Curriculum Vitae

Kavita Gandhi
MD/PhD Candidate
Molecular Epidemiology Track
University of Maryland School of Medicine

Date February 16, 2013

Contact Information
E-mail: kavita.gandhi@som.umaryland.edu

Education

2003	B.S.	University of Florida, Gainesville, FL Microbiology, Philosophy minor (Magna Cum Laude)
2004-Present	MD/PhD	University of Maryland School of Medicine, Baltimore, MD Molecular Epidemiology (Advisor: Christopher Plowe)

Research Experience

2000	Honors Scholarship for Study Abroad Merida, Yucatan Peninsula, Mexico Department of Geology, University of Florida, Gainesville, FL (Advisor: Dr. Mark Brenner)
2002	Environmental Field Studies Abroad Atenas, Costa Rica The School for Field Studies Boston University, Boston, MA (Advisor: Dr. Nolan Quiros)
2003-2004	Pre-IRTA Post-Baccalaureate Fellow National Institute of Allergy and Infectious Disease, NIH, Bethesda, MD (Advisor: Dr. Stephen Leppla)

2006- Present	PhD Candidate Molecular Epidemiology Malaria Section, Center for Vaccine Development Department of Epidemiology, University of Maryland Baltimore, MD (Advisor: Dr. Christopher Plowe, Howard Hughes Medical Institute)

Conferences and Presentations

2001	University of Florida University Scholars Annual Symposiun, Gainesville, FL Oral Paper: Adenoviral Transfection of PDGF-BB Gene in Ischemic Rat Skin
2002	Southeastern Branch American Society for Microbiology Conference, Gainesville, FL Oral Paper: Effects of Transient Gene Therapy in Ischemic Rat Skin. Recipient of American Society for Microbiology President's Award for best undergraduate presentation
2004	Pre-IRTA Poster Day Presentation, National Institutes of Health, Bethesda, MD Poster:Effects of Lethal Toxin on expression of Epithelial Sodium Channel (ENaC) in Alveolar Lung Cells
2008	Doris Duke Charitable Foundation Clinical Scientist Meeting, Newport, RI Abstract: Variation in the Circumsporozoite Protein of *Plasmodium falciparum*: Implications for Malaria Vaccine Development
2009	National MD/PhD Student Conference, Keystone, CO Poster: Variation in the Circumsporozoite Protein of *Plasmodium falciparum*: Implications for Malaria Vaccine Development

2010	Genomic Epidemiology of Malaria Meeting, Cambridge, United Kingdom Abstract: Variation in the Circumsporozoite Protein of *Plasmodium falciparum* Implications for Malaria Vaccine Development
2010	American Society for Tropical Medicine and Hygiene, Atlanta, Georgia Poster: Variation in the Circumsporozoite Protein of *Plasmodium falciparum* Implications for Malaria Vaccine Development
2011	Gordon Conference: Malaria, Barga, Italy Poster: Variation in the Circumsporozoite Protein of *Plasmodium falciparum*: Implications for Malaria Vaccine Development

Honors and Awards

1999	Harvard Prize Book for Academic Excellence
1999-2003	Florida Bright Futures Academic Scholarship
1999-2003	Dean's Award for Academic Excellence
2000	Honors Scholarship for Study Abroad Brigham Young University
2001	Golden Key International Honors Society, May 2001 – present
2001	Anderson's Scholar for Outstanding Academic Accomplishment
2001	University Scholar Research Stipend Recipient
2002	Presidents Award-Best Undergraduate Presentation at SE Branch AMS Conference
2003-2004	Post-baccalaureate Intramural Research Training Awardee
2007	Doris Duke Malaria Research Fellowship Awardee 2007
2008-2009	American Medical Student Association Global Health Scholar

Publications

1. Kavita Gandhi, Mahamadou A. Thera, Drissa Coulibaly, Karim Traore´, Ando B. Guindo, Ogobara K. Doumbo. Next Generation Sequencing to Detect Variation in the Plasmodium falciparum Circumsporozoite Protein. *Am. J. Trop. Med. Hyg.,* 86(5), 2012, pp. 775–781

Techniques

Polymerase Chain Reaction Assays
Sanger sequencing
Pyrosequencing
454 next generation sequencing
Malaria smear prep
Parasite speciation via microscopy
Gel Electrophoresis
Real Time RT-PCR
High resolution electrophoresis
Cell culture
Molecular cloning
Gene transfection

Relevant Course Work

Graduate Epidemiology Courses

Principles of Epidemiology
Principles of Biostatistics
Infectious disease Epidemiology
Molecular Epidemiology
Research Practicum I/II
Computational Analysis
Observational Studies in Epidemiology
Statistical Methods in Epidemiology
Regression Analysis
Statistics for Molecular Biology
Genetic Epidemiology

Medical School Courses

Introduction to Clinical Medicine I & II
Structure and Development
Cell and Molecular Biology
Neuroscience
Functional Systems
Host Defenses and Infectious Diseases
Pathophysiology and Therapeutics I & II
Pediatric Clerkship
Obstetrics and Gynecology Clerkship
Surgical Clerkship
Internal Medicine Clerkship
Neurology Clerkship
Family Medicine Clerkship
Psychiatry Clerkship
Emergency Medicine Clerkship-UMD

Emergency Medicine Clerkship-Mercy
Shock Trauma Critical Care Sub-Internship
Internal Medicine Sub-internship-VA Hospital Baltimore

Organizations

2007	Global Health Student Organization President, Co-founder University of Maryland, Baltimore, MD
2005-Present	Medicine for Mali Fund and Student Elective Co-founder University of Maryland School of Medicine, Baltimore, MD
2006-2007	University of Maryland Chapter of Physicians for Human Rights President University of Maryland School of Medicine, Baltimore, MD

Title of Dissertation: Variation in the Circumsporozoite Protein of *Plasmodium falciparum*: Implications for Vaccine Development.

Kavita Gandhi, Doctor of Philosophy, 2013

Dissertation directed by: Christopher Plowe, MD, MPH

Professor

Department of Medicine

Background: A leading malaria vaccine candidate, RTS,S/AS01, is based on immunogenic regions of *Plasmodium falciparum* circumsporozoite protein (CSP) from the 3D7 variant, and has shown modest efficacy against clinical disease in African children. It is unclear, however, what aspect(s) of the immune response elicited by this vaccine are protective. Better understanding of how diversity in the immunogenic regions of CSP (T-cell and B-cell epitopes) may relate to clinical immunity is needed to evaluate and improve the efficacy of vaccines based on CSP.

Objectives: The goals of this study were to measure diversity in immunogenic regions of CSP in a natural population of parasites and identify associations between variation in amino acid sequences in CSP and the risk of infection and clinical disease caused by *P. falciparum* in African children.

Methods: One hundred children were selected from those who had participated in a prospective cohort study designed to measure incidence of malaria infection in Bandiagara, Mali. DNA was extracted from 769 parasite-positive blood samples corresponding to both acute clinical malaria episodes and asymptomatic infections detected in monthly surveys and B- and T-cell epitope-encoding portions of the *cs* gene were sequenced. Non-synonymous SNP data were generated via 454, a next generation sequencing technology, for the T-cell epitopes and repeat length data was generated for the B-cell epitopes of the *cs* gene. Cox proportional hazards models were used to determine the effect of sequence variation in consecutive infections occurring within individuals on the time to new infection and new clinical malaria episode.

Conclusions: Extensive diversity was found in the T-cell epitopes, but no associations were found between sequence variation in either the T-cell epitopes or the repeat region, and hazard of infection or clinical malaria, suggesting that naturally acquired immunity to CSP may not be allele-specific.

Variation in the Circumsporozoite Protein of *Plasmodium falciparum*: Implications for Vaccine Development

by
Kavita Gandhi

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland Baltimore in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2013

DEDICATION

I would like to dedicate this body of work to all those who struggle to level the playing field, whether you are a Paul Farmer, a Mother Theresa, or my uncle who has dedicated his time as a physician to fit the limbless with prosthetics in India. However small your contribution, and however insurmountable the task may seem, one life touched is a triumph one thousand times over.

ACKNOWLEDGEMENTS

I am very grateful to my mentor Dr. Christopher Plowe for his dedication to my growth as a scientist over the past seven years. This study could not have come to fruition without his support, intellect, and passion. I am also very thankful for the opportunities he provided me during my dissertation including my experiences in Mali, which have shaped me forever, and the chance to present my work to a global audience, which instilled me with confidence in my scientific abilities.

I would like to thank Dr. Shannon Takala-Harrison for her mentorship and willingness to provide much needed advice. Without the elegant groundwork she laid through her own dissertation work, my study would not have been possible. She continued to inspire me with her brilliance and creativity throughout my years in the malaria section. I would not have been able to reach my goals without her assistance.

I would like to thank the Institute for Genome Sciences for their help in sequencing the CS gene, especially Dr. Jaques Ravel for providing the barcode sequences which made this task possible. I also appreciate the assistance of Christine Jones who spent many hours with me helping me set up sequencing runs.

I am grateful to my committee for their individual contributions to this project. Dr. Colin Stine provided me with invaluable advice on the development of my sequence assays, Dr. Zhan, was instrumental in trouble shooting my Cox models and helping with SAS, and Dr. Jon Furuno gave me key insights into my experimental design and dedicated hours of help as one of my readers.

I would like to thank the MSTP program at the University of Maryland, including Dr. Terry Rogers, Nancy Malson, and Jane Bacon for YEARS of support, encouragement, and selfless dedication to my success in becoming a physician scientist.

Last but certainly not least, I would like to thank my mother Ketki Gandhi, for her unwavering belief in my ability, my father Kiran Gandhi for his pride in me, and my sister Neha Gandhi for putting up with nine years of ups and downs. I would like to thank my best friend and partner Peter Surgent for listening, encouraging, and helping me through the hardest parts—I love you all more than you know.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

454	Next generation sequencing platform
3D7	Lab variant of *P. falciparum*
AMA	Apical membrane protein
ACT	Artesunate combination therapy
AS01/2	Adjuvant system including in the RTS,S vaccine
bp	base pair
cs	gene encoding the circumsporozoite protein
CSP	Circumsporozoite protein of *P. falciparum*
Dd2	Lab variant of *P. falciparum*
DDT	dichlorodiphenyltrichloroethane, an organochlorine insecticide
DNA	Deoxyribonucleic acid
FMP1	Subunit vaccine based on the merozoite surface protein 1 of *P. falciparum*
FMP2.1	Subunit vaccine based on the apical membrane antigen 1 of *P. falciparum*
Hb3	Lab variant of *P. falciparum*
HLA	Human leukocyte antigen
IGS	Institute for Genome Sciences
ITN	Insecticide treated bed net
MSP	Merozoite surface protein
NANP	Tetramer repeat of targeted by antibodies against the

NVDP	Tetramer repeat of targeted by antibodies against the
PCR	Polymerase chain reaction
RTS,S	Subunit vaccine targeting CSP
SNP	Single nucleotide polymorphism
SP	Sulfadoxine-pyramethamine
Th2R	Helper T-cell epitope of the circumsporozoite protein
Th3R	Helper T-cell epitope of the circumsporozoite protein

I. INTRODUCTION AND OBJECTIVES

A. Introduction

In the quest for an effective malaria vaccine, substantial resources are being invested in clinical trials to evaluate the safety, immunogenicity and efficacy of vaccines targeting specific immunogenic antigens of *Plasmodium falciparum*. To date, vaccine development and testing has generally not been informed by molecular epidemiological evidence of how genetic diversity in these antigens in parasite populations may affect vaccine efficacy. For example, vaccines that confer variant-specific protection may not be effective in a parasite population in which the vaccine variant is rare. Furthermore, vaccines may create a selective advantage favoring non-vaccine variant types, by changing allele distribution, thus compromising vaccine efficacy.[1]

Vaccines that target the different stages of the parasite's life cycle are currently in development. The most clinically advanced vaccine to date, RTS,S/AS01, targets the sporozoite stage of the malaria parasite's life cycle, which will be discussed further in the background section. The antigen that this vaccine is based on is the circumsporozoite protein (CSP) of *P. falciparum*. This vaccine has been tested in Phase 2 efficacy trials, yielding modest efficacy in African children, and is the only malaria vaccine to be evaluated in a Phase 3 trial. It is still not clear, however, which aspects of the immune response elicited by this vaccine are protective, or what factors affect efficacy. One of these factors may be variation in the amino acid sequence of the antigen in the parasite population.[2, 3] Understanding how diversity in the immunogenic regions of CSP (both T-cell and B-cell epitopes) may affect host infection and clinical disease can aid in the correct evaluation and improvement of vaccines based on CSP.

B. Study goals and Implications

The main objectives of this study were to describe the diversity present in the two immunogenic regions of CSP in a natural parasite population and to determine if particular polymorphic sites in these regions may be important in determining naturally acquired allele-specific immunity to CSP. The results of this study contribute to a body of evidence that aids in interpreting efficacy data from clinical trials of vaccines based on CSP.

C. Research question

Is variation in immunogenic regions of the circumsporozoite protein (CSP) of the malaria parasite, *Plasmodium falciparum,* associated with the risk of malaria infection and clinical disease?

D. Specific aim 1 and hypotheses

To understand the dynamics of polymorphism in the Th2R and Th3R epitopes of CSP in the parasite population at the study site, and to determine if specific polymorphic sites in these epitopes are important in determining allele-specific immunity to CSP.

1. Sub-aim 1

To describe the diversity present in the Th2R and Th3R epitopes of CSP in the parasite population represented by samples collected in a malaria incidence study in Bandiagara, Mali and to determine if the distribution of this diversity differs among important study covariates.

2. Hypothesis for sub-aim1

The prevalence of polymorphisms found in the Th2R and Th3R epitopes of CSP will not be different between age groups, clinical and non-clinical cases, or change significantly over time within the study period.

3. Sub-aim 2

To determine if changes in the predominant amino acid at specific polymorphic sites in the T-cell epitopes in an individual's consecutive infections are associated with an increased hazard of new infection and clinical disease.

4. Hypothesis for sub-aim 2

Changes in the predominant amino acid at specific polymorphic sites in T-cell epitopes are significantly associated with and increased hazard of new infection and clinical disease.

E. Specific aim 2 and hypotheses

To understand the dynamics of polymorphism in the in the B-cell epitopes of the CSP protein in the parasite population at the study site and to determine if polymorphism in these epitopes are important in determining allele-specific immunity to CSP.

1. Sub-aim 1

To describe the diversity present in the B-cell epitopes of CSP in the parasite population represented by samples collected in a malaria incidence study in Bandiagara, Mali and to determine if the distribution of this diversity differs among important study covariates.

2. Hypothesis for sub-aim 1

The prevalence of polymorphisms found in the B-cell epitopes of CSP will not be different between age groups, clinical and non-clinical cases, or change significantly over time within the study period.

3. Sub-aim 2

To determine if changes in the amino acid sequence or repeat length of the B-cell epitopes in an individual's consecutive asymptomatic infections are associated with an increased hazard of new infection and clinical disease.

4. Hypothesis for sub-aim 2

Changes in the amino acid sequence or repeat length of the B-cell epitopes are significantly associated with an increased hazard of new infection and clinical disease.

II. BACKGROUND AND SIGNIFICANCE

A. The malaria problem

Malaria is a life-threatening disease that has a significant effect on morbidity and mortality world-wide. The World Health Organization estimated 655,000 deaths from malaria world-wide in 2010,[4] but a recent study reports that this number may be a gross underestimation. The Institute for Health Metrics and Evaluation at the University of Washington recently estimated that the total mortality may be closer to 1.24 million, taking into account the deaths of children over the age of 5.[5] Approximately half of the world's population lives in malaria endemic regions, with sub-Saharan African bearing the brunt of the incidence and mortality of the disease.[6] Efforts to control malaria have

largely been directed towards *P. falciparum* malaria because it is the most common of the five species of Plasmodia that affect humans and causes the most severe forms of the disease. Reducing the burden of malaria is one of eight Millennium Development Goals, set forth by the United Nations to improve the quality of life for people around the world.[7] Even though goals have been set, progress towards curbing the global effects of malaria has been slow due to many factors including drug and insecticide resistance of the parasite and the vector respectively, as well as lack of an effective vaccine.

Parasite resistance to chloroquine, once the first line treatment for *falciparum* malaria throughout the world, was first observed in Asia 50 years ago, and now exists almost everywhere that *P. falciparum* does.[8] The emergence and spread of chloroquine resistance prompted a shift in standard therapy to other drugs, including sulfadoxine-pyrimethamine (SP)[9], and now most recently artesunate combination therapies or ACTs.[10] Unfortunately, resistance to SP also arose rapidly in Asia and from there, it has been demonstrated that it spread to Africa[1112], and more recently within the past 5 years reports of artesunate resistance have been emerging from Asia.[13, 14] Treatment failure due to parasite resistance to the mostly commonly used antimalarial drugs is a huge public health problem and continues to make efforts to control malaria more difficult.

Vector resistance to insecticides is another serious issue impeding the progress of malaria control. In 1955, a massive campaign to eradicate malaria world-wide was undertaken and mainly involved indoor residual spraying of the insecticide DDT to kill the mosquito vector that transmits malaria as well as use of chloroquine to treat cases. Although this resulted in successes in more temperate regions, exhaustion of resources, lack of follow up, and widespread DDT and chloroquine resistance prompted a shift in

the goal from eradication back to control.[15] Since then resistance has followed with the switch to different insecticides including organophosphates, carbamates and pyrethroids.[16] Reports of pyrethroid resistance in west and south Africa[17-19] are perhaps the most problematic as this is the insecticide of choice used in insecticide treated bed nets (ITN's), one of the most effective interventions in preventing malarial disease.[20]

Recently, the concept of "vaccine resistant malaria" has been described as an incipient problem that could ultimately prove to be as serious an obstacle to 21st century efforts to control malaria as drug and insecticide resistance were to antimalarial campaigns in the 20th century.[1] This concept will be discussed in depth in the following section.

B. Malaria life cycle and vaccine development

With resistance limiting the success of treatment and vector control strategies, an effective vaccine could represent a critical tool for reducing the burden of malaria and reaching the eventual goal of elimination. However, progress towards a vaccine has been slow, in part due to the complicated biology of the malaria parasite and its life cycle. Malaria is caused by protozoan parasites of the genus *Plasmodium* that have both diploid and haploid stages in their insect vectors and vertebrate hosts. The *P. falciparum* genome consists of approximately 5,000 genes encoded on 14 chromosomes, and the antigens expressed on the surface of the parasite and on the host cells it invades are different in each stage of its life. When an infected mosquito bites a susceptible host, the sporozoite form of the parasite is inoculated into the host's blood stream. The sporozoite is coated with the circumsporozoite protein (CSP), encoded by the *cs* gene.[21] The sporozoite

quickly invades hepatocytes and forms liver schizonts which burst to release 10-30,000 merozoites which then infect red blood cells and cause anemia.

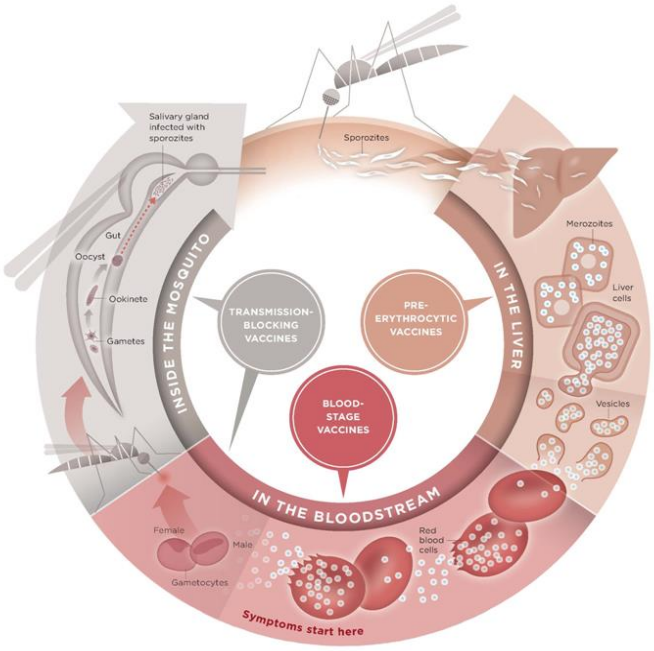

Figure II.1 Malaria life cycle stages targeted by interventions [22]

Vaccines that target the different stages of the parasite's life cycle (Figure II.1) are currently in development. The most clinically advanced of these are described in Figure II.2. Prominent antigens of the blood stage (or erythrocytic stage) that have been the focus of vaccine development include apical membrane antigen-1 (AMA1), and merozoite surface protein-1, (MSP1).[23] Vaccines targeting AMA1 and MSP1, aim to reduce parasitemia by boosting immunity to merozoites in the bloodstream. These vaccines could help ameliorate the burden of clinical disease in populations living in malaria endemic regions.[3] Other vaccines including RTS,S/AS01 and an irradiated whole-organism sporozoite vaccine[24] target the pre-erythrocytic stage of the parasite life cycle by inducing immunity to the sporozoite. Of all the malaria vaccines currently in

development, RTS,S/AS01, developed in a partnership between GlaxoSmithKline Biologicals, the U.S. Army and the PATH Malaria Vaccine Initiative,[25] is the most advanced in clinical development. The goal of this vaccine, as well as of other vaccines targeting the pre-erythrocytic stage, is to prevent liver invasion or parasite escape from the liver and thereby block infection.[26,27] Ideally, vaccines such as these, which have the potential to induce sterilizing immunity to *P. falciparum,* would prevent infection.

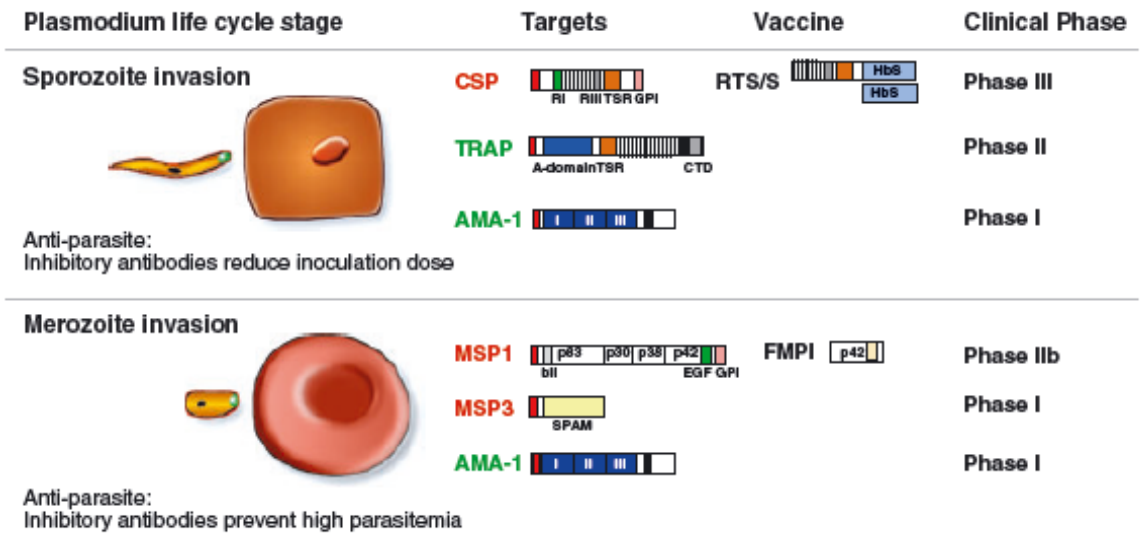

Figure II.2 Malaria vaccine target antigens by life cycle stage [28]

C. Challenges facing vaccine efficacy

No pre-erythrocytic vaccine has yet demonstrated this potential, in terms of being able to elicit an immune response that is completely effective in preventing infection in study participants when tested in malaria endemic regions. One factor that may contribute to this failure to achieve the ideal of complete prevention of infection is the genetic diversity of the parasite. All malaria vaccine antigens being targeted to date are polymorphic to varying degrees. It has been demonstrated that immunity to certain

malaria antigens may be allele-specific, meaning that the immune response elicited is only effective against specific genetic variants (which can be thought of as variants or serotypes). A study conducted in Gambian children found that naturally acquired antibodies against MSP3, a blood stage malaria antigen, protected in an allele-specific way.[29] If a vaccine based on only one genetic variant of this antigen was tested in this population, and if this genetic variant was rare, it follows that the vaccine could have poor efficacy. This concept is further evidenced by a study conducted in Mali[30] of variation in the MSP1 vaccine antigen, on which the FMP1/AS02 malaria vaccine candidate is based. This study found the MSP1 variant corresponding to the vaccine variant to be a minority variant in the population. Additionally this study identified specific polymorphic sites within MSP1 that appeared to play a role in determining allele specific immunity against this protein. A randomized, double blind, Phase IIb trial of this vaccine in Kenya showed no efficacy in preventing clinical cases of malaria in children aged 12 to 47 months as compared to a rabies vaccine.[31] It is yet to be determined whether the low frequency of the vaccine target allele in the parasite population contributed to the lack of observable clinical efficacy. Yet another clinical trial reporting evidence of allele specificity focused on the efficacy of the FMP2.1/AS02 vaccine based on the AMA1 protein of the 3D7 variant of FMP1.[32] In this double-blind randomized trial 400 Malian children were randomized to receive either the FMP2.1/AS02 vaccine or the rabies vaccine. In this trial the primary endpoint was clinical malaria, defined as a febrile illness with greater than or equal to 2500 parasites on microscopy. The overall efficacy in preventing the primary endpoint was only 17.4% and not significant, but efficacy against

clinical malaria caused by a parasite identical to the vaccine variant was 64.3%, indicating that immunity to AMA1 is, at least in part, allele-specific.

It follows that a vaccine based on a variable protein that confers allele-specific immunity may select for non-vaccine variants in the population, as occurred in a trial of a vaccine based on another erythrocytic vaccine antigen, MSP2, in Papua New Guinean children.[2] This issue of "vaccine resistance" is potentially very important issue because if the vaccine only protects against the variant on which it is based, and it is selecting for non-vaccine type variants in vaccinated populations, then both initial and long-term vaccine efficacy may be compromised. The findings of the studies described above highlight the need for molecular epidemiological studies of vaccine antigens that will help guide vaccine development and interpret efficacy data from vaccine trials.

D. Description of the CSP antigen and the RTS,S vaccine

The immunogenic regions of CSP on which the RTS,S/AS02 vaccine is based are polymorphic and can have a high degree of variability in natural parasite populations.[33] This surface protein contains a central repeat region that is important in antibody responses, and two T-cell epitopes that are important in CD8+ and CD 4+ responses.[34, 35]

The central repeat region can vary in both length and in sequence of tetrameric amino acid repeats. Several studies of diversity in this region have been conducted. One study that considered diversity in the repeat regions from 75 samples from geographically diverse regions found that the most common repeats were NANP and NVDP.[36] This study also found a range of 37 to 49 repeats with an average of 41 repeats per allele. Another study explored the diversity in a set of 25 samples, mostly from Asia, found that there was considerable conservation of how the repeats were organized along the

sequences and found that the NVDP tetramer is always preceded by the NANP tetramer.[37] This region is the target of antibodies produced by B-cells. Numerous studies have been done to demonstrate that an antibody response to CSP is important in protection against the pre-erythrocytic malaria parasites. [38, 39]

The T-cell epitopes are polymorphic in that they contain several single nucleotide polymorphisms (SNPs) that code for changes in amino acid sequence. The level of diversity in this region increases with increasing malaria transmission, as evidenced by studies from regions of varying malaria endemicity. In order of increasing levels of malaria transmission, four Th2R–Th3R haplotypes were observed in Peru (n = 139)[40], 20 in Vietnam (n = 142) and 24 haplotypes were observed in a smaller set of samples from the Gambia (n = 44)[41]. Cell mediated (T-cell) immune responses have also been shown to play an important role in protection to pre-erythrocytic parasites.[42 43, 44] If immunity to polymorphic regions of CSP is either partly or fully allele-specific, a vaccine based on a specific clone may protect against infection with parasites that carry homologous CSP but may be less effective against parasites with divergent CSP sequence.

The RTS,S/AS01 vaccine contains the 3D7 variant of each of these polymorphic regions. 3D7 is a common laboratory variant of *P. falciparum*, the genome of which has been sequenced and assembled, and has been incorporated into a searchable database on the National Center for Biotechnology and Information (NCBI) website.[45] A schematic representation of the vaccine is shown in Figure II.3. The amino acid sequence of the Th2R and Th3R in the 3D7 reference genome are KHIKEYLNKIQNSL and NKPKDELDYAND, respectively. The central repeat region coding for the B-cell epitopes contain 37 tetramers of the amino acid sequence NANP, and 4 tetramers of

NVDP. In initial clinical trials, formulations containing CSP alone were poorly immunogenic, so it was fused to the hepatitis B surface antigen. The fusion protein greatly improved immunogenicity.[46] The vaccine is also formulated with the AS01 adjuvant system which contains the immunostimulants monophosphoryl lipid A and saponin.[47]

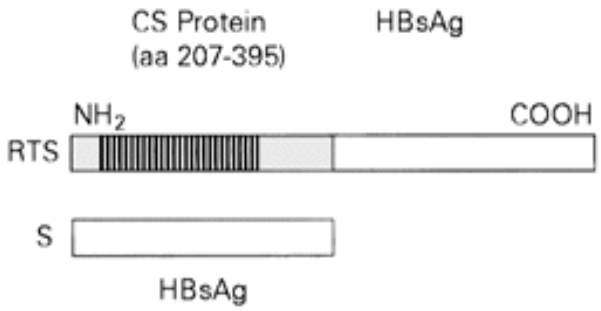

Figure II.3 Schematic representation of the antigen contained in the RTS,S vaccine.[48]

E. Clinical trials of RTS,S

Several clinical trials of this CSP-based vaccine in malaria endemic regions have tested its safety, immunogenicity and efficacy. One trial was conducted in semi-immune adults in The Gambia, who have acquired partial immunity to clinical disease through repeated exposure to the malaria parasite during childhood and young adulthood[49], another was conducted in children between the ages of 1 and 4 years in Mozambique[50], one in infants in Mozambique [51], and the most recent in children ages five through 17 months in Kenya and Tanzania.[52] In all trials, the vaccine proved to be safe and showed measurable but limited efficacy in preventing infection and disease. The trial in The Gambia showed that infections occurred significantly earlier in the control group, and overall, the vaccine showed an efficacy of 34% in preventing infection during the 15 week follow-up period. In the trial in children in Mozambique the vaccine had 45%

efficacy against infection, 30% efficacy against uncomplicated malaria, and 58% efficacy in preventing severe malaria over a 6-month period. The trial in infants showed almost 65% efficacy against infection; this higher efficacy estimate may be due to a shorter follow-up period (3 months) and/or to lower baseline immunity in infants. Finally, the trial in Kenya and Tanzania was conducted in order to determine what formulation of the RTS,S vaccine should enter large scale, multi-site phase III trial in Africa. Previous evidence indicated that a reformulation of the AS0 adjuvant system as a liposomal rather than oil in water suspension may make the vaccine more clinically effective.[53-55] This study concluded that the overall efficacy of the vaccine against clinical malaria with the liposomal adjuvant, AS01, was higher (49%), than estimates of efficacy with the AS02 adjuvant from previous trials in this age group. The authors recommended that the RTS, S/AS01 be further tested in a Phase III trial, which is currently in progress at 11 African sites.

Preliminary efficacy data in children 5 to 17 months from the Phase III trial is similar to that previously reported in Phase II trials, however efficacy in infants was lower. Intention to treat efficacy in preventing uncomplicated malaria and severe malaria was 50.4% and 45.1% in children aged 5 to 17 months.[56] Intention to treat efficacy in infants 6 to 12 weeks of age was 30.1% against uncomplicated malaria, and 26% against severe malaria.[57]

A follow-up study to determine whether vaccination selected for non-vaccine types was conducted for the pediatric trial in Mozambique.[58] In samples from the trial in children, sequence variation was examined in TH2R and TH3R regions of the *cs* gene from infections occurring in both vaccinated and control individuals. No significant

differences were found in the distribution of CSP alleles between the groups. It was concluded from this information that RTS,S/AS02 does not select for non-vaccine alleles in the trial participants. A second follow-up study to a Phase 2 trial in adults in Kenya which tested two formulations of RTS,S was also looked at selection of non-vaccine variants in vaccinated individuals.[59] This study reported two polymorphic sites (D371 and Q339) at which there was a statistical difference in the proportion of non-vaccine variant alleles between vaccinated and control groups. At D371, authors found a lower proportion of non-vaccine variant alleles in the vaccinated group, and at Q339 they found a higher proportion of non-vaccine variant alleles in one of the RTS,S groups. Since the effects were in opposite directions and selection at Q339 was only found in one of the vaccine groups the authors concluded that this evidence was not strong enough to suggest vaccine selection.

Limitations in the study design and methods, however, may have reduced both these studies' ability to detect evidence of allele-specific efficacy. First, the majority of samples used in the follow-up study to the Phase 2 trial in children, were from asymptomatic infections. Since RTS,S has the potential to prevent both infection as well as disease, it is important to consider whether selection is occurring in asymptomatic as well as clinical cases. The total number of clinical infections analyzed in this study was 45. Considering the large amount of diversity in Th2R and Th3R this sample size may have been insufficient to detect differences between study groups.

In addition, both studies excluded polyclonal infections (those infections containing more than one genetic variant of *falciparum* parasites), which made up one-third or greater of the total sample size in both cases. This omission may have biased

their findings if selection happens more frequently in polyclonal infections, as would be expected. Furthermore, neither author specified what criteria they used to determine, single allele, majority allele, or polyclonal infections. If criteria were not strict enough, misclassification bias could have occurred.

Finally, the study did not consider the repeat region, which codes for the B-cell epitopes and is important in eliciting an antibody response to CSP. Titers of antibodies to CSP are used as immunogenicity endpoints in all of the vaccine trials described above. Even though the repeat region is thought to be functionally "invariant" because the dominant epitope NANPNANP is always present,[60] it is reasonable to hypothesize that length polymorphism and such sequence polymorphism as exists in the repeat region may be driven by immune selection pressure, and this region of the *cs* gene should therefore be included in an evaluation of vaccine induced selection.

Further analysis is required to ascertain whether that RTS,S/AS01 confers exclusively allele-specific protection and exerts strong selection for non-vaccine alleles in a parasite population with regard to T-cell epitopes. Moreover, studies of allele-specific efficacy with this modestly efficacious vaccine do not address the question of how diversity in TH2R and TH3R may affect naturally acquired immunity. The within host dynamics of polymorphism in the immunogenic regions of CSP in the context of naturally acquired protective immunity to malaria infection and disease have not been studied. An understanding of these dynamics is key in evaluating efficacy of a vaccine based on these regions. Furthermore, studies have not been done to assess how variation in the B-cell epitopes of CSP affects immunity. The reason for this may be three-fold. First, the role that antibodies play in protection is undefined, and some researchers assert

that cell-mediated immunity is more important.[61,58] Secondly, the repeat region has been notoriously difficult to sequence and align and technological limitations may have prevented closer investigation of this region.[37] Finally, the predominant epitope in the B-cell region does not seem to vary from parasite to parasite from the same or different geographic regions. However, results from RTS,S/AS01 efficacy trials have shown a positive correlation with anti-circumsporozoite antibodies induced by administration of the vaccine and decreased hazard of infection.[49-52] It has been hypothesized that even if these epitopes are not involved directly in protection, they may be involved in structural stability of the protein[36] and the exposure of important antigenic sites to the immune system. In addition to this, we now have a variety of technological tools, which are elaborated upon in the next section, which may help circumvent the difficulties of studying the repeat region

III. STUDY DESIGN AND METHODS

A. Parent study design

The study in which the samples to be analyzed were collected was a prospective cohort study designed to measure the age-specific incidence of malaria infection and disease in children and young adults in Bandiagara, Mali. [62]

1. Study site

The study was conducted in Bandiagara, a rural town of approximately 13,000 inhabitants located in Central Eastern Mali. Transmission of *falciparum* malaria is intense with the peak coinciding with the rainy season from July-October. Children aged less than 10 years experience a mean of two clinical episodes a year and prevalence of parasitemia at the beginning of the study was 17%.[62]

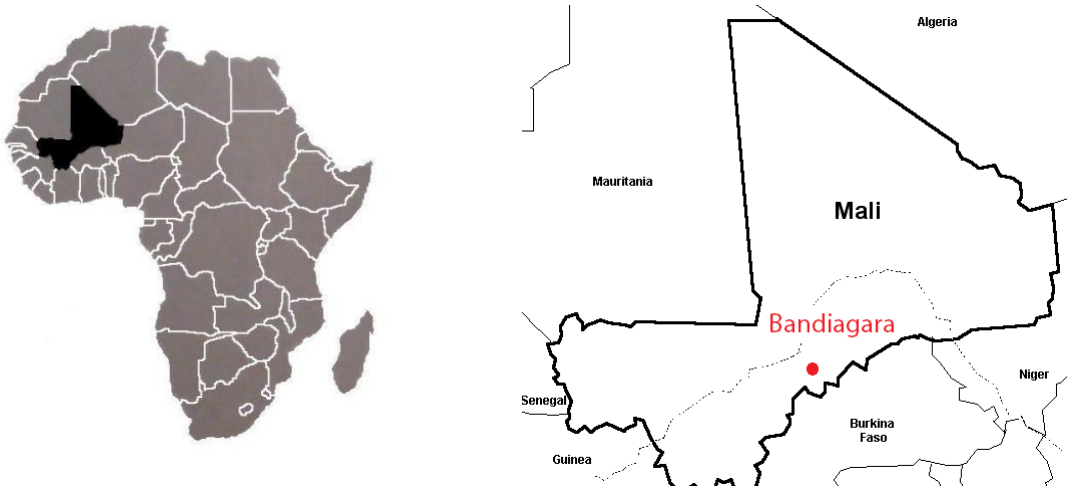

Figure III.1 Field study site

2. IRB Approval

Samples were collected under protocols reviewed and approved by Institutional Review Boards of the University of Maryland School of Medicine and the University of Bamako Faculty of Medicine. Informed consent was obtained from study participants or their guardians. Permission to work in the community was obtained from local officials, elders, and traditional healers. Individual written informed consent was obtained by subjects, parents or guardians.

3. Subjects

A complete population census was conducted in Bandiagara prior to the initiation of the study and as part of the development of a site for testing vaccines. The dominant ethnic group in Bandiagara is Dogon (80%) with 10% being Peuhl, 3% Bambara and 7% falling in the other category. Subjects were sampled in proportion of the 8 districts comprising the town of Bandiagara. Study subjects were aged ≥ 3 months to 20 years. Age groups were defined as <2, 2-4, 5-7, 8-10, 11-14, and 15-20 years of age.

Recruitment of study subjects took place in randomly selected households until the target number of subjects in each age group was achieved.

4. Parent study methods

The study was conducted prospectively during the years 1999, 2000, and 2001. From July to January of each year, blood samples were collected on 3MM Whatman filter paper at least monthly and at every episode of clinical malaria. Clinical malaria episodes were detected both by passive surveillance through provision of around-the-clock free clinical care, as well as by active weekly follow-up of the children in the study by study physicians working at the study site. Clinical episodes were defined as a blood smears positive for *P. falciparum* asexual parasites and symptoms consistent with malaria including, fever, anemia headache, body aches, cough, diarrhea, and abdominal pain. Infections are defined as the microscopic presence of *P. falciparum* parasites in the blood, with or without syptoms.[62]

B. Present study description

Samples from 100 children with at least two years of follow up during the malaria incidence study were used. These children were randomly selected within three age strata. Thirty children aged ≤5 years, 32 children aged 6 to 10 years, and 38 children aged ≥11 years were selected. Blood samples (n=2309) corresponding to all monthly surveys (n=1801) and clinical episodes (n=508) occurring during the transmission season of the three years of the incidence study underwent DNA extraction (QIAamp DNA Mini Kit, Qiagen, Valencia, California). Of these samples 769 were determined to be parasite positive by microscopy and these were selected as the samples to be analyzed in this study.

C. Molecular Methods

In order to substantiate 454, a next generation, high-throughput, technology that generates many parallel sequences from a single sample, as viable method for sequencing the immunogenic regions of the *cs* gene, a pilot study was performed. In this study 454 was compared to Sanger sequencing, a method that has been used to generate sequence data in previous studies for both the Th region (containing both Th2R and Th3R) as well as the repeat region. A set of 45 randomly selected samples from the incidence study described above (but not from the 100 children selected for inclusion in the main study) were sequenced by both 454 as well as Sanger sequencing.

1. PCR amplification for 454

Two nested PCR assays were designed to amplify a region of the *cs* gene containing both Th2R and Th3R, and the central repeat region. The primary PCR was designed to amplify a portion of the gene which contains both regions, and the secondary PCRs amplify Th2R and Th3R, as well as the central repeat region individually. The primary forward and reverse PCR primers were GTTGAGGCCTTTTCCAGGAATACCAG and GTACAACTCAAACTAAGATGTGTTC. Primary PCR conditions are as follows: 30 cycles of 95 °C for 30s, 52 °C for 30s, 72 °C for 1 min. Secondary PCR conditions for the repeat regions and ThRs were 30 cycles of 95 °C for 30s, 55 °C for 30s, 72 °C for 1min 30s, and 25 cycles of 95 °C for 30s, 58 °C for 30s, and 72 °C for 1min respectively. PCR products were amplified using HotStar Taq (Qiagen, Valencia, California).

Secondary PCR primers for both the Th region as well as the repeat region contained specific adapters necessary for the emPCR[63] step of 454 sequencing, followed

by a 12 base pair barcode, and finally the sequence specific primer. Forward and reverse adapter sequences were: CCATCTCATCCCTGCGTGTCTCCGACTCAG and CCTATCCCCTGTGTGCCTTGGCAGTCTCAG. Forty-five unique barcodes were used to tag PCR products from the 45 study samples. Forward and reverse sequence specific primers were AAATGCTAATGCCAACAGTGC and GACCTTGTCCATTACCTTG. The same barcode was used to amplify the Th region and the repeat region from the same sample. The concentration of each PCR product was determined by band intensity compared to a standard of similar molecular weight using geneSNAP software, and 100 ng of each product was then pooled. PCR products for each region were pooled separately into two tubes.

2. 454 Sequencing

Pooled PCR products were sequenced at the University of Maryland School of Medicine Genomic Resource Center on the GS FLX Titanium 454 Platform (Roche Diagnostics, Branford, CT).

Sequences were aligned using gsAmplicon (Roche Diagnostics, Branford, CT) software. For samples containing more than one allele at a polymorphic site, predominance was determined if the majority allele was present in 71% or more of all reads obtained for that sample. If a majority allele could not be determined, that polymorphic site was considered mixed. Haplotype information, however, was still obtained for samples with mixed polymorphic sites. A sensitivity analysis was performed to determine a threshold for inclusion of minor alleles in the total number of SNPs discovered in for both 454 and Sanger sequencing. Based on the curves generated (see manuscript 1 in Results section), the largest numbers of new SNPs were found between

the frequencies of 0.025 and 0.01 for 454, and between 0.2 and .15 for Sanger sequencing. As the discovery of rare SNPs was not a goal of the study a conservative threshold of 0.1 was selected for 454, and 0.2 was selected for Sanger sequencing.

3. Repeat region troubleshooting

After several rounds of troubleshooting, complete sequence data could not be generated for the repeat region using 454 in this study. Methods attempted to improve read quality and length included sequencing the region from both the forward and reverse directions using two different sequencing kits (GS FLX and GS Titanium). It was determined that the current technology has limitations that preclude the generation of complete reads for this region, as outlined in the discussion section.

4. PCR amplification for Sanger sequencing

The primary PCR primers and conditions for Sanger sequencing were identical to those used for 454 sequencing. Secondary sequence specific primers as well as PCR conditions for the Th regions were also the same as those described for 454 sequencing. The repeat region was not subjected to Sanger sequencing in this validation study as there were no 454 sequencing results to compare to. PCR products were loaded on a 2% agarose gel, stained with ethidium bromide, and run at 100 Volts for 1 hour. Bands were detected using geneSNAP (Synoptics LTD, Cambridge, UK) gel imaging software.

5. Sanger sequencing

Once amplification was verified by gel electrophoresis, PCR Products were purified by vacuum filtration in Excela Pure (Edge Biosystems, Gaithersburg, MD) 96-

well plates. Purified PCR product was then amplified and sequenced on an ABI3730 xl at the University of Maryland School of Medicine Biopolymer Lab.

Sequences were aligned to the 3D7 reference genome using Sequencher (Gene Codes Corp, Ann Arbor, MI) software. For samples containing more than one allele at a polymorphic site, a predominant allele was designated if the secondary peak height was less than or equal to 40% of the height of the primary peak on the chromatogram for that sample. If the secondary peak was greater than 40% of the height of the primary peak the polymorphic site and sample were designated as mixed and haplotypes were not constructed for these samples. Minor alleles that were not represented by a peak that was at least 20% of the primary peak height were not included in the total number of SNPs discovered in Sanger sequencing output.

6. Standardized mixed infections

A mixture of PCR product containing Th2R and Th3R amplified from laboratory variants (3D7, Hb3, and Dd2) for which the sequences are known was created, quantified and diluted to concentrations of 100 ng/µl, 50 ng/µl, 25 ng/µl, 12.5 ng/µl, and 6.25 ng/µl. 3D7 comprised 60% of each mixture, Hb3 comprised 30%, and Dd2 comprised 10%. Each of the five mixtures was sequenced by both 454 and Sanger sequencing to test the ability of each technology to quantitate the different alleles in a mixture. The observed allele frequencies for the 454 sequencing method were determined by calculating the percentage of reads that contained each type of allele at a given polymorphic site for each concentration. The observed allele frequencies for Sanger sequencing were determined by calculating the relative peak heights of the major and minor allele at each polymorphic site at each concentration. These frequencies were subtracted from the expected

frequency for allele, and the sums of these differences were averaged for each concentration.

D. Molecular methods specific aim 1

The same nested PCR described in the previous section was used to amplify the Th region for specific aim 1. The only difference was that the secondary 454 primers contained different adapter sequences that were used to generate both forward and reverse sequences for each PCR product.

1. Multiplexing and pooling samples

To sequence all 706 samples in a single run, samples were tagged with unique barcodes. Since the 454 sequencing plate can be divided into 16 regions which are physically separated, 16 groups of samples were created. Each group contained a set of barcodes that was used only once within that group. Once sequencing data was obtained for a specific region of the plate, each sample could be uniquely identified by the barcode used to tag it. A total of 96 primers containing 96 unique barcode sequences were used to amplify this region from study samples. Primers were identical with the exception of the barcode sequence. Once PCR products were generated for each sample, the concentration of each sample was determined using the Qiaxcel capillary gel imaging system (Qiagen, Valencia, California). Approximately 100 ng of PCR product was added to its respective pool. Pooled products were submitted to the sequencing core at the Institute for Genome Sciences (IGS) at the University of Maryland.

2. 454 Sequence Analysis

Sequences were aligned using gsAmplicon (Roche Diagnostics, Branford, CT) software. For samples containing more than one allele at a polymorphic site, predominance was determined if the majority allele was present in 71% or more of all reads obtained for that sample. This cut-off correlates to the predominance cut-off chosen for Sanger sequencing (no secondary peak >= 40% or primary peak height in the chromatogram). If a majority allele could not be determined, that polymorphic site was considered mixed. Haplotype information, however, was still obtained for samples with mixed polymorphic sites. Non-synonymous SNPs (those which result in a change in the amino acid sequence of the Th regions were determined by translating the DNA sequence using web based software such as EBI's EMBOSS Transeq program.[64] Sequences that represented at least 10% of the total number of reads in polyclonal infections were resolved into haplotypes. This cut-off was determined to be an appropriate conservative threshold for minority allele inclusion by a sensitivity analysis that was performed in the methods validation study. This will be described in depth in manuscript I in the Results section.

E. Statistical methods for specific aim 1

1. Descriptive analysis

Prevalences of individual polymorphisms as well as haplotypes for TH2R and Th3R were calculated from all of the successfully sequenced samples. Fisher's exact tests were used to make comparisons between clinical and non-clinical *falciparum* cases, the three age groups, and seasons.

2. Cox proportional hazards models

Cox proportional hazards models were used to model the effect of a change in the predominant amino acid present at a specific polymorphic site from one clinical episode or asymptomatic infection to the next on the time interval between those consecutive episodes while taking into account the significant covariates, age and season. Changes at polymorphic sites were coded as binary predictor variables, 0 indicating no change and 1 indicating a change. Twenty-three polymorphic sites between both Th2R and Th3R were identified as predictor variables. Only changes in the nucleotide sequence that resulted in a change in the amino acid sequence (non-synonymous SNPs) were modeled. No adjustments for multiple comparisons were made. This goal of this analysis is to determine which polymorphic sites are associated with increased hazard of infection or disease.

The time of origin for the Cox model of time to new clinical episode and time to new infection was the time of previous clinical episode of *falciparum* malaria. In using this time of origin parasites were cleared from the blood soon after treatment with chloroquine or sulfadoxine pyramethamine and the next consecutive infection with *falciparum* parasites is a new infection. Consecutive intervals shorter than 2 weeks were excluded from the analysis to exclude treatment failures from the analysis. A fixed effects partial likelihood method[65] was used to determine the effect of a change at polymorphic sites when having repeated events from the same individual. These methods allowed for separate baseline hazard functions for each individual or each event and were evoked in SAS by using the STRATA statement in the PHREG procedure.[66] The significant covariates age and season were included in the model.

3. Logistic regression

A secondary analysis using logistic regression to model the log odds of individuals having a change at a certain polymorphic site in intervals in which an asymptomatic parasitemia was followed by a symptomatic one, versus individuals having the same change in intervals in which they are consecutively asymptomatic, was performed. Due to the high level of variability at certain polymorphic sites, changes at these sites may occur frequently in consecutive infections or clinical episodes, and it is necessary to distinguish between baseline variability at a given polymorphic site, and variability that is actually associated with the outcome. This analysis was done to help clarify whether changes at polymorphic sites are occurring by chance or whether they are truly associated with increased risk of clinical disease. Significant covariates of age, and time were included in the model.

F. Molecular methods for specific aim 2

1. PCR amplification

Primary PCR primers and conditions were the same as those used for specific aim 1, and were described in the validation of study methods section. The secondary forward and reverse PCR primers for the repeat region were TGGGAAACAGGAAAATTGG and GCACTGTTGGCATTAGCATTT. Secondary PCR conditions for the repeat region were 30 cycles of 95 °C for 30s, 55 °C for 30s, 72 °C for 1min 30s.

2. Length determination via capillary gel analysis

To determine the length of the repeat region, PCR products were run on a high resolution gel cartridge on a Qiaxcel capillary gel imaging system (Qiagen, Valencia,

California) using the OM500 analysis method capable of resolving size within 10 base pairs (bp). To validate this method, a 96-well plate containing 15ul of repeat region PCR product, for which the exact length of the repeat region is known, amplified from 100 ng/ul 3D7 genomic DNA per well, was run using the method listed above. Once size data were obtained for each well, they were compared to the known length of the 3D7 repeat region. A systematic underestimation of 10bp was observed in the experimental data. A 10bp correction was therefore applied to length values obtained for PCR products from study samples.

3. Sanger sequencing

Repeat region PCR products for samples that were determined to be 'single clone' (no secondary allele present in a frequency greater than 20% in 454 reads) with respect to Th2R and Th3R, were subjected to Sanger sequencing. The 20% cut-off was determined based on a sensitivity analysis performed to determine the cut-off point for real vs. erroneous SNPs using Sanger sequencing. Once amplification was verified by capillary gel, PCR Products were purified by vacuum filtration in Excela Pure (Edge Biosystems, Gaithersburg, MD) 96-well plates. Purified PCR product was then sequenced on an ABI3730 xl at the University of Maryland School of Medicine Biopolymer Lab.

G. Statistical methods for specific aim 2

1. Descriptive analysis

Prevalences of B-cell size polymorphisms were calculated and compared across seasons, age groups, and between clinical and non-clinical *falciparum* infections. . Fisher's exact tests were used to make comparisons between clinical and non-clinical

falciparum cases, the three age groups, and seasons. Haplotypes were calculated from all of the successfully sequenced samples.

2. Cox proportional hazards models

The same Cox proportional hazards methods as described for specific aim 1 were used for aim 2, except change in the length of the repeat region was used as the predictor variable. The significant covariates age and season were included in the model.

3. Logistic regression

The same logistic regression model described in statistical methods for aim 1 was used for the repeat region. The log odds of individuals having a change in repeat length in intervals in which an asymptomatic parasitemia was followed by a symptomatic one, versus individuals having the same change in intervals in which they are consecutively asymptomatic was modeled. Significant covariates of age, and time were included in the model.

H. Sample flow

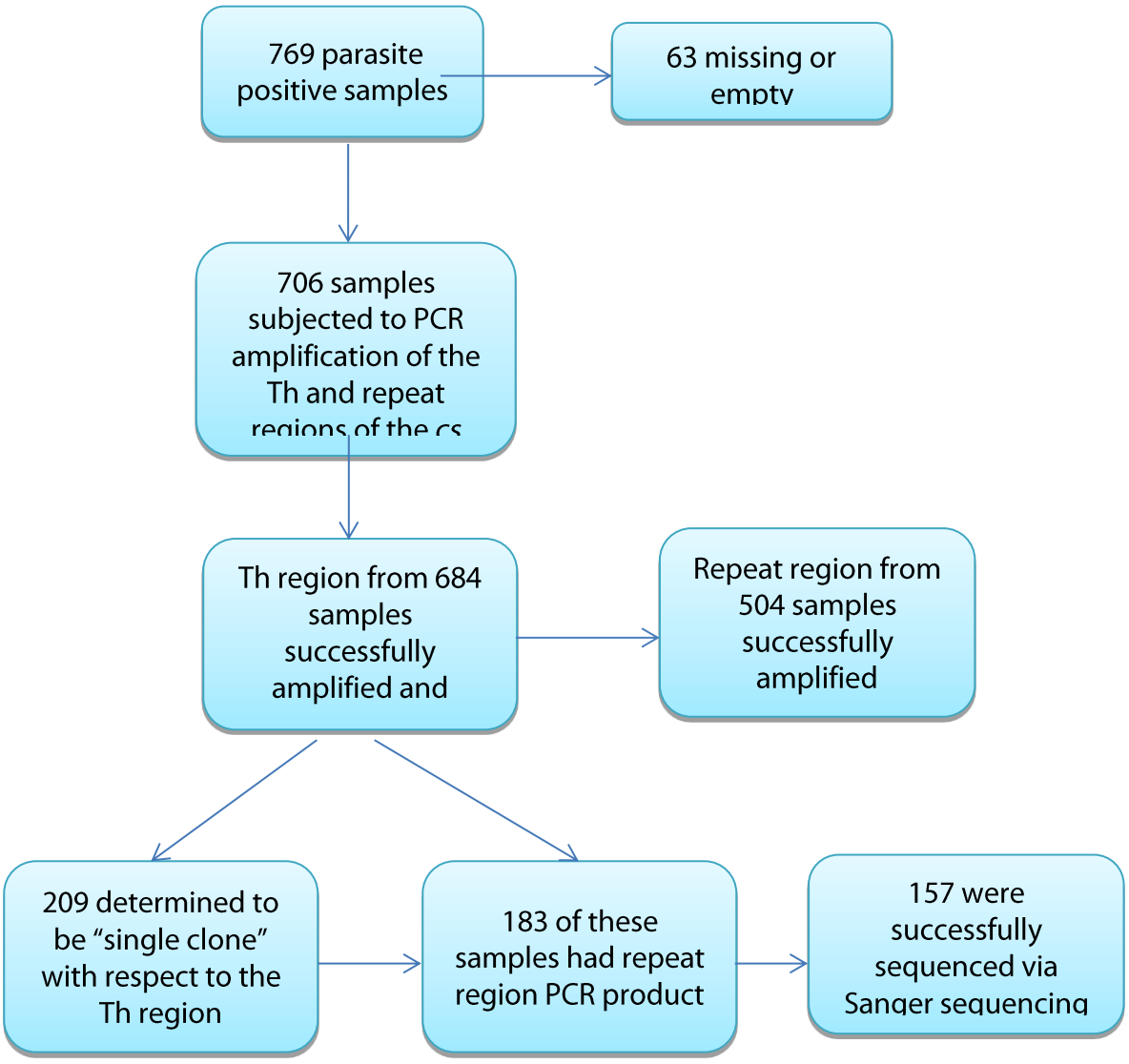

Figure III.2 Sample sizes for each step of data generation

I. Power and sample size calculation

To briefly illustrate that this study was able to detect a reasonable range of hazard ratios a power analysis was conducted using PS power calculation software.[67] Based on an outcome of time to first infection, a median control group survival time of 3 months, one month accrual time and at least 23 more months of additional follow up time for a sample size of 100 children, a significance level (alpha) of 0.05, and 80% power, the following chart was constructed:

Ratio of the number of those without polymorphic change to the number of those with polymorphic change	Minimum Detectable Hazard Ratio
1	1.25
1.5	1.28
2.3	1.36
4	1.51
9	1.86

Table III.1Power calculation showing that study sample size is adequate to detect a reasonable range of hazard ratios.

IV. NEXT GENERATION SEQUENCING TO DETECT VARIATION IN THE PLASMODIUM FALCIPARUM CIRCUMSPOROZOITE PROTEIN

A. Abstract

The malaria vaccine RTS,S/AS01, based on immunogenic regions of the *Plasmodium falciparum* circumsporozoite protein (CSP), has partial efficacy against clinical malaria in African children. Understanding how sequence diversity in CSP T and B-cell epitopes relates to naturally acquired and vaccine-induced immunity may be useful in efforts to improve the efficacy of CSP-based vaccines. However, limitations in sequencing technology have precluded thorough evaluation of diversity in the immunogenic regions of this protein. In this study 454, a next-generation sequencing technology, was evaluated as a method for assessing diversity in these regions. Portions of the circumsporozoite gene (*cs*) from samples collected in a study in Bandiagara, Mali were sequenced both by 454 and by direct sequencing. 454 detected more SNPs and haplotypes in the T-cell epitopes than direct sequencing and was better able to resolve genetic diversity in samples with multiple infections, but failed to generate sequence for the B-cell epitopes

B. Introduction

Substantial resources are being invested in clinical trials to evaluate the potential of vaccines targeting specific immunogenic antigens of *Plasmodium falciparum*, including the circumsporozoite protein (CSP) encoded by the *cs* gene. However, malaria vaccine development and testing has generally not been informed by molecular epidemiological evaluations of how genetic diversity in vaccine antigens in parasite populations may affect vaccine efficacy. For example, vaccines that confer variant-

Kavita Gandhi, Mahamadou A. Thera, Drissa Coulibaly, Karim Traoré, Ando B. Guindo, Ogobara K. Doumbo, Shannon Takala-Harrison, Christopher V. Plowe. Published in the American Journal of Tropical Medicine and Hygiene, May 2012

specific protection may not be effective in a parasite population in which the vaccine variant is rare. Furthermore, vaccines may create a selective advantage for non-vaccine variant types, compromising vaccine efficacy.[1]

The *cs* gene is polymorphic, with diversity in regions that code for epitopes recognized by the human immune system. The central repeat region of the *cs* gene contains tetrameric repeats that vary in both the sequence and number of tetramers.[36] This region codes for epitopes recognized by anti-CSP antibodies.[38, 39] The 3' regions of the *cs* gene, Th2R and Th3R, encode epitopes that are recognized by CD8+ and CD4+ T-cells.[43] The diversity in these regions, which occurs in the form of non-synonymous SNPs, increases as malaria transmission increases across distinct geographic areas,[40, 41]with the highest diversity occurring in Africa. Molecular surveys in Sierra Leone and the Gambia found 42 haplotypes in 99 samples and 24 haplotypes in 44 samples for the region containing Th2R and Th3R respectively.[68,41] The current leading malaria vaccine candidate, RTS,S/AS01, which is based on the immunogenic regions of CSP, has shown modest efficacy in Phase 2 trials[49,50-52] and is currently being evaluated in a multicenter Phase 3 trial in 11 countries in Africa.[69]

A follow-up study to a Phase 2 trial of the vaccine concluded that there was no selection of non-vaccine variants in vaccinated children vs. non-vaccinated children.[58] However, this study, which used direct sequencing to detect polymorphism in the regions coding for the T-cell epitopes, Th2R and Th3R, excluded samples that could not be resolved into predominant alleles from the analysis. Furthermore, diversity in the central repeat region of the *cs* gene which codes for the B-cell epitopes of CSP and which is also

included in the vaccine, was not considered, presumably owing to the limitations in direct sequencing technology.

Direct sequencing is limited in its ability to detect multiple parasite types in a mixed infection because this method depends on reading major and minor peaks on a chromatogram to determine allele presence or absence in a sample. The proportion of these peaks may not correlate well with the actual proportion of parasite DNA in a sample, as incorporation of dye-labeled dideoxynucleotide can vary within a sample and be influenced by flanking sequence.[70] Moreover, complete haplotypes for each unique parasite clone that is in a sample cannot be determined by direct sequencing. It is also impossible to resolve diversity with respect to repetitive DNA sequences in mixed infections using this sequencing method.

New, more powerful sequencing technologies may have the potential to address some of these limitations. 454, a next-generation sequencing platform, generates massively parallel DNA sequences from PCR products, potentially making it possible to resolve diversity in complex infections. With longer read lengths than other next-generation sequencing platforms, 454 might permit sequencing the variable-length central repeat region of *cs* that has defied other sequencing methods on field samples. Furthermore, by providing massively parallel sequence of the target region that permits quantification of reads with different variants, this technology may help determine which alleles are predominant at polymorphic sites more reliably than direct sequencing.

C. Materials and Methods

Standardized mixed infections

Mixtures of PCR product containing Th2R and Th3R amplified from laboratory variants (3D7, Hb3, and Dd2), for which the sequences are known, were created, quantified and diluted to concentrations of 100 ng/µl, 50 ng/µl, 25 ng/µl, 12.5 ng/µl, and 6.25 ng/µl. 3D7 comprised 60% of each mixture, Hb3 comprised 30%, and Dd2 comprised 10%. The PCR products for each variant were generated in triplicate, and three mixtures were made and serially diluted in parallel. Each of the mixtures was sequenced by both 454 and Sanger sequencing to test the ability of each technology to quantitate the different alleles in a mixture. The sequence output for each dilution was combined for both 454 and Sanger sequencing. The observed allele frequencies for the 454 sequencing method were determined by calculating the percentage of reads that contained each type of allele at each of seven polymorphic sites for each concentration from the three parallel dilutions. The observed allele frequencies for Sanger sequencing were determined by calculating the relative peak heights of the major and minor allele at each polymorphic site at each concentration from the three parallel dilutions. These frequencies were subtracted from the expected frequency for each allele, and the sums of the absolute value of these differences were averaged for each concentration.

A sensitivity analysis was performed on sequence output generated from the standardized mixed infections from each technology to determine a threshold for inclusion of minor alleles from the clinical samples. Based on the curves generated (Figure IV.1), the largest numbers of erroneous SNPs were found between the frequencies of 0.025 and 0.01 for 454, and between 0.2 and .15 for Sanger sequencing.

However due to the fact that erroneous SNPs were still present at a frequency of 0.05, a conservative threshold of 0.1 was selected for 454. The threshold of 0.2 was selected for Sanger sequencing.

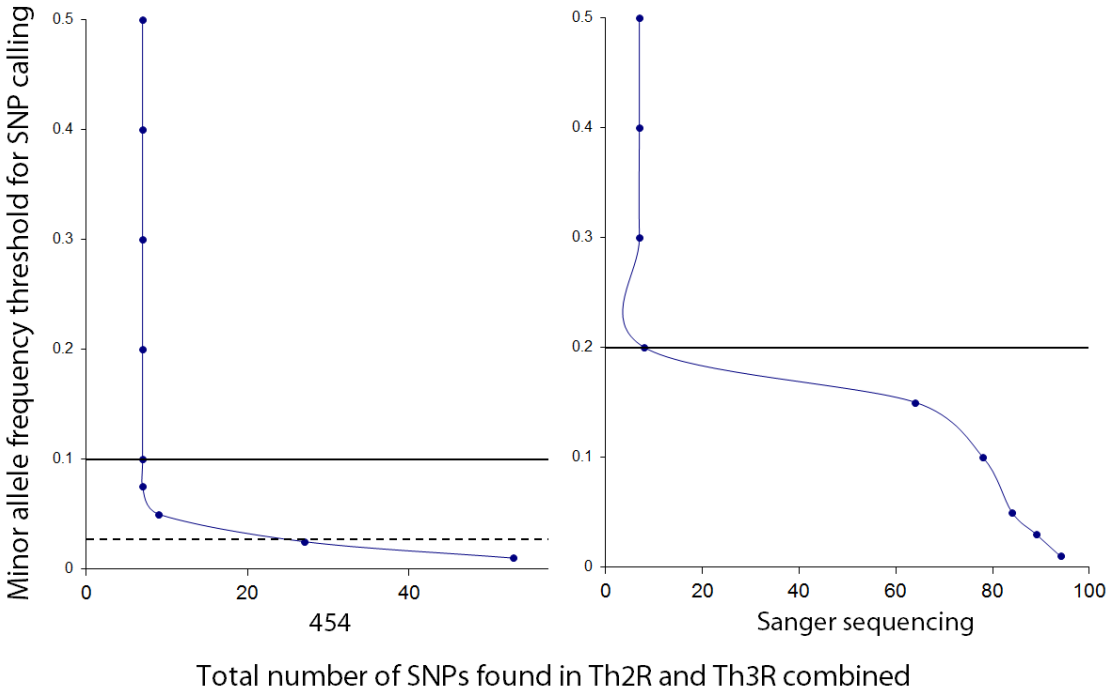

Figure IV.1 Determination of minor allele frequency threshold for 454 and Sanger sequencing

Clinical sample selection

DNA extracted from forty-five parasite positive filter paper blood samples was used to compare the ability of 454 and Sanger sequencing to detect CSP diversity in field samples. Samples were randomly selected from among participants in an incidence study conducted in Bandiagara, Mali from the years 1999 to 2001[62] and represent both clinical and asymptomatic infections detected through passive and active surveillance.

Sanger sequencing

Two nested PCR assays were designed to amplify a region of the *cs* gene containing both Th2R and Th3R, and the central repeat region. The primary PCR was

designed to amplify the region of the gene which contains both regions, and the secondary PCRs amplify Th2R and Th3R, as well as the central repeat region individually. The primary forward and reverse PCR primers were GTTGAGGCCTTTTCCAGGAATACCAG and GTACAACTCAAACTAAGATGTGTTC. Primary PCR conditions are as follows: 30 cycles of 95 °C for 30s, 52 °C for 30s, 72 °C for 1 min. Secondary PCR conditions for the repeat regions and ThRs were 30 cycles of 95 °C for 30s, 55 °C for 30s, 72 °C for 1min 30s, and 25 cycles of 95 °C for 30s, 58 °C for 30s, and 72 °C for 1min respectively. Secondary primers were the same as the sequence specific primers shown in Figure IV.2. Expected product sizes for the Th region and repeat region were 214 and 516 base pairs respectively, based on the 3D7 variant of *P. falciparum.* PCR products were amplified using HotStar Taq (Qiagen, Valencia, California). PCR products were loaded on a 2% agarose gel, stained with ethidium bromide, and run at 100 Volts for 1 hour. Bands were detected using geneSNAP (Synoptics LTD, Cambridge, UK) gel imaging software.

Once amplification was verified by gel electrophoresis, PCR Products were purified by vacuum filtration in Excela Pure (Edge Biosystems, Gaithersburg, MD) 96-well plates. Purified PCR product was then amplified and sequenced on an ABI3730 xl at the University of Maryland School of Medicine Biopolymer Lab.

Sequences were aligned to the 3D7 reference genome using Sequencher (Gene Codes Corp, Ann Arbor, MI) software. For samples containing more than one allele at a polymorphic site, a predominant allele was designated if the secondary peak height was less than or equal to 40% of the height of the primary peak on the chromatogram for that sample. If the secondary peak was greater than 40% of the height of the primary peak the

polymorphic site and sample were designated as mixed and haplotypes were not constructed for these samples. Minor alleles that were not represented by a peak that was at least 20% of the primary peak height were not included in the total number of SNPs discovered in Sanger sequencing output.

454 sequencing

The primary PCR used for 454 sequencing was identical to that used for Sanger sequencing. Secondary PCR primers contained specific adapters necessary for the emPCR[63] step of 454 sequencing, as well as unique barcodes to identify sequences from individual samples (Figure IV.2).

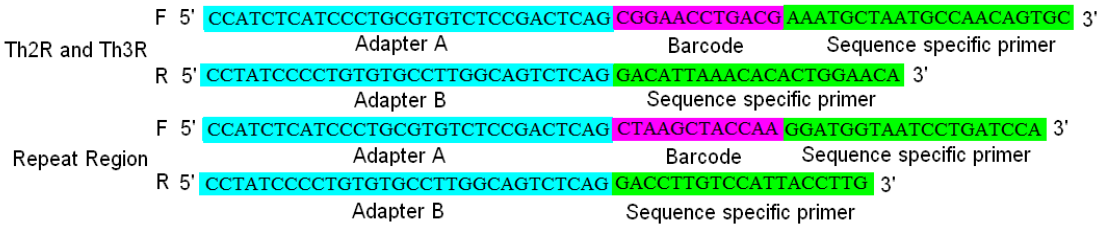

Figure IV.2 Primers used for amplification of PCR products for 454 sequencing

The concentration of each PCR product was determined by band intensity compared to a standard of similar molecular weight using geneSNAP software, and 100 ng of each product was then pooled. PCR products for each region were pooled separately. Pooled PCR products were sequenced at the University of Maryland School of Medicine Genomic Resource Center on the GS FLX Titanium 454 Platform (Roche Diagnostics, Branford, CT). Sequences were aligned using gsAmplicon (Roche Diagnostics, Branford, CT) software. For samples containing more than one allele at a polymorphic site, predominance was determined if the majority allele was present in 60%

or more of all reads obtained for that sample. If a majority allele could not be determined, that polymorphic site was considered mixed. Haplotype information, however, was still obtained for samples with mixed polymorphic sites.

D. Results

Detection of allele frequencies in standardized mixed infections

The average difference between observed and expected allele frequencies for 454 was less than 0.1 for each concentration. The highest difference was 9% which occurred at a concentration of 100 ng/µl, and decreased with decreasing concentration leveling off at 12.5 ng/µl. The average difference between observed and expected allele frequencies for Sanger sequencing also decreased from high to low concentrations with the highest, 0.38, occurring at 100 ng/µl, and the lowest 0.14 occurring at 6.25 ng/µl. Overall the difference between observed and expected allele frequencies was lower at each concentration for 454 than for Sanger sequencing (Figure IV.3). Statistical significance was calculated using at student's t-test to compare the average difference between observed and expected allele frequencies for both technologies at each concentration and is denoted by an asterisk. The percentage of predominant alleles that were correctly identified was 91% and 75% for 454 and Sanger sequencing respectively.

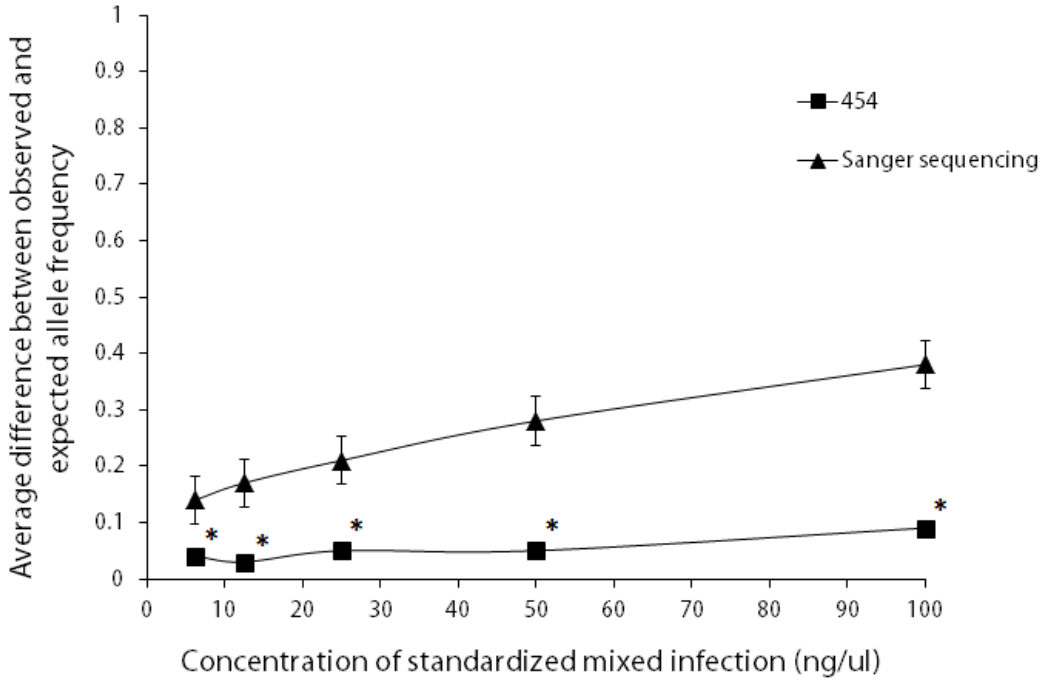

Figure IV.3 Accuracy of allele quantification in standardized mixed infections by 454 and Sanger sequencing.

SNP detection

A total of 17 and 9 SNPs were detected by Sanger sequencing (2x coverage, forward and reverse) in Th2R and Th3R respectively from the 45 samples selected for sequencing. A total of and 24 and 14 SNPs were detected by 454 in Th2R and Th3R respectively (Figure IV.4). The average coverage of the Th regions in 454 sequence output was ~500x, with a range of ~200x to ~1000x.

Haplotype detection

The total number of haplotypes found in Th2R and Th3R respectively was 24 and 10 in Sanger sequencing output, and 72 and 14 in 454 output (Figure IV.4). Only

haplotypes representing at least 10% of all 454 reads obtained for a sample were included. Samples which contained polymorphic sites with more than one allele in Sanger sequencing output could not be resolved into haplotypes and were therefore excluded from haplotype analyses. The proportion of unique haplotypes to the total number of haplotypes detected was 0.53 (24/45) for Sanger sequencing and 0.49 (72/147) for 454.

Mixed infections

Most samples (39 out of 45) contained more than one distinct parasite type based on 454 data, whereas only 20 samples had more than one haplotype detected by Sanger sequencing with respect to Th2R, the Th region with the most diversity (Figure IV.4).

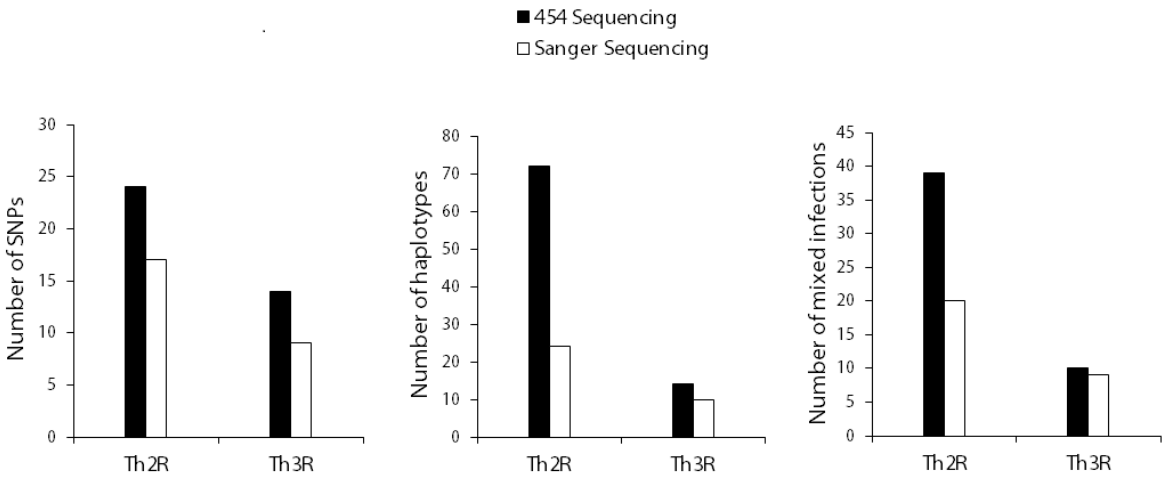

Figure IV.4 Number of SNPs, haplotypes, and mixed infections detected in Th2R and Th3R by 454 and Sanger sequencing

Determination of majority alleles

Of the SNPs identified as majority alleles by either 454 or Sanger sequencing, approximately 74% were identified as majority alleles by both technologies, 24% were identified as a majority allele by one technology and not the other, but were detected by

both, and 2% were identified as the majority allele by 454 but not detected by Sanger sequencing, with respect to Th2R. In the case of Th3R, approximately 77% were identified as majority alleles by both technologies, 18% were identified as a majority allele by one technology and not the other, but were detected by both, and 4% were identified as the majority allele by 454 but not detected by Sanger sequencing (Figure IV.5). There were no samples in which an allele was identified as predominant in Sanger sequencing output and not detected by 454.

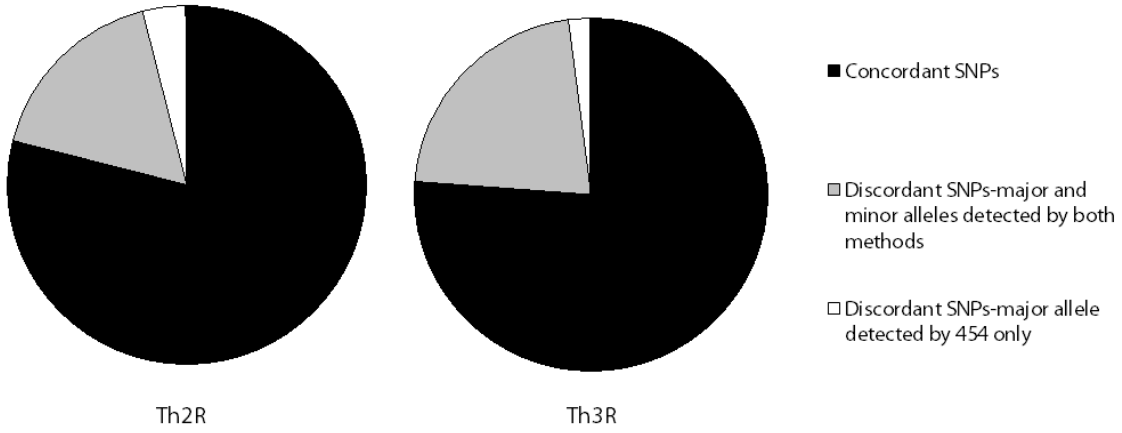

Figure IV.5 Concordance between direct sequencing and 454 in determination of majority alleles in the Th2R and Th3R regions of the circumsporozoite (*cs*) gene.

Haplotype diversity

The number of distinct haplotypes found within each sample with respect to Th2R and Th3R was explored in 454 data. An average of 3.5 parasite types were found with

respect to Th2R (range 1 to 8) and an average of 2.5 parasite types were found with respect to Th3R (range 1 to 4) (Table IV.1)

	Th2R	Th3R
Mean	3.5	1.6
Median	4.5	2.5
Range	1 to 8	1 to 4

Table IV.1 Haplotype diversity with respect to Th2R and Th3R as measured by 454 sequencing

Sequencing the repeat region

Complete sequences for the region of the cs gene coding for B-cell epitopes were not obtained, and therefore diversity and haplotype data could not be generated for this region.

E. Discussion

454 was the more sensitive means of assessing diversity in *cs*, detecting approximately 41% more SNPs in Th2R and 64% more SNPs in Th3R than Sanger sequencing. 454 was also more accurate in identifying the diversity in a known mixture of parasite clones. A previous study has shown that adjustments can be made to peak height in chromatograms to diminish inaccuracies caused by dye-effects in Sanger sequencing when sequencing genomic DNA.[71] However, it is unclear how well these methods would work for sequencing field samples, particularly filter paper blood samples, which tend to have DNA of poorer quality and variable quantity.

A relatively high minor allele frequency was used for defining SNPs (at least 10% of all reads) to strengthen confidence that the SNPs included in the analyses were genuine and not products of PCR or 454 sequencing error. The 10% limit is very conservative, being well above both the lower minor allele frequency thresholds typically used for SNP discovery and the error rates reported by the manufacturers: 0.4% error for HotStar Taq polymerase, and for GS FLX Titanium a 1% error rate for read lengths of 400bp and better for smaller read lengths and it was twice the frequency at which erroneous SNPs were discovered in the sensitivity analysis performed on the standardized mixed infections. Additionally, one base pair indels were excluded from the analysis, since one of the limitations of pyrosequencing is that it has a higher error rate when sequencing homopolymers (stretches of a single base such as, AAAA). This exclusion was done to help ensure erroneous SNPs were excluded from the analysis however it does raise the possibility that some real SNPs were also excluded.

Since this study was not focused on detecting rare variants the conservative threshold for SNP calling would not have affected the conclusions drawn. For 454 applications where detecting infrequent alleles is important, a lower minor allele frequency threshold could likely be identified using more rigorous SNP-calling algorithms. The proportion of unique haplotypes to total haplotypes detected was reassuringly similar for Sanger sequencing (0.53) and 454 (0.49); however, the data from this study showed that using 454 more than tripled the number of evaluable haplotypes that were generated from a sample set. In 45 samples 72 haplotypes were found in Th2R alone, whereas in earlier studies using Sanger sequencing on samples from other West African settings with similar malaria epidemiology, only 24 haplotypes were found in 44

Gambian samples[41], and just 42 haplotypes were found[68] for Th2R and Th3R combined in 99 samples from Sierra Leone.

The high number of SNPs and unique haplotypes in our study is consistent with this region of the *cs* gene being under diversifying selection pressure.[72] Although no evidence has been reported of selection of non-vaccine variants of CSP following immunization with CSP-based vaccines, this evidence of diversifying selection supports the notion that genetic variation in CSP may be driven by the human immune system, and could be important in naturally acquired and vaccine-induced immunity.

The results of the sensitivity analysis revealed different rates of error for the highest concentration of DNA for both technologies. A possible explanation for this finding may be that the high concentration of DNA results in signal interference. In output from Sanger sequencing, base ambiguities may result from overlap between peaks. In 454 output, light signals that occur when bases are incorporated can bleed into signals from surrounding reactions, and this may result in errors.

There were a few alleles that were determined to be majority alleles by 454 but were not detected by Sanger sequencing. Two possible explanations for this finding are either 1) the majority alleles were actually not majority alleles and were the result of PCR bias or 2) these alleles were on the lower end of the cut-off for a majority allele by 454 and they were detected as a minority allele in Sanger sequencing but excluded due to the minority allele cut-off for this method. Given the variability found in the results from the sensitivity analysis performed on lab variants for direct sequencing, the second explanation is more likely.

Despite several attempts and extensive troubleshooting, complete sequence data could not be generated for the repeat region, which appears not to be amenable to sequencing in filter paper samples using current 454 technology. Longer read lengths are required to get full coverage across this ~450 to 550 base pair region. Although read lengths in this range are possible on the GS FLX Titanium 454 Platform, they are at the upper limit of what is routinely obtained. In addition, a known limitation of pyrosequencing is difficulty in reliably generating data on long repetitive DNA sequences due to nucleotide exhaustion resulting in premature termination of read synthesis (Luke Tallon, IGS, personal communication, March 10, 2011). It was initially thought that since the *cs* gene repeats with 12 nucleotides are longer than many short tandem repeats, 454 could still be a viable option for this region. However, despite the longer length of the repeat, premature termination still occurred. Because diversity in the *cs* repeat region may be an important driver of allele-specific natural and vaccine-induced immunity, technology development efforts that will enable sequencing this region are warranted.

The results of these initial studies demonstrate that there is more extensive polymorphism in the regions of the *cs* gene coding for T-cell epitopes than has been previously described in this geographic region. Further examination of polymorphism in CSP in vaccine trials and epidemiologicalal studies may elucidate the contribution of CSP immunity to clinical protection against malaria and inform the development of improved CSP-based vaccines. As read lengths continue to improve and costs decline, 454 and other next-generation and third generation sequencing platforms may be better suited to handle the long repeats of the *cs* gene, so that important diversity in this important region can be examined. Until the entire *cs* gene including the repeat regions

can be fully sequenced in a high throughput fashion, the role, if any, of allelic diversity in limiting the efficacy of RTS,S and other CSP-based vaccines will remain uncertain.

Acknowledgements

We would like to thank Jacques Ravel of the Institute for Genome Sciences at the University of Maryland, Baltimore, for providing barcodes used in this project.

This research was supported by contract N01AI85346 and cooperative agreement U19AI065683 from the National Institute of Allergy and Infectious Diseases (NIAID), grant D43TW001589 from the Fogarty International Center, National Institutes of Health, and contract W81XWH-06-1-0427 from the U.S. Department of Defense and the U.S. Agency for International Development. CVP is supported by a Distinguished Clinical Scientist Award from the Doris Duke Charitable Foundation and by the Howard Hughes Medical Institute. STH is supported by the University of Maryland Multidisciplinary Clinical Research Career Development Program (NIH grant K12RR023250).

Authors' addresses: Kavita Gandhi, Shannon Takala-Harrison, Christopher V. Plowe, Center for Vaccine Development, University of Maryland, 685 West Baltimore Street, Room 480, Baltimore, USA. Telephone: 001-410-706-2491. Fax: 001-410-706-1204. Email: Christopher.plowe@medicine.umaryland.edu. Mahamadou A. Thera, Drissa Coulibaly, Karim Traoré, Ando B. Guindo, Ogobara K. Doumbo, Malaria Research and Training Center, University of Bamako, Bamako, PO Box 1805, Point G, Bamako, Mali. Telephone: Fax: E-mail: okd@mrtcbko.org.

V. VARIATION IN THE CIRCUMSPOROZOITE PROTEIN OF PLASMODIUM FALCIPARUM: IMPLICATIONS FOR VACCINE DEVELOPMENT

A. Abstract

Background

A leading malaria vaccine candidate, RTS,S/AS01, is based on immunogenic regions of *Plasmodium falciparum* circumsporozoite protein (CSP) from the 3D7 variant, and has shown modest efficacy against clinical disease in African children. It is unclear, however, what aspect of the immune response elicited by this vaccine is protective. Better understanding of how diversity in the immunogenic regions of CSP (T-cell and B-cell epitopes) may relate to clinical immunity is needed to evaluate and improve the efficacy of vaccines based on CSP. The goal of this study is to measure diversity in these immunogenic regions and identify associations between variation in amino acid sequences in CSP and the risk of infection and clinical disease caused by *P. falciparum*.

Methods

The present study includes 100 children from a prospective cohort study designed to measure incidence of malaria infection in children in Bandiagara, Mali. From these 100 children 769 parasite-positive blood samples corresponding to both acute clinical malaria episodes and asymptomatic infections detected in monthly surveys were examined in this study. Non-synonymous SNP data was generated via 454, a next generation sequencing technology, for the T-cell epitopes and repeat length data was generated for the B-cell epitopes of the *cs* gene. Cox proportional hazards models were

used to determine the effect of sequence variation in consecutive infections occurring within individuals on the time to new infection and new clinical malaria episode.

Results

Diversity in Th2R and Th3R remained stable throughout seasons, between age groups and between clinical and asymptomatic infections with the exception of a higher proportion of 3D7 haplotypes found in the oldest age group. No associations between sequence variation and hazard of infection or clinical malaria were detected.

Interpretation

The lack of association between sequence variation and hazard of infection or clinical malaria suggests that naturally acquired immunity to CSP may not be allele-specific.

B. Introduction

As part of the effort to reduce the global malaria burden substantial resources are being invested in the development of vaccines targeting specific immunogenic antigens of *P. falciparum*, including the circumsporozoite protein (CSP), encoded by the *cs* gene. To date, malaria vaccine development and testing has generally not been informed by molecular epidemiological evaluations of how genetic diversity in vaccine antigens in parasite populations may affect vaccine efficacy. For example, vaccines that confer allele-specific protection may not be effective in a parasite population in which the vaccine allele is rare, and may create a selective advantage favoring non-vaccine alleles, compromising vaccine efficacy.[1] Furthermore, studies of naturally acquired immunity to two *P. falciparum* blood stage antigens found that immune responses to these antigens may be allele-specific, and a field trial of a vaccine based on one of these antigens reported allele-specific efficacy.[30, 73, 74]

The *cs* gene is polymorphic, with diversity in regions that code for epitopes recognized by the human immune system. The central repeat region of the *cs* gene contains tetrameric repeats that vary in both the sequence and number of tetramers. This region codes for epitopes recognized by anti-CSP antibodies.[38, 39] It has been suggested that the length of the repeat region may play a role in the stability of the protein, and may therefore affect how B-cell epitopes are displayed to the immune system. [36] The 3' regions of the *cs* gene, Th2R and Th3R, encode epitopes that are recognized by CD8+ and CD4+ T-cells.[43] The diversity in these regions, which occurs in the form of non-synonymous SNPs, increases as malaria transmission increases across distinct geographic areas,[40, 41] with the highest diversity occurring in Africa. Molecular surveys in Sierra Leone and the Gambia found 42 haplotypes in 99 samples and 24 haplotypes in 44 samples for the region containing Th2R and Th3R respectively.[68,41] The current leading malaria vaccine candidate, RTS,S/AS01, which is based on the immunogenic regions of CSP, has shown modest efficacy in Phase 2 trials[49,50-52] and Phase 3 trials.[56, 57]

Follow-up studies to Phase 2 trials of the vaccine report conflicting evidence of selection of non-vaccine variants in vaccinated vs. non-vaccinated study participants. The first study was a follow-up to a Phase 2 trial of RTS,S in children in Mozambique, and reported no evidence of selection of non-vaccine variants in vaccinated children.[58] The second study was a follow-up to a Phase 2 trial in adults in Kenya.[59] This study reported two polymorphic sites at which a statistically different proportion of non-vaccine variant alleles were found between control and vaccine groups. At one site a lower proportion of non-vaccine variant alleles was found in the vaccinated group, and at the other site a higher proportion was found in the vaccinated group. Since the effects were in opposite

directions the authors concluded that this evidence was not strong enough to suggest vaccine selection.

Both of these studies, however, had two main issues. First, both studies used direct sequencing to detect polymorphism in the regions coding for the T-cell epitopes, Th2R and Th3R, excluded samples that could not be resolved into predominant alleles from the analysis. Second, diversity in the central repeat region of the *cs* gene which codes for the B-cell epitopes of CSP and which is also included in the vaccine, was not considered.

To design more effective malaria vaccines and to help interpret efficacy data from vaccine trials, an understanding of the dynamics of polymorphism in a vaccine antigen, and the factors that are driving that polymorphism are necessary. With the aid of next generation sequencing to help circumvent some of the issues encountered by previous studies of genetic diversity in CSP, this study has examined both the population level and within-host dynamics of polymorphism in this vaccine antigen.

C. Materials and Methods

Parent study description

The study was conducted in Bandiagara, a rural town of approximately 13,000 inhabitants located in central eastern Mali. Transmission of *falciparum* malaria is intense with the peak coinciding with the rainy season from July-October. Children aged less than 10 years experience, on average, two clinical episodes a year, and the prevalence of parasitemia at the beginning of the study before the onset of malaria transmission was 17%.[62]

A complete population census was conducted in Bandiagara before study initiation. Study participants were sampled in proportion to the population size in each of the eight districts comprising the town of Bandiagara. Study subjects were aged ≥ 3 months to 20 years. Age groups were defined as <2, 2-4, 5-7, 8-10, 11-14, and 15-20 years of age. Recruitment of study subjects took place in randomly selected households until the target number of subjects in each age group was achieved.

The study was conducted prospectively during the years 1999, 2000, and 2001. From July to January of each year, blood samples were collected on 3MM Whatman filter paper monthly and at every episode of clinical malaria. Clinical malaria episodes were detected both by passive surveillance through provision of around-the-clock free clinical care, as well as by active weekly follow-up of the children in the study by physicians working at the study site. Clinical episodes were defined as a blood smears positive for *P. falciparum* asexual parasites and symptoms consistent with malaria including, fever, anemia headache, body aches, cough, diarrhea, or abdominal pain. Infections were defined as the presence of *falciparum* parasites in the blood, with or without syptoms.[62]

Present study description

One hundred children with at least two years of follow up during the malaria incidence study were chosen.[30] These children were randomly selected within three age strata. Thirty children aged ≤5 years, 32 children aged 6 to 10 years, and 38 children aged ≥11 years were selected. Blood samples (n=2309) corresponding to all monthly surveys (n=1801) and clinical episodes (n=508) occurring during the transmission season of the three years of the incidence study underwent DNA extraction (QIAamp DNA Mini Kit,

Qiagen, Valencia, California) as described previously[30, 73]. Of these, 769 parasite positive samples were subjected to sequencing analysis.

PCR amplification

The primary forward and reverse PCR primers were GTTGAGGCCTTTTCCAGGAATACCAG and GTACAACTCAAACTAAGATGTGTTC. These primers were designed to amplify the region of the *cs* gene that contained both the repeat region and the Th regions. Primary PCR conditions are as follows: 30 cycles of 95 °C for 30s, 52 °C for 30s, 72 °C for 1 min. The secondary forward and reverse PCR primers for the repeat region were TGGGAAACAGGAAAATTGG and GCACTGTTGGCATTAGCATTT. Secondary PCR conditions for the repeat regions and ThRs were 30 cycles of 95 °C for 30s, 55 °C for 30s, 72 °C for 1min 30s, and 25 cycles of 95 °C for 30s, 58 °C for 30s, and 72 °C for 1 min respectively. PCR products were amplified using HotStar Taq (Qiagen, Valencia, California). Secondary PCR primers for the ThRs contained specific adapters necessary for the emPCR[63] step of 454 sequencing, as well as unique barcodes to identify sequences from individual samples (Figure V.1). A total of 96 primers containing 96 unique barcode sequences were used to amplify this region from study samples. Primers were identical with the exception of the barcode sequence.

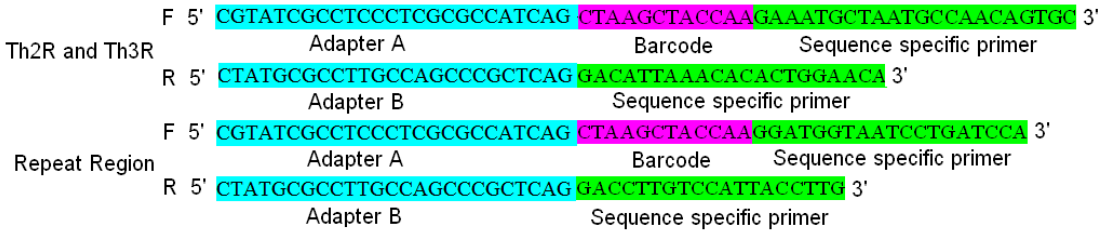

Figure V.1 454 primers used for amplification of the Th region.

454 sequencing

The concentration of each PCR product containing Th2R and Th3R was determined by band intensity measurements taken on the Qiaxcel capillary gel imaging system (Qiagen, Valencia, California), and 100 ng of each product was then pooled. Identical barcodes were used to tag more than one sample. PCR products were separated into sixteen separate pools containing approximately 45 samples each ensuring that no pool contained more than one sample with the same barcode. Each pool was physically separated within a 454 sequencing run. Pooled PCR products were sequenced at the University of Maryland School of Medicine Genomic Resource Center at the Institute of Genome Sciences on the GS FLX Titanium 454 Platform (Roche Diagnostics, Branford, CT). Sequences were aligned using gsAmplicon (Roche Diagnostics, Branford, CT) software. For samples containing more than one allele at a polymorphic site, predominance was determined if the majority allele was present in 71% or more of all reads obtained for that sample. This cut-off was determined in the methods validation study comparing 454 and Sanger sequencing described below. If a majority allele could not be determined, that polymorphic site was considered polyclonal. Haplotype information, however, was still obtained for samples with polyclonal polymorphic sites, because complete sequence reads were available for each variant detected. A methods

validation study was performed by the same authors to establish the comparability of these methods with direct sequencing as well as comparable majority/minority allele cut-off values. The results indicated that 454 was more sensitive at detecting minor alleles and more accurate in the quantitation of these alleles than direct sequencing.[75]

Determination of repeat region length

To determine the length of the repeat region, PCR products were run on a high resolution gel cartridge on a Qiaxcel capillary gel imaging system (Qiagen, Valencia, California) using the OM500 analysis method capable of resolving size within 10 base pairs (bp). To validate this method, a 96-well plate containing 15ul of repeat region PCR product amplified from 100 ng/ul 3D7 genomic DNA per well, was run using the method listed above. A systematic underestimation of 10bp was observed in the experimental data. A 10bp correction was therefore applied to length values obtained for PCR products from study samples.

Repeat region PCR products for samples that were determined to be 'single clone' (no secondary allele present in a frequency greater than 20% in 454 reads) with respect to Th2R and Th3R, were subjected to Sanger sequencing. Once amplification was verified by gel electrophoresis, PCR Products were purified by vacuum filtration in Excela Pure (Edge Biosystems, Gaithersburg, MD) 96-well plates. Purified PCR product was then sequenced on an ABI3730 xl at the University of Maryland School of Medicine Biopolymer Lab.

Sequences were viewed in Sequencher (Gene Codes Corp, Ann Arbor, MI) software, and then copied to Transeq web software[76] to determine the amino acid

sequence and length of the repeat region. Size determinations made via the Qiaxcel for these ‘single clone’ samples were checked against corresponding Sanger sequencing data.

Determination of mixed infections

Mixed infections with respect to the ThRs were defined as samples with no clear majority allele at a given polymorphic site. Mixed infections with respect to the repeat regions were defined as samples which contained more than one clear band in high-resolution gel analysis.

Statistical methods

Fisher’s exact tests were used to compare haplotype frequencies between seasons, age groups, and between clinical and non-clinical episodes. Cox proportional hazards models were used to examine the relationship between diversity in Th2R and Th3R and the development of naturally acquired immunity in the context of symptomatic (clinical) and asymptomatic parasitemia (infection). In the clinical model, the association between changes at polymorphic sites in Th2R and Th3R which occurred between consecutive clinical episodes and the hazard of clinical disease was examined. In the infection model, the association between these changes which occurred between a clinical episode and a consecutive asymptomatic infection was examined. Intervals starting with an asymptomatic episode were included in the model if there was an intermediate time point at which the study participant was parasite negative verified by microscopy. To account for the possibility of treatment failure and to allow for time for allele-specific antibodies to the first of two paired consecutive infections to be present by the time of the second infection, time intervals two weeks or less between consecutive episodes were excluded from the analyses. A logistic regression analysis was performed to determine the odds of

a change in the predominant amino acid at polymorphic residues in intervals including an asymptomatic episode followed by a symptomatic one, to intervals including consecutive asymptomatic episodes. Age and time between consecutive intervals were included as covariates that in this model.

D. Results

Of the 769 parasite-positive samples, 63 were missing or did not have enough material for successful PCR amplification. With respect to the Th regions, DNA was successfully amplified and sequenced from 684 (97%) of the samples. With respect to repeat region PCR products and fragment length data were successfully generated for 504 (74%) of the samples. Of these 504 samples, 157 were successfully sequenced via Sanger sequencing.

Diversity in the Th regions

A total of 31 and 14 nonsynonymous SNPs were detected by 454 sequencing with an average coverage of ~600x (range of 100x-1,300x) in Th2R and Th3R, respectively (Table V.1). There were 17 polymorphic amino acid positions in total between the Th regions.

The number of unique haplotypes found in Th2R and Th3R respectively was 87 and 23 (Table V.1). Of the haplotypes detected in Th2R, 51.7% occurred only once in the three year study period. Of those detected in Th3R, 37.5% occurred once in the three year study period. Only haplotypes representing at least 10% of all 454 reads obtained for a sample were considered for analysis.

For the purposes of this study, polyclonal infections were defined as samples that contained no clear majority allele (no allele that was present in 71% or more of 454

reads). With respect to Th2R, 34.9% of samples were mixed, and with respect to Th3R, 29.8% of samples were mixed.

	Number of haplotypes	Polyclonal infections
Th2R	87	34.9%
Th3R	23	29.8%
Repeat region	20	16.9%

Table V.1 Haplotypes and polyclonal infections detected in Th2R, Th3R and the repeat region.

Haplotype distributions did not vary significantly by age group, study year, or presence of clinical symptom for either Th2R or Th3R, with the exception of the Th2R 3D7 haplotype, which had a significantly greater prevalence in the oldest age group compared to the youngest age group ($p<0.03$, Figure V.2 and V.3).

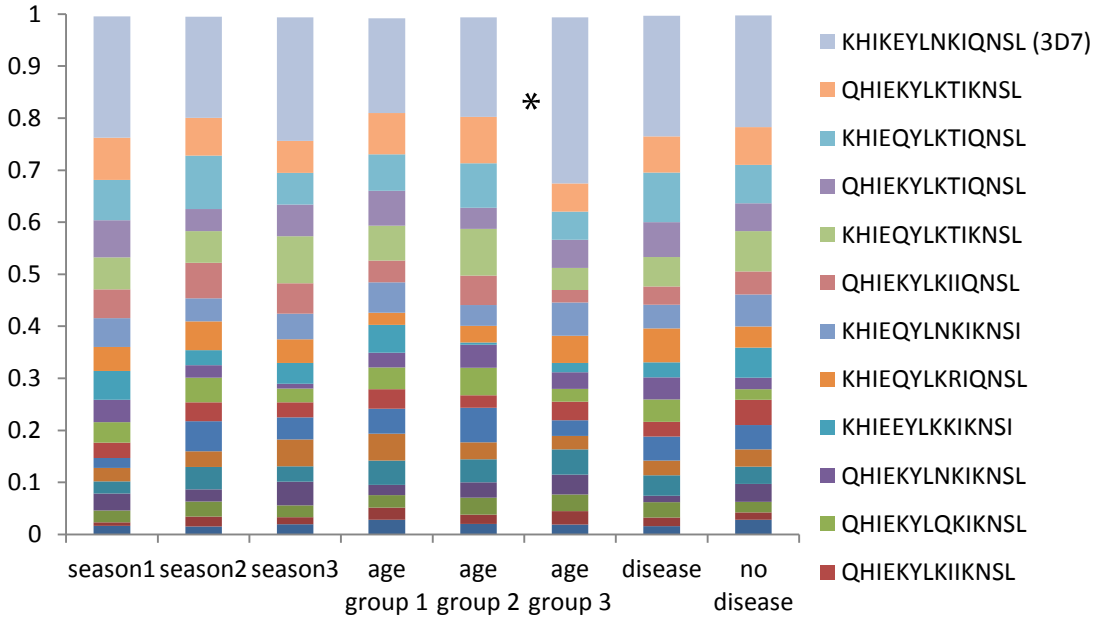

Figure V.2 Distribution of Th2R haplotypes across seasons, age groups, and clinical and non-clinical *Plasmodium falciparum*.

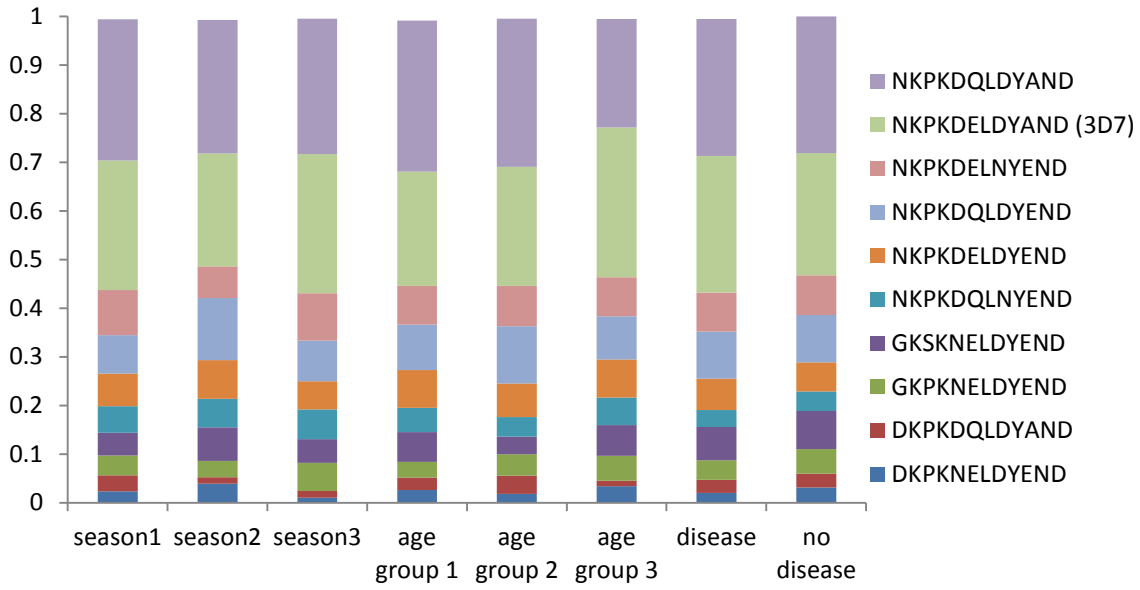

Figure V.3 Distribution of Th3R haplotypes across seasons, age groups, and clinical and non-clinical *Plasmodium falciparum*.

Diversity in the repeat region

The prevalence of different repeat region sizes did not vary significantly by age group, study year, or presence of clinical symptoms (Figure V.4). The most common tetrameric repeat number found in this study was 40 (34.9%), (range of 38 to 43), with the two highest numbers of repeats, 42 and 43, appearing least frequently (3.4% and 0.68%).

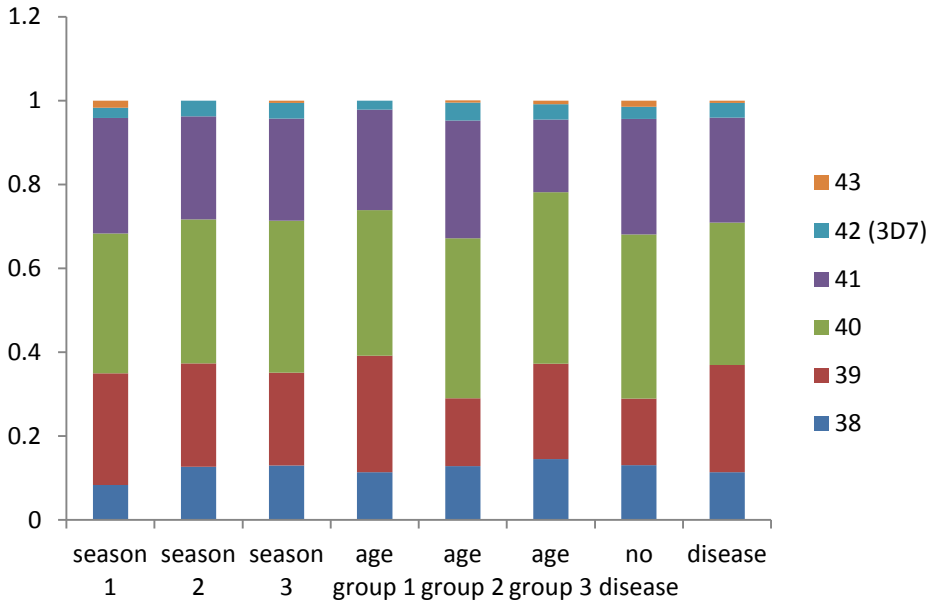

Figure V.4 Distribution of repeat region size polymorphisms across season, age groups and clinical and non-clinical *Plasmodium falciparum* infections

Among the 157 samples for which Sanger sequencing data were available, 20 unique haplotypes were detected, of which 6 were detected only once in the three year study period. The majority of samples (67.5%) had one of three most prevalent haplotypes (Figure V.5). When Sanger sequencing results were compared to adjusted size data generated on the Qiaxcel, there was 94% agreement between the two methods.

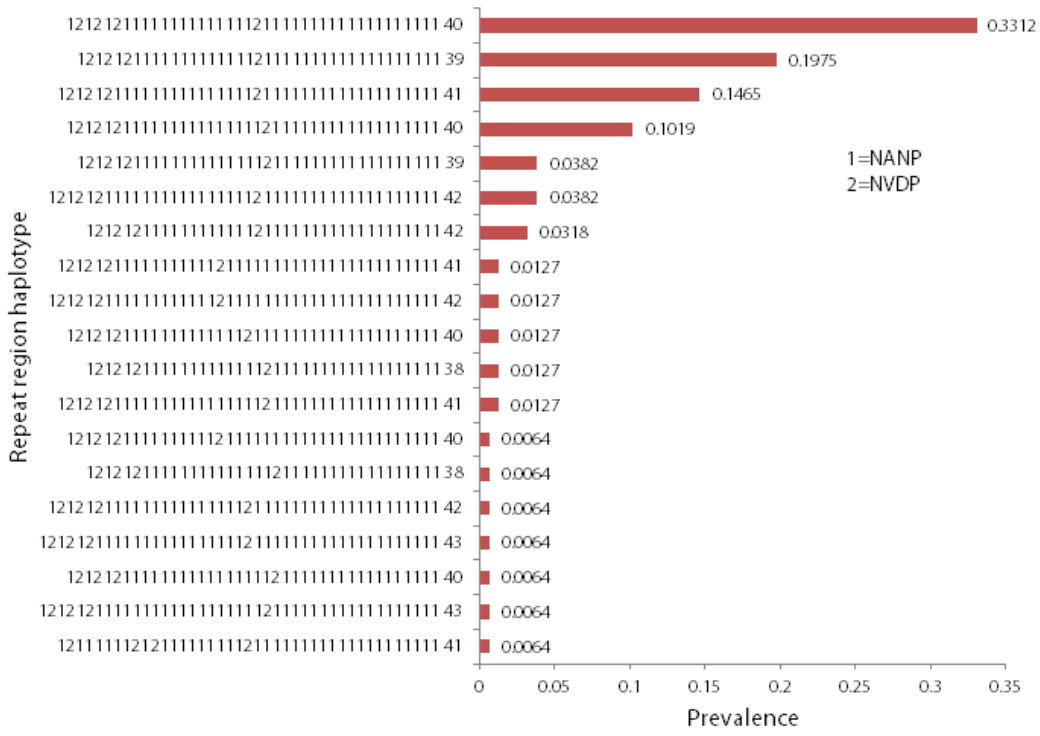

Figure V.5 Repeat region haplotype prevalences

Effect of variation on the risk of infection and disease

There was no significant change in hazard of infection or clinical disease when a change occurred at any of the 17 polymorphic amino acid positions identified in Th2R and Th3R, or when a change occurred in the size of the repeat region (Figures V.6 and V.7). The number of changes between consecutive episodes was also not significantly associated with hazard of infection or disease. Asterisks denote polymorphic amino acid positions that lacked power to detect a hazard ratio below 1.51.

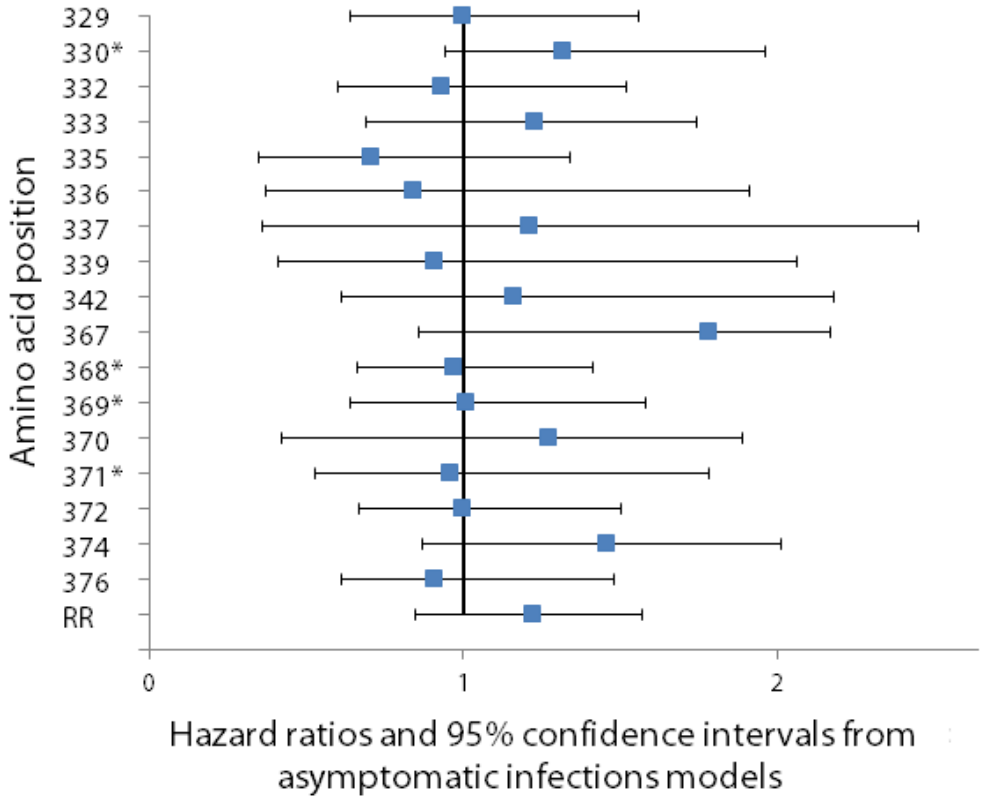

Figure V.6 Association between change in the predominant amino at a polymorphic site and the hazard of *Plasmodium falciparum* infection.

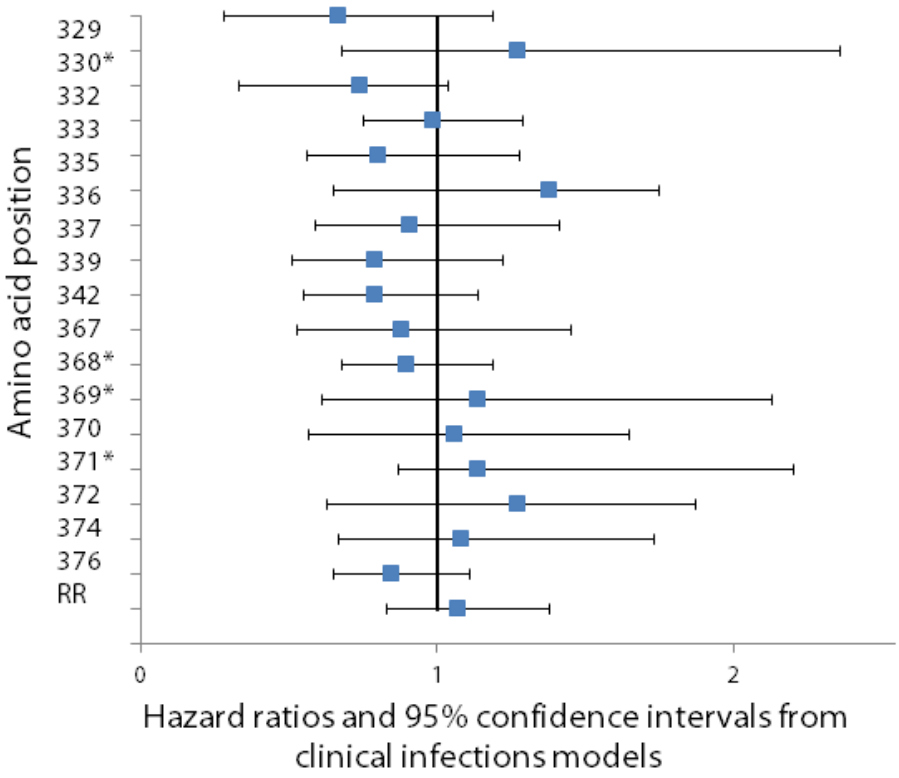

Figure V.7 Association between change in the predominant amino at a polymorphic site and the hazard of *Plasmodium falciparum* infection.

There was no significant difference in the odds of a change occurring between an asymptomatic episode followed by a symptomatic one compared to consecutive asymptomatic episodes, taking into account age and length of interval between episodes, for either the Th regions or the repeat region (Table V.3). Four of the seventeen polymorphic amino acid positions in the Th region had too few observations to model.

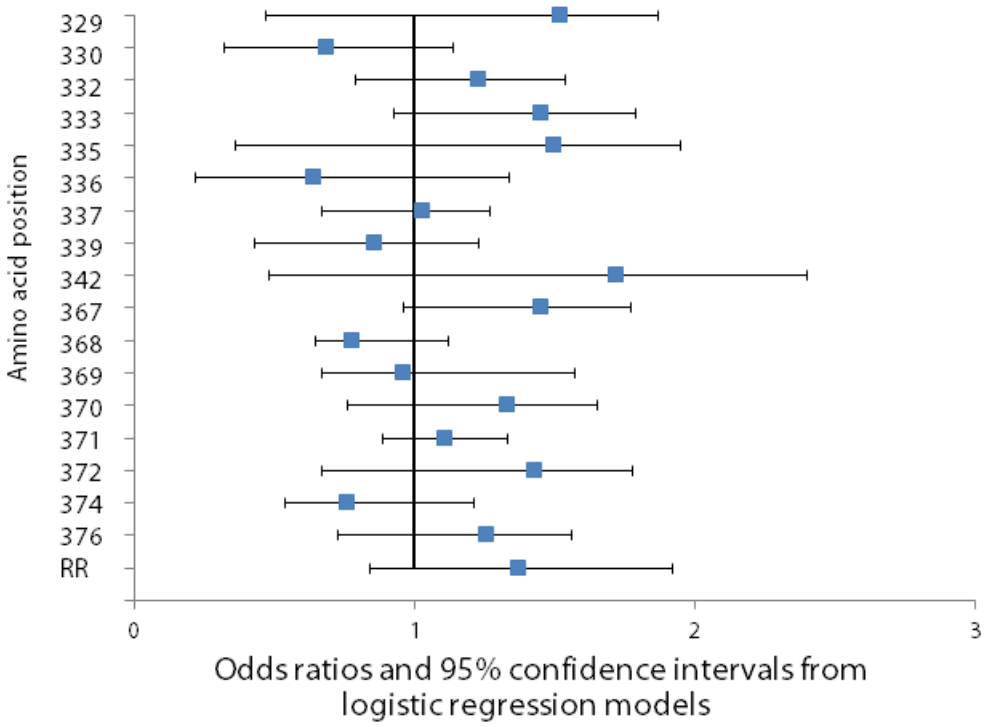

Figure V.8 Association between change in the predominant amino acid at a polymorphic site and clinical disease

E. Discussion

Studies such as this are important in evaluating the potential of polymorphic malaria vaccine antigens. In using this unique sample set in which study participants had at least two years of continuous follow-up, the diversity in the immunogenic regions of the CS protein could be examined in detail over time. No significant associations between within-host dynamics of CSP diversity and the risk of clinical disease or infections were detected in this study. These findings imply that immunity to this protein may not be allele-specific, and have direct implications for the development of vaccines that are based on CSP.

In using 454, the present study was able to include samples containing more than one parasite type, and in doing so circumvented one of the main problems previous studies evaluating the allele-specificity of immunity to CSP have faced. Complete sequence reads were available for each parasite type detected by 454 so haplotypes could be constructed. In 684 samples, 87 haplotypes were found in Th2R, whereas in earlier studies using Sanger sequencing on samples from other West African settings with similar malaria epidemiology, 24 haplotypes were found in 44 Gambian samples[41], and 42 haplotypes were found for Th2R and Th3R combined in 99 samples from Sierra Leone.[68] These numbers likely represent a plateauing of the total amount of diversity present in the Th regions of CSP.

With respect to the repeat region, size determination was very accurate with 94% agreement between lengths obtained by direct sequencing and those determined by high-resolution gel analysis. For all of the remaining 6% of samples length estimates that did not agree were within 12bp (the length of one tetrameric repeat) of eachother. Using this method made it possible to examine the association between repeat region size and increased risk of infection and disease. Although the number of samples that were successfully sequenced was limited, results from one-third of parasite positive samples showed that 67% of successfully sequenced samples had one of three haplotypes with different lengths. This finding suggests that length polymorphism may serve as a reasonable surrogate for haplotype in the Cox and logistic regression models.

This study also had several limitations which may have affected the results. Due to a large degree of polymorphism in the ThRs and a limited sample size, smaller changes in the hazard ratios between consecutive episodes in which there was a change versus no

change for both clinical disease and infection may have been missed. For 7 of the 23 polymorphic sites that were modeled there may not have been enough observations to fully examine associations with infection and disease. Additionally, effects of sequence variation on hazard of clinical disease or infection were not examined with respect to the repeat region due to the limited number of sequences generated. Finally, determination of polyclonal infections for the repeat region was done via capillary gel analysis, which may have been less sensitive than sequencing analysis.

The large number of SNPs and haplotypes in our study is consistent with this region of the *cs* gene being under diversifying selection.[72] Studies have used population genetics to demonstrate positive selection on the T-cell epitope region of the CS protein. One study found that a higher rate of non-synonymous SNPs occurred in the T-cell epitopes of the CSP of P. *falciparum* than would be expected by chance, compared to P. cynomolgi, a malaria parasite that does not infect vertebrate hosts.[77] Although no definitive evidence has been reported of selection of non-vaccine variants of CSP following immunization with CSP-based vaccines, this idea supports the notion that genetic variation in CSP may be driven by the human immune system, implying that naturally acquired and vaccine-induced immunity may be at least to some degree allele-specific. However, the present study as well as a follow-up studies to RTS,S/AS01 vaccine trials[59, 78] all suggest that immunity to CSP may not be allele-specific, in sharp contrast to other diverse vaccine antigens such as the apical membrane antigen 1.[73]

If the diversity in these immunogenic regions is not the target of allele-specific immune responses, the question of what is driving this diversity remains to be answered. In the case of the repeat region, the idea that diversity in this region may be the product

of an immune evasion mechanism by the parasite in which the host mounts non-protective immune response against this region while the parasite escapes into the liver.[79] With respect to the variation in Th regions, the theory that these SNPs may be results of adaptation to the salivary glands of mosquitoes of different species has been proposed.[33] The immunological phenomenon of altered peptide ligand antagonism has also been considered as a possible explanation for diversity in the Th regions.[80] In this process, competing non-protective epitopes in the Th regions may distract Th cells from recognizing the protective epitopes. Related studies have noted that recognition of T-cell epitopes within CSP can be restricted by an individual's human leukocyte antigen (HLA) type, and individuals with certain HLA types are able to recognize a broader variety of the polymorphic epitopes than others.[81, 82] This means individuals with certain HLA types may be better equipped to fight off infection or disease caused by *falciparum* parasites. Combined molecular-immuno-epidemiologicalal studies may be required to test these hypotheses.

VI. DISCUSSION

As the global health community rallies behind the goal of malaria eradication, the need for more effective treatment and prevention strategies is more pressing than ever.[83] Vaccines have been used in nearly all previous successful campaigns to eliminate communicable diseases, and all such unsuccessful campaigns to date were attempted without effective vaccines.[84]

Previous studies of subunit vaccines based on immunogenic antigens from the blood stage of *P. falciparum* have demonstrated limited overall efficacy against clinical malaria. More importantly these studies have highlighted the allele-specific nature of immunity to these antigens raising concerns about the potential for subunit vaccines to be broadly cross-protective and about the danger of these vaccines selecting for non-vaccine variants in a parasite population.

As RTS,S continues to move toward clinical approval, it is important to understand how immunity to this antigen may be achieved. Studies such as those reported here are necessary to understand whether vaccines based on CSP may be subject to the pitfalls of previously studied single-variant blood stage vaccines.

A. Summary of study findings

To address the question of whether or not immunity to CSP may be allele-specific, an accurate representation of parasite diversity present in naturally occurring *P. falciparum* infections had to be generated. Specifically, a reliable way of determining the number and sequence of unique Th2R and Th3R haplotypes in a clinical sample, as well as which haplotype was the predominant one in infection were the most important considerations when developing a technique to evaluate *cs* sequences.

The methods validation study showed that 454 sequencing was more sensitive in assessing diversity in the *cs* gene vs. Sanger sequencing, which has been regarded as the gold standard for de novo sequencing. 454 detected approximately 41% more SNPs in Th2R and 64% more SNPs in Th3R than Sanger sequencing. 454 was also more accurate in identifying the diversity in a known mixture of parasite clones. The percentage of predominant alleles that were correctly identified was 91% and 75% for 454 and Sanger sequencing respectively. These results indicate that 454 is a superior method for detecting parasite diversity as well as a more sensitive and accurate way of detecting majority alleles in a mixed infection with respect to the Th regions.

Once methods for sequence generation were established, study samples were sequenced and analyzed. Prevalences of the most common haplotypes for Th2R, Th3R and the repeat region were calculated and Fisher's exact tests were performed to evaluate whether there were any associations between haplotype prevalence and season, age, and presence of clinical disease. No significant differences in haplotype distribution were noted between any of these covariates. These descriptive analyses were important in determining the baseline diversity present in a natural population of parasites and whether this diversity varies over time, changes with age, or is different between infections and clinical cases of *falciparum* malaria. All of these factors could affect vaccine efficacy, if it were partially or completely allele-specific. Additionally, these data are relevant in cataloguing diversity from different geographic regions for the prospect of designing a multivalent vaccine that may have broader cross-protectivity and efficacy. Interestingly, Th2R and Th3R haplotypes identical to those of the 3D7 variant of *P. falciparum*, which is the variant upon which the RTS,S vaccine is based, were among the most prevalent.

This again could have implications for the efficacy of RTS,S in Mali and settings with similar prevalences of *cs* haplotypes.

Due to limitations of current sequencing technologies, full length sequences for the repeat region could not be successfully generated for every sample. Therefore, haplotype prevalence data available for these samples are limited. From the data available through direct sequencing, only 19 distinct haplotypes were detected from 157 successfully sequenced samples. In the methods validation study 72 unique haplotypes were detected with respect to Th2R and 14 unique haplotypes were detected with respect to Th3R from a total of 50 samples. This discrepancy may be due to the power of 454 in detecting minor haplotypes that cannot be extrapolated through Sanger sequencing. The limitations with regard to data generation for the repeat region will be discussed more fully in the Advantages and Limitations section.

Once the descriptive analyses were done, the main study question of whether there is evidence to support the hypothesis that immunity to CSP is allele-specific was addressed. Cox proportional hazards models were used to examine associations between changes at polymorphic sites and hazard of infection or clinical disease. The goals of these models were to capture the effect of host exposure to a parasite variant to which they have previously not mounted an immune response, as well as identify which polymorphic sites may be more important in allele-specific immunity. Of the 23 polymorphic sites that were identified in both Th2R and Th3R, none had statistically significant results in the Cox proportional hazards models. Logistic regression models were also constructed to calculate the log odds of individuals having a change at a certain polymorphic site in intervals in which an asymptomatic parasitemia was followed by a

symptomatic one, versus individuals having the same change in intervals in which they are consecutively asymptomatic. This analysis was done to help clarify whether changes at polymorphic sites occur by chance or whether they are truly associated with increased risk of clinical disease. No significant associations with changes at polymorphic sites and clinical disease were found in this analysis either. These findings imply that immunity to CSP may not be allele-specific. The associations between the within-host diversity of the immunogenic regions of CSP and clinical and non-clinical *Plasmodium falciparum* infections have never before been examined.

B. Advantages and limitations of study

This study addressed several important questions that have previously not been explored in depth with regard to diversity in CSP. The number of sequences generated for the Th regions allowed for evaluation of the polymorphism to a degree that has previously not been done. In 684 samples, 87 haplotypes were found in Th2R, whereas in earlier studies using Sanger sequencing on samples from other West African settings with similar malaria epidemiology, 24 haplotypes were found in 44 Gambian samples[41], and 42 haplotypes were found for Th2R and Th3R combined in 99 samples from Sierra Leone.[68] These numbers likely represent a plateauing of the total amount of diversity present in the Th regions of CSP.

The next generation sequencing technology used in this study had significant advantages over traditional sequences methods. In using 454 to sequence clinical samples, the diversity present in complex infections could be examined. Complete sequence reads are available for each parasite type detected by 454 so haplotypes can be constructed. A conservative threshold for haplotype inclusion was set in this study to

reduce the possibility of including erroneous sequences in the analysis. In doing so, rare variants in the population may have been missed and therefore associations between the population level and within-host dynamics at these polymorphic sites would not have been examined. However, in a study that is focused on evaluating immunity to predominant alleles, excluding rare variants should not influence study findings.

The lack of sequences for the repeat region represented a significant obstacle to the accurate evaluation of diversity in this region. In order to address this issue, a high resolution gel system was used to determine size polymorphism in this region. This method proved to be very accurate with 94% agreement between lengths obtained by direct sequencing and those determined by high-resolution gel analysis, and allowed for the inclusion of the repeat region in the Cox proportional hazards and logistic regression models. It was also determined that 68% of the samples had one of three haplotypes which differ in length. This finding indicates that size variation may serve as a reasonable surrogate for sequence polymorphism.

Although significant strides were made in describing diversity in the immunogenic regions of CSP as well as addressing the question of whether immunity to CSP is allele-specific, several limitations may have affected these results. First, limited sample size may have influenced both the descriptive and within-host analyses. There were 63 missing or empty sample vials, 85 samples from which no Th region could be amplified, and 265 samples from which no repeat region could be amplified. This may have been due to the age of these samples and breakdown of the extracted parasite DNA over time. As the repeat region is a much larger stretch of DNA, it is more susceptible to fragmentation than the Th region. Additionally, only 157 sequences were generated for

the repeat region. This number was not larger because only the repeat region from "single clone" samples (as determined by 454 analysis of the Th regions) could be sequenced via Sanger sequencing.

This limited sample size as well as large degree of polymorphism in the ThRs, means that smaller changes in the hazard ratios between consecutive episodes in which there was a change versus no change for both clinical disease and infection may have been missed. For 7 of the 23 polymorphic sites that were modeled there may not have been enough observations to fully examine associations with infection and disease. Additionally, effects of sequence variation on hazard of clinical disease or infection were not examined with respect to the repeat region. Although size polymorphism for the repeat region was studied, since capillary gel analysis is less sensitive than sequencing analysis, the estimate for the number of polyclonal infections with respect to the repeat region may be low.

C. Implications for vaccine design

This study has several important implications for vaccine design. First, the descriptive analysis suggests that polymorphism in both Th regions as well as size variation in the repeat region appears to be stable over time. This indicates that if immunity to CSP were at least in part allele-specific the most common parasite types with regard to the immunogenic regions of CSP do not appear to change from season to season. Secondly, the large number of SNPs and haplotypes found in this study is consistent with this region being under diversifying selection.[19] One theory of why genetic variation in CSP exists is that it may be driven by the human immune system,[85] implying that naturally acquired and vaccine-induced immunity may be at least to some

degree allele-specific. However, the present study as well as follow-up studies to RTS,S/AS01 vaccine trials[59, 78] all suggest that immunity to CSP may not be allele-specific, in sharp contrast to other diverse vaccine antigens such as the apical membrane antigen 1.[73]If this is the case, then vaccines based on this protein may not need to include multiple variants in order to achieve broad protection. However, despite being the largest study of *cs* genetic diversity, it is possible that this study was unable to detect more subtle effects on the modeled outcomes due to limited sample size. It is also possible that naturally acquired and vaccine induced immunity work in different ways. A large-scale molecular epidemiology study such as this performed in the context of a CSP vaccine trial is warranted.

D. Diversity in the circumsporozoite protein

If the diversity in the immunogenic regions of CSP is not the target of allele-specific immune responses, the question of what is driving this diversity remains to be answered. In the case of the repeat region, the idea that diversity in this region may be the product of an immune evasion mechanism by the parasite in which the host mounts non-protective immune responses against this region while the parasite escapes into the liver.[79] In this process, the immunodominant epitopes in the repeat region of CSP stimulate the expansion of B-cells that are specific to this region, and inhibit the expansion of B-cells, specific to adjacent epitopes, that may confer a protective immune response. With respect to the variation in Th regions, a similar theory has been proposed. The immunological phenomenon of altered peptide ligand antagonism has also been considered as a possible explanation for diversity in the Th regions.[80] In this process, competing non-protective epitopes in the Th regions may distract Th cells from

recognizing the protective epitopes. Related studies have noted that recognition of T-cell epitopes within CSP can be restricted by an individual's human leukocyte antigen (HLA) type.[81, 82] Furthermore, individuals with certain HLA types are able to recognize a broader variety of the polymorphic epitopes than others. This means individuals with certain HLA types may be better equipped to fight off infection or disease caused by *falciparum* parasites. Another possibility that has been raised is that SNPs in the Th regions may be results of adaptation to the salivary glands of mosquitoes of different species.[33]

E. Future directions

Although this study did not find any evidence of allele-specific immunity to CSP with respect to the repeat region, the lack of complete sequence data for this region prevents firm conclusions from being drawn. As the read length capability of 454 increases, complete sequencing of the repeat region may soon be possible so this question can be addressed more fully.

In addition to sequencing the repeat region as well as polymorphic regions N-terminal to the repeat region that we did not examine because they are not included in the RTS,S vaccine, combining the molecular data generated from a study such as this with actual measurements of antibody titers as well as T-cell responses would provide a definitive answer as to whether immunity to CSP is allele-specific. Furthermore, HLA typing of infected individuals would help assess, first, whether some HLA types are more or less susceptible to infection or disease, and second, which parasite types are recognized by which HLA types.

Finally, to assess whether selective pressure is being exerted upon the *falciparum* parasite by the mosquito, mosquito susceptibility studies to parasites that vary only with respect to the Th regions could help test this theory.

VII. REFERENCES

1. Takala SL, Plowe CV. Genetic diversity and malaria vaccine design, testing and efficacy: preventing and overcoming 'vaccine resistant malaria'. *Parasite Immunol*. 2009;31:560-573.

2. Genton B, Betuela I, Felger I, et al. A recombinant blood-stage malaria vaccine reduces Plasmodium falciparum density and exerts selective pressure on parasite populations in a phase 1-2b trial in Papua New Guinea. *J Infect Dis*. 2002;185:820-827.

3. Guevara Patino JA, Holder AA, McBride JS, Blackman MJ. Antibodies that inhibit malaria merozoite surface protein-1 processing and erythrocyte invasion are blocked by naturally acquired human antibodies. *J Exp Med*. 1997;186:1689-1699.

4. World Health Organization. WHO World Malaria Report 2011. Available at: http://www. who.int/malaria/world_malaria_ report_2011/en/. Accessed February, 26th, 2013.

5. Murray CJ, Rosenfeld LC, Lim SS, et al. Global malaria mortality between 1980 and 2010: a systematic analysis. *Lancet*. 2012;379:413-431.

6. WHO | Malaria Available at: http://www.who.int/mediacentre/factsheets/fs094/en/index.html. Accessed 10/28/2009, 2009.

7. G8 Summitt Documents 2007: Declaration of Growth and Responsibility in Africa. Available at: http://www.g-8.de/Content/DE/Artikel/G8Gipfel/Anlage/Abschlusserkl_C3_A4rungen/WV-afrika-en,templateId=raw,property=publicationFile.pdf/WV-afrika-en.pdf.

8. Wongsrichanalai C, Pickard AL, Wernsdorfer WH, Meshnick SR. Epidemiology of drug-resistant malaria. *Lancet Infect Dis*. 2002;2:209-218.

9. Talisuna AO, Bloland P, D'Alessandro U. History, dynamics, and public health importance of malaria parasite resistance. *Clin Microbiol Rev*. 2004;17:235-254.

10. World Health Organization. *Guidelines for the Treatment of Malaria.* Second ed. Geneva, Switzerland: WHO Press; 2010.

11. Roper C, Pearce R, Nair S, Sharp B, Nosten F, Anderson T. Intercontinental spread of pyrimethamine-resistant malaria. *Science*. 2004;305:1124.

12. Mita T, Venkatesan M, Ohashi J, et al. Limited geographical origin and global spread of sulfadoxine-resistant dhps alleles in Plasmodium falciparum populations. *J Infect Dis*. 2011;204:1980-1988.

13. Noedl H, Se Y, Schaecher K, et al. Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med*. 2008;359:2619-2620.

14. Dondorp AM. Artemisinin resistance in Plasmodium falciparum malaria. *N Engl J Med*. 2009;361:455.

15. History | CDC Malaria Available at: http://www.cdc.gov/malaria/history/index.htm. Accessed 10/30/2009, 2009.

16. Hemingway J, Ranson H. Insecticide resistance in insect vectors of human disease. *Annu Rev Entomol*. 2000;45:371-391.

17. WHO. Atlas of insecticide resistance in malaria vectors of the WHO African region. Available at: http://www.afro.who.int/des/phe/publications/atlas_final_version.pdf. Accessed 10/30/2009, 2009.

18. Pinto J, Lynd A, Vicente JL, et al. Multiple origins of knockdown resistance mutations in the Afrotropical mosquito vector Anopheles gambiae. *PLoS One*. 2007;2:e1243.

19. Casimiro S, Coleman M, Mohloai P, Hemingway J, Sharp B. Insecticide resistance in Anopheles funestus (Diptera: Culicidae) from Mozambique. *J Med Entomol*. 2006;43:267-275.

20. Kelly-Hope L, Ranson H, Hemingway J. Lessons from the past: managing insecticide resistance in malaria control and eradication programmes. *Lancet Infect Dis*. 2008;8:387-389.

21. Reeder JC, Brown GV. Antigenic variation and immune evasion in Plasmodium falciparum malaria. *Immunol Cell Biol*. 1996;74:546-554.

22. Baer K. Release of Hepatic Plasmodium yoelii Merozoites into the Pulmonary Microvasculature. *PLOS pathogens*. 2007;3:e171.

23. Jones TR, Hoffman SL. Malaria vaccine development. *Clin Microbiol Rev*. 1994;7:303-310.

24. Hoffman SL, Billingsley PF, James E, et al. Development of a metabolically active, non-replicating sporozoite vaccine to prevent Plasmodium falciparum malaria. *Hum Vaccin*. 2010;6:97-106.

25. Bill and Melinda Gates Annual Report. Available at: http://www.gatesfoundation.org/nr/public/media/annualreports/annualreport03/HTML/global.html.

26. National Institute of Allergy and Infectious Diseases. Malaria Life Cycle Diagram. Available at: http://www3.niaid.nih.gov/labs/aboutlabs/lmiv/productGroups/default.htm. Accessed 11/2/2009, 2009.

27. Anders RF, Saul A. Malaria vaccines. *Parasitol Today*. 2000;16:444-447.

28. Matuschewski K, Mueller AK. Vaccines against malaria - an update. *FEBS J*. 2007;274:4680-4687.

29. Polley SD, Tetteh KK, Lloyd JM, et al. Plasmodium falciparum merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. *J Infect Dis*. 2007;195:279-287.

30. Takala SL, Coulibaly D, Thera MA, et al. Dynamics of polymorphism in a malaria vaccine antigen at a vaccine-testing site in Mali. *PLoS Med*. 2007;4:e93.

31. Ogutu BR, Apollo OJ, McKinney D, et al. Blood stage malaria vaccine eliciting high antigen-specific antibody concentrations confers no protection to young children in Western Kenya. *PLoS One*. 2009;4:e4708.

32. Thera M, Doumbo O, Coulibaly D, et al. A field trial to assess a blood-stage malaria vaccine. *N Engl J Med*. 2011;365:1004-1013.

33. Kumkhaek C, Phra-Ek K, Renia L, et al. Are extensive T cell epitope polymorphisms in the Plasmodium falciparum circumsporozoite antigen, a leading sporozoite vaccine candidate, selected by immune pressure? *J Immunol*. 2005;175:3935-3939.

34. Lalvani A, Moris P, Voss G, et al. Potent induction of focused Th1-type cellular and humoral immune responses by RTS,S/SBAS2, a recombinant Plasmodium falciparum malaria vaccine. *J Infect Dis*. 1999;180:1656-1664.

35. Sun P, Schwenk R, White K, et al. Protective immunity induced with malaria vaccine, RTS,S, is linked to Plasmodium falciparum circumsporozoite protein-specific CD4+ and CD8+ T cells producing IFN-gamma. *J Immunol*. 2003;171:6961-6967.

36. Escalante AA, Grebert HM, Isea R, et al. A study of genetic diversity in the gene encoding the circumsporozoite protein (CSP) of Plasmodium falciparum from different transmission areas--XVI. Asembo Bay Cohort Project. *Mol Biochem Parasitol*. 2002;125:83-90.

37. Rich SM, Hudson RR, Ayala FJ. Plasmodium falciparum antigenic diversity: evidence of clonal population structure. *Proc Natl Acad Sci U S A*. 1997;94:13040-13045.

38. Egan JE, Hoffman SL, Haynes JD, et al. Humoral immune responses in volunteers immunized with irradiated Plasmodium falciparum sporozoites. *Am J Trop Med Hyg*. 1993;49:166-173.

39. Nardin EH, Nussenzweig RS, McGregor IA, Bryan JH. Antibodies to sporozoites: their frequent occurrence in individuals living in an area of hyperendemic malaria. *Science*. 1979;206:597-599.

40. Chenet SM, Branch OH, Escalante AA, Lucas CM, Bacon DJ. Genetic diversity of vaccine candidate antigens in Plasmodium falciparum isolates from the Amazon basin of Peru. *Malar J*. 2008;7:93.

41. Weedall GD, Preston BM, Thomas AW, Sutherland CJ, Conway DJ. Differential evidence of natural selection on two leading sporozoite stage malaria vaccine candidate antigens. *Int J Parasitol*. 2007;37:77-85.

42. Nardin EH, Nussenzweig RS. T cell responses to pre-erythrocytic stages of malaria: role in protection and vaccine development against pre-erythrocytic stages. *Annu Rev Immunol*. 1993;11:687-727.

43. de Groot AS, Johnson AH, Maloy WL, et al. Human T cell recognition of polymorphic epitopes from malaria circumsporozoite protein. *J Immunol*. 1989;142:4000-4005.

44. Good MF, Pombo D, Quakyi IA, et al. Human T-cell recognition of the circumsporozoite protein of Plasmodium falciparum: immunodominant T-cell domains map to the polymorphic regions of the molecule. *Proc Natl Acad Sci U S A*. 1988;85:1199-1203.

45. NCBI Genome. Available at: http://www.ncbi.nlm.nih.gov/sites/genome. Accessed 11/24/2009, 2009.

46. Gordon DM, McGovern TW, Krzych U, et al. Safety, immunogenicity, and efficacy of a recombinantly produced Plasmodium falciparum circumsporozoite protein-hepatitis B surface antigen subunit vaccine. *J Infect Dis*. 1995;171:1576-1585.

47. Stoute JA, Slaoui M, Heppner DG, et al. A preliminary evaluation of a recombinant circumsporozoite protein vaccine against Plasmodium falciparum malaria. RTS,S Malaria Vaccine Evaluation Group. *N Engl J Med*. 1997;336:86-91.

48. Soares IS, Rodrigues MM. Malaria vaccine: roadblocks and possible solutions. *Braz J Med Biol Res*. 1998;31:317-332.

49. Bojang KA, Milligan PJ, Pinder M, et al. Efficacy of RTS,S/AS02 malaria vaccine against Plasmodium falciparum infection in semi-immune adult men in The Gambia: a randomised trial. *Lancet*. 2001;358:1927-1934.

50. Alonso PL, Sacarlal J, Aponte JJ, et al. Efficacy of the RTS,S/AS02A vaccine against Plasmodium falciparum infection and disease in young African children: randomised controlled trial. *Lancet*. 2004;364:1411-1420.

51. Aponte JJ, Aide P, Renom M, et al. Safety of the RTS,S/AS02D candidate malaria vaccine in infants living in a highly endemic area of Mozambique: a double blind randomised controlled phase I/IIb trial. *Lancet*. 2007;370:1543-1551.

52. Bejon P, Lusingu J, Olotu A, et al. Efficacy of RTS,S/AS01E vaccine against malaria in children 5 to 17 months of age. *N Engl J Med*. 2008;359:2521-2532.

53. Garcon N, Chomez P, Van Mechelen M. GlaxoSmithKline Adjuvant Systems in vaccines: concepts, achievements and perspectives. *Expert Rev Vaccines*. 2007;6:723-739.

54. Mettens P, Dubois PM, Demoitie MA, et al. Improved T cell responses to Plasmodium falciparum circumsporozoite protein in mice and monkeys induced by a novel formulation of RTS,S vaccine antigen. *Vaccine*. 2008;26:1072-1082.

55. Stewart VA, McGrath SM, Dubois PM, et al. Priming with an adenovirus 35-circumsporozoite protein (CS) vaccine followed by RTS,S/AS01B boosting significantly improves immunogenicity to Plasmodium falciparum CS compared to that with either malaria vaccine alone. *Infect Immun*. 2007;75:2283-2290.

56. First Results of Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Children. *N Engl J Med*. 2011;365:1863-1875.

57. Agnandji ST. A phase 3 trial of RTS,S/AS01 malaria vaccine in African infants. *N Engl J Med*. 2012;367:2284.

58. Alloueche A, Milligan P, Conway DJ, et al. Protective efficacy of the RTS,S/AS02 Plasmodium falciparum malaria vaccine is not strain specific. *Am J Trop Med Hyg*. 2003;68:97-101.

59. Waitumbi J. Impact of RTS,S/AS02A and RTS,S/AS01B on Genotypes of P. falciparum in Adults Participating in a Malaria Vaccine Clinical Trial. *PLoS clinical trials*. 2009;6:1.

60. Arnot D. Circumsporozoite Protein of Plasmodium vivax: Gene Cloning and Characterization of the Immunodominant Epitope. *Science*. 1985;230:815.

61. Jalloh A, van Thien H, Ferreira MU, et al. Sequence variation in the T-cell epitopes of the Plasmodium falciparum circumsporozoite protein among field isolates is temporally stable: a 5-year longitudinal study in southern Vietnam. *J Clin Microbiol*. 2006;44:1229-1235.

62. Coulibaly D, Diallo DA, Thera MA, et al. Impact of preseason treatment on incidence of falciparum malaria and parasite density at a site for testing malaria vaccines in Bandiagara, Mali. *Am J Trop Med Hyg*. 2002;67:604-610.

63. Williams R, Peisajovich SG, Miller OJ, Magdassi S, Tawfik DS, Griffiths AD. Amplification of complex gene libraries by emulsion PCR. *Nat Methods*. 2006;3:545-550.

64. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A.* 1988;85:2444-2448.

65. Allison PD. Fixed-Effect Partial Likelihood for Repeated Events. *Sociological Methods & Research*. ;25:207-222.

66. SAS Institute I. SAS/STAT 9.2 user's guide. In: Cary, North Carolina: SAS Institute, Inc; 2008.

67. Fleiss JL. *Statistical Methods for Rates and Proportions.* Second ed. New York: John Wiley & Sons; 1981.

68. Jalloh A, Jalloh M, Matsuoka H. T-cell epitope polymorphisms of the Plasmodium falciparum circumsporozoite protein among field isolates from Sierra Leone: age-dependent haplotype distribution? *Malar J*. 2009;8:120.

69. World's largest malaria vaccine trial now underway in seven African countries. Available at: http://www.gsk.com/media/pressreleases/2009/2009_pressrelease_10124.htm.

70. Carr IM, Robinson JI, Dimitriou R, Markham AF, Morgan AW, Bonthron DT. Inferring relative proportions of DNA variants from sequencing electropherograms. *Bioinformatics*. 2009;25:3244-3250.

71. Hunt P, Fawcett R, Carter R, Walliker D. Estimating SNP proportions in populations of malaria parasites by sequencing: validation and applications. *Mol Biochem Parasitol*. 2005;143:173-182.

72. Ochola LI, Tetteh KK, Stewart LB, Riitho V, Marsh K, Conway DJ. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in Plasmodium falciparum. *Mol Biol Evol*. 2010;27:2344-2351.

73. Takala SL, Coulibaly D, Thera MA, et al. Extreme polymorphism in a vaccine antigen and risk of clinical malaria: implications for vaccine development. *Sci Transl Med*. 2009;1:2ra5.

74. Ouattara A. Molecular basis of allele-specific efficacy of a blood-stage malaria vaccine: vaccine development implications. *J Infect Dis*. 2013;207:511-519.

75. Gandhi K. Next generation sequencing to detect variation in the Plasmodium falciparum circumsporozoite protein. *Am J Trop Med Hyg*. 2012;86:775.

76. Rice P. Longden, A. Bleasby. EMBOSS: the European Molecular Biology Open Software Suite. *Trends in Genetics 16(6): 276-277.*

77. Hughes AL. Circumsporozoite protein genes of malaria parasites (Plasmodium spp.): evidence for positive selection on immunogenic regions. *Genetics*. 1991;127:345-353.

78. Enosse S. RTS,S/AS02A malaria vaccine does not induce parasite CSP T cell epitope selection and reduces multiplicity of infection. *PLoS clinical trials*. 2006;1:e5.

79. Schofield L. The circumsporozoite protein of Plasmodium: a mechanism of immune evasion by the malaria parasite? *Bull World Health Organ*. 1990;68 Suppl:66-73.

80. Gilbert SC. Association of malaria parasite population structure, HLA, and immunological antagonism. *Science*. 1998;279:1173-1177.

81. Doolan DL. HLA-DR-promiscuous T cell epitopes from Plasmodium falciparum pre-erythrocytic-stage antigens restricted by multiple HLA class II alleles. *The journal of immunology*. 2000;165:1123.

82. Doolan DL. Degenerate cytotoxic T cell epitopes from P. falciparum restricted by multiple HLA-A and HLA-B supertype alleles. *Immunity*. 1997;7:97.

83. Plowe CV. The potential role of vaccines in the elimination of falciparum malaria and the eventual eradication of malaria. *J Infect Dis*. 2009;200:1646-1649.

84. Henderson DA. Lessons from the eradication campaigns. *Vaccine*. 1999;17 Suppl 3:S53-5.

85. Putaporntip C. Natural selection maintains a stable polymorphism at the circumsporozoite protein locus of Plasmodium falciparum in a low endemic area. *Infection, genetics and evolution*. 2009;9:567-573.