



MONASH University



Burnet Institute
Medical Research. Practical Action.

Computational approaches for examining the impact of protein structure on malaria immunity

Andrew John Guy

BSc(ScSchProg)(Hons)

A thesis submitted for the degree of Doctor of Philosophy at

Monash University in 2018

Burnet Institute

Department of Immunology and Pathology

COPYRIGHT NOTICE

© Andrew John Guy (2018).

I certify that I have made all reasonable efforts to secure copyright permissions for third-party content included in this thesis and have not knowingly added copyright content to my work without the owner's permission.

*‘Knowledge is invariably a matter of degree: you cannot put your
finger upon even the simplest datum and say “this we know.”’*

- T. S. Eliot

ABSTRACT

Malaria remains a major global health burden, and there is a need to develop an effective vaccine to fast track malaria elimination efforts. There is a vast amount of genome and proteome level data for *Plasmodium* species but there is a need for bioinformatics tools and approaches which can identify immune targets within *Plasmodium* species. Protein structure is a major factor in the recognition of antigen by the adaptive immune system, and this thesis seeks to integrate multiple data sources to explore the role of antigen 3-dimensional (3D) structure in the generation of adaptive immune responses against the malaria parasite.

The aims of this thesis include i) investigating the role of intrinsically disordered protein (IDP) antigens from *Plasmodium* species as targets of adaptive immune, ii) performing a proteome-wide computational analysis of structured malaria proteins to identify key determinants of immune recognition and selection pressure, and iii) development of novel computational approaches to integrate protein structural information into measures of immune selection pressure, and application of these approaches to leading malaria vaccine candidates.

IDPs are a class of proteins that exist as a dynamic ensemble of structurally distinct species, and are particularly enriched within apicomplexan parasites. In Chapter 2, several computational approaches were used to explore if regions of protein disorder were also predicted to be involved in adaptive immune responses. IDPs were shown to be particularly enriched within immunologically-exposed subcellular compartments of *P. falciparum*. Additionally, tandem repeat regions and non-synonymous single nucleotide polymorphisms were found to be strongly associated with regions of disorder. Importantly, IDP regions were shown to contain a paucity of major histocompatibility (MHC) class I- and II-binding peptides, potentially restricting recognition of IDP antigen by CD4+ and CD8+ T-cells.

The role of protein 3D structure in immunity against malaria was examined in Chapter 3 with experimentally characterised and modelled protein structures across the *P. falciparum* proteome. A key finding of this study was the strong propensity for polymorphic residues to be surface exposed and enriched within certain secondary structure elements. Predicted MHC class II binding peptides were mapped onto protein 3D structures. These MHC binding peptides were primarily buried within the core of the protein and in general were not polymorphic. A novel 3D sliding window approach was also used to identify regions within leading vaccine candidates that are likely to be under immune-mediated selection pressure. This 3D sliding window approach was developed into a Python package called BioStructMap, which is made available via an online web interface, and presented in Chapter 4. The BioStructMap tool was further applied in Chapter 5 to two leading

vaccine candidates from *P. vivax*: PvAMA1 and PvDBP. Selection pressures on these antigens were investigated across multiple geographic locations, with similar structural patterns of diversity across most populations.

This work has identified new bioinformatics approaches that highlight the prevalence and importance of IDPs as well as new approaches for mapping critical characteristics onto 3D structures. These approaches have been applied to specific vaccine candidates, highlighting important regions to assist with vaccine design.

PUBLICATIONS DURING ENROLMENT

Publications produced during candidature relevant to the thesis

Guy, A. J., V. Irani, C. A. MacRaild, R. F. Anders, R. S. Norton, J. G. Beeson, J. S. Richards, and P. A. Ramsland. 2015. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *PLoS One* 10: e0141729.

Guy, A. J., V. Irani, J. G. Beeson, B. Webb, A. Sali, J. S. Richards, and P. A. Ramsland. 2018. Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures. *Sci Rep.* 8: 4355.

Guy, A. J., V. Irani, J. S. Richards, and P. A. Ramsland. 2018. Structural patterns of selection and diversity for *Plasmodium vivax* antigens DBP and AMA1. *Malar J.* 17: 183.

Additional publications during candidature

Charnaud, S. C., R. McGready, A. Herten-Crabb, R. Powell, **A. Guy,** C. Langer, J. S. Richards, P. R. Gilson, K. Chotivanich, T. Tsuboi, D. L. Narum, M. Pimanpanarak, J. A. Simpson, J. G. Beeson, F. Nosten, and F. J. I. Fowkes. 2016. Maternal-foetal transfer of *Plasmodium falciparum* and *Plasmodium vivax* antibodies in a low transmission setting. *Sci. Rep.* 6: 20859.

Irani, V., **A. J. Guy,** D. Andrew, J. G. Beeson, P. A. Ramsland, and J. S. Richards. 2015. Molecular properties of human IgG subclasses and their implications for designing therapeutic monoclonal antibodies against infectious diseases. *Mol. Immunol.* 67: 171–182.

Irani, V., P. A. Ramsland, **A. J. Guy,** P. M. Siba, I. Mueller, J. S. Richards, and J. G. Beeson. 2015. Acquisition of Functional Antibodies That Block the Binding of Erythrocyte-Binding Antigen 175 and Protection Against *Plasmodium falciparum* Malaria in Children. *Clin. Infect. Dis.* 61: 1244–1252.

THESIS INCLUDING PUBLISHED WORKS DECLARATION

I hereby declare that this thesis contains no material which has been accepted for the award of any other degree or diploma at any university or equivalent institution and that, to the best of my knowledge and belief, this thesis contains no material previously published or written by another person, except where due reference is made in the text of the thesis.

This thesis includes three original papers published in peer reviewed journals and one submitted publication. The core theme of the thesis is the relationship between protein structure and adaptive immune responses to the malaria parasite. The ideas, development and writing up of all the papers in the thesis were the principal responsibility of myself, the student, working within the Burnet Institute and the Department of Immunology and Pathology, Monash University under the supervision of Paul A. Ramsland, Jack S. Richards and James G. Beeson.

(The inclusion of co-authors reflects the fact that the work came from active collaboration between researchers and acknowledges input into team-based research.)

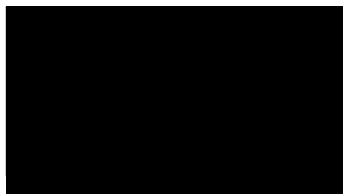
In the case of Chapters 2, 3, 4 and 5 my contribution to the work involved the following:

Thesis Chapter	Publication Title	Status	Nature and % of student contribution	Co-author name(s) Nature and % of Co-author's contribution	Co-author(s), Monash student Y/N
2	Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions	Published	65%. Concept, performed experiments, analysed data, wrote manuscript	<p>Vashti Irani: input 5%, critical analysis of the data and the manuscript</p> <p>Christopher A. MacRaild: input 4%, critical analysis of the data and the manuscript</p> <p>Robin F. Anders: input 2%, critical analysis of the data and the manuscript</p> <p>Raymond S. Norton: input 2%, critical analysis of data and the manuscript</p> <p>James G. Beeson: input 2%, critical analysis of the data and the manuscript</p> <p>Jack S. Richards: input 10%, Concept, research supervision, manuscript preparation</p>	<p>No</p> <p>No</p> <p>No</p> <p>No</p> <p>No</p> <p>No</p>

				Paul A. Ramsland: input 10%, Concept, research supervision, manuscript preparation	No
3	Proteome-wide mapping of immune features onto <i>Plasmodium</i> protein three-dimensional structures	Published	70%. Concept, performed experiments, analysed data, wrote manuscript	<p>Vashti Irani: input 5%, critical analysis of the data and the manuscript</p> <p>Benjamin Webb: input 2%, contributed data, commented on manuscript</p> <p>Andrej Sali: input 2%, contributed data, commented on manuscript</p> <p>James G. Beeson: input 1%, commented on manuscript</p> <p>Jack S. Richards: input 10%, Concept, research supervision, manuscript preparation</p> <p>Paul A. Ramsland: input 10%, Concept, research supervision, manuscript preparation</p>	<p>No</p> <p>No</p> <p>No</p> <p>No</p> <p>No</p>
4	BioStructMap: A Python tool for integration of protein structure and sequence-based features	Returned for revision	85%. Concept, implementation, wrote manuscript	<p>Vashti Irani: input 5%, manuscript preparation</p> <p>Jack S. Richards: input 5%, research supervision, manuscript preparation</p> <p>Paul A. Ramsland: input 5%, research supervision, manuscript preparation</p>	<p>No</p> <p>No</p> <p>No</p>
5	Structural patterns of selection and diversity for <i>Plasmodium vivax</i> antigens DBP and AMA1	Published	85%. Concept, performed experiments, analysed data, wrote manuscript	<p>Vashti Irani: input 5%, manuscript preparation</p> <p>Jack S. Richards: input 5%, research supervision, manuscript preparation</p> <p>Paul A. Ramsland: input 5%, research supervision, manuscript preparation</p>	<p>No</p> <p>No</p> <p>No</p>

I have renumbered sections of submitted or published papers in order to generate a consistent presentation within the thesis.

Student signature:

A large black rectangular box redacting the student's signature.

Date: 10 Jan 2018

The undersigned hereby certify that the above declaration correctly reflects the nature and extent of the student's and co-authors' contributions to this work. In instances where I am not the responsible author I have consulted with the responsible author to agree on the respective contributions of the authors.

Main Supervisor signature:

A large black rectangular box redacting the main supervisor's signature.

Date: 10 Jan 2018

ACKNOWLEDGMENTS

This research was supported by an Australian Government Research Training Program (RTP) Scholarship. I would also like to thank Monash University and the Burnet Institute for giving me the opportunity to pursue a PhD.

I am deeply grateful for the support, encouragement and dedication of my supervisors: Paul Ramsland, Jack Richards and James Beeson. Paul, thank you for your passion for scientific knowledge, your calm approach to everything, and for knowing how to walk the line between pragmatism and perfection. Your attitude towards both science and life is an inspiration, and I am deeply grateful for your continual support and encouragement over the last few years. Jack, thank you for your unquenchable enthusiasm and optimism for both basic science and the translational outcomes that come from it. Your faith in the people around you has always been appreciated, and I am immensely grateful for the numerous opportunities you have given me. James, thank you for your wisdom and seemingly endless knowledge of all things malaria.

A special thanks to Vashti Irani, my desk-neighbour and scientific co-conspirator during my time at the Burnet. Your infectious enthusiasm and willingness to go above-and-beyond for everyone is something to aspire to. Thank you for the countless conversations and collaborations. A big thanks to Alicia Chenoweth and Jessica Anania for the many tea breaks which undoubtedly helped keep me sane during the course of the PhD. Thank you also to Alistair McLean, Xi Zen Yap, Kerry Moore, Katherine O'Flaherty, Liriye Kurtovic and Leanna Surrao for sharing in both the joys and tribulations of PhD life.

Thank you to all of the members of the Richards, Beeson and Fowkes Labs, past and present. From serious scientific discussions during lab meetings, to not-so-serious discussions over cake, I have always appreciated both the scientific insights and the excellent humour that everyone brings to the table.

Thank you to others at the Burnet who have also contributed to my scientific development over my time there. A particular thanks to Paul Sanders, Bruce Wines and Dyson Simmons for their willingness to provide help and advice on a range of topics.

I am also grateful to the various scientific collaborators who have contributed to my work during my PhD: Andrej Šali, Ben Webb, Robin Anders, Ray Norton and Chris MacRaild. Thank you for the many insightful comments and suggestions along the way.

To my friends, family and housemates who have both supported me and put up with me over the last few years, thank you for the meals, conversations, games and various adventures.

I would particularly like to thank my parents, Peta and Tony, for fostering a sense of curiosity about the world, and always encouraging me in whatever I chose to do. Thank you for the endless love and support over all the years.

And to Amy, thank you for your love, cheerfulness and endless encouragement.

Finally, I am immensely grateful to the countless people whose work this thesis builds upon.

ABBREVIATIONS

3D	3-Dimensional
ACT	Artemisinin-based Combination Therapy
AMA1	Apical Membrane Antigen 1
ASA	Accessible Surface Area
BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of Protein Structure
CeTOS	Cell-Traversal Protein for Ookinetes and Sporozoites
CSP	Circumsporozoite Protein
DI/II/III	Domain I/II/III
DARC	Duffy Antigen/Receptor for Chemokines
DBL	Duffy Binding-like
DBP	Duffy Binding Protein
EBA	Erythrocyte Binding Antigen
EBL	Erythrocyte Binding-like
FRET	Fluorescence Resonance Energy Transfer
FTIR	Fourier-transform Infrared Spectroscopy
G6PD	Glucose-6-phosphate-dehydrogenase
GLURP	Glutamate-rich Protein
HIV	Human Immunodeficiency Virus
HKA test	Hudson-Kreitman-Aguade test
HLA	Human Leukocyte Antigen
IDP	Intrinsically Disordered Protein
IDPR	Intrinsically Disordered Protein Region
IgG	Immunoglobulin G
LSA	Liver Stage Antigen
MAF	Minor Allele Frequency

MHC	Major Histocompatibility Complex
MK test	McDonald-Kreitman test
MPQS	ModPipe Quality Score
MS	Mass Spectrometry
MSP	Merozoite Surface Protein
MSPDBL1/2	Merozoite Surface Protein Duffy Binding-like 1/2
NANP	Asparagine-Alanine-Asparagine-Proline
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
PfEMP1	<i>Plasmodium falciparum</i> Erythrocyte Membrane Protein 1
PfPI3K	<i>Plasmodium falciparum</i> Phosphatidylinositol-3-kinase
PIESP2	Parasite-infected Erythrocyte Surface Protein 2
PV	Parasitophorous Vacuole
RII	Region II
RIII-V	Region III-V
RVI	Region VI
RBC	Red Blood Cell
RBL	Reticulocyte Binding-like
RON	Rhoptry Neck
RSA	Relative Solvent Accessibility
SAXS	Small-angle X-ray Scattering
SNP	Single Nucleotide Polymorphism
TRAP	Thrombospondin-related Adhesive Protein
TSR	Thrombospondin-like Type Repeat
VSA	Variable Surface Antigen
WHO	World Health Organization

TABLE OF CONTENTS

Abstract	iv
Publications during Enrolment	vi
Thesis Including Published Works Declaration	vii
Acknowledgments	x
Abbreviations	xii
Table of Contents	xiv
Chapter 1: Introduction	1
1.1 Malaria Background	1
1.1.1 Malaria distribution, prevalence and economic impact	1
1.1.2 Parasite life cycle	2
1.1.3 Treatment and therapy	7
1.1.4 Parasite drug resistance	7
1.1.5 Vaccine development	8
1.2 Stage-specific Immune Responses	8
1.2.1 Pre-erythrocytic stages	8
1.2.2 Blood stage infection	9
1.2.2.1 Antibody responses against the infected red blood cell	10
1.2.2.2 Antibody responses against merozoites	10
1.2.2.3 Cellular immune responses against blood stage malaria	12
1.2.3 Sexual stages	13
1.3 Intrinsically Disordered Proteins	14
1.3.1 Disordered proteins overview	14
1.3.2 Antibody responses against disordered proteins	16
1.3.2.1 Conformation of IDPs when bound to antibodies	16
1.3.2.2 Antibody affinity against IDPs	17
1.3.2.3 Entropic considerations of antibody binding	18

1.3.3 Cellular immune responses against disordered proteins	19
1.4 Identifying Immune Targets in Structured Malaria Antigens	20
1.4.1 Antibody-antigen interactions for malaria antigens	20
1.4.2 Genetic markers of immune selection pressure	25
1.4.3 Identification of immune pressure on malaria genes	26
1.4.4 Integration of protein structural information with polymorphism data	27
1.4.4.1 <i>Pf</i> AMA1	27
1.4.4.2 EBL family proteins	28
1.4.4.3 VAR2CSA	28
1.4.4.4 <i>Pf</i> CSP	29
1.4.4.5 MSPDBL1 & MSPDBL2	29
1.4.4.6 TRAP	29
1.4.5 Current tools for examining protein structure and applicability to immune-related polymorphisms	30
1.5 Thesis Hypothesis and Aims	31
1.6 Summary of Chapters	32
1.7 References	32
Chapter 2: The Impact of Protein Disorder on Adaptive Immunity	50
2.1 Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Prediction	53
2.2 Supplementary Figures and Tables	75
Chapter 3: Structural Features of <i>Plasmodium</i> Antigens: Linking 3D Structure to Immune Mediated Selection Pressure	84
3.1 Proteome-wide Mapping of Immune Features onto <i>Plasmodium</i> Protein Three-Dimensional Structures	87
3.2 Supplementary Figures and Tables	103

Chapter 4: Tools for Spatial Aggregation of Protein Data	118
4.1 BioStructMap: A Python Tool for Integration of Protein Structure and Sequence-based Features	120
4.2 Supplementary Figures	123
4.3 BioStructMap Usage Guide	124
Chapter 5: Selection Pressures on Key <i>P. vivax</i> Antigens: A Structural Perspective	138
5.1 Structural Patterns of Selection and Diversity for <i>Plasmodium vivax</i> antigens DBP and AMA1	141
5.2 Supplementary Figures	156
Chapter 6: Discussion	166
6.1 Summary of findings	166
6.2 Experimental validation of T-cell responses against IDPs	169
6.3 Limitations and future directions for B-cell epitope prediction in IDPs	170
6.4 Integrating structural features into tests of immune selection pressure	171
6.5 Development of an online platform to explore structure and immunology data for <i>Plasmodium</i> species	172
6.6 Concluding remarks	175
6.7 References	175

1. INTRODUCTION

1.1 Malaria Background

Malaria is an infectious, mosquito-borne disease responsible for an estimated 445,000 deaths globally in 2016 [1]. Malaria is caused by various apicomplexan *Plasmodium* species, with *P. falciparum* the major contributor to global malaria mortality. *P. vivax* also contributes significantly to the global burden of disease. Other species capable of infecting humans include *P. ovale*, *P. malariae* and *P. knowlesi*. The clinical symptoms of malaria range from mild to potentially life threatening, including periodic fever, anaemia, chills, headache, fatigue, convulsions and coma. The current recommended treatment for uncomplicated *P. falciparum* malaria is artemisinin-based combination therapy (ACT). The development of drug-resistant malaria parasites is a major concern, with evidence of resistance to artemisinin based therapies throughout Southeast Asia [2,3]. Within a malaria endemic setting, clinical symptoms and associated complications are worst in young children and pregnant women. The gradual development of partial immunity following repeated exposure protects most individuals, while pregnant women are particularly susceptible due to pregnancy-related immunomodulation and parasite sequestration in the placenta [4]. Immunity to clinical malaria develops slowly, and is believed to be partly dependent on antibody responses [5–7]. The exact determinants of this clinical protection are unclear, although it is likely that a complex mix of antibody responses against a number of antigens is required for effective protection from clinical malaria. The high burden of disease, together with increasing evidence of drug resistance against current antimalarial drugs, highlights the urgent need for an efficacious and cost-effective vaccine against malaria.

1.1.1 Malaria distribution, prevalence and economic impact

There were an estimated 216 million cases of malaria in 2016, with roughly 445,000 malaria deaths [1]. Of these, approximately 91% occurred within the World Health Organization (WHO) African region. Importantly, there has been a significant reduction in malaria morbidity and mortality in the last ten years, considering there were over 1 million estimated malaria deaths in 2005 [8]. Aside from the direct impact on human health and wellbeing, the economic impact of malaria disease is considerable. It has been estimated that malaria costs African economies a total of US\$12 billion per year [9]. The economic impact outside of the African region is also significant, with estimates that malaria costs India roughly US\$100 million annually [10].

The majority of malaria cases in the WHO African region are attributed to *P. falciparum*, whereas *P. vivax* is much more prevalent within South-East Asia and South America [1,11] (**Figure 1.1**). Historically, the global distribution of malaria was much wider, encompassing much of Europe, North-America and Northern Australia [11,12]. A combination of increased sanitation and development, alongside dedicated eradication programs has restricted the global distribution of malaria significantly. An ambitious WHO malaria elimination campaign in the 1950's-1960's involving mass insecticide spraying and chemoprophylaxis yielded large reductions in malaria incidence in numerous countries including India and Sri Lanka, with elimination of endemic malaria from a total of 37 countries worldwide by 1978 [13]. However, this campaign failed to address malaria in most of sub-Saharan Africa, and many countries that saw large reductions in malaria incidence during this campaign have seen a subsequent resurgence in malaria incidence in the absence of concerted control efforts.

Encouragingly, recent efforts to address the impact of malaria have been relatively successful. There has been a renewed push towards malaria control, elimination and eventual global eradication [14], beginning with the World Health Organization's Roll Back Malaria initiative announced in 1998 [15]. Target 6C of the Millennium Development Goals was to "have halted by 2015 and begun to reverse the incidence of malaria and other major diseases", and this has been successfully achieved. This current success in malaria control has been achieved using a combination of insecticide-treated bed nets, indoor residual spraying, vector control, mass drug administration, and ready access to screening and treatment of both clinical and subclinical malaria [16]. Moving forward, Goal 3 of the United Nations Sustainable Development Goals includes the goal of ending the global epidemic of malaria by 2030. A detailed pathway towards this is outlined in the WHO *Global Technical Strategy for Malaria 2015-2030* which aims to reduce global malaria case incidence by 90% by 2030 compared to 2015 levels, as part of a strategy towards malaria global elimination [14]. However, it is clear that global malaria elimination will not be achieved with the current tools available, and development of an efficacious malaria vaccine is likely to be required for a successful elimination strategy.

1.1.2 Parasite life cycle

The malaria life cycle is complex, and involves both human and mosquito stages (**Figure 1.2**). Within the human host, there are well-defined pre-erythrocytic, erythrocytic and sexual stages. When an infected *Anopheles* mosquito bites a human host, a number of malaria sporozoites are released from the salivary glands of the mosquito and enter the dermis (although some sporozoites may enter the epidermis or subcutaneous tissue, depending on the injection site) of the human host [20,21]. The number of sporozoites released varies, but is believed to be less than 100 in most

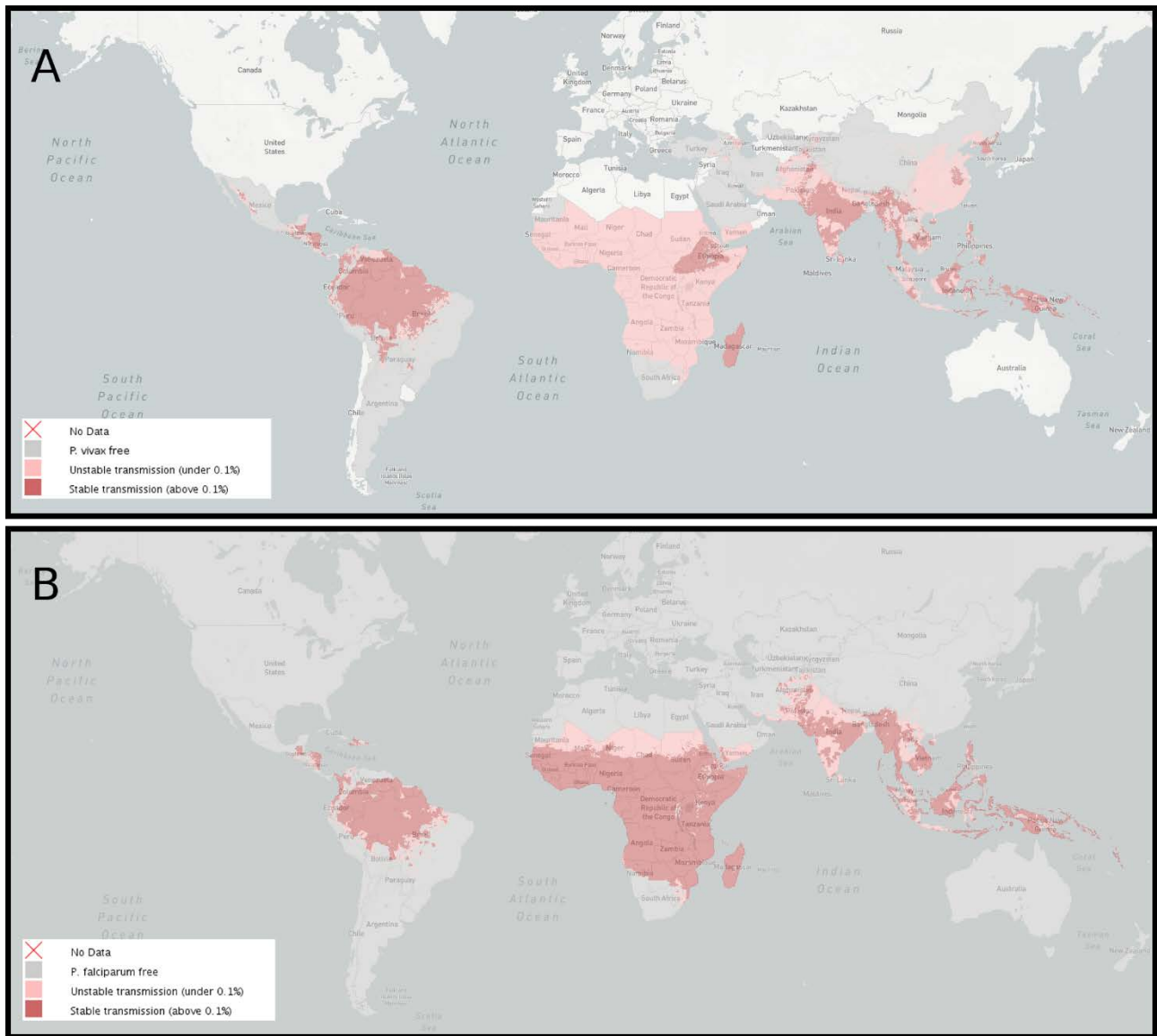


Figure 1.1: Worldwide malaria transmission for *P. vivax* (A) and *P. falciparum* (B) in 2010. Areas of stable malaria transmission (local annual case incidence greater than 0.1%) are shown in dark red. Areas of unstable malaria transmission (local annual case incidence less than 0.1%) are shown in light red. Maps were obtained from the Malaria Atlas Project (<https://map.ox.ac.uk/>) [17], with *P. vivax* and *P. falciparum* transmission data from Gething *et al.* (2012) [18] and Gething *et al.* (2011) [19] respectively.

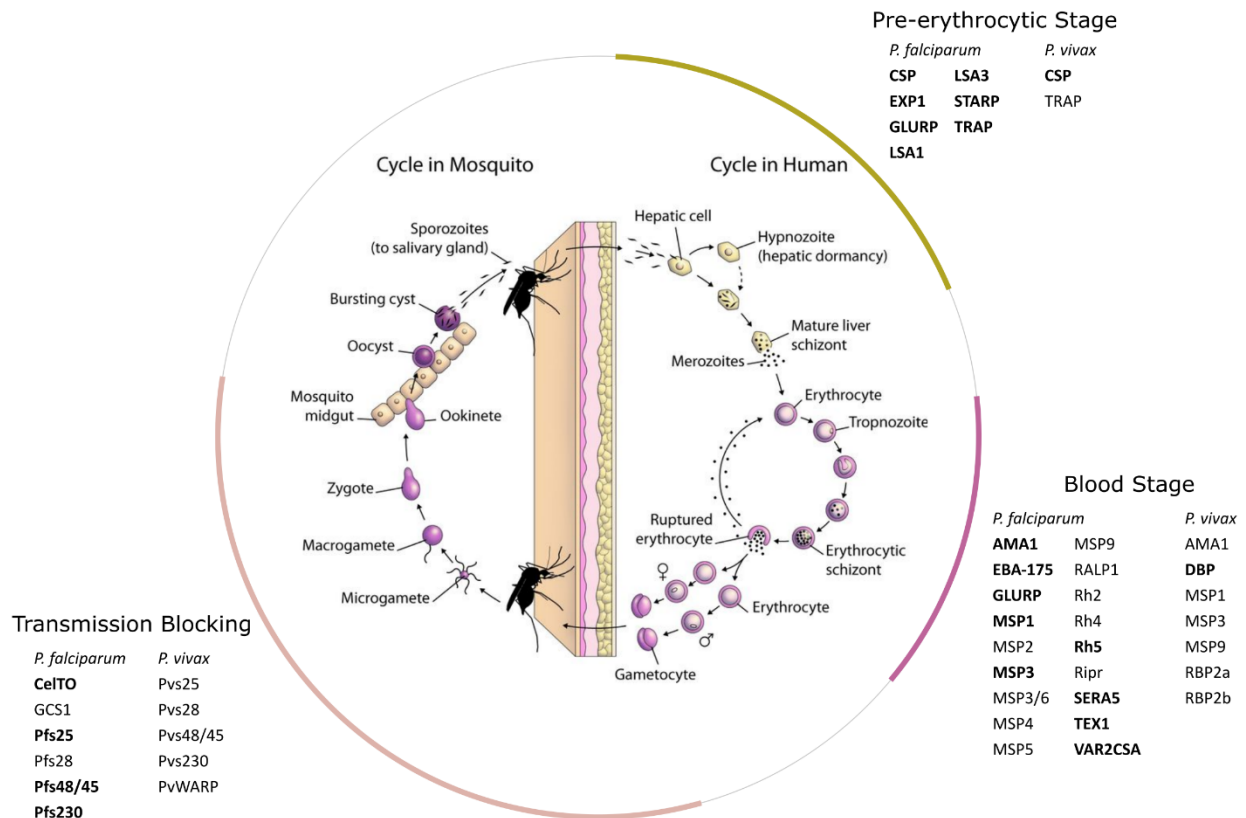


Figure 1.2: Malaria parasite life cycle and associated vaccine candidates for *P. falciparum* and *P. vivax* at each life cycle stage. Antigens listed in the WHO Malaria Vaccine Rainbow Tables [27] as currently in clinical or preclinical development are shown in bold (updated 17/7/17). For further information on stage-specific vaccine candidates refer to [28]. Image adapted from "Life cycle of the malaria parasite" from Epidemiology of Infectious Diseases. Available at: <http://ocw.jhsph.edu>. Copyright © Johns Hopkins Bloomberg School of Public Health. Creative Commons BY-NC-SA.

cases [22]. From the initial site of entry, sporozoites migrate to the circulation, where they eventually reach the liver and begin invasion of liver hepatocytes. This is followed by a period of asexual replication within hepatocytes, which occurs over 9-14 days for *P. falciparum* or 12-17 days for *P. vivax* [23]. This stage of the lifecycle is largely asymptomatic for the human host. Following this period of asexual replication, infected liver cells then release large numbers of merozoite-filled merozoites into the liver sinusoids [24]. These merozoites are formed through budding of the hepatocyte membrane, encasing large numbers of merozoites, and are believed to minimise host immune responses against the newly released merozoites [25]. Once in the circulation, merozoites rapidly invade red blood cells. Upon entering a red blood cell, a merozoite undergoes maturation and division over a 48 hour period (72 hours for *P. malariae*; 24 hours for *P. knowlesi*), before the red blood cell ruptures and releases new merozoites into the circulation [26]. This begins another cycle of red blood cell invasion, replication and subsequent red blood cell rupture which characterizes the symptomatic blood-stage of the *Plasmodium* life cycle.

Merozoite invasion of red blood cells is a tightly coordinated, multi-stage process (reviewed in detail in [29]). After initial contact with a red blood cell, merozoites undergo reorientation such that the apical end of the parasite is in contact with the red blood cell membrane. This is followed by formation of a tight junction between the merozoite and red blood cell, and this tight junction moves along the body of the merozoite as it enters the red blood cell [30]. Once the parasite has entered the red blood cell, the red blood cell undergoes transient echinocytosis, characterised by dramatic membrane morphological changes with the formation of many spiky projections in the red cell membrane, before returning to its normal shape within a few minutes [31]. The initial infected red blood cell is known as a ring stage parasite. Over the next 48 hours the parasite begins a process of asexual replication, moving from ring stage to trophozoite to schizont. Once schizonts are matured, they rupture, releasing large numbers of merozoites that then proceed to infect fresh red blood cells. This cycle of invasion, replication and rupture is responsible for an exponential increase in parasitemia that causes many of the clinical symptoms associated with malaria. A small proportion of ring stage parasites commit to becoming sexual stage gametocytes, and are taken up by a mosquito vector that bites the human host (reviewed in [32]). During the mosquito stage of the life cycle, the female and male gametocytes mate and form ookinetes that further mature into oocysts that embed in the mosquito midgut lumen. The oocysts then rupture and release sporozoites into the salivary glands of the mosquito that can infect another person during a subsequent mosquito blood meal, continuing the life cycle and transmission [33].

A number of GPI-anchored merozoite surface proteins are thought to be involved in mediating initial attachment of the merozoite to red blood cells, including merozoite surface protein 1 (MSP1),

MSP6 and MSP7 [34]. Following this initial attachment, the erythrocyte binding-like (EBL) and reticulocyte binding-like (RBL) families of proteins are involved in further binding events, with the merozoite also undergoing reorientation so the apical end of the parasite is adjacent to the red blood cell membrane [35]. The EBL family of proteins include erythrocyte binding antigen 175 (EBA-175), EBA-140, EBA-181 and EBL-1 within *P. falciparum*, and the Duffy-binding protein (DBP) protein within *P. vivax* [36]. In *P. falciparum*, the EBL family of proteins share a common domain architecture, with region II (RII) containing tandem DBL domains termed F1 and F2, regions III-V (RIII-V) predicted to be intrinsically disordered, a cysteine-rich structured region VI (RVI), a single transmembrane domain, and a cytoplasmic C-terminal tail. EBL family proteins bind to host cell surface glycoproteins via the DBL domain(s) contained within RII: EBA-175 recognises Glycophorin A [37], EBA-140 recognises Glycophorin C [38], the receptor for EBA-181 is unknown, EBL-1 binds to Glycophorin B [39], while PvDBP binds the Duffy antigen/receptor for chemokines (DARC) on reticulocytes [40]. The RBL family of proteins also plays a role during this attachment process, with binding of *Pf*RH5 to basigin on the red blood cell a critical step in the invasion process [41]. *Pf*RH5 is likely anchored to the parasite membrane via the GPI-anchored merozoite protein P133 [42]. Following binding of *Pf*RH5 to basigin, apical membrane antigen 1 (AMA1) binds to RON2 on the erythrocyte membrane to initiate formation of a tight junction between parasite and host-cell membranes. RON2 is a rhoptry protein that forms part of the rhoptry neck (RON) complex, which is translocated into the red blood cell membrane prior to binding to AMA1 [43]. AMA1 is a type 1 transmembrane protein which contains three extracellular domains termed Domain I (DI), II (DII) and III (DIII) [44]. DI contains a conserved hydrophobic binding groove that is the site of RON2 binding [45]. DI also contains a large number of polymorphic residues, presumed to be the result of immune pressure on this region [45,46].

Features of blood stage infection which contribute to clinical symptoms include red blood cell rosetting and adhesion to vasculature, which can result in vascular occlusion. This is driven by presentation of adhesins such as *P. falciparum* erythrocyte membrane protein 1 (PfEMP1) on the surface of infected red blood cells which promote sequestration of infected cells in the vasculature, helping the parasite avoid splenic clearance. PfEMP1 is encoded by ~60 *var* genes, with only a single gene typically expressed at a single time. There is a significant level of diversity between different PfEMP1 variants, which limits the ability to generate cross-reactive antibodies that recognise multiple PfEMP1 variants [47]. PfEMP1 is also implicated in the development of cerebral malaria, with a subset of *var* genes (Group A) associated with sequestration of parasites in the brain microvasculature and associated cerebral malaria pathology [48–50]. These sequestered parasites cause localised hypoxia, inflammation and tissue damage [51,52]. Sequestration of infected erythrocytes is also a particular issue during pregnancy associated malaria, as the expression of a

pregnancy-specific PfEMP1 variant called VAR2CSA allows sequestration of *P. falciparum* infected erythrocytes in the placenta. This is mediated by binding of VAR2CSA to Chondroitin Sulfate-A (CSA) on the placenta, and ultimately leads to a number of adverse outcomes for both the mother and child [47].

1.1.3 Treatment and therapy

The primary goal of malaria treatment is the complete elimination of parasites from the body, with the associated reduction and prevention of pathology associated with blood stage infection. This also has a broader public health benefit in preventing continued transmission. The current recommended treatment for uncomplicated *P. falciparum* malaria is one of several Artemisinin-based Combination Therapies (ACT): artemether and lumefantrine; artesunate and amodiaquine; artesunate and mefloquine; dihydroartemisinin and piperaquine; or artesunate and sulfadoxine-pyrimethamine [53]. For treatment of uncomplicated *P. vivax*, *P. ovale*, *P. malariae* or *P. knowlesi* infection, treatment with either an ACT or chloroquine is recommended. Additionally, to prevent relapse of *P. vivax* or *P. ovale* malaria as a result of dormant liver-stage parasites, an additional 14-day course of primaquine is recommended, provided the patient is not glucose-6-phosphate-dehydrogenase (G6PD) deficient. Finally, intravenous or intramuscular artesunate is recommended for treatment of severe malaria, until the patient can tolerate oral ACT therapy. Despite the increasing spread of artemisinin resistant parasites, ACT currently remains the recommended treatment option [53].

1.1.4 Parasite drug resistance

The emergence of drug resistant parasites poses a major public health issue. Whilst chloroquine was the drug of choice for many years for treatment of malaria, widespread chloroquine resistance amongst *P. falciparum* parasites has prompted a shift to ACT as a first-line treatment [54,55]. However, evidence of artemisinin resistance began to emerge in western Cambodia in 2008, with resistance now spread to large parts of mainland Southeast Asia [56]. Resistance to artemisinin based drugs has been associated with mutations in the *kelch13* gene. This mutation has been suggested to increase levels of an associated phosphatidylinositol-3-kinase (*PfPI3K*), which appears to be a primary target for artemisinin [57]. The effects of drug resistance are countered by administering combination therapies, limiting the ability of parasites to develop resistance to any single drug. However, there have also been reports of additional parasite resistance to partner drugs such as piperaquine, limiting the effectiveness of combination therapies [58]. The inevitable emergence and spread of drug-resistant parasites highlights the need to develop new drugs and an efficacious malaria vaccine.

1.1.5 Vaccine development

To date, only a single vaccine, RTS,S/AS01, has progressed to Phase III clinical trials. RTS,S is a recombinant protein vaccine formed by fusion of the *P. falciparum* circumsporozoite surface protein (CSP) with the Hepatitis B S-antigen, and has been paired with the AS01 adjuvant [59,60]. Results from Phase III trials suggest that over a 12-month period, RTS,S/AS01 reduced incidence of clinical malaria by 50.4% in children aged 5-17 months at enrolment [61]. Despite this success, there is a general consensus that greater efficacy is needed in a vaccine if global eradication of malaria is to be achieved [62–64]. Additionally, protection induced by the RTS,S vaccine appears to be short-lived, with a rapid decline over time [65]. Despite this limited efficacy, RTS,S has been approved for use in the Malaria Vaccine Implementation Programme in Malawi, Ghana and Kenya [66]. However, there remains a need for a long lasting, efficacious and affordable vaccine, especially if wide spread malaria eradication is to be achieved within a realistic time frame. A number of pre-erythrocytic, blood stage and sexual stage vaccines are currently in development [67–69], and key antigens that are currently being assessed in clinical or preclinical trials are shown in **Figure 1.2**. These include the *P. falciparum* blood stage antigens AMA1 and EBA-175, and the *P. vivax* blood stage antigen DBP, all of which are discussed in more detail in Section 1.2.2.

1.2 Stage-specific Immune Responses

Given the current lack of a highly efficacious vaccine for malaria, the continued development of both existing and new vaccine candidates is needed. An understanding of the nature of protective immune responses against the malaria parasite is therefore essential to guide the development of any new vaccine. This section discusses current knowledge regarding the nature of immune responses against various stages of the parasite life cycle.

1.2.1 Pre-erythrocytic stages

Experiments in both animals and humans involving vaccination with radiation-attenuated sporozoites have provided solid evidence that sterile immunity against pre-erythrocytic infection is possible, despite there being little evidence that this occurs naturally following typical exposure to malaria. Radiation-attenuated sporozoites are able to infect hepatocytes, but undergo aborted development within these hepatocytes, and are unable to progress to viable blood stage infection [70]. Although the exact determinants of sterile protective immunity following immunisation with radiation attenuated parasites are still unclear, CD8⁺ T-cells appear to be critical in the protective response, as ablation of CD8⁺ T-cells in both murine and human-primate models removes sterile immunity [71–74]. Furthermore, it appears that protection is being mediated at the liver stage,

rather than against sporozoites in the circulation, as adoptively transferred cytolytic CD8⁺ T-cells specific for CSP antigen were shown to act upon the hepatocytic stage of the lifecycle [75]. While CD8⁺ responses against epitopes from CSP are capable of mediating protection, there is also a role for other antigens such as the thrombospondin-related adhesive protein (TRAP) in protective cellular responses against the liver stage parasite [76]. Interestingly, responses against CSP do not seem to be absolutely required for induction of protective CD8⁺ T-cell immunity, although it is apparent that CSP is immunodominant in a radiation-attenuated parasite mouse model, and protection is reduced in the absence of anti-CSP responses [77]. There is also a potent antibody response against sporozoite surface antigens following immunisation with radiation-attenuated sporozoites, with a high proportion of the antibody response against CSP, although very high antibody responses against TRAP are also observed [78]. However, antibody responses against CSP are not required for sterile immunity following immunisation with radiation-attenuated parasites, and there is suggestion that the immunodominant nature of CSP may be a parasite strategy to direct antibody responses away from more productive targets [79]. Immunisation with RTS,S, which contains the structurally disordered tandem repeat region and structurally ordered C-terminal region of CSP, is moderately effective in preventing blood stage infection in the short-term, although protective immunity wanes quickly. The determinants of protection following RTS,S delivery appear to be both immunoglobulin G (IgG) responses and CD4⁺ T-cell responses [80,81]. It is interesting to note that unlike vaccination with radiation-attenuated sporozoites, vaccination with RTS,S/AS01 produces minimal CD8⁺ T-cell responses.

Other potentially important pre-erythrocytic antigens include TRAP, cell-traversal protein for ookinetes and sporozoites (CeTOS), RON-4 and liver-stage antigen 1 (LSA-1). TRAP is a sporozoite surface protein that is essential for sporozoite gliding motility [82,83], infection of hepatocytes [83–86], and invasion of mosquito salivary glands [87,88]. IgG responses against TRAP have been observed following natural infection [89], but appear to be minimal in areas of low or intermittent transmission [90–93]. Interestingly, induction of significant levels of anti-TRAP IgG and associated memory B-cells was not observed during experimental immunization with infectious sporozoites followed by chloroquine chemoprophylaxis, even after challenged with infectious sporozoites. However, elevated levels of anti-TRAP antibodies were observed following challenge with infected erythrocytes, suggesting an additional role/presence for TRAP during erythrocytic stages [94].

1.2.2 Blood stage infection

Natural immunity to clinical blood stage malaria develops following repeated exposure within a malaria endemic setting. This protection limits the progression of clinical disease, but does not

constitute sterilising immunity. Following natural exposure, individuals tend to develop protection from clinical malaria, but can still harbor a non-symptomatic level of parasitemia, which contributes heavily to transmission levels. This naturally-occurring, non-sterile immunity against symptomatic malaria develops with repeated exposure, and is directed against the blood stage of the malaria parasite. Antibody responses have been shown to constitute an important component of protective immunity against clinical malaria, as passive transfer of gamma-globulin from immune individuals from hyperendemic malaria areas was shown to effectively reduce clinical symptoms and accompanying parasitemia in children suffering from severe *P. falciparum* or *P. malariae* malaria [7].

1.2.2.1 Antibody responses against the infected red blood cell

Although a number of variable surface antigens (VSAs) are expressed on infected red blood cells, including PfEMP1, RIFIN, SURFIN and STEVOR proteins, PfEMP1 appears to be the major target of protective antibody responses to infected erythrocytes [95]. The high diversity between PfEMP1 variants limits the generation of broadly cross-reactive antibodies, and introduces significant challenges to the design of an effective vaccine targeting VSAs [47]. VAR2CSA is a member of the PfEMP1 family, and mediates attachment of infected erythrocytes to the placenta. There may be a role for a vaccine targeting VAR2CSA to limit the impact of pregnancy associated malaria, although this is also hampered by extensive sequence variability in the VAR2CSA gene [96]. There are currently two vaccine candidates for placental malaria in early phase Ia/b clinical trials, with the two vaccines using different domains of VAR2CSA [97,98].

1.2.2.2 Antibody responses against merozoites

Merozoite invasion of red blood cells is a tightly regulated, complex process, and our understanding of the immune response against this stage of the parasite life cycle is incomplete. However, a number of merozoite antigens have emerged as promising candidates for further vaccine development. Antibody responses against the invading merozoite have been reviewed extensively elsewhere [67], and we highlight salient points here.

Numerous studies have demonstrated the importance of antibodies against many merozoite antigens. EBA-175 RII is a prospective vaccine candidate, with a recombinant RII construct having undergone phase I clinical trials [99]. There is considerable evidence for the importance of antibodies targeting EBA-175 RII. Monoclonal antibodies (mAbs) against RII have been shown to inhibit merozoite invasion *in vitro* [100], with some of the most potent invasion inhibitory antibodies shown to inhibit engagement of EBA-175 RII with its Glycophorin A receptor via blocking of the dimerization and receptor binding interface [101]. Additionally, naturally occurring

antibodies against EBA-175 RII are capable of inhibiting parasite growth *in vitro* [102]. Furthermore, naturally occurring antibodies against RII that specifically inhibit engagement of RII with Glycophorin A were shown to be protective against symptomatic malaria [103]. There is also evidence for the importance of antibodies against other EBL family proteins, with antibodies against EBA-175, EBA-140 and EBA-181 from naturally exposed individuals shown to be important in mediating *in vitro* parasite growth inhibition using a series of protein knockout parasites that lack the respective invasion ligands [104]. Region III-V (RIII-V) of EBA-175 is also suggested to be an important target of protective antibody responses, with relatively low levels of circulating anti-EBA-175 RIII-V antibodies associated with protection from clinical malaria [105]. Furthermore, rabbit antibodies raised against recombinant EBA-175 RIII-V were shown to potently inhibit parasite invasion *in vitro* [106].

AMA1 is a major vaccine candidate, with two different vaccine constructs having been tested in phase II clinical trials [107,108], although with no evidence of significant protection against clinical malaria. A number of phase I trials have also been conducted with different AMA1 constructs (reviewed in [67]). Both polyclonal [46,109,110] and monoclonal [111] antibodies against AMA1 have been shown to inhibit invasion *in vitro*. Whilst there is substantial polymorphism within AMA1 [45], and cross-strain specificity has been suggested to be a hurdle for vaccine development [107], there is also evidence for substantial antigenic overlap between different *Pf*AMA1 strains [112], supporting the design of a multi-allele vaccine construct for AMA1. The most polymorphic region of AMA1 is DI [45], which is also the location of the RON2 binding cleft [113].

A number of population studies have been performed to identify protective antibody responses against merozoites, and these have been reviewed by Healer *et al.* [114]. In a systematic meta-analysis that examined associations between clinical *P. falciparum* malaria and antibodies against merozoite antigens, it was observed that antibodies against MSP3 (C-terminus) and MSP1₁₉ were strongly associated with reduced risk, whilst antibodies against AMA1 and glutamate-rich protein (GLURP) were associated to a lesser extent [115]. Antibodies against MSP2 and MSP1 N-terminal regions were not significantly associated with reduced risk of clinical malaria [115]. A study by Richards *et al.* examined antibody responses against 46 merozoite antigens in a prospective cohort of children from a malaria endemic region of Papua New Guinea [116]. This study identified a number of merozoite antigens which were strongly associated with protection from clinical malaria episodes, including a number of targets that have previously received little attention as vaccine candidates. These antigens included both EBA-140 RIII-V and EBA-175 RIII-V, both of which are predicted to be intrinsically disordered [117]. A prospective cohort study of Kenyan children measured associations between antibody responses to merozoite antigens and protection from

clinical malaria, with highest ranked antigens including MSP2, RhopH3, MSP11, MSP3, AMA1 and MSRP1 [118]. This study also noted a couple of previously uncharacterised merozoite antigens were highly associated with protection, including PF3D7_1136200 and PF3D7_0606800. Another study in Kenya used a protein microarray system to measure antibody responses against *P. falciparum* antigens [119]. In this study, the antibody responses against MSP10, MSP2, liver-stage antigen 3 (LSA3) and PfEMP1 had the highest associations with protection. Another protein microarray study in the Peruvian Amazon identified MSP1, HSP70, ring-infected erythrocyte surface antigen (RESA), LSA3, MSP11 and parasite-infected erythrocyte surface protein 2 (PIESP2) as the top antigens associated with protection from symptomatic malaria [120]. Whilst there appear to be a number of vaccine candidates that are repeatedly associated with protective immune responses, there are also considerable discrepancies between the various studies that have examined associations between antibody responses and protection from symptomatic malaria. Differences between these studies could be attributable to numerous factors, including differences in antigen quality and preparation, variation in experimental protocols between laboratories or real differences in parasite antigen usage between populations [114].

1.2.2.3 Cellular immune responses against blood stage malaria

The role of both CD4⁺ T-cells and CD8⁺ T-cells in protective responses against blood stage malaria is uncertain. Despite CD4⁺ T-cells playing an important role in mediating protective immune responses against blood stage infection [121–124], depletion of antigen-specific CD4⁺ T-cells is observed to occur during malaria infection [125]. Furthermore, CD4⁺ T-cell depletion has been shown to be mediated by CD4⁺CD25⁺ regulatory T-cells, and may serve as a mechanism to limit CD4⁺ T-cell associated pathology [126]. In contrast to the suppressed nature of CD4⁺ T-cell responses in normal *Plasmodium* infection, it has been shown that immunisation with low doses of killed blood stage parasites leads to protection from homologous and heterologous blood stage challenge in a CD4⁺ T-cell dependent manner [127]. Furthermore, vaccination with chemically attenuated blood stage parasites has also been shown to induce long-lived, CD4⁺ T-cell dependent protection, with this protection also dependent on red blood cell (RBC) membrane integrity being maintained during preparation of the attenuated parasites for vaccination [128]. Taken together, the role of CD4⁺ T-cells in malaria responses appears to be complex, and may serve to either enhance or limit pathology in a context dependent manner.

While it is widely accepted that CD8⁺ T-cells are crucial in the development of immunity to liver-stage infection, there is still a level of uncertainty and controversy regarding their importance in blood stage immunity. Due to the lack of major histocompatibility complex (MHC) class I expression on RBCs, CD8⁺ T-cells have been thought to contribute little to protective responses

against blood stage infection. However, there is increasing evidence that CD8⁺ T-cells are induced during malaria infection [129,130] and play a crucial role in controlling blood stage infection [131–134]. From a mechanistic standpoint, malaria-specific CD8⁺ T-cells have been shown to recognise infected erythroblasts, which still retain high levels of MHC class I expression, promoting increased phagocytosis of infected cells [132]. This finding, alongside evidence for the important role of IFN- γ in controlling blood stage infection [133], may explain the seemingly paradoxical role for CD8⁺ T-cells in immunity to blood-stage malaria.

There is a clear knowledge gap with regards to the effective antigenic targets that determine both antibody-mediated protection from clinical malaria as well as cell-mediated immunity against blood stage infection following subpatent exposure to blood-stage parasites, although this picture has become clearer in more recent years. More work is required to fully understand the requirements for an effective blood-stage vaccine, both with regards to antigenic targets and the specific immunological mechanisms underlying protection. Furthermore, there is a requirement for long-lasting immune responses following vaccination, which is in contrast to naturally acquired protection to blood stage malaria, which is typically short-lived once the individual is removed from periodic re-exposure.

1.2.3 Sexual stages

Although antibody responses against sexual stage antigens do not reduce parasite burden in the host, they can prevent development of the parasite in the subsequent mosquito host, hence reducing or blocking transmission between human hosts. As such, development of a transmission blocking vaccine would be a useful tool in the pathway to malaria elimination. Antibody responses against the sexual stage of the *P. falciparum* life cycle primarily involves responses against gametocyte antigens Pfs230 and Pfs48/45 that are expressed on the surface of mosquito stage gametes [135,136]. Other key sexual stage vaccine candidates include Pfs28 and Pfs25, although these are not expressed until later sporogony and antibody responses to these antigens are not observed following natural infection [135]. There have been two phase I clinical trials to-date for transmission blocking vaccine candidates [137,138]. Immunisation with a *P. vivax* Pvs25 construct was well tolerated and elicited antibody responses that were capable of partially blocking transmission in a membrane-feeding assay [137]. A later phase 1 trial with Pvs25 and Pfs25 recombinant antigens adjuvanted with Montanide ISA 51 demonstrated immunogenicity but also elicited significant adverse reactions in participants, halting the trial [138]. In comparison to blood stage antigens, naturally occurring immune responses against sexual stage antigens are minimal, and appear to be short lived in naturally exposed populations, correlating with recent exposure rather than age [136].

1.3 Intrinsically Disordered Proteins

Disordered proteins are a unique class of proteins characterised by a high degree of flexibility and lack of a well-defined 3-dimensional structure [139]. Despite this lack of structure, disordered proteins have been shown to play significant roles in many cellular processes. These roles include protein-ligand binding, DNA and RNA binding roles, flexible linkage and roles linked to the utilisation of entropic effects such as entropic clocks and springs [140–146]. Disordered proteins have also been implicated in several neurodegenerative diseases including Parkinson's, Alzheimer's and Huntington's disease [147,148]. Disordered proteins are also enriched in the proteomes of malaria parasites as compared to many other eukaryotic species [149], and may represent attractive vaccine candidates. A number of malaria vaccine candidates are predicted to contain large intrinsically disordered regions, including MSP2, EBA-175 RIII-V, EBA-181 RIII-V, EBA-140 RIII-V, GLURP, MSP3 and TEX1. Despite this, very little is known regarding the nature of immune responses against disordered proteins and how they differ from responses to structured antigens, and this section summarises current knowledge regarding adaptive immune responses against disordered antigens.

1.3.1 Disordered proteins overview

There are a number of terms which have been used to describe intrinsically disordered proteins (IDPs), and these include intrinsically unstructured proteins, unfolded proteins, natively denatured proteins, intrinsically unfolded proteins and natively disordered proteins [150,151]. While most of these terms are generally synonymous, there are a few important distinctions to be made. In general, IDP refers to any protein which does not form a stable structure with well-defined equilibrium values for both residue positions within the protein and polypeptide main-chain conformations. This includes proteins that adopt an extended conformation with high levels of solvent accessibility for most side-chains, as well as proteins that exist in a molten-globule state with mostly buried side chains. In contrast, natively unfolded proteins are typically taken to mean only the prior group [150]. The term IDP shall be used for the remainder of this work, in line with most recent literature.

IDPs are a unique group of proteins which do not adopt a well-defined structure under normal physiological conditions, and generally have a higher level of charged and hydrophilic residues that contribute to this behaviour [152,153]. IDPs can be fully unfolded, collapsed into a condensed state, or somewhere between these two extremes. In the unfolded state, the behaviour of IDPs can sometimes be loosely approximated as a statistical random coil, although steric interactions between side chains and the limits of backbone torsion angles limits this approximation somewhat. Even for IDPs which exist in an unfolded state, there may be levels of structural organization with the

protein, either in transient formation of secondary structure elements, more frequent sampling of particular conformational states, or long-range interactions between non-adjacent regions of the polypeptide chain [154]. It is clear that IDPs can also exist in a collapsed, molten-globule-like state, with extensive, dynamic interactions between side chains and a lack of stable tertiary structure [152,155]. It is also noted that some proteins contain intrinsically disordered domains which are flanked on one or both sides by structured domains, and these are often termed intrinsically disordered protein regions (IDPRs) [156] or intrinsically disordered regions (IDRs) [155].

There are numerous roles attributed to IDPs, and they appear to have critical functions in many cellular processes. These include roles in signalling networks [157–159], in nuclear transcriptional regulation and activation [160–162], cells stress pathways [163,164] and metal binding [165–169]. On a mechanistic level, IDPs can act as flexible linkers between structured protein domains, as entropic springs or bristles, or as entropic clocks [170,171]. The accessible nature of IDPs means they are often utilised to provide sites for efficient post-translational modifications such as phosphorylation, acetylation or ubiquitination [159,171–174]. Similarly, many sites of proteolytic processing are contained within IDPs [171,175]. IDPs also appear to be enriched within proteins known to be molecular chaperones, with several well-defined examples of disordered proteins acting as both protein and RNA chaperones [176,177].

A number of experimental techniques exist to identify and characterise IDPs, with protein disorder usually confirmed by a number of complementary techniques. Some of these techniques include solution-state nuclear magnetic resonance (NMR), small-angle X-RAY scattering (SAXS), circular dichroism, analytical ultracentrifugation, dynamic light scattering, fluorescence resonance energy transfer (FRET), Fourier-transform infrared spectroscopy (FTIR) or atomic force microscopy [154,178,179]. In addition to these techniques, there are a considerable number of computational prediction algorithms which have been developed to predict protein disorder on a per-residue level [180,181]. Most of these algorithms utilise some sort of machine-learning approach using a training dataset of known disordered proteins, with varying success rates across implementations. In recent years, the performance of disorder prediction software has been evaluated by the Critical Assessment of Protein Structure (CASP) experiments [182,183]. A number of predictors have emerged as leading performers in these assessments, including DISOPRED3 [184] and PrDOS [185]. These tools have been trained on X-RAY crystallography data (where missing residues are considered as disordered) [184, 185] and datasets of experimentally characterised IDPs [184]. There is no single authoritative experiment for the characterization of IDPs, and experimental techniques such as circular dichroism, SAXS and NMR probe different biophysical features of the protein under investigation. Missing residues from a crystal structure were used to define disorder in the

CASP9 and CASP10 experiments, and while there are other phenomena which may give rise to missing residues, it is considered that this definition of disorder is an acceptable and necessary compromise to enable independent benchmarking of disorder predictors [184].

1.3.2 Antibody responses against disordered proteins

It has been previously shown that disordered proteins are more common in eukaryotic species as compared to prokaryotes [186,187]. Furthermore, there appears to be a particular enrichment of disordered proteins in the proteomes of apicomplexan parasites such as *Plasmodium spp.* and *Toxoplasma gondii* [149]. There is limited work examining the nature of immune responses against disordered protein antigens. A small number of studies have examined examples of antibody binding to disordered protein regions, and pertinent results from these studies will be discussed in this section.

1.3.2.1 Conformation of IDPs when bound to antibodies

It is apparent from a number of studies that IDPs are not restricted to one common conformation when binding to multiple different antibodies. Chu *et al.* have shown that a disordered C ϵ mX domain from membrane bound IgE is capable of binding two mAbs (h47H4 and h4B12) in distinct conformations [188]. Both mAbs bound to overlapping regions within the target protein, with the peptide that encompassed both epitopes shown to form two distinct conformations when bound to respective mAbs. This peptide was shown to be intrinsically disordered by circular dichroism when in solution, suggesting limited or no formation of structural elements in the unbound state. An older study by Saad *et al.* demonstrates the existence of a discontinuous epitope within an IDP, with a monoclonal antibody binding to two discontinuous regions of the disordered apo-cytochrome c protein [189]. Another monoclonal against this protein recognised only a single region of the two. This again demonstrates the flexible nature of antibody binding to IDPs, with no requirement for an exclusive conformation for the bound IDP.

In some cases, a disordered epitope has a preferred conformation when bound to a protein partner. One key example is the human immunodeficiency virus (HIV) Tat protein, which contains an intrinsically disordered domain in the N-terminal region. Binding of a monoclonal antibody to a peptide from this region induces folding of this peptide into a beta-turn conformation. This beta-turn conformation mimics the binding of Tat to its human protein partner pTEFb [190]. This particular example suggests that some disordered regions may have an energetically preferred binding configuration, which may be shared either between antibodies, or between an antibody and binding to another protein partner. This is perhaps more likely to be the case when the disordered protein has a cognate binding partner as part of its functional role and an accompanying

energetically preferred conformation. Another example that suggests a preferred conformation for a disordered epitope is the *P. falciparum* CSP asparagine-alanine-asparagine-proline (NANP) repeat region. Two human antibodies against the NANP repeat region have been crystallised in complex with an (NPNA)₃ peptide. While the CSP peptide formed distinct conformations when bound to the two different antibodies termed Fab311 and Fab317, a common type-1 β -turn was observed in 1 and 3 of the NPNA repeats respectively [191]. This suggests an energetically preferred configuration for the repeat regions of the CSP protein, while retaining the plasticity associated with IDPs.

Two studies examining antibodies against the disordered malaria antigen MSP2 highlight structural constraints that may have to be considered for IDP based vaccines. The N-terminal region of MSP2 interacts with the parasite plasma membrane, despite being intrinsically disordered *in vitro*, adopting an alpha-helical structure on the lipid surface [192]. A mAb that recognises the N-terminal region of recombinant MSP2 fails to recognise MSP2 on the parasite surface, and this is attributed to the conformation of the lipid-bound N-terminus being incompatible with antibody binding by that particular mAb [193]. Similarly, the C-terminal region of MSP2 has also been shown to interact with the parasite membrane [194]. Two mAbs that recognise overlapping epitopes in this C-terminal domain have differing levels of binding to native MSP2 on the parasite surface, and this is likely the result of lipid interactions within the C-terminal domain either partially blocking one of the epitopes or forcing structural transitions that don't favour mAb binding [194,195].

1.3.2.2 Antibody affinity against IDPs

Another key consideration for the generation of antibodies against IDPs is the range of possible binding affinities for IDPs, and whether these differ compared to the affinity of antibodies against structured antigens. Existing work that has examined the affinity of antibodies against IDPs will be discussed in the following section, with an aim to assessing their suitability as potential vaccine candidates.

Alpha-synuclein is a 140aa, 14.5 kDa protein that is the main component of Parkinson Disease-associated Lewy bodies. Mutations in the alpha-synuclein gene have been reported to lead to an increased incidence of Parkinson Disease, although the normal biological role of alpha-synuclein is unclear. It has been shown that alpha-synuclein is intrinsically disordered *in vitro*, and possibly also *in vivo*, and can form alpha-helical structures upon binding to lipid micelles and vesicles [196]. A study by De Genst *et al.* generated a camelid nanobody against a C-terminal peptide from alpha-synuclein, with a reported binding K_D of 100 nM [197]. This affinity is reported as being typical of other *in vivo* matured camelid antibodies. Measurement of the thermodynamic properties for the binding of this camelid antibody to the peptide antigen from alpha-synuclein revealed a favourable

enthalpic contribution to binding affinity, with a large entropic cost associated with binding, especially at physiological temperatures.

Merozoite surface protein 2 (MSP2) is an essential *P. falciparum* merozoite protein and a leading vaccine candidate. It has been shown that MSP2 is predominantly disordered, while having a short structured domain at the C-terminal end, and an N-terminal region which forms an alpha-helical structure upon binding to lipid membranes. Reddy *et al.* used Surface Plasmon Resonance (SPR) to examine antibody affinity (dissociation rate) against MSP2 and PfAMA1 using sera from naturally exposed individuals [198]. In a cohort of 219 individuals from Kenya and Uganda, the median affinity for MSP2 was significantly lower than the median affinity for AMA1 [198]. When considering MSP2 monoclonal antibodies, it is of interest that the highest affinity antibodies were those directed against the C-terminal region of MSP2 [198], and all of them bind in a region flanked by two cysteine residues [199]. These two cysteine residues form a disulphide linkage, and while this disulphide link is not required for antibody binding [199], it is possible that this forms a region of reduced entropy resulting in an associated increase in monoclonal antibody affinity compared to unstructured regions.

When considering the binding kinetics of antibody-antigen interactions, there are some unique considerations which may dramatically affect overall affinity. Although perhaps an unusual example, a disordered region of the human papillomavirus E7 protein has been shown to bind to a monoclonal antibody with an initial on-rate dependent on a proline isomerization event within the recognised epitope [200]. The proline isomerization required for full antibody binding was shown to occur over minute time-scales, with only ~10% of the protein population having the proline *cis* isomer required for monoclonal binding at any one time. While proline isomerization is unlikely to be important for most antibody binding events, it highlights an additional structural consideration which may need to be taken into account when dealing with otherwise disordered regions. It is important to note that while a proline residue will generally exist as a single isomer within structured proteins, a proline residue will transition between the two possible isomers within unfolded or disordered proteins [201,202].

1.3.2.3 Entropic considerations of antibody binding

Considerations of conformational entropy are also important when dealing with antibody binding to disordered protein domains. It has been suggested that the generation of high affinity antibodies against disordered antigens is limited based on the high entropic cost associated with transition to a bound state [198,203]. However, recent evidence suggests that this effect is not as great as has previously been hypothesised, and that high affinity antibodies are routinely generated against

disordered regions. There have been numerous suggestions that protein disorder enables high-specificity, low-affinity binding interactions, due to the entropic cost associated with the loss of conformational freedom for a disordered polypeptide chain following binding [152,170,204–206]. However, there is little support for this as a general case for all binding events involving disordered proteins [207]. The typical range of binding affinities for disordered proteins binding to other protein partners is similar to that of ordered proteins, and both on-rates and off-rates are comparable to those observed for ordered proteins [207]. In many cases, it is apparent that the large reduction in configurational entropy upon binding is adequately balanced by a similar increase in entropy due to desolvation of the binding interface. Large entropic gains due to desolvation of binding interfaces are particularly associated with hydrophobic residues/regions [207].

A recent study by MacRaild *et al.* suggests that high affinity binding of antibodies to IDPs is possible, and is generally obtained by limiting the number of contact residues while maximising the epitope buried surface area [208]. This is driven by the observation that disordered epitopes typically involve fewer residues in the interaction surface than epitopes within structured proteins, and that disordered epitopes typically bind to a concave paratope with an increased proportion of buried surface area for residues within the epitope. Limiting the number of contact residues is suggested to limit the entropic penalty associated with binding, with a highly specific binding site further compensating for any entropic penalty. Nevertheless, this study did note a small but significant decrease in binding affinity for disordered antigens.

Despite the limited number of studies specifically examining antibody responses against IDPs, there is considerable evidence that IDPs can be targets of humoral immune responses following natural exposure. In the context of malaria infection, population-level studies show that a number of predicted IDPs or IDPRs are targets of naturally acquired antibody responses, with many of these antigens associated with protection from clinical malaria [115,116]. These antigens include GLURP, RAMA, EBA-175 RIII-V and MSP2.

1.3.3 Cellular immune responses against disordered proteins

It is also worth considering the role of cellular immune responses against IDPs, as both CD8⁺ and CD4⁺ T-cell responses have been shown to be important for immunity against various stages of the malaria parasite. Unlike antibody recognition of antigen, CD8⁺ and CD4⁺ T-cells recognise peptide antigen in the context of MHC class I and MHC class II molecules, respectively. It is also important to note that CD4⁺ T-cells are crucial for the development of T-dependent antibody responses that lead to affinity maturation and class switching. There is some evidence that a number of highly disordered nuclear autoantigens make poor CD4⁺ T-cell epitopes, suggested to be the result of either increased flexibility, masking by nucleic acids or increased susceptibility to proteolysis [209].

However, there has been little work examining the recognition of IDPs by CD8⁺ T-cells, nor any work examining recognition of IDPs in the context of infectious disease, and this represents a large knowledge gap that will be addressed in this thesis.

1.4 Identifying Immune Targets in Structured Malaria Antigens

IDPs fall on one end of a spectrum of protein structural behaviour; at the other end are the many proteins that fold into well-defined 3D structures. These proteins are often amenable to crystallisation and subsequent characterisation by X-ray crystallography. There are a number of methods for examining targets of immune responses within structured antigens, including crystallisation of antibody-antigen complexes. Additionally, there are a number of population genetics approaches that also provide a level of insight into the targets of adaptive immunity within a naturally exposed population. This section reviews current knowledge regarding antibody targets in relation to structured *Plasmodium* antigens, as well as current efforts to integrate protein structural information into tests based on population genetics. Finally, Section 1.4.5 examines current computational tools that could be used to integrate structural information into such tests.

1.4.1 Antibody-antigen interactions for malaria antigens

Protein crystallography remains perhaps the most definitive method for defining the epitopes recognised by monoclonal antibodies. For antigens that are derived from *Plasmodium* species, there are a limited number of antigen-antibody complexes in the Protein Data Bank (PDB) (**Table 1.1**). There are a total of 33 PDB structures representing 27 distinct mAbs, and covering 13 unique protein antigens from various *Plasmodium* species. Out of these PDB structures, most epitopes are found within structured domains, with the exception of mAbs against MSP2 and the CSP repeat region which are both predicted to be intrinsically disordered antigens. Additionally, only 11 of the 27 mAbs crystallised against *Plasmodium* antigens are of human origin, and cover only two distinct antigens, CSP and Pfs25. The rest of the crystallised antibodies are of murine origin, with the exception of a single rat antibody. It is important to note that there are distinct differences in the level of somatic hypermutation and amino acid composition in the variable heavy (VH) CDR3 domain of mouse antibodies compared to humans, and it is therefore unclear if the epitope specificity of mouse antibodies accurately reflects what is likely from a naturally exposed human population [210]. A number of example antigen-antibody interfaces are shown in **Figure 1.3**, covering both ordered and disordered antigens. It is interesting to note the concave paratope for the MSP2-binding antibody 6D8, in line with the observation by MacRaild *et al.* that paratopes for disordered antigens tend to have an increased interface curvature [208]. To expand on the work by MacRaild *et al.*, a further analysis of the buried surface area between antigen and antibodies reveals significant differences between ordered and disordered epitopes for malaria antigens (**Figure 1.4**,

Table 1.1: Experimentally determined antibody-antigen complexes for *Plasmodium* antigens (retrieved 21/12/2017).

Antigen	Epitope-containing Region/Domain	Predicted		Organism	PDB IDs (<i>mAb ID</i>)	Antibody Source	References
		Disordered (Y/N)					
CSP	C-terminal aTSR domain	N		<i>P. falciparum</i>	6B0S (1710)	Human	[211]
CSP	NANP repeat region	Y		<i>P. falciparum</i>	5BK0 (663), 6AZM (580), 6AXK (311), 6AXL (317)	Human	[191,212]
CyRPA	-	N		<i>P. falciparum</i>	5TIH (8A7), 5EZO (C12)	Mouse	[213,214]
EBA-175	Region II	N		<i>P. falciparum</i>	4QEX (R217), 4K2U (R218)	Mouse	[101]
MSP1	MSP1-19	N		<i>P. falciparum</i>	1OB1 (G17.12)	Mouse	[215]
MSP2	-	Y		<i>P. falciparum</i>	5TBD (4D11), 4QXT (6D8), 4QY8 (6D8), 4QYO (6D8), 4R3S (6D8)	Mouse	[193,195]
PfAMA1	Domain I	N		<i>P. falciparum</i>	2Q8A (1F9), 2Q8B (1F9)	Mouse	[216]
PfAMA1	Domain III	N		<i>P. falciparum</i>	2J5L (F8.12.19)	Mouse	[217]
Pfs25	-	N		<i>P. falciparum</i>	6B08 (1276), 6B0A (1269), 6B0E (1260), 6B0H (1262), 6AZZ (1190), 6B0G (1245)	Human	[218]
PfSUB1	-	N		<i>P. falciparum</i>	4LVN (NIMP.M7), 4LVO (NIMP.M7)	Mouse	[219]
RH5	-	N		<i>P. falciparum</i>	5MI0 (9AD4), 4U0R (9AD4), 4U1G (QA1)	Mouse	[220,221]
PkAMA1	Domain I/II	N		<i>P. knowlesi</i>	4UAO (R31C2)	Rat	[222]
DBP	RII (subdomain III)	N		<i>P. vivax</i>	5F3J (2D10)	Mouse	[223]
PvAMA1	Domain III	N		<i>P. vivax</i>	2J4W (F8.12.19)	Mouse	[217]
Pvs25	-	N		<i>P. vivax</i>	1Z3G (2A8)	Mouse	[224]

Note: Predictions of protein disordered were performed with DISOPRED3 [184]

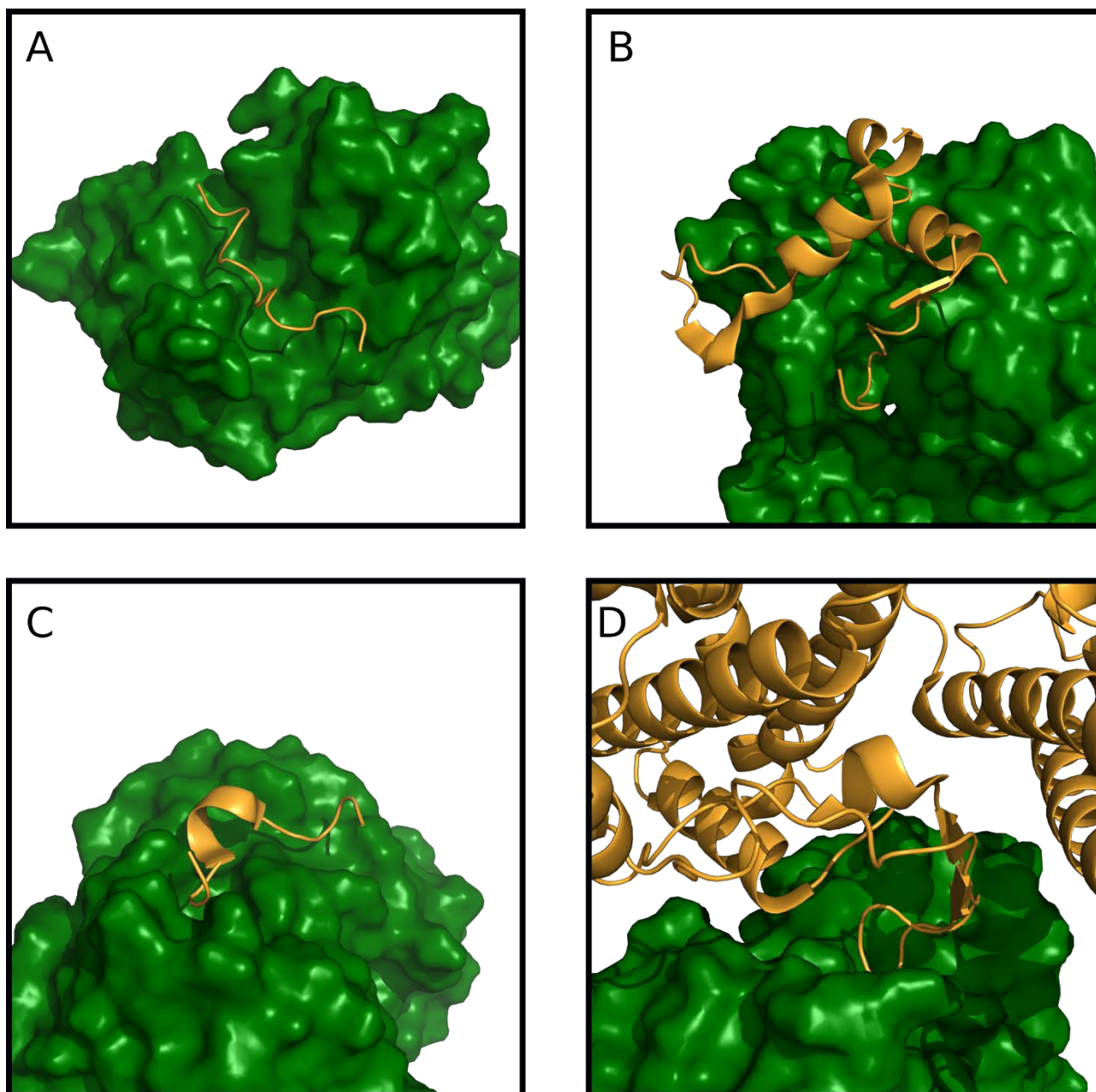


Figure 1.3: Antibody binding to ordered and disordered malaria antigens. Antibody Fab regions are shown in dark green as a surface representation. Antigens are shown in orange as cartoon representations. A) Mouse antibody (Fv region) bound to a disordered fragment of MSP2, PDB structure 4QY8 [193]. MSP2 residues 33-46 (3D7 sequence) are shown. B) Growth inhibitory mouse antibody (Fab region) bound to AMA1 DI, PDB structure 2Q8A [216]. AMA1 residues outside the epitope region are omitted for visual clarity. C) Human antibody (Fab region) bound to CSP NANP repeat region, PDB structure 6AXL [191]. This NANP repeat region adopts an extended, flexible structure in solution [225], and is predicted to be intrinsically disordered [226]. D) Mouse antibody R217 (Fab region) bound to EBA-175 RII, PDB structure 4QEX [101].

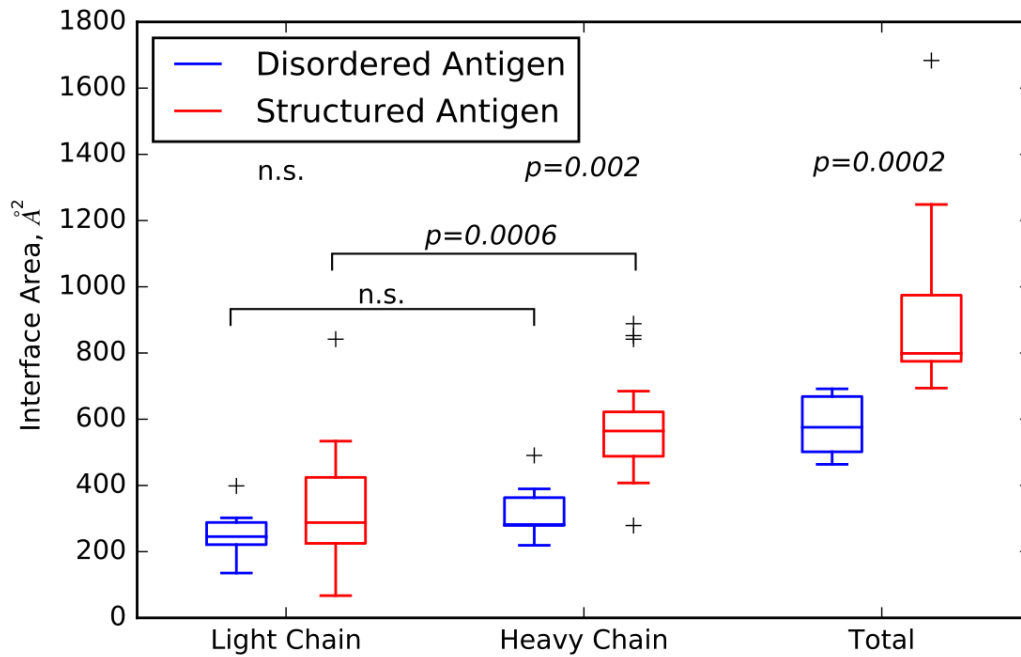


Figure 1.4: Interface surface area between antibody and antigen for disordered and structured *Plasmodium* antigen-antibody complexes. All *Plasmodium* antibody-antigen complexes were extracted from the PDB (accessed 21/12/17), with antigens classified as ordered or disordered as given in Table 1.1 (predicted by DISOPRED3 [184]). Interface surface area between heavy/light chains and the antigen was calculated using PISA [227]. P-values were calculated using the Wilcoxon rank-sum test for comparisons between ordered and disordered antigens, and using the Wilcoxon signed-rank test for comparisons between heavy and light chain interface areas.

Table 1.2: Interface surface area for *Plasmodium* antigen-antibody complexes.

Organism	Antigen	mAb ID	Disordered	Antigen-Antibody Interface Area, Å ²		
				Heavy Chain	Light Chain	Total
<i>P. falciparum</i>	CSP	663	Y	280.1	213.5	493.6
<i>P. falciparum</i>	CSP	580	Y	219.1	244.6	463.7
<i>P. falciparum</i>	CSP	311	Y	490.6	135.4	626.0
<i>P. falciparum</i>	CSP	317	Y	389.6	302.1	691.7
<i>P. falciparum</i>	CSP	1710	N	481.6	229.8	711.4
<i>P. falciparum</i>	CyRPA	8A7	N	572.6	315.7	888.3
<i>P. falciparum</i>	CyRPA	C12	N	407.5	522.6	930.1
<i>P. falciparum</i>	EBA-175	R217	N	452.5	248.6	701.1
<i>P. falciparum</i>	EBA-175	R218	N	468.2	228.2	696.4
<i>P. falciparum</i>	MSP1	G17.12	N	278.9	509.2	788.1
<i>P. falciparum</i>	MSP2	4D11	Y	278.9	246.4	525.3
<i>P. falciparum</i>	MSP2	6D8	Y	284.1	398.6	682.7
<i>P. falciparum</i>	PfAMA1	1F9	N	852.5	396.0	1,248.5
<i>P. falciparum</i>	PfAMA1	F8.12.19	N	507.5	268.5	776.0
<i>P. falciparum</i>	Pfs25	1276	N	574.2	533.7	1,107.9
<i>P. falciparum</i>	Pfs25	1269	N	888.5	351.0	1,239.5
<i>P. falciparum</i>	Pfs25	1260	N	684.7	202.2	886.9
<i>P. falciparum</i>	Pfs25	1262	N	500.7	369.5	870.2
<i>P. falciparum</i>	Pfs25	1190	N	620.6	516.4	1,137.0
<i>P. falciparum</i>	Pfs25	1245	N	611.9	186.8	798.7
<i>P. falciparum</i>	PfSUB1	NIMP.M7	N	490.4	307.1	797.4
<i>P. falciparum</i>	RH5	9AD4	N	593.0	205.8	798.8
<i>P. falciparum</i>	RH5	QA1	N	523.1	263.1	786.2
<i>P. knowlesi</i>	PkAMA1	R31C2	N	841.7	841.7	1,683.4
<i>P. vivax</i>	DBP	2D10	N	N/A*	N/A*	716.7
<i>P. vivax</i>	PvAMA1	F8.12.19	N	556.2	215.4	771.6
<i>P. vivax</i>	Pvs25	2A8	N	626.8	67.2	694.0

Note: All *Plasmodium* antibody-antigen complexes were extracted from the PDB (accessed 21/12/17), with antigens classified as ordered or disordered as predicted by DISOPRED3 [184]. Interface surface area between heavy/light chains and the antigen was calculated using PISA [227].

*mAb 2D10 against *Pv*DBP is a single-chain Fv construct, and hence does not have discrete heavy and light chains defined in the PDB file.

Table 1.2). Whilst light chain contributions to the interface surface area was similar between ordered and disordered antigens, the interface between antigen and heavy chain was much larger for structured antigens. Indeed, for disordered antigens, there is no significant difference between the interaction area with light or heavy chains. In contrast, the heavy chain makes a significantly greater contribution to buried surface area for structured antigens. This is in line with the work of MacRaid *et al.*, which suggests that disordered epitopes have lower buried surface area [208]. The equal contributions between heavy and light chains for disordered antigens suggests a binding mode in which the disordered antigen binds in a groove or a pocket between the two antibody chains, as is observed in **Figure 1.3A**. However, the small number of antigen-antibody complexes for *Plasmodium* species, with even fewer antibodies derived from human sources, limits the ability to draw strong conclusions from this data. The small number of structures also highlights the need to examine other potential markers of antibody recognition such as antibody-driven selection pressure and resultant polymorphic variation.

1.4.2 Genetic markers of immune selection pressure

Immune selection pressure on an antigen can occur as a result of antibody or T-cell recognition of a particular epitope, conferring reduced fitness to strains that are recognised by these antibodies or T-cells, and increased fitness to strains with mutations that reduce or abolish binding of antibodies or T-cells to that particular epitope. Within an individual host, such selection pressure can lead to generation of polymorphic variants that are then transmitted to other individuals. At a population level, and in an endemic setting, immunity is most often generated against high-frequency alleles (i.e. the most common alleles), and this provides a selective advantage to low-frequency alleles in the parasite population. This is an example of balancing selection, in which multiple alleles are maintained within a population at higher frequency than would be expected under a neutral model of selection (i.e. genetic drift alone). As such, balancing selection in *Plasmodium* species is typically assessed within specific populations and not across geographically distant populations.

A number of tests exist for identifying balancing selection, and these will be briefly introduced before discussing their application to *Plasmodium* species. One of the most commonly used tests for identifying balancing selection is Tajima's D [228]. This test involves comparing two estimates of genetic diversity, Watterson's theta (θ) and nucleotide diversity (π), with the difference between these two estimators forming the basis for Tajima's D. A positive Tajima's D value is indicative of balancing selection or recent population contraction, whereas a negative Tajima's D is indicative of a recent selective sweep or population expansion after a bottleneck. Other statistics that can be used to detect evidence of balancing selection include the Hudson-Kreitman-Aguade (HKA) test [229], the McDonald-Kreitman (MK) test [230], and Fu and Li's D and F [231], although these are less

commonly used for *Plasmodium* species as compared to Tajima's D. These tests are often applied either as a single test over the entire gene, or as a sliding window over the gene sequence. Two major drawbacks are apparent from these approaches when trying to identify antibody-mediated selection pressure. Firstly, when applied as a whole-gene metric, information on the particular protein domains under selection is lost and strong signatures of selection are potentially diluted. Secondly, sliding window analyses for these approaches are based on the linear genetic sequence and do not reflect the translated protein in 3D space. Additionally, the width of each sliding window is not based on the structural characteristics of the antigens being tested.

1.4.3 Identification of immune pressure on malaria genes

A number of studies have examined *Plasmodium* genes for evidence of immune selection pressure. The most widely studied gene in all *Plasmodium* species with regards to balancing selection is AMA1. This section summarises existing work that examines selection pressures on *Plasmodium* antigens, and highlights instances in which a sliding window analysis has been performed to identify domain- or region-specific selection pressures.

Within *P. falciparum*, the entire AMA1 gene has been identified as being under balancing selection in studies in Thailand [232,233], Gambia [232,234,235], Cambodia [232], Guinea [234], Iran [236], Nigeria [237] and Venezuela [238]. Additionally, particular regions of AMA1 have been identified as being under immune selection pressure using Tajima's D. Sliding window analyses have identified regions within DI [233,237,239,240], DII [240] and DIII [233,237,239,240] as being under balancing selection. Other studies have estimated selection pressures on particular domains of AMA1 (without a sliding window), with evidence for immune selection pressure on DI [236,241,242] and DIII [236] of *Pf*AMA1.

There has also been a considerable amount of work performed examining selection pressures on *Pv*AMA1. The entire *Pv*AMA1 gene has been identified as under balancing selection in Venezuela [238], whilst other studies have shown DI of *Pv*AMA1 to be the sole region under significant immune selection pressure using both a sliding window approach [238,239,243] and a whole domain analysis [244]. In contrast, a study of *Pv*AMA1 from South Korean isolates showed no evidence of balancing selection on any domain [245], possible as a result of recent population expansion. Similarly, analysis of DI *Pv*AMA1 sequences from Myanmar showed evidence of diversifying selection [246] rather than balancing selection.

In summary, whilst AMA1 in both *P. falciparum* and *P. vivax* appears to be under immune selection pressure in most populations, there appears to be a clear difference in the targets of this selection pressure; both DI and DIII are consistently shown to be under balancing selection in

*Pf*AMA1, compared to only DI for *Pv*AMA1. In contrast to what has been observed in both *P. falciparum* and *P. vivax*, there was no evidence for immune selection pressure on *P. knowlesi* AMA1 in a Malaysian population using Tajima's D or Fu and Li's F and D [247].

Although AMA1 remains the most widely studied gene for evidence of immune selection pressure, a number of other studies have identified balancing selection on other genes. Within *P. falciparum*, there is evidence for balancing selection on genes for *Pf*TRAP [248,249], SURFIN 4.2 [250], SURFIN 4.1 [251,252], *Pf*38 (Domain I) [253], *Pf*MSPDBL1 [252], *Pf*MSPDBL2 [252], *Pf*CSP [254], MSP3 [255] and EBA-175 (Region II) [256]. In *P. vivax*, balancing selection has been observed for *Pv*41 [257], *Pv*MSP1 [258,259], *Pv*CSP [258], *Pv*TRAP [260], *Pv*DBP [260,261], *Pv*MSP5 [262] and *Pv*MSP3-alpha [263]. Additionally, a number of studies have examined balancing selection on a genomic scale for *P. falciparum* [232,234,235]. Notable genes under a high degree of immune selection pressure (as indicated by high Tajima's D values) identified by these studies include MSPDBL1, AMA1 [232,234,235], PHISTa, PHISTb, MSPDBL2 [234,235], EBA-175 and MSP1 [232].

Nearly all of the studies examining immune selection pressure in *Plasmodium* species utilise Tajima's D to identify genes under balancing selection, whilst some studies have also employed the HKA test [252], MK test [237,242,248,258] or Fu and Li's D and F [237,250]. Whilst a number of studies have applied a sliding window analysis to various antigens, there is a notable lack of integration with structural data, despite protein structures being known for several antigens such as AMA1, EBA-175 and *Pv*DBP. The integration of both structural and genetic data forms a major focus of this thesis, and the following section discusses the work that has been performed examining the location of polymorphisms in relation to *Plasmodium* protein structures.

1.4.4 Integration of protein structural information with polymorphism data

There are a limited number of studies that have examined polymorphisms in the context of protein structure for malaria proteins. The majority of these studies have displayed polymorphisms in the context of the 3D protein structure, and again the major focus of most studies has been AMA1, although other antigens such as EBA-175, EBA-140, EBA-181, *Pv*DBP, CSP, VAR2CSA, MSPDBL1, MSPDBL2 and TRAP have also been examined. Salient results from these studies will be discussed below.

1.4.4.1 *Pf*AMA1

For *Pf*AMA1, it has been observed that a large number of polymorphic residues fall within Domain I, although DII and DIII are also moderately polymorphic [239,243,247,264,265]. Additionally,

there are minimal polymorphisms observed on the so-called ‘silent face’ of *Pf*AMA1 [239,243,247,264,265]. Many of the polymorphic residues within *Pf*AMA1 DI are clustered around the conserved RON2 binding cleft [264]. In a comparison between species, *Pf*AMA1 generally had a higher number of polymorphic residues in the populations examined [239,247], whilst *Pk*AMA1 had relatively few polymorphic residues [247]. Predicted and known CD8+ T-cell epitopes have also been mapped onto the *Pf*AMA1 structure, and it is interesting to note that many of these epitopes were localised to the conserved face of the AMA1 structure [266].

1.4.4.2 EBL family proteins

EBA-140, EBA-181 and EBA-175 all belong to the EBL family of *Plasmodium* proteins, with EBA-140 and EBA-175 binding to Glycophorin C and Glycophorin A respectively. The receptor for EBA-181 is unknown [267]. A functional analysis of polymorphisms in the DBL domains of these three EBL proteins show that polymorphisms in EBA-181 and EBA-140 negatively impact on receptor binding strength, but do not affect receptor specificity. In contrast, polymorphisms in EBA-175 did not impact binding strength or receptor specificity, despite EBA-175 being the most polymorphic of the three EBL proteins examined [268]. In EBA-175, none of the polymorphisms intersected with the proposed glycan binding site, either in the F1 or F2 domains, although some polymorphisms fell close to the glycan binding site within the F2 domain [268]. Another study of epitopes within EBA-175 RII has shown that the most inhibitory antibodies target a region involved in the dimerization interface, and hence prevention of dimerization around the Glycophorin A receptor is a likely mechanism of action for these inhibitory antibodies [269].

*Pv*DBP is the sole EBL family protein within *P. vivax*, and binds to DARC on human reticulocytes, forming a homodimer during this process. Polymorphic residues in *Pv*DBP sequences from Brazilian isolates were observed to flank the dimerization and DARC binding interface, whilst the dimerization and DARC binding site itself was conserved [270]. Visualisation of fractional Shannon entropy values (a measure of sequence variation at each residue) over the *Pv*DBP structure highlighted variation mostly within subdomain 2 (bordering the dimerization interface) and on some residues within subdomain 3 [271].

1.4.4.3 VAR2CSA

Variant surface antigen 2-CSA (VAR2CSA) is a *P. falciparum* antigen that is part of the PfEMP1 family of proteins. VAR2CSA binds to chondroitin sulfate A (CSA) in the placenta during pregnancy associated malaria. VAR2CSA is composed of 6 extracellular DBL domains and a transmembrane region. A study examining antibody reactivity against peptide epitopes from each of the 6 DBL domains that make up VAR2CSA noted that antibodies predominantly recognised

subdomains 1 and 2 of each DBL domain, while subdomain 3 was generally unreactive [272]. A structural analysis of the VAR2CSA DBL3X-DBL4 ϵ domains shows that residues within the interface between DBL3X and DBL4 ϵ are primarily conserved, with only a few polymorphic residues situated on the outer edges of this interface [273]. This work highlights the structural constraints that help dictate the location of polymorphic residues, even in the context of antigens which are naturally highly polymorphic.

1.4.4.4 *PfCSP*

CSP is a sporozoite stage antigen that is conserved across *Plasmodium* species, and contains three extracellular domains: a C-terminal thrombospondin-like type repeat (TSR), an N-terminal domain that mediates attachment to liver cells via heparin sulfate binding, and a central repeat region composed of a large number of tandem NANP repeats as well as some variant repeat sequences. A study of the CSP C-terminal TSR region suggests that polymorphic residues from this region have potentially arisen as a result of intermolecular interactions at the protein surface [274]. Within this analysis, it was observed that polymorphic residues were predominantly located on one side of the protein, with multiple polymorphisms found within two T-cell epitopes known as TH2 and TH3. Additionally, polymorphisms were located surrounding, but not part of, a hydrophobic pocket in the TSR region structure [274].

1.4.4.5 *MSPDBL1* & *MSPDBL2*

Merozoite surface protein DBL-1 (MSPDBL1) and merozoite surface protein DBL-2 (MSPDBL2) are two closely related proteins that are localised to the merozoite surface and thought to bind to a currently unknown red blood cell receptor [275]. These two proteins both contain single DBL domains. A study examining polymorphisms within the DBL domains of both MSPDBL2 and MSPDBL1 demonstrated that these antigens are highly polymorphic, although MSPDBL2 was observed to have more polymorphic residues in this study [276]. Polymorphisms were generally distributed evenly over the face of both proteins, with the exception of a conserved cleft between subdomain 2 and 3 in both proteins, suggesting some functional importance for this region, possibly in binding to host-cell receptors [276].

1.4.4.6 *TRAP*

Thrombospondin-related adhesive protein (TRAP) is a *Plasmodium* protein involved in the invasion of hepatocytes and mosquito salivary glands, and is localised to the parasite micronemes and sporozoite surface. TRAP contains both a TSR domain and a von Willebrand factor A (vWA)-like N-terminal A domain. Mapping of polymorphisms onto the structure of the TRAP A domain

showed that polymorphisms were clustered on one face of the protein [277]. Furthermore, there was a lack of polymorphic residues on the region surrounding residues that were deemed to be functionally important in directed mutagenesis studies [277].

In summary, tests for immune selection pressure such as Tajima's D have been applied as sliding windows to identify domain-specific selection pressure in a number of different studies of *Plasmodium* antigens. With regards to identification of polymorphic regions using protein structural information, a number of studies have displayed polymorphisms on a site-by-site basis, with one study also displaying relative Shannon entropy values on a site-by-site basis. However, statistical tests for selection pressure, such as Tajima's D, are likely to be more sensitive in detecting particular sites under immune pressure, and we note that there is a notable absence of studies examining selection pressure in the context of protein structure. An exploration of this forms the basis of Chapters 3, 4 and 5, and in the following section we review current software tools that may allow the integration of structural data into tests for selection pressure.

1.4.5 Current tools for examining protein structure and applicability to immune-related polymorphisms

A number of tools exist for the visualisation and manipulation of protein structures and associated data [278–287]. With regards to mapping tests of immune selection pressure onto protein structures, this section briefly discusses the main features of currently available software tools, and highlights gaps in the capabilities of available tools.

A number of tools have been developed to map polymorphic variants onto protein structures, including MuPIT Interactive [281], G23D [282], COMBOSA3D [284]. MuPIT is a web server that allows for mapping polymorphic variants from genomic data onto protein structures, and mutations are automatically retrieved from UniProtKB [281]. Similarly, G23D is a web server that also allows for mapping of single-nucleotide polymorphisms onto protein structures, and includes a number of additional structural analysis tools, including stability analysis and analysis of comparative contacts between mutant and wild-type variants [282]. COMBOSA3D is another web server allows for colouring of a protein 3D structure based on information from a multiple sequence alignment, such as residue conservation, highly polymorphic residues and sites of indels [284].

Several other tools exist that map sequence based annotations or other sequence-based features onto protein structures, including 3DBIONOTES [278], ConSurf [285,288], Motif3D [283] and POLYVIEW-3D [286]. The 3DBIONOTES v2.0 web server allows automatic annotation onto protein structures, drawing information from a large range of possible annotation sources. Sequences are aligned from UniProt sequences, and annotations mapped onto the 3D structure

[278]. Similarly, POLYVIEW-3D is a web server that enables annotation of sequence and structure features on protein 3D structures, also enabling high quality rendering of annotated structures [286]. ConSurf allows for the calculation of evolutionary conservation of residues and visualisation on the relevant protein structure [285,288]. Motif3D allows for the visualisation of sequence motifs over a protein 3D structure, and was available via a web interface, although it is noted that this is currently not available (access attempted 24/11/17) [283].

Finally, there are a number of tools for general protein structure visualisation and manipulation, including web-based libraries such as 3Dmol.js [279] and PV viewer [289], and offline tools such as UCSF Chimera [280] and PyMOL [287]. Both 3Dmol.js [279] and PV viewer [289] are JavaScript libraries for web-based visualisation of protein structures, and both use native WebGL 1.0 to allow hardware acceleration of rendered graphics. UCSF Chimera [280] is a high-powered, fully-featured program for the visualisation, manipulation and exploration of protein structures. However, it does not have the capability to map spatially-averaged data onto the structure. PyMOL is another offline program for the manipulation, visualisation and analysis of protein structures, and can be extended with Python scripts [287].

While a number of these tools allow for simple mapping of polymorphisms or other sequence-based annotations onto protein structures, there is no utility in any of these tools for the application of a 3D sliding window over a protein structure. This remains a key gap in current approaches to identifying selection pressures on structured antigens, in both the available software and subsequent application to organisms such as *P. falciparum* or *P. vivax*.

1.5 Thesis Hypothesis and Aims

This thesis explores the role of protein structure in the generation of adaptive immune responses against *Plasmodium* species. The overarching hypothesis for this thesis is that both structured and intrinsically disordered proteins are important antigens in adaptive immune responses against the malaria parasite, and consideration of protein structure will yield additional insights into protein regions that are targets of adaptive immunity. This thesis addresses the following specific aims:

- I. To investigate the role of intrinsically disordered protein antigens from *Plasmodium* species as targets of adaptive immune responses using a variety of computational approaches (Chapter 2).
- II. To perform a proteome-wide computational analysis of structured malaria proteins to identify key determinants of immune recognition and selection pressure (Chapter 3).

- III. To develop novel computational approaches to integrate protein structural information into measures of immune selection pressure, and apply these approaches to leading malaria vaccine candidates (Chapters 3, 4, 5).

1.6 Summary of Chapters

Chapter 2 of this thesis implements a proteome-wide computational analysis of intrinsically disordered proteins (IDPs) within *Plasmodium* species. It was shown that IDPs are particularly enriched within immunologically exposed subcellular compartments of the malaria parasite and contain an increased proportion of tandem repeat regions and non-synonymous single nucleotide polymorphisms. Of particular relevance to adaptive immune responses, IDP regions contained very few MHC class I- and II-binding peptides, which may limit the ability of CD8+ and CD4+ T-cells to recognise these regions. Chapter 3 explores regions with known or predicted protein structures within the *P. falciparum* proteome and their relation to known and predicted immune responses. Polymorphic residues were shown to be primarily surface exposed and enriched within turns. This work introduced a new 3D sliding window approach to identify regions within leading vaccine candidates that are likely to be under immune-mediated selection pressure. As part of this analysis, a region with *PfAMA1* DII/III was identified as having particularly strong signatures of balancing selection. The 3D sliding window analysis introduced in Chapter 3 was developed into a software package called BioStructMap, described in Chapter 4 and also released as an online web tool. Finally, Chapter 5 applies the BioStructMap tool to leading vaccine candidates from *P. vivax*, and examines structural patterns of selection and diversity across a number of geographic populations. This work highlights notable similarities in structural patterns of diversity across multiple populations. Furthermore, it was noted that regions of diversity tended to surround conserved binding interfaces. Finally, to integrate the various levels of data that were used and generated as part of this thesis, we have begun the development of an online platform called PlasmoSIP (*Plasmodium* Structure, Immunology and Polymorphisms) that collates the data generated in Chapter 2 alongside the structural mapping approaches outlined in Chapters 3, 4 and 5, and this is discussed briefly in Chapter 6.

1.7 References

1. World Health Organization. World Malaria Report 2017 Available: <http://www.who.int/malaria/publications/world-malaria-report-2017/en/> Accessed 21 Dec 2017.
2. Noedl H, Se Y, Schaecher K, Smith BL, Socheat D, Fukuda MM, et al. Evidence of artemisinin-resistant malaria in western Cambodia. *N Engl J Med*. 2008;359: 2619–2620.
3. Imwong M, Hien TT, Thuy-Nhien NT, Dondorp AM, White NJ. Spread of a single multidrug resistant malaria parasite lineage (*PfPailin*) to Vietnam. *Lancet Infect Dis*. 2017;17: 1022–1023.

4. Schantz-Dunn J, Nour NM. Malaria and pregnancy: a global health perspective. *Rev Obstet Gynecol*. 2009;2: 186–192.
5. Langhorne J, Ndungu FM, Sponaas A-M, Marsh K. Immunity to malaria: more questions than answers. *Nat Immunol*. 2008;9: 725–732.
6. Teo A, Feng G, Brown GV, Beeson JG, Rogerson SJ. Functional Antibodies and Protection against Blood-stage Malaria. *Trends Parasitol*. 2016; doi:10.1016/j.pt.2016.07.003
7. Cohen S, McGregor IA, Carrington S. Gamma-Globulin and Acquired Immunity to Human Malaria. *Nature*. 1961;192: 733–737.
8. World Health Organization and UNICEF. World Malaria Report 2005. 2005.
9. Gallup JL, Sachs JD. The economic burden of malaria. *Am J Trop Med Hyg*. 2001;64: 85–96.
10. World Health Organization. Defeating malaria in Asia, the Pacific, Americas, Middle East and Europe. (Progress & Impact Series, n. 9). 2012.
11. Mendis K, Sina BJ, Marchesini P, Carter R. The neglected burden of *Plasmodium vivax* malaria. *Am J Trop Med Hyg*. 2001;64: 97–106.
12. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis*. 2004;4: 327–336.
13. World Health Organization. Global Malaria Control and Elimination: report of a technical review. 2008. Available: http://whqlibdoc.who.int/publications/2008/9789241596756_eng.pdf Accessed 21 Dec 2017.
14. World Health Organization. Global technical strategy for malaria 2016-2030. 2015. Available: http://apps.who.int/iris/bitstream/10665/176712/1/9789241564991_eng.pdf Accessed 21 Dec 2017.
15. Nabarro D. Roll Back Malaria. *Parassitologia*. 1999;41: 501–504.
16. World Health Organization. World Malaria Report 2015. 2015. Available: http://apps.who.int/iris/bitstream/10665/200018/1/9789241565158_eng.pdf Accessed 21 Dec 2017.
17. Hay SI, Snow RW. The malaria Atlas Project: developing global maps of malaria risk. *PLoS Med*. 2006;3: e473.
18. Gething PW, Elyazar IRF, Moyes CL, Smith DL, Battle KE, Guerra CA, et al. A long neglected world malaria map: *Plasmodium vivax* endemicity in 2010. *PLoS Negl Trop Dis*. 2012;6: e1814.
19. Gething PW, Patil AP, Smith DL, Guerra CA, Elyazar IRF, Johnston GL, et al. A new world malaria map: *Plasmodium falciparum* endemicity in 2010. *Malar J*. 2011;10: 378.
20. Vanderberg JP, Frevert U. Intravital microscopy demonstrating antibody-mediated immobilisation of *Plasmodium berghei* sporozoites injected into skin by mosquitoes. *Int J Parasitol*. 2004;34: 991–996.
21. Matsuoka H, Yoshida S, Hirai M, Ishii A. A rodent malaria, *Plasmodium berghei*, is experimentally transmitted to mice by merely probing of infective mosquito, *Anopheles stephensi*. *Parasitol Int*. 2002;51: 17–23.
22. Medica DL, Sinnis P. Quantitative dynamics of *Plasmodium yoelii* sporozoite transmission by infected anopheline mosquitoes. *Infect Immun*. 2005;73: 4363–4369.
23. Warrell DA. Essential Malariology, 4Ed. 4 edition. CRC Press; 2002.
24. Sturm A, Amino R, van de Sand C, Regen T, Retzlaff S, Rennenberg A, et al. Manipulation of host hepatocytes by the malaria parasite for delivery into liver sinusoids. *Science*. 2006;313: 1287–1290.

25. Graewe S, Rankin KE, Lehmann C, Deschermeier C, Hecht L, Froehlke U, et al. Hostile Takeover by *Plasmodium*: Reorganization of Parasite and Host Cell Membranes during Liver Stage Egress. *PLoS Pathog*. 2011;7: e1002224.
26. Baron S, editor. *Medical Microbiology*. Galveston (TX): University of Texas Medical Branch at Galveston; 2011.
27. World Health Organization. Malaria Vaccine Rainbow Tables. Available: http://www.who.int/vaccine_research/links/Rainbow/en/index.html Accessed 21 Dec 2017.
28. Barry AE, Arnott A. Strategies for designing and monitoring malaria vaccines targeting diverse antigens. *Front Immunol*. 2014;5: 359.
29. Cowman AF, Crabb BS. Invasion of red blood cells by malaria parasites. *Cell*. 2006;124: 755–766.
30. Aikawa M, Miller LH, Johnson J, Rabbege J. Erythrocyte entry by malarial parasites. A moving junction between erythrocyte and parasite. *J Cell Biol*. 1978;77: 72–82.
31. Gilson PR, Crabb BS. Morphology and kinetics of the three distinct phases of red blood cell invasion by *Plasmodium falciparum* merozoites. *Int J Parasitol*. 2009;39: 91–96.
32. Baker DA. Malaria gametocytogenesis. *Mol Biochem Parasitol*. 2010;172: 57–65.
33. Josling GA, Llinás M. Sexual development in *Plasmodium* parasites: knowing when it's time to commit. *Nat Rev Microbiol*. 2015;13: 573–587.
34. Koch M, Baum J. The mechanics of malaria parasite invasion of the human erythrocyte - towards a reassessment of the host cell contribution. *Cell Microbiol*. 2016;18: 319–329.
35. 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.
36. Adams JH, Blair PL, Kaneko O, Peterson DS. An expanding ebl family of *Plasmodium falciparum*. *Trends Parasitol*. 2001;17: 297–299.
37. Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science*. 1994;264: 1941–1944.
38. Maier AG, Duraisingh MT, Reeder JC, Patel SS, Kazura JW, Zimmerman PA, et al. *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med*. 2003;9: 87–92.
39. Mayer DCG, Cofie J, Jiang L, Hartl DL, Tracy E, Kabat J, et al. Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proc Natl Acad Sci USA*. 2009;106: 5348–5352.
40. Horuk R, Chitnis CE, Darbonne WC, Colby TJ, Rybicki A, Hadley TJ, et al. A receptor for the malarial parasite *Plasmodium vivax*: the erythrocyte chemokine receptor. *Science*. 1993;261: 1182–1184.
41. Crosnier C, Bustamante LY, Bartholdson SJ, Bei AK, Theron M, Uchikawa M, et al. Basigin is a receptor essential for erythrocyte invasion by *Plasmodium falciparum*. *Nature*. 2011;480: 534–537.
42. Galaway F, Drought LG, Fala M, Cross N, Kemp AC, Rayner JC, et al. P113 is a merozoite surface protein that binds the N terminus of *Plasmodium falciparum* RH5. *Nat Commun*. 2017;8: 14333.
43. Lamarque M, Besteiro S, Papoin J, Roques M, Normand BV-L, Morlon-Guyot J, et al. The RON2-AMA1 Interaction is a Critical Step in Moving Junction-Dependent Invasion by Apicomplexan Parasites. *PLoS Pathog*. 2011;7: e1001276.
44. Hodder AN, Crewther PE, Matthew ML, Reid GE, Moritz RL, Simpson RJ, et al. The disulfide bond

- structure of *Plasmodium* apical membrane antigen-1. J Biol Chem. 1996;271: 29446–29452.
45. 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 USA. 2005;102: 12736–12741.
 46. Mugenyi CK, Elliott SR, McCallum FJ, Anders RF, Marsh K, Beeson JG. Antibodies to Polymorphic Invasion-Inhibitory and Non-Inhibitory Epitopes of *Plasmodium falciparum* Apical Membrane Antigen 1 in Human Malaria. PLoS One. 2013;8: e68304.
 47. Beeson JG, Chan J-A, Fowkes FJI. PfEMP1 as a target of human immunity and a vaccine candidate against malaria. Expert Rev Vaccines. 2013;12: 105–108.
 48. Bernabeu M, Danziger SA, Avril M, Vaz M, Babar PH, Brazier AJ, et al. Severe adult malaria is associated with specific PfEMP1 adhesion types and high parasite biomass. Proc Natl Acad Sci USA. 2016;113: E3270–E3279.
 49. Tembo DL, Nyoni B, Murikoli RV, Mukaka M, Milner DA, Berriman M, et al. Differential PfEMP1 expression is associated with cerebral malaria pathology. PLoS Pathog. 2014;10: e1004537.
 50. Almelli T, Ndam NT, Ezimegnon S, Alao MJ, Ahouansou C, Sagbo G, et al. Cytoadherence phenotype of *Plasmodium falciparum*- infected erythrocytes is associated with specific pfemp-1 expression in parasites from children with cerebral malaria. Malar J. 2014;13: 1–9.
 51. Buffet PA, Safeukui I, Deplaine G, Brousse V, Prendki V, Thellier M, et al. The pathogenesis of *Plasmodium falciparum* malaria in humans: insights from splenic physiology. Blood. 2011;117: 381–392.
 52. Idro R, Marsh K, John CC, Newton CRJ. Cerebral malaria: mechanisms of brain injury and strategies for improved neurocognitive outcome. Pediatr Res. 2010;68: 267–274.
 53. World Health Organization. Guidelines for the treatment of malaria, 3rd edition. 2015. Available: http://apps.who.int/iris/bitstream/10665/162441/1/9789241549127_eng.pdf Accessed 21 Dec 2017.
 54. Wellems TE, Plowe CV. Chloroquine-Resistant Malaria. J Infect Dis. 2001;184: 770–776.
 55. Yeung S, Pongtavornpinyo W, Hastings IM, Mills AJ, White NJ. Antimalarial Drug Resistance, Artemisinin-based Combination Therapy, and the Contribution of Modeling to Elucidating Policy Choices. Am J Trop Med Hyg. 2004;71: 179–186.
 56. Ashley EA, Dhorda M, Fairhurst RM, Amaratunga C, Lim P, Suon S, et al. Spread of Artemisinin Resistance in *Plasmodium falciparum* Malaria. N Engl J Med. 2014;371: 411–423.
 57. Mbengue A, Bhattacharjee S, Pandharkar T, Liu H, Estiu G, Stahelin RV, et al. A molecular mechanism of artemisinin resistance in *Plasmodium falciparum* malaria. Nature. 2015;520: 683–687.
 58. Amaratunga C, Lim P, Suon S, Sreng S, Mao S, Sopha C, et al. Dihydroartemisinin-piperaquine resistance in *Plasmodium falciparum* malaria in Cambodia: a multisite prospective cohort study. Lancet Infect Dis. 2016;16: 357–365.
 59. Campo JJ, Dobaño C, Sacarlal J, Guinovart C, Mayor A, Angov E, et al. Impact of the RTS,S Malaria Vaccine Candidate on Naturally Acquired Antibody Responses to Multiple Asexual Blood Stage Antigens. PLoS One. 2011;6: e25779.
 60. Stoute JA, Slaoui M, Heppner DG, Momin P, Kester KE, Desmons P, et al. A Preliminary Evaluation of a Recombinant Circumsporozoite Protein Vaccine against *Plasmodium falciparum* Malaria. N Engl J Med. 1997;336: 86–91.
 61. The RTS SCTP. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. N Engl J Med. 2012;367: 2284–2295.

62. Fowkes FJI, Simpson JA, Beeson JG. Implications of the licensure of a partially efficacious malaria vaccine on evaluating second-generation vaccines. *BMC Med.* 2013;11: 1–8.
63. Thera MA, Plowe CV. Vaccines for malaria: how close are we? *Annu Rev Med.* 2012;63: 345–357.
64. Liu J, Modrek S, Gosling RD, Feachem RGA. Malaria eradication: is it possible? Is it worth it? Should we do it? *Lancet Glob Health.* 2013;1: e2–e3.
65. Olotu A, Fegan G, Wambua J, Nyangweso G, Leach A, Lievens M, et al. Seven-Year Efficacy of RTS,S/AS01 Malaria Vaccine among Young African Children. *N Engl J Med.* 2016;374: 2519–2529.
66. PATH and GSK welcome progress toward RTS,S malaria vaccine pilot implementation with selection of countries. In: MVI PATH Malaria Vaccine Initiative. 24 Apr 2017. Available: <http://www.malariavaccine.org/news-events/news/path-and-gsk-welcome-progress-toward-rtss-malaria-vaccine-pilot-implementation> Accessed 21 Dec 2017.
67. 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.
68. Mueller I, Shakri AR, Chitnis CE. Development of vaccines for *Plasmodium vivax* malaria. *Vaccine.* 2015;33: 7489–7495.
69. Takashima E, Morita M, Tsuboi T. Vaccine candidates for malaria: what’s new? *Expert Rev Vaccines.* 2016;15: 1–3.
70. Silvie O, Semblat JP, Franetich JF, Hannoun L, Eling W, Mazier D. Effects of irradiation on *Plasmodium falciparum* sporozoite hepatic development: implications for the design of pre-erythrocytic malaria vaccines. *Parasite Immunol.* 2002;24: 221–223.
71. Weiss WR, Jiang CG. Protective CD8+ T lymphocytes in primates immunized with malaria sporozoites. *PLoS One.* 2012;7: e31247.
72. Weiss WR, Sedegah M, Beaudoin RL, Miller LH, Good MF. CD8+ T cells (cytotoxic/suppressors) are required for protection in mice immunized with malaria sporozoites. *Proc Natl Acad Sci USA.* 1988;85: 573–576.
73. Schofield L, Villaquiran J, Ferreira A, Schellekens H, Nussenzweig R, Nussenzweig V. Gamma interferon, CD8+ T cells and antibodies required for immunity to malaria sporozoites. *Nature.* 1987;330: 664–666.
74. Krzych U, Dalai S, Zarling S, Pichugin A. Memory CD8 T cells specific for plasmodia liver-stage antigens maintain protracted protection against malaria. *Front Immunol.* 2012;3: 370.
75. Rodrigues MM, Cordey A-S, Arreaza G, Corradin G, Romero P, Maryanski JL, et al. CD8+ cytolytic T cell clones derived against the *Plasmodium yoelii* circumsporozoite protein protect against malaria. *Int Immunol.* 1991;3: 579–585.
76. Tsuji M. A retrospective evaluation of the role of T cells in the development of malaria vaccine. *Exp Parasitol.* 2010;126: 421–425.
77. Kumar KA, Sano G-I, Boscardin S, Nussenzweig RS, Nussenzweig MC, Zavala F, et al. The circumsporozoite protein is an immunodominant protective antigen in irradiated sporozoites. *Nature.* 2006;444: 937–940.
78. Trieu A, Kayala MA, Burk C, Molina DM, Freilich DA, Richie TL, et al. Sterile protective immunity to malaria is associated with a panel of novel *P. falciparum* antigens. *Mol Cell Proteomics.* 2011;10: M111.007948.
79. Grüner AC, Mauduit M, Tewari R, Romero JF, Depinay N, Kayibanda M, et al. Sterile protection against malaria is independent of immune responses to the circumsporozoite protein. *PLoS One.* 2007;2: e1371.

80. Moorthy VS, Ballou WR. Immunological mechanisms underlying protection mediated by RTS,S: a review of the available data. *Malar J.* 2009;8: 1–7.
81. White MT, Bejon P, Olotu A, Griffin JT, Riley EM, Kester KE, et al. The relationship between RTS,S vaccine-induced antibodies, CD4⁺ T cell responses and protection against *Plasmodium falciparum* infection. *PLoS One.* 2013;8: e61395.
82. Ejigiri I, Ragheb DRT, Pino P, Coppi A, Bennett BL, Soldati-Favre D, et al. Shedding of TRAP by a rhomboid protease from the malaria sporozoite surface is essential for gliding motility and sporozoite infectivity. *PLoS Pathog.* 2012;8: e1002725.
83. Sultan AA, Thathy V, Frevert U, Robson KJ, Crisanti A, Nussenzweig V, et al. TRAP is necessary for gliding motility and infectivity of *Plasmodium* sporozoites. *Cell.* 1997;90: 511–522.
84. Akhouri RR, Sharma A, Malhotra P, Sharma A. Role of *Plasmodium falciparum* thrombospondin-related anonymous protein in host-cell interactions. *Malar J.* 2008;7: 1–11.
85. Müller HM, Reckmann I, Hollingdale MR, Bujard H, Robson KJ, Crisanti A. Thrombospondin related anonymous protein (TRAP) of *Plasmodium falciparum* binds specifically to sulfated glycoconjugates and to HepG2 hepatoma cells suggesting a role for this molecule in sporozoite invasion of hepatocytes. *EMBO J.* 1993;12: 2881–2889.
86. Robson KJ, Frevert U, Reckmann I, Cowan G, Beier J, Scragg IG, et al. Thrombospondin-related adhesive protein (TRAP) of *Plasmodium falciparum*: expression during sporozoite ontogeny and binding to human hepatocytes. *EMBO J.* 1995;14: 3883–3894.
87. Wengelnik K, Spaccapelo R, Naitza S, Robson KJ, Janse CJ, Bistoni F, et al. The A-domain and the thrombospondin-related motif of *Plasmodium falciparum* TRAP are implicated in the invasion process of mosquito salivary glands. *EMBO J.* 1999;18: 5195–5204.
88. Aly ASI, Vaughan AM, Kappe SHI. Malaria parasite development in the mosquito and infection of the mammalian host. *Annu Rev Microbiol.* 2009;63: 195–221.
89. Scarselli E, Tolle R, Koita O, Diallo M, Müller HM, Früh K, et al. Analysis of the human antibody response to thrombospondin-related anonymous protein of *Plasmodium falciparum*. *Infect Immun.* 1993;61: 3490–3495.
90. Longley RJ, Reyes-Sandoval A, Montoya-Díaz E, Dunachie S, Kumpitak C, Nguitragool W, et al. Acquisition and Longevity of Antibodies to *Plasmodium vivax* Preerythrocytic Antigens in Western Thailand. *Clin Vaccine Immunol.* 2016;23: 117–124.
91. Ambrosino E, Dumoulin C, Orlandi-Pradines E, Remoue F, Toure-Baldé A, Tall A, et al. A multiplex assay for the simultaneous detection of antibodies against 15 *Plasmodium falciparum* and *Anopheles gambiae* saliva antigens. *Malar J.* 2010;9: 1–12.
92. Doodoo D, Hollingdale MR, Anum D, Koram KA, Gyan B, Akanmori BD, et al. Measuring naturally acquired immune responses to candidate malaria vaccine antigens in Ghanaian adults. *Malar J.* 2011;10: 168.
93. Sarr JB, Orlandi-Pradines E, Fortin S, Sow C, Cornelie S, Rogerie F, et al. Assessment of exposure to *Plasmodium falciparum* transmission in a low endemicity area by using multiplex fluorescent microsphere-based serological assays. *Parasit Vectors.* 2011;4: 1–8.
94. Nahrendorf W, Scholzen A, Bijker EM, Teirlinck AC, Bastiaens GJH, Schats R, et al. Memory B-cell and antibody responses induced by *Plasmodium falciparum* sporozoite immunization. *J Infect Dis.* 2014;210: 1981–1990.
95. Chan J-A, Howell KB, Reiling L, Ataide R, Mackintosh CL, Fowkes FJI, et al. Targets of antibodies against *Plasmodium falciparum*-infected erythrocytes in malaria immunity. *J Clin Invest.* 2012;122: 3227–3238.

96. Fried M, Duffy PE. Designing a VAR2CSA-based vaccine to prevent placental malaria. *Vaccine*. 2015;33: 7483–7488.
97. Chêne A, Houard S, Nielsen MA, Hundt S, D'Alessio F, Sirima SB, et al. Clinical development of placental malaria vaccines and immunoassays harmonization: a workshop report. *Malar J*. 2016;15: 476.
98. Nielsen MA, Resende M, de Jongh WA, Ditlev SB, Mordmüller B, Houard S, et al. The Influence of Sub-Unit Composition and Expression System on the Functional Antibody Response in the Development of a VAR2CSA Based *Plasmodium falciparum* Placental Malaria Vaccine. *PLoS One*. 2015;10: e0135406.
99. El Sahly HM, Patel SM, Atmar RL, Lanford TA, Dube T, Thompson D, et al. Safety and immunogenicity of a recombinant nonglycosylated erythrocyte binding antigen 175 Region II malaria vaccine in healthy adults living in an area where malaria is not endemic. *Clin Vaccine Immunol*. 2010;17: 1552–1559.
100. Kim Lee Sim B, Narum DL, Chattopadhyay R, Ahumada A, David Haynes J, Fuhrmann SR, et al. Delineation of Stage Specific Expression of *Plasmodium falciparum* EBA-175 by Biologically Functional Region II Monoclonal Antibodies. *PLoS One*. 2011;6: e18393.
101. 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.
102. Badiane AS, Bei AK, Ahouidi AD, Patel SD, Salinas N, Ndiaye D, et al. Inhibitory humoral responses to the *Plasmodium falciparum* vaccine candidate EBA-175 are independent of the erythrocyte invasion pathway. *Clin Vaccine Immunol*. 2013;20: 1238–1245.
103. Irani V, Ramsland PA, Guy AJ, Siba PM, Mueller I, Richards JS, et al. Acquisition of Functional Antibodies That Block the Binding of Erythrocyte-Binding Antigen 175 and Protection Against *Plasmodium falciparum* Malaria in Children. *Clin Infect Dis*. 2015;61: 1244–1252.
104. Persson KEM, Fowkes FJI, McCallum FJ, Gicheru N, Reiling L, Richards JS, et al. Erythrocyte-binding antigens of *Plasmodium falciparum* are targets of human inhibitory antibodies and function to evade naturally acquired immunity. *J Immunol*. 2013;191: 785–794.
105. Chiu CY, White MT, Healer J, Thompson JK, Siba PM, Mueller I, et al. Different Regions of *Plasmodium falciparum* Erythrocyte-Binding Antigen 175 Induce Antibody Responses to Infection of Varied Efficacy. *J Infect Dis*. 2016;214: 96–104.
106. 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.
107. Thera MA, Doumbo OK, Coulibaly D, Laurens MB, Ouattara A, Kone AK, et al. A Field Trial to Assess a Blood-Stage Malaria Vaccine. *N Engl J Med*. 2011;365: 1004–1013.
108. Sagara I, Dicko A, Ellis RD, Fay MP, Diawara SI, Assadou MH, et al. A randomized controlled phase 2 trial of the blood stage AMA1-C1/Alhydrogel malaria vaccine in children in Mali. *Vaccine*. 2009;27: 3090–3098.
109. 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 USA*. 2011;108: 13275–13280.
110. Hodder AN, Crewther PE, Anders RF. Specificity of the protective antibody response to apical membrane antigen 1. *Infect Immun*. 2001;69: 3286–3294.
111. 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.
112. Drew DR, Hodder AN, Wilson DW, Foley M, Mueller I, Siba PM, et al. Defining the Antigenic Diversity of *Plasmodium falciparum* Apical Membrane Antigen 1 and the Requirements for a Multi-Allele Vaccine against Malaria. *PLoS One*. 2012;7: e51023.
 113. Tonkin ML, Roques M, Lamarque MH, Pugn  re M, Douguet D, Crawford J, et al. Host cell invasion by apicomplexan parasites: insights from the co-structure of AMA1 with a RON2 peptide. *Science*. 2011;333: 463–467.
 114. Healer J, Chiu CY, Hansen DS. Mechanisms of naturally acquired immunity to *P. falciparum* and approaches to identify merozoite antigen targets. *Parasitology*. 2017; 1–9.
 115. 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.
 116. Richards JS, Arumugam TU, Reiling L, Healer J, Hodder AN, Fowkes FJI, et al. Identification and prioritization of merozoite antigens as targets of protective human immunity to *Plasmodium falciparum* malaria for vaccine and biomarker development. *J Immunol*. 2013;191: 795–809.
 117. 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.
 118. Osier FH, Mackinnon MJ, Crosnier C, Fegan G, Kamuyu G, Wanaguru M, et al. New antigens for a multicomponent blood-stage malaria vaccine. *Sci Transl Med*. 2014;6: 247ra102.
 119. Dent AE, Nakajima R, Liang L, Baum E, Moormann AM, Sumba PO, et al. *Plasmodium falciparum* Protein Microarray Antibody Profiles Correlate With Protection From Symptomatic Malaria in Kenya. *J Infect Dis*. 2015;212: 1429–1438.
 120. Torres KJ, Castrillon CE, Moss EL, Saito M, Tenorio R, Molina DM, et al. Genome-level determination of *Plasmodium falciparum* blood-stage targets of malarial clinical immunity in the Peruvian Amazon. *J Infect Dis*. 2015;211: 1342–1351.
 121. Taylor-Robinson AW, Phillips RS, Severn A, Moncada S, Liew FY. The role of TH1 and TH2 cells in a rodent malaria infection. *Science*. 1993;260: 1931–1934.
 122. S  ss G, Eichmann K, Kury E, Linke A, Langhorne J. Roles of CD4- and CD8-bearing T lymphocytes in the immune response to the erythrocytic stages of *Plasmodium chabaudi*. *Infect Immun*. 1988;56: 3081–3088.
 123. Meding SJ, Langhorne J. CD4+ T cells and B cells are necessary for the transfer of protective immunity to *Plasmodium chabaudi chabaudi*. *Eur J Immunol*. 1991;21: 1433–1438.
 124. Brake DA, Long CA, Weidanz WP. Adoptive protection against *Plasmodium chabaudi adami* malaria in athymic nude mice by a cloned T cell line. *J Immunol*. 1988;140: 1989–1993.
 125. Xu H, Wipasa J, Yan H, Zeng M, Makobongo MO, Finkelman FD, et al. The mechanism and significance of deletion of parasite-specific CD4(+) T cells in malaria infection. *J Exp Med*. 2002;195: 881–892.
 126. Nie CQ, Bernard NJ, Schofield L, Hansen DS. CD4+ CD25+ regulatory T cells suppress CD4+ T-cell function and inhibit the development of *Plasmodium berghei*-specific TH1 responses involved in cerebral malaria pathogenesis. *Infect Immun*. 2007;75: 2275–2282.
 127. Pinzon-Charry A, McPhun V, Kienzle V, Hirunpetcharat C, Engwerda C, McCarthy J, et al. Low doses of killed parasite in CpG elicit vigorous CD4+ T cell responses against blood-stage malaria in mice. *J*

Clin Invest. 2010;120: 2967–2978.

128. Good MF, Reiman JM, Rodriguez IB, Ito K, Yanow SK, El-Deeb IM, et al. Cross-species malaria immunity induced by chemically attenuated parasites. *J Clin Invest.* 2013; doi:10.1172/JCI66634
129. Chandele A, Mukerjee P, Das G, Ahmed R, Chauhan VS. Phenotypic and functional profiling of malaria-induced CD8 and CD4 T cells during blood-stage infection with *Plasmodium yoelii*. *Immunology.* 2011;132: 273–286.
130. Lundie RJ, de Koning-Ward TF, Davey GM, Nie CQ, Hansen DS, Lau LS, et al. Blood-stage *Plasmodium* infection induces CD8+ T lymphocytes to parasite-expressed antigens, largely regulated by CD8alpha+ dendritic cells. *Proc Natl Acad Sci USA.* 2008;105: 14509–14514.
131. Horne-Debets JM, Faleiro R, Karunarathne DS, Liu XQ, Lineburg KE, Poh CM, et al. PD-1 dependent exhaustion of CD8+ T cells drives chronic malaria. *Cell Rep.* 2013;5: 1204–1213.
132. Imai T, Ishida H, Suzue K, Taniguchi T, Okada H, Shimokawa C, et al. Cytotoxic activities of CD8+ T cells collaborate with macrophages to protect against blood-stage murine malaria. *Elife.* 2015;4. doi:10.7554/eLife.04232
133. Imai T, Shen J, Chou B, Duan X, Tu L, Tetsutani K, et al. Involvement of CD8+ T cells in protective immunity against murine blood-stage infection with *Plasmodium yoelii* 17XL strain. *Eur J Immunol.* 2010;40: 1053–1061.
134. Horne-Debets JM, Karunarathne DS, Faleiro RJ, Poh CM, Renia L, Wykes MN. Mice lacking Programmed cell death-1 show a role for CD8+ T cells in long-term immunity against blood-stage malaria. *Sci Rep.* 2016;6: 26210.
135. Bousema JT, Drakeley CJ, Sauerwein RW. Sexual-stage antibody responses to *P. falciparum* in endemic populations. *Curr Mol Med.* 2006;6: 223–229.
136. Ouédraogo AL, Roeffen W, Luty AJF, de Vlas SJ, Nebie I, Ilboudo-Sanogo E, et al. Naturally Acquired Immune Responses to *Plasmodium falciparum* Sexual Stage Antigens Pfs48/45 and Pfs230 in an Area of Seasonal Transmission. *Infect Immun.* 2011;79: 4957–4964.
137. Malkin EM, Durbin AP, Diemert DJ, Sattabongkot J, Wu Y, Miura K, et al. Phase 1 vaccine trial of Pvs25H: a transmission blocking vaccine for *Plasmodium vivax* malaria. *Vaccine.* 2005;23: 3131–3138.
138. Wu Y, Ellis RD, Shaffer D, Fontes E, Malkin EM, Mahanty S, et al. Phase 1 trial of malaria transmission blocking vaccine candidates Pfs25 and Pvs25 formulated with montanide ISA 51. *PLoS One.* 2008;3: e2636.
139. Tompa P. Intrinsically unstructured proteins. *Trends Biochem Sci.* 2002;27: 527–533.
140. Uversky VN, Dunker AK. The case for intrinsically disordered proteins playing contributory roles in molecular recognition without a stable 3D structure. *F1000 Biol Rep.* 2013;5: 1.
141. Beckham GT, Bomble YJ, Matthews JF, Taylor CB, Resch MG, Yarbrough JM, et al. The O-glycosylated linker from the *Trichoderma reesei* Family 7 cellulase is a flexible, disordered protein. *Biophys J.* 2010;99: 3773–3781.
142. Buske PJ, Levin PA. A flexible C-terminal linker is required for proper FtsZ assembly in vitro and cytokinetic ring formation in vivo. *Mol Microbiol.* 2013;89: 249–263.
143. Csizmók V, Bokor M, Bánki P, Klement E, Medzihradsky KF, Friedrich P, et al. Primary contact sites in intrinsically unstructured proteins: the case of calpastatin and microtubule-associated protein 2. *Biochemistry.* 2005;44: 3955–3964.
144. Meador WE, Means AR, Quijcho FA. Modulation of calmodulin plasticity in molecular recognition on the basis of x-ray structures. *Science.* 1993;262: 1718–1721.

145. Oldfield CJ, Cheng Y, Cortese MS, Romero P, Uversky VN, Dunker AK. Coupled folding and binding with alpha-helix-forming molecular recognition elements. *Biochemistry*. 2005;44: 12454–12470.
146. von Ossowski I, Eaton JT, Czjzek M, Perkins SJ, Frandsen TP, Schüle M, et al. Protein Disorder: Conformational Distribution of the Flexible Linker in a Chimeric Double Cellulase. *Biophys J*. 2005;4;88: 2823–2832.
147. Raychaudhuri S, Dey S, Bhattacharyya NP, Mukhopadhyay D. The role of intrinsically unstructured proteins in neurodegenerative diseases. *PLoS One*. 2009;4: e5566.
148. Babu MM, van der Lee R, de Groot NS, Gsponer J. Intrinsically disordered proteins: regulation and disease. *Curr Opin Struct Biol*. 2011;21: 432–440.
149. Feng Z-P, Zhang X, Han P, Arora N, Anders RF, Norton RS. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol Biochem Parasitol*. 2006;150: 256–267.
150. Dunker AK, Babu MM, Barbar E, Blackledge M, Bondos SE, Dosztányi Z, et al. What's in a name? Why these proteins are intrinsically disordered. *Intrinsically Disord Proteins*. 2013;1: e24157.
151. Uversky VN. Intrinsically disordered proteins from A to Z. *Int J Biochem Cell Biol*. 2011;43: 1090–1103.
152. Uversky VN. Natively unfolded proteins: a point where biology waits for physics. *Protein Sci*. 2002;11: 739–756.
153. Uversky VN, Gillespie JR, Fink AL. Why are “natively unfolded” proteins unstructured under physiologic conditions? *Proteins*. 2000;41: 415–427.
154. Tompa P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem Sci*. 2012;37: 509–516.
155. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. Classification of intrinsically disordered regions and proteins. *Chem Rev*. 2014;114: 6589–6631.
156. Oldfield CJ, Dunker AK. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu Rev Biochem*. 2014;83: 553–584.
157. Wright PE, Jane Dyson H. Intrinsically disordered proteins in cellular signalling and regulation. *Nat Rev Mol Cell Biol*. 2014;16: 18–29.
158. Bürgi J, Xue B, Uversky VN, van der Goot FG. Intrinsic Disorder in Transmembrane Proteins: Roles in Signaling and Topology Prediction. *PLoS One*. 2016;11: e0158594.
159. Das RK, Huang Y, Phillips AH, Kriwacki RW, Pappu RV. Cryptic sequence features within the disordered protein p27Kip1 regulate cell cycle signaling. *Proc Natl Acad Sci USA*. 2016;113: 5616–5621.
160. Meng F, Na I, Kurgan L, Uversky VN. Compartmentalization and Functionality of Nuclear Disorder: Intrinsic Disorder and Protein-Protein Interactions in Intra-Nuclear Compartments. *Int J Mol Sci*. 2016;17. doi:10.3390/ijms17010024
161. Frege T, Uversky VN. Intrinsically disordered proteins in the nucleus of human cells. *Biochem Biophys Rep*. 2015;5;1: 33–51.
162. Liu J, Perumal NB, Oldfield CJ, Su EW, Uversky VN, Dunker AK. Intrinsic disorder in transcription factors. *Biochemistry*. 2006;45: 6873–6888.
163. Altmeyer M, Neelsen KJ, Teloni F, Pozdnyakova I, Pellegrino S, Grøfte M, et al. Liquid demixing of intrinsically disordered proteins is seeded by poly(ADP-ribose). *Nat Commun*. 2015;6: 8088.
164. Mitrea DM, Kriwacki RW. Phase separation in biology; functional organization of a higher order. *Cell*

Commun Signal. 2016;14: 1.

165. He Z, Dunker AK, Wesson CR, Trumble WR. Ca(2+)-induced folding and aggregation of skeletal muscle sarcoplasmic reticulum calsequestrin. The involvement of the trifluoperazine-binding site. *J Biol Chem.* 1993;268: 24635–24641.
166. Perticaroli S, Nickels JD, Ehlers G, Mamontov E, Sokolov AP. Dynamics and rigidity in an intrinsically disordered protein, β -casein. *J Phys Chem B.* 2014;118: 7317–7326.
167. Wojtas M, Hołubowicz R, Poznar M, Maciejewska M, Ożyhar A, Dobryczycki P. Calcium ion binding properties and the effect of phosphorylation on the intrinsically disordered Starmaker protein. *Biochemistry.* 2015;54: 6525–6534.
168. Breydo L, Uversky VN. Role of metal ions in aggregation of intrinsically disordered proteins in neurodegenerative diseases. *Metallomics.* 2011;3: 1163–1180.
169. Faller P, Hureau C, La Penna G. Metal ions and intrinsically disordered proteins and peptides: from Cu/Zn amyloid- β to general principles. *Acc Chem Res.* 2014;47: 2252–2259.
170. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. Intrinsically disordered protein. *J Mol Graph Model.* 2001;19: 26–59.
171. Tompa P. *Structure and Function of Intrinsically Disordered Proteins.* Florida: Chapman & Hall/CRC; 2010.
172. Kurotani A, Tokmakov AA, Kuroda Y, Fukami Y, Shinozaki K, Sakurai T. Correlations between predicted protein disorder and post-translational modifications in plants. *Bioinformatics.* 2014; doi:10.1093/bioinformatics/btt762
173. Bah A, Forman-Kay JD. Modulation of Intrinsically Disordered Protein Function by Post-translational Modifications. *J Biol Chem.* 2016;291: 6696–6705.
174. Kragelund BB, Schenstrøm SM, Rebula CA, Panse VG, Hartmann-Petersen R. DSS1/Sem1, a Multifunctional and Intrinsically Disordered Protein. *Trends Biochem Sci.* 2016;41: 446–459.
175. de Opakua AI, Merino N, Villate M, Cordeiro TN, Ormaza G, Sánchez-Carbayo M, et al. The metastasis suppressor KISS1 is an intrinsically disordered protein slightly more extended than a random coil. *PLoS One.* 2017;12: e0172507.
176. Tompa P, Kovacs D. Intrinsically disordered chaperones in plants and animals. *Biochem Cell Biol.* 2010;88: 167–174.
177. Kovacs D, Tompa P. Diverse functional manifestations of intrinsic structural disorder in molecular chaperones. *Biochem Soc Trans.* 2012;40: 963–968.
178. Eliezer D. Biophysical characterization of intrinsically disordered proteins. *Curr Opin Struct Biol.* 2009;19: 23–30.
179. Riback JA, Bowman MA, Zmyslowski AM, Knoverek CR, Jumper JM, Hinshaw JR, et al. Innovative scattering analysis shows that hydrophobic disordered proteins are expanded in water. *Science.* 2017;358: 238–241.
180. Li J, Feng Y, Wang X, Li J, Liu W, Rong L, et al. An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014. *Int J Mol Sci.* 2015;16: 23446–23462.
181. Meng F, Uversky VN, Kurgan L. Comprehensive review of methods for prediction of intrinsic disorder and its molecular functions. *Cell Mol Life Sci.* 2017;74: 3069–3090.
182. Monastyrskyy B, Fidelis K, Moulton J, Tramontano A, Kryshchuk A. Evaluation of disorder predictions in CASP9. *Proteins.* 2011;79 Suppl 10: 107–118.

183. Monastyrskyy B, Kryshchuk A, Moulton J, Tramontano A, Fidelis K. Assessment of protein disorder region predictions in CASP10. *Proteins*. 2014;82 Suppl 2: 127–137.
184. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2015;31: 857–863.
185. Ishida T, Kinoshita K. PrDOS: prediction of disordered protein regions from amino acid sequence. *Nucleic Acids Res*. 2007;35: W460–4.
186. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*. 2004;337: 635–645.
187. Pancsa R, Tompa P. Structural disorder in eukaryotes. *PLoS One*. 2012;7: e34687.
188. Chu H-M, Wright J, Chan Y-H, Lin C-J, Chang TW, Lim C. Two potential therapeutic antibodies bind to a peptide segment of membrane-bound IgE in different conformations. *Nat Commun*. 2014;5. doi:10.1038/ncomms4139
189. Saad B, Corradin G, Bosshard HR. Monoclonal antibody recognizes a conformational epitope in a random coil protein. *Eur J Biochem*. 1988;178: 219–224.
190. Serrière J, Dugua J-M, Bossus M, Verrier B, Haser R, Gouet P, et al. Fab'-induced folding of antigenic N-terminal peptides from intrinsically disordered HIV-1 Tat revealed by X-ray crystallography. *J Mol Biol*. 2011;405: 33–42.
191. Oyen D, Torres JL, Wille-Reece U, Ockenhouse CF, Emerling D, Glanville J, et al. Structural basis for antibody recognition of the NANP repeats in *Plasmodium falciparum* circumsporozoite protein. *Proc Natl Acad Sci USA*. 2017;114: E10438–E10445.
192. MacRaild CA, Pedersen MØ, Anders RF, Norton RS. Lipid interactions of the malaria antigen merozoite surface protein 2. *Biochim Biophys Acta*. 2012;1818: 2572–2578.
193. Morales RAV, MacRaild CA, Seow J, Krishnarjuna B, Drinkwater N, Rouet R, et al. Structural basis for epitope masking and strain specificity of a conserved epitope in an intrinsically disordered malaria vaccine candidate. *Sci Rep*. 2015;5: 10103.
194. Das SC, Morales RAV, Seow J, Krishnarjuna B, Dissanayake R, Anders RF, et al. Lipid interactions modulate the structural and antigenic properties of the C-terminal domain of the malaria antigen merozoite surface protein 2. *FEBS J*. 2017;284: 2649–2662.
195. Seow J, Morales RAV, MacRaild CA, Krishnarjuna B, McGowan S, Dingjan T, et al. Structure and Characterisation of a Key Epitope in the Conserved C-Terminal Domain of the Malaria Vaccine Candidate MSP2. *J Mol Biol*. 2017;429: 836–846.
196. Alderson TR, Markley JL. Biophysical characterization of α -synuclein and its controversial structure. *Intrinsically Disord Proteins*. 2013;1: 18–39.
197. De Genst EJ, Guillemins T, Wellens J, O'Day EM, Waudby CA, Meehan S, et al. Structure and properties of a complex of α -synuclein and a single-domain camelid antibody. *J Mol Biol*. 2010;402: 326–343.
198. Reddy SB, Anders RF, Beeson JG, Färnert A, Kironde F, Berenzon SK, et al. High affinity antibodies to *Plasmodium falciparum* merozoite antigens are associated with protection from malaria. *PLoS One*. 2012;7: e32242.
199. 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.
200. Fassolari M, Chemes LB, Gallo M, Smal C, Sánchez IE, de Prat-Gay G. Minute time scale prolyl isomerization governs antibody recognition of an intrinsically disordered immunodominant epitope. *J*

- Biol Chem. 2013;288: 13110–13123.
201. Wüthrich K, Grathwohl C. A novel approach for studies of the molecular conformations in flexible polypeptides. FEBS Lett. 1974;43: 337–340.
 202. William J. Wedemeyer, Ervin Welker, Scheraga* HA. Proline Cis–Trans Isomerization and Protein Folding. Biochemistry. 2002;41: 14637–14644.
 203. MacRaild CA, Zachrdla M, Andrew D, Krishnarjuna B, Nováček J, Židek L, et al. Conformational dynamics and antigenicity in the disordered malaria antigen merozoite surface protein 2. PLoS One. 2015;10: e0119899.
 204. Jane Dyson H, Wright PE. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol. 2005;6: 197–208.
 205. Zhou H-X. Intrinsic disorder: signaling via highly specific but short-lived association. Trends Biochem Sci. 2012;37: 43–48.
 206. Gsponer J, Babu MM. The rules of disorder or why disorder rules. Prog Biophys Mol Biol. 2009;99: 94–103.
 207. Dogan J, Gianni S, Jemth P. The binding mechanisms of intrinsically disordered proteins. Phys Chem Chem Phys. 2014;16: 6323–6331.
 208. MacRaild CA, Richards JS, Anders RF, Norton RS. Antibody Recognition of Disordered Antigens. Structure. 2016;24: 148–157.
 209. Carl PL, Temple BRS, Cohen PL. Most nuclear systemic autoantigens are extremely disordered proteins: implications for the etiology of systemic autoimmunity. Arthritis Res Ther. 2005;7: R1360.
 210. Roy SW. The *Plasmodium gaboni* genome illuminates allelic dimorphism of immunologically important surface antigens in *P. falciparum*. Infect Genet Evol. 2015;36: 441–449.
 211. Scally SW, Murugan R, Bosch A, Triller G, Costa G, Mordmüller B, et al. Rare PfCSP C-terminal antibodies induced by live sporozoite vaccination are ineffective against malaria infection. J Exp Med. 2017; doi:10.1084/jem.20170869
 212. Triller G, Scally SW, Costa G, Pissarev M, Kreschel C, Bosch A, et al. Natural Parasite Exposure Induces Protective Human Anti-Malarial Antibodies. Immunity. 2017;47: 1197–1209.e10.
 213. Chen L, Xu Y, Wong W, Thompson JK, Healer J, Goddard-Borger ED, et al. Structural basis for inhibition of erythrocyte invasion by antibodies to *Plasmodium falciparum* protein CyRPA. Elife. 2017;6. doi:10.7554/eLife.21347
 214. Favuzza P, Guffart E, Tamborrini M, Scherer B, Dreyer AM, Rufer AC, et al. Structure of the malaria vaccine candidate antigen CyRPA and its complex with a parasite invasion inhibitory antibody. Elife. 2017;6. doi:10.7554/eLife.20383
 215. Pizarro JC, Chitarra V, Verger D, Holm I, Pêtres S, Darteville 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. J Mol Biol. 2003;328: 1091–1103.
 216. Coley AM, Gupta A, Murphy VJ, Bai T, Kim H, Foley M, et al. Structure of the malaria antigen AMA1 in complex with a growth-inhibitory antibody. PLoS Pathog. 2007;3: 1308–1319.
 217. Igonet S, Vulliez-Le Normand B, Faure G, Riottot M-M, Kocken CHM, Thomas AW, et al. Cross-reactivity studies of an anti-*Plasmodium vivax* apical membrane antigen 1 monoclonal antibody: binding and structural characterisation. J Mol Biol. 2007;366: 1523–1537.
 218. Scally SW, McLeod B, Bosch A, Miura K, Liang Q, Carroll S, et al. Molecular definition of multiple sites of antibody inhibition of malaria transmission-blocking vaccine antigen Pfs25. Nat Commun.

- 2017;8: 1568.
219. Withers-Martinez C, Strath M, Hackett F, Haire LF, Howell SA, Walker PA, et al. The malaria parasite egress protease SUB1 is a calcium-dependent redox switch subtilisin. *Nat Commun.* 2014;5: 3726.
 220. Campeotto I, Goldenzweig A, Davey J, Barfod L, Marshall JM, Silk SE, et al. One-step design of a stable variant of the malaria invasion protein RH5 for use as a vaccine immunogen. *Proc Natl Acad Sci USA.* 2017;114: 998–1002.
 221. Wright KE, Hjerrild KA, Bartlett J, Douglas AD, Jin J, Brown RE, et al. Structure of malaria invasion protein RH5 with erythrocyte basigin and blocking antibodies. *Nature.* 2014;515: 427–430.
 222. Vulliez-Le Normand B, Faber BW, Saul FA, van der Eijk M, Thomas AW, Singh B, et al. Crystal structure of *Plasmodium knowlesi* apical membrane antigen 1 and its complex with an invasion-inhibitory monoclonal antibody. *PLoS One.* 2015;10: e0123567.
 223. Chen E, Salinas ND, Huang Y, Ntumngia F, Plasencia MD, Gross ML, et al. Broadly neutralizing epitopes in the *Plasmodium vivax* vaccine candidate Duffy Binding Protein. *Proc Natl Acad Sci USA.* 2016;113: 6277–6282.
 224. Saxena AK, Singh K, Su H-P, Klein MM, Stowers AW, Saul AJ, et al. The essential mosquito-stage P25 and P28 proteins from *Plasmodium* form tile-like triangular prisms. *Nat Struct Mol Biol.* 2006;13: 90–91.
 225. Plassmeyer ML, Reiter K, Shimp RL Jr, Kotova S, Smith PD, Hurt DE, et al. Structure of the *Plasmodium falciparum* circumsporozoite protein, a leading malaria vaccine candidate. *J Biol Chem.* 2009;284: 26951–26963.
 226. Guy AJ, Irani V, MacRaid 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.
 227. Krissinel E, Henrick K. Inference of macromolecular assemblies from crystalline state. *J Mol Biol.* 2007;372: 774–797.
 228. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics.* 1989;123: 585–595.
 229. Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. *Genetics.* 1987;116: 153–159.
 230. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature.* 1991;351: 652.
 231. Fu YX, Li WH. Statistical tests of neutrality of mutations. *Genetics.* 1993;133: 693–709.
 232. Samad H, Coll F, Preston MD, Ocholla H, Fairhurst RM, Clark TG. Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. *PLoS Genet.* 2015;11: e1005131.
 233. Polley SD, Chokejindachai W, Conway DJ. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics.* 2003;165: 555–561.
 234. 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.
 235. Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Elizabeth Deerhake M, et al. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet.* 2012;8: e1002992.
 236. Mehrizi AA, Sepehri M, Karimi F, Djadid ND, Zakeri S. Population genetics, sequence diversity and

- selection in the gene encoding the *Plasmodium falciparum* apical membrane antigen 1 in clinical isolates from the south-east of Iran. *Infect Genet Evol.* 2013;17: 51–61.
237. Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics.* 2001;158: 1505–1512.
238. Ord RL, Tami A, Sutherland CJ. ama1 Genes of Sympatric *Plasmodium vivax* and *P. falciparum* from Venezuela Differ Significantly in Genetic Diversity and Recombination Frequency. *PLoS One.* 2008;3: e3366.
239. 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.
240. Osier FHA, Weedall GD, Verra F, Murungi L, Tetteh KKA, Bull P, et al. Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. *Infect Immun.* 2010;78: 4625–4633.
241. Basu M, Maji AK, Mitra M, Sengupta S. Natural selection and population genetic structure of domain-I of *Plasmodium falciparum* apical membrane antigen-1 in India. *Infect Genet Evol.* 2013;18: 247–256.
242. Garg S, Alam MT, Das MK, Dev V, Kumar A, Dash AP, et al. Sequence diversity and natural selection at domain I of the apical membrane antigen 1 among Indian *Plasmodium falciparum* populations. *Malar J.* 2007;6: 154.
243. Arnott A, Mueller I, Ramsland PA, Siba PM, Reeder JC, Barry AE. Global Population Structure of the Genes Encoding the Malaria Vaccine Candidate, *Plasmodium vivax* Apical Membrane Antigen 1 (PvAMA1). *PLoS Negl Trop Dis.* 2013;7: e2506.
244. Zakeri S, Sadeghi H, Mehrizi AA, Djadid ND. Population genetic structure and polymorphism analysis of gene encoding apical membrane antigen-1 (AMA-1) of Iranian *Plasmodium vivax* wild isolates. *Acta Trop.* 2013;126: 269–279.
245. Kang J-M, Lee J, Cho P-Y, Moon S-U, Ju H-L, Ahn SK, et al. Population genetic structure and natural selection of apical membrane antigen-1 in *Plasmodium vivax* Korean isolates. *Malar J.* 2015;14: 455.
246. Moon S-U, Na B-K, Kang J-M, Kim J-Y, Cho S-H, Park Y-K, et al. Genetic polymorphism and effect of natural selection at domain I of apical membrane antigen-1 (AMA-1) in *Plasmodium vivax* isolates from Myanmar. *Acta Trop.* 2010;114: 71–75.
247. Faber BW, Kadir KA, Rodriguez-Garcia R, Remarque EJ, Saul FA, Normand BV-L, et al. Low Levels of Polymorphisms and No Evidence for Diversifying Selection on the *Plasmodium knowlesi* Apical Membrane Antigen 1 Gene. *PLoS One.* 2015;10: e0124400.
248. Ohashi J, Suzuki Y, Naka I, Hananantachai H, Patarapotikul J. Diversifying Selection on the Thrombospondin-Related Adhesive Protein (TRAP) Gene of *Plasmodium falciparum* in Thailand. *PLoS One.* 2014;9: e90522.
249. Weedall GD, Preston BMJ, Thomas AW, Sutherland CJ, Conway DJ. Differential evidence of natural selection on two leading sporozoite stage malaria vaccine candidate antigens. *Int J Parasitol.* 2007;37: 77–85.
250. Kaewthamasorn M, Yahata K, Alexandre JSF, Xangsayarath P, Nakazawa S, Torii M, et al. Stable allele frequency distribution of the polymorphic region of SURFIN(4.2) in *Plasmodium falciparum* isolates from Thailand. *Parasitol Int.* 2012;61: 317–323.
251. Xangsayarath P, Kaewthamasorn M, Yahata K, Nakazawa S, Sattabongkot J, Udomsangpetch R, et al. Positive diversifying selection on the *Plasmodium falciparum* surf4.1 gene in Thailand. *Trop Med Health.* 2012;40: 79–89.

252. Ochola LI, Tetteh KKA, Stewart LB, Riitho V, Marsh K, Conway DJ. Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Mol Biol Evol*. 2010;27: 2344–2351.
253. Reeder JC, Wapling J, Mueller I, Siba PM, Barry AE. Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection. *Malar J*. 2011;10: 126.
254. Putaporntip C, Jongwutiwes S, Hughes AL. Natural selection maintains a stable polymorphism at the circumsporozoite protein locus of *Plasmodium falciparum* in a low endemic area. *Infect Genet Evol*. 2009;9: 567–573.
255. Polley SD, Tetteh KKA, Lloyd JM, Akpogheneta OJ, Greenwood BM, Bojang KA, et al. *Plasmodium falciparum* merozoite surface protein 3 is a target of allele-specific immunity and alleles are maintained by natural selection. *J Infect Dis*. 2007;195: 279–287.
256. Verra F, Chokejindachai W, Weedall GD, Polley SD, Mwangi TW, Marsh K, et al. Contrasting signatures of selection on the *Plasmodium falciparum* erythrocyte binding antigen gene family. *Mol Biochem Parasitol*. 2006;149: 182–190.
257. Wang Y, Ma A, Chen S-B, Yang Y-C, Chen J-H, Yin M-B. Genetic diversity and natural selection of three blood-stage 6-Cys proteins in *Plasmodium vivax* populations from the China-Myanmar endemic border. *Infect Genet Evol*. 2014;28: 167–174.
258. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing Patterns of Selection and Geospatial Genetic Diversity within Two Leading *Plasmodium vivax* Candidate Vaccine Antigens. *PLoS Negl Trop Dis*. 2014;8: e2796.
259. Dias S, Longacre S, Escalante AA, Udagama-Randeniya PV. Genetic diversity and recombination at the C-terminal fragment of the merozoite surface protein-1 of *Plasmodium vivax* (PvMSP-1) in Sri Lanka. *Infect Genet Evol*. 2011;11: 145–156.
260. Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ. Genetic diversity and population structure of genes encoding vaccine candidate antigens of *Plasmodium vivax*. *Malar J*. 2012;11: 68.
261. Premaratne PH, Aravinda BR, Escalante AA, Udagama PV. Genetic diversity of *Plasmodium vivax* Duffy Binding Protein II (PvDBPII) under unstable transmission and low intensity malaria in Sri Lanka. *Infect Genet Evol*. 2011;11: 1327–1339.
262. Putaporntip C, Udomsangpetch R, Pattanawong U, Cui L, Jongwutiwes S. Genetic diversity of the *Plasmodium vivax* merozoite surface protein-5 locus from diverse geographic origins. *Gene*. 2010;456: 24–35.
263. Ord R, Polley S, Tami A, Sutherland CJ. High sequence diversity and evidence of balancing selection in the Pvmsp3alpha gene of *Plasmodium vivax* in the Venezuelan Amazon. *Mol Biochem Parasitol*. 2005;144: 86–93.
264. 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 USA*. 2005;102: 12736–12741.
265. 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.
266. Sedegah M, Kim Y, Peters B, McGrath S, Ganeshan H, Lejano J, et al. Identification and localization of minimal MHC-restricted CD8+ T cell epitopes within the *Plasmodium falciparum* AMA1 protein. *Malar J*. 2010;9: 241.
267. Tham W-H, Healer J, Cowman AF. Erythrocyte and reticulocyte binding-like proteins of *Plasmodium*

- falciparum*. Trends Parasitol. 2012;28: 23–30.
268. Maier AG, Baum J, Smith B, Conway DJ, Cowman AF. Polymorphisms in erythrocyte binding antigens 140 and 181 affect function and binding but not receptor specificity in *Plasmodium falciparum*. Infect Immun. 2009;77: 1689–1699.
269. 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.
270. Sousa TN de, Kano FS, Brito CFA de, Carvalho LH. The Duffy binding protein as a key target for a *Plasmodium vivax* vaccine: lessons from the Brazilian Amazon. Mem Inst Oswaldo Cruz. 2014;109: 608–617.
271. Sampath S, Carrico C, Janes J, Gurumoorthy S, Gibson C, Melcher M, et al. Glycan masking of *Plasmodium vivax* Duffy Binding Protein for probing protein binding function and vaccine development. PLoS Pathog. 2013;9: e1003420.
272. Andersen P, Nielsen MA, Resende M, Rask TS, Dahlbäck M, Theander T, et al. Structural insight into epitopes in the pregnancy-associated malaria protein VAR2CSA. PLoS Pathog. 2008;4: e42.
273. Gangnard S, Lewit-Bentley A, Dechavanne S, Srivastava A, Amirat F, Bentley GA, et al. Structure of the DBL3X-DBL4 ϵ region of the VAR2CSA placental malaria vaccine candidate: insight into DBL domain interactions. Sci Rep. 2015;5: 14868.
274. Aragam NR, Thayer KM, Nge N, Hoffman I, Martinson F, Kamwendo D, et al. Diversity of T cell epitopes in *Plasmodium falciparum* circumsporozoite protein likely due to protein-protein interactions. PLoS One. 2013;8: e62427.
275. Wickramarachchi T, Cabrera AL, Sinha D, Dhawan S, Chandran T, Devi YS, et al. A novel *Plasmodium falciparum* erythrocyte binding protein associated with the merozoite surface, PfDBLMSP. Int J Parasitol. 2009;39: 763–773.
276. Hodder AN, Czabotar PE, Uboldi AD, Clarke OB, Lin CS, Healer J, et al. Insights into Duffy binding-like domains through the crystal structure and function of the merozoite surface protein MSPDBL2 from *Plasmodium falciparum*. J Biol Chem. 2012;287: 32922–32939.
277. 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.
278. Segura J, Sanchez-Garcia R, Martinez M, Cuenca-Alba J, Tabas-Madrid D, Sorzano COS, et al. 3DBIONOTES v2.0: a web server for the automatic annotation of macromolecular structures. Bioinformatics. 2017;33: 3655–3657.
279. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. Bioinformatics. 2015;31: 1322–1324.
280. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem. 2004;25: 1605–1612.
281. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. Hum Genet. 2013;132: 1235–1243.
282. Solomon O, Kunik V, Simon A, Kol N, Barel O, Lev A, et al. G23D: Online tool for mapping and visualization of genomic variants on 3D protein structures. BMC Genomics. 2016;17: 681.
283. Gaulton A, Attwood TK. Motif3D: Relating protein sequence motifs to 3D structure. Nucleic Acids Res. 2003;31: 3333–3336.
284. Stothard PM. COMBOSA3D: combining sequence alignments with three-dimensional structures.

- Bioinformatics. 2001;17: 198–199.
285. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 2010;38: W529–33.
 286. Porollo A, Meller J. Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. *BMC Bioinformatics.* 2007;8: 316.
 287. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
 288. Armon A, Graur D, Ben-Tal N. ConSurf: an algorithmic tool for the identification of functional regions in proteins by surface mapping of phylogenetic information. *J Mol Biol.* 2001;307: 447–463.
 289. Biasini M. pv: v1.8.1. Zenodo; 2015. doi:10.5281/zenodo.20980 Accessed 21 Dec 2017.

Chapter 2

The Impact of Protein Disorder on Adaptive Immunity

Since the determination of the crystal structure of myoglobin by Kendrew *et al.* in 1958 [1], protein crystallisation has become the default method for understanding the nature of a protein's structure, with more than 120,000 protein crystal structures currently deposited in the Protein Data Bank (PDB). The ability to determine a protein's three-dimensional structure has led to major advancements in our understanding of both the structure and associated function of a large number of proteins. As a result, the accepted paradigm for many decades was that a well-folded protein with a defined three-dimensional structure was required for biological function; conversely, loss of a defined three-dimensional structure corresponded to a loss of biological function. This so-called structure-function paradigm [2] is certainly true for many functional proteins—especially enzymes. However, the readiness with which X-ray crystallographic techniques could be applied to many proteins has led to the relative neglect of proteins which do not exist as a homogeneous population with a well-defined three-dimensional structure. Intrinsically disordered proteins (IDPs) fall within this neglected category, being a set of proteins which exist as a dynamic ensemble of structurally distinct species and are hence refractory to crystallisation. It is only in the last decade or two that we have begun to understand the crucial role that IDPs play in many biological systems; it is now abundantly clear that IDPs are not only common within the proteomes of most species, but that many IDPs have functions that do not rely on well-folded protein domains. Indeed, in some cases the biological function of an IDP is directly tied to its disordered state [3].

Previous studies have shown that IDPs are more abundant within eukaryotic species as compared to prokaryotes [4], and that the proteomes of a number of *Apicomplexan* parasites contain particularly high levels of IDPs [5]. Of particular interest is the high proportion of IDPs within *P. falciparum* and *P. vivax*—the species that are responsible for the largest proportion of malaria morbidity and mortality in humans. Given that these parasites have evolved for many thousands of years alongside their primate hosts, and in the presence of immune pressures, it is worth considering whether this increased proportion of IDPs is a result of evolutionary interactions with the host immune response. Very little work has been performed to systematically evaluate the effect of protein disorder on adaptive immune responses, and as such, our current knowledge regarding the antigenic potential of disordered proteins is limited. A number of *P. falciparum* antigens that are key immune targets also happen to be IDPs, including EBA-175 RIII-V, EBA-181 RIII-V, EBA-140 RIII-V and MSP2 [6,7]. However, the relative distribution of IDPs within specific subcellular compartments of *Plasmodium* species is unknown, as is the overall impact of protein disorder on protein antigenicity.

Therefore this chapter aimed to explore the effect on protein disorder on the generation of adaptive immune responses. A proteome-wide computational approach was used to examine the occurrence of IDPs across a number of *Plasmodium* species and compare this with several correlates of immunity. This approach utilised a number of computational prediction algorithms to determine protein disorder, MHC class I and class II binding, predicted linear B-cell epitopes and the existence of tandem repeats. Additionally, published experimental data on known protein localisation was used to identify particular subcellular compartments within *P. falciparum* that may be either enriched or depleted in disordered proteins.

Within this chapter, it was shown that immune responses against IDPs appear to have characteristics distinct from those against structured protein domains, with increased antibody recognition of linear epitopes but some constraints for MHC presentation and increased levels of polymorphisms. Furthermore, it was demonstrated that IDPs are particularly enriched within immunologically-exposed subcellular compartments of *P. falciparum*. MHC presentation of foreign peptides is crucial for the activation of both CD4+ and CD8+ T-cells, and IDP regions were predicted to contain relatively few MHC class I and II binding peptides. This is due to inherent differences in amino acid composition compared to structured domains, and has implications for the generation of T-cell responses, antibody responses and B-cell memory against IDPs. In contrast, linear B-cell epitopes were predicted to be enriched in IDPs, as was expected given the accessible nature of an IDP. Finally, tandem repeat regions and non-synonymous single nucleotide polymorphisms were found to be strongly associated with regions of disorder. This chapter also highlights a number of vaccine candidates that are predicted to contain considerable IDP regions. From this it is clear that a number of IDPs within *P. falciparum* are targeted by functional immune responses and that some of these antigens are realistic vaccine candidates. The results presented in this chapter also suggest that application of an IDP in a vaccine setting would need to involve inclusion of appropriate CD4+ T-cell epitopes within the vaccine construct, as T-cell help is essential for high-affinity, long-lived antibody responses. If the goal of a particular vaccine is to generate a strong antibody response, then the included T-cell epitopes do not need to be from the IDP, and could be from an unrelated antigen. This is one of the important differences between IDPs in a natural setting and in an artificial vaccine construct. In summary, this study provides insights into the role of IDPs in immune evasion and parasite invasion processes, and raises additional considerations when evaluating IDPs for use within any future vaccine constructs.

References

1. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature*. 1958;181: 662–666.

2. Wright PE, Dyson HJ. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol.* 1999;293: 321–331.
3. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradović Z. Intrinsic disorder and protein function. *Biochemistry.* 2002;41: 6573–6582.
4. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol.* 2004;337: 635–645.
5. Feng Z-P, Zhang X, Han P, Arora N, Anders RF, Norton RS. Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol Biochem Parasitol.* 2006;150: 256–267.
6. 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.
7. Adda CG, Murphy VJ, Sunde M, Waddington LJ, Schloegel J, Talbo GH, et al. *Plasmodium falciparum* merozoite surface protein 2 is unstructured and forms amyloid-like fibrils. *Mol Biochem Parasitol.* 2009;166: 159–171.

RESEARCH ARTICLE

Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions

Andrew J. Guy^{1,2}, Vashti Irani^{1,3}, Christopher A. MacRaid⁴, Robin F. Anders⁵, Raymond S. Norton⁴, James G. Beeson^{1,3,6}, Jack S. Richards^{1,3,6,7}*, Paul A. Ramsland^{1,2,8,9}*

1 Centre for Biomedical Research, Burnet Institute, Melbourne, Australia, **2** Department of Immunology, Monash University, Melbourne, Australia, **3** Department of Medicine, University of Melbourne, Melbourne, Australia, **4** Medicinal Chemistry, Monash Institute of Pharmaceutical Sciences, Monash University, Parkville, Australia, **5** Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Australia, **6** Department of Microbiology, Monash University, Melbourne, Australia, **7** Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Australia, **8** Department of Surgery Austin Health, University of Melbourne, Heidelberg, Australia, **9** School of Biomedical Sciences, CHIRI Biosciences, Faculty of Health Sciences, Curtin University, Perth, Australia



OPEN ACCESS

Citation: Guy AJ, Irani V, MacRaid CA, Anders RF, Norton RS, Beeson JG, et al. (2015) Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. PLoS ONE 10(10): e0141729. doi:10.1371/journal.pone.0141729

Editor: Nicholas J Mantis, New York State Dept. Health, UNITED STATES

Received: August 27, 2015

Accepted: October 12, 2015

Published: October 29, 2015

Copyright: © 2015 Guy et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

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

Funding: Funding was provided by the National Health and Medical Research Council of Australia (grants APP637406 and APP1042520; fellowships to RSN, JGB and JSR: APP1059060, APP1077636 and APP1037722), Victoria State Government Operational Infrastructure Support grant, Monash University (Australian Postgraduate Award to AJG) and the University of Melbourne (Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship to VI).

* These authors contributed equally to this work.

* richards@burnet.edu.au (JSR); pramsland@burnet.edu.au (PAR)

Abstract

Malaria remains a significant global health burden. The development of an effective malaria vaccine remains as a major challenge with the potential to significantly reduce morbidity and mortality. While *Plasmodium* spp. have been shown to contain a large number of intrinsically disordered proteins (IDPs) or disordered protein regions, the relationship of protein structure to subcellular localisation and adaptive immune responses remains unclear. In this study, we employed several computational prediction algorithms to identify IDPs at the proteome level of six *Plasmodium* spp. and to investigate the potential impact of protein disorder on adaptive immunity against *P. falciparum* parasites. IDPs were shown to be particularly enriched within nuclear proteins, apical proteins, exported proteins and proteins localised to the parasitophorous vacuole. Furthermore, several leading vaccine candidates, and proteins with known roles in host-cell invasion, have extensive regions of disorder. Presentation of peptides by MHC molecules plays an important role in adaptive immune responses, and we show that IDP regions are predicted to contain relatively few MHC class I and II binding peptides owing to inherent differences in amino acid composition compared to structured domains. In contrast, linear B-cell epitopes were predicted to be enriched in IDPs. Tandem repeat regions and non-synonymous single nucleotide polymorphisms were found to be strongly associated with regions of disorder. In summary, immune responses against IDPs appear to have characteristics distinct from those against structured protein domains, with increased antibody recognition of linear epitopes but some constraints for MHC presentation and issues of polymorphisms. These findings have major implications for vaccine design, and understanding immunity to malaria.

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.

Introduction

Intrinsically disordered proteins (IDPs) are an important class of proteins characterised by a high degree of flexibility and lack of a well-defined three-dimensional structure [1]. They have been shown to play significant roles in many cellular processes, including protein-ligand binding, DNA and RNA binding, and as flexible linkers [2–5]. Other roles for IDPs relate directly to their entropic properties, such as their proposed functions as molecular springs or in the timing of molecular processes (entropic clocks) [6–9]. Whilst many studies have examined the functional roles of disordered proteins, their immunogenic and antigenic properties have received relatively little attention.

Computational studies have shown a higher proportion of IDPs in the proteomes of eukaryotic species as compared to prokaryotes [10–12], with the proteomes of apicomplexan parasites being particularly enriched in IDPs [13]. Of the apicomplexan parasites that infect human hosts, *Plasmodium falciparum* is responsible for the highest number of deaths worldwide, although other species including *P. vivax* also contribute significantly to the global malaria disease burden [14]. There is an urgent need for an effective malaria vaccine, and a major challenge is to identify key antigens that are targeted by protective immune responses and to design vaccine constructs that generate highly effective and long-lasting immunity. Several current vaccine candidates for *P. falciparum* malaria such as CSP, MSP2, MSP3, EBA-175 RIII-V and SERA5 are targets of functional antibody responses [15–20] and are composed partly or almost entirely of disordered regions [16,21–27].

IDPs contain a number of features that may affect adaptive immune responses against *Plasmodium* spp. Firstly, the reduced proportion of bulky, hydrophobic residues in IDPs [12,28] has potential implications for peptide binding to MHC class I and II molecules, as highlighted by recent work suggesting that disordered regions across a number of species contain a paucity of MHC-binding peptides [29]. Secondly, tandem repeat regions are thought to be prevalent within IDPs, with evidence suggesting that evolution of IDPs is sometimes driven by expansion of tandem repeat regions [30,31]. Tandem repeat regions have the potential to be immunodominant [32–34] (e.g. in the sequence of the RTS, S vaccine), with certain repeat motifs capable of inducing both T-cell-dependent and T-cell-independent B-cell responses [35–37]. Finally, the occurrence of non-synonymous single nucleotide polymorphisms (SNPs) in some *P. falciparum* genes has been linked to immune selection pressure [38–40], with evidence from other organisms suggesting that positive selection of non-synonymous SNPs occurs at a higher rate within IDPs [41].

We hypothesised that IDPs are likely to represent major immune targets in *P. falciparum* and are likely to be important vaccine candidates. We sought to determine if characteristics that have been observed for IDPs of other organisms were also found in IDPs of *P. falciparum* and to ascertain the relevance of these characteristics in vaccine construct design. Using a variety of computational techniques, we established that IDPs within the *P. falciparum* proteome are abundant in immunologically-exposed subcellular locations and contain a high proportion of linear B-cell epitopes. We also determined that IDPs have a reduced proportion of MHC-binding peptides compared with ordered domains, which may adversely affect T-cell help. They also have a higher proportion of tandem repeats and polymorphisms, creating additional, but not insurmountable challenges for vaccine construct design. This study has significant implications for understanding the generation of adaptive immune responses, either through natural exposure or vaccination against IDPs, and the development of bioinformatics tools to assist in the development of future vaccine constructs.

Results

Sequences from the entire *P. falciparum* proteome were interrogated using established predictors of protein disorder, MHC class I and II binding, linear B-cell epitopes and tandem repeat regions. Information on protein localisation for *P. falciparum* was obtained from ApiLoc [42] and single nucleotide polymorphisms (SNPs) were obtained from PlasmoDB. Protein sequences for other *Plasmodium* spp. capable of infecting humans (*P. vivax* and *P. knowlesi*) and mice (*P. berghei*, *P. chabaudi*, and *P. yoelii*) were also assessed using these predictors to enable comparison across *Plasmodium* spp. Results from these predictors were stored in a local PostgreSQL database and subjected to further analysis using custom Python and R scripts (Fig 1).

High proportions of the Plasmodium proteome are intrinsically disordered

We considered disorder at both a per-proteome level (i.e. the number of residues across the proteome that fall within disordered regions; expressed as a proportion of residues for the entire proteome) and at a per-protein level (the percentage of predicted disordered residues for each protein). IDPs constituted a significant proportion of the proteomes of the six *Plasmodium* species assessed. On a per-proteome basis, the proportions of the proteomes predicted to be disordered were as follows: *P. falciparum* 32.7%, *P. vivax* 33.2%, *P. knowlesi* 30.6%, *P. berghei* 26.7%, *P. chabaudi* 27.6% and *P. yoelii* 27.5%. The median degree of disorder per-protein for *P. falciparum* was 15.5% (IQR = 6.7–31.6%; Fig 2A). No significant differences between the proportion of disorder per-protein were observed among any of the *Plasmodium* spp. tested ($p > 0.05$, Kruskal-Wallis rank sum test). After combining the results for the six *Plasmodium* spp. tested, the median disorder per-protein was 15.1% (IQR = 7.0–29.7%). Several leading *P. falciparum* vaccine candidates were also assessed to determine the proportion of these proteins that are disordered. There was a significant proportion of disorder among many of these proteins including: 1) pre-erythrocytic antigens: CSP (75.1%), LSA1 (40.6%), TRAP (47.7%); 2) erythrocytic stage antigens: MSP1 (59.1%), MSP2 (72.4%), MSP3 (52.3%), EBA175 (46.6%), AMA1 (21.5%), RESA (50.5%), Rh5 (8.0%), GLURP (95.4%), SERA5 (29.1%); and 3) sexual stage antigens: Pfs25 (5.5%) and Pfs230 (21.3%). The distribution of this disorder is shown for some selected examples, highlighting the heterogeneity of disorder amongst leading vaccine candidates, and demonstrating that vast regions of some these proteins may be almost entirely disordered (Fig 3).

Disordered proteins are abundant within apical organelles, parasitophorous vacuole, exported proteins and the nucleus

Protein localisation data for 451 *P. falciparum* proteins were obtained from the curated ApiLoc database [42], but there was very limited protein localisation data for other *Plasmodium* spp. The prevalence of per-protein disorder in various subcellular locations was highest in nuclear proteins (median = 28.0%), parasitophorous vacuole (PV) proteins (median = 27.7%), exported proteins (median = 27.7%) and apical proteins (median = 23.4%). Median protein disorder was lowest in the endoplasmic reticulum proteins (median = 8.9%) and mitochondrial proteins (median = 8.8%). (Fig 2B, S1 Table). All of these values were significantly different from the median degree of disorder across the whole proteome ($p < 0.001$ for each, Wilcoxon rank-sum test).

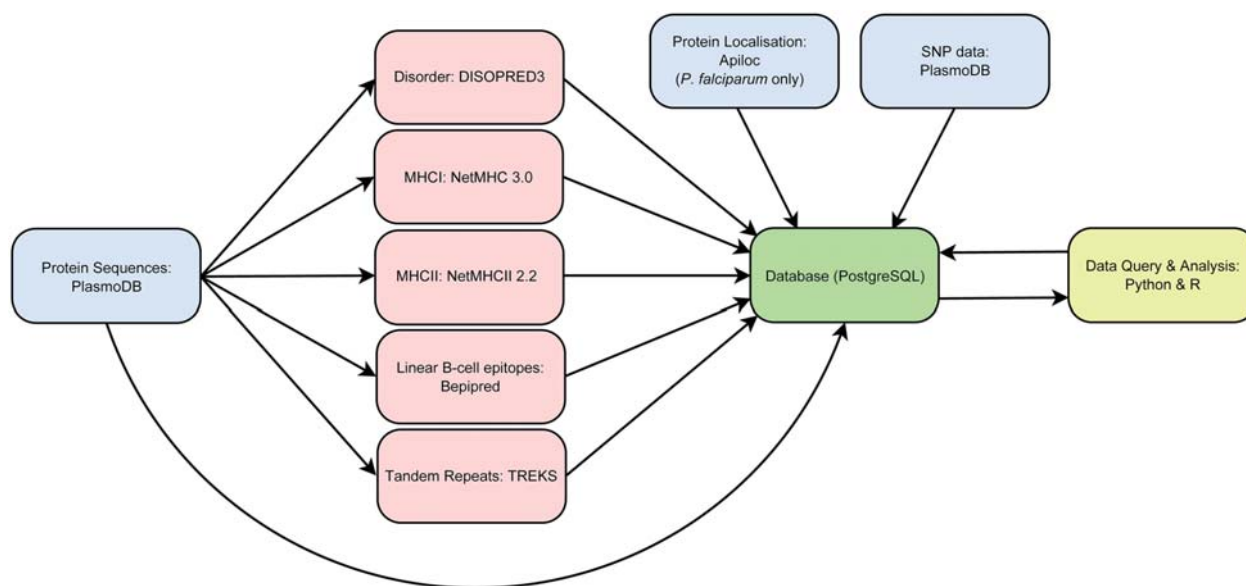


Fig 1. Computational workflow used for analysis of the proteome of *Plasmodium* spp. Protein coding sequences were obtained from PlasmoDB, and submitted to predictors of protein disorder, MHC binding, linear B-cell epitopes and tandem repeats. Protein localisation data for *P. falciparum* was obtained from ApiLoc and non-synonymous single nucleotide polymorphisms (SNPs) were obtained from PlasmoDB. All data were stored in a local PostgreSQL database and queried using custom Python scripts. Statistical analysis and data visualisation were performed using the R statistical computing package.

doi:10.1371/journal.pone.0141729.g001

Disordered proteins contain a biased amino acid composition

It has been shown previously that the amino acid composition of IDPs is distinct from that of structured proteins [12,28]. An assessment of the amino acid composition of ordered and disordered regions in the *P. falciparum* proteome revealed a marked reduction in aromatic residues tryptophan (W), tyrosine (Y) and phenylalanine (F), with a 76%, 45% and 64% reduction respectively. There was also a reduction in hydrophobic residues proline (P), alanine (A), valine (V), leucine (L), and isoleucine (I) in IDPs. Cysteine (C) was also significantly under-represented within disordered domains, with a 53% reduction relative to ordered regions. There was a corresponding increase within disordered regions in the proportion of charged or hydrophilic residues including aspartic acid (D), glutamic acid (E), lysine (K), asparagine (N) and glutamine (Q), with D, E and N being increased at least 50% relative to ordered regions (Fig 4).

Disordered proteins contain fewer predicted MHC binding peptides

To assess the effect of protein disorder on the predicted presentation of *P. falciparum* peptides via MHC class I and MHC class II, we employed *in silico* prediction of peptide binding to MHC. For each HLA allele, we defined the proteome coverage as the percentage of residues across the *P. falciparum* proteome that are part of a predicted high-affinity peptide ($IC_{50} < 50\text{nM}$). The median coverage of high-affinity peptides across all HLA alleles was then calculated. For MHC class I, the median coverage was 3.3% and 1.4% within ordered and disordered regions, respectively ($p < 0.0001$, Wilcoxon rank sum test), equating to a ~2.3-fold decrease within disordered regions (Fig 5). For MHC class II, which is especially important for effective antibody responses, the median coverage was 12.1% and 3.5% within ordered and disordered regions, respectively ($p < 0.0001$, Wilcoxon rank sum test), equating to a ~3.5-fold decrease within disordered regions (Fig 5). When lowering the positive threshold to include predictions for both high and low MHC affinity (predicted $IC_{50} < 500\text{nM}$), the median coverage for MHC

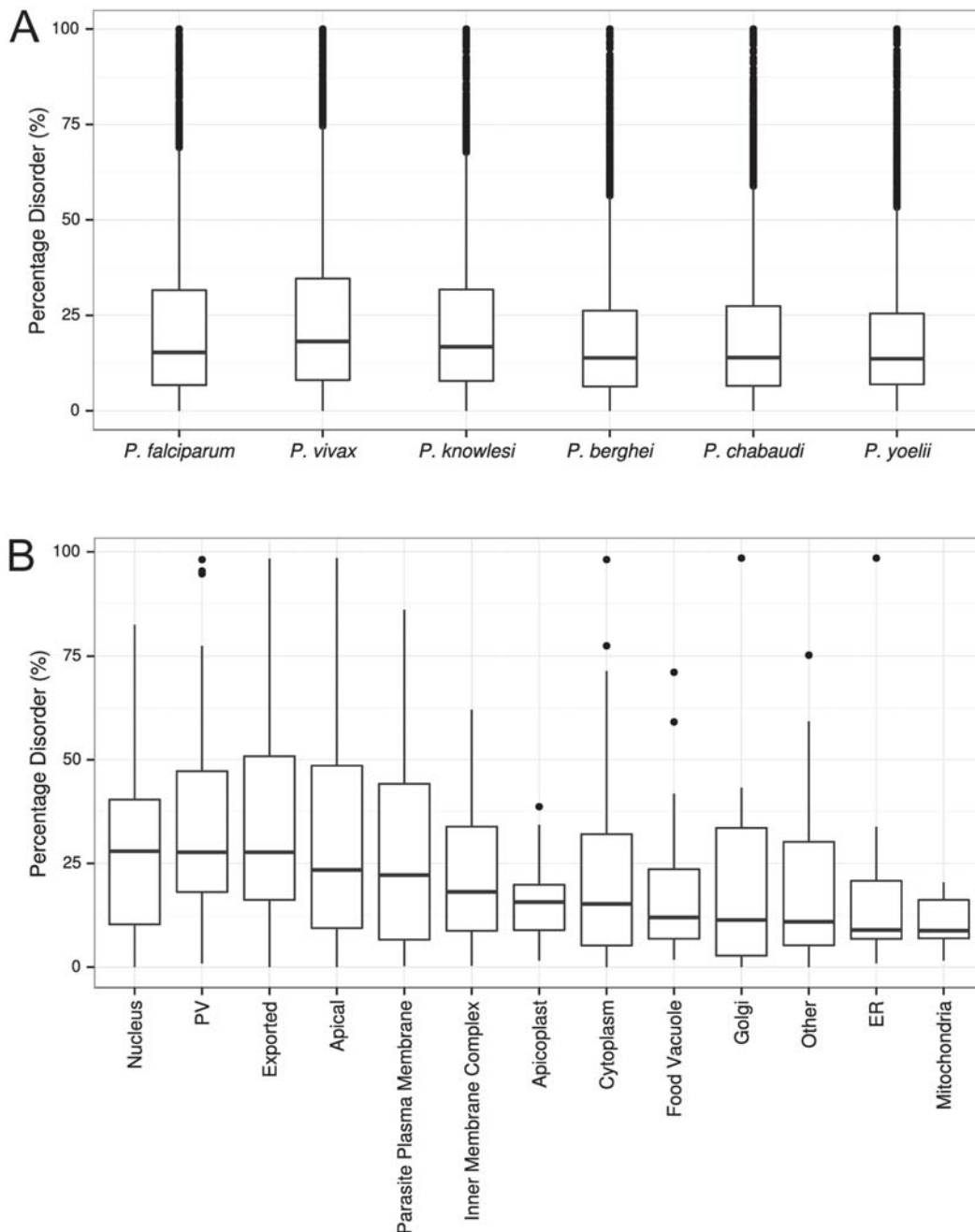
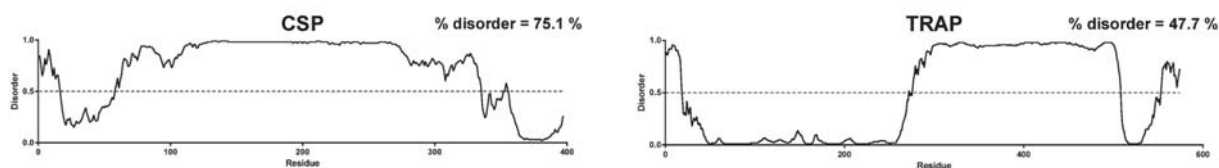


Fig 2. Prediction of protein disorder for various *Plasmodium* spp. and within subcellular locations for *P. falciparum*. **A)** Distribution of protein disorder within the proteome of each *Plasmodium* spp. at the level of individual proteins. **B)** Prediction of protein disorder for *P. falciparum* proteins according to subcellular localisation. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. Percentage disorder was calculated as the proportion of residues predicted to be disordered at the level of individual proteins. Prediction of disorder was performed using DISOPRED3.

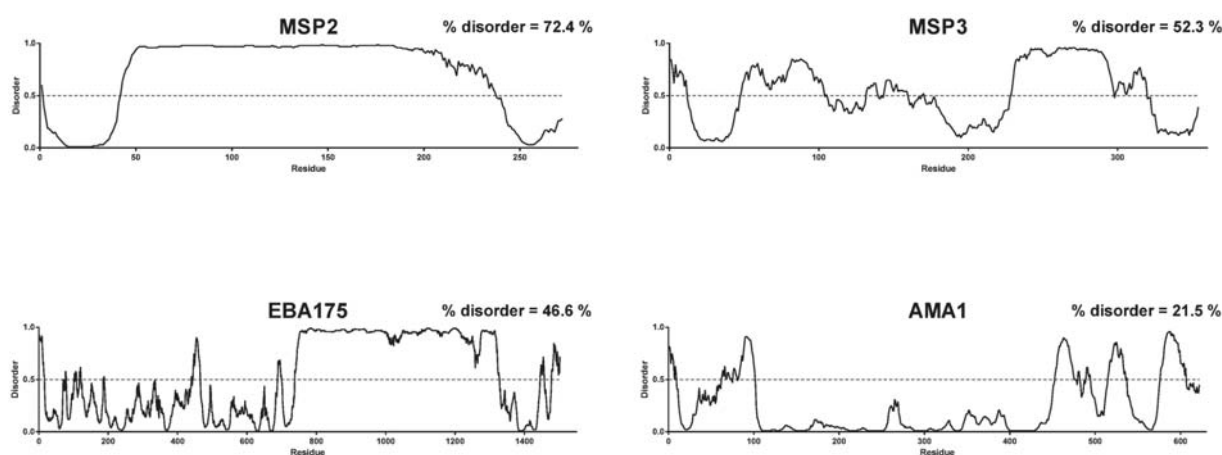
doi:10.1371/journal.pone.0141729.g002

class I was 17.5% within ordered regions and 6.1% within disordered regions ($p < 0.0001$, Wilcoxon rank sum test), while the median coverage for MHC class II was 42.3% within ordered regions and 15.1% within disordered regions ($p < 0.0001$, Wilcoxon rank sum test). When individual HLA haplotypes were assessed, decreased MHC class I and MHC class II epitopes in

Pre-erythrocytic Stage



Erythrocytic Stage



Sexual Stage

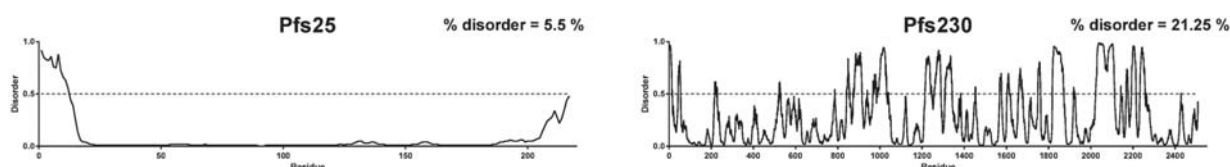


Fig 3. Predicted protein disorder for a number of leading *P. falciparum* vaccine candidates. Disorder predictions were performed using DISOPRED3. A disorder score above 0.5 is indicative of a disordered region (dashed line). All sequences used were from the *P. falciparum* 3D7 strain.

doi:10.1371/journal.pone.0141729.g003

disordered proteins were consistently observed compared with ordered proteins for each haplotype (S1 Fig). These findings presumably reflect the reduced proportion of hydrophobic and aromatic residues within disordered domains, resulting in a reduced ability to bind MHC molecules with high affinity. There was also considerable heterogeneity observed in predicted affinities between different haplotypes, which may have implications for immunity in genetically diverse populations.

To assess the possibility of biased MHC binding within different subcellular compartments, we examined the proportion of MHC class I and MHC class II binding peptides within the subset of proteins described in the ApiLoc database (S2 Fig). Peptides were grouped according to protein disorder and subcellular protein location. No significant difference in MHC class I or MHC class II binding was observed between subcellular locations for high-binding peptides ($p > 0.05$, Kruskal-Wallis rank sum test).

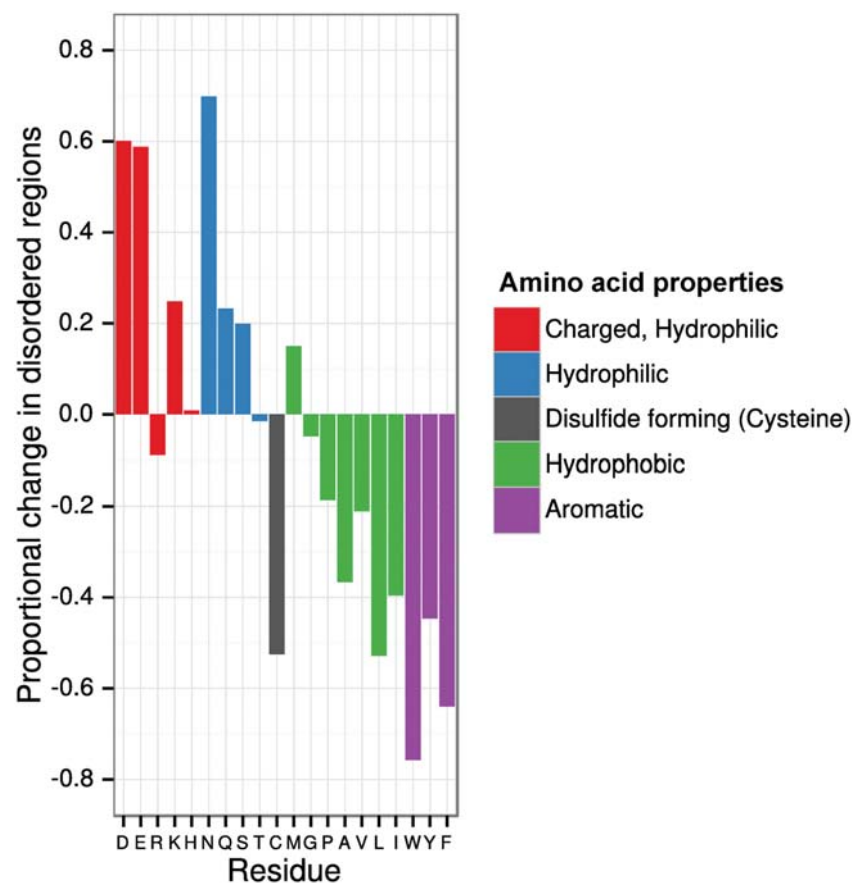


Fig 4. Intrinsically disordered protein domains contain a biased amino acid composition. The relative proportional change in amino acid frequency within disordered regions was calculated for the entire *P. falciparum* proteome, with comparison made to all predicted ordered regions within the proteome.

doi:10.1371/journal.pone.0141729.g004

Reduced MHC binding reflects biased amino acid composition at key peptide anchor points

To identify potential sequence determinants that affect binding to MHC class I and MHC class II molecules, we analysed the position-dependent sequence composition of predicted high-affinity peptides (Fig 6). Peptides were classified as being part of an ordered region, a disordered region, or on the boundary of the two. Inherent differences in sequence composition between ordered and disordered regions (Fig 4) were taken into account when calculating the proportional enrichment of each residue in MHC-binding peptides (see Methods), with data presented as a weighted average across all regions (disordered/ordered/mixed). For MHC class I binding peptides, it was observed that sequence composition differed most from background levels at positions 2 and 9 of all predicted binding peptides (Fig 6A). Importantly, greater than 100% enrichment was observed for methionine (M) and leucine (L) at position 2, and arginine (R), valine (V) and leucine (L) at position 9. There was also a tendency for aromatic residues (W, Y and F) to be enriched at most other peptide positions. For MHC class II binding peptides, we considered only the predicted central core-binding region (as defined by the NetMHCII algorithm [43,44]). It was observed that sequence composition differed most from background levels at position 1 of predicted core-binding regions (Fig 6B). Phenylalanine (F)

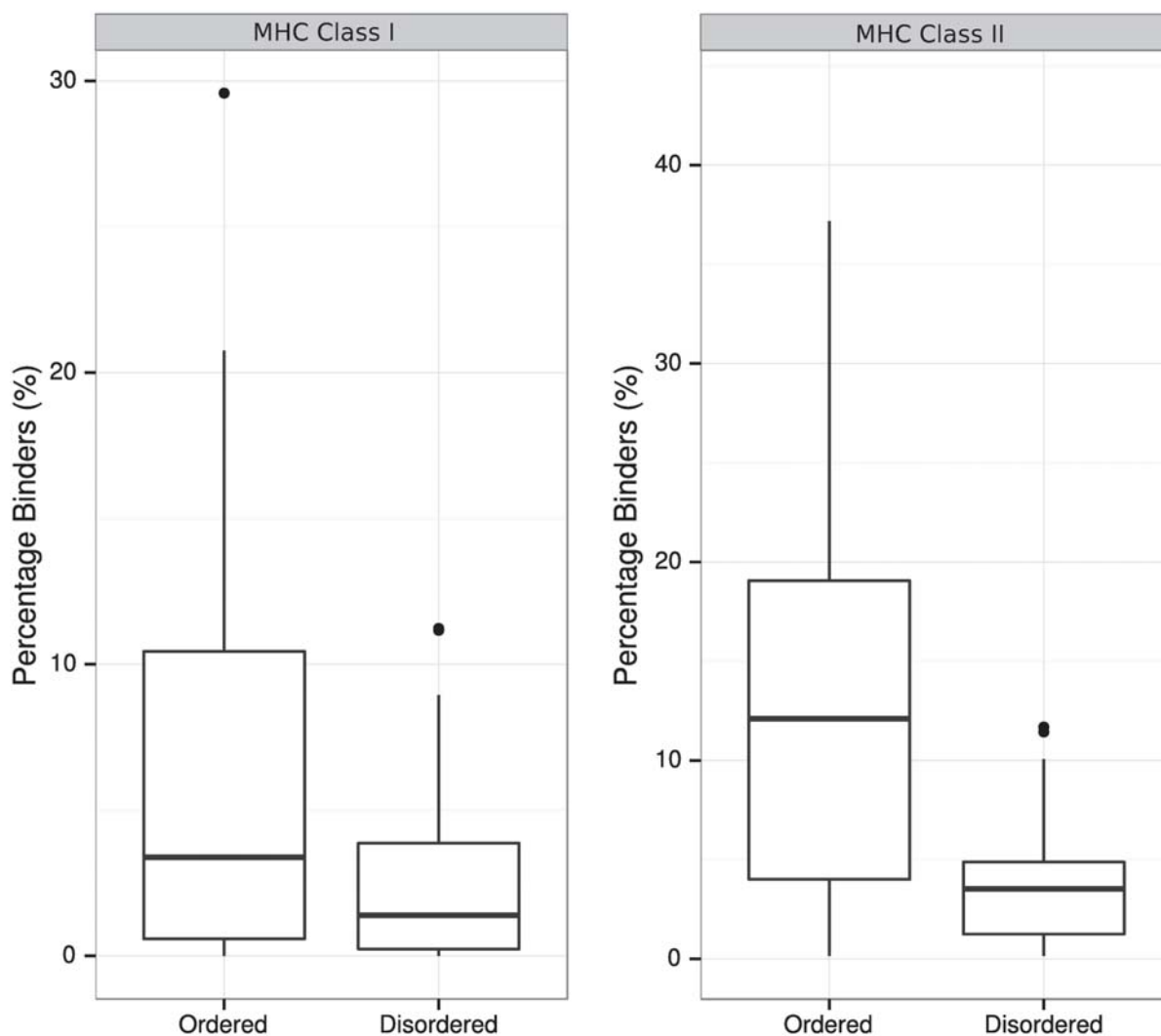


Fig 5. Reduced MHC I and MHC II binding in disordered proteins for *P. falciparum*. The proportion of peptides with predicted high affinity to MHC I and MHC II is significantly higher for peptides within an ordered protein domain. Boxplots represent the distribution of MHC-binding peptides across all MHC alleles tested.

doi:10.1371/journal.pone.0141729.g005

and tyrosine (Y) were particularly enriched at this position, with enrichment of leucine (L), isoleucine (I) and tryptophan (W) observed to a lesser extent. Residues that are enriched within MHC binding peptides are generally found at lower frequency within disordered regions ([S3 Fig](#)); for example, aromatic residues such as F and Y are found at much lower frequency within disordered regions, while being present at much higher frequency in position 1 of predicted MHC class II binding peptides. Similarly, hydrophilic residues are found very rarely at position 1 of MHC class II binding peptides, yet are generally enriched within disordered regions.

To determine if the position of amino acid residues within the MHC binding regions or the amino acid residue characteristics themselves biased these results, we scrambled sequences from within disordered and ordered regions, and submitted scrambled sequences to predictors of MHC I and MHC II binding. There was a small but statistically significant difference between observed MHC binding for the actual sequences, and that of the scrambled sequences (with the exception of MHC II binding within ordered regions) ([S4 Fig](#)). These shifts were

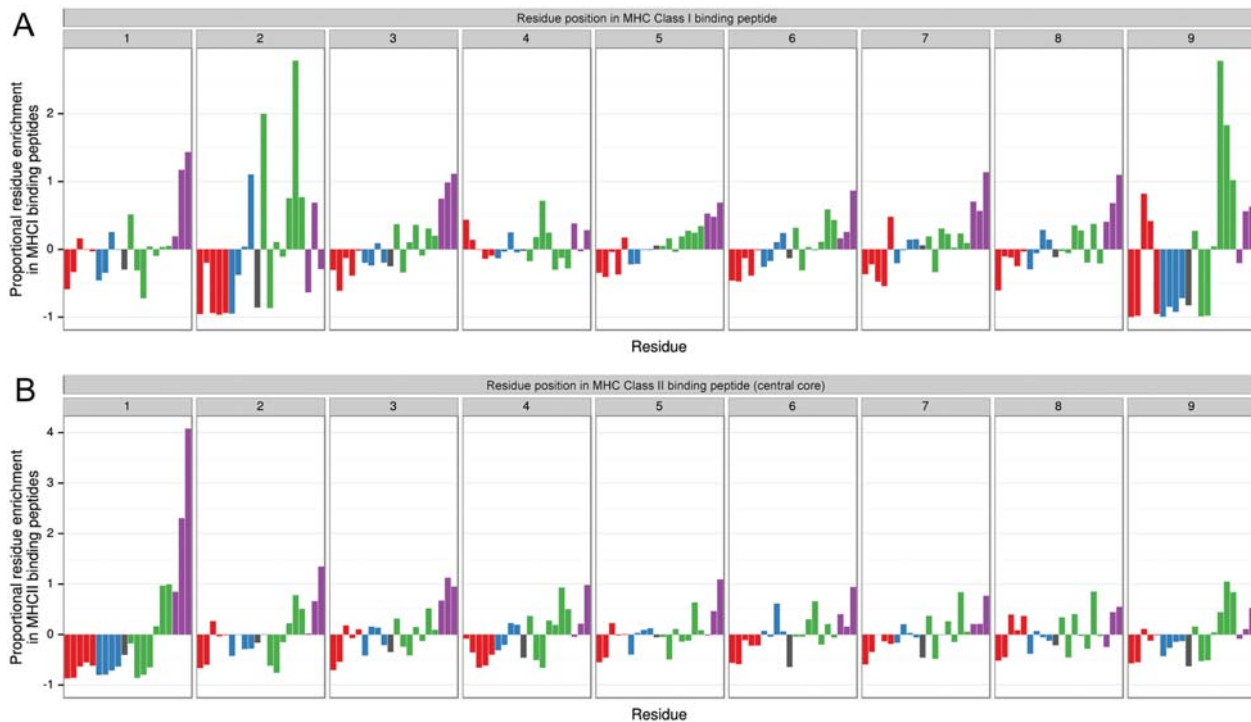


Fig 6. Position-specific enhancement or depletion of residues in MHC binding peptides. Residue abundance in MHC class I (A) and MHC class II (B) binding peptides was calculated relative to the abundance in the background proteome. This calculation was performed for residues at each position in an MHC binding peptide. Adjustment was made for differing sequence composition in disordered versus ordered regions, with results presented as a weighted average across all regions. Amino acid residue labels are omitted for visual clarity—refer to Fig 4 for residue order and colouring.

doi:10.1371/journal.pone.0141729.g006

small compared to the large bias in MHC binding between ordered and disordered regions, suggesting that this bias in MHC binding between regions is predominantly due to sequence composition alone, rather than the result of selective pressure or other functional/structural requirements of MHC-peptide interactions.

Linear B-cell epitopes are more abundant in disordered proteins

The occurrence of linear B-cell epitopes was predicted across the *P. falciparum* proteome and compared to the occurrence of predicted disorder, with comparison made at a per-residue level. The proportion of residues predicted to contain linear B-cell epitopes was assessed with the BepiPred algorithm using a range of different output thresholds that reflect the sensitivity and specificity of detecting a linear B-cell epitope (i.e. lower output threshold has high sensitivity but low specificity, and high output thresholds have low sensitivity but high specificity). Across all these output thresholds, linear B-cell epitopes were predicted to be more common in regions of disorder than in structurally ordered regions (Fig 7; solid line and dashed lines respectively). A comparison with other *Plasmodium* spp. showed that *P. vivax* contained the highest proportion of predicted linear B-cell epitopes, for both ordered and disordered regions, whereas *P. falciparum* contained the second lowest proportion.

When examining the distribution of predicted linear B-cell epitopes within various subcellular compartments, proteins localised to the PV, parasite plasma membrane proteins, exported proteins, apical proteins and nuclear proteins all had a significantly higher percentage of predicted linear B-cell epitopes compared to the background proteome ($p < 0.001$ for all except

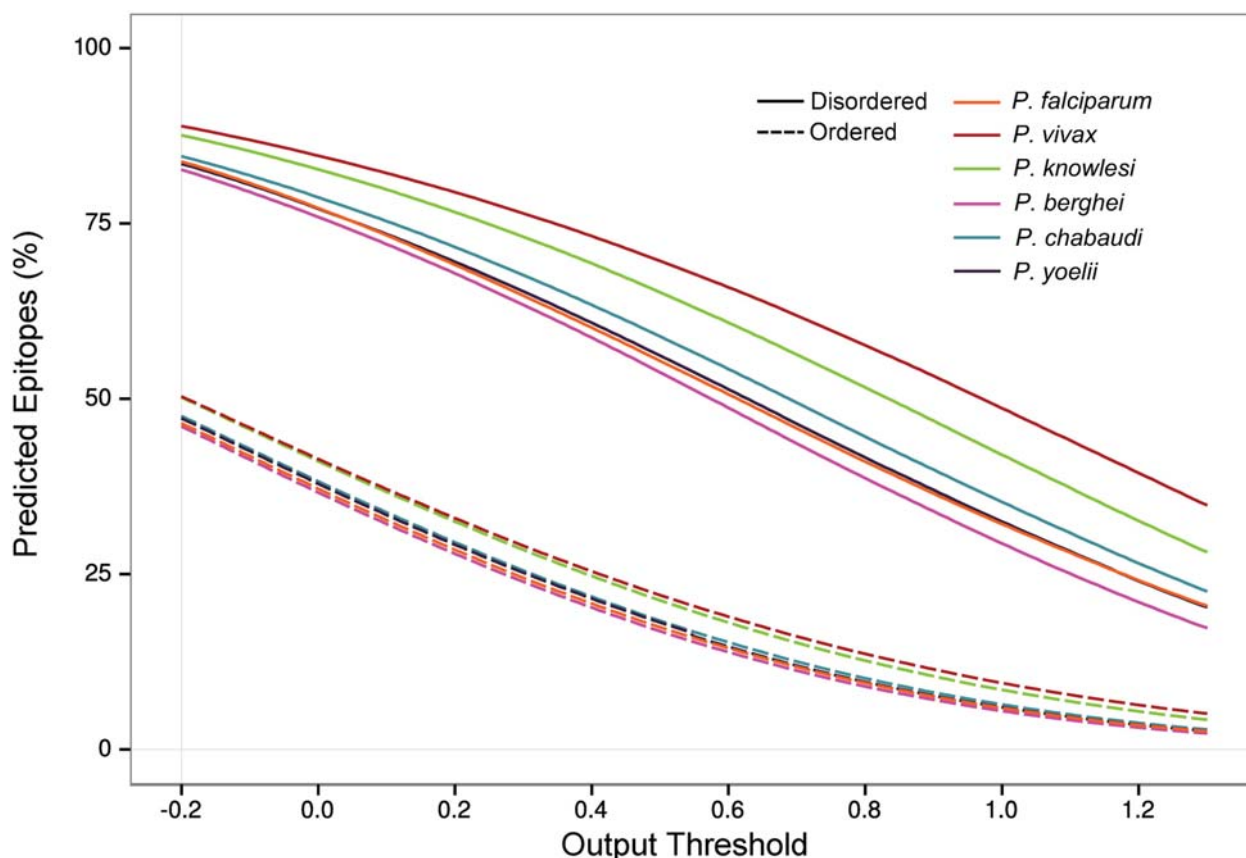


Fig 7. The proportion of predicted linear B-cell epitopes was higher in IDPs for all *Plasmodium* spp. Classification of disorder was achieved using DISOPRED3. BepiPred was used for prediction of linear B-cell epitopes. The number of predicted linear B-cell epitopes as a percentage of all residues is shown across a range of BepiPred output thresholds. The corresponding sensitivity/specificity for each output threshold is given at <http://www.cbs.dtu.dk/services/BepiPred/output.php>. Thresholds range from -0.2 (sensitivity = 0.75, specificity = 0.5) to 1.3 (sensitivity = 0.13, specificity = 0.96).

doi:10.1371/journal.pone.0141729.g007

parasite plasma membrane, $p = 0.005$; Wilcoxon rank sum test). Residues were then grouped according to predicted protein disorder, as predicted linear B-cell epitopes were correlated with predicted disorder. Levels of predicted linear B-cell epitopes remained significantly higher in PV proteins, exported proteins and nuclear proteins after accounting for protein disorder (S5 Fig and S2 Table).

Tandem repeat regions are more common in disordered proteins

The occurrence of tandem repeat sequences within the *P. falciparum* proteome and the relationship to regions of structural disorder was examined to assess the potential role of IDPs in the generation of immunodominant antibody responses. Tandem repeat sequences were identified using T-REKS [45] with a Psim cut-off of 0.8. When grouped according to protein disorder, tandem repeats make up 1.7% of ordered regions, compared to 12.9% of disordered regions. Of all the identified tandem repeat regions, 79% fell within predicted disordered regions. To assess potential bias in the occurrence of tandem-repeat domains within different subcellular compartments, we analysed the occurrence of tandem repeats for the subset of proteins in the ApiLoc database (Fig 8A). Compared to the total *P. falciparum* proteome, exported proteins ($p = 0.004$) and proteins localised to the PV ($p = 0.02$) had a significantly higher

percentage of tandem repeats (Wilcoxon rank-sum test). Lower levels of tandem repeats were observed in proteins in the cytoplasm ($p = 0.01$), endoplasmic reticulum ($p = 0.03$), apicoplast ($p = 0.001$) and mitochondria ($p < 0.001$) (S3 Table).

Non-synonymous single nucleotide polymorphisms are more common in disordered proteins

The occurrence of non-synonymous SNPs in the *P. falciparum* proteome was analysed, with residues grouped according to predicted protein disorder. The percentage of disordered regions that are polymorphic was 2.5%, compared to 1.0% of ordered regions ($p < 0.001$, Pearson's chi-squared test). When proteins were grouped according to subcellular location, an increased proportion of SNPs (as compared to the *P. falciparum* proteome) was observed in exported proteins ($p = 0.001$, Wilcoxon rank sum test), proteins localised to the PV ($p = 0.001$, Wilcoxon rank sum test) and apical proteins ($p < 0.0001$, Wilcoxon rank sum test) (Fig 8B, S4 Table).

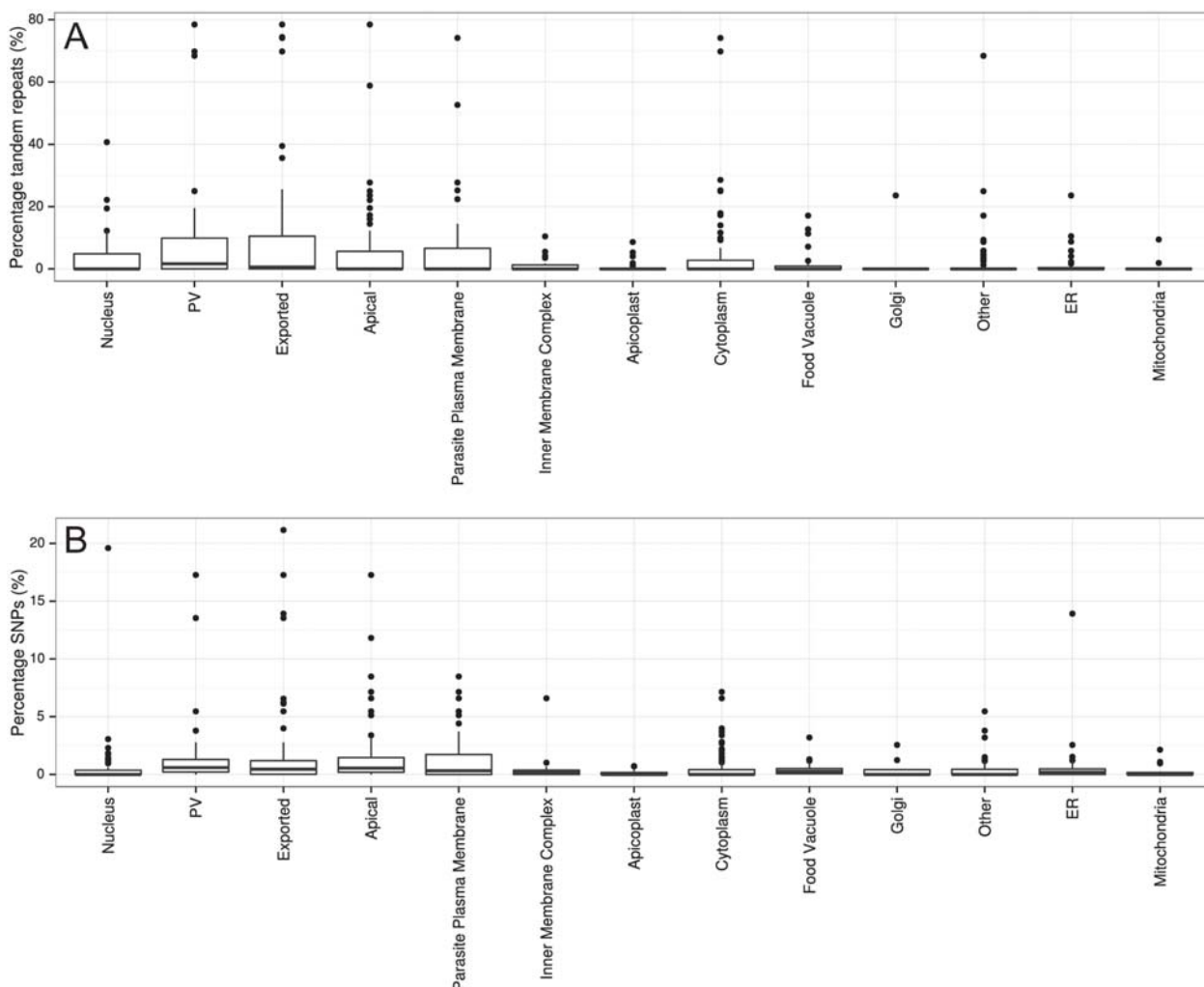


Fig 8. Distribution of tandem repeats and non-synonymous SNPs within different subcellular compartments for *P. falciparum*. (A) Prediction of tandem repeats was performed using T-REKS. Percentage tandem repeats was defined as the percentage of each protein that contains tandem repeat sequences. (B) Percentage SNPs was calculated as the percentage of residues within each protein that contain identified SNPs with a minimum minor allele frequency of 5%. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location.

doi:10.1371/journal.pone.0141729.g008

Discussion

The last few decades have seen an increased understanding of the role of IDPs in various biological systems and an appreciation of their functional importance. Numerous experimental techniques have been employed to identify and characterise IDPs [46], and these have been complemented by a number of computational algorithms developed to predict the occurrence of protein disorder using protein sequence data [47]. Little work has been performed examining the immunological properties of this class of protein, perhaps due to the relative scarcity of IDPs in most bacteria and viruses [10,11]. Eukaryotic organisms, however, are known to contain a relative abundance of IDPs, with apicomplexan parasites including *Plasmodium* and *Toxoplasma* spp. being particularly enriched in IDPs [13]. It is therefore important to understand the potential differences in immune recognition of IDPs compared to ordered protein domains. In this study, we applied a number of computational prediction algorithms at a proteomic level to gain further insight into the role of IDPs as potential antigenic targets.

A high proportion of the *P. falciparum* proteome was predicted to be disordered and this was also the case for other *Plasmodium* spp. IDPs appear to be especially enhanced in apical and exported proteins, suggesting that they may play functional roles in parasite invasion and sequestration, and that they are also likely to be accessible to antibody recognition on intact parasites. Almost nothing is known about the actual functional role of IDPs in *Plasmodium* spp. It is possible that they may play a role as flexible linkers between ordered domains, enabling rapid molecular recognition of host ligands during parasite invasion. This is potentially the case for the erythrocyte binding-like (EBL) family of proteins, which contain a disordered domain termed region III-V (RIII-V) [48]. The EBL family of proteins are found in *P. falciparum*, *P. vivax* and *P. knowlesi*, and includes proteins such as the Erythrocyte Binding Antigens (e.g. EBA-140, EBA-175, EBA-181) and Duffy Binding Protein (DBP). Molecular recognition of erythrocyte receptors by the EBL family of proteins occurs via an N-terminal structured domain termed region II (RII) [39,49,50]. We hypothesise that recognition and binding of erythrocyte receptors via RII is expedited by the flexibility of the adjacent RIII-V domain. Of note, antibodies to the disordered RIII-V of EBA175 can inhibit erythrocyte invasion [18,51]. Similarly, antibodies to the repeat region of CSP can inhibit sporozoite infection of hepatocytes [52], while antibodies against MSP2 can fix complement components to inhibit erythrocyte invasion [53]. These findings indicate the importance of IDPs in host cell invasion and as immune targets.

We observed a general enrichment of charged and hydrophilic residues within disordered regions of *P. falciparum*, with a corresponding decrease in the proportion of aromatic and hydrophobic residues. These observations are consistent with previous studies of IDPs [12,28], although enrichment of D and N was not observed in the study by Dunker *et al.* [12], while enrichment in N was not observed by Radivojac *et al.* [28]. It is important to note that neither of these studies accounted for potential biases in amino acid usage between species, which may explain the observed differences between studies. Our observed reduction in aromatic and hydrophobic residues in IDPs was noted to affect key peptide anchor points for MHC class I and class II binding, supporting recent findings by Mitic *et al.* [29]. Both MHC class I and II presentation require peptides to be anchored in the MHC binding groove through interactions with hydrophobic binding pockets and additional interactions with the floor and walls of the binding channel [54–57]. Reduced MHC class II binding is likely to reduce antigen presentation to helper CD4⁺ T-cells and the acquisition of effective antibody responses, while reduced MHC class I binding is likely to reduce antigen presentation to CD8⁺ T-cells which are required for immune responses against the liver stage infection [58,59]. MHC class I and class II epitopes were identified within IDPs, however, indicating that these potential limitations

may be overcome by careful design of vaccine constructs and a detailed knowledge of the HLA haplotypes of target populations.

Using a sequence-based linear B-cell epitope prediction method [60], it was noted that predicted linear B-cell epitopes were significantly enriched within IDPs. This is not surprising considering that most of the polypeptide chain within an IDP is accessible to antibody binding. Several studies have characterised linear epitopes within IDPs [24,61–64], although there is a notable report describing the existence of a discontinuous epitope with an IDP [65]. We do note that current sequence-based prediction algorithms for linear B-cell epitopes should be used with caution as there is some concern that they perform relatively poorly compared to similar predictors for MHC binding [66]. While this could be due to inherent structural differences between antibody-antigen and MHC-peptide complexes, it has been suggested that current training datasets and classification methods for B-cell epitope predictors are inadequate [67]. Indeed, our recent study of *P. falciparum* MSP2 found that B-cell epitope predictors were poor predictors of individual immunogenic epitopes within this largely disordered protein [68].

It has been previously postulated that IDPs, and tandem repeat regions in particular, may play an important role in the immune evasion of various parasites including *Plasmodium* [37,69,70], *Trypanosoma* [32], *Leishmania* [33,71] and *Ehrlichia* [34]. Tandem repeat regions may induce immunodominant responses that act as immunological decoys, masking responses against functionally important epitopes. This hypothesis is consistent with our finding that a high proportion of IDPs are located in immunologically-exposed subcellular compartments and the observation that tandem repeat regions are predominantly located within IDPs. Repeat protein sequences bear some similarity to the repeated structural motifs found on bacterial polysaccharides that are known to elicit T-cell-independent type 2 B-cell responses. Although such responses have been described against polysaccharide antigens, there is evidence suggesting that some protozoan proteins contain tandem repeats that can act as T-cell-independent type 2 antigens [35–37], negating the need for CD4⁺ T-cell help. Perhaps more importantly, immunodominant responses against protein tandem repeats may develop due to the increased avidity of an antibody to a region in which identical epitopes are located at several adjacent regions of the polypeptide, increasing the apparent epitope concentration within the vicinity of bound antibody. This is likely to lead to rapid antibody re-binding upon dissociation, resulting in high antibody avidity, despite a relatively lower affinity for a single epitope site. In this way, selection of B-cells from the naïve B-cell repertoire is likely to be biased towards cells with reactivity to such tandem repeat domains.

Within *P. falciparum*, there is mixed evidence for the immunodominant nature of tandem repeats. For example, strong antibody responses are acquired naturally against the immunodominant NANP repeat region of CSP [72–74] that is predicted to adopt a coiled-coil structure [75] and antibodies against SERA-5 predominantly target a disordered N-terminal octamer repeat [19]. In contrast, some disordered tandem-repeats within MSP2 (3D7 allele) are poorly immunogenic, which has been attributed to a high degree of conformational flexibility compared to the rest of the sequence [68]. Taken together, protein tandem repeats appear to be immunodominant in some cases, but may be poorly immunogenic in others, possibly due to high flexibility and a large loss of conformational entropy upon antibody binding.

Our observation that IDPs contain a higher proportion of amino acid polymorphisms as a result of non-synonymous SNPs is consistent with previous studies and has important implications for vaccine design. Recent work in *Saccharomyces cerevisiae* showed that IDPs contained a higher proportion of non-synonymous SNPs, with disordered regions shown to be under weaker negative selection than ordered domains [76]. This was attributed to reduced structural constraints for disordered regions, being more tolerant to amino acid changes, especially

changes to amino acids with similar characteristics. Many non-synonymous SNPs within *P. falciparum* appear to be maintained as a result of immune selection pressure, with many of the resulting polymorphisms located on the protein surface and hence accessible to antibody binding [77]. IDPs have a higher proportion of residues accessible for antibody binding which may contribute to the observed increase in non-synonymous SNPs to some degree. Interestingly, only a small number of genes (~100) within the *P. falciparum* proteome have been observed to be under balancing selection [78], and hence it is unlikely that immune pressure alone is responsible for the observed increase in polymorphic residues within IDPs [79].

Conclusions

The role of IDPs as antigenic targets is poorly understood, despite their relative abundance in major human pathogens such as *P. falciparum*. We have shown here that the biased amino acid composition of IDPs can limit their presentation via MHC molecules and may influence the generation of antibody responses and B-cell memory. Furthermore, we have demonstrated that immunologically-exposed subcellular compartments within *P. falciparum* have a higher proportion of IDPs, a greater number of tandem repeat regions, and a higher incidence of non-synonymous SNPs. This indicates that IDPs can be involved in generating immunodominant antibody responses, and that some may play a role in immune evasion. Despite these apparent limitations, it is clear that some IDPs are targeted by functional immune responses and that some of these antigens are realistic vaccine candidates. Indeed, we have shown that a number of leading candidates contain a significant proportion of disordered regions. These findings have major implications for vaccine design, and understanding immunity to malaria.

Methods

Protein sequences

Protein sequences for *P. falciparum* (3D7), *P. vivax* (Sal-1) [80], *P. chabaudi* (chabaudi) [81], *P. berghei* (ANKA) [81] and *P. knowlesi* (Strain H) [82] were obtained from PlasmoDB [83] (<http://plasmodb.org>). All protein-coding sequences were selected for each organism with pseudo-genes excluded, and protein sequences downloaded in FASTA format.

Disorder prediction

DISOPRED3 software was used for prediction of protein disorder [84], and was chosen due to its high ranking in independent benchmarking [85]. The DISOPRED3 algorithm utilises a combination of a support vector machine (SVM), artificial neural network (ANN) and nearest-neighbour classifier to classify residues as disordered/ordered. An initial PSI-BLAST search is also used to create a sequence profile that is then passed to the SVM. Generation of sequence profiles using PSI-BLAST was performed using the UniRef90 protein database and blast-2.2.26 software package from NCBI. PSI-BLAST was run with 3 passes, with an e-value threshold of 0.001 for inclusion in the multi-pass model. PSI-BLAST checkpoint file was saved and used as an input to the DISOPRED2 SVM algorithm (part of the DISOPRED3 prediction workflow). The default threshold was used for DISOPRED3, with a disorder score above 0.5 indicating predicted disorder. DISOPRED3 software is freely available and was obtained from <http://bioinfadmin.cs.ucl.ac.uk/downloads/DISOPRED/> (last accessed 25/06/2015). For analysis of protein disorder, we considered disorder at both a per-proteome level (i.e. the number of residues across the proteome that fall within disordered regions; with no adjustment for protein length) and at a per-protein level (the percentage of residues within each protein predicted to be disordered). A similar approach was taken with all other predictors.

MHC binding prediction

NetMHC 3.0 [86,87] and NetMHCII 2.2 [43,44] were used for prediction of MHC class I and II binding peptides, respectively. Peptide lengths of 9 (NetMHC) and 15 (NetMHCII) residues were used for all predictions. Peptides were grouped according to their predicted binding affinity (IC_{50}): High-affinity, $IC_{50} < 50$ nM; Low-affinity, $50 \text{ nM} < IC_{50} < 500 \text{ nM}$; No-binding, $IC_{50} > 500 \text{ nM}$. Predictions were performed for all human HLA alleles available in each predictor. Both prediction algorithms were downloaded from <http://www.cbs.dtu.dk/services/NetMHCII/> and <http://www.cbs.dtu.dk/services/NetMHC/> (last accessed on 25/06/2015).

For prediction of MHC binding within scrambled sequences from *P. falciparum*, we obtained the predicted disorder scores for each protein, and scrambled sequences within ordered and disordered regions separately, retaining the overall disorder profile for each protein. These scrambled sequences were then submitted to predictors of MHC binding as above. This procedure was repeated four times with the scrambled proteome of *P. falciparum*, with results averaged between repeats.

B-cell epitope prediction

BepiPred 1.0 was used for prediction of linear B-cell epitopes [60]. An output threshold of 0.9 was used (sensitivity = 0.25, specificity = 0.91) for identification of B-cell epitopes (unless stated otherwise). This threshold was chosen to provide a high level of certainty for predicted B-cell epitopes. For comparison of linear B-cell epitopes across *Plasmodium* spp., and between ordered and disordered regions, we used a range of output thresholds, from -0.2 (sensitivity = 0.75, specificity = 0.5) to 1.3 (sensitivity = 0.13, specificity = 0.96). Any residue with an output score above the threshold was considered to be part of a linear B-cell epitope. BepiPred software was obtained from <http://www.cbs.dtu.dk/services/BepiPred/> (last accessed on 25/06/2015).

Protein localisation

Protein localisation data were obtained from the ApiLoc database (<http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc>) for *P. falciparum* sequences only (451 proteins) (last accessed on 25/06/2015). While ApiLoc also contains curated localisations for other *Plasmodium* spp., the number of proteins available for other species was too low to enable proper comparison among subcellular localisations (*P. berghei*, 61 proteins; *P. vivax*, 18 proteins; *P. knowlesi*, 6 proteins; *P. chabaudi*, 4 proteins).

Identification of tandem repeat sequences

Tandem repeat sequences were identified using T-REKS, a program for the detection of repeat sequences based on a K-means algorithm [45]. T-REKS software was obtained from <http://bioinfo.montp.cnrs.fr/?r=t-reks/> (last accessed 25/06/2015), and was run as a standalone Java program. The percentage similarity (P_{sim}) threshold was set to 0.8 for all predictions, with filtering of overlapping repeats disabled. For residues that were part of overlapping repeats, only the repeat with the highest P_{sim} value was considered.

Analysis of amino acid substitutions due to non-synonymous point mutations

Data for single nucleotide polymorphisms from *P. falciparum* were downloaded from PlasmoDB. Within PlasmoDB, SNPs were identified based on differences within a group of isolates, with 3D7 chosen as the reference strain. SNPs were selected from isolates obtained from all

available geographic locations. An 80% read frequency threshold was used, with a minimum minor allele frequency of 5%. To examine amino acid substitutions, only non-synonymous SNPs within coding regions of DNA were used.

Workflow and database integration

Protein sequences in FASTA format were submitted to predictors of protein disorder, MHC binding, B-cell epitopes, and tandem repeats. Output files were collated and reduced to a format amenable to storage in an SQL database. Results were stored in a local SQL database (PostgreSQL 9.3.6). Protein localisation data from ApiLoc and SNP data were also stored in the SQL database. Sequence input to various prediction algorithms, output data processing and SQL queries were handled using in-house custom Perl and Python scripts (S1 File). The computational workflow is depicted in Fig 1.

Sequence analysis of MHC binding peptides

As disordered and ordered regions tend to have a different amino acid composition, we accounted for the background amino acid frequency of ordered and disordered regions when examining the enrichment of particular residues within MHC-binding peptides. Peptides that fell on the boundary of disordered and ordered regions were considered to be part of a mixed region. To do this, we defined the proportional enrichment f'_{ijk} (Eq 1) for any particular residue i found at position j within all MHC binding peptides in a region k (disordered, ordered or mixed), such that residues that are neither enriched nor depleted were assigned a value of 0, while residues with a 100% increase/decrease in frequency were assigned a value of +/-1:

$$f'_{ijk} = \frac{n_{ijk}/\sum_i n_{ijk}}{N_{ijk}/\sum_i N_{ijk}} - 1 \quad (1)$$

where:

n_{ijk} = number of times residue i is found at position j within a predicted MHC binding peptide within region k .

N_{ijk} number of times residue i is found at position j within any peptide within region k .

$\sum_i n_{ijk}$ indicates a sum over all residues i at position j for residues found within an MHC binding peptide within region k .

$\sum_i N_{ijk}$ indicates a sum over all residues i at position j for residues found within region k .

To determine the average enrichment across the entire proteome for residues in MHC-binding peptides, we defined an average enrichment value f_{ij} (Eq 2) with correction for sequence composition in each region k as:

$$f_{ij} = \left(\sum_k \frac{n_{ijk}}{N_{ijk}/\sum_i N_{ijk}} \right) / \left(\sum_k \sum_i n_{ijk} \right) - 1 \quad (2)$$

This represents a weighted average of MHC binding residue enrichment across all regions k , with adjustment for amino acid frequency within each region k .

Statistical analysis

All plots and statistical analysis were produced with the R statistical computing package (R Project for Statistical Computing, <http://www.r-project.org/>), with RStudio IDE v. 0.97.551 used for code development. R package ggplot2 was utilised for most plots [88].

Supporting Information

S1 Fig. Predicted MHCI and MHCII binding for the *P. falciparum* proteome, grouped by protein disorder. The proportion of peptides with predicted binding to MHCI (A) and MHCII (B) is significantly higher for peptides that are contained within a structured protein domain. Prediction of protein disorder was performed using DISOPRED3, while predictions of MHC class I and MHC class II binding were performed with NetMHC 3.0 and NetMHCII 2.2. Peptides were grouped according to their predicted binding affinity (IC50): High-affinity, IC50<50nM; Low-affinity, 50nM<IC50<500nM; No-binding, IC50>500nM. (PDF)

S2 Fig. Distribution of high-affinity MHC-binding peptides from *P. falciparum* across a range of HLA alleles, grouped according to predicted protein disorder and known protein localisation. No significant difference in the proportion of MHCI (A) or MHCII (B) binding peptides was observed between different subcellular locations ($p > 0.05$, kruskal-wallis rank sum test). Boxplots represent the distribution of MHC-binding peptides across all MHC alleles tested. Prediction of protein disorder was performed using DISOPRED3, while prediction of MHC class I and MHC class II binding was performed with NetMHC 3.0 and NetMHCII 2.2. Peptides with predicted high binding affinity are shown (IC50<50nM). (PDF)

S3 Fig. Residues that are enriched within MHC binding peptides are generally found at lower frequency within disordered regions. The position specific enhancement of each residue in both MHC class I (A) and MHC class II (B) binding peptides (IC50 < 50nM) was plotted against the proportional enrichment of that residue in disordered regions. Prediction of protein disorder was performed using DISOPRED3, while prediction of MHC class I and MHC class II binding was performed with NetMHC 3.0 and NetMHCII 2.2. (PDF)

S4 Fig. Predicted MHC binding for scrambled sequences from *P. falciparum*, compared to predicted MHC binding of native sequence. Sequences within disordered and ordered regions of each *P. falciparum* protein were scrambled, and the resultant scrambled proteome was submitted to predictors of MHC class I (A) and MHC class II (B) binding. Sequence scrambling was performed 4x, with results from MHC predictors averaged across all repeats. Prediction of disorder was performed with DISOPRED3. (PDF)

S5 Fig. Distribution of linear B-cell epitopes within *P. falciparum* proteins, grouped according to subcellular localisation and predicted protein disorder. Classification of disorder was achieved using DISOPRED3. BepiPred was used for prediction of linear B-cell epitopes. A threshold of 0.9 was used for BepiPred predictions. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. (PDF)

S1 File. Computational scripts used to generate data, perform analysis and generate figures. (ZIP)

S1 Table. Summary statistics for predicted protein disorder of *P. falciparum* proteins, grouped according to subcellular localisation. Protein localisation was classified using the ApiLoc resource. Prediction of disorder was performed using DISOPRED3. A total of 451 proteins were assigned a location. Percentage disorder was calculated as the proportion of residues

predicted to be disordered at the level of individual proteins.
(DOCX)

S2 Table. Summary statistics for percentage linear B-cell epitopes within *P. falciparum* proteins, grouped according to subcellular localisation. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage of residues predicted to be part of a linear B-cell epitope for each protein in that location, to the distribution within the entire *P. falciparum* proteome. Residues were grouped according to predicted protein disorder, and statistical analysis applied to each group (ordered/disordered).
(DOCX)

S3 Table. Summary statistics for predicted tandem repeats within *P. falciparum* proteins, grouped according to subcellular localisation. Protein localisation was classified using the ApiLoc resource. Prediction of tandem repeats was performed using TREKS, with a PSIM cut-off of 0.8. A total of 451 proteins were assigned a location. Percentage tandem repeats was calculated as the proportion of residues predicted to be part of a tandem repeat at the level of individual proteins. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage tandem repeats for proteins within each respective location to the distribution of percentage tandem repeats within the entire *P. falciparum* proteome.
(DOCX)

S4 Table. Summary statistics for SNPs within *P. falciparum* proteins, grouped according to subcellular localisation. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage of residues targeted by non-synonymous SNPs for each protein in that location, to the distribution of SNPs within the entire *P. falciparum* proteome.
(DOCX)

Author Contributions

Conceived and designed the experiments: AJG JSR PAR. Performed the experiments: AJG. Analyzed the data: AJG. Wrote the paper: AJG JSR PAR. Provided critical analysis of the data and the manuscript: VI CAM RFA RSN JGB JSR PAR.

References

1. van der Lee R, Buljan M, Lang B, Weatheritt RJ, Daughdrill GW, Dunker AK, et al. (2014) Classification of intrinsically disordered regions and proteins. *Chem Rev* 114: 6589–6631. doi: [10.1021/cr400525m](https://doi.org/10.1021/cr400525m) PMID: [24773235](https://pubmed.ncbi.nlm.nih.gov/24773235/)
2. Asano N, Atsuumi H, Nakamura A, Tanaka Y, Tanaka I, Yao M (2014) Direct interaction between EFL1 and SBDS is mediated by an intrinsically disordered insertion domain. *Biochem Biophys Res Commun* 443: 1251–1256. doi: [10.1016/j.bbrc.2013.12.143](https://doi.org/10.1016/j.bbrc.2013.12.143) PMID: [24406167](https://pubmed.ncbi.nlm.nih.gov/24406167/)
3. Hisaoka M, Nagata K, Okuwaki M (2014) Intrinsically disordered regions of nucleophosmin/B23 regulate its RNA binding activity through their inter- and intra-molecular association. *Nucleic Acids Res* 42: 1180–1195. doi: [10.1093/nar/gkt897](https://doi.org/10.1093/nar/gkt897) PMID: [24106084](https://pubmed.ncbi.nlm.nih.gov/24106084/)
4. Ramos I, Fernandez-Rivero N, Arranz R, Aloria K, Finn R, Arizmendi JM, et al. (2014) The intrinsically disordered distal face of nucleoplasmin recognizes distinct oligomerization states of histones. *Nucleic Acids Res* 42: 1311–1325. doi: [10.1093/nar/gkt899](https://doi.org/10.1093/nar/gkt899) PMID: [24121686](https://pubmed.ncbi.nlm.nih.gov/24121686/)
5. van Leeuwen HC, Strating MJ, Rensen M, de Laat W, van der Vliet PC (1997) Linker length and composition influence the flexibility of Oct-1 DNA binding. *The EMBO journal* 16: 2043–2053. PMID: [9155030](https://pubmed.ncbi.nlm.nih.gov/9155030/)

6. Magidovich E, Fleishman SJ, Yifrach O (2006) Intrinsically disordered C-terminal segments of voltage-activated potassium channels: a possible fishing rod-like mechanism for channel binding to scaffold proteins. *Bioinformatics* 22: 1546–1550. PMID: [16601002](#)
7. Dunker AK, Brown CJ, Lawson JD, Iakoucheva LM, Obradovic Z (2002) Intrinsic disorder and protein function. *Biochemistry* 41: 6573–6582. PMID: [12022860](#)
8. Zandany N, Marciano S, Magidovich E, Frimerman T, Yehezkel R, Shem-Ad T, et al. (2015) Alternative splicing modulates Kv channel clustering through a molecular ball and chain mechanism. *Nat Commun* 6: 6488. doi: [10.1038/ncomms7488](#) PMID: [25813388](#)
9. Zandany N, Lewin L, Nirenberg V, Orr I, Yifrach O (2015) Entropic clocks in the service of electrical signaling: 'Ball and chain' mechanisms for ion channel inactivation and clustering. *FEBS Lett* doi: [10.1016/j.febslet.2015.06.010](#)
10. Xue B, Dunker AK, Uversky VN (2012) Orderly order in protein intrinsic disorder distribution: disorder in 3500 proteomes from viruses and the three domains of life. *J Biomol Struct Dyn* 30: 137–149. doi: [10.1080/07391102.2012.675145](#) PMID: [22702725](#)
11. Ward JJ, Sodhi JS, McGuffin LJ, Buxton BF, Jones DT (2004) Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol* 337: 635–645. PMID: [15019783](#)
12. Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P, Oh JS, et al. (2001) Intrinsically disordered protein. *J Mol Graph Model* 19: 26–59. PMID: [11381529](#)
13. Feng ZP, Zhang X, Han P, Arora N, Anders RF, Norton RS (2006) Abundance of intrinsically unstructured proteins in *P. falciparum* and other apicomplexan parasite proteomes. *Mol Biochem Parasitol* 150: 256–267. PMID: [17010454](#)
14. WHO (2014) World Malaria Report. Geneva, Switzerland.
15. McCarthy JS, Marjason J, Elliott S, Fahey P, Bang G, Malkin E, et al. (2011) A phase 1 trial of MSP2-C1, a blood-stage malaria vaccine containing 2 isoforms of MSP2 formulated with Montanide(R) ISA 720. *PLoS One* 6: e24413. doi: [10.1371/journal.pone.0024413](#) PMID: [21949716](#)
16. Singh S, Soe S, Mejia J- P, Roussilhon C, Theisen M, Corradin M, et al. (2004) Identification of a conserved region of *Plasmodium falciparum* MSP3 targeted by biologically active antibodies to improve vaccine design. *J Infect Dis* 190: 1010–1018. PMID: [15295710](#)
17. Oeuvray C, Bouharoun-Tayoun H, Grass-Masse H, Lepers JP, Ralamboranto L, Tartar A, et al. (1994) A novel merozoite surface antigen of *Plasmodium falciparum* (MSP-3) identified by cellular-antibody cooperative mechanism antigenicity and biological activity of antibodies. *Mem Inst Oswaldo Cruz* 89 Suppl 2: 77–80. PMID: [7565137](#)
18. Healer J, Thompson JK, Riglar DT, Wilson DW, Chiu YH, Miura K, et al. (2013) Vaccination with Conserved Regions of Erythrocyte-Binding Antigens Induces Neutralizing Antibodies against Multiple Strains of *Plasmodium falciparum*. *PLoS ONE* 8: e72504. doi: [10.1371/journal.pone.0072504](#) PMID: [24039774](#)
19. Yagi M, Bang G, Tougan T, Palacpac NM, Arisue N, Aoshi T, et al. (2014) Protective epitopes of the *Plasmodium falciparum* SERA5 malaria vaccine reside in intrinsically unstructured N-terminal repetitive sequences. *PLoS One* 9: e98460. doi: [10.1371/journal.pone.0098460](#) PMID: [24886718](#)
20. Foquet L, Hermesen CC, van Gemert GJ, Van Braeckel E, Weening KE, Sauerwein R, et al. (2014) Vaccine-induced monoclonal antibodies targeting circumsporozoite protein prevent *Plasmodium falciparum* infection. *J Clin Invest* 124: 140–144. PMID: [24292709](#)
21. Burgess BR, Schuck P, Garboczi DN (2005) Dissection of merozoite surface protein 3, a representative of a family of *Plasmodium falciparum* surface proteins, reveals an oligomeric and highly elongated molecule. *J Biol Chem* 280: 37236–37245. PMID: [16135515](#)
22. Tsai CW, Duggan PF, Jin AJ, Macdonald NJ, Kotova S, Lebowitz J, et al. (2009) Characterization of a protective *Escherichia coli*-expressed *Plasmodium falciparum* merozoite surface protein 3 indicates a non-linear, multi-domain structure. *Mol Biochem Parasitol* 164: 45–56. doi: [10.1016/j.molbiopara.2008.11.006](#) PMID: [19073223](#)
23. Kulangara C, Luedin S, Dietz O, Rusch S, Frank G, Mueller D, et al. (2012) Cell biological characterization of the malaria vaccine candidate trophozoite exported protein 1. *PLoS ONE* 7: e46112. doi: [10.1371/journal.pone.0046112](#) PMID: [23056243](#)
24. Adda CG, MacRaild CA, Reiling L, Wycherley K, Boyle MJ, Kienzle V, et al. (2012) Antigenic characterization of an intrinsically unstructured protein, *Plasmodium falciparum* merozoite surface protein 2. *Infect Immun* 80: 4177–4185. doi: [10.1128/IAI.00665-12](#) PMID: [22966050](#)
25. Adda CG, Murphy VJ, Sunde M, Waddington LJ, Schloegel J, Talbo GH, et al. (2009) *Plasmodium falciparum* merozoite surface protein 2 is unstructured and forms amyloid-like fibrils. *Mol Biochem Parasitol* 166: 159–171. doi: [10.1016/j.molbiopara.2009.03.012](#) PMID: [19450733](#)

26. Olugbile S, Kulangara C, Bang G, Bertholet S, Suzarte E, Villard V, et al. (2009) Vaccine potentials of an intrinsically unstructured fragment derived from the blood stage-associated *Plasmodium falciparum* protein PFF0165c. *Infect Immun* 77: 5701–5709. doi: [10.1128/IAI.00652-09](https://doi.org/10.1128/IAI.00652-09) PMID: [19786562](https://pubmed.ncbi.nlm.nih.gov/19786562/)
27. Zhang X, Perugini MA, Yao S, Adda CG, Murphy VJ, Low A, et al. (2008) Solution conformation, backbone dynamics and lipid interactions of the intrinsically unstructured malaria surface protein MSP2. *J Mol Biol* 379: 105–121. doi: [10.1016/j.jmb.2008.03.039](https://doi.org/10.1016/j.jmb.2008.03.039) PMID: [18440022](https://pubmed.ncbi.nlm.nih.gov/18440022/)
28. Radivojac P, Iakoucheva LM, Oldfield CJ, Obradovic Z, Uversky VN, Dunker AK (2007) Intrinsic disorder and functional proteomics. *Biophys J* 92: 1439–1456. PMID: [17158572](https://pubmed.ncbi.nlm.nih.gov/17158572/)
29. Mitic NS, Pavlovic MD, Jandric DR (2014) Epitope distribution in ordered and disordered protein regions—part A. T-cell epitope frequency, affinity and hydrophathy. *J Immunol Methods* 406: 83–103. doi: [10.1016/j.jim.2014.02.012](https://doi.org/10.1016/j.jim.2014.02.012) PMID: [24614036](https://pubmed.ncbi.nlm.nih.gov/24614036/)
30. Jorda J, Xue B, Uversky VN, Kajava AV (2010) Protein tandem repeats—the more perfect, the less structured. *The FEBS journal* 277: 2673–2682. doi: [10.1111/j.1742-464X.2010.07684.x](https://doi.org/10.1111/j.1742-464X.2010.07684.x) PMID: [20553501](https://pubmed.ncbi.nlm.nih.gov/20553501/)
31. Tompa P (2003) Intrinsically unstructured proteins evolve by repeat expansion. *Bioessays* 25: 847–855. PMID: [12938174](https://pubmed.ncbi.nlm.nih.gov/12938174/)
32. Goto Y, Carter D, Reed SG (2008) Immunological dominance of *Trypanosoma cruzi* tandem repeat proteins. *Infect Immun* 76: 3967–3974. doi: [10.1128/IAI.00604-08](https://doi.org/10.1128/IAI.00604-08) PMID: [18625739](https://pubmed.ncbi.nlm.nih.gov/18625739/)
33. Goto Y, Coler RN, Guderian J, Mohamath R, Reed SG (2006) Cloning, characterization, and serodiagnostic evaluation of *Leishmania infantum* tandem repeat proteins. *Infect Immun* 74: 3939–3945. PMID: [16790767](https://pubmed.ncbi.nlm.nih.gov/16790767/)
34. Luo T, Zhang X, McBride JW (2009) Major species-specific antibody epitopes of the *Ehrlichia chaffeensis* p120 and *E. canis* p140 orthologs in surface-exposed tandem repeat regions. *Clin Vaccine Immunol: CVI* 16: 982–990. doi: [10.1128/CVI.00048-09](https://doi.org/10.1128/CVI.00048-09) PMID: [19420187](https://pubmed.ncbi.nlm.nih.gov/19420187/)
35. Kumar N, Zheng H (1998) Evidence for epitope-specific thymus-independent response against a repeat sequence in a protein antigen. *Immunology* 94: 28–34. PMID: [9708183](https://pubmed.ncbi.nlm.nih.gov/9708183/)
36. Goto Y, Carter D, Guderian J, Inoue N, Kawazu S, Reed SG (2010) Upregulated expression of B-cell antigen family tandem repeat proteins by *Leishmania amastigotes*. *Infect Immun* 78: 2138–2145. doi: [10.1128/IAI.01102-09](https://doi.org/10.1128/IAI.01102-09) PMID: [20160013](https://pubmed.ncbi.nlm.nih.gov/20160013/)
37. Schofield L (1990) The circumsporozoite protein of *Plasmodium*: a mechanism of immune evasion by the malaria parasite? *Bulletin of the World Health Organization* 68 Suppl: 66–73. PMID: [1709835](https://pubmed.ncbi.nlm.nih.gov/1709835/)
38. Ochola LI, Tetteh KK, Stewart LB, Riitho V, Marsh K, Conway DJ (2010) Allele frequency-based and polymorphism-versus-divergence indices of balancing selection in a new filtered set of polymorphic genes in *Plasmodium falciparum*. *Mol Biol Evol* 27: 2344–2351. doi: [10.1093/molbev/msq119](https://doi.org/10.1093/molbev/msq119) PMID: [20457586](https://pubmed.ncbi.nlm.nih.gov/20457586/)
39. Maier AG, Baum J, Smith B, Conway DJ, Cowman AF (2009) Polymorphisms in erythrocyte binding antigens 140 and 181 affect function and binding but not receptor specificity in *Plasmodium falciparum*. *Infect Immun* 77: 1689–1699. doi: [10.1128/IAI.01331-08](https://doi.org/10.1128/IAI.01331-08) PMID: [19204093](https://pubmed.ncbi.nlm.nih.gov/19204093/)
40. Healer J, Murphy V, Hodder AN, Masciantonio R, Gemmill AW, Anders RF, et al. (2004) Allelic polymorphisms in apical membrane antigen-1 are responsible for evasion of antibody-mediated inhibition in *Plasmodium falciparum*. *Mol Microbiol* 52: 159–168. PMID: [15049818](https://pubmed.ncbi.nlm.nih.gov/15049818/)
41. Nilsson J, Grahn M, Wright AP (2011) Proteome-wide evidence for enhanced positive Darwinian selection within intrinsically disordered regions in proteins. *Genome Biol* 12: R65. doi: [10.1186/gb-2011-12-7-r65](https://doi.org/10.1186/gb-2011-12-7-r65) PMID: [21771306](https://pubmed.ncbi.nlm.nih.gov/21771306/)
42. ApiLoc—A database of published protein sub-cellular localization in Apicomplexa.
43. Nielsen M, Lund O (2009) NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC bioinformatics* 10: 296. doi: [10.1186/1471-2105-10-296](https://doi.org/10.1186/1471-2105-10-296) PMID: [19765293](https://pubmed.ncbi.nlm.nih.gov/19765293/)
44. Nielsen M, Lundegaard C, Lund O (2007) Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC bioinformatics* 8: 238. PMID: [17608956](https://pubmed.ncbi.nlm.nih.gov/17608956/)
45. Jorda J, Kajava AV (2009) T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics* 25: 2632–2638. doi: [10.1093/bioinformatics/btp482](https://doi.org/10.1093/bioinformatics/btp482) PMID: [19671691](https://pubmed.ncbi.nlm.nih.gov/19671691/)
46. Receveur-Brechot V, Bourhis JM, Uversky VN, Canard B, Longhi S (2006) Assessing protein disorder and induced folding. *Proteins* 62: 24–45. PMID: [16287116](https://pubmed.ncbi.nlm.nih.gov/16287116/)
47. Tham WH, Wilson DW, Reiling L, Chen L, Beeson JG, Cowman AF (2009) Antibodies to reticulocyte binding protein-like homologue 4 inhibit invasion of *Plasmodium falciparum* into human erythrocytes. *Infect Immun* 77: 2427–2435. doi: [10.1128/IAI.00048-09](https://doi.org/10.1128/IAI.00048-09) PMID: [19307208](https://pubmed.ncbi.nlm.nih.gov/19307208/)

48. Blanc M, Coetzer TL, Blackledge M, Haertlein M, Mitchell EP, Forsyth VT, et al. (2014) Intrinsic disorder within the erythrocyte binding-like proteins from *Plasmodium falciparum*. *Biochim Biophys Acta* doi: [10.1016/j.bbapap.2014.09.023](https://doi.org/10.1016/j.bbapap.2014.09.023)
49. Tolia NH, Enemark EJ, Sim BKL, Joshua-Tor L (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. [Erratum appears in *Cell*. 2005 Aug 12;122(3):485]. *Cell* 122: 183–193. PMID: [16051144](https://pubmed.ncbi.nlm.nih.gov/16051144/)
50. Sim BKL, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH (1994) Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science* 264: 1941–1944. PMID: [8009226](https://pubmed.ncbi.nlm.nih.gov/8009226/)
51. Lopaticki S, Maier AG, Thompson J, Wilson DW, Tham W- H, Triglia T, et al. (2011) Reticulocyte and erythrocyte binding-like proteins function cooperatively in invasion of human erythrocytes by malaria parasites. *Infect Immun* 79: 1107–1117. doi: [10.1128/IAI.01021-10](https://doi.org/10.1128/IAI.01021-10) PMID: [21149582](https://pubmed.ncbi.nlm.nih.gov/21149582/)
52. Hollingdale MR, Nardin EH, Tharavanij S, Schwartz AL, Nussenzweig RS (1984) Inhibition of entry of *Plasmodium falciparum* and *P. vivax* sporozoites into cultured cells; an in vitro assay of protective antibodies. *J Immunol* 132: 909–913. PMID: [6317752](https://pubmed.ncbi.nlm.nih.gov/6317752/)
53. Boyle MJ, Reiling L, Feng G, Langer C, Osier FH, Aspelting-Jones H, et al. (2015) Human antibodies fix complement to inhibit *Plasmodium falciparum* invasion of erythrocytes and are associated with protection against malaria. *Immunity* 42: 580–590. doi: [10.1016/j.immuni.2015.02.012](https://doi.org/10.1016/j.immuni.2015.02.012) PMID: [25786180](https://pubmed.ncbi.nlm.nih.gov/25786180/)
54. Rammensee HG (1995) Chemistry of peptides associated with MHC class I and class II molecules. *Curr Opin Immunol* 7: 85–96. PMID: [7772286](https://pubmed.ncbi.nlm.nih.gov/7772286/)
55. Nelson CA, Fremont DH (1999) Structural principles of MHC class II antigen presentation. *Rev Immunogenet* 1: 47–59. PMID: [11256572](https://pubmed.ncbi.nlm.nih.gov/11256572/)
56. Natarajan K, Li H, Mariuzza RA, Margulies DH (1999) MHC class I molecules, structure and function. *Rev Immunogenet* 1: 32–46. PMID: [11256571](https://pubmed.ncbi.nlm.nih.gov/11256571/)
57. Matsumura M, Fremont DH, Peterson PA, Wilson IA (1992) Emerging principles for the recognition of peptide antigens by MHC class I molecules. *Science* 257: 927–934. PMID: [1323878](https://pubmed.ncbi.nlm.nih.gov/1323878/)
58. Weiss WR, Sedegah M, Beaudoin RL, Miller LH, Good MF (1988) CD8+ T cells (cytotoxic/suppressors) are required for protection in mice immunized with malaria sporozoites. *Proc Natl Acad Sci U S A* 85: 573–576. PMID: [2963334](https://pubmed.ncbi.nlm.nih.gov/2963334/)
59. Schofield L, Villaquiran J, Ferreira A, Schellekens H, Nussenzweig R, Nussenzweig V (1987) Gamma interferon, CD8+ T cells and antibodies required for immunity to malaria sporozoites. *Nature* 330: 664–666. PMID: [3120015](https://pubmed.ncbi.nlm.nih.gov/3120015/)
60. Larsen JE, Lund O, Nielsen M (2006) Improved method for predicting linear B-cell epitopes. *Immunome Res* 2: 2. PMID: [16635264](https://pubmed.ncbi.nlm.nih.gov/16635264/)
61. De Genst EJ, Guillems T, Wellens J, O'Day EM, Waudby CA, Meehan S, et al. (2010) Structure and properties of a complex of alpha-synuclein and a single-domain camelid antibody. *J Mol Biol* 402: 326–343. doi: [10.1016/j.jmb.2010.07.001](https://doi.org/10.1016/j.jmb.2010.07.001) PMID: [20620148](https://pubmed.ncbi.nlm.nih.gov/20620148/)
62. Serriere J, Dugua JM, Bossus M, Verrier B, Haser R, Gouet P, et al. (2011) Fab'-induced folding of antigenic N-terminal peptides from intrinsically disordered HIV-1 Tat revealed by X-ray crystallography. *J Mol Biol* 405: 33–42. doi: [10.1016/j.jmb.2010.10.033](https://doi.org/10.1016/j.jmb.2010.10.033) PMID: [21035463](https://pubmed.ncbi.nlm.nih.gov/21035463/)
63. Fassolari M, Chemes LB, Gallo M, Smal C, Sanchez IE, de Prat-Gay G (2013) Minute time scale prolyl isomerization governs antibody recognition of an intrinsically disordered immunodominant epitope. *J Biol Chem* 288: 13110–13123. doi: [10.1074/jbc.M112.444554](https://doi.org/10.1074/jbc.M112.444554) PMID: [23504368](https://pubmed.ncbi.nlm.nih.gov/23504368/)
64. Chu HM, Wright J, Chan YH, Lin CJ, Chang TW, Lim C (2014) Two potential therapeutic antibodies bind to a peptide segment of membrane-bound IgE in different conformations. *Nat Commun* 5: 3139. doi: [10.1038/ncomms4139](https://doi.org/10.1038/ncomms4139) PMID: [24457896](https://pubmed.ncbi.nlm.nih.gov/24457896/)
65. Saad B, Corradin G, Bosshard HR (1988) Monoclonal antibody recognizes a conformational epitope in a random coil protein. *Eur J Biochem* 178: 219–224. PMID: [2462497](https://pubmed.ncbi.nlm.nih.gov/2462497/)
66. El-Manzalawy Y, Honavar V (2010) Recent advances in B-cell epitope prediction methods. *Immunome Res* 6.
67. Kringelum JV, Lundegaard C, Lund O, Nielsen M (2012) Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol* 8: e1002829. doi: [10.1371/journal.pcbi.1002829](https://doi.org/10.1371/journal.pcbi.1002829) PMID: [23300419](https://pubmed.ncbi.nlm.nih.gov/23300419/)
68. MacRaid CA, Zachrdla M, Andrew D, Krishnarajuna B, Novacek J, Zidek L (2015) Conformational dynamics and antigenicity in the disordered malaria antigen merozoite surface protein 2. *PLoS One* 10: e0119899. doi: [10.1371/journal.pone.0119899](https://doi.org/10.1371/journal.pone.0119899) PMID: [25742002](https://pubmed.ncbi.nlm.nih.gov/25742002/)
69. Kemp DJ, Coppel RL, Anders RF (1987) Repetitive proteins and genes of malaria. *Annu Rev Microbiol* 41: 181–208. PMID: [3318667](https://pubmed.ncbi.nlm.nih.gov/3318667/)

70. Hisaeda H, Yasutomo K, Himeno K (2005) Malaria: immune evasion by parasites. *Int J Biochem Cell Biol* 37: 700–706. PMID: [15694829](#)
71. Goto Y, Coler RN, Reed SG (2007) Bioinformatic identification of tandem repeat antigens of the *Leishmania donovani* complex. *Infect Immun* 75: 846–851. PMID: [17088350](#)
72. Zavala F, Cochrane AH, Nardin EH, Nussenzweig RS, Nussenzweig V (1983) Circumsporozoite proteins of malaria parasites contain a single immunodominant region with two or more identical epitopes. *J Exp Med* 157: 1947–1957. PMID: [6189951](#)
73. Dame JB, Williams JL, McCutchan TF, Weber JL, Wirtz RA, Hockmeyer WT, et al. (1984) Structure of the gene encoding the immunodominant surface antigen on the sporozoite of the human malaria parasite *Plasmodium falciparum*. *Science* 225: 593–599. PMID: [6204383](#)
74. Enea V, Ellis J, Zavala F, Arnot DE, Asavanich A, Masuda A, et al. (1984) DNA cloning of *Plasmodium falciparum* circumsporozoite gene: amino acid sequence of repetitive epitope. *Science* 225: 628–630. PMID: [6204384](#)
75. Plassmeyer ML, Reiter K, Shimp RL Jr, Kotova S, Smith PD, Hurt DE, et al. (2009) Structure of the *Plasmodium falciparum* circumsporozoite protein, a leading malaria vaccine candidate. *J Biol Chem* 284: 26951–26963. doi: [10.1074/jbc.M109.013706](#) PMID: [19633296](#)
76. Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM (2015) Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions. *Genome Biol Evol* doi: [10.1093/gbe/evv105](#)
77. Bai T, Becker M, Gupta A, Strike P, Murphy VJ, Anders RF, et al. (2005) Structure of AMA1 from *Plasmodium falciparum* reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc Natl Acad Sci U S A* 102: 12736–12741. PMID: [16129835](#)
78. Mobegi VA, Duffy CW, Amambua-Ngwa A, Loua KM, Laman E, Nwakanma DC, et al. (2014) Genome-wide analysis of selection on the malaria parasite *Plasmodium falciparum* in West African populations of differing infection endemicity. *Mol Biol Evol* 31: 1490–1499. doi: [10.1093/molbev/msu106](#) PMID: [24644299](#)
79. Haerty W, Golding GB (2011) Increased polymorphism near low-complexity sequences across the genomes of *Plasmodium falciparum* isolates. *Genome Biol Evol* 3: 539–550. doi: [10.1093/gbe/evr045](#) PMID: [21602572](#)
80. Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. (2008) Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature* 455: 757–763. doi: [10.1038/nature07327](#) PMID: [18843361](#)
81. Otto TD, Bohme U, Jackson AP, Hunt M, Franke-Fayard B, Hoeijmakers WA, et al. (2014) A comprehensive evaluation of rodent malaria parasite genomes and gene expression. *BMC Biol* 12: 86. doi: [10.1186/s12915-014-0086-0](#) PMID: [25359557](#)
82. Pain A, Bohme U, Berry AE, Mungall K, Finn RD, Jackson AP, et al. (2008) The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* 455: 799–803. doi: [10.1038/nature07306](#) PMID: [18843368](#)
83. Aurecochea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. (2009) PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res* 37: D539–543. doi: [10.1093/nar/gkn814](#) PMID: [18957442](#)
84. Jones DT, Cozzetto D (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 31: 857–863. doi: [10.1093/bioinformatics/btu744](#) PMID: [25391399](#)
85. Monastyrskyy B, Kryshchovych A, Moul J, Tramontano A, Fidelis K (2013) Assessment of protein disorder region predictions in CASP10. *Proteins* doi: [10.1002/prot.24391](#)
86. Buus S, Lauemoller SL, Wornig P, Kesmir C, Frimurer T, Corbet S, et al. (2003) Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach. *Tissue Antigens* 62: 378–384. PMID: [14617044](#)
87. Nielsen M, Lundegaard C, Wornig P, Lauemoller SL, Lamberth K, Buus S, et al. (2003) Reliable prediction of T-cell epitopes using neural networks with novel sequence representations. *Protein Sci* 12: 1007–1017. PMID: [12717023](#)
88. Wickham H (2009) ggplot2: elegant graphics for data analysis: Springer New York.

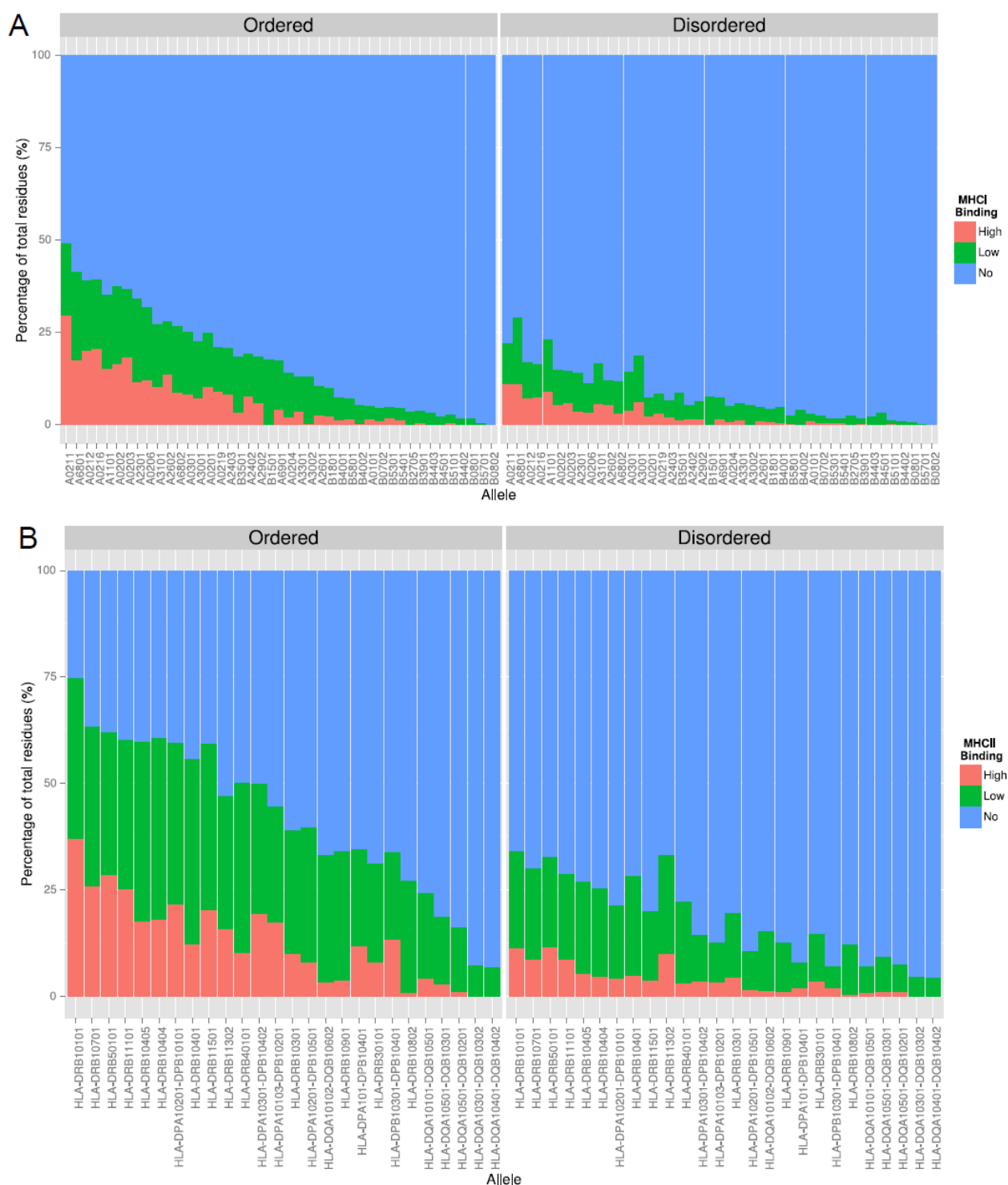


Figure S1: Predicted MHC I and MHC II binding for the *P. falciparum* proteome, grouped by protein disorder. The proportion of peptides with predicted binding to MHC I (A) and MHC II (B) is significantly higher for peptides that are contained within a structured protein domain. Prediction of protein disorder was performed using DISOPRED3, while predictions of MHC class I and MHC class II binding were performed with NetMHC 3.0 and NetMHCII 2.2. Peptides were grouped according to their predicted binding affinity (IC_{50}): High-affinity, $IC_{50} < 50nM$; Low-affinity, $50nM < IC_{50} < 500nM$; No-binding, $IC_{50} > 500nM$.

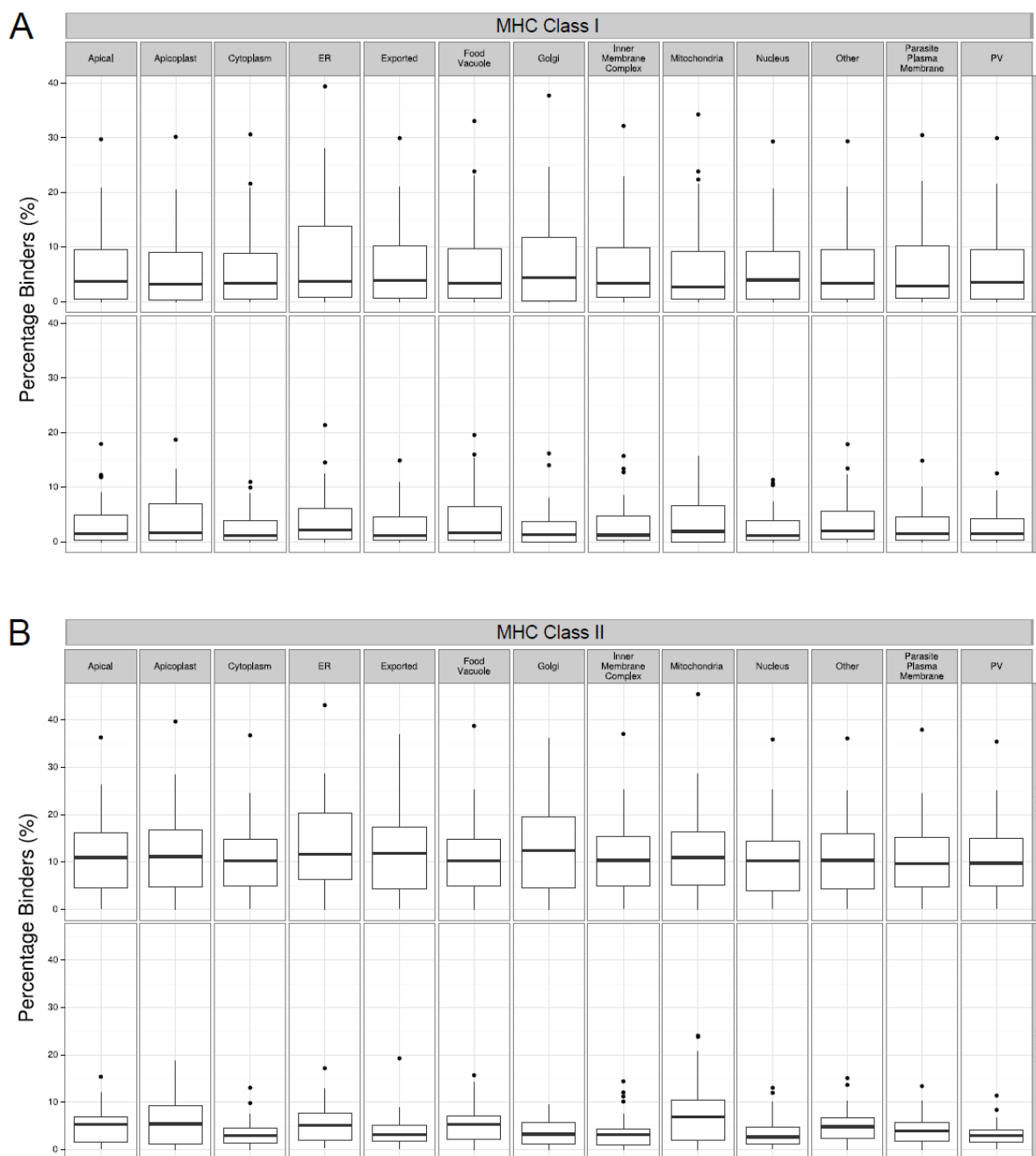


Figure S2: Distribution of high-affinity MHC-binding peptides from *P. falciparum* across a range of HLA alleles, grouped according to predicted protein disorder and known protein localisation. No significant difference in the proportion of MHCI (A) or MHCII (B) binding peptides was observed between different subcellular locations ($p > 0.05$, kruskal-wallis rank sum test). Boxplots represent the distribution of MHC-binding peptides across all MHC alleles tested. Prediction of protein disorder was performed using DISOPRED3, while prediction of MHC class I and MHC class II binding was performed with NetMHC 3.0 and NetMHCII 2.2. Peptides with predicted high binding affinity are shown ($IC_{50} < 50nM$)

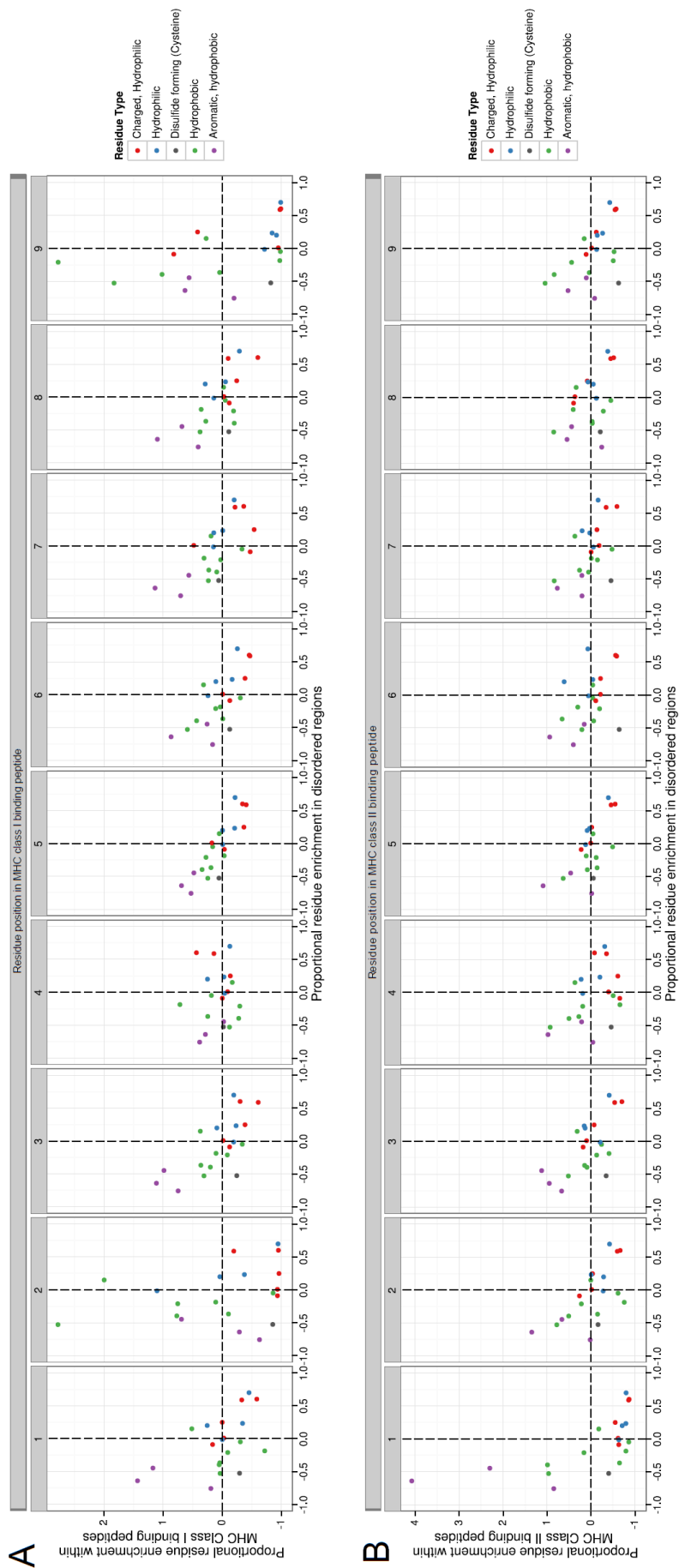


Figure S3: Residues that are enriched within MHC binding peptides are generally found at lower frequency within disordered regions. The position specific enhancement of each residue in both MHC class I (A) and MHC class II (B) binding peptides (IC50 < 50nM) was plotted against the proportional enrichment of that residue in disordered regions. Prediction of protein disorder was performed using DISOPRED3, while prediction of MHC class I and MHC class II binding was performed with NetMHC 3.0 and NetMHCII 2.2..

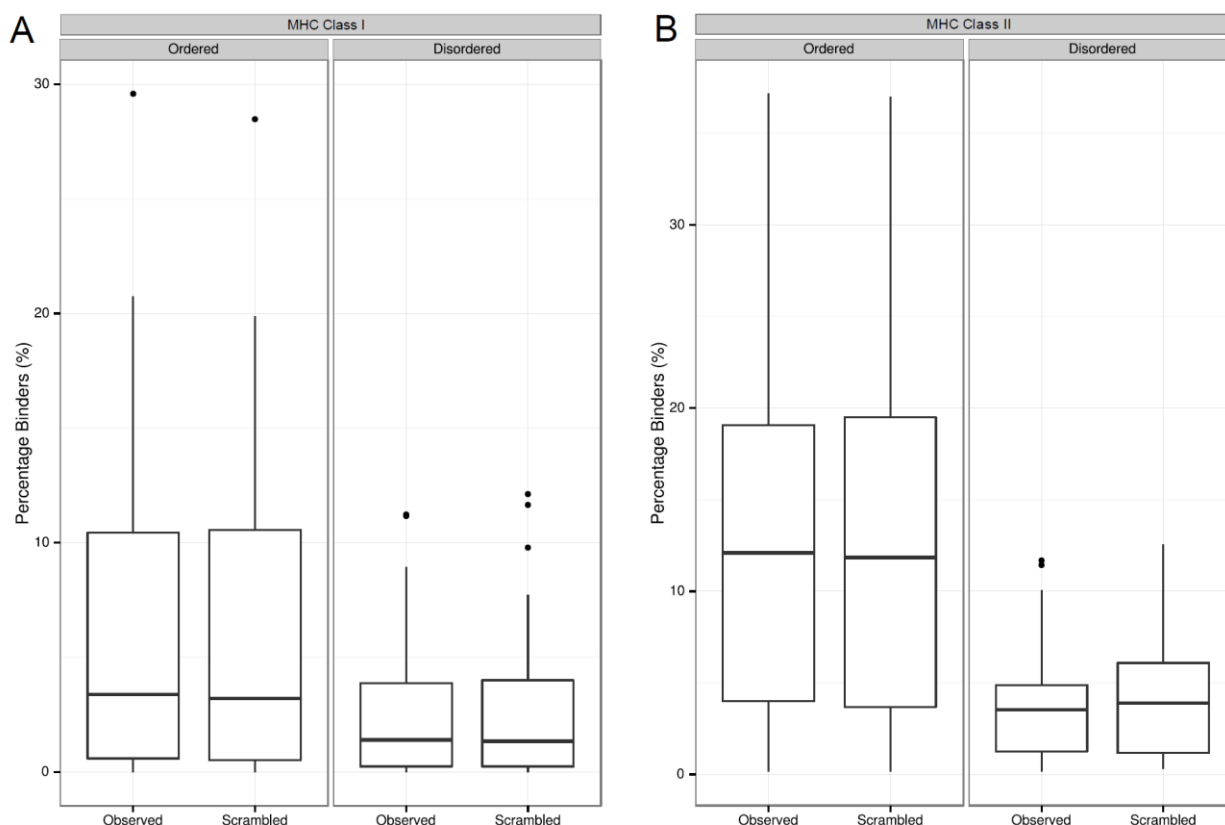


Figure S4: Predicted MHC binding for scrambled sequences from *P. falciparum*, compared to predicted MHC binding of native sequence. Sequences within disordered and ordered regions of each *P. falciparum* protein were scrambled, and the resultant scrambled proteome was submitted to predictors of MHC class I (A) and MHC class II (B) binding. Sequence scrambling was performed 4x, with results from MHC predictors averaged across all repeats. Prediction of disorder was performed with DISOPRED3.

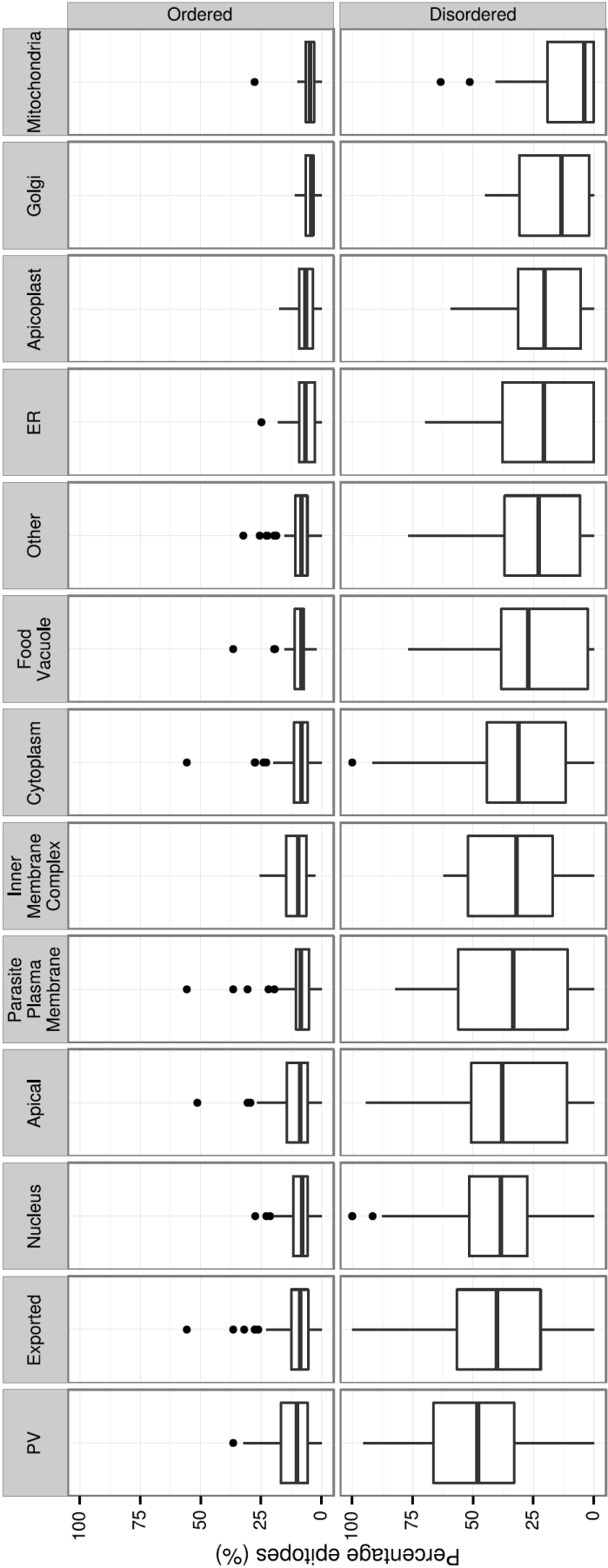


Figure S5: Distribution of linear B-cell epitopes within *P. falciparum* proteins, grouped according to subcellular localisation and predicted protein disorder. Classification of disorder was achieved using DISOPRED3. BepiPred was used for prediction of linear B-cell epitopes. A threshold of 0.9 was used for BepiPred predictions. Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location.

Table S1: Summary statistics for predicted protein disorder of *P. falciparum* proteins, grouped according to subcellular localisation.

Location	Median	IQR	W statistic	df	p-value
Apical	24.3	38.9	273081.5	82	0.0003
PV	27.7	29.1	172672	47	<0.0001
Exported	28.0	34.1	285521	80	<0.0001
Nucleus	28.1	28.7	248017	73	0.0002
Inner Membrane Complex	18.1	25.1	52308	17	0.51
Golgi	14.1	33.3	25671	9	0.84
Parasite Plasma Membrane	22.2	37.6	187876.5	62	0.11
Cytoplasm	15.4	26.3	266415.5	102	0.59
Other	11.4	24.8	146830.5	60	0.19
Apicoplast	15.7	11.0	99806.5	39	0.48
Food Vacuole	12.0	16.8	69140.5	27	0.50
Mitochondria	8.8	9.2	55932.5	28	0.01
ER	8.9	14.0	67192.5	31	0.04

Protein localisation was classified using the ApiLoc resource. Prediction of disorder was performed using DISOPRED3. A total of 451 proteins were assigned a location. Percentage disorder was calculated as the proportion of residues predicted to be disordered at the level of individual proteins.

Table S2: Summary statistics for percentage linear B-cell epitopes within *P. falciparum* proteins, grouped according to subcellular localisation.

Location	Disorder	Median	IQR	W statistic	df	p-value
PV	Ordered	10.2	11.2	163005	47	0.002
PV	Disordered	48.0	33.3	179525	47	<0.0001
Exported	Ordered	8.7	7.0	260513	81	0.005
Exported	Disordered	40.0	34.6	261212.5	80	0.001
Apical	Ordered	8.8	8.8	274922.5	83	0.001
Apical	Disordered	38.0	39.7	249020	82	0.054
Parasite Plasma Membrane	Ordered	8.4	5.6	193397	62	0.058
Parasite Plasma Membrane	Disordered	33.3	45.3	183593	62	0.215
Nucleus	Ordered	8.0	5.9	235542	74	0.014
Nucleus	Disordered	38.4	24.1	243684.5	73	0.001
Inner Membrane Complex	Ordered	9.5	8.4	63922	17	0.020
Inner Membrane Complex	Disordered	31.9	34.9	50334.5	17	0.733
Food Vacuole	Ordered	8.3	3.7	92857	27	0.035
Food Vacuole	Disordered	27.2	35.8	67155.5	27	0.347
Cytoplasm	Ordered	8.2	5.7	325156.5	106	0.024
Cytoplasm	Disordered	31.3	32.5	267458.5	102	0.620
Other	Ordered	8.3	5.1	195127	61	0.023
Other	Disordered	22.7	31.3	134864	60	0.020
Golgi	Ordered	4.2	3.2	19591.5	10	0.052
Golgi	Disordered	13.3	28.9	17258	9	0.052
ER	Ordered	6.5	6.6	77775	31	0.337
ER	Disordered	20.5	37.8	64689	31	0.017
Apicoplast	Ordered	6.5	5.7	97777	39	0.309
Apicoplast	Disordered	20.3	25.9	78019.5	39	0.003
Mitochondrion	Ordered	4.7	3.5	54307.5	28	0.005
Mitochondrion	Disordered	3.8	19.2	40344	28	<0.0001

Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage of residues predicted to be part of a linear B-cell epitope for each protein in that location, to the distribution within the entire *P. falciparum* proteome. Residues were grouped according to predicted protein disorder, and statistical analysis applied to each group (ordered/disordered).

Table S3: Summary statistics for predicted tandem repeats within *P. falciparum* proteins, grouped according to subcellular localisation.

Location	Mean	SD	Median	IQR	W statistic	df	p-value
Exported	9.61	18.58	0.61	10.51	257829	81	0.004
PV	8.54	17.69	1.67	9.89	152788.5	47	0.016
Parasite Plasma Membrane	5.62	12.48	0.00	6.63	177916	62	0.479
Apical	5.43	11.97	0.00	5.62	244030	83	0.178
Cytoplasm	3.79	10.92	0.00	2.79	266778.5	106	0.125
Nucleus	3.38	6.26	0.00	4.86	212537	74	0.406
Other	2.70	9.48	0.00	0.00	140416.5	61	0.014
Golgi	2.14	7.11	0.00	0.00	21433.5	10	0.072
Food Vacuole	1.98	4.46	0.00	0.89	67304	27	0.258
ER	1.84	4.75	0.00	0.42	72449	31	0.075
Inner Membrane Complex	1.41	2.84	0.00	1.29	41423.5	17	0.222
Apicoplast	0.61	1.70	0.00	0.00	81310	39	0.002
Mitochondrion	0.39	1.78	0.00	0.00	51823	28	0.0004

Protein localisation was classified using the ApiLoc resource. Prediction of tandem repeats was performed using TREKS, with a PSIM cutoff of 0.8. A total of 451 proteins were assigned a location. Percentage tandem repeats was calculated as the proportion of residues predicted to be part of a tandem repeat at the level of individual proteins. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage tandem repeats for proteins within each respective location to the distribution of percentage tandem repeats within the entire *P. falciparum* proteome.

Table S4: Summary statistics for SNPs within *P. falciparum* proteins, grouped according to subcellular localisation.

Location	Mean	SD	Median	IQR	W statistic	df	p-value
Exported	1.67	3.71	0.46	1.17	268938.5	81	0.001
PV	1.47	3.14	0.60	1.08	164441.5	47	0.001
Apical	1.46	2.65	0.55	1.27	290436.5	83	<0.0001
Parasite Plasma Membrane	1.22	1.91	0.32	1.72	193234.5	62	0.054
ER	0.83	2.46	0.19	0.48	83620	31	0.744
Nucleus	0.60	2.30	0.00	0.37	174068	74	0.030
Inner Membrane Complex	0.57	1.53	0.17	0.37	43346	17	0.410
Cytoplasm	0.54	1.19	0.00	0.42	254718	106	0.029
Food Vacuole	0.47	0.67	0.27	0.43	79508	27	0.624
Golgi	0.45	0.79	0.00	0.41	26268.5	10	0.491
Other	0.44	0.95	0.00	0.45	142575	61	0.037
Mitochondrion	0.19	0.46	0.00	0.16	52161.5	28	0.001
Apicoplast	0.11	0.18	0.00	0.17	75237.5	39	0.001

Protein localisation was classified using the ApiLoc resource. A total of 451 proteins were assigned a location. A Wilcoxon Rank-Sum test was performed on proteins from each subcellular location, comparing the percentage of residues targeted by non-synonymous SNPs for each protein in that location, to the distribution of SNPs within the entire *P. falciparum* proteome.

Chapter 3

Structural Features of *Plasmodium* Antigens: Linking 3D Structure to Immune Mediated Selection Pressure

Chapter 2 investigated the role of intrinsically disordered proteins as targets of natural immune responses. Protein disorder is at one end of a spectrum of structural behaviour, and we turn our attention here to proteins which are known to adopt a well-defined 3D structure. Antibodies primarily target protein antigens in their native state, and hence the conformational configuration of an antigen is a crucial factor that determines the location of potential epitopes. Furthermore, structural conformation is also an important consideration when examining selection pressures that arise as a result of antibody binding to antigen. It is thought that the majority of antibody epitopes are conformational in nature, with discontinuous regions of the linear protein sequence involved in the antibody-antigen interface [1–3]. The exact proportion of conformational to linear epitopes is unclear, and is likely to be highly dependent on the nature of the pathogen.

Antibody binding to a functional epitope can lead to a number of effector functions, including antibody-dependent cellular cytotoxicity (ADCC), complement activation and inhibition of receptor engagement via neutralisation and steric hindrance. This can lead to selective pressures on the target antigens, driving mutation of epitope sites to minimise antibody binding and subsequent effector function. This can occur within an individual (more common for rapidly mutating viruses) or at a population level. At a population level, immune-mediated selection pressure gives a selective advantage to low-frequency alleles/variants. This leads to balancing selection, in which multiple alleles are maintained within a population at higher frequency than would be expected under neutral selection processes. Sites of immune mediated selection pressure can be identified using metrics such as Tajima's D, with balancing selection giving rise to a Tajima's D value > 0 . Tajima's D has been used previously to identify malaria genes under balancing selection using genome-wide data [4,5] as well as identifying particular protein domains under balancing selection [6–13].

Traditional approaches to identify sites under selection pressure apply a function to either a whole gene or segment of chromosome, or as a sliding window across a gene/genome. In some cases selection pressures are also considered on a site-by-site basis (i.e. per nucleotide position). A sliding window approach involves applying a function or statistical test to successive 'windows' along the DNA sequence, with each window containing a fixed number of nucleotides. A single numerical result is reported for each window. This allows identification of particular regions or domains that

are subject to selection pressure. However, none of these approaches take into account the arrangement of a protein in 3D space.

Given that antibody-mediated selection pressures occur at the level of protein 3D structure, we proposed a modification to a standard sliding window approach that allows incorporation of protein structural information—performing a 3D sliding window over a protein structure. Such an approach allows aggregation of spatially linked data, and could be more sensitive in identifying sites of selection (or immunodominant functional antibody epitopes). To this end, a novel tool was developed and implemented in Python, allowing integration of structural information into tests of immune selection pressure. The development and usage of this tool is discussed in detail in Chapter 4. This tool was used extensively in the proteome-wide analysis of structured *P. falciparum* proteins in this chapter.

This chapter examines regions with experimentally characterised structures within *P. falciparum*, using polymorphism data from a study in The Gambia [5]. A number of immune-related features were mapped to known and predicted structures. Polymorphic residues had a strong propensity to be surface exposed and were enriched within certain secondary structure elements. Predicted MHC class II binding peptides were also mapped onto protein structure. These MHC binding peptides were primarily buried within the core of the protein and in general were not polymorphic. Sites of immune-mediated selection pressure were identified using Tajima's D applied to spatially aggregated data. With this novel approach, a region within *PfAMA1* involving both Domain II and Domain III was identified that had a high Tajima's D value relative to the rest of the protein; this site did not stand out when using a conventional sliding window approach. This result highlights additional epitopes of interest for consideration in future vaccine development work for malaria.

References

1. Van Regenmortel MHV. Mapping Epitope Structure and Activity: From One-Dimensional Prediction to Four-Dimensional Description of Antigenic Specificity. *Methods*. 1996;9: 465–472.
2. Barlow DJ, Edwards MS, Thornton JM. Continuous and discontinuous protein antigenic determinants. *Nature*. 1986;322: 747–748.
3. Forsström B, Axnäs BB, Rockberg J, Danielsson H, Bohlin A, Uhlen M. Dissecting antibodies with regards to linear and conformational epitopes. *PLoS One*. 2015;10: e0121673.
4. 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.
5. Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Elizabeth Deerhake M, et al. Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet*. 2012;8: e1002992.

6. Wang Y, Ma A, Chen S-B, Yang Y-C, Chen J-H, Yin M-B. Genetic diversity and natural selection of three blood-stage 6-Cys proteins in *Plasmodium vivax* populations from the China-Myanmar endemic border. *Infect Genet Evol.* 2014;28: 167–174.
7. 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.
8. Parobek CM, Bailey JA, Hathaway NJ, Socheat D, Rogers WO, Juliano JJ. Differing Patterns of Selection and Geospatial Genetic Diversity within Two Leading *Plasmodium vivax* Candidate Vaccine Antigens. *PLoS Negl Trop Dis.* 2014;8: e2796.
9. Arnott A, Mueller I, Ramsland PA, Siba PM, Reeder JC, Barry AE. Global Population Structure of the Genes Encoding the Malaria Vaccine Candidate, *Plasmodium vivax* Apical Membrane Antigen 1 (Pv AMA1). *PLoS Negl Trop Dis.* 2013;7: e2506.
10. Xangsayarath P, Kaewthamasorn M, Yahata K, Nakazawa S, Sattabongkot J, Udomsangpetch R, et al. Positive diversifying selection on the *Plasmodium falciparum* surf4.1 gene in Thailand. *Trop Med Health.* 2012;40: 79–89.
11. Reeder JC, Wapling J, Mueller I, Siba PM, Barry AE. Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection. *Malar J.* 2011;10: 126.
12. Moon S-U, Na B-K, Kang J-M, Kim J-Y, Cho S-H, Park Y-K, et al. Genetic polymorphism and effect of natural selection at domain I of apical membrane antigen-1 (AMA-1) in *Plasmodium vivax* isolates from Myanmar. *Acta Trop.* 2010;114: 71–75.
13. Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics.* 2001;158: 1505–1512.

SCIENTIFIC REPORTS

OPEN

Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures

Andrew J. Guy^{1,2}, Vashti Irani^{1,3}, James G. Beeson^{1,3,4}, Benjamin Webb⁵, Andrej Sali⁵, Jack S. Richards^{1,3,6,7} & Paul A. Ramsland^{1,2,8,9}

Humoral immune responses against the malaria parasite are an important component of a protective immune response. Antibodies are often directed towards conformational epitopes, and the native structure of the antigenic region is usually critical for antibody recognition. We examined the structural features of various *Plasmodium* antigens that may impact on epitope location, by performing a comprehensive analysis of known and modelled structures from *P. falciparum*. Examining the location of known polymorphisms over all available structures, we observed a strong propensity for polymorphic residues to be exposed on the surface and to occur in particular secondary structure segments such as hydrogen-bonded turns. We also utilised established prediction algorithms for B-cell epitopes and MHC class II binding peptides, examining predicted epitopes in relation to known polymorphic sites within structured regions. Finally, we used the available structures to examine polymorphic hotspots and Tajima's D values using a spatial averaging approach. We identified a region of PfAMA1 involving both domains II and III under a high degree of balancing selection relative to the rest of the protein. In summary, we developed general methods for examining how sequence-based features relate to one another in three-dimensional space and applied these methods to key *P. falciparum* antigens.

Malaria is an infectious mosquito-borne disease caused by *Plasmodium* species, responsible for an estimated 429,000 deaths in 2015¹. *P. falciparum* is the major cause of malaria-related mortality in humans, with *P. vivax* also contributing significantly to morbidity. A more complete description of the determinants for effective immune responses against the parasite is a crucial step in the development of a highly efficacious malaria vaccine. The structural state of an antigen is an important factor that contributes to selection of immunodominant epitopes². Protein conformational states spans a continuum between rigid, well-defined 3-dimensional (3D) structures and completely disordered states^{3–5}. Previously, we have explored the role that intrinsically disordered proteins play as potential antigens within *Plasmodium* species, with disordered domains displaying marked differences to structured domains including containing a paucity of MHC binding peptides, an increased number of tandem repeat segments and an increased proportion of polymorphisms⁶. In this study, we turn our attention to epitope location within structured protein domains. In particular, we utilise established B-cell epitope predictors and predictors of MHC binding, examining these features in relation to the location of immunologically relevant polymorphisms over regions of experimentally determined or modelled structure. Additionally, we incorporate structural information into a test for balancing selection, allowing for more powerful identification of structured regions under immune selection pressure.

¹Life Sciences, Burnet Institute, Melbourne, Australia. ²Department of Immunology, Monash University, Melbourne, Australia. ³Department of Medicine, University of Melbourne, Melbourne, Australia. ⁴Department of Microbiology, Monash University, Clayton, Victoria, Australia. ⁵University of California, San Francisco, San Francisco, California, USA. ⁶Department of Infectious Diseases, Monash University, Melbourne, Australia. ⁷Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Australia. ⁸School of Science, RMIT University, Bundoora, Australia. ⁹Department of Surgery Austin Health, University of Melbourne, Heidelberg, Australia. Correspondence and requests for materials should be addressed to J.S.R. (email: jack.richards@burnet.edu.au) or P.A.R. (email: paul.ramsland@rmit.edu.au)

Immunity against clinical malaria develops naturally following repeated exposure, with antibodies known to play a key role in this process^{7,8}. Within a naturally exposed population, immune selection pressure on the malaria parasite helps drive the occurrence of high-frequency polymorphisms on key malaria antigens. The development of a humoral immune response requires recognition of antigen in its native state. As a result, antigen structure plays a large role in the determination of epitopes for a humoral immune response. In other words, immune selection pressure driven by antibody-antigen interactions also occurs at the level of three-dimensional (3D) protein structure. Thus, examination of polymorphic regions in the context of protein 3D structure may help illuminate particular structural regions that are important targets of natural immunity. A number of studies have explored the relationship between protein structure and immune responses within *Plasmodium* species, including work on AMA1 from various species^{9–13}, CSP^{14–16}, EBA-175¹⁷, MSPDBL2¹⁸ and MSP2¹⁹. The majority of these studies have examined the location of polymorphic residues on a protein structure for single antigens, which likely arise as the result of immune selection pressure on particular epitopes. Polymorphisms can also arise as a result of T-cell driven selection pressure, as has been described for key T-cell epitopes within the C-terminal domain of CSP^{20,21}. Other tests of immune selection pressure include Tajima's D, which can help identify departure from a neutral model of selection²². A number of studies have examined *Plasmodium* proteins under immune selection pressure (balancing selection) using a sliding window approach^{9,10,23–26}, although all of these studies examine Tajima's D in the context of the linear sequence and do not consider the spatial proximity of residues (i.e., residues that are distant in the linear sequence may be proximal in the 3D structure). Here, we incorporate residue spatial information into measures of immune pressure, using both known and modelled protein structures. We demonstrate that the consideration of protein structural information can give extra insights into the regions of a protein under immune selection pressure.

In summary, we show that polymorphic residues within *P. falciparum* are usually surface exposed and are enriched within secondary structure turn elements. Predicted B-cell epitopes are also typically located on highly surface exposed regions. In contrast, predicted MHC class II binding peptides are generally buried within the core of a protein, and do not seem to overlap with polymorphic residues to a significant extent, which suggests that high frequency polymorphisms are more likely driven by humoral immune responses rather than cellular immunity. Antibodies often recognise discontinuous epitopes, therefore it is important to consider the spatial arrangement of residues when examining antigenicity. Accordingly, we incorporate structural information into a modified Tajima's D test, and assessed two polymorphic vaccine candidates, EBA-175 and AMA1. We identified strong signatures of balancing selection for a discontinuous region of PfAMA1 bordering domains II & III.

Methods

Data sources. Protein sequences for *Plasmodium* species were obtained from PlasmoDB, v28 (www.plasmodb.org)²⁷. *Plasmodium* genomes used were *P. falciparum* 3D7, *P. knowlesi* Strain H, *P. yoelii* 17X, *P. chabaudi* chabaudi, *P. vivax* Sal-1, *P. berghei* ANKA and *P. reichenowi* CDC. Coordinates for experimentally determined structures were obtained from the Protein Data Bank (PDB) from the Research Collaboratory for Structural Bioinformatics (RCSB) website (www.rcsb.org)²⁸, accessed on April 20, 2017. Data on polymorphisms from 65 Gambian isolates were obtained from PlasmoDB²⁴.

Identification of matching PDB structures. For each *Plasmodium* species examined, matching PDB structures were identified using a BLAST search against the PDB database, with an e-value cut-off of 10.0. A sequence identity threshold >90% was used, normalized to the length of the shorter sequence in the comparison. The NCBI-blast+ 2.3.0 package was used for all BLAST searches²⁹. Redundancy in PDB structures was removed using a sequence identity cut-off of 90% to group similar structures using precomputed sequence similarity clusters available on the RCSB PDB database (<http://www.rcsb.org/pdb/>).

Python BioStructMap package - Spatial averaging of data over a protein structure. The BioStructMap Python package was developed to map various features on a protein structure. BioStructMap contains methods that take a PDB file as input, alongside a set of data that is aligned to a reference sequence. The output is a map of this data on a three-dimensional (3D) structure using either a predefined or user-defined function, either with or without some level of spatial averaging. The package makes use of the BioPython *Bio.PDB* module for PDB file parsing and manipulation³⁰ and the DendroPy package for calculation of Tajima's D³¹. The source code is available at <https://github.com/andrewguy/biostructmap>. For all structural mapping work using BioStructMap, protein chains were treated as monomers (i.e. interactions between chains in multimeric complexes were ignored).

Comparative structure modelling of *P. falciparum* structures. Template-based models of *P. falciparum* structures were created using ModPipe³², an automated software pipeline that utilises MODELLER for the generation of comparative protein structure models³³. Models are accessible via ModBase (<https://modbase.compbio.ucsf.edu/>)³⁴. *P. falciparum* 3D7 sequences were used to compute all models. Structural models were deemed to be reliable if they had a ModPipe Quality Score (MPQS) greater than 1.1. The MPQS accounts for sequence coverage, sequence identity, gaps in the alignment, the compactness of the model and various statistical potential Z-scores³⁴.

Calculation of Tajima's D. Tajima's D is a statistical test used to identify regions of sequence evolving under non-neutral selection²². Tajima's D was used here to identify regions of sequence subject to balancing selection, in which a higher level of sequence diversity is maintained within a population than would be expected under a neutral model of selection. Balancing selection can arise as a result of immune selection pressure within a population, and is indicated by positive values of Tajima's D. Tajima's D was calculated using the DendroPy Python package, version 4.2.0³¹. For calculation of Tajima's D using a standard sliding window approach, the protein coding

region for each gene was selected based on the 3D7 reference strain, and the corresponding multiple sequence alignment used for a sliding window calculation of Tajima's D. A window size of 102 base pairs (bp) and step size of 3 bp was used unless otherwise specified. For calculation of Tajima's D with incorporation of protein structural information (referred to as spatially derived Tajima's D), the relevant PDB file sequence was aligned to the 3D7 reference strain, and a radius of 15 Å used to extract surrounding residues for each central residue, with Tajima's D calculated using codons for these surrounding residues. For a detailed description, see the BioStructMap documentation at <https://github.com/andrewguy/biostructmap>.

Sequences for calculation of Tajima's D were generated using polymorphism data obtained from PlasmoDB. Polymorphic sites were included on the proviso that at least 50 of the 65 sequences had a reliable base call at that position (percentage isolates with a base call $\geq 76\%$), in line with the original study²⁴. A read frequency threshold of 70% and a minimum read depth of 5 was used to identify reliable base calls.

Calculation of relative solvent accessibility and secondary structure. Relative solvent accessibility (RSA) was calculated using the DSSP program (accessed via BioPython)³⁵, using the maximum accessible surface area (ASA) values from Rost & Sander³⁶. When considering a two-state definition of solvent accessibility, an RSA threshold of 20% was used, similar to approaches taken in previous studies³⁷. Note that this threshold is very close to the median RSA value for the proteins considered in this study, effectively splitting the dataset in two. Secondary structure assignment was also performed using the DSSP program using the eight DSSP secondary structure classes³⁵: H, Alpha helix; B, Beta bridge; E, Strand; G, 3-Turn Helix; I, 5-Turn Helix; T, Turn; S, Bend; –, Other/Coil.

Amino acid propensity scales. Average hydrophilicity and hydrophobicity within a given radius were calculated using the Hopp-Woods hydrophilicity scale³⁸ and the Kyte-Doolittle hydrophobicity scale³⁹, respectively.

Prediction of B-cell epitopes. Potential B-cell epitopes were assessed using both BepiPred 1.0⁴⁰ and BepiPred 2.0⁴¹, as we have previously used BepiPred 1.0 when examining potential epitopes within disordered *P. falciparum* proteins⁶ but wished to utilise the more accurate BepiPred 2.0 algorithm. BepiPred 2.0 uses a linear protein sequence to predict conformational B-cell epitopes, and is trained on epitope data from experimental crystal structures.

Analysis of MHC binding peptides. Prediction of peptide binding to MHC was performed with NetMHCII 2.2^{42,43}, with a 15 residue peptide length and default settings (threshold = -99.9 , P1 amino acid residue preference turned off). Selection of MHC II alleles for analysis was performed according to known MHC haplotype frequencies from within a Gambian population⁴⁴, obtained from www.allelefreqencies.net⁴⁵.

Prediction of membrane-proximal regions. To assess the potential impact of membrane proximity, transmembrane regions were predicted using TMHMM v2.0⁴⁶. Known and predicted GPI-anchored proteins were obtained from Gilson *et al.*⁴⁷, and the position of the GPI omega site within these proteins was predicted using PredGPI⁴⁸. Structured regions were considered to be proximal to a transmembrane region or GPI anchor if they were within 10 residues of either.

Visualisation of protein structures. Protein structures and features mapped on them were visualised using PyMol⁴⁹; feature values were saved into the B-factor column of a PDB file, and visualised using the *spectrum* command.

Data analysis and statistics. The majority of data analysis was performed using the Anaconda distribution of Python 3.5. Data was stored in an SQLite database, and accessed using Python SQLAlchemy. Plotting was performed with the Python Matplotlib package, version 1.5.1⁵⁰. Statistical analysis was performed using SciPy⁵¹.

Results

A Python package for spatial mapping of features onto PDB structures. *Plasmodium* structures identified in the PDB were used for a series of spatial mapping analyses using our BioStructMap Python package. When dealing with protein sequence-level data, such as sequence polymorphisms or measures of evolutionary selection pressure, it is common practice to apply a function over a sliding window on the linear sequence. However, spatial proximity of residues within a 3D folded protein structure is often important, motivating an application of a “3D sliding window” across a protein structure as implemented in BioStructMap (Fig. 1). We were particularly interested in assessing immune-mediated selection pressure that occurs as a result of antibody recognition of a dominant epitope. Such selection pressure often applies to multiple residues within the epitope. Importantly, these residues are often not continuous within a linear sequence; therefore, consideration of protein structural information may enhance identification of regions of immunological importance. In this setting, the selection of a radius for calculation of adjacent residues should be reflective of the potential binding surface for an antibody. While the interaction dimensions for antibody-antigen interaction surfaces vary depending on the sequence of both antigen and antibody, the mean maximum paratope dimension were estimated as 28 Å⁵² or 30 Å⁵³. We have used the upper of these two estimates when choosing a radius of 15 Å for all spatial averaging presented in this paper. Using data on sequence polymorphism from a study in The Gambia²⁴, we mapped known polymorphisms onto *P. falciparum* crystal structures, correlating the occurrence of polymorphisms with residue surface exposure³⁶, average hydrophilicity³⁸ and hydrophobicity³⁹, and predicted B-cell epitopes⁴¹ and MHC class II binding peptides^{42,43}. We used the BioStructMap package to identify polymorphic hotspots within regions with known or modelled structure. Finally, we used this tool to include spatial information into a calculation of Tajima's D and applied this to key antigens.

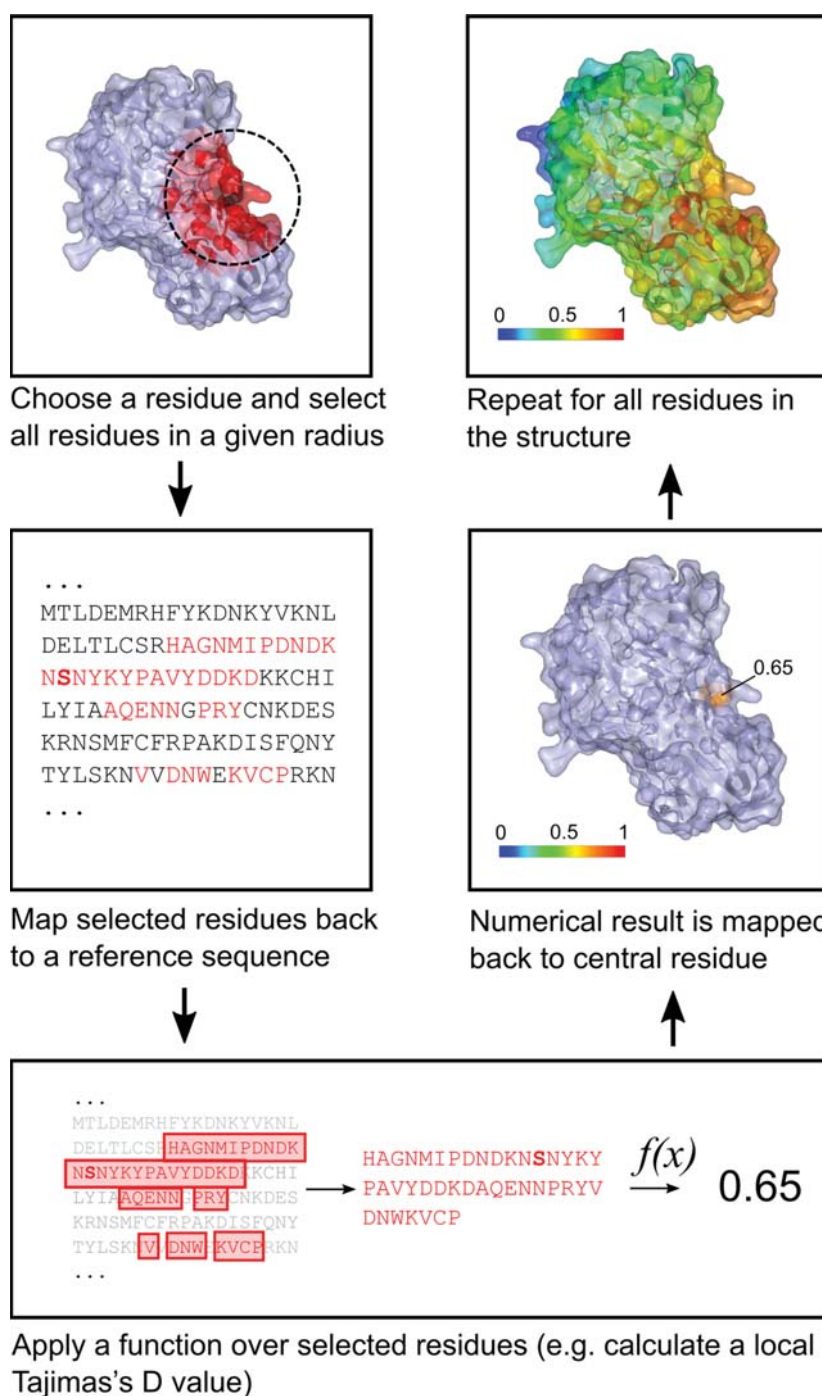


Figure 1. Overview of the structural mapping with spatial averaging approach as used in the Python BioStructMap package. For any given residue within a PDB structure, all residues within a specified radius of the given residues are identified. The location of these residues within a given reference sequence is then found, with the assumption that user-provided data will be aligned to this reference sequence. Using selected residues and the corresponding subset of user-provided data, a function is called, returning (usually) a numerical value. For example, this function may return the mean of the respective data. Note that the provided data and the mapping function may take a diverse number of forms. This includes functions which apply some statistical test over a multiple-sequence alignment of genetic sequences (e.g. Tajima's D). In this case, the function would apply the statistical test over the subset of codons which code for the selected residues. The returned value is then assigned to the original residue in the PDB structure. This process is repeated for all residues within the PDB structure. Results can be viewed as a heatmap displayed over the PDB structure.

Species	Total protein coding genes	No. of proteins with PDB matches	Unique PDB Structures		
			All	X-ray Diffraction	NMR
<i>P. falciparum</i>	5398	275	368	229	31
<i>P. vivax</i>	5530	157	234	150	13
<i>P. knowlesi</i>	5316	124	176	99	10
<i>P. reichenowi</i>	5707	267	353	216	27
<i>P. berghei</i>	4952	108	143	81	8
<i>P. chabaudi</i>	5200	109	145	83	8
<i>P. yoelii</i>	5928	109	145	82	10

Table 1. Number of proteins within various *Plasmodium* species which have at least one matching PDB structure, and number of unique PDB structures matched to various *Plasmodium* species grouped by experimental technique. Note: Matching PDB structures were identified using a BLAST search against the PDB database, with an e-value cutoff of 10.0. A BLAST identity score of 90% was used as a cutoff for identifying close matches. When a PDB structure matched to multiple proteins within an organism, only the match with the highest identity score was considered for this table. Redundancy in PDB structures was removed using a Sequence Identity Cutoff of 90% to group similar structures using precomputed sequence identity clusters available on the RCSB PDB database (<http://www.rcsb.org/pdb/>). Plasmodium genomes used were *P. falciparum* 3D7, *P. knowlesi* Strain H, *P. yoelii* 17X, *P. chabaudi* chabaudi, *P. vivax* Sal-1, *P. berghei* ANKA, *P. reichenowi* CDC.

Overview of experimental protein structures within Plasmodium species. To determine the number of proteins from various *Plasmodium* species that have structures in the PDB, we performed a BLAST search of proteins from major human, primate and murine *Plasmodium* spp. against a database of all PDB sequences. BLAST matches were grouped according to experimental technique (X-ray diffraction, NMR, other) and redundancy across PDB files was accounted for by clustering redundant PDB sequences using a 90% sequence identity threshold. For *Plasmodium falciparum* 3D7 sequences, there were a total of 368 non-redundant structures (total of 1,111 PDB files), equating to a total of 275 proteins (5.1% of the proteome) with at least one known structure in the PDB (Table 1). The majority of these structures were determined using X-ray crystallography. Similarly, for *P. vivax* Sal-1 sequences, 234 non-redundant structures (810 PDB files) were found, covering a total of 157 proteins (2.8% of the proteome).

We examined the set of matching PDB structures for cross-species homology (Fig. 2). When comparing three human malaria species (*P. falciparum*, *P. vivax*, *P. knowlesi*), there was a large core of 146 non-redundant structures (596 PDB files) common to all 3 species, with another 206 non-redundant structures (490 PDB files) that matched only proteins in *P. falciparum*, while only 60 non-redundant structures (136 PDB files) matched proteins in *P. vivax* alone. For the 3 rodent malaria species examined (*P. yoelii*, *P. chabaudi*, *P. berghei*), 133 of the 155 non-redundant structures (602 of the 635 PDB files) matched proteins in all 3 species. Similarly, there was a large overlap between *P. falciparum* and *P. reichenowi*, with only 23 of the 372 non-redundant structures (58 of the 1127 total PDB files) matching proteins in only one of these two species. When considering all 7 species examined here, 81.2% (388 of 478) of non-redundant structures that matched proteins in any *Plasmodium* sequence also matched proteins in two or more *Plasmodium* species.

To determine the number of structured regions that were proximal to either transmembrane regions or GPI anchors, we predicted transmembrane regions using TMHMM v2.0⁴⁶ and extracted known and predicted GPI-anchored proteins from Gilson *et al.*⁴⁷. Only 5% (13 of 275) of proteins with known structures had structured regions within 10 residues of predicted transmembrane domains or GPI omega sites. Given the minimal number of membrane proximal structures identified, we have not treated membrane proximal regions differently in the following analysis.

Polymorphic residues are predominantly surface-exposed. We first correlated the occurrence of non-synonymous single nucleotide polymorphisms with the relative solvent accessibility (RSA) of the corresponding amino acid residues; RSA is a proportional measure of exposure of a particular amino acid residue to bulk solvent. Based on a few known examples^{10,54}, it was expected that polymorphic residues that evolved to evade humoral immune responses would be located predominantly on the surface. We indeed observed that polymorphic residues generally had some level of solvent exposure, with higher overall RSA values than residues that contained no underlying nucleotide polymorphisms (Fig. 3). Polymorphic residues had a significantly higher median RSA value of 0.47 than the background median RSA level of 0.20 ($p < 0.0001$, Mann-Whitney U test). Residues with underlying synonymous SNPs did not have significantly different RSA values to the background distribution ($p = 0.79$, Mann-Whitney U test). Immunologically relevant polymorphisms are expected to be maintained at a high frequency within a population. A minor allele frequency (MAF) threshold of 5% is commonly used to distinguish between high- and low-frequency polymorphisms^{55–58}. Correspondingly, polymorphic residues with a MAF $\geq 5\%$ had significantly higher RSA than all polymorphic residues ($p = 0.04$, Mann-Whitney test), with a median RSA of 0.52.

Average hydrophilicity and hydrophobicity in relation to polymorphic residues. A number of epitope prediction methods have utilised amino acid residue propensity scales based on residue biophysical properties such as hydrophobicity and hydrophilicity in an attempt to predict targets of adaptive immune

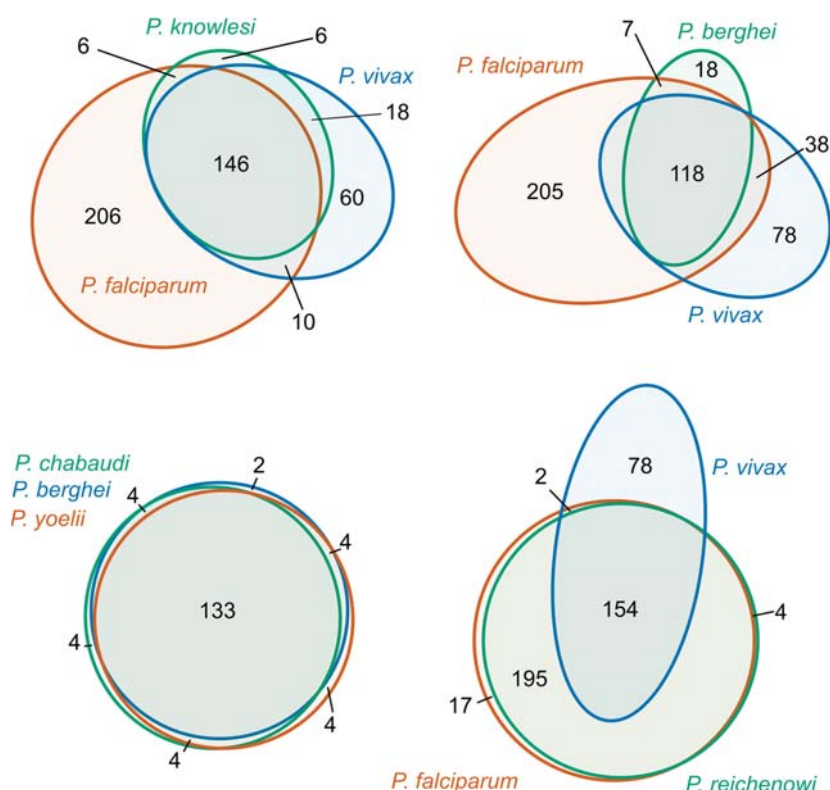


Figure 2. Comparison of unique protein structures which match to *Plasmodium* sequences with >90% identity. Euler diagrams show the number of PDB structures which match to single/multiple/all species. For clarity, only four representative combinations of species are shown. Matching PDB structures were identified using a BLAST search against the PDB database, with an e-value cutoff of 10.0. A BLAST identity score cutoff of 90% was used for included matches. Redundancy in PDB structures was removed using a Sequence Identity Cutoff of 90% to group similar structures using precomputed sequence identity clusters available on the RCSB PDB database (<http://www.rcsb.org/pdb/>). Only a single representative structure from each group of redundant structures was counted when generating Euler diagrams.

responses^{38,40,59–62}. We considered here the effect of average residue hydrophobicity and hydrophilicity within a 15 Å radius on polymorphism location, but observed only a small increase in hydrophilicity and corresponding small decrease in hydrophobicity (Figure S1a,c). Given that average hydrophilicity and hydrophobicity are correlated weakly with RSA, it is likely that the small shift in average hydrophilicity and hydrophobicity is reflective of the much larger shift observed for solvent accessibility between polymorphic and non-polymorphic residues. To explore this hypothesis, we re-ran the spatial averaging algorithm and restricted it to surface exposed residues (RSA ≥ 0.2), showing only a small, albeit significant, difference between polymorphic and non-polymorphic residues (Figure S1b,d).

Polymorphic residues occur preferentially in turns. We also investigated whether or not polymorphic residues are more likely to occur in any particular secondary structure motifs. When considering all polymorphic residues, we observed a significant reduction in polymorphic residues within β-strand elements (E), from 19% to 11% ($p = 0.006$) (Figure S2). When considering only those polymorphisms with MAF ≥ 5%, we observed a large reduction in the proportion of residues within β-strand elements (E) ($p = 0.0009$), from 19% to 6%, alongside a large increase in the proportion of residues within turns (T), from 9% to 22% ($p < 0.0001$).

Prediction of B-cell epitopes and MHC class II binding peptides. Polymorphisms that arise as a result of immune-mediated selection pressure may be driven by antibody or T-cell responses. T-cells recognise peptide antigen in the context of MHC molecules, whereas antibodies typically recognise antigen in its native state. In the context of a humoral immune response, CD4+ T-cell responses are important for the development of a T-dependent B-cell response, recognising peptide antigen presented by B-cells on MHC class II molecules. To examine the location of predicted B-cell epitopes in relation to regions of known structure in *P. falciparum* we used the newly released BepiPred 2.0⁴¹ (Figure S3). Although a number of tools exist for predicting B-cell epitopes based on structural data^{63,64}, we wished to assess the utility of a state-of-the-art method that only uses the linear protein sequence as input, as this will be applicable to the many *Plasmodium* proteins with unknown structures. As we have previously utilised BepiPred 1.0 in a study examining disordered proteins in *P. falciparum*⁶, we provide a comparison to the older BepiPred 1.0⁴⁰ algorithm here (Figure S4). Our subsequent analysis was

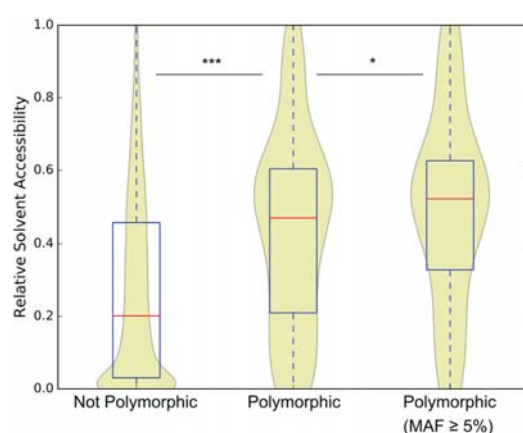


Figure 3. Polymorphic residues within known *P. falciparum* structures are predominantly surface exposed. Relative solvent accessibility (RSA) is shown for residues with and without identified polymorphisms. RSA represents the proportional surface area of a residue that is exposed to solvent, relative to the maximum possible exposure for that amino acid. RSA was calculated using the maximum accessible surface area (ASA) values from Rost & Sanders³⁶. Box-and-whisker plots show median (red line) and interquartile range (box) of residue RSA values for each group. Violin plots show the smoothed distribution of RSA values for each group (violin plots employ a Kernel Density Estimation to compute an empirical probability distribution for each group). Polymorphic residues shown are both those with underlying non-synonymous SNPs regardless of allele frequency ($n = 204$), and those with underlying non-synonymous SNP with a minor allele frequency (MAF) $\geq 5\%$ ($n = 105$). The majority of residues in the dataset did not have underlying polymorphisms ($n = 28,869$). Sequence polymorphisms were obtained from 65 Gambian isolates from Amambua-Ngwa *et al.*²⁴, accessed via PlasmoDB. Polymorphic residues had significantly higher RSA values than the background RSA levels ($p < 0.0001$, Mann-Whitney U test), and polymorphic residues with a MAF $\geq 5\%$ had significantly higher RSA than all polymorphic residues ($p = 0.04$, Mann-Whitney U test).

only performed on proteins that contained at least one high frequency polymorphic residue within a region of known structure, as it was impractical to run the full proteome through the BepiPred 2.0 online interface. Epitopes predicted with BepiPred 2.0 were predominantly surface exposed, with increasing BepiPred thresholds predicting increasingly surface exposed residues (Figure S3a). BepiPred 2.0 epitope scores were significantly higher for polymorphic residues ($p < 0.0001$; Mann-Whitney U test), even when restricting analysis to surface exposed residues ($\text{RSA} \geq 0.2$) ($p < 0.0001$; Mann-Whitney U test) (Figure S3b). At the default threshold, 50% of non-polymorphic surface-exposed residues fell within predicted epitopes, compared to 75% of polymorphic surface-exposed residues.

We also examined the location of predicted MHC class II binding peptides, specifically looking at MHC class II haplotypes, again using those that have been shown to be present at high frequency within the Gambian population: HLA-DPA1*02:01-DPB1*01:01 and HLA-DQA1*05:01-DQB1*03:01⁴⁴. MHC class II binding peptides were observed to be predominantly buried within the core region of proteins, with lower overall surface exposure than the set of non-binders (Figure S3c, d). Additionally, polymorphic residues were found at lower frequency within MHC class II binding peptides compared to non-binding regions. For the HLA-DQA1*05:01-DQB1*03:01 haplotype, 0.20% of residues involved in high-binding peptides were polymorphic (MAF $\geq 5\%$), compared to 0.74% of residues not involved in a predicted MHC binding peptide. Similarly for the HLA-DPA1*02:01-DPB1*01:01 haplotype, 0.21% of residues involved in high-binding peptides were polymorphic (MAF $\geq 5\%$), compared to 0.71% of residues not involved in a predicted MHC binding peptide. Thus, MHC class II binding peptides are typically buried within the hydrophobic core of proteins, and do not appear to be significant targets of immune selection pressure.

Comparative structural modelling of *Plasmodium falciparum* proteins. To extend the structural coverage, the entire proteome of the *P. falciparum* (3D7 strain) was subjected to comparative structural modelling using ModPipe³² (<https://salilab.org/modpipe/>) based on structures in the PDB. A MPQS threshold of 1.1 was used to filter out low-quality models before further analysis. A total of 1575 reliable models were created, covering 923 proteins or 17% of the proteome. The majority of the models in the filtered dataset covered the entire length of the corresponding *P. falciparum* 3D7 protein sequence (median coverage = 95%, mean coverage = 87%) (Figure S5). We then used both known and modelled structures to identify polymorphic hotspots within *P. falciparum*.

Identification of polymorphic hotspots. Polymorphic regions have often been thought to relate to potential antigenicity of malaria proteins, with antigenic diversity a contributing factor to parasite evasion of host immune responses^{65,66}. With this in mind, spatial averaging of polymorphisms was performed to identify regions of proteins that have clusters of high-frequency polymorphisms and are hence likely to be under some level of immune selection pressure. Given that polymorphic residues tend to be surface exposed, we have restricted this

analysis to surface exposed residues with $RSA \geq 0.2$. Polymorphic hotspots on *P. falciparum* structures were identified using the set of all PDB crystal structures that matched proteins in *P. falciparum* (Table S1), and the set of all modelled structures with $MPQS > 1.1$ (Table S2). Protein polymorphisms from the Gambian population were used, and a MAF threshold of 5% employed to restrict polymorphisms to those with some immunological relevance. A threshold of either 10% or 20% of surrounding residues being polymorphic was used to identify regions of protein structure that are particularly polymorphic. For both experimental and modelled sets of structures, most proteins identified are known antigens, and include AMA1, CSP, TRAP, PfEMP1, DBL-MSP2, MSP1 and EBA-175. Mapping of polymorphisms and predicted B-cell epitopes and MHC class II binding peptides on these structures is shown in Figures S6–S10 & Figs 4 and 5. We have used known structures where there are no missing residues, and modelled structures when the relevant experimental structures have unresolved residues. Although the density of polymorphic residues differs between antigens, all proteins examined here have regions of particularly dense polymorphisms that in most cases overlap with B-cell epitopes predicted using BepiPred 2.0. In many cases, the most polymorphic regions are surface exposed protrusions.

Key parameters mapped to PfAMA1 & EBA-175. From the proteins identified above, we focus here on two major vaccine candidates for *P. falciparum* malaria: Apical Membrane Antigen 1 (AMA1) and Erythrocyte Binding Antigen 175 (EBA-175) Region II (RII) (Figs 4 and 5). A number of crystal structures exist for PfAMA1, but they only encompass domains I and II. Previous studies suggest that domain III may also be a significant target of protective antibody responses^{10,67}, and hence we used a modelled structure of all three domains to calculate polymorphic hotspots and spatially derived Tajima's D values. There are two models for PfAMA1 generated by ModPipe with $MPQS > 1.1$; one of these models uses a PfAMA1 structure as a template and hence does not include domain III, whereas the other model uses a PvAMA1 structure as a template and includes all three domains. However, this second model fails to accurately place a loop (S345–Y397; 3D7 sequence) that is unresolved within the PvAMA1 template but resolved within known PfAMA1 structures. Previously published work has manually modelled the full domains I–III of PfAMA1 using a combination of *P. falciparum* and *P. vivax* templates, and we have used this model for examining PfAMA1¹⁰. While the crystal structure for EBA-175 RII has been solved⁶⁸, we have used a ModPipe homology model for analysis here as this model includes a number of residues that are unresolved in the experimental structure. These unresolved residues occur in loops that are no more than 4 residues long, and as such the model is likely not to be inaccurate.

For PfAMA1, the majority of polymorphic residues are surface exposed, with most polymorphisms falling within Domain I (Fig. 4a). The most polymorphic region (Fig. 4b) is a highly surface-exposed loop formed by residues T194–D212, which has been termed the C1-L cluster^{69,70}. Potential B-cell epitopes were predicted using BepiPred 2.0 (Fig. 4c). The predicted epitopes (especially at higher thresholds) predominantly fell on highly surface-exposed regions. In contrast, predicted MHC class II binding peptides were predominantly buried within the core of the protein, especially for predicted high-affinity peptides (Fig. 4d,e). For MHC class II haplotypes with high frequency within the Gambian population (e.g., the HLA-DPA1*02:01-DPB1*01:01 haplotype), 50% of residues involved in predicted high-binding peptides were located in the core of the protein ($RSA < 0.2$), while 57% of residues involved in low-binding peptides were also buried. For HLA-DQA1*05:01-DQB1*03:01, 83% and 43% of residues from predicted high- and low-binding peptides, respectively, were buried in the protein core (47% of all residues are buried).

When considering EBA-175 RII polymorphisms (Fig. 5a), two main hotspots were identified, with both being surface-exposed loops (Fig. 5b). These two loops are located on opposing faces of the structure, with one loop within the F1 domain (residues N252–V266) and the other within the F2 domain (residues S432–N442). The F2 surface loop is involved in the formation of a two-strand antiparallel β -sheet between identical residues (N433–H436) upon EBA-175 RII dimer formation⁶⁸. This region is in the center of the RII dimer, with residues K439 and K442 also likely involved in binding to glycans in the glycophorin A receptor. The F2 β -finger (residues C476–C488) that is the target of inhibitory monoclonal antibodies R217 and R215^{17,71} was also identified as a polymorphic hotspot, although to a lesser extent. Similar to PfAMA1, predicted B-cell epitopes were located on highly surface-exposed regions of EBA-175 RII (Fig. 5c) and include the known F2 β -finger epitope and an epitope (H303–Q315) that is the target of the R218 monoclonal antibody⁷². Conversely, MHC class II binding peptides were mostly buried (Fig. 5d,e). For the HLA-DPA1*02:01-DPB1*01:01 haplotype, 70% of residues involved in predicted high-binding peptides were located in the core of the protein ($RSA < 0.2$), while 58% of residues involved in low-binding peptides were also buried. For HLA-DQA1*05:01-DQB1*03:01, 50% and 65% of residues from predicted high- and low-binding peptides respectively were buried in the protein core (44% of all residues are buried). In summary, a number of highly polymorphic surface exposed regions were identified for both PfAMA1 and EBA-175 RII, which in most cases overlapped with predicted B-cell epitopes. In contrast, predicted MHC II binding peptides were predominantly buried within the core of the protein.

Incorporating protein spatial information into a genetic test for immune selection pressure.

To assist in the identification of regions of protein under immune selection pressure, we developed a modified calculation of Tajima's D that includes protein structural information using the spatial averaging approach introduced earlier in this study. We compared this approach to application of a standard sliding-window over the linear sequence to assess whether our spatially derived Tajima's D calculation improved the ability to detect sites under immune selection pressure. These two methods were applied to PfAMA1 and EBA-175 (Fig. 6). For AMA1, there is evidence for balancing selection within DI when calculating Tajima's D using a traditional sliding window approach (Fig. 6c), as has been observed previously in other populations¹⁰. In contrast to the linear sliding-window approach, the new spatial averaging approach reveals a surface exposed region on the border of DII and DIII as the area with the highest Tajima's D values; parts of DI also appear to be under balancing selection (Fig. 6a). As expected, the so-called 'silent face' of PfAMA1 had Tajima's D values that were negative or close to

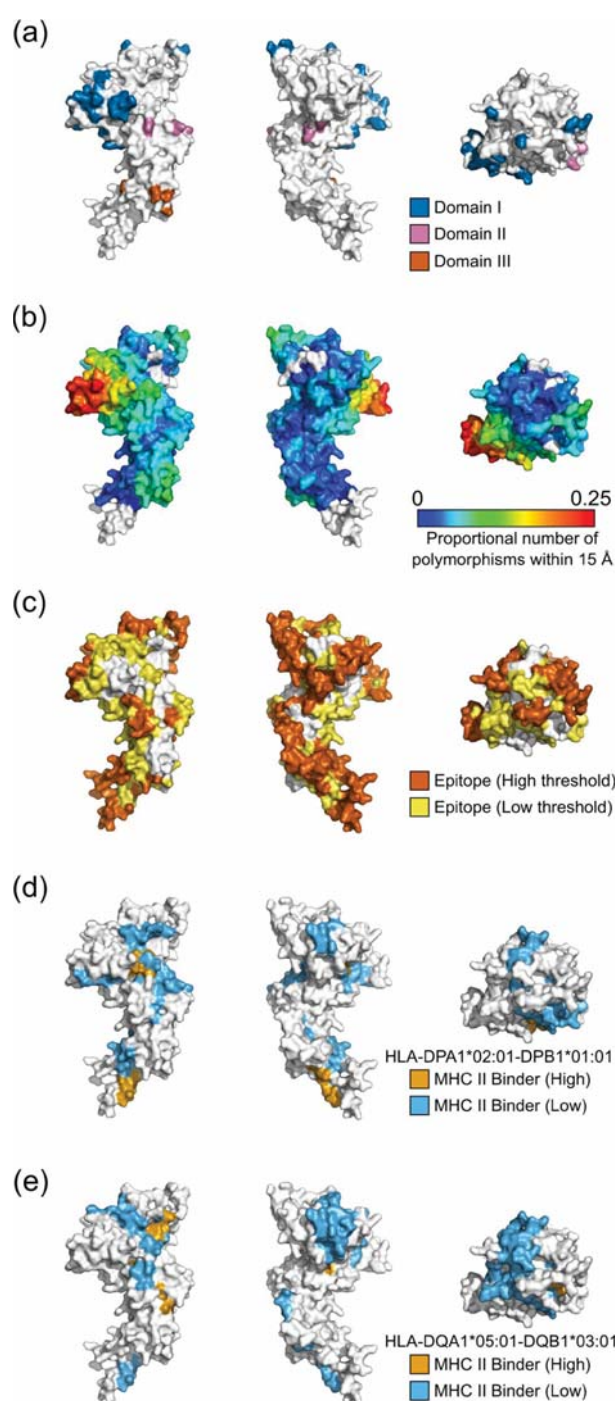


Figure 4. Location of immunologically relevant features mapped onto a *PfAMA1* structural model. Each panel shows the front, back and top view of the modelled *PfAMA1* structure. (a) Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown colored according to location within domain I (blue), domain II (magenta) or domain III (orange). Sequence polymorphisms were obtained from 65 Gambian isolates²⁴. (b) Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. (c) Bepipred 2.0 predictions are shown over the *PfAMA1* structure, with epitopes shown for two Bepipred thresholds—predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). (d,e) The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (d) and HLA-DQA1*05:01-DQB1*03:01 (e) alleles. Residues involved in a low binding peptide (50 nM $<$ IC₅₀ $<$ 500 nM) are shown in light blue, while residues involved in a high binding peptide (IC₅₀ $<$ 50 nM) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

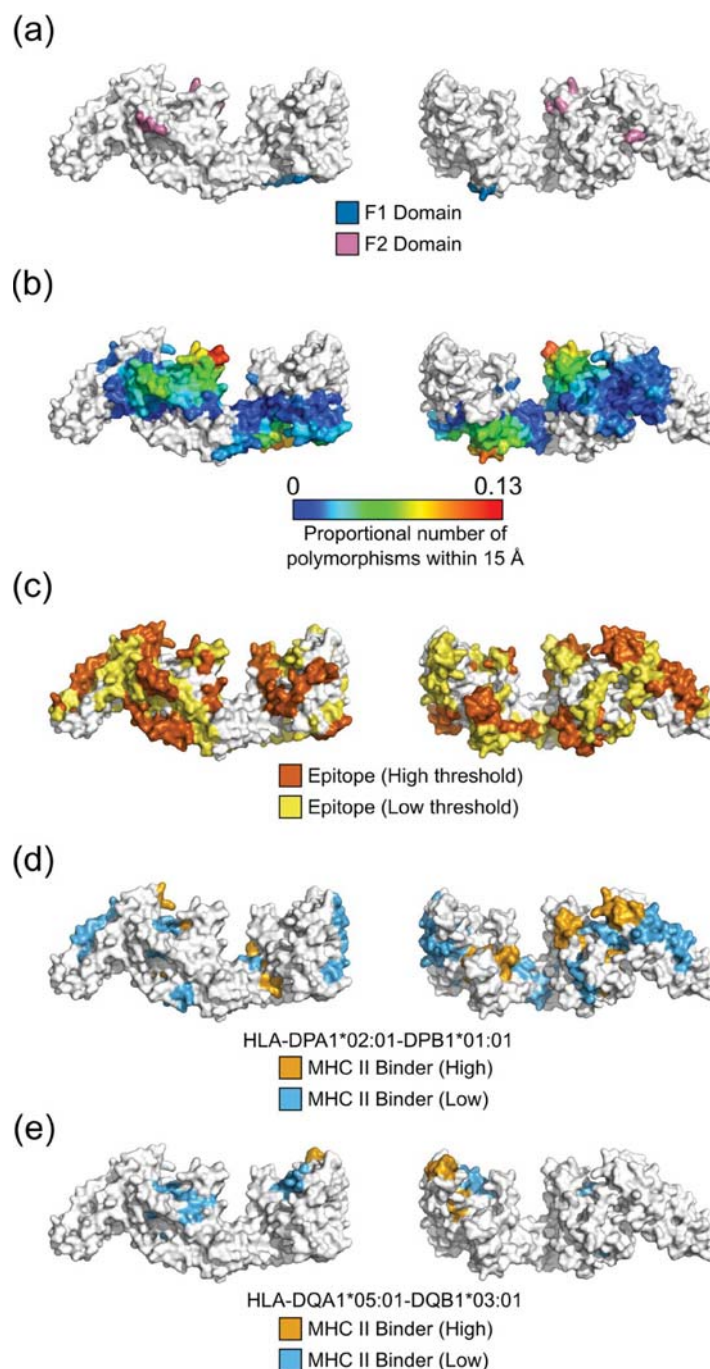


Figure 5. Location of immunologically relevant features mapped onto an EBA-175 RII homology model. Each panel shows the front and back view of the modelled EBA-175 structure. The homology model was modelled from the 1ZRO PDB structure using ModPipe. (a) Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown colored according to location within Region I (blue) and Region II (magenta). Sequence polymorphisms were obtained from 65 Gambian isolates²⁴. (b) Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF \geq 5%. (c) Bepipred 2.0 predictions are shown over the EBA-175 structure, with epitopes shown for two Bepipred thresholds: predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). (d,e) The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (d) and HLA-DQA1*05:01-DQB1*03:01 (e) alleles. Residues involved in a low binding peptide (50 nM < IC₅₀ < 500 nM) are shown in light blue, while residues involved in a high binding peptide (IC₅₀ < 50 nM) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

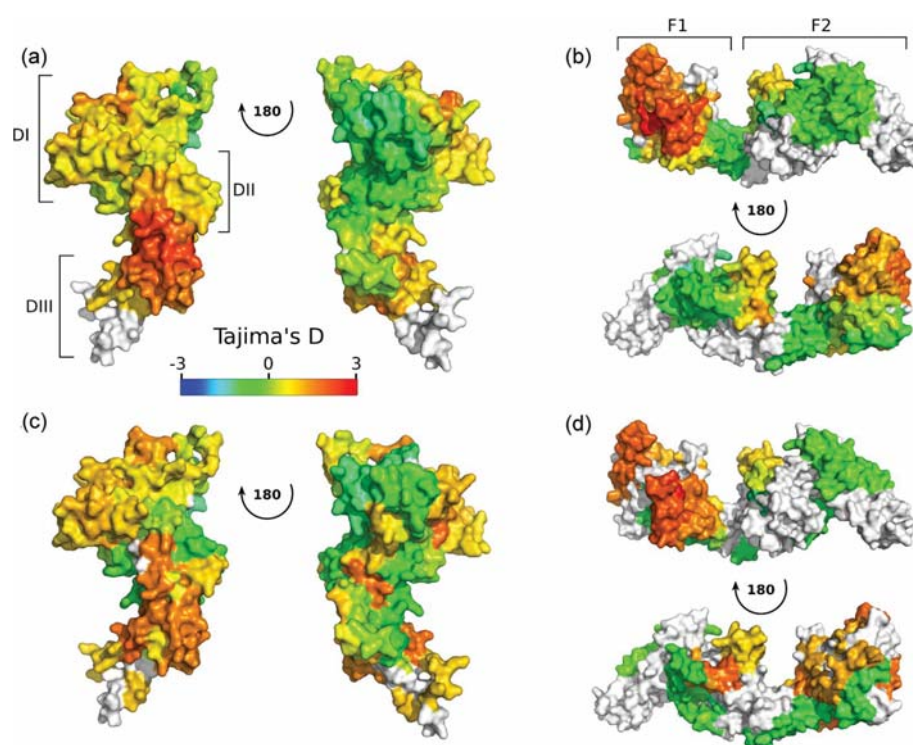


Figure 6. Calculation of Tajima's D for *PfAMA1* and EBA-175, both with and without incorporation of protein structural information. (a, b) Spatial information incorporated into a calculation of Tajima's D using modelled protein structures for AMA1 (a) and EBA-175 RII (b). Tajima's D values for each residue were calculated using only those codons which were mapped to residues within a 15 Å radius of the central residue. (c,d) Tajima's D was calculated over a sliding window of 102 bp and a step size of 3 bp, without incorporation of protein structural information. Tajima's D values for central codons are displayed on the modelled protein structures for AMA1 (c) and EBA-175 RII (d). Data on sequence polymorphisms was obtained from PlasmoDB using sequences from 65 Gambian isolates²⁴. The structural model for *PfAMA1* was manually modelled, previously published in Arnott *et al.*¹⁰, and covers domains I-III of *PfAMA1*. The structural model for EBA175 RII was created using Modpipe, with the PDB structure 1ZRO used as a template. Structures are colored according to the calculated value of Tajima's D mapped to each residue, with residues without a defined Tajima's D value shown in white.

zero. For EBA-175, a large portion of the F1 domain appears to be under balancing selection, as is a surface loop in the F2 domain (residues S432 - N442) (Fig. 6b,d). The region with highest calculated spatially derived Tajima's D values is contained within the F1 domain of RII, comprised predominantly of residues E266-D289, P314-Q322 and L382-L400. This site is also part of the dimerization interface formed between two molecules of EBA-175 RII as it binds to its glycoprotein A receptor; during dimerization this site makes contact with the F2 domain of the other dimer pair⁶⁸. It has previously been suggested that antibodies that block dimerization of EBA175 may negatively impact on glycoprotein A engagement⁷¹, and antibodies that inhibit binding of EBA-175 to glycoprotein A have also been shown to be associated with protection from clinical malaria⁷². Results from spatially derived Tajima's D and conventional sliding window Tajima's D analysis are similar to each other for EBA-175. This structure is predominantly alpha-helical, and sites with high Tajima's D values are mostly continuous stretches of protein sequence.

Previous analyses of the cross-reactivity of human antibodies, and vaccine-induced antibodies in rabbits, to different AMA1 alleles^{73,74} and mutation studies of C1 residues⁷⁵ suggested that polymorphisms in the C1 cluster did not explain a large component of antigenic differences between alleles. Furthermore standard analysis of polymorphisms in linear sequences did not correlate highly with antigenic differences^{74,75}. Therefore, additional approaches that consider structure may be required to yield further insights. Interestingly, our analysis of AMA1 incorporating structural considerations identified a site within DII/DIII of AMA1 that stands out as having a high spatially derived Tajima's D score. This site is composed of four distinct regions of continuous sequence (P303 - G313; L419 - I426; V437 - I454; D483 - F505), which together make up a surface exposed face of DII/DIII (Fig. 7). The spatial proximity of these regions is not accounted for when performing a sliding window analysis, and it is only when using spatial information that we observe the highest Tajima's D score of 2.39, compared to 1.84 for a linear sliding window calculation.

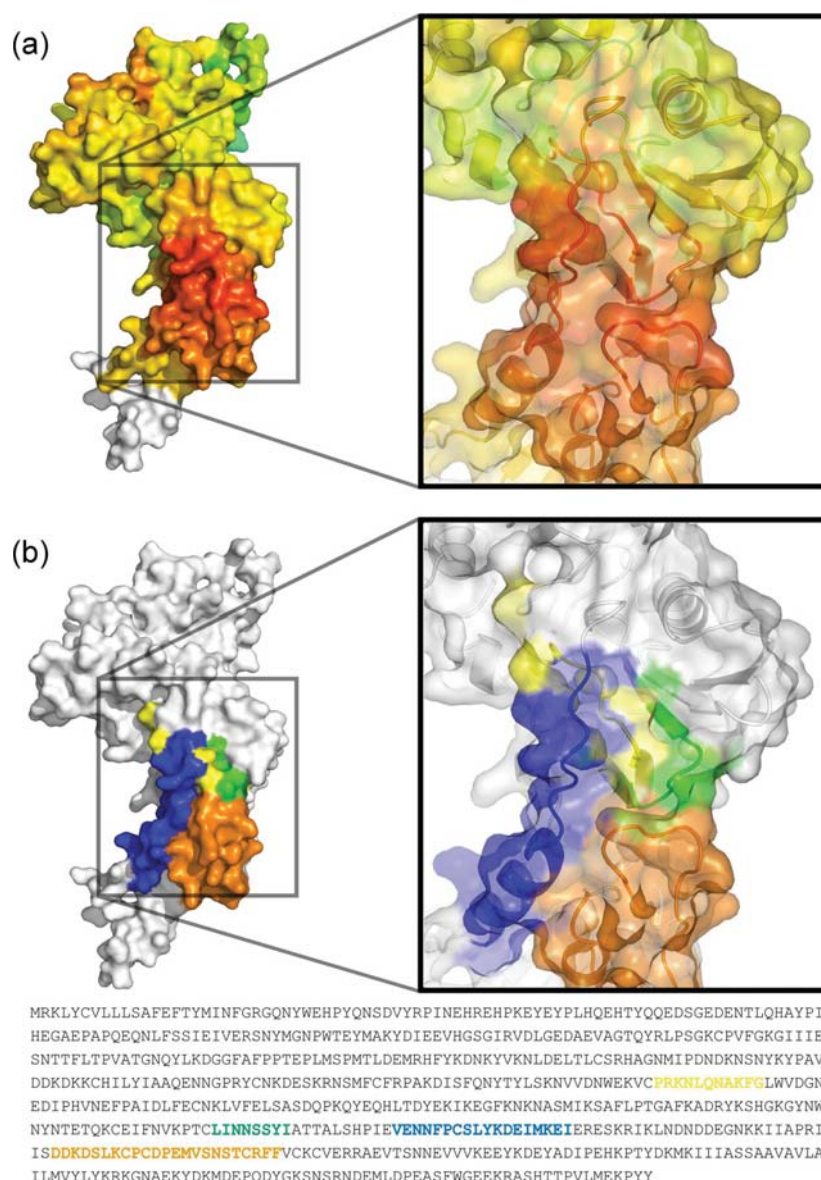


Figure 7. Four discontinuous stretches of sequence make up a region of *PfAMA1* with high Tajima's D values as calculated using spatial mapping. (a) Detailed view of a region of *PfAMA1* with high Tajima's D values as calculated using spatial averaging. The protein structure is colored according to the color scale presented in Fig. 6, with red residues corresponding to the highest Tajima's D values. Without the use of spatial averaging, there is a maximum Tajima's D value of 1.84 within this region, whereas a maximum value of 2.39 is observed when incorporating spatial information. (b) Four discontinuous regions of sequence contribute to the set of surface exposed residues with highest Tajima's D values. These four regions are shown in yellow (P303 - G313; DII), green (L419 - I426; DII), blue (V437 - I454; DII/III) and orange (D483 - F505; DIII).

Discussion

This study has examined structured protein domains from *P. falciparum* and the importance of structural features in assessing potential targets of humoral immunity. Antigenic polymorphism data from a Gambian population was used to identify regions under immune selection pressure, with polymorphic residues found to be primarily surface exposed and enriched within turns as compared to other secondary structure elements. A number of studies have observed that polymorphism variation in buried residues is generally more detrimental to protein function than polymorphic variation in surface exposed regions^{76–78}. This observation, coupled with antibody-driven selection pressure on surface-exposed residues, likely explains our finding that high-frequency polymorphic residues were highly surface exposed. Similarly, our observation that polymorphic residues were proportionally enriched within turns is supported by previous work examining the structural features of antigen-antibody interfaces^{79–81}. These studies used a three-state definition of secondary structure (helices, sheets, loops), and observed

an increased proportion of loop elements (turns, bends or coils) within epitopes; we only observed an increased number of polymorphisms within turns.

Prediction of B-cell epitopes is a challenging problem, with early attempts based on amino acid residue propensity scales only a little better than chance⁸². Despite many advances in recent years, the performance of current B-cell epitope prediction algorithms still falls behind that of predictors of other immunological features, such as MHC and T-cell epitope predictors⁸³. With this in mind, the results obtained in our study are encouraging. The sequence-based BepiPred 2.0 algorithm⁴¹ predominantly identifies surface exposed regions as potential epitopes, despite only using the linear protein sequence as an input. Additionally, when restricting our analysis to only surface exposed residues, ~75% of polymorphic residues were placed within predicted epitopes compared to ~50% of non-polymorphic residues (at the default threshold). Thus, we suggest that BepiPred 2.0 is an informative tool for initial selection of potential epitopes, particularly in the absence of a known protein structure.

Another approach to identifying potential epitopes explored in this study was the identification of polymorphisms clustered within a potential antibody binding radius. All of the proteins identified using this approach are known highly polymorphic antigens, including AMA1, PfEMP1, EBA-175, TRAP, DBLMS2, CSP and MSP1^{84–89}. Additionally, our relatively conservative homology modelling approach did not yield additional proteins of interest when examining polymorphic hotspots, despite a 3-fold increase in the coverage of the proteome. Future work could extend the approach used in this study using structural models derived from other methods such as threading and fold recognition.

Other methods for determining likely immune targets within a pathogen include measures of balancing selection such as Tajima's D. Typically, Tajima's D is applied as either a single metric over a whole gene, or as a sliding window over the genome. A number of studies of malaria antigens have also applied Tajima's D as a sliding window over particular genes to identify regions of the protein under immune-mediated selection pressure^{9,10,23–26}. However, for antibody-mediated selection pressure, a conformational epitope may contain residues that are distant in the linear protein sequence. Thus, we hypothesized that incorporation of structural information into a sliding window calculation of Tajima's D may improve detection of regions under immune selection pressure. We developed a Python tool (BioStructMap) for spatial averaging, and applied it to both PfAMA1 and EBA-175. While application of a spatially derived Tajima's D to EBA-175 did not yield any major differences compared to a standard linear sliding window method, we revealed a region in PfAMA1 bordering DII and DIII with a high spatially derived Tajima's D value that was not observed using a traditional sliding window (Fig. 6). Interestingly, this region is composed of four distinct segments of protein that combine to form a surface-exposed face, which explains why a linear sliding window method failed to identify this discontinuous epitope (Fig. 7).

Previous studies also suggest that DIII may be an additional target of humoral immune responses. In a comparison of Tajima's D between PfAMA1 and PvAMA1, the highest Tajima's D values for PfAMA1 were observed in DIII, in contrast to DI for PvAMA1¹⁰, suggesting that DIII may play a significant role as a target for protective immune responses against *P. falciparum*. It has also been observed that a monoclonal antibody (1E10) against PfAMA1 DIII acts synergistically with antibodies against other distant parts of the protein to inhibit merozoite growth, despite not having potent inhibitory capabilities on its own⁶⁷. This suggests a potential role for antibodies targeting DIII in the context of a broader response against PfAMA1, despite an anti-DIII response not being inhibitory in isolation. Population studies examining antibody levels against PfAMA1 domains suggest that antibody responses against both DII and DIII are relatively rare⁹⁰, however the recombinant protein constructs used in these studies fail to account for the considerable interaction between domains. Indeed, the region of PfAMA1 with the strongest signature of balancing selection in our study was composed of residues from both DII and DIII, and antibodies targeting this epitope may not be identified in assays that use recombinant DII or DIII constructs. Taken together, these studies support our observation that a region bordering DII and DIII of PfAMA1 may be an important target of humoral responses in the context of natural infection, as well as supporting the validity of our spatially derived Tajima's D approach.

In this study, we have developed an approach to identify structured regions of *P. falciparum* proteins that are likely under some level of immune selection pressure, showing that polymorphic sites are predominantly surface exposed and enriched within turns. We applied a spatially derived Tajima's D calculation to key antigens, identifying a region of PfAMA1 between DII and DIII that was under a high degree of balancing selection. These methods and accompanying results have utility in the identification of proteins under balancing selection, furthering our understanding of functional immune targets during malaria infection. These approaches are also broadly applicable to other pathogens.

References

1. World Health Organization. *World Malaria Report 2016*. (2016).
2. Dai, G., Carmicle, S., Steede, N. K. & Landry, S. J. Structural basis for helper T-cell and antibody epitope immunodominance in bacteriophage T4 Hsp10. Role of disordered loops. *J. Biol. Chem.* **277**, 161–168 (2002).
3. Dunker, A. K. *et al.* Intrinsically disordered protein. *J. Mol. Graph. Model.* **19**, 26–59 (2001).
4. Oldfield, C. J. & Dunker, A. K. Intrinsically disordered proteins and intrinsically disordered protein regions. *Annu. Rev. Biochem.* **83**, 553–584 (2014).
5. Tompa, P. Intrinsically disordered proteins: a 10-year recap. *Trends Biochem. Sci.* **37**, 509–516 (2012).
6. Guy, A. J. *et al.* Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *PLoS One* **10**, e0141729 (2015).
7. Cohen, S., McGregor, I. A. & Carrington, S. Gamma-Globulin and Acquired Immunity to Human Malaria. *Nature* **192**, 733–737 (1961).
8. Crompton, P. D. *et al.* Malaria immunity in man and mosquito: insights into unsolved mysteries of a deadly infectious disease. *Annu. Rev. Immunol.* **32**, 157–187 (2014).
9. Arnott, A. *et al.* Global Population Structure of the Genes Encoding the Malaria Vaccine Candidate, Plasmodium vivax Apical Membrane Antigen 1 (Pv AMA1). *PLoS Negl. Trop. Dis.* **7**, e2506 (2013).

10. Arnott, A. *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.* **13**, 233 (2014).
11. Vulliez-Le Normand, B. *et al.* Crystal structure of *Plasmodium knowlesi* apical membrane antigen 1 and its complex with an invasion-inhibitory monoclonal antibody. *PLoS One* **10**, e0123567 (2015).
12. Chesne-Seck, M.-L. *et al.* Structural comparison of apical membrane antigen 1 orthologues and paralogues in apicomplexan parasites. *Mol. Biochem. Parasitol.* **144**, 55–67 (2005).
13. Sedegah, M. *et al.* Identification and localization of minimal MHC-restricted CD8+ T cell epitopes within the *Plasmodium falciparum* AMA1 protein. *Malar. J.* **9**, 241 (2010).
14. Doud, M. B. *et al.* Unexpected fold in the circumsporozoite protein target of malaria vaccines. *Proc. Natl. Acad. Sci. USA* **109**, 7817–7822 (2012).
15. Prato, S., Fleming, J., Schmidt, C. W., Corradin, G. & Lopez, J. A. Cross-presentation of a human malaria CTL epitope is conformation dependent. *Mol. Immunol.* **43**, 2031–2036 (2006).
16. Aragam, N. R. *et al.* Diversity of T cell epitopes in *Plasmodium falciparum* circumsporozoite protein likely due to protein-protein interactions. *PLoS One* **8**, e62427 (2013).
17. Ambroggio, X. *et al.* 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* **8**, e56326 (2013).
18. Hodder, A. N. *et al.* Insights into Duffy binding-like domains through the crystal structure and function of the merozoite surface protein MSPDBL2 from *Plasmodium falciparum*. *J. Biol. Chem.* **287**, 32922–32939 (2012).
19. Morales, R. A. V. *et al.* Structural basis for epitope masking and strain specificity of a conserved epitope in an intrinsically disordered malaria vaccine candidate. *Sci. Rep.* **5**, 10103 (2015).
20. Neafsey, D. E. *et al.* Genetic Diversity and Protective Efficacy of the RTS,S/AS01 Malaria Vaccine. *N. Engl. J. Med.* **373**, 2025–2037 (2015).
21. Doolan, D. L., Houghten, R. A. & Good, M. F. Location of human cytotoxic T cell epitopes within a polymorphic domain of the *Plasmodium falciparum* circumsporozoite protein. *Int. Immunol.* **3**, 511–516 (1991).
22. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595 (1989).
23. Parobek, C. M. *et al.* Differing Patterns of Selection and Geospatial Genetic Diversity within Two Leading *Plasmodium vivax* Candidate Vaccine Antigens. *PLoS Negl. Trop. Dis.* **8**, e2796 (2014).
24. Amambua-Ngwa, A. *et al.* Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet.* **8**, e1002992 (2012).
25. Reeder, J. C., Wapling, J., Mueller, I., Siba, P. M. & Barry, A. E. Population genetic analysis of the *Plasmodium falciparum* 6-cys protein Pf38 in Papua New Guinea reveals domain-specific balancing selection. *Malar. J.* **10**, 126 (2011).
26. Osier, F. H. A. *et al.* Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. *Infect. Immun.* **78**, 4625–4633 (2010).
27. Aurrecochea, C. *et al.* PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res.* **37**, D539–43 (2009).
28. Berman, H. M. *et al.* The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242 (2000).
29. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
30. Hamelryck, T. & Manderick, B. PDB file parser and structure class implemented in Python. *Bioinformatics* **19**, 2308–2310 (2003).
31. Sukumaran, J. & Holder, M. T. DendroPy: a Python library for phylogenetic computing. *Bioinformatics* **26**, 1569–1571 (2010).
32. Eswar, N. *et al.* Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **31**, 3375–3380 (2003).
33. Sali, A. & Blundell, T. L. Comparative protein modelling by satisfaction of spatial restraints. *J. Mol. Biol.* **234**, 779–815 (1993).
34. Pieper, U. *et al.* ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res.* **39**, D465–74 (2011).
35. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
36. Rost, B. & Sander, C. Conservation and prediction of solvent accessibility in protein families. *Proteins* **20**, 216–226 (1994).
37. Chen, H. & Zhou, H.-X. Prediction of solvent accessibility and sites of deleterious mutations from protein sequence. *Nucleic Acids Res.* **33**, 3193–3199 (2005).
38. Hopp, T. P. & Woods, K. R. Prediction of protein antigenic determinants from amino acid sequences. *Proc. Natl. Acad. Sci. USA* **78**, 3824–3828 (1981).
39. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
40. Larsen, J. E. P., Lund, O. & Nielsen, M. Improved method for predicting linear B-cell epitopes. *Immunome Res.* **2**, 2 (2006).
41. Jespersen, M. C., Peters, B., Nielsen, M. & Marcatili, P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* <https://doi.org/10.1093/nar/gkx346> (2017).
42. Nielsen, M. & Lund, O. NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction. *BMC Bioinformatics* **10**, 296 (2009).
43. Nielsen, M., Lundegaard, C. & Lund, O. Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method. *BMC Bioinformatics* **8**, 238 (2007).
44. Kasahara, M. *et al.* HLA-DQ haplotypes in 15 different populations. https://doi.org/10.1007/978-4-431-65868-9_31 (2000).
45. González-Galarza, F. F. *et al.* Allele frequency net 2015 update: new features for HLA epitopes, KIR and disease and HLA adverse drug reaction associations. *Nucleic Acids Res.* **43**, D784–8 (2015).
46. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* **305**, 567–580 (2001).
47. Gilson, P. R. *et al.* Identification and stoichiometry of glycosylphosphatidylinositol-anchored membrane proteins of the human malaria parasite *Plasmodium falciparum*. *Mol. Cell. Proteomics* **5**, 1286–1299 (2006).
48. Pierleoni, A., Martelli, P. L. & Casadio, R. PredGPI: a GPI-anchor predictor. *BMC Bioinformatics* **9**, 392 (2008).
49. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8 (2015).
50. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science Engineering* **9**, 90–95 (2007).
51. der Walt, S., van, Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science Engineering* **13**, 22–30 (2011).
52. Ramaraj, T., Angel, T., Dratz, E. A., Jesaitis, A. J. & Mumey, B. Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta* **1824**, 520–532 (2012).
53. Davies, D. R., Padlan, E. A. & Sheriff, S. Antibody-antigen complexes. *Annu. Rev. Biochem.* **59**, 439–473 (1990).
54. Coley, A. M. *et al.* The most polymorphic residue on *Plasmodium falciparum* apical membrane antigen 1 determines binding of an invasion-inhibitory antibody. *Infect. Immun.* **74**, 2628–2636 (2006).
55. Ochola-Oyier, L. I. *et al.* Comparison of allele frequencies of *Plasmodium falciparum* merozoite antigens in malaria infections sampled in different years in a Kenyan population. *Malar. J.* **15**, 261 (2016).
56. Samad, H. *et al.* Imputation-based population genetics analysis of *Plasmodium falciparum* malaria parasites. *PLoS Genet.* **11**, e1005131 (2015).
57. Mobegi, V. A. *et al.* Genome-Wide Analysis of Selection on the Malaria Parasite *Plasmodium falciparum* in West African Populations of Differing Infection Endemicity. *Mol. Biol. Evol.* <https://doi.org/10.1093/molbev/msu106> (2014).

58. Ayodo, G. *et al.* Combining evidence of natural selection with association analysis increases power to detect malaria-resistance variants. *Am. J. Hum. Genet.* **81**, 234–242 (2007).
59. Saha, S. & Raghava, G. P. S. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* **65**, 40–48 (2006).
60. Odorico, M. & Pellequer, J.-L. BEPITOPE: predicting the location of continuous epitopes and patterns in proteins. *J. Mol. Recognit.* **16**, 20–22 (2003).
61. Alix, A. J. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine* **18**, 311–314 (1999).
62. Parker, J. M., Guo, D. & Hodges, R. S. New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* **25**, 5425–5432 (1986).
63. El-Manzalawy, Y. & Honavar, V. Recent advances in B-cell epitope prediction methods. *Immunome Res.* **6**(Suppl 2), S2 (2010).
64. Yao, B., Zheng, D., Liang, S. & Zhang, C. Conformational B-Cell Epitope Prediction on Antigen Protein Structures: A Review of Current Algorithms and Comparison with Common Binding Site Prediction Methods. *PLoS One* **8**, e62249 (2013).
65. Good, M. F. & Doolan, D. L. Malaria Vaccine Design: Immunological Considerations. *Immunity* **33**, 555–566 (2010).
66. Deroost, K., Pham, T.-T., Opdenakker, G. & Van den Steen, P. E. The immunological balance between host and parasite in malaria. *FEMS Microbiol. Rev.* **40**, 208–257 (2016).
67. Dutta, S. *et al.* Overcoming antigenic diversity by enhancing the immunogenicity of conserved epitopes on the malaria vaccine candidate apical membrane antigen-1. *PLoS Pathog.* **9**, e1003840 (2013).
68. Tolia, N. H., Enemark, E. J., Sim, B. K. L. & Joshua-Tor, L. Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell* **122**, 183–193 (2005).
69. Dutta, S., Lee, S. Y., Batchelor, A. H. & Lanar, D. E. Structural basis of antigenic escape of a malaria vaccine candidate. *Proc. Natl. Acad. Sci. USA* **104**, 12488–12493 (2007).
70. Harris, K. S. *et al.* Binding hot spot for invasion inhibitory molecules on *Plasmodium falciparum* apical membrane antigen 1. *Infect. Immun.* **73**, 6981–6989 (2005).
71. Chen, E., Paing, M. M., Salinas, N., Kim Lee Sim, B. & Tolia, N. H. Structural and Functional Basis for Inhibition of Erythrocyte Invasion by Antibodies that Target *Plasmodium falciparum* EBA-175. *PLoS Pathog.* **9**, e1003390 (2013).
72. Irani, V. *et al.* Acquisition of Functional Antibodies That Block the Binding of Erythrocyte-Binding Antigen 175 and Protection Against *Plasmodium falciparum* Malaria in Children. *Clin. Infect. Dis.* **61**, 1244–1252 (2015).
73. Drew, D. R. *et al.* A novel approach to identifying patterns of human invasion-inhibitory antibodies guides the design of malaria vaccines incorporating polymorphic antigens. *BMC Med.* **14**, 144 (2016).
74. Terheggen, U. *et al.* Limited antigenic diversity of *Plasmodium falciparum* apical membrane antigen 1 supports the development of effective multi-allele vaccines. *BMC Med.* **12**, 183 (2014).
75. Drew, D. R. *et al.* Defining the Antigenic Diversity of *Plasmodium falciparum* Apical Membrane Antigen 1 and the Requirements for a Multi-Allele Vaccine against Malaria. *PLoS One* **7**, e51023 (2012).
76. Chasman, D. & Adams, R. M. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* **307**, 683–706 (2001).
77. Saunders, C. T. & Baker, D. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. *J. Mol. Biol.* **322**, 891–901 (2002).
78. Yue, P., Li, Z. & Moult, J. Loss of protein structure stability as a major causative factor in monogenic disease. *J. Mol. Biol.* **353**, 459–473 (2005).
79. Dalkas, G. A., Teheux, F., Kwasigroch, J. M. & Rooman, M. Cation- π , amino- π , π - π , and H-bond interactions stabilize antigen-antibody interfaces. *Proteins* **82**, 1734–1746 (2014).
80. Rubinstein, N. D. *et al.* Computational characterization of B-cell epitopes. *Mol. Immunol.* **45**, 3477–3489 (2008).
81. Pellequer, J. L., Westhof, E. & Van Regenmortel, M. H. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol. Lett.* **36**, 83–99 (1993).
82. Blythe, M. J. & Flower, D. R. Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* **14**, 246–248 (2005).
83. Nielsen, M. & Marcatili, P. Prediction of Antibody Epitopes. *Methods Mol. Biol.* **1348**, 23–32 (2015).
84. Bai, T. *et al.* Structure of AMA1 from *Plasmodium falciparum* reveals a clustering of polymorphisms that surround a conserved hydrophobic pocket. *Proc. Natl. Acad. Sci. USA* **102**, 12736–12741 (2005).
85. Newbold, C. I. *et al.* PfEMP1, polymorphism and pathogenesis. *Ann. Trop. Med. Parasitol.* **91**, 551–557 (1997).
86. Mayer, D. C. G., Mu, J.-B., Feng, X., Su, X.-Z. & Miller, L. H. Polymorphism in a *Plasmodium falciparum* erythrocyte-binding ligand changes its receptor specificity. *J. Exp. Med.* **196**, 1523–1528 (2002).
87. Ohashi, J., Suzuki, Y., Naka, I., Hananantachai, H. & Patarapotikul, J. Diversifying Selection on the Thrombospondin-Related Adhesive Protein (TRAP) Gene of *Plasmodium falciparum* in Thailand. *PLoS One* **9**, e90522 (2014).
88. Crosnier, C. *et al.* Binding of *Plasmodium falciparum* Merozoite Surface Proteins DBLMSP and DBLMSP2 to Human Immunoglobulin M Is Conserved among Broadly Diverged Sequence Variants. *J. Biol. Chem.* **291**, 14285–14299 (2016).
89. Escalante, A. A., Lal, A. A. & Ayala, F. J. Genetic polymorphism and natural selection in the malaria parasite *Plasmodium falciparum*. *Genetics* **149**, 189–202 (1998).
90. Cortés, A. *et al.* Allele specificity of naturally acquired antibody responses against *Plasmodium falciparum* apical membrane antigen 1. *Infect. Immun.* **73**, 422–430 (2005).

Acknowledgements

Funding was provided by the National Health and Medical Research Council (NHMRC) of Australia including Fellowships to JSR (APP1037722) and JGB (APP1077636), Project (APP1125788) and Program (APP1092789) grants. Support was provided through Monash University (Australian Postgraduate Award to AJG) and the University of Melbourne (Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship to VI). The Burnet Institute is supported by the NHMRC Independent Research Institutes Infrastructure Support Scheme, and a Victoria State Government Operational Infrastructure Support grant. Funding was also provided by National Institutes of Health grants P41 GM109824 and R01 GM083960.

Author Contributions

A.J.G., J.S.R. and P.A.R. designed the research; A.J.G. performed the research; A.J.G., V.I., J.S.R., and P.A.R. discussed and interpreted the data; B.W., and A.S. generated and contributed the template-based models; A.J.G. wrote the manuscript; All authors revised, commented and read the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-22592-3>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018

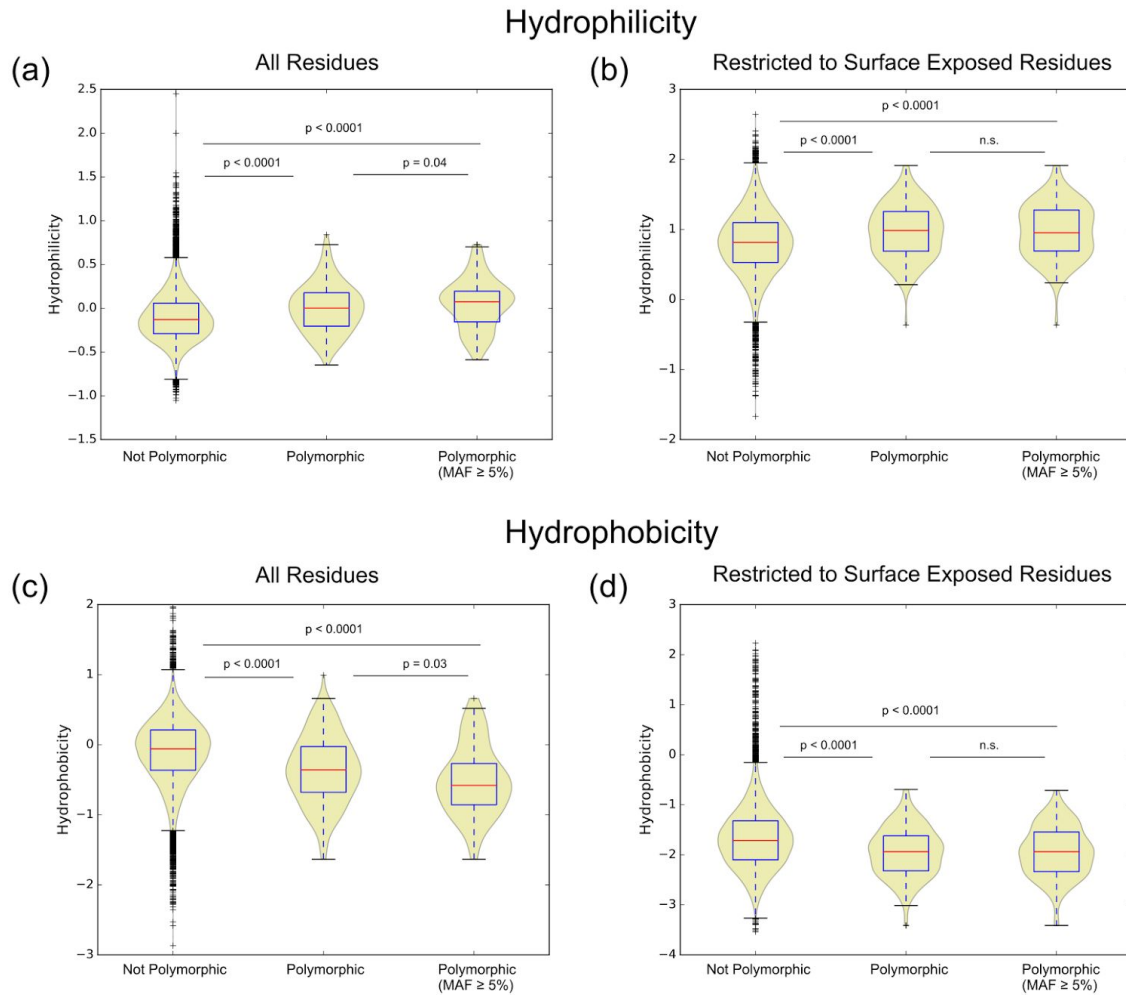


Figure S1: Average hydrophilicity and hydrophobicity are only weakly associated with residue polymorphisms once surface exposure is taken into account. A radius of 15 Å was used for spatial averaging of hydrophilicity and hydrophobicity values, and polymorphic residues identified by the presence of underlying non-synonymous SNPs, with additional filter using a minor allele frequency threshold of 5% also applied. **a)** Average hydrophilicity calculated over a 15 Å radius, grouped by residue polymorphism. **b)** Average hydrophilicity calculated over a 15 Å radius, with calculations restricted to surface exposed residues (RSA > 0.2) and grouped by residue polymorphism. **c)** Average hydrophobicity calculated over a 15 Å radius, grouped by residue polymorphism. **d)** Average hydrophobicity calculated over a 15 Å radius, with calculations restricted to surface exposed residues (RSA > 0.2) and grouped by residue polymorphism. Mann-Whitney U test used for comparison between groups.

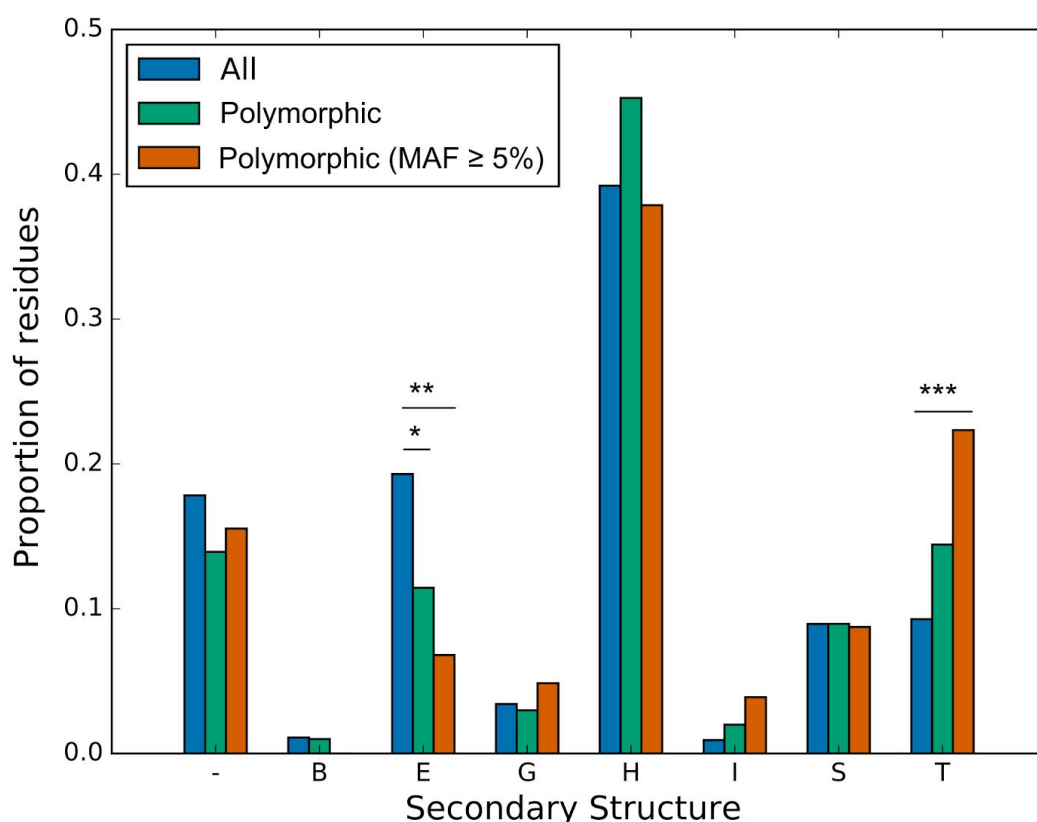


Figure S2: Polymorphic residues are more common within secondary structure turn elements and reduced within β -strand elements. Secondary structure elements for polymorphic residues were compared to the background set of protein structures from *P. falciparum*. When considering all polymorphic residues (without a MAF cutoff), a decreased proportion of polymorphic residues within β -strand (E) elements was observed ($p = 0.03$ after Bonferroni correction for multiple comparison; binomial test). When restricting polymorphic residues to those with a minor allele frequency (MAF) $\geq 5\%$, an increased proportion of polymorphic residues were contained in turn (T) elements ($p = 0.005$ after Bonferroni correction for multiple comparison; binomial test), and a reduced number of polymorphic residues found within within β -strand (E) elements ($p < 0.001$ after Bonferroni correction for multiple comparison; binomial test). Polymorphic residues with underlying non-synonymous SNPs with a minor allele frequency (MAF) $\geq 5\%$ are shown in orange, whereas all polymorphic residues (no MAF threshold) are shown in green. Secondary structure was classified using the DSSP program. Secondary structure assignments are coded as follows: H, Alpha helix; B, Beta bridge; E, Strand; G, 3-Turn Helix; I, 5-Turn Helix; T, Turn; S, Bend; -, Other/Coil.

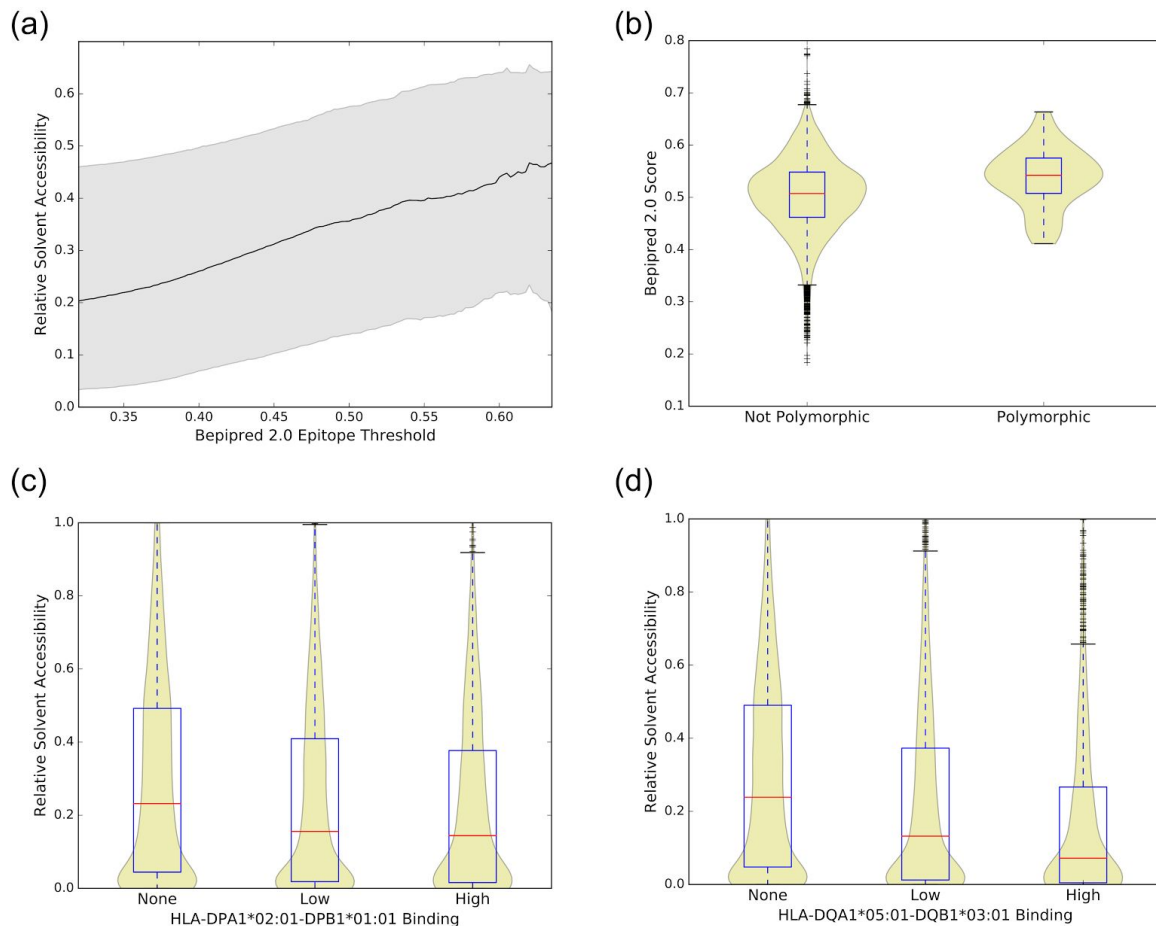


Figure S3: Predicted B-cell epitopes by Bepipred 2.0 strongly favor surface exposed residues, while predicted MHC II binding peptide are generally not surface exposed. a) Median relative solvent accessibility values were plotted for residues with an epitope score above any given threshold score for Bepipred 2.0. The location of the upper and lower quartiles are also shown in grey. We have shown results for thresholds between the 5th and 95th percentile of all scores for Bepipred 2.0. b) The distribution of Bepipred 2.0 scores was compared between polymorphic and non-polymorphic ($MAF \geq 5\%$) residues, with analysis restricted to surface exposed residues (relative solvent accessibility ≥ 0.2). c, d) The location of MHC II binding peptides were assessed in relation to surface exposure, with the majority of residues involved in an MHC II binding peptide having relatively little surface exposure. Haplotypes HLA-DPA1*02:01-DPB1*01:01 (c) and HLA-DQA1*05:01-DQB1*03:01 (d) were assessed here, as these haplotypes have been shown to be common within a Gambian population and are also available for prediction using the NetMHCII tool. Prediction of peptide binding was performed using NetMHCII 2.2, with a 15 aa peptide length and default settings.

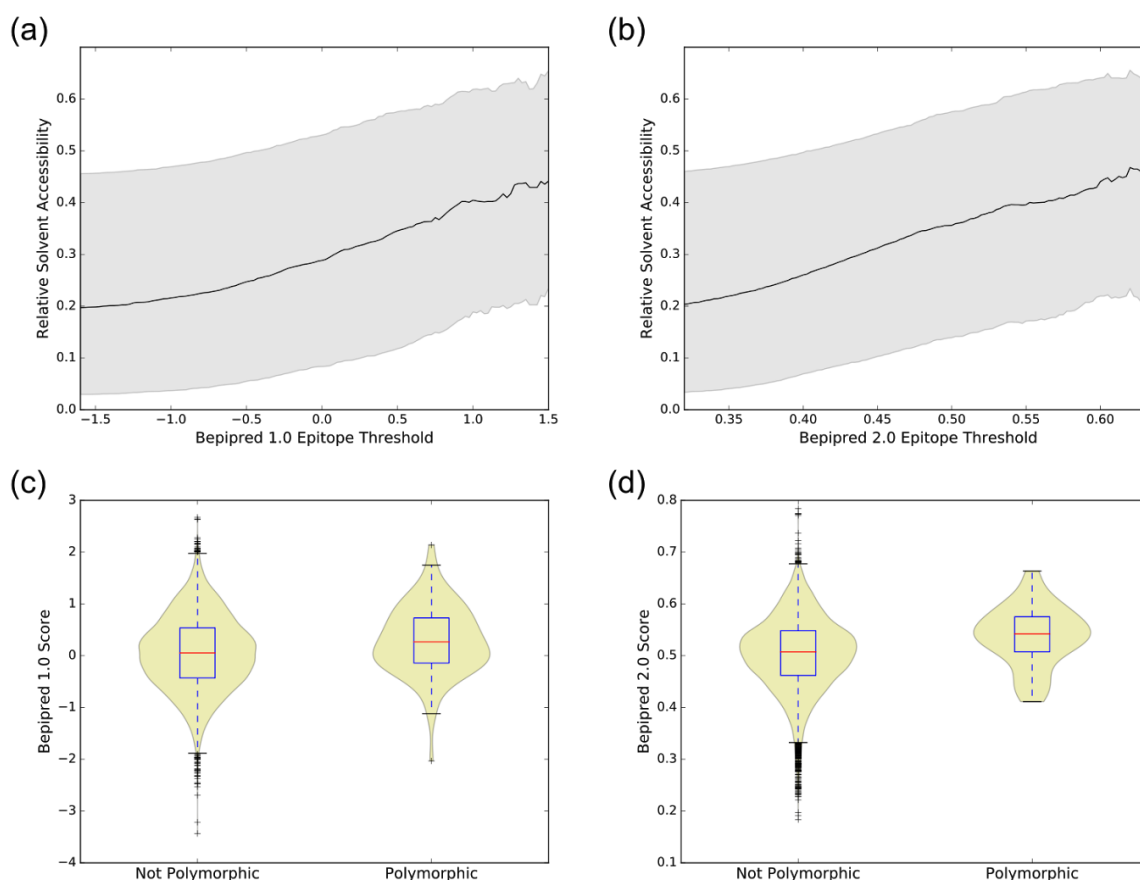


Figure S4: Predicted B-cell epitopes by both Bepipred 1.0 and Bepipred 2.0 both strongly favour surface exposed residues. Bepipred 2.0 performs better than Bepipred 1.0 in assigning polymorphic residues a high epitope probability. **a, b)** Median relative solvent accessibility values were plotted for residues with an epitope score above any given threshold score for Bepipred 1.0 **(a)** and Bepipred 2.0 **(b)**. The location of the upper and lower quartiles are also shown in grey. Note that the epitope scores used by Bepipred 1.0 and Bepipred 2.0 are not directly relatable, and hence we have shown results for thresholds between the 5th and 95th percentile of all scores for each predictor. Both Bepipred 1.0 and Bepipred 2.0 predicted epitopes were predominantly surface exposed, with increasing Bepipred thresholds predicting increasingly surface exposed residues. **c, d)** The distribution of Bepipred 1.0 **(c)** and Bepipred 2.0 **(d)** scores were compared between polymorphic and non-polymorphic (MAF $\geq 5\%$) residues, with analysis restricted to surface exposed residues (relative solvent accessibility ≥ 0.2). When considering the epitopes scores given to polymorphic residues, both Bepipred 1.0 and 2.0 scores were significantly higher for polymorphic residues (Bepipred 1.0, $p = 0.007$; Bepipred 2.0, $p < 0.0001$; Mann-Whitney U test), even when restricting analysis to surface exposed residues (relative solvent accessibility ≥ 0.2). It is noted that Bepipred 2.0 scores for polymorphic residues had the greater shift relative to the background distribution of epitope scores, with the median Bepipred 2.0 score for polymorphic residues shifted higher by 0.40 of the interquartile range (IQR) of the background distribution, compared to 0.22 of the IQR for Bepipred 1.0.

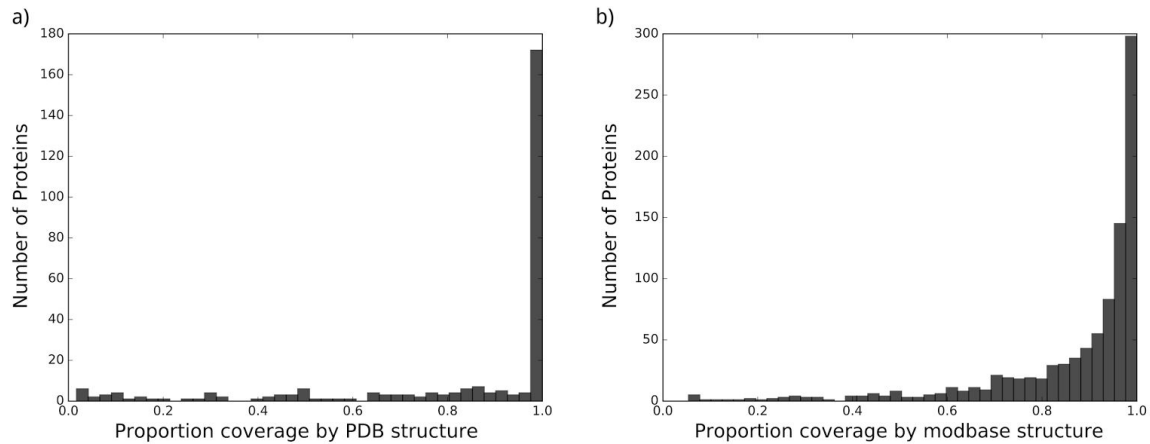


Figure S5: Most *P. falciparum* PDB structures and Modpipe models cover the majority of the reference protein sequence. (a) Proportion of *P. falciparum* reference protein sequence that is covered by a matching PDB structure (sequence similarity > 90%). A total of 275 *P. falciparum* proteins have at least one matching PDB structure. Proteins with no matching structure are not represented in this histogram. (b) Proportion of *P. falciparum* reference protein sequence that is covered by a high-quality ModPipe model (MPQS > 1.1). A total of 923 *P. falciparum* proteins have at least one high-quality ModPipe model. Proteins with no high-quality ModPipe model are not represented in this histogram.

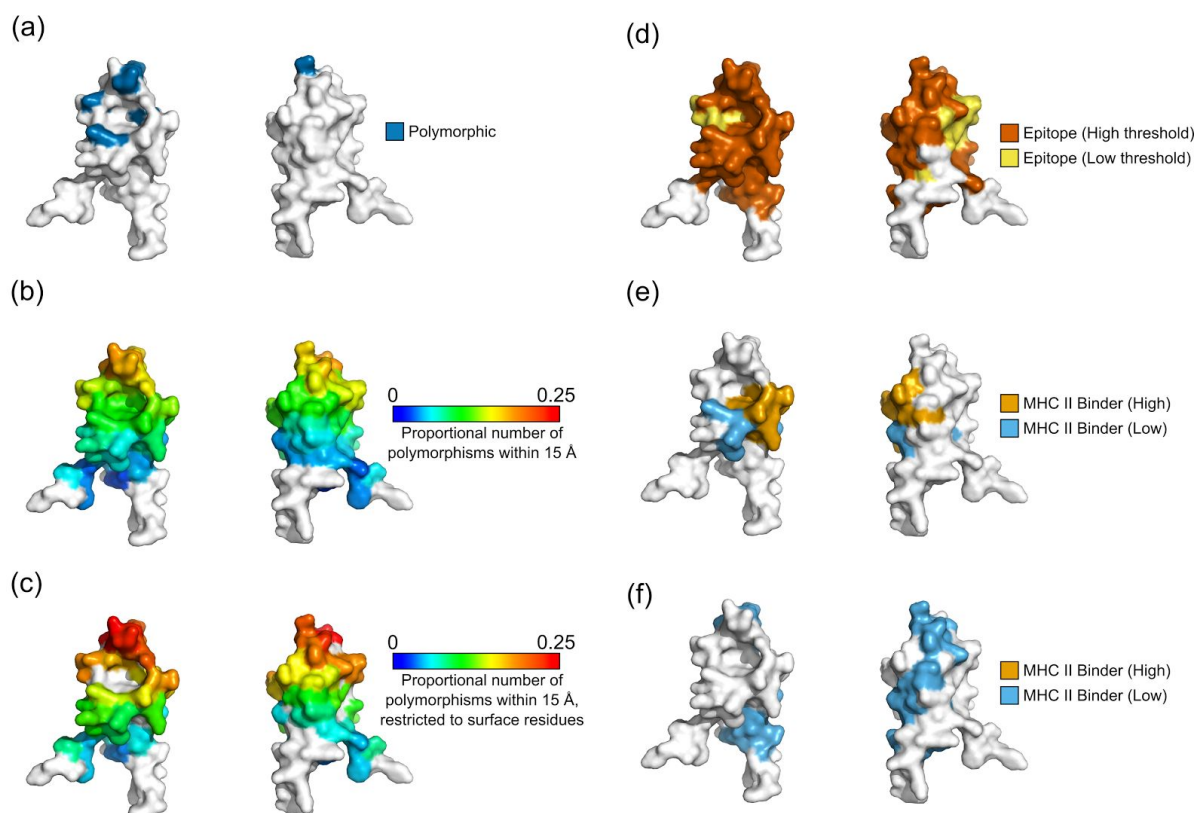


Figure S6: Location of immunologically relevant features mapped onto a CSP structure (PDB code: 3VDJ). Each panel shows the front and back and top view of the CSP structure. **a)** Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown in blue. Sequence polymorphisms were obtained from 65 Gambian isolates¹. **b)** Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. **c)** Spatial averaging of polymorphic residues, restricted to surface exposed residues. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$ and surface exposed residues considered to be those with RSA ≥ 0.2 . **d)** Bepipred 2.0 predictions, with epitopes shown for two Bepipred thresholds — predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). **e, f)** The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (e) and HLA-DQA1*05:01-DQB1*03:01 (f) alleles. Residues involved in a low binding peptide ($50 \text{ nM} < \text{IC}_{50} < 500 \text{ nM}$) are shown in light blue, while residues involved in a high binding peptide ($\text{IC}_{50} < 50 \text{ nM}$) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

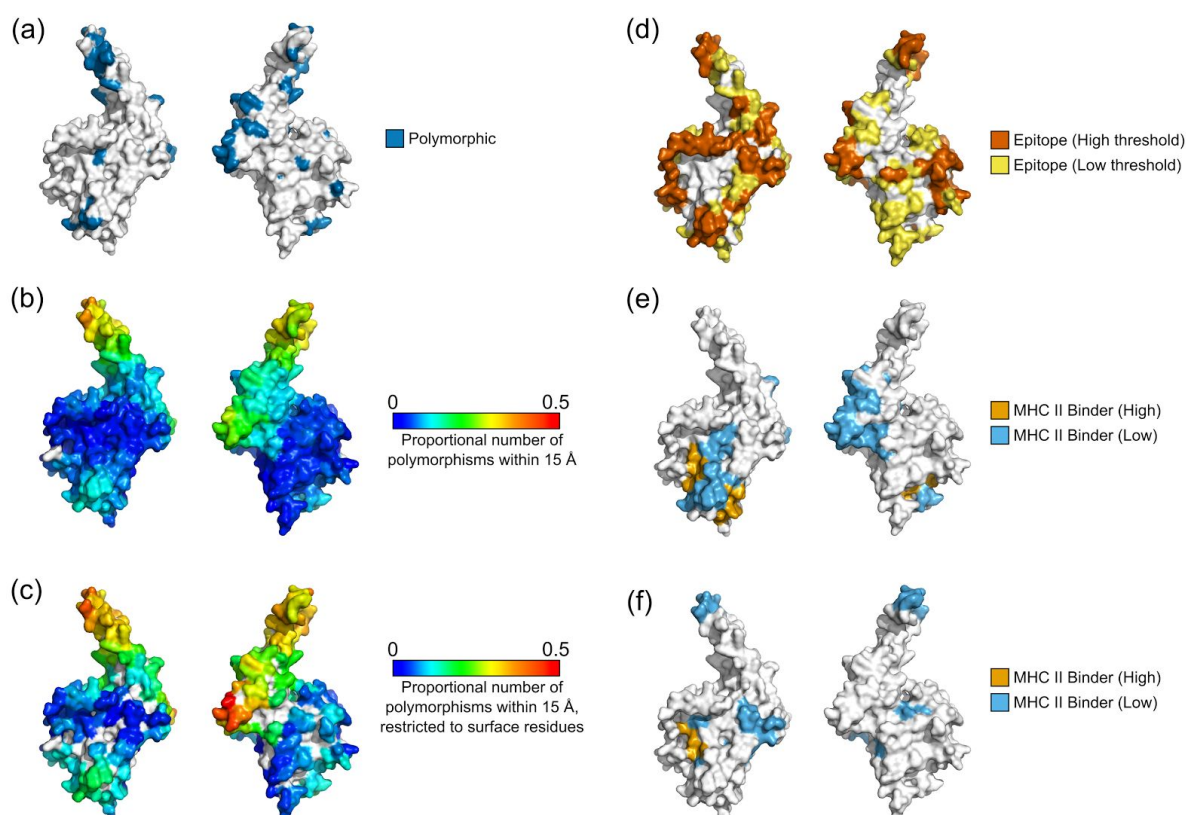


Figure S7: Location of immunologically relevant features mapped onto an MSPDBL2 structural model (*ModPipe ID: 45d80d0181671431333fb05b6011a7e4*). Each panel shows the front and back and top view of the MSPDBL2 structure. **a)** Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown in blue. Sequence polymorphisms were obtained from 65 Gambian isolates¹. **b)** Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. **c)** Spatial averaging of polymorphic residues, restricted to surface exposed residues. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$ and surface exposed residues considered to be those with RSA ≥ 0.2 . **d)** Bepipred 2.0 predictions, with epitopes shown for two Bepipred thresholds — predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). **e, f)** The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (**e**) and HLA-DQA1*05:01-DQB1*03:01 (**f**) alleles. Residues involved in a low binding peptide (50 nM $<$ IC₅₀ $<$ 500 nM) are shown in light blue, while residues involved in a high binding peptide (IC₅₀ $<$ 50 nM) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

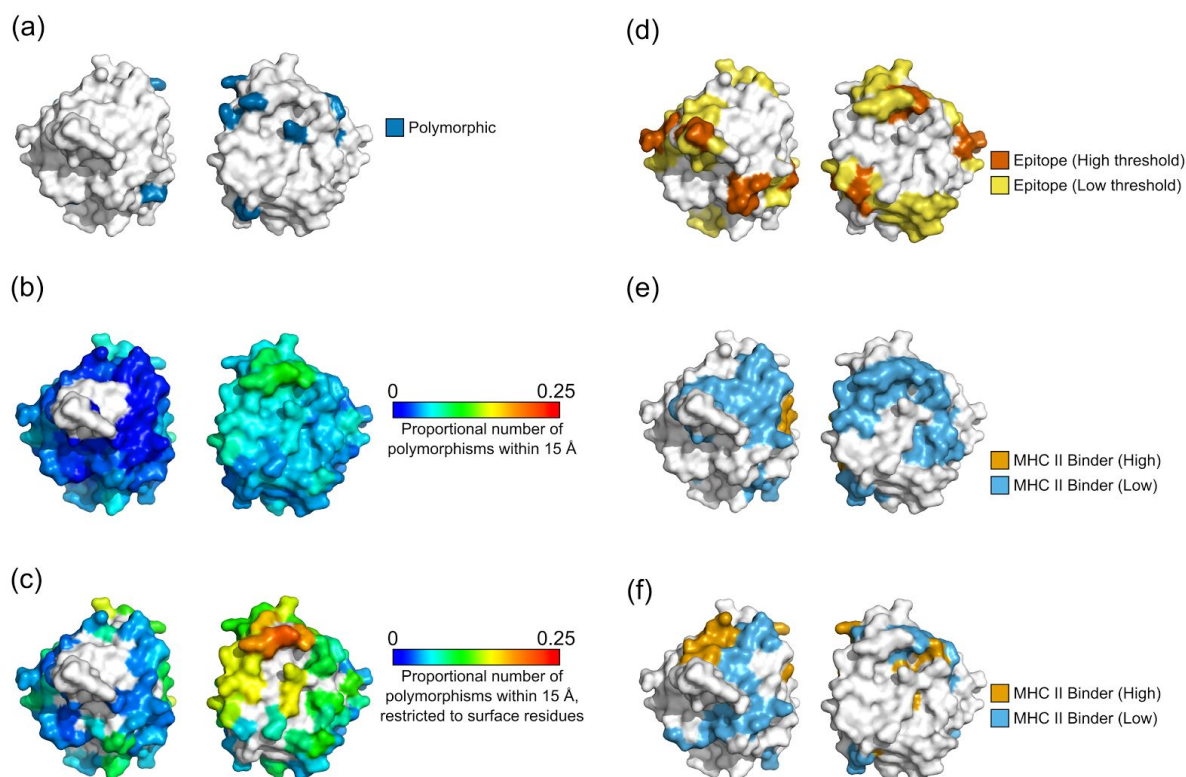


Figure S8: Location of immunologically relevant features mapped onto a TRAP structure (PDB code: 4F1J). Each panel shows the front and back and top view of the TRAP structure. **a)** Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown in blue. Sequence polymorphisms were obtained from 65 Gambian isolates¹. **b)** Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. **c)** Spatial averaging of polymorphic residues, restricted to surface exposed residues. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$ and surface exposed residues considered to be those with RSA ≥ 0.2 . **d)** Bepipred 2.0 predictions, with epitopes shown for two Bepipred thresholds — predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). **e, f)** The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (**e**) and HLA-DQA1*05:01-DQB1*03:01 (**f**) alleles. Residues involved in a low binding peptide ($50 \text{ nM} < \text{IC}_{50} < 500 \text{ nM}$) are shown in light blue, while residues involved in a high binding peptide ($\text{IC}_{50} < 50 \text{ nM}$) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

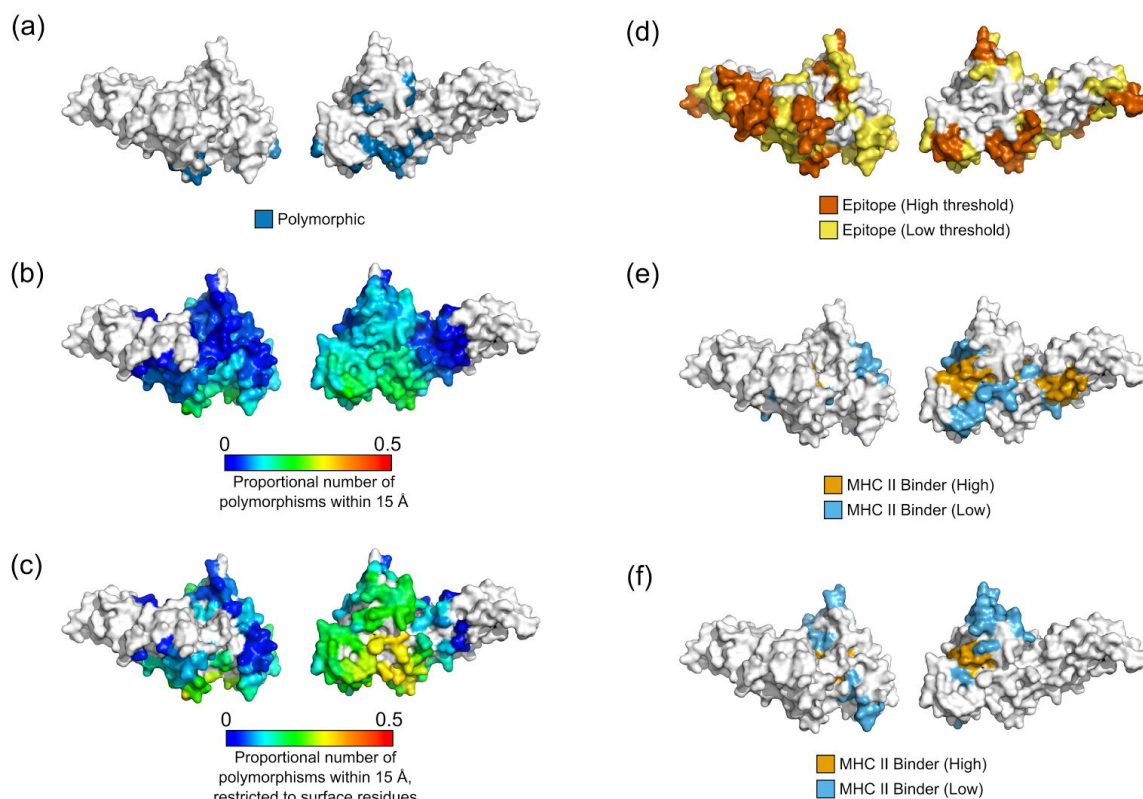


Figure S9: Location of immunologically relevant features mapped onto an PfEMP1 DBL6 structural model (*ModPipe ID: 5f1239b7c2f20f9a5455c9a43f0fa2d7*). Each panel shows the front and back and top view of the PfEMP1 DBL6 structure. **a)** Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown in blue. Sequence polymorphisms were obtained from 65 Gambian isolates¹. **b)** Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. **c)** Spatial averaging of polymorphic residues, restricted to surface exposed residues. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$ and surface exposed residues considered to be those with RSA ≥ 0.2 . **d)** Bepipred 2.0 predictions, with epitopes shown for two Bepipred thresholds — predicted epitopes are shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). **e, f)** The location of predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (**e**) and HLA-DQA1*05:01-DQB1*03:01 (**f**) alleles. Residues involved in a low binding peptide ($50 \text{ nM} < \text{IC}_{50} < 500 \text{ nM}$) are shown in light blue, while residues involved in a high binding peptide ($\text{IC}_{50} < 50 \text{ nM}$) residue are shown in orange. Only the core binding region of each peptide binder is indicated on each structure.

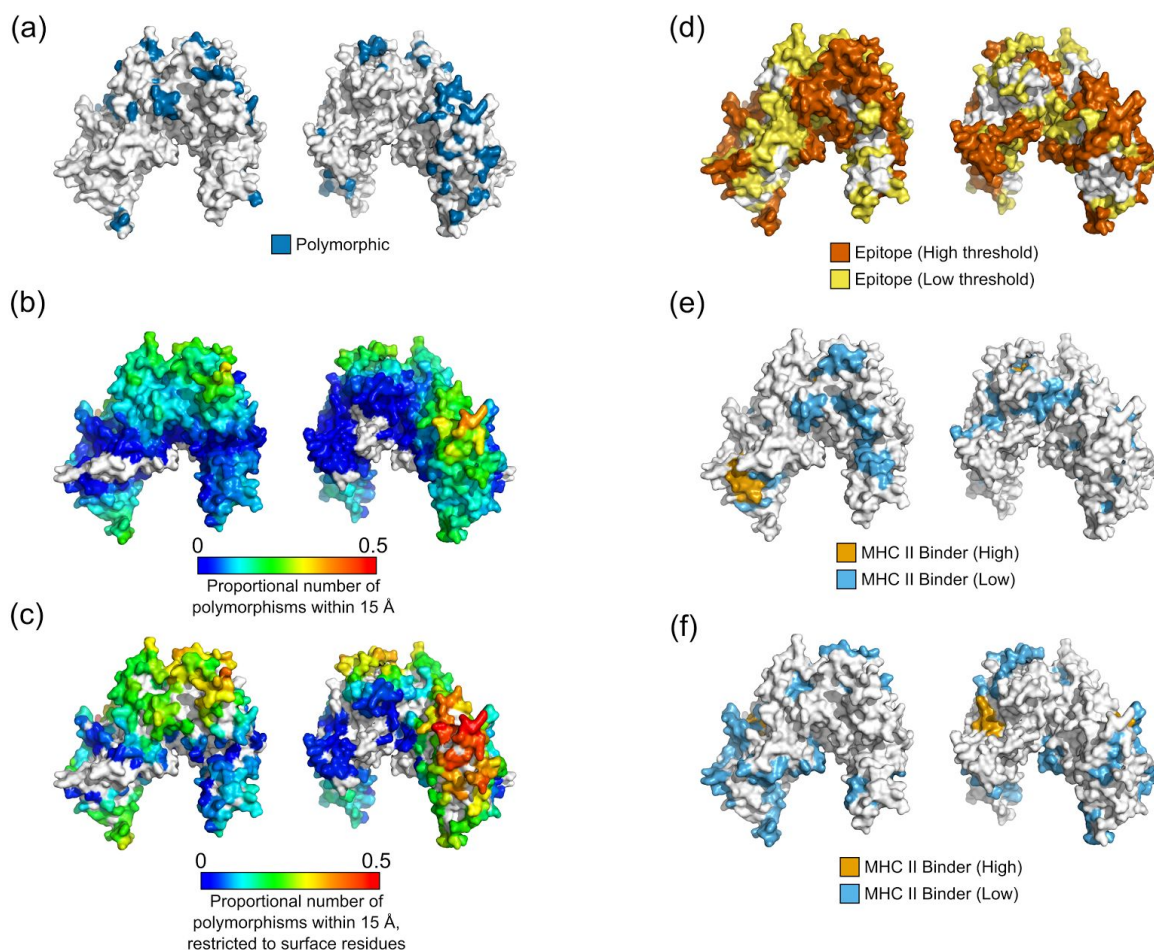


Figure S10: Location of immunologically relevant features mapped onto an PfEMP1 DBL3x-4e structural model (*ModPipe ID: 75fcc4663ff09a106096e5fd1e8ad13a*). Each

panel shows the front and back and top view of the PfEMP1 DBL3x-4e structure. **a)**

Polymorphic residues with an underlying minor allele frequency (MAF) greater than 5% are shown in blue. Sequence polymorphisms were obtained from 65 Gambian isolates¹. **b)**

Spatial averaging of polymorphic residues highlights polymorphic hotspots. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$. **c)** Spatial averaging of polymorphic residues, restricted to surface exposed residues. The proportion of polymorphic residues within 15 Å is shown for each central residue, with polymorphic residues defined as those with a MAF $\geq 5\%$ and surface exposed residues considered to be those with RSA ≥ 0.2 . **d)** Bepipred 2.0

predictions, with epitopes shown for two Bepipred thresholds — predicted epitopes are

shown in yellow for a threshold of 0.5 (specificity = 0.57, sensitivity = 0.59) and in dark orange for a threshold of 0.55 (specificity = 0.81, sensitivity = 0.29). **e, f)** The location of

predicted MHC class II binding peptides are shown for the HLA-DPA1*02:01-DPB1*01:01 (**e**) and HLA-DQA1*05:01-DQB1*03:01 (**f**) alleles. Residues involved in a low binding

peptide ($50 \text{ nM} < \text{IC}_{50} < 500 \text{ nM}$) are shown in light blue, while residues involved in a high binding peptide ($\text{IC}_{50} < 50 \text{ nM}$) residue are shown in orange. Only the core binding region of

each peptide binder is indicated on each structure.

Table S1: Polymorphic hotspots across all *P. falciparum* structures in the PDB, using data on polymorphisms from 65 Gambian isolates. Analysis was restricted to surface exposed residues (RSA > 0.2). A minor allele frequency threshold of 5% was used to identify immunologically relevant polymorphisms.

Gene ID	Description	Location of residues with more than _% polymorphic residues within a 15Å radius	
		>10% polymorphic	>20% polymorphic
PF3D7_1335900	thrombospondin-related anonymous protein (TRAP)	57, 60-63, 65-66, 73, 77, 94, 98-99, 104-105, 109, 112-113, 116-117, 119-129, 152	125
PF3D7_0731500	erythrocyte binding antigen-175 (EBA175)	399-400, 402-407, 481-482, 577-585, 588, 592, 595	-
PF3D7_1036300	merozoite surface protein (DBLMSP2)	162, 166-167, 169-173, 176, 178, 180, 197-198, 200, 216-218, 221-227, 229, 232, 236, 251-252, 255, 261, 287, 290-291, 294-296, 316, 332-333, 335, 339, 342, 344, 355-356, 359-360, 362-364, 366-367, 369-371, 373-374, 388-399, 401-402, 404-406, 408-409, 412-413, 416, 419-427, 429-430, 432, 435, 438-440, 442-443, 445, 447-449, 452-453, 456-457	344, 359, 364, 366-367, 369-371, 373-374, 388-399, 401, 404, 416, 419-427, 429-430, 432, 435, 438-440, 443, 445, 447-449, 452-453, 456-457
PF3D7_1133400	apical membrane antigen 1 (AMA1)	115-116, 121, 124, 148, 164, 167-168, 174-175, 184-189, 192, 194-197, 199-201, 203-207, 209-213, 216, 219, 223-224, 235, 242-246, 281-283, 285-287, 293, 295-297, 299-301, 304-305, 329-330, 332, 339-340, 343-344, 404-405, 407, 424, 435-436, 479-480,	187-189, 192, 194-197, 199-201, 203-207, 209-213, 216, 219, 223, 242-246, 282, 285-286, 296

		483-484, 486-487, 489, 498-504, 508, 510, 512	
PF3D7_0206800	merozoite surface protein 2 (MSP2)*	35-37, 41-43, 45-46	-
PF3D7_0304600	circumsporozoite (CS) protein (CSP)	312-314, 317, 321-322, 324-331, 345, 347-349, 351-359, 361-362, 365-366	325, 349, 352-357
PF3D7_1200600	erythrocyte membrane protein 1, PfEMP1 (VAR2CSA)	2350, 2357-2358, 2360-2362, 2364, 2373-2374, 2376, 2378, 2380-2382, 2385-2386, 2388-2393, 2395-2396, 2399-2400, 2403-2404, 2407-2408, 2410-2417, 2419, 2422, 2425, 2437-2438, 2442-2443, 2447-2448, 2454, 2457-2459, 2472, 2483-2484, 2487-2488, 2570-2571, 2574, 2578-2579	2357-2358, 2360-2361, 2373-2374, 2376, 2378, 2380-2382, 2385, 2393, 2404, 2407-2408, 2442-2443, 2447-2448
PF3D7_0930300	merozoite surface protein 1 (MSP1)	1614, 1639, 1669, 1671-1672, 1674-1676, 1679, 1681, 1687-1688	-
PF3D7_1115700	cysteine proteinase falcipain 2a	257	-
PF3D7_1115300	cysteine proteinase falcipain 2b	255	-
PF3D7_0906500	arginase	154	-

*The MSP2 crystal structure used here is a fragment of MSP2 in complex with an antibody Fv fragment (PDB code = 4QY8) and is not likely to be representative of the native structure of this protein—this region of MSP2 is thought to interact with the parasite plasma membrane ².

Table S2: Polymorphic hotspots across all modelled *P. falciparum* structures with MPQS > 1.1, using data on polymorphisms from 65 Gambian isolates. Analysis was restricted to surface exposed residues (RSA > 0.2). A minor allele frequency threshold of 5% was used to identify immunologically relevant polymorphisms.

Gene ID	Description	Location of residues with more than _% polymorphic residues within a 15Å radius	
		>10% polymorphic	>20% polymorphic
PF3D7_0304600	circumsporozoite (CS) protein (CSP)	317-318, 321-322, 324-330, 345, 347-349, 351-359, 361-362, 365-366	349, 352-354, 356-357, 362
PF3D7_0731500	erythrocyte binding antigen-175 (EBA-175)	229, 329-330, 333, 336, 396, 399-400, 403-407, 409, 412, 577-584, 588, 592, 595	-
PF3D7_0930300	merozoite surface protein 1 (MSP1)	1671-1672, 1674, 1676	-
PF3D7_1335900	thrombospondin-related anonymous protein (TRAP)	39, 43, 46, 57, 60-62, 65-66, 73, 98-100, 104-106, 109, 112-113, 115-117, 119-120, 122-126, 128-129, 148, 152, 238-239, 290	125
PF3D7_1036300	duffy binding-like merozoite surface protein 2 (DBLMSP2)	166-167, 169-170, 172-173, 183, 197-198, 200, 216-218, 221-227, 229, 232, 252, 255, 287, 290-291, 294-296, 332-333, 339, 342, 344, 356, 359-360, 362-364, 366-370, 373-395, 397-399, 401-402, 405-406, 408-409, 412-413, 416, 419-427, 429-430, 432, 435, 438-440, 442, 445, 447-449, 452-453, 456-457	344, 359, 363-364, 366-370, 373-395, 397-399, 401, 416, 419-427, 429-430, 432, 435, 438-440, 445, 447-449, 452-453, 456
PF3D7_1133400	apical membrane antigen 1 (AMA1)	115-116, 118, 121, 124, 148, 164, 167-168, 171, 184-190, 192, 194, 196-197, 199-201, 203-207, 209-213, 219, 223-224, 235, 242-246, 267-268, 270, 281-283, 285-286, 293, 295-297, 299-301, 304-305, 332, 335, 339-340, 343-344, 404-405, 407-408, 423-424, 435-438, 440, 483-486, 489, 492-493,	187-190, 192, 194, 196-197, 199-201, 203-207, 209-213, 219, 223-224, 242-246, 282-283, 285-286, 436

		495-498, 501, 503, 508	
PF3D7_1200600	erythrocyte membrane protein 1, PfEMP1	1209-1215, 1217, 1219-1220, 1222-1223, 1231-1233, 1267, 1280-1281, 1283-1286, 1289, 1300, 1304, 1307-1308, 1315, 1317-1323, 1325-1327, 1329-1333, 1340, 1343, 1368-1369, 1372-1373, 1377-1379, 1381-1392, 1395-1396, 1399-1400, 1403-1404, 1407, 1410, 1414, 1417-1425, 1427-1429, 1432-1434, 1437-1438, 1461-1462, 1464-1465, 1467-1469, 1471-1478, 1480-1482, 1484-1486, 1488-1489, 1491-1492, 1495, 1502, 1505, 1524, 1526, 1534, 1539-1542, 1544-1546, 1548, 1557, 1559-1561, 1564-1572, 1574, 1579-1580, 1582, 1624, 1627-1631, 1633-1634, 1653, 1655, 1657, 1662-1663, 1718, 1722, 1725-1742, 1744, 1751, 1758, 1762-1771, 1773, 1776, 1779, 1809, 1817-1827, 1829-1830, 1837, 1872, 1911-1913, 1915-1916, 2350-2351, 2353-2355, 2357-2362, 2364, 2373-2374, 2376, 2378, 2380-2383, 2385-2386, 2388-2393, 2395-2397, 2400, 2403-2404, 2407, 2410-2417, 2419, 2422, 2425, 2437-2438, 2442-2443, 2447-2448, 2451, 2457-2459, 2480, 2483-2485, 2487-2488, 2521, 2525, 2567, 2570-2571, 2574, 2578-2579	1209-1215, 1217, 1284-1286, 1289, 1300, 1304, 1307-1308, 1317-1323, 1325-1327, 1329-1333, 1340, 1377-1379, 1381-1391, 1400, 1403-1404, 1407, 1410, 1414, 1417-1425, 1427-1429, 1432-1433, 1461, 1464, 1468, 1472, 1474-1477, 1484, 1488, 1526, 1539-1542, 1544-1546, 1548, 1559-1561, 1564-1566, 1568-1572, 1574, 1580, 1629-1631, 1634, 1653, 1726-1742, 1744, 1751, 1762-1771, 1773, 2353-2355, 2357-2362, 2373-2374, 2376, 2378, 2380-2383, 2393, 2397, 2407, 2411-2412, 2442-2443, 2447-2448
PF3D7_1115700	cysteine proteinase falcipain 2a	257	-
PF3D7_1115300	cysteine proteinase falcipain 2b	255	-
PF3D7_0519200	V-type proton ATPase 16 kDa proteolipid	5	-

	subunit		
PF3D7_0919500	major facilitator superfamily domain-containing protein, putative	236	-

SUPPLEMENTARY REFERENCES

1. Amambua-Ngwa, A. *et al.* Population Genomic Scan for Candidate Signatures of Balancing Selection to Guide Antigen Characterization in Malaria Parasites. *PLoS Genet.* 8, e1002992 (2012).
2. Adda, C. G. *et al.* Antigenic characterization of an intrinsically unstructured protein, *Plasmodium falciparum* merozoite surface protein 2. *Infect. Immun.* 80, 4177–4185 (2012).

Chapter 4

Tools for Spatial Aggregation of Protein Data

A large number of online and offline tools exist for the visualisation and manipulation of protein 3D structures. These include Pymol [1], UCSF Chimera [2] and 3Dmol.js [3] for the visualisation of structures. Additionally, there are a number of tools that allow mapping of sequence alignments, variant sites or other parameters onto protein 3D structures. Such tools include POLYVIEW-3D [4], ConSurf [5], Motif3D [6], COMBOSA3D [7], G23D [8] and MuPIT [9]. However, there is a distinct lack of tools that allow application of custom data aggregation functions to sequence-aligned data that has been mapped over a protein 3D structure. Such a tool would be useful to investigate protein-protein interactions, examine selection pressures on structured proteins or simply map desired data onto a protein structure. This chapter presents a computational tool designed to fill this gap, which has been developed as both a Python package (for flexibility and extensibility) and an online web server (for ease-of-use). This tool has been termed BioStructMap (Biological Structure Mapping), and allows for complex functions to be applied to spatially aggregated data.

A sliding window analysis is an approach that is often used in genomics, in which a statistic or metric of interest is applied to successive ‘windows’ of fixed width along the genome. Such an approach is also commonly used in the analysis of protein sequences. At its core, a sliding window analysis applies a data aggregation function to particular subsets of data, where data subsets are defined by proximity along a particular dimension (the independent variable). In a typical sliding window analysis over a protein sequence, this means that residues within a certain distance of (i.e. number of residues away from) a particular position are included in a window centred on that position. For standard sliding window approaches used for both protein and nucleotide sequences, sliding windows are defined according to the linear sequence. However, this neglects the arrangement of a protein in 3D space. The BioStructMap tool presented in this chapter allows application of a ‘3D sliding window’, taking into account residue proximity in 3D space.

The BioStructMap tool applies a 3D sliding window over a protein structure, passing underlying residue data from each window to a data aggregation function, with the resultant value mapped back to the central residue for that window. In this way, protein structural information can be integrated into data analyses which would otherwise use a sliding window applied over the protein sequence or underlying DNA sequence. The data aggregation functions that can be applied by BioStructMap are varied, and include: i) application of a mathematical function to selected residue data (e.g. mean, median, max, min etc), ii) calculation of average amino acid propensity scales (e.g.

hydrophobicity or hydrophilicity), iii) calculation of population genetics statistics such as Tajima's D or Watterson's theta, or iv) calculation of other measures commonly applied in population genetics such as nucleotide diversity or dN/dS. Aside from these built-in functions, spatially aggregated data can also be passed to external command line tools for processing, or users can extend this tool by writing their own Python functions to process data for each window. Examples of this are provided in this chapter.

This chapter formally introduces the new BioStructMap tool (used in Chapter 3) via a short application note, and also includes detailed usage examples across various applications. BioStructMap is available on GitHub (<https://github.com/andrewguy/biostructmap>), via the Python Package Index or via an online server (<https://biostructmap.burnet.edu.au>). The source code has been released under the permissive MIT licence, allowing free modification and reuse. The GitHub repository is the recommended source for up-to-date code containing any updates and bug fixes.

In summary, the BioStructMap package allows us to incorporate protein structural information into tests of selection pressure, allowing for more sensitive identification of protein regions under selection. This is likely to be an important tool in an era in which protein structures are available for a large number of proteins (>130,000 structures in the PDB) and large sequencing datasets allow for identification of particular protein regions under selection pressures.

References

1. Schrödinger, LLC. The PyMOL Molecular Graphics System, Version 1.8. 2015.
2. Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, et al. UCSF Chimera--a visualization system for exploratory research and analysis. *J Comput Chem.* 2004;25: 1605–1612.
3. Rego N, Koes D. 3Dmol.js: molecular visualization with WebGL. *Bioinformatics.* 2015;31: 1322–1324.
4. Porollo A, Meller J. Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. *BMC Bioinformatics.* 2007;8: 316.
5. Ashkenazy H, Erez E, Martz E, Pupko T, Ben-Tal N. ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.* 2010;38: W529–33.
6. Gaulton A, Attwood TK. Motif3D: Relating protein sequence motifs to 3D structure. *Nucleic Acids Res.* 2003;31: 3333–3336.
7. Stothard PM. COMBOSA3D: combining sequence alignments with three-dimensional structures. *Bioinformatics.* 2001;17: 198–199.
8. Solomon O, Kunik V, Simon A, Kol N, Barel O, Lev A, et al. G23D: Online tool for mapping and visualization of genomic variants on 3D protein structures. *BMC Genomics.* 2016;17: 681.
9. Niknafs N, Kim D, Kim R, Diekhans M, Ryan M, Stenson PD, et al. MuPIT interactive: webserver for mapping variant positions to annotated, interactive 3D structures. *Hum Genet.* 2013;132: 1235–1243.

Structural bioinformatics

BioStructMap: A Python tool for integration of protein structure and sequence-based features

Andrew J. Guy^{1,2,*}, Vashti Irani^{1,4}, Jack S. Richards^{1,3,4,6} and Paul A. Ramsland^{1,2,5,7}

¹Life Sciences, Burnet Institute, Melbourne, Australia; ²Departments of Immunology and ³Infectious Diseases, Monash University, Melbourne, Australia; ⁴Departments of Medicine and ⁵Surgery Austin Health, University of Melbourne, Melbourne, Australia; ⁶Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Australia; ⁷School of Science, RMIT University, Bundoora, Australia

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: A sliding window analysis over a protein or genomic sequence is commonly performed, and we present a Python tool, BioStructMap, that extends this concept to three-dimensional (3D) space, allowing the application of a 3D sliding window analysis over a protein structure. BioStructMap is easily extensible, allowing the user to apply custom functions to spatially aggregated data. BioStructMap also allows mapping of underlying genomic sequences to protein structures, allowing the user to perform genetic-based analysis over spatially linked codons—this has applications when selection pressures arise at the level of protein structure.

Availability and implementation: The Python BioStructMap package is available at <https://github.com/andrewguy/biostructmap> and released under the MIT License. An online server implementing standard functionality is available at <https://biostructmap.burnet.edu.au>.

Contact: andrew.guy@burnet.edu.au

Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Consideration of three-dimensional (3D) protein structure is important in many areas of research, including antibody-antigen interactions, protein-protein interactions and drug interactions with proteins. For example, antibody recognition of a dominant epitope can lead to selection pressures on residues associated with that epitope; these residues may be distant in the linear protein sequence despite being spatially connected. In immunology, these non-linear sequence-structure relationships are referred to as discontinuous or conformational epitopes. There are a number of pre-existing online tools that allow for visualisation and mapping of pre-defined features onto protein structures (Porollo and Meller, 2007; Baker and Porollo, 2016; Ashkenazy et al., 2010; Segura et al., 2017), however none of these tools allow for application of a 3D sliding window over a protein structure using user-defined functions. There are many settings in which sliding window analysis is applied to genomic or

protein sequences, and we demonstrate that this sliding window approach can be extended to 3D protein structures.

2 Methods

We present here a Python package named BioStructMap that allows mapping of sequence-associated data onto a protein structure. This tool also allows for the application of a 3D ‘sliding window’ over a protein structure. The user can apply a variety of functions to spatially aggregated data, mapping the result back to the central residue within each window. The user must provide sequence-aligned data, a reference sequence, and PDB format coordinates over which to process data. For each residue in the structure, all residues within a user-defined radius are selected. Data corresponding to these residues (i.e. specific characteristics of interest for these residues) is then passed to a function that returns a numerical value, which is then mapped back to the central residue (Figure S1). A number of predefined functions are included in the BioStructMap package. Users can also supply their own function for data

processing. Data is output as a Python dictionary of residues and associated values, written to a PDB file in the B-factor column, or as a text file. Results can be viewed using PyMOL or similar programs.

BioStructMap uses the Biopython Bio.PDB module for handling PDB files, and can accept both PDB and mmCIF files as input. Sequence alignments are performed using either the NCBI BLAST+ package or the Biopython Bio.pairwise2 module. Alignment of DNA sequences to protein sequences is performed using Exonerate (Slater and Birney, 2005) which allows handling of intron-containing sequences and reverse-sense translation. Calculation of Tajima's D is performed using the Python DendroPy package (Sukumaran and Holder, 2010).

The source code for BioStructMap is available on GitHub (<https://github.com/andrewguy/biostructmap>) or via the Python Package Index (PyPI). A simple web-server interface is also available at <https://biostructmap.burnet.edu.au>, using the JavaScript NGL viewer for visualisation of protein structures (Rose and Hildebrand, 2015). Results can be viewed in the browser or downloaded as PDB files. Further details on BioStructMap use are available in Supplementary Material.

3 Usage example

In areas of endemic malaria, immune selection pressure on the malaria parasite can lead to balancing selection, in which low-frequency alleles are maintained at a higher proportion than would otherwise be expected under a neutral model of selection. Tajima's D (Tajima, 1989) is one statistic that has been used to identify regions under balancing selection within the malaria genome, and has previously been applied as a sliding window over genes of interest (Arnott et al., 2013, 2014). We have also previously applied the BioStructMap tool to key vaccine candidates from *P. falciparum* and *P. vivax*, incorporating protein structural information into calculations of selection pressures and diversity (Guy et al., 2018a, 2018b). We illustrate here one of the potential uses for the BioStructMap tool, applying a 3D sliding window calculation of Tajima's D over the protein structure of *Plasmodium falciparum* EBA-175 Region II (RII), a leading malaria vaccine candidate (Figure 1). This approach groups data that are spatially connected but are distant in the linear sequence. Nucleotide sequences for EBA-175 RII were extracted from GenBank, originally deposited from a study examining signatures of selection in *P. falciparum* strains from Kenya and Thailand (Verra et al., 2006). Since known structures contain a number of unresolved residues, ModPipe (Eswar et al., 2003) was used to generate a comparative structural model for EBA-175 RII. A radius of 15 Å was selected for each window as this is the typical maximum-dimension for an antibody-antigen interface (Ramaraj et al., 2012). When analyzed, a surface exposed loop with a high spatially derived Tajima's D value is identified in both Kenyan and Thai isolates. Importantly, this region is involved in the dimerisation of EBA-175 RII around its glycoporphin A binding partner on the surface of the human red blood cell (Tolia et al., 2005), and antibodies that target the dimerization interface of EBA-175 RII have previously been shown to be highly effective at inhibiting parasite entry into red blood cells (Chen et al., 2013). A region within the F1 domain is also identified as having high Tajima's D values within Thai samples, but to a much lesser extent in Kenyan samples. Further experimental work would be required to validate this region as a target of functional antibody responses.

4 Concluding remarks

The BioStructMap package and associated web interface allow for visualisation of sequence-aligned data over a 3D protein structure, as well as

allowing the incorporation of protein structural information into sequence-based metrics using a 3D sliding window approach. This tool is

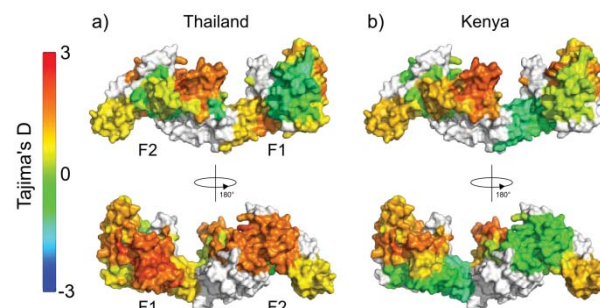


Fig. 1: Tajima's D calculation applied as a 3D sliding window over the protein structure of *P. falciparum* EBA-175 Region II. The F1 and F2 domains are indicated on the monomeric structure. Nucleotide sequences were obtained from *P. falciparum* isolates from (a) Thailand ($n = 48$) and (b) Kenya ($n = 39$) (Verra et al., 2006). The BioStructMap Python package was used to apply Tajima's D calculations using a 3D sliding window with a radius of 15 Å. The structural model is available via ModBase, accession number: ed998157a605f5e58ed66e198e0ae1ab. Structures were visualised with PyMOL.

applicable to a variety of problems, including identification of regions under various forms of genetic, immunological or drug selection pressure, and spatial mapping of residue characteristics that may affect immunogenicity, solubility, binding interaction, etc. The tool is easily extensible, allowing users to define their own functions to apply to spatially aggregated data.

Acknowledgements

We thank Dyson Simmons and Andrew Walter for support with development and deployment of the BioStructMap web-based server.

Funding

This work is supported by the National Health and Medical Research Council (NHMRC) of Australia [APP1037722 & APP1125788 to J.S.R.], and an Australian Postgraduate Award to A.J.G. Burnet Institute received funding from the NHMRC Independent Research Institutes Infrastructure Support Scheme, and the Victorian State Government Operational Infrastructure Support Scheme.

Conflict of Interest: none declared.

References

- Arnott, A. et al. (2014) 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.*, **13**, 233.
- Arnott, A. et al. (2013) Global Population Structure of the Genes Encoding the Malaria Vaccine Candidate, *Plasmodium vivax* Apical Membrane Antigen 1 (Pv AMA1). *PLoS Negl. Trop. Dis.*, **7**, e2506.
- Ashkenazy, H. et al. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res.*, **38**, W529–33.
- Baker, F.N. and Porollo, A. (2016) CoeViz: a web-based tool for coevolution analysis of protein residues. *BMC Bioinformatics*, **17**, 119.
- Chen, E. et al. (2013) Structural and functional basis for inhibition of erythrocyte invasion by antibodies that target *Plasmodium falciparum* EBA-175. *PLoS Pathog.*, **9**, e1003390.
- Eswar, N. et al. (2003) Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.*, **31**, 3375–3380.
- Guy, A.J. et al. (2018a) Proteome-wide mapping of immune features onto *Plasmodium* protein three-dimensional structures. *Sci. Rep.*, **8**, 4355.

BioStructMap

- Guy, A.J. *et al.* (2018b) Structural patterns of selection and diversity for *Plasmodium vivax* antigens DBP and AMA1. *Malar. J.*, **17**, 183.
- Hamelryck, T. and Manderick, B. (2003) PDB file parser and structure class implemented in Python. *Bioinformatics*, **19**, 2308–2310.
- Porollo, A. and Meller, J. (2007) Versatile annotation and publication quality visualization of protein complexes using POLYVIEW-3D. *BMC Bioinformatics*, **8**, 316.
- Ramaraj, T. *et al.* (2012) Antigen-antibody interface properties: composition, residue interactions, and features of 53 non-redundant structures. *Biochim. Biophys. Acta*, 1824, 520–532.
- Rose, A.S. and Hildebrand, P.W. (2015) NGL Viewer: a web application for molecular visualization. *Nucleic Acids Res.*, **43**, W576–9.
- Segura, J. *et al.* (2017) 3DBIONOTES v2.0: a web server for the automatic annotation of macromolecular structures. *Bioinformatics*, **33**, 3655–3657.
- Slater, G.S.C. and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
- Sukumaran, J. and Holder, M.T. (2010) DendroPy: a Python library for phylogenetic computing. *Bioinformatics*, **26**, 1569–1571.
- Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*, **123**, 585–595.
- Tolia, N.H. *et al.* (2005) Structural basis for the EBA-175 erythrocyte invasion pathway of the malaria parasite *Plasmodium falciparum*. *Cell*, **122**, 183–193.
- Verra, F. *et al.* (2006) Contrasting signatures of selection on the *Plasmodium falciparum* erythrocyte binding antigen gene family. *Mol. Biochem. Parasitol.*, **149**, 182–190.

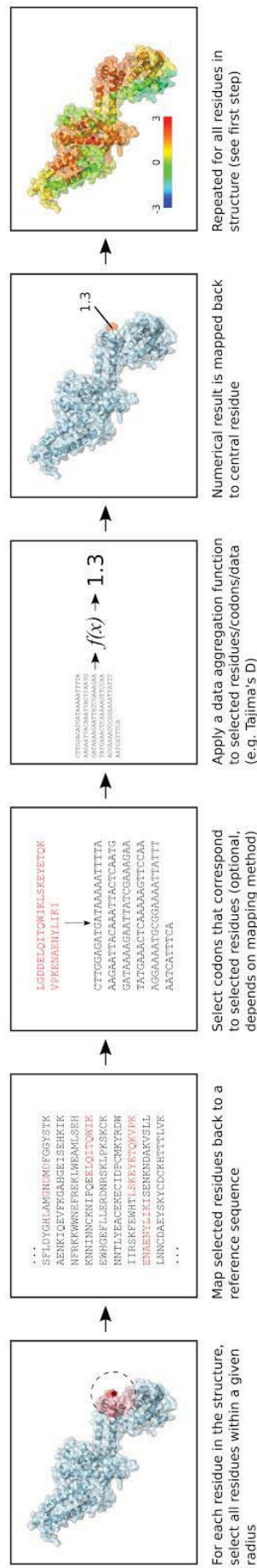


Fig. S1: Application of a 3D sliding window over a protein structure, as implemented in BioStructMap. For each residue in the protein structure, residues within a given radius are identified, and the data corresponding to these residues is extracted. This data is user-supplied, and could be sequence-aligned numerical data, the location of polymorphic residues or genomic sequences. This data is passed to a data aggregation function (e.g. calculate Tajima's D for codons which map to selected residues), and the result from this function mapped back to the central residue. This process is repeated for every residue in the protein structure. BioStructMap is highly flexible with regards to data aggregation functions and user-supplied data—users can write their own data aggregation functions to process specific sequence-aligned data.

BioStructMap Usage Guide

— Andrew Guy, 2017

1. Introduction

1.1 Rationale

Often in genomics or protein biology, a statistical test or data aggregation function is applied as a sliding window over a gene or protein sequence. For example, to identify regions under balancing selection, Tajima's D has been applied to the *Plasmodium falciparum* AMA1 gene as a sliding window (Arnott et al. 2014). Similarly, a sliding window analysis can be performed over a protein sequence, often using some amino acid propensity scale, such as the Kyte & Doolittle hydrophobicity scale (Kyte and Doolittle 1982). However, these analyses fail to account for the arrangement of a protein in 3D space. In the case of the Tajima's D analysis of *P. falciparum* AMA1 mentioned before, balancing selection is thought to arise as a result of immune selection pressure on this particular antigen. Additionally, it is likely that this immune selection pressure is antibody mediated, and hence the result of interactions between antibodies and the structured antigen. Many of the potential interaction sites in such an interaction involve discontinuous regions of the protein sequence i.e. form a conformational/discontinuous epitope. As a result, using a sliding window over the protein sequence is unlikely to fully capture the selection pressures on complex structural epitopes.

With this in mind, we have proposed the application of a 3D sliding window over a protein structure, analogous to the standard 2D sliding window analysis that is often applied over a protein or gene sequence. This 3D sliding window analysis has been implemented in a Python package called BioStructMap.

1.2 Overview

The BioStructMap tools allow for the application of a 3D sliding window analysis over a protein structure. To achieve this, the user must supply a set of sequence-aligned data and a corresponding reference sequence. This reference sequence is used to map data to the protein structure. A set of 3D windows are then created (one for each residue in the structure) with a user-defined radius, and data from each window is passed to a data aggregation function. The result from this data aggregation function is then mapped back to the central residue within each window. These results can then be viewed over the protein structure using a program such as PyMOL (<https://pymol.org>). This procedure is analogous to the traditional 2D sliding window analysis that is often performed over a protein or gene sequence, but also captures information on the spatial arrangement of residues in 3D space.

1.3 Availability and installation

The BioStructMap package is available via the Python Package Index (PyPI), which means that installation with `pip` is as simple as:

```
pip install biostructmap
```

Alternatively, the latest source code can be downloaded from GitHub and installed:

```
git clone https://github.com/andrewguy/biostructmap.git biostructmap
cd ./biostructmap
python setup.py install
```

It is recommended that users install `numpy` before installing `biostructmap`.

BioStructMap also has soft dependencies on the NCBI BLAST+ tool (<https://www.ncbi.nlm.nih.gov/guide/howto/run-blast-local/>) and Exonerate (<https://www.ebi.ac.uk/about/vertebrate-genomics/software/exonerate>). If you choose not to install these, or don't want to use them, all sequence alignments will be performed using the Biopython `Bio.pairwise2` module. This should work just as well if your reference sequence is reasonably similar to the sequence of the PDB file. If this is not the case, then we suggest that a better approach may be to build a homology model using MODELLER (Webb and Sali 2016) and use this instead of using the poorly aligned PDB structure.

If either BLAST+ or Exonerate are not installed, you should indicate this by setting the relevant flags during BioStructMap usage:

```
import biostructmap

biostructmap.seqtools.LOCAL_BLAST = False
biostructmap.seqtools.LOCAL_EXONERATE = False
```

Some functions within BioStructMap also require installation of the DSSP tool (<http://swift.cmbi.ru.nl/gv/dssp/>). These include calculation of relative solvent accessibility and secondary structure determination. If you wish to use these functions, you must have DSSP installed.

2. Basic Usage

Although the BioStructMap package contains several modules, most of these work behind the scenes. The `biostructmap` module should be the only module that needs to be directly used in most cases.

2.1 The Structure class

The main class within the `biostructmap` module is the `Structure` class. This is initialised as such:

```
from biostructmap import biostructmap

my_structure = biostructmap.Structure(pdbfile='./1zrl.pdb', pdbname='1ZRL', mmCIF=False)
```

The `pdbfile` argument can be either a string of the file path to the PDB file of interest, while the `pdbname` argument is an optional descriptive string that is used when naming output files. The optional `mmCIF` flag is used to indicate if the input file is in mmCIF format. If you are using an mmCIF file, you would instead run:

```
my_structure = biostructmap.Structure(pdbfile='./1zrl.cif', pdbname='1ZRL', mmCIF=True)
```

2.2 Mapping data over a structure

The `Structure` class contains a number of methods, the most important being the `map()` method. This method allows for the mapping of data over a protein structure, with the ability to also apply some sort of spatial aggregation to data. The `map` method takes a number of arguments, the most important of which are `data`, `method`, `ref` and `radius`.

The `Structure.map` method returns a dictionary-like object, which can also be used to write output data to a PDB file or CSV file.

data The `data` parameter is a dictionary mapping individual chains within the protein structure to relevant data objects. The exact form of the data object will depend on the `method` argument selected. For example, if you are mapping the location of polymorphic residues onto a structure, the data would be a list of polymorphic sites:

```
data = {('A', 'B'): [1, 34, 56, 77, 120, 121, 125],
        ('C', ): [5, 34, 67, 122]}
}
```

In this example, identical chains A and B are both assigned the same set of data, whilst the unique chain C is assigned another set of polymorphic residues. Note that the given polymorphic sites for each chain are aligned to the reference sequence supplied - see below for more detail.

method The `method` parameter is either a string representing a method for mapping data (one of a number of pre-defined methods), or a custom function for mapping data (explained in more detail below). For example, to map polymorphic sites onto a protein structure, set `method=snps`.

ref The `ref` parameter is a set of reference sequences for all chains, and is used to align the user-supplied data to the protein structure. All data supplied via the `data` argument should align to these reference sequences.

For example, with identical chains A and B, and a unique chain C, we would have:

```
ref = {'A': 'KTQEDKL...DJSKJK',
       'B': 'KTQEDKL...DJSKJK',
       'C': 'NAPNLEV...KLWELW'}
# Note: sequences have been condensed for readability
# You need to provide the full-length sequence!
```

All data provided in the `data` parameter should align to these sequences.

Also note the subtle difference in the keys needed for the `ref` dictionary vs. the `data` dictionary. The `ref` dictionary should have a sequence provided for every chain being evaluated, with the dictionary key being the *string* identifier for that chain. The `data` dictionary requires a *tuple* of chain identifiers for each related data value. This difference arises so that it is possible to map the same data over multiple chains and subsequently ensure we don't duplicate identical data points that might fall within the same radius (from different chains). This will be discussed in more detail in a later section.

If the supplied data is a genomic multiple sequence alignment, then the provided reference sequence should also be a genomic sequence. In this case `biostructmap` will align this genomic sequence to the protein sequence for the relevant chain in the structure. If a genomic sequence is supplied, then the `map_to_dna` argument should also be set to `True`.

If the `ref` argument is not provided, then the sequences for each chain in the structure are used.

radius The radius (in Angstrom) over which to select nearby residues for inclusion within each 3D window. This defaults to 15 Angstrom, which is the typical maximum dimension for an antibody epitope. If you simply want to map data to individual residues (eg. to display polymorphic sites on a protein structure), set `radius=0`.

selector When determining which residues fall within a given radius of a central residue, there are a number of ways in which to compute distances

between residues. The default behaviour is to compute the minimum distance between any two atoms in each pair of residues. The `selector` argument allows the user to specify other atoms by which to compute residue distance. By default this argument is `'all'`, which gets all non-heterologous atoms. Other potential options include `'CA'`, `'CB'` etc. If an atom is not found within a residue object, then the selection method reverts to using `'CA'`.

rsa_range If the user wishes to restrict analysis to residues that fall within a given range of relative solvent accessibility (RSA) values (eg. only surface exposed residues), they can provide a tuple to the `rsa_range` argument. This argument takes a tuple in the form `(minimum, maximum)`, where `minimum` and `maximum` are float values between 0 and 1.

If any residue falls outside the given range of RSA values, then this residue will be ignored in all calculations.

RSA is calculated using the DSSP software. If this is not installed and available on the user's PATH, then any attempt to use the `rsa_range` argument will fail, throwing an exception.

map_to_dna The `map_to_dna` argument is a binary flag to indicate if the reference sequence to be aligned is a DNA sequence. This needs to be set to `True` if the reference sequence is a DNA sequence (e.g. when using the Tajima's D method).

method_params In order to make `biostructmap` flexible and extensible, the `map` method also takes additional arguments that will be passed to the data aggregation method. These arguments should be provided to the `method_params` argument in a dictionary of keyword arguments (key) and associated values (value).

To provide a concrete example of this, we can consider the `'default_mapping'` method that applies a data aggregation method to sequence-aligned numerical data. By default, this method calculates the mean of all data within each radius. However, we can apply other data aggregation functions (e.g. calculate the median) to each 3D window by passing a `method` argument to the `default_mapping` function:

```
import numpy as np

my_structure.map(..., method_params={'method': np.median})
```


2.3 Basic Usage examples

2.3.1 Mapping polymorphic hotspots One usage of the `biostructmap` tool is to determine polymorphic hotspots on a protein structure. This requires the user to provide a list of all polymorphic residues of interest and an associated reference sequence. In this example we have a single-chain structure with the PDB file `1zrl.pdb` and a reference sequence in FASTA format in the file `reference.fasta`. We will use the Biopython `SeqIO` module to read in the reference sequence from file. Polymorphic residues are residues 3, 67, 78, 99, 100, 120 and 121, relative to the reference sequence (where the first residue is number 1).

If we were interested in averaging the number of polymorphisms within a 10 Angstrom radius, we would run:

```
import biostructmap
from Bio import SeqIO

reference_seq = SeqIO.read("reference.fasta", "fasta")

my_structure = my_structure = biostructmap.Structure(pdbfile='./1zrl.pdb',
                                                    pdbname='1ZRL')
hotspots = my_structure.map(data={('A',): [3, 67, 78, 99, 100, 120, 121]}
                           method='snps',
                           ref={'A': reference_seq},
                           radius=10
                           )
```

2.3.2 Amino acid propensity scales We can also apply a 3D sliding window to calculation of amino acid propensity scales. In this example we will apply the Kyte & Doolittle index of hydrophobicity to the protein structure initialized in the above example. We will also demonstrate how to apply a custom amino acid scale as a 3D sliding window.

We can obtain the Kyte & Doolittle scale from the Biopython package:

```
from Bio.SeqUtils import ProtParamData
kd_scale = ProtParamData.kd
```

For the `'aa_scale'` method, the `data` argument should be a dictionary representing the amino acid scale of interest. In this example we will use a window size of 15 Angstrom, and only consider surface exposed residues ($RSA > 0.2$).


```
mean_hydrophobicity = my_structure(data=ProtParamData.kd_scale,
                                   method='aa_scale',
                                   ref={'A': reference_seq},
                                   radius=15,
                                   rsa_range=(0.2, 1)
                                   )
```

To use a custom amino acid propensity scale, we just need to provide a dictionary of numerical values for all amino acids. We will apply the 'relative mutability scale' defined by (Dayhoff, Schwartz & Orcutt, 1978). Again, we are only considering surface exposed residues.

```
relative_mutability = {
    'A': 100, 'R': 65, 'N': 134, 'D': 106, 'C': 20, 'Q': 93,
    'E': 102, 'G': 49, 'H': 66, 'I': 96, 'L': 40, 'K': 56,
    'M': 94, 'F': 41, 'P': 56, 'S': 120, 'T': 97, 'W': 18,
    'Y': 41, 'V': 74
}

mean_mutability = my_structure(data=relative_mutability,
                              method='aa_scale',
                              ref={'A': reference_seq},
                              radius=15,
                              rsa_range=(0.2, 1)
                              )
```

2.3.3 Calculation of Tajima's D Tajima's D is a statistical test used to determine if a sequence is evolving under non-neutral selection pressure. Here we will apply Tajima's D as a 3D sliding window over our protein structure. We need to supply a multiple sequence alignment, using the `biostructmap.SequenceAlignment` class. The multiple sequence alignment is initially supplied as a FASTA file.

In this case, the reference sequence is taken as the first sequence in the multiple sequence alignment. Note the need to set `map_to_dna=True`.

```
msa = biostructmap.SequenceAlignment('./alignment.fasta', file_format='fasta')
reference_seq = str(msa_data[0].seq)

tajimas_d = my_structure(data=({'A',}): msa),
                      method='tajimasd',
                      ref= {'A': reference_seq},
                      radius=15,
                      map_to_dna=True
                      )
```


2.3.4 Nucleotide diversity Nucleotide diversity is a metric that is used to quantify the degree of diversity within a particular window on a gene. We can extend this here to a 3D window over a structure to get a sense of the particular regions of the protein structure that are most diverse within a population (at a genomic level).

Again, we need to supply a multiple sequence alignment.

```
msa = biostructmap.SequenceAlignment('./alignment.fasta', file_format='fasta')
reference_seq = str(msa_data[0].seq)

nucleotide_diversity = my_structure(data={'A': msa},
                                   method='nucleotide_diversity',
                                   ref= {'A': reference_seq},
                                   radius=15,
                                   map_to_dna=True
                                   )
```

2.3.5 Applying a custom data aggregation function The 'default_mapping' method allows the user to apply a custom data aggregation function to data within each window. For example, you could calculate the arithmetic mean of data within a window, calculate the maximum or minimum value within a radius, or apply some other metric to data. We will illustrate with a simple calculation of the maximum data value within a 5 Angstrom window.

Note the use of the additional keyword argument `method_params`, which takes a dictionary of additional parameters to pass to the `default_mapping` method. In this case, `default_mapping` takes the keyword argument `method`, which should be a function that can be used to aggregate a list of data points. This `default_mapping` method can be quite useful when constructing custom mapping procedures!

```
data_values = list(range(1000)) # Just some placeholder data

maximum_values = my_structure(data=data_values,
                              method='default_mapping',
                              ref={'A': reference_seq},
                              radius=5,
                              method_params={
                                  'method': max
                              }
                              )
```


2.4 Results

The results for each mapping call are returned in a dictionary-like object (`DataMap` class - a simple class that extends the `dict` class by adding a couple of additional methods to deal with writing results to files).

The main method that is likely to be used from the `DataMap` object is the `write_data_to_pdb_b_factor` method. This writes all data to the B-factor column of a PDB file, allowing easy visualisation in a program such as PyMOL.

We demonstrate the use of this following a simple calculation of average hydrophobicity (see section 2.3.2).

```
import biostructmap
from Bio import SeqIO
from Bio.SeqUtils import ProtParamData

kd_scale = ProtParamData.kd
reference_seq = SeqIO.read("reference.fasta", "fasta")

mean_hydrophocity = my_structure(data=ProtParamData.kd_scale,
                                method='aa_scale',
                                ref={'A': reference_seq},
                                radius=15,
                                rsa_range=(0.2, 1)
                                )

mean_hydrophocity.write_data_to_pdb_b_factor(fileobj='./1ZRL_hydrophocity.pdb')
```

For the `write_data_to_pdb_b_factor` method, the `fileobj` keyword argument can be either an output file name as a string, or a file-like object to write output data to. Additionally keyword arguments for this method are `default_no_value` and `scale_factor`. The `default_no_value` argument is used to specify the numerical value written to the B-factor column if the value for this residue is `None` (non-numerical values can't be written to the B-factor column). The `scale_factor` argument is used to scale output values in situations where they are either too big or small to fit within the B-factor column. For example, it is usually sensible to scale nucleotide diversity values by a factor of 1000 (`scale_factor=1000`).

3. Extending BioStructMap

`BioStructMap` can be extended by providing custom functions with which to process data within each 3D sliding window. We will briefly discuss the format required for these custom data processing functions.

Each data processing function has the format:


```
def some_method(structure, data, residues, ref, **kwargs):
    ...
    return final_data
```

where `structure` is the parent `biostructmap.Structure` object from which the `map` method has been called, `data` is the `data` argument supplied to the `map` method (no filtering has been applied to this object yet!), `residues` is a list of PDB residues within that particular window, and `ref` is a dictionary mapping PDB residue numbers to reference sequence indices.

3.1 Data aligned to a protein sequence

If the custom method being written needs to deal with data that is aligned to a protein sequence, then the key steps that need to be followed are:

1. From the list of PDB residues within the window, extract the positions of these residues in the corresponding reference sequence.
2. Construct a list of applicable data points, given the list of reference-sequence aligned residues.
3. Perform a data aggregation function over these data points, returning a single value (a single, numerical return value is not absolutely required, although writing data to a PDB file will not be possible if data is non-numerical).

We illustrate these steps with a function to calculate the mean of selected data points:

```
import numpy as np

def calculate_mean(_structure, data, residues, ref, ignore_duplicates=True):
    # Step 1: Get a list of all keys from the data object.
    chains = data.keys()

    # Step 1: Extract position of residues in the reference sequence
    ref_residues = [ref[x] for x in residues if x in ref]

    # Step 1: A little bit more manipulation, as each reference residue is given
    # by a (chain, residue number) tuple, while the data keys are tuples that
    # can contain multiple chains (to enable mapping data between several chains).
    # The list of residues below will look like:
    # [(('A', 'B'), 1), (('A', 'B'), 2), etc.]
    residues = [(chain, x[1]) for chain in chains
```



```
        for x in ref_residues if x[0] in chain]

    # If two separate chains both contain the same data point, and both
    # within the same window, then we might want to de-duplicate this data
    # point, to prevent skewing of the result.
    if ignore_duplicates:
        residues = set(residues)

    # Step 2: Get applicable data points
    data_points = [data[res[0]][res[1]] for res in residues]

    # Step 3: If there is any data that maps to residues within the given
    # window, then we apply some data aggregation method to this data.
    # Note that for this mapping procedure, we define this method to be the
    # arithmetic mean by default.
    if data_points:
        result = np.mean(data_points)
    else:
        result = None
    return result
```

This can then be used by passing this method to the `map` function of a `Structure` object. Note that the data passed to the `map` function should be aligned to the reference sequence.

```
data_values = list(range(1000)) # Just some placeholder data
mean_values = my_structure.map(data={'A': data_values}
                               method=calculate_mean,
                               ref={'A': reference_seq},
                               radius=15
                               )
```

Alternatively, if the user simply wants to apply an aggregation function to selected data points (as in the above example), then the `default_mapping` method provides a nice interface to perform this via the additional keyword argument `method_params`:

```
data_values = list(range(1000)) # Just some placeholder data
mean_values = my_structure.map(data={'A': data_values}
                               method='default_mapping',
                               ref={'A': reference_seq},
                               radius=15,
                               method_params={'method': np.mean}
                               )
```


3.2 Genomic multiple sequence alignment data

If the user wants to perform a statistical test or data aggregation on codons that map to residues within each window, then the simplest option is to use the `_genetic_test_wrapper` function defined in the `biostructmap.map_functions` module. This is a simple wrapper function that constructs a multiple sequence alignment from all codons that map to residues within a window. For example, to define a test to calculate nucleotide diversity:

Firstly, we define a function to calculate nucleotide diversity from a multiple sequence alignment. We are going to use the DendroPy library to perform this.

```
def diversity_from_dendropy(sequence_alignment):
    # Just make sure the alignment is a string in fasta format.
    if not isinstance(alignment, str):
        data = alignment.format('fasta')
    else:
        data = alignment

    # If the alignment doesn't exist, then return None.
    if not alignment or len(alignment[0]) == 0:
        return None

    # Construct the relevant DenroPy data structure
    seq = dendropy.DnaCharacterMatrix.get(data=data, schema='fasta')

    # Calculate diversity
    diversity = dendropy.calculate.popgenstat.nucleotide_diversity(seq)

    return diversity
```

Now we can use the `_genetic_test_wrapper` function to pass alignments from each window to the `diversity_from_dendropy` function:

```
from biostructmap.map_functions import _genetic_test_wrapper

def _calculate_nucleotide_diversity(_structure, alignments, residues, ref):
    nucleotide_diversity = _genetic_test_wrapper(_structure, alignments,
                                                residues, ref,
                                                diversity_from_dendropy)

    return nucleotide_diversity
```

We can then use this function:

```
msa = biostructmap.SequenceAlignment('./alignment.fasta', file_format='fasta')
```



```
reference_seq = str(msa_data[0].seq)

nucleotide_diversity = my_structure(data={'A': msa},
                                    method=_calculate_nucleotide_diversity,
                                    ref= {'A': reference_seq},
                                    radius=15,
                                    map_to_dna=True
                                    )
```

3.3 Genomic data passed to a command line tool

It is also possible to pass data to a command line tool for processing using the general format outline above.

We firstly need to write a temporary file containing a multiple sequence alignment for each window, and then call our command line tool. In this example we will use 'possum' (<https://github.com/jeetsukumaran/possum>), which calculates a number of population statistics from a multiple sequence alignment.

```
def call_possum(alignment)
    #Run external tool over sub alignment.
    with tempfile.NamedTemporaryFile(mode='w') as seq_file:
        seq_file.write(alignment)
        # Make sure data is actually written to file.
        seq_file.flush()
        process = subprocess.run(["/opt/bin/possum", "-f", "dnafasta", "-q",
                                "-v", seq_file.name], stdout=subprocess.PIPE)
    try:
        # Just need to parse the output data to get a numerical value.
        tajd = float(process.stdout.decode().strip().split('\t')[-1])
    except ValueError:
        tajd = None
    return tajd
```

Once we have a function that will process a single multiple sequence alignment, we can wrap this in the `_genetic_test_wrapper` function:

```
def tajimas_d_from_possum(_structure, alignments, residues, ref):
    result = _genetic_test_wrapper(_structure, alignments, residues, ref,
                                   call_possum)

    return result
```

Fianlly, we can use the final function to map Tajima's D over a structure:


```
msa = biostructmap.SequenceAlignment('./alignment.fasta', file_format='fasta')
reference_seq = str(msa_data[0].seq)

tajimas_d = my_structure(data={'A': msa},
                          method=tajimas_d_from_possum,
                          ref= {'A': reference_seq},
                          radius=15,
                          map_to_dna=True
                          )
```

4. References

- Arnott,A. et al. (2014) 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.*, **13**, 233.
- Kyte,J. and Doolittle,R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.*, **157**, 105–132.
- Dayhoff M.O., Schwartz R.M. and Orcutt B.C. (1978) Atlas of Protein Sequence and Structure *National Biomedical Research Foundation*, **5**, 345-352.
- Webb,B. and Sali,A. (2016) Comparative Protein Structure Modeling Using MODELLER. *Curr. Protoc. Bioinformatics*, **54**, 5.6.1–5.6.37.

Chapter 5

Selection pressures on key *P. vivax* antigens: a structural perspective

Previous chapters have explored the use of a novel computational tool that allows incorporation of protein structural information into measures of selection pressure. In this chapter this tool is applied to two key *P. vivax* antigens, and examine selection pressures across multiple geographic locations. In particular, *PvAMA1* and *PvDBP* region II were examined, using homology models based on known protein structures, and compared spatially-derived nucleotide diversity and Tajima's D across populations.

AMA1 is a major vaccine candidate in both *P. falciparum* and *P. vivax*, and is present in all *Plasmodium* species [1]. It is released from the micronemes onto the surface of the merozoite during the invasion process, and binds to RON2, assisting in the formation of a tight junction between the malaria parasite and host-cell membranes [2–4]. AMA1 has been shown to be an important target of anti-parasite immune responses in both *P. falciparum* [5,6] and *P. vivax* [7], supporting its consideration as a vaccine candidate. AMA1 is highly polymorphic, which may present an impediment to developing strain-transcending immunity from a single vaccine construct, although there is suggestion that strain-transcending immunity is achievable [8].

Similar to *PvAMA1*, *PvDBP* is a micronemal protein involved in the invasion of red blood cells. *PvDBP* binds to a receptor called Duffy antigen/receptor for chemokines (DARC) on the reticulocyte surface via a domain termed Region II (RII) [9,10]. *PvDBP* is a member of the EBL family of proteins, which also includes the *P. falciparum* proteins EBA-175, EBA-181 and EBA-140 [11]. Within *P. falciparum*, these EBL family proteins bind to a variety of receptors and facilitate a level of redundancy in the invasion process. However, *PvDBP* is the sole EBL family protein in *P. vivax* and loss of DARC on reticulocytes is protective against *P. vivax* infection, making *PvDBL* an attractive vaccine target [12]. Additionally, population studies [13,14] and *in vitro* work [15,16] have both highlighted the protective effect of anti-*PvDBP* antibody responses.

To gain a further understanding of the distribution of polymorphisms and possible immune-mediated selection pressures over the protein structures of *PvAMA1* and *PvDBP*, these proteins were analysed using the BioStructMap tool outlined in Chapter 4 (and used in Chapter 3). Genomic sequences were extracted from GenBank from a large number of isolates from geographically diverse locations. With these sequences grouped according to geographic location, structural patterns of nucleotide diversity and Tajima's D were examined across all populations. Nucleotide diversity is a measure of the degree of genetic variation across isolates, while Tajima's D is a test

statistic commonly used to identify regions of a gene under immune selection pressure (Tajima's $D > 0$).

Structural patterns of nucleotide diversity were similar across all populations examined, suggesting a commonality in key epitopes targeted across all populations. Domain I of *PvAMA1* was observed to have the highest nucleotide diversity and displayed significant signatures of immune selection pressure. There was no evidence of immune selection pressure on domains II or III of *PvAMA1*. This is in contrast to the results observed for *PfAMA1* in Chapter 3, in which signatures of immune selection pressure were observed on the border of domains II and III. Nucleotide diversity for *PvDBP* was highest bordering the dimerisation and DARC-binding interface, although there was less evidence of immune selection pressure on this antigen.

References

1. Peterson MG, Marshall VM, Smythe JA, Crewther PE, Lew A, Silva A, et al. Integral membrane protein located in the apical complex of *Plasmodium falciparum*. *Mol Cell Biol*. 1989;9: 3151–3154.
2. Healer J, Crawford S, Ralph S, McFadden G, Cowman AF. Independent translocation of two micronemal proteins in developing *Plasmodium falciparum* merozoites. *Infect Immun*. 2002;70: 5751–5758.
3. Lamarque M, Besteiro S, Papoin J, Roques M, Normand BV-L, Morlon-Guyot J, et al. The RON2-AMA1 Interaction is a Critical Step in Moving Junction-Dependent Invasion by Apicomplexan Parasites. *PLoS Pathog*. 2011;7: e1001276.
4. Richard D, MacRaild CA, Riglar DT, Chan J-A, Foley M, Baum J, et al. Interaction between *Plasmodium falciparum* apical membrane antigen 1 and the rhoptry neck protein complex defines a key step in the erythrocyte invasion process of malaria parasites. *J Biol Chem*. 2010;285: 14815–14822.
5. Mugenyi CK, Elliott SR, McCallum FJ, Anders RF, Marsh K, Beeson JG. Antibodies to Polymorphic Invasion-Inhibitory and Non-Inhibitory Epitopes of *Plasmodium falciparum* Apical Membrane Antigen 1 in Human Malaria. *PLoS One*. 2013;8: e68304.
6. 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.
7. Longley RJ, Sattabongkot J, Mueller I. Insights into the naturally acquired immune response to *Plasmodium vivax* malaria. *Parasitology*. 2016;143: 154–170.
8. Drew DR, Hodder AN, Wilson DW, Foley M, Mueller I, Siba PM, et al. Defining the Antigenic Diversity of *Plasmodium falciparum* Apical Membrane Antigen 1 and the Requirements for a Multi-Allele Vaccine against Malaria. *PLoS One*. 2012;7: e51023.
9. Wertheimer SP, Barnwell JW. *Plasmodium vivax* interaction with the human Duffy blood group glycoprotein: identification of a parasite receptor-like protein. *Exp Parasitol*. 1989;69: 340–350.
10. Horuk R, Chitnis CE, Darbonne WC, Colby TJ, Rybicki A, Hadley TJ, et al. A receptor for the malarial parasite *Plasmodium vivax*: the erythrocyte chemokine receptor. *Science*. 1993;261: 1182–1184.
11. Adams JH, Blair PL, Kaneko O, Peterson DS. An expanding ebl family of *Plasmodium falciparum*. *Trends Parasitol*. 2001;17: 297–299.

12. Salinas ND, Tolia NH. Red cell receptors as access points for malaria infection. *Curr Opin Hematol*. 2016;23: 215–223.
13. Cole-Tobian JL, Michon P, Biasor M, Richards JS, Beeson JG, Mueller I, et al. Strain-specific duffy binding protein antibodies correlate with protection against infection with homologous compared to heterologous *Plasmodium vivax* strains in Papua New Guinean children. *Infect Immun*. 2009;77: 4009–4017.
14. Xainli J, Baisor M, Kastens W, Bockarie M, Adams JH, King CL. Age-dependent cellular immune responses to *Plasmodium vivax* Duffy binding protein in humans. *J Immunol*. 2002;169: 3200–3207.
15. Chootong P, Ntumngia FB, VanBuskirk KM, Xainli J, Cole-Tobian JL, Campbell CO, et al. Mapping epitopes of the *Plasmodium vivax* Duffy binding protein with naturally acquired inhibitory antibodies. *Infect Immun*. 2010;78: 1089–1095.
16. Grimberg BT, Udomsangpetch R, Xainli J, McHenry A, Panichakul T, Sattabongkot J, et al. *Plasmodium vivax* invasion of human erythrocytes inhibited by antibodies directed against the Duffy binding protein. *PLoS Med*. 2007;4: e337.

RESEARCH

Open Access



Structural patterns of selection and diversity for *Plasmodium vivax* antigens DBP and AMA1

Andrew J. Guy^{1,2}, Vashti Irani^{1,3}, Jack S. Richards^{1,3,4,5*} and Paul A. Ramsland^{1,2,6,7*}

Abstract

Background: *Plasmodium vivax* is a significant contributor to the global malaria burden, and a vaccine targeting vivax malaria is urgently needed. An understanding of the targets of functional immune responses during the course of natural infection will aid in the development of a vaccine. Antibodies play a key role in this process, with responses against particular epitopes leading to immune selection pressure on these epitopes. A number of techniques exist to estimate levels of immune selection pressure on particular epitopes, with a sliding window analysis often used to determine particular regions likely to be under immune pressure. However, such analysis neglects protein three-dimensional structural information. With this in mind, a newly developed tool, BioStructMap, was applied to two key antigens from *Plasmodium vivax*: PvAMA1 and PvDBP Region II. This tool incorporates structural information into tests of selection pressure.

Results: Sequences from a number of populations were analysed, examining spatially-derived nucleotide diversity and Tajima's D over protein structures for PvAMA1 and PvDBP. Structural patterns of nucleotide diversity were similar across all populations examined, with Domain I of PvAMA1 having the highest nucleotide diversity and displaying significant signatures of immune selection pressure (Tajima's D > 0). Nucleotide diversity for PvDBP was highest bordering the dimerization and DARC-binding interface, although there was less evidence of immune selection pressure on PvDBP compared with PvAMA1. This study supports previous work that has identified Domain I as the main target of immune-mediated selection pressure for PvAMA1, and also supports studies that have identified functional epitopes within PvDBP Region II.

Conclusions: The BioStructMap tool was applied to leading vaccine candidates from *P. vivax*, to examine structural patterns of selection and diversity across a number of geographic populations. There were striking similarities in structural patterns of diversity across multiple populations. Furthermore, whilst regions of high diversity tended to surround conserved binding interfaces, a number of protein regions with very low diversity were also identified, and these may be useful targets for further vaccine development, given previous evidence of functional antibody responses against these regions.

Keywords: *Plasmodium vivax*, Protein structure, Immune selection, Malaria, Population genetics

*Correspondence: jack.richards@burnet.edu.au;

paul.ramsland@rmit.edu.au

¹ Life Sciences, Burnet Institute, 85 Commercial Road, Melbourne, VIC 3004, Australia

⁶ School of Science, RMIT University, Plenty Road, Bundoora, VIC 3083, Australia

Full list of author information is available at the end of the article



© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

Plasmodium vivax infected an estimated 8.55 million people in 2016 and is a significant contributor to global malaria morbidity, with the majority of *P. vivax* cases occurring within South-East Asia [1]. There remains a significant need for a vaccine against *P. vivax*, and an understanding of the targets of natural immune responses following *P. vivax* infection is likely to aid such an effort. A key challenge in vaccine development is the identification of specific antigens and epitopes that are targets of protective antibody responses. It is possible to use population genetic data to identify regions of proteins that are under immune-mediated selection pressure, which gives rise to balancing selection within that protein region. Tajima's D is one test statistic that is often used to identify departures from a neutral model of selection, and has been applied to malaria genes both on a per-gene basis [2, 3], or as a sliding window analysis along a gene [4–7]. A number of studies have previously examined *P. vivax* proteins such as apical membrane antigen 1 (*PvAMA1*) and Duffy-binding protein (*PvDBP*) for evidence of immune selection pressure using these approaches [4, 8–11]. However, because a sliding window analysis is typically performed over the linear gene sequence, it does not take into account the impact of the three-dimensional (3D) structural constraints of the protein in the calculation of selection pressures. A new method that allows incorporation of protein structural information into tests for selection pressure has recently been described [12], and has been applied here to two leading *P. vivax* vaccine candidates: *PvAMA1* and *PvDBP*.

AMA1 is a type I transmembrane protein present in all *Plasmodium* species [13]. It is localized to the parasite micronemes, and is released onto the surface of the merozoite prior to invasion of red blood cells [14]. AMA1 binds to RON2 during the formation of the tight junction between parasite and host–cell membranes [15, 16] and is a target of protective immune responses [17–20]. The ectodomain of *Plasmodium* AMA1 proteins is divided into three domains, termed Domains I (DI), II (DII) and III (DIII) (Additional file 1) [21]. DI is considered to be the most polymorphic, and is also the site of RON2 binding [22]. RON2 binds a conserved hydrophobic cleft that is surrounded by a number of highly polymorphic regions, the most notable being the C1L loop, a surface exposed loop with high variability that is suggested to define strain-specificity in anti-AMA1 responses in *Plasmodium falciparum* infection [23]. While DI is generally considered to be the most important for functional antibody responses, there is evidence that DII and DIII may also be targets of functional antibody responses in *P. falciparum* [8, 24]. A number of studies have investigated selection pressures on *PvAMA1*. Evidence for balancing

selection within Domain I has been observed in a Venezuelan population [9], in a number of Papua New Guinean populations (Madang and Wosera, Madang and East Sepik) [4, 8], an Iranian population [10] and in a Peruvian population [25]. In contrast, two other studies examining *PvAMA1* sequences in isolates from Korea [26] and Myanmar [27] did not find any evidence of balancing selection, but instead observed evidence of recent population bottleneck and expansion in those populations.

PvDBP is an important micronemal protein that binds to the Duffy antigen/receptor for chemokines (DARC) on human reticulocytes during invasion [28, 29]. Whilst there is evidence that *PvDBP* is not absolutely essential for invasion of reticulocytes [30–32], Duffy-negative individuals are largely resistant to *P. vivax* infection, and hence *PvDBP* makes an attractive vaccine target [33]. *PvDBP* is part of the erythrocyte-binding like (EBL) family of proteins, which include EBA175, EBA181 and EBA140 in *P. falciparum* [34]. *PvDBP* is the sole EBL family protein in *P. vivax* [35]. EBL family proteins are composed of a number of distinct domains, with Region II (RII) being a cysteine-rich Duffy-binding like (DBL) domain that is involved in binding to erythrocytes. EBL family proteins each recognize a different receptor via their respective DBL domains [29, 36–38]; *PvDBP* binds to DARC via its DBL domain (RII) [39]. During this process two *PvDBP* molecules form a dimer around two DARC molecules [39, 40]. RII of *PvDBP* has been divided into a number of subdomains (subdomains 1–3) [41] (Additional file 2), and it is subdomain 2 that contains both the dimer interface and DARC binding residues [39]. Immune responses against *PvDBP* have been associated with protection from clinical malaria in naturally exposed cohorts [42, 43] whilst antibodies against *PvDBP* RII epitopes have been found to inhibit both attachment of *PvDBP* RII to erythrocytes [44] and in vitro invasion of erythrocytes [45]. With regards to immune selection pressure on *PvDBP*, a study of 100 Sri Lankan isolates found no evidence of significant selection pressure on this region using Tajima's D, dN/dS or Fu and Li's D and F statistics [11]. Another study examining genetic diversity of *PvDBP* RII across multiple populations showed a significantly positive value of dN/dS in this region, suggesting that this region may be under immune selection pressure [25].

In this study, selection pressures on *PvAMA1* and *PvDBP* Region II were examined in the context of protein structure, using a newly developed tool called BioStructMap [12]. BioStructMap enables the application of a 3D sliding window over a protein structure. This allows incorporation of protein structural information into tests such as Tajima's D or nucleotide diversity that are traditionally performed as a linear 2D sliding window over a

protein or nucleotide sequence. A previous study identified a discontinuous region of *Pf*AMA1 bordering DII and DIII that had a strong signature of balancing selection when considering spatially derived Tajima's D [12]. Given that other studies have identified DI of *Pv*AMA1, rather than DII or DIII, as being under balancing selection, it was considered that incorporation of protein structural information might yield additional insights into other regions under immune selection pressure. Genomic sequences from a number of populations were analysed, and spatially-derived nucleotide diversity and Tajima's D were examined using protein structural information for *Pv*AMA1 and *Pv*DBP Region II. Structural patterns of nucleotide diversity were similar across all populations examined, with Domain I of *Pv*AMA1 having the highest nucleotide diversity and displaying significant signatures of balancing selection (Tajima's $D > 0$). Nucleotide diversity for *Pv*DBP was highest bordering the dimerization and DARC-binding interface, although there was less evidence of immune selection pressure on this antigen.

Methods

Data sources

Reference sequences for Sal-1 *Pv*DBP (PVX_110810) and *Pv*AMA1 (PVX_092275) were obtained from PlasmoDB, v34 (<http://www.plasmoDB.org>) [46]. Genomic sequences from field isolates were extracted from GenBank, restricted to *P. vivax*, and with the condition that sequences had to cover >95% of the structured domains examined here (i.e. partial fragments from these regions were excluded). Sequences from non-human hosts were excluded, as were sequences without a known geographic location described either in the associated literature or clearly annotated in the sequence record. Single isolate populations were also excluded. For *Pv*DBP, bases 766–1659 were used, while for *Pv*AMA1 bases 121–1422 were used.

Sequence polymorphism analysis

The DendroPy Python library [47] was used to calculate Tajima's D, Watterson's Theta, mean pairwise differences and nucleotide diversity. Haplotype diversity was calculated using:

$$H = \frac{n}{n-1} \left(1 - \sum_i x_i^2 \right)$$

where x_i is the relative frequency of the i th haplotype and n is the number of samples.

Normalized Shannon entropy [48] for each position in the protein sequence was calculated using:

$$S = - \sum_i \frac{p_i \log_2 p_i}{\log_2 20}$$

where S is the normalized Shannon entropy and p_i is the frequency at that position of the i th amino acid in the standard 20 amino acid alphabet. The normalized Shannon entropy is a measure of sequence diversity, and takes values between 0 and 1, where 0 indicates perfect sequence conservation at that site, and 1 indicates an even distribution of all possible amino acids at that site.

Phylogenetic analysis

Sequences from all isolates were aligned using MUSCLE v3.8.31 [49], and alignments manually adjusted to minimize gaps. Maximum likelihood phylogenetic trees were constructed for AMA-1 and DBP sequences using IQ-TREE v1.3.11.1 [50]. The ultrafast bootstrap estimation (UFBoot) [51] was used with 5000 bootstrap replicates and a best fit model was chosen according to the Bayesian Inference Criterion [52]. Phylogenetic trees were visualized with iTOL [53].

Incorporation of structural information into sequence analysis

A Python package, BioStructMap, which allows for the application of a 3D sliding window over a protein structure has been previously described [12]. BioStructMap was used to compute spatially derived Tajima's D and nucleotide diversity (π) values for both *Pv*AMA1 and *Pv*DBP structures, with a radius of 15 Å for each window. BioStructMap was also used to calculate Normalized Shannon Entropy on a per-residue basis.

Protein structural models

Known *Pv*DBP structures (i.e. 4NUU, 3RRC, 4YFS, 4NUV, 5F3J) all contain a number of unresolved residues, and to ensure complete coverage of DBP Region II in this analysis, a template-based model of *P. vivax* DBP was generated for use with BioStructMap. This model was created using ModPipe [54], an automated software pipeline that utilizes MODELLER for the generation of comparative protein structure models [55]. The PDB structure 4NUU was used to generate this comparative model. The generated model is accessible via ModBase (<https://modbase.compbio.ucsf.edu/>; ModPipe model ID f7602e019fac5be4a79c4cca6751b392) [56].

The *Pv*AMA1 model used has been previously described [4], and uses a chimeric template to generate a structural model of *Pv*AMA1.

Comparing patterns of selection and diversity between populations

To compare structural patterns of nucleotide diversity and Tajima's *D* between all populations considered in this study, Spearman's rank correlation coefficient was computed for both *PvAMA1* and *PvDBP* residue data between each pair of populations. Residues with missing data in one or both populations (i.e. Tajima's *D* was undefined) were excluded from analysis for that pair of populations.

Data analysis, statistics and other software used

The majority of data analysis was performed using the Anaconda distribution of Python 3.5. Plotting was performed with the Python Matplotlib package, version 1.5.1 [57]. Statistical analysis was performed using SciPy [58]. Protein structures were visualized using PyMol [59].

Results

Population structure of *PvAMA1* and *PvDBP* sequences

This study aimed to examine selection pressures on key structured domains of *PvDBP* and *PvAMA1*. Genomic sequences for each antigen were extracted from GenBank, with a total of 505 *PvAMA1* sequences and 243 *PvDBP* sequences obtained, belonging to 10 and 12 distinct populations, respectively. Sequences which did not cover at least 95% of the structured domains examined (*PvAMA1*, PVX_092275: nucleotides 121–1422; *PvDBP*, PVX_110810: nucleotides 766–1659) were excluded from analysis, as were sequences from single-isolate populations or non-human hosts. Maximum likelihood phylogenetic trees were constructed for both *PvAMA1* and *PvDBP* (Fig. 1) using aligned sequences from all populations. Some populations were generally contained on their own branch (e.g. South Korea for *PvAMA1*, Mexico and Papua New Guinea for *PvDBP*), while other branches contained a mix of populations. This intermixing was particularly evident for populations that are geographically close, such as Thailand, Myanmar and Papua New Guinea for *PvAMA1*.

Traditional measures of selection pressure and diversity for *PvAMA1* and *PvDBP*

Key population parameters for *PvAMA1* and *PvDBP* sequences were computed, as well as several measures of diversity and selection (Tables 1, 2). A total of 259 haplotypes were observed for *PvAMA1* (haplotype diversity = 0.99), while 84 haplotypes were observed for *PvDBP* RII (haplotype diversity = 0.96).

For *PvAMA1*, nucleotide diversity (π) and mean number of pairwise differences (k) were highest in the three Thai populations examined, with a maximum π of 10.09×10^{-3} in the 2007 Tak Province, Thailand samples. Nucleotide diversity was lowest in the South Korean population ($\pi = 5.90 \times 10^{-3}$), which may be explained by a recent bottleneck in this population, limiting overall diversity; this is supported by the negative Tajima's *D* value observed for *PvAMA1* in this South Korean population ($D = -1.20$). Tajima's *D* is often used to identify regions under balancing selection (Tajima's $D > 0$), which can be the result of immune selection pressure, and it is noted that for *PvAMA1*, only the Venezuelan population had a significantly positive Tajima's *D* value ($D = 2.12$, $p < 0.05$), as per the confidence limits outlined in Tajima [60], although these limits may be overly conservative [61]. Most other populations also had a positive Tajima's *D* value, such as Papua New Guinea (Madang: $D = 1.59$; East Sepik: $D = 1.09$) and Thailand (Chantaburi: $D = 1.22$), with the sole exception of South Korea (discussed above).

When examining *PvDBP*, nucleotide diversity was highest in South Korea ($\pi = 10.53 \times 10^{-3}$), Bangkok, Thailand ($\pi = 10.37 \times 10^{-3}$) and Myanmar ($\pi = 10.17 \times 10^{-3}$). Interestingly, the South Korean population had a negative Tajima's *D* value that was statistically significant, suggestive of a recent population bottleneck and expansion. This population also had the highest number of polymorphic sites (58 non-synonymous). This high variability is probably due to 2 or 3 divergent isolates (Fig. 1b), and haplotype diversity is relatively low for this population ($H_d = 0.85$). Tajima's *D* values for *PvDBP* were mostly close to zero, with the highest value observed for a Colombian population ($D = 1.72$).

These observations agree with previous studies that have generally observed greater signatures of immune selection pressure on *PvAMA1* as compared to *PvDBP* [25].

Diversity and selection on the *PvAMA1* structure

We then examined selection pressures and polymorphisms in the context of protein 3D structure using BioStructMap, a tool that allows for incorporation of protein structural information into a sliding window analysis. To quantify nucleotide diversity over the *PvAMA1* structure both as a spatially averaged value and a per-residue value, nucleotide diversity was calculated using a 3D sliding

(See figure on next page.)

Fig. 1 Population structure of *PvAMA1* and *PvDBP* RII sequences. Maximum-likelihood phylogenetic trees are shown for *PvAMA1* (a) and *PvDBP* Region II (b). Leaves are coloured according to the geographic location for each strain. The location of the Sal-1 reference strain is also indicated on each figure

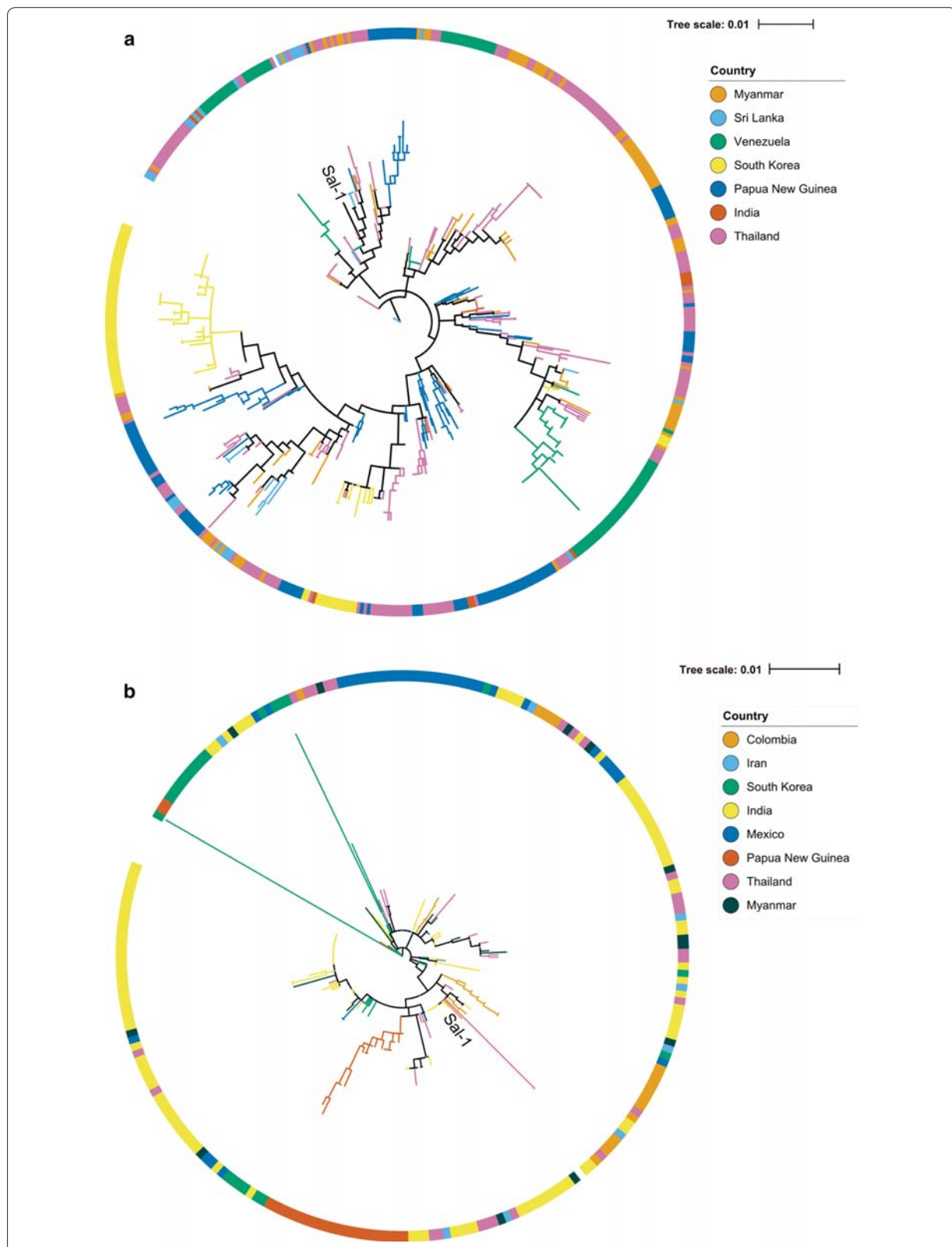


Table 1 Population genetics parameters for PvAMA1 sequences from various geographic locations

Country	n	Number of polymorphic sites	NS sites	S sites	$\pi (\times 10^{-3})$	k	θ	Tajima's D	H	Hd
Myanmar	73	43	31	12	8.55	10.90	8.85	0.75	36	0.935
Sri Lanka	23	34	29	5	7.78	10.13	9.21	0.38	15	0.949
Papua New Guinea (Madang)	61	36	33	4	8.76	11.41	7.69	1.59	51	0.992
Papua New Guinea (East Sepik)	41	34	27	7	8.02	10.45	7.95	1.09	37	0.995
Thailand (Tak; 1996)	58	48	36	12	9.69	12.62	10.37	0.73	52	0.995
Thailand (Tak; 2007)	44	46	33	13	10.09	13.13	10.57	0.85	31	0.982
Thailand (Chantaburi)	56	44	34	10	10.02	13.04	9.58	1.22	25	0.895
India	10	27	22	5	8.31	10.82	9.54	0.64	7	0.911
South Korea	66	57	36	21	5.90	7.68	11.98	-1.20	17	0.921
Venezuela	73	29	22	7	7.66	9.98	5.97	2.12*	17	0.908

n number of isolates, NS sites number of sites with non-synonymous nucleotide polymorphisms, S sites number of sites with synonymous nucleotide polymorphisms, π nucleotide diversity, k mean number of pairwise differences, θ Watterson's theta, H number of haplotypes, Hd haplotype diversity

* $p < 0.05$, indicating rejection of the null hypothesis of a neutral mutation model (confidence limits from Tajima, 1989)

Table 2 Population genetics parameters for PvDBP sequences from various geographic locations

Country	n	Number of polymorphic sites	NS sites	S sites	$\pi (\times 10^{-3})$	k	θ	Tajima's D	H	Hd
Papua New Guinea	23	18	14	5	5.24	4.69	4.88	-0.14	11	0.925
Thailand (Bangkok)	25	46	39	9	10.37	9.01	12.18	-1.00	19	0.977
Colombia	17	14	12	2	6.71	6.00	4.14	1.72	16	0.993
Myanmar	12	32	25	7	10.17	9.09	10.60	-0.64	10	0.970
Mexico	35	13	9	4	3.38	3.02	3.16	-0.14	8	0.556
India (Panna)	20	22	20	2	5.89	5.23	6.20	-0.60	9	0.858
India (Chennai)	20	18	16	2	6.60	5.86	5.07	0.58	8	0.842
India (Delhi)	20	19	16	3	6.64	5.90	5.36	0.38	11	0.889
India (Nadiad)	20	21	17	4	6.54	5.81	5.92	-0.07	10	0.921
India (Kamrup)	20	23	20	3	7.81	6.94	6.48	0.27	12	0.942
Iran	8	21	18	3	8.81	7.57	8.10	-0.34	8	1.000
South Korea	23	74	58	17	10.53	9.36	20.05	-2.12*	11	0.854

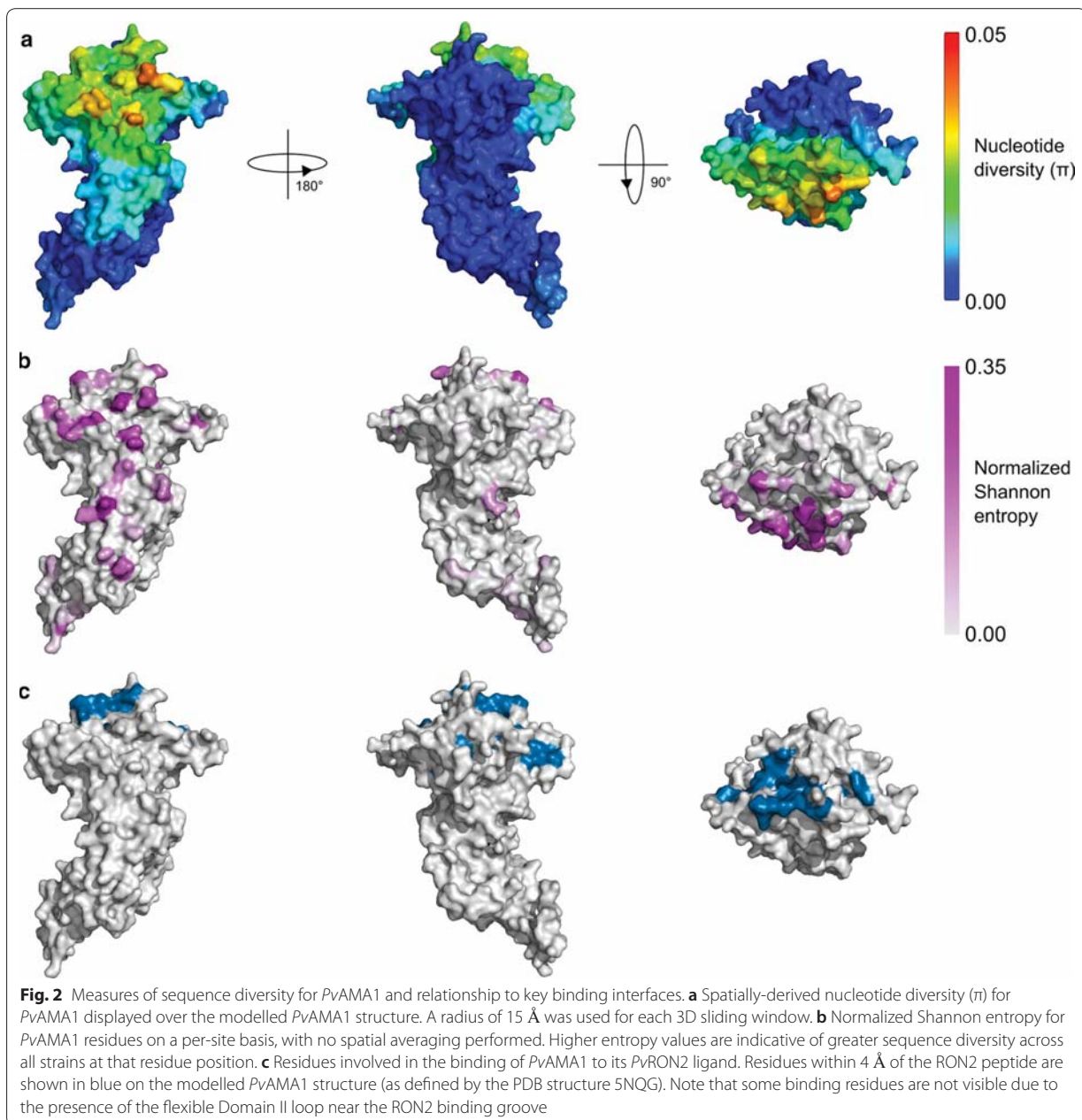
n number of isolates, NS sites number of sites with non-synonymous nucleotide polymorphisms, S sites number of sites with synonymous nucleotide polymorphisms, π nucleotide diversity, k mean number of pairwise differences, θ Watterson's theta, H number of haplotypes, Hd haplotype diversity

* $p < 0.05$, indicating rejection of the null hypothesis of a neutral mutation model (confidence limits from Tajima, 1989)

window (Fig. 2a) and per-residue normalized Shannon entropy (Fig. 2b) using sequences from all populations. Normalized Shannon entropy is a measure of sequence diversity, taking values between 0 and 1, where 0 indicated perfect sequence conservation at that position, while 1 indicates an even distribution of all possible amino acids at that position. For PvAMA1, nucleotide diversity was highest in DI on one side of the RON2 binding cleft (Fig. 2a, c). There was limited diversity within DII, and very little observed in DIII. This pattern of nucleotide diversity appeared to be maintained when examining patterns of nucleotide diversity in individual populations (Additional file 3), with the exception of the

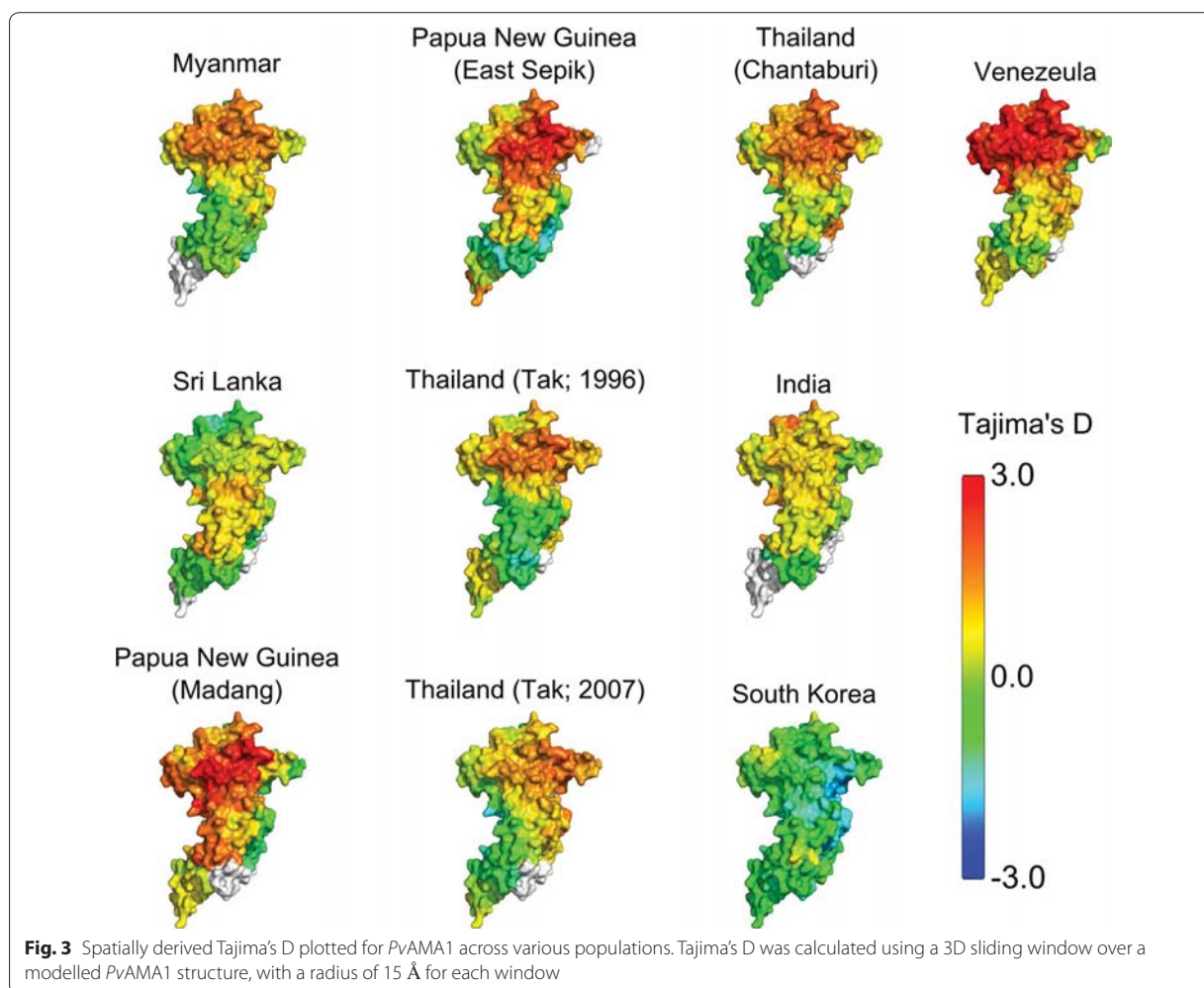
South Korean population, in which even DI displays limited diversity. This is possibly due to recent bottleneck and expansion, as suggested by the negative Tajima's D value for this population. It was also observed that the so-called 'silent face' of PvAMA1 had very low diversity in all populations examined, which is in line with a number of other studies that have noted a distinct lack of polymorphisms on this face both in *P. falciparum* [8, 62–64] and *P. vivax* [4].

To test for evidence of immune-mediated selection pressure on PvAMA1, Tajima's D was calculated as a 3D sliding window over the modelled protein structure (referred to as a spatially-derived Tajima's D). A positive



Tajima's D value provides evidence for balancing selection, which can arise as a result of immune pressure. It is noted that departures from neutrality can also arise as a result of population structure. In particular, sampling of strains across multiple distinct populations can potentially give rise to similar signatures to those of balancing selection. For this reason Tajima's D values were analysed within distinct geographic and temporal populations. Spatially-derived Tajima's D was highest in DI for

nearly all populations examined, with the exception of Sri Lanka and South Korea (Fig. 3; Additional file 4). As was observed for nucleotide diversity, Tajima's D was highest on one side of DI, with the other 'silent-face' typically having Tajima's D values close to zero. This is in agreement with several studies, both in *P. vivax* and *P. falciparum*, in which polymorphisms on AMA1 are focused on one side of the protein structure, with minimal polymorphic variation on the other face [4, 8, 62–64]. It has



been hypothesized that this silent face of the protein is not exposed to the immune system during the invasion process due to interaction with other parasite proteins, or is otherwise inaccessible to antibody binding [63].

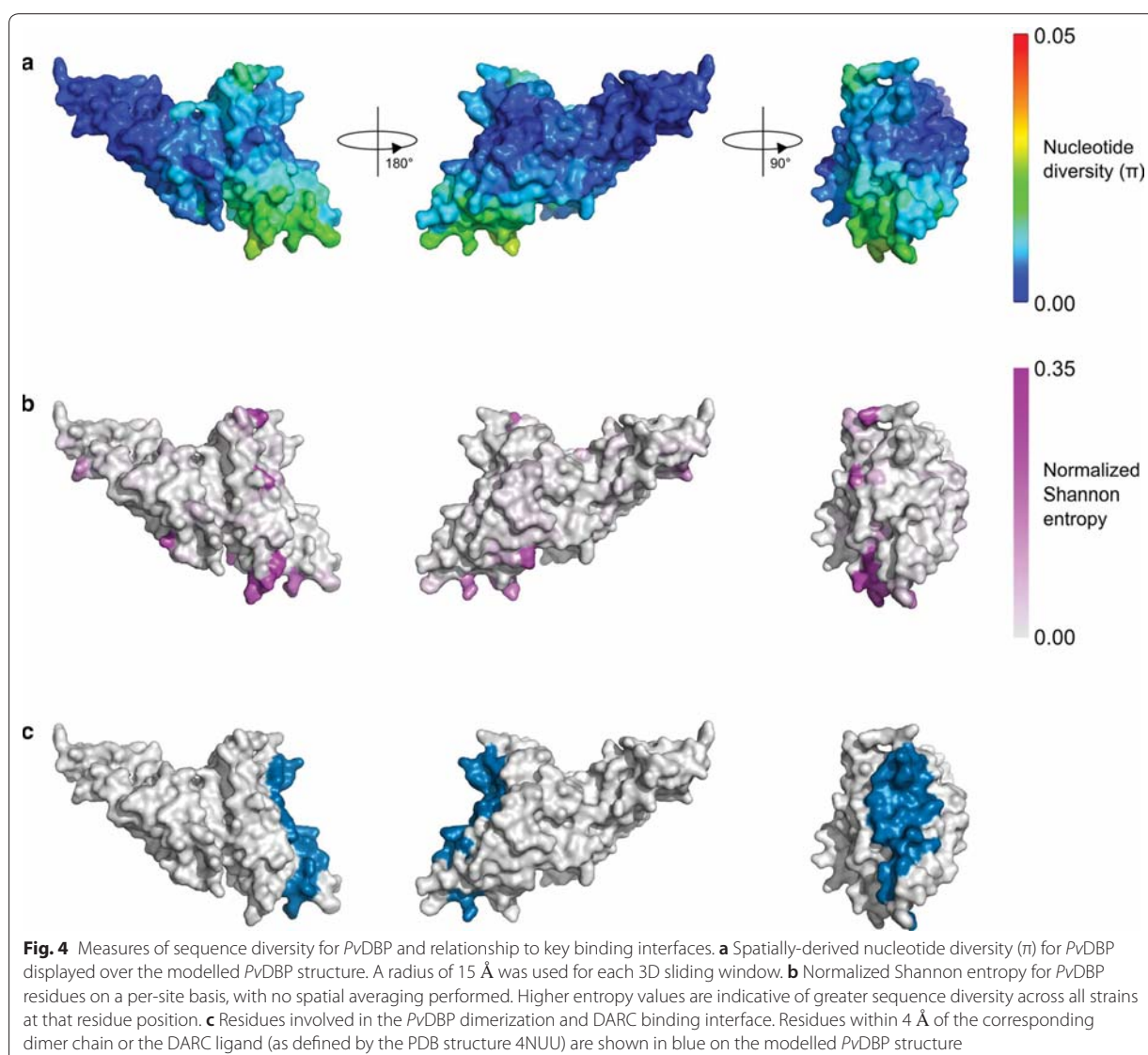
Diversity and selection on the *PvDBP* structure

A similar analysis was performed for *PvDBP*, calculating spatially-derived nucleotide diversity, normalized Shannon entropy and spatially-derived Tajima's D over the modelled *PvDBP* structure. Global nucleotide diversity was highest within subdomain 2 of *PvDBP*, with the most polymorphic residues clustering around the dimerization and DARC binding interface (Fig. 4). Residues that are directly involved in DARC binding and dimerization were nearly all highly conserved. When examining patterns of nucleotide diversity across individual geographic locations (Additional file 5), it is noted that nucleotide diversity within *PvDBP* RII was universally highest in a region that corresponds to a previously identified inhibitory

epitope, termed the DEK epitope [33, 39, 44], which corresponds to residues 338–353 (DEKAQRRKQW-WNESK) of the Sal-1 reference sequence. Correspondingly, the opposite end of the *PvDBP* protein (subdomain 3) had very low nucleotide diversity across all populations. Spatially derived Tajima's D values were also calculated for *PvDBP*, and observed that most populations had Tajima's D values that were negative or close to zero over most of the structure (Fig. 5; Additional file 6); Tajima's D values did not reach statistical significance in any population (with the exception of a single residue in samples from Chennai, India) (Additional file 7). However, it was noted that spatially derived Tajima's D values were generally highest in subdomain 2.

Amino acid mutations within binding interfaces

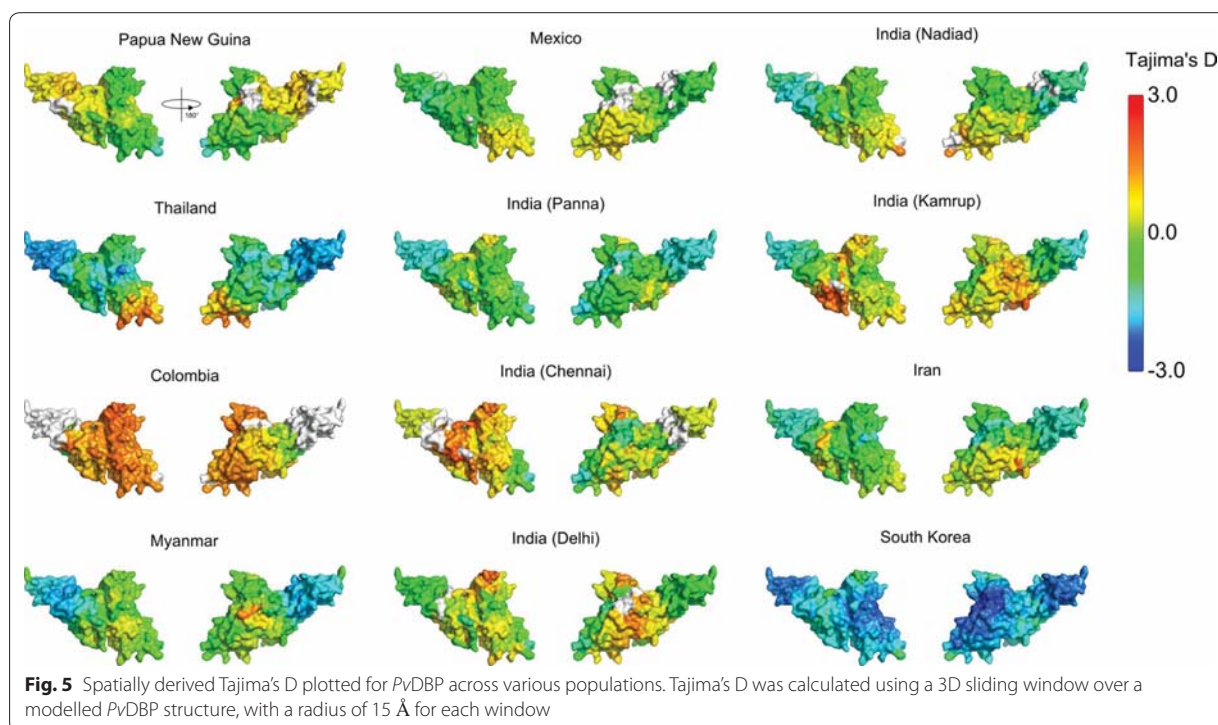
While binding interfaces were generally observed to have low nucleotide diversity, there were some polymorphic residues located within these binding interfaces



for both PvAMA1 and PvDBP (Additional file 8). There were a total of 9 polymorphic residues within the PvAMA1:RON2 interface, although most of these were present at very low frequencies within the global population. The three most frequent polymorphisms within the PvAMA1:RON2 interface fall within a short stretch from residues 130–133 (N132D, N130K, D133N). The most frequent of these polymorphisms (N132D) is a relatively conservative change from asparagine to aspartic acid, and is the only one of the three residues whose side chains are directly involved in hydrogen bonds with RON2; the side chains from both N130 and D133 are only involved in intramolecular hydrogen bonds. It has previously been shown that a PvRON2 peptide binds well

to PfAMA1 [22]. With this in mind, it is noted that the region corresponding to N130–D133 in PvAMA1 is not conserved in PfAMA1, and that this region is also not involved in the binding between PvRON2 and PfAMA1 (PDB structure 5NQF) [22]. These observations, coupled with the observation that a number of polymorphisms fall within this region, suggests that these residues are not a major determinant for binding to PvRON2 and are hence amenable to polymorphic variation.

There were 7 polymorphic residues observed within the PvDBP dimerization and DARC binding interface, although again most of these were present at low frequencies. The highest frequency polymorphism within this region is R263S, with a minor allele frequency of



20%. In the PDB structure 4NUU, the backbone of R263 is involved in hydrogen bonds to DARC, while the side chain is involved in a single intra-molecular hydrogen bond. This residue is also part of a loop which is disordered when not bound to DARC (loop 254–267) [39], suggesting a level of structural plasticity. The other polymorphic residue of note is T359R, which was identified as a major contact within the DARC binding interface by Batchelor et al. [39], although this polymorphism was observed at a relatively low minor allele frequency of 4.1% across all populations. However, this residue is not conserved between *PvDBP* and *PkDBP*; *PkDBP* has an arginine at this position and is also able to bind DARC, supporting the viability of the T359R mutation with regards to maintaining DARC binding ability.

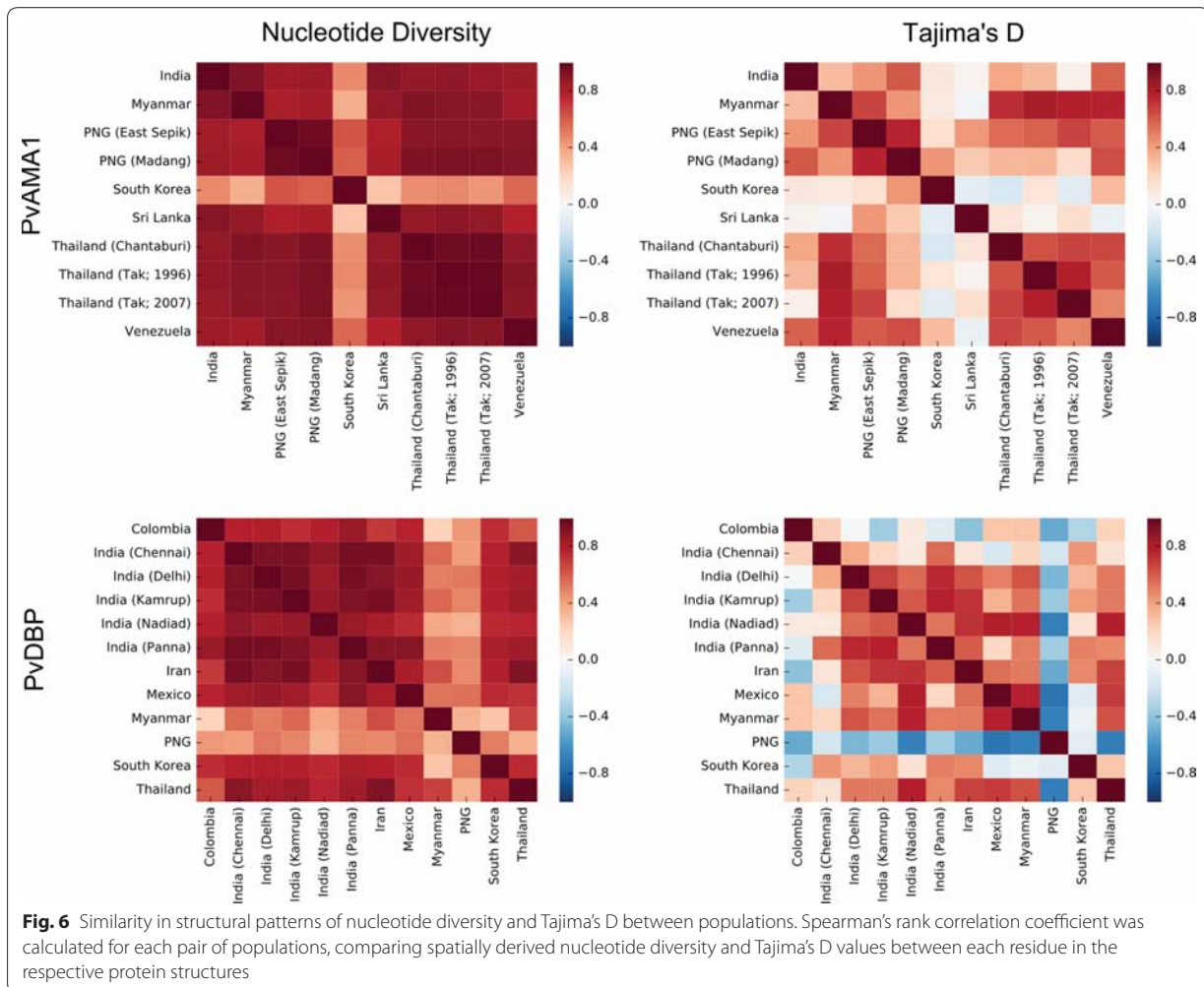
Comparison between 3D and linear sliding window approaches

While a number of previous studies have utilized a linear sliding window approach when calculating Tajima's D or nucleotide diversity [4–7, 65], this study is one of the first to utilize a spatially derived Tajima's D. As such, the Tajima's D values obtained using a 3D sliding window were compared with those obtained using a conventional linear sliding window, using a window size of 102 base pairs and a step size of 3 base pairs for the linear sliding window (Additional files 9, 10). In general there was

broad correspondence between the two approaches for both *PvDBP* RII and *PvAMA1*, although there were some additional regions within *PvAMA1* with Tajima's D values above the threshold for significance when using a 3D sliding window approach that were not identified with a linear sliding window. However, these regions were generally structurally contiguous with other stretches of sequence that were above the threshold for significance when using a linear sliding window.

Comparison of spatially derived nucleotide diversity and Tajima's D between populations

When examining differences in the patterns of nucleotide diversity and Tajima's D values between various populations worldwide, it appeared that most populations had similar structural patterns of diversity/selection, with a few exceptions that could be due to other population effects such as recent bottleneck and expansion events. To quantify the degree of similarity between structural patterns of selection, we computed Spearman's rank correlation coefficient for nucleotide diversity and Tajima's D values between every pair of populations (Fig. 6). Structural patterns of nucleotide diversity were generally highly correlated between populations, with the exception of South Korea for *PvAMA1*, and Myanmar and Papua New Guinea for *PvDBP*. In contrast, there was less agreement in the



structural patterns of Tajima's D over each structure. There was reasonable positive correlation between most populations for spatially derived Tajima's D over the *PvAMA1* structure, with the exception of South Korean and Sri Lankan populations. For *PvDBP*, the Papua New Guinean population was the major outlier when considering spatially derived Tajima's D values, with the highest values observed on subdomain 3, furthest from the dimerization interface. The lower level of apparent immune selection pressure on *PvDBP* may explain the more discordant results observed between populations for *PvDBP*; nearly all Tajima's D values observed for *PvDBP* do not meet the threshold for statistical significance ($p < 0.05$) as defined by Tajima [60] and, therefore, tests of correlation are more sensitive to small variations in Tajima's D. In contrast, a number of populations had significantly positive spatially derived Tajima's D values for *PvAMA1*, including both Papua

New Guinean populations (Madang, East Sepik) and the Venezuelan population (Additional file 7).

Discussion

In this study, patterns of nucleotide diversity and selection were examined over the protein structures for the *P. vivax* antigens *PvAMA1* and *PvDBP*. A number of major observations stand out from this work. Firstly, patterns of diversity on both *PvAMA1* and *PvDBP* were remarkably similar across multiple geographic populations, despite phylogenetic trees for both *PvAMA1* and *PvDBP* sequences suggesting a level of clustering according to geographic location. The only exception to this for *PvAMA1* was the South Korean population which displayed evidence of a recent bottleneck and expansion. This similarity in patterns of diversity is important when trying to extend conclusions made from studies from single geographic locations to a

worldwide population, and these observations suggest a universality with regards to major epitopes on these antigens. It is also interesting to note that highly polymorphic residues for both *PvAMA1* and *PvDBP* tended to fall within regions surrounding, but not a part of, ligand binding interfaces. For *AMA1*, RON2 binds in a hydrophobic cleft in DI, and polymorphic residues fall on one side of this hydrophobic cleft, but generally not within residues that make contact with the RON2 peptide. Similarly with *PvDBP*, contact with DARC occurs primarily via subdomain 2, and the most polymorphic regions were near the DARC binding and dimerization interface. In contrast, the residues directly involved in the *PvDBP* binding interface were highly conserved. These results are suggestive of two things. Firstly, the residues that make up the key binding interfaces in these two antigens have limited capacity for polymorphic variation due to functional constraints, as has been previously suggested by other studies [39, 63]. This makes them attractive vaccine targets, as potential epitopes within these binding sites would have very limited antigenic diversity, and are also less likely to undergo extensive mutations to evade immune responses. Secondly, the high degree of polymorphism around these interfaces suggests that antibody responses that target these polymorphic sites are capable of inhibiting parasite invasion. This inhibition is likely the result of steric hindrance preventing receptor binding and/or dimerization. Future efforts could involve epitope focusing techniques [66, 67] to direct antibody responses to these key conserved interfaces.

For *PvAMA1*, we observed balancing selection primarily on DI in all populations examined. Additionally, while the 3D sliding window approach highlighted additional residues under balancing selection as compared to a linear sliding window approach, nearly all of these regions fell within DI. This agrees with a number of other studies in which *PvAMA* DI is the only domain found to be under significant balancing selection [4, 8–10]. This is in contrast to selection pressures observed on *PfAMA1*, in which both DI and DIII have been observed to be under balancing selection [5–9, 68]. Previous work of ours has applied spatially-derived Tajima's *D* calculations over a *PfAMA1* structure and identified strong balancing selection in a region bordering DII and DIII [12], lending further evidence to DIII being under immune selection pressure in *P. falciparum* but not *P. vivax* *AMA1*. The biological reasons for such a difference are unclear, as *AMA1* has a conserved role between *Plasmodium* species, although it is possible that DIII of *PvAMA1* is less immunogenic than the corresponding *PfAMA1* domain due to structural or sequence differences between the two antigens.

Although individuals in malaria endemic areas develop antibodies to *PvAMA1* [69–71], there are no comprehensive studies on how these antibodies interact with the different domains of *PvAMA1*. However, *AMA1* is functionally conserved across *Plasmodium* species, and there is evidence that *PvAMA1* is functionally equivalent in a *P. falciparum* transgenic line in which *PfAMA1* is replaced by *PvAMA1* [72]. As such, comparisons can be drawn with antibody studies on *PfAMA1*. Dutta et al. [24] generated a panel of monoclonal antibodies (mAbs) to *PfAMA1* and observed that their strain specificity and functional activity was determined by the diversity of the epitope sequence. The limited diversity in *PvAMA1* DIII observed in this current study aligns with the observations that mAbs to *PfAMA1* DIII were the most strain transcending. Similarly, we observed that the polymorphic face of *PvAMA1* DI had the highest diversity, in line with the observation by Dutta et al. that mAbs that bound the polymorphic face of *PfAMA1* DI were strain specific, compared to the others that bound the conserved face [24]. Importantly, mAbs that bound to the conserved face of *PfAMA1* still showed strong growth inhibitory activity suggesting that epitopes on the conserved face can be targets of neutralizing antibodies, despite being under less immune pressure than immunodominant polymorphic regions.

For *PvDBP*, one of the regions of high diversity across all populations was a previously identified epitope within subdomain 2 termed the DEK epitope [33]. Others have observed that the DEK epitope is highly polymorphic and immunodominant [44]. However, due to the polymorphic nature of this epitope, cross-strain specificity is a concern when creating a *PvDBP* RII-based vaccine. Recent work has characterized the location of several conserved epitopes within *PvDBP* RII that are the target of inhibitory mAbs 2D10, 2H2 and 2C6 [73, 74], and all of these conserved epitopes fall within subdomain III, which had the lowest overall nucleotide diversity in this current study. This highlights some of the limitations of using population-level genomic data for identification of functionally important targets of antibody responses—the possibility that conserved regions may contain potential epitopes that are the targets of inhibitory antibodies cannot be excluded.

Several attempts have been made to divert immune responses away from these highly polymorphic regions of *PvDBP* RII and towards conserved epitopes. One attempt to focus away from polymorphic regions involved mutating residues in the DEK epitope to reduce its immunogenicity, and these DEKnull mutants induce antibodies that bind *PvDBP* and can inhibit the interaction with Duffy Binding Ligand [33]. More recently, further epitope focusing techniques have been employed with *PvDBP*

RII, with one strategy involving mutation of all polymorphic residues to alanine, threonine or serine residues [75]. This 'DEKnull-2' recombinant PvDBP RII construct was shown to elicit broadly neutralizing antibodies following mouse immunization, with some naturally exposed individuals also shown to recognize the conserved epitopes on this construct [75]. Other recent efforts towards the development of a PvDBP RII vaccine include a Phase 1a trial of a prime-boost viral-vectored vaccine that demonstrated both safety and immunogenicity, with cross-strain inhibition demonstrated for the single heterologous strain tested [76].

Given the success using epitope focusing techniques for PvDBP RII, we suggest that such an approach could be applied to AMA1 to focus antibody responses towards conserved epitopes on the silent face of DI or within DIII. This would involve mutating major polymorphic residues to reduce the immunodominance of epitopes within the polymorphic face of DI; the most polymorphic residues identified in this study (Additional file 11) could serve as starting point for this work. Alternatively, it has been shown for PfAMA1 that immunization with multiple heterologous strains of PfAMA1 is capable of inducing strain transcending antibody responses [24], and this approach could also be applied to PvAMA1. Other epitope-focusing approaches also exist, including the use of small protein scaffolds to mimic native epitopes [66, 67], although these might be challenging given the discontinuous nature of many potential epitopes within AMA1. The approaches used in this work could also be applied to other antigens such as PvRBP2b, which has recently been identified as a ligand for reticulocyte invasion via binding to transferrin receptor 1 (TfR1) [77].

Conclusions

In this work, signatures of diversity and selection were identified on PvAMA1 and PvDBP, and it was shown that the regions of high diversity and balancing selection on the protein structure are remarkably similar across a number of populations. This suggests that dominant epitopes are the same across multiple human populations, which has positive implications for the development of a universal *P. vivax* vaccine. Furthermore, polymorphisms were observed to cluster around binding interfaces on both PvDBP and PvAMA1, suggesting a level of immune pressure on residues surrounding these key binding interfaces. Large regions with very low diversity were also identified for both antigens, and it is suggested that these areas may also be useful targets to focus on for further vaccine development, given previous evidence of functional antibody responses against these conserved regions.

Additional files

Additional file 1. Domains/subdomains of PvAMA1.

Additional file 2. Domains/subdomains of PvDBP RII.

Additional file 3. Spatially-derived nucleotide diversity for PvAMA1 across multiple populations.

Additional file 4. Spatially-derived Tajima's D for PvAMA1 across multiple populations.

Additional file 5. Spatially-derived nucleotide diversity for PvDBP across multiple populations.

Additional file 6. Spatially-derived Tajima's D for PvDBP across multiple populations.

Additional file 7. Location of statistically significant ($p < 0.05$) Tajima's D values on modelled PvAMA1 (a) and PvDBP (b) structures.

Additional file 8. Sequence diversity at interfaces.

Additional file 9. Comparison of spatially derived Tajima's D and conventional linear sliding window calculation of Tajima's D for PvAMA1.

Additional file 10. Comparison of spatially derived Tajima's D and conventional linear sliding window calculation of Tajima's D for PvDBP RII.

Additional file 11. Table of highly polymorphic residues within PvAMA1.

Authors' contributions

AJG designed the study, performed the research, analysed the results and wrote the manuscript; PAR contributed the modelled PvAMA1 structure; VI, JSR and PAR provided critical interpretation of the data. All authors revised, commented and read the manuscript. All authors read and approved the final manuscript.

Author details

¹ Life Sciences, Burnet Institute, 85 Commercial Road, Melbourne, VIC 3004, Australia. ² Department of Immunology, Monash University, Melbourne, Australia. ³ Department of Medicine, University of Melbourne, Melbourne, Australia. ⁴ Department of Infectious Diseases, Monash University, Melbourne, Australia. ⁵ Victorian Infectious Diseases Service, Royal Melbourne Hospital, Melbourne, Australia. ⁶ School of Science, RMIT University, Plenty Road, Bundoora, VIC 3083, Australia. ⁷ Department of Surgery Austin Health, University of Melbourne, Heidelberg, Australia.

Acknowledgements

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Availability of data and materials

The datasets used and/or analysed during the current study are available from the corresponding author on reasonable request.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Funding

Funding was provided by the National Health and Medical Research Council (NHMRC) of Australia including an Early Career Research Fellowship (APP1037722) and Project Grant (APP1125788) to JSR; and an Australian Postgraduate Award to support AJG through Monash University. Burnet Institute received funding from the NHMRC Independent Research Institutes Infrastructure Support Scheme, and the Victorian State Government Operational Infrastructure Support Scheme.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Received: 12 January 2018 Accepted: 18 April 2018

Published online: 02 May 2018

References

- WHO. World Malaria Report 2017. Geneva: World Health Organization; 2017. <http://www.who.int/malaria/publications/world-malaria-report-2017/en/> Accessed 21 Dec 2017.
- 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;31:1490–9.
- Amambua-Ngwa A, Tetteh KKA, Manske M, Gomez-Escobar N, Stewart LB, Elizabeth Deerhake M, et al. Population genomic scan for candidate signatures of balancing selection to guide antigen characterization in malaria parasites. *PLoS Genet*. 2012;8:e1002992.
- Arnott A, Mueller I, Ramsland PA, Siba PM, Reeder JC, Barry AE. Global population structure of the genes encoding the malaria vaccine candidate, *Plasmodium vivax* apical membrane antigen 1 (Pv AMA1). *PLoS Negl Trop Dis*. 2013;7:e2506.
- Osier FHA, Weedall GD, Verra F, Murungi L, Tetteh KKA, Bull P, et al. Allelic diversity and naturally acquired allele-specific antibody responses to *Plasmodium falciparum* apical membrane antigen 1 in Kenya. *Infect Immun*. 2010;78:4625–33.
- Polley SD, Chokejindachai W, Conway DJ. Allele frequency-based analyses robustly map sequence sites under balancing selection in a malaria vaccine candidate antigen. *Genetics*. 2003;165:555–61.
- Polley SD, Conway DJ. Strong diversifying selection on domains of the *Plasmodium falciparum* apical membrane antigen 1 gene. *Genetics*. 2001;158:1505–12.
- 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.
- Ord RL, Tami A, Sutherland CJ. Ama1 genes of sympatric *Plasmodium vivax* and *P. falciparum* from Venezuela differ significantly in genetic diversity and recombination frequency. *PLoS ONE*. 2008;3:e3366.
- Zakeri S, Sadeghi H, Mehrizi AA, Djadid ND. Population genetic structure and polymorphism analysis of gene encoding apical membrane antigen-1 (AMA-1) of Iranian *Plasmodium vivax* wild isolates. *Acta Trop*. 2013;126:269–79.
- Premaratne PH, Aravinda BR, Escalante AA, Udagama PV. Genetic diversity of *Plasmodium vivax* Duffy Binding Protein II (PvDBPII) under unstable transmission and low intensity malaria in Sri Lanka. *Infect Genet Evol*. 2011;11:1327–39.
- 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.
- Peterson MG, Marshall VM, Smythe JA, Crewther PE, Lew A, Silva A, et al. Integral membrane protein located in the apical complex of *Plasmodium falciparum*. *Mol Cell Biol*. 1989;9:3151–4.
- Healer J, Crawford S, Ralph S, McFadden G, Cowman AF. Independent translocation of two micronemal proteins in developing *Plasmodium falciparum* merozoites. *Infect Immun*. 2002;70:5751–8.
- Lamarque M, Besteiro S, Papoin J, Roques M, Normand BV-L, Morlon-Guyot J, et al. The RON2-AMA1 interaction is a critical step in moving junction-dependent invasion by apicomplexan parasites. *PLoS Pathog*. 2011;7:e1001276.
- Richard D, MacRaild CA, Riglar DT, Chan J-A, Foley M, Baum J, et al. Interaction between *Plasmodium falciparum* apical membrane antigen 1 and the rhoptry neck protein complex defines a key step in the erythrocyte invasion process of malaria parasites. *J Biol Chem*. 2010;285:14815–22.
- 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 USA*. 2011;108:13275–80.
- Mugenyi CK, Elliott SR, McCallum FJ, Anders RF, Marsh K, Beeson JG. Antibodies to polymorphic invasion-inhibitory and non-inhibitory epitopes of *Plasmodium falciparum* apical membrane antigen 1 in human malaria. *PLoS ONE*. 2013;8:e68304.
- Hodder AN, Crewther PE, Anders RF. Specificity of the protective antibody response to apical membrane antigen 1. *Infect Immun*. 2001;69:3286–94.
- Fowkes FJL, 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.
- Pizarro JC, Vulliez-Le Normand B, Chesne-Seck M-L, Collins CR, Withers-Martinez C, Hackett F, et al. Crystal structure of the malaria vaccine candidate apical membrane antigen 1. *Science*. 2005;308:408–11.
- Vulliez-Le Normand B, Saul FA, Hoos S, Faber BW, Bentley GA. Cross-reactivity between apical membrane antigen 1 and rhoptry neck protein 2 in *P. vivax* and *P. falciparum*: a structural and binding study. *PLoS ONE*. 2017;12:e0183198.
- Dutta S, Dlugosz LS, Clayton JW, Pool CD, Haynes JD, Gasser RA 3rd, et al. Alanine mutagenesis of the primary antigenic escape residue cluster, c1, of apical membrane antigen 1. *Infect Immun*. 2010;78:661–71.
- Dutta S, Dlugosz LS, Drew DR, Ge X, Ge X, Ababacar D, et al. Overcoming antigenic diversity by enhancing the immunogenicity of conserved epitopes on the malaria vaccine candidate apical membrane antigen-1. *PLoS Pathog*. 2013;9:e1003840.
- Chenet SM, Tapia LL, Escalante AA, Durand S, Lucas C, Bacon DJ. Genetic diversity and population structure of genes encoding vaccine candidate antigens of *Plasmodium vivax*. *Malar J*. 2012;11:68.
- Kang J-M, Lee J, Cho P-Y, Moon S-U, Ju H-L, Ahn SK, et al. Population genetic structure and natural selection of apical membrane antigen-1 in *Plasmodium vivax* Korean isolates. *Malar J*. 2015;14:455.
- Moon S-U, Na B-K, Kang J-M, Kim J-Y, Cho S-H, Park Y-K, et al. Genetic polymorphism and effect of natural selection at domain I of apical membrane antigen-1 (AMA-1) in *Plasmodium vivax* isolates from Myanmar. *Acta Trop*. 2010;114:71–5.
- Wertheimer SP, Barnwell JW. *Plasmodium vivax* interaction with the human Duffy blood group glycoprotein: identification of a parasite receptor-like protein. *Exp Parasitol*. 1989;69:340–50.
- Horuk R, Chitnis CE, Darbonne WC, Colby TJ, Rybicki A, Hadley TJ, et al. A receptor for the malarial parasite *Plasmodium vivax*: the erythrocyte chemokine receptor. *Science*. 1993;261:1182–4.
- Cavasini CE, de Mattos LC, Couto AAD, Bonini-Domingos CR, Valencia SH, de Neiras WC, et al. *Plasmodium vivax* infection among Duffy antigen-negative individuals from the Brazilian Amazon region: an exception? *Trans R Soc Trop Med Hyg*. 2007;101:1042–4.
- Ménard D, Barnadas C, Bouchier C, Henry-Halldin C, Gray LR, Ratsimbasa A, et al. *Plasmodium vivax* clinical malaria is commonly observed in Duffy-negative Malagasy people. *Proc Natl Acad Sci USA*. 2010;107:5967–71.
- Ntumngia FB, Thomson-Luque R, de Torres LM, Gunalan K, Carvalho LH, Adams JH. A novel erythrocyte binding protein of *Plasmodium vivax* suggests an alternate invasion pathway into Duffy-positive reticulocytes. *MBio*. 2016;7:e011261.
- Ntumngia FB, Adams JH. Design and immunogenicity of a novel synthetic antigen based on the ligand domain of the *Plasmodium vivax* Duffy binding protein. *Clin Vaccine Immunol*. 2012;19:30–6.
- Adams JH, Blair PL, Kaneko O, Peterson DS. An expanding EBL family of *Plasmodium falciparum*. *Trends Parasitol*. 2001;17:297–9.
- Carlton JM, Adams JH, Silva JC, Bidwell SL, Lorenzi H, Caler E, et al. Comparative genomics of the neglected human malaria parasite *Plasmodium vivax*. *Nature*. 2008;455:757–63.
- Sim BK, Chitnis CE, Wasniowska K, Hadley TJ, Miller LH. Receptor and ligand domains for invasion of erythrocytes by *Plasmodium falciparum*. *Science*. 1994;264:1941–4.
- Maier AG, Duraisingh MT, Reeder JC, Patel SS, Kazura JW, Zimmerman PA, et al. *Plasmodium falciparum* erythrocyte invasion through glycophorin C and selection for Gerbich negativity in human populations. *Nat Med*. 2003;9:87–92.
- Mayer DCG, Cofie J, Jiang L, Hartl DL, Tracy E, Kabat J, et al. Glycophorin B is the erythrocyte receptor of *Plasmodium falciparum* erythrocyte-binding ligand, EBL-1. *Proc Natl Acad Sci USA*. 2009;106:5348–52.

39. Batchelor JD, Malpede BM, Omattage NS, DeKoster GT, Henzler-Wildman KA, Tolia NH. Red blood cell invasion by *Plasmodium vivax*: structural basis for DBP engagement of DARC. *PLoS Pathog*. 2014;10:e1003869.
40. Batchelor JD, Zahm JA, Tolia NH. Dimerization of *Plasmodium vivax* DBP is induced upon receptor binding and drives recognition of DARC. *Nat Struct Mol Biol*. 2011;18:908–14.
41. Singh SK, Hora R, Belhali H, Chitnis CE, Sharma A. Structural basis for Duffy recognition by the malaria parasite Duffy-binding-like domain. *Nature*. 2006;439:741–4.
42. Cole-Tobian JL, Michon P, Biasor M, Richards JS, Beeson JG, Mueller I, et al. Strain-specific Duffy binding protein antibodies correlate with protection against infection with homologous compared to heterologous *Plasmodium vivax* strains in Papua New Guinean children. *Infect Immun*. 2009;77:4009–17.
43. Xainli J, Baisor M, Kastens W, Bockarie M, Adams JH, King CL. Age-dependent cellular immune responses to *Plasmodium vivax* Duffy binding protein in humans. *J Immunol*. 2002;169:3200–7.
44. Chootong P, Ntumngia FB, VanBuskirk KM, Xainli J, Cole-Tobian JL, Campbell CO, et al. Mapping epitopes of the *Plasmodium vivax* Duffy binding protein with naturally acquired inhibitory antibodies. *Infect Immun*. 2010;78:1089–95.
45. Grimberg BT, Udomsangpetch R, Xainli J, McHenry A, Panichakul T, Sattabongkot J, et al. *Plasmodium vivax* invasion of human erythrocytes inhibited by antibodies directed against the Duffy binding protein. *PLoS Med*. 2007;4:e337.
46. Aurrecoechea C, Brestelli J, Brunk BP, Dommer J, Fischer S, Gajria B, et al. PlasmoDB: a functional genomic database for malaria parasites. *Nucleic Acids Res*. 2009;37:D539–43.
47. Sukumaran J, Holder MT. DendroPy: a Python library for phylogenetic computing. *Bioinformatics*. 2010;26:1569–71.
48. Shannon CE. A mathematical theory of communication. *Bell Syst Tech J*. 1948;27:379–423.
49. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
50. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol*. 2015;32:268–74.
51. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Le SV. UFBoot2: improving the ultrafast bootstrap approximation. *Mol Biol Evol*. 2018;35:518–22.
52. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermini LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods*. 2017;14:587–9.
53. Letunic I, Bork P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res*. 2016;44:W242–5.
54. Eswar N, John B, Mirkovic N, Fiser A, Ilyin VA, Pieper U, et al. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res*. 2003;31:3375–80.
55. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. *J Mol Biol*. 1993;234:779–815.
56. Pieper U, Webb BM, Barkan DT, Schneidman-Duhovny D, Schlessinger A, Braberg H, et al. ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Res*. 2011;39:D465–74.
57. Hunter JD. Matplotlib: a 2D graphics environment. *Comput Sci Eng*. 2007;9:90–5.
58. van der Walt S, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. *Comput Sci Eng*. 2011;13:22–30.
59. Schrödinger LLC. The PyMOL molecular graphics system, version 1.8. 2015.
60. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics*. 1989;123:585–95.
61. Fu YX, Li WH. Statistical tests of neutrality of mutations. *Genetics*. 1993;133:693–709.
62. Faber BW, Kadir KA, Rodriguez-Garcia R, Remarque EJ, Saul FA, Normand BV-L, et al. Low levels of polymorphisms and no evidence for diversifying selection on the *Plasmodium knowlesi* apical membrane antigen 1 Gene. *PLoS ONE*. 2015;10:e0124400.
63. 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 USA*. 2005;102:12736–41.
64. Takala SL, Coulbaly 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.
65. Escalante AA, Grebert HM, Chaiyaroj SC, Magris M, Biswas S, Nahlen BL, et al. Polymorphism in the gene encoding the apical membrane antigen-1 (AMA-1) of *Plasmodium falciparum*. *X. Asembo Bay Cohort Project. Mol Biochem Parasitol*. 2001;113:279–87.
66. Oscherwitz J. The promise and challenge of epitope-focused vaccines. *Hum Vaccin Immunother*. 2016;12:2113–6.
67. Correia BE, Bates JT, Loomis RJ, Baneyx G, Carrico C, Jardine JG, et al. Proof of principle for epitope-focused vaccine design. *Nature*. 2014;507:201.
68. Mehrizi AA, Sepehri M, Karimi F, Djadid ND, Zakeri S. Population genetics, sequence diversity and selection in the gene encoding the *Plasmodium falciparum* apical membrane antigen 1 in clinical isolates from the south-east of Iran. *Infect Genet Evol*. 2013;17:51–61.
69. Fowkes FJL, McGready R, Cross NJ, Hommel M, Simpson JA, Elliott SR, et al. New insights into acquisition, boosting, and longevity of immunity to malaria in pregnant women. *J Infect Dis*. 2012;206:1612–21.
70. Cutts JC, Powell R, Agius PA, Beeson JG, Simpson JA, Fowkes FJL. Immunological markers of *Plasmodium vivax* exposure and immunity: a systematic review and meta-analysis. *BMC Med*. 2014;12:150.
71. McLean ARD, Boel M, McGready R, Aitaie R, Drew D, Tsuboi T, et al. Antibody responses to *Plasmodium falciparum* and *Plasmodium vivax* and prospective risk of *Plasmodium* spp. infection postpartum. *Am J Trop Med Hyg*. 2017;96:1197–204.
72. Drew DR, Sanders PR, Weiss G, Gilson PR, Crabb BS, Beeson JG. Functional conservation of the AMA1 host-cell invasion ligand between *P. falciparum* and *P. vivax*: a novel platform to accelerate vaccine and drug development. *J Infect Dis*. 2018;217:498–507.
73. Chen E, Salinas ND, Huang Y, Ntumngia F, Plasencia MD, Gross ML, et al. Broadly neutralizing epitopes in the *Plasmodium vivax* vaccine candidate Duffy Binding Protein. *Proc Natl Acad Sci USA*. 2016;113:6277–82.
74. Ntumngia FB, Schloegel J, Barnes SJ, McHenry AM, Singh S, King CL, et al. Conserved and variant epitopes of *Plasmodium vivax* Duffy binding protein as targets of inhibitory monoclonal antibodies. *Infect Immun*. 2012;80:1203–8.
75. Ntumngia FB, Pires CV, Barnes SJ, George MT, Thomson-Luque R, Kano FS, et al. An engineered vaccine of the *Plasmodium vivax* Duffy binding protein enhances induction of broadly neutralizing antibodies. *Sci Rep*. 2017;7:13779.
76. Payne RO, Silk SE, Elias SC, Milne KH, Rawlinson TA, Llewellyn D, et al. Human vaccination against *Plasmodium vivax* Duffy-binding protein induces strain-transcending antibodies. *JCI Insight*. 2017;2:93683 (**Epub ahead of print**).
77. Gruszczyk J, Kanjee U, Chan L-J, Menant S, Malleret B, Lim NTY, et al. Transferrin receptor 1 is a reticulocyte-specific receptor for *Plasmodium vivax*. *Science*. 2018;359:48–55.

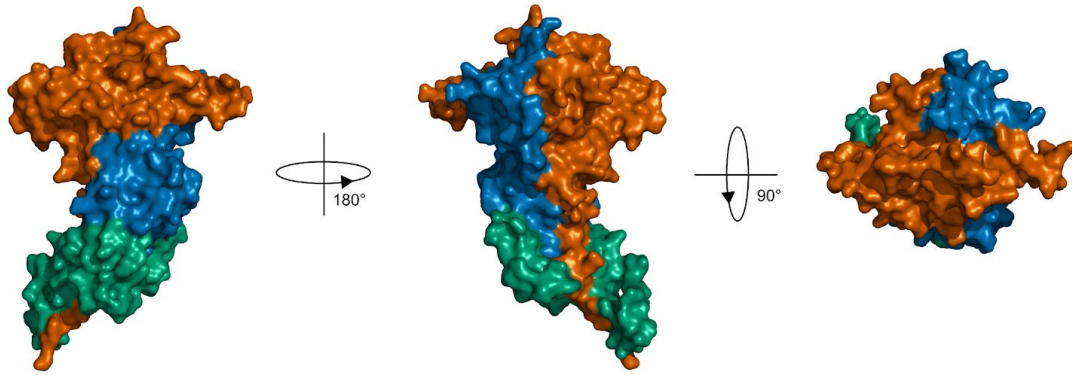
Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

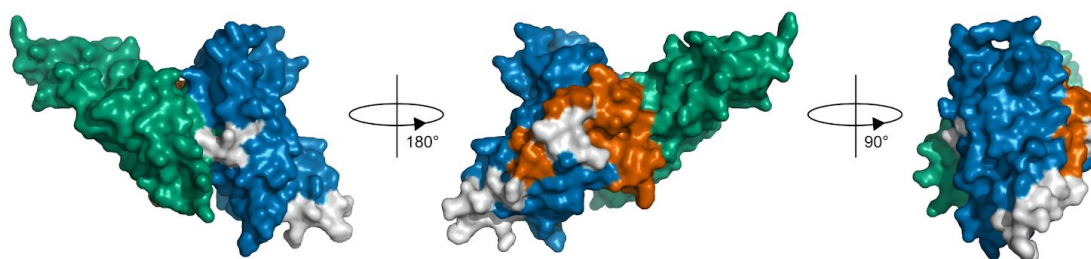
At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

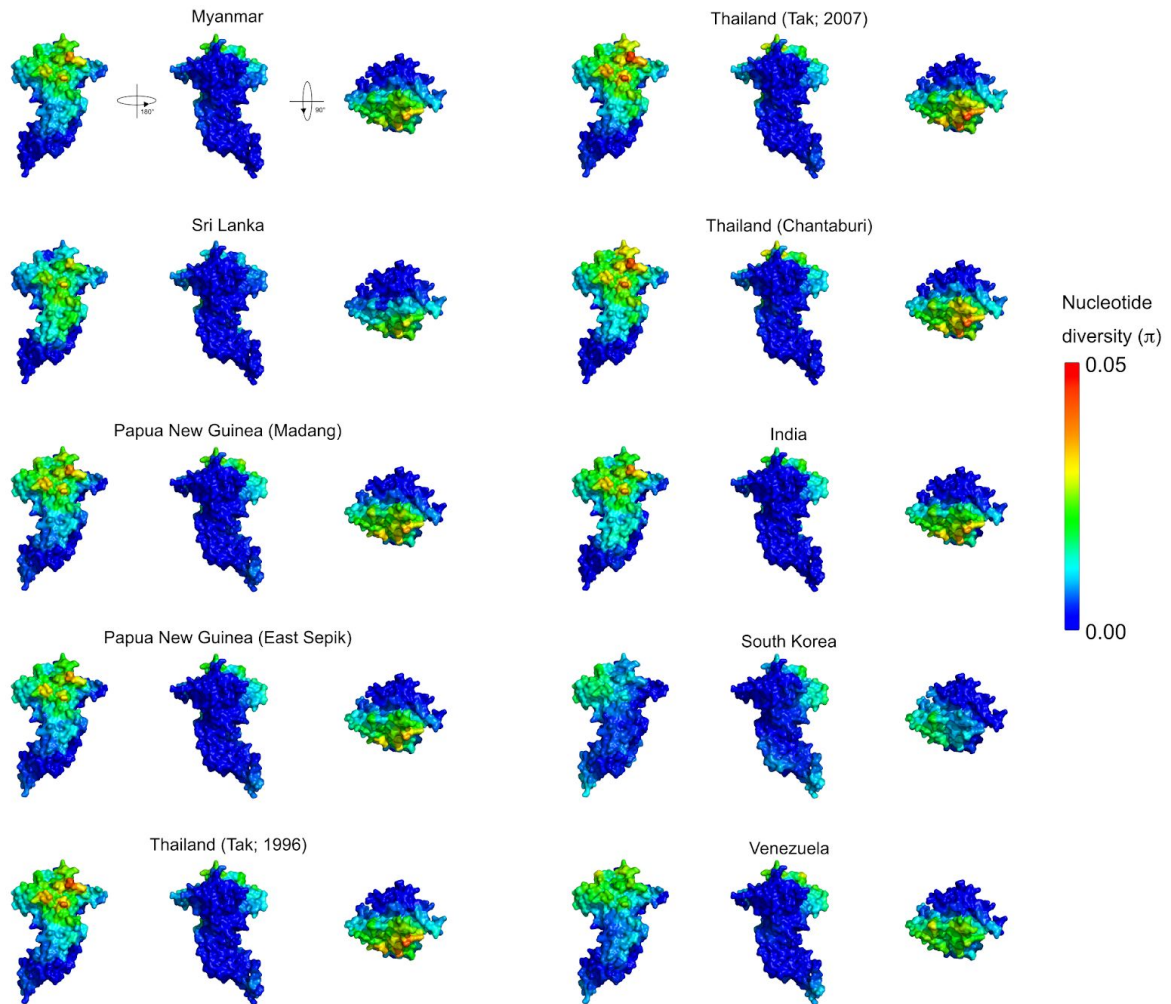




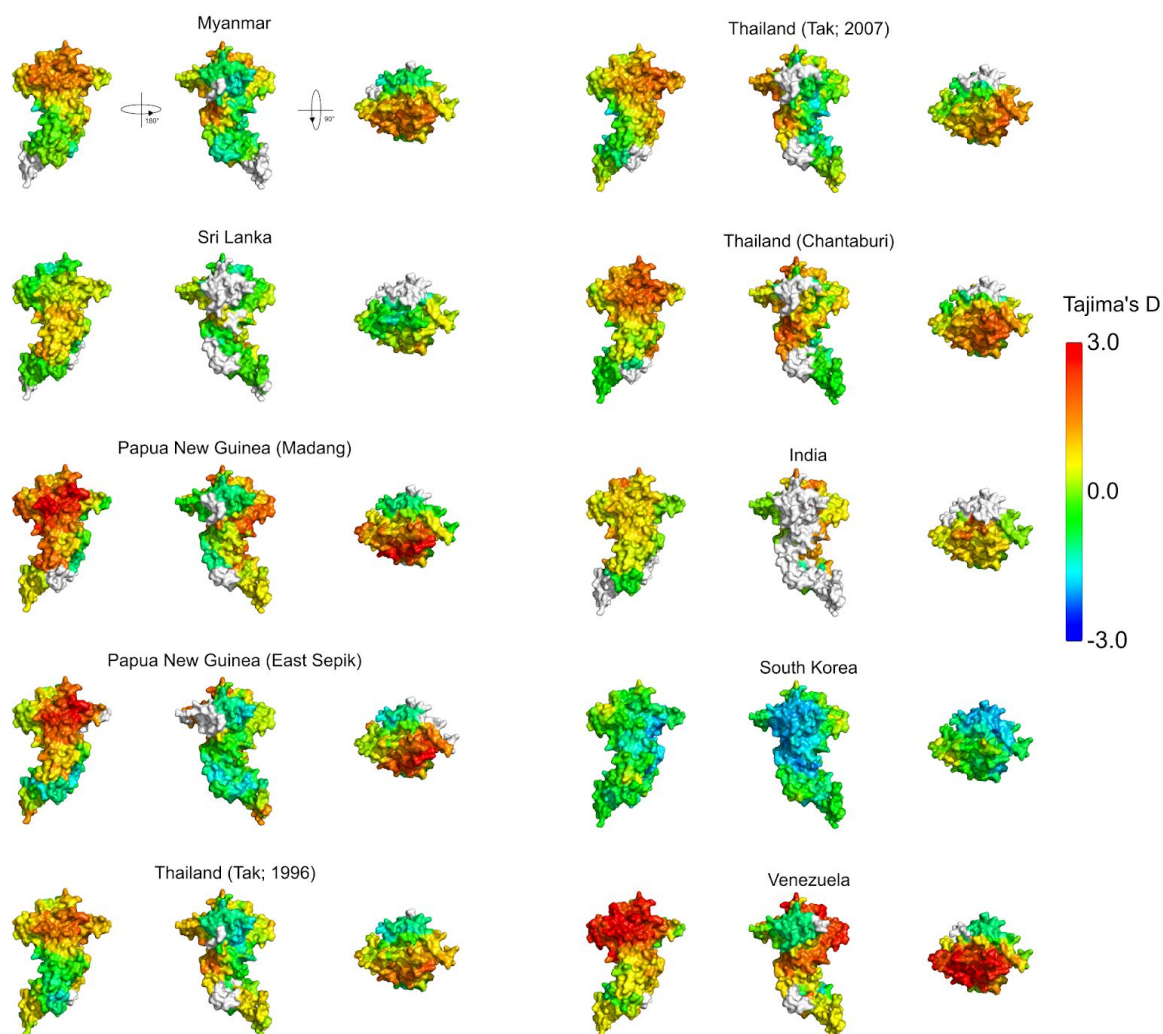
Additional File 1: Domains/subdomains of *PvAMA1*. *PvAMA1* has been divided into three domains, termed Domain I (DI), II (DII) and III (DIII), shown in orange, blue and green, respectively. Domain assignment follows that outlined by Pizarro *et al.* [21], and corresponds to the following residues in the Sal-1 reference sequence: 41-250 (DI); 251-387 (DII); 388-474 (DIII).



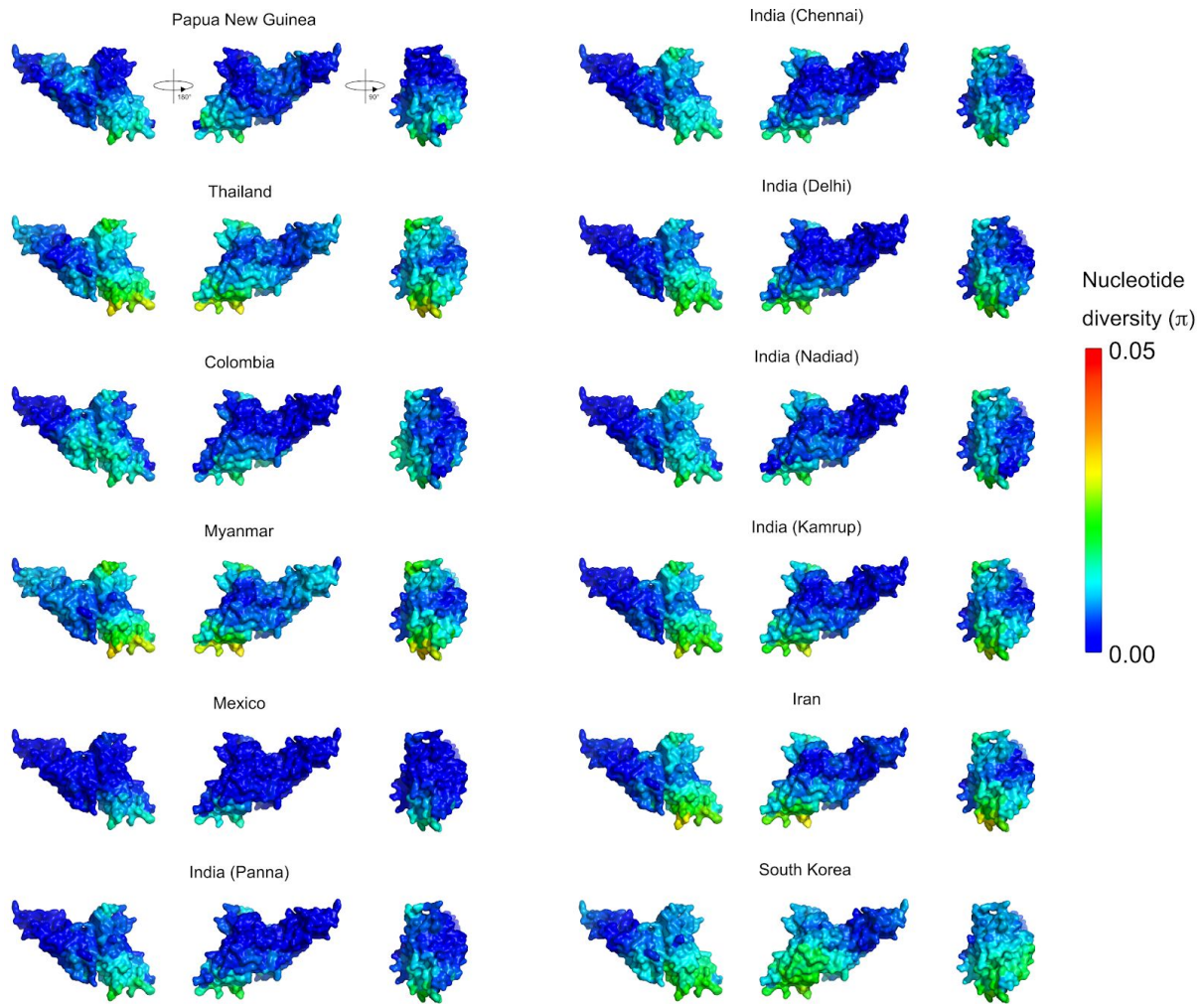
Additional File 2: Domains/subdomains of *PvDBP* RII. *PvDBP* RII has been divided into three subdomains, indicated in orange (subdomain 1), blue (subdomain 2) and green (subdomain 3). Subdomain assignment follows that outlined by Singh *et al.* [41], corresponding to the following residues in the Sal-1 reference sequence: 216-253 (subdomain 1); 265-381 (subdomain 2); 387-508 (subdomain 3). Linkers between subdomains are shown in white.



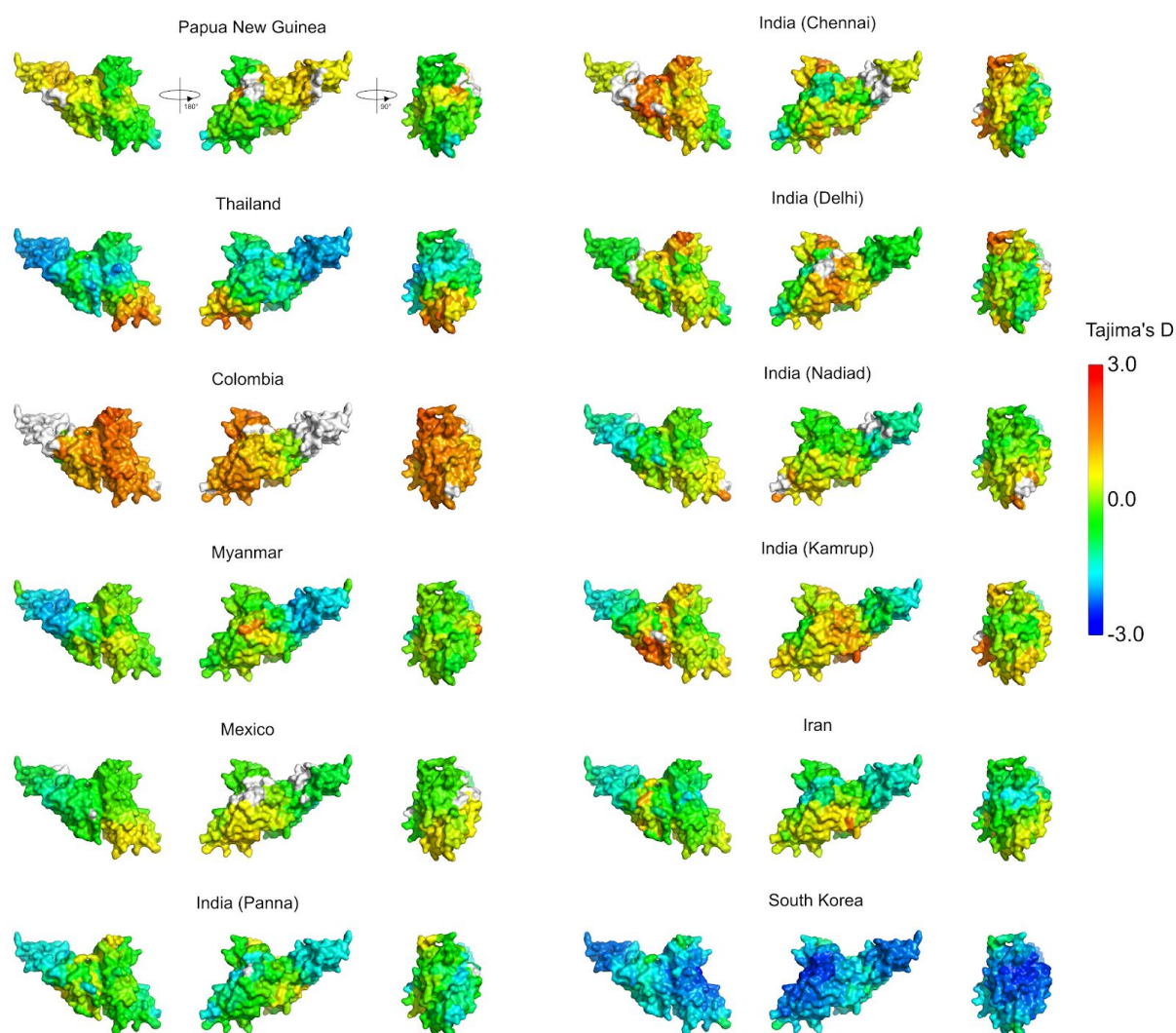
Additional File 3: Spatially-derived nucleotide diversity for *PvAMA1* across multiple populations. Nucleotide diversity was calculated using a 3D sliding window over a modelled *PvAMA1* structure, with a radius of 15 Å for each window.



Additional File 4: Spatially-derived Tajima's D for *PvAMA1* across multiple populations. Tajima's D was calculated using a 3D sliding window over a modelled *PvAMA1* structure, with a radius of 15 Å for each window.

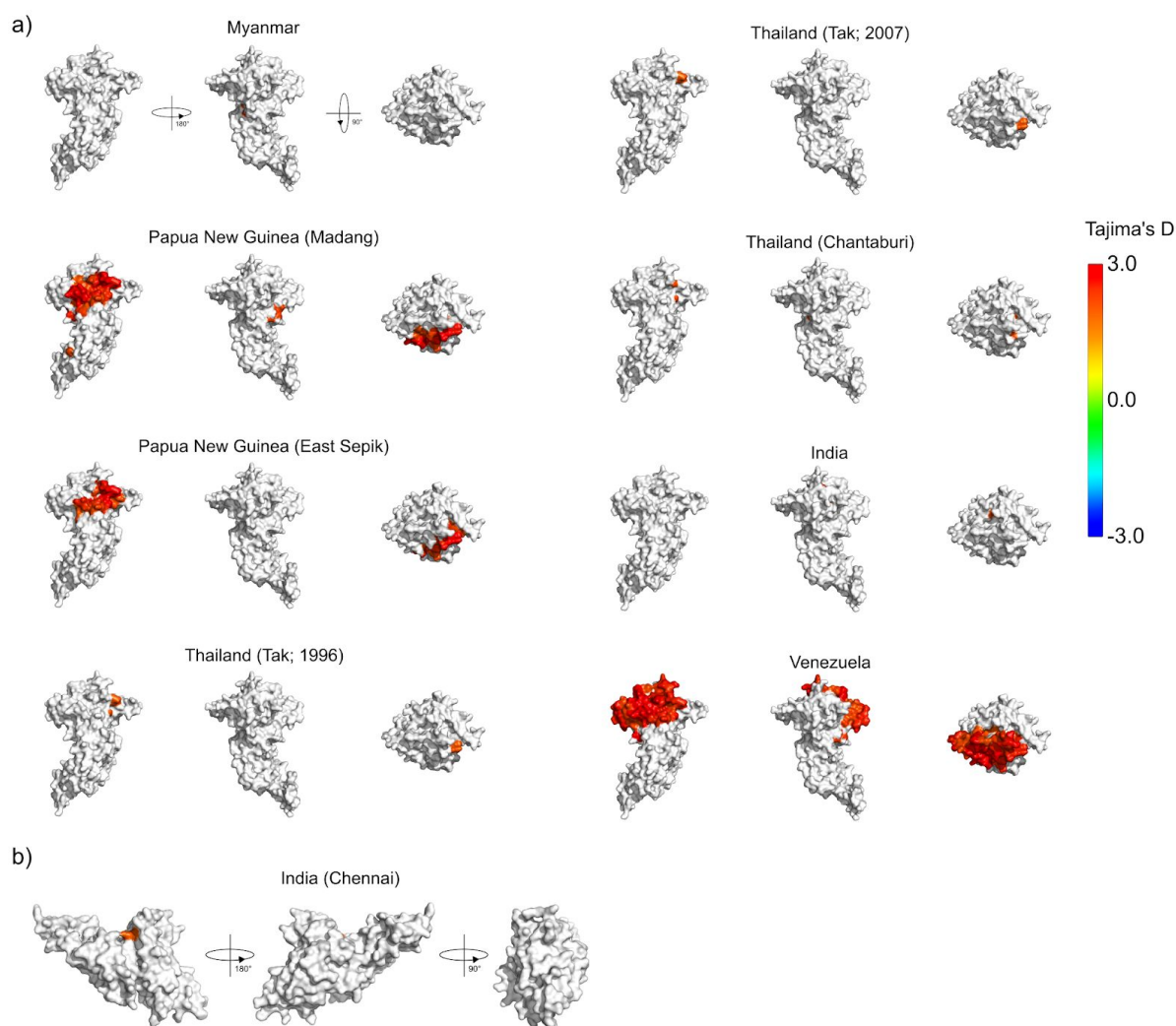


Additional File 5: Spatially-derived nucleotide diversity for *PvDBP* across multiple populations. Nucleotide diversity was calculated using a 3D sliding window over a modelled *PvDBP* structure, with a radius of 15 Å for each window.



Additional File 6: Spatially-derived Tajima's D for *PvDBP* across multiple populations.

Tajima's D was calculated using a 3D sliding window over a modelled *PvDBP* structure, with a radius of 15 Å for each window.



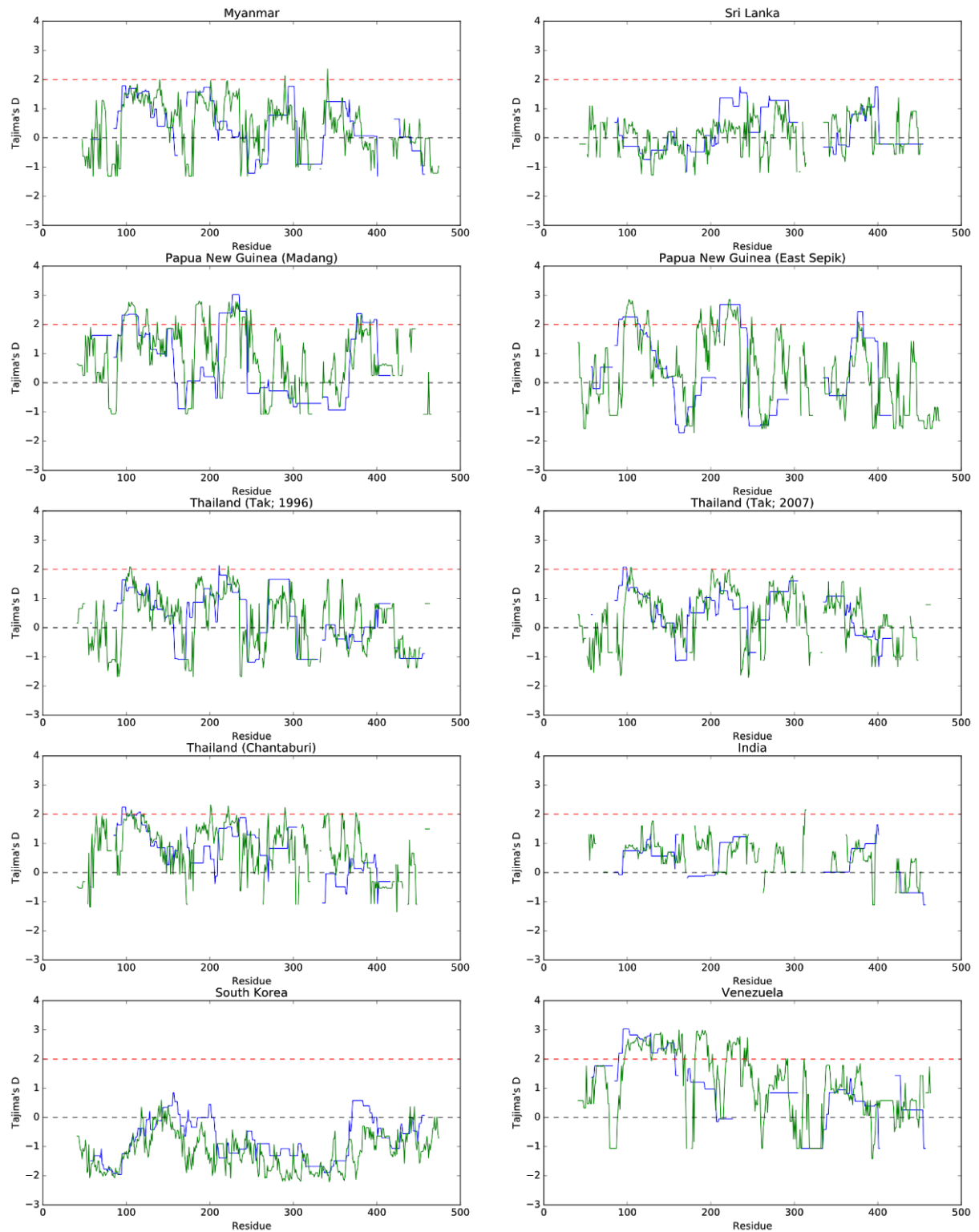
Additional File 7: Location of statistically significant ($p < 0.05$) Tajima's D values on modelled *PvAMA1* (a) and *PvDBP* (b) structures. Confidence limits are those defined by Tajima [60]. Only populations with significant Tajima's D values are shown here.

Additional File 8: Sequence diversity at interfaces.

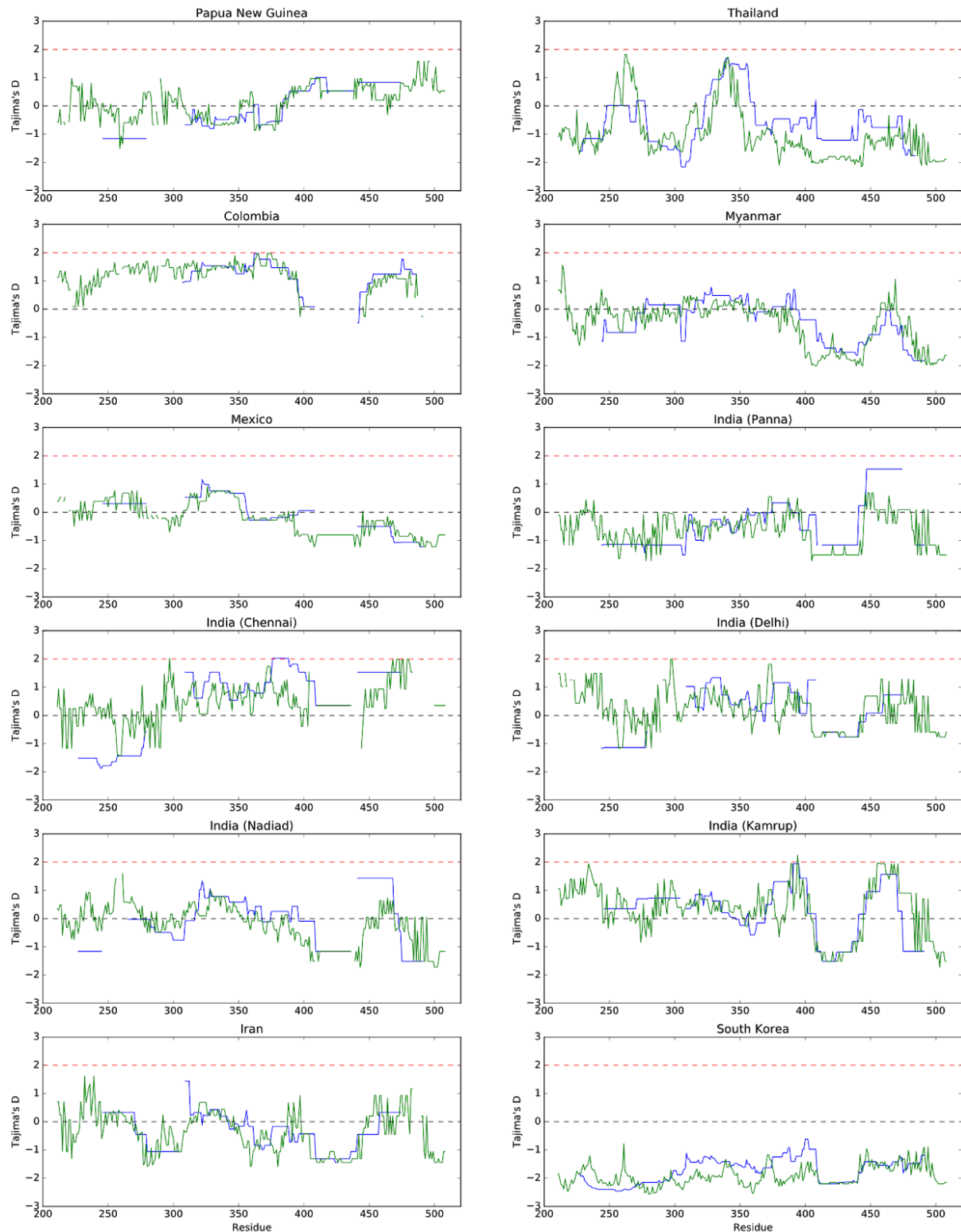
<i>Pv</i>DBP		
Polymorphism	Minor Allele Count	Minor Allele Frequency (%)
R263S	49	20.1
T359R	10	4.1
V240L	3	1.2
V240I	1	0.4
I265L	1	0.4
K366N	1	0.4
Y271S	1	0.4
<i>Pv</i>AMA1		
Polymorphism	Minor Allele Count	Minor Allele Frequency (%)
N132D	235	46.4
N130K	57	11.3
D133N	17	3.4
M153T	12	2.4
G117R	12	2.4
K86R	3	0.6
A172T	3	0.6
N132G	2	0.4
R88G	1	0.2

Note: The presence of polymorphisms was determined using sequences from all populations.

Polymorphisms with a minor allele frequency > 5% are shown in bold. Any polymorphic residues within 4 Å of the *Pf*AMA1:RON2 interface or the *Pv*DBP dimerization/DARC binding interface are included in the above table.



Additional File 9: Comparison of spatially derived Tajima's D and conventional linear sliding window calculation of Tajima's D for *PvAMA1*. Tajima's D values using a linear sliding window results are shown in blue, whilst spatially derived Tajima's D values are shown in green. The threshold for significance ($p < 0.05$) as defined by Tajima [60] is shown as a dotted red line.



Additional File 10: Comparison of spatially derived Tajima's D and conventional linear sliding window calculation of Tajima's D for *PvDBP* RII. Tajima's D values using a linear sliding window results are shown in blue, whilst spatially derived Tajima's D values are shown in green. The threshold for significance ($p < 0.05$) as defined by Tajima [60] is shown as a dotted red line.

6. DISCUSSION

The exact determinants of protective immunity to malaria remain unclear, and it is important to further our knowledge of malaria immunity as part of an ongoing effort to develop an efficacious and long-lasting vaccine against malaria. While the newly licensed RTS,S vaccine offers partial protection from infection, this protection is of limited efficacy and is short-lived. As such, there is a need for ongoing vaccine development and an improved understanding of the targets of protective immunity is likely to aid such an effort. This thesis employed a number of computational approaches to examine the impact of protein structure on the development of adaptive immune responses against the malaria parasite. As part of this work, differences between structured and intrinsically disordered proteins (IDPs) were examined, as well as investigating new ways of integrating protein structural information into population genetics analyses to identify potential targets of protective immunity.

6.1 Summary of findings

Firstly, the role of IDPs in humoral and cellular immune responses was investigated (Chapter 2). *Plasmodium* species contain a disproportionate number of IDPs compared to most other eukaryotes, which raises the question if this may be linked to their evasion and modulation of the host immune response. While there did not seem to be discernible differences in the number of IDPs between different *Plasmodium* species, immunologically exposed subcellular compartments of the parasite had increased levels of IDPs. This suggests some role for this class of proteins either in the invasion process or as targets of immunity. Indeed, a number of leading vaccine candidates contain sizeable regions of disorder, suggesting an important role for these regions [1–3]. It was also observed that disordered protein regions contained a higher frequency of polymorphisms than structured domains, as well as an increased number of tandem repeat regions. Many polymorphisms likely arise as a means of evading established immune responses, although it is also noted that disordered regions are less constrained in terms of sequence conservation [4]. Both the accessibility and plasticity of disordered regions likely contributes to their increased polymorphic variation, and this is something that will need to be considered for vaccines that include IDPs.

As part of an adaptive immune response, T-cells recognise peptide antigen presented in the context of MHC molecules; CD4⁺ T-cells recognise MHC class II/peptide complexes and CD8⁺ T-cells recognise MHC class I/peptide complexes. It was shown in Chapter 2 that IDPs contain very few peptides that are capable of being presented on MHC molecules. This was found to be an inherent property of IDPs, rather than the result of sequence-specific mutations to avoid immune recognition

via MHC molecules. This has implications for the development of any vaccines composed of IDPs, as there will need to be careful consideration to include adequate T-cell help within any vaccine constructs. It is noted that the repeat region of CSP is predicted to be intrinsically disordered, and it is this region which is included in the RTS,S vaccine, along with the C-terminal domain. This construct is coupled to the hepatitis B surface antigen. Both the C-terminal domain of CSP [5] and the Hepatitis B surface antigen [6] contain appropriate CD4⁺ T-cell epitopes, making up for the lack of CD4⁺ T-cell epitopes within the repeat region.

Chapter 3 of this thesis focused on regions of known structure within *Plasmodium* species, in particular *P. falciparum*, and examined markers of immunity and selection pressure in the context of various structural features. This work employed a proteome-wide approach, with the hope of identifying novel structured regions that are significant targets of immune responses. Within this study, it was noted that polymorphic residues that existed at a high frequency within the population were likely to be surface exposed, and were enriched within secondary structure turns, which is in line with previous work on non-malaria antigens [7]. Additionally, as targets of adaptive immunity, there are numerous examples of antibodies targeting surface exposed turns, such as the anti-EBA175 RII mAb R217 which targets the beta-finger on EBA-175 (**Figure 1.3D**) [8] and the anti-pfs25 mAb 1190 which interacts with a surface-exposed turn on Pfs25 via its heavy chain CDR2 and CDR3 loops [9]. This supports the theory that polymorphic variation on many *Plasmodium* antigens is the result of immune selection pressure. As part of this study, we also employed a novel method that allows integration of protein structural information into population genetic tests for selection pressures. Traditionally, tests for selection pressure are applied as a sliding window along a linear gene sequence, allowing identification of particular regions which exhibit strong signatures of selection compared to the rest of the genome. However, given that many antibodies interact with protein antigen in its native, structured form, we considered the need to include information on residue proximity in 3D space into tests for selection pressure. This enabled identification of a region in *PfAMA1* bordering DII and DIII that was under a high degree of balancing selection, which may be the result of immune selection pressure on this region. Interestingly, although DI is the most polymorphic region of *PfAMA1* and is the target of several inhibitory antibodies, there is evidence that mAbs against DIII are also capable of inhibiting parasite growth [10]. This suggests a need to focus on *PfAMA1* domains that contain functional epitopes whilst having minimal polymorphic variation.

Chapter 4 showed the development of this 3D sliding window method into an open-source Python package called BioStructMap, which is now also available as a user friendly, web-based interface. This tool represents a novel method for inclusion of structural data into tests for selection pressure

that have traditionally only utilised genomic data. Additionally, the BioStructMap package can also map other sequence-aligned data over a protein structure, and hence has applications beyond tests of selection pressure.

In Chapter 5, the BioStructMap tool was applied to two leading *P. vivax* vaccine candidates, PvAMA1 and PvDBP. Genomic sequences for these antigens were collated from a number of different global populations, and patterns of diversity and selection on these antigens were compared across populations. Interestingly, very similar patterns of diversity were observed across most populations worldwide, with particular regions of the protein structure having high levels of diversity in all populations, whereas other regions had universally low levels of diversity. This suggests a commonality in the epitopes targeted during the course of naturally acquired immune responses to these antigens. For vaccine development, this is encouraging, as it means that studies from a few populations should be able to be readily extrapolated. Interestingly, in this study we observed balancing selection on DI alone for PvAMA1, in contrast to our previous study on PfAMA1; the latter showing strong signatures of balancing selection on the border of DII and DIII. Balancing selection can arise as the result of immune selection pressure on a protein region, and this highlights potential differences in the immune targeting of PfAMA1 vs PvAMA1. The reason for this difference is not clear, although it is possible that sequence differences between the two species may alter the immunodominant epitopes targeted during the course of natural infection. Indeed, for the two *P. vivax* antigens examined in Chapter 5, several regions that have been shown to be the targets of inhibitory mAbs have very low diversity, suggesting that they may not be targeted to any significant extent during natural exposure. Alternatively, they may be functionally restricted in terms of polymorphic variation due to binding interactions with other proteins, although this is less likely as there is no current evidence for additional binding partners to either AMA1 or DBP within these regions of low diversity. Also, known binding regions within AMA1 and DBP (RON2 binding and DARC binding respectively) are both surrounded by clusters of polymorphic residues, and there is no clear cluster of polymorphisms surrounding other conserved regions within these proteins, again suggesting that these regions may not be targeted during natural exposure. This increases the likelihood that these regions might be able to be targeted in a vaccine construct using epitope focusing techniques, or by reducing the immunodominance of polymorphic regions [11]. Indeed, that has already been investigated for PvDBP in a study that mutated all polymorphic residues to non-reactive amino acids, and found that antibodies raised against these protein constructs generally targeted conserved regions of the protein and were able to inhibit parasite invasion [12,13].

Taken together, the results of this thesis highlight the need to carefully consider a range of different structural characteristics when selecting potential vaccine candidates, whether they are IDPs or structured proteins. This thesis also highlights the wealth of data that already exists for many *Plasmodium* proteins and the value in developing novel bioinformatic tools that can bring combined insights into the genetic, proteomic and immunological aspects of these antigens.

6.2 Experimental validation of T-cell responses against IDPs

The computational approaches outlined in Chapter 2 suggest that disordered domains contain very few MHC-binding peptides, which has the potential to limit the generation of humoral and cellular immune responses against this class of proteins. There are additional peptide processing steps within the MHC processing pathways that likely further restrict the set of peptides that can be presented, and it is not clear if these processing steps would preferentially select ordered or disordered peptides. A number of prediction tools exist for the processing of peptides in the MHC class I pathway [14], and these could also be used to further explore the biases in the MHC class I processing and presentation pathway. However, the processing of peptides in MHC class II pathways is less well understood, and there remains a significant lack of accurate predictive tools for identification of proteasomal cleavage products in endosomes as part of the MHC class II presentation pathway [15].

Acknowledging the limits of predictive approaches, the observations in Chapter 2 would benefit from further experimental validation, and there are a number of approaches that could be used to investigate this. To directly compare the ability to present peptides from ordered vs disordered regions of proteins, a number of mass spectrometry (MS) based approaches could be employed. A number of techniques have been previously employed to investigate the peptides presented on both MHC class I and MHC class II molecules, and these usually involve extraction of MHC-peptide complexes from antigen presenting cells followed by analysis of bound peptide using MS based techniques [16]. There are limitations with some MS based approaches, including identification of only a small percentage of bound peptides and biases towards peptides with particular characteristics [16,17]. However SWATH-MS based approaches have emerged recently as useful tools for characterisation of the immunopeptidome of many cell types [16,18]. SWATH-MS based immunopeptidomics techniques have yet to be applied to presentation of peptides from *Plasmodium* species, and combined with the predictions of disorder used in Chapter 2, these techniques could be used to quantify the *in vitro* difference in presentation of peptides from disordered or ordered regions, compared to our *in silico* predictions. Such experiments would also provide a rich data source for the identification of potential T-cell epitopes for inclusion in vaccine constructs. A

further step that was not fully explored within Chapter 2 is the subsequent activation of T-cells upon binding to MHC-peptide complexes. Given the huge diversity of possible T-cell receptors, prediction of dominant T-cell binding peptides is difficult, and this is a further area that could benefit from experimental work. T-cell activation assays could be performed using cells from naturally exposed individuals, using antigen from a range of ordered and disordered antigens from *P. falciparum* or *P. vivax*. Indeed, a number of studies have already been performed, mostly focusing on CSP, and it is noted that the majority of T-cell epitopes identified fall within the structured C-terminal domain [19–22]. One study which focused on genome-wide epitope profiling of CD8⁺ T-cell responses to *P. berghei* liver stage parasites identified two main CD8⁺ T-cell epitopes within *Pb*TRAP (PBANKA_1349800) and *Pb*S20 (PBANKA_1429200) that were correlated with protection [23]. Both of these epitopes falls within structured regions of *Pb*TRAP and *Pb*S20. However, given the small number of epitopes identified by this study, it is difficult to draw any conclusions from this observation.

6.3 Limitations and future directions for B-cell epitope prediction in IDPs

Whilst prediction of MHC binding peptides is relatively accurate, prediction algorithms for B-cell epitopes is less so [24]. The work presented in Chapters 2 and 3 used predictors of B-cell epitopes to identify putative epitopes within both disordered and structured antigens. Within this work, it was recognised that current B-cell epitope prediction methods are limited in their prediction accuracy, which is partly due to the difficulty in comprehensively defining all possible epitopes for a given antigen [25]. However, when dealing with IDPs, another limitation is evident, as the large majority of IDP regions are predicted to contain B-cell epitopes, and hence these predictions do not provide means of identifying the few immunodominant epitopes within an IDP. This is likely the result of B-cell epitope prediction algorithms being trained on structured antigens, as data is often obtained from antigen-antibody complexes, as is the case for BepiPred 2.0 [26], or from linear peptide array data from a structured antigen. Epitopes within these structured antigens are often found within surface exposed loops, as was observed in Chapter 3, and these may have a similar sequence composition to IDPs. While it is likely that antibodies could be generated against most of the predicted epitopes within IDPs, there is evidence that antibodies are preferentially raised against a few immunodominant epitopes within any given IDP, although recognition of a wide range of epitopes is possible and does occur within a population setting [27,28]. This observation suggests a need for a B-cell epitope prediction algorithm that is specific for disordered antigens, trained on sets of known immunodominant epitopes within IDPs. Such an algorithm should not identify all potential epitopes within an IDP, but should rather focus on identifying the most likely epitopes

within a given protein sequence. This may enable faster and more targeted *in vitro* screening of potential epitopes for further vaccine development.

6.4 Integrating structural features into tests of immune selection pressure

Chapters 3, 4 and 5 introduced and utilised a novel method (BioStructMap) for integrating structural data into a variety of tests that would traditionally be applied as a sliding window over a linear gene sequence. The integration of 3D protein structural data with genomic data, with respect to identifying regions under immune selection pressure, polymorphic hotspots etc., should be a major priority given the large number of structures now available and the increasingly accurate predictions of protein structure by tools such as I-TASSER [29]. The approach introduced in this thesis could easily be extended, for example by examining surface patches rather than spherical windows. This would account for effects such as lack of surface accessibility for some residues within a spherical window, and inclusion of non-contiguous regions on the protein surface as a result of using the simple spherical window. However, the spherical averaging approach did provide additional information compared to a linear sliding window approach, and is likely adequate in most cases.

It would also be worth exploring the impact of antigen dynamics on adaptive immune responses, as it may be that flexible surface loops are able to bind to a variety of paratope shapes, and are also less constrained in terms of sequence conservation. This may mean that these regions are more likely to be polymorphic, as well as serving as good ‘distractors’ for antibody responses, shifting antibody responses away from critical functional epitopes. This was hinted at by the evidence presented in Chapter 3 that polymorphic residues were enriched within turn elements as well as strongly correlated with residue surface exposure. Turn elements are likely to form part of surface exposed loops, and it would be of interest to perform both *in silico* molecular dynamics modelling of key antigens and associated biophysical experiments such as NMR spectroscopy [30] to validate predictions of flexibility and dynamic behaviour. This could be coupled with data on population-level polymorphic variation to investigate the role of single nucleotide polymorphisms on overall protein stability. This would be of interest when coupled with data on the frequency of polymorphisms within a population, as polymorphisms that disrupt protein stability may be less likely to propagate through a population.

6.5 Development of an online platform to explore structure and immunology data for *Plasmodium* species

To integrate the multiple levels of data developed as a part of this thesis, a second preliminary web-based platform has been developed that integrates the data generated in Chapter 2 alongside the structure-based approaches developed in Chapters 3 and 4. This has initially been applied to the *P. falciparum* proteome, but will be extended to other *Plasmodium* species. This platform has been called PlasmoSIP (*Plasmodium* Structure, Immunology and Polymorphisms), and collates a number of experimental data and computational predictions into a single database. This is accessible via an easy-to-use website (<https://plasmosip.burnet.edu.au>) and can be used to aid vaccine design, selection of biomarkers, or as an educational tool. PlasmoSIP currently allows the visualisation of the following protein features: the existence of specific MHCI and MHCII binding epitopes; the occurrence of tandem repeats; protein localisation within the parasite; protein function and accessibility; protein structure; and location of polymorphisms. Within this package we have also included an interactive ‘Proteome Explorer’ that enables users to view summary statistics (i.e. percentage disorder, percentage tandem repeats, number of SNPs, Tajima’s D etc.) for proteins within the proteome, and then proceed to design custom filters to narrow down the selection of proteins for further analysis.

Within PlasmoSIP, individual protein searches enable users to view parameters particular to each protein (**Figure 6.1**). These include predicted structural features that may impact on immune responses (disorder, tandem repeats, location of polymorphisms), as well as potential T-cell epitopes indicated by predicted MHCI and MHCII binding peptides, and predicted linear B-cell epitopes. At a proteomic level, the ‘Proteome Explorer’ enables users to view the distribution of summary metrics for each protein in an interactive manner (**Figure 6.2**). Selectable windows allow the selection of a subset of proteins in each histogram, and all histograms and scatter plots update upon subset selection to display only the updated selection. There is also the utility to restrict the analysis to proteins localised within particular subcellular locations, although this only covers proteins that have known experimental locations as annotated in ApiLoc [31]. Within the Proteome Explorer, selection of individual proteins for detailed viewing is performed by either clicking on the relevant protein within the scatter plot summary, or selecting a protein within the results table. Upon selection of an individual protein, a summary plot is generated for this protein, displaying features such as predicted disorder, tandem repeats, predicted linear B-cell epitopes and known SNPs. A BLAST query is also performed, matching the selected protein sequence against all *Plasmodium* PDB protein structure entries. Suitable PDB matches are then displayed for viewing

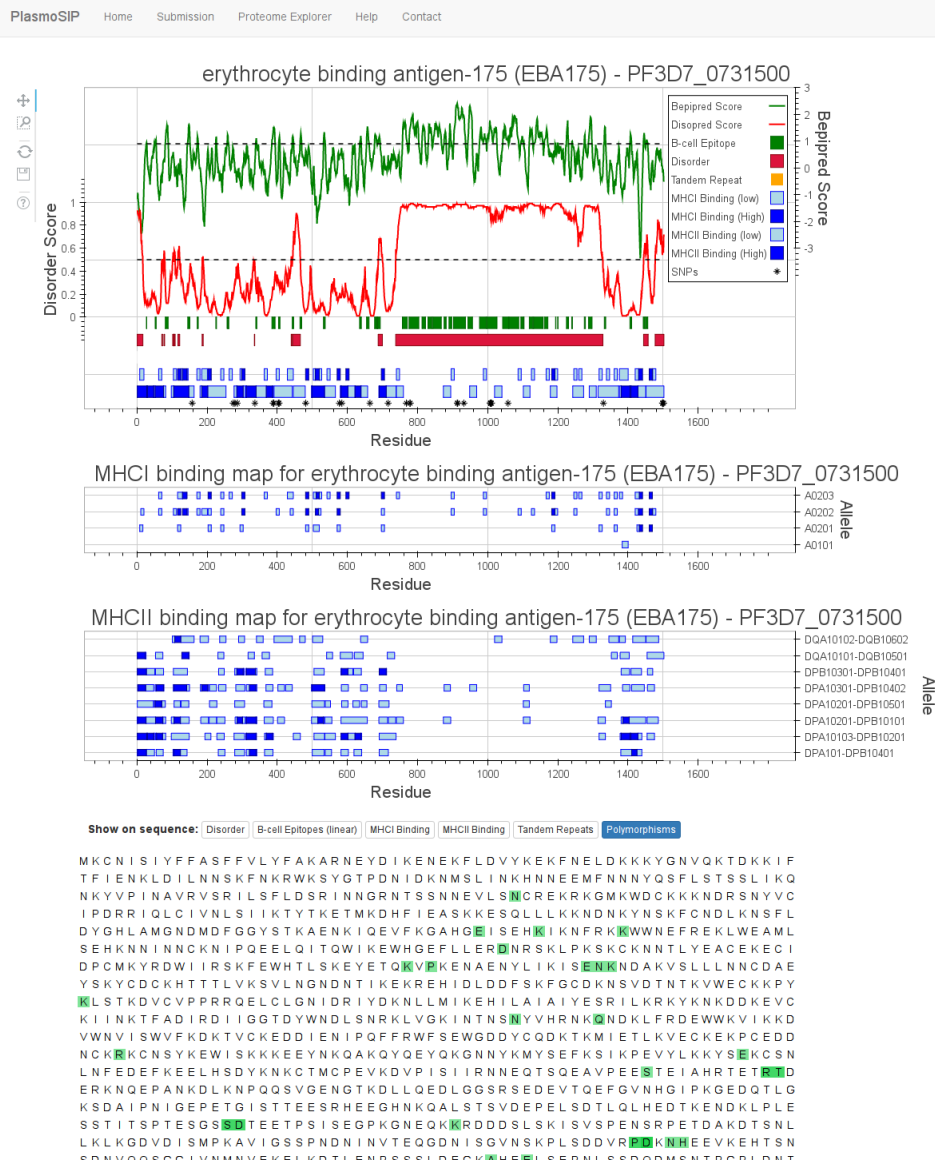


Figure 6.1: Individual protein view within PlasmoSIP showing predictions of disorder, linear B-cell epitopes, MHC binding and polymorphic residues for EBA-175. A large number of MHC class I and class II alleles can be selected, with a representative set shown here. Polymorphic residues are also highlighted on the protein sequence for EBA-175.

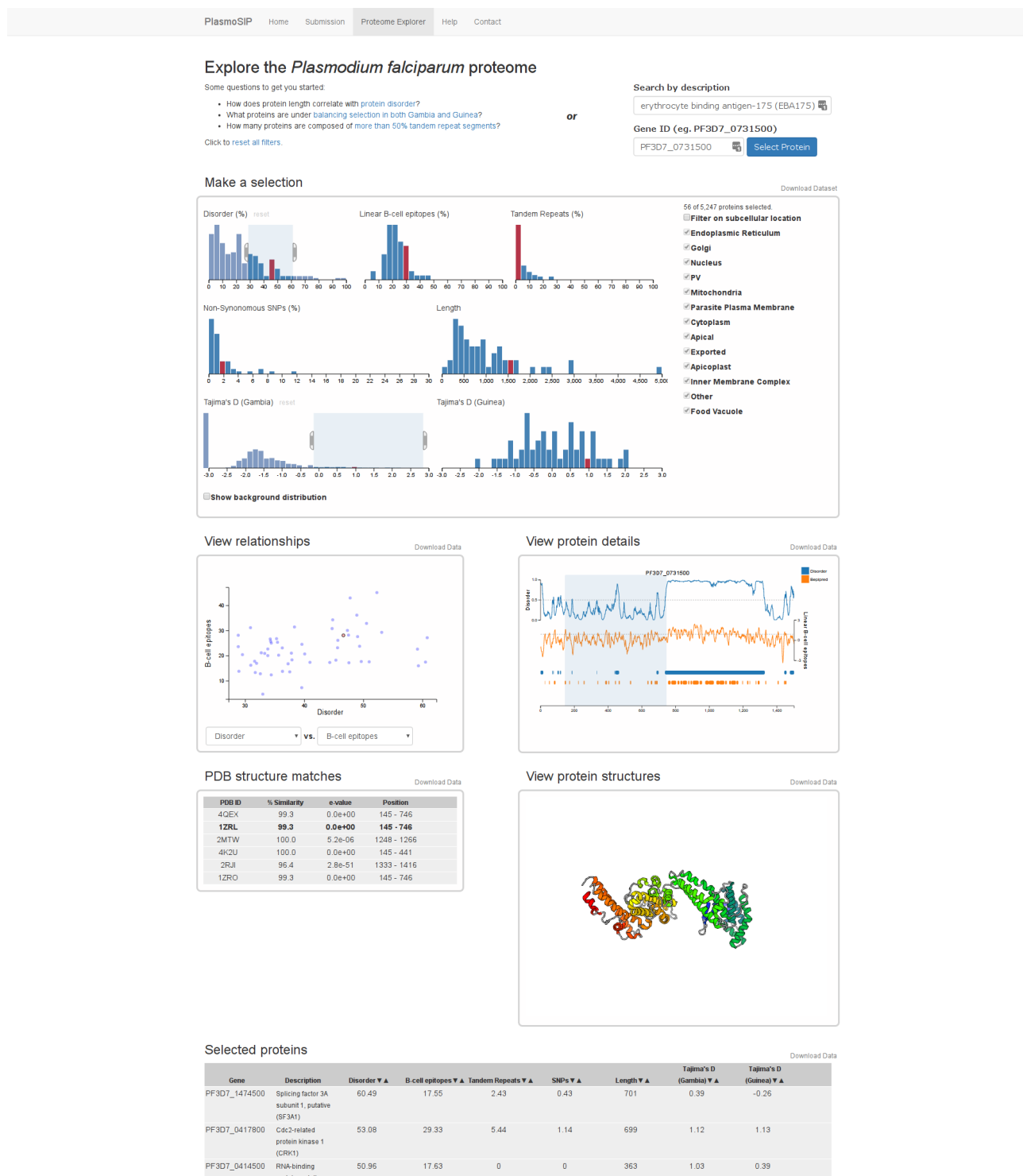


Figure 6.2: PlasmoSIP Proteome Explorer. The Proteome Explorer tool allows the interactive visualisation of various structural and immunological parameters that are applicable to vaccine design and biomarker selection. In this example, the location of EBA-175 is highlighted in histograms of several summary statistics, and a detailed view of the regions of disorder and potential epitopes within EBA-175 is also shown. Matching PDB structures are identified, and protein structures can be viewed in an embedded viewer. A summary list of all proteins within the current selection (as defined by adjustable windows within the proteome-level histograms) is also provided at the bottom of the page.

within an in-browser viewer, with the location of the currently selected PDB structure shown on the individual summary plot.

Whilst not currently implemented, future development of PlasmoSIP will involve the addition and integration of the BioStructMap tool introduced in Chapter 4. This will be paired with polymorphism data from PlasmoDB and structural data from the PDB enabling automated mapping of known polymorphisms from a variety of geographic locations onto known structures. Additionally, this could also utilise modelled structures from the MODBASE database [32], or structures predicted using other tools such as I-TASSER [29]. Future development will also extend the PlasmoSIP tool to other *Plasmodium* species that infect humans such as *P. vivax* and *P. knowlesi*, and species that are used in animal models such as *P. berghei*, *P. chabaudi*, *P. yoelii* and *P. reichenowi*.

6.6 Concluding remarks

This thesis has explored the relationship between protein structure and the adaptive immune system within the context of malaria infection. The overarching hypothesis for this work was that both disordered and ordered proteins are important antigens in the adaptive immune response against malaria, and consideration of protein structure will yield additional insights into targets of adaptive immunity. It was discovered that IDPs are dramatically different to structured antigens in terms of antigen presentation to T-cells, and this raises important considerations for the development of future vaccines based around disordered antigens. Consideration of protein structure is also important when it comes to structured antigens, and protein 3D structural information was used to enhance the identification of regions under possible immune pressure for several leading vaccine candidates. The techniques pioneered in this thesis have broad applicability to other diseases and should enhance awareness of the importance of protein structure when examining selection pressures that arise as the result of protein-protein interactions. Key areas for further research include the experimental validation of some of the results from this thesis, as well as the application of the strategies employed in this thesis to other pathogens such as HIV, influenza or tuberculosis.

6.7 References

1. 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.
2. Adda CG, Murphy VJ, Sunde M, Waddington LJ, Schloegel J, Talbo GH, et al. *Plasmodium falciparum* merozoite surface protein 2 is unstructured and forms amyloid-like fibrils. *Mol Biochem Parasitol*. 2009;166: 159–171.

3. 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.
4. Khan T, Douglas GM, Patel P, Nguyen Ba AN, Moses AM. Polymorphism Analysis Reveals Reduced Negative Selection and Elevated Rate of Insertions and Deletions in Intrinsically Disordered Protein Regions. *Genome Biol Evol*. 2015;7: 1815–1826.
5. Sinigaglia F, Guttinger M, Kilgus J, Doran DM, Matile H, Etlinger H, et al. A malaria T-cell epitope recognized in association with most mouse and human MHC class II molecules. *Nature*. 1988;336: 778–780.
6. Desombere I, Gijbels Y, Verwulgen A, Leroux-Roels G. Characterization of the T cell recognition of hepatitis B surface antigen (HBsAg) by good and poor responders to hepatitis B vaccines. *Clinical & Experimental Immunology*. 2000;122: 390–399.
7. Pellequer JL, Westhof E, Van Regenmortel MH. Correlation between the location of antigenic sites and the prediction of turns in proteins. *Immunol Lett*. 1993;36: 83–99.
8. 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.
9. Scally SW, McLeod B, Bosch A, Miura K, Liang Q, Carroll S, et al. Molecular definition of multiple sites of antibody inhibition of malaria transmission-blocking vaccine antigen Pfs25. *Nat Commun*. 2017;8: 1568.
10. Dutta S, Dlugosz LS, Drew DR, Ge X, Ge X, Ababacar D, et al. Overcoming antigenic diversity by enhancing the immunogenicity of conserved epitopes on the malaria vaccine candidate apical membrane antigen-1. *PLoS Pathog*. 2013;9: e1003840.
11. Oscherwitz J. The promise and challenge of epitope-focused vaccines. *Hum Vaccin Immunother*. 2016;12: 2113–2116.
12. Ntumngia FB, Adams JH. Design and immunogenicity of a novel synthetic antigen based on the ligand domain of the *Plasmodium vivax* duffy binding protein. *Clin Vaccine Immunol*. 2012;19: 30–36.
13. Ntumngia FB, Pires CV, Barnes SJ, George MT, Thomson-Luque R, Kano FS, et al. An engineered vaccine of the *Plasmodium vivax* Duffy binding protein enhances induction of broadly neutralizing antibodies. *Sci Rep*. 2017;7: 13779.
14. Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci*. 2005;62: 1025–1037.
15. Nielsen M, Lund O, Buus S, Lundegaard C. MHC class II epitope predictive algorithms. *Immunology*. 2010;130: 319–328.
16. Caron E, Kowalewski DJ, Chiek Koh C, Sturm T, Schuster H, Aebersold R. Analysis of Major Histocompatibility Complex (MHC) Immunoepitomes Using Mass Spectrometry. *Mol Cell Proteomics*. 2015;14: 3105–3117.
17. Escobar H, Reyes-Vargas E, Jensen PE, Delgado JC, Crockett DK. Utility of Characteristic

- QTOF MS/MS Fragmentation for MHC Class I Peptides. *J Proteome Res.* 2011;10: 2494–2507.
18. Caron E, Espona L, Kowalewski DJ, Schuster H, Ternette N, Alpízar A, et al. An open-source computational and data resource to analyze digital maps of immunopeptidomes. *Elife.* 2015;4. doi:10.7554/eLife.07661
 19. Sedegah M, Kim Y, Ganeshan H, Huang J, Belmonte M, Abot E, et al. Identification of minimal human MHC-restricted CD8⁺ T-cell epitopes within the *Plasmodium falciparum* circumsporozoite protein (CSP). *Malar J.* 2013;12: 185.
 20. Takita-Sonoda Y, Tsuji M, Kamboj K, Nussenzweig RS, Clavijo P, Zavala F. *Plasmodium yoelii*: peptide immunization induces protective CD4⁺ T cells against a previously unrecognized cryptic epitope of the circumsporozoite protein. *Exp Parasitol.* 1996;84: 223–230.
 21. Arévalo-Herrera M, Valencia AZ, Vergara J, Bonelo A, Fleischhauer K, González JM, et al. Identification of HLA-A2 restricted CD8(+) T-lymphocyte responses to *Plasmodium vivax* circumsporozoite protein in individuals naturally exposed to malaria. *Parasite Immunol.* 2002;24: 161–169.
 22. Franke ED, Sette A, Sacchi J Jr, Southwood S, Corradin G, Hoffman SL. A subdominant CD8(+) cytotoxic T lymphocyte (CTL) epitope from the *Plasmodium yoelii* circumsporozoite protein induces CTLs that eliminate infected hepatocytes from culture. *Infect Immun.* 2000;68: 3403–3411.
 23. Hafalla JCR, Bauza K, Friesen J, Gonzalez-Aseguinolaza G, Hill AVS, Matuschewski K. Identification of targets of CD8⁺ T cell responses to malaria liver stages by genome-wide epitope profiling. *PLoS Pathog.* 2013;9: e1003303.
 24. El-Manzalawy Y, Honavar V. Recent advances in B-cell epitope prediction methods. *Immunome Res.* 2010;6 Suppl 2: S2.
 25. Kringelum JV, Lundegaard C, Lund O, Nielsen M. Reliable B cell epitope predictions: impacts of method development and improved benchmarking. *PLoS Comput Biol.* 2012;8: e1002829.
 26. Jespersen MC, Peters B, Nielsen M, Marcatili P. BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes. *Nucleic Acids Res.* 2017; doi:10.1093/nar/gkx346
 27. MacRaild CA, Zachrdla M, Andrew D, Krishnarajuna B, Nováček J, Židek L, et al. Conformational dynamics and antigenicity in the disordered malaria antigen merozoite surface protein 2. *PLoS One.* 2015;10: e0119899.
 27. Guy AJ. Defining the Structural Features of *Plasmodium falciparum* merozoite antigen EBA-175 RIII-V and Identification of Corresponding B-cell Epitopes. Honours Thesis, Monash University. 2012.
 29. Yang J, Yan R, Roy A, Xu D, Poisson J, Zhang Y. The I-TASSER Suite: protein structure and function prediction. *Nat Methods.* 2015;12: 7–8.
 30. Zandarashvili L, Esadze A, Iwahara J. Chapter Two - NMR Studies on the Dynamics of Hydrogen Bonds and Ion Pairs Involving Lysine Side Chains of Proteins. In: Christov CZ, editor. *Adv Protein Chem Struct Biol.* 2013. pp. 37–80.

31. ApiLoc. Available: <http://apiloc.biochem.unimelb.edu.au/apiloc/apiloc> Accessed 21 Dec 2017
32. Pieper U, Eswar N, Davis FP, Braberg H, Madhusudhan MS, Rossi A, et al. MODBASE: a database of annotated comparative protein structure models and associated resources. *Nucleic Acids Res.* 2006;34: D291–5.