PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Global diversity and balancing selection of 23 leading *Plasmodium falciparum* candidate vaccine antigens

Myo T. Naung[1,2,3], Elijah Martin[1,4], Jacob Munro[1], Somya Mehra[1,4], Andrew J. Guy[5], Moses Laman[6], G. L. Abby Harrison[1,2], Livingstone Tavul[6], Manuel Hetzel[7,8], Dominic Kwiatkowski[9,10‡], Ivo Mueller[1,2,11], Melanie Bahlo[1,2], Alyssa E. Barry[1,2,3,4]*

1 Population Health and Immunity Division, Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria, Australia, **2** Department of Medical Biology, University of Melbourne, Carlton, Victoria, Australia, **3** Institute of Mental and Physical Health and Clinical Translation (IMPACT), School of Medicine, Deakin University, Geelong, Victoria, Australia, **4** Life Sciences Discipline, Burnet Institute, Melbourne, Victoria, Australia, **5** School of Science, RMIT University, Melbourne, Victoria, Australia, **6** Vector Borne Diseases Unit, Papua New Guinea Institute of Medical Research, Madang, Papua New Guinea, **7** Swiss Tropical Public Health Institute, Basel, Switzerland, **8** University of Basel, Basel, Switzerland, **9** Sanger Institute, Hinxton, United Kingdom, **10** Big Data Institute, University of Oxford, Oxford, United Kingdom, **11** Division of Parasites and Insect Vectors, Pasteur Institute, Paris, France

‡ On behalf of the MalariaGEN *P. falciparum* Community Project.
* a.barry@deakin.edu.au

Citation: Naung MT, Martin E, Munro J, Mehra S, Guy AJ, Laman M, et al. (2022) Global diversity and balancing selection of 23 leading *Plasmodium falciparum* candidate vaccine antigens. PLoS Comput Biol 18(2): e1009801. https://doi.org/10.1371/journal.pcbi.1009801

Editor: Anders Wallqvist, US Army Medical Research and Materiel Command: US Army Medical Research and Development Command, UNITED STATES

Received: July 4, 2021

Accepted: January 3, 2022

Published: February 2, 2022

Data Availability Statement: All relevant data and code are cited within the manuscript and its Supporting Information files.

Funding: This research was funded by a Project Grant (GNT1161066) from the National Health and Medical Research Council (NHMRC) of Australia. Samples were collected with support from the Asia Pacific International Centre of Excellence in Malaria Research Program funded by the National

Abstract

Investigation of the diversity of malaria parasite antigens can help prioritize and validate them as vaccine candidates and identify the most common variants for inclusion in vaccine formulations. Studies of vaccine candidates of the most virulent human malaria parasite, *Plasmodium falciparum*, have focused on a handful of well-known antigens, while several others have never been studied. Here we examine the global diversity and population structure of leading vaccine candidate antigens of *P. falciparum* using the MalariaGEN Pf3K (version 5.1) resource, comprising more than 2600 genomes from 15 malaria endemic countries. A stringent variant calling pipeline was used to extract high quality antigen gene 'haplotypes' from the global dataset and a new R-package named *VaxPack* was used to streamline population genetic analyses. In addition, a newly developed algorithm that enables spatial averaging of selection pressure on 3D protein structures was applied to the dataset. We analysed the genes encoding 23 leading and novel candidate malaria vaccine antigens including *csp*, *trap*, *eba175*, *ama1*, *rh5*, and *CelTOS*. Our analysis shows that current malaria vaccine formulations are based on rare haplotypes and thus may have limited efficacy against natural parasite populations. High levels of diversity with evidence of balancing selection was detected for most of the erythrocytic and pre-erythrocytic antigens. Measures of natural selection were then mapped to 3D protein structures to predict targets of functional antibodies. For some antigens, geographical variation in the intensity and distribution of these signals on the 3D structure suggests adaptation to different human host or mosquito vector populations. This study provides an essential framework for the diversity of

Institutes of Health, USA (U19AI129392). MB and IM were supported by NHMRC Research Fellowships (GNT1102971 and GNT1155075). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

P. falciparum antigens to be considered in the design of the next generation of malaria vaccines.

Author summary

Highly effective malaria vaccines are important for the sustainable elimination of malaria. However, the diversity of parasite antigens targeted by malaria vaccines has been largely overlooked, with most vaccine formulations based only on a single antigen variant. Failure to accommodate this diversity may result in vaccines only being effective against vaccine-like variants, resulting in limited protective efficacy. Investigation of the diversity of genes encoding parasite antigens can help prioritize and validate them as vaccine candidates as well as to identify the most common variants for inclusion in the next generation of malaria vaccines. Here we measure the diversity of 23 vaccine antigens of *Plasmodium falciparum*, using the publicly available MalariaGEN Pf3K (version 5.1) resource comprising more than 2600 genomes from 15 malaria endemic countries. We find that variants found in current vaccine formulations are rare and thus may target only a small proportion of circulating malaria parasite strains. Variation in intensity of immune selection in parasites from different geographic areas suggests adaptation to different human host or vector populations. This study provides an essential framework for the design of the next generation of malaria vaccines, in addition to providing novel insights into malaria biology.

Introduction

Plasmodium falciparum, the most lethal human malaria parasite, has been co-evolving with its human host for tens of thousands of years [1]. Along this evolutionary timeline, host-parasite interactions have left landmarks of selection on both genomes [2,3]. These landmarks can help to identify parasite surface proteins targeted by host immune responses and therefore antigens that could be used in "subunit" vaccines to protect against infection and disease. However, malaria parasites evade host immune responses through the accumulation of mutations, thus increasing the repertoire of variants circulating in the parasite population, and therefore this diversity needs to be considered in vaccine formulations [4,5]. Otherwise, vaccines may only be partially effective resulting in a "sieve effect", with higher efficacy against vaccine-like strains and low to no efficacy against antigenically distinct strains [6,7,8]. Following vaccination, vaccine-distinct variants might become dominant, requiring new vaccine formulations to be developed. One approach to overcome parasite diversity is to include multiple variants currently circulating in the global population, as is the case for the influenza vaccine which is reformulated annually [9]. Malaria differs from the influenza virus however as malaria parasites have multiple lifecycle stages, hundreds of surface antigens that could be considered as vaccine candidates, and a lack of understanding of how parasite antigen diversity influences host immune responses. With the availability of thousands of malaria parasite genomes [10], characterisation of antigen diversity on a global scale is highly feasible and a crucial step in contemporary malaria vaccine development. A meta-population genetic analysis of antigens would provide a catalogue of common variants to be included in vaccine formulations and provide a basis for predicting vaccine effectiveness in different populations.

Malaria antigen diversity has been largely overlooked with most subunit vaccines based on a single antigen variant, predominantly those of the reference strain 3D7. As a result, very few

have shown significant protective efficacy in human clinical trials, and protection is mostly short-lived. Of the numerous vaccine candidates that have been tested in clinical trials, only RTS, S, has completed Phase III clinical trials. RTS,S is based on the C-terminal and NANP repeat region of the Circumsporozoite Protein (CSP) of the African strain 3D7 [6,11]. The limited efficacy of this vaccine [12] has been attributed to polymorphism of the target antigen, PfCSP, suggesting a diversity-covering approach could be more effective [6,7,13]. This is supported by a study in Africa where the one year cumulative vaccine efficacy against homologous strains was 50.3% compared to 33.4% against heterologous strains [4]. Furthermore, in a phase 2b trial in Papua New Guinean (PNG) children, the "Combination B" malaria vaccine (multi-antigen vaccine, composed of MSP2, RESA and a fragment of MSP1) showed 62% protective efficacy against vaccine (3D7) like strains and limited efficacy against vaccine dissimilar strains [8]. Of the vaccine trials that have been powered to measure allele specific outcomes (comparing infection rates of vaccine versus non-vaccine variants in vaccinees and controls), several have identified variant-specific efficacy (reviewed in [14]). Vaccine development would therefore benefit from a better understanding of the global diversity and distribution of antigen variants, to select common variants that could maximise vaccine efficacy.

The presence of intermediate variant frequencies for a given antigen in a parasite population indicates balancing selection, which is driven by protective immune responses [15–17]. Balancing selection can be measured using the *Tajima's D* statistic using a 'sliding window' approach to identify hotspots of selection, along the length of a gene [18]. However, standard Tajima's D analyses of linear gene sequences do not consider the spatial distribution of polymorphic residues (i.e. residues that are distant on the linear sequence may be proximal on the three-dimensional (3D) structure). Despite the availability of high quality whole genome sequence data from thousands of worldwide *P. falciparum* isolates, there has been no population genetic analysis of malaria vaccine candidate antigens using this data, and only a few studies have mapped selection hotspots onto experimentally-determined or modelled 3D protein structures [16,17,19]. In addition, only a few studies have analysed the diversity of existing or emerging *P. falciparum* vaccine candidate antigens and the majority have focused on few antigens and geographic areas.

In this study, we aimed to conduct a meta-population genetic analysis to characterise the global genetic diversity of 23 emerging and established *P. falciparum* vaccine candidates from different life cycle stages and sub-cellular localizations. Antigens were included if they were mentioned in the WHO Malaria Rainbow Table and in previous pre-clinical or clinical vaccine development studies [2,6,20]. They include antigens that have well-defined three dimensional (3D) structures as well as disordered proteins that are not well characterised [21]. The analysis was based on *P. falciparum* WGS from the Pf3K (version 5.1) global data resource [10] which comprises more than 2,600 WGS from 15 countries and 156 additional WGS from PNG [22,23] generated in collaboration with the MalariaGEN *P. falciparum* Community Project. These comprehensive population genetic analyses of malaria vaccine candidates will aid the development of more broadly effective vaccines.

Results

WGS data from a total of 2661 *P. falciparum* isolates from 16 countries from the Asia-Pacific and Africa were processed to obtain gene sequences for population genetic analysis. The pipeline filters the dataset to remove high complexity infections as determined by F_{ws} [24]. This step aimed to ensure the correct assignment of variants to a single major clone within the sample (S2 Fig). We found a large proportion (greater than 30%) of field isolates from African regions predicted to have more than two clones, especially in Malawi and Ghana, likely

reflecting high transmission intensity in those areas. However, the number of genomes in the final dataset (n = 1499) remained high, even after removing these samples. From these genomes, gene sequences for 23 antigens (Table 1) were obtained. This included a mean of 1374 (1079–1499) sequences per antigen, and 107 (26–433) sequences per country (Table 1, Table A in S1 Text). Nigeria was included in overall genetic diversity analyses but removed from further analyses due to very low sample size (n = 4).

High haplotype diversity and balancing selection indicates antigens that are natural targets of host immunity

Overall genetic diversity analyses based on Nei's nucleotide diversity (π), haplotype diversity and Tajima's D analyses indicated high proportions of SNPs that were non-synonymous (dN) and regions of balancing selection in most antigens [25] (Table A in S1 Text, Fig 1). In addition, the antigen subdomains, *csp*-C-terminal, *eba175* -RII and *pfs48/45* -6C were as diverse as the respective full-length gene (Fig 1, Table A in S1 Text). Different patterns of diversity across parasite populations (defined as all samples from a particular country) were observed for the antigens. For instance, *trap* had high haplotype diversity yet low to moderate nucleotide diversity indicating that many haplotypes were the result of very rare polymorphisms (Fig 1, Table A in S1 text). Whereas *csp* had high haplotype diversity and high (albeit variable amongst populations) nucleotide diversity, which might be related to differences in transmission intensity in different countries (Fig 1). In addition, *celtos* (a gametocyte antigen) also had high

Table 1. Vaccine Candidate Antigens. P. falciparum antigens and associated domains selected for the analysis were functionally important for malaria biology and key candidates of malaria vaccines trials.

Lifecycle Stage	Name	Gene ID	Genomic Length (bp)	Domains	3D Structure
Pre-Erythrocytic	*Csp*	PF3D7_0304600	1194	C-terminal	PDB Code: 3VDJ
	Ctrp	PF3D7_0315200	6345	Full-length	NA
	Trap	PF3D7_1335900	1725	Ectodomain	PDB Code: 4HQF.A
	Exp1	PF3D7_1121600	937	Full-length	NA
	Starp	PF3D7_0702300	1970	Full-length	NA
	Glurp	PF3D7_1035300	3702	Full-length	NA
	Trep	PF3D7_1442600	10632	Full-length	NA
Erythrocytic	*Ama1*	PF3D7_1133400	1869	I—III	Manually developed model [32]
	Ron2	PF3D7_1452000	6570	Full length	NA
	Eba175	PF3D7_0731500	4875	RII, RIII-V	Model based on 1ZRO template
	Rh5	PF3D7_0424100	1788	Full-length	PDB Code: 6MPV.B
	Ripr	PF3D7_0323400	3261	Full-length	NA
	Cyrpa	PF3D7_0423800	1188	Full-length	PDB Code: 6MPV.A
	Msp1	PF3D7_0930300	5163	*Msp1-19*	Model based on 1OB1 template
	Msp3	PF3D7_1035400	1065	Full-length	NA
	Msp4	PF3D7_0207000	963	Full-length	NA
	Msp6	PF3D7_1035500	1116	Full-length	NA
	Ralp1	PF3D7_0722200	2250	Full-length	NA
	Resa	PF3D7_0102200	3464	Full-length	NA
	Sera5	PF3D7_0207600	3435	C-terminal	PDB Code: 2WBF
	Sera8	PF3D7_0207300	2536	Full-length	NA
Gametocyte	*Tramp*	PF3D7_1218000	1059	Full-length	NA
	Pfs48/45	PF3D7_1346700	1347	Full-length, 6C	PDB Code: 6E62.A
	Celtos	PF3D7_1216600	549	Full-length	Model Based on 5TSZ template

https://doi.org/10.1371/journal.pcbi.1009801.t001

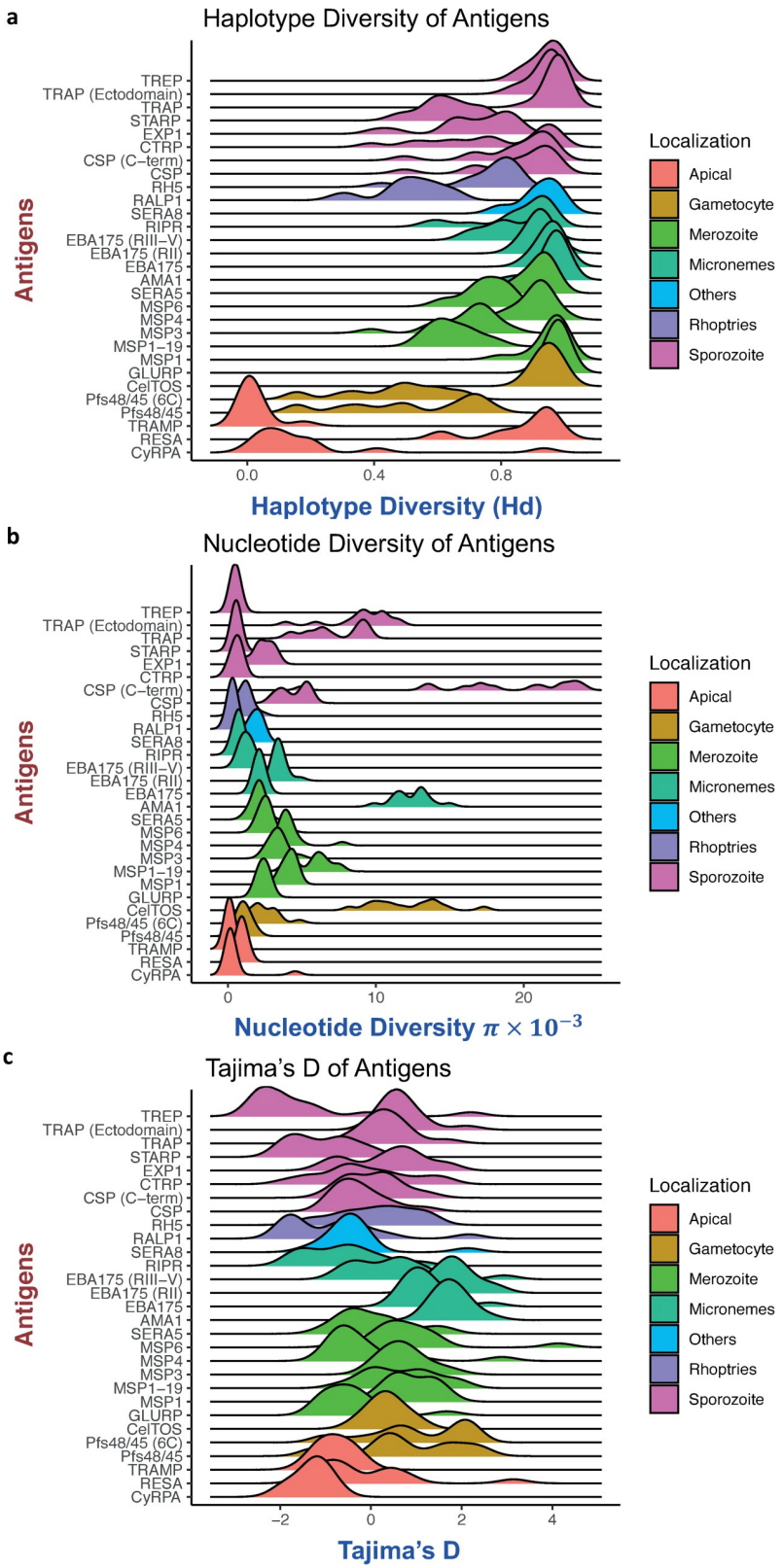

Fig 1. Distribution of haplotype diversity, nucleotide diversity, and Tajima's D values amongst countries for each antigen. The lines from ridgeline plots indicate the range and distribution of respective diversity parameters for each antigen across different parasite populations (countries). Tramp, and cyrpa were conserved across all populations,

whilst trap, ama1, eba175, and celtos were diverse and showed evidence of diversifying selection across all parasite populations.

https://doi.org/10.1371/journal.pcbi.1009801.g001

haplotype diversity but moderate to high (and variable) nucleotide diversity, which may indicate adaptation to mosquito vectors from different geographical regions (Fig 1, Table A in S1 Text). Moderate haplotype diversity, but low nucleotide diversity for *rh5* indicates lower diversity among distributed *rh5* haplotypes (Fig 1, Table A in S1 Text).

Genetic diversity and evidence of diversifying selection on the antigens was not related to their subcellular localization [26] (Fig 1). However, the diversity of some antigens (e.g. CSP) varies by geographic origin indicating that transmission intensity or other location-specific factors may play a role as previously shown in Barry *et al*. 2009 [15].

The use of full-length or full-domain *Tajima's D* values can mask variable patterns within a gene when there are discrete stretches of positive *D* values linked by regions with neutral or highly negative D values [27]. Further analyses were therefore performed on these antigens for full-length or their functionally important domains.

Antigens under balancing selection have large numbers of low frequency non-3D7 haplotypes

We identified the relationships amongst non-synonymous SNP haplotypes using network analysis for 17 of the 23 antigens. Six antigens were excluded because of lack of diversity. Many rare haplotypes are present within African populations, and relatively common haplotypes within Asia-Pacific countries, most likely owing to higher transmission in Africa. Generally, for antigens under balancing selection such as CelTOS, TRAP, AMA1, EBA175, MSP1 and GLURP (Fig 1), high haplotype diversity (Hd $>$ 0.97) was ubiquitous amongst countries with no predominant haplotype (Table B in S1 Text). The majority of these are blood stage antigens suggesting they are dominant targets of natural host immunity.

Haplotypes were often found in many countries (i.e. multiple colours for each node) suggesting minimal geographic variation (Fig 2). However, geographical clustering of haplotypes was observed for CTRP, TRAP (ectodomain), CSP (C-term), GLURP, EBA175 (RII), CelTOS, and Pfs48/45. Haplotypes from the same geographical region (e.g. Africa) were more closely related than those from different regions (e.g. Africa versus Asia, Fig 2). Most of these antigens are pre-erythrocytic or gametocyte antigens, consistent with previous observations [15] and may reflect adaptation of the parasite to local human and mosquito hosts.

Of all analysed antigens, the 3D7 haplotype was dominated only within CyRPA, MSP1-19, MSP3, MSP4, Pfs48/45, RALP1, RH5, RIPR, STARP, TRAMP, and TREP antigens (Fig 2, Tables A and B in S1 Text). For most antigens with multiple balancing selection hotspots such as AMA1, CSP (C-term), EBA175 (RII), CelTOS, the most common haplotypes were only remotely related to 3D7 vaccine haplotype (Table B in S1 Text). No 3D7 haplotypes were observed in the dataset for full length EBA175, MSP1 and SERA8 (Table A in S1 Text, Fig 2). This suggests that a 3D7-based malaria vaccine may be effective against only a small proportion of natural parasite populations.

Polymorphic resides are surface exposed

Relative solvent accessibility (RSA) analysis gives an indication of the regions of a protein that are exposed to the extracellular environment, and thus may be targeted by immune responses [16,28,29]. The RSA of all 23 antigens was compared for 792 polymorphic and 14,789 conserved residues. A higher RSA score suggests presence of a particular amino acid on the surface

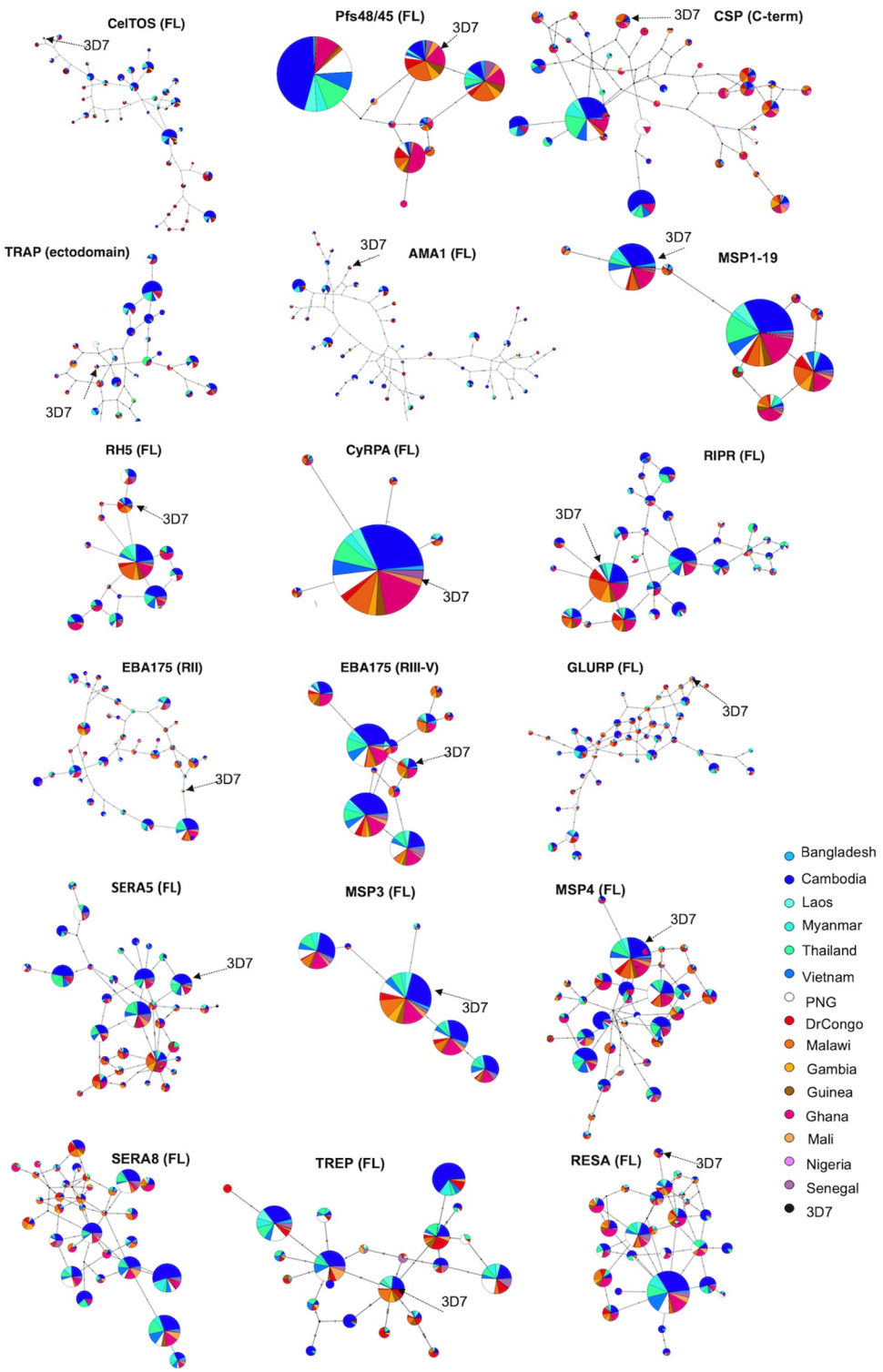

Fig 2. Haplotype network for malaria vaccine antigens. Templeton, Crandall, and Sing (TCS) network summarizing the global diversity of selected antigens using only common haplotypes (> 0.5% of all haplotypes) based on non-synonymous SNPs for full length respective domain of each antigen. Circles represent unique haplotypes, and circles are scaled according to the prevalence of the observed haplotypes. The number of non-synonymous SNP differences between each haplotype was shown by the number of hatch marks on the branches. The vaccine strain 3D7 (arrowed) was included for reference.

https://doi.org/10.1371/journal.pcbi.1009801.g002

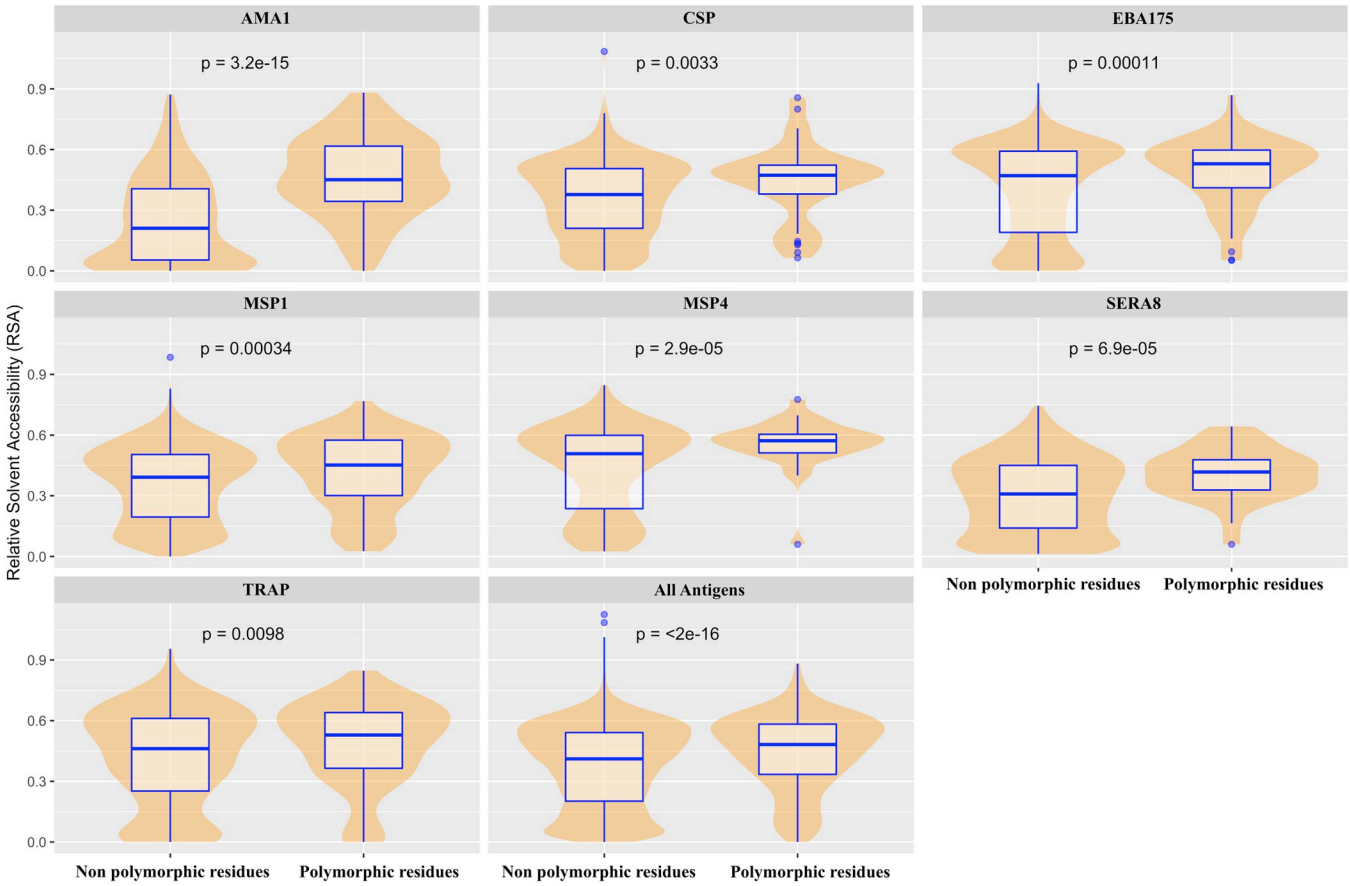

Fig 3. Relative solvent accessibility of polymorphic versus conserved residues. Relative solvent accessibility (RSA) was calculated for all residues for 23 antigens. RSA was calculated using neural network based NetSurfP1.1 program or DSSP program respectively based on the presence of known PDB or homology-modelled structures [30,31]. Polymorphic residues from more than 1000 sequences regardless of minor allele frequency (MAF) were included in the analysis. Box and whisker plots show the median (blue line), and interquartile range (blue box) of RSA values for each residue from respective group. The violin plot (which uses Kernel Density Estimation to compute an empirical probability distribution) shows a smooth distribution of RSA values for most of the calculated group. RSA scores for individual antigens as well as for the combination of all antigens were calculated. Only significant p-values are shown.

https://doi.org/10.1371/journal.pcbi.1009801.g003

and thus making it able to interact with the host environment. Overall, when combining all antigens, most of the polymorphic residues had significantly higher mean RSA scores of 0.46 than conserved residues with a mean RSA score of 0.38 (p value $< 2.0 * 10^{-16}$, t-test) (Fig 3). When individual antigens were analysed, only AMA1, CSP, EBA175, MSP1, MSP4, SERA8, and TRAP had significantly higher RSA scores than that of their residues without underlying polymorphisms ($p < 0.05$, t-test).

Polymorphic residues are predominantly found at functionally important interfaces

Shannon Entropy is a measure of the variability of individual amino acid residues within a protein. We calculated the normalized Shannon Entropy using all available sequences for each antigen that had an available 3D protein structure (Table 2) to investigate the potential functional impact of site-specific diversity. Except for CyRPA, SERA5 and MSP1-19, the normalized Shannon Entropy scores were high (>0.2) among the residues situated at host-receptor binding interface, within previously described immunological epitopes or under balancing selection (Fig 4, Table 2). For instance, residues situated at the surface exposed c1L loop of

Table 2. Residues with Shannon Entropy Scores greater than 0.20 for each calculated antigen.

Antigens	Amino Acid Residuesa*
TRAP (ectodomain)	83, 90, 92, 98, 134
CSP (C-terminal)	318, 322, 352, 357
AMA1	182, 187, 190, 196, 197, 200, 201, 204, 225, 230, 242, 243, 267, 283, 300, 308, 404, 405, 439, 451, 485, 496, 503, 512
EBA175 (RII)	274, 279, 286, 388, 390, 584, 664
RH5	197
CyRPA	NA
SERA5	NA
MSP1-19	NA
Pfs48/45 (6C domain)	304, 322
CelTOS	318, 322, 352, 357

*Residue numbers were based on 3D7 sequence

https://doi.org/10.1371/journal.pcbi.1009801.t002

AMA1, and residues situated at the dimerization interface of EBA175 RII had high normalized Shannon Entropy scores. Mutations at functionally important interfaces may enable parasites to escape from host immune responses.

Evidence of balancing selection at functionally important interfaces for TRAP, AMA1, RIPR for all geographic populations but identification of geographically variable balancing selection hotspots for EBA175 (RII), RH5, CSP, CelTOS, and MSP1-19

An ideal malaria vaccine should be immunogenic and effective against naturally circulating strains from worldwide populations, but it is not feasible to include all haplotypes for each antigen we have described above. This is especially the case for highly diverse antigens such as CSP, TRAP, AMA1, EBA175, MSP1 (Table A in S1 Text). Therefore, for each antigen, it is important to identify the most immunologically relevant polymorphisms and gene regions, that are targets of host immunity and could influence vaccine efficacy. Typically, the *Tajima's D* statistic has been calculated as a single metric encompassing the entire gene or domain of interest, or more recently, using a sliding window approach to identify hotspots of balancing selection along the length of a gene. Functional antibodies are critical for protective immunity from naturally acquired infection [6,32–37] however, more than 90% of functional antibody epitopes are discontinuous (non-linear) epitopes [38,39]. Thus, distant gene segments may be brought into proximity in three-dimensional (3D) protein structures. Consideration of structural features in 3D space are thus important, especially for antigens with highly surface exposed polymorphic residues such as AMA1, CSP, EBA175, and TRAP. Previous studies have suggested that polymorphic residues located on the surface of protein evolved to escape host immune responses [16,27,40]. We therefore examined selective pressure over the 3D protein structures of vaccine candidate antigens in the following analyses where possible.

A spatially derived approach to *Tajima's D* (D^*) was applied to available 3D structures (predicted or experimentally defined) for a subset of parasite populations covering all major endemic regions, including Southeast Asia, Africa, and Oceania (PNG). Structures included well-studied antigens like CSP (C-terminal), TRAP (ectodomain), AMA1, EBA175 (RII), MSP1, RH5, CyRPA, SERA5, CelTOS, and MSP1-19. We observed moderate to high D^* scores (1.0–2.0), high scores (2.1–3.0), and extremely high scores (> 3.1) within most of these

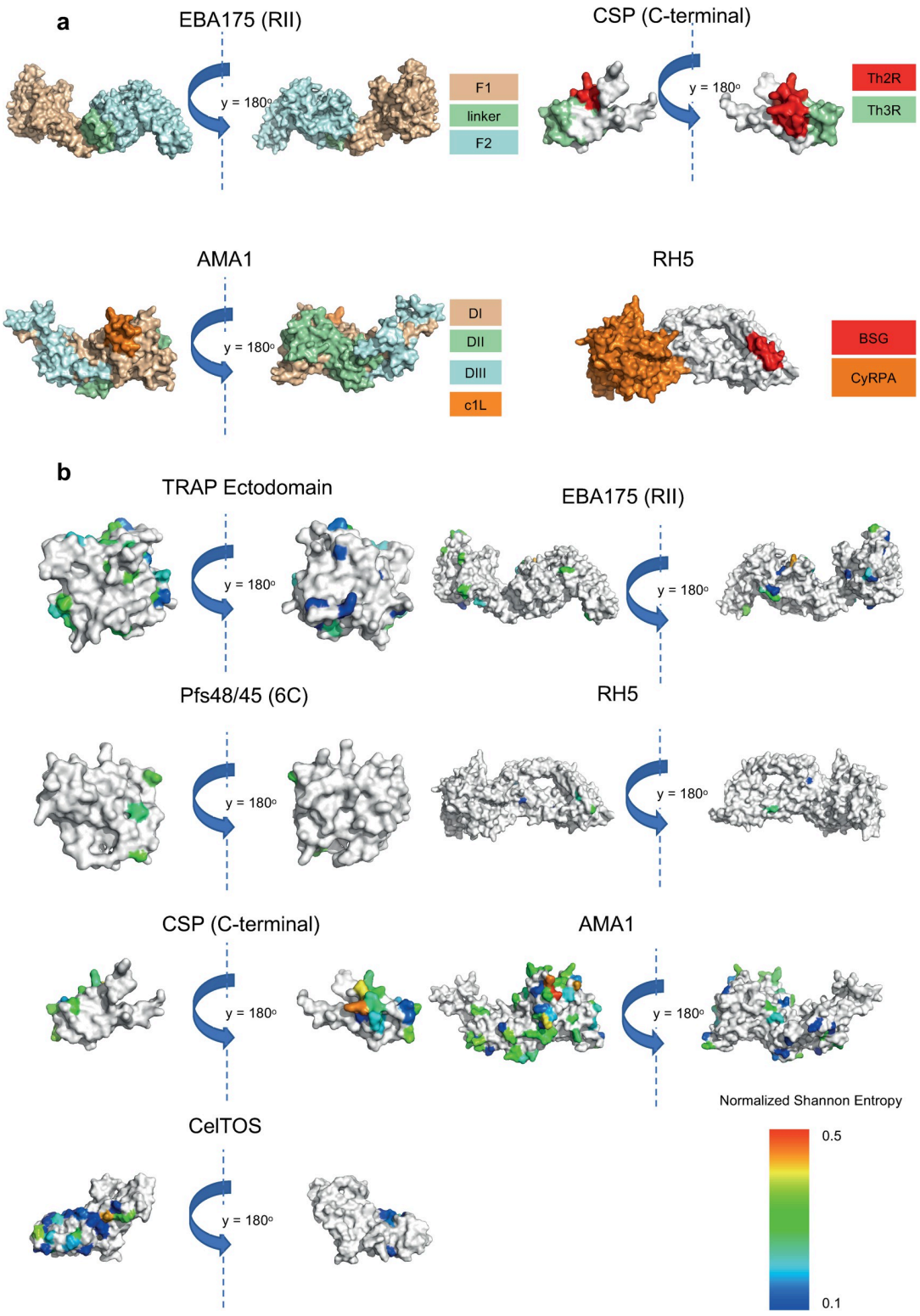

Fig 4. Mutations at functionally important interfaces. (a) Available domain and epitope information for AMA1, EBA175 (RII), RH5, and CSP (C-terminal). **(b)** Site specific diversity measure for CelTOS, Pfs48/45 (6C domain), TRAP (ectodomain), CSP (C-terminal), AMA1, EBA175 (RII) and RH5. Normalized Shannon Entropy was calculated per residue for these antigens (unavailable residues were coloured in white). Higher entropy values indicate higher diversity across all populations for a particular residue. Residues from the CSP Th2R epitope and AMA1 C1L loop have the highest entropy values. Very low entropy values across all populations were observed for SERA5, and CyRPA.

https://doi.org/10.1371/journal.pcbi.1009801.g004

antigens, which is an indication of balancing selection focused on specific regions of the protein. Nearly neutral *Tajima's D* values were found throughout the 3D structure of CyRPA, and SERA5 (C-terminal) for all parasite populations (S4 Fig). In general, we found a dichotomous pattern of balancing selection on some antigens for countries in the Asia-Pacific versus African populations.

The TRAP ectodomain (PDB Code: 4HQF.A) which is comprised of tandem von Willebrand factor A (VWA) and thrombospondin type I repeat (TSR) domains (AA residues: E41 –K240, 3D7 sequence) were included in our analyses [41]. In all populations, the suspected heparin binding interface of TRAP [42] has moderately higher *Tajima*'s D^* scores than the opposite interface of the protein, which acts like the silent face of TRAP (ectodomain) (Fig 5A). Hence, residues Y89 –I100, and I115 –D146 showed moderate spatially derived D^* scores (1.0–1.2) in all observed countries. Residues S123, T124, and N125 form a minor protrusion on the surface of TRAP, and residues R130, R141, K142 are thought to be mediated in heparin binding according to previous study [42]. However, within the active face of TRAP, the intensity of balancing selection (D^* scores) appears to be geographically variable. Minimal nucleotide diversity variation was observed amongst geographical regions (Fig 1).

High D^* scores (2.0–2.4) were found in all populations at the C1-L cluster [41,42] of AMA1 Domain I (AA residue: T194 –D212, 3D7 sequence), known to be associated with immune escape [16,43–45] (Fig 5B). Most of these residues were identified as discontinuous epitopes by the IEDB epitope prediction resource [46]. As expected, the entire AMA1 interface with hydrophobic binding cleft and RON2 interacting sites shows high balancing selection [47]. Similar to previous spatially derived analysis [16], the region on the border of DII and DIII domains (AA residues: P303 –F312, S432 –Y446, I479 –K508, 3D7 sequence) appears to have the highest D^* scores (> 3) in all populations. This region was previously predicted as the surface exposed face of DII/DIII and suggests that these residues might be the targets of protective host immune responses within AMA1. A monoclonal antibody (1E10) against AMA1 DIII functions synergistically with antibodies against distant parts of AMA1 to inhibit merozoite growth [48]. This indicates that DIII of AMA1 plays a significant role as a target of functional antibody responses against *P. falciparum* in the context of natural infections. Consistent with previous findings [16,27,49], the spatially derived *Tajima's D* values are unevenly distributed with high D^* values mostly exclusive to one face of the AMA1 molecule, and close to zero on the opposing face of AMA1, which has been previously been described as the "silent face" [49,50] except in PNG where moderately high D^* scores (1.9) were found at AA residues V524 –Y532. This is in line with the previous hypothesis that silent face of AMA1 has minimal exposure to the immune system [45,49]. The high level of spatial Tajima's D^* and nucleotide diversity was similar amongst populations for AMA1 suggesting that this antigen experiences similar selective pressures across different populations (Figs 5B, and S5).

Ripr was analysed in the context of the linear nucleotide sequence due to the poor predicted structure. The N-terminal region of RIPR contains 2 EGF-like domains, while the C-terminal region contains 8 EGF-like domains [51]. We found some degree of balancing selection (moderately high *D* values = 1.5) and high nucleotide diversity proximal to C-terminal EGF 7–10 region, located away from the CyRPA:RIPR interface [51] (nucleotide residues based on coding region 2900–2980, 3D7 sequence) of *ripr* at most of the populations (Fig 5C). These results were aligned with the anticipated RIPR-specific monoclonal antibody binding sites identified in a previous study [52]. This suggests the C-terminal region of RIPR is a target of host immunity in most populations.

While the crystal structure of EBA-175-RII has been solved [53], we used a 3D7-based *ModPipe* homology structure for analysis, as this model includes a number of functionally important residues that were not resolved in the experimental structure [16]. For D^* analysis on

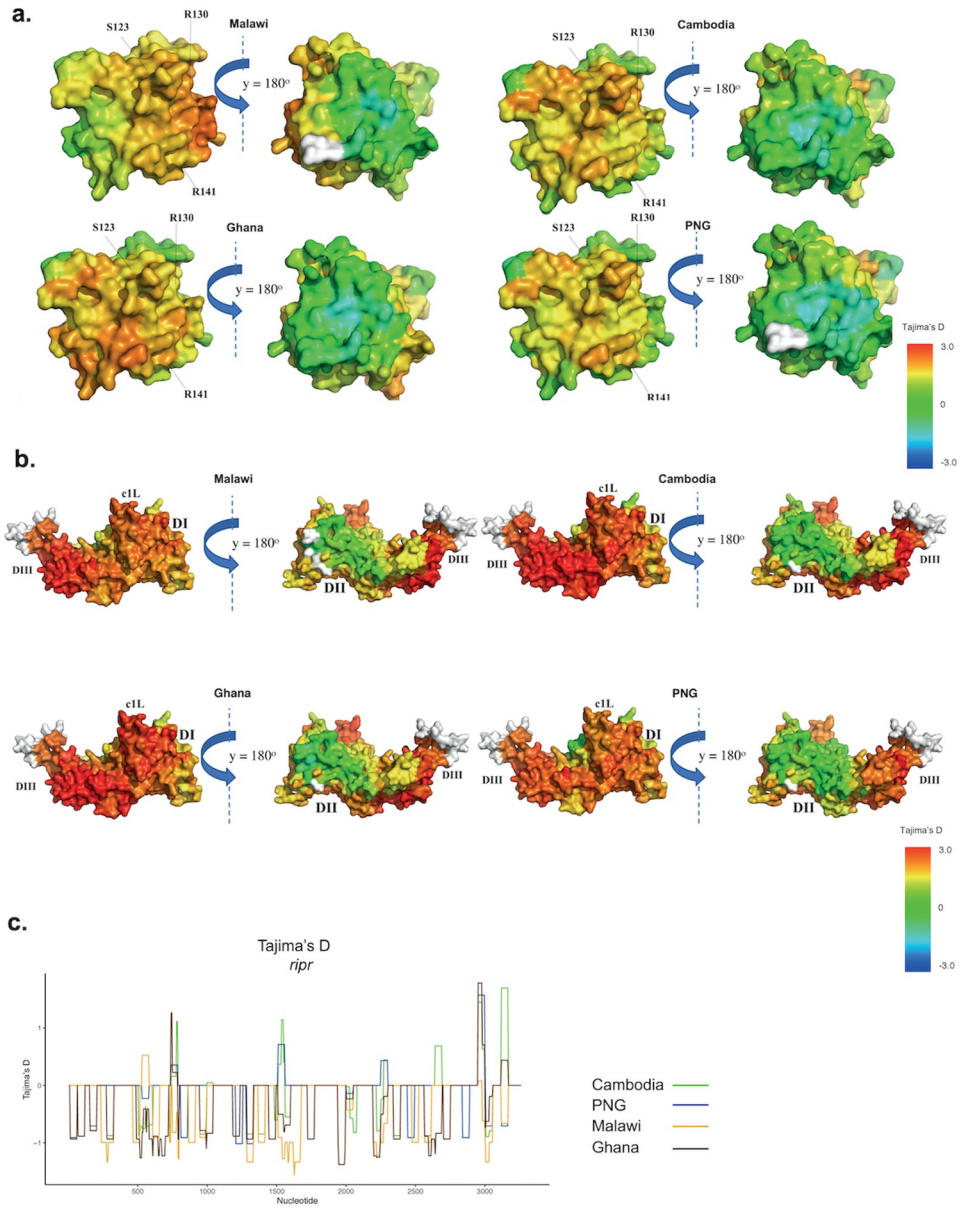

Fig 5. Antigens displaying geographically conserved balancing selection at functionally important interfaces. Antigens were not normalised based on their sizes. a. Spatially derived Tajima's D (D^*) score calculation for TRAP (Ectodomain) with incorporation of protein structural information using a 15Å window. TRAP (ectodomain) (PDB Code: 4HQF.A) was used. The structure was coloured according to D^* scores mapped to each residue, and undefined D^* are shown in white. Residues S123, R130 and R140 are involved in mediating heparin binding. Sample sizes: Malawi (n = 133), Ghana (n = 238), Cambodia (n = 430), and PNG (n = 112). b. Spatially derived Tajima's D (D^*) calculations for AMA1 with incorporation of protein structural information using 15 Å window. The manually modelled structure of AMA1 was used based on published results [27]. The structure was coloured according to D^* scores mapped to each residue with undefined D^* scores were shown in white. The DI, DII, DIII, and surface exposed c1L loop are indicated. Sample sizes: Malawi (n = 139), Ghana (n = 243), Cambodia (n = 433), and PNG (n = 112). c. Polymorphism and evidence of selection for ripr. Tajima's D statistic calculated for disordered regions of RIPR in samples from Cambodia, PNG, Malawi, and Ghana. Tajima's D is calculated with a sliding window approach (a window size of 50 bp and a step size of 5 bp). Nucleotide positions based on coding region are shown in the x-axis. Sample size for each respective population are as follows: Malawi (n = 137), Ghana (n = 246), Cambodia (n = 428), and PNG (n = 111).

https://doi.org/10.1371/journal.pcbi.1009801.g005

EBA-175-RII, a large region of the F1 domain which predominantly consisted of AA residues E226—L294, I312 –K324, and W377 –I400 was under balancing selection (i.e. high D^* scores of 2.8–3.0) in all parasite populations. This cysteine-rich region is also involved in the dimerization interface formation (helix linker and disulphide bridges)[53] between two molecules of EBA-175-RII as it binds to human receptor glycophorin A. During dimerization, this region also makes contact with another cysteine rich F2 domain of the other dimer pair in a 'handshake' interaction [53]. However, slight variations between African and Asia-Pacific populations can also be found within the site from F2 domains, which are predominantly comprised of AA residues C476–C488 and Y710 –F722 where high D^* scores (1.8–2.0) were observed within most Asia-Pacific countries, but not African countries (D^* scores < 0.4) (Fig 6A). The F2 domain makes most of the contact with glycan [54] and residues C476–C488 are a part of F2 β-finger domain that is also a target of inhibitory antibodies, R215 and R217 [54,55]. The linker region was found to be conserved within all populations (Fig 6A). However, the limited nucleotide diversity was observed amongst geographical regions.

The cryo-EM structure of full length RH5 (PDB Code: 6MPV, chain B) was used in our analysis [56]. Evidence of balancing selection was limited within African countries parasite populations based on the spatially derived *Tajima's* D^* scores. In contrast, within Asia-Pacific countries, moderate to high D^* scores (1.5–2.0) were consistently observed at the Basigin (BSG or CD147) binding sites and appear to be under balancing selection (AA residues: S189 – D207, 3D7 sequence) (Fig 6B). Additionally, the residues near CyRPA binding sites which predominantly consist of residues I386 –F421 have moderate evidence of balancing selection (D^* score of 1.5) within PNG population (Fig 6B). This suggests the BSG binding site may be a key target of host immunity and aligns with previous findings from Alanine *et al.* (2019) [57]. Non-synonymous AA polymorphisms at residues Y147 and H148 (nucleotide positions 439, and 442, 3D7 sequence) were under balancing selection in most Asia populations but appear to be under adaptative selection in some African populations according to the linear sliding window analysis with *Tajima*'s *D* values around 1.7 (S3 Fig). These residues did not have PDB coordinates (physical proximity to N-terminal disordered region), and therefore cannot be detected via spatially derived 3D analyses. Nucleotide diversity of RH5 was slightly higher in Asia-Pacific populations than African populations (S3 Fig).

We observed limited evidence of balancing selection (near neutral) within the thrombospondin receptor domain of CSP [58,59] for Asia-Pacific populations (Cambodia and PNG). However, moderately high D^* scores (1.2–1.3) were observed at Th2R residues (AA residues E310 –L327, 3D7 sequence) at Malawi and Ghana populations which suggests some evidence of balancing selection (Fig 6C). Similarly, nucleotide diversity of some of the residues comprising thrombospondin receptor domain of CSP were relatively low in Asia-Pacific populations compared to African populations (S7 Fig). A moderately high to high degree of balancing selection (D^* scores of 1.0 and 2.9) for Th3R residues (AA residues G341 –I364) was also observed within African populations (Fig 6C).

Moderately high D^* scores (1.0–2.0) are found within residues Q1612 –F1625, and E1632 – C1647 from African populations (Malawi and Ghana), but not in Asia-Pacific populations (Fig 6D). Most of these residues were a part of conformational epitopes and involved in interaction with potent monoclonal antibody (G17.12) [60] and inter-domain interactions. In addition, nucleotide diversity of these residues is relatively high in African populations compared to Asia-Pacific parasite populations (S8 Fig). However, different part of MSP1-19 with residues P1651 –C1682 were under balancing selection in most of the Asia-Pacific populations (only shown here for Cambodia and PNG populations) (Fig 6D).

Hotspots of balancing selection were found within N-terminal AA residues F87 and D93 –S104 (3D7 sequence) of CelTOS [61] for most of the populations with moderately high to high spatially

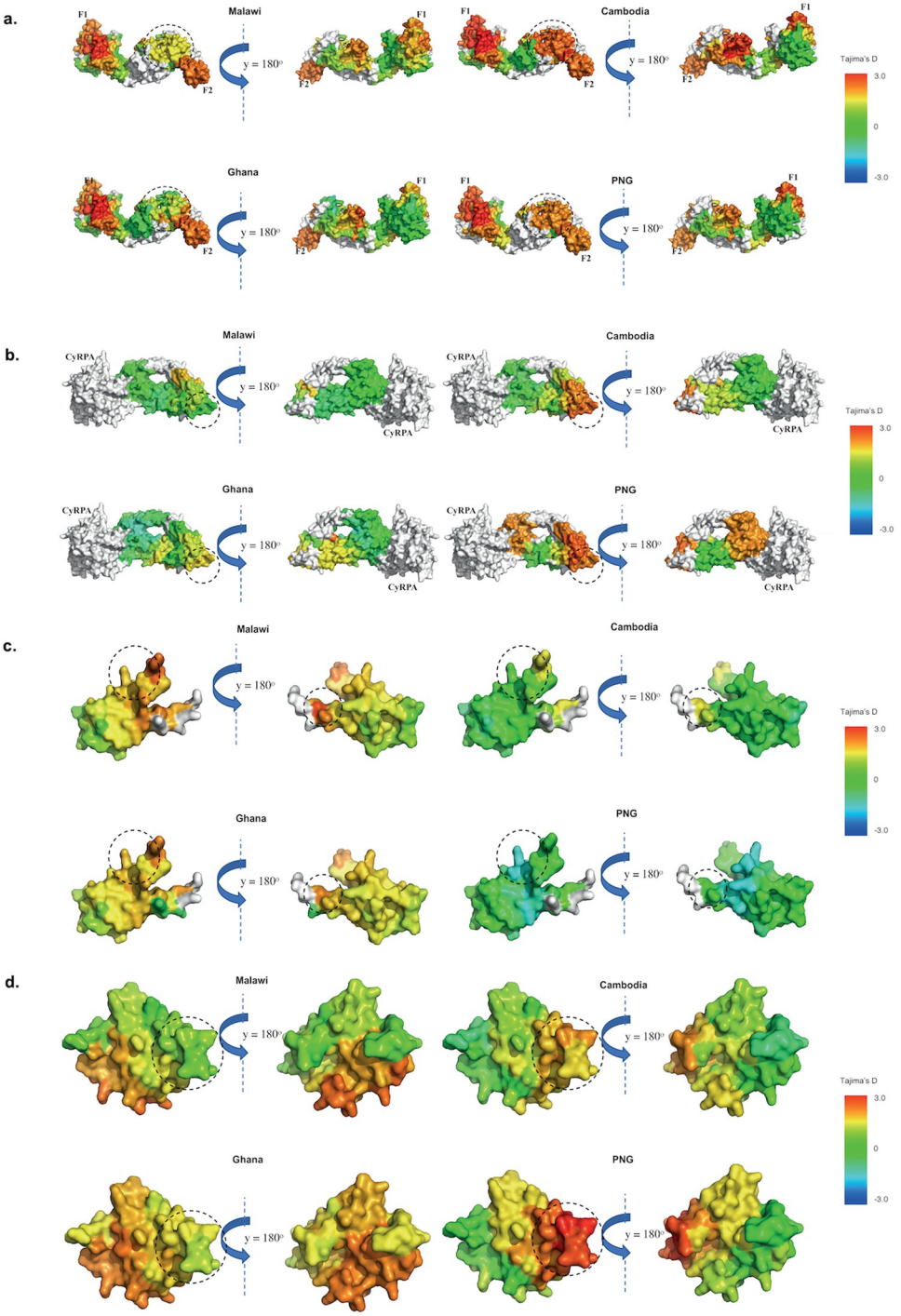

Fig 6. Antigens displaying geographically variable balancing selection hotspots. Antigens were not normalised based on their sizes. a. Spatially derived Tajima's D (D^*) for EBA175 with incorporation of protein structural information using 15 Å window. 3D7-based ModPipe model of EBA175 (RII) based on 1ZRO template was used. Structure was coloured according to D^* scores mapped to each residue with undefined D^* were shown in white. The highlighted region (in circle) shows different D^* scores between Asia-Pacific and African countries. Sample sizes: Malawi (n = 136), Ghana (n = 237), Cambodia (n = 432), and PNG (n = 112). b. Spatially derived Tajima's D (D^*) calculation for countries from Asia-Pacific and Africa for RH5 with incorporation of protein structural information using 15Å window. Cryo-EM structure of RH5-CyRPA complex (PDB code: 6MPV.B) was used. Structure was coloured according to D^* scores mapped to each residue with undefined D^* and CyRPA were shown in white. The circle indicates Basigin-binding sites where different D^* scores were observed between Asia-Pacific and African

countries. Sample sizes: Malawi (n = 142), Ghana (n = 249), Cambodia (n = 433), and PNG (n = 112). c. Spatially derived Tajima's D (D*) calculations for populations from Asia-Pacific and African regions for CSP (C-terminal) with incorporation of protein structural information using 15 Å window. The crystal structure of the thrombospondin receptor (TSR) domain of CSP [58] (PDB code: 3VDJ, chain A, AA residues: Y306—H376, 3D7 sequence), which consists of Th2R and Th3R (T-cell epitopes) was used [59]. Structure was coloured according to D* scores mapped to each residue with undefined D* were shown in white. The highlighted region (in circle) shows different D* scores between Asia-Pacific and African countries. Sample sizes: Malawi (n = 135), Ghana (n = 223), Cambodia (n = 431), and PNG (n = 111). d. Tajima's D (D*) calculation for geographic area or countries from the Asia-Pacific and African regions for $MSP1_{19}$ with incorporation of protein structural information using 15 Å window. The structured region of MSP1 (MSP1-19, AA residue: N1607—S1699) is based on the ModPipe homology model using template (PDB code: 1OB1) [60]. The structure was coloured according to D* scores mapped to each residue with undefined D* were shown in white. The circle indicates variable balancing selection hotspots between Asia-Pacific and African countries. Sample sizes: Malawi (n = 101), Ghana (n = 183), Cambodia (n = 270), and PNG (n = 72).

https://doi.org/10.1371/journal.pcbi.1009801.g006

derived *Tajima's D** scores (1.0–2.2) (S7A Fig). Within the Malawi population, we found high *D** scores (1.8–2.0) at AA resides D131 –I138 (3D7 sequence), but limited balancing selection was found for CelTOS within PNG (*D** scores < 0.4) (S7A Fig). Nucleotide diversity was also relatively low in PNG populations compared to others (S7B Fig). However, the presence and distribution of *D** values were variable amongst geographic locations (S7A Fig). Limited functional information was available for CelTOS therefore the significance of these findings is not able to be postulated.

Intrinsically disordered proteins may be targets of immune selection

Intrinsically disordered proteins (IDP) and intrinsically disordered regions (IDR) are a class of proteins, which lack secondary and tertiary rigid structure (under physiological conditions) but possess active roles in many biological processes [62–71]. Recent predictions have shown an abundance of IDPs within the *Plasmodium* proteome including several vaccine candidate antigens [28]. This is not surprising given that disordered regions create larger intermolecular interfaces, which increases the chances of interaction with potential binding partners, even without tight binding, providing flexibility for binding diverse ligands and other proteins including functional antibodies [29]. IDPs can have a diverse range of functions, although our understanding of the role of IDPs within biological systems is still incomplete [64,68]. Due to the lack of well-defined three-dimensional structure for antigens with extensive disordered regions, we calculated linear *Tajima's D* values to determine balancing selection. Highly disordered proteins were defined as those with more than 50% residues with disorder scores higher than 0.4. This includes several emerging and established vaccine candidate antigens such as CTRP, EXP1, TRAP (C-terminal), STARP, GLURP, EBA175 (RIII-V), MSP3, MSP4, MSP6, RALP1, SERA5 and RESA. For these antigens except STARP and RALP1, evidence of balancing selection was consistently observed within disordered regions with only slight variations amongst geographical regions (Figs 7 and S10). Intrinsically disordered regions (IDR) were previously shown to be enriched with predicted linear B-cell epitopes [28]. We found disordered residues within these antigens (i.e. N-term of SERA5 compared to its C-term [72,73]) contributing to balancing selection (*Tajima's D* values >1) at most of the analysed geographic area or countries (Fig 7). This suggests IDRs as potential immune targets. However, it has been estimated that IDRs have limited proportion of MHC-binding peptide compared to other domains, which may affect T-cell dependent immune responses [28]. Further *in vitro* and *in vivo* investigations are needed to characterize antigens enriched with disordered proteins as potential vaccine targets.

Discussion

Here we have conducted a systematic meta-population genetic analysis of 23 malaria vaccine antigens using a global dataset of *P. falciparum* WGS (after filtration for high complexity

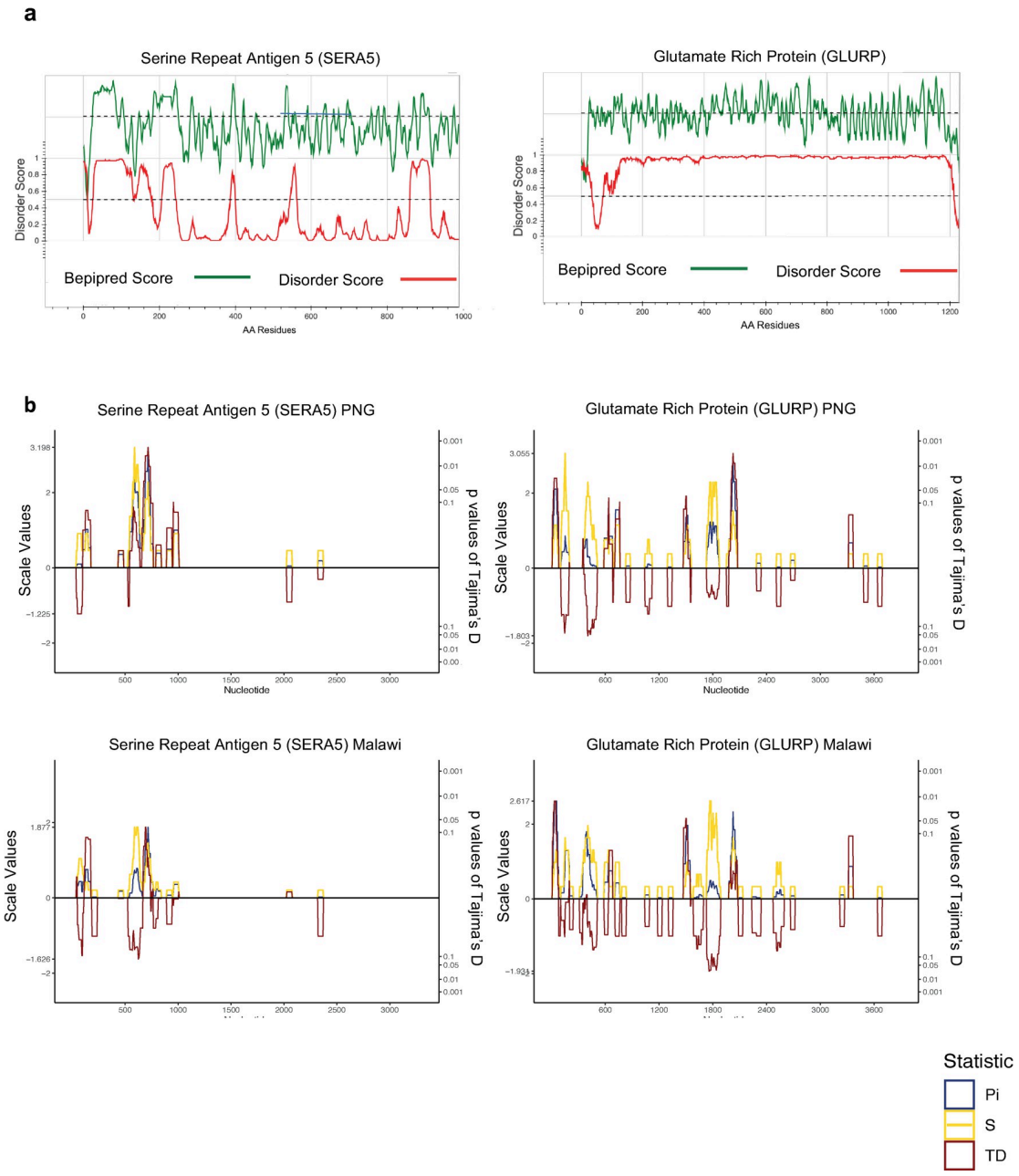

Fig 7. Diversity and selection of SERA5 and GLURP in Asia-Pacific and African regions. a) Computational predictions of protein disorder and B-cell epitopes in SERA5 and GLURP. The green line represents the linear B-cell epitope mapping scores and the red line shows the protein disorder score, respectively. b) Diversity statistics along the sera5 and glurp genes in samples from in Asia-Pacific and African regions, represented by Tajima's D (red line), nucleotide diversity (blue line) and number of segregating sites (yellow line). It is calculated in the context of linear sequence level based on coding region with the sliding window approach (a window size of 50 bp and a step size of 5 bp). Nucleotide positions based on coding region are shown in the x-axis. Sample sizes: Malawi (n = 106), Ghana (n = 208), Cambodia (n = 405), and PNG (n = 108).

https://doi.org/10.1371/journal.pcbi.1009801.g007

infections) covering 15 malaria-endemic countries [22]. Our results demonstrate low frequencies of variants included in malaria vaccines under development, and variable levels of diversity and natural selection amongst different antigens and also amongst different geographic regions. Overall, the results demonstrate that the current 'one size fits all' approach to vaccine

development may result in limited vaccine efficacy and variable efficacy across geographic regions [74]. Different approaches to vaccine design may include focusing on the most common haplotypes circulating in the population, and not limiting to the reference strain by for example, basing the vaccine on the most common haplotypes as shown in Table B in S1 Text.

The variable success of malaria vaccines may be due to the high diversity of vaccine candidate antigens [7,15,75] and strain-specific immune responses. Merozoite antigens, particularly AMA1, and EBA175 have high haplotype and nucleotide diversity across different populations, consistent with strong immune selection. Variable levels of nucleotide diversity for different populations were observed for CSP (C-term) and CelTOS which may result from variable transmission intensity in different regions. Low nucleotide diversity and moderate haplotype diversity across populations for STARP, RH5, and RALP1 suggests that although there may be low polymorphism, immune selection maximises the chance of new infections carrying novel haplotypes. Only limited haplotype and nucleotide diversity was observed within CyRPA, TRAMP, and Pfs48/45 suggesting strong conservation and limited immune selection. For STARP, RH5, RALP1, and CyRPA antigens, low expression throughout the erythrocytic stage (based on abundance of *in vivo* mRNA transcript) suggests minimal opportunities for host evolutionary pressure to be exerted upon those antigens [68,76].

Previous studies have indicated that parasite antigens are differentially immunogenic [77] and this manifests in varying intensity of immune (balancing) selection. We have detected extremely high balancing selection across different populations for AMA1 and EBA175 (RII), moderate to high balancing selection was found for TRAP (ectodomain) and RIPR. Whereas CyRPA, and SERA5 (C-term) appear to have consistently low diversity, and are neutrally evolving in most populations, which implies they have highly conserved epitopes or weak immunogenicity. Variable selection amongst different populations was found for RH5, CSP (C-term), CelTOS, MSP-19 and Pfs48/45 (S9 Fig). High-affinity, receptor-specific binding between parasite ligand and host receptor (such as EBA175 –Glycophorin A) or low-affinity general cell binding (such as TRAP–heparan sulphate proteoglycans (HSPGs)) may explain different balancing selection intensity on these antigens amongst populations [42,55,78]. The high relative solvent accessibility (RSA) of polymorphic residues on some of these antigens suggests they may be antibody targets, supported by considerable amino acid variability, as measured by Shannon Entropy, at host-parasite interfaces, particularly for the TRAP ectodomain, CSP C-term, AMA1, EBA175 RII, RH5, Pfs48/45 6C domain, and CelTOS. All of these antigens have selection hotspots at these interfaces in at least one population.

Hotspots of balancing selection were common within functionally important 3D interfaces of AMA1 and TRAP across populations suggesting sharing of dominant epitopes which may allow the development of broadly efficacious vaccines using a few common variants. On the other hand, we identified differential patterns of selection pressure between African and Asia-Pacific countries at the ligand binding sites of EBA175 RII, RH5, MSP1-19 and CSP C-terminal end. Even more discordant patterns between geographic regions were found for gametocyte antigens such as CelTOS and Pfs48/45. This might reflect parasite adaptation to different host environments. These results were aligned with haplotype network analysis. For blood stage antigens, EBA175 and RH5, this could be driven by different isoforms of the respective host receptors such as glycophorin-A (GPA) [79] and Basigin [80]. GPA encoded by the *GYPA* gene is under selective pressure in different malaria endemic areas, with the MN blood group in particular showing protective associations against malaria in PNG [79,81]. Associations between different isoforms of BSG with malaria outcomes have not yet been identified however *in vitro* studies have shown a reduced RH5 interaction with isoform 2 of BSG [80]. Despite the very low diversity of RH5 there are multiple common haplotypes, with strong balancing selection at the BSG binding site. However, this balancing selection was only found

within Asia-Pacific populations. This suggests that RH5 may have adapted to an unidentified BSG isoform in the region, or that some other factor might be contributing to selection at this BSG-binding interface. Similarly, CSP C-terminal and CelTOS shows unique selection hotpots on the 3D structure as these proteins are predominantly expressed in the sporozoites, this might reflect adaptation to different Anophelene species in Africa and Asia. This geographically variable selection was not easily observed with sliding window analysis of linear sequences showing the enhanced resolution of the 3D analysis.

Immune responses may also be generated to conserved regions of antigens [82] such as that in the novel universal influenza B-cell vaccine approach [9]. This would increase vaccine efficacy for diverse strains by eliciting broadly cross-reactive functional immune responses such as neutralising antibodies directed at highly conserved regions [83–86]. Conserved regions of malaria vaccine candidate antigens have been revealed within RH5, CyRPA, or SERA5 C-terminal end (S8 Fig) and a lack of selection pressure in the C-terminal region of SERA5 and CyRPA. Therefore, not only does this study catalogue and characterise global malaria vaccine antigen diversity, it can also guide the identification of conserved regions. This study therefore provides critical practical information for the next generation of malaria vaccines.

The immunogenic properties of disordered proteins are not well understood. Lower immunogenicity of a disordered region compared to a structured region within MSP2 has been observed [87] suggesting that disordered proteins fail to activate the production of high affinity, specific antibodies, because they adopt diverse conformations depending on their ligands [88]. Therefore, polymorphic residues on disordered proteins may be a 'smoke screen' by diverting protective immune responses from more ordered targets. Nonetheless, numerous B-cell epitopes have been found within disordered regions [28] and many of these appear to elicit functional immune responses [70,71,75,89–92]. For instance, the protective effect of RTS,S appears to be predominantly mediated by generation of antibody responses to the disordered central NANP repeat region [90,93]. Our observation of immune-mediated selection pressure in the N-terminal disordered region of SERA5, is also supported by the finding that *in vitro* parasite growth was inhibited by anti-SERA5 antibodies raised in squirrel monkeys (*Saimiri sciureus*) [72]. In addition, we found balancing selection in the disordered EBA175 RIII-V domain which may facilitate the molecular binding of GYPA to EBA175 RII [94,95]. Future work could extend to characterise disordered proteins into different subclasses based on their immunogenicity.

This study presents a detailed analysis of the genetic diversity and immune selection of 23 candidate vaccine antigens and a framework for analysis of further antigens which has already contributed critical information about the diversity of novel vaccine candidates [96–98]. In addition, the incorporation of sequence-based metrics such as Tajima's *D* into protein structures enhances biological understanding, as evidenced by the finding of unique patterns of immune selection in different geographic areas. However, the dataset mostly includes single census of clinical samples from different collection timepoints, which prevent the exploration of patterns under different clinical presentations and changes in transmission. Further investigations could be done using sequencing datasets from longitudinal cohorts in association with epidemiological parameters to identify specific polymorphisms that enable the parasite to evade host immune responses. Nevertheless, *in silico* findings from our study support *in vitro* and *in vivo* studies and might uncover novel targets of host immunity. The development of a novel population genetic analysis software 'VaxPack' was crucial to automate the data analysis. The approach used in this work is also applicable to a wide variety of problems from different organisms.

Materials and methods

Data sources

WGS data from *P. falciparum* clinical isolates (n = 2512) from multiple countries including Bangladesh, Cambodia, Laos, Myanmar, Thailand, Vietnam, Congo (DRC), Malawi, Gambia, Ghana, Guinea, Mali, Nigeria and Senegal were obtained from the MalariaGEN Pf3k project release 5.1 as unmapped BAM (uBAM) files [10]. WGS data from *P. falciparum* isolates from PNG (n = 156) were also obtained from the MalariaGEN *P. falciparum* Community Project [23] giving a total of 2668 WGS for data analysis. All samples were sequenced using Illumina short read technology at the Sanger Research Institute, UK, and the Broad Institute, Boston, USA [10,24]. The sequences from these 15 countries cover four geographical regions (South-east Asia, Central Africa, West Africa, and Oceania) with variable malaria transmission intensities [99].

Processing and analysis of whole genome sequence data

P. falciparum read alignments from all 2668 samples were mapped to an indexed 3D7 reference genome using uBAM files as input. We followed the standard best practice from *Genome Analysis Toolkit (GATK)* version 4.0.12.0 implemented in *nextflow* (https://github.com/gatk-workflows/gatk4-germline-snps-indels) with some minor changes as follows: (i) running the pipeline twice—the first without base quality score recalibration (BQSR) and (ii) using the pass variants from the first run for BQSR and hard filtering of variants instead of variant quality score recalibration (VQSR) as there is no external variant call set available. Variants were determined using GATK's *haplotypeCaller* to generate haploid genotype calls (*P falciparum* blood stages are haploid) and all isolates were joint genotyped using gvcf files. Variants that had been processed with the GATK pipeline were functionally annotated with *SnpEff for* genomic variant annotations and functional effect prediction (version 4.3T) [100]. We restricted our analyses to single nucleotide polymorphisms (SNPs) from the 14 nuclear chromosomes only, excluding indels and variants from hypervariable regions, as they evolve by different mechanisms to SNPs and have much less impact on antigen genes (S1 Fig). To obtain a high-quality variant dataset for downstream analyses, we further developed stringent variant filters using GATK's SelectVariants and VariantFiltration modules such as read coverage statistics, relative distribution of reads [101] and multiplicity of infection (MOI) (S1 Fig). Variants were removed if they met at least one of the following filtering criteria: QD $<$ 20.0, MQ $<$ 50.0, MQRankSum $<$ -2 (if annotated), ReadPosRankSum (less than -4 or greater than 4 if annotated), SOR $>$ 1. Further filtration was performed based on read depth, and missingness (S1 Fig). We estimated the clonal proportions for each sample using R package, *moimix* (available at https://github.com/bahlolab/moimix), a robust alternative (B)-allele frequency-based estimation. We only kept data from samples with single infections (MOI = 1) or two infections (MOI = 2; i.e. Fws $>$ 0.80) [24]. MOI = 2 isolates were treated as diploid genomes permitting heterozygous genotype calls (S2 Fig). Multi-allelic sites were also permitted in our bioinformatics framework (S1 Fig). All samples that had F_{ws} greater than 0.90 were treated as MOI = 1. For MOI = 2 samples, phasing of haplotype was performed using read-depth and B-allele frequency (S1 Fig). Thus, the final dataset contained high quality data from 948 MOI = 1 and 551 MOI = 2 isolates. Nucleotides that did not have clear major or minor alleles were classified as 'undefined' bases (S1 Fig). Variants within gene coding regions were extracted in fastA format using a custom R-script (https://github.com/myonaung/Naung-et-al-2021). After processing the sequence data, identifying high quality variants and removing complex infections, gene sequences were extracted and compiled as a fastA formatted file.

Processing of antigen sequences

Antigens were selected based on their inclusion as vaccine candidates in the WHO Rainbow Table [20] or whether they had been previously identified as novel candidates based on previous literature [6]. Details of regions analysed, and other details are summarised in Table 1. The raw fastA output for each gene was further edited using customized python script (https://github.com/myonaung/Naung-et-al-2021) by removing sequences that contain undefined or low-quality bases and sequences arising from minor clones in the co-existence of different genotypes (phased by read depth (S1 Fig)). FastA formatted files of respective genes can be found in https://github.com/myonaung/Naung-et-al-2021/tree/master/Raw_FASTA. As Illumina sequencing is subject to some miscalled bases, singleton variants (i.e. minor alleles seen only once in the entire dataset) were assumed to be artefacts and converted back to reference alleles to prevent false positive variants.

The presence of at least two isolates with the same minor allele was considered independent validation of the existence of these alleles in natural parasite populations. Apart from removing singletons, no other minor allele frequency threshold was applied to the dataset. In the case of multiple clones being identified ($0.90 \geq \text{Fws} > 0.80$), only the defined major alleles were included for further analyses. Samples collected from the same country were combined and assumed to have been sampled from a single population.

Population genetic analyses

Population genetic parameters were calculated using a customized in-house R package (source code available at https://github.com/BarryLab01/vaxpack). This program takes aligned multi-fastA files excluding intronic regions as input, alongside reference sequence of the same length. As measure of diversity we defined the polymorphisms for each antigen by the number of polymorphic sites (S), the number of synonymous (dS) SNPs, number of nonsynonymous (dN) SNPs, dN-based haplotypes, the Nei's nucleotide diversity (π) calculated as

$$\pi = \sum_{ij} x_i x_j \pi_{ij} \quad (1)$$

where x_i and x_j are the respective frequencies of the i^{th} and j^{th} sequences, π_{ij} is the number of nucleotide differences per nucleotide site between i^{th} and j^{th} sequences [92]. The haplotype diversity (Hd) was defined as

$$\text{Hd} = [\text{n}/(\text{n}-1)][(1 - \Sigma(\text{f}_\text{i})^2)] \quad (2)$$

where n is the sample size and f is the frequency of the i^{th} allele. Immune mediated selection due to parasite adaptations to evade antibody recognition of dominant epitopes on a particular antigen is indicated by the presence of balancing selection. To measure selection, *Tajima's D* (*D*) statistics were calculated [102,103]. Under neutral conditions, *Tajima's D* values are expected to be approximately zero [104,105], positive *D* values indicate an excess of intermediate frequency polymorphisms most likely due to balancing (immune) selection, and negative values indicate purifying (directional) selection. However, measurement of *Tajima's D* across genetically differentiated populations with varying allele frequencies (i.e. due to population structure) can produce false positive signals of balancing selection. Thus, all *Tajima's D* analyses were calculated separately for each country. Analyses were first carried out with the complete global dataset (n = 1499 samples) for all antigens from Table 1 to investigate the global range of diversity of each antigen as well as the frequency of reference (3D7) haplotypes (frequently used in vaccine development), while the population specific dataset was used to investigate the range and distribution of diversity for countries (23 antigens). RON2

(PF3D7_1452000) was excluded from analyses as no polymorphism was found. The full-length antigen genes as well as functionally important domains were included in the country level.

For each analysed antigen gene, we investigated relationships amongst haplotypes by constructing haplotype networks using the Templeton, Crandall, and Sing (TCS) method in *PopArt* version 1.7 [106]. To focus the analysis on antigen diversity, we based our analysis on non-synonymous polymorphism (via amino acid translation) as synonymous polymorphisms do not change the protein structure. It is very important to note that haplotypes are simply the combination of polymorphic sites without a particular weight upon a specific or functionally important polymorphic residue. All nonsynonymous SNP haplotypes with a frequency greater than 0.5% of total population were included in the analysis using 'Nexus' file format as input. The TCS network was constructed using an agglomerative approach. This allows for visualization of the relationships amongst haplotypes and their geographic distribution. Generally, less frequent recombinant haplotypes connect major (high frequency) haplotypes.

The Relative Solvent Accessibility (RSA) derived from the solvent accessible surface area (ASA) was predicted for all antigens for each amino acid residue based on primary 3D7 protein sequence using neural network based NetSurfP1.1 program [30,107]. It is given by the following equation [30,107]:

$$\mathrm{RSA} = \mathrm{ASA}/\mathrm{ASA}_{\mathrm{MAX(Gly-X-Ala)}} * 100\% \quad (3)$$

RSA values for structured regions with known PDB or homology-modelled structures were calculated with DSSP [31] using maximum allowed solvent accessibilities of residues from Tien *et al.* 2013 [107]. Therefore, RSA represents the ratio of parts of a biomolecule exposed to solvent or accessible surface area (ASA) of a given residue observed in the three-dimensional state over maximum surface area of a residue with potential exposure to solvent ($\mathrm{ASA}_{\mathrm{MAX}}$) within an extended tripeptide flanked with either glycine or alanine residue. The data was stored in SQLite database using RSQlite package on RStudio.

Selection pressure may apply to residues that are non-continuous on the linear sequence but proximal in 3D space. Therefore, application of a more accurate spatially derived approach to compute *Tajima's D** and nucleotide diversity is applicable. We chose the radius of 15 Å to reflect the ideal potential antibody antigen interaction as used by Guy *et al.* (2018) [16,17]. However, our spatial averaging approach was limited to the availability of 3D coordinates for each amino acid residue, and therefore were not applied to highly disordered regions or proteins.

3D structure-based and surface accessibility analysis

For each antigen included in Table 1, the 3D structure for full length or specific domains was first derived from the Protein Data Bank (PDB) from the Research Collaboratory for Structural Bioinformatics (RCSB) website (www.rcsb.org). For each antigen that does not have experimentally determined or complete structure, the possibility of comparative structure modelling was determined based on the distribution of intrinsic disordered regions available at https://plasmosip.burnet.edu.au/submission [28]. Except for AMA1, which was modelled manually, template-based models of 3D7 reference strain for these antigens were predicted using *ModPipe*, an automated software pipeline that utilises the program *MODELLER* [108] for the generation of comparative proteins based on previous studies [16]. We used a manually modelled structure of AMA1, which includes full domains I-III based on a combination of *P. falciparum* and *P. vivax* templates according to previously published work [27]. The modelled structures were assumed to be reliable if the *ModPipe* Quality Score (MPQS) is greater than 1.1 This included for antigens–EBA175 (RII), $\mathrm{MSP1}_{19}$, and CelTOS which is 45% sequence identity to

P. vivax CelTOS structure (PDB Code: 5TSZ) [61]. These structure models are accessible via *ModBase* (https://modbase.compbio.ucsf.edu/) [108]. The *BiostructMap* Python package available at https://github.com/andrewguy/biostructmap was used to calculate nucleotide diversity (π), and *Tajima's D* with consideration of spatial information using a 3D sliding window with a radius of 15 Å (D^* = spatially derived D). Spatial *Tajima's D* (D^*) with 3D sliding window is analogous to traditional sliding window analysis that should not be interpreted at individual residues since it averages out contributions from individual residues. Since the spatial *Tajima's D* (D^*) calculation also considers the spatial information of each amino acid residue from the protein, the results represent accurate detection of selection pressures arising at the level of protein structure but might be variant from the traditional linear *D* calculation as shown in Guy *et al.* (2018) [16,17]. D^* scores of 1.0 to 1.9 were considered as moderately high and D^* scores of above 2 are high. Protein structures and features were visualised using *PyMol* with feature values saved into the B-factor column of a PDB file and using the *spectrum* command. We estimated disordered scores using the PlasmoSIP online tool (https://plasmosip.burnet.edu.au), which contains precalculated results from DISOPRED3 [66]. Due to the lack of well-defined three-dimensional structure for antigens with extensive disordered regions, we calculated linear *Tajima's D* values to determine balancing selection.

Normalized Shannon Entropy for each amino acid residue mapped onto the available protein structure was calculated using

$$S = \sum_i \frac{p_i log_2 p_i}{log_2\ 20} \tag{4}$$

where S is the normalized Shannon Entropy, and p_i is the frequency of i^{th} amino from the specific position [90]. The normalized Shannon Entropy is a site-specific diversity measure ranging between 0 to 1 where 0 represents the conservation of a specific amino acid position, and 1 represents an even distribution of all possible standard 20 amino acids (random distribution) at the specific position. The analysis was limited into antigens with available or predicted 3D protein structure.

Supporting information

S1 Fig. Variant Processing Framework. A pre-processing phase specifies GATK's Haplotype caller's SNPs variants from the regions of interest and performs quality control on these variants. The hard-filtering phase selects only high-quality variants from the variants that passed the pre-processing phase. The integrative phase removes polyclonal (> MOI 2), performs haplotype phasing, and extracts sequences for gene of interest.
(TIFF)

S2 Fig. F_{ws} output from moimix R-package for each country. $F_{ws} > 0.90$ assumed as MOI 1 isolates are highlighted in red, and $0.90 \geq Fws > 0.80$ assumed as MOI 2 isolates are highlighted in blue. Samples with F_{ws} below 0.80 are excluded from the analysis.
(TIFF)

S3 Fig. Polymorphism and selection of full-length rh5 in the context linear sequence level for different populations. The sliding window analyses (a window size of 50 bp and a step size of 5 bp) calculated for segregation sites (S, yellow lines), nucleotide diversity (π, blue lines) and Tajima's D (D, red lines) for each geographic area or country. The results were plotted together and scaled to Tajima's D values. Nucleotide positions based on coding region are shown in the x-axis. The significant values for Tajima's D was determined based on sample

size.
(TIFF)

S4 Fig. Geographically varied selection pressure for SERA5. Tajima's D (D^*) calculation for geographic area or countries from Asia-Pacific and African regions for SERA5 (C-terminal) with incorporation of protein structural information using 15˚A window. The structured region of SERA5 based on experimentally defined structure PDB code: 2WBF was used. The structure was coloured according to D^* scores mapped to each residue with undefined D^* were shown in grey. Only Malawi (n = 106), and PNG (n = 108) populations were shown.
(TIFF)

S5 Fig. Geographically conserved spatially derived nucleotide diversity for full-length AMA1. Nei's nucleotide diversity calculation for geographic area or countries from Asia-Pacific and African regions for AMA1 with incorporation of protein structural information using 15˚A window. Structure was coloured according to nucleotide diversity mapped to each residue. Sample size for each respective population are as follows: Malawi (n = 139), Ghana (n = 243), Cambodia (n = 433), and PNG (n = 112). Similar to selection pressure (determined by D^*), silent face of AMA1 has low nucleotide diversity.
(TIFF)

S6 Fig. Geographically variable spatially derived nucleotide diversity for CSP (C-term). Nei's nucleotide diversity calculation for geographic area or countries from Asia-Pacific and African regions for CSP (C-term) with incorporation of protein structural information using 15˚A window. Structure was coloured according to nucleotide diversity mapped to each residue. Sample size for each respective population are as follows: Malawi (n = 135), Ghana (n = 223), Cambodia (n = 431), and PNG (n = 111).
(TIFF)

S7 Fig. Geographically variable selection for CelTOS. A. Tajima's D (D^*) calculations for populations from Asia-Pacific and African regions for CelTOS with incorporation of protein structural information using 15 Å window. Structure was coloured according to D^* scores mapped to each residue with undefined D^* were shown in white. 3D7-based ModPipe model of the P. vivax CelTOS based on 5TSZ template was used. Sample sizes: Malawi (n = 142), Ghana (n = 245), Cambodia (n = 433), and PNG (n = 112). B. Nei's nucleotide diversity calculation for geographic area or countries from Asia-Pacific and African regions for CelTOS with incorporation of protein structural information using 15˚A window. Structure was coloured according to nucleotide diversity mapped to each residue. Sample size for each respective population are as follows: Malawi (n = 142), Ghana (n = 245), Cambodia (n = 433), and PNG (n = 112).
(TIFF)

S8 Fig. Geographically variable spatially derived nucleotide diversity for MSP1-19. Nei's nucleotide diversity calculation for geographic area or countries from Asia-Pacific and African regions for MSP1-19 with incorporation of protein structural information using 15˚A window. Structure was coloured according to nucleotide diversity mapped to each residue. Sample size for each respective population are as follows: Malawi (n = 101), Ghana (n = 183), Cambodia (n = 270), and PNG(n = 72).
(TIFF)

S9 Fig. Geographically varied selection pressure for Pfs48/45. The sliding window analyses (a window size of 50 bp and a step size of 5 bp) calculated for Tajima's D (D, red lines) for each population. Nucleotide positions based on coding region are shown in the x-axis. Significant

value for Tajima's D was determined by sample size. Sample size for each respective population are as follows: Malawi (n = 142), Ghana (n = 247), Cambodia (n = 433), and PNG (n = 112).
(TIFF)

S10 Fig. Selection of disordered proteins in Asia-Pacific and African regions. a) Computational predictions of protein disorder and B-cell epitopes in EBA175, MSP3, MSP4, MSP6, RESA, TRAP, EXP1 and CTRP. The green line represents the linear B-cell epitope mapping scores and the red line shows the protein disorder score, respectively. b) Tajima's D statistics along the disordered antigens in samples from Cambodia, PNG, Malawi, and Ghana. It is calculated in the context of linear sequence level based on coding region with the sliding window approach (a window size of 50 bp and a step size of 5 bp). Nucleotide positions based on coding region are shown in the x-axis. Sample size for each respective population are as follows: Malawi (n = 106), Ghana (n = 208), Cambodia (n = 405), and PNG (n = 108).
(TIFF)

S11 Fig. Nucleotide sequence variability of all antigens included in the study based on all available sequences. Sliding window analysis of sequence variability was calculated using algorithm from Proutski and Holmes *et al*., (1997) [1] implemented in MBEToolbox [2] using default parameters on MATLAB (version R2020a). Mean (red line), and standard deviation (green dotted line) within each antigen are shown. Nucleotide positions based on coding region are shown in the x-axis.
(TIFF)

S1 Text. Supporting Materials. Table A: Antigen Diversity Summary (Full length or specific domain). **Table B:** Prevalence of haplotypes in dataset for analysed antigens and their proximity to 3D7 vaccine haplotype.
(DOCX)

Acknowledgments

We are grateful to the communities and researchers of malaria endemic countries that enabled the collection and available of the *P. falciparum* genome sequences used in this project, made publically available through the MalariaGEN *P. falciparum* Community Project. We also thank Z. Pava for critical reading of the manuscript.

Author Contributions

Conceptualization: Alyssa E. Barry.

Data curation: Myo T. Naung, Jacob Munro, Somya Mehra, Melanie Bahlo.

Formal analysis: Myo T. Naung, Jacob Munro, Somya Mehra.

Funding acquisition: Alyssa E. Barry.

Investigation: Myo T. Naung.

Methodology: Myo T. Naung, Elijah Martin, Jacob Munro, Somya Mehra, Alyssa E. Barry.

Project administration: Alyssa E. Barry.

Resources: Jacob Munro, Moses Laman, G. L. Abby Harrison, Livingstone Tavul, Manuel Hetzel, Dominic Kwiatkowski, Ivo Mueller, Melanie Bahlo, Alyssa E. Barry.

Software: Elijah Martin, Andrew J. Guy.

Supervision: Ivo Mueller, Melanie Bahlo, Alyssa E. Barry.

Writing – original draft: Myo T. Naung, Alyssa E. Barry.

Writing – review & editing: Jacob Munro, Andrew J. Guy, Melanie Bahlo, Alyssa E. Barry.

References

1. Prugnolle F, Durand P, Ollomo B, Duval L, Ariey F, Arnathau C, et al. A Fresh Look at the Origin of Plasmodium falciparum, the Most Malignant Malaria Agent. PLOS Pathog. 2011; 7: e1001283. https://doi.org/10.1371/journal.ppat.1001283 PMID: 21383971
2. Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, et al. Genome-Wide Analysis of Selection on the Malaria Parasite Plasmodium falciparum in West African Populations of Differing Infection Endemicity. Mol Biol Evol. 2014; 31: 1490–1499. https://doi.org/10.1093/molbev/msu106 PMID: 24644299
3. Leffler EM, Band G, Busby GBJ, Kivinen K, Le QS, Clarke GM, et al. Resistance to malaria through structural variation of red blood cell invasion receptors. Science. 2017;356. https://doi.org/10.1126/science.aam6393 PMID: 28522690
4. Good MF, Doolan DL. Malaria vaccine design: immunological considerations. Immunity. 2010; 33: 555–566. https://doi.org/10.1016/j.immuni.2010.10.005 PMID: 21029965
5. Deroost K, Pham T-T, Opdenakker G, Van den Steen PE. The immunological balance between host and parasite in malaria. FEMS Microbiol Rev. 2016; 40: 208–257. https://doi.org/10.1093/femsre/fuv046 PMID: 26657789
6. Beeson JG, Kurtovic L, Dobaño C, Opi DH, Chan J-A, Feng G, et al. Challenges and strategies for developing efficacious and long-lasting malaria vaccines. Sci Transl Med. 2019; 11: eaau1458. https://doi.org/10.1126/scitranslmed.aau1458 PMID: 30626712
7. Pringle JC, Carpi G, Almagro-Garcia J, Zhu SJ, Kobayashi T, Mulenga M, et al. RTS,S/AS01 malaria vaccine mismatch observed among Plasmodium falciparum isolates from southern and central Africa and globally. Sci Rep. 2018; 8: 6622. https://doi.org/10.1038/s41598-018-24585-8 PMID: 29700348
8. Genton B, Betuela I, Felger I, Al-Yaman F, Anders RF, Saul A, 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. https://doi.org/10.1086/339342 PMID: 11920300
9. Nachbagauer R, Feser J, Naficy A et al. A chimeric hemagglutinin-based universal influenza virus vaccine approach induces broad and long-lasting immunity in a randomized, placebo-controlled phase I trial. Nature Medicine 27, 106–114 (2021). https://doi.org/10.1038/s41591-020-1118-7 PMID: 33288923
10. Pf3k pilot data release 5 | MalariaGEN. 2020. Available: https://www.malariagen.net/data/pf3k-5
11. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. N Engl J Med. 2012; 367: 2284–2295. https://doi.org/10.1056/NEJMoa1208394 PMID: 23136909
12. Preston MD, Campino S, Assefa SA, Echeverry DF, Ocholla H, Amambua-Ngwa A, et al. A barcode of organellar genome polymorphisms identifies the geographic origin of Plasmodium falciparum strains. Nat Commun. 2014; 5: 1–7. https://doi.org/10.1038/ncomms5052 PMID: 24923250
13. Ouattara A, Barry AE, Dutta S, Remarque EJ, Beeson JG, Plowe C V. Designing malaria vaccines to circumvent antigen variability. Vaccine. 2015; 33: 7506–7512. https://doi.org/10.1016/j.vaccine.2015.09.110 PMID: 26475447
14. Barry AE, Alicia A. Strategies for Designing and Monitoring Malaria Vaccines Targeting Diverse Antigens. Frontier in Immunology. 2014;359. https://doi.org/10.3389/fimmu.2014.00359 PMID: 25120545
15. Barry AE, Schultz L, Buckee CO, Reeder JC. Contrasting population structures of the genes encoding ten leading vaccine-candidate antigens of the human malaria parasite, Plasmodium falciparum. PLoS One. 2009; 4: e8497. https://doi.org/10.1371/journal.pone.0008497 PMID: 20041125
16. Guy AJ, Irani V, Beeson JG, Webb B, Sali A, Richards JS, et al. Proteome-wide mapping of immune features onto Plasmodium protein three-dimensional structures. Sci Rep. 2018; 8: 4355. https://doi.org/10.1038/s41598-018-22592-3 PMID: 29531293
17. Guy AJ, Irani V, Richards JS, Ramsland PA. Structural patterns of selection and diversity for Plasmodium vivax antigens DBP and AMA1. Malar J. 2018; 17: 183. https://doi.org/10.1186/s12936-018-2324-3 PMID: 29720179
18. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics. 1989; 123: 585–595. https://doi.org/10.1093/genetics/123.3.585 PMID: 2513255

19. Guy AJ, Irani V, Richards JS, Ramsland PA. BioStructMap: A Python tool for integration of protein structure and sequence-based features. Valencia A, editor. Bioinformatics. 2018. https://doi.org/10.1093/bioinformatics/bty474 PMID: 29931276

20. Schwartz L, Brown GV, Genton B, Moorthy VS. A review of malaria vaccine clinical projects based on the WHO rainbow table. Malaria Journal. 2012; 9;11:11. https://doi.org/10.1186/1475-2875-11-11 PMID: 22230255; PMCID: PMC3286401.

21. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK. Intrinsic Disorder and Functional Proteomics. Biophys J. 2007; 92: 1439–1456. https://doi.org/10.1529/biophysj.106.094045 PMID: 17158572

22. Tessema SK, Nakajima R, Jasinskas A, Monk SL, Lekieffre L, Lin E, et al. Protective Immunity against Severe Malaria in Children Is Associated with a Limited Repertoire of Antibodies to Conserved PfEMP1 Variants. Cell Host Microbe. 2019; 26: 579–590.e5. https://doi.org/10.1016/j.chom.2019.10.012 PMID: 31726028

23. Malaria GEN, Ahouidi A, Ali M et al. An open dataset of *Plasmodium falciparum* genome variation in 7,000 worldwide samples [version 2; peer review:2approved]. *Wellcome Open Res*. 2021; 6:42 https://doi.org/10.12688/wellcomeopenres.16168.2 PMID: 33824913

24. Henden L, Lee S, Mueller I, Barry A, Bahlo M. Identity-by-descent analyses for measuring population dynamics and selection in recombining pathogens. PLOS Genet. 2018; 14: e1007279. https://doi.org/10.1371/journal.pgen.1007279 PMID: 29791438

25. Early AM, Lievens M, MacInnis BL, Ockenhouse CF, Volkman SK, Adjei S, et al. Host-mediated selection impacts the diversity of Plasmodium falciparum antigens within infections. Nat Commun. 2018; 9: 1–10. https://doi.org/10.1038/s41467-017-02088-w PMID: 29317637

26. ApiLoc. 2020. Available: http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc/gene/Neosporacaninum/NCLIV_048060

27. Arnott A, Wapling J, Mueller I, Ramsland PA, Siba PM, Reeder JC, et al. Distinct patterns of diversity, population structure and evolution in the AMA1 genes of sympatric Plasmodium falciparum and Plasmodium vivax populations of Papua New Guinea from an area of similarly high transmission. Malar J. 2014; 13: 233. https://doi.org/10.1186/1475-2875-13-233 PMID: 24930015

28. Guy AJ, Irani V, MacRaild CA, Anders RF, Norton RS, Beeson JG, et al. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. PLoS One. 2015; 10: e0141729. https://doi.org/10.1371/journal.pone.0141729 PMID: 26513658

29. MacRaild CA, Zachrdla M, Andrew D, Krishnarjuna B, Nováček J, Žídek L, et al. Conformational Dynamics and Antigenicity in the Disordered Malaria Antigen Merozoite Surface Protein 2. PLoS One. 2015; 10: e0119899. https://doi.org/10.1371/journal.pone.0119899 PMID: 25742002

30. Petersen B, Petersen T, Andersen P, Nielsen M, Lundegaard C. A generic method for assignment of reliability scores applied to solvent accessibility predictions. BMC Struct Biol. 2009; 9: 51. https://doi.org/10.1186/1472-6807-9-51 PMID: 19646261

31. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22: 2577–2637. https://doi.org/10.1002/bip.360221211 PMID: 6667333

32. Sabchareon A, Burnouf T, Ouattara D, Attanath P, Bouharoun-Tayoun H, Chantavanich P, et al. Parasitologic and clinical human response to immunoglobulin administration in falciparum malaria. Am J Trop Med Hyg. 1991; 45: 297–308. https://doi.org/10.4269/ajtmh.1991.45.297 PMID: 1928564

33. Beeson JG, Drew DR, Boyle MJ, Feng G, Fowkes FJI, Richards JS. Merozoite surface proteins in red blood cell invasion, immunity and vaccines against malaria. FEMS Microbiol Rev. 2016; 40: 343–372. https://doi.org/10.1093/femsre/fuw001 PMID: 26833236

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: e1000218. https://doi.org/10.1371/journal.pmed.1000218 PMID: 20098724

35. Barlow DJ, Edwards MS, Thornton JM. Continuous and discontinuous protein antigenic determinants. Nature. 1986; 322: 747–748. https://doi.org/10.1038/322747a0 PMID: 2427953

36. Day KP, Marsh K. Naturally acquired immunity to Plasmodium falciparum. Immunol Today. 1991; 12: A68–A71. https://doi.org/10.1016/s0167-5699(05)80020-9 PMID: 2069680

37. Bull PC, Lowe BS, Kortok M, Molyneux CS, Newbold CI, Marsh K. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. Nat Med. 1998; 4: 358–360. https://doi.org/10.1038/nm0398-358 PMID: 9500614

38. Haste Andersen P, Nielsen M, Lund O. Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. Protein Sci. 2006; 15: 2558–2567. https://doi.org/10.1110/ps.062405906 PMID: 17001032

39. Kariuki SN, Williams TN. Human genetics and malaria resistance. Hum Genet. 2020. https://doi.org/10.1007/s00439-020-02142-6 PMID: 32130487

40. Coley AM, Parisi K, Masciantonio R, Hoeck J, Casey JL, Murphy VJ, et al. The Most Polymorphic Residue on Plasmodium falciparum Apical Membrane Antigen 1 Determines Binding of an Invasion-Inhibitory Antibody. Infect Immun. 2006; 74: 2628–2636. https://doi.org/10.1128/IAI.74.5.2628-2636.2006 PMID: 16622199

41. Song G, Koksal AC, Lu C, Springer TA. Shape change in the receptor for gliding motility in Plasmodium sporozoites. Proc Natl Acad Sci. 2012; 109: 21420–21425. https://doi.org/10.1073/pnas.1218581109 PMID: 23236185

42. Pihlajamaa T, Kajander T, Knuuti J, Horkka K, Sharma A, Permi P. Structure of Plasmodium falciparum TRAP (thrombospondin-related anonymous protein) A domain highlights distinct features in apicomplexan von Willebrand factor A homologues. Biochem J. 2013; 450: 469–476. https://doi.org/10.1042/BJ20121058 PMID: 23317521

43. Spensley K.J., Wikramaratna P.S., Penman B.S. et al. Reverse immunodynamics: a new method for identifying targets of protective immunity. Sci Rep. 2019. https://doi.org/10.1038/s41598-018-37288-x PMID: 30770839

44. Harris KS, Casey JL, Coley AM, Masciantonio R, Sabo JK, Keizer DW, et al. Binding Hot Spot for Invasion Inhibitory Molecules on Plasmodium falciparum Apical Membrane Antigen 1. Infect Immun. 2005; 73: 6981–6989. https://doi.org/10.1128/IAI.73.10.6981-6989.2005 PMID: 16177378

45. Takala SL, Coulibaly D, Thera MA, Batchelor AH, Cummings MP, Escalante AA, et al. Extreme Polymorphism in a Vaccine Antigen and Risk of Clinical Malaria: Implications for Vaccine Development. Sci Transl Med. 2009; 1: 2ra5. https://doi.org/10.1126/scitranslmed.3000257 PMID: 20165550

46. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. 2019; 47: D339–D343. https://doi.org/10.1093/nar/gky1006 PMID: 30357391

47. Srinivasan P, Beatty WL, Diouf A, Herrera R, Ambroggio X, Moch JK, et al. Binding of Plasmodium merozoite proteins RON2 and AMA1 triggers commitment to invasion. Proc Natl Acad Sci U S A. 2011; 108: 13275–13280. https://doi.org/10.1073/pnas.1110303108 PMID: 21788485

48. Dutta S, Dlugosz LS, Drew DR, Ge X, Ababacar D, Rovira YI, et al. Overcoming Antigenic Diversity by Enhancing the Immunogenicity of Conserved Epitopes on the Malaria Vaccine Candidate Apical Membrane Antigen-1. Blackman MJ, editor. PLoS Pathog. 2013; 9: e1003840. https://doi.org/10.1371/journal.ppat.1003840 PMID: 24385910

49. Bai T, Becker M, Gupta A, Strike P, Murphy VJ, Anders RF, et al. Structure of AMA1 from Plasmodium falciparum reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. Proc Natl Acad Sci U S A. 2005; 102: 12736–12741. https://doi.org/10.1073/pnas.0501808102 PMID: 16129835

50. Dutta S, Lee S Y, Batchelor A H, and Lanar D E. Structural basis of antigenic escape of a malaria vaccine candidate. Proceedings of the National Academy of Sciences. 2007; 104 (30) 12488–12493; https://doi.org/10.1073/pnas.0701464104 PMID: 17636123

51. Chen L, Lopaticki S, Riglar DT, Dekiwadia C, Uboldi AD, Tham W-H, et al. An EGF-like Protein Forms a Complex with PfRh5 and Is Required for Invasion of Human Erythrocytes by Plasmodium falciparum. PLOS Pathog. 2011; 7: e1002199. https://doi.org/10.1371/journal.ppat.1002199 PMID: 21909261

52. Healer J, Wong W, Thompson JK, He W, Birkinshaw RW, Miura K, et al. Neutralising antibodies block the function of Rh5/Ripr/CyRPA complex during invasion of Plasmodium falciparum into human erythrocytes. Cell Microbiol. 2019; 21: e13030. https://doi.org/10.1111/cmi.13030 PMID: 30965383

53. Tolia NH, Enemark EJ, Sim BKL, Joshua-Tor L. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite Plasmodium falciparum. Cell. 2005; 122: 183–193. https://doi.org/10.1016/j.cell.2005.05.033 PMID: 16051144

54. Ambroggio X, Jiang L, Aebig J, Obiakor H, Lukszo J, Narum DL. The epitope of monoclonal antibodies blocking erythrocyte invasion by Plasmodium falciparum map to the dimerization and receptor glycan binding sites of EBA-175. PLoS One. 2013; 8: e56326. https://doi.org/10.1371/journal.pone.0056326 PMID: 23457550

55. Chen E, Paing MM, Salinas N, Sim BKL, Tolia NH. Structural and Functional Basis for Inhibition of Erythrocyte Invasion by Antibodies that Target Plasmodium falciparum EBA-175. PLOS Pathog. 2013; 9: e1003390. https://doi.org/10.1371/journal.ppat.1003390 PMID: 23717209

56. Wong W, Huang R, Menant S, Hong C, Sandow JJ, Birkinshaw RW, et al. Structure of Plasmodium falciparum Rh5-CyRPA-Ripr invasion complex. Nature. 2019; 565: 118–121. https://doi.org/10.1038/s41586-018-0779-6 PMID: 30542156

57. Alanine DGW, Quinkert D, Kumarasingha R, Mehmood S, Donnellan FR, Minkah NK, et al. Human Antibodies that Slow Erythrocyte Invasion Potentiate Malaria-Neutralizing Antibodies. Cell. 2019; 178: 216–228.e21. https://doi.org/10.1016/j.cell.2019.05.025 PMID: 31204103

58. Doud MB, Koksal AC, Mi L-Z, Song G, Lu C, Springer TA. Unexpected fold in the circumsporozoite protein target of malaria vaccines. Proc Natl Acad Sci. 2012; 109: 7817–7822. https://doi.org/10.1073/pnas.1205737109 PMID: 22547819

59. Groot AS de Johnson AH, Maloy WL Quakyi IA, Riley EM, Menon A, et al. Human T cell recognition of polymorphic epitopes from malaria circumsporozoite protein. J Immunol. 1989; 142: 4000–4005. Available: https://www.jimmunol.org/content/142/11/4000 PMID: 2469729

60. Pizarro JC, Chitarra V, Verger D, Holm I, Pêtres S, Dartevelle S, et al. Crystal Structure of a Fab Complex Formed with PfMSP1-19, the C-terminal Fragment of Merozoite Surface Protein 1 from Plasmodium falciparum: A Malaria Vaccine Candidate. Journal of Molecular Biology. 2003; 328: 1091–1103. https://doi.org/10.1016/s0022-2836(03)00376-0 PMID: 12729744

61. Jimah JR, Salinas ND, Sala-Rabanal M, Jones NG, Sibley LD, Nichols CG, et al. Malaria parasite CelTOS targets the inner leaflet of cell membranes for pore-dependent disruption. Elife. 2016;5. https://doi.org/10.7554/eLife.20621 PMID: 27906127

62. Dunker AK, Silman I, Uversky VN, Sussman JL. Function and structure of inherently disordered proteins. Curr Opin Struct Biol. 2008; 18: 756–764. https://doi.org/10.1016/j.sbi.2008.10.002 PMID: 18952168

63. Tompa P. Unstructural biology coming of age. Curr Opin Struct Biol. 2011; 21: 419–425. https://doi.org/10.1016/j.sbi.2011.03.012 PMID: 21514142

64. Uversky VN. Intrinsically disordered proteins from A to Z. 2011. Available: https://pubag.nal.usda.gov/catalog/996094

65. Tompa P. Intrinsically disordered proteins: a 10-year recap. Trends Biochem Sci. 2012; 37: 509–516. https://doi.org/10.1016/j.tibs.2012.08.004 PMID: 22989858

66. Blanc M, Coetzer TL, Blackledge M, Haertlein M, Mitchell EP, Forsyth VT, et al. Intrinsic disorder within the erythrocyte binding-like proteins from Plasmodium falciparum. Biochim Biophys Acta. 2014; 1844: 2306–2314. https://doi.org/10.1016/j.bbapap.2014.09.023 PMID: 25288451

67. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics. 2015; 31: 857–863. https://doi.org/10.1093/bioinformatics/btu744 PMID: 25391399

68. Davies Heledd M., Nofal Stephanie D., Emilia J. McLaughlin, Andrew R. Osborne Repetitive sequences in malaria parasite proteins. FEMS Microbiology Reviews. 2017;923–940, https://doi.org/10.1093/femsre/fux046 PMID: 29077880

69. Cohen S, McGREGOR IA, Carrington S. Gamma-globulin and acquired immunity to human malaria. Nature. 1961; 192: 733–737. https://doi.org/10.1038/192733a0 PMID: 13880318

70. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. Protein Sci A Publ Protein Soc. 2002; 11: 739–756. https://doi.org/10.1110/ps.4210102 PMID: 11910019

71. Olugbile S, Kulangara C, Bang G, Bertholet S, Suzarte E, Villard V, et al. Vaccine potentials of an intrinsically unstructured fragment derived from the blood stage-associated Plasmodium falciparum protein PFF0165c. Infect Immun. 2009; 77: 5701–5709. https://doi.org/10.1128/IAI.00652-09 PMID: 19786562

72. Yagi M, Bang G, Tougan T, Palacpac NMQ, Arisue N, Aoshi T, et al. Protective epitopes of the Plasmodium falciparum SERA5 malaria vaccine reside in intrinsically unstructured N-terminal repetitive sequences. PLoS One. 2014; 9: e98460. https://doi.org/10.1371/journal.pone.0098460 PMID: 24886718

73. Gunasekaran K, Tsai C-J, Kumar S, Zanuy D, Nussinov R. Extended disordered proteins: targeting function with less scaffold. Trends Biochem Sci. 2003; 28: 81–85. https://doi.org/10.1016/S0968-0004(03)00003-3 PMID: 12575995

74. Stanisic DI, Barry AE, Good MF. Escaping the immune system: How the malaria parasite makes vaccine development a challenge. Trends in Parasitology. 2013 Dec; 29(12):612–22. https://doi.org/10.1016/j.pt.2013.10.001 Epub 2013 Oct 29. PMID: 24176554.

75. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. Lancet. 2015; 386: 31–45. https://doi.org/10.1016/S0140-6736(15)60721-8 PMID: 25913272

76. Toenhake CG, Fraschka SA-K, Vijayabaskar MS, Westhead DR, van Heeringen SJ, Bártfai R. Chromatin Accessibility-Based Characterization of the Gene Regulatory Network Underlying Plasmodium falciparum Blood-Stage Development. Cell Host Microbe. 2018; 23: 557–569.e9. https://doi.org/10.1016/j.chom.2018.03.007 PMID: 29649445

77. Paing MM, Tolia NH. Multimeric Assembly of Host-Pathogen Adhesion Complexes Involved in Apicomplexan Invasion. Knoll LJ, editor. PLoS Pathog. 2014; 10: e1004120. https://doi.org/10.1371/journal.ppat.1004120 PMID: 24945143

78. Ntumngia FB, King CL, Adams JH. Finding the sweet spots of inhibition: understanding the targets of a functional antibody against Plasmodium vivax Duffy binding protein. Int J Parasitol. 2012; 42: 1055–1062. https://doi.org/10.1016/j.ijpara.2012.09.006 PMID: 23068913

79. Bigham AW, Magnaye K, Dunn DM, Weiss RB, Bamshad M. Complex signatures of natural selection at GYPA. Hum Genet. 2018; 137: 151–160. https://doi.org/10.1007/s00439-018-1866-3 PMID: 29362874

80. Wanaguru M, Liu W, Hahn BH, Rayner JC, Wright GJ. RH5-Basigin interaction plays a major role in the host tropism of Plasmodium falciparum. Proc Natl Acad Sci U S A. 2013; 110: 20735–20740. https://doi.org/10.1073/pnas.1320771110 PMID: 24297912

81. Malaria Genomic Epidemiology Network, Band G, Rockett KA, Spencer CCA, Kwiatkowski DP. A novel locus of resistance to severe malaria in a region of ancient balancing selection. Nature. 2015; 526: 253–257. https://doi.org/10.1038/nature15390 PMID: 26416757

82. Quiñones-Parra S, Loh L, Brown LE, Kedzierska K, Valkenburg SA. Universal immunity to influenza must outwit immune evasion. Front Microbiol. 2014; 5: 285. https://doi.org/10.3389/fmicb.2014.00285 PMID: 24971078

83. Kinyanjui SM, Mwangi T, Bull PC, Newbold CI, Marsh K. Protection against clinical malaria by heterologous immunoglobulin G antibodies against malaria-infected erythrocyte variant surface antigens requires interaction with asymptomatic infections. J Infect Dis. 2004; 190: 1527–1533. https://doi.org/10.1086/424675 PMID: 15478055

84. Marsh K, Howard RJ. Antigens induced on erythrocytes by P. falciparum: expression of diverse and conserved determinants. Science. 1986; 231: 150–153. https://doi.org/10.1126/science.2417315 PMID: 2417315

85. Doolan DL, Dobaño C, Baird JK. Acquired immunity to malaria. Clin Microbiol Rev. 2009; 22: 13–36, Table of Contents. https://doi.org/10.1128/CMR.00025-08 PMID: 19136431

86. Bruce MC, Galinski MR, Barnwell JW, Donnelly CA, Walmsley M, Alpers MP, et al. Genetic diversity and dynamics of Plasmodium falciparum and P. vivax populations in multiply infected children with asymptomatic malaria infections in Papua New Guinea. Parasitology. 2000; 121: 257–272. https://doi.org/10.1017/s0031182099006356 PMID: 11085246

87. Anders RF. Multiple cross-reactivities amongst antigens of Plasmodium falciparum impair the development of protective immunity against malaria. Parasite Immunol. 1986; 8: 529–539. https://doi.org/10.1111/j.1365-3024.1986.tb00867.x PMID: 3543808

88. Pavlović MD, Jandrlić DR, Mitić NS. Epitope distribution in ordered and disordered protein regions. Part B—Ordered regions and disordered binding sites are targets of T- and B-cell immunity. J Immunol Methods. 2014; 407: 90–107. https://doi.org/10.1016/j.jim.2014.03.027 PMID: 24726865

89. Adda CG, MacRaild CA, Reiling L, Wycherley K, Boyle MJ, Kienzle V, et al. Antigenic characterization of an intrinsically unstructured protein, Plasmodium falciparum merozoite surface protein 2. Infect Immun. 2012; 80: 4177–4185. https://doi.org/10.1128/IAI.00665-12 PMID: 22966050

90. Foquet L, Hermsen CC, van Gemert G-J, Van Braeckel E, Weening KE, Sauerwein R, et al. Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent Plasmodium falciparum infection. J Clin Invest. 2014; 124: 140–144. https://doi.org/10.1172/JCI70349 PMID: 24292709

91. Lopaticki S, Maier AG, Thompson J, Wilson DW, Tham W-H, Triglia T, et al. Reticulocyte and erythrocyte binding-like proteins function cooperatively in invasion of human erythrocytes by malaria parasites. Infect Immun. 2011; 79: 1107–1117. https://doi.org/10.1128/IAI.01021-10 PMID: 21149582

92. Reiling L, Boyle MJ, White MT, Wilson DW, Feng G, Weaver R, et al. Targets of complement-fixing antibodies in protective immunity against malaria in children. Nat Commun. 2019; 10: 1–13. https://doi.org/10.1038/s41467-018-07882-8 PMID: 30602773

93. Bell G., Agnandji S., Asante K., Ghansah A., Kamthunzi P., Emch M. and Bailey J. Impacts of Ecology, Parasite Antigenic Variation, and Human Genetics on RTS,S/AS01e Malaria Vaccine Efficacy. Current Epidemiology Reports. 2021; 8(3), pp.79–88.

94. Healer J, Thompson JK, Riglar DT, Wilson DW, Chiu Y-HC, Miura K, et al. Vaccination with conserved regions of erythrocyte-binding antigens induces neutralizing antibodies against multiple strains of Plasmodium falciparum. PLoS One. 2013; 8: e72504. https://doi.org/10.1371/journal.pone.0072504 PMID: 24039774

95. Mu J, Awadalla P, Duan J, McGee KM, Keebler J, Seydel K, et al. Genome-wide variation and identification of vaccine targets in the Plasmodium falciparum genome. Nat Genet. 2007; 39: 126–130. https://doi.org/10.1038/ng1924 PMID: 17159981

96. Valencia-Hernandez A., Ng W., Ghazanfari N., Ghilas S., de Menezes M., Holz L., Huang C., English K., Naung M., et al. A Natural Peptide Antigen within the Plasmodium Ribosomal Protein RPL6 Confers Liver TRM Cell-Mediated Immunity against Malaria in Mice. Cell Host & Microbe. 2020 27 (6), pp.950–962.e7. https://doi.org/10.1016/j.chom.2020.04.010 PMID: 32396839

97. Shah Z, Naung M, Moser KA, Adam M, Buchwald AG, Dwivedi A., et al. Whole-genome analysis of Malawian *Plasmodium falciparum* isolates identifies potential targets of allele-specific immunity to clinical malaria. PLOS Genetics. 2021; 17(5):e1009576. https://doi.org/10.1371/journal.pgen.1009576 PMID: 34033654

98. Tichkule S, Myung Y, Naung M, et al. VIVID: a web application for variant interpretation and visualisation in multidimensional analyses bioRxiv [Preprint]. 2021 [cited 2022 Jan 10]. Available from: https://www.biorxiv.org/content/10.1101/2021.11.16.468904v1

99. Nakakana U.N., Mohammed I.A., Onankpa B.O. et al. A validation of the Malaria Atlas Project maps and development of a new map of malaria transmission in Sokoto, Nigeria: a cross-sectional study using geographic information systems. Malaria Journal. 2020. https://doi.org/10.1186/s12936-020-03214-8 PMID: 32268904

100. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w ^{1118}; iso-2; iso-3. Fly (Austin). 2012; 6: 80–92. https://doi.org/10.4161/fly.19695 PMID: 22728672

101. DePristo MA, Banks E, Poplin R, Garimella K V, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43: 491–498. https://doi.org/10.1038/ng.806 PMID: 21478889

102. Nei M, Li WH. Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci. 1979; 76: 5269–5273. https://doi.org/10.1073/pnas.76.10.5269 PMID: 291943

103. Nei M, Tajima F. DNA polymorphism detectable by restriction endonucleases. Genetics. 1981; 97: 145–163. https://doi.org/10.1093/genetics/97.1.145 PMID: 6266912

104. Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, et al. History Population and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. Cardon L, editor. PLoS Biol. 2004; 2: e286. https://doi.org/10.1371/journal.pbio.0020286 PMID: 15361935

105. Stahl EA, Dwyer G, Mauricio R, Kreitman M, Bergelson J. Dynamics of disease resistance polymorphism at the Rpm1 locus of Arabidopsis. Nature. 1999; 400: 667–671. https://doi.org/10.1038/23260 PMID: 10458161

106. Leigh JW, Bryant D. popart: full-feature software for haplotype network construction. Nakagawa S, editor. Methods Ecol Evol. 2015; 6: 1110–1116. https://doi.org/10.1111/2041-210X.12410

107. Tien MZ, Meyer AG, Sydykova DK, Spielman SJ, Wilke CO. Maximum Allowed Solvent Accessibilites of Residues in Proteins. PLOS ONE. 2013; 8: e80635. https://doi.org/10.1371/journal.pone.0080635 PMID: 24278298

108. Webb B, Sali A. Comparative Protein Structure Modeling Using MODELLER. Curr Protoc Bioinforma. 2016; 54: 5.6.1–5.6.37. https://doi.org/10.1002/cpbi.3 PMID: 27322406