J. E. Shim, Y. Ko, J. Shin, H. Kim, A. Cho, E. Kim, T. Lee, et al. Trrust: a reference database of human transcriptional regulatory interactions. *Scientific reports*, 5, 2015.

[10] C. Joyner, A. Moreno, E. V. Meyer, M. Cabrera-Mora, J. C. Kissinger, J. W. Barnwell, and M. R. Galinski. Plasmodium cynomolgi infections in rhesus macaques display clinical and parasitological features pertinent to modelling vivax malaria pathology and relapse infections. *Malaria Journal*, 15(1):451, 2016.

[11] K. D. Kim, J. Zhao, S. Auh, X. Yang, P. Du, H. Tang, and Y.-X. Fu. Adaptive immune cells temper initial innate responses. *Nature medicine*, 13(10):1248–1252, 2007.

[12] W. Krotoski, R. Bray, P. Garnham, R. Gwadz, R. Killick-Kendrick, C. Draper, G. Targett, D. Krotoski, M. Guy, L. Koontz, et al. Observations on early and late post-sporozoite tissue stages in primate malaria. ii. the hypnozoite of plasmodium cynomolgi bastianellii from 3 to 105 days after infection, and detection of 36-to 40-hour pre-erythrocytic forms. *The American journal of tropical medicine and hygiene*, 31(2):211–225, 1982.

[13] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27(12):1739–1740, 2011.

[14] W. L. Mandala, C. L. Msefula, E. N. Gondwe, M. T. Drayson, M. E. Molyneux, and C. A. MacLennan. Cytokine profiles in malawian children presenting with uncomplicated malaria, severe malarial anemia, and cerebral malaria. *Clinical and Vaccine Immunology*, 24(4):e00533–16, 2017.

[15] B. Mbengue, B. Niang, M. S. Niang, M. L. Varela, B. Fall, M. M. Fall, R. N. Diallo, B. Diatta, D. Gowda, A. Dieye, et al. Inflammatory cytokine and humoral responses to

plasmodium falciparum glycosylphosphatidylinositols correlates with malaria immunity and pathogenesis. *Immunity, inflammation and disease*, 4(1):24–34, 2016.

[16] I. Mueller, M. R. Galinski, J. K. Baird, J. M. Carlton, D. K. Kochar, P. L. Alonso, and H. A. del Portillo. Key gaps in the knowledge of plasmodium vivax, a neglected human malaria parasite. *The Lancet infectious diseases*, 9(9):555–566, 2009.

[17] F. Noulin, C. Borlon, J. Van Den Abbeele, U. DAlessandro, and A. Erhart. 1912–2012: a century of research on plasmodium vivax in vitro culture. *Trends in parasitology*, 29(6):286–294, 2013.

[18] N. K. Nyakoe, R. P. Taylor, J. N. Makumi, and J. N. Waitumbi. Complement consumption in children with plasmodium falciparum malaria. *Malaria journal*, 8(1):7, 2009.

[19] W. H. Organization et al. World malaria report 2016. *Geneva: WHO. Embargoed until*, 13, 2016.

[20] D. Prakash, C. Fesel, R. Jain, P.-A. Cazenave, G. C. Mishra, and S. Pied. Clusters of cytokines determine malaria severity in plasmodium falciparum–infected patients from endemic areas of central india. *Journal of Infectious Diseases*, 194(2):198–207, 2006.

[21] J. A. Stoute, A. O. Odindo, B. O. Owuor, E. K. Mibei, M. O. Opollo, and J. N. Waitumbi. Loss of red blood cell–complement regulatory proteins and increased levels of circulating immune complexes are associated with severe malarial anemia. *Journal of Infectious Diseases*, 187(3):522–525, 2003.

[22] K. Sun and D. W. Metzger. Inhibition of pulmonary antibacterial defense by interferon-γ during recovery from influenza infection. *Nature medicine*, 14(5):558–564, 2008.

[23] S.-I. Tachibana, S. A. Sullivan, S. Kawai, S. Nakamura, H. R. Kim, N. Goto, N. Arisue, N. M. Palacpac, H. Honma, M. Yagi, et al. Plasmodium cynomolgi genome sequences

provide insight into plasmodium vivax and the monkey malaria clade. *Nature genetics*, 44(9):1051–1055, 2012.

[24] K. J. Torres, E. Villasis, J. Bendezú, J. Chauca, J. M. Vinetz, and D. Gamboa. Relationship of regulatory t cells to plasmodium falciparum malaria symptomatology in a hypoendemic region. *Malaria journal*, 13(1):108, 2014.

[25] T. M. Tran, M. B. Jones, A. Ongoiba, E. M. Bijker, R. Schats, P. Venepally, J. Skinner, S. Doumbo, E. Quinten, L. G. Visser, et al. Transcriptomic evidence for modulation of host inflammatory responses during febrile plasmodium falciparum malaria. *Scientific Reports*, 6, 2016.

[26] M. Warren, J. Skinner, and E. Guinn. Biology of the simian malarias of southeast asia. i. host cell preferences of young trophozoites of four species of plasmodium. *The Journal of parasitology*, pages 14–16, 1966.

[27] A. P. Waters, D. G. Higgins, and T. McCutchan. Evolutionary relatedness of some primate models of plasmodium. *Molecular biology and evolution*, 10(4):914–923, 1993.

[28] M. N. Wykes, J. M. Horne-Debets, C.-Y. Leow, and D. S. Karunarathne. Malaria drives t cells to exhaustion. *Frontiers in microbiology*, 5:249, 2014.

Chapter 6

Quantification of Healthy Red Blood Cell Removal and Preferential Invasion of Reticulocytes in *Macaca mulatta* during *Plasmodium cynomolgi* Infection

6.1 Introduction

The goal of this article is to quantify the removal of healthy red blood cells and preferential invasion of reticulocytes during malaria infection, and find their associations with molecular phenomena through a modeling approach. The disease malaria is caused by *Plasmodium* parasites. Out of the five human *Plasmodium* species capable of causing malaria, *Plasmodium falciparum* and *Plasmodium vivax* account for the majority of human malaria infections. *P. falciparum* is responsible for the majority of malaria-related mortality and is most prevalent in sub-Saharan Africa [9].

The *Plasmodium* life cycle is comprised of several stages. The infection process in humans starts with the injection of sporozoites by mosquitoes into the skin of the host. This is followed by the liver stage, during which the inoculated sporozoites grow and multiply asexually within hepatocytes for 1-2 weeks to produce merozoites. The newly produced merozoites emerge from the liver and enter the blood stream. The blood-stage infection starts immediately after the hepatic stage. The parasite's blood-stage infection in both human and non-human primates generally has a regular cycle of 24 to 72 hours depending on the species of the *Plasmodium* parasite [1, 4]. The parasites invade healthy red blood cells (RBCs) and replicate asexually, remodeling and ultimately destroying the RBCs in the process. The destruction of RBCs during blood-stage malaria infection sometimes results in severe anemia, which is one of the major complications of malaria and a leading cause of mortality.

During blood-stage malaria infection, the majority of red blood cell loss has been attributed to the removal of healthy red blood cells [3, 7, 11]. The hemodynamic model described by Yan *et al* [15] characterizes this phenomenon through a mechanistic model taking into account the interaction between healthy RBCs, infected RBCs, cells form the innate immune system, and cells from the adaptive immune system. This model captures the clinical outcomes of a malaria infection: resistance to the disease, susceptibility, and resilience (chronic infection with mild symptoms). That model has been simplified to account for experimental constraints, and it has been calibrated in this article with the data of *Macaca mulatta* infected with *P. cynomolgi* described by Joyner *et al* [8]. *Plasmodium cynomolgi* is a malaria parasite that infects old world monkeys; it is physiologically and evolutionarily similar to *P. vivax* [13, 14].

This paper is organized as follows: Section 2 provides the description of the experimental data. Section 3 contains the detailed derivation and description of our hemodynamic model. Section 4 describes the model calibration process. Section 5 provides an overview of the calibration results. Section 6 integrates our modeling result with transcriptomic data collected during the experiment. Section 7 describes an adjusted model based on our calibration results. Section 8 offers some conclusions, and discusses the biological significance of our results.

6.2 Experimental Description

The experimental design is shown in Fig 6.1. The infected red blood cell, total red blood cell, and reticulocyte concentration were measured daily over the entire experiment. The four subjects, S1, S2, M1 and M2 are referred as Subject $1 \sim 4$ within this paper. Each red diamond corresponds to a specific time point where transcriptome, cytokine and immune profiling was conducted. Time point 2 (TP2) corresponds to the acute primary infection when parasite counts peaked. A detailed description of the experiment that collected this data can be found in Joyner *et al.* [8].

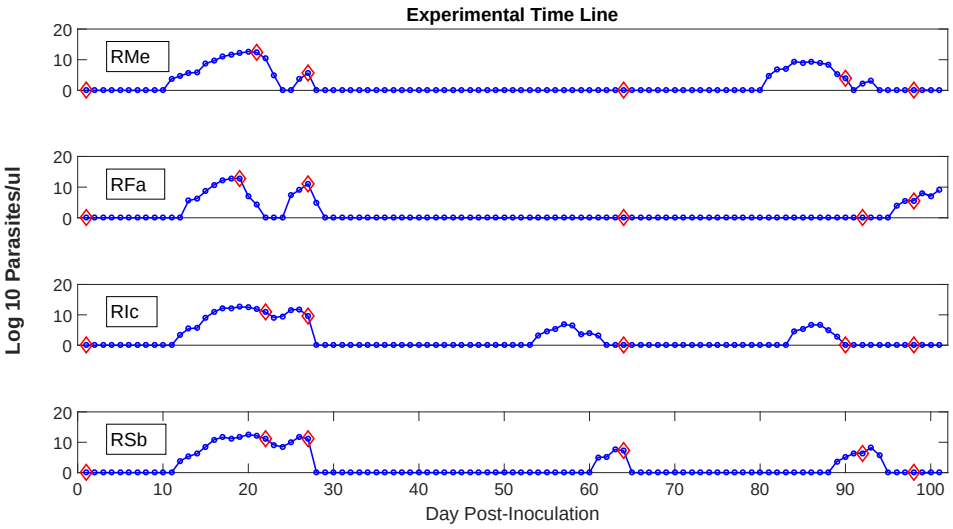

Figure 6.1: Parasite count over the entire experiment. Transcriptome, cytokine, and immune profiling were conducted during the 6-time points marked with the red diamond. Subject S1 and S2 experienced severe malaria symptoms. Subject M1 and M2 experienced mild malaria symptoms.

6.3 Matching a Theoretical Model with Experimental Constraints

The mechanistic partial differential equation model described by Yan *et al.* [15], has the form presented below; the detailed definition of variables, parameters and functions is presented in that paper, and also for completeness in the Appendix.

$$\begin{aligned}
\frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} &= -\left(\sum_{i=1}^{T} w_i(t)\theta_i + \sum_{i=1}^{Q} s_i(t)\psi_i + h(a)\right) u(a,t) \\
&\quad - \gamma\kappa v(\alpha_{max}, t) p(u(a,t)), \\
\frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} &= -\left(\sum_{i=1}^{T} w_i(t)\phi_i + \sum_{i=1}^{Q} s_i(t) b_i(t)\right) v(\alpha, t), \\
\frac{dw_i}{dt} &= o_i(t) - \beta_i w_i(t), \\
\frac{ds_i}{dt} &= l_i(t) - \delta_i s_i(t),
\end{aligned} \tag{6.1}$$

subject to the following initial and boundary conditions:

$$\begin{aligned} u(a,0) =& g(a), \\ u(0,t) =& f(t,\varphi(t)), \\ v(\alpha,0) =& c(\alpha), \\ v(0,t) =& \gamma\kappa \int_{a_0}^{a_{max}} v(\alpha_{max},t)p(u(a,t))\,da. \end{aligned}$$

With the experimental constraints, it was not possible to measure the delay in the erythropoietic response, and hence it was removed from this model. Thus the boundary condition $u(0,t) = x_4 e^{x_5(T_0 - T(t))}$, where x_4 is the baseline production of red blood cells and x_5 is a parameter controlling the speed of erythropoiesis response. T_0 is the sum of the steady-state concentration of reticulocytes and mature red blood cell and T is the sum of the concentration of reticulocytes and mature red blood cell at time t.

Under the explicit assumption, that the immune cell populations do not interact with the infected red blood cell population (v) and healthy red blood cell population (u) ($\phi_i = 0$, $\theta_i = 0$, $\psi_i = 0$, for all i and $b_i(t) = 0$ for all t), the system reduces to:

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial u}{\partial a} =& - h(a)u(a,t) - \gamma\kappa v(\alpha_{max},t)p(u(a,t)), \\ \frac{\partial v}{\partial t} + V\frac{\partial v}{\partial \alpha} =& 0. \end{aligned}$$

Furthermore, under the explicit assumption that the initial age distribution of v is uniform, $v(\alpha,0) = C$. Let

$$V(t) = \int_0^{\alpha_{max}} v(\alpha,t)d\alpha,$$

where $V(t)$ denote the total amount of iRBC at time t. Then,

$$V(0) = \alpha_{max}C.$$

Additionally, let $t_2 = t_1 + \epsilon, \epsilon > 0$, then:

$$\begin{aligned} \int_0^{\alpha_{max}} v(\alpha,t_2)d\alpha = & \int_0^{\alpha_{max}} v(\alpha,t_1)d\alpha, \\ & + \int_{t_1}^{t_2} \gamma\kappa \int_{a_0}^{a_{max}} v(\alpha_{max},t)p(u(a,t))dadt, \\ & - \int_{t_1}^{t_2} v(\alpha_{max},t)dt. \end{aligned}$$

Because $\int_{a_0}^{a_{max}} p(u(a,t)) = 1$ for all t, the system further simplifies to:

$$\begin{aligned}\int_0^{\alpha_{max}} v(\alpha, t_2)d\alpha = &\int_0^{\alpha_{max}} v(\alpha, t_1)d\alpha, \\ &+ \int_{t_1}^{t_2} \gamma\kappa v(\alpha_{max}, t)dt, \\ &- \int_{t_1}^{t_2} v(\alpha_{max}, t)dt.\end{aligned} \tag{6.2}$$

Substituting V into (6.2), and rearranging the equation:

$$V(t_2) - V(t_1) = (\gamma\kappa - 1)\int_{t_1}^{t_2} v(\alpha_{max}, t)dt.$$

Thus,

$$\int_{t_1}^{t_2} \dot{V}dt = (\gamma\kappa - 1)\int_{t_1}^{t_2} v(\alpha_{max}, t)dt$$

and

$$\dot{V} = (\gamma\kappa - 1)\int_{a_0}^{a_{max}} v(\alpha_{max}, t)p(u(a,t))\, da. \tag{6.3}$$

Because the age distribution of red blood cells are unknown throughout the experiment, and only reticulocyte and mature red blood cell populations are measured, we define a_1 as the precise time point where reticulocytes become mature red blood cells. We further define the total population of reticulocytes at time point t as $R(t)$ and the total population of mature red blood cells at time point t as $U(t)$. $R(t)$ and $U(t)$ have the following form:

$$R(t) = \int_0^{a_1} u(a,t)da,$$

and

$$U(t) = \int_{a_1}^{a_{max}} u(a,t)da,$$

thus

$$R(t) + U(t) = \int_0^{a_{max}} u(a,t)da.$$

Therefore, (6.3) can be rewritten as:

$$\dot{V} = (\gamma\kappa - 1)v(\alpha_{max}, t)\left(\frac{R}{R+U} + \frac{U}{R+U}\right).$$

Under the assumption that on average a small fixed percentage of infected red blood cells are bursting, then

$$v(\alpha_{max}, t) = C_2 V.$$

Which means:

$$\dot{V} = (\gamma\kappa - 1)C_2 V \left(\frac{R}{R+U} + \frac{U}{R+U} \right).$$

Additionally, knowing that the infection rate of reticulocyte is different from mature red blood cells, we can modify the system to reflect this fact:

$$\dot{V} = (\gamma\kappa_1 - 1)C_2 V \frac{R}{R+U} + (\gamma\kappa_2 - 1)C_2 V \frac{U}{R+U}.$$

Let

$$x_3 N = (\gamma\kappa_1 - 1)C_2,$$

and

$$x_6 N = (\gamma\kappa_2 - 1)C_2.$$

Then

$$\dot{V} = x_3 N \frac{UV}{R+U} + x_6 N \frac{RV}{R+U}.$$

The derivative of $R(t)$ can be derived similarly. Let:

$$\begin{aligned} \int_0^{a_1} u(a, t_2)da = &\int_0^{a_1} u(a, t_1)da \\ &+ \int_{t_1}^{t_2} x_4 e^{x_5(T_0 - T)} dt \\ &- \int_{t_1}^{t_2} u(a_1, t)dt \\ &- \int_{t_1}^{t_2} x_6 N \frac{RV}{R+U}. \end{aligned}$$

Because the much shorter life span of reticulocytes in comparison to mature red blood cells [2] and the fact that the survival rate of reticulocytes to be close to 1 [6], the hazard function, h(a), of reticulocyte is ignored. Substituting R, and rearrange the equations, we obtain:

$$R(t_2) - R(t_1) = \int_{t_1}^{t_2} x_4 e^{x_5(T_0 - T(t))} - u(a_1, t) - x_6 N \frac{RV}{R+U} dt.$$

Thus,

$$\dot{R} = x_4 e^{x_5(T_0 - T)} - u(a_1, t) - x_6 N \frac{RV}{R+U}.$$

Under the assumption that on average a small fixed percentage of reticulocyte are aging into mature red blood cell at any given time,

$$\dot{R} = x_4 e^{x_5(T_0 - T)} - x_1 R - x_6 N \frac{RI}{R+U}.$$

The derivative of $U(t)$ is also derived similarly. let:

$$\begin{aligned}\int_{a_1}^{a_{max}} u(a, t_2)da = &\int_{a_1}^{a_{max}} u(a, t_1)da \\ &+ \int_{t_1}^{t_2} u(a_1, t)dt \\ &- \int_{t_1}^{t_2} u(a_{max}, t)dt. \\ &- \int_{t_1}^{t_2} \int_{a_1}^{a_{max}} h(a)u(a, t)dadt \\ &- \int_{t_1}^{t_2} x_3 N \frac{UV}{R+U}.\end{aligned}$$

Assuming that on average, a fixed percentage of red blood cells are removed due to random chance, substituting U then the equation becomes:

$$U(t_2) - U(t_1) = \int_{t_1}^{t_2} u(a_1, t) - x_2 U - x_6 N \frac{RI}{U+R} dt.$$

Thus:

$$\dot{U} = x_1 R - x_2 U - x_6 N \frac{RI}{U+R}.$$

In conclusion, under the following explicit assumptions:

- The change of infected red blood cell population and healthy red blood cell population is independent of the immune cell populations,
- the erythropoiesis response does not have a delay,
- on average, a fixed percentage of iRBCs are bursting at any given moment and a fixed percentage of reticulocytes are transitioning into mature red blood cells,

the original PDE model (6.1) is simplified to the following system of ordinary differential equations that captures the change in the total infected red blood cell, red blood cell, and reticulocyte population:

$$\dot{U} = x_1 R - x_2 U - x_3 N \frac{UI}{T},$$
$$\dot{R} = x_4 e^{x_5(T_0 - T)} - x_1 R - x_6 N \frac{RI}{T},$$
$$\dot{I} = x_3 N \frac{UI}{U+R} + x_6 N \frac{RI}{U+R}.$$

Where U denote the concentration of healthy mature red blood cells (RBCs), R denotes healthy reticulocytes (RT) and I denote the concentration of infected red blood cells (iRBCs). Let $T = U + R$ and $T_0 = R_0 + U_0$ where R_0 and U_0 denote the steady state concentration of RTs and RBCs in the absence of malaria infection. N denote the average merozoites produced per infected red blood cell. $x_1 R$ describes the aging of RTs to become RBCs. $x_2 U$ denote the random removal of mature red blood cells. x_3 and x_6 denote the infection success rate of RBCs and RTs respectively. x_4 denotes the baseline production of RTs and the term $e^{x_5(T_0 - T)}$ describes the host erythropoiesis response.

6.4 Parameter Estimation

Under the assumption of steady state, where R_0 is the steady state concentration of RTs and U_0 is the steady state concentration of RBCs, the following equality is established:

$$x_4 = x_1 R_0 = x_2 U_0.$$

The average number of merozoite produced is set to be 20. Fonseca *et al.* [6] estimated that the baseline production of RTs in *M mulatta*, x_4, is $192,00$ cells per day [6]. R_0 and U_0 are estimated for each subjects using the average of the red blood cell and reticulocytes counts during the first ten days of the experiment where no iRBCs were detected. The lower bound and upper bound of the other three parameters, x_3, x_6 and x_5 were also estimated. All three parameters are positive. The upper bound of x_5 was set to allow a maximum of

2-fold increase in RTs production in each subject. Upper bound of x_3 and x_6 was set to be 0.05 and 1 respectively. When $x_3 > 0.05$ and/or $x_6 > 1$, the system becomes numerically unstable to solve due to stiffness using odesolver45 in MATLAB environment. The upper bound of x_3 and x_6 also ensures that the $\dot{I} < 0.5NI$ throughout the simulation, meaning that the number of infected red blood cells produced over a two day period can not exceed the number of merozoites produced.

A multiple objective genetic algorithm [5] was used to estimate the parameters (x_3, x_6, x_5) for each subjects using the estimated lower and upper bound of each parameter. The model was fitted to the cellular data with starting date corresponding to the first appearance of iRBCs and ending date corresponding to time point 2. The two objectives minimized were the average percent error (APE) of predicted RTs and iRBCs concentration. APE have the following form for a specific variable such as RTs:

$$APE_{retic} = \frac{100}{n} \sum_{i=1}^{n} \frac{r_i - R_i}{R_i},$$

where r_i is the observed RTs concentration at time point i and R_i is the model predicted RTs concentration. Contrary to single objective minimization, a Pareto front for both objectives is estimated during each application of the genetic algorithm. The Pareto front refers to a set of possible values of both object function such that the decrease in one objective function necessitates the increase of the other [5]. Each application of a genetic algorithm is terminated when the change in the estimated Pareto front is less than the predefined tolerance (0.0001). Due to the stochastic nature of the genetic algorithm, 1,000 runs were applied to estimate the parameters for each subject. $> 99\%$ of the application of genetic algorithm terminated due to the convergence of Pareto front, the rest did not converge during the maximum allowed run time of 60 seconds.

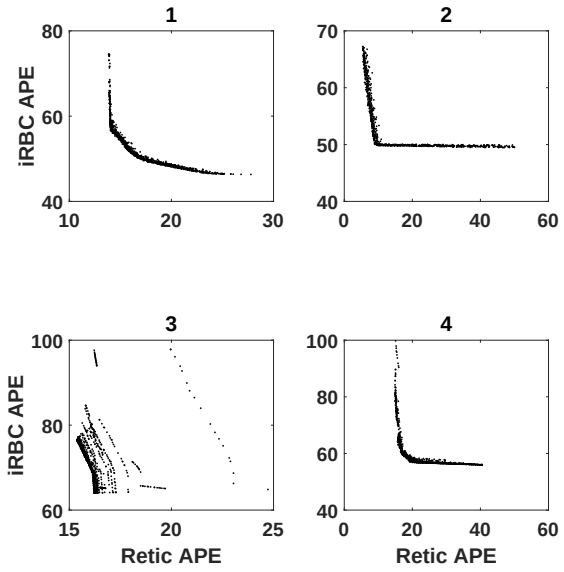

Figure 6.2: Estimated Pareto Front of Average Percentage Error (APE) of iRBCs and RTs for each of the four subjects.

6.5 Results

6.5.1 Result Overview

The Pareto front of iRBC APE and RT APE for each subject are shown in Figure 6.2. For subject 1, 2 and 4, there exist a sub-region on the Pareto front such that the iRBC APE and RT APE have a negative linear relationship, which means that $iRTAPE + RTAPE$ is close to constant. The Pareto front estimated for subject 3 does not contain such a region, indicating a lack of model fit for that specific subject. The iRBC APE for all four subjects has a range of $(45\% \sim 90\%)$ where the RT APE have a range of $(15\% \sim 50\%)$. The simulation of the top 100 model with the lowest iRBCs APE are shown in Figure (6.3, 6.4, 6.5, 6.6).

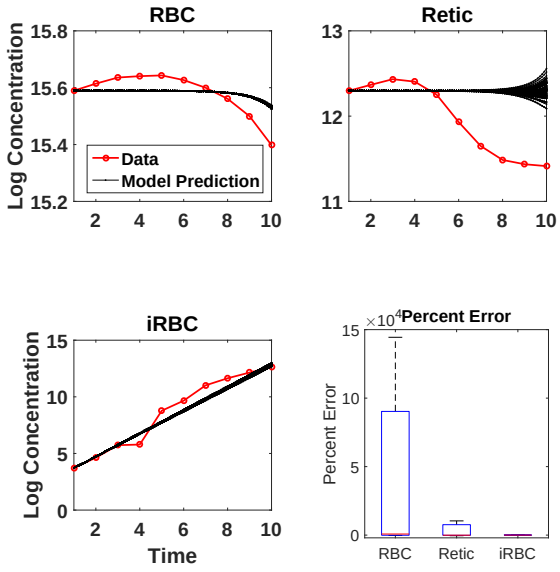

Figure 6.3: Simulation of the top 100 models with the lowest iRBCs APE and distribution of percent error of subject 1

6.5.2 Quantification of The Preferential Infection of Reticulocyte by *P. cynomolgi*

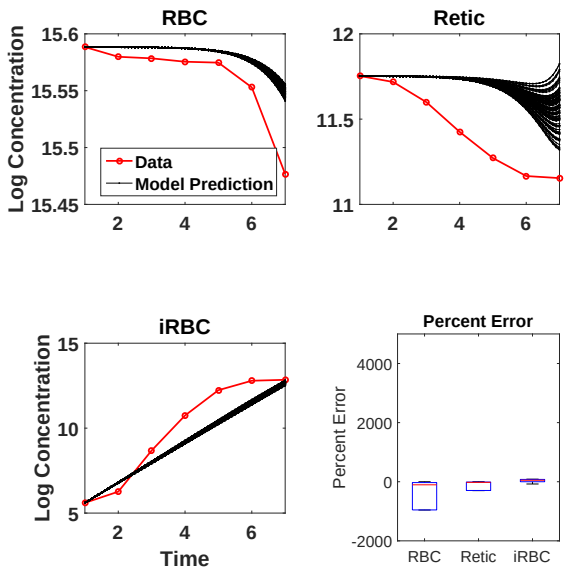

Figure 6.4: Simulation of the top 100 models with the lowest iRBCs APE and distribution of percent error of subject 2

To quantify the preferential infection of RTs by *P cynomolgi*, the ratio of x_6 to x_3 in the top 100 model with the lowest iRBCs APE for each subject was calculated. The ratio $\frac{x_6}{x_3}$ is interpreted as the likelihood of *P cynomolgi* infecting RTs over infecting RBCs. The distribution of $\frac{x_6}{x_3}$ is shown in Figure 6.7. The means of the four distribution are 13, 17, 28 and 18 respectively.

6.5.3 Quantification of Removal of Healthy Red Blood Cell (hRBC)

The loss of healthy RBC (hRBC) is also calculated based on the 100 model prediction with the lowest iRBC APE. Total Loss of hRBC and Ratio of loss of hRBC and parasitized RBC are shown in Figure 6.8. On average, 1.5 million hRBCs per μl are cleared by the host through out the onset of the disease till acute primary infection. Furthermore, our model predicted that for each RBC parasitized, the host removes 3 hRBC. The quantification of

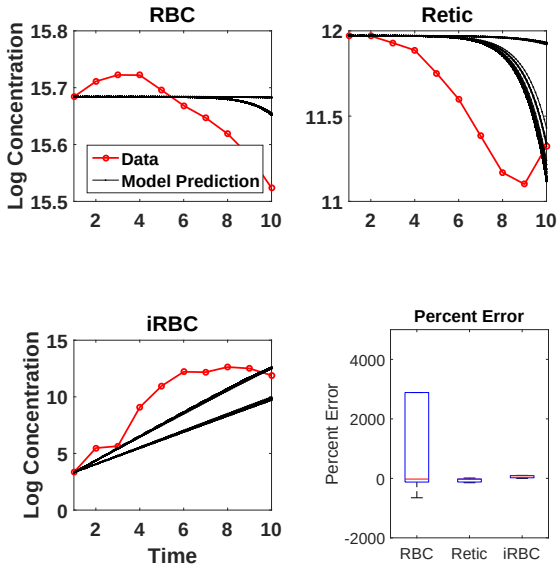

Figure 6.5: Simulation of the top 100 models with the lowest iRBCs APE and distribution of percent error of subject 3

the removal of hRBCs over the entire infection is shown in Figure 6.9. The model fit of RBC and retic Population after adjusting for removal of hRBCs are shown in Figure 6.10.

6.6 Correlation and Enrichment Analysis of Rate of hRBC Removal

The rate of hRBC removal was estimated for all four subjects (Fig 6.11) based on the estimated loss of hRBCs (Fig 6.9). Pearson's correlation between hRBC removal rate and their corresponding transcript, immune cell and cytokine abundance was calculated. The p-value of the correlation was adjusted for false discovery rate and the transcript, immune cell and cytokine exhibiting significant correlation with hRBC removal rate (q-val ≤ 0.05) are shown in Fig 6.12. Several pro-inflammatory ctokines such as IL6 and IL1B are positively correlated with the rate of hRBC removal. On the other hand, CD8 T cell population displayed significant negative correlation with hRBC removal rate.

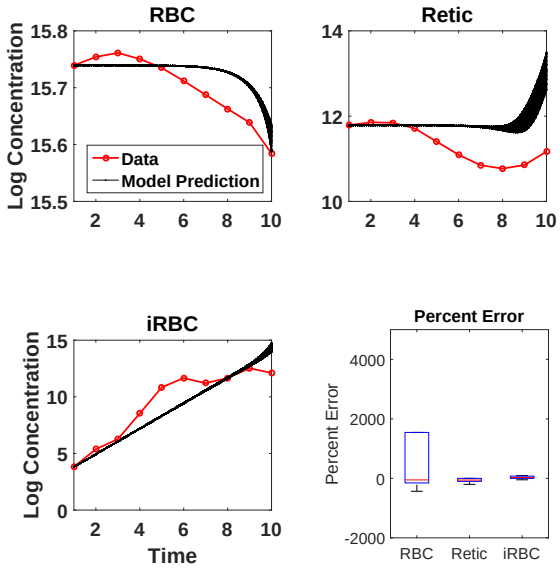

Figure 6.6: Simulation of the top 100 models with the lowest iRBCs APE and distribution of percent error of subject 4

Only seven genes: MYO3B, GAN, DNAJB4, TRIM45, TMEM150A, IL23R and BMF have shown significant correlation with the rate of hRBC removal. To fully explore the association of transcriptome change and rate of hRBC removal, Gene Set Enrichment Analysis [12] was applied to the correlation ranked gene lists. The most significantly enriched GO gene sets and pathways are shown in table 6.1 and table 6.2. Innate immune related gene sets and pathways are positively correlated with hRBC removal rate, which corresponds to the positive correlation between hRBC removal and inflammatory cytokines. On the other hand, transcripts that have negative correlation with hRBC removal are enriched in RNA and protein processing related pathways.

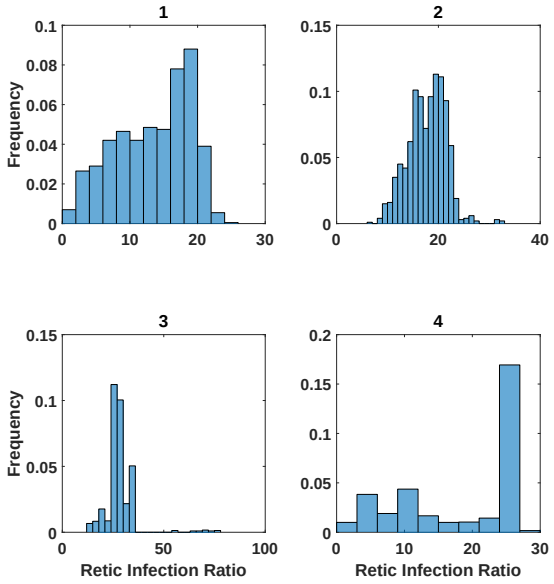

Figure 6.7: Distribution of $\frac{x_6}{x_3}$ for subjects 1 4.

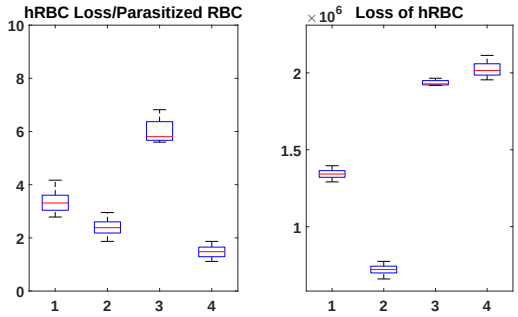

Figure 6.8: Estimated total loss of hRBC and ratio of loss of hRBC and parasitized RBC in subject 1-4

6.7 Empirical Model Adjustment

The simulation of the top 100 model with the lowest iRBC APE for subject 1,2 and 3 (Figure 6.3, 6.4, 6.5, 6.6) failed to capture the increase of RBC population during the early

Table 6.1: Enrichment of GO Gene Sets

Positive Enriched GO Gene Sets	Negative Enriched GO Gene Sets
INNATE IMMUNE RESPONSE	NCRNA PROCESSING
IMMUNE EFFECTOR PROCESS	NCRNA METABOLIC PROCESS
ACTIN FILAMENT BASED PROCESS	RIBOSOME BIOGENESIS
REGULATION OF BODY FLUID LEVELS	RRNA METABOLIC PROCESS
WOUND HEALING	RIBONUCLEOPROTEIN COMPLEX BIOGENESIS
RESPONSE TO BACTERIUM	TRANSLATIONAL INITIATION
DEFENSE RESPONSE TO OTHER ORGANISM	AMIDE BIOSYNTHETIC PROCESS
INFLAMMATORY RESPONSE	PEPTIDE METABOLIC PROCESS
ACTIVATION OF IMMUNE RESPONSE	MULTI ORGANISM METABOLIC PROCESS
REGULATION OF RESPONSE TO WOUNDING	NONSENSE MEDIATED DECAY
HEMOSTASIS	PROTEIN LOCALIZATION
REGULATED EXOCYTOSIS	PROTEIN LOCALIZATION TO ENDOPLASMIC RETICULUM
SECRETION	PROTEIN TARGETING TO MEMBRANE
RESPONSE TO WOUNDING	TRNA METABOLIC PROCESS
CELLULAR RESPONSE TO NITROGEN COMPOUND	RNA CATABOLIC PROCESS
ENDOCYTOSIS	ORGANIC CYCLIC COMPOUND CATABOLIC PROCESS
RESPONSE TO VIRUS	TRNA PROCESSING
CELL CELL ADHESION	MITOCHONDRIAL TRANSLATION
REGULATION OF INFLAMMATORY RESPONSE	TRANSLATIONAL ELONGATION

Table 6.2: Enrichment of Pathways

Positive Enriched Pathways	Negative Enriched Pathways
SYSTEMIC LUPUS ERYTHEMATOSUS	3 UTR MEDIATED TRANSLATIONAL REGULATION
HEMOSTASIS	TRANSLATION
CYTOKINE SIGNALING IN IMMUNE SYSTEM	PEPTIDE CHAIN ELONGATION
CHEMOKINE SIGNALING PATHWAY	RIBOSOME
INTERFERON ALPHA BETA SIGNALING	INFLUENZA VIRAL REPLICATION
PID PDGFRB PATHWAY	PROTEIN TARGETING TO MEMBRANE
REGULATION OF ACTIN CYTOSKELETON	NONSENSE MEDIATED DECAY
PLATELET ACTIVATION SIGNALING AND AGGREGATION	INFLUENZA LIFE CYCLE
AMYLOIDS	METABOLISM OF PROTEINS
PID VEGFR1 2 PATHWAY	METABOLISM OF RNA
RNA POL I PROMOTER OPENING	FORMATION OF THE TERNARY COMPLEX
NABA MATRISOME ASSOCIATED	ACTIVATION OF THE MRNA BINDING TO 43S
RESPONSE TO ELEVATED PLATELET CYTOSOLIC CA2	METABOLISM OF MRNA
NABA SECRETED FACTORS	MITOCHONDRIAL PROTEIN IMPORT
INTERFERON SIGNALING	RNA POLYMERASE
RHO PATHWAY	RNA POL III TRANSCRIPTION INITIATION
LEUKOCYTE TRANSENDOTHELIAL MIGRATION	RESPIRATORY ELECTRON TRANSPORT ATP SYNTHESIS
FOCAL ADHESION	PYRIMIDINE METABOLISM
JAK STAT SIGNALING PATHWAY	TCA CYCLE AND RESPIRATORY ELECTRON TRANSPORT

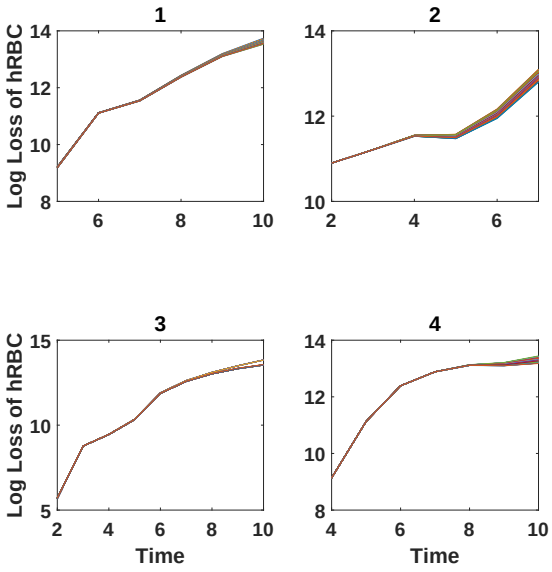

Figure 6.9: Time series of estimated loss of hRBCs.

stages of malaria infection. Considering that the difference between RBC population and steady state RBC population is minimal during the early stage of malaria infection, this discrepancy suggest that RBCs are released during the early stage of blood stage malaria. Additionally, our correlation analysis identified several cell types, cytokines and transcripts (q-value ≤ 0.05) that are linearly correlated with hRBC removal rate. This finding suggest the possibility of using these entities to predict hRBC removal rate. Taking these findings into consideration, our original model can be further expanded to:

$$\dot{U} = x_1 R - x_2 U - x_3 N \frac{UI}{T} - \sum \gamma_i E_i + g(M),$$
$$\dot{R} = x_4 e^{x_5(T_0 - T)} - x_1 R - x_6 N \frac{RI}{T},$$
$$\dot{I} = x_3 N \frac{UI}{U+R} + x_6 N \frac{RI}{U+R},$$

is linearly explained by a combination of the abundances of inflammatory cytokines. Where each E_i denote the abundance of a cytokine exhibiting linear correlation with the rate of

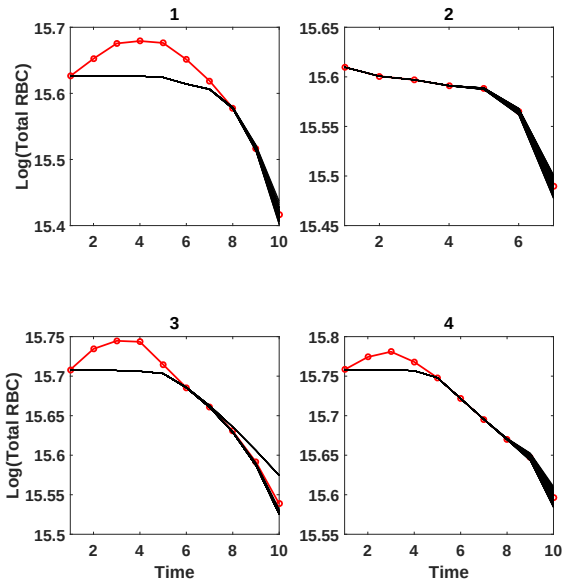

Figure 6.10: Time series of model predicted total red blood cell adjusted by the mean of estimated loss of hRBC.

hRBC removal (Fig: 6.12) and each γ_i quantifies the linear dependency. The term $g(M)$ denote the increase of RBC population during the early stage of malaria infection, that is dependent on some molecular quantity M, possibly pathogen related. The estimated form of $g(M)$ are shown in Figure 6.13 and the post-hoc fitted model is shown in Figure 6.14.

6.8 Discussion

In this chapter, a simplified Ordinary Differential Equation model was fitted to the time series data collected from the MaHPIC experiment. Our simplified model was derived from the earlier PDE system under three explicit assumptions. The simplification of the original model reduced the unknown parameters to 3 parameters with well-defined lower and upper bounds.

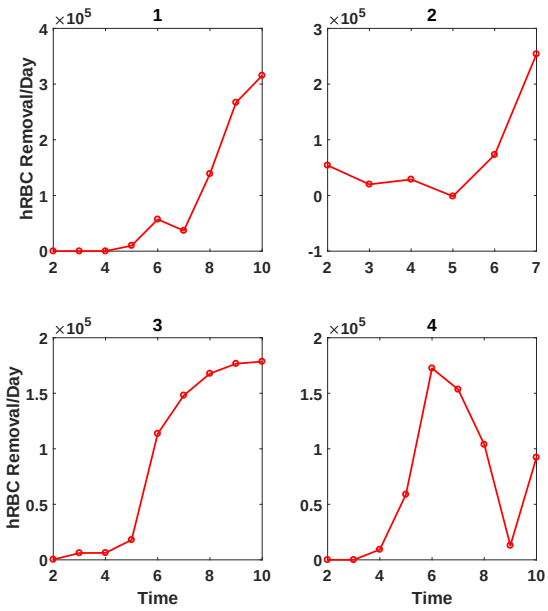

Figure 6.11: Time series of model predicted hRBC removal rate for all four subjects.

The model was fitted using multiple objective genetic algorithm and the Pareto front estimated (Fig: 6.2) demonstrate that the simplified model describes the dynamics of RTs much better than that of iRBCs as shown by the 2-fold difference in the range of RT APE and iRBC APE. The iRBC APE for all four subjects has a range of ($45\% \sim 90\%$) where as the RT APE have a range of ($15\% \sim 20\%$). This result suggests that the immune function omitted in our model have a large impact on parasite population during the primary infection.

The best-fitted models were used to quantify the preferential infection of RTs by *P. cynomolgi*. Our model predicts that, on average, *P. cynomolgi* merozoites are 20 times more likely to infect RTs than RBC (Fig 6.7). The experimental verification of our prediction is difficult due to the lack of an in vitro system to study *P. cynomolgi*, but our prediction provides a general range of preferential infection of RTs, which can be utilized in future experimental designs and modeling studies.

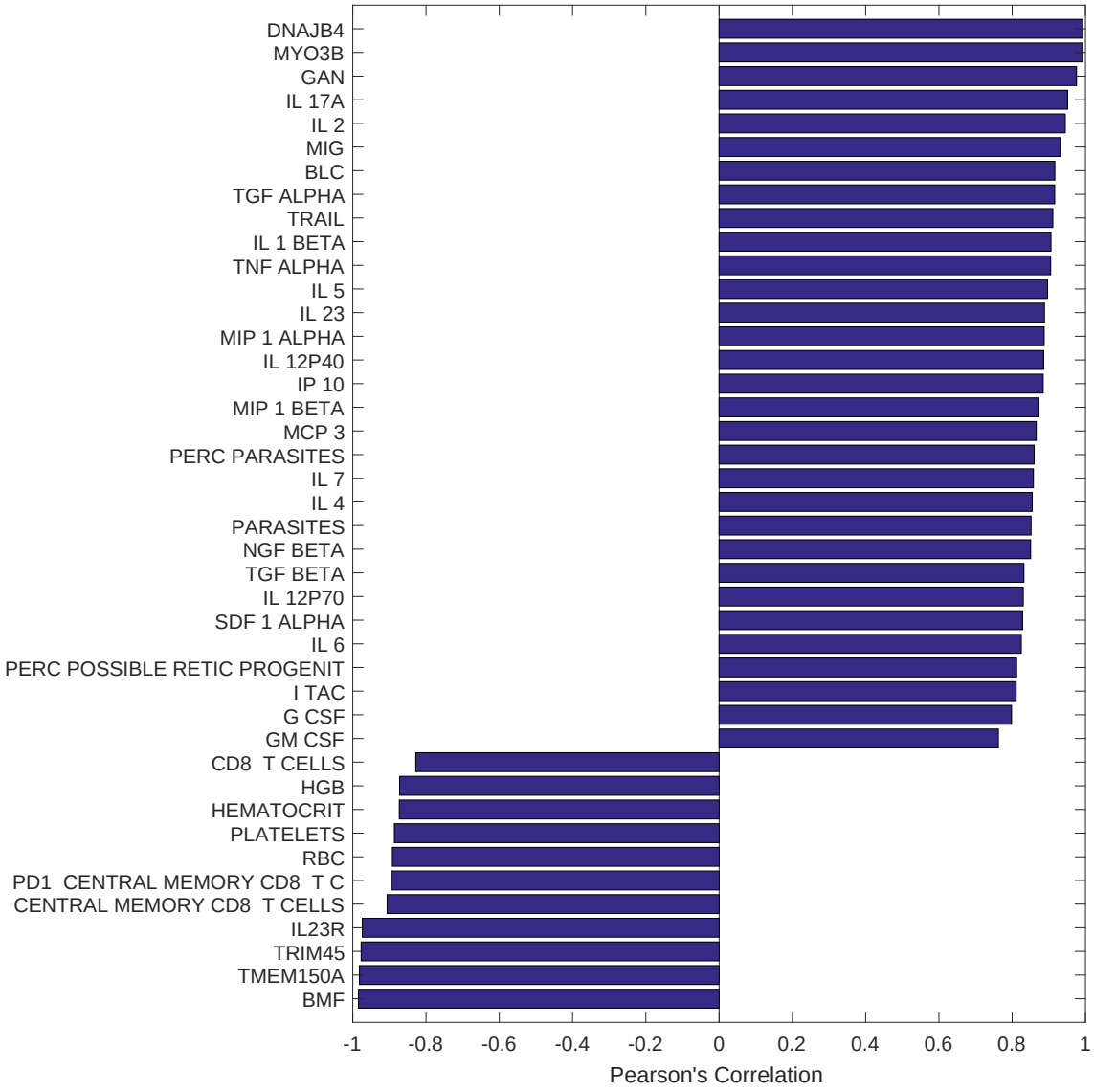

Figure 6.12: Transcripts, immune cell population and cytokines that are highly correlated (q-val ≤ 0.05) with hRBC removal rate.

Additionally, in all the best fit models (Fig 6.3, 6.4, 6.5, 6.6), our prediction of RBC concentration and RT concentration are higher than observed. This over estimation of RBC

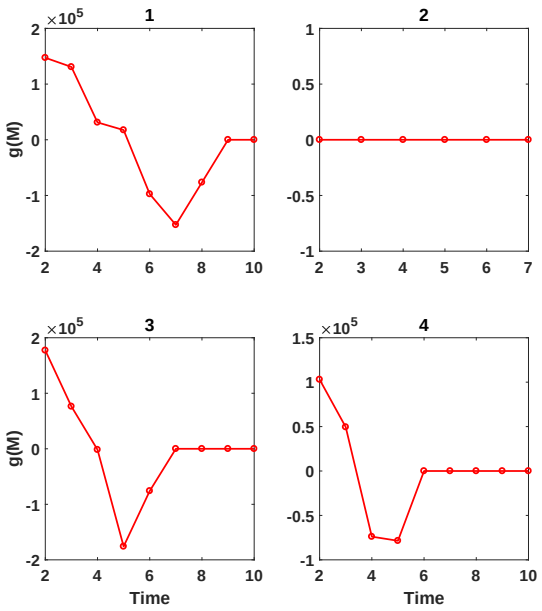

Figure 6.13: Time series of g(M) for subject 1 through 4.

and RT concentration constitutes the majority of APE. Furthermore, our iRBC concentration predictions are on average lower than the observed value. We utilized this fact to estimate the lower bound of the amount of hRBCs removed by the host. Our model estimates that at least 50% $\sim$ 80% of RBC loss during the primary infection of *P. cynomolgi* are due to the removal of hRBCs 6.8. Finally, our model shows that the speed at which hRBCs are removed increases throughout the infection.

Using the estimated hRBC removal rate, a correlation study was conducted to identify transcript, immune cell and cytokine abundance that has significant correlation with hRBC removal rate. Interestingly, innate immune related gene sets (Interferon Response, Rho Pathway and JAK-STAT pathway *et al.*) along with pro-inflammatory cytokines (IL-1B, IL-6 *et al.*) displayed significant positive correlation with hRBC removal. The association of severe malaria anemia with pro-inflammatory response has long been studied [10]; our analysis provides a list of possible cytokine biomarkers for the estimation of host clearance

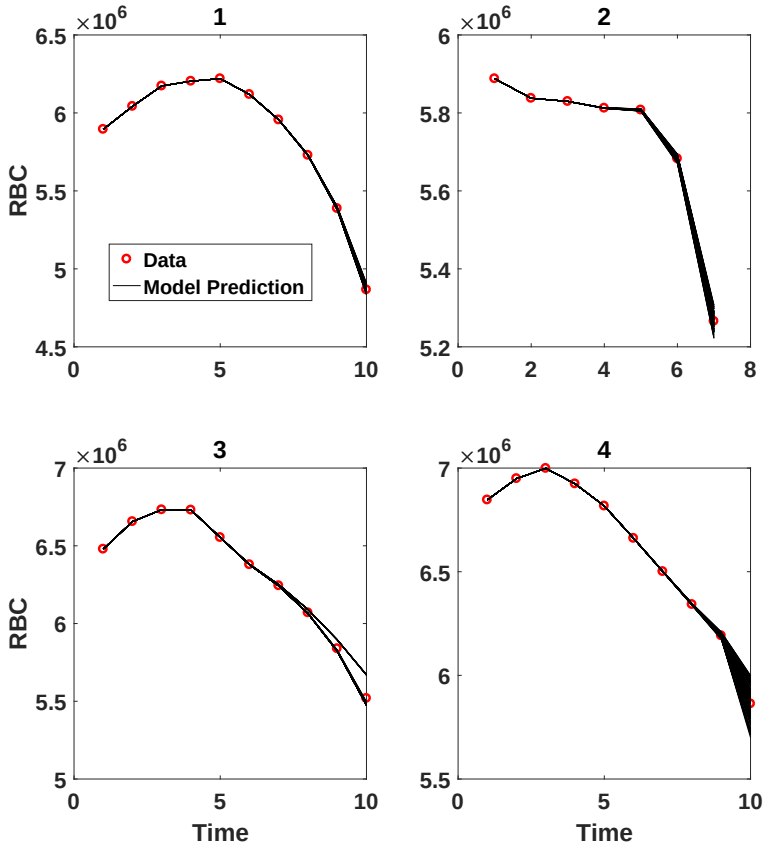

Figure 6.14: Final adjusted model prediction of RBC population for subjects 1 through 4.

of hRBCs. Due to the experimental constraint where cytokine profiling was only conducted during peak parasitemia, our analysis lacks the resolution to provide a mechanistic explanation for the correlation between pro-inflammatory cytokines and host removal of hRBC. Furthermore, the observation that inflammation related genes and cytokines are differentially up-regulated in the severe subjects (Subject 1 and 2), along with the fact that severe subjects have a higher rate of hRBC removal during peak parastemia, suggest the possible role of inflammation associated hRBC clearance and clinical severity.

In conclusion, we have demonstrated that a simplified model with only three unknown parameters can be used to predict RT concentration with an APE of ($15\% \sim 20\%$). Despite

the model's relatively poor performance at predicting iRBC dynamics (APE 50% $\sim$ 70%), it can be used to estimate the preferential infection of RTs and hRBC removal during malaria infection. The estimation of the hRBC removal rate using our model along with the downstream enrichment analysis reveals associations of hRBC removal and both the inflammatory response and CD 8 T-cell response. Application of this model to more time series data sets of malaria infection involving a variety of malaria species is necessary to validate our findings.

6.9 Reference

Bibliography

[1] Z. Bozdech, M. Llinás, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of plasmodium falciparum. *PLoS biology*, 1(1):e5, 2003.

[2] V. L. Clark and J. A. Kruse. Clinical methods: the history, physical, and laboratory examinations. *JAMA*, 264(21):2808–2809, 1990.

[3] W. E. Collins, G. M. Jeffery, and J. M. Roberts. A retrospective examination of anemia during infection of humans with plasmodium vivax. *The American journal of tropical medicine and hygiene*, 68(4):410–412, 2003.

[4] W. E. Collins, M. Warren, J. S. Sullivan, and G. G. Galland. Plasmodium coatneyi: observations on periodicity, mosquito infection, and transmission to macaca mulatta monkeys. *The American journal of tropical medicine and hygiene*, 64(3):101–110, 2001.

[5] K. Deb. *Multi-objective optimization using evolutionary algorithms*. Wiley, 2005.

[6] L. L. Fonseca, H. S. Alezi, A. Moreno, J. W. Barnwell, M. R. Galinski, and E. O. Voit. Quantifying the removal of red blood cells in macaca mulatta during a plasmodium coatneyi infection. *Malaria Journal*, 15(1):410, 2016.

[7] G. Jakeman, A. Saul, W. Hogarth, and W. Collins. Anaemia of acute malaria infections in non-immune patients primarily results from destruction of uninfected erythrocytes. *Parasitology*, 119(02):127–133, 1999.

[8] C. Joyner, A. Moreno, E. V. Meyer, M. Cabrera-Mora, J. C. Kissinger, J. W. Barnwell, and M. R. Galinski. Plasmodium cynomolgi infections in rhesus macaques display

clinical and parasitological features pertinent to modelling vivax malaria pathology and relapse infections. *Malaria Journal*, 15(1):451, 2016.

[9] W. H. Organization et al. World malaria report 2016. *Geneva: WHO. Embargoed until*, 13, 2016.

[10] D. J. Perkins, T. Were, G. C. Davenport, P. Kempaiah, J. B. Hittner, and J. M. Ong'echa. Severe malarial anemia: innate immunity and pathogenesis. *International journal of biological sciences*, 7(9):1427, 2011.

[11] R. N. Price, J. A. Simpson, F. Nosten, C. Luxemburger, L. Hkirjaroen, F. ter Kuile, T. Chongsuphajaisiddhi, and N. J. White. Factors contributing to anemia after uncomplicated falciparum malaria. *The American journal of tropical medicine and hygiene*, 65(5):614–622, 2001.

[12] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.

[13] M. Warren, J. Skinner, and E. Guinn. Biology of the simian malarias of southeast asia. i. host cell preferences of young trophozoites of four species of plasmodium. *The Journal of parasitology*, pages 14–16, 1966.

[14] A. P. Waters, D. G. Higgins, and T. McCutchan. Evolutionary relatedness of some primate models of plasmodium. *Molecular biology and evolution*, 10(4):914–923, 1993.

[15] Y. Yan, B. Adam, M. Galinski, J. C. Kissinger, A. Moreno, and J. B. Gutierrez. Mathematical model of susceptibility, resistance, and resilience in the within-host dynamics between a plasmodium parasite and the immune system. *Mathematical biosciences*, 270:213–223, 2015.

Chapter 7

Conclusion

1. **Coupled age-structured partial differential equation model can be used to capture the disease dynamics during blood-stage malaria.**

   A Partial Differential Equation model was constructed based on existing biological knowledge. This model's ability to capture the general behavior of two major outcomes of infection with *Plasmodium*: (i) the depletion of host RBCs in a naive individual and (ii) the coexistence of RBCs, iRBCs, and the immune system in individuals with previous exposure, was demonstrated both analytically and through numerical simulation.

2. **Differential ℓ_1 distances between transcriptomic time series can be detected and has the potential to reveal underlying biological perturbation.**

   A novel statistical method (MPATS), was created to detect differential pairwise ℓ_1 distances between two groups. The statistical power of the method was demonstrated through simulated data. Its application to existing data set exploring anti-viral and anti-bacterial responses of IFNα dendritic cells complemented existing time series analysis by identifying novel perturbed gene sets.

3. **Single metric to detect differential network is insufficient, and ensemble method can be used to reveal underlying biological change overlooked by traditional differential expression methods.**

   We developed an ensemble differential network analysis tool (eDiNA). Its statistical power and ability to outperform single pair-wise dynamics metric were demonstrated

through its application to simulated data. The identification of differential networks provides context for downstream functional analysis. We applied this method to explore differential transcriptomic network resulting from *P. cynolmogi* infection. We discovered specific pathways that characterize bone marrow and whole blood transcriptional response to *P. cynolmogi* infection. Specifically, our analysis identified 17 gene sets that were uniquely enriched in whole blood transcriptome response to malaria infection. The spliceosome related perturbation is uniquely observed in the whole blood transcriptome data but not the bone marrow. Furthermore, our analysis of bone marrow transcriptional response suggests the activation of the immune response in the bone marrow during *P. cynolmogi* infection.

4. **Severity of clinical outcome during *P. cynomolgi* infection of *Macaca mulatta* is characterized by the differential up-regulation of proinflammatory genes and cytokines.**

   Our analysis of the transcriptomic, cellular and cytokine data of *Macaca mulatta* infected with *P. cynomolgi* reveals differentially up-regulation of cytokines and transcripts in the severe hosts. The downstream pathway analysis demonstrates that severe malaria within our experiment is associated with elevated pro-inflammatory cytokine levels (MIP, IL6, MIG and SICAM) and the up-regulation of innate immunity and anti-viral responses related gene sets (TLR Pathway, NOD Pathway, and RIG-I Pathway). Furthermore, correlation analysis of these differentially up-regulated genes, cell population, and cytokine abundance demonstrates that differential up-regulation of innate immunity and anti-viral responses related gene sets are accompanied by the increase in pro-inflammatory cytokine abundances in hosts experiencing severe malaria.

5. **Quantificaiton of *P. cynolmogi* preferential infection of reticulocytes and estimation of loss of healthy red blood cells during primary infection**

An ordinary differential equation model was fitted to experimental data to quantify the preferential infection of reticulocytes by *P. cynomolgi*. Our model predicts that *P. cynomolgi* merozoites are 20 times more likely to infect reticulocytes than mature red blood cells, taking into account the population difference between red blood cells and reticulocytes. Furthermore, the same model predicted that at least $50\% \sim 80\%$ of red blood cell loss during the primary infection of *P. cynomolgi* are due to the removal of healthy red blood cells. Additionally, analysis of the estimated rate of healthy red blood cell removal reveals its association with pro-inflammatory cytokine abundances.