Charged Repetitive Protein Sequences in the Modification of the Host Red Blood Cell by the Malaria Parasite

Heledd Mai Davies

Thesis is submitted for the degree of Doctor of Philosophy

Institute of Structural and Molecular Biology, University College London

Declaration of Authorship

I confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Heledd Mai Davies

"Don't worry - so what if there is no life after death? There is life after Garp, believe me."

— **John Irving, The World According to Garp**

Abstract

During the symptomatic asexual stage of its life cycle, the malaria parasite *P. falciparum* exports proteins into the host erythrocyte to modulate its properties. Many exported proteins are targeted to the cell periphery where they affect cell deformability, adhesion, and nutrient import. These changes are implicated in causing severe manifestations of the disease. We have identified novel protein modules which are capable of targeting to the periphery of the red blood cell and which may be involved in modulating these important changes. These were all composed of lysine-rich repetitive sequences, which were shown to be functional in ten proteins altogether, including one protein from the zoonotic *P. knowlesi* species. Some were found in proteins known to modulate erythrocyte rigidity and cytoadhesion while others were previously uncharacterised.

One protein, the glutamic acid-rich protein (GARP), contains three lysine-rich repeating sequences with a peripheral-targeting function. Such repetitive sequences are highly enriched in *P. falciparum* yet very few have been shown to be functional. Targeting efficiency was shown to be directly affected by the number of repeats present; suggesting that expansion of short non-functional lysine-rich sequences can lead to the *de novo* formation of localisation sequences. This is observed by comparing the protein sequences of closely related *Plasmodium* species where repeat expansion has led to functional targeting sequences in some but not others. Additionally, we show that the length of many functional repetitive sequences is highly variable between parasite isolates. Repeat expansion and contraction may allow the parasite to rapidly adapt to selective pressure.

Another domain was identified in proteins involved in modulating cytoadhesion and rigidity of host cells in *P. falciparum* and is conserved across primate-infecting *Plasmodium* species. Identifying the roles of these novel modules will expand our understanding of parasite-induced erythrocyte modifications and the underlying mechanisms responsible for severe malaria.

Acknowledgements

Table of Contents

List of Figures

List of Tables

Abbreviations

ACT	Artemisinin-based Combination Therapy
AMA1	Apical Membrane Antigen 1
ATS	Acidic Terminal Segment
BSD	Blasticidin
CD36	Cluster of Differentiation 36
COX1	Cytochrome Oxidase subunit 1
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CSA	Chondroitin Sulfate A
CSP	Circumsporozoite Protein
gDNA	Genomic Deoxyribonucleic Acid
DDT	Dichlorodiphenyltrichloroethane
DMSO	Dimethyl sulfoxide
dNTP	Deoxynucleotide
DTT	Dithiothreitol
EDTA	Ethylenediamine Tetra-acetic Acid
EKAL	EMP3/KAHRP-like Domain
EM	Electron Microscopy
TEM	Transmission Electron Microscopy
SEM	Scanning Electron Microscopy
ER	Endoplasmic Reticulum
FIKK	Exported Kinases containing a 'FIKK' motif
GARP	Glutamic Acid Rich Protein
GEXP	Gametocyte-Stage Exported Protein
GFP	Green Fluorescent Protein
HAP	Histoaspartic Protease
HRP	Histidine Rich Protein
HYP	Hypothetical Exported Protein
HSP	Heat Shock Protein
HT	Host Targeting
ICAM	Intercellular Adhesion Molecule

IOV	Inside-Out Vesicles
KAHRP	Knob Associated Histidine Rich Protein
LB	Luria Bertani
LRR	Leucine-Rich Repeat Protein
LYMP	Lysine-Rich Membrane-Associated PHIST Protein
MAHRP	Maurer's Clefts-Associated Histidine-Rich Protein
MC	Maurer's Clefts
MESA	Mature Erythrocyte Surface Antigen
MEC	MESA Erythrocyte Cytoskeleton-Binding Motif
MHC	Major Histocompatibility Complex
MSP	Merozoite Surface Protein
PACBIO	Pacific Biosciences
PAGE	Polyacrylamide Gel Electrophoreisis
PAO	Phenylarsine-Oxide
PBS	Phosphate Buffered Saline
PCR	Polymerase Chain Reaction
PEXEL	*Plasmodium* Export Element
PfCAM	*Plasmodium falciparum* Calmodulin
PfCDPK1	Calcium-Dependent Protein Kinase
PfEMP	*Plasmodium falciparum* Erythrocyte Membrane Protein
PfRH	*Plasmodium falciparum* Reticulocyte Binding-Like Proteins
PfEBL	*Plasmodium falciparum* Erythrocyte Binding-Like Proteins
PHIST	*Plasmodium* Helical Interspersed Subtelomeric Proteins
PI	Isoelectric Point
pINT	*Plasmodium* Integrase
PIP_2	Phosphatidylinositol 4,5-bisphosphate
PM	Plasma Membrane
PNEP	PEXEL-Negative Exported Proteins
PRESAN	*Plasmodium* RESA-Like Domain
PTP	PfEMP1 Trafficking Protein
PV	Parasitophorous Vacuole

PVM	Parasitophorous Vacuole Membrane
PTEX	*Plasmodium* Translocon of Exported Proteins
RBC	Red Blood Cell
RESA	Ring-infected Erythrocyte Surface Antigen
REX	Ring Exported Protein
RIFIN	Repetitive Interspersed Families of Polypeptides
RON	Rhoptry Neck Proteins
RPMI	Roswell Park Memorial Institute (Media)
SBP	Skeleton Binding Protein
SDS	Sodium Dodecyl Sulfate
SICAvar	Schizont Infected Cell Agglutination Variant Antigen
SN	Supernatant
SOB	Super Optimal Broth
SS	Signal Sequence
STEVOR	Subtelomeric Variable Open Reading Frame
Strep	Streptavidin
SUB1	Subtilisin-Like Protease 1
SURFIN	Surface-Associated Interspersed Genes
TM	Transmembrane
TRX2	Thioredoxin 2
TVM	Tubovesicular Network
VCAM	Vascular Cell Adhesion Protein
VIR	*P. vivax* Interspersed Repeats

Chapter 1: Introduction

1.1 History of Malaria

1.1.1 Origins of the Parasite

Malaria is one of the most widespread and destructive diseases experienced on Earth, with nearly half the world's population currently at risk. Its dominance is a result of millennia of evolution by one of the most prolific parasite species on the planet, the unicellular eukaryotic organisms of the genus *Plasmodium*. *Plasmodium* parasite species are transmitted between vertebrate hosts by *Anopheles* mosquito vectors. They belong to the phylum apicomplexa, which is defined by a distinct apical structure in the invasive stage and includes the parasites responsible for diseases such as toxoplasmosis and babesiosis (Levine, 1971). Most apicomplexan parasites contain a plastid-like organelle called the apicoplast which likely originated from a secondary endosymbiosis event from red algae (McFadden et al., 1996, Kohler et al., 1997). *Plasmodium* parasites have been discovered in amber-preserved insects believed to date back thirty million years (Poinar, 2005); in the intervening years the parasite has diversified to infect most land-based vertebrates including birds, reptiles, rodents and primates. Human-infecting malaria parasites are believed to have originated from a crossover event from chimpanzee or gorilla-infecting parasites in Africa (Liu et al., 2010, Escalante and Ayala, 1994).

Five species of *Plasmodium* infect humans: *P. falciparum*, *P. vivax*, *P. knowlesi*, *P. malariae*, and *P. ovale*. Infection with *Plasmodium falciparum* causes the most severe form of malaria and is responsible for the majority of deaths attributed to the disease. The symptoms of *P. falciparum* infection can vary with disease severity and include fever, severe anaemia, breathing difficulty, and cerebral malaria which can lead to coma and is often fatal. Many malaria cases outside Africa are caused by *P. vivax*; the symptoms of infection are not as severe as those of *P. falciparum*, with mild anaemia and fever commonly reported with few fatalities, although instances of severe disease have also been documented (Anstey et al., 2012). *P. vivax* parasites can remain dormant and cause disease relapse when triggered, making treatment difficult. *P. knowlesi* is a zoonotic malaria parasite which primarily infects macaques in Southeast Asian countries, but can cause severe malaria if transmitted to

15

humans (Cox-Singh et al., 2008). Developments in diagnostic techniques have demonstrated that *P. knowlesi* is far more common than first thought, and is increasing in prevalence due to deforestation leading to increased contact with macaques (Singh et al., 2004a). The other human-infecting malaria parasites, *P. ovale* and *P. malariae* are less common and typically cause mild symptoms (Collins and Jeffery, 2005, Collins and Jeffery, 2007). Individuals in many malaria endemic regions are often infected by more than one parasite species at a time, complicating diagnostic efforts (Snounou et al., 1993).

1.1.2 The Influence of Malaria on Human History

As malaria has evolved to survive within a human host, it has also had a significant influence on the evolution of the human genome. Genetic mutations which protect against malaria are common in populations historically affected by malaria as they confer a selective advantage. For example, sickle cell trait, thalassaemia, and glucose-6-phosphate dehydrogenase deficiency can all result in severe complications and are usually under negative selection, but are selected for in malaria endemic regions as they protect from the severe manifestations of the disease (Kwiatkowski, 2005). Indeed, *P. vivax* has nearly disappeared from West Africa due to the loss of the Duffy antigen from the red blood cells of the population, which is required for the parasite to invade (Miller et al., 1976).

Malaria has also had a significant influence on human history. Malaria-like symptoms have been recorded in patients at least 10,000 years ago, and many ancient civilizations such as the ancient Egyptians and Greeks were plagued by the disease (Neghina et al., 2010). It has been implicated in shaping historical events by causing mass-fatalities in armies (Landers, 1998) and is believed to have contributed to the deaths of numerous historical figures such as Alexander the Great (Atkinson and Truter, 2009). Malaria, translated as 'bad air', was previously believed to be associated with pollution or fumes from swamps. Only in 1880 did the French physician Alphonse Laveran identify the parasite in the blood of infected patients, a discovery which led to the British medical doctor Ronald Ross attributing its transmission to mosquitoes in 1897. Both were eventually awarded the Nobel Prize in medicine for their work.
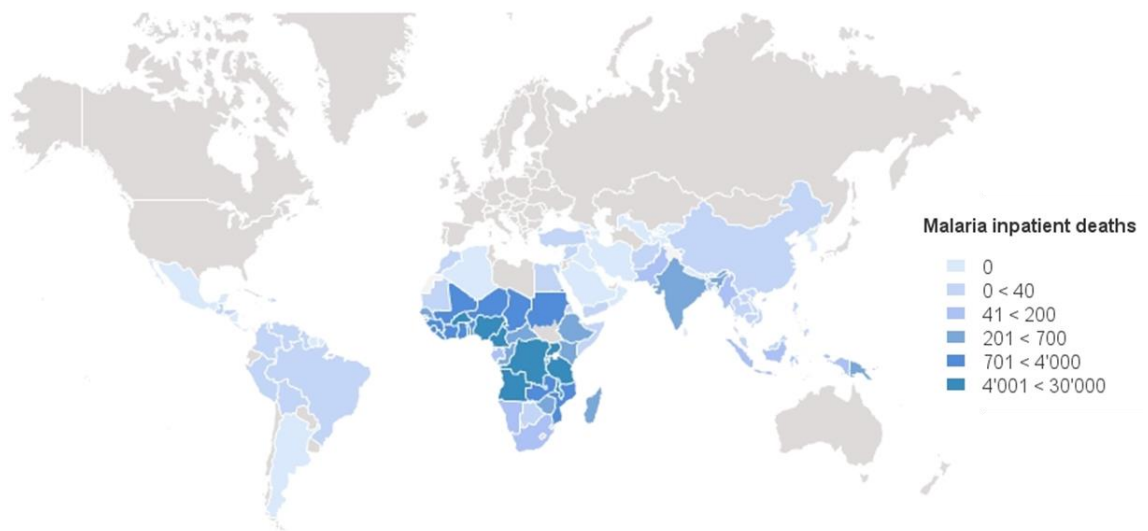
Figure 1.1 – Distribution of malaria infection and number of fatalities caused by the disease per year. Map was created using the 'build a map' tool on the World Health Organisation website.

Today, *P. falciparum* malaria results in more than 400,000 deaths per year (WHO, 2015). Most of these occur in sub-Saharan Africa in children under the age of five who have not developed immunity to the disease (**Fig 1.1**). Severe malaria also occurs in pregnant women due to the build-up of infected erythrocytes in the placenta (Matteelli et al., 1997). Tropical regions are most heavily affected by malaria as they provide an ideal habitat for the *Anopheles* mosquitoes responsible for transmission. In addition to the high number of fatalities, all species of malaria also cause a severe burden on the resources of countries where it is endemic, with loss of working hours, impaired school performance and the cost of medicines being particularly prohibitive in the poorer countries most severely affected (Hotez et al., 2014, Vitor-Silva et al., 2009). As many malaria-endemic countries lack the infrastructure necessary to overcome these problems, eradication of malaria is considered a particularly difficult challenge in world health. However, significant gains have been made in reducing fatalities from the disease in recent years (WHO, 2015).

1.1.3 Treating the Disease

The current front-line treatment against *P. falciparum* malaria is a combination of drugs including the compound artemisinin and its derivatives (Artemisinin-based combination therapy –ACT) (WHO, 2015) (**Fig 1.2**). This compound is derived from the *Artemisia annua* plant, which was used in traditional Chinese medicine for centuries as a remedy against the intermittent fevers typical of malaria. Only in 1977 did Professor Youyou Tu discover the active compound, and was jointly awarded the Nobel Prize for medicine in 2015 (Tu, 2011). The first drugs used against malaria was quinine and other related alkaloids derived from the chinchona tree (**Fig 1.2**). The synthesis of quinine is particularly complex and the compound is still derived from its natural source, limiting manufacturing capacity (Woodward and Doering, 1944). A related but easily-synthesised compound, chloroquine, was synthesised in 1930, and was widely used against malaria worldwide until the emergence of widespread resistance less than 20 years following its deployment (Payne, 1987). Other compounds such as mefloquine, doxycycline and atavaquone/proguanil are used for prophylaxis for travellers to endemic regions (**Fig 1.2**), however due to growing resistance these are not used alone in antimalarial treatment regimens (WHO, 2015).
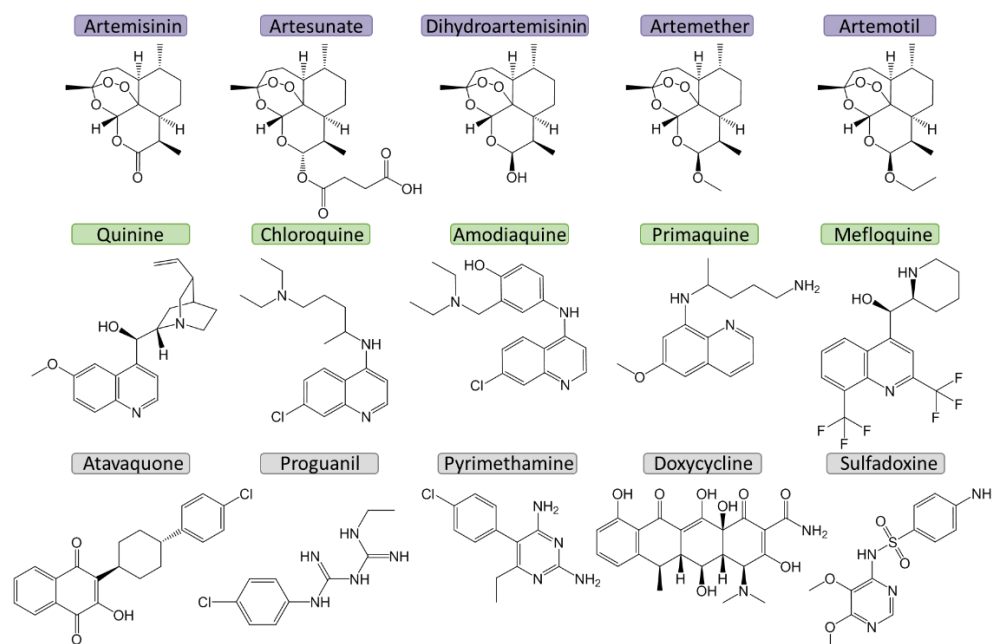
Figure 1.2 – Antimalarial compounds and their structures. Structures were made using Chemdraw (Clark, 2010). Names of Artemisinin-derived compounds are coloured purple and quinine-based compounds coloured green.

Resistance to antimalarial treatment is a growing concern for the future. The first cases of reduced clearance rates for artemisinin-based treatments were reported on the Thai-Cambodian border in 2008 (Noedl et al., 2008), where resistance to chloroquine and many other compounds also originated (Payne, 1987). It is believed that an unregulated drug market distributing artemisin monotherapy as well as sub-theraputic levels of partner drugs, as well as the specific genetic background of *P. falciparum* parasites all contribute to creating a drug-resistance hot-spot in this region. Resistance to artemisinin has been linked to point mutations in the kelch13 gene, which is believed to be involved in protein turnover (Ariey et al., 2014). Artemisinin is activated by the Fe^{2+} ion in heme and is believed to kill the parasite by covalently binding to numerous essential proteins and resulting in increased oxidative stress within the cell (Wang et al., 2015, Becker et al., 2004). The low half-life of the compound in the body, however, means that other longer lasting drugs must be administered at the same time. By affecting protein turnover the parasite may be able to withstand the short efficacy window of artemisinin, while resistance to the other drugs is able to take hold.

A vaccine against malaria would be highly advantageous in overcoming the threat of artemisinin resistance. After several decades, a vaccine has now passed through stage 3 clinical trials and is being distributed in selected African countries (Tinto et al., 2015). This vaccine is composed of the circumsporozoite protein which surrounds the invasive sporozoite stage of the parasite, as well as the hepatitis B surface antigen. A highly-repetitive region of the protein has previously been shown to be highly antigenic and antibodies against this region block invasion of the liver (Stewart et al., 1986). Additionally, a non-repetitive region of the protein is included in the vaccine as the repetitive domain is not capable of forming T-cell epitopes. This vaccine, called RTS,S has been shown to protect against severe malaria in vaccinated children in approximately 30% of cases (Mian-McCarthy et al., 2012). This reduction in severe cases may bring malaria control to a manageable level in endemic countries; however there is some concern that the variability of the CSP protein sequence will allow resistance to develop against the vaccine (Neafsey et al., 2015). A vaccine composed of irradiated whole sporozoites has also entered clinical trials; however there are some practical difficulties in extraction of these invasive

stages from mosquitoes which may hinder widespread use (Seder et al., 2013). Early indications of high protection rates are promising, however.

The most practical malaria eradication schemes used to date have involved vector control. Distribution of insecticide-treated bed-nets to malaria endemic countries is at an all-time high, and has been shown to be one of the most effective methods of lowering infection rates (WHO, 2015). The use of DDT insecticide saw large gains in malaria eradication but these have been partly reversed since the widespread use of the compound was banned for environmental reasons (Mabaso et al., 2004). While some countries still rely on DDT insecticide treatment, resistance in the mosquito population is spreading. Methods for preventing mosquito breeding have contributed to malaria eradication in some countries; mosquito larvae develop while floating on open water, and blocking access to standing water or adding detergent to disrupt surface tension can be effective in lowering insect populations. This has an added advantage of reducing the cases of other diseases transmitted by insects such as dengue fever and zika disease. Genetic methods for creating infertile mosquitoes have also been developed and may be deployed in the near future (Gantz et al., 2015, Hammond et al., 2016).

1.2 The parasite life cycle

1.2.1 Sporozoites

The malaria parasite undergoes several morphological changes as it completes its life cycle, with multiple differentiation steps within both insect and vertebrate hosts (**Fig 1.3**). Sporozoites are injected into the host as the female *Anopheles* mosquito takes a blood meal. These are elongated motile forms of the parasite which are able to glide through the dermis into the bloodstream and migrate to the liver **(Fig 1.3A)** (Amino et al., 2006). Sporozoites actively penetrate the liver and invade hepatocytes, where they grow and multiply into tens of thousands of haploid merozoite forms **(Fig 1.3B)** (Frevert et al., 2005). These are released into the bloodstream as the hepatocytes burst, where they are able to invade erythrocytes. While most *Plasmodium* species invade mature erythrocytes, *P. vivax* invades undifferentiated reticulocytes (Moreno-Perez et al., 2013). Both *P.*

vivax and *P. ovale* may also form dormant hypnozoites in the liver, which form merozoites once re-activated (Mueller et al., 2009).
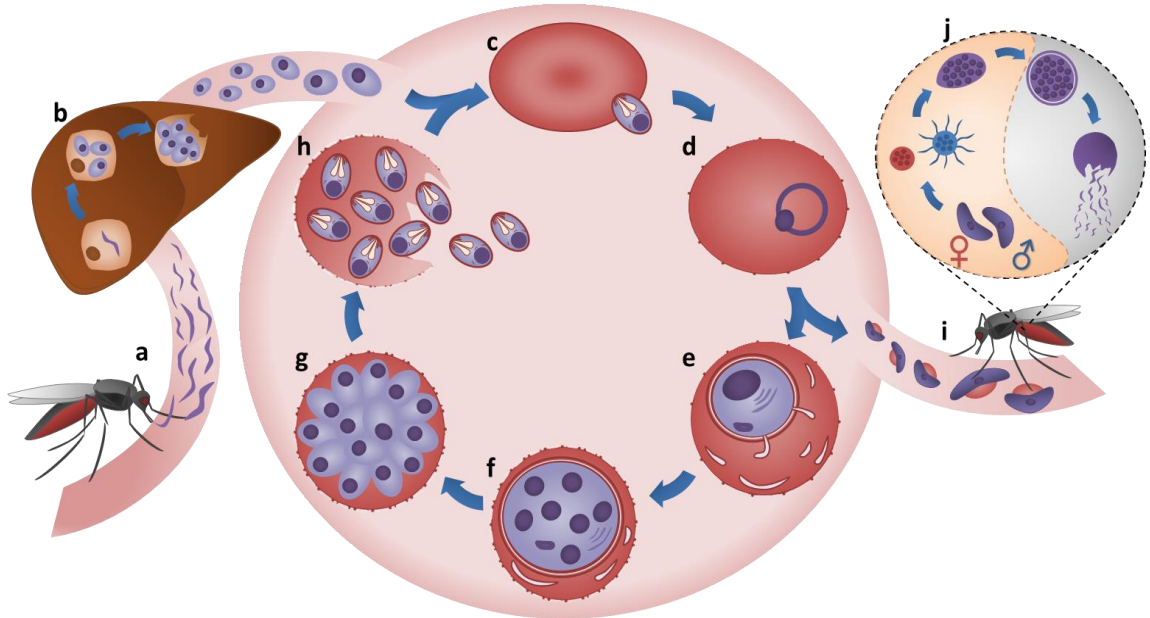
Figure 1.3 - The *Plasmodium falciparum* life cycle. A) Sporozoites are injected from the salivary glands of the *Anopheles* mosquito into the human circulatory system. B) These migrate to the liver cells where they multiply asexually into merozoites, which are released into the bloodstream as the cells rupture. C) Merozoites insert themselves into the host erythrocytes to enter the asexual blood phase. D) Ring stage - parasites start exporting proteins into the red blood cell, and structures called 'knobs' become visible on the surface. E) Early trophozoite – the parasite grows inside its parasitophorous vacuole and new organelles such as Maurer's clefts and tubulovesicular networks are formed. F) Late trophozoite – the parasite replicates its DNA within the confines of the original cell. G) Schizont stage – the cell segments into roughly 16-20 separate membrane-bound merozoites. H) The red blood cell ruptures to release the merozoites which then go on and infect new cells. I) A small proportion of cells will leave the asexual cycle and differentiate into the sexual stage where male and female gametocytes are formed. J) The mosquito phase - The gametocytes are ingested into the mosquito's gut as it takes a blood meal, where they sexually reproduce into zygotes which then form ookinetes which leave the gut and develop into oocysts. These then rupture to release sporozoites from the salivary glands into a new host.

1.2.2 Invasion

The invasion of erythrocytes is a multi-step process (**Fig 1.4**). First, the merozoite attaches itself to the surface of the erythrocyte, an interaction believed to be mediated by the coat of merozoite surface proteins (MSP) which cover the parasite (Moss et al., 2012). The polar merozoite must then reorient itself so that the apical end is in contact with the erythrocyte surface (Gilson and Crabb, 2009). Many proteins involved in this process are stored in organelles within the merozoite such as the rhoptries, micronemes, and dense granules, and are secreted and trafficked to the surface when required (Singh et al., 2010). For example, two families of merozoite surface proteins, the PfRH and erythrocyte binding antigen (PfEBA) families, are released from the rhoptries and micronemes, respectively (Tham et al., 2012, Triglia et al., 2001, Sim et al., 1992). Many of these bind to erythrocyte surface receptors; EBA-175 binds glycophorin A and PfRH5 binds to basigin, for example (Orlandi et al., 1992, Bartholdson et al., 2013). The storage of these proteins within the merozoite until required is thought to limit exposure to the host immune system, while the exposed MSP proteins are variable suggesting they are under selective pressure (Wright and Rayner, 2014).
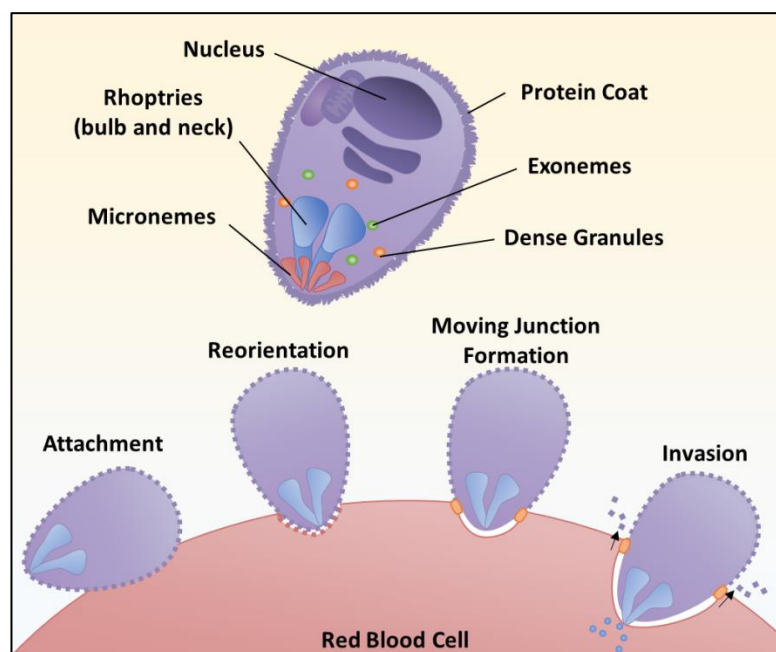
Figure 1.4 – Invasion of merozoites into erythrocytes. Attachment is mediated by the MSP coat. Secretory organelles (e.g. micronemes) release receptors to mediate reorientation and moving junction formation. Proteins are released into the host cell upon invasion, as the protein coat is shed and the parasitophorous vacuole is formed.

22

Once the merozoite is orientated correctly, additional parasite proteins are involved in forming a moving junction composed of an actin and myosin motor which drives invasion into the host cell (**Fig 1.4**) (Baum et al., 2006). A complex of rhoptry proteins known as the RON complex is inserted into the erythrocyte membrane, where its interaction with the apical membrane antigen (AMA1) on the merozoite surface is believed to contribute to the formation of a tight moving junction (Riglar et al., 2011). Additional rhoptry proteins are released into the host cell at this point, which are involved in essential erythrocyte modifications (Counihan et al., 2013, Ling et al., 2004, Nguitragool et al., 2011, Vincensini et al., 2008). As the merozoite enters the host cell, the merozoite surface proteins are shed and the erythrocyte membrane forms a vacuole around the parasite, known as the parasitophorous vacuole (Lingelbach and Joiner, 1998). This remains intact as the parasite develops.

1.2.3 Erythrocyte Stages and Egress

Within the erythrocyte the parasite undergoes several morphological changes (**Fig 1.3C-H**). The early asexual stage is referred to as the ring stage due to its appearance with giemsa staining. The parasite then enters the trophozoite stage, where it continues to grow by ingesting and digesting haemoglobin from the cytosol of the erythrocyte for protein production (Goldberg et al., 1990). The parasite enters the schizont stage and asexually replicates its genome before dividing to form 16-20 daughter merozoites to complete the 48 hour cycle (Bannister et al., 2000).

Egress from the erythrocyte is mediated by a complex cascade of phosphorylation and proteolysis events where a number of proteins are activated in turn and lead to rupture of the host cell (Alam et al., 2015). Many key players in this process have recently been elucidated; protein kinase G triggers the secretion of the subtilisin-like protease 1 (SUB1) from the exonemes of merozoites (Collins et al., 2013) which leads to the subsequent release of proteins from the micronemes. Additionally, the calcium dependent protein kinases PfCDPK1 and PfCDPK5 respond to the calcium influx observed prior to egress and also contribute to secretion of proteins from the micronemes (Dvorin et al., 2010, Bansal et al., 2013). Several proteases lead to lysis of the parasitophorous vacuole and the erythrocyte membrane, likely through a

mechanical disruption of the spectrin cytoskeleton by activated MSP1 (Das et al., 2015). Once egress is triggered, the rupture of the host cell is an explosive event, with merozoites propelled outwards to invade uninfected erythrocytes and continue the cycle.

1.2.4 Sexual Development

During schizogony, a fraction of the parasite population becomes committed to differentiate into gametocytes, the sexual stage of the parasite life cycle. Morphological differences become apparent after the ring stage of the subsequent cycle, where male and female gametocytes begin to form and become increasingly elongated. These stages mature over a period of about a week in *P. falciparum* , and are sequestered within the bone marrow of the host until they reach maturity and are released into the blood stream to be picked up by a mosquito during a blood meal (**Fig 1.3I**). Once within the mosquito midgut changes to the pH and temperature and the presence of xanthurenic acid (Billker et al., 1998) trigger egress from the erythrocyte and exflagellation of male gametocytes to form gametes which fertilize female gametes to form a diploid zygote form (Kuehn and Pradel, 2010). The zygote subsequently differentiates into motile ookinetes, which are able to penetrate the mosquito midgut and form an oocyst (**Fig 1.3J**). Oocysts burst to release thousands of sporozoites, which must then travel to the salivary glands in order to be injected into the host. The cycle restarts once the mosquito bites a new host.

1.3 The Unusual Genome of *P. falciparum*

1.3.1 AT-Bias and Amino Acid Enrichment

The first genome sequence of *Plasmodium falciparum* was completed in 2002 (Gardner et al., 2002); since then the sequences of many other *Plasmodium* species and field strains have become available, allowing greater insight into the evolution and variation of the parasite. The *P. falciparum* genome encodes 5777 gene products, of which 2209 are annotated as 'proteins of unknown function' with little homology to other eukaryotic proteins. One striking observation of the genome is that it is extremely rich in A-T base pairs, with an overall A/T composition of 80.6% (Gardner et al., 2002), and 76.2% within

coding regions. The only other sequenced genome containing a comparable bias belongs to the slime mold *Dictyostelium discoideum* with an AT composition of 77% (Szafranski et al., 2005); why and how the two unrelated eukaryotes independently evolved such extreme genome compositions is not known. As well as a bias towards adenine or thymine at the third position of codons encoding the same amino acids, *P. falciparum* is also enriched in residues encoded by AT-rich codons such as lysine, isoleucine and asparagine, with a resulting reduction in GC-rich residues such as Glycine, alanine and arginine (DePristo et al., 2006).

1.3.2 Protein Repeats

The *P. falciparum* genome is also highly enriched in tandemly repeated protein sequences. At the DNA level, trinucleotide repeats are believed to form from strand slippage during DNA replication in the mitotic cycle, or unequal crossover of chromosomes during meiosis (DePristo et al., 2006). This is a dynamic process with repetitive sequences often expanding and contracting at a greater rate than that of single nucleotide mutation (**Fig 1.5**) (Gemayel et al., 2012). Repeating sequences are relatively common in eukaryotes; single amino acid repeats such as glutamine and asparagine tracts are mainly known for their tendency to aggregate and are often associated with neurodegenerative disorders such as Huntington's disease (Nelson et al., 2013). Tandemly repeated sequences with longer unit lengths, on the other hand, often form modular arrays of folded domains, as seen in solenoid protein domains such as ankyrin repeats (Andrade et al., 2001). These structural tandem repeats often mediate protein-protein interactions across diverse protein families. While many studies illustrate the importance of these two classes of repeats, little is known about sequences containing short repetitive motifs which have no defined secondary structure. These are particularly enriched in *P. falciparum,* with over 30% of proteins containing repetitive sequences (Mendes et al., 2013).

The lack of amino acid diversity in tandemly repeated short motifs characterises them as low complexity sequences. Generally, hydrophobic residues are underrepresented in the repeating sequences of *P. falciparum* proteins, and many are therefore also predicted to be intrinsically disordered (Verra and Hughes, 1999). These extended flexible sequences may be inserted into folded

protein domains which are otherwise highly conserved with homologues in other species, while other proteins are composed almost entirely of short repeats. Notably, long tracts of asparagine are present in a quarter of *P. falciparum* proteins (Singh et al., 2004b); in other eukaryotes such sequences are prone to prion-like aggregation which can be highly disadvantageous to the cell, however additional chaperones are encoded by the parasite to attenuate this effect during febrile episodes (Muralidharan et al., 2012). The fact that the parasite has needed to evolve to compensate for the negative impacts of trinucleotide repeat propagation suggests there may be some selective advantage to such sequences.
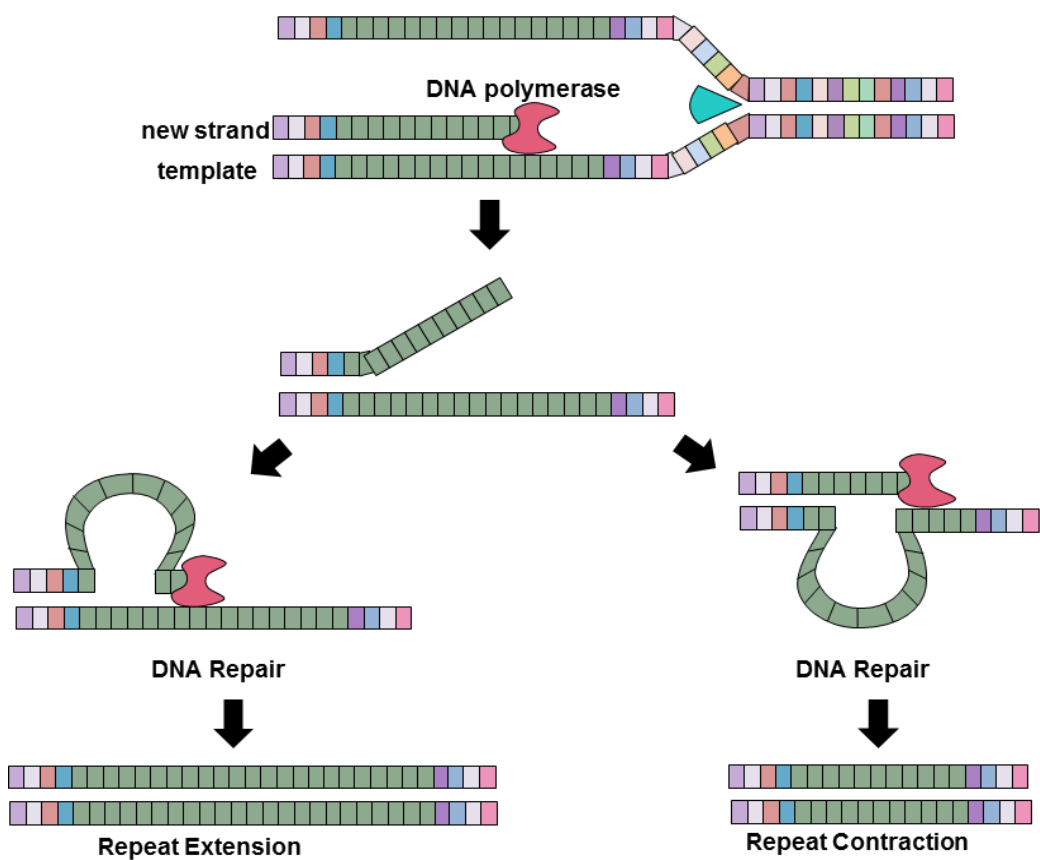
Figure 1.5 – Tandem repeats are highly variable sequences. During DNA replication, dissociation of the strands may result in the formation of secondary structure such as DNA loops in either the template or the newly synthesised strand. In the case of repetitive sequences, the new strand may re-anneal in the incorrect position, leading to extension or contraction of the repetitive sequences.

1.3.3 The Function of Low complexity repeating Sequences

Many perceive repetitive sequences to be 'junk DNA' with no function, and they are often excluded from genomic analysis as they are difficult to resolve (Gemayel et al., 2012). However, deletion of the repetitive sequences of two *P. falciparum* genes results in a clear phenotype; the 'NANP' repeats of the circumsporozoite protein (CSP) are required for invasion of liver cells, while removal of repeats from the REX1 protein affects the separation of Maurer's clefts (Ferguson et al., 2014, McHugh et al., 2015). Additionally, a poly-histidine sequence has been shown to target the knob-associated histidine-rich protein (KAHRP) to Maurer's clefts and the erythrocyte membrane (Wickham et al., 2001). The precise role of the repeats in these cases are unknown, however, and deletion of asparagine repeats from the essential proteasome protein RPN6 had no discernible effect on protein function (Muralidharan et al., 2011).

One prominent hypothesis is that repetitive sequences play a role in immune evasion. Many repetitive sequences are highly antigenic, with antibodies found in the sera of infected individuals against many different repetitive epitopes (Kaur et al., 1990). Notably, the NANP/NVDP repeat region of the circumsporozoite protein (CSP1) now forms the basis of the malaria vaccine RTS, S, and the merozoite surface protein (MSP1) is a vaccine candidate due to protection offered by antibodies against its repetitive region (Conway et al., 2000, Moss et al., 2012, Ferguson et al., 2014). However, the majority of antibodies against repeated regions are not protective, leading many to hypothesise the proteins are instead contributing to the lack of long-term protection against malaria within previously-infected individuals. Atypical memory B-cells have been associated with malaria; these are impaired in proliferation, cytokine production and antibody secretion, suggesting the parasite may actively suppress the host immune response (Portugal et al., 2015).

One theory is that the parasite may flood the immune system with highly-antigenic cross-reactive epitopes to overload B-cells and from a smokescreen to prevent affinity maturation of B-cells towards more protective epitopes (Kemp et al., 1987). Alternatively, it has been shown that highly repetitive sequences may crosslink B-cell receptors and activate the cell in a T-cell independent manner which is less protective (Schofield, 1991, Guy et al., 2015). This is

supported by a computational analysis indicating repetitive sequences are excellent B-cell epitopes, but that the lack of hydrophobic residues in disordered sequences leads to reduced MHC binding and thus weak T-cell stimulation (Guy et al., 2015). These more general roles may explain why such diverse protein families contain repetitive sequences, including many cell-surface proteins which are exposed to the immune system (Mendes et al., 2013). The dynamic expansion and contraction of repeats may therefore allow the parasite to rapidly adapt to selective pressure from the host immune system (Mendes et al., 2013).

1.4 Erythrocyte Modifications

1.4.1 Protein Export

The red blood cell is terminally differentiated and contains no organelles of its own, and so provides an ideal environment for the malaria parasite to reside undetected. The parasite compensates for the lack of protein-making machinery within the erythrocyte by exporting its own set of proteins to perform essential tasks such as nutrient uptake and immune evasion (Haldar and Mohandas, 2007, Nguitragool et al., 2011). These proteins affect mechanical properties such as erythrocyte deformability and adhesion to the vascular system, which contribute to disease severity (Marti and Spielmann, 2013, Spillman et al., 2015, Boddey and Cowman, 2013).

To enter the host cell, proteins must cross both the plasma membrane of the parasite as well as the membrane of the parasitophorous vacuole which surrounds it (**Fig 1.6**) (Marti and Spielmann, 2013). Proteins destined for export must therefore first enter the secretory pathway; N-terminal signal peptides direct many exported proteins to the endoplasmic reticulum (ER), while others contain transmembrane domains. Within the ER, proteins which are to be exported into the host cell must be differentiated from those to be retained within the parasitophorous vacuole. A host targeting (HT) motif, otherwise known as the *Plasmodium* export element (PEXEL), has been shown to be sufficient for enabling proteins to cross the parasitophorous vacuole (Hiller et al., 2004, Marti et al., 2004). This motif, which has the canonical sequence RxLxE/D/Q, is cleaved after the third leucine by the ER-resident protease

Plasmepsin V (Boddey et al., 2010, Russo et al., 2010, Chang et al., 2008). The new N-terminus of the protein is then N-acetylated by an unidentified N-acetyl transferase. Formation of the cleaved N-terminus by a capsid protease is also sufficient for export (Tarr et al., 2013, Gruring et al., 2012), indicating that Plasmepsin V cleavage is not coupled to the transfer of proteins to the export pathway but that the Ac-xE/D/Q N-terminus is instead recognised by components of the export machinery. A significant number of PEXEL-negative exported proteins (PNEPs) have also been identified, which contain transmembrane domains but do not have a clear conserved signal for export (Heiber et al., 2013, Spielmann and Gilberger, 2010).
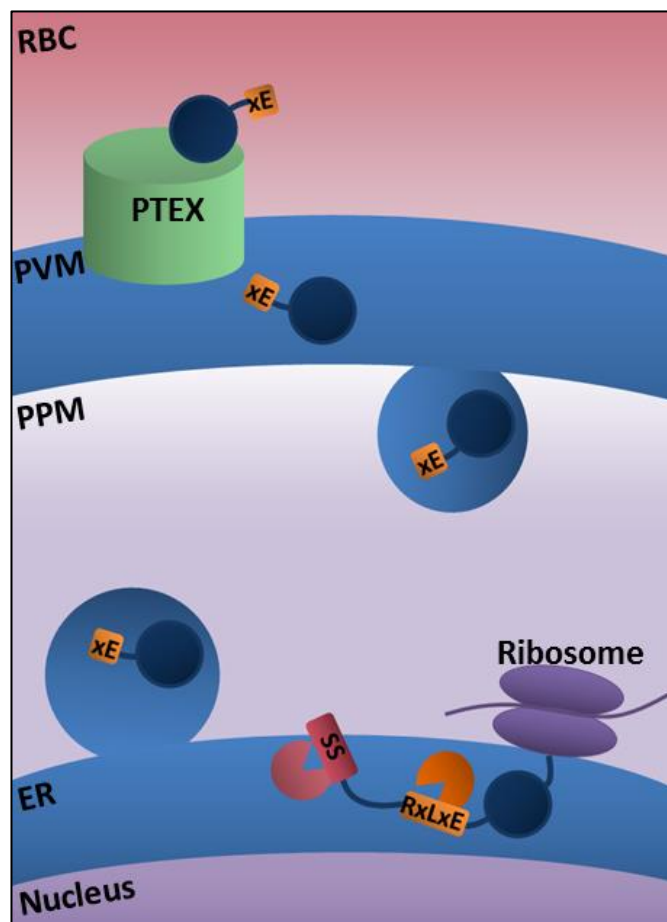
Figure 1.6 – Protein Export in *P. falciparum*. On insertion into the endoplasmic reticulum (ER) Signal Peptidase (red) and Plasmepsin V (orange) cleave the signal sequence (SS) and HT/PEXEL motif (RxLxE), respectively. The cleaved acetylated protein is trafficked *via* vesicular transport to the parasite plasma membrane, then exported into the host cell across the parasitophorous vacuole membrane (PVM) *via* the *Plasmodium* translocon of exported proteins (PTEX).

Once mature exported proteins have passed through the golgi and entered the parasitophorous vacuole, they must traverse the PVM to enter the host cell. A complex termed the *Plasmodium* translocon of exported proteins (PTEX) has been implicated in this process (de Koning-Ward et al., 2009). Within this complex, the EXP2 protein is thought to form a pore within the parasitophorous vacuole membrane (Kalanon et al., 2016, Mesen-Ramirez et al., 2016), while the AAA+ ATPase protein HSP101 is believed to function as a chaperone and may be involved in unfolding proteins to feed them through the pore (Beck et al., 2014). Disulphide bonds must be broken for efficient export, and a thioredoxin-like protein TRX2 is also thought to be involved in the complex (Matthews et al., 2013). HT/PEXEL proteins fused to a DHFR domain, which is stabilised by the addition of WR and cannot be unfolded, were trapped in the PV and immunoprecipitated putative components of the PTEX translocon (Mesen-Ramirez et al., 2016). Additional pull-down experiments and spatial association in cells have indicated the PTEX complex is composed of EXP2, HSP101, TRX2, and two proteins of unknown function named PTEX150 and PTEX88 (Riglar et al., 2013, de Koning-Ward et al., 2009). EXP2, HSP101, and PTEX150 are all refractory to gene disruption in *P. berghei* (Matthews et al., 2013), and a conditional knockdown of the proteins results in a block of protein export for both PEXEL-positive proteins and PNEPs (Elsworth et al., 2014). Other proteins have recently been associated with the PTEX complex, although their roles have not been elucidated (Elsworth et al., 2016).

Some have hypothesised that proteins destined for export are trafficked to segregated compartments of the PV which contain the PTEX translocon, as suggested by the 'necklace of beads' conformation observed in the PV for fluorescently tagged exported proteins (Wickham et al., 2001). Others believe the translocon itself confers selectivity for PEXEL proteins or PNEPs by recognising signals such as the cleaved HT/PEXEL N-terminus. Altogether, more than 400 proteins have been shown to contain a signal sequence and PEXEL/HT motif at the N-terminus or to belong to known families of PNEPs (Hiller et al., 2004, Marti et al., 2004, Heiber et al., 2013). Evidence that the fifth position of the HT/PEXEL motif may tolerate other residues other than E/D/Q suggests an even greater proportion of proteins are exported into the host cell (Schulze et al., 2015, Tarr et al., 2013, Boddey et al., 2013).

Within the erythrocyte cytoplasm new structures called Maurer's clefts are formed soon after parasite invasion (**Fig 1.7**). These membranous structures are hypothesised to behave as an external secretory pathway so that the parasite has the capability to sort proteins outside the confines of its own cell (Bhattacharjee et al., 2008, Hanssen et al., 2008, Przyborski et al., 2003). Vesicles have been observed surrounding these organelles and it is hypothesised that membrane proteins may be trafficked in this fashion between the parasite, the clefts and the erythrocyte membrane (Cyrklaff et al., 2011, Hanssen et al., 2008). Many proteins have been observed in electron-dense clusters in the erythrocyte cytosol, including dots associated with HSP70 and HSP40 proteins, indicating that chaperones may instead be responsible for trafficking proteins within the host cell (Botha et al., 2007, Kulzer et al., 2012). A 'tubovesicular network' extending from the parasitophorous vacuole is also observed in some cells; it may be involved in trafficking proteins or nutrients into or out of the parasite (**Fig 1.7**). Many exported proteins are then inserted into the erythrocyte membrane, while others interact with the unique cytoskeleton of the host cell. Many are associated with the 'knob' protrusions which form of the erythrocyte surface soon after infection **(Fig 1.7)**. However, the function and localisation of the majority of predicted exported proteins is yet to be revealed, with many likely to play roles in immune evasion and nutrient import which are essential for parasite survival.
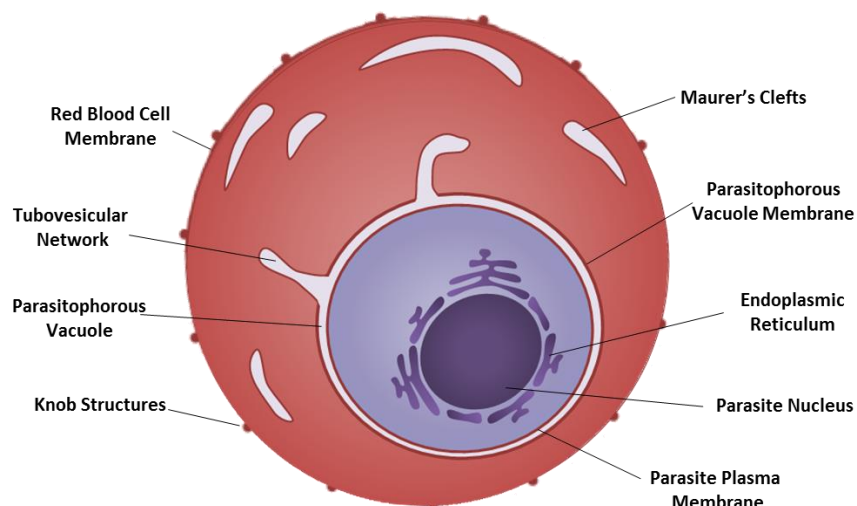
Figure 1.7 – Modifications to the parasite-infected erythrocyte. Membranous structures called Maurer's clefts appear in the host cell cytosol soon after infection, and 'knob' structures form on the cell surface. Additionally, a 'tubovesicular network' extends from the parasitophorous vacuole of the parasite.

1.4.2 The Erythrocyte Cytoskeleton

Erythrocytes are highly specialised for their role in circulating oxygen throughout the body. Their concave structures are highly deformable in order to facilitate their passage through narrow capillaries and splenic sinusoids, many of which have a diameter smaller than that of the erythrocytes themselves (**Fig 1.8A**). This flexibility is conferred by the unique cytoskeleton of the cell, which is composed of a network of spectrin filaments **(Fig 1.8B)** (Mankelow et al., 2012). Alpha and beta-spectrin are large (>250 KDa) proteins composed of a chain of folded domains known as spectrin repeats, of which there are 22 in alpha and 17 in beta spectrin. These repeats each form a three-helix bundle in a coiled-coil conformation, and the repeat domains of the alpha and beta filaments associate to form a heterodimer. The dimers then interact in a head to head manner to form spectrin tetramers. At the tetramerisation site, Ankyrin and Band 4.2 connect spectrin to the membrane pore protein Band 3, while at the other end of the filament, the junctional complex forms a hub of spectrin filaments connected to transmembrane glycophorin proteins (Mohandas and Gallagher, 2008). This is composed of short actin filaments with adducin, dematin, band 4.1, p55, tropomodulin, and tropomyosin (**Fig 1.8C**) (Mankelow et al., 2012). The composition of the red blood cell cytoskeleton has been extensively studied due to the simplicity of the organelle-free erythrocyte, and there exist many protocols for extracting the various different fragments (Black et al., 2008). It is believed that the relatively weak interaction between spectrin dimers at the tetramerisation site contributes to the deformability of the cell, as they may dissociate under shear stress (An et al., 2002). The filaments themselves may also alternate between a tightly-coiled 'chinese finger trap' conformation and an elongated form (Brown et al., 2015).

1.4.3 Modification of the Cytoskeleton

Upon infection with *Plasmodium falciparum*, the erythrocyte cytoskeleton becomes increasingly rigid and cannot deform through narrow openings (Nash et al., 1989). Several parasite proteins have been implicated in rigidifying the host cell; the ring-infected erythrocyte surface antigen (RESA) has been shown to associate with spectrin and its deletion results in a decrease in cell rigidity (Foley et al., 1991, Pei et al., 2007b). Interestingly, this effect is accentuated

when cells are heated to 41 °C, suggesting that RESA may stabilise the cytoskeleton during febrile episodes induced by malaria disease (Mills et al., 2007). Erythrocytes remain rigid once the temperature is decreased, indicating that RESA permanently modifies the cytoskeleton, possibly with the assistance of other heat-activated proteins.
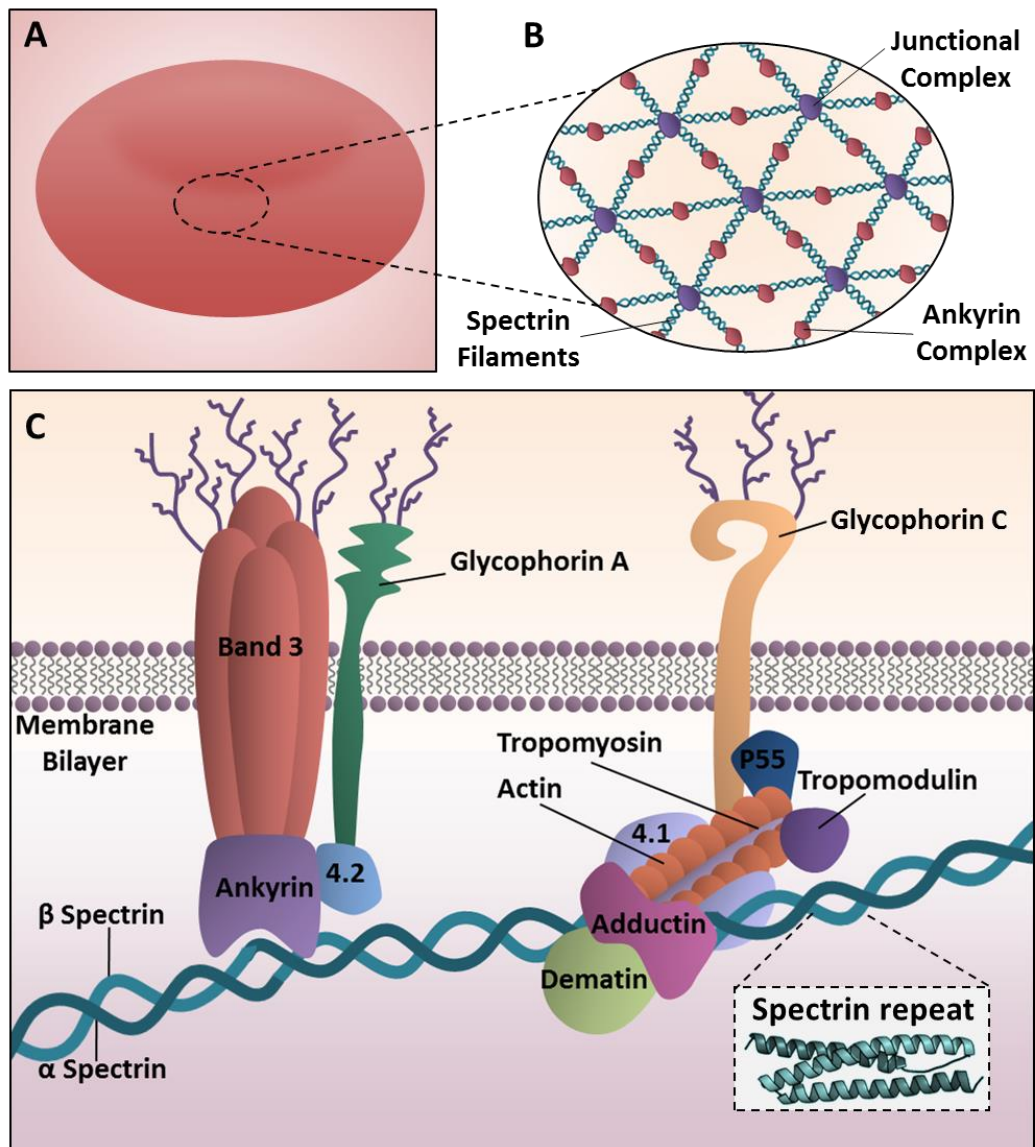
Figure 1.8 – The erythrocyte cytoskeleton and membrane. A) Erythrocytes form a concave sphere which can easily deform through narrow openings. B) The cytoskeleton forms a hexagonal network below the membrane surface. C) Alpha and beta spectrin filaments form tetramers which are connected to the membrane bilayer by two junctional complexes: the Ankyrin complex is composed of the transmembrane protein Band 3 linked to spectrin by Ankyrin, while the junctional complex is composed of short actin filaments capped by various components and linked to the membrane by glycosylated glycophorin proteins. Inset: each spectrin repeat forms a coiled-coil bundle.

Other proteins which have been implicated in modifying cell deformability include PfEMP3 and Pf332, which interact with spectrin and actin, respectively (Waller et al., 2010, Waller et al., 2007a). There is conflicting evidence on whether PfEMP3 destabilises or stabilises the cytoskeleton; deletion of the gene results in a decrease in rigidity (Glenister et al., 2002), however when a short fragment of the protein is incorporated into the cell, the shear resistance is reduced (Pei et al., 2007a). Deletion or truncation of the megadalton protein Pf332 resulted in an increase in cell rigidity and abnormal Maurer's cleft structures, suggesting it may be involved in protein trafficking (Glenister et al., 2009). The mature erythrocyte infected surface antigen (MESA) interacts with protein 4.1, which is present at both the junctional complex and Ankyrin complex (Bennett et al., 1997). This interaction with protein 4.1 may disrupt its binding to p55, however a clear phenotype for MESA has not been established (Waller et al., 2003). Other proteins were shown to contain a motif similar to the MEC motif and also interact with protein 4.1 *in vitro* (Kilili and LaCount, 2011).

Electron microscopy of infected cells has indicated that the actin filaments of the cytoskeleton may be repurposed by the parasite (Cyrklaff et al., 2011). Short capped actin filaments are present in the junctional complex, however much longer filaments are observed extending from the cytoskeleton in infected cells. These may stabilise Maurer's clefts at the cell periphery (Rug et al., 2014); the organelles are motile for the early stages of the parasite but become fixed adjacent to the plasma membrane in later stages (McMillan et al., 2013). Alternatively, these filaments may provide a means of trafficking vesicles between these structures and the plasma membrane. The molecular basis for the reorganisation of actin is unclear, however mutations in haemoglobin which are protective against malaria reduced actin polymerisation (Cyrklaff et al., 2011). The protein MAHRP2 which is localised to Maurer's clefts has also been implicated in tethering the organelles to the erythrocyte cytoskeleton (Pachlatko et al., 2010).

1.4.4 Nutrient Import

The parasite acquires many of the lipids, proteins and nutrients required for growth from the host cell. These are acquired from the erythrocyte cytosol by means of double-membrane invaginations in the PVM and PV called

cytostomes (Lazarus et al., 2008). These fill with haemoglobin and other proteins then fuse with the parasite food vacuole, where proteases break down the haemoglobin to leave toxic heme which is polymerized into pigmented crystals of hemozoin (Goldberg et al., 1990). The amino acids released by the degradation of haemoglobin and other proteins are recycled for protein synthesis.

Some small molecules required by the parasite cannot be acquired from the erythrocyte, however. New permeation pathways are therefore established in the erythrocyte plasma membrane in order for the parasite to import sugars, amino acids, nucleosides, and ions from the blood plasma (Desai, 2014). It is unclear whether the parasite modifies endogenous erythrocyte membrane channels or creates its own pores in the membrane. The cytoadherence-linked asexual gene 3 (CLAG3) is the only protein identified to date which contributes to nutrient import; it is released into the host cell from the rhoptries upon merozoite invasion, where it localises to the erythrocyte plasma membrane (Nguitragool et al., 2011). An inhibitor of membrane channel conductance was found to be selective for DD2 parasites, and was attributed to a divergent CLAG3 sequence (Nguitragool et al., 2011). The protein may form a novel channel in the membrane or may play a modulatory role.

Others have suggested that known anion channels in the erythrocyte are modified by the parasite. The patch-clamp technique, which directly measures conductance across a membrane, was used to evaluate the properties of uninfected and malaria infected erythrocytes, indicating that the anion selectivity and conductance observed for activated endogenous channels were very similar to those of infected cells (Decherf et al., 2004). The erythrocyte kinase PKA was able to activate the channel by phosphorylation, suggesting that the parasite may activate erythrocyte PKA or directly phosphorylate the endogenous anion channel (Decherf et al., 2004). The balance of Na+ and K+ cations is also disrupted in erythrocytes infected with *Plasmodium* , with a greater concentration of sodium in infected cells (Ginsburg et al., 1986). The implications of such changes to the cell composition for the parasite are yet to be established, and it is likely that many more proteins are involved in internalizing required lipids and nutrients into the parasite.

1.4.5 Cytoadhesion

It is believed that the reduction in erythrocyte deformability allows the spleen to detect infected cells, and the parasite has therefore developed a strategy to avoid the organ altogether. In addition to the sequestration of infected erythrocytes in narrow capillaries through increased rigidity, the parasite also expresses adherent proteins on its surface which can interact with receptors on endothelial cells in various organs (Baruch et al., 1996). Infected erythrocytes can also adhere to uninfected erythrocytes and cluster together in 'rosettes', further impeding their journey though the microvasculature (Ho et al., 1991). The sequestration of infected erythrocytes in the circulatory system is believed to be a main factor in the pathogenesis of severe malaria.

Cytoadhesion is mainly mediated by the PfEMP1 family of proteins encoded by var genes (Crabb et al., 1997, Watermeyer et al., 2016, Baruch et al., 1996). There are 60 genes within the family, with only one expressed at a time. Sub-populations of the parasite may switch between expression of PfEMP1 antigens, which bind to different receptors and may contribute to different manifestations of the disease. For example, the VAR2CSA gene encodes a PfEMP1 variant capable of binding to the CSA receptor on placental cells, and results in severe malaria in pregnant women (Viebig et al., 2005). Sequestration of parasites within the brain by association to brain endothelial receptors can result in cerebral malaria, which is a leading cause of death from the disease (Craig et al., 2012b). Variation of the surface antigen protects the parasite from detection by the immune system.

In order to maximise the avidity of binding between PfEMP1 antigens and their respective receptors, the protein is clustered in protrusions known as knob structures on the erythrocyte membrane (**Fig 1.9**) (Crabb et al., 1997). Electron microscopy of these knob protrusions has indicated that they are composed of a spiral-shaped structure lying under the membrane surface, surrounded by electron dense material (**Fig 1.9B-G**) (Watermeyer et al., 2016). The best-studied protein known to be involved in the formation of this structure is the knob-associated histidine rich protein (KAHRP); deletion of KAHRP results in a knobless phenotype (**Fig 1.9B-C**), and results in a significant reduction in adherence to endothelial receptors under flow conditions (Crabb et al., 1997). Other proteins also play a role in the trafficking and stabilising of the variable

surface antigen, as it appears to be transported *via* the Maurer's clefts before insertion into the cell membrane. Deletion of Maurer's cleft proteins such as MAHRP1 and SBP1 result in decreased PfEMP1 on the erythrocyte surface (Cooke et al., 2006, Spycher et al., 2008).
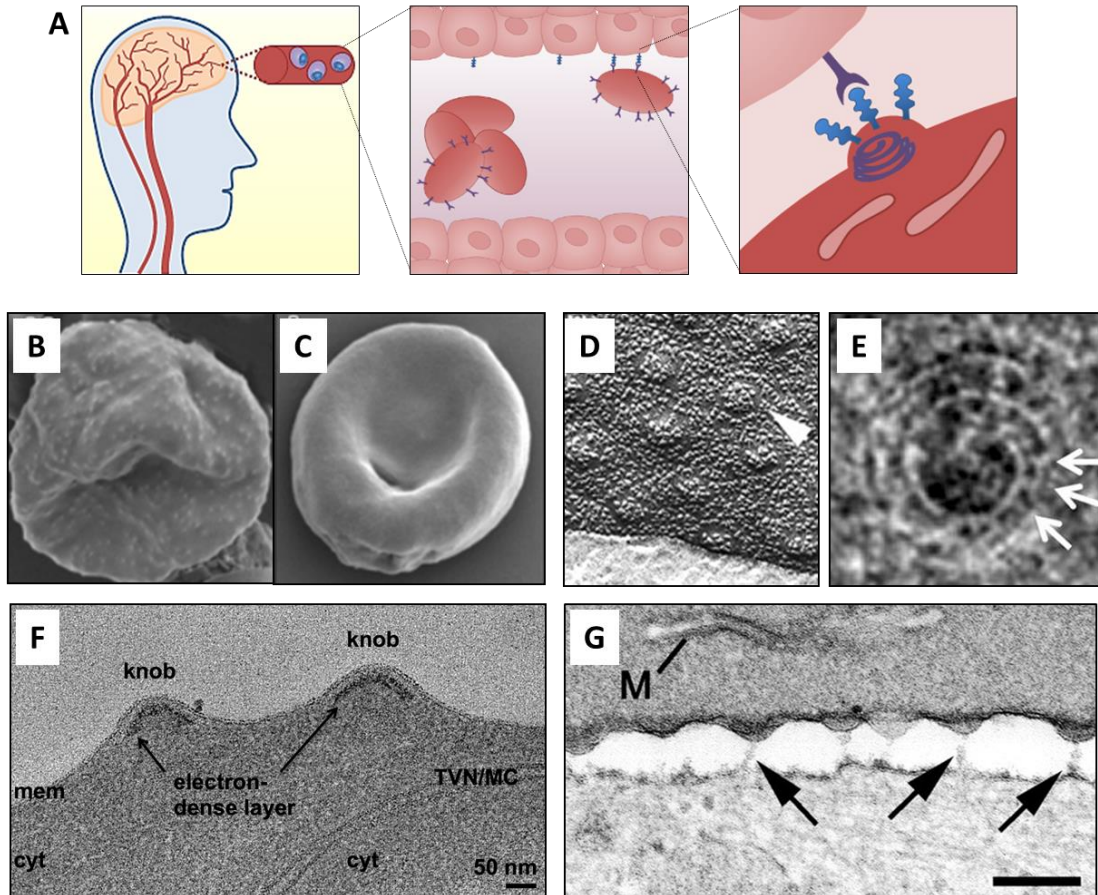
Figure 1.9 – Cytoadhesion of infected erythrocytes is mediated by the PfEMP1 surface antigen presented on knob structures. A) A schematic of cytoadhesion in cerebral malaria; PfEMP1 binds brain endothelial cell receptors, or rosettes form through adhesion to other erythrocytes. These sequester infected cells in the microvasculature. B-C) Scanning electron microscopy (SEM) of KAHRP+ and KAHRP-infected cells; deletion of KAHRP abolishes knob structures (Rug et al., 2006). D) Freeze fracture EM of knobs on the erythrocyte surface. Membrane protein foci are excluded from the area directly surrounding the knobs (Horrocks et al., 2005). E) TEM of knob structures from cytoplasmic side, indicating a spiral structure below membrane (Watermeyer et al., 2016). F) TEM cross-section of an infected erythrocyte, with the electron dense knob structure clearly seen below the membrane (Watermeyer et al., 2016). G) TEM of PfEMP1 presented on the knob structures of an infected cell interacting with receptors on an endothelial cell surface. M = Maurer's cleft (Horrocks et al., 2005). Electron micrographs have been truncated from the original versions by Rug et al., Watermeyer et al., and Horrocks et al.

Another protein family called the Poly-helical interspersed sub-telomeric (PHIST) proteins which contain a *Plasmodium* RESA-like domain (PRESAN) have been implicated in anchoring PfEMP1 to the erythrocyte surface (Oberli et al., 2016). This family is divided into the PHISTA, PHISTB and PHISTC groups based on their sequence motifs (Oakley, 2006, Sargeant et al., 2006). A large number of PHISTB proteins were shown to target the erythrocyte cytoskeleton by an extended PRESAN domain which is not present in the other subgroups (Tarr et al., 2014). One PHISTb protein, known as the lysine-rich membrane-associated PHISTb protein (LYMP) has been shown to directly interact with various PfEMP1 proteins *via* its PRESAN domain, and to interact with the cytoskeletal component Band3 with its C- terminus, suggesting an anchoring role (Oberli et al., 2014, Oberli et al., 2016). Deletion of LYMP resulted in a decrease in cytoadhesion (Proellocks et al., 2014). A PHISTc protein was also shown to interact with the ATS domain of the same PfEMP1 protein as LYMP but with a much lower affinity (Oberli et al., 2014). While the cytosolic ATS domain is largely conserved between PfEMP1 proteins it is possible that different PRESAN domains may exhibit affinities for different surface antigens. Interactions between other PHIST proteins and PfEMP1 are yet to be confirmed, however.

Other families of parasite proteins are displayed on the erythrocyte surface. Like PfEMP1, the RIFIN and STEVOR family proteins are clonally variant (Abdel-Latif et al., 2002, Niang et al., 2009); however their roles are not as well established. Both families have however been shown to mediate rosetting of infected cells to uninfected erythrocytes (Niang et al., 2014, Goel et al., 2015). RIFINS preferentially bind to blood group A receptors while STEVORs bind glycophorin C and may also be involved in invasion of erythrocytes (Goel et al., 2015, Niang et al., 2014). Increased expression of STEVOR proteins also decreases the deformability of the erythrocyte in the late gametocyte stages of the parasite (Naissant et al., 2016). Some members of the SURFIN family are also surface exposed, although little is known about their role in the erythrocyte (Alexandre et al., 2011).

Other proteins may modulate rigidity or cytoadhesion without directly associating with the cytoskeleton or membrane. Members of a unique family of 21 exported kinases known as FIKK kinases have also been shown to modulate

mechanical properties of the erythrocytes (Schneider and Mercereau-Puijalon, 2005). Deletion of FIKK7.1 and FIKK12 resulted in differential phosphorylation of cytoskeletal components (Nunes et al., 2010), while deletion of FIKK4.2 resulted in decreased rigidity and abnormal knob structures with reduced cytoadhesion (Kats et al., 2014). Most FIKK kinases studied to date are localised to the Maurer's clefts, although FIKK12 also associates with the cell periphery (Nunes et al., 2007). The extent of post-translational modifications modifying the erythrocyte is yet to be established; several cytoskeletal components such as spectrin, band 3, and protein 4.1 are differentially phosphorylated in infected erythrocytes (Doerig et al., 2015, Pantaleo et al., 2011, Wu et al., 2009). Phosphorylation may regulate the timing of erythrocyte modification in different life stages; cytoskeleton-targeting proteins such as MESA and KAHRP contain many phosphorylation sites, for example. Other post-translational modifications such as myristoylation and palmitoylation are known to occur in *P. falciparum* yet very little is known about the functional significance of these changes (Doerig et al., 2015).

It is clear that a complex network of proteins is involved in erythrocyte modification. These may be directly associated with the cytoskeleton or are involved in trafficking *via* Maurer's clefts or regulation though post-translational modification. Very few exported proteins have been studied in detail, and elucidating the key players in modulating cell rigidity, cytoadhesion and nutrient import remains an important challenge in elucidating the cause of severe malaria.

1.5 Project Aims and Objectives

The aim of this project was to further our understanding of how the malaria parasite modifies its host erythrocyte. Host cell rigidification and cytoadhesion to endothelial cells leads to the sequestration of infected cells in the microvasculature and contributes to a severe manifestation of the disease. While sequestration allows the parasite to evade the host immune system, changes to the nutrient permeability of host cells are necessary for parasite survival and growth. Greater insights into these processes may therefore lead to new vaccine candidates or drug targets for antimalarial treatment.

It is likely that most proteins directly involved in modulating cytoskeleton rigidity, surface adhesion or the formation of new membrane channels are localised to the periphery of the host cell. Identifying proteins with domains capable of targeting this area is therefore an important first step in explaining the molecular basis of these modifications. A binding interaction with cytoskeletal components has been elucidated for only a handful of proteins to date; these were mostly validated *in vitro* and confirmation of targeting within the cell is required to ensure these interactions are biologically relevant. We aimed to find new proteins involved in modifying the host cell by identifying novel protein domains capable of targeting the erythrocyte periphery. Previously, other conserved peripherally-targeting domains have been identified across exported proteins, such as the PRESAN domain of PHISTb proteins or the MESA erythrocyte cytoskeleton-binding (MEC) motif. It is therefore possible that other uncharacterised binding motifs are used by large groups of proteins with previously unidentified roles in erythrocyte modification.

This project will mainly focus on proteins from *P. falciparum* as it is responsible for the most severe form of malaria. However other *Plasmodium* species such as *P. knowlesi* present an increasing threat to world health and there is currently little understanding of this neglected parasite. Comparisons between *Plasmodium* species will therefore be made to identify conserved mechanisms for survival, as well as to gain insights into the evolution of the species.

Chapter 2: Materials and Methods

2.1 Cloning GFP-tagged constructs of Malaria Proteins

2.1.1 Primer Design

The gene sequences for all malaria proteins used were amplified by PCR from *P. falciparum* strain 3D7 gDNA (catalogue number MRA845 from Bei Resources (Adjalley et al., 2010)), *P. reichenowi* gDNA or *P. knowlesi* gDNA (from strain A1H.1 (Moon et al., 2013)). Primers for restriction ligation cloning were designed to anneal at a temperature of between 55 and 65 °C where possible, with a 3' G or C base. NETPRIMER was used to check for possible primer-dimer formation. Primers were ordered from Integrated DNA Technologies (IDT). To clone repetitive fragments, primers were designed to bind the closest unique sequence to the repetitive DNA sequence to ensure the primers did not anneal at the wrong position in a repeating array (**Fig 2.1A**). Alternatively, an additional homologous sequence was included 5' to the restriction site to target the primer to the end of the repetitive sequence, without including the flanking sequences in the final construct **(Fig 2.1B)**. Overlap PCR was used to fuse two sequences together before ligation (**Fig 2.1C**). For the short GARP repeats and the *P. gaboni* GARP fragment where no genomic DNA was available, the sequence was built up by multiple overlapping primers in sequential PCR reactions (**Fig 2.1D**). All primer sequences are presented in **Appendix Table 4**.

2.1.2 Polymerase Chain Reaction

PCR was performed using phusion DNA polymerase (NEB) in high fidelity Phusion buffer with 200 µM dNTP mix (NEB), 0.5 µM forward and reverse primers and 250 ng template DNA in a 50 µL volume mix. Thirty cycles of denaturation at 97 °C, annealing at 50-65 °C, and extension at 68 °C were performed in a typical PCR reaction. Lower extension temperatures were used for particularly AT rich sequences, as low as 60 °C if necessary, with a longer extension time used to compensate for any effect on polymerase activity. The amplified DNA was purified with Promega DNA cleanup kit according to the manufacturer's instructions.
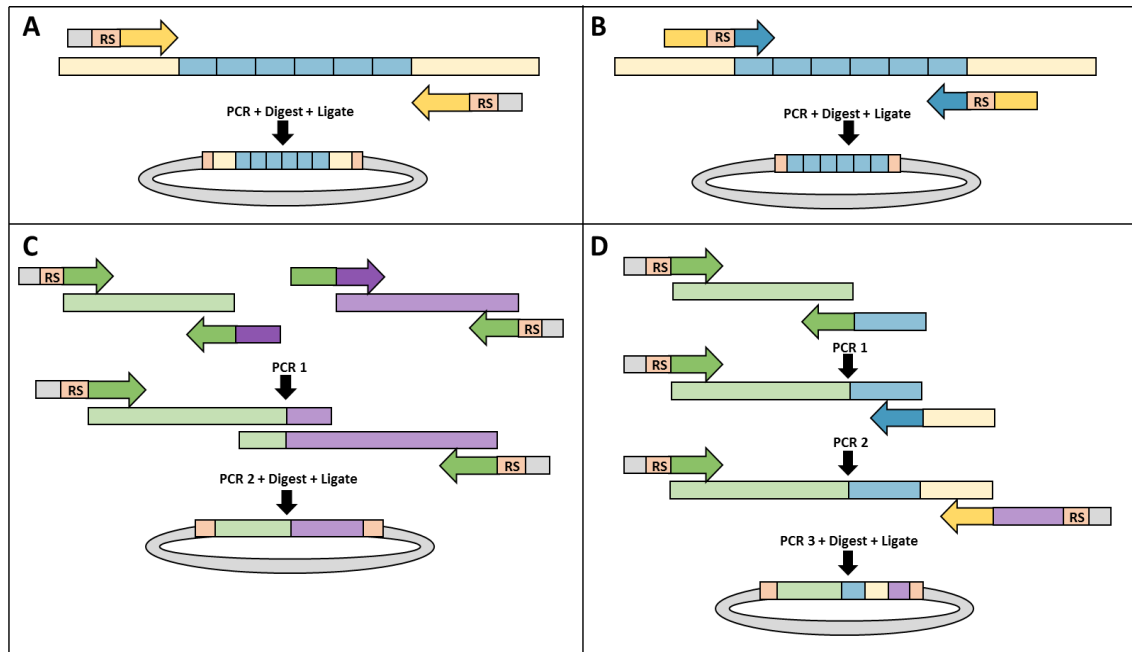
Figure 2.1 – Primer design for restriction-digestion cloning. (A-D) Primers are represented with arrows in matching colours to homologous DNA sequences. Restriction sites are in orange. (A and B) DNA repeats are shown in blue. (C) Overlap PCR. (D) Gene synthesis using primer building blocks.

2.1.3 Restriction-Ligation into expression vectors

Restriction digestion was performed on PCR products and vectors using restriction endonucleases from either New England Biolabs (NEB) or Promega. DNA fragments encoding full-length proteins were inserted into *P. falciparum* expression vector PA446 upstream of a GFP tag and C-terminal strep tag, replacing the Rex3 sequence already present in the plasmid (**Fig 2.2**). Restriction sites AvrII and AatII or BglII were used for full-length genes. For the initial GARP truncations, fragments containing the Rex3 N-terminus and segments of the GARP gene were fused using overlap PCR and ligated into AvrII and BglII sites. For expression of exported protein fragments fused C-terminal to the GFP tag with an N-terminal Rex3 sequence, the PA446 plasmid was modified to remove the linker between Rex3 and GFP and add a NheI site for the insertion of gene fragments after GFP. An extended linker and an AatII site was added between the GFP and the protein fragments (sequence: LESGSGTGASDV) for all but the initial GARP truncations (those shown in **Fig**

3.3). Expression was controlled by pfCAM 5′ and pbDT 3' regions and plasmids contained a Blasticidin resistance gene. An attP site within the plasmid allowed for specific integration into the parasite genome at attB sites, mediated by a Bxb1 integrase (see **2.1.4**) (Adjalley et al., 2010). The PfCAM promoter was replaced by the endogenous promoter for some proteins by cloning the promoter into the BlpI and AvrII sites or inserting both the promoter and full-length gene into the BlpI and BglIII sites. Digested vectors were either purified by gel extraction in 1% agarose or treated with calf intestinal alkaline phosphatase, followed by purification with Promega gel extraction cleanup kit in either case. DNA fragments were ligated into their respective vectors using T4 DNA ligase.

Five constructs were partly or fully cloned by Andrew Osborne: the original full-length GARP gene; the full-length gene for PF3D7_0402000; and the endogenous promoter sequences for GARP, PF3D7_1102300 and PF3D7_1476200.
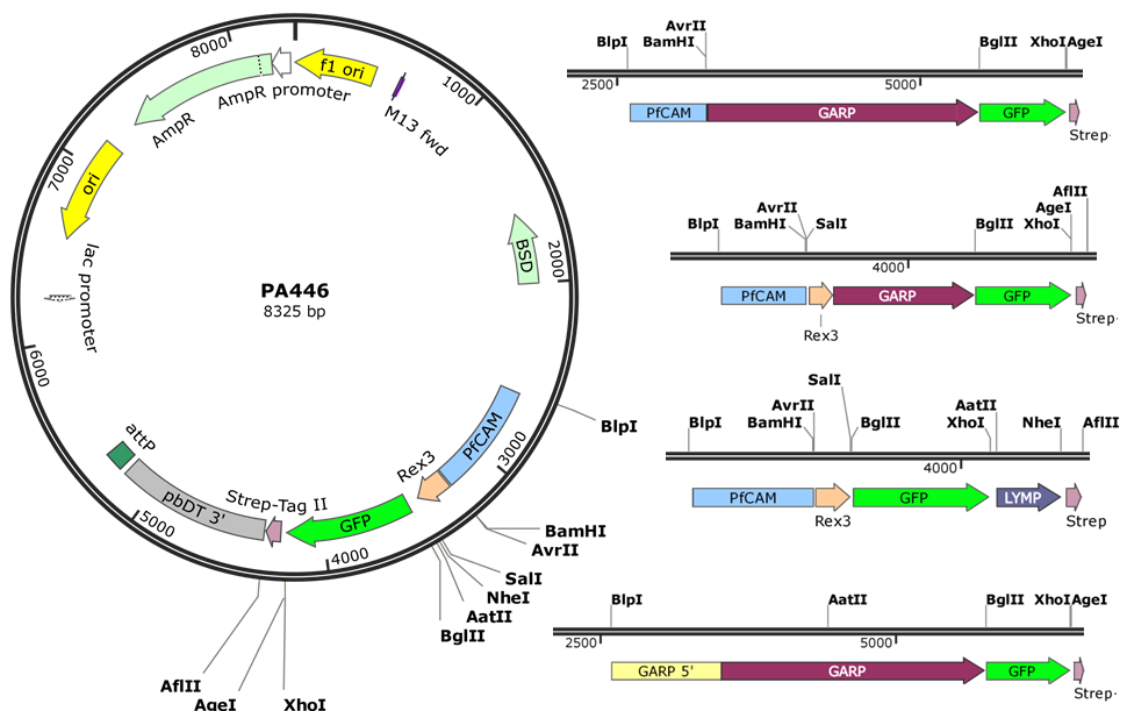
Figure 2.2 – Construct design for expression of recombinant genes by *P. falciparum* parasites. Plasmid map for PA446 is represented on the left with key components highlighted. Examples of final plasmids are shown on the right: full length genes, gene fragments 5' to GFP, gene fragments 3' to GFP and genes expressed under their endogenous promoters.

2.1.4 Purification of DNA

E. coli NEB 10-beta cells were used for ligation as they are particularly suited for large constructs. They were made electro-competent by growing up a single colony in 250mL super optimal broth (SOB containing 20 g/L Tryptone, 5 g/L Yeast Extract, 2.4 g/L MgSO4, 0.5g/L NaCl, 0.186 g/L KCl, from SIGMA) to an optical density of 0.6, then centrifuging at 2300 x g and washing three times in 10% glycerol before storing at -80 °C. These cells were transformed with 5 µL of ligation reaction mix by electroporation with the following settings: 25µF, 200ohms, 2.5KV. After 1 h of outgrowth in SOB, cells were plated on LB agar plates. A single colony was selected and grown overnight in LB broth (containing 10g/L Tryptone, 5 g/L Yeast Extract, 5 g/L NaCl, from SIGMA), before plasmid purification using a miniprep kit according to the manufacturer's instructions (SIGMA). DNA yields were analysed using a Nanodrop spectrophotometer. Correct integration of inserts was confirmed by restriction digest or colony PCR; plasmids were then sequenced by Source Biosciences or GATC sequencing services.

E. coli NEB DH5alpha cells were selected for midiprep as they are well-suited to routine cloning. Like 10-beta cells they are depleted in endonucleases preventing plasmid degradation and cannot utilise homologous recombination so genes can't be easily lost. The cells were made chemically competent cells were made by growing 250 mL of *E.coli* to OD 0.45 then washing with 100 mM $MgCl_2$ for 10 mins followed by 100 mM $CaCl_2$ for 2 h, on ice. Cells were frozen in 100 mM $CaCl_2$ with 10% glycerol. Plasmids were transformed into these cells by heat-shock for 30 s at 42 °C, before plating on LB-agar plates. Single colonies were picked and grown overnight in LB broth to saturation, then purified using Promega midi or maxiprep kits according to the manufacturer's instructions. The final plasmids were then co-precipitated with a pINT no Neomycin plasmid (Adjalley et al., 2010, van Ooij et al., 2013); 10 µg each for transfection using the AMAXA kit and 100 µg of each for BIORAD pre-loaded erythrocyte transfection. The pINT plasmid encoded a Bxb1 integrase, which integrates the plasmid attP site into an attB site in the parasite genome (Adjalley et al., 2010).

2.2 Protocols for working with *Plasmodium falciparum*

2.2.1 Culture of *P. falciparum* in Red Blood Cells

P. falciparum 3D7 strain MRA845 attB parasites (Adjalley et al., 2010) were grown in human red blood cells of various blood groups at approximately 2.5% haematocrit in RPMI containing 0.5% albumax (Trager and Jensen, 1976)(From ThermoFisher). The MRA845 parasite strain contains an attB site for integration of plasmids using Bxb integrase; this is selected for by the hDHFR gene which confers resistance to WR. Media was therefore supplemented with 2.5 nM WR. Parasite cultures were kept in chambers gassed daily for 4 minutes with 5% O_2, 5% CO_2 and balanced N_2 in 37 ˚C incubators (BOC). Parasitemia levels were regularly determined by examination of giemsa stained parasites; a drop of parasite culture was deposited on glass slides and smeared, before being fixed for 30s in 100% methanol and stained with Giemsa stain (15% Giemsa stain in 10 mM phosphate buffer, pH 7.4) (From SIGMA). Parasites were viewed using a 100x oil immersion lens.

2.2.2 Transfection of Plasmids into *P. falciparum* cells

For AMAXA transfection, the parasites were synchronised by overlaying 5 mL of culture onto a 3mL 65% percoll/sorbitol gradient (65% percoll, 3% sorbitol in PBS) and centrifuged at 2300 x g for 10 minutes to give ~95% segmenting schizont stage parasites. These were allowed to invade fresh erythrocytes for 2 h before an additional percoll step to remove schizonts which did not egress in this time. This process was repeated at the point of egress over several days to acquire tightly synchronised parasites. The pINT and expression plasmids were combined and resuspended in miliQ water and transferred to chilled cuvettes with percoll-purified parasites (segmenting schizonts) and AMAXA nucleofactor solution. Parasites were transfected by electroporation with the Amaxa Nucleofector II kit using program U33 before being added to complete media with 2% erythrocytes.

For BIORAD transfection, the expression plasmid and pINT plasmid were rehydrated in 30 μL H2O and 30 μL cytomix solution (containing 120 mM KCl0.15 mM CaCl2, 2 mM EGTA, 5 mM MgCl2, 10 mM K2HPO4/KH2PO4 pH 7.6, and 25 mM HEPES). 1 mL of uninfected erythrocytes were pelleted and

resuspended in 10mL of cytomix, before a second centrifugation step (1800 x g). 400 µL of the compacted red blood cells were added to the plasmids and cuvettes were chilled on ice for 30 minutes before electroporation with the BIORAD GenePulser Xcell kit set at 0.31KV and 950 µF. The pre-loaded erythrocytes were resuspended in complete media and mixed-stage parasites were added to a parasitemia of approximately 1%.

For both transfection methods, drug treatment with 2 µg/ml BSD was applied after 24 h to select for successfully transfected cells.

2.2.3 Microscopy of GFP-tagged fluorescent Proteins within the Parasite

Parasites were cultured to around 5% parasitemia, and fresh RPMI was added a day prior to imaging. A drop of the iRBC/RPMI media mix was placed onto glass slides and kept in place with a coverslip. Coverslips were sealed in place with 15% pig skin gelatin. A Zeiss Axiovert 200M microscope equipped with a HBO100 lamp and a 100X oil lens with a numerical aperture of 1,30 was used to observe the GFP-tagged parasites, with phase contrast and fluorescent images taken at room temperature with an AxioCam MR camera using AxioVision software release 4.8.2. Fluorescent trophozoite parasites were selected by eye, and photographed with an exposure time of 1 second for all fluorescent images and 0.1 seconds for phase contrast. Z-stacks of images were collected and deconvolved by iterative restoration (confidence limit: 95%, iteration limit: 10) using Volocity; a single image from the Z-stack is presented. Automatic brightness and contrast settings were applied to all images using ImageJ.

2.2.4 Quantification and Statistical Analysis of Images

Deconvolved images of fluorescent malaria parasites were analysed to establish the fluorescence intensity at the erythrocyte periphery relative to the erythrocyte cytosol for each parasite line (Tarr et al., 2014). A 40-pixel plot profile was taken across the erythrocyte membrane from the outside of the cell into the erythrocyte cytosol, with the cell membrane at the midpoint. Twenty images were quantified for each construct, taken from at least two imaging sessions on different days. The background and cytoplasmic signals were taken

as the average intensity of the first 5 pixels and pixels 30-35, respectively. Peripheral fluorescence was determined as the maximal signal intensity between pixels 15-25. The background signal was subtracted from both cytoplasmic and peripheral signal and the ratio of the normalised membrane to cytosolic signal was calculated to give the relative peak height for fluorescence intensity at the periphery, and the standard deviation (See **Fig 3.5** for an example and **Appendix Table 1** for quantification of all parasite lines).

To establish whether relative fluorescence intensity at the periphery was significantly enhanced relative to control protein, statistical analysis was performed on each parasite line by comparison to the GFP-tagged REX3$_{1-61}$ construct, which is efficiently exported and remains cytosolic. One-way ANOVA analysis with multiple comparisons was performed, with no correction for multiple comparisons and assuming a Gaussian distribution. Significance is indicated by asterisks, with a P-value < 0.0001 being extremely significant (****) a P-value of 0.0001 to 0.001 is extremely significant (***), P-value of 0.001 to 0.01 is very significant (**) and P-value 0.01 to 0.05 is significant (*). P-values above 0.05 are not considered to be significant (ns). Pairs of constructs were also compared in the same way to establish significant differences (e.g. for GARP truncations). See **Appendix Table 1** for statistical analysis of all parasite lines.

2.2.5 Western Blotting of Transfected proteins

Percoll-purified schizonts were prepared and resuspended in SDS-PAGE sample buffer before heating to 95 °C for 5 minutes. Approximately 2×10^6 schizonts were loaded onto each well before running the gel in running buffer composed of 25 mM Tris, 192 mM Glycine, and 0.05% SDS (SIGMA). The gel was transferred to a nicrocellulose membrane (ThermoFisher) at 100 V for 1 h with transfer buffer containing 25 mM Tris, 192 mM Glycine, 0.05% SDS and 10% methanol. The membrane was blocked with a 3% milk suspension in 50 mM Tris, 150 mM NaCl with 0.1% TWEEN for 1 hour, before adding either rabbit α-GFP monoclonal antibody (Torrey Pines, Catalog Number: TP401, Lot Number: 071519) at a dilution of 1:4000 or mouse α-histoaspartic protease (HAP – Monoclonal Antibody 3F10-6, catalogue number MRA-811A) diluted 1:10000 (Liu et al., 2005) followed by three washes with milk solution, then the

addition of goat α-rabbit (ThermoFisher, Catalog number: 35568, Lot number: OK195926) or goat α-mouse secondary antibodies (ThermoFisher, Catalog number: 35521, Lot number: LB143097) containing a fluorescent tag, at a dilution of 1:10,000, with subsequent washing. The membrane was imaged with a LI-COR Odyssey imager with excitation at 700 nm or 800 nm depending on the antibody used. All western blots which are not shown in the main text are shown in **Appendix Figure 1.**

2.2.6 Extraction of Proteins from Erythrocytes by their Solubility Profile

Parasites were synchronised using a 65% percoll gradient to enrich for schizonts. To make red blood cell ghosts and remove the soluble components of the erythrocyte, the schizonts were resuspended in 50 μL 7.5 mM sodium phosphate pH 7.4 with 1 mM EDTA (SIGMA) in order to hypotonically lyse the cells. They were centrifuged at 18,500 x g for 10 min at 4°C and the supernatant was removed from pelleted 'ghosts'. To make triton shells, the ghost pellet was re-suspended in 50 μL chilled 1% triton X-100 in 50 mM Tris pH 8, 150 NaCl, 2 mM EDTA (all SIGMA), and once again centrifuged at 18,500 x g for 10 min before separating the supernatant and insoluble pellet (Black et al., 2008).

The pellets were resuspended to the same volume as the supernatant in buffer (i.e. 50 μL), and SDS-PAGE loading buffer was added before heating to 95 °C for 10 min then applying the equivalent of 2×10^6 schizonts to each lane of an SDS-PAGE gel. The gel was transferred onto nitrocellusose at 100 mV, 400 A over 1 h transfer time in running buffer containing 10% methanol. Blots were probed with rabbit anti-GFP (Torrey Pines), rat anti- β spectrin or mouse anti-band 3 to confirm successful extraction of fractions.

2.2.7 Preparation of Gametocytes

Asexual parasites in RPMI were adjusted to 0.1 % parasitemia and 6% haematocrit in 12.5 mL of RPMI media (Day 1). The RPMI was exchanged for 12.5 mL of fresh media on day 3, and exchanged for 25 mL of fresh media on days 4-11. Parasites were not passaged into new blood and were allowed to reach very high parasitema (~50%). Once a sufficient number of gametocytes had developed (Day 9), 2.8 ml of 0.5 M N-acetyl glucosamine (NAG -SIGMA)

was added to the culture to remove asexual parasites. This was replenished along with fresh RPMI each day for 3 days. On day 12, gametocytes were enriched by applying to a 65% percoll/sorbitol density gradient with centrifugation at 2300 x g for 10 minutes (Tanaka et al., 2013). The number of purified gametocytes was counted using a hemocytometer and normalised with sample buffer to the concentration of asexual parasites in a control culture, before applying 2×10^6 parasites to each well for SDS-PAGE gel electrophoreisis and western blotting.

2.2.8 Lipid Kinase Inhibitors

Wortmannin and phenyl-arsine oxide (SIGMA) were dissolved in DMSO and added to parasites expressing the full-length GARP-GFP construct at final concentrations of 10 µM and 2 µM, respectively. Bright field and fluorescence images of parasites were taken 5, 30 and 80 minutes following the addition of each compound.

2.3 Bioinformatics

2.3.1 Protein and DNA sequence analysis

Sequence disorder predictions were preformed using DISOPRED, and secondary structure predictions with PSIPRED (Buchan et al., 2013). PROTPARAM was used for calculating the predicted isoelectric point as well as amino acid percentages in individual proteins. Proteins were aligned in Jalview using standard TCOFFEE alignment (Waterhouse et al., 2009) .

Plasmids were designed in APE and plasmid maps were built with SNAPGENE. Artemis software was used for genomic analysis.

2.3.2 Comparison of GARP genes from Closely-Related Parasite Species

A GARP homologue from *P. gaboni* was assembled from two incomplete protein sequences deposited in the NCBI, which were found using the inbuilt BLAST server (GenBank accession numbers: KYN95113.1 and KYN95116.1); the connecting region was assembled from sequence reads, which were also found using BLAST (GenBank Biosamples SAMN04053641 and SAMN04053639) (Sundararaman et al., 2016). The *P. reichenowi* GARP gene (PRCDC_0111200) was acquired from PlasmoDB (Version 26).

2.3.3 Identification of putative, exported, lysine-rich, repeating protein sequences

Protein coding sequences from *P. falciparum*, *P. vivax, P. knowlesi, P. cynomolgi, P. reichenowi, P. berghei, P. chabaudi* and *P. yoelii (17X)* were downloaded from PlasmoDB (Version 26). Putative exported proteins where identified by the presence of either a signal sequence (defined by SignalP (Petersen et al., 2011)), or a transmembrane domain within the first 100 residues (defined using MPEX translocon TM analysis (Snider et al., 2009)) , and an RxL motif (where x is any amino acid) in the 50 residues following the signal sequence/transmembrane segment. Proteins containing more than 4 transmembrane segments within the coding sequence are unlikely to be exported and were excluded from further analysis (see **Appendix Text 1** for perl script).

A custom perl script utilising a sliding-window algorithm was used to identify proteins containing stretches of amino acids of at least 30 residues in length and with a lysine content of at least 20%. Within the set of lysine-rich sequence fragments, repeating protein sequences were identified using the tandem repeat predictor T-REKS (standalone version, using a PSIM cut-off of 0.65) (Jorda and Kajava, 2009) (See **Appendix Text 2** for perl script). Another custom perl script was used to select proteins in which the sequence region comprised of repeats was over 30 residues in length with an individual repeat unit length below 30 residues. Multiple lysine-rich repeat sequences were found in some proteins (See **Appendix Text 3** for perl script). The program XSTREAM (Newman and Cooper, 2007) was used to define the consensus sequence of each repeated array, the consensus error value for each repeat array, and the position of the repeated array within each protein. Parameters for XSTREAM were as follows; min word match = 0.6, min consensus match = 0.65, miss penalty = -3, and gap penalty = -5. Max period value was set to 40 residues unless shorter repeats were apparent within the predicted consensus sequence; other parameters were set to default values. Theoretical isoelectric point values were predicted by PROTPARAM (Gasteiger et al., 2005).

2.3.4 Quantification of amino acid usage and repetitive sequences across species

For the quantification of amino acid usage across *Plasmodium* species, protein sequences for *P. falciparum, P. reichenowi, P. vivax, P. knowlesi, P. cynomolgi, P, berghei, P. chabaudi*, and *P. yoelii* were downloaded from PlasmoDB (version 26). A perl script was written to scan all protein sequences for each of the 20 common amino acids, using global regular expression to search for single letter codes. The number of each amino acid was counted and the total printed out for each species.

The same fasta protein sequences were used to calculate the number of proteins containing repetitive sequences across all *Plasmodium* species. The protein sequences were input into the repeat prediction program T-REKS. The output of T-REKS was analysed by a variation of the perl script 'pickrepeats.pl' as described earlier. This counted proteins containing repetitive sequences over 30 residues in length, and printed out the results for each species.

2.3.5 Sequence analysis of proteins from different parasite isolates

The protein sequences of lysine-rich proteins from different *P. falciparum* parasite strains were extracted from unassembled long-read PACBIO genome sequencing data obtained from the Pf3K consortium. Five lab isolates were included (3D7, DD2, IT, 7G8, and HB3) and eleven field isolates from Gabon, Guinea, UK, Kenya, Mali, Sudan, Senegal, Democratic Republic of the Congo, Togo, and Cambodia. No KAHRP genes were found in the DD2 or Kenyan isolates. LYMP was not found in one of the two Cambodian isolates. All alignments were created with T-COFFEE (Notredame et al., 2000), and represented with 'Multiple Align Show' (Stothard, 2000). 10 out of 141 gene sequences, indicated in **Appendix Table 2**, contain frameshift point mutations. It is unclear if these represent genuine mutations or sequencing errors in database sequences; for the purpose of sequence alignment the reading frames were restored.

2.3.6 Sequence analysis of the KAHRP conserved domain

Proteins with homology to the conserved domain of KAHRP and PfEMP3 were identified by HMMer (Finn et al., 2011). Additionally, homologous sequences within the *P. ovale* genome were identified from unassembled sequence reads acquired from the Sanger Institute through the use of the in-built Blast server. Sequence reads from *P. ovale* proteins containing EKAL domains were assembled using the SEQman Ngen software (DNASTAR). Introns were manually annotated within genes from *P. ovale*, *P. fragile*, *P. inui* and *P. cynomolgi* where necessary. Potential sequencing errors resulting in frameshift mutations were corrected, and introns were annotated based on known *Plasmodium* splice sites. These modifications were made in five proteins from *P. inui* and *P. cynomolgi* (see **Appendix Table 3**). Sequences were aligned with T-COFFEE (Notredame et al., 2000) in Jalview (Waterhouse et al., 2009). Maximum likelihood estimation with TREE-PUZZLE (Schmidt et al., 2002) was used to create phylogenic trees based on an extended conserved domain, and assembled with FigTree v1.2.4 (Rambaut, 2014). Secondary structure predictions and disorder predictions were made by PSIPRED (Buchan et al., 2013) and DISOPRED (Ward et al., 2004), respectively.

Chapter 3: The Glutamic-Acid Rich Protein (GARP): A Highly Charged Repetitive Protein Targeted to the Erythrocyte Periphery

3.1 Abstract

In a screen of GFP-tagged *P. falciparum* proteins predicted to be exported into the host cell, the glutamic acid-rich protein (GARP – gene identifier PF3D7_0113000) was found to localise to the periphery of the host erythrocyte. By expressing short fragments of GARP, we reveal that this targeting is conferred by the lysine-rich repetitive sequences of the protein, with three sequences composed of different repeated motifs each independently sufficient for localising to the periphery. Preliminary experiments indicate these sequences likely interact with components of the erythrocyte cytoskeleton. Truncation of one lysine-rich repeat indicated that targeting efficiency was dependent on the number of repeats present, indicating an avidity-based interaction. While *P. falciparum* GARP has expanded three functional repetitive sequences, the closely related species *P. reichenowi* and *P. gaboni* have expanded different non-targeting motifs or lack some repeats completely. We present a model for the *de novo* formation of functional targeting sequences in *P. falciparum* by expansion of lysine-rich motifs.

3.2 Introduction

The identification of proteins involved in erythrocyte modification is complicated by the difficulty in modifying the parasite genome. Gene deletions were traditionally conducted using single or double-crossover homologous recombination to substitute or modify the gene of interest (de Koning-Ward et al., 2015). Following transfection and drug selection, live parasites are usually observed after a period of between one and three months using this method, due to the rarity of integration events (Maier et al., 2006). One large screen of this kind has been conducted to date; the genes of several proteins containing a HT/PEXEL motif as well as proteins which were not predicted to be exported were deleted, and the effect on erythrocyte properties such as rigidity and

cytoadhesion were measured (Maier et al., 2008). This indicated that several genes are essential for asexual parasite growth; however it is also possible that a failure to recombine the plasmid was responsible for this phenotype (Maier et al., 2008). For successfully deleted genes it is often difficult to establish what constitutes a significant change in phenotype; several aspects such as the life stage of the parasites, parasitemia, and the levels of PfEMP1 surface antigen presented on the cell surface may affect deformability and cytoadhesive properties of the host cell (Kriek et al., 2003, Nash et al., 1989). Tightly synchronised parasites selected for a particular PfEMP1 antigen are therefore required, which is complicated by the long timescales required for parasite levels to recover following transfection.

New technologies have greatly facilitated genome editing of *P. falciparum* parasites. The CRISPR/Cas9 system allows rapid and efficient site-directed modification of the genome (Ghorbal et al., 2014), while several conditional gene knock-down or deletion systems have enabled essential proteins to be probed. Proteins may be fused to a ligand-dependent FKBP protein destabilization domain (ddFKBP) (Armstrong and Goldberg, 2007) or genes may be flanked with LoxP sites and rapidly deleted by the addition of Cre recombinase (Jones et al., 2016). However, the cloning process is often laborious for these techniques, with multiple steps and recodonization of DNA often required (Jones et al., 2016). A simpler method for analysing protein localisation involves transfecting parasites with recombinant GFP-tagged proteins in plasmids which may be maintained episomally or integrated into the genome at a specific site. The p-INT integrase, for example, integrates plasmids containing an attP sequence into attB sites previously inserted into the parasite genome (Adjalley et al., 2010, Nkrumah et al., 2006). These proteins may be cloned with their endogenous promoters or with well-established strong promoters such as the PfCAM promoter for a clear fluorescent signal (Polson and Blackman, 2005).

Prior to my arrival in the lab, a number of proteins predicted to be exported into the host cell were GFP-tagged and expressed from the PfCAM promoter. This screen identified a number of proteins which were localised to the periphery of the host cell, indicating an interaction with the erythrocyte membrane or the underlying cytoskeleton where important virulence-related changes are known

to occur. It is therefore likely given their localisation that these proteins may be involved in these modifications. One protein identified was the glutamic acid-rich protein (GARP), which was strongly targeted to the periphery of infected erythrocytes (**Fig 3.1A**). As no function has previously been assigned to GARP, we aimed to characterise its role in the host cell.

GARP is an 80 KDa protein expressed by the parasite during the early trophozoite stage of the life cycle (Triglia et al., 1988). Like many *Plasmodium falciparum* proteins, it is highly repetitive and it is also enriched in charged residues; lysine, glutamate and aspartate constitute over half the amino acids in the protein. GARP is expected to be exported into the host erythrocyte; it contains an N-terminal signal sequence followed by a HT/PEXEL motif which is predicted to be cleaved in the endoplasmic reticulum, targeting the protein into the host cell. GARP contains five tandemly repeating sequences; the first four repetitive sequences contain mostly basic lysine residues while an acidic 100-residue sequence at the C-terminus is predominantly composed of glutamic acid residues, giving the protein its name (**Fig 3.1B**). The lack of hydrophobic stretches within the protein suggests that it does not form a folded globular domain. Indeed, the protein disorder prediction program DISOPRED indicates the entire mature GARP sequence is intrinsically disordered (**Fig 3.1C**) (Ward et al., 2004).

As with many repetitive sequences encoded by the parasite, GARP appears to be highly antigenic. Due to its reactivity to sera from adults in malaria endemic regions in Papua New Guinea, the protein sequence was deciphered in 1988 (Triglia et al., 1988), however it has not been studied in detail since. A number of genome-wide experiments have indicated GARP may play an important role in the host cell; analysis of parasites infecting children with severe malaria indicated that GARP RNA transcripts were upregulated fivefold relative to parasites extracted from adults (Vignali et al., 2011), and different expression levels were also found in parasites with different PfEMP1 genes expressed (Ralph et al., 2005). GARP may therefore be involved in modulating cytoadhesion of infected erythrocytes. The large gene deletion study previously mentioned did not observe a change in cytoadhesion upon deletion of GARP however, although some decrease in cytoskeleton rigidity was observed (Maier et al., 2008).
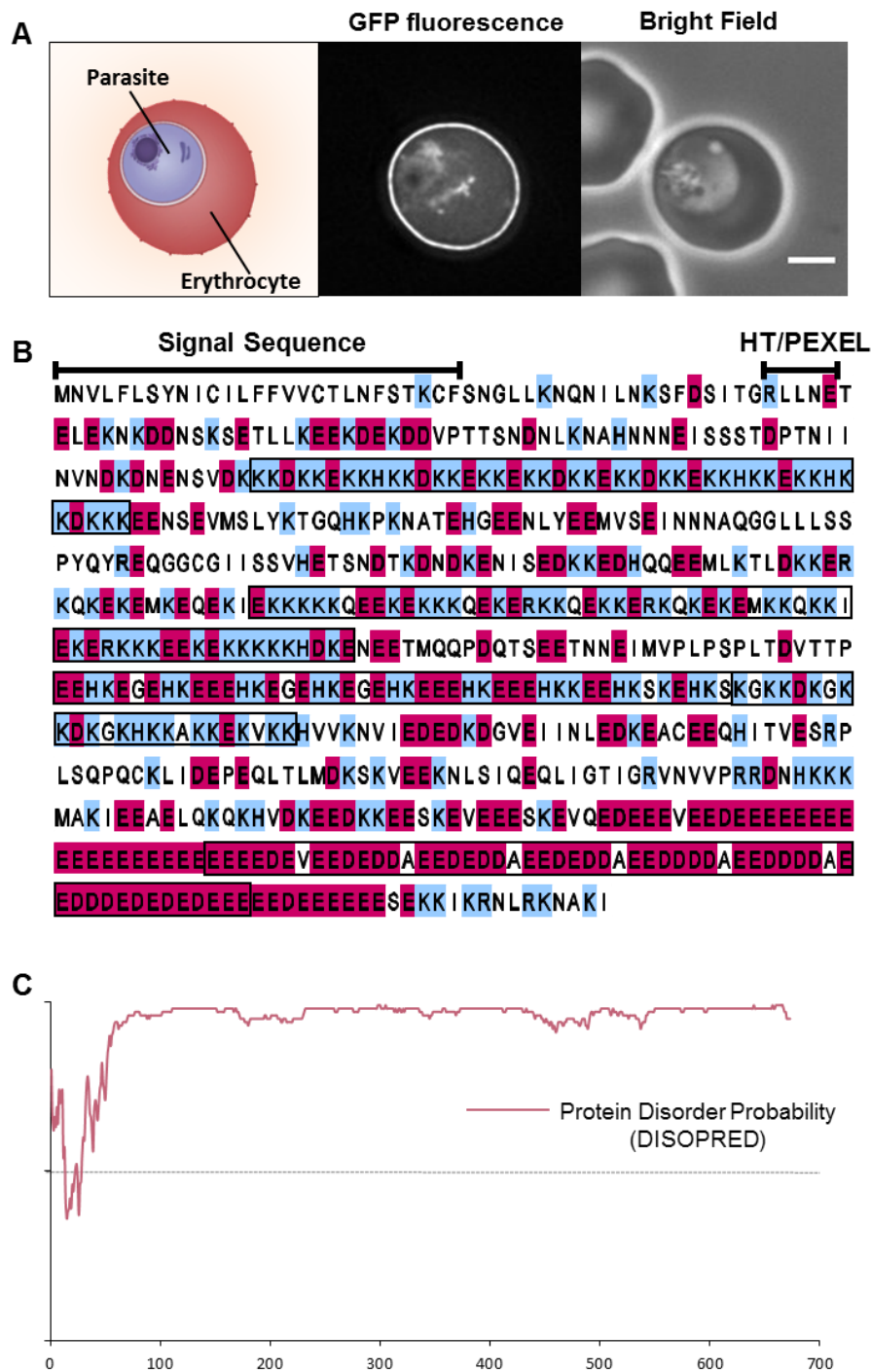
Figure 3.1 – The Glutamic Acid-rich protein (GARP) localises to the periphery of the host erythrocyte. (A) GFP-tagged GARP is localised to the periphery of the infected erythrocyte. Schematic of the infected cell, fluorescent image and phase contrast images are shown from left to right. Scale bar – 2µm. (B) The protein sequence of GARP. Basic residues are shaded blue and acidic residues in red. The five repetitive sequences are boxed and the signal sequence and HT/PEXEL motif are indicated. (C) Disorder prediction for GARP (using DISOPRED) (Ward et al., 2004). Amino acids are considered disordered if they have a confidence score over 0.5, represented by a dotted line.

To gain a greater understanding of the role of GARP and the domains responsible for its localisation at the erythrocyte periphery, multiple truncations of the protein were GFP-tagged and expressed in *P. falciparum*. These indicated several novel targeting sequences in the protein, and comparison with closely related species presents new insights into the evolution of functional disordered and low complexity sequences in the parasite.

3.3 Results

3.3.1 Truncations of GARP indicate multiple targeting sequences are present in the protein

To narrow down the protein sequence responsible for targeting GARP to the erythrocyte periphery, the sequence was split into three segments and cloned in-frame with a C-terminal GFP tag. As fragments lacking the N-terminal HT/PEXEL motif would not be exported into the host cell, these sequences were fused to the N-terminal 61 residues of the REX3 protein which has been used previously to mediate export and does not interact with any erythrocyte components on its own (Tarr et al., 2014, Sargeant et al., 2006). The N-terminal 174 residues of the protein, $GARP_{1-174}$, comprised the signal sequence, the endogenous HT/PEXEL motif and the first lysine-rich repeating sequence. The consensus sequence for this repeated motif, as defined by the repeat prediction program XSTREAM, is EKK (Newman and Cooper, 2007). This indicates the most common residues in each position; however the first position alternates between E, D, H and K (see sequence logo, **Fig 3.2A**). When expressed in *P. falciparum* parasites, this fragment was robustly exported into the host cell where it localised at the periphery of the erythrocyte, as with the full-length protein (**Fig 3.2B**). Fluorescence was also observed within the parasite; this is common for exported proteins which must be trafficked *via* the endoplasmic reticulum and are often internalised into the food vacuole of the parasite as it feeds on haemoglobin from the erythrocyte cytosol (Wickham et al., 2001).

A construct containing the middle fragment, $GARP_{175-523}$, encompassed another three tandemly repeated regions. The first contains seven repeats of the imperfectly repeated motif E-KE--K-KKQ-; dashes indicate the length of the repeated motif varies and that gaps are the most common feature within a position. The third and fourth repeats are adjacent to each other and represent nine EEHKE motifs followed by five repetitions of KGKKD. This construct encompassing all three repeats also localised GFP to the erythrocyte periphery, indicating there are multiple targeting sites within GARP (**Fig 3.2C**).

The C-terminal truncation, $GARP_{524-673}$, contained the acidic C-terminus of the protein, including 8 iterations of the sequence EEDEDDA. The GFP

fluorescence in this case was seen in the erythrocyte cytosol only, with no indication of peripheral targeting by this fragment (**Fig 3.1D**).
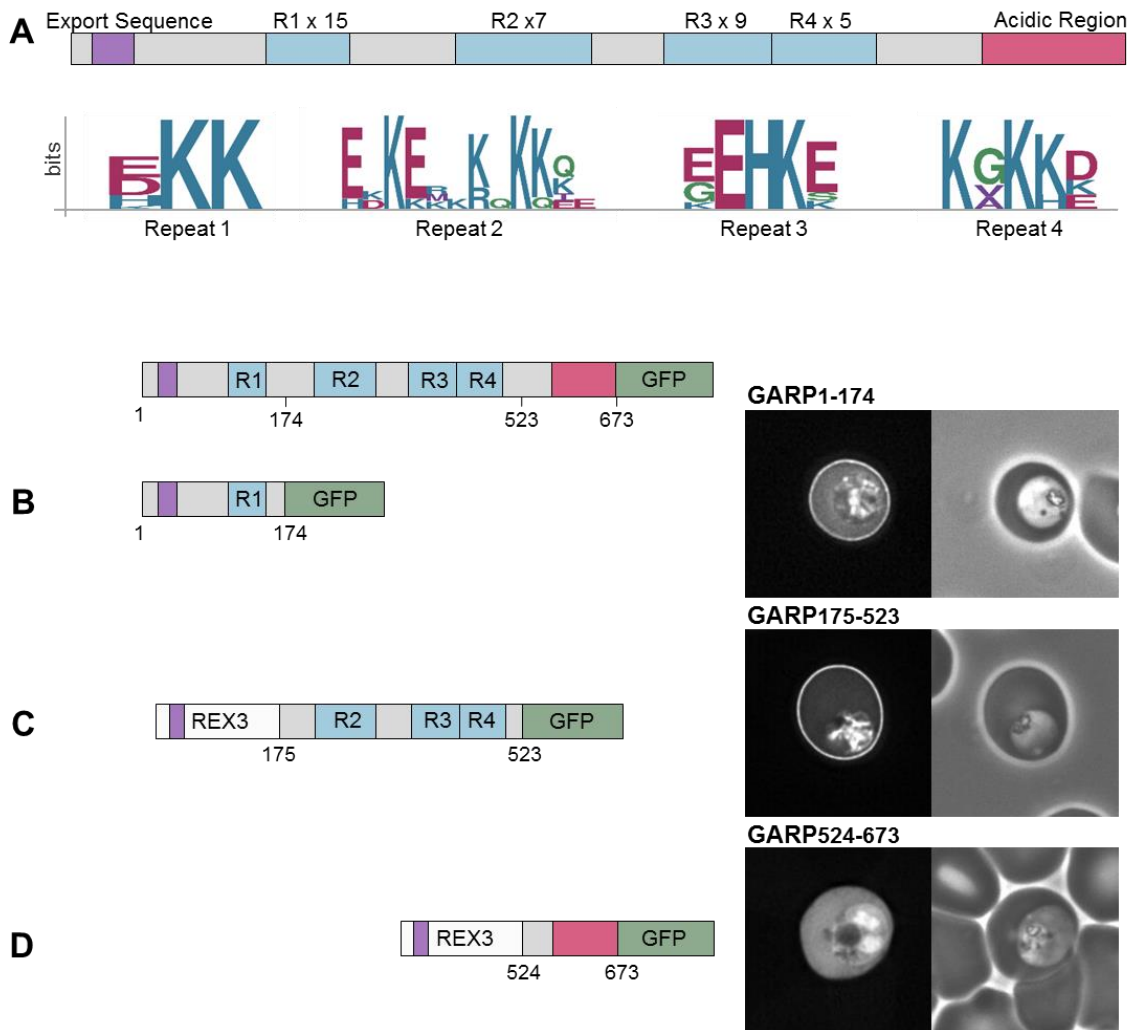
Figure 3.2 – Multiple sequences within the Glutamic acid-rich protein (GARP) target the erythrocyte periphery. (A) Representation of GARP. The four lysine-rich repeating regions are highlighted in blue, and labelled R1-R4. The number of iterations of each tandem repeat is also shown. The export sequence represents both the ER-targeting signal sequence and the HT motif for export. Sequence logos for the four lysine-rich repeats are shown below with conservation of each residue position represented on y-axis (bits). (B-D) GFP-tagged GARP truncations expressed in *P. falciparum* parasites. Construct representation shown to the left, GFP fluorescence and bright field images are shown on the left and the right panels, respectively.

3.3.2 Lysine-rich repeating sequences within GARP localise to the infected erythrocyte periphery

To establish whether the first fragment, $GARP_{1-174}$, was directed to the erythrocyte periphery *via* its lysine-rich repeats, two new constructs were created by splitting the sequence into two halves: one, $GARP_{50-118}$, extended from the HT/PEXEL motif of the protein to the start of the repeats, while the other, $GARP_{119-163}$, contained the first lysine-rich repeat sequence (see **Fig 3.2A**). Both these constructs and all N-terminally truncated proteins hereafter were fused C-terminal to the GFP tag as the positioning of basic residues immediately adjacent to the REX3 HT-motif was shown to hinder export, possibly by impeding insertion into the ER. The construct configuration was therefore $REX3_{1-61}$: GFP: GARP fragment: Strep tag. Additional constructs were made encompassing the second lysine-rich repeat region in insolation as well as the sequence comprising the third and fourth repeat regions (constructs $GARP_{372-446}$ and $GARP_{253-340}$, respectively). When expressed by *P. falciparum* parasites, all three of the sequences containing lysine-rich repeats exhibited strong peripheral fluorescence (**Fig 3.3B-D**). Conversely, the non-repetitive and uncharged N-terminus of the protein, $GARP_{50-118}$, was localised in the erythrocyte cytosol (**Fig 3.3A**). To confirm that the position of the GFP tag did not influence its targeting, the acidic C-terminus of GARP ($GARP_{535-673}$) was also cloned after the GFP tag, and remained cytosolic as observed previously (**Fig 3.3E**).

The targeting of GARP to the erythrocyte periphery can therefore be attributed to three lysine-rich repeating sequences. Each sequence is composed of different repeating motifs, suggesting that this may be a general phenomenon which may apply to other similar sequences. The repetitive but acidic C-terminus did not localise to the erythrocyte periphery, confirming that this is not a characteristic of all repeating sequences, although this sequence may play another role at the periphery of the host cell.
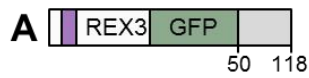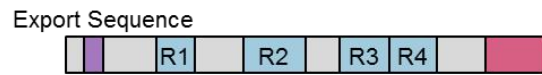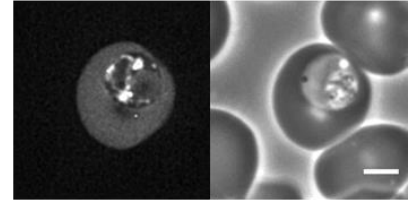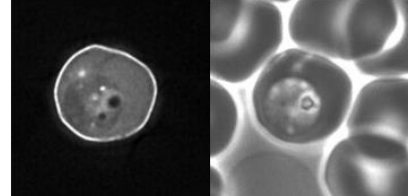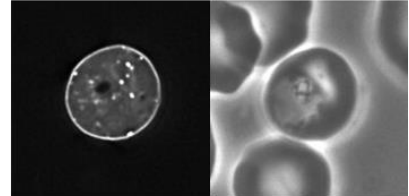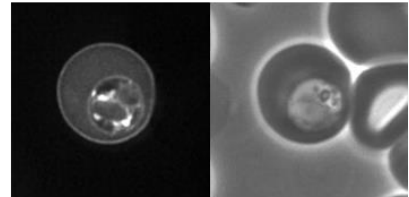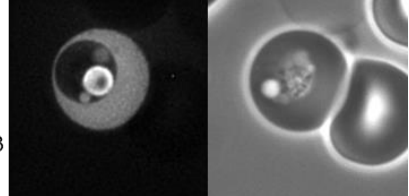
Figure 3.3 – The Lysine-rich repetitive sequences of GARP target the erythrocyte periphery. (A-E) GFP-tagged truncations of GARP expressed in *P. falciparum* parasites. A schematic of each construct, GFP fluorescence and bright field images are shown from left to right. The export sequence represents the signal sequence and HT/PEXEL motif. Lysine-rich repeats are shaded blue and labelled according to position. Scale bar – 2µm.

3.3.3 The targeting efficiency of lysine-rich repeat sequences is length-dependent

The repetitive nature of all peripherally-localised sequences within GARP suggests that this may be an avidity-based interaction where each motif contributes to the overall affinity for a binding partner at the cell periphery. In this case, the length of the sequence may directly affect the targeting efficiency. To approximately define the minimum length required for peripheral localisation, the first lysine-rich repeat sequence of GARP was truncated from 45 residues to 30 and 15 residue segments; containing fifteen, ten and five repeats of the (EDKH)KK motif, respectively **(Fig 3.4)**. These three constructs contained a REX3 export sequence followed by GFP, with the GARP fragment at the C-terminus. An extended linker sequence was inserted between GFP and the GARP fragment, to ensure no protein interactions were disrupted by the proximity of their binding site to GFP. GARP$_{119-163}$(+linker) exhibited a clear localisation at the erythrocyte periphery, as seen for the same sequence without the linker, indicating the linker did not affect targeting (**Fig 3.4A**). A ring of GFP fluorescence at the membrane was also apparent for the truncated fragment GARP$_{134-163,}$ although the peripheral targeting was less efficient than that of the longer fragment; a greater amount of GFP fluorescence was seen in the erythrocyte cytosol in this case **(Fig 3.4B)**. The shortest fragment of 15 residues, GARP$_{149-163}$, was not efficiently recruited to the periphery relative to the larger fragments (**Fig 3.4C**).

To confirm these observations, twenty images of parasites expressing each GFP-tagged fragment were analysed using ImageJ. Plots of fluorescence intensity across the membrane are shown in **Fig 3.5**. A peak in fluorescence is apparent at the periphery for the fifteen and ten-repeat fragments (**Fig 3.5A-B**), whereas the shorter fragment shows a more uniform distribution throughout the erythrocyte (**Fig 3.5C**). The intensity of the fluorescent signal at the periphery relative to the erythrocyte cytosol was quantified for each image, and normalised to the background signal. The fold difference at the periphery relative to the cytosol was calculated for each image and the average is shown for all three constructs. This indicates a clear and significant reduction in targeting efficiency as the number of repeats decreases (**Fig 3.5D**). While some

fluorescence was observed at the periphery for the 5-repeat fragment, this was not significant and likely results from very low affinity for peripheral components.

These data suggest that in the context of the first GARP repeat, a sequence of approximately 30 amino acids in length is necessary for robust peripheral targeting. This suggests that each repeated unit has some affinity for the binding partner, collectively contributing to an increased avidity of the interaction.
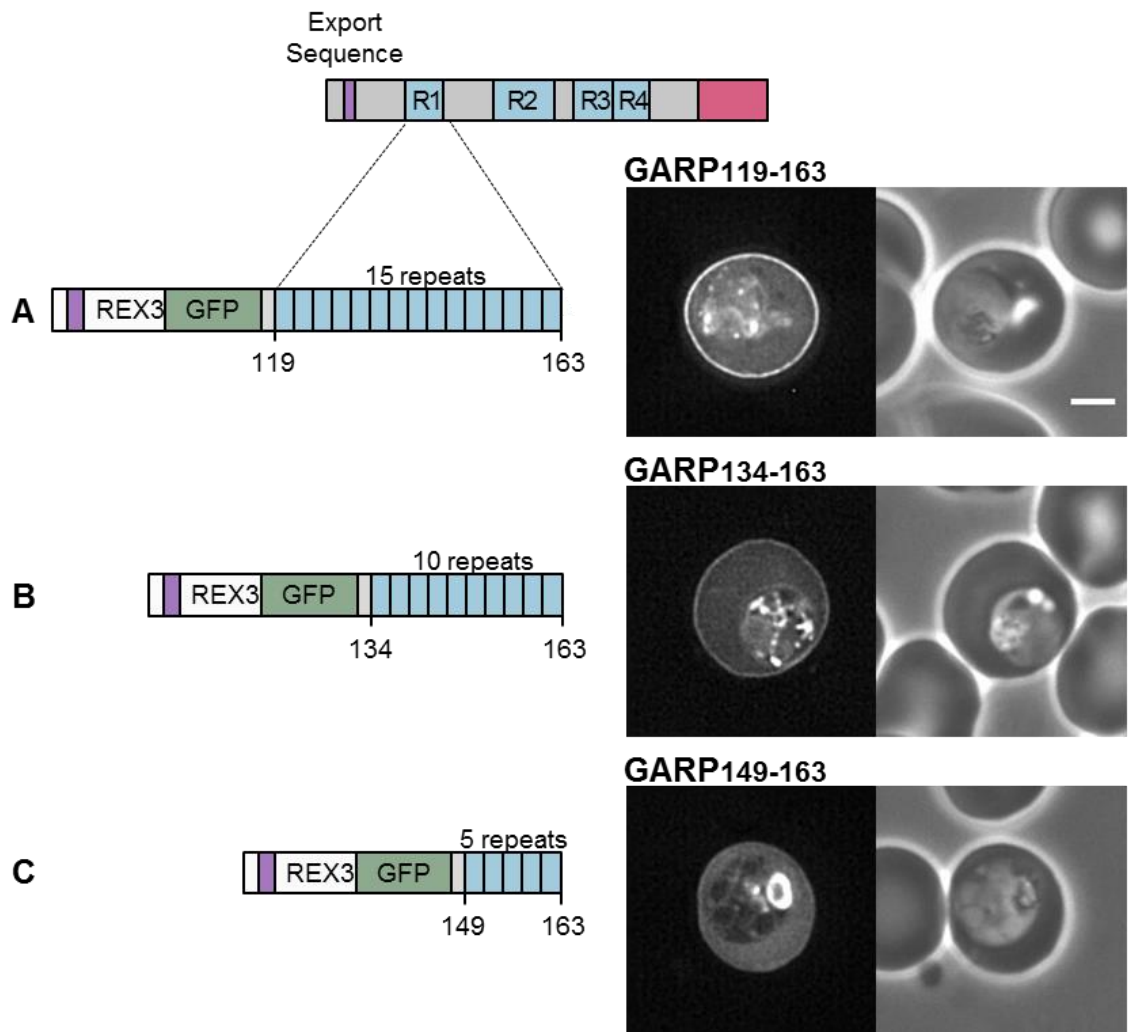
Figure 3.4 – Truncating the first charged repeat of GARP decreases targeting efficiency. (A-C) Truncated fragments of the first lysine-rich repeat sequence (blue) of GARP, GFP-tagged and expressed in *P. falciparum* parasites. Expressed proteins are shown on the left, with dotted lines indicating the region cloned from the full-length protein. The export sequence (purple) of REX3 was used to drive export of each fragment. An additional linker was placed between the GFP and GARP fragments. GFP fluorescence and phase contrast images are shown in the left and right panels, respectively. Scale bar – 2µm.
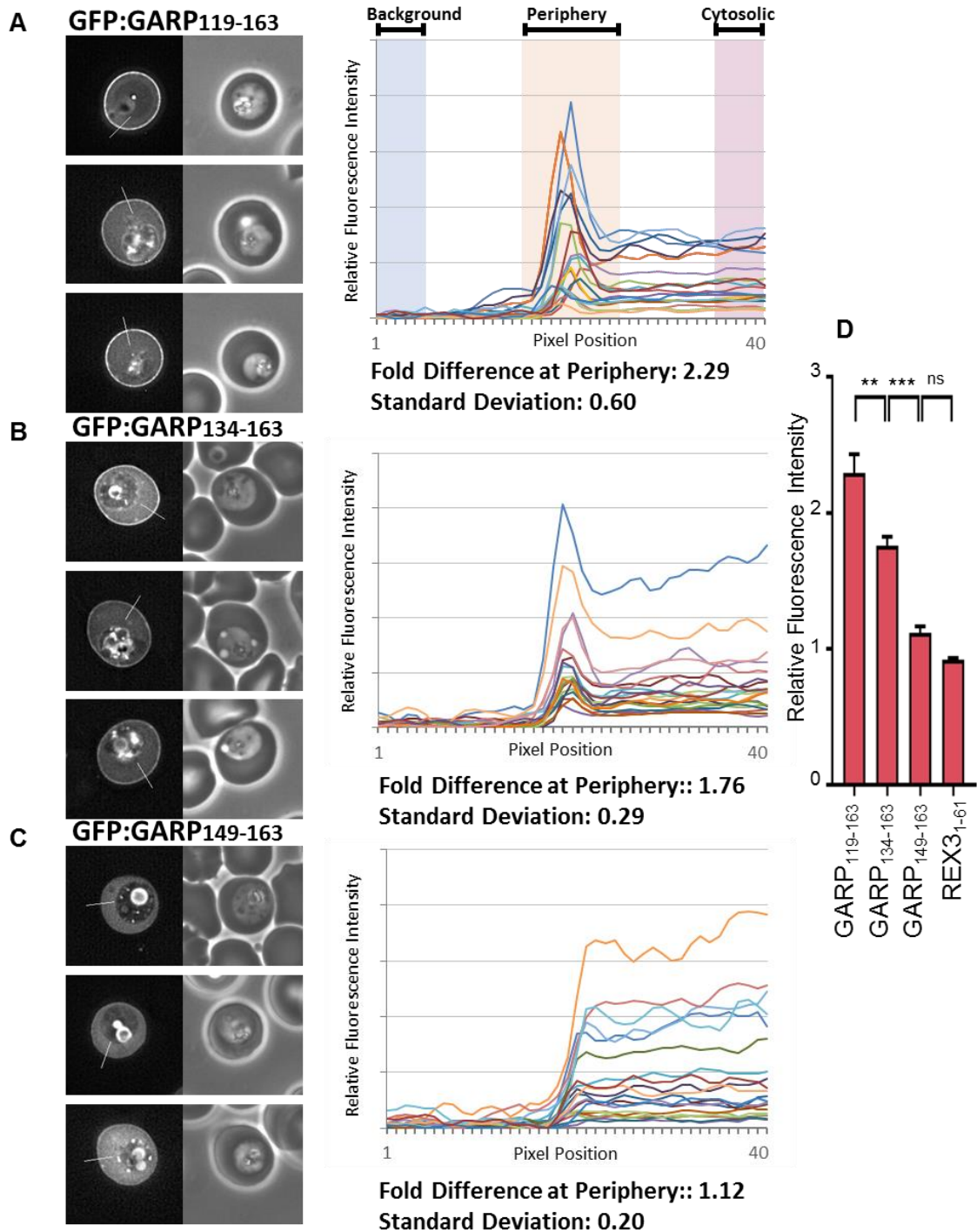
Figure 3.5 - Quantification of GFP fluorescence at the periphery of infected erythrocytes. A 40-pixel plot profile was taken across the erythrocyte periphery. The background and cytoplasmic signals were taken as the average intensity of pixels 1-5 and 30-35, respectively. Peripheral fluorescence was determined as the maximal signal intensity between pixels 15-25. The ratio of the normalised peripheral to cytosolic signal was calculated. (D) Statistical analysis by one-way ANOVA with multiple comparisons was performed, with significance indicated by asterisks with a P-value < 0.0002 (*** - extremely significant), P-value < 0.0021 (** - very significant) and ns – not significant. Error bars: SEM.

3.3.4 Expansion of repeating lysine-rich sequences can generate sequences with a targeting function in exported parasite proteins

Expansion and contraction of repeating sequences commonly occurs through DNA slippage during replication or unequal crossover during meiosis (DePristo et al., 2006). All targeting sequences within GARP are composed of repetitive sequences, with targeting efficiency shown to be dependent on the number of repeats. It is therefore likely that repeat expansion may have contributed to the evolution of these sequences.

To investigate this, the protein sequences of *Plasmodium* species closely related to *P. falciparum* were compared. The chimpanzee-infecting species *P. reichenowi* and *P. gaboni* have recently undergone full genome sequencing, and are some of the closest relatives to *P. falciparum* found to date (Sundararaman et al., 2016, Otto et al., 2014). Both contain a homolog of GARP, which is not found in any other sequenced *Plasmodium* genomes. In *P. reichenowi*, the syntenic GARP sequence is mostly conserved with *P. falciparum* GARP; it contains an acidic C-terminus and the second, third and fourth lysine-rich repeat regions are similar. However, within the region of the first lysine-rich repeats the sequence diverges: the homologous region of *P. reichenowi* GARP contains only five repeats of the (EDHK)KK motif whereas *P. falciparum* GARP has fifteen **(Fig 3.6 and 3.7A)**. Instead, a more acidic DE(TK) repeat has expanded in *P. reichenowi* GARP.

In *P. gaboni*, more differences are apparent. While the acidic C-terminus and second lysine-rich repeat region of the protein are mostly intact, the other repetitive regions have undergone major changes: the third and fourth repeats have not expanded in this parasite species and instead an alternative lysine rich repeat with the consensus motif D(KQ)KKEQTEEHKHD has repeated five times near the C-terminus of the protein. Another insertion consists of five repetitions of the eight-residue sequence KP(AT)D(EVA)A(QL)H. Additionally, the first lysine rich repeat region has expanded differently again in *P. gaboni* compared to *P. falciparum* and *P. reichenowi*; a (HDN)KN repeat has expanded in this region in addition to four repeats which conform to the *P. falciparum* (EKDH)KK motif **(Fig 3.6 and 3.7A).** Analysis of the DNA sequence encoding the first lysine-rich repeat region indicates that the variation in repeated motifs is the

result of divergent expansion of different sequences rather than a frame-shift leading to an alternative open reading frame.

```
P. falciparum GARP   MNVLFLSYNICILFFVVCTLNFSTKCFSNGLLKNQNILNKSFDSITGRLL   50
P. reichenowi GARP   MNVLFLSYNICILFFVVCTLNFSTKCFSNDLLKNQNILNKSFAFITGRLL   50
   P. gaboni GARP    MNVLFLSYNICILFLFVCTLKFSTKCFSNDLLKNQNIVNKTFNFKSRRLL   50

P. falciparum GARP   NETELEKNKDDNSKSETLLKEEKDEKDDVPTTSNDNLKNAHNNNEISSST   100
P. reichenowi GARP   NEAELEKNKDSTNSETLLKDEKDET---------------------    76
   P. gaboni GARP    DETELEKNKDNSTSESILKSEKDKKDNTHGTSNDNSKDFYENNKFFLST   100

                                                                        R1
P. falciparum GARP   DPTNIINVNDKDNENSVD------------KKKDKKEKHKKDKKEKKE   137
P. reichenowi GARP   -----------------DETDEKDETDEKDETDETDEKDEKHKKH   104
   P. gaboni GARP    DQIDITDVNDKDNENSASGTNNNVNKKKKKNHKNDKNSKNSKNNKNSKNN   150

P. falciparum GARP   KKDKKEKKDKKEKKHKKEKKHKKDKKKEENSEVMSLYKTGQHKPKNATEH   187
P. reichenowi GARP   KKDKKETNEKEENVKESSGKKKKEKKTKQNNEMMSLYKTGQHKPKNATEH   154
   P. gaboni GARP    KNNKNNKKNNYDKNNKNDKKKKKGKKKKQNNEVMSLYKTGQHKPKNARDN   200

P. falciparum GARP   GEENLYEEMVSEINNNAQGGLLLLSSPYQYREQGGCGIISSVHETSNDTKD   237
P. reichenowi GARP   AEQNLHNEMVSEINNNEQSESLLLSSPCGYSEQGGCGTLSSVHETSNNIND   204
   P. gaboni GARP    DEQNIYEKFINELTNNLESTLLLSSLYPYSEQGGCGIISTVHDVFYDIKE   250

                                              R2
P. falciparum GARP   NDKENISE-----DKKEDHQQEEMLKTLDKKERKQKEKEMKEQEKIEKKN   282
P. reichenowi GARP   NDKENKSE-----DKKDDPKQEEMLKTLDKKERKQKEKEMRKQEKIEKKN   249
   P. gaboni GARP    NDYENKDEENKDVDKKYEQKQEETLKTLDKKERRQKEKEMKEHQK---KQ   297

P. falciparum GARP   KKQEEKEKKKQEKERKKQEKKERKKQKEKEMKKQKKIEKERKKKEEKEKKK   332
P. reichenowi GARP   KKKEEKEKKKQEKERKKQEKKERKKQKEKEMRKQKIEKKKKKEEKEN-   297
   P. gaboni GARP    KEKEMKEHQKKQKERKKKEDKEKKKLEKEKKMKEVTEKKKQKKQT-----   342

                                                        R3
P. falciparum GARP   KKHDKENEETMQQPDQTSEETNNEIMVPLPSPLTDVTTPEEHKEGEHKEE   382
P. reichenowi GARP   KKHDKENEETMQQQEQTSEENNNENMVPLPSPVTDVSTPEEDKTEQDKTE   347
   P. gaboni GARP    -KQIEQNKEKIQQQQQTSEETNNEIMAPLPSPVTDVSTTQEHKTEEKKE-   390

                                                            R4
P. falciparum GARP   EHKEGEHKEGEHKEEEHKEEEHKKEEHKSKEHKSKGKKDKGKKDKGKHK-   431
P. reichenowi GARP   EDKTEQ-----DKTEQDKTEQDKTEQDKSKEKKQKEKRKKGKK----GK-   387
   P. gaboni GARP    EGK----------------------------------------HKI   396

P. falciparum GARP   -KAKKEKVKKHVVKNVIEDEDKDGVEIINLEDKEACEEQHITVESRPLSQ   480
P. reichenowi GARP   -KQKKEKQKKHVVKNVMGDEDKDDVEIINLEDKKACEEQHITVESRPLYQ   436
   P. gaboni GARP    DTKEKEKEETHILKNFIGTEHEDNIQIITLEDKKTCEQQHITVKNKPLYQ   446

P. falciparum GARP   PQCKLI-------------------------------DEPEQLT   493
P. reichenowi GARP   QQYKLM-------------------------------DEPEQLS   449
   P. gaboni GARP    SQQKIIQPKPADEAQHKPADEAQHKPTDVALHKPTDVALHKPTDAAQQLT   496

P. falciparum GARP   LMDKSKVEEKNLSIQEQLIGTIGRVNVVPRRDNHKKKMAKIEEAELQKQK   543
P. reichenowi GARP   SKDKSKVEEKHLSIQEQLIGTIGRVNVVPRRDNKKNKMTNKDETEVQEQK   499
   P. gaboni GARP    LTDKTDVKEKHLSIQEQLIGNIGKVNVGPKKDDKIKNITKKHEKKVQKHT   546

P. falciparum GARP   HVDKEE-------------------------------------   549
P. reichenowi GARP   DGDKEE-------------------------------------   505
   P. gaboni GARP    HVDKKKEQTEEHKHDDQKKEQTEEHKHDDKKKEQTEEHKHDDKKKEQTEE   596

P. falciparum GARP   ------------------------DKKEESKEVEEESKEV----QEDEEEVEE   574
P. reichenowi GARP   ------------------------DKKEESKEEDEEEEE----ENDEEEEEE   530
   P. gaboni GARP    HKHDDKKKEQTEEHKHDDKKEDKKEESEEDSDEESEEESEEGSDEESEEE   646

P. falciparum GARP   DEEEE---EEEEEEEEEEEE---------EEDEVEDEDDAEEDE   611
P. reichenowi GARP   EEENEE---EEENEEEEENEEE---------ENEEESEEE--------   559
   P. gaboni GARP    SEEGSEEESEEESEEGSEEESEGEDDDEDEEDEEDDEDEEDDDEDEEDD   696

P. falciparum GARP   DDAEEDE-DDAEEDDDDAEEDDDDAEEDDDEDEDEDEEEEDEEEEEESE   660
P. reichenowi GARP   --------------------EESEEESE-EENDDEESEEESEEEEESE   586
   P. gaboni GARP    EDEDDEEDDDDENDDDENDDDEDDDEEDDEDEEDEDDDEEEDEDNEESE   746

P. falciparum GARP   KKIKRNLRKNAKI   673
P. reichenowi GARP   KKIKRNLRKNAEI   599
   P. gaboni GARP    KNKKTNLRKN--A   757
```

Figure 3.6 – Alignment of the GARP protein sequence from *P. falciparum, P. reichenowi and P. gaboni.* Residues with 50% conservation or greater are shaded and the lysine-rich repeats (R1-R4) of *P. falciparum* GARP are boxed.

Once established, repetitive sequences often expand and contract dynamically, however it is uncommon for repeats to disappear completely; mutations in the sequence often fix sequence elements in place and expansion is generally more common than contraction in short tandem repeat sequences (Wheeler et al., 2007). As the only repetitive sequences common to all three closely-related *Plasmodium* species are the second lysine-rich sequence and acidic C-terminus, it is likely these were the only repetitive sequences present within the common ancestor of these species.

As our previous truncations of the first lysine-rich repeat sequence in *P. falciparum* suggests that five repetitions of the consensus motif is not sufficient for robust peripheral localisation, the homologous sequences within *P. reichenowi* and *P. gaboni* may not be sufficient to target the erythrocyte periphery. To test this, the N-terminus of the *P. reichenowi* GARP protein was GFP-tagged for expression in *P. falciparum* parasites; this construct contained the region of the first repetitive sequence only (**Fig 3.7C**). As predicted, the GFP-tagged *P. reichenowi* fragment $GARP_{71-130}$ did not localise to the periphery of the erythrocyte, but remained cytoplasmic with some fluorescence retained within the parasite food vacuole (**Fig 3.7C**). This was compared to the previously mentioned homologous fragment of *P. falciparum* $GARP_{119-163}$ which exhibits strong localisation at the periphery (**Fig 3.7B**). To confirm that *P. gaboni* GARP has not expanded a functional targeting domain at the homologous region to the third and fourth repeats in *P. falciparum* GARP, a sequence encompassing this region of both proteins was also GFP-tagged. As expected, the *P. gaboni* sequence remained in the erythrocyte cytosol (**Fig 3.7E**) while the *P. falciparum* sequence was robustly targeted to the periphery (**Fig 3.7D**).

While the acidic C-terminus and second lysine-rich repeat region of both *P. reichenowi* and *P. gaboni* GARP are largely intact, the loss of additional targeting sequences may affect protein function. In this case it is apparent that repeat expansion has led to the *de novo* formation of a targeting module in *P. falciparum*; such expansive changes to protein sequences on a relatively short timescale may provide a means of rapid evolution for the parasite.
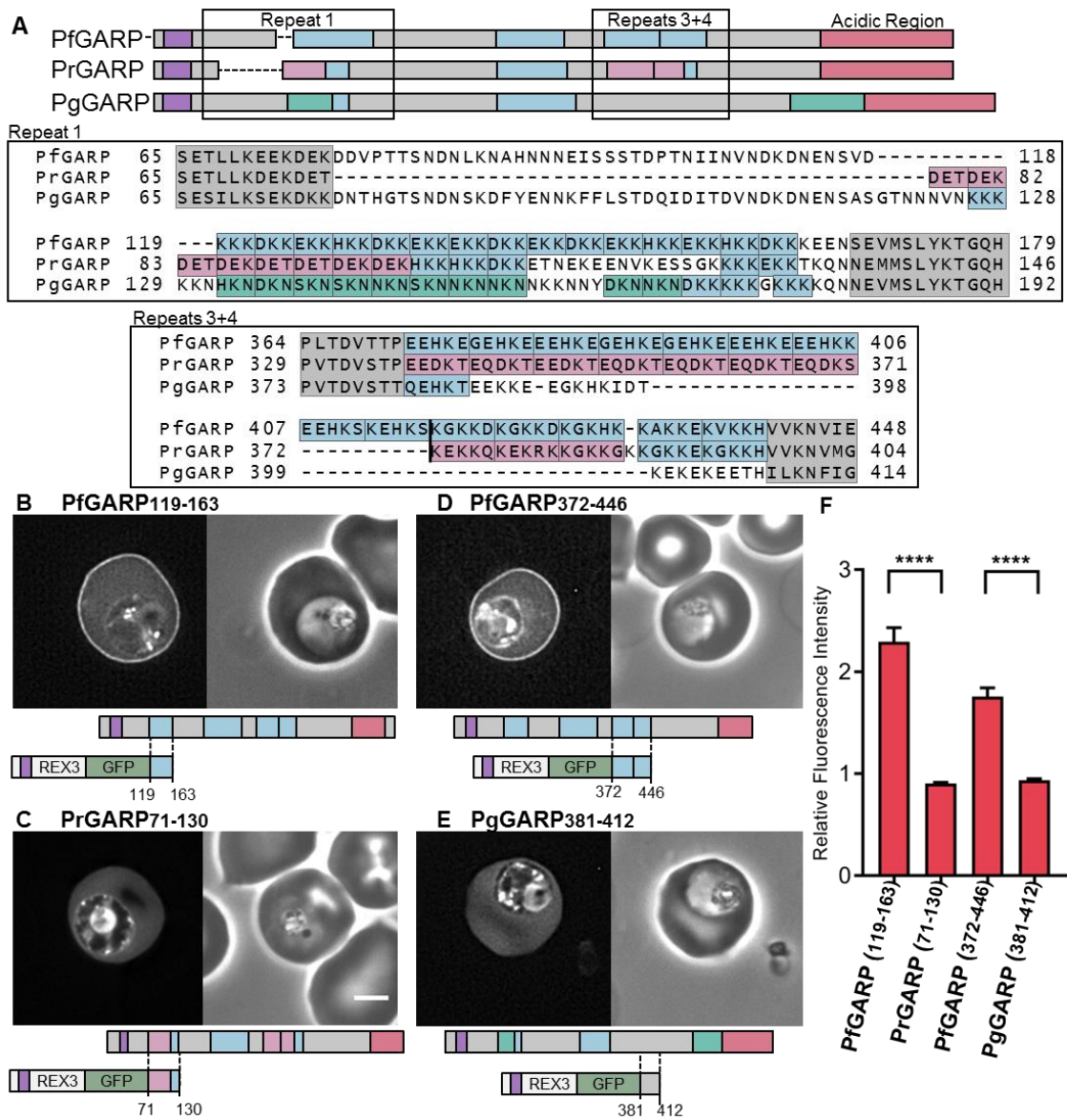
Figure 3.7 – Repeat expansion may generatate functional targeting sequences. (A) Representation of GARP from *P. falciparum* (upper), *P. reichenowi* (middle), and *P. gaboni* (lower) with the export motif in purple and the acidic C-terminus in red. Lysine-rich repeats matching those of *P. falciparum* are in blue. Alternative repeated motifs from *P. reichenowi* and *P. gaboni* are shown in pink and turquoise, respectively. Alignment of two lysine-rich repeat regions of PfGARP and its homologous regions in PrGARP and PgGARP are shown below. Flanking conserved sequences are in grey. (B and D) GFP-tagged PfGARP, (C) PrGARP, and (D) PgGARP fragments (schematics of cloned region below and fluorescence and phase contrast images to left and right). Scale bar – 2μm. (F) Comparison of peripheral fluorescence with statistical analysis (**** - extremely significant). Error bars represent SEM.

3.3.5 Western Blotting confirms the expression of the correctly-sized protein fragments

To confirm the expression of the correct GFP-tagged GARP fragments, erythrocytes infected with schizont-stage parasites were purified and analysed by western blotting (**Fig 3.8**). Anti-GFP was used to probe protein fragments, with anti-histoaspartic protease used as a loading control as this protein is highly expressed throughout the parasite life cycle. Bands at approximately the correct sizes were observed for most parasite lines, with degradation bands visible for several GARP constructs, possibly corresponding to protease cleavage sites within the protein. GFP-only bands at around 25 KDa are also commonly seen for GFP-tagged constructs. Many proteins also ran higher than expected, and the full-length protein often appeared as a blurred band much higher than the expected size of 100 KDa. This size discrepancy is common in repetitive or highly charged proteins, while blurry bands often correspond to post-translational modifications. While modifications to the protein cannot be dismissed, the unusual sequence composition of GARP likely affects its behaviour during SDS-PAGE electrophoresis.
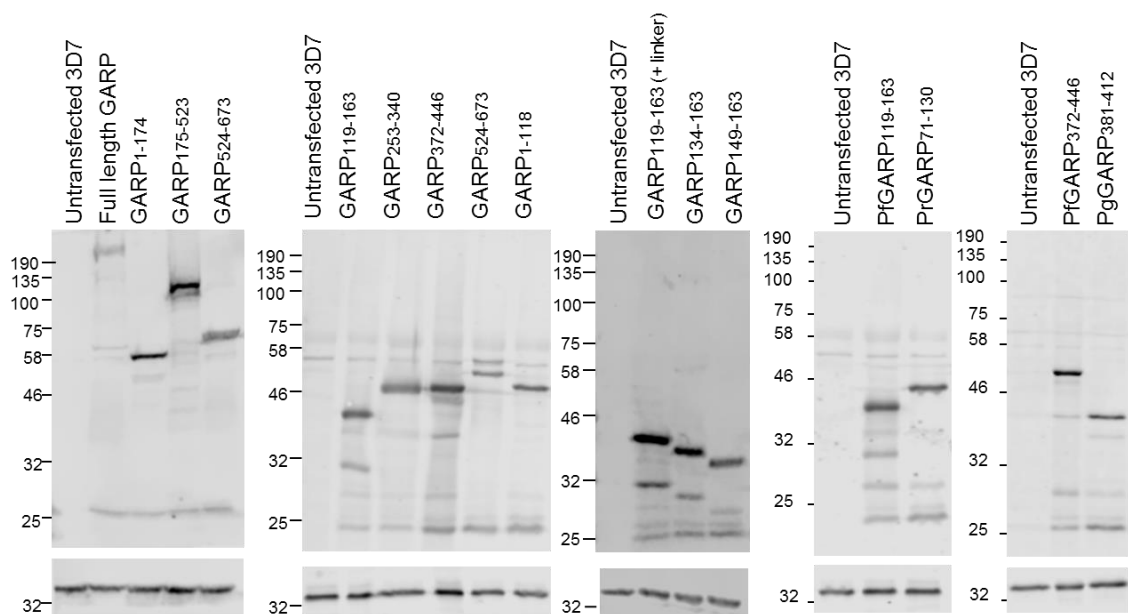
Figure 3.8 – Western blots of GFP-tagged GARP constructs. Anti-GFP western blot (top panel) of transfected parasites expressing GFP-tagged GARP constructs. Approximately 2 x 10^6 percoll-purified schizonts were loaded per lane, with anti-HAP used to confirm equal loading (lower panel).

3.3.6 Solubility Extraction Suggests GARP is bound to the Cytoskeleton

Various components of the erythrocyte periphery may provide a binding surface for the lysine-rich sequences of GARP. The protein may interact with lipids in the erythrocyte membrane or it may bind to components of the unique erythrocyte cytoskeleton or other *Plasmodium* proteins localised to this region. The cytoskeleton is composed of a network of spectrin filaments linked to the membrane by both the Ankyrin and junctional complexes (Mankelow et al., 2012). The cytoskeletal framework is known to be detergent insoluble, along with any proteins which are bound to its core components (Black et al., 2008). This was utilised to establish whether GARP interacts with the cytoskeleton, which is known to be extensively modified by the parasite.

Solubility extraction experiments were performed with erythrocytes infected with $GARP_{1-174}$ as this fragment was easily visualised by western blot compared to the full-length protein. Components of red cells were sequentially extracted according to their solubility in different buffers (**Fig 3.9A**). Hypotonic hemolysis of the infected cells with a low-salt buffer was first performed in order to release the soluble components of the erythrocyte cytoplasm to create red blood cell 'ghosts', where membrane and cytoskeleton components remain within the cell pellet. Following this, a 1% Triton detergent solution was used to solubilise the lipids and proteins within the erythrocyte membrane. Cytoskeletal transmembrane proteins such as Band 3 are also partially removed by this process but most cytoskeleton-associated proteins remain in the pellet fraction. Anti-spectrin and anti-band 3 antibodies were used to probe a western blot of the treated cells to confirm the correct extraction of proteins.

Bands corresponding to GFP-tagged $GARP_{1-174}$ were observed by western blot in both the pellet and supernatant of hypotonically lysed cells (**Fig 3.9B**); this is not surprising as some cytosolic fluorescence was also observed in erythrocytes infected with parasites expressing this fragment. Subsequent treatment with triton solution consistently yielded a similar distribution of protein in both insoluble and soluble fractions in three independent experiments, however an additional wash with triton solution was not able to extract any more of the GARP protein and a high proportion remained in the insoluble pellet (**Fig 3.9B**). This suggests that at least a fraction of the GFP-tagged $GARP_{1-174}$

protein is interacting with the erythrocyte cytoskeleton. The extraction of roughly half the protein during both hypotonic lysis and triton extraction may indicate that a proportion of the protein is only weakly associated with the cytoskeletal framework. However, while spectrin remained in the triton insoluble pellet as expected, Band 3 was only partially extracted; this may indicate incomplete solubilisation of proteins under the conditions used.
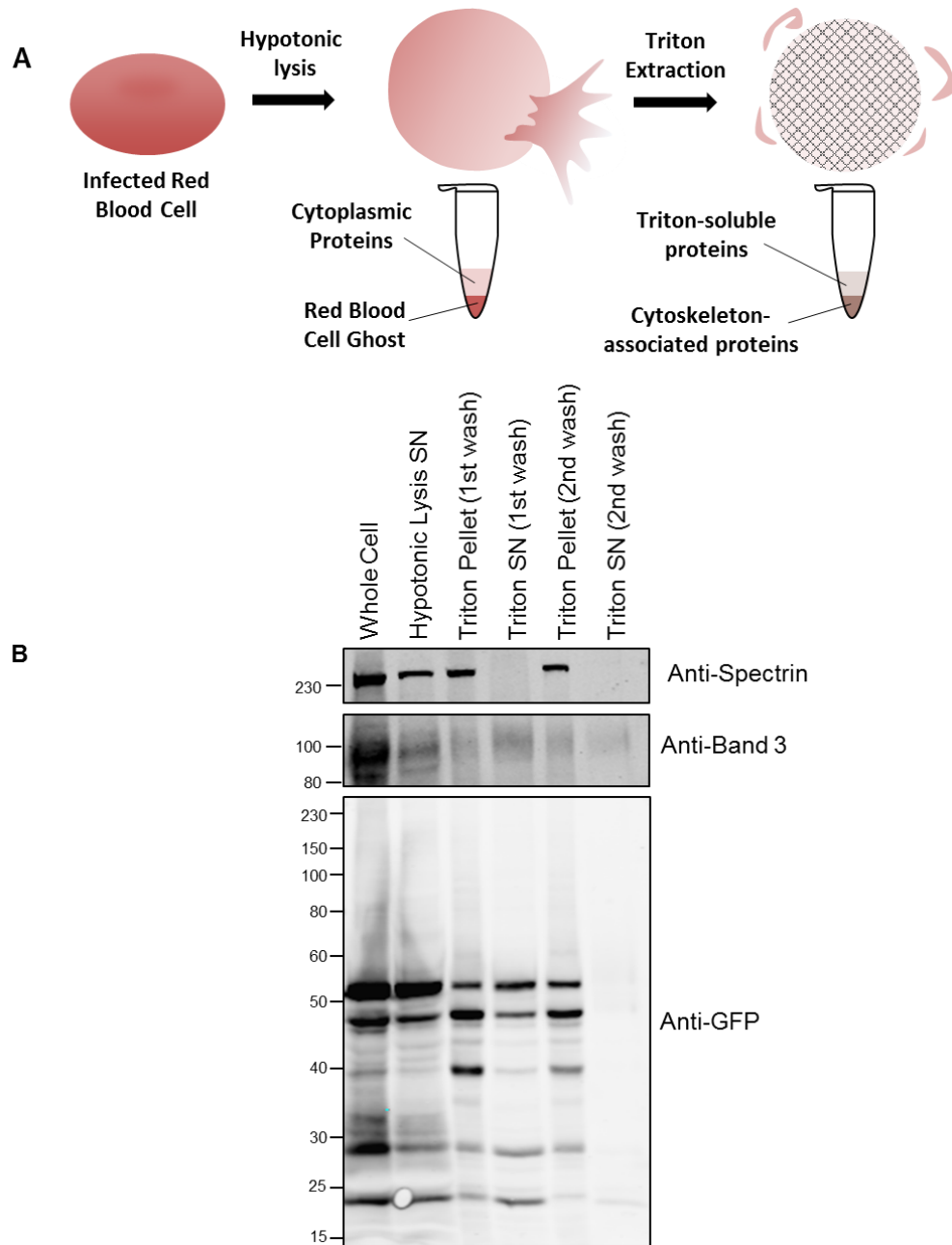
Figure 3.9 –Triton solubility extraction of red blood cells expressing the first GARP repeat sequence. (A) Depiction of erythrocyte components extracted at each step. Cells were first hypotonically lysed followed by extraction in 1% triton detergent. (B) Western blot with anti-spectrin, anti-Band 3 and anti-GFP used to probe the localisation of proteins at each step.

Smaller bands were also observed in extracted fractions for GFP-tagged GARP$_{1-174}$; these are likely to be degradation products of the protein which were observed to a lesser extent previously. Interestingly, while the full-length protein fragment is evenly distributed between triton soluble and insoluble fractions, a fragment of around 40 KDa is predominantly insoluble. As the lysine-rich C-terminal end of the protein fragment is GFP-tagged and the blot is probed with anti-GFP antibodies, this fragment likely contains the lysine-rich repeats of the protein. This may indicate an association between this sequence and the cytoskeleton, although it may also be a result of higher rates of proteolysis within the triton pellet.

As the behaviour of GARP$_{1-174}$ was similar to that of Band 3 when extracted with triton, a sodium carbonate extraction was also performed to extract all proteins which are not integral to membrane bilayers (Fujiki et al., 1982). Spectrin is therefore solubilised by sodium carbonate whereas Band 3 is not (**Fig 3.10A**). The lack of hydrophobic regions within GARP makes its incorporation into membranes highly unlikely, however when cells expressing the GARP$_{119-163}$ fragment were extracted with sodium carbonate, a large proportion of the protein remained in the insoluble fragment (**Fig 3.10B**). This may be explained by incomplete solubilisation of proteins contained within the parasite; incomplete extraction of GARP present in the secretory pathway or food vacuole of the parasite may account for this band. To support this, the extraction of the parasite-localised soluble protein histo-aspartic protease (HAP) was simultaneously probed using anti-HAP (Liu et al., 2005); this protein was also distributed between both soluble and insoluble fractions during both hypotonic lysis and carbonate extraction. Unlike GARP, HAP was also fully extracted by triton solution. It is therefore likely that GARP is not incorporated into the membrane and likely associates with the cytoskeleton, however the results are difficult to interpret due to the unusual sequence composition of GARP, which may affect its solubility in different solutions.
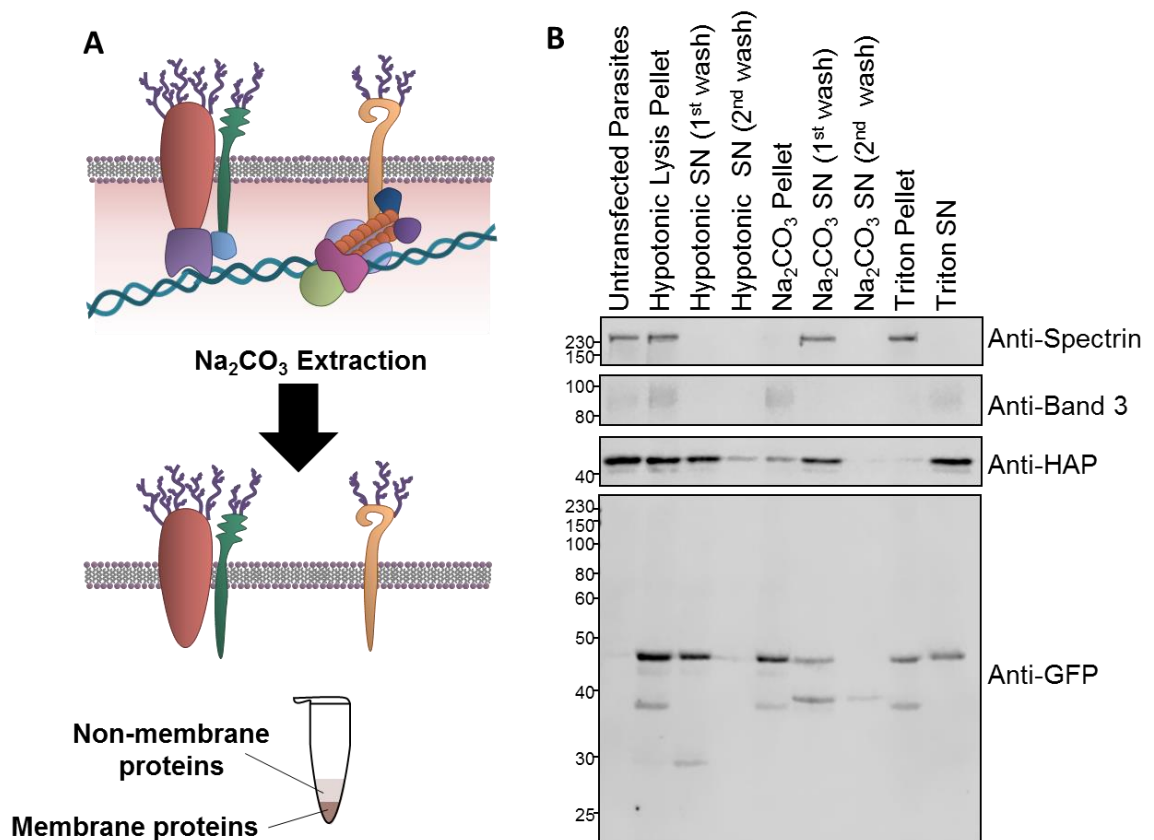
Fig 3.10 – Na$_2$CO$_3$ extraction of proteins from parasites expressing the first GARP repeat sequence. (A) Integral membrane proteins are insoluble in Na$_2$CO$_3$. (B) Anti-spectrin, Anti-Band 3, Anti-HAP and Anti-GFP were used to probe the localisation of proteins at each step.

3.3.7 Lipid Kinase Inhibitors do not affect the Localisation of GARP

While the previous experiments suggested that GARP likely interacts with the erythrocyte cytoskeleton, the basic nature of the protein suggests it may have a natural affinity for negatively charged head groups on phospholipids, for example. Interestingly, lipid rafts enriched in PIP$_2$ have been shown to form in the erythrocyte membrane upon *P. falciparum* infection (Murphy et al., 2007). To test whether GARP associates with these rafts, we added two known lipid kinase inhibitors to the cells to disrupt the balance of phosphatidylinositol phosphorylation. Wortmannin inhibits phosphoinositide 3-kinases (PI3Ks) which convert PIP$_2$ into PIP$_3$, while phenylarsine-oxide (PAO) inhibits phosphatidylinositol 4-kinases (PI4Ks) which are required for the formation of PIP$_2$.

Both compounds were dissolved in DMSO and added to parasite cultures at final concentrations of 10 μM and 2 μM for wortmannin and PAO, respectively. DMSO alone was added to control cells. Images of parasites expressing the full-length GARP GFP-tagged construct were taken 5, 30 and 80 minutes following the addition of each compound. No effect on protein targeting was seen in this time for either inhibitor, suggesting that GARP does not associate with phosphatidylinositol lipids (**Fig 3.11)**. After 24 hours the PAO-treated parasites were almost completely dead, suggesting that PIP_2 formation may be essential. Inhibition of PI4Ks with other compounds has previously been shown to result in the death of the parasite (McNamara et al., 2013); however PAO is also a known tyrosine phosphatase inhibitor and this and other off-target effects may be responsible for its anti-malarial activity. Parasites treated with wortmannin were unaffected after 24 hours, although this is likely due to the rapid degeneration of the drug in aqueous conditions. Further experiments are necessary to validate the effect of these inhibitors.
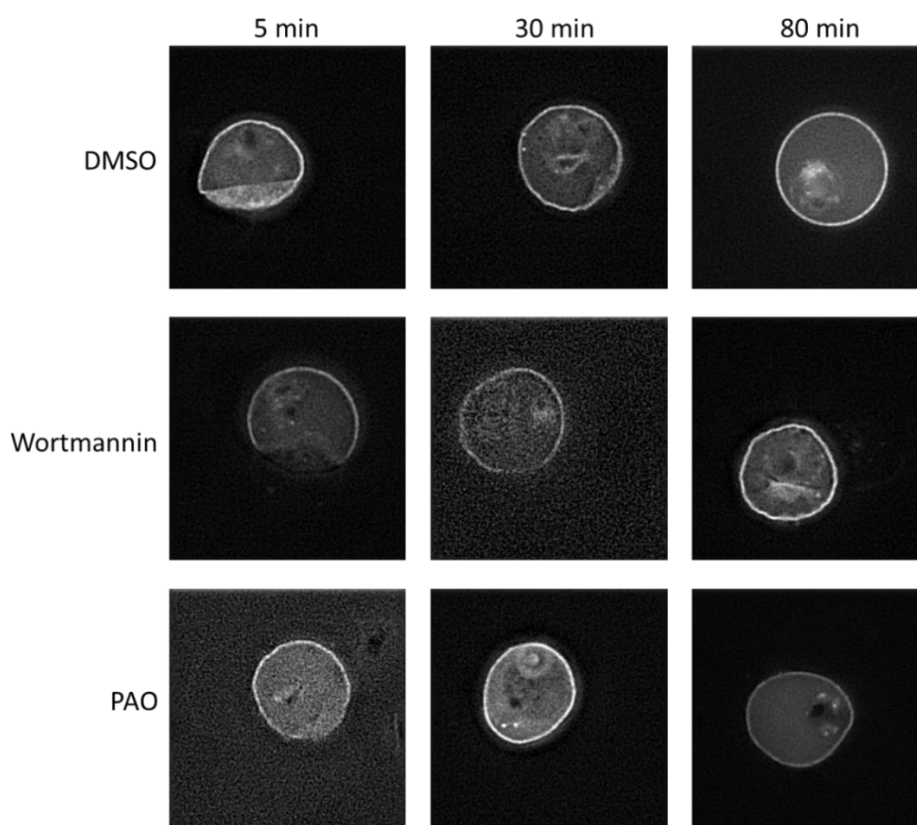
Figure 3.11 – Lipid kinase inhibitors do not affect peripheral targeting. Parasites expressing GFP-GARP were treated with a DMSO control, wortmannin, or PAO. They were imaged after 5, 30 and 80 minutes to observe any effect on localisation.

3.4 Discussion

As modifications to the erythrocyte cytoskeleton and membrane are essential for parasite survival within the host, the identification of a novel protein localising to this region provides a new avenue of investigation for explaining such changes. GARP is a particularly unusual protein; such highly charged and repetitive proteins have previously been dismissed as nonsense protein sequences with no function (Muralidharan et al., 2011). The independent localisation of three lysine-rich repeating sequences to the erythrocyte periphery indicates this is not the case, and that expansion of repeats in *Plasmodium* may provide a means for creating novel targeting modules.

The idea that repeat expansion may contribute to protein evolution has been suggested previously (Muralidharan and Goldberg, 2013, Moxon et al., 1994, Andrade et al., 2001, Levinson and Gutman, 1987, Verstrepen et al., 2005, DePristo et al., 2006, Gemayel et al., 2010, Gemayel et al., 2012, Toll-Riera et al., 2012). Single amino acid repeats are highly mutable and their expansion may create space for a novel domain to emerge (Muralidharan and Goldberg, 2013). We provide direct evidence that expansion of a disordered and highly conserved repetitive sequence can lead to a gain of function by comparing closely related parasite species. While the *P. reichenowi* GARP sequence has expanded a non-functional acidic repetitive region at its N-terminus, *P. falciparum* has expanded a lysine-rich functional targeting sequence. In *P. gaboni*, a peripheral-targeting lysine-rich repeat is missing entirely. These divergent expansions may be in response to selective pressure within the chimpanzee or human hosts. Alternatively, repeats may expand randomly with non-functional repeats retained as neutral spacer sequences.

The peripheral-targeting function of the lysine-rich repetitive sequences in GARP suggests they may contribute to rigidification of the cytoskeleton or the anchoring of cytoadhesive components to the cell membrane. Both these modifications are essential for the parasite to avoid the host immune system by sequestering infected erythrocytes in the microvasculature (Craig et al., 2012b). As the majority of GARP is composed of charged repetitive regions, its function is likely to be directly conferred by these sequences and their adhesion to peripheral components. Solubility extraction of the first lysine-rich repeat of GARP did not conclusively confirm an interaction with the erythrocyte

cytoskeleton; the highly charged nature of the protein likely affects its solubility in detergents and makes interpretation of these results difficult. However, as a large proportion of the protein remained in the insoluble fraction after multiple washes with triton buffer, some association with the triton-insoluble cytoskeleton is probable. This is supported by the fact that lipid kinase inhibitors do not affect localisation, making an interaction with lipid rafts unlikely, although it cannot be ruled out completely based on these preliminary experiments.

Multiple sequences with a targeting function were found within GARP, suggesting the protein may crosslink various components of the erythrocyte cytoskeleton. This may be involved in stabilising protein complexes or preventing spectrin filaments from deforming, for example. This may explain the slight decrease in erythrocyte rigidity observed on deletion of the GARP gene (Maier et al., 2008). Interestingly, several degradation bands are present in Western blots of GARP, suggesting the protein may be cleaved by proteases. This cleavage may be a necessary step towards allowing merozoites to egress from the cell, by dismantling the rigid cytoskeleton. Alternatively, the protein may be degraded in the food vacuole or during cell lysis. The blurred band observed for the protein may also indicate heterogeneous post-translational modification to the protein resulting in a broader size range.

The three lysine-rich sequences within GARP are composed of different repeated motifs. Repetitive sequences are present in more than 30% of the proteins encoded by the parasite (Mendes et al., 2013), and many are enriched in lysine residues due to the AT-rich nature of the parasite genome (DePristo et al., 2006). Similar sequences to those within GARP are therefore common within the parasite, and may confer the same targeting function. As proteins localised to the periphery of the erythrocyte play important roles in host-cell modification, the discovery of a novel targeting domain provides a significant step towards identifying other proteins localising to this region.

Chapter 4: Multiple Lysine-rich repeating Proteins with a Targeting Function

4.1 Abstract

In addition to the glutamic acid rich protein (GARP), several other proteins encoded by *P. falciparum* contain lysine-rich repetitive sequences. To establish whether these also form functional targeting sequences, an algorithm was written to identify proteins with a repetitive sequence containing over 20% lysine residues. Eleven proteins were selected to be GFP-tagged, including two which fell below the assigned criteria of the algorithm. Of these, eight repetitive sequences were localised at the erythrocyte periphery. Additionally, the Hyp12 protein contained a lysine-rich sequence capable of targeting the periphery in isolation but targeting was inhibited by the presence of an acidic sequence in the full-length protein. Four proteins containing functional lysine-rich sequences were previously uncharacterised, while others were taken from proteins containing other domains capable of targeting the erythrocyte periphery or binding to cytoskeletal components (including KAHRP, MESA and LYMP) One novel protein was only weakly expressed in asexual stages under its endogenous promoter but was upregulated in gametocytes, indicating a role in modifying erythrocytes during the sexual stage of the life cycle. Analysis of long-read sequence data of parasite field-isolates and lab strains indicates the length of some functional lysine-rich repeat sequences are highly variable, suggesting that the parasite may modulate the targeting efficiency of proteins by expanding and contracting repetitive sequences.

4.2 Introduction

The protein export predictor EXPORTPRED estimates that 458 proteins encoded by the malaria parasites are exported into the host cell, based on the presence of a signal sequence and a host targeting (HT/PEXEL) motif RxLxE/D/Q (Sargeant et al., 2006). It has since become apparent that an N-terminal transmembrane domain may be utilised instead of a signal sequence to direct proteins to the ER, and that some mutations to the fifth position of the HT/PEXEL motif still allow cleavage by plasmepsin V and export into the host cell (Tarr et al., 2013, Boddey et al., 2013, Schulze et al., 2015). Lysine may

also replace arginine at the first position of the motif, although this has not been investigated in detail and it is not clear whether other residues at the fifth position are tolerated in this case (Schulze et al., 2015). Additionally, a significant number of PEXEL-negative exported proteins (PNEPs) have been identified (Heiber et al., 2013). It is therefore likely that the number of proteins involved in modifying the host cell is greater than first imagined. Within the current 'exportome' of the parasite, many proteins are characterised as 'conserved proteins of unknown function', whose protein sequence bears little homology to proteins from other species and which do not contain any known functional domains. It is likely given the presence of an export element that many of these proteins play an important role within the host cell. The identification of several peripheral targeting sequences within the previously uncharacterised protein GARP suggests that many other peripherally-localised proteins may remain undiscovered, and that other protein domains may also be capable of targeting this important area of the cell.

The screen of uncharacterised exported proteins which identified GARP also uncovered another peripheral targeting domain in the PHISTb family of proteins (Tarr et al., 2014). Protein fragments encompassing the *Plasmodium* RESA N-terminal (PRESAN) domain of several of these proteins were found to independently localise to the erythrocyte periphery when exported by *P. falciparum* parasites (Tarr et al., 2014). The PRESAN domain of one PHISTb protein, named the lysine-rich membrane-associated PHISTb protein (LYMP), has been shown to associate with the ATS domain of PfEMP1 (Oberli et al., 2016, Oberli et al., 2014), however it is not known whether the other proteins within this family also adhere to variant surface antigens. The C-terminus of LYMP is capable of binding to the cytoskeletal component Band 3 (Oberli et al., 2016), possibly anchoring the surface antigen to the erythrocyte cytoskeleton. Deletion of LYMP reduces cytoadhesion of infected cells. The Band 3-binding C-terminus of the protein contains a 100 residue stretch followed by a lysine-rich repeat sequence; however it is not clear whether the lysine-rich repeats themselves are responsible for this interaction.

A peripheral-targeting module is also found in the mature-erythrocyte infecting surface antigen (MESA) (Bennett et al., 1997, Black et al., 2008). Contrary to its name it is not presented on the erythrocyte surface but instead binds to protein

4.1, a component of the junctional complex connecting the membrane to the cytoskeleton (Lustigman et al., 1990). While no function has yet been found for MESA, it has been characterised in detail and the protein 4.1 binding site was found to be a 20-residue sequence near the N-terminus of the protein, named the MESA erythrocyte cytoskeleton-binding (MEC) motif (Bennett et al., 1997). A search for other proteins containing this binding motif indicated that as many as 13 proteins contain similar sequences, with several shown to be capable of binding 4.1 *in vitro*, as well as adhering to inside out erythrocyte vesicles (Kilili and LaCount, 2011). While binding was weak in many cases, it is possible that many of these proteins also localise to the erythrocyte periphery *in vivo*. A yeast two hybrid screen also identified a PHISTa protein (PF3D7_0402000) capable of binding protein 4.1 *via* its PRESAN domain (Parish et al., 2013).
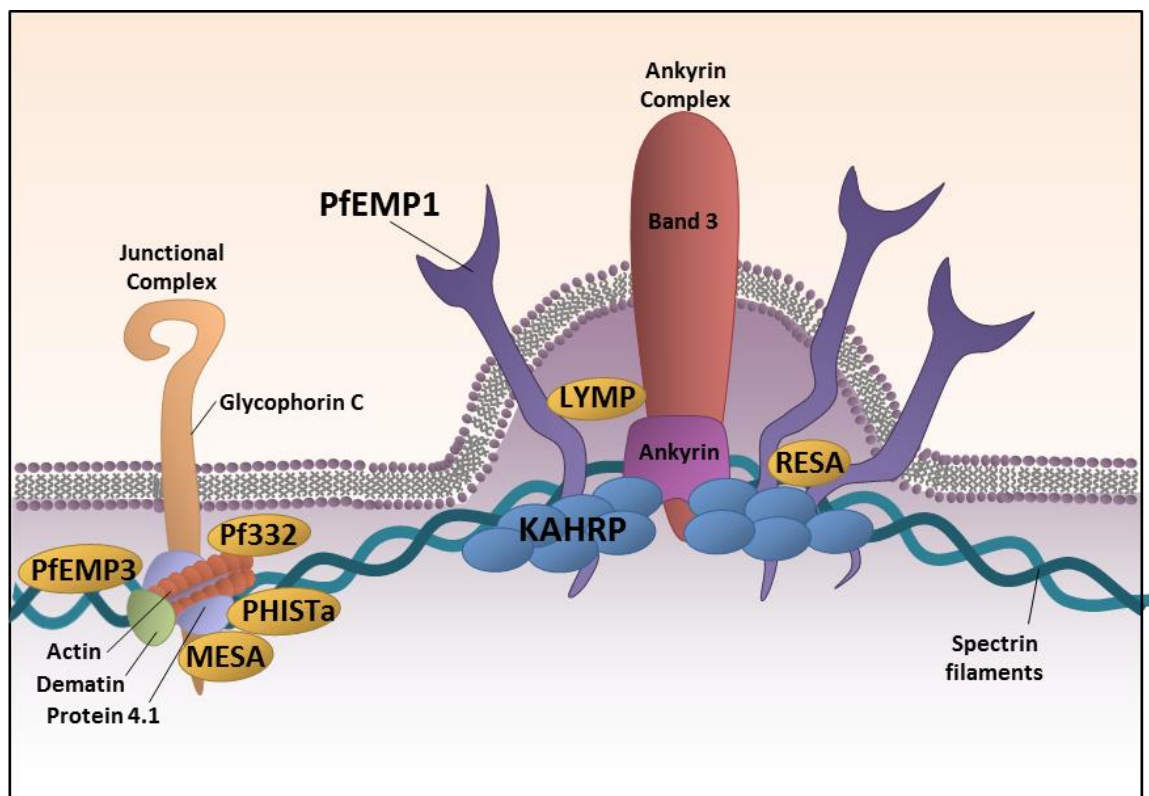
Figure 4.1 – The erythrocyte cytoskeleton and membrane is modified by parasite proteins. Depiction of knob structures within the erythrocyte membrane, with the *Plasmodium* protein KAHRP in blue, PfEMP1 in purple, and cytoskeleton binding proteins LYMP, MESA, PfEMP3, Pf332, RESA and PHISTa protein PF3D7_0402000 in yellow.

Like LYMP, both MESA and PF3D7_0402000 contain repetitive and highly charged sequences; MESA is almost entirely composed of repeats of varying charge, while the PHISTa protein PF3D7_0402000 contains a repetitive lysine-rich C-terminus. Other proteins previously found to localise to the periphery also contain highly charged repetitive sequences. For example, the ring-stage surface antigen (RESA) and Pf332 both contain acidic repeats rich in di-glutamate motifs (Cowman et al., 1984); while the majority of the PfEMP3 protein is composed of glutamate and lysine-rich repeats. As with MESA and PF3D7_0402000, the cytoskeleton targeting motif for each of these was found in non-repetitive and uncharged segments (Knuepfer et al., 2005, Pei et al., 2007b, Waller et al., 2007b, Waller et al., 2010, Bennett et al., 1997, Parish et al., 2013, Foley et al., 1991); PfEMP3, RESA, and Pf332 bind to α-spectrin and actin, β-spectrin, and actin, respectively (Foley et al., 1991, Pei et al., 2007b, Pei et al., 2007a, Waller et al., 2007a, Waller et al., 2010) and all three proteins are involved in modulating the rigidity of the cytoskeleton (Glenister et al., 2009, Hodder et al., 2009, Glenister et al., 2002, Pei et al., 2007a, Pei et al., 2007b, Mills et al., 2007). A representation of the erythrocyte cytoskeleton including the *P. falciparum* proteins binding to its components is shown in **Fig 4.1**.

Due to the enrichment of lysine residues in the parasite genome, many lysine-rich repetitive sequences are found in exported proteins which have no known function. The discovery that three different lysine-rich repeats within GARP are capable of targeting the periphery suggests that this may be a conserved phenomenon. We sought to investigate similar sequences in previously uncharacterised proteins as well as those with known roles at the erythrocyte periphery (**Fig 4.1**).

4.3 Results

4.3.1 Prediction of Exported Proteins by a new Algorithm

To investigate whether sequences similar to those in GARP are capable of targeting to the erythrocyte periphery, we searched for putative exported proteins containing lysine-rich repetitive sequences. Exported proteins contain a HT/PEXEL motif with the consensus sequence RxLxE/D/Q; however it has been shown that other residues may occur at the fifth position of the motif. As the available programs for predicting protein export do not account for this variation in the fifth position, a new algorithm was written for this purpose. All 5777 *P. falciparum* proteins were downloaded from the parasite database, PlasmoDB (version 26) in fasta format. As entry into the ER may be mediated by a signal sequence or an N-terminal transmembrane domain, all proteins were first analysed by the signal peptide predictor SignalP (Petersen et al., 2011) and the transmembrane domain identification software MPEX translocon TM analysis (Snider et al., 2009). Of all programs tested, MPEX proved to be the most successful in predicting N-terminal transmembrane domains in proteins known to be exported into the host cell.

A perl script (Exportfind.pl – see **Fig 4.2** for data flow chart) was written to analyse the output of these programs, and proteins which contained either a signal sequence or a transmembrane domain within the first 50 residues were selected. The sequences of these proteins were then searched using regular expression for RxL motifs (where x is any residue) within 100 residues of the signal peptide cleavage site or the end of the transmembrane domain. Proteins conforming to the above criteria were considered to be likely candidates for export, and were output in fasta format. Proteins with more than 4 transmembrane domains were later removed from this exported protein subset, as they are unlikely to be targeted to the periphery as soluble proteins, and are therefore not of interest.

Using our 'Exportfind' algorithm, 573 proteins are predicted to be exported. The 'Exportpred' prediction software built into the PlasmoDB website predicts that 458 proteins are exported (When Exportpred score threshold = 1) (Sargeant et al., 2006). Of these, 376 were also predicted by our algorithm (**Fig 4.3**). Proteins which were not predicted to be exported by our algorithm were mainly

from the VAR gene family (56 proteins) which are PEXEL-negative exported proteins (PNEPs) (Kriek et al., 2003). Other proteins predicted by 'exportpred' but not 'exportfind' include those containing a KxLxE/D/Q host-targeting motif, which has also been shown to be a functional export signal in some instances (Schulze et al., 2015). Considering KxLxEDQ motifs in our algorithm resulted in 728 proteins predicted to be exported, including several proteins which are likely to function within the parasite, such as Acyl-CoA synthase (ACS6); the smaller RxL dataset was therefore used for detecting functional lysine-rich repeats to reduce false positives. Some additional proteins predicted by our algorithm are known to be exported, suggesting that it is less likely to miss potential candidates, although verification of export in some selected proteins will be required.
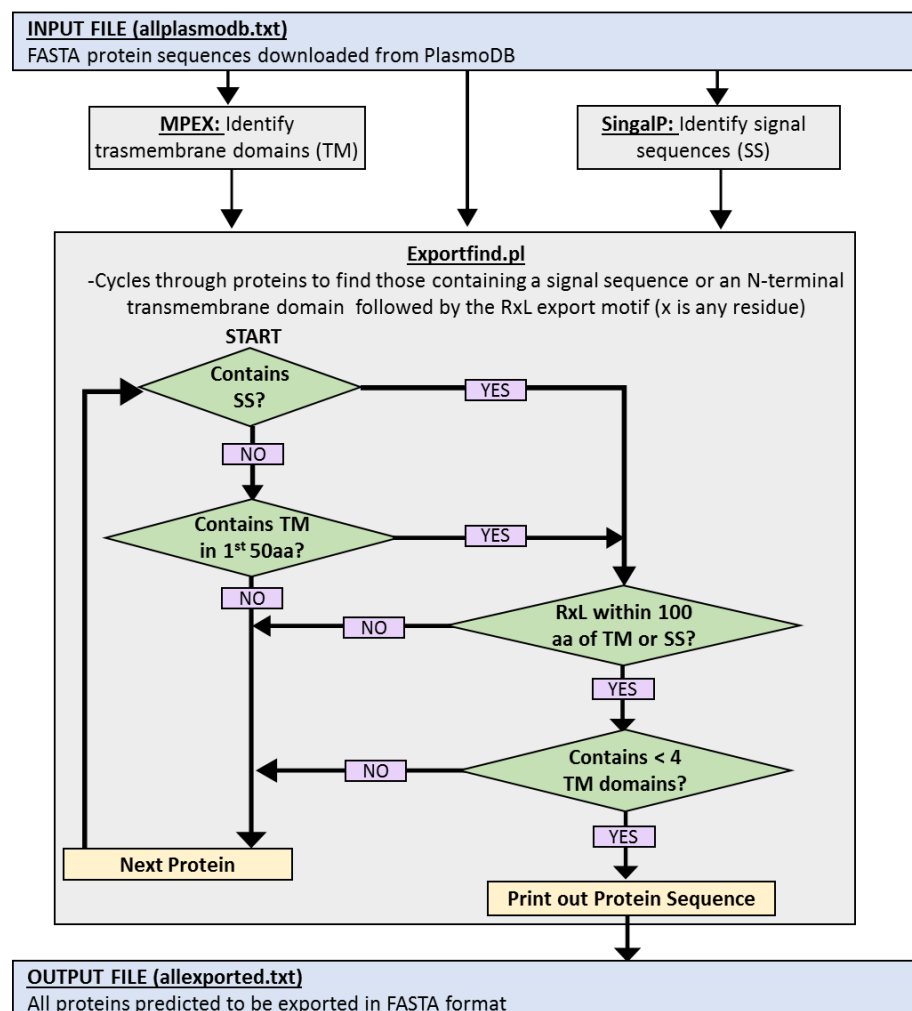
Figure 4.2 - PERL script for the prediction of proteins exported from the parasite into the host cell. Programs are shaded grey and INPUT/OUTPUT text files in blue. Within the algorithm, green diamonds represent a YES/NO decision and yellow boxes represent processes.
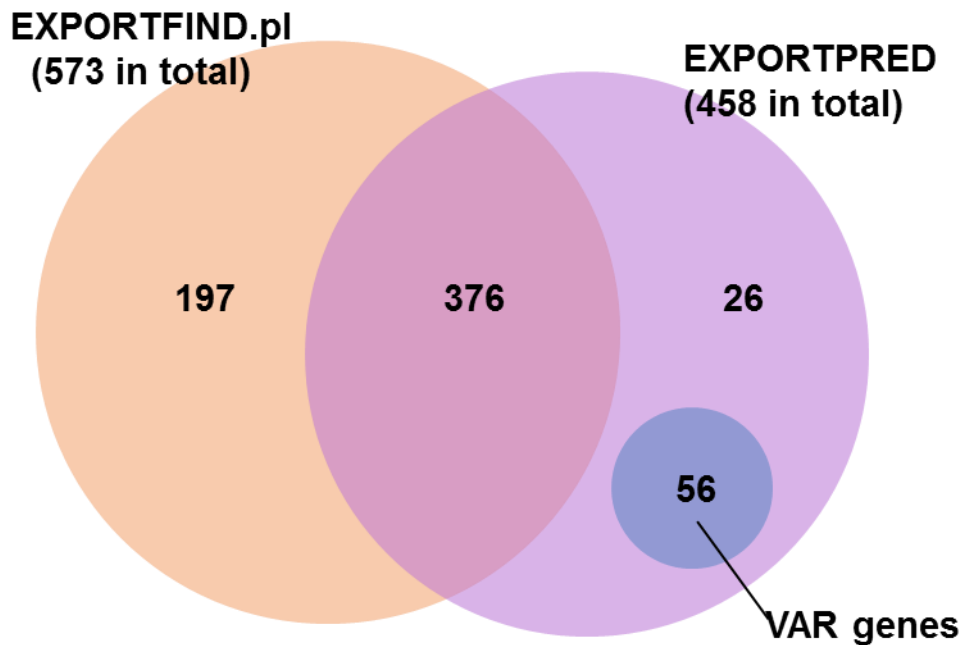
Figure 4.3 – Venn diagram comparing proteins predicted to be exported by the parasite into the erythrocyte by two algorithms. Exported proteins predicted by 'exportfind' are represented by the orange circle and those predicted by established program 'exportpred' are in purple. Of 82 proteins predicted by exportpred and not exportfind, 56 are VAR genes, which are not exported by the canonical export pathway.

4.3.2 Many Exported Proteins Contain Lysine-rich Repetitive Sequences

Lysine-rich sequences were identified in exported protein sequences using a sliding window algorithm written for this purpose (Chargesearch.pl – see **Fig 4.4**). The proteins were scanned for sequences where 4 or more lysine residues were present in a 20 residue window. If 30 or more consecutive windows reached this threshold, the corresponding protein sequence was considered lysine-rich and output in fasta format. Multiple lysine-rich sequences may therefore be found in each protein. A range of window lengths and lysine thresholds were tested; 20% lysine in a 20 residue window gave the largest number of protein sequences which were considered likely candidates by comparison to the GARP sequences. Altogether, 242 lysine-rich sequences were identified by 'Chargesearch'.

INPUT FILE (allexported.txt)
FASTA sequences of all *P. falciparum* proteins predicted to be exported

CHARGESEARCH.pl
- Sliding window algorithm which searches for protein sequences over 30 aa in length containing over 20% lysine residues
Variables
Window Length = 20 aa
Residue to Count = K
Residue Count Threshold = 4 aa
Sequence Length Threshold = 30 aa
(K-Seq – Sequence in which each 20 aa window contains 4 K residues)

START
Next Protein

Window Contains > 4 lysine?

NO

K-Seq > 30 aa?

Next Window

YES

NO

YES

Add 1 aa to K-Seq

NO

Print out K-Seq

End of Protein?

YES

K-Seq > 30 aa?

YES

NO

Print out K-Seq

End of file?

YES

OUTPUT FILE (krichseq.txt)
FASTA file of all sequences with > 20% K and > 30 aa length, e.g.
```
>GARP1
 DNENSVDKKKDKKEKKHKKDKKEKKEKKDKKEKKDKKEKKHKKEKKHKKDKKKEENSEVMSLY
>GARP2
QEEMLKTLDKKERKQKEKEMKEQEKIEKKKKQEEKEKKKQEKERKKQEKKERKQKEKEMKKQKKIEKERKKKEEKEKKKKKHDKENEET
>GARP3
EEEHKEGEHKEGEHKEEEHKEEEHKKEEHKSKEHKSKGKKDKGKKDKGKHKKAKKEKVKKHVVKNVIED
>LYMP1
NTKVRGENKKVRGANKKVRGANKKVGSKNTKVGGENKKAGSSNKKAGNEKEKAGTENKKAGTENKKAGTENKKAGTENTKDK
...
```

Figure 4.4 - PERL script for the identification of lysine-rich sequences. The programs MPEX and SignalP were used to predict transmembrane domains (TM) and signal sequences (SS), respectively. Programs are shaded grey, and INPUT/OUTPUT text files in blue. Within the CHARGESEARCH.pl algorithm, green diamonds represent a YES/NO decision and yellow boxes represent processes.

As the peripheral targeting sequences of GARP were all repetitive, the program T-REKS was used to find tandem repeats within lysine-rich protein sequences (Jorda and Kajava, 2009). The program assigns a PSIM score to each repetitive sequence by considering the variation from the consensus motif, which represents the most common residue found at each position within the repeated sequence. A PSIM cut-off of 0.65 was used, with some lysine-rich sequences containing more than one different repeated motif. A perl script (pickrepeats.pl – see **Fig 4.5** for data flow) was written to analyse the output of T-REKS and to select the longest repetitive sequence for each lysine rich repeat; repetitive sequences longer than 30 residues were selected as possible peripheral-targeting protein sequences. The program retrieved the sequence of the lysine-rich input fragment and the full length protein for each repetitive sequence. An alternative repeat prediction program, X-STREAM, was used to assign the consensus sequence, the motif length, the position of the repeat within the protein, and the error from consensus (analogous to psim) for each protein (Newman and Cooper, 2007).

Thirty sequences, including those within GARP, were found to conform to the above criteria, with some proteins containing multiple repeating lysine-rich sequences resulting in 21 proteins altogether (**Table 4.1**). Additionally, exported proteins which were predicted by 'Exportpred' but not by our algorithm were analyzed, resulting in the addition of PHISTc protein PF3D7_0936800 to the list, which contains a KxLxE/D/Q HT/PEXEL motif (Schulze et al., 2015). A search of all known PNEPs did not find any proteins with repetitive sequences containing over 20% lysine. When all *P. falciparum* proteins were used as input for the 'chargesearch.pl' and 'pickrepeats.pl' programs, 410 lysine-rich sequences were found in 343 proteins, indicating that these sequences are not reserved for exported proteins. Some of these, for example the merozoite adhesive erythrocytic binding protein (MAEBL) (Blair et al., 2002) or the rhoptry-associated membrane antigen (RAMA) may be released into the host cell upon invasion to interact with host-cell components. Altogether, 5.9% of all parasite proteins contain lysine-rich repeats which fulfil our criteria, whereas they were found in only 3.9% of proteins predicted by our algorithm to be exported. This suggests there is no particular enrichment for these repeats in proteins exported into the host cell. It is possible expansion of lysine-rich repeats is selected

against in exported proteins which are not required to be targeted to the periphery, whereas within the parasite they are retained as neutral spacer sequences.

INPUT FILE (krichseq.txt)
K-rich sequences selected by chargesearch.pl in FASTA format

T-REKS: Identify all protein sequences containing tandem repeats

OUTPUT FILE (repeats.txt)
: List of tandemly repeated motifs within each K-rich sequence, e.g.
```
>GARP2
Length: 7 residues – nb: 8  from  27 to 84 – Psim:0.65625 region Length:58
EKKKKKQE
EKEKKKQ-
EKERKKQ-
EKKERKQK
EKEMKKQ-
-KKIEKER
KKKEEK-E
-KKKKHD
***********
```

PICKREPEATS.pl
-Uses output from T-REKS to print out all lysine-rich sequences which contain a repetitive region over 30 aa in length, as well as the full length protein.
Region Length Threshold = 30 aa
START

Region Length > 30 aa? — YES → Retreive K-rich Sequence from krichseq.txt → Retreive full-length Sequence from allexported.txt → Print out K-rich sequence and full-length sequence

NO → End of Protein? — NO → Next Repeats

YES → End of File? — NO → Next Protein

YES →

OUTPUT FILE (replist.txt)
All sequences with >20% K and >30 aa in length which contain tandem repeats, e.g.
```
>GARP1
Repeating Sequence:
DNENSVDKKKDKKEKKHKKDKKEKKEKKDKKEKKDKKEKKHKKEKKHKKDKKKEENSEVMSLY
Full Sequence:
MNVLFLSYNICILFFVVCTLNFSTKCFSNGLLKNQNILNKSFDSITGRLLNETELEKNKDDNSKSETLLKEEKDEKDDVPTTSNDNLKNA
HNNNEISSSTDPTNIINVNDKDNENSVDKKKDKKEKKHKKDKKEKKEKKDKKEKKDKKEKKHKKEKKHKKDKKKEENSEVMSLYKTGQHK
PKNATEHGEENLYEEMVSEINNNAQGGLLLSSPYQYREQGGCGIISSVHETSNDTKDNDKENISEDKKEDHQQEEMLKTLDKKERKQKEK
...
```

Figure 4.5 - PERL script for the identification of tandemly repeated motifs within lysine-rich protein sequences. The program T-REKS was used to predict short tandemly repeated protein sequences. Programs are shaded grey, and INPUT/OUTPUT text files in blue. Within the PICKREPEATS.PL algorithm, green diamonds represent a YES/NO decision and yellow boxes represent processes.

| Gene ID (Alias/Family)	Consensus Sequence	Position Within Protein	Repeat Unit Length	Number of Repeat Units	Error from Cons.	PI
PF3D7_0113000 (GARP)	EKK	119-163	3	15	0.2	10.20
	E-KE--K-KKQ-	279-338	7	7.57	0.28	10.05
	EEHKE	372-415	5	8.8	0.14	5.42
	KGKKD	417-440	5	4.8	0.21	10.44
PF3D7_0114200	KDHMKDDTKDDT*	136-232	12	8.08	0.2	5.12
PF3D7_0201900 (PfEMP3)	KNKELQNKGSEGLKENAEL	1063-1249	19	9.84	0.04	8.93
	NKDISNKDMKNKELL	1263-1315	15	3.53	0.08	8.94
PF3D7_0202000 (KAHRP)	SKKH--KD-HDGE-KKK	363-424	13	4.46	0.21	9.76
	ATKEASTSKE*	543-599	10	5.7	0.09	8.26
PF3D7_0402000 (PHISTa)	KQGGKKEEV	322-426	9	11.67	0.07	9.51
PF3D7_0404800	NNNTQ-MKGKQ	209-271	10	6.2	0.22	9.76
PF3D7_0424500	KEKSKKKHRDDKFNK	85-145	15	4.07	0.2	9.97
PF3D7_0500800 (MESA)	EKND-	474-544	20	3.55	0.07	4.73
	KEKEEV	914-964	21	8.5	0.19	4.86
	KEKEEV	1053-1091	6	6.5	0	4.90
PF3D7_0532400 (LYMP)	NKKVRGA	431-462	7	4.57	0.12	11.78
	ENKKAGT	472-517	7	6.57	0.13	9.80
PF3D7_0701900	LKKEEAKPT-D-	551-627	10	7.3	0.24	9.62
PF3D7_0726200 (FIKK 7.1)	DLLKNKEG	84-237	8	19.25	0.2	5.18
	EDKNCMKKTHENKAECEKN	255-313	19	3.11	0.05	5.94
PF3D7_1038400 (Pf11-1)	EKD	338-386	3	16.33	0.04	4.28
	PK-KEKVP-A-	9245-9547	8	37.62	0.18	10.07
PF3D7_1102300	ERKEREEREKK	134-227	11	8.55	0.2	9.37
	EREKREKKEKE	260-409	11	13.64	0.13	9.57
PF3D7_1148700	KECVPNECMK	262-337	10	7.6	0.17	8.90
PF3D7_1201000	EKDEK	296-387	5	18.4	0.08	4.71
PF3D7_1249600	NKKEDGD	560-638	7	11.29	0.18	4.68
PF3D7_1401200	KPSKYDDIRCFGEPAQKKK	76-138	19	3.32	0.21	9.88
PF3D7_1476200	KEQEKEKERKRKE	451-497	13	3.62	0.18	9.80
PF3D7_1476300	KKEEDI	372-398	6	4.5	0.19	4.88
PF3D7_1476600	KEES	461-503	4	10.75	0.14	4.63
	NKEE	512-533	4	5.5	0.09	4.61
PF3D7_1478600	KLDSQNGKNEKNEKSIPN	659-814	18	8.67	0.1	9.43
PF3D7_0936800	PKVQQQKVQP	296-352	10	5.7	0.11	11.08

Table 4.1 – *P. falciparum* proteins with charged repeat sequences predicted to target to the erythrocyte periphery. Proteins selected to be GFP-tagged and expressed in *P. falciparum* are highlighted in green. An asterisk indicates proteins which did not localise to the erythrocyte periphery. The consensus sequence, position within the protein, repeat unit length, number of repeat units, and the error from consensus were defined by XSTREAM (Newman and Cooper, 2007). Theoretical pI calculated by PROTPARAM is shown for each fragment (Gasteiger et al., 2005).

4.3.3 Lysine-rich Repeats from Multiple Known Cytoskeleton-binding Proteins are capable of Targeting the Erythrocyte Periphery

Sequences encoding lysine-rich repeat sequences from nine predicted exported proteins (**Table 4.1** – highlighted in green) were selected for testing in *P. falciparum* infected erythrocytes. GFP-tagged genes fragments were assembled and co-transfection of the plasmids with the pINT plasmid allowed integration of the DNA into an attB site within the parasite genome (Nkrumah et al., 2006). Gene expression was driven by the PfCAM promoter, and protein localisation was assessed by fluorescence microscopy of live cells. As several proteins identified were previously known to target the erythrocyte periphery by other cytoskeleton-binding domains, the lysine-rich repeating sequences of these proteins were independently cloned in-frame with a REX3 export sequence.

The lysine-rich C-terminus of LYMP has previously been implicated in interactions with the erythrocyte. The entire C-terminus of the protein was shown to bind inside-out erythrocyte vesicles (Proellocks et al., 2014); however the lysine-rich repeats alone were not sufficient for this interaction *in vitro* (Proellocks et al., 2014). The C-terminus was later shown to interact with cytoskeleton component band 3 by co-immunoprecipitation and *in vitro* binding assays, but the role of the repeats in this interaction was not tested (Oberli et al., 2016). To establish whether the lysine-rich repeat sequence of LYMP is capable of targeting the erythrocyte periphery independently *in vivo*, the sequence encoding the repeats, $LYMP_{419-528}$, was cloned at the C-terminus of GFP and the Rex3 export sequence. When expressed by *P. falciparum*, this construct resulted in a clear ring of fluorescence at the periphery of the infected erythrocyte (**Fig 4.6A**). While it is not clear whether this localisation is due to an interaction with band 3, it indicates an interaction between the lysine-rich repeats of LYMP and peripheral components. This occurs independently of the interaction between the PHISTb PRESAN domain of LYMP with the ATS domain of PfEMP1 (Oberli et al., 2014, Oberli et al., 2016).

Another PHISTb protein containing a lysine-rich repeating C-terminus, PF3D7_1476200, has previously been shown by the Osborne group to localise to the erythrocyte periphery (Tarr et al., 2014). It is unclear whether this is also due to an interaction with PfEMP1. When the lysine-rich repeats alone, $PF3D7\_1476200_{443-512}$, were expressed with a GFP tag, the fluorescent signal

was concentrated at the periphery (**Fig 4.6B**). The similarity between the two peripheral-targeting domains in LYMP and PF3D7_1476200 suggests they may play a similar role at the periphery.
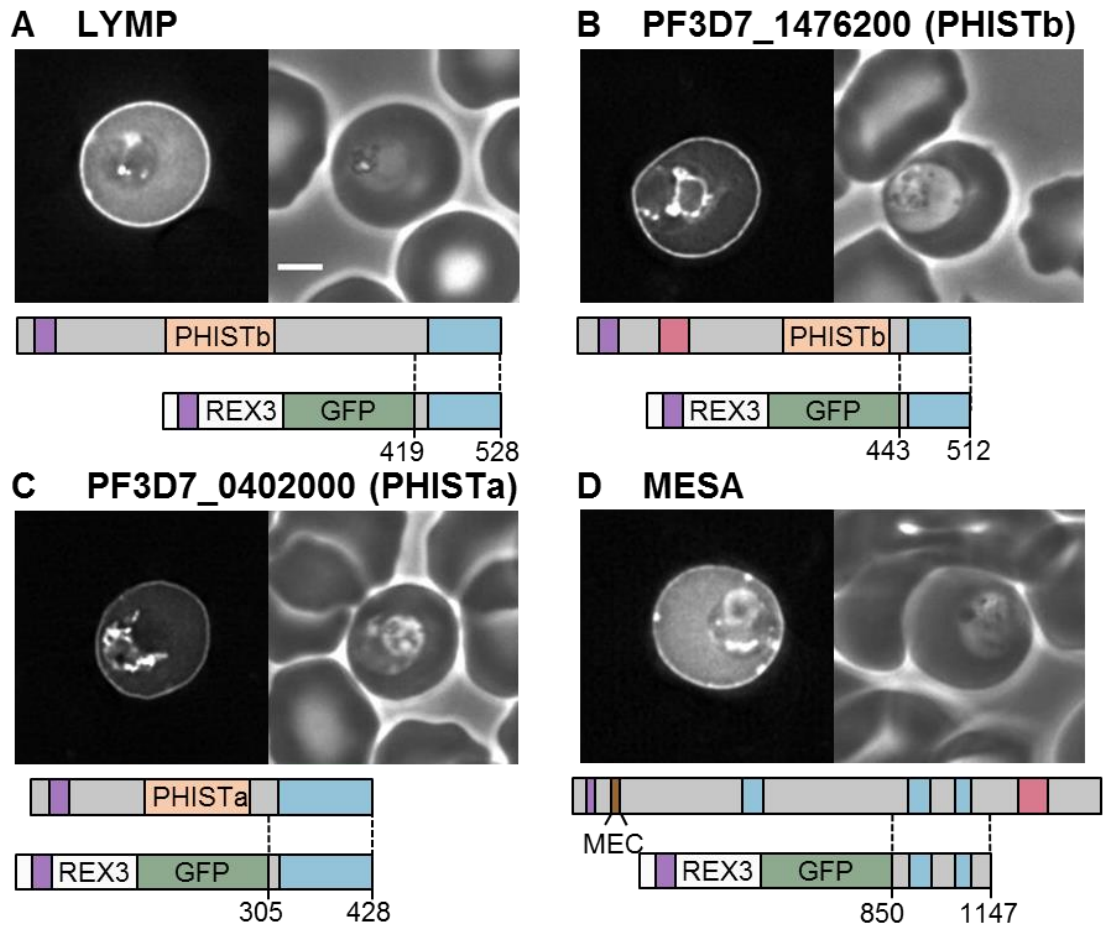
Figure 4.6 - Lysine-rich repeating sequences are present in proteins known to target the erythrocyte periphery. (A-D) Identities of proteins are shown above each image. The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of the full-length protein is shown below each image, with the lysine-rich repeat regions in blue, export sequences (signal sequence and PEXEL-HT motif) in purple, PRESAN/PHIST domains in orange, acidic sequences in red, and the MESA erythrocyte cytoskeleton-binding (MEC) motif in brown. Dotted lines indicate the protein sequence cloned into each GFP-tagged construct, shown below the full-length protein. Protein schematics are approximately to scale, with MESA downscaled by ½. Scale bar - 2 µm.

The PRESAN domain of the PHISTa protein PF3D7_0402000, unlike other PRESAN domains characterised to date, has been shown to interact with the junctional complex component band 4.1 (Parish et al., 2013). Immunofluorescence localisation of the protein indicated its co-localisation with protein 4.1 at the parasitophorous vacuole (Parish et al., 2013); however, this protein contains a canonical HT/PEXEL motif and would therefore be expected to be exported and to bind protein 4.1 at the erythrocyte cytoskeleton. Once again, the lysine-rich repeat sequence at the C-terminus of the protein, PF3D7_0402000$_{305-428}$, was robustly localised to the periphery of the host cell (**Fig 4.6C**).

MESA has also been shown to interact with protein 4.1, this time *via* a MESA erythrocyte cytoskeleton-binding (MEC) motif near the N-terminus of the protein (Bennett et al., 1997). In these experiments, no other fragments of MESA appeared to interact with inside out erythrocyte vesicles (IOVs) *in vitro* (Bennett et al., 1997). The majority of the protein is composed of highly charged repeat sequences, three of which contain over 20% lysine residues. The second lysine-rich repeat sequence and flanking sequence has duplicated to form the third repeat. A GFP fusion protein that contains both of these sequences (MESA$_{850-1147}$) was expressed by *P. falciparum* parasites. This fragment also localised to the periphery of the infected cell (**Fig 4.6D**), despite not binding IOVs *in vitro*.

A lysine-rich repeat of KAHRP has been previously shown to bind spectrin (Pei et al., 2005); a detailed look at the lysine-rich repeats of KAHRP will be given in Chapter 6.

4.3.4 Novel Lysine-rich Repeating Proteins Localise to the Erythrocyte Periphery

While the above-mentioned proteins have previously been shown to interact with cytoskeletal components or to localise to the erythrocyte periphery, other candidates containing lysine-rich repeat sequences are uncharacterised. Some contain PRESAN domains which have not been previously shown to behave as functional targeting domains. The protein PF3D7_1201000 contains two PRESAN domains belonging to the PHISTb and PHISTc subgroups, and previous unpublished work by our group indicated the PHISTb domain may weakly interact with the cytoskeleton. The protein PF3D7_1148700 contains a

PHISTc domain and is annotated as a gametocyte exported protein (GEXP12) as protein fragments have been detected by mass spectrometry in gametocyte stages of the parasite as well as in asexual stages. The lysine-rich repeats of both proteins were fused to GFP, and both exhibited fluorescence at the erythrocyte periphery. In the case of $PF3D7\_1201000_{292-397}$, the peripheral signal was relatively weak and only clearly visible in 50-80% of infected cells (**Fig 4.7A**). GFP-tagged $GEXP12_{231-370}$, however, demonstrated more robust targeting of the C-terminal repeats (**Fig 4.7B**).
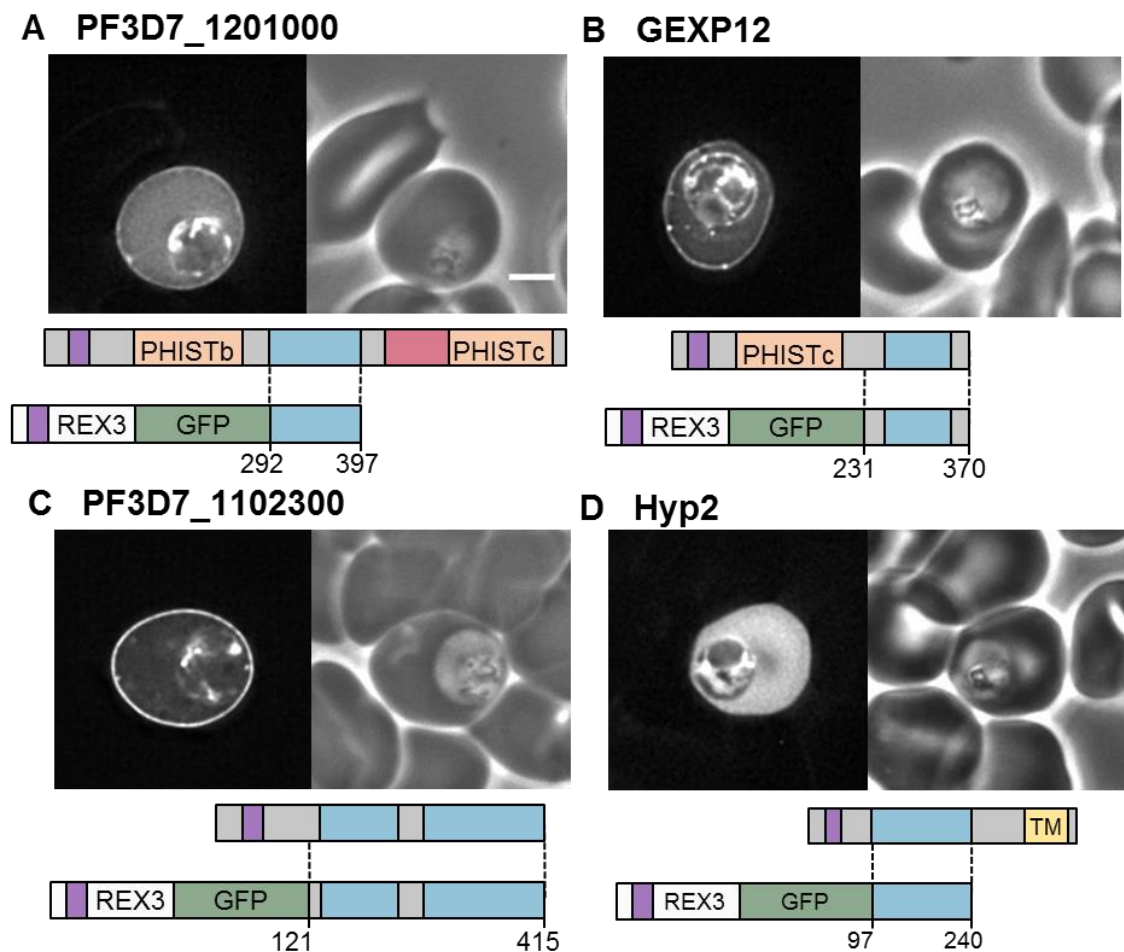
Figure 4.7 – Multiple novel proteins from *P. falciparum* target the erythrocyte periphery. (A-D) Identities of proteins are shown above each image. The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of the full-length protein is shown below each image, with the lysine-rich repeat regions in blue, export sequences (signal sequence and PEXEL-HT motif) in purple, PRESAN/PHIST domains in orange, acidic sequences in red, and transmembrane domains in yellow. Dotted lines indicate the protein sequence cloned into each GFP-tagged construct, shown below the full-length protein. Protein schematics are approximately to scale. Scale bar - 2 µm.

Unlike most lysine-rich repeat proteins discussed previously in the chapter which contain a folded domain, the PF3D7_1102300 protein, like GARP, is predicted to be entirely intrinsically disordered; the majority of the sequence is lysine-rich and repeating and no other potential cytoskeleton-binding domains are present. A fusion protein comprising the N-terminus of Rex3, GFP, and the lysine-rich sequence of PF3D7_1102300, (PF3D7_1102300$_{121-415}$), was expressed in *P. falciparum*. GFP fluorescence was again localised at the periphery of the infected cell (**Fig 4.7C**); some brighter foci of fluorescence are also seen in some cells.

Protein PF3D7_0114200 (Hyp2) is predicted to contain a C-terminal transmembrane domain as well as a lysine-rich repeat sequence. To test whether this lysine-rich repeating region can independently target the periphery, the transmembrane domain was removed and the repetitive sequence, Hyp2$_{97-302}$, was GFP-tagged. In this case the fluorescence remained localised within the cytosol of the erythrocyte, with no peripheral targeting (**Fig 4.7D**). The repeat sequence of the protein consists of eight repeats of the consensus motif KDHMKDDTKDDT, giving it an overall acidic character. The aspartate residues may therefore interfere with the targeting of lysine-residues to the periphery.

Of the nine repetitive sequences with a lysine content of greater than 20%, eight were localised to the erythrocyte periphery (see chapter 6 for analysis of the KAHRP lysine-rich repeats), indicating a low rate of false positives for predicting functional lysine-rich repeat sequences using the criteria of 20% lysine in a moderately conserved repetitive sequence. To test the appropriateness of the thresholds for lysine-richness and repeat conservation selected for the algorithm, two proteins which fell short of these criteria were selected for testing. PF3D7_1149100.1 contains six repetitions of a 40-residue motif, but is composed of only 17% lysine residues. PF3D7_1301400 (Hyp12) contains a lysine-rich C-terminus which is not predicted to be repetitive by T-REKS, however lowering the consensus match threshold in XSTREAM predicts a degenerate motif with the consensus sequence K-KKEK-QE repeated 8 times with a consensus error of 0.36. Interestingly, a GFP fusion of the lysine-rich C-terminus, Hyp12$_{297-381}$ was robustly localised to the periphery of the erythrocyte, suggesting that highly conserved repeats may not be necessary for targeting (**Fig 4.8A**). However, GFP-tagged PF3D7_1149100.1$_{121-450}$ was observed only

in the erythrocyte cytosol, indicating that a certain level of lysine-richness may be required for peripheral localisation within the erythrocyte (**Fig 4.8B**).
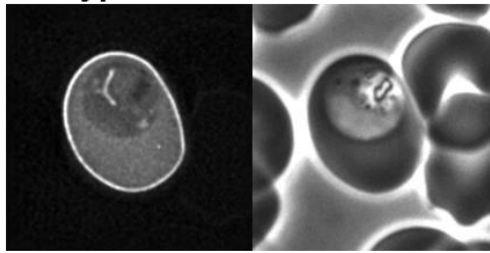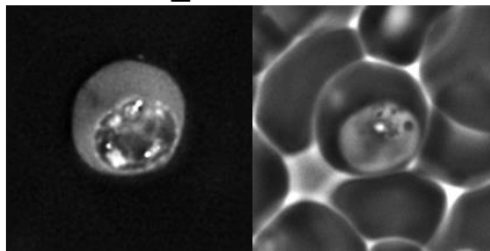
Figure 4.8 - Testing the algorithm criteria for peripheral targeting. (A) Protein Hyp12 contains a lysine-rich sequence with highly degenerate repeats below the consensus error threshold of 0.7, while (B) Protein Pf3D7_1149100.1 contains 17% lysine residues – below the 20% assigned threshold. The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of the full-length protein is to the right, with the lysine-rich repeat regions in blue, acidic sequence in red and export sequences (signal sequence and PEXEL-HT motif) in purple. Dotted lines indicate the protein sequence cloned into each GFP-tagged construct. Protein schematics are approximately to scale. Scale bar - 2 μm.

4.3.5 Characterising Full-length Peripherally-Localised Proteins

Many proteins which localise to the erythrocyte periphery are trafficked *via* Maurer's clefts, and some proteins containing lysine-rich repeats predicted to target the periphery have been shown to be retained in this compartment (Nunes et al., 2010). To confirm that the endogenous versions of all proteins tested are robustly exported and localised to the periphery in the same manner as the lysine-rich fragments, the full-length proteins were tagged with a C-terminal GFP-tag and expressed under the PfCAM promoter. MESA, KAHRP, LYMP, and the PHISTb protein PF3D7_1476200 have all been localised previously using antibodies or full-length proteins, and were not tested (Tarr et al., 2014, Horrocks et al., 2005, Coppel et al., 1988).
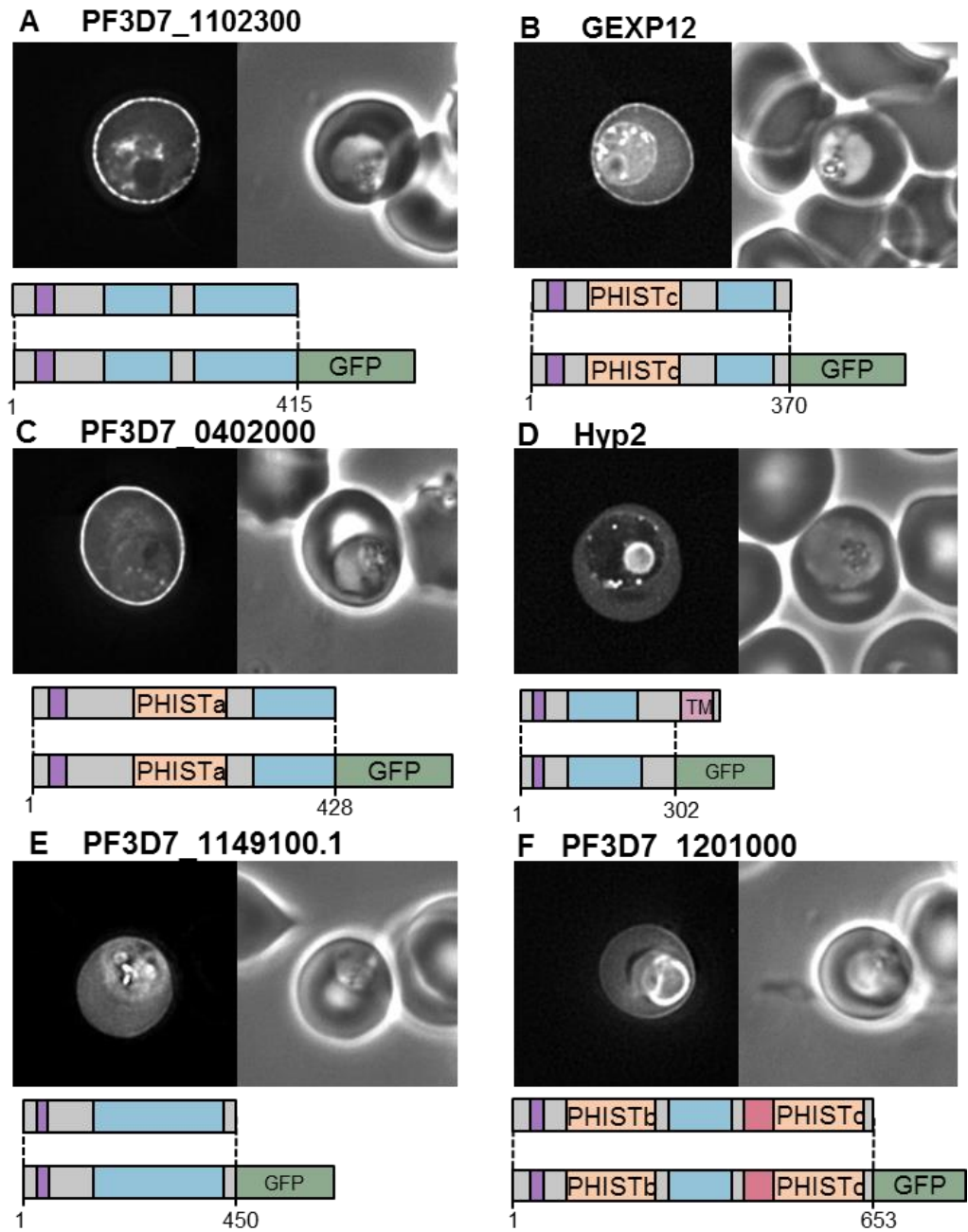
Figure 4.9 - Expression of full-length lysine-rich proteins. (A-F) Identities of proteins are shown above each image. The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of each protein is shown below each image, with the lysine-rich repeat regions in blue, export sequences (signal sequence and PEXEL-HT motif) in purple, PRESAN/PHIST domains in orange, acidic sequences in red, and transmembrane domains in pink. Dotted lines indicate the protein sequence cloned into each GFP-tagged construct, shown below the full-length protein. Protein schematics are approximately to scale.

Proteins PF3D7_1102300 and GEXP12 were robustly exported and localised to the erythrocyte periphery (**Fig 4.9A-B**). In some cases, particularly in high-parasitemia cultures, the GFP signal was also observed in numerous puncta at the periphery (not shown). The PHISTa protein PF3D7_0402000 was also robustly localised to the periphery (**Fig 4.9C**). The HYP2 protein was again expressed without the transmembrane C-terminus and was efficiently exported using its endogenous HT/PEXEL motif, as was the full-length protein PF3D7_1149100.1. Both remained in the erythrocyte cytosol, concurring with the behaviour of the individual repeating fragments (**Fig 4.9D-E**). The two-PRESAN protein PF3D7_1021000 contains a non-canonical HT/PEXEL motif 'RILSS' and the full-length protein was poorly exported into the erythrocyte, with a high proportion of the GFP signal seen in the parasitophorous vacuole. Some localisation at the periphery was observed in a number of cells (**Fig 4.9F**), however this localisation was not significantly increased relative to the REX3 sequence alone. The isolated lysine-rich fragment of this protein was also weakly targeted to the periphery.

4.3.6 Acidic and Basic Sequences in Hyp12 Interact and Inhibit Targeting

In our previous experiments we demonstrated that the lysine-rich C-terminus of Hyp12 was strongly recruited to the periphery of the host cell. Surprisingly, the full-length protein expressed with a C-terminal GFP-tag remained in the erythrocyte cytosol (**Fig 4.10A**). We hypothesised that the GFP tag may interfere with localisation by shielding interactions with the lysine-rich sequences, so the Hyp12 sequence following the predicted cleavage site of the HT-PEXEL motif was cloned at the C-terminus of GFP. The GFP signal remained mainly cytosolic in this case, although some weak peripheral signal was seen in some cells (**Fig 4.10B**). Hyp12 contains a long acidic sequence at its N-terminus mainly composed of glutamate and aspartate residues; to investigate whether this sequence may block the interaction between the basic C-terminus and the erythrocyte periphery, this sequence was removed. Remarkably, in this case the fluorescent signal was robustly localised at the erythrocyte periphery, at levels comparable to the lysine-rich sequence alone (**Fig 4.10C-D**). Given the highly-charged nature of the sequences, it is likely that the acidic sequence directly interacts with the basic C-terminus, preventing it from interacting with peripheral components. Whether some trigger may

separate these sequences and allow the protein to be targeted to the periphery remains to be seen (**Fig 4.10E**).
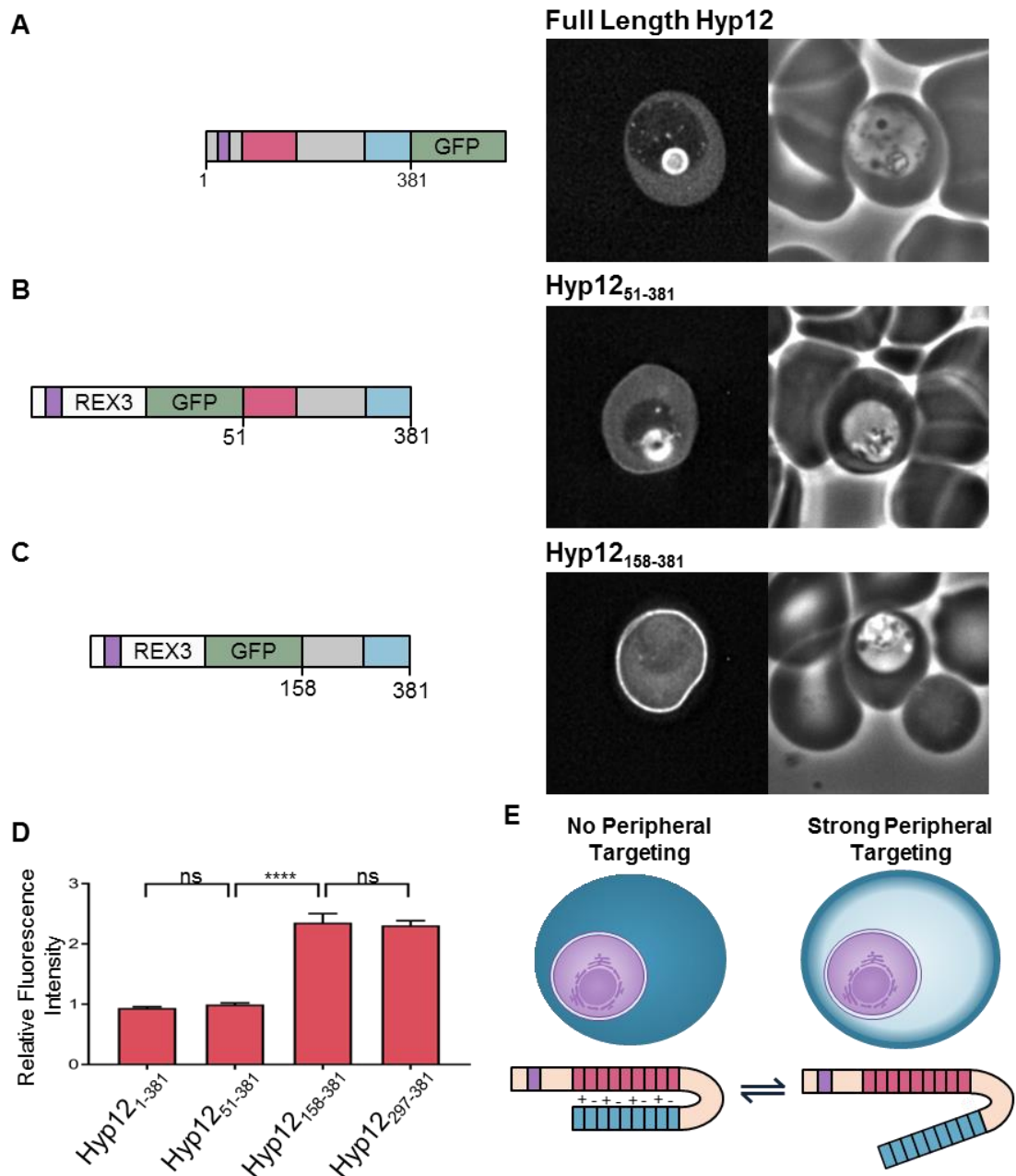
Figure 4.10 – The Hyp12 protein contains an acidic sequence which inhibits targeting to the erythrocyte periphery. (A-C) The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of the full-length protein is to the left, with acidic sequences in red, basic sequences in blue, and export sequences (signal sequence and PEXEL-HT motif) in purple. (D) Quantification of the relative fluorescent signal at the erythrocyte periphery. Error bars represent SEM. ns – Not significant (P-value > 0.5) **** - Extremely significant (P-value < 0.0001). Error bars represent SEM. (E) A potential model for dissociation of the charged sequences with an unknown trigger, which may lead to peripheral localisation.

4.3.7 A Peripherally-Targeted Protein is Upregulated in Gametocytes

A number of transcriptomic and proteomic studies have indicated that the majority of proteins tested are expressed in the asexual stage of the parasite life cycle (Oehring et al., 2012, Le Roch et al., 2004). However it has been shown that overexpression or mis-timed expression may lead to mis-localisation of proteins in the parasite; the PfCAM promoter is a strong promoter which is active from early trophozoite stages of the parasite. To confirm that proteins containing lysine-rich repeats were capable of targeting the periphery at endogenous expression levels, the PfCAM promoter in some plasmids was replaced with the endogenous promoter sequence. The promoter sequences of GARP and PF3D7_1476200 were cloned from 1084 bp, and 926 bp upstream of the genes, respectively. Strong GFP fluorescence was observed for both proteins from early trophozoite stages onward (**Fig 4.11A-B**), concurring with several microarray studies indicating a high level of transcription at this stage for both genes (Le Roch et al., 2004). The proteins were robustly exported and clearly located at the erythrocyte periphery, as seen for the full-length proteins under the PfCAM promoter.

A   GARP + Promoter
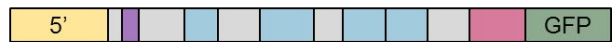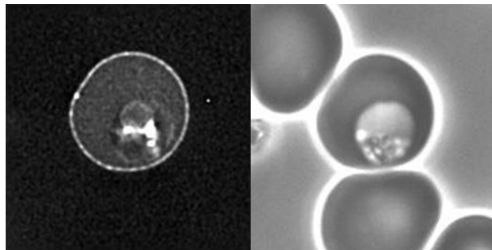
B   PF3D7_1476200 + Promoter
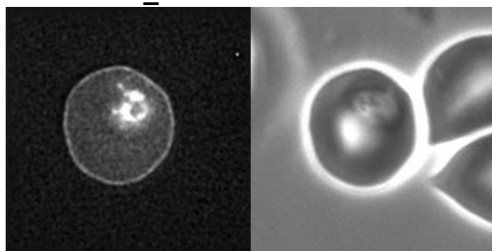
Figure 4.11 – Both GARP and PF3D7_1476200 are expressed in asexual parasites under their endogenous promoters. (A and B) GFP-fluorescence and phase contrast images are shown on the left and right panels, respectively. A schematic of the protein is shown with lysine-rich repeats in blue, the export sequence in purple, the PHISTb domain in orange, acidic stretches in red, and the promoter in yellow.

Fragments of all proteins apart from protein Pf3D7_1102300 have been detected in asexual stages by mass spectrometry analysis. This protein is almost entirely composed of lysine-rich repeats which are expected to be digested by trypsin, possibly explaining why the protein isn't detected by mass spectrometry. However, as transcription levels for the protein are also low (Le Roch et al., 2004), we decided to express the GFP-tagged full length protein under the endogenous promoter to confirm whether it is present in asexual parasites. The PfCAM promoter was replaced with the 967 nucleotides upstream of the Pf3D7_1102300 gene in the vector used to express the full-length protein.

Very low levels of fluorescence were detected from the protein in asexual parasites. A faint signal was detected above the background signal in some cells, indicating that the protein is expressed to some extent; in these cases a clear ring was observed at the erythrocyte periphery even in very weakly fluorescent cells (**Fig 4.12A**). A slightly stronger signal was observed in very late parasites. A very small proportion of young parasites displayed strong fluorescence which was robustly exported and localised at the periphery of the host cell. Interestingly, gametocytes which were present in the culture displayed a much stronger signal, indicating the protein may be expressed in sexual stage parasites (**Fig 4.11B-C**). The transcription data for the protein supports this, with an increase in transcription in sexual stage parasites relative to asexual stages (Pelle et al., 2015). It is possible that the young fluorescent parasites in the culture are early-stage gametocytes which have yet to become elongated. In later-stage gametocytes where some elongation of the parasite is evident, the fluorescent signal is clearly localised at the erythrocyte periphery (**Fig 4.11B**). The protein is still expressed in the highly elongated late stage gametocytes, although the distribution is more disperse and foci of fluorescence can be seen at the tips of the cell in some cases (**Fig 4.11C**).

A previous study has indicated that Pf3D7_1102300 is upregulated during heat shock in asexual stages (Oakley, 2006). We tested this theory by incubating the parasites at 41°C for 12-24 hours before imaging. Many more young fluorescent parasites were observed, although the signal was still low relative to the PfCAM promoter. Quantification of the protein signal by western blot confirms an increase in expression after heat shock and in gametocytes (**Fig 4.11D**). It is

therefore possible that Pf3D7_1102300 plays a role in gametocytes and is upregulated during malarial fever.
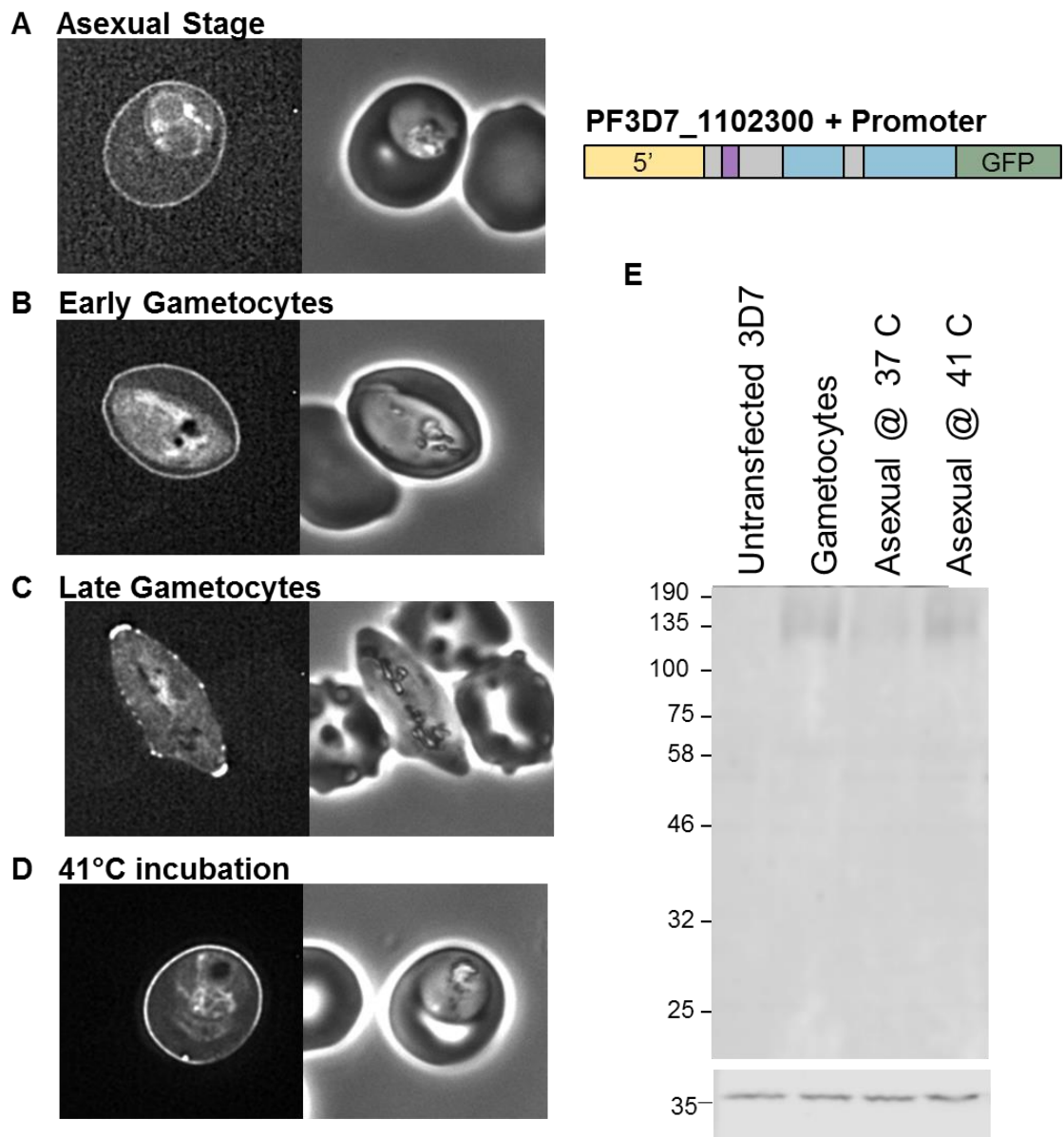
Figure 4.12 – The lysine-rich repeating protein Pf3D7_1102300 is upregulated in gametocytes and during heat shock. (A-D) Parasites expressing Pf3D7_1102300 in various life stages and under heat shock. GFP-fluorescence and phase contrast images are shown on the left and right panels, respectively. A schematic of the protein is shown with lysine-rich repeats in blue, the export sequence in purple and the promoter in yellow. (E) Anti-GFP western blot of the parasites.

These data indicate that many diverse repetitive lysine-rich sequences, in which the size of the repeated unit can vary from 3-30 residues in length, have a propensity to localise to the periphery of the infected erythrocyte in asexual and sexual stage parasites. In Hyp12, a separate highly acidic sequence completely blocks the interaction of the lysine-rich sequence with the cell periphery, likely due to an interaction between the sequences. Although many of the repeating sequences contain both acidic and basic residues, most sequences capable of targeting the erythrocyte periphery had a theoretical isoelectric point value of over 9 (**Table 4.1**). The two exceptions, MESA (pI: 4.90) and the PHISTb/c protein PF3D7_1476200 (pI: 4.71), both display the least prominent peripheral targeting, and the aspartate-rich repeats of PF3D7_0114200 (pI: 5.12) remained entirely cytosolic. It is likely given the basic nature of lysine residues that the interaction with peripheral components is conferred by electrostatic interactions, and acidic residues within the repeats may therefore interfere with binding. Further work is required to elucidate the specific requirements for peripheral localisation.

4.3.8 Variation in length between lysine-rich repeat regions in different *P. falciparum* strains

The length of repeat sequences often varies between different parasite strains (Tan et al., 2010) and the preceding experiments with GARP suggest that variation in length of lysine-rich repeats may influence the efficiency with which these sequences can target proteins to the erythrocyte periphery. Using 'long-read' sequence data generated by the PF3K consortium, we determined the extent of length variation in lysine-rich repeating sequences in the proteins shown to target the erythrocyte periphery. We compared the gene sequences of five laboratory strains of *P. falciparum* parasites (3D7, DD2, HB3, IT, and 7G8) as well as 11 parasites isolated from infected people from diverse geographic locations (Gabon, Guinea, UK, Kenya, Mali, Sudan, Senegal, Democratic Republic of the Congo, Togo, and Cambodia).
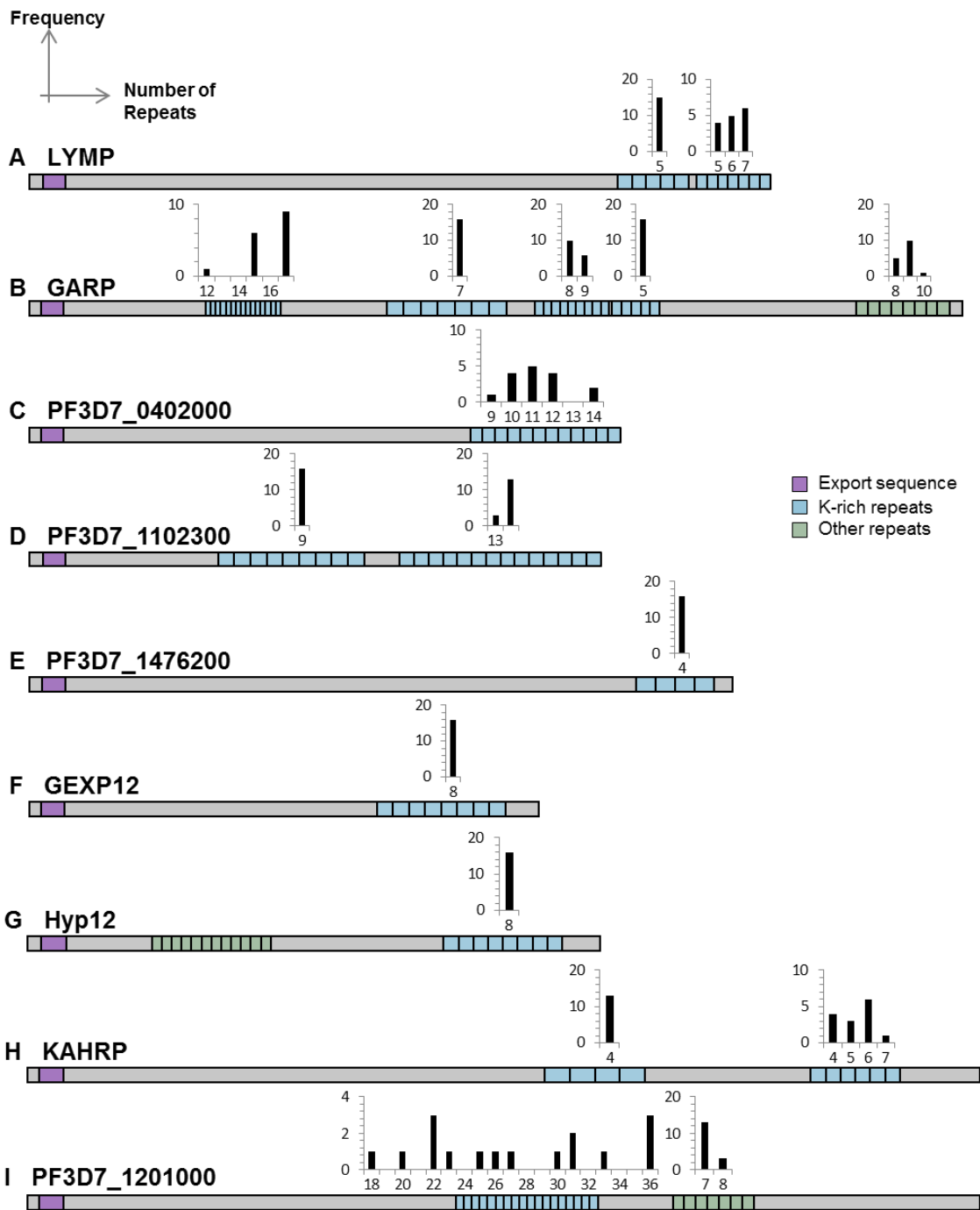
Figure 4.13 – The number of repeating units in peripheral-targeting proteins varies between different field isolates and lab strains. (A-H) Alignment of proteins containing lysine-rich repeating regions indicates considerable variation in length between strains. Histograms represent the frequency (y-axis) with which a given number of repeats (x-axis) is observed across all isolates. Schematics show the repeat architecture of the proteins from the PF3D7 strain. The export sequences (Signal sequence and PEXEL-HT motif) are in purple, lysine-rich repeats are in blue (>20% lysine), and other repeats in green. Schematics are approximately to scale.

Significant variation in repeat number is seen in many lysine-rich targeting sequences. The repeating sequences of LYMP, PHISTa protein PF3D7_0402000, and the first repeat of GARP, contain 5 to 7, 9 to 15, and 12 to 17 copies of repeating motifs, respectively (**Fig 4.13A-C**). Although unlikely to lead to a complete loss or gain of function, these changes may modulate the targeting efficiency of these protein sequences. The C-terminal repeat region of PF3D7_1102300   contains either 13 or 14 copies of the repeat motif EREKREKKEKE but the repeat sequences of the PHISTB protein PF3D7_1476200, GEXP12 and Hyp12 are invariant (**Fig 4.13D-G**). The 5' repeats of KAHRP do not vary but variations in repeat number are observed for the 3' repeats, as has been reported previously (**Fig 4.13H**) (Hirawake et al., 1997, Triglia et al., 1987). In 3D7 parasites, the protein PF3D7_1201000, contains two PRESAN domains, which are separated by 18 units of the sequence DEKEK. In all isolates the repeat unit number has expanded; in some cases by as much as twofold (**Fig 4.13I**).

In 3D7 parasites, MESA contains five repeat sequences; the first, third, fourth and fifth repeats vary significantly **(Fig 4.14A)**. In the 3D7 genome, the sequence encoding the third repeat region, which is itself variable in length (**Fig 4.14B**), is duplicated to form the fourth repeat. In other genomes the sequence is further duplicated resulting in 3 or 4 copies of this repeat sequence and its flanking regions (**Fig 4.14C**). GFP-tagged lysine-rich sequences from both MESA and PF3D7_1201000 displayed a weak fluorescence signal at the erythrocyte periphery; duplication and extension of the repeat regions may increase targeting efficiency of these sequences.
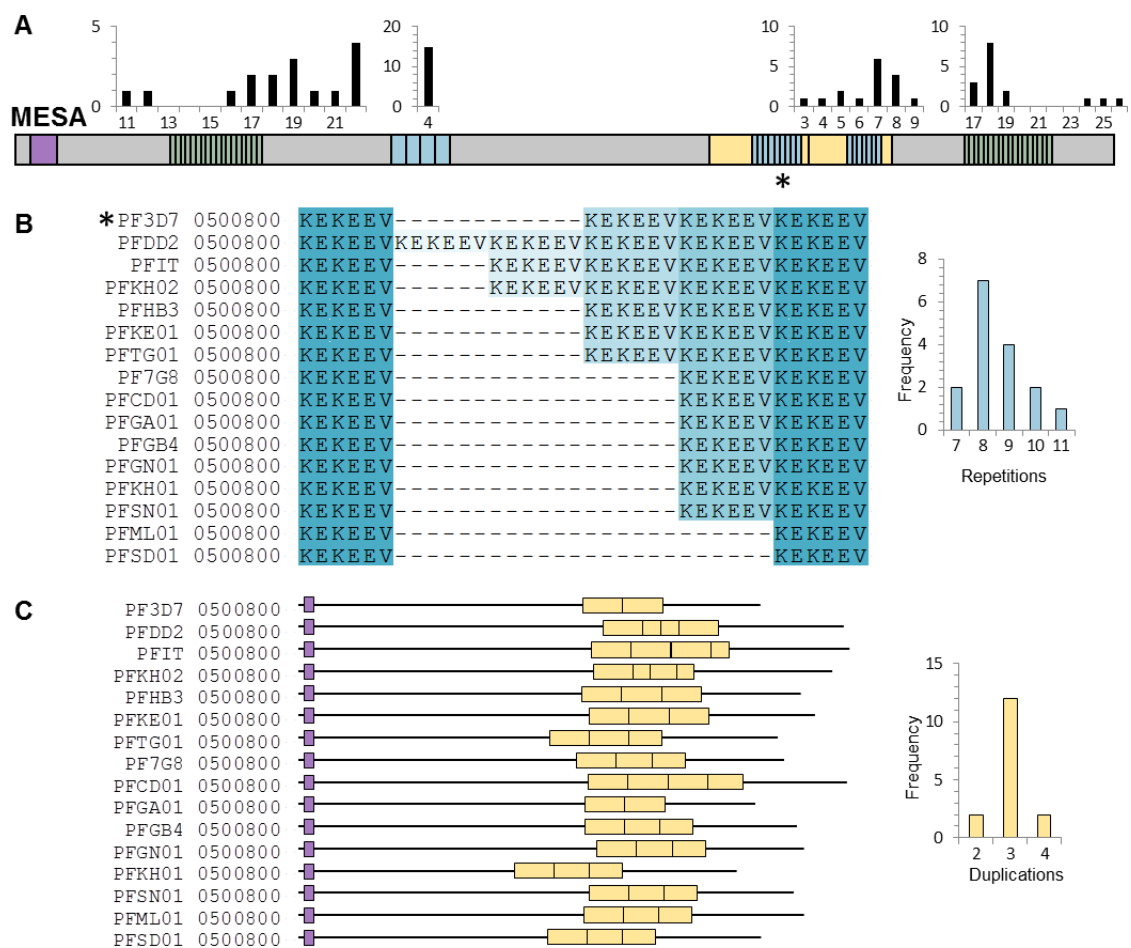
Fig 4.14 – Variation between field isolates in the repetitive sequences of MESA.
(A) Schematic of the MESA protein sequence from the Pf3D7 strain. Histograms above each repeat sequence represent the frequency (y-axis) with which a given number of repeats (x-axis) is observed across all isolates. The export sequences (Signal sequence and PEXEL-HT motif) are in purple, lysine-rich repeats are in blue (>20% lysine), and other repeats in green. (B) Variation in the first KEKEEV repeats of MESA between isolates (origin indicated by country code in gene name – see methods). (C) A large protein domain (yellow) is duplicated twice in Pf3D7 and 3 or 4 times in field isolates.

4.4 Discussion

Repetitive protein sequences are present in more than a third of *P. falciparum* proteins, yet very few examples of functional repeats have been identified to date. In addition to GARP, we have now identified three previously uncharacterised proteins which contain lysine-rich repetitive sequences capable of targeting the erythrocyte periphery: the entirely-charged protein PF3D7_1102300, PHISTc protein GEXP12, and protein PF3D7_1201000 which contains both PHISTb and PHISTc domains. These proteins may play important roles in modifying the erythrocyte membrane or cytoskeleton, and add to an increasingly complex network of proteins which localise to this important area. Elucidating how these proteins interact with the erythrocyte and with each other will greatly contribute to our understanding of the wholesale takeover of host cells by the parasite.

Interestingly, both PF3D7_1102300 and GEXP12 are expressed in gametocytes (Le Roch et al., 2004). Dramatic modifications to the host cell occur as parasites differentiate into the sexual stage; knob structures are lost in the early stages yet the parasites somehow sequester in the bone marrow and spleen, before being released back into circulation in later stages (Pelle et al., 2015). The gametocyte-infected erythrocytes initially become even more rigid than during asexual stages, but towards late gametocyte stages they become highly deformable in order to re-enter the circulatory system to be ingested by a mosquito (Tiburcio et al., 2012). These processes are largely unexplained and whether the lysine-rich proteins are involved in modulating these changes remains to be seen.

Several proteins previously shown to interact with peripheral components of the erythrocyte are present in our cohort of lysine-rich repeating proteins. Of those tested, only LYMP and KAHRP have been previously attributed with functional roles in the erythrocyte; gene deletion of both proteins resulted in a loss of cytoadhesion (Proellocks et al., 2014, Crabb et al., 1997, Glenister et al., 2002), and KAHRP is also known to modulate the rigidity of the host cell and form cytoadherent knob structures (Crabb et al., 1997, Glenister et al., 2002, Rug et al., 2006). It is hypothesised that LYMP contributes to adhesion by anchoring the PfEMP1 surface antigen to the host cell surface, mediated by its PRESAN domain and a C-terminus which contains lysine-rich repeats (Proellocks et al.,

2014, Oberli et al., 2014, Oberli et al., 2016). While it remains unclear whether the lysine-rich repeats are directly involved, their localisation at the periphery suggests they may contribute to PfEMP1 stabilisation.

Three additional proteins which contain functional lysine-rich repeats also contain a PRESAN domain, and may therefore play a similar role to LYMP at the cell surface. About 60 variants of the PfEMP1 surface antigen have been found in *P. falciparum*, and it is possible that different PRESAN domains may bind the ATS domains of different versions of the protein. Interestingly, the PHISTc PRESAN domain of another lysine-rich protein on our list of candidates, PF3D7_0936800, has also been shown to interact with PfEMP1 but with a lower affinity than LYMP (Mayer et al., 2012); whether the lysine-rich repeats are functional in this case remains to be tested. Several PHISTb proteins contain PRESAN domains capable of localising to the periphery (Tarr et al., 2014); the lysine-rich repeat protein PF3D7_1476200 contains a PHISTb domain which may also interact with PfEMP1. However the PRESAN domain of PHISTa protein PF3D7_0402000 binds to protein 4.1 *in vitro*, suggesting the PRESAN domain may instead represent a general protein-interaction domain (Parish et al., 2013). As the lysine-rich repeats of this protein are also capable of targeting the erythrocyte periphery, it may play a role in crosslinking protein 4.1 to other cytoskeletal components.

Four of the proteins tested have been included in a large-scale gene deletion screen (Maier et al., 2008). Individually deleting the genes encoding GARP, PF3D7_1102300, PHISTa protein PF3D7_0402000 and PHISTb/c protein PF3D7_1201000 does not result in a striking phenotype; however some decrease in infected cell rigidity is observed (Maier et al., 2008). Given the similarity between many of the lysine-rich proteins that we have characterized, it is likely that individual genes may be functionally redundant and that deletion of single genes may not be sufficient to reveal a phenotype. A more pronounced phenotype may be observed *in vivo* where fever and components of human serum have both been shown to affect the ability of the parasite to modulate the host cell; RESA strongly rigidifies the cytoskeleton under febrile temperatures (Mills et al., 2007), and previously knob-negative parasites can gain knob structures after injection into a human host (Stanisic et al., 2016). Similar simulation may be required to reveal a phenotype for these proteins.

Interestingly, the protein PF3D7_1102300, which is almost entirely composed of repeats, is upregulated fivefold during heat-shock (Oakley, 2006), and also in parasites selected for adhesion to a brain receptor (Claessens et al., 2012). As GARP is upregulated in children suffering from severe malaria and in parasites expressing different variant surface antigens (Vignali et al., 2011, Ralph et al., 2005), an important role for these proteins may only become apparent in the context of a malarial infection.

The full-length Hyp12 protein is localised in the erythrocyte cytosol, as shown by us and others previously (Petersen et al., 2015). However the presence of a C-terminal lysine-rich sequence in the protein which is capable of robustly targeting the erythrocyte periphery suggests that a particular trigger may 'switch on' this function. It may be possible to disrupt the interaction of the basic sequence with the inhibitory acidic sequence which is required to block localisation. This may potentially represent an additional level of adaptability for the parasite, allowing it to retain proteins in the erythrocyte cytosol until they are required at the periphery. Two other proteins belong to the Hyp12 family of 'hypothetical' exported proteins, yet neither of the other two proteins contains either the lysine-rich C-terminus or the acidic N-terminus, suggesting these oppositely charged sequences expanded at the same time. The conserved sequence between the three proteins lies between the two charged sequences, and is predicted to form a folded domain. Further investigation is required to elucidate the role of this protein and its homologues in the host cell.

Many of the lysine-rich repeating sequences in *P. falciparum* appear to be undergoing constant expansion and contraction. A high level of conservation between units in a repetitive sequence suggests a recent expansion has occurred, while mutations which build up over time can fix a repeat sequence in place. More degenerate repeats are therefore usually less prone to expand and contract (Jorda et al., 2010, Hughes, 2004). This is also observed in the PACBIO sequencing data, where the more degenerate repeats of the PHISTb protein PF3D7_1476200, the second repeat of GARP and the first lysine-rich repeat of KAHRP do not vary between isolates whereas the almost-perfect repeats of PF3D7_1201000 and PHISTa protein PF3D7_0402000 are highly dynamic. However, the highly conserved repeats of GEXP12 and PF3D7_1102300 are invariant, suggesting a more complex mechanism for

driving repeat mutability. While flexibility in the length of a repeat sequence may suggest its role is not essential for the parasite, retaining repeat adaptability may confer an advantage in different host environments.

As highly degenerate repeats such as those within Hyp12 are capable of targeting the periphery, the pattern of lysine residues within a sequence may not be important for targeting; indeed, the lengths of repeating units vary between 3 and 30 residues in the proteins tested. Repeat expansion may therefore be utilised for gaining longer lysine-rich sequences but conservation of repeats does not appear to be necessary for function. The disordered nature of the sequences likely allows sufficient flexibility for lysine residues to be positioned in relation to a binding partner. Alternatively, lysine-rich sequences may interact with different components at the periphery depending on the repetitive motif present.

There is also no apparent consensus in sequence composition between functional lysine-rich sequences; despite the high predicted isoelectric points of most of the sequences, many peripherally-localised repeats also contain acidic residues, and glutamine residues are also common. This variability between lysine-rich repeat sequences makes accurate prediction of sequences with a targeting function difficult. Despite this, the algorithm was successful in predicting a targeting function in eight out of nine proteins selected. Some untested proteins such as PfEMP3, the FIKK kinases FIKK4.1 and FIKK7.1, and the megadalton repeat protein Pf11-1 are also implicated in modulating cytoskeletal properties (Scherf et al., 1992, Brandt and Bailey, 2013, Nunes et al., 2010, Knuepfer et al., 2005, Pei et al., 2007a, Waller et al., 2007b). However, FIKK4.1 and the PfEMP1 trafficking protein (PTP3) have been shown to localise to Maurer's clefts or the erythrocyte cytoplasm, respectively (Nunes et al., 2007, Maier et al., 2008). The lysine-rich repeats of MESA and LYMP were not able to independently associate with inside out erythrocyte vesicles *in vitro* (Bennett et al., 1997, Proellocks et al., 2014), suggesting that post-translational modifications may be involved. Testing of other lysine-rich repeating sequences may elucidate more details on the criteria required for targeting.

Chapter 5: Potential Domains for Modulating Cytoadhesion are Conserved between *Plasmodium* Species

5.1 Abstract

While erythrocyte modifications are known to be important for parasite survival for *P. falciparum*, their role in other *Plasmodium* species is not well established. We show that lysine-rich repetitive sequences are present in all *Plasmodium* species whose genomes have been sequenced, and that one such sequence from a *P. knowlesi* protein is capable of targeting the erythrocyte periphery when expressed by *P. falciparum* parasites. This protein is particularly interesting as it also contains a domain homologous to the N-terminus of *P. falciparum* KAHRP, which is known to play important roles in the formation of knob structures. We confirm that the first lysine-rich repeat of KAHRP is capable of targeting the cytoskeleton as previously indicated *in vitro*, and that the conserved domain also forms GFP puncta at the host cell periphery, indicating it may play a clustering role. This conserved domain was shown to be present in more than twenty proteins across primate-infecting *Plasmodium* species, including the well-studied PfEMP3 protein which was previously shown to modulate erythrocyte rigidity. The identification of several conserved targeting sequences across different species of *Plasmodium* suggests the mechanisms for modifying the erythrocyte are more conserved than previously thought.

5.2 Introduction

Five *Plasmodium* species are known to infect humans, with *P. falciparum* causing the most severe form of malaria. *P. vivax* is the most prevalent outside sub-Saharan Africa; it is not as virulent as *P. falciparum* but can prove fatal in some cases (Anstey et al., 2012). *P. vivax* parasites can remain dormant in the liver as hypnozoites which can lead to relapse of the disease, making eradication difficult. The species *P. ovale* and *P. malariae* also infect humans but are not widespread and have comparatively mild symptoms (Collins and Jeffery, 2007, Collins and Jeffery, 2005). The fifth malaria species known to infect humans, *Plasmodium knowlesi*, is a zoonotic malaria species which primarily infects macaques but may also be transferred to humans, where it can be fatal (Cox-Singh et al., 2008). Relative to *P. falciparum*, little is known about

these other malaria species. However, the genomes of all human-infecting *Plasmodium* species apart from *P. malariae* have now been fully sequenced, allowing new insights into these important species (Ansari et al., 2016, Gardner et al., 2002, Carlton et al., 2008, Pain et al., 2008).

It was originally believed that *Plasmodium falciparum* originated from the closely related chimpanzee-infecting species *P. reichenowi*, possibly at the point our human ancestors diverged from chimpanzees more than 5 million years ago (Escalante and Ayala, 1994). Recently however, mitochondrial and nuclear DNA from parasites found in the faeces of African gorillas and chimpanzees has been sequenced, indicating that a gorilla-infecting parasite species, named *P. prae*falciparum , is the true ancestor of *P. falciparum* (Liu et al., 2010). Within this clade, named Lavernia, other closely-related parasite species were found to infect chimpanzees and gorillas in a host-dependent manner; *P. reichenowi*, *P. gaboni* and *P. billcollinsi* infect chimpanzees while *P. adleri, P. blacklocki* and *P. praefalciparum* infect gorillas (Liu et al., 2010). Phylogenetic analysis of mitochondrial gene sequences such as cytochrome oxidase subunit 1 (COX1) indicates these species are highly divergent from other human-infecting *Plasmodium* species such as *P. vivax* and *P. knowlesi* (**Fig 5.1**).

The other human-infecting *Plasmodium* species cluster with species infecting macaques and other Asian old-world monkeys; *P. vivax*, *P. knowlesi*, *P. inui*, *P. fragile*, *P. hylobati* and *P. cynomolgi* are all closely related, while *P. malariae* and *P. ovale* are more divergent. This concurs with phylogenic analysis performed on other protein sequences (Escalante and Ayala, 1994, Perkins and Schall, 2002). Malaria is also found in many other animal species; *P. berghei*, *P. chabaudi*, and *P. yoelii* infect mice and rats and have been instrumental in developing *in vivo* models of malaria disease in laboratory mice (Craig et al., 2012a). Over 250 other *Plasmodium* species have been identified, for example *P. gallinaceum* which infects birds and *P. mexicanum* which infects lizards.
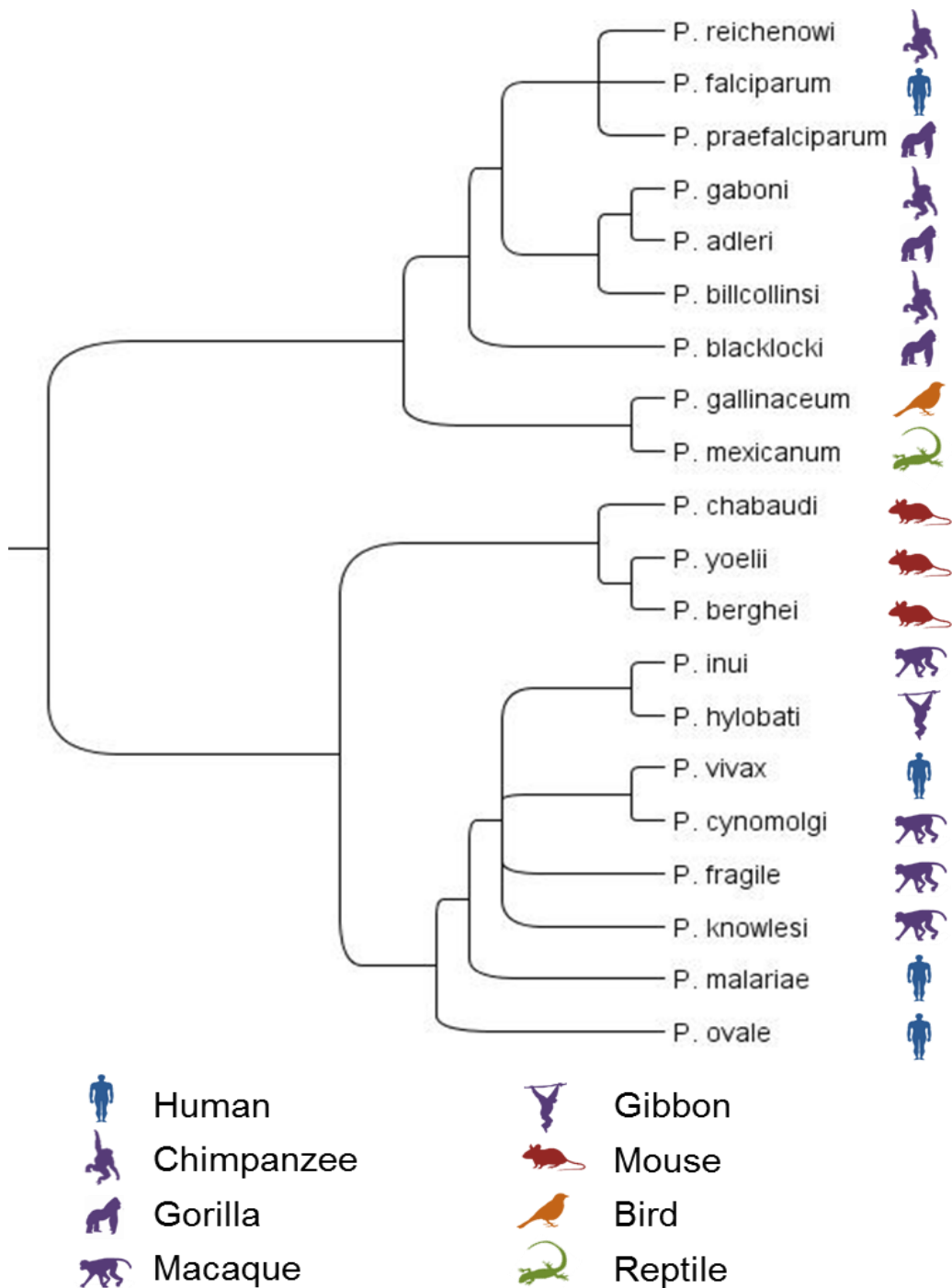
Figure 5.1 – Phylogenic tree of selected *Plasmodium* species infecting primates, birds, reptiles and rodents. A sequence fragment from the mitochondrial protein COX1 of each species was aligned by TCOFFEE and phylogeny was calculated by TREEPUZZLE maximum likelihood analysis and assembled using FigTree (Schmidt et al., 2002). Branch lengths are not to scale. The main vertebrate species invaded by each parasite species is represented on the right.

In *P. falciparum*-infected patients, sequestration of parasites in the brain microvasculature is believed to be responsible for symptoms such as seizures, coma and respiratory distress associated with cerebral malaria (Macpherson et al., 1985). Patients infected with other *Plasmodium* species rarely become comatose, however other symptoms of cerebral malaria have now been documented in both *P. vivax* and *P. knowlesi* infected patients (Anstey et al., 2012, Cox-Singh et al., 2010). In an autopsy of a patient who died following a case of severe *P. knowlesi* malaria, infected erythrocytes were found to be congested in the blood vessels of the brain, suggesting that parasite sequestration may be occurring (Cox-Singh et al., 2010). In *P. vivax*-infected patients, late-stage parasites have been observed in peripheral circulation in some patients but not in others (Lopes et al., 2014). It is not known whether cytoadhesion in these parasites is important for virulence, or whether the molecular mechanisms involved in antigen presentation and variation are conserved with *P. falciparum*.

The role of the knob-associated histidine-rich protein KAHRP in cytoadhesion is well-established; immuno-gold labelling of KAHRP in electron micrographs demonstrates a clear localisation within the electron dense knob protrusions (Culvenor et al., 1987, Taylor et al., 1987), and deletion of the KAHRP gene results in a loss of knobs and reduced cytoadhesion, as well as a 50% decrease in erythrocyte rigidity (Crabb et al., 1997, Glenister et al., 2002). Proteins with domains homologous to KAHRP were previously found in *P. knowlesi* and *P. vivax (Sargeant et al., 2006)*; however they were assumed not to be exported and their role has not been investigated. Other domains which are conserved in exported proteins from different species have been found to behave similarly; PHISTb domains in both *P. knowlesi* and *P. vivax* proteins were capable of targeting the erythrocyte periphery, for example (Tarr et al., 2014). To establish whether other erythrocyte-modulating mechanisms may be conserved between *Plasmodium* species, we investigated the role of lysine-rich repeats and other domains in a KAHRP homolog from *P. knowlesi*. Our findings indicate much greater conservation between species than previously thought.

5.3 Results

5.3.1 Peripheral Targeting of Lysine-rich Repeating Sequences is conserved between *Plasmodium* Species

To investigate whether the targeting of lysine-rich repeat regions to the erythrocyte periphery is conserved within *Plasmodium* species, we searched other parasite genomes for putative exported proteins that contain lysine-rich repeating sequences. The protein sequences for the primate-infecting species *P. reichenowi*, *P. vivax*, *P. knowlesi*, and *P. cynomolgi*, as well as mouse-infecting species *P. berghei*, *P. chabaudi* and *P. yoelii* were downloaded from the malaria database PlasmoDB. The 'exportfind.pl', 'chargesearch.pl' and 'pickrepeats.pl' programs were used with the same parameters as applied to *P. falciparum*. The proteins predicted to contain sequences with a targeting function are shown in **Table 5.1**.

The largest numbers of potential periphery-targeting sequences were found in the *P. reichenowi* genome, with 21 proteins containing lysine-rich repeats, most of which are syntenic to those identified in *P. falciparum*. The genomes of three closely-related species that infect primates; *P. knowlesi, P. vivax* and *P. cynomolgi* contained 18, 15 and 6 lysine-rich repetitive proteins, respectively. These species do not have a strong A/T-rich bias relative to *P. falciparum*, and codon usage is more balanced meaning lysine residues are not as enriched (**Fig 5.2A**). Most *P. cynomolgi* proteins were also conserved in *P. knowlesi* and *P. vivax*, however other protein families have expanded in these two human-infecting species. Notably, lysine-rich repeats are present in four *P. vivax* proteins from the PRESAN-domain family Pv-fam-E (RAD), which form an AT-rich gene cluster on chromosome 5 (highlighted in blue, **Table 5.1**). These may play a similar role to *P. falciparum* proteins such as LYMP. Likewise, in *P. knowlesi* a group of 8 homologous proteins contains 6 exported lysine-rich repeating proteins with the consensus motif 'KKEE' (highlighted in green, **Table 5.1**). These divergent gene families may play a human-specific role in the erythrocyte. Generally, fewer peripheral-targeting lysine-rich repeats were predicted in *Plasmodium* species infecting rodents; *P. yoelli, P. chabaudi* and *P. berghei* contain 10, 5 and 3, respectively. While lysine is similarly enriched in these parasites relative to *P. falciparum*, tandem repeats are less common (See **Fig 5.2B**).

Gene ID	Alias	Consensus Sequence	Position	Repeat	Number	Error from
PBANKA_0524700		EIYMENKKEEQQQSKKKEKVI	466-577	21	5.33	0
PBANKA_0711200		EIKKTPTT-	233-355	8	15.25	0.09
PBANKA_1145400			217-295	24	3.29	0.14
PCHAS_0404100	CSP	PGDK	79-175	4	24.25	0.11
PCHAS_0521700		KENGEEKVT	677-714	9	4.22	0.03
PCHAS_0524900		KKKETVIEIYMEDKKGEQQHP	437-493	21	2.71	0.12
PCHAS_1201300		KEKQEK-ERKE	957-1043	10	8.7	0.23
		QKPTD	814-960	5	29.4	0.12
PCHAS_1370100		DGKKIFEEKKES	169-201	12	2.75	0.15
		EKKSSNEKKTPG	199-231	12	2.75	0.06
PCYB_051130		EKKAEKET	237-278	8	5.25	0.1
PCYB_052240	RAD	KESKPNV	375-425	7	7.29	0.2
PCYB_063210	EBP	EGDKG	285-313	5	5.8	0.17
PCYB_081160	TRA	KKSPIIES	228-279	8	6.5	0.15
PCYB_084720		PKKGAE	299-361	6	10.5	0.19
PCYB_115490		NEKPKE	234-287	6	9	0.02
		KGKD	216-227	4	3	0.15
PKNH_0100400		KEEK	51-75	4	6.25	0
PKNH_0200400		KEEV	442-486	4	11.25	0.16
PKNH_0400300		KEEV	191-288	4	24.5	0.07
PKNH_0623000		GKKECPFKAQNSESDKCA	462-703	18	13.44	0.04
PKNH_0807900		KKEE	69-110	4	10.5	0.1
PKNH_0841300		EKGAQKPGQKKVQEKKDSN	276-320	19	2.37	0
		KKEE	69-170	4	25.5	0.06
PKNH_0900500		KYENG	162-361	5	40	0.17
PKNH_1100500		KYE-NG	162-263	5	20.2	0.17
PKNH_1149200		KEK-E-QEKK-	209-270	8	7.25	0.21
PKNH_1149700		KKEE	69-142	4	18.5	0.11
PKNH_1246800		GNKYENKHEEKL	355-379	12	2.08	0.16
		YNDK	322-350	4	7.25	0.03
PKNH_1247400		GAQKPAQQKVQEKKDSNEK	344-491	19	7.79	0.03
		EEKK	77-232	4	39	0.1
PKNH_1304600		PPKGTKKKTPTEETEQQA	153-188	18	2	0.03
PKNH_1313400		PTPKKE	266-299	6	5.67	0.09
PKNH_1325700	KAHRP	PTVSQPPK	304-363	8	7.5	0.08
		EQAKK	364-432	5	13.8	0
PKNH_1325800		ETEKQDKPKYTYGSYKYPTVK	313-404	21	4.38	0.12
		KKEKEKKDKKE	917-959	11	3.91	0.2
PKNH_1441900		KKKEKEKEKE	242-271	10	3	0.1
PKNH_1473200		KEEK	51-111	4	15.25	0
PRCDC_0053100	RIFIN	KRQKHKEQRDKNIQKIIEKDKR	82-125	22	2	0
PRCDC_0060600	PHISTB	KENNDNE	256-269	7	2	0.14
PRCDC_0111200	GARP	KKERKQKEKEMKKQEKIEKK--	229-296	20	3.4	0.19
PRCDC_0112400	EPF3	DHMK	105-212	4	27	0.14
PRCDC_0201000	EMP3	GLKENAELKNKELRNKGSD	694-793	19	5.26	0.03
		KNKDI	796-818	5	4.6	0.26
PRCDC_0201100	KAHRP	GE-KKKSKKNKD-	362-452	27	3.3	0.08
		KGATKEASTS	545-613	10	6.9	0.09
PRCDC_0500100	MESA	EKND-EKKDKVLG-EGDKEDVK	402-473	20	3.5	0.17
PRCDC_0500500	PIESP2	KHKEDH	184-232	6	8.17	0.06
PRCDC_0506400	SUB3	KNNDS	246-294	5	9.8	0.14
PRCDC_0531400	LYMP	ENKKAGS	437-494	7	8.29	0.09
PRCDC_0723000	FIKK 7.1	KKEDKSCMKKTHGNKAEDE	226-305	19	4.21	0.08
		DLIKNKEG	84-176	8	11.62	0.14
PRCDC_0727700	PTP4	FVDNKEKTLGKHE-HHEEHVKGK	1210-	22	10.68	0.2
PRCDC_1001500	PTP5	NETEKKTDQ	224-262	9	4.33	0.05
PRCDC_1037500	GSP	EKEEKIKKKKVIEKKK	1513-	16	2.19	0.28
		E-PK-KEK--AP	1623-	8	18.75	0.21
		KDVKAKHK	1550-	8	2.12	0.12
		EEKFLK	370-381	6	2	0.17
		D-EK	331-366	3	11.67	0.11
PRCDC_1100800		ERKEREEREKQ	134-392	11	23.55	0.26

113

PRCDC_1146600	PHISTc	KECIPKECIK	263-332	10	7	0.2
PRCDC_1249000	LRR	DKKEDVDNEKYG	529-593	12	5.42	0.18
PRCDC_1400500		QKKKKPSKYDDIRRFGEPT	73-139	19	3.53	0.13
PRCDC_1431700	LRR5	KNNE	604-656	4	13.25	0.15
PRCDC_1475300	PHISTB	KKEEDV	372-404	6	5.5	0.09
PRCDC_1475600		NKEENKDN	471-505	8	4.38	0.06
PVX_002507	Pv-fam-	GAMKNDTKKTPAKR	85-282	14	14.14	0.08
PVX_002535	PHIST	LEEKLNVKKLQELVKLKD	87-143	18	3.17	0.19
PVX_003535		NEMGK	192-247	5	11.2	0.11
PVX_081440		KKRLKEEE	121-142	8	2.75	0.18
		RKERK	92-103	5	2.4	0.17
PVX_081835	KAHRP	KKETK	526-564	5	7.8	0.08
		EINTE	563-642	5	16	0.14
		EKKK-	686-713	4	6.75	0.07
PVX_086900		RSHKKD	713-784	6	12	0.18
PVX_089435	RAD	KKPTA-QV	436-483	7	6.86	0.12
		EKKPDGK	484-504	7	3	0.19
		GKPVE	503-515	5	2.6	0.08
PVX_089790	RAD	KGKTPD	238-370	6	22.17	0.09
PVX_089795	RAD	KGEAK	264-318	5	11	0.07
PVX_089810	RAD	TKPKAG	238-265	6	4.67	0.04
PVX_097575	TRA	PQSKAKQQ	962-1054	8	11.62	0.14
PVX_110825	Pv-fam-	KNDDKDSFISGKS	1010-	13	4	0.13
PVX_110835		EGDQ--D-GK-EDKGEEDEDGK	258-297	18	2	0.22
		CPYKDQSVDKKE	772-823	12	4.33	0.1
		KKTANVKKGAEP	1200-	12	2.25	0.04
		DK-D-KDDK	293-337	7	6	0.32
		EEEAKKL	1146-	7	7.14	0.16
PVX_118682	EMP3	EAKKPEVKKT	1001-	10	3	0.17
PVX_119225		KKAAAP	307-362	6	9.33	0.14
PY17X_0114200	fam-b	KKADVND	284-377	7	13.43	0.11
PY17X_0114400	fam-b	DNKLDDK	175-198	7	3.43	0
PY17X_0216300		KTEKIKNEVSN	603-688	11	7.82	0.13
PY17X_0405400	CSP	KDDLPKEEK	89-122	9	3.78	0.12
PY17X_0526100		EKVIEIYMEDKKGKEQESKKK	462-766	21	14.52	0.25
PY17X_0711400		EIKKAPTSTEIKKASTST	233-307	18	4.17	0.13
PY17X_0932500	Tyr-	EELKN	391-445	5	11	0.07
PY17X_1112100		EIDKSIKKEEEHIKK-	120-173	15	3.6	0.26
PY17X_1203700		QVTDK	1242-	5	14.4	0.15
		QVSDK	677-768	5	18.4	0.14
		QVTDK	787-1070	5	56.8	0.12
PY17X_1440700	GyrA	KDE	125-164	3	13.33	0.15

Table 5.1 – Proteins from multiple parasite species contain lysine-rich repeat sequences predicted to target to the erythrocyte periphery. Gene identifiers as follows: *P. berghei* – PBANKA, *P. chabaudi* – PCHAS, *P. cynomolgi* – PCYB, *P. knowlesi* – PKNH, *P. reichenowi* – PRCDC, *P. vivax* – PVX and *P.* yoelli – PY17X. The consensus sequence, position within the protein, repeat unit length, number of repeat units, and the error from consensus were defined by XSTREAM (51). The RAD family of *P. vivax* proteins is coloured blue and a conserved group of *P. knowlesi* proteins are coloured green. The *P. knowlesi* KAHRP-like protein is in red.
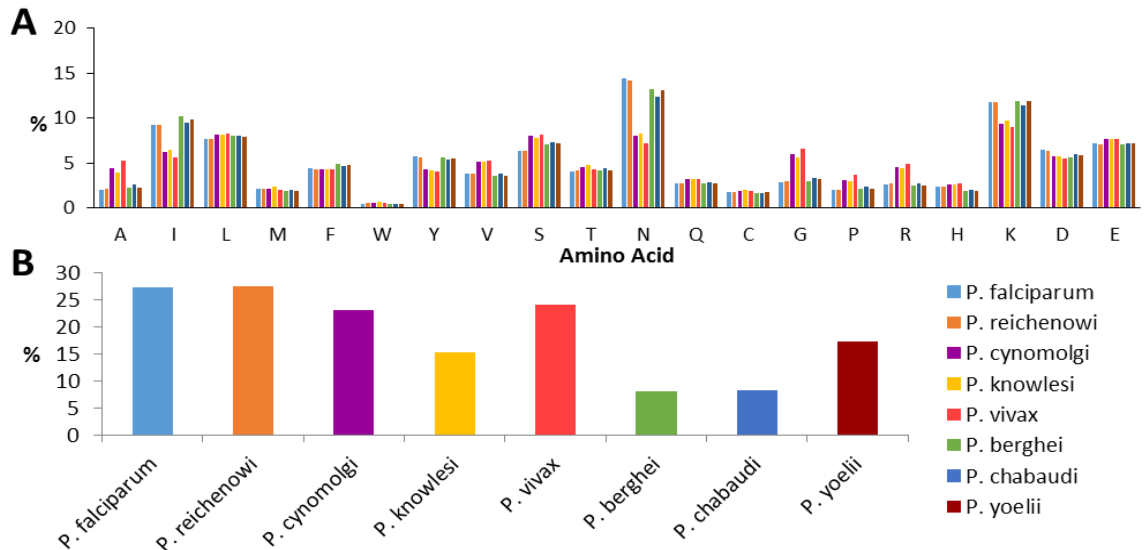
Figure 5.2 – Amino acid bias and repetitive sequences across *Plasmodium* species. A) Percentages of each amino acid represented within the open reading frames of all species. B) Percentage of proteins containing at least one repetitive sequence over 30 residues in length, as predicted by T-REKS.

5.3.2 A *P. knowlesi* Protein contains lysine-rich repeats with a targeting function

To test whether lysine-rich sequences from parasites other than *P. falciparum* have targeting functions, we selected a protein from *P. knowlesi* to be tested. PKNH_1325700 (highlighted in red, **Table 5.1**) contains a non-canonical HT-PEXEL sequence, 'RSLSV', and two repetitive lysine-rich stretches at its C-terminus (**Fig 5.3A**). Notably, it also contains an N-terminal 70-residue sequence which is predicted to form a folded domain and is homologous to the N-terminus of *Plasmodium falciparum* KAHRP (Sargeant et al., 2006). The KAHRP-like domain has been identified in PKNH_1325700 previously, as well as in three *P. vivax* proteins; however it was assumed that these proteins would not be exported as they did not contain the canonical HT/PEXEL sequence RxLxE/D/Q (Sargeant et al., 2006).

When expressed in *P. falciparum* -infected erythrocytes with a GFP tag, full-length PKNH_1325700 was efficiently exported into the host cell, indicating that its PEXEL sequence is functional and confirming that the classical protein export prediction methods may lead to false negatives. Interestingly, the GFP signal appears as a partially punctuate distribution around the periphery of the

115

red blood cell (**Fig 5.3B**). In younger parasites, fewer of these puncta were present and a continuous line of fluorescence was apparent around the periphery of the cell (**Fig 5.3C**). It is not clear whether these puncta represent knob structures; knobs have not been previously identified in *P. knowlesi*-infected cells, however the punctuate pattern is similar to that seen with anti-KAHRP antibodies in *P. falciparum* (Dixon et al., 2011, Waller et al., 2008).
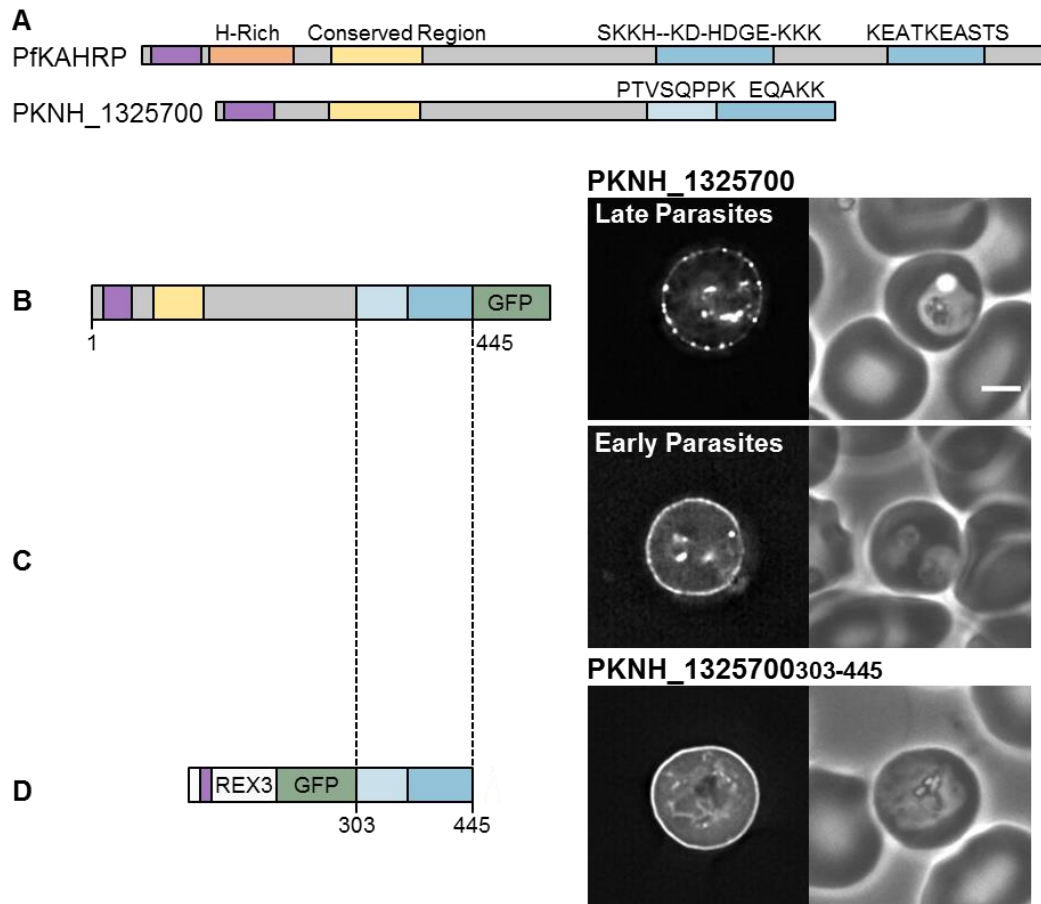
Fig 5.3 – The *P. knowlesi* protein PKNH_1325700 contains a C-terminal peripheral-targeting repetitive sequence and an N-terminal domain also found in PfKAHRP. (A) Representation of *P. falciparum* KAHRP (upper) and *P. knowlesi* protein PKNH_1325700 (lower), with lysine-rich repeat regions shown in blue and their consensus motifs shown above. The first repeat of PKNH_1325700 contains 12.5% lysine residues and is coloured light blue. The conserved region found in both proteins is in orange, the histidine-rich region in yellow and the export sequence in purple. (B-C) *P. falciparum* expressing the GFP-tagged full-length PKNH_1325700 in late parasites and early parasites, respectively. (D) The GFP-tagged C-terminal repeat region of PKNH_1325700. A schematic of the protein, a GFP fluorescence image, and a phase contrast image are shown from left to right. Scale bar – 2μm.

To test whether the lysine-rich sequence alone is able to target to the erythrocyte periphery, Rex3 and GFP were fused to residues 303-445 of PKNH_1325700; this includes the first and second lysine-rich repeat region which have lysine contents of 12.5% and 40%, respectively. This GFP tagged protein formed a continuous ring at the erythrocyte periphery (**Fig 5.3D**). The lack of GFP puncta for the lysine-rich repeats indicates they are not responsible for the apparent self-association of the *P. knowlesi* KAHRP protein. These data confirm that lysine-rich sequences from multiple parasite species may form modules with a targeting function.

5.3.3 The role of *P. falciparum* and *P. knowlesi* KAHRP domains in the host cell

Given that KAHRP is a key cytoskeleton-associated protein involved in sequestration of *P. falciparum* -infected erythrocytes, we sought to investigate the roles of the different protein domains within the *P. falciparum* and *P. knowlesi* proteins. *P. falciparum* KAHRP is composed of a signal sequence and a HT/PEXEL motif, followed by a histidine-rich region. The conserved folded domain separates this sequence from the two repetitive lysine-rich sequences in the protein, commonly named the 5' and 3' repeats. The 5' repeats are made up of four degenerate copies of the consensus sequence SKKH--KD-HDGE-KKK, while the 3' repeats are highly conserved and compose of 6 iterations of the ATKEASTSKE motif. C-terminal truncations of the protein have indicated that both repeats are required for the formation of normal knob structures and have some influence on erythrocyte rigidification (Rug et al., 2006). The histidine-rich repeats of the protein are however sufficient for localisation to the periphery of the host cell (Wickham et al., 2001). The 5' repeats of KAHRP have been shown to interact with PfEMP1 by nuclear magnetic resonance spectroscopy (Waller et al., 1999, Ganguly et al., 2015), although this could not be replicated by another group (Mayer et al., 2012). Others have indicated the 5' repeats interact with spectrin, although an extended sequence following the repeating sequence was also required *in vitro* (Pei et al., 2005).

To establish whether the repeats are capable of targeting the periphery of the host cell *in vivo*, we expressed a GFP-tagged KAHRP fragment encompassing both repeats, as well as each repetitive lysine-rich repeat individually. KAHRP$_{351-598}$, which includes both repeating sequences, was exported by the

REX3 export signal and localised robustly to the periphery of the host cell (**Fig 5.4A**). Truncation of this fragment to include the 5' repeats only, KAHRP$_{363-428}$, resulted in a similar distribution at the erythrocyte periphery; however there was a reduction in the targeting efficiency for the shorter fragment with an increase in cytosolic GFP fluorescence relative to the peripheral signal (**Fig 5.4B and D**). The 3' repeats did not demonstrate any localisation at the erythrocyte periphery (**Fig 5.4C and D**). These data indicate that the 5' lysine-rich repeats are sufficient for peripheral targeting but that the intervening sequence may also contribute to this interaction. A series of bands were seen for the 3' repeats of KAHRP in western blots, indicating it may be truncated or modified (**Fig 5.4E**).
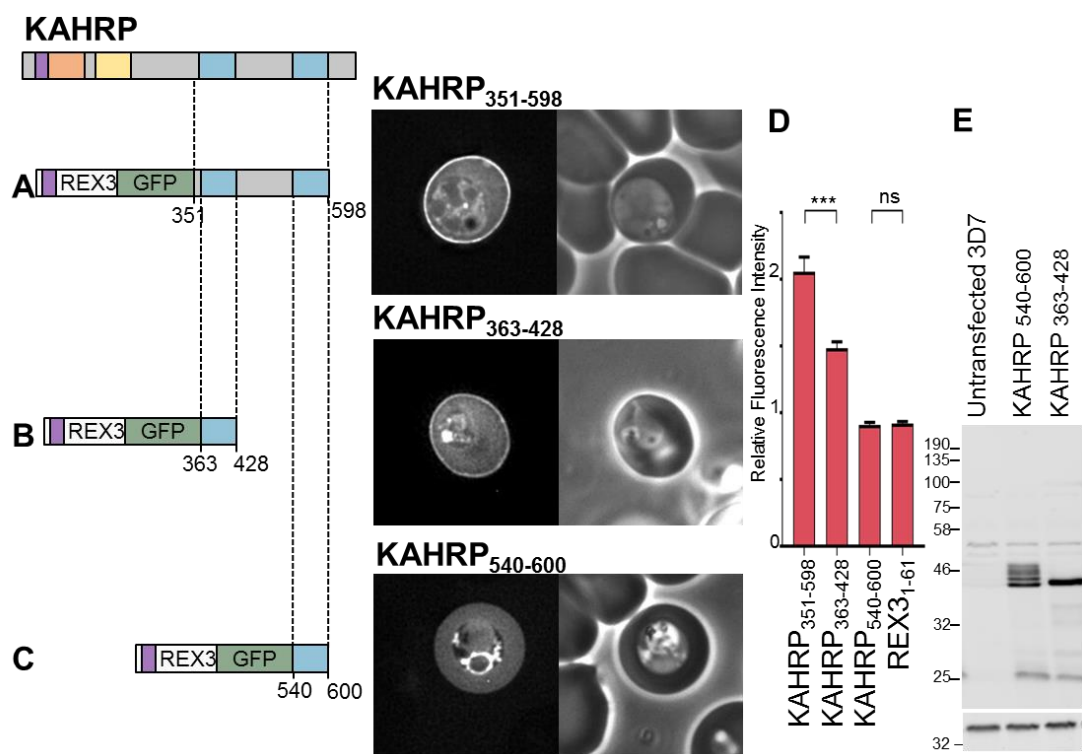
Figure 5.4 – The 5' but not the 3' repeats of KAHRP target the periphery. Representation of *P. falciparum* KAHRP (upper) and GFP-tagged truncated fragments (lower) with lysine-rich repeat regions shown in blue. The conserved region is in yellow, the histidine-rich region in orange, and the export sequence in purple. (A-C) *P. falciparum* expressing the GFP-tagged KAHRP fragments. A schematic of the protein, a GFP fluorescence image, and a phase contrast image are shown from left to right. Scale bar – 2μm. (D) Quantification of fluorescence at the periphery relative to the cytosol for each parasite line (***- extremely significant. Ns – not significantly localised to the periphery relative to the REX3 N-terminus alone). Error bars - SEM. (E) Anti-GFP western blot confirms expression of correct fragment, with anti-HAP loading control (below).

To establish whether the domain which is conserved between *P. falciparum* KAHRP and the *P. knowlesi* homolog PKNH_1325700 is involved in interactions with components of the host cell, we expressed GFP and REX3 fusions of the conserved domain. For both $PfKAHRP_{136-263}$ and $PKNH\_1325700_{75-232}$, puncta of GFP fluorescence were seen at the periphery of the host cell (**Fig 5.5A-D**). These were generally not as numerous as those seen for the full length *P. knowlesi* PKNH_1325700 protein. The number of GFP puncta varied between cells; in some parasites fewer puncta were present and fluorescence was mainly cytosolic (**Fig 5.5A**), whereas others exhibited a similar distribution to the full length protein with spots of GFP forming a rim around the erythrocyte (**Fig 5.5B and D**). Additionally, some weak non-punctuate peripheral localisation was also observed in some cases (**Fig 5.5C**). The localisation of the conserved domain was indistinguishable between the *P. falciparum* and *P. knowlesi* proteins, suggesting a similar function for the domain in both proteins.
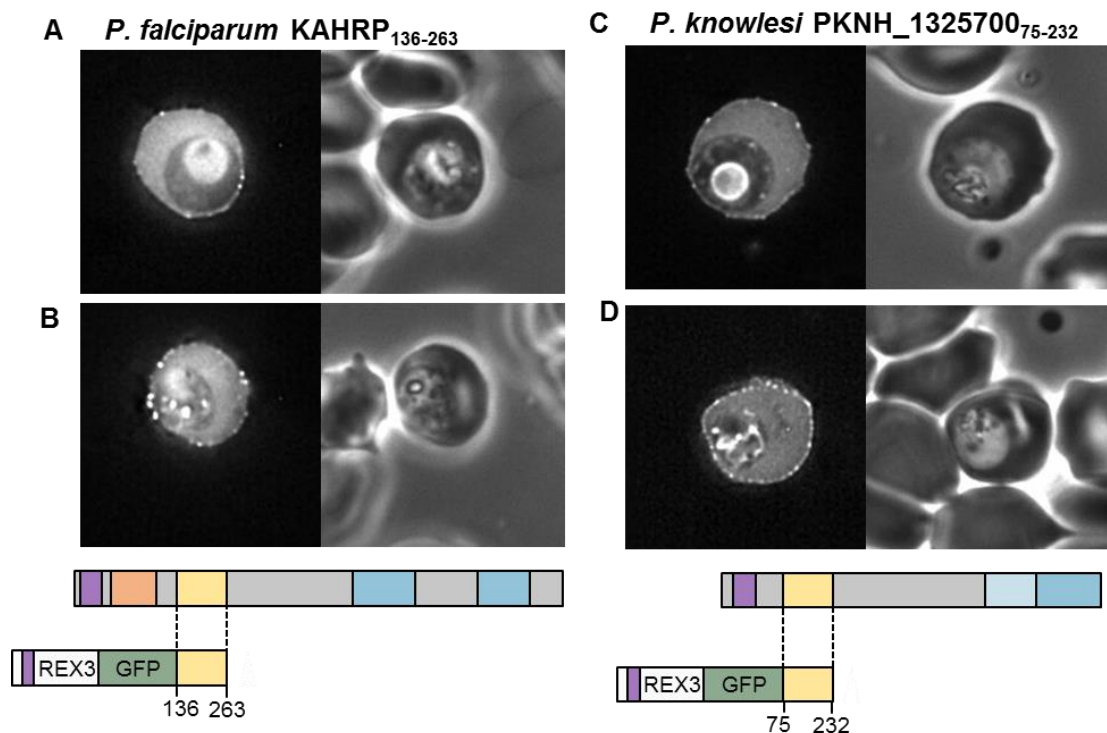
Figure 5.5 – The conserved domain of PfKAHRP and PKNH_1325700 form GFP puncta at the erythrocyte periphery. (A-B) Two representative cells are shown for each construct. The left- and right-hand images show GFP localisation and a phase contrast image, respectively. A representation of the full-length protein and cloned construct is shown below each image, with lysine-rich repeat regions shown in blue, conserved region in yellow, the histidine-rich region in orange, and the export sequence in purple. Scale bar – 2µm.

5.3.4 A conserved KAHRP-like domain in other *Plasmodium* species

The domain conserved between *P. falciparum* KAHRP and the *P. knowlesi* protein PKNH_1325700 was also previously observed in three *P. vivax* proteins. Using the hidden Markov model program HMMer, we identified a homologous domain in several additional proteins from the primate-infecting species *P. reichenowi, P. knowlesi, P. vivax, P. cynomolgi, P. fragile, P. ovale,* and *P. inui* (**Fig 5.6**). Notably, in *P. falciparum* , the conserved domain is also found at the N-terminus of the erythrocyte cytoskeleton-associated PfEMP3 protein; like KAHRP, PfEMP3 has been implicated in modulating the rigidity and cytoadhesion of the host erythrocyte (Waterkeyn et al., 2000, Glenister et al., 2002).

As the conserved domain is present in both EMP3 and KAHRP we refer to it as the EMP3-KAHRP-like domain or EKAL domain. It is predicted to form a folded domain and contains several conserved acidic residues as well as a fully-conserved tyrosine residue (**Fig 5.7A and B**). Twenty-four proteins containing an EKAL domain were found altogether within parasite species for which genome sequencing data was available. These proteins can be grouped into seven branches; five branches are closely related to PfKAHRP while two represent homologs of the EMP3 protein (**Fig 5.7C**). Remarkably, each parasite genome encodes at least one protein with a KAHRP-like EKAL domain and a lysine-rich repeating sequence that may target the protein to the periphery of the infected host cell (**Fig 5.7C**). Other repetitive sequences which are not lysine-rich are also present in many of the proteins, however only *P. falciparum* KAHRP and its *P. reichenowi* homolog contain a histidine-rich repetitive sequence at the N-terminus. PfEMP3 is also mainly composed of repetitive sequences, some of which are predicted to target the erythrocyte periphery by our algorithm (See **Table 4.1**). Although sequence homology in PfEMP3- and KAHRP-like proteins is largely restricted to the EKAL domain, it is likely that the lysine-rich sequences confer a similar peripheral-targeting function in many of the EKAL-domain proteins.

```
PF3D7_0202000 (KAHRP)  TVAN-PPSNEPVVKTQ--VFREARPGGGFKAYEEKYESKHYKLKENVVDG 47
       PRCDC_0201100  AVAN-PPSNEPVVRTQ--VFREARPGGGFKAYEEKYESKHYKLKENVVDG 47
           PO_EKAL1  AVAR-SGRKNPLLKSR--LVTEITH-GGFKEYEEKYESKHYKLKENVNDG 46
           PO_EKAL3  A-GF-GGFSSSPFRTR--FINETIH-GGYKEYEEKRESRRHTLTEDMSDF 45
           PO_EKAL2  AHRRTGNNKRPILRKR--VIKEVIN-GGFKEYEEKYESKHYKLSENVEGD 47
         PVX_081835  AMGR-RVMRKPMLRSR--VVTEFKS-GGLKEYQENYETKHYKLKENVVDG 46
        PCYB_001100  AIGR-RIIRKPMVRSP--VVTEIKR-GGIKEYQENYETKHYKFKENVVDG 46
      PI_C922_04319  ALCR-RVTRKPMVHKR--LVKEIKR-GGLKGYQENYQKKHYKLKENVVDG 46
         PVX_003520  SVVRRGGKYGPPLRTR--VVKEVTR-GGFKEYEEKYESKHYKLKENVEDG 47
        PCYB_042850  AVVRRPGKYNPPLRTR--VVKEVTR-GGFKEYEEKYESKHYKLKENVEDG 47
        PKNH_1325700  SVVRRMGKYNAPLKTR--IVKEVTR-GGFK-VEEKYETKHYTLKENVDEN 46
      PFR_AK88_04566  AVVRRMGKHNPPLRTR--VVKEVTR-GGFKEYEEKYESKHYKLKENVDDK 47
         PVX_003525  ALKYANSNNSPVFKTR--IVEEITP-AGFKKYDENYESKLFKFMESVDNG 47
        PCYB_042840  ALKYAKSKDVPVVKTR--IVEEFTP-AGFKEYNENYESKCFKFKENVENG 47
      PI_C922_02878  AVKYTKSKKTSVFKTR--IVEEFTP-AGIKKYSENYESKHFKLKETVNNG 47
        PKNH_1325800  AMRNANPGRNRIPRTR--LVEEIIP-AGYKEYTENYEVKRFKLKESVDNG 47
      PFR_AK88_04565  AMKHSNAPRNRVTRRR--IVEEIIP-GGYKEYVENYESKRFRSKESIDNG 47
 PF3D7_0201900 (EMP3)  EIKDKGDGYEEIVETKFYGMRENAL-GELDEYEERYEKKRYYLKEDGEGD 49
       PRCDC_0201000  EIKDKGDGYEEIVETKFYGMRENAL-GELDEYEERYEKKRYYLKEGGEGD 49
         PVX_118682  --VE-RIERTVQTVFC--ALKENDT-GEVEEYIETHETKRYKLKADVKDG 44
        PCYB_127900  --ID-QIERIVHTVFR--GLKENNA-GDIEEYIETHETKRYKLEADVKDG 44
        PKNH_1246500  --VD-SIKKTLETVVC--ALKENEA-GDLVEYIETHTTKHYELKTDVKDG 44
      PFR_AK88_03629  --VD-QIERTVETVVC--ALKENDK-GNIDEYIETHTTRRYKLKADVIDG 44
           PO_EKAL4  --IN-RYEKSVTTLFR--AFKENEE-GELAEYIETEGTKTYKLKNSVKDG 44

PF3D7_0202000 (KAHRP)  KKDCDEKYEAANYAFSEECPYTV-NDYSQENGPNIFALRKRFPLGMNDE- 95
       PRCDC_0201100  KKDCDEKYEAANYAFNEECPYTV-NDYSQENGPNIFALRKRFPLGMNDD- 95
           PO_EKAL1  NKECDEKYEAANYGFREKCPYEV-DQTAGPAGPDIFALRKRFPFGSNEE- 94
           PO_EKAL3  CGNCDEKYEGAKYGYRERCPYSS-DQNKKACGSTTYTLSGPLPFQNGFTN 94
           PO_EKAL2  NKDCDEKYEAANYGFREKCPYEM-EQNGDYKGPNIFELRKRFANMMNFSE 96
         PVX_081835  NKECDEKYEAAQYGFKEKCPYDV-GNYGENAGPDIFALRKKFPFGNEKKK 95
        PCYB_001100  NKECDEKYEAAQYGFKEKCPYDV-ENFGENVGPDIFELRKNFPSGKEKKK 95
      PI_C922_04319  SKESDEKYEAAQYGFKEKCSYNV-DNSSGSGGPYLLALRKRFPYGNDKEE 95
         PVX_003520  NKDCDEKYEAANYGFREKCPYEV-NPYTGANGPDIFVLKKRFHQALNKND 96
        PCYB_042850  NKDCDEKYEAANYGFREKCPYEV-NPYTGANGPDIFLLKKRFHMNLNKRQ 96
        PKNH_1325700  NKDCDEKYEAANYGFREKCPYEV-NPYTGATGPNILLLKKRFHQNLKGEE 95
      PFR_AK88_04566  NKDCDEKYEAANYGFREKCPYEV-DPYNRGQGPDILLLKKMFHQNLKRDE 96
         PVX_003525  KRDCDEKYEGSRYGFRESSPYEV-PWKGGYQGPNLIKLRNNFPMRLNFSD 96
        PCYB_042840  KRKCDEKYEGSEYGFREFSPHEV-PRKGGYRGPNLIKLRHHFPMKLNFSD 96
      PI_C922_02878  ERNCDEKYEGSQYGFKESTPYEM-PMEGGYQGPNRTKLRHNFPMRLNFSD 96
        PKNH_1325800  KRECDEKYEASGYGFKEGAPYEM-PRAGGYRGPNLRKLRDNFPMGRNSSN 96
      PFR_AK88_04565  KKDCDEKYEAANYGFREKCPYDM-PRNHGYGGPNLTVLRHNFPMRSTFYD 96
 PF3D7_0201900 (EMP3)  LKDVEEKLEETGYGFREKFPTTR-ILVKRKRNKEQKKLKEDKEKKLIAAE 98
       PRCDC_0201000  LKDVEEKLEETGYGFREKFPTTR-ILVKKKKNKEHKKLKEDKEKNLIAAE 98
         PVX_118682  VKEYDEKFEKASYGFREKLPVTEVKEYDVVE-------R---------- 76
        PCYB_127900  IKEYDLKFEKAGYGYREKLPVTEVQEYDVVE--------R---------- 76
        PKNH_1246500  IKQYDEKLENASYGFREKLPVKEVKEYDVVE--------R---------- 76
      PFR_AK88_03629  VKQYDEKFEKAAYGFREKLPVTEVQEYDVVE--------R---------- 76
           PO_EKAL4  KQVHDETFEKSIYGFRDKLPVKDVEVIDVEE--------I---------- 76
```

Figure 5.6 – Alignment of conserved domain across primate-infecting *Plasmodium* species. Proteins were aligned using T-COFFEE (54). Residues with over 70% identity or similarity are shaded in dark grey and light grey, respectively, using Multiple Align Show (55). A black line above the alignment represents the highly conserved EMP3-KAHRP-like (EKAL) domain and a dotted line represents an extended conserved domain used for assembling a phylogenic tree (**Fig 5.7**).

Figure 5.7 – The EMP3-KAHRP-like (EKAL) domain is present within multiple repeat-containing *Plasmodium* proteins. (A) Secondary structure prediction for the EKAL domain of PfKAHRP (PSIPRED (72)). (B) Sequence logo of the EKAL domain derived from all 24 proteins. Residue position is shown on the x-axis and conservation is represented on the y-axis (bits). (C) Left: phylogenic tree of KAHRP and EMP3 homologs in *P. falciparum* (Pf3D7), *P. reichenowi* (PRCDC), *P. vivax* (PVX), *P. knowlesi* (PKNH), *P. cynomolgi* (PCYB), *P. fragile* (PFR), *P. inui* (PI), and *P. ovale* (PO). Proteins with potential frameshift mutations are indicated with an asterisk. *P. ovale* proteins were assembled de novo and have been named EKAL1-4. Numbers at each node represent Quartet Puzzling (QP) support values predicted by TREEPUZZLE, where values represent the reliability of groupings (70). Right: diagrams representing each protein sequence. PCYB_001100 and PFR A0A0D9QJA3 may be truncated due to gaps in the assembled sequences.

The phylogenic relationship between EKAL domain proteins is broadly similar to that of other conserved proteins such as mitochondrial protein COX1 (**Fig 5.1**); *Plasmodium falciparum* and *P. reichenowi* form a divergent clade while proteins from *P. vivax*, *P. knowlesi*, *P. cynomolgi*, *P inui* and *P. fragile* cluster together (**Fig 5.7C**). These proteins are syntenic to each other but not to any *P. falciparum* EKAL domain proteins. In the *P. falciparum* and *P. reichenowi* genomes KAHRP and PfEMP3 are adjacent to each other in the subtelomeric region of chromosome 2; interestingly, two conserved EKAL domain proteins present in *P. knowlesi*, *P. vivax*, *P. fragile* and *P. cynomolgi* also neighbour each other on chromosome 4. Neither of these proteins belongs to the PfEMP3 clade, however, and the open reading frames lie on different strands. Only one of the syntenic pair was found in *P. inui*; this may be due to incomplete sequencing or gene loss. A group of three EKAL domain proteins are only present in *P. ovale* linages, supporting its divergence from other primate-infecting parasites. The EKAL domain was not found in any proteins encoded by mouse-infecting malaria species, indicating it may only confer a selective advantage in primates.

Our identification of KAHRP and PfEMP3 homologs in seven different primate-infecting species of *Plasmodium* suggests that the erythrocyte modifications mediated by these proteins in *P. falciparum* may also apply to the other species. KAHRP is associated with cytoadhesion and severe malaria; identifying the roles of different EKAL domain proteins may elucidate the causes of severe disease in other *Plasmodium* species.

5.4 Discussion

Cytoadhesion of *P. falciparum* -infected erythrocytes to receptors on endothelial cells is believed to be essential for parasite virulence (Smith et al., 1995, Viebig et al., 2005, Craig et al., 2012b, Fatih et al., 2012). Erythrocytes infected by malaria parasites are substantially less deformable than uninfected erythrocytes, and may therefore be detected and cleared by the spleen (Buffet et al., 2006, Safeukui et al., 2008, Buffet et al., 2009, Huang et al., 2014). The sequestration of these parasites in the brain and other organs prevents parasite clearance, and is believed to be a major cause of cerebral malaria and death (Macpherson et al., 1985). It was previously believed that only *P. falciparum* caused adhesion of infected erythrocytes, however evidence is growing that other human-infecting parasite species may behave similarly (Carvalho et al., 2010, Fatih et al., 2012). Our observation that several protein domains which are likely to play a role in erythrocyte cytoadhesion and rigidity are conserved between *Plasmodium* species supports this theory. We have identified proteins with similar domain architecture to LYMP, KAHRP and PfEMP3 in multiple primate-infecting parasite species, indicating that erythrocyte modifications may occur in a conserved manner.

Cytoadhesion is mediated by the interaction of variant members of the PfEMP1 family of surface proteins with endothelial cell receptors such as CD36, VCAM and ICAM-1 (Pasloske and Howard, 1994). Recently it has been shown that *P. knowlesi*-infected erythrocytes extracted from patients are able to bind to recombinantly expressed VCAM and ICAM-1, but not to CD36 (Fatih et al., 2012). Likewise, *P. vivax*-infected erythrocytes were also capable of binding to placental cryosections and endothelial cells, as well as isolated ICAM-1 but not CD36 (Carvalho et al., 2010). Interestingly, both VCAM and ICAM-1 require induction by cytokines for expression while CD36 is constitutively expressed, suggesting that *P. knowlesi* and *P. vivax* may only cytoadhere during an inflammatory response. While the PfEMP1 protein family is not present in the other parasite species, both *P. knowlesi* and *P. vivax* contain gene families which have been linked to cytoadhesion. In *P. vivax*, antibodies against the VIR family of proteins were able to block adhesion to endothelial cells (Carvalho et al., 2010). In *P. knowlesi*, the schizont Infected Cell Agglutination variant antigen (SICAvar) family is presented on the surface of infected erythrocytes

and has been shown to be antigenically variant (Corredor et al., 2004, Barnwell et al., 1983, Howard et al., 1983). While there is no sequence homology between these gene families and PfEMP1, it remains to be seen whether they perform similar functions.

The clustering of the PfEMP1 surface antigen in 'knob' protrusions on the erythrocyte surface is essential for cytoadhesion under flow conditions as experienced within the host (Crabb et al., 1997, Baruch et al., 1996). To date however, no mechanism for surface antigen clustering has been observed in other human-infecting parasite species, and no knob structures have been documented on the cell surface of infected erythrocytes. KAHRP is a major component of these knob structures; deletion of the gene results in an absence of knobs and reduced cytoadhesion under flow conditions (Crabb et al., 1997). Our identification of a protein domain homologous to KAHRP in all known primate-infecting *Plasmodium* species indicates that antigen clustering may also be conserved. While this conserved domain was previously identified in *P. knowlesi*, *P. vivax* and *P. fragile*, the significance for antigen presentation was not discussed as it was assumed the proteins would not be exported (Sargeant et al., 2006). We show that the full length *P. knowlesi* homolog is robustly exported into the host cell using a non-canonical HT/PEXEL motif. Similar motifs are present in other EKAL-domain proteins indicating they are also likely to be exported into the host cell. The *P. fragile* homolog identified in the previous study was not found in a more recently sequenced genome of the parasite; the previous identification of this protein is therefore assumed to be a contaminant from *P. falciparum* infection as the sequences are almost identical.

When expressed independently, the EKAL domains from *P. knowlesi* and *P. falciparum* KAHRP formed puncta of GFP at the erythrocyte periphery. It is not clear whether this is due to self-association between the EKAL domains themselves or whether the domain clustered by association with other proteins localised to the erythrocyte periphery, such as complete knob structures. It may be of interest to express the fragment in *P. falciparum* DD2 parasites which lack knob structures due to a deletion of the KAHRP gene. Full-length KAHRP has been previously shown to self-associate *in vitro*, and it is believed this is responsible for the assembly of the protrusions (Oh et al., 2000). As the lysine-rich repeats of both *P. falciparum* and *P. knowlesi* KAHRP proteins form a

uniform distribution at the periphery, it is possible the EKAL domain is responsible for the punctuate distribution of the full-length *P. knowlesi* protein. The lysine-rich repeats may anchor the protein to the cytoskeleton while the EKAL domain self-associates into clusters. The lack of electron-dense protrusions in electron micrographs of erythrocytes infected by other *Plasmodium* species may be explained by the absence of a histidine-rich region in these KAHRP homologs; histidine-rich motifs may bind metal ions to give an electron dense appearance. A closer inspection of these cells may therefore elucidate some form of protein clustering. The remainder of each of the EKAL domain proteins is highly divergent and contains many different repeat sequences; these may have adapted to interact with different surface antigens.

The binding partners of *P. falciparum* KAHRP at the periphery are controversial; the 5' repeats have been shown to interact with PfEMP1 by some groups (Oh et al., 2000, Ganguly et al., 2015) but not others (Mayer et al., 2012), and they have also been shown to bind to spectrin (Pei et al., 2005). A sequence prior to these repeats has been shown to bind Ankyrin *in vitro* (Magowan et al., 2000, Weng et al., 2014), although a larger fragment of the protein was shown to bind to a complex of actin, spectrin and protein 4.1 (Oh et al., 2000), which is located at the opposite end of the spectrin filaments to Ankyrin. An *in vivo* analysis of KAHRP domains was previously done by C-terminally truncating GFP-tagged KAHRP (Rug et al., 2006). While this indicated that the 5' and 3' repeats are both involved in the formation of complete knob structures, they did not elucidate a targeting function for these sequences as the N-terminal histidine-rich sequence was present in all fragments tested and is itself sufficient for localisation to Maurer's clefts and the erythrocyte periphery (Waller et al., 1999).

We show that the 5' repeats but not the 3' repeats are capable of localising to the erythrocyte periphery. A fragment encompassing both repeats demonstrated stronger peripheral targeting; this concurs with *in vitro* protein-binding experiments between KAHRP fragments and spectrin which required an extra sequence following the repeats (Pei et al., 2005). The sequence immediately adjacent to the 5' repeats is also lysine rich and likely originated from the same repeated sequence as some motifs are still present, however the sequence has degenerated over time. A series of bands were seen for the 3' repeats of KAHRP in western blots, these may represent degradation of the

protein upon cell lysis. Alternatively, differential phosphorylation of the protein may affect the electrophoretic mobility of the fragment; each KEATKEAST repeat has previously been shown to contain at least two phosphorylation sites (Lasonder et al., 2012). These may modulate protein targeting; however we observed no peripheral localisation of these repeats under any conditions.

We identify for the first time a homologous domain between KAHRP and PfEMP3. As the genes are adjacent to each other at the subtelomeric region of chromosome 2, it is likely a gene duplication event occurred and the homologous domain was retained whereas the rest of the gene diverged. The proteins are expressed at similar times in young parasites and are transported to the erythrocyte cytoskeleton *via* the Maurer's clefts (Knuepfer et al., 2005). PfEMP3 is made up of several repeating sequences, some of which are predicted by our algorithm to target the erythrocyte periphery. C-terminal truncations of the protein have been GFP-tagged previously and indicated a region between residues 120-500 is responsible for targeting to Maurer's clefts and the erythrocyte periphery within infected erythrocytes (Knuepfer et al., 2005). This region includes the EKAL domain but is also rich in lysine residues therefore it is not clear which sequence is responsible for targeting. Others have shown that a 14-residue N-terminal fragment of the protein binds to spectrin and actin *in vitro* (Waller et al., 2007a); however this sequence includes the HT/PEXEL motif of the protein and is therefore cleaved in the mature form. In this study no other regions of PfEMP3 were found to bind to inside-out erythrocyte vesicles (Waller et al., 2007a), indicating other malaria proteins or post-translational modifications may be required for the localisation of the protein at the periphery. Like KAHRP, PfEMP3 has been implicated in modulating rigidity (Glenister et al., 2002). It has been suggested that blocking PfEMP3 in Maurer's clefts blocks PfEMP1 transport to the cell surface (Waterkeyn et al., 2000), but unlike KAHRP, deletion of the gene had no significant impact on the formation of knob structures. The expansion of different repetitive sequences in the two proteins may therefore have generated different functions.

We confirm that lysine-rich repeats capable of targeting the erythrocyte periphery are conserved across *Plasmodium* species. The repetitive C-terminus of *P. knowlesi* KAHRP robustly localises to the periphery when expressed in *P.*

falciparum -infected erythrocytes. Localisation of the protein in *P. knowlesi* – infected erythrocytes will be necessary to confirm that no other *P. falciparum* proteins mediate the targeting function, however. Many other lysine-rich repeats were predicted to be functional across all *Plasmodium* species. While the sequences of these repeats differ, it is possible they all behave as variable-length targeting modules as seen in *P. falciparum*, and may diversify the behaviour of other functional domains such as the EKAL or PRESAN domains, which are also conserved between *Plasmodium* species. Indeed, PHISTb proteins in both *P. knowlesi* and *P. vivax* have also been shown to target the erythrocyte periphery (Tarr et al., 2014). Several proteins with domain arrangements similar to those within LYMP were observed in other species. For example, many of the RAD family proteins of *P. vivax* also contain a PRESAN domain and lysine-rich C-terminal repeats. Like KAHRP, the lysine-rich repeats of LYMP are likely to be functionally important due to the protein's role in cytoadhesion, and these RAD proteins may play a similar role in anchoring surface antigens to the erythrocyte surface.

While we have demonstrated that many similarities exist between Plasmodium species, there are also many differences. The varying levels of repeat expansion and A/T bias in the genomes of the different *Plasmodium* clades may confer a selective advantage in particular hosts. For example, much fewer lysine-rich repeats are present in the genomes of mouse-infecting malaria species, possibly indicating a primate-specific role for these proteins. Conversely to *P. falciparum*, *P. vivax* parasites do not rigidify the host reticulocytes they invade (Suwanarusk et al., 2004, Handayani et al., 2009). This may allow *P. vivax*-infected cells to evade the spleen in an alternative way. Vesicular caveolae known as Schüffner's dots (Aikawa et al., 1975)are also visible at the periphery of *P. vivax and P. ovale*-infected cells (Udagama et al., 1988); no function has been assigned to these structures. Comparing peripherally-localised proteins between species may explain the differences observed and lead to a greater understanding the mechanisms underlying severe malaria in some *Plasmodium* species.

Chapter 6: General Discussion

Repetitive protein sequences have recently been shown to be crucial for protein function in many organisms from yeast to dogs (Verstrepen et al., 2005, Fondon and Garner, 2004)(reviewed in (Gemayel et al., 2012)). For example, the selectivity filter of the nuclear pore complex is formed by polypeptide segments comprised of phenylalanine and Glycine rich repeats (Schmidt and Gorlich, 2016), repetitive sequences in many RNA-binding proteins are important for protein localization in RNA granules (Kato et al., 2012), glutamine-rich repeats modulate transcription factors activity (Gemayel et al., 2015), serine/arginine rich repeats form splicing activation domains (Philipps et al., 2003), and histidine repeats serve as a nuclear speckle localization sequences (Salichs et al., 2009). Yet despite containing more repetitive proteins than any other species whose genomes have been sequenced to date, there are currently few functions assigned to repeats in *Plasmodium* (McHugh et al., 2015, Ferguson et al., 2014). We show that several proteins from *P. falciparum* contain lysine-rich tandemly repeating sequences that confer a peripheral localisation in the infected erythrocyte. Four of the nine proteins identified were previously uncharacterised, including the glutamic acid-rich protein (GARP) which contains three distinct lysine-rich repeat sequences with a targeting function.

The underlying mechanism behind the prolific expansion of repetitive sequences in *P. falciparum* relative to other species is not known. We show that novel repeats have expanded in GARP proteins from closely related species on a relatively short timescale. In general, repeats are thought to expand and contract due to the dissociation of DNA strands during replication, with DNA secondary structures leading to re-annealing of the new strand at the incorrect position (see **section 1.3.2**) (DePristo et al., 2006). This process may occur more frequently in the highly AT-rich genome of the parasite, with weaker base-pairing interactions facilitating the dissociation of DNA strands. Our analysis of the genome of the similarly AT-rich organism *Dictyostelium discoideum* (Szafranski et al., 2005) reveals that it also contains a high number of repetitive sequences relative to other species (13% of all protein sequences – not shown), suggesting that AT-richness may be sufficient to drive repeat protein expansion. However this is much lower than the 27% of proteins which contain repeats in *P. falciparum*, and not all repetitive sequences are located in AT-rich regions of

the *P. falciparum* genome. Additionally, *P. vivax* and *P. cynomolgi* contain a much lower bias towards AT-rich codons yet are also enriched in repetitive sequences, while the AT-rich genomes of *P. berghei* and *P. chabaudi* contain relatively few repetitive proteins. This may be due to selective pressure within these species.

Expansion of repeats has been studied in detail in the case of Huntington's disease where the expansion of CAG repeats forming poly-glutamine tracts is related to disease onset. A number of DNA repair pathways are implicated in this process (Jonson et al., 2013b, Lin and Wilson, 2007, Zhang et al., 2012); it is therefore interesting to note that *P. falciparum* does not utilise the canonical non-homologous end joining machinery to repair double strand breaks, instead recruiting its own repair machinery (Kirkman et al., 2014). This pathway has not been characterised however it may contribute to trinucleotide repeat formation and expansion by allowing greater dissociation of DNA strands. Oxidative stress also contributes to the extension of CAG repeats, an effect attributed to an increase in DNA breakages (Jonson et al., 2013a). Reactive oxygen species are produced by the parasite upon infection (Atamna and Ginsburg, 1993), which may also result in an increase in repeat formation.

Repeat expansion is often considered detrimental for protein function; the aggregation of asparagine and glutamine rich repeats is associated with neurological disorders and prion disease (Halfmann et al., 2011, Perutz et al., 2002). In *P. falciparum*, a quarter of proteins contain asparagine-rich tracts which require additional chaperones to inhibit aggregation during malarial fever (Muralidharan et al., 2012). The mystery of why the malaria parasite has evolved a unique propensity to expand repeats in this manner has perplexed scientists for decades, particularly as so few repetitive sequences appear to be essential for protein function. Our observation that several of these sequences represent localisation sequences may in part explain why they appear to be selected for. The utilisation of repeat expansion for the formation of variable high-avidity binding sequences may represent enough of an advantage for the parasite to overcome any adverse effects.

By comparing the GARP sequences of the closely related parasites *P. falciparum*, *P. reichenowi* and *P. gaboni*, we demonstrate that repeat expansion can generate functional sequences *de novo*. Peripheral-targeting sequences

have expanded in *P. falciparum* whereas they are absent or non-functional in the other two species. It has been hypothesised previously that repetitive sequences may play important roles in protein evolution (Kashi and King, 2006, Andrade et al., 2001, DePristo et al., 2006, Gemayel et al., 2010, Gemayel et al., 2012, Jansen et al., 2012, Toll-Riera et al., 2012). Many repetitive sequences are present in eukaryotic proteins however these are usually highly conserved between protein homologues. Most form modular arrays of folded domains such as Ankyrin repeats, kelch repeats or leucine-rich repeats, which are collectively referred to as solenoid domains (Andrade et al., 2001). While it is likely that solenoid domains originated from repeat expansion, the sequence conservation between each repeat has been lost in most cases (Kajava, 2012, Jorda et al., 2010). These sequences are therefore unable to expand and contract and evidence of their *de novo* formation in the distant past has not been found. We present evidence of *P. falciparum* actively generating new functional domains, and the high level conservation between most repeats indicates that very few have yet to be fixed in place and likely retain the capability to expand and contract (Jorda et al., 2010).

As well as allowing the parasite to generate new targeting sequences, the dynamic nature of repetitive sequences may provide the parasite a means of rapid adaption to selective pressure. We show that many of the sequences capable of targeting the erythrocyte periphery contain a variable number of repeated motifs between the fifteen isolates tested. The truncations of the GARP lysine-rich repeat demonstrated that targeting efficiency is directly proportional to the number of repeats present; therefore the propensity to target the periphery of the host cell may naturally vary for the repetitive proteins investigated here. While no repetitive sequences appeared to be contracted to the point where they would no longer be functional, long-read sequencing data for additional isolates may uncover greater variation. Within proteins which modulate key properties of the host cell such as rigidity, cytoadhesion and nutrient import, changes to repeat length and therefore targeting efficiency may directly modulate such effects. We present a model in which repeat expansion and contraction can generate novel targeting modules or modulate the targeting efficiency of exported parasite proteins (**Fig 6.1**).
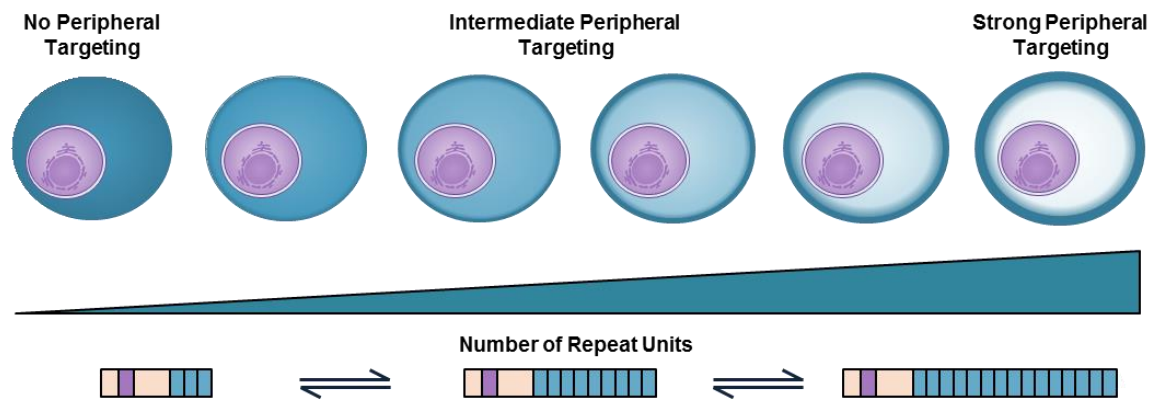
Figure 6.1 – Evolution of novel targeting domains and modulation of targeting efficiency by repeat expansion.

Several studies indicate that alteration of repeat sequence length can modulate protein function (Gemayel et al., 2010). For example, the length of the repeated sequences in the cell surface adhesion protein Flo1 modulates adhesion of yeast cells (Verstrepen et al., 2005). Changes in the lengths of repeating sequences can also modulate the activity of transcription factors (Gemayel et al., 2015, Katti et al., 2000) and changes in dog skull morphology can be correlated with altered lengths of repetitive sequences in transcription factors (Fondon and Garner, 2004). Repeats in the carboxy-terminal domain (CDT) of RNA polymerase II have been shown to directly influence the protein's function by recruiting transcription factors (Meinhart et al., 2005). The length of the CTD in all mammals is tightly regulated with exactly 52 repeats present in all species; truncation of repeats has been shown to impair function (Chapman et al., 2004, Rosonina and Blencowe, 2004). In the RNA polymerase II of *Plasmodium* species, however, the number of repeats is highly variable, with more repeats in primate-infecting species and even substantial variation in CTD lengths between *P. falciparum* isolates (Kishore et al., 2009). This may modulate transcription of variable surface antigens (Ukaegbu and Deitsch, 2015). The variability in this important protein suggests that sequences prone to repeat expansion and contraction may still play important roles despite being highly changeable.

The functions of these sequences remain to be elucidated. While KAHRP and LYMP play important roles in mediating cytoadhesion (Crabb et al., 1997, Proellocks et al., 2014), the fact that no significant change in this property was

observed upon deletion of four of the other proteins tested suggests they may play diverse roles in the cell (Maier et al., 2009). As the lysine-rich repeats of KAHRP and LYMP have been implicated in binding to spectrin (Pei et al., 2005) and band 3 (Oberli et al., 2016), respectively, the different motifs within each protein may all mediate binding to different cytoskeletal components (**Fig 6.2A**).

The function of proteins such as GARP and PF3D7_1102300, which do not contain any folded domains, is likely to be conferred by the repeats themselves. For example, the proteins may prevent the extension of spectrin filaments by linking spectrin repeats (**Fig 6.2B**). This could explain the slight decrease in rigidity observed upon deletion of the proteins (Maier et al., 2009). The expression of PF3D7_1102300 in gametocytes suggests it may be involved in the additional rigidification observed for early sexual-stage parasites as they are sequestered in the bone marrow (Tiburcio et al., 2012), while the slight upregulation seen during heat shock may serve a similar purpose to RESA in stabilising the cell during febrile episodes (Diez-Silva et al., 2012).

The identification of multiple peripheral targeting sequences in several proteins suggests they may be crosslinking different cytoskeletal components. For example, MESA may stabilise the interaction between protein 4.1 and other cytoskeletal components.   Other lysine-rich repeats may simply serve as targeting sequences for localising other important domains to the erythrocyte periphery, or for recruiting other proteins to this region (**Fig 6.2C**). The various PRESAN domain proteins may be involved in the trafficking or stabilisation of PfEMP1 on the erythrocyte surface, for example.

Electrostatic interactions between the basic lysine residues and a negatively-charged partner are likely responsible for the peripheral localisation of the repeating sequences. There are many potential candidates; the sequences may non-specifically bind to the negatively charged head groups of membrane lipids or to phosphorylation sites or acidic residues on the surface of cytoskeletal components. Triton extraction of GARP indicated that an interaction with the insoluble cytoskeleton is likely. However, many peripherally-located parasite proteins are highly acidic; RESA, Pf332 and PfEMP3 all contain glutamate-rich repeats, and interactions between the basic and acidic sequences of these proteins may form a network of proteins at the cell periphery (**Fig 6.2D**).
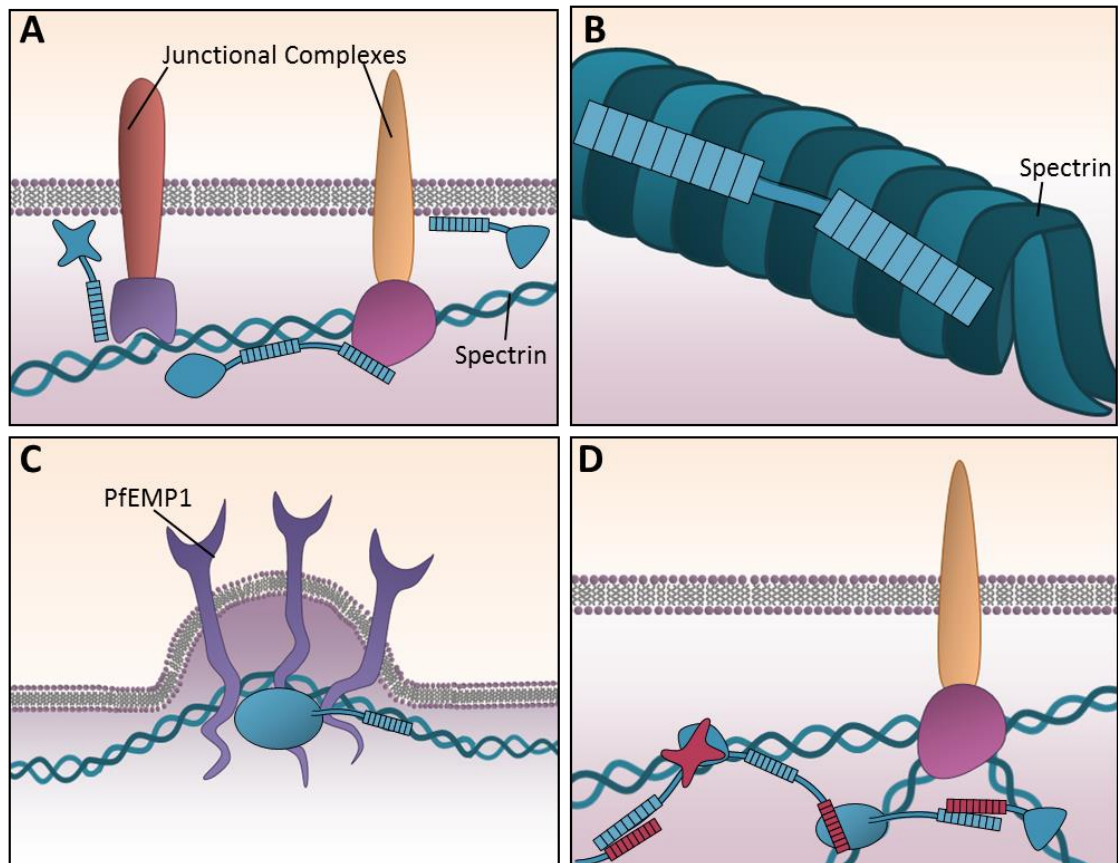
Figure 6.2 - Potential roles for lysine-rich repeats at the cell periphery. (A) Lysine-rich repeats may all bind different components of the erythrocyte cytoskeleton. (B) Some repeating sequences may bind spectrin repeats and modulate their flexibility or (C) others may also serve to target functional domains such as PHIST domains to the periphery. (D) Alternatively, a protein interaction network may be formed through the interaction of acidic and basic sequences.

The interaction between the acidic and basic sequences of Hyp12 indicates that such an association between oppositely charged proteins is possible. To our knowledge, such an interaction between two highly charged sequences predicted to be entirely disordered has not been seen previously. A signal from the parasite such as the release of calcium ions may trigger the dissociation of the sequences and 'switch on' peripheral localisation. Alternatively, the parasite may have expanded acidic sequences to 'switch off' the targeting function of lysine-rich repeats which do not confer an advantage. GARP, MESA, PHISTb protein PF3D7_1476200, and the PHISTb/c protein PF3D7_1201000 all contain highly acidic sequences which may also interact with their lysine-rich repeats. While the full-length GARP, MESA and PF3D7_1476200 protein were robustly

localised to the periphery, the additional targeting domains in these proteins may overcome the effects of blocking targeting by the lysine-rich sequence. The targeting efficiency of full-length PF3D7_1201000 was reduced relative to the lysine-rich fragment alone, however the poor export of the protein made further characterisation of this effect difficult. The cytosolic localisation of the acidic C-terminus of GARP suggests that this sequence does not interact with the peripherally-localised lysine-rich repeats, however.

The variation in peripheral targeting observed between different lysine-rich sequences likely indicates that peripheral targeting is mediated by a relatively weak interaction in most cases and that either decreasing the number of repeats or decreasing the affinity of each individual repeat unit may reduce targeting efficiency. To narrow down the constraints which are important for targeting, some of the characteristics of the sequences tested have been plotted against the fluorescence intensity at the periphery (Fig 6.3). There is no significant correlation between the length of the sequence fragment and the targeting efficiency, suggesting the composition of the repeat unit itself plays the most important role **(Fig 6.3A)**. Indeed, there is a significant correlation between the targeting efficiency and the isoelectric point of the sequence **(Fig 6.3B)** as well as the number of lysine residues **(Fig 6.3C)**. The most significant correlation to targeting efficiency is with the net amount of basic residues after subtracting the number of acidic residues **(Fig 6.3D)**. This suggests that negative residues generally decrease the targeting efficiency of a lysine-rich sequence; sequences containing more basic than acidic sequences display a stronger signal at the erythrocyte periphery and this increases with the increasing number of basic residues over acidic residues. The one outlier in this comparison is the full length GARP sequence; this may be due to the fact that multiple lysine-rich sequences are present within the protein.
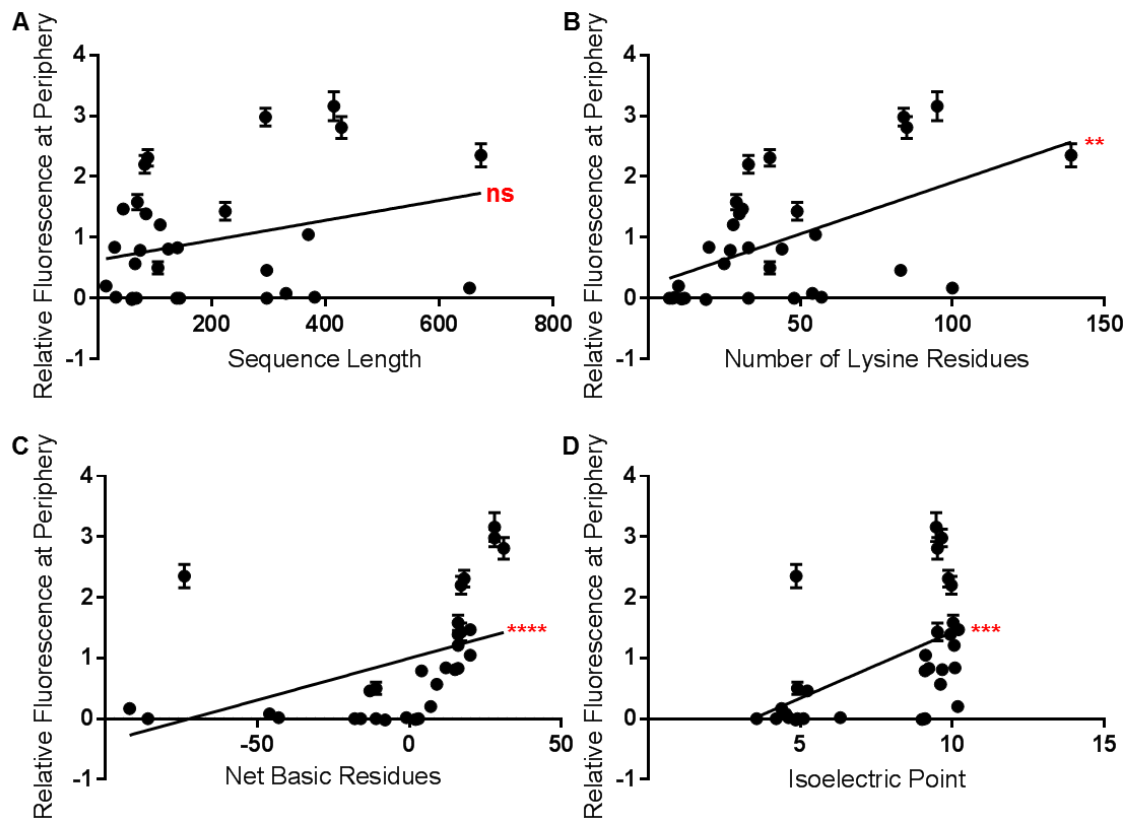
Figure 6.3 – Correlation between sequence characteristics and the efficiency of peripheral targeting. Error bars represent SEM. Significance of correlation (Spearman R) was calculated with Graphpad Prism and indicated by asterisks with a P-value < 0.0001 (**** - extremely significant), P-value < 0.0002 (*** - extremely significant), P-value < 0.0021 (** - very significant) and ns – not significant. The (A) Sequence length, (B) Number of lysine residues, (C) Net basic residues (basic – acidic), and (D) isoelectric point was plotted on the X-axis against the relative fluorescence intensity at the periphery on the Y-axis.

Few proteins in other species contain sequences as charged and repetitive as those of *P. falciparum*; however there are some examples of similar sequences with known functions. These are mainly found in structural proteins; the arginine and glutamate-rich protein trichohyalin is involved in crosslinking keratin in hair and nail fibres (Steinert et al., 2003), while the lysine and glutamate-rich axoneme-associated protein mst101 was shown to be an essential component of sperm flagella in fruit flies (Neesen et al., 1999). Covalent linkages are involved in fortifying the structures of both these proteins, with disulphide bonds linking mst101 filaments together (Neesen et al., 1999) while lysine and glutamine residues in trichohyalin are cross-linked by transglutaminase (Tarcsa et al., 1997). That the erythrocyte cytoskeleton is rigidified in a similar way has

been suggested previously (Adini et al., 2001), and such post-translational modifications may explain the smeared band observed for many of the proteins by western blot.

Notably, other charged repetitive proteins have been shown to bind cytoskeletal components. Both the human microtubule associated protein (MAP1B) and the *C. elegans* proteins CYLC-1 and CYLC-2 are rich in lysine residues and associated with microtubule assembly (Lacroix et al., 2016, Noble et al., 1989), while the sperm protein cylicin has been shown to bind to actin filaments in boars (Rousseaux-Preost et al., 2003). While the erythrocytes of non-mammalian vertebrates contain a tubulin skeleton surrounding the cytoskeleton, tubulin fibres are lost upon maturation of erythrocytes in humans, although a pool of soluble tubulin is believed to remain (Amaiden et al., 2011). Redistribution of tubulin may affect erythrocyte deformability (Amaiden et al., 2012). Actin is mainly present at the junctional complex of the erythrocyte cytoskeleton in short capped filaments. These are remodelled upon parasite invasion and can be seen in electron micrographs linking the cytoskeleton to Maurer's cleft structures (Cyrklaff et al., 2011). It is possible that the charged repeat proteins are involved in this process, or they may bind to other filamentous components of the cytoskeleton such as spectrin.

The localisation of so many charged proteins to the erythrocyte periphery is likely to drastically modify the electrophysical environment surrounding the cytoskeleton. Knob structures have been shown to be positively charged (Aikawa et al., 1996), which may contribute to their interaction with receptors. Other basic residues may confer a similar localisation to lysine-rich repeats; a poly-histidine sequence in KAHRP also targets the erythrocyte periphery (Waller et al., 1999), and the histidine rich protein II may also interact with cytoskeletal components, potentially in a pH-dependent manner (Akompong et al., 2002, Benedetti et al., 2003). Arginine residues are under-represented in the AT-rich *P. falciparum* genome, and no arginine-rich repeats were found in our search with the exception of the protein PF3D7_1102300 which was also enriched for lysine. Some arginine-rich sequences were however found in the *P. vivax, P knowlesi* and *P. cynomolgi* genomes, which do not contain such an extreme codon bias.

In addition to the lysine-rich repeat sequences, we also identify a conserved folded domain which may target the erythrocyte periphery in KAHRP and PfEMP3-like proteins. The localisation of the EKAL domain in puncta at the periphery suggests it may be a component of knob structures, although it is unclear whether it is capable of targeting the periphery in the absence of pre-formed knobs. Both KAHRP and PfEMP3 contain other domains capable of localising to the periphery (Knuepfer et al., 2005, Waller et al., 2007a, Weng et al., 2014, Oh et al., 2000, Rug et al., 2006, Waller et al., 1999, Pei et al., 2005) and most other EKAL domain proteins contain lysine rich repeats predicted to have a targeting function. It is therefore possible that this domain serves another purpose, perhaps self-associating to cluster surface antigens on the surface. Further work will be required to elucidate the function of this domain, and establish its role in the other primate-infecting parasite species.

6.1 Conclusions and Further Work

The aim of this thesis was to identify new protein domains capable of targeting the erythrocyte periphery, where important virulence-related changes are known to occur. We have successfully identified two modules capable of localising to this region: the tandemly repeated lysine-rich sequences and the EKAL domain. The functional lysine-rich repeats described here are not homologous in the traditional sense as the repeated motifs vary in length and composition and likely originated from individual repeat expansion events. Nevertheless, the criteria used for our algorithm resulted in the identification of several peripherally-targeting sequences, suggesting that the common features of these sequences are sufficient to confer a targeting function. While the lysine-rich repeats are predicted to be intrinsically disordered, the EKAL domain is predicted to form a conserved folded structure; differences in phenotype between KAHRP and PfEMP3 deletions suggest the function of the domain may have diverged, however.

In addition to the identification of novel targeting modules, we have also gained an insight into the evolution of the parasite. The adaptability of *P. falciparum* is considered to be one of the main hurdles to designing new antimalarial

treatments, with resistance is an ever-growing problem. The ability to from new functional targeting domains at much greater rates than through point mutations likely accelerated the evolution of *Plasmodium* species and contributed to the spread of malaria to most animal species on the planet. Rapid expansion and contraction of established peripheral-targeting repeating sequences may also allow additional flexibility in modulating the properties of the host erythrocyte in different hosts. While repetitive sequences are generally highly divergent between *Plasmodium* species, the identification of the conserved EKAL domain in primate-infecting species may indicate similar mechanisms are used for surface receptor presentation and cytoadhesion in these species. A closer look at the evolutionary differences between species may lead to a greater understanding of the causes of severe malaria.

Many of the previously-uncharacterised lysine-rich repeat proteins may play important roles at the erythrocyte periphery, and our identification of additional targeting modules in proteins with known cytoskeleton-binding motifs may elucidate their precise roles at this important region. Some proteins present interesting avenues for further research; the entirely charged protein PF3D7_1102300 may play a role in gametocyte development while the Hyp12 protein appears to contain a functional lysine-rich sequence which is 'switched off' and may require a trigger to become functional.

Identifying the binding partners of the lysine-rich repeats will be essential for understanding their function. We have made some progress in designing a protein-binding assay to identify cytoskeletal components which may associate with our proteins of interest. This is based on the AVEXIS assay which was used for identifying surface receptors involved in erythrocyte binding upon invasion, and requires the *in vitro* recombinant expression of biotinylated cytoskeleton baits and reporter-tagged prey proteins. This system has been tested with known binding partners within the erythrocyte cytoskeleton, and has also been used to test for novel interactions between the PfEMP1 surface antigen and the PRESAN domain of different proteins. Once the system has been optimised this new assay will be used to test for binding interactions between lysine-rich repeat proteins and the cytoskeleton.

As some interactions may involve additional parasite proteins or post-translational modification, a technique for identifying binding partners *in vivo* is

also being explored. BioID allows for the biotinylation of proteins in close proximity to the protein of interest by attaching it to a promiscuous biotin ligase. Biotinylated proteins are then isolated and identified by mass spectrometry. Both GARP and KAHRP fragments have been tagged and preliminary work suggests the biotin ligase is capable of biotinylating erythrocyte components within the host cell. Both these techniques are currently in development and we hope they will soon be able to further elucidate the roles of the peripheral-targeting proteins identified here.

In addition to identifying the binding sites of the proteins identified, a phenotypic analysis will be necessary to elucidate their roles within the cell. Using new conditional genome engineering techniques uncharacterised genes may be deleted and their functions analysed by testing erythrocyte properties such as rigidity and cytoadhesion. While the deletion of multiple genes may be required for a clear phenotype to appear in some cases, the removal of the repetitive sequences from proteins with known functions such as KAHRP and LYMP may demonstrate the importance of these sequences for protein function. Simulating *in vivo* conditions may be necessary to observe a function for some proteins, as they may behave differently under heat shock or at different pH levels, or at different life stages such as during sexual stage development.

The discovery of novel domains targeting the erythrocyte periphery demonstrates that we still do not possess the full picture where host cell modification is concerned. It is likely that many other proteins play a role in this region, and establishing how they interact together will require both a holistic analysis of the cytoskeletal network as well as a more detailed look at the structure and function of individual proteins. We hope that this work contributes to this goal, and that this understanding will eventually lead to new treatments for preventing the most severe manifestations of malaria.

Appendix

Construct	Fold-difference in Fluorescence at Membrane relative to Cytosol	Standard Deviation	P-value (Significance)
Full-Length GARP	3.27	0.86	<0.0001 (****)
$GARP_{119-163}$	2.39	0.43	<0.0001 (****)
$GARP_{253-340}$	3.23	0.61	<0.0001 (****)
$GARP_{372-446}$	1.71	0.29	<0.0001 (****)
$GARP_{524-673}$	0.92	0.07	0.9988 (ns)
$GARP_{50-118}$	0.92	0.06	0.9548 (ns)
GARP (+ promoter)	2.90	0.72	<0.0001 (****)
$GARP_{119-163}$ (+linker)	2.29	0.60	<0.0001 (****)
$GARP_{134-163}$	1.76	0.29	<0.0001 (****)
$GARP_{149-163}$	1.12	0.20	0.2323 (ns)
$GARP_{372-446}$ (+linker)	1.76	0.37	<0.0001 (****)
P. reichenowi $GARP_{71-130}$	0.90	0.06	0.9181 (ns)
P. gaboni $GARP_{381-412}$	0.94	0.07	0.9248 (ns)
$PF3D7\_1102300_{121-415}$	3.90	0.65	<0.0001 (****)
$GEXP12_{231-370}$	1.75	0.35	<0.0001 (****)
$LYMP_{419-528}$	2.13	0.34	<0.0001 (****)
$PF3D7\_1476200_{443-512}$	2.50	0.57	<0.0001 (****)
$PF3D7\_0402000_{305-428}$	1.73	0.40	<0.0001 (****)
$PF3D7\_1201000_{292-397}$	1.42	0.45	0.0016 (**)
$MESA_{850-1147}$	1.38	0.20	0.0062 (**)
$KAHRP_{363-428}$	1.49	0.21	0.0007 (***)
$PF3D7\_0114200_{97-240}$	0.92	0.04	0.9845 (ns)
$PF3D7\_1149100.1_{120-416}$	0.92	0.05	0.9917 (ns)
$Hyp12_{297-381}$	2.31	0.33	<0.0001 (****)
$PKNH\_1325700_{303-445}$	3.12	0.67	<0.0001 (****)
PF3D7_1102300 (FL)	4.08	1.06	<0.0001 (****)
GEXP12 (FL)	1.97	0.44	<0.0001 (****)
PF3D7_0402000 (FL)	3.73	0.80	<0.0001 (****)
PF3D7_1201000 (FL)	1.09	0.20	0.2920 (ns)
$PF3D7\_0114200_{1-302}$	0.99	0.08	0.6517 (ns)
PF3D7_1149100.1 (FL)	1.05	0.17	0.4333 (ns)
Hyp12 (FL)	0.94	0.08	0.9064 (ns)
$Hyp12_{51-381}$	1.00	0.10	1.82403 (ns)
$Hyp12_{158-381}$	0.65	2.35	<0.0001 (****)
PF3D7_1102300 (+	4.01	1.82	<0.0001 (****)

Appendix Table 1 - Quantification and statistical analysis of GFP fluorescence at the periphery of infected erythrocytes. See materials and methods section 2.3.7 for details.

Gene Name	Location of Potential Error	Nature of Potential Error
PFCD01_GARP	360	Frame shift (Insertion)
	766	Frame shift (Insertion)
PFGN01_GARP	781	Frame shift (Insertion)
PFML01_KAHRP	1020	Frame shift (Insertion)
PFML01_MESA	3030	Frame shift (Insertion)
PFCD01_MESA	1190	Frame shift (Insertion)
PFGN01_PF3D7_1102300	978	Frame shift (Insertion)
PFSD01_PF3D7_1102300	1013	Frame shift (Insertion)
PFGA01_PF3D7_0402000	1082	Frame shift (Deletion)
PFML01_PF3D7_1201000	995	Frame shift (Insertion)
PFTG01_PF3D7_1201000	995	Frame shift (Insertion)

Appendix Table 2 – Annotation of introns and location of potential frameshift mutations in PACBIO sequencing. (A) Sequences containing frameshift mutations in PACBIO sequences were restored. Where multiple sequences were found (for PfML01_KAHRP, PfML01_MESA, PfTG01_MESA and PfTG01_1476200), the sequence containing the fewest frameshift mutations was used (either one or zero mutations).

Gene Name	Location of Potential Error	Nature of Potential Error
PI C922_04319	110-295	Un-annotated intron
	365	Point mutation (Stop codon)
	749	Point mutation (Stop codon)
	1244	Point mutation (Stop codon)
PI C922_02878	112-243	Un-annotated intron
	74	Frame shift (Deletion)
	1003	Point mutation (Stop codon)
	1672	Point mutation (Stop codon)
PFR AK88_04565	112-256	Un-annotated intron
	996	Assembly gap (skipped)
	1625	Assembly gap (skipped)
	2737	Assembly gap (incomplete sequence)
PCYB_001100	109-284	Un-annotated intron
	3	Point mutation in start codon
	109-284	Un-annotated intron
	639	Point mutation (Stop codon)
	1217	Assembly gap (incomplete sequence)
PCYB_127900	97-246	Un-annotated intron
	46	Frame shift (deletion)
PCYB_042840	112-242	Un-annotated intron
	37	Frame shift (deletion)
	1080	Frame shift (deletion)
	1716-2157	Reverse complement
	2129	Frame shift (insertion)

Appendix Table 3 – Annotation of introns and location of potential frameshift mutations in EKAL-domain proteins. Introns were annotated manually for all sequences indicated. For protein PFR AK88_04565, apparent assembly gaps were present in the protein sequence, which were skipped in accordance to the protein annotation within the European Nucleotide Archive (ENA). Assembly gaps within PFR AK 88_04565 and PCYB_001100 result in a truncated sequence with no stop codon. A section of gene PCYB_042840 appears to be reverse complemented within the assembled sequence, which was modified in-frame with flanking regions. All mutations may be caused by sequencing errors or may be true mutations.

Construct Name	Primer Name	Primer Sequence
GARP$_{1-174}$	HD1_GARP_174_R	TCCTTTAGATCTTCCTGAATATAAAGACATCACTTCAC
GARP175-523	HD4_GARP_HT_175_R (OL)	GGTTTATGTTGACCCGTTTTTGTTTTTTTTAAATCCTGTT
GARP$_{524-673}$	HD5_GARP_HT_524_R (OL)	TTTTTCTTATGATTATCTCTTGTTTTTTTTAAATCCTGTT
REX3	HD6_HT_F	TCCTTTCCTAGGTAAAAAATGCAAACCCGTAAATATAA
GARP$_{524-673}$	HD7_GARP_HT_175_F (OL)	AACAGGATTTAAAAAAAACAAAAACGGGTCAACATAAACC
GARP$_{524-673}$	HD8_GARP_HT_524_F (OL)	AACAGGATTTAAAAAAAACAAGAGATAATCATAAGAAAAAAATGGC
Multiple GARP	AO828GARP_F	CTAAGCCCTAGGTAAAAAATGAATGTGCTATTTCTTTC
Multiple GARP	AO829GARP_R	GTTTAGAGATCTTCCTGAAATTTTGGCATTTTTTCTCA
GARP$_{1-118}$	HD9_GARP6_118_R	TCCTTTGACGTCTCCTGAATCTACAGAGTTTTCATTAT
GARP$_{119-163}$	HD23_GARP14_119_F	ATAATGAAAACTCTCTCGAGAAAAAAAAAGATAAAAAAGAAAAAAAGC
GARP$_{119-163}$	HD26_GARP14_163_R	ACTTCACTGTTTTCGCTAGCTTTTTTATCTTTTTTATG
GARP$_{253-340}$	HD27_GARP16_253_F	CTAAGCCTCGAGCAAGAAGAAATGTTGAAAACACT
GARP$_{253-340}$	HD28_GARP16_340_R	TCCTTTGCTAGCTTCATTTTCCTTATCATGTTTTTTC
GARP$_{372-446}$	HD29_GARP17_372_F	TGACAGACGTAACTCTCGAGGAAGAACACAAAGAAGGAGA
GARP$_{372-446}$	HD30_GARP17_446_R	TCCTTTGCTAGCAACATTTTTAACTACGTGTTTTTTTAC
PF3D7_1102300 (FL)	HD53_PF3D7_1102300_F	CTAAGCCCTAGGTAAAAAATGAAAAATTGTAGGAATATATTAATCG
PF3D7_1102300 (FL)	HD54_PF3D7_1102300_R	TCCTTTGACGTCTCCTGAGACATTTTCCATTTTCATTTTTTC
PF3D7_1149100.1 (FL)	HD55_PF3D7_1149100.1_F	CTAAGCCCTAGGTAAAAAATGATGAAGCTTTTTAATAATTTTACATG
PF3D7_1149100.1 (FL)	HD56_PF3D7_1149100.1_R	TCCTTTGACGTCTCCTGATTGAGCTCTTCTTAATTTATCATTTG
Hyp12 (FL)	HD57_PF3D7_1301400_F	CTAAGCCCTAGGTAAAAAATGTTTTCTTCTACAACAAATATATTTTTTTC
Hyp12 (FL)	HD58_PF3D7_1301400_R	TCCTTTGACGTCTCCTGAGTGTCTTAAGCTAAGAAGCTTATTTCTTC
PF3D7_0402000 1-302	HD59_PF3D7_0114200_1-302_F	CTAAGCCCTAGGTAAAAAATGAAACACCTTCTGTCACTTTTTAAG
PF3D7_0402000 1-302	HD60_PF3D7_0114200_1-302_R	TCCTTTGACGTCTCCTGATACAAATTCTGGTGAAAAAATACCTC
GEXP12 (FL)	HD61_PF3D7_1148700_F	CTAAGCCCTAGGTAAAAAATGAAGCAACATATAACCTTTAAGAAG

GEXP12 (FL)	HD62_PF3D7_1148700_R	TCCTTTGACGTCTCCTGATTTGCTTGGTAATCTTTTTTTTTC
LYMP$_{419-528}$	HD63_LYMP_419-end_F	GGCGCTTCTGACGTCTTAACTGAAAATCAGATAAATGAAG
LYMP$_{419-528}$	HD64_LYMP_linker_F	CTAAGCCTCGAGTCAGGATCTGGTACGGGCGCTTCTGACGTC
LYMP$_{419-528}$	HD65_LYMP_419-end_R	TCCTTTGCTAGCATTTTTTTTTTTATTTTCTTTTCC
PF3D7_0402000$_{305-428}$	HD66_PF3D7_0402000_305-428_F	CTAAGCGACGTCGATACCAGTAAGAAAGATATGAAG
PF3D7_0402000$_{305-428}$	HD67_PF3D7_0402000_305-428_R	TCCTTTGCTAGCATTGTTTTTTTTTAATTCTTTTTTC
KAHRP$_{540-600}$	HD70_KAHRP(3'repeats)_540-600_F	CTAAGCGACGTCACTAAGGGAGCAACTAAAGAAG
KAHRP$_{540-600}$	HD71_KAHRP(3'repeats)_540-600_R	TGTAGTTGAGCTAGCAGTAGTACTTGCTCCTTTAG
MESA$_{850-1147}$	HD72_MESA(lysrepeat)_851-1146_F	CTAAGCGACGTCAAAGATAAAGTGACAAAACAGG
MESA$_{850-1147}$	HD73_MESA(lysrepeat)_851-1146_R	TCCTTTGCTAGCGTTTACTCTTTTGTTAACTTCTTTG
GARP$_{119-163}$ (+linker)	HD80_GARP39_119-163_F	ATAATGAAAACTCTGACGTCAAAAAAAAAGATAAAAAAGAAAAAAAGC
GARP134-163 and GARP149-163	HD81_GARP40_134–163_F	CTAAGCGGATCCTAGGATGCAAAC
GARP$_{134-163}$	HD82_GARP40_134–163_R1	TTTTTTCTTTTTTTTCGACGTCAGAAGCGCC
GARP$_{134-163}$	HD83_GARP40_134–163_R2 (OL)	TTTTCTTTTTTATCTTTTTTTTTCCTTTTTATCTTTTTTTTTCTTTTTTTTTCGACG
GARP$_{134-163}$	HD84_GARP40_134–163_R3 (OL)	TTTATGTTTTTTTTCTTTTTTATGCTTTTTTTCTTTTTTATCTTTTTTTTCC
GARP134-163 and GARP149-163	HD85_GARP40_134–163_R4	TCCTTTGCTAGCTTTTTTATCTTTTTTATGTTTTTTTTTCTTTTTTATGC
GARP$_{149-163}$	HD86_GARP41_149-163 _R1	TTTATGTTTTTTTTCTTTTTTATGCTTTTTTTTCGACGTCAGAAGCGCC
PF3D7_1476200$_{443-512}$	HD105_PF3D7_1476200_443-512_F	CTAAGCGACGTCGATCAAGCAAGCAAGAAAAAAAAAAAG
PF3D7_1476200$_{443-512}$	HD106_PF3D7_1476200_443-	TCCTTTGCTAGCTTTCCAAGTTTTAATTTTATTTTTAGTTTC

	512_R	
PF3D7_1102300$_{121-415}$	HD115_PF3D7_1102300_121-415_F	CTAAGCGACGTCGATAGAGAAAAAGATGATCAAAACG
PF3D7_1102300$_{121-415}$	HD116_PF3D7_1102300_121-415_R	TCCTTTGCTAGCGACATTTTCCATTTTCATTTTTTC
PKNH_1325700	HD134_PkKahrp_full_R	GTACCAGATCCTGAGACGTCTATTTTATTTTTCTTAGATTTGTCATTTTT
GEXP12$_{231-370}$	HD143_PF3D7_1148700_C-term_F	CTAAGCGACGTCTCTGATAAAATATGCAAAGAAAATAAC
GEXP12$_{231-370}$	HD144_PF3D7_1148700_C-term_R	TCCTTTGCTAGCTTTGCTTGGTAATCTTTTTTTTTTC
PKNH_1325700$_{303-445}$	HD145_PkKahrp_C-term_F	CTAAGCGACGTCCCCACCGTAGGTCAATACTC
PF3D7_1201000$_{292-397}$	HD146_PF3D7_1201000_F	CTAAGCGACGTCAAAAAAGAAACGGAAACGG
PF3D7_1201000$_{292-397}$	HD147_PF3D7_1201000_R	TCCTTTGCTAGCGCTCTGAGATTATTTTTAATATTGCTTAC
PKNH_1325700	HD148_PkKahrp_full_F (OL)	GGTTCCTAAATTGTTCTAATTACGAAAAATGTGATAGGAGGAGTG
PKNH_1325700	HD149_PkKahrp_full_R (OL)	CACTCCTCCTATCACATTTTTCGTAATTAGAACAATTTAGGAACC
PKNH_1325700$_{75-232}$	HD150_PkKahrp_cons_F	CTAAGCGACGTCCTTAGAACCCGTATAGTAAAGGAAGTTACC
PKNH_1325700$_{75-232}$	HD151_PkKahrp_cons_R	TCCTTTGCTAGCTCTTGGACTATCTGGTCCTCC
PKNH_1325700	HD152_PkKahrp_full_F2	CTAAGCCCTAGGTAAAAAATGGCTATTCTTAAGGGATACAG
KAHRP$_{363-428}$	HD165_KAHRP5(2)_F	CTAAGCGACGTCTCTAAAAAGCATGAAGGAAATG
KAHRP$_{363-428}$	HD166_KAHRP5(2)_R1 (OL)	TCCTCGTTATCCTTATGCTTCTTAGATTTTTTCTTTTCACAATCGTGG
KAHRP$_{363-428}$	HD167_KAHRP5(2)_R2	TCCTTTGCTAGCTGCGTCCTCGTTATCCTTATGCTTCTTAG
GARP$_{50-118}$	HD223_GARP50-118_F	CTAAGCCTCGAGAACGAAACCGAATTAGAAAAAATAAAG
GARP$_{50-118}$	HD224_GARP50-118_R	TCCTTTGCTAGCATCTACAGAGTTTTCATTATCTTTATCATTAAC
PF3D7_0114200$_{97-240}$	HD225_PF3D7_0114200reps_F	CTAAGCGACGTCAATAATATGGAAAAGGAGATGGG
PF3D7_0114200$_{97-240}$	HD226_PF3D7_0114200reps_R	TCCTTTGCTAGCTTTCAATATATAGTCCATATATTCTTTCG
PF3D7_1149100.1$_{120-416}$	HD227_PF3D7_1149100.1reps_F	CTAAGCGACGTCATCCAAAATGAAGATTCTGAAAAAG
PF3D7_1149100.1$_{120-416}$	HD228_PF3D7_1149100.1reps_	TCCTTTGCTAGCATCATTCGTTGAATTTTTTGTAG

	R	
Hyp12$_{297-381}$	HD229_PF3D7_1301400reps_F	CTAAGCGACGTCAGAGAAAAAAAAAAGAAAGAAAAAC
Hyp12$_{297-381}$	HD230_PF3D7_1301400reps_R	TCCTTTGCTAGCGTGTCTTAAGCTAAGAAGCTTATTTC
GARP (+ promoter)	HD242_GARPpromo1kb_F	CAAAAAGCTAAGCTATACATAGAAATATAAATAATTCTAATCCCATAG
PF3D7_1102300 (+ promoter)	HD244_FOGpromo1kb_F	CAAAAAGCTAAGCCTCTATACATGATAATATGAAGATCATTTC
GARP$_{535-673}$	HD254_ GARPCterm_F	TAAAAACTCGAGGAGGAAGCTGAACTTCAAAAAC
GARP$_{535-673}$	HD255_ GARPCterm_R	TAAAAAGCTAGCAATTTTGGCATTTTTTCTCA
Hyp12$_{51-381}$	HD256_ FulllengthHyp12_F	TAAAAAGACGTCGCAGAACCATCTTCACATGGCTC
Hyp12$_{158-381}$	HD257_ WOacidicHyp12_F	TAAAAAGACGTCTTAATACCTATTCTACCAAATAGCTCG
P. reichenowi GARP$_{71-130}$	HD258_ ReichafterGFP_F	GAAACGTTGGACGTCGATGAAAAGATGAAAC
P. reichenowi GARP$_{71-130}$	HD259_ ReichafterGFP_R	TAAAAAGCTAGCTTTTTTTTTCTTTTTTCTTTTTTTCCAC
P. gaboni GARP$_{381-412}$	HD260_GaboniGARP_R1 (OL)	TTTTATGTTTTCCTTCCTCTTTCTTTTCTTCTGTTTTATGTTCTTGGACGTCAGAAGCGC
P. gaboni GARP$_{381-412}$	HD261_GaboniGARP_R2 (OL)	ATGTGTGTTTCTTCTTTTTCTTTTTCTTTTGTATCTATTTTATGTTTTCCTTCCTCTTTC
P. gaboni GARP$_{381-412}$	HD262_GaboniGARP_R3	TAAAAAGCTAGCAAAGTTTTTAAGTATGTGTGTTTCTTCTTTTTCTTTTTC
GARP$_{372-446}$ (+linker)	HD263_ *Falciparum* GARPlink_F	TGACAGACGTAACTGACGTCGAAGAACACAAAGAAGGAGA

Appendix Table 4 – Primer sequences for all constructs used in this project. The protein name and cloned fragment boundaries are displayed in the first column. Most primer names also contain details of the gene. F and R denote forward and reverse primers, respectively, and OL designates primers used for overlap PCR. Forward and reverse primers are mainly adjacent but may be separate where new primers were required. Some primers may be used for multiple constructs. Restriction sites are coloured red.
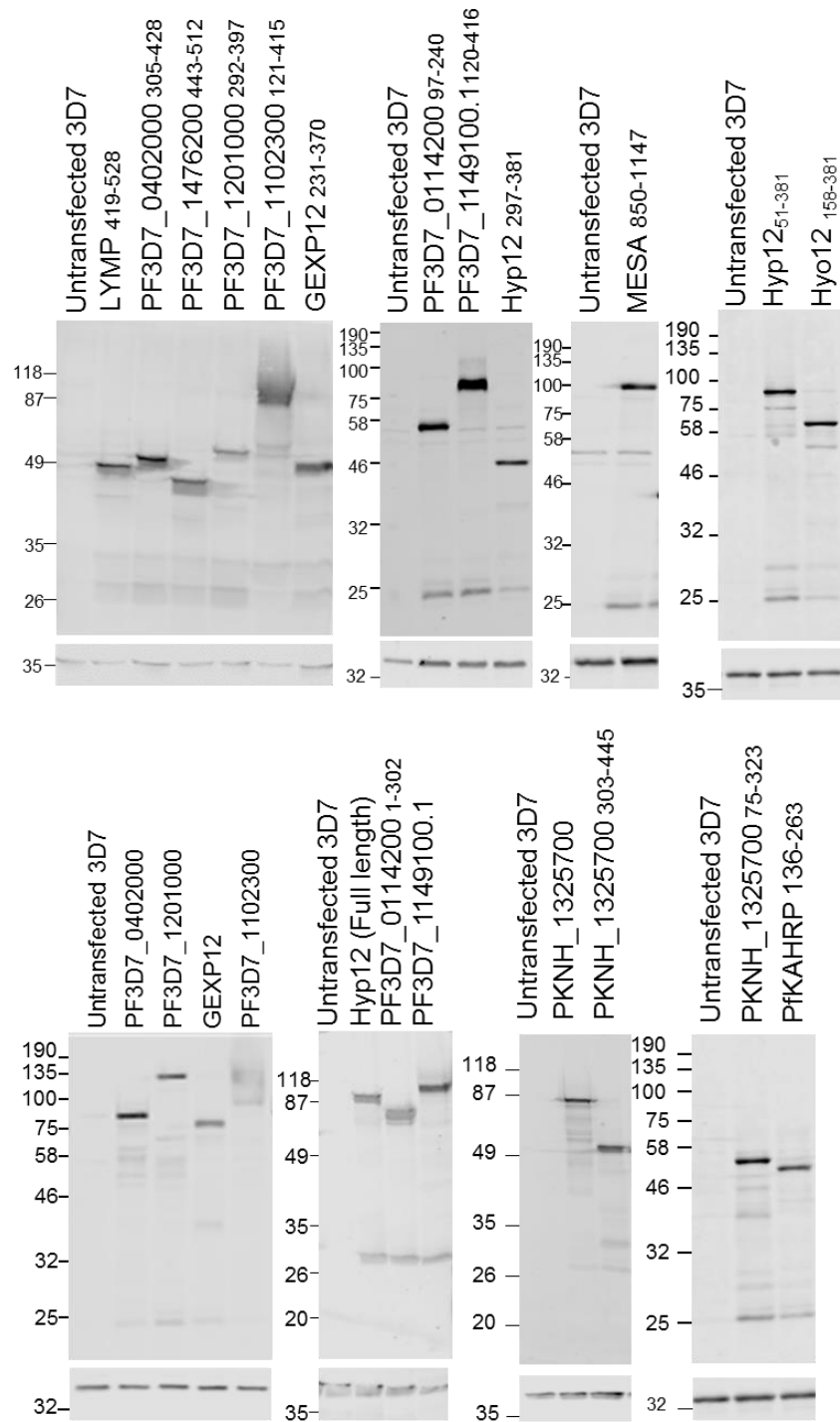
Appendix Figure 1 – Western blots of all GFP-tagged constructs not shown in main text. Anti-GFP western blot (top panel) of transfected parasites expressing GFP-tagged GARP constructs. Approximately 2 x 10^6 percoll-purified schizonts were loaded per lane, with anti-HAP used to confirm equal loading (lower panel).

Appendix Text 1 - 'exportfind.pl' perl script - Subroutines are included below. Input sequences were generated by PlasmoDB as well as MPEX and SignalP. Comments are in red.

```perl
#!usr/bin/perl

# File:  exportfind.pl
# Version 1
# This script takes output from the MPEX JAVA program and the SignalP
web server.
# Output must be copied from the JAVA applet and web servers and saved
as a .txt file.
# The script predicts proteins which are exported into the host
erythrocyte by Plasmodium parasites.
# it extracts protein sequences from a FASTA file (e.g. malaria
proteins downloaded from Plasmodb.com).
# It counts how many transmembrane (TM) domains there are from an MPEX
output file for the same gene set.
# If there are less than 4 TM domains, it checks the SignalP output
for a signal sequence (SS).
# If no SS are present it looks for an N-terminal TM domain in the
MPEX file.
# If either a SS or N-terminal TM domain is present, a regular
expression checks for a PEXEL motif by looking for an RxL motif
# within 100aa of the SS cleavage site or TM.
# Proteins containing a PEXEL are considered likely to be exported and
printed out.
# Both of these are printed out with the name of the protein and the
length of the repeat sequence.
# The number of sequences selected is printed at the end (multiple
sequences may be from the same protein).
# If scoring is turned on, more positively charged proteins are
printed first (highest score first).

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;
use get_mpex;
use get_sigp;
use Getopt::Long;

my $count = 0;
my $genename;
my %genes;
my $length = 30;
my $score = '';
my $help = '';
my $name;
my @data;
my $repseq;
my $repscore;
my $fullseq;
my $replength;
my $psim;
my $seqlength;
my %result;
my $cut;
my $start;
my $end;
my $pexel;
```

```perl
# Open list of proteins from PlasmoDB, as well as the output from MPEX
and SignalP for the same gene set.

my $file = 'allproteins.txt';
my $outfile = 'allexported.txt';

my $mpexinfile = 'mpexresults.txt';
my $psiginfile = 'sigpresults.txt';

open(INFILE, $file) or die "can not open $file:$!\n";
open(OUT, ">>$outfile") or die "can not open $outfile:$! \n";

# Stores all gene sequences within a hash.
while(<INFILE>) {
    chomp;
    if (/^>(\w{2,7}_\d{7})\s.+/) {
        $genename = $1;
    }
    elsif (/^([AILMFWYVSTNQCUGPRHKDE]+)/i) {
        $genes{$genename} .= $1;
    }
}
# Loops through each protein sequence, calling subroutines to check
for TM and SS.
# get_mpex checks the number and position of all TM domains. Proteins
with <4 TM are selected.
# get_sigP checks for the presence of a signal sequence; if present
the cut site is noted.
# If no SS is present, it checks whether the first TM domain is within
the first 100 aa.
foreach my $protein ( keys %genes ) {
    if (length($genes{$protein}) >100 ) {
        my $mpexref = get_mpex($protein, $mpexinfile);
        if (${$mpexref}[0] < 4) {
            my ($sigpos, $sigyes) = get_sigp($protein, $psiginfile);
            if ($sigyes =~ /Y/) {
                $cut = $sigpos;
            }
            elsif ($sigyes =~ /N/) {
                if (${$mpexref}[0] > 0) {
                    if (${$mpexref}[1] < 100) {
                        $cut = ${$mpexref}[1];
                    }
                }
            }
        }
# If a SS or TM was found, a RxL motif is searched for within 100aa of
the cut site/TM domain.
# Proteins containing the motif are printed out with their protein
sequence.
        if (defined $cut) {
            $start = $cut + 5;
            $end = $cut + 100;
            print "$start, $end\n";
            if ($genes{$protein} =~ /^\w{${start},${end}}R\wL.+/ ) {
                print "pex\n";
                print "$protein\n$genes{$protein}\n";
                print OUT "$protein\n$genes{$protein}\n";
            }
        }
    }
# Parameters are undefined for the next protein.
    undef $cut;
```

```perl
        undef $start;
        undef $end;
}
```

Subroutine get_mpex

```perl
#!/usr/bin/perl

# File: get_mpex.pm
# Version 1
# A subroutine to analyse the output of the transmembrane-domain
prediction program MPEX.
# Genenames are passed from the exportfind.pl program and are searched
for in the output file.
# If a TM domain is found, the number of TM domains present and the
position of the 1st is returned to exportfind.pl

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;

sub get_mpex
{
    my ($ID, $file) = @_;
    open(INFILE, $file) or die "can not open $file:$!\n";

    my $genenamem;
    my %genesm;
    my $match;

    # The number of TM domains and the position of the first TM domain
are recorded for each gene.

while(<INFILE>) {
    chomp;
    if (/^Results.+\:\s(\w{2,8}\_\d{7}.+)/) {
        $genenamem = $1;
    }
    if ($genenamem =~ /$ID/) {
        $match = "yes";
    }
    elsif ($genenamem !~ /$ID/) {
        $match = "no";
    }
    if ($match =~ /yes/ &&
/^Number\sof\stranslocon\sTM\spredicted\ssegments:\s(\d+)/i) {
        push @{$genesm{$genenamem}}, $1;
        }
    if ($match =~ /yes/ && /^(\d+)\-(\d+)\s\(.+/i) {
            push @{$genesm{$genenamem}}, $2;
        }
}

    for my $valuem (keys %genesm) {
        return \@{$genesm{$valuem}};
    }
}
1;
```

Subroutine get_sigp

```perl
#!/usr/bin/perl

# File: get_sigp.pm
# Version 1
# A subroutine to establish whether a gene contains a signal sequence.
# Genenames are passed from exportfind.pl and the output from SignalP
web server is searched.
# The presence and position of a signal sequence is returned to
exportfind.pl.

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;

sub get_sigp
{
    my ($ID, $file) = @_;
    open(INFILE, $file) or die "can not open $file:$!\n";

    my $genenamep;
    my $position;
    my $yesno;
    my $geneID2;
    my @split2;

# The signalP output file is searched for the gene.
# The position and presence of a SS is extracted from the same line
and returned to exportfind.pl.

while(<INFILE>) {
    chomp;
    if
(/^(\w+)\s+0\.\d+\s+\d+\s+0\.\d+\s+(\d+)\s+0\.\d+\s+\d+\s+0\.\d+\s+0\.
\d+\s+([YN])\s+0\.\d+\s+SignalP.+/i) {
                $genenamep = $1;
                $position = $2;
                $yesno = $3;
            if ($genenamep eq $ID) {
            return ($position, $yesno);
            last;
        }
    }
}

}

1;
```

Appendix Text 2 - 'chargesearch.pl' perl script. Input was generated by 'exportfind.pl'. Comments are in red.

```perl
#!usr/bin/perl

# File:  chargesearch.pl
# Version 1
# A script for detecting regions within a protein which are rich in
certain amino acids (e.g. highly charged regions).
# It begins by extracting protein sequences from a FASTA file (e.g.
malaria proteins downloaded from Plasmodb.com).
# A sliding window of a certain length (default = 20 aa) scans each
sequence in turn.
# The number of certain predefined residues (e.g. lysine, K) within
the window are counted by regular expression.
# A threshold can be set for the desired number of a certain residue
in a window (e.g. 4 lysines out of 20 aa).
# If above this threshold the sequence within the window is
temporarily stored, and the window moves forward 1 aa.
# Each subsequent window which is also above the threshold adds to the
growing sequence as it iterates.
# Once a window does not contain the requisite number of residues, the
current sequence is then considered complete.
# If the complete sequence is above a certain length (e.g. 30 aa), the
sequence is stored in a hash for that protein.
# The sliding window will continue to scan the sequence, ignoring any
windows which do not pass the threshold.
# Any number of sequences can be picked out of a protein, all stored
in an array within the protein's hash key.
# These are printed out in FASTA format with each of the sequences
within the same protein numbered.
# This output can then be used for further processing (e.g. T-Reks for
finding tandem repeats within the sequences).

# Getopt::Long is used to allow users to change the values for all
parameters through the command line.
# Options for Getopt::Long: (single letters can be used for each
option, e.g. -w instead of --window)

#   infile = select input fasta file (filename)
#   outfile = select name for output file (filename)
#   window = modify window length (integer)
#   residue = type of residues to be counted (string)
#   count = threshold of residues within the window in order for that
window to be added to the sequence (integer)
#   length = length of the sequence required for it to be stored and
printed (integer)

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;
use Getopt::Long;

my $genename;
my %genes;
my %lyscount;
my $regseq;
my %region;
my @res;
my $parameter;
my $score = 0;
```

```perl
my $help = '';

# Default options are set for window size, residue count and length of
sequence.

my $winsize = 20;
my $count = 4;
my $lengthcount = 30;

my $file = 'allexported.txt';
my $outfile = 'resreg.txt';

# Default options can be modified through the command line (see intro)

GetOptions ("infile=s" => \$file,
            'help' => \$help,
            "outfile=s" => \$outfile,
            "window=i" => \$winsize,
            "count=i" => \$count,
            "residue=s{0,3}" => \@res,
            "length=i" => \$lengthcount)
            or die ("Error in command line arguments\n");

# If no option is entered for "residue", the default is set to 'K'

if (not defined ($res[0])){
    $res[0] = "K";
    }

# To look for residues in a row, inputs will be passed into an array
which can then be used to create a search parameter.
# The parameter will be inserted into the regular expression later on.

if ((not defined $res[1]) && (not defined $res[2])) {
    $parameter = "[$res[0]]";
}
elsif ((defined $res[1]) && (not defined $res[2])) {
    $parameter = "[$res[0]][$res[1]]";
}
elsif ((defined $res[1]) && (defined $res[2])) {
    $parameter = "[$res[0]][$res[1]][$res[2]]";
}

# The name of the protein and its sequence from the FASTA file are
held in a hash.
# The sequence may span several lines in file, and will be
concatenated.

open(INFILE, $file) or die "can not open $file:$!\n";
open(OUT, ">>$outfile") or die "can not open $outfile:$! \n";

while(<INFILE>) {
    chomp;
    if ((/^>(.+)/)) {
        $genename = $1;
    }
    elsif (/^([AILMFWYVSTNQCUGPRHKDE]+)/i) {
        $genes{$genename} .= $1;
    }
}
# The sliding window moves along each protein sequence in turn.
# $score counts how many instances of the chosen residue occur within
the window.
```

```perl
# If $score is greater than the predefined threshold ($count) the
sequence qualifies to be stored within $regseq.
# In order to capture the whole sequence, the first instance of a
qualifying window stores the first half of the sequence.
# Subsequent qualifying windows add one residue (the center) each,
whereas the final window adds its latter half.
# Once a window which does not qualify is encountered, the qualifying
sequence up to that point can be stored.
# A sequence is stored in the %region hash if it is over the
predefined length ($lengthcount) plus half the window size.
# Sequences can be stored either once a non-qualifying window is
encountered or on the last window within the protein.

foreach my $protein ( keys %genes ) {
    my $sequence = $genes{$protein};
    my $lengthseq = length($sequence);
    my $regseq = undef;

    for(my $i=0;$i<=($lengthseq-$winsize);$i++) {
        my $window = substr($sequence, $i, $winsize);
        $score = 0;

        while ($window =~ /$parameter/gi)  {
            $score ++;
        }

        if ($score >= $count) {
                $regseq .= substr($window, $winsize/2, 1);
        }
        elsif ($score < ($count)) {
            if (length($regseq) >= ($lengthcount)) {
                push(@{$region{$protein}}, $regseq);
                $regseq = undef;
            }
            elsif(length($regseq) < ($lengthcount)) {
            $regseq = undef;
            }
        }

        if($i ==($lengthseq-$winsize)) {
            if(length($regseq) > ($lengthcount)) {
                push(@{$region{$protein}}, $regseq);
            }
            $regseq = undef;
        }
    }
}

# The regions of the protein high in certain amino acids are stored
within the %region hash, and are printed out.
# A number is added to the name of each protein to represent multiple
sequences within the same protein.

foreach my $value (keys %region) {
    for my $i ( 0 .. $#{$region{$value}} ) {
        print OUT ">$value $i\n$region{$value}[$i]\n";
    }
}
```

Appendix Text 3 - 'pickrepeats.pl' perl script. Subroutines are included below. Input file is the output of the TREKS java program. Comments are in red.

```perl
#!usr/bin/perl

# File:  pickrepeats.pl
# Version 1
# This script takes output from the T-Reks JAVA program.
# (T-reks: identification of tandem repeats in sequences with a k-
means based algorithm.
# Jorda J, Kajava AV(2009). Bioinformatics 25 (20), 2632-2638)

# Output must be copied from the JAVA applet and saved as a .txt file.
# See example output in 'repeatexample.txt'.

# The script creates a hash of the protein name, and the length of its
longest repeating sequence.
# (multiple repeating sequences must be separated before inputting
into T-Reks using 'chargesearch.pl')
# The hash is cycled through, printing out repeating sequences which
are longer than a predefined threshold.
# The subroutine 'get_rep' accesses the full length repeat from the
'resreg.txt' file - the input for T-Reks.
# The subroutine 'get_full' then gets the full length protein sequence
from the initial fasta file.
# Both of these are printed out with the name of the protein and the
length of the repeat sequence.
# The number of sequences selected is printed at the end (multiple
sequences may be from the same protein).
# If scoring is turned on, more positively charged proteins are
printed first (highest score first).

# Getopt::Long is used to allow users to change the threshold repeat
length through the command line.
# Options for Getopt::Long: (single letters can be used for each
option, e.g. -l instead of --length)

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;
use get_rep;
use get_full;
use get_repscore;
use Getopt::Long;

# The default threshold length is set to 30 residues.

my $count = 0;
my $genename;
my %genes;
my $length = 30;
my $score = '';
my $help = '';
my $name;
my @data;
my $repseq;
my $repscore;
my $fullseq;
my $replength;
my $psim;
my $seqlength;
```

```perl
my %result;

my $file = 'repeats.txt';
my $outfile = 'replistr.txt';

my $repinfile = 'resreg.txt';
my $fullinfile = 'allexported.txt';

# Default options can be modified through the command line.

GetOptions ('help' => \$help,
            "length=i" => \$length,
            'score!' => \$score,
            "infile=s" => \$file,
            "outfile=s" => \$outfile,
            "repfile=s" => \$repinfile,
            "fullfile=s" => \$fullinfile)
            or die ("Error in command line arguments\n");

# Bring up usage instructions if --help option is entered, and end
script.

if ($help) {
    die ("\nUSAGE:
[-h or --help]      bring up usage instructions
[-i or --input]     filename of T-REKS output file for input
[-o or --output]    filename for output
[-r or --repfile]   filename of input to the 'get_rep.pm' subroutine
[-f or --fullfile]  filename of input to the 'get_full.pm' subroutine
[-l or --length]    the threshold length for a stored sequence
[-s or --score]     turns on scoring by the 'get_repscore.pm'
subroutine
[-nos or --noscore] turns off scoring by the 'get_repscore.pm'
subroutine\n
See README file for more details on usage with examples\n");
}

# The input file 'repeats.txt' contains the output from the T-Reks
JAVA program.

open(INFILE, $file) or die "can not open $file:$!\n";
open(OUT, ">>$outfile") or die "can not open $outfile:$! \n";

# The full name (and any other information contained in this line) is
stored in a hash.
# Multiple repeating sequences may exist for one input sequence - only
the length of the longest is stored.

while(<INFILE>) {
    chomp;
    if ((/^>(.+)/)) {
    $genename = $1;
    }
    elsif (/^Length:\s(\d+)\sresidues.+Psim:(0\.\d+).+Length:(\d+)/) {
        $replength = $1;
        $psim = $2;
        $seqlength = $3;
        if (defined $genes{$genename}) {
            if ($seqlength > $genes{$genename}[2]) {
            undef $genes{$genename};
            $genes{$genename} = ["$replength", "$psim", "$seqlength"];
            }
        }
```

157

```perl
        elsif (not defined $genes{$genename}[2]) {
            $genes{$genename} = ["$replength", "$psim", "$seqlength"];
        }
    }
}


# Repeating sequences over a certain threshold length ($length) are
printed out.
# The 'get_rep' subroutine sources the repeating sequence while
'get_full' sources the full protein sequence.
# Both these are printed along with the protein name and the length of
the repeat.
# If no files are submitted for the subroutines, no sequences are
printed out - just protein names.
# the 'get_repscore' subroutine scores each sequence according to
percentage positive charge.
# If scoring is turned off, sequences are printed out randomly (all
scores are set to 1).

foreach my $protein ( keys %genes ) {
    if ($genes{$protein}[2] >= $length) {
        if (defined $repinfile) {
            $repseq = get_rep($protein, $repinfile);
            if ($score && $repseq) {
                $repscore = get_repscore($repseq);
            }
            else {
                $repscore = "1";
            }
        }
        elsif (not defined $repinfile) {
            $repseq = "Sequence input file not supplied - See README";
            $repscore = "1";
        }
        if (defined $fullinfile) {
            @data = split ('\s', $protein, 2);
            $name = $data[0];
            $fullseq = get_full($name, $fullinfile);
        }
        elsif (not defined $fullinfile) {
            $fullseq = "Full protein input file not supplied - See
README";
        }
        $result{$protein} = ["$repscore", "$genes{$protein}[2]",
"$repseq", "$fullseq"];
        $count ++;
    }
}

# Sequences are printed out highest score first. This may be disabled.

foreach my $name (sort {$result{$b}[0] <=> $result{$a}[0]} keys
%result) {
print OUT ">$name\nLength of Repeat
sequence:$result{$name}[1]\nScore:$result{$name}[0]\nRepeating
Sequence\n$result{$name}[2]\nFull Sequence:\n$result{$name}[3]\n";
}
# The number of repeat sequences in the output is printed at the end
of the file ($count).

print OUT "Count: $count";
```

Subroutine get_rep

```perl
#!/usr/bin/perl

# File: get_rep.pm
# Version 1
# A subroutine to get the protein sequences entered into T-Reks
program.
# The input file is the output of 'chargesearch.pl' which is in the
FASTA format.
# Protein names and sequences are collected into a hash.
# Repeat sequences which pass the threshold length from
'pickrepeats.pl' will have their names
# searched within the input file.
# The full repeat sequence is returned to 'pickrepeats.pl'.

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;

sub get_rep
{
# The full protein name as well as the number given by
'chargesearch.pl' is passed from
# 'pickrepeats.pl', and also the input file name.

    my ($ID, $file) = @_;
    open(INFILE, $file) or die "can not open $file:$!\n";

    my $genename;
    my %genes;

# A hash is built of protein names and sequences.

    while(<INFILE>) {
        chomp;
        if (/^>(.+)/) {
            $genename = $1;
        }
        elsif (/^([AILMFWYVSTNQCUGPRHKDE]+)/i) {
            $genes{$genename} = $1;
        }
    }
# Protein names within the hash are searched for names passed from
'pickrepeats.pl'
# If a match is found, the full sequence for that repeat is returned.

    foreach my $value ( keys %genes ) {
        if ($value eq $ID) {
        return $genes{$value};
        }
    }
}

1;
```

Subroutine get_full

```perl
#!/usr/bin/perl

# File: get_full.pm
# Version 1
# A subroutine to get a full protein sequence from the FASTA file used
as input for 'chargesearch.pl'.
# Protein names and sequences are collected into a hash.
# Repeat sequences which pass the threshold length from
'pickrepeats.pl' will have their names
# searched within the input file.
# The full protein sequence is returned to 'pickrepeats.pl'.

# Date: 16.07.2015

use strict;
use warnings;
use Data::Dumper;

sub get_full
{
    # The protein name (first word only) and the input file name are
passed from 'pickrepeats.pl'.

    my ($ID, $file) = @_;
    open(INFILE, $file) or die "can not open $file:$!\n";

    my $genenamef;
    my %genesf;

    # A hash is built of protein names and sequences. Multiple lines
are concatenated.

    while(<INFILE>) {
        chomp;
        if (/^>(\w{13})\s.+/) {
            $genenamef = $1;
        }
        elsif (/^([AILMFWYVSTNQCUGPRHKDE]+)/i) {
            $genesf{$genenamef} .= $1;
        }
    }
    # Protein names within the hash are searched for the first word
passed from 'pickrepeats.pl'
    # If a match is found, the full sequence for that protein is
returned.

    foreach my $valuef ( keys %genesf ) {
        if ($valuef eq $ID) {
        print "YES\n";
        return $genesf{$valuef};
        }
    }
}

1;
```

Bibliography

ABDEL-LATIF, M. S., KHATTAB, A., LINDENTHAL, C., KREMSNER, P. G. & KLINKERT, M. Q. 2002. Recognition of variant Rifin antigens by human antibodies induced during natural Plasmodium falciparum infections. *Infect Immun,* 70**,** 7013-21.

ADINI, A., KRUGLIAK, M., GINSBURG, H., LI, L. L., LAVIE, L. & WARBURG, A. 2001. Transglutaminase in Plasmodium parasites: activity and putative role in oocysts and blood stages. *Molecular and Biochemical Parasitology,* 117**,** 161-168.

ADJALLEY, S. H., LEE, M. C. & FIDOCK, D. A. 2010. A method for rapid genetic integration into Plasmodium falciparum utilizing mycobacteriophage Bxb1 integrase. *Methods Mol Biol,* 634**,** 87-100.

AIKAWA, M., KAMANURA, K., SHIRAISHI, S., MATSUMOTO, Y., ARWATI, H., TORII, M., ITO, Y., TAKEUCHI, T. & TANDLER, B. 1996. Membrane knobs of unfixed Plasmodium falciparum infected erythrocytes: New findings as revealed by atomic force microscopy and surface potential spectroscopy. *Experimental Parasitology,* 84**,** 339-343.

AIKAWA, M., MILLER, L. H. & RABBEGE, J. 1975. CAVEOLA-VESICLE COMPLEXES IN PLASMALEMMA OF ERYTHROCYTES INFECTED BY PLASMODIUM-VIVAX AND PLASMODIUM-CYNOMOLGI - UNIQUE STRUCTURES RELATED TO SCHUFFNERS DOTS. *American Journal of Pathology,* 79**,** 285-300.

AKOMPONG, T., KADEKOPPALA, M., HARRISON, T., OKSMAN, A., GOLDBERG, D. E., FUJIOKA, H., SAMUEL, B. U., SULLIVAN, D. & HALDAR, K. 2002. trans expression of a Plasmodium falciparum histidine-rich protein II (HRPII) reveals sorting of soluble proteins in the periphery of the host erythrocyte malarial food vacuole. *Journal of Biological Chemistry,* 277**,** 28923-28933.

ALAM, M. M., SOLYAKOV, L., BOTTRILL, A. R., FLUECK, C., SIDDIQUI, F. A., SINGH, S., MISTRY, S., VISKADURAKI, M., LEE, K., HOPP, C. S., CHITNIS, C. E., DOERIG, C., MOON, R. W., GREEN, J. L., HOLDER, A. A., BAKER, D. A. & TOBIN, A. B. 2015. Phosphoproteomics reveals malaria parasite Protein Kinase G as a signalling hub regulating egress and invasion. *Nature Communications,* 6**,** 15.

ALEXANDRE, J. S. F., YAHATA, K., KAWAI, S., TORII, M. & KANEKO, O. 2011. PEXEL-independent trafficking of Plasmodium falciparum SURFIN4.2 to the parasite-infected red blood cell and Maurer's clefts. *Parasitology International,* 60**,** 313-320.

AMAIDEN, M. R., MONESTEROLO, N. E., SANTANDER, V. S., CAMPETELLI, A. N., ARCE, C. A., PIE, J., HOPE, S. I., VATTA, M. S. & CASALE, C. H. 2012. Involvement of membrane tubulin in erythrocyte deformability and blood pressure. *Journal of Hypertension,* 30**,** 1414-1422.

AMAIDEN, M. R., SANTANDER, V. S., MONESTEROLO, N. E., CAMPETELLI, A. N., RIVELLI, J. F., PREVITALI, G., ARCE, C. A. & CASALE, C. H. 2011. Tubulin pools in human erythrocytes: altered distribution in hypertensive patients affects Na+, K+-ATPase activity. *Cellular and Molecular Life Sciences,* 68**,** 1755-1768.

AMINO, R., THIBERGE, S., MARTIN, B., CELLI, S., SHORTE, S., FRISCHKNECHT, F. & MENARD, R. 2006. Quantitative imaging of Plasmodium transmission from mosquito to mammal. *Nature Medicine,* 12**,** 220-224.

AN, X. L., LECOMTE, M. C., CHASIS, J. A., MOHANDAS, N. & GRATZER, W. 2002. Shear-response of the spectrin dimer-tetramer equilibrium in the red blood cell membrane. *Journal of Biological Chemistry,* 277**,** 31796-31800.

ANDRADE, M. A., PEREZ-IRATXETA, C. & PONTING, C. P. 2001. Protein repeats: Structures, functions, and evolution. *Journal of Structural Biology,* 134**,** 117-131.

ANSARI, H. R., TEMPLETON, T. J., SUBUDHI, A. K., RAMAPRASAD, A., TANG, J., LU, F., NAEEM, R., HASHISH, Y., OGUIKE, M. C., BENAVENTE, E. D., CLARK, T. G., SUTHERLAND, C. J., BARNWELL, J. W., CULLETON, R., CAO, J. & PAIN, A. 2016. Genome-scale comparison of expanded gene families in Plasmodium ovale wallikeri and Plasmodium ovale curtisi with Plasmodium malariae and with other Plasmodium species. *Int J Parasitol*.

ANSTEY, N. M., DOUGLAS, N. M., POESPOPRODJO, J. R. & PRICE, R. N. 2012. Plasmodium vivax: Clinical Spectrum, Risk Factors and Pathogenesis. *In:* HAY, S. I., PRICE, R. & BAIRD, J. K. (eds.) *Advances in Parasitology, Vol 80: Epidemiology of Plasmodium Vivax: History, Hiatus and Hubris, Pt A.* San Diego: Elsevier Academic Press Inc.

ARIEY, F., WITKOWSKI, B., AMARATUNGA, C., BEGHAIN, J., LANGLOIS, A. C., KHIM, N., KIM, S., DURU, V., BOUCHIER, C., MA, L., LIM, P., LEANG, R., DUONG, S., SRENG, S., SUON, S., CHUOR, C. M., BOUT, D. M., MENARD, S., ROGERS, W. O., GENTON, B., FANDEUR, T., MIOTTO, O., RINGWALD, P., LE BRAS, J., BERRY, A., BARALE, J. C., FAIRHURST, R. M., BENOIT-VICAL, F., MERCEREAU-PUIJALON, O. & MENARD, D. 2014. A molecular marker of artemisinin-resistant Plasmodium falciparum malaria. *Nature,* 505**,** 50-+.

ARMSTRONG, C. M. & GOLDBERG, D. E. 2007. An FKBP destabilization domain modulates protein levels in Plasmodium falciparum. *Nature Methods,* 4**,** 1007-1009.

ATAMNA, H. & GINSBURG, H. 1993. ORIGIN OF REACTIVE OXYGEN SPECIES IN ERYTHROCYTES INFECTED WITH PLASMODIUM-FALCIPARUM. *Molecular and Biochemical Parasitology,* 61**,** 231-241.

ATKINSON, J. & TRUTER, E. 2009. ALEXANDER'S LAST DAYS: MALARIA AND MIND GAMES? *Acta Classica,* 52**,** 23-46.

BANNISTER, L. H., HOPKINS, J. M., FOWLER, R. E., KRISHNA, S. & MITCHELL, G. H. 2000. A brief illustrated guide to the ultrastructure of Plasmodium falciparum asexual blood stages. *Parasitology Today,* 16**,** 427-433.

BANSAL, A., SINGH, S., MORE, K. R., HANS, D., NANGALIA, K., YOGAVEL, M., SHARMA, A. & CHITNIS, C. E. 2013. Characterization of Plasmodium falciparum Calcium-dependent Protein Kinase 1 (PfCDPK1) and Its Role in Microneme Secretion during Erythrocyte Invasion. *Journal of Biological Chemistry,* 288**,** 1590-1602.

BARNWELL, J. W., HOWARD, R. J., COON, H. G. & MILLER, L. H. 1983. Splenic requirement for antigenic variation and expression of the variant antigen on the erythrocyte membrane in cloned Plasmodium knowlesi malaria. *Infect Immun,* 40**,** 985-94.

BARTHOLDSON, S. J., CROSNIER, C., BUSTAMANTE, L. Y., RAYNER, J. C. & WRIGHT, G. J. 2013. Identifying novel Plasmodium falciparum erythrocyte invasion receptors using systematic extracellular protein interaction screens. *Cellular Microbiology,* 15**,** 1304-1312.

BARUCH, D. I., GORMLEY, J. A., MA, C., HOWARD, R. J. & PASLOSKE, B. L. 1996. Plasmodium falciparum erythrocyte membrane protein 1 is a parasitized erythrocyte receptor for adherence to CD36, thrombospondin, and intercellular adhesion molecule 1. *Proceedings of the National Academy of Sciences of the United States of America,* 93**,** 3497-3502.

BAUM, J., RICHARD, D., HEALER, J., RUG, M., KRNAJSKI, Z., GILBERGER, T. W., GREEN, J. L., HOLDER, A. A. & COWMAN, A. F. 2006. A conserved molecular motor drives cell invasion and gliding motility across malaria life cycle stages and other apicomplexan parasites. *Journal of Biological Chemistry,* 281**,** 5197-5208.

BECK, J. R., MURALIDHARAN, V., OKSMAN, A. & GOLDBERG, D. E. 2014. PTEX component HSP101 mediates export of diverse malaria effectors into host erythrocytes. *Nature,* 511**,** 592-+.

BECKER, K., TILLEY, L., VENNERSTROM, J. L., ROBERTS, D., ROGERSON, S. & GINSBURG, H. 2004. Oxidative stress in malaria parasite-infected erythrocytes: host-parasite interactions. *International Journal for Parasitology,* 34**,** 163-189.

BENEDETTI, C. E., KOBARG, J., PERTINHEZ, T. A., GATTI, R. M., DE SOUZA, O. N., SPISNI, A. & MENEGHINI, R. 2003. Plasmodium falciparum histidine-rich protein II binds to actin, phosphatidylinositol 4,5-bisphosphate and erythrocyte ghosts in a pH-dependent manner and undergoes coil-to-helix transitions in anionic micelles. *Molecular and Biochemical Parasitology,* 128**,** 157-166.

BENNETT, B. J., MOHANDAS, N. & COPPEL, R. L. 1997. Defining the minimal domain of the Plasmodium falciparum protein MESA involved in the interaction with the red cell membrane skeletal protein 4.1. *Journal of Biological Chemistry,* 272**,** 15299-15306.

BHATTACHARJEE, S., VAN OOIJ, C., BALU, B., ADAMS, J. H. & HALDAR, K. 2008. Maurer's clefts of Plasmodium falciparum are secretory organelles that concentrate virulence protein reporters for delivery to the host erythrocyte. *Blood,* 111**,** 2418-2426.

BILLKER, O., LINDO, V., PANICO, M., ETIENNE, A. E., PAXTON, T., DELL, A., ROGERS, M., SINDEN, R. E. & MORRIS, H. R. 1998. Identification of xanthurenic acid as the putative inducer of malaria development in the mosquito. *Nature,* 392**,** 289-292.

BLACK, C. G., PROELLOCKS, N. I., KATS, L. M., COOKE, B. M., MOHANDAS, N. & COPPEL, R. L. 2008. In vivo studies support the role of trafficking and cytoskeletal-binding motifs in the interaction of MESA with the membrane skeleton of Plasmodium falciparum-infected red blood cells. *Molecular and Biochemical Parasitology,* 160**,** 143-147.

BLAIR, P. L., KAPPE, S. H. I., MACIEL, J. E., BALU, B. & ADAMS, J. H. 2002. Plasmodium falciparum MAEBL is a unique member of the ebl family. *Molecular and Biochemical Parasitology,* 122**,** 35-44.

BODDEY, J. A., CARVALHO, T. G., HODDER, A. N., SARGEANT, T. J., SLEEBS, B. E., MARAPANA, D., LOPATICKI, S., NEBL, T. & COWMAN, A. F. 2013. Role of Plasmepsin V in Export of Diverse Protein Families from the Plasmodium falciparum Exportome. *Traffic,* 14**,** 532-550.

BODDEY, J. A. & COWMAN, A. F. 2013. Plasmodium Nesting: Remaking the Erythrocyte from the Inside Out. *In:* GOTTESMAN, S. (ed.) *Annual Review of Microbiology, Vol 67.* Palo Alto: Annual Reviews.

BODDEY, J. A., HODDER, A. N., GUNTHER, S., GILSON, P. R., PATSIOURAS, H., KAPP, E. A., PEARCE, J. A., DE KONING-WARD, T. F., SIMPSON, R. J., CRABB, B. S. & COWMAN, A. F. 2010. An aspartyl protease directs malaria effector proteins to the host cell. *Nature,* 463**,** 627-U52.

BOTHA, M., PESCE, E. R. & BLATCH, G. L. 2007. The Hsp40 proteins of Plasmodium falciparum and other apicomplexa: Regulating chaperone power in the parasite and the host. *International Journal of Biochemistry & Cell Biology,* 39**,** 1781-1803.

BRANDT, G. S. & BAILEY, S. 2013. Dematin, a human erythrocyte cytoskeletal protein, is a substrate for a recombinant FIKK kinase from Plasmodium falciparum. *Molecular and Biochemical Parasitology,* 191**,** 20-23.

BROWN, J. W., BULLITT, E., SRISWASDI, S., HARPER, S., SPEICHER, D. W. & MCKNIGHT, C. J. 2015. The Physiological Molecular Shape of Spectrin: A Compact Supercoil Resembling a Chinese Finger Trap. *Plos Computational Biology,* 11**,** 20.

BUCHAN, D. W. A., MINNECI, F., NUGENT, T. C. O., BRYSON, K. & JONES, D. T. 2013. Scalable web services for the PSIPRED Protein Analysis Workbench. *Nucleic Acids Research,* 41**,** W349-W357.

BUFFET, P. A., MILON, G., BROUSSE, V., CORREAS, J. M., DOUSSET, B., COUVELARD, A., KIANMANESH, R., FARGES, O., SAUVANET, A., PAYE, F., UNGEHEUER, M. N., OTTONE, C., KHUN, H., FIETTE, L., GUIGON, G., HUERRE, M., MERCEREAU-PUIJALON, O. & DAVID, P. H. 2006. Ex vivo perfusion of human spleens maintains clearing and processing functions. *Blood,* 107**,** 3745-3752.

BUFFET, P. A., SAFEUKUI, I., MILON, G., MERCEREAU-PUIJALON, O. & DAVID, P. H. 2009. Retention of erythrocytes in the spleen: a double-edged process in human malaria. *Current Opinion in Hematology,* 16**,** 157-164.

CARLTON, J. M., ADAMS, J. H., SILVA, J. C., BIDWELL, S. L., LORENZI, H., CALER, E., CRABTREE, J., ANGIUOLI, S. V., MERINO, E. F., AMEDEO, P., CHENG, Q., COULSON, R. M. R., CRABB, B. S., DEL PORTILLO, H. A., ESSIEN, K., FELDBLYUM, T. V., FERNANDEZ-BECERRA, C., GILSON, P. R., GUEYE, A. H., GUO, X., KANG'A, S., KOOIJ, T. W. A., KORSINCZKY, M., MEYER, E. V. S., NENE, V., PAULSEN, I., WHITE, O., RALPH, S. A., REN, Q. H., SARGEANT, T. J., SALZBERG, S. L., STOECKERT, C. J., SULLIVAN, S. A., YAMAMOTO, M. M.,

HOFFMAN, S. L., WORTMAN, J. R., GARDNER, M. J., GALINSKI, M. R., BARNWELL, J. W. & FRASER-LIGGETT, C. M. 2008. Comparative genomics of the neglected human malaria parasite Plasmodium vivax. *Nature,* 455**,** 757-763.

CARVALHO, B. O., LOPES, S. C. P., NOGUEIRA, P. A., ORLANDI, P. P., BARGIERI, D. Y., BLANCO, Y. C., MAMONI, R., LEITE, J. A., RODRIGUES, M. M., SOARES, I. S., OLIVEIRA, T. R., WUNDERLICH, G., LACERDA, M. V. G., DEL PORTILLO, H. A., ARAUJO, M. O. G., RUSSELL, B., SUWANARUSK, R., SNOUNOU, G., RENIA, L. & COSTA, F. T. M. 2010. On the Cytoadhesion of Plasmodium vivax-Infected Erythrocytes. *Journal of Infectious Diseases,* 202**,** 638-647.

CHANG, H. H., FALICK, A. M., CARLTON, P. M., SEDAT, J. W., DERISI, J. L. & MARLETTA, M. A. 2008. N-terminal processing of proteins exported by malaria parasites. *Mol Biochem Parasitol,* 160**,** 107-15.

CHAPMAN, R. D., PALANCADE, B., LANG, A., BENSAUDE, O. & EICK, D. 2004. The last CTD repeat of the mammalian RNA polymerase II large subunit is important for its stability. *Nucleic Acids Research,* 32**,** 35-44.

CLAESSENS, A., ADAMS, Y., GHUMRA, A., LINDERGARD, G., BUCHAN, C. C., ANDISI, C., BULL, P. C., MOK, S., GUPTA, A. P., WANG, C. W., TURNER, L., ARMAN, M., RAZA, A., BOZDECH, Z. & ROWE, J. A. 2012. A subset of group A-like var genes encodes the malaria parasite ligands for binding to human brain endothelial cells. *Proceedings of the National Academy of Sciences of the United States of America,* 109**,** E1772-E1781.

CLARK, D. 2010. ChemBioDraw Ultra 12.0. *Chemistry World,* 7**,** 55-+.

COLLINS, C. R., HACKETT, F., STRATH, M., PENZO, M., WITHERS-MARTINEZ, C., BAKER, D. A. & BLACKMAN, M. J. 2013. Malaria Parasite cGMP-dependent Protein Kinase Regulates Blood Stage Merozoite Secretory Organelle Discharge and Egress. *Plos Pathogens,* 9**,** 13.

COLLINS, W. E. & JEFFERY, G. M. 2005. Plasmodium ovale: Parasite and disease. *Clinical Microbiology Reviews,* 18**,** 570-+.

COLLINS, W. E. & JEFFERY, G. M. 2007. Plasmodium malatiae: Parasite and disease. *Clinical Microbiology Reviews,* 20**,** 579-+.

CONWAY, D. J., CAVANAGH, D. R., TANABE, K., ROPER, C., MIKES, Z. S., SAKIHAMA, N., BOJANG, K. A., ODUOLA, A. M. J., KREMSNER, P. G., ARNOT, D. E., GREENWOOD, B. M. & MCBRIDE, J. S. 2000. A principal target of human immunity to malaria identified by molecular population genetic and immunological analyses. *Nature Medicine,* 6**,** 689-692.

COOKE, B. M., BUCKINGHAM, D. W., GLENISTER, F. K., FERNANDEZ, K. M., BANNISTER, L. H., MARTI, M., MOHANDAS, N. & COPPEL, R. L. 2006. A Maurer's cleft-associated protein is essential for expression of the major malaria virulence antigen on the surface of infected red blood cells. *Journal of Cell Biology,* 172**,** 899-908.

COPPEL, R. L., LUSTIGMAN, S., MURRAY, L. & ANDERS, R. F. 1988. MESA IS A PLASMODIUM-FALCIPARUM PHOSPHOPROTEIN ASSOCIATED WITH THE ERYTHROCYTE-MEMBRANE SKELETON. *Molecular and Biochemical Parasitology,* 31**,** 223-232.

CORREDOR, V., MEYER, E. V. S., LAPP, S., CORREDOR-MEDINA, C., HUBER, C. S., EVANS, A. G., BARNWELL, J. W. & GALINSKI, M. R. 2004. A SICAvar switching event in Plasmodium knowlesi is associated with the DNA rearrangement of conserved 3 ' non-coding sequences. *Molecular and Biochemical Parasitology,* 138**,** 37-49.

COUNIHAN, N. A., KALANON, M., COPPEL, R. L. & DE KONING-WARD, T. F. 2013. Plasmodium rhoptry proteins: why order is important. *Trends in Parasitology,* 29**,** 228-236.

COWMAN, A. F., COPPEL, R. L., SAINT, R. B., FAVALORO, J., CREWTHER, P. E., STAHL, H. D., BIANCO, A. E., BROWN, G. V., ANDERS, R. F. & KEMP, D. J. 1984. The ring-infected erythrocyte surface antigen (RESA) polypeptide of Plasmodium falciparum contains two separate blocks of tandem repeats encoding antigenic epitopes that are naturally immunogenic in man. *Molecular biology & medicine,* 2**,** 207-21.

COX-SINGH, J., DAVIS, T. M. E., LEE, K. S., SHAMSUL, S. S. G., MATUSOP, A., RATNAM, S., RAHMAN, H. A., CONWAY, D. J. & SINGH, B. 2008. Plasmodium knowlesi malaria in humans is widely distributed and potentially life threatening. *Clinical Infectious Diseases,* 46**,** 165-171.

COX-SINGH, J., HIU, J., LUCAS, S. B., DIVIS, P. C., ZULKARNAEN, M., CHANDRAN, P., WONG, K. T., ADEM, P., ZAKI, S. R., SINGH, B. & KRISHNA, S. 2010. Severe malaria - a case of fatal Plasmodium knowlesi infection with post-mortem findings: a case report. *Malaria Journal,* 9**,** 7.

CRABB, B. S., COOKE, B. M., REEDER, J. C., WALLER, R. F., CARUANA, S. R., DAVERN, K. M., WICKHAM, M. E., BROWN, G. V., COPPEL, R. L. & COWMAN, A. F. 1997. Targeted gene disruption shows that knobs enable malaria-infected red cells to cytoadhere under physiological shear stress. *Cell,* 89**,** 287-296.

CRAIG, A. G., GRAU, G. E., JANSE, C., KAZURA, J. W., MILNER, D., BARNWELL, J. W., TURNER, G., LANGHORNE, J. & HINXTON RETREAT MEETING ANIM, M. 2012a. The Role of Animal Models for Research on Severe Malaria. *Plos Pathogens,* 8**,** 9.

CRAIG, A. G., KHAIRUL, M. F. M. & PATIL, P. R. 2012b. Cytoadherence and severe malaria. *The Malaysian journal of medical sciences : MJMS,* 19**,** 5-18.

CULVENOR, J. G., LANGFORD, C. J., CREWTHER, P. E., SAINT, R. B., COPPEL, R. L., KEMP, D. J., ANDERS, R. F. & BROWN, G. V. 1987. Plasmodium falciparum - Identification and Localization Of A Knob Protein Antigen Expressed by a CDNA   Clone. *Experimental Parasitology,* 63**,** 58-67.

CYRKLAFF, M., SANCHEZ, C. P., KILIAN, N., BISSEYE, C., SIMPORE, J., FRISCHKNECHT, F. & LANZER, M. 2011. Hemoglobins S and C Interfere with Actin Remodeling in Plasmodium falciparum-Infected Erythrocytes. *Science,* 334**,** 1283-1286.

DAS, S., HERTRICH, N., PERRIN, A. J., WITHERS-MARTINEZ, C., COLLINS, C. R., JONES, M. L., WATERMEYER, J. M., FOBES, E. T., MARTIN, S. R., SAIBIL, H. R., WRIGHT, G. J., TREECK, M., EPP, C. & BLACKMAN, M. J. 2015. Processing of Plasmodium falciparum Merozoite Surface Protein MSP1 Activates a Spectrin-Binding Function Enabling Parasite Egress from RBCs. *Cell Host & Microbe,* 18**,** 433-444.

DE KONING-WARD, T. F., GILSON, P. R., BODDEY, J. A., RUG, M., SMITH, B. J., PAPENFUSS, A. T., SANDERS, P. R., LUNDIE, R. J., MAIER, A. G., COWMAN, A. F. & CRABB, B. S. 2009. A newly discovered protein export machine in malaria parasites. *Nature,* 459**,** 945-U66.

DE KONING-WARD, T. F., GILSON, P. R. & CRABB, B. S. 2015. Advances in molecular genetic systems in malaria. *Nature Reviews Microbiology,* 13**,** 373-387.

DECHERF, G., EGEE, S., STAINES, H. M., ELLORY, J. C. & THOMAS, S. L. 2004. Anionic channels in malaria-infected human red blood cells. *Blood Cells Molecules and Diseases,* 32**,** 366-371.

DEPRISTO, M. A., ZILVERSMIT, M. M. & HARTL, D. L. 2006. On the abundance, amino acid composition, and evolutionary dynamics of low-complexity regions in proteins. *Gene,* 378**,** 19-30.

DESAI, S. A. 2014. Why do malaria parasites increase host erythrocyte permeability? *Trends in Parasitology,* 30**,** 151-159.

DIEZ-SILVA, M., PARK, Y., HUANG, S., BOW, H., MERCEREAU-PUIJALON, O., DEPLAINE, G., LAVAZEC, C., PERROT, S., BONNEFOY, S., FELD, M. S., HAN, J., DAO, M. & SURESH, S. 2012. Pf155/RESA protein influences the dynamic microcirculatory behavior of ring-stage Plasmodium falciparum infected red blood cells. *Scientific Reports,* 2**,** 7.

DIXON, M. W. A., KENNY, S., MCMILLAN, P. J., HANSSEN, E., TRENHOLME, K. R., GARDINER, D. L. & TILLEY, L. 2011. Genetic ablation of a Maurer's cleft protein prevents assembly of the Plasmodium falciparum virulence complex. *Molecular Microbiology,* 81**,** 982-993.

DNASTAR SeqMan NGen®. Version 12.0 ed. Madison, WI.

DOERIG, C., RAYNER, J. C., SCHERF, A. & TOBIN, A. B. 2015. Post-translational protein modifications in malaria parasites. *Nature Reviews Microbiology,* 13**,** 160-172.

DVORIN, J. D., MARTYN, D. C., PATEL, S. D., GRIMLEY, J. S., COLLINS, C. R., HOPP, C. S., BRIGHT, A. T., WESTENBERGER, S., WINZELER, E., BLACKMAN, M. J., BAKER, D. A., WANDLESS, T. J. & DURAISINGH, M. T. 2010. A Plant-Like Kinase in Plasmodium falciparum Regulates Parasite Egress from Erythrocytes. *Science,* 328**,** 910-912.

ELSWORTH, B., MATTHEWS, K., NIE, C. Q., KALANON, M., CHARNAUD, S. C., SANDERS, P. R., CHISHOLM, S. A., COUNIHAN, N. A., SHAW, P. J., PINO, P., CHAN, J. A., AZEVEDO, M. F., ROGERSON, S. J., BEESON, J. G., CRABB, B. S., GILSON, P. R. & DE KONING-WARD, T. F. 2014. PTEX is an essential nexus for protein export in malaria parasites. *Nature,* 511**,** 587-+.

ELSWORTH, B., SANDERS, P. R., NEBL, T., BATINOVIC, S., KALANON, M., NIE, C. Q., CHARNAUD, S. C., BULLEN, H. E., DE KONING WARD, T. F., TILLEY, L., CRABB, B. S. & GILSON, P. R. 2016. Proteomic analysis reveals novel proteins associated with the Plasmodium protein exporter PTEX and a loss of complex stability upon truncation of the core PTEX component, PTEX150. *Cell Microbiol*.

ESCALANTE, A. A. & AYALA, F. J. 1994. PHYLOGENY OF THE MALARIAL GENUS PLASMODIUM, DERIVED FROM RIBOSOMAL-RNA GENE-SEQUENCES. *Proceedings of the National Academy of Sciences of the United States of America,* 91**,** 11373-11377.

FATIH, F. A., SINER, A., AHMED, A., WOON, L. C., CRAIG, A. G., SINGH, B., KRISHNA, S. & COX-SINGH, J. 2012. Cytoadherence and virulence - the case of Plasmodium knowlesi malaria. *Malaria Journal,* 11**,** 6.

FERGUSON, D. J. P., BALABAN, A. E., PATZEWITZ, E. M., WALL, R. J., HOPP, C. S., POULIN, B., MOHMMED, A., MALHOTRA, P., COPPI, A., SINNIS, P. & TEWARI, R. 2014. The Repeat Region of the Circumsporozoite Protein is Critical for Sporozoite Formation and Maturation in Plasmodium. *Plos One,* 9**,** 25.

FINN, R. D., CLEMENTS, J. & EDDY, S. R. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research,* 39**,** W29-W37.

FOLEY, M., TILLEY, L., SAWYER, W. H. & ANDERS, R. F. 1991. The Ring-Infected Erythrocyte Surface-Antigen of Plasmodium-falciparum Associates with Spectrin in the Erythrocyte-Membrane. *Molecular and Biochemical Parasitology,* 46**,** 137-148.

FONDON, J. W. & GARNER, H. R. 2004. Molecular origins of rapid and continuous morphological evolution. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 18058-18063.

FREVERT, U., ENGELMANN, S., ZOUGBEDE, S., STANGE, J., NG, B., MATUSCHEWSKI, K. & LIEBES, L. 2005. Intravital observation of Plasmodium berghei sporozoite infection of the liver. *Plos Biology,* 3**,** 1034-1046.

FUJIKI, Y., HUBBARD, A. L., FOWLER, S. & LAZAROW, P. B. 1982. Isolation of intracellular membranes by means of sodium carbonate treatment: application to endoplasmic reticulum. *J Cell Biol,* 93**,** 97-102.

GANGULY, A. K., RANJAN, P., KUMAR, A. & BHAVESH, N. S. 2015. Dynamic association of PfEMP1 and KAHRP in knobs mediates cytoadherence during Plasmodium invasion. *Scientific Reports,* 5**,** 9.

GANTZ, V. M., JASINSKIENE, N., TATARENKOVA, O., FAZEKAS, A., MACIAS, V. M., BIER, E. & JAMES, A. A. 2015. Highly efficient Cas9-mediated gene drive for population modification of the malaria vector mosquito Anopheles stephensi. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** E6736-E6743.

GARDNER, M. J., HALL, N., FUNG, E., WHITE, O., BERRIMAN, M., HYMAN, R. W., CARLTON, J. M., PAIN, A., NELSON, K. E., BOWMAN, S., PAULSEN, I. T., JAMES, K., EISEN, J. A., RUTHERFORD, K., SALZBERG, S. L., CRAIG, A., KYES, S., CHAN, M. S., NENE, V., SHALLOM, S. J., SUH, B., PETERSON, J., ANGIUOLI, S., PERTEA, M., ALLEN, J., SELENGUT, J., HAFT, D., MATHER, M. W., VAIDYA, A. B., MARTIN, D. M. A., FAIRLAMB, A. H., FRAUNHOLZ, M. J., ROOS, D. S., RALPH, S. A., MCFADDEN, G. I., CUMMINGS, L. M., SUBRAMANIAN, G. M., MUNGALL, C., VENTER, J. C., CARUCCI, D. J., HOFFMAN, S. L.,

NEWBOLD, C., DAVIS, R. W., FRASER, C. M. & BARRELL, B. 2002. Genome sequence of the human malaria parasite Plasmodium falciparum. *Nature,* 419**,** 498-511.

GASTEIGER, E., HOOGLAND, C., GATTIKER, A., DUVAUD, S., WILKINS, M. R., APPEL, R. D. & BAIROCH, A. 2005. Protein Identification and Analysis Tools on the ExPASy Server. *In:* WALKER, J. M. (ed.) *The Proteomics Protocols Handbook.* Humana Press.

GEMAYEL, R., CHAVALI, S., POUGACH, K., LEGENDRE, M., ZHU, B., BOEYNAEMS, S., VAN DER ZANDE, E., GEVAERT, K., ROUSSEAU, F., SCHYMKOWITZ, J., BABU, M. M. & VERSTREPEN, K. J. 2015. Variable Glutamine-Rich Repeats Modulate Transcription Factor Activity. *Molecular Cell,* 59**,** 615-627.

GEMAYEL, R., CHO, J., BOEYNAEMS, S. & VERSTREPEN, K. J. 2012. Beyond Junk-Variable Tandem Repeats as Facilitators of Rapid Evolution of Regulatory and Coding Sequences. *Genes,* 3**,** 461-480.

GEMAYEL, R., VINCES, M. D., LEGENDRE, M. & VERSTREPEN, K. J. 2010. Variable Tandem Repeats Accelerate Evolution of Coding and Regulatory Sequences. *In:* CAMPBELL, A., LICHTEN, M. & SCHUPBACH, G. (eds.) *Annual Review of Genetics, Vol 44.* Palo Alto: Annual Reviews.

GHORBAL, M., GORMAN, M., MACPHERSON, C. R., MARTINS, R. M., SCHERF, A. & LOPEZ-RUBIO, J. J. 2014. Genome editing in the human malaria parasite Plasmodium falciparum using the CRISPR-Cas9 system. *Nature Biotechnology,* 32**,** 819-821.

GILSON, P. R. & CRABB, B. S. 2009. Morphology and kinetics of the three distinct phases of red blood cell invasion by Plasmodium falciparum merozoites. *International Journal for Parasitology,* 39**,** 91-96.

GINSBURG, H., HANDELI, S., FRIEDMAN, S., GORODETSKY, R. & KRUGLIAK, M. 1986. EFFECTS OF RED-BLOOD-CELL POTASSIUM AND HYPERTONICITY ON THE GROWTH OF PLASMODIUM-FALCIPARUM IN CULTURE. *Zeitschrift Fur Parasitenkunde-Parasitology Research,* 72**,** 185-199.

GLENISTER, F. K., COPPEL, R. L., COWMAN, A. F., MOHANDAS, N. & COOKE, B. M. 2002. Contribution of parasite proteins to altered mechanical properties of malaria-infected red blood cells. *Blood,* 99**,** 1060-1063.

GLENISTER, F. K., FERNANDEZ, K. M., KATS, L. M., HANSSEN, E., MOHANDAS, N., COPPEL, R. L. & COOKE, B. M. 2009. Functional alteration of red blood cells by a megadalton protein of Plasmodium falciparum. *Blood,* 113**,** 919-928.

GOEL, S., PALMKVIST, M., MOLL, K., JOANNIN, N., LARA, P., AKHOURI, R. R., MORADI, N., OJEMALM, K., WESTMAN, M., ANGELETTI, D., KJELLIN, H., LEHTIO, J., BLIXT, O., IDESTROM, L., GAHMBERG, C. G., STORRY, J. R., HULT, A. K., OLSSON, M. L., VON HEIJNE, G., NILSSON, I. & WAHLGREN, M. 2015. RIFINs are adhesins implicated in severe Plasmodium falciparum malaria. *Nature Medicine,* 21**,** 314-+.

GOLDBERG, D. E., SLATER, A. F. G., CERAMI, A. & HENDERSON, G. B. 1990. HEMOGLOBIN DEGRADATION IN THE MALARIA PARASITE PLASMODIUM-FALCIPARUM - AN ORDERED PROCESS IN A UNIQUE ORGANELLE. *Proceedings of the National Academy of Sciences of the United States of America,* 87**,** 2931-2935.

GRURING, C., HEIBER, A., KRUSE, F., FLEMMING, S., FRANCI, G., COLOMBO, S. F., FASANA, E., SCHOLER, H., BORGESE, N., STUNNENBERG, H. G., PRZYBORSKI, J. M., GILBERGER, T. W. & SPIELMANN, T. 2012. Uncovering Common Principles in Protein Export of Malaria Parasites. *Cell Host & Microbe,* 12**,** 717-729.

GUY, A. J., IRANI, V., MACRAILD, C. A., ANDERS, R. F., NORTON, R. S., BEESON, J. G., RICHARDS, J. S. & RAMSLAND, P. A. 2015. Insights into the Immunological Properties of Intrinsically Disordered Malaria Proteins Using Proteome Scale Predictions. *Plos One,* 10**,** 22.

HALDAR, K. & MOHANDAS, N. 2007. Erythrocyte remodeling by malaria parasites. *Current Opinion in Hematology,* 14**,** 203-209.

HALFMANN, R., ALBERTI, S., KRISHNAN, R., LYLE, N., O'DONNELL, C. W., KING, O. D., BERGER, B., PAPPU, R. V. & LINDQUIST, S. 2011. Opposing Effects of Glutamine and Asparagine Govern Prion Formation by Intrinsically Disordered Proteins. *Molecular Cell,* 43**,** 72-84.

HAMMOND, A., GALIZI, R., KYROU, K., SIMONI, A., SINISCALCHI, C., KATSANOS, D., GRIBBLE, M., BAKER, D., MAROIS, E., RUSSELL, S., BURT, A., WINDBICHLER, N., CRISANTI, A. & NOLAN, T. 2016. A CRISPR-Cas9 gene drive system-targeting female reproduction in the malaria mosquito vector Anopheles gambiae. *Nature Biotechnology,* 34**,** 78-83.

HANDAYANI, S., CHIU, D. T., TJITRA, E., KUO, J. S., LAMPAH, D., KENANGALEM, E., RENIA, L., SNOUNOU, G., PRICE, R. N., ANSTEY, N. M. & RUSSELL, B. 2009. High Deformability of Plasmodium vivax-Infected Red Blood Cells under Microfluidic Conditions. *Journal of Infectious Diseases,* 199**,** 445-450.

HANSSEN, E., SOUGRAT, R., FRANKLAND, S., DEED, S., KLONIS, N., LIPPINCOTT-SCHWARTZAND, J. & TILLEY, L. 2008. Electron tomography of the Maurer's cleft organelles of Plasmodium falciparum-infected erythrocytes reveals novel structural features. *Molecular Microbiology,* 67**,** 703-718.

HEIBER, A., KRUSE, F., PICK, C., GRURING, C., FLEMMING, S., OBERLI, A., SCHOELER, H., RETZLAFF, S., MESEN-RAMIREZ, P., HISS, J. A., KADEKOPPALA, M., HECHT, L., HOLDER, A. A., GILBERGER, T. W. & SPIELMANN, T. 2013. Identification of New PNEPs Indicates a Substantial Non-PEXEL Exportome and Underpins Common Features in Plasmodium falciparum Protein Export. *Plos Pathogens,* 9**,** 16.

HILLER, N. L., BHATTACHARJEE, S., VAN OOIJ, C., LIOLIOS, K., HARRISON, T., LOPEZ-ESTRANO, C. & HALDAR, K. 2004. A host-targeting signal in virulence proteins reveals a secretome in malarial infection. *Science,* 306**,** 1934-1937.

HIRAWAKE, H., KITA, K. & SHARMA, Y. D. 1997. Variations in the C-terminal repeats of the knob-associated histidine-rich protein of Plasmodium falciparum. *Biochimica Et Biophysica Acta-Molecular Basis of Disease,* 1360**,** 105-108.

HO, M., DAVIS, T. M. E., SILAMUT, K., BUNNAG, D. & WHITE, N. J. 1991. ROSETTE FORMATION OF PLASMODIUM-FALCIPARUM-INFECTED ERYTHROCYTES FROM PATIENTS WITH ACUTE MALARIA. *Infection and Immunity,* 59**,** 2135-2139.

HODDER, A. N., MAIER, A. G., RUG, M., BROWN, M., HOMMEL, M., PANTIC, I., PUIG-DE-MORALES-MARINKOVIC, M., SMITH, B., TRIGLIA, T., BEESON, J. & COWMAN, A. F. 2009. Analysis of structure and function of the giant protein Pf332 in Plasmodium falciparum. *Molecular Microbiology,* 71**,** 48-65.

HORROCKS, P., PINCHES, R. A., CHAKRAVORTY, S. J., PAPAKRIVOS, J., CHRISTODOULOU, Z., KYES, S. A., URBAN, B. C., FERGUSON, D. J. P. & NEWBOLD, C. I. 2005. PfEMP1 expression is reduced on the surface of knobless Plasmodium falciparum infected erythrocytes. *Journal of Cell Science,* 118**,** 2507-2518.

HOTEZ, P. J., WOC-COLBURN, L. & BOTTAZZI, M. E. 2014. Neglected tropical diseases in Central America and Panama: Review of their prevalence, populations at risk and impact on regional development. *International Journal for Parasitology,* 44**,** 597-603.

HOWARD, R. J., BARNWELL, J. W. & KAO, V. 1983. ANTIGENIC VARIATION IN PLASMODIUM-KNOWLESI MALARIA - IDENTIFICATION OF THE VARIANT ANTIGEN ON INFECTED ERYTHROCYTES. *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences,* 80**,** 4129-4133.

HUANG, S., AMALADOSS, A., LIU, M., CHEN, H. C., ZHANG, R., PREISER, P. R., DAO, M. & HAN, J. 2014. In Vivo Splenic Clearance Correlates with In Vitro Deformability of Red Blood Cells from Plasmodium yoelii-Infected Mice. *Infection and Immunity,* 82**,** 2532-2541.

HUGHES, A. L. 2004. The evolution of amino acid repeat arrays in Plasmodium and other organisms. *Journal of Molecular Evolution,* 59**,** 528-535.

JANSEN, A., GEMAYEL, R. & VERSTREPEN, K. J. 2012. Unstable Microsatellite Repeats Facilitate Rapid Evolution of Coding and Regulatory Sequences. *In:* GARRIDORAMOS, M. A. (ed.) *Repetitive DNA.* Basel: Karger.

JONES, M. L., DAS, S., BELDAL, H., COLLINS, C. R., BLACKMAN, M. J. & TREECK, M. 2016. A versatile strategy for rapid conditional genome engineering using loxP sites in a small synthetic intron in Plasmodium falciparum. *Scientific Reports,* 6**,** 9.

JONSON, I., OUGLAND, R., KLUNGLAND, A. & LARSEN, E. 2013a. Oxidative stress causes DNA triplet expansion in Huntington's disease mouse embryonic stem cells. *Stem Cell Research,* 11**,** 1264-1271.

JONSON, I., OUGLAND, R. & LARSEN, E. 2013b. DNA Repair Mechanisms in Huntington's Disease. *Molecular Neurobiology,* 47**,** 1093-1102.

JORDA, J. & KAJAVA, A. V. 2009. T-REKS: identification of Tandem REpeats in sequences with a K-meanS based algorithm. *Bioinformatics,* 25**,** 2632-2638.

JORDA, J., XUE, B., UVERSKY, V. N. & KAJAVA, A. V. 2010. Protein tandem repeats - the more perfect, the less structured. *Febs Journal,* 277**,** 2673-2682.

KAJAVA, A. V. 2012. Tandem repeats in proteins: From sequence to structure. *Journal of Structural Biology,* 179**,** 279-288.

KALANON, M., BARGIERI, D., STURM, A., MATTHEWS, K., GHOSH, S., GOODMAN, C. D., THIBERGE, S., MOLLARD, V., MCFADDEN, G. I., MENARD, R. & DE KONING-WARD, T. F. 2016. The Plasmodium translocon of exported proteins component EXP2 is critical for establishing a patent malaria infection in mice. *Cellular Microbiology,* 18**,** 399-412.

KASHI, Y. & KING, D. G. 2006. Simple sequence repeats as advantageous mutators in evolution. *Trends in Genetics,* 22**,** 253-259.

KATO, M., HAN, T. N. W., XIE, S. H., SHI, K., DU, X. L., WU, L. C., MIRZAEI, H., GOLDSMITH, E. J., LONGGOOD, J., PEI, J. M., GRISHIN, N. V., FRANTZ, D. E., SCHNEIDER, J. W., CHEN, S., LI, L., SAWAYA, M. R., EISENBERG, D., TYCKO, R. & MCKNIGHT, S. L. 2012. Cell-free Formation of RNA Granules: Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell,* 149**,** 753-767.

KATS, L. M., FERNANDEZ, K. M., GLENISTER, F. K., HERRMANN, S., BUCKINGHAM, D. W., SIDDIQUI, G., SHARMA, L., BAMERT, R., LUCET, I., GUILLOTTE, M., MERCEREAU-PUIJALON, O. & COOKE, B. M. 2014. An exported kinase (FIKK4.2) that mediates virulence-associated changes in Plasmodium falciparum-infected red blood cells. *International Journal for Parasitology,* 44**,** 319-328.

KATTI, M. V., SAMI-SUBBU, R., RANJEKAR, P. K. & GUPTA, V. S. 2000. Amino acid repeat patterns in protein sequences: Their diversity and structural-functional implications. *Protein Science,* 9**,** 1203-1209.

KAUR, P., SHARMA, P., KUMAR, A. & CHAUHAN, V. S. 1990. Synthetic, Immunological And Structural Studies on Repeat Unit Peptides of *Plasmodium-Falciparum* Antigens. *International Journal of Peptide and Protein Research,* 36**,** 515-521.

KEMP, D. J., COPPEL, R. L. & ANDERS, R. F. 1987. REPETITIVE PROTEINS AND GENES OF MALARIA. *Annual Review of Microbiology,* 41**,** 181-208.

KILILI, G. K. & LACOUNT, D. J. 2011. An Erythrocyte Cytoskeleton-Binding Motif in Exported Plasmodium falciparum Proteins. *Eukaryotic Cell,* 10**,** 1439-1447.

KIRKMAN, L. A., LAWRENCE, E. A. & DEITSCH, K. W. 2014. Malaria parasites utilize both homologous recombination and alternative end joining pathways to maintain genome integrity. *Nucleic Acids Research,* 42**,** 370-379.

KISHORE, S. P., PERKINS, S. L., TEMPLETON, T. J. & DEITSCH, K. W. 2009. An Unusual Recent Expansion of the C-Terminal Domain of RNA Polymerase II in Primate Malaria Parasites Features a Motif Otherwise Found Only in Mammalian Polymerases. *Journal of Molecular Evolution,* 68**,** 706-714.

KNUEPFER, E., RUG, M., KLONIS, N., TILLEY, L. & COWMAN, A. F. 2005. Trafficking determinants for PfEMP3 export and assembly under the Plasmodium falciparum-infected red blood cell membrane. *Molecular Microbiology,* 58**,** 1039-1053.

KOHLER, S., DELWICHE, C. F., DENNY, P. W., TILNEY, L. G., WEBSTER, P., WILSON, R. J. M., PALMER, J. D. & ROOS, D. S. 1997. A plastid of probable green algal origin in apicomplexan parasites. *Science,* 275**,** 1485-1489.

KRIEK, N., TILLEY, L., HORROCKS, P., PINCHES, R., ELFORD, B. C., FERGUSON, D. J. P., LINGELBACH, K. & NEWBOLD, C. I. 2003. Characterization of the pathway for transport of the cytoadherence-mediating protein, PfEMP1, to the host cell surface in malaria parasite-infected erythrocytes. *Molecular Microbiology,* 50**,** 1215-1227.

KUEHN, A. & PRADEL, G. 2010. The Coming-Out of Malaria Gametocytes. *Journal of Biomedicine and Biotechnology***,** 11.

KULZER, S., CHARNAUD, S., DAGAN, T., RIEDEL, J., MANDAL, P., PESCE, E. R., BLATCH, G. L., CRABB, B. S., GILSON, P. R. & PRZYBORSKI, J. M. 2012. Plasmodium falciparum-encoded exported hsp70/hsp40 chaperone/co-chaperone complexes within the host erythrocyte. *Cellular Microbiology,* 14**,** 1784-1795.

KWIATKOWSKI, D. P. 2005. How malaria has affected the human genome and what human genetics can teach us about malaria. *American Journal of Human Genetics,* 77**,** 171-192.

LACROIX, B., RYAN, J., DUMONT, J., MADDOX, P. S. & MADDOX, A. S. 2016. Identification of microtubule growth deceleration and its regulation by conserved and novel proteins. *Molecular Biology of the Cell,* 27**,** 1479-1487.

LANDERS, J. 1998. Armies of pestilence: the effects of pandemics on history. *Medical History,* 42**,** 128-128.

LASONDER, E., GREEN, J. L., CAMARDA, G., TALABANI, H., HOLDER, A. A., LANGSLEY, G. & ALANO, P. 2012. The Plasmodium falciparum Schizont Phosphoproteome Reveals Extensive Phosphatidylinositol and cAMP-Protein Kinase A Signaling. *Journal of Proteome Research,* 11**,** 5323-5337.

LAZARUS, M. D., SCHNEIDER, T. G. & TARASCHI, T. F. 2008. A new model for hemoglobin ingestion and transport by the human malaria parasite Plasmodium falciparum. *Journal of Cell Science,* 121**,** 1937-1949.

LE ROCH, K. G., JOHNSON, J. R., FLORENS, L., ZHOU, Y. Y., SANTROSYAN, A., GRAINGER, M., YAN, S. F., WILLIAMSON, K. C., HOLDER, A. A., CARUCCI, D. J., YATES, J. R. & WINZELER, E. A. 2004. Global analysis of transcript and protein levels across the Plasmodium falciparum life cycle. *Genome Research,* 14**,** 2308-2318.

LEVINE, N. D. 1971. UNIFORM TERMINOLOGY FOR PROTOZOAN SUBPHYLUM APICOMPLEXA. *Journal of Protozoology,* 18**,** 352-&.

LEVINSON, G. & GUTMAN, G. A. 1987. SLIPPED-STRAND MISPAIRING - A MAJOR MECHANISM FOR DNA-SEQUENCE EVOLUTION. *Molecular Biology and Evolution,* 4**,** 203-221.

LIN, Y. F. & WILSON, J. H. 2007. Transcription-induced CAG repeat contraction in human cells is mediated in part by transcription-coupled nucleotide excision repair. *Molecular and Cellular Biology,* 27**,** 6209-6217.

LING, I. T., FLORENS, L., DLUZEWSKI, A. R., KANEKO, O., GRAINGER, M., YIM LIM, B. Y., TSUBOI, T., HOPKINS, J. M., JOHNSON, J. R., TORII, M., BANNISTER, L. H., YATES, J. R., 3RD, HOLDER, A. A. & MATTEI, D. 2004. The Plasmodium falciparum clag9 gene encodes a rhoptry protein that is transferred to the host erythrocyte upon invasion. *Mol Microbiol,* 52**,** 107-18.

LINGELBACH, K. & JOINER, K. A. 1998. The parasitophorous vacuole membrane surrounding Plasmodium and Toxoplasma: An unusual compartment in infected cells. *Journal of Cell Science,* 111**,** 1467-1475.

LIU, J., GLUZMAN, I. Y., DREW, M. E. & GOLDBERG, D. E. 2005. The role of Plasmodium falciparum food vacuole plasmepsins. *Journal of Biological Chemistry,* 280**,** 1432-1437.

LIU, W., LI, Y., LEARN, G. H., RUDICELL, R. S., ROBERTSON, J. D., KEELE, B. F., NDJANGO, J.-B. N., SANZ, C. M., MORGAN, D. B., LOCATELLI, S., GONDER, M. K., KRANZUSCH, P. J., WALSH, P. D., DELAPORTE, E., MPOUDI-NGOLE, E., GEORGIEV, A. V., MULLER, M. N., SHAW, G. M., PEETERS, M., SHARP, P. M., RAYNER, J. C. & HAHN, B. H. 2010. Origin of the human malaria parasite Plasmodium falciparum in gorillas. *Nature,* 467**,** 420-425.

LOPES, S. C. P., ALBRECHT, L., CARVALHO, B. O., SIQUEIRA, A. M., THOMSON-LUQUE, R., NOGUEIRA, P. A., FERNANDEZ-BECERRA, C., DEL PORTILLO, H. A., RUSSELL, B. M.,

RENIA, L., LACERDA, M. V. G. & COSTA, F. T. M. 2014. Paucity of Plasmodium vivax Mature Schizonts in Peripheral Blood Is Associated With Their Increased Cytoadhesive Potential. *Journal of Infectious Diseases,* 209**,** 1403-1407.

LUSTIGMAN, S., ANDERS, R. F., BROWN, G. V. & COPPEL, R. L. 1990. THE MATURE-PARASITE-INFECTED ERYTHROCYTE SURFACE-ANTIGEN (MESA) OF PLASMODIUM-FALCIPARUM ASSOCIATES WITH THE ERYTHROCYTE-MEMBRANE SKELETAL PROTEIN, BAND-4.1. *Molecular and Biochemical Parasitology,* 38**,** 261-270.

MABASO, M. L. H., SHARP, B. & LENGELER, C. 2004. Historical review of malarial control in southern African with emphasis on the use of indoor residual house-spraying. *Tropical Medicine & International Health,* 9**,** 846-856.

MACPHERSON, G. G., WARRELL, M. J., WHITE, N. J., LOOAREESUWAN, S. & WARRELL, D. A. 1985. HUMAN CEREBRAL MALARIA - A QUANTITATIVE ULTRASTRUCTURAL ANALYSIS OF PARASITIZED ERYTHROCYTE SEQUESTRATION. *American Journal of Pathology,* 119**,** 385-401.

MAGOWAN, C., NUNOMURA, W., WALLER, K. L., YEUNG, J., LIANG, J., VAN DORT, H., LOW, P. S., COPPEL, R. L. & MOHANDAS, N. 2000. Plasmodium falciparum histidine-rich protein 1 associates with the band 3 binding domain of ankyrin in the infected red cell membrane. *Biochimica Et Biophysica Acta-Molecular Basis of Disease,* 1502**,** 461-470.

MAIER, A. G., BRAKS, J. A. M., WATERS, A. P. & COWMAN, A. F. 2006. Negative selection using yeast cytosine deaminase/uracil phosphoribosyl transferase in Plasmodium falciparum for targeted gene deletion by double crossover recombination. *Molecular and Biochemical Parasitology,* 150**,** 118-121.

MAIER, A. G., COOKE, B. M., COWMAN, A. F. & TILLEY, L. 2009. Malaria parasite proteins that remodel the host erythrocyte. *Nature Reviews Microbiology,* 7**,** 341-354.

MAIER, A. G., RUG, M., O'NEILL, M. T., BROWN, M., CHAKRAVORTY, S., SZESTAK, T., CHESSON, J., WU, Y., HUGHES, K., COPPEL, R. L., NEWBOLD, C., BEESON, J. G., CRAIG, A., CRABB, B. S. & COWMAN, A. F. 2008. Exported proteins required for virulence and rigidity of Plasmodium falciparum-infected human erythrocytes. *Cell,* 134**,** 48-61.

MANKELOW, T. J., SATCHWELL, T. J. & BURTON, N. M. 2012. Refined views of multi-protein complexes in the erythrocyte membrane. *Blood Cells Molecules and Diseases,* 49**,** 1-10.

MARTI, M., GOOD, R. T., RUG, M., KNUEPFER, E. & COWMAN, A. F. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. *Science,* 306**,** 1930-1933.

MARTI, M. & SPIELMANN, T. 2013. Protein export in malaria parasites: many membranes to cross. *Current Opinion in Microbiology,* 16**,** 445-451.

MATTEELLI, A., CALIGARIS, S., CASTELLI, F. & CAROSI, G. 1997. The placenta and malaria. *Annals of Tropical Medicine and Parasitology,* 91**,** 803-810.

MATTHEWS, K., KALANON, M., CHISHOLM, S. A., STURM, A., GOODMAN, C. D., DIXON, M. W. A., SANDERS, P. R., NEBL, T., FRASER, F., HAASE, S., MCFADDEN, G. I., GILSON, P. R., CRABB, B. S. & DE KONING-WARD, T. F. 2013. The Plasmodium translocon of exported proteins (PTEX) component thioredoxin-2 is important for maintaining normal blood-stage growth. *Molecular Microbiology,* 89**,** 1167-1186.

MAYER, C., SLATER, L., ERAT, M. C., KONRAT, R. & VAKONAKIS, I. 2012. Structural Analysis of the Plasmodium falciparum Erythrocyte Membrane Protein 1 (PfEMP1) Intracellular Domain Reveals a Conserved Interaction Epitope. *Journal of Biological Chemistry,* 287**,** 7182-7189.

MCFADDEN, G. I., REITH, M. E., MUNHOLLAND, J. & LANGUNNASCH, N. 1996. Plastid in human parasites. *Nature,* 381**,** 482-482.

MCHUGH, E., BATINOVIC, S., HANSSEN, E., MCMILLAN, P. J., KENNY, S., GRIFFIN, M. D. W., CRAWFORD, S., TRENHOLME, K. R., GARDINER, D. L., DIXON, M. W. A. & TILLEY, L. 2015. A repeat sequence domain of the ring-exported protein-1 of Plasmodium falciparum controls export machinery architecture and virulence protein trafficking. *Molecular Microbiology,* 98**,** 1101-1114.

MCMILLAN, P. J., MILLET, C., BATINOVIC, S., MAIORCA, M., HANSSEN, E., KENNY, S., MUHLE, R. A., MELCHER, M., FIDOCK, D. A., SMITH, J. D., DIXON, M. W. A. & TILLEY, L. 2013. Spatial and temporal mapping of the PfEMP1 export pathway in Plasmodium falciparum. *Cellular Microbiology,* 15**,** 1401-1418.

MCNAMARA, C. W., LEE, M. C. S., LIM, C. S., LIM, S. H., ROLAND, J., NAGLE, A., SIMON, O., YEUNG, B. K. S., CHATTERJEE, A. K., MCCORMACK, S. L., MANARY, M. J., ZEEMAN, A. M., DECHERING, K. J., KUMAR, T. R. S., HENRICH, P. P., GAGARING, K., IBANEZ, M., KATO, N., KUHEN, K. L., FISCHLI, C., ROTTMANN, M., PLOUFFE, D. M., BURSULAYA, B., MEISTER, S., RAMEH, L., TRAPPE, J., HAASEN, D., TIMMERMAN, M., SAUERWEIN, R. W., SUWANARUSK, R., RUSSELL, B., RENIA, L., NOSTEN, F., TULLY, D. C., KOCKEN, C. H. M., GLYNNE, R. J., BODENREIDER, C., FIDOCK, D. A., DIAGANA, T. T. & WINZELER, E. A. 2013. Targeting Plasmodium PI(4)K to eliminate malaria. *Nature,* 504**,** 248-+.

MEINHART, A., KAMENSKI, T., HOEPPNER, S., BAUMLI, S. & CRAMER, P. 2005. A structural perspective of CTD function. *Genes & Development,* 19**,** 1401-1415.

MENDES, T. A. O., LOBO, F. P., RODRIGUES, T. S., RODRIGUES-LUIZ, G. F., DAROCHA, W. D., FUJIWARA, R. T., TEIXEIRA, S. M. R. & BARTHOLOMEU, D. C. 2013. Repeat-Enriched Proteins Are Related to Host Cell Invasion and Immune Evasion in Parasitic Protozoa. *Molecular Biology and Evolution,* 30**,** 951-963.

MESEN-RAMIREZ, P., REINSCH, F., SOARES, A. B., BERGMANN, B., ULLRICH, A. K., TENZER, S. & SPIELMANN, T. 2016. Stable Translocation Intermediates Jam Global Protein Export in Plasmodium falciparum Parasites and Link the PTEX Component EXP2 with Translocation Activity. *Plos Pathogens,* 12**,** 28.

MIAN-MCCARTHY, S., AGNANDJI, S. T., LELL, B., FERNANDES, J. F., ABOSSOLO, B. P., METHOGO, B., KABWENDE, A. L., ADEGNIKA, A. A., MORDMULLER, B., ISSIFOU, S., KREMSNER, P. G., SACARLAL, J., AIDE, P., LANASPA, M., APONTE, J. J., MACHEVO, S., ACACIO, S., BULO, H., SIGAUQUE, B., MACETE, E., ALONSO, P., ABDULLA, S., SALIM, N., MINJA, R., MPINA, M., AHMED, S., ALI, A. M., MTORO, A. T., HAMAD, A. S., MUTANI, P., TANNER, M., TINTO, H., D'ALESSANDRO, U., SORGHO, H., VALEA, I., BIHOUN, B., GUIRAUD, I., KABORE, B., SOMBIE, O., GUIGUEMDE, R. T., OUEDRAOGO, J. B., HAMEL, M. J., KARIUKI, S., ONEKO, M., ODERO, C., OTIENO, K., AWINO, N., MCMORROW, M., MUTURI-KIOI, V., LASERSON, K. F., SLUTSKER, L., OTIENO, W., OTIENO, L., OTSYULA, N., GONDI, S., OTIENO, A., OWIRA, V., OGUK, E., ODONGO, G., BEN WOODS, J., OGUTU, B., NJUGUNA, P., CHILENGI, R., AKOO, P., KERUBO, C., MAINGI, C., LANG, T., OLOTU, A., BEJON, P., MARSH, K., MWANBINGU, G., OWUSU-AGYEI, S., ASANTE, K. P., OSEI-KWAKYE, K., BOAHEN, O., DOSOO, D., ASANTE, I., ADJEI, G., KWARA, E., CHANDRAMOHAN, D., GREENWOOD, B., LUSINGU, J., GESASE, S., MALABEJA, A., ABDUL, O., MAHENDE, C., LIHELUKA, E., MALLE, L., LEMNGE, M., THEANDER, T. G., DRAKELEY, C., ANSONG, D., AGBENYEGA, T., ADJEI, S., BOATENG, H. O., RETTIG, T., BAWA, J., SYLVERKEN, J., SAMBIAN, D., SARFO, A., et al. 2012. A Phase 3 Trial of RTS,S/AS01 Malaria Vaccine in African Infants. *New England Journal of Medicine,* 367**,** 2284-2295.

MILLER, L. H., MASON, S. J., CLYDE, D. F. & MCGINNISS, M. H. 1976. The resistance factor to Plasmodium vivax in blacks. The Duffy-blood-group genotype, FyFy. *N Engl J Med,* 295**,** 302-4.

MILLS, J. P., DIEZ-SILVA, M., QUINN, D. J., DAO, M., LANG, M. J., TAN, K. S. W., LIM, C. T., MILON, G., DAVID, P. H., MERCEREAU-PUIJALON, O., BONNEFOY, S. & SURESH, S. 2007. Effect of plasmodial RESA protein on deformability of human red blood cells harboring Plasmodium falciparum. *Proceedings of the National Academy of Sciences of the United States of America,* 104**,** 9213-9217.

MOHANDAS, N. & GALLAGHER, P. G. 2008. Red cell membrane: past, present, and future. *Blood,* 112**,** 3939-3948.

MOON, R. W., HALL, J., RANGKUTI, F., HO, Y. S., ALMOND, N., MITCHELL, G. H., PAIN, A., HOLDER, A. A. & BLACKMAN, M. J. 2013. Adaptation of the genetically tractable

malaria pathogen Plasmodium knowlesi to continuous culture in human erythrocytes. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 531-536.

MORENO-PEREZ, D. A., RUIZ, J. A. & PATARROYO, M. A. 2013. Reticulocytes: Plasmodium vivax target cells. *Biology of the Cell,* 105**,** 251-260.

MOSS, D. K., REMARQUE, E. J., FABER, B. W., CAVANAGH, D. R., ARNOT, D. E., THOMAS, A. W. & HOLDER, A. A. 2012. Plasmodium falciparum 19-Kilodalton Merozoite Surface Protein 1 (MSP1)-Specific Antibodies That Interfere with Parasite Growth In Vitro Can Inhibit MSP1 Processing, Merozoite Invasion, and Intracellular Parasite Development. *Infection and Immunity,* 80**,** 1280-1287.

MOXON, E. R., RAINEY, P. B., NOWAK, M. A. & LENSKI, R. E. 1994. ADAPTIVE EVOLUTION OF HIGHLY MUTABLE LOCI IN PATHOGENIC BACTERIA. *Current Biology,* 4**,** 24-33.

MUELLER, I., GALINSKI, M. R., BAIRD, J. K., CARLTON, J. M., KOCHAR, D. K., ALONSO, P. L. & DEL PORTILLO, H. A. 2009. Key gaps in the knowledge of Plasmodium vivax, a neglected human malaria parasite. *Lancet Infectious Diseases,* 9**,** 555-566.

MURALIDHARAN, V. & GOLDBERG, D. E. 2013. Asparagine Repeats in Plasmodium falciparum Proteins: Good for Nothing? *Plos Pathogens,* 9**,** 4.

MURALIDHARAN, V., OKSMAN, A., IWAMOTO, M., WANDLESS, T. J. & GOLDBERG, D. E. 2011. Asparagine repeat function in a Plasmodium falciparum protein assessed via a regulatable fluorescent affinity tag. *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 4411-4416.

MURALIDHARAN, V., OKSMAN, A., PAL, P., LINDQUIST, S. & GOLDBERG, D. E. 2012. Plasmodium falciparum heat shock protein 110 stabilizes the asparagine repeat-rich parasite proteome during malarial fevers. *Nature Communications,* 3**,** 10.

MURPHY, S. C., FERNANDEZ-POL, S., CHUNG, P. H., MURTHY, S. N. P., MILNE, S. B., SALOMAO, M., BROWN, H. A., LOMASNEY, J. W., MOHANDAS, N. & HALDAR, K. 2007. Cytoplasmic remodeling of erythrocyte raft lipids during infection by the human malaria parasite Plasmodium falciparum. *Blood,* 110**,** 2132-2139.

NAISSANT, B., DUPUY, F., DUFFIER, Y., LORTHIOIS, A., DUEZ, J., SCHOLZ, J., BUFFET, P., MERCKX, A., BACHMANN, A. & LAVAZEC, C. 2016. Plasmodium falciparum STEVOR phosphorylation regulates host erythrocyte deformability enabling malaria parasite transmission. *Blood,* 127**,** E42-E53.

NASH, G. B., OBRIEN, E., GORDONSMITH, E. C. & DORMANDY, J. A. 1989. Abnormalities In The Mechanical-Properties Of Red Blood-Cells Caused by Plasmodium-Falciparum. *Blood,* 74**,** 855-861.

NEAFSEY, D. E., JURASKA, M., BEDFORD, T., BENKESER, D., VALIM, C., GRIGGS, A., LIEVENS, M., ABDULLA, S., ADJEI, S., AGBENYEGA, T., AGNANDJI, S. T., AIDE, P., ANDERSON, S., ANSONG, D., APONTE, J. J., ASANTE, K. P., BEJON, P., BIRKETT, A. J., BRULS, M., CONNOLLY, K. M., D'ALESSANDRO, U., DOBANO, C., GESASE, S., GREENWOOD, B., GRIMSBY, J., TINTO, H., HAMEL, M. J., HOFFMAN, I., KAMTHUNZI, P., KARIUKI, S., KREMSNER, P. G., LEACH, A., LELL, B., LENNON, N. J., LUSINGU, J., MARSH, K., MARTINSON, F., MOLEL, J. T., MOSS, E. L., NJUGUNA, P., OCKENHOUSE, C. F., OGUTU, B. R., OTIENO, W., OTIENO, L., OTIENO, K., OWUSU-AGYEI, S., PARK, D. J., PELLE, K., ROBBINS, D., RUSS, C., RYAN, E. M., SACARLAL, J., SOGOLOFF, B., SORGHO, H., TANNER, M., THEANDER, T., VALEA, I., VOLKMAN, S. K., YU, Q., LAPIERRE, D., BIRREN, B. W., GILBERT, P. B. & WIRTH, D. F. 2015. Genetic Diversity and Protective Efficacy of the RTS, S/AS01 Malaria Vaccine. *New England Journal of Medicine,* 373**,** 2025-2037.

NEESEN, J., HEINLEIN, U. A. O., GLATZER, K. H. & BUNEMANN, H. 1999. Proteins with tandemly arranged repeats of a highly charged 16-amino-acid motif encoded by the Dhmst101 gene family are structural components of the outer sheath of the extremely elongated sperm tails of Drosophila hydei. *Development Growth & Differentiation,* 41**,** 93-99.

NEGHINA, R., NEGHINA, A. M., MARINCU, I. & IACOBICIU, I. 2010. Malaria, a Journey in Time: In Search of the Lost Myths and Forgotten Stories. *American Journal of the Medical Sciences,* 340**,** 492-498.

NELSON, D. L., ORR, H. T. & WARREN, S. T. 2013. The Unstable Repeats-Three Evolving Faces of Neurological Disease. *Neuron,* 77**,** 825-843.

NEWMAN, A. M. & COOPER, J. B. 2007. XSTREAM: A practical algorithm for identification and architecture modeling of tandem repeats in protein sequences. *Bmc Bioinformatics,* 8**,** 19.

NGUITRAGOOL, W., BOKHARI, A. A. B., PILLAI, A. D., RAYAVARA, K., SHARMA, P., TURPIN, B., ARAVIND, L. & DESAI, S. A. 2011. Malaria Parasite clag3 Genes Determine Channel-Mediated Nutrient Uptake by Infected Red Blood Cells. *Cell,* 145**,** 665-677.

NIANG, M., BEI, A. K., MADNANI, K. G., PELLY, S., DANKWA, S., KANJEE, U., GUNALAN, K., AMALADOSS, A., YEO, K. P., BOB, N. S., MALLERET, B., DURAISINGH, M. T. & PREISER, P. R. 2014. STEVOR Is a Plasmodium falciparum Erythrocyte Binding Protein that Mediates Merozoite Invasion and Rosetting. *Cell Host & Microbe,* 16**,** 81-93.

NIANG, M., YAN YAM, X. & PREISER, P. R. 2009. The Plasmodium falciparum STEVOR multigene family mediates antigenic variation of the infected erythrocyte. *PLoS Pathog,* 5**,** e1000307.

NKRUMAH, L. J., MUHLE, R. A., MOURA, P. A., GHOSH, P., HATFULL, G. F., JACOBS, W. R., JR. & FIDOCK, D. A. 2006. Efficient site-specific integration in Plasmodium falciparum chromosomes mediated by mycobacteriophage Bxb1 integrase. *Nat Methods,* 3**,** 615-21.

NOBLE, M., LEWIS, S. A. & COWAN, N. J. 1989. THE MICROTUBULE BINDING DOMAIN OF MICROTUBULE-ASSOCIATED PROTEIN MAP1B CONTAINS A REPEATED SEQUENCE MOTIF UNRELATED TO THAT OF MAP2 AND TAU. *Journal of Cell Biology,* 109**,** 3367-3376.

NOEDL, H., SE, Y., SCHAECHER, K., SMITH, B. L., SOCHEAT, D., FUKUDA, M. M. & CONSORTIUM, A. R. C. S. 2008. Evidence of Artemisinin-Resistant Malaria in Western Cambodia. *New England Journal of Medicine,* 359**,** 2619-2620.

NOTREDAME, C., HIGGINS, D. G. & HERINGA, J. 2000. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology,* 302**,** 205-217.

NUNES, M. C., GOLDRING, J. P. D., DOERIG, C. & SCHERF, A. 2007. A novel protein kinase family in Plasmodium falciparum is differentially transcribed and secreted to various cellular compartments of the host cell. *Molecular Microbiology,* 63**,** 391-403.

NUNES, M. C., OKADA, M., SCHEIDIG-BENATAR, C., COOKE, B. M. & SCHERF, A. 2010. Plasmodium falciparum FIKK Kinase Members Target Distinct Components of the Erythrocyte Membrane. *Plos One,* 5**,** 8.

OAKLEY, M. S. M. 2006. Molecular factors and biochemical pathways induced by febrile temperature in Plasmodium falciparum parasites. *American Journal of Tropical Medicine and Hygiene,* 75**,** 279-280.

OBERLI, A., SLATER, L. M., CUTTS, E., BRAND, F., MUNDWILER-PACHLATKO, E., RUSCH, S., MASIK, M. F. G., ERAT, M. C., BECK, H. P. & VAKONAKIS, I. 2014. A Plasmodium falciparum PHIST protein binds the virulence factor PfEMP1 and comigrates to knobs on the host cell surface. *Faseb Journal,* 28**,** 4420-4433.

OBERLI, A., ZURBRUGG, L., RUSCH, S., BRAND, F., BUTLER, M. E., DAY, J. L., CUTTS, E. E., LAVSTSEN, T., VAKONAKIS, I. & BECK, H. P. 2016. Plasmodium falciparum PHIST Proteins Contribute to Cytoadherence and Anchor PfEMP1 to the Host Cell Cytoskeleton. *Cell Microbiol*.

OEHRING, S. C., WOODCROFT, B. J., MOES, S., WETZEL, J., DIETZ, O., PULFER, A., DEKIWADIA, C., MAESER, P., FLUECK, C., WITMER, K., BRANCUCCI, N. M. B., NIEDERWIESER, I., JENOE, P., RALPH, S. A. & VOSS, T. S. 2012. Organellar proteomics reveals hundreds of novel nuclear proteins in the malaria parasite Plasmodium falciparum. *Genome Biology,* 13**,** 21.

OH, S. S., VOIGT, S., FISHER, D., YI, S. J., LEROY, P. J., DERICK, L. H., LIU, S. C. & CHISHTI, A. H. 2000. Plasmodium falciparum erythrocyte membrane protein 1 is anchored to the actin-spectrin junction and knob-associated histidine-rich protein in the erythrocyte skeleton. *Molecular and Biochemical Parasitology,* 108**,** 237-247.

ORLANDI, P. A., KLOTZ, F. W. & HAYNES, J. D. 1992. A MALARIA INVASION RECEPTOR, THE 175-KILODALTON ERYTHROCYTE BINDING ANTIGEN OF PLASMODIUM-FALCIPARUM RECOGNIZES THE TERMINAL NEU5AC(ALPHA-2-3)GAL- SEQUENCES OF GLYCOPHORIN-A. *Journal of Cell Biology,* 116**,** 901-909.

OTTO, T. D., RAYNER, J. C., BOHME, U., PAIN, A., SPOTTISWOODE, N., SANDERS, M., QUAIL, M., OLLOMO, B., RENAUD, F., THOMAS, A. W., PRUGNOLLE, F., CONWAY, D. J., NEWBOLD, C. & BERRIMAN, M. 2014. Genome sequencing of chimpanzee malaria parasites reveals possible pathways of adaptation to human hosts. *Nature Communications,* 5**,** 9.

PACHLATKO, E., RUSCH, S., MULLER, A., HEMPHILL, A., TILLEY, L., HANSSEN, E. & BECK, H. P. 2010. MAHRP2, an exported protein of Plasmodium falciparum, is an essential component of Maurer's cleft tethers. *Molecular Microbiology,* 77**,** 1136-1152.

PAIN, A., BOHME, U., BERRY, A. E., MUNGALL, K., FINN, R. D., JACKSON, A. P., MOURIER, T., MISTRY, J., PASINI, E. M., ASLETT, M. A., BALASUBRAMMANIAM, S., BORGWARDT, K., BROOKS, K., CARRET, C., CARVER, T. J., CHEREVACH, I., CHILLINGWORTH, T., CLARK, T. G., GALINSKI, M. R., HALL, N., HARPER, D., HARRIS, D., HAUSER, H., IVENS, A., JANSSEN, C. S., KEANE, T., LARKE, N., LAPP, S., MARTI, M., MOULE, S., MEYER, I. M., ORMOND, D., PETERS, N., SANDERS, M., SANDERS, S., SARGEANT, T. J., SIMMONDS, M., SMITH, F., SQUARES, R., THURSTON, S., TIVEY, A. R., WALKER, D., WHITE, B., ZUIDERWIJK, E., CHURCHER, C., QUAIL, M. A., COWMAN, A. F., TURNER, C. M. R., RAJANDREAM, M. A., KOCKEN, C. H. M., THOMAS, A. W., NEWBOLD, C. I., BARRELL, B. G. & BERRIMAN, M. 2008. The genome of the simian and human malaria parasite Plasmodium knowlesi. *Nature,* 455**,** 799-U7.

PANTALEO, A., FERRU, E., CARTA, F., MANNU, F., GIRIBALDI, G., VONO, R., LEPEDDA, A. J., PIPPIA, P. & TURRINI, F. 2011. Analysis of changes in tyrosine and serine phosphorylation of red cell membrane proteins induced by P. falciparum growth. *Proteomics Clinical Applications,* 5**,** 190-190.

PARISH, L. A., MAI, D. W., JONES, M. L., KITSON, E. L. & RAYNER, J. C. 2013. A member of the Plasmodium falciparum PHIST family binds to the erythrocyte cytoskeleton component band 4.1. *Malaria Journal,* 12**,** 9.

PASLOSKE, B. L. & HOWARD, R. J. 1994. MALARIA, THE RED-CELL, AND THE ENDOTHELIUM. *Annual Review of Medicine,* 45**,** 283-295.

PAYNE, D. 1987. SPREAD OF CHLOROQUINE RESISTANCE IN PLASMODIUM-FALCIPARUM. *Parasitology Today,* 3**,** 241-246.

PEI, X., GUO, X., COPPEL, R., MOHANDAS, N. & AN, X. 2007a. Plasmodium falciparum erythrocyte membrane protein 3 (PfEMP3) destabilizes erythrocyte membrane skeleton. *Journal of Biological Chemistry,* 282**,** 26754-26758.

PEI, X. H., AN, X. L., GUO, X. H., TARNAWSKI, M., COPPEL, R. & MOHANDAS, N. 2005. Structural and functional studies of interaction between Plasmodium falciparum knob-associated histidine-rich protein (KAHRP) and erythrocyte spectrin. *Journal of Biological Chemistry,* 280**,** 31166-31171.

PEI, X. H., GUO, X. H., COPPEL, R., BHATTACHARJEE, S., HALDAR, K., GRATZER, W., MOHANDAS, N. & AN, X. L. 2007b. The ring-infected erythrocyte surface antigen (RESA) of Plasmodium falciparum stabilizes spectrin tetramers and suppresses further invasion. *Blood,* 110**,** 1036-1042.

PELLE, K. G., OH, K., BUCHHOLZ, K., NARASIMHAN, V., JOICE, R., MILNER, D. A., BRANCUCCI, N. M. B., MA, S. Y., VOSS, T. S., KETMAN, K., SEYDEL, K. B., TAYLOR, T. E., BARTENEVA, N. S., HUTTENHOWER, C. & MARTI, M. 2015. Transcriptional profiling defines dynamics of parasite tissue sequestration during malaria infection. *Genome Medicine,* 7**,** 20.

PERKINS, S. L. & SCHALL, J. J. 2002. A molecular phylogeny of malarial parasites recovered from cytochrome b gene sequences. *Journal of Parasitology,* 88**,** 972-978.

PERUTZ, M. F., POPE, B. J., OWEN, D., WANKER, E. E. & SCHERZINGER, E. 2002. Aggregation of proteins with expanded glutamine and alanine repeats of the glutamine-rich and asparagine-rich domains of Sup35 and of the amyloid beta-peptide of amyloid plaques. *Proceedings of the National Academy of Sciences of the United States of America,* 99**,** 5596-5600.

PETERSEN, T. N., BRUNAK, S., VON HEIJNE, G. & NIELSEN, H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods,* 8**,** 785-786.

PETERSEN, W., MATUSCHEWSKI, K. & INGMUNDSON, A. 2015. Trafficking of the signature protein of intra-erythrocytic Plasmodium berghei-induced structures, IBIS1, to P. falciparum Maurer's clefts. *Molecular and Biochemical Parasitology,* 200**,** 25-29.

PHILIPPS, D., CELOTTO, A. M., WANG, Q. Q., TARNG, R. S. & GRAVELEY, B. R. 2003. Arginine/serine repeats are sufficient to constitute a splicing activation domain. *Nucleic Acids Research,* 31**,** 6502-6508.

POINAR, G. 2005. Plasmodium dominicana n. sp (Plasmodiidae : Haemospororida) from Tertiary Dominican amber. *Systematic Parasitology,* 61**,** 47-52.

POLSON, H. E. J. & BLACKMAN, M. J. 2005. A role for poly (dA)poly(dT) tracts in directing activity of the Plasmodium falciparum calmodulin gene promoter. *Molecular and Biochemical Parasitology,* 141**,** 179-189.

PORTUGAL, S., TIPTON, C. M., SOHN, H., KONE, Y., WANG, J., LI, S. P., SKINNER, J., VIRTANEVA, K., STURDEVANT, D. E., PORCELLA, S. F., DOUMBO, O. K., DOUMBO, S., KAYENTAO, K., ONGOIBA, A., TRAORE, B., SANZ, I., PIERCE, S. K. & CROMPTON, P. D. 2015. Malaria-associated atypical memory B cells exhibit markedly reduced B cell receptor signaling and effector function. *Elife,* 4**,** 21.

PROELLOCKS, N. I., HERRMANN, S., BUCKINGHAM, D. W., HANSSEN, E., HODGES, E. K., ELSWORTH, B., MORAHAN, B. J., COPPEL, R. L. & COOKE, B. M. 2014. A lysine-rich membrane-associated PHISTb protein involved in alteration of the cytoadhesive properties of Plasmodium falciparum-infected red blood cells. *Faseb Journal,* 28**,** 3103-3113.

PRZYBORSKI, J. M., WICKERT, H., KROHNE, G. & LANZER, M. 2003. Maurer's clefts - a novel secretory organelle? *Molecular and Biochemical Parasitology,* 132**,** 17-26.

RALPH, S. A., BISCHOFF, E., MATTEI, D., SISMEIRO, O., DILLIES, M. A., GUIGON, G., COPPEE, J. Y., DAVID, P. H. & SCHERF, A. 2005. Transcriptome analysis of antigenic variation in Plasmodium falciparum-var silencing is not dependent on antisense RNA. *Genome Biology,* 6**,** 12.

RAMBAUT, A. 2014. *FigTree v1.4.2* [Online]. Available: http://tree.bio.ed.ac.uk/software/figtree/ [Accessed].

RIGLAR, D. T., RICHARD, D., WILSON, D. W., BOYLE, M. J., DEKIWADIA, C., TURNBULL, L., ANGRISANO, F., MARAPANA, D. S., ROGERS, K. L., WHITCHURCH, C. B., BEESON, J. G., COWMAN, A. F., RALPH, S. A. & BAUM, J. 2011. Super-Resolution Dissection of Coordinated Events during Malaria Parasite Invasion of the Human Erythrocyte. *Cell Host & Microbe,* 9**,** 9-20.

RIGLAR, D. T., ROGERS, K. L., HANSSEN, E., TURNBULL, L., BULLEN, H. E., CHARNAUD, S. C., PRZYBORSKI, J., GILSON, P. R., WHITCHURCH, C. B., CRABB, B. S., BAUM, J. & COWMAN, A. F. 2013. Spatial association with PTEX complexes defines regions for effector export into Plasmodium falciparum-infected erythrocytes. *Nature Communications,* 4**,** 13.

ROSONINA, E. & BLENCOWE, B. J. 2004. Analysis of the requirement for RNA polymerase IICTD heptapeptide repeats in pre-mRNA splicing and 3 '-end cleavage. *Rna-a Publication of the Rna Society,* 10**,** 581-589.

ROUSSEAUX-PREOST, R., LECUYER, C., DROBECQ, H., SERGHERAERT, C., DACHEUX, J. L. & ROUSSEAUX, J. 2003. Characterization of boar sperm cytoskeletal cylicin II as an actin-binding protein. *Biochemical and Biophysical Research Communications,* 303**,** 182-189.

RUG, M., CYRKLAFF, M., MIKKONEN, A., LEMGRUBER, L., KUELZER, S., SANCHEZ, C. P., THOMPSON, J., HANSSEN, E., O'NEILL, M., LANGER, C., LANZER, M., FRISCHKNECHT, F., MAIER, A. G. & COWMAN, A. F. 2014. Export of virulence proteins by malaria-infected erythrocytes involves remodeling of host actin cytoskeleton. *Blood,* 124**,** 3459-3468.

RUG, M., PRESCOTT, S. W., FERNANDEZ, K. M., COOKE, B. M. & COWMAN, A. F. 2006. The role of KAHRP domains in knob formation and cytoadherence of P falciparum-infected human erythrocytes. *Blood,* 108**,** 370-378.

RUSSO, I., BABBITT, S., MURALIDHARAN, V., BUTLER, T., OKSMAN, A. & GOLDBERG, D. E. 2010. Plasmepsin V licenses Plasmodium proteins for export into the host erythrocyte. *Nature,* 463**,** 632-636.

SAFEUKUI, I., CORREAS, J. M., BROUSSE, V., HIRT, D., DEPLAINE, G., MULE, S., LESURTEL, M., GOASGUEN, N., SAUVANET, A., COUVELARD, A., KERNEIS, S., KHUN, H., VIGAN-WOMAS, I., OTTONE, C., MOLINA, T. J., TRELUYER, J. M., MERCEREAU-PUIJALON, O., MILON, G., DAVID, P. H. & BUFFET, P. A. 2008. Retention of Plasmodium falciparum ring-infected erythrocytes in the slow, open microcirculation of the human spleen. *Blood,* 112**,** 2520-2528.

SALICHS, E., LEDDA, A., MULARONI, L., ALBA, M. M. & DE LA LUNA, S. 2009. Genome-Wide Analysis of Histidine Repeats Reveals Their Role in the Localization of Human Proteins to the Nuclear Speckles Compartment. *Plos Genetics,* 5**,** 18.

SARGEANT, T. J., MARTI, M., CALER, E., CARLTON, J. M., SIMPSON, K., SPEED, T. P. & COWMAN, A. F. 2006. Lineage-specific expansion of proteins exported to erythrocytes in malaria parasites. *Genome Biology,* 7**,** 22.

SCHERF, A., CARTER, R., PETERSEN, C., ALANO, P., NELSON, R., AIKAWA, M., MATTEI, D., DASILVA, L. P. & LEECH, J. 1992. Gene Inactivation Of Pf11-1 of *Plasmodium-falciparum* by Chromosome Breakage and Healing - Identification of a Gametocyte-Specific Protein With a Potential Role in Gametogenesis. *Embo Journal,* 11**,** 2293-2301.

SCHMIDT, H. A., STRIMMER, K., VINGRON, M. & VON HAESELER, A. 2002. TREE-PUZZLE: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics,* 18**,** 502-504.

SCHMIDT, H. B. & GORLICH, D. 2016. Transport Selectivity of Nuclear Pores, Phase Separation, and Membraneless Organelles. *Trends in biochemical sciences,* 41**,** 46-61.

SCHNEIDER, A. G. & MERCEREAU-PUIJALON, O. 2005. A new Apicomplexa-specific protein kinase family: multiple members in Plasmodium falciparum, all with an export signature. *Bmc Genomics,* 6**,** 12.

SCHOFIELD, L. 1991. On The Function of Repetitive Domains in Protein Antigens of Plasmodium and other Eukaryotic Parasites. *Parasitology Today,* 7**,** 99-105.

SCHULZE, J., KWIATKOWSKI, M., BORNER, J., SCHLUTER, H., BRUCHHAUS, I., BURMESTER, T., SPIELMANN, T. & PICK, C. 2015. The Plasmodium falciparum exportome contains non-canonical PEXEL/HT proteins. *Molecular Microbiology,* 97**,** 301-314.

SEDER, R. A., CHANG, L. J., ENAMA, M. E., ZEPHIR, K. L., SARWAR, U. N., GORDON, I. J., HOLMAN, L. A., JAMES, E. R., BILLINGSLEY, P. F., GUNASEKERA, A., RICHMAN, A., CHAKRAVARTY, S., MANOJ, A., VELMURUGAN, S., LI, M. L., RUBEN, A. J., LI, T., EAPPEN, A. G., STAFFORD, R. E., PLUMMER, S. H., HENDEL, C. S., NOVIK, L., COSTNER, P. J. M., MENDOZA, F. H., SAUNDERS, J. G., NASON, M. C., RICHARDSON, J. H., MURPHY, J., DAVIDSON, S. A., RICHIE, T. L., SEDEGAH, M., SUTAMIHARDJA, A., FAHLE, G. A., LYKE, K. E., LAURENS, M. B., ROEDERER, M., TEWARI, K., EPSTEIN, J. E., SIM, B. K. L., LEDGERWOOD, J. E., GRAHAM, B. S., HOFFMAN, S. L. & TEAM, V. R. C. S. 2013. Protection Against Malaria by Intravenous Immunization with a Nonreplicating Sporozoite Vaccine. *Science,* 341**,** 1359-1365.

SIM, B. K., TOYOSHIMA, T., HAYNES, J. D. & AIKAWA, M. 1992. Localization of the 175-kilodalton erythrocyte binding antigen in micronemes of Plasmodium falciparum merozoites. *Mol Biochem Parasitol,* 51**,** 157-9.

SINGH, B., SUNG, L. K., MATUSOP, A., RADHAKRISHNAN, A., SHAMSUL, S. S. G., COX-SINGH, J., THOMAS, A. & CONWAY, D. J. 2004a. A large focus of naturally acquired Plasmodium knowlesi infections in human beings. *Lancet,* 363**,** 1017-1024.

SINGH, G. P., CHANDRA, B. R., BHATTACHARYA, A., AKHOURI, R. R., SINGH, S. K. & SHARMA, A. 2004b. Hyper-expansion of asparagines correlates with an abundance of proteins with prion-like domains in Plasmodium falciparum. *Molecular and Biochemical Parasitology,* 137**,** 307-319.

SINGH, S., ALAM, M. M., PAL-BHOWMICK, I., BRZOSTOWSKI, J. A. & CHITNIS, C. E. 2010. Distinct External Signals Trigger Sequential Release of Apical Organelles during Erythrocyte Invasion by Malaria Parasites. *Plos Pathogens,* 6**,** 14.

SMITH, J. D., CHITNIS, C. E., CRAIG, A. G., ROBERTS, D. J., HUDSONTAYLOR, D. E., PETERSON, D. S., PINCHES, R., NEWBOLD, C. I. & MILLER, L. H. 1995. SWITCHES IN EXPRESSION OF PLASMODIUM-FALCIPARUM VAR GENES CORRELATE WITH CHANGES IN ANTIGENIC AND CYTOADHERENT PHENOTYPES OF INFESTED ERYTHROCYTES. *Cell,* 82**,** 101-110.

SNIDER, C., JAYASINGHE, S., HRISTOVA, K. & WHITE, S. H. 2009. MPEx: A tool for exploring membrane proteins. *Protein Science,* 18**,** 2624-2628.

SNOUNOU, G., VIRIYAKOSOL, S., JARRA, W., THAITHONG, S. & BROWN, K. N. 1993. IDENTIFICATION OF THE 4 HUMAN MALARIA PARASITE SPECIES IN FIELD SAMPLES BY THE POLYMERASE CHAIN-REACTION AND DETECTION OF A HIGH PREVALENCE OF MIXED INFECTIONS. *Molecular and Biochemical Parasitology,* 58**,** 283-292.

SPIELMANN, T. & GILBERGER, T. W. 2010. Protein export in malaria parasites: do multiple export motifs add up to multiple export pathways? *Trends in Parasitology,* 26**,** 6-10.

SPILLMAN, N. J., BECK, J. R. & GOLDBERG, D. E. 2015. Protein Export into Malaria Parasite-Infected Erythrocytes: Mechanisms and Functional Consequences. *In:* KORNBERG, R. D. (ed.) *Annual Review of Biochemistry, Vol 84.* Palo Alto: Annual Reviews.

SPYCHER, C., RUG, M., PACHLATKO, E., HANSSEN, E., FERGUSON, D., COWMAN, A. F., TILLEY, L. & BECK, H.-P. 2008. The Maurer's cleft protein MAHRP1 is essential for trafficking of PfEMP1 to the surface of Plasmodium falciparum-infected erythrocytes. *Molecular Microbiology,* 68**,** 1300-1314.

STANISIC, D. I., GERRARD, J., FINK, J., GRIFFIN, P. M., LIU, X. Q., SUNDAC, L., SEKULOSKI, S., RODRIGUEZ, I. B., PINGNET, J., YANG, Y., ZHOU, Y., TRENHOLME, K. R., WANG, C. Y., HACKETT, H., CHAN, J. A., LANGER, C., HANSSEN, E., HOFFMAN, S. L., BEESON, J. G., MCCARTHY, J. S. & GOOD, M. F. 2016. Infectivity of Plasmodium falciparum in malaria-naive individuals is related to knob expression and cytoadherence of the parasite. *Infect Immun*.

STEINERT, P. M., PARRY, D. A. D. & MAREKOV, L. N. 2003. Trichohyalin mechanically strengthens the hair follicle - Multiple cross-bridging roles in the inner root sheath. *Journal of Biological Chemistry,* 278**,** 41409-41419.

STEWART, M. J., NAWROT, R. J., SCHULMAN, S. & VANDERBERG, J. P. 1986. PLASMODIUM-BERGHEI SPOROZOITE INVASION IS BLOCKED INVITRO BY SPOROZOITE-IMMOBILIZING ANTIBODIES. *Infection and Immunity,* 51**,** 859-864.

STOTHARD, P. 2000. The sequence manipulation suite: JavaScript programs for analyzing and formatting protein and DNA sequences. *Biotechniques,* 28**,** 1102-+.

SUNDARARAMAN, S. A., PLENDERLEITH, L. J., LIU, W., LOY, D. E., LEARN, G. H., LI, Y., SHAW, K. S., AYOUBA, A., PEETERS, M., SPEEDE, S., SHAW, G. M., BUSHMAN, F. D., BRISSON, D., RAYNER, J. C., SHARP, P. M. & HAHN, B. H. 2016. Genomes of cryptic chimpanzee Plasmodium species reveal key evolutionary events leading to human malaria. *Nat Commun,* 7**,** 11078.

SUWANARUSK, R., COOKE, B. M., DONDORP, A. M., SILAMUT, K., SATTABONGKOT, J., WHITE, N. J. & UDOMSANGPETCH, R. 2004. The deformability of red blood cells parasitized by Plasmodium falciparum and P-vivax. *Journal of Infectious Diseases,* 189**,** 190-194.

SZAFRANSKI, K., LEHMANN, R., PARRA, G., GUIGO, R. & GLOCKNER, G. 2005. Gene organization features in A/T-rich organisms. *Journal of Molecular Evolution,* 60**,** 90-98.

TAN, J. C., TAN, A., CHECKLEY, L., HONSA, C. M. & FERDIG, M. T. 2010. Variable Numbers of Tandem Repeats in Plasmodium falciparum Genes. *Journal of Molecular Evolution,* 71**,** 268-278.

TANAKA, T. Q., DEHDASHTI, S. J., NGUYEN, D. T., MCKEW, J. C., ZHENG, W. & WILLIAMSON, K. C. 2013. A quantitative high throughput assay for identifying gametocytocidal compounds. *Molecular and Biochemical Parasitology,* 188**,** 20-25.

TARCSA, E., MAREKOV, L. N., ANDREOLI, J., IDLER, W. W., CANDI, E., CHUNG, S. I. & STEINERT, P. M. 1997. The fate of trichohyalin - Sequential, post-translational modifications by peptidyl-arginine deiminase and transglutaminases. *Journal of Biological Chemistry,* 272**,** 27893-27901.

TARR, S. J., CRYAR, A., THALASSINOS, K., HALDAR, K. & OSBORNE, A. R. 2013. The C-terminal portion of the cleaved HT motif is necessary and sufficient to mediate export of proteins from the malaria parasite into its host cell. *Molecular Microbiology,* 87**,** 835-850.

TARR, S. J., MOON, R. W., HARDEGE, I. & OSBORNE, A. R. 2014. A conserved domain targets exported PHISTb family proteins to the periphery of Plasmodium infected erythrocytes. *Molecular and Biochemical Parasitology,* 196**,** 29-40.

TAYLOR, D. W., PARRA, M., CHAPMAN, G. B., STEARNS, M. E., RENER, J., AIKAWA, M., UNI, S., ALEY, S. B., PANTON, L. J. & HOWARD, R. J. 1987. LOCALIZATION OF PLASMODIUM-FALCIPARUM HISTIDINE-RICH PROTEIN-1 IN THE ERYTHROCYTE SKELETON UNDER KNOBS. *Molecular and Biochemical Parasitology,* 25**,** 165-174.

THAM, W. H., HEALER, J. & COWMAN, A. F. 2012. Erythrocyte and reticulocyte binding-like proteins of Plasmodium falciparum. *Trends in Parasitology,* 28**,** 23-30.

TIBURCIO, M., NIANG, M., DEPLAINE, G., PERROT, S., BISCHOFF, E., NDOUR, P. A., SILVESTRINI, F., KHATTAB, A., MILON, G., DAVID, P. H., HARDEMAN, M., VERNICK, K. D., SAUERWEIN, R. W., PREISER, P. R., MERCEREAU-PUIJALON, O., BUFFET, P., ALANO, P. & LAVAZEC, C. 2012. A switch in infected erythrocyte deformability at the maturation and blood circulation of Plasmodium falciparum transmission stages. *Blood,* 119**,** E172-E180.

TINTO, H., D'ALESSANDRO, U., SORGHO, H., VALEA, I., TAHITA, M. C., KABORE, W., KIEMDE, F., LOMPO, P., OUEDRAOGO, S., DERRA, K., OUEDRAOGO, F., OUEDRAOGO, J. B., BALLOU, W. R., COHEN, J., GUERRA, Y., HEERWEGH, D., JONGERT, E., LAPIERRE, D., LEACH, A., LIEVENS, M., OFORI-ANYINAM, O., OLIVIER, A., VEKEMANS, J., AGNANDJI, S. T., LELL, B., FERNANDES, J. F., ABOSSOLO, B. P., KABWENDE, A. L., ADEGNIKA, A. A., MORDMULLER, B., ISSIFOU, S., KREMSNER, P. G., LOEMBE, M. M., BACHE, E., ALABI, A., OWUSU-AGYEI, S., ASANTE, K. P., BOAHEN, O., DOSOO, D., ASANTE, I., YIDANA, Z., ANIM, J., ADENIJI, E., YAWSON, A. K., KAYAN, K., CHANDRAMOHAN, D., GREENWOOD, B., ANSONG, D., AGBENYEGA, T., ADJEI, S., BOATENG, H. O., RETTIG, T., SYLVERKEN, J., SAMBIAN, D., BADU-PREPAH, A., KOTEY, A., BUABENG, P., PAINTSIL, V., ENIMIL, A., HAMEL, M. J., KARIUKI, S., ONEKO, M., ODERO, C., OTIENO, K., AWINO, N., MUTURI-KIOI, V., OMOTO, J., SANG, T., ODHIAMBO, S., LASERSON, K. F., SLUTSKER, L., OTIENO, W., OTIENO, L., OTSYULA, N., GONDI, S., OCHOLA, J., OKOTH, G., MABUNDE, D. C., WANGWE, A., OTIENO, A., OYIEKO, J., COWDEN, J., OGUTU, B., NJUGUNA, P., MARSH, K., AKOO, P., KERUBO, C., MAINGI, C., BEJON, P., OLOTU, A., CHILENGI, R., TSOFA, B., LANG, T., GITAKA, J., AWUONDO, K., MARTINSON, F., HOFFMAN, I., MVALO, T., KAMTHUNZI, P., NKOMO, R., et al. 2015. Efficacy and safety of RTS,S/AS01 malaria vaccine with or without a booster dose in infants and children in Africa: final results of a phase 3, individually randomised, controlled trial. *Lancet,* 386**,** 31-45.

TOLL-RIERA, M., RADO-TRILLA, N., MARTYS, F. & ALBA, M. M. 2012. Role of Low-Complexity Sequences in the Formation of Novel Protein Coding Sequences. *Molecular Biology and Evolution,* 29**,** 883-886.

TRAGER, W. & JENSEN, J. B. 1976. HUMAN MALARIA PARASITES IN CONTINUOUS CULTURE. *Science,* 193**,** 673-675.

TRIGLIA, T., STAHL, H. D., CREWTHER, P. E., SCANLON, D., BROWN, G. V., ANDERS, R. F. & KEMP, D. J. 1987. The Complete Sequence of the Gene For The Knob-Associated Histidine-Rich Protein From *Plasmodium-falciparum*. *Embo Journal,* 6**,** 1413-1419.

TRIGLIA, T., STAHL, H. D., CREWTHER, P. E., SILVA, A., ANDERS, R. F. & KEMP, D. J. 1988. Structure of a Plasmodium-falciparum Gene That Encodes a Glutamic Acid-Rich Protein (GARP). *Molecular and Biochemical Parasitology,* 31**,** 199-201.

TRIGLIA, T., THOMPSON, J., CARUANA, S. R., DELORENZI, M., SPEED, T. & COWMAN, A. F. 2001. Identification of proteins from Plasmodium falciparum that are homologous to reticulocyte binding proteins in Plasmodium vivax. *Infect Immun,* 69**,** 1084-92.

TU, Y. Y. 2011. The discovery of artemisinin (qinghaosu) and gifts from Chinese medicine. *Nature Medicine,* 17**,** 1217-1220.

UDAGAMA, P. V., ATKINSON, C. T., PEIRIS, J. S. M., DAVID, P. H., MENDIS, K. N. & AIKAWA, M. 1988. IMMUNOELECTRON MICROSCOPY OF SCHUFFNER DOTS IN PLASMODIUM-VIVAX-INFECTED HUMAN-ERYTHROCYTES. *American Journal of Pathology,* 131**,** 48-52.

UKAEGBU, U. E. & DEITSCH, K. W. 2015. The Emerging Role for RNA Polymerase II in Regulating Virulence Gene Expression in Malaria Parasites. *Plos Pathogens,* 11**,** 7.

VAN OOIJ, C., WITHERS-MARTINEZ, C., RINGEL, A., COCKCROFT, S., HALDAR, K. & BLACKMAN, M. J. 2013. Identification of a Plasmodium falciparum Phospholipid Transfer Protein. *Journal of Biological Chemistry,* 288**,** 31971-31983.

VERRA, F. & HUGHES, A. L. 1999. Biased amino acid composition in repeat regions of Plasmodium antigens. *Molecular Biology and Evolution,* 16**,** 627-633.

VERSTREPEN, K. J., JANSEN, A., LEWITTER, F. & FINK, G. R. 2005. Intragenic tandem repeats generate functional variability. *Nature Genetics,* 37**,** 986-990.

VIEBIG, N. K., GAMAIN, B., SCHEIDIG, C., LEPOLARD, C., PRZYBORSKI, J., LANZER, M., GYSIN, J. & SCHERF, A. 2005. A single member of the Plasmodium falciparum var multigene family determines cytoadhesion to the placental receptor chondroitin sulphate A. *Embo Reports,* 6**,** 775-781.

VIGNALI, M., ARMOUR, C. D., CHEN, J. Y., MORRISON, R., CASTLE, J. C., BIERY, M. C., BOUZEK, H., MOON, W., BABAK, T., FRIED, M., RAYMOND, C. K. & DUFFY, P. E. 2011. NSR-seq transcriptional profiling enables identification of a gene signature of Plasmodium falciparum parasites infecting children. *Journal of Clinical Investigation,* 121**,** 1119-1129.

VINCENSINI, L., FALL, G., BERRY, L., BLISNICK, T. & BRAUN BRETON, C. 2008. The RhopH complex is transferred to the host cell cytoplasm following red blood cell invasion by Plasmodium falciparum. *Mol Biochem Parasitol,* 160**,** 81-9.

VITOR-SILVA, S., REYES-LECCA, R. C., PINHEIRO, T. R. A. & LACERDA, M. V. G. 2009. Malaria is associated with poor school performance in an endemic area of the Brazilian Amazon. *Malaria Journal,* 8**,** 7.

WALLER, K. L., COOKE, B. M., NUNOMURA, W., MOHANDAS, N. & COPPEL, R. L. 1999. Mapping the binding domains involved in the interaction between the Plasmodium falciparum knob-associated histidine-rich protein (KAHRP) and the cytoadherence ligand P-falciparum erythrocyte membrane protein 1 (PfEMP1). *Journal of Biological Chemistry,* 274**,** 23808-23813.

WALLER, K. L., MCBRIDE, S. M., KIM, K. & MCDONALD, T. V. 2008. Characterization of two putative potassium channels in Plasmodium falciparum. *Malaria Journal,* 7**,** 11.

WALLER, K. L., NUNOMURA, W., AN, X. L., COOKE, B. M., MOHANDAS, N. & COPPEL, R. L. 2003. Mature parasite-infected erythrocyte surface antigen (MESA) of Plasmodium

falciparum binds to the 30-kDa domain of protein 4.1 in malaria-infected red blood cells. *Blood,* 102**,** 1911-1914.

WALLER, K. L., STUBBERFIELD, L. M., DUBLJEVIC, V., BUCKINGHAM, D. W., MOHANDAS, N., COPPEL, R. L. & COOKE, B. M. 2010. Interaction of the exported malaria protein Pf332 with the red blood cell membrane skeleton. *Biochimica Et Biophysica Acta-Biomembranes,* 1798**,** 861-871.

WALLER, K. L., STUBBERFIELD, L. M., DUBLJEVIC, V., NUNOMURA, W., AN, X., MASON, A. J., MOHANDAS, N., COOKE, B. M. & COPPEL, R. L. 2007a. Interactions of Plasmodium falciparum erythrocyte membrane protein 3 with the red blood cell membrane skeleton. *Biochim Biophys Acta,* 1768**,** 2145-56.

WALLER, K. L., STUBBERFIELD, L. M., DUBLJEVIC, V., NUNOMURA, W., AN, X. L., MASON, A. J., MOHANDAS, N., COOKE, B. M. & COPPEL, R. L. 2007b. Interactions of Plasmodium falciparum erythrocyte membrane protein 3 with the red blood cell membrane skeleton. *Biochimica Et Biophysica Acta-Biomembranes,* 1768**,** 2145-2156.

WANG, J., ZHANG, C. J., CHIA, W. N., LOH, C. C. Y., LI, Z. J., LEE, Y. M., HE, Y. K., YUAN, L. X., LIM, T. K., LIU, M., LIEW, C. X., LEE, Y. Q., ZHANG, J. B., LU, N. C., LIM, C. T., HUA, Z. C., LIU, B., SHEN, H. M., TAN, K. S. W. & LIN, Q. S. 2015. Haem-activated promiscuous targeting of artemisinin in Plasmodium falciparum. *Nature Communications,* 6**,** 11.

WARD, J. J., MCGUFFIN, L. J., BRYSON, K., BUXTON, B. F. & JONES, D. T. 2004. The DISOPRED server for the prediction of protein disorder. *Bioinformatics,* 20**,** 2138-2139.

WATERHOUSE, A. M., PROCTER, J. B., MARTIN, D. M. A., CLAMP, M. & BARTON, G. J. 2009. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics,* 25**,** 1189-1191.

WATERKEYN, J. G., WICKHAM, M. E., DAVERN, K. M., COOKE, B. M., COPPEL, R. L., REEDER, J. C., CULVENOR, J. G., WALLER, R. F. & COWMAN, A. F. 2000. Targeted mutagenesis of Plasmodium falciparum erythrocyte membrane protein 3 (PfEMP3) disrupts cytoadherence of malaria-infected red blood cells. *Embo Journal,* 19**,** 2813-2823.

WATERMEYER, J. M., HALE, V. L., HACKETT, F., CLARE, D. K., CUTTS, E. E., VAKONAKIS, I., FLECK, R. A., BLACKMAN, M. J. & SAIBIL, H. R. 2016. A spiral scaffold underlies cytoadherent knobs in Plasmodium falciparum–infected erythrocytes. *Blood,* 127**,** 343-351.

WENG, H. B., GUO, X. H., PAPOIN, J., WANG, J., COPPEL, R., MOHANDAS, N. & AN, X. L. 2014. Interaction of Plasmodium falciparum knob-associated histidine-rich protein (KAHRP) with erythrocyte ankyrin R is required for its attachment to the erythrocyte membrane. *Biochimica Et Biophysica Acta-Biomembranes,* 1838**,** 185-192.

WHEELER, V. C., PERSICHETTI, F., MCNEIL, S. M., MYSORE, J. S., MYSORE, S. S., MACDONALD, M. E., MYERS, R. H., GUSELLA, J. F., WEXLER, N. S. & GRP, U. S. V. C. 2007. Factors associated with HD CAG repeat instability in Huntington disease. *Journal of Medical Genetics,* 44**,** 695-701.

WHO 2015. World Malaria Report 2015. Geneva, Switzerland: World Health Organisation.

WICKHAM, M. E., RUG, M., RALPH, S. A., KLONIS, N., MCFADDEN, G. I., TILLEY, L. & COWMAN, A. F. 2001. Trafficking and assembly of the cytoadherence complex in Plasmodium falciparum-infected human erythrocytes. *Embo Journal,* 20**,** 5636-5649.

WOODWARD, R. B. & DOERING, W. E. 1944. The total synthesis of quinine. *Journal of the American Chemical Society,* 66**,** 849-849.

WRIGHT, G. J. & RAYNER, J. C. 2014. Plasmodium falciparum Erythrocyte Invasion: Combining Function with Immune Evasion. *Plos Pathogens,* 10**,** 7.

WU, Y., NELSON, M. M., QUAILE, A., XIA, D., WASTLING, J. M. & CRAIG, A. 2009. Identification of phosphorylated proteins in erythrocytes infected by the human malaria parasite Plasmodium falciparum. *Malaria Journal,* 8**,** 16.

ZHANG, T. Y., HUANG, J., GU, L. Y. & LI, G. M. 2012. In vitro repair of DNA hairpins containing various numbers of CAG/CTG trinucleotide repeats. *DNA Repair,* 11**,** 201-209.