Temporal Haplotype Changes of Vaccine Candidates for Malaria

Ijeoma Meremikwu

Thesis Advisor: Dr. Karamoko Niaré

Second Reader: Dr. Jeffrey Bailey

This thesis is submitted in partial fulfillment of the requirements for the degree

Bachelor of Science with Honors

Center of Computational Molecular Biology

Brown University

April 2022

Table of Contents

ACKNOWLEDGMENTS 3
INTRODUCTION 4
Background on Malaria **4**
Vaccine Development Efforts **6**
PROTECTIVE IMMUNITY 8
Naturally Acquired Immunity **8**
Allele-Specific Immunity **9**
HYPOTHESIS AND OBJECTIVES 10
..... 10
STUDY DESIGN AND METHODOLOGY 11
MalariaGEN v6.0 Data **11**
Generation of Unphased and Phased VCFs **12**
Sliding Window FST Analysis and FST Peak Identification **13**
Sliding Window Tajima's D Analysis and Tajima's D Peak Identification **13**
T-cell Epitope Prediction with NetMHC4.1pan and NetMHCII4.0pan **13**
Bioinformatic and Visualization Tools **14**
..... 14
RESULTS 15
DISCUSSION 17
APPENDIX 18
Allele Frequency Analysis of Vaccine Candidates **18**
Sliding Window FST Analysis of Vaccine Candidates **19**
Sliding Window Tajima's D Analysis of Vaccine Candidates **20**
Haplotype Frequency Across Years of Vaccine Candidates **21**
T-cell Epitope prediction across HLA supertypes **24**
..... 25
REFERENCES 26

Acknowledgments

I would like to thank Dr. Karamoko Niaré and Dr. Jeffrey Bailey for their mentorship and guidance throughout my thesis research process and training. I am immensely grateful for the time, support, and meetings done for this project. Thank you so much. I would like to thank CCMB Department and faculty members for supporting me throughout my research training and undergraduate career for my coursework. I want to thank Dr. Sohini Ramachandran for her advice and support during this process and Nathaniel Gill for reliable, administrative support and answers throughout registration and deadlines for the Honors Process. I would like to thank MOSAIC+ for providing me with a space and a community within the CS department. Thank you to Amanda Elyssa Ruiz for being a great peer mentor within the Pathology and Laboratory Medicine Department. Thank you to my family (Mom, Dad, Uzoma, Ezinne) and friends for their support within both remote and physical spaces.

CENTER FOR
Computational
Molecular Biology

Introduction

Background on Malaria

Malaria is a life-threatening, feverish disease that presents a major global health concern. There are 241 million documented malaria cases across the globe[1]. About 95% of worldwide cases are found across the African countries, with 55% of these cases in six of these countries. Children under the age of 5 experience severe malaria symptoms and have a high mortality rate[2] [3]. Severe clinical symptoms such as cerebral malaria, respiratory distress syndrome, organ failures, anemias, destruction, and rupture of blood cells (erythrocytes) have been documented in infected patients[4]. Rupturing of erythrocytes and signaling calls for failure to reproduce more erythrocytes for body homeostatic functions. Furthermore, malaria has caused both environmental and psychological distress for affected communities within endemic regions. This is a public health concern as it creates limitations to pursuing activities essential for patients' quality of life[5] [6].

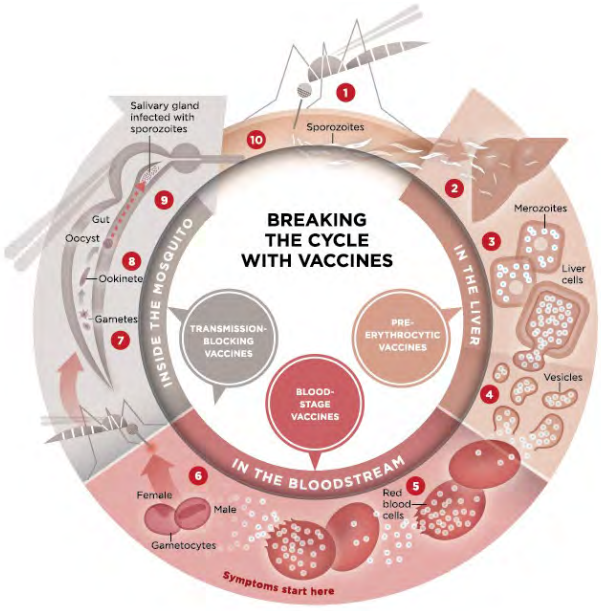

Figure 1 Infographic representing the *Plasmodium falciparum* life cycle and specific targets of the cycle for three different vaccine strategies.

(Infographic Source: https://www.malariavaccine.org/malaria-and-vaccines/vaccine-development/life-cycle-malaria-parasite)

Malaria is caused by *Plasmodium* parasite species infection. The *Plasmodium* parasite species are carried by female-sex *Anopheles* mosquitos. Within the mosquito the *Plasmodium* species undergoes the sexual life cycle and produces sporozoites. These sporozoites can live within the mosquito's saliva. The sporozoites enter the host through mosquito bites and infect the hosts' hepatocytes. Sporozoites creates the parasitophorous vacuole, which is the space where they feed and undergo the asexual life cycle. Intracellularly, these sporozoites mature into

trophozoites. Within the vacuole, the trophozoite will into a schizont containing 8-24 merozoites[4]. In the erythrocytic (blood) stage, these merozoites burst from cells, and enter the bloodstream to infect other erythrocytes. The asexual life cycle continues to produce more infecting merozoites.

About 90% of worldwide malaria cases are caused by *Plasmodium falciparum*[7] [8]. *P. falciparum* is an Apicomplexan, obligate parasite with 14 chromosomes and eukaryotic and bacterial organelles[9]. In endemic regions, *P. falciparum* infections happen during malaria seasons. The seasonal malaria transmission varies based on geographic locations due to rainy and dry seasons [10] [11] [12]. This seasonal malaria transmission poses both a disadvantage to our immune systems and an advantage for the parasite. The immune response in one malaria season would not provide protective immunity against another season. The genetic diversity in *P. falciparum* produces protein variants that can evade host immune responses across different seasons.

However, adults who have lived in endemic, high transmission regions have been documented to develop naturally acquired immunities (NAI). NAI is the development of protective immunity that increases with age and exposure to the parasites[13] . Naturally acquired immunity (NAI) is still not well understood. It has been hypothesized to act as a strain-specific and/or strain-transcending protective immunity [14] [15]. Furthermore, some studies indicate that the NAI response may not involve strong adaptive T-cell immune responses, with evidence of low parasite burden still seen in these individuals. However, what is clear is that allele-specific immunity is gained against antigenic variants within one season. This can protect the host from developing severe clinical symptoms [16] [17] [18]. Therefore, the relationship between allele-specific immune targets and NAI could provide insight into long-term immune protection against *P. falciparum*. The goal is to develop an effective vaccine strategy based on antigens that elicit strong immune responses.

Antigens of important proteins involved in the successful invasion of host cells have been shown to elicit both adaptive immune and complement system responses. Studies have identified that pre-erythrocytic and erythrocytic stage antigens are connected to specific type of immune responses [19] [20]. Antigens within pre-erythrocytic stage proteins can activate both complement and Th1-mediated responses [21]. Targeting the pre-erythrocytic proteins can deter further development of the parasite, and potentially start the clearance of the parasite before entering the erythrocytic stage. This also could prevent the development of clinical symptoms, as invasion and parasite replication are stopped.

Antigens of erythrocytic-stage proteins can elicit CD4+ T-cell mediated, B-cell mediated, and complement system responses [21] [23]. However, finding these antigens in this stage may be hindered by the smart parasite life cycle. The burst size in the liver stage produces an abundant amount of merozoites, enough to start the widespread infection of erythrocytes. On the other hand, the burst size in the erythrocytic stage, produces a much smaller amount of merozoites (8 – 24). In the erythrocytic stage, the adaptive immune system would have a smaller sample of merozoites to identify antigens from. However, the real challenge in identifying antigen epitopes (peptide portions of antigens) lies in the role that the *P. falciparum* genome plays. The parasite's genome diversity creates antigenic variants across multiple seasons.

Vaccine Development Efforts

P. falciparum genome garners mutations due to natural selection pressures and drug resistance. Through these mutations, antigenic variation increases and can evade antibody recognition. This antigenic variation presents one of the current challenges to vaccine development against malaria parasites. Immune responses developed against antigens of the leading vaccine candidates have been documented in previous studies [24] [25] . AMA1, CSP, RH5, MSP1, MSP2, and MSP3 are erythrocytic stage proteins with identified B-cell and T-cell epitopes[21] [26]. PfSPZ is an attenuated sporozoite vaccine, with chemoprophylaxis and genetically attenuated versions [27] . This vaccine showcased efficacy in lab experiments, but logistic challenges in the deployment of such a vaccine stopped further progress in this direction [28] [29] . Sub-unit adjuvant vaccines have also been developed based on CSP in the past decade and are undergoing clinical trials [30]. CSP's NANP-repeat and TH2/TH3 regions contain B-cell and T-cell epitopes and have been used to develop subunit RTS, S AS01E vaccine [31] [32] [33] . While there are numerous efforts to develop vaccines for CSP and pre-erythrocytic candidates, there are current research efforts to further explore more vaccine candidates to develop blood stage vaccines based on the 5200-gene library for *P. falciparum*. Anti-merozoite vaccines currently under development are based on AMA1, MSP3, and GLURP proteins, but still do not show evidence of protection and efficacy in individuals [34] [35] [36].

In addition, clinical trials are testing the efficacy of transmission-blocking vaccines developed based on gametocyte proteins such as Pfs230, Pfs25, Pfs28, and Pfs48/45 in *P. falciparum* [37]. These focus on stopping necessary mechanisms and proteins involved in the sexual stage of *P. falciparum*, which would prevent sporozoites from being able to enter the host (preventing infection altogether). There are pre-fertilization (Pfs230 and fs48/45) and post-fertilization antigens (Pfs25 and Pfs28) identified in the last decade that present potential antibody-antigen targets. The main challenges with transmission-blocking vaccines are the lack of antibody production and maintenance of their immune responses. In addition, there are difficulties with commercial and lab production of their recombinant protein versions. Whole C-terminal, N-terminal, and specific regions of genes are being tested for vaccine efficacy. However, with the genetic diversity creating numerous protein variants, new methods that predict regions as targets for immunity must be explored.

With seasonal variation that allows for selection pressures on alleles, the vaccine research community has shifted goals into creating a vaccine developed on multiple epitopes with high binding affinity and capability of eliciting strong immune responses. The goal is to understand the genomic diversity of *P. falciparum* isolates across malaria seasons to identify targets of allele-specific immunity and find epitopes that could elicit T-cell and B-cell mediated responses.

While the entire genome of *P. falciparum* has been interpreted to be under negative selection, there could be regions within the vaccine candidate genes that are under balancing selection pressures [38]. Loci that are under balancing selection could have variants that would be competing within the population. Perhaps, these could be protein variants created to evade and trick our immune systems. So, to understand how the parasite creates these variants, it is important to look at sites within the genes that allow for competition between alleles. This may provide information on where targets of allele-specific immunity are located. Finding genomic

targets of allele-specific immunity and looking at predicted epitopes would increase our knowledge of which regions to target for vaccine development.

Protective Immunity

Naturally Acquired Immunity

NAI has been found in adult populations exposed to repeated malaria infections. NAI does not sterilize the adult individual from malaria infections but prevents the development of severe clinical symptoms. This is an example of premunition. In premunition, the low, persistent parasitemia provides antigens for development of antibodies. Children in their early years go from a high resistance to parasitemia at birth to later paroxysms of anemia and fevers. Children also have a higher risk for severe and cerebral malaria at ages 2 to 5 [39]. The Immunoglobulin G (IgG) are found in the first six months of life and is thought to confer resistance to high parasitemia [40]. IgG is one of the main immunoglobulins transferred from mother to fetus.

Previous studies have noted that some traits of NAI are specific at certain stages, develop after constant uninterrupted exposure, and can wane if there is an interruption in exposure. Development also depends on the degree of parasite exposure [41] [42] [43]. However, NAI acting in strain-specific or strain-transcending immunity is a trait of current interest.

It is common for one person to have multiple mosquito bites and be infected by heterologous *P. falciparum* strains. Anopheles mosquitos' ability to infect differs across *P. falciparum strains*. Populations with NAI develop immunity against heterologous *P. falciparum* strains depending on exposure. Therefore, it is of current interest to see if NAI acts as a strain-transcending and/or strain-specific immunity. Strain transcending immunity is a type of immune protection against several malaria strains, while strain-specific immunity is protection against one strain [44]. Learning the mechanism of NAI provides an opportunity to create a potential vaccine that works at the erythrocytic stage to prevent the development of clinical symptoms.

Barua et. al investigated NAI hypothesis within a population of Malawian children [45]. They followed 601 Malawian children from 6 – 18 months of age and measured antibody levels at 18 months. The specific antibodies measured were the ones against MSP1 19kD fragment, MSP2, (erythrocyte binding antigen 175) EBA175, variant surface antigens (VSA-very diverse) expressed by *P. falciparum* parasites. and Rh2A9. These were all antigens that would be expressed during the blood stage of the *P. falciparum* cycle. They found that the children who had high antibody levels presented with parasitemia, with antibody levels being higher in children with documented clinical malaria episodes. Perhaps early infection after the age of 6 months may have contributed to an increase in antibody levels. This antibody production may be key in the development of NAI. Furthermore, the number of documented malaria episodes had no difference with measured-antibody levels.

Scorza et. al created an expression library to test antigens that could develop cross-protective immunity against different strains found in *Plasmodium c. adami* in mice [15]. They identified antigens with conserved T-cell and B-cell epitopes that produced antibodies inducing moderate strain-transcending protective immunity against *P. c. adami* DS and SK strains. Strain transcending protective immunity is a type of immune protection against all strains within one infecting malaria parasite species. Strain-transcending protective immunity may require us to find conserved regions T-cell and B-cell epitopes across all strains. This type of protective

immunity could be potentially generated through a multi-epitope vaccine that could counteract the problem of malaria parasite immune evasion.

Allele-Specific Immunity

There is antibody production against epitopes of antigenic variants in one malaria season. However, the following season, new antigenic variants occur, and the last population of antibodies and plasma cells are unable to confer immune protection. The antigenic variation across seasons allows the parasite to evade immune recognition. Previous studies have tried to investigate vaccine candidates that are targets of allele-specific immunity. Vaccine candidates are genes that are being targeted for vaccine development [46] [47]. Successful targets have been shown in adults with multiple malaria infections presenting with none or mild clinical symptoms [48] [49] . We can hypothesize that a conglomeration of antigens that are targets of allele-specific immunity may present a more effective immune protection across seasons. However, targets of allele-specific immunity or NAI in adults are understudied. Previous studies have attempted to do this with small datasets that cover one or two malaria seasons. Our study utilizes the MalariaGEN Pfk6 VCF dataset comprising over 7000 samples across 27 Asian and African countries, collected between 2001 and 2015. Our study will focus on specific cities within countries to identify potential targets of allele-specific immunity across known multiple vaccine candidates [50]. It is our hope that this research would guide the development of a multi-epitope vaccine.

Hypothesis and Objectives

Antibodies produced in one malaria season do not confer protective immunity in the next season due to antigenic variation found in exposed *P. falciparum* proteins. We hypothesize that within the MalariaGEN Pfk6 dataset, would identify targets of allele-specific immunity. There are three main objectives to investigate this hypothesis.

1) To identify signatures of selection pressure in coding sequences of vaccine candidate genes.
2) To measure the seasonal variation of haplotype frequencies in different malaria transmission sites.
3) To determine epitopes and their types found in regions under selection pressure through in silico methods (T-cell and B-cell prediction algorithms)

The study focuses on 10 leading vaccine candidates.

Gene	Chromosome	Start	End
AMA1	11	1292966	1296696
CSP	3	221112	223145
MSP1	9	1201305	1207576
MSP2	2	271576	274917
MSP3	10	1403807	1406234
Pfs230	2	369351	380156
Pfs25	10	1253373	1254412
Pfs28	10	1251151	1251807
Pfs48_45	13	1875452	1878087
RH5	4	1082005	1084464

Table 1: Genomic location of 10 leading vaccine candidate genes and genomic location in 3D7 *P. falciparum* strain.

Study Design and Methodology

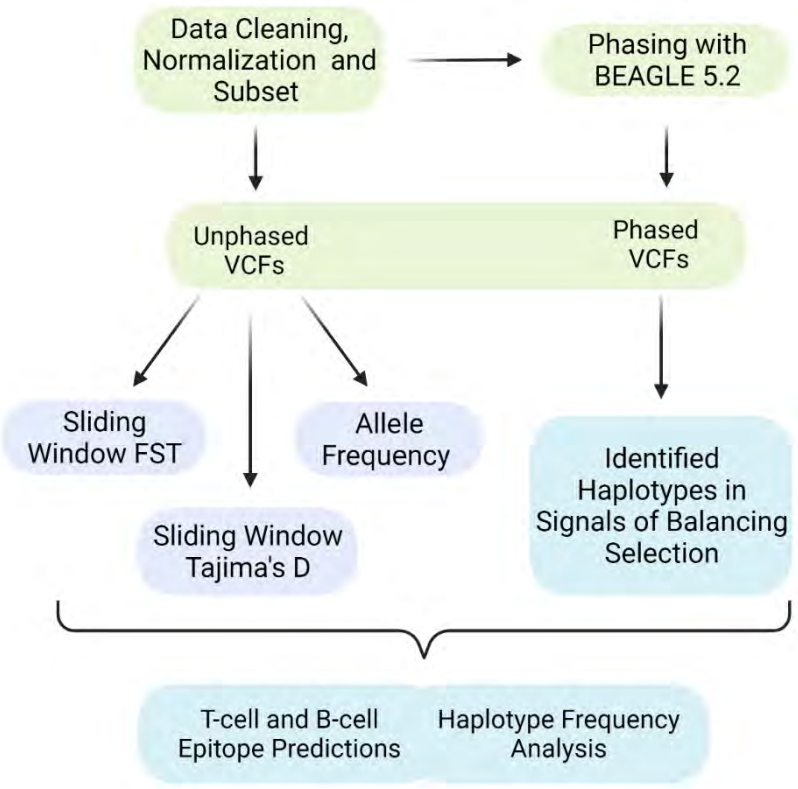

Figure 2: The flowchart for the data analysis pipeline of the Pf6k of variant call format (vcfs) files. Meta-analysis was done on the samples and sites were selected based on the representation of data across different years. Variants analyzed were from collected samples from one site in Cambodia, the Democratic Republic of Congo (DRC), and Ghana.

MalariaGEN v6.0 Data

The MalariaGEN Pfk6 dataset has 7113 samples that collected from 27 countries within the continents of Africa and Asia. The data was released in 2020 and has VCFs, gene differentiation estimates, mappings of drug resistance markers, drug resistance genotypes, and measures of complexity of infection through the consortium's pipeline. Within our analysis we decided to focus on the following vaccine candidates within *P. falciparum*: RH5, CSP, AMA1, MSP1, MSP2, MSP3, Pfs28, Pfs25, Pfs230, and Pfs48/45. Current clinical trials are developing

vaccines based on these candidates. The current field is interested in identification of T-cell and B-cell epitopes among the antigenic variants (See **Vaccine Development**). We looked at the variation within the following geographical sites available within the data: 1) Kinshasa, Democratic Republic of Congo 2) Pursat, Cambodia 3) Navrongo, Ghana **(Table 2)**. We picked sites that overall had well-spread sample data across the years. We compared time populations in each geographical site **(Figure 3)**.

Figure 3: This is a simplistic representation of the time populations that will be compared in downstream population genetics analysis. Each circle represents a VCF file covering gene variants of *P. falciparum* samples found within that year.

Country	Site (Sample Size)	Years	Transmission	Area Context	Rainy Season
Cambodia	Pursat (511)	2010 - 2014	High	Rural	June - Dec
DRC	Kinshasa (366)	2012-2015	Variable	Urban and Rural	October – May
Ghana	Navrongo (782)	2009-2013, 2015	High	Rural	July – Sept

Table 2: Summary of VCF data and malaria transmission information for the geographical sites are indicated within the table.

Generation of Unphased and Phased VCFs

bcftools v1.9 and Beagle v5.2 were for data VCF normalization and phasing, respectively. We generated VCF file for each vaccine candidate in each geographical site. The

VCF files of the vaccine candidates were normalized and phased using the 3D7 *P. falciparum* reference genome.

Sliding Window FST Analysis and FST Peak Identification

vcftools v0.1.16 is a bioinformatic command line package that handles population genetics statistics. Sliding Window FST command was utilized to compare the time populations for each candidate gene. The window size was 10 nucleotides, and the sliding window was 5 nucleotides. The output was a table of weighted FST values for each window. The genome of *P. falciparum* generally had negative FST values. We want to identify regions with high positive FST values. This allele-frequency-based technique can indicate how different sample populations are from each other throughout the entire window of the vaccine candidate. Higher FST values in regions of the genes indicate the selection process of variants across the different years. This was then graphed through the ggplot2 R package. Peaks that were found in the coding sequence region are shaded in gray.

Sliding Window Tajima's D Analysis and Tajima's D Peak Identification

VCFKit v0.2.9 tool was used to calculate the Tajima's D across the years for each vaccine candidate gene. We used as window size of 20 nucleotides and a sliding step of 5 nucleotides. Signals of balancing selection within specific regions can be identified through this method. This was run through the VCF file of a gene with all-time populations for each site. Output text file with calculated Tajima's D values for each geographical site was graphed with ggplot2 R package.

T-cell Epitope Prediction with NetMHC4.1pan and NetMHCII4.0pan

Identification of epitopes within vaccine candidate genes and comparison of epitopes to signals of balancing selection may help in our understanding of where allele-specific immunity targets are located or where changes are located within the vaccine candidates. NetMHCpan and NetMHCIIpan are bioinformatic tools developed for predicting T-cell epitopes given amino acid sequences. Trained on IEDB data and HLA alleles, the neural networks predict epitopes across different HLA Class I and Class II alleles. From the PlasmoDB database, amino acids sequences of the vaccine candidates were collected for 3D7 strain. Each strain-specific fasta sequence was run through NetMHCIIpan 4.0 and NetMHCpan4.1 to predict binding affinities for each amino acid within the sequences. Binding Affinity Rank for each amino acid sequence were then graphed. The binding affinity rank was -log() transformed and plotted with ggplot2 R package.

Bioinformatic and Visualization Tools

Analysis	Tools (Version)
Allele Frequency Calculations per Variant	bcftools (1.9)
Sliding Window FST Analysis	vcftools (v0.1.16), bcftools (1.9)
Sliding Window Tajima's D Analysis	VCFKit(v0.2.9), bcftools (1.9)
Haplotype Frequency Analysis	PEGAS (v1.1)
Visualizations in R	ggplot2, ggpubr, ggsci, gridExtra
T-cell Epitope Predictions	NetMHCpan4.1 (v4.1), netMHCIIpan4.0 (v4.0)

Table 3: List of bioinformatics tools used in data wrangling of VCFs, population statistical analyses, and in silico predictions.

Results

Our data set across the three different geographical sites had samples spanning consecutive years (Table 2). We investigated the frequency and types of mutations found across the vaccine candidate genes throughout the time populations. We found indels, synonymous SNPs, non-synonymous SNPs (missense mutations), and intergenic SNPs across each vaccine candidate (Appendix: Figure 6). We found there is a bias for non-synonymous SNPs compared to synonymous SNPs within each vaccine candidate. Furthermore, AMA1, MSP1, and Pfs230 had qualitatively more non-synonymous SNPs. Common alleles (allele frequency > 5%) were maintained within the population across all vaccine candidates. The next step was to see if common non-synonymous alleles were contributing to genetic differentiation between the different time populations.

In the sliding window FST, we identified peaks within both coding sequence and non-coding regions of all vaccine candidate. Within Kinshasa, DRC, we found peaks in MSP1, CSP, Pfs25, Pfs28 and Pfs48_45. Within Pursat, Cambodia in all genes except Pfs48_45. Within Navrongo, Ghana, we found peaks in all genes except AMA1. Some of peaks were only found in one geographical site, and not the others.

CSP had FST peaks found within NANP repeat regions but had Tajima's D peaks found within the N-terminal region. FST peaks found within the NANP may be due to the insertions and deletions, which are common variants found there, which our allele-frequency plots also confirm. AMA1 has peaks identified in Cambodia's geographical site, as those peaks. Pfs25 and Pfs28 are the smallest genes on our candidate's list and were found to have signals across genes. Overall, the common non-synonymous alleles overlap with Tajima's D Peaks in each vaccine candidate. For example, MSP 2 has identified a peak within its coding region near 274800bp, and there is a non-synonymous mutation found near that position according to our Allele-Frequency plots. FST plots and Tajima's D peaks identified are relatively the same (Table 4).

Regions with positive Tajima's D are evidence of balancing selection. Missense mutations that are found within these regions would potentially contribute to the generation of antigen variants. To confirm this, we identified amino acid haplotypes and their frequency within identified positive peaks of Tajima's D. We calculated the frequencies of each haplotype across the time populations within each geographical site. Common haplotypes that achieved at least 25% frequency in at least one of the years are shown in Figure 7a-c. These haplotypes change in frequency across the years. MSP3, CSP, and AMA1 had significant haplotype frequencies found in Navrongo, Ghana, a high malaria transmission area (Appendix: Figure 7a). MSP3 had significant haplotype frequencies across the years in Pursat, Cambodia, which is a low transmission setting (Appendix: Figure 7b). RH5 had significant haplotype frequency within the in Kinshasa, DRC, which has variable malaria transmission (Appendix: Figure 7c).

Common haplotypes identified from Tajima's D Peaks were evaluated for being potential epitopes. NetMHC4.1pan are T-cell prediction algorithms' neural networks trained on curated epitopes (IEDB) to identify epitopes and evaluate binding affinity and ligand elution towards HLA alleles. HLA supertypes A and B were selected for T-cell epitope prediction. A script to

call NETMHC4.1pan was created to run epitope predictions on the 3D7 strain amino acid sequence of the vaccine candidates. 3D7 strain is a common malaria strain found in homologous and heterologous infections across countries within Africa. Amino acid positions with -log(Binding Affinity Rank) > 0.3 is predicted to have strong binding epitopes for specified HLA alleles (Appendix: Figure 8). We also see that within a vaccine candidate, strength of epitope regions can change depending on HLA supertype.

Vaccine Candidate	# Of FST Peaks Identified	# Of Tajima's D Peaks Identified
CSP	22	18
AMA1	52	84
MSP1	111	83
MSP2	25	5
MSP3	20	38
Pfs25	6	5
Pfs28	12	4
Pfs230	52	66
Pfs48/45	9	6
RH5	6	5

Table 4: Summary of Tajima's D and FST Peaks identified for each candidate gene across Geographical Sites.

Discussion

We completed a sliding window FST analysis to determine what regions contribute to the genetic differentiation between the time populations. We that that measuring the genetic differentiation across time populations would be an indirect measurement of selection pressure. It was unclear whether the common non-synonymous alleles contributed to genetic differentiation. FST analysis did not show the contribution of non-synonymous alleles to the genetic differentiation across time populations. We believe the indel mutations contributed more differentiation compared to SNPs and showed higher FST values in those regions. FST analysis is not specific in indicating whether there is balancing selection or directional selection as shown with the allele-frequency graphs.

We calculated sliding window Tajima's D across the whole nucleotide sequence of the vaccine candidates. Tajima's D is a direct measurement of selection pressure. Positive Tajima's D peaks indicate balancing selection and negative Tajima's D peaks indicate directional selection (either positive or negative). There were peaks identified within Tajima's D peaks for all vaccine candidates. To investigate the contribution of non-synonymous SNPs in generation of antigenic variants, we looked at amino acid haplotypes derived from the variant. Interestingly, the haplotype frequency was insignificant between years in some vaccine genes. Geographical different transmission settings may affect what significant haplotype frequencies could be found. For significant haplotypes had little change in haplotype frequencies across the years. This means that this haplotype is being conserved throughout the balancing selection pressure. Significant haplotypes that do not show drastic differences in frequency across the years, may hold conserved epitopes that can targets of allele-specific immunity.

Investigating signals of balancing selection can elucidate significant alleles that contribute to different haplotypes. Immunological assays to identify antigens and antibodies are costly in both economics and time. Within this study we show that utilizing in silico prediction algorithms on large sampled, and longitudinal datasets may elucidate areas of genes with strong epitopes that can be targets for allele-specific immunity. In addition, we can compare the strength of the epitopes across different HLA allele supertypes curated or HLA-typed from lab with NETMHC4.0pan and NETMHC4.1pan tools. One future direction is to map the haplotype found in these signals and perform in silico prediction of epitopes on vaccine candidate protein sequence.

Appendix

Allele Frequency Analysis of Vaccine Candidates

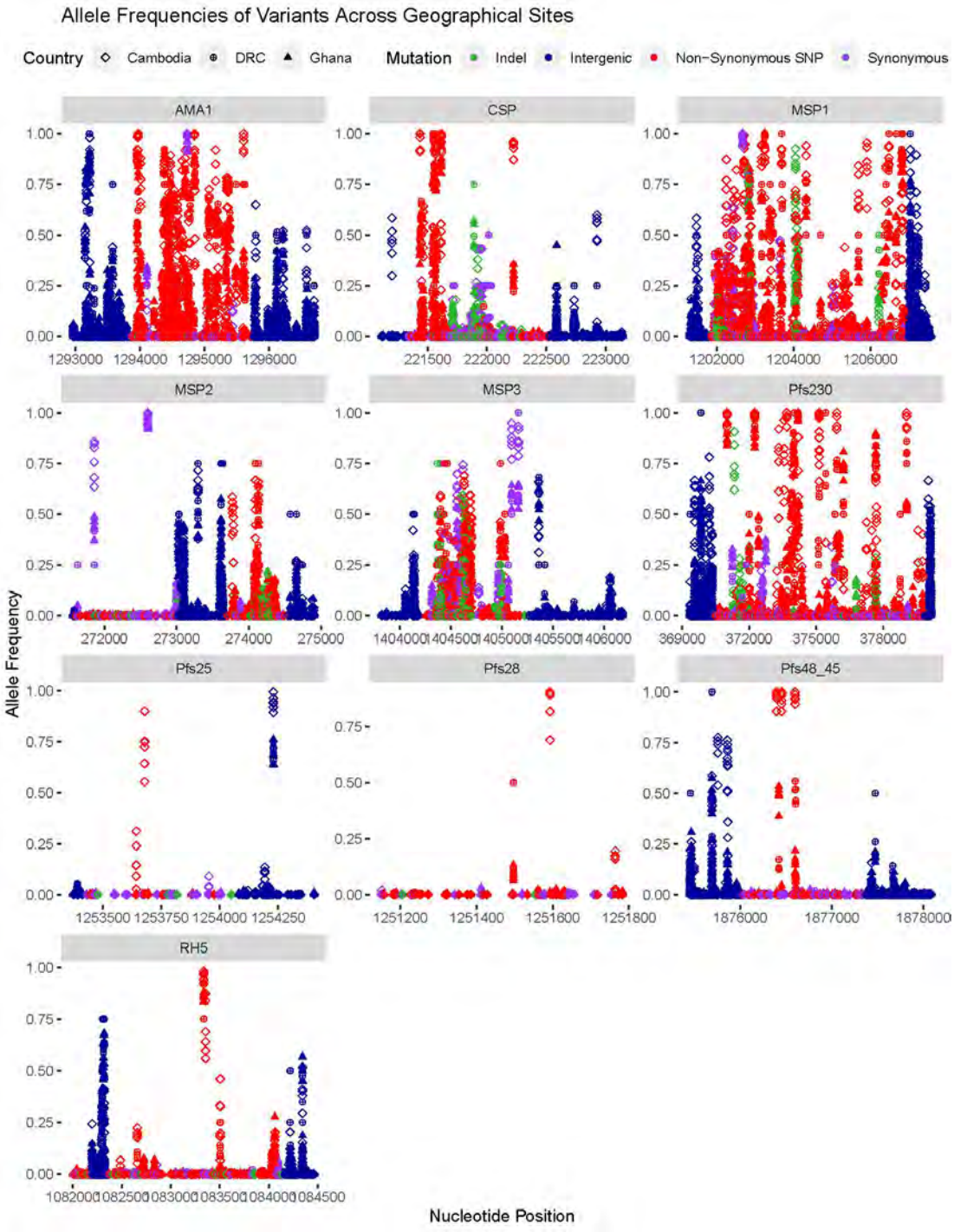

Figure 5: Allele frequencies across different time populations for each geographical site are plotted. Common and rare specific types of alleles are shown across the vaccine candidate genes.

Sliding Window FST Analysis of Vaccine Candidates

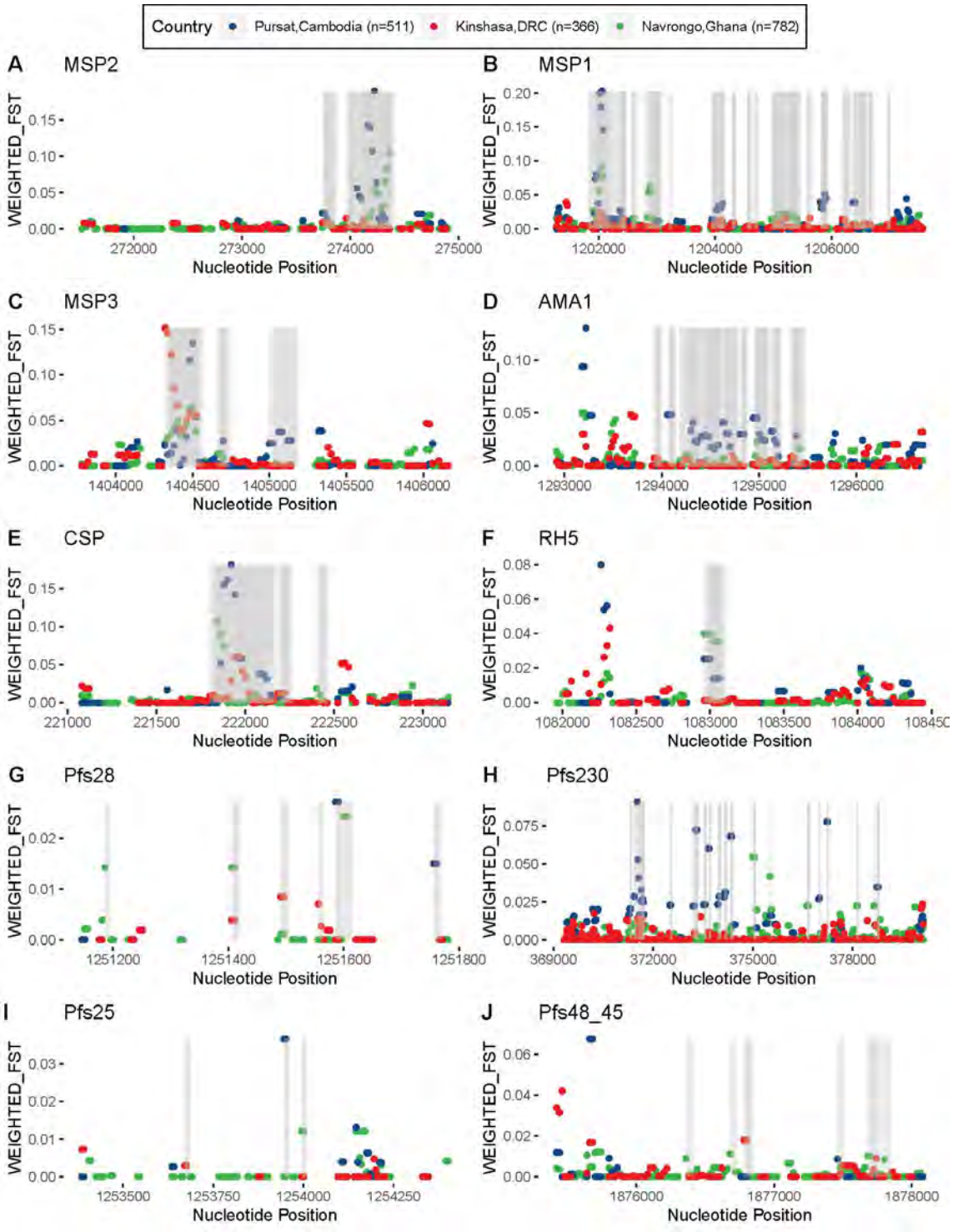

Figure 6: Calculated FST values for specific windows of 10 nucleotides and a step size of 5 nucleotides are shown. Peaks of FST values were determined by two main criteria 1) the Peak is found in the coding sequence region and 2) The peak is greater than neighboring FST values. within geographical site of itself.

Sliding Window Tajima's D Analysis of Vaccine Candidates

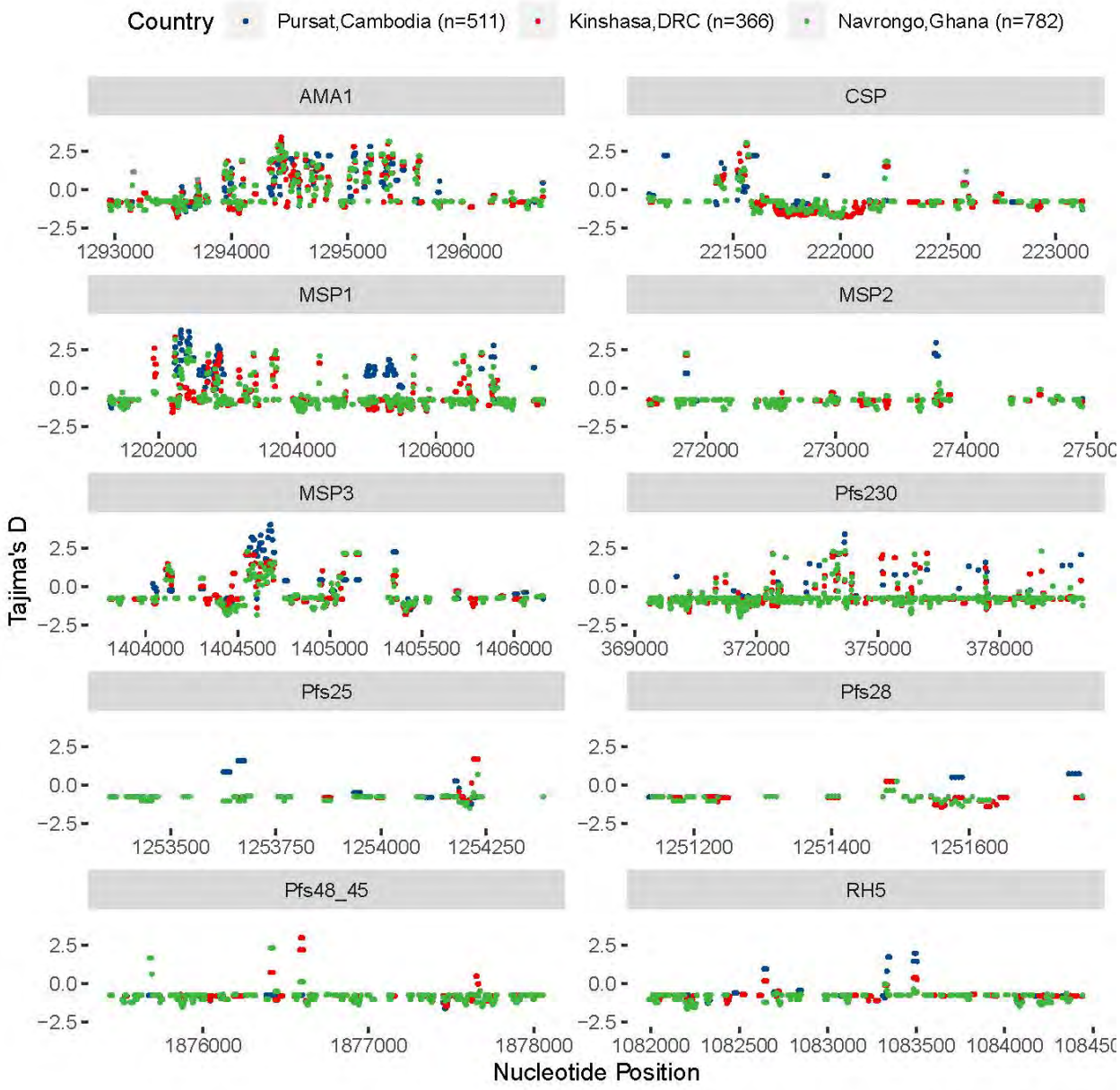

Figure 7: Tajima's D was calculated for a window size of 20 nucleotides and a step size of 5 base pairs using a sliding window approach. Tajima's D peaks were determined by two main criteria 1) Peak is found in the coding sequence region and 2) The peak is greater than neighboring FST values. within geographical site of itself.

Haplotype Frequency Across Years of Vaccine Candidates

Figure 7a: The x-axis shows the frequency of common haplotypes across the years within Kinshasa, Democratic Republic of Congo (DRC). Haplotypes were found and compared within 2012, 2013, and 2014 years. Significance was calculated using the Kruskal-Wallis test against each frequency and year. A haplotype is determined as common if one of the year's frequencies is found to be above >25% for at least one year.

Figure 7b: The x-axis shows the frequency of common haplotypes across the years within Navrongo, Ghana. Haplotypes were found and compared in 2010, 2011, 2012, and 2013 years. Significance was calculated using the Kruskal-Wallis test against each frequency and year. A haplotype is determined as common if one of the year's frequencies is found to be above >25% for at least one year.

Figure 7c: The x-axis shows the frequency of common haplotypes across the years within Navrongo, Ghana. Haplotypes were found and compared in 2009, 2010, 2011, 2012, and 2013 years. Significance was calculated using the Kruskal-Wallis test against each frequency and year. A haplotype is determined as common if one of the year's frequencies is found to be above >25% for at least one year.

T-cell Epitope prediction across HLA supertypes

Figure 8: This shows the predicted -log(Binding Affinity rank) across the amino acid position of 3D7 *P. falciparum* strain vaccine candidates. In addition, the predicted binding affinity rank (BA Rank) was calculated for each specific HLA supertype shown. NETMHC4.1pan algorithm was run to predict BA Rank across 8, 9, 10, and 11mer sliding windows to generate. This indicates specific regions with strong binding CD8+ T-cell epitopes, which activate cytotoxic T-cell-mediated response. The red dashed line at -log(2) is a threshold were strong binding amino acids. Predicted strong CD8+ T-cell epitopes are in regions above the threshold. In addition, Certain regions may not pose strong epitopes when against different HLA supertypes.

The figure is shown on the next page.

CSP
MSP1
MSP2
MSP3
Pfs230
Pfs25
Pfs28
Pfs48_45
RH5
HLA-A01:01
HLA-A02:01
HLA-A03:01
HLA-A24:02
HLA-A26:01
HLA-B07:02
HLA-B08:01
HLA-B15:01
HLA-B27:05
HLA-B39:01
HLA-B46:01
-Log(BA_rank)

References

1. World malaria report 2021. Accessed April 16, 2022. https://www.who.int/publications-detail-redirect/9789240040496

2. Ashley EA, Phyo AP, Woodrow CJ. Malaria. *The Lancet*. 2018;391(10130):1608-1621. doi:10.1016/S0140-6736(18)30324-6

3. Chhim S, Piola P, Housen T, Herbreteau V, Tol B. Malaria in Cambodia: A Retrospective Analysis of a Changing Epidemiology 2006–2019. *Int J Environ Res Public Health*. 2021;18(4):1960. doi:10.3390/ijerph18041960

4. CDC - DPDx - Malaria. Published October 26, 2020. Accessed April 17, 2022. https://www.cdc.gov/dpdx/malaria/index.html

5. Van Damme-Ostapowicz K, Krajewska-Kułak E, Rozwadowska E, Nahorski WL, Olszański R. Quality of life and satisfaction with life of malaria patients in context of acceptance of the disease: quantitative studies. *Malar J*. 2012;11:171. doi:10.1186/1475-2875-11-171

6. Talapko J, Škrlec I, Alebić T, Jukić M, Včev A. Malaria: The Past and the Present. *Microorganisms*. 2019;7(6):179. doi:10.3390/microorganisms7060179

7. Snow RW. Global malaria eradication and the importance of Plasmodium falciparum epidemiology in Africa. *BMC Med*. 2015;13:23. doi:10.1186/s12916-014-0254-7

8. Garcia LS. Malaria. *Clin Lab Med*. 2010;30(1):93-129. doi:10.1016/j.cll.2009.10.001

9. Amos B, Aurrecoechea C, Barba M, et al. VEuPathDB: the eukaryotic pathogen, vector and host bioinformatics resource center. *Nucleic Acids Res*. 2022;50(D1):D898-D911. doi:10.1093/nar/gkab929

10. Oduro AR, Wak G, Azongo D, et al. Profile of the Navrongo Health and Demographic Surveillance System. *Int J Epidemiol*. 2012;41(4):968-976. doi:10.1093/ije/dys111

11. Paton RS, Kamau A, Akech S, et al. Malaria infection and severe disease risks in Africa. *Science*. 2021;373(6557):926-931. doi:10.1126/science.abj0089

12. Diouf I, Rodriguez Fonseca B, Caminade C, et al. Climate Variability and Malaria over West Africa. *Am J Trop Med Hyg*. 2020;102(5):1037-1047. doi:10.4269/ajtmh.19-0062

13. Long CA, Zavala F. Immune Responses in Malaria. *Cold Spring Harb Perspect Med*. 2017;7(8):a025577. doi:10.1101/cshperspect.a025577

14. Doolan DL, Dobaño C, Baird JK. Acquired Immunity to Malaria. *Clin Microbiol Rev*. 2009;22(1):13-36. doi:10.1128/CMR.00025-08

15. Scorza T, Grubb K, Smooker P, Rainczuk A, Proll D, Spithill TW. Induction of Strain-Transcending Immunity against Plasmodium chabaudi adami Malaria with a Multiepitope DNA Vaccine. *Infect Immun*. 2005;73(5):2974-2985. doi:10.1128/IAI.73.5.2974-2985.2005

16. Osier FHA, Weedall GD, Verra F, et al. Allelic Diversity and Naturally Acquired Allele-Specific Antibody Responses to Plasmodium falciparum Apical Membrane Antigen 1 in Kenya. *Infect Immun*. 2010;78(11):4625-4633. doi:10.1128/IAI.00576-10

17. Thera MA, Doumbo OK, Coulibaly D, et al. Safety and Allele-Specific Immunogenicity of a Malaria Vaccine in Malian Adults: Results of a Phase I Randomized Trial. *PLOS Clin Trials*. 2006;1(7):e34. doi:10.1371/journal.pctr.0010034

18. Shah Z, Naung MT, Moser KA, et al. Whole-genome analysis of Malawian Plasmodium falciparum isolates identifies possible targets of allele-specific immunity to clinical malaria. *PLOS Genet*. 2021;17(5):e1009576. doi:10.1371/journal.pgen.1009576

19. Abuga KM, Jones-Warner W, Hafalla JCR. Immune responses to malaria pre-erythrocytic stages: Implications for vaccine development. *Parasite Immunol*. 2021;43(2):e12795. doi:10.1111/pim.12795

20. Doolan DL, Martinez-Alier N. Immune response to pre-erythrocytic stages of malaria parasites. *Curr Mol Med*. 2006;6(2):169-185. doi:10.2174/156652406776055249

21. Bucşan AN, Williamson KC. Setting the stage: The initial immune response to blood-stage parasites. *Virulence*. 2020;11(1):88-103. doi:10.1080/21505594.2019.1708053

22. Hafalla JCR, Bauza K, Friesen J, Gonzalez-Aseguinolaza G, Hill AVS, Matuschewski K. Identification of Targets of CD8+ T Cell Responses to Malaria Liver Stages by Genome-wide Epitope Profiling. *PLOS Pathog*. 2013;9(5):e1003303. doi:10.1371/journal.ppat.1003303

23. Schmidt NW, Podyminogin RL, Butler NS, et al. Memory CD8 T cell responses exceeding a large but definable threshold provide long-term immunity to malaria. *Proc Natl Acad Sci U S A*. 2008;105(37):14017-14022. doi:10.1073/pnas.0805452105

24. Tandel N, Dalai SK. *T Cell-Based Vaccines: Hope for Malaria Elimination*. IntechOpen; 2021. doi:10.5772/intechopen.96767

25. Chan JA, Fowkes FJI, Beeson JG. Surface antigens of Plasmodium falciparum-infected erythrocytes as immune targets and malaria vaccine candidates. *Cell Mol Life Sci*. 2014;71(19):3633-3657. doi:10.1007/s00018-014-1614-3

26. Ntege EH, Takashima E, Morita M, Nagaoka H, Ishino T, Tsuboi T. Blood-stage malaria vaccines: post-genome strategies for the identification of novel vaccine candidates. *Expert Rev Vaccines*. 2017;16(8):769-779. doi:10.1080/14760584.2017.1341317

27. Mugo RM, Mwai K, Mwacharo J, et al. Seven-year kinetics of RTS, S/AS01-induced anti-CSP antibodies in young Kenyan children. *Malar J*. 2021;20(1):452. doi:10.1186/s12936-021-03961-2

28. Chatterjee D, Cockburn IA. The challenges of a circumsporozoite protein-based malaria vaccine. *Expert Rev Vaccines*. 2021;20(2):113-125. doi:10.1080/14760584.2021.1874924

29. Laurens MB. The Promise of a Malaria Vaccine-Are We Closer? *Annu Rev Microbiol*. 2018;72:273-292. doi:10.1146/annurev-micro-090817-062427

30. Artaud C, Kara L, Launay O. Vaccine Development: From Preclinical Studies to Phase 1/2 Clinical Trials. *Methods Mol Biol Clifton NJ*. 2019;2013:165-176. doi:10.1007/978-1-4939-9550-9_12

31. Duffy PE, Patrick Gorres J. Malaria vaccines since 2000: progress, priorities, products. *Npj Vaccines*. 2020;5(1):1-9. doi:10.1038/s41541-020-0196-3

32. Laurens MB. RTS,S/AS01 vaccine (Mosquirix^{TM}): an overview. *Hum Vaccines Immunother*. 2019;16(3):480-489. doi:10.1080/21645515.2019.1669415

33. Han L, Hudgens MG, Emch ME, et al. RTS,S/AS01 Malaria Vaccine Efficacy is Not Modified by Seasonal Precipitation: Results from a Phase 3 Randomized Controlled Trial in Malawi. *Sci Rep*. 2017;7:7200. doi:10.1038/s41598-017-07533-w

34. Fowkes FJI, Richards JS, Simpson JA, Beeson JG. The Relationship between Anti-merozoite Antibodies and Incidence of Plasmodium falciparum Malaria: A Systematic Review and Meta-analysis. *PLOS Med*. 2010;7(1):e1000218. doi:10.1371/journal.pmed.1000218

35. Garcia-Senosiain A, Kana IH, Singh SK, et al. Peripheral Merozoite Surface Proteins Are Targets of Naturally Acquired Immunity against Malaria in both India and Ghana. *Infect Immun*. 2020;88(4):e00778-19. doi:10.1128/IAI.00778-19

36. Turner L, Wang CW, Lavstsen T, et al. Antibodies against PfEMP1, RIFIN, MSP3 and GLURP Are Acquired during Controlled Plasmodium falciparum Malaria Infections in Naïve Volunteers. *PLoS ONE*. 2011;6(12):e29025. doi:10.1371/journal.pone.0029025

37. Chaturvedi N, Bharti PK, Tiwari A, Singh N. Strategies & recent development of transmission-blocking vaccines against Plasmodium falciparum. *Indian J Med Res*. 2016;143(6):696-711. doi:10.4103/0971-5916.191927

38. Naung MT, Martin E, Munro J, et al. Global diversity and balancing selection of 23 leading Plasmodium falciparum candidate vaccine antigens. *PLOS Comput Biol*. 2022;18(2):e1009801. doi:10.1371/journal.pcbi.1009801

39. Wamae K, Wambua J, Nyangweso G, et al. Transmission and Age Impact the Risk of Developing Febrile Malaria in Children with Asymptomatic Plasmodium falciparum Parasitemia. *J Infect Dis*. 2019;219(6):936-944. doi:10.1093/infdis/jiy591

40. Ndiabamoh CM cho, Ekali GL, Esemu L, et al. The immunoglobulin G antibody response to malaria merozoite antigens in asymptomatic children co-infected with malaria and intestinal parasites. *PLoS ONE*. 2020;15(11):e0242012. doi:10.1371/journal.pone.0242012

41. Dechavanne C, Sadissou I, Bouraima A, et al. Acquisition of natural humoral immunity to P. falciparum in early life in Benin: impact of clinical, environmental and host factors. *Sci Rep*. 2016;6(1):33961. doi:10.1038/srep33961

42. Hviid L. Naturally acquired immunity to Plasmodium falciparum malaria in Africa. *Acta Trop*. 2005;95(3):270-275. doi:10.1016/j.actatropica.2005.06.012

43. Pinkevych M, Petravic J, Chelimo K, Kazura JW, Moormann AM, Davenport MP. The Dynamics of Naturally Acquired Immunity to Plasmodium falciparum Infection. *PLOS Comput Biol*. 2012;8(10):e1002729. doi:10.1371/journal.pcbi.1002729

44. Brown IN, Brown KN, Hills LA. Immunity to malaria: the antibody response to antigenic variation by Plasmodium knowlesi. *Immunology*. 1968;14(1):127-138.

45. Barua P, Beeson JG, Maleta K, Ashorn P, Rogerson SJ. The impact of early life exposure to Plasmodium falciparum on the development of naturally acquired immunity to malaria in young Malawian children. *Malar J*. 2019;18(1):11. doi:10.1186/s12936-019-2647-8

46. Takala SL, Plowe CV. Genetic diversity and malaria vaccine design, testing, and efficacy: Preventing and overcoming "vaccine resistant malaria." *Parasite Immunol*. 2009;31(9):560-573. doi:10.1111/j.1365-3024.2009.01138.x

47. Ajibaye O, Osuntoki AA, Balogun EO, et al. Genetic polymorphisms in malaria vaccine candidate Plasmodium falciparum reticulocyte-binding protein homologue-5 among populations in Lagos, Nigeria. *Malar J*. 2020;19(1):6. doi:10.1186/s12936-019-3096-0

48. Barry AE, Arnott A. Strategies for Designing and Monitoring Malaria Vaccines Targeting Diverse Antigens. *Front Immunol*. 2014;5. Accessed April 17, 2022. https://www.frontiersin.org/article/10.3389/fimmu.2014.00359

49. Bergmann-Leitner ES, Duncan EH, Mease RM, Angov E. Impact of pre-existing MSP142-allele specific immunity on potency of an erythrocytic Plasmodium falciparum vaccine. *Malar J*. 2012;11(1):315. doi:10.1186/1475-2875-11-315

50. An open dataset of Plasmodium falciparum... | Wellcome Open Research. Accessed April 17, 2022. https://wellcomeopenresearch.org/articles/6-42