Title: Repetitive sequences in malaria parasite proteins

Authors: Heledd M. Davies1,2*, Stephanie D. Nofal1,3*, Emilia J. McLaughlin1, Andrew R. Osborne1

Summary: Repetitive sequences are widespread in Plasmodium falciparum proteins; the importance of these sequences is increasingly apparent.

Keywords: Malaria, Plasmodium falciparum, Protein repeats, low-complexity, host-pathogen interaction, protein evolution

ABSTRACT

Five species of parasite cause malaria in humans with the most severe disease caused by Plasmodium falciparum. Many of the proteins encoded in the P. falciparum genome are unusually enriched in repetitive low complexity sequences containing a limited repertoire of amino acids. These repetitive sequences expand and contract dynamically and are among the most rapidly changing sequences in the genome. The simplest repetitive sequences consist of single amino acid repeats such as poly-asparagine tracts which are found in approximately 25% of P. falciparum proteins. More complex repeats of two or more amino acids are also common in diverse parasite protein families. There is no universal explanation for the occurrence of repetitive sequences and it is possible that many confer no function to the encoded protein and no selective advantage or disadvantage to the parasite. However, there are increasing numbers of examples where repetitive sequences are important for parasite protein function. We discuss the diverse roles of low complexity repetitive sequences throughout the parasite life cycle, from mediating protein-protein interactions to enabling the parasite to evade the host immune system.

INTRODUCTION

Repetitive nucleotide sequences expand and contract via strand slippage during DNA replication and unequal crossing over during meiosis (Fig. 1)(Gemayel et al., 2010). Within the genome of the P. falciparum strain 3D7, repetitive sequences are found on average once every 467 nucleotides (Tan et al., 2010). These repeat encoding sequences tend to be highly dynamic; the rate of insertion and deletion mutations in genomes of parasites propagated in the asexual blood stage is approximately 10 fold greater than the rate of single base pair substitution (Hamilton et al., 2016) and within coding sequences, the majority of insertion and deletion mutations occur in short tandemly repeated sequences (Hamilton et al., 2016, Miles et al., 2016). The variability of these sequences is also highlighted by the observation that 9% of genes contain repetitive sequences that vary in length between P. falciparum strains 3D7, Dd2, and HB3 (Tan et al., 2010). Perfect repeats are predicted to have expanded more recently than degenerate repeats and are particularly common in Plasmodium species (Mendes et al., 2013).

Many short tandemly repeated sequences within P. falciparum proteins are predicted to be intrinsically disordered (Feng et al., 2006, Guy et al., 2015). In isolation, such sequences are likely dynamic and do not fold into stably folded structures but exist as an ensemble of conformations (Oldfield and Dunker, 2014). In many cases, these sequences may not have a function per se but simply form disordered loops or termini protruding from folded domains. However, repeat sequences can also significantly change the properties of proteins and over evolutionary time expansion and contraction of repeats may fine-tune protein activity or even generate sequences de novo that can confer novel function. Indeed, low complexity
sequences, which include many repetitive sequences, are highly diverse. Whilst many are evolving neutrally others are likely under selective pressure (Battistuzzi et al., 2016, Zilversmit et al., 2010).

**Diverse functions for repetitive protein sequences in diverse organisms**

Functional repetitive proteins from many species have been extensively characterized and illustrate many of the concepts that will also apply to this class of polypeptides from *Plasmodium*. Stretches of single amino acid repeats are often prone to self-association, leading to the formation of protein aggregates or fibres (Oma et al., 2004). The detrimental effects of poly-glutamine and poly-alanine sequence aggregation are well-understood as they are associated with many human diseases (Gatchel and Zoghbi, 2005, Shoubridge et al., 2007). Expansion of these repeat-encoding sequences increases their aggregation propensity and contributes to disease onset. Expansion of single amino acid repeats is also important in non-pathological contexts; the activity of many yeast transcription factors can be fine-tuned over evolutionary time by expansion or contraction of poly-glutamine sequences (Gemayel et al., 2015). Artificial homo-polymERIC sequences composed of different amino acids can direct proteins to different cellular compartments in mammalian cells (Oma et al., 2004) and poly-histidine repeats in several human DNA-binding proteins can confer localization to nuclear speckles (Salichs et al., 2009). *P. falciparum* contains a striking number of asparagine-rich tracts; similar sequences have been shown to form fibres in yeast prion proteins which are believed to regulate protein activity and allow the cell to respond to environmental stimuli (Chernova et al., 2014, Halfmann et al., 2011, Halfmann et al., 2010, Peters and Huang, 2007).

Notably, there are several examples in which repeat length can influence the phenotype of an organism. For example, experimentally changing the number of repeats in the *S. cerevisiae* cell surface proteins Flo1p and Flo11p proportionally changes the adhesion and hydrophobicity of the cells, respectively (Fidalgo et al., 2006, Verstrepen et al., 2005). In the case of Flo1p, a central repeat sequence forms a linker that allows an N-terminal adhesion domain to extend away from the cell surface; expansion of this sequence increases the efficiency of cell adherence. The repetitive sequence in Flo11p is hydrophobic and repeat expansion therefore increases the overall hydrophobicity of the yeast cell surface, allowing cells expressing Flo11p with longer repeats to float on the surface of a liquid medium. Repeat sequences can also contribute to altered phenotypes in more complex organisms; changes in repeat sequences in morphogen proteins are likely responsible for some of the different morphological features that have been selected for in domesticated dogs (Fondon and Garner, 2004, Fondon and Garner, 2007).

These examples establish several key concepts: firstly, repeat sequences can be important for protein function *per se*; secondly, expansion or contraction of repetitive sequences can cause partial or complete loss or gain of a protein function; thirdly, changes in protein activity arising from changes to repeat length can cause phenotypic changes, the magnitude of which can be proportional to the repeat length; and finally, expansion and contraction of repeat sequences can potentially facilitate adaption of an organism over evolutionary time.

**Protein interactions and structures of repetitive protein sequences**

Although relatively few repetitive sequences from *Plasmodium* proteins have been extensively characterized many of the applicable concepts relating to their structure and function have been well established in the context of repetitive proteins from other organisms.
In isolation many repetitive sequences are intrinsically disordered but upon interaction with a binding partner will assume a structured conformation. The structural plasticity and repeating modular architecture of repetitive low complexity sequences results in binding interactions that can differ from those typically described for globular proteins.

Repetitive protein sequences can potentially recruit multiple copies of an interacting partner. For example, multiple copies of transcriptional terminator proteins can interact cooperatively with neighbouring repeats in the C-terminal domain (CTD) of yeast RNA polymerase II (Lunde et al., 2010). Interaction with multiple copies of a binding partner can also increase interaction avidity. For example, the *Staphylococcus aureus* protein FnBPA contains short repeated motifs that interact with tandemly repeated β-sheet domains in human fibronectin. Although in isolation FnBPA repeats are intrinsically disordered, each repeat within a single FnBPA protein interacts with up to four consecutive fibronectin domains by contributing an additional β-strand to each domain (Bingham et al., 2008). As FnBPA contains multiple fibronectin-binding repeats it can interact with multiple fibronectin proteins which is important for adhesion to and invasion of endothelial cells (Edwards et al., 2010).

Alternatively, multiple repeats within a sequence can simultaneously interact with a single non-repetitive protein as illustrated by the interaction of the RNA polymerase II CTD with the mediator complex; here the repeats interact with four structurally non-identical binding sites with each identical repeat assuming a different conformation (Robinson et al., 2012). In addition, the CTD of RNA polymerase II interacts with many other proteins in an orchestrated manner that is regulated by its post-translational modification. This illustrates the key paradigm that repetitive sequences can function as regulated protein interaction hubs (Cumberworth et al., 2013). In this context the role of the *P. falciparum* RNA polymerase II CTD is discussed in a later section.

Although many repetitive sequences acquire structure only upon binding to a structured binding protein some form structured domains in isolation. For example, the repeated three amino acid motifs in collagen assemble to form a right-handed triple helix (Shoulders and Raines, 2009). Recent data also indicate that some repetitive sequences have a propensity to self-associate and undergo phase separation and thus partition into membraneless organelles within the cell (Kato et al., 2012). Although in most cases the functions and interactions of repetitive sequences in *Plasmodium* proteins are not well understood it is likely that these same principles will be relevant to many proteins that are expressed throughout the parasite lifecycle.

THE MALARIA PARASITE LIFECYCLE

Repetitive proteins are expressed throughout the malaria parasite lifecycle during which the parasite undergoes several differentiation steps within both insect and vertebrate hosts (Fig. 2). The repetitive proteins discussed in this review, their repeat sequences, the variability of repeats between different *P. falciparum* strains, and time of expression during the lifecycle are summarized in table 1.

An infection is initiated when haploid sporozoites are injected into the human host as a female *Anopheles* mosquito takes a blood meal. Having migrated via the bloodstream to the liver, sporozoites invade hepatocytes and multiply into thousands of haploid merozoite forms before being released back into the blood (reviewed in (Lindner et al., 2012)). The released merozoites invade erythrocytes and replicate intracellularly (reviewed in (Cowman et al., 2016)). After replication within the erythrocyte, daughter merozoites egress and after a brief
extracellular period invade new erythrocytes, undergoing further rounds of replication, egress and reinvasion. During the blood stage a fraction of the parasite population becomes committed to differentiate into gametocytes, the sexual stage of the parasite lifecycle. Gametocytes can be taken up by a mosquito during a blood meal (reviewed in (Bennink et al., 2016)). Within the mosquito midgut male gametocytes undergo exflagellation to form motile microgametes and female gametocytes form macrogametes. Gamete fusion leads to formation of a diploid zygote that differentiates into a motile oocinete which penetrates the mosquito midgut and forms an oocyst. Haploid sporozoites develop within the oocyst and are eventually released into the mosquito hemolymph before travelling to the salivary glands from where they can be injected into the next host (reviewed in (Aly et al., 2009)). After gamete fusion the parasite is briefly diploid but following a meiotic division the parasite returns to a haploid state and replicates mitotically for the remainder of the lifecycle. It is likely that expansion and contraction of gene sequences encoding repetitive proteins occurs in both meiotic and mitotic phases of the lifecycle.

FUNCTIONS OF REPEAT PROTEINS ON THE PARASITE SURFACE, WITHIN THE PARASITE AND IN THE INTERACTION WITH HOST CELLS

Whilst many proteins expressed throughout the lifecycle contain repetitive sequences the roles of these sequences are in most cases poorly understood. In the following sections we will discuss how repetitive proteins interact with the host immune system as well as specific proteins and groups of proteins where the function or importance of repeat sequences have been established.

**Repetitive Plasmodium proteins and the immune response**

Within the human host many repetitive parasite proteins are recognised by the immune system. 79% of tandem repeats in parasite proteins fall within protein sequences predicted to be intrinsically disordered (Guy et al., 2015). These sequences, in comparison to those from globular folded proteins, have a biased amino acid composition which results in fewer predicted MHC I and II peptides but a greater number of predicted B-cell epitopes (Guy et al., 2015, Hughes, 2004, Verra and Hughes, 1999). Notably, antibody interactions with disordered epitopes are predicted to be more sensitive to epitope sequence changes than interactions with structured epitopes. Given the propensity of disordered sequences to tolerate sequence polymorphism, this may allow effective evasion of antibody responses against these sequences (MacRaild et al., 2016).

Although sera from individuals living in endemic regions contain high levels of antibodies against repetitive proteins it is unclear whether these confer protection against disease. It has been speculated that repetitive proteins can even be detrimental to an effective immune response by crosslinking B-cell receptors and inducing an inferior T-cell independent immune response (Schofield, 1991). Although the antibody responses to most Plasmodium repetitive antigens have not been characterized, in mice there is a T-cell independent B-cell response to a repetitive sequence in the circumsporozoite protein (discussed later) but the bulk of the B-cell response is T-cell dependent (Fisher et al., 2017). It has also been proposed that repetitive antigens can generate an ‘antigenic smokescreen’ in which numerous cross-
reactive repetitive antigens interfere with the affinity maturation of antibodies against protective antigens (Kemp et al., 1987).

Nonetheless, many repetitive sequences are recognised by the immune system, and in particular those located on the parasite surface potentially represent important vaccine components and targets of host immunity. For the majority of the forty eight hour asexual replication cycle in the blood stage the parasite resides within the host erythrocyte. However, after egress merozoites are extracellular for up to several minutes before they invade new erythrocytes. During this period many repetitive merozoite surface proteins including MSP1, MSP2, GLURP, and SERA5 (SERA5 is a parasitophorous vacuole protein important for egress (Collins et al., 2017) but remains associated with the merozoite surface (Li et al., 2002)) and others are exposed to the host immune system (Beeson et al., 2016). SERA5 antibodies isolated from serum of Ugandan adults predominantly recognise the repetitive sequences within the protein and also show parasite growth inhibition in vitro in the presence of monocytes (Yagi et al., 2014). Similarly, antibodies that recognise the repetitive sequences in MSP1 and GLURP also show parasite growth inhibition (Galamo et al., 2009, Theisen et al., 1998). Alleles of the repetitive sequence in MSP1 appear to be under balancing selection which is consistent with this sequence being under immune selection (Conway et al., 2000). The repeating motifs of both MSP1 and MSP2 vary between isolates (Table 1), suggesting that there is not a strong functional constraint on these sequences and that they might easily change in response to immune pressure.

Circumsporozoite protein (CSP) is the major protein on the surface of the sporozoite which is also extracellular in the human host (Sinnis and Nardin, 2002). Both T-cell and antibody responses are elicited against CSP during infection. The RTS,S vaccine, which is currently the most successful malaria vaccine, provides partial protection against disease (Hoffman et al., 2015, Rts, 2015) and is based on a fragment of the circumsporozoite protein comprising tandemly repeated NANP motifs and a C-terminal domain. The central repetitive sequence contains immunodominant antibody epitopes (Zavala et al., 1985). Significantly, the repeat sequence in CSP is variable between parasite species. The repeating sequence motif in P. berghei CSP differs from those found in the P. falciparum and P. vivax proteins but despite this the repeat sequence from P. berghei CSP can be replaced with the repeat sequence from the P. vivax or P. falciparum protein without compromising function (Espinosa et al., 2013, Persson et al., 2002). Also, the number of repeats varies from 37 to 44 in different strains of P. falciparum (Neafsey et al., 2015). Although variation in the number of repeat units in CSP does not significantly affect vaccine efficacy (Neafsey et al., 2015) significant sequence variation can clearly be tolerated in the repeat sequences and their mutation may allow the parasite to evade B-cell responses without disrupting the function of CSP.

Circumsporozoite protein – a repetitive sequence essential for sporozoite development

Circumsporozoite protein, which is essential for parasite survival (Ferguson et al., 2014), is linked via a C-terminal glycophaspatidylidinositol (GPI) anchor to the sporozoite plasma membrane. It consists of an N-terminal domain that binds heparin sulphate proteoglycans, a repeating central sequence, and a C-terminal Thrombospondin repeat domain (Sinnis and Nardin, 2002). Deletion of the central repeating sequence in the P. berghei CSP gene leads to a defect in sporozoite development in the mosquito. Although oocyst formation appears to be normal, the development of sporozoites is defective and sporozoites fail to appear in either the hemolymph or salivary glands of the mosquito (Ferguson et al., 2014).
While many repetitive sequences are predicted to be entirely disordered, biophysical analyses are consistent with the CSP repeats adopting an elongated but flexible cylinder-like structure that is about 18nm in length and 1.5nm in diameter and composed of β-turn structures (Plassmeyer et al., 2009). The model is based on the propensity of short NANP peptides to adopt a β-turn conformation (Ghasparian et al., 2006, Nanzer et al., 1997). NMR and circular dichroism studies suggest that CSP repeat peptides are in equilibrium between folded and extended states (Dyson et al., 1990, Fisher et al., 2017). Consistent with this, experiments examining the force required to unfold the CSP repeat region support a model in which CSP repeats can exist in a state that is partially structured, requiring relatively low force to unfold; or in an unstructured state which requires almost no force to unfold (Patra et al., 2017). Antibody accessibility studies also support a model in which CSP is conformationally dynamic undergoing significant conformational changes in the oocyst, mosquito salivary gland, and human liver (Coppi et al., 2011, Herrera et al., 2015). The CSP repeat structure may serve to extend the N-terminal domain away from the plasma membrane of the parasite. It is possible that deletion of the repeat sequence (Ferguson et al., 2014) results in a shorter rod-like structure or alternatively it may prevent the entire protein from folding correctly. Such differences in an important protein on the surface of the sporozoite may be detrimental to development in the mosquito.

Liver stage antigen-1 – a network of covalently crosslinked repeats

Liver stage antigen-1 is localised in the parasitophorous vacuole during the liver stage of the lifecycle and is required for progression of *P. falciparum* through this stage (Fidock et al., 1994, Mikolajczak et al., 2011). The repeats in Liver stage antigen-1 are rich in glutamine and lysine residues and are an *in vitro* substrate for transglutaminase which catalyses the covalent crosslinking of polypeptides. Although transglutaminase-crosslinked peptides can be detected in liver stage parasites by microscopy, it is unclear whether LSA-1 itself is crosslinked *in vivo* (Nicoll et al., 2011), and whether this is important for liver-stage development.

RNA Polymerase – a hub for regulated recruitment of binding partners

In eukaryotes, three different RNA Polymerases (RNA Pol) are required to transcribe the nuclear genome; RNA Pol I transcribes ribosomal RNA (rRNA), RNA Pol II transcribes messenger RNA (mRNA) and RNA Pol III transcribes 5S rRNA and transfer RNA (tRNA) (Vannini and Cramer, 2012). As discussed previously, the C-terminus of the largest subunit (RbpI) of RNA Pol II contains a repetitive sequence known as the C-terminal Domain (CTD) (Corden et al., 1985). This domain consists of a tandemly repeated Tyr-Ser-Pro-Thr-Ser-Pro-Ser/Lys sequence which undergoes reversible phosphorylation during the transcription cycle. Changes in the phosphorylation pattern of repeats within the CTD are thought to orchestrate the recruitment of various nuclear factors that couple processing of the nascent mRNA to transcription (Howe, 2002)(Fig. 3).

The CTD is essential for cell viability in yeast, drosophila and mice (Litingtung et al., 1999, Nonet et al., 1987, Zehring et al., 1988). The overall length or number of the heptad repeats is also important as decreasing the number of repeats in the CTD of yeast from 26 to below 8 causes loss of viability (West and Corden, 1995). Although small truncations of the yeast
CTD are tolerated, these result in reduced transcript levels of certain genes. Truncation of the CTD in yeast to 17, 14, 13 and 11 repeats leads to the production of 58%, 8%, 5% and 2% of normal transcript levels of specific genes, respectively, signifying the importance of the repetitive nature of the CTD (Liao et al., 1991).

Although the consensus sequence and number of heptad repeats in RNA pol II varies from one organism to another, it is strictly conserved along evolutionary lineages; for example all mammalian CTDs possess 52 heptapeptide repeats with little variation in the consensus (Chapman et al., 2008). Strikingly, there is considerable variation in the number of repeats found across the *Plasmodium* genus. Rodent and bird-infecting species, including *P. berghei* and *P. gallinaceum*, respectively, possess 8 repeats, while the CTD of some primate-infecting species has expanded resulting in 13-15 repeats. This expansion appears to have arisen twice during *Plasmodium* evolution; once in the lineage that has given rise to *P. falciparum* and once in the lineage that has given rise to *P. vivax* (Kishore et al., 2009). Primate infecting parasites appear to have preferentially expanded a heptad repeat containing a lysine residue in the 7th position (Kishore et al., 2009). Furthermore, this plasticity extends not only to different *Plasmodium* species but also to different isolates within a given species, making *Plasmodium* one of the few organisms that show intra-species variation in the number of heptad repeats in its RNA pol II CTD (Kishore et al., 2009). Although the significance of the repeat expansions is unclear it has been suggested that repeat expansion may confer additional functions to the CTD (Ukaegbu and Deitsch, 2015).

In higher eukaryotes, it is known that the CTD recruits proteins with diverse roles in transcription and gene expression. These include factors such as histone modifying enzymes (Yoh et al., 2008). The CTD of *P. falciparum* RNA pol II interacts with PfsSET2 (also referred to as PfsSETvs), a histone methyl transferase which tri-methylates histone H3 at lysine 36 (H3K36me3) (Cui et al., 2008, Jiang et al., 2013, Ukaegbu et al., 2014). This histone modification is found in telomeric and subtelomeric chromatin and associated with genomic regions encoding the PfEMP1 var gene family (Jiang et al., 2013, Ukaegbu et al., 2014). Although there are more than 60 var genes in the *P. falciparum* genome, mRNA from only a single var gene is produced at any one time and consequently a single PfEMP1 protein is displayed on the surface of an infected erythrocyte. By switching var gene expression the parasite is able to evade an immune response directed against a particular PfEMP1 protein. Although mRNA from only a single var gene is transcribed at any one time the RNA pol II also transcribes non-coding RNAs from var genes. It is likely that RNA pol II recruits PfsSET2 to var genes during the latter process. Although the precise mechanism by which PfsSET2 and H3K36 tri-methylation regulates var gene silencing is unclear, deletion of PfsSET2 or expression of a dominant-negative version of PfsSET2 which interacts with the RNA pol II CTD, result in dis-regulation of var gene expression, strongly indicating a role for the CTD and its interacting partner PfsSET2 in this process (Jiang et al., 2013, Ukaegbu et al., 2014).

**Homopolymeric Asparagine-rich repeats and protein aggregation**

One of the best studied examples of repetitive protein sequences modulating protein properties comes from the study of poly-glutamine (Gln) tracts. Expansion of these homopolymeric tracts leads to protein aggregation and is associated with ‘poly-glutamine diseases’ such as Huntington’s disease in humans (Gatchel and Zoghbi, 2005). However, it has also been proposed that such poly-Gln repetitive sequences modulate protein function and that expansion and contraction of repeats may fine-tune protein function and phenotype over evolutionary time (Gemayel et al., 2015).
Single amino acid repeats of the amino acid asparagine (Asn) are prevalent in the *P. falciparum* genome. Like poly-Glutamine sequences, poly-Asn sequences are also prone to aggregation particularly at elevated temperatures (Halfmann et al., 2011) which the parasite would experience during fever. Given that protein aggregation is often toxic to cells it is remarkable that the parasite proteome maintains so many potentially toxic Asn-rich repeats; more than a quarter of its proteins contain such sequences. Expression of a GFP-tagged version of the Asn-rich protein from *P. falciparum*, PF3D7_0923500, in human cells leads to accumulation of fluorescent protein aggregates as does expression of the N-terminal prion forming domain of *S. cerevisiae* asparagine/glutamine-rich Sup35 (Muralidharan et al., 2012). Co-expression of the human Hsp110 or *Plasmodium* homologue PfHsp110c inhibits aggregate formation; PfHsp110c does so approximately 10-15 fold more efficiently than the human chaperone (Muralidharan et al., 2012). This suggests that *Plasmodium* chaperones may be particularly efficient at suppressing the aggregation of Asn-rich repeat sequences. Although PfHsp110c is essential for survival, a reduction of expression leads to increased aggregation of Asn-rich proteins and parasites cannot recover from short periods of heat shock (Muralidharan et al., 2012). These data are consistent with a model in which PfHsp110c plays a key role in maintaining the solubility of Asn-rich proteins under heat shock conditions (Muralidharan et al., 2012).

The function of Asn-rich repeats in the parasite is unclear and in many cases these sequences have likely evolved without cost or benefit. Notably, deletion of the Asn-rich repeat sequence in the proteasome protein Rpn6 does not compromise its function (Muralidharan et al., 2011). However, several transcription factors in *S. cerevisiae* contain poly-glutamine repeats and variation of the repeat length can fine-tune transcription of target genes and resulting phenotypes (Gemayel et al., 2015). This is at least partially due to changes in protein solubility and interaction partners. Notably, the majority of AP-2 transcription factors encoded in the parasite genome have very high asparagine content of between 20% and 30%. The *P. falciparum* protein AP2-g, which triggers the transition to gametocytogenesis (Kafsack et al., 2014, Sinha et al., 2014), comprises 2432 residues, 29.9% of which are asparagine. Asparagines are found both in homopolymeric repeats, the longest of which is 33 residues in length, as well as in short tandemly repeated sequences. It is unclear if variations in the AP2-g Asn-rich repeats seen in different strains (Campino et al., 2016) fine tune its solubility/activity and hence the propensity of parasites to form gametocytes. Changes in Poly-Asn sequences in other AP2 transcription factors might also modulate their activity leading to small changes in progression through the lifecycle; even small delays in the asexual cycle of the parasite could lead to a selective advantage. Additionally, given the propensity of asparagine-rich repeats to aggregate, particularly under heat shock, it is possible that solubility/activity of transcription factors might change during fever.

**Asparagine repeats and protein translation**

The rate of translation of a protein is determined by the concentration of available aminoacylated tRNAs. Given the unusually high occurrence of asparagine residues in the *P. falciparum* genome, particularly in long low complexity repetitive sequences, the parasite might require large amounts of Asn:tRNA to efficiently translate Asn-rich proteins. In many organisms there is redundancy in the genes encoding tRNAs for each amino acid. For example, in *C. elegans* there are 46 genes encoding tRNAs that recognise the four glycine codons (Duret, 2000). Given the abundance of Asn in encoded proteins, it might be expected that the *P. falciparum* genome would encode multiple tRNAs that recognize Asn codons thus allowing the parasite to make high levels of Asn:tRNA. However, only a single gene encodes Asn:tRNA (Gardner et al., 2002) and levels of Asn:tRNA are not proportionally more
abundant. This has led to a model in which translation is slowed down when the ribosome encounters long low complexity Asn-repeat sequences due to the relatively lower concentrations of Asn-tRNA (Filisetti et al., 2013, Frugier et al., 2010). This could slow the rate of polypeptide emergence from the ribosome tunnel and allow N-terminal domains that have already been synthesised more time to fold before emergence of subsequent domains. This may increase the overall efficiency of protein folding.

MODIFICATION OF THE HOST ERYTHROCYTE BY REPETITIVE EXPORTED PROTEINS

The human erythrocyte is a relatively simple terminally differentiated cell that lacks internal organelles. During the blood stage of infection the parasite profoundly remodels its host cell, establishing new permeability pathways for acquiring nutrients, modifying the erythrocyte cytoskeleton to increase rigidity, and generating knob-like structures on the surface of the host cell which allow infected cells to cytoadhere to endothelial cells in the microvasculature (Boddey and Cowman, 2013, de Koning-Ward et al., 2016, Przyborski et al., 2016, Spielmann and Gilberger, 2015, Spillman et al., 2015). The parasite brings about these changes by exporting over 400 proteins into the host cell; these proteins can be localised to the erythrocyte cytoplasm, plasma membrane, the underlying spectrin cytoskeleton, or to Maurer’s clefts which are membranous structures assembled by the parasite (Fig. 4). Many exported proteins are characterized by the presence of an N-terminal signal sequence followed by a short motif referred to as a Protein Export Element (PEXEL) or Host targeting motif (HT) (Boddey et al., 2013, Hiller et al., 2004, Marti et al., 2004, Schulze et al., 2015). Another class of exported proteins lack a PEXEL sequence and are referred to as PEXEL negative exported proteins (PNEPS) (Heiber et al., 2013). Notably, many exported proteins contain repetitive sequences that play key roles in diverse processes including cytoadhesion, formation of Maurer’s clefts, protein localisation, and detoxification of heme (Fig. 4).

Lysine-rich repeats in exported proteins – Changes in repeat length can influence protein localization

There are several examples of repetitive proteins, such as the yeast proteins Flo1p and Flo11p (discussed in previous section), where the function of the protein can be proportionally modulated by repeat length (Fidalgo et al., 2006, Verstrepen et al., 2005). It is conceptually clear in such cases how relatively small changes in the number of tandemly repeated sequences could be important in facilitating adaption of an organism to its environment. In Plasmodium, many proteins including exported proteins, contain low complexity lysine-rich repetitive sequences. In some cases these sequences are adjacent to a folded domain but they can also be found in proteins that are predicted to be entirely intrinsically disordered. In several proteins which are exported into the host cell by a PEXEL motif, lysine-rich sequences can target the protein to the periphery of the infected erythrocyte (Davies et al., 2016). Comprising relatively simple tandemly repeated motifs of as few as three residues (e.g. Lys-Lys-Glu), these sequences are likely to interact with either lipids or proteins in the erythrocyte plasma membrane or the underlying erythrocyte spectrin cytoskeleton.

The exported P. falciparum protein glutamic acid-rich protein (GARP) contains multiple lysine-rich sequences that function in this manner. Significantly, the targeting propensity of a repetitive lysine-rich sequence from GARP is proportional to its length - increasing the number of tandemly repeated sequences proportionally shifts the protein from being diffusely
localised in the erythrocyte cytoplasm to being highly concentrated adjacent to the plasma membrane of the infected cell (Fig. 5) (Davies et al., 2016). At least nine exported parasite proteins contain lysine-rich sequences that independently target to the periphery of the erythrocyte. The lengths of repeating lysine-rich sequences from many of these exported proteins vary between parasite strains (Davies et al., 2016). Additionally, in the protein MESA, a block of sequence made up of tandemly repeated lysine-rich motifs has itself been duplicated with two to four copies of the sequence in different parasite strains (Davies et al., 2016). Although the phenotypic consequences of repeat expansion and duplication have not been determined experimentally, several of these proteins are known to modulate rigidity and adhesion of infected cells. Given this, expansion and contraction of lysine-rich sequences may proportionally change the concentration of proteins localised to the periphery of the host cell, fine-tuning processes such as cell rigidity or adhesion, and facilitating adaption of the parasite to selective pressure over evolutionary time.

A comparison of the protein sequence of GARP from three different species, *P. falciparum*, *P. reichenowi*, and *P. gaboni*, indicates that repeat expansion can also lead to de novo formation of a targeting domain (Davies et al., 2016). There are three stretches of lysine-rich repetitive sequence within *P. falciparum* GARP. The first sequence in the *P. falciparum* protein is able to target a GFP reporter to the periphery of the host cell. However, the equivalent lysine-rich sequence in the *P. reichenowi* protein is shorter and does not efficiently target to the periphery of the host cell. The second lysine-rich sequence is conserved and is likely to be functional in all three species. The third lysine-rich sequence, that comprises two different repeated motifs in *P. falciparum* GARP, can target a protein to the erythrocyte periphery but the equivalent sequence in *P. gaboni* comprises only one or two repeated lysine-rich motifs and cannot target GFP to the periphery of the infected erythrocyte. These observations show not only that repeat expansion of lysine-rich sequences can modulate protein localization but also that functional targeting sequences can be generated de novo by repeat expansion over evolutionary time.

The localisation of Hyp12 (PF3D7_1301400), which modulates the rigidity of the infected cell (Maier et al., 2008), appears to be defined by two conflicting sequences; a repetitive lysine-rich sequence and a repetitive acidic sequence (Davies et al., 2016). Hyp12 is unusual in that the lysine-rich repeat sequence alone is able to target GFP to the periphery of the infected cell but the GFP-tagged full-length protein is localised in the erythrocyte cytoplasm (Davies et al., 2016, Petersen et al., 2015). Removal of the acidic sequence results in the protein being localised to the red cell periphery suggesting that the acidic sequence inhibits the targeting function of the lysine-rich repeat sequence. It is possible that a specific stimulus might unmask the sequence during an infection. Over evolutionary time, it is likely that either expansion of the lysine-rich region or contraction of the acidic sequence could shift the protein from the erythrocyte cytoplasm to the periphery of the erythrocyte.

**Knob-associated histidine-rich protein – repeat protein interactions and parasite virulence**

Knob-associated histidine-rich protein (KAHRP) is one of the best-characterised repeat-containing proteins. It is a key component of structured spiral-like ‘knob’ protrusions which the parasite assembles at the surface of the infected erythrocyte (Crabb et al., 1997, Watermeyer et al., 2016). Knobs are involved in displaying PfEMP1 proteins on the surface of the infected cell where they mediate adhesion to specific ligands allowing parasites to become sequestered in tissues thus avoiding passage through the spleen (Crabb et al., 1997). KAHRP interacts with the host cytoskeletal components spectrin (Kilejian et al., 1991, Pei et
and deletion of the gene causes a loss of knob structures, decreases adhesion under flow conditions (Crabb et al., 1997), and reduces erythrocyte rigidity (Glenister et al., 2002). KAHRP contains an N-terminal repetitive polyhistidine stretch that will target GFP to the periphery of the infected erythrocyte (Rug et al., 2006). It also has two unique lysine-rich repetitive regions, known as the 5' and 3' repeats; only the 5’ repeat sequence is sufficient to independently target GFP to the erythrocyte periphery (Davies et al., 2016). Progressive truncation of the C-terminus of KAHRP indicates that the 5' repeats are important for the role of KAHRP in cytoadhesion (Rug et al., 2006). Several studies indicate that these repeats interact with spectrin (Cutts et al., 2017, Kilejian et al., 1991, Pei et al., 2005). Most recently, it has been demonstrated in vitro that the repeats interact with β-spectrin domains 10-14 and more weakly with α-spectrin domains 12-16. Significantly, progressive removal of repeats from the KAHRP repeat sequence leads to a corresponding reduction of binding affinity for β-spectrin domains 10-14 i.e. the binding affinity is proportional to the number of KAHRP repeats (Fig. 5) (Cutts et al., 2017). This again highlights the potential importance of repeat expansion and contraction in modulating protein function. Modelling and molecular dynamics simulations support a model whereby the KAHRP repeats adopt an extended conformation of the surface of spectrin, interacting with charge complementary surfaces on multiple β-spectrin domains (Cutts et al., 2017).

Although knob structures are unique to P. falciparum, KAHRP-like proteins are found in all primate-infecting Plasmodium species (Davies et al., 2016, Sargeant et al., 2006). These proteins share an N-terminal EKAL domain (EMP3 KAHRP-like domain) followed by lysine-rich repeating sequences but the function of the proteins in other species is not known. The repeated lysine-rich motifs in the P. falciparum and P. knowlesi proteins both contain consecutive lysine residues but otherwise differ significantly. However, when expressed in P. falciparum, both can target a protein to the periphery of the infected cell (Davies et al., 2016). It is likely that both these repetitive lysine-rich sequences interact with a negatively charged surface on spectrin. Presumably the binding interface can tolerate somewhat diverse conformations of positively charged interacting peptides. Notably, there does not appear to be a strong consensus sequence among all of the different lysine-rich targeting sequences identified thus far. Identification of the binding partners, which could be proteins or lipids, for each protein and further structural studies will be required to understand the functions of these diverse repetitive sequences.

Although KAHRP contains a conserved EKAL domain of unknown function at its N-terminus (Davies et al., 2016, Sargeant et al., 2006), the remainder of the protein is predicted to be intrinsically disordered and much of the sequence is repetitive. Likewise, PfEMP3 contains a conserved N-terminal EKAL domain and several different repetitive sequences at its C-terminus but the repetitive sequences are very different from those of KAHRP. It is likely that the PfEMP3 gene arose as a duplication of the KAHRP gene or vice versa yet gene knockout and functional studies show that the two proteins are not functionally equivalent (Crabb et al., 1997, Glenister et al., 2002, Waterkeyn et al., 2000). Likely, after gene duplication the dynamic repetitive sequences have diverged dramatically and consequently the function of the two proteins has likewise also diverged. The dynamic nature of repetitive sequences together with the telomeric location of many of the genes encoding exported proteins suggests that gene duplication and rapid diversification have occurred frequently during the evolution of the parasite; repeat expansion likely plays an important role in the latter process.

PHIST/PRESAN domain proteins
A large group of functionally diverse exported parasite proteins belong to the PHIST (Plasmodium helical interspersed subtelomeric) or PRESAN (Plasmodium RESA N-terminal) domain family (Oakley et al., 2007, Sargeant et al., 2006, Warncke et al., 2016). The proteins are characterized by the presence of a conserved alpha helical domain (Mayer et al., 2012) and are exported into the infected erythrocyte during the asexual erythrocyte stages and during gametocytogenesis (Oakley et al., 2007, Sargeant et al., 2006, Silvestrini et al., 2010, Tiburcio et al., 2015). One of the best characterised proteins in this family is LyMP (lysine-rich membrane-associated PHISTb), which modulates erythrocyte cytoadhesion by linking PfEMP1 to the erythrocyte cytoskeleton (Oberli et al., 2014, Oberli et al., 2016, Proellocks et al., 2014, Tarr et al., 2014). LyMP contains an N-terminal PHIST/PRESAN domain, which interacts with the cytoplasmic tail of PfEMP1 (Oberli et al., 2014), and a C-terminal sequence that interacts with the erythrocyte cytoskeletal protein Band 3 (Oberli et al., 2016). The extreme C-terminus of LyMP contains a lysine-rich repetitive sequence which when GFP-tagged and exported into the infected cell and do not independently confer a cytoskeletal localization. Although the fragment of LyMP that has been shown to interact with Band 3 in vitro contains the lysine-rich sequence, it is unclear whether this sequence interacts with Band 3 or another component of the erythrocyte cytoskeleton.

LyMP is an example of a gene that encodes a folded domain (PHIST/PRESAN domain) but has expanded a repetitive lysine-rich sequence which likely allows the encoded protein to interact with an additional protein. Similarly, GEXP12, PF3D7_1201000, PF3D7_1476200, and PF3D7_0402000 contain a PHIST/PRESAN domain but also contain lysine-rich sequences that alone localise to the periphery of the infected erythrocyte (Davies et al., 2016). The PHIST/PRESAN domain of PF3D7_0402000 interacts with Band 4.1 and ankyrin that are components of the erythrocyte cytoskeleton (Parish et al., 2013, Shakya et al., 2017, Vignali et al., 2008), but the protein also contains a C-terminal lysine-rich repetitive targeting sequence (Davies et al., 2016). Other PHIST/PRESAN proteins, such as PFD80 (Tarr et al., 2014) and RESA (Ring expressed surface antigen) (Brown et al., 1985) localize to the erythrocyte cytoskeleton, and have expanded repetitive sequences that are not lysine-rich and do not independently confer a cytoskeletal localization. RESA modulates the rigidity of the infected cell and deletion of the gene does not affect cytoadhesion (Maier et al., 2008).

Although the role of PFD80 is unclear it is likely essential as attempts to delete the gene have been unsuccessful (Maier et al., 2008). The localization and likely functions of PHIST proteins are diverse and in most cases the functions of their repetitive sequences are unclear. However, as with KAHRP and PfEMP3, it is possible that the functions of the protein family members have been diversified by duplication of genes encoding a protein with a conserved domain followed by expansion of different repetitive sequences which may contribute to modulating or changing the function of the protein.

**Repetitive Maurer’s cleft proteins**

There are numerous exported proteins with repetitive sequences that are not lysine-rich. In some cases the function of these sequences is well understood. Ring exported protein-1 (REX1) is a PEXEL negative exported protein which has a repetitive sequence that is essential for protein function (McHugh et al., 2015). The protein is peripherally associated with Maurer’s clefts (Hawthorne et al., 2004) which are flattened membrane vesicles in the infected erythrocyte and are essential for the trafficking of PfEMP1 to the surface of the infected erythrocyte. Deletion of genes encoding several Maurer’s cleft proteins disrupts PfEMP1 trafficking (Cooke et al., 2006, Maier et al., 2007, Rug et al., 2014, Spycher et al., 2008).
Deletion of the repetitive region of REX1 (residues 371 - 579) reduces the number of Maurer’s clefts detected in the erythrocyte. On average 13 punctae corresponding to Maurer’s clefts are seen in erythrocytes infected with wildtype parasites but only 2 large punctae are seen in cells infected with parasites expressing REX1 that is missing the repetitive sequence (McHugh et al., 2015). Electron microscopy reveals that these Maurer’s clefts have an unusual stacking phenotype with an average of six clefts stacked on top of each other. (McHugh et al., 2015). Although PfEMP1 is trafficked to the abnormal stacked Maurer’s clefts, transfer to the erythrocyte plasma membrane is less efficient when REX1 is either knocked down or the repetitive region is deleted (McHugh et al., 2015). Removal of the repetitive region also reduces the adhesion of infected erythrocytes to CD36 cells under flow conditions (McHugh et al., 2015).

Electron microscopy reveals that these Maurer’s clefts have an unusual stacking phenotype with an average of six clefts stacked on top of each other. (McHugh et al., 2015). Although PfEMP1 is trafficked to the abnormal stacked Maurer’s clefts, transfer to the erythrocyte plasma membrane is less efficient when REX1 is either knocked down or the repetitive region is deleted (McHugh et al., 2015). Removal of the repetitive region also reduces the adhesion of infected erythrocytes to CD36 cells under flow conditions (McHugh et al., 2015). This indicates that the repetitive region of REX1 is functionally important for both Maurer’s cleft architecture and efficient trafficking of PfEMP1.

It is unlikely that the repeat region is simply forming a linker sequence between the N- and C-termini of the protein, as the C-terminal residues are not required for protein function and can be deleted. This suggests that the repeat sequences play an active role in protein function, possibly forming a key structural element or recruiting other proteins to the Maurer’s cleft membrane. Notably, the sequence and length of the repetitive domain varies between different parasites; this could contribute to the formation of stacked cleft structures in some strains (Wickert and Krohne, 2007) and may modulate the propensity of parasites to cytoadhere (McHugh et al., 2015).

Skeleton binding protein-1 (SBP1) is an exported PNEP which contains a single transmembrane segment which is integrated within the Maurer’s cleft membrane. Deletion of the gene results in reduced trafficking of PfEMP1 to the erythrocyte surface and reduced cytoadhesion (Chan et al., 2016, Cooke et al., 2006, Maier et al., 2007). SBP1 contains repetitive sequences located N- and C-terminally on either side of the transmembrane segment. The N-terminal repeats can self-associate (Saridaki et al., 2009) while the glutamine-rich repeats at its C-terminus associate with the cytoskeletal components spectrin and 4.1 in vitro (Kats et al., 2015). Although these interactions likely contribute to the tethering of Maurer’s clefts to the erythrocyte cytoskeleton, the C-terminus of SBP1 can be replaced by an unrelated protein sequence without altering the location of Maurer’s clefts or cytoadhesion of infected cells. However, as it is known that there are multiple mechanisms for tethering of Maurer’s clefts there may be redundancy in this function of SBP1 (Cyrklaff et al., 2011, Hanssen et al., 2010, Pachlatko et al., 2010, Rug et al., 2014).

Pf332 is a large exported protein that localises to Maurer’s clefts and is mainly composed of several different glutamate-rich repeats. Pf332 deletion and truncation of its repetitive C-terminus have been shown to have varying effects on the rigidity and cytoadhesion of infected cells; differences between studies may be a result of using different parasite strains and the particular deletion strategies (Glenister et al., 2009, Hodder et al., 2009).

Exported Repetitive histidine-rich proteins

Histidine-rich Protein 2 (HRP2) is a *P. falciparum* protein comprised almost entirely of repetitive Histidine-rich sequences. HRP2 is exported into the erythrocyte cytoplasm during the blood stage of infection and released into the blood during parasite egress (Howard et al., 1986) where it can accumulate to high levels (Ramutton et al., 2012). Several functions for HRP2 have been proposed including the detoxification of heme. The parasite internalizes haemoglobin from the cytoplasm of the host erythrocyte and digests it within the food vacuole; a fraction of HRP2 is also localised in the food vacuole (Papalexis et al., 2001,
Within the food vacuole the toxic heme moiety, a by-product of haemoglobin degradation, is converted into inert hemozoin. *In vitro*, HRP2 is able to catalyse the conversion of heme to hemozoin as is a similar histidine-rich repetitive protein HRP3 (Sullivan et al., 1996). HRP2 is also able to promote the degradation of heme (Papalexis et al., 2001). However, *P. falciparum* parasites lacking genes encoding both HRP2 and HRP3 are able to produce hemozoin suggesting that there are multiple mechanisms by which hemozoin is generated *in vivo* or that HRP2 plays a relatively minor role in this process. Indeed, other protein-based (Jani et al., 2008, Sullivan et al., 1996) and non-protein based mechanisms have been proposed (reviewed in (Egan, 2008)).

HRP2 has also been implicated in regulation of thrombin inhibition (Ndonwi et al., 2011). *In vitro*, anti-thrombin inhibits the activity of thrombin and factor 10a and the inhibition is potentiated by glycosaminoglycans such as heparin. The potentiation can be antagonized by the interaction of HRP2 with glycosaminoglycans (Ndonwi et al., 2011). Thus HRP2 may behave as a pro-coagulant during infection.

More recent studies indicate that HRP2 may play a role in cerebral malaria by disrupting the integrity of brain endothelial cell barriers. Using a system that comprises an *in vitro* cultured monolayer of human endothelial cells which mimics the blood brain barrier, it has been shown that HRP2 can disrupt the integrity of an endothelial barrier (Pal et al., 2016). Addition of parasites that produce HRP2 (but not parasites that do not make HRP2) or recombinant HRP2 protein to the monolayer of endothelial cells results in redistribution of tight junction and adherens junction proteins, and disruption of the barrier formed by the cell monolayer. Blocking of key signalling factors in the host cell prevents the disruption caused by HRP2 supporting a model of inflammasome mediated activation of the innate immune system in this process. Exposure to HRP2 also increases the number of VCAM-1 and ICAM-1 receptors on the surface of endothelial cells. Both receptors can be used by infected erythrocytes to adhere to endothelial cells suggesting that activation of an inflammatory response may enhance cytoadherence and sequestration of parasites (Pal et al., 2016). Although mouse malaria parasites lack an HRP2 homolog, injection of HRP2 protein reduces survival of infected mice (Pal et al., 2017). It remains to be determined whether the incidence of cerebral malaria in humans is lower in infections caused by HRP2-negative parasites.

Maurer’s cleft associated histidine-rich protein 1 (MAHRP-1) is an exported protein localized to Maurer’s clefts. The protein contains a single transmembrane segment and a repetitive histidine-rich C-terminus, and is involved in protein trafficking within the infected erythrocyte (Spycher et al., 2008). Like HRP2, the repetitive histidine-rich sequence of MAHRP is implicated in heme detoxification (Papalexis et al., 2001). The repetitive sequence is able to bind ferric heme, which is toxic to cells, and accelerates its degradation mediated by hydrogen peroxide. This suggests that the histidine-rich sequence of MAHRP-1 may play a role in scavenging ferric heme and detoxification of reactive oxygen species within the infected erythrocyte (Spycher et al., 2003). Notably, the binding stoichiometry between ferric heme and MAHRP-1 is proportional to the number of repeats within the protein. This suggests that the potential protective function of MAHRP-1 could be influenced by repeat expansion (Spycher et al., 2003).

**Histidine-rich repetitive proteins and diagnosis of malaria**

The World Health Organisation now recommends diagnostic testing of people with suspected malaria before treatment (WHO, 2016). Rapid diagnostic tests (RDTs) represent an effective tool in this context. Most RDTs utilize antibodies to detect the HRP2 protein in blood. The mature protein, after cleavage of the PEXEL/HT motif (Chang et al., 2008), comprises
histidine-rich repeated sequences which form the epitopes recognized in RDTs (Wellems and Howard, 1986). More than 20 sequence variants of these repeat sequences have been identified (Baker et al., 2010, Deme et al., 2014). There is significant polymorphism in overall length, sequence composition of the repeated motifs, and the number of repeated motifs, between different P. falciparum strains (Baker et al., 2010, Baker et al., 2005, Deme et al., 2014, Ramutton et al., 2012) but there is currently no indication that sequence polymorphism can cause false negative RDT tests (Baker et al., 2010, Ramutton et al., 2012). However, increasing RDT use may place a selective pressure on the HRP2 gene and a contraction or mutation of repeats recognised by antibodies in RDTs may result in reduced test sensitivity or failure. Parasites lacking the HRP2 gene, initially identified in Peru (Gamboa et al., 2010) and subsequently other regions, do yield negative HRP2 based RDT tests (Bharti et al., 2016, Amoah et al., 2016, Li et al., 2015, Murillo Solano et al., 2015, Akinyi Okoth et al., 2015, Abdallah et al., 2015, Kumar et al., 2013, Wurtz et al., 2013, Maltha et al., 2012, Houze et al., 2011, Koita et al., 2012, Dorado et al., 2016, Baldeviano et al., 2015, Cheng et al., 2014).

**CONCLUSION**

In summary, repetitive sequences are widespread in Plasmodium genomes (Mendes et al., 2013) and are particularly abundant in P. falciparum genes. These dynamic sequences can arise, expand, or contract rapidly. Due to the assumption that they represent non-functional ‘junk DNA’ and are challenging to accurately assemble from short-read sequence data, they are mostly excluded from genomic analyses. Indeed, it is reasonable to assume in many cases that repeats simply expand and contract rapidly due to frequent errors in DNA replication, and evolve neutrally. Although it is difficult to provide a universal explanation for the presence of these sequences in the parasite genome, it appears that relatively simple homopolymeric repeats tend to evolve neutrally but more complex sequences appear to be under selection suggesting that they are functional (Battistuzzi et al., 2016, Zilversmit et al., 2010).

The fraction of Plasmodium repetitive protein sequences that are functionally important remains unclear and most remain uncharacterized. It is evident that some sequences within open reading frames encode repetitive polypeptide sequences that can modulate protein activity or are integral to protein function forming key structural elements, binding sites for other proteins and ligands, and localisation signals. Repeat expansion and contraction can likely modulate and diversify protein function resulting in gain or loss of function phenotypes that allow the parasite to adapt under selective pressure. Despite the diverse functions of repetitive sequences it is conceptually clear in many cases how changes in repeat length might influence protein function and parasite phenotype. The development of new genetic tools will now allow the importance of these sequences to be tested. In the simplest cases, relatively small changes in length of diverse repeat types may alter protein folding efficiency, stability, or aggregation leading to selective advantage for a parasite. In cases such as the CTD of RNA Polymerase II, where repeat sequences have very specific functions in the regulated recruitment of protein binding partners, changes in repeat number may fine tune protein function in a manner proportional to the number of repeats. The abundance of repetitive sequences that are recognized by the immune system and the occurrence of functional repetitive sequences in proteins such as KAHRP and REX1 that are central players in the pathology of severe malaria, highlight why an understanding of these sequences is important to our understanding of severe disease, its treatment, and prevention.
FUNDING:

This work was supported by the Wellcome Trust [099764 Z/12/Z] the Biotechnological and Biological Sciences research council [BB/M009513/1].

Footnote:

Corresponding author: Andrew Osborne

1- Institute of Structural and Molecular Biology, Department of Biological Sciences, Birkbeck and University College London, London, UK
E-mail a.osborne@ucl.ac.uk; Telephone (+44) 0207 679 3155; Fax (+44) 0207 679 7046.

* - authors contributed equally
2 - current address - The Francis Crick Institute, London, United Kingdom, UK
3 - current address - London School of Hygiene & Tropical Medicine, London, UK


circumsporozoite protein has two functional domains, each with distinct roles as sporozoites journey from mosquito to mammalian host. *J Exp Med*, 208, 341-56.


homozygous deletion in the C-terminal domain of RNA polymerase II. Mol Gen Genet, 261, 100-5.


detoxification of the by-products of haemoglobin degradation. Mol Biochem Parasitol, 115, 77-86.


Figure legends:

Figure 1. Repeat expansion and contraction during DNA replication and unequal crossing over in meiosis.

Figure 2. The Plasmodium falciparum life cycle. A) Sporozoites are injected from the salivary glands of the Anopheles mosquito and migrate to the host liver. B) Upon invasion of hepatocytes, thousands of merozoites are produced and subsequently released into the bloodstream. C) Merozoites invade host erythrocytes and cycle through the ring, trophozoite and schizont stages before daughter merozoites egress from the host erythrocyte and invade new cells. D) A small proportion of parasites differentiate into male and female gametocytes. E) Gametocytes are ingested into the mosquito’s midgut as it takes a blood meal where they form gametes and fuse to form a diploid zygote. The zygote differentiates into an ookinete which develops into an oocyst within the midgut wall. Upon oocyst rupture, sporozoites are released into the hemolymph and travel to the mosquito salivary gland. Proteins containing repetitive/low complexity sequences are labelled at the approximate life stage where they are utilised.

Figure 3. RNA Polymerase II C-terminal domain – phosphorylation and recruitment of proteins by the RNA pol II CTD during transcription.

Figure 4. Repetitive proteins exported into the host cell. Many proteins are exported into the host erythrocyte by P. falciparum. These are involved in creating new structures within the host cell: membranous structures called Maurer’s clefts (MC) appear in the host cell cytoplasm, a tubovesicular network’ (TVN) extends from the parasitophorous vacuole (PV), and ‘knob’ protrusions form on the cell surface which present the cytoadhesive surface protein PfEMP1. Many repetitive proteins interact with the spectrin cytoskeleton of the erythrocyte resulting in increased cell rigidity.

Figure 5. Expansion and contraction of lysine-rich repetitive sequences and changes in protein localization or protein binding affinity. The top panel illustrates the correlation between the length of a lysine-rich sequence from the protein GARP and the efficiency with which it is targeted to the erythrocyte periphery.
The lower panel illustrates the changes in binding affinity between β-spectrin and KAHRP repeats as the number of repeats in KAHRP is altered.
<table>
<thead>
<tr>
<th>Gene Identifier</th>
<th>Alias</th>
<th>Consensus Motif in 3D7</th>
<th>Number of Repeats</th>
<th>Localisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF3D7_0930300</td>
<td>MSP1</td>
<td>GASAQS (GSGGSVASG/SGGVSVT)*</td>
<td>2-6 (3-5/4)</td>
<td>Merozoite surface</td>
</tr>
<tr>
<td>PF3D7_0206800</td>
<td>MSP2</td>
<td>AGGS (TTTESNSPSPPiPAGAGASGNP/AGAGASGS/GASGSG/SGSAGG/GAGAS/SGSAG) *</td>
<td>7-11 (4 6/8 13/6 9/3)</td>
<td>Merozoite surface</td>
</tr>
<tr>
<td>PF3D7_0207600</td>
<td>SERA5</td>
<td>QGSTGASP</td>
<td>3-12</td>
<td>PV + Merozoite surface</td>
</tr>
<tr>
<td>PF3D7_1035300</td>
<td>GLURP</td>
<td>EILPEDKNEKVQHEIVEVESEKSVSEPAEHVEIV</td>
<td>4-14 3</td>
<td>Merozoite surface</td>
</tr>
<tr>
<td>PF3D7_1036400</td>
<td>LSA1</td>
<td>DLEQERLAKEKLIQEQQS</td>
<td>32-93</td>
<td>Liver Stage</td>
</tr>
<tr>
<td>PF3D7_0318200</td>
<td>RNA polymerase 2</td>
<td>YSPTSPK</td>
<td>6-11</td>
<td>Nucleus</td>
</tr>
<tr>
<td>PF3D7_0831800</td>
<td>HRP2</td>
<td>AHHAAD</td>
<td>38-46</td>
<td>RBC cytoplasm</td>
</tr>
<tr>
<td>PF3D7_1222600</td>
<td>AP2-G**</td>
<td>KNN</td>
<td>7-9</td>
<td>Nucleus</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DTYN</td>
<td>16-18</td>
<td></td>
</tr>
<tr>
<td>PF3D7_1370300</td>
<td>MAHRP1</td>
<td>HDHD</td>
<td>3-5</td>
<td>Maurer's Clefts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>DHG</td>
<td>6-12</td>
<td></td>
</tr>
<tr>
<td>PF3D7_0304600</td>
<td>CSP</td>
<td>NANP</td>
<td>40-46</td>
<td>Sporozoite Surface</td>
</tr>
<tr>
<td>PF3D7_0935900</td>
<td>REX1</td>
<td>KPQAEKDAKLTTTYDQTEKV</td>
<td>3-6</td>
<td>Maurer's Clefts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NKETKPKQNDKYTL</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>PF3D7_0501300</td>
<td>SBP1</td>
<td>ASGIGNLVGDA</td>
<td>5-6</td>
<td>Maurer's Clefts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QNAQ</td>
<td>7-14</td>
<td></td>
</tr>
<tr>
<td>PF3D7_1149000</td>
<td>Pf332</td>
<td>PVEEKNVSEEI</td>
<td>6-11</td>
<td>Maurer's Clefts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FVTGELPEEDIINEKVQEEEE</td>
<td>3-5</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>EESASEEIVEDEGSV</td>
<td>5-7</td>
<td></td>
</tr>
<tr>
<td>Protein ID</td>
<td>Antigen</td>
<td>Peptide Sequence</td>
<td>Start</td>
<td>End</td>
</tr>
<tr>
<td>---------------------</td>
<td>---------</td>
<td>---------------------------</td>
<td>-------</td>
<td>-----</td>
</tr>
<tr>
<td>PF3D7_0401800</td>
<td>PFD80</td>
<td>ENVEEKTMDEEIVDQGSVV</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEEVVEEGSV</td>
<td>5</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EEEVVEGSAT</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EEIVEEESSS</td>
<td>11</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EEIVEEEGSVV</td>
<td>10</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SVTEELVDEG</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEEIVED-EGSF</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NEEILEEEEGSY</td>
<td>6</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>GSATDFVQGQSDNEEIIIEE</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>IKEEQLDSEE</td>
<td>6</td>
<td>20</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EEEVSVDD</td>
<td>2</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EEEIEIESVT</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>TEDVEEVSS</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>PF3D7_0102200</td>
<td>RESA</td>
<td>STA</td>
<td>9</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RSASAASTT</td>
<td>8</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>STSTTQSPST</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>PF3D7_0113000</td>
<td>GARP</td>
<td>EEPTVADEHV</td>
<td>6</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>EENV</td>
<td>41</td>
<td>47</td>
</tr>
<tr>
<td>PF3D7_0201900</td>
<td>PfEMP3</td>
<td>LEEYNETDLAKGKEVTNKAHEN</td>
<td>17</td>
<td>19</td>
</tr>
<tr>
<td></td>
<td></td>
<td>KNKELQNGKSEGKLKENAEL</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NKDISNKDMKHNKELL</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>QQNTGLKNTPSEG</td>
<td>54</td>
<td>87</td>
</tr>
<tr>
<td>PF3D7_0202000</td>
<td>KAHRP</td>
<td>SKKHKDHDGEKKK</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ATKEASTSKE</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>PF3D7_0402000</td>
<td>PHISTa</td>
<td>KQGGKKEEV</td>
<td>9</td>
<td>14</td>
</tr>
<tr>
<td></td>
<td>Protein</td>
<td>Gene</td>
<td>Repeat Consensus</td>
<td>Phase</td>
</tr>
<tr>
<td>-------------</td>
<td>---------</td>
<td>-------</td>
<td>------------------</td>
<td>-------</td>
</tr>
<tr>
<td>PF3D7_0500800</td>
<td>MESA</td>
<td>GESKET</td>
<td>EKNDEKKDKVLGEGDKEDVK</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KEKEEV</td>
<td>7-11</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>KEKEEV</td>
<td>3-9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ESEE</td>
<td>17-26</td>
</tr>
<tr>
<td>PF3D7_0532400</td>
<td>LYMP</td>
<td>NKKVRGA</td>
<td>EKKDKVLGEGDKEDVK</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>ENKKAGT</td>
<td>5-7</td>
</tr>
<tr>
<td>PF3D7_1102300</td>
<td>N/A</td>
<td>ERKEREEREKK</td>
<td>EKKDKVLGEGDKEDVK</td>
<td>9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EREKREKEKEKE</td>
<td>13-14</td>
</tr>
<tr>
<td>PF3D7_1148700</td>
<td>GEXP12</td>
<td>KECPNEMC</td>
<td>KEKREEREKK</td>
<td>8</td>
</tr>
<tr>
<td>PF3D7_1201000</td>
<td>PHISTb/c</td>
<td>KECVPNECMK</td>
<td>KEKREEREKK</td>
<td>18-36</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>EKDEK</td>
<td>7-8</td>
</tr>
<tr>
<td>PF3D7_1476200</td>
<td>PHISTb</td>
<td>KEQKEKERRK</td>
<td>KEQKEKERRKKE</td>
<td>4</td>
</tr>
<tr>
<td>PF3D7_1301400</td>
<td>HYP12</td>
<td>KKEKQE</td>
<td>KKEKQE</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>NEDE</td>
<td>12</td>
</tr>
</tbody>
</table>

Table 1. Repetitive sequences in *Plasmodium falciparum* proteins

The repeat consensus for each protein was determined from *P. falciparum* 3D7 strain sequences using the programme XSTREAM, with gaps removed for clarity. The phases of some repeats were adjusted to those described in the literature. (*) Where the repeating motif differs between isolates the alternative repeating motifs are represented after the 3D7 consensus sequence. Variation in repeat numbers seen in 16 different *P. falciparum* strains (3D7, IT, HB3, DD2, 7G8, CD1, GA1, GB4, GN1, KE1, KH1, KH2, ML1, SD1, SN1, TG1, where two-letter codes represent the country of origin for the 11 field isolates) were determined manually using Jalview with TCOFFEE used to align the sequences. (**) AP2-G also contains several poly-N sequences which vary by up to 12 residues. Long-read PACBIO genome sequencing data was obtained from the Pf3K consortium.
A. Strand slippage during DNA replication

DNA polymerase

new strand

template strand

Dissociation

Re-annealing

DNA synthesis

Expansion of repeats in new strand

Contraction of repeats in new strand

B. Unequal crossing over between homologous chromosomes during meiosis

Unequal crossing over between repeats R2 and R3

Repeat contraction

Repeat expansion

Davies et al. Figure 1
Davies et al. Figure 3
Davies et al. Figure 4
Davies et al. Figure 5